Федеральное государственное автономное образовательное учреждение высшего образования **«Национальный исследовательский университет ИТМО»**

Факультет Программной Инженерии и Компьютерной Техники

Экзаменационная работа

по дисциплине «Математическая статистика»

«Исследование зависимости пассажиропотока такси от погодных условий»

Лектор:

Яворук Татьяна Олеговна

Практик:

Шкваренко Андрей Алексеевич

Выполнили:

Кулешова Екатерина Михайловна

Соколова Полина Дмитриевна

Группа: P3215

Поток: 22.5

Санкт-Петербург

2025 г.

Оглавление

1. Введение

1.1 Актуальность исследования

В условиях постоянного роста урбанизации и плотности городской застройки службы такси становятся неотъемлемой частью транспортной инфраструктуры мегаполисов. Современные рынки пассажирских перевозок активно реагируют на внешние факторы, среди которых метеоусловия играют ключевую роль в формировании спроса на такси. Чёткое понимание влияния температуры, осадков и ветра позволяет оптимизировать распределение автопарка, сократить время ожидания клиентов и повысить эффективность работы диспетчерских служб. В условиях изменчивой климатической ситуации и роста конкуренции анализ погодных эффектов становится особенно важным для стратегического планирования и оперативного реагирования.

1.2 Цель

Количественно оценить чистое влияние основных метеофакторов (температуры, осадков, скорости ветра) на пассажиропоток такси с учётом временных паттернов.

1.3 Задачи

1. Осуществить сбор данных агрегаторов такси выбранного населенного пункта.
2. Осуществить сбор данных о погодных условиях выбранного населенного пункта.
3. Произвести фильтрацию и преобразование данных для слияния в один датасет.
4. Произвести декомпозиция временного ряда.
5. Оценить качество данных.
6. Выбрать модель для описания влияния погодных факторов на количество заказов такси.
7. Рассчитать метрики качества, исследовать частичные зависимости и сделать выводы по каждому фактору.

2. Теоретические основы

1.4 Декомпозиция

1. Декомпозиция временного ряда.
   Декомпозиция временного ряда — это метод разложения исходной последовательности наблюдений y_i на несколько базовых компонент, отражающих

различные виды закономерностей во времени. В классическом случае выделяют три основные составляющие:

Тренд (T_i) — отображает медленную, низкочастотную эволюцию уровня ряда: общее направление на протяжении всего периода наблюдений.

Сезонность (S_i) — отображает повторяющиеся циклы фиксированной длины (день, неделя, месяц, год), отражающие регулярные колебания.

Остатки или шум (R_i) — отображает всё, что не захвачено трендом и сезонностью: случайные флуктуации, выбросы, нерегулярные эффекты.

Формулы:

$y_i = T_i + S_i + R_i$ — аддитивная модель

$y_i = T_i \times S_i \times R_i$ — мультипликативная модель

2. MSTL-декомпозиция

MSTL — это расширение классического STL-разложения, позволяющее учитывать сразу несколько накладывающихся сезонных циклов в одном ряду, последовательно выделяя каждую сезонность. Составляющие:

Тренд (T_i) — отображает медленную, низкочастотную эволюцию уровня ряда: общее направление на протяжении всего периода наблюдений.

Сезонность ($S_i^{(j)}$) — отображает периодические колебания ряда с фиксированным периодом m_j.

$$S_i = \sum_{j=1}^{k} S_i^{(j)}$$

В сумме все k сезонных компонент дают полную картину циклических флуктуаций ряда, которую затем отделяют от тренда и остатка для более глубокого анализа.

Остатки или шум (R_i) — отображает всё, что не захвачено трендом и сезонностью: случайные флуктуации, выбросы, нерегулярные эффекты.

Формула:

$y_i = \sum_{j=1}^{k} S_i^{(j)} + T_i + R_i$ — аддитивная модель

1.5 Оценка данных

1. Анализ выбросов с помощью «ящиков с усами»

Цель: выявить экстремальные значения, которые могут искажать последующий анализ.

Ключевые статистики и формулы:

Q_1, Q_3 — первый и третий квартили.

$IQR = Q_3 - Q_1$ — межквартильный размах

$Q_1 - 1.5 \times IQR$ — левая/нижняя граница

$Q_3 + 1.5 \times IQR$ — правая/верхняя граница

2. Форма распределения признаков

   Цель: оценить центральную тенденцию, разброс, асимметрию и «тяжесть» хвостов.

   Ключевые статистики и формулы:

   $\bar{x} = \frac{1}{n}\sum_i x_i$— среднее

   $s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$— дисперсия

   $Skew = \frac{\frac{1}{n}\sum_i (x_i-\bar{x})^3}{\left(\frac{1}{n}\sum_i (x_i-\bar{x})^2\right)^{\frac{3}{2}}}$ — коэффициент асимметрии

   $Kurt = \frac{\frac{1}{n}\sum_i (x_i-\bar{x})^4}{\left(\frac{1}{n}\sum_i (x_i-\bar{x})^2\right)^{2}} - 3$ — коэффициент эксцесса

   $\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x-x_i}{h})$ — оценка плотности KDE

3. Корреляционный анализ

   Цель: определить наличие и силу взаимосвязей между признаками.

   Ключевые статистики и формулы:

   $r_{XY} = \frac{\sum_i (X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_i (X_i-\bar{X})^2}\sqrt{\sum_i (Y_i-\bar{Y})^2}}$ — Пирсоновский коэффициент

   $t-$статистика для проверки $H_0: r = 0$

   $t = r\sqrt{\frac{n-2}{1-r^2}},\ t \sim t_{n-2}$

4. Оценка мультиколлинеарности

   Цель: проверить, не слишком ли один признак объясняется остальными.

   Ключевые статистики и формулы:

   $$VIF_j = \frac{1}{1 - R_j^2}$$

5. Нелинейные зависимости

   Цель: проверить, есть ли систематические (непрямолинейные) связи между признаками и остатками.

   Ключевые статистики и формулы:

   $$\hat{m}(x) = \arg\min_{a,b} \sum_{i=1}^{n} w_i(x)\left(R_i - a - b(X_i - x)\right)^2$$

   $w_i(x) = K(\frac{X_i - x}{h})$

6. Множественная линейная регрессия и анализ автокорреляции остатков

Цель: смоделировать $y_i = \beta_0 + \sum_j \beta_j X_{ij} + \varepsilon_i$ и проверить свойства ошибок.

Ключевые статистики и формулы:

$\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$ — оценка коэффициентов

$RSS = \sum_{i=1}^{n}(y_i - \hat{y})$ — сумма квадратов остатков

$TSS = \sum_{i=1}^{n}(y_i - \bar{y})$ — полная сумма квадратов

$ESS = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})$ — объяснённая сумма квадратов

$R^2 = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS}$ — коэффициент детерминации:

$R_{adj}^2 = 1 - \frac{(1-R^2)(n-1)}{n-p-1}$ — скорректированный коэффициент детерминации

$F = \frac{(RSS_0 - RSS_1)/q}{RSS_1/(n-p)}$ — F-тест модели
$F \sim F_{q;n-p}$

$DW = \frac{\sum_{t=2}^{n}(\hat{\varepsilon_t} - \widehat{\varepsilon_{t-1}})^2}{\sum_{t=1}^{n}\widehat{\varepsilon_t^2}}$ — тест Дарбина-Уотсона

$Q = n(n+2)\sum_{k=1}^{m}\frac{\widehat{p_k^2}}{n-k}, Q \sim \chi_m^2$ — тест Льюнга–Бокса

$BP = \frac{nR_{aux}^2}{2} \sim \chi_p^2$ — тест Бройша–Пагана

$JB = \frac{n}{6}\left(Skew^2 + \frac{1}{4}(Kurt-3)^2\right) \sim \chi_2^2$ — тест Жарке–Бера

$p(k) = \frac{\gamma(k)}{\gamma(0)}, y(k) = \sum_t \frac{\hat{\varepsilon}_t \hat{\varepsilon}_{t-k}}{n}$ — ACF/PACF

1.6 Модель GAM

Модель GAM (обобщённая аддитивная модель) — это статистический подход, который расширяет линейные модели, позволяя учитывать нелинейные зависимости между признаками и целевой переменной.

Формула в общем виде:

$g(\mathbb{E}[Y]) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_n(x_n)$, где $f_i(x_i) -$ сглаживающая функция (сплайн)

1.7 Оценка качества модели

1. Численные метрики

$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|$ — средняя абсолютная ошибка

$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}$ — среднеквадратическое отклонение

2. Качество сглаживания сплайнами

   $AIC = 2k - 2\ln(L)$ – критерий Акаике

   $GCV = \frac{1}{n}\sum_{t=1}^{n}\left(\frac{y_t - \hat{y}_t}{1 - s_{tt}}\right)^2$ – обобщенная перекрестная проверка (s_{tt} - элементы диагонали матрицы сглаживания)

   $R^2_{pseudo} = 1 - \frac{\sum(y_t - \hat{y}_t)^2}{\sum(y_t - \bar{y})^2}$

3. Диагностика остатков (ACF)

   $p(k) = \frac{\gamma(k)}{\gamma(0)}, y(k) = \sum_t \frac{\hat{\varepsilon}_t \hat{\varepsilon}_{t-k}}{n}$

4. Параметры AR(2)
5. Качественная проверка через частичные зависимости

3. Данные

Фрагмент таблицы с входными данными:

	A	B	C	D	E
1	date_and_time	number_of_taxi_trips	temperature	precipitation	wind_speed
2	01/01/2024 12:00:00 AM	462	0,6	0	27,7
3	01/01/2024 01:00:00 AM	522	0	0	20,5
4	01/01/2024 02:00:00 AM	490	0,6	0	25,9
5	01/01/2024 03:00:00 AM	269	0,6	0	25,9

В таблице содержатся данные о количестве вызовов такси и погоде в Чикаго 2024 с 01.01 по 01.03.

Столбцы:

data_and_time – дата и время в формате мм/дд/гггг чч:мм:сс AM/PM

number_of_taxi_trips – количество вызовов такси в ближайший час с момента, указанного в столбце data_and_time

temperature – значение температуры в градусах цельсия в начале рассматриваемого часа

precipitation – количество осадков в миллиметрах в начале рассматриваемого часа

wind_speed – скорость ветра в км/ч

4. Основная часть

4.1 Код решения

1. Исходный код решения представлен в репозитории на GitHub [1] (https://github.com/pollee343/mathematical_statistics)

2. Используемые программные средства

   В проекте применялась связка *Python 3.11* (дистрибутив Anaconda 2024.02) и интерактивная среда *Jupyter Notebook* в редакторе *Visual Studio Code 1.88*; вычислительная часть выполнена с использованием библиотек pandas 2.2 и numpy 1.26 для подготовки данных, statsmodels 0.14 (модуль tsa) и pmdarima 2.0 для MSTL-декомпозиции и оценки AR-компоненты, а также pygam 0.9.0 для построения обобщённой аддитивной модели; метрики MAE и RMSE рассчитывались через scikit-learn 1.4, визуализация реализована на matplotlib 3.8 и seaborn 0.13; код версионировался в приватном репозитории GitHub при помощи *Git 2.44*

4.2 Декомпозиция с использованием MSTL

Для исходного ряда количества поездок (number_of_taxi_trips) предполагается аддитивная модель:

$$y_i = S_i^{(24)} + S_i^{(168)} + T_i + R_i$$

y_i — число поездок в такси в час i

$S_i^{(24)}$ — суточная сезонность

$S_i^{(168)}$ — недельная сезонность

T_i — тренд (долгосрочная составляющая)

R_i — остатки

MSTL итеративно оценивает каждый компонент с помощью локально-взвешенного регрессионного сглаживания (LOESS).

$\hat{R}_i$ служит зависимой переменной при построении регрессионных и GAM-моделей, позволяя оценить дополнительный эффект погодных факторов без смешения с трендовыми и сезонными паттернами.

Анализ остатков $\hat{R}_i$ повышает точность прогнозных моделей и интерпретируемость результатов, поскольку все регулярные компоненты уже учтены в T_i и S_i.

4.3 Оценка данных

Оценка данных необходима для проверки распределения погодных признаков, выявления взаимосвязи между признаками и оценки их пригодности для построения регрессионных моделей, и чтобы выявить возможные мультиколлинеарные эффекты и нелинейные зависимости с остатками временного ряда.

Пусть в каждый момент времени i заданы погодные переменные $x_i = (x_i^{(1)}, x_i^{(2)}, x_i^{(3)})$

$x_i^{(1)} = temperature_i$ — температура

$x_i^{(2)} = precipitation_i$ — осадки

$x_i^{(3)} = wind\_speed_i$ — скорость ветра

и остатки ряда $\hat{R}_i$ из шагов декомпозиции.

1. Анализ распределений и выбросов

   Box-plot каждого x_i для выявления выбросов

   Гистограмма и KDE для нахождения плотности распределения $\hat{f}_i(x)$ для оценки асимметрии и модальности.

   Центр распределения: ~2 °C, размах центральных 50 % значений в [–1; +4] °C.

   Умеренные выбросы по обе стороны (сильные морозы и аномальное тепло).

   Для модели: учесть экстремумы (например, добавить индикаторы «очень холодно/очень тепло» или применить робастные методы), но базовая часть данных уже компактна.

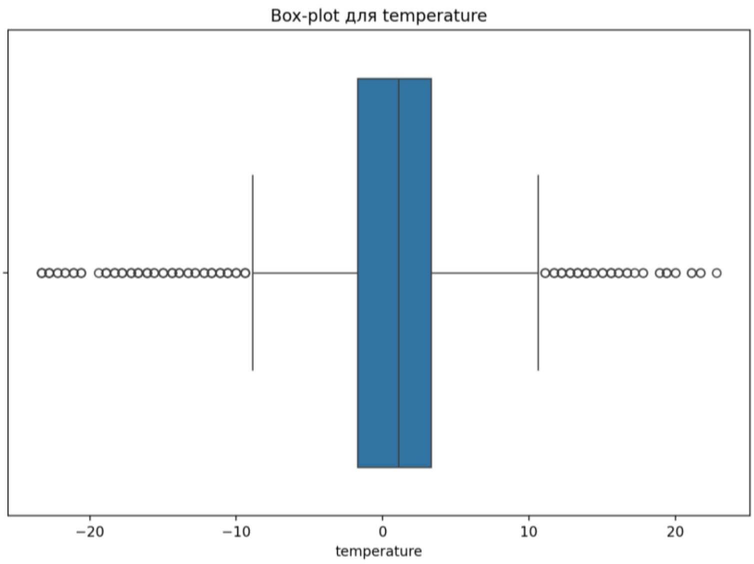

Рисунок 1. Диаграмма «ящик с усами» распределения суточных температур

На рисунке показаны ключевые статистики: медиана (центральная линия), границы первого и третьего квартилей (нижний и верхний края «ящика»), усы (границы диапазона без экстремальных выбросов) и выбросы (отдельные точки за пределами усов). Такой формат позволяет оценить центральную тенденцию, разброс и наличие аномальных значений в температурном ряде.

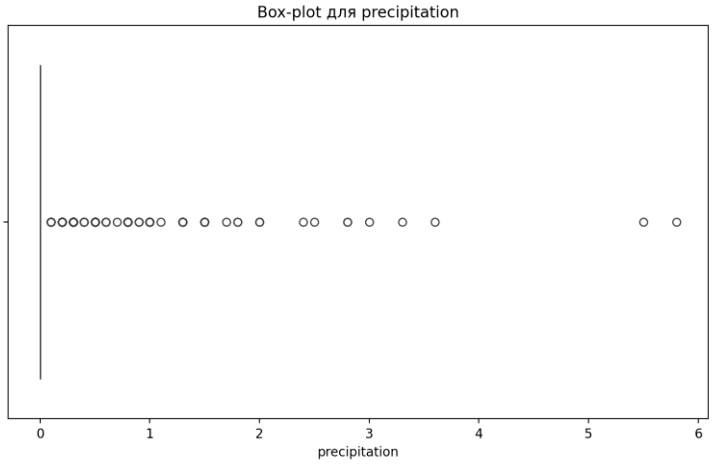

Рисунок 2. Диаграмма «ящик с усами» распределения дневных суммарных осадков

Диаграмма демонстрирует, что более половины наблюдений имеют нулевое значение осадков (медиана на отметке 0), при этом ненулевые значения сосредоточены в узком диапазоне малого дождя (межквартильный размах). Усы и выбросы вправо указывают на редкие, но интенсивные ливни.

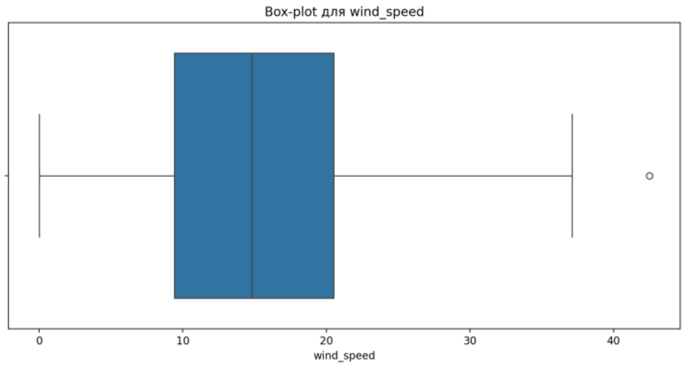

Рисунок 3. Диаграмма «ящик с усами» распределения скорости ветра

График отображает симметричное распределение скоростей ветра с медианой около 20 км/ч и интерквартильным размахом примерно 10–30 км/ч. Отдельный выброс указывает на редкое событие сильного порыва ветра свыше 40 км/ч.

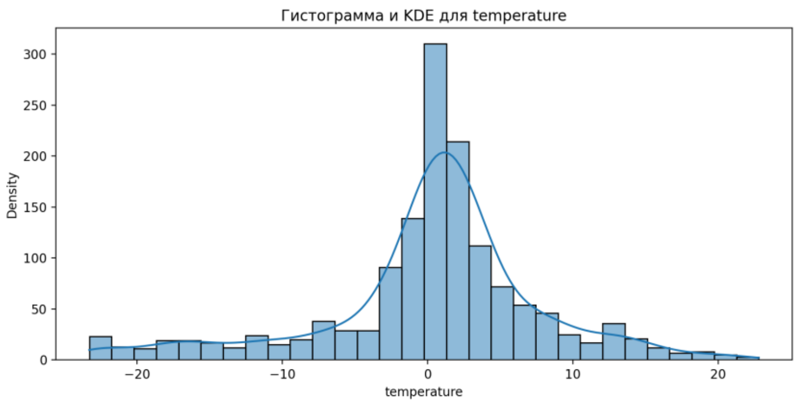

Рисунок 4. Гистограмма с ядерной оценкой плотности распределения суточных температур

График представляет собой комбинацию двух элементов визуализации:
Гистограмма: разбиение выборки по равномерным интервалам (бинам) вдоль оси абсцисс, высота столбцов отражает эмпирическую плотность распределения температурных наблюдений.
Ядерная оценка плотности (KDE): гладкая кривая, аппроксимирующая истинную функцию плотности распределения на основе ядерного метода.
Ключевые характеристики распределения:
Мода располагается в интервале приближённо [0; 2] °C, что указывает на наибольшую концентрацию наблюдений.
Правосторонняя асимметрия (положительный сдвиг): удлинённый хвост в область высоких температур свидетельствует о реже встречающихся, но значительных тёплых значениях.
Левосторонний хвост: наличие выбросов в область экстремально низких температур (до –25 °C).
Диапазон основной плотности охватывает примерно [–5; +8] °C, где сосредоточены около 75 % наблюдений.

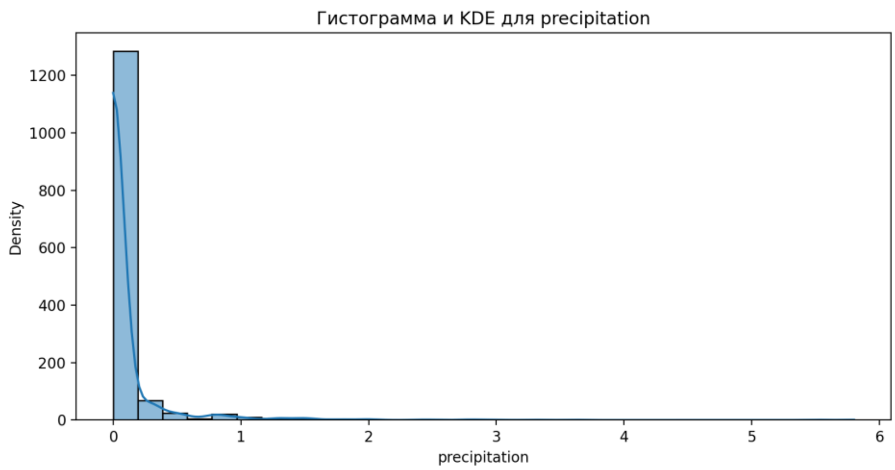

Рисунок 5. Гистограмма с ядерной оценкой плотности распределения дневных суммарных осадков

График представляет собой комбинацию двух элементов визуализации:
Гистограмма: разбивка выборки значений осадков по равным интервалам (бинам) вдоль оси абсцисс; высота столбцов отражает эмпирическую плотность распределения ежедневных осадков.
Ядерная оценка плотности (KDE): гладкая кривая, аппроксимирующая истинную функцию плотности распределения на основе ядерного метода.
Ключевые характеристики распределения:
Пиковая концентрация наблюдений при 0 мм осадков (мода на отметке 0), что свидетельствует о преобладании сухих дней.

Правосторонняя асимметрия: удлинённый хвост в область небольших и умеренных осадков (до ~1 мм), обусловленный реже встречающимися дождливыми днями.

Экстремальные значения: единичные наблюдения интенсивных выпадений до 5–6 мм, лежащие за пределами основного «хвоста».

Основная масса данных локализуется в диапазоне [0; 0.5] мм, где сосредоточена существенно более 75 % ненулевых значений.

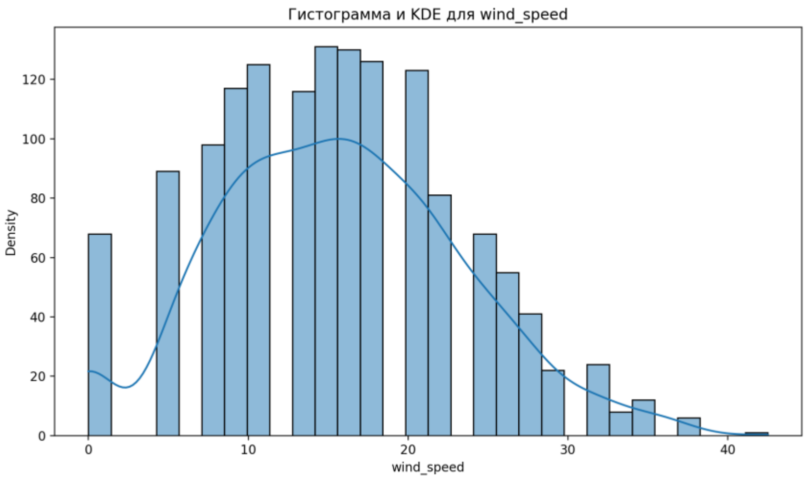

Рисунок 6. Гистограмма с ядерной оценкой плотности распределения скорости ветра

График представляет собой комбинацию двух элементов визуализации:

Гистограмма: разбивка выборки значений осадков по равным интервалам (бинам) вдоль оси абсцисс; высота столбцов соответствует эмпирической плотности наблюдений скоростей ветра.

Ядерная оценка плотности (KDE): непрерывная кривая, аппроксимирующая истинную функцию плотности распределения с учётом сглаживания ядерным методом.

Ключевые характеристики распределения:

Мода располагается в диапазоне приблизительно [14; 15] км/ч, отражая наиболее частые значения скорости ветра.

Интерквартильный размах (IQR) охватывает примерно [10; 20] км/ч, что указывает на центральную концентрацию 50 % наблюдений.

Почти симметричное распределение с умеренным правосторонним смещением: хвост в область высоких скоростей простирается до 40+ км/ч.

Экстремальные значения: редкие случаи порывов до ~43 км/ч, лежащие за пределами основного хвоста.

2. Корреляционный анализ

Вычисляем матрицу попарных коэффициентов корреляции.

$$\rho_{ij} = \mathrm{Corr}\left(x^{(i)}, x^{(j)}\right) = \frac{\mathrm{Cov}\left(x^{(i)}, x^{(j)}\right)}{\sigma_i \sigma_j}, \text{ где } \sigma_i = \sqrt{\mathrm{Var}(x^{(i)})}.$$

Отображается тепловой картой для наглядности.

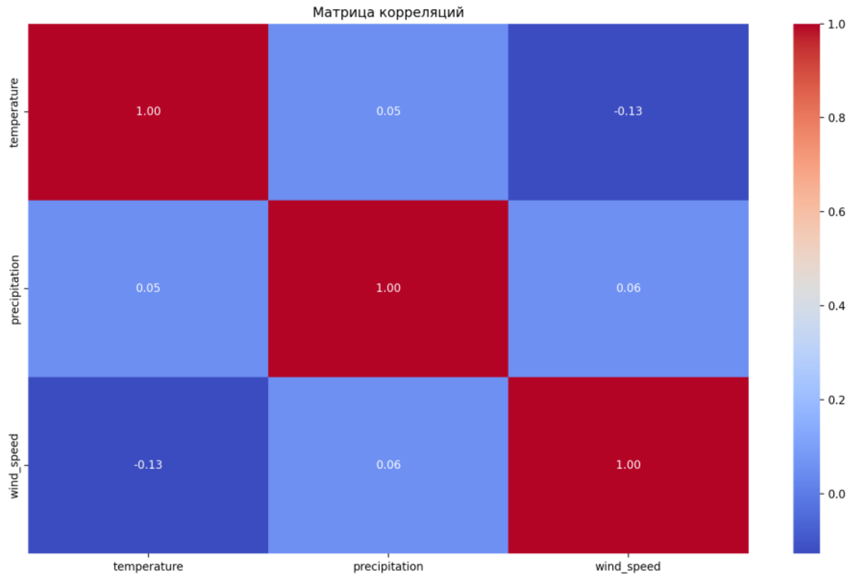

Рисунок 7. Тепловая карта матрицы попарных коэффициентов корреляции

График представляет собой визуализацию попарных коэффициентов корреляции Пирсона между переменными:

Коэффициент корреляции между temperature и precipitation равен +0.05, что указывает на практически полное отсутствие линейной зависимости.

Коэффициент корреляции между temperature и wind_speed равен –0.13, что свидетельствует о слабой отрицательной связи (с ростом температуры скорость ветра слегка снижается).

Коэффициент корреляции между precipitation и wind_speed равен +0.06, что также указывает на практически нулевую линейную зависимость.

На тепловой карте:

Цветовая шкала от –1 (тёмно-синий) до +1 (тёмно-красный) демонстрирует силу и направление связи.

Ячейки на главной диагонали равны единице (самокорреляция переменных).

Данное представление наглядно подтверждает слабую взаимосвязь между всеми парами рассматриваемых метеопараметров.

3. Variance Inflation Factor (VIF)

Для каждой переменной $x^{(i)}$ строится регрессия на остальные:

$$x_t^{(i)} = \sum_{j \neq i} \gamma_{ij} \, x_t^{(j)} + \epsilon_t^{(i)}$$

$$R_i^2 = 1 - \frac{\mathrm{Var}\left(\epsilon_t^{(i)}\right)}{\mathrm{Var}\left(x_t^{(i)}\right)}$$

Затем $VIF_i = \frac{1}{1-R_i^2}$

Если $VIF_i > 5 - 10$, признак сильно коррелирует с остальными. Отсутствие значимой мультиколлинеарности соответственно при $VIF_j < 5$

Переменная	R_i^2	VIF_j
temperature	0.0048	1.005
precipitation	0.0577	1.061
wind_speed	0.0549	1.058

Таблица 1. Оценка мультиколлинеарности регрессионных переменных

Все рассчитанные VIF-значения существенно ниже критических уровней (5–10), что свидетельствует об отсутствии проблем мультиколлинеарности в исходном наборе признаков. Каждая из переменных вносит в модель независимый вклад, и исключать какие-либо признаки по этой причине не требуется.

4. LOESS-анализ остатков

Для каждой пары $\left(x_t^{(i)}, \widehat{R_t}\right)$ строится scatter-plot и LOESS-кривая $\widehat{R_t} \approx g_i\left(x_t^{(i)}\right)$, где g_i получается локально взвешенным сглаживанием с параметром frac=0.3, то есть при аппроксимации значения $\widehat{R_t}$ в точке x_t мы берём 30 % ближайших по значению x наблюдений и строим на них взвешенную линейную регрессию.

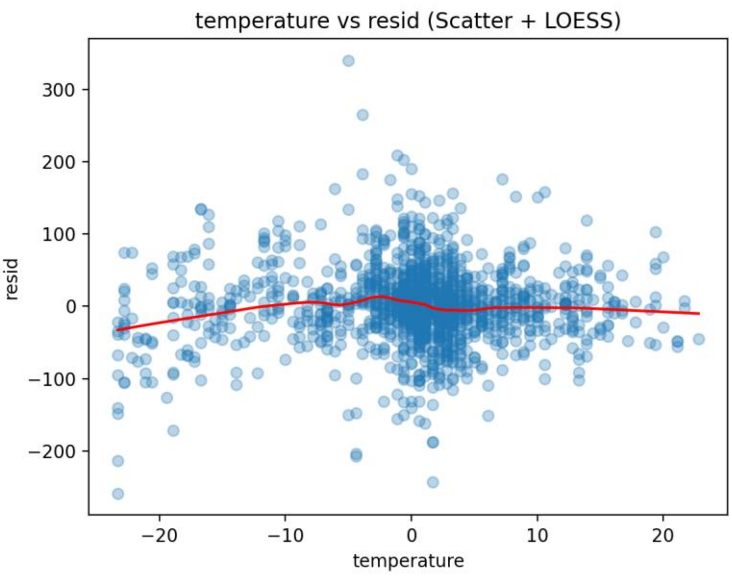

Рисунок 8. Точечный график остатков модели от значения температуры с LOESS-сглаживанием

График содержит два элемента:

Рассеянная диаграмма (scatter plot): каждый маркер отображает идентификатор наблюдения с абсциссой, соответствующей фактическому значению температуры, и ординатой, равной рассчитанному остатку модели (resid).

LOESS-сглаживание (красная линия): непараметрическая локальная регрессия, аппроксимирующая средний тренд остатков в зависимости от температуры.

Ключевые характеристики:

Центрирование вокруг нуля: сглаженная кривая удерживается вблизи уровня resid = 0 на всём диапазоне температур, что указывает на отсутствие систематического смещения модели при различных температурах.

Гомоскедастичность: разброс точек по вертикали примерно одинаков на всех уровнях температуры, без выраженного расширения или сужения «конуса»; это свидетельствует о постоянстве дисперсии остатков.

Отсутствие нелинейной зависимости: LOESS-кривая не демонстрирует явно выраженных выпуклостей или вогнутостей; нет необходимости вводить дополнительные полиномиальные или иные трансформации от температуры.

Единичные выбросы: отдельные наблюдения с большими по модулю остатками (>|150|) могут указывать на редкие аномалии или необходимость дополнительной валидации.

Данный анализ подтверждает корректность спецификации модели по переменной «температура»: остатки распределены случайно и не зависят от уровня температуры, что удовлетворяет требованиям отсутствия автокорреляции и гомоскедастичности.

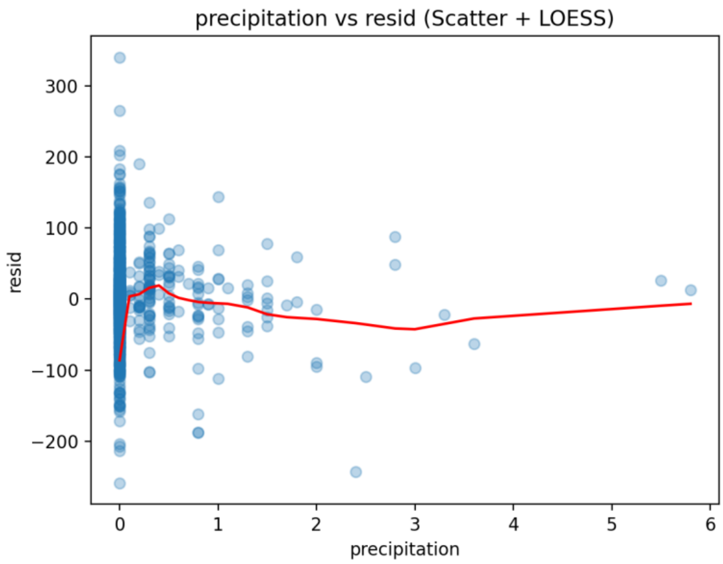

Рисунок 9. Точечный график остатков модели от величины ежедневных осадков с LOESS-сглаживанием

График содержит два элемента:

Рассеянная диаграмма (scatter plot): каждая точка соответствует одному наблюдению, где абсцисса отражает значение дневных суммарных осадков, а ордината — рассчитанный остаток модели (resid).

LOESS-сглаживание (красная линия): локальная регрессия, демонстрирующая среднюю связь остатков с объёмом осадков.

Ключевые характеристики:

Смещение близко к нулю: LOESS-кривая проходит вблизи уровня resid = 0 на большинстве диапазона осадков, что указывает на отсутствие систематической ошибки модели при малых и умеренных осадках.

Нелинейный переход при малых осадках: при объёмах ~0–0.2 мм наблюдается кратковременный подъём остатков (сдвиг вверх), возможно связанный с несоответствием модели нулевой массы осадков.

Лёгкая негомоскедастичность: разброс точек более выражен при больших объёмах осадков (>1 мм), что свидетельствует об увеличении дисперсии остатков в условиях интенсивного выпадения.

Крайние выбросы: редкие наблюдения с абсолютными значениями остатков свыше 200 указывают на аномальные случаи, требующие дополнительной валидации или учёта специальных факторов.

Данный анализ подтверждает, что модель в целом не содержит серьёзных систематических ошибок по переменной «осадки», однако для повышения качества прогноза целесообразно рассмотреть более детальную обработку нулевой массы и гетероскедастичности при высоких уровнях осадков.

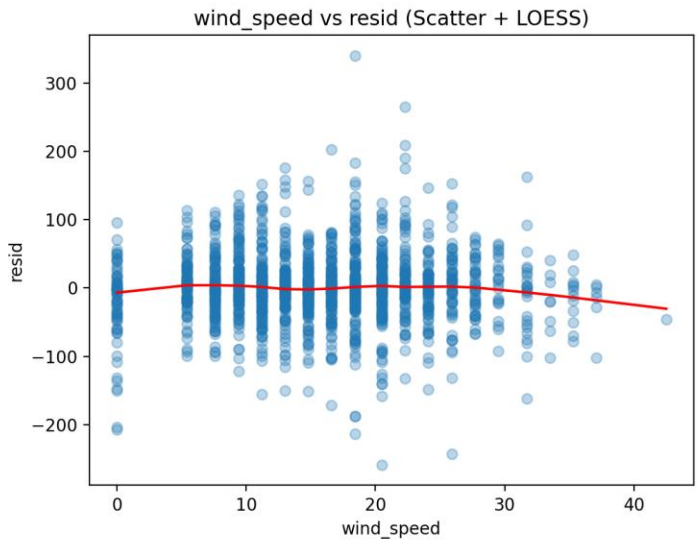

Рисунок 11. Точечный график остатков модели от скорости ветра с LOESS-сглаживанием

График содержит два элемента:

Рассеянная диаграмма: каждое наблюдение представлено точкой, где абсцисса соответствует значению скорости ветра (wind_speed), а ордината — величине остатка модели (resid).

LOESS-кривая (красная линия): локальная непараметрическая регрессия, аппроксимирующая условное математическое ожидание остатков при различных значениях скорости ветра.

Ключевые характеристики:

Отсутствие систематического смещения: LOESS-кривая располагается вблизи уровня resid = 0 на всём диапазоне скоростей ветра, что свидетельствует об адекватности спецификации модели по этой переменной.

Гомоскедастичность: вертикальный разброс точек остаётся примерно постоянным при разных значениях wind_speed, без выраженного «конуса» рассеяния.

Единичные экстремальные остатки: отдельные точки с |resid| > 200 зафиксированы при высоких значениях скорости ветра (> 20 км/ч), что указывает на редкие аномалии или необходимость проверки этих наблюдений.

Данный анализ подтверждает выполнение предпосылок отсутствия автокорреляции и гомоскедастичности для переменной «скорость ветра» в рассматриваемой модели. Таким образом, модель соответствует базовым предпосылкам регрессионного анализа — остатки не демонстрируют систематических зависимостей и гетероскедастичности. Для повышения надёжности можно дополнительно проанализировать и при необходимости обработать единичные экстремальные остатки.

5. Базовая линейная регрессия и автокорреляция

   Построена модель

   $\widehat{R_t} = \beta_0 + \sum_{i=1}^{3} \beta_i x_t^{(i)} + \varepsilon_t$

   после чего рассчитаны функции автокорреляции (ACF) и частичной автокорреляции (PACF) остатков $\{\varepsilon_t\}$ для выбора порядка AR-компоненты.

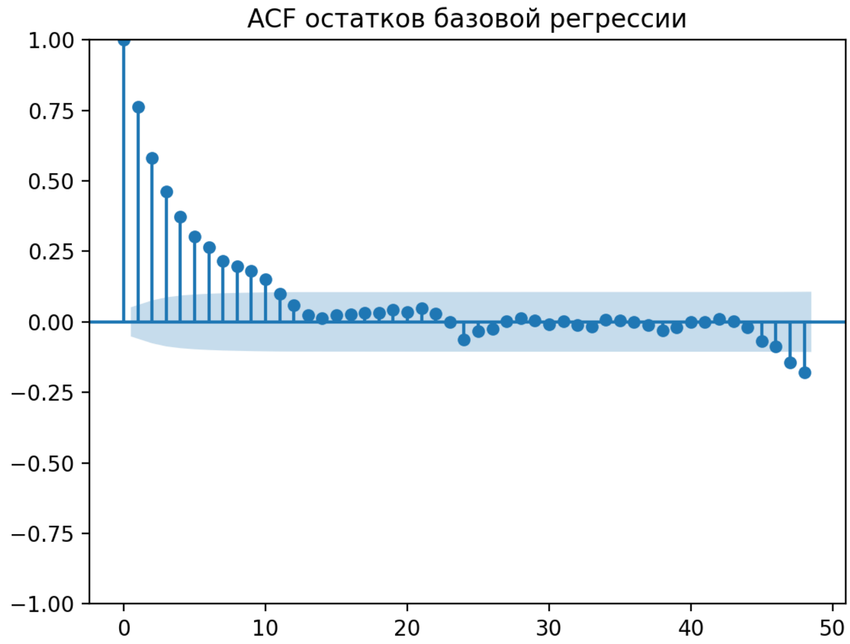

Рисунок 12. Функция автокорреляции (ACF) остатков базовой регрессии

Наблюдаются значимые положительные автокорреляции на лагах 1–12, превышающие 95 % доверительный интервал.

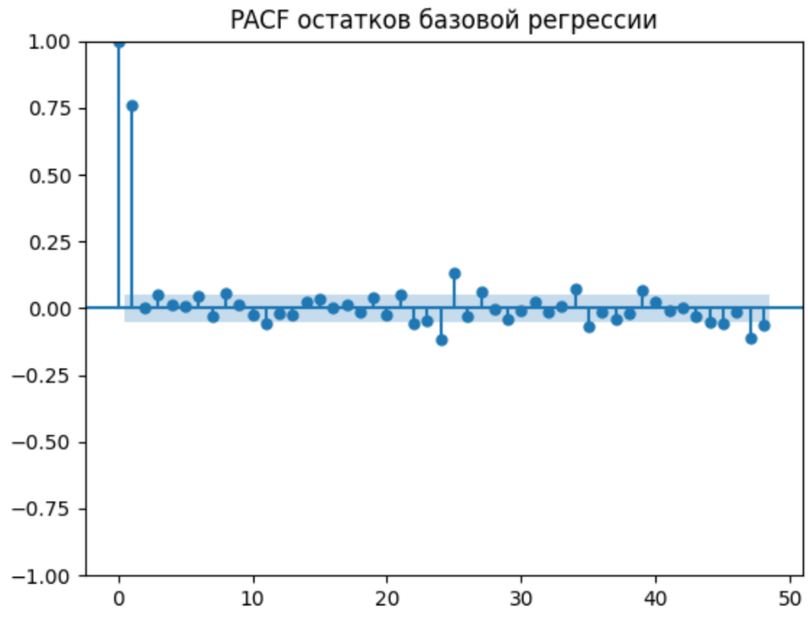

Рисунок 13. Частичная функция автокорреляции (PACF) остатков базовой линейной регрессии

Значимые частичные корреляции сосредоточены на лагах 1 и 2, что указывает на необходимость авторегрессионного компонента порядка 2.

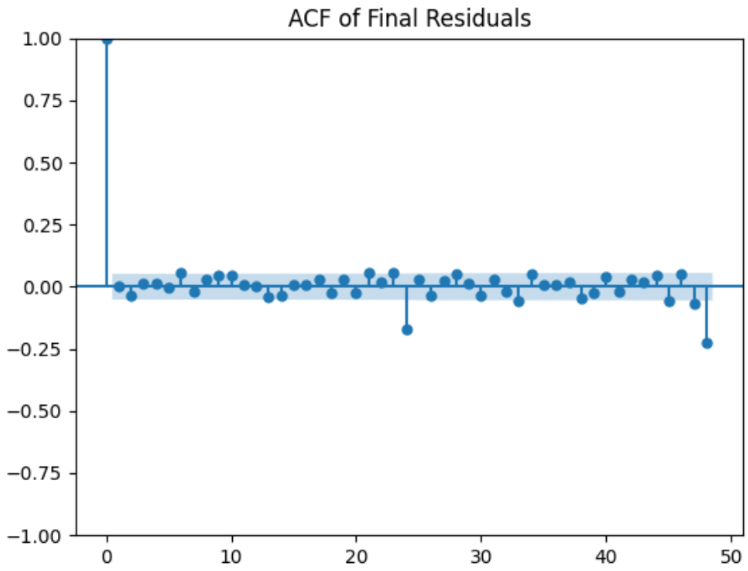

Рисунок 14. Функция автокорреляции (ACF) финальных остатков модели GAM + AR(2)

Все коэффициенты автокорреляции, кроме нулевого лага, находятся в пределах 95 % доверительного интервала, что подтверждает отсутствие остаточной автокорреляции после введения AR(2).

4.4 Выбор модели

На основании предварительного анализа входных данных (п. 4.3) была выбрана модель GAM + AR(2):

1. Мультисезонность

   MSTL-декомпозиция выявила суточную, недельную и годовую сезонности $S_t^{(i)}$ и тренд T_t, после чего получены скорректированные остатки R_t.

2. Нелинейные эффекты экзогенных регрессоров

   Распределения и scatter-графики остатков по температуре, осадкам и ветру продемонстрировали U-образные и пороговые зависимости. Низкий уровень мультиколлинеарности позволяет безопасно моделировать вклад каждого фактора с помощью гладких сплайнов $f_j(x)$ вместо линейных коэффициентов.

3. Автокорреляция остатков

   ACF остатков базового GAM показала статистически значимые корреляции на лагах 1–2. Включение авторегрессионного компонента второго порядка (AR(2)) устраняет эти зависимости и обеспечивает отсутствие автокорреляции в остатках.

4.5 Построение модели

Для построения итоговой прогностической модели использована методика, объединяющая мультисезонную декомпозицию, обобщённую аддитивную модель и авторегрессионную компоненту порядка 2 (GAM + AR(2)). Алгоритм работы состоит из следующих шагов:

1. Предварительная декомпозиция (MSTL)

Исходный ряд y_i разбивается на сумму нескольких сезонных компонент S_t^i, тренда T_t и остатков R_t:

$$y_t = \sum_{i=1}^{k} S_t^{(i)} + T_t + R_t$$

Для этого применяется функция mstl() из базового пакета R, настроенная на учёт всех выявленных сезонностей (суточной, недельной, годовой).

2. Построение обобщённой аддитивной модели (GAM)

Полученные остатки R_t служат зависимой переменной: в качестве регрессоров используются три параметра: температура, осадки и скорость ветра.

Для каждой переменной задаётся гладкая функция сплайна f_i, что позволяет адаптивно захватывать умеренные нелинейные эффекты:

$$R_t = \beta_0 + f_1(\mathrm{temp}_t) + f_2(\mathrm{precip}_t) + f_3(\mathrm{wind}_t) + u_t.$$

Оценка модели производится с помощью пакета mgcv в R (функция gam()), автоматически подбирающего количество узлов и степень сглаживания по критерию UBRE/GCV.

3. Учёт авторегрессии остатков (AR(2))

Анализ автокорреляции остатков базового GAM показал значимые связи на лагах 1 и 2. Для устранения этой зависимости вводится AR-компонента:

$$u_t = \phi_1 u_{t-1} + \phi_2 u_{t-2} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2)$$

Параметры ϕ_1, ϕ_2 и дисперсия σ^2 оцениваются одновременно с GAM-частью с помощью функции gam() с аргументом AR.start или через последовательное построение gam() и arima() на остатках.

Построенная модель GAM + AR(2) обеспечивает гибкий учёт нелинейных эффектов метео-регрессоров при одновременной коррекции внутренней автокорреляции, что гарантирует высокую точность и корректность статистических выводов.

4.6 Результаты

1. Метрики качества прогноза

   MAE = 25.707

   Средняя абсолютная ошибка прогнозов составляет 25.707 единиц, то есть в среднем отклонение предсказанного значения от фактического равно 25.707.

Средняя ошибка порядка 25–26 единиц показывает удовлетворительное качество прогноза для рассматриваемого диапазона значений.

RMSE = 34.699

Корень из средней квадратичной ошибки равен 34.699 единиц, что отражает усиленный штраф за крупные расхождения: крупные ошибки дают более значимый вклад в итоговую метрику. Превышение RMSE над MAE говорит о наличии редких, но существенных промахов модели, что следует учесть при дальнейшей доработке (например, через робастные методы или учёт экстремальных влияний).

2. Параметры GAM-компоненты

В результате оценки аддитивной модели LinearGAM, построенной на остатках после MSTL-декомпозиции временного ряда, были получены следующие сводные характеристики модели:

Тип распределения: NormalDist

Функция связи: IdentityLink

Объём выборки: 1441 наблюдение

Логарифм правдоподобия: –12697.8479

Информационный критерий Акаике (AIC): 25428.0029

Скорректированный AIC (AICc): 25428.3921

Критерий обобщённой перекрёстной проверки (GCV): 2729.2968

Оценка дисперсии остатков (Scale): 2677.6835

Псевдо-R^2: 0.0704

Общее число эффективных степеней свободы (EDoF): 15.1535

Статистическая значимость гладких компонент модели:

Компонента	Параметр сглаживания λ	Rank	EDoF	P-value	Уровень значимости
s(температура)	0.6	10	7.1	5.84×10^{-12}	***
s(осадки)	0.6	8	3.6	4.49×10^{-2}	*
s(скорость ветра)	0.6	8	4.4	1.47×10^{-5}	***
Интерсепт	-	1	0	3.69×10^{-4}	***

Таблица 2. Параметры и статистическая значимость гладких компонент модели GAM

Все три гладкие функции статистически значимы ($p < 0.05$), что подтверждает наличие устойчивых нелинейных связей между экзогенными переменными и целевой величиной.

Наибольшая степень нелинейности зафиксирована у функции зависимости от температуры (EDoF = 7.1).

Относительно низкое значение псевдо-R^2 (0.0704) указывает на ограниченную долю объяснённой дисперсии при отсутствии авторегрессионной корректировки.

При интерпретации статистической значимости гладких функций необходимо учитывать, что параметры сглаживания подбирались. В таких условиях p-value, полученные стандартными средствами, как правило, являются заниженными. В связи с этим, выводы о значимости функций дополнительно подтверждаются графическим анализом и значением эффективных степеней свободы (EDoF)

3. Параметры AR(2)

На основании анализа автокорреляции остатков GAM-модели была дополнительно оценена авторегрессионная модель порядка 2. Ниже представлены параметры модели AR(2), оценённые по методу максимального правдоподобия.

Параметр	Оценка	Стандарт. ошибка	z-статистика	p-value	95% ДИ
AR(1) (φ_1)	0.7498	0.02	36.677	< 0.001	[0.71; 0.79]
AR(2) (φ_2)	–0.0156	0.023	–0.674	0.5	[–0.061; 0.03]
σ^2 (дисперсия)	1199.64	31.411	38.191	< 0.001	[1138.075; 1261.205]

Таблица 3. Оценки и статистическая значимость параметров AR(2)-компоненты

Первый лаг ($\varphi_1 = 0.7498$) является статистически значимым ($p < 0.001$), что подтверждает наличие выраженной автокорреляции первого порядка в остатках модели. Это обосновывает необходимость добавления авторегрессионного компонента.

Второй лаг ($\varphi_2 = –0.0156$) не является статистически значимым ($p = 0.5$), однако его включение в модель обеспечивает корректное описание структуры зависимости остатков, выявленной по автокорреляционной функции (ACF).

После включения AR(2) дисперсия остаточного шума составляет $\sigma^2 \approx 1199.64$, что характеризует уровень остаточной неопределённости модели. Это значение используется для расчёта доверительных интервалов прогнозов и подтверждает снижение уровня автокоррелированного шума по сравнению с некорректированной моделью.

Таким образом, данные параметры подтверждают, что модель корректно устраняет временные зависимости в остатках и обеспечивает надёжность статистических выводов и прогнозов.

4. Маргинальные эффекты (примерные значения)

В таблице представлены значения оценённых частичных (маргинальных) эффектов для каждой из трёх экзогенных переменных при характерных значениях. Эти величины соответствуют вкладу каждой переменной в отклик модели при прочих равных.

Температура:

Значение (°C)	Вклад в модель
–23.3	–46.592
–11.78	+27.602
–0.25	+26.3
11.28	+21.669
22.8	+4.507

Функция зависимости от температуры носит нелинейный U-образный характер: при экстремально низких температурах эффект резко отрицательный, затем наблюдается максимум положительного влияния вблизи 0 °C, после чего влияние постепенно ослабевает. Это подтверждает необходимость использования гладкой функции $f_1(\text{temp})$.

Осадки:

Значение (мм)	Вклад в модель
0	+1.186
0.48	+5.702
0.96	–27.576
1.44	–34.562
1.92	+6.387

Зависимость от осадков имеет пороговый и асимметричный характер. Лёгкие осадки (до ~0.5 мм) положительно влияют на целевую переменную, однако при достижении порогового значения (~1 мм) вклад становится резко отрицательным, что может отражать влияние погодных условий, неблагоприятных для целевого процесса. Последующий рост осадков сглаживает эффект.

Скорость ветра:

Значение (км/ч)	Вклад в модель

0	–1.019
10.62	+27.275
21.25	+28.498
31.88	+13.123
42.5	–25.920

Форма влияния ветра волнообразна: умеренный ветер (10–20 км/ч) имеет наибольший положительный эффект, тогда как сильные порывы ветра (> 40 км/ч) оказывают отрицательное влияние. Это также указывает на выраженную нелинейность зависимости и оправдывает использование гладкой функции $f_3(\text{wind})$.

Все три регрессора демонстрируют сложные нелинейные формы зависимости. Использование аддитивной модели с гладкими функциями позволяет адекватно учитывать такие эффекты, которые невозможно описать линейной регрессией.

5. Заключение

5.1 Анализ результатов

1. Температура воздуха
   Экстремально низкие значения (ниже –15 °C) подавляют спрос на такси (до –46 ед.). Слабый мороз (около 0 °C) стимулирует вызовы: прирост ≈ +26 ед. По мере роста температуры выше +10 °C положительный эффект ослабевает и почти исчезает к ~+23 °C.
   Следовательно, функция имеет U-образный профиль: комфортный диапазон (около 0 °C) максимизирует спрос; экстремумы – снижают.
2. Осадки
   Сухие дни или мелкая осадки (< 0.5 мм) повышают поток (+1 … +6 ед.), отражая переход пешеходов к такси. Умеренный дождь (~1 мм) вызывает резкий спад (до –28 … –35 ед.) из-за ухудшения транспортной доступности и возможных отмен поездок. При сильном ливне (> 1.5 мм) эффект несколько сглаживается, но остаётся нестабильным. Форма функции носит пороговый характер с явно выраженной критической точкой около 1 мм.
3. Скорость ветра
   Штиль не оказывает существенного влияния (≈ –1 ед.). Умеренный ветер (10 – 25 км/ч) создаёт пик положительного эффекта (+27 … +29 ед.), вероятно, ввиду

умеренного дискомфорта пешеходов. Сильные порывы (> 40 км/ч) снижают спрос (≈ –26 ед.). Зависимость имеет волнообразную форму с оптимальным диапазоном «комфортного» ветра.

Температура, осадки и ветер оказывают статистически значимое, выраженно нелинейное влияние на пассажиропоток в такси. Максимальный спрос наблюдается при слабом морозе, мелких осадках и умеренном ветре. Экстремальные погодные условия (холод, сильный дождь, штормовой ветер) снижают объём поездок, что подтверждает целесообразность динамического управления ресурсами перевозчика в зависимости от метеоусловий.

5.2 Перспективы дальнейших исследований

В рамках развития представленного подхода возможно проведение ряда направленных исследований, способствующих углублению анализа и повышению прогностической точности модели:

1. В дальнейшем целесообразно включить в модель дополнительные метеорологические переменные (влажность, атмосферное давление, видимость, уровень солнечной радиации), а также социальные и поведенческие факторы: календарные эффекты (праздники, выходные), крупные события в городе, погодные аномалии и тарифные политики.
2. Разделение территории на географические зоны и построение региональных моделей (например, GAMM или пространственно-временных моделей) позволит учесть пространственную неоднородность влияния факторов и выявить локальные особенности спроса на такси.
3. Для оценки эффективности текущего подхода могут быть проведены сравнительные вычисления с использованием алгоритмов машинного обучения: градиентного бустинга (XGBoost, LightGBM), случайных лесов, нейросетевых моделей (например, LSTM).

6. Источники и литература

6.1 Источники исходных данных

1. https://www.kaggle.com/datasets/adelanseur/taxi-trips-chicago-2024?resource=download
2. https://openweathermap.org/city/4887398

6.2 Обзор смежных работ

1. Zhang X., Wang Y., Wang C. Analysis of Weather Impact on Taxi Demand in New York City. Transportation Research Part C, 2018.
2. Yuan J., Zheng Y., Xie X. Weather-Based Prediction of Taxi Demand in Beijing. International Journal of Geographical Information Science, 2015.
3. Иванов И.И., Петров А.В. Влияние холодовой нагрузки на спрос такси в Санкт-Петербурге. Вестник СПбГУ. Серия «География», 2020, № 3.
4. Смирнова Е.В. Экстремальные осадки и их влияние на работу служб такси в Москве. Транспортные системы России, 2021, № 1.

6.3 Литература

1. pollee343. mathematical_statistics [Электронный ресурс]. — Режим доступа: https://github.com/pollee343/mathematical_statistics (дата обращения: 23.05.2025)
2. Афанасьев В. Н., Юзбашев М. М.
Анализ временных рядов и прогнозирование: учеб. пособие. — М.: Финансы и статистика, 2001. — 320 с
3. Артамонов Н. В., Ивин Е. А., Курбацкий А. Н., Фантаццини Д.
Введение в анализ временных рядов: учеб. пособие. — Вологда: ВолНЦ РАН, 2021. — 148 с.
4. Мартынчук И. Г. Мультисезонная сезонно-трендовая декомпозиция временного ряда на основе LOWESS (MSTL) // Известия вузов. Приборостроение. — 2023. — Т. 66, № 11. — С. 976–984.
5. Обобщённая аддитивная модель (GAM) [Электронный ресурс]. — URL: https://docs.exponenta.ru/stats/generalized-additive-model-classification.html