Анализ временных рядов. Современные подходы

Статистический подход

Фундамент анализа данных как научной дисциплины.

Априорные представления о реальности проверяют с помощью известных методов и процедур. Представления основаны на теории вероятностей и математической статистике.

Программные пакеты: STATISTICA, SPSS, Statgraphics…

Синергетический подход

В основе – теория хаоса, теория циклов, теория волн Эллиота.

Опирается на фрактальные и синергетические концепции, методы технического анализа.

Программные пакеты: Metastock, Tradestation, Elwawe…

Кибернетический подход

Методы искусственного интеллекта: методы и технологии экспертных систем, нейронных сетей, генетических алгоритмов, нечёткой логики.

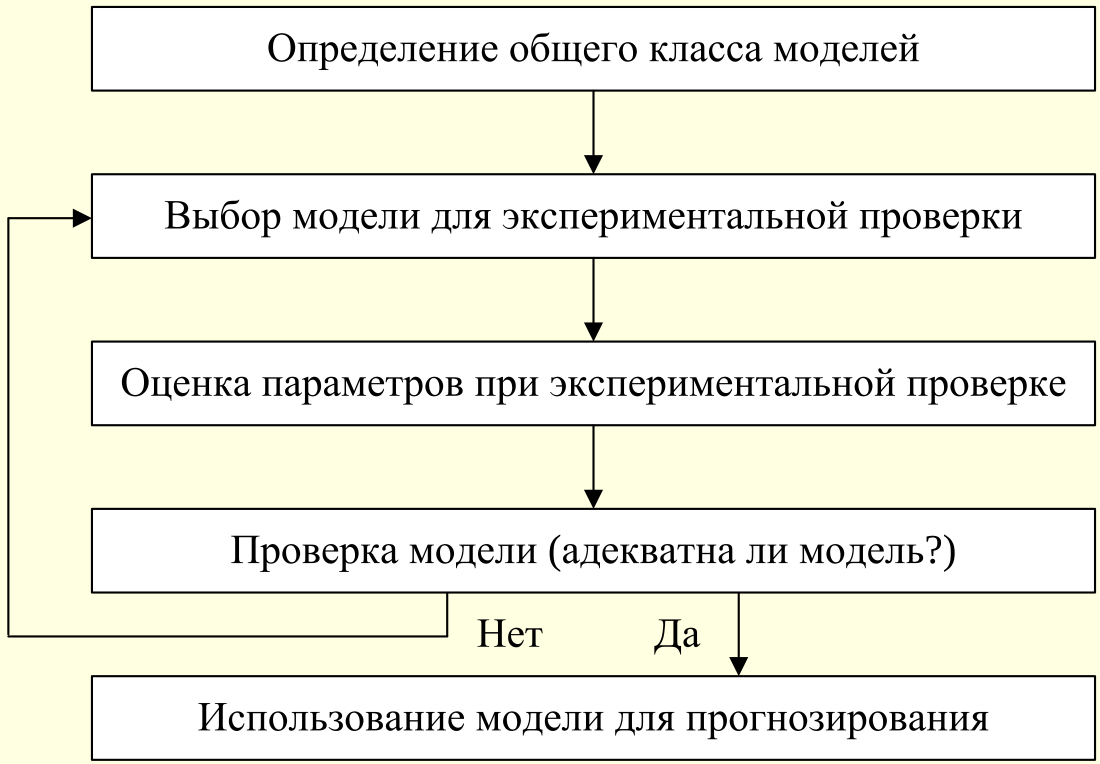
В основе – механизмы поиска подходящих алгоритмов.

Программные пакеты: Statistica Neural Networks (нейронные сети), Genehunter (генетические алгоритмы), Matlab Fuzzy Logic Toolbox (нечеткая логика)

Статистическая обработка временных рядов и прогнозирование

- Стационарные и нестационарные временные ряды
- Обнаружение нестационарности
- Модели стационарных временных рядов

Статистическая обработка временных рядов и прогнозирование



Стационарность ряда

Ряд называется строго стационарным, если совместное распределение вероятностей $Y_{t_1},\dots,Y_{t_m}$ не зависит от сдвига по времени, т. е. совпадает с распределением вероятностей $Y_{t_1+L},\dots,Y_{t_m+L}$ для любых $L, t_1,\dots,t_m$.

Слабая стационарность

Ряд называется слабо стационарным или просто стационарным, если средние, дисперсии и ковариации не зависят от времени *t*.

Таким образом, для стационарного ряда

$$MY(t) = \mu$$

$$DY(t) = \nu_0 = \sigma^2$$

$$Cov(Y_t, Y_{t+L}) = \nu_L$$

$$ACF(t, L) = \varphi(L)$$

Автокорреляционная функция ряда

$$ACF(L) = \frac{Cov(Y_t, Y_{t+L})}{DY_t} = \frac{\nu_L}{\nu_0} = \varphi_L$$

Выборочная автокорреляционная функция временного ряда называется ***коррелограммой***

и определяется следующим образом:

$$ACF(L) = \frac{\sum_{t=L+1}^{T} (Y_t - \overline{Y})(Y_{t+L} - \overline{Y})}{\sum_{t=1}^{T} (Y_t - \overline{Y})^2}$$

Свойства коэффициента автокорреляции

1. Строится по аналогии с линейным коэффициентом корреляции и характеризует тесноту **только линейной** связи текущего и предыдущего уровней ряда.

2. По знаку коэффициента автокорреляции **нельзя** сделать вывод о возрастающей или убывающей тенденции в уровнях ряда

Автокорреляционная функция ряда

- **Автокорреляция элементов временного ряда** – корреляционная зависимость между последовательными элементами временного ряда.
- **Лаг** – число периодов, по которым рассчитывается коэффициент автокорреляции между парами элементов ряда.
- **Автокорреляционная функция временного ряда** – последовательность коэффициентов автокорреляции с лагами, равными 1, 2, 3 ….

Свойства коэффициента автокорреляции

Если наиболее высоким оказался коэффициент автокорреляции первого порядка, исследуемый ряд содержит только тренд (тенденцию).

Если наиболее высоким оказался коэффициент автокорреляции порядка τ , ряд содержит циклические колебания с периодичностью в τ моментов времени.

Свойства коэффициента автокорреляции

Если ни один из коэффициентов автокорреляции не является значимым, можно сделать одно из двух предположений:

Ряд не содержит тренда и циклических колебаний и имеет случайную структуру.

Ряд содержит сильный нейтральный тренд, для выявления которого нужно провести дополнительный анализ.

Частная автокорреляционная функция

ϕ_{kk}

– «чистая корреляция» между Y_t и Y_{t+L} при исключении влияний промежуточных значений $Y_{t+1},\ldots,Y_{t+L-1}$

$$\rho_L = \frac{E\left[(Y_t-\mu)(Y_{t+L}-\mu)\right]}{\sqrt{E\left[(Y_t-\mu)^2\right]E\left[(Y_{t+L}-\mu)^2\right]}} \qquad \mu = E\left[Y_t\right]$$

$$\rho_j = \phi_{L1}\rho_{j-1} + \ldots + \phi_{L(L-1)}\rho_{j-L+1} + \phi_{LL}\rho_{j-L}$$

j–й коэффициент автокорреляционной функции

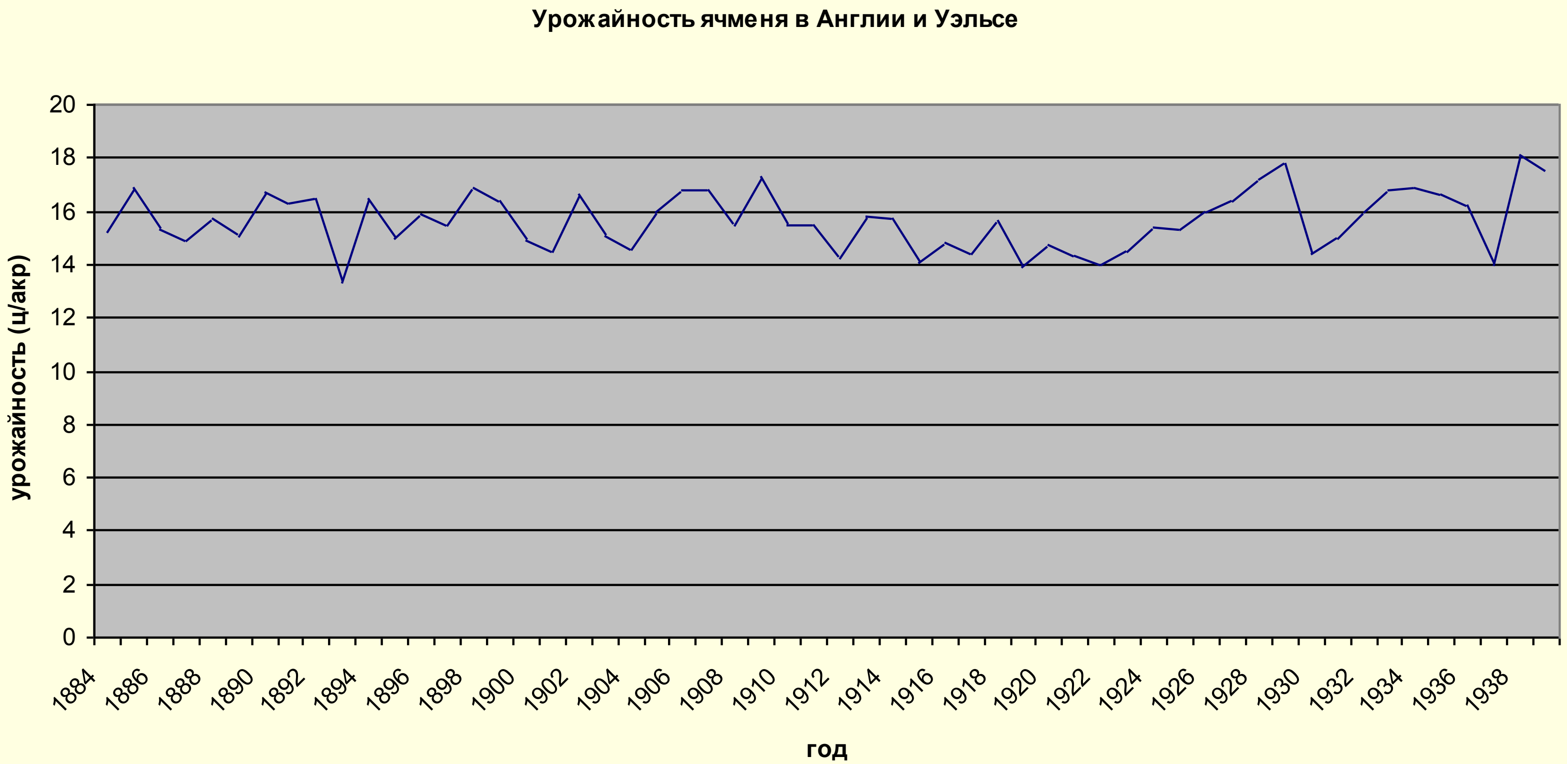
Примеры временных рядов

1) «белый шум» – все уровни временного ряда распределены одинаково.

$$Y_t = \varepsilon_t \qquad \text{где} \qquad \varepsilon_t \sim iid(0, \sigma^2)$$

$MY_t = 0$, $DY_t = \sigma^2$ $\qquad ACF(L) = \begin{cases} 1, L = 0; \\ 0, L > 0 \end{cases}$

Белый шум


Урожайность ячменя в Англии и Уэльсе
урожайность (ц/акр)
20
18
16
14
12
10
8
6
4
2
0
1884
1886
1888
1890
1892
1894
1896
1898
1900
1902
1904
1906
1908
1910
1912
1914
1916
1918
1920
1922
1924
1926
1928
1930
1932
1934
1936
1938
год

Примеры временных рядов (продолжение)

2) .авторегрессионный ряд первого порядка (AR(1))

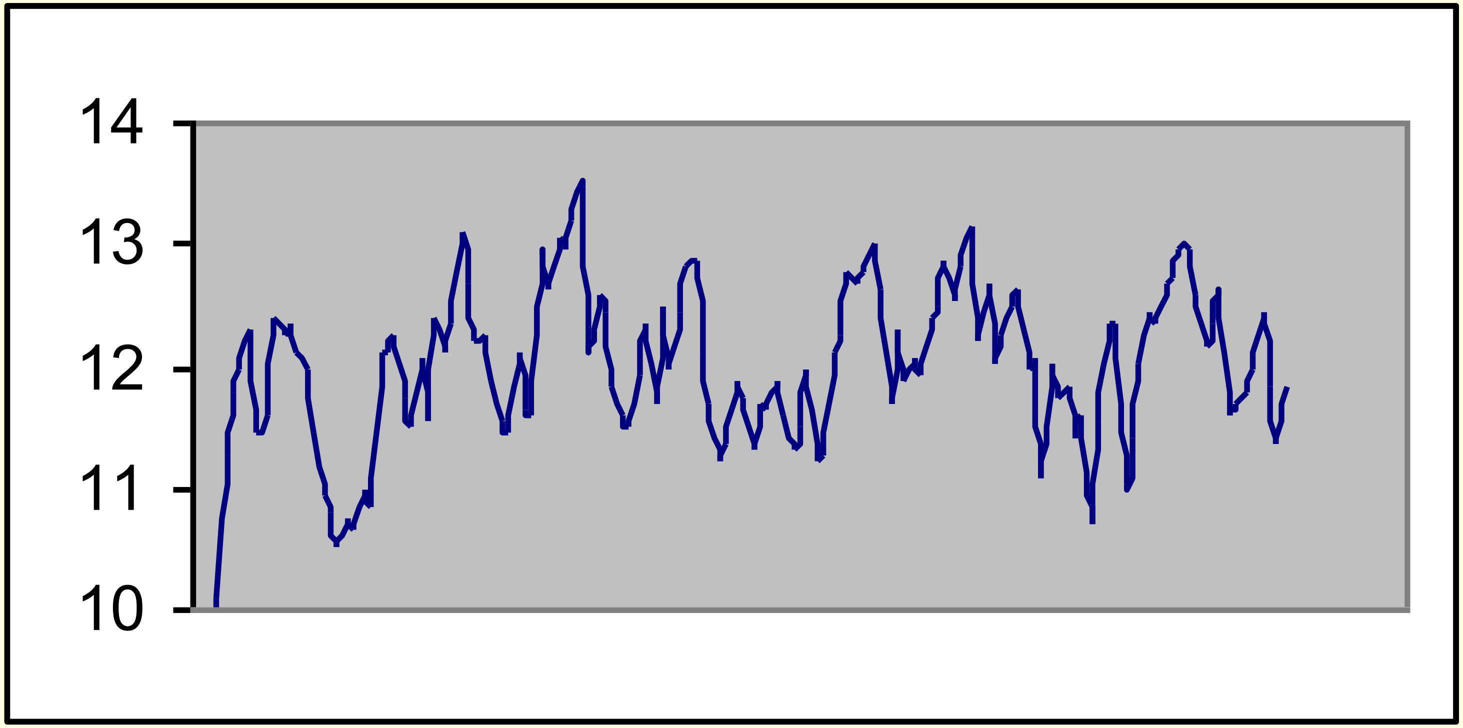
$$Y_t = \alpha + \rho Y_{t-1} + \varepsilon_t \qquad \varepsilon_t \sim iid(0, \sigma^2)$$

$$MY_t = \alpha + \rho MY_{t-1} = \alpha + \rho\alpha + \rho^2 MY_{t-2} = \alpha(1 + \rho + \rho^2 + ..) = \frac{\alpha}{1-\rho}$$

$$DY_t = \frac{\sigma^2}{1-\rho} \qquad \nu_L = \frac{\rho^L \sigma^2}{1-\rho^2}$$

если $|\rho| < 1$ ряд стационарен.

AR(1)



Примеры временных рядов (продолжение)

3) «случайное блуждание»

$$Y_t = Y_{t-1} + \varepsilon_t \qquad \varepsilon_t \sim iid(0, \sigma^2)$$

$$MY_t = MY_{t-1}$$

$$DY_t = DY_{t-1} + \sigma^2 \qquad DY_1 = \sigma^2$$

$$DY_t = t\sigma^2$$

Дисперсия ряда неограниченно возрастает со временем
Ряд не является стационарным.

Примеры временных рядов (продолжение)

4) ряд с трендом, например, линейным.

$$Y_t = \alpha + \beta t + \varepsilon_t$$

$$\varepsilon_t \sim iid(0, \sigma^2)$$

$$MY_t = \alpha + \beta t$$

ряд не является стационарным.

Примеры временных рядов (продолжение)

5) ряд с сезонной компонентой не является стационарным.

$$Y_t = S(t) + \varepsilon_t$$

$$\varepsilon_t \sim iid(0, \sigma^2)$$

Ряд с трендом и сезонностью

Трендовая нестационарность

В случае 4) и 5) методы моделирования стационарных временных рядов применяются к остаткам регрессии или к сглаженным уровням временного ряда, т.е. к уровням, очищенным от тренда, циклической и сезонной составляющей.

Обнаружение нестационарности

1. Визуальный анализ временного ряда. Возможно, временной ряд содержит видный на глаз временной тренд и сезонность (периодичную компоненту). Возможно, что разброс значений возрастает или убывает со временем (признак «случайного блуждания»). Это может служить указанием на зависимость среднего и, соответственно, дисперсии от времени. Во всех трех случаях ряд, скорее всего, не будет стационарным.

Обнаружение нестационарности (продолжение)

2. Построить график выборочной автокорреляционной функции или коррелограмму. Коррелограмма стационарного временного ряда быстро убывает со временем, быстро уходит почти в ноль после нескольких первых значений – «влияние предыдущих уровней затухает». Если график показывает, что ACF убывает медленно, с колебаниями, то ряд, скорее всего, будет нестационарным.

Избавление от нестационарности

- Выделить тренд и сезонность, т. е. неслучайную составляющую временного ряда.
- Если ряд представляет «случайное блуждание», то взятие последовательных разностей делает ряд стационарным. На практике порядок разностей, как правило, не больше двух.

Модели стационарных временных рядов

Модели Бокса-Дженкинса.

Модели авторегрессии и скользящего среднего ARMA(p,q)

ARMA(p,q)

$$Y_t = \alpha + \rho_1 Y_{t-1} + \ldots + \rho_p Y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_1 - \ldots - \theta_q \varepsilon_{t-q}$$

$$\varepsilon_t \sim iid(0, \sigma^2)$$

процесс авторегрессии порядка *p* и скользящего среднего порядка *q* –ARMA(p,q)

$$\sum_{\tau=1}^{p} \rho_\tau Y_{t-\tau}$$ -авторегрессионный член порядка *p*

$$\sum_{\tau=1}^{q} \theta_\tau \varepsilon_{t-\tau}$$ -член скользящего среднего порядка *q*

Анализ стационарных временных рядов

- Спецификация ARMA-моделей.
- Оценивание модели.
- Проверка адекватности модели.
- Прогнозирование на основе построенной модели.

Спецификация ARMA-моделей.

- Определение *p* и *q*.

Для этого можно построить графики автокорреляционной функции и частной автокорреляционной функции ряда для того, чтобы определить подходящие значения *p* и *q*. Автокорерляционные функции и частные автокорреляционные функции процессов *AR*(1), *AR*(2), *MA*(1), *MA*(2) и *ARMA*(1,1) легко опознать.

Спецификация ARMA-моделей (продолжение)

На этом этапе мы можем сформулировать несколько гипотез относительно возможных значениях порядков *p* и *q*. Для подходящих значений автокорреляционная функция остатков *ARMA*(*p*,*q*) – модели похожа на «белый шум».

Оценивание модели

В современные пакеты встроены различные методы оценивания *ARIMA* – моделей, такие как линейный или нелинейный МНК. полный или условный метод максимального правдоподобия.

Проверка адекватности модели

Необходимо проверить правильность предположений относительно параметров модели. Для этого

- проверяем статистическую значимость коэффициентов модели должны достоверно отличатся от нуля
- проверяем отсутствие автокорреляции в остатках $ARMA(p,q)$ – модели

Вопросы

- Что такое временной ряд?
- Определение сильно стационарного ряда.
- Определение слабо стационарного ряда.
- Что такое автокорреляционная функция ряда?
- Что такое тренд?
- Приведите примеры стационарных и нестационарных временных рядов.
- Как проверить стационарность ряда?
- Какие вы знаете типы нестационарных рядов? Приведите примеры.
- Чем нам грозит регрессия одного стационарного ряда на другой.
- Что такое ARMA представление стационарного ряда?
- Как подобрать адекватную ARMA модель ряда?

Введение

Временные ряды отличаются от обычных данных об одном временном срезе в том отношении, что в случае временных рядов сама последовательность наблюдений несет в себе важную информацию.

Теперь чтобы охарактеризовать совокупность данных в целом, **уже недостаточно** знать лишь типичное значение этих данных (среднее значение) или даже изменчивость этой совокупности данных (дисперсия). В этом случае желательно знать, что, скорее всего, произойдет дальше. **НУЖЕН ПРОГНОЗ!**

Введение

ПРИМЕР. Составить бюджет на следующий квартал.

Требуется достоверная оценка ожидаемого объема продаж.

Прогноз – основа для прогнозирования других показателей бюджета (возможно, с помощью регрессионного анализа).

Проанализировав временной ряд фактических квартальных объемов продажи за последние несколько лет, можно сделать прогноз, который будет представлять наиболее достоверную оценку, базирующуюся на общих тенденциях продаж, с учетом любых сезонных колебаний спроса.

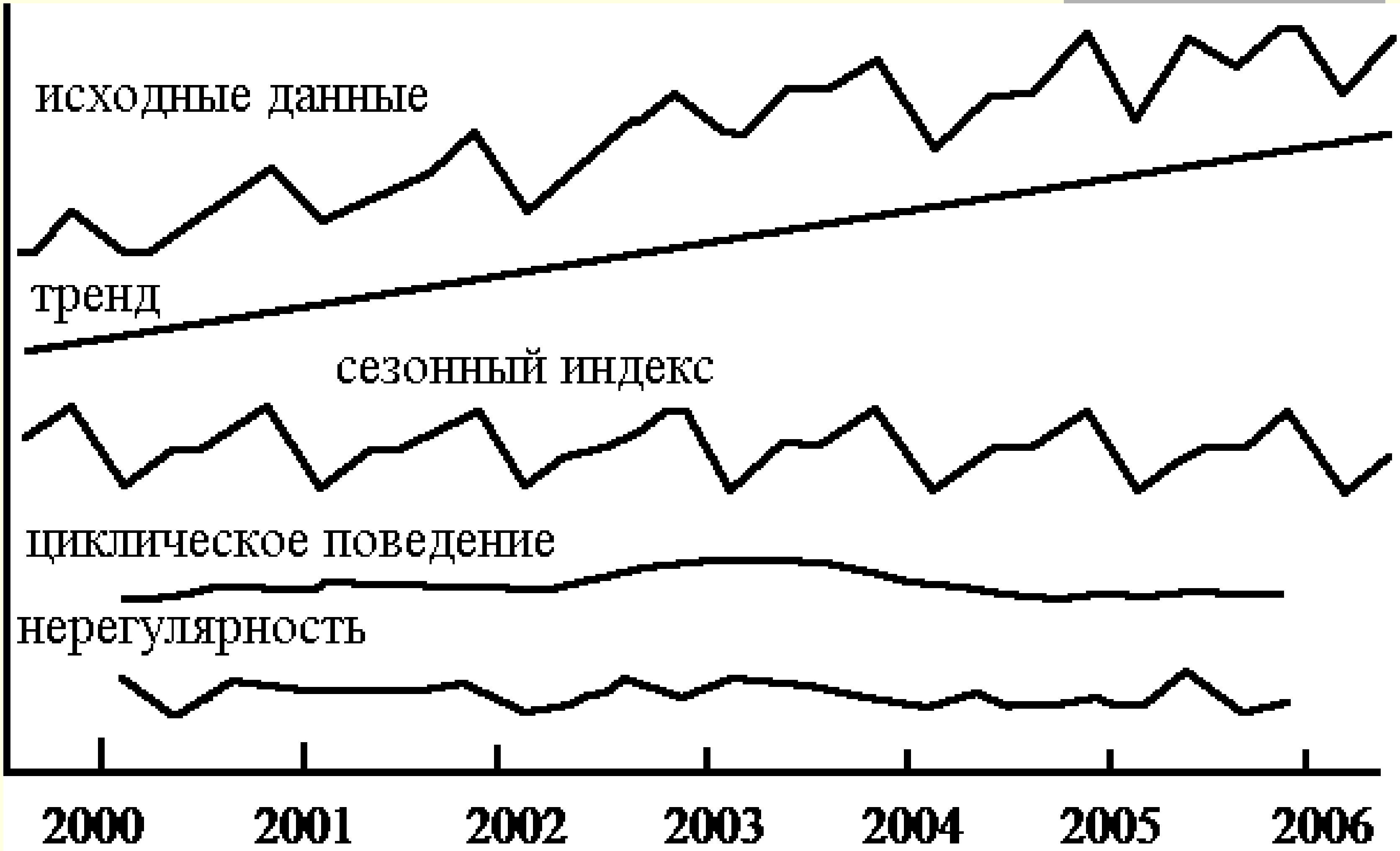
Анализ трендов и сезонности

Анализ трендов и сезонности представляет собой непосредственный, интуитивный подход к оцениванию четырех базовых компонентов помесячных или поквартальных временных рядов: долгосрочный тренд (тенденция), сезонность, циклическая вариация и нерегулярный компонент.

Базовая модель временного ряда представляет числа в этом ряде в виде произведения, получаемого путем умножения перечисленных компонентов.

$$\text{Данные} = \text{Тренд} \times \text{Сезонность} \times \text{Цикличность} \times \text{Нерегулярность}$$

Анализ трендов и сезонности



Анализ трендов и сезонности

Тренд и циклическая компонента: скользящее среднее

Скользящее среднее представляет новый ряд, полученный путем усреднения соседних наблюдений временного ряда и перехода к следующему периоду времени – в итоге получается более гладкий ряд.

$$\text{Скользящее среднее} = \text{Тренд} \times \text{Цикличность}$$

Анализ трендов и сезонности

Сезонный индекс: среднее значение отношения к скользящему среднему отражает сезонное поведение

Чтобы выделить сезонное поведение, прежде всего, следует получить отношение исходных значений к скользящему среднему. Полученный результат будет включать сезонную и нерегулярную компоненты, поскольку скользящее среднее исключает из данных тренд и циклическую компоненту.

$$(\text{Сезонность})(\text{Нерегулярность}) = \frac{\text{Данные}}{\text{Скользящее среднее}}$$

Анализ трендов и сезонности

Затем, чтобы устранить нерегулярную компоненту, надо усреднить эти значения для каждого сезона. Сезонный компонент проявляется, поскольку он присутствует ежегодно, тогда как нерегулярную компоненту, как правило, удается усреднить.

$$\text{Сезонный индекс} = \text{Среднее значение}\left(\frac{\text{Данные}}{\text{скользящее среднее}^{*}}\right)$$

$*$ - за соответствующий сезон

Анализ трендов и сезонности

Поправка на сезон: деление ряда на сезонный индекс.

Поправка на сезонные колебания устраняет из результатов измерения ожидаемый сезонный компонент (путем деления ряда на сезонный индекс для соответствующего периода), что позволяет нам непосредственно сравнивать один квартал или месяц с другим (после внесения поправки на сезон), выявляя те или иные скрытые тенденции.

$$\text{Значение с поправкой на сезон} = \left(\frac{\textit{Данные}}{\textit{Сезонный индекс}} \right) =$$

$$= \textit{Тренд} \times \textit{Цикличность} \times \textit{Нерегулярность}$$

Анализ трендов и сезонности

Долгосрочный тренд и прогноз с поправкой на сезонные колебания: линия регрессии

Когда временной ряд демонстрирует **долгосрочную** линейную тенденцию к нарастанию или снижению, для оценки этой тенденции и прогнозирования будущего можно воспользоваться **регрессионным анализом**.

Анализ трендов и сезонности

Прогноз: тренд с учетом сезонности

Чтобы прогнозировать будущее, надо учесть сезонность в долгосрочном тренде, вернув ему ожидаемую сезонную вариацию. Для этого достаточно умножить значение тренда на значение сезонного индекса для того периода времени, который вы прогнозируете. Этот процесс является обратным по отношению к внесению поправки на сезонные колебания. Результирующий прогноз включает долгосрочный тренд и сезонную вариацию.

Моделирование циклического поведения с помощью ARIMA-процессов Бокса-Дженкинса

ARIMA-процессы Бокса-Дженкинса представляют собой семейство линейных статистических моделей, основанных на нормальном распределении, которые позволяют имитировать поведение множества различных реальных временных рядов путем комбинирования процессов авторегрессии, процессов интегрирования и процессов скользящего среднего.

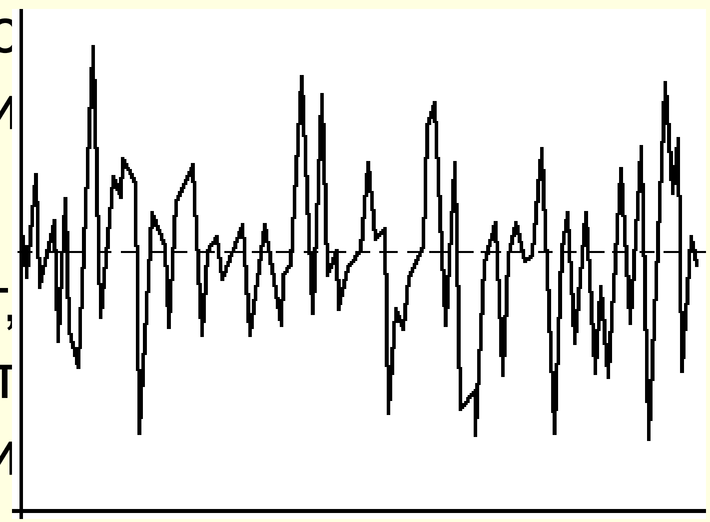
ARIMA - сокращение от **A**utoregressive **I**ntegrated **M**oving **A**verage

Моделирование циклического поведения с помощью ARIMA-процессов Бокса-Дженкинса

Процесс случайного шума не обладает памятью: отправная точка

Процесс случайного шума состоит из случайной выборки (независимых наблюдений) из нормального распределения с постоянным средним и стандартным отклонением.

$$Y_t = \mu + \xi_t$$



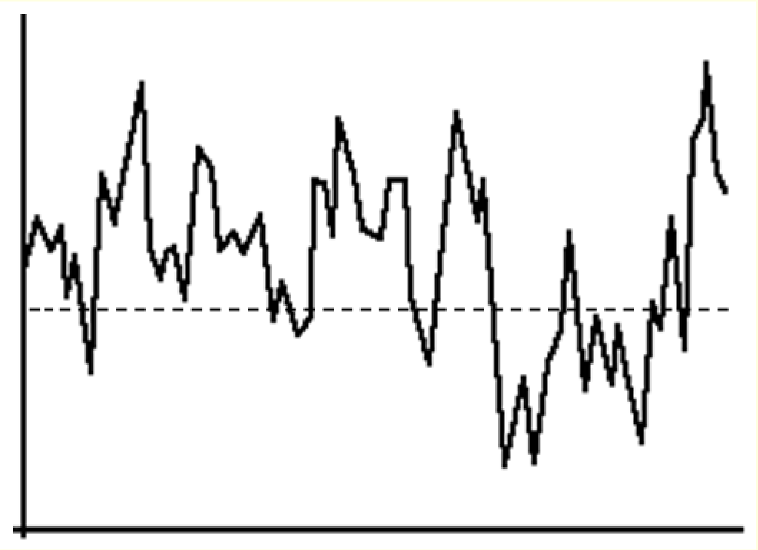
Тенденции (тренды) отсутствуют, наблюдения не содержат информации о прошлом поведении ряда.

Моделирование циклического поведения с помощью ARIMA-процессов Бокса-Дженкинса

Процесс авторегрессии (AR) обладает памятью о своем прошлом

Любое наблюдение процесса авторегрессии (AR) – линейная функция от предыдущего наблюдения плюс случайный шум.

$$Y_t = \delta + \varphi Y_{t-1} + \xi_t$$



Процесс авторегрессии «помнит» о своем предыдущем состоянии и использует эту информацию для определения своего дальнейшего поведения.

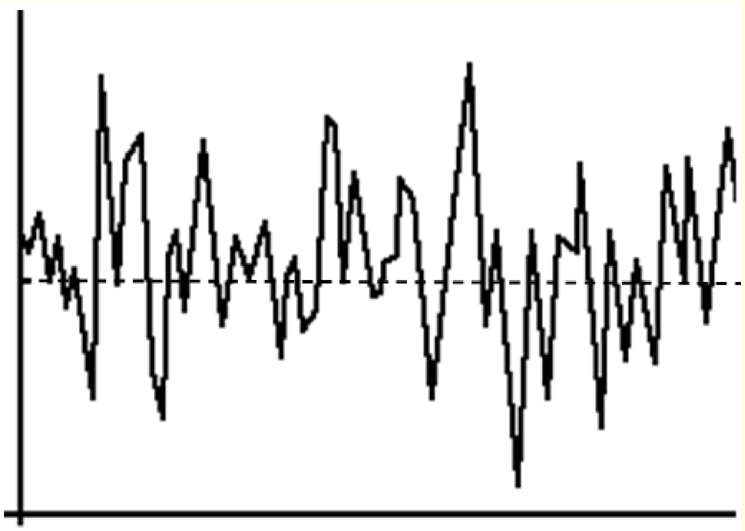
Моделирование циклического поведения с помощью ARIMA-процессов Бокса-Дженкинса

Процесс скользящего среднего (MA) имеет ограниченную память

Любое наблюдение МА состоит из константы (долгосрочное среднее значение процесса) плюс независимый случайный шум минус часть предыдущего случайного шума.

$$Y_t = \mu + \xi_t - \theta\xi_{t-1}$$

Процесс МА «не помнит» в точности своего прошлого, но помнит компонент случайного шума того состояния, в котором он (процесс) находился.



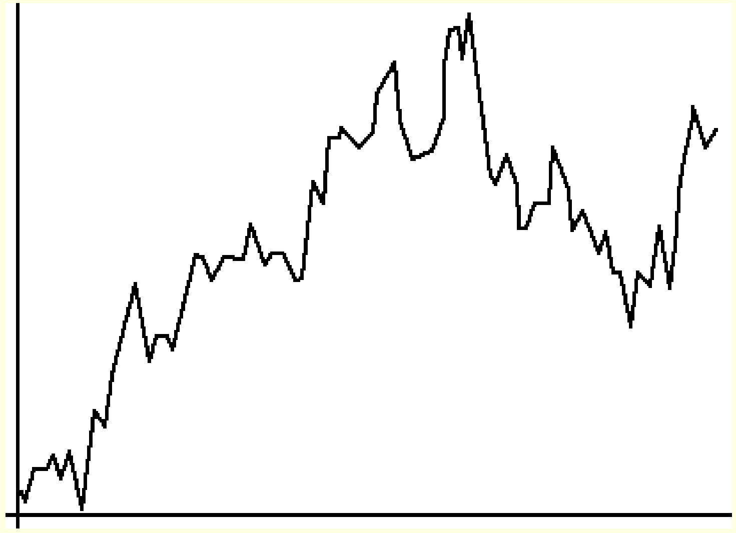
Память ограничена одним шагом в будущее; за пределами этого шага для процесса все начинается заново.

Моделирование циклического поведения с помощью ARIMA-процессов Бокса-Дженкинса

Чистый интегрированный (I) процесс помнит, где он находился, и затем движется случайно

$$Y_t = \delta + Y_{t-1} + \xi_t$$

Каждое наблюдение чистого интегрированного (I) процесса (pure integrated (I) process), называемого также случайным блужданием, заключается в случайном шаге в сторону от текущего наблюдения. Процесс знает, где он находится, но забыл, как он попал туда.

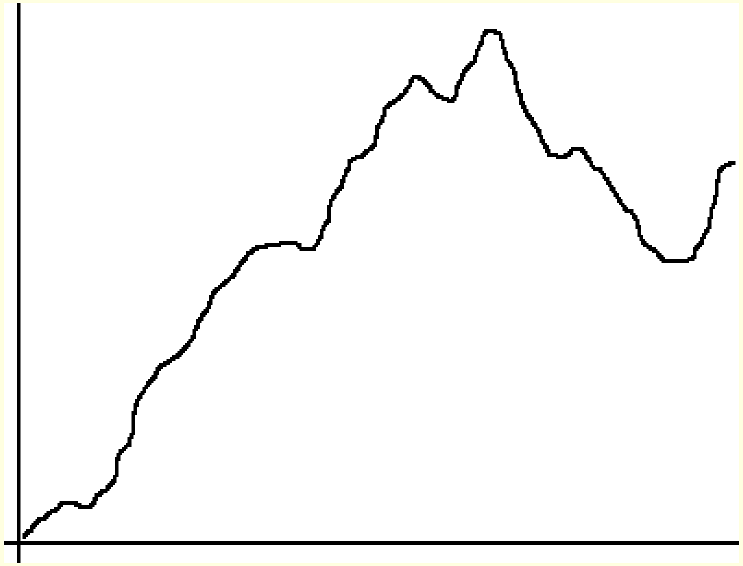


Моделирование циклического поведения с помощью ARIMA-процессов Бокса-Дженкинса

Процесс авторегрессионного интегрированного скользящего среднего (ARIMA)

$$Y_t - Y_{t-1} = \delta + \varphi(Y_{t-1} - Y_{t-2}) + \xi_t - \theta\xi_{t-1}$$

Состоит из линейной функции предыдущего изменения плюс независимый случайный шум минус определенная доля предыдущего случайного шума. Этот процесс знает, где он находится, помнит, как он попал в это состояние, и помнит даже часть предыдущего шумового компонента.

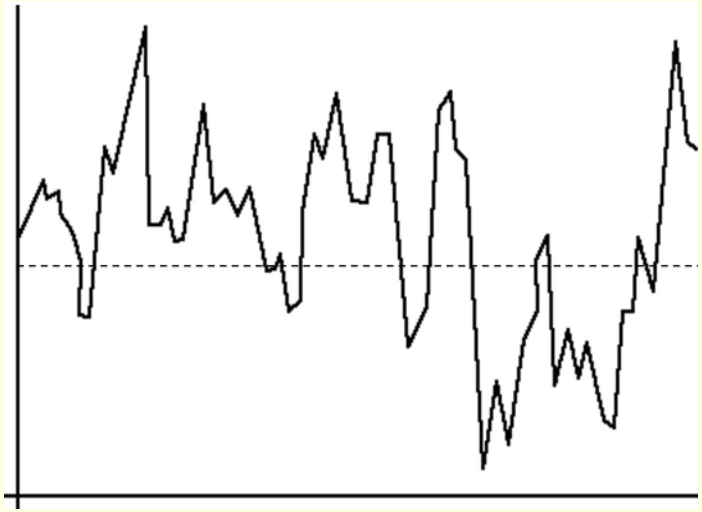


Моделирование циклического поведения с помощью ARIMA-процессов Бокса-Дженкинса

Процесс авторегрессии и скользящего среднего (ARMA)

Состоит из линейной функции от предыдущего наблюдения плюс независимый случайный шум минус некоторая доля предыдущего случайного шума.

$$Y_t = \delta + \varphi Y_{t-1} + \xi_t - \theta \xi_{t-1}$$



Запоминает как свое предыдущее состояние, так и компонент случайного шума предыдущего состояния.

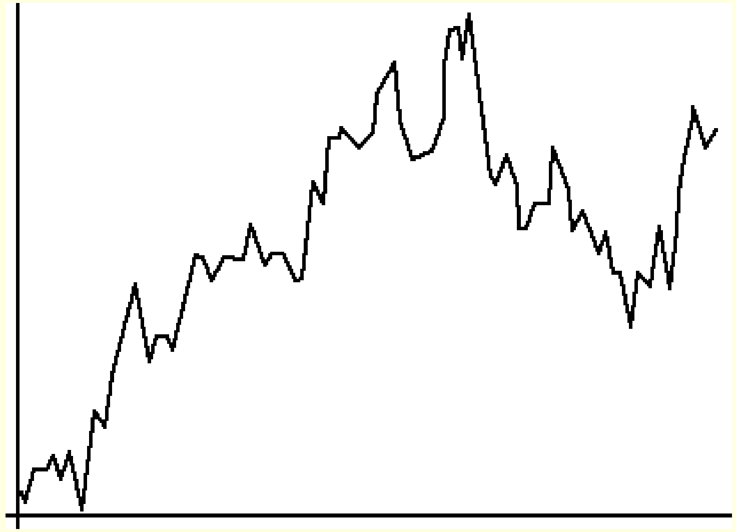
Сочетает в себе память процесса авторегрессии с памятью процесса скользящего среднего.

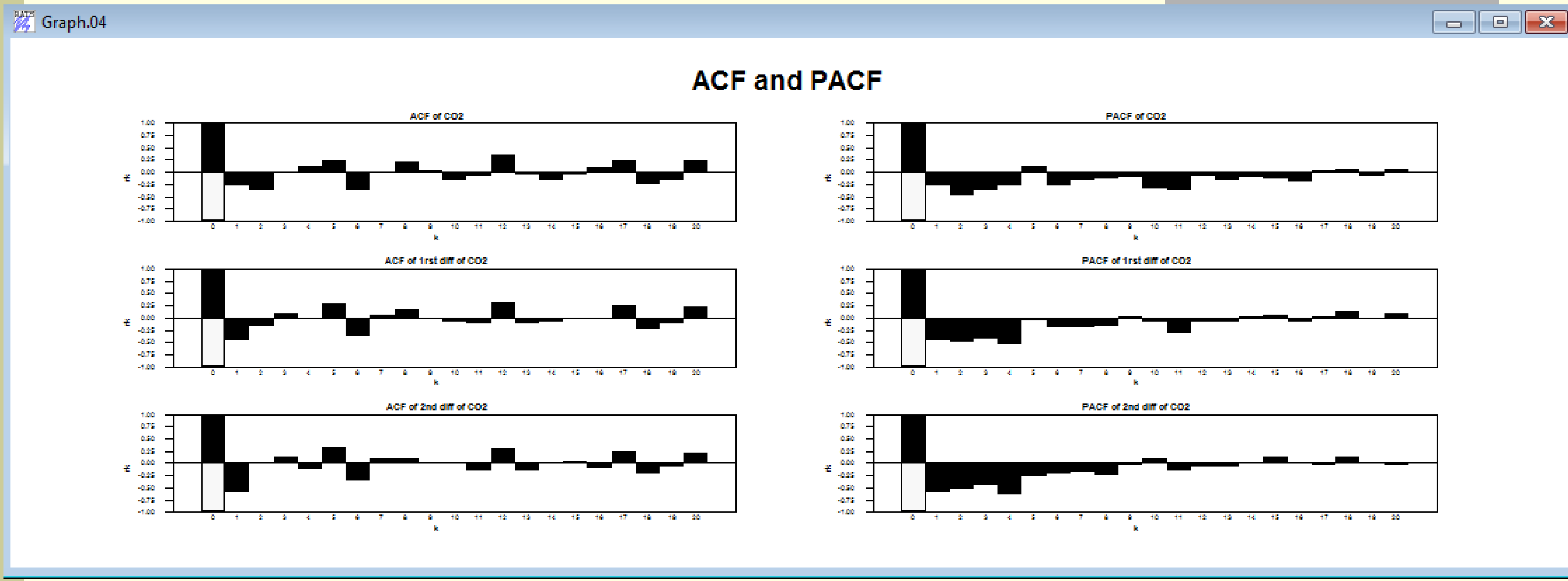
Моделирование циклического поведения с помощью ARIMA-процессов Бокса-Дженкинса

Чистый интегрированный (I) процесс помнит, где он находился, и затем движется случайно

$$Y_t = \delta + Y_{t-1} + \xi_t$$

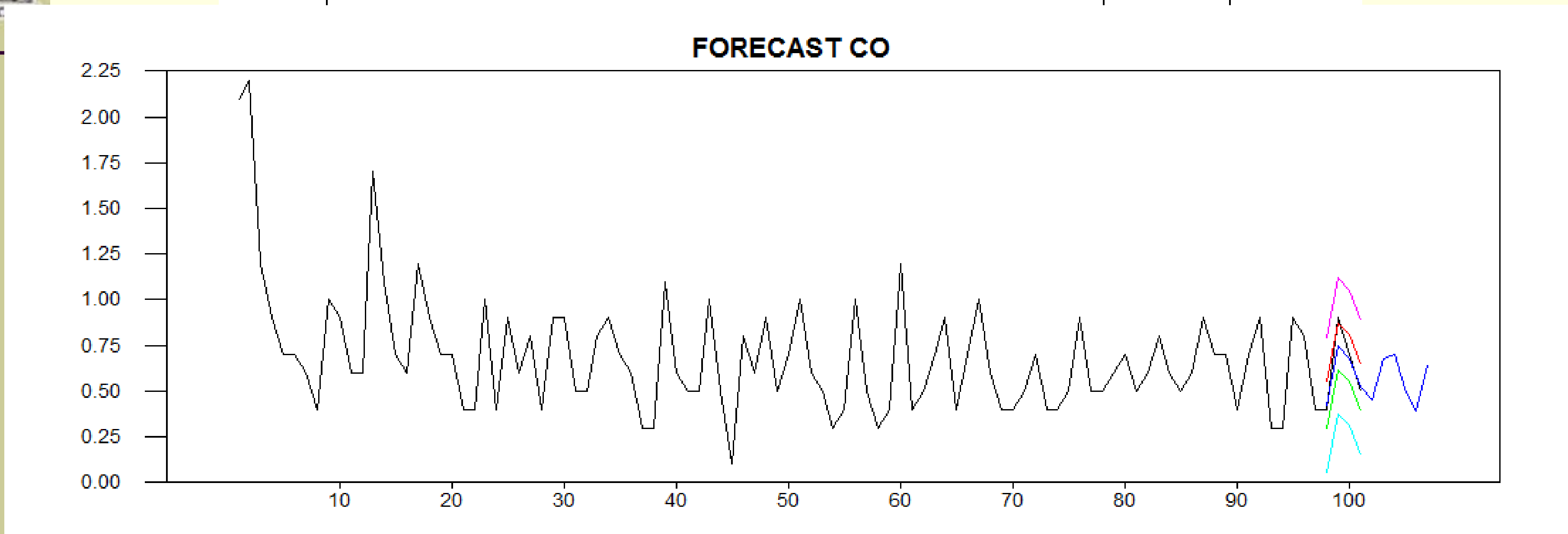
Каждое наблюдение чистого интегрированного (I) процесса (pure integrated (I) process), называемого также случайным блужданием, заключается в случайном шаге в сторону от текущего наблюдения. Этот процесс знает, где он находится, но забыл, как он попал туда.

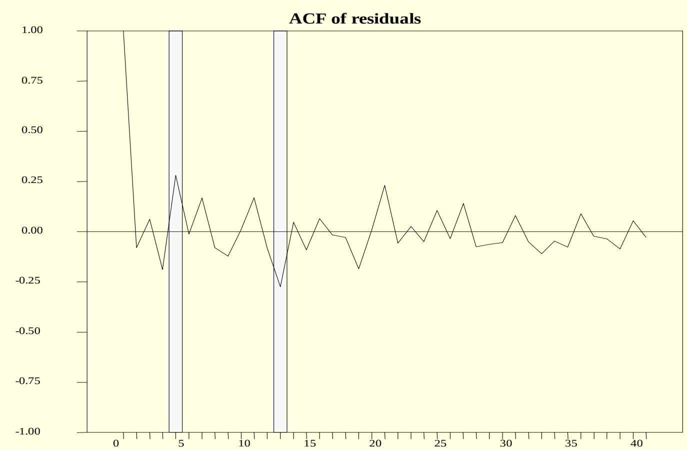




Раз-ность	$\varpi_i \pm \hat{\sigma}(\varpi)$	$\hat{\sigma}^2_{\omega}$	Оценки параметров	Идентифицирован-ные модели
0	0.069 ± 0.03	0.110	$\hat{\phi}_1 = 0.953\ (\pm 0.018)$ $\hat{\theta}_1 = -0.608\ (\pm 0.084)$	ARMA (1,0,1)

98	$\nabla^2 y_i - 0.354\nabla^2 y_{i-1} + 0.506\nabla^2 y_{i-4} - 0.186\nabla^2 y_{i-11} = \varepsilon_i - 0.785\varepsilon$ $(0.099) \quad (0.095) \quad (0.095) \quad (0.070)$	50.2009	40





Мультипликативная модель временных рядов

- Основное предположение: факторы, влияющие на исследуемый объект в настоящем и прошлом, будут влиять на него и в будущем.
- Таким образом, основные цели анализа временных рядов заключаются в идентификации и выделении факторов, имеющих значение для прогнозирования.

Компоненты мультипликативной модели

- Долговременная тенденция называется трендом (trend).
- Тренд является компонентой временного

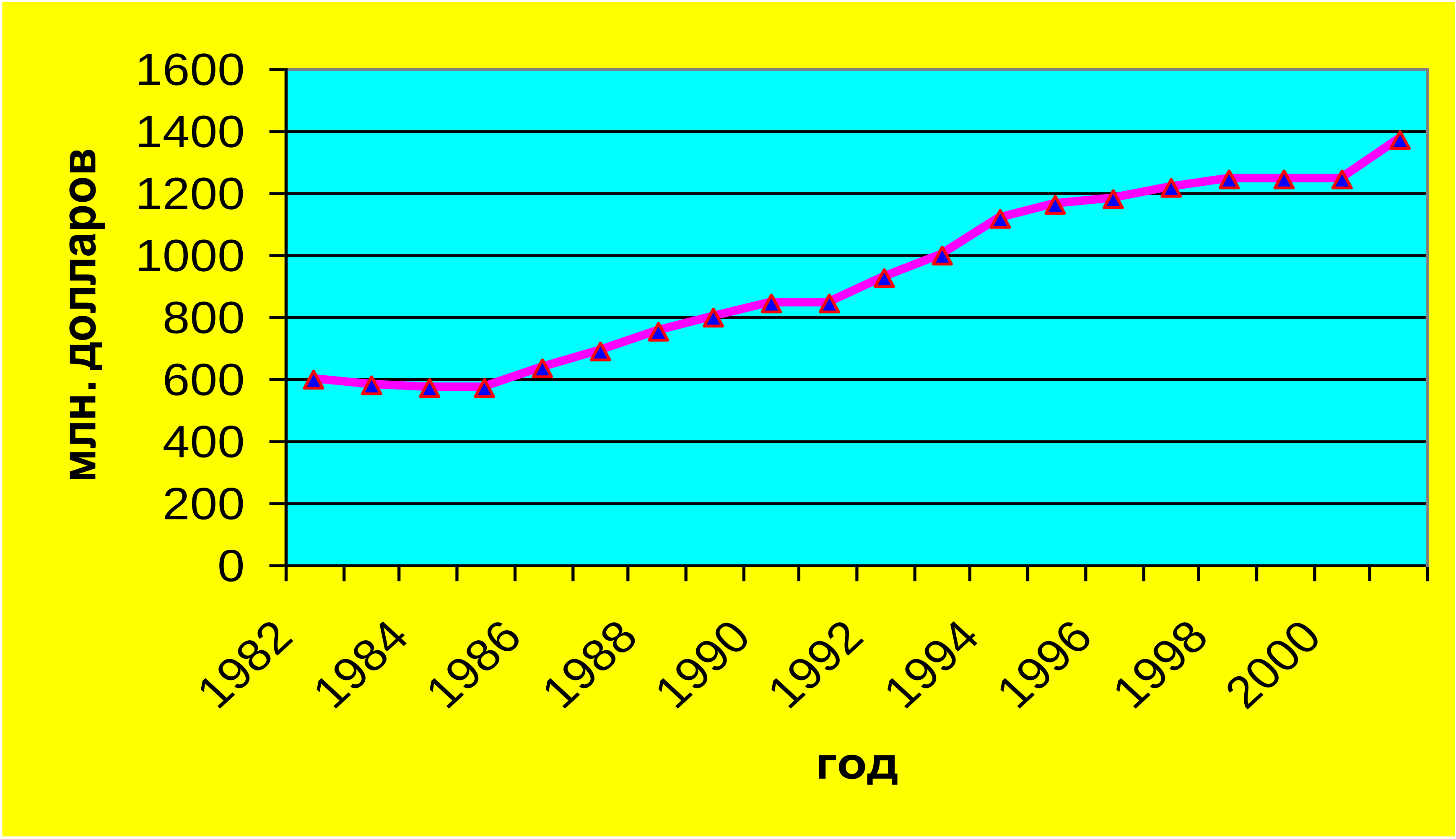
ряда.

- Циклический компонент (cyclical component) описывает колебание данных вверх и вниз. Его длина изменяется в интервале от 2 до 10 лет.

Компоненты мультипликативной модели

- Любые наблюдаемые данные, не лежащие на кривой тренда и не подчиняющиеся циклической зависимости, называются случайными компонентами (random component)

Фактический валовой доход компании WWC за период с 1982 по 2001 годы

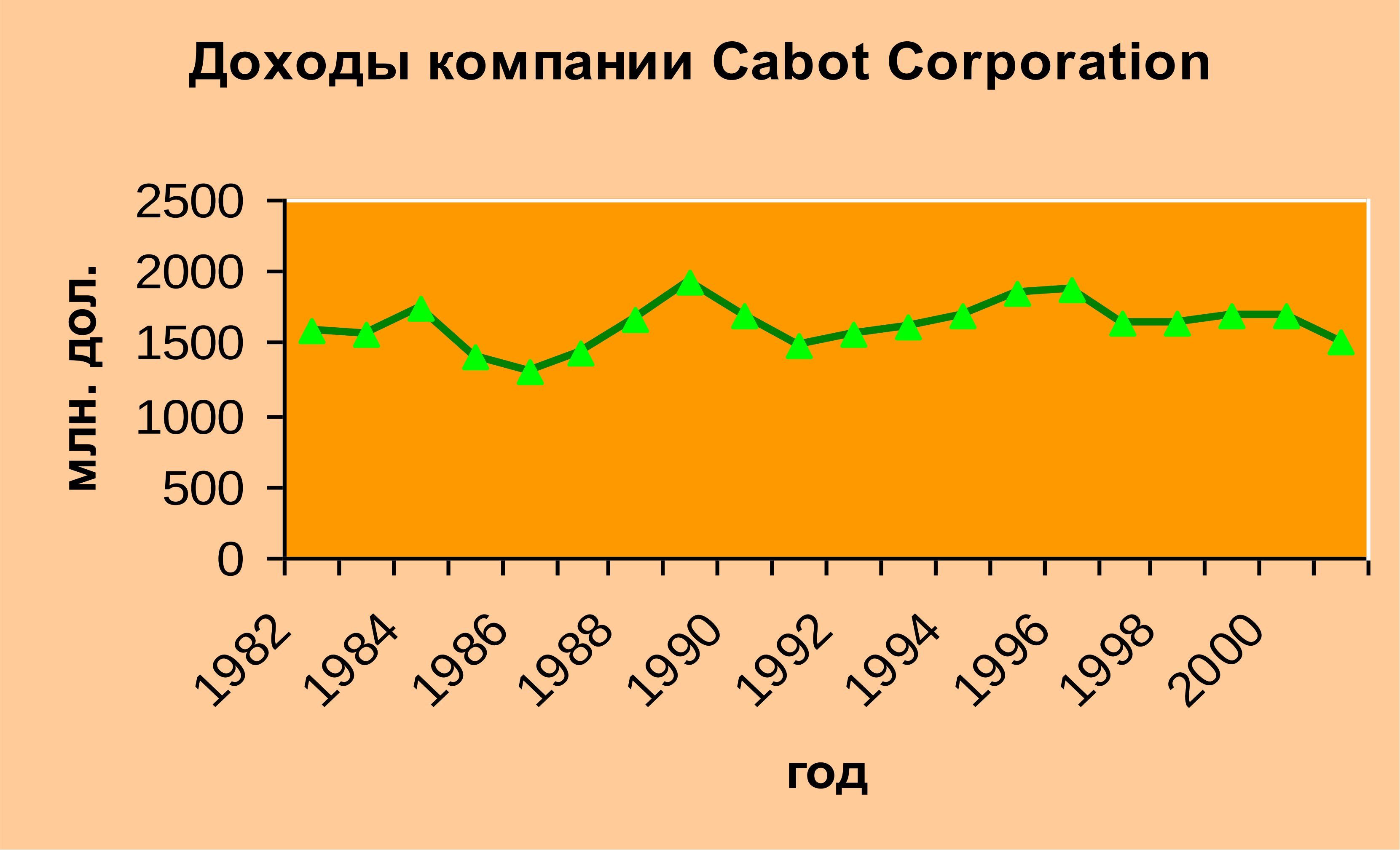


Классическая мультипликативная модель временного ряда для ежегодных данных

$$Y_i = T_i * C_i * I_i$$

- T_i – значение тренда;
- C_i – значение циклического компонента в i-том году
- I_i – значение случайного компонента в i-том году

Сглаживание годовых временных рядов:
скользящее среднее и экспоненциальное среднее



Скользящее среднее

Moving average

подвижная динамическая средняя, которая исчисляется по ряду при последовательном передвижении на один интервал, т.е. сначала вычисляют средний уровень из определенного числа первых по порядку уровней ряда, затем - средний уровень из такого же числа членов, начиная со второго.

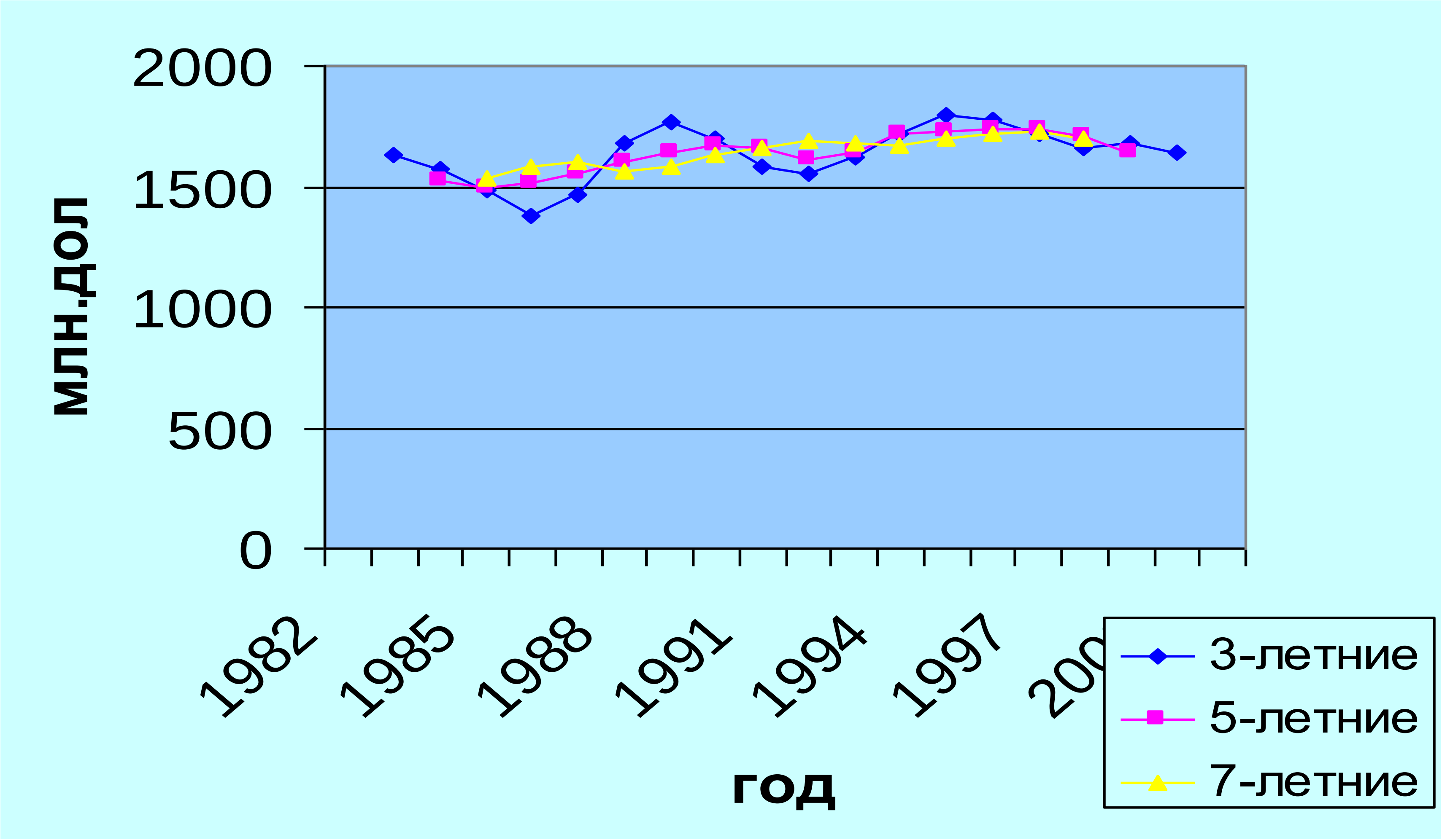
$$\bar{y}_1 = \frac{y_1 + y_2 + y_3}{3} \qquad \bar{y}_2 = \frac{y_2 + y_3 + y_4}{3}$$

$$\bar{y}_3 = \frac{y_3 + y_4 + y_5}{3}$$

Скользящие средние с продолжительностью периода, равной 3

- Для того, чтобы исключить циклические колебания, длина периода должна быть целым числом, кратным средней длине цикла.
- Для трехлетнего периода невозможно выполнить вычисления для первого и последнего года, а при пятилетнем периоде сглаживания - первых двух и последних двух лет.

Скользящие средние для доходов Cabot Corporation



Экспоненциальное сглаживание

Exponential smoothing

Этот метод позволяет делать краткосрочные прогнозы (в рамках одного периода), когда наличие долговременных тенденций остается под вопросом.

При экспоненциальном сглаживании веса, присвоенные наблюдаемым значениям, убывают со временем, поэтому после выполнения вычислений наиболее часто встречающиеся значения получат наибольший вес, а редкие величины - наименьший

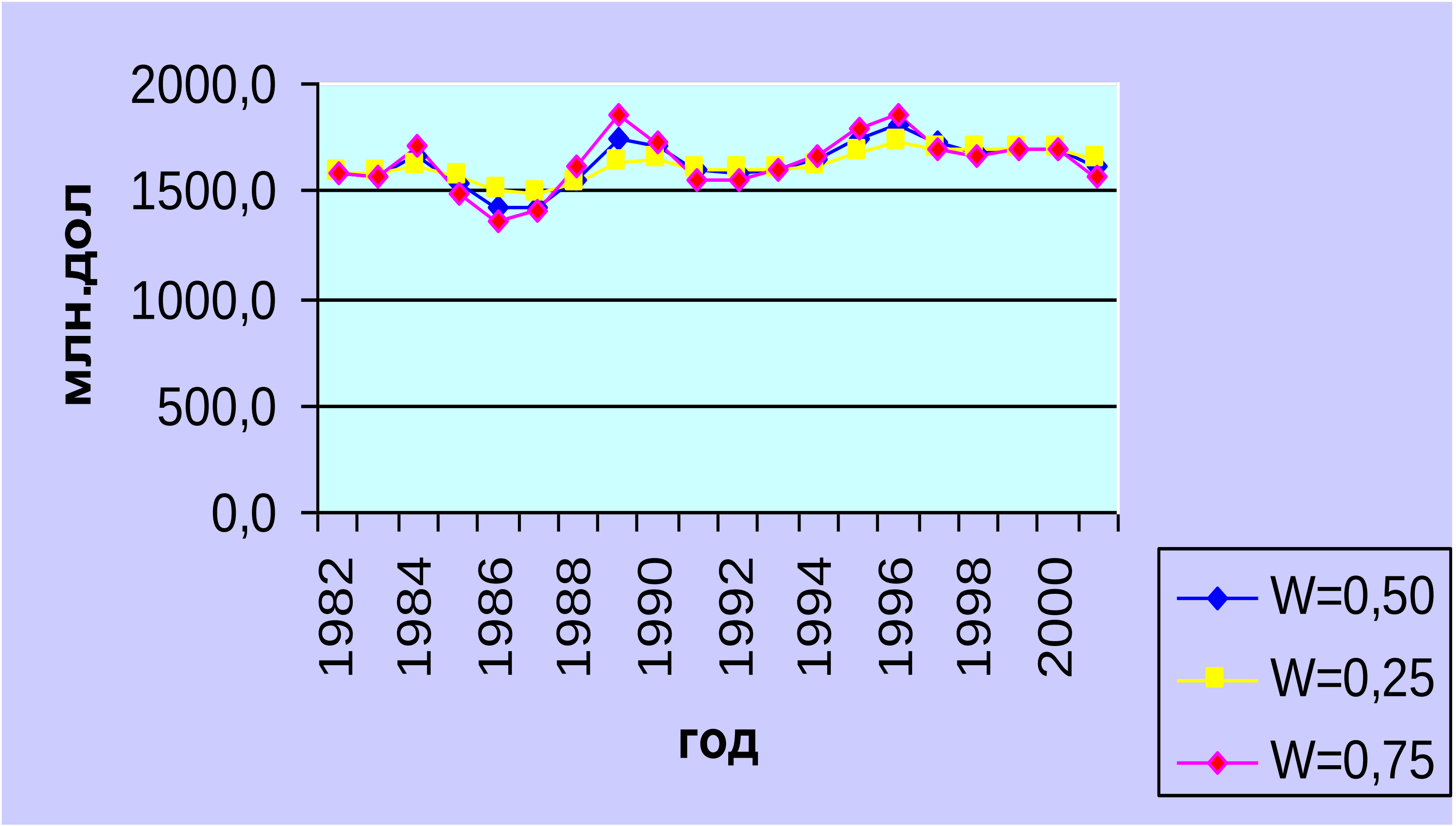
Вычисление экспоненциально сглаженного значения в i-том периоде времени

$$E_i = Y_i$$

$$E_i = WY_i + \left(1 - W\right)E_{i-1} \qquad i = 2,3,4.....$$

- Где E_i- значение экспоненциально сглаженного ряда, вычисленное для i-го периода;
- E_{i-1}- значение экспоненциально сглаженного ряда, вычисленное для предшествующего периода
- Y_i- наблюдаемое значение временного ряда в i-ом периоде;
- W – субъективный вес или сглаживающий коэффициент (0<W<1)

Экспоненциально сглаженный временной ряд для доходов Cabot Corporation



Экспоненциально сглаженный временной ряд для доходов Cabot Corporation

- Допустим, что коэффициент сглаживания равен 0,25. Первое наблюдаемое значение $Y_{1982} = 1587{,}7$ одновременно является первым сглаженным значением $E_{1982} = 1587{,}7$.
- Используя значение временного ряда для 1983 года получаем:

  $E_{1983} = WY_{1983}+(1-W)E_{1982} =$

  $= 0{,}25*1558{,}0+0{,}75*1587{,}7 = 1580{,}3$

Экспоненциально сглаженный временной ряд для доходов Cabot Corporation

- Сглаженное значение временного ряда для 1984 года:

  $E_{1984} = WY_{1984} + (1\text{-}W)E_{1983} =$

  $= 0{,}25*1752{,}5+0{,}75*1580{,}3 = 1623{,}3$ и т.д.

Этот процесс продолжается до тех пор пока не будут сглажены все значения вариационного ряда.

Прогнозирование значений для (i+1)-го интервала

$$\widehat{Y}_{i+1} = E_i$$

Для предсказания доходов компании Cabot Corporation в 2002 году можно использовать сглаженное значение, вычисленное для 2001 года