УДК 004.67 DOI: 10.46573/2658-5030-2026-3-121-129

РЕШЕНИЕ ЗАДАЧИ ПРОГНОЗИРОВАНИЯ ДЛЯ ПОСЛЕДОВАТЕЛЬНОСТЕЙ СО СВЕРХБОЛЬШИМ КОЛИЧЕСТВОМ ВРЕМЕННЫХ РЯДОВ

С.Л. ГОРОБЧЕНКО[1], канд. техн. наук, В.А. СОКОЛОВА[1], канд. техн. наук,
Н.Н. ВЕРНЕР[2], канд. техн. наук, С.А. ВОЙНАШ[3], ассистент,
С.В. АЛЕКСЕЕВА[4], канд. техн. наук

[1]Санкт-Петербургский государственный университет промышленных технологий и дизайна, 191186, Санкт-Петербург, ул. Большая Морская, 18,
e-mail: sgorobchenko@yandex.ru
[2]Санкт-Петербургский государственный лесотехнический университет им. С.М. Кирова, 194021, Санкт-Петербург, Институтский пер., 5,
e-mail: wernern@mail.ru
[3]Российский университет дружбы народов имени Патриса Лумумбы**,** 117198, Москва, ул. Миклухо-Маклая, 6,
e-mail: sergeyvoinash@yandex.ru
[4]Санкт-Петербургский государственный архитектурно-строительный университет, 190005, Санкт-Петербург, 2-я Красноармейская ул., 4,
e-mail: pum222@mail.ru

Статья посвящена решению задачи прогнозирования для последовательностей, содержащих сверхбольшое количество временных рядов (до 300 тысяч и более) и характеризующихся нестационарностью, разнородностью и наличием сезонности.

Рассмотрена проблема выбора оптимальной методики прогнозирования на примере данных о продажах крупной распределительной сети. Проведен сравнительный анализ нескольких моделей временных рядов, реализованных в библиотеках языка Python (pmdarima, statsmodels, skforecast, statsforecast, sktime), включая ARIMA, SARIMA и их модификации с экзогенными факторами (ARIMAX, SARIMAX). Обоснован выбор наиболее эффективной модели SARIMAX на основе минимизации ошибки RMSE. Для повышения вычислительной эффективности при работе с массивами данных большого объема предложен подход, основанный на предварительной кластеризации рядов с использованием методов k-средних, метрик Евклида и DTW, а также оптимизации числа кластеров методами «локтя» и «силуэта». Показано, что прогнозирование на основе центроидов кластеров с последующей обратной нормировкой позволяет значительно сократить время вычислений без критической потери точности. Результаты работы подтверждают эффективность предложенного подхода для построения достоверных прогнозов в системах управления поставками.

Ключевые слова: временные ряды, прогнозирование, большие данные, кластеризация, экзогенные переменные, метод k-средних, машинное обучение, Python, оптимизация, анализ данных.

ВВЕДЕНИЕ

Среди множества видов временных рядов, которые встречаются в различных задачах анализа больших баз данных, выделяются последовательности со сверхбольшим количеством временных рядов [1–6]. Они могут обладать элементами

нестационарности, случайности, разной длительности, зависимости, иерархичности, сложности и наблюдаемости, переменностью и регулярностью структуры и мультикорреляции [7, 8].

Подобные характеристики делают задачу анализа и выработки прогноза достаточно сложной, поскольку предполагают необходимость изначальной обработки и предпрогнозного построения массива данных с применением разнообразных методик, таких как нелинейная динамика, определение оптимального объема выборки и группирование [9, 10]. Предобработанные массивы данных должны соответствовать выбору методики, которая в свою очередь должна отвечать особенностям самого массива данных [11].

Области данных, обладающих в основном числовым форматом и достаточно легко поддающихся цифровой обработке, лежат в сфере статистических и математических видов обработки больших данных. К ним можно отнести общестатистические и математические методы. К наиболее популярным, часто упоминаемым в литературе, относятся регрессионные и другие методы (АРПСС, SSA, SARIMA), а также различные виды спектрального анализа, разложения в ряды Фурье и преобразований (F-преобразование).

Выбор той или иной методики определяется возможностью получения достоверного и верифицируемого прогноза. В частности, необходимо исходить из метрики ряда (одномерный/многомерный) и требуемой точности, что определит выбор высокоточных или робастных методик. В связи с этим часто используются комплексные методы и методы гибридного моделирования.

Для решения задач обработки данных и прогнозирования для последовательностей со сверхбольшим количеством временных рядов, выражаемых в числовом формате, оптимально выбирать статистические методы. В работе демонстрируется выбор и применение статистических методов библиотеки Python для прогнозирования указанных последовательностей. Целью проведенного исследования является прогнозирование продаж крупной компании с распределенной торговой сетью с количеством рядов, превышающих 300 тысяч, и объемом данных в несколько десятков терабайт со сроком накопления данных свыше трех лет.

МЕТОДИКА ВЫБОРА МОДЕЛИ

Для работы с таким огромным массивом данных предварительно выберем наилучшую из стандартных моделей. Выбор будем осуществлять на меньшей компании с 917 рядами, априорно предполагая стационарность всех предоставленных рядов. Анализ стационарности проводится заранее исходя из общих соображений о характере компаний. Дополнительно на некоторой выборке проводятся тесты Квятковского – Филлипса – Шмидта – Шина, а также расширенный тест Дики – Фуллера и Филлипса – Перрона. Они должны подтвердить или опровергнуть предложенную гипотезу о стационарности рядов.

Выбор будем осуществлять из нескольких стандартных, имеющихся в библиотеках Python.

1. Модель AUTO-ARIMA из библиотеки pmdarima

Используем модель ARIMA (p, q), которая описывается равенством

$$y_t = a_1 y_{t-1} + \ldots + a_p y_{t-p} + e_t + b_1 e_{t-1} + \ldots + b_q e_{t-q}, \qquad (1)$$

где a_i, b_i – параметры модели; e_t – гауссовский белый шум.

Опишем функцию AUTO-ARIMA из библиотеки pmdarima, что позволит провести выбор параметров модели.

1.1. AUTO-ARIMA

Функция AUTO-ARIMA стремится определить оптимальные параметры для ARIMA-модели. Чтобы найти наилучшую модель, AUTO-ARIMA оптимизирует для заданного информационного критерия его значение и возвращает ARIMA, которая минимизирует это значение (в качестве информационного критерия был выбран AIC). Кроме автоматического подбора параметров модели, преимущество AUTO-ARIMA состоит также в том, что она проводит дифференциальные тесты для определения порядка дифференцирования, в связи с чем постоянно анализируется и подтверждается стационарность исследуемых рядов.

1.2. График

В качестве прогноза AUTO-ARIMA было взято среднее значение по ряду. Визуальная оценка показывает, что ряд неудачный, поэтому был проведен переход к модели SARIMA.

2. Модель SARIMA из библиотеки statsmodels

Нетрудно заметить, что во временном ряду прослеживаются сезонные колебания с периодичностью в один год. Поэтому можно ожидать, что прогноз модели SARIMA, учитывающей сезонные колебания, будет лучше, чем прогноз модели AUTO-ARIMA. Необходимо подобрать параметры (p, d, q) и (P, D, Q) в модели SARIMA:

$$(\mathrm{p}, \mathrm{d}, \mathrm{q}, \mathrm{P}, \mathrm{D}, \mathrm{Q}, \mathrm{s} = 52) \quad (2)$$

2.1. Optuna

Подбор параметров будем осуществлять с помощью библиотеки optuna в Python. Разобьем данные на две группы: обучающую выборку и тестовую, используя соотношения:

$$\mathrm{p}, \mathrm{q} \in \{0 \dots 4\}\, \mathrm{d}, \mathrm{D} \in \{0 \dots 2\}\, \mathrm{P}, \mathrm{Q} \in \{0 \dots 3\} \quad (3)$$

Обучающая выборка представляет собой начальный 80%-й отрезок ряда, а тестовая – оставшиеся 20 %. Обучая нашу модель на первой группе данных, с помощью optuna добиваемся минимизаций RMSE на оставшейся тестовой выборке, получая тем самым оптимальные значения:

2.2. График

При добавлении сезонности прогноз в достаточной степени реагирует на сезонные колебания, но плохо чувствует величину скачков.

3. Модель ARIMAX и SARIMAX из библиотеки skforecast

3.1. Построение экзогенной переменной

Заметим, что с 2021–2023 года в конце декабря и в начале января наблюдается сильное снижение продаж, поэтому в качестве экзогенных факторов в нашем случае возьмем новогодние праздники, а именно: выходные дни с 1 и 2 января и укороченные дни 31 декабря и с 3 по 7 января. Из-за того, что недельные данные с праздничными днями могут приходиться на разные недели, для построения экзогенного ряда каждый непраздничный и неукороченный день принимается за 1, праздничный – за 0, а укороченный – за 0,5, после чего проводится понедельное усреднение для получения требуемой экзогенной переменной.

3.2. Подбор параметров и графики

Для подбора параметров модели используем, как и в предыдущем пункте, обучающую и тестовую выборку продаж с тем же процентом (80 %) и библиотеку

optuna для минимизации RMSE. Отличие состоит лишь в том, что добавляется обучающая и тестовая выборки экзогенного ряда.

В отличие от моделей AUTO-ARIMA из библиотеки pmdarima и SARIMA из statsmodels, модель ARIMAX из skforecast уже учитывает экзогенные факторы, которые проявляются в виде просадки.

4. Модель ARIMAX из библиотеки statsforecast

Рассмотрим модель ARIMAX из библиотеки statsforecast. Ее функционал позволяет добавить сезонность, поэтому, учитывая прошлый опыт, будем сразу строить ее сезонную модификацию. Проделав все то же самое, что и в случае модели SARIMAX из библиотеки skforecast, получим, что, в отличие от других SARIMA из ранее рассмотренных библиотек, сезонность выражена очень слабо. Из-за этого она похожа на обычную модель ARIMA с незначительным добавлением сезонности. Эта модель, очевидно, проигрывает модели SARIMAX из библиотеки skforecast как визуально, так и по метрике RMSE.

Сравнив между собой все указанные выше модели, можно признать наилучшей первую модель SARIMAX из библиотеки skforecast. Именно ее и будем использовать для дальнейшего прогноза.

5. Применение выбранной модели SARIMAX ко всему первому массиву временных рядов

Применив выбранную модель ко всем 917 рядам, получили, что первые два временных ряда, на которые прогноз лучше, похожи друг на друга визуально, а количество продаж товара на них выше, чем на двух следующих. Модель также «уловила» сезонное колебание среднего значения *y* ряда.

6. Кластеризация

6.1. Обработка данных

Предыдущие примеры показывают, что перед использованием выбранной модели желательно предварительно разбить временные ряды на «похожие», чтобы учесть их специфику, тем самым обеспечить возможность их кластеризации. Предварительно проведем обработку данных, удалив из 300 тысяч временных рядов те 180 тысяч, на графиках которых отсутствовали продажи за последние 4 месяца. В имеющихся временных рядах продаж относительно новых товаров пропуски в них дозаполним нулями, чтобы все временные ряды можно было рассматривать на одном и том же временном промежутке. Для того чтобы считать расстояние между рядами в выбранной метрике, предварительно проведем нормирование рядов. Это можно делать двумя способами: смещением на минимум и последующим делением временного ряда на максимум или последующей нормировкой на отрезке $[-1; 1]$.

6.2. Метод k-средних

Для кластеризации оставшихся 120 тысяч рядов после обработки воспользуемся методом k-средних. Метод заключается в следующем:

1. Выбирается число кластеров k.

2. Из исходного множества временных рядов случайным образом выбираются k, которые будут служить начальными центрами кластеров.

3. Для каждого из исходных временных рядов определяется ближайший к нему центр кластера (в выбранной метрике).

4. Проводится перевычисление центра каждого кластера с нахождением среднего между рядами в каждом кластере.

5. Проводится повторное разбиение рядов на кластеры в соответствии с тем, какой из новых центров оказался ближе в выбранной метрике.

Выход из алгоритма осуществляется, если изменение расстояний между центрами незначительное или ограничивается количество итераций.

6.3. Оптимизация числа кластеров

Для того чтобы избежать «недообучения» или «переобучения», в алгоритме кластеризации необходимо подобрать оптимальное число кластеров. Для этого воспользуемся методами «локтя» и «силуэта».

6.3.1. Метод «локтя»

Метод применяется к алгоритму k-средних и заключается в том, что для каждого натурального числа k применяется метод k-средних, а далее вычисляется WCSS-метрика, равная сумме квадратов расстояний каждого временного ряда до его центра в кластере. Затем строится график зависимости значения WCSS от количества кластеров, затем ищется точка излома «локтя», которая и указывает на оптимальное количество кластеров.

6.3.2. Метод «силуэта»

Для каждого временного ряда в кластере вводится величина ширины силуэта *S*, которая вычисляется по формуле

$$S(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))}\ , \qquad (4)$$

где i – временной ряд, а $a(i)$ и $b(i)$ – среднее расстояние между временным рядом i и всеми остальными рядами в этом кластере и среднее расстояние между i и всеми остальными рядами из других кластеров соответственно.

Затем строится график зависимости среднего значения ширины силуэта рядов от количества кластеров и выбирается оптимальное значение k.

6.3.3. График метода «локтя» и метода «силуэта»

Применив указанные методы для евклидовой и DTW-метрики на меньшей компании, получили, что оба метода указывают на оптимальное значение k, примерно равное 24 в случае евклидовой метрики и 20 в случае метрики DTW.

Метод «локтя» для большей компании: оптимальное значение k составляет 36.

6.4. Кластеры

Для выбранного оптимального значения k, применив метод k-средних и взяв в качестве метрики евклидову и DTW-метрику, получим кластеризацию, показанную на рис. 1.

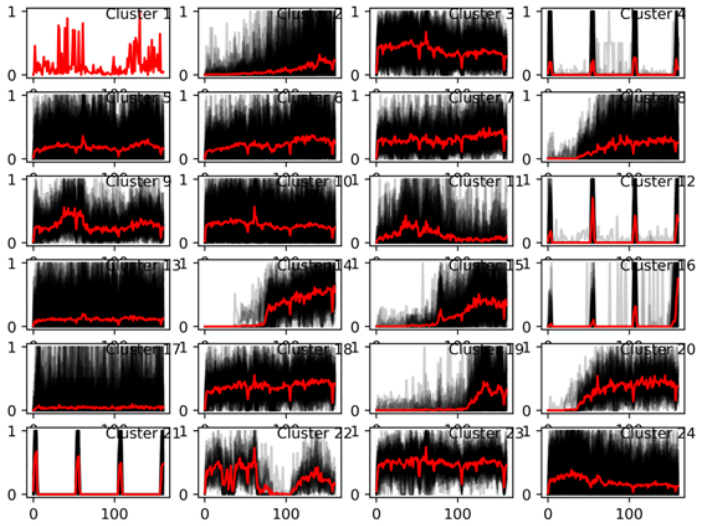
6.4.1. Прогноз

Взяв центроид в каждом кластере и применив для него выбранную нами модель SARIMAX, получим прогноз на кластер, который поднимается до прогноза на каждый отдельный элемент соответствующего класса обратной нормировкой.

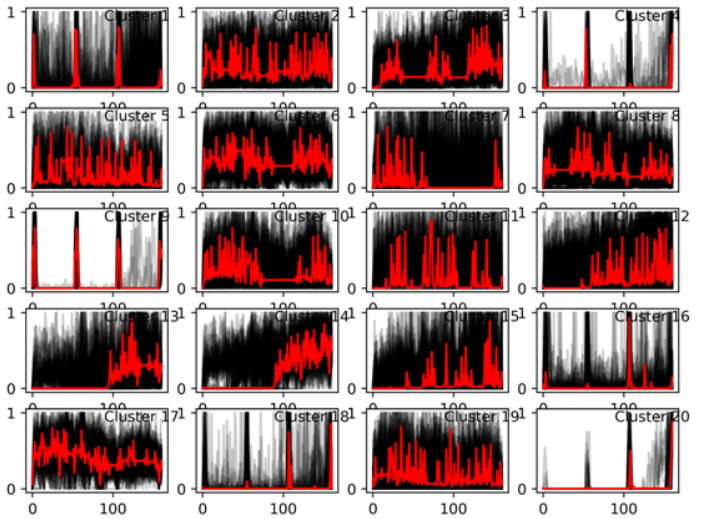
6.4.2. Векторный метод кластеризации

Рассмотрим векторный метод кластеризации временных рядов. Сопоставим каждому временному ряду вектор из 700 его параметров, а затем, применив метод k-средних к векторам параметров, получим кластеризацию по вектору и, соответственно, кластеризацию рядов, которые соответствуют данным векторам параметров.

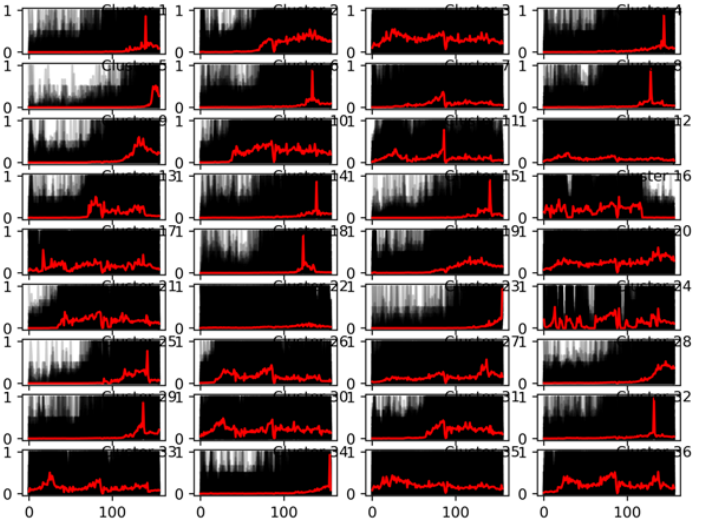
Сложность заключается в том, что для 300 тысяч рядов этот процесс весьма трудоемок, поэтому необходимо сократить размерность векторов параметров и счет (т.е. количество расчетов, которые проводятся на компьютере) может достигать до месяца на компьютерах средней мощности. Осуществим этот процесс с помощью метода главных компонент.



(а)



(б)



(в)

Рис. 1. Кластеризация для временных рядов большей компании с евклидовой метрикой: (а)–(в) – этапы кластеризации (графиками изображены центры кластеров)

7. Результаты

Рассмотрены модели ARIMAX, SARIMAX, реализованные от разных библиотек Python. Установлена наилучшая модель для прогноза еженедельных продаж. Для работы с большими данными сгенерирован алгоритм, который ускоряет вычисление прогноза.

Проведем сравнение оценок при разном подходе к кластеризации временных рядов (рис. 2 и 3).

	Unnamed: 0	r2	rmse
count	110012.000000	1.100120e+05	110012.000000
mean	55005.500000	-5.559598e+01	4.913733
std	31757.873244	1.026420e+04	49.072667
min	0.000000	-3.033997e+06	0.020535
25%	27502.750000	-7.893860e-01	0.619812
50%	55005.500000	-2.128069e-01	1.023013
75%	82508.250000	-3.199674e-02	1.809221
max	110011.000000	7.045325e-01	4545.227650

Рис. 2. Распределение без учета новинок

	rmse
count	20565.000000
mean	14.256572
std	97.359820
min	0.162232
25%	1.754793
50%	2.807351
75%	4.679743
max	4097.407157

(а)

	rmse
count	20355.000000
mean	8.855974
std	60.257557
min	0.000000
25%	1.075583
50%	1.745273
75%	3.474619
max	1917.837488

(б)

Рис. 3. Распределение с учетом новинок (а)
и предложенные компанией оценки, полученные с помощью прогноза
на каждом временном ряду (б)

Из представленных оценок следует, что оценки полученных прогнозов с помощью кластеризации незначительно ухудшают RMSE, но время прогнозирования при этом существенно сокращается. Такие модели могут использоваться в системе для прогноза поставок в точки продаж.

ЗАКЛЮЧЕНИЕ

Задачи нахождения решений для последовательностей со сверхбольшим количеством временных рядов по-прежнему остаются достаточно сложными и трудоемкими. Определяющим для этих целей можно считать выбор методики, который желательно проводить итерационным способом с расчетом отклика модели и ее верификацией по каждому варианту. Одним из важных моментов является разбиение общего объема выборки на рабочую и тестовую. Это дает возможность сравнивать результаты расчетов и моделирования на той же самой выборке. Полученные результаты позволяют спрогнозировать снижение затратности проведения расчетов, иметь достоверные и верифицируемые данные также и для расчета подобных выборок во многих областях техники и бизнеса.

ЛИТЕРАТУРА

1. Сотников А.И. Анализ временных рядов с использованием алгоритмов Big data // *Информационно-технологический вестник*. 2021. № 1 (27). С. 118–124.

2. Безрукавный Д.С. Анализ и управление данными в виде временных рядов: дис. … канд. техн. наук. Москва, 2007. 115 с.

3. Тюнина Е.С. Методы анализа и прогнозирования временных рядов // *Современные технологии в науке и образовании – СТНО-2019*: *Сборник трудов II Международного научно-технического форума*. В 10 т. / под общ. ред. О.В. Миловзорова. Рязань: ИП Коняхин А.В., 2019. Т. 4. С. 148–151.

4. Кричевский А.М. Прогнозирование временных рядов с долговременной корреляционной зависимостью: дис. … канд. техн. наук. Санкт-Петербург, 2008. 179 с.

5. Сергеев А.В. Выявление и анализ тенденций временных рядов. *Международная конференция по мягким вычислениям и измерениям*. В 2 т. СПб.: Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина). 2019. Т. 1. С. 468–471.

6. Гончарова Н.Д., Терехова Ю.С. Анализ и моделирование статистических рядов: учебное пособие. Новосибирск: СибГУТИ, 2016. 97 с.

7. Зубов А.В., Зубов И.В., Зубов С.В., Стрекопытов И.С., Стрекопытова М.В. Аналитическая природа случайных последовательностей // *Научно-технические ведомости Санкт-Петербургского государственного политехнического университета. Физико-математические науки.* 2010. № 3 (104). С. 84–89.

8. Zubov A.V., Zubov S.V. Vector of control observation, synthesis, number of entrances and exits, structure, aggregate, coefficient, multiple // *Middle Volga Mathematical Society Journal.* 2014. V. 16. № 1. P. 168–176.

9. Росенко А.П. Методика обработки массива данных, полученных экспертным путем // *Доклады Томского государственного университета систем управления и радиоэлектроники.* 2012. № 1-2 (25). С. 192–197.

10. Ташполатова Б.Б. Методы анализа случайных данных и их алгоритмизация. *Международная конференция по мягким вычислениям и измерениям*. 2018. В 2 т. СПб.: Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина). Т. 2. С. 521–524.

11. Рычка О.В. Сравнительный анализ методов интеллектуальной обработки данных и повышения качества прогнозных моделей // *Информатика и кибернетика*. 2023. № 2 (32). С. 36–42.

Для цитирования: Горобченко С.Л., Соколова В.А., Вернер Н.Н., Войнаш С.А., Алексеева С.В. Решение задачи прогнозирования для последовательностей со сверхбольшим количеством временных рядов // Вестник Тверского государственного технического университета. Серия «Технические науки». 2026. № 3 (31). С. 121–129.

SOLVING THE FORECASTING PROBLEM FOR SEQUENCES WITH AN ULTRA-LARGE NUMBER OF TIME SERIES

S.L. GOROBCHENKO[1], Cand. Sc., V.A. SOKOLOVA[1], Cand. Sc.,
N.N. WERNER[2], Cand. Sc., S.A. VOINASH[3], Assistant, S.V. ALEKSEEVA[4], Cand. Sc.

[1]Saint Petersburg State University of Industrial Technologies and Design,
18, Bolshaya Morskaya St., Saint Petersburg, 191186, e-mail: sgorobchenko@yandex.ru
[2]Saint Petersburg State Forest Technical University named after S.M. Kirov,
5, Institutsky Lane, Saint Petersburg, 194021, e-mail: wernern@mail.ru
[3] Peoples' Friendship University of Russia named after Patrice Lumumba,
6, Miklukho-Maklaya St., Moscow, 117198, e-mail: sergeyvoinash@yandex.ru
[4] Saint Petersburg State University of Architecture and Civil Engineering,
4, 2-ya Krasnoarmeyskaya St., Saint Petersburg, 190005, e-mail: pum222@mail.ru

This article addresses the problem of forecasting sequences containing an extremely large number of time series (up to 300,000 or more) characterized by non-stationarity, heterogeneity, and seasonality. The problem of choosing the optimal forecasting method is considered using sales data from a large distribution network as an example. The study provides a comparative analysis of several time series models implemented in Python libraries (pmdarima, statsmodels, skforecast, statsforecast, sktime), including ARIMA, SARIMA, and their modifications with exogenous factors (ARIMAX, SARIMAX). The choice of the most effective SARIMAX model is justified based on minimizing the RMSE error. To improve computational efficiency when working with large data sets, an approach is proposed based on preliminary clustering of series using the k-means, Euclidean, and DTW methods, as well as optimizing the number of clusters using the elbow and silhouette methods. It is shown that forecasting based on cluster centroids followed by inverse normalization significantly reduces computation time without critically losing accuracy. The results confirm the effectiveness of the proposed approach for generating reliable forecasts in supply chain management systems.

Keywords: time series, forecasting, big data, clustering, exogenous variables, k-means method, machine learning, Python, optimization, data analysis.

Поступила в редакцию/received: 17.04.2026; после рецензирования/revised: 21.04.2026; принята/accepted: 28.04.2026