Исследование и предсказание выполнненого пассажирооборота на внутренних рейсах в Российской Федерации

Даниил Лёгенький
Механико-математический факультет,
МГУ им.Ломоносова
г.Москва, Россия

Егор Дранов
ФПМИ,
МФТИ
г. Долгопрудный, Россия
dranov.em@phystech.edu

Денис Пискун
Механико-математический факультет,
МГУ им.Ломоносова
г.Москва Россия
denis.piskun.01@mail.ru

Дмитрий Литовка
Механико-математический факультет,
МГУ им.Ломоносова
г.Москва Россия
dmitriy.litovkaa@gmail.com

Аннотация—Рассматриваются данные выполненного пассажирооборота внутренних авиационных рейсов. Исследуются его свойства, строится несколько моделей и сравниваются их метрики. Наилучшей моделью, предсказывающей ряд, является модель SARIMA$(0, 1, 3)(4, 1, 0)_{12}$.

Index Terms—пассажирооборот, выполненный пассажирооборот, SARIMA

I. Введение и описание ряда

В данной работе исследуются данннные выполненного пассажирооборота на внутренних авиационных рейсах в РФ. Данные были получены с сайта государственной статистики ЕМИСС.

Выполненный пассажирооборот - сумма произведений от умножения числа перевезенных пассажиров на каждом этапе полета на протяженность этапа, соответственно по каждому виду сообщений.

Данные по показателю сводятся из представляемых авиапредприятиями данных по форме федерального государственного статистического наблюдения 12-ГА "Сведения о перевозках пассажиров и грузов".

Периодичность сбора данных является месячной, единица измерения значений представлены в тысячах пассажиро-километров. Рассматривается ряд в период с 2009-01-31 до 2020-01-01.

На рис.1 представлен график временного ряда, как можно видеть на нем присутсвует сезонность. Действительно, летние периоды, как правило, характеризуются большим числом авиарейсов из-за отпусков. Данный факт явно виден на приведенном графике.

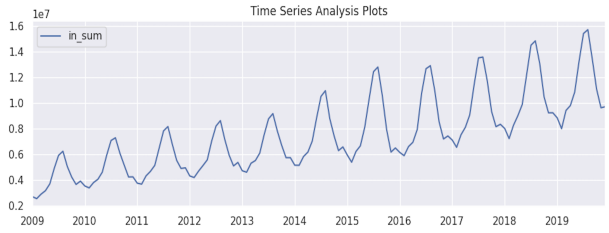

Рис. 1. Выполненный пассажирооборот, внутренние рейсы.

II. Преобразование ряда

A. Проверка на стационарность

На рис.1 можно наблюдать ежегодный рост, что говорит о нестационарности ряда. Проверим нестационарность ряда с помощью теста Дики-Фуллера. Тест выдает значение P-Value = 0.992, значит нулевая гипотеза о нестационарности ряда не отвергается.

Так же рассмотрим графики автокорреляционной и частитчно автокорреляционной функций для временного ряда (рис.2 и рис.3). Можно наблюдать, что ACF представляет из себя синусоидальную функцию с периодом примерно 12 месяцев.

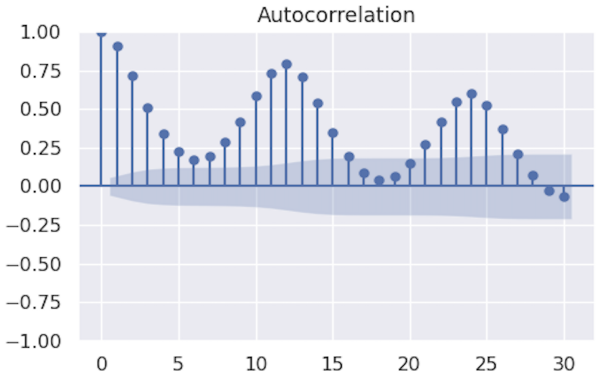

Рис. 2. ACF ряда на рис.1.

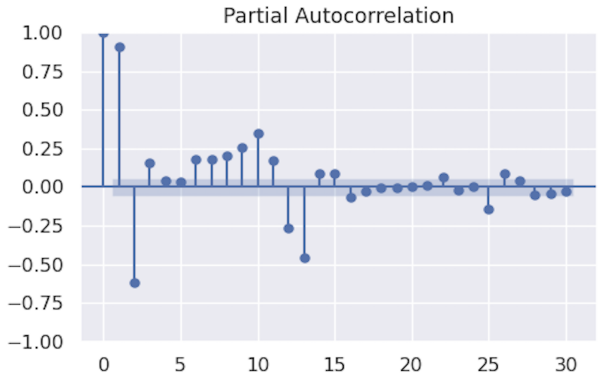

Рис. 3. PACF ряда на рис.1.

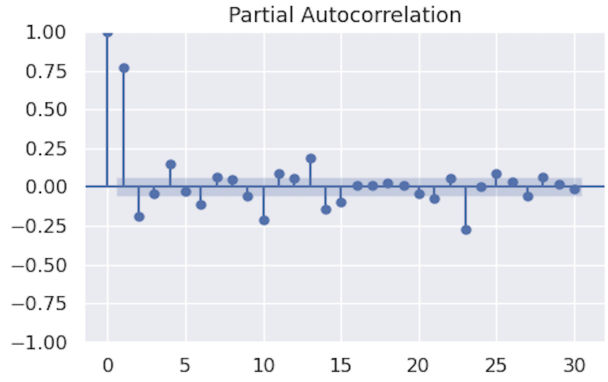

Рис. 6. PACF ряда на рис.4.

B. Построение стационарного ряда

Для того, чтобы посроить модель прогноза необходимо поучить стационарный ряд. Критерием стационарности в данной работе является тест Дики-Фуллера, для его выполнения необходимо отвергнуть нулевую гипотезу о нестационарности ряда. Будем преобразовывать ряд, для выполнения теста.

Для начала сделаем преобразование Бокса-Кокса. Как правило это используется для стабилизации дисперсии. Оптимальным параметром преобразования $\lambda = 0.113$. При этом P-Value в тесте Дики-Фуллера для нового ряда стал равен 0.635, что по прежнему не позволяет нам отвергнуть нулевую гипотезу. Далее уберем в ряде сезонность.

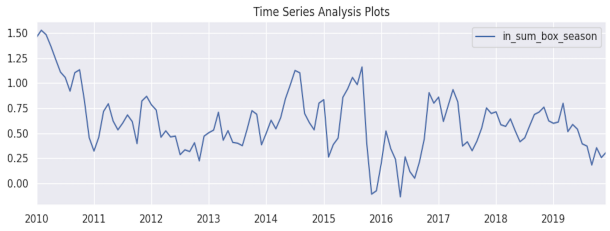

Рис. 4. Ряд без сезонности.

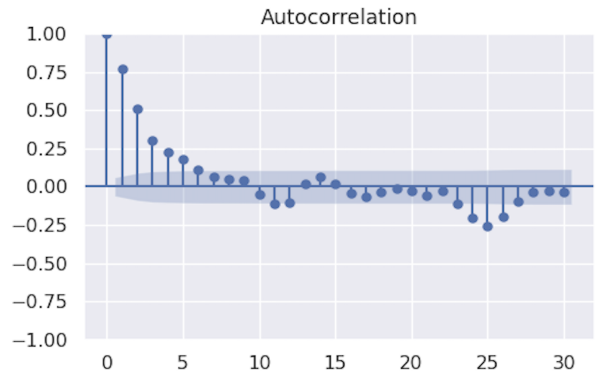

Рис. 5. ACF ряда на рис.4.

Ряд стал намного больше походил на стационарный, однако все равно можно наблюдать некоторое небольшое смещение среднего значения. Для устранения этого эффекта рассчитаем первую разность.

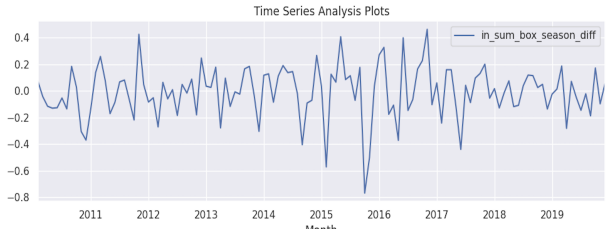

Рис. 7. Первая разность ряда.

График стал намного больше походж на стационарный. В отличии от рис.4, на рис.7 значение среднего значения со временем не меняется. Для более точной проверки проведем тест Дики-Фуллера. Здесь значение P-Value = 0.001, что меньше заданной границы, а значит нулевая гипотеза о нестационарноти отвергается.

Далее вся работа по построению модели и по предсказанию будет строится именно для ряда, который изображен на рис.7. Наблюдая за ACF и PACF уже можно говорить о некоторых параметрах будущей модели.

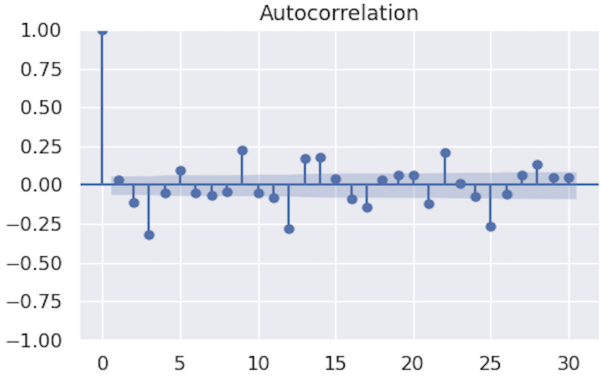

Рис. 8. ACF ряда на рис.7.

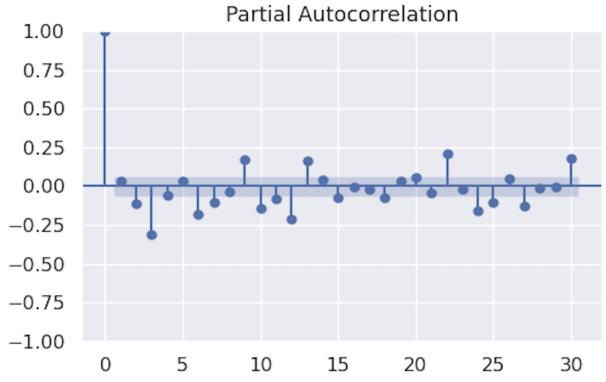

Рис. 9. PACF ряда на рис.7.

III. Построение модели

Поскольку в модели явно присутсвует сезонность, то для лучшего прогноза будем использовать модель SARIMA. Для подбора оптимальных параметров воспользуемся возможностями Python, который перебирает параметры на сетке и выбирает их по величине AIC.

Наилучшей моделью стала SARIMA$(0, 1, 3)(4, 1, 0)_{12}$, со значением AIC = -69.107. Параметры принимают в ней следующие значения.

	coef
ma.L1	−0.0647
ma.L2	−0.0932
ma.L3	−0.3187
ar.S.L12	−0.3981
ar.S.L24	−0.3159
ar.S.L36	−0.2536
ar.S.L48	−0.2581
sigma2	0.0271

Рис. 10. Parametrs for auto SARIMA.

Построим для данной модели график предсказания. Сравним полученный результат с наивным прогнозом, а так же с реальными данными.

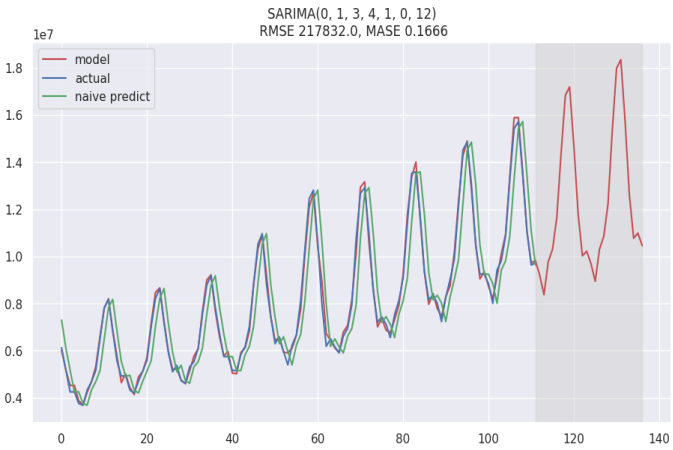

Рис. 11. SARIMA$(0, 1, 3)(4, 1, 0)_{12}$.

Как можно видеть, модель достаточно хорошо превосходит наивный прогноз, что видно по метрике MASE = 0.1666. Данная модель была построена по тому, какое у нее значение AIC. Если поменять параметры, то модель сможет улучшить метрику MASE. Например, SARIMA$(3, 0, 3)(4, 1, 0)_{12}$.

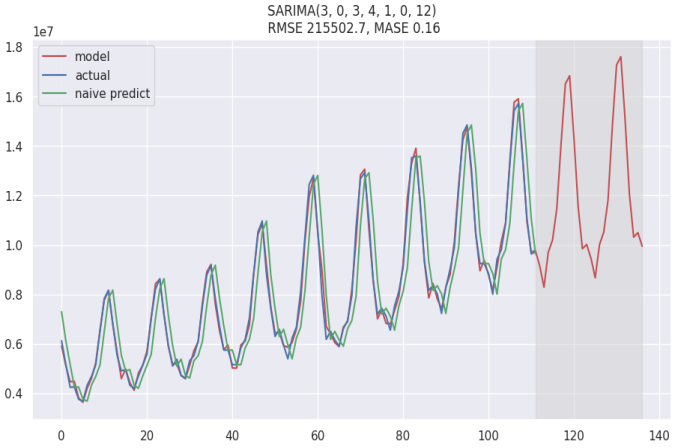

Рис. 12. SARIMA$(3, 0, 3)(4, 1, 0)_{12}$.

Здесь MASE = 0.16, что говорит о том, что текущая модель лучше 'бьет' наивный прогноз, чем auto-SARIMA.

IV. Вывод

Проведен анализ временного ряда выполнненого пассажирооборота на внутренних рейсах в РФ. Изначальный ряд не был стационарным, в силу своей сезонности. Сезонность объясняется тем, что во время летних отпусков число перелетов пассажиров растет, так как большинство выезжают на отдых именно в это время. Так же происходит рост среднего значения выполнненого пассажирооборота. Основным объяснением может служить рост популярности авиаперевозок, развитие авиационной инфраструктуры в РФ, строительство аэропортов, доступность авиабилетов для более широкой категории граждан.

Удалось получить стационарный ряд из нестационарного начального. Поскольку ряд являлся сезонным, то рассматривалась модель, которая ее учитывает, а именно SARIMA. Auto-SARIMA дала неплохие результаты, однако изменение параметров позволило улучшить метрику MASE, для SARIMA$(3, 0, 3)(4, 1, 0)_{12}$ она равна 0.16.

Таким образом, в работе была получена модель, которую можно использовать для предсказания значений выполнненого пассажирооборота на внутренних рейсах.