МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Чорноморський національний університет
імені Петра Могили
Факультет комп'ютерних наук
Кафедра інтелектуальних інформаційних систем

ДОПУЩЕНО ДО ЗАХИСТУ
Завідувач кафедри інтелектуальних
інформаційних систем, д-р техн. наук, проф.
_____Ю. П. Кондратенко
«____»_____2023 р.

БАКАЛАВРСЬКА КВАЛІФІКАЦІЙНА РОБОТА

ІНФОРМАЦІЙНА СИСТЕМА ПРОГНОЗУВАННЯ ЧАСОВИХ РЯДІВ НА ОСНОВІ МЕТОДОЛОГІЇ GAM

Спеціальність 122 «Комп'ютерні науки»

122 – БКР – 402.21910219

Виконала студентка 4-го курсу, групи 402
_____ Д. П. Пожар
«19» червня 2023 р.

Керівник: канд. техн. наук, доцент
_____І. О. Калініна
«19» червня 2023 р.

Миколаїв – 2023

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Чорноморський національний університет ім. Петра Могили
Факультет комп'ютерних наук
Кафедра інтелектуальних інформаційних систем

Рівень вищої освіти **бакалавр**
Спеціальність **122 «Комп'ютерні науки»**
(шифр і назва)
Галузь знань **12 «Інформаційні технології»**
(шифр і назва)

ЗАТВЕРДЖУЮ

Завідувач кафедри інтелектуальних інформаційних систем, д-р техн. наук, проф.
_____ Ю. П. Кондратенко
«____» _____2022 р.

З А В Д А Н Н Я
на виконання кваліфікаційної роботи

Видано студентці групи 402 факультету комп'ютерних наук Пожар Дарині Петрівні.

1. Тема кваліфікаційної роботи «Інформаційна система прогнозування часових рядів на основі методології GAM».

Керівник роботи Калініна Ірина Олександрівна, канд. техн. наук, доцент.

Затв. наказом Ректора ЧНУ ім. Петра Могили від «___» _____20__р. № _____

2. Строк представлення кваліфікаційної роботи студентом «___» ___ 20__р.

3. Вхідні (початкові) дані до роботи: набір даних готельних номерів за певний період часу.

Очікуваний результат: прогнозування вартості готельних номерів та підбір моделей для прогнозування значень часового ряду, аналіз результатів та вибір оптимальної моделі.

4. Перелік питань, що підлягають розробці (зміст пояснювальної записки):

– дослідження та аналіз сучасного стану статистики щодо вартості готельних номерів;

– аналіз прогнозування на основі методології GAM;

– аналіз інструментів та програмного забезпечення для прогнозування;

– дослідження питань трансформації вхідних даних у необхідні формати, заповнення пропусків;

– аналіз методів декомпозиції часових рядів та побудови моделей;

– прогнозування та оцінювання метрик для визначення якості моделей та прогнозування.

5. Перелік графічних матеріалів: презентація.

6. Завдання до спеціальної частини: «Опис основних питань охорони праці, пов'язаних з професійною діяльністю та використанням комп'ютерів для роботи в офісі».

7. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис
Спеціальна частина з охорони праці	Боженко А. Л., викладач	

Керівник роботи канд. техн. наук, доцент Калініна І. О.

(наук. ступінь, вчене звання, прізвище та ініціали)

_____

(підпис)

Завдання прийнято до виконання Пожар Д. П.

(прізвище та ініціали)

_____

(підпис)

Дата видачі завдання « 23 » листопада 2022 р.

КАЛЕНДАРНИЙ ПЛАН
виконання бакалаврської кваліфікаційної роботи

Тема: Інформаційна система прогнозування часових рядів на основі методології GAM

№	Найменування роботи	Початок	Закінчення	Примітки
1	Подання заяви на затвердження теми та керівників БКР	29.10.2022	30.10.2022	Виконано
2	Отримання завдання на виконання БКР	08.11.2022	12.11.2022	Виконано
3	Складання календарного плану роботи на весь період виконання БКР	02.12.2022	02.12.2022	Виконано
4	Отримання завдання на переддипломну практику	15.03.2023	15.03.2023	Виконано
5	Проходження переддипломної практики, збір та аналіз матеріалів до БКР	01.05.2023	14.05.2023	Виконано
6	Розробка звіту з переддипломної практики	15.05.2023	17.05.2023	Виконано
7	Виконання БКР: аналіз сучасного стану прогнозування часових рядів, огляд існуючих технологій, розробка ІС	18.05.2023	16.06.2023	Виконано
8	Попередній захист БКР на засіданні комісії кафедри	29.05.2023	30.05.2023	Виконано
9	Доробка та остаточне оформлення БКР	01.06.2023	17.06.2023	Виконано
10	Подання БКР рецензенту	18.06.2023	19.06.2023	Виконано
11	Подання БКР, її електронної копії та інших документів (відгуку, рецензії) до захисту	20.06.2023	22.06.2023	Виконано
12	Захист БКР перед екзаменаційною комісією (ЕК)	26.06.2023	29.06.2023	Виконано

Розробила студентка Пожар Д. П. _____
(прізвище, ім'я, по батькові студента) (*підпис*)

Керівник роботи канд. техн. наук, доцент Калініна І. О. _____
(*посада, прізвище, ім'я, по батькові)* *(підпис)*

«_08_» _____12_____ 2022 р.

АНОТАЦІЯ

бакалаврської кваліфікаційної роботи студентки групи 402 ЧНУ ім. Петра Могили

Пожар Дарини Петрівни

Тема: «Інформаційна система прогнозування часових рядів на основі методології GAM»

Актуальність: прогнозування часових рядів є важливою задачею в багатьох галузях, включаючи економіку, фінанси, логістику, метеорологію та інші. Також інформаційні системи прогнозування допомагають покращити якість прогнозів та прийняття рішень на основі наявних даних.

Об’єктом роботи є дані вартості готельних номерів за певний період часу.

Предмет роботи – інтелектуальна система прогнозування вартості готельних номерів на основі методології GAM.

Метою роботи є прогнозне оцінювання послідовності спостережень та динаміки змін цін у майбутньому.

Пояснювальна записка складається зі вступу, трьох розділів, висновків та додатків.

У першому розділі розглядається аналіз наявних робіт на тему прогнозування вартості готельних номерів на основі методології GAM .

У другому розділі розглядаються математичні моделі, методи, інформаційні технології, що використовуються для прогнозування вартості готельних номерів.

У третьому розділі описано проєктування моделей та їх програмна реалізація з подальшим прогнозуваннями та їх результатами.

Бакалаврська кваліфікаційна робота містить 84 сторінки, 34 рисунки, 3 таблиці, 40 джерел та 3 додатки.

Ключові слова: прогнозування, часові ряди, машинне навчання, декомпозиція, prophet, GAM, моделі, метрики, R.

ABSTRACT

for bachelor's qualification work of a student of 402 group at Petro Mohyla Black Sea National University

Daryna Pozhar

on topic: "Time series forecasting information system based on GAM methodology"

Relevance: forecasting time series is an important task in many industries, including economics, finance, logistics, meteorology and others. Also forecasting information systems help to improve the quality of forecasts and decision-making based on available data.

The object of work is the cost data of hotel rooms for a certain period of time.

The subject of the work is an intelligent system for predicting the cost of hotel rooms based on the GAM methodology.

The purpose of the work is to forecast the sequence of observations and the dynamics of price changes in the future.

The explanatory note consists of an introduction, four sections, conclusions and appendices.

The first section discusses the analysis of existing works on the topic of forecasting the cost of hotel rooms based on the GAM methodology.

The second section discusses mathematical models, methods, information technologies used to predict the cost of hotel rooms.

The third section describes the design of models and their software implementation with subsequent predictions and their results.

Bachelor qualification paper contains 84 pages, 34 figures, 3 tables, 42 sources and 3 annexes

Keywords: prediction, time series, machine learning, decomposition, prophet, GAM, models, metrics, R.

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ

GAM – Generalized Additive Models

MSE – mean squared error

MAE – mean absolute error

RMSE – root mean squared error

MAPE – mean absolute percentage error

ML – machine learning

DL – deep learning

NLP – natural language processing

USD – United States dollar

ВСТУП

У сучасному світі готельний бізнес є однією з найбільш розвинутою галуззю туристичної індустрії. Зростання попиту на готельні послуги та конкуренція на ринку змушують готелі пристосовуватися до змінних потреб клієнтів, а також зосереджувати увагу на ефективному управлінні своїми ресурсами.

Одним з найважливіших аспектів управління готельним бізнесом є встановлення адекватних цін на готельні номери. Вартість готельних номерів має бути розрахована таким чином, щоб вона відповідала якості послуг, рівню комфорту та забезпечувала готелю достатній рівень прибутковості.

У цьому контексті, використання математичних моделей і методологій для розрахунку вартості номерів є важливим інструментом для готелів. Одним з таких методів є генералізовані аддитивні моделі (Generalized Additive Models, GAM), які дозволяють враховувати нелінійність і взаємозв'язки між різними факторами, впливаючими на вартість готельних номерів.

Метою даної роботи є прогнозне оцінювання послідовності спостережень та динаміки змін цін готельних номерів у майбутньому на основі методології GAM. Дослідження включатиме аналіз факторів, що впливають на вартість номерів, врахування нелінійності та взаємозв'язків, а також розробку математичної моделі на основі GAM. Дана модель дозволить готелям більш точно і адекватно встановлювати ціни на готельні номери з урахуванням різних факторів, таких як сезонність, зручність розташування, рівень послуг та вплив змін на ринку та попит клієнтів. Це дозволить готелям бути більш конкурентоспроможними, забезпечувати оптимальну використання своїх ресурсів та досягати стійкого росту прибутковості.

Виконання даної кваліфікаційної роботи сприятиме подальшому розвитку готельної галузі, спрощенню процесу встановлення цін на готельні номери та забезпеченню готелів зручним та ефективним інструментом для управління

прибутковістю. Результати цього дослідження будуть корисними для готельних підприємств, маркетингових відділів та фахівців з управління готельним бізнесом.

Отже, дана бакалаврська кваліфікаційна робота має важливе значення для розвитку готельної індустрії та пропонує новий підхід до розрахунку вартості готельних номерів на основі методології GAM. Використання генералізованих адитивних моделей дозволяє ураховувати складність взаємозв'язків між різними факторами, такими як сезонність, рівень затовареності, локація готелю, зручності та комфортність номерів, а також інші чинники, що впливають на вартість.

Одним з ключових переваг використання методології GAM є здатність моделі враховувати нелінійність залежностей між факторами та вартістю номерів. Традиційні лінійні моделі не завжди здатні адекватно описати такі складні залежності, тоді як GAM дозволяють врахувати криволінійні та нестабільні залежності, що зустрічаються у готельному бізнесі.

Розробка методики розрахунку вартості готельних номерів на основі методології GAM має практичне застосування. Вона дозволить готелям більш точно визначати оптимальну цінову політику, забезпечуючи належну рентабельність бізнесу. Готелі зможуть адаптувати свої ціни до змін на ринку та змінних умов, що дозволить їм залучати більше клієнтів та збільшувати свою конкурентоспроможність.

Предметом дослідження даної кваліфікаційної роботи є розробка і впровадження нового підходу до розрахунку вартості готельних номерів, що базується на методології GAM. Для досягнення цієї цілі необхідно провести детальний аналіз факторів, що впливають на вартість номерів, побудувати і адаптувати генералізовану адитивну модель, зібрати дані та перевірити ефективність розробленої моделі на реальних даних готельного підприємства.

Очікується, що результати дослідження допоможуть готелям вдосконалити свою стратегію ціноутворення та оптимізувати доходи. Застосування методології GAM дозволить готелям враховувати багатофакторні залежності, а також

адаптувати ціни до змін у попиті та конкурентному середовищі. Покращений розрахунок вартості готельних номерів позитивно вплине на прибутковість готельних підприємств та підвищить їх конкурентоспроможність на ринку.

Крім того, результати цієї роботи можуть мати значення для академічного та дослідницького співтовариства, які зацікавлені в області управління готельним бізнесом, економіки туризму та застосування математичних моделей. Бакалаврська кваліфікаційна робота може слугувати основою для подальших досліджень і вдосконалення методології розрахунку вартості готельних номерів з використанням GAM.

В цілому, розробка методики розрахунку вартості готельних номерів на основі методології GAM є актуальною та значущою для готельної індустрії. Вона сприятиме покращенню управління готельними ресурсами, збільшенню прибутковості готельних підприємств та забезпеченню задоволення потреб клієнтів.

1 АНАЛІЗ ПРОГНОЗУВАННЯ ВАРТОСТІ ГОТЕЛЬНИХ НОМЕРІВ НА ОСНОВІ МЕТОДОЛГІЇ GAM

1.1 Опис предметної сфери

Розрахунок вартості готельних номерів є важливим етапом в готельному бізнесі, який дозволяє встановити ціну за проживання гостей [1]. Цей процес базується на різних факторах, що впливають на вартість кімнати та включають в себе розмаїті аспекти.

Один з ключових факторів, що враховується при розрахунку вартості готельних номерів – це категорія кімнати. Готелі зазвичай мають кілька категорій номерів, таких як стандарт, покращений, люкс, сімейний тощо. Категорія номера залежить від таких факторів, як розмір кімнати, наявність зручностей та комфортні умови проживання. Чим вища категорія номера, тим вища його вартість.

Розташування готелю є іншим важливим фактором, що впливає на вартість готельних номерів. Готелі, розташовані в центральних районах або в престижних локаціях, зазвичай мають вищі ціни порівняно з готелями на околицях міста або в менш популярних районах. Близькість до визначних місць, туристичних атракцій та зручний доступ до інфраструктури можуть також впливати на вартість номера.

Сезонність є ще одним чинником, який впливає на розрахунок вартості готельних номерів. У піковий туристичний сезон або під час проведення важливих заходів, попит на готельні номери зазвичай зростає, що призводить до підвищення цін. У той же час, в менш популярні періоди, коли попит знижується, готелі можуть пропонувати знижки та спеціальні пропозиції, щоб привернути більше клієнтів. Крім того, тривалість проживання також має вплив на розрахунок вартості готельного номера. Готелі можуть пропонувати знижки або спеціальні тарифи для довгострокового проживання, що може знизити загальну вартість номера. Зазвичай чим триваліше проживання, тим менша вартість за одну добу.

Послуги та зручності, що надаються гостям, також враховуються при розрахунку вартості готельних номерів. Наявність таких послуг, як безкоштовний

сніданок, обслуговування номерів, фітнес-центр, басейн, паркування чи бездротовий доступ до Інтернету, може вплинути на загальну ціну номера. Чим більше додаткових послуг надається, тим вища може бути вартість номера [2].

В сучасному світі прогнозування є необхідною складовою планування, оскільки якість прогнозів безпосередньо впливає на прийняття ефективних рішень щодо виробництва, транспортування та персоналу. Воно має стратегічне значення для багатьох компаній, оскільки може впливати на різні сфери діяльності. Прогнозування передбачає коротко-, середньо- та довгострокові прогнози, залежно від поставленої задачі. Хоча деякі події можуть бути точно передбачуваними, наприклад, час сходу сонця, а ось виграш у лотерею є непередбачуваним. Прогнозованість залежить від різних факторів, включаючи розуміння цих факторів, доступність даних, схожість минулого до майбутнього та можливий вплив прогнозів на саму подію. Успішні прогнози враховують реальні моделі та зв'язки, які спостерігаються в історичних даних, однак вони не просто повторюють минулі події, які не повторяться у майбутньому.

Часові ряди постійно знаходяться у стані змін, і прогнозування відображає цю динаміку. При прогнозуванні необхідно враховувати, що середовище буде продовжувати змінюватися у майбутньому, навіть якщо воно є змінним в даний момент. Модель прогнозування повинна відображати ці зміни, а не просто поточний стан речей. Прогнозування націлене на фіксування руху, а не поточне положення. Цитата Авраама Лінкольна підкреслює, що знання про наше місце в даний момент і куди ми йдемо допомагає нам приймати кращі рішення.

При прогнозуванні часових рядів метою є оцінка того, як послідовність спостережень буде розвиватися у майбутньому. Прості методи прогнозування використовують лише інформацію про саму змінну, не враховуючи фактори, що впливають на її поведінку. Вони прогнозують тренд та сезонні зміни, але ігнорують іншу важливу інформацію, таку як маркетингові ініціативи, конкурентна активність, економічні зміни та інше. Прогнозування вимагає використання аналітичних моделей, які враховують усі особливості конкретного прогнозування.

Сучасні методи прогнозування часових рядів дозволяють налаштовувати параметри моделей під конкретний набір даних, що розглядається. Такі системи потребують визначення параметрів моделей, вибору базових методів прогнозування та комбінування методів для досягнення якісних прогнозів.

Ефективність методів прогнозування залежить від наявних даних [3]. Якщо є обмежені або неповні дані – необхідно використовувати якісні методи прогнозування. Кількісне прогнозування може бути успішним, якщо ми маємо доступ до числової інформації про минуле і можемо припустити, що деякі аспекти минулих моделей будуть продовжуватися у майбутньому. Існує широкий спектр кількісних методів прогнозування, розроблених для конкретних дисциплін та цілей. Кожен метод має свої властивості, точність і вартість, які слід враховувати при виборі методу.

Більшість задач прогнозування включають використання часових рядів, які зібрані протягом регулярних інтервалів часу, або перехресних даних, які зібрані в один момент часу. Прогнозування з використанням моделей часових рядів є важливим інструментом для планування, оскільки якість прогнозу безпосередньо впливає на якість прийнятих рішень щодо планування виробництва, транспортування та персоналу. Використання аналітичних моделей прогнозування дозволяє враховувати різні фактори та залежності, що впливають на часові ряди, а також надає можливість налаштування параметрів для досягнення більшої точності в прогнозуванні.

Процес прогнозування включає п'ять основних кроків: визначення проблеми, збір даних, попередній (пошуковий) аналіз, вибір моделей, використання та оцінка моделі прогнозування [4] (див. табл. 1.1).

Таблиця 1.1 – Основні етапи завдання прогнозування

Визначення кроку	Опис кроку
Визначення проблеми	Ретельне визначення проблеми є ключовим етапом прогнозування, і це часто найскладніша частина. Важливо розуміти, як саме будуть використовуватися прогнози, для кого вони потрібні і як функція прогнозування впишеться в організацію, що потребує прогнозів. Прогнозисту необхідно спілкуватися з усіма зацікавленими сторонами, які беруть участь у зборі даних, підтримці баз даних і використанні прогнозів для майбутнього планування. Це вимагає витрати часу та зусиль, але є важливою передумовою для успішного прогнозування.
Збір даних	Завжди потрібні принаймні два типи інформації для прогнозування: (а) статистичні дані і (б) накопичений досвід фахівців, які збирають дані та використовують прогнози. Іноді може бути складно отримати достатньо історичних даних для створення якісної статистичної моделі. Також можуть виникати ситуації, коли старі дані стають менш корисними через структурні зміни в системі, і тоді ми можемо спиратися лише на найновіші дані. Проте варто пам'ятати, що надійні статистичні моделі враховують еволюцію системи і не варто відкидати добрі дані без належної потреби.

Закінчення таблиці 1.1

Визначення кроку	Опис кроку
Попередній (пошуковий) аналіз	Перед початком аналізу завжди рекомендується побудувати графік даних. Це дозволяє виявити послідовні закономірності, наявність суттєвого тренду, сезонність та докази наявності бізнес-циклів. Також важливо виявити можливі викиди даних, які потребують пояснення від експертів. Такий аналіз дозволяє оцінити силу зв'язків між доступними для аналізу змінними. Для цього аналізу розроблено різні інструменти.
Вибір моделей	Вибір найкращої моделі для використання залежить від наявності історичних даних, потужності зв'язків між прогнозованою змінною та пояснювальними змінними, а також від цілей використання прогнозів. Зазвичай проводять порівняльний аналіз двох-трьох потенційних моделей. Кожна модель є штучною конструкцією, що базується на наборі припущень (явних і неявних) і зазвичай містить один або кілька параметрів, які потрібно оцінити за допомогою відомих історичних даних.
Використання та оцінка моделі прогнозування	Після вибору моделі та оцінки її параметрів, модель використовується для складання прогнозів. Оцінка ефективності моделі може бути проведена тільки після отримання даних за прогнозований період. Існує ряд методів для оцінки точності прогнозів. Крім того, є організаційні питання, пов'язані з використанням та виконанням прогнозів. При практичному застосуванні моделі прогнозування виникають практичні питання, такі як робота з відсутніми значеннями та викидами, або робота з короткими часовими рядами.

1.2 Огляд та аналіз наявних публікацій

За останні роки модель генералізованої аддитивної моделі (Generalized Additive Model, GAM) стала популярним підходом для прогнозування за часовими рядами. GAM є гнучким статистичним методом, який може враховувати нелінійність, тренди, сезонність та інші структури в часових рядах. У цьому огляді ми розглянемо деякі з провідних публікацій, в яких використовується модель GAM для прогнозування за часовими рядами.

1. "Time series forecasting with generalized additive models: A review" (2017) - Ця публікація авторства Shenghai Yang та Xiaofeng Shao включає огляд застосування моделі GAM для прогнозування за часовими рядами. Вона досліджує різні варіанти GAM, включаючи моделі з різними функціями впливу, наприклад, лінійними, нелінійними, сезонними і т.д. Робота надає важливі вказівки щодо вибору параметрів та налаштування моделі GAM для прогнозування часових рядів [5].

2. "Forecasting daily electricity demand using generalized additive models" (2018) - В цій статті автори, Pau Fonseca і Casas та Andrea Lázaro, досліджують застосування моделі GAM для прогнозування щоденного попиту на електроенергію. Вони показують, як GAM може бути використана для моделювання сезонної залежності, додаткових факторів впливу та нелінійності в часових рядах попиту. Результати дослідження свідчать про високу точність прогнозів, отриманих за допомогою моделі GAM [6].

3. "Generalized additive models for seasonal forecasting of infectious diseases" (2020) - Ця стаття авторства Sara Tartof, Ashleigh R. Tuite та Джеймса О. Ллойда-Сміта досліджує застосування моделі GAM для сезонного прогнозування [7].

4. "Generalized Additive Models for Time Series Forecasting: An Application to Air Pollution Prediction" (2019) - У цій публікації автори Івано Саліні та Сільвія Спара використовують модель GAM для прогнозування забруднення повітря. Вони досліджують вплив сезонності, тенденції та метеорологічних факторів на рівень

забруднення повітря. Результати показують, що модель GAM здатна добре узгоджуватись з даними та точно прогнозувати майбутні значення забруднення повітря [8].

5. "Forecasting Tourism Demand with Generalized Additive Models" (2020) - У цій статті автори Анна Сантамарія та Мігель Анхель Мартінез-Еспіньо досліджують застосування моделі GAM для прогнозування попиту в туристичній галузі. Вони використовують різні фактори, такі як погода, святкові дні, економічні показники тощо, для побудови моделі прогнозування. Результати показують, що модель GAM є ефективним інструментом для прогнозування туристичного попиту [9].

6. "Generalized Additive Models for Time Series Forecasting: A Bayesian Approach" (2021) - У цій роботі автори Міхал Калінський та Лукашз Бурда розглядають байєсівський підхід до моделювання та прогнозування за допомогою моделі GAM. Вони використовують байєсівський кадр для оцінки параметрів моделі та прогнозування майбутніх значень. Результати показують, що байєсівський підхід до моделі GAM може покращити точність прогнозів у порівнянні з традиційними методами [10].

Загалом, публікації, що використовують модель GAM для прогнозування за часовими рядами, підтверджують ефективність цього підходу в аналізі та прогнозуванні часових рядів. Модель GAM дозволяє враховувати різноманітні структури даних, такі як тренди, сезонність, нелінійність та вплив зовнішніх факторів. Це робить її потужним інструментом для прогнозування в різних галузях, включаючи фінанси, економіку, енергетику, екологію та інші [11].

Оглянуті публікації також підкреслюють важливість належного вибору параметрів та налаштування моделі GAM для досягнення найкращих результатів. Деякі з них досліджують різні функції впливу та специфічні особливості домену дослідження, щоб досягти оптимального прогнозування.

Застосування моделі GAM для прогнозування часових рядів може мати велике значення для прийняття рішень та планування в різних сферах. Вона

дозволяє отримувати точні прогнози та розуміти взаємозв'язки між змінними у часі [12].

Незважаючи на успіхи моделі GAM у прогнозуванні часових рядів, все ще існують виклики, такі як врахування додаткових факторів, управління високою розмірністю даних та робота з недостатніми або неточними даними. Ці аспекти потребують подальшого дослідження та вдосконалення моделі GAM для прогнозування за часовими рядами [13].

Узагальнюючи, модель GAM є потужним інструментом для прогнозування за часовими рядами і її застосування в різних галузях є активною та перспективною галуззю досліджень. Дослідники продовжують вдосконалювати цю модель та її методику застосування для досягнення ще кращих результатів в прогнозуванні за часовими рядами. З використанням моделі GAM можна отримати гнучкі та точні прогнози, а також зрозуміти динаміку та вплив різних факторів на часовий ряд [14].

Одним із основних переваг використання моделі GAM є її здатність автоматично виявляти та враховувати нелінійні залежності між змінними. Це робить її особливо ефективною в ситуаціях, коли простіші моделі, такі як лінійна регресія, не можуть достатньо точно описати поведінку часового ряду [15].

Крім того, модель GAM дозволяє включати в аналіз сезонність, яка є важливою властивістю багатьох часових рядів. Вона може автоматично виявляти та моделювати повторювані патерни, пов'язані з різними сезонами, що дозволяє зробити більш точний прогноз майбутніх значень [16].

Однак, важливо враховувати, що успішність моделі GAM залежить від правильного вибору параметрів та управління ризиком перенавчання. Необхідно ретельно налаштовувати модель, використовувати методи крос-валідації та інші стратегії для оцінки її ефективності [17].

Крім того, в ряді публікацій вказується на необхідність вдосконалення моделі GAM для роботи з великими обсягами даних, врахування залежності між спостереженнями в часі та розв'язання проблеми недостатніх даних.

Усе ж, модель GAM продовжує розвиватись та застосовуватись в широкому спектрі галузей для прогнозування за часовими рядами. З ростом обсягу даних та збільшення обчислювальної потужності, можна очікувати подальше покращення моделі GAM та розширення її застосування [18].

Загалом, модель GAM є важливим інструментом для прогнозування за часовими рядами, який дозволяє враховувати складні структури даних і отримувати точні прогнози. Використання моделі GAM у дослідженнях і практичних застосуваннях вже показало свою ефективність. Однак, додаткові дослідження та розробки в галузі вдосконалення методології моделі GAM для прогнозування за часовими рядами допоможуть розширити її потенціал та зробити її ще більш корисною для практичного використання.

Враховуючи широкий спектр застосувань та активний розвиток досліджень в цій області, можна очікувати, що модель GAM залишиться важливим інструментом прогнозування за часовими рядами і буде продовжувати привертати увагу дослідників і практиків у майбутньому [19].

1.3 Постановка задачі

Для вирішення завдання прогнозування вартості готельних номерів існують певні кроки, які потрібно виконати. По-перше, необхідно завантажити дані і підготувати їх для подальшого аналізу. Далі слід провести візуальну оцінку даних та перевірити їх на наявність викидів або пропусків. Якщо ці проблеми виявлено потрібно їх вирішити.

Наступним кроком є візуалізація відредагованого набору даних. Потім варто оцінити нормальність отриманих значень і розділити набір даних на навчальну та тестову вибірки. Після цього необхідно перевірити дані на наявність нелінійності, автокореляції та нестаціонарності. Також слід здійснити декомпозицію часового ряду і перевірити кожну його компоненту на наявність впливу на дані.

Далі, використовуючи ручні налаштування моделей, слід створити кілька моделей, кожна з яких використовуватиме окрему складову [20]. Наприклад, одна

модель може використовувати тренд, друга - сезон з мультиплікативною складовою, а третя - комбінацію складових. Під час тестування моделей з різними значеннями параметрів слід вибрати кілька найкращих моделей, які найточніше відображають дані вибірки. Знову ж таки, необхідно перевірити залишки моделей.

Після перевірки моделей можна переходити до прогнозування. По-перше, після прогнозування потрібно візуалізувати отримані значення. Після графічної оцінки даних слід провести тести на точність, включаючи основні метрики похибок, такі як MSE, MAE, RMSE, MAPE, а також оцінити покриття. Після цього слід перевірити результати та визначити, наскільки найкращі дві моделі відрізняються від ідеальних значень (значень тестової вибірки). На основі цього можна створити комбіновану модель, яка буде додатково перевірена та використана для прогнозування.

Після завершення всіх вищезазначених етапів слід зробити висновки щодо проведеної роботи і визначити, яка з моделей показала себе найкраще. Також можна проаналізувати роботу цієї моделі на вибірці. Якщо необхідно, можна вдосконалити модель шляхом модифікації або комбінування з іншою моделлю. Загалом, використання різних підходів до розробки прогнозуючих моделей в єдиній інформаційній системі дозволяє отримати ефективний прогноз вартості готельних номерів.

Висновки до розділу 1

У рамках аналізу прогнозування вартості готельних номерів на основі методології GAM було встановлено, що цей підхід є ефективним і потенційно корисним для прогнозування вартості готельних номерів.

Методологія GAM дозволяє моделювати нестаціонарність та неоднорідність в даних шляхом використання гладких функцій залежності. Вона комбінує переваги лінійних та нелінійних моделей, дозволяючи враховувати як лінійну, так і нелінійну залежність між змінними.

Прогнозування вартості готельних номерів на основі методології GAM дозволяє отримати більш точні результати порівняно з традиційними методами. Це може бути корисно для готельних власників і менеджерів, які бажають оптимізувати свою стратегію ціноутворення, збільшити доходи та покращити задоволеність клієнтів.

Враховуючи всі переваги методології GAM в аналізі прогнозування вартості готельних номерів, рекомендується готелям розглянути застосування цього підходу для покращення своєї стратегії управління, ефективного ціноутворення та досягнення більшого успіху на ринку готельних послуг.

Прогнозування в сучасному світі є надзвичайно важливою задачею у плануванні бізнес-процесів. Від якості прогнозування залежить ефективність прийнятих рішень щодо планування виробництва, транспортування та управління персоналом. Прогнозування є невід'ємною складовою прийняття стратегічних рішень, оскільки його важлива роль проявляється у багатьох сферах діяльності компаній.

Сучасні методи створення аналітичних моделей прогнозування часових рядів дозволяють налаштовувати параметри кожної моделі залежно від конкретного набору даних. Це свідчить про необхідність системного підходу при розробці інформаційної системи прогнозування, яка передбачає комбінування кількох прогнозних моделей для отримання якісних прогнозів. Такі системи вимагають точного визначення параметрів моделей, базових методів прогнозування, а також вибору відповідної комбінації методів для вирішення проблеми.

Отже, можна зробити висновок, що прогнозування виступає ключовим елементом у плануванні бізнес-процесів, а розробка систем прогнозування потребує уважного врахування параметрів моделей та вибору оптимальних методів для досягнення якісних прогнозів.

2 МАТЕМАТИЧНІ МОДЕЛІ, МЕТОДИ, ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ, ЩО ВИКОРИСТОВУЮТЬСЯ ДЛЯ ПРОГНОЗУВАННЯ ВАРТОСТІ ГОТЕЛЬНИХ НОМЕРІВ

2.1 Машинне навчання

Машинне навчання (ML) прагне автоматично вивчати суттєві зв'язки та шаблони на основі прикладів та спостережень [21-22]. Прогрес у галузі ML призвів до значного зростання інтелектуальних систем з людськими когнітивними можливостями, які проникають у різні сфери нашого ділового та особистого життя. Вони формують мережеву взаємодію на електронних ринках шляхом використання різних методів, що дозволяють компаніям покращувати процеси прийняття рішень для збільшення продуктивності, залученості та задоволення співробітників. Крім того, системи помічників, які можна навчити та адаптувати до індивідуальних уподобань користувачів, роблять наше повсякденне життя більш зручним. Навіть торгові агенти здатні змінювати традиційні ринки фінансової торгівлі за допомогою інноваційних підходів, що використовують методи ML.

Крім розкрученого зовнішнього вигляду, науковці та професіонали мають необхідність глибокого розуміння основних концепцій, процесів та викликів, пов'язаних з впровадженням такої технології. На цьому тлі ідея полягає в передачі розуміння машинного навчання та глибокого навчання (DL) в контексті електронних ринків. Це дозволить спільноті отримати користь від цих технологічних досягнень, будь то для аналізу великих обсягів даних, зібраних у цифрових екосистемах, або для розробки нових інтелектуальних систем для електронних ринків. Щоб забезпечити глибоке розуміння галузі, необхідно чітко розрізняти кілька відповідних термінів та концепцій один від одного. (див. рис. 2.1) [23].

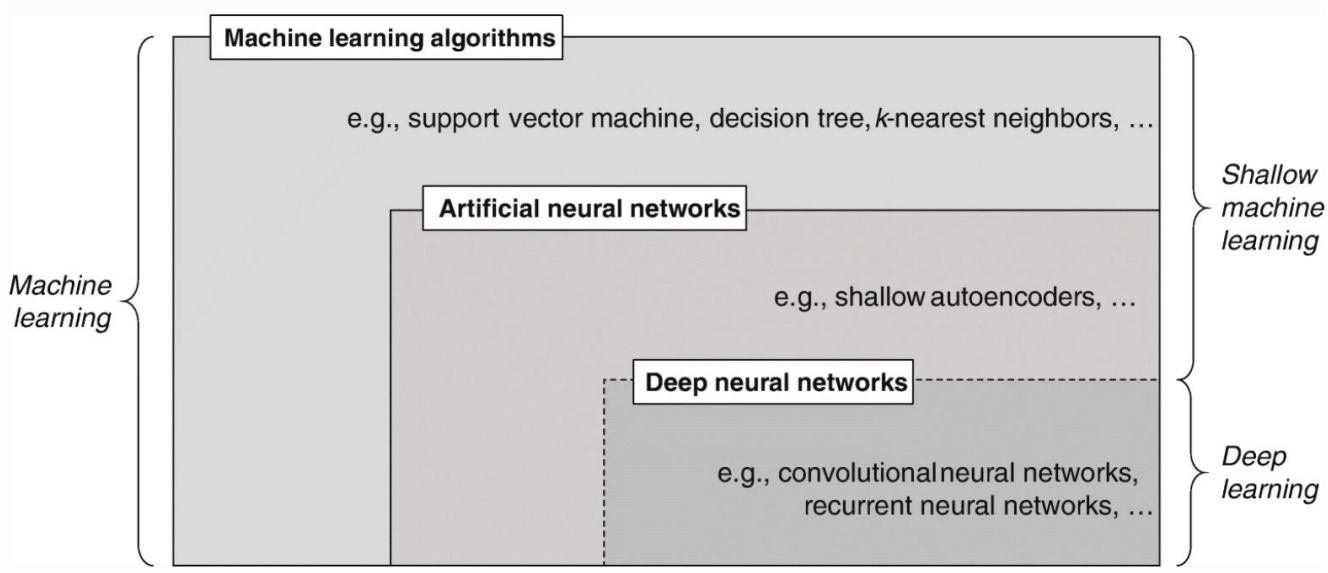

Рисунок 2.1 – Діаграма Венна концепцій і класів машинного навчання

У широкому розумінні штучний інтелект охоплює всі техніки, які дозволяють комп'ютерам імітувати людську поведінку і здійснювати прийняття рішень на рівні, що перевищує людські здібності, для розв'язання складних завдань самостійно або з мінімальною людською участю [24]. Це відноситься до різних ключових проблем, таких як представлення знань, мислення, навчання, планування, сприйняття та комунікація, і використовує різноманітні інструменти і методи (наприклад, кейс-метод, правила базових систем, генетичні алгоритми, нечіткі моделі, мультиагентні системи) [25].

Первинні дослідження в галузі штучного інтелекту переважно зосереджувалися на жорстко закодованих висловлюваннях у формальних мовах, які потім комп'ютер міг автоматично обчислювати за допомогою правил логічного виведення. Цей підхід відомий як підхід на основі бази знань. Однак така парадигма стикається з деякими обмеженнями, оскільки людям часто складно пояснити всі свої неявні знання, необхідні для виконання складних завдань [26].

Машинне навчання вирішує ці обмеження, вносячи нові можливості. Узагальнюючи, ML означає, що продуктивність комп'ютерних програм може покращуватись з досвідом, набутим у виконанні певного класу завдань і досягненні

ефективних результатів [27-29]. Його головна мета полягає в автоматизації побудови аналітичної моделі для вирішення когнітивних завдань, таких як виявлення об'єктів або машинний переклад природними мовами. Це досягається за допомогою ітеративного вивчення вхідних даних, пов'язаних з конкретною проблемою, що дозволяє комп'ютерам виявляти приховані концепції та складні шаблони без явного програмування [30]. Особливо в сферах, пов'язаних з великими обсягами даних, таких як класифікація, регресія та кластеризація, ML демонструє високу застосовність. Шляхом аналізу попередніх обчислень та виявлення закономірностей у великих базах даних він допомагає знаходити надійні та повторювані рішення. Саме з цієї причини алгоритми ML успішно використовуються у багатьох галузях, таких як виявлення шахрайства, оцінка кредитоспроможності, аналіз найкращих наступних пропозицій, розпізнавання мови та зображень або обробка природної мови (NLP).

Виходячи з поставленої проблеми та наявних даних, ми можемо виділити три типи ML: навчання з вчителем, навчання без вчителя та навчання з підкріпленням (див. табл. 2.1).

Таблиця 2.1 – Огляд типів машинного навчання

Тип	Опис
Навчання з вчителем	Навчання з учителем передбачає використання навчального набору даних, який містить приклади вхідних даних разом з мітками або цільовими значеннями для вихідних даних. Наприклад, це може включати прогнозування кількості активних користувачів, які підписалися на ринкову платформу протягом місяця (це вважається цільовою змінною або y), з використанням різних вхідних характеристик, таких як кількість проданих продуктів або позитивні відгуки користувачів (часто позначаються як вхідні функції або змінні

Продовження таблиці 2.1

Тип	Опис
	х). Потім пари вхідних і вихідних даних з навчального набору використовуються для налаштування параметрів моделі машинного навчання. Після успішного навчання модель може бути використана для прогнозування цільової змінної у на основі нових або невидимих точок вхідних даних х. Залежно від типу задачі навчання з учителем, можна виділити два основних варіанти: регресія, де передбачається числове значення (наприклад, кількість користувачів), і класифікація, де результат передбачення належить до певної категорії (наприклад, "глядачі" або "покупці»).
Навчання без вчителя	Навчання без учителя відбувається тоді, коли система навчання має виявляти шаблони і структуру в навчальних даних без наявності попередніх міток або специфікацій. В такому випадку навчальні дані складаються лише зі змінних х з метою виявлення цікавих інформаційних взаємозв'язків, таких як групи елементів, що мають спільні властивості (відомі як кластеризація), або проектування даних з великої розмірності в нижчу (відоме як зменшення розмірності). Прикладом неконтрольованого навчання на електронних ринках є використання методів кластеризації для групування клієнтів або ринків у сегменти з метою більш цілеспрямованого спілкування з цільовою аудиторією.
Навчання з підкріпленням	У системі навчання з підкріпленням, замість надання пар вхідних і вихідних даних, ми визначаємо поточний стан системи, встановлюємо ціль, надаємо перелік доступних дій та обмежень середовища для результатів цих дій, і дозволяємо моделі машинного навчання

Закінчення таблиці 2.1

Тип	Опис
	експериментувати та навчатися самостійно шляхом проб і помилок з метою максимізації отриманої винагороди. Моделі навчання з підкріпленням з успіхом використовуються в закритих середовищах, таких як ігри, але також є застосовними для багатоагентних систем, наприклад, електронних ринків.

2.2 Статистичний аналіз

Статистичний аналіз є процесом, який включає збір та аналіз даних з метою виявлення закономірностей і тенденцій [31]. Він є важливою складовою аналітики даних і використовується для усунення упередженості в оцінюванні даних за допомогою числового аналізу. Цей метод допомагає здобути інтерпретації досліджень, розробити статистичні моделі і здійснити планування опитувань та досліджень. Статистичний аналіз може бути застосований в різних ситуаціях, таких як збір інтерпретацій досліджень, статистичне моделювання або планування опитувань та досліджень. Він також є незамінним інструментом для бізнес-аналітики, особливо при роботі з великими обсягами даних.

Статистичний аналіз є науковим інструментом, який використовується для збору та аналізу великих обсягів даних з метою виявлення загальних закономірностей і тенденцій, а також перетворення їх на значиму інформацію. Він є ключовим інструментом аналізу даних, який допомагає робити важливі висновки з необроблених та неструктурованих даних. Завдяки статистичному аналізу можна зробити висновки, що полегшують прийняття рішень та допомагають підприємствам прогнозувати майбутні тенденції на основі минулих даних. Він представляє собою науку збору та аналізу даних для виявлення тенденцій і закономірностей і перетворення їх на поняття. Статистичний аналіз вимагає роботи

з числовими даними та широко застосовується підприємствами та організаціями для отримання важливої інформації

Нижче наведено 6 типів статистичного аналізу (див. табл. 2.2) [32]:

Таблиця 2.2 – Типи статистичного аналізу

Тип	Опис
Описовий аналіз	Описовий статистичний аналіз включає процес збору, інтерпретації, аналізу та узагальнення даних з метою їх подання у вигляді діаграм, графіків і таблиць. Його основна мета полягає не в зробленні висновків, а в тому, щоб складні дані стали доступними для читання та зрозуміння.
Інференційний аналіз	Інференційний статистичний аналіз є невід'ємною складовою наукового дослідження та прийняття обґрунтованих рішень на основі аналізу набору даних. Він використовується для отримання значущих висновків та узагальнень, що дозволяють зрозуміти зв'язок між різними змінними та робити прогнози для всієї цільової популяції.
Прогнозний аналіз	Прогнозний статистичний аналіз представляє собою вид статистичного аналізу, що здійснює детальний розгляд даних з метою виявлення минулих тенденцій та прогнозування майбутніх подій на їх основі. Цей підхід надається за допомогою розумових алгоритмів, основаних на машинному навчанні, аналізу даних, моделюванні даних і штучному інтелекту, що дозволяє проводити комплексний статистичний аналіз накопичених даних. Завдяки цьому, прогнозний статистичний аналіз дозволяє отримати цінні

Закінчення таблиці 2.2

Тип	Опис
	прогнози та визначити потенційні ризики та можливості на основі наявних даних.
Наказовий аналіз	Наказовий аналіз проводить аналіз даних і призначає найкращий курс дій на основі результатів. Це тип статистичного аналізу, який допомагає прийняти обґрунтоване рішення.
Дослідницький аналіз даних	Дослідницький аналіз схожий на інференційний аналіз, але відмінність полягає в тому, що він включає дослідження невідомих асоціацій даних. Він аналізує потенційні зв’язки в даних.
Причинно-наслідковий аналіз	Причинно-наслідковий статистичний аналіз — це спеціальний підхід до аналізу даних, який зосереджується на встановленні зв'язку між причинами та наслідками у недопрацьованих даних. Його основна мета полягає у виявленні причин, які призводять до певних явищ, а також у визначенні впливу цих явищ на інші змінні. Простими словами, цей методологічний підхід дозволяє з'ясувати, чому певні події відбуваються, і розуміти, як вони впливають на інші аспекти аналізу. Причинно-наслідковий статистичний аналіз використовується компаніями та організаціями для встановлення причин провалів або невдач. Він надає можливість глибоко проаналізувати дані та ідентифікувати фактори, які призводять до небажаних результатів. Це дозволяє підприємствам прийняти необхідні заходи для усунення виявлених причин і покращення своєї продуктивності або ефективності.

Статистичний аналіз володіє низкою переваг, які роблять його справді корисним інструментом для людства, незалежно від його масштабу та сфери застосування. Нижче перераховано лише кілька причин, чому варто розглянути інвестування в статистичний аналіз:

- допомога визначити фінансові показники: завдяки статистичному аналізу ви зможете зрозуміти щомісячний, щоквартальний або щорічний прибуток від продажів і витрат. Це надає вам необхідну основу для прийняття добре обґрунтованих та обізнаних рішень;
- обґрунтовані рішення: статистичний аналіз забезпечує вам необхідні дані і інформацію для прийняття обґрунтованих рішень. Ви можете підтвердити свої гіпотези, провести дослідження і засновані на даних рішення, що допоможе вам досягти бажаних результатів;
- виявлення причин та виправлення проблем: статистичний аналіз дозволяє виявляти причини проблем або збоїв і розробляти стратегії для їх виправлення. Наприклад, він може допомогти вам виявити причину зростання загальних витрат і зменшити непотрібні витрати;
- розуміння ринку та розробка маркетингових стратегій: за допомогою статистичного аналізу ви можете провести детальний аналіз ринку і розробити ефективну стратегію маркетингу та продажів. Ви отримаєте цінну інформацію про споживачів, конкурентів та потенційні можливості;
- підвищення ефективності процесів: статистичний аналіз допомагає виявляти недоліки та оптимізувати різні процеси. Ви можете використовувати дані та статистику для ідентифікації ефективних підходів, зменшення часу, зусиль і ресурсів, необхідних для досягнення певних цілей. Це дозволяє вам збільшити продуктивність та ефективність вашої діяльності, забезпечуючи оптимальне використання ресурсів і досягнення найкращих результатів.

Таким чином, статистичний аналіз володіє великим потенціалом для поліпшення діяльності індивідів та організацій. Він допомагає зрозуміти причинно-наслідкові зв'язки, забезпечує об'єктивні дані для прийняття рішень і виявлення

проблем. Інвестування в статистичний аналіз може стати важливою складовою успіху і забезпечити конкурентну перевагу в динамічному світі сучасних технологій та бізнесу.

2.3 Середовище розробки

Для вирішення поставленої задачі було обрано середовище розробки RGui, що базується на мові програмування R. R є мовою програмування та програмним середовищем для статистичних обчислень, аналізу та візуалізації даних. У розробці R значний вплив мали мови програмування S з семантикою, успадкованою від Scheme. Назва R походить від перших літер імен засновників - Роса Іхаки та Роберта Джентлмена, які працювали в Оклендському Університеті в Новій Зеландії.

Використання статистичного аналізу, заснованого на R [33], має свої переваги:

– доступність різних методів та функцій статистичного аналізу дозволяє здійснювати широкий спектр досліджень і виконувати різноманітні аналітичні завдання;

– безкоштовна природа R зменшує витрати на придбання програмного забезпечення, що робить його доступним для всіх користувачів;

– можливість внесення змін у R дає можливість адаптувати його під конкретні потреби дослідника або організації;

– активна спільнота розробників R забезпечує постійне оновлення та підтримку програмного середовища;

– великий вибір додаткових пакетів розширює функціональність R і надає можливість виконувати спеціалізовані завдання і аналізи;

– командний рядок R дозволяє повну гнучкість і контроль над аналітичним процесом, а також забезпечує повторюваність результатів та документованість аналізу.

R надає значні можливості для проведення статистичного аналізу, включаючи лінійну і нелінійну регресію, стандартні статистичні тести, аналіз часових рядів, кластерний аналіз та багато іншого. Використання додаткових функцій і пакетів, які доступні на Comprehensive R Archive Network, що дозволяє легко розширювати можливості R.

Інші представники тієї чи іншої сфери також мають багато з вищевказаних особливостей середовища розробки/мови програмування. Тому важливо вказати саме ті переваги, які відрізняють R від усіх інших конкурентів. Перш за все, R спеціалізується на статистичному аналізі, що спрощує роботу з різними видами числової інформації. Наприклад, відсутність необхідності обробляти базові завдання з масивами за допомогою циклів у R дозволяє безпосередньо працювати зі стовбцями та рядками даних. По-друге, R є сучасною мовою програмування з великою спільнотою користувачів. Існують мови, які спеціалізуються на математичній та статистичній обробці даних, але вони застарілі або не здатні використовувати сучасні засоби обробки типів наборів даних. По-третє, R має потужну графічну базу, яка дозволяє ефективно візуалізувати дані та їх складові. Якщо вбудованих можливостей недостатньо, в R є безкоштовні бібліотеки, які не тільки посилюють його функціональність, але також забезпечують зручність візуалізації в разі потреби.

Саме завдяки такому набору характеристик, було вирішено скористатися середовищем та мовою R для виконання поставленої задачі. В наступних розділах буде приділено більше уваги бібліотекам, які дозволяють зручно та якісно налаштовувати моделі для прогнозування. Вибір цих бібліотек також базується на сучасних методах обробки та трансформації даних, включаючи tidyverse. Перетворення даних у цей формат вимагає певної попередньої роботи, проте ці зусилля оправдовуються у довгостроковій перспективі. Завдяки чітким даним та акуратним інструментам, які надають пакети tidyverse, витрати часу на подальше пересування даних з одного формату в інший зменшуються, що дає змогу більше часу приділяти аналітичним запитанням.

2.4 Математичні моделі

Загальна лінійна модель є важливою моделлю в прикладних і соціальних дослідженнях для багатьох застосувань статистичного аналізу. Статистичні моделі забезпечують математичну основу для інтерпретації та дослідження параметрів, а також визначають ролі та порівняльну важливість різних змінних для конкретних процесів [34]. На рис. 2.2 представлено різні доступні моделі на основі регресії, що підкреслює їх основні переваги порівняно з іншими моделями.

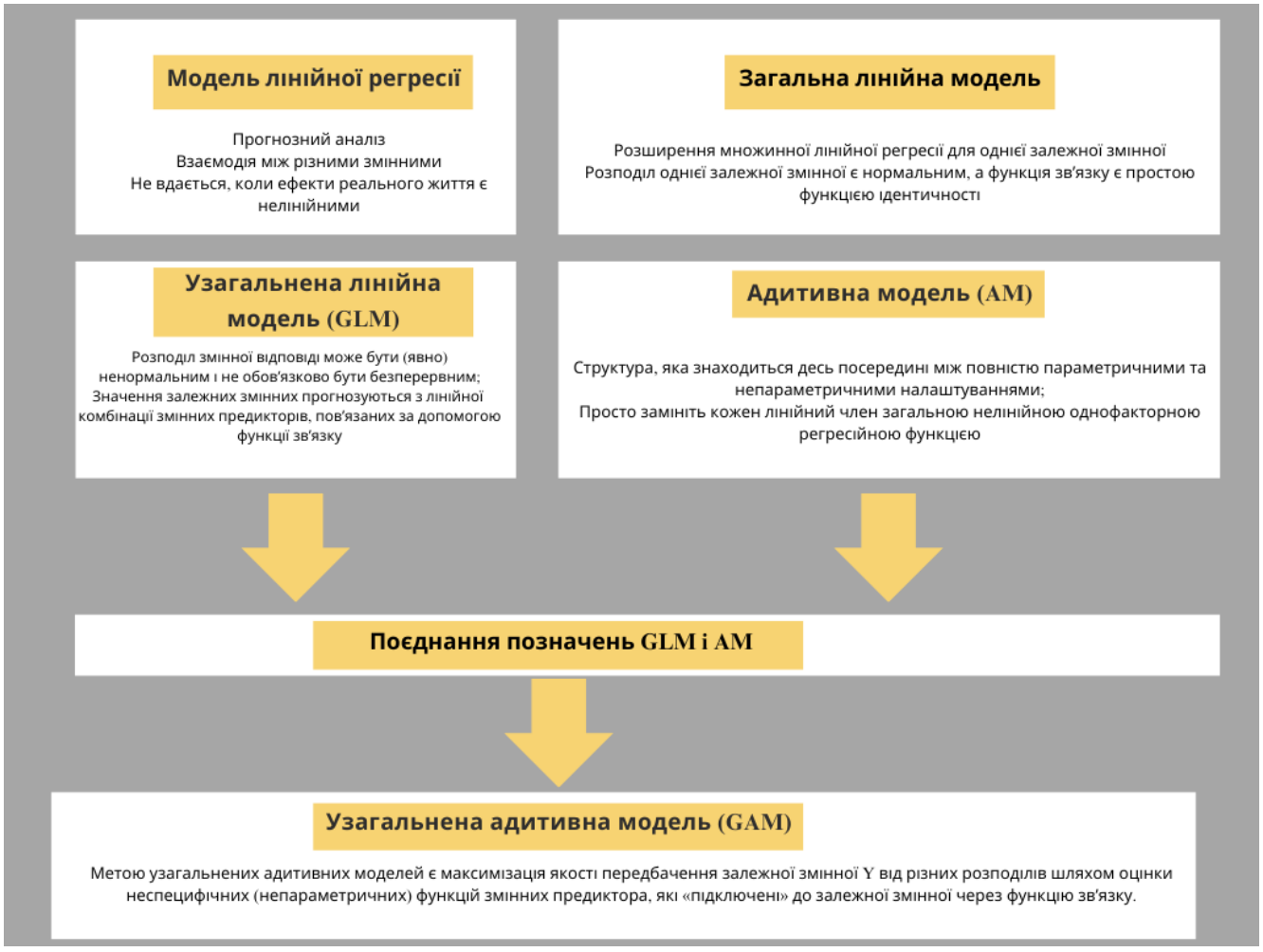

Рисунок 2.2 – Загальний огляд і зв'язок між моделями регресії

Можна узагальнити математичні підходи множинної регресії, перетворивши їх у загальні лінійні моделі та уточнюючи їх відповідно [35-37]. Наприклад,

лінійний метод найменших квадратів використовується для прогнозування залежної змінної Y на основі набору предикторів або змінних X у лінійній регресії.

Іншими словами, це означає передбачення найкращої оцінки для предикторів і адитивної моделі у випадку невизначеної моделі. Непараметрична функція предиктора визначається як заміщення одного коефіцієнта для кожної змінної. Узагальнення математичних підходів множинної регресії можна перетворити у загальні лінійні моделі та далі уточнювати їх. Зокрема, лінійний метод найменших квадратів застосовується для прогнозування залежної змінної Y на основі набору предикторів або змінних X у лінійній регресії.

Іншими словами, метою є передбачення найкращої оцінки для предикторів і адитивної моделі в контексті невизначеної моделі; при цьому непараметрична функція предиктора визначається як заміщення одного коефіцієнта для кожної змінної.

GAM (генералізовані аддитивні моделі) є важливим доповненням до загальних лінійних моделей [38-39]. Інші дослідження показали, що GLM (загальні лінійні моделі) з лінійним предиктором може взаємодіяти з сумою гладких функцій коваріатів [40]. GAM надає структуру для узагальнення загальної лінійної моделі, дозволяючи додавання адитивних нелінійних функцій змінних.

Однією з переваг GAM є можливість обмежити помилку передбачення залежної змінної Y для різних розподілів шляхом оцінки гладких функцій, які зв'язані за допомогою функції зв'язку зі змінною.

GAM надає гнучку специфікацію відповіді, визначаючи модель у термінах гладкої функції, яка замінює детальні параметричні зв'язки на коваріатах. Ця гнучкість і доцільність досягаються за рахунок використання гладких функцій у стандартному форматі та вибору рівня гладкості.

Однією з недавніх пропозицій є модель Prophet, яка доступна через пакет fable.prophet. Ця модель була розроблена компанією Facebook (S. J. Taylor & Letham, 2018) спочатку для прогнозування щоденних даних з урахуванням тижневої та річної сезонності, а також святкових ефектів. Пізніше вона була

розширена для врахування інших типів сезонності. Модель найкраще справляється з часовими рядами, які мають виражену сезонність та кілька сезонів історичних даних.

Prophet можна вважати нелінійною регресійною моделлю виду

$$yt = g(t) + s(t) + h(t) + \varepsilon t, \qquad (2.1)$$

де $g(t)$ – описує кусково-лінійний тренд (або «член зростання»);

$s(t)$ – описує різні сезонні закономірності;

$h(t)$ – фіксує ефект відпустки;

εt – є членом помилки білого шуму.

Якщо не вказано прямо, автоматично вибираються вузли (або точки зміни) для кусково-лінійного тренду. За бажанням можна застосувати логістичну функцію для встановлення верхньої межі тенденції. Сезонна складова складається з Фур'є-членів, що відповідають відповідним періодам. За замовчуванням, для річної сезонності використовується порядок 10, а для тижневої сезонності - порядок 3. Святкові ефекти вводяться у формі простих фіктивних змінних.

Для оцінювання моделі використовується байєсівський підхід, щоб забезпечити автоматичний вибір точок зміни та інших характеристик моделі.

Висновки до розділу 2

Машинне навчання (ML) має за мету автоматичне вивчення суттєвих залежностей та шаблонів на основі прикладів і спостережень. Прогрес в галузі ML призвів до зростання інтелектуальних систем з людськими когнітивними здібностями, які проникають в наше бізнесове та особисте життя. Вони формують мережеву взаємодію на електронних ринках усіма можливими способами, допомагаючи компаніям поліпшувати процеси прийняття рішень для підвищення продуктивності, залученості та збереження співробітників. Такі системи помічників можна навчити адаптуватись до індивідуальних вподобань користувачів, а торгові агенти змінюють традиційні фінансові ринки.

Статистичний аналіз є науковим інструментом, який допомагає збирати та аналізувати великі обсяги даних з метою виявлення загальних закономірностей, тенденцій і отримання значущої інформації. У простих словах, статистичний аналіз допомагає робити важливі висновки на основі необроблених і неструктурованих даних.

R має значні можливості для здійснення статистичного аналізу, таких як лінійна і нелінійна регресія, класичні статистичні тести, аналіз часових рядів (серій), кластерний аналіз і багато іншого. З використанням додаткових функцій і пакетів, доступних на сайті Comprehensive R Archive Network, R може легко розширюватись.

GAM має перевагу в тому, що вона дозволяє обмежити помилку передбачення залежної змінної Y від різних розподілів шляхом оцінки неспецифічних функцій, які пов'язані залежністю за допомогою функції зв'язку.

GAM дозволяє гнучко визначати відповідь, моделюючи її у термінах гладкої функції замість детальних параметричних зв'язків на коваріатах. Ця гнучкість та ефективність досягаються за рахунок представлення гладких функцій у стандартному шаблоні та вибору рівня гладкості.

3 МОДЕЛЮВАННЯ ТА ПРОГРАМНА РЕАЛІЗАЦІЯ МОДЕЛЕЙ І ПРОГНОЗІВ. ДОСЛІДЖЕННЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ

3.1 Аналіз вхідного набору даних

Набір даних представлений вартістю готельних номерів за 2012-2013 роки. Завантаживши дані, можна побачити, що вони представляють собою часовий ряд, визначений 3 змінними: id номера, дата та час, ціна номера в USD.

Першим кроком у прогнозуванні є підготовка даних у правильному форматі. Цей процес може включати завантаження даних, визначення відсутніх значень, фільтрацію часових рядів та інші завдання попередньої обробки.

Багато моделей мають різні вимоги до даних; деякі вимагають, щоб ряди були в порядку часу, інші вимагають відсутності пропущених значень. Перевірка ваших даних є важливим кроком для розуміння їх характеристик, і її слід завжди робити перед оцінкою моделей.

Після завантаження та трансформації в формат Дати та часу перетворюємо розглянутий набір даних в формат tsibble за допомогою функції as_tsibble ().

Мінлива *price_usd* - це вартість room (в USD), зазначена в день та час date_time. Таблиця room містить тільки один часовий ряд. Тому при перетворенні цієї таблиці в формат tsibble аргументу key було присвоєно значення NULL - таким чином, ми повідомили програмі, що в таблиці немає групуючих змінних.

Зобразимо на малюнку динаміку вартості room в розглянутий період часу (див. рис. 3.1).

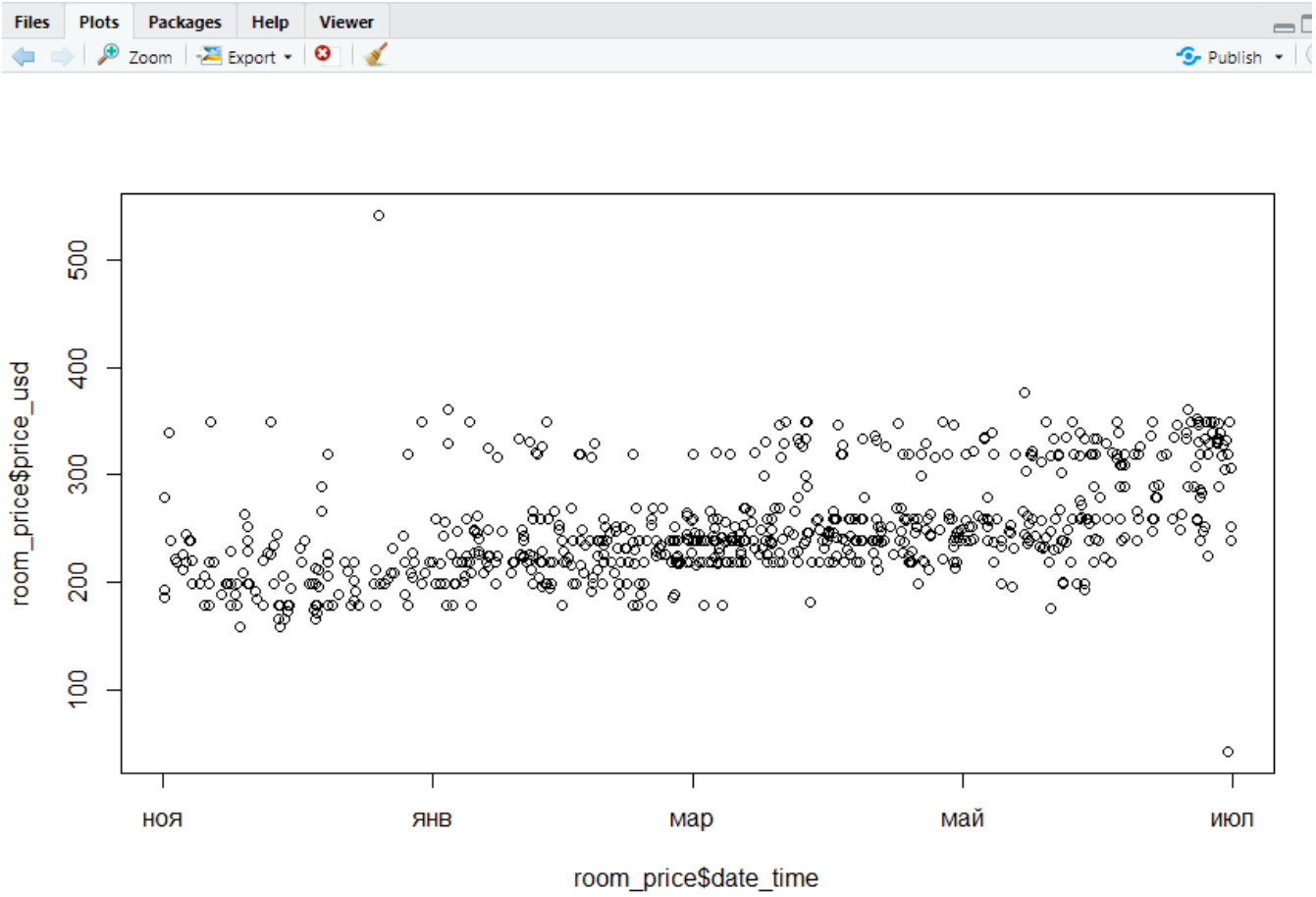

Рисунок 3.1 – Динаміка вартості

Існує кілька методів декомпозиції часових рядів. Одним з найбільш широко використовуваних є розкладання на тренд і сезонну складову за допомогою локальної поліноміальної регресії. Найпростіше з цим методом можна розібратися, застосувавши його до конкретних даних (див. рис. 3.2).

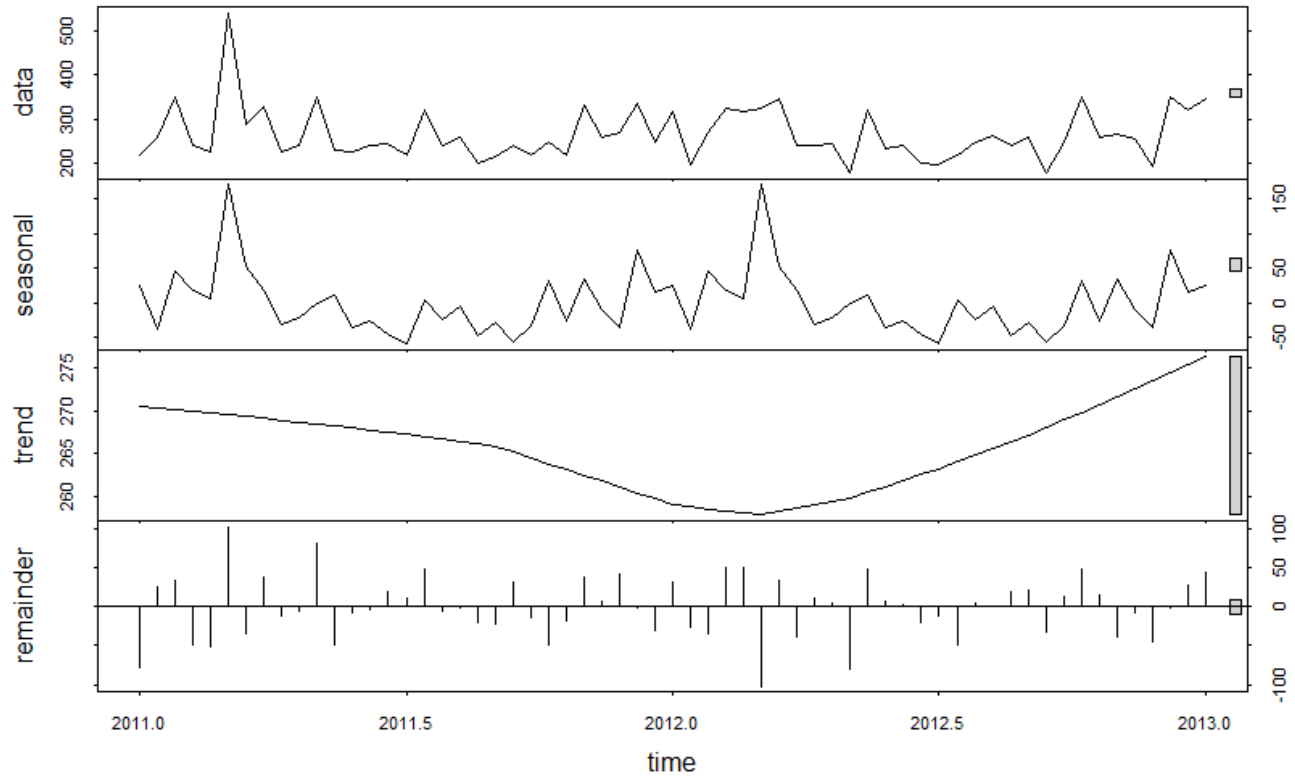

Рисунок 3.2 – Декомпозиція

На рис. 3.2, виділені за допомогою методу STL, компоненти часового ряду показані на трьох нижніх графіках (trend - тренд, seasonal - компонента сезонності, remainder - залишки). Якщо їх підсумувати, то отримаємо вихідний ряд y (наведено на верхньому графіку - data).

З наведених графіків видно, що вплив trend, seasonal та remainder не є одинаковим, це свідчить про те, що часовий ряд нестаціонарний.

Далі ми переходимо до тестування часового ряду.

Перевірка на нестаціонарність виконується трьома тестами: Розширений тест Дікі-Фуллера, Тест KPSS на рівень стаціонарність, Тест Філіпса-Перрона на одиничний корінь

3.1.1 Дікі-Фуллера (ADF)

У статистиці та економетриці розширений тест Дікі-Фуллера (ADF) перевіряє нульову гіпотезу про наявність одиничного кореня у вибірці часових

рядів. Альтернативна гіпотеза відрізняється залежно від того, яка версія тесту використовується, але, як правило, це стаціонарність або тенденція-стаціонарність. Це доповнена версія тесту Дікі-Фуллера для більшого і складного набору моделей часових рядів.

```
> adf.test(tsData[,1])

        Augmented Dickey-Fuller Test

data:  tsData[, 1]
Dickey-Fuller = -3.2798, Lag order = 3, p-value =
0.08327
alternative hypothesis: stationary
```

Рисунок 3.3 – Розширений тест Дікі-Фуллера

Якщо значення p> 0,05 під час тесту ADF (доповнений Дікі-Фуллер) часового ряду, тоді серія називається нестаціонарною і приймає гіпотезу NULL. Якщо значення p <= 0,05, воно відхиляє гіпотезу NULL, яка символізується як H0, і вона називається стаціонарною, коли дані не мають одиничного кореня. Альтернативна гіпотеза символізується як H1. З результату виконання тесту можна сказати, що часовий ряд нестаціонарний(так як p > 0.05).

3.1.2 KPSS

В економетриці тести Квятковського – Філліпса – Шмідта – Шіна (KPSS) використовуються для перевірки нульової гіпотези про те, що спостережуваний часовий ряд є стаціонарним навколо детермінованої тенденції (тобто тенденції, стаціонарної) проти альтернативи одиничному кореню.

```
> kpss.test(tsData[,1])

        KPSS Test for Level Stationarity

data:  tsData[, 1]
KPSS Level = 0.10748, Truncation lag parameter =
3, p-value = 0.1
```

Рисунок 3.4 – Тест Квятковського – Філліпса – Шмідта – Шіна

З результату виконання тесту можна сказати, що часовий ряд нестаціонарний(так як p > 0.05).

3.1.3 PP

У статистиці тест Філіпса – Перрона (названий на честь Пітера С. Б. Філліпса та П'єра Перрона) є одиничним тестом. Тобто він використовується для аналізу часових рядів для перевірки нульової гіпотези про те, що часовий ряд інтегрований із порядку.

```
> pp.test(tsData[,1])

        Phillips-Perron Unit Root Test

data:  tsData[, 1]
Dickey-Fuller Z(alpha) = -61.293, Truncation lag
parameter = 3, p-value = 0.01
alternative hypothesis: stationary
```

Рисунок 3.5 – Тест Філіпса – Перрона

З результату виконання тесту можна сказати, що часовий ряд стаціонарний (так як p < 0.05).

Встановлено великий вплив тренду та нестаціонарність часового ряду.

3.2 Моделювання GAM

Припустимо, що нам необхідно зробити прогноз вартості номера на наступні 90 днів. Для зниження дисперсії виконаємо логарифмування значень вартості номера price_usd. Розіб'ємо вихідну вибірку на навчальну (всі спостереження за винятком останніх 90 днів) і перевірочну (останні 90 днів).

```
> room_train <- room_price %>%
+ mutate(price_usd=log(price_usd)) %>%
+ slice(1:(n()-90)) %>%
+ as.data.frame()

room_test <- room_price %>%
+ mutate(price_usd=log(price_usd), date_time=as.Date(date_time)) %>%
+ tail(90) %>%
+ as.data.frame()

> room_train %>%
+ ggplot(., aes(date_time, price_usd))+geom_line()+theme_minimal()

> room_test %>%
+ ggplot(., aes(date_time, price_usd))+geom_line()+theme_minimal()
```

Підгонку моделі будемо виконувати на навчальних даних (room_train). Перевірочна вибірка (room_test) стане в нагоді в самому кінці процесу моделювання, щоб з'ясувати наскільки наші очікування щодо якості обраної оптимальної моделі відповідають дійсності.

На рис. 3.6 та 3.7 відображена вибірка room після розбиття на тестову та перевірочну вибірку.

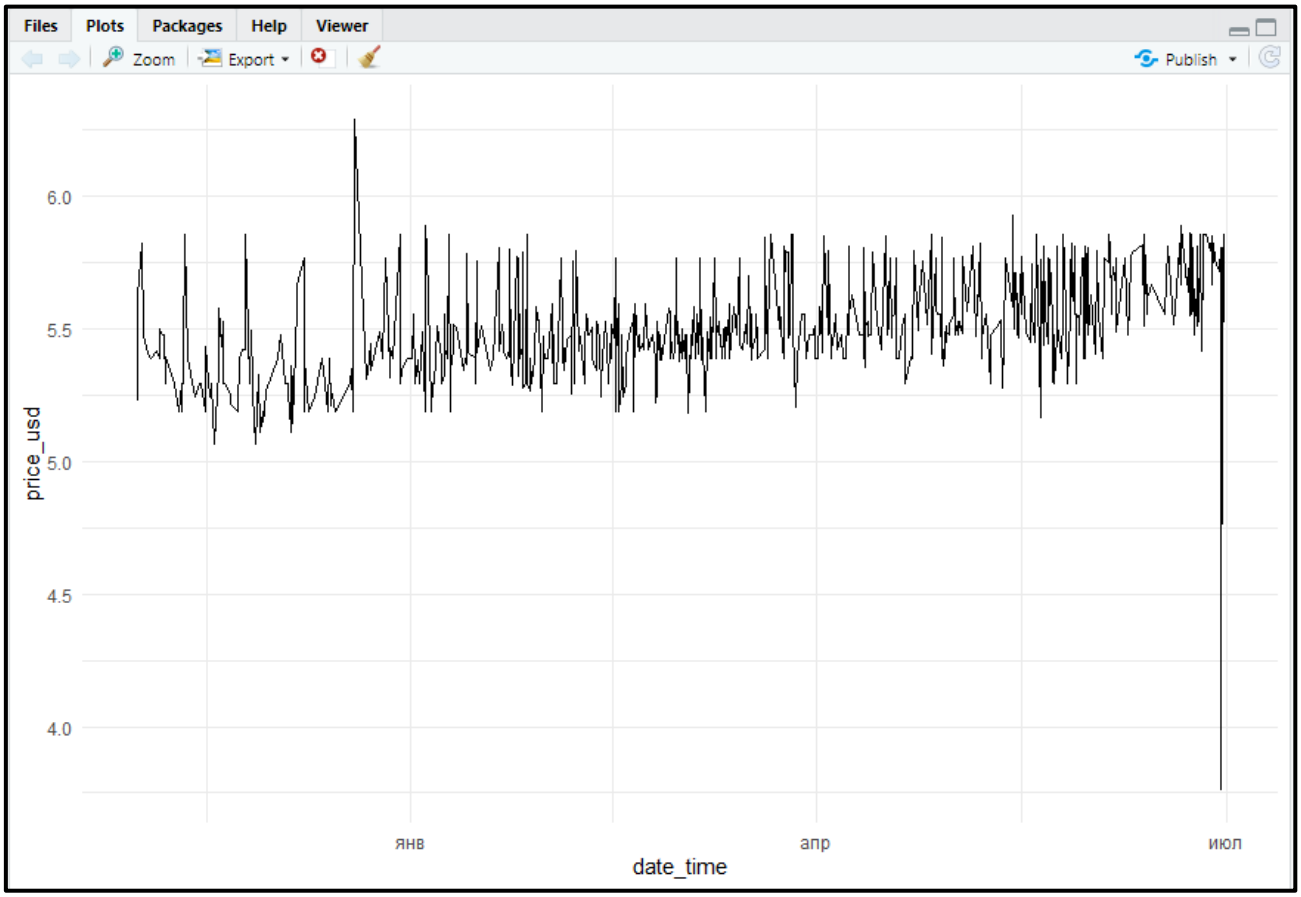

Рисунок 3.6 – Візуалізація room_train

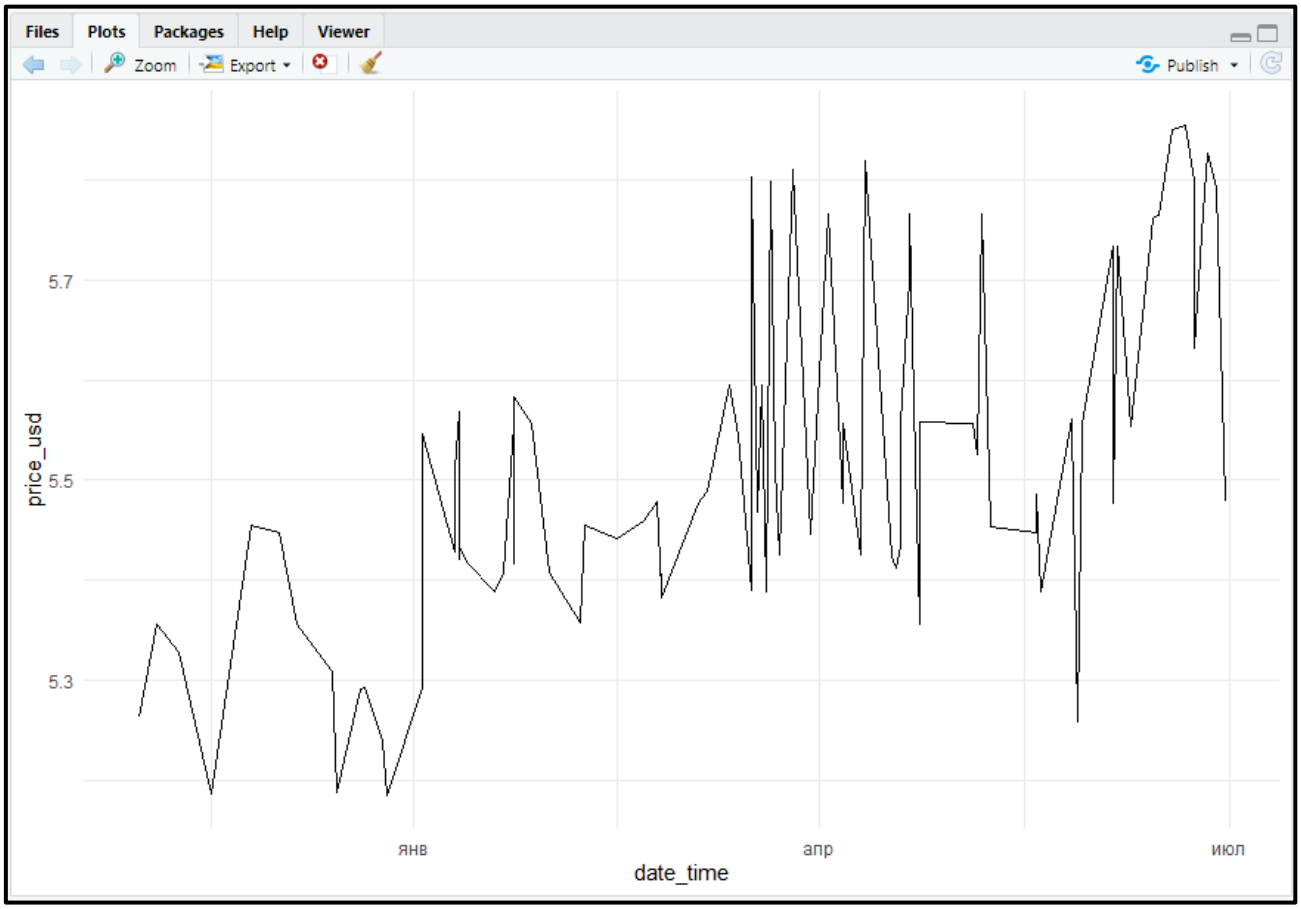

Рисунок 3.7 – Візуалізація room_test

3.2.1 Прогнозування

Побудуємо модель для прогнозування (позначимо її M0) з використанням параметрів, прийнятих в prophet за замовчуванням.

```
> names(room_train)[1] <- "y"
> names(room_train)[2] <- "ds"
> M0 <- prophet(room_train, yearly.seasonality=TRUE, daily.seasonality=TRUE)
> str(M0)
```

На рис. 3.8 та 3.9 відображена структура моделі M0 за допомогою команди str(M0).

```
> str(M0)
List of 32
 $ growth                 : chr "linear"
 $ changepoints           : POSIXct[1:25], format: "2012-11-15 10:28:33" ...
 $ n.changepoints         : num 25
 $ changepoint.range      : num 0.8
 $ yearly.seasonality     : logi TRUE
 $ weekly.seasonality     : chr "auto"
 $ daily.seasonality      : logi TRUE
 $ holidays               : NULL
 $ seasonality.mode       : chr "additive"
 $ seasonality.prior.scale: num 10
 $ changepoint.prior.scale: num 0.05
 $ holidays.prior.scale   : num 10
 $ mcmc.samples           : num 0
 $ interval.width         : num 0.8
 $ uncertainty.samples    : num 1000
 $ specified.changepoints : logi FALSE
 $ start                  : POSIXct[1:1], format: "2012-11-01 11:40:14"
 $ y.scale                : num 6.29
 $ logistic.floor         : logi FALSE
 $ t.scale                : num 20833263
 $ changepoints.t         : num [1:25] 0.0579 0.0993 0.1414 0.1826 0.2484 ...
 $ seasonalities          :List of 3
  ..$ yearly:List of 5
  .. ..$ period        : num 365
  .. ..$ fourier.order : num 10
  .. ..$ prior.scale   : num 10
  .. ..$ mode          : chr "additive"
  .. ..$ condition.name: NULL
  ..$ weekly:List of 5
  .. ..$ period        : num 7
  .. ..$ fourier.order : num 3
  .. ..$ prior.scale   : num 10
  .. ..$ mode          : chr "additive"
  .. ..$ condition.name: NULL
  ..$ daily :List of 5
  .. ..$ period        : num 1
  .. ..$ fourier.order : num 4
  .. ..$ prior.scale   : num 10
  .. ..$ mode          : chr "additive"
  .. ..$ condition.name: NULL
 $ extra_regressors       : list()
 $ country_holidays       : NULL
 $ stan.fit               :List of 4
  ..$ par        :List of 6
  .. ..$ k         : num 0.0516
  .. ..$ m         : num 0.841
  .. ..$ delta     : num [1:25(1d)] -1.47e-06 3.31e-08 -2.72e-08 -1.85e-09 -3.96e-10 ...
  .. ..$ sigma_obs: num 0.0268
  .. ..$ beta      : num [1:34(1d)] -0.0158 0.0411 -0.0342 -0.0488 0.0434 ...
  .. ..$ trend     : num [1:648(1d)] 0.841 0.841 0.842 0.842 0.842 ...
  ..$ value      : num 2022
  ..$ return_code: int 0
  ..$ theta_tilde: num [1, 1:710] 5.16e-02 8.41e-01 -1.47e-06 3.31e-08 -2.72e-08 ...
  .. ..- attr(*, "dimnames")=List of 2
  .. .. ..$ : NULL
```

Рисунок 3.8 – str(M0)

```
 .. .. ..$ : chr [1:710] "k" "m" "delta[1]" "delta[2]" ...
$ params                :List of 6
 ..$ k         : num 0.0516
 ..$ m         : num 0.841
 ..$ delta     : num [1, 1:25] -1.47e-06 3.31e-08 -2.72e-08 -1.85e-09 -3.96e-10 ...
 ..$ sigma_obs: num 0.0268
 ..$ beta      : num [1, 1:34] -0.0158 0.0411 -0.0342 -0.0488 0.0434 ...
 ..$ trend     : num [1:648(1d)] 0.841 0.841 0.842 0.842 0.842 ...
$ history               :'data.frame':        648 obs. of  5 variables:
 ..$ y       : num [1:648] 5.23 5.63 5.83 5.48 5.4 ...
 ..$ ds      : POSIXct[1:648], format: "2012-11-01 11:40:14" ...
 ..$ floor   : num [1:648] 0 0 0 0 0 0 0 0 0 0 ...
 ..$ t       : num [1:648] 0 0.000747 0.004239 0.006084 0.010334 ...
 ..$ y_scaled: num [1:648] 0.83 0.895 0.926 0.87 0.859 ...
$ history.dates         : POSIXct[1:648], format: "2012-11-01 11:40:14" ...
$ train.holiday.names   : NULL
$ train.component.cols  :'data.frame':        34 obs. of  5 variables:
 ..$ additive_terms      : int [1:34] 1 1 1 1 1 1 1 1 1 1 ...
 ..$ daily               : int [1:34] 0 0 0 0 0 0 0 0 0 0 ...
 ..$ weekly              : int [1:34] 0 0 0 0 0 0 0 0 0 0 ...
 ..$ yearly              : int [1:34] 1 1 1 1 1 1 1 1 1 1 ...
 ..$ multiplicative_terms: num [1:34] 0 0 0 0 0 0 0 0 0 0 ...
$ component.modes       :List of 2
 ..$ additive      : chr [1:6] "yearly" "weekly" "daily" "additive_terms" ...
 ..$ multiplicative: chr [1:2] "multiplicative_terms" "extra_regressors_multiplicative"
$ fit.kwargs            : list()
- attr(*, "class")= chr [1:2] "prophet" "list"
```

Рисунок 3.9 – str(M0)

Для отримання прогнозу на основі цієї моделі необхідно спочатку скористатися функцією make_future_dataframe () і створити таблицю з датами, які охоплюють необхідний часовий проміжок в майбутньому ("горизонт"), а потім подати цю таблицю разом з модельним об'єктом на функцію predict ():

```
> future_df<-make_future_dataframe(M0,periods = 90)
> forecast_M0<-predict(M0,future_df)
```

Об'єкт forecast_M0 - це звичайна таблиця, в якій зберігаються значення декількох розрахованих на основі моделі M0 величин, включаючи компоненти моделі, передбачені значення відгуку, а також верхні і нижні межі довірчих інтервалів відповідних величин. Перші кілька передбачених значень вартості room і їх (прийняті за замовчуванням) 80% (див. рис. 3.10):

```
> head(forecast_M0)
                   ds    trend additive_terms
1 2012-11-01 11:40:14 5.295560      0.2142817
2 2012-11-01 15:59:28 5.295802      0.2416482
3 2012-11-02 12:11:58 5.296937      0.1989354
4 2012-11-02 22:52:48 5.297537      0.1897785
5 2012-11-03 23:28:34 5.298919      0.1288658
6 2012-11-04 08:05:45 5.299403      0.1107633
  additive_terms_lower additive_terms_upper        daily
1            0.2142817            0.2142817 -0.011857852
2            0.2416482            0.2416482  0.018770665
3            0.1989354            0.1989354 -0.007820396
4            0.1897785            0.1897785  0.016681327
5            0.1288658            0.1288658  0.013408794
6            0.1107633            0.1107633 -0.011244301
   daily_lower  daily_upper       weekly weekly_lower
1 -0.011857852 -0.011857852 -0.002179176 -0.002179176
2  0.018770665  0.018770665  0.001641946  0.001641946
3 -0.007820396 -0.007820396  0.015978235  0.015978235
4  0.016681327  0.016681327 -0.003298835 -0.003298835
5  0.013408794  0.013408794 -0.032047075 -0.032047075
6 -0.011244301 -0.011244301 -0.016683676 -0.016683676
  weekly_upper    yearly yearly_lower yearly_upper
1 -0.002179176 0.2283187    0.2283187    0.2283187
2  0.001641946 0.2212356    0.2212356    0.2212356
3  0.015978235 0.1907775    0.1907775    0.1907775
4 -0.003298835 0.1763960    0.1763960    0.1763960
5 -0.032047075 0.1475041    0.1475041    0.1475041
6 -0.016683676 0.1386913    0.1386913    0.1386913
  multiplicative_terms multiplicative_terms_lower
1                    0                          0
2                    0                          0
3                    0                          0
4                    0                          0
5                    0                          0
6                    0                          0
  multiplicative_terms_upper yhat_lower yhat_upper
1                          0   5.297136   5.738930
2                          0   5.327700   5.740922
3                          0   5.264930   5.706217
4                          0   5.272697   5.713946
5                          0   5.200654   5.649100
6                          0   5.192411   5.632026
  trend_lower trend_upper     yhat
1    5.295560    5.295560 5.509841
2    5.295802    5.295802 5.537451
3    5.296937    5.296937 5.495873
4    5.297537    5.297537 5.487316
5    5.298919    5.298919 5.427784
6    5.299403    5.299403 5.410166
```

Рисунок 3.10 – forecast_M0

Таблицю forecast_M0 і об'єкт M0 далі можна подати на функцію plot(), щоб зобразити підігнані модель і прогнозні значення на графіку (див. рис. 3.11):

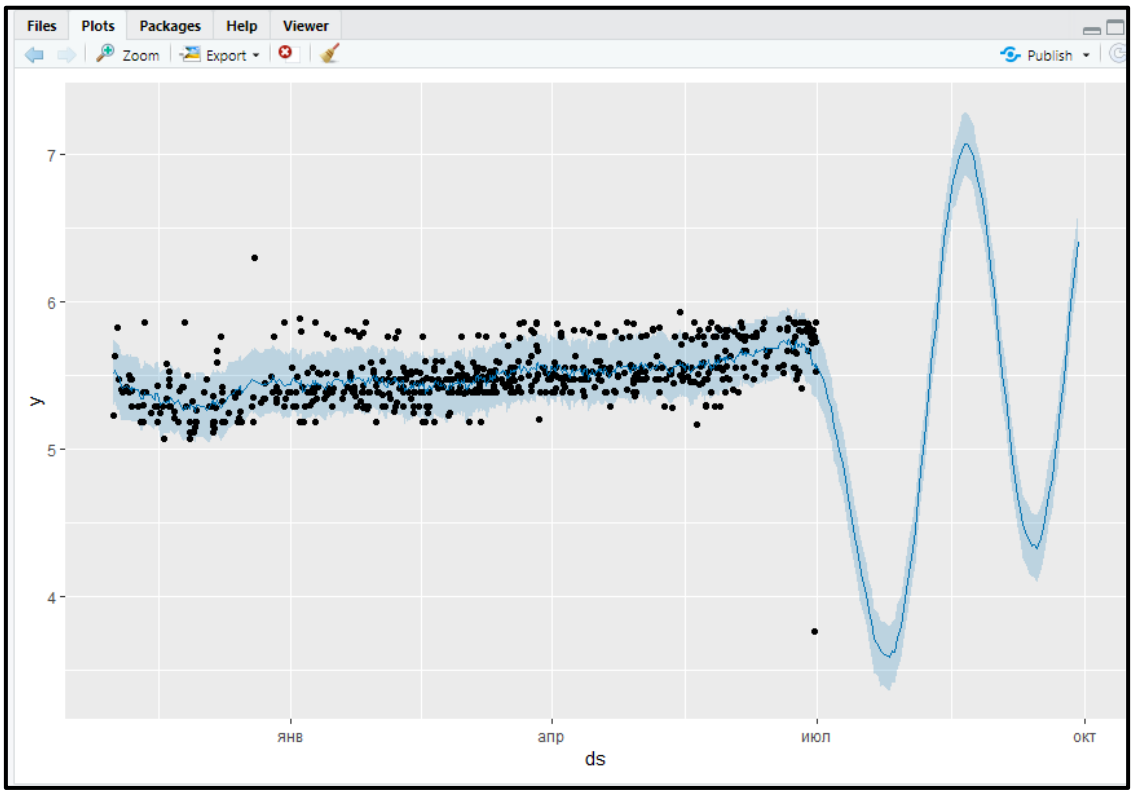

Рисунок 3.11 – Прогноз

Точки на рис. 3.11 відповідають (логарифмічному) значенням вартості room з навчальної вибірки.

Хмарно-блакитна лінія - це передбачені моделлю значення вартості, а огинає цю лінію світло-блакитна "стрічка" позначає 80% - довірчі кордону передбачених значень. Прогнозні значення у наступні 90 днів видно в правій частині графіка. З графіку видно, що прогноз незадовільний, так як він недостатньо точно передає структуру вибірки. Вірогідно причина в тому, що вибірка недостатньо насичена даними.

Змінимо розмір тренувальної вибірки з 90 днів до 240 (див. рис. 3.12). З графіку видно, що прогноз став дещо точніше повторювати структуру тренувальної вибірки, але все одно незадовільно.

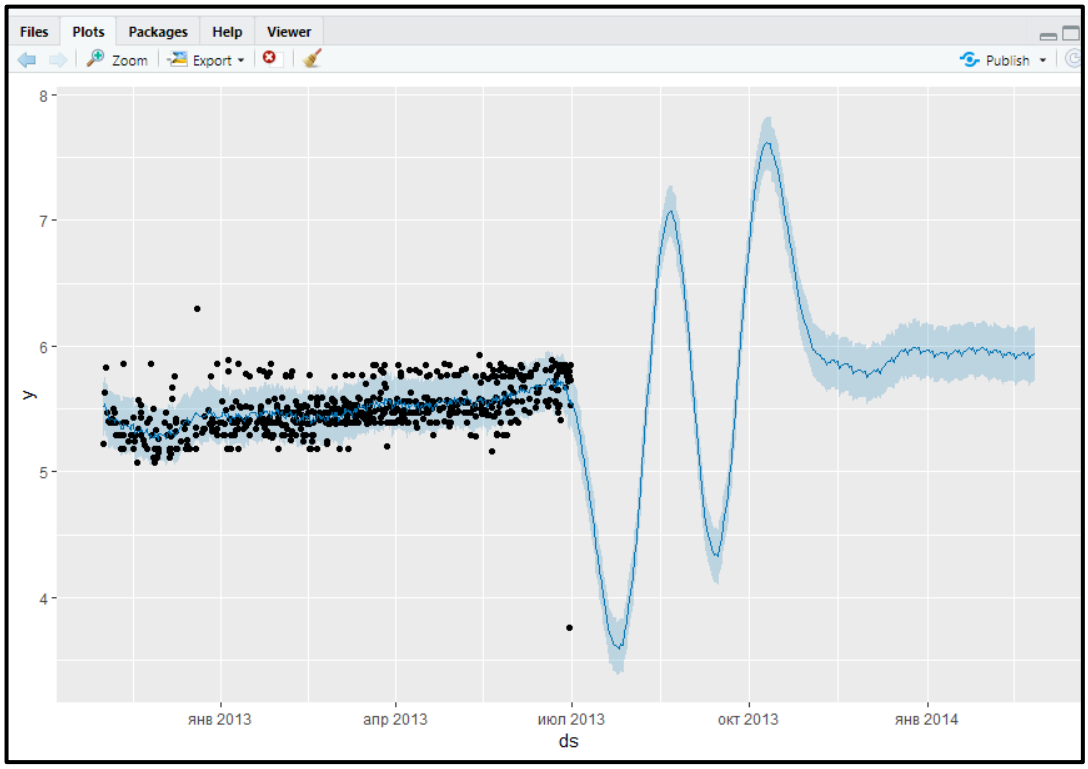

Рисунок 3.12 – Прогноз - 240 днів

Зобразимо окремі компоненти моделі (див. рис. 3.13):

```
prophet_plot_components(M0, forecast_M0)
```

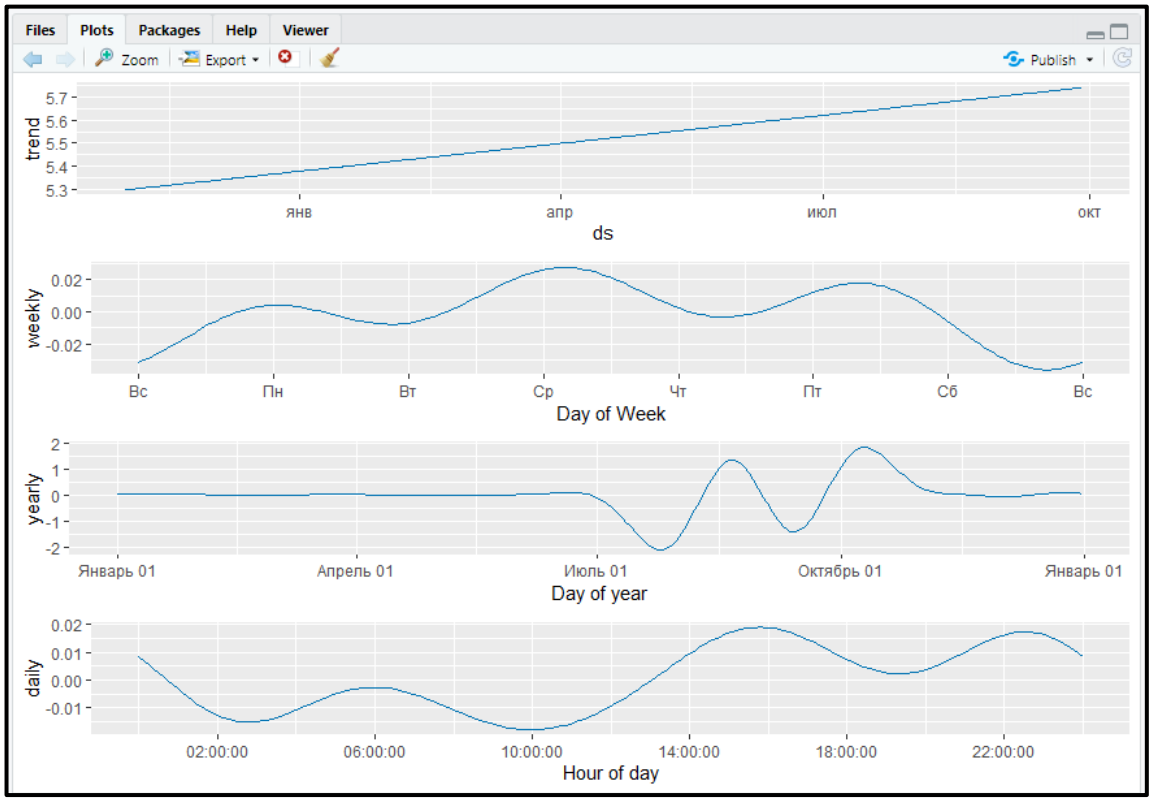

Рисунок 3.13 – Окремі компоненти моделі – 90 днів

На рис. 3.13 видно, що модель M0 добре передає наявний в даних складний тренд. Видно також, що в цьому часовому ряду є дуже слабо виражені внутрішньорічні коливання і практично неіснуючі коливання в межах тижня. Шкали ординат цих трьох графіків допомагають оцінити внесок кожної з компонент. Отримана модель досить добре передає властивості цього ряду. Проте, якість прогнозу M0 недостатньо добра. На даному етапі моделювання головною ознакою незадовільної якості прогнозів M0 є надмірно розширені довірчі кордону прогнозних значень (див. рис. 3.13).

3.2.2 Зміна параметрів

Змінюючи значення основних аргументів функції prophet () отримаємо альтернативні моделі для прогнозування. У зв'язку зі складністю лінії тренду тимчасового ряду поставимо крапки зламу тренда. Це завдання можна виконати

двома способами: поставити самостійно або довіритися їх автоматичного виявлення.

В автоматичному режимі при ініціалізації моделі 25 потенційних точок зламу рівномірно розподіляються в межах інтервалу, який охоплює перші 80% спостережень з навчальної вибірки. Це сталося, коли була побудована модель M0. Однак ці 25 точок - лише передбачувані місця істотних змін в тренді: в більшості випадків на практиці тренд часового ряду не змінюється так часто. Тому в ході підгонки моделі спрацьовує механізм регуляризації, в результаті чого вибирається мінімально необхідну кількість точок зламу.

Зобразимо ці автоматично виявлені точки зламу за допомогою функції add_changepoints_to_plot (). Так, для моделі M0 отримуємо (див. рис. 3.14):

```
> plot(M0,forecast_M0)+add_changepoints_to_plot(M0)
```

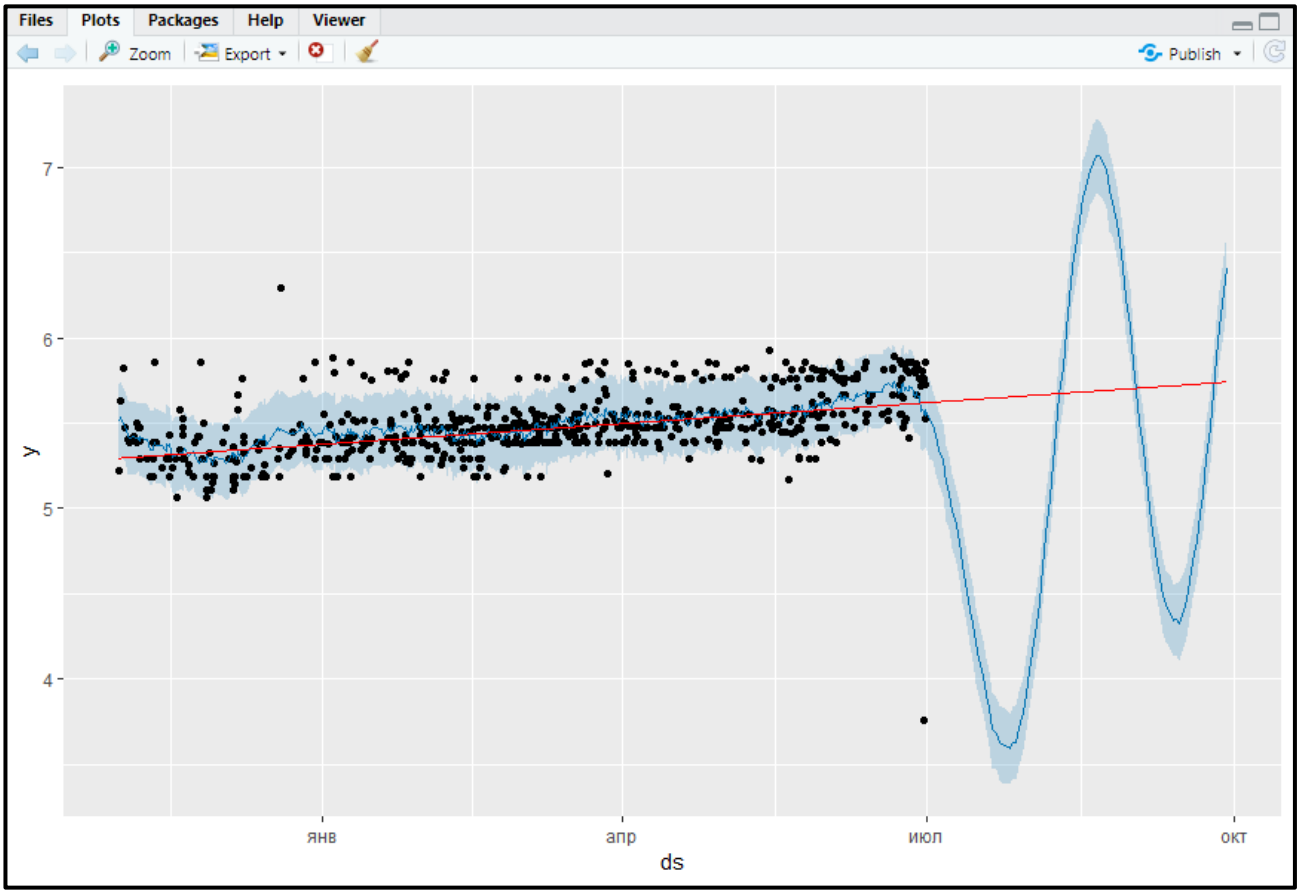

Рисунок 3.14 – Автоматично виявлені точки зламу

Судячи по отриманому графіку, модель M0 недооцінює кількість "переломних моментів" в тренді.

Побудуємо нову модель M1, яка буде ініціалізована з меншим початковим кількістю потенційних точок зламу (15 замість 25 за замовчуванням) (див. рис. 3.15):

```
> M1<-prophet(room_train,n.changepoints = 15, yearly.seasonality=TRU
E)
> forecast_M1<-predict(M1,future_df)
> plot(M1,forecast_M1)+add_changepoints_to_plot(M1)
```

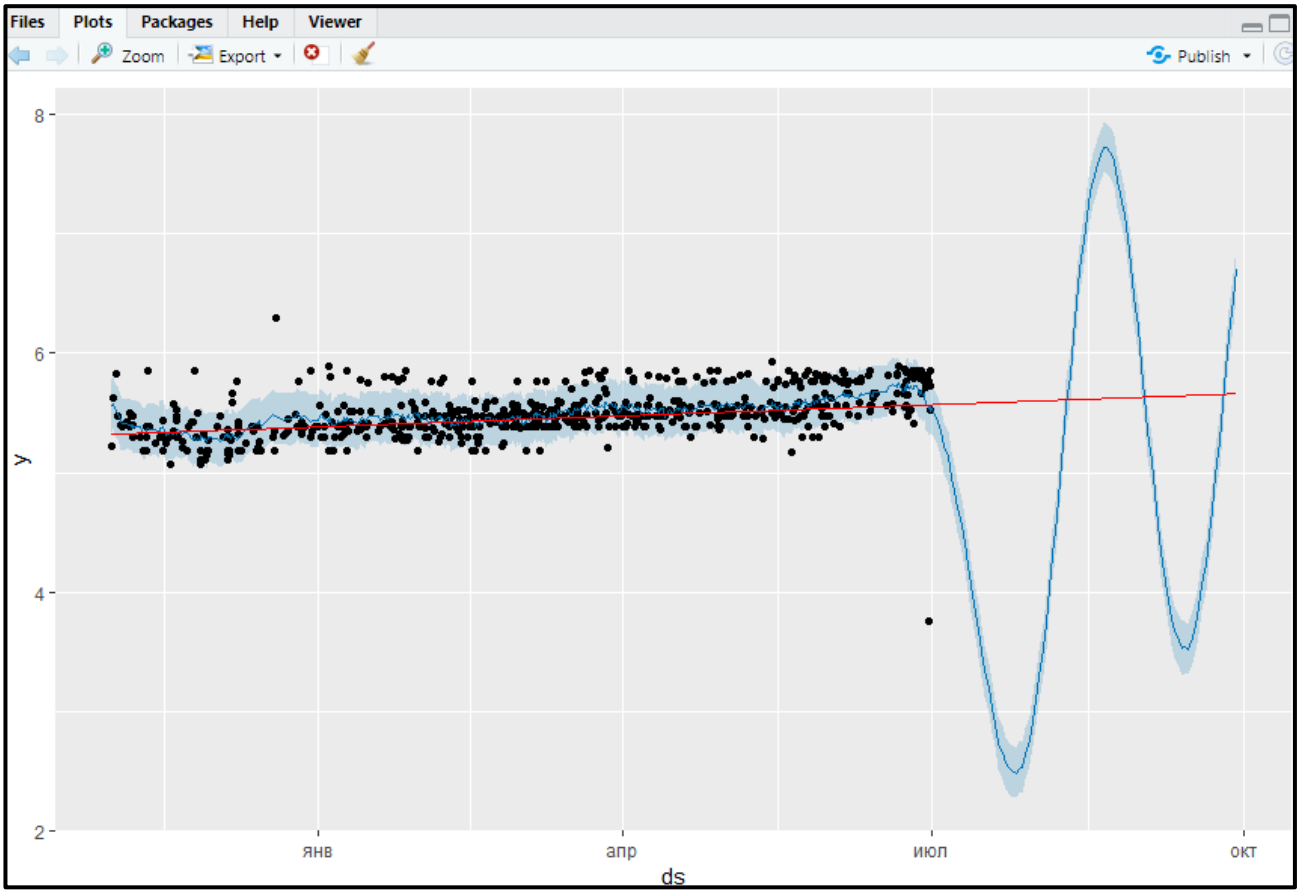

Рисунок 3.15 – Автоматично виявлені точки зламу з меншим початковим кількістю

Знову оцінений тренд вийшов менш точним, ніж в моделі M0.

Побудуємо ще одну модель для прогнозування (позначимо її M2). Крім зміни початкового кількості потенційних точок зламу тренда змінимо також часовий

інтервал, в межах якого відбувається їх оцінювання. Збільшимо інтервал до 90%, скориставшись аргументом changepoint.range і одночасно збільшимо кількість потенційних точок зламу з 15 до 20, оскільки на більшій проміжку часу можна очікувати більше перепадів в тренді (див. рис. 3.16):

```
> M2<-prophet(room_train,n.changepoints = 20,changepoint.range = 0.9
, yearly.seasonality=TRUE)
> forecast_M2<-predict(M2,future_df)
> plot(M2,forecast_M2)+add_changepoints_to_plot(M2)
```

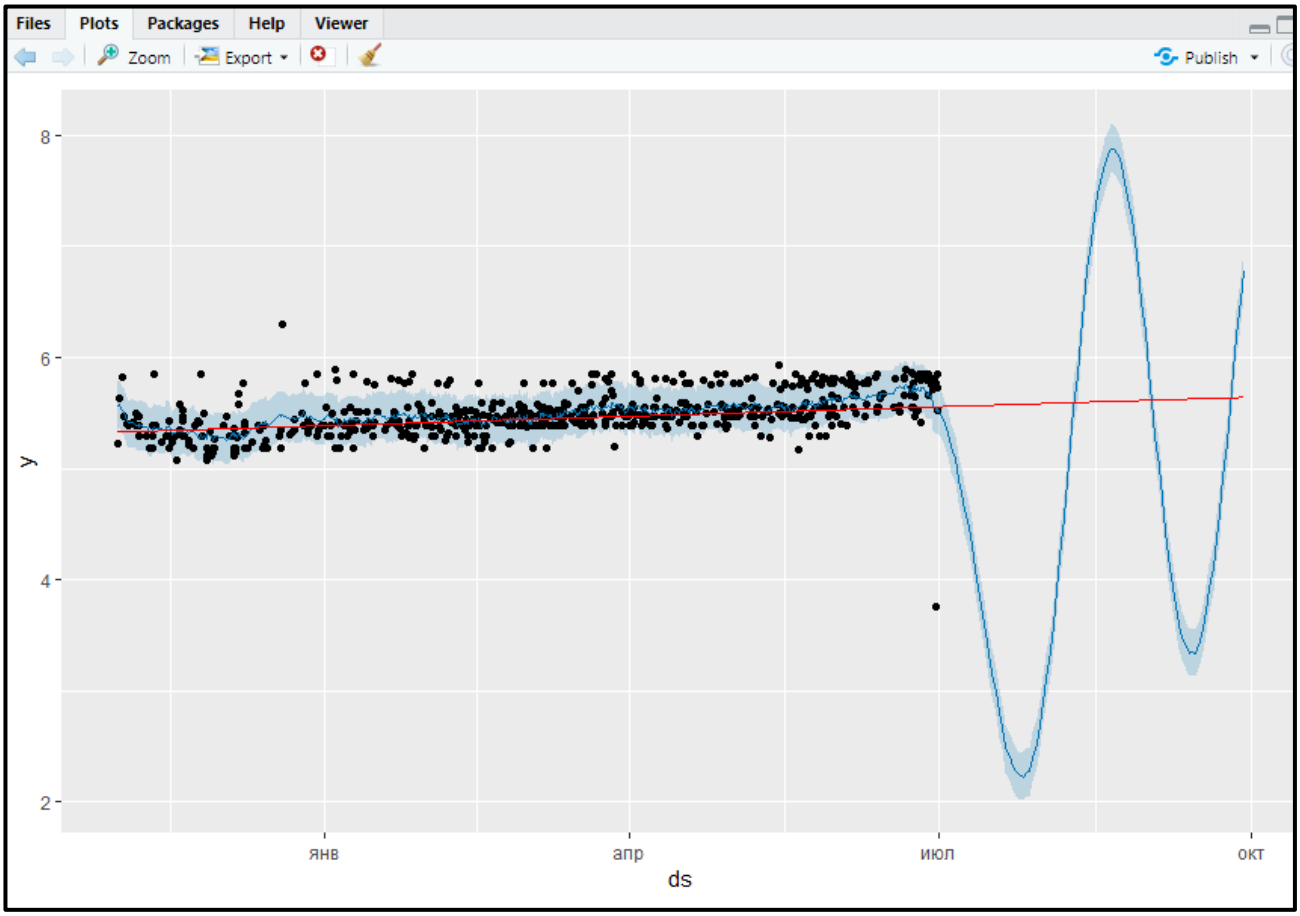

Рисунок 3.16 – Автоматично виявлені точки зламу з меншим початковим кількістю потенційних точок зламу та інтервалом 90%

Як видно на рис. 3.16, отримана модель M2 трохи краще передає властивості аналізованого часового ряду. Це стосується і одержуваного з її допомогою прогнозу.

Побудуємо наступну альтернативну модель для прогнозування, змінюючи параметр для налаштування гладкості тренда в моделюючому тимчасовому ряду - це changepoint.prior.scale. Чим більше значення цього параметра (в порівнянні з прийнятим за замовчуванням значенням 0.05), тим більше точок зламу залишиться в отриманої моделі. У моделі M3 збільшуємо інтервал, в межах якого оцінюються точки зламу тренда (до 90%), одночасно збільшуючи рівень регуляризації за допомогою параметра changepoint.prior.scale. Початкова кількість потенційних точок зламу залишимо рівним значенню, прийнятому за замовчуванням (25) (див. рис. 3.17):

```
> M3<-prophet(room_train, changepoint.range = 0.9, changepoint.prior
.scale = 0.50, yearly.seasonality=TRUE)
> forecast_M3<-predict(M3,future_df)
> plot(M3,forecast_M3)+add_changepoints_to_plot(M3)
```

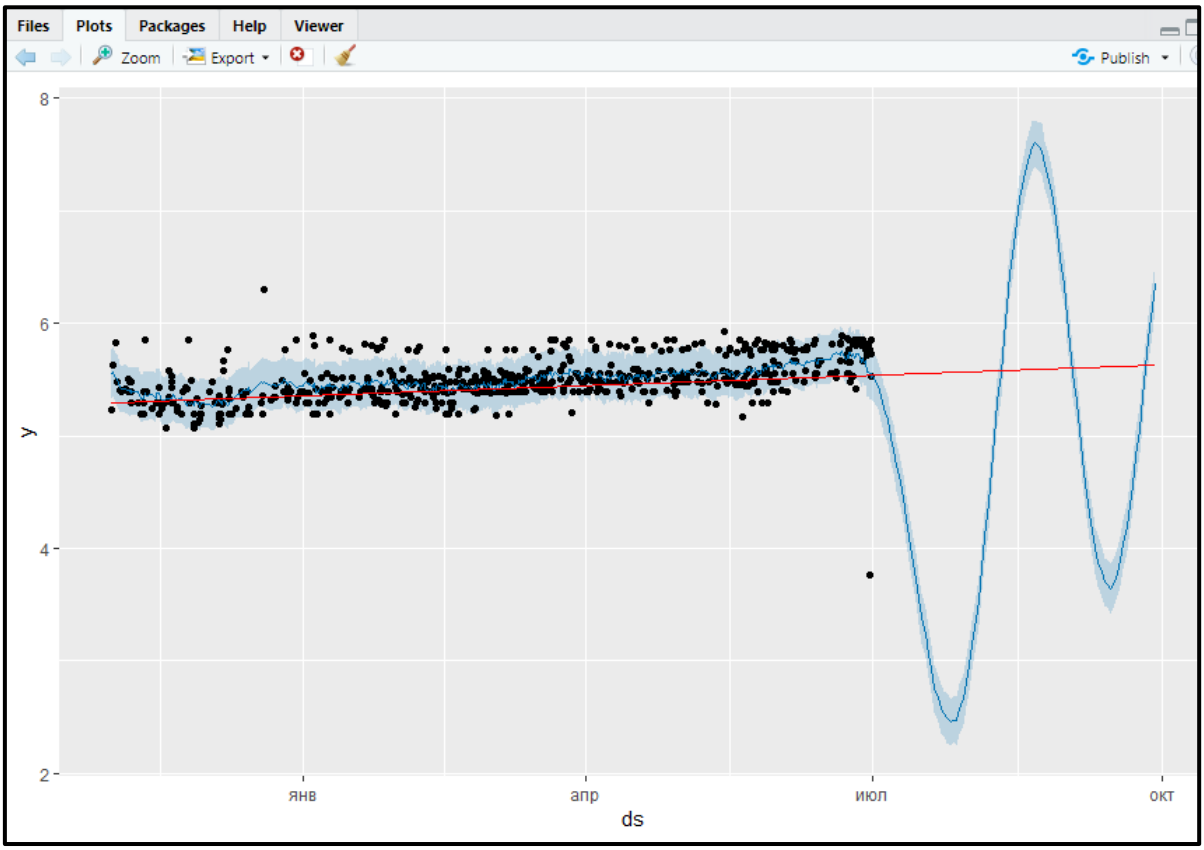

Рисунок 3.17 – Автоматично виявлені точки зламу з інтервалом 90% з більшим параметром гладкості

Видно, що моделі M3 трохи кращі результати ніж попередні моделі.

Побудуємо модель M4 і на це раз поставимо крапки зламу тренда "вручну", а не в автоматичному режимі. Для цього завдання скористаємося аргументом changepoints. Ставлячи точки зламу тренда самостійно врахуємо, що вибір дат, що подаються на аргумент changepoints, заснований на візуальному аналізі навчальних даних, а також змінемо параметр для налаштування гладкості тренда в модельованному тимчасовому ряду (див. рис. 3.18).

```
> M4<-prophet(room_train, changepoint.prior.scale = 0.50, yearly.sea
sonality=TRUE, changepoints = c(
+ "2012-12-31", "2013-03-20", "2013-03-20",
+ "2013-03-03", "2013-04-03", "2013-04-03",
+ "2013-04-11", "2013-06-01", "2013-02-08",
+ "2013-04-19", "2013-04-13"
+ ))
```

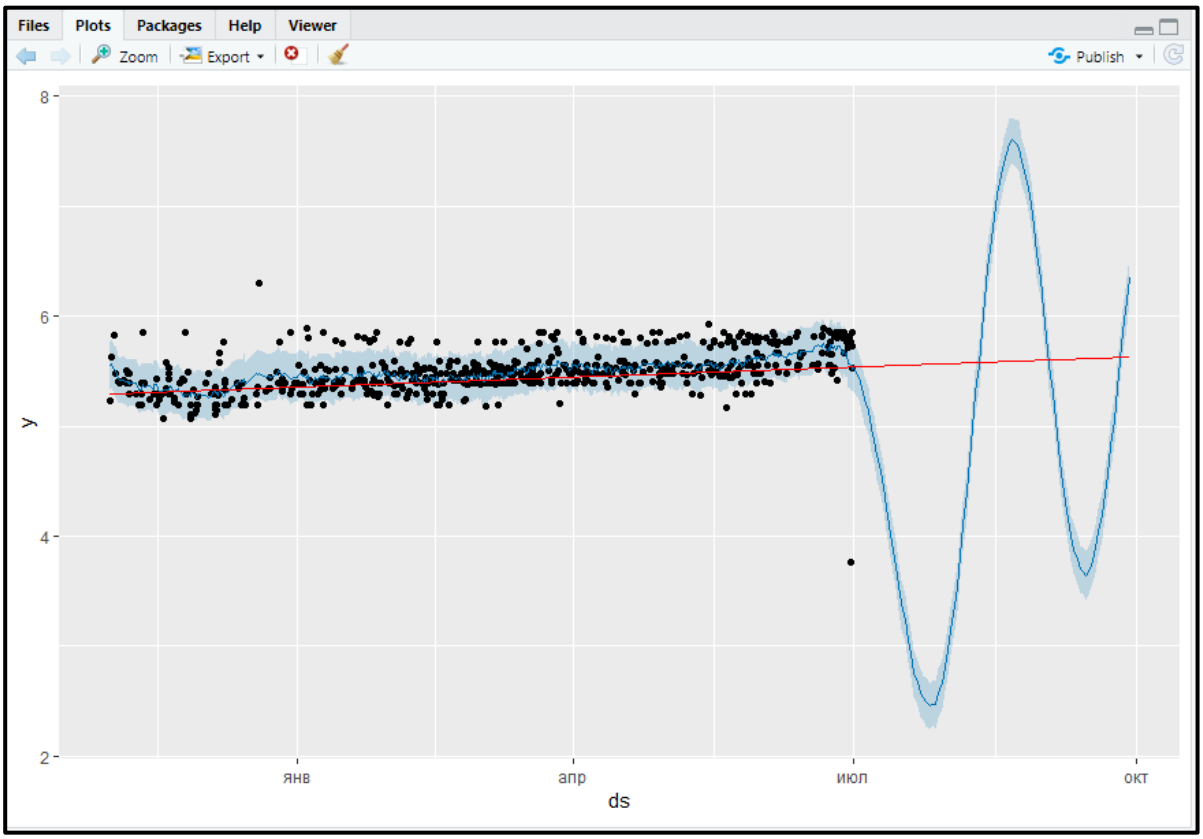

Рисунок 3.18 – Власні точки злому

Модель M4 непогано описує тренд в уже згадуваному тимчасовому ряду.

3.2.3 Річна, тижнева і денна компоненти

Сезонна компонента є однією зі складових при підгонки тимчасового ряду за допомогою адитивних регресійних моделей. Сезонні компоненти апроксимуються за допомогою часткових сум ряду Фур'є, число членів якого (порядок) визначає гладкість відповідної функції.

Функція prophet () має три аргументи, за допомогою яких можна контролювати гладкість функцій річний, тижневої та денної сезонності: yearly.seasonality, weekly.seasonality і daily.seasonality. Збільшення значень цих аргументів призведе до підгонки менш гладких функцій відповідних компонент, що одночасно збільшить ризик перенавчання моделі. Збільшивши значення аргументу yearly.seasonality до 20, отримаємо наступну модель (див. рис. 3.19):

```
> M3B <- prophet(room_train,
+ yearly.seasonality = 20,
+ changepoint.range = 0.9,
+ changepoint.prior.scale = 0.02)
> prophet:::plot_yearly(M3B)
```

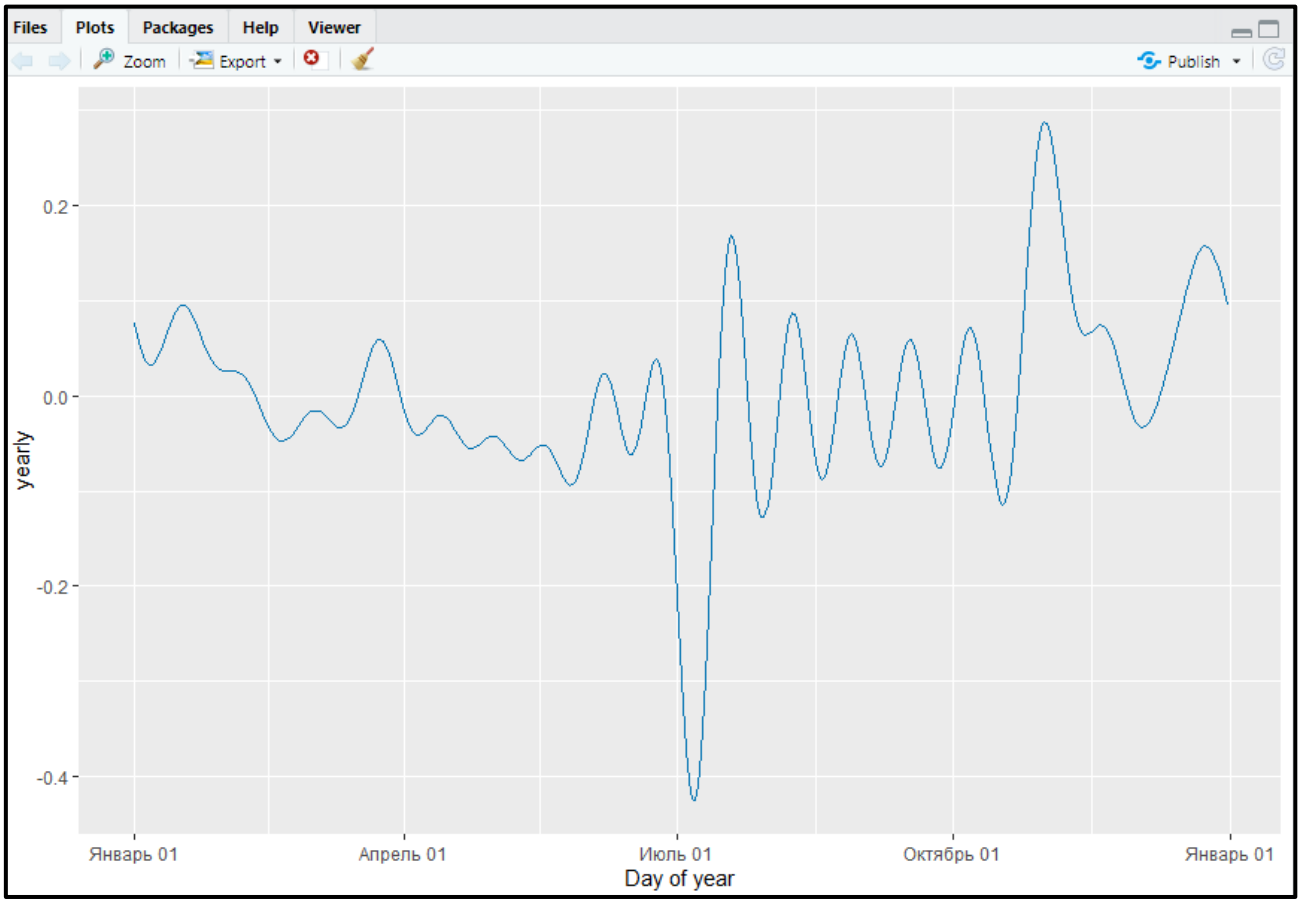

Рисунок 3.19 – Функція річної сезонності, оцінена за допомогою моделі M3B

3.2.4 Призначені для користувача сезонні компоненти

Для даних, що охоплюють як мінімум два роки, функція prophet () автоматично додає в модель компоненти річний і тижневої сезонності. Якщо гранулярність даних перевищує денну (наприклад, коли є погодинні спостереження залежної змінної), то в модель автоматично буде додана також і компонента денний сезонності. Крім цього, є можливість додати і будь-які інші сезонні компоненти за допомогою функції add_seasonality () (наприклад, годинну, місячну, квартальну і т.п.).

У наведеному нижче коді спочатку відключаємо автоматично додається в модель тижневу сезонність і замість неї додаємо місячну (допустивши, що один місячний період становить 30.5 днів). На рис. 3.20 представлені всі сезонні компоненти отриманої моделі.

```
> M10 <- prophet(weekly.seasonality = FALSE, yearly.seasonality= TRUE)
> M10 <- add_seasonality(m = M10,
+ name = "monthly",
+ period = 30.5,
+ fourier.order = 5)
> M10 <- fit.prophet(M10, room_train)
> forecast_M10 <- predict(M10, future_df)
> prophet_plot_components(M10, forecast_M10)
```

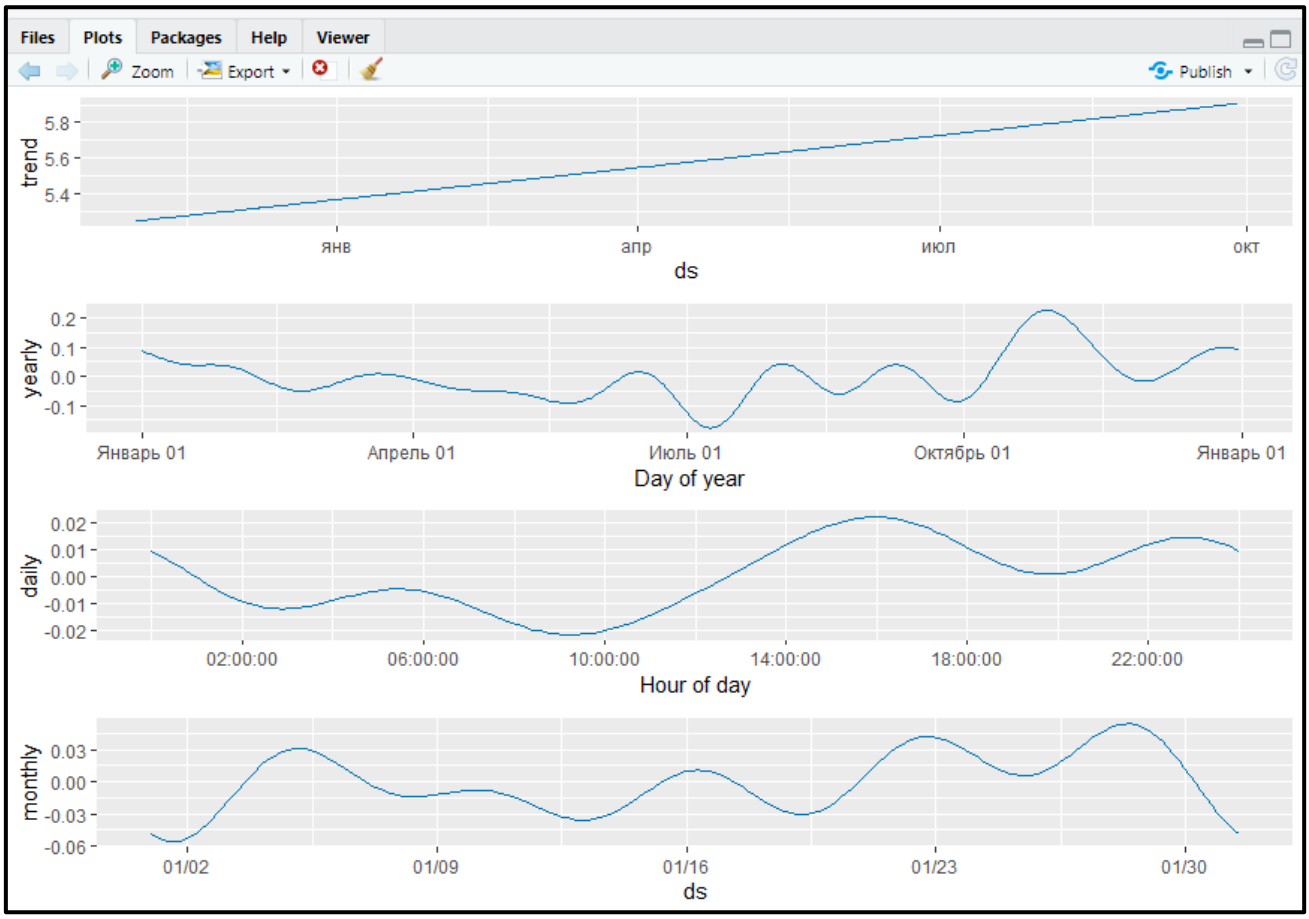

Рисунок 3.20 – Компоненти моделі M10

Відповідно до отриманої моделі чітко видно місячні, денні, рокові коливання та тренд. Проте, якість прогнозу M10 залишає бажати кращого. На даному етапі моделювання головною ознакою незадовільної якості прогнозів M10 є надмірно розширюються довірчі кордону прогнозних значень (див. рис. 3.20).

Аналогічним чином замість компоненти місячних коливань додаємо компоненту квартальної сезонності (задавши період довжиною 365.25 / 4 днів). На рис. 3.21 представлені всі сезонні компоненти отриманої моделі.

```
> M11 <- prophet(weekly.seasonality = FALSE, yearly.seasonality= TRUE)
> M11 <- add_seasonality(m = M11,
+ name = "quarter",
+ period = 365.25/4,
+ fourier.order = 2)
> M11 <- fit.prophet(M11, room_train)
> forecast_M11 <- predict(M11, future_df)
> prophet_plot_components(M11, forecast_M11)
```

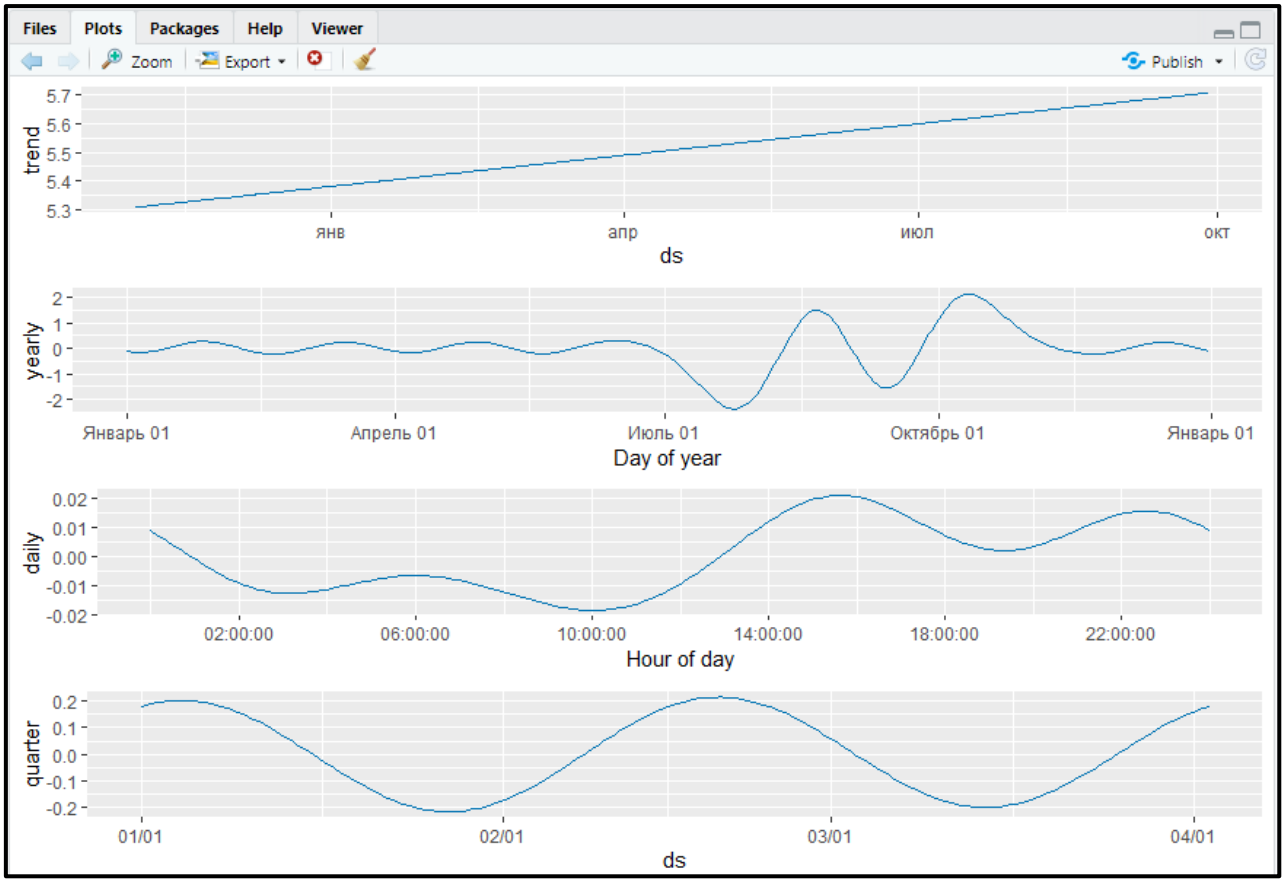

Рисунок 1.21 – Компоненти моделі M11

Відповідно до отриманої моделі чітко видно коливання по кварталу яке дуже гладке і це не дуже добре та тренд. В порівнянні з M10 тренд майже не змінився (див. рис. 3.21).

3.2.5 Умовні режими сезонності

У ряді випадків функція, що апроксимує ту чи іншу сезонну складову, може змінювати свої властивості в залежності від якихось сторонніх чинників. Наприклад, коливання протягом робочих днів можуть мати характер, сильно відрізняється від такого у вихідні дні. Пакет prophet дозволяє моделювати такі умовні режими сезонності (тобто режими, які залежать від сторонніх чинників) за допомогою аргументу condition.name функції add_seasonality (). На цей аргумент подається ім'я (булевої) змінної, яка визначає відповідний режим. Такі змінні повинні зберігається в тій же таблиці, що і основні дані по тимчасовому ряду.

Як приклад припустимо, що тижневі коливання вартості stellar в літні місяці відрізняються від таких в інші місяці. Щоб змоделювати таке розходження додаємо в таблицю з даними stellar_train дві нові індикаторні змінні: summer (приймає значення TRUE в літні місяці і FALSE в інші місяці) і not_summer (TRUE в нелітні місяці і FALSE влітку). Важливо пам'ятати, що такі ж змінні потрібно додати і в таблицю з майбутніми датами future_df - інакше прогнозні значення розрахувати не вийде:

```
> is_summer <- function(ds) {
+     month <- as.numeric(format(ds, '%m'))
+     return(month > 5 & month < 9)
+ }
> room_train$summer <- is_summer(room_train$ds)
> room_train$not_summer <- !room_train$summer
> future_df$summer <- is_summer(future_df$ds)
> future_df$not_summer <- !future_df$summer
> M12 <- prophet(weekly.seasonality = FALSE)
> M12 <- add_seasonality(M12, name = 'weekly_summer',
+                        period = 7,
+                        fourier.order = 3,
+                        condition.name = 'summer')
> M12 <- add_seasonality(M12, name = "weekly_not_summer",
+                        period = 7,
+                        fourier.order = 3,
```

```
+                              condition.name = "not_summer")
> M12 <- fit.prophet(M12, room_train)
> forecast_M12 <- predict(M12, future_df)
> prophet_plot_components(M12, forecast_M12)
```

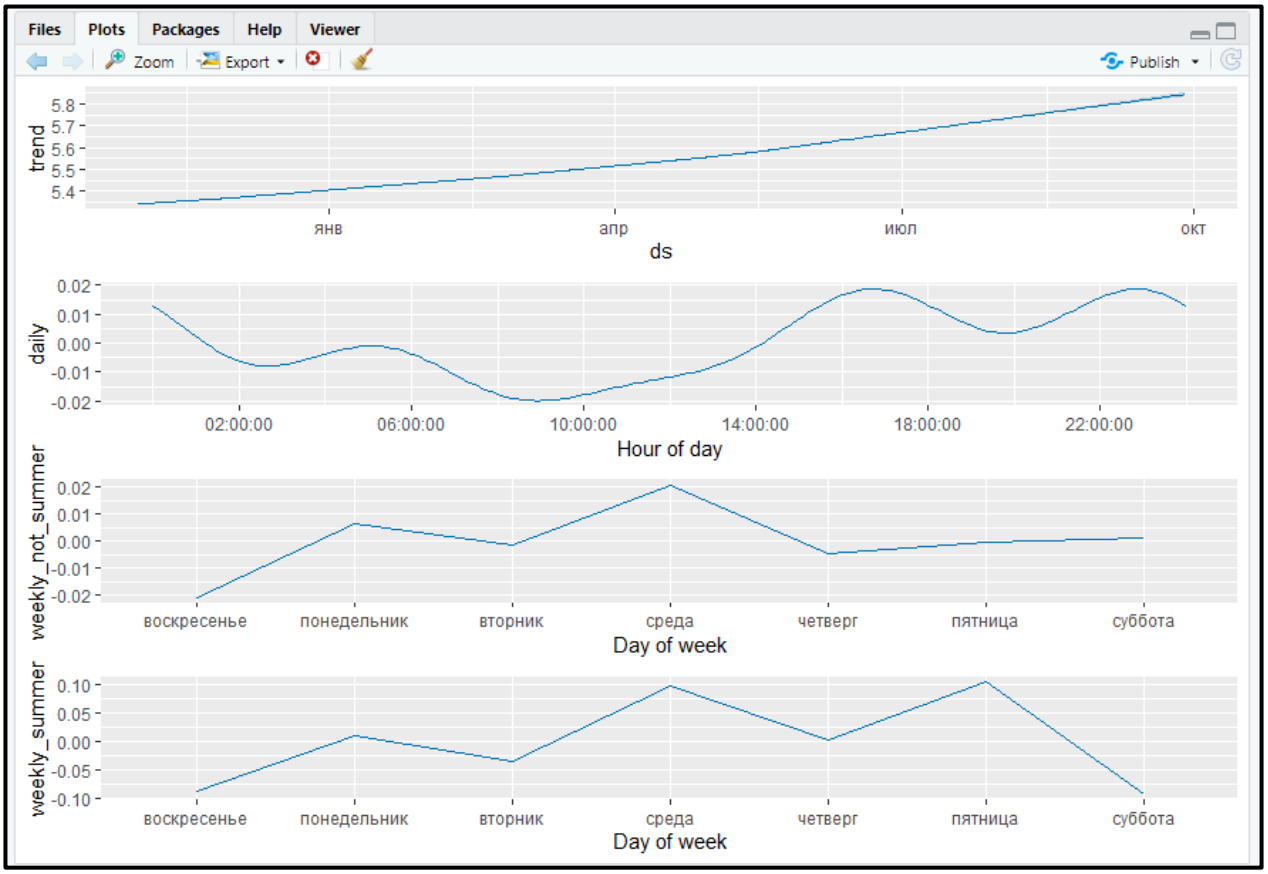

Рисунок 3.22 – Компоненти моделі M12

Відповідно до отриманої моделі, в нелітні місяці вартість room протягом тижня зазвичай досягає максимуму по середам, тоді як в літні місяці по середам та п’ятницям. (див. рис. 3.22).

3.2.6 Адитивна і мультиплікативна сезонності

За характером функціонального зв'язку між своїми компонентами моделі часових рядів діляться на два основних типи - адитивні і мультиплікативні. Перший з них застосовується у випадках, коли амплітуда сезонних коливань приблизно постійна. Якщо ж ця амплітуда помітно змінюється в часі (зазвичай зростає), то будують мультиплікативну модель.

У пакеті prophet за замовчуванням підганяються адитивні моделі часових рядів. У мультиплікативний моделях, як випливає з їх назви, сезонна компонента множиться на тренд (в зв'язку з цим внесок сезонних коливань моделюється у вигляді частки (%) від рівня тренда).

Припустимо, що амплітуда всіх сезонних компонент істотно змінюється в часі. Для підгонки відповідних моделей скористаємося аргументом seasonality.mode функції prophet ():

```
> M14 <- prophet(room_train, seasonality.mode = "multiplicative"
)
> forecast_M14 <- predict(M14, future_df)
> plot(M14, forecast_M14)
```

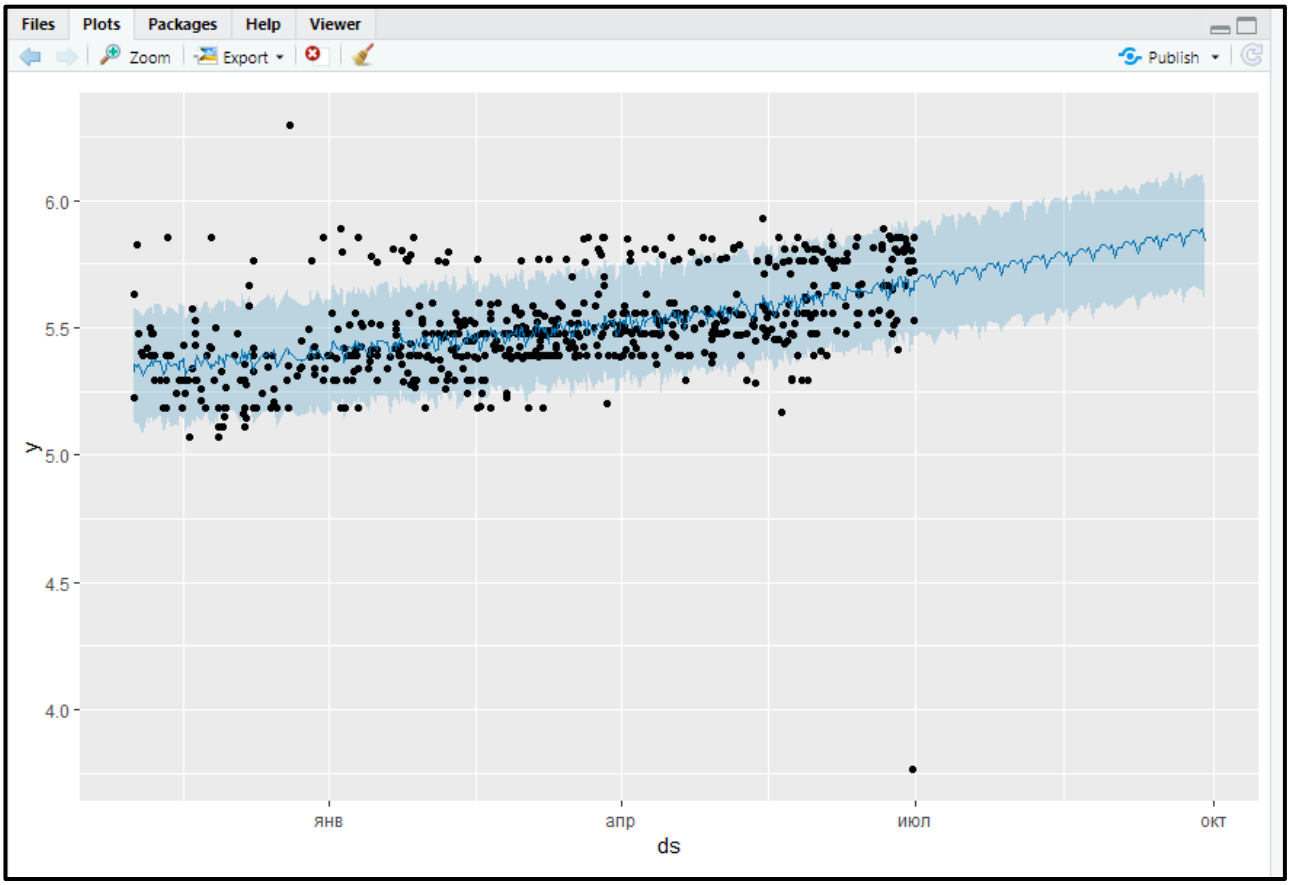

Рисунок 3.23 – Прогноз вартості room, отриманий на основі моделі M14

З графіку видно, що прогноз незадовільний, так як він недостатньо точно передає структуру вибірки. Вірогідно причина в тому, що вибірка недостатньо насичена даними (див. рис. 3.23).

Побудуємо модель, в якій тижнева коливання представлені в адитивному вигляді, а річні - в мультиплікативному. Для цього застосовується функція add_seasonality ():

```
> M15 <- prophet(yearly.seasonality = FALSE)
> M15 <- add_seasonality(M15, name = 'yearly',
+                        period = 365.25,
+                        fourier.order = 10,
+                        mode = "multiplicative")
> M15 <- fit.prophet(M15, room_train)
>
> forecast_M15 <- predict(M15, future_df)
> prophet_plot_components(M15, forecast_M15)
```

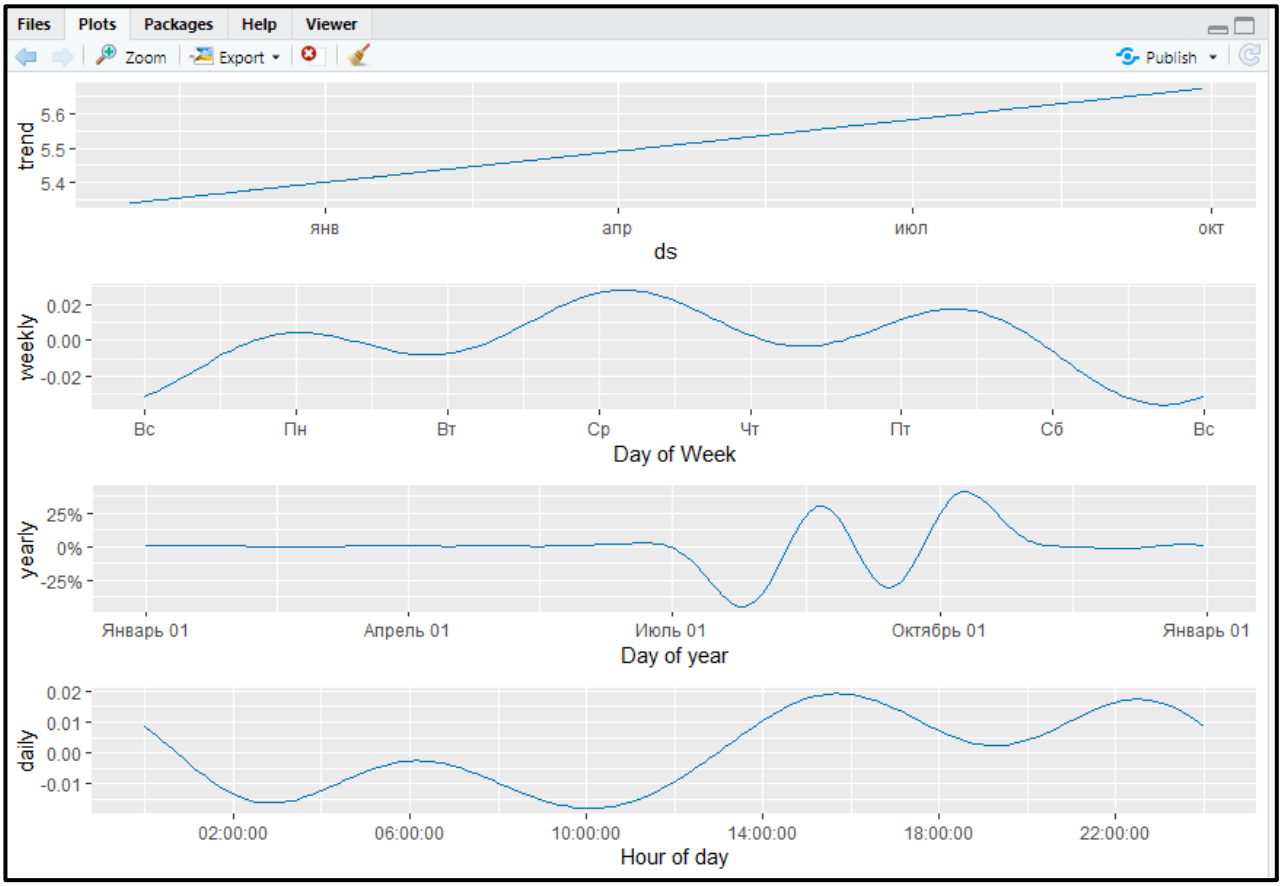

Рисунок 3.24 – Компоненти моделі M15

Відповідно до отриманої моделі, видно, що тижнева компонента різко змінилась порівнянно з M12, а ось річна не дивлячись на те, що стала гладкою зберегла загальну форму з M12 (див. рис. 3.24).

3.2.7 Виконання перехресної перевірки

Перехресна перевірка за методом імітованих історичних прогнозів виконується за допомогою функції cross_validation().

Функція cross_validation () повертає таблицю з (y) і оціненими (yhat) значеннями моделюється змінної, а також довірчими межами передбачених значень (yhat_lower і yhat_upper) для кожної точки відліку cutoff і кожної дати ds відповідного прогнозного періоду:

```
> M3_cv <- cross_validation(M3, initial = 180,
+ period = 60,
+ horizon = 60,
+ units = "days")
Making 1 forecasts with cutoffs between 2013-05-01 14:41:17 and 2013-05-01 14:41:17
> head(M3_cv)
          y                   ds     yhat yhat_lower
1 5.767258 2013-05-01 16:56:28 5.489377   5.302122
2 5.476464 2013-05-01 19:54:46 5.518202   5.342440
3 5.529429 2013-05-02 11:21:29 5.457510   5.257806
4 5.475543 2013-05-02 17:07:52 5.465598   5.273830
5 5.529429 2013-05-03 00:13:04 5.442170   5.245641
6 5.481430 2013-05-03 09:12:59 5.448224   5.260908
  yhat_upper              cutoff
1   5.677654 2013-05-01 14:41:17
2   5.714800 2013-05-01 14:41:17
3   5.657069 2013-05-01 14:41:17
4   5.664044 2013-05-01 14:41:17
5   5.653335 2013-05-01 14:41:17
6   5.648030 2013-05-01 14:41:17
```

Рисунок 3.25 – Head

3.2.8 Метрики якості моделі

Для оцінки якості прогнозів моделей розглянемо наступні показники:

- Ефективне значення помилка (mean squared error, MSE);

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\,, \tag{3.1}$$

- Квадратний корінь з середньоквадратичної помилки (root mean squared error, RMSE);

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}\,, \tag{3.2}$$

- Середня абсолютна помилка (mean absolute error, MAE);

$$MAE = \sum_{i=1}^{n}|y_i - \hat{y}_i|\,, \tag{3.3}$$

- Середня абсолютна питома помилка (mean absolute percentage error, MAPE);

$$MAPE = \sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|, \tag{3.4}$$

- "Покриття" (coverage): частка істинних значень модельованої змінної, які знаходяться в межах довірчих границь прогнозу.

У наведених формулах y_i та $\hat{y}_i$ - це справжнє і передбачене значення модельованої змінної відповідно, а n - кількість спостережень.

Функція performance_metrics () має такі аргументи:

- df - таблиця, отримана за допомогою функції cross_validation ();
- metrics - вектор з назвами метрик якості моделі (за замовчуванням цей аргумент приймає значення NULL, що призводить до розрахунку всіх перерахованих вище метрик, тобто c ("mse", "rmse", "mae", "mape", " coverage "));
- rolling_window - розмір "ковзного вікна", в межах якого відбувається усереднення кожної метрики (за замовчуванням приймає значення 0.1, тобто 10% від довжини прогнозного горизонту).

Застосуємо функцію performance_metrics () для розрахунку середньоквадратичної помилки прогнозу моделі M3:

```
> performance_metrics(M3_cv, metrics = "mse",
+                     rolling_window = 0.1) %>% head()
```

	horizon	mse
1	149.2614 hours	0.04971235
2	152.0100 hours	0.04499001
3	170.1089 hours	0.04641615
4	194.6967 hours	0.04663727
5	196.5764 hours	0.04690691
6	237.1867 hours	0.04776721

Рисунок 3.26 – M3_cv with window 0.1

Як видно з отриманого результату, перше усереднене значення MSE доводиться на 6-й день прогнозного горизонту, оскільки довжина цього горизонту для моделі M3 становить 60 днів, а 6 - це 10% від цієї довжини (розмір ковзаючого вікна, що задається аргументом rolling_window).

Якщо аргументу rolling_window привласнити значення 0, то запитувані метрики якості будуть розраховані для кожної дати прогнозного горизонту (тобто розмір ковзного вікна в даному випадку фактично дорівнює 1):

```
> performance_metrics(M3_cv, metrics = "mse",
+                     rolling_window = 0) %>% head()
```

	horizon	mse
1	2.253056 hours	7.721770e-02
2	5.224722 hours	1.742135e-03
3	20.670000 hours	5.172409e-03
4	26.443056 hours	9.889576e-05
5	33.529722 hours	7.614173e-03
6	42.528333 hours	1.102677e-03

Рисунок 2.27 – M3_cv with window 0

Якщо ж аргументу rolling_window привласнити значення 1, то запитувані метрики якості будуть усереднені по усьому прогнозному горизонту:

```
> performance_metrics(M3_cv, metrics = "mse",
+                     rolling_window = 1) %>% head()
```

	horizon	mse
1	1440 hours	0.4670725

Рисунок 3.28 – M3_cv with window 1

Метрики якості моделей, отримані в ході перехресної перевірки, можна візуалізувати за допомогою функції plot_cross_validation_metric(), яка має такі аргументи:

- df_cv - таблиця, отримана за допомогою функції cross_validation ();
- metric - назва метрики;
- rolling_window - розмір "ковзного вікна", в межах якого відбувається усереднення метрики.

Функція plot_cross_validation_metric () повертає об'єкт класу ggplot:

```
> plot_cross_validation_metric(M3_cv, metric = "mse",rolling_win
dow = 0.1)
```

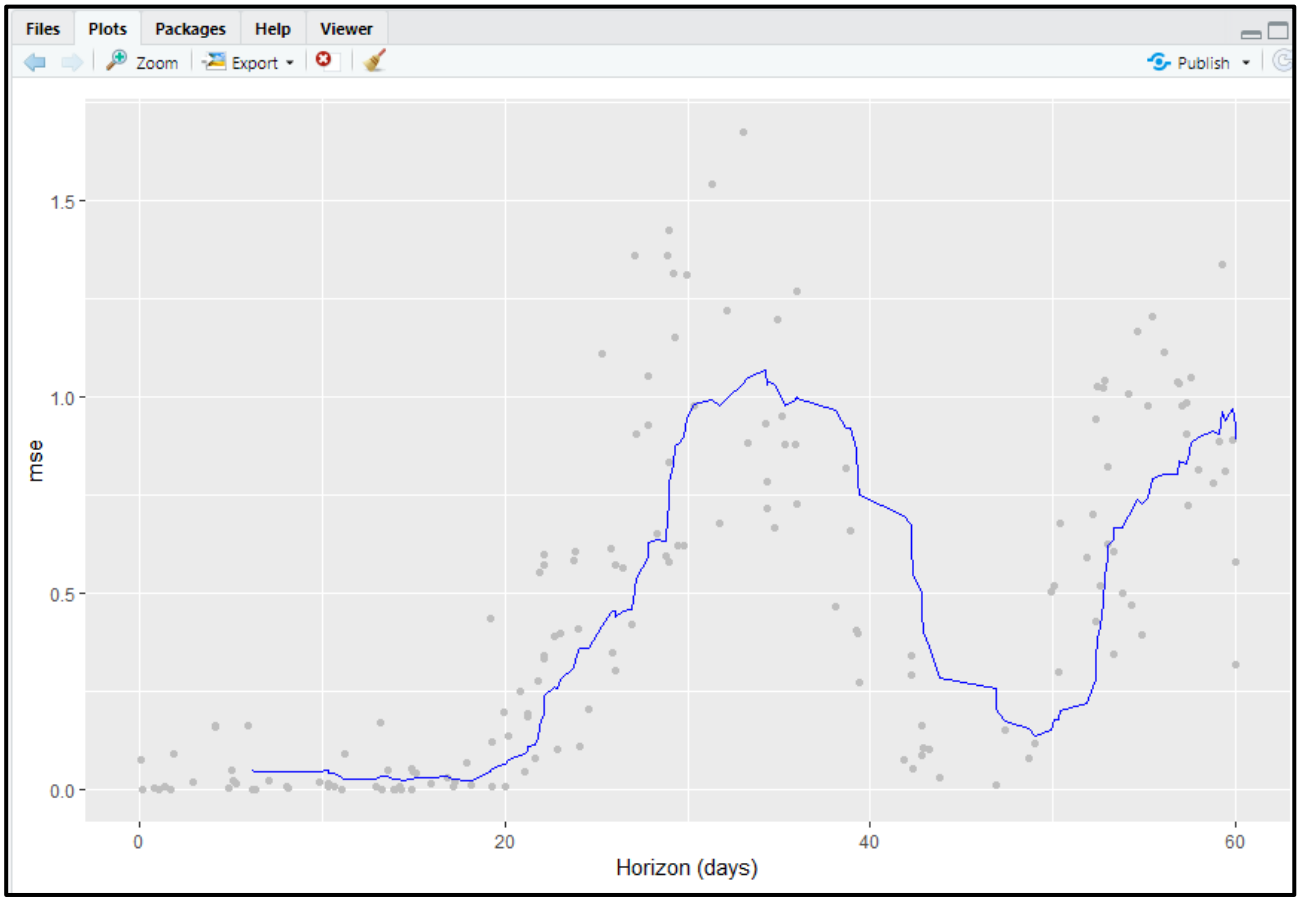

Рисунок 3.29 – Візуалізація метрики якості передбачень (MSE), отриманої за результатами перехресної перевірки моделі M3

На рис. 3.29 наведені оцінки MSE для кожної з дат прогнозного горизонту (H = 60) кожного з K = 5 блоків даних, які брали участь в перехресній перевірці. Блакитна лінія відповідає усередненим значенням в межах кожного ковзного вікна розміром в 6 спостережень. Судячи з великого розкиду отриманих оцінок MSE, якість моделі M3 не найкраще.

3.2.9 Вибір оптимальної моделі

Застосуємо методологію виконання перехресної перевірки для вибору оптимальної моделі з декількох альтернативних. Припустимо, що перед нами стоїть завдання вибрати оптимальну модель вартості room з побудованих раніше моделей M4, M12 і M15. Для опису якості цих моделей скористаємося всіма

метриками: MSE, RMSE, MAE, MAPE і покриття. Для спрощення прикладу припустимо також, що нас цікавить якість прогнозів в цілому для 60-денного прогнозного горизонту (тобто нам нецікаві окремі дати цього горизонту). Розрахуємо обидві метрики якості для кожної з моделей-кандидатів:

```
> M4_cv <- cross_validation(M4, initial = 180,period = 120,horiz
on = 60,units = "days")
> M12_cv <- cross_validation(M12, initial = 180,period = 120,hor
izon = 60,units = "days")
> M15_cv <- cross_validation(M15, initial = 180,period = 120,hor
izon = 60,units = "days")
>
> M4_perf <- performance_metrics(M4_cv,metrics = c("mse", "rmse"
, "mae", "mape", "coverage"), rolling_window = 1)
> M12_perf <- performance_metrics(M12_cv,metrics = c("mse", "rms
e", "mae", "mape", "coverage"),rolling_window = 1)
> M15_perf <- performance_metrics(M15_cv,metrics = c("mse", "rms
e", "mae", "mape", "coverage"),rolling_window = 1)
>
```

```
> M4_perf
     horizon      mse      rmse       mae      mape
1 1440 hours 0.549967 0.7415976 0.6301036 0.1123622
   coverage
1 0.2189349
> M12_perf
     horizon        mse      rmse       mae       mape
1 1440 hours 0.04902704 0.2214205 0.1520675 0.02787672
   coverage
1 0.6923077
> M15_perf
     horizon       mse      rmse       mae       mape
1 1440 hours 0.2777545 0.5270241 0.4515766 0.08070329
   coverage
1 0.2189349
>
```

Рисунок 3.30 – Результати performance для M4_perf, M12_perf та M15_perf

Як бачимо, M12 краще за інші моделі по всім вибраним метрикам якості. Це можна бачити також з графіків, побудованих за допомогою функції plot_cross_validation_metric() (див. рис. 3.30):

```
> M4_cv_plot <- plot_cross_validation_metric(M4_cv,metric =
"mape",rolling_window = 0.1) +
```

```
     +     ylim(c(0, 0.15)) + ggtitle("M4")
     > M12_cv_plot <- plot_cross_validation_metric(M12_cv,metric =
"mape",rolling_window = 0.1) +
     +     ylim(c(0, 0.15)) + ggtitle("M12")
     > M15_cv_plot <- plot_cross_validation_metric(M15_cv,metric =
"mape",rolling_window = 0.1) +
     +     ylim(c(0, 0.15)) + ggtitle("M15")
     >
     > gridExtra::grid.arrange(M4_cv_plot, M12_cv_plot, M15_cv_plot,
ncol = 3)
```

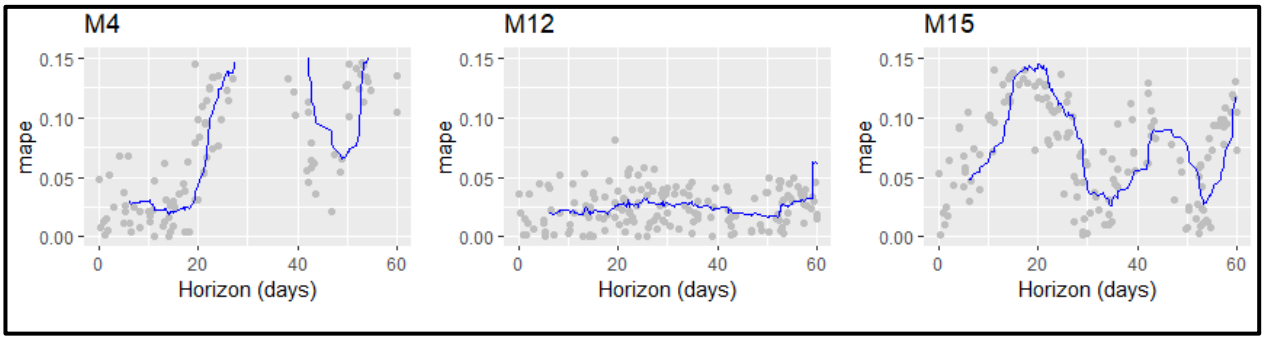

Рисунок 3.31 – Порівняння якості передбачень трьох моделей по метриці MAPE

На рис. 3.31 видно, що якість моделей недостатньо задовільна, скоріш за все із-за малої кількості даних.

До сих пір ми будували всі моделі по навчальних даних з таблиці room_train. Однак у нас є і перевірки набір даних - room_test. Подивимося, як обрана оптимальна модель M12 спрацює на цій перевірочної вибірці. На рис. 3.32 представлені навчальні дані і істинні значення вартості room в прогнозному періоді. Блакитна суцільна лінія на цьому графіку відповідає передбаченим моделлю значень, а світло-блакитна смуга навколо неї - 80% -ної довірчою області передбачених значень:

```
     > plot(M12, forecast_M12) +
     + coord_cartesian(xlim = c(as.POSIXct("2012-12-21"),as.POSIXct(
"2013-04-19")))
```

```
    + geom_point(data = room_test, aes(as.POSIXct(date_time), price
_usd), col = "red")
```

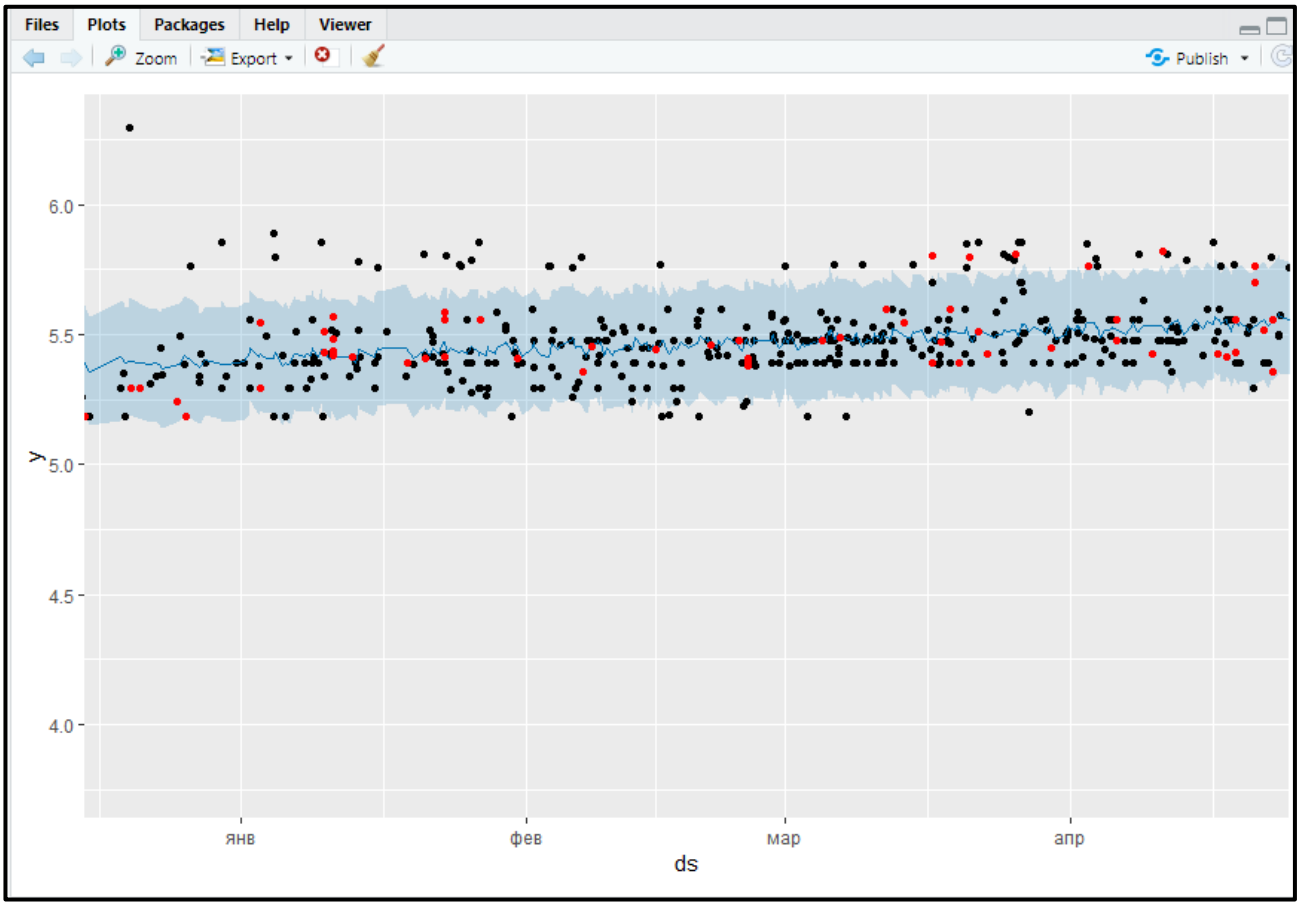

Рисунок 3.32 – Порівняння дійсних значень вартості номерів і прогнозних значень, отриманих за допомогою моделі M12

Хоча обрана в якості оптимальної модель M12 не змогла вірно передбачити деякі локальні коливання вартості room в прогнозному періоді, в цілому вона має достатньо задовільний результат: більшість справжніх значень вартості виявилося в межах 80%-ної довірчої смуги.

Висновки до розділу 3

На основі проведених досліджень і аналізу результатів можна зробити висновок про ефективність інформаційної системи прогнозування часових рядів на основі методології GAM на прикладі готельних номерів.

Під час роботи з даною системою було здійснено обробку вхідних даних, включаючи перевірку їх на відповідність вимогам моделей прогнозування. Виявлено, що дані представляють собою часовий ряд, складений з трьох змінних: ідентифікатор номера, дата та час, ціна номера в USD. Враховано особливості цих даних, такі як відсутність чітко вираженої річної сезонності та зміна тренду з плином часу.

Для прогнозування значень часового ряду було застосовано методологію GAM (Generalized Additive Models), яка дозволяє враховувати тренд, циклічність та сезонність в даних. Цей підхід дозволяє виявити шаблони та закономірності у часовому ряді та забезпечує гнучкість у моделюванні.

В результаті аналізу була вибрана оптимальна модель прогнозування часових рядів на основі методології GAM. Ця модель здатна відповідно відтворити шаблони та закономірності в даних готельних номерів, що дозволяє зробити більш точні прогнози щодо цін на номери в майбутньому.

У цілому, інформаційна система прогнозування часових рядів на основі методології GAM є ефективним інструментом для аналізу та прогнозування цін на готельні номери. Вона дозволяє здійснити комплексний аналіз даних, виявити закономірності та забезпечити точні прогнози, що може бути використано для прийняття рішень у готельній індустрії та планування бізнес-процесів.

ВИСНОВКИ

У даній кваліфікаційній роботі була розроблена інформаційна система прогнозування часових рядів на основі методології GAM (Generalized Additive Models). Застосування даної методології дозволило покращити точність прогнозування та забезпечити більш гнучкий підхід до моделювання складних тенденцій та сезонності в часових рядах.

У процесі роботи була проведена літературний огляд з питань прогнозування часових рядів та методології GAM, що дозволило зрозуміти сутність методу та його потенціал у вирішенні задач прогнозування. Також було проведено аналіз існуючих інформаційних систем для прогнозування часових рядів та виявлено їхні обмеження та недоліки, що підкреслило актуальність розробки нового підходу.

На основі теоретичних основ методології GAM було розроблено та реалізовано інформаційну систему, яка включає в себе такі компоненти, як збір та підготовка даних, побудова моделі GAM, оцінка точності прогнозів та візуалізація результатів. Система була реалізована з використанням сучасних програмних інструментів та технологій, що забезпечило її ефективність та зручність використання.

Отже, розроблена інформаційна система прогнозування часових рядів на основі методології GAM є ефективним інструментом для готельного бізнесу, дозволяючи забезпечити оптимальне управління кількістю доступних номерів та планування ресурсів. Її використання може призвести до підвищення доходів готелів, зниження недохідності та поліпшення задоволеності клієнтів шляхом оптимального використання готельного обладнання та персоналу.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Формування цінової політики та механізмів ціноутворення на ринку готельних послуг: веб-сайт. URL: https://vseosvita.ua/library/embed/01007ag1-389e.docx.html (дата звернення: 02.05.2023).

2. Business hotels: features, benefits, requirements: https://ribashotelsgroup.ua/en/blog/biznes-oteli-osobennosti-preimushtestva-trebovaniya/ (дата звернення: 03.05.2023).

3. Forecasting data and methods. *Otexts* : веб-сайт. URL: https://otexts.com/fpp3/data-methods.html (дата звернення: 04.05.2023).

4. The basic steps in a forecasting task. *Otexts* : веб-сайт. URL: https://otexts.com/fpp3/basic-steps.html (дата звернення: 10.02.2023).

5. Dynamic generalised additive models (DGAMs) for forecasting discrete ecological time series веб-сайт. URL: https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/2041-210X.13974 (дата звернення: 04.05.2023).

6. Forecasting high resolution electricity demand data with additive models including smooth and jagged components URL: https://www.sciencedirect.com/science/article/abs/pii/S0169207020300583 (дата звернення: 04.05.2023).

7. Dynamic generalised additive models (DGAMs) for forecasting discrete ecological time series URL: https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/2041-210X.13974 (дата звернення: 05.05.2023).

8. On the Use of Generalized Additive Models in Time-Series Studies of Air Pollution and Health. URL: https://academic.oup.com/aje/article/156/3/193/71628 (дата звернення: 05.05.2023).

9. Forecasting tourism growth with State-Dependent Models. URL: https://www.sciencedirect.com/science/article/pii/S0160738322000366 (дата звернення: 06.05.2023).

10. Forecasting with Bayesian Dynamic Generalized Linear Models in Python. URL: https://towardsdatascience.com/forecasting-with-bayesian-dynamic-generalized-linear-models-in-python-865587fbaf90 (дата звернення: 06.05.2023).

11. Generalized Additive Model. URL: https://www.sciencedirect.com/topics/social-sciences/generalized-additive-model (дата звернення: 06.05.2023).

12. What Is Time Series Forecasting? URL: https://thenewstack.io/what-is-time-series-forecasting/ (дата звернення: 06.05.2023).

13. Doing magic and analyzing seasonal time series with GAM (Generalized Additive Model) in R. URL: https://petolau.github.io/Analyzing-double-seasonal-time-series-with-GAM-in-R/ (дата звернення: 07.05.2023).

14. Dynamic Generalised Additive Models (DGAM) for forecasting discrete ecological time series. URL: https://www.biorxiv.org/content/10.1101/2022.02.22.481550v1.full (дата звернення: 07.05.2023).

15. When to use generalized additive models. URL: https://crunchingthedata.com/when-to-use-generalized-additive-models/ (дата звернення: 07.05.2023).

16. Time Series Analysis with Generalized Additive Models. URL: https://www.kdnuggets.com/2017/04/time-series-analysis-generalized-additive-models.html (дата звернення: 09.05.2023).

17. Cross-validation: evaluating estimator performance. URL: https://scikit-learn.org/stable/modules/cross_validation.html (дата звернення: 09.05.2023).

18. Inference and computation with generalized additive models and their extensions. URL: https://link.springer.com/article/10.1007/s11749-020-00711-5 (дата звернення: 09.05.2023).

19. Comprehensive and Comparative Analysis of GAM-Based PV Power Forecasting Models Using Multidimensional Tensor Product Splines against Machine Learning Techniques: веб-сайт. URL: https://www.mdpi.com/1996-1073/14/21/7146 (дата звернення: 10.05.2023).

20. Janiesch C., Zschech P., Heinrich K. Machine learning and deep learning. *Electron Markets.* 2021. Vol. 31, P. 685–695. https://doi.org/10.1007/s12525-021-00475-2.

21. Young T., Hazarika D., Poria S., Cambria E. Recent Trends in Deep Learning Based Natural Language Processing. *IEEE Computational Intelligence Magazine*. 2018. Vol. 13. no. 3. P. 55-75. doi: 10.1109/MCI.2018.2840738.

22. Goodfellow I., Bengio Y., Courville A. Deep Learning. *MIT Press*. 2016. URL: https://www.deeplearningbook.org (дата звернення: 17.05. 2023).

23. Russell S. J., Norvig P. Artificial Intelligence: A Modern Approach : навч. посіб. Вид. 3-е. 2010. 1151 с. URL: https://zoo.cs.yale.edu/classes/cs470/materials/aima2010.pdf (дата звернення: 19.05.2023).

24. Serena H. C., Anthony J. J., John P. N. Artificial Intelligence techniques: An introduction to their use for modelling environmental systems. *Mathematics and Computers in Simulation*. 2008. Vol. 78. P. 379-400. https://doi.org/10.1016/j.matcom.2008.01.028.

25. Brynjolfsson E., McAfee A. The Business of Artificial Intelligence. *Harvard Business Review*. 2017. URL: https://hbr.org/2017/07/the-business-of-artificial-intelligence (дата звернення: 19.05.2023).

26. Jordan M. I., Mitchel T. M. Machine learning: Trends, perspectives, and prospects. *Science*. 2015. Vol. 349. P. 255-260. DOI: 10.1126/science.aaa8415.

27. Hastie T., Tibshirani R., Wainwrith M. Statistical Learning with Sparsity : монографія. 2015.

28. Muthukrishnan R., Rohini R. LASSO: A feature selection technique in predictive modeling for machine learning. *IEEE International Conference on*

Advances in Computer Applications (ICACA). 2016. P. 18-20, doi: 10.1109/ICACA.2016.7887916.

29. Bishop C. M. Pattern recognition and machine learning : підручник. Information Science and Statistics. 2006. 758 с.

30. What is statistical analysis? *Whatis* : веб-сайт. URL: https://www.techtarget.com/whatis/definition/statistical-analysis (дата звернення: 19.05.2023).

31. What is Statistical Analysis? Types, Methods and Examples. *Simplilearn* : веб-сайт. URL: https://www.simplilearn.com/what-is-statistical-analysis-article (дата звернення: 19.05.2023).

32. What is R? *Webarchive* : веб-сайт. URL: https://web.archive.org/web/20080724195808/http://wiki.r-project.org/rwiki/doku.php?id=getting-started:what-is-r:what-is-r (дата звернення: 19.05.2023).

33. Autoregressive integrated moving average. *Wikipedia* : веб-сайт. URL: https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average (дата звернення: 20.05.2023).

34. Ravindra K., Rattan P., Mor S., Aggarwal A. N. Generalized additive models: Building evidence of air pollution, climate change and human health. *Environment International.* 2019. Vol. 132.

35. Dehghan, A., Khanjani, N., Bahrampour, A. The relation between air pollution and respiratory deaths in Tehran, Iran- using generalized additive models. *BMC Pulm Med.* Vol. **18.** 2018. https://doi.org/10.1186/s12890-018-0613-9

36. Ravindra K. Emission of black carbon from rural households kitchens and assessment of lifetime excess cancer risk in villages of North India. *Environment International*. 2019. Vol. 122. P. 201-212. https://doi.org/10.1016/j.envint.2018.11.008.

37. Hastie T. J. Generalized Additive Models : навч. посіб. 1992. 59 с.

38. James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning. *Springer New York*. 2021. 607 с. https://doi.org/10.1007/978-1-0716-1418-1.

39. Wood S.N. Generalized Additive Models: An Introduction with R. *Chapman and Hall/CRC*. 2006. https://doi.org/10.1201/9781420010404.

40. Moritz S., Bartz-Beielstein T. ImputeTS: Time Series Missing Value Imputation in R. URL: https://cran.r-project.org/web/packages/imputeTS/vignettes/imputeTS-Time-Series-Missing-Value-Imputation-in-R.pdf.

ДОДАТОК А

Приклад коду для завантаження та перетворення даних у необхідний формат

```
> setwd('F:\\університет\\4 курс\\диплом')
> hotels_price <- read.csv("hotels_price.csv")
> fix(hotels_price)
> str(hotels_price)
'data.frame':          1647 obs. of  3 variables:
 $ prop_id  : num  83045 13252 73738 73738 13252 ...
 $ date_time: Factor w/ 1411 levels "2012-11-01T10:16:16Z",..: 263 745 745 612 831 831 269 1347 8
91 1196 ...
 $ price_usd: num  219 160 189 189 186 161 259 349 239 219 ...
> hotels_price $date_time=strptime(hotels_price $date_time, format = '%Y-%m-%dT%H:%M:%SZ')
> unique(hotels_price$prop_id)
[1] 83045 13252 73738
> room_price <- dplyr::select(dplyr::filter(hotels_price, prop_id == "83045"), price_usd, date_time)
> summary(room_price)
   price_usd        date_time
 Min.   : 43.06   Min.   :2012-11-01 11:40:14
 1st Qu.:219.00   1st Qu.:2013-01-24 16:34:48
 Median :239.00   Median :2013-03-13 08:44:50
 Mean   :249.73   Mean   :2013-03-13 01:04:32
 3rd Qu.:264.79   3rd Qu.:2013-05-05 17:28:51
 Max.   :541.00   Max.   :2013-06-30 20:09:46
> room_price $date_time=ymd_hms(room_price $date_time)
> str(room_price)
'data.frame':          738 obs. of  2 variables:
 $ price_usd: num  219 259 349 239 226 ...
 $ date_time: POSIXct, format: "2012-12-31 08:59:22" ...
> attach(room_price)
> as_tsibble(room_price, key=NULL, index=date_time)
# A tsibble: 738 x 2 [1s] <UTC>
   price_usd date_time
       <dbl> <dttm>
 1      186  2012-11-01 11:40:14
 2      193. 2012-11-01 15:40:09
 3      279  2012-11-01 15:59:28
 4      339  2012-11-02 12:11:58
 5      239  2012-11-02 22:52:48
 6      222. 2012-11-03 23:28:34
 7      219  2012-11-04 08:05:45
 8      226. 2012-11-05 15:20:23
 9      212  2012-11-05 18:38:32
10       219. 2012-11-06 13:11:44
# ... with 728 more rows
> str(room_price)
'data.frame':          738 obs. of  2 variables:
 $ price_usd: num  219 259 349 239 226 ...
 $ date_time: POSIXct, format: "2012-12-31 08:59:22" ...
```

```
> glimpse(room_price, width = 60)
Rows: 738
Columns: 2
$ price_usd <dbl> 219.00, 259.00, 349.00, 239.00, 226.00, ~
$ date_time <dttm> 2012-12-31 08:59:22, 2013-01-01 20:12:4~
> detach(room_price)

> plot(room_price$date_time, room_price$price_usd)
```

ДОДАТОК Б

Приклад коду прогнозування

```
> names(room_train)[1] <- "y"
> names(room_train)[2] <- "ds"

> M0 <- prophet(room_train, yearly.seasonality=TRUE, daily.seasonality=TRUE)
> str(M0)
> future_df<-make_future_dataframe(M0,periods = 90)
> forecast_M0<-predict(M0,future_df)
> prophet_plot_components(M0, forecast_M0)
```

ДОДАТОК В

Приклад коду для вибору оптимальної моделі

```
> M3_cv <- cross_validation(M3, initial = 180,
+ period = 60,
+ horizon = 60,
+ units = "days")
> head(M3_cv)
> performance_metrics(M3_cv, metrics = "mse",
+                     rolling_window = 0.1) %>% head()
> performance_metrics(M3_cv, metrics = "mse",
+                     rolling_window = 0) %>% head()
> performance_metrics(M3_cv, metrics = "mse",
+                     rolling_window = 1) %>% head()
> plot_cross_validation_metric(M3_cv, metric = "mse",rolling_window = 0.1)
> M4_cv <- cross_validation(M4, initial = 180,period = 120,horizon = 60,units = "days")
> M12_cv <- cross_validation(M12, initial = 180,period = 120,horizon = 60,units = "days")
> M15_cv <- cross_validation(M15, initial = 180,period = 120,horizon = 60,units = "days")
>
> M4_perf <- performance_metrics(M4_cv,metrics = c("mse", "rmse", "mae", "mape", "covera
ge"), rolling_window = 1)
> M12_perf <- performance_metrics(M12_cv,metrics = c("mse", "rmse", "mae", "mape", "cove
rage"),rolling_window = 1)
> M15_perf <- performance_metrics(M15_cv,metrics = c("mse", "rmse", "mae", "mape", "cove
rage"),rolling_window = 1)
>
> M4_cv_plot <- plot_cross_validation_metric(M4_cv,metric = "mape",rolling_window = 0.1)
+
+    ylim(c(0, 0.15)) + ggtitle("M4")
> M12_cv_plot <- plot_cross_validation_metric(M12_cv,metric = "mape",rolling_window =
0.1) +
+    ylim(c(0, 0.15)) + ggtitle("M12")
> M15_cv_plot <- plot_cross_validation_metric(M15_cv,metric = "mape",rolling_window =
0.1) +
```

```
+     ylim(c(0, 0.15)) + ggtitle("M15")
>
> gridExtra::grid.arrange(M4_cv_plot, M12_cv_plot, M15_cv_plot, ncol = 3)
> plot(M12, forecast_M12) +
+ coord_cartesian(xlim = c(as.POSIXct("2012-12-21"),as.POSIXct("2013-04-19")))
+ geom_point(data = room_test, aes(as.POSIXct(date_time), price_usd), col = "red")
```