Grupiranje država prema njihovoj mjeri sreće klasterskom analizom

Karačić, Lucija

Grupiranje država prema njihovoj mjeri sreće klasterskom analizom

Karačić, Lucija

SVEUČILIŠTE U ZAGREBU

PRIRODOSLOVNO–MATEMATIČKI FAKULTET

MATEMATIČKI ODSJEK

Lucija Karačić

GRUPIRANJE DRŽAVA PREMA NJIHOVOJ MJERI SREĆE KLASTERSKOM ANALIZOM

Diplomski rad

Voditelj rada:
prof. dr. sc. Anamarija Jazbec

Zagreb, prosinac 2020.

Ovaj diplomski rad obranjen je dana _____ pred ispitnim povjerenstvom u sastavu:

1. _____, predsjednik
2. _____, član
3. _____, član

Povjerenstvo je rad ocijenilo ocjenom _____.

Potpisi članova povjerenstva:

1. _____
2. _____
3. _____

“Ištite i dat će vam se! Tražite i naći ćete! Kucajte i otvorit će vam se! Doista, tko god ište, prima; i tko traži, nalazi; i onomu koji kuca otvorit će se.” (Mt 7,7-8)

Hvala mojoj obitelji na podršci i razumijevanju. Hvala mojim prijateljima za sve trenutke bodrenja i razveseljivanja. Hvala prof. dr. sc. Anamariji Jazbec na pomoći i savjetima. Hvala mojem Jakovu na beskonačnoj potpori, strpljenju i ljubavi.
Bez vas bi sve ovo bilo mnogo teže.

Sadržaj

Uvod

Živimo u svijetu punom podataka. Svaki dan ljudi se susreću s novim informacijama koje interpretiraju kao podatke za daljnju analizu. Kako bi naučio novi pojam, objekt ili fenomen, čovjek traži obilježja koja ga opisuju kako bi ga dalje mogao usporediti s nekim poznatim pojmom, objektom ili fenomenom na temelju sličnosti ili različitosti. Stoga je klasifikacija jedna od osnovnih sposobnosti čovjeka koju primjenjuje u svim životnim aspektima.

Za razliku od klasifikacije gdje je unaprijed poznata struktura grupa, a cilj je rasporediti podatke u postojuće grupe, klasterska analiza grupira elemente u klastere tako da su elementi unutar svakoga klastera najsličniji mogući. Danas se klasterska analiza primjenjuje u raznim područjima, poput biologije, medicine, marketinga, psihologije. Njezinom primjenom štedimo na vremenu, a dobiveni rezultati prikladni su za algoritme vezane uz daljnju obradu podataka.

Glavna tema ovog rada jest klasterska analiza, a podijeljena je u dva poglavlja. U prvom poglavlju susrest ćemo se s osnovnim pojmovima, mjerama sličnosti i udaljenosti te algoritmima klasteriranja. U drugom poglavlju koristeći bazu koja sadržava podatke o 146 država svijeta iz 2017. godine, hijerarhijskom i nehijerarhijskom klasterskom analizom analizirat ćemo promatrane države koje su najsličnije u odnosu prema promatranim varijablama.

Poglavlje 1

Klasterska analiza

1.1 Opis metode

Klasterska analiza jedna je od metoda multivarijatne analize koja se temelji na načelima multivarijatne statistike, odnosno na promatranju i analizi dviju ili više varijabli istodobno. Klasterska analiza ima razne ciljeve, ali se svi svode na to da za odabrani skup podataka S odredi podskupove $C_i, i \in \mathbb{N}$ gdje podatci u istom podskupu trebaju biti međusobno slični, a uzorci iz različitih podskupova ne bi trebali biti [10]. Podskupove $C_i, i \in \mathbb{N}$ dobivene klasterskom analizom koji su homogeni i/ili dobro separirani nazivamo klasteri. Skup svih klastera $C_i, i \in \mathbb{N}$ čini klastering $\mathbf{C}$ za dani skup podataka S. Prema [13], najčešće se koriste hijerarhijski i particijski klastering:

- klastering $\mathbf{C} = \{C_1, C_2, \ldots, C_k\}$, $k \in \mathbb{N}$ je hijerarhija ako za $\forall\, C_i, C_j \in \mathbf{C}, i, j = 1, 2, \ldots, k$ vrijedi
$$C_i \cap C_j \in \{C_i, C_j, \emptyset\}. \tag{1.1}$$
- klastering $\mathbf{C} = \{C_1, C_2, \ldots, C_k\}$, $k \in \mathbb{N}$ je particija ako vrijede sljedeća svojstva:
$$C_i \neq \emptyset, i = 1, 2, \ldots, k \tag{1.2}$$
$$\bigcup_{i=1}^{k} C_i = S \tag{1.3}$$
$$C_i \cap C_j = \emptyset, i, j = 1, 2, \ldots, k, i \neq j \tag{1.4}$$

Kod particijskoga klasteringa svaki podatak iz skupa S pripada točno jednom klasteru C_i. Međutim, podatak može biti član svih klastera s određenom vjerojatnošću. Ovakvu vrstu klasteringa nazivamo meki klastering (engl. *fuzzy clustering*). Postupak klasterske analize sastoji se od pet osnovnih koraka:

1. **Početna analiza**
   Na početku istraživanja za dani skup podataka moramo definirati cilj klasterske analize. Nakon toga analiziramo podatke: uklanjamo neispravne vrijednosti, odabiremo varijable koje su povezane s ciljem klasterske analize koji se želi postići, radimo standardizaciju varijabli ako je potrebno, itd.

2. **Odabir mjere sličnosti ili udaljenosti**
   Ovisno o našim podatcima odabiremo odgovarajuću mjeru sličnosti ili različitosti. Gotovo svi algoritmi klasteriranja eksplicitno su ili implicitno povezani s odabranom mjerom.

3. **Odabir ili kreiranje algoritma**
   Ovisno o cilju klasterske analize, odabiremo ili kreiramo pripadni algoritam za klasteriranje.

4. **Validacija**
   Različiti pristupi daju nam različita rješenja; čak i za iste algoritme možemo dobiti različita rješenja (npr. promjenom redoslijeda ulaznih podataka). Stoga je bitno provesti validaciju, proces vrednovanja rezultata klasteriranja objektivno i kvantitativno.

5. **Interpretacija**
   Detaljno proučavamo svaki klaster na temelju njegovih obilježja i deskriptivne statistike, a rade se i daljnje analize ovisno o dobivenim informacijama iz klastera.

1.2 Mjere sličnosti i udaljenosti

Za zadani skup podataka S odaberimo dva proizvoljna objekta $x, y \in S$. Svaki od njih ima oblik $x = (x_1, x_2, \ldots, x_n)$ i $y = (y_1, y_2, \ldots, y_n)$, $n \in \mathbb{N}$, gdje su varijable x_i, y_i ,$\forall i \in \{1, 2, \ldots, n\}$ atributi objekata x i y.

Klasterskom analizom želimo pronaći grupe među podatcima tako da su objekti unutar grupe međusobno slični, a različiti od objekata iz drugih grupa. Dva objekta su 'blizu' ako im je sličnost velika, a udaljenost (različitost) mala. Želimo pronaći način prema kojem ćemo odrediti udaljenost ili sličnost između dva objekta. Određivanje prikladne mjere između dva objekta skupa S svodi se na određivanje mjere između njihovih atributa (varijabli), stoga će odabir mjere ovisiti o vrsti podataka. Prema [13], udaljenost je funkcija koja svakom paru objekata (x, y) iz S pridružuje realan broj, tj. $d : (x, y) \to \mathbb{R}$. te mora zadovoljavati sljedeće uvjete:

- $d(x, y) \geq 0$ (nenegativnost),

- $d(x, x) = 0$,
- $d(x, y) = d(y, x)$ (simetričnost).

Ako dodatno vrijedi i:

- $d(x, y) = 0 \Leftrightarrow x = y$,
- $(\forall z \in S)\ d(x, y) \leq d(x, z) + d(z, y)$ (nejednakost trokuta),

kažemo da je d metrika.

Prema [13], na sličan način definiramo i sličnost, funkciju koja svakom paru objekata (x, y) iz S pridružuje realan broj, tj. $s : (x, y) \to \mathbb{R}$ te mora zadovoljavati sljedeće uvjete:

- $0 \leq s(x, y) \leq 1$,
- $s(x, x) = 1$,
- $s(x, y) = s(y, x)$ (simetričnost).

Uglavnom mjere udaljenosti koristimo za numeričke varijable, a mjere sličnosti za kategorijske varijable.

1.2.1 Numeričke varijable

Numeričke su varijable kvantitativne, tj. poprimaju vrijednosti iz skupa realnih brojeva $\mathbb{R}$. Dijelimo ih na diskretne i neprekidne; diskretne mogu poprimiti samo konačno ili prebrojivo mnogo vrijednosti, a neprekidne poprimaju vrijednost iz cijelog $\mathbb{R}$ ili iz nekog intervala realnih brojeva. Primjer diskretnih varijabli jest broj djece u obitelji ili broj učenika na nastavi, a težina ili temperatura učenika primjer su neprekidnih varijabli.

Ako varijable imaju različite mjerne jedinice ili veliku standardnu devijaciju (varijabilnost), potrebno je napraviti standardizaciju. Neka su $x^{(1)}, x^{(2)}, \ldots, x^{(n)} \in S$ proizvoljni objekti oblika $x^{(j)} = (x_1^{(j)}, x_2^{(j)}, \ldots, x_k^{(j)})$ i $1 \leq k \leq n$ proizvoljan. Prema [11], postoji nekoliko različitih mogućnosti za standardizaciju atributa, ali dvije najčešće korištene su:

1. z-vrijednost (engl. *z-score*)

$$z_k^{(j)} = \frac{x_k^{(j)} - \mu_k}{\sigma_k}, 1 \leq j \leq n, \tag{1.5}$$

gdje je μ_k aritmetička sredina i σ_k standardna devijacija, za $\forall k$

2. jedinični interval (engl. *unit interval*)

$$z_k^{(j)} = \frac{x_k^{(j)} - \min x^{(j)}}{\max x^{(j)} - \min x^{(j)}}, 1 \leq j \leq n. \tag{1.6}$$

Neka su $x, y \in S$, $x = (x_1, x_2, \ldots, x_n), y = (y_1, y_2, \ldots, y_n), n \in \mathbb{N}$ i za $\forall i \in \{1, 2, \ldots, n\}$ x_i, y_i su numeričke varijable. Prema [3], razlikujemo sljedeće mjere udaljenosti:

- **Minkowski udaljenost**

$$d(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}, \ p > 0, \tag{1.7}$$

- **Euklidska udaljenost**

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}, \tag{1.8}$$

- **Manhattan udaljenost**

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|, \tag{1.9}$$

- **Prosječna udaljenost**

$$d(x, y) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)^2}, \tag{1.10}$$

- **Čebiševljeva udaljenost**

$$d(x, y) = \max_{1 \leq i \leq n} |x_i - y_i|, \tag{1.11}$$

- **Mahalanobis udaljenost**

$$d(x, y) = (x - y)^T S^{-1} (x - y), \tag{1.12}$$

  pri čemu je S kovarijacijska matrica.

Primijetimo, Manhattanova udaljenost poseban je slučaj minkowski udaljenosti za $p = 1$, kao i euklidska za $p = 2$ i Čebiševljeva za $p \to +\infty$. Euklidska i Manhattan udaljenost jesu metrike. Minkowski, Manhattanova i euklidska udaljenosti su osjetljive na netipične vrijednosti (engl. *outliers*), vrijednosti koje su jako različite od ostalih. Problem rješavamo standardizacijom podataka.

1.2.2 Kategorijske varijable

Kategorijske varijable vrste su podataka koje se mogu podijeliti u kategorije; vrijednosti koje poprimaju jesu imena ili oznake. Dijelimo ih na nominalne i ordinalne.

Među kategorijama nominalnih varijabli ne možemo uspostaviti prirodni poredak. Primjer nominalnih varijabli jest boja očiju (smeđa, zelena, plava, siva...) ili krvna grupa (A, B, AB, 0). Mjera sličnosti između objekata x i y, čiji su atributi nominalni, dana je sa

$$s(x, y) = \frac{m}{p}, \tag{1.13}$$

pri čemu je m broj poklapanja (broj atributa za koje x i y imaju iste vrijednosti,), a p ukupan broj atributa. Alternativno, mjeru udaljenosti možemo izračunati kao

$$d(x, y) = 1 - s(x, y) \tag{1.14}$$

Kod ordinalnih varijabli među kategorijama možemo uspostaviti prirodni poredak. Primjer ordinalnih varijabli jesu školske ocjene (nedovoljan, dovoljan, dobar, vrlo dobar, odličan) ili dobne kategorije sportaša (početnici, mlađi kadeti, kadeti, mlađi juniori, juniori, mlađi seniori, seniori, veterani). Neka je i proizvoljni ordinalni atribut, a M_i broj stanja koja može poprimiti. Prema [9, str. 74], računanje sličnosti između ordinalnih varijabli uključuje sljedeće korake:

1. Budući da su stanja M_i uređena, svaku vrijednost supstituiramo s odgovarajućim rangom $r_i \in \{1, 2, \ldots, M_i\}$.

2. Budući da svaki ordinalni atribut može imati različit broj stanja, potrebno je svako stanje prebaciti u interval vrijednosti $[0.0, 1.0]$. Takvu normalizaciju podataka dobivamo definiranjem
$$z_i^{(j)} = \frac{r_i^{(j)} - 1}{M_i - 1}. \tag{1.15}$$

3. Mjeru udaljenosti izračunamo tako da bilo koju mjeru udaljenosti za numeričke varijable primijenimo na $z_i^{(j)}$. Sličnost možemo izračunati preko (1.14).

1.2.3 Binarne varijable

Binarne ili dihotomne varijable jesu varijable koje imaju točno dva različita stanja, a to su: *laž* ili *istina* (koje uglavnom označavamo s 0 i 1). Kad obrađujemo binarne podatke, vrlo je bitno što su vrijednosti 0 i 1. Stoga, na temelju značenja vrijednosti 0, binarne varijable dijelimo na simetrične i asimetrične.

Kažemo da je binarna varijabla simetrična ako *istina* označava prisutnost određenog atributa, a *laž* prisutnost drugog atributa. Varijabla spol sastoji se od dvaju stanja: muško i žensko. U ovom slučaju je svejedno koje stanje označimo s 0, a koje s 1. Kod asimetričnih varijabli *istina* označava prisutnost određenog atributa, a *laž* odsutnost istog atributa. Kao primjer pogledajmo rezultat SARS-Cov-2 testa: 1 je Covid pozitivna osoba, a 0 Covid negativna.

Neka su $x, y \in S$, $x = (x_1, x_2, \ldots, x_n), y = (y_1, y_2, \ldots, y_n), n \in \mathbb{N}$ i za $\forall i \in \{1, 2, \ldots, n\}$ x_i, y_i binarne varijable. Zbog definicije i interpretacije mjera sličnosti binarne varijable prikazujemo preko tablice frekvencija:

Tablica 1.1: Tablica frekvencija

(x, y)	1	0	$\sum$
1	a	b	$a + b$
0	c	d	$c + d$
$\sum$	$a + c$	$b + d$	n

pri čemu je $n = a + b + c + d$ i za $\forall i \in \{1, 2, \ldots, n\}$:

- a = broj varijabli takvih da je $x_i = y_i = 1$
- d = broj varijabli takvih da je $x_i = y_i = 0$
- b = broj varijabli takvih da je $x_i = 1, y_i = 0$
- c = broj varijabli takvih da je $x_i = 0, y_i = 1$

Mjere sličnosti za simetrične varijable uzimaju u obzir slučaj (0, 0), a asimetrične ne. Prema [7], najčešće korištene mjere su:

- **Simple Matching koeficijent**

$$s(x, y) = \frac{a + d}{n} \tag{1.16}$$

- **Jaccard koeficijent**

$$s(x, y) = \frac{a}{a + b + c} \tag{1.17}$$

- **Rogers-Tanimoto koeficijent**

$$s(x, y) = \frac{a + d}{a + 2(b + c) + d} \tag{1.18}$$

- **1. Gower-Legendre koeficijent**

$$s(x,y) = \frac{a+d}{a+\frac{1}{2}(b+c)+d} \tag{1.19}$$

- **2. Gower-Legendre koeficijent**

$$s(x,y) = \frac{a}{a+\frac{1}{2}(b+c)} \tag{1.20}$$

- **1. Sneath-Sokal koeficijent**

$$s(x,y) = \frac{a}{a+2(b+c)} \tag{1.21}$$

- **2. Sneath-Sokal koeficijent**

$$s(x,y) = \frac{2(a+d)}{2a+b+c+2d} \tag{1.22}$$

1.2.4 Mješovite varijable

U praksi klastersku analizu treba primijeniti na različitim vrstama podataka pa se samim time nameće pitanje kako izračunati mjeru sličnosti za takvu vrstu podataka. Prema [7, str. 54], jedna od mogućnosti je da skaliramo sve podatke tako da svi budu na istoj skali, tj. zamijenimo vrijednosti varijabli njihovim pozicijama u objektima. Nakon toga primijenimo neku od mjera udaljenosti za numeričke varijable.

Neka su $x, y \in S$, $x = (x_1, x_2, \ldots, x_n), y = (y_1, y_2, \ldots, y_n), n \in \mathbb{N}$ i za $\forall i \in \{1, 2, \ldots, n\}$ x_i, y_i varijable različitog tipa. Prema [7], 1971. godine Gower predlaže mjeru sličnosti

$$s(x,y) = \frac{1}{n}\sum_{i=1}^{n} \delta(x_i, y_i), \tag{1.23}$$

pri čemu je $\delta(x_i, y_i)$ mjera sličnosti:

- za binarne i nominalne varijable

$$\delta(x_i, y_i) = \begin{cases} 1, & x_i = y_i \\ 0, & x_i \neq y_i \end{cases}. \tag{1.24}$$

- za numeričke i ordinalne varijable

$$\delta(x_i, y_i) = 1 - \frac{|x_i - y_i|}{r_i}, \tag{1.25}$$

  pri čemu je r_i raspon varijable i.

Kao i kod kategorijskih varijabli, mjeru udaljenosti možemo izračunati preko (1.14).

1.3 Algoritmi klasteriranja

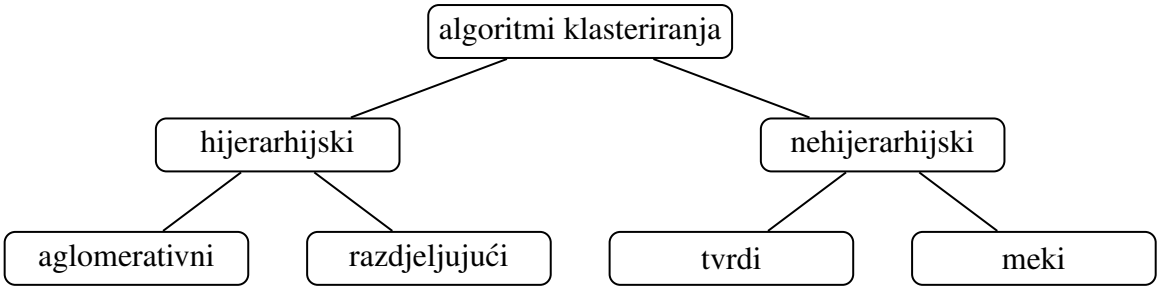
Odabir optimalnog algoritma ovisi o ciljevima analize i strukturi podataka. Primjerice, jedan algoritam može biti učinkovit na nekom skupu podataka, ali na skupu koji ima drukčiju strukturu, veličinu ili dimenziju nije učinkovit. Prema [9, str. 446], dobar algoritam trebao bi imati sljedeća obilježja:

- **skalabilnost**
  Mnogo se algoritama izvodi nad manjim skupovima podataka s manje od sto objekata. S obzirom na to da u praksi radimo s velikim bazama podataka gdje imamo ogroman broj objekata, bitna je funkcionalnost pod tim uvjetima.
- **sposobnost klasteriranja podataka velikih dimenzija**
  Susrećemo se sa skupovima podataka koji imaju velik broj varijabli, a većina algoritama radi s malim brojem, što je velik izazov.
- **minimalni zahtjevi nad ulaznim parametrima**
  Razni algoritmi klasteriranja traže osnovne informacije u obliku ulaznih parametara, poput željenog broja klastera. Poželjno je da algoritam prima samo nužne parametre kako bi što manje utjecali na konačni rezultat.
- **dobro podnošenje devijacija**
  Devijacije definiramo kao objekte koji odstupaju od generalnog ponašanja podataka i odnose se kao netipične vrijednosti.
- **neovisnost o redoslijedu unosa podataka**
- **sposobnost izvođenja na različitim tipovima varijabli**

Algoritme klasteriranja kategoriziramo prema tome kako tvore klastere. Dijelimo ih na hijerarhijske i nehijerarhijske (particijske). Hijerarhijske algoritme dijelimo na aglomerativne i razdjeljujuće, a particijske na tvrde i mekane algoritme (engl. *hard and soft clustering*).

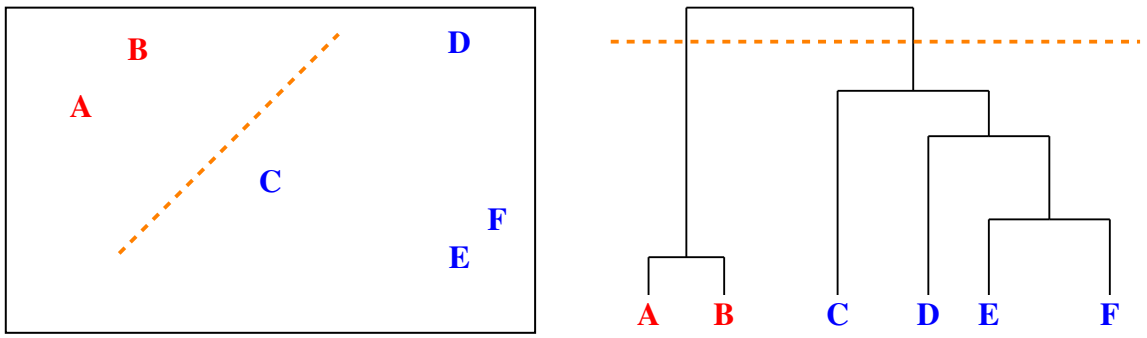
1.3.1 Hijerarhijski algoritmi

Hijerarhijski algoritmi temelje se na izgradnji hijerarhije klastera, koju grafički prikazujemo pomoću strukture stabla koja se naziva dendrogram. Aglomerativni algoritam počinje s $n \in \mathbb{N}$ klastera, od kojih svaki sadržava točno jedan podatak (engl. *singleton*) te nastavlja s postupnim spajanjem klastera sve dok svi nisu spojeni u jedan klaster. U nekim slučajevima algoritam završava kada dobijemo $k \in \mathbb{N}$ klastera, pri čemu je k unaprijed zadan.



Slika 1.1: Podjela algoritama klasteriranja

Dendrogram je matematički i grafički prikaz cjelovitoga hijerarhijskoga klasteriranja. Sastoji se od korijena, čvorova, grana i listova. Korijen je skup svih podataka, čvorovi su klasteri, a svaki je list podatak. Visina dendrograma jest sličnost (udaljenost) između dva objekta. Povlačenjem vodoravne linije na određenoj visini dendrograma dobivamo klastere koje očitujemo preko sjecišta. Primjerice, povlačenjem vodoravne linije kao na Slici 1.2 dobivamo klastere $\{A, B\}$ i $\{C, D, E, F\}$.



Slika 1.2: Skup podataka i njegov dendrogram

Razdjeljujući algoritam djeluje suprotno; na početku imamo klaster od n elemenata koji se postupno separira na manje sve dok ne dobijemo n klastera s jednim podatkom.

Za dani skup podataka $S = \{x_1, x_2, \ldots, x_n\}, n \in \mathbb{N}$, aglomerativni algoritam sastoji se od sljedećih koraka:

1. Svaki podatak stavi u zaseban klaster, tj. $C_i = \{x_i\}, i = \{1, 2, \ldots, n\}$
2. Izračunaj matricu udaljenosti D, $n \times n$ matricu čiji su elementi $d_{ij} = d(x_i, x_j), 1 \leq i, j \leq n$, tj.

$$D = \begin{bmatrix} 0 & d_{12} & \cdots & d_{1n} \\ d_{21} & 0 & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & 0 \end{bmatrix} \tag{1.26}$$

Zbog simetričnosti mjere udaljenosti vrijedi $d_{ij} = d_{ji}, \forall i, j$.

3. Odredi najbliži par klastera C_p i C_q, tj.

$$d(C_p, C_q) = \min_{i,j} d(C_i, C_j) = \min_{i,j} d_{ij}$$

4. Definiraj novi klaster $C_r = C_p \cup C_q$
5. Neka je $C_p = C_r$ i $C_q \in S \setminus C_p$
6. Ponavljaj dokle god postoje barem dva klastera.

Ključni korak algoritma odabir je najbližeg para klastera. Na temelju različitih definicija udaljenosti između dva klastera razlikujemo nekoliko aglomerativnih algoritama. Za proizvoljne klastere C_p i C_q, njihovu udaljenost određujemo pomoću sljedećih metoda:

- **Metoda minimuma**

$$d(C_p, C_q) = \min_{i,j}\{d(x_i, x_j) \mid x_i \in C_p, x_j \in C_q\} \tag{1.27}$$

Udaljenost između dva klastera definira se kao najmanja udaljenost između elemenata koji pripadaju tim klasterima.

- **Metoda maksimuma**

$$d(C_p, C_q) = \max_{i,j}\{d(x_i, x_j) \mid x_i \in C_p, x_j \in C_q\} \tag{1.28}$$

Udaljenost između dva klastera definira se kao najveća udaljenost između elemenata koji pripadaju tim klasterima. Metoda maksimuma stvara klastere koji su slične veličine tijekom cijelog procesa aglomeracije te zato dobiveni dendrogram izgleda uravnoteženije.

- **Metoda prosjeka**

$$d(C_p, C_q) = \frac{1}{n_p n_q} \sum_{x_i \in C_p} \sum_{x_j \in C_q} d(x_i, x_j), \tag{1.29}$$

pri čemu je n_p broj elemenata u C_p, a n_q broj elemenata u C_q. Udaljenost između dva klastera definira se kao aritmetička sredina udaljenosti između svih mogućih parova objekata koji pripadaju različitim klasterima.

- **Centroid metoda**
$$d(C_p, C_q) = d^2(t_p, t_q), \tag{1.30}$$
pri čemu je t_p centar klastera C_p, a t_q centar klastera C_q.

- **Ward metoda**
$$d(C_p, C_q) = \frac{n_p n_q}{n_p + n_q} d^2(t_p, t_q). \tag{1.31}$$
Elemente povezuju u klaster tako da varijanca unutar klastera bude minimalna.

Za dani skup podataka $S = \{x_1, x_2, \ldots, x_n\}, n \in \mathbb{N}$, razdjeljujući algoritam sastoji se od sljedećih koraka:

1. Svi podatci danog skupa S čine jedan klaster, tj. $C = \{x_1, x_2, \ldots, x_n\}$
2. Klaster podijelite na dva najmanja slična klastera koristeći partitivni algoritam, tj. $C = C_p \cup C_q$
3. Jedan od dobivenih klastera podijeli na dva najmanje slična klastera
4. Ponavljaj dok svaki podatak nije u zasebnom klasteru, tj. dok ne dobijemo n klastera $C_i = \{x_i\}, i = 1, 2, \ldots, n$.

Najpoznatiji razdjeljujući algoritmi jesu MONA i DIANA.

1.3.2 Nehijerarhijski algoritmi

Primjenom tvrdog nehijerarhijskog algoritma na skup S od $n \in \mathbb{N}$ podataka, dobit ćemo $k \leq n$ particija, gdje je svaka particija klaster. Svaki element skupa S pripada točno jednom klasteru, a svaki klaster sadržava barem jedan element. Najčešće korišteni algoritmi jesu k-sredina (engl. *k-means*), PAM, CLARA i CLARANS.

Cilj k-sredina algoritma jest stvoriti unaprijed zadani broj klastera k iz skupa podataka $S = \{x_1, x_2, \ldots, x_n\}, n \in \mathbb{N}$, tako da je vrijednost funkcije cilja minimalna. Funkcija cilja koristi se za procjenu kvalitete particioniranja tako da su objekti unutar istoga klastera međusobno slični, ali različiti (udaljeniji) od objekata iz drugih klastera. Za funkciju cilja uzimamo sumu kvadrata pogrešaka (SSE) između točaka skupa S i centroida klastera.

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - c_i\|^2 \tag{1.32}$$

Uočimo, zapravo minimiziramo varijancu klastera; što je manje varijacija unutar klastera, podatci su homogeniji unutar istoga klastera. Centroid c_i je centar klastera C_i te se računa kao aritmetička sredina objekata koji pripadaju istom klasteru, tj.

$$c_i = \frac{1}{n_i} \sum_{x \in C_i} x, \tag{1.33}$$

pri čemu je n_i broj elemenata klastera $C_i, i = 1, 2, \ldots, k$.

K-sredina sastoji se od sljedećih koraka:

1. Odaberi broj klastera k
2. Inicijaliziraj slučajnim odabirom centroide $\{c_1, c_2, \ldots, c_k\}$
3. Za svaki objekt iz skupa S pronađi najbliži centar c_i i pridruži ga klasteru C_i
4. Za svaki klaster C_i izračunaj novi centroid c_i
5. Ponavljaj 3. i 4. korak sve dok se centroidi ne mijenjaju.

Prema [6], postoji nekoliko metoda za određivanje optimalnog broja klastera k, no među popularnijima je metoda lakta. Što je broj klastera veći, vrijednost funkcije cilja monotono pada; prikažemo li grafički ovisnost funkcije cilja o broju klastera, dobit ćemo graf koji izgleda kao lakat (odaberimo k za koji vrijednost funkcije cilja naglo pada).

Prednost ovog algoritma jest jednostavna implementacija, učinkovitost nad velikim skupom podataka te relativno mala složenost algoritma. Konvergencija je zajamčena, no u većini slučajeva konvergira prema lokalnom optimumu. Jedan od nedostataka ovog algoritma jest da rezultati mogu ovisiti o inicijaliziranim centroidima. Također, netipične vrijednosti mogu utjecati na centroide i umjesto da budu ignorirani, svaki od njih može dobiti svoj klaster. Umjesto centroida, možemo promatrati medoid, tj. primjeniti algoritam k-medoida (engl. *k-medoids*) koji minimizira funkciju cilja (1.34). Prema [7], medoid o_i je objekt s minimalnom apsolutnom udaljenosti u odnosu prema drugim članovima klastera.

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - o_i||^2 \tag{1.34}$$

Struktura algoritma se u odnosu prema algoritmu k-sredina ne mijenja; početne medoide također inicijaliziramo slučajnim odabirom. PAM algoritam (engl. *the Partitioning Around Medoids*) je popularna realizacija k-medoid algoritma; djelotvoran je nad manjim skupom podataka, ali se ne prilagođava dobro nad velikim skupom podataka. Stoga primjenjujemo

CLARA algoritam (engl. *Clustering LARge Applications*), koji uzima manje uzorke te primjenjuje PAM algoritam nad njima. Optimalni rezultat među uzorcima uzimamo kao konačno rješenje. Kvalitetu i skalabilnost CLARA algoritma možemo poboljšati CLARANS algoritmom (engl. *Clustering Large Applications based upon RANdomized Search*). Prema [9, str. 451-457], nasumično odabere k objekata u skupu podataka kao trenutačne medoide te jedan medoid x i objekt y koji nije medoid. Ako zamjenom x sa y dobijemo manju vrijednost funkcije cilja, napravi se zamjena. Skup medoida nakon $l \in \mathbb{N}$ koraka smatra se lokalno optimalnim.

Za razliku od tvrdoga klasteriranja gdje svaki podatak skupa S pripada točno jednom klasteru, meki algoritam dopušta podatcima da pripadaju svim klasterima s nekom određenom težinom. Za zadani k definirajmo varijable $m_1, m_2, \ldots, m_k$ kao vjerojatnost kojom podatak x_i pripada klasteru $C_i, i = 1, 2, \ldots, k$. Kod tvrdoga klasteriranju, jedna od ovih vrijednosti jest 1, a ostale su 0. Prema definiciji vjerojatnosti, vrijedi

$$m_i \in [0, 1], \sum_{i=1}^{k} m_i = 1. \tag{1.35}$$

Najčešće korišten algoritam jest meke k-sredine (engl. *fuzzy k-means*). Prema [4], sastoji se od sljedećih koraka:

1. Odaberi broj klastera k
2. Slučajnim odabirom svakom podatku x_i pridruži vjerojatnosti $m_1^{(i)}, m_2^{(i)}, \ldots, m_k^{(i)}$
3. Izračunaj centroide c_i svakoga klastera C_i, tj.

$$c_i = \frac{\sum_{j=1}^{k} \sum_{x \in C_i} (m_j^{(i)})^m \cdot x}{\sum_{j=1}^{k} (m_j^{(i)})^m}, i = 1, 2, \ldots, k \tag{1.36}$$

   pri čemu je $1 \leq m < \infty$ parametar zamućenosti.

4. Za svaki podatak x_i izračunaj nove vjerojatnosti, tj.

$$m_j^{(i)} = \frac{1}{\sum_{l=1}^{k} \sum_{x \in C_i} \left(\frac{\|x - c_j\|}{\|x - c_l\|} \right)^{\frac{2}{m-1}}} \; j = 1, 2, \ldots, k \; , i = 1, 2, \ldots, n \tag{1.37}$$

5. Ponavljaj 3. i 4. korak sve dok se centroidi ne mijenjaju.

Najčešće se koristi u biologiji i medicini, primjerice, pri dijagnosticiranju bolesti i klasteriranju tumorskih stanica.

Poglavlje 2

Primjer

2.1 Opis podataka i metode

Za primjenu klasterske analize koristimo bazu `https://www.kaggle.com/rabbitsusan/world-happiness-report`. Baza je dobivena iz Svjetskog izvještaja o sreći za 2017. godinu. Izvještaj se temelji na projektu koji je istraživao kako socijalno, urbano i prirodno okružje utječu na sreću ljudi. Baza sadržava podatke o 146 država svijeta. Detaljni opis i izračun svake varijable možemo pronaći u [11] i [12]. Navodimo varijable koje su korištene u primjeni klasterske analize te njihove kratke opise:

- **Država**
  Ime države
- **Regija**
  Regija kojoj država pripada: Australija i Novi Zeland, Sjeverna Amerika, Južna Amerika i Karibi, zapadna Europa, središnja i istočna Europa, istočna Azija, jugoistočna Azija, južna Azija, Bliski istok i sjeverna Afrika ili supsaharska Afrika.
- **Mjera sreće**
  Mjera sreće je varijabla koja se računa pomoću podataka iz Gallupove ankete (vidi [11]). Temelji se na odgovorima na pitanje o procjeni života koje je postavljeno u anketi. Svaki ispitanik zamisli ljestvicu od 0 do 10, gdje je 0 najgori mogući život, a 10 najbolji mogući život. Nakon toga ispitanik ocjenjuje svoj život na toj ljestvici te se promatra prosjek odgovora. Takav pristup nazivamo Cantrilove ljestve.
- **Ekonomija**
  Varijabla koja se računa pomoću bruto domaćeg proizvoda i pariteta kupovne moći (vidi [11, str. 1,2]).

- **Obitelj**
  Odnosi se na prosjek binarnih odgovora (0 ili 1) na pitanje iz Gallupove ankete: „U slučaju da si u nevolji, imaš li obitelj ili prijatelje na koje se možeš osloniti u bilo kojem trenutku?"

- **Zdravlje**
  Varijabla koja daje podatke o zdravom životnom vijeku; računa se pomoću podataka WHO-a i WDI-a (vidi [11, str. 2,3]).

- **Sloboda**
  Odnosi se na prosjek binarnih odgovora (0 ili 1) na pitanje iz Gallupove ankete: „Jesi li zadovoljan ili nezadovoljan slobodom donošenja životnih odluka?"

- **Darežljivost**
  Varijabla koja daje podatke o velikodušnosti i donacijama (vidi [12]).

- **Povjerenje**
  Odnosi se na prosjek binarnih odgovora (0 ili 1) na dva pitanja iz Gallupove ankete: „Je li korupcija rasprostranjena u vlasti ili ne?" i „Je li korupcija rasprostranjena u gospodarstvu ili ne?"

Za obradu i analizu podataka korišten je SAS, programski sustav koji se koristi za statističku analizu i vizualizaciju podataka. Nad originalnim podatcima provedena je deskriptivna statistika. Nakon toga provodimo klastersku analizu u odnosu prema varijablama *Mjera sreće, Ekonomija, Obitelj, Zdravlje, Sloboda, Darežljivost* i *Povjerenje*. Najprije je provedeno hijerarhijsko (aglomerativno) klasteriranje pomoću Wardove metode i metode maksimuma. Zatim je provedeno nehijerarhijsko klasteriranje, tj. algoritam k-sredina. S obzirom na to da su sve varijable numeričke, u oba je slučaja kao mjera udaljenosti odabrana euklidska udaljenost. Svi algoritmi primijenjeni su nad originalnim i standardiziranim podatcima.

2.2 Deskriptivna statistika

U Tablicama 2.1, 2.3, 2.5, 2.7 i 2.9 prikazani su podatci iz naše baze. Deset najsretnijih država redom su Norveška, Danska, Island, Švicarska, Finska, Nizozemska, Kanada, Novi Zeland, Švedska i Australija; prevladavaju države zapadne Europe. S druge strane, Burundi, Tanzanija, Sirija, Ruanda, Togo, Gvineja, Liberija, Jemen, Haiti i Madagaskar redom su najmanje sretne države.

Tablica 2.1: Podatci za države Australije i Novog Zelanda te srednje i istočne Europe (SAS ispis)

		Mjera_srece	Ekonomija	Obitelj	Zdravlje	Sloboda	Darezljivost	Povjerenje
Regija	Drzava							
Australia and New Zealand	Australia	7.28399992	1.48441494	1.51004195	0.84388679	0.60160738	0.47769925	0.30118373
	New Zealand	7.31400013	1.40570605	1.54819512	0.81675971	0.61406213	0.50000513	0.38281670
Central and Eastern Europe	Albania	4.64400005	0.99619275	0.80368525	0.73115975	0.38149864	0.20131294	0.03986422
	Armenia	4.37599993	0.90059674	1.00748372	0.63752443	0.19830327	0.08348809	0.02667442
	Azerbaijan	5.23400021	1.15360177	1.15240026	0.54077578	0.39815584	0.04526934	0.18098751
	Belarus	5.56899977	1.15655756	1.44494522	0.63771427	0.29540026	0.15513751	0.15631382
	Bosnia and Herzegovina	5.18200016	0.98240942	1.06933594	0.70518631	0.20440318	0.32886750	0.00000000
	Bulgaria	4.71400023	1.16145909	1.43437946	0.70821768	0.28923172	0.11317769	0.01105153
	Croatia	5.29300022	1.22255623	0.96798301	0.70128852	0.25577229	0.24800298	0.04310311
	Czech Republic	6.60900021	1.35268235	1.43388522	0.75444400	0.49094617	0.08810676	0.03687293
	Estonia	5.61100006	1.32087934	1.47667110	0.69516832	0.47913143	0.09889081	0.18324892
	Georgia	4.28599978	0.95061266	0.57061493	0.64954698	0.30941004	0.05400882	0.25166664
	Hungary	5.32399988	1.28601193	1.34313309	0.68776345	0.17586352	0.07840166	0.03663694
	Kazakhstan	5.81899977	1.28455627	1.38436902	0.60604155	0.43745428	0.20196442	0.11928289
	Kosovo	5.27899981	0.95148438	1.13785350	0.54145205	0.26028794	0.31993145	0.05747162
	Kyrgyzstan	5.00400019	0.59622008	1.39423859	0.55345780	0.45494339	0.42858037	0.03943918
	Latvia	5.84999991	1.26074863	1.40471494	0.63856697	0.32570791	0.15307479	0.07384273
	Lithuania	5.90199995	1.31458235	1.47351611	0.62894994	0.23423179	0.01016466	0.01186564
	Macedonia	5.17500019	1.06457794	1.20789301	0.64494818	0.32590598	0.25376096	0.06027779
	Moldova	5.83799982	0.72887063	1.25182557	0.58946520	0.24072905	0.20877913	0.01009129
	Montenegro	5.23699999	1.12112904	1.23837650	0.66746467	0.19498906	0.19791102	0.08817419
	Poland	5.97300005	1.29178786	1.44571197	0.69947535	0.52034211	0.15846597	0.05930781
	Romania	5.82499981	1.21768391	1.15009129	0.68515831	0.45700374	0.13351992	0.00438790
	Russia	5.96299982	1.28177810	1.46928239	0.54734933	0.37378311	0.05226382	0.03296288
	Serbia	5.39499998	1.06931758	1.25818980	0.65078467	0.20871553	0.22012588	0.04090378
	Slovakia	6.09800005	1.32539356	1.50505924	0.71273291	0.29581747	0.13654448	0.02421085
	Slovenia	5.75799990	1.34120596	1.45251882	0.79082823	0.57257581	0.24264909	0.04512898
	Tajikistan	5.04099989	0.52471364	1.27146328	0.52923513	0.47156671	0.24899764	0.14637715
	Turkmenistan	5.82200003	1.13077676	1.49314916	0.43772608	0.41827193	0.24992499	0.25927034
	Ukraine	4.09600020	0.89465195	1.39453757	0.57590395	0.12297478	0.27006146	0.02302947
	Uzbekistan	5.97100020	0.78644109	1.54896915	0.49827263	0.65824866	0.41598365	0.24652822

Tablica 2.2: Deskriptivna statistika regija Australija i Novi Zeland te srednja i istočna Europa (SAS ispis)

The MEANS Procedure

Regija=Australia and New Zealand

Variable	N	Mean	Std Dev	Minimum	Maximum
Mjera_srece	2	7.2990000	0.0212134	7.2839999	7.3140001
Ekonomija	2	1.4450605	0.0556556	1.4057060	1.4844149
Obitelj	2	1.5291185	0.0269784	1.5100420	1.5481951
Zdravlje	2	0.8303232	0.0191817	0.8167597	0.8438868
Sloboda	2	0.6078348	0.0088068	0.6016074	0.6140621
Darezljivost	2	0.4888522	0.0157726	0.4776993	0.5000051
Povjerenje	2	0.3420002	0.0577232	0.3011837	0.3828167

Regija=Central and Eastern Europe

Variable	N	Mean	Std Dev	Minimum	Maximum
Mjera_srece	29	5.4099310	0.5916445	4.0960002	6.6090002
Ekonomija	29	1.0920510	0.2271075	0.5247136	1.3526824
Obitelj	29	1.2822854	0.2296362	0.5706149	1.5489692
Zdravlje	29	0.6360897	0.0824626	0.4377261	0.7908282
Sloboda	29	0.3466092	0.1320132	0.1229748	0.6582487
Darezljivost	29	0.1861161	0.1065284	0.0101647	0.4285804
Povjerenje	29	0.0796197	0.0784476	0	0.2592703

Tablice deskriptivne statistike dobivene su procedurom means. Preko naredbe var definiramo varijable za koje želimo deskriptivnu statistiku, a naredbom by radimo zasebnu analizu za svaku vrijednost varijable *Regija*. Korišten je sljedeći SAS kod:

```
proc means data=final;
    var Mjera_srece Ekonomija Obitelj Zdravlje Sloboda
    Darezljivost Povjerenje;
    by Regija;
run;
```

Australija i Novi Zeland imaju visoke vrijednosti svih varijabli te spadaju među najsretnije države svijeta. Među državama srednje i istočne Europe posebno se ističu Češka, Slovačka, Poljska, Uzbekistan i Rusija, koje imaju iznadprosječne vrijednosti *Mjere sreće*. Pogledamo li države Balkana, Rumunjska i Slovenija imaju iznadprosječne vrijednosti varijable *Mjera sreće*, a Hrvatska ima ispodprosječnu vrijednost. Zanimljivo je uočiti kako Bosna i Hercegovina ima vrijednost 0 za varijablu *Povjerenje*; sve zemlje Balkana imaju ispodprosječnu vrijednost varijable *Povjerenje*. Ekonomska situacija u Sloveniji, Hrvatskoj i Rumunjskoj bolja je u odnosu prema ostalim državama Balkana.

Tablica 2.3: Podatci za države istočne Azije, Sjeverne Amerike te Južne Amerike i Kariba (SAS ispis)

		Mjera_srece	Ekonomija	Obitelj	Zdravlje	Sloboda	Darezljivost	Povjerenje
Regija	Drzava							
Eastern Asia	China	5.27299976	1.08116579	1.16083741	0.74141550	0.47278771	0.02880684	0.02279428
	Japan	5.92000008	1.41691518	1.43633783	0.91347587	0.50562555	0.12057277	0.16376074
	Mongolia	4.95499992	1.02723587	1.49301124	0.55778348	0.39414397	0.33846423	0.03290229
	South Korea	5.83799982	1.40167844	1.12827444	0.90021408	0.25792167	0.20667437	0.06328267
Latin America and Caribbean	Argentina	6.59899998	1.18529546	1.44045115	0.69513708	0.49451920	0.10945706	0.05973989
	Bolivia	5.82299995	0.83375657	1.22761905	0.47363025	0.55873293	0.22556073	0.06047773
	Brazil	6.63500023	1.10735321	1.43130601	0.61655235	0.43745375	0.16234990	0.11109276
	Chile	6.65199995	1.25278461	1.28402495	0.81947970	0.37689528	0.32666242	0.08228798
	Colombia	6.35699987	1.07062233	1.40218294	0.59502792	0.47748742	0.14901447	0.04666874
	Costa Rica	7.07900000	1.10970628	1.41640365	0.75950927	0.58013165	0.21461323	0.10010659
	Dominican Republic	5.23000002	1.07937384	1.40241671	0.57487375	0.55258983	0.18696785	0.11394525
	Ecuador	6.00799990	1.00082040	1.28616881	0.68563622	0.45519820	0.15011247	0.14013465
	El Salvador	6.00299978	0.90978450	1.18212509	0.59601855	0.43245253	0.07825799	0.08998096
	Guatemala	6.45400000	0.87200195	1.25558519	0.54023999	0.53131062	0.28348839	0.07722328
	Haiti	3.60299993	0.36861026	0.64044982	0.27732113	0.03036986	0.48920378	0.09987215
	Honduras	5.18100023	0.73057312	1.14394498	0.58256948	0.34807986	0.23618887	0.07334545
	Jamaica	5.31099987	0.92557931	1.36821806	0.64102238	0.47430724	0.23381834	0.05526778
	Mexico	6.57800007	1.15318382	1.21086216	0.70997900	0.41273001	0.12099043	0.13277412
	Nicaragua	6.07100010	0.73729920	1.28721571	0.65309596	0.44755185	0.30167422	0.13068798
	Panama	6.45200014	1.23374844	1.37319255	0.70615614	0.55002683	0.21055694	0.07098392
	Paraguay	5.49300003	0.93253732	1.50728488	0.57925069	0.47350779	0.22415066	0.09106591
	Peru	5.71500015	1.03522527	1.21877039	0.63016611	0.45000288	0.12681972	0.04704909
	Trinidad and Tobago	6.16800022	1.36135590	1.38022852	0.51998329	0.51863074	0.32529646	0.00896482
	Uruguay	6.45400000	1.21755970	1.41222787	0.71921682	0.57939225	0.17509693	0.17806187
	Venezuela	5.25000000	1.12843120	1.43133760	0.61714423	0.15399712	0.06501963	0.06449112
North America	Canada	7.31599999	1.47920442	1.48134899	0.83455765	0.61110091	0.43553972	0.28737152
	United States	6.99300003	1.54625928	1.41992056	0.77428663	0.50574052	0.39257878	0.13563879

Tablica 2.4: Deskriptivna statistika regija istočna Azija, Sjeverna Amerika te Južna Amerika i Karibi (SAS ispis)

The MEANS Procedure

Regija=Eastern Asia

Variable	N	Mean	Std Dev	Minimum	Maximum
Mjera_srece	4	5.4964999	0.4615731	4.9549999	5.9200001
Ekonomija	4	1.2317488	0.2062874	1.0272359	1.4169152
Obitelj	4	1.3046152	0.1867369	1.1282744	1.4930112
Zdravlje	4	0.7782222	0.1664567	0.5577835	0.9134759
Sloboda	4	0.4076197	0.1102166	0.2579217	0.5056255
Darezljivost	4	0.1736296	0.1317208	0.0288068	0.3384642
Povjerenje	4	0.0706850	0.0643919	0.0227943	0.1637607

Regija=Latin America and Caribbean

Variable	N	Mean	Std Dev	Minimum	Maximum
Mjera_srece	21	5.9579048	0.7694693	3.6029999	7.0790000
Ekonomija	21	1.0116954	0.2243284	0.3686103	1.3613559
Obitelj	21	1.3000960	0.1811528	0.6404498	1.5072849
Zdravlje	21	0.6186672	0.1135142	0.2773211	0.8194797
Sloboda	21	0.4445413	0.1345238	0.0303699	0.5801317
Darezljivost	21	0.2093000	0.0987473	0.0650196	0.4892038
Povjerenje	21	0.0873439	0.0384834	0.0089648	0.1780619

Regija=North America

Variable	N	Mean	Std Dev	Minimum	Maximum
Mjera_srece	2	7.1545000	0.2283955	6.9630000	7.3160000
Ekonomija	2	1.5127319	0.0474150	1.4792044	1.5462593
Obitelj	2	1.4506348	0.0434365	1.4199206	1.4813490
Zdravlje	2	0.8044221	0.0426180	0.7742866	0.8345577
Sloboda	2	0.5584207	0.0745010	0.5057405	0.6111009
Darezljivost	2	0.4140593	0.0303780	0.3925788	0.4355397
Povjerenje	2	0.2115052	0.1072912	0.1356388	0.2873715

Japan i Južna Koreja sretnije su države u odnosu prema Kini. U odnosu prema ostatku regije Kina ima iznimno male vrijednosti za varijable *Darežljivost* i *Povjerenje*. Među državama Južne Amerike i Kariba odmah primjećujemo Kostariku i Haiti. Naime, osim po *Darežljivosti*, Haiti je ispodprosječna država (posebice po *Slobodi*). Kostarika se ističe po velikoj vrijednosti varijable *Mjera sreće*, ali nije dominantna u odnosu prema ostalim varijablama. Ako pogledamo države Sjeverne Amerike, Kanada je dominantnija u odnosu prema Americi; jedinu razliku čini varijabla *Ekonomija*.

Tablica 2.5: Podatci za države Bliskog istoka i sjeverne Afrike te južne Azije (SAS ispis)

		Mjera_srece	Ekonomija	Obitelj	Zdravlje	Sloboda	Darezljivost	Povjerenje
Regija	Drzava							
Middle East and Northern Africa	Algeria	5.87200022	1.09186447	1.14621747	0.61758465	0.23333581	0.06943665	0.14609611
	Bahrain	6.08699989	1.48841226	1.32311046	0.65313304	0.53674692	0.17266849	0.25704217
	Egypt	4.73500013	0.98970181	0.99747139	0.52018726	0.28211016	0.12863144	0.11438137
	Iran	4.69199991	1.15687311	0.71155125	0.63933319	0.24932261	0.38724291	0.04876107
	Iraq	4.49700022	1.10271049	0.97861320	0.50118047	0.28855553	0.19963726	0.10721576
	Israel	7.21299982	1.37538242	1.37628996	0.83840400	0.40598860	0.33008266	0.08524210
	Jordan	5.33599997	0.99101239	1.23908889	0.60459006	0.41842115	0.17217046	0.11980327
	Kuwait	6.10500002	1.63295245	1.25969875	0.63210571	0.49633759	0.22828980	0.21515955
	Lebanon	5.22499991	1.07498753	1.12962425	0.73508108	0.28851599	0.26445076	0.03751383
	Libya	5.52500010	1.10180306	1.35756433	0.52016902	0.46573323	0.15207367	0.09261021
	Morocco	5.23500013	0.87811458	0.77486444	0.59771067	0.40815833	0.03220996	0.08776318
	Palestinian Territorie	4.77500010	0.71624923	1.15564716	0.56566697	0.25471106	0.11417317	0.08928260
	Qatar	6.37500000	1.87076569	1.27429688	0.71009809	0.60413098	0.33047387	0.43929926
	Saudi Arabia	6.34399986	1.53062356	1.28667760	0.59014833	0.44975057	0.14761601	0.27343226
	Syria	3.46199989	0.77715313	0.39610261	0.50053334	0.08153945	0.49366373	0.15134713
	Tunisia	4.80499983	1.00726581	0.86835146	0.61321205	0.28968069	0.04969336	0.08672315
	Turkey	5.50000000	1.19827437	1.33775318	0.63760561	0.30074060	0.04669304	0.09967158
	United Arab Emirates	6.64799976	1.62634337	1.26641023	0.72679824	0.60834527	0.36094195	0.32448956
	Yemen	3.59299994	0.59168345	0.93538225	0.31008092	0.24946372	0.10412521	0.05676742
Southern Asia	Afghanistan	3.79399991	0.40147722	0.58154333	0.18074678	0.10617952	0.31187093	0.06115783
	Bangladesh	4.60799980	0.58668298	0.73513174	0.53324103	0.47835666	0.17225535	0.12371786
	Bhutan	5.01100016	0.88541639	1.34012651	0.49587929	0.50153768	0.47405455	0.17338039
	India	4.31500006	0.79222125	0.75437260	0.45542762	0.46998701	0.23153849	0.09222689
	Nepal	4.96199989	0.47982019	1.17928326	0.50413078	0.44030595	0.39409617	0.07297555
	Pakistan	5.26900005	0.72688353	0.67269069	0.40204778	0.23521526	0.31544602	0.12434807
	Sri Lanka	4.44000006	1.00985014	1.25997639	0.62513083	0.56121326	0.49086356	0.07365397

Tablica 2.6: Deskriptivna statistika regija Bliski istok i sjeverna Afrika te južna Azija (SAS ispis)

The MEANS Procedure

Regija=Middle East and Northern Africa

Variable	N	Mean	Std Dev	Minimum	Maximum
Mjera_srece	19	5.3696842	0.9852456	3.4619999	7.2129998
Ekonomija	19	1.1685354	0.3402122	0.5916834	1.8707657
Obitelj	19	1.0955114	0.2635823	0.3961026	1.3762900
Zdravlje	19	0.6059801	0.1124622	0.3100809	0.8384040
Sloboda	19	0.3637678	0.1397995	0.0815394	0.6083453
Darezljivost	19	0.1991723	0.1302980	0.0322100	0.4936637
Povjerenje	19	0.1490843	0.1060060	0.0375138	0.4392993

Regija=Southern Asia

Variable	N	Mean	Std Dev	Minimum	Maximum
Mjera_srece	7	4.6284286	0.4997323	3.7939999	5.2690001
Ekonomija	7	0.6974788	0.2198504	0.4014772	1.0098501
Obitelj	7	0.9318749	0.3150805	0.5815433	1.3401265
Zdravlje	7	0.4566577	0.1397324	0.1807468	0.6251308
Sloboda	7	0.3989708	0.1645394	0.1061795	0.5612133
Darezljivost	7	0.3414464	0.1189807	0.1722554	0.4908636
Povjerenje	7	0.1030658	0.0397237	0.0611578	0.1733804

Među državama Bliskog istoka i sjeverne Afrike uočavamo nekoliko jako sretnih država s dobrom ekonomskom situacijom, primjerice, Izrael, Ujedinjeni Arapski Emirati i Katar. S druge strane nalaze se Sirija i Jemen, jedne od najmanje sretnih država svijeta. Zanimljivo je uočiti kako je Šri Lanka dominantnija zemlja u odnosu prema Indiji. Zbog ratne situacije za Afganistan bilježimo očekivano niske vrijednosti svih varijabla.

Tablica 2.7: Podatci za države zapadne Europe i jugoistočne Azije (SAS ispis)

		Mjera_srece	Ekonomija	Obitelj	Zdravlje	Sloboda	Darezljivost	Povjerenje
Regija	Drzava							
Western Europe	Austria	7.00600004	1.48709726	1.45994496	0.81532842	0.56776619	0.31647232	0.22106037
	Belgium	6.89099979	1.46378076	1.46231270	0.81809187	0.53977072	0.23150334	0.25134313
	Cyprus	5.62099981	1.35593808	1.13136327	0.84471470	0.35511154	0.27125430	0.04123798
	Denmark	7.52199984	1.48238301	1.55112159	0.79256553	0.62600672	0.35528049	0.40077007
	Finland	7.46899986	1.44357193	1.54024673	0.80915767	0.61795086	0.24548277	0.38261154
	France	6.44199991	1.43092346	1.38777685	0.84446585	0.47022212	0.12976231	0.17250243
	Germany	6.95100021	1.48792338	1.47252035	0.79895073	0.56251138	0.33626917	0.27673194
	Greece	5.22700024	1.28948748	1.23941457	0.81019890	0.09573125	0.00000000	0.04328978
	Iceland	7.50400019	1.48063302	1.61057401	0.83355212	0.62716264	0.47554022	0.15352656
	Ireland	6.97700024	1.53570664	1.55823112	0.80978262	0.57311034	0.42785832	0.29838815
	Italy	5.96400023	1.39506662	1.44492328	0.85314435	0.25645071	0.17278965	0.02802809
	Luxembourg	6.86299992	1.74194360	1.45758367	0.84508950	0.59662789	0.28318098	0.31883442
	Malta	6.52699995	1.34327984	1.48841167	0.82194424	0.58876705	0.57473058	0.15306607
	Netherlands	7.37699986	1.50394464	1.42893922	0.81069613	0.58538449	0.47048983	0.28266183
	North Cyprus	5.80999994	1.34691131	1.18630338	0.83464724	0.47120363	0.26684570	0.15535335
	Norway	7.53700018	1.61646318	1.53352356	0.79666650	0.63542259	0.36201224	0.31596384
	Portugal	5.19500017	1.31517530	1.36704302	0.79584354	0.49846530	0.09510271	0.01586945
	Spain	6.40299988	1.38439786	1.53209090	0.88896060	0.40878123	0.19013357	0.07091410
	Sweden	7.28399992	1.49438727	1.47816217	0.83087516	0.61292410	0.38539925	0.38439873
	Switzerland	7.49399996	1.56497955	1.51691175	0.85813129	0.62007058	0.29054928	0.36700729
	United Kingdom	6.71400023	1.44163394	1.49646008	0.80533594	0.50819004	0.49277416	0.26542807
Southeastern Asia	Cambodia	4.16800022	0.60176510	1.00623834	0.42978340	0.63337582	0.38592297	0.06810595
	Indonesia	5.26200008	0.99553859	1.27444470	0.49234572	0.44332346	0.61170459	0.01531714
	Malaysia	6.08400011	1.29121542	1.28464603	0.61878443	0.40226498	0.41660893	0.06560071
	Myanmar	4.54500008	0.36711055	1.12323594	0.39752257	0.51449204	0.83807516	0.18881621
	Philippines	5.42999983	0.85769922	1.25391758	0.46800906	0.58521467	0.19351342	0.09933189
	Singapore	6.57200003	1.69227767	1.35381436	0.94949240	0.54984057	0.34596598	0.46430779
	Thailand	6.42399979	1.12786877	1.42579246	0.64723903	0.58020073	0.57212311	0.03161274
	Vietnam	5.07399988	0.78854758	1.27749133	0.65216899	0.57105559	0.23496805	0.08763324

Tablica 2.8: Deskriptivna statistika regija zapadna Europa i jugoistočna Azija (SAS ispis)

The MEANS Procedure

Regija=Western Europe

Variable	N	Mean	Std Dev	Minimum	Maximum
Mjera_srece	21	6.7037143	0.7568905	5.1950002	7.5370002
Ekonomija	21	1.4574109	0.1060254	1.2894875	1.7419436
Obitelj	21	1.4449457	0.1235719	1.1313633	1.6105740
Zdravlje	21	0.8246735	0.0246752	0.7925655	0.8889606
Sloboda	21	0.5151253	0.1379214	0.0957313	0.6354226
Darezljivost	21	0.3034967	0.1420512	0	0.5747306
Povjerenje	21	0.2189994	0.1269483	0.0158695	0.4007701

Regija=Southeastern Asia

Variable	N	Mean	Std Dev	Minimum	Maximum
Mjera_srece	8	5.4448750	0.8659408	4.1680002	6.5720000
Ekonomija	8	0.9652529	0.4131530	0.3671106	1.6922777
Obitelj	8	1.2499476	0.1307879	1.0062383	1.4257925
Zdravlje	8	0.5819182	0.1788124	0.3975226	0.9494924
Sloboda	8	0.5349710	0.0776498	0.4022650	0.6333758
Darezljivost	8	0.4498603	0.2137196	0.1935134	0.8380752
Povjerenje	8	0.1275907	0.1457593	0.0153171	0.4643078

U zapadnoj su Europi države koje su iznimno sretne. Norveška, kao država s najvećom mjerom sreće, predvodnica je ove regije. Ekonomska situacija u svim je državama iznimno dobra, posebice u Luksemburgu. Prema Tablici 2.7, Grčka ima vrijednost 0 za varijablu *Darežljivost*; ljudi nisu skloni donirati u dobrotvorne svrhe. Singapur u odnosu prema drugim državama jugoistočne Azije ima mnogo bolju ekonomsku situaciju, a korumpiranost je svedena na minimum.

Tablica 2.9: Podatci za države supsaharske Afrike (SAS ispis)

		Mjera_srece	Ekonomija	Obitelj	Zdravlje	Sloboda	Darezljivost	Povjerenje
Regija	Drzava							
Sub-Saharan Africa	Angola	3.79500008	0.85842818	1.10441196	0.04986867	0.00000000	0.09792649	0.06972034
	Benin	3.65700007	0.43108541	0.43529984	0.20993021	0.42596278	0.20794846	0.06092902
	Botswana	3.76600003	1.12209415	1.22155500	0.34175551	0.50519633	0.09934845	0.09858320
	Burkina Faso	4.03200007	0.35022771	1.04328001	0.21584426	0.32436785	0.25086469	0.12032811
	Burundi	2.90499997	0.09162257	0.62979358	0.15161079	0.05990075	0.20443519	0.08414795
	Cameroon	4.69500017	0.56430537	0.94601822	0.13289212	0.43038875	0.23629846	0.05130663
	Chad	3.93600011	0.43801299	0.95385587	0.04113472	0.16234203	0.21611385	0.05358188
	Congo (Brazzaville)	4.29099989	0.80896425	0.83204436	0.28995743	0.43502587	0.12085213	0.07961813
	Congo (Kinshasa)	4.28000021	0.09210235	1.22902346	0.19140703	0.23596135	0.24645583	0.06024136
	Ethiopia	4.46000004	0.33923385	0.86466920	0.35340971	0.40884274	0.31265074	0.16545571
	Gabon	4.46500015	1.19821024	1.15562022	0.35657859	0.31232858	0.04378538	0.07604679
	Ghana	4.11999989	0.66722482	0.87366474	0.29563773	0.42302629	0.25692394	0.02533637
	Guinea	3.50699997	0.24454993	0.79124469	0.19412914	0.34858751	0.26481509	0.11093762
	Ivory Coast	4.17999983	0.60304892	0.90478003	0.04864217	0.44770619	0.20123747	0.13006178
	Kenya	4.55299997	0.56047946	1.06795073	0.30998835	0.45276377	0.44486031	0.06464132
	Liberia	3.53299999	0.11904179	0.87211794	0.22991820	0.33288118	0.26654989	0.03894825
	Madagascar	3.64400005	0.30580869	0.91302037	0.37522331	0.18919677	0.20873253	0.06723198
	Malawi	3.97000003	0.23344204	0.51256883	0.31508958	0.46691465	0.28717047	0.07271165
	Mali	4.19000006	0.47618049	1.28147340	0.16936567	0.30661374	0.18335420	0.10497025
	Mauritania	4.29199982	0.64845729	1.27203083	0.28534928	0.09609804	0.20187002	0.13695701
	Mauritius	5.62900019	1.18939555	1.20956099	0.63800746	0.49124733	0.36093375	0.04218156
	Niger	4.02799988	0.16192533	0.99302501	0.26850501	0.36365870	0.22867385	0.13857295
	Nigeria	5.07399988	0.78375626	1.21577048	0.05691573	0.39495257	0.23094720	0.02612157
	Rwanda	3.47099996	0.36874589	0.94570702	0.32642481	0.58184385	0.25275603	0.45522001
	Senegal	4.53499985	0.47930902	1.17969191	0.40936285	0.37792227	0.18346889	0.11546045
	Sierra Leone	4.70900011	0.36842093	0.98413605	0.00556475	0.31869769	0.29304090	0.07109518
	South Africa	4.82900000	1.05469871	1.38478863	0.18708007	0.47924674	0.13936238	0.07250950
	Sudan	4.13899994	0.65951669	1.21400857	0.29092082	0.01499586	0.18231745	0.08984752
	Tanzania	3.34899998	0.51113588	1.04198980	0.36450928	0.39001778	0.35425636	0.06603511
	Togo	3.49499989	0.30544472	0.43188253	0.24710557	0.38042614	0.19689615	0.09566502
	Uganda	4.08099985	0.38143072	1.12982774	0.21763261	0.44318596	0.32576606	0.05706972
	Zambia	4.51399994	0.63640678	1.00318730	0.25783590	0.46160349	0.24958015	0.07821355
	Zimbabwe	3.87500000	0.37584654	1.08309591	0.19676375	0.33638421	0.18914349	0.09537538

Tablica 2.10: Deskriptivna statistika regije supsaharska Afrika (SAS ispis)

The MEANS Procedure

Regija=Sub-Saharan Africa

Variable	N	Mean	Std Dev	Minimum	Maximum
Mjera_srece	33	4.1211818	0.5482662	2.9050000	5.6290002
Ekonomija	33	0.5281380	0.3048920	0.0916226	1.1982102
Obitelj	33	0.9915483	0.2371578	0.4318825	1.3847886
Zdravlje	33	0.2431625	0.1278524	0.0055648	0.6380075
Sloboda	33	0.3454027	0.1441928	0	0.5818439
Darezljivost	33	0.2284647	0.0808043	0.0437854	0.4448603
Povjerenje	33	0.0931855	0.0728133	0.0253364	0.4552200

U regiji supsaharska Afrika prevladavaju države koje su najmanje sretne. U ovoj se regiji nalazi i najmanje sretna država, Burundi. Prema Tablici 2.7, Angola ima vrijednost 0 za varijablu *Sloboda*; naime, već se godinama krše osnovna ljudska prava kao što su sloboda okupljanja, udruživanja, govora i medija. Po *Mjeri sreće* ističu se Mauricijus i Nigerija; bolja ekonomska situacija prevladava na Mauricijusu.

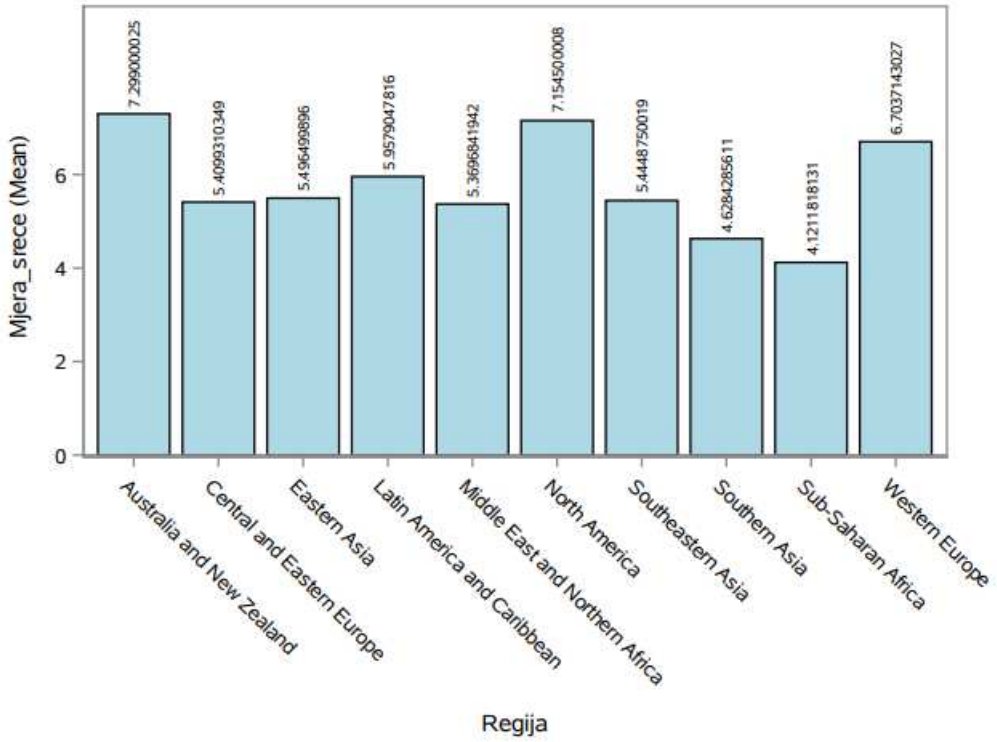
Tablica 2.11: Deskriptivna statistika originalnih podataka (SAS ispis)

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
Mjera_srece	146	5.3949041	1.1186029	2.9050000	7.5370002
Ekonomija	146	1.0040986	0.4039941	0.0916226	1.8707657
Obitelj	146	1.2059496	0.2693977	0.3961026	1.6105740
Zdravlje	146	0.5652676	0.2233314	0.0055648	0.9494924
Sloboda	146	0.4078774	0.1507146	0	0.6582487
Darezljivost	146	0.2464319	0.1373115	0	0.8380752
Povjerenje	146	0.1217934	0.1025196	0	0.4643078

Usporedimo države Balkana s Tablicom 2.11. Kosovo je jedina država Balkana koja je po svim varijablama ispodprosječna država. Slovenija, Srbija i Rumunjska iznadprosječno su sretne države. Ako pogledamo ekonomsku situaciju, jedino su Albanija te Bosna i Hercegovina ispodprosječne. Unatoč tome, sve države imaju dobre temelje za razvoj zdravog životnog vijeka. Hrvatska se s Makedonijom te Bosnom i Hercegovinom ističe po velikodušnosti, tj. donacijama u dobrotvorne svrhe. Zanimljivo je uočiti kako su jedino Slovenija i Rumunjska zadovoljne slobodom donošenja životnih odluka. Sve države imaju ispodprosječnu vrijednost varijable *Povjerenje*; pripadaju skupini korumpiranijih država svijeta.

Na Slici 2.1 vidimo da regija Australija i Novi Zeland imaju najveću prosječnu vrijednost varijable *Mjera sreće*, a u stopu je slijede Sjeverna Amerika i zapadna Europa. Međutim, regijama Australija i Novi Zeland te Sjeverna Amerika pripadaju samo dvije države, a u zapadnoj je Europi 21 država i u njoj se nalazi većina najsretnijih država svijeta.



Slika 2.1: Aritmetička sredina varijable *Mjera sreće* po regijama

2.3 Hijerarhijsko klasteriranje

Broj klastera određuje donositelj odluke; ovisi na kojoj će visini odsjeći dendrogram. Općenito je pogreška koristiti dendrograme kao alat za određivanje točnog broja klastera u podatcima. Kada znamo da postoji točan broj klastera, to će u većini slučajeva biti vidljivo na dendrogramu. Aglomerativno klasteriranje provodi se korištenjem procedure cluster. Za Wardovu metodu i metodu maksimuma, pri čemu je udaljenost euklidska, korišten je sljedeći SAS kod nad originalnim podatcima:

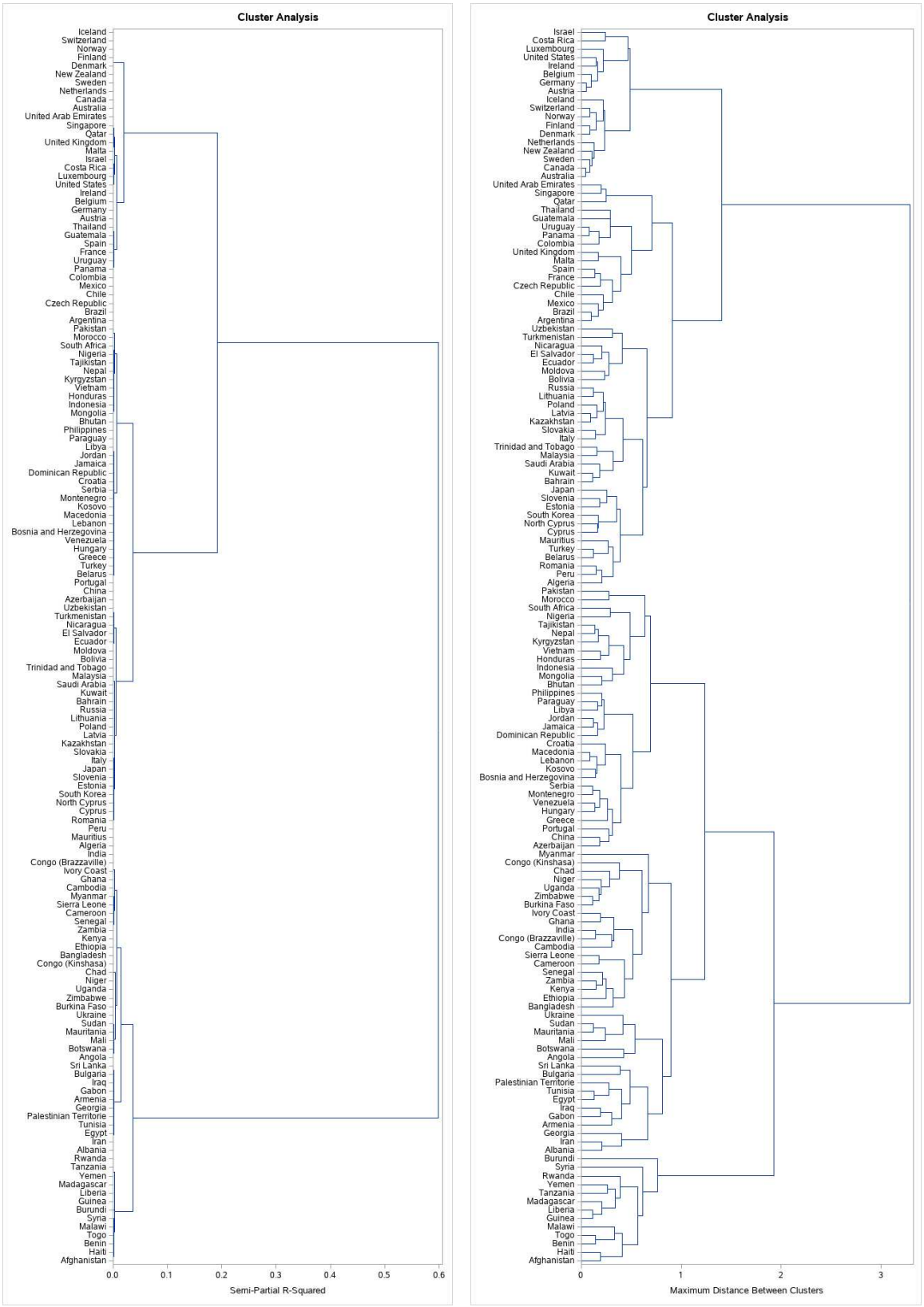
```
proc distance data=final out=dist nostd method=Euclid;
    var interval(Mjera_srece Ekonomija Obitelj Zdravlje Sloboda
    Darezljivost Povjerenje);
```

```
    id Drzava;
run;

proc cluster data=dist method=ward outtree=Tree;
    id Drzava;
run;

proc cluster data=dist method=com outtree=Tree;
    id Drzava;
run;
```

Za odabranu mjeru udaljenosti, koju definiramo naredbom method, procedura distance računa matricu udaljenosti te je sprema u bazu out. Nakon toga procedura cluster provodi hijerarhijsko klasteriranje nad bazom out, ovisno o metodi koja je određena naredbom method. Preko naredbe id definiramo varijablu po kojoj ćemo razlikovati opažanja u dendrogramu.

Na Slici 2.2 prikazani su dendrogrami dobiveni gornjim kodom; lijevo se nalazi dendrogram dobiven Wardovom metodom, a desno dendrogram dobiven metodom maksimuma. Za Hrvatsku možemo primijetiti da se u oba dendrograma pojavljuje s nekim državama Balkana, točnije sa Srbijom, Crnom Gorom, Kosovom, Makedonijom te Bosnom i Hercegovinom. Međutim, s njima se pojavljuje i Libanon, koji spada pod regiju Bliski istok i sjeverna Afrika. Slovenija i Rumunjska pojavljuju se s Estonijom, Južnom Korejom, Sjevernim Ciprom i Ciprom, a Bugarska i Albanija pojavljuju se s Palestinom, Egiptom, Armenijom, Iranom, Irakom, Gruzijom i Gabonom.

Također je zanimljivo uočiti pridruživanje Turske Balkanu u odnosu prema različitim metodama: primjenom Wardove metode Turska je bliže Hrvatskoj, Srbiji, Crnoj Gori, Kosovu, Makedoniji te Bosni i Hercegovini, dok je primjenom metode maksimuma bliže Sloveniji i Rumunjskoj. Pogledamo li deset država koje imaju najveću mjeru sreće (Norveška, Danska, Island, Švicarska, Finska, Nizozemska, Kanada, Novi Zeland, Švedska i Australija), vidimo da se na oba dendrograma pojavljuju zajedno.



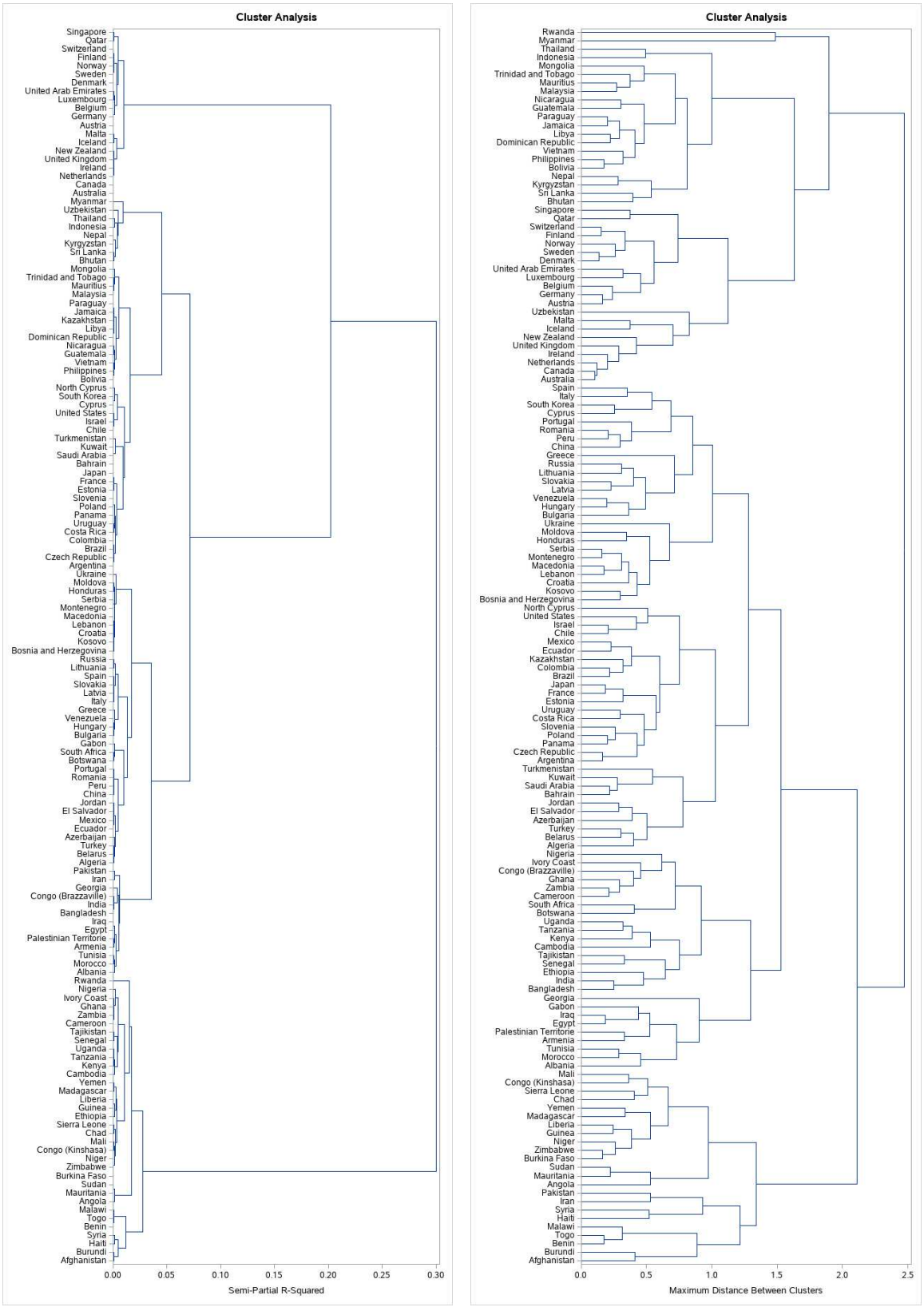
Slika 2.2: Dendrogram originalnih podataka dobiven Wardovom metodom i metodom maksimuma (SAS ispis)

Wardovu metodu i metodu maksimuma primjenjujemo nad standardiziranim podatcima, pri čemu je odabrana euklidska udaljenost. Standardizacija je obavljena procedurom stdize, koja dobivene podatke sprema u bazu out. Naredbom method određujemo metodu standardizacije. Korišten je sljedeći SAS kod:

```
proc stdize data=final out=standardizirani method=std nomiss;
    var Mjera_srece Ekonomija Obitelj Zdravlje Sloboda
    Darezljivost Povjerenje;
run;

proc distance data=standardizirani out=dist nostd method=Euclid;
    var interval(Mjera_srece Ekonomija Obitelj Zdravlje Sloboda
    Darezljivost Povjerenje);
    id Drzava;
run;

proc cluster data=dist method=ward outtree=Tree;
    id Drzava;
run;

proc cluster data=dist method=com outtree=Tree;
    id Drzava;
run;
```

Na Slici 2.3 prikazani su dendrogrami dobiveni primjenom Wardove metode i metode maksimuma nad standardiziranim podatcima. U oba dendrograma, kao i kod originalnih podataka, Hrvatska se pojavljuje sa Srbijom, Crnom Gorom, Kosovom, Makedonijom te Bosnom i Hercegovinom, a s njima se pojavljuje i Libanon. Ako to usporedimo s originalnim podatcima, Bugarska i Albanija se ne pojavljuju zajedno; Bugarska je bliža Hrvatskoj, a Albanija se u oba dendrograma pojavljuje s Marokom, Tunisom, Armenijom, Palestinom i Egiptom. Zanimljivo je uočiti da se Slovenija neovisno o metodi pojavljuje s Francuskom, Estonijom, Poljskom, Urugvajom, Panamom i Kostarikom, a Rumunjska se pojavljuje s Portugalom, Kinom i Peruom.

Za razliku od originalnih podataka gdje se deset najsretnijih zemalja pojavljuju zajedno, na oba dendrograma se Švicarska, Finska, Norveška, Švedska i Danska pojavljuju sa Singapurom i Katarom, a Island, Novi Zeland, Nizozemska, Kanada i Australija pojavljuju se s Irskom i Ujedinjenim Kraljevstvom.



Slika 2.3: Dendrogram standardiziranih podataka dobiven Wardovom metodom i metodom maksimuma (SAS ispis)

2.4 Nehijerarhijsko klasteriranje

Nakon hijerarhijskoga klasteriranja provedeno je nehijerarhijsko klasteriranje. Proveden je algoritam k-sredina nad originalnim i standardiziranim podatcima. S obzirom na to da većina dendrograma sugerira 5 do 7 klastera, proveden je algoritam za $k = 5, 6, 7$. Za $k = 5$ korišten je sljedeći SAS kod:

```
proc fastclus data=final out=klaster5 maxclusters=5 nomiss maxiter=300;
    var Mjera_srece Ekonomija Obitelj Zdravlje Sloboda
    Darezljivost Povjerenje;
run;
```

Procedura fastclus provodi algoritam k-sredina nad bazom data te rezultat klasteriranja sprema u bazu out. Naredbom maxclusters definiramo broj klastera, dok naredbom nomiss u analizu ne ulaze varijable koje nedostaju. Za mjeru udaljenosti procedura fastclus uzima Euklidsku udaljenost. Naredbom maxiter definiramo maksimalni broj iteracija. Analogni kod primjenjujemo i za standardizirane podatke te za $k = 6, 7$.

U Tablicama 2.12, 2.13, 2.14 i 2.15 prikazani su rezultati klasteriranja algoritmom k-sredina primijenjenog na originalne i standardizirane podatke za $k = 5, 6, 7$. Baza podataka može sadržavati više obilježja i ta se obilježja mogu međusobno bitno razlikovati zato što, primjerice, dolaze iz različitih domena. Stoga razlike možemo bolje uočiti ako standardiziramo podatke; grafički prikaz aritmetičkih sredina bit će mnogo bolji i jasniji.

Tablica 2.12: Rezultati klasteriranja algoritmom k-sredina (SAS ispis)

Obs	Drzava	k=5 orig	k=5 stand	Drzava_1	k=6 orig	k=6 stand	Drzava_2	k=7 orig	k=7 stand
1	Afghanistan	1	4	Moldova	1	1	Pakistan	1	1
2	Haiti	1	4	Azerbaijan	1	1	Iran	1	2
3	Syria	1	4	Bosnia and Herzegovina	1	1	Albania	1	2
4	Benin	1	4	Croatia	1	1	Bosnia and Herzegovina	1	2
5	Burundi	1	4	Greece	1	1	Bulgaria	1	2
6	Guinea	1	4	Honduras	1	1	Egypt	1	2
7	Malawi	1	4	Hungary	1	1	Honduras	1	2
8	Togo	1	4	Kosovo	1	1	Kosovo	1	2
9	Tanzania	1	2	Lebanon	1	1	Lebanon	1	2
10	Liberia	1	4	Macedonia	1	1	Macedonia	1	2
11	Madagascar	1	4	Montenegro	1	1	Morocco	1	2
12	Yemen	1	4	Morocco	1	1	Palestinian Territorie	1	2
13	Rwanda	1	2	Serbia	1	1	Tunisia	1	2
14	Georgia	2	4	Turkey	1	1	Mongolia	1	3
15	Armenia	2	5	Venezuela	1	1	Vietnam	1	3
16	Gabon	2	5	Bolivia	1	1	South Africa	1	4
17	Iraq	2	5	Cyprus	1	1	Nigeria	1	4
18	Ukraine	2	5	Estonia	1	1	Bhutan	1	5
19	Botswana	2	2	Mauritius	1	1	Indonesia	1	5
20	Bangladesh	2	2	Peru	1	1	Kyrgyzstan	1	5
21	Burkina Faso	2	2	Belarus	1	1	Nepal	1	5
22	Cameroon	2	2	China	1	1	Tajikistan	1	5
23	Congo (Brazzaville)	2	2	Dominican Republic	1	1	Algeria	2	2
24	Congo (Kinshasa)	2	2	Jamaica	1	1	Lithuania	2	2
25	Ethiopia	2	2	Jordan	1	1	Moldova	2	2
26	Ghana	2	2	Libya	1	1	South Korea	2	2
27	India	2	2	Mongolia	1	1	Azerbaijan	2	2
28	Ivory Coast	2	2	Paraguay	1	1	Croatia	2	2
29	Mali	2	2	Philippines	1	1	Greece	2	2
30	Niger	2	2	Portugal	1	1	Hungary	2	2
31	Senegal	2	2	Vietnam	1	1	Montenegro	2	2
32	Sierra Leone	2	2	Indonesia	1	3	Serbia	2	2
33	Uganda	2	2	Pakistan	2	2	Turkey	2	2
34	Zambia	2	2	Iran	2	2	Venezuela	2	2
35	Zimbabwe	2	2	Armenia	2	1	Bolivia	2	3
36	Angola	2	4	Gabon	2	1	Cyprus	2	3
37	Chad	2	4	Iraq	2	1	El Salvador	2	3

Tablica 2.13: Rezultati klasteriranja algoritmom k-sredina - nastavak (SAS ispis)

Obs	Drzava	k=5 orig	k=5 stand	Drzava_1	k=6 orig	k=6 stand	Drzava_2	k=7 orig	k=7 stand
38	Mauritania	2	4	Albania	2	1	Estonia	2	3
39	Sudan	2	4	Bulgaria	2	1	Kazakhstan	2	3
40	Cambodia	2	1	Egypt	2	1	Latvia	2	3
41	Kenya	2	1	Palestinian Territorie	2	1	Mauritius	2	3
42	Myanmar	2	1	Tunisia	2	1	North Cyprus	2	3
43	Sri Lanka	2	1	South Africa	2	1	Peru	2	3
44	Algeria	3	5	Bangladesh	2	4	Romania	2	3
45	Lithuania	3	5	Cameroon	2	4	Slovenia	2	3
46	Moldova	3	5	Senegal	2	4	Turkmenistan	2	3
47	South Korea	3	5	Sierra Leone	2	4	Belarus	2	3
48	Kuwait	3	3	Zambia	2	4	China	2	3
49	Saudi Arabia	3	3	Nigeria	2	4	Dominican Republic	2	3
50	Bahrain	3	3	Kenya	2	3	Jamaica	2	3
51	Argentina	3	5	Myanmar	2	3	Jordan	2	3
52	Bolivia	3	5	Sri Lanka	2	3	Libya	2	3
53	Brazil	3	5	Bhutan	2	3	Paraguay	2	3
54	Chile	3	5	Kyrgyzstan	2	3	Philippines	2	3
55	Colombia	3	5	Nepal	2	3	Portugal	2	3
56	Cyprus	3	5	Tajikistan	2	1	Kuwait	3	3
57	Czech Republic	3	5	Algeria	3	1	Saudi Arabia	3	3
58	Ecuador	3	5	Lithuania	3	1	Bahrain	3	3
59	El Salvador	3	5	South Korea	3	1	Argentina	3	3
60	Estonia	3	5	Kuwait	3	1	Brazil	3	3
61	France	3	5	Saudi Arabia	3	1	Chile	3	3
62	Guatemala	3	5	Bahrain	3	6	Colombia	3	3
63	Italy	3	5	Argentina	3	1	Czech Republic	3	3
64	Japan	3	5	Brazil	3	1	Ecuador	3	3
65	Kazakhstan	3	5	Chile	3	1	France	3	3
66	Latvia	3	5	Colombia	3	1	Guatemala	3	3
67	Malaysia	3	5	Czech Republic	3	1	Italy	3	3
68	Mauritius	3	5	Ecuador	3	1	Japan	3	3
69	Mexico	3	5	El Salvador	3	1	Malaysia	3	3
70	Nicaragua	3	5	France	3	1	Mexico	3	3
71	North Cyprus	3	5	Guatemala	3	1	Nicaragua	3	3
72	Panama	3	5	Italy	3	1	Panama	3	3
73	Peru	3	5	Japan	3	1	Poland	3	3
74	Poland	3	5	Kazakhstan	3	1	Russia	3	3

Tablica 2.14: Rezultati klasteriranja algoritmom k-sredina - nastavak (SAS ispis)

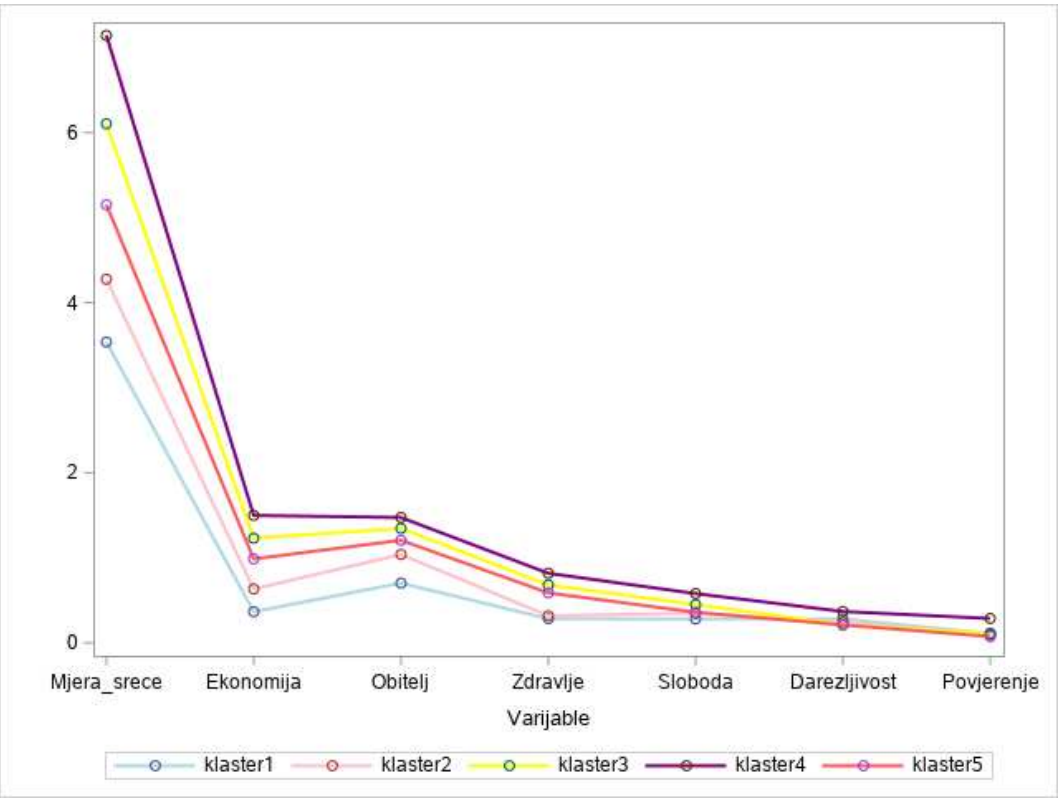
Obs	Drzava	k=5 orig	k=5 stand	Drzava_1	k=6 orig	k=6 stand	Drzava_2	k=7 orig	k=7 stand
75	Romania	3	5	Latvia	3	1	Slovakia	3	3
76	Russia	3	5	Malaysia	3	1	Spain	3	3
77	Slovakia	3	5	Mexico	3	1	Trinidad and Tobago	3	3
78	Slovenia	3	5	Nicaragua	3	1	Uruguay	3	3
79	Spain	3	5	North Cyprus	3	1	Thailand	3	5
80	Trinidad and Tobago	3	5	Panama	3	1	Uzbekistan	3	5
81	Turkmenistan	3	5	Poland	3	1	Malta	3	7
82	Uruguay	3	5	Romania	3	1	Qatar	3	7
83	Thailand	3	1	Russia	3	1	Singapore	3	7
84	Uzbekistan	3	3	Slovakia	3	1	United Arab Emirates	3	7
85	Malta	3	3	Slovenia	3	1	Georgia	4	2
86	Qatar	3	3	Spain	3	1	Armenia	4	2
87	Costa Rica	4	5	Trinidad and Tobago	3	1	Gabon	4	2
88	Israel	4	5	Turkmenistan	3	1	Iraq	4	2
89	Australia	4	3	Uruguay	3	1	Ukraine	4	2
90	Austria	4	3	Thailand	3	3	Botswana	4	4
91	Belgium	4	3	Uzbekistan	3	3	Congo (Brazzaville)	4	4
92	Canada	4	3	Malta	3	6	India	4	4
93	Denmark	4	3	Qatar	3	6	Angola	4	4
94	Finland	4	3	Georgia	4	2	Sudan	4	4
95	Germany	4	3	Ukraine	4	1	Sri Lanka	4	5
96	Iceland	4	3	Botswana	4	1	Afghanistan	5	1
97	Ireland	4	3	Burkina Faso	4	4	Haiti	5	1
98	Luxembourg	4	3	Congo (Brazzaville)	4	4	Syria	5	1
99	Netherlands	4	3	Congo (Kinshasa)	4	4	Benin	5	1
100	New Zealand	4	3	Ethiopia	4	4	Burundi	5	1
101	Norway	4	3	Ghana	4	4	Guinea	5	1
102	Singapore	4	3	India	4	4	Malawi	5	1
103	Sweden	4	3	Ivory Coast	4	4	Togo	5	1
104	Switzerland	4	3	Mali	4	4	Tanzania	5	4
105	United Arab Emirates	4	3	Niger	4	4	Liberia	5	4
106	United Kingdom	4	3	Uganda	4	4	Madagascar	5	4
107	United States	4	3	Zimbabwe	4	4	Yemen	5	4
108	Pakistan	5	4	Angola	4	4	Rwanda	5	6
109	Iran	5	4	Chad	4	4	Bangladesh	6	4
110	Albania	5	5	Mauritania	4	4	Burkina Faso	6	4
111	Azerbaijan	5	5	Sudan	4	4	Cameroon	6	4

Tablica 2.15: Rezultati klasteriranja algoritmom k-sredina - nastavak (SAS ispis)

Obs	Drzava	k=5 orig	k=5 stand	Drzava_1	k=6 orig	k=6 stand	Drzava_2	k=7 orig	k=7 stand
112	Bosnia and Herzegovina	5	5	Cambodia	4	3	Congo (Kinshasa)	6	4
113	Bulgaria	5	5	Afghanistan	5	2	Ethiopia	6	4
114	Croatia	5	5	Haiti	5	2	Ghana	6	4
115	Egypt	5	5	Syria	5	2	Ivory Coast	6	4
116	Greece	5	5	Benin	5	4	Mali	6	4
117	Honduras	5	5	Burundi	5	4	Niger	6	4
118	Hungary	5	5	Guinea	5	4	Senegal	6	4
119	Kosovo	5	5	Malawi	5	4	Sierra Leone	6	4
120	Lebanon	5	5	Togo	5	4	Uganda	6	4
121	Macedonia	5	5	Tanzania	5	4	Zambia	6	4
122	Montenegro	5	5	Liberia	5	4	Zimbabwe	6	4
123	Morocco	5	5	Madagascar	5	4	Chad	6	4
124	Palestinian Territorie	5	5	Yemen	5	4	Mauritania	6	4
125	Serbia	5	5	Rwanda	5	5	Cambodia	6	5
126	Tunisia	5	5	Costa Rica	6	1	Kenya	6	5
127	Turkey	5	5	Israel	6	1	Myanmar	6	5
128	Venezuela	5	5	Australia	6	6	Costa Rica	7	3
129	Belarus	5	5	Austria	6	6	Israel	7	3
130	China	5	5	Belgium	6	6	Australia	7	7
131	Dominican Republic	5	5	Canada	6	6	Austria	7	7
132	Jamaica	5	5	Denmark	6	6	Belgium	7	7
133	Jordan	5	5	Finland	6	6	Canada	7	7
134	Libya	5	5	Germany	6	6	Denmark	7	7
135	Mongolia	5	5	Iceland	6	6	Finland	7	7
136	Paraguay	5	5	Ireland	6	6	Germany	7	7
137	Philippines	5	5	Luxembourg	6	6	Iceland	7	7
138	Portugal	5	5	Netherlands	6	6	Ireland	7	7
139	Vietnam	5	5	New Zealand	6	6	Luxembourg	7	7
140	South Africa	5	2	Norway	6	6	Netherlands	7	7
141	Nigeria	5	2	Singapore	6	6	New Zealand	7	7
142	Bhutan	5	1	Sweden	6	6	Norway	7	7
143	Indonesia	5	1	Switzerland	6	6	Sweden	7	7
144	Kyrgyzstan	5	1	United Arab Emirates	6	6	Switzerland	7	7
145	Nepal	5	1	United Kingdom	6	6	United Kingdom	7	7
146	Tajikistan	5	2	United States	6	6	United States	7	7

Tablica 2.16: Aritmetičke sredine varijabli po klasterima originalnih podataka za $k = 5$ (SAS ispis)

Cluster Means							
Cluster	**Mjera_srece**	**Ekonomija**	**Obitelj**	**Zdravlje**	**Sloboda**	**Darezljivost**	**Povjerenje**
1	3.537153812	0.365369306	0.702084816	0.283278697	0.280252612	0.280186447	0.109305471
2	4.275566673	0.633200931	1.038918497	0.317789273	0.347792603	0.243012797	0.094586358
3	6.103558130	1.230381095	1.344818759	0.680306183	0.449644799	0.212246150	0.108411603
4	7.143238090	1.498273458	1.473283654	0.817519912	0.578557884	0.368106622	0.285708700
5	5.152435914	0.986729050	1.205328587	0.586966244	0.358681791	0.209985314	0.073376917



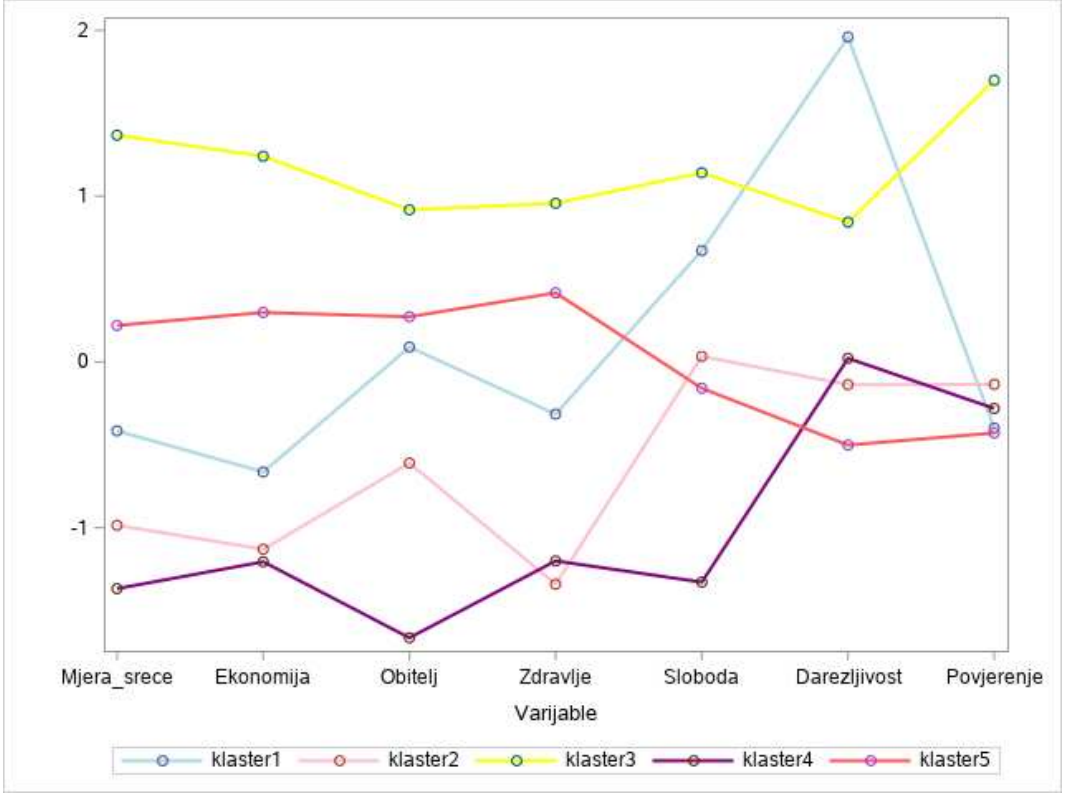
Slika 2.4: Aritmetičke sredine 5 dobivenih klastera po analiziranim varijablama za originalne podatke (SAS ispis)

Slovenija i Rumunjska pripadaju klasteru 3, a ostale države Balkana pripadaju klasteru 5. Klaster 5 je dominantniji u odnosu prema klasteru 3 po svim varijablama. Pogledajmo prosječne vrijednosti varijable *Mjera sreće* na Slici 2.4; najsretnije su države u klasteru 4, zatim u klasterima 3, 5, 2 i 1. Takav trend uočavamo i za varijable *Ekonomija, Obitelj, Zdravlje* i *Sloboda*; klasteri se međusobno slijede u stopu.

Tablica 2.17: Aritmetičke sredine varijabli po klasterima standardiziranih podataka za $k = 5$ (SAS ispis)

Cluster Means							
Cluster	Mjera_srece	Ekonomija	Obitelj	Zdravlje	Sloboda	Darezljivost	Povjerenje
1	-0.415710694	-0.663600919	0.089805637	-0.314396084	0.671806796	1.960177082	-0.399055035
2	-0.985193708	-1.129797333	-0.609473153	-1.340950099	0.033552555	-0.136957767	-0.135477003
3	1.367863379	1.240479856	0.917407825	0.957145792	1.141400479	0.843194631	1.698885351
4	-1.367592635	-1.205335261	-1.663901572	-1.200240970	-1.328112123	0.022075783	-0.280533682
5	0.219940844	0.298777610	0.272433213	0.416752217	-0.158519599	-0.501468788	-0.428479696

Na Slici 2.5 uočavamo dominantnost klastera 3; jedinu razliku čini vrijednost varijable *Darežljivost*. Njegova je dominantnost posebno izražena varijablom *Povjerenje*; kod klastera 1, 2, 4 i 5 razlika je jako mala. Pogledamo li tablice rezultata klasteriranja, vidimo da je riječ o najsretnijim državama svijeta. Iznadprosječnu vrijednost varijable *Darežljivost* uočavamo u klasteru 1; posebno se ističu Mjanmar, Indonezija i Tajland, države jugoistočne Azije. Klaster 4 ima ispodprosječnu vrijednost varijabli *Obitelj* i *Sloboda*; to su uglavnom države supsaharske Afrike, primjerice, Mauretanija, Sudan, Angola, Čad, Malavi. Unatoč tome što su države klastera 5 sretnije i u boljoj ekonomskoj situaciji, imaju ispodprosječnu vrijednost varijable *Darežljivost* u odnosu prema državama iz klastera 4. Među državama klastera 5 nalaze se sve države Balkana. Slika 2.5 također sugerira da najveću razliku među klasterima rade varijable *Mjera sreće* i *Obitelj*.

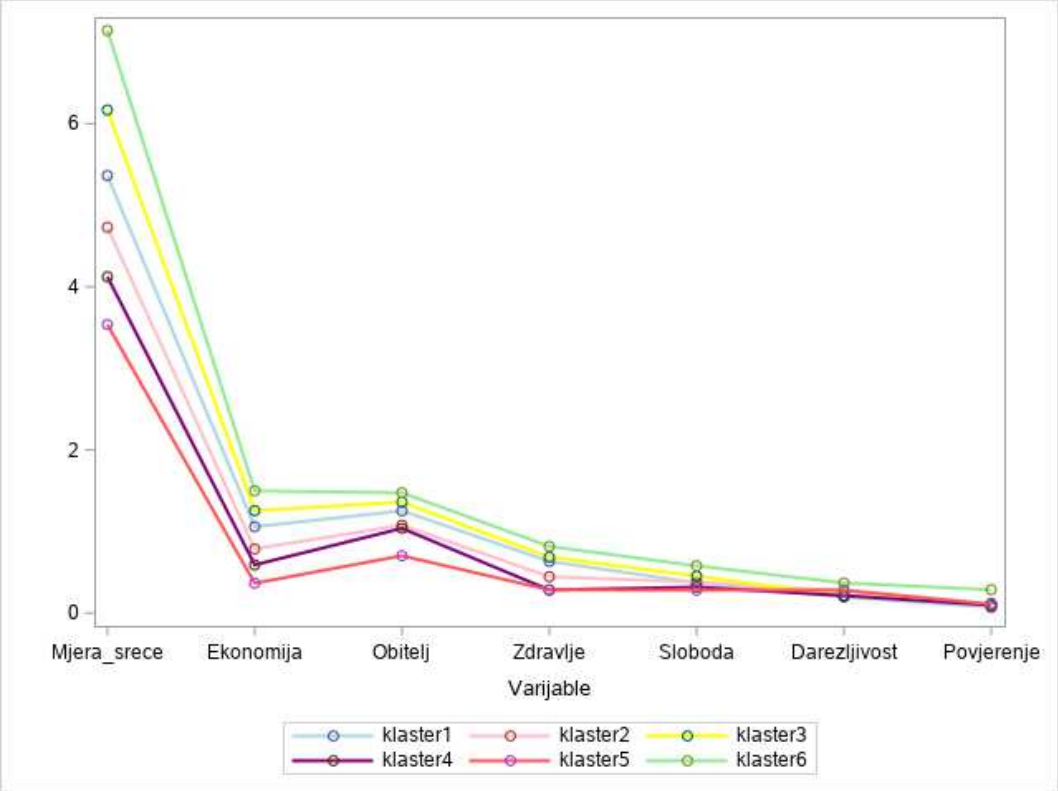


Slika 2.5: Aritmetičke sredine 5 dobivenih klastera po analiziranim varijablama za standardizirane podatke (SAS ispis)

Tablica 2.18: Aritmetičke sredine varijabli po klasterima originalnih podataka za $k = 6$ (SAS ispis)

Cluster Means							
Cluster	Mjera_srece	Ekonomija	Obitelj	Zdravlje	Sloboda	Darezljivost	Povjerenje
1	5.363343760	1.058351494	1.252943058	0.633009753	0.369975683	0.195498432	0.068774093
2	4.728833357	0.785555553	1.077937201	0.444722724	0.379956185	0.269211223	0.084252394
3	6.167999989	1.255197883	1.359767466	0.686000374	0.452966789	0.211739081	0.115605739
4	4.122631575	0.590628697	1.038237343	0.282466122	0.317850147	0.213422652	0.097934138
5	3.537153812	0.365369306	0.702084816	0.283278697	0.280252612	0.280186447	0.109305471
6	7.143238090	1.498273458	1.473283654	0.817519912	0.578557884	0.368106622	0.285708700

Albanija i Bugarska pripadaju klasteru 2, Slovenija i Rumunjska klasteru 3, a ostale države Balkana klasteru 1. Ako usporedimo s rezultatima za $k = 5$, primjećujemo da su se Albanija i Bugarska izdvojile; pridružene su klasteru u kojem prevladavaju države supsaharske Afrike te države Bliskog istoka i sjeverne Afrike. Prema Tablici 2.18, države u klasteru 3 sretnije su i u boljoj ekonomskoj situaciji u odnosu prema državama u klasterima 1 i 2.

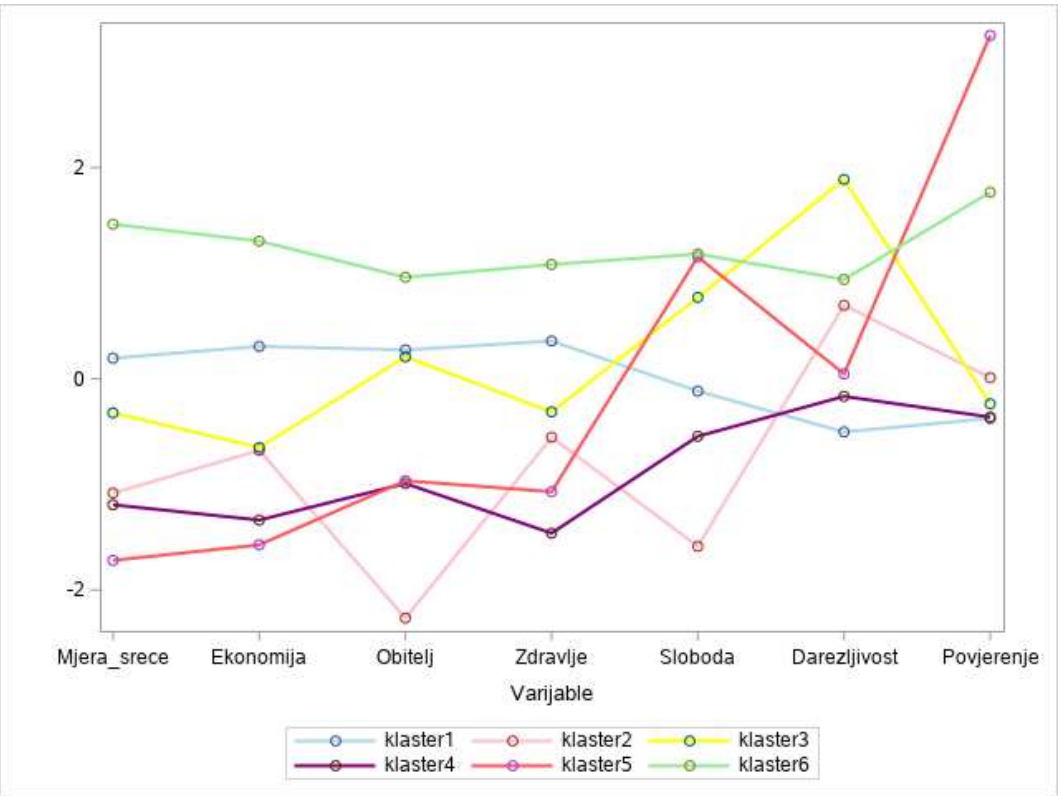


Slika 2.6: Aritmetičke sredine 6 dobivenih klastera po analiziranim varijablama za originalne podatke (SAS ispis)

Tablica 2.19: Aritmetičke sredine varijabli po klasterima standardiziranih podataka za $k = 6$ (SAS ispis)

Cluster Means							
Cluster	**Mjera_srece**	**Ekonomija**	**Obitelj**	**Zdravlje**	**Sloboda**	**Darezljivost**	**Povjerenje**
1	0.195331608	0.306527100	0.273757526	0.357938997	-0.116971454	-0.503763136	-0.375688741
2	-1.082216848	-0.677807505	-2.266008536	-0.553793198	-1.587135883	0.695310498	0.010391845
3	-0.322638229	-0.651117225	0.208153409	-0.312954486	0.770748901	1.887639014	-0.237480310
4	-1.193307750	-1.338092052	-0.992051036	-1.461916129	-0.544705453	-0.166712973	-0.362581400
5	-1.719917066	-1.572678001	-0.966016231	-1.069454716	1.154277592	0.046056754	3.252320815
6	1.463559050	1.304130461	0.961942496	1.082633630	1.182227921	0.940765489	1.766618467

Pogledamo li tablice rezultata klasteriranja, vidimo da su klasteru 6 pridružene najsretnije države svijeta; uglavnom su to države zapadne Europe. Prema Slici 2.7, u klasteru 2 varijable *Obitelj* i *Sloboda* imaju mnogo niže vrijednosti u odnosu prema ostalim klasterima; klasteru 2 pripadaju Afganistan, Haiti, Sirija, Gruzija, Iran i Pakistan. Primijetimo kako klasteri 2 i 4 imaju skoro jednake vrijednosti za varijablu *Mjera sreće*, dok za ostale varijable bilježimo razlike. *Povjerenje* je kod klastera 1, 2, 3 i 4 vrlo slično, dok klasteri 5 i 6 odskaču, posebice klaster 5. Klasteru 5 pridružena je samo jedna država, Ruanda. Vrijednosti varijabla *Ekonomija, Zdravlje* i *Darežljivost* različite su kod svih klastera. Sve države Balkana nalaze se u klasteru 1; iznadprosječne vrijednosti uočavamo za varijable *Mjera sreće, Ekonomija, Obitelj, Zdravlje* i *Sloboda*.

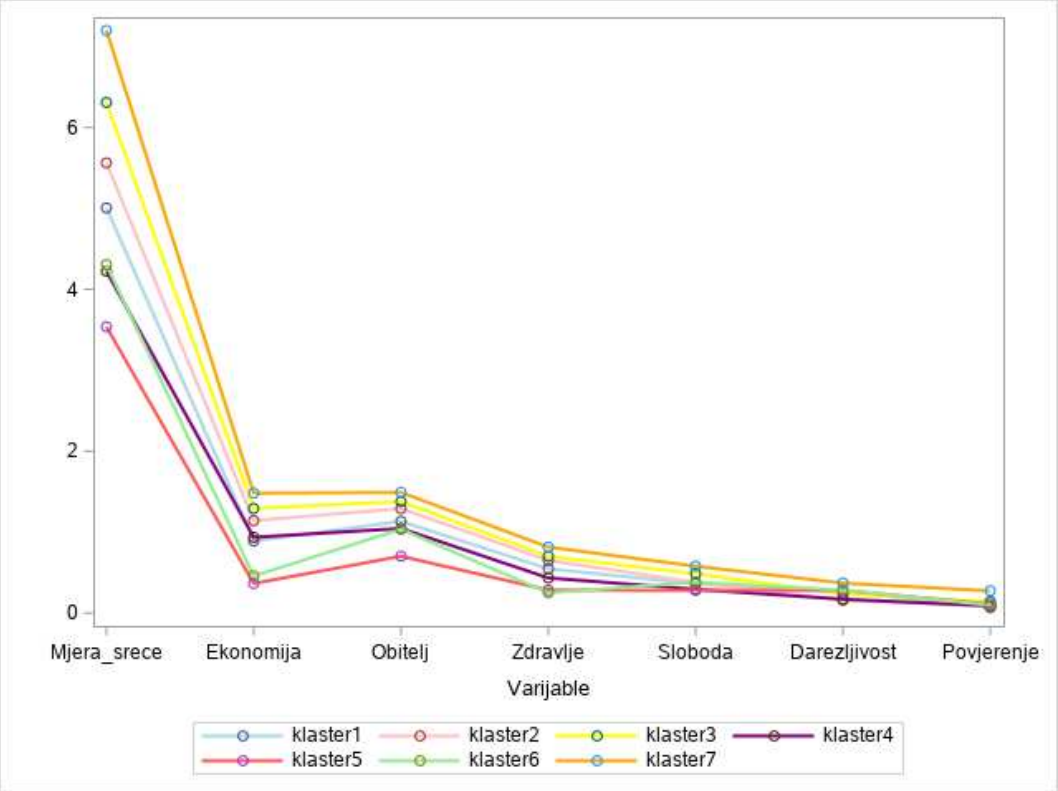


Slika 2.7: Aritmetičke sredine 6 dobivenih klastera po analiziranim varijablama za standardizirane podatke (SAS ispis)

Tablica 2.20: Aritmetičke sredine varijabli po klasterima originalnih podataka za $k = 7$ (SAS ispis)

Cluster Means							
Cluster	**Mjera_srece**	**Ekonomija**	**Obitelj**	**Zdravlje**	**Sloboda**	**Darezljivost**	**Povjerenje**
1	5.005590937	0.889669072	1.133541400	0.546171338	0.362190793	0.265739192	0.068065448
2	5.562363610	1.141623376	1.290514975	0.650993992	0.376833534	0.159027758	0.082334485
3	6.308172423	1.291763797	1.376908315	0.700778237	0.485441309	0.238922116	0.141300626
4	4.224181847	0.936168795	1.044839865	0.433981391	0.292544592	0.170347962	0.089843919
5	3.537153812	0.365369306	0.702084816	0.283278697	0.280252612	0.280186447	0.109305471
6	4.305315783	0.457798483	1.035490337	0.250520153	0.379778294	0.285081912	0.097331981
7	7.199368427	1.481322188	1.490459586	0.815348817	0.578501565	0.369649007	0.274267650

Albanija, Bosna i Hercegovina, Bugarska, Kosovo i Makedonija pripadaju klasteru 1, zajedno s državama supsaharske Afrike, južne Azije te Bliskog istoka i sjeverne Afrike. Ostale države Balkana pridružene su klasteru 2. Prema Tablici 2.20 najveću razliku između klastera 1 i 2 čine varijable *Mjera sreće* i *Ekonomija*.

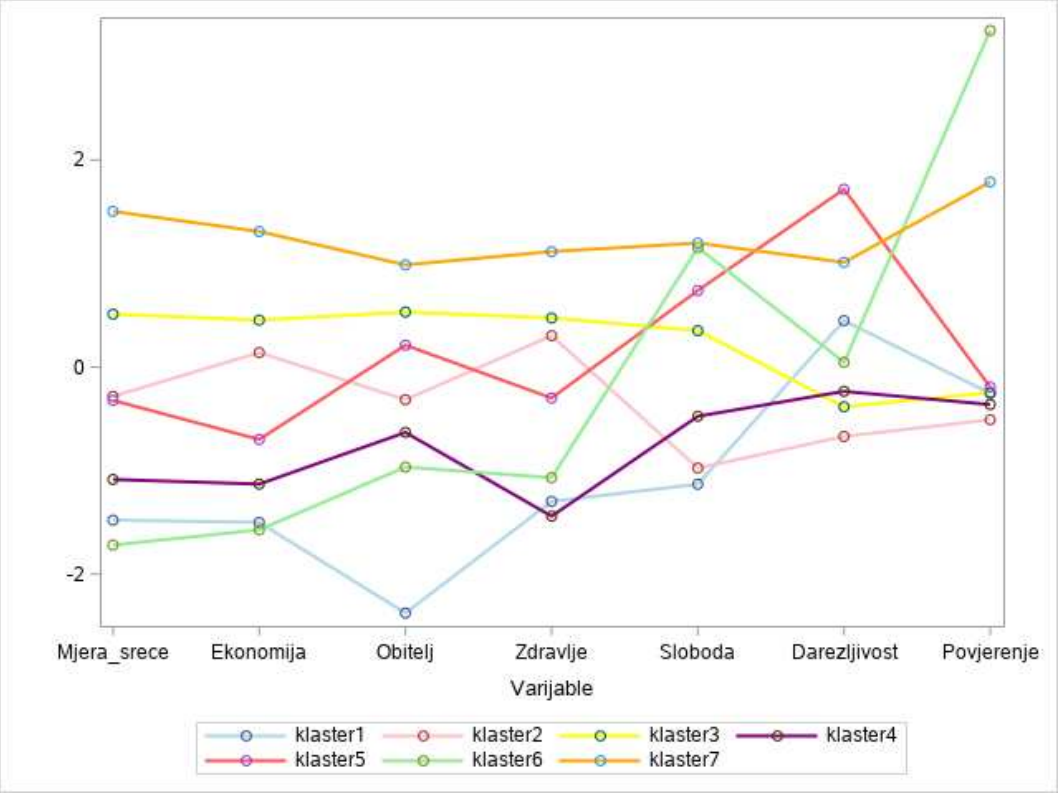


Slika 2.8: Aritmetičke sredine 7 dobivenih klastera po analiziranim varijablama za originalne podatke (SAS ispis)

Tablica 2.21: Aritmetičke sredine varijabli po klasterima standardiziranih podataka za $k = 7$ (SAS ispis)

Cluster Means							
Cluster	**Mjera_srece**	**Ekonomija**	**Obitelj**	**Zdravlje**	**Sloboda**	**Darezljivost**	**Povjerenje**
1	-1.479239841	-1.500741925	-2.376482562	-1.297968632	-1.132235780	0.447937454	-0.254720549
2	-0.282995197	0.140290339	-0.315217098	0.302926780	-0.975760932	-0.670367291	-0.508682422
3	0.510935015	0.453084156	0.531187769	0.473633592	0.352044305	-0.383957214	-0.246567214
4	-1.085991338	-1.130791651	-0.631626417	-1.440805185	-0.475897927	-0.234667045	-0.362376666
5	-0.322069351	-0.699798716	0.211338179	-0.299171427	0.739097407	1.717734151	-0.194091631
6	-1.719917066	-1.572678001	-0.966016231	-1.069454716	1.154277592	0.046056754	3.252320815
7	1.503789744	1.309145546	0.987039784	1.115452813	1.197807494	1.011144688	1.787921836

Prema Slici 2.9, *Povjerenje* je kod klastera 1, 2, 3, 4 i 5 vrlo slično, dok klasteri 6 i 7 odskaču, posebice klaster 6. Klasteru 6 pridružena je samo jedna država, Ruanda; jednak je klasteru 5 za $k = 6$. Klaster 1, osim za *Darežljivost*, poprima ispodprosječne vrijednosti, posebice za varijablu *Obitelj*. Pogledamo li tablice rezultata klasteriranja, vidimo da je uglavnom riječ o državama supsaharske Afrike. Slovenija i Rumunjska pridružene su klasteru 3, koji je dominantniji u odnosu prema klasteru 2, gdje se nalaze ostale države Balkana. Najveću razliku između klastera 2 i 3 uočavamo za varijable *Mjera sreće, Obitelj* i *Sloboda.* Primijetimo kako klasteri 2 i 5 imaju skoro jednake vrijednosti za varijablu *Mjera sreće*, dok za ostale varijable bilježimo razlike.



Slika 2.9: Aritmetičke sredine 7 dobivenih klastera po analiziranim varijablama za standardizirane podatke (SAS ispis)

Pogledajmo rasprostranjenost regija u odnosu prema klasterima. Za originalne podatke iz Tablica 2.22, 2.23 i 2.24 uočavamo sljedeće:

- Neovisno o broju k, Australija i Novi Zeland te Sjeverna Amerika uvijek se nalaze u istom klasteru.
- Većina država srednje i istočne Europe raspoređena je u dva klastera. Međutim, porastom broja k dolazi do raspodjele tih klastera, tj. povećava se rasprostranjenost.
- Za $k = 5, 6$, države istočne Azije dolaze u parovima; Kina i Mongolija te Japan i Južna Koreja. No za $k = 7$, u paru dolaze Kina i Južna Koreja, dok Japan i Mongolija čine zasebne klastere.
- Kod država Južne Amerike i Kariba porastom broja k raste rasprostranjenost država među klasterima, ali uvijek postoji klaster u kojem se nalazi većina država, tj. grupirane su države Argentina, Brazil, Čile, Kolumbija, Ekvador, Meksiko, Panama, Nikaragva, Urugvaj te Trinidad i Tobago. Haiti je u svim slučajevima odvojen od ostalih država Južne Amerike i Kariba.
- Porastom broja k raste rasprostranjenost država Bliskog istoka i sjeverne Afrike.
- Države jugoistočne Azije rasprostranjene su po klasterima u svim slučajevima.
- Zanimljivo je primijetiti kako prelaskom broja k s 5 na 6 dolazi do grupiranja država južne Azije u jedan klaster, no za $k = 7$ dolazi do raspodjele.
- Porastom broja k raste rasprostranjenost država supsaharske Afrike među klasterima.
- Države zapadne Europe raspoređene su u tri klastera; 61,90 % država nalazi se u jednom klasteru, dok su Portugal i Grčka te Španjolska, Italija, Francuska, Malta, Sjeverni Cipar i Cipar u zasebnim klasterima. Prelaskom na $k = 6$, Cipar se pridružuje Portugalu i Grčkoj, a zatim i Sjeverni Cipar za $k = 7$.

Tablice dobivamo pomoću procedure freq, gdje se koriste podatci dobiveni algoritmom k-sredina:

```
proc freq data=klaster5;
    tables Regija*Cluster;
run;
```

Tablica 2.22: Tablica frekvencija originalnih podataka za $k = 5$ (SAS ispis)

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of Regija by CLUSTER

Regija	CLUSTER(Cluster) 1	2	3	4	5	Total
Australia and New Zealand	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	2 1.37 100.00 9.52	0 0.00 0.00 0.00	2 1.37
Central and Eastern Europe	0 0.00 0.00 0.00	3 2.05 10.34 10.00	13 8.90 44.83 30.23	0 0.00 0.00 0.00	13 8.90 44.83 33.33	29 19.86
Eastern Asia	0 0.00 0.00 0.00	0 0.00 0.00 0.00	2 1.37 50.00 4.65	0 0.00 0.00 0.00	2 1.37 50.00 5.13	4 2.74
Latin America and Caribbean	1 0.68 4.76 7.69	0 0.00 0.00 0.00	14 9.59 66.67 32.56	1 0.68 4.76 4.76	5 3.42 23.81 12.82	21 14.38
Middle East and Northern Africa	2 1.37 10.53 15.38	1 0.68 5.26 3.33	5 3.42 26.32 11.63	2 1.37 10.53 9.52	9 6.16 47.37 23.08	19 13.01
North America	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	2 1.37 100.00 9.52	0 0.00 0.00 0.00	2 1.37
Southeastern Asia	0 0.00 0.00 0.00	2 1.37 25.00 6.67	2 1.37 25.00 4.65	1 0.68 12.50 4.76	3 2.05 37.50 7.69	8 5.48
Southern Asia	1 0.68 14.29 7.69	3 2.05 42.86 10.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	3 2.05 42.86 7.69	7 4.79
Sub-Saharan Africa	9 6.16 27.27 69.23	21 14.38 63.64 70.00	1 0.68 3.03 2.33	0 0.00 0.00 0.00	2 1.37 6.06 5.13	33 22.60
Western Europe	0 0.00 0.00 0.00	0 0.00 0.00 0.00	6 4.11 28.57 13.95	13 8.90 61.90 61.90	2 1.37 9.52 5.13	21 14.38
Total	13 8.90	30 20.55	43 29.45	21 14.38	39 26.71	146 100.00

Tablica 2.23: Tablica frekvencija originalnih podataka za $k = 6$ (SAS ispis)

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of Regija by CLUSTER

Regija	CLUSTER(Cluster) 1	2	3	4	5	6	Total
Australia and New Zealand	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	2 1.37 100.00 9.52	2 1.37
Central and Eastern Europe	11 7.53 37.93 34.38	5 3.42 17.24 20.83	11 7.53 37.93 29.73	2 1.37 6.90 10.53	0 0.00 0.00 0.00	0 0.00 0.00 0.00	29 19.86
Eastern Asia	2 1.37 50.00 6.25	0 0.00 0.00 0.00	2 1.37 50.00 5.41	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	4 2.74
Latin America and Caribbean	7 4.79 33.33 21.88	0 0.00 0.00 0.00	12 8.22 57.14 32.43	0 0.00 0.00 0.00	1 0.68 4.76 7.69	1 0.68 4.76 4.76	21 14.38
Middle East and Northern Africa	5 3.42 26.32 15.63	5 3.42 26.32 20.83	5 3.42 26.32 13.51	0 0.00 0.00 0.00	2 1.37 10.53 15.38	2 1.37 10.53 9.52	19 13.01
North America	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	2 1.37 100.00 9.52	2 1.37
Southeastern Asia	3 2.05 37.50 9.38	1 0.68 12.50 4.17	2 1.37 25.00 5.41	1 0.68 12.50 5.26	0 0.00 0.00 0.00	1 0.68 12.50 4.76	8 5.48
Southern Asia	0 0.00 0.00 0.00	5 3.42 71.43 20.83	0 0.00 0.00 0.00	1 0.68 14.29 5.26	1 0.68 14.29 7.69	0 0.00 0.00 0.00	7 4.79
Sub-Saharan Africa	1 0.68 3.03 3.13	8 5.48 24.24 33.33	0 0.00 0.00 0.00	15 10.27 45.45 78.95	9 6.16 27.27 69.23	0 0.00 0.00 0.00	33 22.60
Western Europe	3 2.05 14.29 9.38	0 0.00 0.00 0.00	5 3.42 23.81 13.51	0 0.00 0.00 0.00	0 0.00 0.00 0.00	13 8.90 61.90 61.90	21 14.38
Total	32 21.92	24 16.44	37 25.34	19 13.01	13 8.90	21 14.38	146 100.00

Tablica 2.24: Tablica frekvencija originalnih podataka za $k = 7$ (SAS ispis)

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of Regija by CLUSTER

Regija	CLUSTER(Cluster) 1	2	3	4	5	6	7	Total
Australia and New Zealand	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	2 1.37 100.00 10.53	2 1.37
Central and Eastern Europe	7 4.79 24.14 31.82	14 9.59 48.28 42.42	5 3.42 17.24 17.24	3 2.05 10.34 27.27	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	29 19.86
Eastern Asia	1 0.68 25.00 4.55	2 1.37 50.00 6.06	1 0.68 25.00 3.45	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	4 2.74
Latin America and Caribbean	1 0.68 4.76 4.55	7 4.79 33.33 21.21	11 7.53 52.38 37.93	0 0.00 0.00 0.00	1 0.68 4.76 7.69	0 0.00 0.00 0.00	1 0.68 4.76 5.26	21 14.38
Middle East and Northern Africa	6 4.11 31.58 27.27	4 2.74 21.05 12.12	5 3.42 26.32 17.24	1 0.68 5.26 9.09	2 1.37 10.53 15.38	0 0.00 0.00 0.00	1 0.68 5.26 5.26	19 13.01
North America	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	2 1.37 100.00 10.53	2 1.37
Southeastern Asia	2 1.37 25.00 9.09	1 0.68 12.50 3.03	3 2.05 37.50 10.34	0 0.00 0.00 0.00	0 0.00 0.00 0.00	2 1.37 25.00 10.53	0 0.00 0.00 0.00	8 5.48
Southern Asia	3 2.05 42.86 13.64	0 0.00 0.00 0.00	0 0.00 0.00 0.00	2 1.37 28.57 18.18	1 0.68 14.29 7.69	1 0.68 14.29 5.26	0 0.00 0.00 0.00	7 4.79
Sub-Saharan Africa	2 1.37 6.06 9.09	1 0.68 3.03 3.03	0 0.00 0.00 0.00	5 3.42 15.15 45.45	9 6.16 27.27 69.23	16 10.96 48.48 84.21	0 0.00 0.00 0.00	33 22.60
Western Europe	0 0.00 0.00 0.00	4 2.74 19.05 12.12	4 2.74 19.05 13.79	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	13 8.90 61.90 68.42	21 14.38
Total	22 15.07	33 22.60	29 19.86	11 7.53	13 8.90	19 13.01	19 13.01	146 100.00

Prema Tablicama 2.25, 2.26 i 2.27, za standardizirane podatke uočavamo sljedeće:

- Neovisno o broju k, Australija i Novi Zeland te Sjeverna Amerika uvijek se nalaze u istom klasteru, kao i kod originalnih podataka.
- Za $k = 5$, 86,21 % država srednje i istočne Europe smješteno je u jedan klaster; Kirgistan, Tadžikistan, Uzbekistan i Gruzija smješteni su u zasebne klastere. Prelaskom na $k = 6$, Tadžikistan je pripojen prethodnom velikom klasteru, dok Kirgistan i Uzbekistan imaju svoj klaster, kao i Gruzija. Za $k = 7$ dolazi do spajanja Gruzije s Kirgistanom i Uzbekistanom te do raspodjele velikoga klastera na dva klastera.
- Za $k = 5, 6$, sve države istočne Azije nalaze se u jednom klasteru, dok se za $k = 7$ Mongolija odvaja.
- Za $k = 5, 6$, 95,24 % država Južne Amerike i Kariba unutar je istoga klastera (Haiti se odvojio). Prelaskom na $k = 7$, Venezuela i Honduras se odvajaju u zaseban klaster.
- Porastom broja k raste i broj klastera u kojima se nalaze države Bliskog istoka i sjeverne Afrike. Za $k = 5$, 57,89 % država nalazi se u jednom klasteru i prelaskom na $k = 6$ tom se klasteru pridružuju Kuvajt i Saudijska Arabija. No, za $k = 7$ dolazi do raspodjele tih država.
- Države jugoistočne Azije jednako su raspoređene u tri klastera za sva tri slučaja. Isto primjećujemo i za države južne Azije.
- Za $k = 5$, većina država supsaharske Afrike raspoređena je u dva klastera, no prelaskom na $k = 6$ dolazi do grupiranja u klaster u kojem se nalazi 81,82 % država supsaharske Afrike.
- Države zapadne Europe za $k = 5, 6$ raspoređene su u dva klastera u omjeru $2 : 1$, no za $k = 7$ iz manjega klastera odlazi Grčka u zasebni klaster.

Tablica 2.25: Tablica frekvencija standardiziranih podataka za $k = 5$ (SAS ispis)

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of Regija by CLUSTER

Regija	CLUSTER(Cluster) 1	2	3	4	5	Total
Australia and New Zealand	0 0.00 0.00 0.00	0 0.00 0.00 0.00	2 1.37 100.00 8.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	2 1.37
Central and Eastern Europe	1 0.68 3.45 11.11	1 0.68 3.45 4.55	1 0.68 3.45 4.00	1 0.68 3.45 5.56	25 17.12 86.21 34.72	29 19.86
Eastern Asia	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	4 2.74 100.00 5.56	4 2.74
Latin America and Caribbean	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 0.68 4.76 5.56	20 13.70 95.24 27.78	21 14.38
Middle East and Northern Africa	0 0.00 0.00 0.00	0 0.00 0.00 0.00	5 3.42 26.32 20.00	3 2.05 15.79 16.67	11 7.53 57.89 15.28	19 13.01
North America	0 0.00 0.00 0.00	0 0.00 0.00 0.00	2 1.37 100.00 8.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	2 1.37
Southeastern Asia	4 2.74 50.00 44.44	0 0.00 0.00 0.00	1 0.68 12.50 4.00	0 0.00 0.00 0.00	3 2.05 37.50 4.17	8 5.48
Southern Asia	3 2.05 42.86 33.33	2 1.37 28.57 9.09	0 0.00 0.00 0.00	2 1.37 28.57 11.11	0 0.00 0.00 0.00	7 4.79
Sub-Saharan Africa	1 0.68 3.03 11.11	19 13.01 57.58 86.36	0 0.00 0.00 0.00	11 7.53 33.33 61.11	2 1.37 6.06 2.78	33 22.60
Western Europe	0 0.00 0.00 0.00	0 0.00 0.00 0.00	14 9.59 66.67 56.00	0 0.00 0.00 0.00	7 4.79 33.33 9.72	21 14.38
Total	9 6.16	22 15.07	25 17.12	18 12.33	72 49.32	146 100.00

Tablica 2.26: Tablica frekvencija standardiziranih podataka za $k = 6$ (SAS ispis)

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of Regija by CLUSTER

Regija	CLUSTER(Cluster) 1	2	3	4	5	6	Total
Australia and New Zealand	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	2 1.37 100.00 9.09	2 1.37
Central and Eastern Europe	26 17.81 89.66 33.77	1 0.68 3.45 16.67	2 1.37 6.90 20.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	29 19.86
Eastern Asia	4 2.74 100.00 5.19	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	4 2.74
Latin America and Caribbean	20 13.70 95.24 25.97	1 0.68 4.76 16.67	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	21 14.38
Middle East and Northern Africa	13 8.90 68.42 16.88	2 1.37 10.53 33.33	0 0.00 0.00 0.00	1 0.68 5.26 3.33	0 0.00 0.00 0.00	3 2.05 15.79 13.64	19 13.01
North America	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	2 1.37 100.00 9.09	2 1.37
Southeastern Asia	3 2.05 37.50 3.90	0 0.00 0.00 0.00	4 2.74 50.00 40.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 0.68 12.50 4.55	8 5.48
Southern Asia	0 0.00 0.00 0.00	2 1.37 28.57 33.33	3 2.05 42.86 30.00	2 1.37 28.57 6.67	0 0.00 0.00 0.00	0 0.00 0.00 0.00	7 4.79
Sub-Saharan Africa	4 2.74 12.12 5.19	0 0.00 0.00 0.00	1 0.68 3.03 10.00	27 18.49 81.82 90.00	1 0.68 3.03 100.00	0 0.00 0.00 0.00	33 22.60
Western Europe	7 4.79 33.33 9.09	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	14 9.59 66.67 63.64	21 14.38
Total	77 52.74	6 4.11	10 6.85	30 20.55	1 0.68	22 15.07	146 100.00

Tablica 2.27: Tablica frekvencija standardiziranih podataka za $k = 7$ (SAS ispis)

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of Regija by CLUSTER

Regija	CLUSTER(Cluster) 1	2	3	4	5	6	7	Total
Australia and New Zealand	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	2 1.37 100.00 9.52	2 1.37
Central and Eastern Europe	0 0.00 0.00 0.00	15 10.27 51.72 51.72	11 7.53 37.93 22.92	0 0.00 0.00 0.00	3 2.05 10.34 27.27	0 0.00 0.00 0.00	0 0.00 0.00 0.00	29 19.86
Eastern Asia	0 0.00 0.00 0.00	1 0.68 25.00 3.45	3 2.05 75.00 6.25	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	4 2.74
Latin America and Caribbean	1 0.68 4.76 11.11	2 1.37 9.52 6.90	18 12.33 85.71 37.50	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	21 14.38
Middle East and Northern Africa	1 0.68 5.26 11.11	9 6.16 47.37 31.03	6 4.11 31.58 12.50	1 0.68 5.26 3.70	0 0.00 0.00 0.00	0 0.00 0.00 0.00	2 1.37 10.53 9.52	19 13.01
North America	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	2 1.37 100.00 9.52	2 1.37
Southeastern Asia	0 0.00 0.00 0.00	0 0.00 0.00 0.00	3 2.05 37.50 6.25	0 0.00 0.00 0.00	4 2.74 50.00 36.36	0 0.00 0.00 0.00	1 0.68 12.50 4.76	8 5.48
Southern Asia	2 1.37 28.57 22.22	0 0.00 0.00 0.00	0 0.00 0.00 0.00	2 1.37 28.57 7.41	3 2.05 42.86 27.27	0 0.00 0.00 0.00	0 0.00 0.00 0.00	7 4.79
Sub-Saharan Africa	5 3.42 15.15 55.56	1 0.68 3.03 3.45	1 0.68 3.03 2.08	24 16.44 72.73 88.89	1 0.68 3.03 9.09	1 0.68 3.03 100.00	0 0.00 0.00 0.00	33 22.60
Western Europe	0 0.00 0.00 0.00	1 0.68 4.76 3.45	6 4.11 28.57 12.50	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	14 9.59 66.67 66.67	21 14.38
Total	9 6.16	29 19.86	48 32.88	27 18.49	11 7.53	1 0.68	21 14.38	146 100.00

2.5 Zaključak

Radi primjene klasterske analize korištena je baza podataka koja sadržava informacije o 146 država; sve države ulaze u daljnju analizu. Svaka država pripada određenoj regiji; razlikujemo deset regija. Provedeno je hijerarhijsko i nehijerarhijsko klasteriranje da bi se odredilo kako je Hrvatska u odnosu prema varijablama *Mjera sreće, Ekonomija, Obitelj, Zdravlje, Sloboda, Darežljivost* i *Povjerenje* grupirana s obzirom na ostale države Balkana.

Hijerarhijsko (aglomerativno) klasteriranje provedeno je nad originalnim i standardiziranim podatcima. Korištena je Wardova metoda i metoda maksimuma. Hrvatska se u svim slučajevima pojavljivala sa Srbijom, Crnom Gorom, Kosovom, Makedonijom te Bosnom i Hercegovinom; Albanija, Bugarska, Slovenija i Rumunjska udaljenije su od njih. Zatim je primijenjen algoritam k-sredina nad originalnim i standardiziranim podatcima, uz ograničenje od 5, 6 i 7 klastera.

Tablica 2.28: Rezultati algoritma k-sredina za države Balkana

<table>
<tr><th></th><th>originalni podatci</th><th>standardizirani podatci</th></tr>
<tr><td rowspan="2">k=5</td><td>Slovenija i Rumunjska</td><td rowspan="2">Albanija, Bugarska, BiH, Crna Gora, Hrvatska, Kosovo, Makedonija, Rumunjska, Slovenija, Srbija</td></tr>
<tr><td>Albanija, Bugarska, BiH, Crna Gora, Hrvatska, Kosovo, Makedonija, Srbija</td></tr>
<tr><td rowspan="3">k=6</td><td>Slovenija i Rumunjska</td><td rowspan="3">Albanija, Bugarska, BiH, Crna Gora, Hrvatska, Kosovo, Makedonija, Rumunjska, Slovenija, Srbija</td></tr>
<tr><td>Albanija, Bugarska</td></tr>
<tr><td>BiH, Crna Gora, Hrvatska, Kosovo, Makedonija, Srbija</td></tr>
<tr><td rowspan="2">k=7</td><td>Albanija, Bugarska, BiH, Kosovo, Makedonija</td><td>Slovenija i Rumunjska</td></tr>
<tr><td>Crna Gora, Hrvatska, Rumunjska, Slovenija, Srbija</td><td>Albanija, Bugarska, BiH, Crna Gora, Hrvatska, Kosovo, Makedonija, Srbija</td></tr>
</table>

Prema Tablici 2.28 zaključujemo da za originalne podatke algoritam k-sredina pronalazi razlike između država Balkana te ih stoga raspoređuje u različite klastere. Naime, za $k = 5, 6$ Slovenija i Rumunjska se odvajaju od ostalih država Balkana. Pogledamo li podatke u Tablici 2.1 za države Balkana, vidimo da je Slovenija iznadprosječna s obzirom na ostale zemlje Balkana, ali Rumunjska je najsretnija i iznadprosječna s obzirom na ekonomsku situaciju i slobodu donošenja životnih odluka. Pogledamo li klastere 3 i 5 u Tablici 2.16 te klastere 3 i 1 u Tablici 2.18, možemo zaključiti da su Slovenija i Rumunjska odvojene zbog iznadprosječnih vrijednosti varijabla *Mjera sreće* i *Ekonomija* s obzirom na ostatak Balkana; upravo te varijable rade najveću razliku između klastera. Suprotni rezultat dobiven je za standardizirane podatke; tek za $k = 7$ algoritam pronalazi razlike između država, gdje se ponovno odvajaju Slovenija i Rumunjska. Pogledamo li klastere 2 i 3 na Slici 2.9, uz varijable *Mjera sreće* i *Ekonomija*, razliku radi i varijabla *Sloboda* po kojoj se Slovenija i Rumunjska ističu u odnosu prema ostatku Balkana.

Primjenom algoritma k-sredina nad standardiziranim podatcima regije su raspoređene po klasterima, uz nekoliko iznimki. Pogledamo li dobivene klastere, primjećujemo da većinski dio država neke regije uvijek pripada određenom klasteru. Primjerice, to vrijedi za države Bliskog istoka i sjeverne Afrike kada je $k = 5, 6$, no za $k = 7$ ne možemo reći da države većinski pripadaju nekom klasteru. Iznimku čine države jugoistočne i južne Azije. Naime, ni u jednom se klasteru ne nalazi više od 50 % država južne Azije.

Raspoređenost regija po klasterima za originalne podatke manja je u odnosu prema standardiziranim podatcima. Naime, države istočne Azije, jugoistočne Azije, srednje i istočne Europe te Bliskog istoka i sjeverne Afrike rasprostranjene su po gotovo svim klasterima. Države supsaharske Afrike samo za $k = 5$ većinski pripadaju određenom klasteru, kao i države južne Azije za $k = 6$. Stoga, ako gledamo raspoređenost regija po klasterima, dobit ćemo bolje rezultate za standardizirane podatke.

Pogledamo li Slike 2.5, 2.7 i 2.9, možemo zaključiti da varijabla *Mjera sreće* ne utječe na klasteriranje kao ostale varijable; razliku između klastera uglavnom rade varijable *Obitelj, Zdravlje, Sloboda, Darežljivost* i *Povjerenje*. Također možemo zaključiti da ekonomska situacija nije presudni faktor za sreću. Hrvatska je unatoč relativno dobroj ekonomskoj situaciji ispodprosječno sretna; do izražaja dolazi manjak potpore i suosjećajnosti unutar obitelji i društva. Uz to, pripada korumpiranijim državama svijeta, što ostavlja utjecaj i na velikodušnost građana; nisu skloni čestim donacijama zbog raširenosti korupcije u raznim područjima.

Bibliografija

[1] *SAS/STAT® 13.2 User's Guide The FASTCLUS Procedure*, https://support.sas.com/documentation/onlinedoc/stat/132/fastclus.pdf, (pristupljeno: studeni 2020.).

[2] *SAS/STAT® 9.2 User's Guide The CLUSTER Procedure*, https://support.sas.com/documentation/cdl/en/statugcluster/61777/PDF/default/statugcluster.pdf, (pristupljeno: studeni 2020.).

[3] S. Aghabozorgi, A. S. Shirkhorshidi i T. Y. Wah, *A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data*, Public Library of Science since (2015.), 645–678.

[4] S. Bhirudd, M. Chandanec, P. Nerurkara i A. Shirkeb, *Empirical Analysis of Data Clustering Algorithms*, Procedia Computer Science (2018.), 1–10.

[5] S. Bittanti, D. L. Boley, G. Gazzaniga i S. M. Savaresi, *Choosing the cluster to split in bisecting divisive clustering algorithms*, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.25.3507&rep=rep1&type=pdf, (pristupljeno: listopad 2020.).

[6] S. S. Dimov, C. D. Nguyen i D. T. Pham, *Selection of K in K-means clustering*, Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science (2005.), 103–119.

[7] B. S. Everitt, S. Landau, M. Leese i D. Stahl, *Cluster Analysis*, WILEY, 2011.

[8] J. Friedman, T. Hastie i R. Tibshirani, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, 2008.

[9] J. Han, M. Kamber i J. Pei, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2012.

[10] P. Hansen i B. Jaumard, *Cluster Analysis and Mathematical Programming*, Mathematical Programming **79** (1997), 191–215.

[11] J. F. Helliwell, H. Huang i S. Wang, *Statistical Appendix for "The social foundations of world happiness"*, https://s3.amazonaws.com/happiness-report/2017/StatisticalAppendixWHR2017.pdf, (pristupljeno: studeni 2020.).

[12] J. F. Helliwell, R. Layard i J. Sachs, *World happiness report 2017*, https://worldhappiness.report/ed/2017/, (pristupljeno: studeni 2020.).

[13] R. Xu i D. Wunsch II, *Survey of Clustering Algorithms*, IEEE Transactions on Neural Networks and Learning Systems **16** (2005.), 645–678.

Sažetak

U ovom radu opisana je klasterska analiza i njezina primjena na bazu podataka koja sadržava informacije o 146 država. Klasterska analiza je metoda grupiranja objekata u klastere. Svaki je klaster grupa objekata najveće sličnosti. Ovisno o vrsti objekata i cilju istraživanja, odabiremo odgovarajuću mjeru sličnosti ili različitosti (udaljenosti). Nakon toga odabiremo algoritam klasteriranja; razlikujemo hijerarhijske i nehijerarhijske algoritme.

Korištena baza dobivena je iz Svjetskog izvještaja o sreći (*World Happiness Report*) za 2017. godinu. Svakoj državi pridružene su vrijednosti varijabla *Mjera sreće, Ekonomija, Obitelj, Zdravlje, Sloboda, Darežljivost* i *Povjerenje*. S obzirom na to da su sve varijable numeričke, kao mjera udaljenosti odabrana je euklidska udaljenost. Najprije je provedeno hijerarhijsko (aglomerativno) klasteriranje pomoću Wardove metode i metode maksimuma. Zatim je provedeno nehijerarhijsko klasteriranje, tj. algoritam k-sredina za $k = 5, 6, 7$. Za obradu i analizu podataka korišten je programski sustav SAS.

Summary

This paper describes cluster analysis and its application to a database containing information on 146 countries. Cluster analysis is a method of grouping objects into clusters. Each cluster is a group of objects with the greatest similarity. Depending on the type of objects and the goal of the research, we choose the appropriate measure of similarity or difference (distance). After that, it is necessary to select the clustering algorithm; there are two types, hierarchical and non-hierarchical algorithms.

Used database was obtained from the 2017 World Happiness Report. Value of variables *Measure of Happiness, Economy, Family, Health, Freedom, Generosity* and *Trust*, are assigned to each country. Since all variables are numerical, Euclidean distance was chosen as a measure of distance. Firstly, hierarchical (agglomerative) clustering was conducted using the Ward method and the maximum method. Next, non-hierarchical clustering was made with k-means algorithm for $k = 5, 6, 7$. The SAS software system was used for data processing and analysis.

Životopis

Rođena sam 19. svibnja 1994. godine u Zagrebu. Osnovnu školu Medvedgrad u Zagrebu upisala sam 2001. godine. Svoje srednjoškolsko obrazovanje započela sam 2009. godine u II. gimnaziji u Zagrebu. Tijekom osnovne i srednje škole bila sam aktivna sportašica. Nakon završene srednje škole, 2013. godine upisujem preddiplomski sveučilišni studij Matematika na Prirodoslovno-matematičkom fakultetu u Zagrebu na kojem 2018. godine stječem titulu *univ. bacc. math.* Iste godine upisujem diplomski studij Matematičke statistike, također na Prirodoslovno-matematičkom fakultetu u Zagrebu. Početkom akademske 2017./2018. godine postajem članica Studentskog zbora Prirodoslovno-matematičkog fakulteta; godine 2017./2018. izabrana sam za zamjenicu predsjednika Studentskog zbora, a akademske 2018./2019. godine izabrana sam za predsjednicu Studentskog zbora. Sudjelovala sam i u organizaciji projekta STEM Games 2018, međunarodnog studentskog natjecanja u sportu i znanju. Tijekom svojeg studiranja obavljala sam različite studentske poslove. Trenutačno radim kao *Data Scientist* u IT tvrtki Qualia.