МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
СХІДНОУКРАЇНСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМ. В. ДАЛЯ
ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ ТА ЕЛЕКТРОНІКИ
КАФЕДРА КОМП'ЮТЕРНИХ НАУК ТА ІНЖЕНЕРІЇ

До захисту допускається
в.о. завідувача кафедри
_____ Рязанцев О.І.
«____»_____ 20__ р.

МАГІСТЕРСЬКА РОБОТА

НА ТЕМУ:

Методи прогнозування і кластеризації медичних даних

Освітній рівень"Магістр"
Спеціальність 123 "Комп'ютерна інженерія"

Науковий керівник роботи: _____ (підпис) В.М.Барбарук (ініціали, прізвище)

Консультант з охорони праці: _____ (підпис) Я.О.Критська (ініціали, прізвище)

Студент: _____ (підпис) Д.О. Зубенко (ініціали, прізвище)

Група: КІ-19дм

Сєвєродонецьк 2020

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
СХІДНОУКРАЇНСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВОЛОДИМИРА ДАЛЯ

Факультет Інформаційних технологій та електроніки
Кафедра Комп'ютерних наук та інженерії
Освітній рівень магістр
Напрям підготовки _____
(шифр і назва)
Спеціальність 123 "Комп'ютерна інженерія"
(шифр і назва)

ЗАТВЕРДЖУЮ:

Т.в.о. завідувача кафедри _____
_____ В.С.Кардашук
«_____» _____20____р.

З А В Д А Н Н Я
НА МАГІСТЕРСЬКУ РОБОТУ СТУДЕНТУ

Зубенку Дмитру Олеговичу
(прізвище, ім'я, по батькові)

1. Тема роботи Методи прогнозування і кластеризації медичних даних

керівник проекту (роботи) Барбарук Віктор Миколайович, к.т.н., доц.
(прізвище, м.'я, по батькові, науковий ступінь, вчене звання)
затверджені наказом вищого навчального закладу від «5» 10 2020 р. № 140/15.15
2. Строк подання студентом роботи 10.01.2021
3. Вихідні дані до роботи Матеріали науково-дослідної практики, математичні моделі кластеризації даних, перелік використовуваних програмних засобів, теоретичні відомості про методи кластеризації даних з пропусками, прогнозування при обробці часових рядів

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити)

охорона праці та безпека в надзвичайних ситуаціях, висновки

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень)
Електронні плакати

6. Консультанти розділів проекту (роботи)

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Охорона праці та безпека в надзвичайних ситуаціях	Критська Я.О. ст. викл. кафедри КНІ		

7. Дата видачі завдання 14.10.2020

Керівник _____
(підпис)

Завдання прийняв до виконання _____
(підпис)

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів дипломного проекту (роботи)	Строк виконання етапів проекту (роботи)	Примітка
1	Розробка технічного завдання	02.09.2020-15.09.2020	
2	Аналіз літератури	16.09.2020-22.09.2020	
3	Розробка методів прогнозування	23.09.2020-25.09.2020	
4	Реалізація методів	26.09.2020-06.10.2020	
5	Аналіз результатів дослідження	07.10.2020-25.11.2020	
6	Розробка частини проекту "Охорона праці та безпеки в надзвичайних ситуаціях"	26.11.2020-1.12.2020	
7	Оформлення пояснювальної записки, автореферату та презентації	2.12.2020-09.01.2021	

Студент _____ Зубенко Д.О.
(підпис) (прізвище та ініціали)

Науковий керівник _____ Барбарук В.М.
(підпис) (прізвище та ініціали)

АНОТАЦІЯ

Зубенко Д.О. Методи прогнозування і кластеризації медичних даних

Об’єктом дослідження є часові ряди з пропущеними значеннями, що піддаються кластеризації для прогнозування. Метою є розгляд існуючих методів позбавлення від пропусків в даних в задачах кластеризації та доцільність їх використання в реальних задачах. Вирішується задача кластеризації та прогнозування часових рядів.

Розглянуто найпоширеніші методи кластеризації даних та позбавлення від пропусків в задачах кластеризації даних. Дослідження проблеми кластеризації даних, що містять пропущені значення та розгляд методів, які дозволяють розв’язати цю задачу. Проведення експериментів та порівняння результатів кожного з методів, висновки про доцільність використання того чи іншого методу та побічні ефекти.

У результаті роботи проведено аналіз методів кластеризації та прогнозування часових рядів, що можуть містити пропуски.

Ключові слова: кластерний аналіз, неповні дані, обробка та аналіз даних, data mining, fcm, методи відновлення даних, часові ряди, прогнозування.

ABSTRACT

Zubenko D. Medical data forecasting and clustering system

The object of the study is time series with missing values, which are clustered for forecasting. The goal is to consider existing methods for getting rid of data gaps in clustering problems and the expediency of their use in real-world problems. The task of clustering and forecasting time series is being solved.

The most common methods of data clustering and getting rid of gaps in data clustering problems are considered. Investigation of the problem of clustering data containing missing values, and consideration of methods that can solve this problem. Conducting experiments and comparing the results of each of the methods, conclusions about the advisability of using a particular method and side effects.

As a result of implemented, the analysis of clustering methods and forecasting of time series, which may contain omissions.

Keywords: cluster analysis, incomplete data, data processing and analysis, data mining, fcm, data recovery methods, time series, forecasting.

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ ТА УМОВНИХ ПОЗНАЧЕНЬ

КА – кластерний аналіз

ПО – предметна область

НМ – нечітка множина

DTW – dynamic time warping

ВСТУП

Все більший інтерес породжує сфера обробки і аналізу даних. Якщо раніше обчислювальні машини, а разом з ними і комп'ютерні науки розвивалися досить повільно і основний інтерес становив розвиток саме апаратної бази: збільшення пам'яті, як оперативної так і жорстких дисків, загальна швидкість обчислень та інше. Розвиток програмного забезпечення був обмежений цими характеристиками, необхідно було розв'язувати проблеми працездатності програми в таких умовах, думати про те як скоротити ресурси, що використовуються, зменшити розміри коду. Це ті задачі, які ставилися і вирішувались раніше і були тоді найбільш популярними.

З доволі швидким розвитком комп'ютерних технологій і наук пов'язаних з тим, задачі, що вирішуються, стали змінюватися. На сьогодні проблема з пам'яттю або обчислювальними ресурсами не стоїть так гостро, апаратна база стрімко розвивається і більше не становить основний інтерес. Ще однією рушійною силою є те, що комп'ютери стали все більш доступними і вже доволі тривалий час не малий вклад в розвиток роблять прості користувачі. Враховуючи це, збільшується кількість інформації, що потрапляє в комп'ютер і безпосередньо в Інтернет.

Зі збільшення об'ємів даних стали виникати нові задачі, все більший інтерес становить робота з даними і розв'язання проблем, що пов'язані з їх обробкою і подальшим аналізом. Оскільки можливості створення нового контенту мають всі бажаючи – обсяги даних непомірно зростають, а їх впорядкованість досить слабка. З'являється необхідність пошуку необхідних даних, в тому числі зображень, їх обробки, для маркетингових і статистичних досліджень, використанню у інших сферах, а також для простого користувача.

Найпопулярнішими напрямками досліджень наразі все частіше стають: Big Data, Data Mining, Machine Learning. Постають питання добування даних їх глобальний і інтелектуальний аналіз. Серед таких актуальних завдань знаходять своє місце і поняття класифікації та кластеризації.

Задачі подібні, але основна відмінність полягає у тому, що кластеризація передбачає розбиття за умови початкової невизначеності щодо конкретних груп, вона може мати критерії щодо кількості кінцевих кластерів, але не їх зміст, загалом, можна сказати, що це навчання без вчителя.

Отже кластеризація становить інтерес, як спосіб попередньої обробки даних, для більш зручного їх подальшого аналізу. Отримавши необхідні групи, а також їх центроїди

можна продовжувати роботу вже з конкретними представниками, а не з усім набором даних, що особливо актуально в умовах безкінечно зростаючого об’єму інформації. Такий підхід дозволяє: краще зрозуміти дані, шляхом використання для кожного кластеру найбільш підходящого алгоритму аналізу; провести стиснення, виділивши найбільш типових представників, за умов збитковості даних; виявлення новизни, шляхом виділення об’єктів, що не потратили до жодного з кластерів.

Так, наприклад, однією з проблем кластеризації можна виділити роботу з пропущеними значеннями, а саме роботу з часовими рядами. Існує безліч методів, але вони не передбачають відсутності якоїсь кількості інформації. Але, як тільки ми виходимо за рамки тестових даних і переходимо до обробки реальних - стикаємося з проблемою з цією проблемою, адже в дійсності ідеальних даних не існує і всі вони містять шуми (некорисну інформацію, яка не вплине або навіть зашкодить результату), пропуски невідповідні формати та інше.

Таким чином є доцільним розглянути варіанти рішення проблеми кластеризації даних з пропусками, які існують методи та їх доцільність.

1 ОГЛЯД ОСНОВНИХ МЕТОДІВ КЛАСТЕРИЗАЦІЇ ЧАСОВИХ РЯДІВ ТА РОБОТИ З ПРОПУСКАМИ

За тривалий час розвитку людства та, безпосередньо, комп’ютерної техніки було накопичено досить велику кількість інформації. Слід зазначити, що ті дані, які на сьогодні, є в нашому розпорядженні не завжди добре структуровані та підготовлені одразу. Найчастіше, та інформація, що доступна є досить хаотичною, містить шуми (тобто некорисну інформацію, яка, скоріше за все, нам не знадобиться), дефекти. Тому зараз актуальними питаннями є зберігання та обробка цих даних.

Зважаючи на сучасні потреби все більш актуальними та затребуваними є розробки в галузі аналізу накопичених даних, адже ми живемо в той час коли кількість інформації, якою ти володієш, вже менш значуща ніж її якість і можливість обробити, зробити висновки на цьому підґрунті.

Одне з місць в обробці таких даних займає кластеризація. Кластеризація – це процес розбиття заданої вибірки об'єктів (спостережень) на підмножини (як правило, непересічні), які називаються кластерами, так, щоб кожен кластер складався з схожих об'єктів, а об'єкти різних кластерів істотно відрізнялися [1)]. Вона може бути корисною на початкових етапах проведення аналізу.

Можна виділити наступні основні завдання кластерного аналізу:

– розробка типології або класифікації;

– дослідження корисних концептуальних схем групування об'єктів;

– породження гіпотез на основі дослідження даних;

– перевірка гіпотез для визначення, чи дійсно типи (групи), виділені тим чи іншим способом, присутні в наявних даних.

На рисунку 1.1 наведено результат кластеризації, на якому можна побачити, що було сформовано три кластери (виділені різними кольорами), що характеризують найбільш наближені між собою точки.

Таким чином, в результаті отримуємо кластери, які характеризують свою групу і надають нам можливість працювати не з величезною кількістю різної інформації, а з одним представником, за яким згодом можна буде зробити висновок про весь кластер і кожну його складову.

Отже, на початку проведення аналізу дуже корисним буде провести кластеризацію, яка надалі спростить розрахунки. Набагато легше виділити групи схожих об’єктів,

вивчити їх особливості і побудувати для кожної групи окрему модель, ніж створювати одну загальну модель на всіх даних.

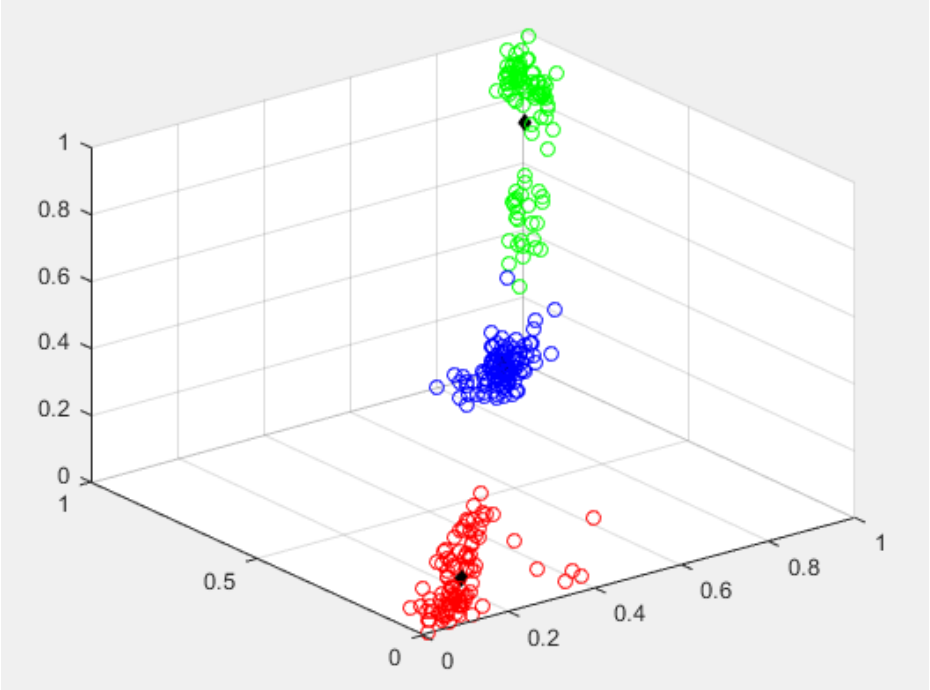


Рисунок 1.1 – Приклад результату кластеризації тестових даних у форматі 3D

Більш детально розглянемо приведений процес кластеризації [1)]:

Позначимо множину об’єктів, що спостерігається, як $\tau = \{x_i\}_{i=1}^{n}$ і складається з набору атрибутів $x_i = \{t_1^i, t_2^i, ..., t_m^i\}$, де t_i^v приймає значення із заданої множини T_v^i. Завдання кластеризації полягає в побудові множини $C = \{c_v\}_{v=1}^{k}$ і відображення $F : \tau \to C$ заданої множини об'єктів на множину кластерів.

Кластер містить записи з τ схожі (за заданим критерієм) один на одного

$$x_i \in c_v, x_j \in c_v \Rightarrow d(x_i, x_j) < \varepsilon,$$

де $d(x_i, x_j)$ – міра близькості між об'єктами (відстань), а ε - максимальне значення порога, що формує один кластер.

Сфера застосування може бути доволі широка, використання в сегментації зображень [**Ошибка! Источник ссылки не найден.**], прогнозування, аналіз текстів, оптимізація, машинне навчання. Якщо детальніше розглянути сфери застосування, можна побачити, що вони самі по собі є досить важливими і актуальними, сегментацію зображень можна використовувати для діагностики біологічного або біомедичного матеріалу, так у джерелі [**Ошибка! Источник ссылки не найден.**] наводяться основні

етапи сегментації цитологічних зразків на основі біонічних пікселів. Важливість аналізу та комп`ютерної обробки текстової інформації наводиться в роботі [**Ошибка! Источник ссылки не найден.**, 2)].

Як вже було зазначено, дані не завжди приходять у зручному для обробки, а саме кластеризації, вигляді. Часто виникають ситуації, коли частина інформації відсутня. Це можуть бути пропущені відповіді у анкетуванні, коли різні люди не відповіли на деякі запитання. Можлива ситуація втрати або пошкодження частини отриманої інформації, збій у роботі датчиків, помилки в роботі програми. І такі ситуації це не виключення, а найчастіше саме так і трапляється. Тому виникає необхідність перейти до даних, які не мають пропусків, для подальшої роботи з ними.

1.1 Існуючі методи кластеризації даних з пропусками

Отже, існують декілька можливих розв’язання задачі обробки даних з пропусками. Їх умовно можна поділити на такі, що націлені на попередню підготовку даних і ті, що вимагають адаптації (модифікації) стандартних алгоритмів кластеризації для обробки безпосередньо даних з пропусками. Кожні з цих варіантів має свої недоліки та переваги, і не є універсальними для боротьби з цією проблемою.

При застосуванні першого варіанту позбавлення від пропусків також є декілька підходів. По-перше, при підготовці даних є варіант видалення наявних пропусків. В такому випадку можна позбутися зайвої інформації, яка перешкоджає якісному аналізу даних, однак існує суттєвий недолік, втрачається інформація, яка може бути корисною, і якщо таких пропусків велика кількість, то є ризик видалити занадто багато і зробити подальший аналіз неможливим, або неінформативним. В цьому випадку можна видаляти не всі наявні дані, а лише ті за якими відбувається аналіз, або пропуски в яких є неприйнятними.

По-друге, відбувається заміна пропусків на їх оцінки, найчастіше це може бути середнє значення всієї вибірки, мода, медіана та інше і регресивні моделі, коли невідоме значення ознаки обчислюється за допомогою знайденої функції регресії за відомим ознаками. Широке поширення набув ЕМ-алгоритм, який передбачає вірогідну модель побудови вибірок.

Другий варіант полягає в використанні алгоритмів, що вже існують для застосування у обробці повних даних та адаптацію їх до даних з пропусками. У цьому випадку нема потреби проводити попередню обробку, а саме відновлення, але алгоритми,

що все існують, потребують модифікації. При такому підході не відбувається втрати даних і їх спотворення, на відміну від попереднього випадку, адже коли відсутня інформація заповнюється штучними даними це погіршує правильне отримання результатів. Однак до недоліків слід віднести необхідність розробки нових алгоритмів, або адаптацію наявних для вирішення такої задачі.

На рисунку 1.2 наведено графік існуючих алгоритмів заповнення пропусків в таблицях.

Отже, розглянувши основні методи та можливості обробки, а саме кластеризації даних з пропусками можна стверджувати, що на сьогодні не існує універсальної моделі, яка б розв’язувала цю проблему на будь-яких наборах даних. Слід обирати шляхи вирішення тієї чи іншої проблема попередньо проаналізувавши похідні дані і виходячи з них робити висновок про обраний шлях вирішення задачі.

Результати порівняння різних підходів показують, що при вирішенні практичних завдань апріорі не ясно, який з підходів виявиться найбільш прийнятним. Таким чином, створення нових підходів і алгоритмів кластеризації неповних даних є актуальним завданням.

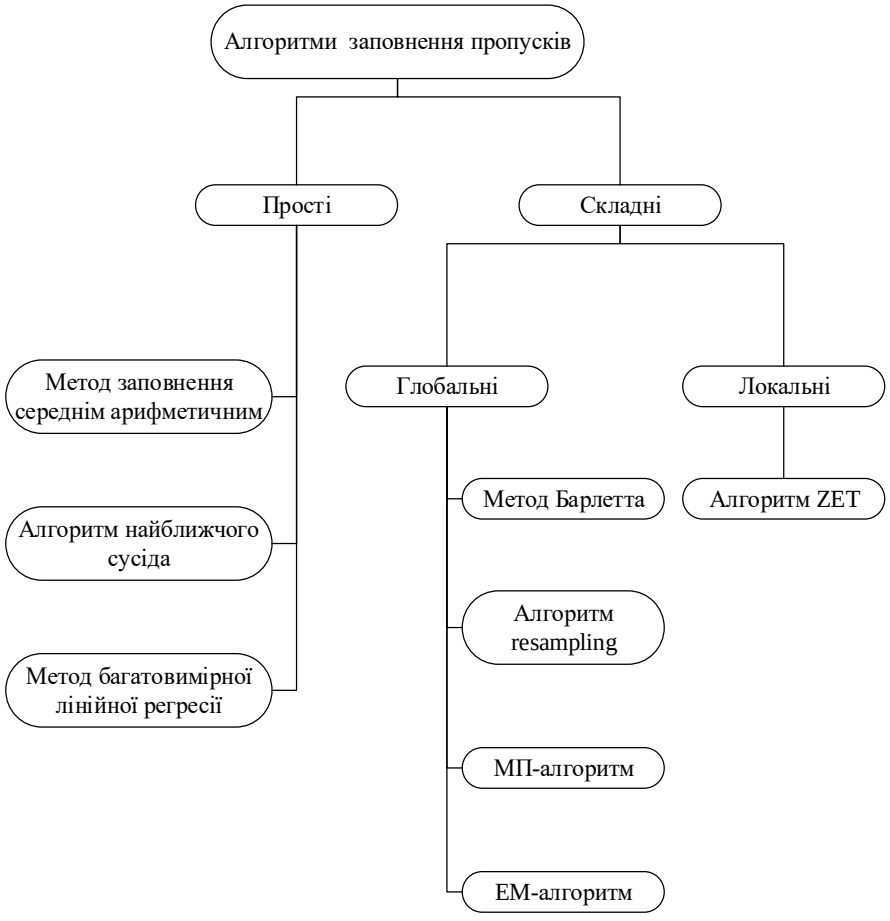


Рисунок 1.2 – Класифікація алгоритмів заповнення пропусків

1.2 Прості методи заповнення пропусків

1.2.1 Виключення рядків з наявністю пропущених значень

Це метод, що легко реалізувати, але він може призвести до суттєвої втрати важливих даних. Його можна використовувати лише тоді, коли пропуски в даних розміщені випадковим чином і їх доволі мало, щоб вплинути на кінцевий результат.

Даний метод з видалення всіх рядків, які містять пропуски, є найгіршим варіантом. Він можливий лише у випадках коли вибірка містить мінімальну кількість пропусків, або тоді коли було попередньо проведено інший вид обробки і відбувається видалення залишків пустих значень.

Такий підхід не дає коректних результатів, якщо кількість пропусків більше ніж 30%, за таких даних не вдається побудувати кластері побідні еталонним. Однак, слід зазначити, що у разі, коли кількість пропусків знаходиться близько 10% формування кластерів відбувається досить схоже на необхідний результат.

Але все ж, даний метод не підходить для аналізу даних зі значною кількістю пропусків, особливо якщо важлива велика точність розрахунків і певність результатів. Це найпростіший метод, який може бути використано у разі крайньої необхідності, де потреба швидкості отримання результатів перевищує їх точність.

1.2.2 Кластеризація із заповненням пропусків вибірковими статистиками

Підхід, що передбачає заповнення невідомих вибірковими статистиками (середнє, медіана, тощо), спирається на, так званий, “ наївний” підхід, який полягає у припущені, що взаємозв'язок між змінними в даному наборі даних відсутній.

В даному випадку, до недоліків слід віднести той факт, що після заповнення розподілення даних буде відрізнятися від похідних значень. , що деякі дані були замінені на “штучні”, безперечно призведе до спотворення результатів. Також можна віднести зменшення дисперсії.

Але попри це, на даних з пропусками приблизно 30% можна сказати, що такий вид обробки пропусків дає хороші результати, адже результати кластеризації подібні до еталонних.

Так, якщо говорити про задачі кластеризації, то практичні результати показують, що за 30% пропусків можна досягти припустимих результатів кластеризації. Але, все ж слід враховувати втрату певної достовірності і точності такого аналізу.

1.2.3 Заповнення пропусків з урахуванням структури зв'язків

Попередній метод передбачав відсутність зав'язків між параметрами, це так званий "наївний" метод. Альтернативою йому можна назвати метод, що враховує зв'язки між параметрами. Наприклад між такими параметрами медичної оцінки стану людини, як trestbps (артеріальний тиск у стані спокою) та віком пацієнта можна встановити кореляційний взаємозв'язок. Цей факт може допомогти відновити пропущені дані в параметрі, використовуючи рівняння простої регресії.

Як і попередній метод потребує випадкове розподілення пропусків і залежить від правильно обраного методу регресійного аналізу.

Серед простих методів боротьби з пропущеними значеннями при кластеризації, шляхом їх попередньої обробки, найкращим вважається метод боротьби з пропусками з урахуванням взаємозв'язків між полями, але не завжди він дає переваги при використанні, не всі набори можуть мати очевидні зв'язки тому значних переваг, у порівнянні з попереднім методом заміни на середні значення, не можна добитися.

Ефективність використання даних методів позбавлення від пропусків в даних для подальшого їх використання в задачах кластеризації розглянуто в роботі [**Ошибка! Источник ссылки не найден.**].

Отже, за результатами можна сказати, що у разі необхідності позбавитися від незначної кількості пропусків, можна використовувати метод їх видалення, у випадках, якщо пропусків близько 30% цей варіант розглядати не варто, слід звернути увагу на два інші методи, що є більш складними, але при цьому збільшують точність відновлення, та якість подальшої кластеризації. Адже аналізуючи отримані графіки і таблиці було зроблено висновок, про подібність результатів до еталонних значень.

1.3 Складні методи заповнення пропусків

Як представлено на рисунку 1.2 складні методи поділяються на локальні та глобальні. Детальний розгляд методів наводиться в роботі [2)].

1.3.1 Приклади глобальних методів

Метод Бардетта – це метод, що базується на рівняннях лінійної регресії. Містить декілька кроків, які дозволяють відновити пропущені значення. Спочатку всі пропуски заповнюються певними значеннями, у якості яких може слугувати те саме середнє значення, для обраних даних. Далі, враховуючи початкове значення, обране на попередньому кроці, будується регресійна модель. Ця модель надалі використовується для прогнозування нового значення для пропуска, що оброблявся.

Наступний метод – Resampling. Він полягає в тому, що на місця пропусків поміщаються довільні дані з тих, що наявні в наборі. Оскільки алгоритм є ітеративним, на кожній з ітерацій будується регресійна модель (отримуються рівняння лінійної регресії). Всі отримані рівняння аналізуються і підсумкове використовується для остаточного прогнозування з метою заповнення пропусків.

Перевагою методу є можливість отримати більш точного прогнозу за рахунок ітераційності зазначеного процесу.

МП-алгоритми - методи, що засновані на оцінках максимальної правдоподібності. Їх суть полягає у побудові моделі породження пропусків, згодом робляться висновки на підґрунті функцій правдоподібності. Дані методи висувають більш слабкі умови до початкових даних, що забезпечується можливістю врахувати специфіку конкретної області. Сам факт необхідності будувати модель пропусків можна віднести до недоліків, як фактор, що ускладнює метод.

ЕМ-алгоритм, назва кого є абревіатурою двох кроків, з яких він складається [1)]:

Е-крок (expectation)

$$g_{i,j}^{0} = g_{i,j};$$

$$g_{i,j} = \frac{\omega_i p_i\left(x_j\right)}{\sum_{v=1}^{k} \omega_v p_v\left(x_j\right)}, i = 1,\ldots,k,\ j = 1,\ldots,n. \quad (1.1)$$

$$\delta = \max\left\{\left|g_{i,j}^{0} - g_{i,j}\right|\right\}. \quad (1.2)$$

Таким чином, на цьому кроці, за допомогою регресії, відбувається оцінка значень, що можуть бути використані для заповнення пропусків.

М-крок (maximization)

$$\sum_{j=1}^{n} g_{i,j} \ln p_i(x_j) \to \max_{\Theta}, i = 1,...,k, \quad (1.3)$$

$$\omega_i = \frac{1}{n}\sum_{j=1}^{n} g_{i,j}, i = 1,...,k. \quad (1.4)$$

На цьому кроці розраховується нова коваріаційна матриця, враховуючи попередні заповнені пропуски.

Кроки Е та М виконуються доти, доки коваріаційна матриця не припинить змінюватися.

1.3.2 Приклади локальних методів

До локальних методів відновлення даних розглянемо ZET алгоритм, він у своїй роботі використовує лише необхідне частину даних, що обробляються.

ZET алгоритм базується на таких гіпотезах [8] як:

Гіпотеза збитковості, що полягає в припущені про наявність збиткових даних в реальних таблицях, в яких рядки можуть бути сході між собою, а стовпці знаходитися у відносинах залежності. Якщо такої збитковості не проявляється – алгоритм втрачає свою актуальність.

Гіпотеза локальної компактності. Припущення полягає у тому, що для прогнозування деякого об’єкта і-го рядка j-го стовпця необхідно використовувати лише певну частину даних, що складається з елементів рядків, що схожі на і-й рядок, та елементів стовпців, що схожі на j-й стовпець.

Гіпотеза лінійних залежностей, полягає у припущені про лінійність зв’язків між стовпцями та строками. Для кожного окремого пропуска є тільки певна кількість схожих рядків та стовпців.

Отже, на початку, для кожного пропуска, відбувається пропуск необхідних рядків та стовпців, враховуючи наведені умови. Потім використовується лінійна регресія задля прогнозування значень пропущених даних.

1.4 Методи кластеризації даних, алгоритм k-середніх

Виходячи з [4)], можна сказати, що методи кластеризації поділяються на ієрархічні та алгоритми розподілу (неієрархічні) алгоритми.

В процесі ієрархічної кластеризації відбувається злиття і поділ кластерів під час побудови дерева вкладених кластерів (дендрограми). На рисунку 1.4 наведено результат ієрархічної кластеризації та приклад дендрограми. Ієрархічні методи також поділяються на:

– агломеративні, на початку алгоритму кожен елемент – це окремий кластер, подальші кроки кластери поєднуються в один. Таким чином суть метода полягає в об’єднанні кластерів, зменшенні їх кількості;

– дивизивні (роз'єднувлаьні), що передбачають один кластер на початку і подальше його розділення на більшу кількість кластерів.

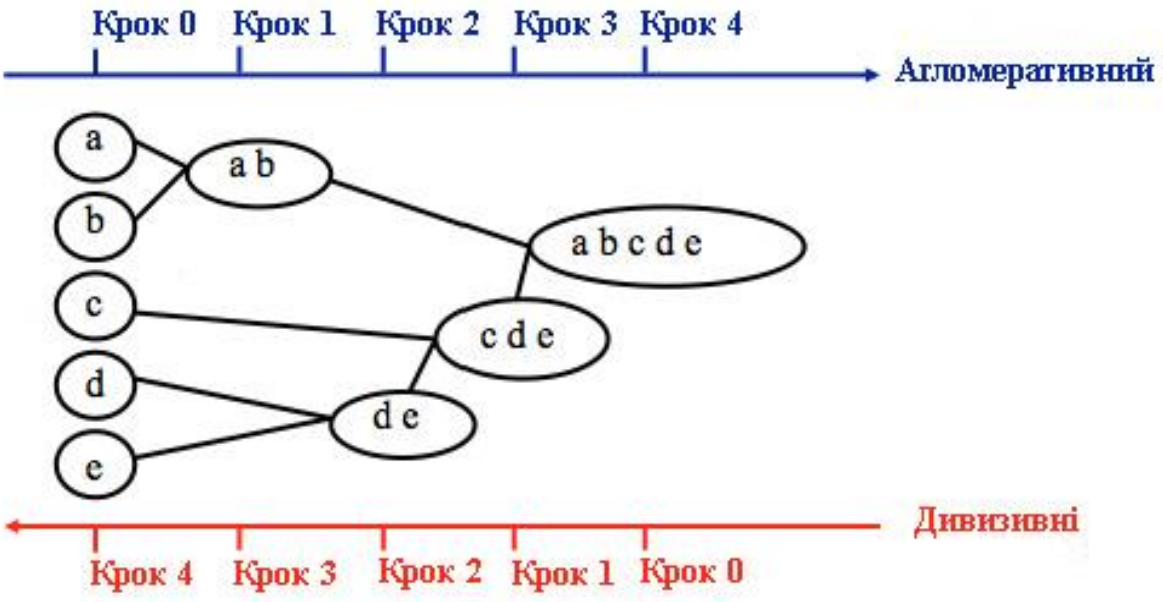


Рисунок 1.3 – Ієрархічні методи кластерного аналізу

При використанні агломеративних ієрархічних алгоритмів можна використовувати одне з правил об’єднання двох кластерів [4)]:

– метод ближнього сусіда (одиночний зв'язок);
– метод найбільш віддаленого сусіда (повний зв'язок);
– незважений центроїдний метод;
– зважений центроїдний метод;
– метод невиваженого попарного середнього;
– метод зваженого попарного середнього;
– метод Варда (Ward's method).

Метод ближнього сусіда полягає у об’єднанні двох найбільш близьких об’єктів серед різних кластерів, близькість визначається схожістю об’єктів у матриці збігів. На кожному з кроків продовжується пошук найбільш схожих сусідів і це відбувається до тих пір, поки матриця збігів не буде вичерпана.

До переваг методу можна віднести його простоту серед подібних методів, а також можливість виділяти кластери форми будь якої складності. Однак, метод містить і недоліки, він потребує критерій порогового рівня подібності за яким буде припинено об’єднання, та, як побічний ефект може утворитися один загальний кластер, що не відповідає цілям кластеризації.

$$\rho_{\min}(S_l, S_m) = \min_{x_i \in S_l; x_j \in S_m} \rho(x_i, x_j). \tag{1.5}$$

Метод найбільш віддаленого сусіда полягає у визначенні відстані між кластерами за найбільшою відстанню між будь-якими двома об’єктами в різних кластерах. Метод, на відміну від попереднього, не варто використовувати, якщо кластери мають подовжену форму.

$$\rho_{\min}(S_l, S_m) = \max_{x_i \in S_l; x_j \in S_m} \rho(x_i, x_j). \tag{1.6}$$

За незваженим центроїдним методом, у якості відстань між двома кластерами, визначається відстань між їх центрами ваги

$$\rho_{gc}(S_l, S_m) = \rho(\overline{x_j}, \overline{x_m}). \tag{1.7}$$

де $\overline{x_j}, \overline{x_m}$ - вектори середніх S_l та S_m відповідно.

Зважений центроїдний методаналогічний попередньому, але передбачає розрахунок з урахуванням того, що при визначенні розмірів кластерів використовуються ваги. Тому, у випадках, коли відомо про відмінності у розмірах кластерів бажаніше використовувати саме цей метод, у порівнянні з попереднім.

За методом невиваженого попарного середнього відстань між кластерами розраховується як середня відстань між усіма парами об’єктів в цих кластерах.

$$\overline{\rho}(S_l, S_m) = \frac{1}{n_l n_m} \sum_{x_i \in S_l} \sum_{x_j \in S_m} \rho(x_i, x_j). \tag{1.8}$$

Метод добре працює при використанні у роботі з кластерами протяжної форми та у випадках коли відомо про різні розміри кластерів.

Метод зваженого попарного середнього аналогічний попередньому, відрізняється тим, що при розрахунку розмір кластерів береться за ваговий коефіцієнт. Так само може бути використаний у випадках наявності кластерів різних розмірів, підходить для такої ситуації краще ніж попередній метод.

На відміну від попередніх методів метод Варда, для оцінки відстаней між кластерами використовує методи дисперсійного аналізу. В цьому випадку відстань між кластерами визначається як приріст суми квадратів відстаней об'єктів до центрів кластерів. Це досить ефективний метод, але за результатами його використання можна отримати кластери малого розміру.

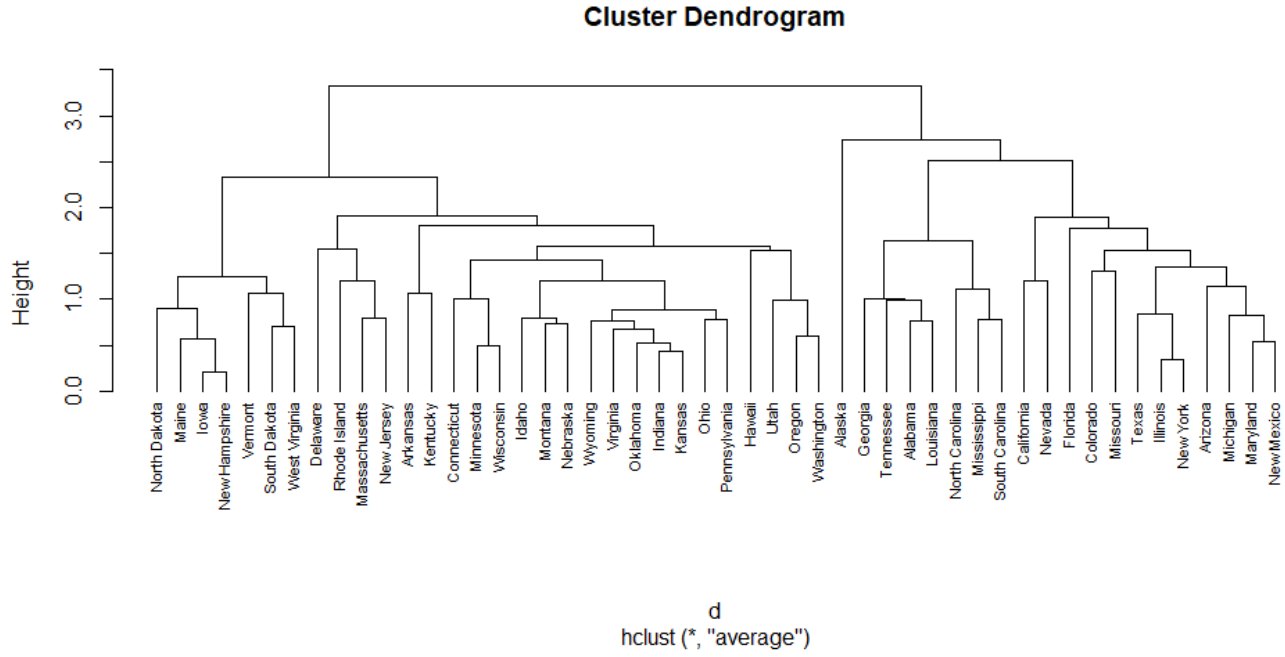


Рисунок 1.4 – Приклад дендрограми

Неієрархічні методи відрізаються тим, що потребують наявність умови зупинки і кількість кластерів. Це ітеративний процес поділу на кластери до тих пір поки не буде виконана умова зупинки. При великій кількості спостережень неієрархічні методи КА більш зручні оскільки у такому випадку дендограми стають перевантаженими і втрачають наочність.

Існує два підходи до визначення границь кластерів, при використанні неієрархічних алгоритмів. За першим кластери визначаються в місцях найбільшого скупчення точок, другий підхід полягає у мінімізації міри відмінності об'єктів.

Сам алгоритм розбиття, в загальному випадку, уявляє собою ітераційний процес з певним переліком кроків:

1) на першому кроці обирається/задається певна кількість k кластерів, які необхідно отримати із вхідних даних за результатом КА;

2) на першій ітерації, випадковим чином, обираються k центри кластерів, у якості яких слугують початкові записи;

3) далі, для кожного запису визначається найближчий до нього центр кластера;

4) наступним кроком виконується обчислення нових центроїдів (центрів тяжіння). Після цього центри кластерів, з попередніх кроків, замінюються на розраховані центроїди.

Останні два кроки виконуються ітераційно до тих пір, поки не буде виконано умову зупинки, у якості якої може бути перетин заданої максимальної кількості ітерацій,

або похибка менше заданої ε, коли границі кластерів і розташування центроїдів перестають змінюватися.

Найбільш популярним методом неієрархічної кластеризації є метод найменших квадратів:

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left|x_i - s_j\right|^2 \to \min. \tag{1.9}$$

і його чисельна реалізація називається методом k-середніх.

Алгоритм k –середніх полягає у тому, що на початку обирається k довільних центри далі, за цими центрами решта множини розбивається на групи. На наступному кроці обчислюються нові центри для отриманих кластерів таким чином, щоб [5)] квадрат евклідової відстані від елемента кластера до його центроїда був меншим ніж відстань до центроїдів решти кластерів.

У якості міри близькості зазвичай використовується Евклідова відстань:

$$\rho(x,y) = \|x - y\| = \sqrt{\sum_{p=1}^{n}\left(x_p - y_p\right)^2}. \tag{1.10}$$

де $x, y \in R^n$.

1.4.1 Метрики й відстані

Для формування кластерів в задачах КА використовуються метрики, які уявляють собою функцію, що визначає відстань між точками і класами в просторі R^p (метричний простір).

У роботі [6)] наводяться визначення відстані і метрики. Так, $d_{i,j} = d\left(x_i, x_j\right)$ називається відстанню, якщо воно задовольняє наступним умовам:

– невід'ємність: для всіх x_i та x_j має виконуватись умова $d\left(x_i, x_j\right) > 0$

$$\forall x_i, x_j: \quad d\left(x_i, x_j\right) > 0. \tag{1.11}$$

– шдентичність: $d\left(x_i, x_j\right)=0$, тоді й тільки тоді, коли $x_i = x_j$ тобто об’єкти співпадають

$$x_i = x_j \Leftrightarrow d\left(x_i, x_j\right)=0. \tag{1.12}$$

– симетричність: умова воркнується тоді, коли відстань від x_i до x_j дорівнює відстані від x_j до x_i

$$d\left(x_i, x_j\right)=d\left(x_j, x_i\right). \tag{1.13}$$

Для будь-яких трьох p-мірних точок x_i, x_j та x_l відстань $d\left(x_i, x_j\right)$ називається метрикою, якщо виконується додаткова умова:

– нерівність трикутника: відстань від x_i до x_j завжди менше або дорівнює сумі відстаней від кожної з них до x_l.

$$d\left(x_i, x_j\right) \le d\left(x_i, x_l\right)+d\left(x_l, x_j\right). \tag{1.14}$$

Після розрахунку відстані робиться висновок чи належить об’єкти одному кластеру таким чином, що при $d\left(x_i, x_j\right)<\sigma$ об’єкти розміщуються в один кластер, а у протилежному випадку об’єкти різні і будуть належати до різних кластерів.

У базовому алгоритмі нечітких k-середніх відстань між об'єктом і центром кластера розраховується через стандартну норму Евкліда., що наведено у формулі (1.6), яка є найбільш використовуваною для таких розрахунків, але існують і інші міри, що підбираються в залежності від даних, з якими необхідно працювати. Наприклад, Евклідова відстань використовується для роботи з кількісними даними, для них же можна застосувати Манхеттенську метрику (1.15) або відстань Махаланобіса (1.16) [8)]:

$$d_{ik}^{(1)} = \sum_{j=1}^{N}\left|x_{ij} - x_{kj}\right|. \tag{1.15}$$

$$d_{ik}^{M} = \left(x_{ij} - x_{kj}\right)^T W^{-1}\left(x_{ij} - x_{kj}\right). \tag{1.16}$$

де W – коваріаційна матриця вибірки $X = \left\{X_1, X_2, \ldots, X_n\right\}$.

Норма Махаланобіса дозволяє виділяти кластери у вигляді гіпереліпсоїдів, вісі яких можуть бути орієнтовані в довільних напрямах.

Для номінальних (якісних) даних використовується міра подібності Хеммінга (1.17) та міра подібності Роджерса-Танімото (1.18) [8)]:

$$\mu_{ij}^{H} = \frac{n_{ik}}{N}. \tag{1.17}$$

де n_{ik} – це число співпадаючих ознак у зразків X_i та X_k.

$$\mu_{ij}^{R-T} = n_{ik}^{N}\left(n_i^{'} + n_k^{'} - n_{ik}^{''}\right). \tag{1.18}$$

де $n_i^{'}$, $n_k^{'}$ - загальне число одиничних ознак у зразків X_i та X_k відповідно; $n_{ik}^{''}$ - число одиничних ознак, що збігаються, у зразків X_i та X_k.

Для змішаних типів ознак можна використовувати відстань Журавльова [8)]:

$$d_{ik} = \sum_{j=1}^{N} I_{ik}^{j}. \tag{1.19}$$

де $I_{ik}^{j} = \begin{cases} 1, якщо \left|x_{ij} - x_{kj}\right| < \varepsilon \\ 0, у іншому випадку \end{cases}$

Для наочного розгляду деяких наведених метрик на рисунку 1.6 наведено ізолінії деяких норм.

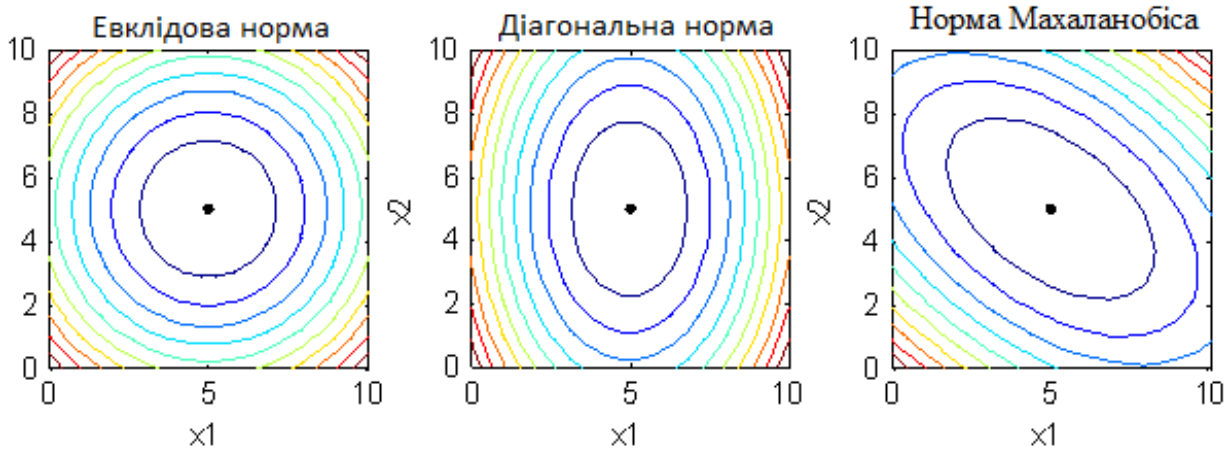


Рисунок 1.5 - Ізолінії різних норм

1.4.2 Міри близькості

Відстань $d\left(x_i, x_j\right)$ є різновидом міри близькості, однак існує ряд додаткових мір, які також можуть представляти інтерес.

Прикладом міри близькості є:

$$r\left(q_i, q_j\right) = \frac{v\left(q_i, q_j\right)}{v_0}. \tag{1.20}$$

де $v\left(q_i, q_j\right)$ – кількість однакових атрибутів у об'єктах q_i, q_j, що порівнюються; v_0 – загальне число атрибутів.

До мір близькості висуваються такі ж умови, як і розглянуті раніше, для відстаней (1.11-1.14), але на відміну від них, у даному разі виконання умови 1.14 не є обов'язковим.

Умови, яким відповідають міри близькості, аналогічні розглянутим раніше для відстаней:

Після розрахунку робиться висновок: при $r\left(q_i, q_j\right) < \sigma$ об'єкти розміщуються в один кластер, а у протилежному випадку об'єкти різні і будуть належати до різних кластерів.

Існує ряд асоціативних мір близькості:

Міра близькості Жаккара – це бінарна міра, що заснована на використанні інформації про мнодину загальних символів. Цю міру можна назвати першою відомою мірою близькості.

$$r_J\left(q_i, q_j\right) = \frac{c}{a+b+c}. \tag{1.21}$$

Міра близькості Сьоренсена – це статистичний коефіцієнт.

$$r_S\left(q_i, q_j\right) = \frac{2c}{a+b}. \tag{1.22}$$

де a і b число видів в зразках a і b, відповідно; c — число спільних для обох зразків видів.

Міра близькості Сокала-Сніта:

$$r_{SS}\left(q_i,q_j\right)=\frac{c}{2\left(a+b\right)+c}. \tag{1.23}$$

Міри близькості Дейка:

$$r_{SS}\left(q_i,q_j\right)=\frac{c}{2\left(a+b\right)+c}; \tag{1.24}$$

$$r_{D_2}\left(q_i,q_j\right)=\frac{c-\min\left(a,b\right)}{c+\min\left(a,b\right)}. \tag{1.25}$$

Міра близькості Кульчинського:

$$r_K\left(q_i,q_j\right)=\frac{a+b}{2ab}. \tag{1.26}$$

Міра незгоди Танімото:

$$r_T\left(q_i,q_j\right)=\frac{a+b}{a+b+c}. \tag{1.27}$$

Міра близькості Браун-Бланке:

$$r_B\left(q_i,q_j\right)=\frac{c}{\max\left(a,b\right)}. \tag{1.28}$$

1.4.3 Метод k-середніх

Метод k-середніх розділяє m спостережень на k груп (кластерів), щоб мінімізувати сумарне квадратичне відхилення точок кластерів від центроїдів кластерів:

$$min\left[\sum_{i=1}^{k}\sum_{x^{(j)}\in S_i}\left\|x^{(j)}-\mu_i\right\|^2\right]. \tag{1.29}$$

де $x^{(j)}\in R^n, \mu_i\in R^n$; μ_i – центроїд для кластера S_i.

Також, при використанні алгоритму слід звернути увагу на те, що k-means володіє низкою обмежень, таких як:

$u_{ij}^{(l)}\{0,1\}; \sum_{j=1}^{c} u_{ij} = 1; 0 < \sum_{j=1}^{d} u_{ij} < d$, тобто кожен вектор даних може належати лише одному кластеру і кожен кластер повинен містити не менше одного елемента і не більше загальної кількості елементів.

Використання алгоритму k-means для кластеризації супроводжується рядом проблем, а саме:

– необхідність заздалегідь визначити кількість результуючих кластерів, що не завжди може бути зручним, адже загалом необхідно експериментальним шляхом підбирати прийнятну, для даного набору даних, кількість кластерів, і для кожного оптимальна кількість буде своя. Існують додаткові методи, які дозволяють обчислювати цей параметр перед, безпосередньо, початком КА;

– алгоритм дуже чутливий до вибору початкових центрів кластерів, оскільки на першому кроці центри кластерів обираються випадковим чином – це призводить до зростання ймовірності похибки і можливості результатів відмінних між собою при повторному (багаторазовому) запуску алгоритму.

Також складнощі викликають ситуації приналежності об’єкта однаково до декількох кластерів або до жодного зі сформованих.

Попри недоліки k-means можна назвати досить простим алгоритмом, добре придатним для розуміння загальні процесів кластеризації і хорошою основою для побудови, на його базі, розширених версій, що покликані вирішувати більш вузько направлені задачі.

Одною з таких модифікацій цього методу можна вважати fuzzy k-means або c-means [9)], який відрізняється тим, що кожен елемент кластеру належить до нього з певною ймовірністю, і його не обов’язково можна чітко віднести до одного з кластерів, тобто групи можуть перетинатися. Що загалом вирішує проблему приналежності об’єкта однаково до декількох кластерів або до жодного зі сформованих, оскільки визначає саме ступінь цієї приналежності.

Неієрархічні методи більш стійкі до шумів, неправильної метрики та наявності незначимих параметрів у порівнянні з ієрархічними методами, які виграють, у випадках з невизначеною кількістю кластерів, ітерацій, або умови зупинки. Також ієрархічні методи дозволяють більш детально вивчити структуру даних.

1.5 Особливості кластеризації часових рядів

Часовий ряд – це орієнтована у часі або хронологічна послідовність за предметною область, що становить інтерес. Це спосіб представлення статистичних даних, з якими найчастіше може зіткнутися аналітик. Данні часових рядів використовуються в найрізноманітніших сферах людської діяльності, таких як фондові ринки, дані датчика, контроль стану машини, екологічні застосування або медичні дані.

Наприклад [9)], на рисунку 1.7 показана ринкова прибутковість казначейських цінних паперів США з 10-річним постійним терміном погашення з квітня 1953 року по грудень 2006 року. Цей графік називається графіком часового ряду.

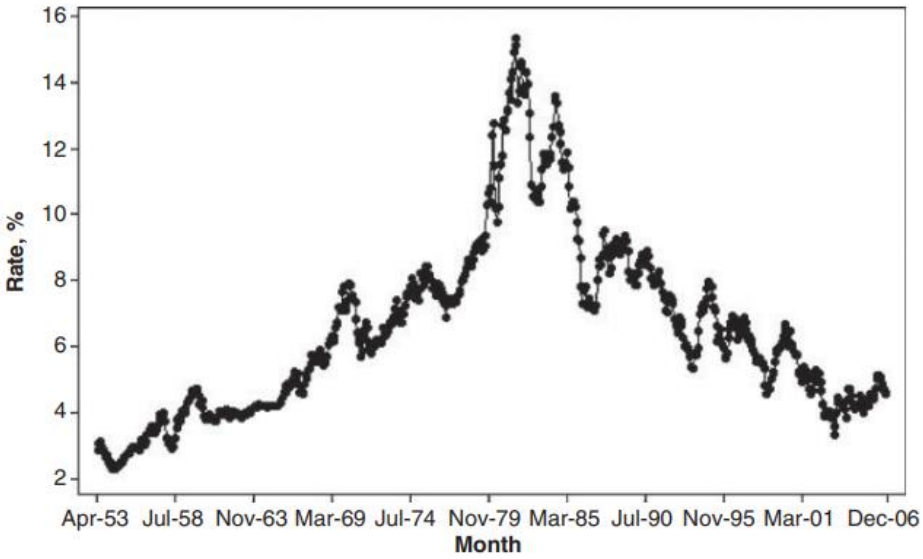


Рисунок 1.6 – Приклад графіку часового ряду

Працюючи з часовими рядами можна зіткнутися з типовими складнощами data science, такими, як велика розмірність вхідних даних, наявність шуму та пропущені дані. Розглядаючи, безпосередньо, кластеризацію часових рядів слід звернути увагу на додаткові незручності:

– ряди можуть містити різну кількість відліків;

– більше ступенів свободи для визначення схожості одного об’єкта на інший;

– при виборі метрик та статистик слід звертати уваг на локальну залежність даних.

Важливою задачею, що вирішується при роботі з часовим рядом є визначення близькості, що буде використовуватись при кластеризації:

– близькість за часом. Полягає у тому, що необхідно знайти особливі точки і інтервали, що відповідають одне одному у часі, повна відповідність не вимагається, головне загальна схожість;

– близькість за формою. Полягає у знаходженні однакових характерні особливості, які можуть бути рознесені за часом або розтягнуті та таке інше;

– близькість за структурою. Полягає у знаходженні послідовностей з однаковим законом змінювання.

Часові ряді розглядаються і аналізуються з метою:

– визначення природи ряду;

– прогнозування майбутніх значень ряду.

Як вже зазначалося дані можуть бути в неідеальній, для аналізу, формі. Часові ряди також володіють недоліками, містять аномальні значення, що вимагає проведення попередньої обробки і згладжуванні ряду. Якщо цього не зробити аномальні дані можуть призвести до спотворення результатів, що будуть отримані.

Існує два види аномальних явищ:

– явища першого роду, що викликані технічними збоями і можуть бути виявлені і усунуті;

– до явиш другого роду відносься ті, що мають епізодичний характер і не можуть бути усунуті.

Для того, щоб виявити такі значення використовується критерій Ірвіна, за ним аномальною точкою вважається Y_t, що відстає від попередньої точки Y_{t-1} на величину, більшу середньоквадратичного відхилення:

$$\lambda_i = \frac{|Y_t - Y_{t-1}|}{\sigma}. \tag{1.30}$$

де λ_i - критерій Ірвіна, а σ - середньоквадратичне відхилені:

$$\sigma = \sqrt{\frac{\sum_{t=1}^{n}(Y_t - \overline{Y})^2}{n-1}}. \tag{1.31}$$

Розрахувавши λ_i робиться висновок про аномальність точки, яка вважається такою у разі виконання умови: $\lambda_i > \lambda_{таб}$, де $\lambda_{таб}$ - це табличне значення критерію Ірвіна, частину якої наведено у таблиці 1.2.

Таблиця 1.1 – Табличні значення критерію Ірвіна

λ_1	λ_3	λ_5	λ_7	λ_9
	1,443582	1,443582	1,443582	1,443582
λ_{12}	λ_{14}	λ_{16}	λ_{18}	λ_{20}
1,443582	1,443582	1,443582	1,443582	1,443582

Також існують випадкові коливання часового ряду, впоратися з ними може допомогти згладжування [11)], яке має декілька методів:

Проста (середньоарифметична) ковзаюча середня:

$$Y_t = \frac{\sum_{i=t-p}^{t+p} Y_i}{2p+1}; p < t < n - p. \quad (1.32)$$

Зважена (середньозважена) ковзаюча середня:

$$Y_t = \frac{\sum_{i=t-p}^{t+p} \rho_i Y_i}{\sum_{i=t-p}^{t+p} \rho_i}; p < t < n - p. \quad (1.33)$$

Середньохронологічна:

$$Y_t = \frac{\frac{Y_{t-T/2}}{2} \sum_{i=t-T/2+1}^{t+T/2-1} Y_i + \frac{Y_{t+T/2}}{2}}{T}; \frac{T}{2} < t < n - \frac{T}{2}. \quad (1.34)$$

Зазвичай T має значення 4 – квартали, або 12 – місяців, що дорівнює року.

Експоненційне згладжування:

$$Y_t = \frac{\sum_{i=1}^{t} \rho Y_i}{\sum_{i=1}^{t} \rho_i}. \quad (1.35)$$

Дане згладжування для даної точки враховує значення всіх попередніх точок.

1.6 Метод нечіткої кластеризації даних k-середніх (FCM)

Традиційні принципи кластерного аналізу припускають, що виділяються групи, де кожен об'єкт може належати тільки до однієї групи. Обмеженість такого підходу часто призводить до аналітичної невизначеності, і тому розумною альтернативою є інтерпретація компонентів систем як нечітких об'єктів. У роботі [12)] розглядається застосування нечіткої кластеризації для діагностики хвороби Альцгеймера, і робиться наголос, що чітка логіка дозволяє лише визначити, хворий пацієнт або ні, тоді як використання нечіткої логіки потенційно дає змогу оцінити «на скільки» хворий.

Цей алгоритм є узагальненням алгоритму k-means, в разі, якщо кластери є нечіткими множинами, і, елемент може належати різним кластерам з різним ступенем ймовірності. Дозволяє розбити наявну множину елементів потужністю N на задане число НМ k.

Як все зазначалося основною відмінністю даного методу від попереднього є можливість приналежності елемента даних до декількох кластерів із певною вірогідністю, що визначається за допомогою функції належності. Функція належності – це поняття із задач, пов’язаних з нечіткими множинами, вона дозволяє визначити ступінь з яким елемент належить тій чи іншій множині (у даному випадку кластеру). Існує безліч методів для визначення цієї функції, такі як експертна оцінка, параметричній ідентифікації нечітких моделей, також продовжують пропонуватися і нові методи, наприклад [13)] детальніше визначає ці поняття. Функція належності може приймати значення на інтервалі [0, 1] у порівнянні з класичною логікою, яка однозначно визначає 0 або 1. Використання такої логіки значно розширяє можливості застосування її у вирішенні різноманітних задач (наприклад, кластеризації).

Алгоритм fuzzy k-means [14)]:

Для того щоб вирішити певні проблеми під час кластеризації, такі як кількість кластерів і початкові центри кластерів, було введено термін для цільової функції FKM. Кластеризація X на c кластерів за цим алгоритмом повинна мінімізувати таку цільову функцію:

$$f\left[U,V\right]=\sum_{i=1}^{n}\sum_{k=1}^{c}u_{ik}d_{ik}+\gamma\sum_{i=1}^{n}\sum_{k=1}^{c}u_{ik}\log u_{ik};\tag{1.36}$$

за умови

$$\sum_{k=1}^{c}u_{ik}=1,u_{ik}\in\left(0,1\right],1\le i\le n,1\le k\le c.\tag{1.37}$$

де U – матриця розбиття n на c;

V – матриця розбиття c на m, що містить центри кластерів;

d_{ik} – міра відмінності між k-м центром кластера та i-м об'єктом.

Процедура мінімізації, чергуючи член матриці U і центр кластера матриці V, може бути застосована, наступним чином:

$$u_{ik}=\frac{\exp\left(\frac{-d_{ik}}{\gamma}\right)}{\sum_{s=1}^{c}\exp\left(\frac{-d_{is}}{\gamma}\right)};\tag{1.38}$$

$$v_{k}=\frac{\sum_{i=1}^{n}u_{ik}x_{i}}{\sum_{i=1}^{n}u_{ik}}.\tag{1.39}$$

Задачею FCM кластеризації є пошук матриці належності розмірності $c\times l$:

$$M=\begin{bmatrix}m_{11} & m_{12} & \ldots & m_{1l}\\ m_{21} & m_{22} & \ldots & m_{2l}\\ & & \ldots & \\ m_{c1} & m_{c2} & \ldots & m_{cl}\end{bmatrix}$$

де m_{ij} – ступінь приналежності j-го елемента і-му кластеру.

При цьому матриця М має задовольняти умовам:

– $m_{ij} \in [0,1], i = \overline{1,c};\ j = \overline{1,l}$;

– кожен об’єкт має бути розподіллений між усіма кластерами $\sum_{i=1}^{c} m_{ij} = 1,\ j = \overline{1,l}$;

– жоден з кластерів не має містити всі елементи, а також лишатися пустим $0 < \sum_{j=1}^{l} m_{ij} < l, i = \overline{1,c}$;

Для оцінки якості розбиття використовується критерій розкиду:

$$J = \sum_{i=1}^{c}\sum_{j=1}^{l} \left(m_{ij}\right)^{w} d\left(v_i, x_j\right). \tag{1.40}$$

де $d\left(v_i, x_j\right)$ – Евклідова відстань між *j*-м об’єктом $x_j = \left(x_{j1,} x_{j2}, \ldots, x_{jn}\right)$ та і-м центром кластера $v_i = \left(v_{i1,} v_{i2}, \ldots, v_{in}\right)$, що наведена у формулі (1.10).

$w \in (1, \infty)$ – експонентна вага, що визначає нечіткість, розмитість кластерів.

Також визначається матриця координат центрів кластерів розмірністю $c \times n$:

$$V = \begin{bmatrix} v_{11} & v_{12} & \ldots & v_{1n} \\ v_{21} & v_{22} & \ldots & v_{2n} \\ & & \ldots & \\ v_{c1} & v_{c2} & \ldots & v_{cn} \end{bmatrix}.$$

Елементи цієї матриці визначаються за формулою:

$$v_{ik} = \frac{\sum_{j=1}^{l} \left(m_{ij}\right)^{w} x_{jk}}{\sum_{j=1}^{l} \left(m_{ij}\right)^{w}}, k = \overline{1,n}(v). \tag{1.41}$$

Застосування та актуальність нечіткої кластеризації наводиться в статті [15)]. Головні переваги методу — його простота та швидкість виконання. Даний алгоритм в

адаптованому вигляді, для багатомірних даних, наводиться в [16)], також наводяться експерименти і порівняльний аналіз з класичним методом FCM.

1.7 Постановка задачі дослідження

Зважаючи на сучасні потреби все більш актуальними і затребуваними є розробки в галузі аналізу накопичених даних, адже ми живемо в той час коли кількість інформації, якою ти володієш, вже менш значима ніж її якість і можливість обробити, зробити висновки на цьому підґрунті.

Отже кластеризація становить інтерес для попередньої обробки даних, з метою подальшого прогнозування та більш зручного їх подальшого аналізу. Отримавши необхідні групи і відповідні їм центроїди можна робити подальші висновки та прогнози вже з конкретними представниками, а не з усім набором даних, що особливо актуально в умовах безкінечно зростаючого об’єму інформації. Такий підхід дозволяє: біль усвідомлено обробляти дані, шляхом використання для кожного кластеру найбільш підходящого алгоритму аналізу; виявлення новизни, шляхом виділення об’єктів, що не потрапили до жодного з кластерів; провести стиснення, виділивши найбільш типових представників, за умов збитковості даних.

Предмет дослідження – це методи підготовки та обробки вхідних даних, що містять пропущені значення, для їх подальшого прогнозування, аналізу та використання в задачах кластеризації; розгляд адаптованого класичного методу кластеризації для вирішення проблеми неповних даних.

Метою роботи можна назвати розгляд існуючих методів позбавлення від пропусків в даних в задачах кластеризації та доцільність їх використання і реальних задачах. Аналіз та оцінка адаптованих методів кластеризації для вирішення подібного роду задач, та висновок про їх переваги перед методами попереднього позбавлення від пропусків пере процесом кластеризації.

Завданням даної роботи є аналіз переваг та недоліків кожного з методів, що направлені на відновлення даних, для визначення доцільності використання їх в задачах кластеризації та виділення найбільш придатного до застосування, порівняння їх між собою, оцінка результативності за наслідками порівняння кластеризації відновлених даних з результатами кластеризації еталонних. Особлива увага приділяється методам нечіткої кластеризації FCM, метод видалення всіх рядків, що містять пропуски,

заповненням пропусків вибірковими статистиками, заповнення пропусків з урахуванням структури зв'язків. Наукова новизна – дослідження проблеми кластеризації даних, що містять пропущені значення та розгляд методів, які дозволяють розв'язати цю задачу. Проведення експериментів та порівняння результатів кожного з методів, висновки про доцільність використання того чи іншого методу та побічні ефекти.

Практична значущість роботи полягає у визначені можливості використання в реальних задачах, що зазвичай не є ідеальними і з великою ймовірністю міститимуть пусті значення, методів обробки даних для використання їх в задачах кластеризації.

Говорячи про актуальність і затребуваність даної теми слід дослідити способи застосування даних методів та розробки в даній сфері, зацікавленість серед дослідників. Можна сказати, що класичні методи кластеризації є досить широкою темою для досліджень і модифікації для вирішення різноманітних задач та різних типів даних. В роботі [12] розглядається підходи до кластеризації відео на основі аналізу багатовимірних часових рядів. Розглядаючи цю роботу можна зрозуміти, що КА може бути використано навіть в таких складних задачах, як обробка відео. Оскільки відеоданні є неструктурованими – це ускладнює загалом відеоаналіз, робить його складною науковою галуззю, що саме по собі висвітлює розробки більш цінними і значущими. Також робиться наголос на тому, що кластеризація у даному випадку та її продуктивність знаходиться в прямій залежності від обраної функції.

Проблеми обробки та кластеризації відео так само розглядаються в роботі [13, 14]. В ній піднімаються проблеми створення єдиної структури, що здатна відображати і передавати історії високого рівня. Не зважаючи на те, що сучасні методи здатні впоратись з цими нюансами, залишаються питання потреби досить високошвидкісних процедур.

Так, в роботі [15] наводиться гібридний алгоритм нечіткої кластеризації, він дозволяє обробляти біомедичні дані, що є досить важливим завданням, складні задачі обробки відеопотоків, обробка інформації з датчиків IoT, що стрімко набуває популярності та таке інше. Важливими перевагами є обчислювальна простота і можливість використовувати для різних об'ємів даних, як великих так і маленьких. Крім FCM кластеризації [16] також використовуються on-line нейро-фаззі систему для вирішення задач послідовного нечіткого кластерування даних.

Говорячи про обробку саме часових рядів, можна зробити висновок, що це питання, наразі, є досить актуальним, пошук пропусків та власне кластеризація володіють своїми складнощами. Розробляються методи адаптивної кластеризації [17] для вирішення задачі обробки коротких часових рядів з нерівномірно розподіленими спостереженнями.

Піднімається питання неможливості застосування евклідової метрики, або використання стохастичних критеріїв для визначення відстані між вибірками, наприклад для реалізації, що наведена на рисунку 1.7.

Що доводить досить широкі можливості методів КА які потребують модифікацій задля більш ефективного використання у різних випадках для різних даних.

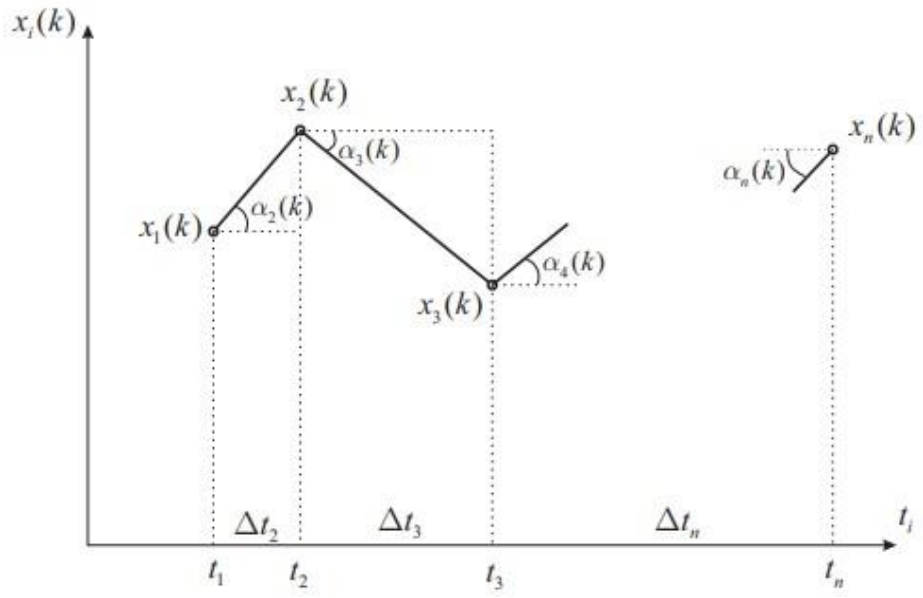


Рисунок 1.7 – Приклад часових рядів з неоднорідним тактом квантування.

Розбиття множини на групи подібних об’єктів – це дуже потужний механізм підготовки даних до подальшого аналізу, але існує проблема обробки даних, що не є повними, тобто мають пропуски. Відсутні дані створюють багато труднощів, оскільки більшість процедур аналізу даних не були призначені для них [18].

Зважаючи на те, що в реальних даних велика ймовірність отримати таку ситуацію, коли інформація, що аналізується не повна, можна сказати, що відмовитися від кластеризації за цією причиною – не вихід, тому така задача потребує вирішення і є досить актуальною.

Отже існують декілька можливих рішень проблеми обробки даних з пропусками, які наведено на рисунку 1.2. Їх умовно можна поділити на такі, що націлені на попередню підготовку даних і ті, що вимагають адаптації (модифікації) стандартних алгоритмів кластеризації для обробки безпосередньо даних з пропусками. Кожен з цих варіантів має свої недоліки та переваги, і не є універсальними для рішення цієї проблеми.

Однак не зважаючи на наявність уже існуючих методів заповнення пропущений значень, ця тема також продовжує свій розвиток у модифікаціях методів задля отримання найкращих результатів у якомога більшій кількості випадків. Так у роботі [18]

розглядається розробка алгоритму заповнення пропущених даних на основі Адаліни. В даній роботі використовується непромережені системи, з паралельно працюючими адаптивними лінійними асоціаторами. Також наводитеся схема (рисунок 1.8), на якій наочно можна побачити, як відбувається позбавлення від пропусків за допомогою однієї з реалізацій.

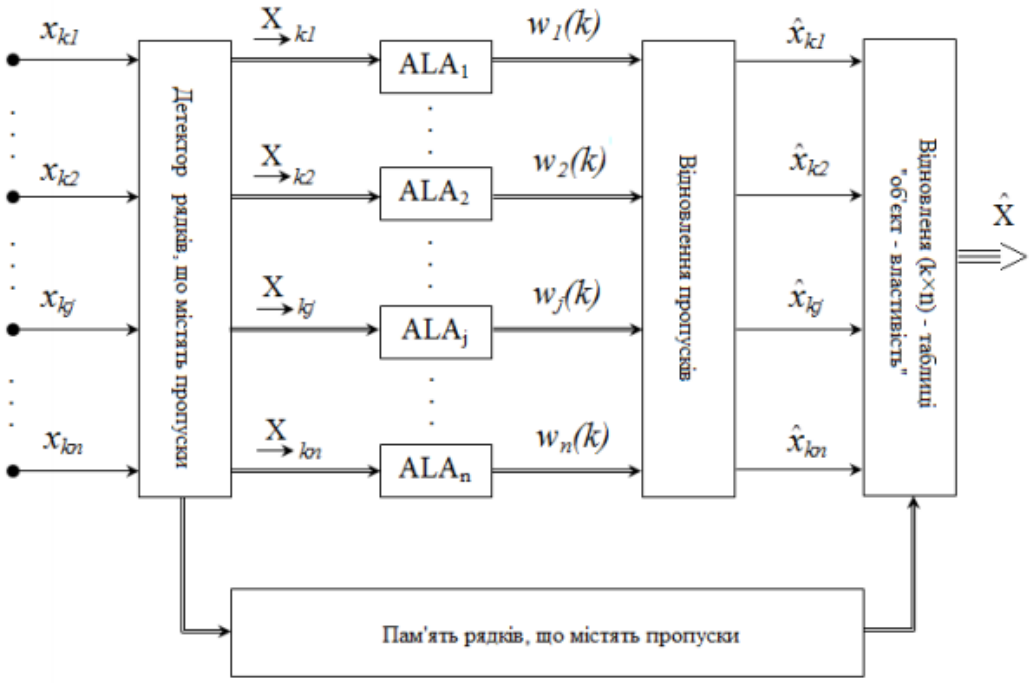


Рисунок 1.8 – Схема, що відображує один з варіантів відновлення пропусків в таблицях даних

2 МАТЕМАТИЧНІ МОДЕЛІ КЛАСТЕРИЗАЦІЇ ТА ПРОГНОЗУВАННЯ ЧАСОВИХ РЯДІВ

Для роботи з часовими рядами, а саме кластеризації, існує ряд можливих алгоритмів:

– k-середніх і його варіанти;

– ієрархічні алгоритми;

– засновані на щільності методи кластеризації, такі як DBSCAN.

Також важливим етапом обробки часових рядів є визначення багатомірний він чи одномірний.

2.1 Алгоритм fuzzy k-means (c-means)

Нечітка кластеризація складається з таких кроків [10, 11]:

Крок 1: Спочатку встановлюються параметри алгоритму, які включають c-кількість кластерів; m – експоненціальна вага; ε – параметр завершення алгоритму, для яких мають виконуватися умови: $c \in [2,\ N),\ (m\ >\ 0)$ та $(\varepsilon\ >\ 0)$.

Можна зазначити, що m, строго більше 1, що вказує на показник ступеня членства, який використовується в критерії відповідності.

Слід зазначити, що $m \to 1$ дає більш чітку кластеризацію, тоді як $m \to Inf$ призводить до повної нечіткості. Значення, занадто близькі до 1, можуть привести до повільної конвергенції. Далі відзначимо, що навіть значення за замовчуванням m = 2 може привести до повної нечіткості, тобто членства $u(i,\ v)\ ==\ 1\ /\ k$. В цьому випадку рекомендується вибрати менше значення m.

Крок 2: Провести, випадковим чином, генерацію матриці нечіткого розбиття $\mu_{ij}^{(0)}\ \sim U(0,\ 1)$ та $let\ t\ =\ 1$.

Крок 3: Розрахувати центри кластерів (v_j), за допомогою формули:

$$v_j = \frac{\sum_{i=1}^{N} \mu_{ij}^m x_i}{\sum_{i=1}^{N} \mu_{ij}^m}, (j = \overline{1,c}). \quad (2.1)$$

Крок 4: Отримати відстані від отриманих центрів (v_j) до об'єктів X.

$$D_{ij} = \sqrt{\|x_i - v_i\|^2}; i = \overline{1,N}; j = \overline{1,c}. \quad (2.2)$$

Крок 5: Оновити елементи матриці нечіткого розбиття μ_{ij} для всіх i, j, якщо $D_{ij} > 0$.

$$\mu ij = \left(\sum_{k=1}^{c} \left(\frac{\|x_i - v_j\|}{\|x_i - v_k\|} \right)^{\frac{2}{m-1}} \right)^{-1}, \left(i = 1, 2, \ldots, N; j = 1, 2, \ldots, C\right). \quad (2.3)$$

Якщо $D_{ij} = 0$, тоді $\mu ij = \begin{cases} 1, якщо\, j = i \\ 0, якщо\, j \neq i \end{cases}$.

Крок 6: Перевірити умову $\left\| \mu^{(t)} - \mu^{(t-1)} \right\|$.

Якщо $\left\| \mu^{(t)} - \mu^{(t-1)} \right\| < \varepsilon$, тобто різниця між матрицею нечіткого розбиття поточної і попередньої ітерації менше ніж заданий на початку розмір похибки ε. Якщо умова виконується, то відбувається зупинка алгоритму і на даному етапі будуть отримані кінцеві результати кластеризації. Якщо ж умова не виконується, то алгоритм знов починає розрахунки, переходячи на крок 3.

Проаналізувавши алгоритм можна сказати, що на першому ж кроці необхідно визначити кількість результуючих кластерів. Це важливий етап, і як було зазначено раніше, саме він може викликати найбільше труднощів і неправильний вибір здатен негативно вплинути на результат.

Цю задачу можна вирішити емпіричним шляхом, тобто експериментально для заданого набору даних підібрати кількість кластерів, що найбільш підходить у даній конкретній ситуації. Наприклад, у разі нечіткої кластеризації, можна зробити висновок про необхідність зменшити кількість кластерів, якщо чітко видно, що одна зі сформованих

груп повністю, або значною частиною, складається з елементів іншої/інших груп. Але існують методи, які дозволяють визначити цей параметр автоматично.

2.2 Алгоритм динамічної трансформації часової шкали (DTW)

Алгоритм DTW визначає оптимальну послідовність трансформації часу між двома часовими рядами та дозволяє знайти оптимальну відстань між ними.

Алгоритм динамічної трансформації часового ряду запропоновано як альтернативу евклідової метрики для обробки часових рядів. Адже говорячи про класичний спосіб розрахунку (евклідову метрику) слід зазначити, що вона володіє значним недоліком при рішенні таких задач. А саме, якщо один з двох рядів, що мають бути визначені, як однакові, буде зміщено у часі, тоді за евклідовою метрикою ряди можуть виявитися різними. Отже, для боротьби з цим недоліком було запроваджено метод DTW.

Різниця між Евклідовою метрикою і метрикою в методі DTW наведено на рисунку 2.1.

Модифікацією алгоритму DTW є алгоритм soft-DTW, при якому DTW відстань є:

$$DTW_\gamma\left(X_i, X_j\right) = -\gamma \log \sum_{k=1}^{K} e^{\left(\frac{d(w_k)}{K\cdot\gamma}\right)}. \qquad (2.4)$$

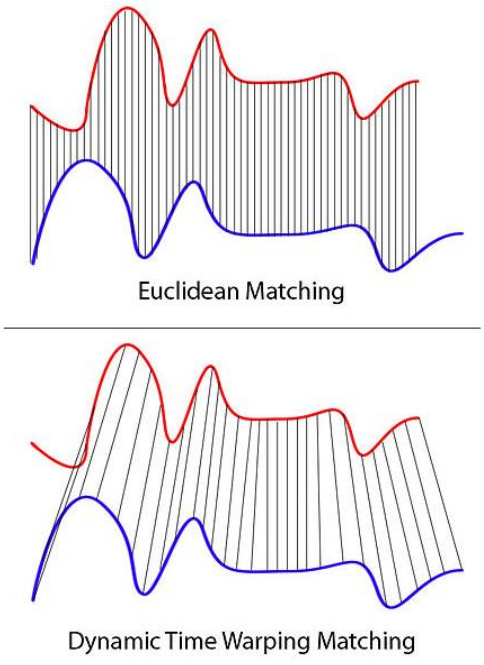


Рисунок 0.1 – Порівняння Евклідової метрики і DTW

Алгоритм складається з таких кроків [22)]:

Крок 1: будується матриця відстаней $D = \left\{ d_{i,j} \right\}$.

Таблиця 2.1 – Приклад матриці відстаней у методі DTW

	y_1	y_2	y_3	y_4	y_5	y_6	y_7
x_1	Δ_{11}	Δ_{12}	Δ_{13}	Δ_{14}	Δ_{15}	Δ_{16}	Δ_{17}
x_2	Δ_{21}	Δ_{22}	Δ_{23}	Δ_{24}	Δ_{25}	Δ_{26}	Δ_{27}
x_3	Δ_{31}	Δ_{32}	Δ_{33}	Δ_{34}	Δ_{35}	Δ_{36}	Δ_{37}
x_4	Δ_{41}	Δ_{42}	Δ_{43}	Δ_{44}	Δ_{45}	Δ_{46}	Δ_{47}
x_5	Δ_{51}	Δ_{52}	Δ_{53}	Δ_{54}	Δ_{55}	Δ_{56}	Δ_{57}

Зазвичай використовується евклідова відстань (1.10), або різниця за координатами, для одного виміру це становить:

$$d_{i,j}\left(x_i, x_j\right) = \left| x_i - x_j \right|. \tag{2.5}$$

Кожен елемент матриці відповідає відстані між точками x_i, x_j.

Крок 2: будуємо матрицю трансформації $D_{DTW} = \left\{ r_{i,j} \right\}$, елементи якої розраховуються за формулою (2.6):

$$r_{i,j} = d_{i,j} + \min\left(D_{i-1,j}, D_{i-1,j-1}, D_{i,j-1} \right). \tag{2.6}$$

Крок 3: побудова оптимального шляху трансформації і DTW відстань.

Шлях трансформації W встановлює відповідність між X_i і X_j, та визначається за формулою:

$$W = \left\{ w_k \right\}_{k=1}^{K}, \quad w_k = \left(i, j\right)_k, \quad \max\left(n_i, n_j\right) \le K < \left(n_i + n_j\right). \tag{2.7}$$

де K – це довжина шляху.

Також, шлях трансформації має задовольняти умовам:

– має містити всі точки обох часових рядів

$$w_1 = (1,1), \quad w_K = (n_i, n_j). \tag{2.8}$$

– безперервність. Шлях трансформації пересувається на один крок за один раз

$$\forall w_k = (w_i, w_j), w_{k+1} = (w_{i+1}, w_{j+1}) : w_i - w_{i+1} \le 1, w_j - w_{j+1} \le 1. \tag{2.9}$$

– монотонність. Обмеження гарантує, що повернення до попередньої точки не відбудеться, тобто індекси не зменшуватимуться.

DTW відстань (вартість шляху) між двома послідовностями розраховується за допомогою формули:

$$DTW(X_i, X_j) = \min\left(\frac{\sum_{k=1}^{K} d(w_k)}{K}\right). \tag{2.10}$$

Загалом методи DTW показали значну ефективність при кластеризації часових рядів. Але, попри це, при великій кількості даних він може призвести до збільшення навантаження. Тому в [23)] розглядається можливість розробити ітеративну процедуру DTW, що здатна скоротити часові послідовності.

2.3 Алгоритм DBSCAN

DBSCAN – це алгоритм, що засновано на щільності. Кластеризація за ним відбувається таким чином, що до однієї групи потрапляють точки ,які розташовані близько одна до одної. Точки ж, які стоять досить далеко від найближчого сусіда вважаються викидами. Цей алгоритм досить часто використовується для кластеризаії даних, в якості альтернативи алгоритму k-means.

При роботі з алгоритмом DBSCAN [24)] використовуються такі поняття:

Околиця:

$$E(x_j) = \{x_i \in X_Q | d(x_i, x_j) \le \sigma\}. \tag{2.11}$$

Кореневим об’єктом, називається об’єкт околиця якого містить не менше m (ступінь) об’єктів.

$$x_j \in X_Q \left| \text{card}E\left(x_j\right) \geq m. \right. \tag{2.12}$$

Об'єкт x_i є безпосередньо щільно-досяжний з об'єкта x_j, якщо x_j - це кореневий об’єкт, а x_i знаходиться на відстані не більшій ніж σ від точки x_j.

Об'єкт x_i є щільно-досяжний з об'єкта x_j, якщо існують

$$\left\{ x_p \in X_Q \right| \quad p = \overline{1,k}, \quad x_1 \equiv x_j \quad x_2 \equiv x_i \Big\}; \tag{2.13}$$

$$\forall p = \overline{1,k-1}: \quad \left\{ x_{p+1} \in E\left(x_p\right) \left| \text{card}E\left(x_p\right) \geq m \right. \right\}. \tag{2.14}$$

Шумом називається множина точок, які не належать до жодного кластеру.

Розглядаючи алгоритм слід зазначити, що точка можу бути в одному з трьох станів:

– не переглянута;

– помічена як та, що не є точкою жодного з кластерів;

– та, що належить деякому кластеру.

Алгоритм DBSCAN складається за таких етапів:

Крок 1: Для кожної околиці σ кожної точки $x_i \in X_Q$ знайти точки та кореневі об’єкти x_j, які мають більше ніж m сусідів.

Крок 2: Для ядрових точок (кореневих об’єктів) сусідніх графів визначити компоненти зв'язності.

Крок 3: Для кожної не ядрової точки знайти кластер, який знаходиться в околиці σ. Якщо такий кластер знайдено помістити точку в нього, якщо ж ні тоді задана точка є шумом.

Алгоритм завершується тоді, коли сі точки переглянуті, інакше обирається наступна точка, що ще не була переглянута.

На рисунку 2.21 [15] наведено приклад, за яким видно формування кластерів при застосуванні методів, що використовують щільність

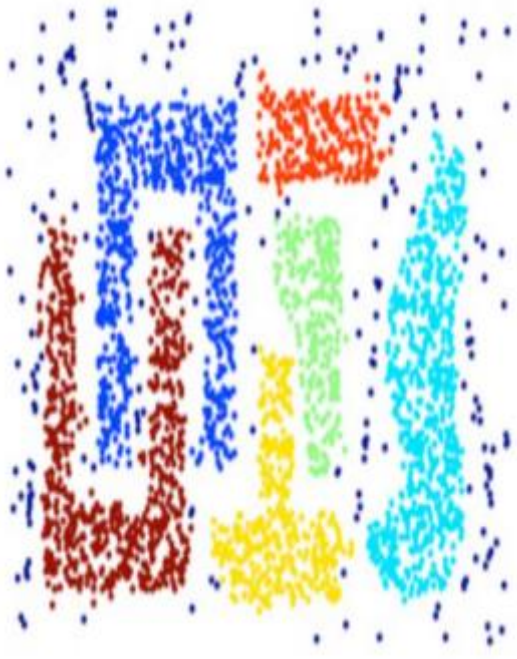


Рисунок 2.2 - Приклад кластеризації на основі щільності

До переваг алгоритму слід віднести те, що немає необхідності на початку визначати кількість кластерів, на відміну від k-means, де це значення є параметром функції. Це досить суттєво, як все було сказано раніше, необхідність на початку визначення кількості кластерів ускладнює процес кластеризації, робить його менш зручним, адже природнім було б те, що б кількість кластерів визначалася у залежності від насправді існуючих груп за їх відмінностями.

Однак, DBSCAN вимагає такі параметри, як максимальну відстань $\sigma = d_{max}\left(x_i, x_j\right)$ між сусідніми точками (околиця об’єкта), і мінімальне число точок в околиці - m.

2.4 Робота з пропусками з метою подальшого прогнозування

Відновлення пустих значень вирішує задачу заповнення пропусків в таблицях і часових рядах такими значеннями, що найбільш точно підходять і мають мінімальну похибку.

В роботі [**Ошибка! Источник ссылки не найден.**] розглядається важливість і складність нечіткої кластеризації в реальних умовах, які передбачають вміст пропущених даних, коли похідна таблиця має пусті клітинки, задача може ускладнюватися ще і тим,

що кількість цих значень невідома і хаотична. Таким чином, вирішення цієї проблеми є актуальним і продовжує розглядатися, попри наявність напрацювань у даній сфері.

Нечітка кластеризація досить часто використовується для вирішення проблем пропусків, шляхом адаптування класичного алгоритму, але існують і інші методи, що дозволяють розв’язати цю задачу. Так, наприклад, в роботі [**Ошибка! Источник ссылки не найден.**] розглядається нейронна мережа, яка здатна впоратися с відсутніми значеннями в режимі реального часу з високою швидкістю, при цьому маючи просту реалізацію.

Підходи на основі математичного апарату обчислювального інтелекту, а, насамперед, нейромережах [26)], для вирішення такої задачі, вважаються досить ефективними.

Перспективною є сфера вивчення методів кластеризації з метою прогнозування. Алгоритм полягає в тому, що необхідно на перших кроках відновити, одним з можливих методів, пропущені дані, після чого зробити прогноз (передбачення часових рядів, на основі наявної ділянки, з відновленими заздалегідь даними). Далі провести кластеризацію розширеної таким чином області даних.

Вирішення такої задачі може застосовуватися і убути необхідним у багатьох сферах, таких як наука, економіка, соціальна сфера, виробництво та таке інше. З розвитком ІТ, а також збільшенням кількості даних, що зберігаються, зростає популярність і розвиток методів прогнозування. Вони потребують досить детального вивчення похідних наборів даних, а також, безпосередньо методів, що будуть використовуватися для аналізу.

3 ЕКСПЕРИМЕНТАЛЬНІ ДОСЛІДЖЕННЯ МЕТОДІВ ПРОГНОЗУВАННЯ І КЛАСТЕРИЗАЦІЇ ЧАСОВИХ РЯДІВ

3.1 Розгляд предметної області

В роботі розглядатиметься таблиця, що містить медичну інформацію. Ця предметна область була обрана на сам перед тому, що на сьогоднішній день, обробка саме медичної інформації є доволі актуальною та важливою задачею. В контексті кластеризації даних з пропусками така предметна область є досить доречною, адже в ній є велика ймовірність пустих, незаповнених полів, відсутність певної кількості важливої інформації. Та попри це обробка такого роду даних дуже важлива і необхідна задача, а її достовірність і точність – це беззаперечна необхідність.

Отже, можна зазначити, що ймовірність втрати певної частини інформації в таких ситуаціях не виключена, що зумовлено значним поспіхом, заповненням анкети самим пацієнтом, який не надав значущості якимось питанням та багато інших причин. При цьому, попри дані фактори, необхідно проаналізувати отриману інформації якомога точніше і отримати результат максимально приближений до еталонного, того, що можна отримати при повних даних.

Які цілі можуть переслідуватися в кластеризації медичної інформації? Насамперед це:

– групування пацієнтів за захворюваннями, пошук спільного та відмінного в межах групи, що дозволить покращити якість лікування;

– оцінка критеріїв, що впливають на ту чи іншу групу, здатна покращити розуміння причин захворювань;

– можливість оцінити шкідливі звички, кому вони притаманні, як впливають на здоров’я;

– регіональні відмінності, або відмінності відносно статі, соціальних груп та іншого.

Кластеризація використовується в безлічі різних областей діяльності людини. У медицині кластеризация симптомів і захворювань помітно полегшує аналіз і вид лікування. Вірний розгляд кластерів симптомів шизофренії, параної і інших захворювань є визначальним при веденні успішної терапії [27)].

В роботі [20] розглядається аналіз і обробка мікроскопічних зображень крові, що є складною але важливою задачею. Було використано метод сегментації кольорів.

Всі ці досліди та висновки можуть допомогти визначити найбільш ефективні методи лікування та допоможуть лікарю в його подальшій роботі. Також, в майбутньому, такий аналіз може допомогти у розвитку штучного інтелекту і його застосуванні в медицині, адже формалізація даних дозволяє вирішувати складні питання без навчання і отримання знань від експертів (що є непростою задачею). А організація пацієнтів в групи дає можливість, як лікарю, так і, надалі, штучному інтелекту працювати за даними швидше і ефективніше.

3.2 Практичні результати застосування простих методів обробки даних з пропусками

Перевірка методів кластеризації, а саме можливості отримання допустимих результатів кластеризації в умовах неповних даних проводитиметься на прикладі таблиці з інформацією про хвороби серця пацієнтів.

Ця таблиця містить в собі 303 рядки та 8 стовпців, перші 15 рядків наведено на рисунку 3.1.

1	age	sex	cp	trestbps	fbs	restecg	thalach	exang
2	63	1	3	145	1	0	150	0
3	37	1	2	130	0	1	187	0
4	41	0	1	130	0	0	172	0
5	56	1	1	120	0	1	178	0
6	57	0	0	120	0	1	163	1
7	57	1	0	140	0	1	148	0
8	56	0	1	140	0	0	153	0
9	44	1	1	120	0	1	173	0
10	52	1	2	172	1	1	162	0
11	57	1	2	150	0	1	174	0
12	54	1	0	140	0	1	160	0
13	48	0	2	130	0	1	139	0
14	49	1	1	130	0	1	171	0
15	64	1	3	110	0	0	144	1

Рисунок 3.1 – Приклад таблиці heart

З рисунку 3.1 можна побачити, що таблиця, яка розглядається, містить такі поля, що наведено в таблиці 3.1.

Таблиця 3.1 – Похідні дані для дослідження

№	Назва стовпчика	Опис
1	age	Вік, в роках
2	sex	Стать (1 = чоловік; 0 = жінка)
3	cp	Тип болю в грудях
4	trestbps	Артеріальний тиск у стані спокою (в мм рт.ст. при надходженні в госпіталь)
5	fbs	Рівень цукру в крові натще> 120 мг / дл (1 = істина; 0 = брехня)
6	restecg	Результати електрокардіографії в спокої
7	thalach	Досягнуто максимальний пульс
8	exang	Розвинути стенокардію, викликану фізичним навантаженням (1 = так; 0 = ні)

3.3 Результати застосування методів

Для початку проведемо кластеризацію похідної таблиці, яка не містить пропущених даних. Для кластеризації обрано метод – fuzzy k-means і мову програмування R.

Виконаємо побудову "нечітких" кластерів з використанням функції fanny() з пакету cluster:

```
res.fanny <- fanny(data, k = 3, memb.exp = 1.3,
                   metric = "euclidean", stand = TRUE, maxit = 500)
print(res.fanny$membership,3)
```

– обрано 3 кластери (k = 3);

– memb.exp обрано таким чином, що memb.exp -> 1 дає більш чітку кластеризацію, тоді як memb.exp -> Inf призводить до повної нечіткості;

– metric - рядок символів, який вказує метрику, яка буде використовуватися для розрахунку відмінностей між спостереженнями. Варіанти «euclidean», «manhattan» і «SqEuclidean». Евклідові відстані - це корінь суми квадратів різниць, манхеттенські

відстані – сума абсолютних різниць, а «SqEuclidean», квадрат евклідових відстаней – сума квадратів різниць;

– stand – параметр логічного типу, якщо true, вимірювання в x стандартизируются перед обчисленням відмінностей. Вимірювання стандартизовані для кожної змінної (стовпчик) шляхом вирахування середнього значення змінної і ділення на середнє абсолютне відхилення змінної;

– maxit – максимальна кількість ітерацій, в нашому випадку – 500;

– перші, середні та останні 10 рядків, отримані за результатами кластеризації, наведено на рисунку 3.2.

```
        [,1]   [,2]   [,3]
 [1,] 0.503 0.2476 0.2492
 [2,] 0.203 0.7080 0.0889
 [3,] 0.322 0.5473 0.1304
 [4,] 0.239 0.6528 0.1079
 [5,] 0.323 0.2860 0.3913
 [6,] 0.343 0.3160 0.3407
 [7,] 0.526 0.2766 0.1974
 [8,] 0.138 0.7956 0.0665
 [9,] 0.466 0.2874 0.2469
[10,] 0.403 0.4520 0.1449
                ...
[145,] 0.409 0.2480 0.3427
[146,] 0.449 0.1752 0.3762
[147,] 0.278 0.6023 0.1195
[148,] 0.479 0.3619 0.1593
[149,] 0.154 0.7817 0.0641
[150,] 0.205 0.6947 0.1006
[151,] 0.389 0.1626 0.4487
[152,] 0.366 0.2239 0.4102
[153,] 0.483 0.2632 0.2538
[154,] 0.556 0.2238 0.2203
                ...
[294,] 0.502 0.2211 0.2764
[295,] 0.210 0.3137 0.4762
[296,] 0.189 0.0882 0.7227
[297,] 0.280 0.1620 0.5580
[298,] 0.385 0.1915 0.4240
[299,] 0.265 0.1431 0.5919
[300,] 0.285 0.5300 0.1853
[301,] 0.439 0.2076 0.3530
[302,] 0.142 0.0932 0.7650
[303,] 0.468 0.3800 0.1517
```

Рисунок 3.2 – Результат застосування функції fanny

В результаті виводиться матриця коефіцієнтів приналежності, максимальний з яких визначає цільовий кластер.

Для оцінки міри нечіткості, отриманої класифікації, використовується коефіцієнт поділу Dunn (3.1):

$$F_k = \sum_{i=1}^{n}\sum_{r=1}^{k}\frac{\mu_{ir}^2}{k}. \quad (3.1)$$

Цей коефіцієнт приймає значення 1 в разі чіткої кластеризаії:

dunn_coeff *normalized*

0.4203087 *0.1304630*

В даному випадку Fk = 0.42, а його нормована версія, що змінюється від 0 до 1 – 0.13.

Побудуємо діаграму. На рисунку 3.3 наведено ординаційну діаграму з результатами кластеризації.

fviz_cluster(res.fanny, frame.type = "norm", frame.level = 0.7)

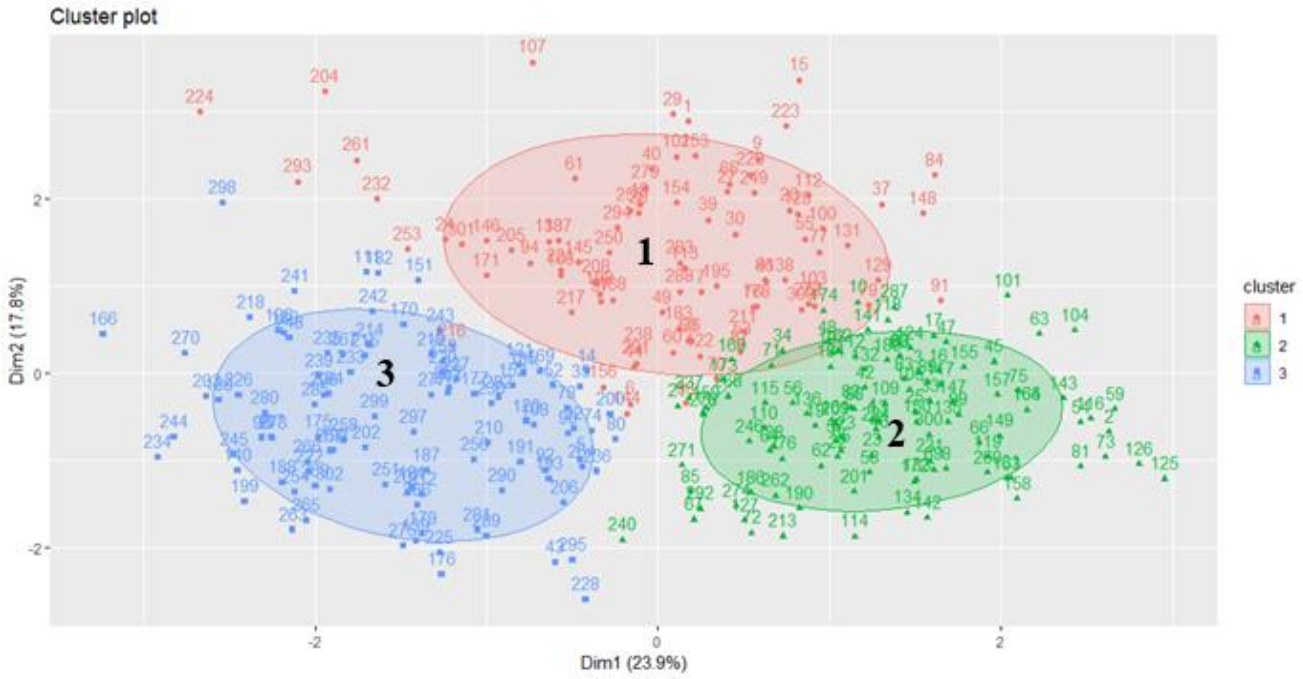


Рисунок 3.3 – Графік результатів нечіткої кластеризації

Для подальшого аналізу внесемо у похідні дані 30% пропусків випадковим чином.

Кластеризація з видаленням пропусків:

heartMissedNO <- subset(heartMissed, exang != "")

Перші, середні та останні 10 рядків, отримані за результатами кластеризації, наведено на рисунку 3.3

	[,1]	[,2]	[,3]
1	0.4942	0.2168	0.2890
2	0.1298	0.7809	0.0893
3	0.2238	0.6263	0.1499
7	0.4347	0.2923	0.2730
10	0.3306	0.4800	0.1894
14	0.3695	0.2501	0.3804
15	0.4660	0.2665	0.2675
16	0.1732	0.7063	0.1205
17	0.2529	0.5813	0.1659
18	0.4144	0.2278	0.3578
...			
160	0.3868	0.3782	0.2349
162	0.2761	0.5348	0.1892
164	0.1312	0.7797	0.0891
171	0.4473	0.1848	0.3678
173	0.3781	0.3545	0.2674
177	0.3981	0.1947	0.4072
178	0.4136	0.3201	0.2663
183	0.4156	0.2709	0.3135
184	0.3178	0.4751	0.2072
185	0.3908	0.1834	0.4258
...			
283	0.4502	0.2238	0.3260
288	0.4696	0.2597	0.2707
289	0.2670	0.1818	0.5512
293	0.4406	0.1686	0.3908
294	0.5121	0.1760	0.3119
297	0.3096	0.1419	0.5485
299	0.3098	0.1351	0.5551
300	0.2257	0.5778	0.1965
301	0.4582	0.1537	0.3882
302	0.2466	0.0995	0.6539

Рисунок 3.4 – Результат застосування функції fanny при видалених 30% даних

Коефіцієнт Dunn приймає значення:

dunn_coeff	*normalized*
0.4153731	*0.1230597*

На рисунку 3.4 наведено ординаційну діаграму з результатами кластеризації.

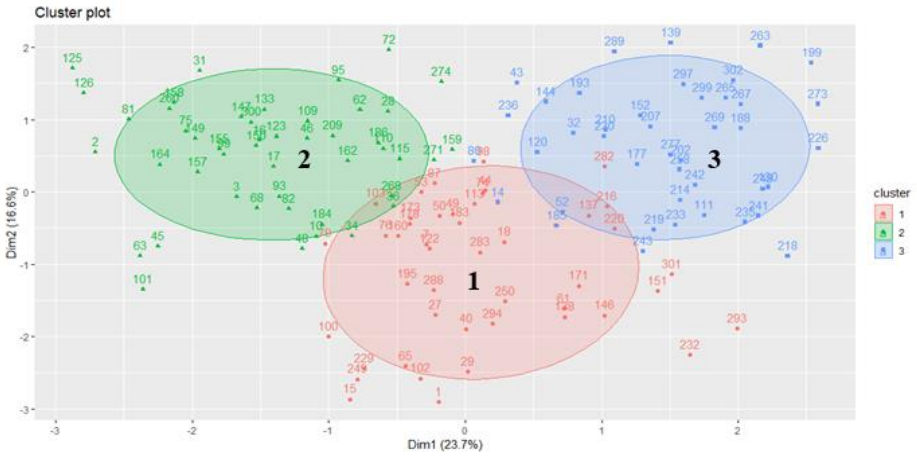


Рисунок 3.5 – Графік результатів нечіткої кластеризації при видалених 30% даних

Кластеризація із заповненням пропусків вибірковими статистиками:

Отже, введемо припущення, що взаємозв'язку між змінними, в даному випадку немає, тоді ефективним способом заповнення пропусків буде використання середніх значень, для цього оберемо медіану.

```
heartMissedProcess<-heartMissed
ind <- apply(heartMissedProcess, 1, function(x) sum(is.na(x))) > 0
heartMissedProcess[ind, 1:8]
pPmI <- preProcess(heartMissedProcess[, 1:8], method = 'medianImpute')
heartMissedProcess[, 1:8] <- predict(pPmI, heartMissedProcess[, 1:8])
(Imp.Med <- heartMissedProcess[ind, 1:8])
```

Перші, середні та останні 10 рядків, отримані за результатами кластеризації, наведено на рисунку 3.5

```
       [,1]  [,2]   [,3]
  [1,] 0.413 0.263 0.3237
  [2,] 0.206 0.660 0.1336
  [3,] 0.303 0.501 0.1959
  [4,] 0.253 0.593 0.1542
  [5,] 0.342 0.270 0.3887
  [6,] 0.354 0.276 0.3703
  [7,] 0.433 0.281 0.2853
  [8,] 0.158 0.739 0.1034
  [9,] 0.204 0.682 0.1133
 [10,] 0.337 0.459 0.2040
             ...
[145,] 0.362 0.304 0.3339
[146,] 0.414 0.181 0.4048
[147,] 0.269 0.555 0.1759
[148,] 0.358 0.416 0.2264
[149,] 0.159 0.744 0.0965
[150,] 0.205 0.657 0.1380
[151,] 0.389 0.168 0.4431
[152,] 0.350 0.273 0.3776
[153,] 0.396 0.341 0.2635
[154,] 0.411 0.314 0.2751
             ...
[294,] 0.432 0.231 0.3372
[295,] 0.282 0.304 0.4141
[296,] 0.309 0.122 0.5691
[297,] 0.334 0.173 0.4935
[298,] 0.370 0.212 0.4187
[299,] 0.327 0.160 0.5131
[300,] 0.285 0.489 0.2261
[301,] 0.394 0.209 0.3967
[302,] 0.265 0.132 0.6034
[303,] 0.423 0.300 0.2771
```

Рисунок 3.6 – Результат застосування функції fanny при заміні 30% значень на медіани

Коефіцієнт Dunn приймає значення:

dunn_coeff normalized

0.38203172 0.07304758

На рисунку 3.6 наведено ординаційну діаграму з результатами кластеризації.

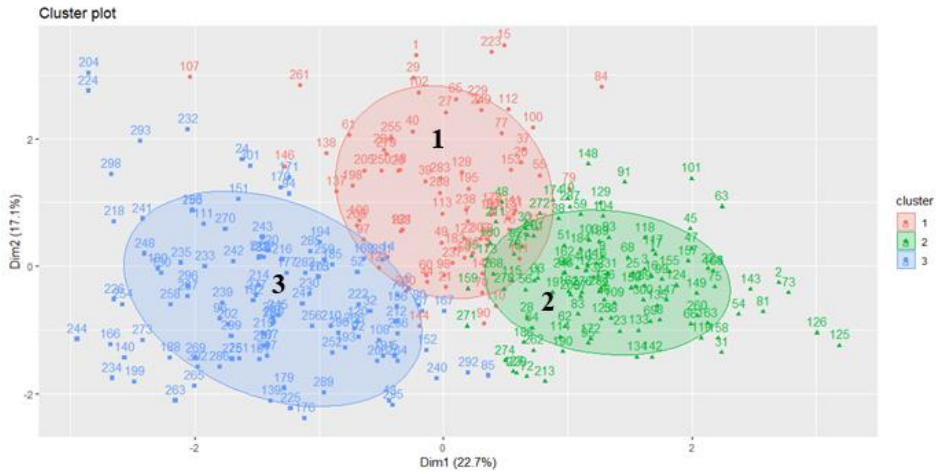


Рисунок 3.7 – Графік результатів нечіткої кластеризації при заміні 30% значень на медіани

Заповнення пропусків з урахуванням структури зв’язків:

Попередній метод передбачав відсутність зав’язків між параметрами, це так званий “наївний” метод. Альтернативою йому можна назвати метод, що враховує зв’язки між параметрами.

Приклад для заповнення поля age:

```
lm(age ~ trestbps, data = heartMissKor)
ageTres <- function(oP) {if (is.na(oP)) return(NA)
  else return(34.7253 + 0.1508   * oP)
}
heartMissKor[is.na(heartMissKor$age), 'age'] <-
  sapply(heartMissKor[is.na(heartMissKor$age), 'trestbps'], ageTres)
heartMissKor[ind, 10]
```

Перші, середні та останні 10 рядків, отримані за результатами кластеризації, наведено на рисунку 3.7.

	[,1]	[,2]	[,3]
1	0.472	0.2742	0.2542
2	0.190	0.7050	0.1051
3	0.292	0.5531	0.1552
4	0.392	0.3199	0.2886
5	0.346	0.2427	0.4117
6	0.337	0.2566	0.4063
7	0.429	0.3461	0.2249
8	0.159	0.7487	0.0926
9	0.401	0.3250	0.2744
10	0.236	0.6512	0.1132
...			
145	0.386	0.3268	0.2872
146	0.394	0.1980	0.4083
147	0.273	0.5945	0.1325
148	0.315	0.5478	0.1371
149	0.143	0.7837	0.0728
150	0.194	0.6897	0.1162
151	0.330	0.1575	0.5122
152	0.359	0.2779	0.3628
153	0.411	0.2994	0.2894
154	0.414	0.2784	0.3075
...			
294	0.424	0.2859	0.2899
295	0.245	0.2601	0.4951
296	0.218	0.0900	0.6924
297	0.301	0.1528	0.5466
298	0.399	0.1970	0.4038
299	0.284	0.1524	0.5633
300	0.298	0.5108	0.1914
301	0.438	0.2158	0.3461
302	0.190	0.1035	0.7063
303	0.424	0.2601	0.3158

Рисунок 3.8 – Результат застосування функції fanny при заміні 30% значень з урахуванням структури зв’язків

Коефіцієнт Dunn приймає значення:

dunn_coeff	*normalized*
0.4116371	*0.1174556*

На рисунку 3.8 наведено ординаційну діаграму з результатами кластеризації при заміні 30% значень з урахуванням структури зв’язків.

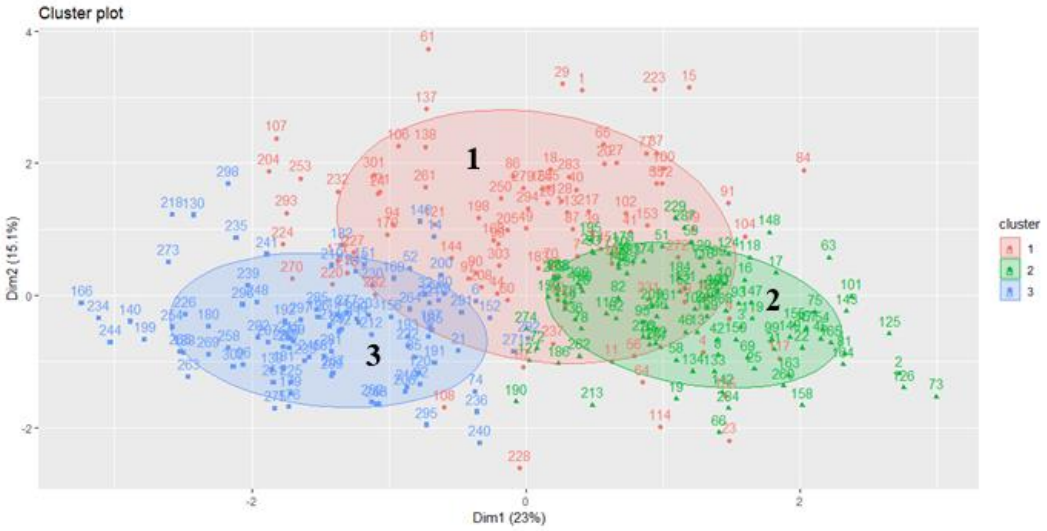


Рисунок 3.9 – Графік результатів нечіткої кластеризації при заміні 30% значень з урахуванням структури зв’язків

Для проведення кластеризації і отримання параметрів на кожному етапі використовувалася функція:

```
kmean <- function(data) {
  res.fanny <- fanny(data, k = 3, memb.exp = 1.3,
                    metric = "euclidean", stand = TRUE, maxit = 500)
  print(res.fanny$membership,3)
  res.fanny$coeff
  print(  res.fanny$coeff)
  Dunn <- res.fanny$membership^2
  fviz_cluster(res.fanny, ellipse.type = "norm", ellipse.level = 0.7)
}
```

Розглядаючи використану функцію можна зазначити, що цільовою таблицею обрано *data* (у даному випадку heart.csv) і, початково обрано 3 кластери. memb.exp – число r, строго більше 1, що вказує на показник ступеня членства, який використовується в критерії відповідності.

Слід зазначити, що $r \to 1$ дає більш чітку кластеризацію, тоді як $r \to Inf$ призводить до повної нечіткості, значення, занадто близькі до 1, можуть привести до повільної конвергенції, навіть значення за замовчуванням r = 2 може привести до повної нечіткості, тобто членства $u\left(i, v\right) == 1 / k$. В цьому випадку рекомендується вибрати менший memb.exp (= r).

metric – рядок символів, який вказує метрику, яка буде використовуватися для розрахунку відмінностей між спостереженнями. Варіанти «euclidean», «manhattan» і «SqEuclidean». Евклідові відстані – це корінь суми квадратів різниць, манхеттенські відстані – сума абсолютних різниць, а «SqEuclidean», квадрат евклідових відстаней – сума квадратів різниць.

stand – параметр логічного типу, якщо true, вимірювання в x стандартизуються перед обчисленням відмінностей. Вимірювання стандартизовані для кожної змінної (стовпчика) шляхом вирахування середнього значення змінної і ділення на середнє абсолютне відхилення змінної.

maxit – максимальна кількість ітерацій.

3.4 Практичні результати прогнозування та кластеризації часових рядів

Для аналізу часових рядів, у якості тестового набору, оберемо дані діастолічної складової артеріального тиску у стані спокою за період у 4 місяці. Графік цього часового ряду наведено на рисунку 3.10.

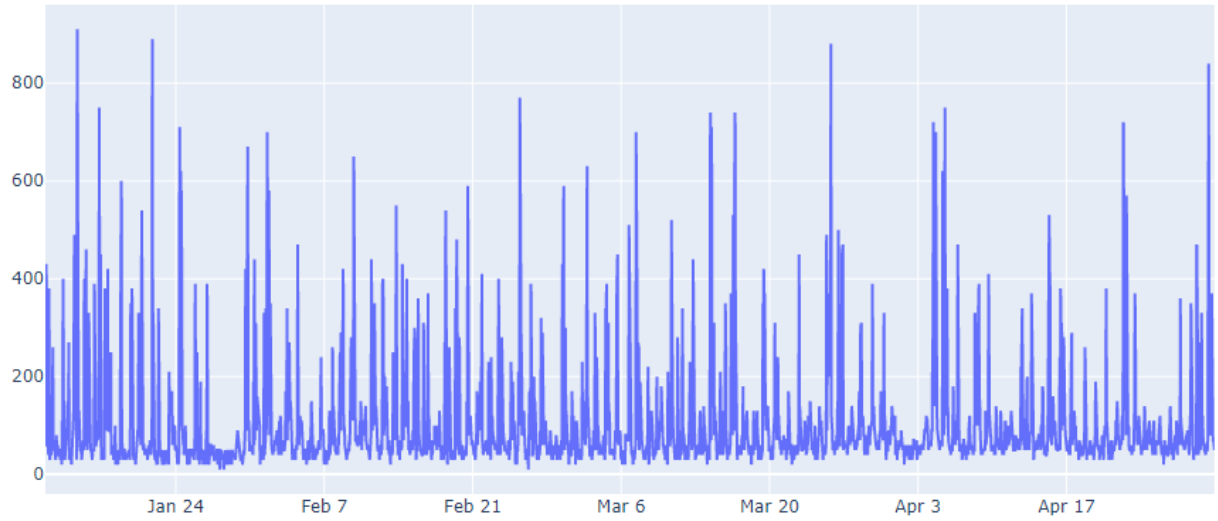


Рисунок 3.10 – Часовий ряд використаної величини

Модель ряду можна представити як:

$$x = m + s + z, \tag{3.2}$$

де m – тренд, s - сезонність, z - похибка.

Модель ряду наведено на рисунку 3.11. З наведеного графіку видно, що випадкова величина z (random) має досить вагомий вплив на часовий ряд, що розглядається.

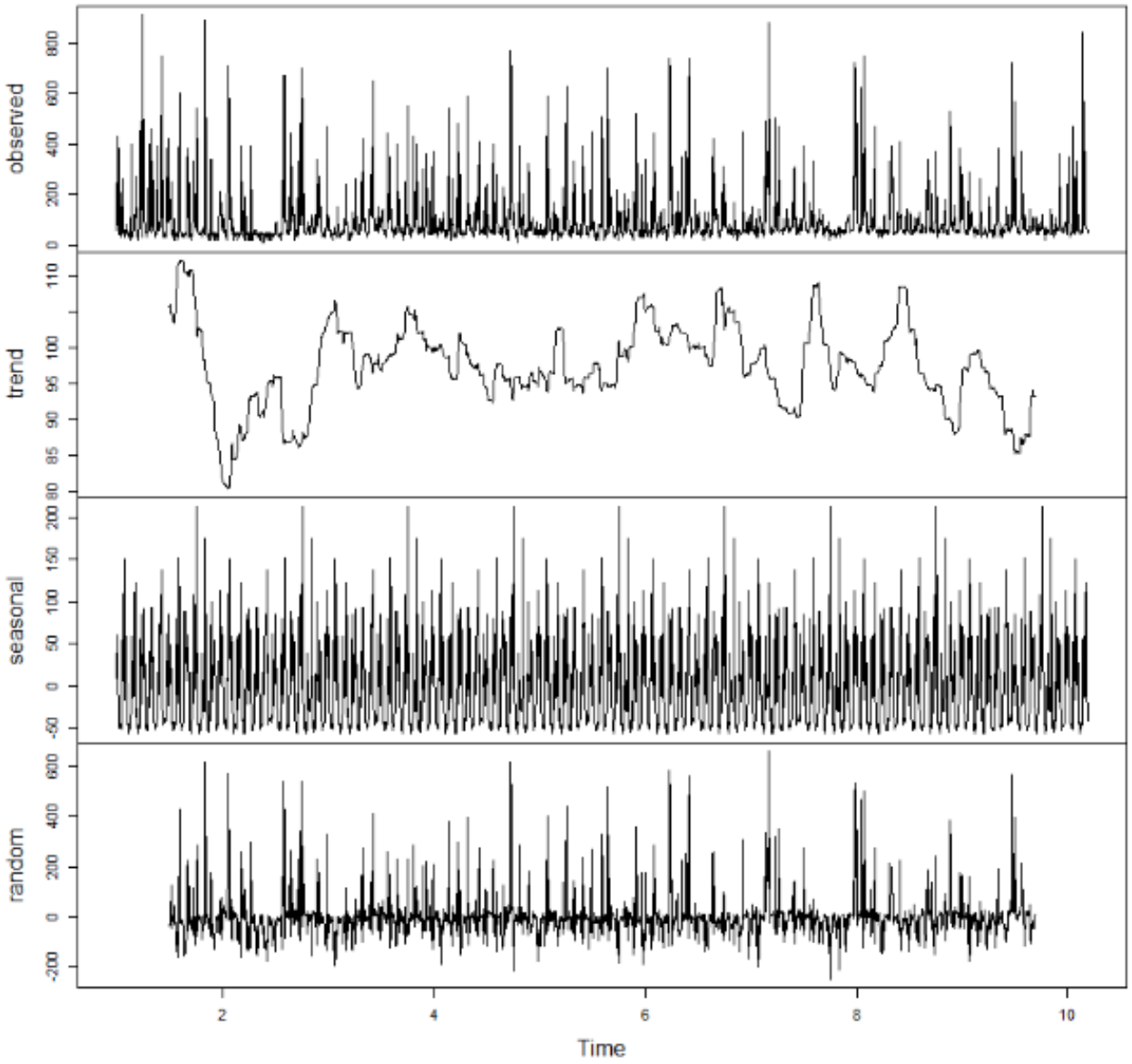


Рисунок 3.11 – Опис часового ряду, що розглядається

3.4.1 Прогнозування ряду та побудова SARIMA моделі

В процесі прогнозування, для того щоб побачити загальні тенденції з урахуванням наявних шумів, проведемо згладжування діастолічної складової артеріального тиску у стані спокою за днями та тижднями, за допомогою ковзної середньої, що наведено на рисунках 3.12, 3.13.

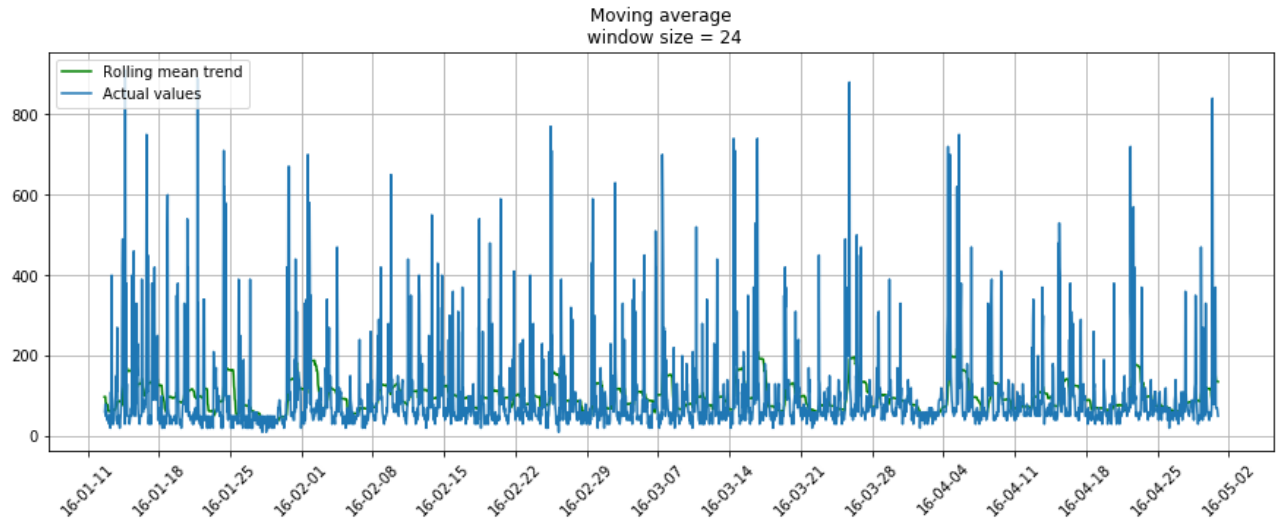


Рисунок 3.12 – Згляджування ряду за днями

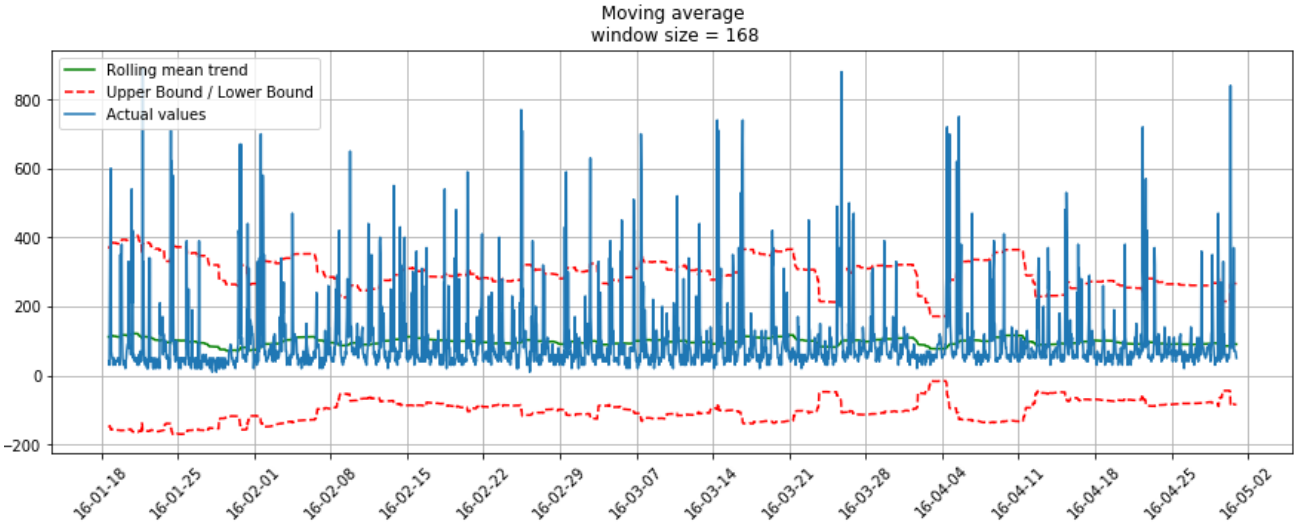


Рисунок 3.13 – Згладжування ряду за тижднями

Дані методи дозволяють прогнозувати на одну точку вперед, але використовуючи подвійне експоненційне згладжування можна побудувати прогноз на дві точки вперед. Такий графік наведено на рисунку 3.14.

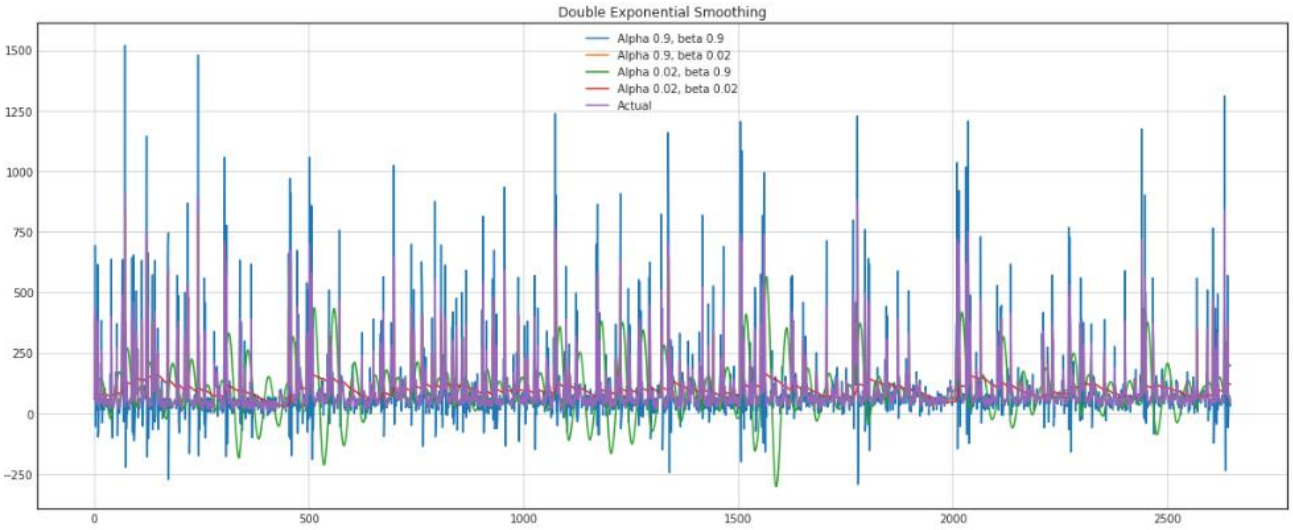


Рисунок 3.14 – Подвійне експоненційне згладжування ряду

Для подальшої обробки з метою отримати прогноз корисним буде використовувати стаціонарний ряд. Стационарним вважається ряд, що з часом не змінює своїх статичних значень (постійні математичне сподівння та дисперсія).

За рисунком 3.15 - 3.17 видно, що заданий ряд таким не є.

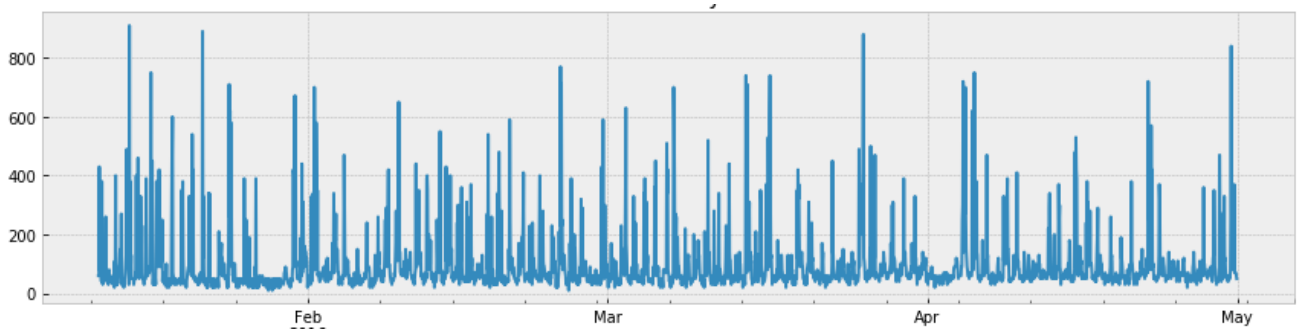


Рисунок 3.15 – Графіки аналізу часового ряду

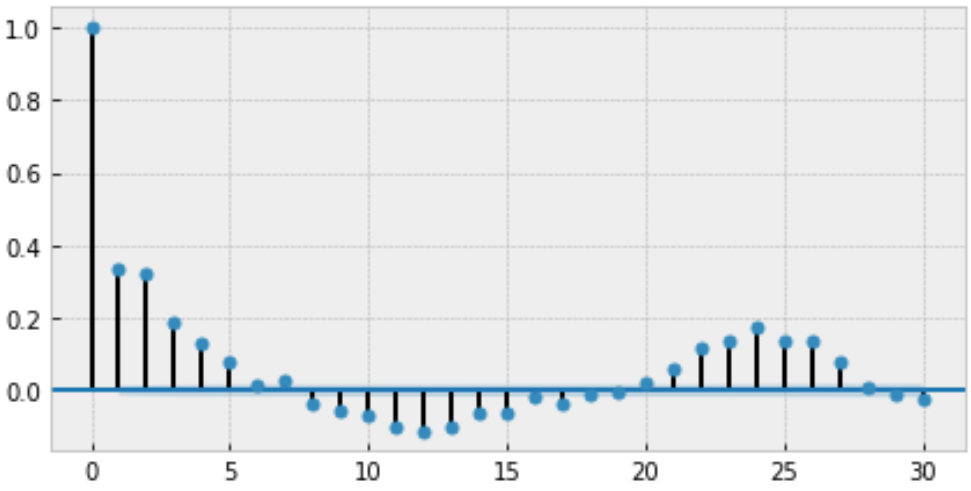


Рисунок 3.16 – Автокореляція часового ряду

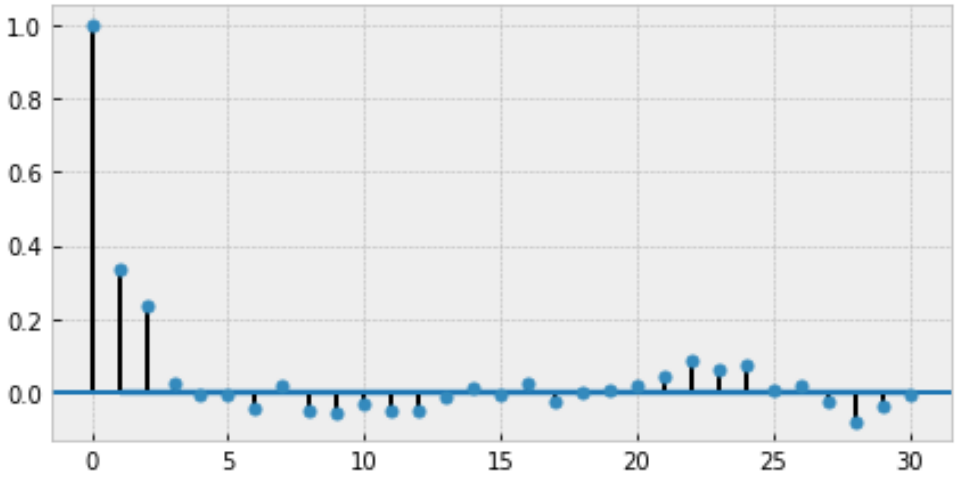


Рисунок 3.17 – Часткова автокореляція часового ряду

Після перетворень отримаємо ряд, що наведено на рисунку 3.18 – 3.20, з яким можна далі працювати та будувати SARIMA модель.

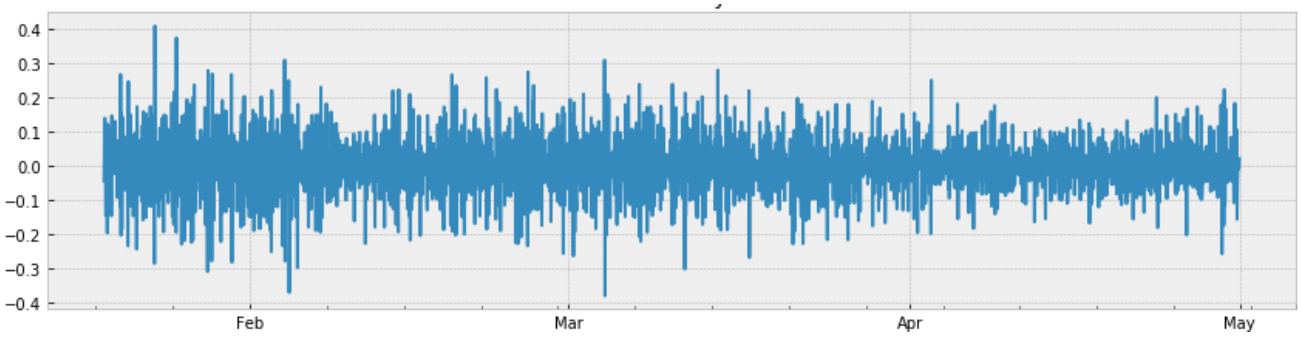


Рисунок 3.18 - Графіки аналізу часового ряду після перетворень

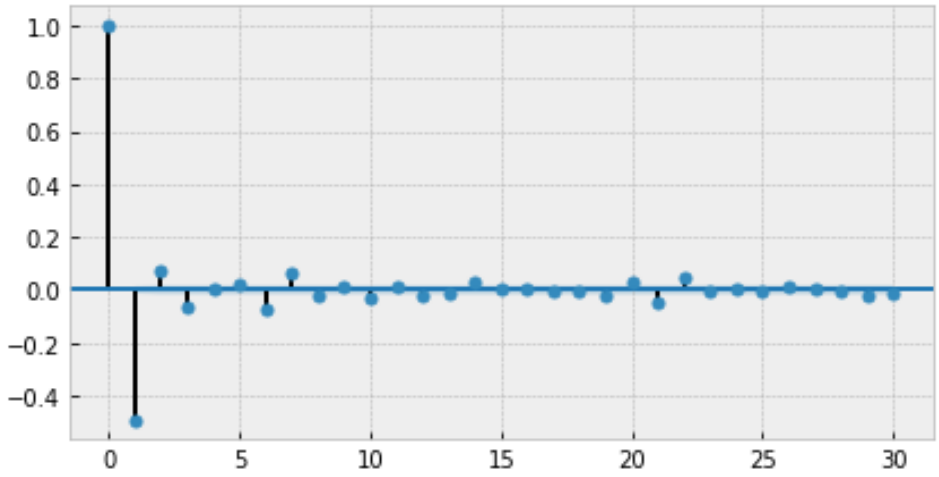


Рисунок 3.19 – Автокореляція часового ряду після перетворень

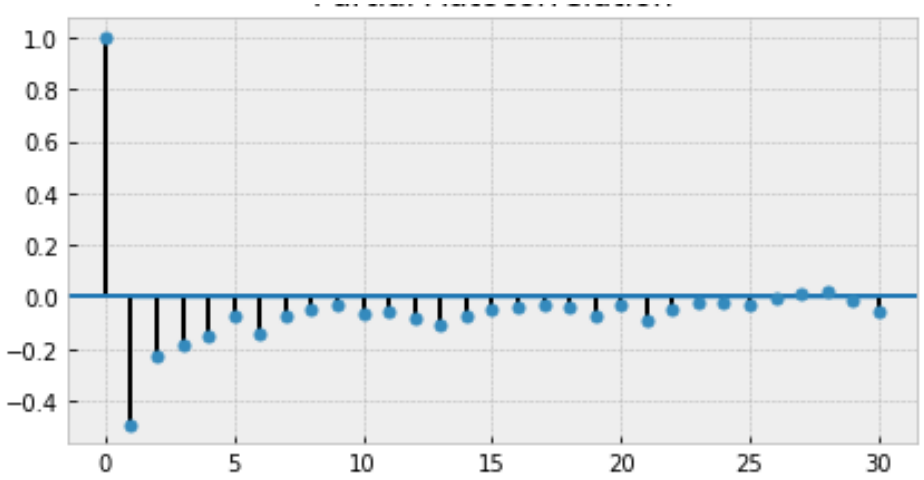


Рисунок 3.20 – Часткова автокореляція часового ряду після перетворень

Побудуємо прогноз по моделі, що вийшла. Модель наведено на рисунку 3.21.

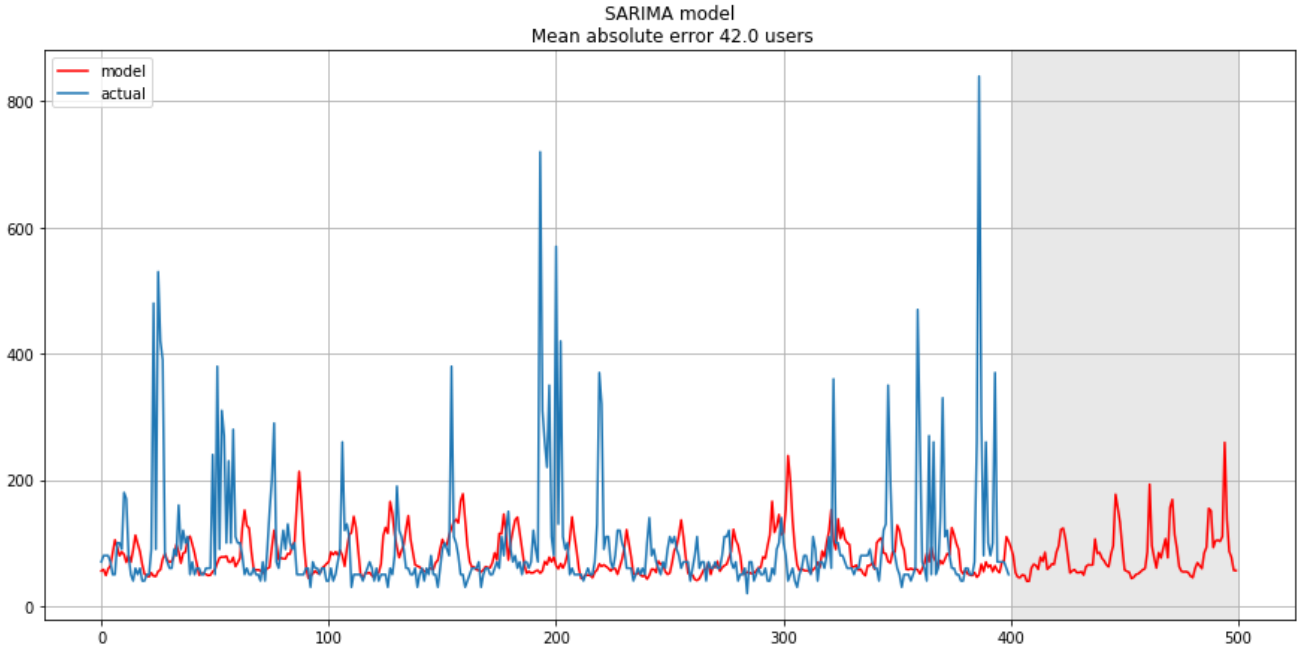


Рисунок 3.21 – Побудована SARIMA модель для похідного ряду

За результатами побудованої моделі можна сказати, що вона вийшла не ідеальною і було отримано не повністю достовірний прогноз, але з урахуванням викидів і шумів можна сказати, що модель не погано відтворила загальну тенденцію.

3.4.2 Кластеризація часового ряду

Наступним кроком проведемо кластеризацію часового ряду. Для цього використаємо: Euclidean k-means, DBA k-means та Soft-DTW k-means кластеризації часових рядів. На рисунках 3.22-3.24 наведено результати такої кластеризації відповідно.

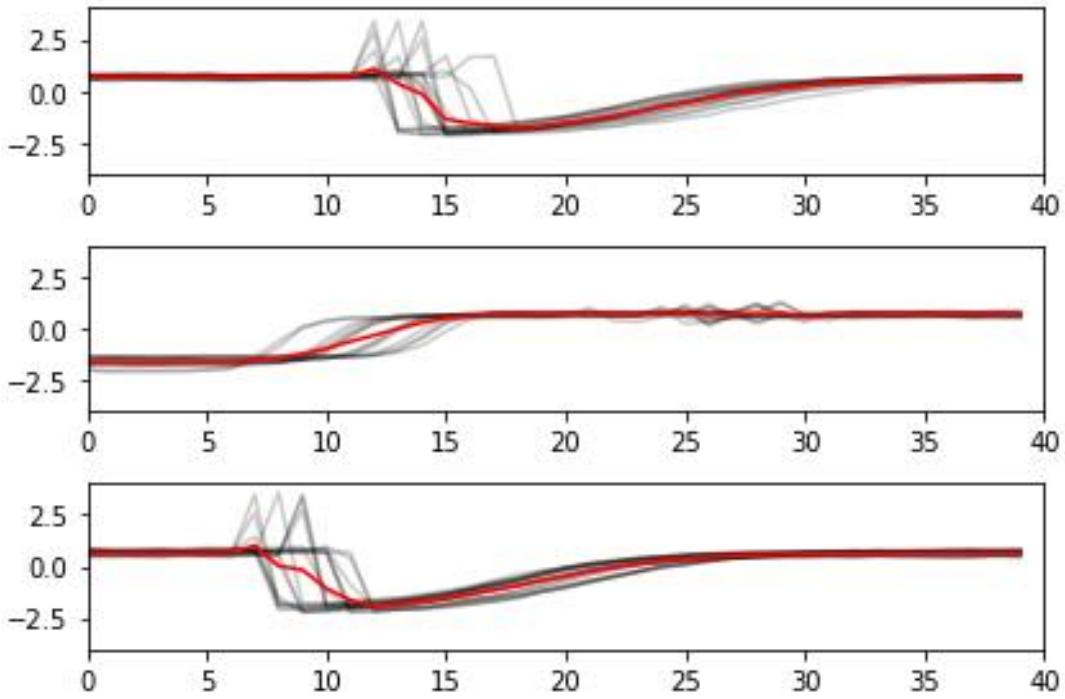


Рисунок 3.22 – Результат кластеризації часового ряду методом Euclidean k-means

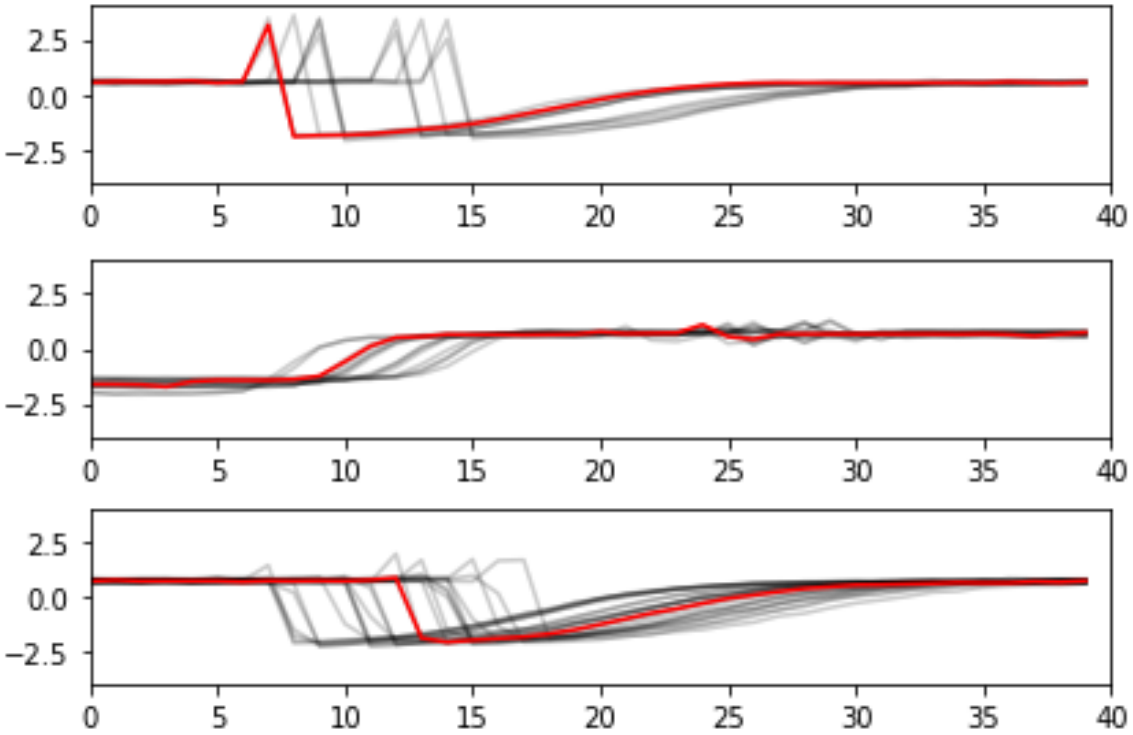


Рисунок 3.23 – Результат кластеризації часового ряду методом DBA k-means

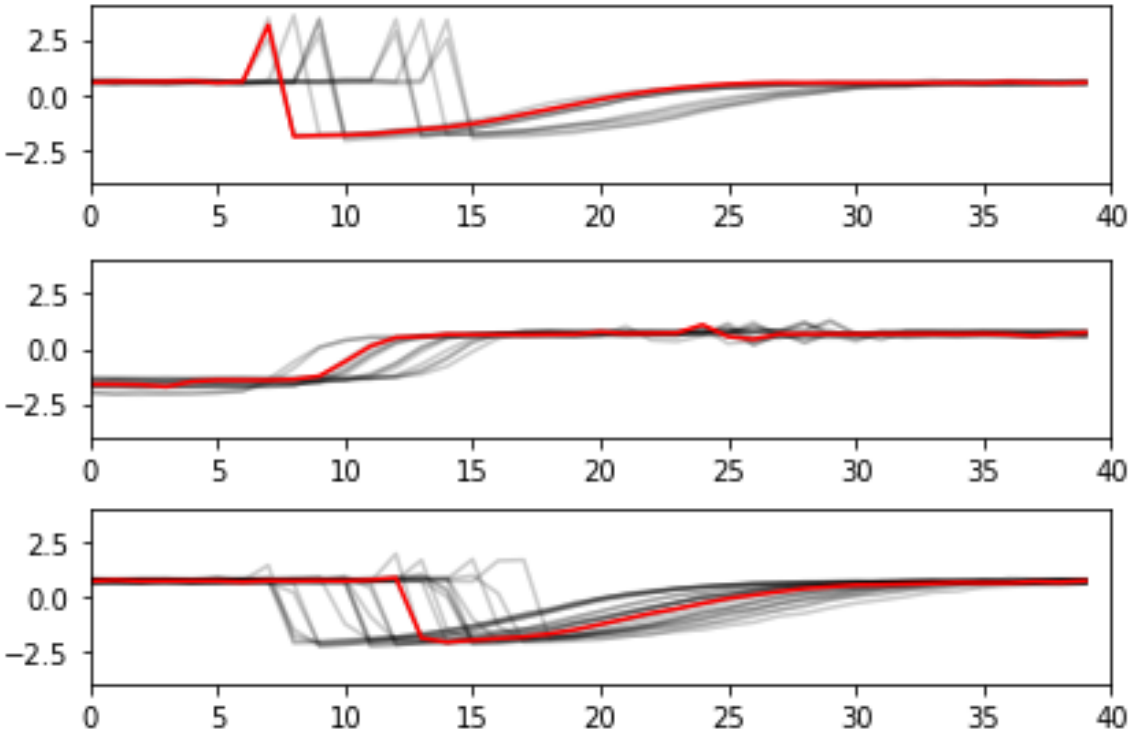


Рисунок 3.24 – Результат кластеризації часового ряду методом Soft-DTW k-means

4 ОХОРОНА ПРАЦІ

4.1 Аналіз потенційних небезпечних і шкідливих виробничих чинників проектованого об’єкту, що мають вплив на персонал

У даному дипломному проекті розробляється програмне забезпечення.

Розроблене програмне забезпечення орієнтоване на роботу з персональним комп'ютером. Експлуатовані для вирішення внутрішньовиробничих завдань ПЕОМ типу IBM PC мають наступні характеристики:

споживана потужність	220 Вт;
робоча напруга	220 В;
напруга джерел живлення	+12 В; - 12 В; +5 В;
робоча частота	50 Гц.

Відповідно до ДСН 3.3.6.042-99 [31] до легкої фізичної роботи відносяться всі види діяльності, виконувані сидячи і ті, що не потребують фізичної напруги. Робота користувача ПК відноситься до категорії 1а.

При роботі на ПЕОМ користувач піддається ряду потенційних небезпек. Унаслідок недотримання правил техніки безпеки при роботі з машиною (невиконання огляду відкритих частин ПЕОМ, що знаходяться під напругою або знятих для ремонту вузлів) для користувача існує небезпека поразки електричним струмом.

Джерелами підвищеної небезпеки можуть служити наступні елементи:

– розподільний щит;
– джерела живлення;
– блоки ПЕОМ і друку, що знаходяться в ремонті.

Ще одна проблема полягає у тому, що спектр випромінювання комп'ютерного монітора включає рентгенівську, ультрафіолетову і інфрачервону області, а також широкий діапазон хвиль інших частот. Небезпека рентгенівського проміння мала, оскільки цей вид випромінювання поглинається речовиною екрану. Проте велику увагу слід приділяти біологічним ефектам низькочастотних електромагнітних полів (аж до порушення ДНК).

Відповідно до ДСанПІН 3.3.2.007-98 [32], при обслуговуванні ПЕОМ мають місце фізичні і психофізичні небезпечні, а також шкідливі виробничі чинники:

– підвищене значення напруги в електричному ланцюзі, замикання

– якої може відбутися через тіло людини;
– підвищений рівень статичної електрики;
– підвищений рівень електромагнітних випромінювань;
– підвищена або знижена температура повітря робочої зони;
– підвищений або знижений рух повітря;
– підвищена або знижена вологість повітря;
– відсутність або недостатність природного світла;
– підвищена пульсація світлового потоку;
– недостатня освітленість робочого місця;
– підвищений рівень шуму на робочому місці;
– розумове перенапруження;
– емоційні навантаження;
– монотонність праці.

4.2 Заходи щодо техніки безпеки

Основним небезпечним чинником при роботі з ЕОМ є небезпека поразки людини електричним струмом, яка посилюється тим, що органи чуття людини не можуть на відстані знайти наявності електричної напруги на устаткуванні.

Проходячи через тіло людини, електричний струм чинить на нього складну дію, що є сукупністю термічної (нагрів тканин і біологічних середовищ), електролітичної (розкладання крові і плазми) і біологічної (роздратування і збудження нервових волокон і інших органів тканин організму) дій.

Тяжкість поразки людини електричним струмом залежить від цілого ряду чинників:

– значення сили струму;
– електричного опору тіла людини і тривалості протікання через нього струму;
– роду і частоти струму;
– індивідуальних властивостей людини і навколишнього середовища.

Розроблений дипломний проект передбачає наступні технічні способи і засоби, що застерігають людину від ураження електричним струмом [33]:

– заземлення електроустановок;
– занулення;

– захисне відключення;

– електричне розділення ятерів;

– використовування малої напруги;

– ізоляція частин, що проводять струм;

– огорожа електроустановок.

Занулення зменшує напругу дотику і обмежує година, протягом якого людина, ткнувшись до корпусу, може потрапити під дію напруги.

Струм однофазного короткого замикання визначається по наближеній формулі:

$$I_K = \frac{\text{Uф}}{\text{Zп} + \frac{\text{Zт}}{3}} \tag{4.1}$$

де Uф - номінальна фазна напруга мережі, В;

Zп - повний опір петлі, створене фазними і нульовими дротами, Ом;

ZT - повний опір струму короткого замикання на корпус, Ом.

Згідно таблиці 4 [34]: $\text{Zт}/3 = 0{,}1$ Ом.

Для провідників і жил кабелю для розрахунку повного опору петлі використовуємо формулу(4.2.) :

$$\text{Zп} = \sqrt{\text{Rп}^2 + \text{Xп}^2}\,, \tag{4.2}$$

де Rп = Rф + Ro - сумарний активний опір фазного Rф і нульового Ro дротів, Ом;

Xп - індуктивний опір паяння дротів, Ом.

Перетин 1 км мідного дроту S = 2.5 мм, тоді згідно таблицям 5 і 6 [34], має такий опір:

Xп = 0,11 Ом;

Rф = 7,55 Ом;

Ro = 7,55 Ом.

Отже, Rп = 7,55 + 7,55 = 15,1 Ом.

Тоді по формулі(4.2) знаходимо повний опір петлі :

$$Zп = \sqrt{15,1^2 + 0,11^2} \approx 15,1_{(Ом)}.$$

Струм однофазного короткого замикання рівний:

$$Iк = \frac{220}{15,1 + 0,1} = 14,47 \text{ (А)}.$$

Дія плавкої вставки на ПЕОМ забезпечується, якщо виконується співвідношення:

$$Iк \geq k * Iн, \quad (4.3)$$

де Iн - номінальний струм спрацьовування плавкої вставки, А;

k - коефіцієнт кратності нелінійного струму Iн, А.

Коефіцієнт кратності нелінійного струму Iн розраховується по формулі(4.4.) :

$$Iн = P / U, \quad (4.4)$$

де P = 220 Вт - споживана потужність;

U = 220 В - робоча напруга;

k = 3 А - для плавких вставок.

Отже, Iн = 220 / 220 = 1 А.

Підставивши значення у вираз(4.3), одержимо:

$$14,47 > 3*1.$$

Таким чином, доведено, що апарат забезпечить спрацьовування (і захист) при підвищенні номінального струму.

4.3 Заходи, що забезпечують виробничу санітарію і гігієну праці

Вимоги до виробничих приміщень встановлюються *ДСП 173-96* [35], СНіП, відповідними ГОСТами і ОСТамі з урахуванням небезпечних і шкідливих чинників, що утворюються в процесі експлуатації електроустаткування.

Підвищення працездатності людини і збереження її здоров'я забезпечується стабільними метеорологічними умовами. Мікроклімат виробничих приміщень [1] визначається діючими на організм людини поєднаннями температури, вологості і швидкості руху повітря, а також температури навколишніх поверхонь. Значне коливання параметрів мікроклімату приводити до порушення систем кровообігу, нервової і потовидільної, що може викликати підвищення або пониження температури тіла, слабкість, запаморочення і навіть непритомність.

Відповідно до ДСН 3.3.6.042-99 [31] встановлюють оптимальну і допустиму температуру, відносну вологість і швидкість руху повітря в робочій зоні . За відсутності надмірного тепла, вологи, шкідливих речовин в приміщенні досить природної вентиляції.

У приміщенні для виконання робіт операторського типу (категорія 1а), пов'язаних з нервово-емоційною напругою, проектом передбачається дотримання наступних нормованих величин параметрів мікроклімату (табл. 4.1).

Таблиця 4.1 - Санітарні норми мікроклімату робочої зони приміщень для робіт категорії 1а.

Пора року	Температура, С	Відносна вологість, %	Швидкість руху повітря, м/с
Холодна	22…24	40…60	0,1
Тепло	23…25	40…60	0,1

У приміщенні, де знаходиться ПЕОМ, повітрообмін реалізується за допомогою природної організованої вентиляції (з пристроєм вентиляційних каналів в перекриттях будівлі і вертикальних шахт) й установленого промислового кондиціонера фірми Mitsubishi, який дозволяє вирішити переважну більшість завдань по створенню та підтримці необхідних параметрів повітряного середовища. Цей метод забезпечує приток потрібної кількості свіжого повітря, визначеного в СНіП (30 $м^3$ в годину на одного працівника).

Шум на виробництві має шкідливу дію на організм людини. Стомлення операторів через шум збільшує число помилок при роботі, призводить до виникнення травм. Для оператора ПЕОМ джерелом шуму є робота принтера. Щоб усунути це джерело шуму, використовують наступні методи. При покупці принтера слід вибирати найбільш шумозахисні матричні принтери або з великою швидкістю роботи (струменеві, лазерні). Рекомендується принтер поміщати в найбільш віддалене місце від персоналу, або застосувати звукоізоляцію та звукопоглинання (під принтер підкладають демпфуючі підкладки з пористих звукопоглинальних матеріалів з листів тонкої повсті, поролону, пеноплену).

При роботі на ПЕОМ, проектом передбачені наступні методи захисту від електромагнітного випромінювання: обмеження часом, відстанню, властивостями екрану.

Обмеження години роботи на ПЕОМ складає 3,5-4,5 години. Захист відстанню передбачає розміщення монітора на відстані 0,4-0,5 м від оператора. Передбачений монітор 20" TFT, Samsung 2043BW відповідає вимогам стандарту [36].

Стандарт [36] пред’являє жорсткі вимоги в таких областях: ергономіка (фізична, візуальна і зручність користування), енергія, випромінювання(електричних і магнітних полів), навколишнє середовище і екологія, а також пожежна та електрична безпека, які відповідають всім вимогам ДСанПІН 3.3.2.007-98 [32].

Для зниження стомлюваності та підвищення продуктивності праці обслуговуючого персоналу в колірній композиції інтер'єру приміщень для ПЕОМ дипломним проектом пропонується використовувати спокійні колірні поєднання і покриття, що не дають відблисків.

У проекті передбачається використання сумісного освітлення. У світлий час доби приміщення освітлюватиметься через віконні отвори, в решту часу використовуватиметься штучне освітлення.

Як штучне освітлення необхідно використовувати штучне робоче загальне освітлення. Для загального освітлення необхідно використовувати люмінесцентні лампи. Вони володіють наступними перевагами: високою світловою віддачею, тривалим терміном служби, хоча мають і недоліки: високу пульсацію світлового потоку.

При експлуатації ПЕОМ виробляється зорова робота. Відповідно до ДБН В.2.5-28:2018 [37] ця робота відноситься до розряду 5а. При цьому нормоване освітлення на робочому місці(Ен) при загальному освітленні дорівнює 200 лк.

Приміщення завдовжки 12 м, шириною 10 м, заввишки 4 м обладнується світильниками типу ЛП02П, оснащеними лампами типу ЛБ зі світловим потоком 3120 лм кожна.

Виконаємо розрахунок кількості світильників в робочому приміщенні завдовжки a=12 м, шириною b=10 м, заввишки з=4 м, використовуючи формулу (4.5) розрахунку штучного освітлення при горизонтальній робочій поверхні методом світлового потоку:

$$n = (E \cdot S \cdot Z \cdot k)/(F \cdot U \cdot M), \quad (4.5)$$

де F - світловий потік = 3120 лм;

E - максимально допустима освітленість робочих поверхонь = 200 лк;

S - площа підлоги = 120 м2;

Z - поправочний коефіцієнт світильника = 1,2;

k - коефіцієнт запасу, що враховує зниження освітленості в процесі експлуатації світильників = 1,5;

n - кількість світильників;

U - коефіцієнт використовування освітлювальної установки = 0,6;

M - кількість ламп у світильнику =2.

З формули(4.5) виразимо n(4.6) і визначимо кількість світильників для даного приміщення:

$$n = (E \cdot S \cdot Z \cdot k)/(F \cdot U \cdot M), \quad (4.6)$$

Отже, $n = (200 \cdot 120 \cdot 1{,}2 \cdot 1{,}5)/(3120 \cdot 0{,}6 \cdot 2) = 12$

Виходячи з цього, рекомендується використовувати 12 світильників. Світильники слід розміщувати рядами, бажано паралельно стіні з вікнами. Схема розташування світильників зображена на рис. 4.1.

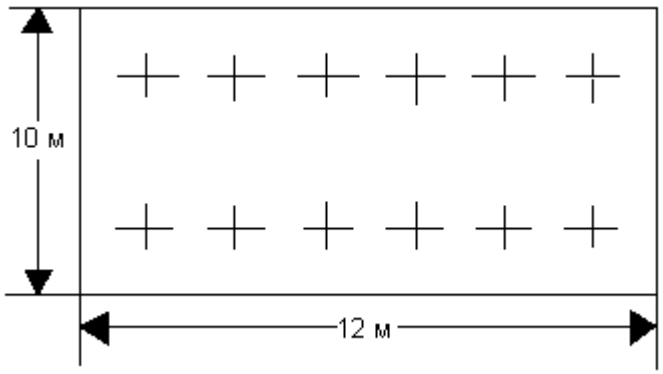


Рисунок 4.1 - Схема розташування світильників

4.4 Рекомендації по пожежній безпеці

Пожежі в приміщеннях, де встановлена обчислювальна техніка, представляють небезпеку для життя людини. Пожежі також пов’язані як з матеріальними втратами, так і з відмовою засобів обчислювальної техніки, що у свою чергу спричиняє за собою порушення ходу технологічного процесу.

Пожежа може виникнути при наявності горючої речовини та внесення джерела запалювання в горюче середовище. Пальними матеріалами в приміщеннях, де розташовані ПЕОМ, є:

– поліамід - матеріал корпусу мікросхеми, горюча речовина, температура самозаймання аерогелю 420 З ;

– полівінілхлорид - ізоляційний матеріал, горюча речовина, температуразапалювання 335 З, температура самозаймання 530 З, кількість енергії, що виділяється при згоранні - 18000 - 20700 кДж/кг;

– стеклотекстоліт ДЦ - матеріал друкарських плат, важкозаймистий матеріал, показник горючості 1.74, не схильний до температурного самозаймання;

– пластика кабельний №489 - матеріал ізоляції кабелю, горючий матеріал, показник горючості більш 2.1;

– деревина - будівельний і обробний матеріал, матеріал з якого виготовлені меблі, горючий матеріал, показник горючості більше 2.1, теплота згорання 18731 - 20853 кДж/кг, температура запалювання 399 З, схильна до самозаймання [34].

Згідно ДСТУ Б В.1.1-36:2016 [38] приміщення відносяться до категорії В (пожежовибухонебезпечним) і згідно правилам побудови електроустановок простір усередині приміщення відноситься до вогненебезпечної зони класу П - IIа (зони, розташовані в приміщеннях, в яких зберігаються тверді горючі речовини).

Потенційними джерелами запалення при роботі ПЕОМ є:

– іскри при замиканні і розмиканні ланцюгів;

– іскри і дуги коротких замикань;

– перегріви від тривалого перевантаження і наявності перехідного опору.

Продуктами згорання, що виділяються при пожежі, є: оксид вуглецю, сірчистий газ, оксид азоту, синильна кислота, акропеін, фосген, хлор та ін. При горінні пластмас, окрім звичайних продуктів згорання, виділяються різні продукти термічного розкладання: хлорангідрідні кислоти, формальдегіди, хлористий водень, фосген, синильна кислота, аміак, фенол, ацетон, стирол та ін., що шкідливо впливають на організм людини.

Для захисту персоналу від дії небезпечних і шкідливих чинників пожежі проектом передбачається застосування промислового протигаза з коробкою марки В (жовта).

Пожежна безпека об’єктів народного господарства регламентується ГОСТ 12.1.004-91 [39] і забезпечується системами запобігання пожежам і протипожежному захисту. Для успішного гасіння пожеж вирішальне значення має швидке виявлення пожежі і своєчасний виклик пожежних підрозділів до місця пожежі.

Зменшити горюче навантаження не представляється можливим, тому проектом передбачається застосувати наступні способи і їх комбінації для запобігання утворенню(внесення) джерел запалення :

– застосування устаткування, що задовольняє вимогам електростатичної безпеки;

– застосування в конструкції швидкодіючих засобів захисного відключення можливих джерел запалення;

– виключення можливості появи іскрового заряду статичної електрики в горючому середовищі з енергією, рівної і вище мінімальної енергії запалення;

– підтримка температури нагріву поверхні машин, механізмів, устаткування, пристроїв, речовин і матеріалів, які можуть увійти до контакту з пальним середовищем, нижче гранично допустимої, становить 80% якнайменшої температури самозаймання пального;

– заміна небезпечних технологічних операцій більш безпечними;

– ізольоване розташування небезпечних технологічних установок і устаткування;

– зменшення кількості пальних і вибухонебезпечних речовин, що знаходяться у виробничих приміщеннях;

– запобігання можливості утворення пальних сумішей на лінії, вентиляційних системах і ін.;

– механізація, автоматизація та справність(потокова) виробництва;

– суворе дотримання стандартів і точне виконання встановленого технологічного режиму;

– запобігання можливості появи в небезпечних місцях джерел запалення;

– запобігання розповсюдженню пожеж і вибухів;

– використовування устаткування і пристроїв, при роботі яких не виникає джерел запалення;

– виконання вимог сумісного зберігання речовин і матеріалів;

– наявність громовідводу;

– організація автоматичного контролю параметрів, що визначають джерела запалення;

– ліквідація можливості самозаймання речовин і матеріалів.

Для запобігання пожежі в обчислювальних центрах проектом пропонується виконання наступних вимог:

– електроживлення ЕОМ повинно мати автоматичне блокування відключення електроенергії на випадок зупинки системи охолоджування і кондиціонування;

– система вентиляції обчислювальних центрів повинна бути обладнана блокуючими пристроями, що забезпечують її відключення на випадок пожежі;

– робочі місця повинні бути оснащені пожежними щитами, сигналізацією, засобами для сповіщення про пожежну небезпеку (телефонами), медичними аптечками для надання першої медичної допомоги, розробленим планом евакуації.

Для зниження пожежної небезпеки в приміщеннях використовуються первинні засоби гасіння пожеж, а також система автоматичної пожежної сигналізації, яка дозволяє знайти початкову стадію загоряння, швидко і точно оповістити службу пожежної охорони про час і місце виникнення пожежі.

Відповідно до НАПБ А.01.001-2014 [40] приміщення категорії В підлягають устаткуванню системами автоматичної пожежної сигналізації. Проектом передбачається застосування датчика типу ІДФ - 1(димовий фотоелектричний датчик), оскільки специфікою пожеж обчислювальної техніки і радіоапаратури є, в першу чергу, виділення диму, а потім - підвищення температури.

При виникненні пожежі в робочому приміщенні обслуговуючий персонал зобов'язаний негайно вжити заходи по ліквідації пожежі. Для ліквідації пожежі використовують вогнегасники (хімічно-пінні, пінні для повітря ОП-5, ОП-6, ОП-9, вуглекислотні ОУ-5), пісок, пожежний інвентар(сокири, ломи, багри, шерстяну або азбестову ковдри) [41]. Як засіб індивідуального захисту проектом передбачається використовування промислового протигаза з маскою, фільтруючої коробки В.

В якості організаційно-технічних заходів рекомендується проводити навчання робочого персоналу правилам пожежної безпеки.

4.5 Охорона навколишнього природного середовища

Діяльність з використання комп’ютерної техніки впливає на навколишнє природне середовище.

Основним екологічним аспектом в процесі діяльності за даними спеціальностями є процеси впливу на атмосферне повітря та процеси поводження з відходами, які утворюються, збираються, розміщуються, передаються на видалення (знешкодження), утилізацію, тощо в ІТ галузі.

Немає впливу на атмосферне повітря при нормальних умовах праці, бо в приміщенні не використовуються сканери, принтери та інші джерела викиду забруднюючих речовин в повітря робочої зони.

В процесі діяльності користувача виникають процеси поводження з відходами ІТ галузі.

Нижче надано перелік відходів, що утворюються в процесі роботи:

– відпрацьовані люмінесцентні лампи - I клас небезпеки;
– змінні носії інформації - IV клас небезпеки;
– макулатура - IV клас небезпеки.

ВИСНОВКИ

В ході роботи було розглянуто методи кластеризації даних та часових рядів з попереднім прогнозуванням. Для проведення кластеризації оброблених даних розглянуто метод кластеризації k-середніх.

За результатами проведеної роботи можна зробити висновок, що позбавляючись від пропусків, найгіршим варіантом є варіант видалення всіх рядків, які містять пропуски. Цей метод можливий лише у випадках коли вибірка містить мінімальну кількість пропусків, або тоді коли було попередньо проведено інший вид обробки і відбувається видалення залишків пустих значень. Найкращим вважається метод боротьби з пропусками з урахуванням взаємозв'язків між полями, але на даній вибірці він не значно перевершує метод заміни на середні значення.

Якщо порівнювати таблиці, що є результатами методу fanny, то можна сказати, що обидва методи впоралися добре на даному наборі даних з урахуванням 30% пропусків. Якщо аналізувати графіки, то можна сказати, що є огріхи в кластеризації, відновлення відбулося не ідеально, тому слід враховувати втрату повної достовірності, при виборі одного з таких методів.

У якості обробки часових рядів було проведено аналіз ряду з метою його прогнозування, та отримано SARIMA модель, що наведено на рисунку 3.21. За цією моделлю можна зробити висновок, що вона побудована не бездоганно, але з урахуванням великої кількості викидів та значного впливу випадної величини, що відображено на рисунку 3.11, можна сказати, що отримана модель не погано відобразила загальну тенденцію ряду.

Також, було проведено КА та отримано результати кластеризації на рисунках 3.22, 3.24. Розглянуто декілька методів кластеризації і можна зробити висновок, що на цьому наборі вони дали досить схожі результати.

Під час виконання розділу «Охорона праці» було проаналізовано умови праці, виявлені причини травматизму і захворювань, можливі небезпечні й шкідливі виробничі фактори. Також було проведено ряд розрахунків щодо виконання вимог охорони праці в приміщенні відділу програмного забезпечення, визначення основних екологічних аспектів впливу на навколишнє середовище та заходи, вжиті для їх подолання. Дотримання цих вимог є важливим для збереження працездатності та здоров'я працівників.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЬ

1) Шумейко, А. А., & Сотник, С. Л. (2012). Интеллектуальный анализ данных (введение в Data Mining). *Днепропетровск: Белая ЕА*, *212*.

2) Деркач, О. І. (2016). Аналітична обробка текстової інформації за допомогою засобів кластеризації. *Young*, *34*(7).

3) Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

4) Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, *31*(3), 264-323.

5) Perret, B., Chierchia, G., Cousty, J., Guimarães, S. J. F., Kenmochi, Y., & Najman, L. (2019). Higra: Hierarchical graph analysis. *SoftwareX*, *10*, 100335.

6) Steinley, D. (2006). K-means clustering: a half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, *59*(1), 1-34.

7) Ackermann, M. R. (2009). *Algorithms for the Bregman k-Median problem* (Doctoral dissertation, University of Paderborn).

8) Khachumov, M. V. (2012). Distances, metrics and cluster analysis. *Scientific and Technical Information Processing*, *39*(6), 310-316.

9) Huang, Z., & Ng, M. K. (1999). A fuzzy k-modes algorithm for clustering categorical data. *IEEE Transactions on Fuzzy Systems*, *7*(4), 446-452.

10) Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2015). *Introduction to time series analysis and forecasting*. John Wiley & Sons.

11) Zhang, J., Zhao, Z., Xue, Y., Chen, Z., Ma, X., & Zhou, Q. (2017). Time series analysis. *Handbook of Medical Statistics*, *269*.

12) Крашений, І. Е., Попов, А. О., Рамірез, Х., Горріз, Х. М., Крашеный, И. Э., Попов, А. А., ... & Горриз, Х. М. (2016). Використання методів кластеризації в системах нечіткого виводу для діагностики хвороби Альцгеймера на основі ПЕТ-зображень.

13) Штовба, С. Д. (2006). Побудова функцій належності нечітких множин за кластеризацією експериментальних даних. *Інформаційні технології та комп'ютерна інженерія*, (2), 92-95.

14) Xu, J., Han, J., Xiong, K., & Nie, F. (2016, July). Robust and Sparse Fuzzy K-Means Clustering. In *IJCAI* (pp. 2224-2230).

15) Gorshkov, Y., Kolodyazhniy, V., & Bodyanskiy, Y. (2009, June). New recursive learning algorithms for fuzzy Kohonen clustering network. In *Proc. 17th Int. Workshop on Nonlinear Dynamics of Electronic Systems* (pp. 58-61).

16) Bodyanskiy, Y. V., Tyshchenko, O. K., & Mashtalir, S. V. (2019, June). Fuzzy Clustering High-Dimensional Data Using Information Weighting. In *International Conference on Artificial Intelligence and Soft Computing* (pp. 385-395). Springer, Cham.

17) Oleg, K., Sergii, M., & Mykhailo, S. (2017, October). Video Clustering via Multidimensional Time-Series Analysis. In *Proceedings of the 9th International Conference on Information Management and Engineering* (pp. 60-63). ACM.

18) Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, *7*(2), 147.

19) Волкова, В. В., & Шафроненко, А. Ю. (2011). Нечітка кластеризація масивів даних з пропущеними значеннями. *Індуктивне моделювання складних систем*.

20) Kesemen, O., Tezel, Ö., & Özkul, E. (2016). Fuzzy c-means clustering algorithm for directional data (FCM4DD). *Expert systems with applications*, *58*, 76-82.

21) Женбинг, Х., Бодянский, Е. В., Тыщенко, А. К., & Ткачев, В. Н. (2017). Fuzzy Clustering Data Arrays with Omitted Observations.

22) Kate, R. J. (2016). Using dynamic time warping distances as features for improved time series classification. *Data Mining and Knowledge Discovery*, *30*(2), 283-312.

23) Hu, Z., Mashtalir, S. V., Tyshchenko, O. K., & Stolbovyi, M. I. (2018). Clustering matrix sequences based on the iterative dynamic time deformation procedure. *International Journal of Intelligent Systems and Applications*, *10*(7), 66-73.

24) Wang, D., Lu, X., & Rinaldo, A. (2017). DBSCAN: Optimal Rates For Density Based Clustering. *arXiv preprint arXiv:1706.03113*.

25) Tiwari, K. K., Raguvanshi, V., & Jain, A. (2016). DBSCAN: An Assessment of Density Based Clustering and It's Approaches.

26) Gautam, C., & Ravi, V. (2015). Data imputation via evolutionary computation, clustering and a neural network. *Neurocomputing*, *156*, 134-142.

27) Пеливан, М. А. (2015). Кластеризация методом наиболее удаленных соседей или методом полной связи. Техногенной безопасности и устойчивого развития, 23.

28) Воронцов, К.В. Алгоритми кластерізації та багатовимірного шкалювання [Текс] / Воронцов К.В. - Курс лекций. МГУ, 2007. – 14с.

29) Jain, A. Data clustering: Review [Text] / A. Jain, M. Murty, P. Flynn – ACM Computing Surveys Vol. 31, 1999 –pp. 540 - 597.

30) Котов, А. Кластерізація даних [Текс] / А. Котов, Н. Красільніков. –206. – 295с.

31) Санітарні норми мікроклімату виробничих приміщень ДСН 3.3.6.042-99. Постанова N 42 від 01.12.99. Режим доступу: www. URL: https://zakon.rada.gov.ua/rada/show/va042282-99

32) Державні санітарні правила і норми роботи з візуальними дисплейними терміналами електронно-обчислювальних машин ДСанПІН 3.3.2.007-98. Затверджено Постановою Головного державного санітарного лікаря України 10 грудня 1998 р. N 7. Режим доступу: www. URL: https://zakon.rada.gov.ua/rada/show/v0007282-98

33) НПАОП 40.1-1.21-98 «Правила безпечної експлуатації електроустановок споживачів». *Наказ від 09.01.98 №4.* Режим доступу: www. URL: https://zakon.rada.gov.ua/laws/show/z0093-98

34) ГОСТ 12.1.044-89 «Система стандартов безопасности труда. Пожаровзрывоопасность веществ и материалов. Номенклатура показателей и методы их определения». Постанова від 12.12.1989 № 3683. Режим доступу: www. URL: http://online.budstandart.com/ru/catalog/doc-page?id_doc=51048

35) *ДСП 173-96 «Державні санітарні правила планування та забудови населених пунктів». Наказ від 19.06.1996 №173.* Режим доступу: www. URL: https://zakon.rada.gov.ua/laws/show/z0379-96

36) TCO'07 Certified Displays. © 2007 Copyright TCO Development AB. Режим доступу: www. URL: https://tcocertified.com/files/2015/11/TCO-Certified-Displays-7.0.pdf

37) ДБН В.2.5-28:2018 «Природне і штучне освітлення». Режим доступу: www. URL: http://www.minregion.gov.ua/wp-content/uploads/ 2018/12/V2528-1.pdf

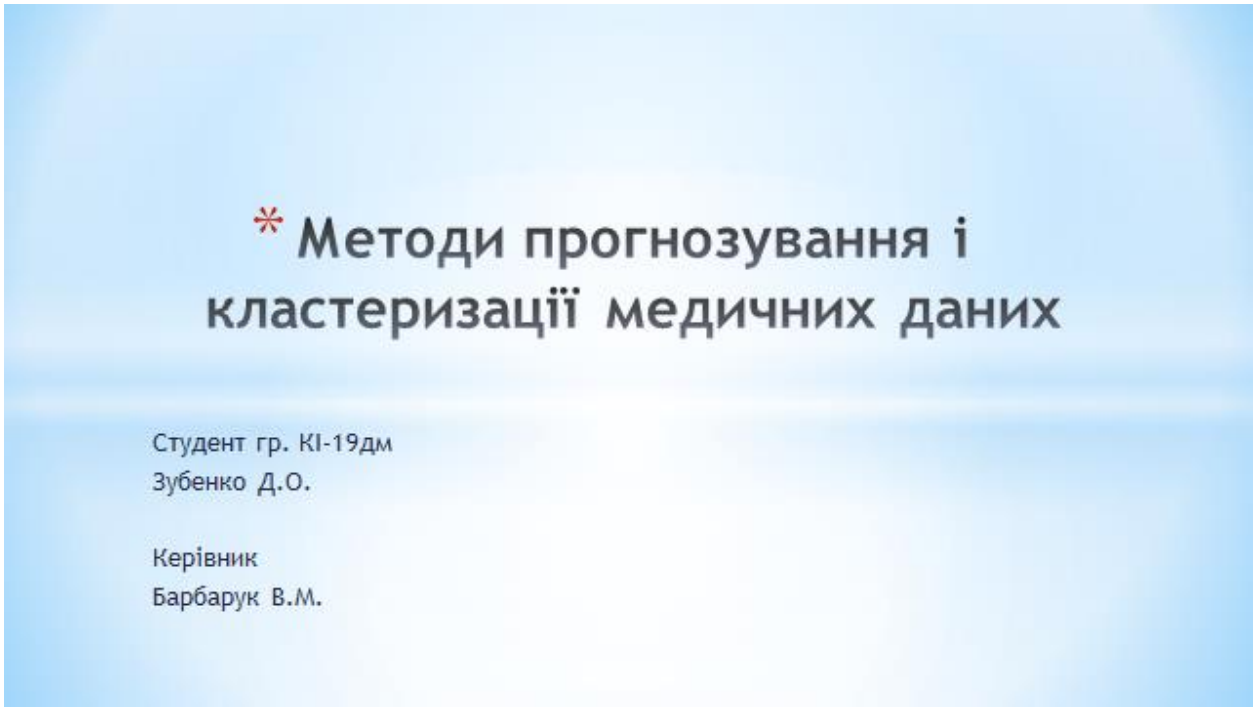
38) ДСТУ Б В.1.1-36:2016 «Визначення категорій приміщень, будинків та зовнішніх установок за вибухопожежною та пожежною небезпекою». Наказ від 15.06.2016 №158. Режим доступу: www. URL: https://zakon.rada.gov.ua/rada/show/v0158858-16

39) ГОСТ 12.1.004-91 «Система стандартов безопасности труда. Пожарная безопасность. Общие требования». Режим доступу: www. URL: http://online.budstandart.com/ua/catalog/doc-page.html?id_doc=48679

40) НАПБ А.01.001-2014 «Правила пожежної безпеки в Україні». Наказ від 30.12.2014 №1417. Режим доступу: www. URL: https://zakon.rada.gov.ua/laws/show/z0252-15

41) НАПБ Б.01.008-2018 «Про затвердження правил експлуатації та типових норм належності вогнегасників». Наказ від 15.01.2018 №25. Режим доступу: www. URL: http://search.ligazakon.ua/ l_doc2.nsf/link1/ RE31677.html

ДОДАТОК А.
ЕЛЕКТРОННІ ПЛАКАТИ

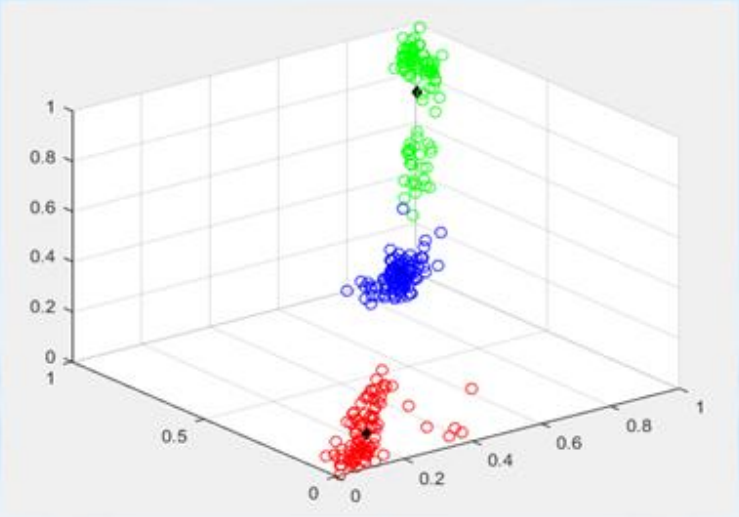


* Постановка задачі

* **Предмет дослідження** - це методи підготовки та обробки вхідних даних, що містять пропущені значення, для їх подальшого прогнозування, аналізу та використання в задачах кластеризації; розгляд адаптованого класичного методу кластеризації для вирішення проблеми неповних даних.
* **Метою роботи** можна назвати розгляд існуючих методів позбавлення від пропусків в даних в задачах кластеризації та доцільність їх використання і реальних задачах. Аналіз та оцінка адаптованих методів кластеризації для вирішення подібного роду задач, та висновок про їх переваги перед методами попереднього позбавлення від пропусків пере процесом кластеризації.

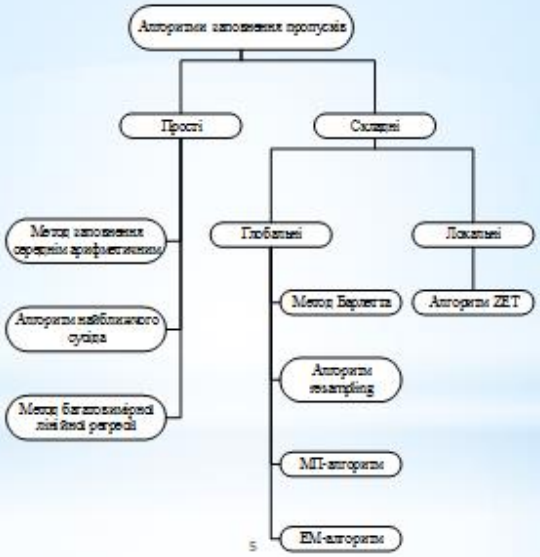
3

* Приклад результату кластеризації тестових даних



4

* Класифікація алгоритмів заповнення пропусків



5

* МАТЕМАТИЧНІ МОДЕЛІ КЛАСТЕРИЗАЦІЇ ТА ПРОГНОЗУВАННЯ ЧАСОВИХ РЯДІВ

- k-середніх і його варіанти;
- ієрархічні алгоритми;
- засновані на щільності методи кластеризації, такі як DBSCAN.

6

* Розгляд предметної області

	age	sex	cp	trestbps	fbs	restecg	thalach	exang
2	63	1	3	145	1	0	150	0
3	37	1	2	130	0	1	187	0
4	41	0	1	130	0	0	172	0
5	56	1	1	120	0	1	178	0
6	57	0	0	120	0	1	163	1
7	57	1	0	140	0	1	148	0
8	56	0	1	140	0	0	153	0
9	44	1	1	120	0	1	173	0
10	52	1	2	172	1	1	162	0
11	57	1	2	150	0	1	174	0
12	54	1	0	140	0	1	160	0
13	48	0	2	130	0	1	139	0
14	49	1	1	130	0	1	171	0
15	64	1	3	110	0	0	144	1

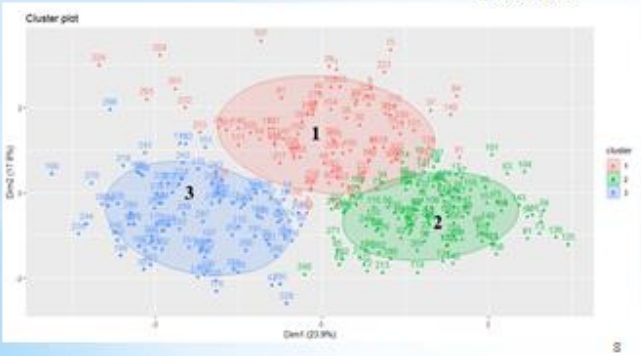
7

* Побудова "нечітких" кластерів з використанням функції fanny()

res.fanny <- fanny(data, k = 3, memb.exp = 1.3,
metric = "euclidean", stand = TRUE, maxit = 500)
print(res.fanny$membership,3)

dunn_coeff
0.4203087

normalized
0.1304630




```
        [,1]    [,2]    [,3]
[1,] 0.503 0.2476 0.2492
[2,] 0.203 0.7080 0.0889
[3,] 0.322 0.5473 0.1304
[4,] 0.239 0.6528 0.1079
[5,] 0.323 0.2860 0.3913
[6,] 0.343 0.3160 0.3407
[7,] 0.526 0.2766 0.1974
[8,] 0.138 0.7956 0.0665
[9,] 0.466 0.2874 0.2469
[10,] 0.403 0.4520 0.1449
                ...
[145,] 0.409 0.2480 0.3427
[146,] 0.449 0.1752 0.3762
[147,] 0.278 0.6023 0.1195
[148,] 0.479 0.3619 0.1593
[149,] 0.154 0.7817 0.0641
[150,] 0.205 0.6947 0.1006
[151,] 0.389 0.1626 0.4487
[152,] 0.366 0.2239 0.4102
[153,] 0.483 0.2632 0.2538
[154,] 0.556 0.2238 0.2203
                ...
[294,] 0.502 0.2211 0.2764
[295,] 0.210 0.3137 0.4762
[296,] 0.189 0.0882 0.7227
[297,] 0.280 0.1620 0.5580
[298,] 0.385 0.1915 0.4240
[299,] 0.265 0.1431 0.5919
[300,] 0.285 0.5300 0.1853
[301,] 0.439 0.2076 0.3530
[302,] 0.142 0.0932 0.7650
[303,] 0.468 0.3800 0.1517
```

* Нечітка кластеризація при видалених 30% даних

```
heartMissedNO <- subset(heartMissed, exang != "")
```

dunn_coeff	normalized
0.4153731	0.1230597

9

```
      [,1]   [,2]   [,3]
1   0.4942 0.2168 0.2890
2   0.1298 0.7809 0.0893
3   0.2238 0.6263 0.1499
7   0.4347 0.2923 0.2730
10  0.3306 0.4800 0.1894
14  0.3695 0.2501 0.3804
15  0.4660 0.2665 0.2675
16  0.1732 0.7063 0.1205
17  0.2529 0.5813 0.1659
18  0.4144 0.2278 0.3578
             ...
160 0.3868 0.3782 0.2349
162 0.2761 0.5348 0.1892
164 0.1312 0.7797 0.0891
171 0.4473 0.1848 0.3678
173 0.3781 0.3545 0.2674
177 0.3981 0.1947 0.4072
178 0.4136 0.3201 0.2663
183 0.4156 0.2709 0.3135
184 0.3178 0.4751 0.2072
185 0.3908 0.1834 0.4258
             ...
283 0.4502 0.2238 0.3260
288 0.4696 0.2597 0.2707
289 0.2670 0.1818 0.5512
293 0.4406 0.1686 0.3908
294 0.5121 0.1760 0.3119
297 0.3096 0.1419 0.5485
299 0.3098 0.1351 0.5551
300 0.2257 0.5778 0.1965
301 0.4582 0.1537 0.3882
302 0.2466 0.0995 0.6539
```

* Кластеризація із заповненням пропусків вибірковими статистиками

```
heartMissedProcess<-heartMissed
ind <- apply(heartMissedProcess, 1, function(x) sum(is.na(x))) > 0
heartMissedProcess[ind, 1:8]
pPmI <- preProcess(heartMissedProcess[, 1:8], method =
'medianImpute')
heartMissedProcess[, 1:8] <- predict(pPmI, heartMissedProcess[, 1:8])
(Imp.Med <- heartMissedProcess[ind, 1:8])
```

10

```
       [,1]  [,2]  [,3]
 [1,] 0.413 0.263 0.3237
 [2,] 0.206 0.660 0.1336
 [3,] 0.303 0.501 0.1959
 [4,] 0.253 0.593 0.1542
 [5,] 0.342 0.270 0.3887
 [6,] 0.354 0.276 0.3703
 [7,] 0.433 0.281 0.2853
 [8,] 0.158 0.739 0.1034
 [9,] 0.204 0.682 0.1133
[10,] 0.337 0.459 0.2040
            ...
[145,] 0.362 0.304 0.3339
[146,] 0.414 0.181 0.4048
[147,] 0.269 0.555 0.1759
[148,] 0.358 0.416 0.2264
[149,] 0.159 0.744 0.0965
[150,] 0.205 0.657 0.1380
[151,] 0.389 0.168 0.4431
[152,] 0.350 0.273 0.3776
[153,] 0.396 0.341 0.2635
[154,] 0.411 0.314 0.2751
            ...
[294,] 0.432 0.231 0.3372
[295,] 0.282 0.304 0.4141
[296,] 0.309 0.122 0.5691
[297,] 0.334 0.173 0.4935
[298,] 0.370 0.212 0.4187
[299,] 0.327 0.160 0.5131
[300,] 0.285 0.489 0.2261
[301,] 0.394 0.209 0.3967
[302,] 0.265 0.132 0.6034
[303,] 0.423 0.300 0.2771
```

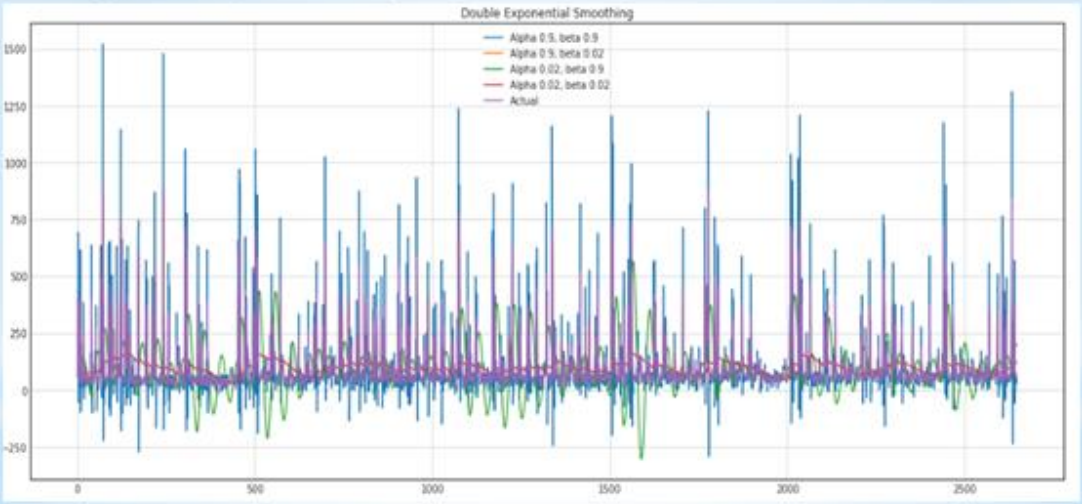
* Заповнення пропусків з урахуванням структури зв'язків

```
lm(age ~ trestbps, data = heartMissKor)
ageTres <- function(oP) {if (is.na(oP)) return(NA)
 else return(34.7253 + 0.1508  * oP)
}
heartMissKor[is.na(heartMissKor$age), 'age'] <-
 sapply(heartMissKor[is.na(heartMissKor$age), 'trestbps'], ageTres)
heartMissKor[ind, 10]
```

11

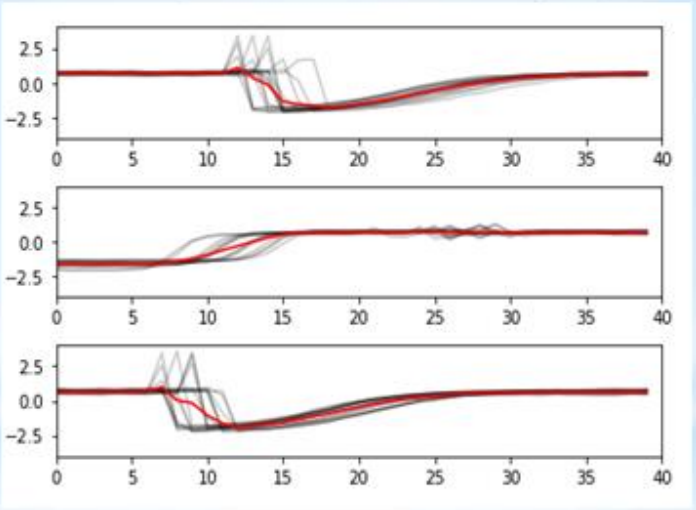
	[,1]	[,2]	[,3]
1	0.472	0.2742	0.2542
2	0.190	0.7050	0.1051
3	0.292	0.5531	0.1552
4	0.392	0.3199	0.2886
5	0.346	0.2427	0.4117
6	0.337	0.2566	0.4063
7	0.429	0.3461	0.2249
8	0.159	0.7487	0.0926
9	0.401	0.3250	0.2744
10	0.236	0.6512	0.1132
		...	
145	0.386	0.3268	0.2872
146	0.394	0.1980	0.4083
147	0.273	0.5945	0.1325
148	0.315	0.5478	0.1371
149	0.143	0.7837	0.0728
150	0.194	0.6897	0.1162
151	0.330	0.1575	0.5122
152	0.359	0.2779	0.3628
153	0.411	0.2994	0.2894
154	0.414	0.2784	0.3075
		...	
294	0.424	0.2859	0.2899
295	0.245	0.2601	0.4951
296	0.218	0.0900	0.6924
297	0.301	0.1528	0.5466
298	0.399	0.1970	0.4038
299	0.284	0.1524	0.5633
300	0.298	0.5108	0.1914
301	0.438	0.2158	0.3461
302	0.190	0.1035	0.7063
303	0.424	0.2601	0.3158

* Прогнозування ряду та побудова SARIMA моделі



12

* Кластеризація часового ряду



13

* ВИСНОВКИ

* В ході роботи було розглянуто методи кластеризації даних та часових рядів з попереднім прогнозуванням. Для проведення кластеризації оброблених даних розглянуто метод кластеризації k-середніх.
* За результатами проведеної роботи можна зробити висновок, що позбавляючись від пропусків, найгіршим варіантом є варіант видалення всіх рядків, які містять пропуски. Цей метод можливий лише у випадках коли вибірка містить мінімальну кількість пропусків, або тоді коли було попередньо проведено інший вид обробки і відбувається видалення залишків пустих значень. Найкращим вважається метод боротьби з пропусками з урахуванням взаємозв'язків між полями, але на даній вибірці він не значно перевершує метод заміни на середні значення.
* Якщо порівнювати таблиці, що є результатами методу fanny, то можна сказати, що обидва методи впоралися добре на даному наборі даних з урахуванням 30% пропусків. Якщо аналізувати графіки, то можна сказати, що є огріхи в кластеризації, відновлення відбулося не ідеально, тому слід враховувати втрату повної достовірності, при виборі одного з таких методів.