Хмельницький національний університет
Факультет програмування
та комп'ютерних і телекомунікаційних систем
Кафедра телекомунікацій, медійних та інтелектуальних технологій

ДИПЛОМНА РОБОТА МАГІСТРА

Застосування кластерного аналізу для визначення факторів ризику інфекційних захворювань COVID-19

Галузь знань :11 – Математика та статистика

Спеціальність:113 – Прикладна математика

Шифр ДРПМ 2019/100.01.05.00

Виконав: студентка II курсу, група ПМм 19 _____ В.Ю.Варгата
Підпис, дата Ініціали, прізвище

Керівник _____ Н.В.Грипинська
Підпис, дата Ініціали, прізвище

До захисту допускаю:
Зав. кафедри ТМІТ _____ С.К.Підченко
Підпис, дата Ініціали, прізвище

14 12 2020 р.

Хмельницький, 2020

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

Факультет Програмування та комп'ютерних і телекомунікаційних систем

Кафедра телекомунікацій, медійних та інтелектуальних технологій

Освітній рівень магістр

Галузь знань 11 Математика та статистика

Спеціальність 113 Прикладна математика

Освітня програма освітньо-професійна програма підготовки магістра

ЗАТВЕРДЖУЮ

Зав. Кафедри, доктор

технічних наук, доцент Підченко

Сергій Констянтинович

" 03 " 09 2020 р.

ЗАВДАННЯ

НА ДИПЛОМНИЙ ПРОЕКТ (РОБОТУ)

Варгата Вікторія Юріївна

Прізвище, ім'я, по батькові студента

1. Тема проекту (роботи) Застосування кластерного аналізу для визначення факторів ризику інфекційних захворювань COVID-19

Керівник проекту (роботи) Грипинська Н.В., к.ф.-м.н., доцент кафедри ТМІТ

Прізвище, ім'я, по батькові, науковий ступінь, вчене звання

Затверджена наказом ректора університету від 01.09.2020 р. № 118

2. Строк подання студентом проекту (роботи) на кафедру 01.12.2020 р.

3. Вихідні дані до проекту (роботи): Модель динаміки, моделі кластерного аналізу, дендрограма кластеризації, карта груп регіонів поширення захворюваностей.

4. Зміст пояснювальної записки (перелік питань, які потрібно розробити) Аналіз предметної області, аналіз статистики захворюваності, аналітичний огляд результатів обробки джерел інформації, розроблення гіпотези дослідження, аналіз статистичних даних, постановка задачі й завдань досліджень, дослідження динаміки захворюваності. Розробка кластерної моделі аналізу

5. Перелік графічного матеріалу (із зазначенням обов'язкових креслень) Презентаційні матеріали, дендрограма кластеризації регіонів України за показниками захворюваності населення (метод Варда), Дендрограма кластеризації регіонів України за показниками захворюваності населення (методом повного зв'язку)

КАЛЕНДАРНИЙ ПЛАН

Назва етапів (розділів) дипломного проекту (роботи)	Строк виконання етапів проекту (роботи)	Примітка
1. Аналіз предметної області	01.06-23.06.2020	Виконала
2. Аналіз статистики захворюваності	25.06-25.07.2020	Виконала
3. Аналітичний огляд результатів обробки джерел інформації	26.07-01.08.2020	Виконала
4.Розроблення гіпотези дослідження	02.08-15.08.2020	Виконала
5.Аналітичний огляд обробки джерел інформації	16.08-01.09.2020	Виконала
6.Аналіз статистичних даних	01.09-25.09.2020	Виконала
7.Методи проведення кластерного аналізу	26.09-02.10.2020	Виконала
8.Виконання індивідуального завдання за обраною тематикою	03.10-15.10.2020	Виконала
9.Постановка задачі й завдань досліджень	15.10-17.10.2020	Виконала
10.Дослідження динаміки захворюваності	18.10-23.10.2020	Виконала
11.Розробка кластерної моделі аналізу	24.10-08.10.2020	Виконала

Студент _____ В.Ю. Варгата

Підпис Ініціали, прізвище

Керівник проекту (роботи) _____ Н.В. Грипинська

Підпис Ініціали, прізвище

АНОТАЦІЯ

Тема дипломної роботи: Застосування кластерного аналізу для визначення факторів ризику інфекційних захворювань COVID-19.

Автор роботи: В.Ю. Варгата

Керівник роботи: Н.В. Грипинська

Загальний обсяг роботи: 100 сторінок, 22 рисунка, 3 таблиці, 3 додатки, 16 посилань.

МЕТОД ВАРДА ФАКТОРИ РИЗИКУ МОДЕЛІ КЛАСТЕРНОГО АНАЛІЗУ, ГІПОТЕЗИ ДОСЛІДЖЕННЯ, COVID-19, МЕТОД РАНГІВ, МЕТОД К-СЕРЕДНІХ.

Метою даної роботи є визначення факторів ризику інфекційних захворювань CoVID-19, методами кластерного аналізу. Отримано моделі кластерного аналізу, виділено фактори ризику, регіони з найбільшою поширеністю, та запропоновано методи протидії.

ANNOTATION

Thesis topic: Application of cluster analysis to determine risk factors for infectious diseases COVID-19.

Author of the work: V.Y. Vargata

Supervisor: N.V. Gripynska

Total volume of work: 100 pages, 22 figures, 3 tables, 3 appendices, 16 links.

WARDO METHOD RISK FACTORS OF CLUSTER ANALYSIS MODELS, RESEARCH HYPOTHESES, COVID-19, RANKING METHOD, K-AVERAGE METHOD.

The aim of this work is to determine the risk factors for infectious diseases CoVID-19, using cluster analysis methods. Cluster analysis models were obtained, according to which risk factors, main causes and methods of solving the problem of disease spread were identified. Regions with the highest prevalence have been identified, and methods of counteraction have been proposed.

14 грудня 2020
(дата/date)

(підпис/signature)

ЗМІСТ

Вступ 6

1 Аналіз предметної області 8

1.1 Аналіз динаміки поширення захворювання 8

1.2 Аналітичний огляд результатів обробки джерел інформації 17

1.3 Розроблення гіпотези дослідження 21

2 Аналітичний огляд обробки джерел інформації 29

2.1 Аналіз статистичних даних 29

2.2 Аналіз чинників та їх вплив на динаміку поширення COVID-19 34

2.3 Методи проведення кластерного аналізу 43

3 Проведення аналізу методами кластеризації 70

3.1 Постановка задачі й завдань досліджень 70

3.2 Дослідження динаміки захворюваності 71

3.3 Розробка математичної моделі кластерного аналізу 76

Висновки 86

Перелік джерел посилання 88

Додаток А. Наукова теза 89

Додаток Б. Презентаційні матеріали 93

Додаток В. Антиплагіат 101

ВСТУП

2020 рік надовго відіб`ється в пам`яті людства роком пандемії нового віруса. 31 грудня 2019 року ВООЗ була проінформована про виявлення випадків пневмонії, викликаної невідомим збудником, 3 січня китайські служби повідомили ВООЗ про 44 випадки пневмонії в місті Ухань провінції Хубей. Патоген виявився новим корона вірусом (нині відомим як SARS-CoV-2, раніше - під тимчасовою назвою 2019 nCoV), який раніше не виявлявся серед людської популяції. 30 січня 2020 року у зв'язку зі спалахом епідемії Всесвітня організація охорони здоров`я оголосила «надзвичайну ситуацію міжнародного значення», а 28 лютого 2020 року ВООЗ підвищила оцінку ризиків на глобальному рівні з високих на дуже високі. 11 березня 2020 року епідемія була визнана пандемією. Пандемія небезпечна тим, що одночасне захворювання інфекцією безлічі людей може привести до перевантаженості системи охорони здоров'я з підвищеною кількістю госпіталізацій і смертей.

Системи охорони здоров'я можуть виявитися не готовими до надзвичайно великої кількості тяжкохворих пацієнтів. Найбільш важливою відповіддю по відношенню до інфекції не є лікувальні заходи, а зниження швидкості її поширення, щоб розтягнути її в часі і знизити, таким чином, навантаження на системи охорони здоров'я. Епідемія закінчиться, як тільки серед населення виробиться достатній колективний імунітет.

Для того, щоб кількість захворювань не зростала щоденно в експоненціальній формі, результатом чого на сьогоднішній день є всесвітня пандемія, важливим етапом у боротьбі з епідемією є вивчення динаміки приросту кількості захворювань. Ефективне дослідження захворюваності, а саме - збір статистичних даних, аналіз та побудова динамічної моделі допоможуть контролювати кількість захворювань, ефективно вводити нові карантинні обмеження, що в результаті пришвидшить кінець пандемії.

Тематикою даної роботи, є застосування кластерного аналізу для визначення факторів ризику інфекційних захворювань COVID-19.

Метою даної роботи є визначення факторів ризику інфекційних захворювань CoVID-19, застосовуючи методи кластерного аналізу.

Завдання даної роботи має наступне формулювання: розробити моделей кластерного аналізу методами Варда, рангів та методом повного зв`язку на основі результатів кластерного аналізу.

Об`єктом дослідження, є фактори які впливають на динаміку поширення захворюваності;

Предмет дослідження: дані про кількість інфікованих осіб, та приросту захворювань;

Згідно отриманих результатів, разом з вірусологами Хмельницької обласної інфекційної лікарні було проведено апробацію запропонованих методів. Робота отримала схвальні відгуки на теоретичному рівні, а запропоновані методи лікарі оцінили як «перспективні в процесі боротьби з пандемією».

РОЗДІЛ 1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

1.1 Аналіз динаміки поширення захворювання

COVID-19 (абревіатура від англ. COronaVIrus Disease 2019) - коронавірусна інфекція 2019 року nCoV – визначена як потенційно важка та гостра респіраторна інфекція, яка викликається новим видом коронавірусу SARS-CoV-2 (2019 nCoV). Являє собою небезпечне распіраторне захворювання, яке може протікати як у формі ГРВІ легкого перебігу так і у важкій формі. Найбільш частим ускладненням захворювання є вірусна пневмонія, здатна призводити до гострого респіраторного дистрес-синдрому і подальшої гострої дихальної недостатності, при яких найчастіше необхідні киснева терапія і респіраторна підтримка. Найбільш поширеними симптомами захворювання є підвищена температура тіла, стомлюваність і сухий кашель. У рідкісних випадках ураження вірусом дітей і підлітків, може призводити до розвитку гострого запального синдрому.

Поширюється вірус переважно шляхом повітряно-крапельним через вдихання в повітрі розпорошених при кашлі, чханні, фізичному контакті або розмові часток з вірусом, а також через потрапляння вірусу на різні поверхні з подальшим занесенням в очі, ніс, рот та будь які інші слизові оболонки тіла. До числа ефективних заходів профілактики відноситься: «дезінфекція рук і дотримання правил респіраторної гігієни». Захворювання викликається новим вірусом, проти якого у людей спочатку немає імунітету до інфекції сприйнятливі люди будь-яких вікових категорій.

На даний момент проти вірусу відсутні будь-які засоби лікування або профілактики (в тому числі і вакцини). У більшості випадків (приблизно в 90%) якогось специфічного лікування не потрібно, і одужання відбудеться саме по собі. Важкі форми хвороби з більшою ймовірністю можуть розвинутися у літніх людей а також у людей з певними супутніми

захворюваннями, що включають астму, діабет і серцеві захворювання. У край важких випадках часто застосовуються деякі засоби для підтримки функцій життєво важливих органів.

У більшості тих, що заразилися інфекція протікає в легкій формі або безсимптомно. Приблизно в 10% випадків захворювання протікає у «важкій» формі з необхідністю застосування кисневої терапії, ще в 5% стан хворих критичне. В цілому по світу на 16 травня летальність захворювання оцінюється приблизно в 6,5%. Згідно з аналізом даних по 1099 пацієнтам станом на 27 лютого 2020 року біля 91,1% пацієнтів з COVID-19 діагностувалася пневмонія. Показники з плином часу можуть змінитися.

У зв'язку з епідемією Всесвітньою організацією охорони здоров'я (ВООЗ) оголошено надзвичайну ситуацію у сфері суспільної охорони здоров'я, яке має міжнародне значення, а ризики на глобальному рівні оцінюються як дуже високі. Ситуація швидко розвивається, щодня збільшується кількість хворих і загиблих. Ведуться різні наукові та клінічні дослідження. Багато наукові і медичні видавництва і організації підписалися під заявою про вільний доступ і обмін інформацією, пов'язаною з новим захворюванням.

ВООЗ опублікувала дані, згідно з яких за станом на 30 березня 2020 роки кількість смертей, поділена на кількість діагностованих випадків, становило 4,7% (29 957/634 835) [4]. До показників смертності відносяться коефіцієнт летальності (CFR), який відображає відсоток діагностованих людей, які помирають від захворювання, і коефіцієнт смертності від інфекції (IFR), який представляє собою відсоткову частку померлих від загального числа інфікованих (діагностованих і не діагностованих). Ці показники не прив'язані до часу і розраховуються після завершення епідемії. Ряд вчених намагалися розрахувати ці цифри для конкретних груп населення [2]. Значення CFR створюють перебільшене враження ризику від хвороби в умовах, коли багато випадків захворювання залишаються непоміченими. З іншого боку, цей показник може применшувати реальний ризик, якщо не

реєструються випадки смерті від захворювання. Показник змінюється в залежності від різних факторів, включаючи охоплення тестування. У Китаї оцінки CFR знизилися з 17,3% (для тих, у кого симптоми почалися 1-10 січня 2020 року) до 0,7% (для тих, у кого симптоми з'явилися після 1 лютого 2020 року).

11 березня дане поширення вірусу було визнано як «пандемія». Ця епідемія є найпершою в історії людства «пандемією», яку можуть взяти під контроль. Урядам має сенс підготувати списки навченого персоналу, який здатний взяти ситуацію під контроль, а також списки медикаментів, засобів індивідуального захисту, припасів і обладнання, необхідних для лікування. ВООЗ закликає країни до підготовки лікарень, забезпечення захисту медичних працівників та до рішення про необхідність прийняття тих чи інших заходів соціального дистанціювання.

31 грудня 2019 року ВООЗ була проінформована про виявлення випадків пневмонії, викликаної невідомою інфекцією, 3 січня китайські служби повідомили ВООЗ про 44 випадках пневмонії в місті Ухань провінції Хубей. Патоген виявився новим коронавірусів (нині відомим як SARS-CoV-2, раніше - під тимчасовою назвою 2019 nCoV), який раніше не виявлявся серед людської популяції. 30 січня 2020 року у зв'язку зі спалахом епідемії ВООЗ оголосила режим надзвичайної ситуації міжнародного значення в галузі охорони здоров'я, а 28 лютого 2020 року ВООЗ підвищила оцінку ризиків на глобальному рівні з високих на дуже високі. 11 березня 2020 року епідемія була визнана пандемією. Пандемія небезпечна тим, що одночасне захворювання інфекцією безлічі людей може привести до перевантаженості системи охорони здоров'я з підвищеною кількістю госпіталізацій і смертей. Системи охорони здоров'я можуть виявитися не готовими до надзвичайно великої кількості тяжкохворих пацієнтів. Найбільш важливою відповіддю по відношенню до інфекції не є лікувальні заходи, а зниження швидкості її поширення, щоб розтягнути її в часі і знизити, таким чином, навантаження на системи охорони здоров'я. Епідемія закінчиться в момент, як тільки серед

населення виробиться достатній колективний імунітет, або буде розроблено дієву вакцину.

Згідно з аналізом 72 314 випадків захворювань, проведеним Центром з контролю і запобігання захворювання Китаю, станом на 11 лютого 2020 року 87% випадків захворювання припадали на вік від 30 до 79 років, 1% - на дітей 9 років і молодше, ще 1% - на дітей і підлітків у віці від 10 до 19 років, а 3% хворих були літніми людьми у віці від 80 років. Співвідношення чоловічої і жіночої статі склало 51% до 49% відповідно. Серед захворілих 4% були медичними працівниками. На прикладі Китаю стало зрозуміло, що поширення інфекції можна обмежувати, зупиняючи спалаху захворюваності.

У США приблизно третина хворих є літніми людьми віком від 65 років. На них припадає майже половина госпіталізацій, 53% переказів в реанімацію і 80% смертей серед хворих COVID-19.

Для правильної оцінки ризиків необхідні додаткові дослідження для виявлення розповсюдженості вірусів серед населення в цілому, у тому числі серед людей без симптомів захворювання. Реальна розповсюдженість COVID-19, спектр його поширення та реальний рівень вмісту залишків. Зареєстровані показники смертності в різних країнах дуже неоднорідно: у Німеччині повідомляється про дуже незначний розмір смерті (CFR 0,35%) за порівнянням з іншими європейськими країнами з аналогічним поселенням та системами здорового харчування, що свідчить про відсутність одноманітних критеріїв оцінки смертних випадків.

Станом на 02.09.2020 р., з моменту початку епідемії, статистика приросту захворюваності має наступний вигляд:

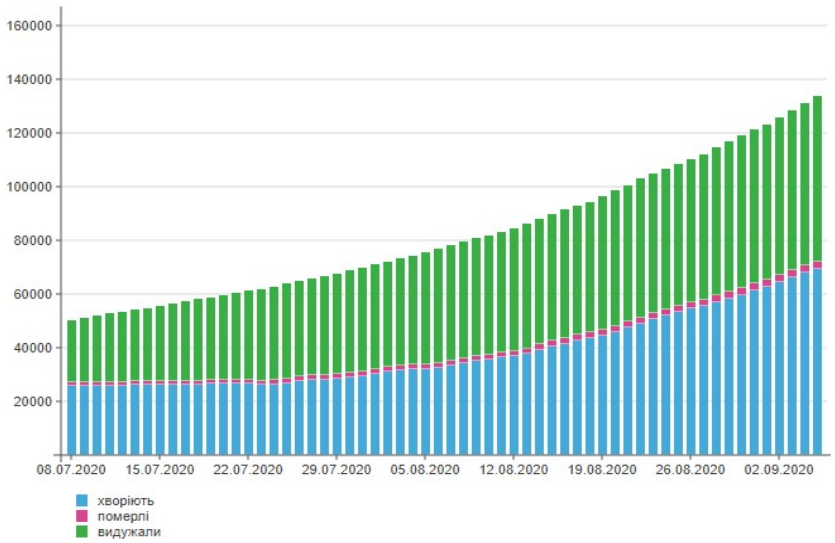


Рисунок 1.1 – Статистика захворюваності станом на 02.09.2020 р.

З кінця вересня, рішенням кабміну було введено умови «Адаптивного» карантину, згідно якого населені пункти, області та регіони поділяють за рівнями небезпеки на зони: «Червона» (найнебезпечніша), «Жовтий» (Середній рівень небезпеки) та «Зелену».

До «червоного» рівня небезпеки віднесено 100 адміністративних одиниць, в той час як він діє у 68 містах та районах. Дане рішення прийняла Державна комісія з питань ТЕБ та ДСНС.

До дев`яти обл. центрів, у яких було встановлено «червоний» рівень, потрапили ще 2 міста – Миколаїв та Рівне. Отже, він діятиме у 11 обласних центрах: Рівному, Чернівцях, Чернігові, Хмельницькому, Харкові, Києві, Миколаєві, Черкаси, Івано-Франківську, Сумах, та Тернополі.

До так званого помаранчевого рівня небезпеки поширення вірусу віднесено 437 АО, що на 37 більше, ніж за тиждень до того. Зокрема, він діятиме у дванадцяти обласних центрах: Луцьку, Вінниці, Житомирі, Дніпро, Львові, Краматорськ, Ужгороді, Одесі, Запоріжжі, Херсоні, Черкасах та у Києві.

«Жовтий» рівень небезпеки задіяно у 54-х АО. До зеленого рівня епідемічної небезпеки віднесено 27 АО, тоді як за тиждень до цього він був встановлений у 54-х містах та районах.

М.В.Степанов звернувся до централізованих органів влади а також уповноважених контролюючих органів з заявою на посилення контролю щодо дотримання обмежувальних карантинних заходів, встановлених у регіонах відповідно до рівня епідемічної небезпеки.

Кластерний аналіз (Cluster analysis) – це аналіз, багатовимірна статистична функція, яка виконує забір даних, що містять інформацію щодо вибірки об'єктів, в результаті впорядковуючи об'єкти в відносно однорідні групи. Основне завдання кластерного аналізування відноситься до статистичної обробки і до широкого класу задач навчання «без учителя».

Кластерний аналіз – це не якийсь заздалегіть визначений алгоритм, це загальна задача, на розв'язання якої використовуються різноманітні підходи. Зокрема, алгоритми побудови кластерів можуть відрізнятись у розумінні того, що саме відносити до одного кластер і як їх ефективно шукати. До найпопулярніших концепцій кластерів відносять групи з елементами, що утворюються ґрунтуючись на відстані між об`єктами, та щільності ділянок у просторовій сфері даних, інтервалах чи на конкретних статистичних розподілах. Сама кластеризація може формулюватися як задача «багатокритеріальної оптимізації». Відповідний алгоритм кластеризації і вибору параметрів (а також включаючи такі параметри, як: функція відстані, порогове значення щільності, кількість очікуваних кластерів) залежать від конкретного набору даних а також мети використання результатів. Кластерний аналіз не є автоматизованим завданням, а навпаки – «ітераційним процесом» виявлення або інтерактивною багатокритеріальною оптимізацією, що містить в собі як «спроби» так і «невдачі». Доволі часто доводиться змінювати процес опрацювання даних і параметри моделі доки не буде отримано результат із заданими наперед властивостями.

Більшість із дослідників та аналітиків схиляються до того, що вперше термін «кластерний аналіз» було запропоновано математиком Рене Тріона. Згодом виник ряд термінів, які на сьогоднішній день прийнято вважати

синонімом терміну «кластерний аналіз»: автоматична класифікація, ботріологія.[1]

Спектр застосувань «кластерного аналізу» доволі широкий: його використовують в таких сферах як: археологія, медицина, психологія, хімія, біологія, державному управлінні, філології, антропології, маркетингу, соціології, геології та інакших дисциплінах. Проте - універсальність цього застосування призвела до появи великої к-сті несумісних термінів, методів та підходів, що ускладнюють однозначне використання та несуперечливу інтерпретацію кластерного аналізу.

Кластерний аналіз використовується для опису просторового та часового порівняння спільнот (сукупностей) організмів у неоднорідних середовищах. Він також використовується в систематиці рослин для генерування штучного філогенію або скупчень організмів (особин) виду, роду чи вищого рівня, що мають низку ознак.

Загальноприйнятої класифікації методів кластерного аналізу не існує, проте можна виділити ряд груп та підходів (деякі методи можна віднести відразу до кількох груп й тому пропонується розглядати дану типізацію як наближення до реальної класифікації методів кластеризації) [9] :

1.Імовірнісний підхід. Він передбачає, що кожний даний об'єкт відноситься до одного із К класів. Деякі автори (наприклад, А.В.Орлов) вважає, що ця група зовсім не відноситься до вирішення задач «кластеризації», також протиставляють її під назвою «дискримінація», мається на увазі вибір віднесення об'єктів до однієї із відомих груп:

- ЕМ-алгоритм;
- дискримінантний аналіз;
- К-середніх;
- К-медіан;
- Алгоритми сімейства FOREL;

2.Підходи засновані на основі систем штучного інтелекту: досить умовна група:

- метод не чіткої кластеризації С-середніх;
- нейромережа Кохоннена;
- генетичний алгоритм;

3.Логічний підхід. Будування дендограми здійснюється за допомогою так званого «дерева рішень».

4.Теоретико-Графовий підхід;

5.Ієрархічний підхід. Цей метод передбачає наявність укладених груп (кластери абсолютно різного порядку). Алгоритми ж в свою чергу поділяються на: агломеративні та дивизивн. За к-стю ознак також виділяють монотетичні та політетичні методи класифікації.

Ієрархічна дивизивна кластеризація або таксономія. Завдання кластеризації розглядаються в кількісної таксономії.

6.Методи, які не ввійшли у попередні групи;.

- статистичний алгоритм кластеризації;
- квартет кластеризаторів;
- алгоритми сімейства KRAB;
- алгоритм, заснований на методі просіювання;
- DBSCAN і ін.

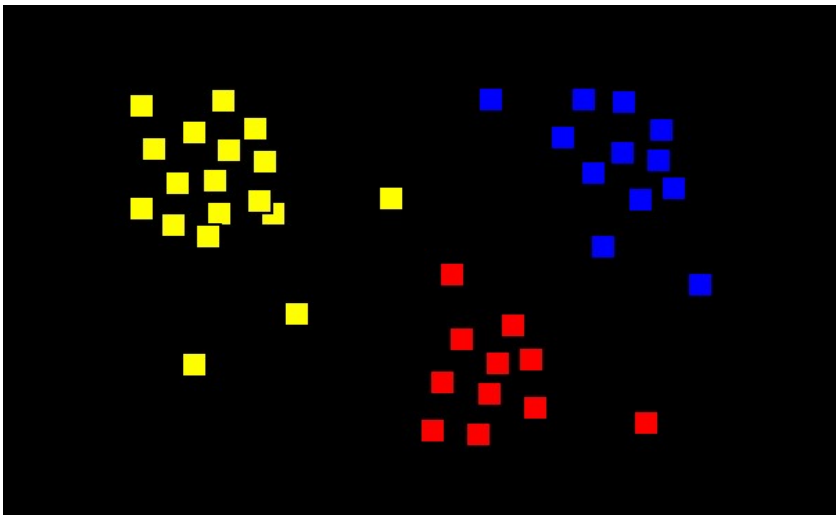


Рисунок 1.2 - Результат кластерного аналізу

Кластерний аналіз виконує наступні основні завдання:

- Розробка типологій та класифікацій.
- Дослідження концептуальних схем для групування об'єктів.
- Породження гіпотез опираючись на основі результатів дослідження.

Перевірка даних гіпотез та дослідження для подальшого визначення, чи дійсно типи, які виділені тим чи інакшим методом, присутні в наявних даних.

Незалежно від предмету вивчення та застосування кластерного аналізу припускаються наступні етапи:

- Відібрання вибірки для «кластеризації». Під даним етапом мається на увазі, що є сенс кластеризувати лише кількісні дані;
- Встановлення змінних, згідно яких будуть оцінюватися об'єкти у вибірці, тобто простору ознак;
- Вичислення значень однієї чи іншої міри подібності (чи відмінності) між обраними об'єктами;
- Застосування даного методу кластеризації для створення груп схожих між собою об'єктів;
- Перевірка правдивості результатів кластерного рішення.

Можна зустріти опис 2-х важливих фундаментальних вимог, що пред'являються до даних – це однорідність та повнота. Однорідність вимагає, щоб всі кластеризаційні суті були однією природи, описувалися подібним набором характеристик. Якщо «кластерному аналізу» передує так званий - факторний аналіз, тоді вибірка не потребує так званого «ремонту» - викладені вимоги виконуються цілком і повністю автоматично самою функцією факторного моделювання (є ще одна велика перевага – це стандартизація без негативних наслідків щодо вибірки; а за умов якщо її проводити виключно для кластеризації, вона може спричиняти за собою зменшення чіткості розділення груп). В другому випадку вибірку необхідно коригувати[4].

У медицині кластеризація має чимало застосувань в різних областях. До прикладу, в біоінформатиці за допомогою кластеризації аналізуються складні мережі взаємодії генів, що складаються часом з сотень або деколи - тисяч досліджуваних одиниць. Кластерний аналіз дозволяє виділяти деякі підмережі, вузькі місця, концентратори та інші схожі властивості досліджуваної системи, яка дає змогу в кінцевому результаті дізнатися внесок кожного гена у формування досліджуваного феномену.

В сфері екології дуже широко застосовується для виділення просторово-однорідних груп простих організмів, спільнот і т.п. Рідше методи кластерного аналізу застосовуються для дослідження спільнот в часі. Гетерогенність структури спільнот призводить до появи нетривіальних методів кластеризації (до прикладу, метод Чекановського).

В загальному вартує відзначити той факт, що історично склалося так, що в якості заходів близькості в біології частіше використовуються міри схожості, а не заходи відмінності (відстані).

Таким чином було досліджено та проаналізовано метод кластерного аналізу, та його застосування у процесі боротьби з епідемією поширення COVID-19.

1.2 Аналітичний огляд результатів обробки джерел інформації

Перед тим, як розпочати розробку моделі динаміки, необхідно обрати джерела статистичної інформації. Джерелами статистичної інформації можуть слугувати інтернет – ресурси, мас – медіа, соціальні мережі, та спеціалізовані офіційні сайти Статистики України.

Одним із таких, є інтернет ресурс https://covid19.gov.ua/.

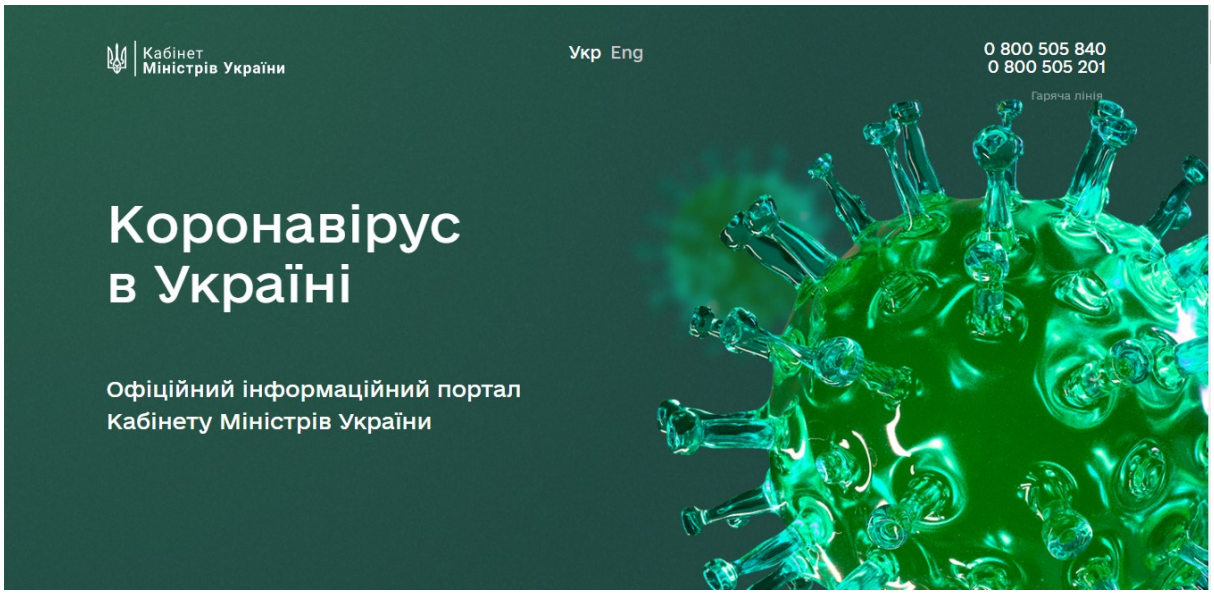


Рисунок 1.3 – Інтерфейс офіційного сайту статистики захворюваності CoVID -19 в Україні

Даний інтернет – ресурс є досить зручним джерелом статистичної інформації. Окремо можна виділити модуль «Аналітичні панелі та відкриті дані»

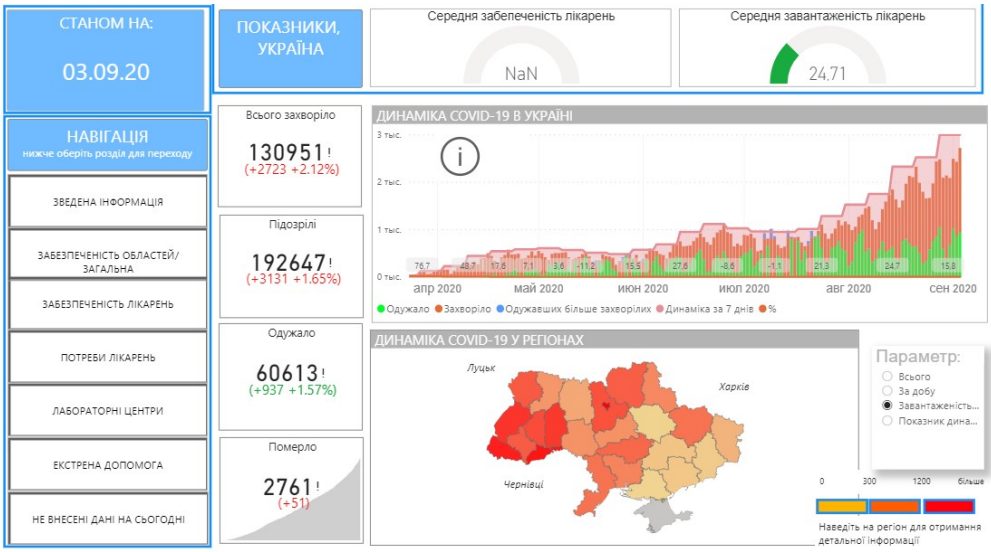


Рисунок 1.4 – Інтерфейс вікна Аналітичні панелі та відкриті дані

Даний інтернет ресурс надає зручний та швидкий доступ до перегляду та збору статистичних даних. У моделі Аналітичної панелі реалізована

інтерактивна карта «Червоних» зон, дані кількості захворівших та одужавших пацієнтів, відсотковий приріст захворюваності та одужуваності.

Також, на сайті реалізована можливість перегляду даних у різні часові проміжки.

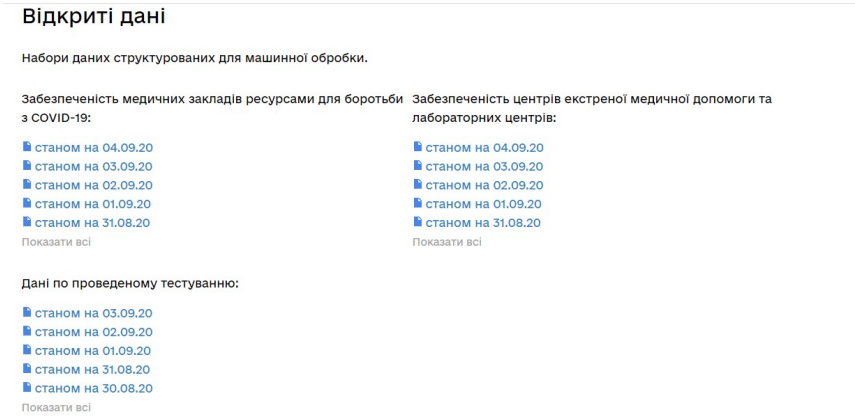


Рисунок 1.5 – Архів відкритих даних

Ще одним потужнім джерелом статистичної інформації є інтернет – ресурс https://news.google.com/, розділ covid 19

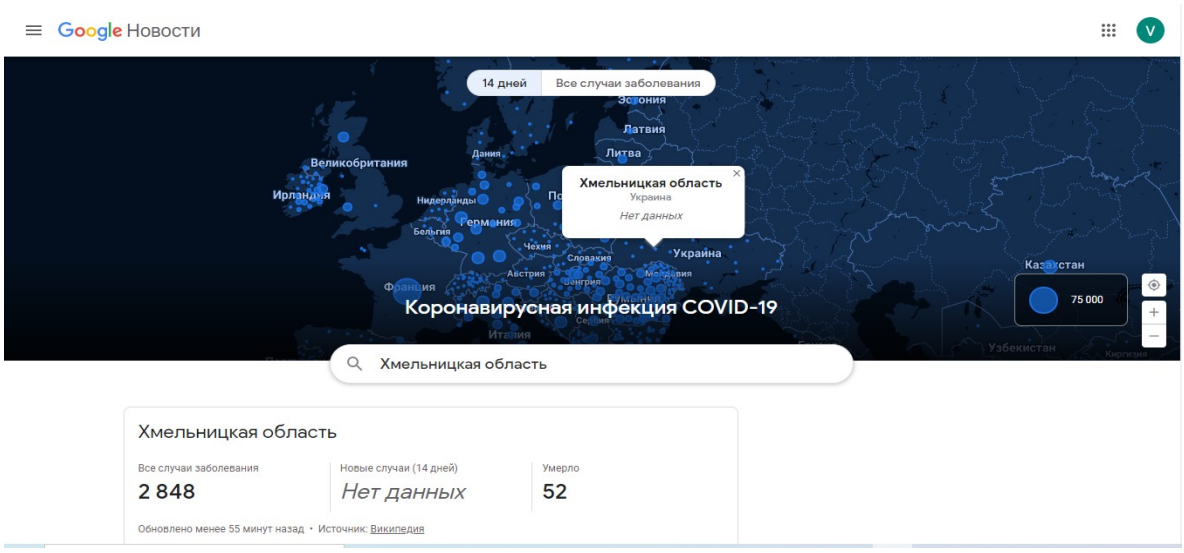


Рисунок 1.6 – Інтерфейс ресурсу google news.

Перевагами даного ресурсу є:

– інтерактивна карта, розділена по регіонах;
– швидкий доступ до статистичних даних;

– зручний інтерфейс.

Ресурс GIS-Hub COVID-19 містить багато інформації, карти, додатки та посилання на ресурси по боротьбі з коронавірусом в Україні та світі.

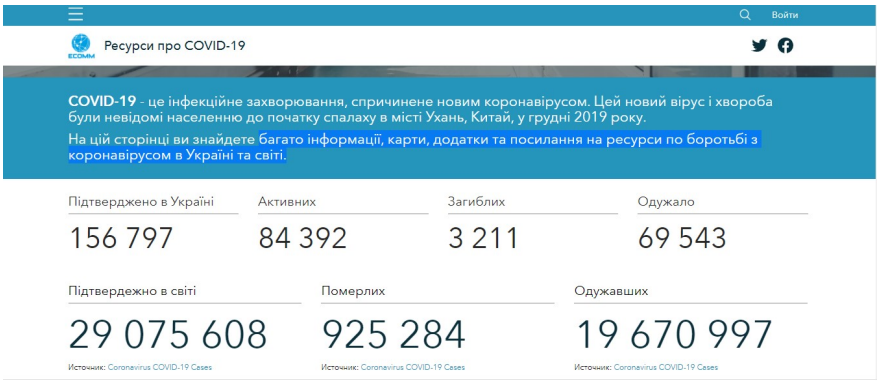


Рисунок 1.7 – Ресурс GIS-Hub COVID-19

Задля досягнення мінімальної похибки при дослідженні динаміки приросту захворюваності, варто брати інформацію з різноманітних джерел. Тому, в процесі розробці динамічної моделі можливе допущення вибору ресурсів, безпосередньо в процесі дослідження та побудови моделі.

Також варто відмітити офіційні джерела статистичної інформації:

COVID-19 Coronavirus latest data visualized – це візуалізація, яка заснована на фактах та даних, яка постійно оновлюється.

WorldHealthOrganization (WHO) – найактуальніша інформація про перебіг пандемії ковіду на всій території планети, відповіді на найпопулярніші запитання, руйнувння міфів про Covid-19, щоденні брифінги ВООЗ, статистика і поради.

Міністерство охорони здоров'я України – це офіційна сторінка у мережі Facebook з актуальною та достовірною інформацію щодо коронавірусу взятих із офіційних джерел МОЗ.

Інформаційний портал Кабінету Міністрів України – портал з відповідями на основні питань про Covid: шляхи передачі, як діагностується, які особи є в групі ризику, як вберегти себе, та яка ситуація в країні.

Центр громадського здоров’я – офіційна сторінка у мережі Facebook. Організація відповідає за збереження і зміцнення різних верств здоров’я

населення, «соціально-гігієнічний» моніторинг поширення захворювань, епідеміологічний нагляд та біологічну безпеку.

Novelle Coronavirusa la Information Center – це безкоштовний доступ до різних медичних досліджень «Elsevier» з приводу поширення нового коронавірусу COVID-19.

COVID-19 Clarivate Analytics - надає доступ до провідних світових досліджень та останніх новин, пов'язаних із новим коронавірусом COVID-19.

Coronavirus disease (COVID-19) – UNICEF – надає поточну інформацію, пояснення для батьків, вчителів, що ґрунтуються на останніх наукових доказах експертів у галузі охорони здоров'я.

Коронавірус_інфо – телеграм-канал інформації коронавірусу в Україні, офіційно верифікований Міністерством охорони здоров'я України. Тут досить оперативно інформують про поточну ситуацію та останні новини навколо поширення вірусу «COVID-19» в країні.

Novelle la Coronavirusa «COVID-19» Data Repository by CSSE – це сховище даних для нової візуальної панелі моніторингу поширення пандемії 2019-2021 років, керованої Університетським центром системних досліджень і інженерії імені Джона Хопкінса (JHU CSSE).

Таким чином було обрано ресурси аналітичних даних, згідно яких можливий збір статистики, та подальше будування динамічної моделі, було проаналізовано та обрано джерела статистичної інформації, згідно яких буде проводитися подальший кластерний аналіз.

1.3 Розроблення гіпотези дослідження

За останні 50 років у світі виникла чимала кількість небезпечних інфекцій – віруси чи бактерії з плином часу змінювали свої властивості в сторону набуття вірусолентності, епідеміологічного розповсюдження, якого раніше не спостерігалося. (За визначенням ВООЗ, емерджентні інфекції –

хвороби та збудники, які виникають досить раптово і цим зумовлюють дещо - надзвичайні епідеміологічні ситуації.)

Коронавіруси відомі ще з 1960-х років, 4 з яких досить швидко включилися в циркуляцію. Під час деяких сезонних інфекцій вони беруть участь у захворюваннях, що називають ГРВІ (є такі дані, що близько 15% всіх ГРВІ спричинені саме коронавірусом).

У 2005 році у Китаї стався сплеск SARS (важкий гострий респіраторний синдром, під час якого захворіли около 10 тисяч осіб, біля 12% з захворювань мали летальні випадки), а вже у 2013 році — MERS (близькосхідний респіраторний синдром, около 2,5 тис. випадків захворювань, з доволі високим показником смертності). SARS і MERS стали так би мовити - підготовкою до того, що ми маємо на сьогоднішній день. Нинішній коронавірус SARS-CoV-2 є 7 у роді коронавірусів людини.

Багато перших хворих мали відношення до ринку Ухань, на якому продаються морепродукти, а також птиці, змії, кажани і сільськогосподарські тварини. Оскільки в ході розшифрування генному коронавіруса в ньому були виявлені складові частини, близькі коронавірусу кажанів і панголінів, то передбачалося, що на просторі Уханьского ринку морепродуктів сталася зустріч кажанів і панголінів, яка створила умови для рекомбінації коронавірусів цих тварин. Вперше версія з'явилася в заяві міської влади Ухань 31 грудня 2019 року, на наступний день після того, як за з'ясування походження нового вірусу взялося керівництво Уханьского інституту вірусології. Згідно муніципальним звітів, кажани ніколи не продавалися на місцевому ринку [9], а панголіни занесені в Червону книгу. Посол КНР в Росії стверджує, що коронавірус був завезений на ринок якимось інфікованою людиною, після чого спалахнула епідемія.

30 грудня 2019 року влада Ухань направили в Уханьський інститут вірусології запит - провести перевірку на предмет того, чи не було в УІВ неправильного поводження з експериментальними матеріалами. 31 грудня Ши Чженлі почала перевірку своєї лабораторії на предмет можливого витоку

з неї коронавіруса. 6 лютого професор бота Сяо опублікував статтю, в якій висловив версію про можливе походження нового коронавіруса в УІВ [9]. 7 лютого Ши Чженлі заявила, що коронавірус не має відношення до її лабораторії.

Пандемія COVID-19 поставила перед наукою нові завдання, вирішення яких, з очевидних причин (короткий часовий інтервал, недостатність вихідних відомостей), на сьогодні не представляється можливим. Дані про походження нової модифікації коронавіруса поки обмежені і продовжують поповнюватися. Велика кількість питань виникає з приводу особливостей його існування у кровоносній руслі людини, молекулярних механізмів взаємодії з ендотелієм кровоносних судин, з клітинами крові, з гемоглобіном. знання в цій галузі – скоріше окремі гіпотези, поки практично позбавлені системних наукових обґрунтувань.

Коронавірус SARS-CoV-2. Новий коронавірус SARS-CoV-2 – один з сімейства Coronaviridae, що включає на січень 2020 року більше 40 видів РНК-вірусів, які вражають людину і тварин. В лютому 2020 р китайські вчені виділили штам коронавіруса і вперше продемонстрували зображення SARS-CoV-2 за допомогою електронного мікроскопа. Знімки отримані в національному сховище патогенних мікроорганізмів, були опубліковані на сайті відомства, представлені ВООЗ і широко ілюструються у відкритій пресі. В літературі з'явилися й інші зображення вірусу, отримані на електронному мікроскопі Білок SARS-CoV-2 був досить ефективний для зв'язування людських клітин, в результаті чого вчені зробили висновок про те, що це результат «природного відбору», а не продукт так званої - генної інженерії. Зазначається, що нині вчені розглядають дві ймовірні теорії саме природного походження самого вірусу. Згідно з першим сценарієм, вірус розвинувся до свого поточного патогенного стану шляхом природного переходу «від тварини до людини».

Саме так у минулому й виникали попередні спалахи поширення коронавірусу, саме коли особи заражалися ним після контакту з Циветами (SARS) і Верблюдами (MERS).

Згідно з іншим сценарієм, прабатько SARS-CoV-2 потрапив до людського організму, придбавши генномні характеристики, шляхом адаптації цитонейрозу під час передачі від людини людині.

Як повідомляв Укрінформ, станом на ранок четверга загальна кількість випадків зараження коронавірусною інфекцією у світі становить 219 367, число померлих – 8 970, а тих, що одужали – 85 749.

Новенький коронавірус був досить швидко ідентифікований після того, як виник перший всплеск у Китаї. Вже 7 січня його було виділено на культурі клітин, а вже 12-го січня (за 6 днів) був секвенований повний геном цього вірусу. В той час як для розшифровки повного генному вірусу SARS (у 2002–2004 роках) знадобилося півтора року

На початку червня виклик, кинутий британським нейробіологом Карлом Фрістоном щодо пандемії COVID-19, реально приголомшив світ.

У разі підтвердження гіпотези, на нас чекають нові відкриття. Зокрема, вони можуть:

− позбавити до 85% населення України та інших країн небезпеки захворюванності при зараженні новим вірусом SARS-CoV-2;

− нівелювати таке твердження, що відмінність у рази між статистичними показниками захворюваності на коронавірус у різних країнах та на різних континентах є наслідком «конкретних дій урядів», оскільки, за гіпотезою, це пояснюється особливостями груп населення, що є менш уразливими для COVID-19.

Згідно з однією з гіпотез, вірус пішов із лабораторії в місті Ухань, де стався початковий спалах захворювання з «нульовим» пацієнтом, є лабораторія з найвищим рівнем біологічної безпеки «BSL-4», в якій вивчили й коронавірус. Проте у журналі «The Lancet» була опублікована заява, яка підписана багатьма вченими, які в свою чергу заперечують цю теорію. Вони

схиляються до того, що вірус має цілком «природне» походження. Поки що на це питання чіткої відповіді не було дано.

Наразі є повідомлення, що вірус має 2 лінії. У природі так і є. Оскільки чим більшу к-сть «хазяїнів» вірус проходить, тим більше він мутує. У SARS-CoV-2 відбувається около двох мутацій на тиждень. Протягом трьох місяців циркуляції він змінюється, проте ці мутації ще не настільки виражені, аби серйозно на щось вплинути.

Отримати вакцину – складно, хоча в теорії й легко. Теперішні методи молекулярної вірусології дають змогу дуже швидко «секвенувати» віруси, визначитися, яка з частин відповідає за той чи інший синтез білків. В цьому випадку захисним антигеном є «білок S», саме для нього й треба виробляти вакцину. На даний момент в багатьох країнах цим займаються.

Наприклад, одна із вакцин вже проходить першу стадію лабораторного дослідження в США на 50 добровольцях. Хочу зазначити, що лише 1 стадія на волонтерах триватиме 13 около місяців. Щоб стати «комерційним препаратом», вакцина має пройти ще 2 стадії. Ми ще не знаємо, наскільки буде ефективною дана вакцина. Наприклад, коли починалася епідемія «Еболи» то у 2014 році, було 6 вакцин, які претендували на комерційний препарат. І лише у грудні минулого року почали реєструвати перші дієві ліки проти захворювання. Отож бо, нам треба лише сподіватися на профілактичні заходи та колективний імунітет. Якщо таки буде вакцина – це, звісно, чудово, але поки ми маємо виходити із ситуації використовуючи свої сили. Скоро, на привеликий жаль, не вийде вакцини. COVID-19 зазвичай починається з ураження «епітелію» на верхніх дихальних шляхах з подальшим розповсюдженням до альвеол та легенів. Поточні дані свідчать про те, що важчий перебіг COVID-19 наявний у пацієнтів з послабленою імунною відповіддю й зниженою здатністю протидіяти поширенню вірусу. Так, у 79% пацієнтів при інфікуванні COVID-19 спостерігаються «легкі» симптоми захворювання, проте у ряду пацієнтів відзначається «важкий» перебіг захворювання, який потребує госпіталізації, інтенсивної терапії, а в деяких

випадках може спричинити навіть летальний кінець. Поточні дані свідчать про те, що підвищена схильність до розвитку ускладнень, важчого протікання інфекції і ризику «летального кінця» спостерігається в пацієнтів із коморібідною патологією, особливо у осіб «похилого» віку чи пацієнтів з «артеріальною гіпертензією», серцево-судинними захворюваннями (СЦЗ), діабетом або з хронічними захворюваннями легень та ВІЛ. Враховуючи цю поширеність нової коронавірусної інфекції, на сьогодні всі міжнародні медичні спільноти знаходяться у стадії у пошуках ефективного лікування COVID-19.

Вірус «SARS-CoV-2» за багатьма х-ми схожий до віруса SARS-CoV-1, що спричинив пандемію у 2002–2003 рр., включно зі схожістю генному у понад 82%, смертністю пацієнтів внаслідок гострого респіраторного дистрес-синдрому а також потраплянням до клітин шляхом зв'язування із прецептором ангіотензин-перетворювального ферменту другого типу. Окрім того, SARS-CoV-2 у порівнянні з SARS-CoV-1 має більшу спорідненість до геному АПФ-2. Внаслідок подібності цих 2-х вірусів попередні дані щодо SARS-CoV-1 можуть сприяти розвитку гіпотез щодо можливих варіантів лікування COVID-19, викликаної SARS-CoV-2, включно з перепрофілюванням фармакологічних препаратів, схвалених задля клінічного використання.

Відомо, що «SARS-CoV-2» інфікує лише клітини людини шляхом зв'язування із прецептором «АПФ-2». Інфекція «SARS-CoV-2» як правило починається у верхніх дихальних шляхах, далі вірус поширюється на нижчі дихальні шляхи, де власне вражає епітеліальні клітини, особливо пневмоцити 2-го типу, які експресують АПФ-2. В додаток до ураження легень «COVID-19» також може викликати ураження у серцево-судинної системи, включно з асоційованим міокардитом, і інших органів, котрі також експресують АПФ-2.

Враховуючи наявні дані про обидва віруси SARS, запропонований наступний патофізіологічний ланцюг ураження при COVID-19

Рисунок 1.8 - Схема ураження SARS-CoV-2.

Згідно отриманих даних, гіпотези подальшого розвитку поширення будуть наступними:

1. Пандемія ніколи не завершиться

Чотири інших штами коронавірусу постійно присутні у світовій популяції. Вони викликають простуду, хоча в окремих випадках навіть можуть призвести до пневмонії і смерті.

Адмеш Адаляля з Центру безпеки охорони здоров'я імені Д. Гопкінса стверджує, що точно таким самим буде і новий вірус.

2. Спільними зусиллями в сфері охорони здоров'я вірус буде знищено.

Вважається, що «Covid-19» схожий на раніше згаданий вірус SARS, який зафіксували ще у 2003 році. Він убив близько 7 сотень людей, а заразив – понад 8 тисяч. Саме тоді експерти та органи охорони здоров'я доклали чимало зусиль, щоб «відслідкувати», діагностувати та ізолювати хворих. Зрештою, хвороба зникла у 2005-му.

Якщо кількість людей, які мають вразливість до нового коронавірусу, впаде нижче заданого порогу, спалах може бути стримано. Проте поточна хвороба є більш заразною, аніж SARS, що робить більш складнішою боротьбу з вірусом.

3.Буде винайдена вакцина.

Таким чином, основними причинами ускладненого перебігу та смертності внаслідок захворювання COVID-19 є, по-перше, ураження легень вірусом з подальшим розвитком дихальної недостатності та ураження

інакгих систем організму, які супроводжуються цитокіновим штормом та, розвитком поліорганної недостатності.

РОЗДІЛ 2 МЕТОДОЛОГІЯ ДОСЛІДЖЕННЯ ТА АНАЛІЗ ОТРИМАНИХ ДАНИХ

2.1 Аналіз статистичних даних

Згідно зібраних статистичних даних, станом на 20 вересня в Україні зафіксовано 3 497 зафіксованих випадків коронавірусу COVID-19. Кількість активних хворих становить: 99 359.р [8]

Загалом в Україні 184 734 лабораторно підтверджені випадки COVID-19, із них 3705 летальних, 81 670 пацієнтів одужали. Проведено 2 102 377 тестувань методом ПЛР. За добу одужало 1 769 пацієнтів.

Наразі коронавірусна хвороба виявлена:

- Вінницька область –5 455 випадків;
- Волинська область –7 132 випадки;
- Дніпропетровська область – 4 394 випадки;
- Донецька область – 2 634 випадки;
- Житомирська область – 5 191 випадок;
- Закарпатська область – 9 343 випадки;
- Запорізька область – 3 149 випадків;
- Івано-Франківська область – 12 811 випадків;
- Кіровоградська область – 1 024 випадки;
- м. Київ – 20 335 випадків;
- Київська область – 8 148 випадків;
- Львівська область – 19 045 випадків;
- Луганська область – 814 випадків;
- Миколаївська область – 2 643 випадки;
- Одеська область – 12 130 випадків;
- Полтавська область – 1 630 випадків;
- Рівненська область – 11 459 випадків;
- Сумська область – 3 280 випадків;

- Тернопільська область – 12 069 випадків;
- Харківська область – 15 908 випадків;
- Херсонська область – 868 випадків;
- Хмельницька область – 5 049 випадків;
- Чернівецька область – 13 448 випадків;
- Черкаська область – 3 238 випадків;
- Чернігівська область – 3 537 випадків.

В середньому, за добу в Україні кількість нових зареєстрованих випадків захворювання виявляється у Києві (332), Харківській області (302), у Львівській області (204), в Одеській області (203), у Тернопільській області (201).

Станом на 27 вересня 2020 року, з початку епідемії одужало вже 69 543 українця (за останню добу - 1267), померло 3264 пацієнти (за добу - 53).

Найбільше хворих за добу з'явилося в Полтавській області (390), найменше – у Херсонській області (17). А лідером за загальною кількістю випадків COVID-19 зараз є Київ (17 714).

Протягом 27 числа в Україні провели 47 621 тест, із них: 26 898 - методом ПЛР і 20 723 - методом ІФА,

Згідно зібраних статистичних даних, за період часу, з дати виявлення першого захворювання по 27 вересня, було побудовано схематичну карту поширення захворюваності.

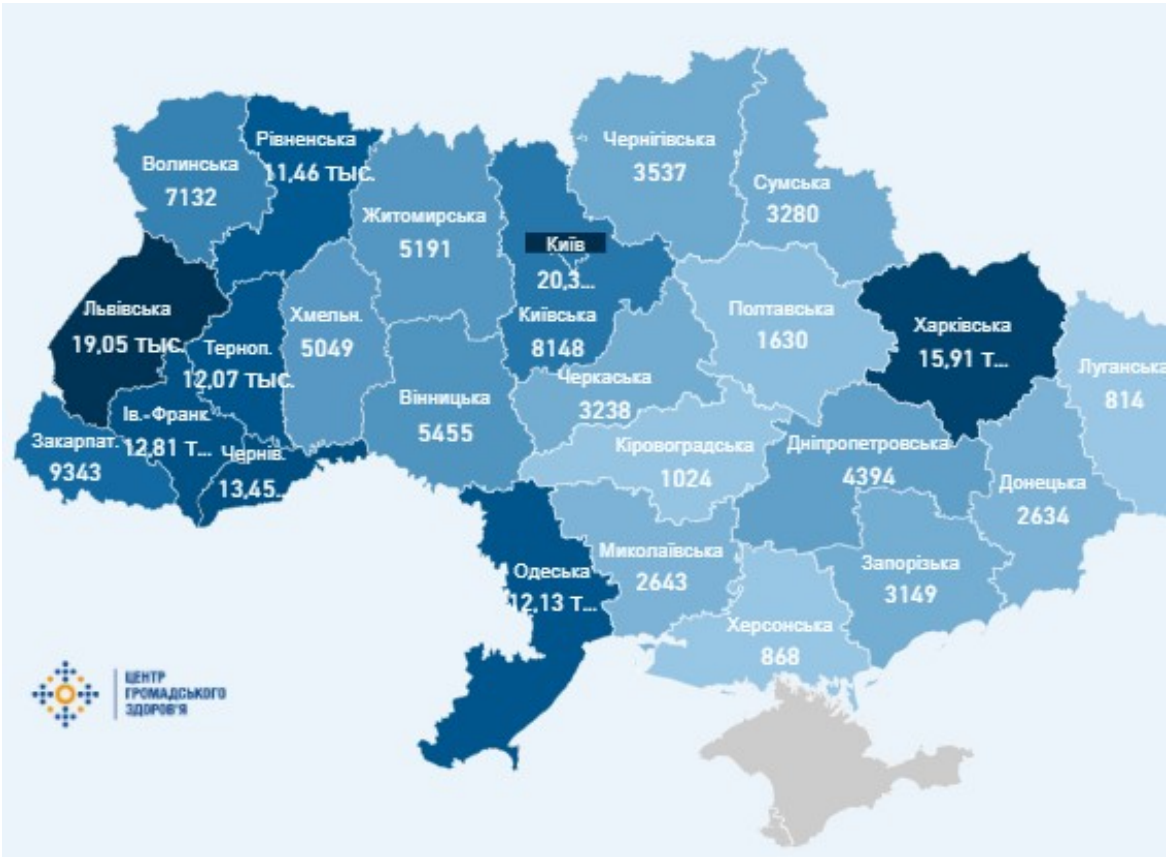


Рисунок 2.3 – карта поширення захворюваності

Поширенню SARS-CoV-2 молодими людьми може сприяти низка факторів. Згідно статистичних даних, та висновків медичних експертів допускається, що представники молодшого покоління вважають себе менш схильними до ризику: таким чином, вони мобільніші і мають більше шансів бути переносниками.

При цьому у молодих людей коронавірус виявляють частіше. До прикладу, аналіз даних, зібраних в американському місті Сіетлі, показав, що більша половина усіх нових випадків зараження ковідом припадає на людей у віці від 22 до 31 років. А згідно з останніми даними американського Центру з контролю та профілактики захворювань, майже 70 відсотків жителів Америки з позитивним діагнозом COVID-19 були молодші за 58 років.

У травні керівник CDC Роберт Редфілд заявив про те, що основні симптоми захворювання вірусом можуть не проявлятися у 25% заражених людей. В свою чергу дослідження, які були проведені у Гонконзі та

Великобританії, показали те - що частка «прихованих» носіїв серед переносників вірусу може сягати до 50% .

Допоки що вчені ще не дійшли консенсусу з приводу того, наскільки «безсимптомні носії» заразні та якою мірою сприяють поширенню SARS-CoV-2. Більшість вірусологів дотримується тієї думки, що ризик зараження від хворого з ознаками коронавірусу все ж дещо вищий, ніж від носія інфекції, у якого вони відсутні.

При цьому фахівці здатні вважати, що пріоритетними у боротьбі з поширенням вірусу як й раніше мають залишатися «профілактичні заходи» – зокрема, соц. дистанціювання та носіння захисних противірусних масок. Томас Ентоні – професор вірусології американського інституту досліджень пропонує: «У деяких регіонах, де недавно були зафіксовані нові спалахи коронавірусу – до прикладу, в Австралії, - влада лише зовсім недавно зобов'язала жителів носити маски у громадських місцях».

В Німеччині, такі правила запровадили ще у кінці квітня. Відповідно до одного із досліджень, носіння масок знижує шанс заразитися SARS-CoV-2 на 42 відсотки.

Поширення захворюваності у світі в цілому відповідає загальній сезонній динаміці грипоподібних захворювань (за винятком Східної Європи)

Рисунок 2.4 – Графік динаміки приросту захворюванності

Регіоном зі значним поширенням «COVID-19» вважають той, у якому є хоча б 1 із наступних ознак:

− навантаженість ліжок в медичних закладах ОЗ, визначених для госпіталізації пацієнтів в яких було підтвердженио та з підозрою на інфікування COVID-19, становить більш як 60 % протягом 5 днів підряд без урахування місць-ліжок, які було відведено для лікування у дитячих відділеннях;

− середня к-сть тестів методом «полімеразної ланцюгової реакції (ПЛР)» та імуноферментного аналізу становить менш ніж 41 тестування на 100 тис. населення протягом останнього тижня;

− показники захворюваності за 14 днів на 100 тис. населення перевищують мінімальний рівень захворюваності.

− базовий рівень захворюваності на вірус становить 45 нових випадків на 100 тис. населення за 2 тижні.

Державна інспекція з питань техногенно-екологічної (ТЕБ) безпеки та НС раз на тиждень приймає рішення щодо встановлення на території регіонів або окремих адміністративно-територіальних одиниць (АТО) регіону рівня епідемічної небезпеки: "жовтий", "помаранчевий" і "червоний".

Рішення щодо «послаблення "червоного", "помаранчевого" та "жовтого" рівня небезпеки не можна переглянути раніше ніж через два тижні з дня встановлення такого рівня небезпеки».

Якщо для окремих АТО регіону, де мешкає около 79% та більше всього населення регіону, визначено «помаранчевий чи червоний рівень небезпеки», помаранчевий рівень визначають для всього визначеного регіону загалом.

Червоний рівень небезпеки на території окремих АТО регіону може бути запроваджено згідно рішення Держ. комісії з питань техногенно-екологічної безпеки і НС у разі:

− навантаженості більше 73% ліжок у мед.закладах охорони здоров’я регіону, які визначені керівником робіт з ліквідації наслідків НС

медико-біологічного характеру державного рівню, пов’язаної з поширенням на території України коронавірусу, спричиненої вірусом «SARS-CoV-2», для подальшої госпіталізації пацієнтів із підтвердженим діагнозом COVID-19 протягом п’яти днів;

– перевищення середнього рівня по країні захворюваності на «COVID-19» (випадків на 110 000 населення за два тижні) більше ніж у пять разів.

З ціллю визначення епідеміологічної доцільності та запровадження окремих протиепідемічних обмежень які передбачено пунктом 15 цієї постанови, на відповідному засідання Державної комісії щодо питань техногенно-екологічної безпеки та НС запрошують представників регіональних комісій із питань ТЕБ та НС, на територіях у яких запроваджують протиепідемічні обмеження червоного рівня епідемічної небезпеки.

Ці висновки можуть мати свої наслідки для пандемії коронавірусу. Адже досі контактною особою вважається та людина, яка контактувала з хворим за два дні або менше до того, коли у неї проявилися перші симптоми. Тепер, скоріш за все, необхідно буде відстежувати усі контакти хворого за набагато більш-триваліший період, що буде неабияким викликом як для осіб, щодо яких це стосується, так і для усіх сервісів для відстеження ланцюгів зараження вірусом.

2.2 Аналіз чинників та їх вплив на динаміку поширення COVID-19

Почати слід з того, що йдеться про кількісне, статистичне дослідження, де фокус – люди.

Як і більшість досліджень такого характеру, воно вимагає вибірки, що є одним із центральних понять статистики і одне із ключових з методологічної точки зору, адже якщо вибірка нерепрезентативна і упереджена, то результат,

скоріше за все, буде так само упередженим або ж таким, що не можна генералізувати на все населення.

Повітряно-крапельний шлях є основним для передачі вірусу, однак, за словами авторів роботи, точних даних про вплив погодних умов на життєздатність інфікованих частинок не було.

Проведене моделювання та гідродинамічні експерименти показали, що динаміка випаровування містяться в повітрі, що видихається найдрібніших крапель слини є ключовим фактором поширення вірусу.

При високій температурі і низькій відносній вологості краплі випаровуються стрімко, що істотно знижує життєздатність патогенів. Однак, при високій вологості радіус переміщення хмари повітря, що видихається і концентрація в ньому вірусних часток зберігаються більш тривалий час.

Швидкість вітру також є дуже важливим фактором для визначення правил соціального дистанціювання, особливо під час осінньо-зимового періоду.

Отримані результати, допомагають пояснити посилення пандемії з липня-місяця в різних густозаселених місцевостях по всьому світу, у яких спостерігалися високі рівні відносної вологості. Крім того, вони служать попередженням другого спалаху захворюваності в період, коли погодні умови підвищать виживаність і передачу вірусу.

Одна із відмінніших рис епідемії полягає в тому, що вірус «Covid-19» є більш стійким у навколишньому середовищі, має різні механізми передання такі як: «повітряно-крапельний», контактний, аліментарний – у нього доволі тривалий інкубаційний період. Якщо щодо для вірусу грипу і ГРВІ це 3-5 днів, то для цього штаму коронавірусу – до 14 днів. Іноді й інкубаційний період може тривати до 24-х днів. У зв’язку з цією ситуацією люди, що переносять легкі та стерті форми захворювання, можуть представляти серйозну епідемічну небезпеку, оскільки хвороба протікає доволі непомітно і людина може продовжувати відвідувати громадські місця і заражати інших.

Також, згідно досліджень дослідників з Бейрута виявилося, що виживання вірусів у повітрі і на поверхнях, а також сприйнятливість до інфекцій і схильність людей збиратися в приміщеннях, залежать від сезонних змін температури і вологості. Порівняно з грипом та іншими респіраторними захворюваннями новий коронавірус має більш високу швидкість передачі через відсутність у більшості людей імунітету. Саме тому чинники, що впливають на сезонність, поки що не можуть зупинити поширення COVID-19 у літні місяці.

Згідно з даними на початок березня кількість випадків щодня перевищувало число випадків в попередній день в 1,12-1,25 разів. Зміна кількості заражених день у день складається з трьох цифр: кількість заражених в певний день (N), середня кількість осіб, з яким заражений може контактувати в певний день (E), і ймовірність кожного контакту стати новим випадком зараження (p). Відповідно, за даними на 6 березня, в середньому кожні 16 днів кількість заражених збільшувалася в 10 раз. Але просто провести висхідний тренд недостатньо: в який момент зростання кривої повинен зупинитися. І важливу роль починають грати змінні E і p – вони повинні знижуватися, щоб зупинити експоненціальне зростання.

Цей вірус передається повітряно-крапельним шляхом (ПКШ) через «вдихання дрібних часток», розпорошених в повітрі при кашлі, чханні або розмові. Частки з вірусом можуть потрапляти на поверхні і предмети, а потім інфікувати доторкується до них людини через наступні дотики до очей, носа або рота. Вірус може залишатися життєздатним протягом декількох годин, потрапляючи на поверхні предметів. На сталевих поверхнях і на пластиці він може зберігатися до 2-3 днів [40]. Дослідження з сильним розпиленням показало, що вірус міг би перебувати в повітрі до декількох годин, однак ВООЗ уточнює, що в природних і медичних умовах розпорошення відбувається іншим способом, а про передачу вірусу по повітрю поки не повідомлялося. За даними Китайського центру з контролю і профілактиці захворювань життєздатний вірус був виявлений у фекаліях хворих COVID-

19, що означає можливість передачі інфекції, наприклад, через контаміновані руки, їжу і воду, однак даний механізм передачі не є основним в випадку COVID-19. Є також повідомлення про те, що вірус виявлявся в крові і слині.

COVID-19 поширюється від людини до людини переважно дихальними шляхами після того, як заражена людина кашляє, чхає, співає, розмовляє або дихає. Нова інфекція відбувається, коли вірусосодержащіе частинки, видихані інфікованою людиною, або крапель дихання, або аерозолі, потрапляють у рот, ніс або очі інших людей, які тісно контактують із зараженою людиною. Під час передачі від людини до людини вважається, що в середньому 1000 інфекційних віріонів SARS-CoV-2 ініціюють нову інфекцію. Чим ближче люди взаємодіють і чим довше вони взаємодіють, тим більша ймовірність передачі COVID-19. Ближчі відстані можуть включати більші краплі (які падають на землю) та аерозолі, тоді як більші відстані включають лише аерозолі. Більші краплі також можуть випаровуватися в аерозолях (відомих як ядра крапель). Відносна важливість більших крапель та аерозолів не зрозуміла станом на листопад 2020 року, проте, як відомо, вірус не передається між приміщеннями на великі відстані, наприклад, через повітроводи. Повітряно-крапельне передавання може особливо відбуватися в приміщенні, в місцях високого ризику, таких як ресторани, хори, тренажерні зали, нічні клуби, офіси та релігійні місця, часто коли вони переповнені чи менш провітрюються. Це також відбувається в закладах охорони здоров'я, часто коли пацієнти, які генерують аерозолі, проводяться на пацієнтах із COVID-19. Соціальне дистанціювання та носіння тканинних масок для обличчя, хірургічних масок, респіраторів та інших покривів для обличчя - це засоби передачі крапель. Передача може бути зменшена в приміщенні з доглянутими системами опалення та вентиляції, щоб підтримувати хорошу циркуляцію повітря та збільшувати використання зовнішнього повітря. Кількість людей, як правило, заражених однією зараженою людиною, варіюється; станом на вересень 2020 року було підраховано, що одна заражена людина в середньому заразить від двох до трьох інших людей. Це

інфекційніше, ніж грип, але менше, ніж кір . Він часто поширюється кластерами, де зараження можна простежити за індексом або географічним розташуванням. Існує головна роль "надпоширених подій", коли багато людей заражаються однією людиною.

В одному із досліджень повідомляється про випадок захворювання всередині сім'ї, де у 2-х членів сім'ї були відсутні симптоми та аномалії на рентгенівських знімках, але проби слизу з верхніх дихальних шляхів все ж показували наявність вірусу. Таким чином, можливі безсимптомні випадки інфекції. «Хоча відомий випадок передачі інфекції при її безсимптомному перебігу піддався критиці, стає все більше свідчень можливої передачі інфекції від безсимптомних носіїв».

Станом на кінець лютого 2020 роки не існує доказів можливості розвитку внутрішньоутробної інфекції або будь-яких ускладнень після неї у новонароджених, якщо у матері виявлено пневмонію на третьому триместрі вагітності. Проте вибірки у поточних досліджень вкрай маленькі, а Національна комісія з охорони здоров'я Китаю дала рекомендації вести моніторинг вагітних в тому числі і після одужання, а також ізолювати дитину від хворої матері після народження як мінімум на 14 днів

М'яка ізоляція виграє у карантину, а майже повна ізоляція – статистично найбільш надійний спосіб зупинити епідемію. Симуляції випадкові, кожне прочитання статті дасть трохи різні результати, але висновок залишиться колишнім. Лише тоді, коли багато людей матимуть стійкий імунітет після імунізації або одужання, вірус стане більш сприйнятливим до сезонних факторів. До цього необхідно дотримуватися суворих заходів щодо обмеження поширення коронавірусної інфекції.

Коронавірус ніколи не зникне, але сотень смертей щодня, як зараз, більше не буде, адже укорінені захворювання поводять дещо інакше, аніж нові.

Науковці впевнено заспокоюють, що по закінченню пандемії настане час, коли життя цілком нормалізується: «завдяки вакцині чи імунітету до

цього вірусу (наразі невідомо, чи його можна отримати, проте інші поширені коронавіруси імунітету не дають)».

Лесслері звертає увагу, що вже після 1 хвилі вразливих людей стане дещо менше, себто поменшає та потенційних жертв коронавірусу. А отже , й відсоток населення, який він атакуватиме.

Проте ця перша хвиля може бути довгою та болючою. Вчений Майкл Де Санта припускає, що лише в Америці вона призведе до 1,5 мільйона госпіталізацій та 150 тисяч смертей. Це у 5-11 разів більше, ніж річні наслідки грипу.

У Китаї передача йде в основному в колі сім'ї, внутрішньолікарняна передача в даній країні для інфекції не характерна.

Є повідомлення про передачу інфекції від людини домашнім кішкам, тиграм і левам. Експериментально з'ясовано, що вірус може легко передаватися між домашніми кішками. Можливість передачі від кішок до людини вимагає подальших досліджень.

Імовірно вірус ефективніше передається в сухих і холодних умовах, а також в тропічних з високою абсолютною вологістю. Поки є лише непрямі свідчення на користь зимової сезонності в північній півкулі. Однак аналіз кореляційних зв'язків між метеорологічними параметрами і швидкістю поширення інфекції в китайських містах не виявив взаємозв'язку швидкості поширення з температурою навколишнього середовища.

Патогенез

Вірус потрапляє в клітину приєднанням білка пепломера до рецептора – ангіотензинперетворюючого ферменту 2 клітини. Цим же шляхом відбувалося проникнення в разі вірусу SARS-CoV, однак структурний 3D-аналіз пепломера на поверхні вірусу в разі SARS-CoV-2 передбачає максимально сильна взаємодія з рецептором [39]. Входу в клітку також сприяє попередня преактивація пепломера фурином, яка була відсутня у вірусу SARS-CoV. Після приєднання до рецептора вірус SARS-CoV-2 використовує рецептори клітини і ендосоми для проникнення. Допомагає

проникненню трансмембранної сировини протеза 2 (TMPRSS2). Після зараження вірус поширюється через слиз по дихальних шляхах, викликаючи великий викид цитокінів та імунну відповідь в організмі. При цьому може спостерігатися зниження кількості лімфоцитів в крові, зокрема Т-лімфоцитів. Деякі дослідження доводять, що на боротьбу з вірусом витрачається дуже велика кількість лімфоцитів. Зниження їх кількості також знижує захисні властивості імуносистеми і може призводити до загострення захворювання.

Високий рівень вірусовиділення в горлі спостерігається в перший тиждень з появи симптомів, досягаючи найбільшого рівня на 4-й день, що передбачає активну реплікацію вірусу в верхніх дихальних шляхах. Тривалість вірусовиділення після зникнення симптомів захворювання оцінюється в 8-20 днів. Однак виявлення РНК вірусу після одужання не означає наявності життєздатного вірусу.

Ускладнення

У більшості COVID-19 протікає в легкій або середній формі, але в деяких випадках COVID-19 викликає сильні запальні процеси, звані ЦИТОКІНОВИЙ штормом, який може привести до смертельної пневмонії і гострого респіраторного дистрес-синдрому. При цьому профілі цитокинового шторму можуть відрізнятися у різних пацієнтів. Зазвичай COVID-19 супроводжується синдромом вивільнення цитокінів, при якому спостерігається підвищений рівень інтерлейкіну-6 (IL-6), що корелює з дихальною недостатністю, гострим респіраторним дистрес-синдромом та ускладненнями. Підвищені рівні протизапальних цитокінів можуть також свідчити про розвиток вторинного гемофагоцитарного лімфогістіоцитозу.

Інші експерти, а саме – колишній директор із Центру контролю і профілактики захворювань США – Тревор Філіпс, стверджує, що «кількість постраждалих може сильно змінюватися. Найгіршим, але неправдоподібним варіантом можна назвати мільйон загиблих у Штатах».

За цей час вірусом можуть «інфікуватися близько 85% (55,1 млн) населення країни, а до 16% – потребуватимуть госпіталізації».

При цьому, навіть кажуть, що раптові спалахи на території усієї країни можуть сильно влучити по системі охорони здоров'я, а лікарні страждають через нестачу обладнання.

Більше того - вчені досить стурбовані майбутніми населення нерозвинутими державами. Зокрема Африки, де в перенаселених містах та глибинах континенту буде важко забезпечити так зване "соціальне дистанціювання" для утриманн епідемії.

На початку березня у ВООЗ обговорювали, що саме стримання повинно залишатись пріоритетом для всіх країн.

Це пов'язано з особливостями новітнього віруса. Важливий показник – так званий «послідовний інтервал», мається на увазі час, після якого інфекція в середині людини може бути потенційно – небезпечною для поширення. Для нового коронавірусу – термін становить близько 5-6 днів. Тоді як для грипу – всього 3.

Дані про тривалість і напруженості імунітету відносно вірусу SARS-CoV-2 в даний час відсутні [6], для визначення тривалості будуть потрібні довгострокові серологічні дослідження імунітету людей, які видужали [48]. Проти коронавірусів, відмінних від SARS-CoV-2, формується гуморальний імунітет, проте часто повідомляється про випадки повторного виникнення інфекції (реінфекції). Виділення РНК вірусу знижується з настанням одужання і може тривати деякий час - від днів до тижнів, однак це не означає наявність життєздатного вірусу. При клінічному одужанні спостерігається вироблення IgM - і IgG -антитіла, що означає розвиток імунітету проти інфекції. З'являються публікації про випадки реінфекції через тривалий період часу після першого інфікування, поки підтверджені випадки реінфекції є рідкісними.

Хоча SARS-CoV-2 має здатність обходу вродженого імунітету, передбачається, що велика кількість легких і асимптоматичних випадків пояснюється роботою адаптивного імунітету внаслідок раніше перенесених захворювань, викликаних тими, які циркулюють серед населення

коронавірусом застуди. У 40% -60% Не перехворіли COVID-19 осіб виявляються крос-реактивні CD4 + Т-клітини, які можуть забезпечувати частковий імунітет від COVID-19.

«Таким чином новий вірус поширюється повільніше ніж грип. І тому, за словами Г. Вітакера - експерта із інфекційних захворювань в Університеті Корнелла, уникнення будь-яких контактів під час пандемії – дієвий спосіб для утримання хвороби.».

Прогнози розвитку захворювання

Летальність і тяжкість захворювання пов'язані з віком пацієнтів та наявністю супутніх захворювань. Основною причиною летальних випадків є дихальна недостатність, що розвивається на тлі гострого респіраторного дистрес-синдрому.

Згідно з аналізом 44 672 підтверджених випадків в Китаї (із загального числа в 72 314 випадків за даними з 31 грудня 2019 року по 11 лютого 2020 року), летальність становить 2,3%. Серед загиблих більше літніх людей віком від 60 років і людей з хронічними хворобами. Серед критично хворих летальність становить 49% [34] [136]. Оскільки ситуація розвивається, показники можуть змінитися.

Рівень летальності може відрізнятися між країнами, в деяких країнах рівень летальності виявився вищим, ніж в Китаї. В цілому по світу за станом на 8 квітня він оцінюється приблизно в 5,85%. Летальність серед госпіталізованих варіюється від 4% до 11%. На відмінності між країнами можуть впливати різні чинники. Наприклад, висока летальність в Італії частково пояснюється великою кількістю населення похилого віку в країні.

У порівнянні з важким гострим респіраторним синдромом і близькосхідним респіраторним синдромом летальність у COVID-19 набагато нижче. Однак захворювання COVID-19 легше поширюється і вже відняло набагато більше життів.

Підсумкова летальність серед пацієнтів без супутніх захворювань в Китаї була набагато нижчою і становила 0,9%.

2.3 Методи проведення кластерного аналізу

Кластерний аналіз з'явився відносно недавно – у 1939 році. Його запропонував вчений К. Трііон. Дослівно термін "кластер" перекладається з англійської "cluster" означає гроно/згусток.

Особливо бурхливий розвиток кластерного аналізу відбувся приблино у 60-х роках. Передумовами цього були поява швидкісних ЕОМ та визнання класифікацій фундаментальним методом наукових досліджень.

Таким чином, суть кластеризації полягає у здійсненні «класифікації» схожих об'єктів за певними ознаками за допомогою дослідження численних обчислювальних процедур. В результаті цього утворюються так звані "кластери" або групки дуже схожих між собою об'єктів. Порівняно з іншими методами, цей підвид аналізу дає можливість класифікувати об'єкти не за однією якоюсь конкретною ознакою, а за декількома одночасно. Для цього вводять відповідні показники, які характеризують міру близькості за усіма класифікаційними параметрами.

Метою кластерного аналізу є пошук наявних структур, що виражається в утворенні груп, яку важко знайти під час візуального обстеженні чи за допомогою експертів.

Кожен із цих кроків грає значну роль у практичному здійсненні кластерного аналізу.

Основними завданнями кластеризації є:

– встановити між собою схожість об'єктів. Водночас його дія полягає також у привнесенні структури у досліджувані об'єкти. Це означає те, що методи кластеризації необхідні щоб виявити структури у розробці типології або класифікації досліджуваних об'єктів;

– дослідити і визначити прийняті концептуальні схеми групування об'єктів;

– висунення гіпотез на підставі результатів дослідження;

– перевірка гіпотез чи справді типи, які були виділені якимось певним чином і мають місце у наявних даних.

Кластерний аналіз потребує проведення наступних послідовних кроків:

1) процес вибірки об'єктів для проведення кластеризації;

2) визначення множини ознак, згідно яких будуть оцінюватися відібрані об'єкти;

3) оцінка міри схожості об'єктів кластеру;

4) застосування кластерного аналізу для створення груп подібних за ознаками об'єктів;

5) перевірка достовірності результатів кластерного аналізу.

Визначення множини ознак, що покладаються в основу оцінки об'єктів, в кластерному аналізі є 1 із чи не найважливіших завдань усього наукового дослідження. Метою даного кроку повинна полягати також у визначенні сукупності змінних ознак, яка найкраще відображається у понятті «схожості». «Ці ознаки мають вибиратися з урахуванням теоретичних положень, які покладені в основу класифікації, та мети дослідження».

«При визначенні міри подібності об'єктів кластерного аналізу використовуються 4 види коефіцієнтів: коефіцієнти кареляції, показники віддаленостей, коефіцієнти асоціативності і ймовірності, коефіцієнти подібності. Кожен із цих показників має свої переваги та недоліки, які попередньо потрібно враховувати. На практиці найбільшого розповсюдження у сфері соціальних та економічних наук здобули так звані коефіцієнти кореляції та віддаленостей.» - пояснює науковець Гордон Фрімен.

В результаті аналізу сукупність вхідних даних створюються так звані «однорідні групи», в такий спосіб, що об'єкти всередині них подібні між собою за деякими критеріями, а об'єкти із різних груп відрізняються 1і від одного.

Кластеризація може здійснюватися 2 основними способами, зокрема за допомогою так званих «ієрархічних чи ітераційних методів».

Ієрархічні процедури – це послідовні дії формування кластерів різного рангу, підпорядкованих за чітко встановленою послідовністю «ієрархією». Найчастіше ієрархічні процедури здійснюються так званим шляхом агломеративних (об'єднувальних) дій. Ці методи передбачають наступні операції:

– послідовне об'єднання схожих об'єктів із утворенням «матриці подібності» об'єктів;

– побудова дендраграми (деревоподібна діаграма), яка відображає послідовне об'єднання об'єктів у окремі кластери;

– формування з досліджуваної вибірки окремих кластерів на початковому етапі аналізу та об'єднання цих об'єктів в одну велику групу на кінцевому етапі аналізу.

Ітераційні процедури полягають в утворенні із первинних даних однорангових (1 рангу) ієрархічно не підвпорядкованих поміж собою кластерів.

Першим із найбільш поширених способів проведення ітераційних процедур ось уже понад 40 років виступає метод к-середніх (розроблений у 1969 р. Джоном МакКуіном). Його застосування потребує здійснення наступних кроків:

– розділення вихідних даних досліджуваної сукупності на деяку задану к-сть кластерів;

– обрахування багатовимірних середніх «центрів тяжіння» у виділених кластерів;

– розрахунок «Евкліптової» відстані кожної одиниці сукупності до наперед визначених центрів тяжіння кластерів та побудова матриці відстаней, кора ґрунтується на метриках відстаней. Використовують різні метрики відстаней, наприклад: Евклідова відстань, «Мінковського, Махалонобіса» «Манхаттенська, Чебишева», тощо;

– визначення нових центів тяжіння і нових кластерів.

Найвідомішим та широко застосовуваними методами формування кластерів є:

– метод Варда.

– одиничного зв'язку;

– середнього зв'язку;

– повного зв'язку;

Метод одиничного зв'язку (або ж метод близького сусіда) передбачає приєднання одиниці сукупності до певного кластера, якщо вона близька (знаходиться на 1 рівні схожості) хоча б до одного представника цього кластеру.

Метод «повного зв'язку» (дальнього сусіда) вимагає певного рівня подібності об'єкта (не менше заданого граничного рівня), який передбачається включити до кластеру, із будь-яким іншим.

Метод «середнього зв'язку» ґрунтується на використанні так званої «середньої відстані» між кандидатом на включення у кластер та представниками наявного кластера.

Згідно з методом Варда приєднання об'єктів до кластерів здійснюється у випадку мінімального приросту внутрішньогрупових сум квадратів відхилень. Завдяки цьому утворюються кластери приблизно похожого розміру, які мають форму «гіперсфер».

Оптимальною варто вважати кількість визначених кластерів, яка визначається як «різниця кількості спостережень ікроків», після якої відстань об'єднання відповідно - збільшується стрибкоподібно.

Кластерний аналіз, так як й інші методи вивчення стохастичного зв'язку, вимагає численних доволі складних розрахунків, котрі краще здійснювати за допомогою сучасних інформаційних систем (ІС), зокрема з використанням програмного продукту Statistica.

Основною метою даної роботи є застосування кластерного аналізу та побудова математичних моделей, згідно результатів, які у разі їхньої

ефективності можна в подальшому застосовувати для боротьби з поширенням пандемії.

Математична модель – математичне уявлення реальності, один з варіантів моделі як системи, дослідження якої дозволяє отримувати інформацію про деяку іншу систему. Математична модель призначена передбачити поведінку реального об'єкта, але завжди є той чи інший ступінь його ідеалізації. Найважливіші математичні моделі зазвичай володіють важливою властивістю універсальності: принципово різні реальні явища можуть описуватися однією і тією ж математичною моделлю. Для створення математичних моделей використовують будь-які мат. засоби — мову диференційних чи інтегральних рівнянь, теорії множин, абстрактна алгебра, математична логіка, теорії ймовірностей, графи і т.д. Процес створення математичної моделі має назву «математичне моделювання». Це найзагальніший і найбільш використовуваний в науці, зокрема, у кібернетиці, метод досліджень. Для розробки математичних моделей широко використовують «диференційне числення», теорія множин, матриці та графи, а також «планування експерименту». Відповідно до цього розрізняють «теоретико-множинні», топологічні, матричні та поліномні мат. моделі.

Техніка «кластеризації» застосовується у найбільш різноманітніших областях. Вченний Джордж Фішер - дав прекрасний огляд різних опублікованих досліджень, які містять результати, отримані методами кластеризації: «До прикладу, в області медицини кластеризація захворювань, лікування захворюваностей чи симптомів захворювань призводить до широко використовуваним таксономія. У області психіатрії правильно застосовувати діагностику кластерів симптомів, таких як «параноя», «шизофренія», «альцгеймер», «суіцидальна параноя», і т.п., є вирішальною для успішної терапії. В сфері археології з допомогою кластерного аналізу дослідники пробують встановити таксономію кам'яних знарядь, похоронних об'єктів і т.п. Також відомі широкі застосування кластеризаційних методів в маркетингових дослідженнях. В загальному, щоразу кожного разу, коли

необхідно класифікувати так звані "гори" інформації до придатних задля подальшої обробки груп, кластеризація є корисною та ефективною.»

Загальноприйнятої класифікації методів кластерного аналізу як такої - не існує, проте можна виділити ряди груп підходів (деякі методи можна віднести одразу до декількох груп й тому пропонується розглядати дану типізацію як деяке наближення відносно реальної класифікації методів кластеризації) :

Ймовірнісний підхід передбачає, що кожен даний об'єкт відноситься до одного із Н класів. Деякі автори (наприклад, А. І. Весемир) вважають: «дана група зовсім не відноситься до кластеризації і протиставляють її під назвою «дискримінація» », тобто вибір віднесення об'єктів до однієї з ведомих груп (навчальних початкових вибірок).

- К-середніх
- К-медіан
- ЕМ-алгоритм
- Алгоритми сімейства FOREL
- дискримінантний аналіз

Підходи засновані на системі штучного інтелекту: досить умовна група, так як методів дуже багато і методично вони дуже різні.

- Логічний підхід. Побудова дендрограми здійснюється за допомогою дерева рішень.
- Метод нечіткої кластеризації С-середніх;
- Генетичний алгоритм;
- нейромережа Кохонена;
- Теоретико-графовий підхід.
- Графові кластеризаційні алгоритми;
- Ієрархічний підхід передбачає наявність вкладених груп (кластерів різного порядку). Алгоритми ж у свою чергу діляться на: агломеративні (об'єднавчі), дивизивні (розділяючі). За к-стю ознак інколи виділяють монотетичні і політетичні методи класифікації.

- Ієрархічна дивизивна кластеризація або таксономія. Завдання кластеризації розглядаються в кількісної таксономії.

Інші методи, що не ввійшли в попередні групи:

- Статистичні алгоритми кластеризації
- ансамбль кластерізаторів
- Алгоритми сімейства KRAB
- Алгоритм, заснований на методі просіювання
- DBSCAN і ін.

Підходи 4 і 5 іноді об'єднують під назвою структурного або геометричного підходу, що володіє більшою формалізованістю поняття близькості.

Незважаючи на значні відмінності між перерахованими методами все вони опираються на вихідну «гіпотезу компактності»: у просторі об'єктів все близькі об'єкти повинні належати до одного кластеру, а все різні об'єкти відповідно повинні знаходитися в різних кластерах.

Методи кластеризаційного аналізу можна класифікувати на такі як чіткі та нечіткі. Чіткі методи - розбивають вихідну множину об’єктів X на декілька «непересічних» підмножин. При цьому любий об’єкт із множини X належить лише до одного кластеру.

Нечіткі методи кластерного аналізу дозволяють будь-якоим екземплярам одночасно належати до всіх раніше визначених кластерів, але з дещо різним ступенем.

Концептуальний зв’язок між кластерним аналізом та теорією нечітких множин заснований на обставині, яка трактується так: «при вирішенні завдань структуризації складених систем більшість формованих класів об’єктів дещо розмиті за своєю природою». Саме ця «розмитість» полягає у тому, що перехід від приналежності до неприналежності елементів множин до даних класів швидше поступовий, ніж так званий «скачкоподібний». Тому найбільш адекватну відповідь в подібних випадках слід шукати не на

питання: “Чи належить даний елемент до певного класу?”, а на питання: “У якій мірі даний елемент належить цьому класу?”.

Вимога знаходження однозначної кластеризації елементів досліджуваної предметної області є доволі грубою та жорсткою, особливо якщо це стосується рішення задач системного аналізу, які слабо структуруються. Методи нечіткої кластеризації послабляють дану вимогу. Послаблення вимоги здійснюється рахунком введення в розгляд нечітких кластерів та відповідних їм функцій «приналежності», які набувають значень з інтервалу [0, 1].

У загальному ж випадку основним завданням «нечіткої кластеризації» є саме знаходження розбиття множини елементів досліджуваної вибірки, які утворюють структури «нечітких» кластерів, які присутні у вхідних даних. Дане завдання зводиться до знаходження міри приналежності елементів, універсальну суму (універсум) шуканим нечітким кластерам, які в сукупності і визначають дане не чітке розбиття фінальної множини елементів.

Приклад 1. Метелик являє собою 16 об’єктів, двовимірне зображення яких нагадує «комаху». При чіткій кластеризації отримуємо два кластери з 8 об’єктів.

Розглянемо також «горизонтальну деревоподібну діаграму». Діаграма починається із кожного об'єкту в класі (у лівій частині діаграми). Тепер уявім, що поступово (дрібними кроками) ви, так би мовити - "послабляєте" ваш критерій про те, котрі саме об'єкти є «унікальними», а які ні. Інакше кажучи , ви знижуєте поріг, що відноситься вже до вирішення про об'єднання 2-х або більше об'єктів до одного кластеру.

Порівняння чіткої та нечіткої кластеризації «метелика»:

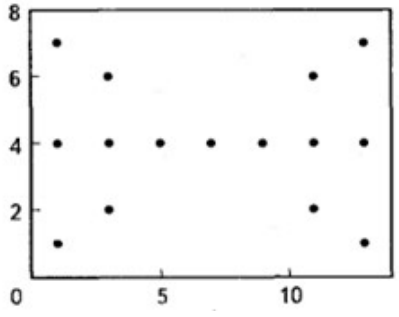


Рисунок 2.5 - вхідні дані.

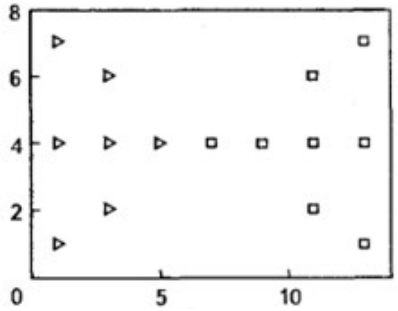


Рисунок 2.6 - чітка кластеризація I.

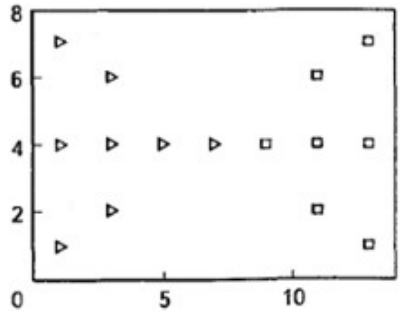


Рисунок 2.7 - чітка кластеризація II.

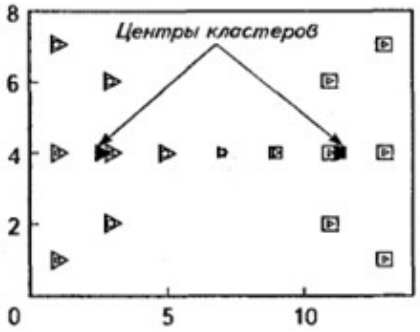


Рисунок 2.8 - нечітка кластеризація

На малюнку об’єкти 1-го кластера позначені «трикутничками», а другого – «квадратиками». Симетричний «метелик» під час чіткої кластеризації розбивається на 2 несиметричні кластери. При нечіткій кластеризації проблемний восьмий об’єкт, який розташований в центрі цього «метелика», одночасно належить двом абсолютно симетричним кластерам з однією й тією ж мірою. На даному малюнку розмір маркерів пропорційний до міри приналежності об’єктів кластера.

Метод «деревоподібної кластеризації» дозволяє побудувати ієрархічне кластер-дерево, що має наступний вигляд:

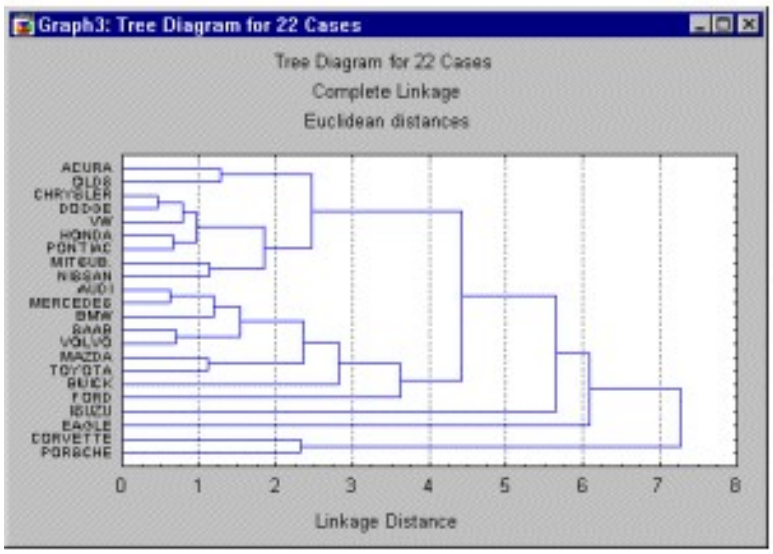


Рисунок 2.9 – Деревоподібна кластеризація.

В цьому випадку розглядається задача з об'єднання розрізнених представників деякої сукупності даних в кластери або групи великих розмірів. Між об'єктами існують деякі подібності або відмінності, які можна виразити як «відстань між об'єктами».

Визначити відстані, які мають чисельні значення не викликає складності. Наприклад, числа об'єднані в десятки, сотні, тисячі і т.д. Один унікальний об'єкт цілком може входити в певну десятку, яка в свою чергу входить в деяку сотню. Відстані між об'єктами визначається як числова одиниця (наприклад, гривня або десять гривень). Таку процедуру можна представити у вигляді діаграми, яка дещо нагадує дерево.

В результаті, ви пов'язуєте між сбою все більше і більше число об'єктів і агрегуєте більше кластерів, які в результаті складаються з більш сильніших розрізнених елементів. В кінцевому результаті, на останньому кроці всі об'єкти об'єднуються разом воєдину. На зображених дендрограмах горизонтальні осі представляють «відстань об'єднання» (у вертикальних деревовидних дендрограмах вертикальні осі представляють відстань об'єднань). Так, для кожного вузла у графі (там, де формується новенький кластерочок) ви можете спостерігати величину відстані, для котрого відповідні «елементи» зв'язують у новий, єдиний кластер. Коли дані мають

чітку "структуру" в термінах кластерів об'єктів, схожих між собою, тоді дана структура, скоріш за все, повинна бути відображена в ієрархічному дереві за допомогою різних гілок гілками. В кінцевому результаті успішного аналізу завдяки методу об'єднання з'являється можливість виявити кластери та інтерпретувати їх.

Метод двохвхідного об`єднання

У даному виді аналізу питання що постає перед дослідником зазвичай виражається у термінах спостережень чи змінних. Виявляється те, що кластеризація, як за спостереженнями, так і за змінними може привести до цікавих результатів. Наприклад, уявіть, що медичний дослідник збирає дані щодо різних характеристик (змінних) станів пацієнтів (спостережень), які страждають серцево-судинними захворюваннями. Дослідник може захотіти кластеризувати ці спостереження для визначення кластерів пацієнтів з схожими симптомами. У той же час дослідник цілком може захотіти «кластеризувати» змінні задля визначення кластерів змінних, котрі пов'язані з батьківською виключно фізичним станом.

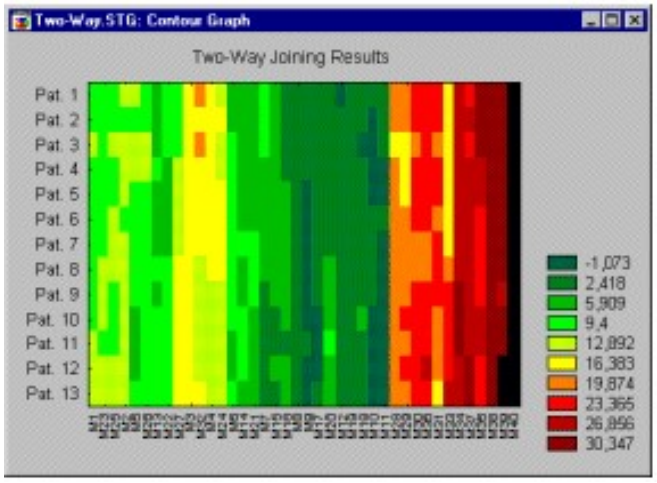


Рисунок 2.10 – Графік двох вхідної кластеризації

Повертаючись до попереднього прикладу, можна припустити, що досліднику необхідно виділити кластери пацієнтів, які схожі по відношенню до певних кластерів характеристики фізичного стану. Труднощі з інтерпретацією отриманих результатів виникає в результаті того, що

подібності між цими «різними» кластерами можуть відбуватися з (чи бути причиною) деякої відмінності підмножин змінних. Тому що кластери є за своєю природою так би мовити «неоднорідними». Це може здатися спочатку дещо «туманним», адже справді, у порівнянні з іншими описаними методами кластеризації, двохвхідне об'єднання є, чи не найменш часто використовуваним методом. Але - деякі дослідники все таки вважають, що даний метод пропонує доволі «потужний» засіб розвідувального аналізу даних (за більш детальною інформацією ви можете скористатися описом цього методу у «Хартігана (Hartigan, 1975)»).

Ієрархічні агломеративні методи («Agglomerative Nesting, AGNES») - характеризується послідовними об’єднаннями початкових елементів і відповідним зменшенням числа кластерів.

Усі ці методи переглядають матрицю схожості розмірностей N N (де N – це кількість об’єктів) та послідовно об’єднують більш схожі об’єкти. Послідовність об’єднань таких кластерів можна подати «візуально» у вигляді дендрограм:

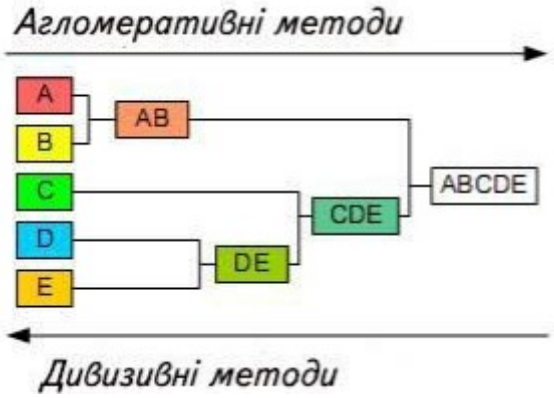


Рисунок 2.11 – Ієрархічний метод.

Ієрархічні агломеративні методи відрізняються зачасту за правилами побудови «кластерів». Також є багато різних правил групувань, кожне з яких породжує специфічний ієрархічний метод. Найпоширеніші чотири з методів: «єдиничного зв’язку», «повного зв’язку», середнього зв’язку та «метод Уорда».

Метод одиночного зв’язку (nearest neighbor analysis) – його ще інколи називають, методом аналізу найближчого сусіда, який є одним з найпростіших ієрархічних методів, не дивлячись на те, що він був запропонований одним з останніх – ще у 1973 р.

В основі цього алгоритму лежить припущення про те, що якщо об’єкти надто близькі за значеннями н-1 властивості, тоді вони близькі й за значеннями n-ї властивості.

Метод Варда було запропоновано у 1963 р. і на сьогоднішній день він залишається одним з найпопулярніших методів кластерного аналізу. Він побудований, щоб оптимізувати найменшу дисперсію всередині самих кластерів. Дана цільова функція відома як внутрішньогрупова сума квадратів (ВСК) або сума квадратів відхилень (СКВ). На 1 кроці, коли кожен окремий кластер складається із єдиного об’єкта, СКВ рівне нулю. За методом Уорда об’єднуються групи чи об’єкти, для котрих СКВ отримує найменший приріст.

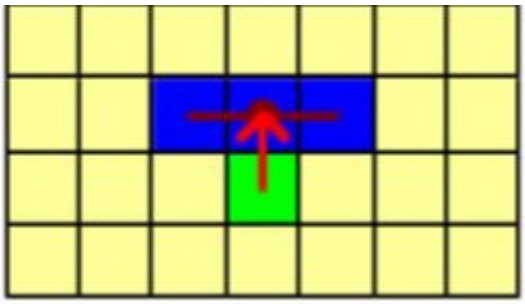


Рисунок 2.12 – принцип методу Варда.

Цей метод має тенденцію до «знаходження» або розробки кластерів приблизно таких рівних розмірів, які мають «гіперсферичну» форму.

Неієрархічні методи мають за основу вже заздалегідь задану к-сть кластерів («k-середніх, PAM кластеризація») або застосовують складні алгоритмічні методи знаходження їх кількості.

Ієрархічна кластеризація виконується переважно за допомогою послідовного об’єднання найменших кластерів до найбільших чи навпаки –

від більших до найменших (дивизивна). На відміну від не ієрархічних, ці алгоритми кластеризації будують розбиття для усіх можливих варіантів.

Із «неієрархічних» методів кластеризації особливої уваги заслуговують так звані - ітеративні методи. Вони функціонують за наступним визначеним алгоритмом:

1) спочатку вихідні дані розбивають на деяку кількість кластерів, пілся чого обчислюють центри тяжіння цих кластерів;

2)далі кожна точка даних поміщується в кластері з найближчим центром тяжіння;

3)після цього обраховуються біільш нові «центри тяжіння» кластерів; кластери не замінюються на нові до поки, поки вони не будуть повністю переглянуті разом з всіма даними;

4) За необхідністю - кроки 2 й 3 повторюються до тих пір, поки не перестануть змінюватись кластери.

На відміну від ієрархічних методів, які потребують обчислень і збереження матриці збіжності між об’єктами розмірністю N ітеративні методи працюють виключно з початковими даними. Тому за їхньою допомогою можна обробляти досить великі обсяги даних. Окрім того, методи ітерації виконують декілька переглядів даних і за рахунок чого компенсують наслідки невдалих вихідних розбивок даних. Дані методи породжують кластери одного рангу, які не «вкладені», і тому не можуть бути частиною ієрархії. Більшість із цих ітеративних методів не дають змогу так званого «перекриття» кластерів. Як правило властивості ітеративних методів групування можуть описуватися за допомогою трьох основних факторів: вибір вихідної розбивки, тип ітерації і статистичний критерій. Ці чинники можуть різним способом поєднуватись, утворюючи алгоритми вибору даних під час визначенні найоптимальнішої розбивки. Різноманітні комбінації ведуть до розробки методів, що породжують різні результати при роботі з одними і тими ж даними [9].

Ітерації згідно наведеними принципами полягають у приєднанні об’єктів у кластер із найбільшим центром тяжіння. Кількість фінальних кластерів фіксована до початку кластеризації. Перерахунок центру тяжіння кластера може здійснюватися як після кожної зміни його складу, так і після того, коли буде завершено перегляд усіх вхідних даних. На сьогодні існує досить багато варіантів цього методу, які відрізняються особливостями роботи.

До найбпростіших і ефективних алгоритмів кластеризації відноситься k-середній метод, запропонований Г. Боллом і Д. Холлом у 1965 р. [10]. Конструктивно алгоритм – це ітераційна процедура, яка складається з кроків:

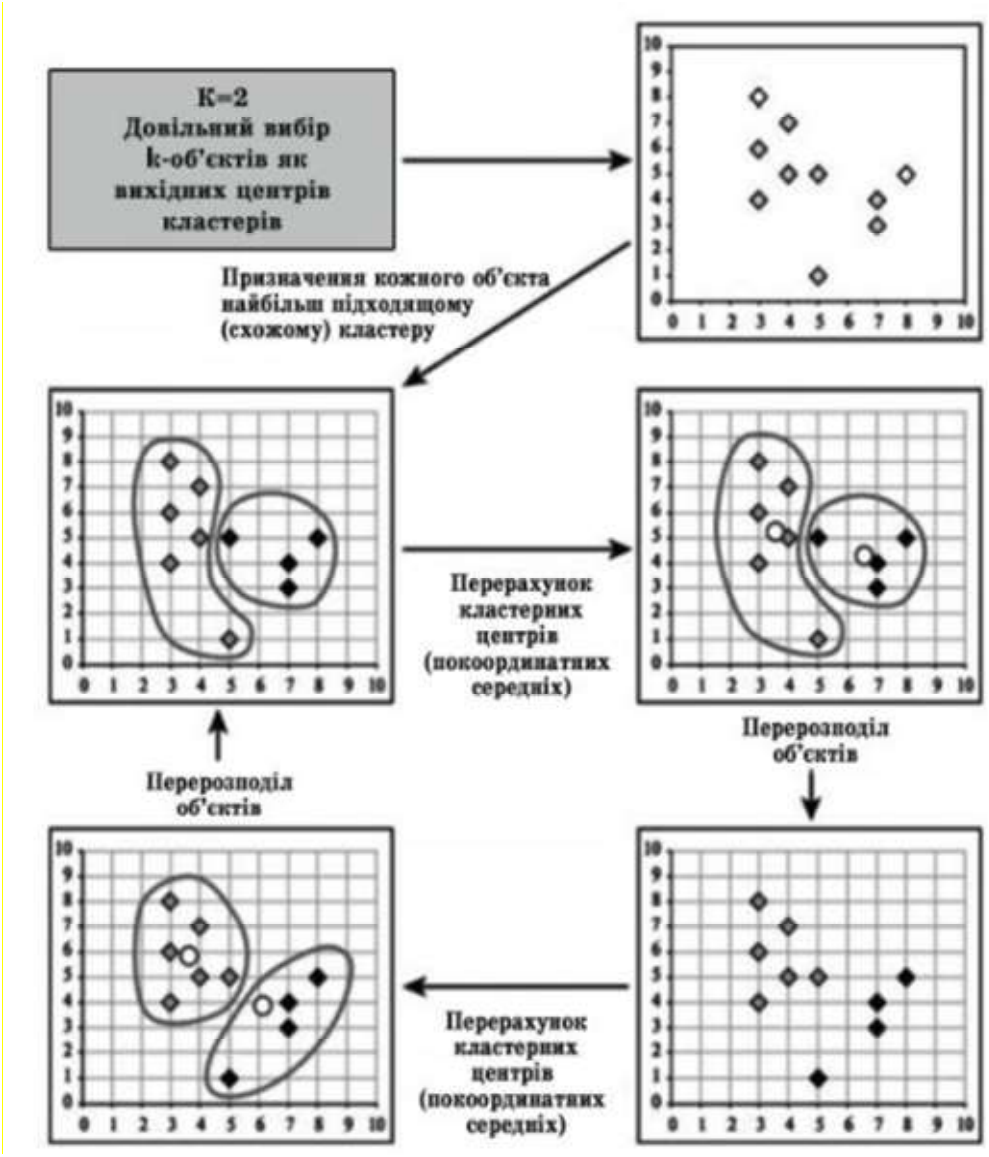


Рисунок 2.13 – Принцип роботи алгоритму К-середніх

1. Задається початкова кількість кластерів k, яка повинна бути заздалегідь сформована з деяких об’єктів вхідної вибірки.

2. Випадковим чином вибирається H записів, що будуть слугувати «початковими центрами» цих кластерів. Початкові точки, з яких потім виростають кластери, також, називають "насінням". Кожен запис являє собою "ембріон" кластера, що складається лише з одного елемента.

3. Для кожного подальшого запису вхідної вибірки визначають найближчий до неї центр кластера.

4. Далі проводиться обчислення «центроїдів» – або «центрів тяжіння» цих кластерів. Це робиться переважно шляхом визначення середнього значення для значень кожної ознаки усіх записів у кожному кластері. До прикладу, якщо в кластер увійшли три записи з наборами ознак (x1, y1), (x2, y2), (x3, y3), то координати його центроїда будуть розраховуватися так:

$$(x, y) \quad \frac{x_1 x_2 x_3}{3}, \frac{y_1 y_2 y_3}{3}. \qquad (2.1)$$

Потім старий «центр кластера» зміщується у його центроід. Таким чином, центроїдні обєкти стають новітніми центрами кластерів для проведення наступної Ітерації алгоритму.

Кроки 3 і 4 повторюються доти, доки виконання алгоритму не буде перервано дослідником або доки не буде виконуватися умова відповідно до певного критерію збіжності.

Зупинка цього алгоритму проводиться в момент, коли границі кластерів та розташування центроїдів перестають змінюватися. На кожній ітерації у кожному кластері залишається один і той набір записів.

Алгоритм k-середніх звичайно знаходить набір стабільних кластерів за кілька 10-в ітерацій.

Також існує ще один аспект, про який варто згадати. Це питання кластеризації усієї вибірки даних або ж із її вибірки. Названий аспект дуже важливий для двох розглянутих груп, методів, проте він більш критичніший

для ієрархічних методів. Ієрархічні методи ніяк не можуть працювати із чималими наборами даних, а використання «деякої» вибірки, мається на увазі - частини даних, могло б дозволити застосовувати ці методи у практиці.

«Одним із недоліків даного методу є порушення умови зв'язності елементів 1 кластера, тому розвиваються різні модифікації даного методу, а також його нечіткі аналоги, у яких на 1-й стадії алгоритму допускається приналежність 1 елемента множини до кількох кластерів».

Самоорганізаційна Карта Кохонена.

Дана карта складається з певних компонентів, які називаються «вузлами» або «нейронами». Їхня кількість задається аналітиком заздалегідь. Кожен з вузлів описується двома основними векторами. Перший – це вектор так званої ваги m, котрий має таку ж розмірність, що і вхідні дані. Другий – це вектор r, який являє собою координати вузла на карті. Ця карта візуально відображається за допомогою осередків прямокутної або багатокутньої форми; остання застосовується більш частіше, оскільки в даному випадку відстані між центрами суміжних осередків однакові, що значно підвищує коректність візуалізації цієї карти.

Для початку відома розмірність цих вхідних даних, по ній певним чином будується початковий варіант цієї «карти». В процесі навчання вектори ваг вузлів наближаються до самих «вхідних даних». Для кожного із спостережень обирається найбільш схожий до вектора вагів узол, а також значення його «вектора ваги» яке наближається до спостереження. Також до спостереження наближаються й вектори ваги декількох окремих вузлів, розташованих поруч один з одним, таким чином якщо у безлічі вхідних даних два спостереження були схожими, на карті їм відповідатимуть близькі вузли. Циклічний процес навчання, який перебирає вхідні дані, закінчується після досягнення картою допустимої похибки, чи після здійснення так званої - заданої к-сті І-терацій. Таким чином, в кінцевому результаті навчання «карта Кохонена» класифікує вхідні дані на відповідні кластери та візуалізує багатовимірні вхідні дані в «двовимірній площині», розподіляючи вектори

близьких ознак в сусідні кластери і розфарбовуючи їх у залежності від аналізованих параметрів.

Метод відстаней.

Об'єднання або метод деревовидної кластеризації використовується під час формування кластерів розбіжності або відстані між об'єктами. Дані відстані можуть визначатися як в одновимірному так і в багатовимірному просторі. До прикладу, якщо ви будете кластеризувати типи їжі у кафе, то можете взяти до уваги к-сть ціну, калорій, оцінку смаку і т.д. Найбільш прямий спосіб обрахування відстаней між двома об'єктами в багатовимірному просторі полягає у обчисленні «Евклідових відстаней». Якщо ви маєте 2-х або 3 простір, то цей захід є реальною геометричною відстанню між об'єктами у просторі. Проте алгоритм об'єднання не «дбає» про те, чи є "надані" для даної відстані справжніми чи деякими іншими похідними заходами відстані, що більш значуще для дослідника; завданням дослідників є підібрати правильний метод для специфічних застосувань.

Методи об'єднання або метод зв'язку

Коли кожний об'єкт представляє собою окремий кластер, відстані між цими об'єктами визначаються заздалегідь обраної мірою. Виникає питання - як визначити відстані між кластерами? Існують різні правила, йменовані методами об'єднання або зв'язку для двох кластерів.

Метод найбільш віддалених сусідів або «повного зв'язку». Тут відстані між кластерами визначаються як найбільша відстань між будь-якими двома об'єктами у різних кластерах (тобто "найбільш віддаленими сусідами"). Метод доцільно застосовувати, коли об'єкти дійсно відбуваються з різних сукупностей. Якщо ж кластери мають у деякому роді «подовжену» форму чи їх природний тип є "ланцюговим", то цей метод не доцільно використовувати.

Метод ближнього сусіда або одиночного зв'язку. Тут відстань між двома кластерами визначається за допомогою відстаней між 2 найбільш близькими об'єктами у різних кластерах. Цей метод дозволяє виділити

кластери будь якої складної форми за умови, що різні частини даних кластерів з'єднані ланцюгами близьких 1 до одного елементів. В результаті роботи даного методу кластери представляються потужніми "ланцюжками" або "волокнистими" кластерами, які "зчепленими разом" тільки окремими елементами, що випадково опинилися ближче інших один до одного.

Метод Варда (Уорда) – згідно цього методу як відстань поміж кластерами беруть приріст суми квадратів відстаней об'єктів до центрів кластерів, який отримується в результаті їхнього об'єднання. На відміну від інакших методів кластерного аналізу для оцінки відстаней між двома кластерами, тут використовуються методи так званого «дисперсійного аналізу». На кожному кроці даного алгоритму об'єднують такі 2 кластери, які призводять до найменшого збільшення в цільовій функції, тобто внутрішньо групової суми квадратів (ВГСК). Даний метод направлений на об'єднання близьких кластерів і "прагне" створювати кластери все меншого розміру.

Метод невиваженого попарного середнього.

Як відстані між двома кластерами беруть середню відстань між усіма парами об'єктів у них. Даний метод слід використати, якщо об'єкти дійсно відбуваються із різних точок, у випадках присутності кластерів ланцюжкового типу, при припущенні кривих розмірів кластерів.

Метод зваженого попарного середнього.

Даний метод схожий на раніше згаданий метод «невиваженого попарного середнього», але різниця полягає лише в тім, що тут у якості вагового коефіцієнта використовується сам розмір кластера (к-сть об'єктів, що містяться в кластері).

Даний метод рекомендується використовувати саме при наявності припущення про кластери з різними розмірами.

Незважений центроїдний метод (метод невиваженого попарного центроїдного усереднення).

Як відстань між двома кластерами в даному методі береться відстані між їх центрами тяжкості.

Зважений центроїдний метод.

Даний метод дуже схожий на попередній, різниця полягає лише в тім, що для обліку різниці поміж розмірами кластерів, використовуються ваги. Цей метод доцільно використовувати у випадках, коли є припущення щодо істотних відмінностей у розмірах деяких кластерів.

При аналізі результатів соц. досліджень рекомендується здійснювати аналізи методами: ієрархічного агломеративного сімейства, методом Варда, при якому в кластерах оптимізується мінімальна дисперсія, і в результаті створюються кластери більш-менш схожих розмірів. Метод Варда найбільш вдалий для проведення аналізу соціологічних даних. В якості запобіжної відмінності краще квадратична Евклідова відстань, яка сприяє збільшенню «контрастності кластерів». «Головним підсумком ієрархічного кластерного аналізу є дендограма. При її інтерпретації дослідники стикаються із проблемою того ж роду, що і тлумачення результатів так званого факторного аналізу – відсутністю однозначних критеріїв виділення кластерів. В якості головних рекомендовано використовувати 2 способи - візуальний аналіз графіків і порівняння результатів кластеризації, виконаної різними методами» - стверджує соціолог Дензіель Вашингтон.

Візуальний аналіз дендрограми передбачає так зване «обрізання» дерева на оптимальному рівні подібності елементів оибірки. «Виноградну гілку» доцільно «різати» на позначці 5 шкал, таким чином буде досягнуто 80% рівня подібності. Якщо виділення кластерів по цій мітці буде утруднено (на ній буде злиття декількох дрібненьких кластерів в один великий), то можна вибрати іншу мітку.

Тепер виникає лише питання стійкості прийнятого рішення. По факту, перевірка стійкості кластеризації зводиться до перевірки достовірностей. Тут є так зване «емпіричне правило» яке трактується як – стійка типологія зберігається лише при зміні методів кластеризації. Результати ієрархічного кластерного аналізу можна перевіряти завдяки ітеративним кластерним аналізом методом k-середніх. Якщо порівнювані класифікаційні груп

респондентів мають масову частку збігів більше ніж 74% (більше 2.5/3 збігів), в такому випадку рішення приймається.

Перевірити адекватність рішення, без допомоги іншого виду аналізу, на жаль не можна. Принаймні, у теоретичному плані ця проблема досі не вирішена. У класичній праці «Кластерний аналіз» докладно розглядаються а в підсумку відкидаються додаткові 5 методів перевірки стійкості:

- тести значущості для зовнішніх ознак придатні тільки для повторних вимірів;
- кофенетична кореляція – не рекомендується і обмежена у використанні;
- методика повторних (випадкових) вибірок, що, тим не менш, не доводить обґрунтованість рішення;
- тести значущості (дисперсійний аналіз) - завжди дають значущий результат;
- методи Монте-Карло дуже складні й доступні тільки досвідченим математикам

Кластеризація на основі зв’язків, також відома як ієрархічна кластеризація, базується на основній ідеї об’єктів, які більше пов’язані з об’єктами поблизу, ніж з об’єктами, що знаходяться далі. Ці алгоритми з'єднують "об'єкти", утворюючи "кластери" на основі їх відстані. Кластер можна описати значною мірою за допомогою максимальної відстані, необхідної для з'єднання частин кластера. На різній відстані утворюються різні кластери, які можна представити за допомогою дендрограми, яка пояснює, від чого загальна назва " ієрархічна кластеризація" походить. Ці алгоритми не забезпечують єдиного розділення набору даних, а натомість забезпечують розгалужену ієрархію кластерів, які зливаються між собою на певній відстані. У дендрограмі вісь у позначає відстань, на якій кластери зливаються, тоді як об’єкти розміщені вздовж осі x так, щоб кластери не змішувалися.

Кластеризація на основі зв’язків – це ціле сімейство методів, які відрізняються способом обчислення відстаней. Окрім звичайного вибору функцій відстані, користувачеві також потрібно визначитися з критерієм зв’язку (оскільки кластер складається з декількох об’єктів, існує кілька кандидатів для обчислення відстані) для використання. Популярний вибір відомий як одне-важільна кластеризація (мінімум об'єкт відстаней), метод повного зв'язку (максимум об'єктних відстаней), і UPGMA або WPGMA("Метод незважених або зважених парних груп із середнім арифметичним значенням", також відомий як середнє кластеризування зв'язків). Крім того, ієрархічна кластеризація може бути агломеративною (починаючи з окремих елементів та об’єднуючи їх у кластери) або розділювальною (починаючи з повного набору даних та ділячи його на розділи).

Ці методи не дадуть унікального розділення набору даних, а ієрархії, з якої користувачеві все одно потрібно вибрати відповідні кластери. Вони не надто надійні щодо викидів, які або відображатимуться як додаткові кластери, або навіть спричинятимуть злиття інших кластерів (відомий як "феномен ланцюга ", зокрема, з кластеризацією з однією зв'язкою). Разом із обробкою даних ці методи визнані теоретичною основою кластерного аналізу, але часто вважаються застарілими. Однак вони дали натхнення для багатьох пізніших методів, таких як кластеризація на основі щільності.

Кластер на основі розподілу.

Модель кластеризації, найбільш тісно пов'язана зі статистикою, базується на моделях розподілу. Кластери тоді можна легкго визначити як об’єкти, що належать, швидше за все, до одного розподіллу. Зручною властивістю цього підходу є те, що це дуже нагадує спосіб генерування штучних наборів даних: шляхом вибірки випадкових об’єктів з розподілу.

Хоча теоретична основа цих методів чудова, вони страждають від однієї ключової проблеми, відомої як переобладнання, якщо тільки не

обмежуються складність моделі. Більш складна модель зазвичай зможе краще пояснити дані, що ускладнює вибір відповідної моделі.

Один з відомих методів відомий як моделі Гаусова суміші (з використанням алгоритму очікування-максимізації). Тут набір даних зазвичай моделюється з фіксованою (щоб уникнути переобладнання) кількістю гауссових розподілів, які ініціалізуються випадковим чином і параметри яких ітеративно оптимізовані для кращого відповідності набору даних. Це сходиться до локального оптимуму, тому багаторазові прогони можуть дати різні результати. Для того щоб отримати жорстку кластеризацію, об'єкти часто потім призначаються гауссовому розподілу, якому вони, швидше за все, належать; для м'яких кластерів це не потрібно.

Кластеризація на основі розподілу створює складні моделі для кластерів, які можуть фіксувати кореляцію та залежність між атрибутами. Однак ці алгоритми створюють додаткове навантаження для користувача: для багатьох реальних наборів даних може не бути стисло визначеної математичної моделі (наприклад, припускаючи, що розподіл Гауса є досить вагомим припущенням про дані).

Кластеризація на основі щільності.

У кластеризації на основі щільності кластери визначаються як області з більшою щільністю, ніж решта набору данних. Об'єктами в розріджених районах - які необхідні для відокремлення скупчень - зазвичай вважаються точки шуму та кордону.

Найпопулярнішим методом кластеризації на основі щільності є DBSCAN. На відміну від багатьох нових методів, він має чітко визначену кластерну модель, яка називається "щільність-доступність". Подібно до кластеризації на основі зв'язків, вона базується на точках з'єднання в межах певних порогів відстані. Однак він з'єднує лише точки, які задовольняють критерію щільності, в оригінальному варіанті визначений як мінімальна кількість інших об'єктів у цьому радіусі. Кластер складається з усіх об’єктів, пов’язаних із щільністю (які можуть утворювати кластер довільної форми, на

відміну від багатьох інших методів), а також усіх об’єктів, що знаходяться в межах діапазону цих об’єктів. Ще однією цікавою властивістю DBSCAN є те, що його складність досить низька - вона вимагає лінійної кількості запитів діапазону в базі даних - і що він виявить по суті однакові результати (це детерміновано для основних і шумових точок, але не для прикордонних) у кожному прогоні, тому немає необхідності запускати його кілька разів. OPTICS - це узагальнення DBSCAN, яке позбавляє потреби вибирати відповідне значення для параметра діапазону і створює ієрархічний результат, пов’язаний з результатом кластеризації зв’язків. DeLi-Clu, Dustity-Link-Clastering поєднує в собі ідеї кластеризації з одним зв’язком та OPTICS, усуваючи параметр повністю та пропонує покращення продуктивності в порівнянні з OPTICS за допомогою індексу дерева R.

Ключовим недоліком DBSCAN та OPTICS є те, що вони очікують певного падіння щільності для виявлення меж кластера. На наборах даних із, наприклад, перекриттями гауссових розподілів – типовим випадком використання штучних даних - межі кластера, створені цими алгоритмами, часто виглядатимуть довільними, оскільки щільність кластера постійно зменшується. На наборі даних, що складається із сумішей Гаусса, ці алгоритми майже завжди перевершують такі методи, як кластеризація EM, які здатні точно моделювати такий тип даних.

Середній зсув – це підхід кластеризації, при якому кожен об’єкт переміщується в найщільнішу область поблизу, на основі оцінки щільності ядра. Згодом об’єкти сходяться до локальних максимумів щільності. Подібно до кластеризації k-середніх, ці "атрактори щільності" можуть служити представниками набору даних, але середній зсув може виявляти кластери довільної форми, подібні до DBSCAN. Через дорогу ітераційну процедуру та оцінку щільності, середній зсув зазвичай повільніший, ніж DBSCAN або k-Means. Крім того, застосовності алгоритму середнього зсуву до багатовимірних даних заважає негладка поведінка оцінки щільності ядра, що призводить до надмірної фрагментації хвостів кластера.

Оцінка (або "перевірка") результатів кластеризації така ж складна, як і сама кластеризація. [33] Популярні підходи передбачають "внутрінє " оцінювання, де кластеризація узагальнюється до єдиного балу якості, " зовнішнє " оцінювання, де кластеризація порівнюється з існуючою класифікацією "основна істина", " ручне " оцінювання експертом-людиною та " непряма " оцінка шляхом оцінки корисності кластеризації за призначенням. [34]

Заходи внутрішньої оцінки страждають від того, що вони представляють функції, які самі по собі можуть розглядатися як мета кластеризації. Наприклад, можна згрупувати дані, встановлені за коефіцієнтом Силует; за винятком того, що для цього не існує відомого ефективного алгоритму. Використовуючи такий внутрішній показник для оцінки, можна порівняти схожість проблем оптимізації і не обов’язково, наскільки корисною є кластеризація.

Зовнішнє оцінювання має подібні проблеми: якщо ми маємо такі мітки "основна істина", то нам не потрібно буде кластеризуватися; а в практичному застосуванні ми зазвичай не маємо таких ярликів. З іншого боку, мітки відображають лише одне можливе розподіл набору даних, що не означає, що не існує іншого, а може, навіть кращого, кластеризації.

Тому жоден із цих підходів не може врешті-решт судити про фактичну якість кластеризації, але для цього потрібна людська оцінка [3], яка є вкрай суб’єктивною. Тим не менше, така статистика може бути досить інформативною для виявлення поганих скупчень [5], але не слід відкидати суб'єктивну оцінку людини.

Коли результат кластеризації оцінюється на основі даних, які були скупчені самі, це називається внутрішнім оцінюванням. Ці методи зазвичай призначають найкращий бал алгоритму, який створює кластери з високою схожістю всередині кластера та низькою схожістю між кластерами. Одним недоліком використання внутрішніх критеріїв у кластерній оцінці є те, що високі бали за внутрішнім показником не обов'язково призводять до

ефективних програм пошуку інформації. [6] Крім того, ця оцінка упереджена до алгоритмів, що використовують ту саму кластерну модель. Наприклад, кластеризація k-означає, природно, оптимізує відстань до об'єкта, і внутрішній критерій на основі відстані, ймовірно, переоцінить результуючу кластеризацію.

Отже, заходи внутрішнього оцінювання найкраще підходять для того, щоб отримати деяке розуміння ситуацій, коли один алгоритм працює краще, ніж інший, але це не означає, що один алгоритм дає більше достовірних результатів, ніж інший. [5] Дійсність, виміряна таким індексом, залежить від твердження, що такий тип структури існує у наборі даних. Алгоритм, розроблений для якихось моделей, не має шансів, якщо набір даних містить кардинально інший набір моделей або якщо оцінка вимірює кардинально інший критерій. [5] Наприклад, k-означає кластеризацію може знаходити лише опуклі кластери, і багато індекси оцінки припускають опуклі кластери. На наборі даних з неопуклими кластерами також не використовується k-значення, ані критерій оцінки, який передбачає опуклість, не є обґрунтованим.

Існує більше десятка заходів внутрішньої оцінки, як правило, заснованих на інтуїції, що елементи в одному кластері повинні бути більш подібними.

При зовнішньому оцінюванні результати кластеризації оцінюються на основі даних, які не використовувались для кластеризації, таких як відомі мітки класів та зовнішні тести. Такі показники складаються з набору попередньо класифікованих предметів, і ці набори часто створюються (експертами) людьми. Таким чином, набори еталонів можна розглядати як золотий стандарт для оцінки. Ці типи методів оцінки вимірюють, наскільки кластеризація наближена до наперед визначених базових класів. Однак нещодавно було обговорено, чи це адекватно для реальних даних, або лише для синтетичних наборів даних з фактичною базовою істиною, оскільки класи можуть містити внутрішню структуру, наявні атрибути можуть не

допускати поділу кластерів або класи можуть містити аномалії. Крім того, з точки зору відкриття знань, відтворення відомих знань не обов'язково може бути запланованим результатом. У спеціальному сценарії обмеженої кластеризації, коли метаінформація (наприклад, мітки класів) використовується вже в процесі кластеризації, затримка інформації для цілей оцінки є нетривіальною.

Ряд заходів адаптовано до варіантів, що використовуються для оцінки класифікаційних завдань. Замість підрахунку кількості випадків, коли клас був правильно призначений одній точці даних (відомий як справжні позитивні дані), такі метрики підрахунку пар оцінюють, чи передбачається, що кожна пара точок даних, яка справді знаходиться в одному кластері, знаходиться в одному і тому ж скупченні.

Результати кластеризаційного аналізу можуть й не мати достатнього статистичного обґрунтування. Проте, з другого боку – під час розв’язання задач кластеризації цілком допустима нестатистична інтерпретація отриманих результатів, а також доволі велика різноманітність варіантів поняття кластера. Така нестатистична інтерпретація дає можливість аналітикам одержувати результати кластеризації, які задовольняють умови, що у разі використання інших методів часто доволі таки складно.

РОЗДІЛ 3. ПРОВЕДЕННЯ АНАЛІЗУ МЕТОДАМИ КЛАСТЕРИЗАЦІЇ

3.1 Постановка задачі й завдань досліджень.

При формуванні постановки задачі, важливо враховувати такі аспекти, як актуальність та мета завдання.

Актуальність дипломної роботи полягає у тому, що при даній динамічної моделі, можливе подальше дослідження зон зараження, динаміку розповсюдження захворюваності та, в кінцевому результаті, будування ефективної методики боротьби з вірусом.

Метою дипломної роботи є дослідження факторів ризику захворюваності на CoVID-29.

Об’єктом наукового дослідження є кількість захворювань, за період 2 кварталу.

Отже, постановка задачі, для даного проекту буде наступною: необхідно розробити модель динамки захворюваності приросту захворюваності на CoVID-19 в період 2 кварталу в усіх регіонах України. Дослідити актуальні статистичні данні, розробити початкову модель динаміки, порівняти динаміку захворюваності в 2 кварталі, з теперішнім (3 квартал), розробити кінцеву модель.

Завдання, до розробки даного проекту буде наступним:

- дослідити актуальні данні;
- побудувати графік приросту захворюваності;
- вивести формулу приросту захворюваності;
- за допомогою отриманих формул, обрахувати приріст в період поточного кварталу, і порівняти результат, з реальними статистичними даними;
- провести роботу над виправленням помилок;
- сформувати кінцеву модель динаміки.

Таким чином, було виділено мету, об`єкт та актуальність результаті чого було сформовано постановку задачі й завдання дослідження.

3.2 Дослідження динаміки захворюваності. Розробка кластерної моделі аналізу

Складну «медико-географічну» ситуацію в Україні визначають досить швидкі темпи поширеності хвороби, надто високий рівень захворюваності населення та суттєві регіональні диспропорції в рівнях захворювання. Разом з погіршенням якості навколишнього середовища (НС), соц.-економічних умов життя (СЕУЖ), рівня якості медичного обслуговування та фінансування мед. закладів регіональна диференціація медико – географічної ситуації посилюється. В результаті це зумовлює актуальність комплексних медико-географічних досліджень.

Аналізу були піддані показники, що характеризують захворюваність і поширеність захворювання, а також смертність від них у регіонах України.

Для оцінки співвідношення між поширеністю захворювань і смертністю від них у щорічних епідеміологічних показниках був використаний вибірковий метод виділення регіонів, що характеризують субпопуляції з малою, середньою і високою поширеністю захворювання. На прикладі даних за березень – вересень 2020 року. показано виділення із усієї сукупності показників трьох вибірок з низькою (субпопуляція А), середньою (субпопуляція В) і високою (субпопуляція С) поширеністю хвороб.

Співвідношення розповсюдженості хвороби і смертностей від них оцінювалось у методичному варіанті, який передбачав оцінку регіонів «за поширеністю хвороби» в поточному році без урахування значень цього показника в попередні роки.

Отримані дані використовувалися для оцінки впливу зростання активності патогенна на збитки від нього (при переході від популяції з низькою поширеністю хвороби до популяції із середнім рівнем цього

показника або від популяції із середньою поширеністю хвороби до популяції з високим рівнем цього показника). У субпопуляціях з різним рівнем поширеності хвороби щороку оцінювались показники летальності, а також коефіцієнт виживання (КВ), який становить відношення субпопуляції, пригніченої захворюванням, яка незважаючи на це зберігає життєздатність, до субпопуляції, що не витримує тягаря хвороби і вмирає протягом року. Цей показник, виражений в умовних одиницях, становить дріб (поширеність хвороби – смертність) / смертність.

Визначалися коефіцієнти лінійної кореляції Пірсона між епідеміологічними показниками.

Епідеміологічні показники протягом 6 місяців свідчать про значну мінливість, якій не притаманна тенденція до зростання чи спаду. Показники, особливо захворюваність, коливаються. Різниця між найбільшими і найменшими показниками поширеності хвороби і захворюваності в різних роках становить відповідно 26,4 і 34,5 %. Щорічна захворюваність складає майже 9/10 (точніше, 87,7 %) від усієї поширеності хвороби. Смертність відносно значна – в середньому 4,00 на 100 тис. населення відповідного віку – це означає, що переважна більшість захворювань (96 %) завершується одужанням. Небезпеку захворювань, їх агресивність для популяції визначає показник летальності, який в середньому становить 0,618, а життєва перспектива популяції, обтяженої захворюванням, характеризується коефіцієнтом виживання, який в різні роки коливається в межах від 138,9 до 198,7 умовних одиниць

Привертають увагу характерні взаємозв’язки між епідеміологічними показниками. Коефіцієнт кореляції між захворюваністю і поширеністю хвороби становить 0,93. Між поширеністю хвороби і смертністю взаємозв’язок відсутній (r=0,07), а між захворюваністю і смертністю дуже слабий і негативний (r=-0,19). Ці взаємини означають, що поширеність хвороби у високій мірі пропорційна захворюваності, проте смертність значною мірою не підпорядкована ані захворюваності, ані поширеності.

Таким чином, згідно із отриманих даних, було побудовано графік загальної кількості захворюваності станом на 27.09.2020 р.

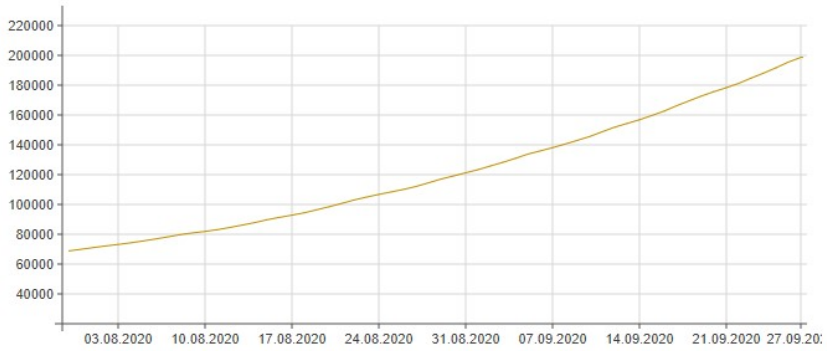


Рисунок 3.1 -Загальна кількість захворювань

Наступним не менш важливим кроком буде виділення кластеру летальних випадків захворювання. Таким чином, станом на 27.09.2020 р. було отримано наступний графік такої динаміки:

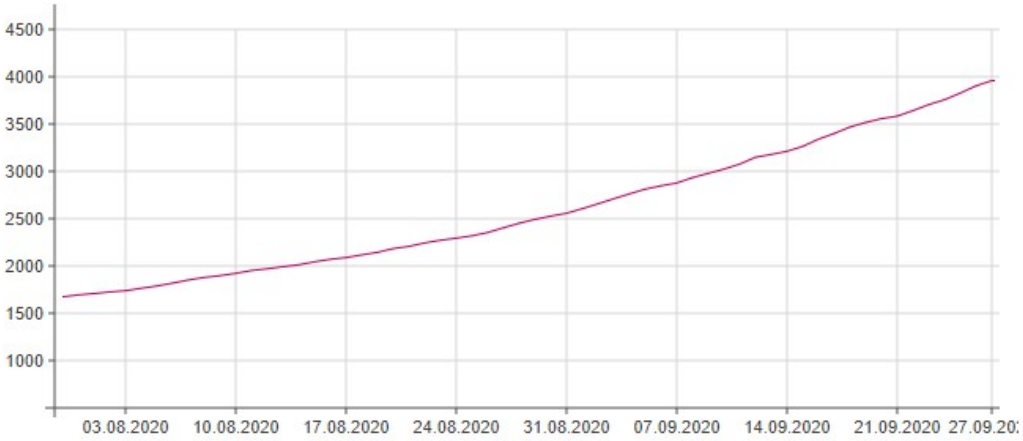


Рисунок 3.2 – Кількість летальних випадків

Наступний кластер дослідження епідеміологічної ситуації, буде складатися з кількості осіб, які одужали:

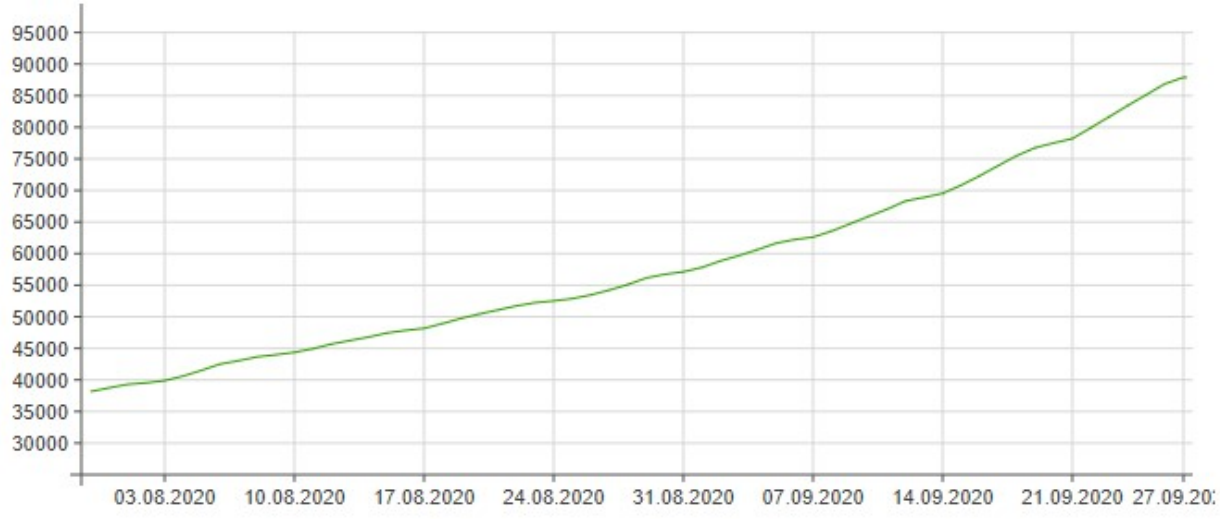


Рисунок 3.3 – Кількість осіб що одужали

Для наглядного порівняння співвідношення кількості одужавших осіб, осіб які хворіють зараз, та летальних випадків, наступним етапом проектування буде створено таблицю динаміки захворюваності.

Таблиця 3.1- Статистика динаміки епідеміологічної ситуації по областям

ОБЛАСТЬ	ЗАРАЖЕНИХ	СМЕРТЕЛЬНІ ВИПАДКИ	ОСОБИ ЯКІ ВИДУЖАЛИ	ЗАРАЗ ХВОРІЮТЬ
Вінницька	5900	113	3778	2009
Волинська	7632	162	5100	2370
Дніпровська	5285	106	2782	2397
Донецька	3016	56	1076	1884
Житомирська	5656	103	3162	2391
Закарпатська	9632	303	4484	4845
Запорізька	3663	58	1223	2382
Івано-Франківська	13335	303	6194	6838
Київська	8714	172	5346	3196
Кіровоградська	1074	56	844	174

Луганська	974	11	516	447
Львівська	19784	555	6625	12604
Миколаївська	2980	73	986	1921
Одеська	12876	187	2520	10169
Полтавська	1910	36	954	920
Рівненська	12032	161	9603	2268
Сумська	3672	63	1614	1995
Тернопільська	12986	157	6960	5869
Харківська	17364	331	4325	12708
Херсонська	1049	23	559	467
Хмельницька	5622	103	2351	3168
Черкаська	3694	53	1598	2043
Чернівецька	14009	352	7959	5698
Чернігівська	3960	61	841	3058
м. Київ	21815	361	6482	14972

Відомо, що перебіг і результат захворювання багато в чому залежать від якості медичного забезпечення та використаних методів лікування. Немає сумніву, що застосування сучасних методів лікування досвідченими лікарями здатне значно знизити смертність від захворювань. Виключно важлива в цьому відношенні якісна організація медичного обслуговування, яка може знизити небезпеку від цих хвороб у тій місцевості, на яку поширюється її вплив. Проте в Україні існує значна нерівність у наданні медичної допомоги інфекційним хворим, що може певною мірою бути причиною істотних відмінностей епідеміологічних показників у регіонах країни.

Однак зв’язати гальмування смертності тільки з особливостями медичної допомоги немає підстав - цьому суперечить нестабільність епідеміологічних показників у регіонах країни і, головне, той факт, що

стабілізація смертності та сприятливі зміни летальності реєструються в регіонах, де зростає поширеність захворювань.

Таким чином було досліджено динаміку розвитку епідеміологічного стану, зібрано та підготовлено необхідні дані для подальшого аналізу.

3.3 Розробка математичної моделі кластерного аналізу

Математична модель може бути подана у вигляді графіка, дендрограми, таблиці даних чи математичної формули. У даній роботі буде наведено дендрограми результатів кластерного аналізу а також відповідні графіки.

Аналіз динаміки захворюваності населення на коронавірусну хворобу дає змогу виділяти сучасні тенденції їхнього поширення, визначити регіональні зміни, що є важливим для здійснення типізації регіонів за рівнем захворюваністю населення.

Просторовий аналіз захворюваності спрямований на ідентифікацію процесів поширення захворюваності для цілей моніторингу поточної ситуації, прогнозування чи виявлення типічних моделей основуючись на вивчення динаміки поширенності епідемії. В просторовому аналіз широко застосовують методи багаторівневого моделювання, які дають змогу виявити вплив індивідуальних та контекстуальних факторів на здоров’я та динаміку захворюваності населення, визначити ризики захворювання людей з деякими ознаками, що проживають у визначених місцях. Використання методів аналізу а також моделювання значно посилюють інтерес до просторового аналізу в медичній географії.

Групування регіонів за показниками, що характеризують стан здоров’я населення, було здійснено із використанням методів кластеризації: рангів, побудови карт самоорганізації Кохонена. Щоб здійснити пошук тісноти зв’язку між соціально-економічними показниками, які характеризують вплив чинників на захворюваність населення, використано кореляційний аналіз”.

На основі результатів аналізу чинників регіональних відмінностей захворюваності, тенденцій динаміки, визначення просторових структур захворюваності запропонованою типізацію регіонів України.

Кластерний аналіз проводився у програмі Statistica агломеративним методом (complete linkage – повного зв’язку) та методом Варда.

У програмі STATISTICA реалізовані агломеративні методи мінімальної дисперсії - деревоподібна кластеризація і двухвходових кластеризація, а також дивізійний метод k-середніх. У методі деревовидної кластеризації передбачені різні правила ієрархічного об'єднання в кластери:

1. Правило одиночній зв'язку. На 1 кроці об'єднуються два найбільш близькі об'єкти, які мають максимальну міру подібності. На наступному кроці до них приєднують об'єкт із максимальною мірою схожості із одним з об'єктів кластера, а саме - для його включення в кластер потрібна максимальна схожість лише із 1-м об`єктом кластера. Метод називають ще й «методом близького сусіда», так як відстань між двома кластерами визначається як відстань між двома найбільш близькими об'єктами в різних кластерах. Це правило «нанизує» об'єкти для формування кластерів. Недолік даного методу - утворення дуже великих довгастих кластерів.

2. Правило повних зв'язків. Метод дає змогу усунути недолік, властивий методу одиночній зв'язку. Суть правила в тому, що “два об'єкти, що належать одній і тій же групі (кластеру), мають коефіцієнт подібності, який більше деякого порогового значення S201D. У термінах евклідова відстані це означає, що відстань між двома точками (об'єктами) кластера не повинно перевищувати деякого граничного значення d. Таким чином, d визначає максимально допустимий діаметр підмножини, утворює кластер. Цей метод називають ще методом найбільш віддалених сусідів, так як при досить великому пороговому значенні d відстань між кластерами визначається найбільшою відстанню між будь-якими двома об'єктами в різних кластерах.

3. Правило невиваженого попарного середнього. Відстань між двома кластерами визначається як «середня відстань між усіма об'єктами» у них. Метод досить ефективний, в тому числі - коли об'єкти в дійсності формують різноманітні групи, проте він працює добре і у випадках протяжних кластерів.

4. Правило зважене попарне середнє. Метод ідентичний попередньому, за винятком того, що при обчисленні розмір відповідних кластерів використовується в Як вагового коефіцієнта. Бажано цей метод використовувати, коли передбачаються різні розміри кластерів.

5. Невиважений центроїдний метод. Відстань між 2 кластерами визначається як «відстань між їх центрами тяжкості».

6. Зважений центроїдний метод. Ідентичний попередньому методу, але за винятком того, що при обчисленнях відстані використовують ваги для обліку різниці між розмірами кластерів. Саме тому, якщо є значні відмінності в розмірах кластерів, цей метод виявляється переважно попереднього.

7. Правило «Варда» (Уорда). У цьому методі в якості цільової функції застосовують «внутрішньо групову суму квадратів відхилень», яка є ні чим іншия, як сума кв. відстаней між кожною точкою і середньої по кластеру точкою, що містить даний об'єкт. “При кожному кроці об'єднують такі 2 кластери, що призводять до найменшого збільшення цільової функції, тобто внутрішньо групової суми квадратів відхилень. Цей метод направлений на об'єднання близько розташованих кластерів. Помічено, що метод Варда призводить до створення кластерів більш-менш рівних розмірів й мають форму гіперсферу”.

Раніше ми розглянули методи кластеризації об'єктів (спостережень), проте іноді кластеризація по змінним може привести до досить цікавих результатів. В модулі Кластерний аналіз також передбачена ефективна двухвходових процедура, яка дасть змогу кластеризувати відразу в двох напрямках - за спостереженнями і змінним.

Метод К-середніх

Припустимо, є гіпотези щодо числа m кластерів (по змінним або спостереженнями). Тоді можна програмно задати створити рівно н кластерів таким чином, щоб вони були настільки різні, наскільки це можливо. “Саме для вирішення завдань цього типу призначений метод k-means (k-середніх). Гіпотеза може ґрунтуватися на теоретичних міркуваннях, результати попередніх досліджень чи здогадки”. Виконуючи це послідовне розбиття на різне число кластерів, можливо порівнювати якість прийнятих рішень.

Програма розпочинає із Н випадково кількостей обраних кластерів, а потім змінює приналежність її об'єктів до цих кластерів, щоб «мінімізувати» мінливість всередині самих кластерів і «максимізувати» мінливість між цими кластерами. “Алгоритм рандомним чином у просторі призначає центри кластерів”. Далі обчислює відстань поміж центрами цих кластерів і кожним його об'єктом, і об'єкт приписується до того кластеру, до якому він є найближче. Завершивши цей крок, алгоритм обраховує середні значення для кожного кластера окремо. Усіх цих середніх буде стільки, скільки використовується змінних для аналізування, - К штук. Набір середніх являє собою координати нового знаходження центру кластера. Алгоритм ще раз вираховує відстань від кожного поточного об'єкта до цих центрів кластерів і приписує об'єкти до найближчих кластерів. Потім знову обчислюються центри тяжкості кластерів, і цей процес повторюється до тих пір, поки центри тяжіння не перестануть, так би мовити - мігрувати в просторі.

Якщо в деревовидній кластеризації можна використовувати категоріальні змінні, то так як в методі k-середніх як метрики використовують евклідову метрику, то перед проведенням кластеризації необхідно стандартизувати змінні. З цієї ж причини в методі передбачається, що змінні безперервні і виміряні як мінімум в інтервального шкалою.

Евклідова відстань - геометричне відстань в багатовимірному просторі. В нашому випадку це відстань між наборами показників (L1-A6) для кожного

підприємства і воно еквівалентно відстані між підприємствами відповідно обраними показниками. Чим менше відстань між об'єктами, тим вони більш схожі. Квадрат евклідової відстані використовують, якщо необхідно надати великі ваги більш віддаленим один від одного об'єктів.

Аналіз дендрограми кластеризації регіонів України за рівнем захворюваності населення COVID-19, отриманої методом Варда, дає підставу виділити шість кластерів регіонів:

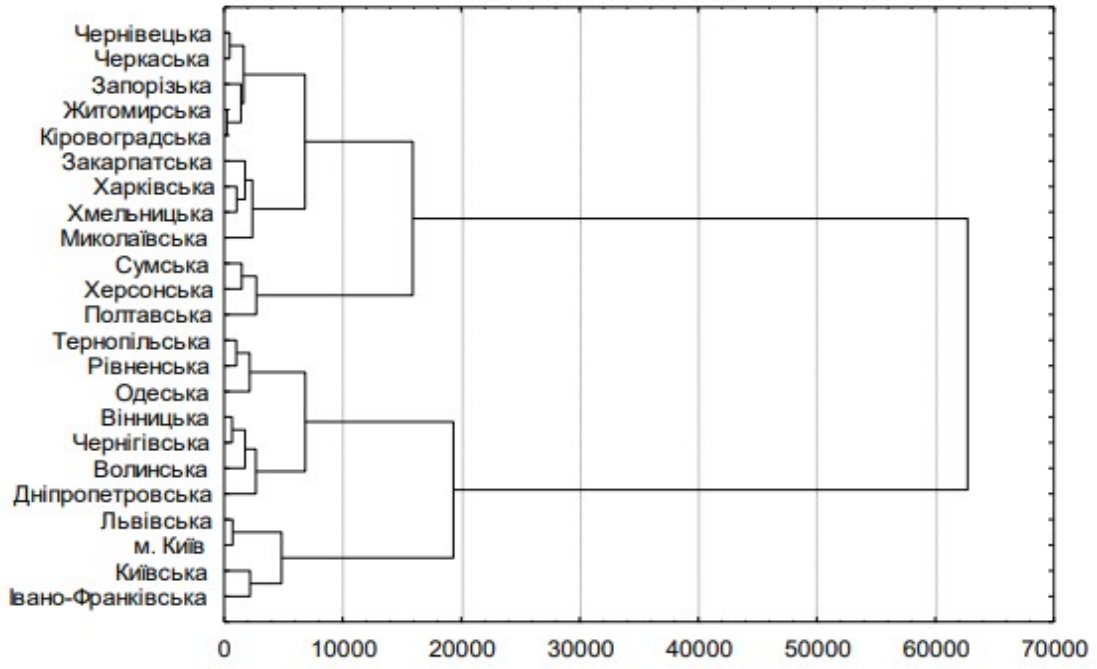


Рисунок 3.4 - Дендрограма регіонів України за показниками захворюваності населення (метод Уорда)

Кластеризація регіонів України за рівнем захворюваності населення, отримана методом повного зв’язку, також визначає виділення шести основнтих кластерів, “склад 3, 4, 5, та 6-го з яких є ідентичним кластерному аналізу, виконаному методом Варда, а склад першого та 2-го кластерів відрізняється лише на 1 регіональну одиницю”:

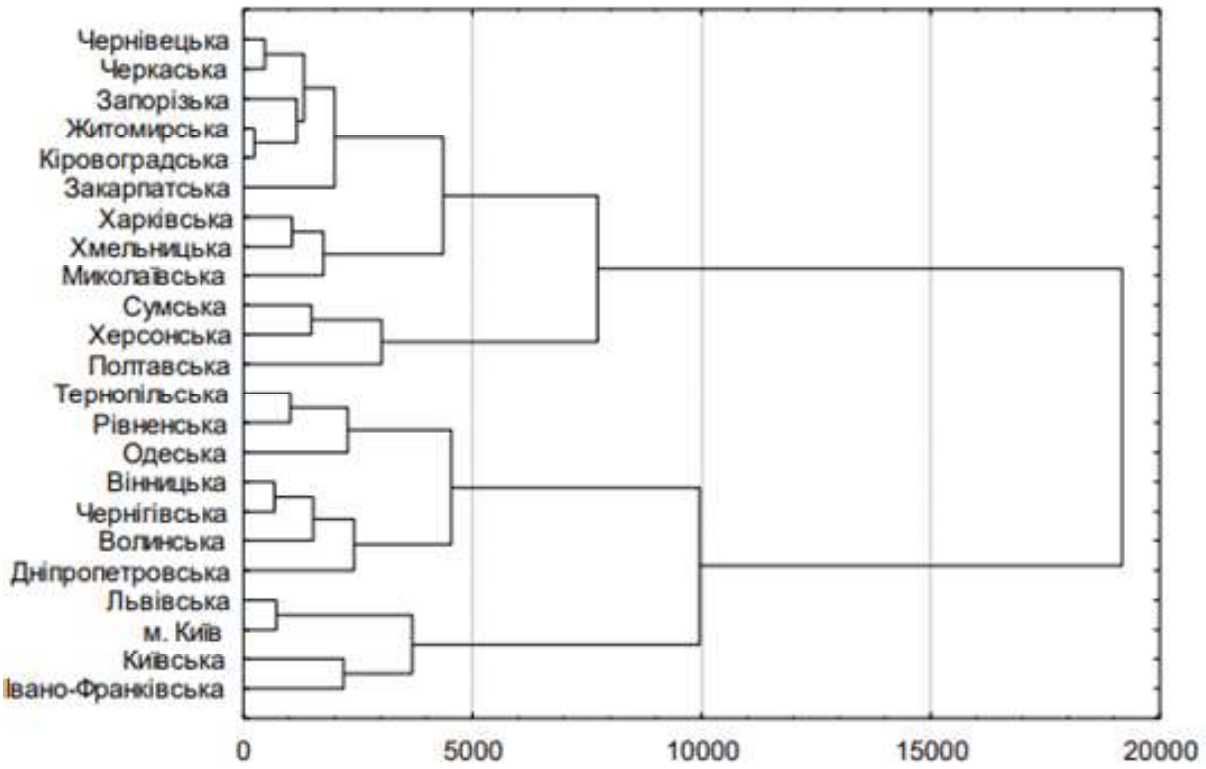


Рисунок 3.5 - Дендрограма регіонів України за показниками захворюваності населення (метод повного зв’язку)

Кластер 1 – Чернівецька, Черкаська, Кіровоградська, Запорізька та Житомирські області;

Кластер 2 – Миколаївська, Закарпатська, Хмельницька та Харківські області;

Кластер 3 – Полтавська, Сумська, Херсонська області;

Кластер 4 – Одеська, Тернопільська, Рівненська, області;

Кластер 5 – Чернігівська, Волинська, Вінницька та Дніпропетровські області;

Кластер 6 – Івано-Франківська, Львівська, Київська області та м. Київ.

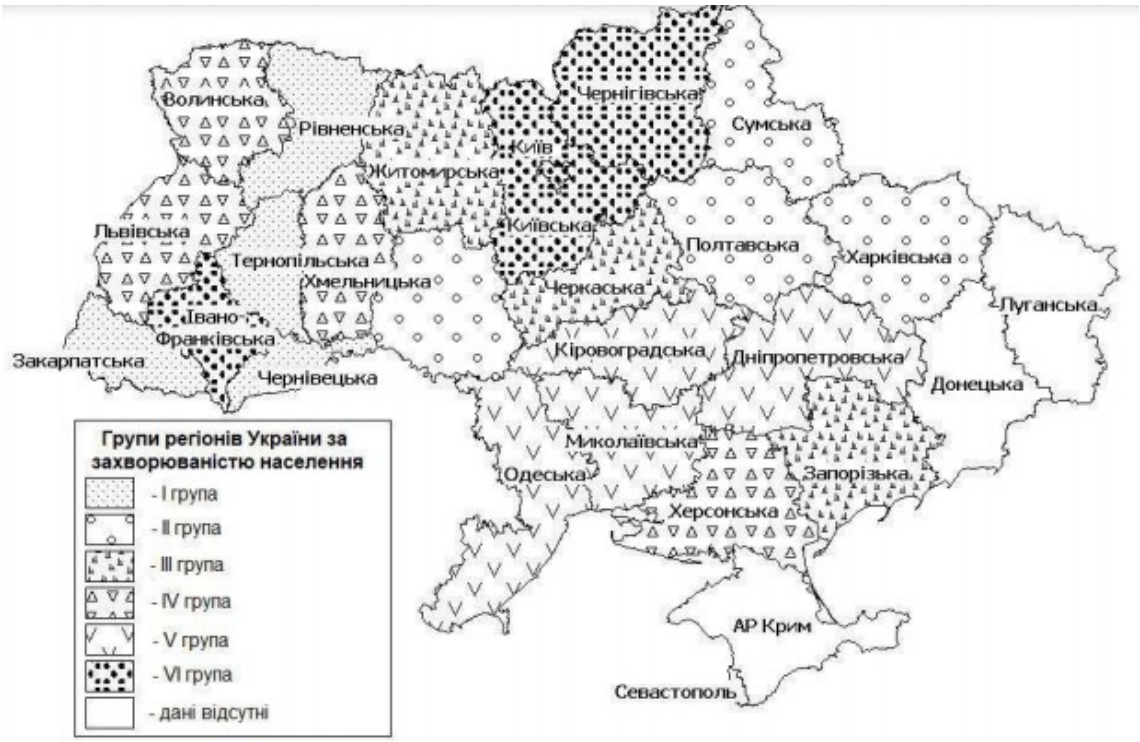


Рисунок 3.6 - Групи регіонів України за захворюваністю населення на окремі види хвороб (метод рангів)

За рівнем захворюваності населення ранговим методом доцільно виділяти п’ять груп регіонів

Група 1 – Закарпатська, Чернівецька, Тернопільська та Рівненська області;

Група 2 – Львівська, Волинська, Сумська, Хмельницька, Вінницька, Харківська та Полтавська області;

Група 3 – Житомирська, Запорізька, Чернігівська, Херсонська, Черкаська, Івано-Франківська області;

Група 4 – Київська область а також м. Київ;

Група 5 – Миколаївська, Кіровоградська, Одеська та Дніпропетровська області.

На основі проведених групувань та кластеризації регіонів України методом Варда, агломерованим методом, ранговим методом а також методом побудови карт самоорганізації Кохонена, аналізу динаміки різних видів поширення захворювання протягом вересня-жовтня місяця, та виявлення її

тенденцій доцільно виділяти такі типи регіонів за поширенням хвороби та рівнем захворюваності населення:

Тип 1 – Тернопільська, Рівненська, Закарпатська, Чернівецька області – для них характерні низькі та нижчі за середній показники поширення захворюваності населення на COVID-19 з переважанням позитивних динамік показників захворюваності населення за виключенням виникнення пневмонії;

Тип 2 – Хмельницька, Харківська, Вінницька та Сумська області – для яких характерні нижчі середніх показники захворюваності населення на COVID-19, середні значення захворюваності населення на пневмонію, та вище середнього рівня захворюваність з летальними випадками із переважанням позитивних показників динаміки захворюваності населення із виключенням стабільного росту захворюваності на легку форму пневмонії.

Тип 3 – Житомирська, Волинська, Полтавська, Чернігівська, Черкаська, Запорізька та Дніпропетровська області – характеризує значно нижче середнього рівня захворюваністю населення на COVID-19, середнім рівнем захворюваності на пневмонію та вище середнього рівнем захворюваності з летальними випадками з переважанням позитивних динамік показників захворюваності населення з виключенням захворюваності з летальними випадками.

Тип 4 – Івано-Франківська, Львівська, Київська області та м. Київ – для якого характерні середні та вище середнього показники захворюваності населення на коронавірус (з виключенням західних регіонів) і високими значеннями щодо показників поширення захворюваності населення на хвороби органів дихання із переважанням позитивних рис динаміки показників захворюваності населення за виключенням захворюваності з летальними випадками.

Тип 5 – Одеська, Кіровоградська, Миколаївська та Херсонська області – для них характерний середній рівень поширення захворюваності населення на хвороби органів дихання і високі рівні захворюваності населення з

летальними наслідками з переважанням негативних рис динаміки показників захворювання населення за виключенням захворюваності на двосторонню пневмонію.

Виходячи з цих статистичних даних, для порівняльного аналізу рівня захворюваності населення України на ковід з іншими країнами світу “репрезентативними є показники дитячої та материнської смертності, а також захворюваність та смертність населенності від коронавірусної хвороби”.

Підсумовуючи власні дослідження захворюваності населення України на COVID-19, отже зупинимось на основних проблемах та можливих шляхах їх вирішення. Проблеми захворюваності та ОЗ населення України доцільно об’єднати у 3 блоки. “1 блок формують такі демографічні проблеми як чинники захворюваності населення України: вік населення, регресивна вікова структура населення, високий рівень смертності населення, насамперед, чоловіків у працездатному віці, помірний рівень дитячої смертності, депопуляція населення”.

Наступній блок охоплює проблеми медичного обслуговування населення України: низьку фізичну (особливо в сільській місцевості) і економічну (для окремих категорій населення) доступність якісної медицини, нераціональну територіальну організацію систем надання медичної допомоги, недостатній рівень кваліфікованих фахівців для медичних закладів (насамперед, у сільській місцевості), неукомплектованість медичних закладів у сільських місцевостях, захворюваність і здоров’я населення в Україні: суспільно-географічний вимірюваність кадрами закладів медичного обслуговування, постаріння кадрів, неконкурентоспроможний рівень ЗП працівників мед. сфери, що впливає на мотивацію праці і якість надання медичних послуг, застарілість матеріальної технічної бази та неукомплектованість мед. установ сучасним обладнанням, приладами та лікарськими засобами, недостатнє та не досить ефективне фінансування медичної сфери. 3-й блок включає проблеми стану здоров’я населення України: високий рівень захворюваності і смертності населення від

коронавірусу, зростання загального рівня захворюваності, недостатню ефективність реалізації заходів протидії COVID-19, швидкі темпи поширеності та високий рівень захворюваності населення на ковід, низьку ефективність лікування, досить пізнє виявлення соціально вразливих захворювань, недостатню забезпеченість хворих необхідними медичними препаратами.

Пріоритетними напрямами у покращення медико-географічної ситуації у регіонах України можуть бути:

- забезпечення медичними закладами «охорони здоров'я» та фахівцями відповідно до потреб (насамперед, в сільській місцевості);
- покращення надання медичної допомоги насамперед - соціально вразливим класам населення;
- оптимізація територіальної структури медичної сфери, для підвищення територіальної доступності, якості «медичної допомоги» насамперед для сільських жителів, в тому числі децентралізації управління та створення територіальних громад;
- удосконалення та комплектація медичних закладів охорони здоров'я сучасною матеріально-технічною базою;
- постійні профілактичні огляди людей з метою виявлення різних видів захворюваності на ранніх стадіях розвитку;
- створення і реалізація «стратегії формування» здорового життя населення;
- розробка ефективної моделі координації роботи в приватних установах охорони здоров'я;
- проведення активної екополітики з метою покращення ситуації із якістю дезінфекції громадских місць, якістю питної води та поводження з відходами.

ВИСНОВКИ

За час роботи над дипломною роботою, дослідженням матеріально бази, дослідження проблематики, емпіричного дослідження та збору аналітичних даних я навчилася проведення науково-дослідної роботи та опрацювання методики, а саме: дослідила динаміку поширеності інфекції, зібрала аналітичні дані, провела аналіз цих даних та виявила основні методи поширення, підготувала дані для кластерного аналізу з метою виявлення основних факторів ризику поширення інфекції, поглибила теоретичні знання в сфері статистичного аналізування, моделювання методом кластерного аналізу, підібрала фактичний матеріал для написання випускної магістерської роботи.

Набула умінь і навичок опрацювання наукових та інформаційних джерел.

На основі аналізу предметної області і аналітичного огляду інформаційних джерел в галузі статистики уточнено тема ДМР: Застосування кластерного аналізу для визначення факторів ризику інфекційних захворювань COVID-19.

Визначено об’єкт, предмет дослідження: група інфікованих осіб, кількість приросту захворювань;

Сформульовано мету ДМР: визначити фактори ризику інфекційних захворювань CoVID-19, застосовуючи методи кластерного аналізу, та завдання: розробити модель на основі результатів кластерного аналізу.

В результаті було досліджено проблематику поширення захворюваності, створено моделі кластерного аналізу, визначено фактори ризику, та запропоновано можливі варіанти рішення, висвітлено основні проблематики, та способи їхнього вирішення.

Перелік джерел посилання

1. Електроніка та інформаційні технології. 2015. Випуск 5. С. 102–113 Electronics and information technologies. 2015. Issue 5. P. 102–113

2. Дзендзелюк О. Автоматизована система моніторингу параметрів довкілля / О.Дзендзелюк, О. Дзелюк, І. Мусійчук, В. Рабик // Теор. електротехніка. – 2010. – Вип. 61. – С. 90–98.

3. Lin Feng. Time Series Forecasting with Neural Networks / Lin Feng, Yu Xing Huo, Gregor Shirley // J. of Complexity International 2. 2015. [Електронний ресурс]. – Режим доступу: http://journal-ci.ccse.monash.edu.au/ci/vol02/cmxhk/cmxhk.htm

4. Riedmiller M. Rprop – Description and Implementation Details [Електронний ресурс] / M. Riedmiller // Technical Report. – 2011. Janeary.. – Режим доступу: www.inf.fuberlin.de/lehre/ WS06/Mustererkennung/Paper/rprop.pdf

6. Гамбаров Г.М. Статистичне моделювання та прогнозування / Г.М.Гамбаров, Н.М. Журавель, Ю.Г. Королев. – М.: С., 2010. – 140 с.

6. Єжов А. А. Комп`ютеринг та прогнозування / А. А. Єжов, С. А. Шумский. – М.: МІФІ, 1998. – 222 с.

7. Галушкін А. И. Теорія прогнозування. Кн.1: Навчальний посібник для вишів. / А. И. Галушкін. // Видавницьке підприємство редакції журналу «Прогноз». – 2010. – С.215

8. Коронавірус в Україні Офіційний інформаційний портал Кабінету Міністрів України— Режим доступу: https://covid19.gov.ua/

9. Хайкин С. Моделювання динамічних структур / С. Хайкин. – М. : Вильямс, 2016. – 1103 с.

10. Мицель А. А. Прогнозирование динамических структур / А. А. Мицель, Е.А. Ефремова. // Известия Томского политехнического университета. – 2017. – №8.

11. Андриенко В. М. Анализ и моделирование динамики / В. М. Андриенко, А. Ш. Тулякова. // Научный журнал «Аспект». – 2011. – №2. – С. 34.

12. Иванов Д. В. Прогнозирование с использованием искусственных нейронных сетей [Електронний ресурс] / Д. В. Иванов – Режим доступу до ресурсу: forex-mmcis.ru./D.Ivanov.

13. Войтенко, В. П., Писарук, А. В., Кошель, Н. М. (2013). Смертність внаслідок інфекційних хвороб населення у містах та сільській місцевості України: медико-демографічні та соціальні аспекти. Пробл. старения и долголетия, (2), 185-201.

14. Gundarow, I. (2013). Noninfectious mechanisms of infectious epidemics. Zdrowie i Społeczeństwo, 3(1), 11-16.

15. Андрейчин, М. А. (2010). Відкриття збудників інфекційних хвороб: сучасні досягнення і перспектива. Нобелівський рух і Україна: Збірник праць Тернопільського осередку Наукового товариства ім. Т. Шевченка, 5. (с. 204-223). Тернопіль: Джура.

16. Андрейчин, М. А. (2012). Медична допомога інфекційним хворим в Україні: проблеми і шляхи їх розв'язання. Інфекційні хвороби, (1), 5-7.

17.Industrial internet reference architecture [Online]. Available: http://www.iiconsortium.org/IIRA.htm.

18. "Implementation Strategy Plattform Industrie 4.0 Results Report", Bitkom e.V., VDMA e.V., ZVEI e.V., Berlin, January 2016.

19. D. Barry. Azure/iot-edge [Online]. Available: https://github.com/Azure/iot-edge.

20. Справочник по математике для экономистов / В.Е. Барбаумов, В.И. Ермаков, Н.Н. Кривенцова и др.; под ред. В.И. Ермакова. 2-е изд., перераб. И доп. М.: Высшая школа, 1997. 384 с.

Додаток А

Наукова теза

Міністерство освіти і науки України
Хмельницький національний університет

ЗБІРНИК НАУКОВИХ ПРАЦЬ
за матеріалами XII всеукраїнської науково-практичної конференції
«Актуальні проблеми комп'ютерних наук АПКН-2020»

9-10 листопада 2020

Хмельницький 2020

УДК 004.4

Варгата В. Ю.

Хмельницький національний університет

ЗАСТОСУВАННЯ КЛАСТЕРНОГО АНАЛІЗУ ДЛЯ ВИЗНАЧЕННЯ ФАКТОРІВ РИЗИКУ ІНФЕКЦІЙНИХ ЗАХВОРЮВАНЬ COVID-19

Розглянуто метод кластерного аналізу та його застосування у сфері дослідження епідеміологічної ситуації. Було досліджено динаміку поширення вірусу COVID-19 на території України, зібрано та проаналізовані актуальні дані, розглянуто епідеміологічний стан України та окремих регіонів. Зібрані дані було досліджені та проаналізовані.

The method of cluster analysis and its application in the field of epidemiological research is considered. The dynamics of the spread of the COVID-19 virus on the territory of Ukraine was studied, current data were collected and analyzed, and the epidemiological situation in Ukraine and individual regions was considered. The collected data were researched and analyzed.

2020 рік надовго відіб`ється в пам`яті людства роком пандемії нового віруса. 31 грудня 2019 року Всесвітнє організація охорони здоров'я була проінформована про виявлення випадків пневмонії, викликаної невідомим збудником. Системи охорони здоров'я можуть виявитися не готовими до надзвичайно великої кількості тяжкохворих пацієнтів. Найбільш важливою відповіддю по відношенню до інфекції не є лікувальні заходи, а зниження швидкості її поширення, щоб розтягнути її в часі і знизити, таким чином, навантаження на системи охорони здоров'я.[1]

Для того, щоб ефективно знизити швидкість поширення вірусної хвороби, та звести небезпеку пандемії до мінімуму, важливим етапом є правильне дослідження та аналіз динаміки поширення захворюваності в Україні, та окремих регіонах. Згідно результатів аналізу та дослідження, стає можливим впровадження ефективних карантинних умов, у кожному регіоні окремо, що в результаті, теоретично, приведе до зниження поширення захворюваності. Ефективним методом аналізу – є метод, так званого «Кластерного аналізу». Таким чином, актуальність розробки даної роботи набуває чинності, оскільки, на даний час, дуже важливо правильно спланувати методи впровадження карантинних обмежень.

Кластерний аналіз (англ. Cluster analysis) - багатовимірна статистична процедура, що виконує збір даних, що містять інформацію про вибірку об'єктів, і потім впорядковує об'єкти в порівняно однорідні групи. Завдання кластеризації відноситься до статистичної обробки, а також до широкого класу задач навчання без учителя [2].

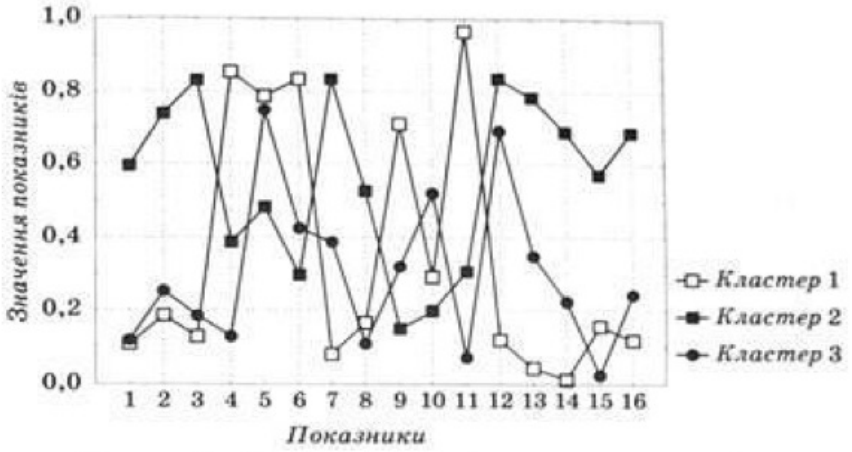


Рисунок 1 – Графік результатів кластерного аналізу

Аналізу були піддані показники, що характеризують захворюваність і поширеність захворювання, а також смертність від них у регіонах України.

Для оцінки співвідношення між поширеністю захворювань і смертністю від них у щорічних епідеміологічних показниках був використаний вибірковий метод виділення регіонів, що характеризують субпопуляції з малою, середньою і високою поширеністю захворювання. Співвідношення розповсюдженості хвороб і смертності від них оцінювалось у методичному варіанті, який передбачав оцінку регіонів «за поширеністю хвороби» в поточному році без урахування значень цього показника в попередні роки. Епідеміологічні показники протягом 6 місяців свідчать про значну мінливість, якій не притаманна тенденція до зростання чи спаду. Показники, особливо захворюваність, коливаються. Різниця між найбільшими і найменшими показниками поширеності хвороби і захворюваності в різних роках становить відповідно 26,4 і 34,5 %. Щорічна захворюваність складає майже 9/10 (точніше, 87,7 %) від усієї поширеності хвороби.

Смертність відносно значна – в середньому 4,00 на 100 тис. населення відповідного віку – це означає, що переважна більшість захворювань (96 %) завершується одужанням. Небезпеку захворювань, їх агресивність для популяції визначає показник летальності, який в середньому становить 0,618, а життєва перспектива популяції, обтяженої захворюванням, характеризується коефіцієнтом виживання, який в різні роки коливається в межах від 138,9 до 198,7 умовних одиниць.

Привертають увагу характерні взаємозв'язки між епідеміологічними показниками. Коефіцієнт кореляції між захворюваністю і поширеністю хвороби становить 0,93. Між поширеністю хвороби і смертністю взаємозв'язок відсутній (r=0,07), а між захворюваністю і смертністю дуже слабий і негативний (r=-0,18).

Згідно з отриманих даних, було побудовано графік загальної кількості захворюваності станом на 27.09.2020 р.

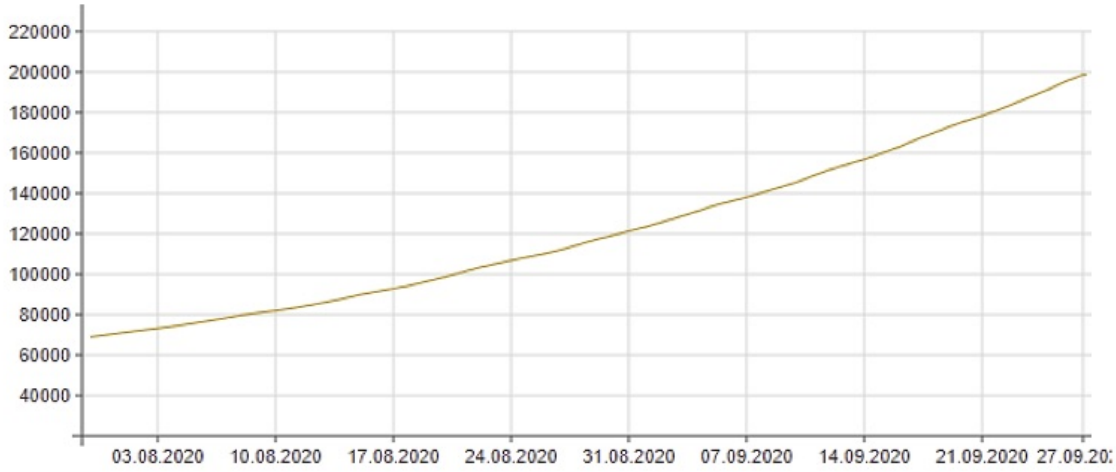


Рисунок 2 – Графік динаміки захворюваності

Таким чином, було зібрано та проаналізовано актуальні дані статистики захворюваності. Наступним етапом в розробці даного проекту буде нормалізація отриманих даних, та створення моделі динаміки згідно результатів «Кластерного аналізу».

Перелік посилань

1. Коронавірус в Україні Офіційний інформаційний портал Кабінету Міністрів України— Режим доступу: https://covid19.gov.ua/
2. Гамбаров Г.М. Статистичне моделювання та прогнозування / Г.М.Гамбаров, Н.М. Журавель, Ю.Г. Королев. – М.: С., 1990. – 140 с.

Додаток Б

Презентаційні матеріали

Постановка проблематики

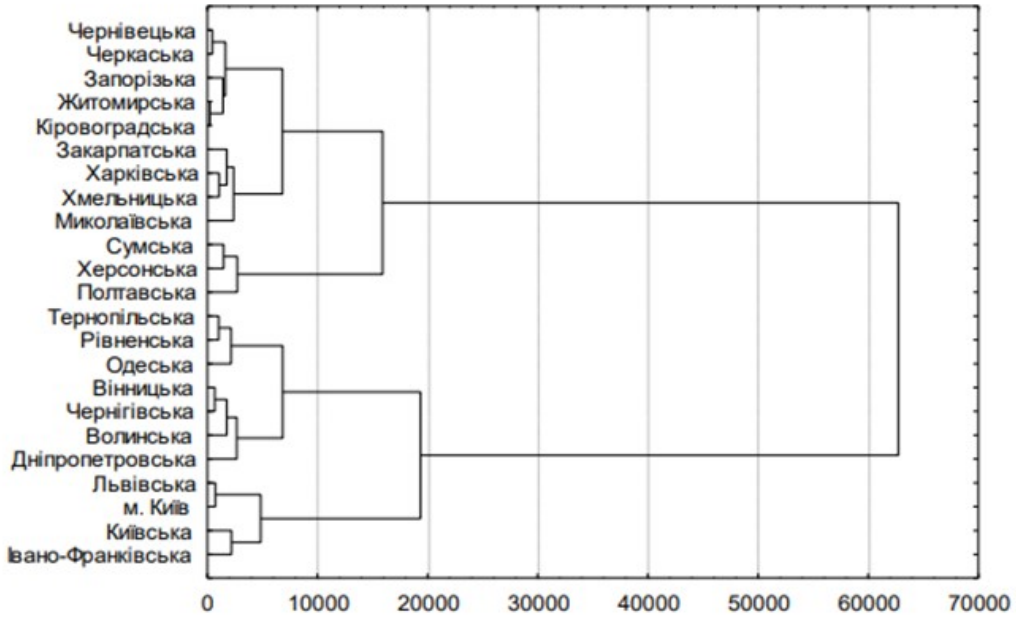
- Об`єктом дослідження, є фактори які впливають на динаміку поширення захворюванності
- Предмет дослідження: дані про кількість інфікованих осіб, та приросту захворювань;

Мета роботи

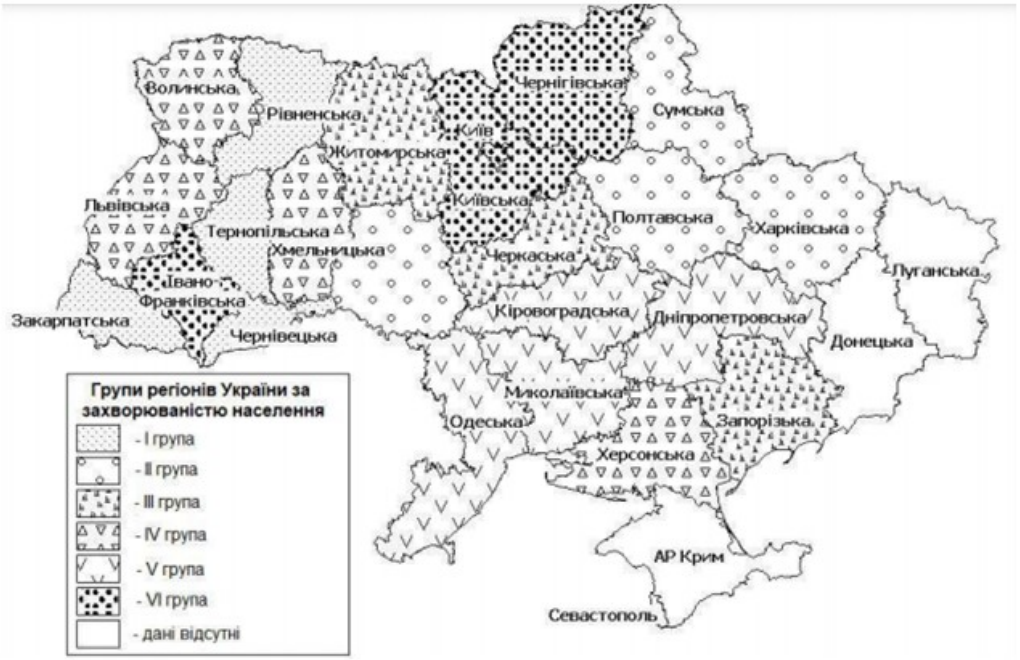
Основна мета даної роботи полягає у визначенні факторів ризику інфекційних захворювань CoVID-19, застосовуючи методи кластерного аналізу

Завдання : розробити моделей кластерного аналізу методами Варда, рангів та методом повного зв`язку на основі результатів кластерного аналізу.

Дендрограма кластеризації регіонів України за показниками захворюваності населення (метод Варда)



Карта поширення захворюванності в окремих регіонах. Поділ регіонів на групи ризику згідно результатів отриманих кластеризацією методом рангів



Новизна розробки

- Новизна розробки: Вперше було отримано моделі кластерного аналізу, згідно яких було виділено фактори ризику, основні причини та методи вирішення проблеми поширення захворюванності.
- Виділено групи ризику, регіони з найбільшою поширеністю, та запропоновано методи протидії.

Висновки

- Було проведено збір статистичних даних та їх нормалізація
- Проведено аналіз та підготовка до кластеризації
- Проведено кластерний аналіз, виявлено групи ризику, фактори поширення та можливі методи боротби з епідеміологією.

ДОДАТОК В
(ОБОВ`ЯЗКОВИЙ)
АНТИПЛАГІАТ

Mon Nov 23 13:08:24 EET 2020, Стецюк Віктор Іванович, Хмельницький національний університет, ХНУ

Anti-Plagiarism v-15.257

Максимальное совпадение с одним документом 4.0%

Словари проверки: en_US, ru_RU, ua_UA. **Ошибок в документах: 12%**

	Документ		Суммарное совпадение по Базе Данных	
ID: 80930 Название: Застосування кластерного аналізу для визначення факторів ризику інфекційних захворювань COVID-19 Добавлено в БД: 2020-11-23 Авторы: Варгата Вікторія Юріївна Руководители: Грипинська Надія Василівна Консультанты: Оппоненты:	Символы	Лексемы	Символы	Лексемы
	123450	929	12162 (10%)	114 (12%)

Источник плагиата

ID	Описание	Наличие плагиата в документе	
		Символы	Лексемы

UNICHECK

Имя пользователя:
Kafedra TMIT KhNU

ID проверки:
1005450183

Дата проверки:
14.12.2020 11:37:49 EET

Тип проверки:
Doc vs Internet

Дата отчета:
14.12.2020 11:54:31 EET

ID пользователя:
100005657

Название файла: Варгата_ПМм-19-1(повторно2)

Количество страниц: 84 Количество слов: 17033 Количество символов: 129013 Размер файла: 2.05 MB ID файла: 1005740511

542 слова помечены как "исключенные" и не учитываются в подсчете слов

5.47% Совпадения

Наибольшее совпадение: 0.99% с Интернет-источником (http://dspace.nuft.edu.ua/jspui/bitstream/123456789/26960/1/..

5.47% Источники из Интернета 128 ..... Страница 86

Поиск совпадений с Библиотекой не производился

4.07% Цитат

Цитаты 13 ..... Страница 87

Не найдено ни одной ссылки

0.05% Исключений

Некоторые источники исключены автоматически (фильтры исключения: количество найденных слов меньш...

0.05% Исключений из Интернета 6 ..... Страница 88

Нет исключенных библиотечных источников

Модификации

Обнаружены модификации текста. Подробная информация доступна в онлайн-отчете.

Замененные символы 12

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

РЕЦЕНЗІЯ НА ДИПЛОМНУ РОБОТУ

Дипломник Варгата Вікторія Юріївна

Тема Застосування кластерного аналізу для визначення факторів ризику інфекційних захворювань COVID-19

Спеціальність 113 – Прикладна математика

Обсяг дипломної роботи:

Кількість листів креслень 12; кількість сторінок записки 86

1.Короткий зміст ДР та прийнятих рішень Проаналізовано предметну область та джерела інформації та статистичні дані. Розроблено гіпотезу досліджень. Розглянуто та обрано методи кластерного аналізу в сфері визначення факторів ризику. Розроблено моделі кластерного аналізу та запропоновано найефективніші методи протидії поширенню пандемії;

2. Висновок про відповідність ДР поставленому завданню Дипломна робота виконана у повному обсязі. Розроблені моделі кластерного аналізу можуть бути запропоновані для подальшого використання у процесі дослідження факторів ризику.

3. Характеристика виконання кожного розділу роботи, ступінь використання останніх досягнень науки і техніки і передових методів роботи: В першому розділі аналізу предметної області було досліджено динаміку поширеності захворювання, розглянуто джерела статистичної інформації та розроблено гіпотезу наукового дослідження. В розділі чітко наведено показники, які впливають на динаміку поширення. В другому розділі було проведено підготовку аналітичних даних до проведення кластерного аналізу та подальшого визначення факторів ризику. Було відібрано найактуальніші, на момент розробки дані, побудовано карту, яка дає змогу візуально оцінити ступінь поширеності захворювання, розглянуто найактуальніші методи кластерного аналізу. В третьому розділі було побудовано моделі кластерного аналізу методами Варда, повного зв'язку та методом рангів. Отримані результати висвітлили найбільш небезпечніші регіони, допомогли визначити фактори ризику – в результаті чого було запропоновано методи протидії.

4. Позитивні сторони роботи: Тематика даної роботи на сьогоднішній день актуальна. Дана робота чітко висвітлює основні фактори ризику, дозволяє візуально оцінити поточний епідеміологічний стан, використовує ефективні методи аналізу та моделювання, та допомагає прийняти найефективніше рішення щодо застосування методів протидії поширенності пандемії.

5. Негативні сторони роботи ефективність запропонованих методів не було досліджено на практиці, необхідна подальша перевірка та коригування даних згідно отриманих результатів.
6. Оцінка графічного оформлення та пояснювальної записки роботи графічний матеріал відповідає результатам аналізу. Графіки та карти чіткі та зрозумілі. Проте кількість графічного матеріалу могла би бути збільшеною, за рахунок поділу аналізу на декілька етапів, що в результаті могло б збільшити точність отриманих даних.
7. Відгук про роботу в цілому робота виконана в повному обсязі, моделі кластерного аналізу допомагають чітко висвітлити основні фактори поширення. В цілому результат можна оцінити як «позитивний»
8. Інші зауваження Відсутні
9. Оцінка дипломної роботи добре
РЕЦЕНЗЕНТ (прізвище, ім'я, по-батькові, посада, місце роботи) Діхтярук Микола Миколайович, кандидат фізико-математичних наук, доцент, доцент кафедри вищої математики та комп'ютерних застосувань

"30" листопада 2020 р. _____
(підпис)

Завідувачу кафедри ТМІТ
д-р.техн.наук Підченку С.К.

Вартова В.Ю.
ПІБ здобувача вищої освіти

ФПКТС, 2 курсу, групи ПМм-19-1

ЗАЯВА

З правилами чинного Положення «Про дотримання академічної доброчесності в Хмельницькому національному університеті» від 26.09.2020 (зі змінами від 26.11.2020), згідно з яким виявлення плагіату є підставою для відмови в допуску кваліфікаційної роботи до захисту та застосування заходів дисциплінарної та академічної відповідальності, ознайомлений (а). Про використання програмно-технічних засобів для перевірки кваліфікаційних робіт здобувачів вищої освіти на плагіатоповіщений (а) та надаю свою згоду на обробку та збереження університетом моєї роботи в інституційному репозитарії університету.

Також надаю університету право на передачу моєї роботи для обробки та збереження в базах даних програмно-технічних засобів(Unicheck та Anti-Plagiarism) та використання роботи для виявлення плагіату в інших роботах, які перевіряються програмно-технічними засобами та користувачами, що мають доступ до цих програмно-технічних засобів, виключно в обмежених цілях для виявлення плагіату в текстах робіт.

Робота для перевірки університетом надається в друкованому та електронному варіанті. Електронна версія моєї роботи збігається (ідентична) з друкованою.

03.12.2020р
дата

підпис

РІШЕННЯ ЕКСПЕРНОЇ КОМІСІЇ

КАФЕДРИ ТЕЛЕКОМУНІКАЦІЙ, МЕДІЙНИХ ТА ІНТЕЛЕКТУАЛЬНИХ ТЕХНОЛОГІЙ

ПРО ДОПУСК КВАЛІФІКАЦІЙНОЇ РОБОТИ ДО ЗАХИСТУ

Підтверджуємо ознайомлення з результатом звіту подібності щодо роботи, генерованого системою виявлення текстових збігів/ідентичності/схожості:

Назва: Застосування кластерного аналізу для визначення факторів ризику інфекційних захворювань COVID-19

Автор: Варгата Вікторія Юріївна

Спеціальність: 113 – прикладна математика

Освітня програма: освітньо-професійна

Науковий керівник: Грипинська Надія Василівна, к.ф.-м.н., доцент

Після аналізу звіту подібності зроблено такий висновок:

№	Висновок	Позначка про відповідність
1	Запозичення, виявлені в роботі, є законними і не є плагіатом. Робота приймається до захисту.	+
2	Виявлені запозичення не є плагіатом, розміщені в розділах, які не описують безпосередньо авторське дослідження, але кількість цитат перевищує обсяг, виправданий поставленою метою роботи Робота приймається до захисту, але має бути відкоригована. Відкоригований варіант має бути поданий на кафедру за 2 дні до захисту, разом із заявою щодо самостійності виконання письмової роботи та ідентичності друкованої та електронної версії роботи	
3	Виявлені запозичення не є плагіатом, але частково розміщені в розділах, які описують безпосередньо авторське дослідження, а кількість цитат перевищує обсяг, виправданий поставленою метою роботи. В зв'язку з цим мета роботи та поставлені завдання не були досягнені. Робота може бути допущена до захисту (наступного року) після того як буде відкоригована та допрацьована і успішно пройде повторну перевірку на академічний плагіат.	
4	Робота містить навмисні текстові спотворення, передбачувані спроби укриття запозичень або інші прояви академічного плагіату. Робота містить фабрикацію або фальсифікацію даних. Робота не допускається до захисту.	

Підтвердження:

Запозичення, виявлені в роботі, є законними і не є плагіатом, оскільки:

1) усі запозичення розміщені в розділах, які не описують безпосередньо авторське дослідження і не стосуються результатів роботи;
2) переважно запозичення є цитатами джерел літератури, що стосуються теми кваліфікаційної роботи, та мають належним чином оформленні посилання;
3) джерело з яким виявлено найбільше співпадінь є належним чином процитоване у роботі та за яким проведено порівняльний аналіз методів;
4) також частина запозичень є сталими виразами кваліфікаційної дипломної роботи.

14.12.20
Дата

Підпис

Підпис