UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

ROZDELENIE EURÓPY POMOCOU ZHLUKOVEJ ANALÝZY

BAKALÁRSKA PRÁCA

2014 Martina HLAVATÁ

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

ROZDELENIE EURÓPY POMOCOU ZHLUKOVEJ ANALÝZY

BAKALÁRSKA PRÁCA

Študijný program:	Ekonomická a finančná matematika
Študijný odbor:	1114 Aplikovaná matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Vedúci práce:	Mgr. Tomáš Miklošovič

Bratislava 2014 **Martina HLAVATÁ**

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Martina Hlavatá
Študijný program: ekonomická a finančná matematika (Jednoodborové štúdium, bakalársky I. st., denná forma)
Študijný odbor: 9.1.9. aplikovaná matematika
Typ záverečnej práce: bakalárska
Jazyk záverečnej práce: slovenský

Názov: Rozdelenie Európy pomocou zhlukovej analýzy / *The division of Europe by cluster analysis*

Cieľ: Cieľom tejto práce bude vytvorenie prehľadu najpoužívanejších metód zhlukovej analýzy. Následne sa jednotlivé metódy aplikujú na ekonomické veličiny štátov a ich domácností za účelom vytvorenia čo najhomogénnejších skupín. Následnou analýzou vytvorených skupín sa pokúsime odpovedať na otázky, ktoré sme nastolili.

Vedúci: Mgr. Tomáš Miklošovič
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Daniel Ševčovič, CSc.

Dátum zadania: 18.10.2013

Dátum schválenia: 14.11.2013

doc. RNDr. Margaréta Halická, CSc.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie Rada by som sa poďakovala v prvom rade svojmu vedúcemu bakalárskej práce Mgr.Tomášovi Miklošovičovi za jeho odborné rady, pripomienky a podporu počas vypracovania práce. Zároveň veľmi ďakujem maminke a ocinkovi, sestričke Ivanke, priateľovi Lukáškovi a ďalším kamarátom za ich neustály záujem a morálnu podporu.

Abstrakt v štátnom jazyku

HLAVATÁ, Martina: Rozdelenie Európy pomocou zhlukovej analýzy [Bakalárska práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: Mgr. Tomáš Miklošovič, Bratislava, 2014, 81 strán.

V predloženej práci sme sa zaoberali teóriou zhlukovej analýzy, pomocou ktorej sme porovnávali rôzne európske krajiny. Naším cieľom bolo predstaviť rôzne typy zhlukovacích metód a následne ich aplikovať na rozdelenie Európy. Využitím troch rôznych algoritmov sme na základe vstupných ekonomických veličín európskych krajín vytvorili potencionálne výstupy. Ich vzájomným porovnaním sme získali takmer homogénne zoskupenia. Analýzy sme uskutočnili pre roky 2005 a 2012, čo nám umožnilo pozorovať aj medziročnú zmenu. Ako bonus sme pripravili podrobný opis priebehu štyroch metód na ilustračnom príklade.

Kľúčové slová: matica vzdialenosti, aglomeratívne a divízne hierarchické metódy, K-means, fuzzy zhlukovanie.

Abstract

HLAVATÁ, Martina: The division of Europe by cluster analysis [Bachelor Thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: Mgr. Tomáš Miklošovič, Bratislava, 2014, 81 pages.

This thesis is focused on the theory of cluster analysis, by which we compared different European countries. Our objective was to introduce various types of clustering methods and apply them on the division of Europe. Based on input economic features of European countries we created potential outputs by using three different algorithms. Comparing them we obtained almost homogenous groups. We carried out analyses for years 2005 and 2012 that provided us with the opportunity to observe an inter-period change. Furthermore, we prepared detailed description of processes of four methods by presenting an illustrative example.

Keywords: distance matrix, agglomerative and divisive methods, K-means, fuzzy clustering.

Obsah

Zoznam obrázkov **9**

Zoznam tabuliek **10**

Úvod **11**

1 Teória zhlukovej analýzy **12**

1.1 Procedúra zhlukovania 12

1.2 Meranie vzdialenosti 14

1.3 Hierarchické metódy 19

1.3.1 Aglomeratívne metódy 20

1.3.2 Divízne metódy 23

1.3.3 Nevýhody a vylepšenia 25

1.4 Nehierarchické metódy 25

1.4.1 K-means 27

1.4.2 Nevýhody a vylepšenia 27

1.5 Fuzzy metódy 29

1.5.1 Fuzzy c-means 29

1.5.2 Nevýhody a vylepšenia 30

2 Ilustračný príklad **32**

2.1 Hierachická aglomeratívna metóda 33

2.2 Hierarchická divízna metóda 36

2.3 Nehierarchická metóda 40

2.4 Fuzzy metóda 43

3 Rozdelenie Európy **47**

3.1 Vládne výdavky 48

3.2 Vládne príjmy 55

3.3 Spotreba domácností 61

3.4 Kvalita života 65

Záver **73**

Zoznam použitej literatúry **74**

Príloha A **75**

Zoznam obrázkov

1 Dendrogram 20
2 Algoritmus aglomeratívnych metód 21
3 Algoritmus divíznych metód 24
4 K-means algoritmus 28
5 Aglomeratívna priemerová metóda - ilustračný príklad 36
6 Divízna metóda - ilustračný príklad 39
7 Wardova metóda - Výdavky v roku 2012 49
8 DIANA - Výdavky v roku 2012 50
9 Wardova metóda - Výdavky v roku 2005 51
10 DIANA - Výdavky v roku 2005 52
11 Wardova metóda - Príjmy v roku 2012 56
12 DIANA - Príjmy v roku 2012 57
13 Wardova metóda - Príjmy v roku 2005 58
14 DIANA - Príjmy v roku 2005 59
15 Wardova metóda - Spotreba v roku 2005 63
16 DIANA - Spotreba v roku 2005 64
17 Wardova metóda - Kvalita v roku 2012 67
18 DIANA - Kvalita v roku 2012 68
19 Wardova metóda - Kvalita v roku 2005 69
20 DIANA - Kvalita v roku 2005 70

Zoznam tabuliek

1 Hodnoty binomických čŕt dvoch objektov 18
2 Aglomeratívne metódy zhlukovej analýzy 22
3 Lekárske fakulty na Slovensku 32
4 1.iterácia K-means - ilustračný príklad 41
5 2.iterácia K-means - ilustračný príklad 41
6 3.iterácia K-means - ilustračný príklad 42
A.1 Distribúcia vládnych prostriedkov v roku 2012 75
A.2 Distribúcia vládnych prostriedkov v roku 2005 76
A.3 Vládny príjem z daní v roku 2012 77
A.4 Vládny príjem z daní v roku 2005 78
A.5 Spotreba priemerných domácností v roku 2005 79
A.6 Kvalita života jednotlivcov v roku 2012 80
A.7 Kvalita života jednotlivcov v roku 2005 81

Úvod

V dnešnej dobe patrí k čoraz viac diskutovaným témam dianie v celej Európe. Záujem o ostatné európske krajiny vzbudzuje hlavne veľmi úzke prepojenie medzi nimi, ktoré je dôsledkom rôznych kooperatívnych združení ako Európska únia, Európske združenie voľného obchodu, Európsky hospodársky priestor alebo napríklad aj Energetická charta. Pri takejto priamej závislosti je pre všetkých prirodzené zaujímať sa o rozdiel v životnej úrovni v domácej krajine oproti zahraničiu. Vznikajú taktiež rôzne hypotézy či napríklad občania južných alebo severných krajín nemajú vhodnejšie podmienky na spokojný život.

Teória zhlukovej analýzy, alebo aj clustering, ktorá vznikla v tridsiatych rokoch 20. storočia je metóda, ktorá vo všeobecnosti zaraďuje skúmané objekty do skupín (zhlukov) tak, aby objekty v jednom zhluku mali podobné vlastnosti a objekty v rôznych odlišné. Zhluková analýza ako taká nie je jednoznačná, existuje mnoho rozdielnych algoritmov, ktoré sa líšia najmä v ponímaní zhlukov, efektivity a v cieli analýzy.

V súčasnosti predstavuje zhluková analýza rozvinutú štatistickú klasifikačnú metódu, ktorá sa však ďalej teoreticky takmer nerozširuje. Existuje veľké množstvo článkov a monografií, ktoré sa ňou zaoberajú, medzi inými aj [2], [8], ktoré nám poslúžili ako podklad teoretickej časti tejto práce. Jej využitie je veľmi obsiahle, používa sa napríklad na segmentáciu trhu spotrebiteľov v marketingovej oblasti, v poisťovníctve, ďalej pri identifikovaní nebezpečných oblastí zemetrasení, tsunami a iných prírodných katastrof. V dnešnej dobe plnej informácií je potrebné informácie taktiež prehľadne zatrieďovať. Na tento účel takisto môže slúžiť clustering, bežne to môžeme vidieť v každej knižnici.

Cieľom tejto bakalárskej práce bolo predstaviť jednotlivé metódy zhlukovej analýzy, pričom sme sa zamerali najmä na hierarchické metódy. Pomocou nich a získaných dát z [7] o spotrebe domácností v jednotlivých štátoch, vládnych výdavkoch na rôzne hospodárske odvetvia, vládnom príjme z rôznych typov daní a o parametroch modelujúcich kvalitu života, sme zatriedili štáty do skupín a analyzovali výsledky.

1 Teória zhlukovej analýzy

S narastajúcim množstvom dát a informácií prišla potreba vyvinúť metódy na ich sprehľadnenie a zatrieďovanie. Okrem iných klasifikačných metód sa začala vyvíjať aj zhluková analýza, nazývaná aj klastrovou analýzou. Výstupom tejto metódy je určitý počet zhlukov, pričom objekty v jednom zhluku majú podobné vlastnosti a objekty v rôznych zhlukoch ich majú čo najviac odlišné. Na zadeľovanie skúmaných objektov do týchto skupín existuje množstvo rôznych algoritmov, ktoré sú podrobne popísané v kapitolách 1.3, 1.4 a 1.5. Okrem zvolenej metódy je dôležité použiť vhodnú normu, ktorá je reprezentovaná funkciou vzdialenosti, respektíve funkciou podobnosti. Normám sa venujeme v kapitole 1.2. Ako podklad k tejto kapitole sme použili [8].

1.1 Procedúra zhlukovania

Vstupom pre klastrovú analýzu býva N objektov, označovaných indexami $1 \leq i \leq N$, ktoré vlastnia d čŕt, označovaných indexami $1 \leq j \leq d$. Tieto údaje sa zvyknú zapisovať do $N \times d$ matice:

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1d} \\ x_{21} & x_{22} & \dots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{Nd} \end{pmatrix} \tag{1}$$

Riadkový d-rozmerný vektor x_i je vektorom čŕt i-teho objektu, pričom prvok x_{ij} označuje hodnotu j-tej črty i-teho objektu.

Teraz si popíšeme rôzne druhy čŕt.

Črty, nazývané aj znaky, klasifikujeme:

1. podľa rozsahu oboru hodnôt ako

   - **spojité**, ak obor hodnôt je nespočítateľná množina.
   - **diskrétne**, ak obor hodnôt je konečná alebo spočítateľná nekonečná množina.
   - **binárne**, ak obor hodnôt má práve dva prvky.

2. podľa úrovne merateľnosti ako

   - **nominálne**, ak čísla slúžia iba na označenie názvov a neexistuje medzi nimi žiadny matematický vzťah.
   - **ordinálne**, ak čísla slúžia iba na označenie názvov a jediný matematický vzťah, ktorý dáva zmysel je porovnávanie.
   - **intervalové**, ak rozdiely medzi číslami sú zmysluplné, avšak neexistuje hodnota 0 a podiel čísel nemá žiadnu výpovednú hodnotu.
   - **podielové**, ak existuje absolútna 0 a podiel medzi hodnotami je zmysluplný.

   Prvé dva typy sa bežne nazývajú kvalitatívne a druhé dva kvantitatívne znaky.

Priebeh zhlukovej analýzy je popisovaný vo všeobecnosti v štyroch krokoch.

1. **Výber a extrahovanie čŕt**

   Vstupom do celej procedúry je získanie matice (1). Niektoré údaje v nej môžu byť pre nás nadbytočné, teda zvyšujú výpočtovú náročnosť a majú malú výpovednú hodnotu. Preto si v tomto kroku zvolíme, ktoré črty objektov nás zaujímajú. Následne z vybraných znakov môžeme odvodiť nové veličiny. Výber a extrahovanie sa líšia vzhľadom na cieľ procedúry a sú často závislé od našej intuícii.

2. **Voľba algoritmu**

   V druhom kroku sa zvolí vhodný algoritmus, funkcia vzdialenosti alebo kritériová funkcia. Po tejto fáze sa spustia simulácie, ktorých výsledkom sa venujú posledné dva kroky.

3. **Overenie správnosti**

   Teraz je potrebné overiť, či sme v predošlej fáze zvolili vhodný algoritmus a vhodnú normu. Na testovanie správnosti existujú tri druhy indexov: vonkajšie, vnútorné a relatívne.

4. **Vyhodnotenie výsledkov**

   Na záver navrhneme hypotézy založené na získaných výsledkoch, ilustrujeme ich v tabuľkách, grafoch alebo dendrogramoch. Dôležité je pamätať si, že zhluková analýza nám neposkytuje finálny výsledok, ale iba potencionálny výstup.

1.2 Meranie vzdialenosti

V nasledujúcej časti sa najprv oboznámime s funkciami vzdialenosti a podobnosti a s ich vlastnosťami. Následne sa zameriame na typy noriem pre spojité, binomické, diskrétne a zmiešané črty. Zároveň uvedieme spôsoby, ako postupovať v prípade, ak niektoré údaje v matici (1) nemáme k dispozícii.

Definícia (Funkcia vzdialenosti)
Funkcia $D : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ sa nazýva *funkcia vzdialenosti*, ak spĺňa nasledujúce 2 vlastnosti:

1. Symetrickosť: $D(x_i, x_j) = D(x_j, x_i)$,
2. Nezápornosť: $\forall x_i, x_j : D(x_i, x_j) \geq 0$.
   Ak aj podmienky
3. Trojuholníková nerovnosť: $\forall x_i, x_j, x_k : D(x_i, x_j) \leq D(x_i, x_k) + D(x_j, x_k)$,
4. Reflektívnosť: $(x_i = x_j) \Rightarrow D(x_i, x_j) = 0$
   sú splnené, nazývame ju *metrická funkcia vzdialenosti*. Ak platí iba reflektívnosť, ide o *semimetrickú funkciu vzdialenosti*. Metrická funkcia je považovaná za *ultrametrickú*, ak vyhovuje nasledovnej silnejšej podmienke:
5. $\forall x_i, x_j, x_k : D(x_i, x_j) \leq \max\{D(x_i, x_k), D(x_j, x_k)\}$.

Poznámka 1: Funkcia vzdialenosti sa definuje ako $D(x_i, x_j) = \sum_{l=1}^{d} d_l(x_{il}, x_{jl})$ pričom d označuje počet čŕt objektov a $d_l(x_{il}, x_{jl})$ vzdialenosť objektov x_i a x_j v l-tej črte. Konkrétny tvar je ovplyvnený voľbou normy.

Definícia (Funkcia podobnosti)
Funkcia $S : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ sa nazýva *funkcia podobnosti*, ak spĺňa nasledujúce vlastnosti:

1. Symetrickosť: $S(x_i, x_j) = S(x_j, x_i)$,
2. Nezápornosť: $\forall x_i, x_j : 0 \leq S(x_i, x_j) \leq 1$.
   Ak aj podmienky
3. $\forall x_i, x_j, x_k : S(x_i, x_j) \cdot S(x_j, x_k) \leq [S(x_i, x_j) + S(x_j, x_k)] \cdot S(x_i, x_k)$,

4. $(x_i = x_j) \Rightarrow (S(x_i, x_j) = 1)$

   platia, nazývame ju *metrická funkcia podobnosti.*

Poznámka 2: Matica vzdialenosti a aj matica podobnosti sú symetrické $N \times N$ matice s prvkami $D_{ij} = D(x_i, x_j)$, respektíve $S_{ij} = S(x_i, x_j)$.

Chýbajúce hodnoty

Doteraz sme považovali za vstup do procedúry $N \times d$ maticu X, ktorej všetky prvky boli číselné a zmysluplné. Môže sa nám však stať, že niektoré hodnoty nebudú dostupné. Uvedieme možné riešenia tohto problému.

- Objekty, ktorým chýba hodnota niektorej črty vylúčime úplne z procesu. Algoritmus prevedieme iba pre tie objekty, ktoré majú všetky údaje dostupné a vo vhodnom tvare.

  Pozor! Toto riešenie obnáša riziko výrazného poklesu počtu objektov a môže znehodnotiť získanú informáciu.

- Modifikujeme funkciu vzdialenosti, tak, aby prípadné chýbajúce hodnoty počítala za absolútnu nulu:

$$D(x_i, x_j) = \frac{d}{d - \sum_{l=1}^{d} \delta_{ijl}} \cdot \sum_{\forall l \ a \ \delta_{ijl}=0} d_l(x_{il}, x_{jl}), \tag{2}$$

  pričom

$$\delta_{ijl} = \begin{cases} 1 & \text{ak } x_{il} \text{ alebo } x_{jl} \text{ chýba.} \\ 0 & \text{inak.} \end{cases} \tag{3}$$

  Popis zvyšných neznámych je uvedený v *Poznámke 1.*

- Chýbajúca hodnota x_{il} spôsobuje, že nie sme schopní určiť vzdialenosť objektu i so žiadným iným objektom $j \in \{1, 2, \ldots, N\} \backslash \{i\}$ v l-tej črte. Ako tretie riešenie problému sa ukázalo nahradenie neznámej vzdialenosti aritmetickým priemerom vzdialeností v danej črte medzi všetkými objektami s dostupnou hodnotou. Tento aritmetický priemer vyrátame podľa:

$$\bar{d}_l = \frac{2}{M \cdot (M-1)} \cdot \sum_{j=2}^{M} \sum_{i=1}^{j-1} d_l(x_{il}, x_{jl}), \tag{4}$$

  kde M je počet objektov s dostupnou hodnotou l-tej črty, pričom objekty sú prečíslované indexami od 1 po M. Potom v pôvodnom označení $\forall j \in \{1, 2, \ldots, N\}$, $j \neq i$ určíme chýbajúcu vzdialenosť $d_l(x_{il}, x_{jl}) = \bar{d}_l$.

Škálovanie

Posledná dôležitá úprava dát, pred tým ako na ne aplikujeme určitý algoritmus, je škálovanie prvkov matice X. V prípade, že hodnoty jednej črty sú rádovo väčšie ako iných čŕt, príslušné vzdialenosti objektov budú ovplyvnené najmä touto črtou, čo zníži dôležitosť informácie z ostatných čŕt. Pokiaľ požadujeme, aby každá črta mala rovnakú váhu, ako ideálne riešenie sa ukazuje aproximatívne transformovanie na normálne rozdelenie so strednou hodnotou 0 a varianciou 1. Túto transformáciu prevedieme pre všetky pôvodné prvky matice X, ktoré vo vzorci označíme ako x^*_{il}, nasledovne:

$$\begin{aligned} x_{il} &= \frac{x^*_{il} - m_l}{s_l}, \\ m_l &= \frac{1}{N}\sum_{i=1}^{N} x^*_{il}, \\ s_l &= \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x^*_{il} - m_l)^2}. \end{aligned} \tag{5}$$

Funkcia vzdialenosti pre spojité črty

Ako sme už spomínali, spojité črty môžu nadobúdať nekonečno hodnôt. V tejto časti uvedieme zoznam noriem, ktoré nám zrátajú vzdialenosť medzi takými črtami. Asi najznámejšou normou je *Euklidovská norma*, ktorá patrí medzi metrické funkcie vzdialenosti:

$$D(x_i, x_j) = (\sum_{l=1}^{d} \mid x_{il} - x_{jl} \mid^2)^{\frac{1}{2}}. \tag{6}$$

Jej všeobecnou formou je *Minkovskeho norma* nazývaná aj p-norma:

$$\forall p > 0 : D(x_i, x_j) = (\sum_{l=1}^{d} \mid x_{il} - x_{jl} \mid^p)^{\frac{1}{p}}. \tag{7}$$

Ak $p = 2$ získame vyššie uvedenú Euklidovskú normu. Dosadením $p = 1$ dostaneme L_1*-normu* nazývanú aj Mannhatanská alebo poštárska norma:

$$D(x_i, x_j) = \sum_{l=1}^{d} \mid x_{il} - x_{jl} \mid . \tag{8}$$

Poslednú významnú normu odvodenú od Minkovskeho, ktorá sa volá L_∞ alebo maximová norma, dostaneme ak $p = \infty$:

$$D(x_i, x_j) = \max_{1 \le l \le d} \mid x_{il} - x_{jl} \mid . \tag{9}$$

Teraz si predstavíme niekoľko menej známych noriem. Prvá v poradí je metrická *štvorcová Mahalanobis norma*:

$$D(x_i, x_j) = \sqrt{(x_i - x_j)^\top S^{-1}(x_i - x_j)}, \tag{10}$$

kde S je kovariančná matica $S = E((x - E(x))(x - E(x))^\top)$ a $E(\cdot)$ označuje strednú hodnotu náhodnej premennej.

Ako druhú uvedieme *bodovo symetrickú normu*. Nech x_r označuje referenta klastru a $\|.\|$ Euklidovskú normu. Tento typ sa využíva v algoritmoch, v ktorých nie je dôležitá vzdialenosť každých dvoch objektov, ale iba vzdialenosť od referenta zhluku, čo je práve jeden objekt v každom klastri.

$$D(x_i, x_r) = \min_{j \in \{1,2,\ldots,N\}, j \neq i} \frac{\|x_i - x_r + x_j - x_r\|}{\|x_i - x_r\| + \|x_j - x_r\|}. \tag{11}$$

Nasledujúca je založená na *Pearsonovom korelačnom koeficiente* r_{ij}:

$$\begin{aligned} D(x_i, x_j) &= \frac{1 - r_{ij}}{2}, \\ r_{ij} &= \frac{\sum_{l=1}^{d}(x_{il} - \bar{x_i})(x_{jl} - \bar{x_j})}{\sqrt{\sum_{l=1}^{d}(x_{il} - \bar{x_i})^2 \sum_{l=1}^{d}(x_{jl} - \bar{x_j})^2}}, \\ \bar{x_i} &= \frac{1}{d}\sum_{l=1}^{d} x_{il}. \end{aligned} \tag{12}$$

Poslednú funkciu vzdialenosti odvodíme cez *kosínusovú funkciu podobnosti*. Čím sú 2 objekty podobnejšie, tým sú viac paralelné a tým je hodnota $\cos\alpha$ väčšia, teda bližšia k hodnote 1:

$$\begin{aligned} S(x_i, x_j) = \cos\alpha &= \frac{x_i^\top x_j}{\|x_i\|\|x_j\|}, \\ D(x_i, x_j) &= 1 - S(x_i, x_j). \end{aligned} \tag{13}$$

Funkcia podobnosti pre diskrétne črty

V prípade diskrétnych čŕt začneme jednoduchším typom a to binárnym. Uvažujme, že tento typ čŕt nadobúda 2 hodnoty: 0, 1. Každý objekt buď má daný znak $\rightarrow 1$ alebo nemá $\rightarrow 0$. Hodnoty dvoch objektov vzhľadom na všetky znaky sa dajú znázorniť tak ako v tabuľke 1, kde n_{11} označuje počet čŕt, ktoré majú oba objekty, n_{00}, ktoré nemá ani jeden, n_{01}, ktoré prvý nemá a druhý má a nakoniec n_{10} počet znakov, ktoré prvý objekt má a druhý nie.

Pomocou týchto štyroch hodnôt vyjadríme podobnosť $S(x_i, x_j)$. Z logického hľadiska rozlíšime 2 prípady.

Tabuľka 1: Hodnoty binomických čŕt dvoch objektov

	$x_i = 1$	$x_i = 0$
$x_j = 1$	n_{11}	n_{10}
$x_j = 0$	n_{01}	n_{00}

1. Ak hodnoty 1 a 0 majú rovnocenný význam (napríklad ak daná črta označuje pohlavie, 1=žena, 0=muž), tak ide o rovnocenné pomenovanie. V tomto symetrickom prípade je funkcia podobnosti:

$$S(x_i, x_j) = \frac{n_{11} + n_{00}}{n_{11} + n_{00} + w(n_{10} + n_{01})}, \tag{14}$$

   kde w označuje parameter váhy zmiešaných hodnôt n_{01} a n_{10}.

2. Ak však iba hodnota 1 je významná a 0 nie je, lebo absencia tejto črty nemá výpovednú hodnotu, tak podobnosť vyrátame nasledovne:

$$S(x_i, x_j) = \frac{n_{11}}{n_{11} + w(n_{10} + n_{01})}. \tag{15}$$

Poznámka 3: Parameter váh w je volený podľa dôležitosti príspevku prítomnosti zmiešanej dvojice n_{01} alebo n_{10} do podobnosti objektov. Jeho zvyčajné hodnoty sú 1/2, 1 alebo 2.

Teraz sa zamerajme na diskrétne znaky, ktoré nadobúdajú viac hodnôt ako 2. Jednou možnosťou je spracovať ich ako viac binárnych čŕt. Avšak, používanejšia a jednoduchšia metóda vychádza z váženého aritmetického priemeru počtu čŕt, v ktorých sa dva objekty zhodujú:

$$S(x_i, x_j) = \frac{1}{d} \sum_{l=1}^{d} S_{ijl}, \tag{16}$$

kde

$$S_{ijl} = \begin{cases} 0 & \text{ak sa } x_i \text{ a } x_j \text{ nezhodujú v } l\text{-tej črte.} \\ w & \text{inak.} \end{cases}$$

Funkcia vzdialenosti pre zmiešané črty

Na záver si ukážeme jeden spôsob ako určiť podobnosť objektov, definovaných rôznymi typmi znakov. Tento vzorec sa nápadne podobá na vyššie uvedený vzorec (16). Líši sa vo výpočte S_{ijl} v závislosti od typu črty a okrem toho, chýbajúce hodnoty nebudeme

nahradzovať aritmetickým priemerom, ale ich jednoducho vynecháme.

$$S(x_i, x_j) = \frac{\sum_{l=1}^{d}(1 - \delta_{ijl})S_{ijl}}{\sum_{l=1}^{d}(1 - \delta_{ijl})} \tag{17}$$

Pre diskrétne črty l definujeme:

$$S_{ijl} = \begin{cases} 0 & \text{ak sa } x_i \text{ a } x_j \text{ nezhodujú v } l\text{-tej črte.} \\ w & \text{inak.} \end{cases}$$

Pre spojité črty l definujeme:

$$S_{ijl} = 1 - \frac{\mid x_{il} - x_{jl} \mid}{\max_{1 \leq m \leq N} x_{ml} - \min_{1 \leq m \leq N} x_{ml}}.$$

Pripomeňme si aj vzorec (3):

$$\delta_{ijl} = \begin{cases} 1 & \text{ak } x_{il} \text{ alebo } x_{jl} \text{ chýba,} \\ 0 & \text{inak,} \end{cases}$$

z ktorého je na prvý pohľad zrejmé, že $1 - \delta_{ijl} = 1$ ak ani x_{il} ani x_{jl} nechýbajú a naopak $1 - \delta_{ijl} = 0$ ak aspoň jednu nepoznáme.

1.3 Hierarchické metódy

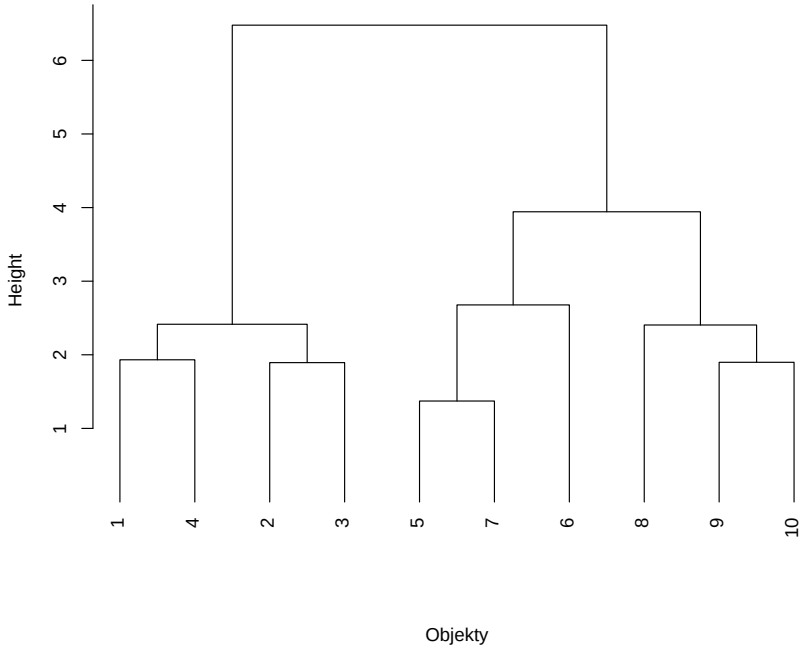
Hierarchické metódy predstavujú prvú skupinu metód zhlukovej analýzy. Na základe matice vzdialenosti, respektíve matice podobnosti, vytvárajú postupnosť hierarchicky vnorených zhlukov, buď od jednoprvkových zhlukov po 1 zhluk obsahujúci všetky objekty, alebo naopak. Prvý z uvedených typov popisuje *aglomeratívne* a druhý *divízne metódy*. Pri aglomeratívnom princípe sa v každom kroku 2 najbližšie zhluky zlúčia do jedného, v divíznom sa naopak 2 najvzdialenejšie prvky jedného zhluku rozdelia na dva zhluky. Priebeh oboch typov sa zobrazuje v dendrograme.

Dendrogram

Dendrogram je hierarchický graf, zobrazujúci postupné spájanie alebo rozpájanie klastrov. Vrchný uzol sa nazýva koreň a reprezentuje zhluk obsahujúci všetky objekty. Koncové uzly sa nazývajú listy, pričom ide o jedno-objektové klastre. Všetky vnútorné uzly predstavujú zhluky, pričom ich výška udáva informáciu o podobnosti objektov v ňom obsiahnutých: h_{ij} označuje výšku najmenšieho klastru, ktorý obsahuje i-ty aj j-ty objekt.

Pri analyzovaní výsledkov hierarchického algoritmu je dendrogram preseknutý horizontálnou priamkou v určitej výške. Úroveň výšky preseknutia môže byť určená podľa rôznych kritérii, napríklad počet zhlukov, do ktorých objekty majú byť zatriedené, maximálna povolená vzdialenosť objektov vnútri jedného zhluku alebo minimálna vzdialenosť objektov v rôznych zhlukoch.

Obr. 1: Dendrogram



1.3.1 Aglomeratívne metódy

Na úvod si popíšeme všeobecný algoritmus aglomeratívnych metód zobrazený na obrázku 2 a následne uvedieme viaceré možnosti na špecifikovanie jednotlivých krokov.

1. Začína sa vždy s N klastrami $(C_i)_{i\in\{1,\dots,N\}}$, pričom $\forall i \in \{1,\dots,N\} : C_i = x_i$. Pre ne vyrátame maticu vzdialenosti D s prvkami $D_{i,j} = D(C_i, C_j)$.

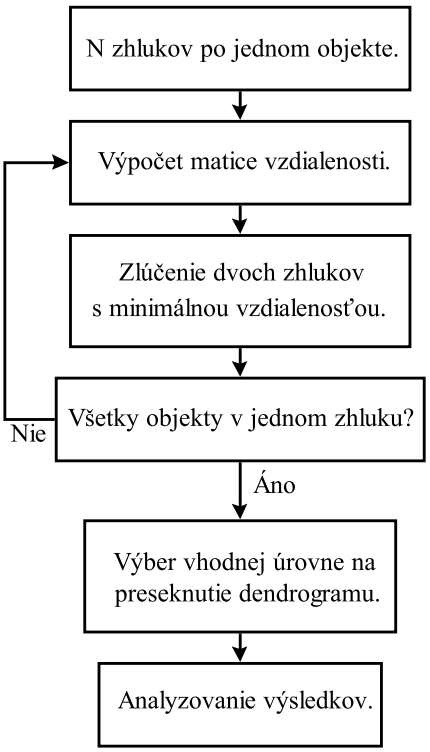
2. V matici D určíme minimálnu hodnotu $D(C_i, C_j) = \min_{1\leq k,l\leq N, k\neq l} D(C_k, C_l)$ a spojíme zhluky C_i a C_j do nového zhluku C_{ij}.

3. Aktualizujeme maticu D, pričom nanovo musia byť určené vzdialenosti medzi

novým zhlukom C_{ij} a ostatnými zhlukmi C_l. Odstránime j-ty riadok aj stĺpec a do i-teho riadku aj stĺpcu zapíšeme vzdialenosti charakterizujúce nový klaster.

4. Opakujeme kroky 2 a 3, až kým nám neostane práve 1 klaster. Následne vytvoríme dendrogram a určíme výšku hladiny jeho preseknutia. Pomocou získaných klastrov analyzujeme výsledky algoritmu.

Obr. 2: Algoritmus aglomeratívnych metód



Na vyrátanie prvkov matice vzdialenosti je potrebné na začiatok zvoliť vhodnú normu v závislosti od typu čŕt objektov (viac v časti 1.2). Avšak na aktualizáciu to nie je postačujúce. Okrem normy treba zvoliť spôsob, akým sa vyráta vzdialenosť medzi dvomi zhlukmi s viacerými objektami. Uvažujme situáciu v kroku 3. Klastre C_i a C_j sa zlúčili do C_{ij}, a teda je potrebné určiť ich vzdialenosť so všetkými ostatnými zhlukmi $D(C_l, C_{ij})$. Nasledujúci vzorec bol navrhnutý v roku 1967:

$$D(C_l, C_{ij}) = \alpha_i D(C_l, C_i) + \alpha_j D(C_l, C_j) + \beta D(C_i, C_j) + \gamma \mid D(C_l, C_i) - D(C_l, C_j) \mid \quad (18)$$

Voľbou rôznych kombinácií parametrov $\alpha_i, \alpha_j, \beta, \gamma$ sú určené viaceré klastrové aglomeratívne metódy. Najznámejšie z nich sú získané zo zdrojov [2] a [8]. Ich prehľad, uvedený v tabuľke 2, teraz postupne popíšeme.

Tabuľka 2: Aglomeratívne metódy zhlukovej analýzy

Algoritmus	α_i	α_j	β	γ
Metóda najbližšieho suseda	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Metóda najvzdialenejšieho suseda	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Priemerová metóda	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	0	0
Vážená priemerová metóda	$\frac{1}{2}$	$\frac{1}{2}$	0	0
Centroidná metóda	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	$-\frac{n_i n_j}{(n_i+n_j)^2}$	0
Mediánová metóda	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0
Flexibilná metóda	$\frac{1}{2}(1-\beta)$	$\frac{1}{2}(1-\beta)$	β	0
Wardova metóda	$\frac{n_i+n_l}{n_i+n_j+n_l}$	$\frac{n_j+n_l}{n_i+n_j+n_l}$	$-\frac{n_l}{n_i+n_j+n_l}$	0

Legenda: symbolom n_i sa označuje počet objektov v zhluku C_i

1. *Metóda najbližšieho suseda* (angl. single linkage), berie do úvahy najkratšiu vzdialenosť dvoch prvkov z rôznych zhlukov:

$$D(C_l, C_{ij}) = \min(D(C_l, C_i), D(C_l, C_j)). \tag{19}$$

   Práve preto môže produkovať veľmi pretiahnuté zhluky obsahujúce 2 úplne odlišné objekty, iba kvôli postupnému zreťazeniu ostatných objektov. Napriek tomuto problému, táto metóda funguje dobre ak sú jednotlivé zhluky od seba dostatočne vzdialené.

2. Naopak *metóda najvzdialenejšieho suseda* (angl. complete linkage) berie do úvahy najdlhšiu možnú vzdialenosť dvoch prvkov z rôznych klastrov:

$$D(C_l, C_{ij}) = \max(D(C_l, C_i), D(C_l, C_j)). \tag{20}$$

   Ide o efektívnu metódu pri hľadaní malých kompaktných skupín.

3. Kompromis medzi týmito dvomi metódami vytvárajú *priemerová a vážená priemerová metóda*, ktoré zohľadňujú vzdialenosť ku obom objektom v novom klastri. Navzájom sa od seba líšia faktom, že priemerová metóda uvažuje aj počet objektov v klastroch C_i a C_j, pričom vážená prikladá rovnakú váhu obom klastrom.

4. *Centroidná metóda* je založená na vzdialenostiach centroidov zhlukov. Centroid je vektor aritmetických priemerov každej jednej črty, teda: $m_i = \frac{1}{n_i} \sum_{x \in C_i} x$. Potom

platí:

$$D(C_l, C_{ij}) = \frac{n_i}{n_i + n_j} D(C_l, C_i) + \frac{n_j}{n_i + n_j} D(C_l, C_j) - \frac{n_i n_j}{(n_i + n_j)^2} D(C_i, C_j) \\ = \|m_l - m_{ij}\|^2, \tag{21}$$

kde $\|.\|$ označuje Euklidovskú normu.

5. Podobne je určená *mediánová metóda*, ktorá však udáva rovnakú váhu obom zhlukom:

$$D(C_l, C_{ij}) = \frac{1}{2} D(C_l, C_i) + \frac{1}{2} D(C_l, C_j) - \frac{1}{4} D(C_i, C_j). \tag{22}$$

6. Ako jediná z uvedených metód závisí od parametra *flexibilná metóda*. Jej cieľom je demonštrovať efekt voľby rôznych parametrov β na tvar výsledného dendrogramu.

7. *Wardova metóda* známa aj ako metóda minimálnej variancie minimalizuje nárast sumy štvorcovej odchýlky objektov od ich centroidov. Nech K je aktuálny počet klastrov a body m_l sú centroidy zhlukov. Potom suma štvorcovej odchýlky objektov od ich centroidov je

$$E = \sum_{l=1}^{K} \sum_{x_i \in C_l} \|x_i - m_l\|^2.$$

Pre zmenu funkcie vzdialenosti je relevantný iba nárast E po zlúčení zhlukov C_i a C_j, reprezentovaný ako:

$$\triangle E_{ij} = \frac{n_i n_j}{n_i + n_j} \|m_i - m_j\|^2.$$

Vzorec (18) má v tomto prípade tvar

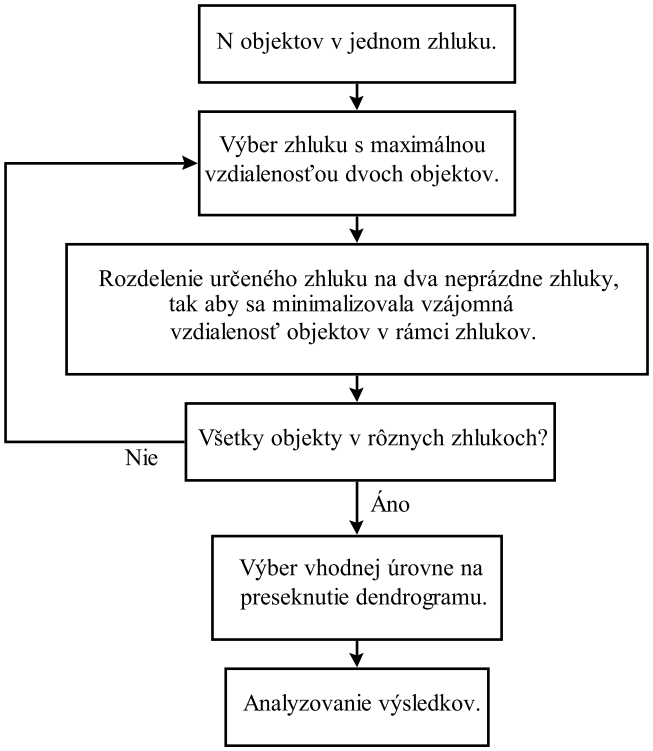
$$D(C_l, C_{ij}) = \frac{n_i + n_l}{n_i + n_j + n_l} D(C_l, C_i) + \frac{n_j + n_l}{n_i + n_j + n_l} D(C_l, C_j) - \frac{n_l}{n_i + n_j + nl} D(C_i, C_j). \tag{23}$$

1.3.2 Divízne metódy

Druhý typ hierarchických metód sú divízne metódy. Na rozdiel od aglomeratívnych, na začiatku sú všetky objekty v jednom klastri, postupne sa rozdeľujú do viacerých až nakoniec získame N klastrov po jednom objekte. Priebeh algoritmu je zobrazený

na obrázku 3. Táto metóda je výpočtovo náročnejšia, keďže na rozdelenie klastru s N objektami na dva neprázdne zhluky je potrebné zvážiť až $2^{N-1} - 1$ možností. Preto je aj menej využívanejšia. Jej hlavnou výhodou však je poskytnutie obrazu hlavnej štruktúry s väčšími zhlukmi, ktorý dostávame na začiatku algoritmu, a preto netrpí akumulovaním chybných rozhodnutí.

Obr. 3: Algoritmus divíznych metód



DIANA

Na zníženie výpočtovej náročnosti bol navrhnutý systém DIANA (Divisive Analysis). Ide o iteračný algoritmus riešiaci, ako vhodne rozdeliť zhluk obsahujúci maximálne vzdialený pár objektov, nech sa volá C_l, na 2 menšie zhluky, nech sa volajú C_i a C_j.

1. Na začiatok sa zatriedia všetky objekty z C_l do C_i, pričom $C_j = \emptyset$.

2. V prvej iterácii vyrátame pre všetky objekty z C_i priemernú vzdialenosť od zvyšných objektov v C_i nasledovne:

$$d(x_m, C_i \backslash \{x_m\}) = \frac{1}{n_i - 1} \sum_{x_p \in C_i, p \neq m} D(x_m, x_p)^1. \tag{24}$$

[1] $D(x_i, x_j)$ je určené zvolenou normou.

3. Objekt s maximálnou vzdialenosťou $d(x_m, C_i \backslash \{x_m\})$ sa presunie do C_j.

4. Vo zvyšných iteráciach určíme pre všetky objekty z C_i rozdiel medzi priemernou vzdialenosťou s objektami v C_i a priemernou vzdialenosťou s objektami v C_j:

$$d(x_m, C_i \backslash \{x_m\}) - d(x_m, C_j) = \frac{1}{n_i - 1} \sum_{x_p \in C_i, p \neq m} D(x_m, x_p) - \frac{1}{n_j} \sum_{x_q \in C_j} D(x_m, x_q). \tag{25}$$

5. Ak maximálna hodnota rozdielu vzdialeností $d(x_m, C_i \backslash \{x_m\}) - d(x_m, C_j)$ je kladná, presunieme príslušný objekt do C_j a opakujeme kroky 4 a 5. V opačnom prípade ukončíme túto iteračnú sériu.

1.3.3 Nevýhody a vylepšenia

Napriek vhodnému upraveniu dát preškálovaním a vysporiadaním sa s chýbajúcimi hodnotami, bývajú klasické hierarchické metódy kritizované najmä kvôli ich trom vlastnostiam. Prvá nevýhoda je ich **výpočtová náročnosť**, najmenej $O(N^2)$, ktorá spôsobuje nevhodnosť uvedených algoritmov pre veľký počet objektov N. Druhou nevýhodou je fakt, že akonáhle je objekt zadelený do určitého klastru, je považovaný za súčasť tohto zhluku a nie ako jednotlivý objekt. To znamená, že prípadná predošlá chyba v klasifikácii nemôže byť opravená. Poslednou nevýhodou je nesenzitívnosť týchto metód k tzv. outliers, teda k objektom, ktoré sú výrazne odlišné od ostatných objektov. Prítomnosť outliers v uvažovanom sete objektov môže výrazne skresliť výsledky.

Reakciou na túto kritiku bol vývoj nových algoritmov snažiacich sa napraviť uvedené nevýhody a pritom zachovať hierarchickú štruktúru výsledkov. Medzi tieto vylepšenia patria algoritmy *BIRCH, CURE, ROCK, Chameleon* a ďalšie. Viac sa dočítate v zdrojoch [5], [6] a [8].

1.4 Nehierarchické metódy

Na rozdiel od hierarchických metód, nehierarchické nevytvárajú žiadnu štruktúru postupne vnorených zhlukov. Ich výstupom je jednoducho rozdelenie N objektov do vopred určeného počtu K klastrov $\{C_1, \ldots, C_K\}$, pričom $K \leq N$. Tieto zhluky musia spĺňať nasledovné 3 podmienky:

- $\forall i \in \{1, \ldots, K\}$: $C_i \neq \emptyset$
- $\bigcup_{i=1}^{K} C_i = X$
- $\forall i, j \in \{1, \ldots, K\}, i \neq j$: $C_i \cap C_j = \emptyset$

Jednotlivé algoritmy sú založené na minimalizácii alebo maximalizácii kritériovej funkcie J. Prvou možnosťou, výpočtovo veľmi náročnou, je postupne odskúšať všetky kombinácie rozdelenia objektov do K zhlukov, ktoré spĺňajú určené tri podmienky. Preto sa vytvorili menej náročné algoritmy, napríklad K-means metóda.

Suma štvorcových chýb

Nech všetky objekty $x_1, x_2, \ldots, x_N \in \mathbb{R}^d$ sú zaradené do K zhlukov $\{C_1, \ldots, C_K\}$. Kritériová funkcia

$$J_S(\Gamma, M) = \sum_{i=1}^{K} \sum_{j=1}^{N} \gamma_{ij} \|x_j - m_i\|^2 \tag{26}$$

kde Γ je matica obsahujúca koeficienty

$$\gamma_{ij} = \begin{cases} 1 & \text{ak } x_j \in C_i, \\ 0 & \text{inak}, \end{cases}$$

$M = [m_1, \ldots, m_K]$ je matica centroidov, kde $m_i = \frac{1}{n_i} \sum_{j=1}^{N} \gamma_{ij} x_j$, sa nazýva suma štvorcových chýb. Jej minimalizovanie patrí k častým prostriedkom pre nehierarchické metódy. Je vhodné najmä pre kompaktné množiny.

Ku ďalším kritériovým funkciám patria v prípade $d = 1$:
$S_T = \sum_{j=1}^{N} (x_j - m)(x_j - m)^\top$ totálna rozptylová matica,
$S_V = \sum_{i=1}^{K} \sum_{j=1}^{N} \gamma_{ij} (x_j - m_i)(x_j - m_i)^\top$ vnútro-zhluková rozptylová matica,
$S_M = \sum_{i=1}^{K} n_i (m_i - m)(m_i - m)^\top$ medzi-zhluková rozptylová matica,
kde $m = \frac{1}{N} \sum_{i=1}^{K} n_i m_i$ je totálny priemer a platí $S_T = S_V + S_M$.

Ak $d > 1$, je potrebné rozptylové matice preškálovať buď ich determinantom alebo stopou (súčet prvkov na hlavnej diagonále). Ako príklad použitia stopy si všimnime, že stopa matice S_V je v skutočnosti spomínaná suma štvorcových chýb (26): $tr(S_V) = J_S(\Gamma, M)$. Platí dokonca $\min J_S(\Gamma, M) \Leftrightarrow \min tr(S_V) \Leftrightarrow \max tr(S_M)$. Ako skalarizovanie pomocou determinantu sa zvykne používať $\min det(S_V)$. Aplikovanie determinantu na maticu S_M nie je vhodné, lebo je singulárna v prípade ak $K \leq d$.

1.4.1 K-means

Najpopulárnejšia nehierarchická metóda sa nazýva K-means, teda metóda K-priemerov. Jej obľúbenosť spočíva v približne lineárnej výpočtovej náročnosti a v jednoduchosti jej algoritmu, ktorého schéma sa nachádza na obrázku 4. Popis ku nej môže byť zhrnutý v nasledovných krokoch:

1. Na úvod sa zvolí K počiatočných centroidov $\{m_1, \ldots, m_K\}$ tvoriacich maticu M, buď celkom náhodne, alebo podľa určitej predošlej znalosti.

2. Metódou najbližšieho suseda priradíme objekty k jednotlivým centroidom a vytvoríme tak K zhlukov. Objekt x_j je zaradený týmto algoritmom do klastru C_i, práve vtedy keď, $\forall l \in \{1, \ldots, K\}, i \neq l$: $\|x_j - m_i\| < \|x_j - m_l\|$.

3. Podľa aktuálneho zadelenia objektov prerátame centroidy zhlukov podľa:

$$m_i = \frac{1}{n_i} \sum_{x_j \in C_i} x_j. \tag{27}$$

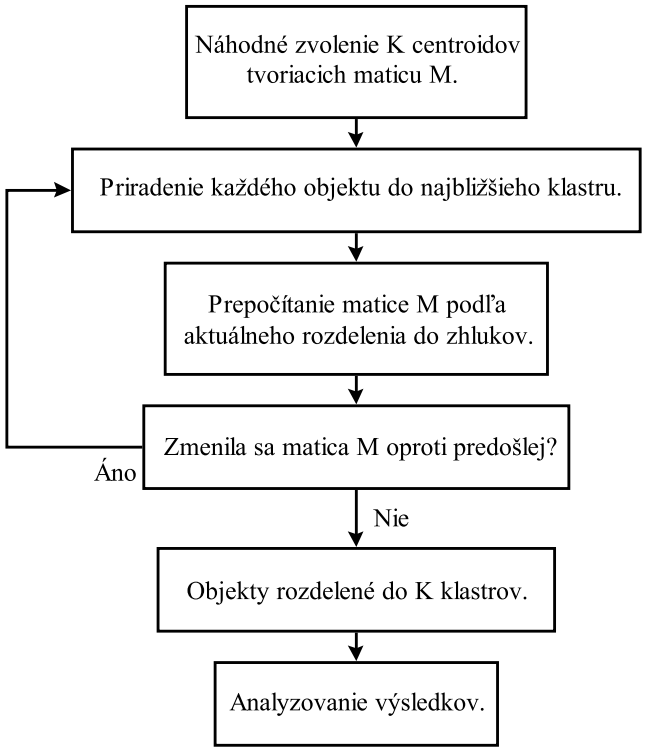
4. Skontrolujeme, či sa matica M zmenila oproti predošlej. Ak áno, je potrebné zopakovať kroky 2 a 3 znova, aby sa centroidná matica ustálila a tým pádom aj zadelenie objektov ku centroidom, ktoré reprezentujú jednotlivé zhluky by bolo definitívne. V prípade, že sa matica M nezmenila, ukončíme algoritmus a vyhodnotíme výsledky zhlukovacieho algoritmu.

1.4.2 Nevýhody a vylepšenia

Výhody K-means metódy sme spomenuli už vyššie. Patria medzi ne nekomplikovanosť algoritmu a výpočtová jednoduchosť. Je teda vhodná aj pre veľkú skupinu dát. Teraz sa pozrieme na nevýhody a na pokusy upraviť túto metódu, aby boli odstránené.

Hlavným problémom je **konvergencia**. Bola síce dokázaná konvergencia danej metódy k extrému, avšak nie ku globálnemu, ale iba k lokálnemu. To znamená, že výsledné optimum môže byť rôzne pre odlišné štartovacie centroidy. Experimentálne bolo dokázané, že náhodná voľba alebo Kaufman-Rousseeuwova metóda výberu počiatocných vektorov je najideálnejšia. Kaufman a Rousseeuw navrhli postupné určovanie centroidov. Prvý m_1 sa nachádza v centre, má najnižšiu sumu vzdialeností so všetkými ob-

Obr. 4: K-means algoritmus



jektami. Ďalší, m_2, sa zvolí tak, aby sme získali, čo najväčší pokles kritériovej funkcie. Ostatné sa postupne určia rovnakým spôsobom ako m_2.

Neskôr bol vyvinutý globálny K-means algoritmus, ktorý nezávisí od počiatočných centroidných bodov. Pozostáva z viacerých K-means procedúr, pričom na začiatku sa určí prvý centroid. Vo zvyšných iteráciach k, $k = \{2, \ldots, K\}$ je predošlých $k-1$ centroidov fixovaných a určí sa podľa pravidiel k-ty počiatočný bod. Keďže tento algoritmus vyžaduje opakovanie K-means metódy N krát pre každé k, vykazuje vysokú výpočtovú náročnosť.

Medzi ďalšie vylepšenia patria *genetický K-means algoritmus (GKA)* a *EBLG algoritmus*, o ktorých sa viac dozviete v [8].

Ako druhý problém spomenieme neurčitosť **voľby čísla K**. V praxi nie je určené, aký počet klastrov je ideálny a neexistujú žiadne presné pravidlá na určenie tohoto neznámeho čísla. Existuje technika ISODATA založená na prahovej konštante ρ, ktorú si však takisto musíme na úvod zvoliť. Podľa ISODATY sa postupne zvyšuje číslo K, vždy keď existuje objekt x_j, ktorý podľa 2. kroku K-means metódy má byť zaradený do klastru C_i, avšak nie je splnená podmienka $\|x_j - m_i\| < \rho$. Teda tento objekt ani

pre najbližší centroid nespĺňa podmienku neprekročenia maximálnej vzdialenosti ρ od centroidu. V momente, kedy pre všetky objekty platí, že ich vzdialenosť od centroidu klastru, do ktorého patria je menšia ako ρ, už nie je potrebné zvyšovať počet zhlukov a teda je určené neznáme číslo K.

1.5 Fuzzy metódy

Hierarchické a nehierarchické metódy sú považované za veľmi neflexibilné, nakoľko každý objekt smie patriť práve do jedného klastru. Uvoľnením tohto obmedzenia získavame fuzzy metódy, ktoré sa vyznačujú hlavne tým, že každý objekt môže patriť do každého zhluku s určitým stupňom členstva. Sú vhodné hlavne pre zhluky, ktorých hranice sú nejasné, lebo nám dovoľujú analyzovať hlbšie vzťahy medzi objektami a určenými klastrami.

1.5.1 Fuzzy c-means

Najznámejšia fuzzy metóda, ktorá vznikla zovšeobecnením mechanizmu ISODATA, sa nazýva Fuzzy c-means(FCM). Jej cieľom je rozdeliť N objektov do c fuzzy zhlukov, tak aby kritériová funkcia

$$J(U, M) = \sum_{i=1}^{c} \sum_{j=1}^{N} (u_{ij})^f \cdot D_{ij}^2 \tag{28}$$

bola minimalizovaná. Zložkami kritériovej funkcie sú:

- Matica priemerov $M = [m_1, m_2, \ldots, m_c]$,
- Parameter fuzifikácie f, ktorého zvyčajná hodnota je $f = 2$, priamo úmerne ovplyvňuje ako veľmi sa jednotlivé zhluky miešajú,
- Funkcia vzdialenosti D_{ij} označuje vzdialenosť objektu x_j od vektoru priemeru m_i, pričom je použitá Euklidovská norma: $D_{ij} = D(x_j, m_i) = \|x_j - m_i\|$,
- Matica $U_{c \times N}$ s prvkami u_{ij} označujúcimi koeficient členstva j-teho objektu v i-tom klastri. Podmienky na koeficienty členstva sú nasledovné:

$$\forall j \in \{1, 2, \ldots, N\} : \sum_{i=1}^{c} u_{ij} = 1, \tag{29}$$

$$\forall i \in \{1, 2, \ldots, c\} : 0 < \sum_{j=1}^{N} u_{ij} < N. \tag{30}$$

Podmienka (29) zabezpečuje rovnakú váhu všetkých objektov a podmienka (30) neprázdnosť všetkých klastrov.

Algoritmus FCM je iteračná procedúra na minimalizáciu kritériovej funkcie (28). Popíšeme ju v piatich krokoch.

1. Nastavíme počítadlo krokov $t = 0$, určíme počet fuzzy zhlukov c, parameter fuzifikácie f a prah presnosti ρ.

2. Zvolíme maticu M buď náhodne alebo podľa určitej vstupnej informácie o objektoch.

3. Vytvoríme alebo aktualizujeme maticu U podľa:

$$u_{ij}^{t+1} = \begin{cases} \frac{1}{\sum_{l=1}^{c} (\frac{D_{lj}}{D_{ij}})^{\frac{2}{1-f}}} & \text{ak } I_j = \emptyset, \\ \frac{1}{|I_j|} & \text{ak } I_j \neq \emptyset \wedge i \in I_j, \\ 0 & \text{ak } I_j \neq \emptyset \wedge i \notin I_j, \end{cases} \tag{31}$$

   kde $I_j = \{i | 1 \leq i \leq c, x_j = m_i\}$, teda indexová množina klastrov, s ktorých priemerom m_i sa zhoduje objekt j v ľubovoľnej črte.

4. Aktualizujeme maticu M:

$$m_i^{t+1} = \frac{\sum_{j=1}^{N} (u_{ij}^{t+1})^f \cdot x_j}{\sum_{j=1}^{N} (u_{ij}^{t+1})^f}. \tag{32}$$

5. Opakujeme kroky 3. a 4. kým Euklidovská veľkosť rozdielu matíc M^ta M^{t+1} je menšia ako zvolená konštanta $\rho : \|M^{t+1} - M^t\| < \rho$.

1.5.2 Nevýhody a vylepšenia

Aj FCM metóda trpí dvomi veľkými nedostatkami. Ako prvý uvedieme **voľbu počiatočnej matice priemerov** M, nakoľko aj tento algoritmus konverguje iba lokálne. Na odstránenie tohoto problému bola navrhnutá iteračná *'Mountain' metóda (MM)*, ktorej cieľom je z kandidujúcich vektorov v_i vybrať vhodnú kombináciu vektorov do matice M. Samozrejme s vyšším počtom kandidujúcich vektorov sa zvyšuje kvalita výberu,

ale aj výpočtova náročnosť. Vhodnosť vektora sa v tejto metóde ohodnocuje pomocou 'Mountain' funkcie:

$$M(v_i) = \sum_{j=1}^{N} e^{-\alpha \cdot D(x_j, v_i)}, \tag{33}$$

kde $\alpha > 0$ a $D(x_j, v_i)$ označuje vzdialenosť j-teho objektu od i-teho vektora. Keďže funkcia je tým väčšia, čím sú jednotlivé objekty k danému vektoru v_i bližšie, v každej iterácii vyberieme vektor v_t^* maximalizujúci funkciu $M^t(v_j)$. Následne aktualizujeme hodnotu Mountain funkcie pre zvyšné vektory, pričom vplyv vybratého vektora eliminujeme nasledovne:

$$M^{t+1}(v_j) = M^t(v_j) - M(v_t^*) \sum_{t=1}^{N} e^{-\beta \cdot D(v_t^*, v_j)}, \tag{34}$$

kde β je konštanta. V publikácii [8] sa dočítate o ďalších možnostiach na zabezpečenie globálnej konvergencie iteračného fuzzy procesu.

Druhým nedostatkom je nízka **citlivosť voči outliers**. Aby vplyv objektov, ktoré sú výrazne odlišné nebol rovnaký ako tých, ktoré sú relatívne podobné, vznikla pravdepodobnostná c-means metóda(PCM), ktorá obmedzenie (29) uvoľní na:

$$\forall j \in \{1, 2, \ldots, N\} : \max_i u_{ij} > 0. \tag{35}$$

PCM interpretuje koeficienty u_{ij} ako zhodnosť objektu j s predstaviteľom klastru i a nie ako koeficient členstva. Navyše podmienka (35) zabezpečuje, že každý objekt patrí aspoň do jedného klastru s nenulovou pravdepodobnosťou a táto pravdepodobnosť nie je ovplyvnená ostatnými objektami. Viac o PCM metóde sa dočítate v [8].

Ako posledná otázka sa vynára nejasnosť **voľby počtu fuzzy zhlukov** c. S týmto problémom sa dá vysporiadať, obdobne ako v prípade nehierarchickej metódy K-means, zvolením maximálnej prahovej vzdialenosti medzi každým objektom a najbližším zhlukovým vektorom priemeru. Podrobnejšie nájdete v sekcii 1.4.2.

2 Ilustračný príklad

V tejto časti názorne aplikujeme niektoré zo spomínaných zhlukových metód na ilustračnom príklade. Použijeme jednu hierarchickú aglomeratívnu, hierarchickú divíznu, nehierarchickú a fuzzy metódu a porovnáme dosiahnuté výsledky.

Príklad: (podľa [1])
Uvažujme 10 slovenských medicínskych, farmaceutických, zdravotníckych alebo ošetrovateľských fakúlt na Slovensku. Celkový posudok týchto fakúlt je ovplyvnený piatimi faktormi:

1. Kvalita vzdelania,
2. Atraktivita štúdia,
3. Veda a výskum,
4. Doktorandské štúdium,
5. Úspešnosť získavania grantov.

Počty bodov v tabuľke 3 odrážajú intenzitu výkonu, nakoľko sa zohľadňuje aj veľkosť fakulty.

Tabuľka 3: Lekárske fakulty na Slovensku

Číslo	fakulta	Vzdelanie	Atraktivita	Veda	Doktorandi	Granty
1	Jesseniova Lekárska Fakulta UK	85	84	65	53	71
2	Lekárska Fakulta UPJŠ	71	68	49	42	86
3	Farmaceutická Fakulta UK	62	66	74	47	55
4	Lekárska Fakulta UK	81	72	58	45	31
5	Fakulta Zdravotníctva a sociál. práce TU	68	33	29	47	24
6	Fakulta Sociál. Vied a zdravotníctva UKF	53	35	65	36	10
7	Fakulta Zdravotníctva KU	64	44	14	42	4
8	VŠ Zdravotníctva a soc. práce sv. Alžbety	38	45	10	27	0
9	Fakulta Zdravotníctva TUAD	41	37	41	0	0
10	Fakulta Zdravotníckych Odborov PU	42	48	8	0	19

Čo presne jednotlivé kritéria zohľadňujú a aj ohodnotenia fakúlt v iných odvetviach je uvedené v [1].

2.1 Hierachická aglomeratívna metóda

Na meranie vzdialenosti si zvolíme klasickú Euklidovskú normu a z rôznych aglomeratívnych metód si vyberieme priemerovú. Teraz krok po kroku vytvoríme hierarchiu zhlukov predstavujúcich slovenské lekárske fakulty.

1. Každá fakulta tvorí samostatný zhluk. Vytvoríme maticu vzdialenosti D:

$$\forall i \in \{1, \ldots, 10\} : D_{i,i} = 0,$$
$$\forall i, j \in \{1, \ldots, 10\}, i \neq j : D_{i,j} = D(x_i, x_j) = \left(\sum_{l=1}^{5} |x_{il} - x_{jl}|^2 \right)^{\frac{1}{2}}.$$

Napríklad

$$\begin{aligned} D_{1,2} &= \sqrt{|85-71|^2 + |84-68|^2 + |65-49|^2 + |53-42|^2 + |71-86|^2} \\ &= \sqrt{1054} \doteq 32.465 \end{aligned}$$

Podobne vyrátame zvyšok matice D a dostávame maticu D^0:

$$\begin{pmatrix} C_1 & C_2 & C_3 & C_4 & C_5 & C_6 & C_7 & C_8 & C_9 & C_{10} \\ 0 & 32.4654 & 35.0143 & 43.2782 & 80.1935 & 86.2264 & 96.1873 & 111.6781 & 112.1205 & 109.1192 \\ 32.4654 & 0 & 41.1825 & 56.8419 & 74.1822 & 86.4928 & 92.5959 & 103.7304 & 105.2853 & 95.7862 \\ 35.0143 & 41.1825 & 0 & 35.1141 & 64.1171 & 57.1752 & 81.9390 & 92.4013 & 87.2067 & 92.6553 \\ 43.2782 & 56.8419 & 35.1141 & 0 & 50.8331 & 52.1920 & 61.2127 & 78.5302 & 78.1025 & 82.2557 \\ 80.1935 & 74.1822 & 64.1171 & 50.8331 & 0 & 42.9185 & \mathbf{28.0535} & 48.7955 & 60.6135 & 59.7997 \\ 86.2264 & 86.4928 & 57.1752 & 52.1920 & 42.9185 & 0 & 53.6190 & 59.4222 & 46.0435 & 70.1142 \\ 96.1873 & 92.5959 & 81.9390 & 61.2127 & \mathbf{28.0535} & 53.6190 & 0 & 30.5614 & 55.5608 & 50.2494 \\ 111.6781 & 103.7304 & 92.4013 & 78.5302 & 48.7955 & 59.4222 & 30.5614 & 0 & 41.9881 & 33.4515 \\ 112.1205 & 105.2853 & 87.2067 & 78.1025 & 60.6135 & 46.0435 & 55.5608 & 41.9881 & 0 & 39.6485 \\ 109.1192 & 95.7862 & 92.6553 & 82.2557 & 59.7997 & 70.1142 & 50.2494 & 33.4515 & 39.6485 & 0 \end{pmatrix}.$$

2. Nájdeme najnižšiu nenulovú hodnotu v matici D, tou je 28.0535. Indikuje nám, že zhluky 5 a 7 sú k sebe najbližšie, a preto ich zlúčime do jedného nového zhluku.

3. Zvolili sme si priemerovú metódu, ktorej všeobecný tvar pre vzdialenosť medzi novým zhlukom $C_{i,j}$ a ľubovoľným iným C_l je:

$$D(C_l, C_{i,j}) = \frac{n_i}{n_i + n_j} \cdot D(C_l, C_i) + \frac{n_j}{n_i + n_j} \cdot D(C_l, C_j).$$

V tejto iterácii platí $i = 5$, $j = 7$, $n_5 = 1$ a $n_7 = 1$, a teda

$$D(C_l, C_{5,7}) = \frac{1}{2} \cdot D(C_l, C_5) + \frac{1}{2} \cdot D(C_l, C_7).$$

Konkrétne

$$D(C_1, C_{5,7}) = \frac{1}{2} \cdot 80.1935 + \frac{1}{2} \cdot 96.1873 = 88.1904.$$

Obdobne vyrátame zvyšné nové vzdialenosti, siedmy riadok aj stĺpec matice D vymažeme a do piateho riadku a stĺpca zapíšeme aktualizované vzdialenosti nového zhluku. Zároveň uložíme do pamäti $n_5 = 2$.

$$D^1 = \begin{pmatrix} C_1 & C_2 & C_3 & C_4 & C_{5,7} & C_6 & C_8 & C_9 & C_{10} \\ 0 & \mathbf{32.4654} & 35.0143 & 43.2782 & 88.1904 & 86.2264 & 111.6781 & 112.1205 & 109.1192 \\ \mathbf{32.4654} & 0 & 41.1825 & 56.8419 & 83.3891 & 86.4928 & 103.7304 & 105.2853 & 95.7862 \\ 35.0143 & 41.1825 & 0 & 35.1141 & 73.0280 & 57.1752 & 92.4013 & 87.2067 & 92.6553 \\ 43.2782 & 56.8419 & 35.1141 & 0 & 56.0229 & 52.1920 & 78.5302 & 78.1025 & 82.2557 \\ 88.1904 & 83.3891 & 73.0280 & 56.0229 & 0 & 48.2688 & 39.6785 & 58.0872 & 55.0245 \\ 86.2264 & 86.4928 & 57.1752 & 52.1920 & 48.2688 & 0 & 59.4222 & 46.0435 & 70.1142 \\ 111.6781 & 103.7304 & 92.4013 & 78.5302 & 39.6785 & 59.4222 & 0 & 41.9881 & 33.4515 \\ 112.1205 & 105.2853 & 87.2067 & 78.1025 & 58.0872 & 46.0435 & 41.9881 & 0 & 39.6485 \\ 109.1192 & 95.7862 & 92.6553 & 82.2557 & 55.0245 & 70.1142 & 33.4515 & 39.6485 & 0 \end{pmatrix}$$

4. Overíme, či už existuje iba jeden zhluk, teda či matica D má rozmery 1×1. Ak nie opakujeme kroky 2 a 3. Najmenšiu nenulovú vzdialenosť zhlukov označíme priamo v matici D a potom ju aktualizujeme:

$$D^2 = \begin{pmatrix} C_{1,2} & C_3 & C_4 & C_{5,7} & C_6 & C_8 & C_9 & C_{10} \\ 0 & 38.0984 & 50.0600 & 85.7897 & 86.3596 & 107.7043 & 108.7029 & 102.4527 \\ 38.0984 & 0 & 35.1141 & 73.0280 & 57.1752 & 92.4013 & 87.2067 & 92.6553 \\ 50.0600 & 35.1141 & 0 & 56.0229 & 52.1920 & 78.5302 & 78.1025 & 82.2557 \\ 85.7897 & 73.0280 & 56.0229 & 0 & 48.2688 & 39.6785 & 58.0872 & 55.0245 \\ 86.3596 & 57.1752 & 52.1920 & 48.2688 & 0 & 59.4222 & 46.0435 & 70.1142 \\ 107.7043 & 92.4013 & 78.5302 & 39.6785 & 59.4222 & 0 & 41.9881 & \mathbf{33.4515} \\ 108.7029 & 87.2067 & 78.1025 & 58.0872 & 46.0435 & 41.9881 & 0 & 39.6485 \\ 102.4527 & 92.6553 & 82.2557 & 55.0245 & 70.1142 & \mathbf{33.4515} & 39.6485 & 0 \end{pmatrix}$$

$$D^3 = \begin{pmatrix} C_{1,2} & C_3 & C_4 & C_{5,7} & C_6 & C_{8,10} & C_9 \\ 0 & 38.0984 & 50.0600 & 85.7897 & 86.3596 & 105.0785 & 108.7029 \\ 38.0984 & 0 & \mathbf{35.1141} & 73.0280 & 57.1752 & 92.5283 & 87.2067 \\ 50.0600 & \mathbf{35.1141} & 0 & 56.0229 & 52.1920 & 80.3930 & 78.1025 \\ 85.7897 & 73.0280 & 56.0229 & 0 & 48.2688 & 47.3515 & 58.0872 \\ 86.3596 & 57.1752 & 52.1920 & 48.2688 & 0 & 64.7682 & 46.0435 \\ 105.0785 & 92.5283 & 80.3930 & 47.3515 & 64.7682 & 0 & 40.8183 \\ 108.7029 & 87.2067 & 78.1025 & 58.0872 & 46.0435 & 40.8183 & 0 \end{pmatrix}$$

$$D^4 = \begin{pmatrix} C_{1,2} & C_{3,4} & C_{5,7} & C_6 & C_{8,10} & C_9 \\ 0 & 44.0792 & 85.7897 & 86.3596 & 105.0785 & 108.7029 \\ 44.0792 & 0 & 64.5255 & 54.6836 & 86.4606 & 82.6546 \\ 85.7897 & 64.5255 & 0 & 48.2688 & 47.3515 & 58.0872 \\ 86.3596 & 54.6836 & 48.2688 & 0 & 64.7682 & 46.0435 \\ 105.0785 & 86.4606 & 47.3515 & 64.7682 & 0 & \mathbf{40.8183} \\ 108.7029 & 82.6546 & 58.0872 & 46.0435 & \mathbf{40.8183} & 0 \end{pmatrix}$$

$$D^5 = \begin{pmatrix} C_{1,2} & C_{3,4} & C_{5,7} & C_6 & C_{8,9,10} \\ 0 & \mathbf{44.0792} & 85.7897 & 86.3596 & 106.2866 \\ \mathbf{44.0792} & 0 & 64.5255 & 54.6836 & 85.1919 \\ 85.7897 & 64.5255 & 0 & 48.2688 & 50.9300 \\ 86.3596 & 54.6836 & 48.2688 & 0 & 58.5266 \\ 106.2866 & 85.1919 & 50.9300 & 58.5266 & 0 \end{pmatrix}$$

$$D^6 = \begin{pmatrix} C_{1,2,3,4} & C_{5,7} & C_6 & C_{8,9,10} \\ 0 & 75.1576 & 70.5216 & 95.7393 \\ 75.1576 & 0 & \mathbf{48.2688} & 50.9300 \\ 70.5216 & \mathbf{48.2688} & 0 & 58.5266 \\ 95.7393 & 50.9300 & 58.5266 & 0 \end{pmatrix}$$

$$D^7 = \begin{pmatrix} C_{1,2,3,4} & C_{5,6,7} & C_{8,9,10} \\ 0 & 73.6123 & 95.7393 \\ 73.6123 & 0 & \mathbf{53.4622} \\ 95.7393 & \mathbf{53.4622} & 0 \end{pmatrix}$$

$$D^8 = \begin{pmatrix} C_{1,2,3,4} & C_{5,6,7,8,9,10} \\ 0 & \mathbf{84.6758} \\ \mathbf{84.6758} & 0 \end{pmatrix}$$

$$D^9 = \begin{pmatrix} C_{1,2,3,4,5,6,7,8,9,10} \\ 0 \end{pmatrix}$$

5. Po deviatej iterácii sa zhlučili všetky objekty do jedného klastru. Vykreslíme si príslušný dendrogram (obrázok 5). Preložme horizontálnu čiaru vo výške 51. Vzniknú nám práve tri zhluky:

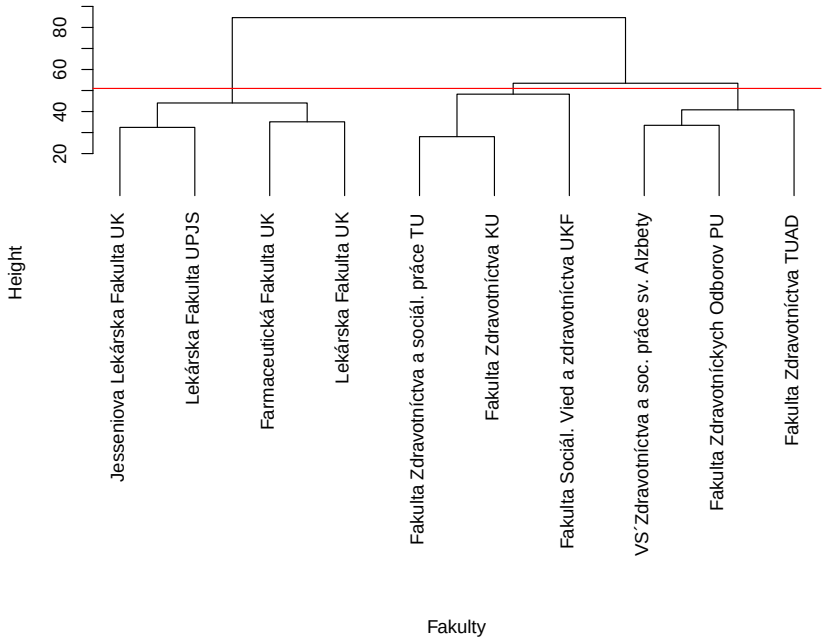
1. Jesseniova Lekárska Fakulta UK, Lekárska Fakulta UPJŠ, Farmaceutická Fakulta UK, Lekárska Fakulta UK,

2. Fakulta Zdravotníctva a sociál. Práce TU, Fakulta Sociál. Vied a zdravotníctva UKF, Fakulta Zdravotníctva KU,

3. VŠ Zdravotníctva a soc. Práce sv. Alžbety, Fakulta Zdravotníctva TUAD, Fakulta Zdravotníckych Odborov PU.

Stručne môžme skonštatovať, že v prvej skupine skončili Lekárske a farmaceutické fakulty, ktorých hodnotenia v takmer všetkých kritériach boli nadpriemerné. Ďalšie dva klastre tvoria zdravotnícke fakulty a vzájomne sa podľa dát líšia hlavne v úrovni vzdelania, grantovej úspešnosti a doktorandskom štúdiu.

Obr. 5: Aglomeratívna priemerová metóda - ilustračný príklad



2.2 Hierarchická divízna metóda

Na začiatku máme všetky objekty v jednom klastri C_1. Postupne ich iteračným systémom DIANA rozdelíme na 10 klastrov po jednom objekte. DIANA pracuje na základe vzdialeností objektov v jednom klastri, pričom vzdialenosti určíme pomocou Euklidovskej normy. Na úvod vytvoríme maticu vzájomných vzdialeností D_1^0 všetkých objektov:

$$\begin{pmatrix}
0 & 32.4654 & 35.0143 & 43.2782 & 80.1935 & 86.2264 & 96.1873 & 111.6781 & \mathbf{112.1205} & 109.1192 \\
32.4654 & 0 & 41.1825 & 56.8419 & 74.1822 & 86.4928 & 92.5959 & 103.7304 & 105.2853 & 95.7862 \\
35.0143 & 41.1825 & 0 & 35.1141 & 64.1171 & 57.1752 & 81.9390 & 92.4013 & 87.2067 & 92.6553 \\
43.2782 & 56.8419 & 35.1141 & 0 & 50.8331 & 52.1920 & 61.2127 & 78.5302 & 78.1025 & 82.2557 \\
80.1935 & 74.1822 & 64.1171 & 50.8331 & 0 & 42.9185 & 28.0535 & 48.7955 & 60.6135 & 59.7997 \\
86.2264 & 86.4928 & 57.1752 & 52.1920 & 42.9185 & 0 & 53.6190 & 59.4222 & 46.0435 & 70.1142 \\
96.1873 & 92.5959 & 81.9390 & 61.2127 & 28.0535 & 53.6190 & 0 & 30.5614 & 55.5608 & 50.2494 \\
111.6781 & 103.7304 & 92.4013 & 78.5302 & 48.7955 & 59.4222 & 30.5614 & 0 & 41.9881 & 33.4515 \\
\mathbf{112.1205} & 105.2853 & 87.2067 & 78.1025 & 60.6135 & 46.0435 & 55.5608 & 41.9881 & 0 & 39.6485 \\
109.1192 & 95.7862 & 92.6553 & 82.2557 & 59.7997 & 70.1142 & 50.2494 & 33.4515 & 39.6485 & 0
\end{pmatrix}.$$

DIANA:

1. V každej iterácii zvolíme zhluk, ktorý obsahuje najvzdialenejší pár objektov, v tomto prípade sú to objekty 1 a 9 v zhluku C_1.

2. Pre všetky objekty vo vybratom zhluku vyrátame priemernú vzdialenosť od zvyš-

ných objektov v tom istom zhluku podľa (24):

$$\begin{pmatrix} d(x_1, C_1\backslash\{x_1\}) \\ d(x_2, C_1\backslash\{x_2\}) \\ d(x_3, C_1\backslash\{x_3\}) \\ d(x_4, C_1\backslash\{x_4\}) \\ d(x_5, C_1\backslash\{x_5\}) \\ d(x_6, C_1\backslash\{x_6\}) \\ d(x_7, C_1\backslash\{x_7\}) \\ d(x_8, C_1\backslash\{x_8\}) \\ d(x_9, C_1\backslash\{x_9\}) \\ d(x_{10}, C_1\backslash\{x_{10}\}) \end{pmatrix} = \begin{pmatrix} \mathbf{78.4759} \\ 76.5070 \\ 65.2006 \\ 59.8178 \\ 56.6118 \\ 61.5782 \\ 61.1088 \\ 66.7288 \\ 69.6188 \\ 70.3422 \end{pmatrix}.$$

3. Objekt s maximálnou priemernou vzdialenosťou od ostatných presunieme do C_2. V tomto príklade ide o objekt 1.

4. Po presunutí jedného objektu do nového klastru, zrátame pre každý zvyšný objekt v pôvodnom zhluku rozdiel medzi priemernou vzdialenosťou s objektami v pôvodnom a v novom zhluku podľa (25):

$$\begin{pmatrix} d(x_2, C_1\backslash\{x_2\}) - d(x_2, C_2) \\ d(x_3, C_1\backslash\{x_3\}) - d(x_3, C_2) \\ d(x_4, C_1\backslash\{x_4\}) - d(x_4, C_2) \\ d(x_5, C_1\backslash\{x_5\}) - d(x_5, C_2) \\ d(x_6, C_1\backslash\{x_6\}) - d(x_6, C_2) \\ d(x_7, C_1\backslash\{x_7\}) - d(x_7, C_2) \\ d(x_8, C_1\backslash\{x_8\}) - d(x_8, C_2) \\ d(x_9, C_1\backslash\{x_9\}) - d(x_9, C_2) \\ d(x_{10}, C_1\backslash\{x_{10}\}) - d(x_{10}, C_2) \end{pmatrix} = \begin{pmatrix} \mathbf{49.5468} \\ 33.9596 \\ 18.6071 \\ -26.5294 \\ -27.7293 \\ -39.4633 \\ -50.5680 \\ -47.8144 \\ -43.6242 \end{pmatrix}.$$

5. Maximálny a kladný rozdiel vzdialeností má druhý objekt, preto ho presunieme do klastru C_2. Teraz zopakujeme kroky 4 a 5 kým vektor rozdielov priemerných vzdialeností nebude obsahovať iba záporné hodnoty, čo značí, že tieto zvyšné objekty sú priemerne bližšie k objektom v pôvodnom zhluku ako k objektom v novom zhluku.

$$\begin{pmatrix} d(x_3, C_1\backslash\{x_3\}) - d(x_3, C_2) \\ d(x_4, C_1\backslash\{x_4\}) - d(x_4, C_2) \\ d(x_5, C_1\backslash\{x_5\}) - d(x_5, C_2) \\ d(x_6, C_1\backslash\{x_6\}) - d(x_6, C_2) \\ d(x_7, C_1\backslash\{x_7\}) - d(x_7, C_2) \\ d(x_8, C_1\backslash\{x_8\}) - d(x_8, C_2) \\ d(x_9, C_1\backslash\{x_9\}) - d(x_9, C_2) \\ d(x_{10}, C_1\backslash\{x_{10}\}) - d(x_{10}, C_2) \end{pmatrix} = \begin{pmatrix} \mathbf{34.8457} \\ 12.5457 \\ -26.4549 \\ -31.8618 \\ -42.7922 \\ -52.6828 \\ -50.2510 \\ -41.2850 \end{pmatrix}$$

Objekt 3 presunieme do zhluku C_2.

$$\begin{pmatrix} d(x_4, C_1\backslash\{x_4\}) - d(x_4, C_2) \\ d(x_5, C_1\backslash\{x_5\}) - d(x_5, C_2) \\ d(x_6, C_1\backslash\{x_6\}) - d(x_6, C_2) \\ d(x_7, C_1\backslash\{x_7\}) - d(x_7, C_2) \\ d(x_8, C_1\backslash\{x_8\}) - d(x_8, C_2) \\ d(x_9, C_1\backslash\{x_9\}) - d(x_9, C_2) \\ d(x_{10}, C_1\backslash\{x_{10}\}) - d(x_{10}, C_2) \end{pmatrix} = \begin{pmatrix} \mathbf{22.1096} \\ -24.3286 \\ -22.5799 \\ -43.6979 \\ -53.8118 \\ -47.8780 \\ -43.2671 \end{pmatrix}$$

Objekt 4 presunieme do C_2.

$$\begin{pmatrix} d(x_5, C_1\backslash\{x_5\}) - d(x_5, C_2) \\ d(x_6, C_1\backslash\{x_6\}) - d(x_6, C_2) \\ d(x_7, C_1\backslash\{x_7\}) - d(x_7, C_2) \\ d(x_8, C_1\backslash\{x_8\}) - d(x_8, C_2) \\ d(x_9, C_1\backslash\{x_9\}) - d(x_9, C_2) \\ d(x_{10}, C_1\backslash\{x_{10}\}) - d(x_{10}, C_2) \end{pmatrix} = \begin{pmatrix} -19.2953 \\ -16.0981 \\ -39.3749 \\ -53.7413 \\ -46.9079 \\ -44.3015 \end{pmatrix}$$

Nakoľko sú všetky hodnoty záporné, ukončíme túto iteráciu. Zrekapitulujeme si rozdelenie objektov, v C_1 sú objekty $5, 6, 7, 8, 9$ a 10, v C_2 objekty $1, 2, 3$ a 4.

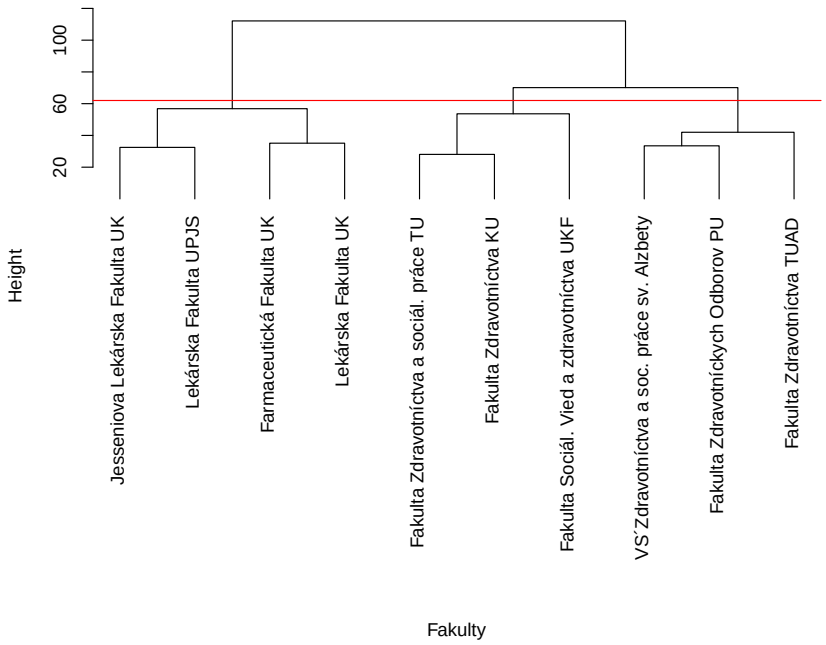
Pre druhú iteráciu si vytvoríme maticu vzdialenosti pre obidva zhluky a zvýrazníme najväčšiu vzdialenosť.

$$D_1^1 = \begin{pmatrix} 0 & 42.9185 & 28.0535 & 48.7955 & 60.6135 & 59.7997 \\ 42.9185 & 0 & 53.6190 & 59.4222 & 46.0435 & \mathbf{70.1142} \\ 28.0535 & 53.6190 & 0 & 30.5614 & 55.5608 & 50.2494 \\ 48.7955 & 59.4222 & 30.5614 & 0 & 41.9881 & 33.4515 \\ 60.6135 & 46.0435 & 55.5608 & 41.9881 & 0 & 39.6485 \\ 59.7997 & \mathbf{70.1142} & 50.2494 & 33.4515 & 39.6485 & 0 \end{pmatrix}$$

$$D_2^1 = \begin{pmatrix} 0 & 32.4654 & 35.0143 & 43.2782 \\ 32.4654 & 0 & 41.1825 & 56.8419 \\ 35.0143 & 41.1825 & 0 & 35.1141 \\ 43.2782 & 56.8419 & 35.1141 & 0 \end{pmatrix}$$

Budeme teda deliť prvý zhluk. Rovnakým iteračným systémom by sme sa dopracovali do situácie desiatich zhlukov po jednom objekte. Kvôli zdĺhavosti postupu, uvádzame iba konečný dendrogram na obrázku 6.

Obr. 6: Divízna metóda - ilustračný príklad



Horizontálnu čiaru preložíme tak, aby sme dostali práve 3 zhluky, aby sme mohli dosiahnutý výsledok porovnať s prvou metódou. Divízna metóda rozdelila fakulty do nasledujúcich troch klastrov:

1. Jesseniova Lekárska Fakulta UK, Lekárska Fakulta UPJŠ, Farmaceutická Fakulta UK, Lekárska Fakulta UK,

2. Fakulta Zdravotníctva a sociál. Práce TU, Fakulta Sociál. Vied a zdravotníctva UKF, Fakulta Zdravotníctva KU,

3. VŠ Zdravotníctva a soc. Práce sv. Alžbety, Fakulta Zdravotníctva TUAD, Fakulta Zdravotníckych Odborov PU,

čo sú presne tie isté ako nám určila aglomeratívna metóda. Toto nie je pravidlo, skôr výnimka. Pri viacrozmerných dátach sa zvykne kumulovať chyba, a preto aglomeratívna metóda idúca "zdola"nemusí dať to isté rozdelenie ako divízna metóda idúca "zhora".

2.3 Nehierarchická metóda

Objekty teraz rozdelíme tak, aby totálna suma štvorcových chýb bola minimalizovaná. Na to využijeme iteračný algoritmus K-means, ktorý zadeľuje objekty do klastrov, reprezentovaných priemerovými vektormi, až kým sa rozdelenie neustáli. Počet zhlukov určíme opäť tri ako v predošlých dvoch metódach, aby sa výsledky dali porovnať.

1. Zvolíme si počiatočné priemerové vektory. Keďže každý objekt disponuje piatimi črtami, tak aj tieto tri priemerové vektory budú mať rozmer 5×1. Z nich vytvoríme následne maticu $M = [m_1, m_2, m_3]_{5\times 3}$. Keďže vieme, že hodnoty jednotlivých čŕt sa pohybujú v rozmedzí 0 až 100, zvolíme náhodné vektory v tomto rozsahu: $m_1 = (53, 38, 80, 67, 98)^\top$, $m_2 = (93, 57, 8, 41, 18)^\top$, $m_3 = (99, 52, 88, 64, 46)^\top$ a teda matica má podobu:

$$M = \begin{pmatrix} 53 & 93 & 99 \\ 38 & 57 & 52 \\ 80 & 8 & 88 \\ 67 & 41 & 64 \\ 98 & 18 & 46 \end{pmatrix}. \tag{36}$$

2. Vypočítame Euklidovskú vzdialenosť každého objektu od každého vektoru priemeru.

$$\begin{aligned} \|x_1 - m_1\| &= \sqrt{(85-53)^2 + (84-38)^2 + (65-80)^2 + (53-67)^2 + (71-98)^2} \\ &= 65.4981 \end{aligned}$$

Zvyšné vzdialenosti sa nachádzajú v tabuľke 4.

Teraz každý objekt začleníme do toho zhluku, ku ktorému je najbližšie: $C_1 = \{x_2\}$, $C_2 = \{x_5, x_7, x_8, x_9, x_{10}\}$, $C_3 = \{x_1, x_3, x_4, x_6\}$.

3. Na základe momentálneho zaradenie objektov zaktualizujeme maticu M podľa (27):

$$M = \begin{pmatrix} 71 & 50.6 & 70.25 \\ 68 & 41.4 & 64.25 \\ 49 & 20.4 & 65.5 \\ 42 & 23.2 & 45.25 \\ 86 & 9.4 & 41.75 \end{pmatrix}.$$

Tabuľka 4: 1.iterácia K-means - ilustračný príklad

fakulta/priemer	m_1	m_2	m_3
1	65.4981	83.6361	**49.9500**
2	**54.3507**	83.1324	68.1542
3	56.1249	82.4803	**46.1628**
4	86.0058	55.2630	**47.0106**
5	93.4184	**41.4005**	74.6726
6	94.5463	73.6342	**70.8096**
7	118.2117	**35.2562**	94.9368
8	127.9766	**60.7701**	115.4946
9	125.5349	**78.7274**	109.5901
10	127.0236	**66.0606**	120.3744

4. Keďže sa matica M oproti predošlej zmenila, vrátime sa na krok 2 a 3.

Vyrátame vzdialenosti objektov od nových vektorov priemeru. Výsledky sa nachádzajú v tabuľke 5.

Tabuľka 5: 2.iterácia K-means - ilustračný príklad

fakulta/priemer	m_1	m_2	m_3
1	**32.4654**	98.3355	39.0320
2	**0**	90.3476	47.4921
3	41.1825	79.0815	**17.9444**
4	56.8419	64.9313	**18.6414**
5	74.1822	**35.0268**	51.3030
6	86.4928	**46.9050**	47.4025
7	92.5959	**24.6957**	67.3573
8	103.7304	**19.5622**	81.0370
9	105.2853	**34.0952**	77.3886
10	95.7862	**30.0280**	83.2676

Aktuálne rozdelenie objektov je nasledovné: $C_1 = \{x_1, x_2\}, C_2 = \{x_5, x_6, x_7, x_8, x_9, x_{10}\}$, $C_3 = \{x_3, x_4\}$.

Matica M po tomto zadelení má nasledujúci tvar:

$$M = \begin{pmatrix} 78 & 51 & 71.5 \\ 76 & 40.3333 & 69 \\ 57 & 27.8333 & 66 \\ 47.5 & 25.3333 & 46 \\ 78.5 & 9.5 & 43 \end{pmatrix}.$$

Pretože sa zmenilo rozdelenie objektov, tak aj matica M sa zmenila. Musíme teda vykonať tretiu iteráciu.

Nové vzdialenosti objektov od priemerových vektorov sa nachádzajú v tabuľke 6.

Tabuľka 6: 3.iterácia K-means - ilustračný príklad

fakulta/priemer	m_1	m_2	m_3
1	**16.2327**	94.8253	35.2314
2	**16.2327**	87.9972	46.4247
3	34.6049	73.8298	**17.5570**
4	47.8383	60.5131	**17.5570**
5	75.5215	**31.9974**	55.1294
6	84.8204	**39.0875**	51.8483
7	93.0027	**26.1119**	70.1587
8	106.5481	**24.5323**	83.9300
9	107.5384	**31.8826**	80.8965
10	101.3780	**35.5692**	85.8327

Vidíme, že objekty ostali po tretej iterácii v tých istých zhlukoch ako po druhej. Matica M sa teda nezmení. Týmto končí iteračná schéma K-means a dáva nám iné rozdelenie ako predošlé dve metódy a to:

1. Jesseniova Lekárska Fakulta UK, Lekárska Fakulta UPJŠ,

2. Farmaceutická Fakulta UK, Lekárska Fakulta UK,

3. Fakulta Zdravotníctva a sociál. Práce TU, Fakulta Sociál. Vied a zdravotníctva UKF, Fakulta Zdravotníctva KU, VŠ Zdravotníctva a soc. Práce sv. Alžbety, Fakulta Zdravotníctva TUAD, Fakulta Zdravotníckych Odborov PU.

Prvé dve skupiny doteraz tvorili jeden zhluk, K-means metóda ich však rozdelila do dvoch skupín, ktoré sa príznačne líšia iba v grantovej úspešnosti. Na druhej strane tretí

zhluk je tvorený až šiestimi fakultami, ktoré sa síce od seba v niektorých črtách ako veda a výskum, doktorandské štúdium či grantová úspešnosť dosť líšia, avšak sú ešte viac odlišné od objektov v pvých dvoch zhlukoch.

Ako sme upozornili v sekcii 1.4.2, tento algoritmus môže konvergovať ku rôznym lokálnym extrémom, v závislosti od počiatočnej matice M. Napríklad ak by sme uvažovali počiatocné priemerové vektory $m_1 = (55, 38, 60, 7, 15)^\top$, $m_2 = (85, 76, 90, 56, 72)^\top$, $m_3 = (70, 63, 35, 49, 43)^\top$, tak by metóda konvergovala do stavu:

1. Jesseniova Lekárska Fakulta UK, Lekárska Fakulta UPJŠ, Farmaceutická Fakulta UK,

2. Lekárska Fakulta UK, Fakulta Zdravotníctva a sociál. Práce TU, Fakulta Sociál. Vied a zdravotníctva UKF,

3. Fakulta Zdravotníctva KU, VŠ Zdravotníctva a soc. Práce sv. Alžbety, Fakulta Zdravotníctva TUAD, Fakulta Zdravotníckych Odborov PU.

Obe tieto rozdelenia majú totálnu sumu štvorcových chýb rovnakú, 22657.1, a teda toto kritérium považuje obe rozdelenia za rovnako správne. Pre inú počatočnú maticu môže teoreticky vzniknúť zasa iné zadelenie objektov do zhlukov. Ak by práve jedno z riešení malo najmenšiu sumu štvorcových chýb, zvolili by sme ho za najlepšie riešenie. Ak je však takých viac, nezískame jednoznačné riešenie. Takejto nejednoznačnosti sa vieme vyhnúť, ak si pred analýzou zistíme viac informácii o dátach a vhodne určíme počiatočnú maticu M.

2.4 Fuzzy metóda

Zvolíme si spomenutý FCM algoritmus, ktorého cieľom je minimalizácia kritériovej funkcie (28). Táto iteračná metóda postupne mení koeficienty členstva v matici U a následne priemerové vektory v matici M. Iteruje až kým Euklidovská maticová norma rozdielu matíc M v dvoch po sebe idúcich iteráciach je menšia ako stanovený prah presnosti ρ, čo naznačuje približne stabilný stav.

1. Stanovíme počítadlo iterácii $t = 0$, počet fuzzy klastrov $c = 3$, aby sa výsledky opäť dobre porovnávali. Ďalej pre parameter fuzifikácie zvolíme štandardnú hodnotu

$f = 2$ a prah presnosti nech je $\rho = 1$. Čím je ρ menšie, tým sú výsledky presnejšie, ale zvyšuje to aj počet potrebných iterácii.

2. Matica M^0 nech je zhodná so vstupnou maticou M do K-means algoritmu, teda (36), aby sme mohli porovnať konvergenciu týchto metód pre rovnaké vstupné hodnoty.

3. Na vytvorenie matice U si najprv určíme indexové množiny každého objektu $I_j = \{i | 1 \leq i \leq c, x_j = m_i\}$.

$$I^0 = \begin{pmatrix} \emptyset & \emptyset & \emptyset & \emptyset & \emptyset & \{1\} & \emptyset & \emptyset & \emptyset & \{2\} \end{pmatrix}^\top$$

Jednotlivé vzdialenosti objektov od vektorových priemerov sa nachádzajú v tabuľke 4. Teraz podľa vzorcu (31) vyrátame koeficienty členstva:

$$u_{11}^1 = \frac{1}{\sum_{l=1}^{3} (\frac{D_{l1}}{D11})^{\frac{2}{1-2}}} = \frac{1}{(\frac{65.4981}{65.4981})^{-2} + (\frac{83.6361}{65.4981})^{-2} + (\frac{49.9500}{65.4981})^{-2}} = 0.3001$$

Ostatné koeficienty členstva sú:

$$U^1 = \begin{pmatrix} 0.3001 & 0.4846 & 0.3400 & 0.1477 & 0.1306 & 1.0000 & 0.0725 & 0.1501 & 0.2060 & 0 \\ 0.1840 & 0.2072 & 0.1574 & 0.3578 & 0.6650 & 0 & 0.8151 & 0.6656 & 0.5237 & 1.0000 \\ 0.5159 & 0.3082 & 0.5026 & 0.4945 & 0.2044 & 0 & 0.1124 & 0.1843 & 0.2703 & 0 \end{pmatrix}$$

4. Na základe aktuálneho fuzzy začlenenie objektov aktualizujeme maticu M podľa vzorca (32).

$$m_1^1 = \frac{\sum_{j=1}^{N} (u_{1j}^1)^2 \cdot x_j}{\sum_{j=1}^{N} (u_{1j}^1)^2} = \frac{0.3001^2 \cdot \begin{pmatrix} 85 \\ 84 \\ 65 \\ 53 \\ 71 \end{pmatrix} + \cdots + 0^2 \cdot \begin{pmatrix} 42 \\ 48 \\ 8 \\ 0 \\ 19 \end{pmatrix}}{0.3001^2 + 0.4846^2 + \cdots + 0.2060^2 + 0^2} = \begin{pmatrix} 58.3090 \\ 45.8911 \\ 61.1217 \\ 37.8694 \\ 28.4308 \end{pmatrix}$$

Spojením stĺpcových vektorov m_1^1, m_2^1, m_3^1 dostaneme:

$$M^1 = \begin{pmatrix} 58.3090 & 52.5637 & 71.3649 \\ 45.8911 & 45.3667 & 67.9218 \\ 61.1217 & 19.4416 & 58.4020 \\ 37.8694 & 23.3085 & 43.5256 \\ 28.4308 & 14.3104 & 48.6356 \end{pmatrix}.$$

5. Vyrátame

$$\|M^1 - M^0\| = \left\| \begin{pmatrix} 5.3090 & -40.4363 & -27.6351 \\ 7.8911 & -11.6333 & 15.9218 \\ -18.8783 & 11.4416 & -29.5980 \\ -29.1306 & -17.6915 & -20.4744 \\ -69.5692 & -3.6896 & 2.6356 \end{pmatrix} \right\|$$
$$= \sqrt{5.3090^2 + 40.4363^2 + \cdots + 3.6896^2 + 2.6356^2}$$
$$= 103.3569$$

Keďže 103.3569 je viac ako $\rho = 1$, nastavíme $t = 1$ a opakujeme kroky 3, 4 a 5. Indexové množiny sú tentoraz všetky prázdne množiny.

$$I^1 = \begin{pmatrix} \emptyset & \emptyset & \emptyset & \emptyset & \emptyset & \emptyset & \emptyset & \emptyset & \emptyset & \emptyset \end{pmatrix}^\top$$

Matica U má po prvej iterácii tvar:

$$U^2 = \begin{pmatrix} 0.1852 & 0.2300 & 0.2089 & 0.2312 & 0.3650 & 0.7240 & 0.1595 & 0.1087 & 0.2773 & 0.1357 \\ 0.0897 & 0.1316 & 0.0503 & 0.0778 & 0.4475 & 0.1547 & 0.7408 & 0.8264 & 0.5924 & 0.7689 \\ 0.7251 & 0.6384 & 0.7408 & 0.6910 & 0.1876 & 0.1213 & 0.0997 & 0.0649 & 0.1303 & 0.0953 \end{pmatrix}.$$

Následne aktualizovaná matica M je:

$$M^2 = \begin{pmatrix} 58.0442 & 48.4121 & 73.8516 \\ 42.4220 & 43.6448 & 71.1541 \\ 54.2569 & 17.6405 & 61.1680 \\ 35.9396 & 21.9186 & 46.3372 \\ 20.4590 & 8.5773 & 57.8482 \end{pmatrix}.$$

$$\|M^2 - M^1\| = 17.3726$$

Po druhej iterácii sa nám rozdiel matíc M^2 a M^1 síce zmenšil, ale nie dostačujúco. Po ďalších iteráciach sa dostaneme do stavu:

$$M^{13} = \begin{pmatrix} 62.7003 & 42.4634 & 74.1456 \\ 39.0293 & 44.0847 & 72.9155 \\ 37.8871 & 18.6994 & 62.0041 \\ 40.9341 & 12.6024 & 47.2202 \\ 16.3557 & 7.5150 & 65.8585 \end{pmatrix}.$$

$$\|M^{13} - M^{12}\| = 1.0083 > \rho$$

Nech $t = 14$.

$$I^{14} = \begin{pmatrix} \emptyset & \emptyset & \emptyset & \emptyset & \emptyset & \emptyset & \emptyset & \emptyset & \emptyset & \emptyset \end{pmatrix}^\top$$

$$U^{14} = \begin{pmatrix} 0.0450 & 0.0919 & 0.1078 & 0.3353 & 0.8654 & 0.6703 & 0.6036 & 0.1555 & 0.2256 & 0.1153 \\ 0.0257 & 0.0580 & 0.0522 & 0.1275 & 0.0884 & 0.2110 & 0.3326 & 0.8138 & 0.7161 & 0.8452 \\ 0.9293 & 0.8501 & 0.8400 & 0.5372 & 0.0462 & 0.1187 & 0.0638 & 0.0307 & 0.0583 & 0.0395 \end{pmatrix}$$

$$M^{14} = \begin{pmatrix} 62.8524 & 42.3755 & 74.1302 \\ 39.0295 & 44.0315 & 72.8970 \\ 37.1811 & 18.9768 & 62.0306 \\ 41.1082 & 12.3027 & 47.2162 \\ 16.3390 & 7.5350 & 65.7763 \end{pmatrix}$$

$$\|M^{14} - M^{13}\| = 0.8591 < \rho$$

Dospeli sme ku záveru FCM algoritmu. Jeho hlavným výstupom je matica U^{14}. Vidíme, že každý objekt patrí s nenulovým koeficientom do každého z troch klastrov. Avšak pre analýzu rozdelenia nás najviac zaujímajú väčšie čísla. Všimnime si najprv najväčšie hodnoty v každom stĺpci, podľa nich by sme dostali zhlučenie, ktoré je zhodné s výsledkami oboch hierarchických metód.

Fuzzy zhlukovanie nám však umožňuje vidieť koeficienty členstva ku všetkým klastrom. Pre tri fuzzy zhluky sú hodnoty menšie ako 0.1 zanedbateľné. Zamerajme sa na tie vyššie a rozanalyzujme ich: Lekárska Fakulta UK nepatrí do tretieho zhluku jednoznačne, má výrazný podiel aj v prvom zhluku, Fakulta Sociál. Vied a zdravotníctva UKF a Fakulta Zdravotníctva KU majú nezanedbateľnú úlohu nielen v prvom, ale aj v druhom zhluku, a zároveň VŠ Zdravotníctva a soc. Práce sv. Alžbety a Fakulta Zdravotníctva TUAD sú mierne podobné s objektami v prvom zhluku.

Keď tieto výsledky porovnáme s K-means metódou, ktorá nám pre rovnakú vstupnú maticu M vytvorila 3 zhluky po 2, 2 a 6 objektoch, vidíme istú analógiu. Zlúčením prvého a druhého zhluku by sme dostali spomínaný zhluk šiestich fakúlt. Pri pohľade na U^{14} vidíme, že naozaj týchto 6 objektov sa výrazne líši od tretej skupiny a väčšina z nich má značný podiel aj v prvom aj v druhom zhluku. Avšak objekty tretieho zhluku podľa FCM sú veľmi podobné a FCM narozdiel od K-means vedie skôr k ich rozdeleniu na 3 a 1 fakultu, nie na 2 a 2 fakulty.

3 Rozdelenie Európy

V tejto časti práce aplikujeme teoretické poznatky z teórie zhlukovej analýzy na rozdelenie Európskych štátov do klastrov na základe získaných údajov. Údaje sme čerpali z [7] a spracovali pomocou softwaru R. Analýzy zameriame na dôležité subjekty, ktoré charakterizujú jednotlivé krajiny, na vládu, domácnosti a aj jednotlivcov. Budeme analyzovať nasledovné vstupné dáta:

1. percentuálne prerozdelenie vládnych výdavkov do rôznych sektorov,
2. percentuálny vládny príjem z rôznych typov daní,
3. normovaná spotreba priemernej domácnosti,
4. kvalita života jednotlivcov.

Porovnáme výsledky z rokov 2005 a 2012, aby sme zaznamenali nastané zmeny a vývoj. V prípade ak dáta pre niektoré krajiny neboli dostupné, tieto krajiny sme vynechali, aby sme zbytočne neskreslovali výsledky. Z tohto dôvodu zúžime naše analýzy na nasledovných 27 štátov: Belgicko, Bulharsko, Česko, Dánsko, Nemecko, Estónsko, Írsko, Grécko, Španielsko, Francúzsko, Taliansko, Cyprus, Lotyšsko, Litva, Luxembursko, Maďarsko, Malta, Holandsko, Rakúsko, Poľsko, Portugalsko, Slovinsko, Slovensko, Fínsko, Švédsko, Veľká Británia, Nórsko.

Spomedzi množstva spomenutých metód sme si vybrali z aglomeratívnych metód Wardovu metódu, minimalizujúcu varianciu, nakoľko sa nám vo všeobecnosti zdá metóda vychádzajúca zo štatistických princípov vhodnejšia. Z divíznych metód použijeme iteračnú schému Diana, kvôli jej nízkej výpočtovej náročnosti. Z nehierarchických metód použijeme najznámejší algoritmus K-means, z dôvodu jeho jednoduchosti a takmer lineárnej výpočtovej náročnosti. Pri vyhodnocovaní výsledkov si musíme dať pozor na jeho lokálnu konvergenciu, v závislosti od vstupných centroidov. Fuzzy metódy vynecháme, kvôli ich obtiažnej interpretácii.

Kvôli jednoduchosti si zvolíme rovnaký počet zhlukov pre všetky analýzy, konkrétne 5. Považujeme ho pri danom počte objektov za vhodný, priemerne bude obsahovať každý zhluk 5 až 6 objektov, ale zároveň nechávame možnosť ukázaniu sa väčších aj výrazne menších skupín, až jednotlivcov.

3.1 Vládne výdavky

Najprv sa zameriame na štruktúru vládnych výdavkov podľa sektoru, kam sú investované. Zaujíma nás percentuálne prerozdelenie vládnych prostriedkov do oblastí: služby pre verejnosť, obrana, verejný poriadok a bezpečnosť, hospodárstvo, ochrana životného prostredia, bývanie a občianska vybavenosť, zdravotníctvo, rekreácia, kultúra a náboženstvo, vzdelávanie a sociálna podpora.

2012

Vychádzame z údajov, ktoré sú uvedené v tabuľke A.1. Na obrázku 7 je zobrazený dendrogram získaný Wardovou aglomeratívnou metódou. Jej výstupom je rozdelenie štátov do nasledovných zhlukov:

- Belgicko, Malta, Poľsko, Maďarsko, Bulharsko, Španielsko
- Česko, Litva, Holandsko, Slovensko, Estónsko, Lotyšsko
- Dánsko, Švédsko, Nemecko, Fínsko, Rakúsko, Francúzsko, Nórsko, Luxembursko, Írsko, Slovinsko, UK
- Grécko, Taliansko, Portugalsko
- Cyprus

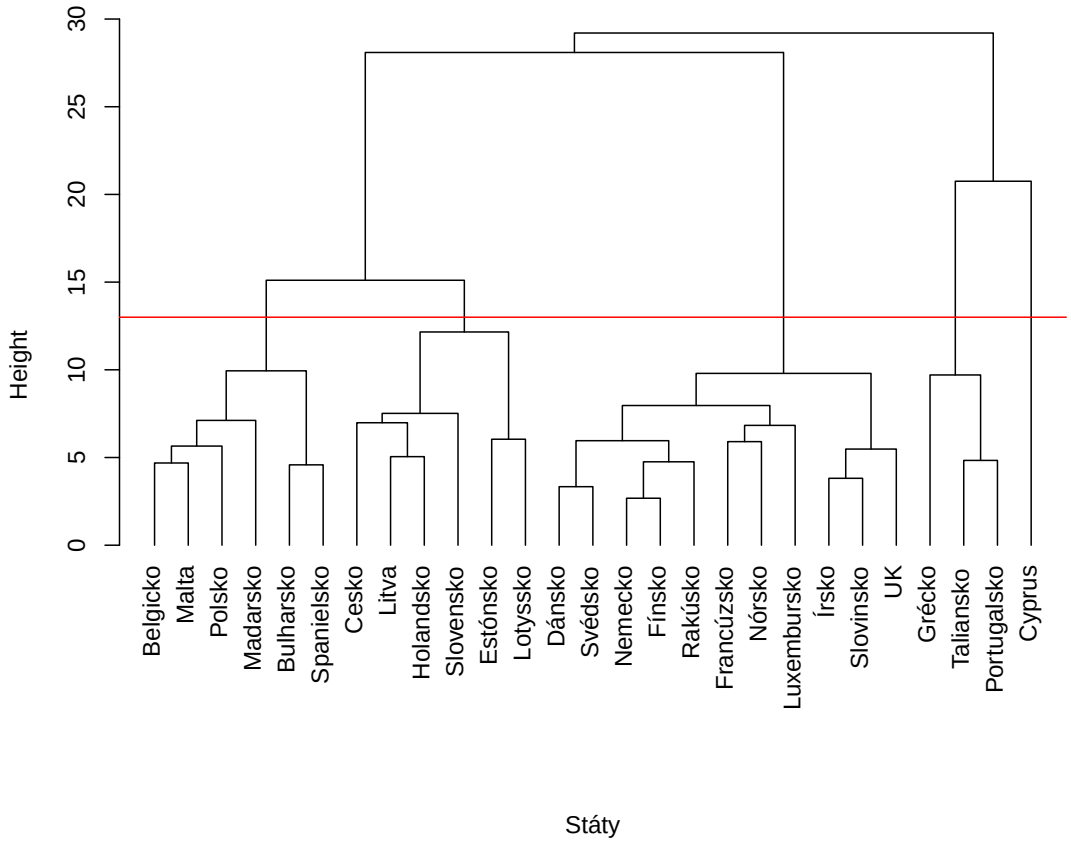
Druhý výstup poskytuje iteračná schéma Diana, ktorej dendrogram je na obrázku 8:

- Belgicko, Malta, Maďarsko, Bulharsko, Španielsko, Lotyšsko
- Česko, Holandsko, Slovensko, Estónsko, Litva
- Dánsko, Švédsko, Nemecko, Fínsko, Rakúsko, Francúzsko, Luxembursko, Írsko, Slovinsko, UK, Nórsko, Poľsko, Taliansko, Portugalsko
- Grécko
- Cyprus

Tretím je výsledok metódy K-means. Z množstva lokálnych riešení sme vybrali to najčastejšie.

- Belgicko, Malta, Maďarsko, Bulharsko, Španielsko, Lotyšsko, Poľsko

Obr. 7: Wardova metóda - Výdavky v roku 2012



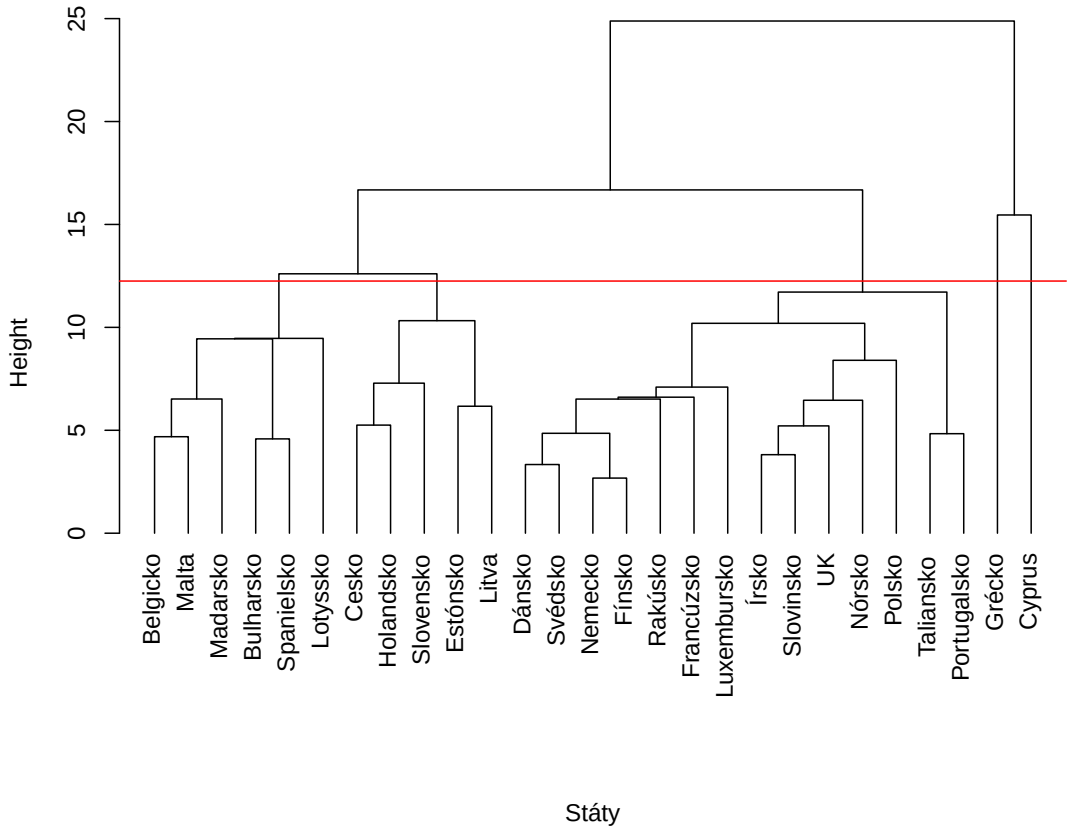
- Česko, Holandsko, Slovensko, Estónsko, Litva
- Dánsko, Švédsko, Nemecko, Fínsko, Rakúsko, Francúzsko, Luxembursko, Írsko, Slovinsko, UK, Nórsko
- Taliansko, Grécko, Portugalsko
- Cyprus

2005

V roku 2005 rozdelíme vládne výdavky do rovnakých sektorov. Percentuálne údaje sú uvedené v tabuľke A.2. Dendrogram Wardovej metódy na obrázku 9 vedie k tesnému rozdeleniu na zhluky:

- Belgicko, Maďarsko, Taliansko, Slovensko, Grécko

Obr. 8: DIANA - Výdavky v roku 2012

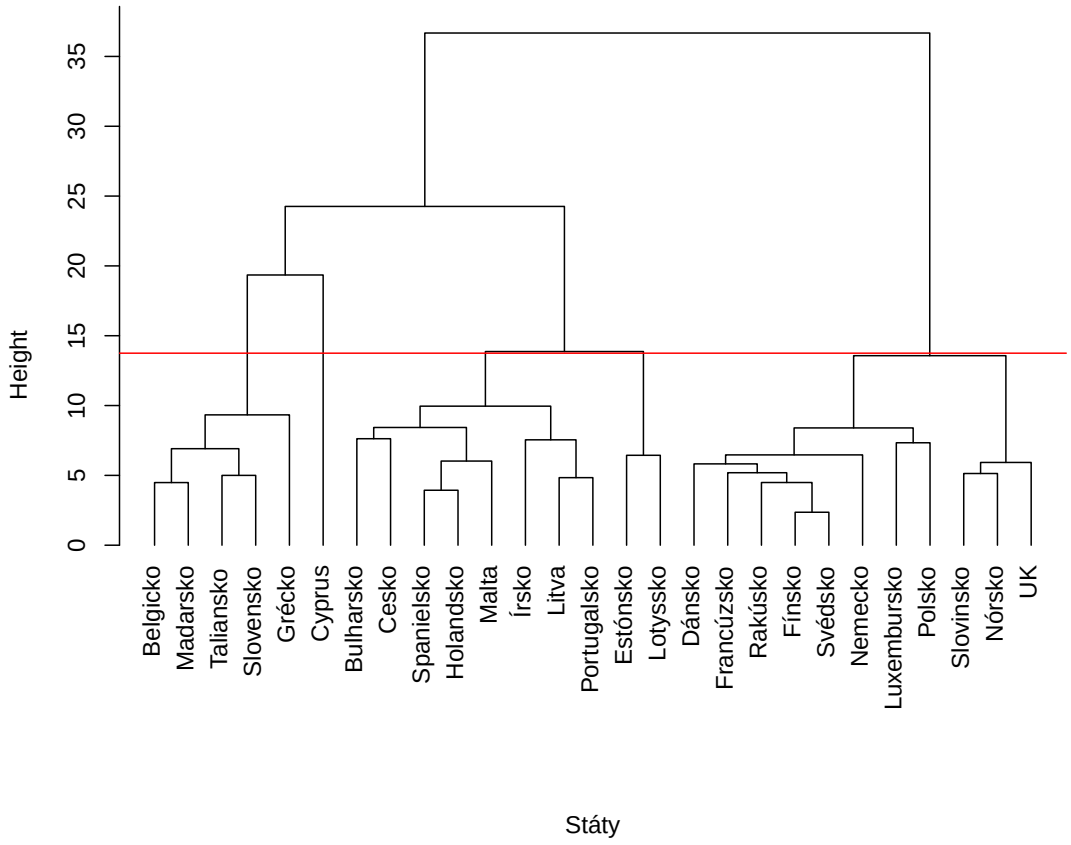


- Cyprus
- Bulharsko, Česko, Španielsko, Holandsko, Malta, Írsko, Litva, Portugalsko
- Estónsko, Lotyšsko
- Dánsko, Francúzsko, Rakúsko, Fínsko, Švédsko, Nemecko, Luxembursko, Poľsko, Slovinsko, Nórsko, UK

Rozdelenie Dianou je podľa obrázku 10 nasledovné:

- Belgicko, Maďarsko, Slovensko, Španielsko, Holandsko, Malta, Portugalsko, Bulharsko
- Grécko

Obr. 9: Wardova metóda - Výdavky v roku 2005

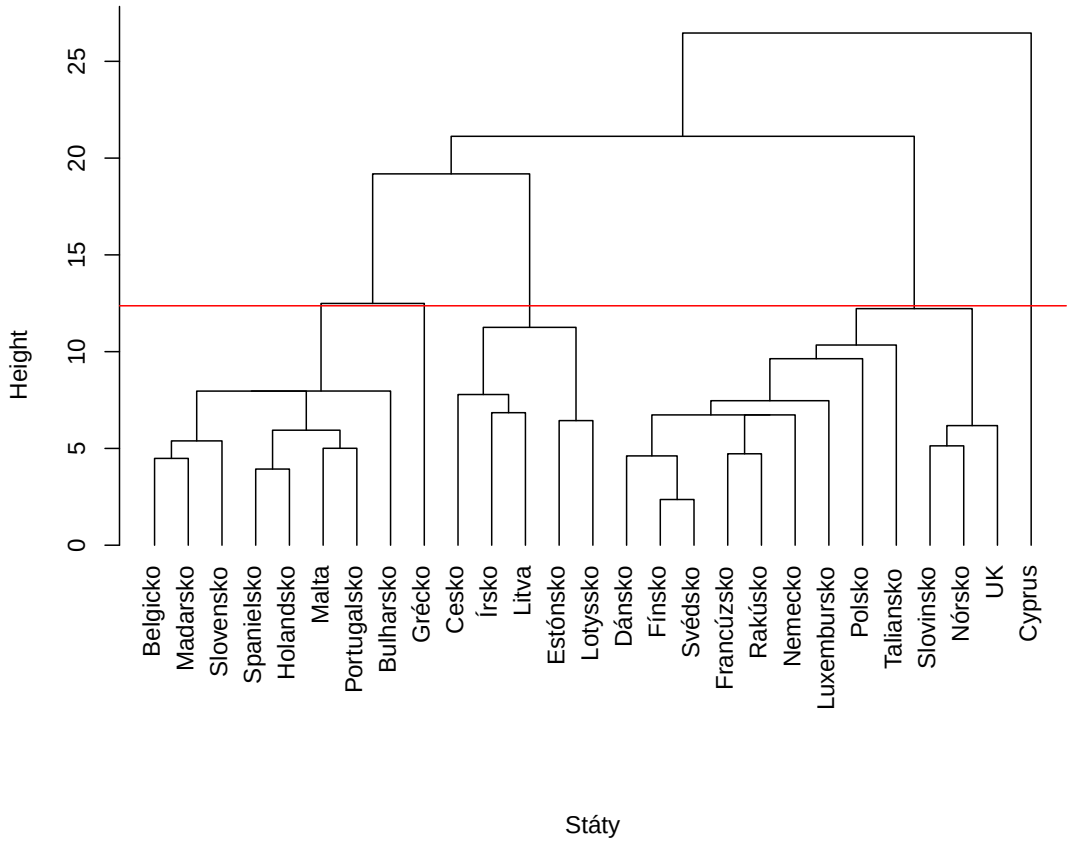


- Česko, Írsko, Litva, Estónsko, Lotyšsko
- Dánsko, Fínsko, Švédsko, Francúzsko, Rakúsko, Nemecko, Luxembursko, Poľsko, Taliansko, Slovinsko, Nórsko, UK
- Cyprus

Algoritmus K-means rozdelil štáty takto:

- Belgicko, Maďarsko, Slovensko, Španielsko, Holandsko, Malta, Portugalsko, Bulharsko, Grécko, Taliansko
- Slovinsko, Nórsko, UK
- Česko, Írsko, Litva, Estónsko, Lotyšsko

Obr. 10: DIANA - Výdavky v roku 2005



- Dánsko, Fínsko, Švédsko, Francúzsko, Rakúsko, Nemecko, Luxembursko, Poľsko,
- Cyprus

Zhrnutie:

Ako prvé si v tabuľkách A.1 a A.2 všimnime, že po siedmich rokoch sa prerozdelenie vládnych finančných prostriedkov do jednotlivých sektorov zmenilo a zmeny neboli rovnaké pre všetky krajiny. Napríklad Belgicko a Bulharsko výraznejšie znížili príspevky do služieb pre verejnosť a zároveň značne navýšili percento venované sociálnej podpore. Grécko a Cyprus taktiež pridelili viac financií na sociálnu podporu, avšak navýšili aj podiel peňazí pre sektor služby pre verejnosť. Tieto navýšenia prišli na úkor iných sektorov. Na základe porovnania údajov v tabuľke by sme vedeli ľahko porovnávať vývoj rozdelenia financií jednotlivých štátov. Zamerajme sa však teraz na porovnanie zhlukov

týchto krajín.

Výsledky jednotlivých metód sa v tomto prípade veľmi nelíšia, to naznačuje jasnú hranicu medzi zhlukmi. Všimnime si ustálené skupiny štátov, ktoré sa často zároveň zhodujú v oboch uvažovaných rokoch:

1. Cyprus
2. Belgicko, Bulharsko, Španielsko, Maďarsko, Malta
3. Dánsko, Nemecko, Francúzsko, Luxembursko, Rakúsko, Fínsko, Švédsko
4. Česko, Estónsko, Litva, Holandsko, Slovensko
5. Slovinsko, UK, Nórsko
6. Grécko, Taliansko, Portugalsko

V žiadnej skupine nevystupujú Poľsko, Írsko a Lotyšsko, buď sa ich zadelenie po siedmich rokoch zmenilo - napríklad Írsko sa v roku 2005 viac radilo do 4. skupiny hoci v roku 2012 patrilo jednoznačne k tretej a piatej skupine spolu. Iný prípad je Poľsko, ktoré v roku 2005 bolo zaradené so štátmi tretej a piatej skupiny, ale v roku 2012 jeho zaradenie nebolo úplne jednoznačné. Svojimi údajmi ho rôzne analýzy zhlučili s odlišnými štátmi. Podobne nejednoznačne je na tom Lotyšsko.

Pozrime sa teraz na skutočné dáta, podľa ktorých boli analýzy vykonané a skúsme nájsť dôvody pre tieto zoskupenia. Dáta sme naschvál nepreškálovali, aby väčšie finančné príspevky mali aj väčšiu váhu. Na prvý pohľad si všimneme, že najviac analýzu ovplyvnia tie sektory, v ktorých je najväčší rozdiel medzi minimálnou a maximálnou hodnotou. Medzi ne patria: služby pre verejnosť, hospodárstvo, zdravotníctvo, vzdelanie a sociálna podpora. Hodnoty v ostatných sektoroch majú menší vplyv, nakoľko sa líšia maximálne o 6 percent. Ďalej si všimnime už spomínané najmenšie a najväčšie hodnoty v stĺpcoch. Tie v podstate vysvetľujú odlúčenosť Cypru od ostatných krajín. V oboch analyzovaných rokoch Cyprus venuje oproti ostatným výrazne najviac financií do služieb pre verejnosť a aj do bývania a občianskej vybavenosti. Vyvažujú to tým, že zasa najmenej percent vládnych prostriedkov ide do zdravotníctva a sociálnej podpory.

Venujme sa teraz vplyvu položky sociálna podpora. V roku 2012 táto črta výborne popisuje skupiny 1 až 6. Tretia skupina do nej venuje najviac, všetci cez 41 percent.

Na ňu nadväzuje s cca 40 percentami piata skupina a aj Írsko. Kombinácia Írska, tretej a piatej skupiny tvoria v roku 2012 jeden zhluk. S podobnými hodnotami v tejto črte vystupuje 6. skupina, pričom si všimnime výsledky divíznej metódy Diana. Okolo 34 až 37 percent prideľuje sociálnej podpore 2. skupina, pričom aj Poľsko sa s 38 percentami blíži k tejto skupine. Na záver okrem Cypru najmenej percent financií venuje do sociálnej podpory 4. skupina - 31 až 35 percent. V roku 2005 je táto črta rovnako dôležitá. Väčšina hodnôt zodpovedá rovnakému rozdeleniu štátov, môžeme však vidieť aj medziročné rozdiely. Šiesta skupina tvorená iba tromi štátmi naberá až 5 percentný rozdiel medzi hodnotami. To môže byť čiastočnou príčinou, prečo tieto štáty neboli zadelené do rovnakých skupín všetkými metódami. Po druhé, Slovensko a Holandsko sa v tejto črte výrazne líšia od zvyšku 4. skupiny, blížia sa skôr ku hodnotám Belgicka, Maďarska a Španielska, čo aj zodpovedá výsledkom Diany a K-means.

Pri analýze ostatných čŕt nie je badateľné žiadne rozintervalovanie ako v prípade sociálnej podpory. Hodnoty jednotlivých členov skupín sa síce poväčšinou podobajú, avšak nastáva podobnosť aj medzi štátmi v rôznych skupinách. Sledujme najprv zdravotníctvo. V roku 2012 hodnoty od 13 do 18 percent nadobúdajú aj 3. aj 4. skupina, pričom analýzy nevykazujú žiadnu podobnosť medzi týmito štátmi. V roku 2005 takéto hodnoty nadobúdala navyše aj 2. skupina. Tu môžeme vidieť slabý súvis medzi miešaním štátov 2. a 4. skupiny. Pri črte hospodárstvo krásne vidno, že 2. a 4. skupina do neho v oboch rokoch venovala percentuálne veľa financií. Relatívne málo doňho investovali štáty v tretej, piatej a šiestej skupiny. Tieto tri skupiny sú často spolu zaradené do zhluku. Nakoniec analyzujme vplyv služieb pre verejnosť. Táto črta nadobúda vysoké hodnoty v prípade šiestej skupiny v roku 2012, v roku 2005 bolo Portugalsko ešte dosť rozdielne. Okolo 13 až 19 percent venuje druhá skupina. 6 percent je síce veľký rozdiel, ale nakoľko ostatné skupiny dávajú pod 13% na služby pre verejnosť, tak to nie je až taký problém. Podobné množstvo financií sem dávajú štáty tretej, štvrtej aj piatej skupiny v oboch rokoch. Hranica medzi nimi žiaľ nie je jasná.

Porovnaním výsledkov klastrovej analýzy a skutočných údajov sa zdá, že najväčší odraz vo výsledkoch majú sektory hospodárstvo a sociálna podpora. Ostatné podporujú výsledky metód, ale neurčujú presné hranice medzi jednotlivými zhlukmi. Ak by sme ešte pomocou dát z tabuliek chceli zaradiť Poľsko, Írsko a Lotyšsko, Poľsko by sme v

roku 2012 zaradili ku druhej a v roku 2005 ku tretej skupine. Írsko v roku 2012 ku piatej a v roku 2005 buď k druhej alebo štvrtej skupine. A na záver Lotyšsko v roku 2012 ku štvrtej skupine hlavne kvôli podobnej hodnote v črte sociálna podpora a v roku 2005 ku Estónsku a teda opäť ku skupine 4.

3.2 Vládne príjmy

V druhej časti ostaneme pri vláde, ale pozrieme sa odkiaľ získava financie. Existuje viacero zdrojov, či už z dlhopisov alebo finančných trhov, my sa však zameriame na rôzne skupiny daní. Daňové systémy sú rôzne v každom štáte a klastrová analýza nám poskytuje možnosť vzájomne ich porovnať. Uvedieme percentuálny vládny príjem z nasledujúcich skupín:

1. Daň z pridanej hodnoty a importu - DPH, dane a clá na dovoz
2. Dane z produktov okrem DPH a importu - napríklad z lotérie, stávkovania, poistnej prémie, finančných transakcií, spotrebné dane
3. Iné dane z produkcie - napríklad z pozemkov a budov, za znečistenie
4. Daň z príjmu
5. Iné bežné dane - bežné dane z kapitálu, z výdavkov, volebné dane, ...
6. Daň z kapitálu
7. Dane zo sociálnych príspevkov

Obdobne ako v sekcii 3.1 uvedieme výsledky troch metód pre každý analyzovaný rok.

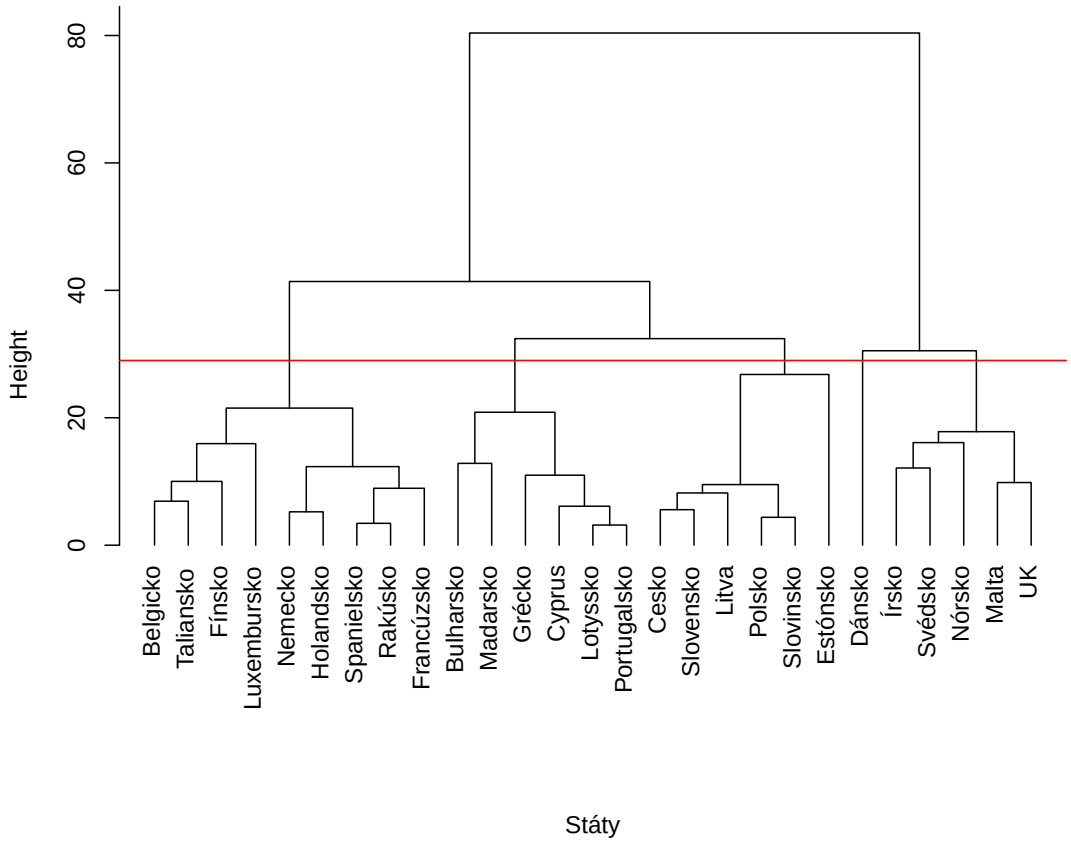
2012

Algoritmy aplikujeme na dáta v tabuľke A.3.

Wardova aglomeratívna metóda (podľa obrázku 11) vedie k rozdeleniu:

- Belgicko, Taliansko, Fínsko, Luxembursko, Nemecko, Holandsko, Španielsko, Rakúsko, Francúzsko
- Bulharsko, Maďarsko, Grécko, Cyprus, Lotyšsko, Portugalsko

Obr. 11: Wardova metóda - Príjmy v roku 2012

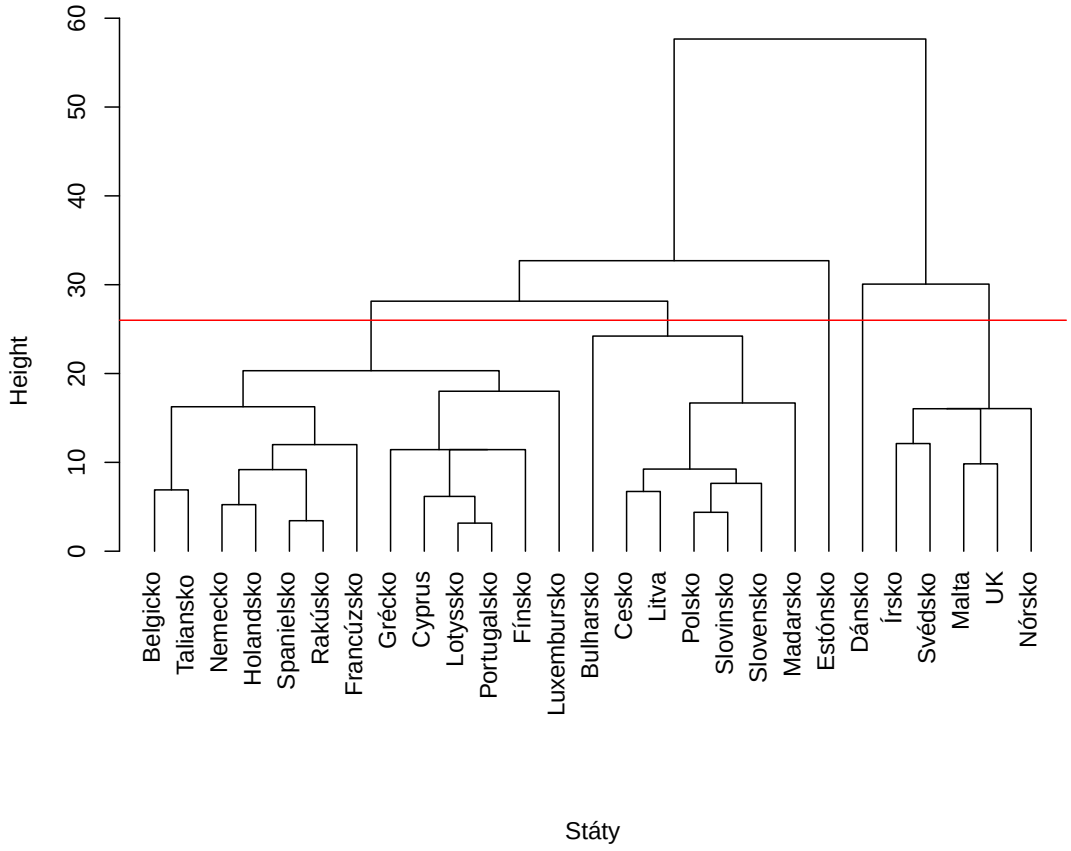


- Česko, Slovensko, Litva, Poľsko, Slovinsko, Estónsko
- Dánsko
- Írsko, Švédsko, Nórsko, Malta, UK

Divízny algoritmus DIANA (podľa obrázku 12):

- Belgicko, Taliansko, Nemecko, Holandsko, Španielsko, Rakúsko, Francúzsko, Grécko, Cyprus, Lotyšsko, Portugalsko, Fínsko, Luxembursko
- Bulharsko, Česko, Litva, Poľsko, Slovinsko, Slovensko, Maďarsko
- Estónsko
- Dánsko

Obr. 12: DIANA - Príjmy v roku 2012



- Írsko, Švédsko, Malta, UK, Nórsko

Nehierarchická iteračná schéma K-means:

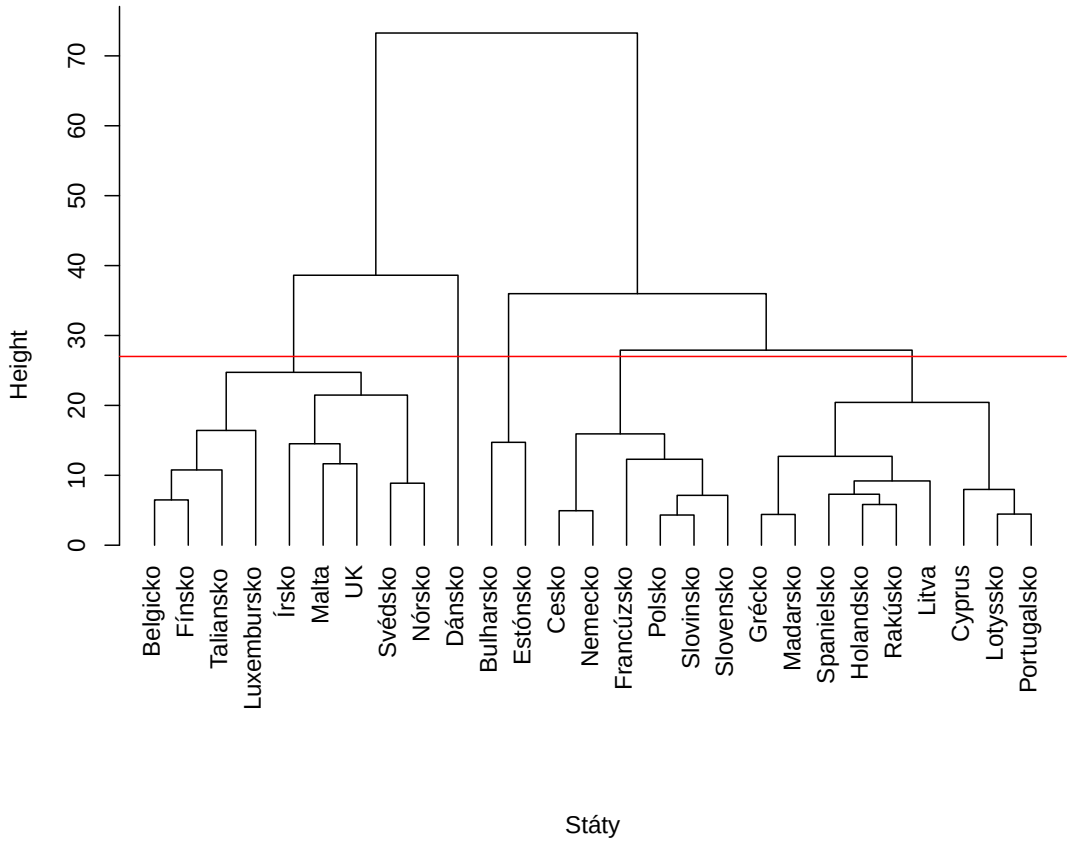
- Belgicko, Taliansko, Nemecko, Holandsko, Španielsko, Rakúsko, Francúzsko
- Bulharsko, Grécko, Cyprus, Lotyšsko, Luxembursko, Portugalsko, Fínsko
- Česko, Litva, Poľsko, Slovinsko, Slovensko, Maďarsko, Estónsko
- Írsko, Švédsko, Malta, UK, Nórsko
- Dánsko

2005

Údaje z roku 2005 sa nachádzajú v tabuľke A.4. Výstupy z nich sú nasledovné.

Wardova aglomeratívna metóda (podľa obrázku 13):

Obr. 13: Wardova metóda - Príjmy v roku 2005

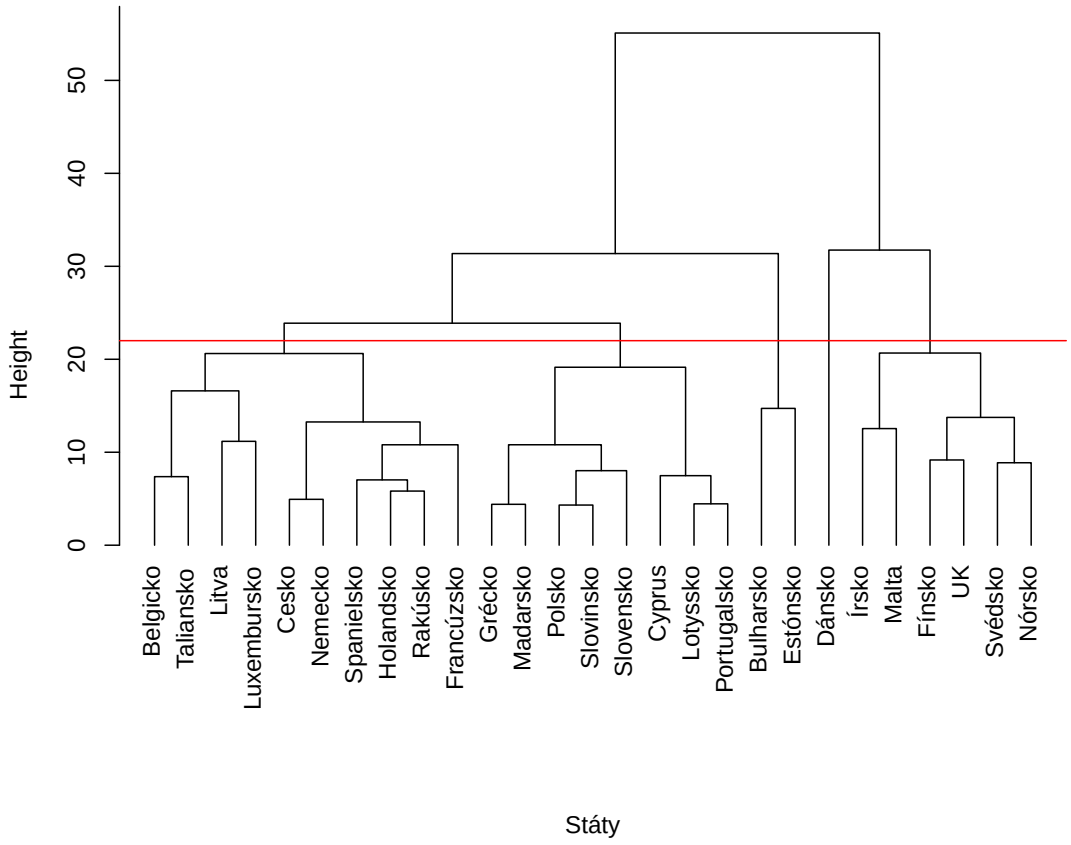


- Belgicko, Fínsko, Taliansko, Luxembursko, Írsko, Malta, UK, Švédsko, Nórsko
- Dánsko
- Bulharsko, Estónsko
- Česko, Nemecko, Francúzsko, Poľsko, Slovinsko, Slovensko
- Grécko, Maďarsko, Španielsko, Holandsko, Rakúsko, Litva, Cyprus, Lotyšsko, Portugalsko

Divízny algoritmus DIANA (podľa obrázku 14):

- Belgicko, Taliansko, Litva, Luxembursko, Česko, Nemecko, Španielsko, Holandsko, Rakúsko, Francúzsko

Obr. 14: DIANA - Príjmy v roku 2005



- Grécko, Maďarsko, Poľsko, Slovinsko, Slovensko, Cyprus, Lotyšsko, Portugalsko
- Bulharsko, Estónsko
- Dánsko
- Írsko, Malta, Fínsko, UK, Švédsko, Nórsko

Nehierarchická iteračná schéma K-means:

- Česko, Nemecko, Francúzsko, Slovinsko, Slovensko, Poľsko
- Grécko, Španielsko, Taliansko, Cyprus, Lotyšsko, Litva, Maďarsko, Holandsko, Rakúsko, Portugalsko
- Dánsko

- Belgicko, Írsko, Malta, Luxembursko, Fínsko, Švédsko, UK, Nórsko
- Bulharsko, Estónsko

Zhrnutie:

Najprv sa zamerajme na výstupy jednotlivých zhlukových metód. Možno pozorovať, že niektoré zhluky sú veľmi podobné aj v roku 2005 aj v roku 2012. To môže byť spôsobené tým, že daňové systémy sa nezvyknú meniť veľmi často a štruktúra obyvateľstva čo sa týka zamestnanosti, sociálnych príspevkov alebo nakupujúcich v obchodoch sa tiež za sedem rokov veľmi nezmení. Ustálené skupiny odpozorované z výsledkov metód v oboch rokoch sú:

1. Dánsko
2. Belgicko, Luxembursko, Fínsko, Holandsko, Rakúsko, Taliansko, Španielsko, Francúzsko, Nemecko
3. Česko, Poľsko, Slovinsko, Slovensko, Litva
4. Bulharsko, Estónsko
5. Írsko, Malta, Švédsko, UK, Nórsko
6. Grécko, Cyprus, Lotyšsko, Portugalsko, Maďarsko

Napriek očakávaným malým zmenám medzi výsledkami v rokoch 2005 a 2012 nás analýza upozornila na nasledovné fakty: Bulharsko a Estónsko mali veľmi podobnú štruktúru vládnych výdavkov iba v roku 2005. V roku 2012 sa Estónsko viac zaraďuje do zhluku s treťou skupinou a Bulharsko viac so šiestou. Ďalej Litva sa v roku 2005 viac zatrieďuje ku niektorým štátom druhej skupiny.

Ako druhé sa sústreďme na samostatné údaje v tabuľkách A.3 a A.4 a pokúsme sa pomocou nich odôvodniť vzniknuté skupiny štátov. Nakoľko nás zaujíma, z ktorých typov daní vláda získava najviac prostriedkov, tieto dáta nenormalizujeme. Všimnime si extramálne hodnoty v každom stĺpci. Vidíme, že daň zo sociálnych príspevkov a daň z príjmu nadobúdajú najväčšie aj rozdiely medzi štátmi a zároveň aj najvyššie hodnoty vôbec. Tieto dve skupiny daní tým pádom najviac vplývajú na našu analýzu. Zanedbateľná nie je úplne ani daň z pridanej hodnoty a importu.

V roku 2012 sú dane zo sociálnych príspevkov takmer príkladne podelené na intervaly. Percentuálne najmenej z tejto dane získava Dánsko (iba 2%). To a nadmerne vysoký príjem z dane z príjmu sú hlavnými príčinami jeho úplne outlierskej pozície. 15 až 23 percent získavajú z tejto dane štáty 5. skupiny. Ďalej v rozmedzí 26 a 40 percent sa nachádzajú druhá a šiesta skupina, čo sa odráža v ich zhlučení podľa Diany. Najvyššie percentá od 40 do 45 sú v prípade tretej skupiny. V roku 2005 nastáva zmena hlavne pre štáty druhej skupiny. Hodnoty tejto črty sa u všetkých okrem Nemecka a Francúzska pohybujú na úrovni $28-35\%$, čo sa výrazne kryje so šiestou skupinou. Nemecko a Francúzsko získavali zo sociálnych príspevkov nad 37 percent, čo sa zhoduje s treťou skupinou. Podobné výsledky nám dávajú v roku 2005 Wardova a K-means metóda.

Analyzujme ešte druhú významnú položku, a to daň z príjmu. V roku 2012 sú opäť výsledky dosť jednoznačné, najviac z dane z príjmu získava Dánsko, ďalej so značným odstupom piata skupina. Stredné hodnoty 24 až 35 percent z nej získavajú druhá a šiesta skupina okrem Maďarska, čo implikuje ich zhlučenie ako Diana. Najnižšie a veľmi podobné percentá získavajú z tejto črty tretia a štvrtá skupina. V roku 2005 nastáva zmena opäť v odlíšení Nemecka a Francúzska od zvyšku druhej skupiny. Na druhej strane nastáva podobnosť medzi zvyškom druhej skupiny a Litvy so svojimi 30 percentami.

Obe tieto črty výborne popisujú výsledky zhlukovej analýzy. Odrážajú aj podobnosť zhlukov po siedmich rokoch, ale aj preskupenie niektorých jednotlivcov. Čo sa týka ostatných čŕt, nevykazujú až také jednoznačné hranice medzi jednotlivými skupinami, nakoľko rozsah, v ktorom sa hodnoty vyskytujú, je menší. Zároveň však tieto hodnoty nevytvárajú ani žiadnu kritiku na výstupy klastrových metód.

3.3 Spotreba domácností

Ako tretie nás budú zaujímať domácnosti a konkrétne na aké tovary a služby míňajú peniaze. Aby sme predišli výkyvom spôsobenými výrazne rozdielnymi platmi v rôznych krajinách, normujeme spotrebu. To znamená, že nás nezaujíma výška jednotlivých položiek, ale iba vzájomný pomer. Preto sme údaje preškálovali na promile (nachádzajú sa v tabuľke A.5). Najnovšie dáta v tejto oblasti sú z roku **2005**, preto v tejto časti

výnimočne analýzu roku 2012 musíme vynechať. V analýze uvažujeme bežné sektory tovarov a služieb pre priemernú domácnosť: jedlo a nealkoholické nápoje, alkoholické nápoje, tabak a narkotiká, oblečenie a obuv, energie, zariaďovanie a údržba domácnosti, zdravie, doprava, komunikácie, rekreácia a kultúra, vzdelanie, reštaurácie a hotely a poistenie.

Analýzu opäť začneme Wardovou metódou, poskytujúcou nám dendrogram na obrázku 15. Pretnutie horizontálnou čiarou nám určí rozdelenie na nasledovných 5 klastrov:

- Belgicko, Francúzsko, Írsko, Rakúsko, Slovinsko, Holandsko, Dánsko, Švédsko, Nemecko, Luxembursko, UK, Fínsko, Nórsko
- Grécko, Portugalsko, Cyprus, Španielsko, Taliansko
- Bulharsko, Estónsko, Slovensko, Poľsko
- Česko, Maďarsko, Malta
- Lotyšsko, Litva

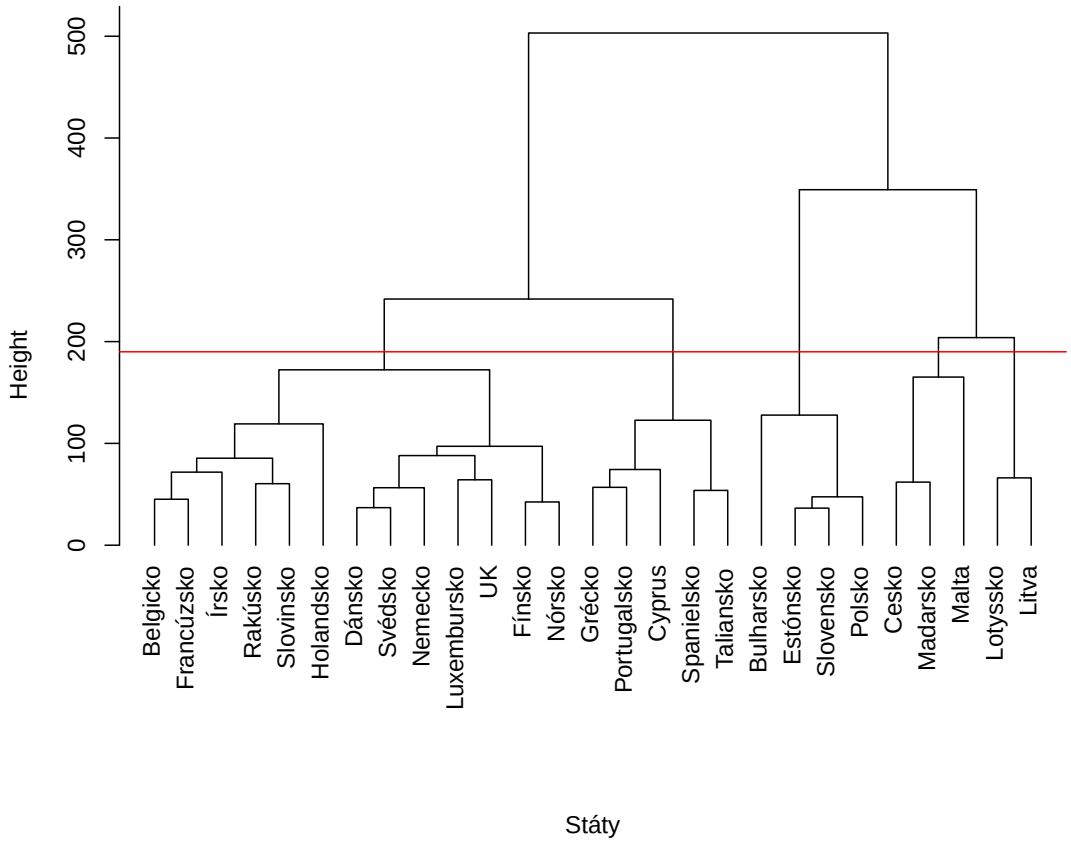
Podobne pomocou divíznej metódy získavame z obrázku 16 zatriedenie:

- Belgicko, Francúzsko, Írsko, Rakúsko, Dánsko, Švédsko, Nemecko, Fínsko, Nórsko, Luxembursko, UK, Holandsko,
- Česko, Maďarsko, Slovinsko, Grécko, Portugalsko, Cyprus, Španielsko, Taliansko
- Malta
- Bulharsko, Estónsko, Slovensko, Poľsko
- Lotyšsko, Litva

Nakoniec nehierarchická metóda K-means poskytla výstup:

- Bulharsko, Estónsko, Slovensko, Poľsko
- Česko, Maďarsko, Malta, Lotyšsko, Litva
- Grécko, Portugalsko, Cyprus, Španielsko, Taliansko

Obr. 15: Wardova metóda - Spotreba v roku 2005



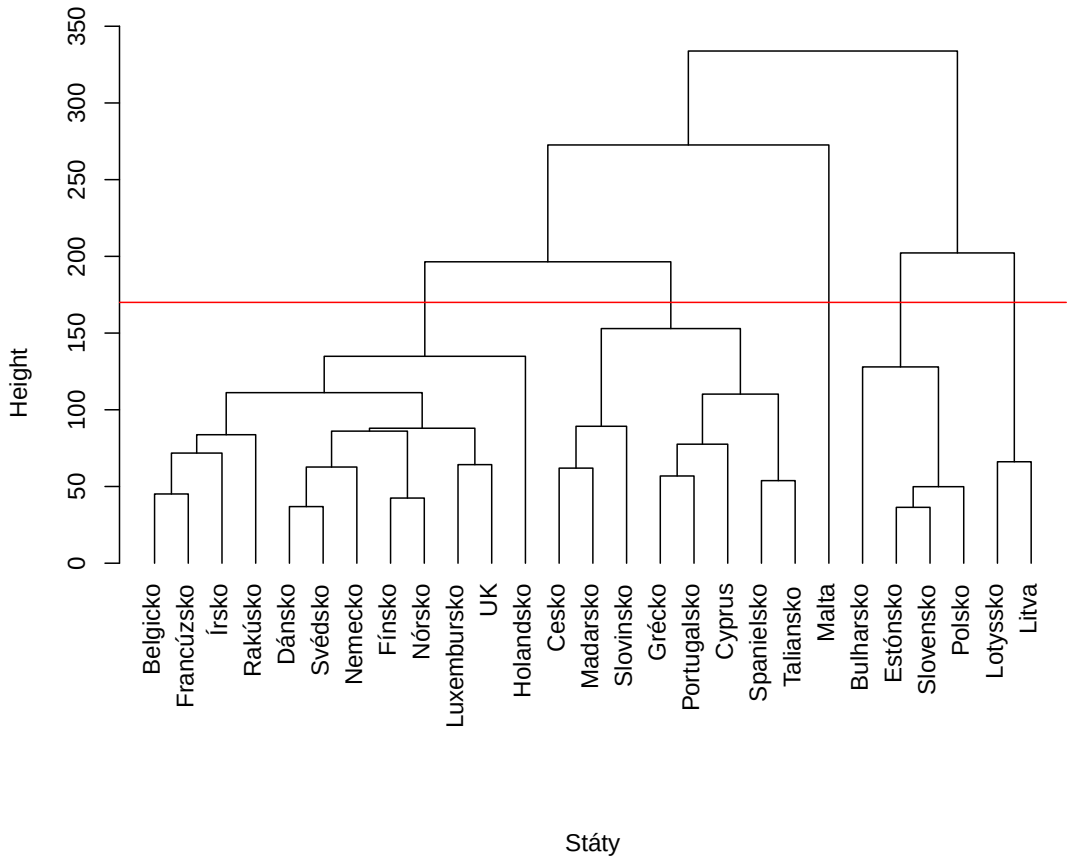
- Belgicko, Francúzsko, Írsko, Rakúsko, Slovinsko, Holandsko
- Dánsko, Švédsko, Nemecko, Luxembursko, UK, Fínsko, Nórsko

Zhrnutie

Zhlukové metódy vytvorili v tomto prípade takmer rovnaké rozdelenie do zhlukov. Na základe nich vytvoríme nasledujúce skupiny štátov:

1. Bulharsko, Estónsko, Slovensko, Poľsko
2. Česko, Maďarsko, Malta
3. Lotyšsko, Litva
4. Grécko, Portugalsko, Cyprus, Španielsko, Taliansko

Obr. 16: DIANA - Spotreba v roku 2005



5. Belgicko, Francúzsko, Írsko, Rakúsko, Slovinsko, Holandsko

6. Dánsko, Švédsko, Nemecko, Luxembursko, UK, Fínsko, Nórsko

Obe hierarchické metódy viedli ku spojeniu prvej a druhej skupiny do jedného veľkého zhluku. Je to podložené aj údajmi v tabuľke, alebo je jasný rozdiel medzi týmito dvomi skupinami? Pozrime sa teda na tabuľku A.5. Keďže novšie údaje nie sú k dispozícii, prichádzame o možnosť pozorovať vývoj, akým sa uberá distribúcia financií domácností. Dáta sú prerátane na promile, čo prislúcha predstave domácnosti, ktorá má práve 1000 na mesiac.

Podobne ako v častiach 3.1 a 3.2 všimnime si stĺpce s najväčšíme rozdielmi - to sú energie, jedlo a nealkoholické nápoje, doprava a nakoniec poistenie. Začnime teda energiami. Na ne minú viac než 310 promile krajiny v 1. skupine. Stredne draho vyjdú

energie v 4., 5. a 6. skupine, pričom podobnosť piatej a šiestej skupiny vykazujú aj zhlukové metódy. Čo sa týka druhej skupiny, podobné hodnoty majú Česko a Maďarsko, Malta sa v tejto črte líši o 40 promile, čo môže byť čiastočný dôvod pre jej oddelenie metódou Diana. Podobné a osobité hodnoty majú aj Lotyšsko a Litva, čo však vôbec nie je prekvapením pri pohľade na všetky stĺpce.

Ako druhý analyzujme vplyv výdavkov na jedlo a nealko. Ten vykazuje obrovskú zhodu s klastrovou metódou. Najviac minie 3. skupina $(300-350)$, potom 1. skupina s $230-320$ promile. Druhá skupina minie $210-230$, štvrtá $160-200$ a na záver piata a šiesta skupina s hodnotami $110-170$ sa výrazne podobajú. Čo sa týka výdavkov na dopravu, výrazne najnižšie hodnoty nadobúda 1. skupina, avšak hranice medzi zvyšnými skupinami vôbec nie sú jednoznačné. Táto črta nám teda výsledky clusteringu veľmi nepodporila. Na záver výdavky na poistenie nám tiež neosvetlia situáciu, nakoľko síce štáty v jednej skupine majú navzájom podobnú hodnotu, ale táto hodnota sa podobá aj so štátmi v iných skupinách.

Na základe pozorovania vplyvov jednotlivých čŕt, usudzujeme, že na klastrovú analýzu mali najvýraznejší vplyv výdavky na energie a na jedlo a nealkoholické nápoje. Ostatné črty nám na prvý pohľad vytvárajú veľmi rôznorodé rozdelenia.

3.4 Kvalita života

V poslednej časti sa pokúsime vymodelovať spokojnosť občanov, alebo kvalitu života jednotlivcov. Pod týmito pojmami si vieme toho predstaviť veľa, spokojnosť so školstvom, zdravotníctvom, pracovnými možnosťami, výšku platu, dostatočné kultúrne možnosti, čisté životné prostredie. Avšak väčšina z týchto ukazovateľov sa iba ťažko vyjadruje v číslach. Takisto platy alebo ceny v obchode nie sú dobré črty, lebo spokojnosť závisí od reálnej hodnoty peňazí, nie nominálnej. Nakoniec sme zvolili nasledovné črty:

1. očakávaná dĺžka života (v rokoch)- vysoká hodnota svedčí o dobrom zdravotníctve a o všeobecne pokojnom, zdravom živote

2. počet zdravých dojčeniec (z 1000 dojčeniec) - takisto je to ukazovateľ kvality zdravotníctva a vysoká hodnota spôsobuje radosť rodičom týchto zdravých detí

3. miera zamestnanosti (v percentách) - čím vyššia, tým viac ľudí má pracovné

miesto, príjem a tým pádom sa cítia istejšie

4. priemerná pracovná doba (v hodinách na týždeň) - čím nižšia tým sú pracujúci ľudia šťastnejší, majú viac času na rodinu a priateľov

5. nezadĺžená časť populácie (v percentách) - chápeme pod tým ľudí s nedoplatkami v účtoch, hypotékach, nájomnom.

6. časť populácie schopná čeliť nečakaným finančným nákladom (v percentách) - posledné dve črty jasne poukazujú na finančnú situáciu jednotlivcov, pričom samozrejme čím viac ľudí nemá finančné ťažkosti, tým sú v priemere šťastnejší.

Konkrétne údaje z roku 2012 sa nachádzajú v tabuľke A.6 a z roku 2005 v tabuľke A.7. Keďže tieto hodnoty nie sú v rovnakých jednotkách, musíme ich preškálovať. Na to použijeme vzorce (5) uvedené v teoretickej časti. Navyše preškálované hodnoty črty priemerná pracovná doba vynásobime $\cdot(-1)$, aby pre všetky črty platilo: čím vyššia hodnota, tým väčšia spokojnosť obyvateľov.

2012

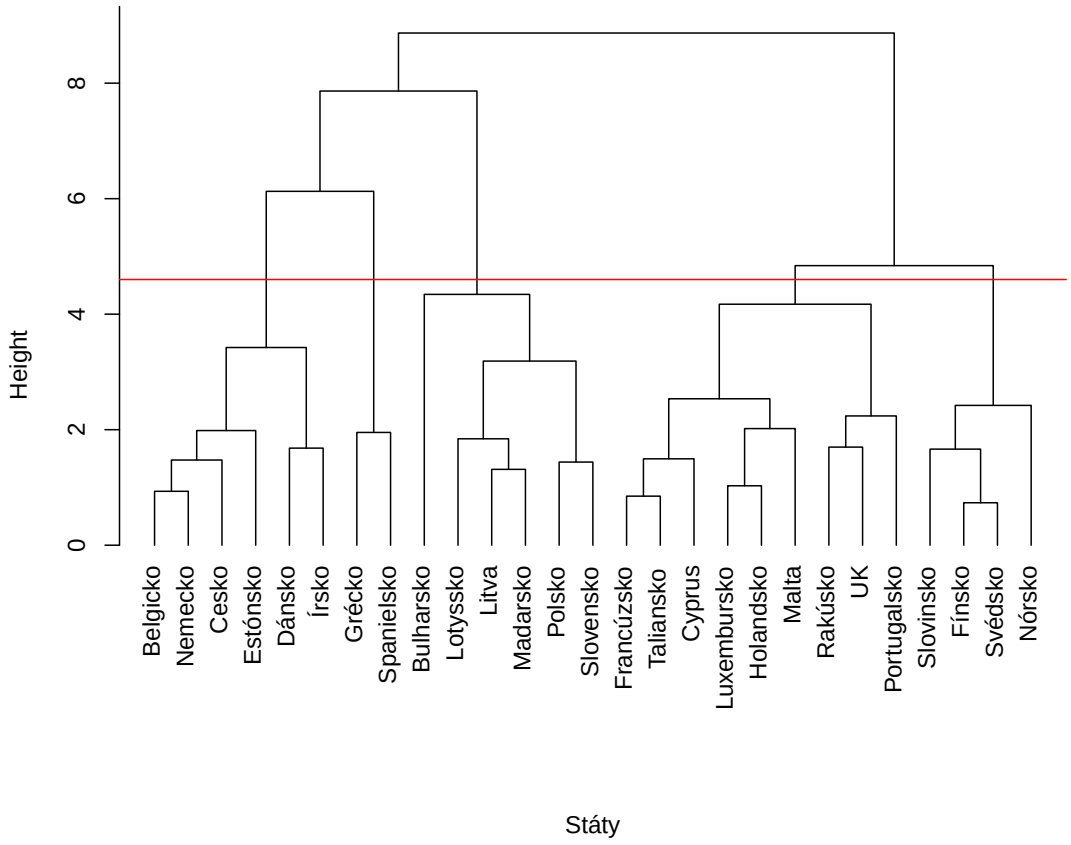
Najprv získavame výsledky z dendrogramu na obrázku 17 z Wardovej metódy:

- Belgicko, Nemecko, Česko, Estónsko, Dánsko, Írsko
- Grécko, Španielsko
- Bulharsko, Lotyšsko, Litva, Maďarsko, Poľsko, Slovensko
- Francúzsko, Taliansko, Cyprus, Luxembursko, Holandsko, Malta, Rakúsko, UK, Portugalsko
- Slovinsko, Fínsko, Švédsko, Nórsko

Ďalej divízna metóda určuje rozdelenie podľa obrázku 18:

- Belgicko, Nemecko, Česko, Dánsko, Írsko, Francúzsko, Taliansko, Cyprus
- Luxembursko, Holandsko, Slovinsko, Fínsko, Švédsko, Nórsko, Malta, Rakúsko, UK, Portugalsko
- Grécko, Španielsko

Obr. 17: Wardova metóda - Kvalita v roku 2012

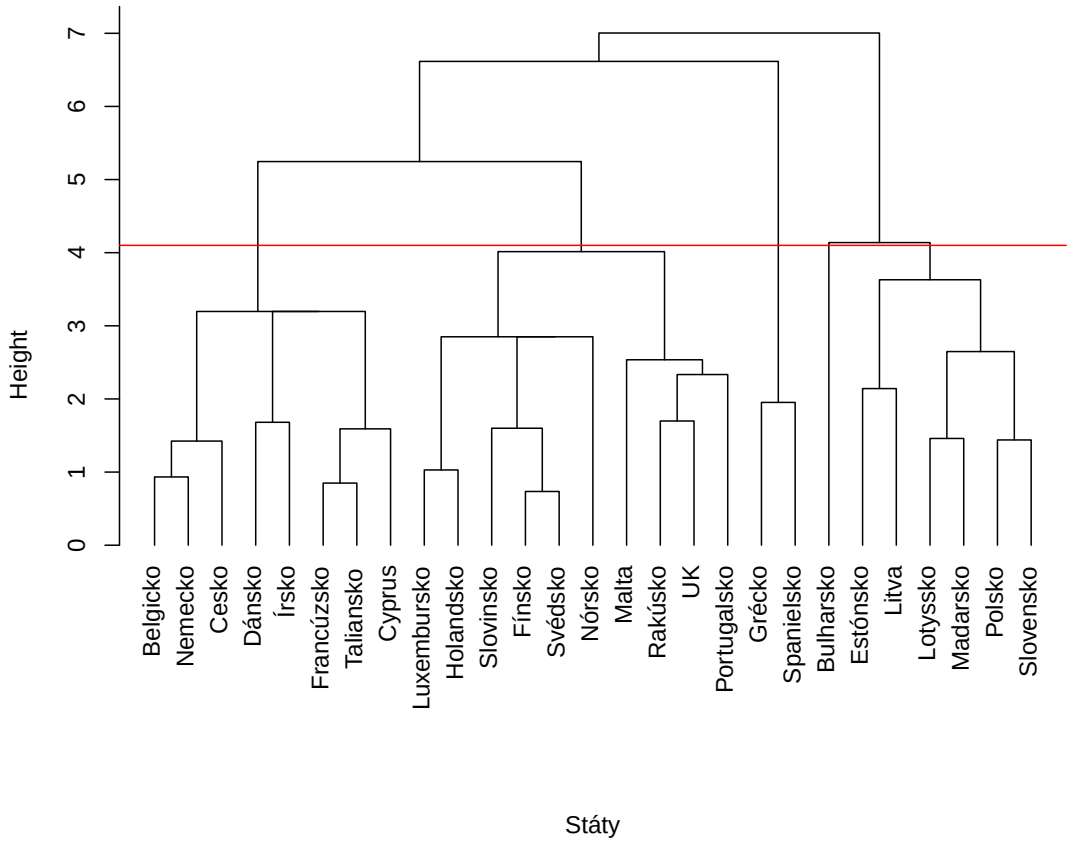


- Bulharsko
- Estónsko, Litva, Lotyšsko, Maďarsko, Poľsko, Slovensko

K-means iteračná metóda nám dáva výstup:

- Grécko
- Španielsko
- Belgicko, Nemecko, Česko, Dánsko, Írsko, Francúzsko, Taliansko, Cyprus, Estónsko
- Litva, Lotyšsko, Maďarsko, Poľsko, Slovensko, Bulharsko
- Luxembursko, Holandsko, Slovinsko, Fínsko, Švédsko, Nórsko, Malta, Rakúsko, UK, Portugalsko

Obr. 18: DIANA - Kvalita v roku 2012



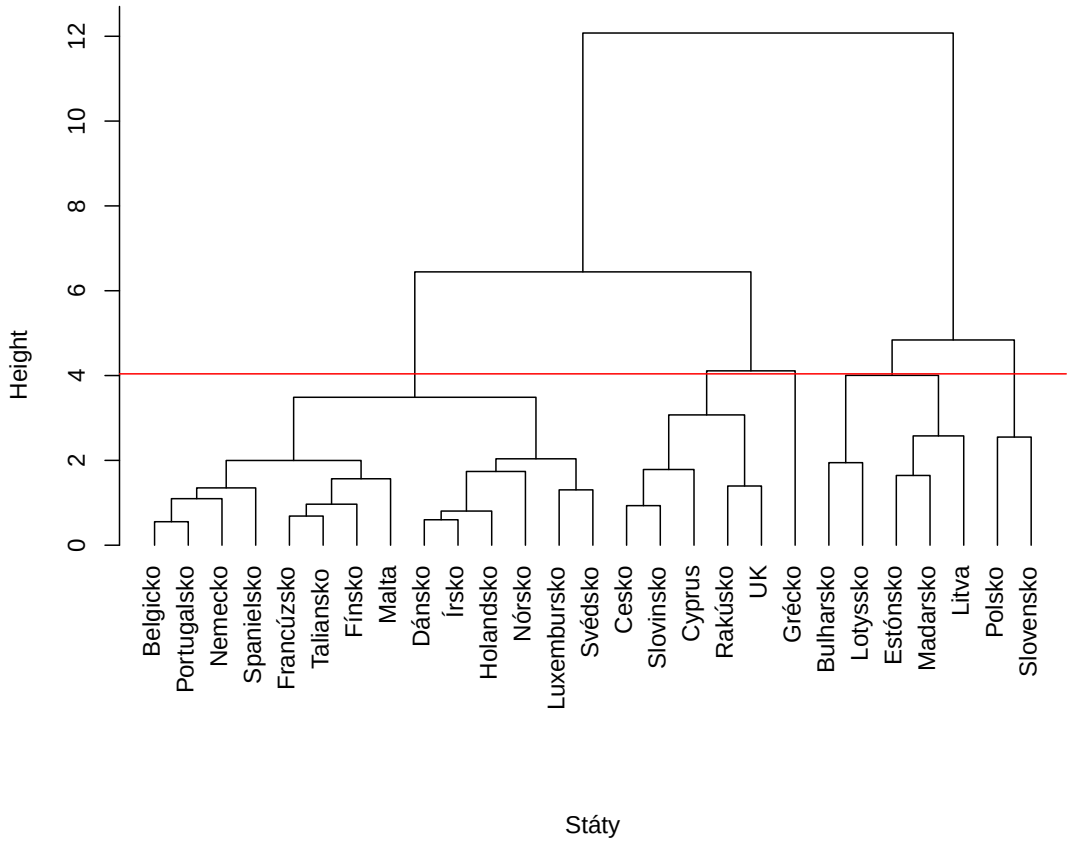
2005

Poslednú analýzu opäť začneme Wardovou metódou, poskytujúcou nám dendrogram na obrázku 19 a teda rozdelenie:

- Belgicko, Portugalsko, Nemecko, Španielsko, Francúzsko, Taliansko, Fínsko, Malta, Dánsko, Írsko, Holandsko, Nórsko, Luxembursko, Švédsko
- Česko, Slovinsko, Cyprus, Rakúsko, UK
- Grécko
- Bulharsko, Lotyšsko, Estónsko, Maďarsko, Litva
- Poľsko, Slovensko

Divízna metóda vedie podľa dendrogramu na obrázku 20 ku klastrom:

Obr. 19: Wardova metóda - Kvalita v roku 2005

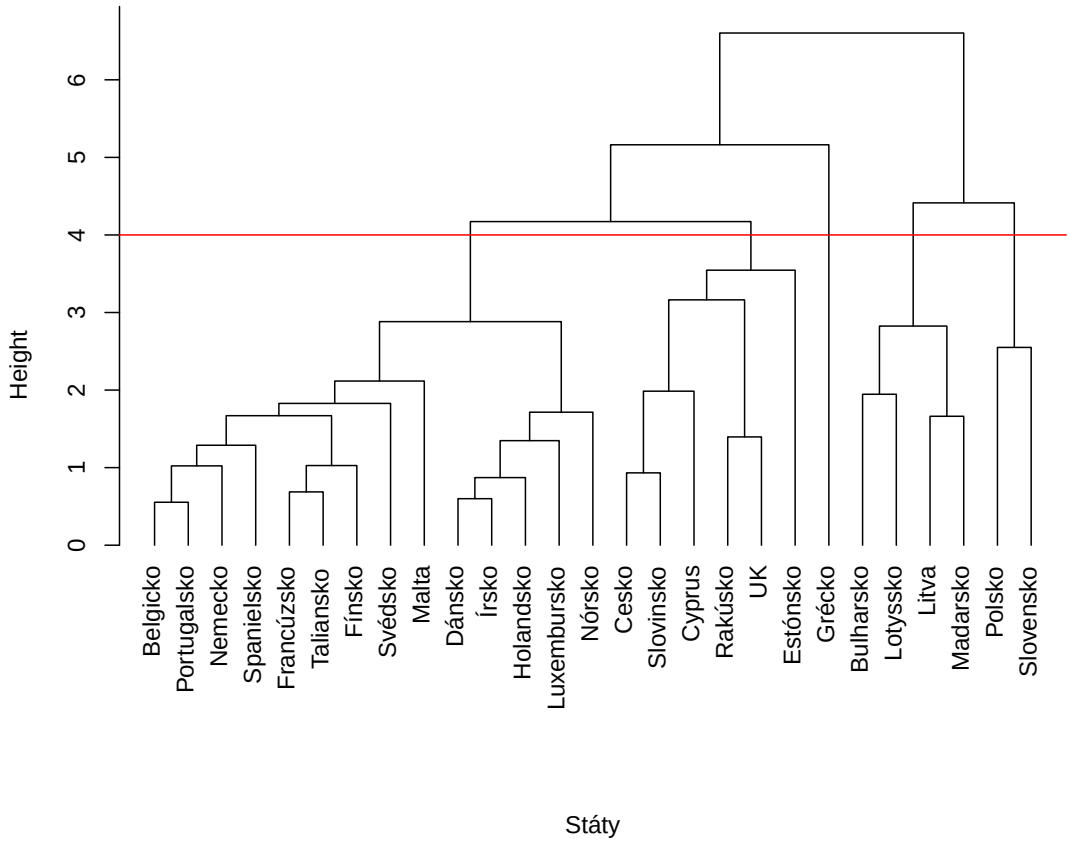


- Belgicko, Portugalsko, Nemecko, Španielsko, Francúzsko, Taliansko, Fínsko, Švédsko, Malta, Dánsko, Írsko, Holandsko, Luxembursko, Nórsko
- Česko, Slovinsko, Cyprus, Rakúsko, UK, Estónsko
- Grécko
- Bulharsko, Lotyšsko, Litva, Maďarsko
- Poľsko, Slovensko

A na záver K-means iteračnou schémou získavame rozdelenie do piatich zhlukov:

- Česko, Slovinsko, Cyprus, Rakúsko, UK,
- Estónsko, Litva, Maďarsko

Obr. 20: DIANA - Kvalita v roku 2005



- Belgicko, Portugalsko, Nemecko, Španielsko, Francúzsko, Taliansko, Fínsko, Švédsko, Malta, Dánsko, Írsko, Holandsko, Luxembursko, Nórsko
- Grécko
- Poľsko, Slovensko, Bulharsko, Lotyšsko

Zhrnutie:

Kvalitu života občanov jednotlivých krajín sme vymodelovali pomocou šiestich čŕt zohľadňujúcich finančné, pracovné a zdravotnícke aspekty života. Pre lepší náhľad na skutočné hodnoty sa v prílohe A nachádzajú údaje ešte v nepreškálovanom tvare. Preškálovanie je však v tomto prípade nevyhnutné, aby sme zamedzili dominantnosti niektorých čŕt, ktorá by v tomto prípade bola úplne neopodstatnená. Práve preto sa v tomto prípade nestačí pozerať iba na niektoré z nich.

Pri pohľade na výsledky klastrovej analýzy v oboch rokoch možno vidieť dosť veľkú zmenu. Približne spoločné ustálené zhluky sú:

1. Belgicko, Česko, Dánsko, Nemecko, Írsko
2. Bulharsko, Estónsko, Lotyšsko, Litva, Maďarsko, Poľsko, Slovensko
3. Grécko, Španielsko
4. Francúzsko, Taliansko, Cyprus
5. Luxembursko, Malta, Holandsko, Rakúsko, Portugalsko, Slovinsko, Fínsko, Švédsko, UK, Nórsko

Údaje v tabuľkách sú celkovo veľmi podobné, v niektorých prípadoch síce možno vidieť odôvodnenosť výsledkov clusteringu, často je však výsledné rozdelenie do zhlukov sprevádzané nejednoznačnosťou.

Podľa očakávania môžeme vidieť v roku 2012 zhlučené Grécko so Španielskom. Zaujímavý je prechod Španielska z 1. skupiny v roku 2005 do tohto zhluku, ktorý je spôsobený výrazným poklesom hodnôt v miere zamestnanosti a v oboch finančných črtách. Podobnosť štátov v druhej skupine spočíva v nízkej očakávanej dĺžke života, vyššej dojčenskej úmrtnosti a v hlavne v roku 2005 nízkej schopnosti čeliť nečakaným finančným nákladom. Prvá skupina sa vyznačuje najmä dobrou finančnou situáciou v oboch rokoch. Spojenie prvej a piatej skupine v roku 2005 sa dá tiež pekne vidieť v podobnosti vo finančných črtách a hlavne je badateľné v tomto smere výrazné zhoršenie piatej skupiny v roku 2012, stúpla zadĺženosť a klesla schopnosť obyvateľstva čeliť problémom s financiami. Štvrtá skupina tvorená tromi štátmi má vzájomne podobné hodnoty vo všetkých črtách, nepatria však nikdy k maximálnym alebo minimálnym hodnotám. Aj to zapríčiňuje zhlučenie tejto skupiny napríklad s 1. alebo 5. skupinou. V roku 2005 tvoria Česko, Cyprus, Rakúsko, Slovinsko a UK veľmi ustálenú skupinu, hlavne kvôli vysokej zamestnanosti, dlhšej pracovnej dobe a priemernej očakávanej dĺžke života. V roku 2012 k podobnému zhlučeniu nevedie ani clustering ani tabuľkové hodnoty.

V tejto kapitole bola zhluková analýza veľmi užitočná, nakoľko vytvorila ustálene zhluky, ktoré by sme porovnávaním tabuľkových hodnôt nezískali. Hlavnou príčinou je rovnocennosť všetkých čŕt, čo značne komplikuje možnú interpretáciu. Pozitívnou

stránkou tejto analýzy je aj, že výsledky sa stretávajú s naším očakávaním, podmienky pre život sú podobné v Strednej Európe a Pobaltí, ďalej v severských krajinách a krajinách Beneluxu a nakoniec je jasné vyčlenenie štátov najviac poznačených finančnou krízou.

Záver

Cieľom bakalárskej práce bolo predstaviť rôzne metódy zhlukovej analýzy a následne ich aplikovať na získané dáta týkajúce sa spotrebiteľov a štátov v Európe. Dokopy sme skúmali 4 odvetvia, a to distribúciu vládnych výdavkov, štruktúru vládnych príjmov, normovanú spotrebu domácností a kvalitu života. Výstupy[2] získané Wardovou, Diana a K-means metódou sme porovnali a zdôvodnili vstupnými dátami. Jednotlivé algoritmy neviedli k totožnému rozdeleniu, ale k veľmi podobnému. Zaujímavé bolo pozorovať zmenu zhlučenia podľa dát z roku 2005 a podľa dát z roku 2012.

Čo sa týka celkového výstupu, výsledky z rôznych analýz majú podľa očakávania určitú zhodu. Medzi homogénnejšie zhluky patria: 1. severské krajiny s Veľkou Britániou, Luxemburskom, Nemeckom, prípadne aj s Írskom, Francúzskom a Rakúskom, 2. Slovensko, Poľsko, Česko, Maďarsko, pobaltské štáty, 3. Taliansko, Španielsko, Portugalsko, Grécko, Cyprus. Existujú však aj štáty, ktoré každá analýza spojila s inými štátmi. Sem patria napríklad Belgicko, Holandsko, Bulharsko, Malta.

Priamo z takéhoto rozdelenia sa nedá odpovedať na nastolené otázky, nevieme povedať či severania sú šetrnejší ako južania, a či stredoeurópania sú menej spokojní so životom. Výstup z týchto metód je iba podobnosť krajín vrámci jedného zhluku a rozdielnosť vrámci rôznych. V kombinácii so vstupnými údajmi[3] majú však výsledky oveľa väčšiu výpovednú hodnotu[4].

Niektoré časti tejto práce boli taktiež spracované v odbornom článku [4].

[2] viď sekcie 3.1, 3.2, 3.3, 3.4

[3] viď príloha A

[4] viď zhrnutia v kapitole 3

Zoznam použitej literatúry

[1] Akademická rankingová a ratingová agentúra (ARRA), dostupné na Internete (14.04.2014): http://www.arra.sk/

[2] Arabie, P. - Hubert, L.J. - De Soete, G.: *Clustering and Classification*, World Scientific Publishing Co. Pte. Ltd., Singapore, 1996.

[3] Miklošovič, T.: osobná komunikácia, FMFI UK, Bratislava, 2013-2014.

[4] Miklošovič, T., Hlavatá, M.: *Zhluková analýza v kontexte európskych krajín*, Forum Statisticum Slovacum, 3/2014, Bratislava, 2014. s. 68-78, dostupné na Internete (28.05.2014): http://www.ssds.sk/casopis/archiv/2014/fss0314.pdf

[5] Sudipto, G. - Rajeev, R. - Kyuseok, S.: *CURE: An efficient clustering algorithm for large databases.*, In Proc. of 1998 ACM-SIGMOD Int. Conf. on Management of Data, 1998.

[6] Sudipto, G. - Rajeev, R. - Kyuseok, S.: *ROCK: A robust clustering algorithm for categorical attributes.*, In Proc. of the 15th Int'l Conf. on Data Eng., 1999.

[7] Štatistická databáza Eurostatu, dostupné na internete (16.05.2014): http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database

[8] Xu, R., Wunsch, D.C.II: *Clustering*, John Wiley&Sons, Inc., Hoboken, New Jersey, 2009.

Príloha A

Tabuľka A.1: Distribúcia vládnych prostriedkov v roku 2012

2012	SpV	O	VPaB	H	OŽP	BaOV	Z	RKaN	V	SP
Belgicko	14,7	1,8	3,4	12,6	1,2	0,7	14,7	2,4	11,5	37,0
Bulharsko	10,3	3,1	6,6	14,4	2,0	2,8	12,9	2,2	9,8	35,9
Česko	11,3	2,0	4,1	12,5	3,1	1,6	17,6	6,1	10,9	31,0
Dánsko	15,1	2,5	1,9	6,2	0,7	0,6	14,5	2,8	13,3	42,5
Nemecko	13,7	2,4	3,5	7,7	1,3	1,0	15,7	1,7	9,7	43,3
Estónsko	9,1	4,7	5,3	11,5	2,2	1,7	13,0	4,4	16,2	32,0
Írsko	13,7	1,0	3,9	8,4	1,9	2,0	16,7	1,9	12,2	38,4
Grécko	25,7	4,4	3,4	6,0	1,0	0,4	10,8	1,2	7,6	39,4
Španielsko	12,8	2,0	4,4	16,1	1,7	0,9	12,9	2,7	9,4	37,2
Francúzsko	10,5	3,4	3,2	6,5	1,9	3,4	14,6	2,6	10,8	43,1
Taliansko	18,0	2,7	3,8	6,7	1,8	1,3	14,5	1,4	8,2	41,5
Cyprus	27,2	4,2	4,7	6,9	0,7	5,0	7,1	2,4	14,6	27,2
Lotyšsko	12,4	2,4	4,9	14,5	2,0	3,3	10,6	4,0	15,0	30,9
Litva	11,9	2,8	5,1	9,1	2,5	0,7	16,5	2,3	15,5	33,6
Luxembursko	11,1	0,9	2,4	9,8	2,9	1,9	11,5	4,0	12,3	43,3
Maďarsko	18,6	1,7	4,0	12,8	1,5	1,9	10,8	3,9	9,8	35,1
Malta	15,4	1,7	3,4	12,0	3,4	0,9	13,4	2,1	13,7	33,9
Holandsko	10,8	2,5	4,1	10,5	3,3	1,2	17,7	3,4	11,5	34,9
Rakúsko	13,0	1,3	2,9	11,3	1,0	1,2	15,3	1,9	10,8	41,3
Poľsko	14,1	2,8	4,1	11,0	1,3	2,0	10,9	2,8	12,9	38,1
Portugalsko	19,2	2,4	3,8	5,9	1,0	1,2	12,8	2,0	11,9	39,6
Slovinsko	12,1	2,2	3,7	8,1	1,5	1,6	14,5	3,7	13,3	39,3
Slovensko	15,9	2,7	6,4	9,3	2,5	2,0	16,3	2,6	10,2	32,0
Fínsko	13,1	2,8	2,7	8,6	0,4	0,8	14,4	2,2	11,2	43,7
Švédsko	13,9	2,8	2,8	8,5	0,7	1,4	13,7	2,1	13,1	41,2
UK	12,0	4,9	5,0	5,8	1,9	1,7	16,6	2,1	12,6	37,5
Nórsko	9,2	3,3	2,2	9,8	1,6	1,5	16,9	2,8	12,6	40,1

Legenda: SpV=služby pre verejnosť, O=obrana, VPaB=verejný poriadok a bezpečnosť, H=hospodárstvo, OŽP=ochrana životného prostredia, BaOV=bývanie a občianska vybavenosť, Z=zdravotníctvo, RKaN=rekreácia kultúra a náboženstvo, V=vzdelávanie, SP=sociálna podpora

Tabuľka A.2: Distribúcia vládnych prostriedkov v roku 2005

2005	SpV	O	VPaB	H	OŽP	BaOV	Z	RKaN	V	SP
Belgicko	17,0	2,1	3,2	13,8	1,3	0,7	13,4	2,5	11,3	34,6
Bulharsko	16,3	5,7	7,2	11,4	1,9	1,8	12,8	2,0	11,5	29,5
Česko	12,0	3,7	4,9	15,2	2,6	3,6	16,0	2,7	10,7	28,6
Dánsko	12,8	2,8	1,9	5,8	1,1	1,0	13,6	3,1	13,9	44,0
Nemecko	12,9	2,2	3,4	7,8	1,2	2,2	14,3	1,7	8,7	45,5
Estónsko	8,5	4,1	6,4	11,8	2,7	0,6	12,1	6,8	17,8	29,1
Írsko	9,7	1,3	4,3	11,1	2,6	4,1	19,2	1,9	13,7	32,1
Grécko	22,5	6,6	3,6	8,1	1,3	0,8	13,6	0,8	8,8	33,9
Španielsko	12,3	2,9	4,7	12,5	2,2	2,3	14,8	3,6	11,1	33,6
Francúzsko	13,0	3,4	3,0	6,8	1,8	3,7	14,5	2,4	10,9	40,6
Taliansko	18,3	2,7	4,1	8,0	1,8	1,4	14,5	1,9	9,7	37,5
Cyprus	25,4	4,4	4,9	9,9	0,7	5,3	7,1	2,7	14,8	24,7
Lotyšsko	10,6	3,4	6,5	14,8	2,1	3,7	12,0	3,5	15,7	27,6
Litva	12,3	4,2	5,2	11,3	1,7	0,9	16,9	2,6	15,8	29,0
Luxembursko	10,9	0,6	2,5	10,6	2,7	1,6	12,5	5,4	11,4	41,8
Maďarsko	19,1	2,6	4,1	11,1	1,2	1,8	11,2	3,3	11,7	33,9
Malta	15,2	2,0	3,5	14,2	3,4	1,7	14,4	1,5	12,7	31,5
Holandsko	13,6	3,2	4,1	10,8	3,7	1,0	12,8	3,9	12,2	34,7
Rakúsko	14,1	1,8	3,0	10,0	1,0	1,2	15,3	1,9	10,5	41,2
Poľsko	14,4	2,3	3,9	8,9	1,4	3,4	10,2	2,4	14,0	39,0
Portugalsko	14,4	2,9	4,2	10,8	1,3	1,3	15,4	2,5	14,7	32,5
Slovinsko	12,8	3,0	3,9	8,2	1,8	1,2	14,0	2,9	14,8	37,4
Slovensko	16,0	4,3	5,4	9,9	1,7	2,0	12,8	2,8	10,4	34,8
Fínsko	13,3	3,1	2,7	9,4	0,6	0,6	13,6	2,2	12,2	42,2
Švédsko	13,9	3,2	2,5	8,0	0,7	1,6	12,5	1,9	13,0	42,7
UK	10,5	5,6	5,9	6,7	1,5	2,4	15,7	2,4	13,6	35,6
Nórsko	10,1	3,6	2,1	9,1	1,4	1,4	17,3	2,5	13,6	38,8

Legenda: SpV=služby pre verejnosť, O=obrana, VPaB=verejný poriadok a bezpečnosť, H=hospodárstvo, OŽP=ochrana životného prostredia, BaOV=bývanie a občianska vybavenosť, Z=zdravotníctvo, RKaN=rekreácia kultúra a náboženstvo, V=vzdelávanie, SP=sociálna podpora

Tabuľka A.3: Vládny príjem z daní v roku 2012

2012	DPH+I	DzPoDPH+I	iDzP	DzP	iBD	DzK	DzSP
Belgicko	16,2	8,2	4,4	35,2	1,6	1,9	32,5
Bulharsko	33,8	19,3	1,9	17,6	0,5	0,9	26,1
Česko	25,3	7,8	1,4	20,4	0,3	0	44,9
Dánsko	20,6	9,6	4,4	57,1	5,9	0,4	2,0
Nemecko	20,1	6,7	1,8	29,9	0,7	0,4	40,4
Estónsko	40,6	0,6	2,0	21,1	0	0	35,7
Írsko	28,3	5,2	5,1	42,7	1,6	1,6	15,4
Grécko	21,0	14,3	1,9	24,5	5,7	0,2	32,3
Španielsko	16,4	9,4	5,7	30,0	1,0	1,1	36,2
Francúzsko	15,3	9,0	10,1	23,9	2,7	1,1	37,9
Taliansko	13,6	11,5	9,0	33,3	1,3	0,2	31,0
Cyprus	25,8	10,6	6,0	29,5	2,0	0	26,1
Lotyšsko	25,3	12,5	3,8	26,4	1,4	0	30,5
Litva	28,3	11,3	1,6	17,9	0,2	0	40,7
Luxembursko	26,5	2,1	4,1	35,4	2,0	0,4	29,5
Maďarsko	23,8	19,7	3,3	17,2	0,7	1,3	34,0
Malta	23,1	15,8	1,3	39,2	1,7	0,7	18,2
Holandsko	19,6	6,7	3,2	25,5	2,9	0,6	41,4
Rakúsko	18,3	8,1	7,5	29,7	1,5	0	34,9
Poľsko	23,0	12,1	4,7	20,7	1,6	0,1	37,9
Portugalsko	27,1	12,3	3,1	27,2	1,6	0,5	28,2
Slovinsko	21,7	13,8	2,8	19,1	1,7	0,1	40,8
Slovensko	21,3	10,6	3,4	18,8	1,1	0	44,8
Fínsko	20,5	11,7	0,6	34,6	1,9	0,6	30,1
Švédsko	20,8	7,0	13,9	40,8	0,5	0	17,0
UK	19,7	12,2	4,8	34,4	6,8	0,6	21,5
Nórsko	18,4	7,4	1,2	48,3	1,8	0,2	22,7

Legenda: DPH+I=daň z pridanej hodnoty a importu, DzPoDPH+I=dane z produktov okrem DPH a importu, iDzP=iné dane z produkcie, DzP=daň z príjmu, iBD=iné bežné dane, DzK=daň z kapitálu, DzSP=dane zo sociálnych príspevkov

Tabuľka A.4: Vládny príjem z daní v roku 2005

2005	DPH+I	DzPoDPH+I	iDzP	DzP	iBD	DzK	DzSP
Belgicko	16,3	9,1	4,0	36,5	1,8	1,4	30,9
Bulharsko	38,5	12,3	2,3	14,7	0,2	0,9	31,1
Česko	21,7	8,3	1,2	24,8	0,3	0,1	43,7
Dánsko	19,8	11,7	3,4	56,8	5,6	0,4	2,3
Nemecko	17,7	7,9	1,8	27,5	0,9	0,5	43,7
Estónsko	40,3	0,7	2,1	23,1	0	0	33,8
Írsko	30,2	10,9	3,1	39,0	0,9	0,5	15,4
Grécko	21,4	14,6	0,9	25,5	1,4	0,6	35,5
Španielsko	17,6	13,7	3,1	29,5	1,1	1,3	33,6
Francúzsko	16,6	9,0	9,8	23,8	2,1	1,2	37,5
Taliansko	14,6	12,1	8,7	32,2	1,0	0,3	31,1
Cyprus	28,9	12,5	5,4	24,6	2,1	2,7	23,8
Lotyšsko	26,7	13,5	3,0	26,8	0,7	0	29,3
Litva	24,3	11,9	1,8	30,9	0,1	0	30,8
Luxembursko	27,3	2,9	5,0	34,6	1,8	0,4	28,0
Maďarsko	22,4	17,7	1,7	23,4	0,8	0,3	33,7
Malta	24,4	18,7	1,7	32,4	2,5	1,1	19,2
Holandsko	21,7	8,9	2,9	27,6	3,1	0,9	34,9
Rakúsko	18,6	8,7	7,3	28,9	1,5	0,1	34,9
Poľsko	24,2	12,1	5,0	19,5	1,6	0,1	37,4
Portugalsko	27,5	16,8	2,2	25,3	1,0	0,1	27,0
Slovinsko	22,1	10,7	7,5	21,8	0,8	0,1	37,1
Slovensko	25,0	12,3	2,5	18,1	1,2	0	40,9
Fínsko	19,5	11,5	0,6	38,4	1,6	0,7	27,7
Švédsko	18,1	7,3	8,0	44,1	0,9	0,1	21,5
UK	17,7	12,4	4,3	37,4	6,4	0,7	21,2
Nórsko	18,3	8,8	1,2	49,4	1,8	0,2	20,4

Legenda: DPH+I=daň z pridanej hodnoty a importu, DzPoDPH+I=dane z produktov okrem DPH a importu, iDzP=iné dane z produkcie, DzP=daň z príjmu, iBD=iné bežné dane, DzK=daň z kapitálu, DzSP=dane zo sociálnych príspevkov

Tabuľka A.5: Spotreba priemerných domácností v roku 2005

2005	J+N	A+T+N	O+O	E	ZaUD	Z	D	K	R+K	V	R+H	P
Belgicko	145,63	23,73	50,70	272,92	60,41	50,70	139,16	31,28	102,48	5,39	67,96	49,62
Bulharsko	324,07	39,09	31,89	357,00	30,86	44,24	51,44	47,33	29,84	5,14	37,04	2,06
Česko	216,39	30,46	58,82	211,13	70,38	21,01	116,60	48,32	111,34	5,25	53,57	56,72
Dánsko	124,87	34,63	51,42	313,75	64,01	28,33	144,81	25,18	119,62	4,20	41,97	47,22
Nemecko	116,55	17,69	49,95	308,01	56,19	37,46	138,40	30,18	115,50	8,32	44,75	77,00
Estónsko	233,64	29,08	57,11	310,49	54,00	27,00	103,84	57,11	66,46	13,50	32,19	15,58
Írsko	130,16	59,26	53,97	247,62	76,19	26,46	122,75	35,98	106,88	20,11	63,49	57,14
Grécko	166,13	36,44	75,03	257,23	66,45	63,24	111,47	40,73	45,02	25,72	92,18	20,36
Španielsko	185,57	23,71	71,13	312,37	48,45	22,68	108,25	27,84	65,98	11,34	95,88	26,80
Francúzsko	142,40	24,44	70,14	279,49	64,82	44,63	143,46	35,07	73,33	6,38	48,88	66,95
Taliansko	194,36	18,81	73,15	309,30	60,61	40,75	124,35	22,99	60,61	7,31	52,25	35,53
Cyprus	159,28	20,04	81,22	227,85	62,24	49,58	154,01	35,86	63,29	42,19	87,55	16,88
Lotyšsko	303,22	32,19	75,80	177,57	54,00	38,42	113,19	60,23	65,42	14,54	55,04	10,38
Litva	350,62	36,31	81,95	196,06	43,57	48,76	84,02	47,72	44,61	11,41	47,72	7,26
Luxembursko	97,79	17,88	67,30	316,51	74,66	27,34	170,35	23,13	78,86	4,21	83,07	38,91
Maďarsko	233,95	37,27	51,76	200,83	48,65	42,44	145,96	67,29	87,99	8,28	33,13	42,44
Malta	223,97	28,39	87,28	95,69	112,51	31,55	174,55	30,49	106,20	12,62	74,66	22,08
Holandsko	109,95	21,99	60,73	268,06	67,02	13,61	114,14	32,46	114,14	10,47	58,64	128,80
Rakúsko	136,55	29,41	58,82	234,24	65,13	32,56	169,12	27,31	132,35	8,40	57,77	48,32
Poľsko	264,52	25,93	47,72	326,76	46,68	47,72	84,02	49,79	65,35	13,49	17,63	10,37
Portugalsko	161,80	24,01	42,80	277,66	50,10	63,67	134,66	31,32	59,50	17,75	112,73	24,01
Slovinsko	173,78	24,97	72,84	239,33	60,35	15,61	162,33	41,62	97,81	8,32	44,75	58,27
Slovensko	255,74	29,23	58,46	317,33	43,84	29,23	86,64	44,89	62,63	8,35	45,93	17,75
Fínsko	139,87	26,43	41,85	299,56	56,17	38,55	172,91	30,84	123,35	2,20	46,26	22,03
Švédsko	118,24	21,40	51,80	333,33	66,44	25,90	146,40	31,53	137,39	0,00	39,41	28,15
UK	103,56	25,10	52,30	309,62	67,99	12,55	141,21	28,24	128,66	14,64	83,68	32,43
Nórsko	122,64	32,49	58,70	274,63	68,13	31,45	189,73	27,25	128,93	3,14	39,83	23,06

Legenda: J+N=jedlo a nealkoholické nápoje, A+T+N=alkoholické nápoje, tabak a narkotiká, O+O=oblečenie a obuv, E=energie, ZaUD=zariaďovanie a údržba domácnosti, Z=zdravie, D=doprava, K=komunikácie, R+K=rekreácia a kultúra, V=vzdelanie, R+H=reštaurácie a hotely, P=poistenie

Tabuľka A.6: Kvalita života jednotlivcov v roku 2012

2012	ODŽ	MZ	NČP	SČNFN	PZD	PPD
Belgicko	80,50	92,40	91,40	74,60	996,20	41,43
Bulharsko	74,40	87,70	90,40	73,60	992,20	41,15
Česko	78,10	93,00	89,40	72,60	997,40	42,13
Dánsko	80,20	92,50	88,40	71,60	996,60	38,75
Nemecko	81,00	94,50	87,40	70,60	996,70	41,85
Estónsko	76,70	90,00	86,40	69,60	996,40	40,88
Írsko	80,90	85,30	85,40	68,60	996,50	39,75
Grécko	80,70	75,70	84,40	67,60	997,10	43,85
Španielsko	82,50	75,00	83,40	66,60	996,90	41,60
Francúzsko	82,10	90,20	82,40	65,60	996,50	41,08
Taliansko	82,40	89,30	80,40	63,60	997,10	40,33
Cyprus	81,10	88,10	79,40	62,60	996,50	42,08
Lotyšsko	74,10	85,00	78,40	61,60	993,70	40,48
Litva	74,10	86,60	77,40	60,60	996,10	39,60
Luxembursko	81,50	94,90	76,40	59,60	997,50	40,78
Maďarsko	75,30	89,10	75,40	58,60	995,10	40,60
Malta	80,90	93,60	74,40	57,60	994,70	41,48
Holandsko	81,20	94,70	73,40	56,60	996,30	40,85
Rakúsko	81,10	95,70	72,40	55,60	996,80	43,48
Poľsko	76,90	89,90	71,40	54,60	995,40	42,30
Portugalsko	80,60	84,10	70,40	53,60	996,60	42,58
Slovinsko	80,30	91,10	68,40	51,60	998,40	41,90
Slovensko	76,30	86,00	67,40	50,60	994,20	41,65
Fínsko	80,70	92,30	66,40	49,60	997,60	40,15
Švédsko	81,80	92,00	65,40	48,60	997,40	40,85
UK	81,00	92,10	64,40	47,60	995,90	42,80
Nórsko	81,50	96,80	62,40	45,60	997,50	39,05

Legenda: ODŽ=očakávaná dĺžka života, MZ=miera zamestnanosti, NČP=nezadĺžená časť populácie, SČNFN=schopnosť čeliť nečakaným finančným nákladom, PZD=počet zdravých dojčeniec, PPD=priemerná pracovná doba

Tabuľka A.7: Kvalita života jednotlivcov v roku 2005

2012	ODŽ	MZ	NČP	SČNFN	PZD	PPD
Belgicko	79,10	91,50	92,50	77,00	996,30	41,15
Bulharsko	72,50	89,90	77,90	23,20	989,60	41,48
Česko	76,10	92,10	89,00	57,10	996,60	42,83
Dánsko	78,30	95,20	93,30	75,50	995,60	40,33
Nemecko	79,40	88,70	94,50	75,30	996,10	41,53
Estónsko	73,00	92,10	88,50	65,00	994,60	41,35
Írsko	79,00	95,60	90,60	77,50	996,20	40,60
Grécko	79,50	90,10	66,90	61,20	996,20	44,20
Španielsko	80,30	90,80	92,80	65,30	996,30	42,28
Francúzsko	80,30	91,10	89,30	64,40	996,20	40,98
Taliansko	80,90	92,30	87,30	72,10	996,20	41,20
Cyprus	78,70	94,70	78,20	56,50	995,40	42,30
Lotyšsko	70,60	90,00	75,20	29,00	992,30	42,70
Litva	71,20	91,50	78,50	34,40	992,90	39,43
Luxembursko	79,60	95,40	96,00	78,60	997,40	40,90
Maďarsko	73,00	92,80	82,40	42,70	993,80	40,98
Malta	79,40	93,10	90,80	66,20	994,60	41,48
Holandsko	79,60	94,70	94,30	74,40	995,10	40,73
Rakúsko	79,50	94,80	97,00	75,10	995,80	44,30
Poľsko	75,00	82,10	73,30	37,40	993,60	43,23
Portugalsko	78,20	91,40	92,90	81,30	996,50	41,63
Slovinsko	77,50	93,50	84,60	57,10	995,90	42,88
Slovensko	74,10	83,60	88,50	40,70	992,80	41,48
Fínsko	79,10	91,60	88,60	67,00	997,00	40,45
Švédsko	80,70	92,30	90,80	85,90	997,60	40,98
UK	79,20	95,20	90,60	69,20	994,90	43,15
Nórsko	80,30	95,50	88,80	70,50	996,90	39,33

Legenda: ODŽ=očakávaná dĺžka života, MZ=miera zamestnanosti, NČP=nezadĺžená časť populácie, SČNFN=schopnosť čeliť nečakaným finančným nákladom, PZD=počet zdravých dojčeniec, PPD=priemerná pracovná doba