

электронный журнал
МОЛОДЕЖНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК

Издатель Общероссийская общественная организация "Академия инженерных наук им. А.М. Прохорова"
ISSN 2307-0609

11, ноябрь 2017

УДК 004.021

**Применение авторегрессионных моделей скользящего среднего в задаче
прогнозирования финансовых временных рядов**

Семиохин С.И., магистр
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Компьютерные системы и сети»
drstep321@mail.ru

Научный руководитель: Самарев Р.С., к.т.н., доцент
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Компьютерные системы и сети»
samarev@actm.org

Аннотация: В данной статье рассмотрены возможности применения авторегрессионных моделей скользящего среднего для прогнозирования финансовых временных рядов на примере валютной пары EUR/USD. Показано, каким образом проводится статистический анализ временного ряда. Рассмотрены способы приведения искомого ряда к стационарному виду. Показан способ подбора оптимальных параметров для интегрированных моделей авторегрессии-скользящего среднего. Проведена оценка качества полученного прогноза. Сформулированы выводы по возможности использования рассматриваемых моделей применительно к финансовым времененным рядам.

Ключевые слова: прогноз (*forecast*), временные ряды (*time series*), стационарность (*stationarity*), авторегрессия (*autoregressive*), скользящее среднее (*moving average*), случайное блуждание (*random walk*), ARIMA, SARIMAX.

Введение

Анализ данных используется для решения широкого спектра задач и нашел свое применение в различных областях. В данной работе рассматривается применение анализа данных для задач прогнозирования финансовых временных рядов.

Основной особенностью финансовых временных рядов, таких как валютные курсы, акции компаний, фондовые индексы, является то, что значение ряда в каждый момент времени отражает в себе все побочные признаки, которые влияют на данный показатель.

Рассмотрим возможность применения авторегрессионных моделей скользящего среднего [1] для прогнозирования финансовых временных рядов на примере валютной пары EUR/USD.

1. Предобработка данных

Анализ временного ряда имеет два основных отличия от стандартной проблемы регрессии: наблюдения зависят от времени и помимо основного тренда ряд может иметь так называемый сезонный тренд, то есть повторяющийся тренд, характерный определенному промежутку времени (примером таких сезонных трендов является снижение цены на зимнюю одежду в начале весны).

Вся дальнейшая обработка будет проводиться на дневных данных валютной пары EUR/USD за период времени с первого января 2008 года по май 2017 (рис.1).

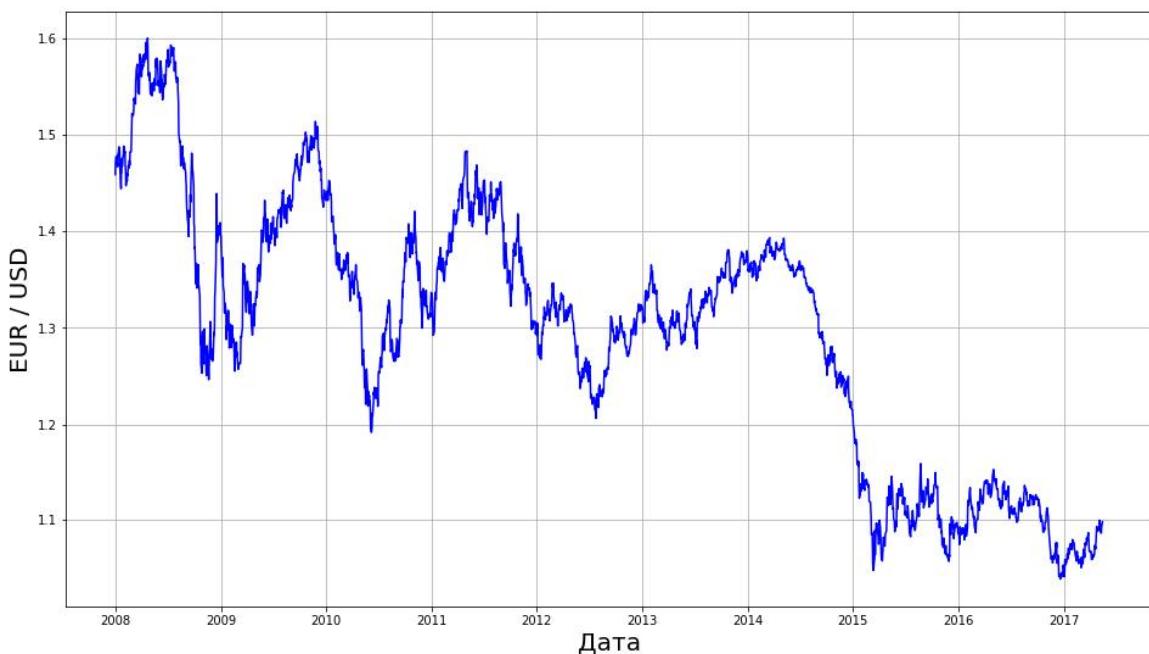


Рис. 1. Данные для построения модели

Данные представлены в формате пары <год-месяц-день>:<значение отношения EUR/USD>.

Для многих статистических (в том числе авторегрессионных) моделей одним из важнейших свойств при анализе временного ряда является его стационарность. Временной ряд называется *строго стационарным* (strictly stationarity) или стационарным в узком смысле, если все его свойства не зависят от времени [2]. Ряд считается слабо стационарным (weak stationarity) или стационарным в широком смысле, если соответствует трем статистическим свойствам [2]:

- 1) Математическое ожидание стационарного ряда является постоянным, то есть среднее значение временного ряда, вокруг которого изменяются уровни, является величиной.

- 2) Дисперсия стационарного ряда является постоянной. Она характеризует отклонение значений временного ряда относительно его среднего значения.
- 3) Автоковариация ряда является постоянной (ковариация зависит только от сдвига, но не от времени). Автоковариация - корреляционная связь между последовательными уровнями одного и того же ряда динамики (сдвинутыми на определенный промежуток времени L - лаг).

Строгая стационарность подразумевает слабую стационарность, но не наоборот.

Стационарность может нарушаться по математическому ожиданию или по дисперсии. В зависимости от выбранной характеристики говорят о стационарности временного ряда относительно среднего значения или относительно дисперсии.

Большая часть моделей, работающих с временными рядами подразумевает, что ряд стационарен.

По визуальному изображению ряда можно предположить, что рассматриваемый временной ряд не является стационарным, однако проведем дополнительные исследования, а именно:

- 1) Отобразим на графике скользящее среднее и скользящую дисперсию, чтобы определить, соответствует ли ряд первым двум критериям стационарности.
- 2) Тест Дики-Фуллера (*Augmented Dickey-Fuller test, ADF*) [3].

Тест Дики-Фуллера работает следующим образом - делается предположение о виде процесса, породившего данный временной ряд, строится вспомогательная модель и проверяются гипотезы о коэффициентах этой модели. Затем делается вывод о стационарности/нестационарности исходного ряда.

При помощи этого теста проверяют значение коэффициента α в авторегрессионном уравнении первого порядка AR(1):

$$y_t = \alpha * y_{t-1} + e_t ,$$

где y_t – рассматриваемый временной ряд,

e_t – шум.

Нулевая гипотеза $H_0: \alpha = 1$ (существует единичный корень, ряд нестационарный).

Альтернативная гипотеза: $H_1: \alpha < 1$ (единичного корня нет, ряд стационарный).

H_0 отвергается на N процентном ($N=1\%, 5\%, 10\%$) уровне значимости, если значение статистики лежит левее критического значения (критические значения — отрицательные).

В противном случае гипотеза не отвергается и процесс может содержать единичные корни, то есть быть нестационарным (интегрированным) временным рядом.

Так же особенностью стационарного временного ряда является то, что скользящее среднее и скользящее стандартное отклонение ведут себя схожим образом. Скользящее среднее равняется среднему значению за период/окно.

Были проведены тесты на стационарность, в качестве периода для скользящего среднего был взят 30-дневный промежуток (рис. 2, таблица 1).



Рис. 2. Временной ряд со скользящим средним и дисперсией

Таблица 1.

Тест Дики-Фуллера на исходном временном ряду

Свойства теста Дики-Фуллера	Исходные ряд
Тестовая статистика	-1.444857
p-уровень значимости	0.560592
Количество наблюдений	2444
Критическое значение (1%)	-3.433028
Критическое значение (5%)	-2.862723
Критическое значение (10%)	-2.567400

И визуальный тест, и статистический говорят о том, что ряд не является стационарным. Так как p-уровень значимости равен 0.5604592 мы не можем отвергнуть гипотезу H_0 и быть уверены в стационарности ряда. Для того, чтобы понять, что делает ряд нестационарным необходимо рассмотреть составляющие ряда.

Каждый временной ряд складывается из следующих основных составляющих (компонентов):

- 1) Тенденции, характеризующей общее направление динамики изучаемого явления. Аналитически тенденция выражается некоторой функцией времени, называемой трендом (T).
- 2) Циклическая компонента (C) – плавно изменяющаяся компонента, описывающая длительные периоды относительного подъема и спада, состоит из циклов, меняющихся по амплитуде и протяженности (в экономике бывает связана со взаимодействием спроса и предложения, ростом и истощением ресурсов, изменением в финансовой и налоговой политике и т.п.).
- 3) Сезонные колебания (S) – периодические колебания, которые имеют определенный и постоянный период (объем продаж накануне Нового Года, объем перевозок пассажиров городским транспортом).
- 4) Случайной составляющей, которая является результатом воздействия множества случайных факторов (e) - шум.

Временной ряд представляет собой:

- либо сумму этих компонент $X=T+C+S+e$ в аддитивной модели,
- либо произведение $X=T*C*S*e$ в мультипликативной модели.

Второй вариант более распространен и сводится к первому логарифмированием.

Сильные изменения скользящего среднего на рисунке 2 обусловлены наличием значимого тренда. Считается, что сезонность также придает ряду свойства нестационарности.

Существует два основных подхода к приведению ряда к стационарности [3]:

- 1) Декомпозиция ряда на компоненты и удаление таких показателей как тренд и сезонность.
- 2) Разностной дифференцирование – вычитание из значения $x(t)$ значения $x(t - 1)$, где 1 – временной лаг.

Искомый временной ряд был разложен на сумму компонентов посредством логарифмирования (рис. 3, рис. 4).

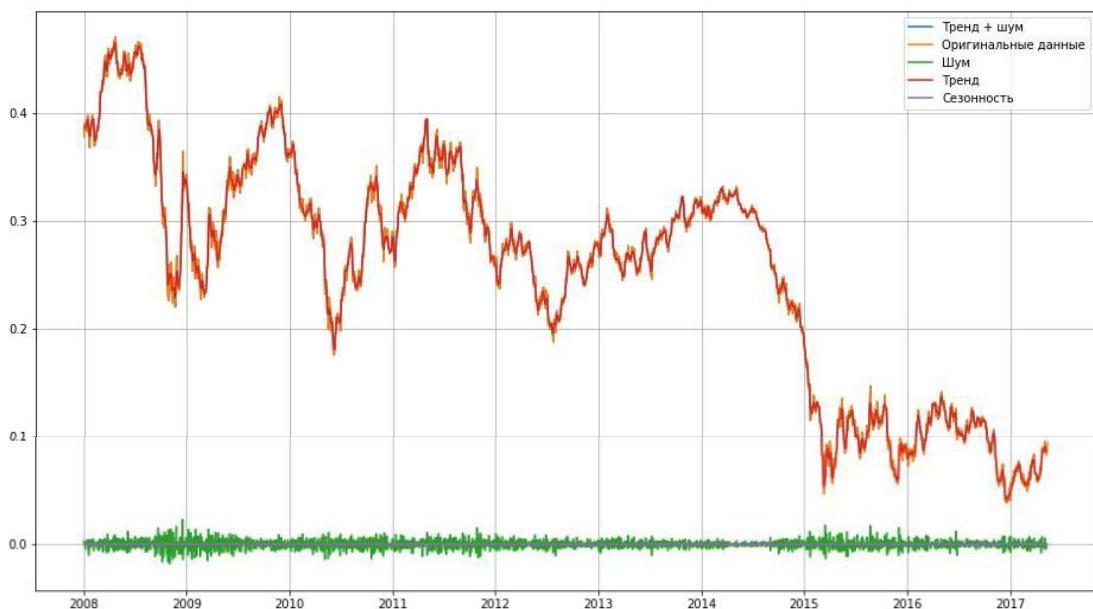


Рис. 3. Разложение временного ряда на компоненты



Рис. 4. Разложение временного ряда на компоненты (приближенное)

Рис. 3 показывает, что временной ряд в таком масштабе практически полностью совпадает с трендом, сезонность отсутствует, а значения шума достаточно небольшие. Увеличив масштаб (рис. 4) видно, что совокупность тренда и шума совпадают с оригинальным временным рядом.

Таким образом, можно говорить, что вариант с удалением сезонности и тренда не подходит, так как тренд является одним из ключевых показателей, которые надо прогнозировать.

Второй способ предполагает дифференцирование ряда. Используем временной лаг равный единице и вычтем его из оригинального ряда, так что $x(t) = x(t) - x(t - 1)$ (рис. 5).

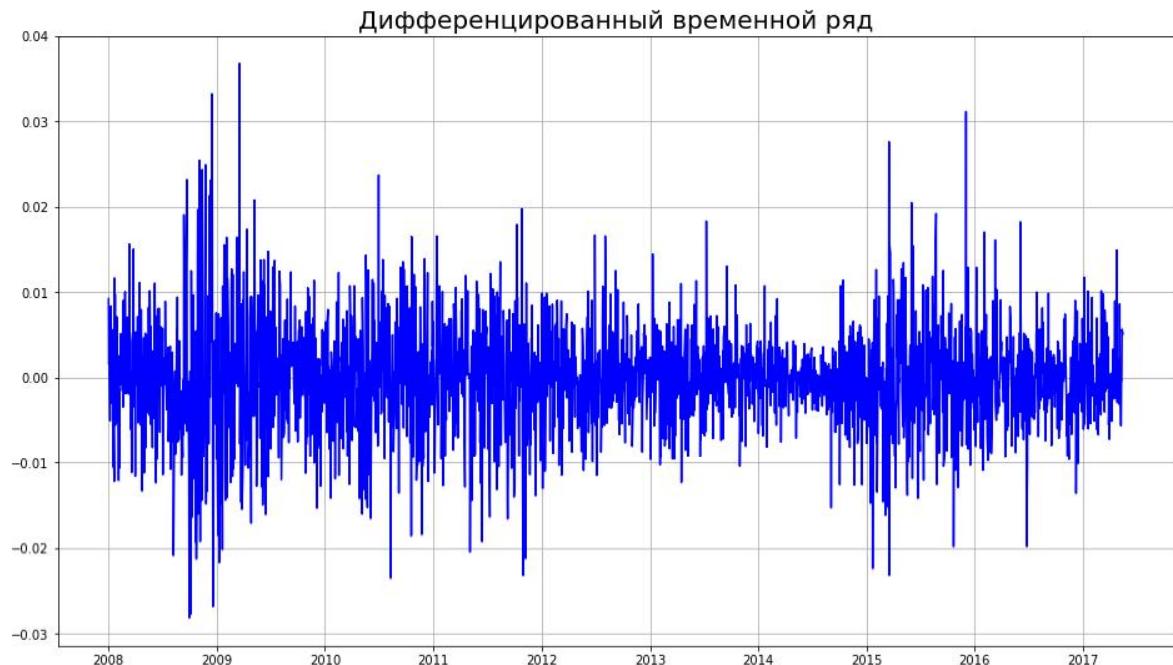


Рис. 5. Дифференцированный временной ряд

Снова были проведены тесты на стационарность (рис.6, таблица 2).

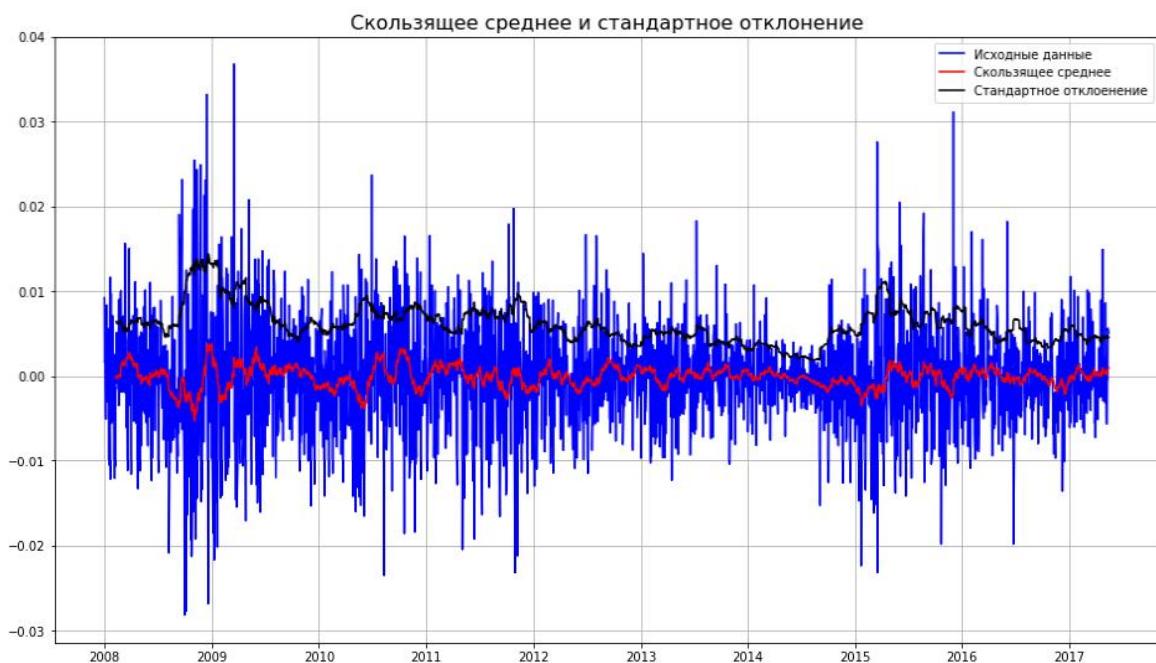


Рис. 6. Визуальная проверка на стационарность за счет скользящего окна и дисперсии

Таблица 2.

Тест Дики-Фуллера для дифференцированного ряда

Свойства теста Дики-Фуллера	Ряд первых разностей
Тестовая статистика	-48.67836
p-уровень значимости	0.00000
Количество наблюдений	2443
Критическое значение (1%)	-3.433030
Критическое значение (5%)	-2.862724
Критическое значение (10%)	-2.567400

Как видно из теста Дики-Фуллера и из визуального представления, есть все основания отвергнуть гипотезу о том, что ряд нестационарный. Таким образом, можно работать с данным рядом как со стационарным.

Исходный ряд получается посредством скользящего прибавления ряда к первому значению, то есть: $x(t) = x(t) + x(t - 1)$

Для использования с конкретной реализацией было принято решение использовать не просто логарифмирование, а преобразование Бокса-Кокса [5], позволяющее преобразовать асимметричные распределения в нормальное, что является важным для многих статистических моделей, включая авторегрессионные. Для исходной последовательности $y = \{y_1, y_2, y_3, \dots, y_n\}$ однопараметрическое преобразование Бокса-Кокса с параметром λ определяется следующим образом:

$$y_i^\lambda = \begin{cases} \frac{y_i^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(y_i^\lambda), & \lambda = 0 \end{cases}$$

Для проведения такого преобразования использовался пакет `scipy.stats`, в котором параметр λ определяется методом максимизации правдоподобия.

2. Построение прогноза

При построении модели ARIMA [6] для полученного ранее ряда первых разностей необходимо подобрать три параметра:

- 1) p – порядок компоненты авторегрессионного процесса,
- 2) d – порядок интегрированного ряда,
- 3) q – порядок компоненты скользящего среднего.

Параметр d равняется 1, так как прогноздается для ряда первых разностей. Для определения параметров p и q рассмотрим авторкорреляционную (autocorrelation function,

ACF) и частично автокорреляционную (partial autocorrelation function, PACF) функции для ряда первых разностей.

Автокорреляционная функция — зависимость взаимосвязи между функцией и ее сдвинутой копией от величины временного сдвига. Для определения значений необходимо представить коррелограмму (график автокорреляции) для ряда.

ACF определяет меру корреляции временного ряда и того же ряда, но со сдвигом. Используется для определения параметра q , так как по ее коррелограмме можно определить количество автокорреляционных коэффициентов сильно отличных от 0 в модели скользящего среднего.

То есть анализ автокорреляционной функции и коррелограммы позволяет найти лаг, при котором автокорреляция наиболее высокая, а, следовательно, и лаг, при котором связь между текущим и предыдущими уровнями временного ряда наиболее тесная (рис. 7).

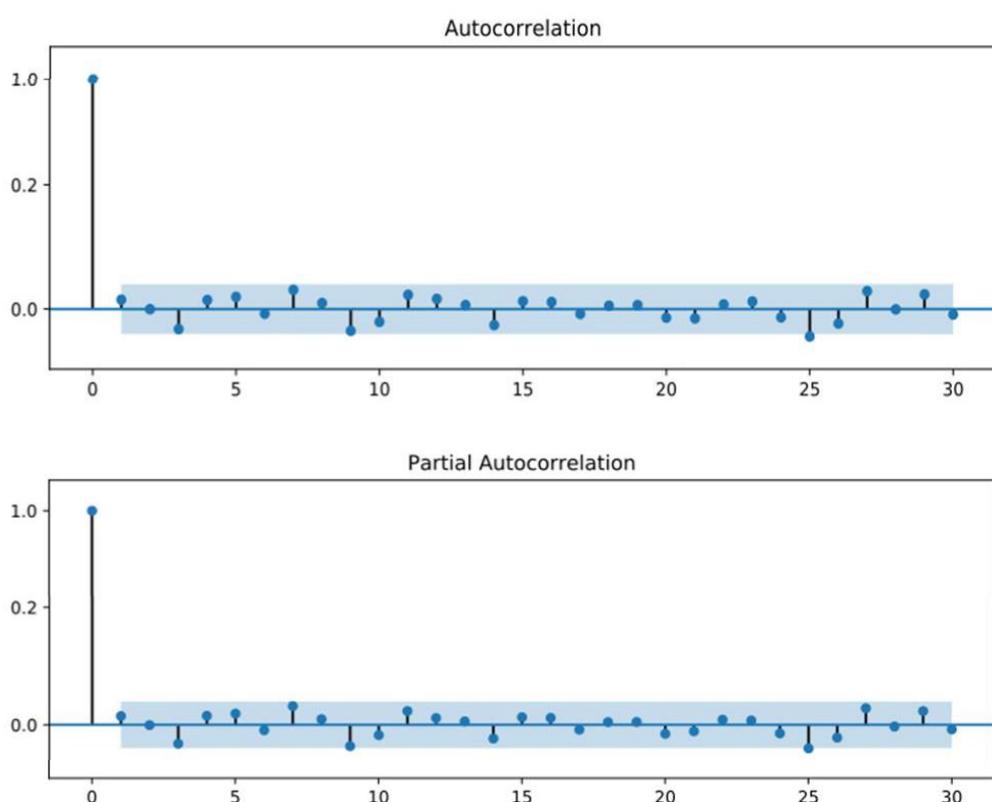


Рис. 7. Автокорреляционная (ACF) и частично автокорреляционная (PACF) функции

PACF также показывает корреляцию между временным рядом и сдвинутой вариацией себя же, но за вычетом влияния всех внутренних значений автокорреляции. В частной автокорреляционной функции устраняется зависимость между наблюдениями внутри лага (промежуточными наблюдениями). По коррелограмме данной функции

определяют максимальный номер коэффициента сильно отличный от 0 в модели авторегрессии.

На рис. 7 представлены коррелограммы, где голубой зоной представлена область, внутри которой значения лага можно считать хаотичными. Выход за эту зону означает корреляцию между лагами. Общее количество лагов для проверки было взято равное 30 (продолжительность месяца).

И для ACF, и для PACF первый лаг, для которого видна корреляция – третий лаг. Таким образом, параметры p и q примем за 3. Полученная модель ARIMA имеет вид: $(0, 1, 0)$. Полученный результат показывает, что рассматриваемый временной ряд описывается такими же параметрами, как процесс случайных блужданий.

Рассмотрим процесс белого шума, который стационарен, имеет математическое ожидание 0 и дисперсию 1:

$$x_t = x_{t-1} + e_t,$$

где e_t – значение процесса шума,

x_{t-1} – предыдущее значение временного ряда.



Рис. 8. Процесс случайного блуждания

Смоделированный процесс случайного блуждания по свойствам близок к рассматриваемому валютному временному ряду, таким образом, прогнозирование с использованием относительно простых подходов (таких как построение линейной регрессии или градиентного бустинга) вряд ли будет эффективно.

Было принято решение провести еще одну проверку на оптимальность подобранных параметров: сравнить значения информационного критерия Акаике (AIC) [7] для разных параметров. Данный критерий применяется для сравнения статистических моделей и в общем случае вычисляется следующим образом:

$$AIC = 2k - 2\ln(L),$$

где k — число параметров в статистической модели,

L — максимизированное значение функции правдоподобия модели.

Критерий вознаграждает за качество приближения, но и штрафует за использование излишнего количества параметров модели. Считается, что наилучшей будет модель с наименьшим значением критерия AIC. Абсолютные значения не несут в себе никакой полезной информации и рассматривать результаты стоит только в сравнении с показателями критериев для других моделей.

Подбор параметров производился подбором по сетке (Grid Search), то есть перебором всевозможных комбинаций (таблица 3).

Таблица 3.

Подбор параметров модели по AIC

Параметры (p, d, q)	Значение AIC
(0, 0, 0)	-3490.3496089
(0, 0, 1)	-5621.853147688794
...
(4, 0, 0)	-12146.496201139362
(4, 0, 1)	-12148.496313907879
(4, 0, 2)	-12143.658228896475
...

Наилучшим набором параметров, полученных таким методом оказался набор (4, 0, 1). Он и был использован для построения модели.

Была построена соответствующая модель (рис. 9).

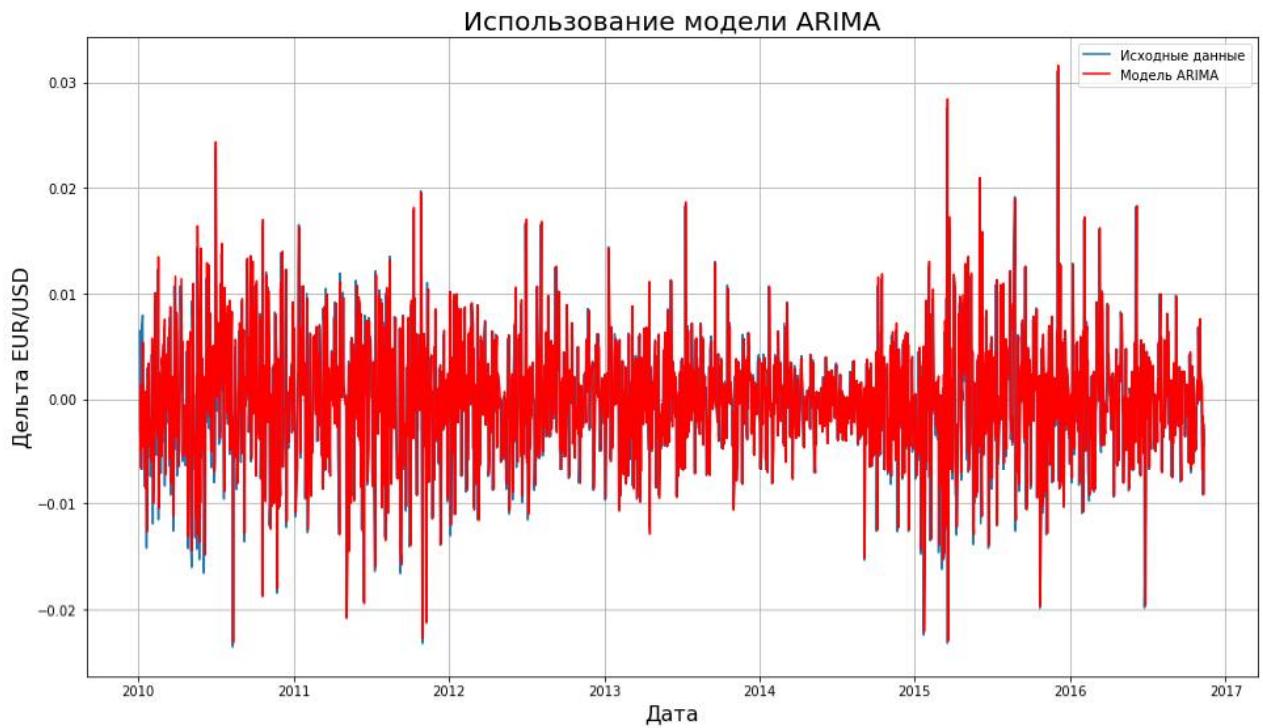


Рис. 9. Модель ARIMA

Затем данные были денормализованы (рис. 10).



Рис. 10. Модель ARIMA (денормализованные данные, приближенное)

На основе построенной модели был сделан единичный прогноз (рис. 11).



Рис. 11. Однодневный прогноз

Кажется, что модель достаточно четко описывает обучающую выборку, однако это не так. Сдвиг на рисунках выше обусловлен тем, что прогноз формируется по факту так же, как и в процессе белого шума: $x_t = x_{t-1} + e_t$, причем e_t крайне мало, это можно увидеть и графически (рис. 12) и по значениям (таблица 4).

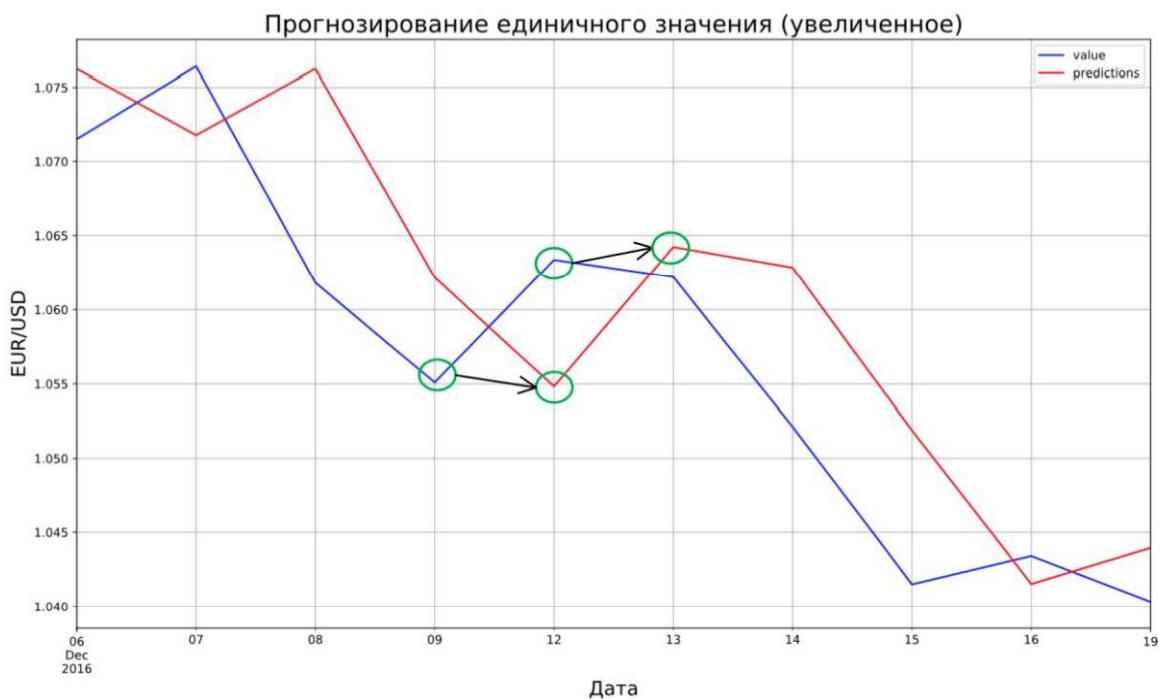


Рис. 12. Однодневный прогноз (приближенное)

Таблица 4.

Однодневный прогноз

Тестовое значение	Прогноз
0.057881	...
0.051840	0.057231
0.054949	0.051786
0.054820	0.054952
0.056089	0.054318
0.060281	0.055908
0.062844	0.059881
...	0.062552

Эту гипотезу подтвердил и многодневный прогноз (рис. 13).

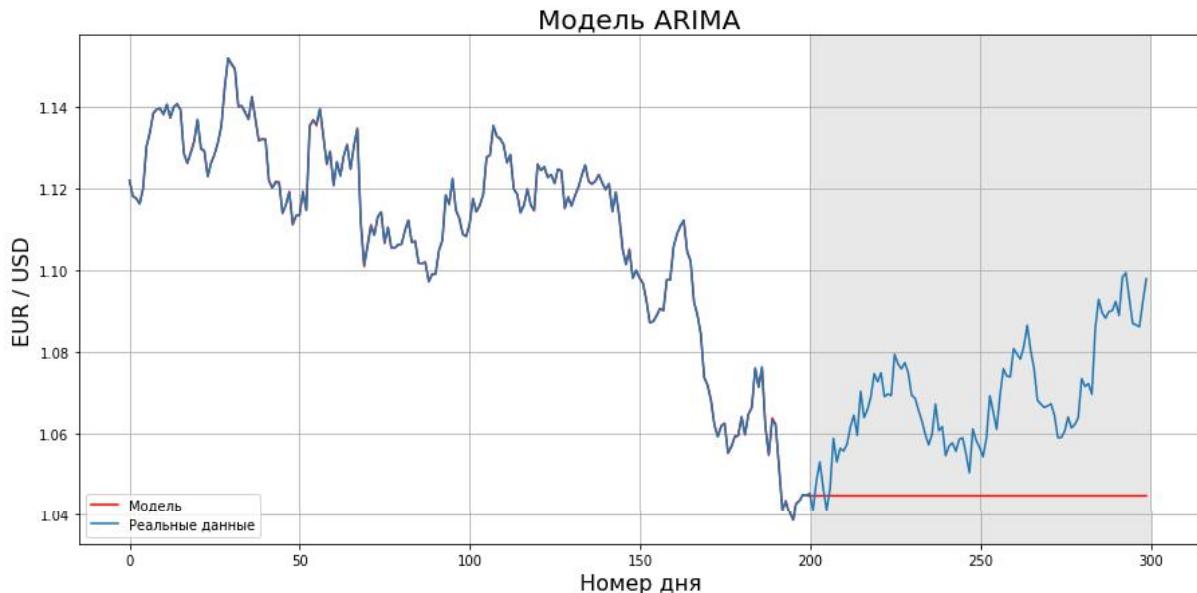


Рис. 13. Многодневный прогноз с ARIMA

Таким образом, гипотеза была подтверждена и модель ARIMA не обладает сколь значимой предсказательной способностью для рассматриваемого ряда.

В качестве альтернативы модели ARIMA было решено попробовать ее сезонную вариацию: модель SARIMAX [8]. Так как автокорреляционные и частично автокорреляционные функции и разложения ряда на сезонность показали, что краткосрочная сезонность в рассматриваемом ряду отсутствует, в качестве сезонной компоненты было взято значение 365 – год, p,d,q параметры же остались такими же, как и для вышеописанной ARIMA-модели (рис. 14).

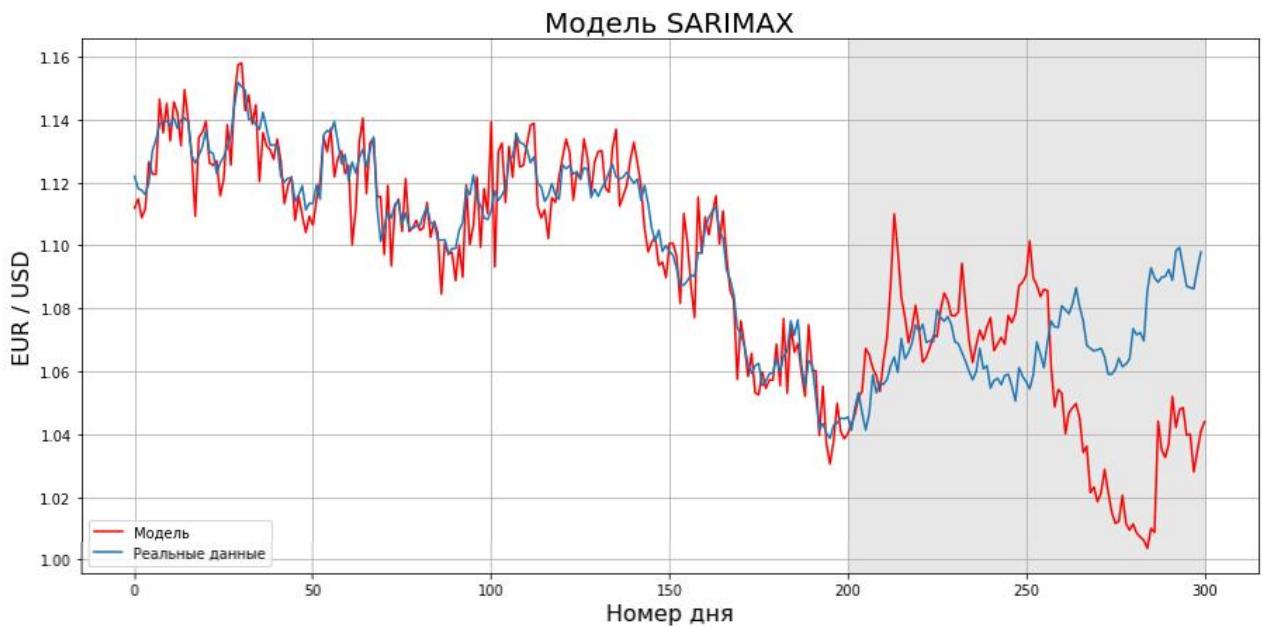


Рис. 14. Многодневный прогноз с SARIMAX

На рисунке 14 видно, что модель SARIMAX не так точно подгоняется под обучающую выборку и имеет больше выбросов, однако в отличие от ARIMA она обладает предсказательной способностью на представленных данных.

Так как модель SARIMAX оказалась значительно менее гладкой, чем оригинальные данные, была проведена попытка сгладить ее посредством скользящего среднего за одну и две недели (рис. 15).

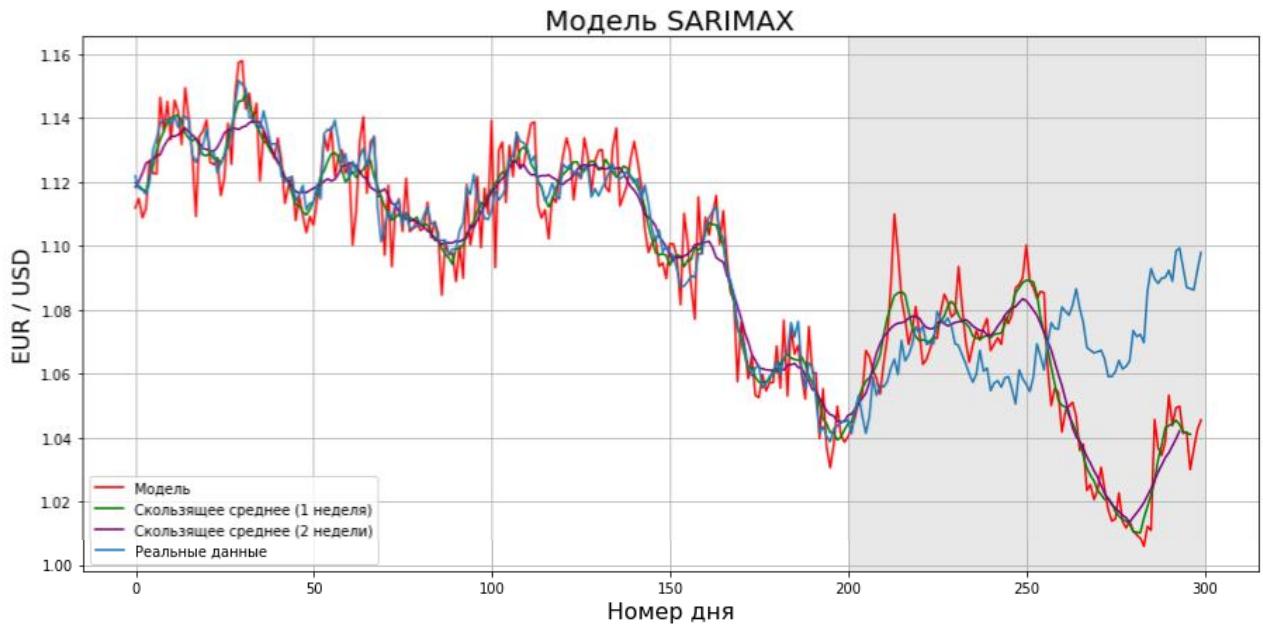


Рис. 15. Многодневный прогноз с SARIMAX со сглаживанием

Если рассмотреть прогноз ближе, то видно, что за исключением одного отрезка, где модель вела себя некорректно, прогноз неплохо отражает реальную ситуацию (рис. 16).



Рис. 16. Многодневный прогноз с SARIMAX (приближенный)

3. Оценка результатов

Для количественной оценки в задачах регрессии используется метрика R^2 [4]:

$$R^2 = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \bar{y})^2},$$

где y_i – реальное значение ряда,

\hat{y}_i – прогнозное значение ряда,

\bar{y} – среднее значение ряда.

Данная метрика является коэффициентом детерминации. R^2 показывает, насколько условная дисперсия полученной модели отличается от дисперсии реальных значений. Если этот коэффициент близок к 1, то условная дисперсия модели достаточно мала и весьма вероятно, что модель неплохо описывает данные. Если же коэффициент R-квадрат значительно меньше, то с большой долей вероятности прогнозная модель не отражает реальное положение вещей.

Основная проблема использования метрики R^2 заключается в том, что ее значение увеличивается (*не уменьшается*) от добавления в модель новых предикторов, даже если они никак не влияют на результат. Однако, в случае с временным рядом единственным предиктором являются предыдущие значения ряда, поэтому в данном случае использование данной метрики уместно.

Для многодневного прогноза с моделью SARIMAX показатель R^2 оказался равен 0,286 , что явилось следствием штрафа за участок с неверным прогнозом.

Заключение

В статье была рассмотрена возможность применения авторегрессионных моделей для прогнозирования финансовых временных рядов на примере валютной пары EUR/USD.

Разработанная модель SARIMAX достаточно полно описывает зависимости для валютной пары EUR/USD и может применяться для многодневного прогноза. Однако, модель достаточно нестабильна и на некоторых временных отрезках ведет себя некорректно, поэтому слепо применять ее нельзя, возможно, более тщательный подбор параметров сезонности позволит улучшить качество модели и уменьшить количество выбросов.

Список литературы

- [1]. Box G.E.P., Hilimer S., Tiao G.C. Analysis and modeling of seasonal time series // Seasonal analysis of economic time series. NBER, 1978. P. 309-344. MLA
- [2]. Бокс Дж., Дженкинс Г. Анализ временных рядов, прогноз и управление. М.: Мир, 1974, кн. 1. 406 с.
- [3]. Dickey D.A., Fuller W.A. Distribution of the estimators for autoregressive time series with a unit root // Journal of the American statistical association. 1979. Т. 74. № 366a. P. 427-431.
- [4]. Box G. E. P., Cox D. R. An analysis of transformations // Journal of the Royal Statistical Society. Series B (Methodological). 1964. С. 211-252.
- [5]. Айвазян С.А. Прикладная статистика. Основы эконометрики. Т. 2. М.: Юнити-Дана, 2001. 432 с.
- [6]. Akaike H. Information theory and an extension of the maximum likelihood principle // Selected Papers of Hirotugu Akaike. Springer New York, 1998. P. 199-213.
- [7]. Bercu S., Proia F. A SARIMAX coupled modelling applied to individual load curves intraday forecasting //Journal of Applied Statistics. 2013. Т. 40. № 6. P. 1333-1348.
- [8]. Cameron A.C., Windmeijer F.A.G. An R-squared measure of goodness of fit for some common nonlinear regression models //Journal of Econometrics. 1997. Т. 77. № 2. P. 329-342.