МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВАДИМА ГЕТЬМАНА

Навчально-науковий інститут
«Інститут інформаційних технологій в економіці»

Кафедра математичного моделювання та статистики

Освітньо-професійна програма «Економічна кібернетика» Галузь знань 05 «Соціальні та поведінкові науки» Спеціальність 051 «Економіка»

Форма навчання: дистанційна

КВАЛІФІКАЦІЙНА БАКАЛАВРСЬКА РОБОТА

на тему **«Аналіз і моделювання демографічних процесів у країнах Азії»**

здобувача Щирби Олександри Олександрівни

Науковий керівник: кандидат економічних наук, доцент

Білик Тетяна Олександрівна

_____
(підпис)

Робота допущена до захисту перед екзаменаційною комісією з атестації здобувачів вищої освіти (ЕК)

В. о. завідувача кафедри: кандидат фізико-математичних наук, професор Великоіваненко Г.І.

(підпис)

Київ 2024

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ

ІМЕНІ ВАДИМА ГЕТЬМАНА

Навчально-науковий інститут
Інститут інформаційних технологій в економіці
Кафедра математичного моделювання та статистики

Освітньо-професійна програма	**Економічна кібернетика**
Галузь знань	**05 Соціальні та поведінкові науки**
Спеціальність	**051 Економіка**

ПОГОДЖЕНО
Керівник проектної групи (гарант)
освітньо-професійної програми

О.М. Притоманова
(підпис)
_____ 20___ р.

ЗАТВЕРДЖУЮ
В.о. завідувача кафедри
математичного моделювання та
статистики
Г.І. Великоіваненко
(підпис)
_____ 20___ р.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ

здобувача вищої освіти __Щирби Олександри Олександрівни_____
(прізвище, ім'я, по батькові)
_____дистанційної_____ **форми навчання**
очної (денної), дистанційної

на підготовку кваліфікаційної бакалаврської роботи

на тему **«**Аналіз і моделювання демографічних процесів у країнах Азії**»**

Тему затверджено наказом ректора Університету від «_____» _____ 20___ р .№ _____

Кваліфікаційна бакалаврська робота виконується на матеріалах
статистичних даних демографічної організації Worldometer
https://www.worldometers.info/ та інших відкритих джерел

План кваліфікаційної бакалаврської роботи

Розділ 1	Основи демографії та демографічні тенденції у країнах Азії
(назва розділу)	
Розділ 2	Моделювання та аналіз індикаторів демографічних показників
(назва розділу)	

Об'єкт дослідження:	демографічні процеси в країнах Азії
Предмет дослідження:	методи та інструментарій моделювання демографічних процесів в країнах Азії
Мета кваліфікаційної бакалаврської роботи:	аналіз і моделювання демографічних процесів в Азії, зокрема, дослідження впливу народжуваності, смертності, міграції та інших факторів на чисельність і структуру населення

Конкретні завдання, які здобувач повинен виконати для досягнення поставленої мети:

У розділі 1	Здійснити критичний аналіз джерел, присвячених проблемам, особливостям та перспективам демографічного розвитку країн Азії. Оцінити вплив демографічних змін на соціально-економічний розвиток регіону.
У розділі 2	Провести аналіз множинної лінійної моделі, що описує демографічні показники країн Азії за певний період. Використати метод найменших квадратів (МНК) для побудови регресійної моделі. Виконати перевірку моделі за допомогою тестів Дарбіна-Уотсона, Бройша-Пагана та Шапіро-Вілка. Здійснити ієрархічний кластерний аналіз для виявлення груп країн з подібними демографічними характеристиками.

Завдання підготував науковий керівник _____ _____

(підпис) *(ініціали, прізвище)*

«_____» _____ 20___ р.

Завдання одержав здобувач _____

(підпис) *(ініціали, прізвище)*

«_____» _____ 20___ р.

Реферат

Кваліфікаційна бакалаврська робота містить 53 сторінки, 1 таблицю, 32 рисунки, перелік використаних джерел з 31 найменувань, додатки.

«Аналіз і моделювання демографічних процесів у країнах Азії»

Об'єктом дослідження є демографічні процеси в країнах Азії.

Предметом дослідження є методи та інструментарій моделювання демографічних процесів в країнах Азії

Мета та завдання виконання дослідження. Метою виконання кваліфікаційної бакалаврської роботи є аналіз і моделювання демографічних процесів в Азії, зокрема, дослідження впливу народжуваності, смертності, міграції та інших факторів на чисельність і структуру населення.

Відповідно до поставленої мети визначені такі *завдання*:

- розкрити сутність демографічних чинників як об’єктів для моделювання та аналізу;
- розкрити тенденції демографічних процесів в країнах Азії за останні десятиліття;
- розробити та проаналізувати множинну лінійну модель для прогнозування демографічних процесів у країнах Азії;
- провести кластерний аналіз на основі демографічних показників країн Азії;
- оцінити результати дослідження та їх вплив на демографічну ситуацію в країнах Азії.

Теоретична, методична та практична значущість отриманих результатів Теоретична, методична та практична значущість отриманих результатів. Під час дослідження систематизовано різноманітні підходи до аналізу та моделювання демографічних процесів, а також проаналізовано їх недоліки та переваги. Застосовано економетричні методи для оцінки параметрів демографічних моделей, включаючи регресійний аналіз, кластеризацію та інші статистичні методи. Проведено аналіз і моделювання демографічних показників країн Азії, що дозволило виявити основні тенденції та фактори, що впливають на демографічні процеси. Практичне значення отриманих результатів полягає в можливості їх використання для вдосконалення демографічної політики країн Азії, а також для розробки стратегій соціально-економічного розвитку, які враховують демографічні особливості регіону. Результати дослідження можуть сприяти підвищенню ефективності управління демографічними процесами та розробці ефективних державних політик. Отримані висновки та рекомендації можуть бути застосовані для аналізу демографічних процесів в інших регіонах світу, що забезпечить більш збалансований соціально-економічний розвиток та стабільність у довгостроковій перспективі.

Рік виконання кваліфікаційної бакалаврської роботи – 2024.

Рік захисту роботи – 2024.

Ключові слова: демографічні процеси, країни Азії, економетричне моделювання, регресійний аналіз, кластерний аналіз, соціально-економічний розвиток.

В і д г у к
про кваліфікаційну бакалаврську роботу
здобувача освітньо-професійної програми «Економічна кібернетика»
Щирби Олександри Олександрівни
на тему: «Аналіз та моделювання демографічних процесів в країнах Азії»

1. Актуальність теми: Важливою передумовою економічного, інтелектуального, політичного розвитку будь-якої країни є демографічний фактор. Актуальність використання економіко-математичних моделей та методів з метою вивчення демографії пов'язана з необхідністю вивчення популяційних та міграційних процесів для розробки ефективної політики, яка сприятиме стабільному розвитку суспільства.
2. Позитивні риси кваліфікаційної бакалаврської роботи: Індивідуальне завдання виконане здобувачем у повному обсязі. Якісно використані моделі та методи дослідження дозволили провести комплексний аналіз демографічних процесів в кранах Азії, досягти мети роботи, сформулювати обґрунтовані висновки та надати рекомендації.
3. Наявність самостійних розробок автора: побудована економетрична модель залежності кількості населення країн Азії від земельної площі, чистої зміни населення та індексу народжуваності та проведено її аналіз, за допомогою методів кластерного аналізу виконано групування та класифікацію країн Азії за вказаними чинниками
4. Цінність теоретичних висновків та практичних рекомендацій полягає у тому, що запропоновані моделі та методи дозволяють ефективно аналізувати та прогнозувати основні чинники демографічних процесів, що, в свою чергу, допомагає вчасно реагувати на негативні демографічні зміни та коригувати демографічну політику.
5. Наявність недоліків: Система показників, що впливають на результуючий фактор моделі не достатньо повна, в майбутньому варто було б дослідити вплив на показник кількості населення інших демографічних чинників.
6. Загальна оцінка кваліфікаційної бакалаврської роботи та її допущення до захисту перед ЕК: Кваліфікаційна бакалаврська робота Щирби О.О. «Аналіз та моделювання демографічних процесів у країнах Азії» є самостійним завершеним дослідженням, відповідає вимогам, що висуваються для даного типу робіт, може бути рекомендована до захисту перед ЕК. Рівень виконання кваліфікаційної бакалаврської роботи «високий».

Науковий керівник:
к.е.н., доцент, доцент кафедри
математичного моделювання та статистики Білик Т.О.

“____”_____20___року

ЗМІСТ

ВСТУП 6

РОЗДІЛ 1 8

ОСНОВИ ДЕМОГРАФІЇ ТА ДЕМОГРАФІЧНІ ТЕНДЕНЦІЇ У КРАЇНАХ АЗІЇ 8

1.1 Поняття демографії та основних демографічних чинників 8

1.2 Особливості демографічного аналізу країн Азії 10

1.3 Демографічні тенденції в країнах Азії 11

РОЗДІЛ 2 16

МОДЕЛЮВАННЯ ТА АНАЛІЗ ІНДИКАТОРІВ ДЕМОГРАФІЧНИХ ПОКАЗНИКІВ 16

2.1 Побудова та аналіз регресійної моделі демографічних процесів 16

2.2 Аналіз кластеризації демографічних показників у країнах Азії 39

ВИСНОВКИ 54

ВСТУП

Демографічні процеси є невід'ємною частиною соціально-економічного розвитку будь-якої країни. Зміни в чисельності населення, його вікова структура, рівень народжуваності та смертності, а також міграційні процеси мають значний вплив на економічне зростання, зайнятість, соціальні послуги та інфраструктуру. Аналіз демографічних процесів є важливим завданням для розробки ефективної політики, яка сприятиме стабільному розвитку суспільства.

В умовах глобалізації та швидких змін демографічних тенденцій у світі, особливу увагу привертають країни Азії. Цей регіон є одним із найдинамічніших у демографічному плані, характеризується високими темпами зростання населення, значними міграційними потоками та різноманітними соціально-економічними викликами. Вивчення демографічних процесів у країнах Азії дозволяє глибше зрозуміти їхні причини, наслідки та вплив на розвиток регіону.

Актуальність теми даної роботи полягає у важливості демографічного фактора як передумови економічного, інтелектуального та політичного розвитку будь-якої країни. Зростаюча потреба у використанні економіко-математичних моделей та методів для вивчення демографічних процесів пов'язана з необхідністю аналізу популяційних та міграційних процесів, що є основою для розробки ефективної політики, спрямованої на стабільний розвиток суспільства.

Метою даної дипломної роботи є аналіз і моделювання демографічних процесів в Азії, зокрема, дослідження впливу народжуваності, смертності, міграції та інших факторів на чисельність і структуру населення.

Для досягнення поставленої мети в роботі вирішуються наступні завдання:

1) розкриття сутності демографічних чинників як об’єктів для моделювання та аналізу;
2) розкриття тенденцій демографічних процесів в країнах Азії за останні десятиліття;

3) розробка та аналіз множинної лінійної моделі для прогнозування демографічних процесів у країнах Азії;
4) проведення кластерного аналізу на основі демографічних показників країн Азії;
5) оцінка результатів дослідження та їх вплив на демографічну ситуацію в країнах Азії.

Об'єктом дослідження є демографічні процеси в країнах Азії.

Предметом дослідження є методи та інструментарій моделювання демографічних процесів в країнах Азії.

Методологічною основою роботи є загальнонаукові методи дослідження: аналіз, системний підхід, а також спеціальні методи статистичного аналізу та економетричного моделювання. У роботі використовуються дані статистичних відомств країн Азії, міжнародних організацій, а також результати наукових досліджень у галузі демографії.

Практичне значення отриманих результатів полягає в можливості їх використання для вдосконалення демографічної політики країн Азії, а також для розробки стратегій соціально-економічного розвитку, які враховують демографічні особливості регіону.

Таким чином, дане дослідження сприятиме глибшому розумінню демографічних процесів в Азії та їхнього впливу на соціально-економічний розвиток, а також допоможе у розробці ефективних політичних рішень у цій сфері.

РОЗДІЛ 1
ОСНОВИ ДЕМОГРАФІЇ ТА ДЕМОГРАФІЧНІ ТЕНДЕНЦІЇ У КРАЇНАХ АЗІЇ

1.1 Поняття демографії та основних демографічних чинників

Демографія – це наука, яка вивчає закономірності відтворення населення, враховуючи його структуру, динаміку та розподіл. Вона досліджує широкий спектр аспектів, пов'язаних із населенням, і використовує статистичні дані для аналізу змін у чисельності населення і його складі. Одними з основних аспектів, що вивчаються демографією, є народжуваність, смертність та міграція населення. Народжуваність та смертність є природними демографічними процесами, міграція є механічним демографічним процесом, а зміни у складі населення, спричинені соціально-економічним та культурним розвитком є соціальним демографічним процесом [1].

Коефіцієнт народжуваності відображає кількість народжених дітей на 1000 осіб за певний період, що допомагає оцінити рівень оновлення населення. Високі показники народжуваності зазвичай спостерігаються у менш розвинених країнах, тоді як у більш розвинених країнах ці показники часто нижчі через вищий рівень освіти, що спричиняє зміну соціальних норм. *Коефіцієнт смертності* визначається кількістю смертей на 1000 осіб за певний період і є ключовим показником здоров'я населення. Цей показник залежить від багатьох факторів, таких як якість медичного обслуговування, рівень життя, екологічні умови та інші соціально-економічні чинники. Зниження рівня смертності зазвичай свідчить про покращення умов життя та медичного обслуговування [2]. *Природний приріст населення* – це різниця між коефіцієнтом народжуваності та коефіцієнтом смертності. Позитивний природний приріст означає зростання чисельності населення, тоді як негативний приріст вказує на його зменшення.

Механічний рух населення пов'язаний з міграцією, тобто переселенням людей з одного місця в інше, що може бути внутрішнім (в межах країни) або

зовнішнім (між країнами) процесом. Внутрішня міграція сприяє урбанізації та економічному розвитку міських центрів, тоді як зовнішня міграція впливає на міжнародні відносини та економічну співпрацю між країнами [3]. *Коефіцієнт міграції* показує інтенсивність міграційних потоків і визначається як кількість мігрантів (іммігрантів або емігрантів) на 1000 осіб населення за певний період, а *чиста міграція* показує різницю між кількістю людей, які приїхали до країни, та кількістю людей, які виїхали з країни за певний період. Якщо чиста міграція позитивна - це означає, що кількість іммігрантів перевищує кількість емігрантів і населення країни зростає за рахунок міграції. Якщо чиста міграція негативна - це означає, що кількість емігрантів перевищує кількість іммігрантів і населення країни зменшується. *Чиста зміна населення* є ще одним важливим показником, що відображає чисту зміну чисельності населення за рік. Цей показник обчислюється як різниця між числом народжених і числом померлих, з урахуванням чистої міграції. Позитивне значення вказує на зростання населення, тоді як негативне значення вказує на його зменшення [4].

Соціальний рух населення відображає зміни у складі населення, спричинені соціально-економічним та культурним розвитком, такими як підвищення рівня освіти, зміни у структурі зайнятості, урбанізація та інші процеси.

Для глибшого розуміння демографічних процесів також використовуються такі показники, як середня тривалість життя, коефіцієнт шлюбності, коефіцієнт розлучуваності та інші. *Середня тривалість життя* показує, скільки років в середньому може прожити людина, народжена в певний період, і є важливим індикатором загального стану здоров'я населення. *Коефіцієнти шлюбності та розлучуваності* відображають частоту укладення та розірвання шлюбів у суспільстві, що впливає на структуру сімей та народжуваність.

Важливим чинником, що впливає на демографічну ситуацію також є *площа землі*, доступна для проживання та ведення господарства. Регіони з обмеженими ресурсами та високою щільністю населення можуть стикатися з проблемами перенаселення, що ускладнює забезпечення необхідних умов для проживання, роботи та розвитку інфраструктури. Водночас великі за площею країни, де ресурси

розподілені нерівномірно, можуть мати значні внутрішні міграційні потоки, коли люди переїжджають до більш сприятливих для життя регіонів [5].

Розуміння демографічних процесів дозволяє урядам та міжнародним організаціям приймати обґрунтовані рішення для підтримки економічного зростання, забезпечення соціального захисту та покращення якості життя населення.

1.2 Особливості демографічного аналізу країн Азії

Для проведення досліджень демографія використовує державні статистичні бази даних, які містять інформацію про народження, смерті, шлюби, розлучення, захворювання та зайнятість. Наприклад, з 1948 року Статистичний відділ ООН веде "Систему демографічних щорічників". Крім того, демографи розробляють прогнози чисельності населення, такі як "Перспективи світового населення", які використовуються для формування різних державних політик, зокрема, у сферах соціального забезпечення, охорони здоров'я, ринку праці, пенсійної та податкової політики. Регулярні переписи, організовані урядами (наприклад, Бюро перепису населення США), є однією з форм соціальних опитувань, що надають цінні демографічні дані. Важливо доповнювати ці статистичні дані знаннями з соціології, економіки та політології для розуміння мотивацій, цінностей, цілей та уподобань різних груп населення [6].

Дослідження демографічних особливостей на регіональному рівні є важливим через специфіку соціально-економічного розвитку регіонів і різноманітність складових. Системний підхід до дослідження демографії дозволяє враховувати взаємозв'язок між процесами відтворення населення та соціально-економічним розвитком, оскільки ці процеси тісно пов'язані і утворюють складну систему причинно-наслідкових зв'язків.

Для комплексного вивчення демографічної ситуації регіону доцільно використовувати методи моделювання. Вони дозволяють виявити закономірності та особливості демографічних процесів і їхній зв'язок із соціально-економічними показниками на різних рівнях. Моделювання забезпечує ефективний аналіз і прогнозування демографічних змін, що є необхідним для планування розвитку регіонів та допомагає виявити закономірності і взаємозв'язки, які не можна знайти іншими методами. Впровадження математичних методів у демографічні дослідження сприяє розвитку моделювання як наукової галузі, особливо в соціальних науках.

Демографія є прикладною наукою, і зібрані та оброблені дані про населення широко використовуються як державними структурами, так і приватними суб'єктами для прогнозування демографічного розвитку регіонів, що є основою для розробки регіональних планів, прогнозування економічного розвитку, оптимізації соціальної сфери [7]. Управління охороною здоров'я, освітою, транспортом неможливе без знання структури населення, як і прийняття інвестиційних рішень без розуміння місцевого ринку праці. Демографи не лише аналізують тенденції, але й надають рекомендації щодо сімейної та міграційної політики. У бізнесі демографічні дані використовуються у маркетингу, управлінні виробництвом та продажем нових продуктів і послуг [8].

1.3 Демографічні тенденції в країнах Азії

Протягом останнього століття світове населення зазнало значних демографічних змін. Демографічний перехід, що відбувся в розвинених західних суспільствах, показує, як ці країни пройшли шлях від високої народжуваності та високої смертності до швидкого зростання населення з падінням смертності. Проте цей період швидкого зростання був тимчасовим. Суспільства згодом досягли нового рівня демографічного рівноваги з низькими та стабільними рівнями

народжуваності і смертності. Разом із цією демографічною еволюцією спостерігаються зміни у віковій структурі населення. Зниження народжуваності та підвищення тривалості життя призвели до старіння населення [9].

У різних субрегіонах Азії існує значна розбіжність у ключових демографічних показниках. Південна Азія, яка включає Афганістан, Бангладеш, Бутан, Індію, Іран, Мальдіви, Непал, Пакистан і Шрі-Ланку, має найбільшу чисельність населення – близько 2 мільярдів у 2022 році. Очікується, що населення цього субрегіону зростатиме в найближчі роки, оскільки його природний приріст становить 1,4%, що вище за середній показник для всієї Азії (0,8%). Серед п'яти основних субрегіонів Азії Центральна Азія має найвищий рівень природного приросту та наймолодшу вікову структуру. У 2022 році третина населення цього субрегіону була молодшою за 15 років. До Центральної Азії належать Казахстан, Киргизстан, Таджикистан, Туркменістан та Узбекистан.

На сьогоднішній день Азія є домом для більшої частини населення світу, складаючи 60% від загальної чисельності. Це пояснюється тим, що п'ять з найбільш населених країн світу – Китай, Індія, Індонезія, Пакистан і Бангладеш – розташовані в Азії. Разом Китай і Індія мають близько 2,8 мільярда людей. Обидві країни демонструють різні демографічні траєкторії. Китай давно визнаний найбільш населеною країною світу, досягнувши мільярдної позначки ще у 1980-х роках. Країна також відома успіхом у стримуванні зростання населення за допомогою агресивної політики однієї дитини, яка діяла з 1980 по 2015 рік.

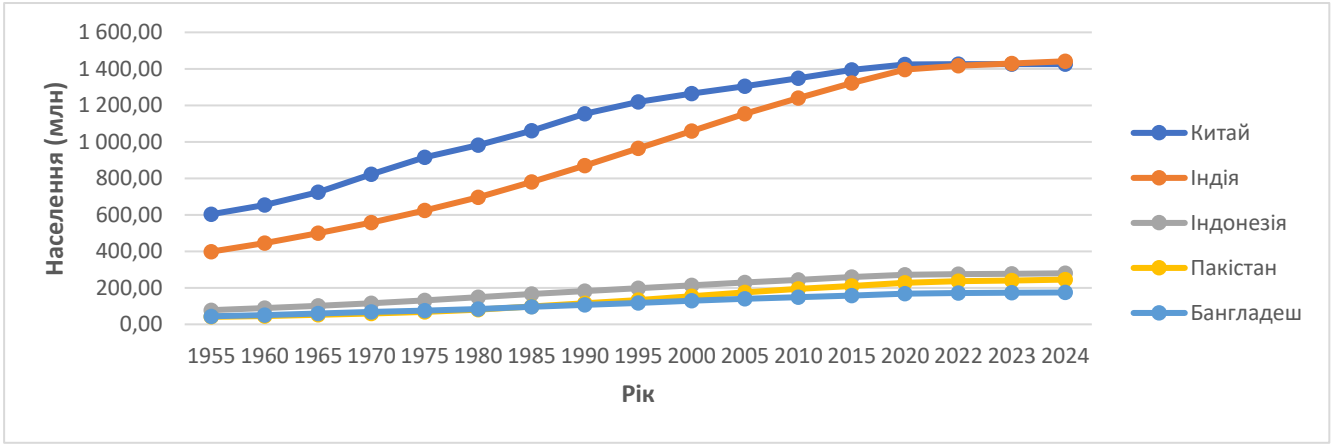

Рисунок 1.1 - Динаміка населення топ-5 країн Азії за кількістю населення (1955-2024 рр.)

Джерело: розроблено автором на основі [10]

Різноманітність культур, мов, релігій та традицій сприяє різнобарвності демографічного профілю Азії. Хоча загалом усі регіони континенту відчувають зниження темпів зростання населення з часом, Східна Азія традиційно реєструє найнижчий рівень зростання, навіть показуючи негативний рівень у останні роки. Найбільш помітним є скорочення населення Японії, яке досягло свого піку в 2010 році з 128 мільйонами.

Таблиця 1.1 - Індикатори популяції в Азії, 2022 р. [11]

	Населення (млн)	*Коефіцієнт природного приросту*	*Населення віком < 15 років (%)*	*Населення віком ≥ 65 років (%)*
Азія	4 730	0.8	24	10
Західна Азія	294	1.3	28	6
Центральна Азія	78	1.9	31	5
Південна Азія	2 008	1.4	27	6
Південно-Східна Азія	676	0.8	25	8
Східна Азія	1 674	0.0	18	15

Демографічна різниця між країнами Азії також помітна у віковій структурі. Наприклад, Японія має найвищий середній вік населення – 48,4 роки, що є другим показником у світі після Монако. Це означає, що половина населення Японії старше 48 років. На протилежному кінці знаходяться молоді країни, такі як Камбоджа і Філіппіни, з середнім віком населення 26,5 і 24,5 роки відповідно [10].

Домінуючі економіки Азії — Китай, Японія, Південна Корея та Тайвань — стикаються з безпрецедентним швидким старінням населення. У 1999 році Японія стала першою великою країною, де середній вік населення перевищив 40 років. У 2021 році середній вік в Японії зріс до 48,4 років, і сусідні країни також демонструють подібні тенденції. Цей процес старіння населення обумовлений кількома ключовими причинами: зростанням тривалості життя завдяки покращенню медичних послуг, зниженням рівня народжуваності через економічні та соціальні зміни, а також зменшенням частки молодих людей унаслідок міграції до інших країн у пошуках кращих економічних можливостей. Водночас більшість населення світу проживає в бідних країнах з молодим середнім віком, включаючи такі країни Індійсько-Тихоокеанського регіону, як Індія, Індонезія, Філіппіни та В'єтнам.

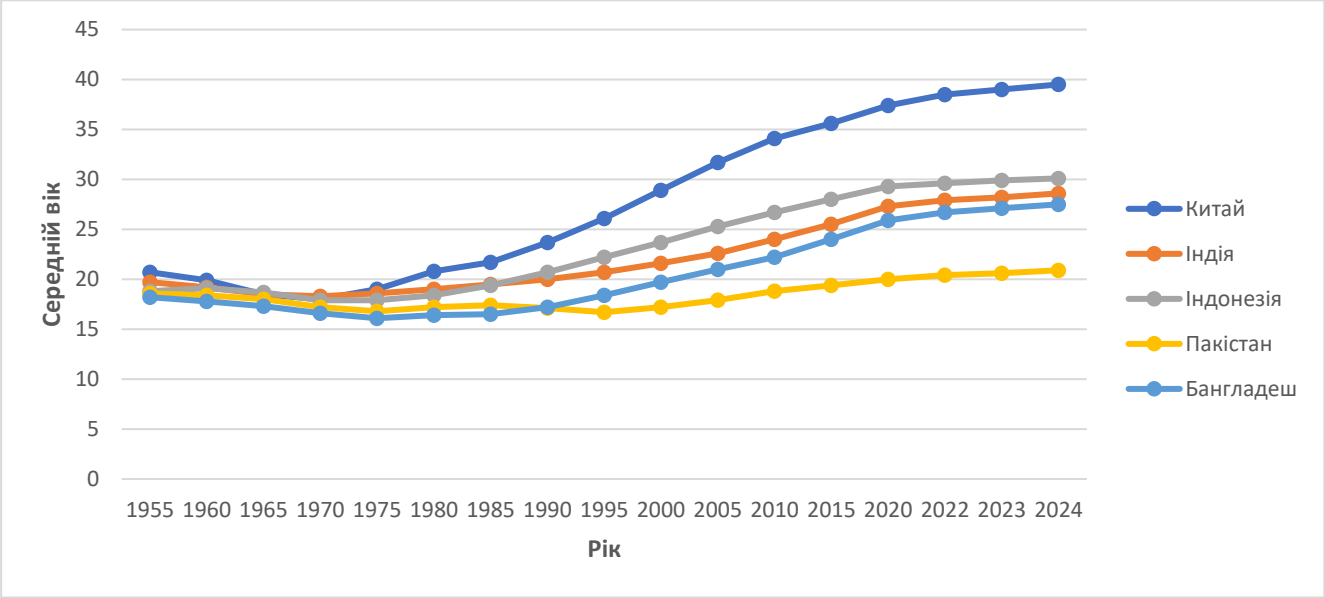

Рисунок 1.2 - Динаміка зміни середнього віку топ-5 країн Азії за кількістю населення (1955-2024 рр.)

Джерело: розроблено автором на основі [10]

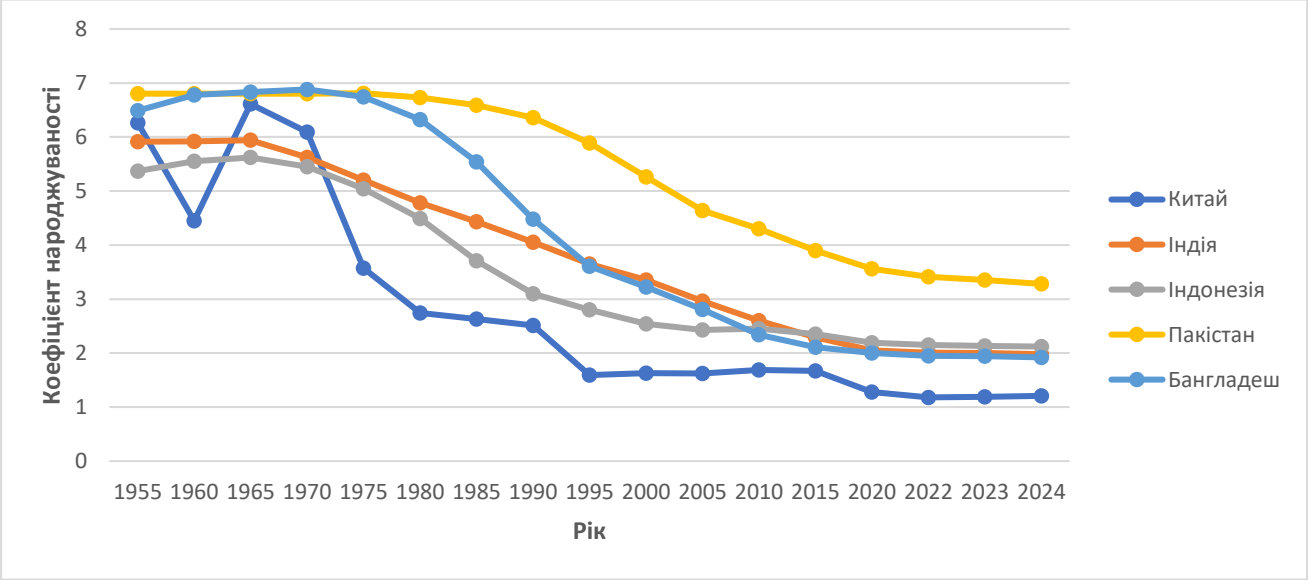

Рисунок 1.3 - Динаміка зміни коефіцієнта народжуваності топ-5 країн Азії за кількістю населення (1955-2024 рр.)

Джерело: розроблено автором на основі [10]

Різноманітні демографічні зміни в Азії створюють нові взаємозв'язки. Японія та Південна Корея, будучи виробничими гігантами, переносять значну частину свого виробництва до країн з молодшим населенням. Молоді країни Індійсько-Тихоокеанського регіону, такі як Індія та Філіппіни, також є одними з найбільших джерел еміграції у світі. Ці країни стикаються з різними викликами, пов'язаними з демографічними змінами, такими як перенаселеність та недостатня кількість робочих місць, що може призвести до політичної нестабільності.

Імміграція є одним із варіантів підтримання працездатного населення. Регіональні країни, такі як Австралія, Нова Зеландія та Сінгапур, використовують цей підхід для компенсації низького рівня народжуваності, що сприяє зростанню населення. Японія, Південна Корея та Тайвань експериментують із залученням іноземних робітників, але поки що ці програми не мають значного впливу на стрімке старіння населення. Для Китаю, з його великою кількістю населення, імміграція не може бути вирішенням проблеми скорочення працездатного населення.

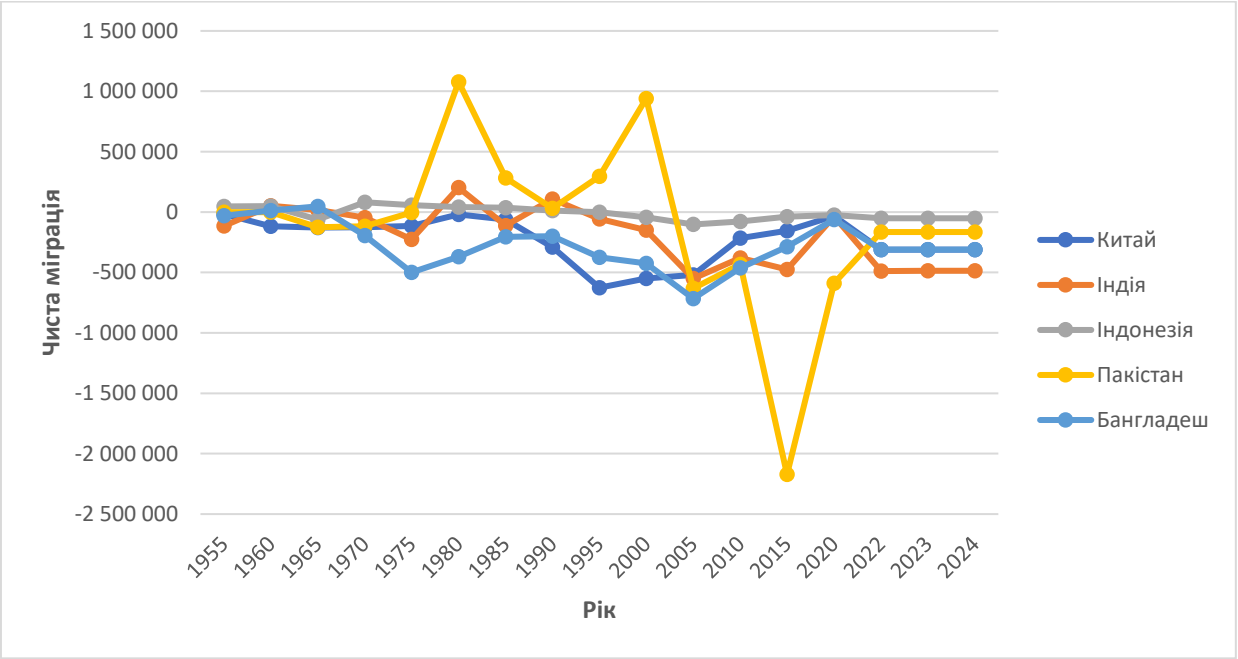

Рисунок 1.4 - Динаміка зміни чистої міграції топ-5 країн Азії за кількістю населення (1955-2024 рр.)

Джерело: розроблено автором на основі [10]

В'єтнам є прикладом країни, яка ледве підтримує стабільне населення без імміграції та швидкого старіння, але також стикається з внутрішніми демографічними змінами. Переміщення населення з сільських районів до міст стимулює ринок праці в містах, що сприяє зростанню експорту В'єтнаму. Медіа та громадськість у В'єтнамі стурбовані зростанням числа молодих робітників, що емігрують. Водночас, зростаюча взаємозалежність між Південною Кореєю та В'єтнамом є чудовим прикладом демографічної взаємопов'язаності в Азії, оскільки Південна Корея продовжує переносити своє виробництво до країн з нижчими витратами [11].

РОЗДІЛ 2
МОДЕЛЮВАННЯ ТА АНАЛІЗ ІНДИКАТОРІВ ДЕМОГРАФІЧНИХ ПОКАЗНИКІВ

2.1 Побудова та аналіз регресійної моделі демографічних процесів

Оцінювання параметрів економетричної моделі та дисперсійний аналіз є ключовими етапами процесу побудови моделі. Покрокова регресія є альтернативним методом оцінювання параметрів моделі, що базується на елементах дисперсійного аналізу.

Згідно з базовим поясненням взаємозв'язку між двома змінними за методом найменших квадратів (МНК), зазвичай акцент робиться на коефіцієнтах кореляції. Оцінка параметрів моделі прямо пропорційна до коефіцієнта парної кореляції. Це означає, що оцінити параметри моделі можна через коефіцієнти кореляції: спочатку оцінити тісноту зв'язку між кожною парою змінних, а потім знайти оцінки параметрів економетричної моделі.

Перш ніж приступити до покрокової регресії, необхідно нормалізувати дані. Це дозволяє усунути вплив різних одиниць вимірювання змінних, зробивши їх порівнянними. Для цього використовується функція «STANDARDIZE» в Excel, яка стандартизує кожне значення змінної, тобто приводить його до стандартизованої форми (Z-оцінки). Формула для нормалізації виглядає так:

$$y_i^* = \frac{y_i - \bar{y}}{\sigma_y}, \quad x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{\sigma_{x_j}} \tag{2.1}$$

де y_i^* - значення нормалізованої залежної змінної;

x_{ij}^* - значення нормалізованих пояснювальних змінних;

$\bar{y}$ - середнє значення залежної змінної;

$\bar{x}_j$ - середнє значення незалежної змінної;

σ_y, σ_{x_j} - *середньоквадратичні відхилення.*

Потім знаходиться кореляційна матриця (матриця парних коефіцієнтів кореляції), де обчислюються парні коефіцієнти кореляції між залежною і пояснювальними змінними. Порівнюючи абсолютні значення коефіцієнтів кореляції, вибирають змінну з найбільшим значенням, яка найтісніше пов'язана з залежною змінною. За допомогою МНК знаходять оцінку параметрів цієї моделі:

$$\hat{y}^* = \hat{\beta} \cdot x_j^* \tag{2.2}$$

де $\hat{\beta}$ *- оцінка параметрів моделі, що базується на нормалізованих даних.*

Процес триває з додаванням наступних змінних до моделі, поки всі значущі змінні не будуть включені.

Цей алгоритм дозволяє побудувати економетричну модель для аналізу демографічних процесів, враховуючи взаємозв'язки між змінними.

Коефіцієнт детермінації показує, яку частку загальної варіації залежної змінної пояснює регресійна модель та є квадратом коефіцієнта кореляції:

$$d = r^2 \tag{2.3}$$

Чим ближче значення d до 1, тим краще модель пояснює варіацію залежної змінної [12].

Для оцінки статистичної значущості коефіцієнта детермінації висувається нульова гіпотеза ($H0$): $d = 0$. Альтернативна гіпотеза (HA) припускає, що принаймні один параметр моделі відрізняється від нуля.

Для перевірки цих гіпотез використовують F-тест Фішера зі ступенями свободи m і $n - m - 1$. На основі значень d, отриманих у моделі, розраховується експериментальне значення F-статистики:

$$F = \frac{(d/m)}{((1-d)/(n-m-1))} \tag{2.4}$$

де n – загальна кількість спостережень або кількість даних у вашій вибірці.

m – кількість незалежних змінних (регресорів) у моделі.

Це значення порівнюється з табличним значенням F-розподілу при заданому рівні значущості (α), який зазвичай дорівнює 0,05 або 0,01. Якщо експериментальне значення *F* перевищує табличне, нульова гіпотеза відхиляється, що свідчить про адекватність моделі. У випадку, коли експериментальне значення менше табличного, модель вважається неадекватною [13].

t-критерій Стьюдента використовується для перевірки статистичної значущості коефіцієнтів регресії в економетричних моделях, допомагаючи визначити, чи є вплив незалежної змінної на залежну змінну значущим. Формула t-статистики виглядає так:

$$t = \frac{\beta}{SE(\beta)} \tag{2.5}$$

де β – оцінка коефіцієнта регресії;

$SE(\beta)$ – стандартна помилка коефіцієнта.

Зазвичай використовується рівень значущості 5% (α=0,05), але він може бути іншим залежно від дослідження. Перевірка гіпотез включає нульову гіпотезу ($H0$): β=0 і альтернативну гіпотезу (HA): $\beta \neq 0$.

Фактичне значення t-статистики порівнюється з табличним значенням t-розподілу зі ступенями свободи $n - m - 1$. Якщо $|t_{експ}| > t_{табл}$, нульова гіпотеза відхиляється, що означає значущість коефіцієнта регресії, інакше нульова гіпотеза не відхиляється [14].

Важливо зрозуміти, що побудова моделі за допомогою методу найменших квадратів - це лише початковий етап. Після отримання початкових оцінок параметрів регресії необхідно переконатися в адекватності моделі, перевіривши основні припущення класичної лінійної регресії [15]. Ці припущення включають нормальність розподілу залишків, відсутність автокореляції та гетероскедастичності.

Тест Шапіро-Вілка використовується для перевірки нормальності розподілу даних. Цей тест є одним із найбільш поширених методів статистичного аналізу, який дозволяє визначити, наскільки близько розподіл вибірки відповідає нормальному розподілу. Нормальність залишків є важливим припущенням у багатьох статистичних методах, зокрема в лінійній регресії, оскільки відхилення від нормальності можуть вплинути на точність і надійність отриманих результатів.

Основна мета тесту Шапіро-Вілка полягає у перевірці гіпотези про те, що дані вибірки походять з нормального розподілу. Тест перевіряє нульову гіпотезу про нормальність розподілу проти альтернативної гіпотези, що розподіл даних відмінний від нормального.

Основні етапи проведення тесту включають:

1) *рангування даних*. Вибірка даних сортується в порядку зростання.
2) *Обчислення середнього значення*. Розраховується середнє значення вибірки.
3) *розрахунок коефіцієнтів*. Обчислюються коефіцієнти, які залежать від розміру вибірки і очікуваних значень при нормальному розподілі.

$$a_i = \frac{m_i}{\sqrt{\Sigma_{i-1}^{n} m_i^2}} \tag{2.6}$$

де mi — очікувані значення порядкових статистик.

4) *обчислення статистики тесту*. Використовуючи розраховані коефіцієнти, обчислюється статистика тесту W:
5)

$$w = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{2.7}$$

де x(i) – i-й спостережуваний порядок статистики,

$\bar{x}$ – середнє значення вибірки,

a_i– коефіцієнти, що залежать від вибірки.

Нульова гіпотеза ($H0$): Дані вибірки походять з нормального розподілу.

Альтернативна гіпотеза **(HA)**: Дані вибірки не походять з нормального розподілу.

Після обчислення статистики тесту W, отримане значення порівнюється з критичним значенням для заданого рівня значущості. Якщо p-значення є меншим за обраний рівень значущості, нульова гіпотеза відхиляється, що свідчить про ненормальність розподілу даних. Якщо p-значення більше α, немає підстав відхиляти нульову гіпотезу, і дані вважаються такими, що походять з нормального розподілу [16].

Тест Дарбіна-Уотсона є одним з найвідоміших методів для перевірки моделі на наявність автокореляції залишків. На відміну від багатьох інших тестів, він включає кілька етапів і зони невизначеності.

На початку розраховується значення d-статистики за формулою:

$$\mathrm{d} = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \tag{2.8}$$

де t - індекс, що позначає спостереження, від 1 до n;

e – залишки регресійної моделі для t-го спостереження [17].

Значення d-статистики Дарбіна-Уотсона можуть варіюватися від 0 до 4.

На другому етапі задаються рівень значущості α і підраховується кількість факторів k у моделі. Далі знаходяться відповідні критичні значення d_L та d_U за таблицею Дарбіна-Уотсона для заданого рівня значущості, кількості факторів і спостережень n.

Якщо $0 < d < d_L$, це вказує на наявність позитивної автокореляції.

Якщо $d_L \leq d \leq d_U$ або $4-d_U \leq d \leq 4-d_L$, то результат знаходиться в зоні невизначеності, і висновки про автокореляцію зробити не можна.

Якщо $d > 4-d_L$, це свідчить про наявність негативної автокореляції.

Якщо $d_U < d < 4-d_U$, автокореляція відсутня.

Якщо тест Дарбіна-Уотсона показав наявність автокореляції залишків, необхідно дослідити причини її виникнення. Модель регресії повинна бути переглянута і побудована так, щоб зменшити загрозу виникнення автокореляції. Це може включати додавання нових змінних, трансформацію даних або використання інших статистичних методів [18].

Тест Бройша-Пагана застосовується для виявлення гетероскедастичності в регресійних моделях, припускаючи, що дисперсія залишків залежить від кількох змінних. Як і в інших тестах, він перевіряє нульову гіпотезу про гомоскедастичність проти альтернативної гіпотези про гетероскедастичність [19].

Процес тестування починається з оцінки параметрів регресії за допомогою методу найменших квадратів (МНК) і обчислення залишків. Далі обчислюється дисперсія залишків:

$$\sigma_e^2 = \frac{RSS}{n} \tag{2.9}$$

де RSS – сума квадратів залишків,

а n – кількість спостережень.

На наступному етапі залишки $e(t)$ діляться на δ_e^2 і регресуються на $z(t)$ за допомогою звичайного методу найменших квадратів (ЗМНК). Це дозволяє отримати оцінки регресії та обчислити суму квадратів помилок (ESS):

$$ESS = \beta^T X^T X \beta - \mu_y \tag{2.10}$$

де μy – середнє значення послідовності X(t).

Статистика тесту розраховується як $\frac{RSS}{2}$ допоміжної регресії і має асимптотично розподіл хі-квадрат з $p-1$ ступенями свободи. Якщо $\frac{RSS}{2}$ перевищує критичне значення хі-квадрат $x^2_{\alpha(p-1)}$, то нульова гіпотеза про гомоскедастичність відхиляється на рівні значущості α [20]. Крім того, можна використати статистику nR^2 для тієї ж допоміжної регресії. Ця статистика також має розподіл хі-квадрат з $p-1$ ступенями свободи при виконанні нульової гіпотези. Обидва підходи є асимптотично еквівалентними [21].

Регресійний аналіз

Після розгляду основ демографії та ключових факторів впливу, а також аналізу значення демографічного аналізу та сучасних демографічних тенденцій в країнах Азії, настав час перейти до практичної частини дослідження. Метою цієї частини є проведення кількісного аналізу демографічних даних за допомогою методу множинної лінійної регресії.

У цьому дослідженні використовується набір даних, що містить широкий спектр демографічних показників для країн Азії за 2023 рік. Таблиця включає інформацію про чисельність населення, річну зміну населення, чисту зміну населення, щільність населення, земельну площу, чисту міграцію, коефіцієнт народжуваності, середній вік та відсоток міського населення (дивитись у додатку А).

Виконаємо покроковий регресійний аналіз у Microsoft Excel, який є потужним інструментом для проведення статистичних аналізів, включаючи нормалізацію даних та побудову кореляційних матриць. Після здійснення нормалізації демографічних показників, що дозволило усунути вплив різних одиниць вимірювання змінних та зробити їх порівнянними будуємо кореляційну матрицю для визначення ступеня та напрямку взаємозв'язку між різними змінними. Це дозволить ідентифікувати незалежні змінні, які мають найбільший вплив на залежну змінну, і є важливим кроком у побудові економетричної моделі.

	Населення (млн)	*Чиста зміна населення (тис)*	*Щільність населення (осіб/км^2)*	*Земельна площа (тис км^2)*	*Чиста міграція*	*Коефіцієнт народжуваності*	*Середній вік*	*Міське населення (%)*
Населення (млн)	1							
Чиста зміна населення (тис)	0,649253386	1						
Щільність населення (осіб/км^2)	0,178533426	0,222224893	1					
Земельна площа (тис км^2)	0,809139511	0,219354434	-0,048035847	1				
Чиста міграція	-0,525490172	-0,425681181	-0,12223654	-0,356363124	1			
Коефіцієнт народжуваності	-0,1269215	0,158623625	0,001093589	-0,054672631	0,035233037	1		
Середній вік	0,058121435	-0,248145783	0,065621473	0,068122592	0,03122305	-0,870212882	1	
Міське населення (%)	-0,108232751	-0,27122505	0,194172612	-0,004777205	0,114068648	-0,375344654	0,4904151	1

Рисунок 2.1 - Кореляційна матриця демографічних показників для країн Азії
Джерело: розроблено автором

Для побудови економетричної моделі чисельність населення буде обрана як залежна змінна. Аналіз кореляційної матриці дозволяє не тільки визначити взаємозв'язки між залежною змінною та потенційними незалежними змінними, а також оцінити вплив цих незалежних змінних одна на одну з метою виявлення мультиколінеарності.

Аналіз кореляційної матриці показав, що чисельність населення має помірний позитивний зв'язок з чистою зміною населення (кореляція 0,649) та сильно корелює із земельною площею (0,809). Це свідчить про те, що більша площа території країни пов'язана з більшою чисельністю населення. Водночас, чисельність населення має помірний негативний зв'язок з чистою міграцією (-0,525), що вказує на те, що зростання чистої міграції призводить до зменшення чисельності населення. Інші змінні, такі як щільність населення, коефіцієнт народжуваності, середній вік та відсоток міського населення, мають відносно слабкі, або дуже слабкі кореляції з чисельністю населення.

Щільність населення має низький коефіцієнт кореляції з чисельністю населення (0,178), тому вона не додає значної інформації до моделі. Крім того, з наукової точки зору, дослідження взаємозв'язку чисельності населення з іншими змінними, які можуть мати більш специфічний вплив або відображати інші аспекти демографічних процесів, виглядає більш цікавим і значущим. Щільність населення виключається з моделі на користь змінних, які потенційно можуть надати глибші та більш інформативні результати.

Чиста міграція та чиста зміна населення також є суміжними поняттями, і їхній кореляційний зв'язок достатньо високий (-0,426), що може спричинити проблеми з мультиколінеарністю. Мультиколінеарність виникає, коли незалежні змінні мають високі кореляції між собою, що призводить до нестабільності оцінок параметрів регресії. Це може спотворити результати та зробити інтерпретацію моделі складною.

Коефіцієнт народжуваності має сильний зв'язок із середнім віком (-0,870), що свідчить про високий рівень мультиколінеарності між цими змінними. Водночас, середній вік сильно корелює з відсотком міського населення (0,490).

Така ситуація ускладнює використання цих змінних у моделі одночасно, оскільки це може призвести до проблем із точністю та стабільністю моделі.

Враховуючи потребу в різноманітності моделі та з метою уникнення мультиколінеарності, найбільш логічним буде розглянути коефіцієнт народжуваності. Ця змінна не має сильної кореляції з іншими незалежними змінними, що робить її більш стабільною та надійною для використання у моделі. Коефіцієнт народжуваності має найбільший коефіцієнт кореляції з залежною змінною серед тих параметрів, які не були виключені дедуктивним шляхом.

Для остаточного визначення, які змінні включати до моделі, буде проведено парну регресію, що дозволить детальніше дослідити взаємозв'язки між змінними та забезпечити більш точну інтерпретацію результатів. Інші змінні вирішено не включати до моделі через їх схожість з іншими показниками, що знижує їхню інформативність та може призвести до мультиколінеарності.

Проведення парної регресії в Excel дозволить поетапно додавати змінні до моделі, починаючи з тих, що мають найвищі коефіцієнти кореляції з залежною змінною. Це дозволить побудувати модель, яка точно відображатиме вплив різних демографічних та соціально-економічних факторів на чисельність населення.

Для початку аналізу створено першу модель за допомогою функції «LINEST» у Microsoft Excel, де в якості незалежної змінної використано лише змінну «земельна площа». Функція «LINEST» у Microsoft Excel використовується для виконання аналізу залежності між однією залежною змінною і декількома незалежними змінними. «LINEST» фактично реалізує метод найменших квадратів (МНК), який є стандартним методом для оцінки параметрів лінійної регресії. Ця функція повертає статистику, яка описує найкращу лінію, що відповідає даним, і надає інформацію про коефіцієнти регресії, стандартні помилки, значення R^2, F-критерій і інші важливі показники.

0,809139511	-2,42854E-17
0,083945226	0,08311816
0,654706749	0,593582389
92,90836293	49
32,73533744	17,26466256

Рисунок 2.2 - Результати функції LINEST для населення та площі землі
Джерело: розроблено автором

За результатами аналізу МНК коефіцієнт детермінації дорівнює 0.655, що свідчить про те, що приблизно 65.5% варіації в населенні можна пояснити змінами у площі землі.

Статистичне значення F дорівнює 92.91. Для перевірки значущості цього значення F-статистики необхідно порівняти його з табличним значенням F-розподілу для заданого рівня значущості (α=0.05) та відповідних ступенів свободи. У даному випадку кількість незалежних змінних дорівнює 1, а загальна кількість спостережень дорівнює 51. Ступені свободи для чисельника F-розподілу становлять m=1, а для знаменника $n-m-1$=49. Табличне значення F-розподілу для при таких параметрах дорівнює приблизно 4.04 (дивитись у додатку Б). Оскільки експериментальне значення F значно перевищує табличне значення, це дозволяє відхилити нульову гіпотезу та свідчить про статистичну значущість моделі.

t-критерій Стьюдента використовується для перевірки статистичної значущості коефіцієнтів регресії. Для даного випадку t-статистика дорівнює 0.809/0.0839≈9.64. Для перевірки значущості t-статистики порівнюємо її з табличним значенням t-розподілу для заданого рівня значущості (α=0.05) та ступенів свободи. Табличне значення t-розподілу дорівнює приблизно 2.01 (дивитись у додатку В). Оскільки фактичне значення t значно перевищує табличне значення, нульова гіпотеза відхиляється, що означає значущість коефіцієнта регресії.

Загалом, отримані результати показують, що модель добре пояснює варіацію залежної змінної, а високі значення F-статистики і t-статистики підтверджують значущість моделі. Низька стандартна помилка свідчить про високу точність оцінок параметрів, що додатково підтверджує надійність моделі.

Далі додаємо незалежну змінну, яка стоїть на другому місці за значенням коефіцієнта кореляції, а саме «чиста зміна населення». Проводимо аналіз результатів за методом найменших квадратів (МНК). Цей додатковий фактор дозволить отримати більш точний прогноз, враховуючи не лише площу землі, але й чисту зміну населення.

0,49561	0,70042	-2,2514E-17
0,0494	0,0494	0,047717678
0,88852	0,34077	#Н/Д
191,284	48	#Н/Д
44,426	5,57404	#Н/Д

Рисунок 2.3 - Результати функції LINEST для населення, площі землі та чистої зміни населення

Джерело: розроблено автором

На основі аналізу МНК для двох незалежних змінних, отримані наступні результати. Коефіцієнт детермінації дорівнює 0.86852, що означає, що модель пояснює 86.85% варіації залежної змінної. Стандартні помилки для обох коефіцієнтів дорівнюють 0.0494, що вказує на високу точність оцінок.

Для моделі з двома незалежними змінними і загальною кількістю спостережень 51 табличне значення F-розподілу дорівнює приблизно 3.19 (дивитись у додатку Б). Оскільки експериментальне значення F (191.284) значно перевищує табличне значення, це дозволяє відхилити нульову гіпотезу і підтвердити, що модель є статистично значущою.

t-статистика для коефіцієнта регресії площі землі становить 0.49561/0.0494≈10.04, а для чистої зміни населення 0.70042/0.047717678≈14.67. Табличне значення t-розподілу для рівня значущості α=0.05 і ступенів свободи 48 дорівнює приблизно 2.01 (дивитись у додатку В). Оскільки фактичні значення t значно перевищують табличне значення, це підтверджує значущість кожного коефіцієнта.

Порівняно з першою моделлю, де було додано лише одну незалежну змінну, додавання другої змінної значно покращило якість моделі. Тепер можна спробувати додати ще одну змінну, яка за якістю коефіцієнта кореляції йде наступною, а саме «коефіцієнт народжуваності».

-0,1731	0,52663	0,68416	-1E-18
0,04263	0,04363	0,04314	0,04149
0,91746	0,29632	#Н/Д	#Н/Д
174,143	47	#Н/Д	#Н/Д
45,8731	4,12694	#Н/Д	#Н/Д

Рисунок 2.4 - Результати функції LINEST для населення, площі землі, чистої зміни населення та коефіцієнта народжуваності

Джерело: розроблено автором

На основі аналізу МНК для трьох незалежних змінних, встановлено, що коефіцієнт детермінації дорівнює 0.91745, що означає, що модель пояснює 91.75% варіації залежної змінної. Стандартні помилки дорівнюють 0.04263 для площі землі, 0.04363 для чистої зміни населення та 0.04314 для коефіцієнта народжуваності відповідно, що вказує на високу точність оцінок.

Для моделі з трьома незалежними змінними і загальною кількістю спостережень 51 табличне значення F-розподілу дорівнює приблизно 2.80 (дивитись у додатку Б). Експериментальне значення F дорівнює 174.143, що значно перевищує табличне значення для заданого рівня значущості та відповідних ступенів свободи, що дозволяє відхилити нульову гіпотезу і підтвердити, що модель є статистично значущою.

t-статистика для коефіцієнта регресії площі землі становить $-0.1731/0.04263 \approx -4.06$, для чистої зміни населення ≈ 12.07, і для коефіцієнта народжуваності≈ 15.86. Табличне значення t-розподілу для рівня значущості $\alpha=0.05$ і ступенів свободи 47 дорівнює приблизно 2.01 (дивитись у додатку В). Оскільки фактичні значення t значно перевищують табличне значення, це підтверджує значущість кожного коефіцієнта.

На наступному етапі аналізу додається змінна "міське населення (%)", яка є наступною за значущістю коефіцієнта кореляції. Попередні моделі демонстрували високий рівень релевантності, тому цікаво оцінити, чи покращить додавання цієї змінної параметри моделі.

-0,0333	-0,1843	0,51919	0,68502	1,2E-18
0,04674	0,04568	0,04508	0,04338	0,04171
0,91836	0,29788	#Н/Д	#Н/Д	#Н/Д
129,369	46	#Н/Д	#Н/Д	#Н/Д
45,9182	4,08181	#Н/Д	#Н/Д	#Н/Д

Рисунок 2.5 - Результати функції LINEST для населення, площі землі, чистої зміни населення, коефіцієнта народжуваності та міським населенням у %

Джерело: розроблено автором

Коефіцієнт детермінації дорівнює 0.91836, що означає, що модель пояснює 91.84% варіації залежної змінної. Стандартні помилки для площі землі, чистої зміни населення, коефіцієнта народжуваності та міського населення (%) відповідно дорівнюють 0.04674, 0.04568, 0.04508 та 0.04338, що вказує на високу точність оцінок. Експериментальне значення F дорівнює 129.369, що значно перевищує табличне значення для заданого рівня значущості, підтверджуючи статистичну значущість моделі. t-статистика для коефіцієнтів регресії значно перевищують табличне значення, це підтверджує їх значущість.

Поступово додаючи по одній незалежній змінній до моделі, було досягнуто високих показників аналізу моделі. Початкове включення площі землі як першої незалежної змінної надало базову модель з достатньо високим коефіцієнтом детермінації. Додавання другої змінної, такої як чиста зміна населення, ще більше підвищило пояснювальну здатність моделі, що відобразилось у зростанні коефіцієнта детермінації та статистичній значущості коефіцієнтів. Додавання третьої незалежної змінної, такої як коефіцієнт народжуваності, ще більше покращило модель, що призвело до зростання коефіцієнта детермінації до 0.92. F- та t-статистики також підтверджують значущість додаткової змінної.

На основі подальшого аналізу, додавання четвертої змінної "міське населення (%)" майже не покращило значущість моделі. Подальше додавання змінних навряд чи значно покращить модель, оскільки їхній коефіцієнт значущості по відношенню до залежної змінної ще менший. Крім того, інші потенційні змінні не можуть бути додані через наявність мультиколінеарності по відношенню до інших незалежних змінних.

Таким чином, найкращим рішенням буде продовжити дослідження моделі з трьома незалежними змінними: *площа землі, чиста зміна населення та коефіцієнт народжуваності*. Ця модель забезпечує високу точність і пояснювальну здатність, що робить її оптимальною для подальших аналізів.

Для удосконалення моделі застосуємо алгоритм Феррара-Глобера. Важливо запевнитися, що між обраними незалежними змінними відсутня мультиколінеарність, оскільки її наявність може спричинити нестабільність оцінок коефіцієнтів, що ускладнить інтерпретацію результатів моделі. Зокрема, високі рівні мультиколінеарності можуть призвести до значного розширення довірчих інтервалів для коефіцієнтів регресії, що знижує точність прогнозів і робить моделі менш надійними.

Спочатку обчислюється матриця парних кореляцій між незалежними змінними, що дозволяє оцінити взаємозв'язки між ними.

	Column 1	*Column 2*	*Column 3*
Column 1	1	0,21935	0,15862
Column 2	0,21935	1	-0,0547
Column 3	0,15862	-0,05467	1

Рисунок 2.6 - Матриця парних кореляцій між незалежними змінними
Джерело: розроблено автором

Далі визначаємо визначник матриці парних кореляцій, який використовується для розрахунку Хі-квадрат критерію. Хі-квадрат критерію (χ^2 - критерій) — це статистичний тест, який використовується для оцінки відповідності спостережуваних частот очікуваним частотам в одній або більше категоріях. Він допомагає визначити, чи є розбіжності між спостережуваними і очікуваними значеннями випадковими або ж вони вказують на певну закономірність. У цьому контексті χ^2 - критерій допомагає оцінити наявність мультиколінеарності між змінними.

Розраховане значення Хі-квадрат критерію порівнюється з табличним значенням для відповідного рівня значущості та ступенів свободи. Якщо розрахункове значення менше за табличне, це свідчить про відсутність мультиколінеарності між незалежними змінними.

Значення фактичного значення Хі-квадрат критерію розраховується за фомулою:

$$\chi^2 = -\left[n - 1 - \frac{1}{6}(2m + 5)\right] \ln|r_{xx}| \quad (2.11)$$

де n - кількість спостережень у вибірці;

m - кількість незалежних змінних у моделі;

$|r_{xx}|$ *- визначник матриці парних кореляцій між незалежними змінними.*

У Excel для обчислення табличного значення χ^2 (хі-квадрат) використовується функція «CHIINV», яка знаходить табличне значення χ^2 для заданого рівня значущості (зазвичай 0.05) при $\frac{1}{2}m(m-1)$ ступенів свободи [22].

За результатами розрахунків обраної моделі розрахункове значення Хі-квадрат критерію = 4.01996, що є меншим за обчислене табличне значення 7.81473 і свідчить про відсутність мультиколінеарності між незалежними змінними, що дозволяє використовувати їх у моделі без ризику нестабільності оцінок коефіцієнтів.

Оцінка параметрів побудованої регресійної моделі та аналіз результатів

Для побудованої моделі, де залежною змінною є чисельність населення, а незалежними змінними виступають земельна площа, чиста зміна населення та коефіцієнт народжуваності, рівняння лінійної регресії має наступний вигляд:

Населення= β0 + β1 · Земельна площа + β2 · Чиста зміна населення + β3 · Коефіцієнт народжуваності

Для подальшого вдосконалення та перевірки адекватності моделі буде проведена пострегресійна діагностика в середовищі RStudio, яке забезпечує зручні інструменти для аналізу даних, статистичного моделювання та візуалізації результатів. Пострегресійна діагностика включає перевірку нормальності залишків розподілу, автокореляції залишків, гетероскедастичності, аналіз довірчих

інтервалів для коефіцієнтів регресії. Ці кроки дозволяють оцінити якість моделі та впевнитися в коректності її припущень.

Після завантаження Excel-файлу в RStudio, дані були попередньо оброблені шляхом перейменування стовпців для зручності аналізу:

- population – чисельність населення (млн),
- area – земельна площа (тис км^2),
- netchange – Чиста зміна населення (тис),
- birthrate – коефіцієнт народжуваності.

Наступним кроком стала оцінка параметрів регресійної моделі за допомогою функції *lm* в RStudio.

Після побудови моделі за допомогою функції *lm*, застосування функції *summary* надає детальну інформацію про оцінені параметри моделі, їх статистичну значущість та інші важливі метрики. Це дозволяє оцінити, наскільки добре обрана модель пояснює варіацію в залежній змінній, а також визначити вплив кожної з незалежних змінних.

```
Call:
lm(formula = population ~ area + netchange + birthrate, data = dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-266.26  -20.57   11.32   32.22  173.10

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  95.590779  35.115523   2.722 0.009069 **
area          0.133124   0.008394  15.859  < 2e-16 ***
netchange     0.084076   0.006965  12.071 5.25e-16 ***
birthrate   -62.219323  15.326474  -4.060 0.000185 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 82.48 on 47 degrees of freedom
Multiple R-squared:  0.9175,    Adjusted R-squared:  0.9122
F-statistic: 174.1 on 3 and 47 DF,  p-value: < 2.2e-16
```

Рисунок 2.7 - Результати лінійної регресії за допомогою функції lm
Джерело: розроблено автором

Значення перетину (Intercept) становить 95.59. Це показує, яке значення прийме залежна змінна, якщо всі незалежні змінні дорівнюватимуть нулю. В даному контексті перетин має обмежене практичне значення, оскільки не завжди можливо, щоб усі змінні дорівнювали нулю в реальному світі.

Коефіцієнт для змінної area становить 0.133124. Це означає, що збільшення площі на одну одиницю призводить до збільшення чисельності населення на 0.133124 одиниці, за умови, що інші змінні залишаються незмінними. P-value для цієї змінної є дуже малим (менше 2e-16), що вказує на високу статистичну значущість цього коефіцієнта.

Коефіцієнт для змінної netchange дорівнює 0.084076. Це означає, що збільшення чистої зміни населення на одну одиницю призводить до збільшення чисельності населення на 0.084076 одиниці, за умови, що інші змінні залишаються незмінними. P-value для цієї змінної становить 5.25e-16, що свідчить про її високу статистичну значущість.

Коефіцієнт для змінної birthrate становить -62.219323. Це свідчить про те, що збільшення коефіцієнта народжуваності на одну одиницю призводить до зменшення чисельності населення на 62.219323 одиниці, за умови, що інші змінні залишаються незмінними. P-value для цієї змінної становить 0.000185, що вказує на її статистичну значущість.

Стандартна похибка залишків (Residual standard error) у моделі становить 82.48. Цей показник характеризує середню величину відхилень фактичних значень залежної змінної (населення) від передбачених значень, отриманих за допомогою регресійної моделі. Іншими словами, це середнє значення різниці між спостереженими та прогнозованими значеннями, враховуючи ступінь свободи моделі. Чим менша стандартна похибка залишків, тим точніше модель передбачає залежну змінну. Значення 82.48 свідчить про те, що в середньому передбачення моделі відрізняються від фактичних значень чисельності населення на 82.48 одиниць. Враховуючи масштаб даних (чисельність населення в мільйонах), це значення є досить значним, що може вказувати на необхідність подальшого вдосконалення моделі або включення додаткових значущих змінних.

Скорегований коефіцієнт детермінації (Adjusted R-squared) враховує кількість незалежних змінних у моделі та кількість спостережень, коригуючи значення R-squared з урахуванням числа параметрів. Даний показник допомагає уникнути переобчислення моделі, забезпечуючи більш точну оцінку якості регресії

в умовах багатофакторного аналізу. Значення скорегованого коефіцієнта детермінації для даної моделі становить 0.9122, що вказує на те, що після коригування з урахуванням кількості параметрів, модель все ще пояснює 91.22% варіації чисельності населення. Це свідчить про те, що модель є адекватною і добре підходить для аналізу.

Коефіцієнт детермінації (R-squared), F-статистика та t-статистика вже обчислювався для даної моделі під час проведення покрокового аналізу, тому не будемо зупинятись на цих даних.

Загалом результати показують високий рівень пояснювальної здатності та статистичну значущість усіх незалежних змінних обраної моделі. Модель є надійною і може використовуватись для подальшого аналізу та прогнозування.

Однією з найголовніших умов адекватності моделі є нормальність розподілу даних. Для перевірки цієї умови використовується тест Шапіро-Вілка, який допомагає визначити, наскільки розподіл кожної змінної в кластерах відхиляється від нормального розподілу. Виконання цього тесту є важливим кроком для забезпечення коректності статистичних висновків та моделей.

```
        Shapiro-Wilk normality test

data:  residuals_model
W = 0.8709, p-value = 5.133e-05
```

Рисунок 2.8 - Тест Шапіро-Вілка на нормальність розподілу залишків
Джерело: розроблено автором

Отримані результати вказують на статистику тесту W, яка дорівнює 0.8709, і p-значення 5.133e-05. Значення статистики W, близьке до 1, свідчить про відповідність даних нормальному розподілу. У нашому випадку значення 0.8709 вказує на певне відхилення від нормальності. Однак саме по собі це значення ще не є вирішальним.

p-значення, що дорівнює 5.133e-05 (або 0.00005133), є дуже малим і значно меншим за загальноприйнятий рівень значущості $\alpha=0.05$. Низьке p-значення вказує на те, що ми можемо відхилити нульову гіпотезу про нормальність розподілу залишків. Це свідчить про те, що розподіл залишків моделі значно відрізняється від

нормального розподілу. Таким чином, припущення про нормальність розподілу залишків порушується, що може вплинути на коректність і надійність результатів регресійного аналізу.

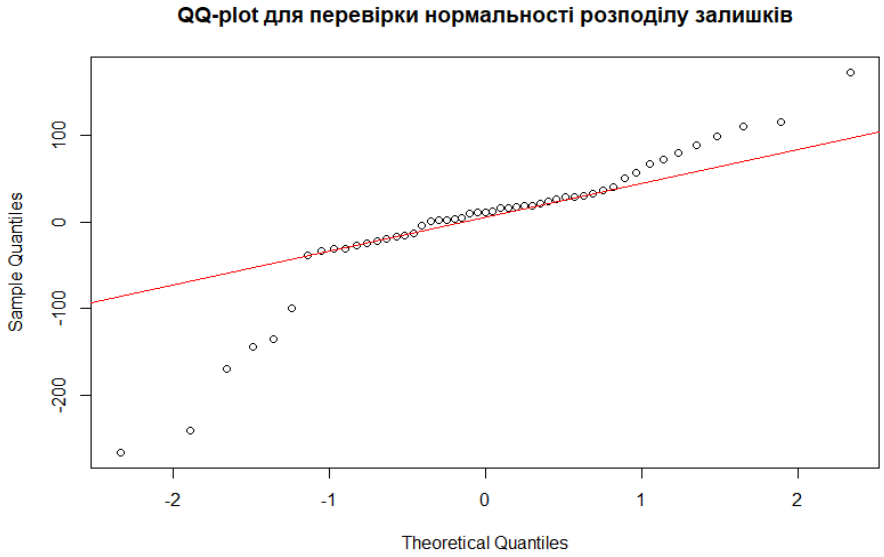

Рисунок 2.9 - QQ-plot для перевірки нормальності залишків
Джерело: розроблено автором

Графік квантиль-квантиль (QQ-plot) для перевірки нормальності розподілу залишків показує, наскільки добре залишки моделі відповідають нормальному розподілу. На графіку можна побачити, що точки значною мірою відхиляються від червоної лінії, яка представляє ідеальний нормальний розподіл. Особливо помітні відхилення в лівій частині графіка, що вказує на наявність значних відхилень від нормальності. Це свідчить про те, що розподіл залишків не відповідає нормальному розподілу.

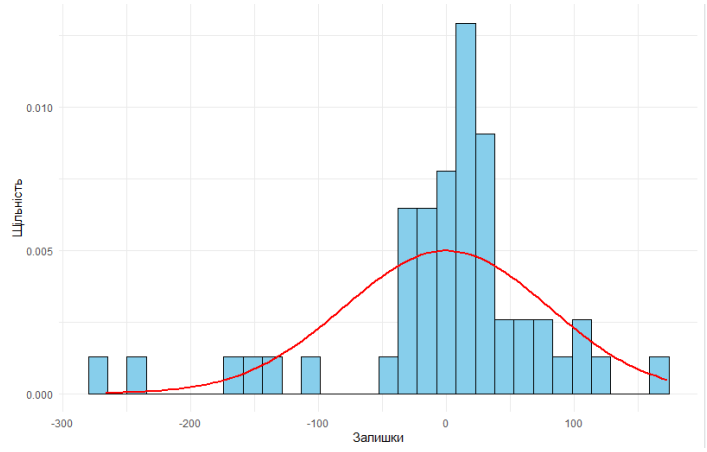

Рисунок 2.10 - Гістограма залишків моделі з накладеною кривою нормального розподілу
Джерело: розроблено автором

Гістограма залишків моделі з накладеною кривою нормального розподілу також показує відхилення залишків від нормального розподілу. Видно, що розподіл залишків асиметричний і має кілька піків, що не відповідає кривій нормального розподілу (червона лінія). Ліва частина гістограми містить кілька великих негативних залишків, а права частина має менші позитивні залишки, що також вказує на порушення припущення нормальності.

Ще одним важливим аспектом для забезпечення коректності регресійної моделі є перевірка відсутність автокореляції залишків, оскільки це може вплинути на оцінки коефіцієнтів регресії та їхню статистичну значущість. Автокореляція означає, що залишки моделі не є незалежними, що порушує одне з основних припущень лінійної регресії. Для виявлення автокореляції використаємо тест Дарбіна-Уотсона, який дозволяє оцінити, чи існує лінійна залежність між послідовними залишками.

```
        Durbin-Watson test

data:  model
DW = 1.8406, p-value = 0.25
alternative hypothesis: true autocorrelation is greater than 0
```

Рисунок 2.11 - Тест Дарбіна-Уотсона для перевірки автокореляції залишків
Джерело: розроблено автором

Для побудованої моделі значення DW становить 1.8406, що досить близьке до 2. У даному тесті це значення використовується як еталон, і вважається, що коли розрахункове значення наближене до 2, то автокореляція відсутня. Критичні значення при рівні значущості 0.05 та кількості факторів 3 становлять dL=1.42 та dU=1.67. Значення DW для моделі знаходиться між dU=1.67та 4−dU=2.33, що підтверджує відсутність автокореляції залишків.

P-значення тесту Дарбіна-Уотсона становить 0.25, що значно вище рівня значущості 0.05. Це означає, що немає підстав відхилити нульову гіпотезу про відсутність автокореляції, тобто залишки не є автокорельованими.

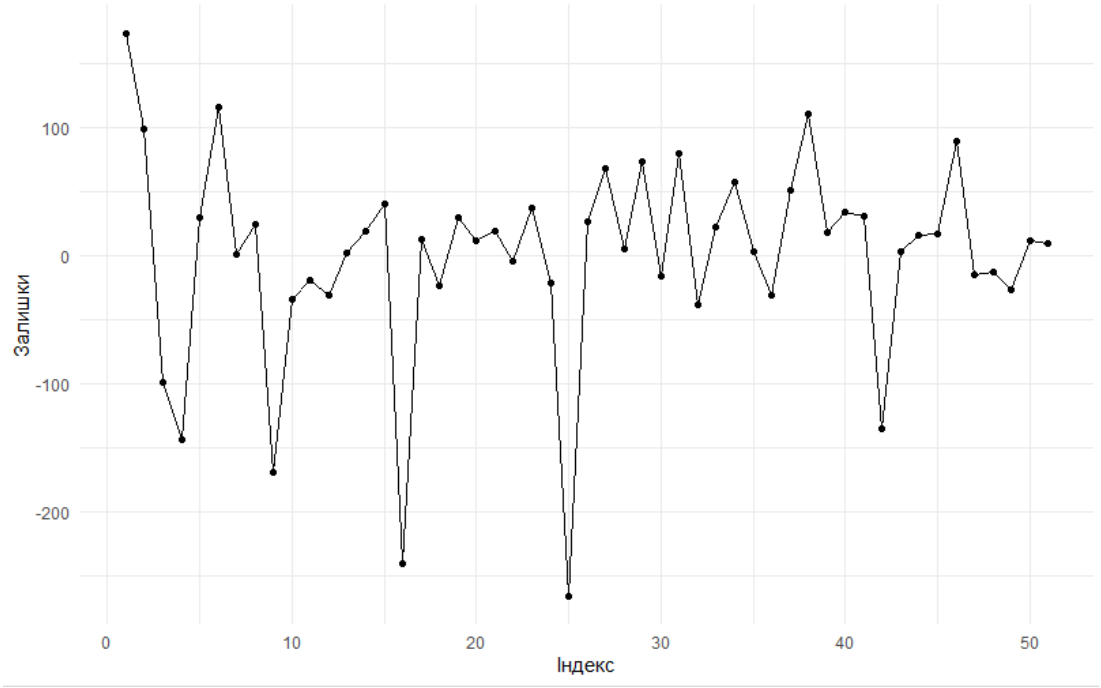

Рисунок 2.12 - Графік залишків моделі для тесту Дарбіна-Уотсона
Джерело: розроблено автором

Графік залишків моделі підтверджує висновки гіпотезу про відсутність автокореляції. Залишки моделі не демонструють чіткої тенденції чи систематичного патерну, що свідчить про випадковий розподіл залишків. Значення залишків коливаються навколо нуля, і відсутні тривалі послідовності позитивних або негативних залишків, що є характерним для випадків з автокореляцією.

Таким чином, результати тесту Дарбіна-Уотсона та аналіз графіку залишків моделі показують, що в моделі немає автокореляції залишків. Це підтверджує адекватність побудованої моделі та свідчить про те, що припущення регресійного аналізу не порушуються.

Після проведення тесту Дарбіна-Уотсона додатково побудуємо графік автокореляційної функції (ACF). Якщо залишки моделі не є автокорельованими, то значення автокореляцій на графіку ACF будуть випадковими і не демонструватимуть систематичної структури. Натомість, якщо автокореляція присутня, графік ACF покаже значущі автокореляції на різних лагах, що може вказувати на необхідність перегляду моделі або додавання нових змінних.

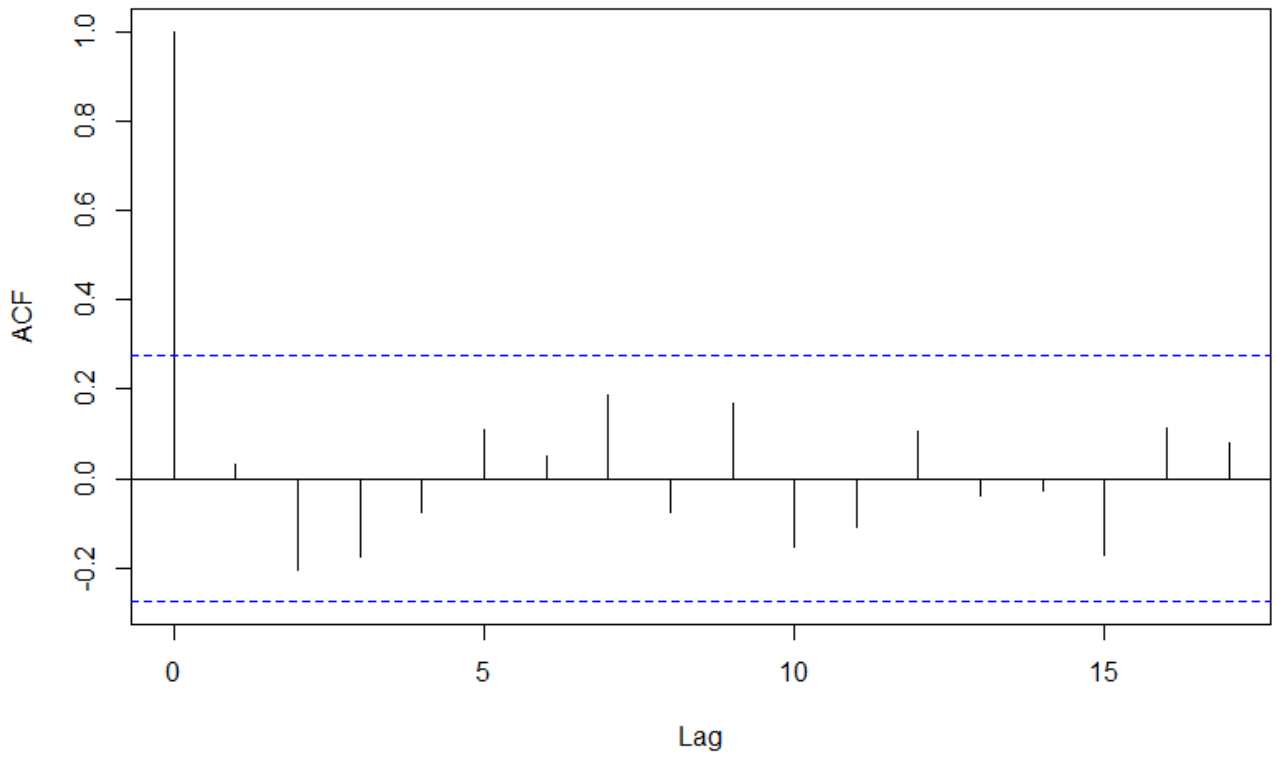

Рисунок 2.13 - Графік автокореляційної функції для залишків моделі
Джерело: розроблено автором

Значення автокореляції на Лаг 0 дорівнює 1, що є нормальним результатом, оскільки це просто кореляція залишків з самими собою. Це підтверджує, що залишки повністю корелюють з собою на нульовій затримці.

Спостерігається, що всі значення автокореляції для різних лагів є близькими до нуля і не виходять за межі довірчих інтервалів. Це вказує на відсутність статистично значущої автокореляції на цих лагах, що означає, що залишки моделі не демонструють систематичних патернів або тенденцій, які могли б вказувати на автокореляцію.

Результати, представлені на графіку, узгоджуються з результатами тесту Дарбіна-Уотсона, який також свідчить про відсутність автокореляції у залишках.

Після дослідження моделі на нормальний розподіл залишків та автокореляцію, наступним кроком є перевірка гетероскедастичності за допомогою тесту Бройша-Пагана. Це допоможе визначити, чи є дисперсія залишків постійною, що є важливою умовою для коректного застосування регресійного аналізу.

```
        studentized Breusch-Pagan test

data:  model
BP = 5.2341, df = 3, p-value = 0.1554
```

Рисунок 2.14 - Результати тесту Бройша-Пагана для моделі регресії
Джерело: розроблено автором

Значення тесту Бройша-Пагана (BP) є мірою відхилення дисперсії залишків від константної. У даному випадку значення BP = 5.2341, що є меншим за критичне значення хі-квадрат розподілу для 3 ступенів свободи при рівні значущості 0.05 (7.815). Це свідчить про відсутність гетероскедастичності, що підтверджується p-значенням 0.1554, яке є більшим за звичайний рівень значущості 0.05. Таким чином, ми не маємо підстав відхилити нульову гіпотезу про відсутність гетероскедастичності. Це означає, що залишки моделі мають постійну дисперсію і гетероскедастичність не є проблемою для даної моделі.

Візуалізуємо результати тесту на графіку:

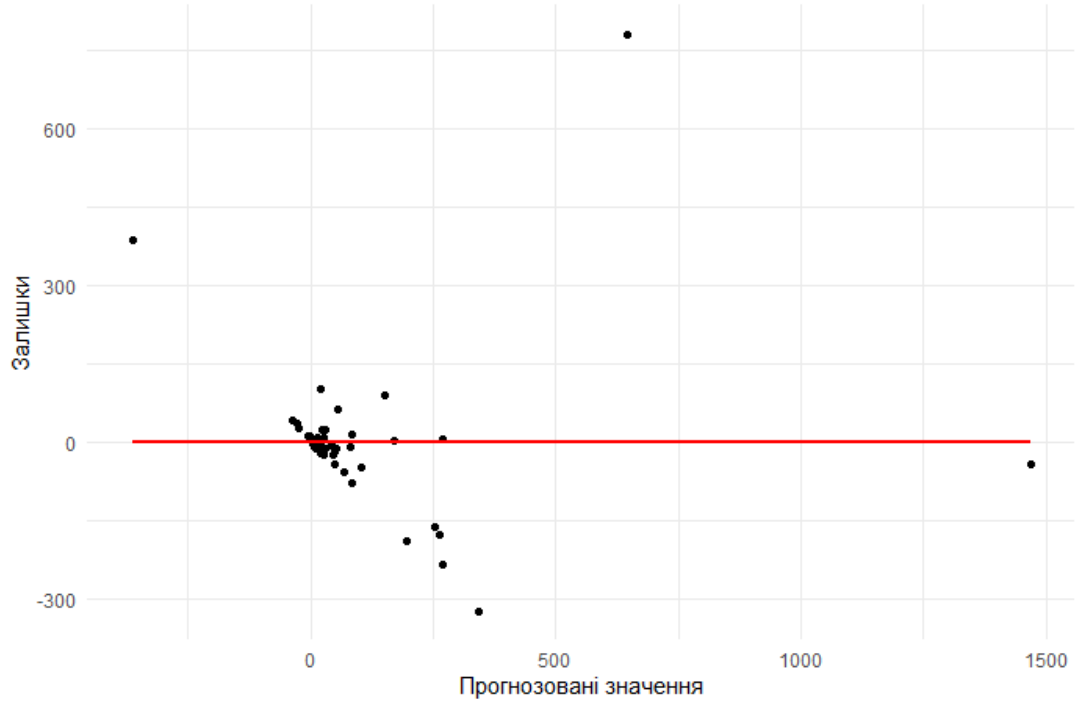

Рисунок 2.15 - Графік залишків моделі регресії для тесту Бройша-Пагана
Джерело: розроблено автором

Графік залишків моделі регресії підтверджує ці висновки. На графіку видно, що залишки моделі розподіляються випадково навколо нуля, без явного патерну або систематичних відхилень. Це свідчить про відсутність гетероскедастичності. Відсутність значних кластерів або систематичних

відхилень підтверджує, що дисперсія залишків є постійною, що є важливою передумовою для коректного застосування методу найменших квадратів (МНК).

Отже, результати тесту Бройша-Пагана та аналіз графіку залишків моделі регресії свідчать про відсутність гетероскедастичності в моделі. Це підтверджує адекватність побудованої моделі та вказує на те, що припущення регресійного аналізу не порушуються.

2.2 Аналіз кластеризації демографічних показників у країнах Азії

Статистичний кластерний аналіз ґрунтується на характеристиках порівнюваних спостережень або наборів даних для їх групування у кластери. У цьому процесі гомогенність та гетерогенність розглядаються як внутрішні та зовнішні властивості кластерів. Це означає, що об'єкти всередині одного кластеру повинні бути схожими між собою, тоді як об'єкти з різних кластерів повинні мати відмінності. Для ефективного проведення кластерного аналізу необхідно обрати відповідний алгоритм кластеризації, визначити міру подібності між об'єктами та правильно інтерпретувати отримані результати. Кластерний аналіз знаходить застосування у багатьох галузях, включаючи маркетинг, де використовується для сегментації ринку, біологію для класифікації видів, соціальні науки для виявлення груп з подібними характеристиками, а також в інших сферах, де потрібна організація та аналіз великих обсягів даних [23]. Особливе значення кластерний аналіз має в економічній сфері, зокрема для дослідження демографічних показників, що дозволяє виявляти важливі тенденції та закономірності у демографічних процесах, що є актуальним для економічного планування та прогнозування.

На відміну від класифікаційних задач, кластерний аналіз не потребує попередніх припущень про структуру даних та не обмежує способи представлення досліджуваних об'єктів. Він дозволяє аналізувати різні типи даних, такі як

інтервальні, частотні та бінарні. Важливо при цьому забезпечити, щоб змінні вимірювались на порівнюваних шкалах.

Кластерний аналіз виконує такі основні завдання:

1) створення типології або класифікації об'єктів;
2) виявлення корисних концептуальних схем для групування об'єктів;
3) формулювання гіпотез на основі аналізу даних;
4) перевірка гіпотез або дослідження для визначення наявності типів (груп) у даних.

Процес кластерного аналізу зазвичай включає такі етапи:

1) вибір вибірки для кластеризації;
2) визначення набору змінних для оцінки об'єктів;
3) розрахунок міри схожості між об'єктами;
4) використання методу кластерного аналізу для формування груп схожих об'єктів;
5) перевірка надійності отриманих результатів [24].

Ієрархічна кластеризація, також відома як кластеризація на основі зв'язків, здійснюється шляхом групування схожих об'єктів у вкладені кластери. Цей метод полягає в ітеративному об'єднанні менших кластерів у більші на основі їхньої схожості або близькості. Перевага ієрархічних методів класифікації полягає в їх наочності. Результати кластеризації відображаються у вигляді дендрограм, що в перекладі з грецької означає "дерево". Дендрограма чітко показує близькість окремих об'єктів і кластерів, а також графічно ілюструє послідовність їх об'єднання. Дендрограму також називають деревоподібною схемою або деревом об'єднання кластерів. На вертикальній осі дендрограми відображаються відстані між об'єктами, а на горизонтальній – самі об'єкти класифікації..

Ієрархічні методи кластеризації поділяються на агломеративні та дивізіональні. У агломеративній кластеризації кожен об'єкт спочатку утворює окремий кластер, після чого найближчі кластери поступово об'єднуються, доки не залишаться лише два кластери. Дивізіональна кластеризація починається з одного великого кластеру, який поступово розділяється на менші підкластерів [25].

Кластерний аналіз дає можливість обчислювати різні типи відстаней між об'єктами, створюючи матриці відстаней, які потім використовуються в процедурі об'єднання. Процедура об'єднання передбачає визначення правила об'єднання або зв'язку між двома кластерами [26].

Одним з найвідоміших методів кластерного аналізу є *метод Варда*. Він вирізняється застосуванням дисперсійного аналізу для оцінки відстаней між кластерами. Метод Варда базується на внутрішньогруповій сумі квадратів відхилень, що є сумою квадратів відстаней між кожним об’єктом і середнім значенням у кластері, до якого належить цей об'єкт. На кожному кроці об’єднуються ті два кластери, об'єднання яких викликає найменше зростання внутрішньогрупової суми квадратів. Цей метод спрямований на об’єднання найближчих кластерів для мінімізації внутрішньогрупової варіації [27].

Подібність між об'єктами визначається шляхом обчислення відстаней між ними: чим менша відстань, тим більша подібність об'єктів. Для адекватного визначення відстані, міра повинна мати такі властивості: симетрія, мінімальна відстань від об'єкта до центру ваги, монотонні зміни в описовому просторі та змістовна інтерпретація.

Однією з використовуваних мір відстаней у кластерному аналізі є квадрат евклідової відстані. Ця міра підносить до квадрату стандартну евклідову відстань, надаючи більшої ваги віддаленішим об'єктам. Формула розрахунку квадрата евклідової відстані виглядає так:

$$d(x, y) = \Sigma_i (x_i - y_i)^2 \qquad (2.12)$$

де x_i — координата об'єкта x за і-ою ознакою;

y_i — координата об'єкта y за і-ою ознакою x[28].

На рис. 2.1 можна побачити, як об'єкти згруповані в кластери на основі подібності, яка визначається за допомогою евклідової відстані:

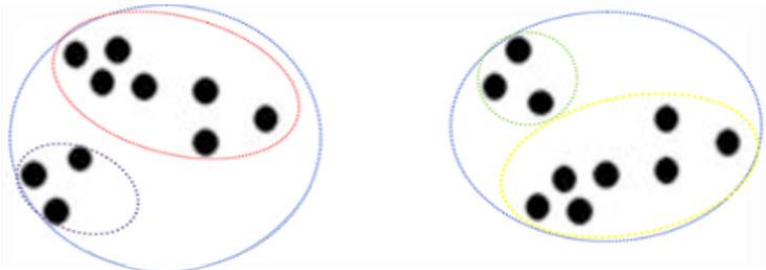

Рисунок 2.16 - Приклади кластеризації даних за допомогою квадрата евклідової відстані [29]

Метод k-середніх є широко використовуваним методом кластеризації, який прагне мінімізувати середню квадратичну відстань між точками в одному кластері. Незважаючи на те, що він не гарантує абсолютної точності, його простота і швидкість це компенсують. Доповнення методу k-середніх простою рандомізованою технікою для вибору початкових точок дозволяє досягти значної конкурентоспроможності з оптимальною кластеризацією. Експерименти показали, що таке доповнення значно підвищує швидкість і точність методу k-середніх. Перевагою цього методу є його зручність для кластеризації великої кількості спостережень, тоді як метод ієрархічного кластерного аналізу може стати перевантаженим і втратити наочність при великій кількості даних.

Основна ідея методу k-середніх полягає в тому, щоб визначити кластери таким чином, щоб мінімізувати загальну варіацію всередині кожного кластеру. Це означає, що елементи в межах одного кластеру повинні бути максимально схожими один на одного. Для досягнення цього методу використовується кілька алгоритмів, але стандартний підхід полягає в тому, щоб обчислити загальну варіацію всередині кла Формула виглядає наступним чином:

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2 \qquad (2.13)$$

де x_i - точка даних, що належить до кластера C_k,

μ_k — середнє значення точок у кластері C_k [30].

Алгоритм методу k-середніх складається з наступних кроків:

1) спочатку випадковим чином обирається набір з k центральних точок (центроїдів). Для визначення оптимального значення k необхідно проаналізувати дані та оцінити кількість окремих груп або кластерів у наборі даних;
2) для кожної точки даних обчислюється відстань між нею та кожною з центральних точок, після чого точка відноситься до групи з найближчим до неї центром;
3) на основі класифікованих точок перераховується центр групи, шляхом обчислення середнього значення всіх векторів групи.
4) повторюються кроки 2 і 3 задану кількість разів або до тих пір, поки зміна центрів груп не буде меншою за наперед задане значення.

Загальна варіація в межах кластера визначається наступним чином:

$$\sum_{k=1}^{K} W(C_k) = \sum_{k=1}^{K} \sum_{x_i \in C_k} (x_i - \mu_k)^2 \qquad (2.14)$$

Цей метод однак має кілька недоліків. По-перше, потрібно заздалегідь визначити кількість кластерів, що не завжди є очевидним. По-друге, результат роботи алгоритму залежить від вибору початкових центрів кластерів, тому виконуючи алгоритм кілька разів, можна отримати різні результати [31].

Після проведення комплексного економетричного аналізу моделі переходимо до проведення кластерного аналізу за допомогою програмного забезпечення STATISTICA, яке широко використовується для статистичного аналізу, візуалізації даних і обробки інформації.

Враховуючи наявні демографічні дані країн Азії, а саме показники чисельності населення, земельної площі, чистої зміни населення та коефіцієнта народжуваності, можна визначити, які країни мають схожі демографічні профілі. Це допоможе виявити основні тенденції та порівняти демографічні особливості різних груп країн.

Аналіз кластерів дозволить не лише зрозуміти подібності між країнами, але й виявити суттєві відмінності. Наприклад, країни з високою чисельністю населення, такі як Китай та Індія, можуть бути класифіковані окремо від країн з меншою чисельністю, але високою щільністю населення, таких як Сінгапур і Гонконг. Це дозволить виявити демографічні аномалії та винятки, що потребують окремого дослідження.

На початку дані були нормалізовані для забезпечення порівнянності між різними показниками. Нормалізація даних необхідна, оскільки демографічні показники мають різні одиниці вимірювання і масштаби. Це дозволяє уникнути ситуацій, коли певні показники домінують над іншими, що могло б призвести до некоректних результатів кластеризації.

На першому етапі дослідження побудовано ієрархічну дендрограму, яка відображає результати кластерного аналізу демографічних даних країн Азії. Дендрограма була створена за допомогою методу Варда, використовуючи евклідову відстань для визначення схожості між об'єктами.

На дендрограмі по вертикальній осі відображено відстань між кластерами (міра зв'язку), а по горизонтальній осі – самі країни. З цієї діаграми можна побачити, як країни групуються у кластери на різних рівнях ієрархії.

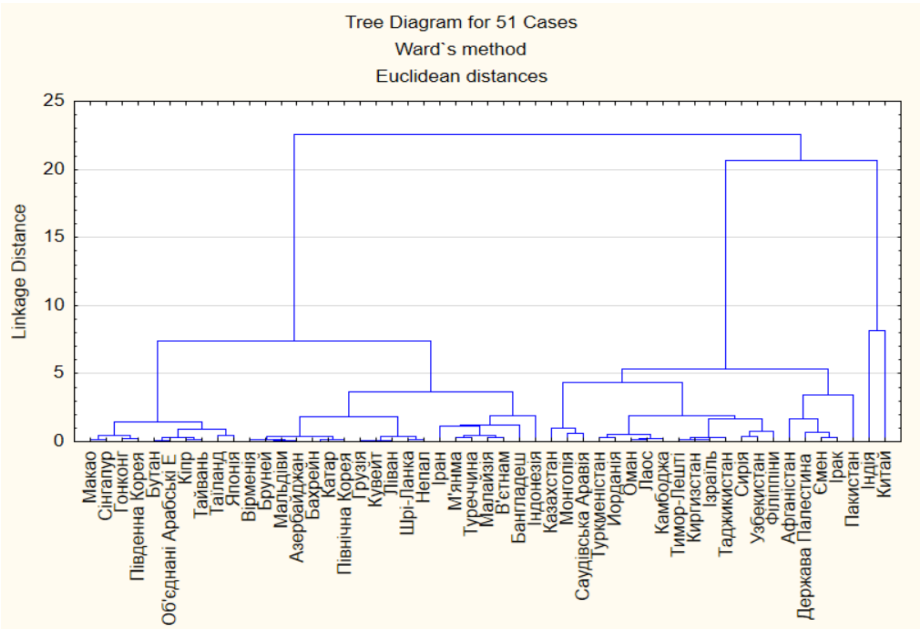

Рисунок 2.17 - Ієрархічна дендрограма кластеризації країн Азії за демографічними показниками

Джерело: розроблено автором

Аналізуючи дендрограму, можна виділити три основні кластери: перший кластер включає країни від Індії до Китаю, другий кластер об'єднує країни від Макао до Індонезії, а третій кластер складається з країн від Казахстану до Пакистану.

Побудуємо графік агломерації, який показує відстані зв'язку між кластерами на кожному кроці об'єднання:

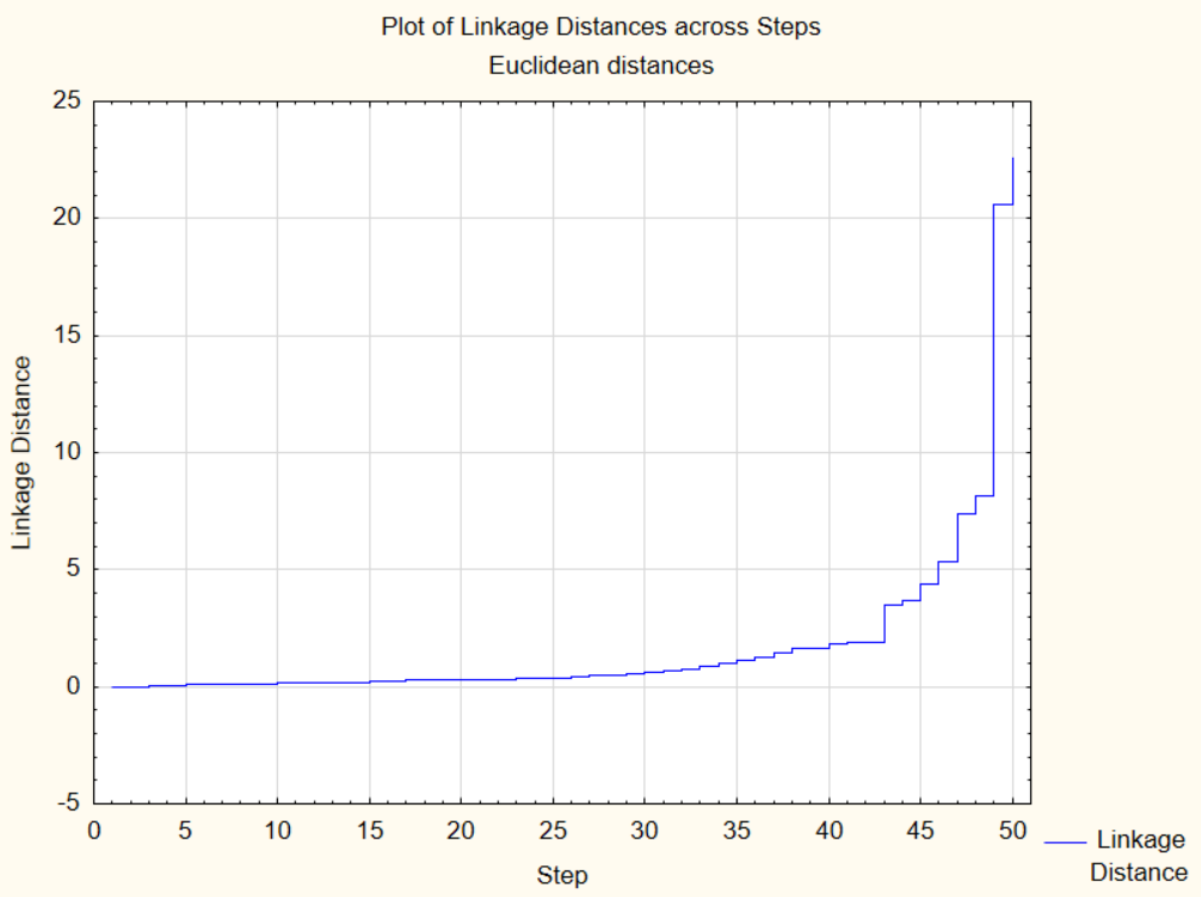

Рисунок 2.18 - Графік агломераційних відстаней на кожному кроці
Джерело: розроблено автором

Побудувавши графік агломерації, який показує відстані зв'язку між кластерами на кожному кроці об'єднання, бачимо, що на початкових етапах відстань збільшується повільно, вказуючи на об'єднання дуже схожих об'єктів. Однак на пізніших етапах відстань зв'язку різко зростає, що свідчить про об'єднання вже значно різних кластерів.

На основі цього графіка агломерації можна зробити висновок, що оптимальна кількість кластерів для даного аналізу знаходиться перед стрімким зростанням відстані, а саме до 45-го кроку.

Графік агломерації підтверджує висновки, зроблені на основі дендрограми, де було визначено три основні кластери. Поєднання цих двох підходів демонструє, що вибір трьох кластерів для методу k-середніх є обґрунтованим. Такий підхід дозволяє врахувати природну структуру даних і забезпечити точні та змістовні результати кластеризації, що відображають основні демографічні тенденції в країнах Азії.

	Means and Standard Deviations (cluster)	
Case No.	Mean	Std.Dev.
Китай	2,308773	3,695875
Індія	3,122096	2,920001
Індонезія	0,566510	0,425869
Пакистан	1,115704	0,994086
Бангладеш	0,076169	0,491138
Японія	-0,468971	0,540259
Філіппіни	0,315410	0,456779
В'єтнам	-0,116077	0,176765
Іран	0,033908	0,528449
Туреччина	-0,075744	0,179983
Таїланд	-0,380463	0,485998
М'янма	-0,070484	0,075248
Південна Корея	-0,619407	0,668338
Ірак	0,390334	0,838601
Афганістан	0,763240	1,444549
Саудівська Аравія	0,292082	0,568112
Узбекистан	0,118030	0,488765
Малайзія	-0,246391	0,139158
Ємен	0,424760	0,975761
Непал	-0,220291	0,077517
Північна Корея	-0,330077	0,088828
Тайвань	-0,549519	0,453452
Шрі-Ланка	-0,283706	0,080969
Сирія	0,113848	0,480807
Казахстан	0,519128	0,893026
Камбоджа	-0,154814	0,236236
Йорданія	-0,064053	0,519681
Азербайджан	-0,388393	0,129658
Таджикистан	0,093003	0,759639
Об'єднані Арабські Е	-0,482875	0,322958
Ізраїль	-0,000952	0,651963
Гонконг	-0,703344	0,692850
Лаос	-0,131820	0,306213
Киргизстан	0,022081	0,633774
Туркменістан	-0,029167	0,420863
Сінгапур	-0,634943	0,566788
Ліван	-0,303724	0,169172
Держава Палестина	0,152207	0,979466
Оман	-0,093651	0,369539
Кувейт	-0,277948	0,151387
Грузія	-0,277662	0,146345
Монголія	0,184511	0,581164
Вірменія	-0,445197	0,177660
Катар	-0,380205	0,062049
Бахрейн	-0,384353	0,061089
Тимор-Лешті	0,006376	0,731330
Кіпр	-0,545200	0,368167
Бутан	-0,508813	0,305571
Макао	-0,611847	0,495609
Мальдіви	-0,419893	0,112810
Бруней	-0,418184	0,113519

Рисунок 2.19 - Середні значення та стандартні відхилення демографічних показників для країн у кожному кластері, визначених за методом Варда

Джерело: розроблено автором

Для детальнішого аналізу ієрархічної дендрограми доцільно побудувати таблицю Середніх значень та стандартних відхилень. Вона показує середні значення демографічних показників та розподіл значень навколо середнього для

кожної країни у кластері. Це корисно для розуміння загальних тенденцій та внутрішньої варіативності в межах кожного кластеру.

Бачимо, що країни першого кластеру (Індія та Китай) мають найвищі середні значення населення, землі та чистої зміни населення, а також найбільші стандартні відхилення, що вказує на значну варіативність цих показників. Другий кластер, який включає країни від Макао до Індонезії, візуально характеризується нижчими середніми значеннями та меншими стандартними відхиленнями, що свідчить про більш однорідні демографічні показники. Третій кластер, що включає країни від Казахстану до Пакистану, має середні значення, які є вищими ніж у другому кластері, але нижчими ніж у першому, і помірні стандартні відхилення, що вказує на певну варіативність даних.

Це ще підтверджує доцільність вибору трьох кластерів для методу k-середніх.

Після оцінки середніх значень та стандартних відхилень демографічних показників для кожної країни за методом Варда, отримано загальне уявлення про варіативність та однорідність даних в межах кластерів, сформованих ієрархічною кластеризацією. Наступним кроком є проведення кластеризації методом k-середніх. Почнему з аналізу описової статистики для кожного кластеру, сформованого цим методом.

Це потрібно для того, щоб:

1) порівняти результати, отримані різними методами кластеризації, та забезпечити більш глибоке розуміння характеристик кожної групи країн;
2) оцінити, чи підтверджуються результати ієрархічної кластеризації методом Варда, та визначити відмінності, які можуть виникнути при застосуванні методу k-середніх;
3) отримати детальнішу картину про внутрішню структуру та варіативність кластерів, сформованих методом k-середніх, що дозволить зробити більш точні та надійні висновки щодо демографічних характеристик кластерів.

Основна різниця між попередньою таблицею середніх значень та стандартних відхилень за методом Варда і аналізом описової статистики за методом k-середніх полягає у представленні даних. У методі Варда таблиця відображала середні значення та стандартні відхилення для кожної країни окремо, що дозволяло оцінити індивідуальні характеристики кожної країни в межах кластерів. Натомість метод k-середніх надає середні значення та стандартні відхилення для кожного кластеру загалом, що дозволяє оцінити характеристики та варіативність всього кластеру в цілому.

Це забезпечить більш цілісне розуміння демографічних тенденцій у кожному кластері та допоможе виявити загальні закономірності, характерні для груп країн, об'єднаних методом k-середніх.

Нижче на рисунках 1, 2 та 3 наведені результати описової статистики для кожного кластеру:

Variable	Descriptive Statistics for Cluster 1 (cluster) Cluster contains 2 cases Mean	Standard Deviation	Variance
Населення (млн)	4,792434	0,007510	0,00006
Чиста зміна населення (тис)	2,880623	4,733234	22,40351
Земельна площа (тис км^2)	3,895319	3,171019	10,05536
Коефіцієнт народжуваності	-0,706638	0,730698	0,53392

Рисунок 2.20 - Описова статистика для кластеру 1 (метод k-середніх)
Джерело: розроблено автором

Variable	Descriptive Statistics for Cluster 2 (cluster) Cluster contains 29 cases Mean	Standard Deviation	Variance
Населення (млн)	-0,215908	0,157442	0,024788
Чиста зміна населення (тис)	-0,252465	0,227746	0,051868
Земельна площа (тис км^2)	-0,288197	0,238969	0,057106
Коефіцієнт народжуваності	-0,648734	0,516204	0,266467

Рисунок 2.21 - Описова статистика для кластеру 2 (метод k-середніх)
Джерело: розроблено автором

Variable	Descriptive Statistics for Cluster 3 (cluster) Cluster contains 20 cases Mean	Standard Deviation	Variance
Населення (млн)	-0,166177	0,277855	0,077203
Чиста зміна населення (тис)	0,078011	0,631543	0,398846
Земельна площа (тис км^2)	0,028354	0,541630	0,293363
Коефіцієнт народжуваності	1,011328	0,662763	0,439254

Рисунок 2.22 - Описова статистика для кластеру 3 (метод k-середніх)
Джерело: розроблено автором

Проаналізувавши описову статистику кожного кластеру бачимо, що кожен кластер має свою специфіку та середні значення, що дозволяє зрозуміти, в чому саме вони відрізняються один від одного.

Перший кластер, що складається з двох країн (Китай та Індія), має найвищі середні значення населення та земельної площі, але значення коефіцієнта народжуваності є найнижчим серед трьох кластерів. Це свідчить про те, що ці країни мають великі території та населення, але їх демографічний ріст є відносно низьким.

Другий кластер, що включає 29 країн, характеризується найменшими середніми значеннями за всіма показниками, зокрема, населенням, чистою зміною населення та земельною площею. Коефіцієнт народжуваності в цьому кластері також нижчий, ніж у третьому кластері, але вищий, ніж у першому. Це означає, що в даний кластер входять країни з відносно низькими демографічними та територіальними показниками.

Третій кластер, який включає 20 країн, має середні значення, що знаходяться між значеннями першого та другого кластерів. Найвищим показником у цьому кластері є коефіцієнт народжуваності, що свідчить про високі темпи демографічного росту в цих країнах, попри те, що середні значення населення та земельної площі є нижчими, ніж у першому кластері.

Тепер, проаналізувавши результати обох методів кластеризації, можемо зробити висновки, що обидва методи кластеризації демонструють подібну тенденцію розподілу країн на кластери з подібними демографічними та територіальними характеристиками. Країни з більшим населенням та площею утворюють окремий кластер, а країни з меншими значеннями цих показників групуються в інші кластери. Це свідчить про узгодженість результатів між методом Варда і методом k-середніх.

Для того, щоб оцінити значущість відмінностей між кластерами, можна побудувати дисперсійний аналіз (ANOVA). Цей аналіз дозволяє визначити, чи існують статистично значущі відмінності між середніми значеннями змінних у різних кластерах. F-значення розраховується як відношення між дисперсією між

групами (міжкласовою) до дисперсії всередині груп (внутрішньокласовою) і високі значення F вказують на значні відмінності між групами. Значення ймовірності (p-value) менші за 0.05 підтверджують статистичну значущість відмінностей між кластерами за всіма показниками. Це може бути корисно для подальшого аналізу і розуміння особливостей кожного кластеру в контексті демографічних показників.

Variable	Analysis of Variance (cluster)					
	Between SS	df	Within SS	df	F	signif. p
Населення (млн)	47,83902	2	2,16098	48	531,3026	0,000000
Чиста зміна населення (тис)	18,56611	2	31,43389	48	14,1754	0,000015
Земельна площа (тис км^2)	32,77177	2	17,22823	48	45,6531	0,000000
Коефіцієнт народжуваності	33,65918	2	16,34082	48	49,4357	0,000000

Рисунок 2.23 - Дисперсійний аналіз (ANOVA) для кластерів
Джерело: розроблено автором

На рис. Бачимо, що значення F для кожної змінної є високими, що свідчить про значну відмінність між групами. Значення p-value для всіх змінних є меншими за 0.05, що підтверджує значущість цих відмінностей.

Для наочності результатів побудуємо графік середніх значень для кожного кластеру. На графіку представлено середні значення всіх змінних для трьох кластерів.

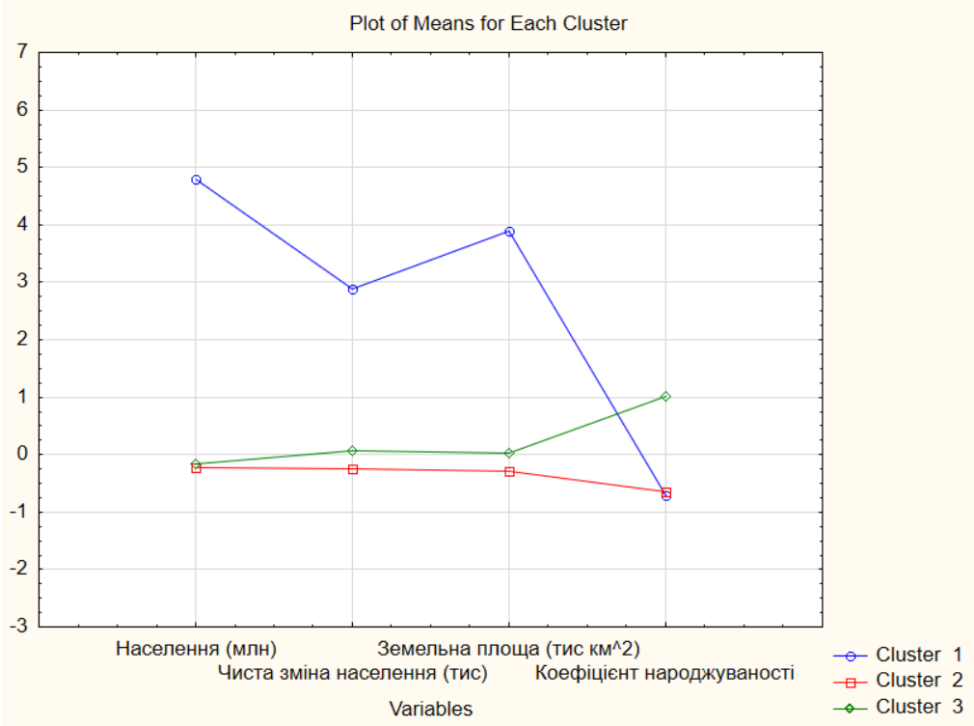

Рисунок 2.24 - Графік середніх значень для кожного кластера
Джерело: розроблено автором

Цей графік чітко демонструє відмінності між кластерами за всіма ключовими змінними. Він підтверджує, що перший кластер включає найбільші за населенням і територією країни з низьким рівнем народжуваності, Кластер 2 - найменші країни за всіма показниками, а третій кластер характеризується середніми значеннями з позитивним коефіцієнтом народжуваності.

Для аналізу значень евклідових відстаней між кластерами звернемося до таблиці евклідових відстаней між кластерами.

Cluster Number	Euclidean Distances between Clusters (cluster) Distances below diagonal Squared distances above diagonal		
	No. 1	No. 2	No. 3
No. 1	0,000000	13,10122	12,58682
No. 2	3,619561	0,00000	0,74192
No. 3	3,547791	0,86135	0,00000

Рисунок 2.25 - Евклідові відстані між кластерами
Джерело: розроблено автором

Евклідові відстані між кластерами допомагають визначити ступінь схожості чи відмінності між ними. Таким чином, з цієї таблиці видно, що Кластер 2 і Кластер 3 є найбільш схожими між собою, тоді як Кластер 1 найбільше відрізняється від Кластера 2. Результати цього аналізу підтверджують висновки, зроблені на основі середніх значень, стандартних відхилень та дисперсійного аналізу.

Для подальшого детального аналізу демографічних та економічних особливостей країн, що входять до кожного з визначених кластерів, звернемося до таблиць, які містять інформацію про членів кластеру та їх відстані від відповідного центру. Ці таблиці містять інформацію, що допоможе зрозуміти рівень однорідності та варіативності всередині кожного кластеру. Аналіз цих даних дозволить оцінити специфічні характеристики країн у кожному кластері та зробити висновки щодо їхньої економічної ситуації, враховуючи їхнє розташування відносно центрів кластерів.

linkage	Members of Cluster Number 1 (cluster) and Distances from Respective Cluster Center Cluster contains 2 cases
	Distance
Китай	2,030789
Індія	2,030789

Рисунок 2.26 - Члени Кластеру 1 та Відстані від Центру Кластеру
Джерело: розроблено автором

linkage	Members of Cluster Number 2 (cluster) and Distances from Respective Cluster Center Cluster contains 29 cases
	Distance
Бангладеш	0,551633
Японія	0,365789
В'єтнам	0,254038
Іран	0,529027
Туреччина	0,290897
Таїланд	0,258370
М'янма	0,343866
Південна Корея	0,486472
Малайзія	0,126241
Непал	0,236675
Північна Корея	0,106081
Тайвань	0,295455
Шрі-Ланка	0,236538
Камбоджа	0,424234
Азербайджан	0,073252
Об'єднані Арабські Емірати	0,169963
Гонконг	0,553569
Сінгапур	0,426169
Ліван	0,316502
Кувейт	0,306729
Грузія	0,305983
Вірменія	0,096594
Катар	0,136739
Бахрейн	0,140111
Кіпр	0,242790
Бутан	0,182543
Макао	0,365657
Мальдіви	0,107139
Бруней	0,105340

Рисунок 2.27 - Члени Кластеру 2 та Відстані від Центру Кластеру
Джерело: розроблено автором

linkage	Members of Cluster Number 3 (cluster) and Distances from Respective Cluster Center Cluster contains 20 cases
	Distance
Індонезія	0,873497
Пакистан	1,203598
Філіппіни	0,378124
Ірак	0,322527
Афганістан	0,955613
Саудівська Аравія	0,629073
Узбекистан	0,129575
Ємен	0,435403
Сирія	0,248056
Казахстан	0,735404
Йорданія	0,320527
Таджикистан	0,268828
Ізраїль	0,286984
Лаос	0,419564
Киргизстан	0,252185
Туркменістан	0,298522
Держава Палестина	0,423110
Оман	0,365163
Монголія	0,408194
Тимор-Лешті	0,315870

Рисунок 2.28 - Члени Кластеру 3 та Відстані від Центру Кластеру
Джерело: розроблено автором

Китай та Індія, що розташовані у першому клатері мають найбільші середні значення населення та земельної площі. Це свідчить про наявність значних людських і земельних ресурсів, що є ключовими факторами для економічного зростання. Однак, низький коефіцієнт народжуваності може вказувати на проблеми з майбутнім демографічним ростом, що може позначитися на ринку праці та

соціальній сфері. Відстані від центрів кластерів для обох країн майже ідентичні (2,030789), що вказує на їх високу схожість за обраними показниками.

Другий кластер включає 29 країн. Найменшу відстань до центру кластеру має Азербайджан (0,073252), а найбільшу - Гонконг (0,553569). Це вказує на те, що Азербайджан є найбільш типовим представником цього кластеру, а Гонконг - найменш типовим. Цей кластер характеризується найнижчими середніми значеннями за всіма показниками. Низькі значення населення та земельної площі свідчать про обмежені ресурси, що можуть впливати на економічний розвиток. Незначні коливання цих показників вказують на відносну стабільність, але й можливу сприйнятливість до зовнішніх економічних впливів.

Третій кластер включає 20 країн. Найменшу відстань до центру кластеру має Узбекистан (0,129575), а найбільшу - Пакистан (1,203598). Це свідчить про те, що Узбекистан є найбільш типовим представником цього кластеру, а Пакистан - найменш типовим. Країни цього кластеру мають середні значення демографічних показників, які знаходяться між значеннями першого та другого кластерів. Високий коефіцієнт народжуваності вказує на потенційний демографічний ріст, що може стимулювати економічний розвиток за умови відповідного управління ресурсами. Проте, варіативність даних може свідчити про економічну нестабільність або нерівномірний розподіл ресурсів.

ВИСНОВКИ

У даній дипломній роботі проведено всебічний аналіз і моделювання демографічних процесів у країнах Азії. Проведене дослідження вказує на загальну тенденцію до зниження рівня народжуваності та старіння населення. Виявлено нерівномірний розподіл молодого працездатного населення, що створює додаткові соціально-економічні виклики для країн регіону. Значні міграційні потоки також впливають на демографічну ситуацію та потребують розробки ефективних політичних рішень для інтеграції мігрантів.

В рамках дослідження побудовано множинну лінійну регресійну модель для оцінки впливу різних факторів на чисельність населення країн Азії. Основними незалежними змінними моделі виступили земельна площа, чиста зміна населення та коефіцієнт народжуваності. Аналіз моделі показав, що найбільший вплив на зміну чисельності населення мають площа землі, чиста зміна населення та коефіцієнт народжуваності в порядку спадання. Модель має високий рівень пояснювальної здатності, що свідчить про її адекватність та надійність.

Проведено кілька статистичних тестів для перевірки адекватності моделі та її припущень, включаючи тест Шапіро-Вілка, тест Дарбіна-Уотсона, тест на гетероскедастичність та інші. Результати тестів вказали на відсутність автокореляції залишків та відсутність гетероскедастичності, що підтверджує надійність моделі. Однак тест Шапіро-Вілка вказав на певні відхилення від нормальності розподілу залишків, що може свідчити про важливість до подальшого аналізу.

За допомогою методів кластерного аналізу виконано групування та класифікацію країн Азії за вказаними чинниками. Виявлено три основні кластери країн за характеристиками демографічних показників: перший кластер включає країни з найбільшим населенням та найбільшою територією, другий кластер охоплює країни з найменшими значеннями демографічних та територіальних показників, а третій кластер характеризується середніми значеннями з високими

темпами демографічного росту, що підтверджує узгодженість результатів між методом Варда і методом k-середніх.

Для підтримки стабільного рівня населення та запобігання негативним тенденціям рекомендується стимулювання народжуваності через соціальні програми підтримки сімей, покращення умов для мігрантів, забезпечення їм доступу до освіти, охорони здоров'я та ринку праці, а також розробка програм соціального захисту для літніх людей. Отримані результати аналізу та моделювання розробленої моделі можуть бути корисними для аналізу демографічних процесів в інших регіонах світу та розробки нових моделей прогнозування демографічних змін.

Таким чином, проведене дослідження дозволило глибше зрозуміти демографічні процеси в країнах Азії та їхній вплив на соціально-економічний розвиток, що сприятиме розробці ефективних політичних рішень у цій сфері.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1) Смоловик Л. Г. Основи демографії: конспект лекцій. Харків: НТУ "ХПІ", 2019. 120 с.
2) Ткаченко І. А. Демографічні процеси: аналіз та прогнозування: навчальний посібник. Київ: МАУП, 2008. 200 с.
3) Demographic statistics: methodology and definitions. Madrid: National Statistics Institute, 2019. P 85.
4) Glossary. DataBank. The World Bank. URL: https://databank.worldbank.org/metadataglossary/population-estimates-and-projections/series/SM.POP.NETM. (accessed: 10.05.2024).
5) Francisco J. Goerlich Gisbert. Life Potential as a Basic Demographic Indicator. Documentos de trabajo (Fundación BBVA), vol. 8, 2012, P. 3-17.
6) Кистовська І. М. Стратегії соціально-економічного розвитку регіонів: монографія. Тернопіль: ТНЕУ, 2008. 340 с.
7) Нємец Л. М., Сегіда К. Ю., Заволока Ю. Ю. Моделювання демографічної ситуації як важливий напрям суспільно-географічного дослідження населення регіону в контексті системного підходу. Вісник Херсонського національного університету. С. 1–6.
8) Саріогло В.Г. Методологія оцінки якості життя населення України. URL: http://dspace.nbuv.gov.ua/bitstream/handle/123456789/11851/04-Sarioglo.pdf (дата звернення: 14.05.2024).
9) Park S.-C., Ogawa N., Kim C. J., Sirivunnabood P., Le T.-H. Demographic Transition and Its Impacts in Asia and Europe. Manila: Asian Development Bank, 2021. P. 382.

10) Countries in Asia by Population. Worldometers. URL: https://www.worldometers.info/population/countries-in-asia-by-population/#google_vignette (дата звернення: 14.05.2024).

11) Cruz G. T., Laguna E. P. Confronting Asia's Challenging Demographic Realities. Asian Productivity Organization. 2023. Vol. 4–5. P. 1-33.

12) Пашко А. О. Статистичний аналіз даних: Методичні матеріали до курсу "Інтелектуальна обробка даних" для студентів. Київ: Київський національний університет імені Тараса Шевченка, 2019. 55 с.

13) Бідюк П. І. Прикладна статистика: Методичні матеріали до курсу. Київ: Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського", 2023. 186 с.

14) T-Test Criteria. EZTests. URL: https://www.eztests.xyz/criteria/ttest/#google_vignette (дата звернення: 18.05.2024).

15) Прикладна економетрика: навчальний посібник у двох частинах. Частина 1 / рецензенти: О. В. Піскунова, В. О. Бабенко. Харків: ХНЕУ ім. С. Кузнеця, 2016. 235 с.

16) Критерій Шапіро-Вілка. Навчальні матеріали. URL: https://westudents.com.ua/glavy/88658-kritery-shapro-vlka-W.html (дата звернення: 18.05.2024).

17) Кондіус І. Г. Електронний підручник "Моделювання та методи і системи машинного інтелекту". URL: https://elib.lntu.edu.ua/sites/default/files/elib_upload/%D0%95%D0%BB%D0%B5%D0%BA%D1%82%D1%80%D0%BE%D0%BD%D0%BD%D0%B8%D0%B9%20%D0%BF%D1%96%D0%B4%D1%80%D1%83%D1%87%D0%BD%D0%B8%D0%BA%20%D0%9A%D0%BE%D0%B

D%D0%B4%D1%96%D1%83%D1%81%20(%D0%9C%D0%9C%D0%86%D0%A1%D0%9C)/page12.html (дата звернення: 18.05.2024).

18) Гурянова Л. С. Економетрика: навчальний посібник. Харків: ХНЕУ ім. Кузнеця, 2015. 384 с.

19) Слоневський Р. Б., Жук І. В. Економетрика: навчальний посібник. Тернопіль: ТНЕУ, 2012. 222 с.

20) Гаврюшин В. А. Методичні вказівки до курсу "Бакалаврська робота": рекомендації для студентів. Київ: Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського", 2021. 52 с.

21) Stav A. Introduction to Applied Econometrics. URL: http://andriystav.cc.ua/Downloads/AppliedEco/01_Intro.pdf (дата звернення: 19.05.2024).

22) Основні етапи економетричного моделювання. URL: http://um.co.ua/5/5-6/5-67031.html#google_vignette (дата звернення: 19.05.2024).

23) Кластерний аналіз. Mind the Graph. URL: https://mindthegraph.com/blog/uk/%D0%BA%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%BD%D0%B8%D0%B9-%D0%B0%D0%BD%D0%B0%D0%BB%D1%96%D0%B7/ (дата звернення: 22.05.2024).

24) Уманець В. С. Створення та порівняльний аналіз кластеризаційних алгоритмів розпізнавання статистичних залежностей: дис. ... магістра наук : 122 "Комп’ютерні науки та інформаційні технології". Київ, 2018. 120 с.

25) Сенієв А., Тіцька А. Лабораторна робота №1 з дисципліни "Методи аналізу даних": Ієрархічний кластерний аналіз. Львів: Львівський національний університет імені Івана Франка, 2024. 15 с.

26) Кластерний аналіз. Тернопільський національний технічний університет імені Івана Пулюя. URL:

https://wiki.tntu.edu.ua/%D0%9A%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%BD%D0%B8%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D1%96%D0%B7 (дата звернення: 23.05.2024).

27) Войтович С. Я., Букало Н. А. Методика застосування кластерного аналізу для ідентифікації образів споживачів освітніх послуг. Вісник економічних досліджень. 2018. № 4. 10 с.

28) Бегун С. Застосування кластерного аналізу для дослідження демографічної ситуації в регіоні. Економічний часопис Східноєвропейського національного університету імені Лесі Українки. 2016. № 2. 7 с.

29) Chapter 4: Data Mining Techniques. Section 1: Introduction. URL: https://www.hypertextbookshop.com/dataminingbook/public_version/contents/chapters/chapter004/section001/blue/page003.html (дата звернення: 23.05.2024).

30) Івохін І. А. Основні поняття кластеризації та постановка задачі: Лекція 11. Київ: Київський національний університет імені Тараса Шевченка. 14 с.

31) Івохін І. А. Стандартні методи кластеризації даних: Лекція 2. Київ: Київський національний університет імені Тараса Шевченка. 14 с.

ДОДАТКИ

Додаток А

№	Країна	Населення (млн)	Річна зміна (%)	Чиста зміна населення (тис)	Щільність населення (осіб/км^2)	Земельна площа (тис км^2)	Чиста міграція	Коефіцієнт народжуваності	Середній вік	Міська популяція (%)
1	Китай	1 425,6714	-0,0002	-215,9850	152	9 388,2110	-310 220	1,2	39	65
2	Індія	1 428,6277	0,0081	11 454,4900	481	2 973,1900	-486 136	2	28	36
3	Індонезія	277,5341	0,0074	2 032,7830	153	1 811,5700	-49 997	2,1	30	59
4	Пакистан	240,4857	0,0198	4 660,7960	312	770,8800	-165 988	3,3	21	35
5	Бангладеш	172,9543	0,0103	1 767,9470	1,329	130,1700	-309 977	1,9	27	41
6	Японія	123,2945	-0,0053	-657,1790	338	364,5550	99 994	1,3	49	94
7	Філіппіни	117,3374	0,0154	1 778,3590	394	298,1700	-69 996	2,7	25	47
8	В'єтнам	98,8590	0,0068	672,0940	319	310,0700	-82 700	1,9	33	40
9	Іран	89,1728	0,0070	622,1970	55	1 628,5500	-39 998	1,7	33	74
10	Туреччина	85,8162	0,0056	474,9580	112	769,6300	-318 067	1,9	32	77
11	Таїланд	71,8013	0,0015	104,2490	141	510,8900	18 999	1,3	40	52
12	М'янма	54,5780	0,0074	398,6910	84	653,2900	-34 998	2,1	30	33
13	Південна Корея	51,7841	-0,0006	-31,7510	533	97,2300	29 998	0,9	44	82
14	Ірак	45,5046	0,0227	1 008,4380	105	434,3200	-6 000	3,4	20	71
15	Афганістан	42,2399	0,0270	1 111,0830	65	652,8600	-65 846	4,4	17	26
16	Саудівська Аравія	36,9470	0,0148	538,2050	17	2 149,6900	28 998	2,4	31	83
17	Узбекистан	35,1639	0,0155	536,2920	83	425,4000	-19 999	2,8	27	49
18	Малайзія	34,3085	0,0109	370,3040	104	328,5500	48 997	1,8	31	78
19	Ємен	34,4498	0,0224	753,2110	65	527,9700	-29 914	3,6	19	37
20	Непал	30,8966	0,0114	349,0100	216	143,3500	-62 012	2	24	22
21	Північна Корея	26,1608	0,0035	91,4050	217	120,4100	-2 000	1,8	36	63
22	Тайвань	23,9233	0,0013	29,8820	676	35,4100	23 999	1,2	42	80
23	Шрі-Ланка	21,8936	0,0028	61,4360	349	62,7100	-77 495	2	33	19
24	Сирія	23,2270	0,0498	1 101,7650	126	183,6300	757 103	2,7	22	53
25	Казахстан	19,6066	0,0108	208,6350	7	2 699,7000	0	3	30	57
26	Камбоджа	16,9448	0,0106	176,9840	96	176,5200	-29 998	2,3	27	26
27	Йорданія	11,3371	0,0045	51,1830	128	88,7800	-157 392	2,7	24	85
28	Азербайджан	10,4127	0,0053	54,5770	126	82,6580	0	1,7	32	57
29	Таджикистан	10,1435	0,0192	190,7560	72	139,9600	-19 999	3,1	22	28
30	Об'єднані Арабські Емірат	9,5169	0,0080	75,7420	114	83,6000	0	1,4	34	94
31	Ізраїль	9,1745	0,0151	136,2110	424	21,6400	9 999	2,9	29	92
32	Гонконг	7,4916	0,0004	2,7440	7,135	1,0500	19 999	0,8	46	93
33	Лаос	7,6338	0,0139	104,3040	33	230,8000	-9 999	2,4	24	37
34	Киргизстан	6,7353	0,0158	104,7240	35	191,8000	-9 999	2,9	24	37
35	Туркменістан	6,5161	0,0133	85,3300	14	469,9300	-4 000	2,6	26	52
36	Сінгапур	6,0147	0,0065	39,0340	8,592	0,7000	26 998	1	43	80
37	Ліван	5,3539	-0,0247	-135,8090	523	10,2300	-177 331	2,1	29	97
38	Держава Палестина	5,3712	0,0231	121,1580	892	6,0200	-5 000	3,4	20	83
39	Оман	4,6444	0,0149	68,0860	15	309,5000	0	2,5	29	73
40	Кувейт	4,3101	0,0097	41,2350	242	17,8200	11 999	2,1	40	89
41	Грузія	3,7283	-0,0043	-16,1030	54	69,4900	-9 999	2,1	37	63
42	Монголія	3,4472	0,0144	48,7910	2	1 553,5600	-850	2,7	27	67
43	Вірменія	2,7780	-0,0009	-2,4990	98	28,4700	-5 000	1,6	35	67
44	Катар	2,7164	0,0079	21,2690	234	11,6100	0	1,8	34	99
45	Бахрейн	1,4855	0,0090	13,2760	1,955	0,7600	0	1,8	34	89
46	Тимор-Лешті	1,3606	0,0144	19,3000	91	14,8700	-5 000	3	21	35
47	Кіпр	1,2601	0,0069	8,6500	136	9,2400	5 000	1,3	39	65
48	Бутан	0,7874	0,0064	4,9690	21	38,1170	300	1,4	29	49
49	Макао	0,7041	0,0129	8,9810	23,472	0,0300	5 000	1,1	39	97
50	Мальдіви	0,5210	-0,0053	-2,7660	1,737	0,3000	-8 652	1,7	32	39
51	Бруней	0,4525	0,0078	3,5220	86	5,2700	0	1,7	33	80

Додаток Б

K_2	Число степенів свободи варіації для більшої дісперсії (K_1)									
	1	3	5	7	9	11	14	20	30	50
1	161	216	230	237	241	243	245	248	250	252
	4052	5403	5764	5928	6022	6082	6142	6208	6258	6302
3	10.3	9.28	9.01	8.88	8.81	8.76	8.71	8.66	8.62	8.58
	34.12	29.46	28.24	27.67	27.34	27.13	26.92	26.69	26.50	26.35
5	6.61	5.41	5.05	4.88	4.78	4.70	4.64	4.56	4.50	4.44
	16.26	12.06	10.97	10.45	10.15	9.96	9.77	9.55	9.38	9.24
7	5.59	4.35	3.97	3.79	3.68	3.60	3.52	3.44	3.38	3.32
	12.25	8.45	7.46	7.00	6.71	6.54	6.35	6.15	5.98	5.85
9	5.12	3.86	3.48	3.29	3.18	3.10	3.02	2.93	2.86	2.80
	10.56	6.99	6.06	5.62	5.35	5.18	5.00	4.80	4.64	4.51
11	4.84	3.59	3.20	3.01	2.90	2.82	2.74	2.65	2.57	2.50
	9.05	6.22	5.32	4.88	4.63	4.46	4.29	4.10	3.94	3.80
13	4.67	3.41	3.02	2.84	2.72	2.63	2.55	2.46	2.38	2.32
	9.07	5.74	4.86	4.44	4.19	4.02	3.85	3.67	3.51	3.37
15	4.54	3.29	2.90	2.70	2.59	2.51	2.43	2.33	2.25	2.18
	8.68	5.42	4.56	4.14	3.89	3.73	3.56	3.36	3.20	3.07
17	4.45	3.20	2.81	2.62	2.50	2.41	2.33	2.23	2.15	2.08
	8.40	5.18	4.34	3.93	3.68	3.52	3.35	3.16	3.00	2.89
19	4.38	3.13	2.74	2.55	2.43	2.34	2.26	2.15	2.07	2.00
	8.18	5.01	4.17	3.77	3.52	3.36	3.19	3.00	2.84	2.70
21	4.32	3.07	2.68	2.49	2.37	2.28	2.20	2.09	2.00	1.93
	8.02	4.87	4.04	3.65	3.40	3.24	3.07	2.88	2.72	2.58
23	4.28	3.03	2.64	2.45	2.32	2.24	2.14	2.04	1.96	1.88
	7.88	4.46	3.94	3.54	3.30	3.14	2.97	2.78	2.62	2.48
25	4.24	2.99	2.60	2.41	2.28	2.20	2.11	2.00	1.92	1.84
	7.77	4.68	3.86	3.46	3.21	3.05	2.89	2.70	2.54	2.40
27	4.21	2.96	2.57	2.37	2.25	2.16	2.08	1.97	1.88	1.80
	7.68	4.60	3.79	3.39	3.14	2.98	2.83	2.63	2.45	2.33
29	4.18	2.93	2.54	2.35	2.22	2.14	2.05	1.94	1.85	1.77
	7.60	4.54	3.73	3.33	3.08	2.92	2.77	2.57	2.41	2.27

Додаток В

Число ступенів свободи f	Рівень значимості α			
	0,10	0,05	0,01	0,001
1	6.31	12.70	63.70	637.00
2	2.92	4.30	9.92	31.60
3	2.35	3.18	5.84	12.90
4	2.13	2.78	4.60	8.61
5	2.01	2.57	4.03	6.86
6	1.94	2.45	3.71	5.96
7	1.89	2.36	3.50	5.40
8	1.86	2.31	3.36	5.04
9	1.83	2.26	3.25	4.78
10	1.81	2.23	3.17	4.59
11	1.80	2.20	3.11	4.44
12	1.78	2.18	3.05	4.32
13	1.77	2.16	3.01	4.22
14	1.76	2.14	2.98	4.14
15	1.75	2.13	2.95	4.07
16	1.75	2.12	2.92	4.01
17	1.74	2.11	2.90	3.96
18	1.73	2.10	2.88	3.92
19	1.73	2.09	2.86	3.88
20	1.73	2.09	2.85	3.85
21	1.72	2.08	2.83	3.82
22	1.72	2.07	2.82	3.79
23	1.71	2.07	2.81	3.77
24	1.71	2.06	2.80	3.74
25	1.71	2.06	2.79	3.72
26	1.71	2.06	2.78	3.71
27	1.71	2.05	2.77	3.69
28	1.70	2.05	2.76	3.66
29	1.70	2.05	2.76	3.66
30	1.70	2.04	2.75	3.65
40	1.68	2.02	2.70	3.55
60	1.67	2.00	2.66	3.46
120	1.66	1.98	2.62	3.37
∞	1.64	1.96	2.58	3.29

Короткий звіт за результатами перевірки кваліфікаційної роботи інтернет-системою Strikeplagiarism:

Назва: Щирба_Білик_антиплагіат

Автор: Щирба.

Дата звіту: 2024-05-31

9,70%
Коефіцієнт подібності № 1:

1,61%
Коефіцієнт подібності № 2:

Довжина фрази для коефіцієнта подібності № 2:

25

Кількість слів: 8 981 **Число знаків:** 67 112

Адреси пропущені при перевірці: