УДК 330.47

О. Л. Тоцька – старший викладач кафедри фінансів підприємств і кредиту Волинського державного університету імені Лесі Українки

Автоматизація ієрархічного методу кластерного аналізу за допомогою програмного пакета StatSoft Statistica 6.0

Кластерний аналіз є сукупністю методів класифікації багатовимірних спостережень, які базуються на визначенні поняття віддалі між досліджуваними об'єктами з наступним виділенням в них подібних груп. Він використовується для розділення сукупності вхідних даних на однорідні групи так, щоб об'єкти всередині групи були подібними між собою згідно з деяким критерієм, а об'єкти із різних груп відрізнялись один від одного [1, 129]. Для автоматизації кластерного аналізу використовуються різноманітні пакети прикладних програм, зокрема StatSoft Statistica 6.0.

Методи кластерного аналізу поділяються на три групи: ієрархічні, варіаційні та методи пошуку "згустків" об'єктів. Ієрархічна класифікація є найбільш популярною, оскільки їй притаманні математична прозорість алгоритму, наочність подання даних тощо [2, 190].

Для автоматизації ієрархічної процедури кластерного аналізу за допомогою пакета StatSoft Statistica 6.0 потрібно виконати наступні дії:

1) завантажити програму

Пуск → Програми → STATISTICA 6.0 → STATISTICA;

2) створити файл з даними

File (Файл) → New (Новий) → у рядку "Number of variables" ввести кількість змінних → у рядку "Number of cases" ввести кількість спостережень (випадків) → OK → заповнити створений файл даними;

3) провести z-перетворення, тобто стандартизацію даних (вона необхідна для усунення різниці в одиницях виміру змінних і розраховується за формулою:

стандартизоване значення $z_i = (x_i - \bar{x})/s$, де середнє $\bar{x} = \left(\sum_{i=1}^{n} x_i\right)/n$, стандартне відхилення $s = \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2/n}$)

виділити всі комірки з введеними даними → Data (Дані) → Standardize (Стандартизувати) → OK;

4) вибрати метод кластерного аналізу

виділити всі комірки зі стандартизованими даними → Statistics (Статистичні обчислення) → Multivariate Exploratory Techniques (Багатовимірні дослідницькі методи) → Cluster Analysis (Кластерний аналіз) → вибрати один з трьох методів кластерного аналізу: Joining (tree clustering) – ієрархічний, K-means clustering – К-середніх, Two-way joining – двовходове об'єднання (у нашому випадку – Joining) → OK;

5) вказати початкові параметри

на вкладці "Advanced" заповнити вказані нижче рядки і натиснути кнопку OK:

у рядку "Input file" (Вхідний файл) вибрати одне з двох значень:

- Raw data – неопрацьовані дані,
- Distance matrix – матриця відстані;

у рядку "Cluster" (Кластер) вказати, що буде класифікуватись:

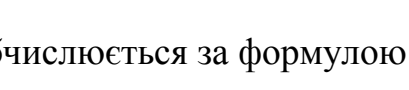
- Variables (columns) – змінні (колонки),
- Cases (rows) – спостереження (рядки);

у рядку "Amalgamation (linkage) rule" (Правило об'єднання) вибрати один з семи алгоритмів:

- Single linkage – одинарне об'єднання (метод ближнього сусіда),
- Complete linkage – повне об'єднання (метод найбільш віддаленого сусіда),
- Unweighted pair-group average – незважене попарне групове середнє,
- Weighted pair-group average – зважене попарне групове середнє,

- Unweighted pair-group centroid – незважений попарний груповий центроїд,
- Weighted pair-group centroid (median) – зважений попарний груповий центроїд (медіана),
- Ward's method – метод Варда (Уорда);

у рядку "Distance measure" (Міра відстані) вказати одну з семи відстаней:

- Squared Euclidean distances – квадрат Евклідових відстаней, який обчислюється за формулою  ,
- Euclidean distances – Евклідові відстані, які розраховуються за формулою  ,
- City-block (Manhattan) distances – відстань міських кварталів (Манхеттенська), яка обчислюється за формулою  ,
- Chebychev distance metric – показник відстані Чебишева, який розраховується за формулою  ,
- Power – степенева відстань, яка обчислюється за формулою  ,
- Percent disagreement – процент незгоди, який розраховується за формулою  ,
- 1-Pearson – коефіцієнт кореляції Пірсона, який обчислюється за формулою  ,

де x_i, y_i – значення двох змінних,

$\bar{x}, \bar{y}$ – їх середні значення,

s_x, s_y – їх стандартні відхилення,

n – кількість пар значень;

у блоці “MD (missing data) deletion” (Вилучення відсутніх даних) поставити перемикач на одній з двох опцій:

- Casewise – мудрий реєстр (видаляє рядки чи стовпці з відсутніми даними),
- Mean substitution – заміна середнім;

у рядку “Batch processing and reporting” (Обробка пакету даних і повідомлення) поставити при потребі прапорець;

6) ознайомитись з результатами

Horizontal hierarchical tree plot – горизонтальний ієрархічний деревовидний графік,

Vertical icicle plot – вертикальний бурульковидний графік,

Rectangular branches – прямокутні гілки,

Scale tree to dlink/dmax*100 – дерево розмірності (у процентному співвідношенні),

Amalgamation schedule – список об’єднання,

Graph of amalgamation schedule – графік списку об’єднання,

Distance matrix – матриця відстані,

Descriptive statistics – описова статистика (середнє і стандартне відхилення),

Matrix – матриця;

7) зберегти результати

File (Файл) → Save (Зберегти) → у рядку “Ім’я файла” ввести назву → Зберегти.

Література

1. Єлейко В. І. Основи економетрії.– Львів: Марка Лтд, 1995.– 192 с.
2. Столяров Г. С., Ємшанов Д. Г., Ковтун Н. В. АРМ статистика: Навч. посібник.– К.: КНЕУ, 1999.– 268 с.