ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 70 за 2022 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

**В.А. Судаков, М.А. Тимофеев**

# Прогнозирование авиационных перевозок методами статистического анализа и машинного обучения



***Рекомендуемая форма библиографической ссылки:*** Судаков В.А., Тимофеев М.А. Прогнозирование авиационных перевозок методами статистического анализа и машинного обучения // Препринты ИПМ им. М.В.Келдыша. 2022. № 70. 14 с.
https://doi.org/10.20948/prepr-2022-70
https://library.keldysh.ru/preprint.asp?id=2022-70

**О р д е н а  Л е н и н а**
**ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ**
**имени М.В.Келдыша**
**Р о с с и й с к о й  а к а д е м и и  н а у к**

**В. А. Судаков, М. А. Тимофеев**

# Прогнозирование авиационных перевозок методами статистического анализа и машинного обучения

**Москва — 2022**

*Судаков В.А., Тимофеев М.А.*

**Прогнозирование авиационных перевозок методами статистического анализа и машинного обучения**

В работе рассматриваются актуальные методы прогнозирования временных рядов на примере внутренних и международных перевозок Российской Федерации в последние годы с учетом влияния внешних факторов. Разработаны модели с применением авторегрессии – скользящего среднего – и с использованием градиентного бустинга. Исследована возможность использования данных по заболеваниям COVID-19 для прогноза.

***Ключевые слова:*** анализ временных рядов, авиаперевозки, ARIMA, SARIMA, градиентный бустинг.

*Vladimir Anatolyevich Sudakov, Maksim Aleksandrovich Timofeev*

**Air traffic forecasting using statistical analysis and machine learning methods**

The paper considers current methods for forecasting time series on the example of domestic and international transportation of the Russian Federation in recent years, taking into account the influence of external factors. Models were developed using autoregressive moving average and using gradient boosting. The possibility of using data on COVID-19 diseases for forecasting was investigated.

***Key words:*** time-series analysis, aviation, ARIMA, SARIMA, gradient boosting.

Исследование выполнено при финансовой поддержке РФФИ и CNPq (Бразилия), Фонда содействия инновациям (Россия), DBT, DST (Индия), MOST, NSFC (Китай), SAMRC (ЮАР) в рамках научного проекта № 20-51-80002.

## Введение

Пассажирооборот – это ключевой индекс для построения долгосрочной стратегии авиакомпании, поэтому качество его прогнозирования критически важно как для частных авиакомпаний, так и для государственных органов, занимающихся нормативным регулированием в сфере воздушного транспорта.

В данной работе рассматриваются актуальные способы прогнозирования месячного пассажирооборота Российской Федерации, также применимые и для других временных рядов.

Существуют различные статистические методы, позволяющие предсказать будущие значения временного ряда [1]. Эти методы используются и для предсказания пассажирооборота. Сезонность и существенные внешние факторы обуславливают применение методов SARIMA и прогнозирование с помощью градиентного бустинга. Далее будут выявлены их особенности и ограничения.

## Подготовка и разведочный анализ данных

Для обучения модели были взяты статистические данные с декабря 2008 года по июль 2022 года. Гранулярность – один месяц. Итого 163 временных периода. Графики временных трендов показаны на рисунке 1. Рассматривается выполненный пассажирооборот в тысячах пассажиро-километров.

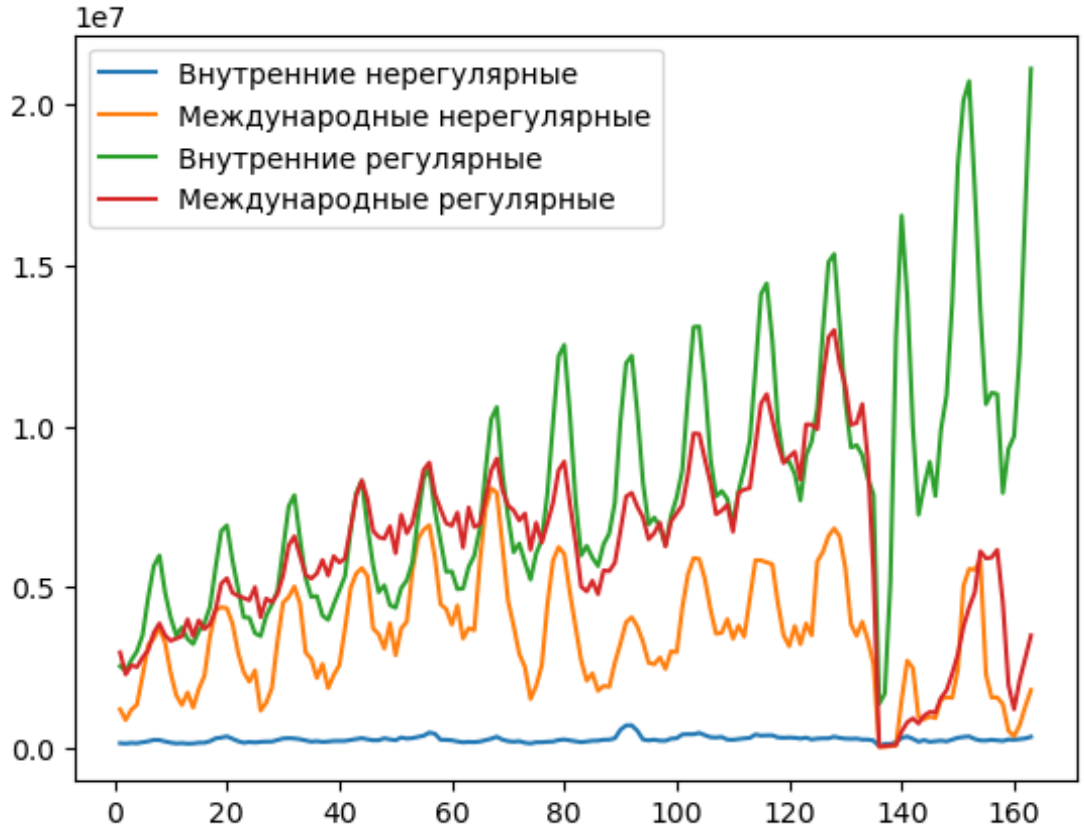


*Рис. 1.* Перевозки пассажиров в тысячах пассажиро-километров с декабря 2008 года по июль 2022 года: суммарные за месяц

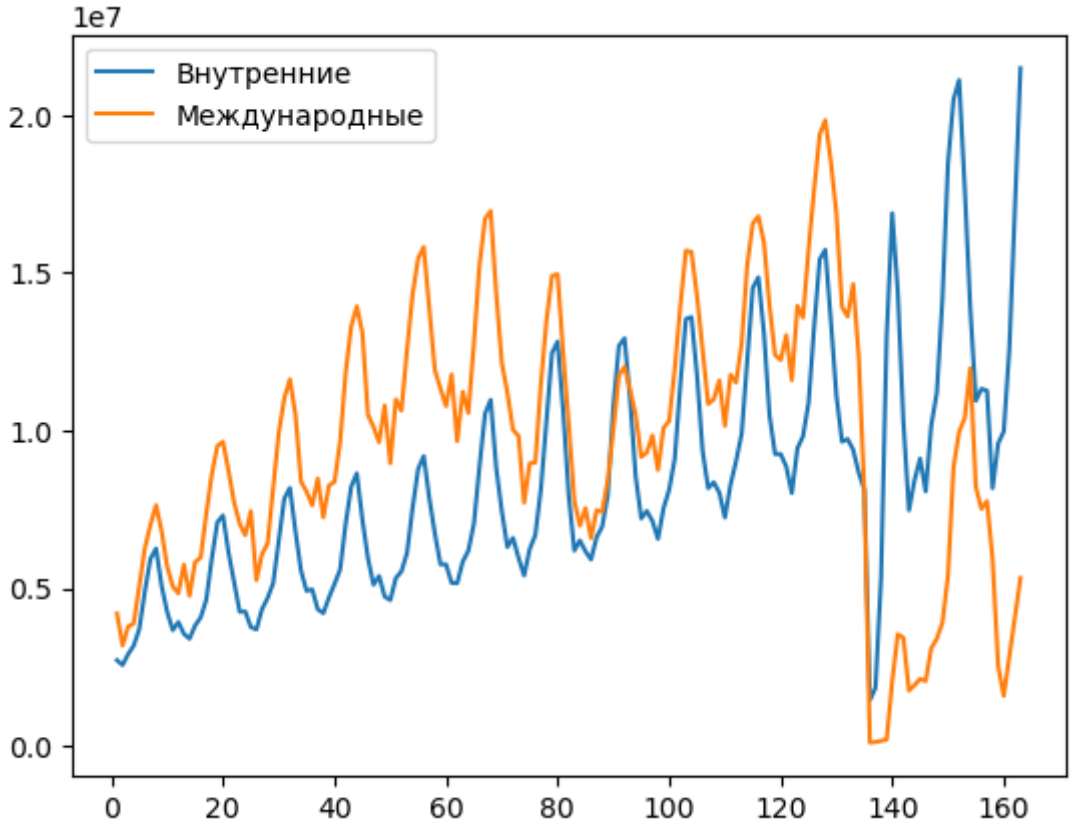


*Рис. 2.* Перевозки пассажиров в тысячах пассажиро-километров с декабря 2008 года по июль 2022 года: суммарные за месяц без детализации по регулярности

Регулярные и нерегулярные перевозки были объединены (см. рис. 2). Во временных рядах прослеживается сезонность. В периоде с марта 2022 года наблюдаем провал, особенно в международных перевозках. Из-за необходимости сделать отложенную выборку по крайней мере 10% от всей выборки для проверки качества модели этот период не попадет в обучающую выборку.

Временной ряд был разделен на три компонента: тренд, сезонность и остатки.

Интерпретация компонентов:

- тренд — общее направление ряда за длительный период времени;
- сезонность — отчетливая повторяющаяся закономерность, наблюдаемая через равные промежутки времени из-за различных сезонных факторов;
- остаток — нерегулярная составляющая, состоящая из колебаний временного ряда после удаления предыдущих составляющих.

Тренды показаны на рисунках 3 и 4.

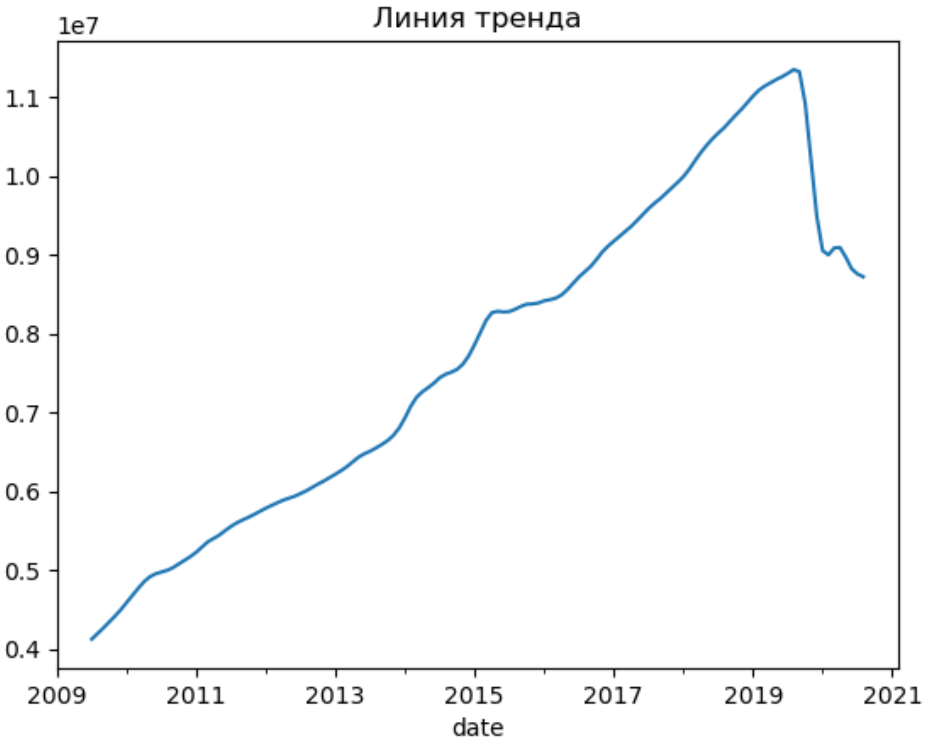


*Рис. 3.* Тренд для пассажирских перевозок по внутренним маршрутам


*Рис. 4.* Тренд для пассажирских перевозок по международным маршрутам

На рисунках 3 и 4 линия тренда резко изменила направление. В отложенной выборке тренд снова изменяется. Для международных перевозок есть один похожий случай в 2014, для внутренних – случай уникальный.

Сезонная компонента показана на рисунке 5. По графику отчетливо виден период в 12 месяцев. Далее вычисляется ряд статистических характеристик рядов.

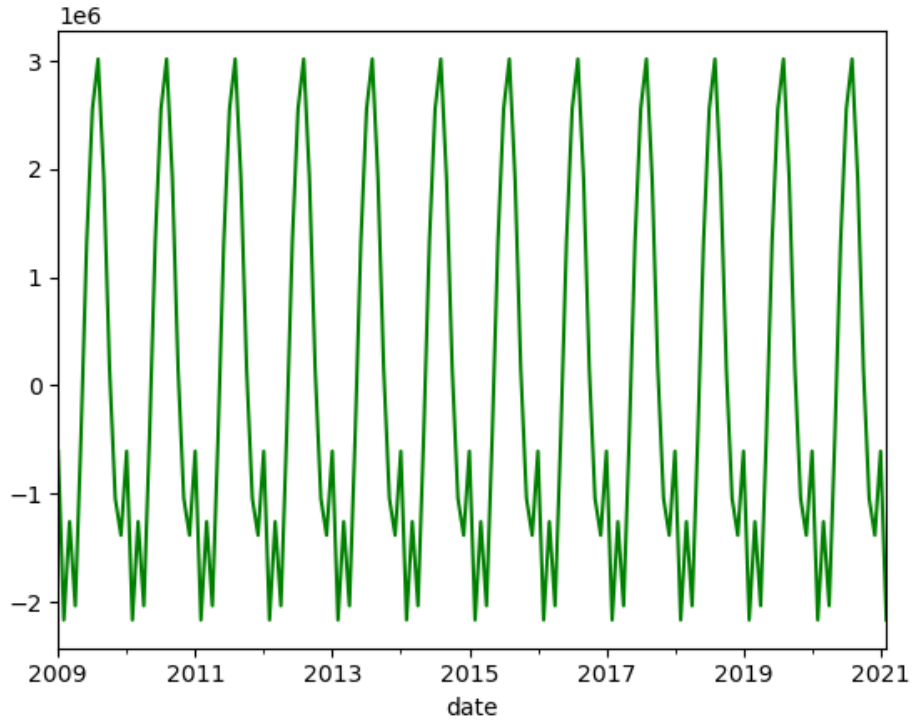


*Рис. 5.* Сезонность для пассажирских перевозок

Статистические характеристики для внутренних рейсов следующие:
1. Тест Дики-Фюллера: -1.5810001088552643.
2. P-значение: 0.4931303610438193.
3. Количество лагов: 14.
4. Критические значения:
   1%: -3.481281802271349;
   5%: -2.883867891664528;
   10%: -2.5786771965503177.

Статистические характеристики для международных рейсов:
1. Тест Дики-Фюллера: -2.3606173620621616;
2. P-значение: 0.1531616792220229.
3. Количество лагов: 14.
4. Критические значения:
   1%: -3.481281802271349;
   5%: -2.883867891664528;
   10%: -2.5786771965503177.

На основе полученных оценок и разведочного анализа можно сделать вывод, что ряды нестационарны, с выраженной сезонностью и, вероятно, будут хорошо описаны моделью SARIMA.

## Модели класса ARIMA

SARIMA – сезонная интегрированная модель авторегрессии – скользящего среднего, применяемая для прогнозирования и анализа временных рядов [2-3]. Является расширением модели ARMA, позволяющим моделировать нестационарные временные ряды с выраженной сезонностью. Этому классу моделей не требуются другие параметры, кроме самого временного ряда, так как следующее значение прогнозируется на основе предыдущих.

Авторегрессионные модели (AR) предполагают, что следующие значения временного ряда линейно зависят от предыдущих. Авторегрессионный процесс определяется формулой:

$X_t = c + \sum_{i=1}^{p} a_i X_{t-i} + \varepsilon_t$.

Модели скользящего среднего (MA) – модели, предполагающие, что ряд описывается следующей формулой:

$X_t = \sum_{j=0}^{q} b_j \varepsilon_{t-j}$.

Модель применима для интегрированных временных рядов, то есть рядов, которые можно привести к стационарным, путем взятия разности некоторого порядка от временного ряда.

Далее показана формула временного ряда модели SARIMA

$\Phi_p(B^m)\varphi(B)\nabla_m^D\nabla^d x_t = \Theta_Q(B^m)\theta(B)\omega_t$,

где:

$\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p$,

$\Phi_p(B^m) = 1 - \Phi_1 B^m - \Phi_2 B^{2m} - \cdots - \Phi_p B^{pm}$,

$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$,

$\Theta_Q(B^m) = 1 + \Theta_1 B^m + \Theta_2 B^{2m} + \cdots + \Theta_Q B^{Qm}$,

$\nabla^d = (1 - B)^d$,

$\nabla_m^D = (1 - B^m)^D$,

$B^k x_t = x_{t-k}$.

Модель определяется семью параметрами:

- $p$ – порядок авторегрессии модели. Временной ряд считается авторегрессивным, если предыдущие наблюдения хорошо описывают последующие;
- $d$ – порядок дифференцирования;
- $q$ – порядок скользящего среднего модели, то есть размер «скользящего окна»;
- $P$ – порядок сезонного компонента авторегрессии;
- $D$ – порядок дифференцирования сезонного компонента.
- $Q$ – порядок скользящего среднего сезонной компоненты модели.
- $m$ – размер сезонной компоненты. Для исследуемого временного ряда $m$=12, так как очевидна годовая сезонность.

Процесс подбора параметров можно автоматизировать, используя библиотеку pmdarima для языка Python. Параметры *d* и *D* определяются путем теста Дики-Фюллера для ряда и его сезонной компоненты соответственно.

Параметры $p$, $P$, $q$ и $Q$ подбираются итерацией по всей области, разрешенной для тестирования.

На рисунке 6 показаны графики характеристик ряда, необходимые для подбора параметров модели SARIMA.

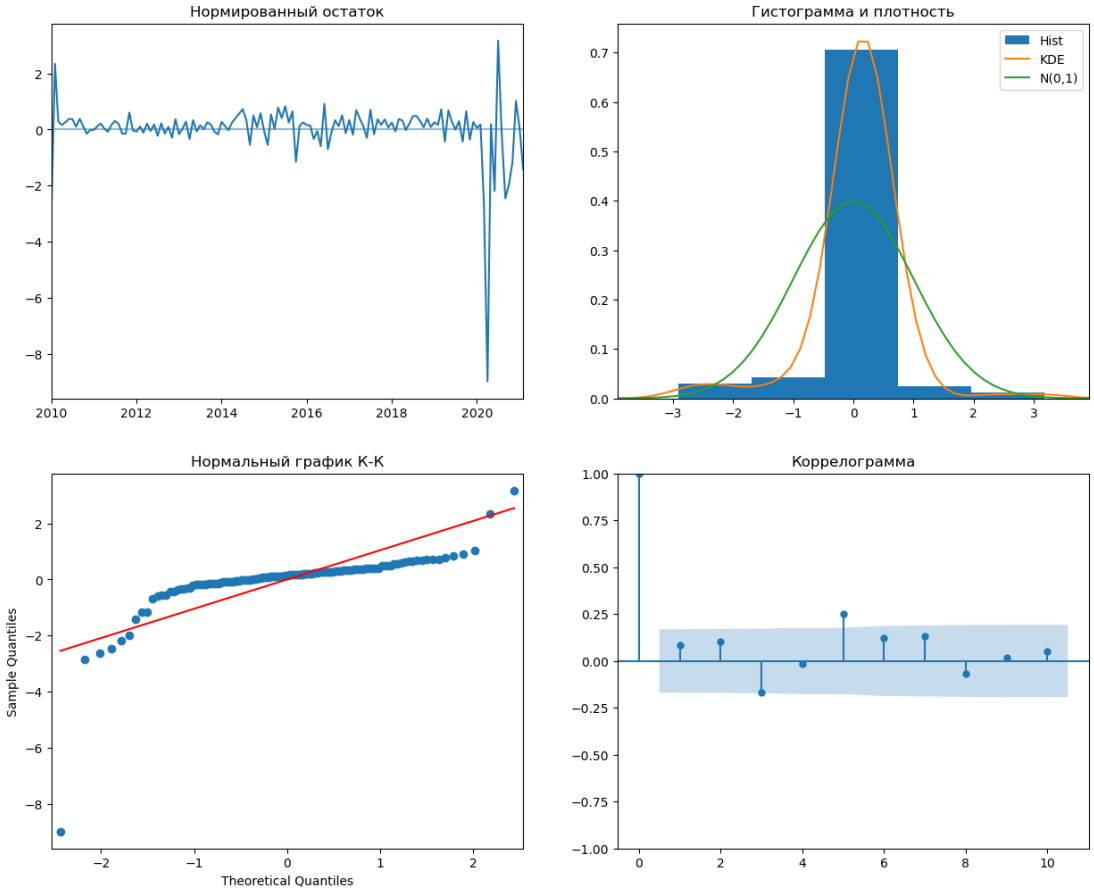


*Рис. 6.* Подбор параметров модели внутренних перевозок

Подбор параметров модели показывает следующие результаты:

- ошибки модели большую часть выборки находятся около нуля, но становятся значительно большими по модулю к концу выборки;
- плотность распределения значений мало соответствует нормальной;
- заметной корреляции в ошибках нет.

Прогноз для внутренних перевозок показан на рисунке 7. На нем видна отчетливая тенденция к постепенному росту, несмотря не некоторое временное падение.

Для оценки качества модели используется Mean Absolute Percentage Error (среднее абсолютное процентное отклонение), являющееся стандартной мерой точности предсказания метода прогнозирования в статистике.

MAPE модели SARIMA для внутренних перевозок равен 0.415.

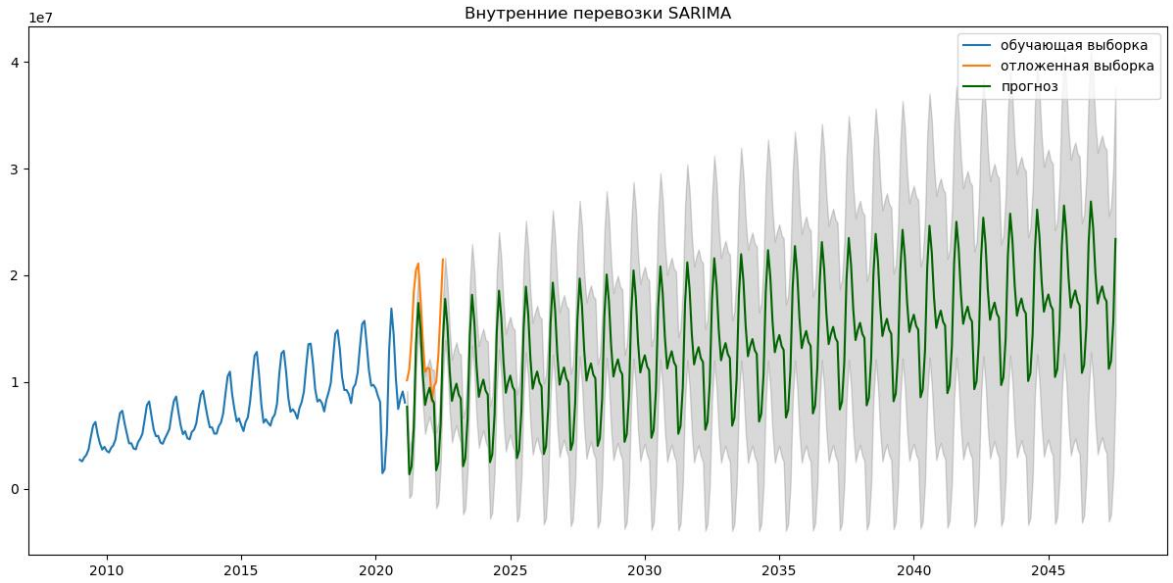


*Рис. 7.* Прогноз на 25 лет для внутренних перевозок с отложенной выборкой для проверки

Для международных перевозок выводы по параметрам модели схожие, но гистограмма значений лучше описывается нормальным распределением (см. рис. 8).

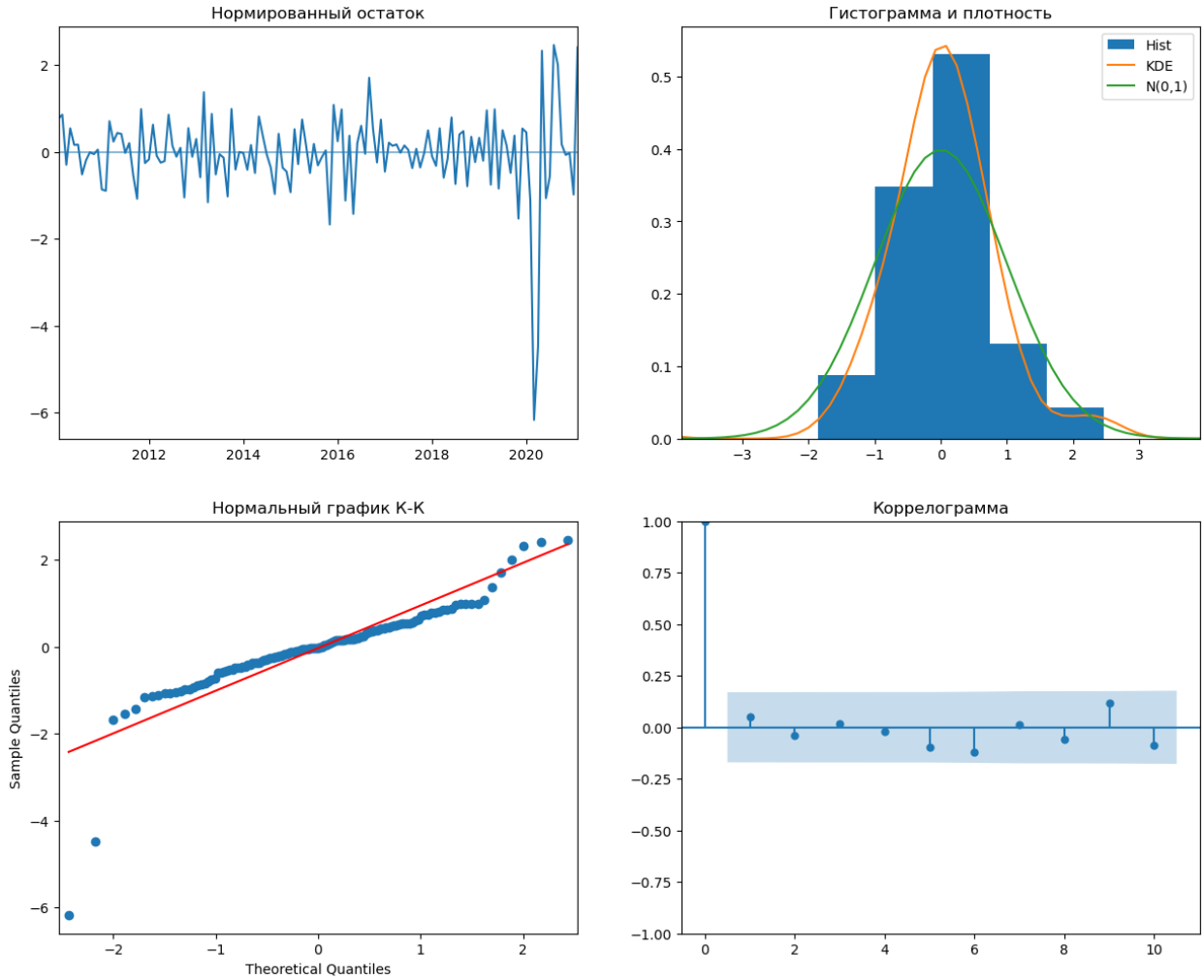


*Рис. 8.* Подбор параметров модели внутренних перевозок

Результат построения прогноза международных перевозок показан на рисунке 9. Модель предполагает нисходящий тренд, что, возможно, несправедливо, так как зависит от множества непрогнозируемых внешних факторов. Несмотря на тенденцию к их уменьшению по наиболее вероятному значению, есть некоторая небольшая вероятность как роста, так и более резкого уменьшения, что показывает конус неопределенности.

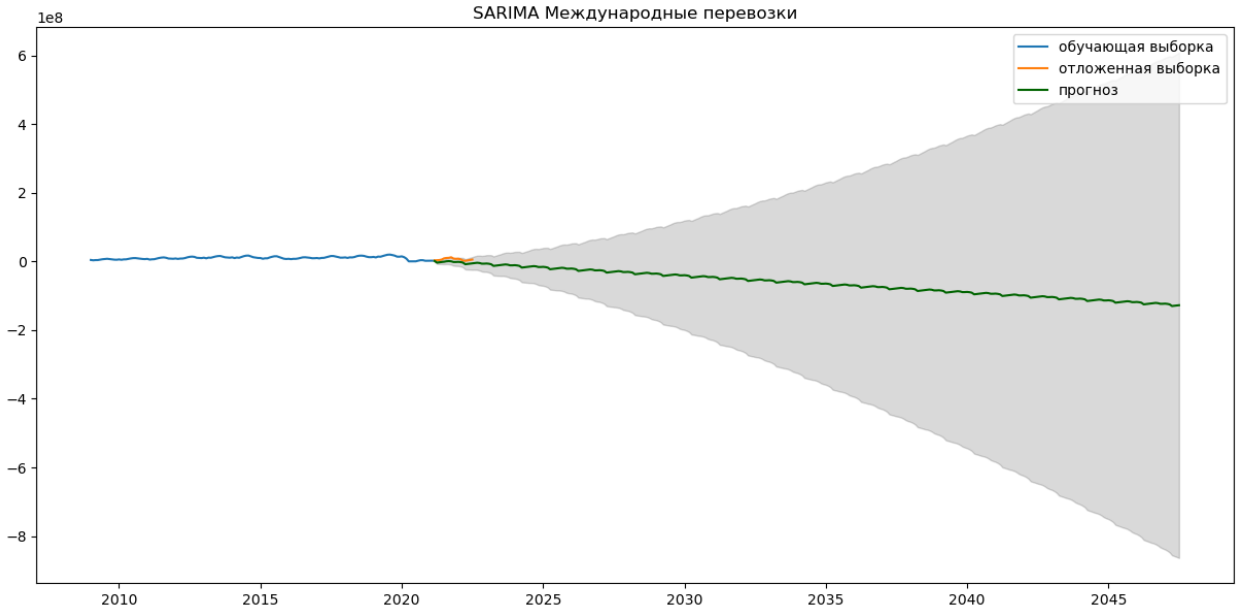


*Рис. 9.* Прогноз на 25 лет для международных перевозок с отложенной выборкой для проверки

Для международных перевозок модель не показывает вразумительных результатов из-за меняющегося тренда, но адаптируется к изменениям, когда новый тренд формируется. Высокая волатильность последних месяцев отражается в большой неуверенности модели и широком 95% доверительном интервале.

## Градиентный бустинг

Для параметрических моделей необходимо сгенерировать признаки, по которым можно будет строить прогноз: например, в недавнем прошлом это количество подтвержденных заболеваний COVID-19, год, месяц, сезон. Для международных перевозок возможным признаком было бы количество санкций, введенных против РФ. Даже в таком случае спрогнозировать эти признаки достаточно хорошо не представляется возможным.

Для моделирования временных рядов также можно использовать параметрические модели [4]. Они лучше всего описывают ряды без четкой внутренней структуры, но с явными факторами, влияющими на значения. В таком случае необходимо создать признаки, по которым будет составляться прогноз. Учитывая гранулярность ряда, для моделирования были выбраны следующие параметры: год, номер месяца, сезон, также данные по COVID-19.

При выборе параметра, отражающего влияние covid-19 на авиаперевозки, есть выбор между двумя параметрами: общее количество заболевших (covid) или месячный прирост заболевших (covid-rolling). Так как ограничения могут вводиться и сниматься, отвечая волнам новых заболеваний, второй параметр может сработать лучше. Для сравнения предсказательной силы параметров была построена коррелограмма.

| | Domestic | International | covid | covid_rolling | year | month |
|---|---|---|---|---|---|---|
| **Domestic** | 1.000000 | 0.323102 | 0.443107 | 0.200854 | 0.691486 | 0.203303 |
| **International** | 0.323102 | 1.000000 | -0.344967 | -0.313481 | 0.013691 | 0.185961 |
| **covid** | 0.443107 | -0.344967 | 1.000000 | 0.659118 | 0.517982 | -0.057882 |
| **covid_rolling** | 0.200854 | -0.313481 | 0.659118 | 1.000000 | 0.408869 | -0.071296 |
| **year** | 0.691486 | 0.013691 | 0.517982 | 0.408869 | 1.000000 | -0.053258 |
| **month** | 0.203303 | 0.185961 | -0.057882 | -0.071296 | -0.053258 | 1.000000 |

*Рис. 10.* Коррелограмма признаков модели со значением ряда

Коррелограмма показывает небольшие корреляции признаков с целевыми значениями. Для обнаружения зависимостей нужен более мощный метод, чем линейная регрессия.

Для параметрического моделирования временных рядов используется градиентный бустинг. Метод заключается в том, что для регрессии строится простое решающее дерево или другой алгоритм машинного обучения, его ошибки моделируются другим алгоритмом. В дальнейшем ошибки всего ансамбля моделируются следующими деревьями. Для реализации была выбрана библиотека Catboost.

Catboost — это библиотека для языка Python, созданная компанией Yandex, использующая градиентный бустинг, основанный на решающих деревьях. Библиотека часто применяется как конечное решение для задач поиска, рекомендательных систем, также применима для прогнозирования временных рядов. В библиотеку уже встроена функция подбора параметров методом кросс-валидации.

MAPE для модели на основе Catboost для внутренних перевозок равен 0.4585. Бустинг лучше повторяет обучающую выборку, но делает менее правдоподобный прогноз (см. рис. 11).

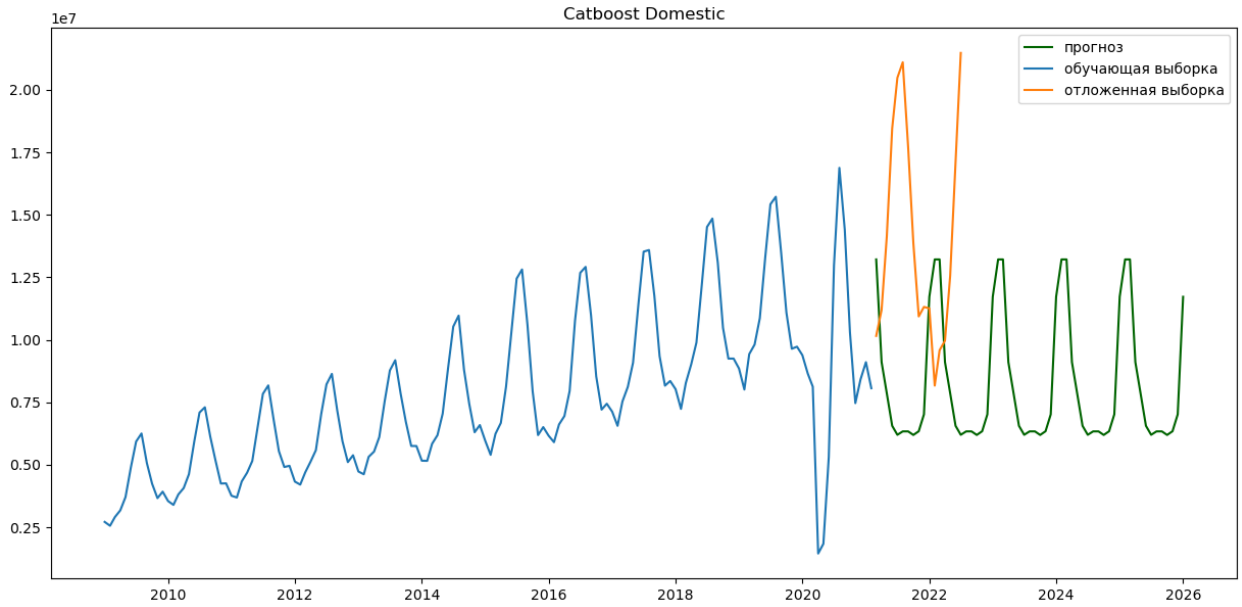


*Рис. 11*. Catboost прогноз для внутренних перевозок

MAPE модели на основе Catboost для международных перевозок: 1.392887193177242.

Бустинг как метод легко подвержен переобучению, в этом случае для долгосрочного прогнозирования не подходит.

К другим минусам можно отнести сложность в сборе признаков. Признаки на основе месяца можно генерировать неограниченно, но влияние глобальных эпидемий и других внешних факторов — отдельная задача для прогноза.

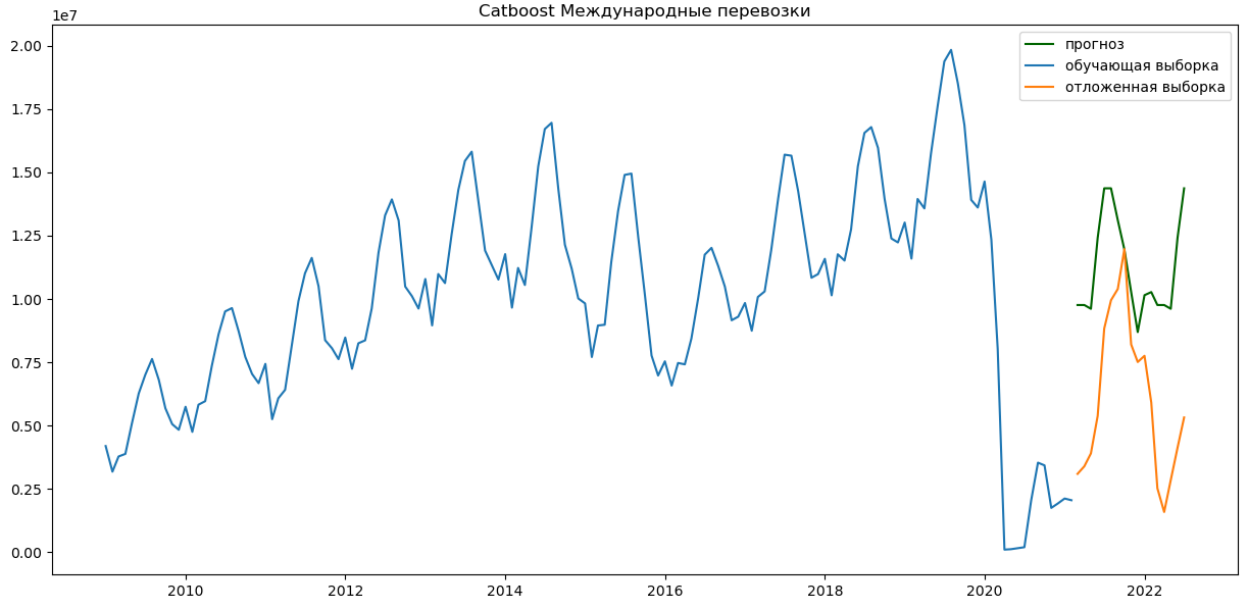


*Рис. 12*. Catboost прогноз для международных перевозок

## Заключение

Были созданы две модели, описывающие временной ряд, рассмотрены сильные и слабые стороны каждого подхода к прогнозированию авиаперевозок.

Градиентный бустинг применительно к временным рядам хорошо повторяет оригинальный ряд, но склонен к переобучению, из-за чего его предсказательная способность снижена. Возможно, для временных рядов, более зависящих от прогнозируемых внешних переменных, его использование будет более оправдано, чем непараметрическое моделирование.

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.

## Библиографический список

1. Xiaojia Guo, Yael Grushka-Cockayne, Bert De Reyck. Forecasting Airport Transfer Passenger Flow Using Real-Time Data and Machine Learning // Manufacturing & Service Operations Management. July 2021. URL: https://doi.org/10.1287/msom.2021.0975.

2. Айвазян С. А. Прикладная статистика. Основы эконометрики. Том 2. — М.: Юнити-Дана, 2001. — 432 с. — ISBN 5-238-00305-6.

3. Tang X., Deng G. Prediction of Civil Aviation Passenger Transportation Based on ARIMA Model. Open Journal of Statistics, 6, 824-834. 2016. URL: https://doi.org/10.4236/ojs.2016.65068.

4. Yoonchul R. Multi-State and Individual State Time Series Model Comparison. Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the Degree of Master of Science URL: https://repository.lib.ncsu.edu/bitstream/handle/1840.20/38621/etd.pdf.

# Оглавление