

Лекция 14. Анализ и прогнозирование временных рядов

Прогнозирование одна из самых востребованных задач. И это неудивительно: зная, пусть даже с определенной погрешностью, характер развития событий в будущем, можно принимать более обоснованные управленческие решения, планировать деятельность, разрабатывать соответствующие комплексы мероприятий, эффективно распределять ресурсы и т.д. С точки зрения технологий анализа данных прогнозированию может рассматриваться как определение некоторой неизвестной величины по набору связанных с ней значений. Поэтому прогнозирование выполняется с помощью таких задач Data Mining, как регрессия, классификация и кластеризация.

Продажи, поставки, заказы это процессы, распределенные во времени. Данные, собираемые и используемые для разработки прогнозов, чаще всего представляют собой временные ряды, то есть описывают развитие того или иного бизнес процесса во времени. Следовательно, прогнозирование в области продаж, сбыта и спроса, управления материальными запасами и потоками обычно связано именно с анализом временных рядов. Это объясняет, почему обработке временных рядов в бизнес аналитике и Data Mining уделяется такое большое внимание.

Все методы прогнозирования можно разделить на три большие группы формализованные, эвристические и комплексные.

Формализованные методы позволяют получать в качестве прогнозов количественные показатели, описывающие состояние некоторого объекта или процесса. При этом предполагается, что анализируемый объект или процесс обладает свойством инерционности, то есть в будущем он продолжит развиваться в соответствии с теми же законами, по которым развивался в прошлом и существует в настоящем.

Недостатком формализованных методов является то, что для прогноза они могут использовать только исторические данные, находящиеся в пределах эволюционного цикла развития объекта или процесса. Поэтому такие методы пригодны лишь для оперативных и краткосрочных прогнозов. К формализованным методам относятся экстраполяционные и регрессионные методы, методы математической статистики, факторный анализ и др.

Эвристические методы основаны на использовании экспертных оценок. Эксперт (группа экспертов), опираясь на свои знания в предметной области и практический опыт, способен предсказать качественные изменения в поведении исследуемого объекта или процесса. Эти методы особенно полезны в тех случаях, когда поведение объектов и процессов, для которых требуется дать прогноз,

характеризуется большой степенью неравномерности, Если формализованные методы в силу присущих им ограничений используются для оперативных и краткосрочных прогнозов, то эвристические методы чаще применяются для среднесрочных и перспективных,

Комплексное прогнозирование использует комбинацию формализованного подхода с экспертными оценками, что в некоторых случаях позволяет добиться наилучшего результата.

Независимо от методики на эффективность прогноза в наибольшей степени влияет то, насколько он полезен для планирования и ведения бизнеса. Прогноз полезен только тогда, когда его компоненты тщательно продуманы и ограничения, содержащиеся в нем, представлены явно.

Прогнозирование очень широкое понятие, как отмечалось ранее, в большинстве случаев оно связывается с предвидением во времени, с предсказанием дальнейшего развития событий. аналитики.

Временной ряд и его компоненты

Временной ряд представляет собой последовательность наблюдений за изменениями во времени значений параметров некоторого объекта или процесса, строго говоря, каждый процесс непрерывен во времени, то есть некоторые значения параметров этого процесса существуют в любой момент времени. Например, если вы запросите в банке текущий курс валют, вам никогда не ответят, что в настоящее время курса валют нет. Возможно, новый курс установился минуту назад, возможно, через час он изменится, но в настоящее время какое-то его значение непременно существует.

Для задач анализа не нужно знать значения параметров объектов в любой момент времени. Интерес представляют *временные отсчеты* значения, зафиксированные в некоторые, обычно равноотстоящие моменты времени. Отсчеты могут браться через различные промежутки: через минуту, час, день, неделю, месяц или год в зависимости от того, насколько детально должен быть проанализирован процесс. В задачах анализа временных рядов мы имеем дело с дискретным временем, когда каждое наблюдение за параметром образует временной отсчет. Непрерывный и дискретный временные ряды иллюстрируются на рис. 1.

Непрерывный ряд – это процесс, значения которого известны в любой момент времени, а дискретный ряд процесс, значения которого известны только в заданные моменты времени (временные отсчеты). Непрерывные данные – это данные, которые могут принимать бесконечное множество значений. Непрерывными могут быть только числовые данные. Дискретные данные могут

принимать ограниченный набор заранее определенных значений (категорий).

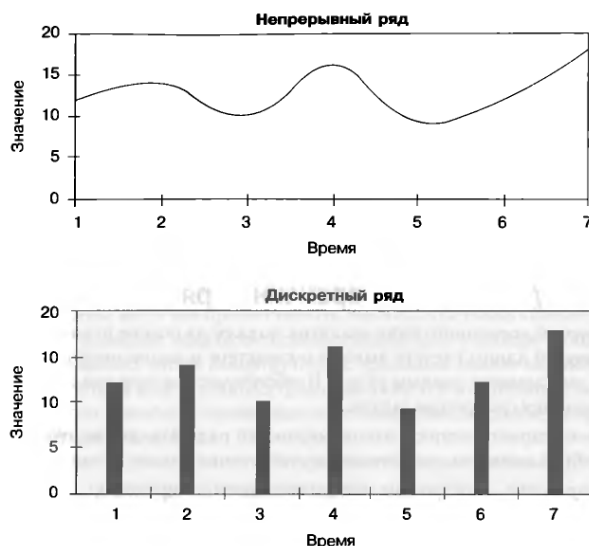


Рис. 1. Непрерывный и дискретный временные ряды

Временные ряды бывают одномерные и многомерные. Одномерные ряды содержат наблюдения за изменением только одного параметра исследуемого процесса или объекта, а многомерные за двумя параметрами или более.

Цели и задачи анализа временных рядов

При изучении временного ряда аналитик должен на основе некоторого отрезка ряда конечной длины сделать выводы о характере и закономерностях процесса, который описывается данным рядом. Наиболее часто в ходе анализа временных рядов решаются следующие задачи:

- описание характеристик и закономерностей ряда. На основе этого описания могут быть выявлены свойства соответствующих бизнес процессов;
- моделирование построение модели исследуемого процесса;
- прогнозирование предсказание будущих значений временного ряда;
- управление. Зная свойства временных рядов, можно выработать методы воздействия на соответствующие бизнес процессы для управления ими.

Наиболее востребованными в бизнес аналитике являются задачи моделирования и прогнозирования. Они взаимосвязаны: чтобы составить прогноз, сначала нужно построить соответствующую модель, проверить степень ее адекватности и правильно применить к имеющимся наблюдениям.

При решении данных задач можно столкнуться с такими проблемами, как: недостаточное количество наблюдений, содержащихся во временном ряду; статистическая изменчивость ряда, которая приводит к тому, что прошлые значения ряда устаревают и обесцениваются с точки зрения выявления текущих закономерностей и возможности прогнозирования.

Детерминированная и случайная составляющая временного ряда

Процесс разрабатывается и направляется осмысленно в соответствии с теми или иными целями, поэтому в его поведении должны присутствовать определенные закономерности. Некоторые процессы протекают равномерно, не отклоняясь от намеченных показателей. Например, завод должен производить ровно столько изделий, сколько требуется заказчикам, поэтому, если спрос стабильный, то и производство необходимо поддерживать на одном уровне. Увеличивать или снижать его нет смысла, поскольку в первом случае будет иметь место работа «на склад», а во втором упущенная прибыль и потерянные заказчики.

Если спрос устойчиво возрастает с течением времени, то и производство должно наращиваться пропорционально. Если спрос характеризуется периодичностью, вызванной сезонными колебаниями, то и производство должно подстраиваться под них. Возьмем мороженое и шипованные автопокрышки: первое лучше продается в теплое время, а вторые в зимнее. Попытка пойти против данной закономерности вряд ли будет способствовать коммерческому успеху.

Но любой бизнес-процесс сталкивается с воздействием различных случайных факторов. Диапазон таких факторов очень широк от стихийных бедствий и пожаров до банального хищения на складе готовой продукции. При этом случайные факторы могут как препятствовать, так и способствовать бизнесу. Например, аномально холодная зима будет способствовать взлету продаж отопительных приборов, а аномально теплая испортит бизнес продавцам зимних автопокрышек.

Общим свойством всех случайных факторов является то, что они не поддаются прогнозированию и вносят во временные ряды изменения, не соответствующие их основным закономерностям.

Можно выделить две составляющие временного ряда *закономерную (детерминированную) и случайную (стохастическую)*.

Закономерная (детерминированная) составляющая временного ряда последовательность значений, элементы которой могут быть вычислены в соответствии с определенной функцией. Закономерная составляющая временного ряда отражает действие известных факторов и величин.

Зная функцию, описывающую закономерность, в соответствии с которой развивается исследуемый процесс, мы можем вычислить значение детерминированной составляющей в любой момент времени.

Случайная (стохастическая) составляющая временного ряда последовательность значений, которая является результатом воздействия на

исследуемый процесс случайных факторов. Случайная составляющая и ее влияние на временной ряд могут быть оценены только с помощью статистических методов.

Пример соотношения детерминированной и случайной составляющих представлен на рис. 2. Здесь детерминированная составляющая изображена плавной темной линией, а случайная быстро изменяющейся светлой линией.

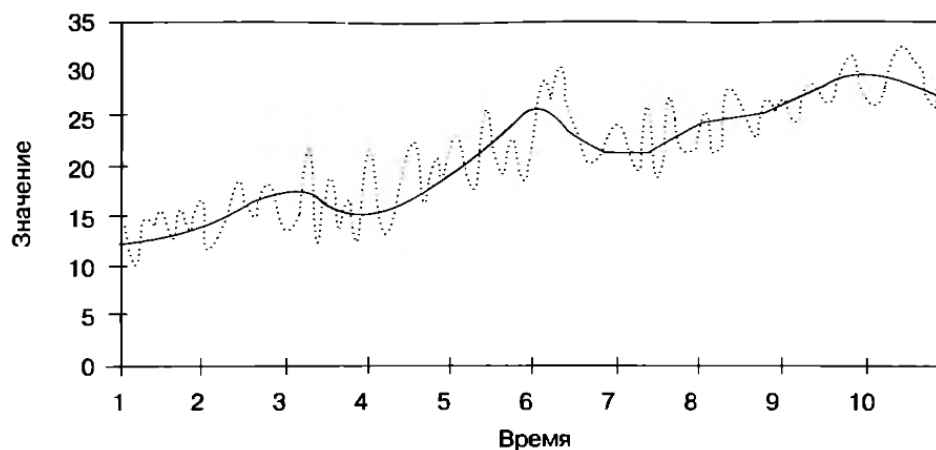


Рис. 2. Детерминированная и случайная составляющие

Случайная составляющая не существует отдельно от детерминированной. Она проявляется только как результат воздействия набора случайных факторов на исследуемый процесс и обычно выражается в повышенной изменчивости временного ряда, а также в отклонении значений детерминированной составляющей. Проще говоря, результирующее значение временного ряда – это результат взаимодействия детерминированной и случайной составляющих. Простейшим видом такого взаимодействия является случай, когда каждое значение временного ряда можно рассматривать как сумму (разность) двух значений, одно из которых обусловлено детерминированной составляющей, а другое случайной.

Модели временных рядов

Наблюдаемые значения временного ряда представляют собой результат взаимодействия детерминированной и случайной составляющих. Различают два вида такого взаимодействия:

аддитивное значения временного ряда получаются как результат сложения детерминированной и случайной составляющих;

мультипликативное значения временного ряда получаются как результат умножения детерминированной и случайной составляющих.

Соответственно модели временных рядов также бывают аддитивные и мультипликативные.

Аддитивная модель имеет вид:

$$x_i = d_i + p_i \text{ ИЛИ } X = D + P$$

где $i = 1, \dots, n$ – номер временного отсчета.

Мультипликативная модель имеет вид:

$$x_i = d_i \times p_i \text{ ИЛИ } X = D \times P$$

Очень важно знать характер взаимодействия детерминированной и случайной составляющих, поскольку методика анализа временных рядов зависит от используемой модели.

Компоненты временного ряда

Количество разнообразных процессов в экономике, управлении, бизнесе, социальной и государственной сфере весьма велико, и поведение временных рядов, описывающих эти процессы, может существенно различаться. Поэтому для описания поведения временных рядов были введены три компоненты, своего рода типовые структуры, которые можно выделить во временном ряду – тренд, сезонная компонента и циклическая компонента.

С учетом указанных компонент детерминированная составляющая ряда может быть записана в виде:

$$d_i = t_i + s_i + c_i$$

где t_i – тренд; s_i – сезонная компонента; c_i – циклическая компонента; $i = 1, \dots, n$ – номер временного отсчета.

Тренд

Тренд – наиболее важная компонента временного ряда. Именно с выделения тренда чаще всего и начинается анализ временного ряда.

Тренд – медленно меняющаяся компонента временного ряда, которая описывает влияние на временной ряд долговременно действующих факторов, вызывающих плавные и длительные изменения ряда.

Действительно, среди всех факторов, влияющих на экономические и бизнес процессы, выделяются быстродействующие и медленнодействующие. Быстродействующие факторы, такие как стихийное бедствие, «обвал» фондового рынка и т. д, могут изменить ситуацию в течение нескольких дней или даже часов. Медленнодействующие факторы могут изменять ситуацию в течение нескольких месяцев и даже лет.

Чтобы представить характер тренда, обычно бывает достаточно взглянуть на график временного ряда. Для описания тренда используются различные модели, наиболее популярными из которых являются следующие:

- *Простая линейная модель:* $t_i = a + b \times i$, где $i = 1, \dots, n$ – номер временного отсчета (элемента ряда). Пример линейного тренда представлен на рис. 3.

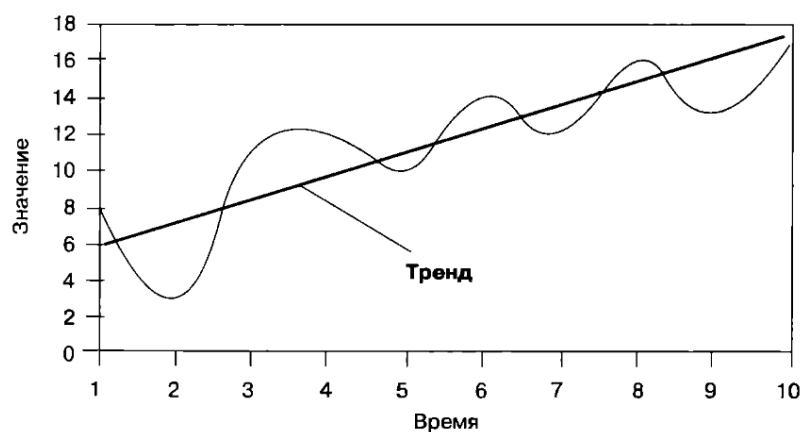


Рис. 3. Линейный тренд

Несмотря на свою простоту, линейная модель тренда часто оказывается полезной при решении реальных задач анализа, поскольку многие бизнес-процессы линейны по своей природе.

- *Полиномиальная модель:* $t_i = a + b_1 \times i + b_2 \times i^2 + \dots + b_n \times i^n$. В большинстве реальных задач степень полинома не превышает 5.
- *Экспоненциальная модель:* $t_i = \exp(a + b \times i)$. Используется в случаях, когда процесс характеризуется равномерным увеличением темпов роста. Пример экспоненциального тренда представлен на рис. 4 (сплошная линия).

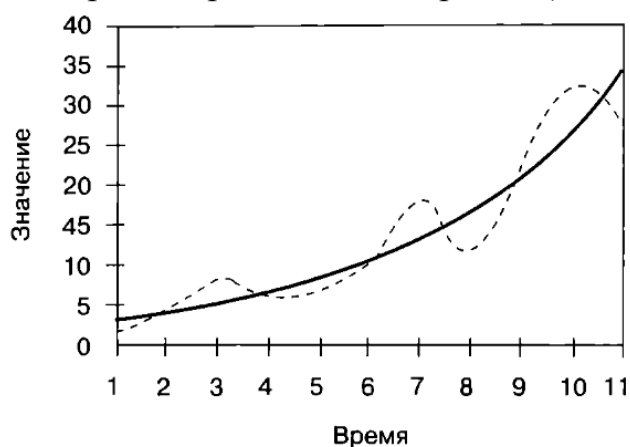


Рис. 4. Экспоненциальный тренд

Логистическая модель: $t_i = a / (1 + b \times e^{-ki})$, где k — константа, управляющая крутизной логистической функции (рис. 5).

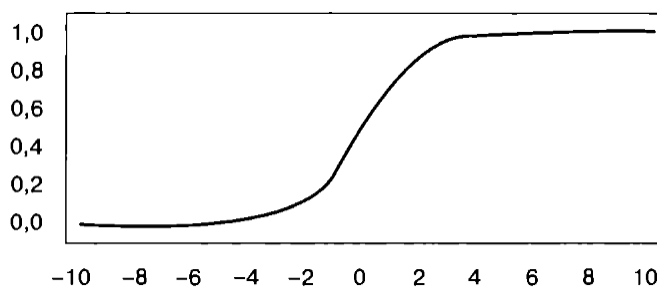


Рис. 5. Логистическая функция (сигмоида)

Такого типа кривые, имеющие S-образную форму, часто называют сигмоидами. Они хорошо описывают процессы с непостоянными темпами роста.

Сезонная компонента

Многим процессам свойственна повторяемость во времени, причем периодичность таких повторений может изменяться в очень широком диапазоне. Например, для экономики стран и регионов характерны подъемы и спады, которые могут длиться десятилетиями. В супермаркете продажи ежедневно увеличиваются в вечернее время, когда люди идут с работы. Подобных примеров можно привести множество.

Очевидно, что для описания таких периодических изменений, присутствующих во временных рядах, тренд непригоден. Поэтому вводится еще одна компонента, называемая сезонностью, или сезонной компонентой.

Сезонная компонента – составляющая временного ряда, описывающая регулярные изменения его значений в пределах некоторого периода и представляющая собой последовательность почти повторяющихся циклов.

Сезонная компонента может быть привязана к определенному календарному временному интервалу: дню, неделе, кварталу, месяцу или году – либо к какому-нибудь событию, которое прямо не соотносится с конкретными календарными интервалами.

Сезонную компоненту с изменяющимся периодом иногда называют «плавающей».

Утверждать, что сезонная компонента описывается периодической функцией, было бы неверно. Понятно, что регулярно повторяющиеся всплески продаж, образующие сезонную компоненту, не будут воспроизводиться абсолютно точно (как, например, значения функций синуса или косинуса). Тем не менее сезонная компонента обычно хорошо прослеживается на графике временного ряда.

Циклическая компонента

Часто временные ряды содержат изменения, слишком плавные и заметные для случайной составляющей. В то же время такие изменения нельзя отнести ни к тренду, поскольку они не являются достаточно протяженными, ни к сезонной компоненте, поскольку они не являются регулярными. Подобные изменения называются циклической компонентой временного ряда. Она занимает промежуточное положение между детерминированной и случайной составляющими временного ряда.

Циклическая компонента временного ряда – интервалы подъема или спада, которые имеют различную протяженность, а также различную амплитуду

расположенных в них значений.

Наличие в рядах данных циклических компонент связано с тем, что в пределах интервалов более глобальных изменений (например, сезонных) могут наблюдаться не имеющие периодичности временные подъемы и спады, которые, в отличие от случайной компоненты, не вызваны действием случайных факторов, а являются особенностями бизнеса и обусловлены общеэкономической ситуацией. На графике 6 циклическая компонента выглядит как плавные волнообразные флуктуации вокруг тренда.

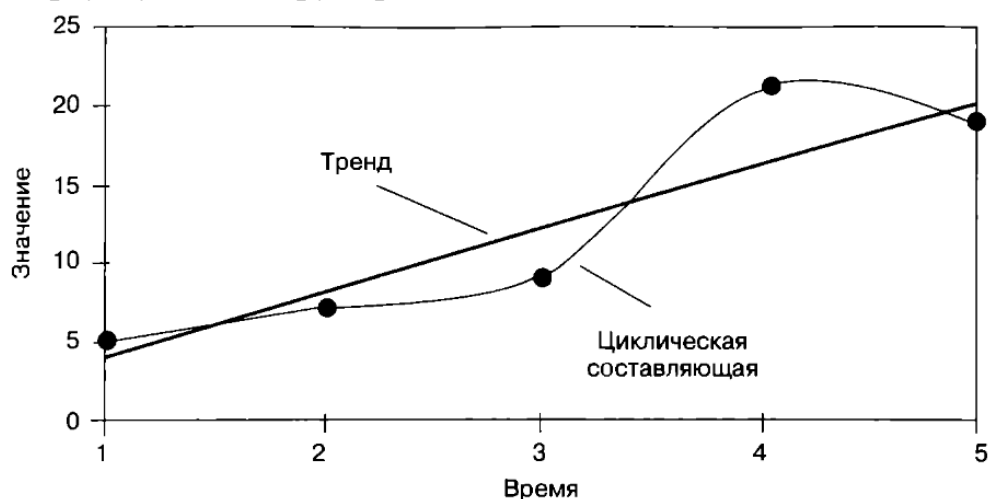


Рис. 6. Циклическая компонента

Выделить во временных рядах циклическую компоненту формальными методами довольно сложно. Обычно для этого используют информацию из других временных рядов, которые связаны с исследуемым. Например, объяснить колебания ряда в пределах сезонных интервалов можно инфляцией, изменениями на рынке трудовых ресурсов и т. д. Но для этого необходимо привлечь соответствующую информацию.

Изучение циклической компоненты часто оказывается полезным для прогнозирования, особенно краткосрочного. Таким образом, временной ряд можно представить, как композицию, состоящую из двух составляющих: случайной и детерминированной. Детерминированная составляющая, в свою очередь, содержит три компоненты: тренд, сезонную и циклическую.

Исследование временных рядов и автокорреляция

Цель анализа временного ряда – построение его математической модели, с помощью которой можно обнаружить закономерности поведения ряда, а также построить прогноз его дальнейшего развития.

Главной проблемой при построении таких моделей является нестационарность ряда.

Временной ряд называется *стационарным*, если его статистические

свойства (математическое ожидание и дисперсия) одинаковы на всем протяжении ряда.

Если статистические свойства для различных интервалов ряда существенно различаются, то такой ряд называется *нестационарным*. Применение к нестационарным рядам различных методов анализа, в том числе статистических, затруднено. Поэтому, прежде чем приступать к построению модели ряда, его стараются свести к стационарному.

При исследовании временного ряда, как правило, ищут ответы на несколько вопросов.

- Является ли ряд данных случайным?
- Содержит ли временной ряд тренд и сезонную компоненту?
- Является ли временной ряд стационарным?

Для ответа используется аппарат корреляционного анализа. *Корреляция* – это понятие математической статистики, которое характеризует степень статистической взаимосвязи между элементами данных. Если взаимосвязь между элементами данных присутствует, то такие данные называются коррелированными, в противном случае некоррелированными.

Если определяется корреляция между двумя временными рядами, то говорят о взаимной корреляции. Когда устанавливается степень статистической зависимости между значениями одного временного ряда, имеет место автокорреляция. В этом случае вычисляется корреляция между временным рядом и его копией, сдвинутой на один или несколько временных отсчетов.

Смысл корреляционного анализа заключается в следующем. Детерминированная составляющая временного ряда, которая описывает закономерности, присущие связанному с ним процессу, характеризуется плавными изменениями значений ряда. То есть соседние значения ряда не должны сильно отличаться, и, следовательно, между ними присутствует взаимная зависимость (элементы ряда коррелированы между собой). Если значения ряда в большей степени обусловлены случайной составляющей и соседние значения могут существенно отличаться друг от друга, то степень их взаимозависимости и, следовательно, корреляция будут меньше.

Рассмотрим понятие автокорреляции на следующем примере. Пусть дан ряд, который содержит последовательность ежемесячных наблюдений за продажами (табл.1).

Таблица 1. Продажи по месяцам

Месяц	Продажи
Январь	125

Февраль	130
Март	140
Апрель	132
Май	145
Июнь	150
Июль	148
Август	155
Сентябрь	157
Октябрь	160
Ноябрь	158
Декабрь	165

Для того чтобы вычислить автокорреляцию ряда, будем использовать его копию, сдвинутую в сторону запаздывания на определенное количество отсчетов (табл. 2).

Таблица 2. Данные для расчета автокорреляционной функции (АКФ)

X	125	130	140	132	145	150	148	155	157	160	158	165
X_{i-1}		125	130	140	132	145	150	148	155	157	160	158
X_{i-2}			125	130	140	132	145	150	148	155	157	160
...
X_{i-n}												125

Тогда ряд X_{i-1} будет представлять собой копию исходного ряда, сдвинутую на один временной отсчет, X_{i-2} – на два временных отсчета и т. д. Для определения степени взаимной зависимости элементов ряда используется величина, называемая коэффициентом автокорреляции r_k , где k – количество отсчетов, на которое был сдвинут временной ряд при вычислении данного коэффициента.

Лаг (сдвиг во времени) определяет порядок коэффициента автокорреляции. Если $L = 1$, то имеем коэффициент автокорреляции 1-го порядка $r_{k,k-1}$. Если $L = 2$, то коэффициент автокорреляции 2-го порядка $r_{k,k-2}$ и т.д.

Коэффициент автокорреляции вычисляется в соответствии с формулой:

$$r_k = \frac{\sum_{i=k+1}^n (x_i - \bar{x})(x_{i-k} - \bar{x})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 (x_{i-k} - \bar{x})^2}}$$

где x_i – значение 1-го отсчета; x_{i-k} – наблюдение x_i со сдвигом на k временных отсчетов; \bar{x} – среднее значение ряда, которое рассчитывается по формуле

$$\bar{x}_{k-l} = \frac{\sum_{t=k}^n x_t}{n-l}$$

Коэффициент корреляции изменяется в диапазоне $[-1; 1]$, где $r_k = 1$ указывает на полную корреляцию.

Составим таблицу 3 вспомогательных расчетов.

Таблица 3. Вспомогательные расчеты

k	x_i	x_{i-1}	$x_i - \bar{x}_i$	$x_{i-1} - \bar{x}_{i-1}$	$(x_i - \bar{x}_i) \times (x_{i-1} - \bar{x}_{i-1})$	$(x_i - \bar{x}_i)^2$	$(x_{i-1} - \bar{x}_{i-1})^2$
1	125	-	-	-	-	-	-
2	130	125	-19	-20,5	385,5	361	420,25
3	140	130	-9	-15,5	139,5	81	240,25
4	132	140	-17	-5,5	93,5	289	30,25
5	145	132	-4	-13,5	54	16	182,25
6	150	145	1	-5,5	-5,5	1	30,25
7	148	150	-1	0,5	-0,5	1	0,25
8	155	148	6	2,5	15	36	6,25
9	157	155	8	4,5	36	64	20,25
10	160	157	11	11,5	126,5	121	132,25
11	158	160	9	14,5	130,5	81	210,25
12	165	158	16	2,5	40	256	6,25
итого	1640	1600	-	-	1014,5	1307	1385,95

Рассчитаем среднее значения

$$\bar{x}_i = \frac{\sum_{t=2}^n x_t}{n-1} = \frac{1640}{11} = 149 \text{ (первое значение 125 не берем)}$$

$$\bar{x}_{i-1} = \frac{\sum_{t=2}^n x_{t-1}}{n-1} = \frac{1600}{11} = 145,5$$

Определим коэффициент автокорреляции уровней ряда первого порядка:

$$r_1 = \frac{\sum_{i=k+1}^n (x_i - \bar{x})(x_{i-1} - \bar{x})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (x_{i-1} - \bar{x})^2}} = \frac{1014,5}{\sqrt{1307 \times 1385,95}} = 0,87$$

Аналогично можно определить коэффициенты автокорреляции второго и более высоких порядков. Например, для второго порядка формулы будут следующими:

$$\bar{x}_2 = \frac{\sum_{t=k}^n x_t}{n-2}$$

$$r_2 = \frac{\sum_{i=k+1}^n (x_i - \bar{x})(x_3 - \bar{x})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=3}^n (x_i - \bar{x})^2}}$$

Если рассчитать коэффициенты автокорреляции для каждого сдвига, то

получим последовательность коэффициентов, называемую автокорреляционной функцией (АКФ).

Результат расчета АКФ для ряда X представлен в табл. 4.

Таблица 4. Результаты расчета АКФ для ряда X

k	0	1	2	3	4	5	6	7	8	9	10
r_k	1	0,87	0,86	0,84	0,79	0,71	0,64	0,53	0,02	0	0

Значение коэффициента автокорреляции при нулевом сдвиге r_0 равно 1, поскольку определяется автокорреляция двух одинаковых копий временного ряда, а ряд полностью коррелирован сам с собой. Также наблюдается высокая степень корреляции $r_k > 0,8$ при сдвиге менее чем на 4 временных отсчета и умеренная при r_k , изменяющемся на 5 - 7 отсчетов. При сдвиге более чем на 7 отсчетов коэффициент корреляции быстро падает. Такой результат вполне объясним. Значения ряда, расположенные близко друг к другу, имеют высокую степень взаимной зависимости, что обеспечивает высокий коэффициент автокорреляции при малом сдвиге (в пределах 4 отсчетов). По мере удаления значений степень их взаимной зависимости падает, что приводит к уменьшению коэффициента автокорреляции при сдвиге более чем на 5 отсчетов. и наконец, при сдвиге более чем на 7 отсчетов коэффициент автокорреляции устремляется к нулю, что говорит о резком уменьшении взаимной зависимости точек, отстоящих друг от друга более чем на 7 отсчетов.

Для ряда на основе АКФ можно сделать вывод, что в нем наблюдается высокая степень взаимной зависимости между соседними значениями. Данный вывод подтверждается визуальным исследованием ряда: в нем присутствует небольшой линейный положительный тренд, отсутствует сезонная компонента, а достаточно высокая гладкость позволяет выдвинуть предположение о малой величине случайной составляющей. Все это хорошо согласуется с выводами, сделанными на основе корреляционного анализа.

Кроме того, корреляционный анализ позволяет прийти к следующим заключениям о поведении временного ряда.

- Если ряд содержит тренд, то степень взаимной зависимости между последовательными значениями ряда и корреляция между ними очень высоки. При этом коэффициент автокорреляции значителен для первых нескольких сдвигов ряда, а в дальнейшем убывает до нуля.

- Если действие случайной компоненты велико, то коэффициенты автокорреляции для любого значения сдвига будут близки к нулю. Высокая изменчивость ряда, являющаяся следствием воздействия случайной компоненты, приводит к уменьшению взаимной связи между последовательными значениями

ряда и, соответственно, к уменьшению коэффициента автокорреляции.

- Если ряд содержит сезонную компоненту, то коэффициент автокорреляции будет большим для значений сдвига, равных периоду сезонной составляющей или кратных ему.

Таким образом, корреляционный анализ один из важнейших инструментов исследования временных рядов. Он позволяет выявлять в ряду тренд и сезонную компоненту, а также определять, насколько поведение ряда обусловлено случайной компонентой. Знание данных свойств временного ряда помогает строить более адекватные модели и выбирать методы прогнозирования.