METODOLOŠKI ZVEZKI ŠT. 4

RAZVRŠČANJE V SKUPINE

Teorija in uporaba v družboslovju

Anuška Ferligoj

Ljubljana 1989

Anuška Ferligoj: RAZVRŠČANJE V SKUPINE

Zbirka METODOLOŠKI ZVEZKI, št. 4

Izdajatelj: Jugoslovansko združenje za sociologijo,
Sekcija za metodologijo in statistiko
Ureja Anuška Ferligoj

Recenzent: prof.dr. Konstantin Momirović

Založnik: Raziskovalni inštitut, Fakulteta za sociologijo, politične vede in novinarstvo, Kardeljeva pl. 5, Ljubljana

Tisk: Edvard Usenik, Kadilnikova 8, Ljubljana

Naklada: 300 izvodov

Po sklepu Komiteja za kulturo SRS št. 4210-31/88 z dne 19.1.1988 je zbirka Metodološki zvezki oproščena temeljnega in posebnega davka od prometa proizvodov.

Kazalo

Predgovor

Metode za sistematično raziskovanje so v družboslovju kakor tudi v drugih znanostih zelo podobne. Te v splošnem obsegajo razpoznavanje in formulacijo problemov, zbiranje ustreznih empiričnih podatkov (preko opazovanja ali eksperimenta) in največkrat uporabo matematičnih in statističnih metod za razkrivanje zvez med podatki ali za preverjanje postavljenih domnev o proučevanih pojavih. Seveda obstajajo specifični problemi in težave v družboslovnih znanostih, ki so (morda) manj izrazite v naravoslovnih vedah, kot sta fizika in kemija. Tako je na primer merjenje v fiziki v splošnem precej preprostejše in zanesljivejše kot v družboslovju, kar je med drugim posledica zapletenosti, nejasnosti, dvoumnosti nekaterih vidikov človekovega vedenja. Zaradi teh specifičnih težav družboslovci v splošnem potrebujejo drugačno, ponavadi kompleksnejše in temu ustrezno zapletenejše analitično orodje za analizo podatkov. Med te sodijo tudi metode za razvrščanje v skupine, ki jih obravnava ta knjiga.

Razvrščanje objektov (ali česa drugega) v skupine, tako da so objekti znotraj skupin kar čim bolj podobni med seboj in objekti različnih skupin kar čim bolj različni med seboj, je zelo star, intuitivno preprost in razumljiv problem. Bolj ali manj vešče so ga reševali že stari Grki in rešujemo ga še danes. Problem

razvrščanja v skupine je bil do pred nekaj desetletji reševan ločeno v posameznih znanstvenih disciplinah, ne da bi se tako dobljeni rezultati povezovali in dopolnjevali. To je značilno za začetne faze izgradnje določene teorije. V šestdesetih letih je zaznati prve poskuse združitve različnih pristopov reševanja problema razvrščanja v skupine in v letu 1963 prvo obsežnejše delo *Sokala in Sneatha* iz tega področja. Od tedaj se področje razvrščanja v skupine razvija kot samostojna disciplina znotraj multivariatne analize.

Revolucijo v analizi podatkov in s tem v razvoju in uporabi kompleksnejših metod so omogočili predvsem računalniki. Pred tridesetimi leti je moral družboslovec, ki je hotel za svoje podatke z manjšim številom enot in spremenljivk uporabiti eno od metod razvrščanja v skupine, računati s tedaj dostopnimi namiznimi računskimi stroji več dni. Danes pa je to delo mogoče opraviti z zelo učinkovitimi statističnimi računalniškimi programskimi paketi, kot so SAS, GENSTAT, SPSS, BMDP itd., takorekoč v hipu. Ta možnost enostavne uporabe zapletenih metod za analizo podatkov tudi brez ali le z delnim poznavanjem uporabljenih metod ima tudi negativne posledice, ki se kažejo predvsem v prepogostem produciranju nesmiselnih in napačnih raziskovalnih rezultatov v družboslovju, pa tudi drugje. Na neke vrste streznitev po evforični uporabi metod, ki jih uporabniki niso dovolj razumeli, kaže povečano zanimanje uporabnikov za razumevanje metod za analizo podatkov. In tem je namenjana ta knjiga.

V tej knjigi, ki obravnava načine reševanja problemov razvrščanja v skupine, so najprej predstavljeni običajni problemi razvrščanja v skupine z osnovnimi pojmi, razvoj področja razvrščanja v skupine ter proces reševanja teh problemov, ki gre od izbora objektov in spremenljivk, preko merjenja podobnosti do izbire primerne metode razvrščanja v skupine. Merjenju podobnosti med enotami je posvečeno posebno poglavje. Po matematizaciji pro-

blema razvrščanja v skupine so obravnavane posamezne metode v smislu njihovih predpostavk, logike metode, lastnosti dobljenih rešitev itd. Podrobneje so predstavljene metode hierarhičnega združevanja v skupine in dva tipa nehierarhičnih metod: metoda prestavljanj in metoda voditeljev. Ob obravnavi posameznih metod je poudarek predvsem na vprašanju, kdaj so posamezne metode uporabne in na konkretni interpretaciji dobljenih rezultatov. S tem upam, da si bo bralec pridobil dovolj znanja, da bodo zaključki njegovih bodočih analiz z metodami razvrščanja v skupine bolj zanesljivi in veljavni. Na razvojni poti teorije razvrščanja v skupine so se izoblikovali različni tipi problemov, ki so pogosti in jih ne moremo reševati s standardnimi metodami. V dveh poglavjih sta obravnavana dva taka tipa problemov, ki sta ožje področje mojega raziskovanja: razvrščanje v skupine z omejitvami in večkriterijsko razvrščanje v skupine. Seveda pa v tej knjigi nisem mogla zajeti vseh znanih metod in pristopov, ki se kar vrstijo v zadnjem desetletju. Na zadnjih straneh tega dela je podan obsežen spisek literature s področja razvrščanja v skupine, ki lahko koristi zahtevnejšemu bralcu pri razkrivanju tega zanimivega dela multivariatne analize.

Ob pisanju te knjige sem imela pred očmi predvsem bralca z nematematično izobrazbo. Zato sem se ves čas trudila, da bi bilo besedilo napisano čim bolj razumljivo. Včasih pa je potrebno uporabiti tudi formalnejši jezik. Tedaj sem s primeri ponazorila predstavljeno misel.

Ta knjiga je rezultat mojega dolgoletnega ukvarjanja s področjem razvrščanja v skupine, tako v smislu razvoja in uporabe metod, kakor tudi dolgoletnega poučevanja na Fakulteti za sociologijo, politične vede in novinarstvo in drugih fakultetah na Univerzi v Ljubljani. Skupaj z Vladimirjem Batageljem že približno petnajst let razkrivava in raziskujeva področje razvrščanja

v skupine. Sredi sedemdesetih let sem za določitev tipologije občin SR Slovenije prvič uporabljala metode razvrščanja v skupine in tedaj dostopne ustrezne računalniške programe. Ob tem sem se Batagelju zasmilila, ker so bili ti programi resnično neprijazni do uporabnika. Pričel je s programiranjem svojega sedaj že znanega paketa programov za razvrščanje v skupine CLUSE. S tem je bilo zgrajeno osnovno orodje za samostojno raziskovalno delo, ki se je začelo z odločitvijo za optimizacijski pristop k razvrščanju, z empirično primerjavo tedaj znanih metod razvrščanja v skupine in raziskovanjem omejitev. Na to raziskovanje so pomembno vplivala spoznanja, do katerih sem prišla ob sodelovanju pri različnih raziskavah na področju družboslovja, medicine in drugod. Razvrščanje v skupine sem predavala večkrat. Študentom sem leta 1982 pripravila zapiske predavanj, ki so bili osnova za pisanje te knjige. Pedagoške izkušnje, še posebej na seminarjih za manjše število vedoželjnih študentov na FSPN, so bile še posebej dragocene pri zasnovi te knjige. Z največjim zadovoljstvom sodelujem tudi s prof. Branislavom Ivanovićem, prof. Konstantinom Momirovićem, prof. Srdjanom Bogosavljevićem in drugimi v okviru Sekcije za klasifikacije Jugoslovanske Zveze statističnih društev in ob drugih priložnostih. To sodelovanje je vedno dragocena vzpodbuda za moje nadaljnje raziskovalno delo na področju razvrščanja v skupine.

Profesorjem Vladimirju Batagelju, Branislavu Ivanoviću, Konstantinu Momiroviću, Srdjanu Bogosavljeviću, mojim študentom in kolegom se najlepše zahvaljujem za vzpodbude in veselje ob delu na področju razvrščanja v skupine.

Posebna zahvala pa gre mami Lauri, ki mi je omogočila pogoje, da sem lahko zbrano pripravila in pisala to knjigo.

Piran, avgust 1989 Anuška Ferligoj

1.

Uvod

1.1 Osnovni pojmi

Urejanje ali razvrščanje podobnih reči v skupine je najbrž ena od najstarejših človekovih mentalnih aktivnosti. V najširšem pomenu je razvrščanje v skupine proces abstrakcije poimenovanja skupin objektov, za katere menimo, da so na nek način podobni med seboj. Proces razvrščanja v skupine je pomembno vplival na razvoj več znanstvenih disciplin. Mogoče so najpomembnejši rezultati tega procesa Darwinova razvojna teorija v biologiji, razvrstitev kemijskih elementov v znano Mendeljejevo tabelo in Marxova zgodovinska periodizacija razrednih družb.

Nalogo razvrščanja v skupine lahko intuitivno zastavimo takole: dane objekte je potrebno razvrstiti v nekaj skupin med seboj (znotraj skupine) podobnih objektov. Množico iskanih skupin imenujemo razvrstitev.

Predno začnemo reševati posamezen problem razvrščanja v skupine, ga moramo čimbolj natančno vsebinsko proučiti in na osnovi tega sestaviti čimbolj ustrezen (formalen) opis problema. V

opisu problema običajno določimo lastnosti, ki ustrezno opisujejo proučevane objekte in na osnovi katerih želimo objekte razvrstiti v skupine. Tako dobljenim opisom objektov pravimo enote. Torej i-ta enota X_i je nabor vrednosti izmerjenih spremenljivk (urejena m-terka)

$$X_i = (x_{i1}, x_{i2}, ..., x_{im})$$

kjer je x_{ij} vrednost j-te spremenljivke za i-ti objekt. Enote predstavljajo izhodišče za nadaljnjo obravnavo problema razvrščanja in zato izbor spremenljivk odločilno vpliva na smiselnost dobljenih razvrstitev glede na zastavljeni problem. Iz značaja problema razvrščanja v skupine je potrebno tudi čimbolj natančno opredeliti, kakšne razvrstitve so smiselne, kakšne najboljše. To lahko storimo tako, da karseda natančno opredelimo množico dopustnih razvrstitev (število skupin, ali gre za prekrivajoče skupine, itd.) in kriterije razvrščanja.

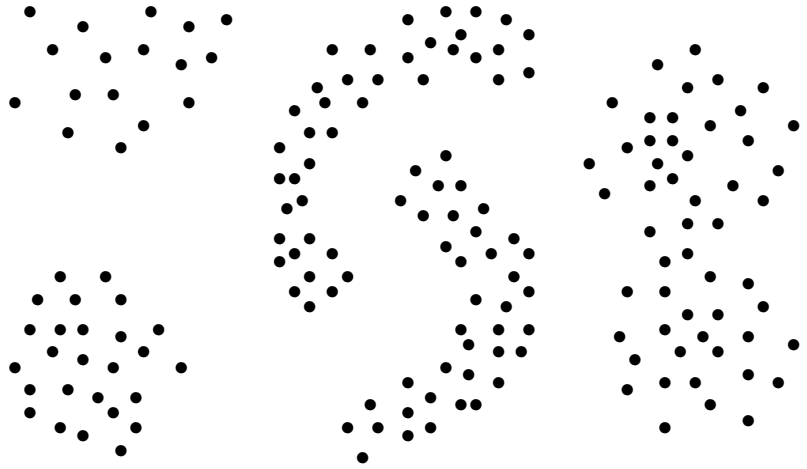
Nalogo razvrščanja v skupine lahko torej zastavimo takole: Množico enot je potrebno po izbranem kriteriju razvrstiti v nekaj skupin. Pri tem je treba običajno določiti, kakšne razvrstitve želimo.

Kot primer vzemimo razvrščanje občin v SR Sloveniji glede na njihovo družbeno-ekonomsko razvitost. Enote torej v tem primeru določajo spremenljivke, ki merijo družbeno-ekonomsko razvitost (npr. družbeni proizvod na prebivalca, število zaposlenih v gospodarstvu na 100 prebivalcev, število zaposlenih v šolstvu, prosveti in kulturi na 100 prebivalcev, število bolniških postelj na 1000 prebivalcev, nataliteta, mortaliteta dojenčkov, število telefonskih naročnikov na 100 prebivalcev). V tem primeru nas lahko zanima le skupina bolj razvitih in skupina manj razvitih območij v Sloveniji. Tedaj iskano razvrstitev določata dve neprekrivajoči se skupini, kjer je vsaka občina razvrščena natanko v eno skupino (taki razvrstitvi pravimo popolna razvrstitev). Občine znotraj

posamezne skupine si morajo biti karseda podobne glede na merjene družbeno-ekonomske spremenljivke. Glede na ta pogoj izberemo ustrezen kriterij, s pomočjo katerega lahko objektivno izberemo najboljšo razvrstitev izmed vseh možnih popolnih razvrstitev v dve skupini.

V primeru razvrščanja občin v skupine glede na njihovo družbeno-ekonomsko razvitost nas lahko zanimajo tudi bolj specifično določene skupine, na primer regije. V tem primeru so smiselne razvrstitve le tiste, kjer so občine v posamezni skupini geografsko sosedne. Tedaj se množica vseh dopustnih razvrstitev, med katerimi iščemo najboljšo, zoži le na tiste, kjer je zadoščen pogoj geografske sosednosti. Brez upoštevanja te pomembne lastnosti iskane razvrstitve je lahko rezultat procesa razvrščanja povsem neuporaben.

Razvrstitve enot, ki so določene z vrednostjo ene ali več (številskih) spremenljivk, je mogoče razbrati iz grafičnih predstavitev enot s točkami v eno- ali več-razsežnem prostoru, kjer je vsaka izmed razsežnosti določna z eno spremenljivko. Skupine lahko v tem prikazu razberemo takole: skupino sestavljajo relativno gosto posejane točke, ki so obkrožene s praznim prostorom ali z relativno redko posejanimi točkami. Tako določenim skupinam pravimo naravne skupine (npr. Everitt 1974, str. 44). Cormack (1971) in kasneje Gordon (1981, str. 5) sta za razkritje naravnih skupin podala dve želeni lastnosti skupin: *interno kohezivnost* (homogenost) in *eksterno izolacijo* (ločenost). Grafično lahko predstavimo ti dve lastnosti na treh tipičnih razvrstitvah, kjer so enote določene z dvema spremenljivkama in prikazane v dvorazsežnem prostoru (glej sliko 1.1). Iz primerov je razvidno, da ni nujno, da skupine zadoščajo obema želenima lastnostima. V primeru (a) sta skupini kohezivni in izolirani, v primeru (b) sta izolirani, a ne kohezivni, ker sta točki na začetku in na koncu 'klobase' bolj oddaljeni kot



(a) (b) (c)

Slika 1.1: Trije tipi razvrstitev

rep.,pok.		ZKJ	ZZBNOV
B i H	B	7.09	2.83
Črna gora	Č	11.00	6.01
Hrvatska	H	6.67	5.77
Makedonija	M	6.48	3.65
Slovenija	S	6.07	6.84
Ožja Srbija	O	9.13	5.64
Kosovo	K	5.17	2.88
Vojvodina	V	9.59	4.85

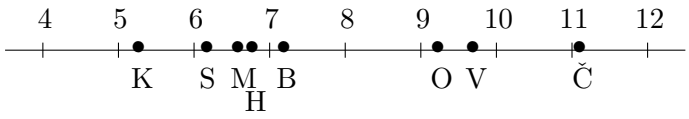
Tabela 1.1: Odstotek članov v ZKJ in ZZBNOV v letu 1978

točki na začetkih obeh 'klobas', v zadnjem primeru (c) pa sta kohezivni, vendar ne izolirani, ker ju veže nekaj točk. V običajnih primerih razvrščanja v skupine gre za razkritje struktur podatkov, podobnih primeru (c), kjer so skupine sicer homogene, vendar ne izrazito ločene med seboj.

Denimo, da želimo razvrstiti republike in pokrajini v skupine glede na odstotek članov ZKJ in ZZBNOV v celotnem prebivalstvu. Podatki za leto 1978 (Vir: Statistični koledar Jugoslavije 1980) so podani v tabeli 1.1.

Najprej razvrstimo republike in pokrajini v skupine glede na odstotek članov v ZKJ, tako da skupine razberemo iz grafičnega prikaza enot s točkami na premici. Iz slike 1.2 je razvidno, da se republike in pokrajini izrazito gostijo v dve skupini in sicer v skupino z relativno manjšim odstotkom članov v ZKJ, ki jo sestavljajo Kosovo, Slovenija, Makedonija, Hrvatska ter Bosna in Hercegovina, in skupino z relativno večjim odstotkom, v kateri so ožja Srbija, Vojvodina in Črna gora.

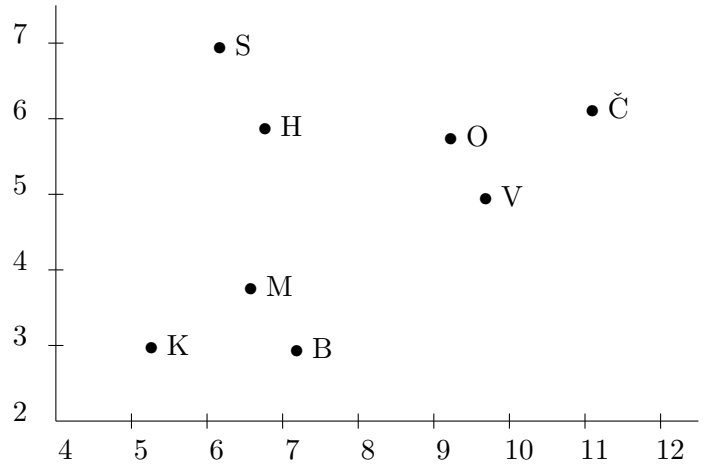
Razvrstitev republik in pokrajin glede na odstotek članov v



Slika 1.2: Republike in pokrajini glede na odstotek članov v ZKJ

obeh družbeno-političnih organizacijah je mogoče razkriti z grafičnim prikazom v dvorazsežnem prostoru, kjer je prva razsežnost določena s prvo spremenljivko (člani v ZKJ) in druga razsežnost z drugo spremenljivko (člani v ZZBNOV). S pomočjo grafične ponazoritve, ki je podana na sliki 1.3, je mogoče republike in pokrajini razvrstiti v tri skupine: skupina z relativno majhnim odstotkom članov v ZKJ in ZZBNOV (Kosovo, Makedonija, Bosna in Hercegovina), skupina z relativno majhnim odstotkom članov v ZKJ in relativno večjim odstotkom v ZZBNOV (Slovenija, Hrvatska) in skupina z relativno večjim odstotkom članov v obeh organizacijah (Črna gora, Ožja Srbija, Vojvodina).

Za razvrščanje enot z več spremenljivkami je opisana grafična metoda neuporabna, ker je težko grafično predstaviti več kot trirazsežni prostor. V takih primerih, ki so sicer najpogostejši, ne vemo, kako se podatki strukturirajo v večrazsežnem prostoru (npr. koliko naravnih skupin se kaže v strukturi, za kakšen tip skupin gre glede na kohezivnost in izoliranost). Razkrivanje neznane strukture proučevanih enot je prav gotovo izziv, vreden resnega raziskovanja. In prav to je področje, ki ga obravnava ta knjiga. V takih primerih je namreč potrebno uporabiti drugačne, multivariatnemu načinu obravnave ustrezne in ponavadi zapletenejše matematično računalniške metode.



Slika 1.3: Republike in pokrajini glede na odstotek članov v ZKJ in ZZBNOV

1.2 Razvoj področja razvrščanja v skupine

Z razvrščanjem v skupine so se ukvarjali že v antiki (Aristotel, Galen,...). Med najpomembnejše dosežke na tem področju sodi prav gotovo drevo živih bitij. Sprva so se z razvojem postopkov razvrščanja v skupine ukvarjali predvsem strokovnjaki s področij, v katerih se je problem razvrščanja pojavljal. Tako so se na primer v biologiji postopki razvrščanja v skupine razvijali že v 18. stoletju (Adanson), v psihologiji sta se med prvimi ukvarjala s postopki razvrščanja v skupine Zubin (1938) in Tryon (1939), v antropologiji pa Driver in Kroeber (1932). Čeprav je problem razvrščanja v skupine zelo star, je prvo delo, ki je urejeno povzelo različne pristope za njegovo reševanje, izšlo šele v letu 1963 (Sokal in Sneath 1963). Področje razvrščanja v skupine se od tedaj izredno hitro razvija. To potrjuje naraščajoči delež člankov s to tematiko v teoretičnih in uporabnih statističnih in drugih revijah v zadnjih desetletjih (naj navedem le najodzivnejše kritične preglede področja: Ball in Hall 1967; Fleiss in Zubin 1969; Cormack 1971; Bailey 1974), več zelo odzivnih knjig, ki obravnavajo področje razvrščanja v skupine (npr. Jardine in Sibson 1971; Sneath in Sokal 1973; Anderberg 1973; Bijnen 1973; Bock 1974; Everitt 1974; Duran in Odel 1974; Ajvazjan, Bežajeva in Staroverov 1974; Hartigan 1975; Clifford in Stephenson 1975; Ivanović 1977; Spath 1977; Elisejeva in Rukavišnikov 1977; Jambu 1978; Gordon 1981; Lerman 1981; Zupan 1982; Lorr 1983; Aldenderfer in Blashfield 1984; Romesburg 1984), ustanovitev posebne revije za področje razvrščanja v skupine Journal of Classification v letu 1984 in ustanovitev Mednarodnega združenja klasifikacijskih društev v letu 1985. Mednarodno združenje prireja vsaki dve leti strokovno srečanje, ki se ga ponavadi udeleži 200 do 300 strokovnjakov s področja razvrščanja v skupine iz celega sveta.

V Jugoslaviji je med prvimi uporabljal in razvijal metode razvrščanja v skupine (kombinacijo metode I-razdalje in Sorensove metode hierarhičnega združevanja) Ivanović za razvrščanje držav glede na njihovo stopnjo družbeno-ekonomske razvitosti za leti 1967 in 1968 ter kasneje za leto 1970 (Ivanović 1971a; 1971b; 1972; 1976). V zadnjem času se metode razvrščanja v skupine razvijajo in uporabljajo vse pogosteje tudi pri nas. Zato je Zveza statističnih društev Jugoslavije leta 1986 ustanovila Sekcijo za klasifikacije, katere prvi predsednik je bil prav prof. Ivanović. Sekcija prireja vsako leto Majsko strokovno srečanje v Mostarju. V tisku je že tretji zbornik tega sedaj že tradicionalnega mostarskega srečanja, ki se ga udeleži okoli trideset strokovnjakov iz Jugoslavije. Od leta 1988 je Sekcija za klasifikacije Zveze statističnih društev Jugoslavije tudi polnopravna (sedma) članica Mednarodnega združenja klasifikacijskih društev.

Menim, da sta predvsem dva poglavitna razloga za tak razcvet področja razvrščanja v skupine v zadnjih dveh desetletjih:

- problem razvrščanja v skupine je bil pred nekaj desetletji reševan ločeno v posameznih znanstvenih disciplinah, ne da bi se tako dobljeni rezultati povezovali in dopolnjevali. To je značilno za začetne faze izgradnje določene teorije. V šestdesetih letih je zaznati prve poskuse združitve različnih pristopov reševanja problema razvrščanja v skupine in v letu 1963 prvo obsežnejše, že omenjeno delo Sokala in Sneatha. Od tedaj se področje razvrščanja v skupine razvija kot samostojna disciplina znotraj multivariatne analize;

- na razvoj teorije razvrščanja v skupine je zelo pomembno vplival razvoj računalniške tehnologije. Računalnik je omogočil uporabo računsko zahtevnejših postopkov in obdelave velikih količin podatkov. Pomembna pa so tudi teoretična

spoznanja v računalništvu, še posebej rezultati teorije zahtevnosti. Šele pred dobrim desetletjem je bilo pokazano, da je problem razvrščanja v skupine računsko zelo zahteven (NP-težek). Zato ni čudno, da se je reševal in se še vedno rešuje z različnimi hevrističnimi pristopi, bolj ali manj prilagojenimi posebnostim reševanega problema.

Avtorji z različnih znanstvenih področij različno poimenujejo področje razvrščanja v skupine. Največkrat se uporabljajo termini 'cluster analiza', taksonomija, klasifikacija pa tudi Q-analiza, tipologija, grupiranje itd. Včasih se uporablja termin klasifikacija za prirejanje enot k že določenim skupinam (npr. pri diskriminantni analizi). Nalogo *razvrščanja* v skupine razlikujemo od naloge *uvrščanja*, kjer so skupine oziroma karakteristike skupin že določene in je potrebno vsako dano enoto prirediti skupini, ki ji je najbolj podobna (najbližja). V tem delu so obravnavani postopki za reševanje nalog razvrščanja v skupine, ki jih je mogoče uporabljati tako za razvrščanje enot, določenih z izbranimi spremenljivkami, kakor tudi spremenljivk, določenih z več enotami v skupine (npr. razvrščanje prostočasnih aktivnosti v skupine, določitev tipologije časopisov glede na njihovo branost).

1.3 Nekateri razlogi za razvrščanje

Enote (ali spremenljivke) razvrščamo v skupine iz več razlogov. Najpogostejši so:

- **pregledovanje podatkov**: z metodami razvrščanja v skupine je mogoče učinkovito pregledati podatke (npr. poiskati tujke (outliers), 'otipati' strukturo v podatkih). V tej fazi analize podatkov gre bolj za postavljanje začetnih delovnih domnev o pojavih, ki jih obravnavamo;

- **zgoščanje podatkov**: namesto vseh enot analiziramo skupine enot ali predstavnike skupin, ki so bile dobljene z ustreznimi metodami razvrščanja v skupine. To pride posebno prav, kadar imamo velike količine podatkov;
- **določitev tipologije**: najpogostejši razlog za razvrščanje v skupine je empirična določitev tipologije pojavov v konkretnem področju raziskovanja in preverjanje domnev o tipologiji, ki jo raziskovalec postavi na osnovi teorije ali že opravljenih analiz podatkov.

1.4 Proces razvrščanja v skupine

Pri razvrščanju v skupine gre, kakor smo že omenili, za določanje skupin podobnih objektov. Čeprav je problem razvrščanja intuitivno zelo preprost, je analitična določitev iskane razvrstitve povezana s celo vrsto problemov. Iskane razvrstitve namreč ni mogoče poiskati z eno metodo ali pristopom z natančno določenimi pravili uporabe. Pri reševanju problema razvrščanja v skupine se je potrebno večkrat tudi intuitivno odločiti, kaj izbrati iz množice možnih izborov v določenem koraku reševanja zastavljenega problema. Osnovni koraki pri reševanju problemov razvrščanja v skupine so (npr. Anderberg 1973, 10-16; Lorr 1983, 11-21; Aldenderfer in Blashfield 1984, 9-12):

1. izbira objektov,
2. določitev množice spremenljivk, ki določajo enote,
3. računanje podobnosti med enotami,
4. uporaba ustrezne metode razvrščanja v skupine,
5. ocena dobljene rešitve.

1.4.1 Izbira objektov, spremenljivk in podobnosti

Proces razvrščanja v skupine se začne z izbiro množice (vzorca ali populacije) objektov in njihovih značilnosti, ki jih merimo na njih. Izbira spremenljivk odločilno vpliva na razvrščanje v skupine. Upoštevati je potrebno spremenljivke, ki kar najbolj ustrezajo danemu problemu. Poskrbeti je potrebno za pravo 'težo' posameznih spremenljivk, in podobno.

Predno nadaljujemo razpravo o problemih pri izbiri ustreznih spremenljivk za opis proučevanih objektov, na hitro obnovimo, katere tipe spremenljivk ponavadi ločimo glede na njihove merske lestvice (npr. Blejec 1973; Momirović 1988). V grobem ločimo atributivne spremenljivke, katerih vrednosti lahko le opišemo z besedami (npr. narodnost), in številske, katerih vrednosti so realna števila (npr. višina čistega mesečnega osebnega dohodka). Atributivne se nadalje delijo na nominalne in ordinalne, številske pa na intervalne in razmernostne spremenljivke. Dve enoti lahko le primerjamo med seboj glede na vrednosti nominalne spremenljivke (npr. spol), glede na vrednosti ordinalne spremenljivke pa lahko enote uredimo (npr. učni uspeh). Intervalna spremenljivka dopušča primerjati razlike med dvema vrednostima (npr. temperatura zraka), razmernostna pa tudi količnike (npr. velikost naselja). Poseben primer nominalne spremenljivke je dihotomna ali binarna spremenljivka, ki ima le dve vrednosti, na primer 'ima določeno lastnost' in 'nima te lastnosti'.

Denimo, da so izbrane spremenljivke, na osnovi katerih želimo razvrščati objekte v skupine, številske. Največkrat se zgodi, da so v povprečju vrednosti ene spremenljivke precej večje kot vrednosti neke druge upoštevane spremenljivke. Tiste s povprečno večjimi vrednostmi imajo največkrat večjo težo pri razvrščanju v skupine kot tiste z manjšimi. Če tega ne želimo, je potrebno pred merjenjem podobnosti med enotami spremenljivke ustrezno

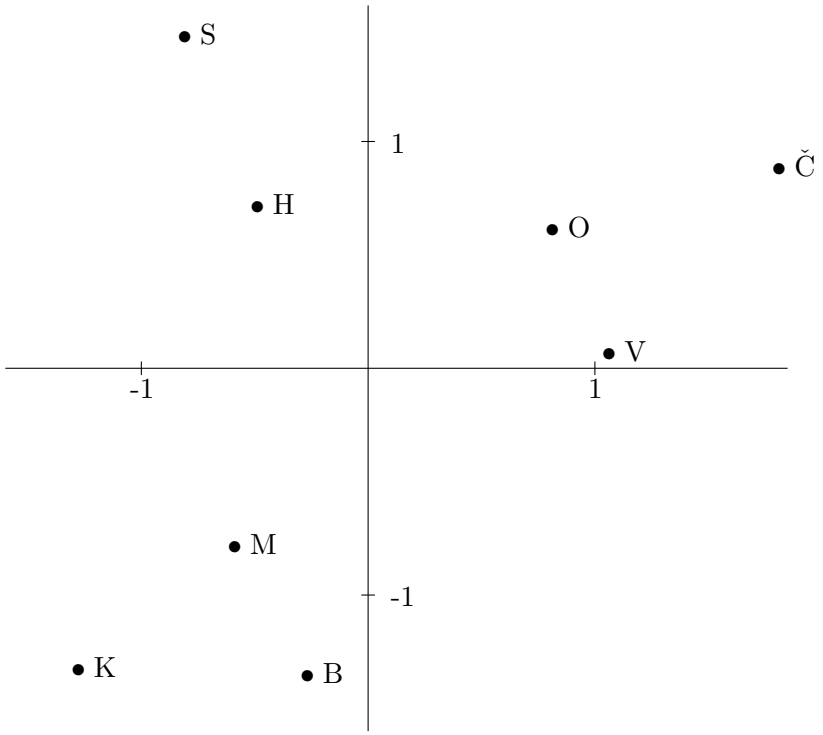
rep.,pok.	ZKJ	ZZBNOV
B i H	-0.30	-1.39
Črna gora	1.78	0.85
Hrvatska	-0.52	0.68
Makedonija	-0.62	-0.82
Slovenija	-0.84	1.43
Ožja Srbija	0.78	0.58
Kosovo	-1.31	-1.36
Vojvodina	1.03	0.03

Tabela 1.2: Standardizirani odstotki članov v ZKJ in ZZBNOV

standardizirati, tako da ima vsaka spremenljivka podobno težo pri razvrščanju v skupine. Znanih je več načinov standardizacije. Najpogosteje se uporablja običajni način standardizacije, kjer se posamezni vrednosti spremenljivke x_{ij} (vrednost j-te spremenljivke X_j za i-to enoto) odšteje njeno aritmetično sredino (μ_j) in deli s standardnim odklonom te spremenljivke (σ_j):

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

V tabeli 1.2 so tako izračunane standardizirane vrednosti za spremenljivki odstotek članov v ZKJ in ZZBNOV, ki sta podani v tabeli 1. Aritmetična sredina in standardni odklon za članstvo v ZKJ sta $\mu_{ZK} = 7.65$ in $\sigma_{ZK} = 1.89$, za članstvo v ZZBNOV pa $\mu_{ZB} = 4.81$ in $\sigma_{ZB} = 1.42$. Na sliki 1.4 je ponovno podan grafični prikaz teh enot v dvorazsežnem prostoru, kjer sta razsežnosti v tem primeru določeni s standardiziranima spremenljivkama. Primerjava prikazane razvrstitve za nestandardizirani spremenljivki (slika 1.3) z razvrstitvijo za standardizirani spremenljivki (slika 1.4) ka-



Slika 1.4: Republike in pokrajini glede na standardizirani spremenljivki

že, da standardizacija v tem primeru ne vpliva bistveno na strukturo enot.

Drugi možni načini standardizacije so na primer še, da posamezno vrednost spremenljivke delimo z njenim standardnim odklonom

$$z_{ij} = \frac{x_{ij}}{\sigma_j}$$

ali njeno maksimalno vrednostjo

$$z_{ij} = \frac{x_{ij}}{\max X_j}$$

ali aritmetično sredino

$$z_{ij} = \frac{x_{ij}}{\mu_j}$$

ali razliko med maksimalno in minimalno vrednostjo te spremenljivke

$$z_{ij} = \frac{x_{ij}}{\max X_j - \min X_j}$$

ali celo takole

$$z_{ij} = \frac{x_{ij} - \min X_j}{\max X_j - \min X_j}$$

Milligan in Cooper (1988) sta podala obsežen pregled različnih možnih standardizacij pri razvrščanju v skupine in jih primerjala.

Glede na težave, ki jih lahko imamo pri merjenju podobnosti med dvema enotama, je tudi ugodno, da so merjene spremenljivke istega tipa merskih lestvic (npr. vse številske ali vse dihotomne).

Često je število izbranih spremenljivk zelo veliko. Koristno je, da pred uporabo metod razvrščanja v skupine število spremenljivk zmanjšamo in v analizo vključimo tiste, za katere smo v predhodni analizi dognali, da imajo zadostno pojasnjevalno moč. To lahko storimo predvsem na osnovi dobrega poznavanja problema,

	1	2	...	j	...	m
1	x_{11}	x_{12}	...	x_{1j}	...	x_{1m}
2	x_{21}	x_{22}	...	x_{2j}	...	x_{2m}
	...	...	...	...	...	...
i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{im}
	...	...	...	...	...	...
n	x_{n1}	x_{n2}	...	x_{nj}	...	x_{nm}

Tabela 1.3: Primer matrike podatkov

statistično-analitično pa se za to najpogosteje uporablja metoda glavnih komponent.

Pri razvrščanju enot v skupine gre za to, da so enote v dobljenih skupinah čim bolj podobne med seboj. Odločiti se moramo torej, kako bomo merili podobnost (ali različnost) med dvema enotama. Različnosti ali podobnosti med enotami so lahko v procesu razvrščanja v skupine direktno ocenjene (npr. izbrana oseba priredi po nekem kriteriju vsakemu paru enot vrednost iz določenega intervala vrednosti, ki določajo, kako močno sta enoti posameznega para podobni med seboj), največkrat pa jih izračunamo na osnovi zbranih podatkov, ki jih ponavadi uredimo v matriko. Primer take matrike, kjer je n enot opisanih z m spremenljivkami, je podan v tabeli 1.3.

Izbira mere podobnosti je odvisna predvsem od zastavljenega problema, ki ga rešujemo, in od tipa merskih lestvic merjenih spremenljivk. Izračunane mere podobnosti s_{ij} med n enotami ponavadi uredimo v matriko podobnosti, ki je ponavadi simetrična. Primer matrike podobnosti je podan v tabeli 1.4. Merjenje podobnosti je podrobneje obravnavano v naslednjem poglavju. Večina

	1	2	...	i	...	n
1	s_{11}	s_{12}	...	s_{1i}	...	s_{1n}
2	s_{21}	s_{22}	...	s_{2i}	...	s_{2n}
	...	...	...	...	...	...
i	s_{i1}	s_{i2}	...	s_{ii}	...	s_{in}
	...	...	...	...	...	...
n	s_{n1}	s_{n2}	...	s_{ni}	...	s_{nn}

Tabela 1.4: Primer matrike podobnosti

metod razvrščanja v skupine predpostavlja, da so med obravnavanimi enotami že izračunane mere podobnosti (npr. metode hierarhičnega združevanja). Nekatere metode pa mere podobnosti (ali različnosti) med enotami računajo korakoma med samim postopkom razvrščanja (npr. metoda voditeljev).

1.4.2 Pregled metod razvrščanja v skupine

V naslednjemu koraku se moramo odločiti, katera od metod razvrščanja v skupine je najprimernejša za reševanje postavljenega problema. Najprej podajmo kratek pregled znanih metod razvrščanja v skupine. Večino metod lahko razvrstimo v tri osnovne skupine: hierahične, nehierarhične in geometrijske metode. Vse ostale, ki jih ne moremo preprosto uvrstiti v te tri skupine, stlačimo v skupino preostalih metod.

Hierarhične metode so najbrž največkrat uporabljene metode za razvrščanje v skupine. Te metode je mogoče deliti na *metode združevanja*, kjer v vsakem koraku postopka združimo dve ali več skupin v novo skupino, in *metode cepitve*, kjer na vsakem koraku izbrano skupino razcepimo na dve ali več skupin. Hierarhične

metode so zelo priljubljene predvsem zato, ker ne zahtevajo od uporabnika, da vnaprej opredeli število skupin iskane razvrstitve. Drugi razlog pa je, da je rezultat postopnega združevanja ali cepitve možno zelo nazorno grafično predstaviti na primer z drevesom združevanja (glej grafične predstavitve v četrtem poglavju). Najobsežnejši razred metod hierarhičnega združevanja v skupine predstavljajo metode, ki temeljijo na zaporednem združevanju dveh skupin v novo skupino. Te metode so podrobneje predstavljene v četrtem poglavju. Metode cepitve, ki so se sicer pokazale za manj učinkovite, lahko delimo na *monotetične*, ki v posameznem koraku postopka cepijo skupine glede na eno izbrano spremenljivko (npr. asociacijska analiza, ki sta jo razvila Lambert in Williams 1962, 1966) in *politetične*, ki ob cepljenju skupin upoštevajo vse dane spremenljivke.

Nehierarhične metode se od hierarhičnih ločijo predvsem v tem, da je potrebno vnaprej podati število skupin iskane razvrstitve. Te metode razvrščajo enote tako, da z izbranim optimizacijskim kriterijem izboljšujejo vnaprej podano začetno razvrstitev. V literaturi je predlaganih več takih kriterijev. Mogoče je najbolj znan kriterij minimizacije vsote kvadratov razdalj posamezne enote do težišča v posamezni skupini (Ward 1963). Nehierarhične metode so največkrat iteracijske: začnejo z začetno razvrstitvijo s podanim številom skupin in tako ali drugače prestavljajo enote iz ene skupine v druge skupine z namenom, da s temi prestavitvami dosežejo zmanjšanje (ali v primeru maksimizacije kriterija povečanje) vrednosti izbrane kriterijske funkcije razvrščanja. Ta proces se nadaljuje, dokler nobena prestavitev enote ne izboljša vrednosti kriterijske funkcije. Te metode v splošnem dajo le lokalno optimalne razvrstitve. Zato je priporočljivo, da razvrščanje s temi metodami ponovimo z več različnimi začetnimi razvrstitvami, po možnosti dobljenimi z različnimi metodami. Najbolj

znani in najpogoteje vključeni nehierarhični metodi v programskih paketih sta metoda prestavljanj in metoda voditeljev (znana tudi pod imenom k-means, metoda dinamičnih oblakov, itd.). Obe metodi sta podrobneje predstavljeni v petem poglavju. Ponavadi te metode razvrščajo v skupine, kjer je vsaka enota natanko v eni skupini (popolne razvrstitve). Vendar ta pogoj ni vedno potreben. Nekatere metode zmorejo poiskati tudi prekrivajoče skupine (npr. metoda voditeljev), nekatere celo 'razmazane' skupine ('fuzzy' skupine, npr. Bezdek 1981; Veledar in Kovalerchuk 1988; Bodjanova 1989).

Geometrijske metode. Če na objektih merimo le dve ali tri spremenljivke, jih lahko predstavimo v dvo- ali trirazsežnem prostoru in s tem ugotovimo njihovo strukturo. Običajno je na obravnavanih objektih merjenih več spremenljivk. Zato je tako preprosto razkrivanje strukture podatkov nemogoče. Geometrijske metode (nekateri jih imenujejo tudi ordinalne metode) omogočajo preslikavo podatkov iz originalnega več razsežnega prostora v manj razsežni, pogosto kar v dvorazsežni prostor, v katerem je lahko grafično ali kako drugače raziskati strukturo podatkov. Najbolj znani geometrijski metodi sta metoda glavnih komponent in večrazsežnostno lestvičenje (Sheppard 1962 a,b; Kruskal 1964 a,b). Med te metode razvrščanja v skupine sodijo tudi metode, ki jih že drugo desetletje razvija Momirović s sodelavci (npr. Momirović in Zakrajšek 1973; Momirović 1978, 1986). Grafične metode je koristno uporabiti pred uporabo drugih metod razvrščanja v skupine, ker lahko iz grafičnega prikaza razberemo, za kakšen tip skupin gre v konkretnem primeru.

Med **ostalimi metodami** omenimo vsaj nekatere metode razvrščanja v skupine, ki so bile v strokovni javnosti deležne pozornosti. Med te prav gotovo sodijo metode, ki temeljijo na teoriji grafov (npr. Hubert 1973; Ivanović 1977). Najpopolnejši pregled

grafovskih metod je podal Matula (1977). Zanimiva je Wishartova metoda modusov (1969), ki išče zgoščišča podatkov. V primeru neizrazite naravne strukture Wishartova metoda da le eno skupino. Lefkovitcheva metoda (1980) pa najprej množico vseh možnih razvrstitev skrči na množico 'obetajočih' razvrstitev in nato z eksaktnimi metodami poišče v tej zreducirani množici najboljšo razvrstitev. Obstaja pa še vrsta drugih zanimivih metod razvrščanja v skupine.

Za konec tega pregleda naj omenimo še pristope, ki rešujejo bolj specifične probleme razvrščanja v skupine, ki pa vendarle niso tako zelo redki. Med te sodita razvrščanje v skupine z omejitvami, ki je obravnavano v šestem poglavju, in večkriterijsko razvrščanje, ki je obravnavano v sedmem.

Kaj izbrati v tej pestri množici metod? V primeru, ko nimamo jasne domneve o številu skupin, lahko izbiramo med hierarhičnimi metodami združevanja ali cepitve. V primeru, ko poznamo število skupin, so primernejše metode nehierarhičnega razvrščanja v skupine (npr. metoda prestavljanj, metoda voditeljev). Tudi število enot je pomembno pri odločanju o ustrezni metodi. Najbolj znane metode razvrščanja v skupine, kot so hierarhične metode združevanja in metoda prestavljanj, so uporabne (tudi če imamo na voljo zelo velike računalnike) le za razvrščanje manjšega števila enot (nekaj sto). Za razvrščanje nekaj tisoč enot je primerna na primer metoda voditeljev ali nekatere druge metode, ki so razvite posebej za večje količine podatkov (npr. Zupan 1982, 1986). Pri izbiranju ustrezne metode je zelo koristno, če raziskovalec ve, kakšen tip skupin želi razkriti v svojih podatkih: ali gre za eliptične (primer (a) na sliki 1) ali verižne skupine (primer (b)), ali za med seboj ločene skupine (primer (a) ali (b)) ali za prekrivajoče (primer (c)), itd. Čim bolj raziskovalec pozna

svoj problem razvrščanja v skupine in svoje podatke, tem ustreznejšo mero podobnosti in metodo razvrščanja v skupine lahko izbere. Ne smemo namreč pozabiti, da vsaka metoda pri iskanju strukture v podatkih vsiljuje strukturo, ki je vgrajena v metodi. Nekatere metode na primer znajo razkriti le krogle, nekatere le dolge 'klobase', ne glede na to ali te v naravni strukturi podatkov so ali niso. Zato je v vsakem primeru potrebno obravnavane enote razvrščati z več različnimi metodami, primerjati dobljene razvrstitve in ob tem ugotavljati stabilnost dobljenih rešitev. Ob teh kritičnih mislih pa je potrebno takoj pribiti, da znajo vse znane metode, ki so obravnavane v tem delu, brez težav razkriti izrazito naravno strukturo z neprekrivajočimi skupinami.

1.4.3 Stabilne in objektivne razvrstitve

Cilj razvrščanja v skupine je poiskati *stabilne* in *objektivne* razvrstitve (npr. Gordon 1981, 8-9; Dunn in Everitt 1982, 2-10). Stabilne v smislu, da se dobljena razvrstitev bistveno ne spremeni (a) z dodajanjem novih objektov v proučevano množico objektov, (b) z dodajanjem nekaj novih spremenljivk med izbrane merjene spremenljvke ali (c) z vsiljenimi napakami na nekaj posameznih vrednostih merjenih spremenljivk. Objektivnost je težje opredeliti. V našem primeru je objektivnost možno opredeliti s *ponovljivostjo* rezultata: neodvisni raziskovalci naj bi prišli z analizo enake množice podatkov z enakim potekom razvrščanja v skupine do enakega (ali vsaj zelo podobnega) rezultata. Prednost objektivnega pristopa je tudi v tem, da omogoča kritiko, kajti potek razvrščanja je tedaj mogoče ponoviti, pri tem je možno tudi ugotoviti pomanjkljivosti in predlagati izboljšave. V primeru razvrščanja v skupine, katerega proces je povezan z več pomembnimi odločitvami, je težko popolnoma zadostiti kriteriju objektivnosti, še pose-

bej če procesa razvrščanja ne poznamo dovolj in ne znamo izbrati ustrezne odločitve v posameznem koraku v procesu razvrščanja v skupine. Raziskovalec mora ob svoji analizi vsekakor težiti k temu, da bo njegova rešitev čim bolj zadostila obema kriterijema.

Rezultat opisanega procesa razvrščanja v skupine je razvrstitev, optimalna glede na merjene spremenljivke in izbran kriterij razvrščanja v skupine. Največkrat nas ob dobljeni razvrstitvi tudi zanima, katere so tipične lastnosti posameznih dobljenih skupin, katere upoštevane spremenljivke najbolj ločijo skupine med seboj in podobno. Na ta vprašanja lahko preprosto odgovorimo tako, da za posamezno skupino izračunamo osnovne statistične karakteristike za vsako spremenljivko posebej (npr. aritmetično sredino, standardni odklon). Lahko pa uporabimo tudi nekatere metode multivariatne analize, kot na primer diskriminantno analizo.

Uvodno poglavje sklenimo z naslednjo Anderbergovo mislijo (1973): Le skrbna in inteligentna uporaba metod razvrščanja v skupine lahko razkrije neznano strukturo v podatkih in s tem odpre nove poglede na proučevane pojave.

2.

Merjenje podobnosti

Pri razvrščanju v skupine gre za to, da tvorimo skupine, ki jih sestavljajo karseda podobne enote glede na izbrane merjene spremenljivke. Vprašanje je, kako razpoznati, da je določena enota bolj podobna eni kot drugi enoti, oziroma, kako meriti podobnost med enotama.

Podobnost količinsko popišemo s preslikavo - *mero podobnosti*, ki vsakemu paru enot (X, Y) priredi neko realno število

$$s : (X, Y) \mapsto R$$

Za mero podobnosti zahtevamo, da je simetrična

a. $s(X, Y) = s(Y, X)$

in da zadošča ali pogoju

b1. $s(X, X) \leq s(X, Y)$

ali pogoju

b2. $s(X, X) \geq s(X, Y)$

Meri podobnosti, ki zadošča pogoju b1, pravimo prema, meri, ki zadošča pogoju b2, pa obratna mera podobnosti.

Mera podobnosti s določa v množici neurejenih parov enot urejenost in glede na to urejenost je mogoče definirati tudi pojem enakovrednosti mer podobnosti. Meri podobnosti sta *enakovredni*, če je urejenost parov enot, dobljena s prvo mero, enaka urejenosti parov enot z drugo mero podobnosti.

Pri premi meri podobnosti je običajno izpolnjen naslednji pogoj

c. $s(X, X) = s^*$

Kadar je izpolnjen ta pogoj, dobimo s predpisom

$$d(X, Y) = s(X, Y) - s^*$$

enakovredno *mero različnosti* d, ki sicer zadošča naslednjim pogojem:

1.	$d(X, Y) \geq 0$	nenegativnost
2.	$d(X, X) = 0$	
3.	$d(X, Y) = d(Y, X)$	simetričnost

Mera različnosti lahko zadošča še nekaterim pogojem. Če zadošča še pogojema

4.	$d(X, Y) = 0 \Longrightarrow X = Y$	razločljivost
5.	$\forall Z : d(X, Y) \leq d(X, Z) + d(Z, Y)$	trikotniška neenakost

ji pravimo *razdalja*.

O merah podobnosti in različnosti, še posebej o enakovrednosti mer, je znanih še veliko zanimivih rezultatov. Bralec, ki ga formalnejši zapisi ne prestrašijo, si lahko nekaj tega najde v delih Batagelja (1985 b, 1988 a).

Poznanih je ogromno bolj ali manj posrečenih mer podobnosti in mer različnosti. Največkrat so obravnavane mere za enote, ki so določene s takimi ali drugačnimi spremenljivkami in ponavadi predstavljene v obliki matrike podatkov, ki je podana v tabeli 1.3. Podrobni pregledi takih mer so podani na primer v delih Clifforda in Stephensona (1975, 49-82), Everitta (1974, 49-59), Gordona (1981, 13-32) in Lorra (1983, 22-44). V literaturi pa je mogoče najti tudi mere, ki so primerne za specifičnejše opise objektov. Tako je Košmeljeva (1986, 1987) predlagala več mer različnosti, primernih za razvrščanje enot, kjer je vsaka izbrana spremenljivka podana s časovno vrsto (gre torej za trirazsežno matriko podatkov). Objekte se vedno ne da preprosto popisati s spremenljivkami. Batagelj (1988) je predlagal mere, ki so primerne za merjenje podobnosti kompleksnejših struktur objektov (npr. nizov, molekul, grafov).

Iz bogate zbirke mer podobnosti, različnosti in razdalj smo izbrali le nekaj najpogosteje omenjenih, ki jih predstavljamo v naslednjih razdelkih. Razvrščene so po tipih spremenljivk, ki določajo enote, kajti prav od tipa spremenljivk je ponavadi odvisno, katero mero podobnosti ali različnosti je primerno izbrati za potrebe razvrščanja v skupine.

2.1 Mere podobnosti za številske podatke

Pri razvrščanju enot, določenih s samimi številskimi spremenljivkami, je najpogosteje uporabljena evklidska razdalja. Za enoti X in Y, opisanimi z m številskimi spremenljivkami

$$X = (x_1, x_2, ..., x_m)$$

$$Y = (y_1, y_2, ..., y_m)$$

je evklidska razdalja med njima definirana takole

$$d(X,Y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2}$$

Pogosto je uporabljena tudi razdalja Manhattan

$$d(X,Y) = \sum_{i=1}^{m}|x_i - y_i|$$

Obe razdalji sta posebna primera razdalje Minkowskega

$$d(X,Y) = (\sum_{i=1}^{m}|x_i - y_i|^r)^{\frac{1}{r}} \quad , \quad r > 0$$

in sicer , če je $r = 1$, gre za razdaljo Manhattan, če je $r = 2$, pa za evklidsko razdaljo. Pri odločanju, katero razdaljo uporabiti v določenem primeru razvrščanja v skupine, je koristno upoštevati naslednjo lastnost razdalje Minkowskega: pri večjih vrednostih r-ja imajo večjo težo pri merjenju razdalje med enotama večje razlike $|x_i - y_i|$. V limiti, to je pri $r = \infty$, je Minkowskijeva razdalja

$$d(X,Y) = \max_i |x_i - y_i|$$

Imenuje se razdalja Čebiševa ali trdnjavska razdalja.

Kot primer izračunajmo evklidske razdalje med republikami in pokrajinama glede na odstotek članov ZKJ in ZZBNOV v celotnem prebivalstvu, pri čemer naj bosta obe spremenljivki standardizirani. Razdalje torej računamo iz podatkov, ki so podani v tabeli 1.2. Evklidsko razdaljo med Bosno in Hercegovino ter Črno goro izračunamo takole

$$d(B,C) = \sqrt{(b_1 - c_1)^2 + (b_2 - c_2)^2} =$$

	B	C	H	M	S	O	K	V
B	0.0	3.1	2.1	0.7	2.9	2.3	1.0	2.0
C		0.0	2.3	3.0	2.7	1.0	3.8	1.1
H			0.0	1.5	0.8	1.3	2.2	1.7
M				0.0	2.3	2.0	0.8	1.9
S					0.0	1.8	2.8	2.3
O						0.0	2.9	0.6
K							0.0	2.7
V								0.0

Tabela 2.1: Evklidske razdalje med republikami in pokrajinama

$$= \sqrt{(-0.30 - 1.78)^2 + (-1.39 - 0.85)^2} = 3.06$$

Tako izračunane evklidske razdalje med vsemi osmimi enotami so zaokrožene na eno decimalno mesto in urejene v matriko razdalj, ki je podana v tabeli 2.1 (simetrične vrednosti so izpuščene).

Poznane so tudi druge razdalje, ki niso posebni primer Minkowskijeve razdalje. Med njimi je najpomembnejša Mahalanobisova posplošena razdalja (1936), ki je definirana takole

$$d(X, Y) = (X - Y)'\Sigma^{-1}(X - Y)$$

Σ je varianačno-kovariančna matrika spremenljivk znotraj skupin. Za razliko od drugih omenjenih razdalj upošteva tudi povezanosti med spremenljivkami. Če so korelacijski koeficienti med spremenljivkami enaki 0, je Mahalanobisova razdalja enaka kvadratu evklidske razdalje.

Za enote, ki imajo samo pozitivne vrednosti spremenljivk, sta poznani še dve meri različnosti: Lance-Williamsova mera različno-

sti (1966)

$$d(X,Y) = \frac{\sum_{i=1}^{m} |x_i - y_i|}{\sum_{i=1}^{m} (x_i + y_i)}$$

in razdalja Canberra (Lance in Williams 1967 a)

$$d(X,Y) = \sum_{i=1}^{m} \frac{|x_i - y_i|}{|x_i + y_i|}$$

ki pa sta zelo občutljivi na majhne spremembe okoli vrednosti blizu 0.

Če so enote opisane s številskimi spremenljivkami, lahko uporabimo tudi zelo znano mero podobnosti - Pearsonov korelacijski koeficient (1926), ki je definiran takole

$$r(X,Y) = \frac{\sum_{i=1}^{m} (x_i - \mu_X)(y_i - \mu_Y)}{\sqrt{\sum_{i=1}^{m} (x_i - \mu_X)^2 \sum_{i=1}^{m} (y_i - \mu_Y)^2}}$$

kjer sta

$$\mu_X = \frac{1}{m} \sum_{i=1}^{m} x_i$$

in

$$\mu_Y = \frac{1}{m} \sum_{i=1}^{m} y_i$$

Lastnost koeficienta korelacije je, da ostane enak, če eno ali drugo enoto linearno transformiramo.

Nesporno je koeficient korelacije primerna mera podobnosti med spremenljivkami. Različna mnenja pa so o smislu njegove uporabe za merjenje podobnosti med enotami prav zaradi omenjene lastnosti tega koeficienta. Denimo, da so vrednosti neke enote izračunane tako, da je vsem vrednostim neke druge enote prišteto neko večje število. 'Profil' teh dveh enot je torej za omenjeno število premaknjen, vendar enak. Enoti sta torej glede na

vrednosti posameznih spremenljivk zelo različni in evklidska razdalja bi to pokazala. Zaradi vzporednosti obeh 'profilov' (ena enota je linearna kombinacija druge enote) pa bi bil izračunani koeficient korelacije 1! Torej popolna podobnost. Ko izbiramo mero podobnosti ali različnosti moramo predvsem vedeti, kakšno podobnost želimo meriti. Če želimo na primer meriti podobnost med 'profiloma' obeh enot, je koeficient korelacije prav gotovo primerna mera.

Znanih je še veliko drugih mer podobnosti, različnosti in razdalj za številske enote. Nekatere so posebej primerne za reševanje specifičnejših problemov razvrščanja v skupine. Tako je za razvrščanje območij (npr. držav ali v našem primeru republik in pokrajin) glede na njihovo družbeno ekonomsko razvitost Ivanović (1963; 1976; 1977; 1982; 1988) razvil družino I-razdalj, ki zmorejo odstraniti prekrivanja med izbranimi indikatorji razvitosti.

2.2 Mere podobnosti za binarne podatke

Za enote, ki so določene s samimi dihotomnimi spremenljivkami, je poznanih več mer podobnosti. Te so določne s frekvencami v asociacijski tabeli za par enot, med katerima merimo podobnost. Asociacijsko tabelo za enoti X in Y, kjer so vrednosti vseh m spremenljivk $+$ in $-$, je naslednja

		enota Y	
		$+$	$-$
enota X	$+$	a	b
	$-$	c	d

Vsota vseh štirih frekvenc je enaka številu vseh merjenih spremenljivk ($a + b + c + d = m$). Frekvenca a pove, na koliko

spremenljivkah imata enoti X in Y hkrati pozitiven odgovor in frekvenca d hkrati negativen odgovor. Frekvenci b in c pa štejeta, na koliko spremenljivkah imata enoti različna odgovora.

Najbrž so za razvrščanje binarnih enot v skupine najprimernejše mere ujemanja. Naštejmo tiste, ki so najbolj znane:

- Sokal-Michenerjeva mera (1958)
  (enake uteži na $++$ in na $--$ ujemanju)
  $$\frac{a+d}{a+b+c+d}$$
- Prva Sokal-Sneathova mera (1963)
  (dvojna utež na $++$ in $--$ ujemanju)
  $$\frac{2(a+d)}{2(a+d)+b+c}$$
- Rogers-Tanimotova mera (1960)
  (dvojna utež na neujemanju)
  $$\frac{a+d}{a+d+2(b+c)}$$
- Russell-Raova mera (1940)
  (le $++$ ujemanje v števcu)
  $$\frac{a}{a+b+c+d}$$
- Jaccardova mera (1908)
  (v števcu in imenovalcu ne upošteva ujemanje na $--$)
  $$\frac{a}{a+b+c}$$

- Czekanowskijeva mera (1913)
  (v števcu in v imenovalcu ni $--$ ujemanja, dvojna utež na $++$ ujemanju)
  $$\frac{2a}{2a+b+c}$$

- Druga Sokal-Sneathova mera (1963)
  (v števcu in v imenovalcu ni $--$ ujemanja, dvojna utež na neujemanju)
  $$\frac{a}{a+2(b+c)}$$

- Kulczynskijeva mera (1927)
  (kvocient med ujemanjem in neujemanjem, kjer $--$ ujemanje ni upoštevano)
  $$\frac{a}{b+c}$$

Vse omenjene mere podobnosti razen zadnje lahko zavzamejo vrednosti v intervalu od 0 do 1. Prve tri omenjene mere ujemanja so glede na definicijo na začetku tega poglavja enakovredne. Prav tako so enakovredne peta, šesta in sedma omenjena mera. Z drugimi besedami to pomeni, da je urejenost vseh parov enot, dobljena z eno od teh treh mer, enaka urejenosti parov enot s preostalima dvema merama.

Pojem enakovrednosti je zelo pomemben pri razvrščanju v skupine. Nekatere metode razvrščanja v skupine namreč proizvedejo enake razvrstitve, če merimo podobnost med enotami s sicer različnimi, vendar enakovrednimi merami (npr. minimalna in maksimalna metoda hierarhičnega združevanja, nemetrično večrazsežnostno lestvičenje).

Za primer vzemimo tri osebe X, Y in Z, ki smo jih povprašali, s katerimi od desetih naštetih aktivnosti se ukvarjajo v prostem

času. Vsaka od teh desetih spremenljivk lahko zavzame le dve vrednosti: se ukvarjam (+) in se ne ukvarjam (−). Torej gre za tri binarne enote. Denimo, da so zbrani podatki naslednji:

	1	2	3	4	5	6	7	8	9	10
X	+	+	−	−	+	+	−	−	−	−
Y	+	+	+	−	+	−	−	−	+	+
Z	+	−	−	+	−	+	−	−	−	−

Asociacijska tabela za enoti X in Y je tedaj

		enota Y	
		+	−
enota X	+	3	1
	−	3	3

Izračunajmo za ti dve enoti Sokal-Michenerjevo in Jaccardovo mero ujemanja:

$$S(X,Y) = \frac{3+3}{3+1+3+3} = 0.60$$

$$J(X,Y) = \frac{3}{3+1+3} = 0.43$$

Ti dve meri ujemanja sta izračunani še za preostala dva para. Rezultati, dobljeni s Sokal-Michenerjevo mero ujemanja, so naslednji:

	X	Y	Z
X	1.00	0.60	0.70
Y		1.00	0.30
Z			1.00

Pare uredimo od para z najmanjšo vrednostjo mere do para z največjo:

$$(Y, Z) < (X, Y) < (X, Z)$$

Jaccardova mera ujemanja pa da naslednje rezulate:

	X	Y	Z
X	1.00	0.43	0.40
Y		1.00	0.13
Z			1.00

Tudi v tem primeru uredimo pare

$$(Y, Z) < (X, Z) < (X, Y)$$

Dobljeni urejenosti parov se razlikujeta. Ti dve meri torej nista enakovredni. Tudi sicer sta ti dve meri zelo različni. Prva meri ujemanje na $++$ in $--$ odgovorih, druga pa le na $++$. V našem primeru, kjer se očitno kažeta bistveno različna rezultata za para (X, Y) in (X, Z), lahko na osnovi podatkov vidimo, da sta enoti X in Z relativno močno povezani med seboj zaradi relativno močnega ujemanja na $-$ vrednostih, kar v našem primeru pomeni na neukvarjanju z naštetimi prostočasnimi aktivnostmi. Po Jaccardovi meri, v kateri je vgrajeno le ujemanje na $+$ odgovorih, se pravi le na ujemanju v smislu ukvarjanja z naštetimi aktivnostmi, pa se bolje odreže par (X, Y). Ta primer zopet kaže, kako pomembna je opredelitev, kaj pravzaprav želimo meriti.

Poznamo pa tudi druge mere podobnosti, ki so določene z omenjemi štirimi frekvencami asociacijske tabele. Omenimo dve verjetnostni meri podobnosti

- Kulczynskijeva povprečna pogojna verjetnost (1927) (++ ujemanje)
$$\frac{a}{2}(\frac{1}{a+b}+\frac{1}{a+c})$$

- Povprečna pogojna verjetnost ujemanja
$$\frac{1}{4}(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{b+d}+\frac{d}{c+d})$$

in še dve meri podobnosti, ki zavzemata vrednosti od -1 do 1 in ki se najpogosteje uporabljata kot meri podobnosti pri razvrščanju dihotomnih spremenljivk v skupine.

- Yulova mera podobnosti
$$\frac{ad-bc}{ad+bc}$$

- Pearsonov koeficient Φ
$$\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

Pokazati se da, da je Pearsonov koeficient korelacije, če ga računamo na binarnih podatkih, enak koeficientu Φ.

2.3 Mere podobnosti za nominalne podatke

Če so enote opisane z nominalnimi spremenljivkami, lahko uporabimo nekatere mere podobnosti, ki smo jih omenili za binarne podatke, če vsako spremenljivko *dihotomiziramo* (vse vrednosti

posamezne spremenljivke smiselno združimo v dve vrednosti) ali pa *'dummyziramo'* (vsaka vrednost nominalne spremenljivke je nova dihotomna spremenljivka, ki ima le dve vrednosti: prisotnost (+) ali odsotnost (−) določene vrednosti nominalne spremenljivke). V slednjem primeru seveda ni primerna Sokal-Michenerjeva mera ujemanja in njej enakovredne, ker je ujemanje −− predvsem posledica 'dummyzacije'.

V posebnih primerih, ko imajo vse nominalne spremenljivke, ki določajo enote, enake vrednosti, je mogoče izbirati med različnimi merami ujemanja. Za primer vzemimo, da želimo poiskati tipologijo držav glede na njihovo glasovanje v OZN o različnih resolucijah v nekem časovnem razdobju. Enote so v tem primeru države, ki so določene z nominalnimi spremenljivkami, to je posameznimi resolucijami. Tako opredeljene spremenljivke imajo naslednje vrednosti: 1 - glas za, 2 - glas proti, 3 - vzdržan, 4 - ni prisoten. Mere ujemanja v primeru nominalnih enot so podobno kot mere podobnosti za binarne enote opredeljene s frekvencami v kontingenčnih tabelah. Velikost kontingenčnih tabel določa število vrednosti nominalne spremenljivke (npr. v omenjenem primeru s štirimi vrednostmi gre za kontingenčne tabele 4x4). V splošnem lahko zapišemo kontingenčno tabelo za enoti X in Y, kjer sta enoti določeni z m nominalnimi spremenljivkami s p različnimi vrednostmi, takole

	1	2	...	p
1	f_{11}	f_{12}	...	f_{1p}
2	f_{21}	f_{22}	...	f_{2p}
...	...	...	...	...
p	f_{p1}	f_{p2}	...	f_{pp}

Vsota vseh frekvenc v zgornji tabeli je seveda enaka številu spre-

menljivk (m). Najpreprostejša mera ujemanja med enotama X in Y je lahko tedaj kvocient med vsoto frekvenc na diagonali kontingenčne tabele za ti dve enoti in številom vseh spremenljivk, ki ti dve enoti opisujejo

$$s(X,Y) = \frac{f_{11} + f_{22} + ... + f_{pp}}{m}$$

V literaturi je mogoče najti še veliko drugih zanimivih mer podobnosti, primernih za nominalne enote (npr. Liebetrau 1983; Reynolds 1984; Momirović 1988).

2.4 Mere podobnosti za mešani tip podatkov

Vse omenjene koeficiente računamo tedaj, ko enote določa le en tip spremenljivk. Za merjenje podobnosti med enotami z različnimi tipi spremenljivk so možni vsaj trije pristopi:

- vse spremenljivke transformiramo v isti tip spremenljivk (npr. v dihotomne spremenljivke). Anderberg (1973) je takim transformacijam posvetil v svoji knjigi celo poglavje (1973, 30-69);
- spremenljivke razvrstimo v skupine z istim tipom spremenljivk in nato enote razvrščamo za vsako skupino spremenljivk posebej. Rešitev problema poizkusimo poiskati s primerjavo dobljenih razvrstitev. V tem primeru je ugodno uporabiti metode večkriterijskega razvrščanja v skupine, ki so predstavljene v sedmem poglavju tega dela;
- uporabimo lahko tudi sestavljene mere, ki so v glavnem kombinacije mer za enote z istim tipom spremenljivk (npr. Es-

tabrook in Rogers 1966; Gower 1971; Legendre in Chodorowski 1977). Najpogosteje je omenjen Gowerjev koeficient podobnosti (1971), ki je definiran takole

$$s_{ij} = \frac{\sum_{k=1}^{m} w_{ijk} s_{ijk}}{\sum_{k=1}^{m} w_{ijk}}$$

kjer je s_{ijk} mera podobnosti med i-to in j-to enoto glede na k-to spremenljivko in ustrezno opredeljena glede na njen merski tip. Utež w_{ijk} zavzame vrednost 1 ali 0 glede na to, ali k-ta spremenljivka dopušča primerljivost med enotama ali ne (npr. $w_{ijk} = 0$, če vrednost k-te spremenljivke na eni ali drugi enoti ni poznana). Pokazati se da, da je v primeru, ko so enote določene s samimi dihotomnimi spremenljivkami, Gowerjeva mera enaka Jaccardovi meri ujemanja.

2.5 Zveze med merami različnosti in podobnosti

Praviloma metode razvrščanja v skupine predpostavljajo, da so med enotami izračunane mere različnosti. Problemu razvrščanja v skupine in tipu merjenih spremenljivk pa je včasih primerno izbrati neko mero različnosti ali razdaljo, včasih pa neko mero podobnosti. Poglejmo, kako je mogoče transformirati mero podobnosti s v mero različnosti d in obratno.

Vedno lahko konstruiramo mere podobnosti iz razdalj. Na primer transformacija

$$s = \frac{1}{1+d}$$

priredi razdalji d mero podobnosti s, ki je definirana na območju $[0, 1]$.

Transformacija iz mer podobnosti v mere različnosti je lahko težavna, če v dani množici enot mera podobnosti posamezne enote s seboj s_{ii} ni za vse enote enaka. Če te mere izvzamemo iz nadaljnjega razmišljanja, lahko izbiramo med več transformacijami. Izbira ustrezne transformacije je odvisna tudi od posameznih lastnosti mere, ki jo želimo transformirati, na primer od njenega definicijskega območja. Mere podobnosti med enotami ponavadi zavzemajo vrednosti med 0 in 1: čim bolj sta si enoti podobni, tem bolj se mera podobnosti približuje 1; čim bolj sta si različni, tem bolj se mera podobnosti bliža 0 (npr. mere ujemanja za binarne enote). V primeru torej, ko je mera podobnosti s definirana na območju $[0, 1]$, je ustrezna transformacija v mero različnosti d naslednja

$$d = 1 - s$$

Gower in Legendre (1986) sta obravnavala metrične lastnosti mer različnosti, ki jih dobimo s to transformacijo. Pokazala sta, da dobimo razdalje s transformacijo večine omenjenih mer ujemanja za binarne podatke. Izjema so prva Sokal-Sneathova, Czekanowskijeva in Kulczynskijeva mera ujemanja. Iz teh in tudi preostalih omenjenih mer podobnosti za binarne podatke dobimo le mere različnosti.

Našteli smo več mer podobnosti, ki so definirane na območju $[-1, 1]$ (npr. Yulova mera podobnosti, Pearsonov koeficient korelacije). V tem primeru lahko uporabimo na primer naslednji transformaciji

$$d = \frac{1 - s}{2} \qquad \text{ali} \qquad d = 1 - |s|$$

To sta zelo različni transformaciji. Posledica prve je, da bo mera različnosti najmanjša (0) pri meri podobnosti $+1$ in največja pri meri podobnosti -1. Po drugi transformaciji pa je najmanjša

mera različnosti, če je mera podobnosti -1 ali $+1$, največja pa pri meri podobnosti 0. Pri tej transformaciji torej dobljena mera različnosti ne loči več, v katero smer sta enoti povezani, pozitivno ali negativno.

Zelo zanimiva je naslednja transformacija

$$d = \sqrt{1 - s}$$

Gower (1971, 1985), Gower in Legendre (1986) ter delno tudi drugi avtorji so pokazali, da so $d_{ij} = \sqrt{1 - s_{ij}}$ evklidske razdalje, če je matrika podobnosti $[s_{ij}]$ pozitivno semi definitna z elementi $0 \leq s_{ij} \leq 1$ in $s_{ii} = 1$. Pokazala sta, da temu izreku zadošča večina omenjenih mer ujemanja. Ponovno sta izjemi prva Sokal-Sneathova in Kulczynskijeva mera. S to transformacijo dobimo evklidsko razdaljo tudi v primeru koeficienta Φ. V merskem smislu se torej od omenjenih mer podobnosti za binarne podatke pri teh transformacijah najslabše odrežejo prva Sokal-Sneathova in Kulczynskijeva mera ujemanja, obe verjetnostni meri in Yulova mera podobnosti.

Že Cronbach in Gleser (1953) sta pokazala, da je možno z zgornjo transformacijo dobiti evklidsko razdaljo med standardiziranima enotama (vsaki vrednosti enote je odšteta aritmetična sredina vseh njenih vrednosti in ta razlika deljena s standardnim odklonom vrednosti), če je med enotama izračunan Pearsonov korelacijski koeficient.

Katero transformacijo izbrati v konkretnem primeru? Predvsem je potrebno izhajati iz problema, ki ga rešujemo, in natančno premisliti, katera mera podobnosti in katera transformacija (če je sploh potrebna) izbrane mere sta najbolj ustrezni. Nekaterim metodam (npr. metodam, ki uporabljajo geometrijski pristop, kjer so enote predstavljene kot točke v prostoru) bolj ustrezajo razdalje med enotami, nekatere so v merskem smislu manj zahtevne. Zato

pri izbiri ustrezne mere podobnosti in transformacije ene mere v drugo upoštevajmo tudi njihove merske lastnosti. Bralcu, ki želi več informacij o merskih lastnostih mer podobnosti, priporočam, da si prebere Gowerjev in Legendrov članek (1986).

3.

Matematizacija problema razvrščanja v skupine

3.1 Osnovni pojmi

Uvodna razprava o problemu razvrščanja v skupine morda daje vtis, da gre pri razvrščanju v skupine za zbirko razmeroma nepovezanih postopkov in metod. Zato v tem poglavju povežimo in poenotimo večji del področja razvrščanja v skupine z optimizacijskim pristopom, ki je rezultat večletnega Batageljevega in mojega raziskovanja (npr. Batagelj 1979, 1985 a, 1986 b; Ferligoj in Batagelj 1980, 1982, 1983).

Opredelimo najprej nekaj pojmov, ki jih potrebujemo pri matematizaciji problema razvrščanja v skupine.

V uvodu smo že opredelili enote, ki glede na obravnavano vsebino ustrezno opisujejo objekte, ki jih želimo razvrstiti v skupine. Do opisa objektov ponavadi pridemo tako, da na vsaki enoti X_i $(i = 1, 2, ..., n)$ izmerimo nekaj lastnosti (spremenljivk). *Množico enot* označimo z $E = \{X_i\}$. *Skupina* enot je neprazna podmnožica

množice enot, ki jo označimo s $C \subseteq E$. *Razvrstitev* pa je množica skupin enot $\mathcal{C} = \{C_i\}$.

Kot primer vzemimo občine SR Slovenije pred cepitvijo občine Maribor na pet občin:

$$E = \{ \text{ 60 občin SR Slovenije } \}$$

Dve skupini sta na primer:

$$\begin{aligned} C_i &= \{ \text{ obalne občine } \} \\ C_j &= \{ \text{ ljubljanske občine } \} \end{aligned}$$

V primeru občin SR Slovenije zapišimo razvrstitev, ki je določena s plansko regionalizacijo občin v 12 regij (skupin občin):

$$\begin{aligned} \mathcal{C} = \{ &\text{ pomurska r., podravska r., koroška r., savinjska r.,} \\ &\text{zasavska r., osrednja slovenska r., spodnja posavska r.,} \\ &\text{dolenjska r., goriška r., obalno-kraška r., kraška r.,} \\ &\text{gorenjska r. } \} \end{aligned}$$

kjer so posamezne regije sestavljene iz naslednjih občin:

$$\begin{aligned} C_1 &= \textbf{pomurska r.} \\ &= \{ \text{ G. Radgona, Lendava, Ljutomer, Murska Sobota } \} \\ C_2 &= \textbf{podravska r.} \\ &= \{ \text{ Lenart, Maribor, Ormož, Ptuj, Sl. Bistrica } \} \\ C_3 &= \textbf{koroška r.} \\ &= \{ \text{ Dravograd, Sl. Gradec, Ravne, Radlje } \} \\ C_4 &= \textbf{savinjska r.} \\ &= \{ \text{ Celje, Laško, Mozirje, Sl. Konjice, Šentjur,} \\ &\quad \text{Šmarje, Velenje, Žalec } \} \\ C_5 &= \textbf{zasavska r.} \\ &= \{ \text{ Hrastnik, Trbovlje, Zagorje } \} \\ C_6 &= \textbf{osrednja sl. r.} \end{aligned}$$

$=$ { Domžale, Grosuplje, Kamnik, Kočevje, Litija,
Lj. Bežigrad, Lj. Center, Lj. Moste-Polje, Lj. Šiška,
Lj. Vič-Rudnik, Logatec, Ribnica, Vrhnika }

$C_7 =$ **spodnje posavska r.**

$=$ { Brežice, Krško, Sevnica }

$C_8 =$ **dolenjska r.**

$=$ { Črnomelj, Metlika, Novo mesto, Trebnje }

$C_9 =$ **goriška r.**

$=$ { Ajdovščina, Idrija, Nova Gorica, Tolmin }

$C_{10} =$ **obalno-kraška r.**

$=$ { Izola, Koper, Piran, Sežana }

$C_{11} =$ **kraška r.**

$=$ { Cerknica, Ilirska Bistrica, Postojna }

$C_{12} =$ **gorenjska r.**

$=$ { Jesenice, Kranj, Radovljica, Škofja Loka, Tržič }

Razvrstitev je *popolna*, če je vsaka enota natanko v eni skupini. Tako je razvrstitev občin v regije popolna razvrstitev z dvanajstimi skupinami. V splošnem razvrstitev ni nujno popolna. Včasih iščemo razvrstitve, ki jih lahko sestavljajo prekrivajoče skupine. V nekaterih primerih dopustimo, da nekatere enote niso razvrščene v nobeno skupino. Razvrstitev pa lahko vnaša tudi določeno strukturo med skupine. Primer take strukture je drevesna ali hierarhična urejenost med skupinami, ki bo podobneje obravnavana v naslednjem poglavju.

Na osnovi vsebine problema razvrščanja v skupine opredelimo, kakšne razvrstitve so smiselne oziroma dopustne (npr. ali gre za popolno ali drevesno razvrstitev, v primeru popolne razvrstitve je potrebno določiti število skupin, itd.). To lahko storimo tako, da kar se da natančno opredelimo množico dopustnih razvrstitev Φ. Med temi razvrstitvami iščemo tiste, ki kar najbolje ustrezajo našim namenom. Ustreznost razvrstitve ponavadi izrazimo s *kri-*

terijsko funkcijo P, ki vsaki razvrstitvi $\mathcal{C}$ iz množice dopustnih razvrstitev Φ priredi neko nenegativno realno število

$$P : \mathcal{C} \mapsto R_0^+$$

3.2 Problem razvrščanja v skupine kot optimizacijski problem

Z vpeljanimi pojmi lahko zastavimo problem razvrščanja v skupine kot optimizacijski problem takole:

> Določi razvrstitev $\mathcal{C}^*$ tako, da bo
>
> $$P(\mathcal{C}^*) = \min_{\mathcal{C} \in \Phi} P(\mathcal{C})$$
>
> kjer je Φ množica (dopustnih) razvrstitev.

To pomeni: če imamo množico razvrstitev Φ in izračunamo za vsako razvrstitev $\mathcal{C} \in \Phi$ vrednost kriterijske funkcije, je najboljša (najprimernejša) razvrstitev ($\mathcal{C}^*$) tista, ki ima najmanjšo vrednost kriterijske funkcije.

3.3 Kriterijske funkcije

S kriterijsko funkcijo torej opišemo, kakšna naj bo želena razvrstitev enot, ki so opisane z izbranimi spremenljivkami. Vanjo torej skušamo glede na zastavljeni problem vgraditi, kakšne naj bodo želene skupine. V tem smislu se spomnimo pojmov interne kohezivnosti in eksterne izolacije skupin. Največkrat nas zanima predvsem interna kohezivnost skupin (homogenost, kompaktnost) in večina znanih kriterijskih funkcij meri predvsem to.

Kriterijska funkcija za posamezno skupino je običajno določena s pomočjo merjenih različnosti med enotami na primer takole:

$$p(C) = \sum_{X,Y \in C} d(X,Y)$$

ali

$$p(C) = \max_{X,Y \in C} d(X,Y)$$

Kriterijska funkcija razvrstitve pa te dele po skupinah največkrat aditivno povzame na primer takole:

$$P(\mathcal{C}) = \sum_{C \in \mathcal{C}} p(C)$$

ali

$$P(\mathcal{C}) = \max_{C \in \mathcal{C}} p(C)$$

Te kriterijske funkcije torej merijo homogenost znotraj skupin, ne pa ločljivosti med skupinami. Vzemimo dve razvrstitvi z enako interno strukturo skupin, vendar naj bodo te skupine prvič bolj skupaj, a še vedno dobro ločene med seboj, drugič bolj narazen. Vrednost katerekoli zgoraj omenjene kriterijske funkcije bo za obe razvrstitvi enaka.

Znanih je še veliko drugačnih kriterijskih funkcij, ki večinoma merijo prav tako le homogenost znotraj skupin. Tako je cela družina kriterijskih funkcij izpeljana iz predpostavke, da je množica enot vzorec, vzet iz mešanice večrazsežnih normalnih porazdelitev (npr. Day 1969; Wolfe 1970). Te kriterijske funkcije ponavadi izhajajo iz matrike celotne razpršenosti T, katere elementi so določeni takole

$$t_{kl} = \sum_{i=1}^{n} (x_{ik} - \bar{X}_k)(x_{il} - \bar{X}_l) \quad , \quad k,l = 1,2,...,m$$

kjer sta $\bar{X}_k$ in $\bar{X}_l$ aritmetični sredini k-te in l-te spremenljivke. Podobno lahko definiramo tudi matriko razpršenosti znotraj skupin

$$W = \sum_{i=1}^{k} W_i$$

kjer je k število skupin in W_i matrika razpršenosti znotraj i-te skupine, to je matrika vsot kvadratov in produktov odklonov od aritmetičnih sredin na enotah i-te skupine. Matrika razpršenosti med skupinami B pa je določena z vsotami kvadratov in produktov odklonov aritmetičnih sredin skupin od aritmetične sredine vseh enot. Te tri matrike razpršenosti lahko povežemo z naslednjo enakostjo (npr. Dillon in Goldstein 1984)

$$T = W + B$$

Kriterijske funkcije so tedaj določene s temi tremi matrikami na primer takole (npr. Friedman in Rubin 1967):

- *Sled matrike* W: $P(\mathcal{C}) = sl(W)$
  Problem minimizacije te kriterijske funkcije je ekvivalenten problemu maksimizacije kriterijske funkcije $P(\mathcal{C}) = sl(B)$.

- *Determinanta matrike* W: $P(\mathcal{C}) = |W|$
  Problem minimizacije te funkcije je ekvivalenten maksimizaciji
  $$P(\mathcal{C}) = |T|/|W| = \prod_j (1 + \lambda_j)$$
  kjer so λ_j lastne vrednosti, dobljene iz enačbe $|B - \lambda W| = 0$.

- *Sled matrike* BW^{-1}
  V tem primeru maksimiziramo kriterijsko funkcijo
  $$P(\mathcal{C}) = sl(BW^{-1}) = \sum_i \lambda_i$$

Izčrpen pregled teh kriterijskih funkcij je podal Marriott (1982).

3.4 Primer

Kot primer vzemimo množico petih enot $E = \{A, B, C, D, E\}$, ki jih določata dve spremenljivki (X in Y):

	A	B	C	D	E
X	1	2	3	5	5
Y	1	3	2	3	5

Enote lahko predstavimo tudi grafično (glej sliko 3.1).

Razvrstimo dane enote v dve skupini (popolna razvrstitev), kjer je kriterijska funkcija določena takole

$$P(\mathcal{C}) = \sum_{C \in \mathcal{C}} \sum_{X \in C} d(X, T_C)$$

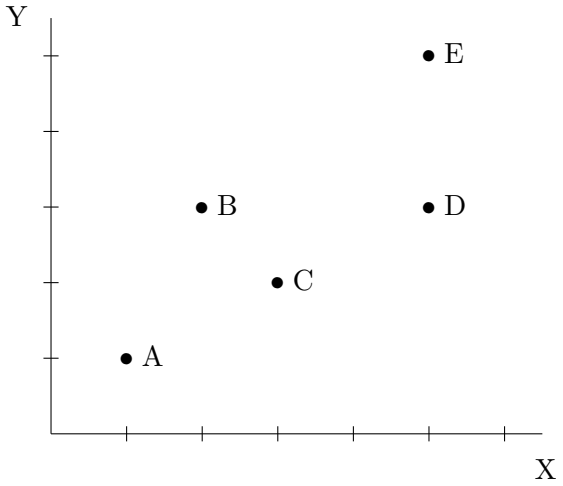
kjer je $T_C = (\bar{X}_C, \bar{Y}_C)$ težišče skupine C ter različnost d evklidska razdalja.

Vse dopustne razvrstitve (popolne razvrstitve z dvema skupinama) ter izračunane vrednosti zgoraj zapisane kriterijske funkcije za vsako razvrstitev so zapisane v tabeli 3.1. Kriterijska funkcija ima najmanjšo vrednost pri zadnji razvrstitvi:

$$P(\mathcal{C}_{15}) = 5.41$$

Najboljša razvrstitev je torej

$$\mathcal{C}^* = \{\{A, B, C\}, \{D, E\}\}$$



Slika 3.1: Grafična predstavitev razvrstitve petih enot

$\mathcal{C}$	C_1	C_2	T_1	T_2	$P(\mathcal{C})$
1	A	$BCDE$	$(1.0; 1.0)$	$(3.75; 3.25)$	6.65
2	B	$ACDE$	$(2.0; 3.0)$	$(3.50; 2.75)$	8.18
3	C	$ABDE$	$(3.0; 2.0)$	$(3.25; 3.00)$	8.67
4	D	$ABCE$	$(5.0; 3.0)$	$(2.75; 2.75)$	7.24
5	E	$ABCD$	$(5.0; 5.0)$	$(2.75; 2.25)$	5.94
6	AB	CDE	$(1.5; 2.0)$	$(4.33; 3.33)$	6.66
7	AC	BDE	$(2.0; 1.5)$	$(4.00; 3.67)$	7.21
8	AD	BCE	$(3.0; 2.0)$	$(3.33; 3.33)$	9.58
9	AE	BCD	$(3.0; 3.0)$	$(3.33; 2.67)$	9.48
10	BC	ADE	$(2.5; 2.5)$	$(3.67; 3.00)$	8.48
11	BD	ACE	$(3.5; 3.0)$	$(3.00; 2.67)$	9.34
12	BE	ACD	$(3.5; 4.0)$	$(3.00; 2.00)$	8.08
13	CD	ABE	$(4.0; 2.5)$	$(2.67; 3.00)$	8.58
14	CE	ABD	$(4.0; 3.5)$	$(2.67; 2.33)$	9.11
15	DE	ABC	$(5.0; 4.0)$	$(2.00; 2.00)$	5.41

Tabela 3.1: Pregled vseh razvrstitev z izračunanimi vrednostmi kriterijske funkcije

Glede na grafični prikaz smo to razvrstitev tudi pričakovali kot rešitev. Iz tega preprostega primera (razvrščamo le pet enot v popolno razvrstitev z dvema skupinama) je razvidno, da je število vseh dopustnih razvrstitev, za katere je potrebno pregledati vrednosti kriterijske funkcije, 15. V splošnem je v primeru razvrščanja n enot v dve neprekrivajoči skupini

$$2^{n-1} - 1$$

vseh dopustnih razvrstitev, kar pomeni, da z naraščanjem števila enot število razvrstitev eksponentno narašča.

V primeru razvrščanja n enot v k skupin je število vseh popolnih razvrstitev enako Stirlingovemu številu druge vrste

$$\mathcal{S}(n,k) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^{k-i}\binom{k}{i}i^n$$

Če bi zgoraj omenjenih pet enot razvrščali v tri skupine, bi morali pregledati 25 različnih dopustnih razvrstitev. Število razvrstitev eksplodira z večanjem števila enot. Na primer: število vseh možnih različnih razvrstitev tridesetih enot v deset skupin je

$$\mathcal{S}(30,10) = 17337334359918936459 4756$$

In trideset enot je zelo majhna množica enot. Ponavadi razvrščamo v skupine nekaj sto, včasih nekaj tisoč enot!

3.5 Reševanje problema razvrščanja

Pokazali smo, da lahko problem razvrščanja v skupine obravnavamo kot optimizacijski problem nad množico (dopustnih) razvrstitev. Ponavadi ga označimo z dvojico (Φ, P). Privzamemo lahko,

da je množica enot E končna. Tedaj je končna tudi množica dopustnih razvrstitev Φ. Problem razvrščanja v skupine je zato, vsaj teoretično, vedno rešljiv.

Če število enot ni preveliko (na primer manjše od 15), lahko problem razvrščanja v skupine rešimo s pregledom vseh možnih dopustnih razvrstitev. Pri večjem številu enot pa je množica dopustnih razvrstitev, v kateri iščemo optimalno, veliko preobsežna, da bi jo lahko z izbrano kriterijsko funkcijo v celoti pregledali. Kako eksponentno raste število popolnih razvrstitev s številom enot, kaže Stirlingovo število, ki smo ga zapisali v prejšnjem razdelku.

Nekatere probleme razvrščanja v skupine je mogoče učinkovito rešiti. Z drugimi besedami to pomeni, da je mogoče pokazati, da so ekvivalentni optimizacijskim problemom, za reševanje katerih so poznani učinkoviti (rešljivi v polinomskem času) eksaktni algoritmi. V splošnem pa lahko rečemo, da je večina problemov razvrščanja v skupine računsko zelo zahtevnih in, kot je videti, sodijo med NP-težke probleme (NP je oznaka za Nedeterminističen Polinomski) (Shamos 1976, str. 272; Brucker 1978; Garey in Johnson 1979, str. 281; Batagelj 1985, 60-73). Iz teh rezultatov torej izhaja, da za večino problemov razvrščanja v skupine ne obstajajo učinkoviti postopki za eksaktno reševanje takih problemov. Zato se moramo zateči k približnim (hevurističnim) postopkom, ki so relativno hitri in ki dajo dobre, a ne vedno najboljše rezultate.

Ta spoznanja teorije zahtevnosti algoritmov in problemov, ki jasneje razkrivajo problem razvrščanja v skupine, so znana šele dobrih deset let. Raziskovalci, ki so hoteli rešiti svoje probleme razvrščanja, pa so zahtevnost teh problemov občutili že desetletja pred omenjenimi teoretičnimi rezultati in si zato pomagali z različnimi hevurističnimi pristopi, ki so bili bolj ali manj prilagojeni reševanim problemom. Večino teh metod razvrščanja v skupine

smo pregledno predstavili v uvodnem poglavju.

Opisani optimizacijski pristop vsekakor predstavlja možno teoretično osnovo razvrščanja v skupine, ker omogoča enoten pogled na razvrščanje v skupine in zmore zastaviti ter analizirati vprašanje zahtevnosti problemov razvrščanja v skupine. S temi spoznanji se lahko tudi vprašamo, v čem je hevristika posamezne metode, kako vendar priti do čim boljše, po možnosti optimalne rešitve itd. Ta vprašanja so obravnavana ob predstavitvah posameznih metod, ki so zajete v naslednjih poglavjih.

4.

Hierarhične metode

4.1 Postopek združevanja v skupine

Najobsežnejši razred metod hierarhičnega razvrščanja v skupine predstavljajo metode, ki temeljijo na zaporednem združevanju (zlivanju) dveh ali več skupin v novo skupino. Med njimi so najpogostejše tiste, ki združijo vsakič po dve skupini. Te lahko v grobem opišemo z naslednjim postopkom:

vsaka enota je skupina:
$C_i = \{X_i\}$, $X_i \in E$, $i = 1, 2, ..., n$

ponavljaj, dokler ne ostane ena sama skupina:
določi najbližji si skupini C_p in C_q:
$d(C_p, C_q) = \min_{u,v} d(C_u, C_v)$;
združi skupini C_p in C_q v skupino $C_r = C_p \cup C_q$;
zamenjaj skupini C_p in C_q s skupino C_r;
določi mere različnosti d med novo skupino C_r in ostalimi.

Denimo, da s tem postopkom razvrščamo n enot v skupine. Postopek začne z razvrstitvijo z n skupinami (vsaka enota je v svoji skupini) in po natančno $n-1$ korakih konča z razvrstitvijo z eno samo skupino (vse enote se zlijejo v eno skupino). S postopkom združevanja v skupine torej dobimo zaporedje n popolnih razvrstitev

$$\mathcal{C}_n \ , \ \mathcal{C}_{n-1} \ , \ \mathcal{C}_{n-2} \ , \ \ldots \ , \ \mathcal{C}_2 \ , \ \mathcal{C}_1$$

kjer je $\mathcal{C}_k$ popolna razvrstitev s k skupinami, $\mathcal{C}_n = \{\{X\}|X \in E\}$ in $\mathcal{C}_1 = \{E\}$. Unijo teh razvrstitev

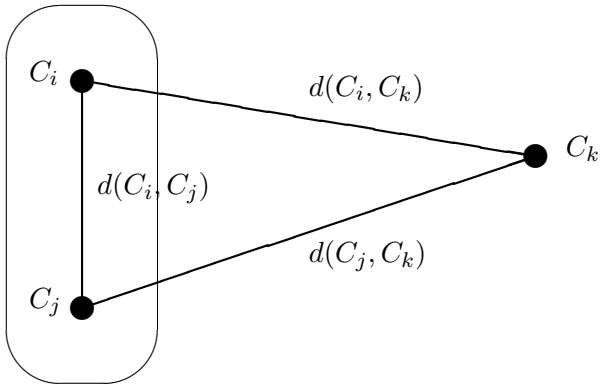
$$\mathcal{T} = \cup_{i=1}^{n} \mathcal{C}_i$$

imenujemo *drevesna razvrstitev*, ki ima naslednje lastnosti

a. $\forall X \in E : \{X\} \in \mathcal{T}$
b. $\forall C, C' \in \mathcal{T} : C \cap C' \in \{C, C', \emptyset\}$
c. $E \in \mathcal{T}$

4.2 Metode hierarhičnega združevanja v skupine

Metode hierarhičnega združevanja v skupine, opisane z zgornjim postopkom, običajno predpostavljajo, da so mere različnosti med enotami že izračunane in na osnovi teh so na različne načine določene mere različnosti med skupinami. Gre torej za dve vrsti mer različnosti: med enotami in med skupinami. Nekateri avtorji jih zato različno označujejo (npr. mero različnosti med enotami z d, med skupinami pa z D). V tem delu obe meri različnosti označujemo z d. Iz posameznih definicij nedvoumno sledi, za kateri tip mer različnosti gre.



Slika 4.1: Tri skupine

Mere različnosti d med novo skupino in ostalimi v postopku združevanja v skupine določamo na več načinov in ti določajo različne metode hierarhičnega združevanja v skupine. Vzemimo, da imamo v nekem koraku postopka tri skupine C_i, C_j in C_k ter podane mere različnosti med njimi približno tako, kot je prikazano na sliki 4.1. Denimo, da sta skupini C_i in C_j najbližji, zato ju združimo v novo skupino $C_i \cup C_j$. Mero različnosti med novo skupino in skupino C_k lahko določimo na primer na naslednje načine:

- **Minimalna metoda** ali tudi enojna povezanost (Florek et al. 1951; Sneath 1957):

$$d(C_i \cup C_j, C_k) = \min(d(C_i, C_k), d(C_j, C_k))$$

- **Maksimalna metoda** ali tudi polna povezanost (McQuitty 1960):

$$d(C_i \cup C_j, C_k) = \max(d(C_i, C_k), d(C_j, C_k))$$

- **McQuittyjeva metoda** (McQuitty 1966, 1967):

$$d(C_i \cup C_j, C_k) = \frac{d(C_i, C_k) + d(C_j, C_k)}{2}$$

Različnosti med novo skupino in ostalimi skupinami lahko določamo tudi na druge, bolj zapletene načine, tako da upoštevamo sestavo posameznih skupin. Naštejmo nekaj takih načinov:

- **Povprečna metoda** (Sokal in Michener 1958):

$$d(C_i \cup C_j, C_k) = \frac{1}{(n_i + n_j)n_k} \sum_{U \in C_i \cup C_j} \sum_{V \in C_k} d(U, V)$$

  Oznaka n_i pomeni število enot v skupini C_i.

- **Gowerjeva metoda** (Gower 1967):

$$d(C_i \cup C_j, C_k) = d^2(T_{ij}, T_k)$$

  kjer s T_{ij} označimo težišče združene skupine $C_i \cup C_j$ in s T_k težišče skupine C_k.

- **Wardova metoda** (Ward 1963):

$$d(C_i \cup C_j, C_k) = \frac{(n_i + n_j)n_k}{n_i + n_j + n_k} d^2(T_{ij}, T_k)$$

Postopek združevanja v skupine po maksimalni metodi prikažimo z združevanjem republik in pokrajin v skupine glede na odstotek članov ZKJ in ZZBNOV v celotnem prebivalstvu (podatki so podani v uvodnem poglavju). Različnost med enotama naj bo merjena z evklidsko razdaljo, pri čemer naj bosta pred računanjem razdalj obe spremenljivki standardizirani. Te razdalje smo že izračunali v poglavju o merjenju podobnosti in prikazali v tabeli 2.1. Zapišimo jih ponovno

	B	C	H	M	S	O	K	V
B	0.0	3.1	2.1	0.7	2.9	2.3	1.0	2.0
C		0.0	2.3	3.0	2.7	1.0	3.8	1.1
H			0.0	1.5	0.8	1.3	2.2	1.7
M				0.0	2.3	2.0	0.8	1.9
S					0.0	1.8	2.8	2.3
O						0.0	2.9	0.6
K							0.0	2.7
V								0.0

Sledimo postopku združevanja, ki smo ga zapisali na začetku tega poglavja. Najbližji enoti sta ožja Srbija in Vojvodina, ker je razdalja med njima najmanjša (0.6). Združimo ju. V naslednjem koraku moramo izračunati razdalje med novo skupino (O,V) in ostalimi. Po maksimalni metodi je razdalja med novo skupino in na primer Bosno in Hercegovino (B) določena takole

$$d((O,V),B) = \max(d(O,B), d(V,B)) = \max(2.3\ ,\ 2.0) = 2.3$$

Podobno izračunamo razdalje med novo skupino in preostalimi republikami in pokrajino. V matriki razdalj moramo razdalje, ki se nanašajo na ožjo Srbijo in Vojvodino zamenjati z izračunanimi razdaljami, ki se nanašajo na združeno ožji Srbiji z Vojvodino. S tem se matrika zmanjša za en en stolpec in eno vrstico. Popravljena matrika je tedaj

	B	C	H	M	S	(O,V)	K
B	0.0	3.1	2.1	0.7	2.9	2.3	1.0
C		0.0	2.3	3.0	2.7	1.1	3.8
H			0.0	1.5	0.8	1.7	2.2
M				0.0	2.3	2.0	0.8
S					0.0	2.3	2.8
(O,V)						0.0	2.9
K							0.0

V drugem koraku postopka združimo Makedonijo ter Bosno in Hercegovino (razdalja med njima je 0.7). Če na novo izračunamo razdalje med novo skupino (B,M) in ostalimi po maksimalni metodi, dobimo naslednjo matriko

	(B,M)	C	H	S	(O,V)	K
(B,M)	0.0	3.1	2.1	2.9	2.3	1.0
C		0.0	2.3	2.7	1.1	3.8
H			0.0	0.8	1.7	2.2
S				0.0	2.3	2.8
(O,V)					0.0	2.9
K						0.0

V tem koraku združimo pri razdalji 0.8 Hrvatsko in Slovenijo - tedaj je matrika razdalj

	(B,M)	C	(H,S)	(O,V)	K
(B,M)	0.0	3.1	2.9	2.3	1.0
C		0.0	2.7	1.1	3.8
(H,S)			0.0	2.3	2.8
(O,V)				0.0	2.9
K					0.0

V četrtem koraku se Kosovo pridruži skupini (B,M) pri razdalji 1.0. Popravljene razdalje so tedaj

	(B,M,K)	C	(H,S)	(O,V)
(B,M,K)	0.0	3.8	2.9	2.9
C		0.0	2.7	1.1
(H,S)			0.0	2.3
(O,V)				0.0

V tem koraku se Črna gora pridruži skupini (O,V) pri razdalji 1.1 in matrika je

	(B,M,K)	(C,O,V)	(H,S)
(B,M,K)	0.0	3.8	2.9
(C,O,V)		0.0	2.7
(H,S)			0.0

V predzadnjem koraku se zlijeta skupini (C,O,V) in (H,S) pri razdalji 2.7 in matrika je

	(B,M,K)	(C,O,V,H,S)
(B,M,K)	0.0	3.8
(C,O,V,H,S)		0.0

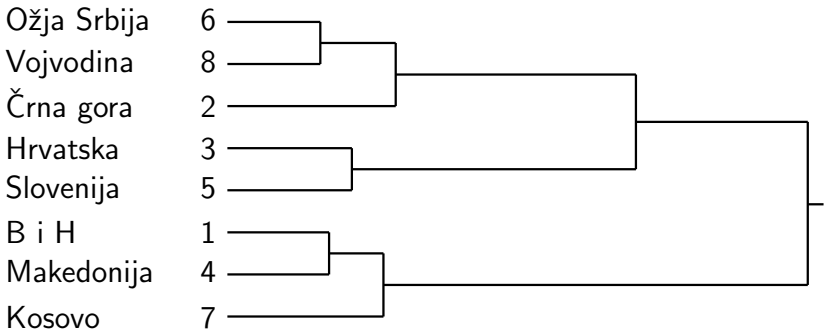
Vse enote pa se zlijejo v eno skupino pri razdalji 3.8 .

4.3 Drevo združevanja

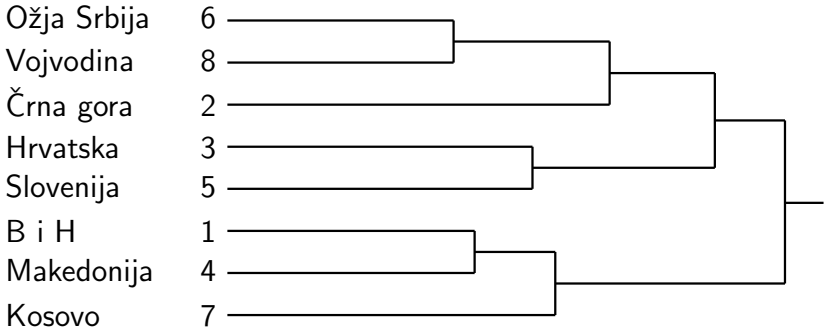
Potek združevanja si lahko grafično ponazorimo z drevesom združevanja - dendrogramom. Listi tega drevesa so enote, točke združitve pa sestavljene skupine: levi in desni naslednik vsake točke sta skupini, iz katerih je nastala. Višina točke, ki jo imenujemo nivo združevanja, je sorazmerna meri različnosti med skupinama. Dendrogram za obravnavani primer, kjer smo združevali republike in pokrajini po maksimalni metodi, je predstavljen na sliki 4.2.

Podobno lahko republike in pokrajini združimo z drugimi omenjenimi metodami. Na sliki 4.3 je podano drevo združevanja, dobljeno po minimalni metodi, in na sliki 4.4 po Wardovi metodi. Očitno je, da se drevesne razvrstitve, dobljene po teh treh metodah, bistveno ne razlikujejo.

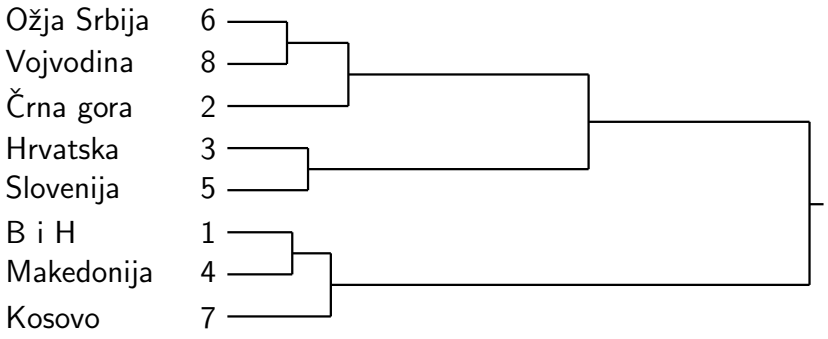
Že večkrat smo omenili, da je pri razvrščanju v skupine težko vnaprej vedeti, koliko izrazitih skupin se skriva v strukturi podatkov. Pri metodah združevanja v skupine na srečo ni potrebno vnaprej podati števila skupin. Še več, te metode omogočajo s pregledom nivojev združevanja analitično določitev primernega števila skupin. To je namreč določeno s številom vej drevesa združevanja, ki jih dobimo z rezanjem drevesa pri največjem skoku



Slika 4.2: Republike in pokrajini - maksimalna metoda



Slika 4.3: Republike in pokrajini - minimalna metoda



Slika 4.4: Republike in pokrajini - Wardova metoda

(prirastku) dveh sosednjih nivojev združevanja. V našem primeru je iz drevesa združevanja, dobljenega z maksimalno metodo, razviden največji skok od nivoja 1.1 na nivo 2.7 . Če drevo režemo na tem intervalu, dobimo tri veje. V tem primeru je torej najprimernejša razvrstitev v tri skupine. Tudi drevesi, dobljeni z minimalno in Wardovo metodo, kažeta, da se republike in pokrajini razvrščajo v tri izrazite skupine. To pravzaprav v tem primeru že vemo: spomnimo se grafičnega prikaza na sliki 3 v uvodnem poglavju. Ta primer torej kaže, da znajo vse tri metode združevanja odlično razkriti strukturo v podatkih.

4.4 Lance in Williamsov obrazec

Lance in Williams (1967) sta pokazala, da je mogoče večino metod hierarhičnega združevanja v skupine predstaviti kot posebne primere metode, pri kateri se nove mere različnosti določajo po nasled-

njem obrazcu

$$d(C_i \cup C_j, C_k) =$$

$$= \alpha_1 d(C_i, C_k) + \alpha_2 d(C_j, C_k) + \beta d(C_i, C_j) + \gamma |d(C_i, C_k) - d(C_j, C_k)|$$

S primerno izbiro koeficientov α_1, α_2, β in γ v zgornjem obrazcu dobimo večino znanih načinov zaporednega združevanja v skupine. Za primer izpeljimo štiri koeficiente za minimalno metodo. V ta namen se spomnimo naslednje enakosti

$$\min(a, b) = \frac{1}{2}(a + b - |a - b|)$$

Mera različnosti $d(C_i \cup C_j, C_k)$ po minimalni metodi je

$$d(C_i \cup C_j, C_k) = \min(d(C_i, C_k), d(C_j, C_k))$$

Na osnovi zgornje enakosti jo lahko zapišemo takole

$$d(C_i \cup C_j, C_k) = \frac{1}{2}(d(C_i, C_k) + d(C_j, C_k) - |d(C_i, C_k) - d(C_j, C_k)|$$

S primerjavo slednjega z Lance in Williamsovim obrazcem razberemo koeficiente za minimalno metodo, ki so:

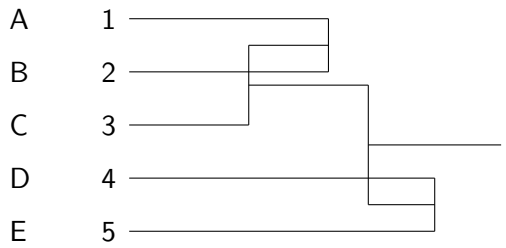
$$\alpha_1 = \frac{1}{2}, \alpha_2 = \frac{1}{2}, \beta = 0, \gamma = -\frac{1}{2}$$

Vrednosti koeficientov za vse omenjene metode hierarhičnega združevanja v skupine so podane v tabeli 4.1.

Jambu (1978) je Lance in Williamsov obrazec še bolj posplošil tako, da mu je dodal še tri koeficiente. Ta posplošitev je pomembna predvsem pri računalniški realizaciji hierarhičnih postopkov (npr. Batagelj 1988 b).

metoda	α_1	α_2	β	γ
minimum	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
maksimum	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
McQuitty	$\frac{1}{2}$	$\frac{1}{2}$	0	0
povprecna	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	0	0
Gower	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0
Ward	$\frac{n_i+n_k}{n_i+n_j+n_k}$	$\frac{n_j+n_k}{n_i+n_j+n_k}$	$-\frac{n_k}{n_i+n_j+n_k}$	0

Tabela 4.1: Vrednosti koeficientov v Lance in Williamsovem obrazcu



Slika 4.5: Nemonotono drevo združevanja

4.5 Monotonost

Na osnovi Lance in Williamsovega obrazca lahko generiramo neskončno mnogo metod združevanja v skupine. Vpašanje je, ali vsak nabor štirih koeficientov določa metodo, ki smiselno razkriva strukturo v podatkih. Na to vprašanje lahko odgovorimo tako, da kreiramo kriterije, s katerimi lahko ocenimo, ali je določena metoda združevanja smiselna. Eden takih kriterijev je *monotonost* metode. Ko združimo skupini C_i in C_j v novo skupino $C_r = C_i \cup C_j$, se namreč lahko zgodi, da je mera različnosti (nivo združevanja), pri kateri združimo skupini C_i in C_j, manjša od mere različnosti, pri katerih smo v prejšnjih korakih združili v skupino C_i oziroma v skupino C_j. Metoda, pri kateri se lahko zgodi tak pojav, zgradi 'nenaravno raščeno drevo' oziroma nemonotono drevo. Primer nemonotonega drevesa je podan na sliki 4.5.

Povedano lahko strnemo v naslednjo definicijo: naj bo h nivo združevanja skupin v drevesu združevanja, ki je definiran na na-

slednji način

$$X \in E \Longrightarrow h(\{X\}) = 0$$

$$C_r = C_i \cup C_j \Longrightarrow h(C_r) = d(C_i, C_j)$$

Drevo združevanja je monotono natanko tedaj, ko za vsako skupino $C_r = C_i \cup C_j$ v drevesu velja

$$h(C_r) \geq \max(h(C_i), h(C_j))$$

Batagelj (1981) je dokazal naslednji izrek: Metoda hierarhičnega združevanja v skupine, osnovana na obrazcu Lancea in Williamsa, zagotavlja monotona drevesa (drevesne razvrstitve) natanko tedaj, ko so izpolnjeni naslednji trije pogoji:

$$\gamma \geq -\min(\alpha_1, \alpha_2)$$

$$\alpha_1 + \alpha_2 \geq 0$$

$$\alpha_1 + \alpha_2 + \beta \geq 1$$

Prvi pogoj je zadoščen pri vseh omenjenih metodah, drugi tudi, tretji pa ni na primer pri Gowerjevi metodi.

4.6 Hevristika metod združevanja v skupine

Kako se kaže hevristika pri metodah združevanja v skupine? Hierarhične metode je mogoče povezati z optimizacijskim pristopom, ki smo ga predstavili v prejšnjem poglavju. Z drugimi besedami: postopek združevanja je mogoče opisati z ustrezno kriterijsko funkcijo (npr. Batagelj 1987, 1988). Iz te izhaja 'požrešna' hevristika, ki pomeni naslednje: postopek se začne z združevanjem enot (rekli

smo skupin) in nato v vsakem koraku združi najbližji skupini. V naslednjih korakih (pri manjšem številu skupin) se lahko izkaže, da bi bilo bolje, ko bi v prejšnjih korakih združeval drugače, vendar pomagati se ne da. Kar je bilo v prejšnjih korakih združeno, je pač združeno. Zato se učinek 'požrešnosti' manj pozna na nižjih nivojih združevanja (pri večjem številu skupin) in bolj pri višjih. To tudi pomeni, da so razvrstitve, dobljene z rezanjem drevesa združevanja na višjih nivojih, v splošnem manj zanesljive.

V primeru izrazite strukture ne pride do učinka 'požrešnosti'. Problem pa je v tem, da ne vemo, kakšno strukturo podatkov imamo. Najlažje ugotovimo, ali je prišlo do slabše rešitve zaradi 'požrešnosti' postopka, če podatke analiziramo še z drugimi nehierarhičnimi metodami razvrščanja v skupine in preverimo stabilnost dobljene razvrstitve.

Pri izraziti naravni strukturi podatkov, kjer so skupine ločene med seboj, dobimo pravo razvrstitev v skupine z vsako omenjeno metodo hierarhičnega združevanja. To pa ne velja, če so skupine prekrivajoče ali zelo specifično oblikovane (npr. prepletajoče klobase). Tedaj se razvrstitve, dobljene z različnimi metodami, razlikujejo med seboj, in sicer toliko bolj, kolikor bolj je naravna struktura podatkov slaba, neizrazita.

4.7 Nekaj lastnosti metod združevanja v skupine

Metode hierarhičnega združevanja v skupine so zelo priljubljene iz več razlogov: postopek je relativno preprost, rezultat združevanja je mogoče nazorno prikazati z drevesom združevanja (dendrogramom), v splošnem postopek zahteva relativno malo računalniškega časa, uporabniku ni potrebno vnaprej določiti števila skupin. Ker

je poznanih več metod hierarhičnega združevanja, se uporabnik sooči s problemom, katero metodo izbrati.

V literaturi so najpogosteje omenjene minimalna, maksimalna in Wardova metoda. Najbrž zato, ker ima vsaka od teh metod zanimive specifične lastnosti. Minimalna metoda se imenuje tudi enojna povezanost (single linkage), ker v vsakem koraku postopka združuje tisti skupini, med katerima obstaja največja povezanost oziroma najmanjša različnost, izmerjena med najbližjima enotama ene in druge skupine. Zato se minimalna metoda zelo obnese pri razkrivanju dolgih 'klobasastih', tudi neeliptičnih struktur. Po drugi strani pa je ta metoda neuporabna pri neizrazito ločenih skupinah. V teh primerih se kaže 'verižni' učinek metode, ko v vsakem koraku združevanja skupini dodaja le posamezno enoto. Veriženje je zaznati tudi pri nekaterih drugih hierarhičnih metodah (npr. pri Gowerjevi metodi). Minimalna metoda je invariantna za vsako transformacijo mere podobnosti, ki ohranja urejenost parov enot, oziroma, če so mere podobnosti enakovredne. To pomeni, da dobimo enake drevesne razvrstitve, če izhajamo iz enakovrednih mer podobnosti med enotami. Ta lastnost velja tudi za maksimalno metodo. Zaželena je tudi monotonost drevesne razvrstitve. Omenili smo že, da Gowerjeva metoda ne zagotavlja monotonih rešitev in je zato manj priporočljiva.

V uvodnem poglavju smo omenili, da sta zaželeni lastnosti skupin interna kohezivnost in eksterna izolacija. Minimalna in maksimalna metoda se v tem smislu obnašata prav nasprotno. Minimalna metoda išče skupine, ki so izrazito ločene med seboj in se ne zmeni za kohezivnost znotraj njih. Maksimalna metoda pa je osredotočena na razkrivanje znotraj kohezivnih skupin. Preostale omenjene metode sledijo bolj ali manj obema lastnostima skupin.

Več avtorjev (npr. Everitt, 1974; Mojena, 1978; Ferligoj in Batagelj, 1980) je empirično primerjalo različne metode hierarhi-

čnega združevanja v skupine na več slučajno generiranih skupinah podatkov in pri tem ugotavljalo primernost posameznih metod. Te empirične primerjave so pokazale, da je Wardova metoda najprimernejša za eliptično strukturirane podatke, medtem ko je minimalna metoda primerjejša za odkrivanje verižno strukturiranih podatkov. Maksimalna metoda pa dobro razkriva okrogle skupine.

4.8 Smeri razvoja

Hierarhične metode so najbrž najbolj zanimivo in najbolj raziskovano področje razvrščanja v skupine, ker drevesna struktura rešitev omogoča razburljive razvojne poti. Naj navedem le nekaj smeri tega razvoja.

Denimo, da smo za dano množico enot na različne načine tvorili drevesne razvrstitve. Kako primerjati drevesne razvrstitve (npr. Day 1985)? Kako najti za dane drevesne razvrstitve eno razvrstitev, ki bo karseda dobro predstavljala skupino drevesnih razvrstitev? V literaturi to razvrstitev imenujejo konsenzna razvrstitev. Konsenznim razvrstitvam in primerjavi drevesnih razvrstitev je bila posvečena posebna številka revije Journal of Classification v letu 1986, kjer sta francoska in ameriška šola predstavili svoje najnovejše dosežke.

Diday (npr. 1986) je posplošil drevesne razvrstitve tako, da je z njimi mogoče grafično predstaviti združevanje prekrivajočih se skupin. Ta posplošena trirazsežna drevesa se imenujejo piramide. Raziskovanje gre v smeri razkrivanja lastnosti teh dreves in iskanja ustreznih metod za njihovo določitev.

V tem delu so omenjeni le najvidnejši dosežki na področju hierarhičnega razvrščanja v skupine in omenjena le največkrat citirana literatura, ki bralcu lahko razkrije tukaj komaj dotaknjene ali neomenjene rezultate. Naj za konec opozorim še na Gordonov

članek (1987), v katerem je zelo dober pregled novejših dosežkov na tem področju.

4.9 Primera uporabe metod združevanja v skupine

4.9.1 Tipologija aktivnosti v prostem času

Za empirično raziskovanje prostočasnih aktivnosti je psihologinja Sadarjeva 116 dijakom tretjih razredov gimnazije Ivana Cankarja v Ljubljani v letu 1974 postavila vprašanje, kaj delajo v prostem času. S tem vprašanjem je dopustila, da je vsak dijak sam določil, kaj so zanj prostočasne aktivnosti. Niti eden od anketiranih dijakov ni navedel kot svojo prostočasno aktivnost domača opravila, gledanje televizije in poslušanje radia. Prostočasne aktivnosti z oznakami in številom dijakov, ki so zapisali določeno aktivnost, so podane v tabeli 4.2. Aktivnosti so urejene od največkrat do najmanjkrat omenjene. Več kot polovica dijakov je navedla branje knjig kot prostočasno aktivnost. Veliko frekvenco imajo tudi druge manj zahtevne aktivnosti: smučanje, polušanje glasbe in ukvarjanje z ročnimi deli.

Zanima nas, katere prostočasne aktivnosti gojijo dijaki hkrati oziroma katere aktivnosti so si bližje, podobnejše. Z drugimi besedami: dobiti želimo tipologijo prostočasnih aktivnosti. Najprej se moramo odločiti, kdaj sta si dve aktivnosti podobni in kako bomo to podobnost (ali različnost) nejustrezneje merili. Z dano aktivnostjo se dijak ukvarja ali pa ne - torej ima posamezna aktivnost (spremenljivka) dve vrednosti: da in ne. Aktivnosti sta si tedaj tem bolj podobni, čim bolj se nagnjenja do obeh aktivnosti ujemajo. Torej: aktivnosti sta si podobni, če dijak, ki goji i-to aktivnost, zelo verjetno goji tudi j-to aktivnost, in seveda, če ne

f_i	OZNAKA	OPIS AKTIVNOSTI
69	BERE	branje knjig
48	SMUCA	smučanje
38	GLASBA	poslušanje glasbe, plošč
35	ROCDEL	ročna dela: šivanje, pletenje, vezenje,...
28	CSPORT	član športnega društva
27	ZOGA	košarka, nogomet, odbojka, rokomet
26	JEZIKKR	jezikovni krožki
25	KINO	kino
25	IGRA	igranje instrumentov
24	DRSA	drsanje
23	TENIS	namizni tenis, tenis, badmington
22	GORE	planinarjenje, alpinizem
21	OSTALOSP	ostale fizične aktivnosti: karate, kolesarjenje, telovadba, veslanje
21	OSTALKR	debatni, naravoslovni, šahovski krožek
19	UMKRE	umetniška kreacija (slikanje, risanje, pisanje pesmi, proze)
18	TABOR	Zveza tabornikov, Planinsko društvo
16	SPREHOD	hodi v naravo, sprehodi
14	PLAVA	plavanje
14	UMETKR	literarni, dramski, likovni krožek
13	GLEDA	gledališče, koncerti
13	PLES	ples
13	PROBLEM	zahtevne mentalne aktivnosti (šah, reševanje problemov, računalništvo,...)
11	OBISK	obiski prijateljev

Tabela 4.2: Prostočasne aktivnosti

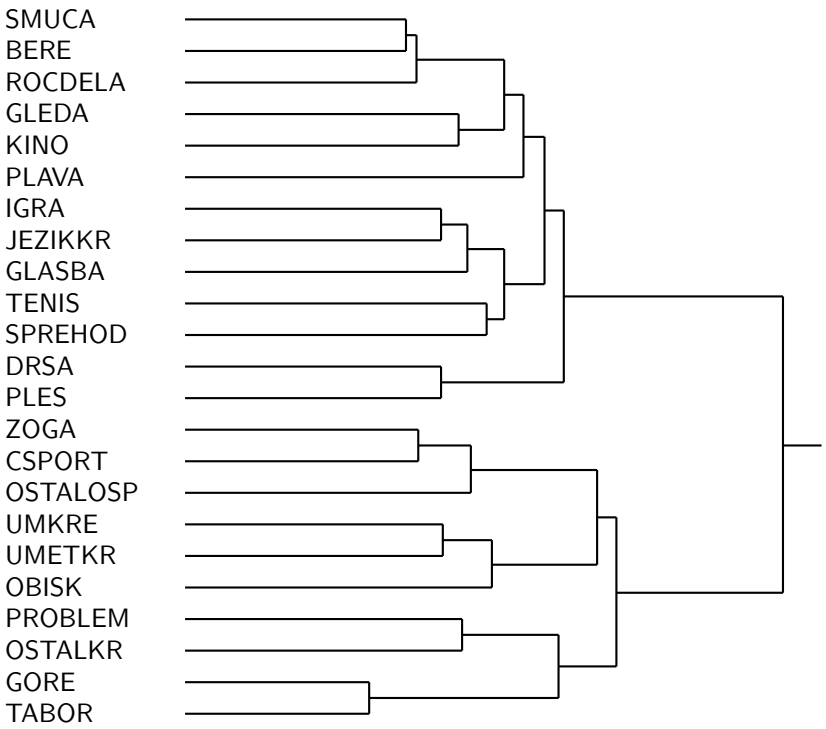
goji i-te aktivnosti, zelo verjetno ne bo gojil j-te. Upoštevati pa je potrebno še način zbiranja podatkov: dijaki so namreč morali sami navesti svoje prostočasne aktivnost, pri čemer so lahko neke aktivnosti pozabili navesti ali pa so se jim zdele manj pomembne. Zato so odgovori 'ne' lahko vprašljivi. V tem primeru je, glede na vse povedano, zelo primerna mera povezanosti med prostočasnimi aktivnostmi Jaccardovo mero ujemanja J (glej razdelek 2.2). Mero različnosti d pa določimo z naslednjo transformacijo

$$d = 1 - J$$

Tako določene različnosti med prostočasnimi aktivnostmi so začetni podatki za hierarhično združevanje v skupine. Iz množice metod hierarhičnega združevanja v skupine smo izbrali Wardovo metodo, ki jo več avtorjev priporoča kot najprimernejšo. Oglejmo si drevo združevanja, ki je prikazano na sliki 4.6. Iz dendrograma sta razvidni dve izraziti skupini aktivnosti in sicer zgornja skupina aktivnosti družabno-zabavnega značaja ter spodnja skupina zahtevnejših intelektualnih aktivnosti z aktivno rekreacijo. Slednja skupina se nadalje deli na tri izrazitejše skupine in sicer: športne aktivnosti, umetniška kreacija in zahtevnejše intelektualne aktivnosti z aktivno rekreacijo.

4.9.2 Tipologija evropskih držav glede na razvojne kazalce

V drugem primeru razvrščamo evropske države v skupine glede na družbeno-ekonomske in demografske spremenljivke. V analizi podatkov smo upoštevali le 27 evropskih držav, ker za majhne evropske države (Andora, Liechtenstein, Vatikan, San Marino, Monaco in Malta) ni vseh podatkov. Izbrali smo naslednje družbeno- ekonomske in demografske spremenljivke (Vir: The Hammond Almanac, 1980):



Slika 4.6: Prostočasne aktivnosti - Wardova metoda

- osebni dohodek na prebivalca
- delež industrijske proizvodnje glede na vso proizvodnjo
- odstotek urbanega prebivalstva
- odstotek prebivalcev v največjem mestu
- gostota prebivalcev
- nataliteta
- mortaliteta
- pričakovano trajanje življenja
- mortaliteta dojenčkov
- število prebivalcev na zdravnika
- število prebivalcev na bolniško posteljo
- število slušateljev višjih in visokih šol na prebivalca
- površina tlakovanih cest na celotno površino
- dolžina železnice na površino
- število motornih vozil na prebivalca
- število radijskih sprejemnikov na prebivalca
- število TV naročnikov na prebivalca
- število telefonskih naročnikov na prebivalca
- število primerkov časopisov na prebivalca

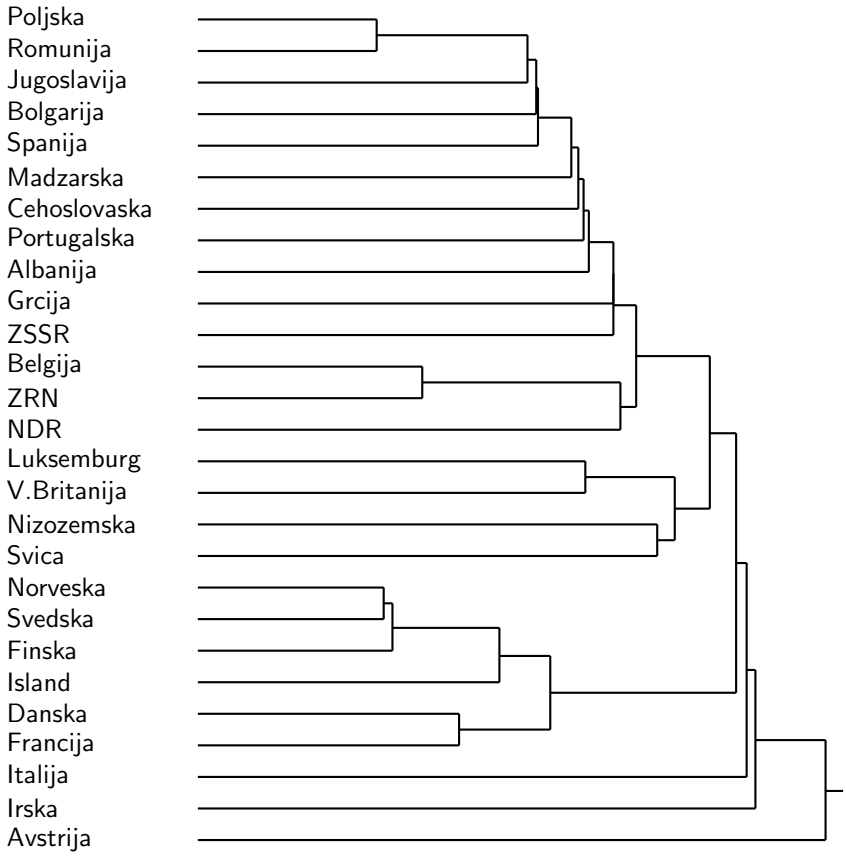
Podobnost med državama smo merili s Pearsonovim korelacijskim koeficientom tako, da smo pred tem spremenljivke standardizirali (vsaki vrednosti spremenljivke smo odšteli njeno povprečje in razliko delili s standardnim odklonom spremenljivke). Merjena je torej podobnost med 'profiloma' držav. Mero različnosti d pa smo izračunali takole:

$$d_{i,j} = \frac{1 - r_{i,j}}{2}$$

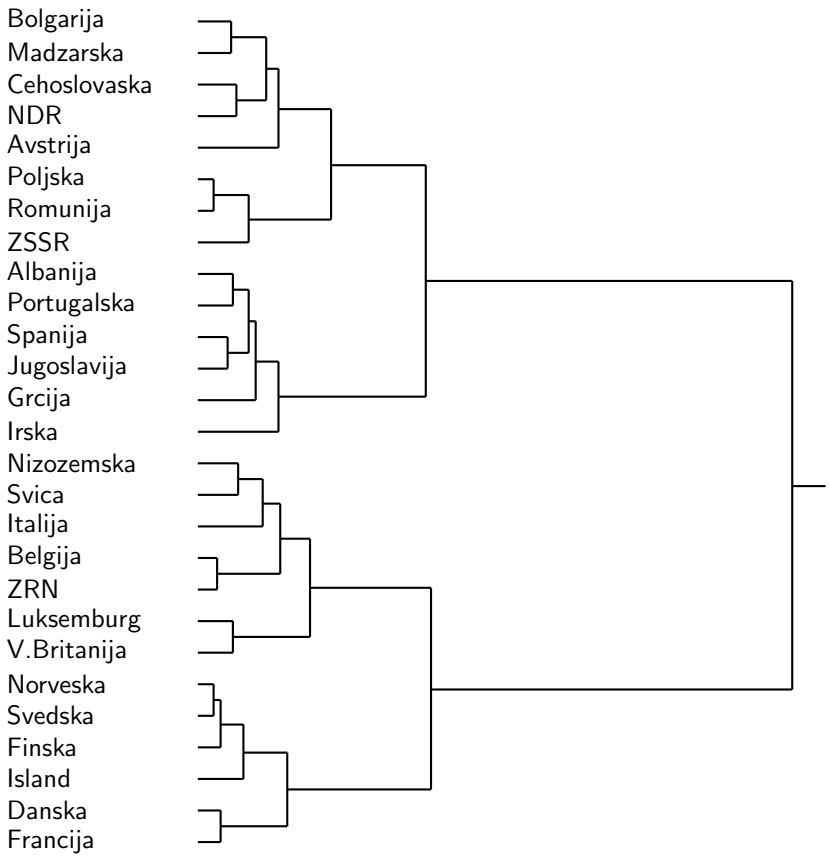
kjer je $r_{i,j}$ Pearsonov korelacijski koeficient za i-to in j-to enoto.

Na naslednjih dveh straneh sta podana dendrograma, ki sta dobljena z minimalno in Wardovo metodo hierarhičnega združevanja v skupine. V zgornjem delu drevesa, dobljenega z minimalno

metodo, se kaže tipični učinek veriženja: v posameznih korakih združevanja se skupinam pridružijo le posamezne enote. Minimalna metoda namreč išče optimalno povezane skupine in ne homogene, kompaktne skupine. Ta metoda lahko združi relativno oddaljeni skupini, če je med njima le nekaj enot. Drevo združevanja, dobljeno z Wardovo metodo, kaže dve izraziti skupini in sicer zgornjo skupino manj razvitih evropskih držav in spodnjo skupino bolj razvitih evropskih držav. Vsaka od teh dveh skupin se nato razcepi in sicer skupina manj razvitih držav v tri izrazitejše skupine: 'Avstro-ogrsko' skupino, skupino vzhodnih držav in 'sredozemsko' skupino; skupina bolj razvitih držav pa v dve izrazitejši skupini: skupino zahodnih držav in 'skandinavsko' skupino.



Slika 4.7: Evropske države - minimalna metoda



Slika 4.8: Evropske države - Wardova metoda

5.

Nehierarhične metode

Nehierarhične metode razvrščanja v skupine se od hierarhičnih ločijo predvsem v tem, da je potrebno vnaprej podati število skupin iskanih razvrstitev. Večina teh metod razvršča enote tako, da glede na izbrani kriterij izboljšuje neko začetno razvrstitev. Ker je pregled vseh dopustnih razvrstitev največkrat nemogoč, si te metode pomagajo tako, da pregledajo le del množice razvrstitev, v upanju, da je optimalna rešitev v njej. V splošnem pa je razvrstitev, dobljena s temi metodami, le lokalno optimalna, zato jim rečemo tudi metode razvrščanja v skupine z lokalno optimizacijo.

Osnova za metode lokalne optimizacije so pravila, s katerimi določimo za vsako razvrstitev iz množice dopustnih razvrstitev, v katere razvrstitve se lahko pomaknemo iz nje. S temi pravili določimo *sosedstveno strukturo* razvrstitev. Tedaj poskušamo problem razvrščanja v skupine rešiti tako, da začnemo v neki (začetni) razvrstitvi in s premikanjem po sosedstveni strukturi razvrstitev skušamo izboljšati vrednost izbrane kriterijske funkcije. S premikanjem nadaljujemo, vse dokler ne pridemo do razvrstitve, ki nima nobenega soseda z manjšo vrednostjo kriterijske

funkcije. Dobljena razvrstitev je lokalno optimalna za dano sosedstveno strukturo.

Pri metodah lokalne optimizacije je potrebno izbrati primerno sosedstveno strukturo razvrstitev in določiti učinkovito strategijo pregledovanja sosednjih razvrstitev. Pri iskanju čim bolj učinkovite metode se je potrebno zavedati, da čim več sosedov ima posamezna razvrstitev, tem večja je možnost, da dobimo globalni ekstrem, a tudi, da moramo opraviti tem več dela v posameznem koraku.

Da bi dobili čim boljšo rešitev in vtis o tem, kako dobra je dobljena rešitev, postopek ponovimo z več različnimi začetnimi razvrstitvami. Empirične primerjave metod kažejo, da se zelo približamo globalnemu ekstremu kriterijske funkcije, če vzamemo kot začetno razvrstitev rezultat razvrščanja z neko drugo metodo razvrščanja v skupine (na primer razvrstitev, dobljeno s postopki hierarhičnega združevanja v skupine).

Problem lokalnih rešitev nekateri avtorji (npr. Bonomi in Lutton 1984) uspešno rešujejo tako, da v posameznih korakih dovolijo premik tudi v razvrstitev s slabšo vrednostjo kriterijske funkcije.

Pri metodah razvrščanja v skupine z lokalno optimizacijo lahko podamo različnosti med enotami na dva načina:

- podamo matriko različnosti med enotami,
- različnosti med enotami sproti računamo.

Med metodami razvrščanja v skupine, ki temeljijo na že izračunanih različnostih med enotami, je najbolj znana metoda prestavljanj. Te metode so primerne za razvrščanje nekaj sto enot v skupine. Drugo skupino pa tvorijo metode, ki so primerne za razvrščanje v skupine tudi večjih količin podatkov (nekaj tisoč). V literaturi so te metode poznane pod imeni metoda voditeljev (npr. Hartigan 1975), K-MEANS (npr. Forgy 1965; Jancey 1966; MacQueen 1967) ali metoda (dinamičnih) oblakov (npr. Diday 1974,

1979). Oba tipa metod sta v nadaljevanju podrobneje obravnavana.

5.1 Metoda prestavljanj

Nehierarhične metode razvrščanja v skupine z lokalno optimizacijo lahko torej v grobem opišemo z naslednjim postopkom:

določi začetno dopustno razvrstitev $\mathcal{C}$

ponavljaj, dokler gre:
če med tekočo razvrstitvijo $\mathcal{C}$ in sosednjimi razvrstitvami obstaja razvrstitev $\mathcal{C}'$, za katero velja
$P(\mathcal{C}') \leq P(\mathcal{C})$
se pomakni v razvrstitev $\mathcal{C}'$.

Pri metodi prestavljanj je sosedstvena struktura razvrstitev določena takole: razvrstitvi $\mathcal{C}$ in $\mathcal{C}'$ iz množice dopustnih razvrstitev Φ sta sosednji, če dobimo razvrstitev $\mathcal{C}'$ tako, da v razvrstitvi $\mathcal{C}$ prestavimo neko enoto iz ene skupine v drugo:

$$\mathcal{C}' = (\mathcal{C} \ - \ \{C_i, C_j\}) \ \cup \ \{C_i - \{X_k\}, C_j \cup \{X_k\}\}$$

ali z zamenjavo dveh enot iz različnih skupin v razvrstitvi $\mathcal{C}$:

$$\mathcal{C}' = (\mathcal{C} \ - \ \{C_i, C_j\}) \ \cup \ \{C_i - \{X_p\} \cup \{X_q\}, C_j - \{X_q\} \cup \{X_p\}\}$$

Žal nam naslonitev na matriko različnosti, če naj bo postopek učinkovit, precej omejuje izbiro primernih kriterijskih funkcij. Pri metodi prestavljanj se v vsakem koraku spremenita le dve skupini. Zato so primerne tiste kriterijske funkcije, ki učinkovito preverjajo

pogoj $P(\mathcal{C}') \leq P(\mathcal{C})$ tako, da v vsakem premiku po sosedstveni strukturi računajo le spremembo v vrednosti kriterijske funkcije, ki je nastala zaradi sprememb v dveh skupinah. To omogoča Wardova kriterijska funkcija, ki jo nekateri avtorji imenujejo tudi kriterij vsote kvadratov (npr. Gordon 1981), kriterij minimalne variance (npr. Späth 1985) ali v francoski literaturi inercija (npr. Bouroche in Saporta 1980). Wardova kriterijska funkcija je definirana takole:

$$P(\mathcal{C}) = \sum_{C \in \mathcal{C}} \sum_{X \in C} d(X, T_C)$$

kjer je T_C težišče skupine C in d kvadrat evklidske razdalje. Edwards in Cavalli-Sforza (1965) sta pokazala, da lahko izračunamo Wardovo kriterijsko funkcijo tudi brez težišč takole:

$$P(\mathcal{C}) = \sum_{C \in \mathcal{C}} \frac{1}{2n_C} \sum_{X,Y \in C} d(X, Y)$$

kjer je n_C število enot v skupini C.

Do nedavnega je bilo mišljeno, da se metode, ki temeljijo na Wardovi kriterijski funkciji, praviloma lahko uporablja, če imamo med enotami merjene kvadrate evklidskih razdalj. V praksi so se te metode uspešno uporabljale tudi za druge mere različnosti. Po Batagelju (1988 b) je uporaba tudi drugih mer različnosti v Wardovi kriterijski funkciji povsem legalna.

Relativno preprosto se da dokazati (npr. Späth 1985, str. 19), da je Wardova kriterijska funkcija, kjer je d kvadrat evklidske razdalje, enaka sledi matrike razpršenosti znotraj skupin W (glej razdelek 3.3):

$$P(\mathcal{C}) = sl(W)$$

Pa naj kdo reče, da ni v multivariatni analizi vse povezano med seboj!

V metodo prestavljanj je mogoče učinkovito vgraditi tudi kriterijske funkcije tipa maksimum (Ferligoj in Batagelj 1980):

$$P(\mathcal{C}) = \sum_{C \in \mathcal{C}} \max_{X,Y \in C} d(X,Y)$$

ali

$$P(\mathcal{C}) = \max_{C \in \mathcal{C}} \max_{X,Y \in C} d(X,Y)$$

Avtorji različic metode prestavljanj vanje vgrajujejo tudi druge bolj ali manj primerne kriterijske funkcije, na primer take, ki izhajajo iz matrik razpršenosti T, W in B.

Na začetku tega poglavja smo omenili, da z metodami lokalne optimizacije razvrščamo v vnaprej podano število skupin. Ponavadi ne vemo, katero število skupin se skriva v strukturi podatkov. Primerno število skupin lahko razberemo s pregledom nivojev združevanja pri hierarhičnem razvrščanju v skupine. Število skupin pa lahko določamo tudi s pregledom vrednosti kriterijskih funkcij dobljenih optimalnih razvrstitev pri različnem številu skupin: najprimernejše je tisto število skupin, pri katerem je največji padec vrednosti upoštevane kriterijske funkcije.

Kot primer poizkusimo razvrstiti republike in pokrajine v tri skupine glede na odstotek članov ZKJ in ZZBNOV metodo prestavljanj z Wardovo kriterijsko funkcijo, kjer so d evklidske razdalje med republikami in pokrajinami. Te so razvidne v tabeli 2.1. Vzemimo, da je začetna razvrstitev $\mathcal{C}_0$ podana z naslednjimi skupinami

$$C_1 = \{B, \check{C}, H\} \ , \ C_2 = \{M, S, O\} \ , \ C_3 = \{K, V\}$$

Vrednost Wardove kriterijske funkcije za to razvrstitev je:

$$P(\mathcal{C}_0) = \frac{3.1 + 2.1 + 2.3}{3} + \frac{2.3 + 2.0 + 1.8}{3} + \frac{2.7}{2} = 5.9$$

Naj bo sosedstvena struktura določena le s premiki posamezne enote iz ene skupine v drugo (in ne s premenami). Sosednjo razvrstitev tedaj dobimo, če na primer prestavimo B iz prve skupine v drugo ali tretjo. Če B prestavimo v drugo skupino, je vrednost kriterijske funkcije za tako dobljeno sosednjo razvrstitev $P(\mathcal{C}_1) = 5.5$. Če pa B prestavimo v tretjo skupino, je vrednost kriterijske funkcije $P(\mathcal{C}_2) = 5.1$. Torej v obeh primerih dobimo boljšo razvrstitev. Ker je $P(\mathcal{C}_2) < P(\mathcal{C}_1)$, premaknemo B v tretjo skupino. Prestavimo naslednjo enoto, to je Č, v drugo ali tretjo skupino in preverimo, če dobimo manjšo vrednost kriterijske funkcije. Izračun pokaže, da premik Č iz prve v drugo skupino zmanjša vrednost kriterijske funkcije, itd. Na ta način dobimo po sedmih korakih razvrstitev, katere sosednje razvrstitve ne dajo manjše vrednosti kriterijske funkcije. Ta razvrstitev $\mathcal{C}$ in njena vrednost kriterijske funkcije je naslednja:

$$\mathcal{C} = \{\{H, S\}, \{\check{C}, O, V\}, \{B, M, K\}\}$$

in

$$P(\mathcal{C}) = 2.2$$

Dobljena razvrstitev je 'stabilna', ker jo dobimo pri različno določenih začetnih razvrstitvah. Dobljeni rezultat je pravzaprav pričakovan, če se spomnimo naravne razvrstitve, ki smo jo prikazali v uvodnem poglavju na sliki 1.3.

5.2 Metoda voditeljev

Med nehierarhične metode lokalne optimizacije sodi tudi metoda voditeljev. Ta metoda je zelo popularna, ker zmore razvrščati v skupine večje število enot. Skoraj vsak statistični računalniški jezik, paket ali program, ki vključuje metode razvrščanja, vsebuje

tudi eno od različic metode voditeljev (več o tem v člankih Juga 1988, 1989).

Metoda voditeljev je iteracijska metoda, kjer se je potrebno odločiti, v koliko skupin razvrščamo enote. Postopek se začne z vnaprej podano množico predstavnikov posameznih skupin - voditeljev. Metoda priredi enote najbližjim voditeljem, poišče centroide (težišča) tako dobljenih skupin - nove voditelje, zopet priredi enote najbližjim voditeljem, itd. Postopek se konča, ko se nova množica voditeljev ne razlikuje od množice voditeljev, dobljene korak pred njo.

Osnovna shema metode voditeljev je tedaj:

določi začetno množico voditeljev $\mathcal{L} = \{L_i\}$

ponavljaj
določi razvrstitev $\mathcal{C}$ tako, da prirediš
vsako enoto njej najbližjemu voditelju
za vsako skupino $C_i \in \mathcal{C}$ izračunaj njeno središče $\bar{C}_i$
in ga določi za novega voditelja L_i skupine C_i
dokler se voditelji ne ustalijo

Ker je množica enot, ki jih razvrščamo, končna, je končna tudi množica vseh razvrstitev. Zato zgornji postopek prej ali slej skonvergira v lokalno optimalno rešitev. Tako kot pri ostalih metodah lokalne optimizacije, poskušamo tudi tu dobiti čim boljšo razvrstitev tako, da postopek ponovimo večkrat z različnimi začetnimi množicami voditeljev.

Začetno množico voditeljev lahko določimo na različne načine. Najpreprosteje je, da so določeni slučajno. Pogosto pa se poskuša število korakov v postopku metode voditeljev zmanjšati tako, da voditelje maksimalno razpršimo med proučevanimi enotami. To lahko storimo tako, da za prvega voditelja izberemo enoto, ki je

v središču vseh enot, za drugega najoddaljenejšo enoto od prvega voditelja, za tretjega najoddaljenejšo enoto od prvih dveh voditeljev itd. Najbolje pa je, da voditelje določimo na osnovi predhodno opravljene analize podatkov in domnev o strukturi proučevanih pojavov.

V vsakem koraku postopka metode voditeljev nas ponavadi zanima, kako dobro razvrstitev smo dobili. To merimo z ustrezno kriterijsko funkcijo. Zanimivo vprašanje je, ali izbrana kriterijska funkcija monotono pada z zaporedjem korakov postopka. Diday s sodelavci (npr. 1979) je med drugim pokazal, da Wardova kriterijska funkcija, kjer je d kvadrat evklidske razdalje, prav gotovo monotono pada. Zato je ponavadi v primeru razvrščanja številskih podatkov v skupine ob vsakem koraku postopka metode voditeljev izračunana prav ta kriterijska funkcija.

Z metodo voditeljev ponavadi iščemo popolno razvrstitev, ne pa nujno. Batagelj (1982, 1989) je na primer vgradil v svoj program CLUSE, ki ga uporabljamo za vse izračune v tej knjigi, tudi naslednjo možnost: naj bo podana množica k voditeljev in realno število $r > 0$. Pripadnost enote X posamezni skupini C_i je določena z naslednjim pogojem:

$$X \in C_i \Longleftrightarrow d(X, L_i) \leq r$$

To z drugimi besedami pomeni, da je enota, ki leži v okolici dveh voditeljev, element obeh pripadajočih skupin. Enota, ki ni v okolici nobenega voditelja, ni razvrščena.

Kot je že bilo rečeno, je poznanih veliko različic metode voditeljev. Tudi poimenovane so na različne načine: K-MEANS, metoda dinamičnih oblakov, razvrščanje k najbližjim središčem itd. Ker zahteva metoda voditeljev za večje količine podatkov relativno veliko računalniškega časa, nekateri programi izračunajo le posamezni korak v opisanemu postopku. Če uporabnik želi popraviti

množico voditeljev, mora na novo pognati ustrezni program. To je še posebej primerno, ko nimamo številskih podatkov, temveč na primer binarne, kjer različnost med enotami in voditelji merimo z eno od mer ujemanj (npr. Sokal-Michenerjeva mera). Med primeri, ki sledijo, je prikazana uporaba tudi take različice metode voditeljev.

Pokažimo na majhnem številu enot, kako deluje metoda voditeljev. V ta namen zopet vzemimo republike in pokrajini in jih razvrstimo v popolno razvrstitev s tremi skupinami glede na odstotek članov v ZKJ in ZZBNOV. Standardizirani podatki so podani v tabeli 1.2, evklidske razdalje pa v tabeli 2.1. Začetne voditelje določimo tako, da so karseda razpršeni v množici enot. Koordinati središča vseh enot (težišča) sta tedaj določeni s povprečnima vrednostima obeh spremenljivk. Ker sta spremenljivki standardizirani, sta ti koordinati enaki $(0, 0)$. Enota, ki je temu središču najbližja, je Hrvatska, ki jo proglasimo za prvega voditelja L_1. Drugi voditelj je najoddaljenejša enota od prvega voditelja. Pregled izračunanih evklidskih razdalj od prvega voditelja (Hrvatske) do preostalih enot kaže, da je najoddaljenejša enota Črna gora ($d(H, \check{C}) = 2.3$), zato jo postavimo za drugega voditelja L_2. Tretji voditelj je najoddaljenejša enota od obeh že določenih voditeljev. Pregled evklidskih razdalj pokaže, da je tedaj tretji voditelj Kosovo. V naslednjem koraku priredimo preostale enote k najbližjemu voditelju. Dobljena razvrstitev je tedaj:

$$\mathcal{C} = \{\{H, S\}, \{\check{C}, O, V\}, \{K, M, B\}\}$$

Že v prvem koraku smo torej dobili optimalno razvrstitev.

5.3 Primeri uporabe metod lokalne optimizacije

5.3.1 Aktivnosti v prostem času

Za prvi primer ponovno razvrščajmo prostočasne aktivnosti v skupine, tokrat z metodo prestavljanj. Pred uporabo metode prestavljanj se moramo odločiti za število skupin, ki je kar najbolj značilno za dane podatke. Eden izmed načinov, kako analitično določiti primerno število skupin, je pregled nivojev združevanja pri hierarhičnem razvrščanju v skupine. V primeru hierarhičnega združevanja prostočasnih aktivnosti po Wardovi metodi (glej razdelek 4.9.1.) pregled drevesa združevanja daje slutiti, da se aktivnosti naravno razvrščajo v dve skupini. Skoki nivojev z manjšimi vrednostmi pa niso več tako izraziti - morda nakazujejo šest skupin. Zato v tem primeru razvrstimo prostočasne aktivnosti z Wardovo metodo lokalne optimizacije najprej v dve, nato pa še v šest skupin. Tudi v tem primeru uporabimo za mero podobnosti Jaccardov koeficient in transformacijo $d = 1 - J$.

Omenili smo, da dobimo z metodo prestavljanj razvrstitev, ki je v splošnem le lokalni minimum kriterijske funkcije za dano sosedstveno strukturo. Da bi dobili čim boljšo razvrstitev (po možnosti globalni minimum funkcije), postopek ponovimo z različnimi začetnimi razvrstitvami. V primeru razvrščanja prostočasnih aktivnosti v dve skupini smo z Wardovo metodo prestavljanj poiskali najboljšo razvrstitev pri dvajsetih različnih slučajno določenih začetnih razvrstitvah, pri čemer smo devetnajstkrat dobili enako končno razvrstitev z vrednostjo kriterijske funkcije $P(\mathcal{C}) = 8.90$ in le eno razvrstitev, ki je imela večjo vrednost funkcije (9.11). Zelo verjetno je dobljena vrednost globalni ekstrem funkcije. Ta rezultat potrjuje, da se obravnavane prostočasne aktivnosti 'narav-

C_1	C_2
Zoga	Plava
Umkre	Smuca
Problem	Drsa
Obisk	Tenis
Gore	Gleda
Umetkr	Kino
Ostalkr	Bere
Csport	Glasba
Tabor	Igra
Ostalosp	Ples
	Sprehod
	Jezikkr
	Rocdel

Tabela 5.1: Prostočasne aktivnosti v dveh skupinah

no' razvrščajo v dve skupini. Najboljša končna razvrstitev prostočasnih aktivnosti, dobljena z Wardovo metodo prestavljanj v dve skupini je podana v tabeli 5.1.

Ta razvrstitev prostočasnih aktivnosti je enaka razvrstitvi, ki smo jo dobili z Wardovo metodo hierarhičnega združevanja v skupine (zahtevnejše intelektualne aktivnosti z aktivno rekreacijo in aktivnosti družabno zabavnega značaja), kar zopet potrjuje, da sta metodi razkrili izrazito 'naravno' strukturo podatkov.

Razvrstimo prostočasne aktivnosti še v šest skupin. Tudi v tem primeru poiščimo najboljšo rešitev pri dvajsetih različnih slučajno določenih začetnih razvrstitvah. V tem primeru smo dobili najmanjšo vrednost Wardove kriterijske funkcije le trikrat. Prostoča-

C_1	C_2	C_3	C_4	C_5	C_6
Smuca	Drsa	Plava	Zoga	Gore	Problem
Tenis	Ples	Umkre	Csport	Tabor	Ostalkr
Gleda		Obisk	Ostalosp		
Kino		Umetkr			
Bere					
Glasba					
Igra					
Sprehod					
Jezikkr					
Rocdel					

Tabela 5.2: Prostočasne aktivnosti v šestih skupinah

sne aktivnosti se torej slabše razvrščajo v šest skupin kot v dve skupini, kar daje slutiti tudi pregled drevesa združevanja. Najboljša dobljena končna razvrstitev prostočasnih aktivnosti v šest skupin ($P(\mathcal{C}) = 6.64$) je podana v tabeli 5.2. Če primerjamo dobljeno razvrstitev z dvema skupinama z razvrstitvijo s šestimi skupinami, vidimo, da se druga skupina (družabno zabavne aktivnosti) skoraj ne spremeni: od trinajstih aktivnosti se iz skupine izločijo le tri aktivnosti PLAVA, DRSA in PLES, od katerih se DRSA in PLES združita v drugo skupino, PLAVA pa se priključi k umetniškim dejavnostim (tretja skupina). Prva skupina pa se razcepi na štiri homogene skupine in sicer: umetniške dejavnosti (tretja skupina), športne aktivnosti (četrta skupina), planinarjenje (peta skupina) in zahtevnejše intelektualne aktivnosti (šesta skupina).

Doslej smo prostočasne aktivnosti razvrščali v skupine z metodami hierarhičnega združevanja in metodo prestavljanj. Poglej-

mo, kaj nam lahko razkrije uporaba poenostavljene neiterativne metode voditeljev. Na osnovi že opravljene analize prostočasnih aktivnosti lahko določimo množico voditeljev, ki jih dodamo množici enot, in nato razvrstimo dijake k najbližjim voditeljem. Podobnost med dijaki in voditelji naj bo merjena s Sokal-Michenerjevo mero ujemanja. Iskana razvrstitev naj bo popolna.

Iz drevesa združevanja, dobljenega z Wardovo metodo hierarhičnega združevanja, je mogoče razbrati pet skupin prostočasnih aktivnosti, ki nastopajo skupaj (to pomeni: če se dijak ukvarja z eno od teh aktivnosti se zelo verjetno ukvarja tudi z drugimi aktivnostmi iz skupine). Te smo ponovno preverjali z metodo prestavljanj. Glede na opravljene analize in postavljene domneve o tipologiji prostočasnih aktivnosti navedimo naslednje tipe aktivnosti:

1. neselektivne aktivnosti: smučanje, kino, branje, ročna dela
2. glasbene aktivnosti: igranje inštrumentov, poslušanje glasbe, sodelovanje v jezikovnih krožkih
3. umetniške aktivnosti: umetniška kreacija, sodelovanje v ustreznih krožkih, obiskovanje prijateljev
4. športne aktivnosti: igranje z žogo in ostale fizične aktivnosti, član športnega društva
5. naravoslovne aktivnosti: planinarjenje, član Zveze taborni-kov ali Planinskega društva, zahtevne mentalne aktivnosti in član ustreznih krožkov

Vsakemu tipu aktivnosti priredimo njegovega predstavnika - voditelja. Ob pripisovanju prostočasnih aktivnosti zgoraj opisanim petim voditeljem moramo še upoštevati, da se lahko nekatere aktivnosti pojavljajo pri več voditeljih (na primer branje knjig), kar

pa ni razvidno iz drevesa združevanja ali iz popolne razvrstitve, dobljene z metodo prestavljanj, kjer je vsaka lastnost natanko v eni skupini. Zato smo v tem smislu določili nekaj množic petih voditeljev in med njimi izbrali najboljšo, to je z najmanjšo vrednostjo kriterijske funkcije metode voditeljev

$$P(\mathcal{C}) = \sum_{C \in \mathcal{C}} \sum_{X \in C} d(X, L_C)$$

kjer je L_C voditelj skupine C in $d = 1 - S$, kjer je S Sokal-Michenerjeva mera ujemanja. Na osnovi pregleda dobljenih razvrstitev smo še spreminjali nekatere lastnosti na posameznih voditeljih, da bi še zmanjšali vrednost kriterijske funkcije. Najboljša dobljena ravrstitev je imela vrednost funkcije $P(\mathcal{C}) = 18.9$. Lastnosti voditeljev in dobljenih skupin te razvrstitve so podane v tabeli 5.3.

V tabeli so v drugem stolpcu (f_i) zapisane frekvence pozitivnih vrednosti na posameznih spremenljivkah (aktivnostih). Tako je na primer 27 dijakov navedlo aktivnost igranje z žogo. V tem stolpcu so ponekod prištete enice oziroma trojka, ker smo k množici dijakov dodali postavljene (umetne) voditelje. Pod posameznimi voditelji (oziroma skupinami) so zapisane frekvence pojavljanja posameznih aktivnosti. Za posameznega voditelja so bile vnaprej navedene tiste aktivnosti, katerih frekvenca je v oklepaju. Iz tabele je razvidno, da je na primer aktivnost branje knjig tipična kar za tri skupine: prvo ('neselektivci'), drugo ('naravoslovci') in četrto ('glasbeniki'). Nekatere lastnosti pa niso tipične za nobeno skupino (na primer plavanje, drsanje, tenis itd.). Predzadnji vrstici povesta, kako močna je posamezna skupina. Največ dijakov (40.5 odstotkov) je navedlo 'neselektivne' aktivnosti, 27.2 odstotkov dijakov je 'športnikov', 17.2 odstotkov dijakov je 'glasbenikov', 13.8 je 'naravoslovcev' in le 5.2 odstotkov 'umetnikov'. V zadnji

aktivnosti	f_i	L_1	L_2	L_3	L_4	L_5
ZOGA	27+1	1	3	(21)	1	2
PLAVA	14	9	0	3	1	1
SMUCA	48+1	(33)	5	8	1	2
DRSA	24	15	1	6	2	0
TENIS	23	10	2	7	4	0
UMKRE	19+1	6	2	3	3	(6)
GLEDA	13	9	1	0	3	0
KINO	25+1	(22)	2	2	0	0
BERE	69+3	(39)	(12)	4	(16)	1
GLASBA	38+1	15	4	5	(14)	1
IGRA	25+1	8	1	2	(15)	0
PLES	13	8	0	2	3	0
PROBLEM	13+1	0	(8)	3	3	0
SPREHOD	16	11	0	2	3	0
OBISK	11+1	1	1	3	3	(4)
GORE	22+1	1	(12)	7	3	0
UMETKR	14+1	3	1	2	4	(5)
JEZIKKR	26+1	10	2	2	(13)	0
OSTALKR	21+1	6	(9)	5	2	0
CSPORT	28+1	5	1	(20)	0	3
TABOR	18+1	2	(9)	4	3	1
ROCDEL	35+1	(30)	1	2	3	0
OSTALSP	21+1	3	0	(17)	2	0
n_i		46+1	15+1	31+1	19+1	5+1
r_i		.30	.30	.35	.30	.22

Tabela 5.3: Uvrstitev dijakov k petim voditeljem

vrstici so podane razdalje med voditeljem in v njegovi skupini najoddaljenejšo enoto. Ta mera torej podaja homogenost oziroma heterogenost skupine. Iz podanih vrednosti na primer sledi, da je najhomogenejša skupina 'umetnikov' in najheterogenejša skupina 'športnikov'.

5.3.2 Evropske države glede na razvojne kazalce

Evropske države smo v razdelku 4.9.2. razvrščali z dvema metodama hierarhičnega združevanja: minimalno in Wardovo. Drevo združevanja, dobljeno z Wardovo metodo , kaže izrazito razvrstitev v dve skupini in malo manj izrazito v pet skupin.

Da bi dobili z metodo prestavljanj čim boljšo razvrstitev v dve skupini (po možnosti globalni minimum kriterijske funkcije), smo tudi v tem primeru postopek ponovili z desetimi različnimi slučajno določenimi začetnimi razvrstitvami. Z Wardovo metodo prestavljanj, kjer je mera različnosti določena tako kot v primeru hierarhičnega združevanja, smo v vseh desetih primerih dobili enako končno razvrstitev z vrednostjo kriterijske funkcije $P(\mathcal{C}) = 4.85$. Dobljeni rezultat torej potrjuje, da se evropske države glede na merjene razvojne kazalce izrazito (zelo stabilno) razvrščajo v dve skupini. Dobljena končna razvrstitev evropskih držav v dve skupini je podana v tabeli 5.4.

Z Wardovo metodo prestavljanj smo torej dobili skupino bolj razvitih evropskih držav (prva skupina) in skupino manj razvitih evropskih držav (druga skupina), ki pa se nekoliko razlikujeta od razvrstitve v dve skupini, dobljene z Wardovo metodo hierarhičnega združevanja. Tako je na primer Avstrija v drevesu združevanja v skupini manj razvitih držav, v razvrstitvi, dobljeni z metodo prestavljanj, pa v skupini bolj razvitih. Metode hierarhičnega združevanja namreč v posameznih korakih združevanja

C_1	C_2
Avstrija	Albanija
Belgija	Bolgarija
Danska	ČSSR
Finska	Grčija
Francija	Madžarska
NDR	Irska
ZRN	Poljska
Island	Portugalska
Italija	Romunija
Luksemburg	Španija
Nizozemska	ZSSR
Norveška	Jugoslavija
Švedska	
Švica	
V.Britanija	

Tabela 5.4: Evropske države v dveh skupinah

C_1	C_2	C_3	C_4	C_5
Albanija	Grčija	Avstrija	Belgija	Danska
Bolgarija	Irska	ČSSR	ZRN	Finska
Poljska	Portugalska	NDR	Italija	Francija
Romunija	Španija	Madžarska	Luksemburg	Island
ZSSR	Jugoslavija		Nizozemska	Norveška
			Švica	Švedska
			V. Britanija	

Tabela 5.5: Evropske države v petih skupinah

upoštevajo podobnosti le tistih skupin, ki jih v danem koraku združujejo ('požrešna' hevristika), medtem ko metode z lokalno optimizacijo za razvrščanje uporabljajo vso informacijo o podobnostih med enotami. Zato so metode hierarhičnega združevanja zelo primerne tedaj, ko še nimamo postavljene domneve o številu skupin (prav s pregledom nivojev združevanja lahko določimo primerno število skupin). Če pa vemo, v koliko skupin želimo razvrščati, so primernejše metode lokalne optimizacije.

Razvrstimo evropske države glede na razvojne kazalce še v pet skupin. Tudi v tem primeru poiščimo najboljšo razvrstitev pri desetih slučajno določenih začetnih razvrstitvah. Pri štirih začetnih razvrstitvah smo dobili enako najmanjšo vrednost Wardove kriterijske funkcije, ki je zelo verjetno globalni ekstrem. Ta rezultat pomeni, da je naravna razvrstitev v pet skupin sicer šibkejša kot v dve skupini, vendar kar izrazita. Najboljša dobljena končna razvrstitev evropskih držav v pet skupin ($P(\mathcal{C}) = 3.14$) je podana v tabeli 5.5.

Skupina bolj razvitih evropskih držav se torej razbije na dve

Bosna in Hercegovina	106	občin
Črna gora	20	”
Hrvaška	105	”
Makedonija	30	”
Slovenija	60	”
Ožja Srbija	113	”
Vojvodina	44	”
Kosovo	22	”

Tabela 5.6: Število občin v posamezni republiki in pokrajini

skupini: skupino zahodnih držav (četrta skupina) in 'skandinavsko' skupino (peta skupina). Skupina manj razvitih držav se tudi razcepi na dve skupini: skupino vzhodnih držav in 'sredozemsko' skupino. Zanimiva je tretja skupina 'avstro-ogrskih' držav, ki je sestavljena iz dveh držav iz skupine bolj razvitih držav (Avstrija in NDR) in iz dveh držav iz skupine manj razvitih držav (Češkoslovaška in Madžarska). To skupino smo že videli v Wardovem drevesu združevanja.

5.3.3 Jugoslovanske občine glede na stanovanjski standard

Kot primer razvrščanja večjega števila enot v skupine z metodo voditeljev vzemimo občine SFR Jugoslavije in njihov stanovanjski standard. Potrebne podatke smo vzeli iz Popisa stanovanj v SFRJ, ki je bil opravljen 31. marca 1971. V tem popisu je zajetih 500 občin. V tabeli 5.6 je podano število občin v posamezni republiki in pokrajini.

Izbrali smo enajst spremenljivk, ki merijo stanovanjski stan-

dard. Opis teh spremenljivk je podan v tabeli 5.7. Prve štiri spremenljivke merijo starost stanovanj, ostalih sedem pa meri kvaliteto stanovanj (opremeljenost stanovanj z napeljavami, kopalnico ter vrsta materiala, iz katerega so stanovanja zgrajena). Vse spremenljivke smo pred razvrščanjem v skupine običajno standardizirali, tako da so vrednosti različnih spremenljivk na posameznih občinah primerljive med seboj.

Tipologijo jugoslovanskih občin glede na njihov stanovanjski standard smo poiskali z metodo voditeljev tako, da smo različnost med enotami in voditelji merili z evklidsko razdaljo. Razvrščali smo v popolne razvrstitve z dvema, tremi, štirimi in petimi skupinami. Začetni voditelji so bili v vseh štirih primerih slučajno izbrani. Vsakič je postopek metode voditeljev skonvergiral v manj kot enajstih korakih. V tabelah 5.8, 5.9 in 5.10 so prikazani voditelji dobljenih najboljših razvrstitev občin z ustreznimi lastnostmi posameznih skupin in razvrstitev.

Iz tabele 5.8 lahko razberemo, da imata voditelja razvrstitve z dvema skupinama naslednje lastnosti: vrednosti spremenljivk na prvem voditelju kažejo, da so stanovanja v občinah, ki se vežejo na tega voditelja, zgrajena pretežno po I. svetovni vojni, da so najverjetneje brez električne napeljave in vodovoda ali le z električno napeljavo, brez kopalnice in z zemeljskim podom, stanovanja so podpovprečno grajena iz trdih materialov. Lastnosti drugega voditelja pa kažejo ravno nasprotno: večji delež stanovanj v občinah, ki pripadajo temu voditelju, je zgrajenih pred I. svetovno vojno, nadpovprečni delež stanovanj ima električno napeljavo in vodovod ter kopalnico, večinoma so grajena iz trdega materiala in nimajo zemeljskega poda. Torej prvi voditelj je predstavnik občin z relativno slabšim, drugi pa z relativno boljšim stanovanjskim standardom.

Za razumevanje dobljenih rezultatov so zanimive tudi stati-

OZNAKA	OPIS SPREMENLJIVK
do 1918	št. stanovanj zgrajenih do leta 1918 na 100 stanovanj
1919-1945	št. stanovanj zgrajenih od leta 1919 do leta 1945 na 100 stanovanj
1946-1960	št. stanovanj zgrajenih od leta 1946 do leta 1960 na 100 stanovanj
po 1960	št. stanovanj zgrajenih po letu 1960 na 100 stanovanj
EL in VO	št. stanovanj z električno napeljavo in vodovodom na 100 stanovanj
EL	št. stanovanj z le električno napeljavo na 100 stanovanj
brez El.VO	št. stanovanj brez električne napeljave in vodovoda na 100 stanovanj
KOPALNICA	št. stanovanj s kopalnico na 100 stanovanj
TRD MAT	št. stanovanj iz trdega materiala na 100 stanovanj
ODPRTO OG	št. stanovanj z odprtim ognjiščem na 100 stanovanj
ZEM POD	št. stanovanj z zemeljskim podom na 100 stanovanj

Tabela 5.7: Spremenljivke, ki merijo stanovanjski standard občin

OZNAKA	L_1	L_2	L_1	L_2	L_3
do 1918	-0.34	1.03	-0.35	1.03	0.81
1919-1945	0.18	-0.56	0.19	-0.52	-0.60
1946-1960	0.28	-0.84	0.28	-0.83	-0.68
po 1960	0.15	-0.44	0.15	-0.45	-0.33
EL in VO	-0.48	1.45	-0.47	1.56	0.05
EL	0.36	-1.10	0.37	-1.15	-0.44
brez El.VO	0.21	-0.64	0.20	-0.73	0.47
KOPALNICA	-0.45	1.35	-0.44	1.45	0.11
TRD MAT	-0.32	0.97	-0.33	0.94	1.02
ODPRTO OG	-0.17	0.50	-0.21	0.03	4.50
ZEM POD	0.28	-0.84	0.29	-0.85	-0.65
n_i	376	124	371	112	17
$r_{i,max}$	6.1	8.9	5.5	6.1	5.6
$\bar{r}_i$	2.5	2.8	2.5	2.4	3.2
$P(\mathcal{C})$		1293			1250

Tabela 5.8: Razvrstitvi občin v dve in tri skupine, dobljeni z metodo voditeljev

OZNAKA	L_1	L_2	L_3	L_4
do 1918	-0.75	0.92	1.40	-0.05
1919-1945	-0.51	-0.46	-0.68	0.66
1946-1960	0.91	-0.79	-1.09	-0.17
po 1960	0.73	-0.34	-0.83	-0.26
EL in VO	-0.38	1.75	0.31	-0.52
EL	-0.20	-1.31	-0.29	0.76
brez El.VO	0.78	-0.79	-0.07	-0.22
KOPALNICA	-0.30	1.63	0.27	-0.53
TRD MAT	-0.03	0.94	1.15	-0.55
ODPRTO OG	-0.21	-0.16	3.02	-0.21
ZEM POD	-0.45	-0.85	-0.77	0.81
n_i	158	94	31	217
$r_{i,max}$	5.6	5.9	6.9	5.6
$\bar{r_i}$	2.2	2.2	3.1	2.1
$P(\mathcal{C})$				1105

Tabela 5.9: Razvrstitev občin v štiri skupine, dobljena z metodo voditeljev

OZNAKA	L_1	L_2	L_3	L_4	L_5
do 1918	-0.78	0.94	1.39	0.07	-0.51
1919-1945	-0.38	-0.46	-0.66	0.82	-0.12
1946-1960	1.17	-0.80	-1.07	-0.30	0.36
po 1960	0.46	-0.36	-0.84	-0.46	0.56
EL in VO	-0.62	1.77	0.29	-0.60	-0.26
EL	-0.73	-1.34	-0.27	0.76	0.54
brez El.VO	1.76	-0.79	-0.06	-0.13	-0.31
KOPALNICA	-0.55	1.65	0.25	-0.60	-0.21
TRD MAT	-0.19	0.94	1.12	-0.98	0.18
ODPRTO OG	-0.17	-0.16	2.88	-0.19	-0.26
ZEM POD	-0.37	-0.85	-0.76	1.19	-0.20
n_i	80	92	33	139	156
$r_{i,max}$	5.1	5.9	6.8	4.0	4.6
$\bar{r}_i$	2.2	2.2	3.1	2.0	1.8
$P(\mathcal{C})$					1039

Tabela 5.10: Razvrstitev občin v pet skupin, dobljena z metodo voditeljev

stične karakteristike, ki so zapisane v spodnjem delu tabel: n_i označuje število enot v skupini C_i, $r_{i,max}$ največjo razdaljo med voditeljem L_i in enotami v skupini C_i, $\bar{r}_i$ povprečno razdaljo med voditeljem L_i in enotami v skupini C_i in $P(\mathcal{C})$ vrednost Wardove kriterijske funkcije v posamezni razvrstitvi.

V skupini občin z relativno slabšim stanovansjkim standardom je precej več občin (376) kot v skupini občin z boljšim stanovanjskim standardom (124). Obe razdalji med enotami skupine in njihovim voditeljem merita, koliko se enote v skupini razlikujejo med seboj. Podatki kažejo, da je skupina z relativno slabšim standardom homogenejša od tiste z boljšim standardom.

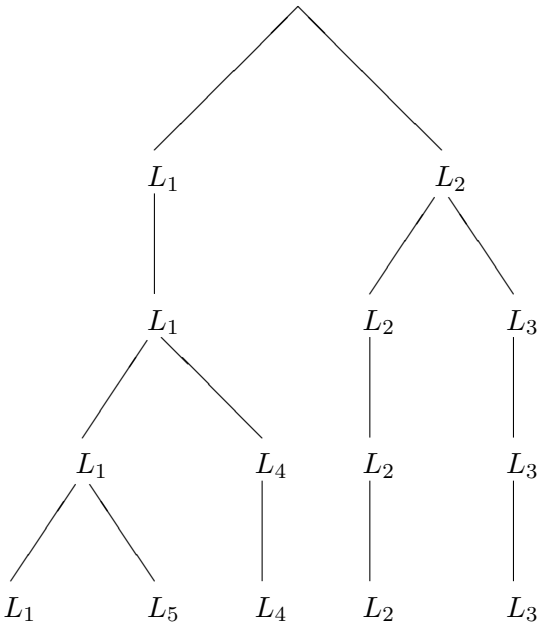
Dobljeni voditelji razvrstitve občin s tremi skupinami kažejo, da se prvi voditelj skoraj ne razlikuje od prvega voditelja razvrstitve z dvema skupinama. Razcepila se je torej heterogenejša skupina z manjšim številom občin in sicer v skupino z izrazito boljšim stanovanjskim standardom, ki jo sestavljajo vse občine SR Slovenije razen Lendave, stari centri drugih republik in pokrajin in dalmatinske občine z večjimi starimi mesti. Tretji voditelj pa kaže, da je v občinah, ki se vežejo nanj, nadpovprečni delež stanovanj, zgrajenih pred I. svetovno vojno, tako kot v drugi skupini, vendar je njihov standard precej nižji. To skupino občin sestavljajo pretežno občine s manjšimi starimi mesti v Istri in Dalmaciji.

Voditelji razvrstitve s štirimi skupinami (tabela 5.9) pa razkrivajo, da se je prva skupina iz razvrstitve s tremi skupinami razcepila na dve skupini z voditeljema L_1 in L_4. Prva skupina občin ima relativno veliko stanovanj, zgrajenih po II. svetovni vojni, vendar z zelo slabim stanovansjkim standardom. Skupino sestavlja 158 občin, večinoma iz Bosne in Hercegovine. Četrto skupino pa sestavljajo občine s stanovanji, ki so večji del zgrajena med obema vojnama. Nadpovprečni delež stanovanj v teh občinah ima le električno napeljavo in zemeljski pod. V tej skupini so vse občine

Vojvodine razen Novega Sada, kmečke občine Kosova, ožje Srbije, Makedonije in Hrvaške. Ta skupina je najštevilnejša, a tudi najhomogenejša.

Razvrstitev s petimi skupinami (tabela 5.10) se od razvrstitve s štirimi razlikuje v tem, da se je v slednji razvrstitvi cepila prva skupina v skupini C_1 in C_5. Voditelj prve skupine kaže še izrazitejše lastnosti kot prvi voditelji doslej. Stanovanja v občinah, ki se vežejo nanj, so pretežno zgrajena po II. svetovni vojni in imajo relativno zelo nizek stanovanjski standard. Največ stanovanj v občinah pete skupine pa je zgrajenih v šestdesetih letih in so relativno kvalitetnejša.

Na sliki 5.1 je povzeta razprava dobljenih skupin in njihovih voditeljev tako, da je nakazana drevesna struktura dobljenih voditeljev razvrstitev z dvema, tremi, štirimi in petimi skupinami.



Slika 5.1: Drevesna razvrstitev voditeljev

6.

Razvrščanje v skupine z omejitvami

6.1 Uvod

Na razvojni poti teorije razvrščanja v skupine so se izoblikovali različni tipi problemov razvrščanja v skupine. Eden izmed teh je problem razvrščanja v skupine z omejitvami. V tem primeru gre za razvrščanje podobnih enot v skupine glede na izbrane karakteristike tako, da mora iskana razvrstitev zadoščati še nekaterim drugim pogojem. Tudi ta problem je relativno star. Eden izmed najpogosteje reševanih konkretnih primerov s tega področja je regionalizacija: potrebno je poiskati skupine podobnih geografskih območij glede na izbrane karakteristike, pri čemer morajo biti območja, ki sestavljajo skupino, geografsko sosednja. V literaturi se je v zadnjem času pojavilo več pristopov za analitično določitev regionalizacije. Webster in Burrough (1972) in kasneje Perruchet (1983) so problem reševali tako, da so priredili običajne metode razvrščanja v skupine, katerih osnova je matrika različnosti. V ma-

triki različnosti so meri različnosti med enotama, ki nista geografsko sosedni, pripisali tako veliko vrednost, da so s tem preprečili združitev teh dveh enot v skupino. Večina avtorjev (Spence 1968; Webster 1973, 1978; Thauront 1976; Openshaw 1977; Lebart 1978; Lefkovitch 1980, Fischer 1980; Ferligoj in Batagelj 1982; Gordon 1987) pa je reševala problem regionalizacije tako, da so priredili običajne metode razvrščanja v skupine (npr. metode hierarhičnega združevanja v skupine in metode lokalne optimizacije) tako, da ob določitvi razvrstitve preverjajo, ali enote, ki sestavljajo posamezno skupino, zadoščajo dodatni zahtevi po geografski sosednosti.

Reševani pa so bili tudi drugačni problemi razvrščanja v skupine s podobnimi omejitvami, ki jih sicer v literaturi imenujemo relacijske omejitve. Geografsko sosednost med območji namreč lahko popišemo z relacijo R takole:

$$X_i \; R \; X_j \equiv enota \;\; X_i \;\; je \;\; geografsko \;\; sosedna \;\; z \;\; enoto \;\; X_j$$

Relacija R je lahko določena tudi drugače, na primer s časovno povezanostjo (enoti sta v relaciji, če sta iz sosednjih časovnih točk: X_t z X_{t-1} ali X_{t+1}). Posebne primere razvrščanja z relacijsko omejitvijo, ki pa niso primeri regionalizacije, sta reševala Gordon (1973, 1980) in Ivanović (1981). Z Batageljem sva najprej reševala problem razvrščanja v skupine z relacijsko omejitvijo za splošno simetrično (1982) nato pa tudi nesimetrično relacijo (1983) (primer nesimetrične relacije je na primer prijateljstvo). Izčrpen pregled rezultatov razvrščanja v skupine s simetrično relacijsko omejitvijo je podal Murtagh (1985 b).

V literaturi je mogoče najti tudi drugačne, nerelacijske dodatne pogoje-omejitve pri razvrščanju v skupine. Tak primer je omejitev funkcije vrednosti dane (omejevalne) spremenljivke v posamezni skupini. Primer take omejitve na spremenljivki pri regionalizaciji je zahteva, da je število prebivalcev (omejevalna spremenljivka) v

skupini nad neko dano mejo. Razvrščanje v skupine z omejevalno spremenljivko je obravnavalo že več avtorjev (npr. Mills 1967; Openshaw 1977; DeSarbo in Mahajan 1984; Ferligoj 1981, 1986).

Tretja skupina omejitev, ki je bila deležna posebne obravnave, pa je optimizacijska zahteva ob razvrščanju v skupine. Če nadaljujemo s primerom regionalizacije, je taka optimizacijska omejitev zahteva, da so centri geografskih območij v skupini čimbolj med seboj povezani oziroma konkretno, da je vsota dolžin cestnih povezav med centri območij kar se da majhna. Doslej so bili taki problemi razvrščanja v skupine reševani predvsem z metodo pragov (Lefkovitch 1985; Ferligoj in Lapajne 1987). Morda pa je najobetavnejši pristop k reševanju problemov razvrščanja v skupine z optimizacijskimi omejitvami večkriterijsko razvrščanje v skupine, ki je obravnavano v naslednjem poglavju te knjige.

Izkazalo se je, da optimizacijski pristop k razvrščanju v skupine (Batagelj 1979; Ferligoj in Batagelj 1982) omogoča enotno obravnavo vseh omenjenih tipov omejitev. V nadaljevanju tega poglavja se bomo pomudili ob splošnem reševanju problema razvrščanja v skupine z omejitvami, zanj priredili metode hierarhičnega združevanja v skupine in metodo prestavljanj ter na nekaj primerih pokazali, kako se obravnavane metode lahko uporablja v konkretnih primerih. Pokazali bomo tudi možne smeri razvoja tega področja.

V tem poglavju si precej pomagamo s teorijo grafov. Da bi bila čim bolj razumljiva, se omejujem le na najpomembnejše rezultate s tega področja. Če bralec potrebuje ali želi podrobnejše spoznati teorijo grafov, je na voljo več ustreznih učbenikov tudi v slovenskem jeziku (npr. Bajc in Pisanski 1985). Podrobnejša razlaga snovi, ki je zajeta v tem poglavju, in dokazi omenjenih izrekov pa so v moji doktorski disertaciji (Ferligoj 1983).

6.2 Problem razvrščanja v skupine z omejitvami

Z optimizacijskim pristopom k razvrščanju v skupine postavimo problem razvrščanja v skupine z omejitvami takole:

Določiti želimo razvrstitev $\mathcal{C}^*$ tako, da bo vrednost kriterijske funkcije P za to razvrstitev najmanjša, če pregledamo vse razvrstitve v množici dopustnih razvrstitev $\mathcal{C} \in \Phi$, ki je določena s pogoji - omejitvami, ki niso zajeti v kriterijski funkciji:

$$P(\mathcal{C}^*) = \min_{\mathcal{C} \in \Phi} P(\mathcal{C})$$

V nadaljevanju je obravnavan vsak od že omenjenih treh tipov omejitev posebej.

6.2.1 Splošna relacijska omejitev

Prvi tip omejitev je relacijska omejitev in ta v splošnem določa naslednje množice razvrstitev:

$$\begin{aligned}\Phi(R) = \quad & \{\mathcal{C} : \text{je razbitje množice enot E in} \\ & \text{vsaka skupina } C \in \mathcal{C} \text{ poraja v grafu } (E, R) \\ & \text{'dogovorjeno povezan' podgraf } (C\ ,\ R \cap C \times C)\}\end{aligned}$$

Z isto relacijo lahko določimo različne tipe množic dopustnih razvrstitev. V primeru simetrične relacije (če je XRY, potem je tudi YRX) lahko zahtevamo, da je vsaka enota v skupini v relaciji z vsako enoto v skupini (v jeziku teorije grafov je taka skupina enot klika). V primeru regionalizacije to pomeni, da vsako območje v posamezni regiji meji z vsakim drugim območjem v tej regiji. Lahko pa postavimo običajno (šibkejšo) zahtevo, ki v primeru regionalizacije pomeni, da posamezno regijo (skupino območij) sestavljajo območja, ki tvorijo geografsko povezano ozemlje, kjer je

možno priti iz vsakega območja posamezne regije v vsako območje te regije tako, da se ne stopi na ozemlje druge regije. V jeziku teorije grafov rečemo, da so enote (območja) v vsaki skupini (regiji) povezane s potmi, ki v celoti ležijo znotraj skupine glede na graf, ki je določen z množico enot E in simetrično relacijo R (geografsko sosednostjo).

Če relacija R ni simetrična, lahko z isto relacijo R določimo še pestrejše tipe množic dopustnih razvrstitev (Ferligoj in Batagelj, 1983). Denimo, da imamo skupino učencev v nekem šolskem razredu. Naj relacija R tedaj popisuje prijateljstvo med učenci: učenec X_i je v relaciji R z učencem X_j, če X_i meni, da je X_j njegov prijatelj. Glede na tako definirano relacijo lahko določimo skupine v razvrstitvi $\mathcal{C} \in \Phi(R)$ takole: v skupini so učenci, ki so vsi med seboj prijatelji (klika). Ali: v skupini naj bodo učenci, kjer je vsak posamezni učenec preko prijateljskih vezi z drugimi učenci povezan z vsakim učencem v skupini. Oziroma natančneje: iz vsake enote (učenca) v določeni skupini lahko pridemo po poti, ki jo določa relacija (prijateljstvo), v vsako enoto te skupine. V tem primeru gre za krepko povezanost. Če pri tem ne upoštevamo smeri relacije, pa dobimo šibko povezanost. Povsem jasno je, da je prvi tip povezanosti (klika) veliko strožji kot preostali tipi povezanosti. Poznanih je še veliko drugačnih tipov povezanosti. Omenimo še enega v primeru prijateljske strukture učencev: učenci naj bodo šibko pvezani med seboj z dodatnim pogojem, da je v vsaki skupini učenec (ali skupina krepko povezanih učencev), za katerega vsi v skupini menijo, da je njihov prijatelj, ali, da od posameznega učenca vodi prijateljska veriga do njega. Tega učenca (ali skupino učencev) lahko zaradi te vloge imenujemo vodjo ali center.

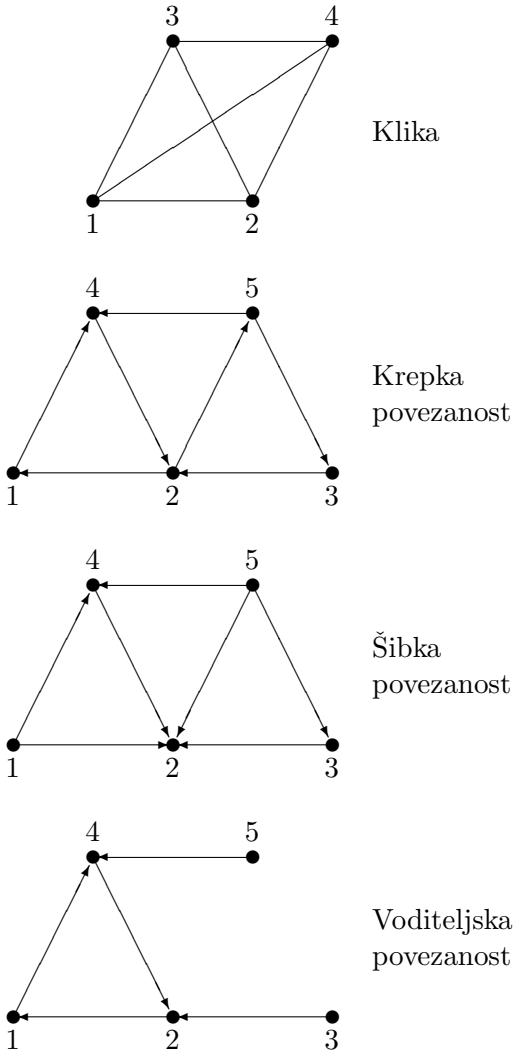
Pri analizi relacij je lahko v veliko pomoč grafična ponazoritev relacije R na množici enot E. Urejen par (E, R) določa graf, ki

tip razvrstitev	tip povezanosti podgrafa
$\Phi^1(R)$	šibka povezanost
$\Phi^2(R)$	šibka povezanost, v vsaki skupini je natanko en center
$\Phi^3(R)$	krepka povezanost
$\Phi^4(R)$	klika
$\Phi^5(R)$	obstaja enostaven sprehod, ki vsebuje vse enote skupine

Tabela 6.1: Nekaj množic dopustnih razvrstitev, določenih z relacijsko omejitvijo R

ga lahko tudi narišemo: vsaki posamezni enoti v ravnini priredimo krožec in za vsak par enot (X, Y), kjer je $X\ R\ Y$, povežemo ustrezni enoti (krožca) z usmerjeno črto od X do Y. Na sliki 6.1 so grafično prikazani štirje omenjeni tipi povezav: klika, krepka, šibka in voditeljska povezanost. V primeru klike so prikazane štiri enote, kjer je vsaka enota v relaciji z vsako. Zato mora biti vsak par enot povezan z nasprotno usmerjenima črtama. To ponavadi poenostavimo tako, da enoti povežemo z neusmerjeno črto. Na drugi sliki, kjer je pet enot krepko povezanih, se lepo vidi, da lahko pridemo po usmerjenih črtah iz vsake enote v vsako enoto. To ne velja v primeru šibke povezanosti. Če ne upoštevamo usmerjenosti črt, pa lahko pridemo iz vsake enote v vsako drugo. V četrtem primeru gre za voditeljsko povezanost. Center določajo tri enote (1,2,4), ki so krepko povezane, iz preostalih enot pa gre usmerjena črta v eno od centralnih enot.

Primeri dopustnih razvrstitev z relacijsko omejitvijo $\Phi^i(R)$ so podani v tabeli 6.1. O teh množicah izvemo veliko z analizo same relacije R. Brez posebnih težav lahko pokažemo, da med



Slika 6.1: Tipi povezanosti

množicami dopustnih razvrstitev $\Phi^1(R)$, $\Phi^2(R)$, $\Phi^3(R)$, $\Phi^4(R)$ in $\Phi^5(R)$ veljajo naslednje štiri zveze:

a. $\Phi^4(R) \subseteq \Phi^3(R) \subseteq \Phi^2(R) \subseteq \Phi^1(R)$;

b. $\Phi^4(R) \subseteq \Phi^5(R) \subseteq \Phi^2(R)$;

c. Če je relacija R simetrična, je

$$\Phi^3(R) = \Phi^1(R);$$

d. Če je relacija R ekvivalenčna, je

$$\Phi^4(R) = \Phi^1(R).$$

Dokazi teh trditev so v Ferligoj (1983, 48-49 in 68).

Iz relacije R lahko razberemo za posamezne množice dopustnih razvrstitev $\Phi^i(R)$ tudi najmanjše možno število skupin v razvrstitvah iz $\Phi^i(R)$

$$\omega^i(R) = \min_{\mathcal{C} \in \Phi^i(R)} |\mathcal{C}|$$

Najmanjše možno število skupin v razvrstitvah iz $\Phi^1(R)$, $\Phi^2(R)$, $\Phi^3(R)$ in $\Phi^4(R)$ je:

$\omega^1(R)$ = število šibko povezanih komponent;
$\omega^2(R)$ = število centrov v množici E;
$\omega^3(R)$ = število krepko povezanih komponent;
$\omega^4(R)$ = moč minimalnega pokritja grafa (E, R) s klikami.

Dokaz tega izreka je prav tako v Ferligoj (1983, 49-52).

6.2.2 Omejevalna spremenljivka

V tem primeru je množica dopustnih razvrstitev določena takole:

$$\Phi_k[a,b] = \{\mathcal{C} : \text{je razbitje množice enot } E \text{ v k skupin in za vsako skupino } C \in \mathcal{C} \text{ velja: } v_C \in [a,b]\}$$

kjer je v_C funkcija vrednosti izbrane (omejevalne) spremenljivke V po enotah v skupini C. Na primer:

$$v_C = \sum_{X \in C} v_X$$

$$v_C = \max_{X \in C} v_X - \min_{X \in C} v_X$$

Vzemimo naslednji primer regionalizacije: razvrstiti je potrebno območja v k skupin tako, da so si znotraj skupine karseda podobna glede na družbeno-ekonomsko razvitost in da število prebivalcev v posamezni skupini območij (regiji) ni majše od a in ni večje od b. V tem primeru je omejevalna spremenljivka *število prebivalcev* in funkcija vsota vrednosti te spremenljivke po enotah v posamezni skupini (prvi primer zgoraj).

Če določimo omejevalno spremenljivko V tako, da vsaki enoti priredimo vrednost 1, je $\Phi[a,b]$ za tako določeno spremenljivko množica razvrstitev z omejitvijo števila enot v posamezni skupini.

Velja naslednja lastnost:

$$[a,b] \subseteq [c,d] \Rightarrow \Phi_k[a,b] \subseteq \Phi_k[c,d]$$

Pred reševanjem konkretnega problema razvrščanja v skupine z omejitvami je priporočljivo analizirati postavljene omejitve. V primeru omejevalne spremenljivke se je potrebno vprašati, ali je omejitveni interval $[a,b]$ glede na funkcijo vrednosti vseh enot v_E in število skupin k smiselno postavljen oziroma, ali omejitev zagotavlja neprazno množico dopustnih razvrstitev $\Phi_k[a,b]$. Ta analiza je odvisna od tipa funkcije.

Poglejmo za primer že omenjeno funkcijo

$$v_C = \sum_{X \in C} v_X$$

V tem primeru mora veljati naslednji (potrebni) pogoj:

$$a \leq \frac{v_E}{k} \leq b$$

Grobi meji, ki v tem primeru vedno zagotavljajo $\Phi_k[a, b] \neq 0$ (zadostni pogoj), pa sta

$$b \geq \lceil \frac{n}{k} \rceil \max_{X \in E} v_X$$

$$a \leq \lfloor \frac{n}{k} \rfloor \min_{X \in E} v_X$$

Včasih je težko poiskati posamezno razvrstitev, ki naj zadošča postavljeni omejitvi na spremenljivki. Tako na primer je problem določitve razvrstitve, za katero naj bo vsota vrednosti omejevalne spremenljivke v posamezni skupini porazdeljena karseda enakomerno, NP-težek (Shamos 1976).

6.2.3 Optimizacijska omejitev

Tretji tip omejitev le omenimo. V tem primeru gre za naslednje množice dopustnih razvrstitev:

$$\begin{aligned} \Phi(F) = \quad & \{\mathcal{C} : \text{je razbitje množice enot } E \text{ in zanjo velja,} \\ & \text{da ima tudi po drugem kriteriju } F \\ & \text{zadovoljivo vrednost: } F(\mathcal{C}) < f\} \end{aligned}$$

Vrednost drugega kriterija f je prag, ki določa število razvrstitev v množici dopustnih razvrstitev. Ponazorimo to misel zopet na

primeru regionalizacije, ki smo ga navedli v uvodu tega poglavja. Denimo, da razvrščamo območja glede na izbrane spremenljivke, ki merijo njihovo družbeno-ekonomsko razvitost. To je običjni problem razvrščanja v skupine, kjer je kriterij razvrščanja določen z merjenimi spremenljivkami. Vzemimo pa še en kriterij, ki naj bo karseda zadoščen ob razvrščanju v skupine: vsota dolžin cestnih povezav med centri območij naj bo karseda majhna. Ta pogoj (vsota dolžin povezav) določa drugi kriterij F. Zgornja definicija množice dopustnih razvrstitev rešuje ta problem tako, da za drugi kriterij smiselno postavimo prag f, ki določa podmnožico razvrstitev, v kateri iščemo optimimalno razvrstitev glede na prvi kriterij (razvrščanje v skupine). Manjši kot je postavljeni prag, bolj bo dobljena rešitev optimalna po drugem kriteriju in manj po prvem. Torej: določitev vrednosti praga je zelo pomembna odločitev.

Seveda lahko relacijske omejitve, omejitve na spremenljivkah, optimizacijske in morda še kakšne omejitve med seboj kombiniramo.

6.3 Reševanje problema razvrščanja v skupine z omejitvami

V splošnem je za reševanje problema razvrščanja v skupine z omejitvami mogoče prirediti vsaj dva tipa metod razvrščanja v skupine: metode hierarhičnega združevanja in metode lokalne optimizacije.

6.3.1 Prirejene metode hierarhičnega združevanja v skupine

Shema prirejenega postopka za hierarhično združevanje v skupine je naslednja:

vsaka enota je skupina:
$C_i = \{X_i\}$, $X_i \in E$, $i = 1, 2, ..., n$

ponavljaj, dokler sta vsaj dve skupini, ki ob združenju dasta dopustno razvrstitev:
določi najbližji par takih skupin C_p in C_q:
$d(C_p, C_q) = \min\{d(C_u, C_v) : C_u$ in $C_v, u \neq v,$ dasta pri združitvi dopustno razvrstitev $\}$;
združi skupini C_p in C_q v skupino $C_r = C_p \cup C_q$;
zamenjaj skupini C_p in C_q s skupino C_r;
določi mere različnosti d med novo skupino C_r in ostalimi.

V tem postopku je potrebno torej ob vsaki združitvi dve skupin preveriti, ali ta združitev zadošča vsem postavljenim omejitvam in zagotavlja dopustno razvrstitev.

Batagelj (1983) je pokazal, da je mogoče uporabiti tako prirejene metode hierarhičnega združevanja v skupine le v primeru, ko ima dana omejitev lastnost deljivosti. Omejitev $T(C)$ je deljiva, če za vsako netrivialno skupino C, ki jo sestavljata vsaj dve enoti, velja:

$$\exists C_1, C_2 \neq \emptyset :$$

$$(\ C_1 \cup C_2 = C \ \text{ in } \ C_1 \cap C_2 = \emptyset \ \text{ in } \ T(C_1) \ \text{ in } \ T(C_2)\ )$$

Omejitev na spremenljivki v splošnem ni deljiva. Pri postopnem združevanju skupin se namreč lahko zgodi, da v nekem koraku, ko je funkcija vrednosti v_C v posameznih skupinah še manjša od postavljene omejitve, ne more združiti nobenih dveh skupin tako, da bi bilo zadoščeno postavljeni omejitvi - v vsakem koraku bi bila funkcija vrednosti večja od postavljene omejitve. Zato za take probleme razvrščanja v skupine metode združevanja niso primerne (Ferligoj 1986). Tudi pri relacijski omejitvi je mogoče najti take vrste povezanosti v podgrafu (na primer krepka povezanost), za katere omejitev ni deljiva.

V primeru relacijske omejitve, ko je mogoče uporabiti metode hierarhičnega združevanja v skupine, je potrebno pri združevanju dveh skupin C_p in C_q popraviti tudi relacijo med združeno skupino $C_r = C_p \cup C_q$ in preostalimi skupinami, ne le različnosti med njimi. Vzemimo najprej primer razvrščanja v skupine s simetrično relacijsko omejitvijo, kjer so enote v vsaki skupini povezane s potmi, ki v celoti ležijo znotraj skupine glede na graf, ki je določen z množico enot E in simetrično relacijo R (npr. pri običajni regionalizaciji). V tem primeru popravljamo relacijo R takole:

$$
\begin{aligned}
R(C_r) &= \{C_r\} \cup R(C_p) \cup R(C_q) - \{C_p, C_q\} \\
R(C_k) &= \begin{cases} R(C_k) \cup \{C_r\} - \{C_p, C_q\} & C_k \in R(C_r) \\ R(C_k) & \text{drugače} \end{cases}
\end{aligned}
$$

kjer je $R(X) = \{Y | X R Y\}$.

Tudi v primeru nesimetrične relacijske omejitve je mogoče dobiti za omenjene tipe razvrstitev ($\Phi^1(R)$, $\Phi^2(R)$, $\Phi^4(R)$ in $\Phi^5(R)$) strategije popravljanja relacije tako, da se v vsakem koraku združevanja v skupine ohranja izbrani tip razvrstitve (Ferligoj 1983, 55-67; Ferligoj in Batagelj 1983) . Tako popustljiva strategija ohranja tip razvrstitve $\Phi^1(R)$, voditeljska $\Phi^2(R)$ in stroga $\Phi^5(R)$. V naslednjem razdelku so na primeru desetih enot, za katere je

podana matrika različnosti in relacija, prikazani učinki prav teh treh strategij.

Hierarhične rešitve problemov razvrščanja v skupine z relacijskimi omejitvami, dobljene z opisanim prirejenim postopkom združevanja, so največkrat nemonotone. Vzemimo najprej primer razvrščanja v skupine s simetrično relacijsko omejitvijo. V tem primeru lahko trdimo naslednje: Metoda združevanja v skupine, ki temelji na Lance in Williamsovem obrazcu zagotavlja monotone drevesne razvrstitve za vsako matriko različnosti D in vsako simetrično relacijsko omejitev R ($R \neq E \times E$, (E, R) je povezan graf) natanko takrat, ko koeficienti $(\alpha_1, \alpha_2, \beta, \gamma)$ na vsakem koraku združevanja zadoščajo pogojem:

$$\alpha_1 + \alpha_2 \geq 0$$

$$\gamma + \min(\alpha_1, \alpha_2) \geq 0$$

$$\min(\alpha_1 + \alpha_2 \ , \ \gamma + \min(\alpha_1, \alpha_2)) + \beta \geq 1$$

Dokaz te trditve je v Ferligoj in Batagelj (1982).

Analiza tega izreka pokaže, da izrek velja tudi za tri načine popravljanja nesimetrične relacije R (strogi, voditeljski in popustljivi).

Izmed poznanih metod hierarhičnega združevanja v skupine le maksimalna metoda zadošča tretjemu pogoju zgornjega izreka.

Doslej smo komaj kdaj obravnavali reševanje problemov razvrščanja v skupine z optimizacijsko omejitvijo. Omenim naj le, da za nekatere tipe optimizacijske omejitve je mogoče z metodo pragov prevesti problem razvrščanja v skupine z optimizacijsko omejitvijo na problem razvrščanja v skupine z relacijsko omejitvijo in ga nato rešujemo z opisanimi prirejenimi metodami hierarhičnega združevanja v skupine (npr. Ferligoj in Lapajne 1987).

	a	b	c	d	e	f	g	h	i	j
a	0	5	7	4	6	6	2	4	2	3
b		0	8	1	3	4	4	5	3	4
c			0	4	5	7	9	3	2	5
d				0	3	2	4	8	6	3
e					0	6	4	6	5	7
f						0	6	8	5	7
g							0	4	8	2
h								0	3	4
i									0	5
j										0

Tabela 6.2: Matrika različnosti

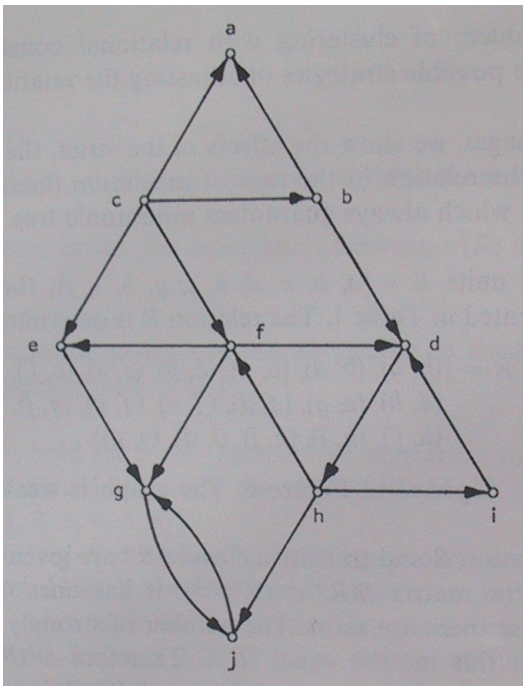
Primer

Na nekaj naslednjih straneh je na preprostem primeru desetih enot prikazan učinek strogega, voditeljskega in popustljivega načina popravljanja relacije pri (hierarhičnem) združevanju v skupine z maksimalno metodo, ki vedno zagotavlja monotona drevesa združevanja. Za množico desetih enot

$$E = \{a, b, c, d, e, f, g, h, i, j\}$$

je v tabeli 6.2 podana (simetrična) matrika različnosti D. Relacija pa je določena z naslednjo množico parov:

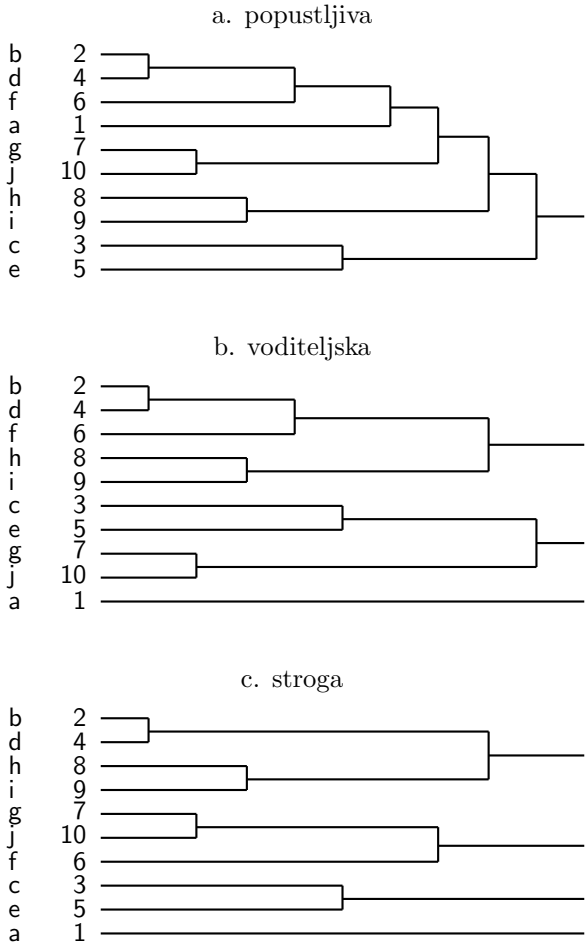
$$\begin{aligned} R = \{&(b,a),(b,d),(c,a),(c,b),(c,e),\\ &(c,f),(d,h),(e,g),(f,d),(f,e),\\ &(f,g),(g,j),(h,f),(h,i),(h,j),\\ &(i,d),(j,g)\} \end{aligned}$$



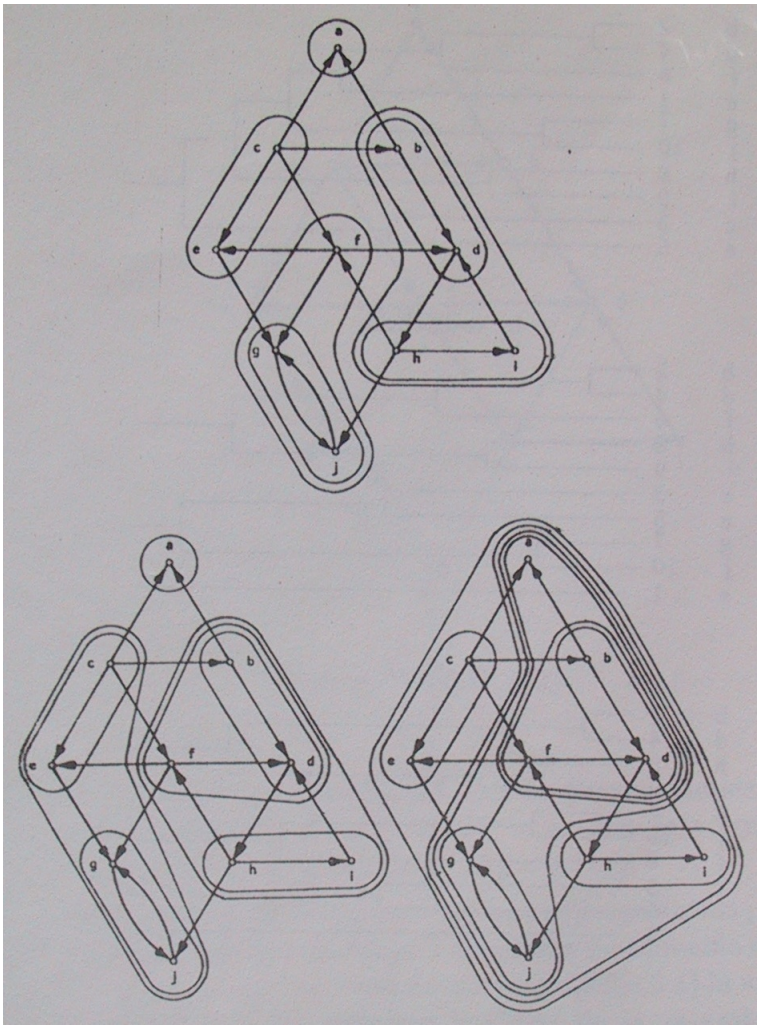
Slika 6.2: Graf (E,R)

Graf (E, R) je prikazan na sliki 6.3. Iz slike je razvidno, da je graf (E, R) šibko povezan. Zato je $\omega^1(R) = 1$. Ker je v tem grafu $\{g, j\}$ edina klika, ki ni enota, je $\omega^4(R) = 9$.

Učinek strogega, voditeljskega in popustljivega načina popravljanja relacije R pri združevanju v skupine z maksimalno metodo je grafično ponazorjen z drevesi združevanja (slika 6.4) in na grafu (E, R) (slika 6.5). Na slednjem je vidno, da je na vsakem koraku združevanja v skupine razviden tip razvrstitve, ki se pri



Slika 6.3: Tri relacijske omejitve



Slika 6.4: Strogi, voditeljski in popustljivi način združevanja

posemeznem načinu popravljanja relacije ohranja: pri strogi različici v vsaki dobljeni skupini obstaja enostaven sprehod, ki vsebuje vse enote te skupine; pri voditeljski različici vsaka skupina vsebuje en center; pri popustljivi pa so enote dobljene skupine šibko povezane. Iz obeh slik je razvidno, da z različnimi načini združevanja v skupine pri isti matriki različnosti in isti relaciji dobimo različne drevesne razvrstitve, ki imajo na koncu združevanja tudi različno število skupin: s strogim načinom štiri skupine, z voditeljskim tri skupine in s popustljivim eno samo skupino. Ti rezultati so skladni z ustreznimi $\omega^i(R)$. Razlike so tudi v sestavi skupin. Tako se na primer enota f po voditeljskem in popustljivem načinu pridruži skupini $\{b, d\}$, po strogem načinu pa ta združitev ni dopustna. Zato se ta enota pridruži na višjem nivoju skupini $\{g, j\}$.

6.3.2 Prirejena metoda prestavljanj

Shema postopka prirejene metode prestavljanj je naslednja:

določi začetno dopustno razvrstitev $\mathcal{C}$

dokler
    obstajata enoti $X_i \in C_p \in \mathcal{C}$ in $X_j \in C_q \in \mathcal{C}$, $p \neq q$
    tako, da je $P(\mathcal{C}') \leq P(\mathcal{C})$, kjer dobimo
    $\mathcal{C}'$ tako, da enoto X_i prestavimo iz skupine C_p
    v skupino C_q v razvrstitvi $\mathcal{C}$ ali enoti X_i
    in X_j medseboj premenjamo in pri tem tako dobljeni
    novi skupini zadoščata omejitvam
ponavljaj
    zamenjaj $\mathcal{C}$ z $\mathcal{C}'$.

V tem postopku ponavadi uporabljamo Wardovo kriterijsko funkcijo ali druge, ki smo jih omenili ob predstavitvi običajne

metode prestavljanj.

Tako prirejene metode lokalne optimizacije so primerne za vse tri tipe omejitev, vendar je pri nadaljnji razdelavi teh metod potrebno razrešiti vsaj dva problema:

- (učinkovito) preverjanje, ali je v vsaki skupini, dobljeni s prestavitvami ali premenami, zadoščeno postavljenim omejitvam,
- generiranje začetne dopustne razvrstitve.

Za nekatere omejitve je lahko že drugi problem NP-težek. Batagelj in Ferligoj (1985) sta ugotovila, da je za splošno relacijsko mejitev prvi problem lahko precej težek, ker nas pripelje do zapletenih problemov teorije grafov. Zato kaže, da je problem razvrščanja v skupine z nesimetrično relacijsko omejitvijo primerneje reševati z metodami hierarhičnega združevanja ali pa sestaviti posebne postopke za dani tip problema. Za omejitve na spremenljivkah pa je ustrezno prirejena metoda prestavljanj zelo primerna.

Tudi za optimizacijske omejitve je mogoče prirediti metodo prestavljanj. Možna je naslednja prirejena varianta te metode (Ferligoj 1987):

določi začetno dopustno razvrstitev $\mathcal{C}$ tako,
    da preveriš, če je $F(\mathcal{C}) > f$, kjer je f vnaprej
    podana vrednost (prag) omejevalnega kriterija F;
ponavljaj dokler gre
    če med tekočo razvrstitvijo $\mathcal{C}$ in sosednjimi
        dopustnimi razvrstitvami, ki jih dobiš s
        prestavitvami in premenami in ki jih preveriš
        na zgornji način, obstaja razvrstitev $\mathcal{C}'$,
        za katero velja $P(\mathcal{C}') \leq P(\mathcal{C})$,
    se pomakni v $\mathcal{C}'$.

Na razvrščanje v skupine z optimizacijsko omejitvijo lahko pogledamo tudi drugače: gre pravzaprav za dvojno optimizacijski problem. Iz teorije večkriterijske optimizacije je mogoče povzeti, da reševanje teh problemov s pragi v splošnem ne da dobrih rešitev (na primer paretovsko učinkovitih rešitev), čeprav problem rešujemo eksaktno. Iz te teorije izhaja, da je primerneje ustrezno kombinirati kriterije, po katerih optimiziramo. O tem več v naslednjem poglavju.

6.4 Koeficient vsiljenosti strukture

Za študij vpliva upoštevanih omejitev na razvrščanje v skupine glede na izbrano kriterijsko funkcijo lahko definiramo naslednji *koeficient vsiljenosti strukture* (Ferligoj 1986):

$$K = \frac{P(\mathcal{C}_c^*) - P(\mathcal{C}^*)}{P(\mathcal{C}_c^*)}$$

kjer je $\mathcal{C}^*$ dobljena razvrstitev brez omejitev in $\mathcal{C}_c^*$ dobljena razvrstitev z omejitvami ($P(\mathcal{C}_c^*) \geq P(\mathcal{C}^*)$). Koeficient K ni definiran, če je $P(\mathcal{C}^*) = 0$. V tem primeru naj bo $K = 0$. Koeficient vsiljenosti strukture K je definiran na intervalu $[0, 1]$. Če je $K = 0$, pomeni, da omejitve ne vplivajo na razvrščanje v skupine glede na izbrano kriterijsko funkcijo, kar pa ne zagotavlja, da sta razvrstitvi identični ($\mathcal{C}^* = \mathcal{C}_c^*$). Koeficient K meri relativni prirast kriterijske funkcije zaradi vpliva upoštevanih omejitev. Čim manjši je koeficient vsiljenosti strukture, tem manj omejitve vplivajo na razvrščanje v skupine glede na izbrano kriterijsko funkcijo.

6.5 Smeri razvoja razvrščanja v skupine z omejitvami

Odprtih vprašanj in problemov ter možnosti nadaljnjega raziskovanja razvrščanja v skupine z omejitvami je precej. Naštejemo jih vsaj nekaj:

- Omejitve, ki smo jih obravnavali doslej, se nanašajo le na enote znotraj posamezne skupine. Zanimivo bi bilo raziskati tudi omejitve, ki se nanašajo na odnose med skupinami.
- Problem razvrščanja v skupine z omejitvami smo reševali predvsem z dvema tipoma metod: hierarhičnim združevanjem v skupine in metodo s prestavljanji. V teoriji razvrščanja v skupine je znanih še veliko drugih zanimivih metod, ki bi jih najbrž bilo mogoče prirediti za reševanje vsaj nekaterih problemov razvrščanja v skupine z omejitvami (npr. metoda voditeljev).
- Obravnavane prirejene metode so primerne za razvrščanje le za manjša števila enot (nekaj sto). Potrebno bi bilo poiskati primernejše metode za razvrščanje v skupine z omejitvami večjega števila enot.
- Nekatere rešitve bi bilo potrebno še podrobneje razdelati: na primer metodo prestavljanj za nesimetrične relacijske omejitve in reševanje problema razvrščanja v skupine z optimizacijsko omejitvijo.

Vse pogostejša uporaba metod razvrščanja v skupine z omejitvami kaže, da problemi razvrščanja v skupine, če jih skrbno postavimo in dobro pretehtamo, postavljajo določene omejitve v iskanju razvrstitev. Neupoštevanje te omejitev privede do neveljavnih (celo nesmiselnih) rešitev.

6.6 Primer regionalizacije občin SR Slovenije

Nekateri pojavi se v danih teritorialnih območjih danem časovnem razdobju zaradi premajnega števila prebivalcev v njih redko pojavljajo, tako da jih ni mogoče statistično obravnavati (npr. samomori, redke bolezni). Problem lahko razrešimo vsaj na dva načina:

- da podaljšamo opazovalno časovno razdobje,
- da z združevanjem podobnih območij opzujemo dovolj velika območja.

Prvi način je manj primeren, ker postane zaradi sprememb v času proučevani pojav preveč heterogen. Drugi način je morda bolj obetaven, vendar je potrebno glede na postavljeni problem določiti kriterije za združevanje teritorialnih enot. To lahko naredimo posredno na osnovi spremenljivk, ki so močno povezane s proučevanimi pojavi. Tedaj lahko postavljeno nalogo združevanja območij zastavimo glede na teorijo razvrščanja v skupine z omejitvami takole: Območja je potrebno razvrstiti v skupine glede na izbrane spremenljivke, ki so povezane s proučevanimi redkimi pojavi, pri čemer je potrebno upoštevati naslednji omejitvi:

1. združene enote morajo biti geografsko sosednje in
2. v skupini mora biti vsaj dano število prebivalcev.

Množica dopustnih razvrstitev je torej omejena s simetrično relacijsko omejitvijo in omejitvijo vsote vrednosti omejevalne spremenljivke (število prebivalcev) v posamezni skupini ($\Phi_k(R)[a, b]$).

Denimo, da proučujemo delež prebivalcev z neko lastnostjo na občinah SR Slovenije in da smo ugotovili, da je za statistično

proučevanje tega deleža potrebno vsaj 29000 prebivalcev v posameznem teritorialnem območju C ($v_C \geq 29000$). Ker je v letih 1964 in 1974 približno dve tretjini občin SR Slovenije z manj prebivalci kot zahteva omejitev, je potrebno občine glede na izbrane spremenljivke razvrstiti v skupine, tako da je zadoščeno geografski sosednjosti in omejitvi števila prebivalcev v posamezni skupini.

Spremenljivke, na osnovi katerih občine razvrščamo v skupine, naj bodo naslednji kazalci družbeno-ekonomske razvitosti:

- družbeni proizvod na prebivalca
- osebni dohodek na prebivalca
- zaposleni v gospodarstvu na 100 prebivalcev
- prodano na drobno v 1000 din na prebivalca
- št. naročnikov dnevnikov (Delo, Lj. Dnevnik, Večer) na 100 prebivalcev
- št. oskrbovancev v zavodih za dnevno varstvo otrok na 100 prebivalcev
- št. obiskovalcev kina na prebivalca
- št. TV sprejemnikov na 100 prebivalcev
- št. telefonskih naročnikov na 100 prebivalcev
- št. osebnih avtomobilov na 100 prebivalcev
- št. priseljenih oseb v občino na 100 prebivalcev
- št. izseljenih oseb iz občine na 100 prebivalcev
- št. slušateljev višjih in visokih šol po stalnem bivališču na 100 prebivalcev
- št. seminarjev in tečajev na delavskih univerzah na 100 prebivalcev
- št. zaposlenih v šolstvu, prosveti in kulturi na 100 prebivalcev

Predno se lotimo razvrščanja občin v skupine, si najprej oglejmo obe omejitvi:

1. Relacijska omejitev je določena z geografsko sosednjostjo občin. Dve občini sta sosednji, če mejita druga na drugo. Relacija je razvidna iz zemljevida SR Slovenije, ki je prikazan na sliki 6.5 ali 6.6.

2. Omejitev prebivalcev v posamezni skupini ($v_C \geq 29000$) lahko preposto analiziramo tako, da občine uredimo glede na število prebivalcev v ranžirno vrsto v obeh letih. Iz pregleda teh podatkov lahko rečemo, da ima kar 44 občin v letu 1964 in 40 občin v letu 1974 manj prebivalcev kot zahteva postavljena omejitev. Ker se je število prebivalcev v Sloveniji od leta 1964 do leta 1974 povečalo od 1 625 933 na 1 782 470 prebivalcev (vendar ne enakomerno po vseh občinah - v kar 12 občinah se je število prebivalcev zmanjšalo), se je število občin, ki imajo manjše število prebivalcev od postavljene omejitve, zmanjšalo od 44 na 40 občin. Zato v letu 1974 omejitev števila prebivalcev manj vpliva na rešitev.

Pred razvrščanjem v skupine se postavi vprašanje, kako določiti čim večje število skupin, tako da razvrstitev še vedno zadošča obema omejitvama. S poizkusi slučajne določitve razvrstitev glede na obe omejitvi pri različnem številu skupin smo ugotovili, da je 35 skupin največje možno število skupin glede na dano število enot (60).

Za razvrščanje občin v skupine glede na 15 izbranih kazalcev družbeno-ekonomske razvitosti pri upoštevanju relacijske omejitve in omejitve števila prebivalcev v posamezni skupini smo uporabili prirejeno metodo s prestavljanji z Wardovo kriterijsko funkcijo, kjer smo različnost merili z evklidsko razdaljo, tako da smo predhodno spremenljivke standardizirali.

Na zemljevidu SR Slovenije (slika 6.5 in 6.6) sta vrisani obe najboljši dobljeni razvrstitvi v 35 skupin z obema omejitvama za leti 1964 in 1974. Razvrstitvi se razlikujeta predvsem v zahodni Sloveniji, kjer občini Škofja Loka in Radovljica v letu 1964 še ne, v letu 1974 pa že zadoščata postavljeni omejitvi števila prebivalcev. Zato lahko v letu 1974 tvorita vsaka zase svojo skupino in s tem vplivata na razvrščanje sosedov.

Primerjava dobljene razvrstitve brez omejitev z razvrstitvijo z obema omejitvama v posameznem letu razkriva tri osnovne tipe skupin:

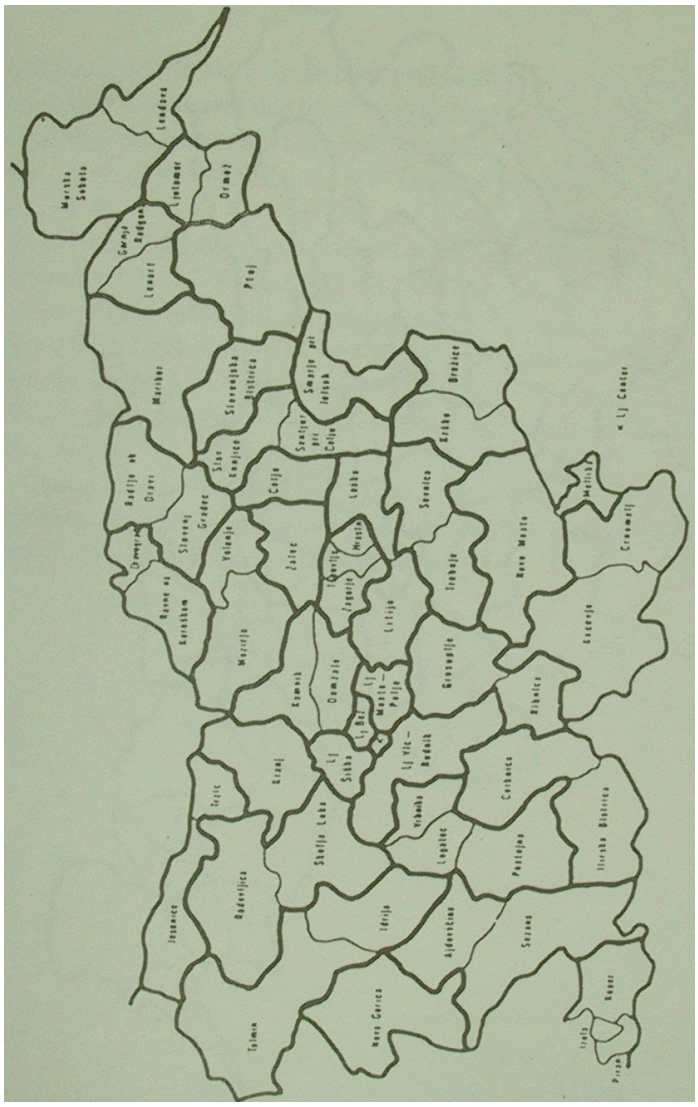
1. Skupine občin, ki so v obeh primerih razvrščanja (brez in z omejitvama) enake. V letu 1964 je takih skupin 10, v letu 1974 pa 9.

2. Skupine občin, ki so v primeru razvrščanja brez omejitev premajhne glede na število prebivalcev in so zato prisiljene k združitvi z najprimernejšimi sosedi glede na družbeno-ekonomsko razvitost. V vsakem letu je takih skupin 14.

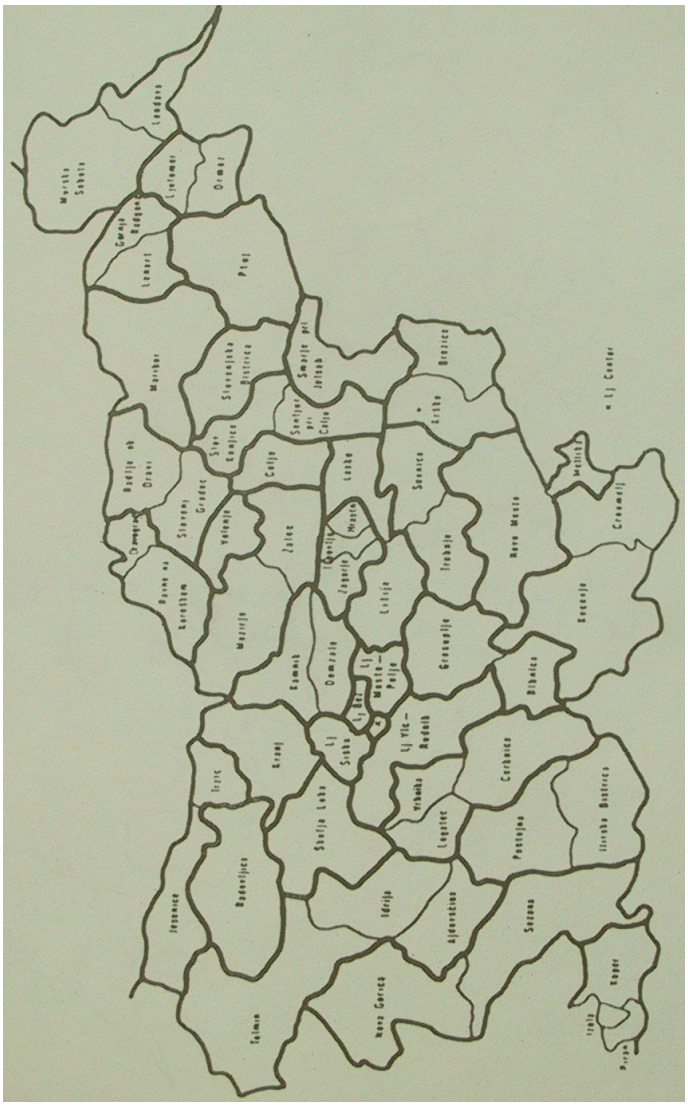
3. Skupine občin, ki so v primeru razvrščanja v skupine brez omejitev velike glede na število prebivalcev in se zato cepijo oziroma zmanjšajo. V letu 1964 jih je 8, v leu 1974 pa 9.

Le treh skupin občin v letu 1964 in ene skupine v letu 1974 ne moremo uvrstiti v zgornjo tipologijo.

Poudariti je potrebno še naslednje: čim večje je število skupin glede na dano število enot, tem bolj je razvrstitev določena le z omejitvama in manj z razvrščanjem glede na izbrane spremenljivke in kriterijsko funkcijo. To trditev lahko na našem primeru pokažemo s pregledom koeficientov vsiljenosti strukture, ki ga v tem primeru interpretiramo takole: čim večja je vrednost koeficienta, tem bolj razvrstitev določajo (vsiljujejo) postavljene omejitve. V



Slika 6.5: Razvrstitev slovenskih občin v letu 1964



Slika 6.6: Razvrstitev slovenskih občin v letu 1974

tabeli 6.3 so zapisani koeficienti vsiljenosti strukture za najboljše dobljene razvrstitve z relacijsko omejitvijo in omejtvijo števila prebivalcev glede na razvrstitev brez omejitev v pet, trideset in petintrideset skupin, dobljene z Wardovo metodo prestavljanj z upoštevanjem izbranih razvojnih kazalcev. Iz tabele je razvidno, da z naraščanjem števila skupin narašča vrednost koeficienta in sicer od $K_5 = 0.11$ pri razvrstitvi s petimi skupinami do $K_{35} = 0.50$ pri razvrstitvi s petintridesetimi skupinami v letu 1964 ter podobno v letu 1974 od $K_5 = 0.11$ do $K_{35} = 0.42$.

Primerjava dobljenih razvrstitev in koeficientov vsiljenosti strukture v letih 1964 in 1974 za razvrstitvi v petintrideset skupin kaže, tako kot že prej analiza omejevalne spremenljivke, da je v letu 1974 omejitev števila prebivalcev manj omejevalna ($K_{64} = 0.11$, $K_{74} = 0.42$).

Opisani metodološki pristop za določitev dovolj velikih območij za proučevanje redkih pojavov je bil uporabljen za analizo rakavih bolezni (Ferligoj in Pompe-Kirn 1988).

1964

št. skupin	5	30	35
brez omejitev	63.9	18.1	13.8
relac. omejitev in omejitev preb.	71.7	30.2	27.9
koef. vsiljenosti strukture (K)	0.11	0.40	0.50

1974

št. skupin	5	30	35
brez omejitev	65.0	19.7	15.3
relac. omejitev in omejitev preb.	73.0	28.7	26.4
koef. vsiljenosti strukture (K)	0.11	0.32	0.42

Tabela 6.3: Vrednosti Wardove kriterijske funkcije

7.

Večkriterijsko razvrščanje v skupine

7.1 Uvod

Velikokrat naletimo na probleme razvrščanja v skupine, ki jih ne moremo rešiti s klasičnimi metodami razvrščanja v skupine, ker zahtevajo optimizacijo po več kriterijih. Kot primer navedimo razvrščanje strokovnih revij z določenega področja v skupine, tako da so si revije znotraj posamezne skupine karseda podobne glede na recenzijsko politiko revije, obliko objavljenih člankov (npr. način citiranja vključenost povzetka v enem od svetovnih jezikov) in drugih lastnosti (prvi kriterij) ter obenem čimpogosteje druga drugo citirajo (drugi kriterij).

V splošnem se optimalne razvrstitve po posameznih kriterijih razlikujejo med seboj. Zato nastane problem, kako poiskati tako najboljšo rešitev, ki čim bolj zadošča vsem obravnavanim kriterijem.

Problem večkriterijskega razvrščanja v skupine lahko rešujemo

na več načinov:

- z redukcijo večkriterijskega problema na problem z enim kriterijem, ki je določen z neko kombinacijo obravnavanih kriterijev;
- z metodami razvrščanja z omejitvami, kjer izbrani kriterij jemljemo kot kriterij razvrščanja v skupine, ostali kriteriji pa določajo omejitve pri razvrščanju oziroma omejujejo množico dopustnih razvrstitev, po kateri optimiziramo (npr. Lefkovitch 1985; Ferligoj in Lapajne 1987);
- z direktnimi metodami večkriterijskega razvrščanja v skupine. Tak pristop je predlagal Hanani (1979), ki je v ta namen priredil metodo dinamičnih oblakov. Z Batageljem sva v ta namen priredila metodo prestavljanj in metode hierarhičnega združevanja v skupine (Ferligoj 1987; Ferligoj in Batagelj 1992).

V tem poglavju predstavljamo problem večkriterijskega razvrščanja v skupine in možnosti njegovega reševanja.

7.2 Problem večkriterijskega razvrščanja v skupine

Omenili smo že, da običajni problem razvrščanja v skupine lahko formuliramo kot optimizacijski problem takole: določiti želimo razvrstitev $\mathcal{C}^*$ tako, da bo

$$P(\mathcal{C}^*) = \min_{\mathcal{C} \in \Phi} P(\mathcal{C})$$

kjer je Φ množica dopustnih razvrstitev, $\mathcal{C}$ razvrstitev in P kriterijska funkcija.

Pri večkriterijskem razvrščanju v skupine imamo več kriterijskih funkcij $P_s, s = 1, \ldots, k$. Naš cilj je določiti razvrstitev $\mathcal{C}_1 \P i$ tako, da bo

$$P_s(\mathcal{C}) \to \min, \qquad s = 1, \ldots, k$$

V idealnem primeru iščemo dominantno razvrstitev. Razvrstitev $\mathcal{C}_0$ je *dominantna*, če za vsak $\mathcal{C} \in \Phi$ in za vsak kriterij P_s velja

$$P_s(\mathcal{C}_0) \leq P_s(\mathcal{C}), \qquad s = 1, \ldots, k$$

Običajno je množica dominantnih razvrstitev prazna. Zato nastane problem, kako najti rešitev večkriterijskega razvrščanja v skupine, ki je čim boljša glede na vse obravnavane kriterije. Vpeljimo pojem učinkovite razvrstitve. Učinkovitost razvrstitve je mogoče definirati na različne načine. Ena od najstrožjih je učinkovitost po Paretu:

Razvrstitev $\mathcal{C}^* \in \Phi$ je *učinkovita po Paretu* ali *paretovska razvrstitev*, če ne obstaja taka razvrstitev $\mathcal{C} \in \Phi$, za katero je

$$P_s(\mathcal{C}) \leq P_s(\mathcal{C}^*), \qquad s = 1, \ldots, k$$

pri čemer mora veljati stroga neenakost za vsaj en kriterij.

Če dominantna množica ni prazna, potem sovpada z množico po Paretu učinkovitih razvrstitev.

V literaturi je podanih še več drugih definicij učinkovitosti rešitev nalog večkriterijske optimizacije (npr. Podinovskij in Nogin 1982, 29-56; Homenjuk 1983, 12-13).

7.3 Reševanje problemov diskretne večkriterijske optimizacije

Večkriterijsko razvrščanje v skupine torej obravnavamo kot večkriterijski optimizacijski problem, ki je bil v zadnjih desetletjih zelo

pogosto obravnavan (npr. Mac Crimon 1973; Zeleny 1974; Podinovskij in Nogin 1982; Homenjuk 1983). V primeru večkriterijskega razvrščanja gre za diskretno večkriterijsko optimizacijo (množica dopustnih razvrstitev je končna), kar pomeni, da več zelo koristnih izrekov s področja večkriterijske optimizacije ne velja ali velja le delno, še posebej tistih, kjer je predpostavljena konveksnost. Ti izreki (povzeti so v Ferligoj 1987) obravnavajo predvsem strategijo konverzije večkriterijskega razvrščanja v skupine v enokriterijsko razvrščanje in problem zagotavljanja po Paretu učinkovite razvrstitve.

Vsekakor je zanimivo vprašanje, koliko paretovskih razvrstitev obstaja za dani večkriterijski problem razvrščanja v skupine. Velja naslednji izrek:

IZREK 7.31 *Naj bo $\mathcal{C}^* \in \Phi$ edina optimalna rešitev za kriterij P_i. Tedaj je ta razvrstitev tudi po Paretu učinkovita rešitev danega večkriterijskega problema.*

DOKAZ: Predpostavimo nasprotno: $\mathcal{C}^* \in \Phi$ ni paretovska razvrstitev. Tedaj obstajata razvrstitev $\mathcal{C} \neq \mathcal{C}^*$ in indeks k, tako da velja

$$P_k(\mathcal{C}) < P_k(\mathcal{C}^*)$$

in za $j \neq k$

$$P_j(\mathcal{C}) \leq P_j(\mathcal{C}^*)$$

Toda za i-ti kriterij je $\mathcal{C}^*$ optimalna razvrstitev. Zato je $P_i(\mathcal{C}) = P_i(\mathcal{C}^*)$, in ker je $\mathcal{C}^*$ edina optimalna rešitev, velja enakost $\mathcal{C} = \mathcal{C}^*$. Tako smo prišli do protislovja. ■

Iz tega izreka sledi, da je ob pogoju, da obstaja za vsak kriterij ena sama optimalna rešitev, minimalno število po Paretu učinkovitih razvrstitev dane naloge večkriterijskega razvrščanja v

skupine enako številu različnih optimalnih rešitev po posameznih kriterijih. Maksimalno število paretovskih razvrstitev je teoretično enako številu dopustnih razvrstitev. Število paretovskih razvrstitev je običajno nekje vmes. Ponavadi so razvrstitve, ki niso optimalne po posameznem kriteriju, zanimivejše. Prav gotovo je koristno, če lahko določimo vse ali večji del paretovskih razvrstitev. Ustrezen pregled teh razvrstitev nam pomaga razkriti rešitve problema večkriterijskega razvrščanja v skupine.

7.4 Reševanje problema večkriterijskega razvrščanja v skupine

V prejšnjem razdelku smo opozorili na nekaj lastnosti rešitev problema večkriterijskega razvrščanja v skupine. Po drugi strani pa smo v tretjem poglavju spoznali, da so običajni problemi razvrščanja v skupine v splošnem NP-težki. Zato jih rešujemo s hevrističnimi pristopi, ki pa ne zagotavljajo optimalne razvrstitve. V primeru večkriterijskega razvrščanja v skupine tedaj nastopata kar dva problema:

- problem lokalnih rešitev (razvrščanje) in
- problem učinkovitosti rešitev (večkriterijskost).

V nadaljevanju sta obravnavana dva tipa metod za reševanje večkriterijskega razvrščanja v skupine: prirejena metoda prestavljanj in metode hierarhičnega združevanja v skupine.

7.4.1 Metoda prestavljanj za večkriterijsko razvrščanje v skupine

Že večkrat se je izkazalo, da je za razvrščanje v skupine prav metoda prestavljanj tista, ki jo je mogoče ustrezno prirediti tudi za

precej zapletene probleme razvrščanja v skupine (npr. razvrščanje z omejevalno spremenljivko, Ferligoj 1986). Seveda ima tudi ta metoda svoje slabosti, ker pri slabši naravni strukturi podatkov kaj rada zaide v lokalno-optimalno razvrstitev. Vendar je z večkratno ponovitvijo postopka mogoče dobiti zelo zadovoljivo rešitev. Ob tem je zanimivo, da pomembni statistični paketi, kot so SPSS, BMDP ali celo SAS, te metode ne vključujejo. Metodo prestavljanj je mogoče učinkovito prirediti tudi za reševanje problema večkriterijskega razvrščanja v skupine.

Postopek metode prestavljanj, prirejen za večkriterijsko razvrščanje v skupine, sloni na sami definiciji po Paretu učinkovite rešitve (Ferligoj 1987) in je naslednji:

določi začetno dopustno razvrstitev $\mathcal{C}$ in zanjo izračunaj vrednosti vseh kriterijskih funkcij $P_s(\mathcal{C}), s = 1, \ldots, k$;

ponavljaj, dokler gre:
če med tekočo razvrstitvijo $\mathcal{C}$ in sosednjimi razvrstitvami obstaja razvrstitev $\mathcal{C}'$,
za katero velja
$$P_s(\mathcal{C}') \leq P_s(\mathcal{C}),\ s = 1, \ldots, k$$
kjer za vsaj en kriterij velja stroga neenakost,
se pomakni v razvrstitev $\mathcal{C}'$.

Sosednjo razvrstitev dobimo, če prestavimo enoto iz ene skupine v drugo skupino ali če enoti iz dveh skupin zamenjamo.

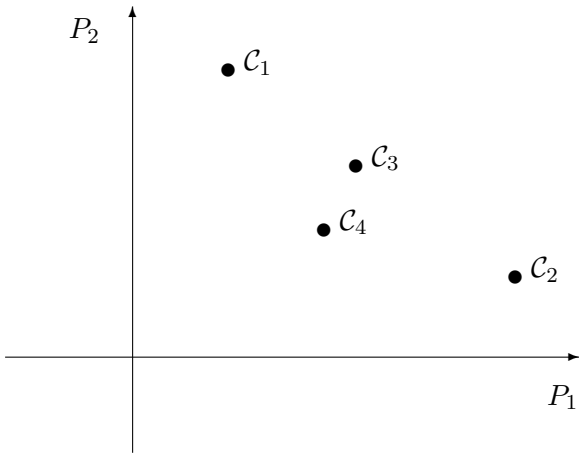
Rešitev je omejena s sosedstveno strukturo razvrstitev in je odvisna od začetne razvrstitve. Zato dobljena rešitev ni nujno po Paretu učinkovita rešitev in jo imenujmo *lokalna paretovska razvrstitev*. Priporočljivo je postopek večkrat ponoviti (nekaj stokrat) in dobljene rešitve pregledati. Učinkovit pregled dobljenih

razvrstitev je mogoče sistematično vgraditi v naslednji postopek:

> določi optimalne razvrstitve po posameznem kriteriju in jih dodaj v množico paretovskih razvrstitev Π ;
>
> **ponavljaj**:
>
> > z lokalno optimizacijo (po metodi prestavljanj) določi tekočo lokalno paretovsko razvrstitev $\mathcal{C}$;
> > če ne obstaja taka razvrstitev iz tekoče množice paretovskih razvrstitev $\mathcal{C}_p \imath \Pi$, za katero je
> > $$P_s(\mathcal{C}_p) \leq P_s(\mathcal{C}),\ s = 1, \ldots, k$$
> > pri čemer mora veljati stroga neenakost za vsaj en kriterij, jo vključi v množico paretovskih razvrstitev Π;
> > če pri tem katera od razvrstitev ni več paretovska, jo izloči.

S tako nadgrajenim meta postopkom za metodo prestavljanj torej najprej vpnemo vse razvrstitve, ki jih dobimo s ponavljanjem postopka metode prestavljanj v kriterijskem prostoru, med robne točke, ki so optimalne razvrstitve po posameznem kriteriju in ki so glede na omenjeni izrek prav gotovo paretovske razvrstitve. Hkrati sproti preverjamo, ali tekoča dobljena razvrstitev sodi v množico paretovskih razvrstitev Π. Z dodajanjem in izločanjem se tekoča množica Π približuje pravi množici paretovskih razvrstitev. Seveda pa je mogoče, da se v njej nahajajo tudi lokalno paretovske razvrstitve, še posebej, če postopka nismo dovoljkrat ponovili.

Postopek ponazorimo z dvokriterijskim primerom (glej sliko 7.1), kjer v i-tem koraku postopka sestavljajo množico paretovskih razvrstitev $\Pi = \{\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3\}$; $\mathcal{C}_1$ je optimalna razvrstitev glede na prvi in $\mathcal{C}_2$ optimalna razvrstitev glede na drugi kriterij. Razvrstitev $\mathcal{C}_3$ se do tega koraka kaže kot paretovska razvrstitev. V naslednjem koraku dobimo razvrstitev $\mathcal{C}_4$, ki zadošča pogoju vključitve



Slika 7.1: Paretovske razvrstitve

v množico paretovskih razvrstitev, pri tem pa je potrebno razvrstitev $\mathcal{C}_3$ izločiti iz tekoče množice Π .

7.4.2 Metode hierarhičnega združevanja za večkriterijsko razvrščanje v skupine

Običajni postopek hierarhičnega združevanja v skupine, ki smo ga podrobneje obravnavali na začetku četrtega poglavja, predpostavlja, da imamo mere različnosti med enotami urejene v simetrični matriki $D = [d_{i,j}]$. V primeru večkriterijskega razvrščanja v skupine pa ponavadi lahko vhodne podatke uredimo v k matrik različnosti $D^s, s = 1, \ldots, k$ (npr. če merimo različnosti med obravnavanimi enotami v k različnih pogojih, časovnih trenutkih, itd.).

V tem primeru želimo poiskati hierarhično razvrstitev, ki bo karseda dobro popisala strukture podatkov, ki se skrivajo v k matrikah različnosti. Prva misel, kako dobiti to rešitev, je naslednja: v koraku postopka hierarhičnega združevanja, ko poiščemo par najbližjih skupin, to storimo tako, da iščemo najbližji par v vseh k matrikah različnosti. Nato združimo najbližji skupini v vsaki matriki različnosti posebej in nadaljujemo z računanjem različnosti nove združene skupine s preostalimi skupinami po izbrani metodi. Če so mere različnosti v posameznih matrikah neprimerljive, je potrebno matrike različnosti ustrezno normalizirati. Seveda je ta korak nepotreben, če gre na primer za enako mero različnosti, ki jo računamo med izbranimi enotami, določnimi z enakimi spremenljivkami, opazovanimi v različnih časih.

Ta pristop lahko posplošimo tako, da definiramo skupno matriko $D = [d_{i,j}]$ na primer takole:

$$d_{i,j} = \max(d^s_{i,j}; s = 1, \ldots, k)$$

$$d_{i,j} = \min(d^s_{i,j}; s = 1, \ldots, k)$$

$$d_{i,j} = \sum_{s=1}^{k} \alpha_s d^s_{i,j} \quad , \quad \sum_{s=1}^{k} \alpha_s = 1$$

Postopek prirejene metode hierarhičnega združevanja v skupine je lahko tedaj naslednji:

vsaka enota je skupina:
$C_i = \{X_i\}$, $X_i \in E$, $i = 1, 2, \ldots, n$

normaliziraj vsako matriko različnosti $D^s, s = 1, \ldots, k$;

ponavljaj, dokler ne ostane ena sama skupina:
določi skupno matriko $D = f(D^s; s = 1, \ldots, k)$;
določi v D najbližji si skupini C_p in C_q:
$d(C_p, C_q) = \min_{u,v} d(C_u, C_v)$;
združi skupini C_p in C_q v skupino $C_r = C_p \cup C_q$;
v vsaki matriki različnosti $D^s, s = 1, \ldots, k$:
zamenjaj skupini C_p in C_q s skupino C_r;
določi mere različnosti d med novo skupino C_r in ostalimi.

S tem postopkom dobimo eno drevesno razvrstitev, ki jo lahko ponazorimo tudi z drevesom združevanja.

Postopek združevanja lahko za večkriterijsko razvrščanje v skupine priredimo tudi tako, da v postopku poiščemo najbližji par skupin po paretovskemu kriteriju takole:

Par skupin (C_p, C_q) je paretovsko najbližji, če ne obstaja drugi par skupin (C_i, C_j), za katerega je

$$d^s_{i,j} \leq d^s_{p,q} \qquad s = 1, \ldots, k$$

pri čemer mora veljati stroga neenakost za vsaj eno matriko različnosti.

V tem postopku lahko v posameznem koraku dobimo več paretovskih parov skupin. To pomeni, da je rezultat tega postopka množica drevesnih razvrstitev, ki jo imenujmo množica *paretovskih drevesnih razvrstitev*. Problem pri tem postopku je, da je lahko ta množica rešitev zelo velika. Potrebno bi bilo najti ustrezen meta

postopek za pregled dobljenih drevesnih rešitev, podoben meta postopku pri prirejeni metodi prestavljanj. Najbrž ni potrebno posebej poudariti, da teh drevesnih razvrstitev ne moremo preprosto predstaviti z drevesom združevanja. Ponavadi rešitve podamo v oklepajski obliki. Pa še ena opomba: v tem postopku ni potrebna normalizacija matrik različnosti.

7.5 Primera večkriterijskega razvrščanja v skupine

Za prikaz obravnavanih metod za večkriterijsko razvrščanje v skupine potrebujemo podatke o določenih enotah, zbrane na različne načine ali ob različnih pogojih. Najlažje je prikazati predloženo metodo na manjši skupini enot, za katero je mogoče rešitve poiskati tudi brez računalnika. Zato najprej ponazorimo predstavljene metode na množici šestih enot, določenih z dvema spremenljivkama v dveh časih. V drugem primeru pa so uporabljeni podatki o politikih iz II. svetovne vojne, ki jih je zbral in analiziral Everitt (1987) z metodami večsmernega večrazsežnostnega lestvičenja. V tem primeru gre za trirazsežno matriko podatkov: posamezna matrika različnosti med enotami je dobljena na več načinov.

Poudariti moramo, da so predložene metode večkriterijskega razvrščanja v skupine uporabne tudi za drugačne tipe podatkov, kjer so posamezne matrike podatkov, na osnovi katerih računamo kriterijske funkcije, lahko povsem različne.

7.5.1 Razvrščanje šestih enot

Kot prvi primer večkriterijskega razvrščanja v skupine vzemimo torej množico šestih enot

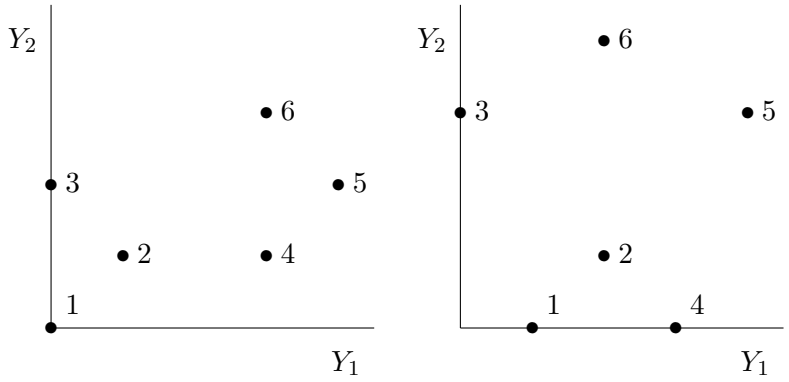
$$E = \{X_1, X_2, X_3, X_4, X_5, X_6\}$$

	1		2	
enote	Y_1	Y_2	Y_1	Y_2
1	0	0	1	0
2	1	1	2	1
3	0	2	0	3
4	3	1	3	0
5	4	2	4	3
6	3	3	2	4

Tabela 7.1: Šest enot v dveh časovnih točkah

ki jih določata dve spremenljivki (Y_1 in Y_2) v dveh časih. Podatki so podani v tabeli 7.1. Poglejmo, kaj nam v tem primeru razkrijejo obravnavane metode za večkriterijsko razvrščanje v skupine. Najprej je potrebno na osnovi podatkov, podanih v tabeli 7.1, izračunati za vsako časovno točko posebej različnosti med šestimi enotami. V tabeli 7.2 so v zgornjem trikotniku prikazani izračunani kvadrati evklidskih razdalj za prvo časovno točko, v spodnjem trikotniku pa za drugo.

Razvrstimo šest enot v popolno razvrstitev z dvema skupinama. V tem primeru gre za majhno število enot, kjer lahko pregledamo vseh 31 dopustnih razvrstitev. Če izračunamo za vsako razrstitev Wardovo kriterijsko funkcijo na osnovi kvadratov evklidksih razdalj, ki so zapisane v tabeli 7.2, dobimo najmanjšo vrednost te funkcije v prvem času za razvrstitev $\{\{1,2,3\},\{4,5,6\}\}$, v drugem času pa za razvrstitev $\{\{1,2,4\},\{3,5,6\}\}$. Ti dve razvrstitvi smo seveda pričakovali, kajti iz grafičnega prikaza na sliki 7.2 sta obe razvrstitvi lepo razvidni. Dobljeni optimalni razvrstitvi po posameznemu kriteriju sta tudi paretovski razvrstitvi. Ali



Slika 7.2: Šest enot v prvem in drugem času

	1	2	3	4	5	6
1	0	2	4	10	20	18
2	2	0	2	4	10	8
3	10	8	0	10	16	10
4	4	2	18	0	2	4
5	18	8	16	10	0	2
6	17	9	5	17	5	0

Tabela 7.2: Kvadrati evklidskih razdalj za obe časovni točki

sta to edini paretovski razvrstitvi? Ponovimo postopek prirejene metode prestavljanj nekaj desetkrat. Z meta postopkom dobimo poleg že omenjenih razvrstitev še eno: $\{\{1,2,3,4\},\{5,6\}\}$. Če v meta postopku le nekajkrat ponovimo metodo prestavljanj, lahko zaidemo tudi v lokalno paretovsko razvrstitev $\{\{1,3,6\},\{2,4,5\}\}$.

Podatke analizirajmo še s prirejenim postopkom hierarhičnega združevanja v skupine z iskanjem parov najbližjih skupin po Paretu. Uporabimo metodo maksimum. Rezultat so naslednje tri drevesne razvrstitve

$$(((1,2),3),(4,(5,6)))$$

$$((((1,2),4),3),(5,6))$$

$$(((1,2),4),(3,(5,6)))$$

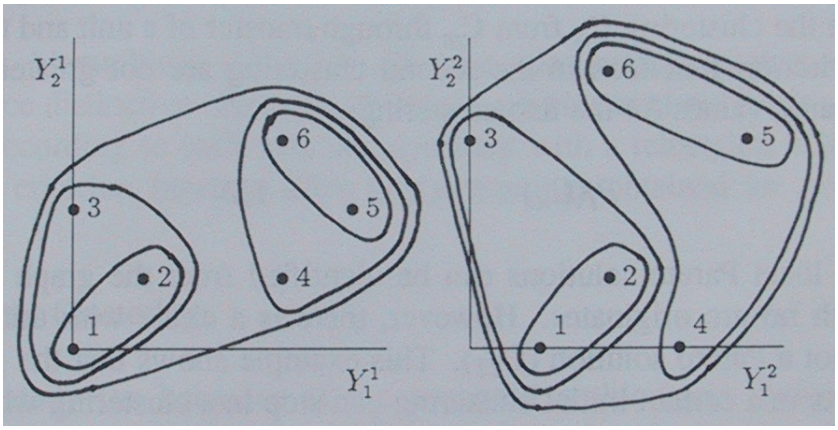
Tri dobljene drevesne razvrstitve so prikazane na slikah 7.3, 7.4 in 7.5. Če se v teh treh drevesnih razvrstitvah osredotočimo na razvrstitve z dvema skupinama, opazimo tri paretovske popolne razvrstitve, ki smo jih dobili s prirejeno metodo prestavljanj. Pri tem je zanimivo tudi to, da smo v primeru metode prestavljanj uporabili Wardovo kriterijsko funkcijo, v primeru hierarhičnega združevanja v skupine pa povsem drugačno funkcijo, namreč:

$$P(\mathcal{C}) = \max_{C\in\mathcal{C}} \max_{X,Y\in C} d^2(X,Y)$$

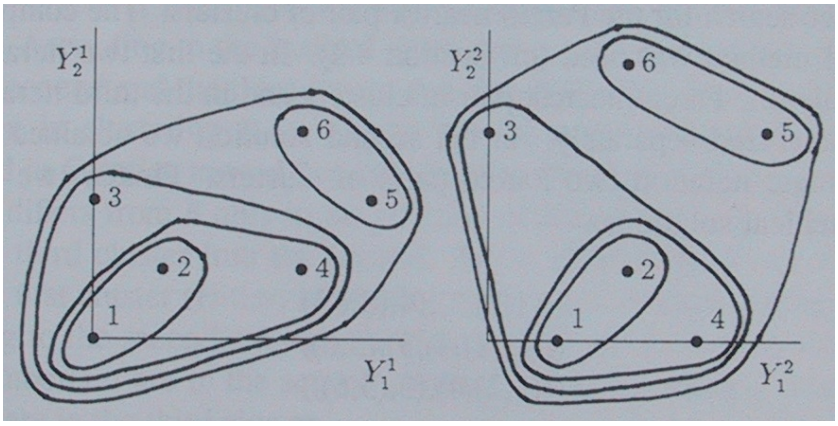
Tudi s popolnim pregledom dopustnih razvrstitev smo dobili tri omenjene paretovske razvrstitve z dvema skupinama.

7.5.2 Politiki iz II. svetovne vojne

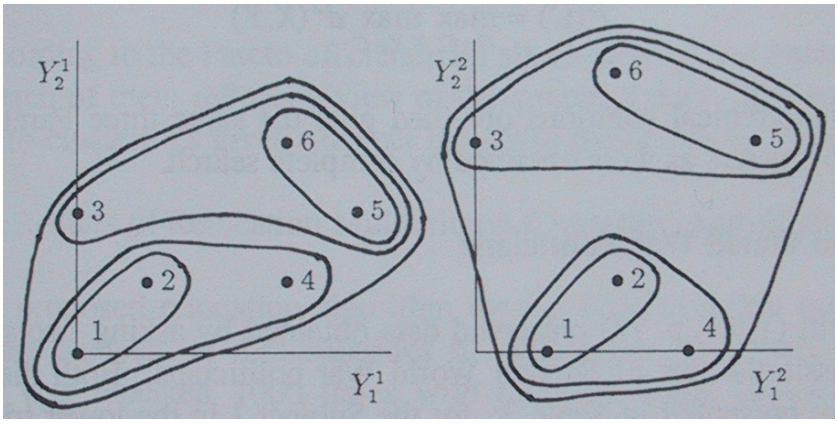
Everitt (1987) je v svojem delu uporabil metodo nemetričnega večrazsežnostnega lestvičenja (MULTISCAL) na podatkih, ki so



Slika 7.3: Drevesna razvrstitev (((1,2),3),(4,(5,6))) v obeh časih



Slika 7.4: Drevesna razvrstitev ((((1,2),4),3),(5,6)) v obeh časih



Slika 7.5: Drevesna razvrstitev (((1,2),4),(3,(5,6))) v obeh časih

bili dobljeni tako, da sta dve osebi v Veliki Britaniji ocenili različnosti za vsak par izbranih znanih politikov iz II. svetovne vojne. Ocene različnosti prve osebe so podane v spodnjem trikotniku, ocene druge osebe pa v zgornjem trikotniku v tabeli 7.3.

Analiza vsake matrike različnosti posebej

Hierarhično združevanje politikov glede na ocene različnosti prve osebe razkrivajo tri izrazite skupine politikov in prav tako tudi analiza ocen druge osebe. Drevesi se delno razlikujeta. S popolnim pregledom vseh 86.526 razvrstitev s tremi skupinami smo z Wardovo kriterijsko funkcijo določili najboljši razvrstitvi, ki sta prikazani v tabeli 7.4.

		1	2	3	4	5	6	7	8	9	10	11	12
1	*Hitler*		2	7	8	5	9	2	6	8	8	8	9
2	*Mussolini*	3		8	8	8	9	1	7	9	9	9	9
3	*Churchill*	4	6		3	5	8	7	2	8	3	5	6
4	*Eisenhower*	7	8	4		8	7	7	3	8	2	3	8
5	*Stalin*	3	5	6	8		7	7	5	6	7	9	5
6	*Attlee*	8	9	3	9	8		9	7	7	4	7	5
7	*Franco*	3	2	5	7	6	7		5	9	8	8	9
8	*DeGaulle*	4	4	3	5	6	5	4		6	5	6	5
9	*MaoTseTung*	8	9	8	9	6	9	8	7		8	8	6
10	*Truman*	9	9	5	4	7	8	8	4	4		4	6
11	*Chamberlain*	4	5	5	4	7	2	2	5	9	5		8
12	*Tito*	7	8	2	4	7	8	3	2	4	5	7	

Tabela 7.3: Matriki različnosti

		1.*oseba*			2.*oseba*
I	1	*Hitler*	*I*	1	*Hitler*
	2	*Mussolini*		2	*Mussolini*
	7	*Franco*		5	*Stalin*
				7	*Franco*
II	3	*Churchill*	*II*	3	*Churchill*
	4	*Eisenhower*		4	*Eisenhower*
	8	*DeGaulle*		8	*DeGaulle*
	10	*Truman*		9	*MaoTseTung*
	11	*Chamberlain*		10	*Truman*
				12	*Tito*
III	5	*Stalin*	*III*	6	*Attlee*
	6	*Attlee*		11	*Chamberlain*
	9	*MaoTseTung*			
	12	*Tito*			
		$P_1 = 17.87$			$P_2 = 18.17$

Tabela 7.4: Najboljši razvrstitvi za vsako osebo posebej

Določitev vseh paretovskih razvrstitev s pregledom vseh razvrstitev s tremi skupinami

Za vseh 86.526 razvrstitev s tremi skupinami so bile izračunane Wardove kriterijske funkcije na osnovi obeh matrik različnosti. S pregledom le-teh smo dobili šest paretovskih razvrstitev, ki so prikazane v tabeli 7.5.

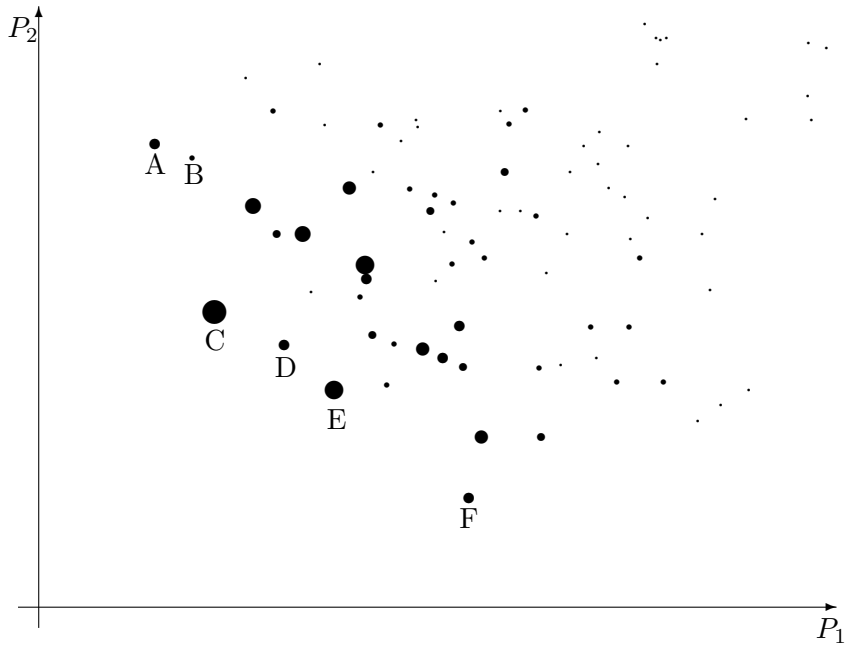
Prve štiri dobljene paretovske razvrstitve imajo enake prve skupine (Hitler, Mussolini in Franco). Glede na optimalno razvrstitev politikov po oceni 1. osebe (prva razvrstitev) se druga in tretja paretovska rešitev razlikujeta le v poziciji enega politika. V drugi razvrstitvi je De Gaulle prestavljen iz druge skupine v tretjo, v tretji pa Attlee iz tretje v drugo skupino. Četrta paretovska razvrstitev se glede na prvo razlikuje v tem, da sta Attlee in Tito prestavljena iz tretje skupine v drugo. Peta paretovska razvrstitev ima enako prvo skupino kot šesta razvrstitev (Hitler, Mussolini, Stalin in Franco), ki je optimalna za izbor druge osebe. V slednji so preostale enote v drugi skupini razen Attleeja in Chamberlaina, v peti razvrstitvi pa sta se v tretjo skupino ločila Mao Tse Tung in Titom. V dobljenih paretovskih razvrstitvah sta torej najpogosteje različno razvrščena Attlee (trikrat) in Tito (dvakrat). V vseh dobljenih razvrstitvah pa vedno nastopajo skupaj Hitler Mussolini in Franco v prvi skupini ter Churchill, Eisenhower in Truman v drugi skupini, ki se ji po petkrat pridružita De Gaulle in Chamberlain.

Rezultati metode prestavljanj, prirejene za večkriterijsko razvrščanje v skupine

Na osnovi obeh matrik različnosti razvrščamo v tri skupine s prirejeno metodo s prestavljanji tako, da za obe matriki računamo Wardovo kriterijsko funkcijo. Postopek smo ponovili s 1000 različnimi

	opt.1					*opt*.2
skupina	*A*	*B*	*C*	*D*	*E*	*F*
I	1	1	1	1	1	1
	2	2	2	2	2	2
	7	7	7	7	5	5
					7	7
II	3	3	3	3	3	3
	4	4	4	4	4	4
	8	10	6	6	6	8
	10	11	8	8	8	9
	11		10	10	10	10
			11	11	11	12
				12		
III	5	5	5	5	9	6
	6	6	9	9	12	11
	9	8	12			
	12	9				
		12				
P_1	17.87	18.47	18.83	19.95	20.75	22.92
P_2	21.97	21.82	20.17	19.81	19.33	18.17

Tabela 7.5: Paretovske razvrstitve



Slika 7.6: Razvrstitve, dobljene s 1000-kratno ponovitvijo metode prestavljanj

začetnimi razvrstitvami. Dobljene razvrstitve lahko zelo nazorno predstavimo v kriterijskem prostoru, ki je v našem primeru dvorazsežen (glej sliko 7.6). Vsaka točka predstavlja eno razvrstitev. Ploščina kroga, ki predstavlja razvrstitev, je sorazmerna s frekvenco pojavitve razvrstitve pri 1000 ponovitvah postopka.

Pri 1000-kratni ponovitvi postopka smo dobili kar 84 različnih razvrstitev. V 69.9% vseh ponovitev smo dobili 78 lokalnih paretovskih razvrstitev. Postopek zlahka razkrije tretjo razvrstitev

(Hitler, Mussolini, Franco), (Churchill, Eisenhower, Attlee, De Gaulle, Truman, Chamberlain), (Stalin, Mao Tse Tung, Tito) (13%) in peto (9.5%), težje četrto (3,2%), obe optimalni razvrstitvi po vsakem kriteriju posebej pa še težje (šesto v 2,7%, prvo pa v 2.2%). Najtežje pa je postopek razkril drugo razvrstitev (0.8% dobljenih razvrstitev od 1000 ponovitev s slučajno začetno razvrstitvijo). Zato je prav gotovo smiselno pred ponovitvami postopka najprej izračunati optimalne razvrstitve po posameznih kriterijih, si jih zapomniti in si od sproti dobljenih razvrstitev zapomniti le tiste, ki so paretovske rešitve glede na že izbrane razvrstitve. Ker so sosednostne strukture razmeroma revne, je postopek za posamezno iskanje rešitve zelo kratek. Zato je potrebno metodo prestavljanj velikokrat ponoviti (vsaj 100 krat). Za primer smo meta-postopek ponovili 25-krat s po 100 ponovitvami. Od teh smo 18-krat dobili vse paretovske razvrstitve, 7-krat pa je manjkala le druga razvrstitev, ki se tudi sicer redko pojavlja.

Literatura

[1] Acketa D. (1986): Odabrana poglavlja teorije prepoznavanja oblika sa primenama. Novi Sad: Institut za matematiku.

[2] Ajvazjan S.A., Bežajeva Z.I. in Staroverov O.V. (1974): Klassifikacija mnogomernih nabljudenij, Moskva: Statistika.

[3] Aldenderfer M.S. in Blashfield R.K. (1984): Cluster Analysis. Series on Quantitative Applications in the Social Sciences, 44, Bevery Hills: Sage.

[4] Anderberg M.R. (1973): Cluster Analysis for Applications. New York: Academic Press.

[5] Arabie P., Carroll J.D., De Sarbo W.S. (1987): Three Way Scaling and Clustering. Series on Quantitative Applications in the Social Sciences, 65, Bevery Hills: Sage.

[6] Bailey K.D. (1974): Cluster analysis. V: D.R. Heise (Ed.): Sociological methodology 1975, San Francisco: Jossey-Bass Publ.

[7] Bajc D. in Pisanski T. (1985): Najnujnejše o grafih. Ljubljana: DMFA.

[8] Ball G. in Hall D.J. (1967): A clustering technique for summarizing multivariate data. Behavioral Science 12, 153-155.

[9] Batagelj V. (1979): Razvrščanje v skupine - osnovni pojmi. Seminar za numerično in računalniško matematiko, 159, Ljubljana: DMFA SRS.

[10] Batagelj V. (1981): Note on ultrametric hierarchical clustering algorithms. Psychometrika, 46 (3), 351-352.

[11] Batagelj V. (1982): CLUSE - programi za razvrščanje v skupine. (priročnik za DEC-10). Ljubljana

[12] Batagelj V. (1984): Notes on the dynamic clusters method. Proceedings of the IV. Conference on Applied Mathematics, Split, 139-146.

[13] Batagelj V. (1985a): Algorithmic aspects of the clustering problem. Proceedings of 7 th International Symposium 'Computer at the University', Zagreb: SRCE, 27-29.5.

[14] Batagelj V. (1985b): Razvrščanje v skupine - nehierarhični postopki. Ljubljana: Fakulteta za elektrotehniko (magistrsko delo).

[15] Batagelj V. (1986 a): On the adding clustering algorithms. V: COMPSTAT 1986, Rim: Universita "La Sapienza", 29-30.

[16] Batagelj V. (1986 b): Razvrščanje in optimizacija. Statistička revija, 1-2, 167-174.

[17] Batagelj V. (1988 a): Similarity measures between structured objects, I. Proceedings of the MATH/CHEM/COMP 88, Dubrovnik.

[18] Batagelj V. (1988 b): Generalized Ward and related clustering problems. V: H.H. Bock (Ed.): Classification and Related Methods of Data Analysis, Amsterdam: North-Holland, 67-74.

[19] Batagelj V. (1989 a): CLUSE za IBM PC - priročnik. Ljubljana.

[20] Batagelj V. (1989 b): Povabilo v LATEX. Seminar za računalniško matematiko, Ljubljana: DMFA SRS.

[21] Beale E.M.L. (1969): Cluster analysis. London: Scientific Control Systems.

[22] Berry B.J.L. in Ray M. (1966): Multivariate socio-economic regionalization: A pilot study in central Canada, Department of Geography, University of Chicago.

[23] Bezdek J. (1981): Pattern Recognition with Fuzzy Objective Functions. New York: Plenum.

[24] Blejec M. (1973): Statistične metode za ekonomiste. Ljubljana: Univerza v Ljubljani.

[25] Bijnen E.J. (1973): Cluster analysis: Survey and evaluation of techniques. Groningen: Tilburg University Press.

[26] Bock H.H. (1974): Automatische Klassifikation. Gottingen: Vandenhoeck and Ruprecht.

[27] Bodjanova S. (1989): Fuzzy aspect in classification and reduction of dimensionality of multivariate observations. Članek na Majskem skupu '89. Sekcije za klasifikacije Saveza statističkih društava Jugoslavije, Mostar.

[28] Bogosavljević S. (1988): Evaluacija klasifikacione strukture. Zbornik radova. Majski skup '87. Sekcije za klasifikacije Saveza statističkih društava Jugoslavije. Beograd: Savezni zavod za statistiku, 28-32.

[29] Bonomi E. in Lutton J.L. (1984): The n-city travelling salesman problem: Statistical mechanics and the Metropolis algorithm. SIAM Review, 26, 551-568.

[30] Bouroche J.M. in Saporta G. (1980): L'analyse des donnes. Que sais-je? Pariz: Presses Universitaires de France.

[31] Brucker P. (1978): On the complexity of clustering problems. V: R. Henn, B. Korte, W. Oettli (Eds.): Optimization and Operations

Research. lecture Notes in Economics and Mathematical Systems, 157, Berlin: Springer-Verlag.

[32] Clifford H.T. in Stephenson W. (1975): An introduction to numerical classification. New York: Academic Press.

[33] Cormack R.M. (1971): A rewiew of classification. Journal of the Royal Stat. Soc. Series A, 134, 321-367.

[34] Cronbach L.J. in Gleser G.C. (1953): Assessing the similarity between profiles. Psychol. Bull., 50, 456-473.

[35] Czekanowski J. (1913): Zarys metod statystycznek. E. Wendego, Warsaw; glej tudi: Coefficient of racial likeness. Anthropol. Anz. 9 (1932), 227-249.

[36] Day N.E. (1969): Estimating the components of a mixture of normal distributions. Biometrika, 56, 463-474.

[37] Day W.H.E. (1986): Foreword: Comparison and consenus of classifications. Journal of Classification, 3, 183-186.

[38] DeSarbo W.S. in Mahajan V. (1984): Constrained classification: The use of a priori information in cluster analysis. Psychometrika, 49, 187-215.

[39] Diday E. (1974): Optimization in nonhierarchical clustering. Pattern Recognition, 6, 17-33.

[40] Diday E. in sodelavci (1979): Optimisation en classification automatique, Tome 1. in 2., Rocquencourt: INRIA.

[41] Diday E. (1986): New kinds of graphical representations in clustering. V: COMPSTAT 1986, Physica-Verlag, Heidelberg, 169-175.

[42] Dillon W.R. in Goldstein M. (1984): Multivariate Analysis: Methods and Applications. New York: Wiley.

[43] Driver H.E. in Kroeber A.L. (1932): Quantitative expression of cultural relationship. University of California, Publications in American Archaeology and Etnology 31, 211-256.

[44] Dubov Ju.A., Travkin S.I., Jakimec V.N. (1986): Mnogokriterial'nye modeli formirovanija i vybora variantov sistem. Moskva: Nauka.

[45] Dunn G. in Everitt B.S. (1982): An Introduction to Mathematical Taxonomy. Cambridge: Cambridge University Press.

[46] Duran B.S. in Odell P.L. (1974): Cluster Analysis - A Survey. Berlin: Springer.

[47] Edwards A.W.F. in Cavalli-Sforza L.L. (1965): A method for cluster analysis. Biometrics, 21, 362-375.

[48] Elisejeva I.I. in Rukavišnikov V.O. (1977): Gruppirovka, korrelacija, raspoznavanje obrazov. Moskva: Statistika.

[49] Estabrook G.F. in Rogers D.J. (1966): A general method of taxonomic description for a computed similarity measure. BioScience, 16, 789-793.

[50] Everitt B.S. (1974): Cluster analysis. London: Heinemann Educational Books.

[51] Everitt B.S. (1987): Introduction to Optimization Methods and their Application in Statistics. London: Chapman and Hall.

[52] Ferligoj A. (1981): Razvrščanje v skupine z omejitvami. Statistička revija, 31, 32-46.

[53] Ferligoj A. (1982): Razvrščanje v skupine (zapiski). Ljubljana: FSPN.

[54] Ferligoj A. (1982): Metode razvrščanja v skupine in možnosti njihove uporabe v demografskem raziskovanju. Ekonomska revija, 3-4, 531-540.

[55] Ferligoj A. (1983): Razvrščanje v skupine z omejitvami. Ljubljana: Ekonomska fakulteta Borisa Kidriča (doktorska disertacija).

[56] Ferligoj A. (1984): Clustering with constraints. Proceedings of 6 th International Symposium 'Computer at the University', Zagreb: SRCE.

[57] Ferligoj A. (1986): Clustering with constraining variable. Journal of Mathematical Sociology, 12, 299-313.

[58] Ferligoj A. (1987): Večkriterijsko razvrščanje v skupine. Zbornik IX. Medjunarodnog simpozija "Kompjuter na sveučilištu", SRCE: Zagreb.

[59] Ferligoj A. (1988): Razvoj in perspektive razvrščanja v skupine z omejitvami. Zbornik radova. Majski skup '87. Sekcije za klasifikacije Saveza statističkih društava Jugoslavije. Beograd: Savezni zavod za statistiku, 44-51.

[60] Ferligoj A. in Batagelj V. (1980): Taksonomske metode v družboslovnem raziskovanju. Poročilo RSS, Ljubljana: RIFSPN.

[61] Ferligoj A. in Batagelj V. (1982): Clustering with relational constraint. Psychometrika, 47, 413-426.

[62] Ferligoj A. in Batagelj V. (1983): Some types of clustering with relational constraint. Psychometrika, 48, 541-552.

[63] Ferligoj A. in Batagelj V. (1992): Direct multicriteria clustering algorithms. Journal of Classification, 9, 43-61.

[64] Ferligoj A. in Lapajne Z. (1986): Razvrščanje srednješolskih programov v skupine. Sodobna pedagogika, 38, 27-37.

[65] Ferligoj A. in Pompe-Kirn V. (1988): Reševanje problema majhnih populacijskih območij pri analizi incidence redkih bolezni z metodami razvrščanja v skupine z omejitvami. Statistička revija, 38, 21-31.

[66] Fischer M.M.(1980): Regional taxonomy. Regional Science and Urban Economics, 10, 503-537.

[67] Fleiss J.L. in Zubin J. (1969): On the methods and theory of clustering. Multivar. Behav. Res., 4, 235-250.

[68] Florek K. et al. (1951): Sur la liaison et la division des points d'un ensemble fini. Colloquium Mathematicum, 2, 282-285.

[69] Forgy E.W. (1965): Cluster analysis of multivariate data: efficiency versus interpretability of classification. Biometrics, 21, 768-769.

[70] Friedman H.P. in Rubin J. (1967): On some invariant criteria for grouping data. JASA, 62, 1159-1178.

[71] Fulgosi A. (1979): Faktorska analiza. Zagreb: Školska knjiga.

[72] Garey M.R. in Johnson D.S. (1979): Computers and intractability (A guide to the theory of NP-completness), San Francisco: Freeman.

[73] Gordon A.D. (1973): Classification in the presence of constraints. Biometrics, 2, 821-827.

[74] Gordon A.D. (1980): Methods of constrained classification. V: R. Tomassone (Ed.), Analyse de Donnee et Informatique. Le Chesnay: INRIA.

[75] Gordon A.D. (1981): Classification. London: Chapman and Hall.

[76] Gordon A.D. (1986): Links between clustering and assignment procedures. V: COMPSTAT 1986, Heidelberg: Physica-Verlag, 149-156.

[77] Gordon A.D. (1987 a): Classification and assignment in soil science. Soil use and management, 3, 3-8.

[78] Gordon A.D. (1987 b): A review of hierarchical classification. J.R.Statist.Soc. A, 150, 119-137.

[79] Gower J.C. (1967): A comparison of some methods of cluster analysis. Biometrics, 23, 623-638.

[80] Gower J.C. (1971): A general coefficient of similarity and some of its properties. Biometrics, 27, 857-872

[81] Gower J.C. (1985): Measure of similarity, dissimilarity, and distance. V: S. Kotz, N.L. Johnson in C.B. Read (Eds.): Encyclopedia of Statistical Sciences, Vol. 5, New York: Wiley, 397-405.

[82] Gower J.C. in Legendre P. (1986): Metric and Euclidean properties of dissimilarities coefficients. Journal of Classification, 3, 5-48.

[83] Guttman L. (1968): A general nonmetric technique for finding the smallest coordinate space for a configuration of points. Psychometrika, 33, 469-506.

[84] Hanani U. (1979): Multicriteria dynamic clustering. Rapport de Recherche No. 358, Rocquencourt: IRIA.

[85] Hartigan J.A. (1975): Cluster algorithms. New York: Wiley.

[86] Hartigan J.A. (1979): Distribution problems in clustering. V: J. Van Ryzin (Ed.): Classification and Clustering. New York: Academic Press.

[87] Hartigan J.A. (1985): Statistical theory in clustering. Journal of Classification, 2, 63-76.

[88] Hubert L. (1973): Min and Max hierarchical clustering using asymetric simmilarity measures. Psychometrika, 38, 63-72.

[89] Hubert L. in Arabie P. (1985): Comparing partitions. Journal of Classification, 2, 193-218.

[90] Homenjuk V.V. (1983): Elementi teorii mnogocelevoj optimizacii. Moskva: Nauka.

[91] Ivanović B. (1963): Diskriminaciona analiza sa primenom u ekonomskim istraživanjima. Beograd: Naučna knjiga.

[92] Ivanović B. (1971 a): Grupiranje zemalja u odnosu na njihov socio-ekonomski profil. Geneva: UNCTAD.

[93] Ivanović B. (1971 b): Analitička studija strukture izvoza zemalja u razvoju. Geneva: UNCTAD.

[94] Ivanović B. (1972): Klasifikacija skupa objekata prema stepenu sličnosti sa primenom u razradi tipologije zemalja prema njihovom socio-ekonomskom profilu. Statistička revija, 1-2.

[95] Ivanović B. (1976): Socio-ekonomski nivo i profil razvijenih zemalja i zemalja u razvoju u 1970. godini. Studije, analize i prikazi, št. 80, Beograd: Zvezni zavod za statistiko.

[96] Ivanović B. (1977): Teorija klasifikacije. Beograd: Institut za ekonomiku industrije.

[97] Ivanović B. (1981): Problemi statističkih selekcija kod odredjivanja grupe najslabijih elemenata jednog skupa. Statistička revija, 31, 47-59.

[98] Ivanović B. (1982): Primena I-korelacije u metodologiji I-odstojanja i nov način odredjivanja redosleda indikatora prema stepenu značajnosti. Statistička revija, 32.

[99] Ivanović B. (1988): Grupisanje obeležja preko metoda automatske klasifikacije. Statistička revija, 38, 11-20.

[100] Jaccard P. (1908): Nouvelles recherches sur la distribution florale. Bulletin de la Societe Vaudoise des Sciences Naturelles, 44, 223-270.

[101] Jambu M. (1978): Classification automatique pour l'analyse des donnes. Vol. 1 in 2, Pariz: Dunod.

[102] Jancey R.C. (1966): Multidimensional group analysis. Austral. J. Botany, 14, 127-130.

[103] Jardine N. in Sibson R. (1971): Mathematical Taxonomy. New York: Wiley.

[104] Jug J. (1988): Metoda voditeljev v programskih paketih CLUSE in SPSS/PC+. Metodološki zvezki, 3, Ljubljana: JUS, 92-105.

[105] Jug J. (1989): Programi za razvrščanje v skupine v programskih paketih CLUSE, SPSS in SAS. Zbornik radova Majski skup '89. Sekcije za klasifikacije Saveza statističkih društava Jugoslavije. Beograd: SZS (v tisku)

[106] Košmelj K. (1986): Two step procedure for clustering time varying data. Journal of Mathematical Sociology, 12.

[107] Košmelj K. (1987): Pregled pristopov za reševanje problema razvrščanja enot z upoštevanjem dodatne dimenzije. Statistička revija,

[108] Kruskal J.B. (1964 a in b): Multidimensional scaling. Psychometrika, 29, 1-27 in 115-129.

[109] Kulczynski S. (1927): Die Pflanzenassoziationen der Pieninen. Bulletin international de l'Academie polonaise des Sciences et des Lettres, Classe des Sciences mathematiques et naturelles, Serie B, Supplement II, 57-203.

[110] Lance G.N. in Williams W.T. (1966): Computer programs for hierarchical polythetic classification ('similarity analyses'). Comput. J., 9, 60-64.

[111] Lance G.N. in Williams W.T. (1967 a): Mixed-data classificatory programs. I. Agglomerative systems. Aust. Comput. J., 1, 15-20.

[112] Lance G.N. in Williams W.T. (1967 b): A general theory of classificatory sorting strategies. Comput. J., 9 (4), 373-380.

[113] Lambert J.M. in Williams W.T. (1962): Multivariate methods in plant ecology. IV. Nodal analysis. J. Ecol., 50, 775-802.

[114] Lambert J.M. in Williams W.T. (1966): Multivariate methods in plant ecology. VI. Comparison of information-analysis and association-analysis. J. Ecol., 54, 635-664.

[115] Lebart L. (1978): Programme d' Agregation avec Contraintes (CAH Contiguite). Les Cahiers d'Analyse des Donnees, 3, 275-287.

[116] Lefkovitch L.P. (1980): Conditional clustering. Biometrics, 36, 43-58.

[117] Lefkovitch L.P. (1985): Multi-criteria clustering in genotype - environment interaction problems. Theoretical and Applied Genetics, 70, 585-589.

[118] Legendre P. in Chodorowski A. (1977): A generalization of Jaccard's association coefficient for Q analysis of multi-state ecological data matrices. Efologia Polska, 25, 297-308.

[119] Lerman I.C. (1981): Classification et analyse ordinale des donnees. Pariz: Dunod.

[120] Liebetrau A.M. (1983): Measures of association. Series on Quantitative Applications in the Social Sciences, 32, Bevery Hills: Sage.

[121] Lorr M. (1983): Cluster Analysis for Social Scientists. San Francisco: Jossey-Bass.

[122] MacCrimmon K.R. (1973): An overview of multiple objective decision making. In: J.L. Cochrane and M. Zeleny (Eds.): Multiple Criteria Decision Making, S. Carolina: University of S. Carolina Press.

[123] Mac Queen J.B. (1967): Some methods of classification and analysis of multivariate observations. Proceedings of 5th Berkley Symposium, 1, 281-297.

[124] Mahalanobis P.C. (1936): On the generalized distance in statistics. Proc. Nat. Inst. Sci. India, 2, 49-55.

[125] Marcotorchino F. (1986): Cross association measures and optimal clustering. V: COMPSTAT 1986, Heidelberg: Physica-Verlag, 188-194.

[126] Marriott F.H.C. (1982): Optimization methods of cluster analysis. Biometrika, 69, 417-421.

[127] Matula D.W. (1977): Graph theoretic techniques for cluster analysis algorithms. V: J. Van Ryzin (ed.): Classification and Clustering. New York: Academic Press.

[128] Mc Quitty L.L. (1960): Hierarchical linkage analysis for the isolation of types. Educ. Psychol. Measur., 20, 55-67.

[129] Mc Quitty L.L. (1966): Similarity analysis by reciprocal pairs for discrete and continuous data. Educ. Psychol. Measur., 26, 825-831..

[130] Mc Quitty L.L. (1967): Expansion of similarity analysis by reciprocal pairs for discrete and continuous data. Educ. Psychol. Measur., 27, 253-255.

[131] Milligan G.W. in Cooper M.C. (1988): A study of standardization of variables in cluster analysis. Journal of Classification, 5, 181-204.

[132] Mills G.(1967): The determination of local government electoral boundaries. Operational Research Quarterly, 18, 243-255.

[133] Mirkin B.G. (1974): Problema gruppovogo vybora. Moskva: Nauka.

[134] Mirkin B.G. (1980): Analiz kačestvenyh priznakov i struktur. Moskva: Statistika.

[135] Mlinar Z. in Ferligoj A. (1977): Sličnosti i razlike u prostorno društvenim promenama. Sociologija, 2-3, 413-438.

[136] Mojena R. (1977): Hierarchical grouping methods and stopping rules: An evaluation. Computer Journal, 20 (4), 359-363.

[137] Momirović K. (1978): XTQ procedures for the determination of polar taxonomic variables. Proceedings of the INFORMATICA 78, Bled.

[138] Momirović K. (1986) COMTAX - Algoritam i program za detekciju i komaraciju polarnih i distinktnih taksona. Statistička revija, 3-4, 141-149.

[139] Momirović K. (1988): Uvod u analizu nominalnih varijabli. Metodološki zvezki, 2, Ljubljana: JUS.

[140] Momirović K. in Zakrajšek E. (1973): Odredjivanje taksonomskih skupina direktnom oblimin transformacijom ortogonaliziranih originalnih i latentnih varijabli. Kineziologija, 3 (1), 83-92.

[141] Murtagh F. (1985a): Multidimensional Clustering Algorithms. COMPSTAT Lectures 4, Dunaj: Physica-Verlag.

[142] Murtagh F. (1985b): A survey of algorithms for contiguity- constrained clustering and related problems. The Computer Journal, 28, 82-88.

[143] Pearson K. (1926): On the coefficient of racial likeness. Biometrika, 18, 105-117.

[144] Perruchet C. (1983): Constrained agglomerative hierarchical classification. Pattern Recognition, 16, 213-217.

[145] Podinovskij V.V. in Nogin V.D. (1982): Pareto-optimalnye rešenija mnogokriterialnyh zadač. Moskva: Nauka.

[146] Reynolds H.T. (1984): Analysis of nominal data. Series on Quantitative Applications in the Social Sciences, 7, Bevery Hills: Sage.

[147] Rogers D.J. in Tanimoto T.T. (1960): A computer program for classifying plants. Science, 132, 1115-1118.

[148] Rohlf F.J (1974): Hierarchical clustering using the minimum spanning tree. Comput. J., 16, 93-95.

[149] Romesburg H.C. (1984): Cluster Analysis for Researchers. Belmont: Lifetime Learning Publications.

[150] Rosenberg S. in Kim M.P. (1975): The method of sorting as a data-gathering procedure in multivariate research. Multivariate Behavioral Research, 10, 489-502.

[151] Russell F.F. in Rao T.R. (1940): On habitat and association of scecies of anopheline larvae in south-eastern Madras. J. Malaria Inst. India, 3, 153-178.

[152] Shepard R.N. (1962 a, b): The analysis of proximities multidimensional scaling with an unknown distance function. Psychometrika, 125-139 in 219-246

[153] Shamos M.I. (1976): Geometry and statistics: Problems at the interface. V: J.F. Traub (Ed.): Algorithms and complexity: New directions and recent results. New York: Academic Press.

[154] Sneath P.H.A. (1957): The application of computers to taxonomy. J. Gen. Microbiol., 17, 201-226.

[155] Sneath P.H.A. in Sokal R.R. (1973): Numerical taxonomy. San Francisco: Freeman.

[156] Sokal R.R. in Michener C.D. (1958): A statistical method for evaluatin systematic relationships. Univ. Kansas Sci. Bull., 38, 1409-1438.

[157] Sokal R.R. in Sneath P.H.A. (1963): Principles of numerical taxonomy. San Francisco: Freeman.

[158] Spath H. (1977): Cluster Analyse Algorithmen zur Objektklassifizierung und Datenreduktion. Munchen: R. Oldenbourg.

[159] Spath H. (1985): Cluster Dissection and Analysis. Chichester: Horwood.

[160] Spence N.A.(1968): A multifactor uniform regionalization of British counties on the basis of employment data for 1961. Regional Studies, 2, 87-104.

[161] Tryon R.C. (1939): Cluster Analysis: Correlation profile and orthometric (factor) analysis for the isolation of units in mind and personality. An Arbor: Edwards Brothers.

[162] Veledar E. in Kovalerchuk B.J. (1988): Vjerovatnostna interpretacija presjeka i unije fuzzy skupova kao osnova za klasifikaciju. luster Zbornik radova. Majski skup '88. Sekcije za klasifikacije Saveza statističkih društava Jugoslavije. Beograd: Savezni zavod za statistiku, 122-126.

[163] Ward J.H. (1963): Hierarchical grouping to optimize an objective function. JASA, 58, 236-244.

[164] Webster R. (1973): Automatic soil-boundary location from transect data. Mahematical Geology, 5, 27-37.

[165] Webster R. (1978): Optimally partitioning soil transects. Journal

[166] Webster R. in Burrough P.A. (1972): Computer-based soil mapping of small areas from sample data II Classification smoothing. Journal of Soil Science, 23, 222-234.

[167] Wishart D. (1969): Mode analysis: A generalization of nearest neighbour which reduces chaining effects (with Discussion). V: A.J. Cole (Ed.): Numerical Taxonomy. London: Academic Press.

[168] Wolfe J.H. (1970): Pattern clustering by multivariate mixture analysis. Multiv. Behavioral Research, 5, 329-350.

[169] Zagorujko N.G. (1972): Metodi raspoznavanja i ih primenenija. Moskva: Sovjetskoje radno.

[170] Zeleny M. (1974): The theory of the displaced ideal. In: M. Zeleny (Ed.): Multiple Criteria Decision Making, Springer- Verlag, New York, 153-206.

[171] Zubin J.A. (1938): A technique for measuring likemindedness. Journal of Abnormal and Social Psychology, 33, 508-516.

[172] Zupan J. (1982): Clustering of Large Data Sets. Letchworth: Research Studies Press.

[173] Zupan J. (1986): Hierarhično grupiranje velikih množic podatkov, Statistička revija, 3-4, 175-181.