

# **SHLUKOVÁ ANALÝZA KRAJŮ ČESKÉ REPUBLIKY DLE VYBRANÝCH CHARAKTERISTIK ZEMĚDĚLSTVÍ V PROGRAMU STATISTICA**

## **CLUSTER ANALYSIS OF REGIONS OF CZECH REPUBLIC BY SELECTED CHARACTERISTICS OF AGRICULTURE IN PROGRAM STATISTICA**

**Louda Zdeněk**

### **Souhrn**

Agrární sektor je jednou z nejdůležitějších součástí každé ekonomiky. Pro plnohodnotné zapojení českého zemědělství do mezinárodního agrárního obchodu je nutnou podmínkou jeho zefektivnění a zajištění konkurenceschopnosti. Nezbytným zdrojem informací pro všechny subjekty na trhu s agrárními produkty by pak měla být statistická analýza dat, která by měla zajistit objektivní informace pro správná marketingová a podnikatelská rozhodnutí. Rozsáhlou a důležitou skupinu metod tvoří metody vícerozměrné statistické analýzy, z nichž jednou z nejužívanějších je shluková analýza. Příspěvek mapuje (na praktickém příkladu seskupování krajů ČR podle vybraných charakteristik rostlinné výroby) možnosti shlukování ve statistickém paketu STATISTICA.

### **Summary**

The agrarian sector is one of the most important parts of each economy. The necessary condition for the full-value integration of the Czech agriculture into the international agrarian market is better effectiveness and competitive advantage. Then the statistical data analysis should be the inevitable information source for all subjects of the agrarian market and it should provide objective information for correct marketing and business decisions. The methods of the multivariate statistical analysis form wide and important set of methods and the cluster analysis is one of the most used of them. The paper surveys - by the practical example of clustering the regions of the Czech Republic by characteristics of vegetable production - possibilities for clustering in statistical program STATISTICA.

### **Klíčová slova**

Vícerozměrné statistické metody, shluková analýza, metody shlukování, kraje České republiky, osevní plochy.

### **Keywords**

Multivariate statistical methods, cluster analysis, clustering methods, regions of the Czech Republic, areas under farm crops.

### **Úvod**

Nezbytným zdrojem informací pro všechny subjekty na trzích s agrárními produkty je (nebo by měla být) statistická analýza dat, která by měla zajistit objektivní informace pro

správná marketingová a podnikatelská rozhodnutí. Důležitou skupinou metod, které umožňují analyzovat složitou realitu zastoupenou velkým množstvím zkoumaných proměnných, tvoří metody vícerozměrné statistické analýzy. Ty používají v algoritmu často rozsáhlý datový soubor tvořící vícerozměrnou veličinu, ale výrazně se navzájem liší v tom, na jaké cílové otázky umožňují získat odpovědi a tudíž i v typu úloh, pro jejichž řešení jsou vhodné. Nejobecnějším rozdělením do dvou skupin je třídění na:

a) **Metody analýzy korelačních a kovariančních struktur:** umožňují posuzovat vzájemné vztahy a závislosti mezi proměnnými tvořícími vícerozměrnou veličinu a podle toho hodnotit informační „důležitost“, jednotlivých proměnných a odvodit z výsledku řadu dalších informací o vlastnostech dat. Tedy z velkého počtu silněji i slaběji navzájem korelovaných proměnných získáme - s co nejmenší ztrátou informace obsažené v původních datech - podstatně redukováný a snadněji vysvětlený výsledek v podobě menšího počtu hypotetických veličin. Do této skupiny metod patří kanonická korelační analýza, analýza hlavních komponent a faktorová analýza;

b) **Metody vícerozměrné statistické klasifikace:** umožňují podle hodnot proměnných tvořících vícerozměrnou veličinu s nejmenší možnou chybovostí rozdělit soubor statistických jednotek do homogenních podsouborů, ve kterých jsou si jednotky navzájem co nejvíce podobné, zatímco jednotky z rozdílných podsouborů navzájem co nejvíce nepodobné. Do této skupiny je zařazována diskriminační analýza a shluková analýza.

## Cíl a metodika

Cílem příspěvku je provést shlukování krajů České republiky - na základě podobnosti v úrovni vybraných charakteristik rostlinné výroby rozřadit kraje do několika specifických skupin. Shluková analýza je prováděna v programu STATISTICA, příspěvek tedy zároveň mapuje možnosti tohoto statistického balíku v oblasti shlukových analýz.

Data byla čerpána z materiálů ČSÚ (Statistická ročenka ČR - 2002). Ukazatele tvořily bilance půdy (pěstební plochy jednotlivých plodin) a údaje o průměrných ročních teplotách a srážkách (předpokládané důležité charakteristiky pro rostlinnou výrobu).

Často je účelné (a někdy i nutné) pracovat v analýzách s relativními údaji, proto byly vypočteny relativní ukazatele v podobě podílu krajů na plochách plodin celé ČR, u ukazatelů meteorologických (teploty a srážky) byl vypočten relativní ukazatel odchylky hodnoty kraje od průměrné hodnoty daného ukazatele v celé ČR.

### Postup shlukové analýzy

1. *Volba vstupní databáze:* zadává se typ dat (a) proměnných (sloupců) analyzovaných objektů (řádků), (b) sloupců matice vzdáleností, (c) sloupců korelační matice.
2. *Volba druhu veličin:* zadává se typ užitečných veličin v datech - (a) intervalová, (b) ordinální, (c) nominální, (d) symetrická binární, (e) asymetrická binární, (f) poměrová.
3. *Název objektů:* pojmenování jednotlivých objektů umístěných v řádcích.
4. *Typ shlukovací techniky:* jednoduchá průměrová (Average), skupinového průměru, centroidní (Centroid), nejbližšího souseda (Single, Nearest), nejvzdálenějšího souseda (Complete, Furthest), mediánová (Median), Wardova a flexibilní.
5. *Druh užitých vzdáleností:* metrika Euklidova (geometrická vzdálenost), Hammingova (Manhattanská vzdálenost), Mahalanobisova a zobecněná Minkowskiho metrika.
6. *Postup linkování a zařazení do shluků:* výpočet vzdáleností mezi objekty a shluky a vytváření dendrogramu - (1) metodou hierarchického shlukování, (2) metodou nejbližších středů, (3) metodou středů-medoidů, (4) metodou fuzzy shlukování.
7. *Výpočet skutečných a predikovaných vzdáleností v dendrogramu:* Jsou porovnány skutečné vzdálenosti mezi objekty a vypočtené vzdálenosti (predikované) v dendrogramu, jejich rozdíl a konečné i procentuální vyjádření tohoto rozdílu.

8. *Hledání nejlepší techniky tvorby dendrogramu*: k sestrojení optimálního dendrogramu lze kombinovat řadu technik - kritériem věrohodnosti jsou především kofenetický korelační koeficient  $CC$ , míry těsnosti proložení  $\delta$ , ale také další kritéria jako mezishluková suma čtverců  $WSS_k$ , procento variace  $PV_k$ , silueta  $s$ , průměrná silueta  $SC$ , Wilkova statistika  $\lambda$ , rozdělovací koeficienty Dunnův  $F(U)$  a Kaufmanův  $D(U)$ .
9. *Vysvětlení nejlepšího dendrogramu podobností objektů*.
10. *Vysvětlení nejlepšího dendrogramu pomocí podobností proměnných*: odhalí souvislosti ve struktuře objektů analyzované databáze a je druhým důležitým cílem shlukové analýzy.

## Výsledky

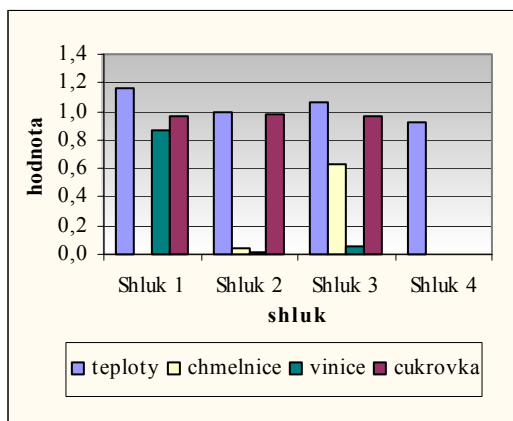
Statistický program STATISTICA obsahuje v oblasti shlukovacích metod proceduru hierarchického shlukování a proceduru shlukování pomocí k-průměrů. S přihlédnutím k rozsáhlejším a detailnějším výstupům, které poskytuje „průměrové“ shlukování je primárně použito této procedury (pro 4, 5 a 6 požadovaných shluků krajů - vzhledem k celkovému počtu krajů ČR jsou právě tyto počty shluků nejvhodnější), ovšem pro porovnání jsou uvedeny i výstupy hierarchického shlukování.

### Shlukování metodou k-průměrů pro 4 shluky krajů ČR

Tab. č. 1: Prům. hodnoty ukazatelů pro 4 shluky

|           | Shluk 1 | Shluk 2 | Shluk 3 | Shluk 4 |
|-----------|---------|---------|---------|---------|
| teploty   | 1,1573  | 0,9957  | 1,0653  | 0,9191  |
| Srážky    | 0,7995  | 1,0861  | 0,9152  | 1,0048  |
| orná půda | 0,1167  | 0,0766  | 0,0608  | 0,0698  |
| chmelnice | 0,0000  | 0,0455  | 0,6364  | 0,0000  |
| Vinice    | 0,8750  | 0,0078  | 0,0625  | 0,0000  |
| TTP       | 0,0290  | 0,0696  | 0,0714  | 0,1142  |
| obiloviny | 1,1283  | 0,9911  | 0,9049  | 0,8488  |
| brambory  | 0,8606  | 0,9558  | 0,7165  | 0,8948  |
| cukrovka  | 0,9694  | 0,9738  | 0,9723  | 0,0000  |
| Řepka     | 0,9225  | 0,9860  | 0,9014  | 0,9988  |
| pícniny   | 1,0051  | 1,0177  | 0,7432  | 0,9025  |

Graf č. 1: Průměry ukazatelů pro 4 shluky



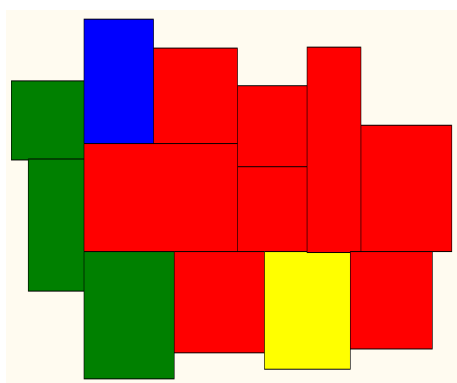
Základním výstupem programu STATISTICA v oblasti shlukování pomocí k-průměrů jsou průměrné hodnoty jednotlivých ukazatelů pro vytvořené shluky (viz. Tabulka č. 1) s dobře patrnými odlišnostmi mezi jednotlivými skupinami krajů. Například chmelnice nejsou vůbec zastoupeny ve shlucích č. 1 a č. 4 a rozhodně největší zastoupení jich je ve shluku č. 3 (logicky v Ústeckém kraji), vinice nejsou zase vůbec zastoupeny ve shluku č. 4, minimálně ve shlucích č. 2 a č. 3 a nejvíce jsou logicky zastoupeny ve shluku č. 1 (v Jihomoravském kraji) apod.

Danou situaci dokresluje i Graf č. 1, kde jsou zobrazeny statisticky významné ukazatele tohoto modelu pro všechny 4 shluky.

**Tab. č. 2: Analýza rozptylu pro 4 shluky**

|                  | Mezisk.       | sv       | Vnitřní       | sv       | F             | p             |
|------------------|---------------|----------|---------------|----------|---------------|---------------|
| <b>Teploty</b>   | <b>0,0486</b> | <b>3</b> | <b>0,0291</b> | <b>9</b> | <b>5,000</b>  | <b>0,0261</b> |
| Srážky           | 0,0933        | 3        | 0,1254        | 9        | 2,232         | 0,1537        |
| Orná půda        | 0,0020        | 3        | 0,0219        | 9        | 0,273         | 0,8428        |
| <b>chmelnice</b> | <b>0,3446</b> | <b>3</b> | <b>0,0661</b> | <b>9</b> | <b>15,634</b> | <b>0,0006</b> |
| <b>Vinice</b>    | <b>0,6931</b> | <b>3</b> | <b>0,0034</b> | <b>9</b> | <b>608,33</b> | <b>0,0000</b> |
| TTP              | 0,0069        | 3        | 0,0063        | 9        | 3,282         | 0,0724        |
| obiloviny        | 0,0761        | 3        | 0,1108        | 9        | 2,061         | 0,1758        |
| brambory         | 0,0569        | 3        | 0,1540        | 9        | 1,108         | 0,3953        |
| <b>cukrovka</b>  | <b>2,1855</b> | <b>3</b> | <b>0,0370</b> | <b>9</b> | <b>177,20</b> | <b>0,0000</b> |
| Řepka            | 0,0108        | 3        | 0,0535        | 9        | 0,602         | 0,6295        |
| pícniny          | 0,0845        | 3        | 0,1732        | 9        | 1,463         | 0,2886        |

**Obr. č. 1: Shluky (4) krajů ČR**



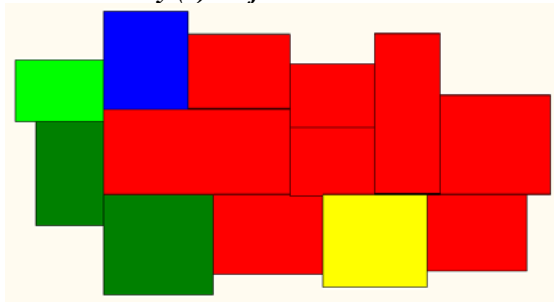
Dalším výstupem v této proceduře je tabulka analýzy rozptylu výsledného shlukovacího modelu s testy významnosti jednotlivých shlukovacích proměnných - zde z deseti použitých proměnných byly významné pouze čtyři (tučně červeně zvýrazněné - viz. Tabulka č. 2), a to teploty, chmelnice, vinice a cukrovka. Ostatní proměnné příliš nepřispívají k roztřídění krajů, nejsou pro shlukování statisticky významné. Po eliminování nevýznamných shlukovacích proměnných by bylo vhodné provést znovu proceduru shlukování pouze s těmi proměnnými, které ke spojování významně přispívají.

Samozřejmým výstupem shlukovací analýzy je výčet členů jednotlivých shluků - zde již převeden do grafické podoby „pseudomapy ČR“ s vyobrazením jednotlivých krajů a jejich příslušností jednotlivým shlukům (viz. Obrázek č. 1).

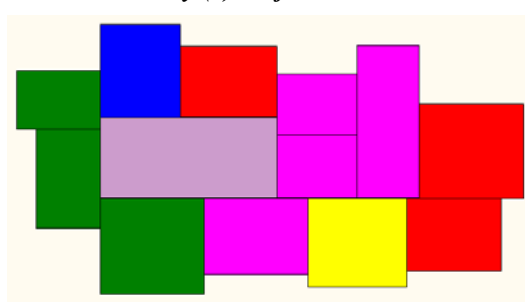
## Shlukování metodou k-průměrů pro 5 a 6 shluků krajů ČR

Pro požadovaný počet 5 shluků opět program STATISTICA nabízí stejnou strukturu výstupů. Oproti předchozímu případu (požadavek 4 shluků) dochází ve složení shluků pouze k oddělení Karlovarského kraje od Jihočeského a Plzeňského (viz. Obrázek č. 2). Analýzou rozptylu jsou opět zjištěny statisticky významné proměnné pro dané shlukování, a to teploty, chmelnice, vinice, trvalé travní porosty a cukrovka.

Obr. č. 2: Shluky (5) krajů ČR



Obr. č. 3: Shluky (6) krajů ČR



Pro požadovaných 6 shluků se již situace ve skupinách krajů značně mění (oproti požadavku 4 shluků), Karlovarský kraj zůstává ve shluku s Jihočeským a Plzeňským, samostatný shluk vytváří Praha + Středočeský kraj, a do samostatné skupiny jsou také zařazeny kraje Královéhradecký, Pardubický, Olomoucký a Vysočina (viz. Obrázek č. 3). Statisticky významnými jsou opět shledány proměnné plochy chmelnic, vinic a cukrovky a nově srážky a plochy orné půdy.

## Porovnání shlukování metodou k-průměrů s hierarchickým

Program STATISTICA umožňuje hierarchickou tvorbu shluků pomocí spojování jednoduchého (metodou nejbližšího souseda), úplného (metodou nejvzdálenějšího souseda), neváženého či váženého průměru skupin dvojic, neváženého či váženého centroidu dvojic a pomocí Wardovy metody, a dále nabízí několik měr vzdálenosti mezi objekty, z nejznámějších Euklidovské, blokové (Manhattanské) a Čebyševovy. Výstupem shlukování je pak klasický horizontální hierarchický stromový graf (dendrogram podobnosti objektů) a rozvrh shlukování včetně grafického znázornění. Z dendrogramu lze jednoduchým „subjektivním vertikálním řezem“ zjistit složení shluků. Největší skoky v grafu rozvrhu shlukování také mohou pomoci při stanovení nejvýhodnějšího počtu shluků. Pro názornost byly vybrány metody spojování jednoduchého, úplného a neváženého průměru dvojic, vše s nejpoužívanější (Euklidovskou) metrikou.

Hierarchickým shlukováním jsou vytvořeny značně podobné shluky jako metodou k-průměrů. Tvorba 4 shluků krajů ČR dopadá naprosto shodně ve všech 3 případech jako metoda k-průměrů, tvorba 5 shluků kromě metody úplného spojení, kde je vytvořen shluk krajů Královéhradeckého, Pardubického, Olomouckého, Vysočiny a Středočeského včetně Prahy, také. Ovšem při tvorbě 6 shluků je u všech tří hierarchických metod extrahován Karlovarský kraj do samostatného shluku (oproti metodě k-průměrů), a dále u metody jednoduchého spojení vzniká samostatný shluk Středočeského kraje včetně Prahy, v ostatních dvou případech pak opět výše zmiňovaný shluk pěti krajů. Pro názornost jsou v tabulce č. 3 zachyceny výsledky všech 3 hierarchických metod shlukování a k-průměrové metody.

**Tabulka č. 3: Porovnání shlukovacích metod při tvorbě 4, 5 a 6 shluků krajů ČR**

| počet<br>shluků | jednoduché spojení |       |       | úplné spojení |       |       | metoda nev. průměru |       |       | metoda k-průměrů |       |       |
|-----------------|--------------------|-------|-------|---------------|-------|-------|---------------------|-------|-------|------------------|-------|-------|
|                 | 4                  | 5     | 6     | 4             | 5     | 6     | 4                   | 5     | 6     | 4                | 5     | 6     |
| Kraje           | KV                 | KV    | KV    | KV            | KV    | KV    | KV                  | KV    | KV    | KV               | KV    | KV    |
|                 | Plz                | Plz   | Plz   | Plz           | Plz   | Plz   | Plz                 | Plz   | Plz   | Plz              | Plz   | Plz   |
|                 | Jč                 | Jč    | Jč    | Jč            | Jč    | Jč    | Jč                  | Jč    | Jč    | Jč               | Jč    | Jč    |
|                 | JM                 | JM    | JM    | JM            | JM    | JM    | JM                  | JM    | JM    | JM               | JM    | JM    |
|                 | Ú                  | Ú     | Ú     | Ú             | Ú     | Ú     | Ú                   | Ú     | Ú     | Ú                | Ú     | Ú     |
|                 | L                  | L     | L     | L             | L     | L     | L                   | L     | L     | L                | L     | L     |
|                 | Zl                 | Zl    | Zl    | Zl            | Zl    | Zl    | Zl                  | Zl    | Zl    | Zl               | Zl    | Zl    |
|                 | MS                 | MS    | MS    | MS            | MS    | MS    | MS                  | MS    | MS    | MS               | MS    | MS    |
|                 | HK                 | HK    | HK    | HK            | HK    | HK    | HK                  | HK    | HK    | HK               | HK    | HK    |
|                 | Ol                 | Ol    | Ol    | Ol            | Ol    | Ol    | Ol                  | Ol    | Ol    | Ol               | Ol    | Ol    |
|                 | Pce                | Pce   | Pce   | Pce           | Pce   | Pce   | Pce                 | Pce   | Pce   | Pce              | Pce   | Pce   |
|                 | V                  | V     | V     | V             | VI    | VII   | V                   | V     | VI    | V                | V     | V     |
|                 | P+Stč              | P+Stč | P+Stč | P+Stč         | P+Stč | P+Stč | P+Stč               | P+Stč | P+Stč | P+Stč            | P+Stč | P+Stč |

## Diskuse

V rámci shlukování pomocí k-průměrů jsou pro všechny stanovené počty shluků separovány zcela specifické kraje Jihomoravský a Ústecký. V prvním případě zde získává na významu proměnná vinic, které jsou právě v tomto regionu zastoupeny nejvíce z celé ČR, ve druhém případě je stejná situace pro proměnnou chmelnic, díky nimž má tento kraj v rámci celé republiky zcela výjimečné postavení. Při metodě jednoduchého spojení hierarchického shlukování při hledání 6 shluků dochází k často se vyskytujícímu negativnímu jevu, a to ke zřetězování, kdy jsou i naprosto odlišné objekty zahrnuty do jednoho shluku (zde v podobě čtyř „jednokrajových shluků“ a s většinou ostatních krajů zahrnutých do jediného shluku). Ostatní hierarchické metody dávají velmi obdobné výsledky jako metoda k-průměrů.

Program STATISTICA nabízí poměrně rozsáhlé uživatelské možnosti při shlukové analýze, avšak hledisko hodnocení kvality vytvořených shlukových modelů je zde zcela opomenuto. Získané výsledky lze tedy hodnotit pouze subjektivně, ovšem i přesto mají (nebo mohou mít) pro své uživatele značnou informační hodnotu.

## Závěry

Dané metody vesměs pomáhají vytvořit skupiny krajů s podobnými vlastnostmi. Ovšem uspokojující závěry lze získat pouze po velmi pečlivém a důkladném výběru shlukovacích proměnných (zvláště důležitá je jejich nekorelovanost) včetně jejich počtu, na významu v počátečních fázích práce nabývá samozřejmost stejně definovaných objektů u všech proměnných. Zdaleka největším problémem seskupování krajů ČR však je značná nedostupnost vhodných dat. Z hlediska samotného shlukování by také bylo asi vhodnější pracovat s větším počtem shlukovaných jednotek (např. vytvářet shluky na úrovni okresů), kde by hrála větší roli specifičnost daných jednotek, přeci jenom v rámci krajů existují specifické regiony, jejichž ojedinělost se pak v celkovém „číselném“ hodnocení kraje ztrácí.

## Literatura

1. Brabenec, V; Šařecová, P: Statistické metody v marketingu a obchodu - vybrané přednášky a příklady, ČZU, Praha, 2001
2. Hebák, P; Hustopecký, J: Vícerozměrné statistické metody s aplikacemi, Praha, SNTL/ALFA, 1987
3. Meloun, M; Militký, J: Kompendium statistického zpracování dat, Praha, Academia, 2002
4. Steel, R. G. D; Torrie, J. H: Principles and Procedures of Statistics, Tokyo, Tosho Printing Co., 1981
5. ČSÚ: Porovnání krajů - vybrané ukazatele 1995 - 2000, Praha, 2001
6. <http://www.czso.cz/cz/cisla/1/10/2002/index.htm>

