МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и стохастического анализа

ПРИМЕНЕНИЕ МОДЕЛИ АВТОРЕГРЕССИИ ДЛЯ АНАЛИЗА ВРЕМЕННЫХ РЯДОВ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студентки 4 курса 412 группы
направления 01.03.02 — Прикладная математика и информатика

механико-математического факультета
Кротовой Юлии Игоревны

Научный руководитель
ст. преп. ______________ Н. В. Сергеева

Заведующий кафедрой
д. ф.-м. н., доцент ______________ С. П. Сидоров

Саратов 2021

ВВЕДЕНИЕ

Актуальность темы. В современном мире каждый день компании собирают огромное количество данных за какие-либо промежутки времени. Эти данные можно интерпретировать как временные ряды с целью их исследования и анализа. В сфере экономики и других прикладных областях актуальны проблемы исследования динамики процессов и их прогнозирования. Это имеет большое практическое значение: инвесторы смогут решить, в какие активы вложить средства, аналитики смогут отслеживать развитие рынка и контролировать процессы, негативно влияющие на экономику, топ менеджеры смогут принять верное решение на благо будущего своей компании. Существует множество различных способов решения проблем в этой предметной области.

Целью бакалаврской работы является моделирование и анализ различных данных, которые можно представить в виде временных рядов, ввиду их дальнейшего прогнозирования. Основная идея, лежащая в основе метода, состоит в следующем: предполагается, что по известному временному ряду мы определяем какая модель подходит наилучшим образом. Задача работы - исследовать данную процедуру и произвести на ее основе прогнозирование реальных экономических процессов.

Объектом исследования являются модели временных рядов.

Предмет исследования является последовательность значений или стоимости различных предметов, измеренной через равные промежутки времени.

Практическая значимость проводимого исследования состоит в том, что на основании построенных моделей можно проводить исследования реально существующих данных, рассчитывать их характеристики. По результатам прогнозов можно делать выводы о будущих значениях данных, которые могут помочь учесть динамику показателя.

Структура и содержание бакалаврской работы. Работа состоит из введения, пяти разделов, заключения, списка использованных источников, содержащего 27 наименований, и одного приложения. Общий объем работы составляет 69 страниц.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обосновывается актуальность темы работы, формулируется цель работы и решаемые задачи, отмечается практическая значимость полученных результатов.

В первом разделе рассмотрены основные понятия временных рядов. Временной ряд — это последовательность упорядоченных во времени числовых показателей, характеризующих уровень состояния и изменения изучаемого явления. Временным рядом обычно называют ряд значений $x_1, x_2, \ldots, x_T$ анализируемой случайной величины, полученных в последовательные моменты времени $t = 1, 2, \ldots, T$. Случайный процесс x_t $(t = 1, 2, \ldots, T)$ может быть описан с помощью t-мерного распределения вероятностей, и, таким образом, отношение между временным рядом и случайным процессом можно рассматривать как отношение между случайной выборкой и «генеральной совокупностью». В связи с наличием ошибок измерения экономических показателей, наличием случайных флуктуаций, свойственных наблюдаемым системам, при исследовании временных рядов широко применяется вероятностно-статистический подход.

В рамках такого подхода наблюдаемый временной ряд понимается как реализация некоторого случайного процесса. Однако существуют два принципиальных различия между моделью «случайная выборка» и моделью «временной ряд». Эти различия состоят в том, что временной ряд имеет структуру, отличающую его от последовательности независимых случайных величин, так что наблюдения не являются набором совершенно независимых числовых значений. Некоторые элементы структуры ряда иногда можно выявить уже на основании простого визуального анализа графика ряда. Это относится, например, к таким компонентам ряда, как тренд и циклы. Вторым различием является то, что элементы временного ряда не являются одинаково распределенными.

Во втором разделе определяются авторегрессионные модели, модели скользящего среднего и линейного фильтра, описывается идентификация моделей и условия их стационарности. Описаны смешанные авторегрессионные модели.

Модель линейного фильтра

Предполагается, что белый шум ε_t можно трансформировать в процесс x_t при помощи линейного фильтра. Операция линейной фильтрации заключается в вычислении взвешенной суммы предыдущих наблюдений [?]. Таким образом, стационарный стохастический процесс x_t с нулевым математическим ожиданием представим в виде линейной комбинации последовательности возмущений $\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots$, а именно:

$$x_t = \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} + \cdots = \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i}, \tag{1}$$

Выражение (1) можно представить в виде:

$$x_t = (1 + \psi_1 L + \psi_2 L^2 + \ldots)\varepsilon_t,$$

где L - лаговый оператор: $L\varepsilon_t = \varepsilon_{t-1}, L^m \varepsilon_t = \varepsilon_{t-m}$, а также $\psi_0 = 1$ и справедливо следующее выражение:

$$\sum_{i=0}^{\infty} |\psi_i| < \infty, \tag{2}$$

то есть ряд абсолютных значений коэффициентов обязан сходиться.

Выражение вида (1) называются моделью линейного фильтра. Линейный оператор

$$\psi(L) = 1 + \psi_1 L + \psi_2 L^2 + \cdots = \sum_{i=0}^{\infty} \psi_i L_i,$$

называется оператором линейного фильтра или передаточной функцией фильтра.

Для того, чтобы модель линейного фильтра имела смысл, должно выполняться ограничение (2), тогда оператор линейного фильтра называется устойчивым, а процесс будет стационарным. Частными случаями модели линейного фильтра (1) являются модели авторегрессии $AR(p)$, скользящего среднего $MA(q)$, смешанные процессы авторегрессии — скользящего среднего $ARMA(p, q)$.

Модель авторегрессии

В модели авторегрессии текущее значение процесса x_t представляется в виде линейной комбинации конечного числа предыдущих значений процесса и импульса ε_t, при этом в качестве x_t рассматриваются центрированные значения, полученные как отклонения исходных уровней временного ряда от их среднего значения:

$$x_t = \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \ldots \varphi_p x_{t-p} + \varepsilon_t.$$

Модель скользящего среднего

Другой частный случай модели линейного фильтра, широко распространенный в анализе временных рядов, — модель скользящего среднего, когда x_t линейно зависит от конечного числа q предыдущих значений ε_t:

$$x_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}.$$

В третьем разделе описаны свойства стационарных рядов и способы сведения ряда к стационарному.

Без предположений о ряде мы не можем его спрогнозировать. За базовое предположение возьмем стационарность. Это свойство характеризует неизменность ряда во времени, т.е. ряд будет стационарным, если:

1. $E(y_t) = \mu = const$,
2. $V(y_t) = E(y_t - \mu)^2 = \sigma^2 = const$,
3. $cov(y_t, y_{t-k}) = \gamma_k$, где функция γ_k называется автоковариационной функцией и зависит только от лага k (сдвига по времени), не зависит от t (конкретного момента времени).

В четвертом разделе описываются методы прогнозирования, меры качества прогнозов и различные подходы к выбору оптимальной модели. Рассмотрены некоторые простые методы прогнозирования, такие как наивный прогноз, прогноз методом дрейфа, наивный сезонный метод. Определены меры качества прогноза, такие как ошибки, зависящие от масштаба, средние проценты отклонения, масштабированные ошибки.

В пятом разделе строится модель нескольких временных рядов, продемонстрированы различные способы сведения ряда к стационарному, подобраны оптимальные модели, которые дают качественные прогнозы.

Рассмотрим ряд ежедневной стоимости акций компании Яндекс за 250 дней на рисунке 1. Данные взяты из открытого интернет-источника. Код анализа ряда представлен в приложении А. Визуально ряд не является стационарным, заметен тренд на увеличение. График похож на процесс случайного блуждания, так как текущее значение сильно зависит от предыдущего.

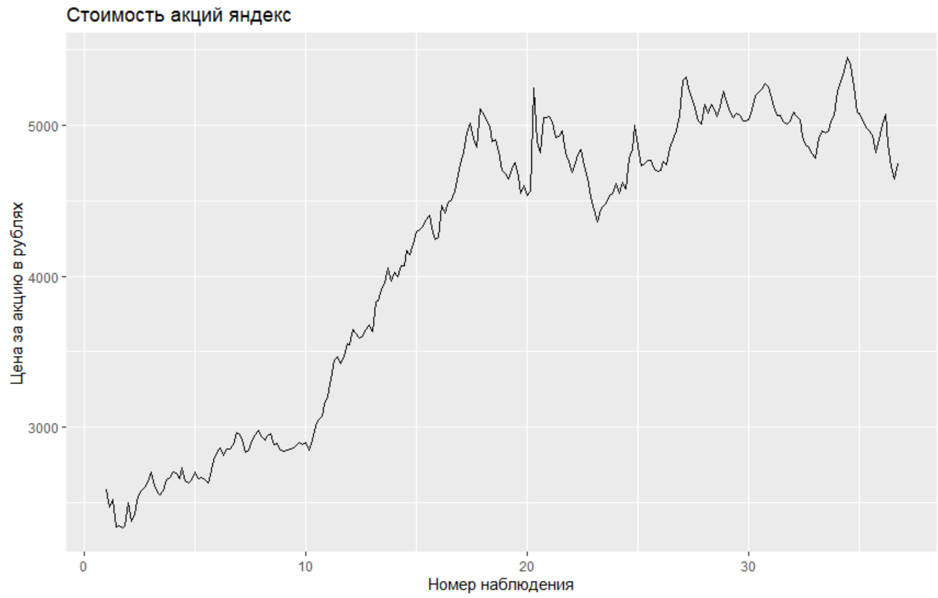

Рисунок 1 – График временного ряда стоимости акций Яндекс

Обратимся к коррелограммам ряда. Наше предположение о нестационарности подтверждается, так как значения автокорреляционной функции монотонно убывают (что опять же характерно для процесса случайного блуждания) и несмотря на большую выборку лагов, так и не входят в доверительный интервал на рисунке 2. Частная автокорреляционная функция значима на первом лаге и не значима на следующих, что характерно для нестационарных процессов (рисунок 3) .

Проверим стационарность тестом единичного корня Дики-Фуллера. Рассмотрим критические значения для рассматриваемого случая. Если статистика теста Дики- Фуллера меньше -2.58, то вывод о стационарности остатков можно сделать с 99 %-ным уровнем надежности, а если меньше -1.95, но больше -2.58, то с 95 %-ным уровнем надежности. Если интересующая нас статистика меньше -1.62, но больше -1.95, то уровень надежности вывода о

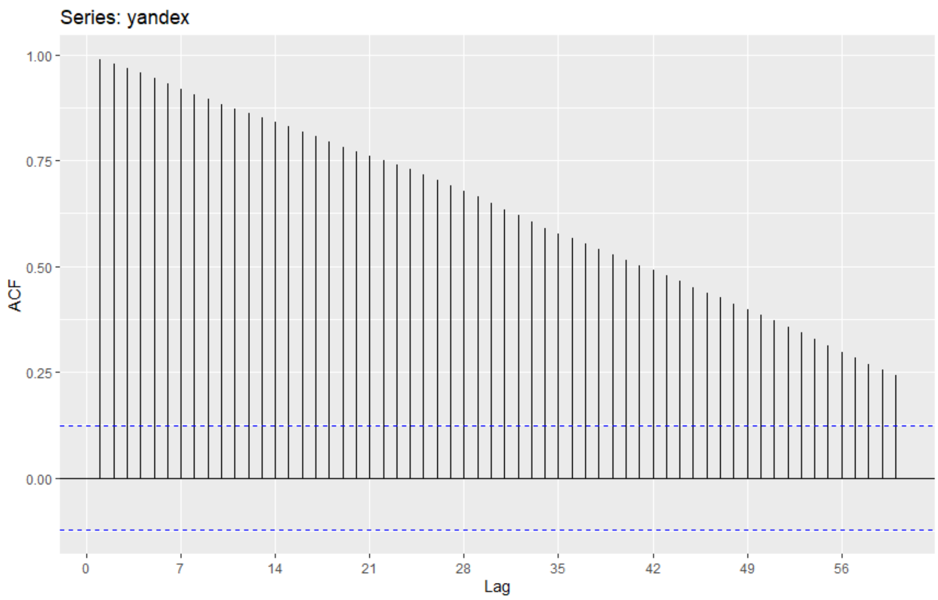

Рисунок 2 – График АКФ стоимости акций Яндекс

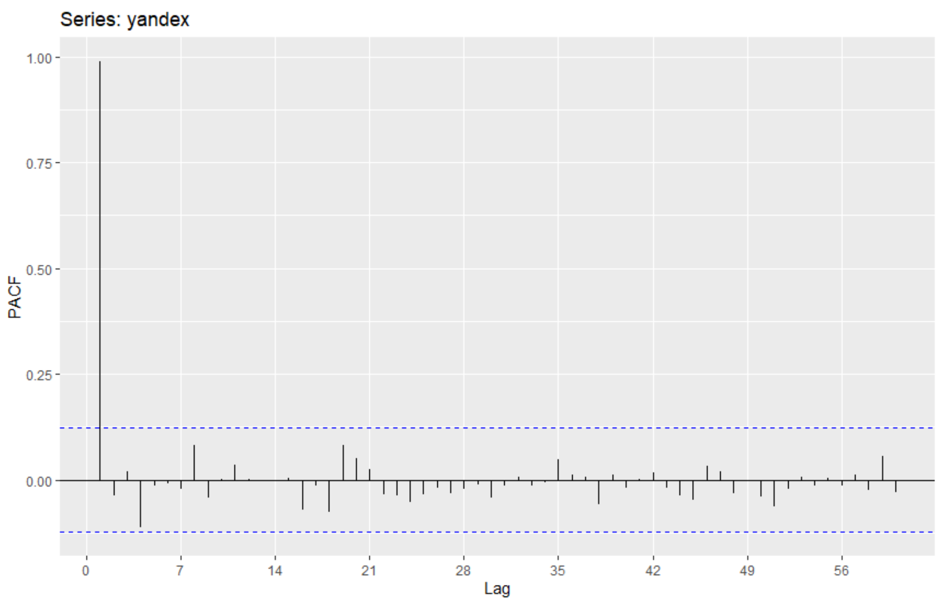

Рисунок 3 – График ЧАКФ стоимости акций Яндекс

стационарности остатков снижается до 90 %. Значение статистики 1.1828, что больше любого критического значения, следовательно, ряд не является стационарным.

Применим дифференцирование первого порядка для того, чтобы убедиться в нашем предположении о процессе случайного блуждания. Ряд примет вид на рисунке 4. Визуально ряд кажется стационарным, проверим это с помощью коррелограмм и тестов проверки на стационарность.

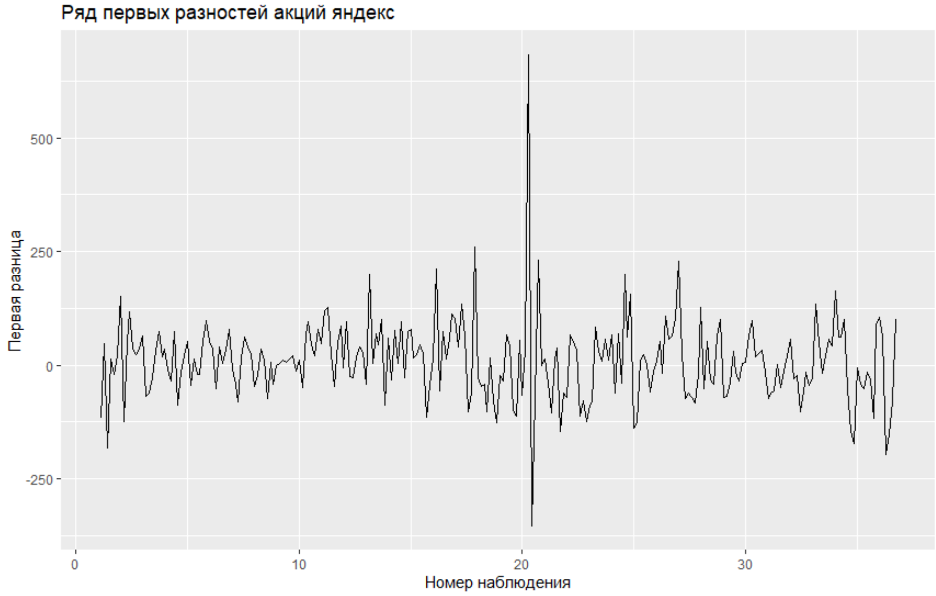

Рисунок 4 – График временного ряда первых разностей стоимости акций Яндекс

Коррелограммы ряда на рисунке 5 внешне похожи на коррелограммамы ряда случайного блуждания, так как среди первых значений практически все равны нулю, значит, наш процесс является процессом случайного блуждания.

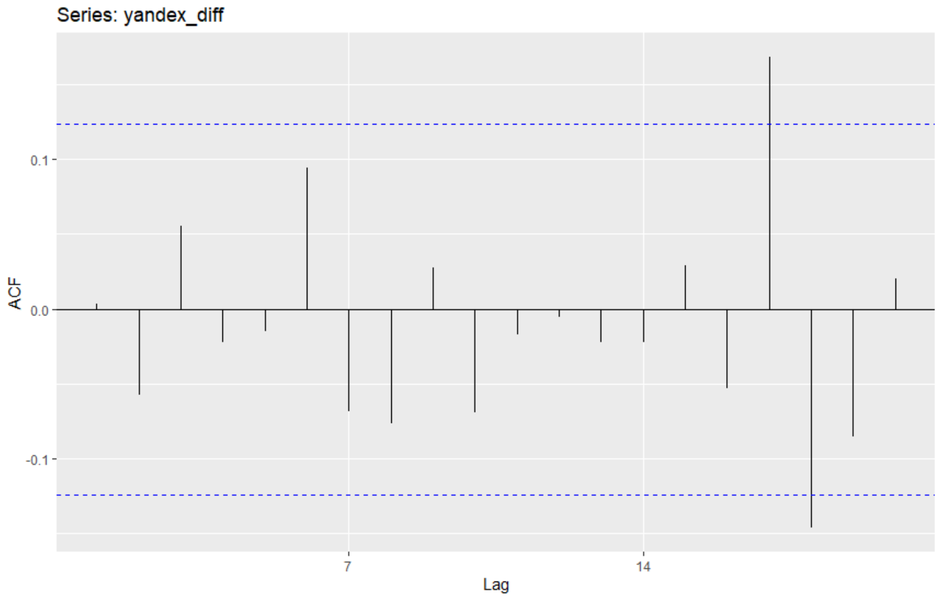

Рисунок 5 – График АКФ первых разностей стоимости акций Яндекс

На графике частной автокорреляционной функции на рисунке 6 не наблюдается значимость первых лагов, что означает, что ряд стал стационарным.

Применяя тест Дики-Фуллера к скорректированным данным получим

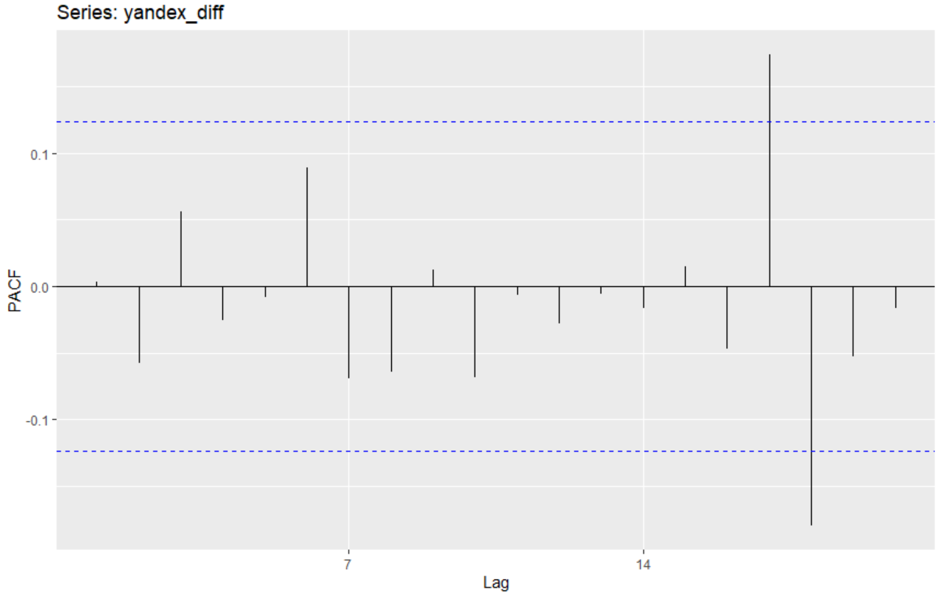

Рисунок 6 – График ЧАКФ первых разностей стоимости акций Яндекс

значение статистики равное -11.5395, что говорит о том, что ряд стал стационарным, так как это значение попадает в 99%-ный интервал.

Значение статистики критерия $KPSS$ стало равным 0.2449, что также входит в доверительные интервалы теста.

Применим простейшие методы прогнозирования (наивный метод, метод прогнозирования средним значением, наивный с трендом) для нашего ряда, результаты представлены на рисунке 7.

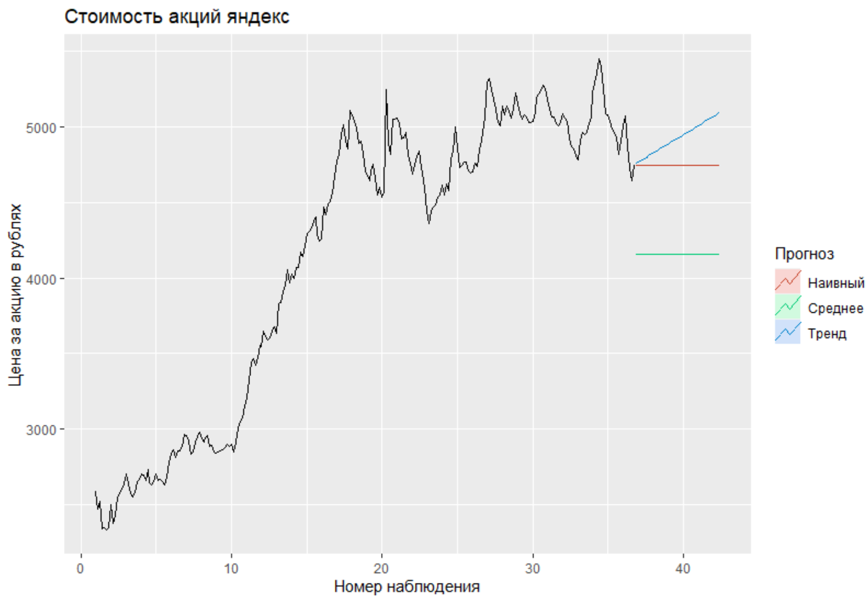

Рисунок 7 – Прогнозирование ряда

В таблице 1 представлены численные показатели качества прогноза. Чем меньше значения ошибок, тем лучше прогноз. Таким образом, наименьшими показателями обладает модель прогноза методом дрейфа (наивного с трендом). Это также визуализируется по графику.

Таблица 1 – Меры качества прогнозов

Название модели	RMSE	MAE	MPE	MAPE	MASE
Среднее значение	974.6837	877.5599	-6.976646	24.66411	4.524384
Наивный	89.67762	63.2096	0.2207514	1.542644	0.325886
Наивный с трендом	89.25935	62.40755	-0.001480629	1.51599	0.3217509

Попробуем рассмотреть несколько различных моделей процессов авторегрессии и скользящего среднего и по критериям выбрать наиболее точную. Рассмотрим модель случайного блуждания $ARIMA(0,1,0)$, которая по нашим предположениям должна наиболее точно прогнозировать ряд. Получим следующее регрессионное уравнение $y_t = y_{t-1} + \varepsilon_t$. Для сравнения рассмотрим модель $ARIMA(1,1,0)$, ее уравнение будет иметь вид

$$y_t = y_{t-1} + 0.0127 \cdot y_{t-1} + \varepsilon_t.$$

В языке R есть встроенная функция подбора модели. Для нашего случая автоматически подобрана модель $ARIMA(0,1,0)(1,0,0)_7$. С помощью лагового оператора уравнение модели $ARIMA(0,1,0)(1,0,0)_7$ запишется следующим образом

$$(1 - 0.0727 \cdot L^7)(1 - L)y_t = 8.7275 + \varepsilon_t.$$

В таблице 2 представлены значения критериев для каждой модели. Наиболее лучшей будет модель с наименьшими значениями критериев. Такой является модель белого шума $ARIMA(0,1,0)$, что подтверждает наши начальные предположения о прогнозе ряда.

Подытоживая расчеты, наилучшей моделью оказывается процесс случайного блуждания, даже автоматический подбор модели справляется хуже. Подсчитанным прогнозным значением является значение 4748.6. График ряда с прогнозированными значениями представлен на рисунке 8.

Таблица 2 – Итоговая таблица полученных результатов

Название модели	AIC	AIC_c	BIC
ARIMA(0,1,0)	2959.58	2959.6	2963.1
ARIMA(1,1,0)	2961.54	2961.59	2968.58
ARIMA(0,1,0)(1,0,0)[7]	2960	2960.09	2970.56

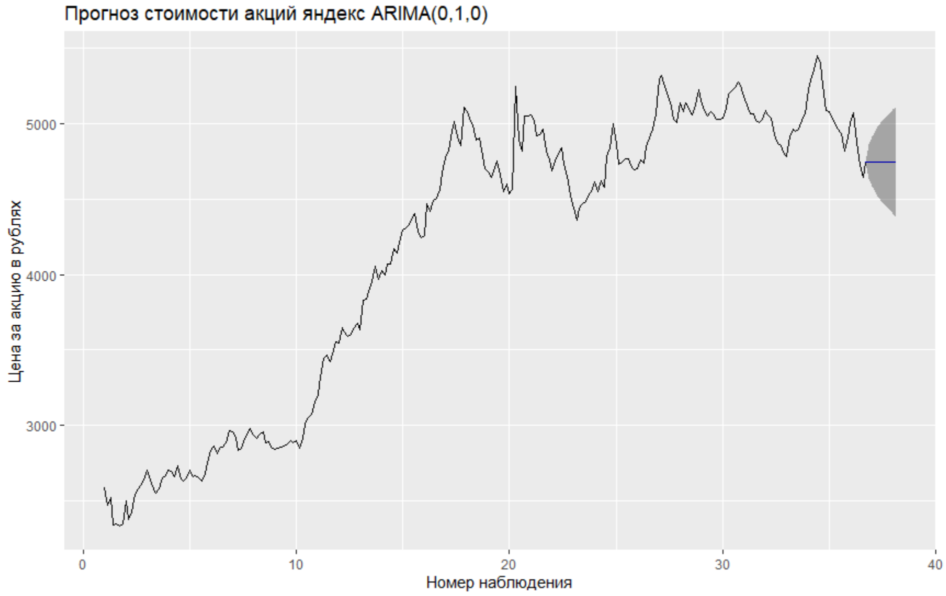

Рисунок 8 – Прогнозирование ряда моделью $ARIMA(0, 1, 0)$

Серым цветом на рисунке 8 показан 80-% доверительный интервал для прогнозных значений, в процессе случайного блуждания именно он представляет практическую ценность для дальнейшего прогнозирования, так как доверительный интервал дает нам информацию о возможном максимальном и минимальном значении прогнозируемой величины.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ

В первом разделе содержится общая характеристика временных рядов, основные определения и обозначения понятий теории временных рядов, вводятся определения автоковариационной и частной автоковариационной функций. Рассмотрены различные способы идентификации таких характеристик рядов, как тренд, сезонность, циклы. Данный раздел является вводным и служит для обеспечения систематизированного и замкнутого изложения основного материала.

Во втором разделе определяются авторегрессионные модели, модели скользящего среднего и линейного фильтра, описывается идентификация моделей и условия их стационарности. Описаны смешанные авторегрессионные модели.

В третьем разделе вводится понятие стационарности, условия стационарности и критерии проверки на стационарность. Понятие стационарности является важнейшей характеристикой временных рядов, так как без предположений о ряде невозможно его спрогнозировать. Также в данном разделе рассмотрены различные способы сведения рядов к стационарным, поскольку большинство реальных данных не имеет стационарной структуры.

В четвертом разделе рассмотрены некоторые методы прогнозирования, а также различные меры качества прогноза и критерии выбора наиболее точных моделей.

В пятом разделе рассмотрены реальные данные, интерпретированные в качестве временных рядов. Приводимые в работе данные имеют различную структуру и частоту наблюдений, что представляет ценность для демонстрации различных подходов к прогнозированию рядов. Поскольку построение моделей авторегрессии и скользящего среднего, получение прогнозных оценок исходного ряда возможно только с использованием программных средств, то для расчета была написана программа на языке R. Был проведен сравнительный анализ и оценка адекватности полученных моделей, выбраны оптимальные модели для каждого изучаемого ряда.