



МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«МИРЭА – Российский технологический университет»
РТУ МИРЭА

Институт информационных технологий (ИИТ)
Кафедра прикладной математики (ПМ)

КУРСОВАЯ РАБОТА

по дисциплине: «Языки программирования для статистической
обработки данных»

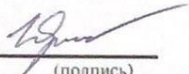
Тема курсовой работы: «Разработка программы для прогнозирования
временного ряда на основе линейных авторегрессионных моделей на
основе данных статусов авиаперевозок»

Студент группы ИМБО-02-22 Ким Кирилл Сергеевич


(подпись)

Руководитель
курсовой работы

старший преподаватель,
Юрченков И.А.


(подпись)

Работа представлена к защите «06» июня 2024 г.

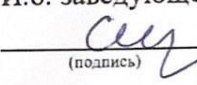
Допущен к защите «06» июня 2024 г.

Москва 2024 г.



МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«МИРЭА – Российский технологический университет»
РТУ МИРЭА

Институт информационных технологий (ИИТ)
Кафедра прикладной математики (ПМ)

Утверждаю
И.о. заведующего кафедрой ПМ
 Смоленцева Т.Е.
(подпись)
«09» февраля 2024 г.

ЗАДАНИЕ
на выполнение курсовой работы
по дисциплине «Языки программирования для статистической обработки
данных»

Студент Ким Кирилл Сергеевич

Группа ИМБО-02-22

Тема «Разработка программы для прогнозирования временного ряда на основе линейных авторегрессионных моделей на основе данных статусов авиаперевозок»

Исходные данные: выбранная студентом задача и алгоритм её решения, а также набор данных

Перечень вопросов, подлежащих разработке, и обязательного графического материала:

Описание решаемой задачи машинного обучения или статистической обработки данных
(математическая формулировка, проблематика, существующие способы решения)

Анализ выбранного алгоритма или метода решения выбранной задачи

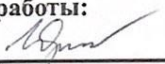
Выбор данных и описание набора данных, описание предикторов и целевых факторов

Построение сценария и логики обработки данных на основе выбранного алгоритма или метода

Оценка качества решения задачи на основе метрик качества

Срок представления к защите курсовой работы:

Задание на курсовую работу выдал


Подпись руководителя

до «24» мая 2024 г.

Юрченков И.А.
(ФИО руководителя)

«09» февраля 2024 г.

Ким К.С.

(ФИО обучающегося)

«09» февраля 2024 г.

Задание на курсовую работу
получил


Подпись обучающегося

СОДЕРЖАНИЕ

ВВЕДЕНИЕ	4
1 ТЕОРЕТИЧЕСКАЯ ЧАСТЬ.....	5
1.1 Основные понятия временных рядов	5
1.2 Линейная авторегрессионная модель.....	9
2 ПРАКТИЧЕСКАЯ ЧАСТЬ	13
2.1 Набор данных, анализ качества данных, предобработка данных, подготовка данных к моделированию	13
2.2 Линейная авторегрессионная модель на основе данных авиаперевозок для прогнозирования временного ряда.....	16
ЗАКЛЮЧЕНИЕ	19
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ	20
Теоретическая часть.....	20
Практическая часть	20
ПРИЛОЖЕНИЯ.....	21

ВВЕДЕНИЕ

В современном мире авиационная отрасль играет ключевую роль в международных и национальных перевозках. Эффективное управление авиаперевозками требует точного прогнозирования временных рядов статусов авиаперевозок, таких как задержки, отмены рейсов и другие изменения. Разработка программы для прогнозирования временного ряда на основе линейных авторегрессионных моделей позволит улучшить планирование и управление авиаперевозками.

Целью данного исследования является разработка программы для прогнозирования временного ряда на основе линейных авторегрессионных моделей на основе данных статусов авиаперевозок. Для достижения этой цели поставлены следующие задачи:

1. Изучить теоретические основы временных рядов и авторегрессионных моделей.
2. Провести анализ данных статусов авиаперевозок.
3. Разработать программу прогнозирования временного ряда на основе полученных данных.

После завершения данного исследования будет создана программа, способная предсказывать статусы авиаперевозок с высокой точностью, что поможет авиакомпаниям оптимизировать свою деятельность и повысить уровень обслуживания пассажиров.

1 ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

1.1 Основные понятия временных рядов

Для разработки программы прогнозирования временного ряда на основе данных статусов авиаперевозок необходимо разобраться с основными понятиями временных рядов. Временной ряд — это последовательность данных $Y(n)$, измеренных в последовательные моменты времени $t(n) = t_0 + n \cdot \Delta t$.

Каждое измерение представляет собой наблюдение для конкретной переменной в определенный момент времени.

Компоненты временного ряда:

1. Тренд (T — trend) — долгосрочное изменение уровня значений временного ряда. Тренд может быть восходящим, нисходящим или стационарным.
2. Сезонность (S — seasonal) — циклическое повторение паттернов в данных с постоянным интервалом времени. Например, продажи игрушек могут иметь сезонное изменение в преддверии праздников.
3. Цикл (C — cyclic) — периодические колебания в данных, обычно с более длительным циклом, чем у сезонности. Например, экономические циклы имеют периодичность в несколько лет.
4. Шум (E — errors) — непредсказуемая случайная переменная, которая не может быть объяснена трендом, сезонностью или циклом. Шум включает в себя случайные флуктуации и ошибки измерения.

Функция $F(T, S, C, E, n)$ описывает временной ряд $Y(n)$, как комбинацию его компонентов.

$$Y(n) = F(T, S, C, E, n),$$

Включение компонентов в модель временного ряда:

1. Аддитивная модель: $Y(n) = T(n) + S(n) + C(n) + E(n)$, Аддитивная модель используется, если амплитуда колебаний более-менее постоянная.
2. Мультипликативная модель: $Y(n) = T(n) * S(n) * C(n) * E(n)$, Мультипликативная — если амплитуда колебаний зависит от значения сезонной компоненты.

Выбор между аддитивной и мультипликативной моделями зависит от того, как компоненты взаимодействуют друг с другом в конкретном временном ряде.

Автокорреляция — это корреляционная зависимость значений временного ряда, которые сменяют друг друга. Появляется в том случае, когда соседствующие между собой значения взаимосвязаны.

Число периодов, по которым рассчитывается называется лагом.

Лаг — это количество моментов, по которым принято рассчитывать коэффициент автокорреляции. Лаговый оператор B сначала берет значение элемента временного ряда и уменьшает его на единицу времени. Если лаговый оператор используется снова, то значение сдвигается еще на несколько временных единиц. Расчет шагов лага происходит по следующей формуле:

$$\begin{aligned}By_t &= y_{t-1}, \\ B(By_t) &= B^2y_t = y_{t-2}, \\ B^py_t &= y_{t-p}.\end{aligned}$$

Обычно временные ряды описывают при помощи следующих критериев:

- математическое ожидание — это средний параметр произвольного размера, измерения которого стремятся к бесконечности;
- дисперсия — это случайная очередность параметров произвольного размера по отношению к математическому ожиданию;

- автокорреляционная функция — это очередность коэффициентов автокорреляции с лагами со случайными значениями не меньше единицы

Временные ряды как правило делят на стационарные и нестационарные.

Временной ряд называется стационарным

- в узком смысле, если для любых t_1, \dots, t_n, τ вектор $(y_{t_1+\tau}, \dots, y_{t_n+\tau})$ совпадает по распределению с $(y_{t_1}, \dots, y_{t_n})$, то есть при сдвиге всех моментов времени на одно и тоже число совместное распределение значений временного ряда в эти моменты времени не поменяется;
- в широком смысле, если
 - $E y_t^2 < +\infty$ для любого t ;
 - $E y_t$ не зависит от t , то есть в среднем значение временного ряда постоянно;
 - $cov(y_{t+\tau}, y_{s+\tau}) = cov(y_t, y_s)$ для любых t, s, τ , то есть значение автокорреляции зависит только от длины отрезка времени между двумя значениями;
- для гауссовских распределений, то есть для случая, когда все векторы вида $(y_{t_1}, \dots, y_{t_n})$ имеют нормальное распределение, определения эквивалентны. Это следует из того, что распределение гауссовского случайного вектора полностью определяется математическим ожиданием и ковариациями;

В нестационарных временных рядах статистические свойства меняются со временем. Они показывают сезонные эффекты, тренды и другие структуры, которые зависят от временного показателя.

У нестационарного ряда есть возможность превращения в стационарный. Для того, нужно воспроизвести следующий алгоритм:

- если у нестационарного временного ряда обнаружится возможность экспоненциального роста, то для него используют простое логарифмирование или логарифмирование цепных индексов:

$$y_t^* = \ln(y_t),$$

$$y_t^* = \ln\left(\frac{y_t}{y_{t-1}}\right) = \ln y_t - \ln y_{t-1}.$$

- следующим шагом является вычисление роста исследуемого временного ряда при помощи следующей функции:

$$y_t^* = \frac{y_t - y_{t-1}}{y_{t-1}} = \frac{y_t}{y_{t-1}} - 1.$$

Интегрирование для порядка d можно представить при помощи следующего уравнения:

$$\Delta^d y_t = \Delta^{d-1} y_t - \Delta^{d-1} y_{t-1}.$$

Для того, чтобы выполнить анализ временного ряда необходимы различные методы аналитики для выборки из него необходимых элементов.

При помощи этого теста Дикки-Фуллера проверяют является ли ряд стационарным или нет. Он проверяет ряд на наличие единичного корня в авторегрессии на один шаг назад. Если говорить конкретно, то проверяется значение коэффициента α в авторегрессионном уравнении первого порядка:

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

где y_t — является временным рядом;

ε_t — ошибка

В том случае, когда значение параметра α приравнивается к единице, процесс имеет единичный корень, а это обозначает что временной ряд не является стационарным.

Если $|\alpha| < 1$, то ряд стационарный. Тест Дикки-Фуллера рассчитывает p -статистику, в случае $p < 0.05$ гипотеза о стационарности ряда не отвергается.

Из-за его простоты тест работает не очень хорошо. Существует довольно много улучшенных тестов таких как:

- расширенный тест Дикки-Фуллера;
- Kwiatkowski–Phillips–Schmidt–Shin (KPSS).

Анализ временных рядов позволяет выявлять закономерности в данных, делать прогнозы и принимать более обоснованные решения на основе исторических данных. Методы анализа временных рядов включают в себя статистические и математические модели, такие как ARIMA (среднее, интегрированное, скользящее среднее), экспоненциальное сглаживание и регрессионный анализ. [1.1]

1.2 Линейная авторегрессионная модель

Авторегрессия (autoregressive model, AR) — это регрессия ряда на собственные значения в прошлом. Другими словами, наши признаки в модели обычной регрессии мы заменяем значениями той же переменной, но за предыдущие периоды.

Когда мы прогнозируем значение в период t с помощью данных за предыдущий период (AR(1)), уравнение будет выглядеть следующим образом.

$$y_t = c + \varphi_1 y_{t-1},$$

где c — это константа;

φ_1 — вес модели;

y_{t-1} — значение в период $t - 1$.

Количество используемых предыдущих периодов определяется параметром p . Обычно записывается как $AR(p)$.

Модель скользящего среднего (moving average, MA) помогает учесть случайные колебания или отклонения (ошибки) истинного значения от прогнозного. Можно также сказать, что модель скользящего среднего — это авторегрессия на ошибку.

Если использовать ошибку только предыдущего наблюдения, то уравнение будет выглядеть следующим образом.

$$y_t = \mu + \varphi_1 \varepsilon_{t-1},$$

где μ — это среднее значение временного ряда;

φ_1 — вес модели;

ε_{t-1} — ошибка в период $t - 1$.

Такую модель принято называть моделью скользящего среднего с параметром $q = 1$ или $MA(1)$. Разумеется, параметр q может принимать и другие значения ($MA(q)$).

ARMA предполагает, что в данных отсутствует тренд и сезонность (данные стационарны). Если данные нестационарны, нужно использовать более сложные версии этих моделей:

Модель $ARMA(p, q)$ по сути является суммой моделей $AR(p)$ и $MA(q)$, иначе говоря, модель есть сумма нескольких предыдущих значений ряда и нескольких предыдущих значений белого шума с некоторыми коэффициентами.

$$y_t = \alpha + \varphi_1 y_{t-1} + \dots + \varphi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_q \varepsilon_{t-q}$$

Эквивалентную запись ряда в терминах оператора сдвига можно получить, рассмотрев два многочлена.

$$a(L)y_t = \alpha + b(L)\varepsilon_t.$$

или

$$y_t = \mu + \frac{b(L)}{a(L)}\varepsilon_t.$$

где $a(z) = 1 - \varphi_1 z - \dots - \varphi_p z^p$;

$$b(z) = 1 + \theta_1 z + \dots + \theta_p z^p.$$

Заметим, что во втором представлении константа α заменена на $\mu = Ey_t$. На самом деле, стационарность такого ряда будет определяться только его $AR(p)$ компонентой, то есть значениями коэффициентов, так ряд в модели $MA(q)$ всегда является стационарным.

- ARIMA, здесь добавляется компонент Integrated (I), который отвечает за удаление тренда (сам процесс называется дифференцированием);
- SARIMA, эта модель учитывает сезонность (Seasonality, S);
- SARIMAX включает еще и внешние или экзогенные факторы (eXogenous factors, отсюда и буква X в названии), которые напрямую не учитываются моделью, но влияют на нее.

Параметров у модели SARIMAX больше. Их полная версия выглядит как $SARIMAX(p, d, q) \times (P, D, Q, s)$. В данном случае, помимо известных параметров p и q , у нас появляется параметр d , отвечающий за тренд, а также набор параметров (P, D, Q, s) , отвечающих за сезонность.

Теперь давайте воспользуемся моделью SARIMAX для прогнозирования авиаперевозок. [1.2]

Сэмплирование для авторегрессии производится скользящим окном с шириной авторегрессионной зависимости.

Следующим этапом является оценивания точности прогноза в одной модели одновременно используют разные метрики с индивидуальными свойствами.

В качестве примера возьмем идентифицированную модель временных рядов, в которой уже построен прогноз. Получим вектор ошибок и представим его в виде разницы фактических и расчетных данных:

$$e = y - \hat{y}.$$

MAE (MAD) — среднее абсолютное отклонение. Также, как и MFE отображает среднее абсолютное отклонение действительных данных от прогнозируемых. Единственным отличием от средней ошибки прогноза (MFE) ошибки с разными значениями не сокращают друг друга. Если значение метрики стремится к нулю, то прогноз будет более точным.

$$MAE = \frac{1}{n} \sum_{i=1}^n |e_i|.$$

MAPE — средняя абсолютная ошибка в процентах. Демонстрирует процент отклонения действительных значений от прогнозируемых, но в этот вариант можно применять только для рядов, со средним значением больше единицы. Точность прогноза данной метрики зависит от её минимального значения. [1.3]

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{e_i}{y_i} \right| * 100\%.$$

Рассмотрим применение линейную авторегрессионную модель на основе данных статусов авиаперевозок для прогнозирования временного ряда.

2 ПРАКТИЧЕСКАЯ ЧАСТЬ

2.1 Набор данных, анализ качества данных, предобработка данных, подготовка данных к моделированию

Для выполнения практической работы были использованы открытые наборы данных о статусах авиаперевозок.

Данные были взяты статистические данные с января 2018 года по декабрь 2020 года.

В наборе данных включает в себя следующие столбцы данных:

1. FlightDate — дата вылета (ггггммдд);
2. Airline — авиакомпания;
3. Cancelled — индикатор отмены рейса (1=Да);
4. DepDelayMinutes — разница в минутах между запланированным и фактическим временем отправления. Для ранних отправок установлено значение 0 и др.

Исходные данные были соединены в одну таблицу и для оптимизации работы с ними на этапах обработки, были исключены столбцы, которые не будут использоваться.

Полученный набор представлен на Рисунок 2.1

FlightDate	Airline	Origin	Dest	Cancelled	Diverted	CRSDepTime	DepTime	DepDelayMinutes	DepDelay	ArrTime	ArrDelayMinutes
2018-01-01	Endeavor Air Inc.	ATL	MOB	False	False	940	935	0	-5	949	0
2018-01-01	Endeavor Air Inc.	MOB	ATL	False	False	1035	1028	0	-7	1304	6
2018-01-01	Endeavor Air Inc.	ATL	OAJ	False	False	2215	2214	0	-1	2330	0
2018-01-01	Endeavor Air Inc.	ATL	MGM	False	False	1503	1457	0	-6	1447	0
2018-01-01	Endeavor Air Inc.	MGM	ATL	False	False	1528	1520	0	-8	1708	0
2018-01-01	Endeavor Air Inc.	DCA	JFK	False	False	900	949	49	49	1059	36

Рисунок 2.1 — Табличное представление данных

Типы данных и общее количество объектов представлены на Рисунок 2.2, общее количество объектов подтверждает тот, факт, что для более быстрой работы – нужно было исключить лишние столбцы.

data1	5689512 obs. of 61 variables
\$ FlightDate	: chr "2018-01-23" "2018-01-24" "2018-01-25" "2018-01-26" ...
\$ Airline	: chr "Endeavor Air Inc." "Endeavor Air Inc." "Endeavor Air Inc." "Endeavor Air Inc."
\$ Origin	: chr "ABY" "ABY" "ABY" "ABY" ...
\$ Dest	: chr "ATL" "ATL" "ATL" "ATL" ...
\$ Cancelled	: chr "False" "False" "False" "False" ...
\$ Diverted	: chr "False" "False" "False" "False" ...
\$ CRSDepTime	: int 1202 1202 1202 1202 1400 1202 1202 1202 1037 ...
\$ DepTime	: num 1157 1157 1153 1150 1355 ...
\$ DepDelayMinutes	: num 0 0 0 0 0 NA 2 0 0 24 ...

Рисунок 2.2 — Типы данных и общее количество

Для выполнения обработки временного ряда на языке программирования R мы будем использовать библиотеку `forecast`, для построения графиков библиотеки `ggplot2` и `dplyr`.

Убирали максимум и минимум значения из столбца `total_delay` (переименованный `DepDelayMinutes`) показано на Рисунке 2.3.

	FlightDate	total_delay
	<chr>	<dbl>
1	2018-01-01	49
2	2018-01-01	141
3	2018-01-01	44
4	2018-01-01	55
5	2018-01-01	56
6	2018-01-01	118
7	2018-01-01	105
8	2018-01-01	13
9	2018-01-01	17
10	2018-01-01	19

Рисунок 2.3 — Набор данных без максимум и минимум

Графики временных трендов показаны на Рисунке 2.4. В верхней части холста показана временная диаграмма. В нижней показаны значения автокорреляционной функции для различных лагов. Значения частных автокорреляций также для различных лагов.

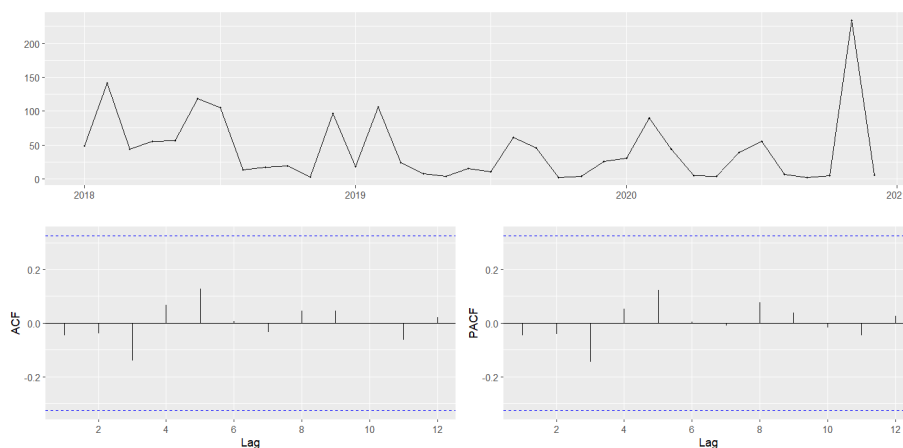


Рисунок 2.4 — Отображение данных

По виду Автокорреляционной функции заметно, что данный процесс является стационарным, можно увидеть по тесту Дикки-Фуллера:

- Dickey-Fuller = -6.0803;
- Truncation lag parameter = 3;
- p-value = 0.01.

Поскольку мы наблюдаем зашумление с трендовыми зависимостями, значения автокорреляций при лагах, в будущем при оценке модели авторегрессии, будем учитывать наличие тренда в модели, а также обращать внимания на частные автокорреляции.

Создаем временной ряд и декомпозируем. На Рисунке 2.5. показан результат создания временного ряда. [2.1]

```
> flight_delays_ts
      Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
2018   49 141  44  55  56 118 105  13  17  19   3  96
2019   18 106  24   8   4  15  11  61  46   2   4  26
2020   31  90  44   5   4  39  55   7   2   5 234   6
```

Рисунок 2.5 — Результат декомпозиции временного ряда

Декомпозиция временного ряда на составляющие: сам ряд, тренд, сезонные колебания, случайные колебания. Все эти характеристики были нами получены исходя из априорного понимания о годовой цикличности. Декомпозиция временного ряда представлена на Рисунке 2.6.

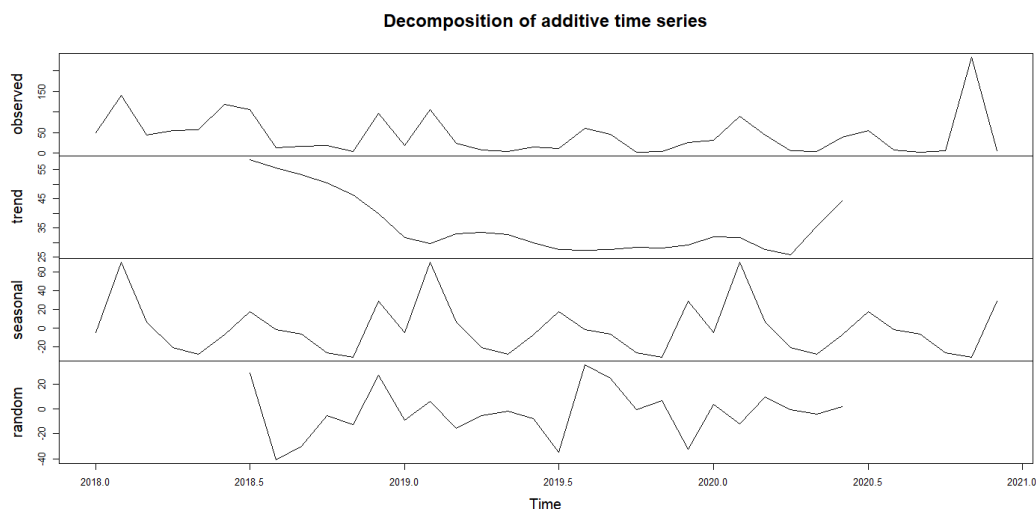


Рисунок 2.6 — Декомпозиция временного ряда

Построим график ряда и тренда друг на друге показано на Рисунке 2.7.



Рисунок 2.7 — Обзор характеристик временного ряда

Были рассмотрены данные временного ряда, тренд достаточно гладкий, видна большая цикличность, нам придется учитывать в прогнозировании ряда.

2.2 Линейная авторегрессионная модель на основе данных авиаперевозок для прогнозирования временного ряда

Для построения модели АР необходимо сначала определить порядок модели, p . Это можно сделать с помощью различных методов, таких как критерий Акаике. Модель "arima5.model" получаем такие коэффициенты (Таблица 2.1):

Таблица 2.1 — Коэффициенты полученной модели

	ar1	ma1
	-0.6451	-1.0000
s.e.	0.1648	0.0979

- среднеквадратической ошибки (σ^2) оценено как 3614;

- логарифмическое правдоподобие (log likelihood) составляет -190.05;
- AIC (критерий Акаике), которое равно 386.1.

Видим, что метрики нехорошие, так как MAPE=334.8823 %. Модель достаточно плохая и получила такие меры по устранению ошибок в обучающем наборе (Таблица 2.2):

Таблица 2.2 — Меры по устранению ошибок

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Набор	3.50843	58.42505	41.05397	-262.0789	334.8823	0.8908177	-0.06795259

Дальше прогнозируем будущие значения, прогнозируемые значения лучше всего увидеть на Рисунке 2.8.

На графике показаны исходные данные о авиаперевозках и прогноз на следующие 12 месяцев.

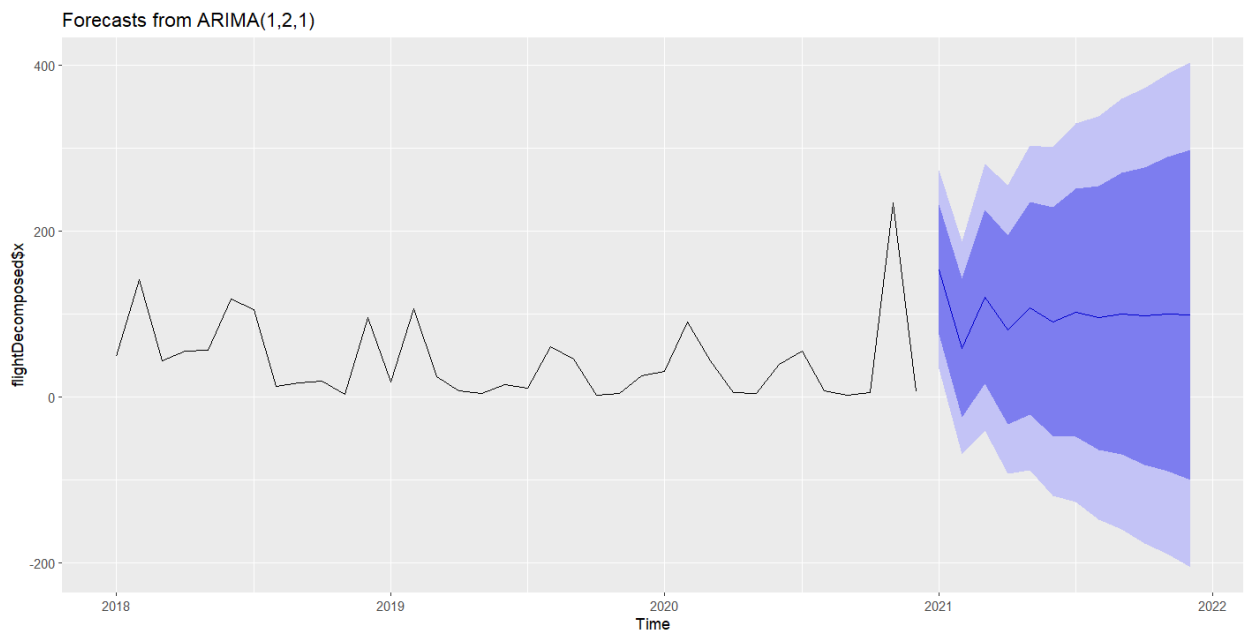


Рисунок 2.8 — ARIMA модель прогноза

Дальше оцениваем на сколько близко прогноз находится к последним данным на Рисунке 2.9.

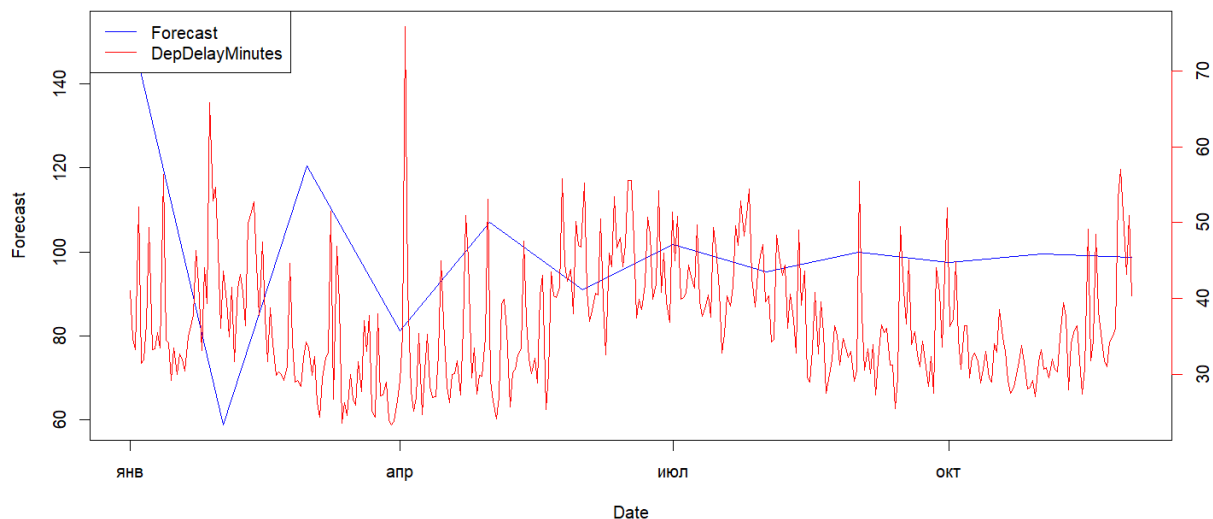


Рисунок 2.9 — Сравнение

Для международных перевозок модель не показывает вразумительных результатов.

Эта модель способна делать прогнозы, однако она не справляется с объемом данных из-за своей относительной слабости. Для достижения более точных прогнозов необходимо внести коррективы и улучшить модель.

ЗАКЛЮЧЕНИЕ

В результате работы была разработана программа для прогнозирования временного ряда на основе линейных авторегрессионных моделей на данных статусов авиаперевозок. Проведен анализ качества прогнозов, была создана модель, описывающая временный ряд.

Модели класса ARIMA хорошо показывают себя при моделировании временных рядов с выраженной внутренней структурой, но плохо адаптируются к внезапным изменениям тренда.

Статистическое моделирование использует накопленные знания об объекте наблюдения, чтобы воспроизвести и предсказать его дальнейшее поведение. В ситуациях, не имеющих аналогов, моделирование должно строиться на экспертных оценках, что особенно актуально для международных перевозок. Степень неуверенности модели отражается в широкой области доверительного интервала, и эта степень тем ниже, чем меньше влияние внешних переменных и чем больше внутренняя структурированность ряда.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

Теоретическая часть

- 1.1. Волков Никита. Аналитика временных рядов / Учебник по машинному обучению [Электронный ресурс].
<https://education.yandex.ru/handbook/ml/article/analitika-vremennyh-ryadov>
- 1.2. Волков Никита. Аналитика временных рядов / Учебник по машинному обучению [Электронный ресурс].
<https://education.yandex.ru/handbook/ml/article/modeli-vida-arima>
- 1.3. Губко Павел. Метрики классификации и регрессии [Электронный ресурс].
<https://education.yandex.ru/handbook/ml/article/metriki-klassifikacii-i-regressii>

Практическая часть

- 2.1 Эйлин Нильсен. Практический анализ временных рядов. Прогнозирование со статической и машинное обучение
- 2.2 Венэбльз У.Н., Смит Д. М. и Рабочая группа разработки R. Введение в R.
- 2.3 Заметки по R: среда программирования для анализа данных и графики. Москва. 2012. 109 с.
- 2.4 Золотарюк А.В. Язык и среда программирования R. Учебное пособие. ИНФРАМ, 2019, 162 с.

ПРИЛОЖЕНИЯ

Приложение А — Листинги кода программы.

Приложение А

Листинг А — Код приложения

```
# Установка и загрузка необходимых библиотек
install.packages("forecast")
install.packages("dplyr")
install.packages("ggplot2")
install.packages("seasonalview")
install.packages("tidyverse")
install.packages("gridExtra")
install.packages("outliers")
install.packages("readr")

library(outliers)
library(stats)
library(quantmod)
library(forecast)
library(ggplot2)
library(dplyr)
library(readr)
library(seasonalview)
library(tidyverse)
library(gridExtra)

# Загрузка данных
data1 <- read.csv('D:\\practice1\\archive\\Combined_Flights_2018.csv')
data2 <- read.csv('D:\\practice1\\archive\\Combined_Flights_2019.csv')
data3 <- read.csv('D:\\practice1\\archive\\Combined_Flights_2020.csv')

# Выбрали столбцы FlightDate и DepDelayMinutes
data11 <- select(data1, FlightDate, DepDelayMinutes)
data22 <- select(data2, FlightDate, DepDelayMinutes)
data33 <- select(data3, FlightDate, DepDelayMinutes)

# Объединили таблицы
data111 <- full_join(data11, data22, by = "FlightDate")
data222 <- full_join(data111, data33, by = "FlightDate")
data222 <- data222[1:2]
data2222 <- na.omit(data222)

flight_delays <- data2222 %>%
  group_by(FlightDate) %>%
```


Продолжение листинга А — Код приложения

```
summarise(total_delay = DepDelayMinutes.x)

flight_delays1 <- na.omit(flight_delays)

# Нашли мин макс и исключили
min_value <- min(flight_delays1$total_delay)
max_value <- max(flight_delays1$total_delay)

flight_delays12345 <- subset(flight_delays1, total_delay > min_value &
total_delay < max_value)
print(flight_delays12345)

# Временной ряд
flight_delays_ts <- ts(flight_delays12345$total_delay, start = c(2018, 1),
end = c(2020, 12), frequency = 12)

(acf(flight_delays_ts, main=""))

plot(stl(flight_delays_ts, s.window="periodic")$time.series, main="")

ggtsdisplay(flight_delays_ts)
ggtsdisplay(diff(flight_delays_ts))
mean(diff(flight_delays_ts))
ggtsdisplay(diff(diff(flight_delays_ts, 12)))
ggtsdisplay(diff(diff(flight_delays_ts)))
mean(diff(diff(flight_delays_ts)))
plot(decompose(flight_delays_ts))
flightDecomposed <- decompose(flight_delays_ts)
plot(flightDecomposed$x,
      main = "Обзор характеристик временного ряда",
      xlab = "Время наблюдения",
      ylab = "Значения")
lines(flightDecomposed$trend, col = "red")

PP.test(flightDecomposed$x)

# ARIMA модель прогноза -----
fit <- auto.arima(flightDecomposed$x)
summary(fit)
armaorder(fit)
arima1.model <- auto.arima(flightDecomposed$x)
arima2.model <- arima(flightDecomposed$x, order = c(2,3,1))
```

Продолжение листинга А — Код приложения

```
arima3.model <- arima(flightDecomposed$x, order = c(3,0,1))
arima4.model <- arima(flightDecomposed$x, order = c(1,4,3))
arima5.model <- arima(flightDecomposed$x, order = c(1,2,1))

AIC(arima1.model, arima2.model, arima3.model, arima4.model, arima5.model)

summary(arima5.model)
arimaorder(arima5.model)

future1 <- forecast(arima5.model, h = 12)
print(future1)
autoplot(future1)

# ARIMA для тренда -----
arima6.model <- auto.arima(flightDecomposed$trend)
future5 <- forecast(arima5.model, h = 12)

plot(future5)
print(future5)
str(future5)
summary(future5)

round(predict(arima5.model,
              n.ahead=12,
              se.fit=TRUE)$se) +
  predict(arima5.model,
          n.ahead=12,
          se.fit=TRUE)$pred

round(-predict(arima5.model,
               n.ahead=12,
               se.fit=TRUE)$se) +
  predict(arima5.model,
          n.ahead=12,
          se.fit=TRUE)$pred

#Оценка-----
data4 <- read.csv('D:\\practice1\\archive\\Combined_Flights_2021.csv')
f44 <- select(data4, FlightDate, DepDelayMinutes)
f444 <- na.omit(f44)
f444
```

Продолжение листинга А — Код приложения

```
min_value1 <- min(f444$DepDelayMinutes)
max_value1 <- max(f444$DepDelayMinutes)
f33312345 <- subset(f444, DepDelayMinutes > min_value1 & DepDelayMinutes <
max_value1)
print(f33312345)

f33312345 <- f33312345 %>%
  group_by(FlightDate) %>%
  summarise(DepDelayMinutes = mean(DepDelayMinutes))

forecast_df <- data.frame(forecast=future1$mean)

f33312345$FlightDate <- as.Date(f33312345$FlightDate)

flight_date <- as.Date(c("2021-01-01", "2021-02-01", "2021-03-01",
                        "2021-04-01", "2021-05-01", "2021-06-01",
                        "2021-07-01", "2021-08-01", "2021-09-01",
                        "2021-10-01", "2021-11-01", "2021-12-01"))

fdata <- data.frame(flight_date, forecast_df)

# Создание первого графика
plot(fdata$flight_date, fdata$forecast, type = "l", col = "blue", xlab =
"Date", ylab = "Forecast")

# Установка параметров для второго графика
par(new = TRUE)

# Создание второго графика
plot(f33312345$FlightDate, f33312345$DepDelayMinutes, type = "l", col =
"red", xlab = "", ylab = "", axes = FALSE)

# Добавление осей координат для второго графика
axis(side = 4, col = "red")

# Добавление легенды
legend("topleft", legend = c("Forecast", "DepDelayMinutes"), col =
c("blue", "red"), lty = 1)

library(ggplot2)
```

Продолжение листинга А — Код приложения

```
# Создание графиков с ggplot2
ggplot() +
  geom_line(data = fdata, aes(x = flight_date, y = forecast), color =
"blue") +
  geom_line(data = f33312345, aes(x = FlightDate, y = DepDelayMinutes),
color = "red") +
  labs(x = "Date", y = "Value") +
  scale_color_manual(values = c("blue", "red"))
```