StatSoft

Novinky *STATISTICA 12*: Jednodušší tvorba shluků

Pokud používáte ve svých analýzách shlukovou analýzu, určitě Vás tato nová funkcionalita potěší. V tomto článku si ukážeme naprosto jednoduchou a použitelnou vychytávku u shlukové analýzy.

Shluková analýza se snaží vytvářet skupiny pozorování nebo skupiny proměnných (podle toho, co potřebujeme shlukovat). V každé skupině (shluku) pak jsou prvky, které jsou si nějakým způsobem blízké (podobné). Metod pro tvorbu shluků je mnoho. My se soustředíme na metody hierarchického shlukování, protože právě zde je nová funkcionalita. Hierarchické shlukování funguje tak, že se snaží spojovat shluky postupně. Na začátku máme každý shluk ve formě jednoho bodu, poté se dva z bodů, které jsou si nejblíže, spojí do jednoho shluku, a tak shlukování pokračuje, dokud nemáme jen jeden velký shluk obsahující všechny body…

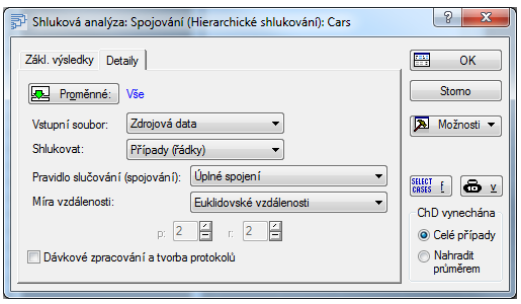
Příklad

Ukažme si vše na příkladu.

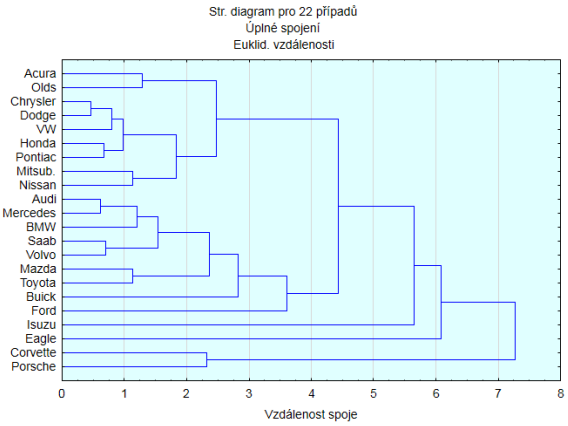
Otevřeme si datový soubor *Soubor->Otevřít příklady…->Datasets->Cars.sta*, ve kterém máme informace o vlastnostech aut. Úlohou by mohlo být podívat se, která auta mají podobné vlastnosti a můžeme je tedy zařadit do stejné skupiny (určitě si dovedeme představit, že v datech mohou být skupiny obsahující například závodní auta, americké sedany, atd.).

	Performance, fuel economy, and approximate price for various automobiles				
	1 PRICE	2 ACCELERATION	3 BRAKING	4 HANDLING	5 MILEAGE
Acura	-0,521	0,477	-0,007	0,382	2,079
Audi	0,866	0,208	0,319	-0,091	-0,677
BMW	0,496	-0,802	0,192	-0,091	-0,154
Buick	-0,614	1,689	0,933	-0,210	-0,154
Corvette	1,235	-1,811	-0,494	0,973	-0,677
Chrysler	-0,614	0,073	0,427	-0,210	-0,154
Dodge	-0,706	-0,196	0,481	0,145	-0,154
Eagle	-0,614	1,218	-4,199	-0,210	-0,677
Ford	-0,706	-1,542	0,987	0,145	-1,724
Honda	-0,429	0,410	-0,007	0,027	0,369
Isuzu	-0,798	0,410	-0,061	-4,230	1,067
Mazda	0,126	0,679	-0,133	0,500	-1,724
Mercedes	1,051	0,006	0,120	-0,091	-0,154
Mitsub.	-0,614	-1,003	0,084	0,382	0,718
Nissan	-0,429	0,073	-0,007	0,263	0,997
Olds	-0,614	-0,734	0,409	0,382	2,114
Pontiac	-0,614	0,679	0,536	0,145	0,195
Porsche	3,454	-2,215	-0,296	0,618	-1,026
Saab	0,588	0,679	0,246	0,263	0,021
Toyota	-0,059	1,218	0,228	0,736	-0,851
VW	-0,706	-0,128	0,102	0,382	0,195
Volvo	0,219	0,612	0,138	-0,210	0,369

Otevřeme dialog shlukové analýzy: *Statistiky->Vícerozměrné průzkumné techniky->Shluková analýza->Spojování (hierarchické shlukování),* proměnné vybereme všechny, v záložce *Detaily* vybereme shlukovat podle řádků a například *Úplné spojení* jako pravidlo pro slučování:



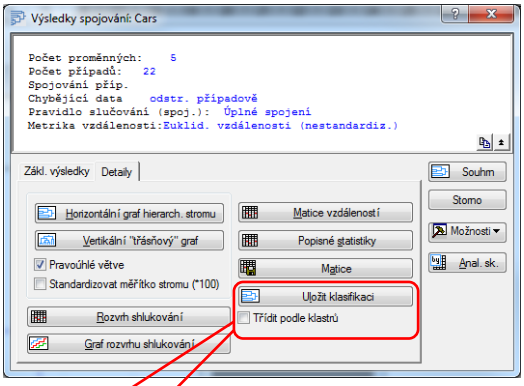
Podívejme se, jak vypadá výstup takovéto metody.

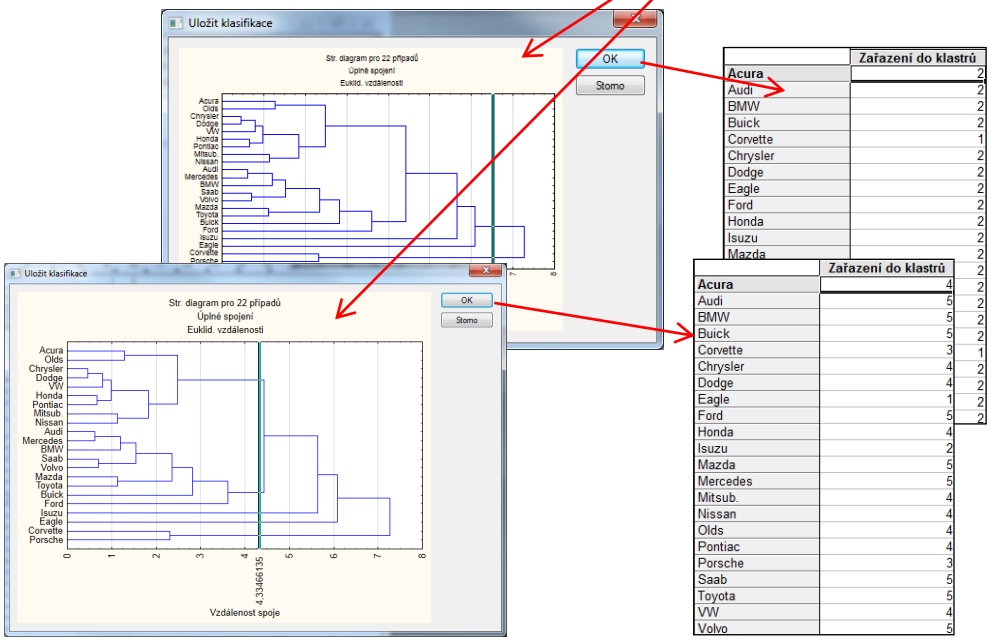


Typický výstup z takovéhoto shlukování je ve formě tzv. dendrogramu, vidíme z něj, jak probíhalo shlukování, kdy se jednotlivé shluky spojily. Nejdříve byly dány do jednoho shluku auta *Dodge* a *Chrysler*, poté *Audi* a *Mercedes* (což se asi dalo čekat), naposledy se s ostatními spojil shluk obsahující *Porsche* a *Corvetta* se zbytkem. Jediné, co zbývá rozhodnout pro rozdělení do konečných shluků je počet shluků. To se u dendrogramu dělá tak, že se zvolí hodnota vzdálenosti spoje (osa x), poté body, které byly spojeny pod touto hodnotou, jsou spojeny a tvoří shluky, spojení provedené ve větší vzdálenosti již ignorujeme.

A to je právě místo, kde může pomoct novinka, ke které se snažíme dostat. Přepněte se na záložku ***Detaily***, zde vidíte novou volbu ***Uložit klasifikaci***, po kliknutí na toto tlačítko se objeví dendrogram, ve kterém je možné manuálně posunout červenou linii dělení do shluků (viz obrázky).

Po kliknutí ***OK*** se objeví výsledné zařazení do shluků. Objeví se také dialog, jestli chcete přidat do souboru nějaké další proměnné. Předem zaškrtnutá volba ***Třídit podle klastrů*** provede seřazení podle přiřazení shluků.





	Zařazení do klastrů
Acura	2
Audi	2
BMW	2
Buick	2
Corvette	1
Chrysler	2
Dodge	2
Eagle	2
Ford	2
Honda	2
Isuzu	2
Mazda	2

	Zařazení do klastrů	
Acura	4	2
Audi	5	2
BMW	5	2
Buick	5	2
Corvette	3	1
Chrysler	4	2
Dodge	4	2
Eagle	1	2
Ford	5	2
Honda	4	
Isuzu	2	
Mazda	5	
Mercedes	5	
Mitsub.	4	
Nissan	4	
Olds	4	
Pontiac	4	
Porsche	3	
Saab	5	
Toyota	5	
VW	4	
Volvo	5	

Závěrem

Novým drobným zlepšením funkcionality shlukové analýzy je tedy grafická volba dělící hodnoty pro shluky. Věříme, že i takováto drobnost stojí za zmínku a bude jistě oceněna uživateli používajícími tuto metodologii.

StatSoft STATISTICA

StatSoft CR s. r. o.
Ringhofferova 115/1, 155 21 Praha 5 – Zličín
t +420 233 325 006, **f** +420 233 324 005
e info@statsoft.cz