



МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

«МИРЭА – Российский технологический университет»
РТУ МИРЭА

Институт информационных технологий
кафедра прикладной математики

КУРСОВАЯ РАБОТА

по дисциплине **Технологии организации, обработки и хранения статистических данных**

(наименование дисциплины)

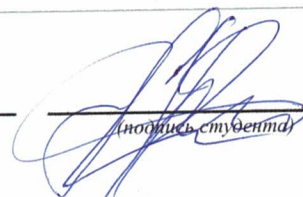
Тема курсовой работы

Статистический анализ валютных курсов

Студент группы

**ИМБО-01-21 Рутковская
Анастасия Алексеевна**


(учебная группа Ф.И.О. студента)


(подпись студента)

Руководитель курсовой работы

доцент, к.п.н. Митина О.А.

(должность, звание, учёная степень Ф.И.О.)


(подпись руководителя)

Рецензент (при наличии)

(должность, звание, учёная степень Ф.И.О.)

(подпись рецензента)

Работа представлена к защите

«14» декабря 2022 г.

Допущен к защите

«22» декабря 2022 г.

Ст. преподаватель **Буданцев А.В.** 

Москва 2022 г.



МИНОБРНАУКИ РОССИИ

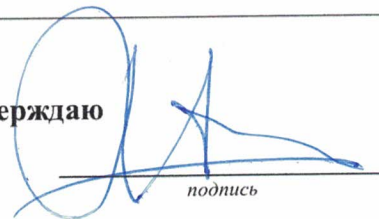
Федеральное государственное бюджетное образовательное учреждение
высшего образования

«МИРЭА – Российский технологический университет»
РТУ МИРЭА

Институт информационных технологий
кафедра прикладной математики

Утверждаю

Заведующий кафедрой
Держинский Роман
Игоревич


подпись

Ф.И.О.

«9» сентября 2022 г.

ЗАДАНИЕ

на выполнение курсовой работы по дисциплине

Технологии организации, обработки и хранения статистических данных

Студент Рутковская Анастасия Алексеевна Группа ИМБО-01-21

Тема: Статистический анализ валютных курсов

Исходные данные: Выборка данных стоимости валют доллара и евро

Перечень вопросов, подлежащих разработке, и обязательного графического материала: _____

1) Исследование ключевых понятий временных рядов и корреляционного анализа

2) Изучение математической модели ARIMA для построения прогнозных значений временных рядов

Срок представления к защите курсовой работы

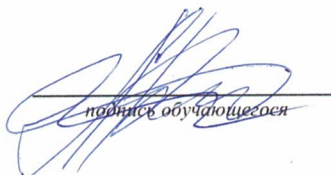
до «22» декабря 2022 г.

Задание на курсовую работу выдал


подпись руководителя

Митина О.А.
Ф.И.О. руководителя

Задание на курсовую работу получил


подпись обучающегося

«09» 09 2022 г.
Рутковская А.А.
Ф.И.О. обучающегося

СОДЕРЖАНИЕ

ВВЕДЕНИЕ	5
1 ТЕОРЕТИЧЕСКАЯ ЧАСТЬ.....	6
1.1 Ключевые понятия временных рядов.....	6
1.2 Основные понятия корреляционного анализа	13
2 ПРАКТИЧЕСКАЯ ЧАСТЬ	19
2.1 Корреляционный анализ на примере временного ряда.....	19
2.2 Прогнозирование временного ряда с помощью ARIMA	22
ЗАКЛЮЧЕНИЕ	32
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ	33
ПРИЛОЖЕНИЯ.....	36

ВВЕДЕНИЕ

Курс валюты всё время меняется. Для этого нужно пристально следить за рынком и строить логические и математические модели, которые подскажут, что может произойти в ближайшем будущем с разными валютами.

Цель работы — провести анализ различных валют и найти между ними зависимость с использованием корреляционного анализа, а после с помощью интегрированной модели авторегрессии ARIMA предсказать дальнейшее поведение временного ряда изменения стоимости валют в рублях (на примере отношения доллара к рублю).

Задачи, решаемые в данной курсовой работе:

- изучение научной и методической литературы по исследуемой проблеме;
- использование знаний математической статистики с использованием современных средств обработки данных: аналитической платформы Loginom;
- построение интегрированной модели авторегрессии ARIMA с использованием наиболее коррелирующих с прогнозируемым рядом рядов;
- обучение качественному оформлению документации.

Данная курсовая работа является актуальной для сферы покупки и продаж различных валют. Корреляционный анализ временных рядов актуален для предпринимателей, занимающихся оценкой рынка валют.

1 ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

1.1 Ключевые понятия временных рядов

Временной ряд (ряд динамики) — это индексированная последовательность точек данных, отражающих развитие во времени некоторого процесса, зафиксированных через равные промежутки времени. Временной ряд состоит из двух элементов: моментов или периодов времени наблюдений, к которым относятся статистические данные, и самих данных, называемых уровнями ряда. Совокупности элементов — время и уровень — называются членами временного ряда.

Временные ряды классифицируются по следующим признакам:

1. В зависимости от способа выражения уровней ряды динамики подразделяются на ряды абсолютных, относительных и средних величин.
2. В зависимости от того, как выражают уровни ряда состояние явления на определенные моменты времени или его величину за определенные интервалы времени, различают моментные и интервальные временные ряды.
3. В зависимости от расстояния между уровнями ряды подразделяются на ряды динамики с равностоящими и неравностоящими уровнями во времени.
4. В зависимости от наличия закономерностей в изучаемом процессе хронологические ряды подразделяются на стационарные и нестационарные. Если математическое ожидание значения признака и дисперсия постоянны (не зависят от времени), то процесс считают стационарным, и временные ряды считают стационарными. Однако экономические процессы во времени обычно не являются стационарными, так как содержат основную тенденцию развития.

5. По числу показателей выделяют изолированные и комплексные (многомерные) временные ряды. [1.5]

Обычно в поведении временного ряда выявляют две основные тенденции — тренд и периодические колебания. При этом под трендом понимают зависимость экономической величины линейного, квадратичного или иного типа, которую выявляют тем или иным способом сглаживания, либо расчетным путем, в частности, с помощью метода наименьших квадратов. Другими словами, тренд — это очищенная от случайностей основная тенденция временного ряда.

Временной ряд обычно колеблется вокруг тренда, причем отклонения от тренда часто обнаруживают правильность. Часто это связано с естественной или назначенной периодичностью. Также ряд динамики подвержен влиянию факторов эволюционного и колебательного характера. Влияния эволюционного, долговременного характера — это изменения, определяющие некое общее направление развития, многолетнюю эволюцию, которая складывается из других математических и случайных колебаний. Влияния колебательного характера — это циклические (конъюнктурные) и сезонные колебания. [1.8]

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

1. Тенденция, характеризующая общее направление динамики изучаемого явления. Аналитическая тенденция выражается некоторой функцией времени, называемой трендом. Будем обозначать эту функцию через $T(t)$.
2. Циклическая, или периодическая, составляющая, характеризующая циклические или периодические колебания изучаемого явления. Колебания представляют собой отклонения фактических уровней ряда от тренда. Сезонные колебания (S) — это периодические колебания, которые имеют определенный и постоянный период, равный годовому промежутку. Конъюнктивные колебания (K) —

колебания, связанные с большими экономическими циклами, период таких колебаний — несколько лет.

3. Случайная составляющая (E), которая является результатом воздействия множества случайных факторов.

Некоторая функция f от указанных составляющих называется уровнем (\tilde{x}) временного ряда:

$$\tilde{x} = f(T, K, S, E).$$

В зависимости от взаимосвязи между составляющими может быть построена либо аддитивная модель временного ряда, которая характеризуется тем, что характер циклических и сезонных колебаний остается постоянным:

$$\tilde{x} = T + K + S + E,$$

либо мультипликативная модель временного ряда, которая характеризуется постоянством циклических и сезонных колебаний только по отношению к тренду:

$$\tilde{x} = T \cdot K \cdot S \cdot E$$

временного ряда. [1.4]

В процессе формирования значений временных рядов не всегда участвуют все четыре компоненты. Однако во всех случаях предполагается наличие случайной компоненты. Следует обратить внимание на то, что в отличие от случайной составляющей, тренд, сезонная и циклическая компоненты являются закономерными, неслучайными. Важнейшей классической задачей при исследовании временных рядов является выявление и статистическая оценка основной тенденции развития изучаемого процесса и отклонений от нее.

Отметим основные этапы анализа временных рядов:

- графическое представление и описание поведения временного ряда;
- выделение и удаление закономерных (неслучайных) составляющих временного ряда (тренда, сезонных и циклических составляющих);
- сглаживание и фильтрация (удаление низкочастотных и высокочастотных составляющих временного ряда);
- исследование случайной составляющей временного ряда, построение и проверка адекватности математической модели для ее описания;
- прогнозирование развития изучаемого процесса на основе имеющегося временного ряда;
- исследование взаимосвязи между различными временными рядами.

Для анализа временных рядов используется математическая модель ARIMA (AutoRegressive Integrated Moving Average), объединяющая в себе интегрированную авторегрессию, скользящее среднее и возможность учёта дополнительных внешних факторов. [1.9]

Модели ARIMA применяются для решения задач, в которых требуется построить прогноз на основе имеющихся данных, то есть вычислить последующие значения ряда на основе предыдущих. Прогноз представляет собой обоснованное суждение о возможном состоянии исследовании процесса или объекта в будущем.

Составляющие модели ARIMA:

1. AR — авторегрессия. Эта модель работает с зависимостью между будущим значением и некоторым количеством запаздывающих значений.
2. I — интегрирующий член. Использование дифференцирования необработанных наблюдений, чтобы сделать временной ряд стационарным.
3. MA — скользящее среднее. Эта модель, вместо использования прошлых значений прогнозируемой переменной в регрессии работает с прошлыми ошибками прогноза в регрессионной модели.

По срокам прогнозы временного ряда делятся на:

- оперативные (до 1 месяца);
- краткосрочные (от 1 месяца до 1 года);
- среднесрочные (от 1 до 5 лет);
- долгосрочные (от 5 до 15 лет);
- дальносрочные (15 лет и более).

Качество построенной модели ARIMA можно судить по нескольким параметрам:

- информационный критерий Акаике (AIC), а также его скорректированная версия (AICc) для рядов с малым обучающим множеством;
- информационный критерий Байеса (BIC);
- коэффициент детерминации (R^2). [1.9]

Информационный критерий Акаике — критерий для выбора лучшей из нескольких статистических моделей, построенных на одном и том же наборе данных и использующих логарифмическую функцию правдоподобия. Предложен Хироцугу Акаикэ в 1974 году. Критерий является не статистическим, а информационным, поскольку основан на оценке потери информации при уменьшении числа параметров модели. Критерий позволяет найти компромисс между сложностью модели (числом параметров) и её точностью. Лучшей признаётся та модель, для которой значение AIC минимально.

В общем случае AIC вычисляется по формуле:

$$AIC = 2k - 2 \ln L$$

где k — число параметров модели;

L — максимизированное значение функции правдоподобия модели.

Так же существует информационный критерий Акаике скорректированный, который применяется для выборок малого размера, когда отношение числа содержащихся в выборке примеров к числу параметров модели меньше 40. Значение критерия вычисляется следующим образом:

$$AIC_c = AIC + 2k \frac{k + 1}{n - k - 1}$$

где k — число параметров модели;

n — объём обучающей выборки;

AIC — информационный критерий Акаике.

Информационный критерий Байеса основан на использовании функции правдоподобия и тесно связан с информационным критерием Акаике. Предложен Гидеоном Шварцем в 1978 году. Однако, поскольку при разработке подхода автор использовал и адаптировал идеи Байеса, за критерием закрепилось два названия: Шварца и Байеса.

В основе подхода лежит тот факт, что при увеличении числа параметров модели значение функции правдоподобия растёт, но при этом возможно наступление эффекта переобучения. Под переобучением модели в данном случае понимается, что когда параметров модели оказывается слишком много, то доле каждого из них в объясняющей способности модели становится малой и они теряют свою значимость.

Поэтому задача выбора модели заключается в том, чтобы включить в неё минимум параметров, которые, тем не менее, вносили бы наибольший вклад в значение функции правдоподобия. Значение критерия вычисляется по формуле:

$$BIC = -2 \ln L + k \ln n$$

где L — максимальное значение функции правдоподобия наблюдаемой выборки с известным числом параметров;

k — число параметров модели;

n — объём обучающей выборки.

Коэффициент детерминации — статистический показатель, отражающий объединяющую способность регрессии (т.е. зависимости математического

ожидания случайной величины от одной или нескольких случайных величин) $f: X \rightarrow Y$ и определяемый как доля дисперсии зависимой переменной, объясненная регрессионной моделью с данным набором независимых переменных. Данный показатель является статистической мерой согласия, с помощью которой можно определить, насколько уравнение регрессии соответствует реальным данным, на которых она построена, и вычисляется по формуле:

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

где $\sum_i (y_i - \hat{y})^2$ — сумма квадратов остатков регрессии;

$\sum_i (y_i - \bar{y})^2$ — сумма квадратов отклонений точек данных от среднего значения.

Коэффициент детерминации изменяется в диапазоне от 0 до 1. Если он равен 1, это соответствует идеальной модели, когда все точки наблюдений лежат точно на линии уравнения, то есть сумма квадратов их отклонений равна 0. Если коэффициент детерминации равен 0, это означает, что связь между переменными регрессионной модели отсутствует и вместо неё для оценки значения выходной переменной можно использовать простое среднее её наблюдаемых значений.

На практике, если коэффициент детерминации близок к 1, это указывает на то, что модель работает очень хорошо (имеет высокую значимость), а если к 0, то это означает очень низкую значимость модели, когда входная переменная плохо «объясняет» поведение выходной, т.е. линейная зависимость между ними отсутствует. Очевидно, что такая модель будет иметь низкую эффективность.

Таким образом, временной ряд — это последовательность значений, описывающих протекающий во времени процесс. Для проведения анализа временных рядов необходимо построение и проверка адекватности математической модели. В процессе формирования значений временных рядов не всегда участвуют все четыре компоненты и для определения состава

компонентов (структуры временного ряда) в модели временного ряда проводят корреляционный анализ и строят автокорреляционную функцию.

1.2 Основные понятия корреляционного анализа

Для определения состава компонентов (структуры временного ряда) в модели временного ряда строят автокорреляционную функцию. Автокорреляцией называется корреляционная связь между последовательными уровнями одного и того же временного ряда, сдвинутыми на определенный промежуток времени (L), иначе говоря, автокорреляция — это связь между рядом x_1, x_2, \dots, x_{n-1} и рядом $x_{1+L}, x_{2+L}, \dots, x_n$, где L — положительное целое число. [1.3]

Автокорреляция может быть измерена коэффициентом автокорреляции:

$$r_{t,t-L} = \frac{\overline{x_t \cdot x_{t-L}} - \bar{x}_t \cdot \bar{x}_{t-L}}{\sigma_t \cdot \sigma_{t-L}},$$

где

$$\overline{x_t \cdot x_{t-L}} = \frac{\sum_{i=1+L}^n x_i \cdot x_{i-L}}{n-L},$$

$$\bar{x}_t = \frac{\sum_{i=1+L}^n x_i}{n-L},$$

— это средний уровень ряда $(x_{1+L}, x_{2+L}, \dots, x_n)$,

$$\bar{x}_{t-L} = \frac{\sum_{i=1+L}^n x_{i-L}}{n-L}$$

— это средний уровень ряда $(x_1, x_2, \dots, x_{n-L})$, σ_t, σ_{t-L} — средние квадратические отклонения для рядов $(x_{1+L}, x_{2+L}, \dots, x_n)$ и $(x_1, x_2, \dots, x_{n-L})$ соответственно. [1.4]

Лаг (сдвиг во времени) определяет порядок коэффициента автокорреляции. Если $L = 1$, то имеем коэффициент автокорреляции первого порядка $r_{t,t-1}$, если $L = 2$, то коэффициент автокорреляции второго порядка $r_{t,t-2}$ и т.д. С увеличением лага число пар значений, по которым рассчитывается коэффициент автокорреляции, уменьшается. Для обеспечения статистической достоверности коэффициентов автокорреляции максимальный лаг должен быть не больше $\frac{n}{4}$, где n — количество временных периодов.

Если вычислены несколько коэффициентов автокорреляции, можно определить лаг (L), при котором автокорреляция ($r_{t,t-L}$) наиболее высокая, выявив тем самым структуру временного ряда.

1. Если наиболее высоким оказывается значение коэффициента автокорреляции первого порядка $r_{t,t-1}$, то исследуемый ряд содержит только тенденцию.
2. Если наиболее высоким оказался коэффициент автокорреляции $r_{t,t-L}$ порядка L , то ряд содержит колебания с периодом L .
3. Если ни один из $r_{t,t-L}$ не является значимым, можно сделать одно из двух предположений:
 - либо ряд не содержит тенденций и циклических колебаний, а его уровень определяется только случайной компонентой;
 - либо ряд содержит сильную нелинейную тенденцию, для выявления которой нужно провести дополнительный анализ.

Изучая явление или процесс во времени, аналитики часто оценивают взаимосвязь изменений уровней двух или более рядов динамики, отличающихся по содержанию, но связанных между собой. Корреляционный анализ — это совокупность методов обработки данных с целью обнаружения статистической взаимосвязи между случайными величинами или признаками.

Впервые элементы корреляционного анализа в научных исследованиях начал применять французский палеонтолог Жорж Кювье, который и ввёл в научных обиход термин «корреляция». Он разработал «закон корреляции» частей и органов живых существ, с помощью которого можно восстановить облик ископаемого животного, имея в распоряжении лишь часть его останков. В статистике слово «корреляция» первым стал использовать английский биолог и статистик Фрэнсис Гальтон в конце XIX века. Значительный вклад в развитие теории корреляционного анализа внесли Карл Пирсон, Чарльз Спирмен, Морис Кендалл и другие.

В самом общем виде принятие гипотезы о наличии корреляции означает, что изменение значения переменной X произойдет одновременно с пропорциональным изменением значения Y . Если с изменением значения одной из переменных вторая может в определенных пределах принимать любые значения с некоторыми вероятностями, но ее среднее значение или иные статистические характеристики изменяются по определенному закону, то связь является статистической.

Корреляционной связью называют важнейший частный случай статистической связи, состоящий в том, что разным значениям одной переменной соответствуют различные средние значения другой.

Корреляционная связь не предполагает причинной зависимости между переменными. Корреляционный анализ может использоваться для определения тесноты и направления связи и в причинных моделях. Инструментами корреляционного анализа являются разнообразные меры связи. Выбор мер (коэффициентов) связи зависит от способов измерения переменных и характера связи между ними. [1.10]

Поскольку корреляционная связь является статистической, первым условием возможности ее изучения является общее условие всякого статистического исследования: наличие данных для достаточно большой совокупности явлений. Вторым условием закономерного проявления

корреляционной связи служит условие, обеспечивающее надежное выражение закономерности в средней величине.

В соответствии с сущностью корреляционной связи ее изучение имеет две цели:

1. Измерение параметров уравнения, выражающего связь средних значений зависимой переменной со значениями независимой переменной (зависимость средних величин результативного признака от значений одного или нескольких факторных признаков).
2. Измерение тесноты связи двух (или большего числа) признаков между собой.

Корреляционный анализ для двух случайных величин включает в себе:

- построение корреляционного поля и составление корреляционной таблицы;
- вычисление выборочных коэффициентов корреляции и корреляционных отношений;
- проверка статистической гипотезы значимости связи.

Рекомендуется методами корреляционного анализа решать следующие задачи [1.1]:

1. Взаимосвязь. Установление наличия зависимости между двумя признаками и определение её силы.
2. Прогнозирование. Предсказание поведения одного признака на основе изменения другого, коррелирующего с первым.
3. Отбор переменных. Корреляционный анализ позволяет производить выбор набора входных переменных для аналитической модели, в наименьшей степени коррелирующих между собой и в наибольшей степени коррелирующих с выходной переменной. Это позволяет сделать работу аналитических моделей более точной и устойчивой.

Сила корреляционной связи между двумя переменными характеризуется с помощью коэффициента корреляции. Если коэффициент корреляции описывает

связь между двумя случайными величинами, то он называется простым, если между одной случайно величиной и их группой, то множественным. [1.11]

Простой коэффициент корреляции (Пирсона) вычисляется по формуле:

$$r = \frac{\sum_{i=1}^n (x_n - \bar{x})(y_n - \bar{y})}{n\sigma_x\sigma_y}$$

где n — число статических переменных;

x и y — случайные переменные. [1.7]

Коэффициент корреляции Пирсона описывает степень линейной связи и применим к непрерывным величинам. Значения коэффициента корреляции расположены в диапазоне от -1 до 1 и интерпретируются следующим образом:

- если коэффициент корреляции близок к 1, то между переменными наблюдается положительная корреляция, т. е. отмечается высокая степень связи между переменными. В данном случае, если значения переменной x будет возрастать, то и выходная переменная также будет увеличиваться;
- если коэффициент корреляции близок к -1, это означает, что между переменными имеет место сильная корреляция, т. е. поведение выходной переменной будет противоположным поведению входной. Если значение x будет возрастать, то y будет уменьшаться, и наоборот;
- промежуточные значения, близкие к 0, будут указывать на слабую корреляцию между переменными и, соответственно, низкую зависимость. Поведение переменной x не будет совсем (или почти совсем) влиять на поведение y (и наоборот).

В Таблице 1 приведена зависимость степени линейной связи от модуля коэффициента корреляции.

Таблица 1 — Значения коэффициента корреляции

Связь	Модуль коэффициента корреляции
отсутствует	0-0,1
слабая	0,1-0,4
заметная (умеренная)	0,4-0,7
сильная (высокая)	0,7-0,9
значительная (очень высокая), близкая к функциональной	0,9-1

Очевидно, что если корреляция между переменными высокая, то, зная поведение входной переменной, проще предсказать поведение выходной, и полученное предсказание будет точнее. Таким образом, на этапе предобработки данных можно отобрать временные ряды, которые наиболее коррелируют с исследуемым временным рядом, и использовать их в интегрированной модели авторегрессии ARIMA в качестве входных данных.

Для прогнозирования развития изучаемого процесса проводится исследование взаимосвязи между различными временными рядами для выявления наиболее коррелируемых. Тогда для нахождения прогнозных значений можно опираться не только на данные исходного временного ряда, но и значительно коррелируемого с ним.

Далее на практике с использованием полученных теоретических знаний проведем корреляционный анализ с целью выявления коррелируемости двух временных рядов и построим прогнозирование одного временного ряда на основании уже имеющихся данных ряда и данных значительно коррелируемого с ним временного ряда. А также с помощью информационных критериев проверим качество построенной математической модели.

2 ПРАКТИЧЕСКАЯ ЧАСТЬ

2.1 Корреляционный анализ на примере временного ряда

Для проведения корреляционного анализа временных необходимо провести поиск подходящих данных для последующей корреляции и выборки, какие данные лучше всего коррелируют с исследуемыми. [2.1]

В качестве исследуемых данных на вход подаётся временной ряд о стоимости евро в рублях, и в качестве поиска коррелируемого с ним - временной ряд стоимости доллара в рублях. [2.3]

Эти данные включают в себе следующие характеристики:

- Date — дата, тип данных дата;
- Open_EUR — стоимость евро на момент открытия торгов, вещественного типа данных;
- Close_EUR — стоимость евро на момент закрытия торгов (именно эти данные и будут предсказываться), вещественного типа данных.
- Open_USD — стоимость доллара на момент открытия торгов, вещественного типа данных;
- Close_USD — стоимость доллара на момент закрытия торгов, вещественного типа данных.

Каждый из двух временных рядов представлен в трёх вариантах: стоимость по дням, по неделям и по месяцам.

Данные, поданные на вход, необходимо исследовать, используя low-code платформу Loginom.

Корреляцию проводим для трех различных случаев: корреляция стоимости евро и доллара на момент закрытия торгов по дням, по неделям и по месяцам. Вначале делаем слияние по дате, для того чтобы получить один общий набор данных, вместо исходных двух. Результат слияния по месяцам представлен на

Рисунке 1, где оранжевым цветом изображен временной ряд для стоимости доллара, а синим для стоимости евро.

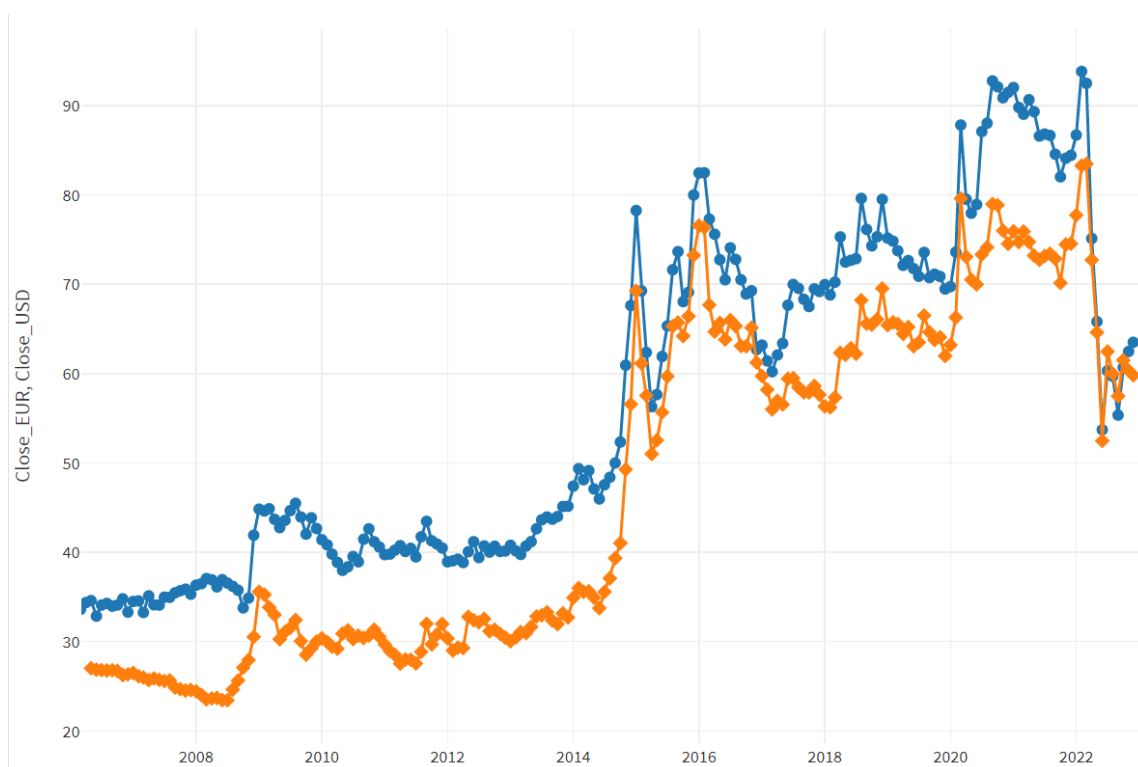


Рисунок 1 — Результат слияния двух наборов данных

Далее проводим фильтрацию строк, чтобы избежать пустых значений. После фильтрации строк построим автокорреляционную функцию и проведем корреляционный анализ.

Лаг определяет порядок коэффициента автокорреляции. Если $L = 1$, то имеем коэффициент автокорреляции первого порядка $r_{t,t-1}$, если $L = 2$, то коэффициент автокорреляции второго порядка $r_{t,t-2}$ и т.д. Для обеспечения статистической достоверности коэффициентов автокорреляции максимальный лаг должен быть не больше $\frac{n}{4}$, где n — количество временных периодов. В нашем случае имеется $n = 208$ временных периодов и лаг равный 38. Так как вычислены несколько коэффициентов автокорреляции, можно определить лаг (L), при котором автокорреляция наиболее высокая, выявив тем самым структуру временного ряда. Наиболее высоким оказывается значение коэффициента автокорреляции первого порядка $r_{t,t-1}$, значит исследуемый ряд содержит только тенденцию.

Далее проведем корреляционный анализ на примере временных рядов по месяцам. Построим корреляционное поле по составленной корреляционной таблице, где по горизонтальной оси будут значения стоимости доллара на момент закрытия торгов, а по вертикальной значения стоимости евро на момент закрытия торгов, и изобразим на Рисунке 2. [2.1]

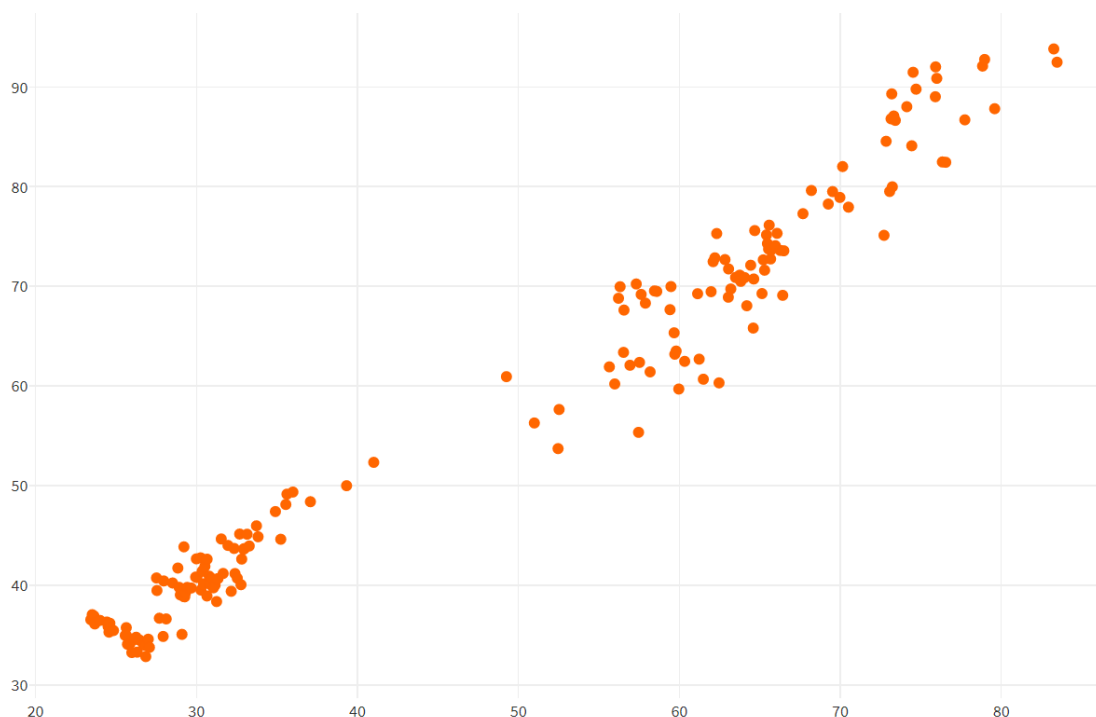


Рисунок 2 — Корреляционное поле

Анализ Рисунка 2 позволяет сделать вывод о наличии сильной линейной статистической связи между значениями стоимости доллара на момент закрытия торгов и значениями стоимости евро на момент закрытия торгов. При этом связь имеет положительную тенденцию, то есть с ростом переменной X наблюдается увеличение отклика Y .

На выход подаётся рассчитанный узлом «Корреляционный анализ» коэффициент корреляции Пирсона пары данных. Такой анализ проводим для каждой пары: дни, недели, месяца. [2.11]

Корреляции временных рядов всеми тремя способами необходимо сгруппировать и найти среднее значение, результат подсчета среднего значения коэффициента Пирсона для трех случаев представлен на Рисунке 3.

ab Поле1.Имя	ab Поле2.Имя	9.0 Пирсона Среднее
Close_EUR	Close_USD	0,98

Рисунок 3 — Значение коэффициента Пирсона

По получившимся данным видно, что временной ряд стоимости евро в рублях имеет близкую к функциональной корреляцию с временным рядом стоимость доллара в рублях. Значение коэффициента Пирсона в среднем равно 0.98, что соответствует очень высокой корреляции.

Следовательно, при прогнозировании временного ряда при помощи математической модели ARIMA можно использовать данные временного ряда стоимости доллара в рублях в качестве входного параметра.

2.2 Прогнозирование временного ряда с помощью ARIMA

Для наиболее точного предсказания строится несколько интегрированных моделей авторегрессии ARIMA, на вход которым подаются различные параметры и настройки которых частично друг от друга отличаются. Для каждого из предсказаний — на 5 дней, на 5 недель и на 5 месяцев для соответствующих наборов данных по дням, неделям и месяцам — используем по 5 моделей ARIMA. [2.9]

Для построения корректного прогноза необходимо достаточное количество данных о предыдущих значениях стоимости евро и доллара на моменты закрытия торгов, но также нужно учитывать то, что совсем старые данные не могут быть использованы, иначе прогноз будет некорректным. Поэтому при построении прогноза в моделях ARIMA будем задавать разные промежутки дат, для обучения моделей на разных входных множествах. Входное поле — поле, соответствующее внешнему фактору, влияющему на прогноз. [2.2]

Для каждого из трех случаев, мы будем использовать следующие модели ARIMA:

- 2006EUR(close,open), которая в качестве прогнозируемого ряда использует Close_EUR, а в качестве входного — Open_EUR;
- 2006EUR_cl_op USD_cl_сезон, которая в качестве прогнозируемого ряда использует Close_EUR, а в качестве входного — Open_EUR и Close_USD, с расчетом сезонности — 48 недель (год);
- 2018EUR_cl_op_USD_cl, которая в качестве прогнозируемого ряда использует Close_EUR, а в качестве входного — Open_EUR и Close_USD;
- 2018EUR_cl_open USD_cl_сезон, которая в качестве прогнозируемого ряда использует Close_EUR, а в качестве входного — Open_EUR и Close_USD, с расчетом сезонности — 16 недель (треть года);
- 2006EUR_cl USD_op, которая в качестве прогнозируемого ряда использует Close_EUR, а в качестве входного — Open_USD.

После автоматического обучения модели на основе входных данных и прогнозируемого ряда получаем варианты прогноза, свои для каждого ряда. Модель строит прогноз не только на новый период, но и на уже имеющиеся данные. В начале таблицы остаются пустые строки в столбцах с прогнозом, так как на этих строках будет проходить обучение AR части. [2.10]

Насколько правдоподобными могут быть варианты прогнозов можно судить по информационным критериям Акаике (AIC), Байеса (BIC) и коэффициенту детерминации (R^2). Модель прогнозирования тем лучше, чем ниже AIC и BIC . [2.4]

На данные критерии положительно влияет уменьшение остаточной дисперсии и отрицательно — количество включенных параметров. Основным различием между ними является степень жесткости, то есть, насколько велик штраф за большое количество параметров в модели. Критерий Байеса является наиболее жестким критерием, причем, как можно увидеть из приведенной формулы его расчета, в отличие от остальных критериев, его жесткость возрастает с ростом n — количеством параметров. [2.6]

Различие в жесткости проистекает из различия в предъявляемых требованиях к моделям прогнозирования. Критерий AIC направлен на достижение высокой точности прогноза: на минимизацию расхождения между плотностью распределения по истинной модели и по выбранной модели. В основе критерия BIC лежит требование максимизации вероятности выбора истинной модели. [2.5]

Рассмотрим информационные критерии AIC и BIC , которые для этих моделей колеблются от самых неправдоподобных значений ($AIC = 17217,73$ и $BIC = 17450,66$ для модели 2006EUR_cl_or USD_cl_сезон, используемой на временных рядах по дням) до правдоподобных ($AIC = 116,12$ и $BIC = 168,98$ для модели 2018EUR_cl_or USD_cl_сезон, используемой на временных рядах по месяцам).

Например, последняя часть предсказанного графика с помощью модели 2018EUR_cl_or USD_cl_сезон на обучающем тестовом множестве из помесечных данных, где на горизонтальной оси расположены даты, а на вертикальной оси: синим — значение стоимости евро на момент закрытия торгов и оранжевым — прогнозные значения, изображена на Рисунке 4.

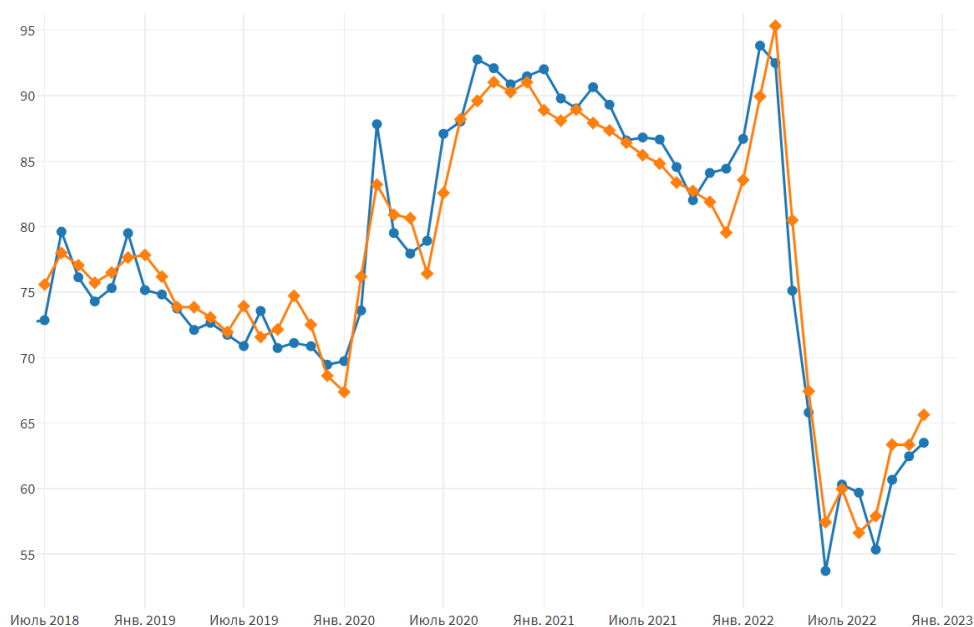


Рисунок 4 — Предсказанные данные и реальные значения

Пусть значение критериев Акаике и Байеса чуть больше ($AIC = 225,37$ и $BIC = 251,22$ — одни из средних показателей среди всех моделей), чем у некоторых других, но там, где модель предсказала падение или подъем стоимости евро, с большой вероятностью произошло падение или подъем в действительности. Это указывает на то, что построенные нами модели ARIMA дают достаточно достоверный прогноз. Покажем это на Рисунке 5, где изобразим прогнозные значения (оранжевым цветом) с ошибкой аппроксимации (синим цветом), которая высчитывается на из разности прогнозного значения и фактической стоимости.

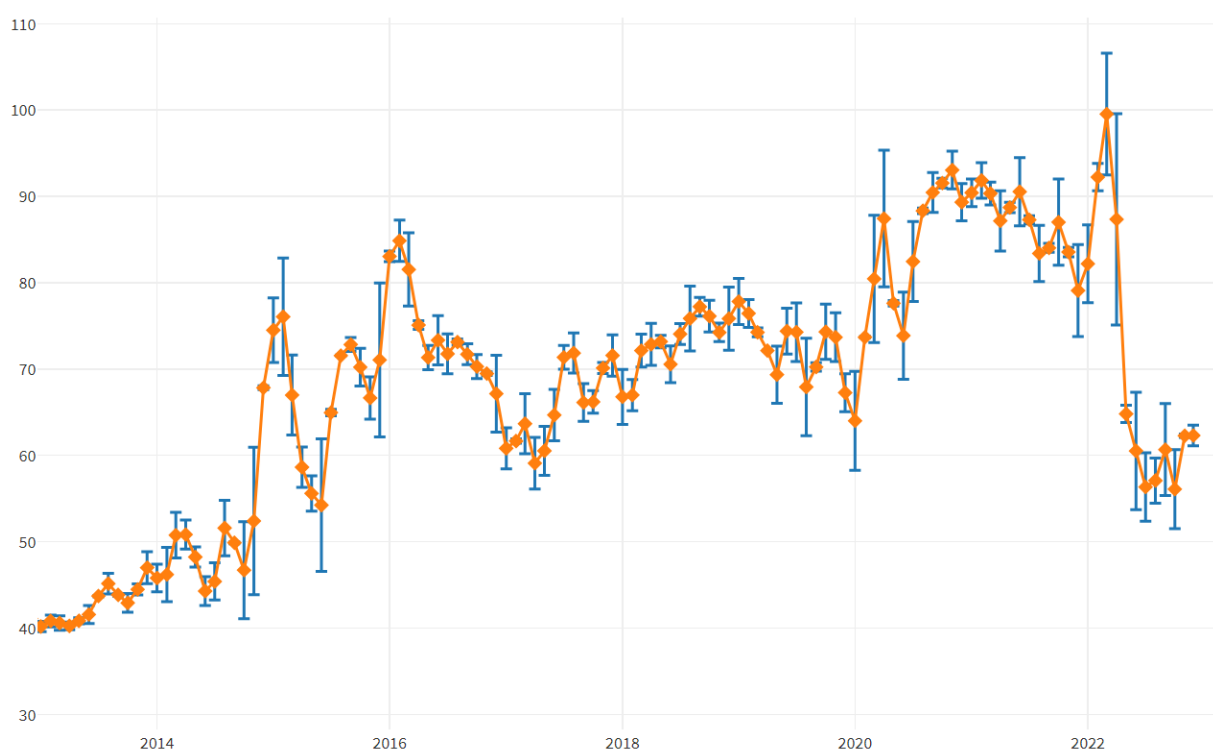


Рисунок 5 — Прогнозные значения с ошибкой аппроксимации

Как видно из данного графика, модель достаточно хорошо предсказывает рост и падение стоимости евро, но при резких изменениях стоимости возникает большая погрешность.

Приведем еще один пример предсказанного графика с помощью модели 2018EUR_c1_or_USD_c1 на обучающем тестовом множестве из понедельных данных, где на горизонтальной оси расположены даты, а на вертикальной оси: синим — ошибка аппроксимации и оранжевым — прогнозные значения стоимости евро, изображена на Рисунке 6.

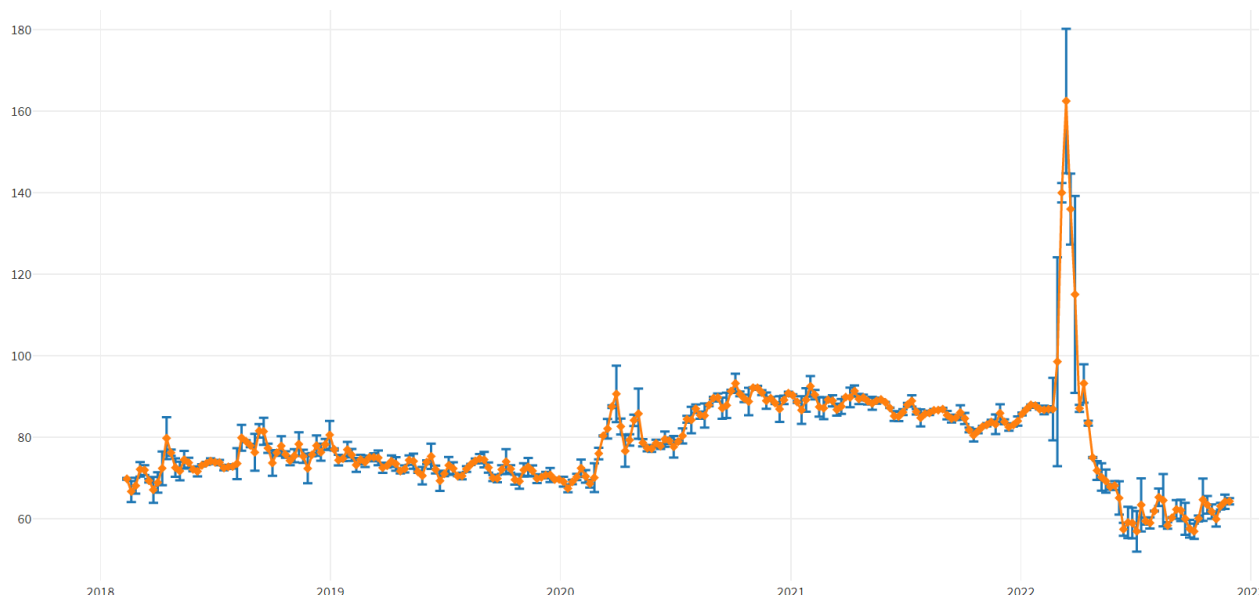


Рисунок 6 – Пример прогнозных значений с ошибкой аппроксимации

Поэтому на практике, однако же, не всегда можно верить лишь одной какой-то модели. Если кто-то предскажет больше, чем нужно, а кто-то — меньше, то с большой долей вероятности ответ будет лежать где-то между ними. Поэтому объединим все результаты для каждого набора данных — дни, недели, месяца — по отдельности и разделим на количество этих результатов.

Воспользуемся узлом «Фильтр строк», чтобы настроить нужный диапазон дат. Также в отдельную таблицу зададим необходимые даты для прогнозируемых значений стоимости евро, исключая выходные дни (в эти дни нет торгов, так как биржа валют закрыта), для этого создадим генератор календаря и на входных полях установим необходимый промежуток времени, после чего с помощью узла «Фильтр строк» убираем выходные дни.

Чтобы объединить данные прогнозов с моделей ARIMA воспользуемся узлом «соединение», в который на порт главной таблицы подадим выход узла «Фильтр строк», в котором были отобраны нужные данные. Получим среднее значение для каждого случая с помощью узла «Калькулятор», в котором на вход подадим прогнозные значения и посчитаем их среднее.

Изобразим полученный результат на графике временного ряда стоимости евро на момент открытия торгов по месяцам, также включим в данный график прогнозируемые значения стоимости евро на пять месяцев. На Рисунке 7 синим

цветом изображены фактические значения стоимости евро за длительный период, а оранжевым цветом приведен краткосрочный прогноз стоимости евро на ближайшие пять месяцев.

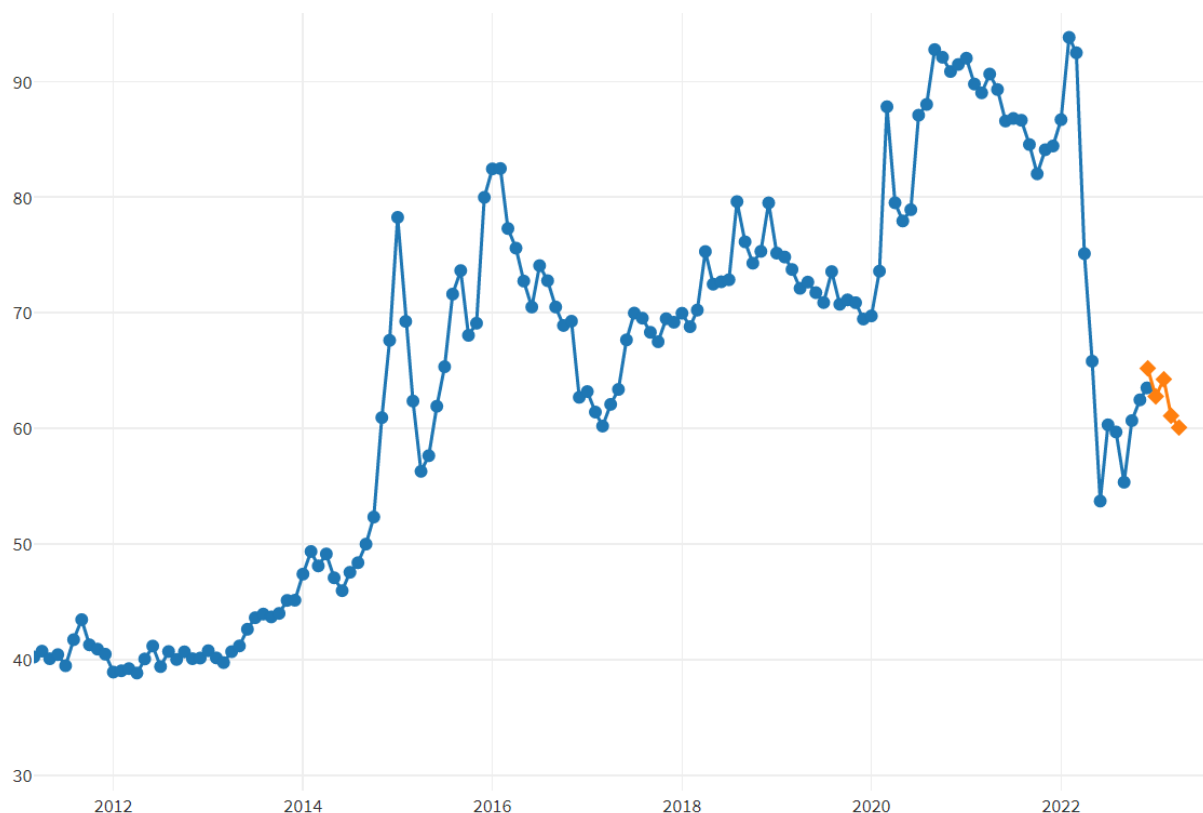


Рисунок 7 — Временной ряд стоимости евро с прогнозируемыми значениями

В результате для того, чтобы визуальное оценить прогнозные значения и их корреляцию с фактическими показателями, построим графики исходных значений стоимости евро на момент закрытия торгов (оранжевым цветом) и прогноза этой величины (синим цветом), полученного в результате использования моделей ARIMA, и проанализируем их. [2.7]

Изобразим на Рисунке 8 результат прогнозирования по месяцам. Здесь голубая зона отвечает за обучение модели на обучающем тестовом множестве из помесечных данных, на этом временном отрезке возможно построение только кривой фактических данных. Красная — за построение прогнозных значений стоимости при наличии фактических значений стоимости евро, на графике присутствуют сразу две кривые, что позволяет визуальное оценить, насколько близки полученные в результате работы модулей ARIMA прогнозные значения к фактическим. И зеленая — за прогнозные значения стоимости евро на пять

месяцев вперед, начиная с 01.12.2022, на графике изображена одна кривая прогноза.

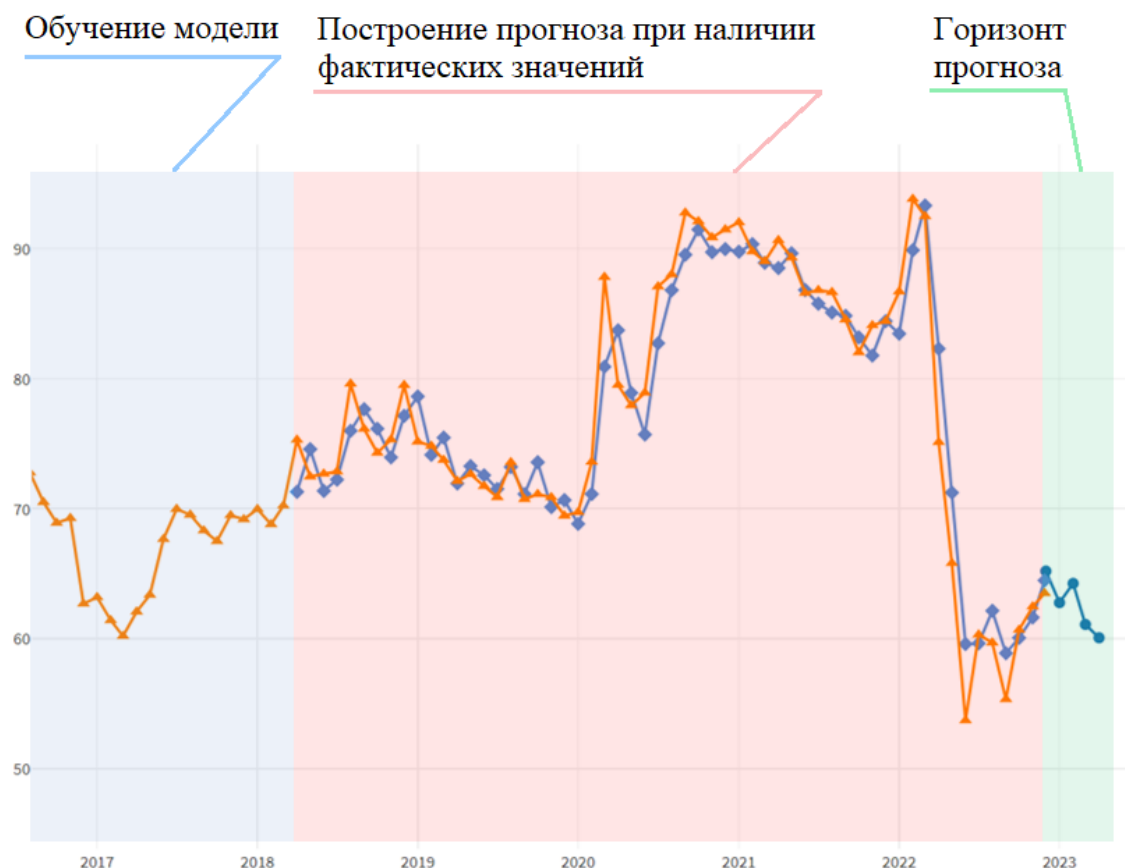


Рисунок 8 — Результат прогнозирования по месяцам

Как видно из данного графика, модель хорошо предсказывает рост и падение стоимости евро, но при резком росте или резком падении значения могут отличаться с большой погрешностью. Также в следующие пять месяцев, судя по предсказаниям временного ряда, ожидается небольшое падение стоимости евро.

Также сделаем прогноз на пять недель вперед и изобразим график прогнозирования на Рисунке 9. Здесь голубая зона отвечает за обучение модели на обучающем тестовом множестве из понедельных данных, на этом временном отрезке возможно построение только кривой фактических данных. Красная зона — за построение прогнозных значений стоимости при наличии фактических значений стоимости евро. И зеленая — за прогнозные значения стоимости евро на пять недель вперед.

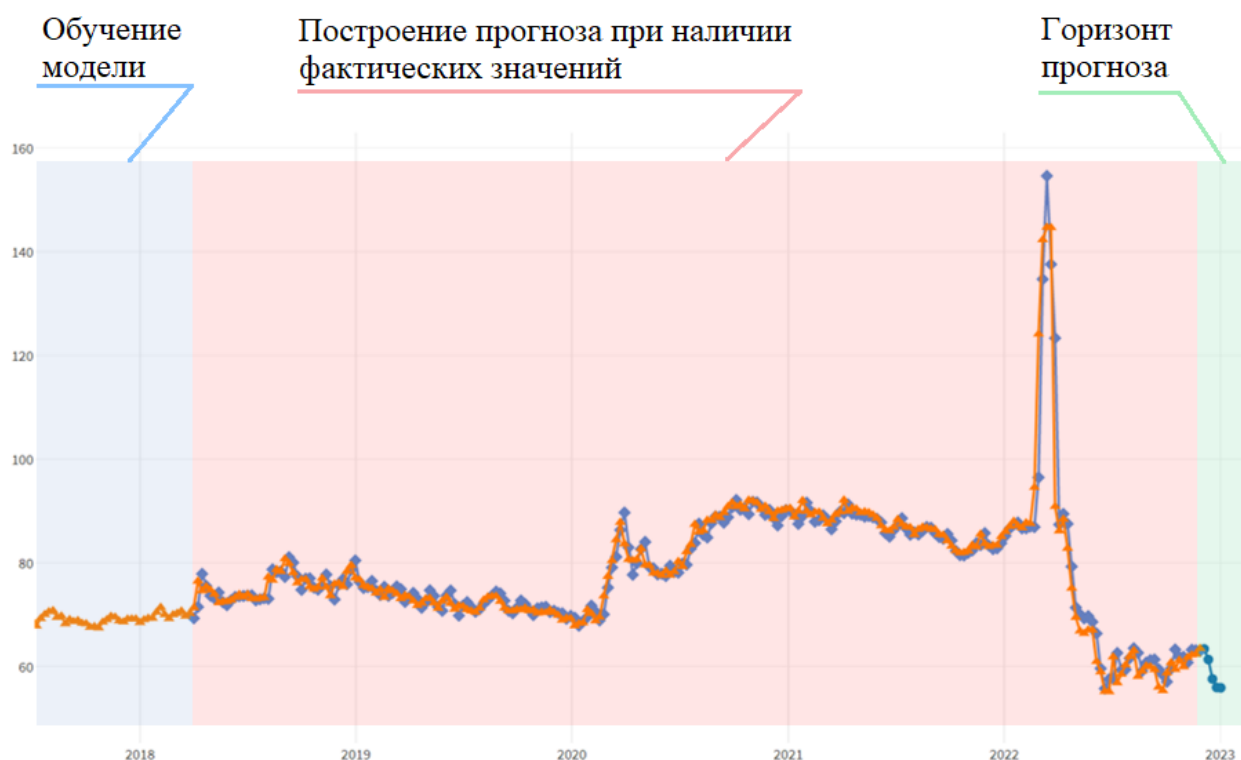


Рисунок 9 — Результат прогнозирования по неделям

Система моделей предсказала, что в течении следующих пяти недель будет небольшое падение, а затем значение стоимости евро выровняется. Можно предположить, что на следующую неделю, шестую, стоимость евро будет примерно равна значению на пятой неделе.

Сделаем оперативный прогноз на ближайшие пять дней и изобразим на Рисунке 10. Такой прогноз будет достаточно точным, так как он формируется на последних данных моделей. Голубая зона графика представляет собой зону обучения модели на тестовом множестве из данных по дням, на этом временном отрезке возможно построение только кривой фактических данных. Красная зона — за построение прогнозных значений стоимости при наличии фактических значений стоимости евро, на графике изображено две кривые — фактическая стоимость евро и предсказание модели. И зеленая зона — за прогнозные значения стоимости евро на пять дней вперед, на графике изображена одна кривая.



Рисунок 10 — Результат прогнозирования по дням

Модели предсказали, что следующие пять дней стоимость евро будет колебаться примерно на одном и том же уровне.

Результаты проведённого корреляционного анализа и прогнозирования временных рядов позволяют сделать выводы, представляющие интерес для исследования стоимости валют. Математическая модель ARIMA позволяет предсказывать падение и подъем стоимости различных валют на основании данных прошлых лет и стоимости других валют, коррелируемых с данной, пример роста и падения стоимости валют показан в Таблице 2.

Таблица 2 — Предсказание подъема и падения стоимости

Дата и время	Стоимость на момент закрытия торгов, руб.	Прогнозная стоимость, руб.
30.06.2022 00:00	53,71	55,91
01.07.2022 00:00	55,17	55,08
04.07.2022 00:00	119,53	57,88
05.07.2022 00:00	57,53	94,83
06.07.2022 00:00	60,22	73,78
07.07.2022 00:00	63,79	62,06

С помощью данной модели можно предсказывать поведение рынка валют в будущем. Настоящее исследование анализирует рынок и прогнозирует рынок валютных курсов на временном отрезке от пяти дней до пяти месяцев, включая понедельное исследование. Пример результата прогнозирования временного ряда стоимости евро на момент закрытия торгов на пять дней на будущее представлен в Таблице 3.

Таблица 3 — Предсказание подъема и падения стоимости

Дата и время	Прогнозное значение стоимости, руб.
29.11.2022 00:00	63,099240267544324
30.11.2022 00:00	63,61404334219581
01.12.2022 00:00	64,6371326600796
02.12.2022 00:00	63,46634061666921
05.12.2022 00:00	63,465976032022766

Исходя из всего вышесказанного, результаты проведённого корреляционного анализа и прогнозирования временных рядов позволяют сделать выводы, представляющие интерес для исследования стоимости валют. Данные, полученные в ходе выполнения работы, являются актуальными для сферы покупки и продаж различных валют. Проведенный корреляционный анализ временных рядов, а также прогнозирование стоимости одной из валют актуальны для предпринимателей, занимающихся оценкой рынка стоимости валют.

ЗАКЛЮЧЕНИЕ

Временные ряды играют очень большую роль в технологиях анализа данных. Анализ временных рядов позволяет обнаруживать тенденции из закономерности в исследуемых процессах, строить прогнозы и предсказывать будущие изменения в стоимостях валют.

В результате данного курсового проекта проведён корреляционный анализ стоимостей доллара к рублю и евро к рублю и была найдена закономерность между ними, а также проведено прогнозирование временного ряда с помощью интегрированной модели авторегрессии ARIMA на основе стоимости евро и сформированы оперативные и краткосрочные прогнозы их будущей стоимости. Рассмотренные в работе темы были реализованы с помощью low-code платформы Loginom, сценарий проекта представлен в Приложения.

Поставленная цель работы достигнута.

В ходе данной курсовой работы выполнены следующие задачи:

- изучена научная и методическая литература по проблеме;
- произведён корреляционный анализ данных;
- использованы знания математической статистики с использованием современных средств обработки данных: аналитической платформы Loginom;
- построены интегрированные модели авторегрессии ARIMA с использованием наиболее коррелирующих с прогнозируемым рядом рядов;
- пройдено обучение качественному оформлению документации.

Данная курсовая работа является актуальной для сферы покупки и продажи различных валют. Проведенный корреляционный анализ временных рядов и прогнозирование стоимости валют актуальны для предпринимателей, занимающихся оценкой рынка стоимости валют.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

- 1.1. Елисеева И.И. «Общая теория статистики»: Учебник / Елисеева И.И., Юзбашев М.М. — 4-е издание, переработанное и дополненное — Москва: Финансы и Статистика, 2002. — 480 с.
- 1.2. Юрченко Т.В. «Эконометрика: временные ряды»: учебное пособие / Т.В. Юрченко. — Санкт-Петербург: ИЭО СПбУТУиЭ, 2022. — с. 99.
- 1.3. Кизбикенов К.О. «Прогнозирование и временные ряды»: учебное пособие / К. О. Кизбикенов. — Барнаул: АлтГПУ, 2017. — с. 13.
- 1.4. Молотникова А.А. «Основы эконометрики»: учебное пособие / А.А. Молотникова. — Санкт-Петербург: Лань, 2022. — с. 95.
- 1.5. Горобцов А.С., Рыжов Е.Н. «Математические основы цифрового анализа временных рядов»: учебное пособие / А.С. Горобцов, Е.Н. Рыжов. — Волгоград: ВолгГТУ, 2022. — с. 10.
- 1.6. Горлач Б.А., Шахов В.Г. «Математическое моделирование. Построение моделей и численная реализация»: учебное пособие для вузов / Б.А. Горлач, В.Г. Шахов. — 4-е изд., стер. — Санкт-Петербург: Лань, 2022. — с. 272.
- 1.7. Ганичева А.В. «Прикладная статистика»: учебное пособие / А.В. Ганичева. — Санкт-Петербург: Лань, 2022. — с. 88.
- 1.8. Федорова Н.П., Миронова З.А. «Статистика. Общая теория статистики»: учебное пособие / Н.П. Федорова, З.А. Миронова. — Ижевск: Ижевская ГСХА, 2019. — с. 37.
- 1.9. Бабёнышев С.В., Матеров Е.Н. «Математические методы и информационные технологии в научных исследованиях»: учебное пособие / С.В. Бабёнышев, Е.Н. Матеров. — Железногорск: СПСА, 2018. — с. 153.

- 1.10. Круценюк К.Ю. «Корреляционно-регрессионный анализ в эконометрических моделях»: учебное пособие / К.Ю. Круценюк. — Норильск: НГИИ, 2018. — с. 6.
- 1.11. Митина Т.В. «Многомерные случайные величины. Корреляционный анализ»: учебное пособие / Т.В. Митина. — Дубна: Государственный университет «Дубна», 2021. — с. 27.

ПРАКТИЧЕСКАЯ ЧАСТЬ

- 2.1. LoginomHelp [Электронный ресурс]: Корреляционный анализ. — Режим доступа: <https://help.loginom.ru/userguide/processors/scrutiny/correlation-analysis.html>
- 2.2. LoginomHelp [Электронный ресурс]: ARIMAX. — Режим доступа: <https://help.loginom.ru/userguide/processors/datamining/arimax.html>
- 2.3. Yahoo!Finance [Электронный ресурс]. Yahoo Finance. — Режим доступа: <https://finance.yahoo.com/quote/EURRUB%3DX/history?p=EURRUB%3DX>
- 2.4. LoginomHelp [Электронный ресурс]: Коэффициент детерминации. — Режим доступа: <https://wiki.loginom.ru/articles/coefficient-of-determination.html?ysclid=lbi8jld7un74470633>
- 2.5. LoginomHelp [Электронный ресурс]: Информационный критерий Акаике (Akaike's information criterion). — Режим доступа: <https://wiki.loginom.ru/articles/aic.html?ysclid=lbi8bgv6ge305890380>
- 2.6. LoginomHelp [Электронный ресурс]: Информационный критерий Байеса (Bayesian information criterion). — Режим доступа: <https://wiki.loginom.ru/articles/bic.html>
- 2.7. LoginomHelp [Электронный ресурс]: Линейная регрессия (Linear regression). — Режим доступа: <https://wiki.loginom.ru/articles/linear-regression.html>

- 2.8. «Шестая научно-техническая конференция студентов и аспирантов «МИРЭА - Российского технологического университета»: Сборник трудов, 24-29 мая 2021 г.»: сборник научных трудов. — Москва: РТУ МИРЭА, 2021. — с. 224.
- 2.9. LoginomHelp [Электронный ресурс]: Модель авторегрессии скользящего среднего (ARIMA). — Режим доступа: <https://wiki.loginom.ru/articles/arima.html>
- 2.10. Бабёнышев С.В., Матеров Е.Н. «Математические методы и информационные технологии в научных исследованиях»/ С.В. Бабёнышев, Е.Н. Матеров. — Железногорск: СПСА, 2018. — с. 159.
- 2.11. LoginomHelp [Электронный ресурс]: Коэффициент корреляции (Correlation coefficient). — Режим доступа: <https://wiki.loginom.ru/articles/correlation-coefficient.html>

ПРИЛОЖЕНИЯ

Приложение А — Графический материал.

Приложение А

Итоговый сценарий проекта на платформе Loginom.

На Рисунке 11 представлен сценарий проекта на low-code платформе Loginom.

Описание сценария:

- импорт 6 файлов с данными евро и доллара по дням, неделям и месяцам;
- корреляционный анализ данных;
- поиск среднего значения коэффициента Пирсона для различных пар;
- импорт 6 файлов для дальнейшей работы с ARIMA;
- подмодели с 5 различными узлами ARIMA;
- объединение прогнозов и изначальных рядов и визуализация.

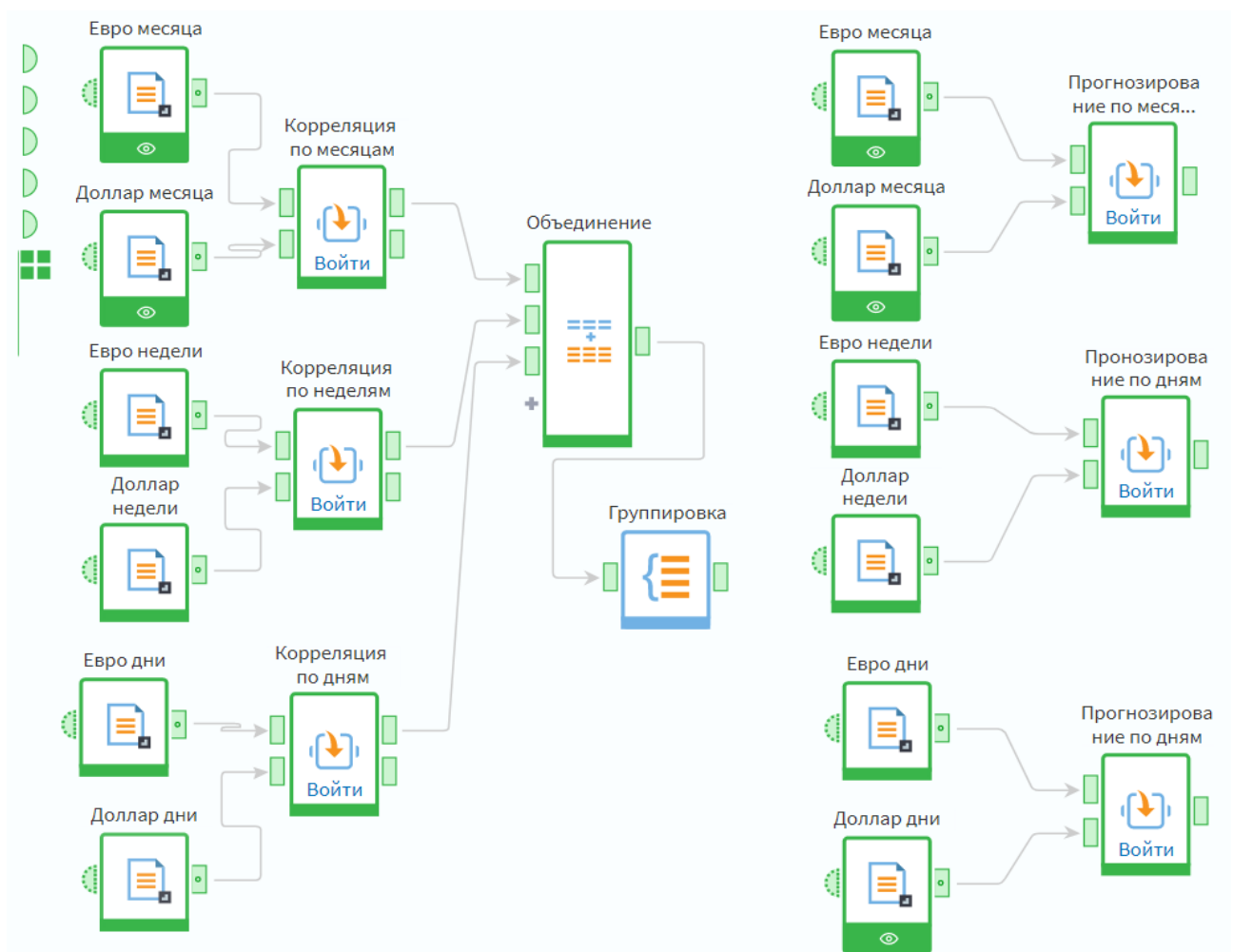


Рисунок 11 – Итоговый сценарий в Loginom