Завада О.П

МОДЕЛІ ТА МЕТОДИ
В ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЯХ

(Текст лекцій)

Львів 2012

У тексті лекцій описані сучасні методи опрацювання інформації. Детально розглянена теорія часових рядів:ипобудова трендів, сезонних та циклічних коливань, аналіз причинно-наслідкових залежностей.

Особлива увага приділена процесам дейтамайнінгу, а також кластерному і факторному аналізу.

ТЕКСТ ЛЕКЦІЙ

ВСТУП. Поняття інформаційного суспільства

На теперішньому етапі людство переходить від індустріального суспільства до інформаційного.

За визначенням комісії Євросоюзу «інформаційне суспільство – це суспільство, в якому діяльність людей здійснюється на основі використання послуг, що надаються за допомогою інформаційних технологій та технологій зв'язку».

У найбільш розвинених країнах світу кількість людей, професія яких пов'язана з інформацією чи інформатикою (програмісти; бібліотекарі; журналісти; аналітики в банках та на підприємствах;) вже перевищила кількість робітників і селян разом узятих. А в США кількість таких людей, діяльність яких пов'язана з обробкою інформації, навіть зрівнялася із сферою обслуговування.

Частка інформаційних технологій у ВВП країн, які входять до Євросоюзу, уже досягла 5%.

Характерними рисами інформаційного суспільства є:

- головною формою розвитку є інформаційна економіка;
- інформаційні технології набирають глобального характеру;
- реалізовано вільний доступ кожної людини до інформаційних ресурсів усієї цивілізації.

Ряд вчених-економістів в Україні вважають, що нашій державі замість того, щоб підтримувати нерентабельні або низькорентабельні індустріальні підприємства варто здійснювати інновації в першу чергу в інформаційну індустрію. Кадровий потенціал для цього в Україні достатній.

ТЕМА 1. СУЧАСНІ СХОВИЩА ДАНИХ.

Трансакційні та аналітичні системи обробки даних.

В 1970-1980-х роках комп'ютерні системи на той час використовувалися, в основному, лише для операційної роботи (проведення таких трансакцій, як оплата за рахунками, розрахунки з постачальниками, купівля квитків на транспорт тощо). Вони називалися системами обробки трансакцій (On-Line Transaction Processing, OLTP). До OLTP-систем можна віднести системи «Сигма-абітурієнт», «Сигма-студент», «Сигма-викладач», які виконують такі трансакції, як видача екзаменаційних відомостей, зміна прізвища студентки, зміна посади чи вченого звання викладача тощо. Звичайно, OLTP-система має змогу давати відповіді на прості запити, наприклад, чи є сьогодні купейні квитки до Києва, видати список студентів групи ЕКЕ-41, підрахувати кількість професорів на економічному факультеті,…

Основною вимогою до OLTP-систем була вимога цілісності (тобто, якщо асистент ставав доцентом, то відповідна зміна повинна була бути одночасно зроблена і в навчальній частині, і у відділі кадрів, і в бухгалтерії. Інформація опрацьовувалася засобами реляційних СУБД, наприклад, MySQL. Запити забезпечувалися мовою SQL.

В 1980-х роках практично повністю завершився процес комп'ютеризації бізнесу. Підприємства та організації накопичили величезні обсяги даних, які стосувалися різних аспектів їхньої діяльності. Виникло розуміння того, що ці масиви можуть бути дуже корисними. На їхній основі можна виконувати глибокий аналіз, виявляти приховані закономірності функціонування економічних систем з метою якісного прийняття управлінських рішень.

Відповідні комп'ютерні системи почали називати системами аналітичної обробки (On-Line Analysis Processing, OLAP). До OLAP-систем аналітики можуть звертатися із складними запитами, такими як «визначити середній час між виставленням рахунків за газ та їх оплатою в розрізі різних груп клієнтів» тощо. Основою OLAP-систем є сховища даних (які можуть зберігати терабайти інформації). Інформація опрацьовується потужними СУБД типу ORACLE. Аналітичні запити частково виконуються мовою SQL, проте часто потребують застосування мови високого рівня типу JAVA чи спеціалізованих пакетів.

Реляційні бази даних

Теорія реляційних баз даних була розроблена в 1970 році Е. Коддом (E. Codd).

Інформація в таких базах зберігається у вигляді таблиць (файлів) спеціального вигляду.

Поля (атрибути, реквізити) кожної із таблиць реляційної БД повинні задовольняти таким умовам:

1) усі поля кожної із таблиць є атомарними (неподільними);

2) одне чи декілька полів (реквізитів) кожної таблиці утворюють ключ (простий або складений). Два рядки з однаковим ключем не допускаються;

3) у випадку складеного ключа кожен не ключовий реквізит повинен повно залежати від ключа (тобто він не може залежати лише від його частини).Наприклад, коли ключем є пара «батько, мати», а не ключовими реквізитами діти, то в реляційних базах не повинно бути «його дітей», «її дітей» та «спільних дітей»;

4) будь яка транзитивна залежність між полями у записах не допускається. Наприклад, у записі «факультет, кафедра, викладач, вчене звання» має місце транзит. При переході викладача на якусь кафедру іншого факультету може відбутися спотворення (втрата цілісності) інформації;

5) такої всі таблиці в реляційних базах даних повинні бути зв’язані за ключами.

При виконанні умов 1-5 кажуть, що база даних знаходиться в третій нормальній формі Кодда. Третя нормальна форма (3НФ) є цінною тим, що вона гарантує цілісність бази. Зазначимо, що до 3НФ перейти можна завжди.

Робота з даними в реляційних БД здійснюються за допомогою мови запитів SQL.

Сховища даних, вітрини

Концепція сховищ даних була сформульована Білом Інмоном (Bill Inmon) в 1992 році в роботі “Building the Data Warehouse”.

Сховище даних (Data Warehouse) – це предметно орієнтована, інтегрована, прив’язана до часу та незмінна сукупність даних, призначена для підтримки прийняття рішень.

Інтегрованість. Дані в сховище надходять з різних джерел, вони можуть мати різні формати та способи кодування. Сховище повинно мати засоби для приведення таких даних в єдиний формат.

Предметна орієнтація. На відміну від оперативних баз даних, де дані організуються відповідно до процесів (відвантаження товару, нарахування зарплати), в сховищах дані організовані відповідно до напрямків діяльності (замовники, постачальники,…).

Підтримка хронології, незмінність. Таким чином в сховищах дані зберігають свою істинність в довільний момент часу, тоді як в OLTP-системах дані при проведеннях трансакцій змінюються. Зокрема, в системі «Сигма» при переході викладача на іншу посаду, на іншу кафедру,.. попередня інформація втрачається. В сховищах же тільки додаються нові дані.

Враховуючи той факт, що при управлінні підприємством чи організацією потрібно забезпечувати як поточні трансакції, так і аналітику, сховище повинно поєднувати можливості як OLTP-, так і OLAP-систем.

Загальну схему сховищ даних представимо на рис.1.1.

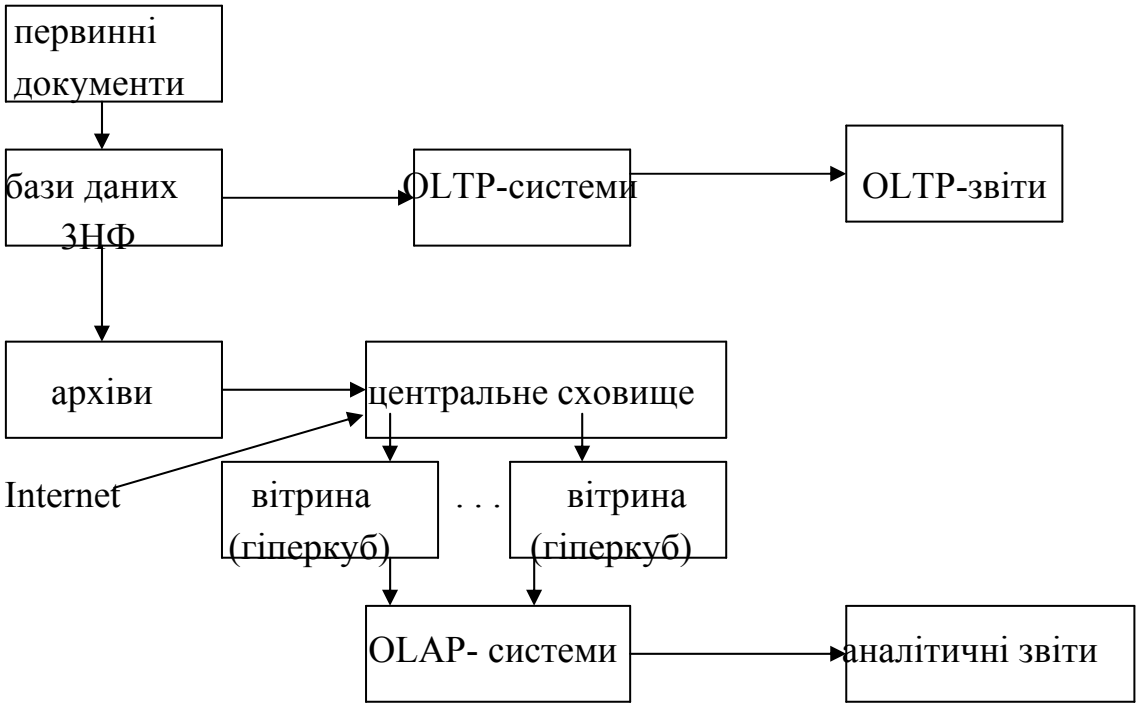

Рис.1.1.

Для баз даних, які підтримують OLTP-системи, важливим є швидкість трансакцій та забезпечення цілісності. Очевидно, найкращим машинним представленням у цьому випадку є реляційна модель з використанням третьої нормальної форми.

Час від час ця оперативна інформація передається в архіви (іноді попередньо вона узагальнюється). Вимога цілісності в архівах забезпечується автоматично, тому третя нормальна форма для них не є обов’язковою.

Центральне сховище повинно мати засоби для перегляду як оперативних даних, так і даних з архівів.

Проте дуже рідко аналітикам потрібна відразу вся інформація з центрального сховища. Для кожних конкретних аналітичних застосувань з цього сховища вирізають так звані кіоски (вітрини, Data Marts). До даних кіосків застосовуються складні аналітичні запити, а також математичні методи (прогнозування, кластерний та факторний аналіз, дерева рішень, нечітка логіка тощо)

Основною вимогою до кіосків є швидкість опрацювання даних. Найкращим тут є представлення даних у вигляді векторів, матриць, гіперкубів вищої розмірності.

Для утворення вітрин використовується мова запитів SQL або алгоритмічна мова.

ТЕМА 2. .ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ СХОВИЩ ДАНИХ (Дейтамайнінг)

Поняття дейтамайнінгу.

Із початку 1990-х менеджери почали виявляти бажання самим мати доступ до нагромаджених даних і самим аналізувати їх.

Класичний статистичний аналіз працює в режимі верифікації. Формулюється статистична гіпотеза, а потім при допомозі математичного апарату вона або підтверджується або спростовується. Проте на практиці виявилося потрібним також виявляти приховані, раніше невідомі зв'язки між даними.

Відповідні інструментальні засоби були створені і отримали назву засобів дейтамайнінгу (Data mining). Інша назва: засоби відкриття, видобуття знань (Knowladge Discovery).

Програмне забезпечення дейтамайнінгу працює в режимі відкриття (discovery mode), виявляючи невідомі раніше взірці (patterns) або шаблони.

У даних лекціях будемо розглядати деякі алгоритми дейтамайнінгу.

Дейтамайнінг – це клас аналітичного прикладного програмного забезпечення, яке підтримує рішення, виконуючи пошук за прихованими взірцями, шаблонами, формами.

Основні характеристики дейтамайнінгу:

- Інформація міститься у сховищах даних великого розміру.
- Середовище дейтамайнінгу звичайно орієнтується на архітектуру клієнт-сервер.
- Від користувача систем дейтамайнінгу не вимагається кваліфікованої програмістської підготовки.
- Інструментальні засоби дейтамайнінгу є сумісними з табличними процесорами.
- Результати дейтамайнінгу часто є несподіваними (а не 'те,що й потрібно було довести')і вимагають від менеджера-користувача вміння творчо мислити.

Типи інформації, які отримуються засобами дейтамайнінгу.

Інтелектуальний дейтамайнінг відкриває (видобуває) інформацію всередині баз і сховищ даних. Основними типами такої інформації є класифікація, кластеризація, асоціація, упорядкування, прогнозування.

Класифікація. Той чи інший об’єкт (наприклад, споживач) відноситься до певного класу.

Наприклад, особа відноситься до середнього класу, якщо вона задовольняє хоча би 5 із таких 8 умов:

- має власний автомобіль;
- має кількість кімнат хоча б на одну більше, ніж кількість членів сім’ї;
- має дачний будиночок;
- кожного літа відпочиває за кордоном;
- не хапається за додатковий підробіток;
- не користується субсидіями;
- має вдома цілодобовий Інтернет;
- має посудомиючу машину.

Кластеризація (групування). Об’єкти розділяються на групи, кластери. При цьому кількість цих груп наперед невідома. Наперед невідомі також характеристики цих груп. Кількість груп та характеристики цих груп є результатом застосування кластерного аналізу до сховища даних.

Наприклад, виявилося, що всі чоловіки поділяються на три кластери:

-однолюби (за все своє життя були близько знайомими лише з 2-4 жінками);

-звичайні чоловіки, таких 80% (20±5 жінок);

-донжуани (сотні жінок).

Асоціація. Виявляються зв’язки між подіями.

Наприклад, при купівлі дитячих памперсів дуже часто одночасно купують і дві пляшки пива.

Упорядкування, послідовність. Виявляються зв’язки в часі.

Приклад. Через 3 місяці після отримання зарплатної пластикової картки 20% клієнтів замовляють послугу ‘інформація при прихід зарплати на мобільний телефон’.

Прогнозування. Оцінюються майбутні значення показників (наприклад, значення попиту) на основі даних із сховища (економетричні методи, методи гнучкого прогнозування). Ці знання потрібні для розробки заходів щодо просування товарів. Наприклад, страхові компанії знають, що одружені живуть довше.

Дейтамайнінг, який виявляє наперед невідомі, нетривіальні та практично корисні знання, є дуже корисним при прийнятті рішень.

Досить часто економічний ефект від прийняття грамотного рішення в 10 разів перевищує вартість проекту дейтамайнінгу, незважаючи на те, що такий проект коштує від 350 тис. до 750 тис. доларів.

Окремими гілками дейтамайнінгу є Text Mining та Web Mining.

Text Mining виконує пошук нових знань в слабо структурованих текстових файлах (в основному, за допомогою комбінації ключових слів). Web Mining виявляє закономірності в поведінці користувачів з метою їх активного залучення. Наприклад, виявлення клієнтів, які здійснюють покупки в Інтернеті. (Ігровий Web-портал, який розглядається в магістерській роботі О. Крамара, планує аналітичне дослідження клієнтів на серверній частині.

Процеси дейтамайнінгу

В класичних OLAP-системах користувач повинен задавати як взірці пошуку, так і гіпотези.

В умовах дейтамайнінгу система бере ініціативу у свої руки і

Огляд програмного забезпечення дейтамайнінгу

Зараз на ринку програмних продуктів пропонуються десятки різних систем детамайнінгу.

Система Oracle 10g Data Mining

Ця система постачається як доповнення до СУБД ORACLE. Вона підтримує: класифікацію, асоціацію (алгоритм Apriori), кластеризацію, регресійний аналіз, а також аналіз важливості атрибутів (факторний аналіз?).

Oracle 10g Data Mining передбачає як програмний, так і графічний інтерфейс. Програмний інтерфейс включає мову JAVA, а також розширення мови запитів SQL із такими функціями як CLUSTER, FEATURE, PREDICTION.

Система має також опцію Anomaly Detection .

Система Clementine компанії SPSS.

В цю систему закладено широкий набір аналітичних алгоритмів, зокрема, класифікацію, прогнозування, пошуку асоціацій (GRI, Apriori, Sequence), метод головних компонент, дерева рішень.

Система Microsoft SQL Server Analysis Services

У 2000 році ця система містила тільки два алгоритми дейтамайнінгу: алгоритм побудови дерев рішень та алгоритм кластеризації. Проте вже в 2005 році вона має, крім того,

-алгоритм Assotiation Rules пошуку закономірностей типу А і Б → С, наприклад, для пошуку перехресних даних з продажів у електронній комерції;

-алгоритм Naive Bayes Це імовірнісна модель яка шукає, наприклад, відмінності між клієнтами, які припинили покупки в нашій фірмі і тими, які продовжують купувати;

-алгоритм Sequence Clustering поєднує прогнозування з кластеризацією.Клієнти діляться на кластери і відбувається прогнозування їхньої поведінки;

-алгоритм Time Series виявляє закономірності в різних часових послідовностях.

Система Microsoft SQL Server також має мову запитів SQL, розширену оператором

Create Mining Model .

В цілому варто зазначити, що зараз кожна фірма, яка розробляє і супроводжує бази та сховища даних, інтенсивно впроваджує у ці СУБД засоби дейтамайнінгу.

Користувач не мусить при цьому володіти відповідним математичним апаратом (економетрією, математичною логікою, матричною алгеброю тощо). Йому (достатньо володіти англійською мовою в обсязі програмістських термінів). Проте варто не забувати, що довільний математичний метод (алгоритм) є коректним тільки при виконанні певних умов (шукати обернену матрицю можна тільки тоді, коли її визначник не є нулем і не є близьким до нуля; будувати множинну регресію можна тільки при відсутності мультиколінеарності,…). Тому на кожній фірмі, установі бажано мати математично грамотну людину, яка знає не лише пункти меню конкретної програми, але й знає алгоритми дейтамайнінгу і умови їх застосування.

ТЕМА 3. КЛАСТЕРНИЙ АНАЛІЗ

Кластерний аналіз –це багатомірна статистична процедура (кластеризація), яка класифікує об'єкти або спостереження в однорідні групи. Набір усіх досліджуваних об'єктів розподіляється по підкласах, які називаються кластерами (cluster, згустками, класами, скупченнями, таксонами).

Синонімами до терміну кластеризація є: кластерний аналіз, сегментаційний аналіз, сегментація, таксономія, розпізнавання без навчання, автоматична класифікація, неконтрольована класифікація.

Основна мета кластеризації – розділити множину початкових даних на такі підмножини, групи, щоб об'єкти всередині кожної групи були подібними до себе, а об'єкти з різних груп – неподібними.

Основним поняттям кластерного аналізу є дистанція (відстань).

Вважаючи, що кожен об'єкт має n атрибутів, задамо його у вигляді точки в n-вимірному просторі.

Для обчислення відстані $dist(A,B)$ між об'єктами $A=(a_1,...,a_j,...,a_m)$ та $B=(b_1,...,b_j,...,b_m)$ можна використати різні формули:

-відстань Матхаттана $dist(A,B) = \sum\left|a_j - b_j\right|$

- Евклідову відстань $dist(A,B) = \sqrt{\sum\left(a_j - b_j\right)^2}$ (3.1)

-зважену Евклідову відстань $dist(A,B) = \sqrt{\sum w_j\left(a_j - b_j\right)^2}$

-коефіцієнт кореляції $dist(A,B) = \frac{\overline{A} * \overline{B}}{\left|\overline{A}\right| * \left|\overline{B}\right|} = \frac{\sum\left(a_j * b_j\right)}{\sqrt{\sum a_j^2 * \sum b_j^2}}$.

У тому випадку, коли значення атрибутів об'єкта не є числовими (наприклад, стать), то або ці значення потрібно перетворити у числові, або використати якусь з якісних мір подібності (коефіцієнти Рао, Хеммінга, Жаккарда тощо)

Нормалізація атрибутів.

Оскільки різні атрибути одного об'єкту можуть мати різну розмірність та різний діапазон, то їх необхідно нормалізувати, стандартизувати. Нехай, наприклад, для деякої групи людей є відомими їх вік, кількість дітей та зарплата. Потрібно визначити міру відстані (близькості) між цими людьми. Без виконання нормалізації класифікація цілком залежатиме тільки від зарплати, оскільки зарплата вимірюється сотнями і тисячами, а кількість дітей одиницями і дуже рідко десятками.

Нормалізація перетворює матрицю

$$A = \begin{pmatrix} a_{11} & \dots & a_{1m} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{nm} \end{pmatrix} \tag{3.2}$$

де n - кількість об'єктів, а m - кількість ознак, в матрицю

$$Z = \begin{pmatrix} z_{11} & \dots & z_{1m} \\ \dots & \dots & \dots \\ z_{n1} & \dots & z_{nm} \end{pmatrix} \tag{3.3}$$

Нормалізація виконується, наприклад, за формулами

$$z_{ij} = \frac{a_{ij} - \min a_j}{\max a_j - \min a_j} \tag{3.4}$$

або за формулами

$$z_{ij} = \frac{a_{ij} - \overline{a}_j}{S_j} \tag{3.5}$$

де $\overline{a}_j = 1/n \sum a_{ij}$, $S_j = \sqrt{1/n \sum \left(a_{ij} - \overline{a}_j\right)^2}$

Тепер у першому випадку всі атрибути $0 \le z_{ij} \le 1$.

У другому випадку значення z_{ij} будуть як додатними, так і відємними. Проте середнє арифметичне кожної ознаки $\overline{z}_j = 0$, а середнє квадратичне відхилення (а, отже, і дисперсія) кожної з ознак дорівнює одиниці.

Матриця відстаней (схожості) між обєктами

Нехай r_{rl} = віддаль між k-им та l-им рядками нормованої матриці(3.3). Тоді згідно (3.1),

$$r_{kl} = \sqrt{\sum \left(z_{kj} - z_{lj}\right)^2}$$

Побудована таким чином квадратна матриця

$$Z = (z_{rl}) = \begin{pmatrix} 0 & r_{12} & \dots & r_{1n} \\ & 0 & \dots & \dots \\ & & 0 & r_{n-1,n} \\ & & & 0 \end{pmatrix}$$

називається матрицею відстаней між об'єктами. Вона є симетричною відносно головної діагоналі.

Відстані між кластерами

На рисунку 3.1. множина із двадцяти об'єктів розбита на чотири кластери. Відстань між кожною парою цих об'єктів можна знаходити за довільною із формул (3.1), наприклад, нею може бути Евклідова відстань.

Центральним питанням в кластерному аналізі є обчислення відстані між кластерами, кожен з яких містить декілька об’єктів.

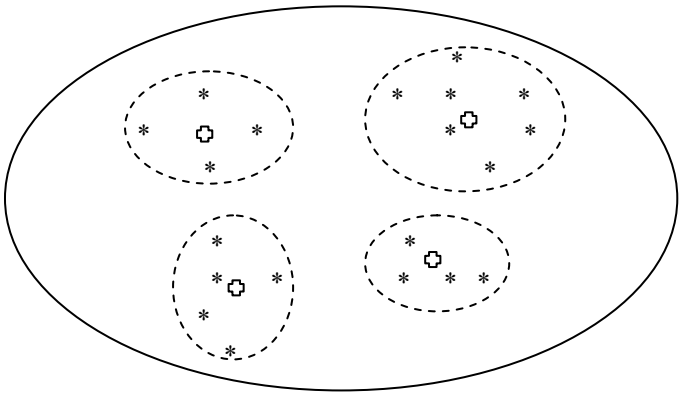

Рис. 3.1.

Існує багато варіантів обчислення такої відстані, наприклад,:

- відстань між найближчими сусідами (Nearest neighbor);
- відстань між найвіддаленішими сусідами (Furthest neighbor);
- відстань між центрами кластерів (Centroid clustering).

Вибір відстані між кластерами потрібно здійснювати на основі змістовного розуміння задачі. Наприклад, за методом «найближчого сусіда» від України до Казахстану набагато ближче, ніж до Німеччини чи Естонії. Проте, коли міряти віддалі між столицями, то найближче до Києва знаходяться Мінськ та Кишинів (отже, не дивно, що економіка нашої держави є най подібнішою до молдавської та білоруської), трохи далі – Вільнюс, ще трохи далі – Варшава і Бухарест, і ще далі – Москва. При вимірюванні між географічними центрами країн висновки знову ж будуть різними. Найкраще у цьому прикладі, мабуть, брати зважену (на кількість населення) евклідову міру між географічними центрами.

Ієрархічний кластерний аналіз.

Ідея методу є такою: на першому кроці всі n об’єктів розподіляються на n кластерів,. На другому кроці шукаються два найближчі кластери (кожен кластер на цьому кроці поки-що складається із одного об’єкта) і об’єднуються в один кластер. Таким чином після виконання другого кроку кількість кластерів стає рівною $n-1$. Така процедура повторюється до тих пір, поки усі кластери обєднаються в один. Як видно, цей алгоритм є навчальним методом без учителя.

Звичайно, обидва крайні випадки (кожен запис бази даних є окремим кластером; уся БД є одним кластером) не є потрібними на практиці.

Основною ідеєю кластеризації є пошук корисних взірців у базі, що має зробити її зрозумілішою при прийнятті рішень.

Тому сам кінцевий користувач повинен зупинитися на тій чи іншій кількості кластерів. Останнє можна вважати як недоліком, так і перевагою кластерного аналізу.

На заключному етапі кластерного аналізу знаходять центри кожного з кластерів та дають економічну інтерпретацію отриманим кластерам.

Результати кластеризації представляють дендрограмою, яка показує, які об'єкти обєдналися на якому етапі (тобто, які об'єкти виявилися найбільш подібними між собою), а також в деяких програмних системах наглядно демонструє відстані між кластерами.

Швидкий кластерний аналіз (метод K-центрів)

Кількість кластерів K аналітик фіксує наперед. Комп'ютер випадково (або аналітик свідомо) вибирає по одному елементу в кожному із цих кластерів. Далі розраховуються відстані від кожного іншого об'єкту до цих кластерів і виконуються об'єднання до тих пір, поки всі елементи будуть класифіковані.

Цей метод є швидшим, ніж ієрархічний. Його застосовують для аналізу великих баз даних. Метод K-центрів дає змогу пояснити економічно кожен кластер, але він не дає відповіді на те, який об'єкт коли приєднався до свого кластера. Дендрограма, отже, також не будується. Проте є багато застосувань, коли аналітика цікавить лише кількість елементів у кожному кластері та характеристики центрів кластерів, і зовсім не цікавлять конкретні об'єкти. Наприклад, виконуючи анонімне анкетування, ми хочемо дослідити населення України на приналежність до вищого, середнього та нижчого класів.

Зазначимо, що кластерний аналіз відноситься до методів дейтамайнінгу, основаних на збережених даних.

Введемо початкові статистичні дані для подальшого аналізу показників конкурентоспроможності одинадцяти країн. Для цього заповнимо таблицю (рис.3.2.).

2001 рік	1 Belarus	2 Bosnia and Herzegovina	3 Bulgaria	4 Croatia	5 Czech Republic	6 Macedonia, FYR	7 Poland	8 Russian Federation	9 Slovak Republic	10 Ukraine	11 Slovenia
ВВП	1239,16	1469,83	1719,17	4471,55	6048,74	1704,91	4975,89	2100,74	3924,55	780,73	9925,54
Середній дохід	349,96	529,13	567,32	1743,9	1875,1	511,47	1542,52	567,2	1295,1	202,99	4069,47
Очікувана тривалість життя	67	73	71	74	75	72	74	65	73	68	76
Темп росту ВВП	97,33	104,29	109,97	109,26	109,55	95,5	111,68	118,34	103,79	122,81	102,22

	Статистика										
	1 Belarus	2 Bosnia and Herzegovina	3 Bulgaria	4 Croatia	5 Czech Republic	6 Macedonia, FYR	7 Poland	8 Russian Federation	9 Slovak Republic	10 Ukraine	11 Slovenia
ВВП	3796,07	3121,22	4092,63	9665,07	13925,8	3053,03	8883,77	6925,89	10212,33	2275,56	18588,52
Середній дохід	988,91	989,81	1334,79	3500,72	3777,57	972,14	2462,43	1602,16	2640,92	640,05	6869,02
Очікувана тривалість життя	68	74	72	75	76	74	75	65	74	68	77
Темп росту ВВП	122,83	113,59	116,51	110,43	114,27	106,77	111,83	129,68	116	124,43	108,24

Рис.3.2. Вигляд електронної таблиці введення вхідних даних в систему STATISTICA за 2001 та 2006 роки

Модуль «Кластерний аналіз» викликається за допомогою **STATISTICA Module Switcher – Перемикач модулів STATISTICA** (рис.3.3).

Вибираємо ієрархічний метод (Joining (tree clustering)) у вікні Define Method of Cluster Analysis (рис.3.4).

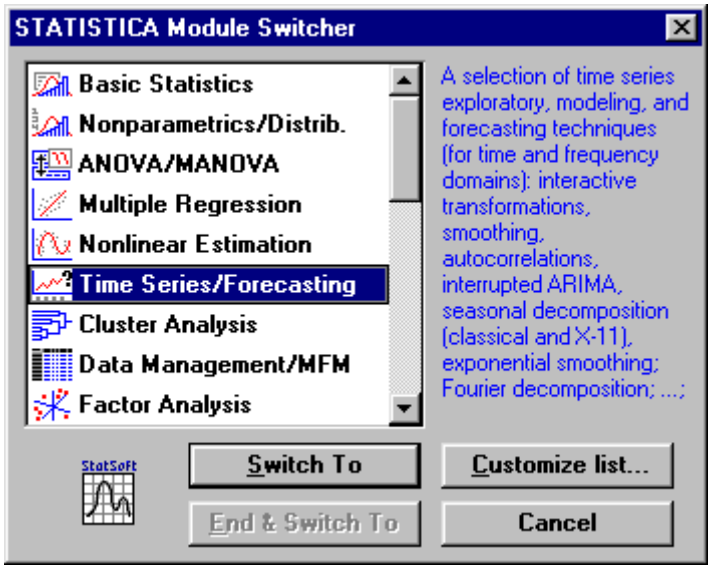

Рис. 3.3. Перемикач Модулів STATISTICA

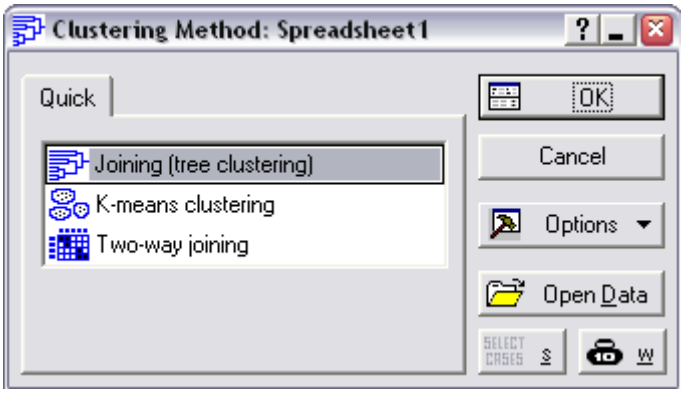

Рис. 3.4. Вибір методу кластерного аналізу.

На екрані буде виведене нове вікно Cluster Analysis Joining (Tree Clustering), в якому відкриваємо закладку Advanced (розширений) → Cluster→ Variables (columns) також вибираємо Complete linkage →Amalgamation (linkage) rule →OK.

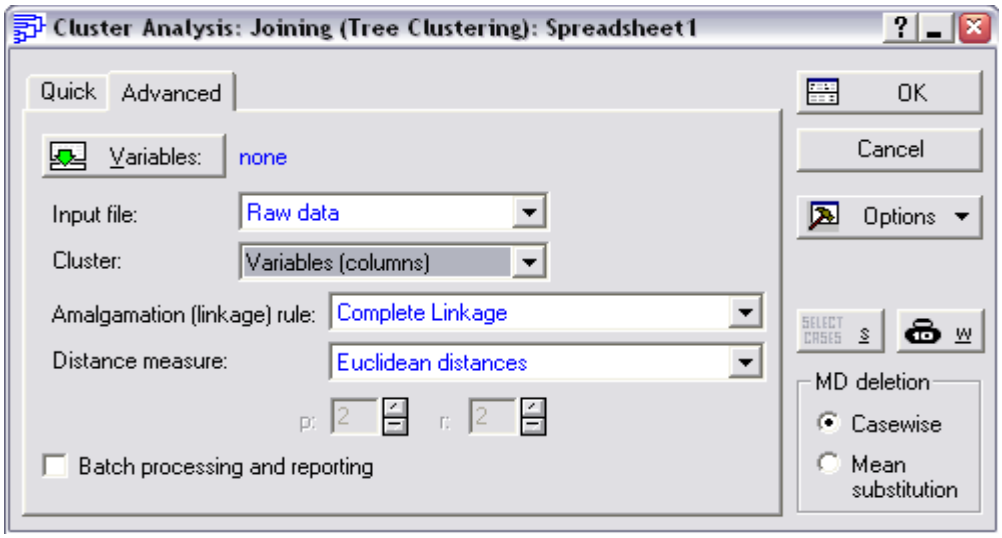

Рис.3.5. Стартова панель модуля кластерного аналізу.

Наступним кроком є вибір змінних (Select variables for the analysis).

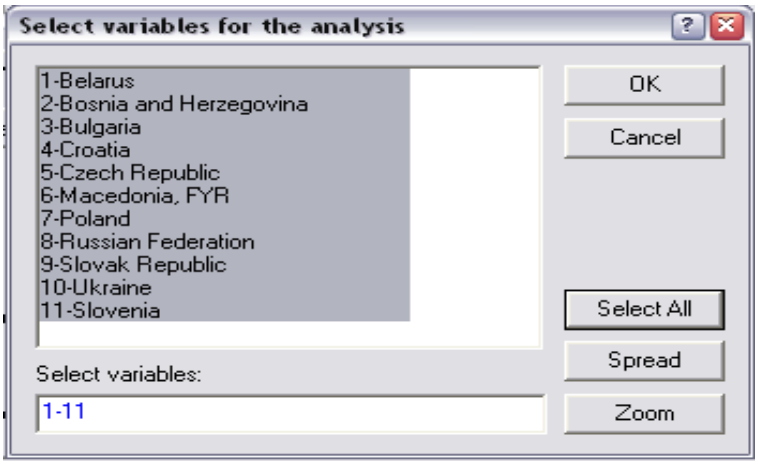

Рис. 3.6. Вікно – Вибрати змінні для кластерного аналізу.

Вікно результатів аналізу складається з двох частин: верхня частина вікна – інформаційна, нижня складається з функціональних кнопок, які дозволяють повністю подивитися на результати аналізу.

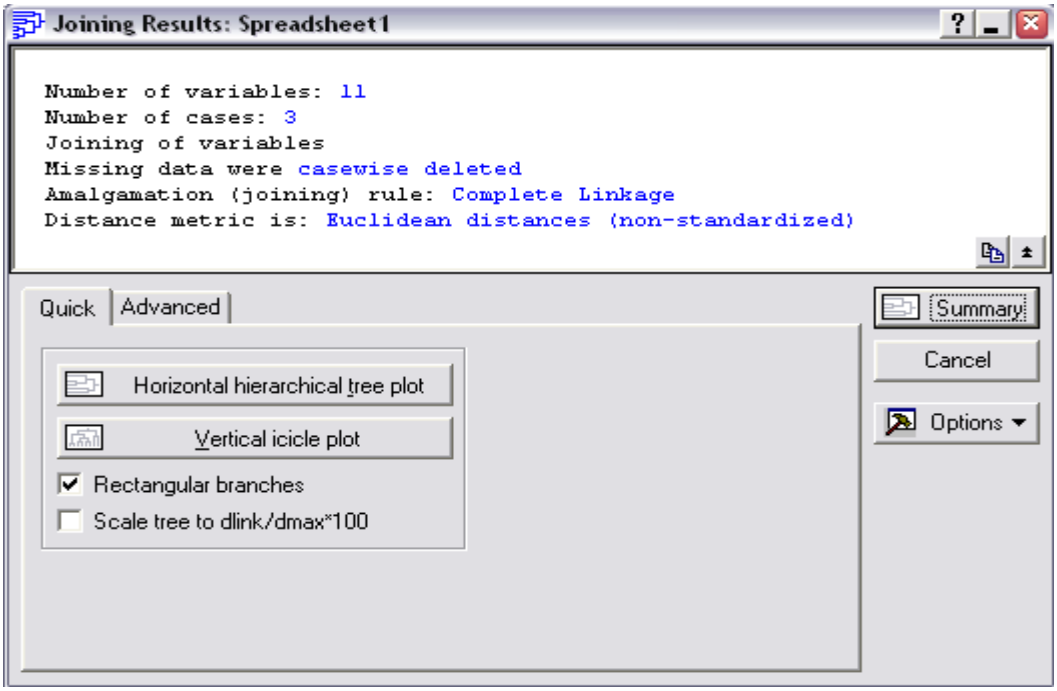

Рис. 3.7. Вікно результату кластерного аналізу.

В нижній частині виберемо кнопку Tree Diagram. Подивимося результат кластерного аналізу на графіку (дендрогамі).

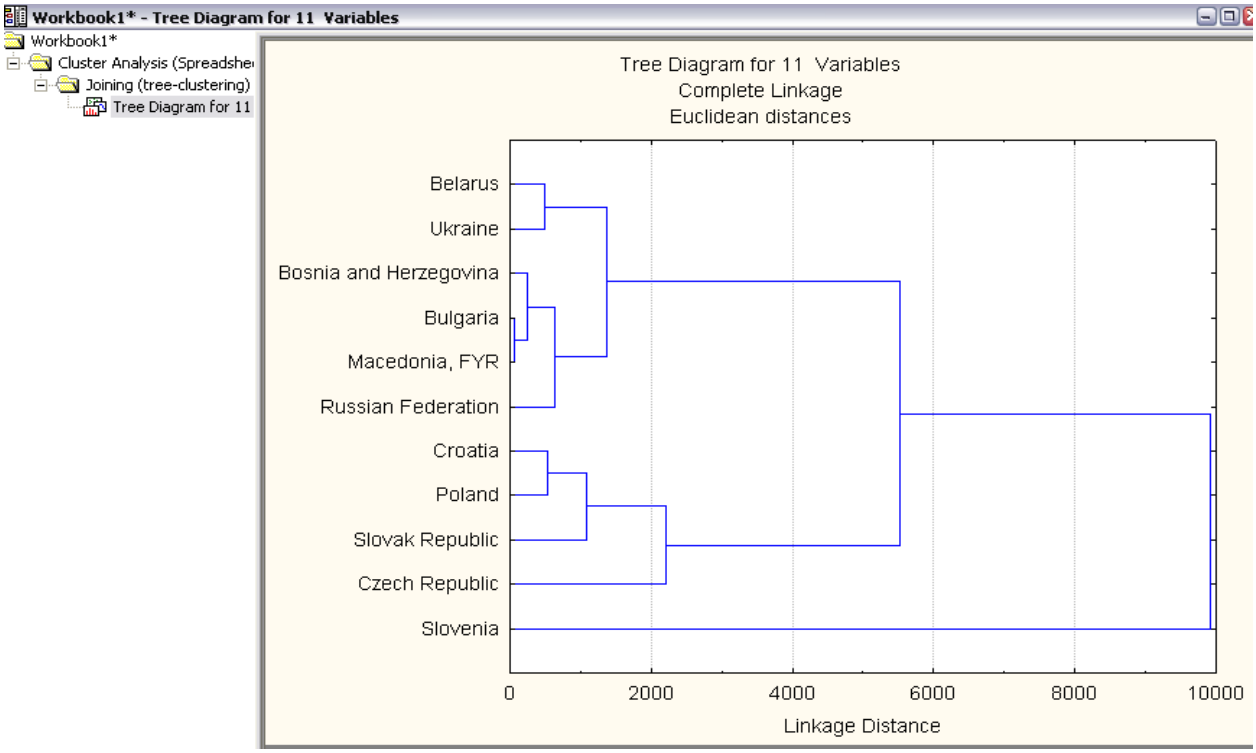

Рис. 3.8. Кластерний результат для даних за 2001 рік.

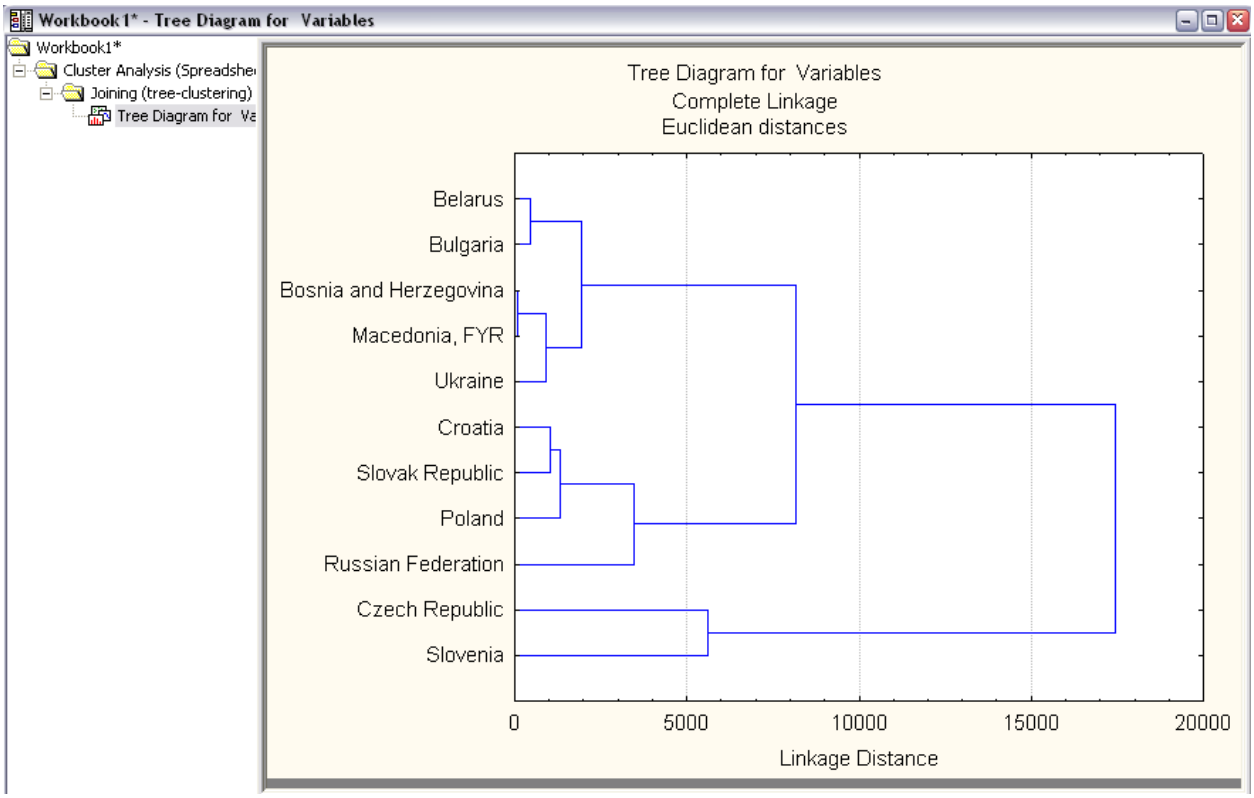

Рис. 3.9. Кластерний аналіз. Результат для даних за 2006 рік.

Проведемо групування країн на основі отриманого ієрархічного дерева. Групування здійснено за формулою Евклідової відстані, тип з'єднання – загальний.

Здійснивши таке групування, визначено наступні групи країн:

У 2001 році:

1. Болгарія, Македонія, Боснія і Герцоговина, Україна, Білорусія, Росія.

2. Хорватія, Польща, Словаччина, Чехія.

3. Словенія.

У 2006 році:

1. Македонія, Боснія і Герцоговина, Білорусія, Болгарія, Україна.

2. Хорватія, Словаччина, Польща, Росія.

3. Чехія та Словенія.

За кожен досліджуваний рік сформовано по три групи, які презентують особливості економічного зростання слов'янських країн.

Побудова кластерів на основі вибраних показників кількох років дозволяє виявити динаміку зміни в процесах економічного зростання.

При аналізі даних за 2001 рік, сформованих в ієрархічному дереві , бачимо, що у першому кластері об'єдалися Болгарія і Македонія, згодом

Боснія та Герцоговина, україна і Білорусія. Останньою приєдналась Росія і ці країни утворили найслабшу групу. Середню групу, тобто другий кластер утворили Хорватія, Польща, Словаччина і на останньому етапі до них приєдналась Чехія. Словенія в свою чергу утворила найсильніший кластер, не об'єднавшись із жодною з країн.

В результаті кластерного аналізу даних за 2006 рік на першому етапі об'єдналися Македонія та Боснія і Герцоговина, потім до них приєдналися Білорусія і Болгарія, а потім ще й Україна, на другому і третьому етапі, відповідно. Вони утворили середній клас. На кожному кроці роботи алгоритму здійснюється об'єднання двох найближчих кластерів і знову будується матриця віддалей , розмірності якої зменшуються на одну одиницю. На другому етапі об'єдналися Хорватія і Словаччина, а на третьому до них приєдналася Польща, на четвертому Росія. Вони утворили сильніший кластер. Чехія і Словенія об'єдналися останніми і утворили найсильніший клас.

Отже, порівняно з 2001 роком, більшими темпами відбувається економічне зростання у Росії, яка з найслабшої групи перейшла у другий кластер та у Чехії, яка разом з Словенією у 2006 році стали лідерами за показниками економічного зростання. Загалом відбулись незначні відмінності в перебігу процесів економічного зростання слов'янських країн, судячи з побудованого дерева ієрархій. Це означає, що для більш точних результатів потрібно зібрати більше статистичних показників, зокрема таких, що стали значущими в умовах глобалізації та НТП.

Метод найближчого сусіда

Метод найближчого сусіда не лише використовується для обчислення відстаней між кластерами, а, крім того, є самостійним методом дейтамайнінгу. Він ґрунтується на збереженні даних. В пам'яті тримається певний набір даних для порівняння з новими елементами.

Наприклад, коли з'являється новий клієнт банку, то його атрибути порівнюються з наявними банківськими клієнтами (вік, освіта, посада, місце проживання…) і виділяється множина клієнтів, найбільш подібних до нього. Цей метод відносять до методів з контрольованим навчанням, оскільки він використовується для передбачення.

Метод найближчого сусіда є стійким (робастим) відносно неякісних та відсутніх даних.

Основне завдання методу – передбачення або прогнозування показників не знаходячи при цьому залежності між показниками. Просто

вибираються історичні записи, подібні до того, який аналізується і на їх основі виконується передбачення.

Даний метод реалізовано в деяких програмних системах.

Запитання до теми

Постановка задачі кластерного аналізу. Відстань між об’єктами. Нормалізація атрибутів у кластерному аналізі. Матриця відстаней між об’єктами. Відстані між кластерами. Ієрархічний кластерний аналіз. Швидкий кластерний аналіз (метод K-центрів). Метод найближчого сусіда. Виконання кластерного аналізу засобами системи STATISTICA.

ТЕМА 4. ФАКТОРНИЙ АНАЛІЗ

Вхідною інформацією для кластерного аналізу була множина із n об'єктів, кожен із яких характеризувався m ознаками:

$$A = \begin{pmatrix} a_{11} & \dots & a_{1m} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{nm} \end{pmatrix} = \left(\overline{A}_1, \overline{A}_2, \dots, \overline{A}_m\right) \qquad (4.1)$$

де $\overline{A}_1 = \begin{pmatrix} a_{11} \\ \dots \\ a_{n1} \end{pmatrix}, \dots, \overline{A}_m = \begin{pmatrix} a_{1m} \\ \dots \\ a_{nm} \end{pmatrix}$.

У базі даних цим об'єктам відповідали записи, а ознакам – поля. Утворивши відповідну матрицю (двовимірний гіперкуб), ми у рядках маємо об'єкти (наприклад, країни), а у стовпцях – ознаки (наприклад, макроекономічні показники).

Кластерний аналіз об'єднує велику кількість об'єктів (яких може бути сотні або тисячі) у невелику кількість кластерів (не більше десяти). Факторний аналіз також здійснює стиск даних (Data reduction) початкової матриці. Він замінює набір ознак (яких може бути десятки) в невелику кількість (2, 3, щонайбільше 4) факторів.

Ознаки, які характеризують той чи інший економічний об'єкт, на практиці завжди є висококорельованими (мультиколінеарними) між собою. Наприклад, ніяк не можуть бути незалежними між собою такі ознаки як «розмір середньої зарплати», «ВВП на душу населення» та «середня тривалість життя».

Термін мультиколінеарність означає, що в регресійній моделі незалежні змінні (ознаки) пов'язані між собою лінійною залежністю. Тому застосовувати у цьому випадку класичний економетричний підхід не є коректним. При побудові економетричних моделей мультиколінеарність має бути відсутньою (коефіцієнт кореляції між довільною парою ознак має бути близьким до нуля; скалярний добуток має бути близьким до нуля).

Факторний аналіз – це обґрунтована заміна великої кількості ознак меншою кількість факторів.

У результаті виконання факторного аналізу матриця (4.1) замінюється матрицею

$$F = \begin{pmatrix} w_{11} & \dots & w_{1m} \\ \dots & \dots & \dots \\ w_{n1} & \dots & w_{nm} \end{pmatrix} = \left(\overline{F}_1, \overline{F}_2, \dots, \overline{F}_m\right), \qquad (4.2)$$

де всі пари $\langle F_k, F_l \rangle$ є ортогональними.

Тут $F_1, F_2, \ldots$ є факторами. Кожен фактор характеризує групу ознак, які мають подібний характер зміни при переході від одного об'єкта до іншого. На практиці обмежуються невеликою кількістю факторів.

Найпоширеніший метод витягу факторів – метод головних компонент. В цьому методі фактори $F_1, F_2, \ldots$ ще називаються головними компонентами (першою головною компонентою, другою головною компонентою,…). Розглянемо цей метод.

Обчислимо за даними матриці A парні кореляції r_{kl} між всіма парами $\{< k,l > | \; k,l = \overline{1,\ldots,m}\}$ ознак за формулами

$$r_{kl} = \frac{\sum (a_{ik} - \overline{a}_k)(a_{il} - \overline{a}_l)}{\sqrt{\sum (a_{ik} - \overline{a}_k)^2 * \sum (a_{il} - \overline{a}_l)^2}} \qquad r_{kl} = r_{lk} \tag{4.3}$$

У результаті отримуємо симетричну кореляційну матрицю

$$R = \begin{pmatrix} 1 & r_{12} & \ldots & r_{1m} \\ r_{12} & 1 & r_{23} & \ldots \\ \ldots & r_{23} & \ldots & r_{m-1,m} \\ r_{1m} & \ldots & r_{m-1,m} & 1 \end{pmatrix} \tag{4.4}$$

Будуємо такий визначник (детермінант) :

$$|\lambda E - R| = \begin{vmatrix} \lambda - 1 & -r_{12} & \ldots & -r_{1m} \\ -r_{12} & \lambda - 1 & \ldots & \\ \ldots & \ldots & \ldots & -r_{m-1,m} \\ -r_{1m} & \ldots & -r_{m-1,m} & \lambda - 1 \end{vmatrix} \tag{4.5}$$

Цей визначник є многочленом m-ого порядку відносно змінної λ . Він називається характеристичним многочленом матриці R .

Цьому многочленові відповідає характеристичне рівняння

$$|\lambda E - R| = 0 \; , \tag{4.6}$$

яке має не більше, ніж m дійсних коренів. Розташуємо ці корені по спаданню:

$$\lambda_1 \geq \lambda_2 \geq \ldots$$

Число λ_1 називається першим власним числом матриці A, число λ_2 - її другим власним числом,…

Кожному власному числу відповідає свій власний вектор кореляційної матриці. Перший власний вектор $\overline{U}_1$ знаходиться як ненульовий розвязок системи рівнянь

$$(\lambda_1 E - R)\overline{U}_1 = 0 \ , \qquad (4.7)$$

другий власний вектор U_2 як ненульовий розвязок системи

$$(\lambda_1 E - R)\overline{U}_2 = 0 \ ,\ldots$$

Власні числа характеризують вклади відповідних головних компонент у загальну дисперсію ознак $\overline{A}_1, \overline{A}_2, ..., \overline{A}_m$. Перша головна компонента має найбільший вплив, друга менший,...

Розглянемо таку таблицю бази даних:

Табл. 4.1

2001 рік	Експорт на душу населення, $	Інвестиції на душу населення, $	ВВП на душу населення, $
Білорусія	18,72	31,2	3090
Боснія і Герцоговина	43,89	133	2749
Болгарія	329	549	3512
Чехія	816	1133	12185
Хорватія	189	402	8751
Македонія	22,1	49	2859
Польща	100,4	271	7943
Росія	31,5	90	5340
Словаччина	301	391	8803
Словенія	175	270	17172
Україна	84,5	165,7	1828

Тут $a_{11} = 18{,}72; a_{12} = 31{,}2; a_{13} = 3090; ...; a_{11,3} = 1828$. Використавши формулу (4.3), отримуємо таку кореляційну матрицю:

$$R = \begin{pmatrix} 1 & 0{,}986275 & 0{,}493442 \\ 0{,}986275 & 1 & 0{,}486078 \\ 0{,}493442 & 0{,}486078 & 1 \end{pmatrix}$$

Зокрема, коефіцієнт кореляції між другою (інвестиції) та третьою (ВВП на душу населення) ознакою дорівнює 0,486078.

Відповідне кубічне характеристичне рівняння є таким:

$$\begin{vmatrix} \lambda - 1 & -0{,}986275 & -0{,}493442 \\ -0{,}986275 & \lambda - 1 & -0{,}486078 \\ -0{,}493442 & -0{,}486078 & \lambda - 1 \end{vmatrix} = 0$$

Знаходимо перші два власні числа як розв'язки останнього рівняння:

$$\lambda_1 = 2{,}34358; \lambda_2 = 0{,}64271$$

Для знаходження першого власного вектора $\overline{U}_1 = \begin{pmatrix} u_{11} \\ u_{21} \\ u_{31} \end{pmatrix}$ потрібно розв'язати таку систему трьох рівнянь з трьома невідомими:

$$(2{,}34358E - R)U_1 = 0 \ ,$$

тобто систему $\begin{pmatrix} 1{,}34358 & -0{,}98628 & -0{,}49344 \\ -0{,}98628 & 1{,}34358 & -0{,}48608 \\ -0{,}49344 & -0{,}48608 & 1{,}34358 \end{pmatrix} * \begin{pmatrix} u_{11} \\ u_{21} \\ u_{31} \end{pmatrix} = 0$

Визначник цієї системи згідно (4.6) дорівнює нулю, отже, вона має безліч розв'язків. Множина всіх цих розв'язків є такою:

$$\overline{U}_1 = \begin{pmatrix} 1{,}54 * z \\ 1{,}57 * z \\ z \end{pmatrix},$$

де z - довільне дійсне число. При $z = 1$ отримуємо шуканий ненульовий розв'язок, тобто, перший власний вектор кореляційної матриці:

$$\overline{U}_1 = \begin{pmatrix} u_{11} \\ u_{21} \\ u_{31} \end{pmatrix} = \begin{pmatrix} 1{,}54 \\ 1{,}57 \\ 1 \end{pmatrix}$$

Аналогічним чином, розв'язавши систему рівнянь

$$(0{,}64271E - R)\overline{U}_2 = 0 \ ,$$

отримуємо другий власний вектор $\overline{U}_2 = \begin{pmatrix} u_{12} \\ u_{22} \\ u_{33} \end{pmatrix} = \begin{pmatrix} -0{,}27 \\ -0{,}28 \\ 1 \end{pmatrix}$

Всі власні вектори є ортогональними між собою.

Третє власне число, як і третій власний вектор не обчислюємо, оскільки нашою задачею є редукція кількості ознак. Крім того, варто відкидати вектори, які відповідають невеликим власним числам (наприклад, числам, меншим від 1).

На основі отриманих власних векторів будуємо вектори (фактори, головні компоненти):

$$\overline{F_1} = \frac{\overline{U_1}}{\left|\overline{U_1}\right|} * \sqrt{\lambda_1} = \frac{\overline{U_1}}{\left|\sqrt{1{,}54^{\,2} + 1{,}57^{\,2} + 1^2}\right|} * \sqrt{2{,}34} = \begin{pmatrix} 0{,}963 \\ 0{,}965 \\ 0{,}254 \end{pmatrix}$$

$$\overline{F_2} = \frac{\overline{U_2}}{\left|\overline{U_2}\right|} * \sqrt{\lambda_2} = \frac{\overline{U_2}}{\left|\sqrt{0{,}27^{\,2} + 0{,}28^{\,2} + 1}\right|} * \sqrt{0{,}643} = \begin{pmatrix} -0{,}258 \\ -0{,}249 \\ 0{,}967 \end{pmatrix}$$

....

Набір усіх цих векторів утворює матрицю факторних навантажень

$$W = \begin{pmatrix} w_{11} & w_{12} & \ldots & w_{1m} \\ w_{21} & w_{22} & \ldots & w_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ w_{m1} & w_{m2} & & w_{mm} \end{pmatrix} = \begin{pmatrix} 0{,}963 & -0{,}258 & \ldots \\ 0{,}965 & -0{,}249 & \ldots \\ 0{,}254 & 0{,}967 & \ldots \end{pmatrix}$$

В математиці доводиться, що $\sum_i w_{ij} = \lambda_i$.У нашому прикладі

$0{,}963^2 + 0{,}965^2 + 0{,}254^2 = \lambda_1 = 2{,}34;\ (-0{,}258)^2 + (-0{,}249)^2 + 0{,}967^2 = \lambda_2 = 0{,}643$.

Елементами матриці факторних навантажень $W = \left(w_{ij}\right)$ є коефіцієнти парної кореляції, які вимірюють тісноту зв'язку між ознаками A_i та факторами F_j .

Отже, згідно методу головних компонент перший та другий фактори є такими:

$$\overline{F_1} = 0{,}963\overline{A_1} + 0{,}965\overline{A_2} + 0{,}254\overline{A_3}$$

$$\overline{F_2} = 0{,}258\overline{A_1} + 0{,}249\overline{A_2} + 0{,}967\overline{A_3} \qquad (4.8)$$

Перший фактор має високі навантаження на першу та другу ознаки (експорт та зовнішні інвестиції), а другий - на третю ознаку (ВВП на душу населення). Тому з метою редукції даних будемо вважати:

$$\overline{F_1} = 0{,}963\overline{A_1} + 0{,}965\overline{A_2}$$

$$\overline{F_2} = 0{,}967\overline{A_3}$$

або навіть:

$$\overline{F_1} = \overline{A_1} + \overline{A_2}$$

$$\overline{F_2} = \overline{A_3}$$

Перший фактор можна інтерпретувати як відкритість економіки. Він є сумою перших двох ознак. Його можна навіть змінити однією з них.

У нашому прикладі редукція трьох ознак до двох факторів була очевидною. Проте при наявності 10-20 ознак тільки математичний апарат факторного аналізу дозволяє грамотно виділити головні фактори.

Пакет STATISTICA дозволяє автоматизувати розрахунки за методом головних компонент(модуль Factor Analysis).

Після появи на екрані стартової панелі модуля Factor Analysis-факторний аналіз вказуємо вхідний файл і виконуємо перший етап факторного аналізу – обчислення матриці кореляцій. Визначимо метод виділення факторів, тобто метод головних компонент (Principal components, головні, ведучі компоненти) у вікні Define Method of Factor Extraction і натиснемо кнопку Correlations.

Data: Spreadsheet05 (11v by 11c)

	Показники конкурентоспроможності країн, 2005р.				
	1 Експорт у.о.)\душ	2 нвкстиції у.о.)/душ	3 ВВП у.о.)/душ	4 Var4	5 Var5
Білорусія	18,7201	31,20016	3090,36		
Боснія і Герцоговина	43,89357	133,0108	2749,498		
Болгарія	329,6235	549,3725	3512,619		
Чехія	816,2351	1133,66	12185,71		
Хорватія	189,1375	402,4202	8751,817		
Македонія	22,09286	49,09524	2859,249		
Польща	100,4655	271,5283	7943,339		
Росія	31,50564	90,01612	5340,562		
Словаччина	301,2205	391,1955	8803,079		
Словенія	175,5861	270,1325	17172,56		
Україна	84,53598	165,7568	1828,718		

Рис 4.1. Введення вхідних даних в систему STATISTICA

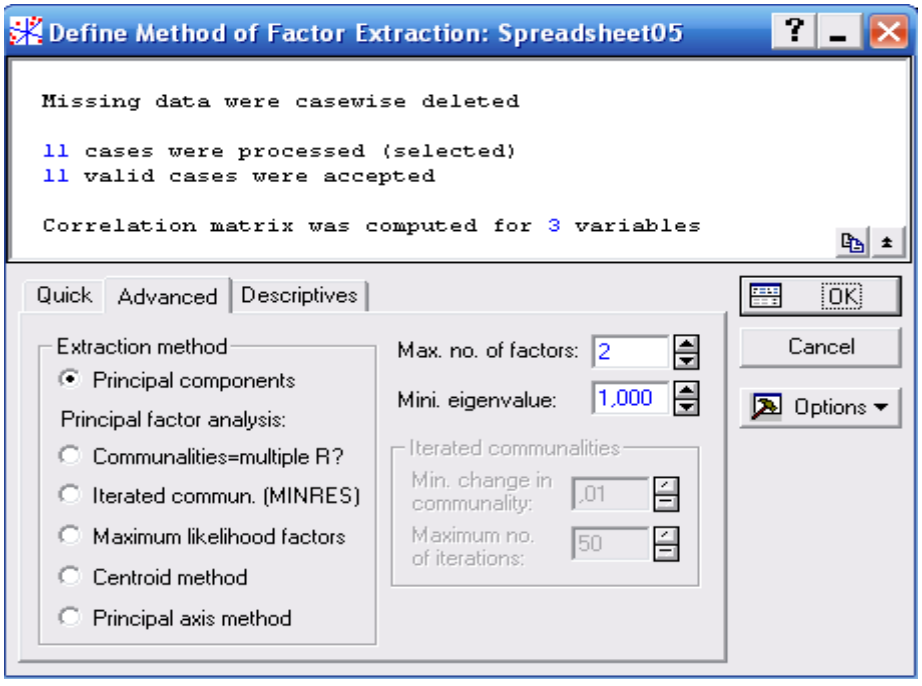

Рис. 4.2. Вигляд вікна Define Method of Factor Extraction – Визначити метод виділення факторів з вибором методу головних компонент.

На екрані буде виведена кореляційна матриця вибраних змінних.

Variable	Correlations (Spreadsheet05) Casewise deletion of MD N=11 Експорт (у.о.)\душу	Інвкстиції (у.о.)/душу	ВВП (у.о.)/душу
Експорт (у.о.)\душу	1,00	0,99	0,49
Інвкстиції (у.о.)/душу	0,99	1,00	0,49
ВВП (у.о.)/душу	0,49	0,49	1,00

Рис 4.3. Кореляційна матриця для даних з початкового файлу

Після вибору методу головних компонент і натискання кнопки OK система виведе на екран результати факторного аналізу (рис 4.4).

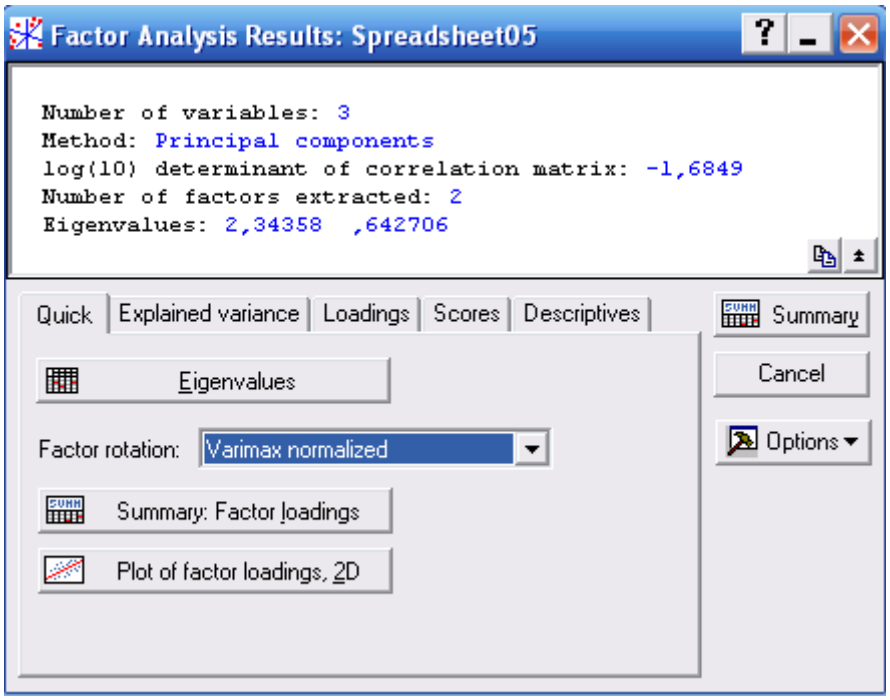

Рис. 4.4. Вікно результатів факторного аналізу

В верхній частині вікна міститься інформація:

- Number of variables – кількість змінних, що аналізуються: 3;
- Method – метод аналізу: головні компоненти;
- Number of factor extraction – кількість виділених факторів: 2 (для нашого випадку це – експорт товарів і послуг і інвестиції);
- Eigenvalues – власні значення: λ_1=2,34358; λ_2=0,642706.

Дослідимо числові факторні навантаження (рис.4.5).

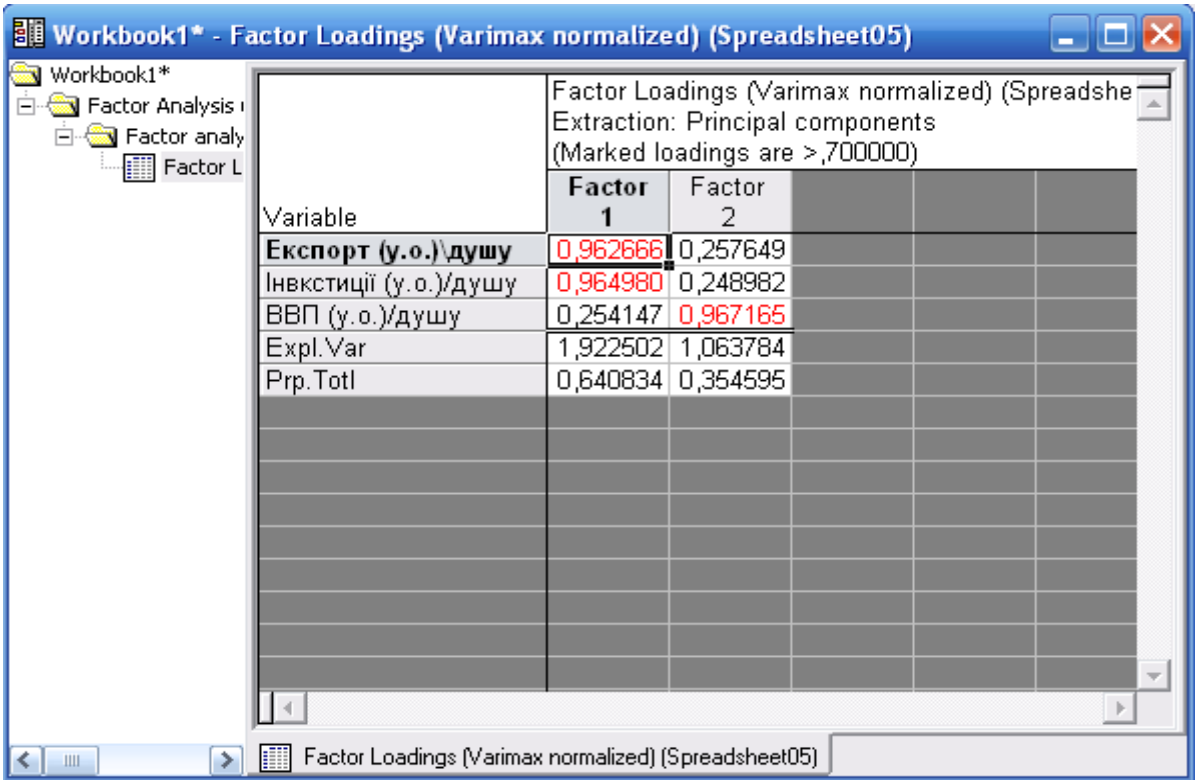

Рис 4.5. Таблиця факторних навантажень для даних з початкового файлу.

Побудуємо Двовимірний графік навантажень (Plot of Loadings 2D, Факторний результат для даних).

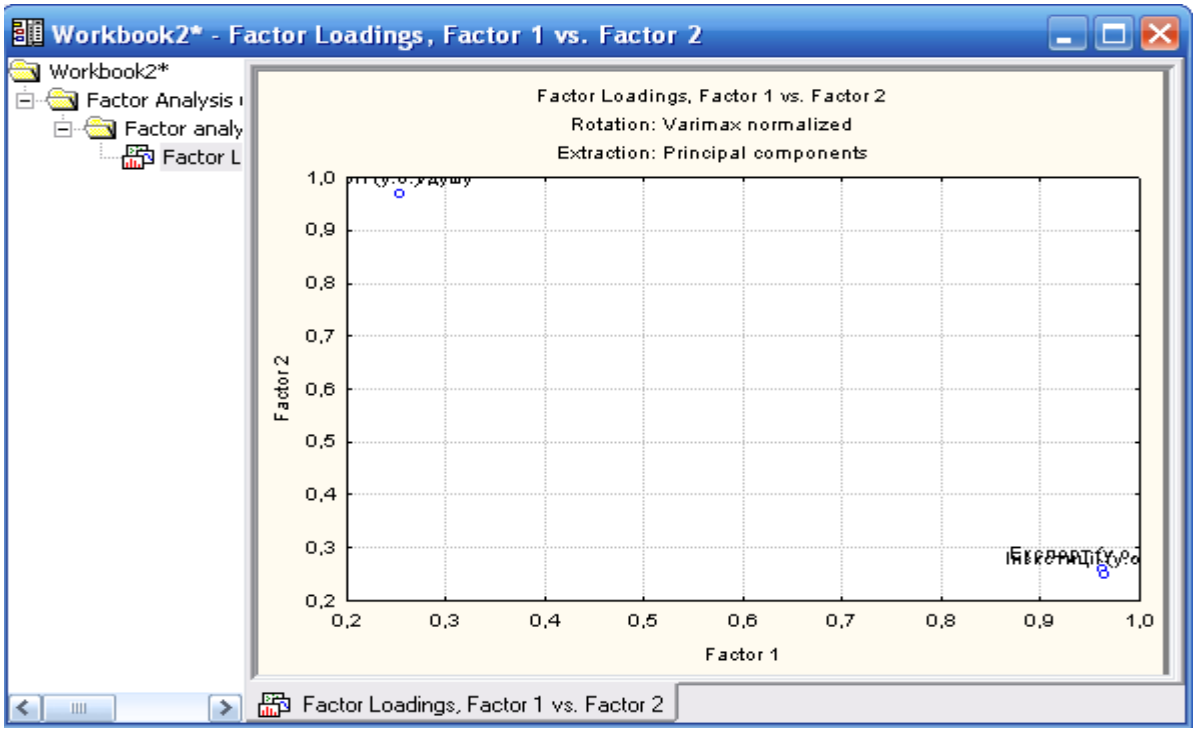

Рис.4.6.

Обертання факторів

На двовимірному графіку осі відповідають першому та другому факторам (головним компонентам). Кожна з m ознак A_i є на цьому графіку вектором, який виходить з точки (0,0) і має координати (w_{i1}, w_{i2}) . Для змістовного розуміння отриманих результатів часто потрібно виконати обертання факторів (наприклад, алгоритмом варімакс) .

На рисунку 4.7 показано результати факторного аналізу (2, с.223) п'яти ознак до і після обертання.

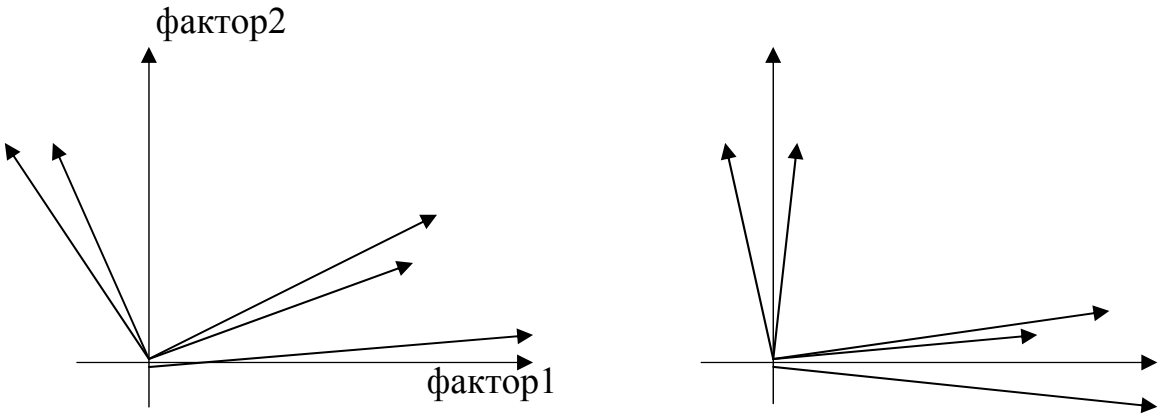

Рис.4.7

Таким чином, аналіз множини об'єктів. Кожен із яких характеризується багатьма ознаками, найкраще виконувати у такій послідовності:

- факторний аналіз з метою виявлення невеликої кількості головних факторів;
- кластерний аналіз (на базі цих факторів) з метою розбиття усієї бази даних на групи подібних між собою об'єктів;
- економетричний аналіз кожного із цих кластерів зокрема.

Запитання до теми

Постановка задачі факторного аналізу. Метод головних компонент. Характеристичний многочлен кореляційної матриці. Власні числа та власні вектори кореляційної матриці. Матриця факторних навантажень. Виконання факторного аналізу засобами системи STATISTICA. Головні компоненти, їх властивості, економічна інтерпретація. Обертання факторів.

ТЕМА 5. АНАЛІЗ ЧАСОВИХ РЯДІВ

Класична економетрія на основі статистичних даних (із бази, із сховища) будує залежність між результатною змінною (залежною змінною, результатним показником) y від змінних (ознак, аргументів, незалежних змінних) у вигляді явної математичної функції

$$y = y(x_1,...,x_n) \qquad n \geq 1 \tag{5.1}$$

В частковому випадку будується лінійна залежність

$$y = a_0 + a_1 x_1 + ... + a_n x_n \tag{5.2}$$

тобто на основі значень $y, x_1,...,x_n$ обчислюються параметри регресії $a_0, a_1,...,a_n$.

Параметрам регресії можна довіряти лише тоді, коли виконується ряд умов.

Одна з таких умов – це відсутність мультиколінеарності між ознаками $x_1,...,x_n$. На практиці ця умова не виконується практично ніколи (хоча, як було показано в попередньому розділі, попереднє застосування факторного аналізу цей недолік усуває).

Іншою обов'язковою умовою побудови регресійного рівняння є перевірка того факту, що вплив всіх неврахованих аргументів на результуючу змінну не є суттєвим. Але ж це перевірити взагалі нереально.

Тому використовувати економетричні методи в економіці слід дуже обережно.

Частковою задачею економетрії є аналіз часових (динамічних) рядів, тобто побудова теоретичного рівняння регресійної залежності

$$y = y(t) \tag{5.3}$$

за емпіричними (статистичними) даними часового ряду (незалежною змінною є час)

$$y_1,..., y_t,..., y_n \tag{5.4}$$

Основною задачею аналізу часових рядів є розклад цього ряду на

- тренд;
- сезонні коливання;
- циклічні (макроекономічні) коливання;
- залишок (випадкові ефекти).

Побудова тренду

Першою задачею аналізу часових (динамічних) рядів є дослідження тенденції, тобто еволюції, напрямку розвитку показника $y = y(t)$. Математична функція. Яка описує таку тенденцію, називається трендом.

Найчастіше тренд шукають у вигляді лінійної

$$y = a + bt$$

або експотенційної функції

$$y = ae^{bt}$$

В першому випадку параметри a та b шукаються як розв'язки такої оптимізаційної задачі:

$$f = f(a,b) = \sum_{t=1}^{n}(a + bt - y_t)^2 \to \min \quad . \tag{5.5}$$

Взявши від функції $f = f(a,b)$ часткові похідні за a та за b, отримуємо систему нормальних рівнянь

$$\begin{cases} 2\sum_{t=1}^{n}(a + bt - y_t) = 0 \\ 2\sum_{t=1}^{n} t(a + bt - y_t) = 0 \end{cases}, \tag{5.6}$$

звідки

$$\begin{cases} b = \dfrac{n\sum_{t=1}^{n} t y_t - \sum_{t=1}^{n} t \sum_{t=1}^{n} y_t}{n\sum_{t=1}^{n} t^2 - \left(\sum_{t=1}^{n} t\right)^2} \\ a = \dfrac{\sum_{t=1}^{n} y_t - b\sum_{t=1}^{n} t}{n} \end{cases} \tag{5.7}$$

Нелінійні залежності зводяться до лінійних за допомогою нескладних математичних перетворень. Так, логарифмуючи експотенційну функцію, отримуємо

$$ln(y) = ln(a) + bt \quad . \tag{5.8}$$

Задача знаходження параметрів a та b експотенційної функції тепер є лінійною відносно змінних t та $ln(y_t)$. Розрахункові формули для цих параметрів легко отримати із (5.7):

$$\begin{cases} b = \dfrac{n\sum_{t=1}^{n} t \ln y_t - \sum_{t=1}^{n} t \sum_{t=1}^{n} \ln y_t}{n\sum_{t=1}^{n} t^2 - \left(\sum_{t=1}^{n} t\right)^2} \\ \ln a = \dfrac{\sum_{t=1}^{n} \ln y_t - b\sum_{t=1}^{n} t}{n} \end{cases} \tag{5.9}$$

Перевірка наявності тренду в часовому ряді

При побудові регресійних рівнянь як між змінними x_i , так і між змінними y_i повинні бути відсутні автокореляції. Інакше кажучи, кожне спостереження не повинно залежати від інших спостережень.

Розглянемо часовий ряд, який містить набір значень деякого економічного показника:

$$y_1,...,y_t,...,y_n \qquad (5.10)$$

Перевірка наявності тренду (тобто перевірка наявності зв'язку між часом $1,\ldots,t,\ldots,n$ та значеннями $y_1,...,y_t,...,y_n$) за допомогою коефіцієнта кореляції не може вважатися науково обгрунтованою, оскільки значення часу $1,\ldots,t,\ldots,n$ аж ніяк не є незалежними між собою (після січня завжди наступить лютий, а не якийсь довільний місяць).

Для перевірки наявності тренду в часовому ряді в останні роки щораз частіше використовується критерій Фостера-Стюарда.

Згідно цього методу будуються дві допоміжні змінні v_t та l_t таким чином: якщо значення y_t за своєю величиною перевищує усі попередні значення, то приймаємо $v_t = 1$, інакше $v_t = 0$. Якщо ж значення y_t за своєю величиною є меншим від усіх попередніх значень, то $l_t = 1$, інакше $l_t = 0$.

Далі визначають величини s^{emp} та d^{emp}:

$$s^{emp} = \sum_{i=1}^{n}(v_t + l_t) \qquad (5.11)$$

$$d^{emp} = \sum_{i=1}^{n}(v_t - l_t)$$

Розподіл випадкових величин s та d збігається до нормального. Для s та d при різних значеннях n підраховані табличні (теоретичні) значення середніх та дисперсій (μ - середнє значення для s; σ_1 - дисперсія для s; 0 – середнє значення для d; σ_2 – дисперсія для d).

Для виявлення тенденції поведінки дисперсії перевіряють гіпотезу: чи можна вважати випадковою різницю $s^{emp} - \mu$. Для виявлення тенденції поведінки середнього перевіряють гіпотезу: чи можна вважати випадковою різницю $d^{emp} - 0$.

Для цього за даними часового ряду обчислюють емпіричні значення

$$T_1^{emp} = \frac{s - \mu}{\sigma_1}; \qquad T_2^{emp} = \frac{|d| - 0}{\sigma_2} \qquad (5.12)$$

та знаходять при заданому рівні довіри (значимості) α теоретичні значення критерію Стьюдента $T^{teor}(\alpha, n-1)$.

При $T_1^{emp} > T^{teor}$ та $T_2^{emp} > T_2^{teor}$ гіпотеза про існування тренду ряду $y_1,...,y_t,...,y_n$ приймається із ступенем довіри 1-α.

Сезонні коливання та перевірка їх наявності

Багато економічних показників (індекси споживчих цін, попит на сезонні товари, виробництво сільськогосподарської продукції тощо) є часовими рядами, що систематично коливаються. Найчастіше період таких коливань становить один рік.

Кожен часовий ряд, як правило, містить дві складові: тренд та коливання. Випадки, коли ряд містить лише тренд або лише коливання, в економіці зустрічаються рідко. Тому дослідження коливань економічних показників є такою ж важливою задачею, як і дослідження трендів.

На наявність сезонних коливань вказує, наприклад, візуальний аналіз графіка часового ряду. Проте найбільш обґрунтованим методом перевірки наявності сезонних коливань слід вважати прийняття рішення про існування тренду на основі аналізу корелограми цього ряду.

Розглянемо поняття автокореляцій та корелограми.

Нехай $y_1,...,y_t,...,y_n$ - деякий часовий ряд. Автокореляцією першого порядку цього ряду називають кореляцію між рядами $y_1,...,y_t,...,y_{n-1}$ та $y_2,...,y_t,...,y_n$:

$$r_1 = \frac{\sum_{t=1}^{n-1}\left[(y_t - \bar{y})(y_{t+1} - \bar{\bar{y}})\right]}{\sqrt{\sum_{t=1}^{n-1}(y_t - \bar{y})^2 \sum_{t=1}^{n-1}(y_{t+1} - \bar{\bar{y}})^2}} \quad , \tag{5.13}$$

де

$$\bar{y} = \frac{1}{n-1}\sum_{t=1}^{n-1} y_t \quad ; \quad \bar{\bar{y}} = \frac{1}{n-1}\sum_{t=1}^{n-1} y_{t+1} \tag{5.14}$$

Автокореляція другого порядку – це кореляція між рядами $y_1,...,y_t,...,y_{n-2}$ та $y_3,...,y_t,...,y_n$. В загальному випадку автокореляція *k*–ого порядку r_k у часовому ряді $y_1,...,y_t,...,y_n$ обчислюється за формулою

$$r_k = \frac{\sum_{t=1}^{n-k}\left[\left(y_t - \frac{1}{n-k}\sum_{t=1}^{n-k} y_t\right)\left(y_{t+k} - \frac{1}{n-k}\sum_{t=1}^{n-k} y_{t+k}\right)\right]}{\sqrt{\sum_{t=1}^{n-k}\left(y_t - \frac{1}{n-k}\sum_{t=1}^{n-k} y_t\right)^2 \sum_{t=1}^{n-k}\left(y_{t+k} - \frac{1}{n-k}\sum_{t=1}^{n-k} y_{t+k}\right)^2}} \tag{5.15}$$

Очевидно, що всі значення r_k знаходяться в проміжку $0 < r_k < 1$.

Графік, на якому по горизонтальній осі відкладені значення k , а по вертикальній осі – значення r_k , називається корелограмою. Аналіз корелограми дає змогу встановити наскільки далеко значення y_t впливають на значення y_{t+k} . Наявність великих значень автокореляції r_{12} свідчить про наявність у часовому ряді сезонних коливань з періодом у 12 місяців.

Побудова регресійного рівняння для сезонних коливань

Розглянемо ряд, у якому є сезонні коливання з періодом T=12 (за кількістю місяців у році):

$$y'_1,...,y'_{12},...,y'_{12\cdot n} \quad : \tag{5.16}$$

Нехай цей ряд не містить трендової компоненти і нехай, крім того, середнє значення його елементів дорівнює нулю. Це виконується за допомогою операції елімінації тренду.

Тоді регресійну залежність, яка теоретично описує поведінку ряду (4.16), можна шукати у вигляді синусоїди з періодом T=12:

$$y(t) = a\sin(\frac{\pi}{6}t) + b\cos(\frac{\pi}{6}t) \tag{5.17}$$

На практиці ряд (5.4) перетворюють у ряд, що містить 12 елементів:

$$y_1,...,y_t,...,y_{12} \quad , \tag{5.18}$$

де

$$y_t = \frac{1}{n}\sum_{i=0}^{n-1} y'_{t+12\cdot i} \quad (t=1,2,\ldots,12) \tag{5.19}$$

Для знаходження параметрів a та b формули (4.17) на основі даних ряду (4.18) розв'язуємо оптимізаційну задачу

$$\sum_{t=1}^{12}(a\sin(\frac{\pi}{6}t) + b\cos(\frac{\pi}{6}t) - y_t)^2 \to \min \ . \tag{5.20}$$

Ця задача зводиться до знаходження розв'язків системи двох лінійних рівнянь

$$\begin{cases} a\sum_{t=1}^{12}(\sin\frac{\pi}{6}t)^2 + b\sum_{t=1}^{12}(\sin\frac{\pi}{6}t)(\cos\frac{\pi}{6}t) = \sum_{t=1}^{12} y_t(\sin\frac{\pi}{6}t) \\ a\sum_{t=1}^{12}(\sin\frac{\pi}{6}t)(\cos\frac{\pi}{6}t) + b\sum_{t=1}^{12}(\cos\frac{\pi}{6}t)^2 = \sum_{t=1}^{12} y_t(\cos\frac{\pi}{6}t) \end{cases} \tag{5.21}$$

звідки отримуються шукані параметри регресії:

$$\begin{aligned} a &= \frac{1}{6}\sum_{t=1}^{12} y_t(\cos\frac{\pi}{6}t) \\ b &= \frac{1}{6}\sum_{t=1}^{12} y_t(\sin\frac{\pi}{6}t) \end{aligned} \tag{5.22}$$

Побудова регресійного рівняння для циклічних коливань

В економіці крім сезонних коливань спостерігаються також так звані довгі (циклічні) коливання. Усі країни з ринковою економікою циклічно проходять як фази підйому, так і фази спаду. Тривалість повного макроекономічного циклу не є однаковою в різні періоди розвитку людства і не є однаковою для різних країн. Дослідження циклічних (макроекономічних, довгих) коливань мають дуже важливе значення, тому що як фіскальна, так і монетарна політика держави повинна бути іншою напередодні зростання і іншою напередодні економічного спаду.

Нехай задано часовий ряд

$$y_1,...,y_t,...,y_n \ , \tag{5.23}$$

з періодом коливань T ($T << n$).

Нехай цей ряд не містить тренду. Нехай також середнє значення елементів ряду (5.23) дорівнює нулю. Розглянемо задачу побудови синусоїди з періодом коливань, який дорівнює величині T:

$$y(t) = a\sin(\frac{2\pi}{T}t)+b\cos(\frac{2\pi}{T}t) \ , \tag{5.24}$$

Ця синусоїда повинна бути близькою до значень ряду (4.23) згідно критерію методу найменших квадратів:

$$f(a,b) = \sum_{t=1}^{n}\left(y_t - a\sin(\frac{2\pi}{T}t)-b\cos(\frac{2\pi}{T}t)\right)^2 \to \min \tag{5.25}$$

Взявши частинні похідні від функції *f(a,b)*, маємо:

$$\begin{cases} \sum_{t=1}^{n}(y_t - a\sin(\frac{2\pi}{T}t)-b\cos(\frac{2\pi}{T}t))\sin(\frac{2\pi}{T}t)=0 \\ \sum_{t=1}^{n}(y_t - a\sin(\frac{2\pi}{T}t)-b\cos(\frac{2\pi}{T}t))\cos(\frac{2\pi}{T}t)=0 \end{cases} \tag{5.26}$$

З цієї системи знаходимо параметри a та b:

$$\begin{aligned} a &= \frac{\sum_{t=1}^{n}\cos^2(\frac{2\pi}{T}t)\cdot\sum_{t=1}^{n}y_t\sin(\frac{2\pi}{T}t)-\sum_{t=1}^{n}y_t\cos(\frac{2\pi}{T}t)\sum_{t=1}^{n}\sin(\frac{2\pi}{T}t)\cos(\frac{2\pi}{T}t)}{\sum_{t=1}^{n}\cos^2(\frac{2\pi}{T}t)\cdot\sum_{t=1}^{n}\sin^2(\frac{2\pi}{T}t)-[\sum_{t=1}^{n}\sin(\frac{2\pi}{T}t)\cos(\frac{2\pi}{T}t)]^2} \\ b &= \frac{\sum_{t=1}^{n}\sin^2(\frac{2\pi}{T}t)\cdot\sum_{t=1}^{n}y_t\cos(\frac{2\pi}{T}t)-\sum_{t=1}^{n}y_t\sin(\frac{2\pi}{T}t)\sum_{t=1}^{n}\sin(\frac{2\pi}{T}t)\cos(\frac{2\pi}{T}t)}{\sum_{t=1}^{n}\cos^2(\frac{2\pi}{T}t)\cdot\sum_{t=1}^{n}\sin^2(\frac{2\pi}{T}t)-[\sum_{t=1}^{n}\sin(\frac{2\pi}{T}t)\cos(\frac{2\pi}{T}t)]^2} \end{aligned} \tag{5.27}$$

Отриману формулу $y(t) = a\sin(\frac{2\pi}{T}t) + b\cos(\frac{2\pi}{T}t)$ можна представити також у вигляді $y(t) = A\sin(\frac{2\pi}{T}t + \varphi)$; $A = \sqrt{a^2 + b^2}$; $\varphi = arctg\frac{a}{b}$. (5.28)

Останній запис дозволяє наглядно оцінити як амплітуду коливань, так і розташування піків сезонної компоненти часового ряду.

Перевірка наявності циклічних коливань

Наявність циклічних коливань та визначення періоду цих коливань теорія часових рядів пропонує виконувати при допомозі спектрального аналізу.

Основою спектрального аналізу є теорема Фур'є, яка стверджує, що довільна неперервна (за винятком, можливо, скінченої кількості точок) функція *y=f(x)* на інтервалі [-π; π] може бути розкладена в ряд

$$f(x) = \sum_{j=1}^{\infty} a_j \sin(jx) + \frac{1}{2}b_0 + \sum_{j=1}^{\infty} b_j \cos(jx) \quad , \qquad (5.28)$$

де
$$a_j = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(jx)dx$$
$$b_j = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(jx)dx$$

Зазначимо, що за допомогою заміни змінних розклад функцій у ряд Фур'є можливий і на довільному іншому відрізку. Якщо при цьому середнє значення функції *y=f(x)* на досліджуваному відрізку дорівнює нулеві, то середній доданок у формулі (4.29) пропадає. Функція *y=f(t)* , отже, представляє собою суму гармонійних компонент (гармонік):

$$a_1\sin(\alpha_1 t) + b_1\cos(\alpha_1 t)$$
. . . . . .
$$a_j\sin(\alpha_j t) + b_j\cos(\alpha_j t)$$
. . . . . .

Амплітуда кожної *j*-ої гармоніки становить $c_j = \sqrt{a_j^2 + b_j^2}$, а частота цієї гармоніки дорівнює α_j , що відповідає періоду коливань $T_j = \frac{2\pi}{\alpha_j}$.

Отже, задача наближення дискретних значень періодичного часового ряду $y_1,...,y_t,...,y_n$ (5.29)
неперервними гармонійними функціями розпочинається із виявлення декількох (часто лише однієї чи двох) гармонік

$$a\sin(\alpha t) + b\cos(\alpha t) \qquad (5.30)$$

з періодом коливань (частотою), близьким до періоду коливань ряду (5.30).

Тісноту зв'язку між рядом (5.29) та функцією (5.30) задає величина

$$I(\alpha)=\frac{1}{\pi n}\left(\sum_{t=1}^{n} y_t \cdot \cos \alpha t\right)^2+\frac{1}{\pi n}\left(\sum_{t=1}^{n} y_t \cdot \sin \alpha t\right)^2 \qquad (5.31)$$

Ця величина в спектральному аналізі називається інтенсивністю зв'язку між часовим рядом та гармонікою з частотою α. Графік, де по осі абсцис відкладені значення частот, а по осі ординат значення інтенсивностей, називають спектром.

Методика побудови спектру часового ряду полягає в знаходженні значень інтенсивності $I(\alpha)$ при різних значеннях α . Визначається те значення частоти α, яке забезпечує найбільше значення інтенсивності. Відповідне цій частоті значення періоду $T=\frac{2\pi}{\alpha}$ приймається за період циклічних (довгих) коливань. Встановивши період T, регресійне рівняння для циклічної складової часового ряду будуємо за формулами (5.27), (5.28).

В найбільш загальному випадку дослідження часових рядів полягає у їхньому розкладі на три компоненти: тренд, циклічну складову, сезонну складову.

Типова схема аналізу часових рядів представлена на рис. 5.1.

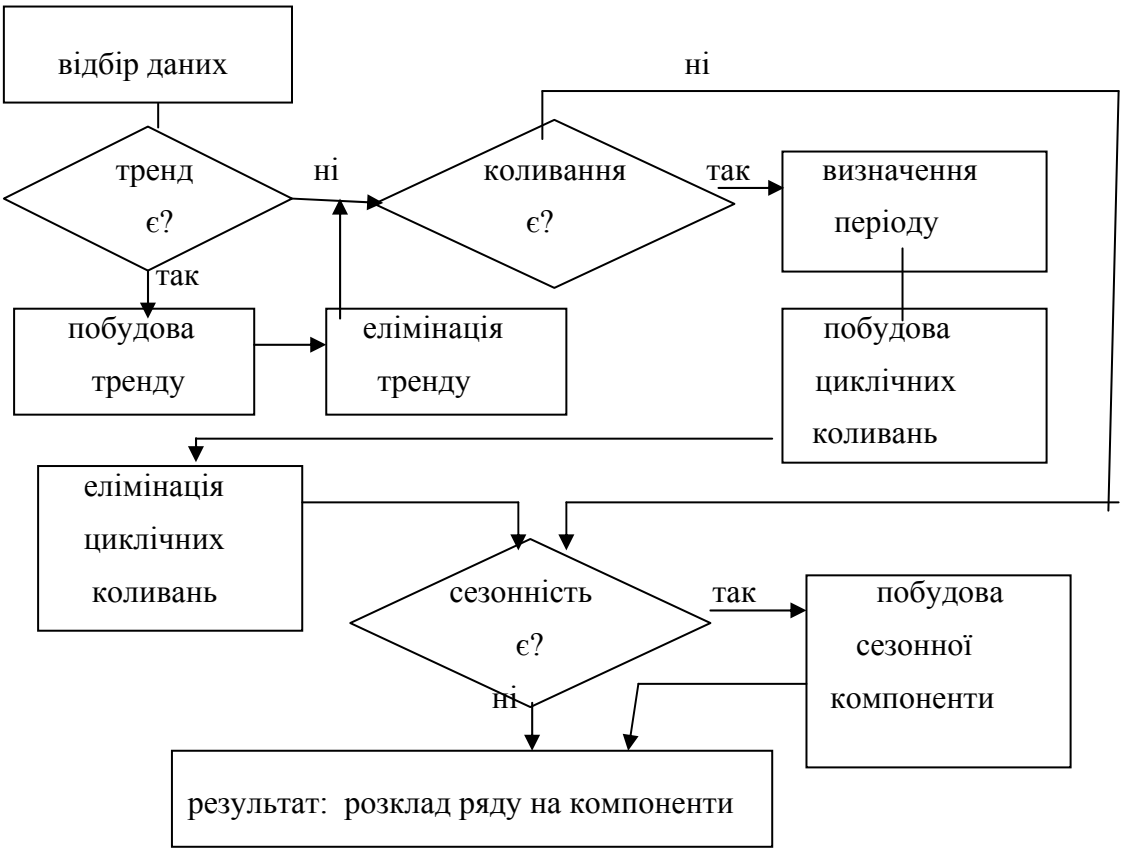

Рис. 5.1.

Розглянемо часовий ряд y_{wage} офіційної середньомісячної зарплатні в Україні за 1995-2004 роки (табл. 5.1). Дані в купоно-карбованцях за період до вересня 1996 року приведені до гривень.

Табл. 5.1

Середньомісячна офіційна заробітна платня в Україні (грн.)

	1995р.	1996р.	1997р.	1998р.	1999р.	2000р.	2001р.	2002р.	2003р.	2004р.
1	32,08	103,28	126,68	136,82	148,16	180,97	253,39	320,76	400,59	444,7
2	41,53	108,95	126,36	137,85	152,03	190,62	263,66	328,7	391,2	461,5
3	50,23	117,55	135,15	149,76	166,61	210,67	281,03	354,81	415,49	489,17
4	53,67	116,28	133,14	146,39	165,53	205,35	288,93	355,78	422,58	494,66
5	60,74	119,44	139,53	148,61	168,87	213,21	302,96	358,88	439,26	498,48
6	71,39	125,47	144,3	158,01	180,76	228,78	317,81	377,41	476,16	550,81
7	75,84	130,46	151,44	159,21	183,27	238,49	327,31	398,1	476	530,82
8	81,39	129,99	147,67	153,21	180,94	247,44	329,33	390,07	489	520,59
9	88,73	132,7	150,34	156,4	186,44	249,04	326,34	391,14	479	556,61
10	95,18	134,87	149,65	156,07	186,85	254,11	335,75	397,49	489	561,87
11	101,22	132,27	147,07	155,54	190,39	257,58	334,44	395,7	498	569,35
12	124,48	151,51	165,81	176,09	218,88	296,26	378,45	442,91	551	607,11

Процес аналізу цих даних виконувався згідно схеми, представленої на рис. 5.1. Методом найменших квадратів (критерій Фостера-Стюарда дає 90% впевненість в існуванні тренду) отримано таку трендову складову:

$$y_{wage}^{trend,1} = 12{,}179 + 4{,}122t \qquad (5.32)$$

Спектр ряду середньої зарплатні, із якого еліміновано тренд, наведено на рис. 5.2.

На основі аналізу цього спектру визначено період циклічних коливань. Найбільшу інтенсивність забезпечує частота α=0,062. Цій частоті відповідає період $T = \frac{2\pi}{\alpha} = 101$ (місяців).

Рис. 5.2

Отримане регресійне рівняння циклічних (довгих) коливань значень середньомісячної зарплатні має вигляд

$$y_{wage}^{cycle} = 31{,}42 \sin\left(\frac{2\pi}{101} t\right) + 36{,}54 \cos\left(\frac{2\pi}{101} t\right) \qquad (5.33)$$

Амплітуда коливань становить $\sqrt{31{,}42^2 + 36{,}54^2} = 48{,}19$.

Корелограма ряду $y_{wage} - y_{wage}^{trend} - y_{wage}^{cycle}$ середньої зарплатні після елімінації тренду та довгих коливань наведена на рис. 5.3.

Рис. 5.3

Високі значення 12-ої та 24-ої автокореляцій та від'ємні і близькі до нуля значення інших автокореляцій вказують на наявність сезонних коливань з періодом 12 місяців.

Сезонна складова, що була отримана, має вигляд:

$$y_{wage}^{seazon} = -9{,}414 \sin\left(\frac{\pi}{6} t\right) - 0{,}150 \cos\left(\frac{\pi}{6} t\right) \qquad (5.34)$$

Отже, остаточне регресійне рівняння часового ряду значень середньої заробітної платні в Україні за 1995 – 2004 роки y_{wage} має вигляд

$$y_{wage}^{teoret} = 12{,}18 + 4{,}122t + 31{,}4 \sin\left(\frac{2\pi}{101} t\right) + 36{,}54 \cos\left(\frac{2\pi}{101} t\right) - 9{,}41 \sin\left(\frac{\pi}{6} t\right) - 0{,}15 \cos\left(\frac{\pi}{6} t\right) \quad (4.36)$$

Графіки емпіричних (статистичних) та теоретичних значень середньої зарплати представлено на рис. 5.4.

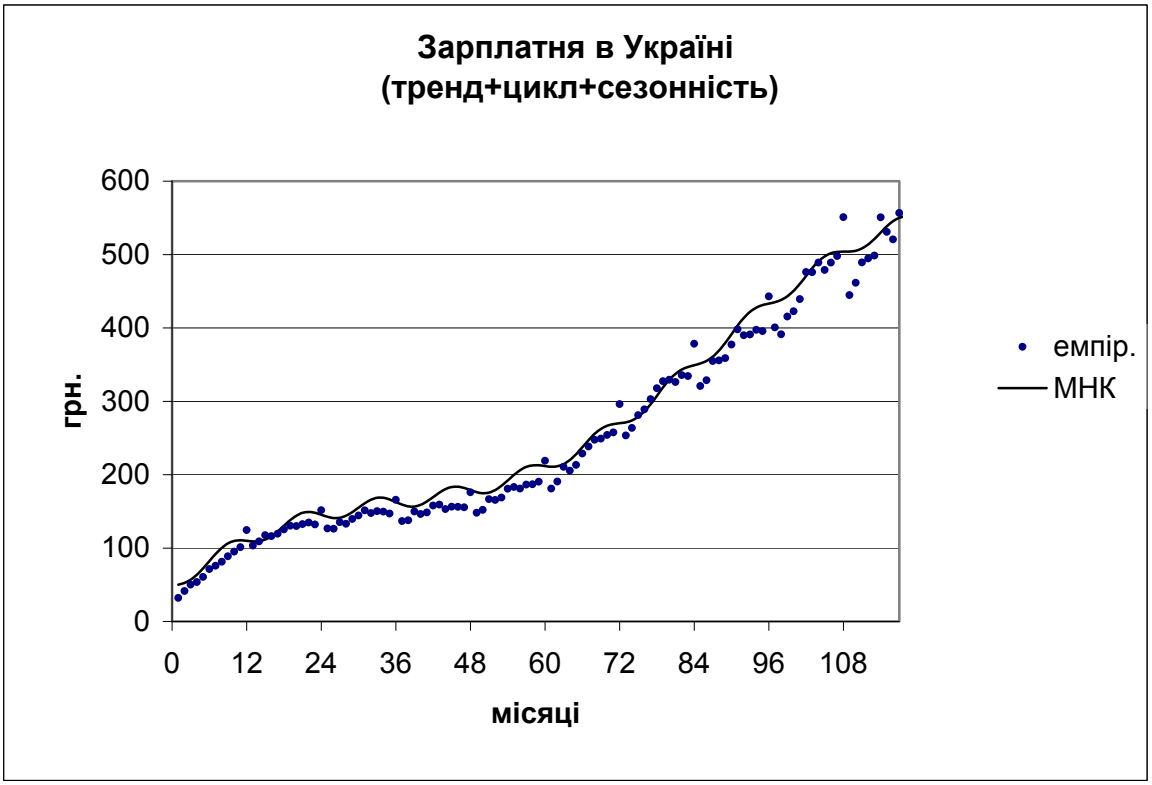

Рис. 5.4

Звичайно, не в кожному ряді усі ці три компоненти є присутніми. Проте сума цих складових має бути близькою до початкового часового ряду (тобто, випадкова компонента не повинна бути великою).

Зазначимо, що аналіз часових рядів методом Фостера-Стюарда та за допомогою корелограми і спектру почали застосовувати в економічних дослідженнях відносно недавно. У більшості програмних систем (зокрема, в системі EXCEL, а також в пакетах Statistica та Statgraphics) ці методи ще не реалізовані.

ТЕМА 6. ШКАЛИ ТА КОРЕЛЯЦІЙНИЙ АНАЛІЗ

Одна із проблем застосування статистичних методів в економіці полягає в тому, що дослідник декількома натисканнями клавіш в тому чи іншому пакеті отримує великий набір числових показників. Дуже часто отримані результати (незважаючи на найточніші математичні формули) використовуються некваліфіковано, оскільки кожен математичний метод і кожну математичну формулу можна використовувати тільки за певних умов. Особливо гостро це питання стоїть у тому випадку, коли якісні ознаки приводяться до кількісних.

Основні типи шкал

Кількісні методи можуть бути застосовані в економічному аналізі тільки після того, як емпіричні дані будуть переведені на мову чисел. Процедура, внаслідок якої виникає числова модель властивостей об'єкта, називається виміром. При вимірі встановлюється відповідність між властивостями об'єкта і властивостями відповідних їм чисел.

Набір властивостей об'єкта і чисел, що відповідають їм, називається шкалою.

Номінальні шкали.

При побудові такої шкали виконується розбиття множини всіх об'єктів на класи, які не перетинаються. Кожен такий клас є окремим пунктом шкали.

Номінальні шкали

При побудові такої шкали множину всіх об'єктів розподіляють на класи, що не перетинаються. Кожен такий клас є окремим пунктом шкали.

Приклад

Картка абітурієнта містить такі ознаки як місце проживання, стать, національність,..

Якісну ознаку «національність» можна зробити кількісною, наприклад, таким чином:

1 –українець

2 -росіянин

3 -поляк

. . . . .,

утворивши шкалу {1,2,3,…}.

Проте у випадку номінальної шкали над елементами множини {1,2,3,…} не можна виконувати ні арифметичних дій, ні навіть порівняння

чисел на > та <. У тому числі не має сенсу і середнє арифметичне. Єдина можлива дія – це порівняння на = та на ≠ . В номінальних шкалах можна також підраховувати кількість об'єктів у кожному класі, їх частоту (процентне співвідношення), знаходити моду (об'єкт, який зустрічається найчастіше).

Іноді розбиття всіх об'єктів на класи не є очевидною задачею і вимагає застосування кластерного аналізу.

Найважливіше те, що у випадку номінальних шкал можна застосовувати лише ті статистичні формули, побудова (виведення. доведення) яких не ґрунтувалося на арифметичних діях над елементами шкали. Зокрема, у випадку номінальних шкал не можна використовувати формулу парної кореляції, оскільки вона включає середнє арифметичне.

Кажуть, що номінальна шкала забезпечує найслабший тип виміру.

Порядкові шкали

Порядкові шкали – це такі шкали, в яких крім відношення (відносин) рівності між об'єктами, встановлено ще й відношення послідовності.

Приклад.

Магістерські роботи випускників економічного факультету розбиваються на три класи згідно педагогічної системи оцінок {3,4,5} та на пять класів згідно шкали {A,B,C,D,E}. Критерії оцінки магістерських такі: «5» може отримати тільки магістрант, який мав хоча б одну публікацію в студентсько-аспірантському збірнику; при відсутності програмної реалізації магістерська робота оцінюється не більше, ніж на «3»;…

Як в першій, так і в другій множині вже можна виконувати операції порівняння на > та < , проте арифметичні дії тут також є недопустимими. Справді, ніяк не можна стверджувати, що магістерська робота на «5» перевищує роботу на «4» на стільки само, як і робота на «4» перевищує роботу на «3». Відстані між елементами таких шкал, отже, встановлювати не можна.

Інтервальні шкали

В інтервальних шкалах є змога визначати віддалі між парами об'єктів. Класичним прикладом інтервальної шкали є термометр (справді, -5 це на 10 градусів холодніше, ніж +5; так само як +20 на 10 градусів холодніше, ніж +30). В інтервальних шкалах допускаються операції додавання та віднімання, проте ні множення, ні ділення (тобто відношення) виконувати не можна. Справді, стверджувати, що 10^0 - це в 5 разів тепліше, ніж 2^0 , а -10^0 - в в 2 рази тепліше, ніж -20^0 не має сенсу.

Класичною (псевдо)інтервальною шкалою є шкала ІСС (індекс соціального самопочуття), який ще називають «термометром соціальної думки».

Шкали відносин (відношень).

В цих шкалах можна виконувати всі попередні дії, а також дії множення і ділення, тобто знаходження відношень.

Найтиповішими прикладами таких шкал є зарплата, дохід тощо. Справді, зарплата у 12000₴ є в три рази більшою та на 8000₴ більшою, ніж зарплата в 3000₴.

Кореляційні та причинно-наслідкові зв'язки

Економіка досліджує, в основному, кореляційні зв»язки між об'єктами, ознаками, показниками. При кореляційних зв'язках при зміні однієї величини змінюється розподіл (зокрема, середнє) іншої.

Проте дуже важливим є на практиці дослідити, чи знайдений кореляційний зв'язок між показником x та показником y є причинно-наслідковим. Справді, дуже важливо вияснити, чи бідність є причиною (ми не хочемо жити за законами Європи, тому й бідні), чи наслідком (підвищувати вартість проїзду в метро хоча б до рівня Росії не можна, бо ми бідні) нерозумних дій.

Побудуємо часові ряди обох ознак: $x = x(t)$ та $y = y(t)$.

Розв'яжемо таку економетричну задачу (тест Гренджера): знайти коефіцієнти α , β , γ та δ , щоб виконувалася така регресійна залежність:

$$\begin{cases} y(t) = \alpha x(t-1) + \beta y(t-1) \\ x(t) = \gamma x(t-1) + \delta y(t-1) \end{cases} \qquad (6.1)$$

При $\alpha >> 0$ та $\beta \approx 0; \delta \approx 0$ робимо висновок про те, що зміна величини x є причиною для зміни величини y . При $\delta >> 0$ та $\alpha \approx 0; \gamma \approx 0$ процес $y(t)$ є причиною процесу $x(t)$.

За дослідження в галузі моделювання причинно-наслідкових зв'язків в економіці Гренджер (Granger) в 2004 році отримав нобелевську премію з економіки.

Зазначимо, проте, що тест Гренджера є вірним тільки тоді, коли всі невраховані в тесті ознаки не справляють суттєвого впливу як на x , так і на y.

Коефіцієнти зв'язку між ознаками у випадку номінальних шкал

Критерій χ^2

Розглянемо набір об'єктів, які характеризуються декількома ознаками. Вони представлені у файлі (таблиці, базі даних):

Табл.6.1.

Обєкти	Ознаки			
	x	. . .	y	. . .
1				
2				
. . .				

Нехай ознака x може приймати чотири значення $\{x_1, x_2, x_3, x_4\}$, а ознака y три значення $\{y_1, y_2, y_3\}$.

На основі цього файлу маємо таку матрицю (двовимірний гіперкуб), яка задає двовимірний розподіл:

Табл. 6.2

Значення ознаки x	Значення ознаки y			
	y_1	y_2	y_3	
x_1	N_{11}	N_{12}	N_{13}	$Nx_1 = \sum N_{1j}$
x_2	N_{21}	N_{22}	N_{23}	Nx_2
x_3	N_{31}	N_{32}	N_{33}	Nx_3
x_4	N_{41}	N_{42}	N_{43}	Nx_4
	$Ny_1 = \sum N_{i1}$	Ny_2	Ny_3	$N = \sum$

Тут N_{ij} - кількість об'єктів, у яких ознака x приймає значення x_i , а ознака y - значення y_j .

При відсутності зв'язку для всіх i та j мала б місце теоретична частота:

$$N_{ij}^{(0)} = 1/N \, N(x_i) \cdot N(y_j)$$

Мірою відхилення емпіричних частот N_{ij} від теоретичних є величина

$$\chi^2_{emp} = \sum_i \sum_j \frac{\left(N_{ij} - N_{ij}^{(0)}\right)}{N_{ij}^{(0)}} \qquad (6.2)$$

Ця міра називається критерієм Пірсона або критерієм χ^2 .

Крім знайденого емпіричного значення χ^2_{emp} за таблицями або за допомогою функції системи EXCEL знаходимо теоретичне значення $\chi^2_{teor}(f,p)$, де f - кількість ступенів вільності (у нашому прикладі $f=6$, оскільки в таблиці незалежними є тільки 6 значень), а p - ступінь довіри. При $\chi^2_{teor}(6;0,95)<\chi^2_{emp}$ з ймовірністю 95% можна стверджувати, що зв'язок між ознаками x та y існує.

Як видно із виведення формули для χ^2., при її виведенні використовувалися лише числа N_{ij}, які відповідають кількостям об'єктів, що мають ознаки x_i та y_j . Дії над самими величинами x_i та y_j не виконувалися.

Критерії, пов'язані з критерієм χ^2

Критерій Пірсона пропонує розглядати коефіцієнт середньої квадратичної спряженості

$$c=\sqrt{\frac{\varphi^2}{1+\varphi^2}}\ , \qquad (6.3)$$

де $\varphi^2=\frac{\chi^2}{N}$.

При відсутності зв'язку коефіцієнт $c=0$.

Часто розглядають коефіцієнт

$$c'=\frac{c}{c_{\max}} \qquad (0\le c'\le 1)$$

Розглядають також коефіцієнт Чупрова

$$T=\sqrt{\frac{\chi^2}{N\sqrt{k-1}\sqrt{l-1}}}\ , \qquad (6.4)$$

де k та l - кількість рядків та стовпців матриці. При $k=l$ маємо $0\le T\le 1$.

Для видовжених таблиць використовують коефіцієнт Крамера

$$T_c=T\sqrt[4]{\frac{\min(k-1,l-1)}{\max(k-1,l-1)}}\ . \qquad (6.5)$$

Дослідження зв'язку між ознаками у випадку номінальних шкал.

У номінальних шкалах ознаки можна порівнювати на > та < . Крім моди, має сенс і поняття медіани. Об'єкти можна впорядковувати за ознаками по зростанню або спаданню. Тим самим кожен об'єкт отримує певний ранг.

Нехай ми маємо N об'єктів. Нехай x та y дві ознаки цих об/єктів, для кожної із яких мають місце відношення «більше-менше».

Утворимо всі можливі пари об'єктів. Цих пар буде $\frac{1}{2}N(N-1)$.Знайдемо ранги кожного об'єкта як за ознакою x , так і за ознакою y.

Нехай ранги об'єкта k відповідно дорівнюють r_x^k та r_y^k .

Нехай також ранги об'єкта l відповідно дорівнюють r_x^l та r_y^l .

Коефіцієнт рангової кореляції Кендала.

Нехай тепер P - кількість таких пар $\langle k,l\rangle$, де ранги пар розташовані в однаковій послідовності (тобто або $r_x^k > r_x^l$ і $r_y^k > r_y^l$,

або $r_x^k < r_x^l$ і $r_y^k < r_y^l$).

Через Q позначимо відповідно кількість таких пар $\langle k,l\rangle$, де ранги розташовані в різних послідовностях ($P+Q=\frac{1}{2}N(N-1)$).

Коефіцієнт Кендела (Kendall)– це відношення

$$\tau = \frac{P-Q}{\frac{1}{2}N(N-1)} \qquad (o \le \tau \le 1) \qquad (6.7)$$

Зазначимо, що при виведенні коефіцієнту τ ми використали тільки ранжування об'єктів на "<" та ">" , не виконуючи арифметичних дій над самими ознаками.

Коефіцієнти рангової кореляції Спірмена

Виконавши те саме ранжування, обчислюємо величину

$$\rho = \sum_{i=1}^{N} d_i^2 = \sum_{i=1}^{N} \left(r_x^i - r_y^i\right)^2 \qquad (-1 \le \rho \le 1) \qquad (6.8)$$

яка називається коефіцієнтом Спірмена.

Коефіцієнти Кандела та Спірмена розраховують, наскільки ко-релюють між собою ранги ознак.

Дослідження зв'язку між ознаками у випадку інтервальних шкал.

Над ознаками в інтервальних шкалах можна виконувати операції додавання та віднімання. Отже, у цьому випадку має сенс знаходити середнє арифметичне, математичне сподівання, середньоквадратичне відхилення, дисперсію, коефіцієнти варіації кожної з ознак.

Коефіцієнт парної кореляції.

У нашому випадку при дослідженні зв'язку між двома ознаками можна всі пари x та y нанести на графік.

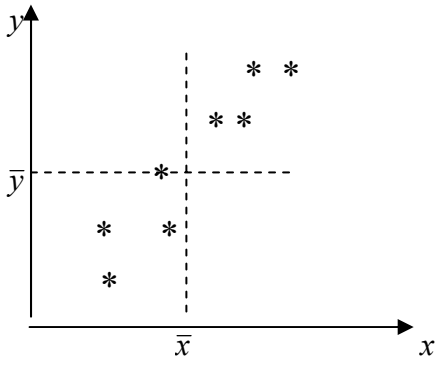

Рис.6.1.

Дослідимо набір значень $x_i - \bar{x}$ та $y_i - \bar{y}$. Якщо значенням однієї ознаки, меншим від її середнього, відповідає значення другої ознаки, теж, менше від середнього, то це свідчить про наявність зв'язку між ознаками. Мірою такого зв'язку може бути кореляція

$$S = \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$

Якщо зв'язку немає, то додатні і відємні складові врівноважаться і сума S буде близькою до нуля.

Для того, щоб міра тісноти зв'язку двох ознак знаходилася в інтервалі [-1;1], потрібно поділити значення S на $S_{\max}$. Отримуємо коефіцієнт парної кореляції:

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2} * \sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N\sigma_x\sigma_y}, \qquad (6.9)$$

де $\sigma_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N}}$ та $\sigma_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{N}}$ - середньоквадратичні відхилення ознак від їхніх середніх значень.

Дослідження зв'язку між ознаками у випадку шкал відносин.

Шкали відносин забезпечують найсильніший тип виміру. В цих шкалах можна використовувати усі дії, що й у попередніх шкалах. Можна використовувати всі розгляне ні раніше міри центральної тенденції (моду, медіану, середнє арифметичне,…). Крім того, за рахунок операції множення має сенс і середнє геометричне.

Для оцінювання тісноти зв'язку між різними ознаками можна використовувати і χ^2-критерій, і коефіцієнти Кендела та Спірмена, і коефіцієнт парної кореляції (табл.6.3).

Таблиця 6.3.

Тип шкали	Статистичні міри	
	Міри центральної тенденції	Міри тісноти звязку
Номінальна	Відсотки, мода	χ^2-критерій Пірсона
Порядкова	Відсотки, мода, медіана	χ^2-критерій Пірсона, коефіцієнт Кандела, коефіцієнт Спірмена
Інтервальна	Відсотки, мода, медіана, середнє арифметичне, дисперсія, коефіцієнт варіації	χ^2-критерій Пірсона, коефіцієнт Кендела, коефіцієнт Спірмена, коефіцієнт парної кореляції
Відносин	Відсотки, мода, медіана, середнє арифметичне, дисперсія, коефіцієнт варіації, середнє геометричне	χ^2-критерій Пірсона, коефіцієнт Кендела, коефіцієнт Спірмена, коефіцієнт парної кореляції

Зазначимо, що при аналізі зв'язку між двома ознаками, які належать до різних шкал, слід використовувати властивості тільки шкали нижчого рівня. Наприклад, при дослідженні зв'язку між національностями (номінальна шкала) та розміром зарплати (шкала відносин) можна використовувати лише χ^2-критерій Пірсона.

Аналіз одновимірних розподілів в пакеті Statistica виконується в пункті меню Analyze. Descriptive statistics. Freguence. Statistics.

Для аналізу двовимірних розподілів виконуємо: Analyze. Descriptive statistics. Grosstabs. Row (Columns). Statistics.

У вікні Statistics маємо змогу вибрати потрібний критерій:

Chi square (χ^2-критерій Пірсона)

Correlation (тут, зокрема, є критерій Спірмена)

Kendal (коефіцієнт Кендела).

ДОДАТОК А

Таблиця офіційної україно-англійської транслітерації
(прийнята 19 квітня 1996 року
Українською Офіційною Термінологічною Комісією,
www.rada.kiev.ua/translit.htm)

Українська літера	Англійська літера	Особливості	Приклади
А	A		Алушта - Alushta
Б	B		Борщагівка - Borschahivka
В	V		Вишгород - Vyshhorod
Г	H, gh	gh – при сполученні ”зг”	Гадяч - Hadiach; Згорани - Zghorany
Ґ	G		Ґалаґан - Galagan
Д	D		Дон - Don
Е	E		Рівне - Rivne
Є	Ye, ie	Ye – на початку слів	Єнакієве - Yenakiieve; Наєнко - Naienko
Ж	Zh		Житомир - Zhytomyr
З	Z		Закарпаття - Zakarpattia
И	Y		Медвин - Medvyn
І	I		Іржава - Irshava
Ї	I	Yi - на початку слів	Їжакевич - Yizhakevych; Кадіївка - Kadiivka
Й	Y, i	Y - на початку слів	Йосипівна - Yosypivka; Стрий - Stryi
К	K		Київ - Kyiv
Л	L		Лебедин - Lebedyn
М	M		Миколаїв - Mykolaiv
Н	N		Ніжин - Nizhyn
О	O		Одеса - Odesa
П	P		Полтава - Poltava
Р	R		Ромни - Romny
С	S		Суми - Sumy
Т	T		Тетерів -Teteriv

Закінчення додатку А

У	U		Ужгород - Uzhhorod
Ф	F		Фастів - Fastiv
Х	Kh		Харків - Kharkiv
Ц	Ts		Біла Церква – Bila Tserkva
Ч	Ch		Чернівці - Chernivtsi
Ш	Sh		Шостка - Shostka
Щ	Sch		Гоща - Hoscha
Ь	‘		Русь – Rus’; Львів – L’viv
Ю	Yu, iu	Yu - на початку слів	Юрій - Yurii; Крюківка - Krukivka
Я	Ya, ia	Ya - на початку слів	Яготин - Yahotyn; Ічня - Ichnia

ДОДАТОК Б. Таблиця офіційної україно-англійської транслітерації

ЗАТВЕРДЖЕНО

постановою Кабінету Міністрів України
від 27 січня 2010 р. N 55

Джерело: http://zakon.rada.gov.ua/cgi-bin/laws/main.cgi?nreg=55-2010-%EF

. 1. Буквосполучення "зг" відтворюється латиницею як "zgh" (наприклад, Згорани - Zghorany, Розгон - Rozghon) на відміну від "zh" - відповідника української літери ж".

2. М'який знак і апостроф латиницею не відтворюються.

3. Транслітерація прізвищ та імен осіб і географічних назв здійснюється шляхом відтворення кожної літери латиницею.

Таблиця транслітерації українського алфавіту латиницею

Український алфавіт	Латиниця	Позиція у слові	Приклади написання	
			українською мовою	латиницею
Аа	Aa		Алушта Андрій	Alushta Andrii
Бб	Bb		Борщагівка Борисенко	Borshchahivka Borysenko
Вв	Vv		Вінниця Володимир	Vinnytsia Volodymyr
Гг	Hh		Гадяч Богдан Згурський	Hadiach Bohdan Zghurskyi
Ґґ	Gg		Ґалаґан Ґорґани	Galagan Gorgany
Дд	Dd		Донецьк Дмитро	Donetsk Dmytro
Ее	Ee		Рівне Олег Есмань	Rivne Oleh Esman
Єє	Ye ie	на початку слова в інших позиціях	Єнакієве Гаєвич Коропʼє	Yenakiieve Haievych Koropie
Жж	Zh zh		Житомир Жанна Жежелів	Zhytomyr Zhanna Zhezheliv
Зз	Zz		Закарпаття Казимирчук	Zakarpattia Kazymyrchuk
Ии	Yy		Медвин Михайленко	Medvyn Mykhailenko
Іі	Ii		Іванків Іващенко	Ivankiv Ivashchenko
Її	Yi i	на початку слова в інших позиціях	Їжакевич Кадиївка Марʼїне	Yizhakevych Kadyivka Marine
Йй	Y i	на початку слова в інших позиціях	Йосипівка Стрий Олексій	Yosypivka Stryi Oleksii
Кк	Kk		Київ Коваленко	Kyiv Kovalenko
Лл	Ll		Лебедин Леонід	Lebedyn Leonid
Мм	Mm		Миколаїв Маринич	Mykolaiv Marynych
Нн	Nn		Ніжин Наталія	Nizhyn Nataliia
Оо	Oo		Одеса Онищенко	Odesa Onyshchenko
Пп	Pp		Полтава Петро	Poltava Petro

http://zakon.rada.gov.ua/cgi-bin/laws/main.cgi?nreg=55-2010-%EF

Український алфавіт	Латиниця	Позиція у слові	Приклади написання	
			українською мовою	латиницею
Рр	Rr		Решетилівка Рибчинський	Reshetylivka Rybchynskyi
Сс	Ss		Суми Соломія	Sumy Solomiia
Тт	Tt		Тернопіль Троць	Ternopil Trots
Уу	Uu		Ужгород Уляна	Uzhhorod Uliana
Фф	Ff		Фастів Філіпчук	Fastiv Filipchuk
Хх	Kh kh		Харків Христина	Kharkiv Khrystyna
Цц	Ts ts		Біла Церква Стеценко	Bila Tserkva Stetsenko
Чч	Ch ch		Чернівці Шевченко	Chernivtsi Shevchenko
Шш	Sh sh		Шостка Кишеньки	Shostka Kyshenky
Щщ	Shch shch		Щербухи Гоща Гаращенко	Shcherbukhy Hoshcha Harashchenko
Юю	Yu iu	на початку слова в інших позиціях	Юрій Корюківка	Yurii Koriukivka
Яя	Ya ia	на початку слова в інших позиціях	Яготин Ярошенко Костянтин Знам'янка Феодосія	Yahotyn Yaroshenko Kostiantyn Znamianka Feodosiia