



ZHLUKOVÁ ANALÝZA V SOFTVÉRI STATISTICA

CLUSTER ANALYSIS IN THE SOFTWARE STATISTICA

Jana HALČINOVÁ, Peter TREBUŇA

Abstract

The article deals with making the cluster analysis in statistical software STATISTICA. For example, a simple procedure is described for the analysis of data. There are two forms of cluster analysis, hierarchical and non-hierarchical methods. Graphical output is a dendrogram of hierarchical methods, which has a horizontal or vertical form.

Key words

Cluster analysis, hierarchical cluster analysis, dendrograph, STATISTICA.

Úvod

Zhluková analýza je súhrnné označenie pre skupinu metód, ktorých cieľom je buď zoskupiť dané objekty do zhlukov alebo vytvoriť hierarchiu zhlukov objektov. Hierarchické metódy zhlukovej analýzy zaraďujú analyzované objekty do hierarchického systému zhlukov. Pri nehierarchických metódach zhlukovej analýzy ide o zaradenie objektov do vopred stanoveného počtu disjunktných zhlukov.

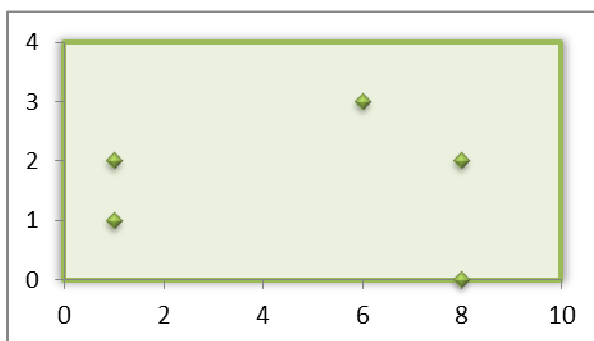
Vstupné údaje zhlukovej analýzy

Vstupom pre zhlukovú analýzu je vstupná matica objektov a výstupom je konkrétna identifikácia zhlukov. Majme vstupnú dátovú maticu o rozmeroch 2x5, na ktorej bude ilustrovaný postup hierarchických metód zhlukovania v štatistickom softvéri STATISTICA.

Matica má tvar

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 6 & 3 \\ 8 & 2 \\ 8 & 0 \end{pmatrix}$$

Rozloženie objektov zhlukovej analýzy v priestore podľa dátovej matice je graficky znázornené na obr.1.



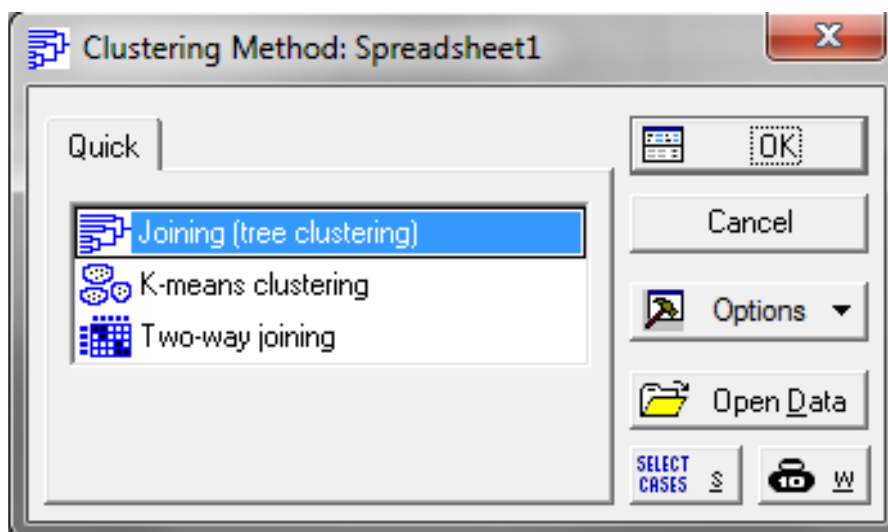
Obr. 1 Znázornenie objektov zhlukovej analýzy



Zhluková analýza v softvéri STATISTICA

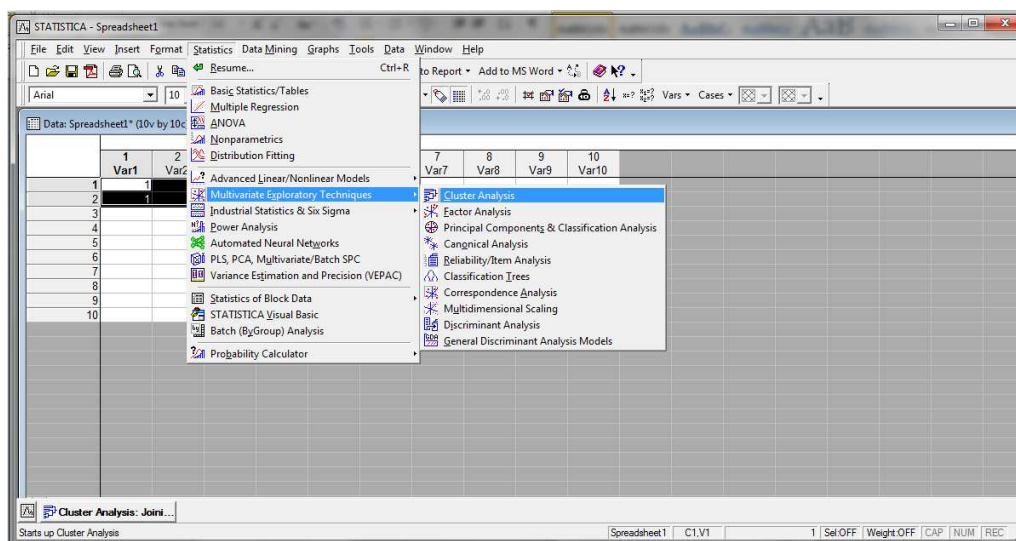
STATISTICA je špičkový systém, ktorý slúži pre podporu strategického a výkonného riadenia financií, vzťahov so zákazníkmi, kvality, rizík, pri výskume, vo výučbe a vývoji. Využíva k tomu rozsiahlu ponuku analytických metód, grafických výstupov, nástrojov pre automatizáciu, spoluprácu, reportovanie, správu dát.

Jednou z analytických metód softvéru STATISTICA je zhluková analýza, ktorá je obsahom tohto článku. Program poskytuje tieto hlavné kategórie zhlukovej analýzy znázornené na obr.2.



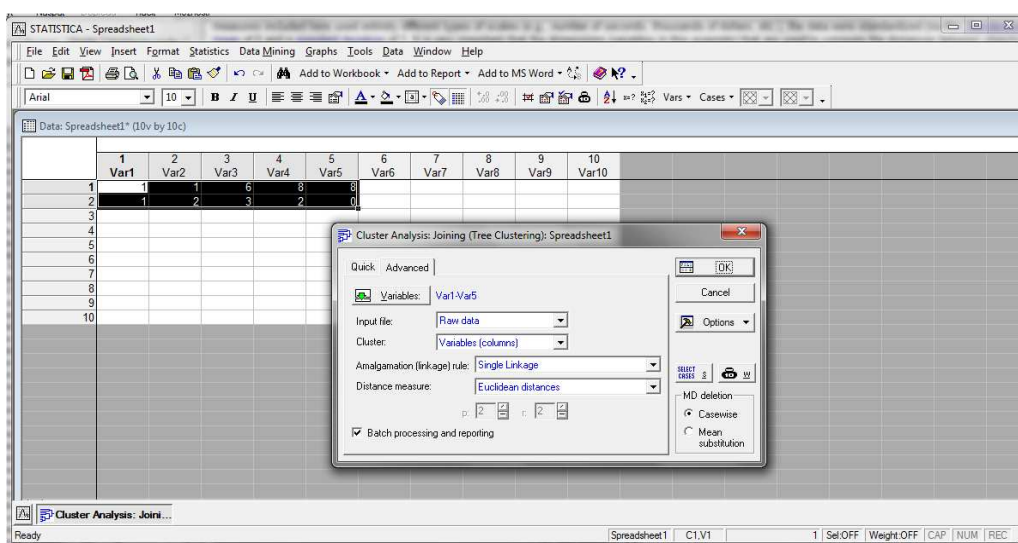
Obr. 2 Hlavné kategórie zhlukovej analýzy

Pre rozsiahlosť problematiky bude v predkladanom článku pozornosť upriamená na „Strom zhlukovej analýzy“ tzn. dendrogram hierarchických metód zhlukovej analýzy. Spôsob výberu záložky „Cluster Analysis“ z množstva nástrojov softvéru STATISTICA je na obr.3.



Obr. 3 Spôsob voľby „Cluster Analysis“

Pre uskutočnenie zhlukovania je potrebné špecifikovať podmienky analýzy znázornené na obr.4.



Obr. 4 Špecifikácia podmienok zhlukovej analýzy

STATISTICA poskytuje tieto možnosti výberu:

Vstupné údaje (*Input file*):

- *Raw data*
- *Distance matrix*

Zhluk (*Cluster*):

- *Variables (columns)*
- *Cases (rows)*

Metóda zhlukovej analýzy (*Amalgamation (linkage) rule*):

- *Single Linkage*
- *Complete Linkage*
- *Unweighted pair-group average*
- *Weighted pair-group average*
- *Unweighted pair-group centroid*
- *Weighted pair-group centroid (median)*
- *Ward's method*

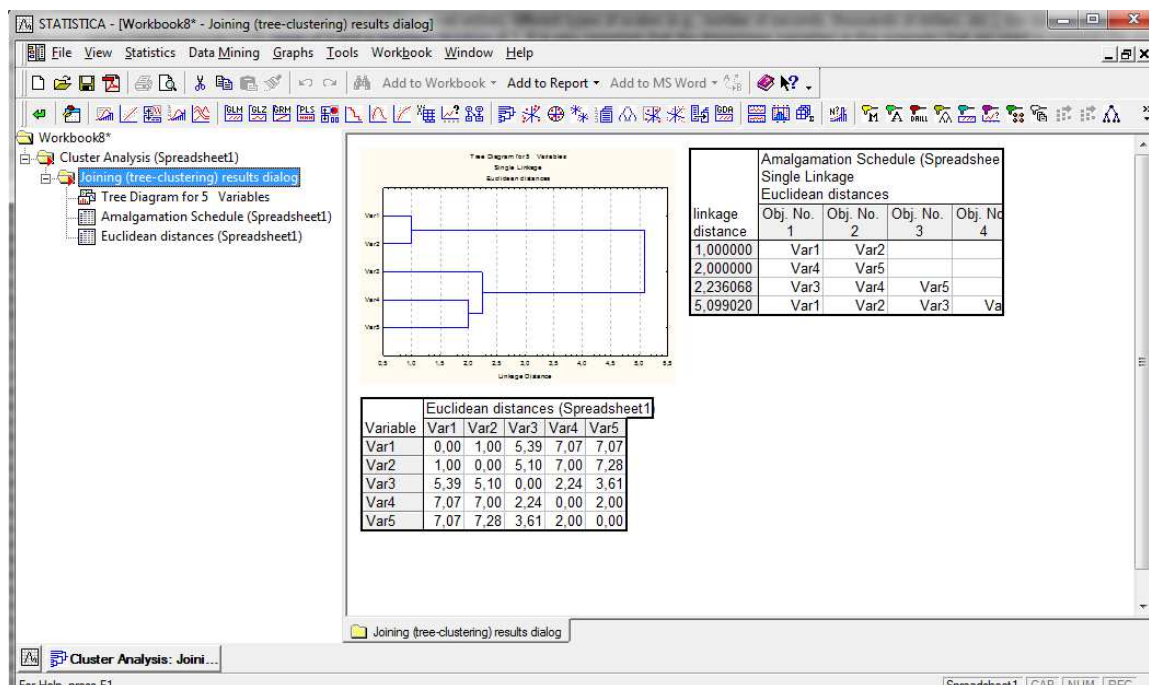
Miera vzdialenosti (*Distance measure*):

- *Squared Euclidean distances*
- *Euclidean distances*
- *City-block (Manhattan) distances*
- *Chebyshev distance metric*
- *Power: $SUM(ABS(x-y)**p)**1/r$*
- *Percent disagreement*
- *1-Pearson r*

Pre ukážku je zvolený príklad aplikovaný na hore uvedenú vstupnú dátovú maticu. Bola použitá euklidovská miera vzdialenosti a ako metóda zhľukovania bola vybraná metóda najbližšieho suseda (Single Linkage). y

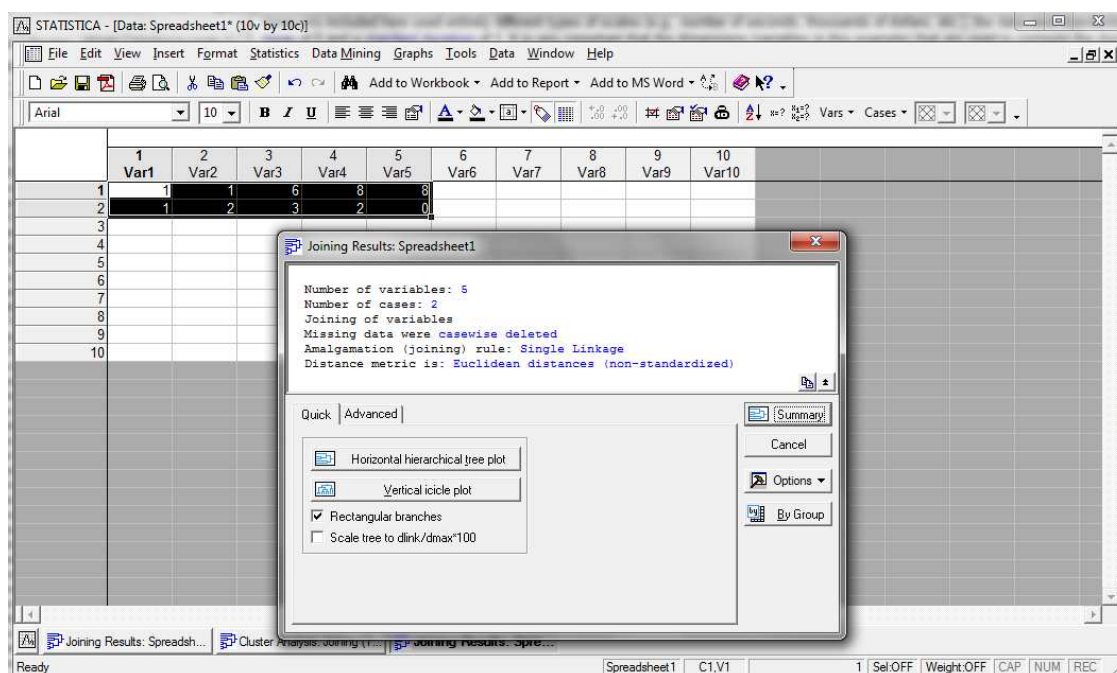


Výstup analýzy – dendrogram a proces zhlukovania tzn. Matica euklidovskej vzdialenosti a plán zhlukovania so vzdialenosťami pre jednotlivé kroky zhlukovania sú znázornené na obr.5.

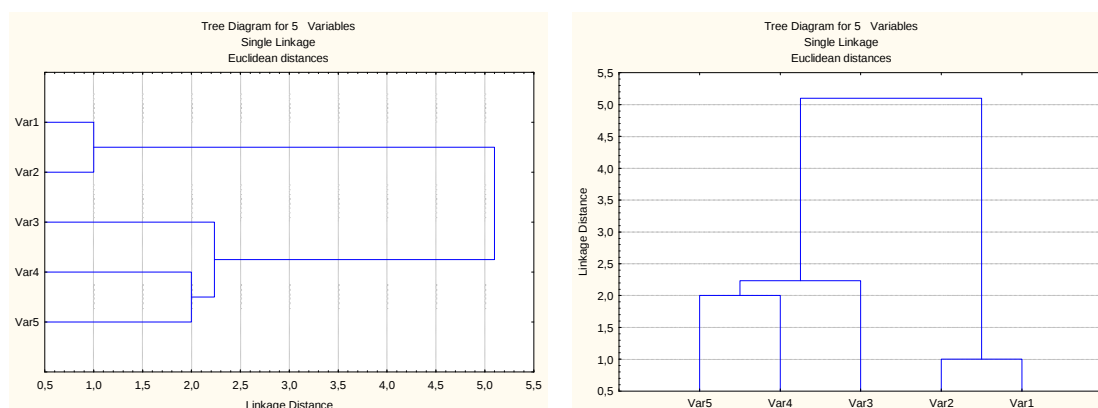


Obr. 5 Dendrogram a proces zhlukovej analýzy

Softvér STATISTICA umožňuje tvorbu dendrogramu v horizontálnej alebo vertikálnej podobe (obr.6). Príklady takto vytvorených dendrogramov sú znázornené na obr.7.



Obr. 6 Možnosť výberu horizontálneho a vertikálneho dendrogramu



Obr. 7 Horizontálny a vertikálny dendrogram

Súhrn

STATISTICA je komplexný systém obsahujúci prostriedky pre správu dát, ich analýzu, vizualizáciu a vývoj užívateľských aplikácií. Poskytuje široký výber základných aj pokročilých techník špeciálne vyvinutých pre podnikanie, data mining, vedu a inžiniersku aplikáciu.

Jedným z nástrojov softvéru STATISTICA je zhuková analýza, ktorej je venovaný tento článok. Na základe vstupných údajov je možné pomocou tohto softvéru vygenerovať dendrogramy hierarchických metód zhukovej analýzy (horizontálne a vertikálne), matice vzdialeností a pod.

Príspevok bol pripravený v rámci riešenia grantového projektu VEGA č. 1/0102/11 Metódy a techniky experimentálneho modelovania vnútropodnikových výrobných a nevýrobných procesov.

Kľúčové slová

Zhluková analýza, hierarchické metódy zhukovej analýzy, dendrogram, STATISTICA.

Použitá literatúra

- [1] ŘEZANKOVÁ, H. – HÚSEK, D. – SNÁŠEL, V.: Shluková analýza dat. Praha: Professional Publishing. 2009. 218 s. ISBN 978-80-86946-81-8,
- [2] <http://www.statsoft.cz/>.

Kontaktná adresa

Ing. Jana Halčinová
doc. Ing. Peter Trebuňa, PhD.
Technická univerzita v Košiciach
Strojnícka fakulta
Katedra priemyselného inžinierstva a manažmentu
Němcovej 32, 040 01 Košice
e-mail: jana.halcinova@tuke.sk, peter.trebuna@tuke.sk