Санкт-Петербургский государственный университет

Кильдякова Юлия Александровна

Выпускная квалификационная работа бакалавра

Прогнозирование иерархических временных рядов

Направление 02.03.02
«Фундаментальная информатика и информационные технологии»
ООП СВ.5003.2018: «Программирование и информационные технологии»

Научный руководитель:
Доктор физ.-мат. наук, профессор
кафедры математической теории игр
и статистических решений:
Парилина Елена Михайловна

Санкт-Петербург
2022

Содержание

Введение

Прогнозирование — это распространенная задача науки о данных, которая помогает организациям в планировании распределения ресурсов, постановке целей и обнаружении аномалий.

Временной ряд — это упорядоченная во времени последовательность значений какого-либо показателя. Существует множество методов и подходов к прогнозированию временных рядов. Главное отличие и преимущество прогнозирования на основе применения технологий машинного обучения перед традиционными статистическими методами в том, что такие решения постоянно сопоставляют сделанный прогноз с фактическими данными – именно это позволяет модели обучаться и повышать точность прогноза в будущем.

Временные ряды часто можно естественным образом дезагрегировать по различным интересующим атрибутам. Так иерархические временные ряды часто возникают из-за географического деления. Например, общие продажи товаров могут быть дезагрегированы по странам, затем в каждой стране по штатам, в каждом штате по региону и так далее до уровня торговых точек.

Обычно на основе дезагрегированных временных рядов составляются дезагрегированные прогнозы, и требуется, чтобы прогнозы складывались так же, как и данные, т. е. удовлетворяли структуре иерархии. Например, прогнозы региональных продаж должны складываться, чтобы давать прогнозы продаж по штатам, которые, в свою очередь, должны суммироваться, чтобы давать прогноз продаж по стране.

В данной выпускной квалификационной работе будут рассмотрены и применены основные подходы к согласованию прогнозов иерархических временных рядов для данных туризма Российской Федерации с географической иерархией.

Постановка задачи

Целью работы является исследование прогнозных моделей и подходов к согласованию прогнозов иерархических временных рядов на данных о численности граждан Российской Федерации, размещенных в коллективных средствах размещения для различных субъектов РФ. Данные получены с официального сайта федерального агентства по туризму.

Задачами работы являются:

- предобработка полученных данных,
- обучение моделей,
- анализ точности моделей,
- определение наиболее эффективных моделей и подходов, их комбинаций,
- получение прогнозов численности граждан РФ, размещенных в коллективных средствах размещения, для каждого уровня географической иерархии на два последующих временных промежутка – года или квартала, в зависимости от разбиения входных данных.

Обзор литературы

Статистика — очень молодая наука. Прогресс в статистике, анализе данных и временных рядов всегда сильно зависел от того, когда, где и как данные были доступны и в каком количестве. Появление анализа временных рядов как дисциплины связано не только с развитием теории вероятностей, но в равной степени и с развитием стабильных национальных государств, где ведение учета впервые стало достижимой и интересной целью. Одним из ориентиров для начала анализа временных рядов как дисциплины является применение авторегрессионных моделей к реальным данным. Этого не было до 1920-х годов.

Существует три основных подхода к построению прогнозов: методы на основе регрессии, методы эвристического сглаживания, и общие модели временных рядов. Все эти подходы полезны и подробно описываются в [1].

[2] представляет собой практическое руководство, охватывающее инновации в анализе данных временных рядов и примеры использования из реального мира.

В данной работе обсуждается прогнозирование больших коллекций временных рядов, которые должны каким-то образом суммироваться. Проблема заключается в том, что нам нужны прогнозы, согласованные по всей структуре агрегирования. То есть требуется, чтобы прогнозы складывались в соответствии со структурой агрегирования набора временных рядов. В [3] дается введение в методы получения согласованных прогнозов для иерархических временных рядов.

[4] представляет хорошее введение в нисходящие подходы. Цель исследования состояла в том, чтобы выявить ситуации, в которых целесообразно составлять дезагрегированные прогнозы, и, если выгодно, какой метод дезагрегации использовать.

Методы оптимального согласования прогнозов были разработаны в серии статей. В [5] описывается метод, основанный на независимом прогнозировании всех рядов на всех уровнях иерархии с последующим

использованием регрессионной модели для оптимального объединения и согласования этих прогнозов.

В [6] рассматривается эффективность пяти подходов к иерархическому прогнозированию: два варианта нисходящего подхода, восходящий метод, на тот момент недавно предложенный нисходящий подход, в котором прогнозы верхнего уровня дезагрегируются в соответствии с прогнозируемыми пропорциями рядов более низкого уровня, и также недавно предложенный подход оптимального согласования.

В [7] показано, что метод наименьших квадратов для согласования прогнозов иерархических временных рядов может быть распространен на гораздо более общие наборы временных рядов с ограничениями агрегирования.

В [8] предлагается новый подход к согласованию прогнозов, который включает информацию из полной ковариационной матрицы ошибок прогнозов для получения набора согласованных прогнозов.

[9] расширяет подход согласования для работы с временными иерархиями.

Если говорить о развитии технологий прогнозирования временных рядов, в 2017 году Facebook была выпущена модель Prophet. В [10] описывается модульная регрессионная модель с интерпретируемыми параметрами, которая лежит в основе библиотеки.

Анализ временных рядов и прогнозирование еще не достигли своего золотого периода, и на сегодняшний день в анализе временных рядов по-прежнему преобладают традиционные статистические методы, а также более простые методы машинного обучения. Нас все еще ждет большой скачок вперед в предсказании будущего.

Глава 1. Подходы к согласованию прогнозов

1.1 Структура иерархического временного ряда

На рис. 1 показан пример двухуровневой иерархической структуры временного ряда. Наблюдение на верхнем уровне иерархии для временного промежутка с номером t обозначается за y_t. Наблюдения на последующих уровнях обозначаются как $y_{j,t}$ для t-ого наблюдения для узла j. Например, для узла A наблюдение с номером t обозначается за $y_{A,t}$.

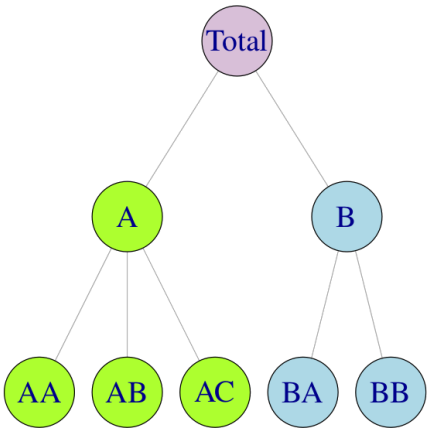

Рис. 1: Древовидная диаграмма двухуровневого временного ряда

Для любого t наблюдения на нижнем уровне иерархии будут суммироваться в наблюдения рядов уровнями выше. Например:

$$y_t = y_{AA,t} + y_{AB,t} + y_{AC,t} + y_{BA,t} + y_{BB,t}, \quad (1)$$

$$y_{A,t} = y_{AA,t} + y_{AB,t} + y_{AC,t}, \quad (2)$$

$$y_{B,t} = y_{BA,t} + y_{BB,t}. \quad (3)$$

Уравнения (1) – (3) также можно представить с помощью матричной записи. Строится суммирующая матрица S размера $m \times n$, где n – общее количество узлов, m – количество узлов нижнего уровня, которая определяет способ агрегирования рядов нижнего уровня.

$$\begin{bmatrix} y_t \\ y_{A,t} \\ y_{B,t} \\ y_{AA,t} \\ y_{AB,t} \\ y_{AC,t} \\ y_{BA,t} \\ y_{BB,t} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_{AA,t} \\ y_{AB,t} \\ y_{AC,t} \\ y_{BA,t} \\ y_{BB,t} \end{bmatrix}$$

или

$$y_t = Sb_t$$

где y_t – n-мерный вектор всех наблюдений в момент времени t, S – суммирующая матрица, b_t – m-мерный вектор всех наблюдений нижнего уровня иерархии в момент t.

1.2 Восходящий подход

Простым методом создания согласованных прогнозов является восходящий подход. Этот подход включает сначала создание прогнозов для каждого ряда на нижнем уровне, а затем их суммирование для получения прогнозов для всех рядов в структуре.

Например, для иерархии, представленной на рис. 1, сначала строятся прогнозы с шагом прогнозирования h для каждого временного ряда нижнего уровня:

$$\hat{y}_{AA,h}, \hat{y}_{AB,h}, \hat{y}_{AC,h}, \hat{y}_{BA,h}, \hat{y}_{BB,h}.$$

Путем их суммирования получаются прогнозы для остальных временных рядов иерархии:

$$\tilde{y}_h = \hat{y}_{AA,h} + \hat{y}_{AB,h} + \hat{y}_{AC,h} + \hat{y}_{BA,h} + \hat{y}_{BB,h},$$

$$\tilde{y}_{A,h} = \hat{y}_{AA,h} + \hat{y}_{AB,h} + \hat{y}_{AC,h},$$

$$\tilde{y}_{B,h} = \hat{y}_{BA,h} + \hat{y}_{BB,h}$$

где $\tilde{y}_{j,h}$ – согласованные прогнозы для узла j с шагом прогнозирования h.

Так же, как и ранее, можно использовать суммирующую матрицу:

$$\begin{bmatrix} \tilde{y}_h \\ \tilde{y}_{A,h} \\ \tilde{y}_{B,h} \\ \tilde{y}_{AA,h} \\ \tilde{y}_{AB,h} \\ \tilde{y}_{AC,h} \\ \tilde{y}_{BA,h} \\ \tilde{y}_{BB,h} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{y}_{AA,h} \\ \hat{y}_{AB,h} \\ \hat{y}_{AC,h} \\ \hat{y}_{BA,h} \\ \hat{y}_{BB,h} \end{bmatrix}$$

или

$$\tilde{y}_h = S\hat{b}_h$$

где $\tilde{y}_h$ – m-мерный вектор согласованных прогнозов с шагом прогнозирования h, $\hat{b}_h$ – m-мерный вектор прогнозов с шагом h на нижнем уровне иерархии.

Преимущество этого подхода в том, что прогнозы создаются на нижнем уровне структуры, поэтому информация не теряется из-за агрегирования. С другой стороны, данные нижнего уровня могут быть довольно зашумленными и более сложными для моделирования и прогнозирования.

1.3 Нисходящий подход

Нисходящий подход включает прогнозирование верхнего уровня иерархии с последующим разбиением прогноза на более детализированные ряды.

Положим $p_1, \ldots, p_m$ – множество пропорций дезагрегации, которые определяют, как должны распределяться прогнозы главного временного ряда для получения прогнозов для каждого ряда на нижнем уровне структуры.

Например, для структуры с рис. 1, используя пропорции $p_1, \ldots, p_5$ получается:

$\tilde{y}_{AA,h} = p_1\hat{y}_h, \tilde{y}_{AB,h} = p_2\hat{y}_h, \tilde{y}_{AC,h} = p_3\hat{y}_h, \tilde{y}_{BA,h} = p_4\hat{y}_h$ и $\tilde{y}_{BB,h} = p_5\hat{y}_h$.

Используя матричную запись:

$$\tilde{b}_h = p\hat{y}_h$$

где $p = (p_1, \ldots, p_m)^T$ – m-мерный вектор пропорций.

После создания прогнозов нижнего уровня на h шагов вперед они суммируются для создания согласованных прогнозов для остальных рядов иерархии. В общем виде, для определенного набора пропорций, нисходящие подходы могут быть представлены как:

$$\tilde{y}_h = Sp\hat{y}_h.$$

Два наиболее распространенных нисходящих подхода определяют пропорции дезагрегации на основе исторических пропорций данных. Они показали хорошие результаты в исследовании Gross & Sohl [4].

Средние исторические пропорции:

$$p_j = \frac{1}{T}\sum_{t=1}^{T}\frac{y_{j,t}}{y_t}, \qquad j = 1,\dots,m.$$

Каждая пропорция p_j отражает среднее значение исторических пропорций ряда нижнего уровня $y_{j,t}$ по отношению к главному ряду y_t за период $t = 1,\dots,T$.

Пропорции исторических средних:

$$p_j = \sum_{t=1}^{T}\frac{y_{j,t}}{T}\Big/\sum_{t=1}^{T}\frac{y_t}{T}, \qquad j = 1,\dots,m.$$

Каждая пропорция p_j отражает среднее историческое значение ряда нижнего уровня $y_{j,t}$ по отношению к среднему значению главного ряда y_t.

Удобством таких нисходящих подходов является их простота. Нужно моделировать и генерировать только прогнозы для наиболее агрегированных рядов верхнего уровня. Недостатком является потеря информации из-за высокого уровня агрегации. Используя такие нисходящие подходы, нет возможности зафиксировать и использовать характеристики отдельных рядов.

1.4 Срединный подход

Срединный подход представляет собой комбинацию двух описанных выше методов и может использоваться только для строго иерархических

временных рядов. В этом подходе выбирается средний уровень и напрямую прогнозируется. Затем для всех уровней выше выбранного используется восходящий подход, а для уровней ниже среднего используется нисходящий подход.

Поскольку это компромисс между двумя разными подходами, итоговые прогнозы не теряют такой объем информации, как в случае нисходящего подхода, и время вычислений не увеличивается, как в случае восходящего подхода.

1.5 Подход оптимального согласования

Три описанных выше подхода сосредоточены на прогнозировании временных рядов на одном уровне, а затем на их использовании для вывода остальных уровней. В отличие от них, в оптимальном методе согласования прогнозируется каждый из уровней, используя всю информацию и взаимосвязи, которые может предложить данная иерархия.

Предположим, что мы прогнозируем все ряды независимо друг от друга, игнорируя ограничения агрегации. Такие прогнозы называются базовыми и обозначаются как $\hat{y}_h$, где h – шаг прогноза.

Тогда все подходы к прогнозированию для иерархических структур можно представить как:

$$\tilde{y}_h = SG\hat{y}_h$$

где G – матрица, преобразующая базовые прогнозы в прогнозы нижнего уровня иерархии, которые с помощью суммирующей матрицы S агрегируются, создавая набор согласованных прогнозов $\tilde{y}_h$.

Например, для нисходящих подходов матрица G будет выглядеть следующим образом:

$$G = \begin{bmatrix} p_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ p_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ p_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ p_4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ p_5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Можно записать выражение (15) в ином виде:

$$\tilde{y}_h = P\hat{y}_h$$

где P является матрицей согласования. Она принимает базовые прогнозы $\hat{y}_h$ и преобразует их в согласованные прогнозы $\tilde{y}_h$.

В описанных в предыдущих параграфах методах не проводилось реального согласования, поскольку эти методы основывались на прогнозах с одного уровня структуры, которые либо агрегировались, либо дезагрегировались для получения прогнозов на всех других уровнях иерархии. Однако, в целом, возможно использовать другие матрицы G, а затем и P, которые бы объединяли и согласовывали все базовые прогнозы для получения согласованных прогнозов.

Оптимальное согласование произойдет, если получится найти матрицу G, минимизирующую ошибку прогноза множества согласованных прогнозов.

В первую очередь следует удостовериться, что прогнозы несмещенные. Если базовые прогнозы $\hat{y}_h$ несмещенные, то согласованные прогнозы $\tilde{y}_h$ также будут несмещенными при условии $SGS = S$. Это накладывает ограничение на матрицу S.

Далее необходимо найти ошибку в прогнозах. Ковариационная матрица ошибок согласованных прогнозов с шагом h определяется выражением

$$V_h = Var[y_{T+h} - \tilde{y}_h] = SGW_hG^TS^T$$

где $W_h = Var[y_{T+h} - \hat{y}_h]$ – ковариационная матрица соответствующих ошибок базовых прогнозов.

Задача состоит в том, чтобы найти матрицу G, которая минимизирует дисперсии ошибок согласованных прогнозов. Они находятся на диагонали матрицы V_h, поэтому сумма всех дисперсий ошибок определяется следом матрицы V_h. Матрица G, которая минимизирует след V_h и удовлетворяет условию $SGS = S$, выражается как:

$$G = (S^TW_h^{-1}S)^{-1}S^TW_h^{-1}.$$

Таким образом оптимально согласованные прогнозы имеют вид:

$$\tilde{y}_h = S(S^T W_h^{-1} S)^{-1} S^T W_h^{-1} \hat{y}_h.$$

Для использования на практике необходимо оценить W_h, дисперсию ошибки базового прогноза с шагом h.

Положим $W_h = k_h diag(\widehat{W}_1)$ для всех h, где константа пропорциональности $k_h > 0$,

$$\widehat{W}_1 = \frac{1}{T}\sum_{t=1}^{T} e_t \, e_t^T,$$

e_t – n-мерный вектор остатков моделей, сгенерировавших базовые прогнозы, сложенные в том же порядке, что и данные.

Эта оценка масштабирует базовые прогнозы с использованием дисперсии остатков, и поэтому она называется оценкой взвешенных наименьших квадратов.

Глава 2. Используемые прогнозные модели

2.1 ARIMA

Если объединить дифференцирование, модель авторегрессии и модель скользящего среднего, будет получена несезонная модель ARIMA.

Дифференцирование временного ряда — вычисление разницы между последовательными наблюдениями. Дифференцирование является одним из способов сделать нестационарный временной ряд стационарным. Данная операция может применяться последовательно несколько раз до тех пор, пока ряд не будет стационарным.

В модели авторегрессии прогнозируется интересующая переменная, с использованием линейной комбинации прошлых значений переменной.

Таким образом авторегрессионная модель порядка p может быть записана как:

$$y_t = c + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots + \varphi_p y_{t-p} + \varepsilon_t, \quad (4)$$

где ε_t – белый шум. Модель, представленная уравнением (4), обозначается как AR(p).

Вместо того, чтобы использовать прошлые значения переменной в регрессии, модель скользящего среднего использует прошлые ошибки прогноза в модели, подобной регрессии:

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}, \quad (5)$$

где ε_t – белый шум. Модель (5) обозначается как MA(q) и называется моделью скользящего среднего порядка q. Каждое значение y_t можно рассматривать как взвешенное скользящее среднее ошибок прогноза.

ARIMA — это аббревиатура от Auto-Regressive Integrated Moving Average (в данном контексте «интеграция» — противоположность дифференцированию). Полная модель может быть записана как:

$$y'_t = c + \varphi_1 y'_{t-1} + \ldots + \varphi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t, \quad (6)$$

где y'_t – продифференцированный временной ряд (может быть продифференцирован больше одного раза).

Выражение (6) называется ARIMA(p, d, q), где:

p – порядок авторегресссионной части,

d – степень дифференцирования,

q – порядок части скользящего среднего.

2.2 SARIMAX

В квартальных данных имеется зависимость количества туристов от квартала года, то есть сезона. Поэтому для них возможно использовать сезонные модели.

Модели ARIMA также способны моделировать широкий диапазон сезонных данных. Сезонная модель ARIMA формируется путем включения дополнительных сезонных условий в модель, которая рассматривалась в предыдущем параграфе. Это может быть записано следующим образом:

$$\text{ARIMA}\ (p, d, q)\ (P, D, Q)_m$$

где (p, d, q) – несезонная часть модели, (P, D, Q) – сезонная часть модели, m – количество наблюдений за год.

В модели ARIMAX, по сравнению с классической моделью ARIMA, учитывается ковариант, отражающий влияние внешних эффектов на значение ряда в момент времени t:

$$y'_t = \beta x_t + \varphi_1 y'_{t-1} + \ldots + \varphi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

где x_t – ковариант в момент времени t [11].

SARIMAX (Seasonal Auto-Regressive Integrated Moving Average with eXogenous factors) — это расширенная версия модели ARIMA. Она включает в себя сезонную компоненту, а также может учитывать экзогенные факторы (таким же образом, как в модели ARIMAX) [12].

2.3 Prophet

Prophet является библиотекой для прогнозирования временных рядов, разработанной в 2017 году Facebook [13].

Внутри библиотеки заложена аддитивная модель, выглядящая следующим образом:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t.$$

где $g(t)$ – компонента тренда. Является кусочно-линейной или логистической функцией. Логистическая функция вида $g(t) = C/(1 + exp(-k(t - b)))$ позволяет моделировать рост с насыщением, когда при увеличении показателя снижается темп его роста. Также с помощью Prophet возможен автоматический выбор точек изменения тренда или задание их вручную.

В представленной модели $s(t)$ – сезонная компонента. У Prophet имеется большее количество инструментов для обработки сезонных данных, чем у SARIMA, что позволяет анализировать временные ряды с различной сезонностью.

Слагаемое $h(t)$ – компонента, отвечающая за аномалии. Такие временные промежутки так же, как и точки изменения тренда, возможно задать вручную.

Слагаемое ε_t – ошибка, которая содержит информацию, не учтенную моделью.

Подробнее про алгоритмы можно прочитать в публикации Sean J. Taylor, Benjamin Letham [10].

Глава 3. Исходные данные

3.1 Описание данных

Данные получены с официального сайта федерального агентства по туризму [14]. В данных представлена информация о численности граждан Российской Федерации, размещенных в коллективных средствах размещения (гостиницах, хостелах и т. п.), для различных субъектов Российской Федерации. Представлены такие субъекты как: федеральные округа, области, города федерального значения, автономные округа, республики, края. Измерения проведены с годовой и квартальной частотой (начиная с 2016 года). Фрагмент данных представлен на рис. 2.

Субъект РФ	ОКАТО	2013 г. (досчет)	2014 г. (досчет)	2015 г. (досчет)	I кв 2016	II кв 2016	III кв 2016	IV кв 2016	2016 г.	2016 г. (досчет)
Российская Федерация	643	32 560 580	33 798 523	43 656 898	8 553 416	10 429 426	13 369 019	10 655 687	43 007 548	48 338 704
Центральный федеральный округ	030	9 418 269	9 660 925	11 921 376	2 517 848	2 838 212	3 201 156	3 187 928	11 745 144	15 229 091
Белгородская область	14000000000	156 801	171 668	151 471	37 568	37 416	55 344	32 664	162 992	233 033
Брянская область	15000000000	113 056	125 824	139 264	13 415	37 204	26 594	66 221	143 434	170 861
Владимирская область	17000000000	357 276	344 029	535 697	66 342	138 676	160 227	127 593	492 838	573 172

Рис. 2: Фрагмент данных

Краткосрочное размещение туристов является одной из самых важных сфер в индустрии туризма. Показатель численности граждан, размещенных в коллективных средствах размещения, может применяться с целью определения туристического потока в разрезе субъектов Российской Федерации. А прогнозирование туристического потока в свою очередь позволяет субъектам РФ планировать распределение бюджета и развитие туристической инфраструктуры.

3.2 Предобработка данных

Данные представлены в формате .xlsx. Для работы моделей необходимо конвертировать данные в формат .csv, а также изменить их внутреннюю структуру. Предобработка данных производилась с использованием библиотеки pandas [15] языка программирования Python.

В файле разные уровни географической иерархии обозначаются различными отступами. В данных имеется более 3 уровней. Так, например,

для Архангельской области подсчеты численности проведены отдельно для Ненецкого автономного округа и остальных территорий области (рис. 3).

22	Северо-Западный федеральный округ	031	4 051 915	4 209 131	5 415 135
23	Республика Карелия	86000000000	227 849	227 977	363 728
24	Республика Коми	87000000000	197 646	178 117	267 678
25	Архангельская область	11000000000	272 495	267 464	345 218
26	Ненецкий автономный округ (Архангельская область)	11100000000	17 943	17 246	20 671
27	Архангельская область (кроме Ненецкого автономного округа)	11001000000	254 552	250 218	324 547

Рис. 3: Четвертый уровень географической иерархии

Так как четвертый уровень встречается крайне редко, было принято решение ограничиться 3 уровнями иерархии, а именно: РФ, округа и области. Соответственно, такие строки 4-го уровня удаляются из данных.

Также удаляются строки федеральных округов и Российской Федерации в целом. Таким образом в датасете остаются только строки с данными по областям, и каждой такой строке присваивается значение федерального округа, к которому она принадлежит. Это необходимо для восстановления иерархии после приведения данных к виду, принимаемому на вход моделями.

Для входных данных с разбиением по годам используются столбцы с пометкой досчёт. Такие ячейки имеют значение больше, чем подсчет суммы всех кварталов, т. к. являются пересчетом на основании дополнительных данных.

Для входных данных с разбиением по кварталам используются столбцы с квартальными значениями.

После всех перечисленных выше преобразований, финальным шагом предобработки данных является приведение таблицы к виду, в котором каждому измерению будет соответствовать одна строка. Данные действия выполняются с помощью операции unpivot. Фрагменты преобразованных данным представлены на рис. 4–5.

	Область	Округ	Год	Кол-во
0	Белгородская область	Центральный федеральный округ	2013	156801.0
1	Брянская область	Центральный федеральный округ	2013	113056.0
2	Владимирская область	Центральный федеральный округ	2013	357276.0
3	Воронежская область	Центральный федеральный округ	2013	247380.0
4	Ивановская область	Центральный федеральный округ	2013	182692.0
...	...	...	...	...
651	Амурская область	Дальневосточный федеральный округ	2020	244300.0
652	Магаданская область	Дальневосточный федеральный округ	2020	61126.0
653	Сахалинская область	Дальневосточный федеральный округ	2020	188127.0
654	Еврейская автономная область	Дальневосточный федеральный округ	2020	26292.0
655	Чукотский автономный округ	Дальневосточный федеральный округ	2020	15582.0

656 rows × 4 columns

Рис. 4: Подготовленные данные с разбиением по годам

	Область	Округ	Год	Кол-во
0	Белгородская область	Центральный федеральный округ	01.01.2016	37568
1	Брянская область	Центральный федеральный округ	01.01.2016	13415
2	Владимирская область	Центральный федеральный округ	01.01.2016	66342
3	Воронежская область	Центральный федеральный округ	01.01.2016	73521
4	Ивановская область	Центральный федеральный округ	01.01.2016	61345
...	...	...	...	...
1963	Амурская область	Дальневосточный федеральный округ	01.10.2021	4081
1964	Магаданская область	Дальневосточный федеральный округ	01.10.2021	5390
1965	Сахалинская область	Дальневосточный федеральный округ	01.10.2021	NaN
1966	Еврейская автономная область	Дальневосточный федеральный округ	01.10.2021	NaN
1967	Чукотский автономный округ	Дальневосточный федеральный округ	01.10.2021	NaN

1968 rows × 4 columns

Рис. 5: Подготовленные данные с разбиением по кварталам

Глава 4. Моделирование

4.1 Описание реализации

В исследовании были использованы реализации моделей и подходов из библиотеки scikit-hts [16] для языка программирования Python. Программный код и полученные графики представлены в GitHub репозитории[1]. Для нисходящего подхода пропорции дезагрегации определяются на основании средних исторических пропорций, а для подхода оптимального согласования дисперсия ошибки базовых прогнозов оценивается с использованием оценки взвешенных наименьших квадратов.

В результате работы моделей на выходе получаются смоделированные значения для каждого представленного в датасете временного промежутка, а также прогноз на два временных промежутка вперед. Данные значения можно отразить на графике вместе с истинно наблюдаемыми значениям, что позволяет наглядно увидеть точность моделей.

Кроме графического представления анализ ошибок модели происходит с помощью мер точности, выраженных средней абсолютной ошибкой (MAE) и средней абсолютной ошибкой в процентах (MAPE). Преимущество процентных ошибок заключается в том, что они не содержат единиц измерения, и поэтому они часто используются для сравнения эффективности прогнозов между наборами данных. В данной работе они необходимы для сравнения точности на годовых и квартальных данных.

4.2 Результаты

Проведем сравнение результатов работы моделей зафиксировав на каждом уровне иерархии конкретный узел, для которого будет происходить сравнение. Для среднего уровня выберем в качестве такого узла Северо-Западный федеральный округ, а для нижнего – город Санкт-Петербург.

[1] https://github.com/JuliaKil/tourism-hts/blob/master/tourism.ipynb

В таблицах 1 и 2 представлены оценки точности моделей и подходов для годовых и квартальных данных соответственно.

Таблица 1: средние абсолютные ошибки моделей и подходов при работе с годовыми данными (в сотнях тысяч и в процентах)

		Уровень иерархии		
Модель	Подход	Верхний	Средний	Нижний
		Российская Федерация	Северо-Западный федеральный округ	Санкт-Петербург
ARIMA	Восходящий подход	6889.64 **15.32%**	1988.28 **29.8%**	352.54 **14.28%**
	Нисходящий подход	6766.26 **15.32%**	1800.43 **27.68%**	587.16 **22.31%**
	Подход оптимального согласования	6745.77 **15.23%**	1034.47 **15.93%**	342.38 **13.16%**
Prophet	Восходящий подход	5682.37 **12.33%**	2008.23 **31.32%**	294.21 **10.98%**
	Нисходящий подход	5703.74 **12.38%**	1940.75 **31.23%**	406.73 **14.41%**
	Подход оптимального согласования	5688.36 **12.35%**	1006.03 **14.94%**	329.53 **11.66%**

Анализируя полученные результаты, можно сказать, что Prophet показывает сравнимые с ARIMA оценки точности на среднем уровне, но лучшие на высшем и низшем уровнях. Что касается подходов, то восходящим подходом ожидаемо демонстрируются лучшие результаты на нижнем уровне. В то время как на среднем уровне ошибки при применении

подхода оптимального согласования значительно ниже, чем при применении остальных подходов. В целом, для годовых данных наименьшие ошибки дает применение подхода оптимального согласования и библиотеки Prophet. На рис. 6–8 представлены графики, полученные при использовании такой комбинации прогнозной модели и подхода согласования для годовых данных.

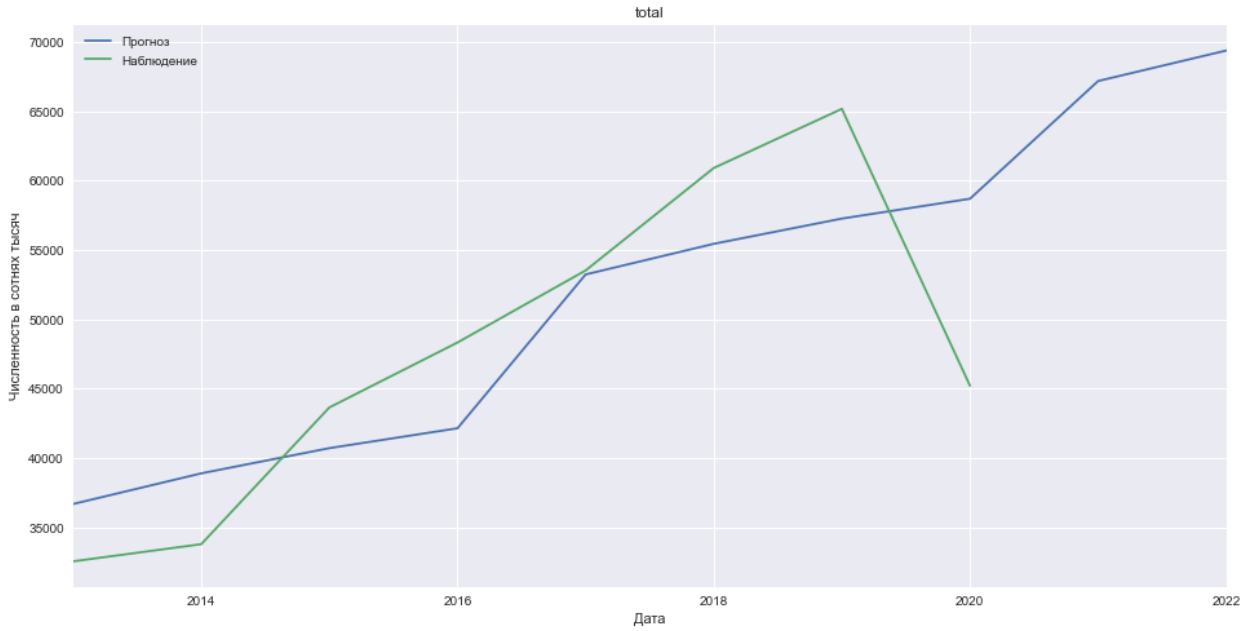

Рис. 6: графическое представление результатов применения подхода оптимального согласования и библиотеки Prophet для годовых данных на верхнем уровне иерархии (Российская Федерация)

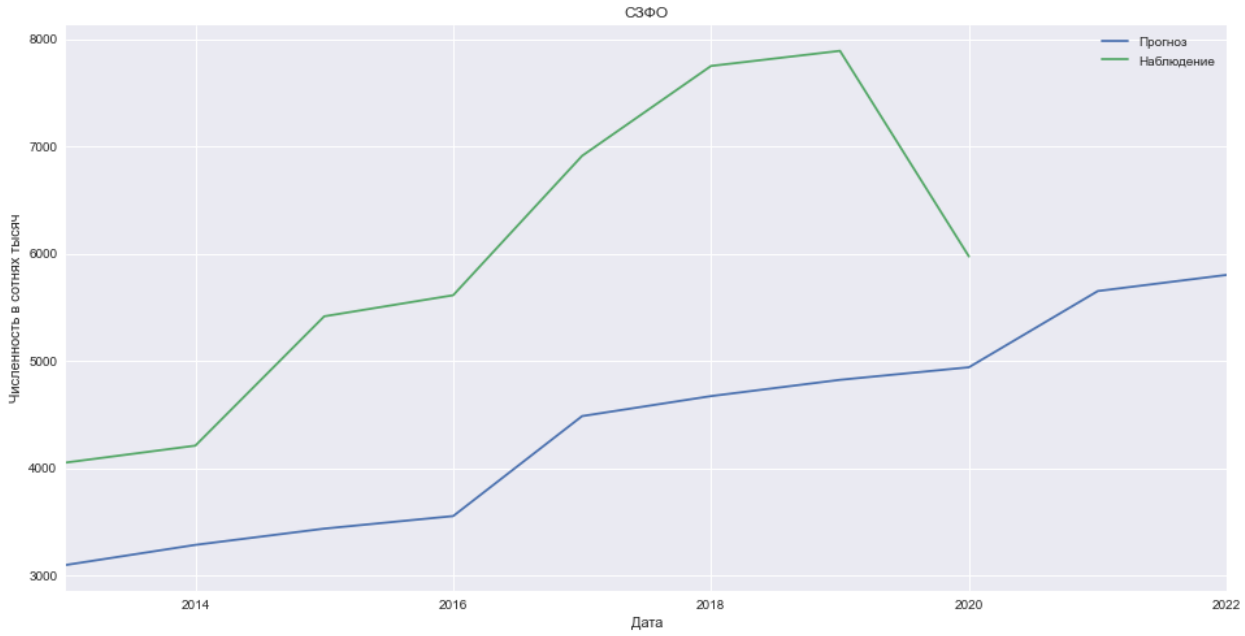

Рис. 7: графическое представление результатов применения подхода оптимального согласования и библиотеки Prophet для годовых данных на среднем уровне иерархии (Северо-Западный федеральный округ)

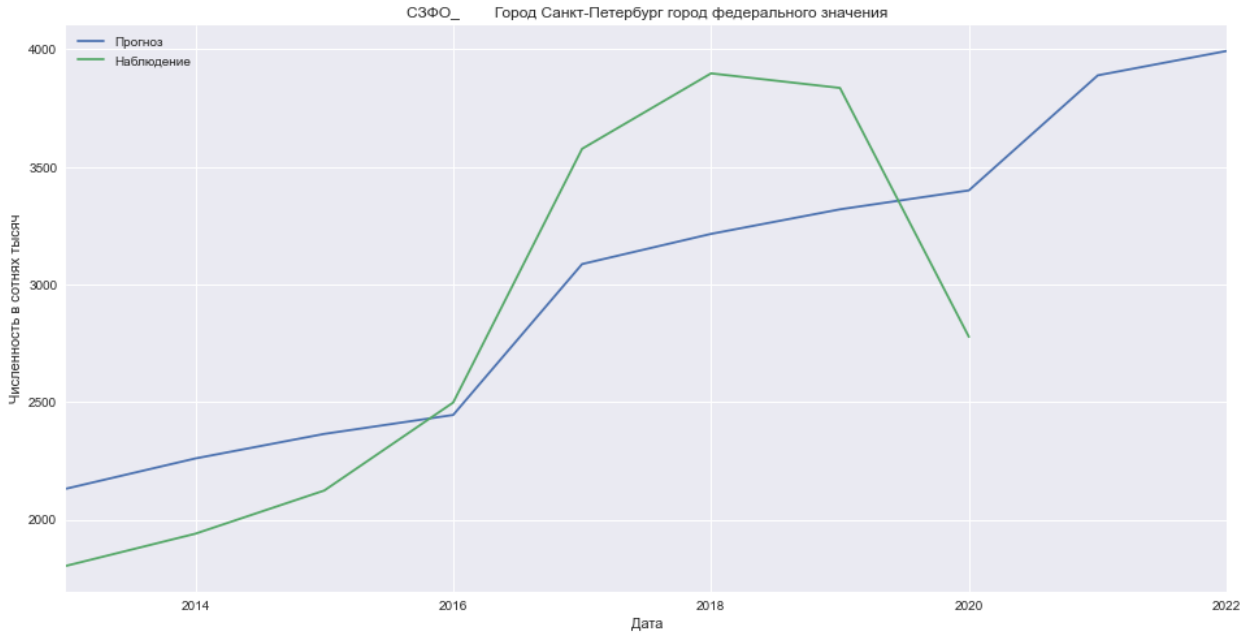

Рис. 8: графическое представление результатов применения подхода оптимального согласования и библиотеки Prophet для годовых данных на нижнем уровне иерархии (Санкт-Петербург)

Таблица 2: средние абсолютные ошибки моделей и подходов при работе с квартальными данными (в сотнях тысяч и в процентах)

		Уровень иерархии		
Модель	Подход	Верхний	Средний	Нижний
		Российская Федерация	Северо-Западный федеральный округ	Санкт-Петербург
ARIMA	Восходящий подход	2807.02 **38.78%**	579.47 **42.36%**	201.62 **50.49%**
	Нисходящий подход	3331.14 **44.04%**	539.07 **40.68%**	239.69 **60.99%**
	Подход оптимального	2921.24 **40.09%**	406.38 **36.62%**	201.15 **48.81%**

	согласования			
SARIMAX	Восходящий подход	4492.53 **45.7%**	764.58 **52.85%**	220.72 **52.68%**
	Нисходящий подход	4683.01 **48.94%**	701.54 **49.95%**	274.14 **59.08%**
	Подход оптимального согласования	4568.08 **47.29%**	528.3 **43.32%**	219.35 **51.17%**
Prophet	Восходящий подход	1839.72 **27.41%**	533.04 **39.52%**	99.47 **29.28%**
	Нисходящий подход	1829.27 **27.3%**	517.06 **39.49%**	178.7 **44.32%**
	Подход оптимального согласования	1833.52 **27.34%**	300.0 **26.76%**	105.81 **29.19%**

Средние абсолютные ошибки на квартальных данных меньше, чем на годовых в абсолютном значении, но больше в процентном. Лучшие результаты на всех уровнях иерархии показывает библиотека Prophet, по сравнению с моделями ARIMA и SARIMAX. Моделью SARIMAX демонстрируются наибольшие значения ошибок среди всех моделей. Ошибки подхода оптимального согласования на нижнем уровне иерархии сопоставимы с ошибками восходящего подхода. Также ошибки этого подхода на среднем уровне меньше ошибок остальных подходов, как и в случае с годовыми данными. И точно так же для квартальных данных наименьшие ошибки дает применение подхода оптимального согласования и библиотеки Prophet. На рис. 9–11 представлены графики, полученные при использовании такой комбинации прогнозной модели и подхода согласования для квартальных данных.

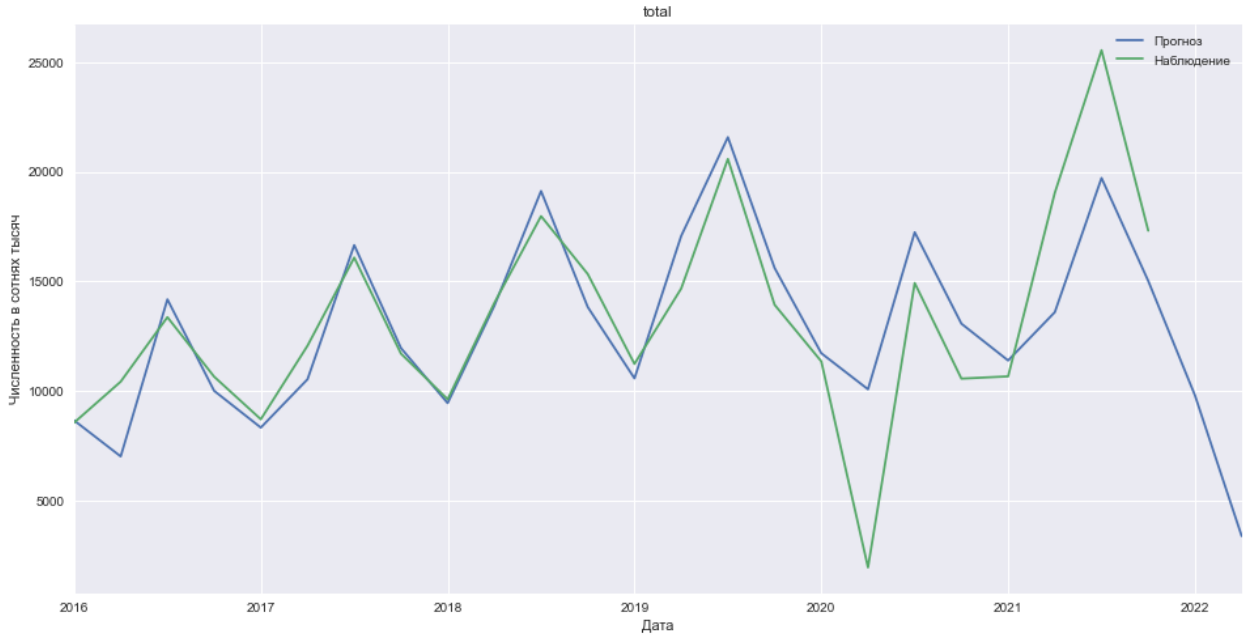

Рис. 9: графическое представление результатов применения подхода оптимального согласования и библиотеки Prophet для квартальных данных на верхнем уровне иерархии (Российская Федерация)

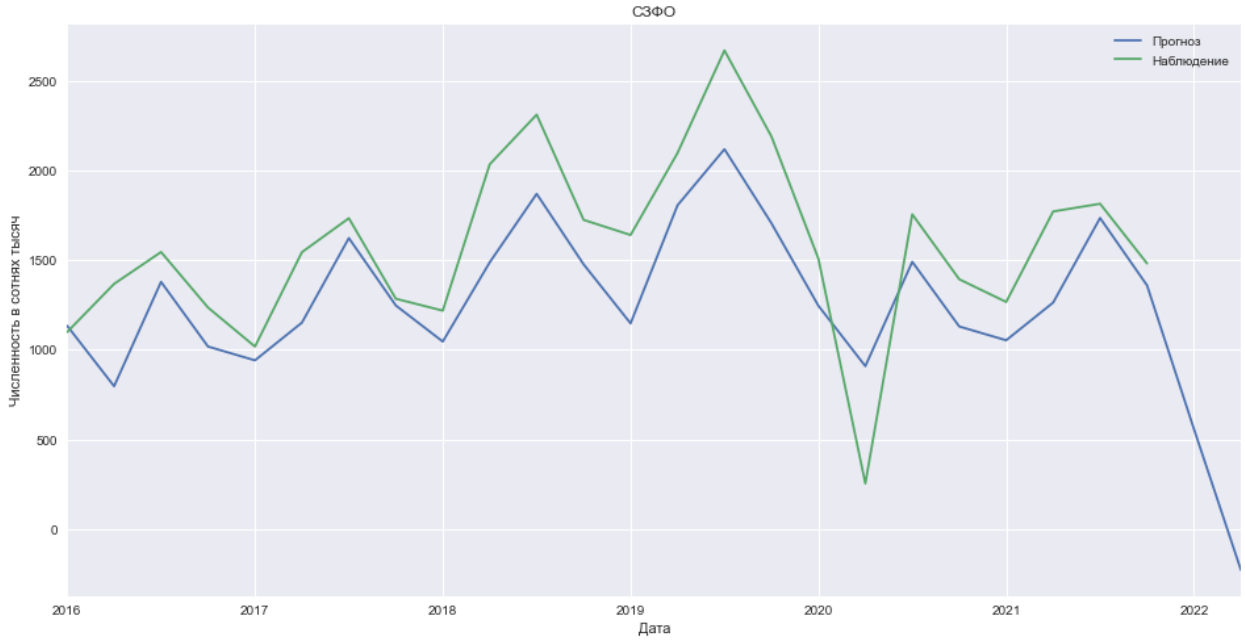

Рис. 10: графическое представление результатов применения подхода оптимального согласования и библиотеки Prophet для квартальных данных на среднем уровне иерархии (Северо-Западный федеральный округ)

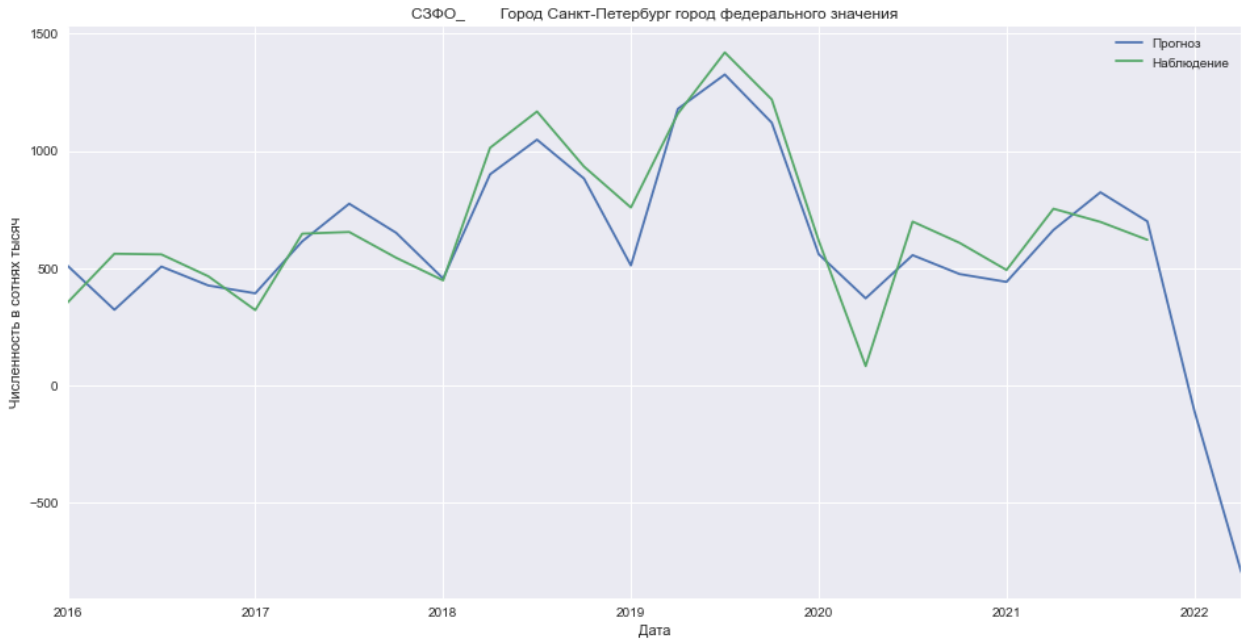

Рис. 11: графическое представление результатов применения подхода оптимального согласования и библиотеки Prophet для квартальных данных на нижнем уровне иерархии (Санкт-Петербург)

В таблицах 3 и 4 представлены прогнозы моделей и подходов для годовых и квартальных данных соответственно.

Таблица 3: прогнозы моделей и подходов для годовых данных (в сотнях тысяч)

			Уровень иерархии		
Модель	Подход	Год	Верхний	Средний	Нижний
			Российская Федерация	Северо-Западный федеральный округ	Санкт-Петербург
ARIMA	Восходящий подход	2021	44891.15	3744.63	1855.66
		2022	43155.20	3685.21	1549.11
	Нисходящий подход	2021	31271.60	2636.08	1814.43
		2022	31950.44	2693.30	1853.82
	Подход	2021	38070.66	3736.97	1747.40

	оптимального согласования	2022	37472.50	3431.40	1467.22
Prophet	Восходящий подход	2021	67335.85	5115.62	4913.73
		2022	70134.84	5346.82	5934.91
	Нисходящий подход	2021	67034.15	5650.73	3889.44
		2022	68821.52	5801.40	3993.15
	Подход оптимального согласования	2021	67194.13	6982.02	4723.17
		2022	69392.11	7338.58	5716.25

Таблица 4: прогнозы моделей и подходов для квартальных данных (в сотнях тысяч)

			Уровень иерархии		
Модель	Подход	Квартал (2022 г.)	Верхний	Средний	Нижний
			Российская Федерация	Северо-Западный федеральный округ	Санкт-Петербург
ARIMA	Восходящий подход	I	14709.02	811.99	665.18
		II	12593.40	497.50	695.17
	Нисходящий подход	I	13414.28	1118.89	701.91
		II	13316.38	1110.72	696.79
	Подход оптимального согласования	I	13471.33	1160.80	641.53
		II	12502.62	1089.04	676.22
SARIM	Восходящий	I	14973.9	1287.09	565.72

AX	подход	II	13067.6	1132.32	515.44
	Нисходящий подход	I	15984.45	1333.27	836.39
		II	14755.95	1230.80	772.11
	Подход оптимального согласования	I	15495.55	1363.93	570.96
		II	13957.31	1258.63	524.83
Prophet	Восходящий подход	I	10188.92	641.76	-111.39
		II	4820.43	46.50	-807.19
	Нисходящий подход	I	9529.89	794.89	498.66
		II	2367.59	197.48	123.89
	Подход оптимального согласования	I	9804.00	557.97	-106.99
		II	3389.81	-227.49	-794.69

Можно заметить, что большинство прогнозов имеет нисходящий тренд. Возможно предположить, что такое поведение моделей связано с аномалиями 2020 года, встречающимися в данных. В некоторых случаях это приводит к отрицательным значениям прогнозов, что, конечно, не реалистично. Причины и способы решения проблемы отрицательных прогнозов описываются в [17].

Выводы

В данной работе была проанализирована эффективность различных подходов к согласованию прогнозов иерархических временных рядов и различных прогнозных моделей применительно к данным туризма Российской Федерации. Как для годового, так и для квартального разбиения данных наибольшую точность показывает применение библиотеки Prophet и подхода оптимального согласования (в исследовании ошибки базовых прогнозов оценивались с использованием оценки взвешенных наименьших квадратов). На квартальных данных полученные ошибки моделей меньше в абсолютных значениях, но при этом больше в процентах.

Получены прогнозы на 2021–2022 годы для годовых данных и на I–II кварталы 2022 года для квартальных данных. Большинство прогнозов моделей имеют нисходящий тренд.

Заключение

Таким образом, были изучены подходы к согласованию прогнозов иерархических временных рядов, все подходы были реализованы в программе, написанной на языке Python с использованием библиотеки scikit-hts, и исследованы применительно к данным о количестве граждан РФ, размещенных в коллективных средствах размещения. Поставленные задачи выполнены, проведен сравнительный анализ эффективности указанных подходов и прогнозных моделей, получены прогнозы на 2021–2022 годы и I–II кварталы 2022 года.

Список литературы

1. Montgomery D. C., Jennings C. L., Kulahci M. Introduction to time series analysis and forecasting. – John Wiley & Sons, 2015.
2. Nielsen A. Practical time series analysis: Prediction with statistics and machine learning. – O'Reilly Media, 2019.
3. Hyndman R. J., Athanasopoulos G. Forecasting: principles and practice. – OTexts, 2018.
4. Gross C. W., Sohl J. E. Disaggregation methods to expedite product line forecasting //Journal of forecasting. – 1990. – Т. 9. – №. 3. – С. 233-254.
5. Hyndman R. J. et al. Optimal combination forecasts for hierarchical time series //Computational statistics & data analysis. – 2011. – Т. 55. – №. 9. – С. 2579-2589.
6. Athanasopoulos G., Ahmed R. A., Hyndman R. J. Hierarchical forecasts for Australian domestic tourism //International Journal of Forecasting. – 2009. – Т. 25. – №. 1. – С. 146-166.
7. Hyndman R. J., Lee A. J., Wang E. Fast computation of reconciled forecasts for hierarchical and grouped time series //Computational statistics & data analysis. – 2016. – Т. 97. – С. 16-32.
8. Wickramasuriya S. L., Athanasopoulos G., Hyndman R. J. Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization //Journal of the American Statistical Association. – 2019. – Т. 114. – №. 526. – С. 804-819.
9. Athanasopoulos G. et al. Forecasting with temporal hierarchies //European Journal of Operational Research. – 2017. – Т. 262. – №. 1. – С. 60-74.
10. Taylor S. J., Letham B. Forecasting at scale //The American Statistician. – 2018. – Т. 72. – №. 1. – С. 37-45.
11. The ARIMAX model muddle //Rob J Hyndman [Электронный ресурс] — URL: https://robjhyndman.com/hyndsight/arimax/ (дата обращения: 12.05.2022)

12. SARIMAX: Introduction //statmodels [Электронный ресурс] — URL: https://www.statsmodels.org/dev/examples/notebooks/generated/statespace_sarimax_stata.html (дата обращения: 12.05.2022)
13. Библиотека Prophet [Электронный ресурс] — URL: https://facebook.github.io/prophet/ (дата обращения: 22.02.2022)
14. Численность граждан Российской Федерации, размещенных в коллективных средствах размещения (Росстат) //Федеральное агентство по туризму [Электронный ресурс] — URL: https://tourism.gov.ru/contents/analytics/statistics/chislennost-grazhdan-rossiyskoy-federatsii-razmeshchennykh-v-kollektivnykh-sredstvakh-razmeshcheniya/ (дата обращения: 11.05.2022)
15. Документация библиотеки pandas [Электронный ресурс] — URL: https://pandas.pydata.org/docs/ (дата обращения: 16.04.2022)
16. Документация библиотеки для работы с иерархическими временными рядами scikit-hts [Электронный ресурс] — URL: https://scikit-hts.readthedocs.io/en/latest/ (дата обращения: 03.12.2021)
17. Wickramasuriya S. L., Turlach B. A., Hyndman R. J. Optimal non-negative forecast reconciliation //Statistics and Computing. – 2020. – Т. 30. – №. 5. – С. 1167-1182.