Agrovoc descriptors: statistical methods, optimization methods, weather data, cartography

Agris category code: U10, A01, P40

COBISS koda 1.02

Metode za razvrščanje enot v skupine; osnove in primer

Katarina KOŠMELJ[1], Lidija BRESKVAR ŽAUCER[2]

Prispelo 3. maja 2006; sprejeto 10. avgusta 2006.
Received May 3, 2006; accepted August 10, 2006.

IZVLEČEK

Članek predstavlja osnove hierarhičnih in optimizacijskih metod za razvrščanje enot v skupine. Podana je ilustracija na enostavnem zgledu ter primer, ki se nanaša na razvrščanje meteoroloških postaj. Članek je namenjen študentom in raziskovalcem, ki se prvič srečujejo s to problematiko.

Ključne besede: hierarhične metode razvrščanja, optimizacijske metode razvrščanja

ABSTRACT

METHODS FOR CLUSTER ANALYSIS; INTRODUCTION AND A CASE STUDY

The paper deals with cluster analysis. It is focused on hierarchical and optimization clustering methods. A simple example is worked-out for illustration, a case-study on 19 meteorological stations in Slovenia is presented. The paper presents elementary ideas and is meant for beginners.

Key words: hierarhical clustering methods, optimization clustering methods

1 UVOD

Namen razvrščanja enot v skupine je uvrstiti enote v skupine po principu podobnosti tako, da so znotraj posamezne skupine enote, ki so si glede na vnaprej določen kriterij podobne, znotraj različnih skupin pa enote, ki so si glede na ta kriterij različne. Vsaka enota je uvrščena v samo eno skupino; torej se skupine ne prikrivajo.

Postopek razvrščanja v skupine ima naslednje korake:

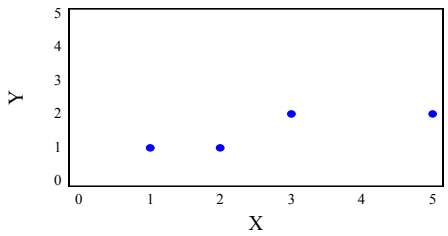
1. Izbira enot, izbira njihovih lastnosti.
2. Standardizacija spremenljivk, če je le-ta potrebna.
3. Izbira ustrezne razdalje (različnosti) med enotami. Ta je odvisna od vrste podatkov in od kriterija podobnosti.

[1] red. prof., dr. znan., SI-1111 Ljubljana, Jamnikarjeva 101, p.p 2995
[2] mlada raziskovalka, univ. dipl. inž. kraj. arh., SI-1111 Ljubljana, Jamnikarjeva 101, p.p 2995

4. Uporaba različnih metod razvrščanja.
5. Analiza rezultatov.

Namen tega članka je predstavitev osnov za metode razvrščanja. Za ilustracijo bomo uporabili enostaven primer: imamo 4 enote označene A, B, C in D, za vsako enoto podatek za lastnost X in za lastnost Y. Lastnosti sta primerljivi, standardizacija ni potrebna.

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



Slika 1. Podatki in njihov grafični prikaz.
Figure 1. Data and their graphical representation.

Na osnovi podatkov izračunamo matriko razdalj, ki predstavlja vhod za metode razvrščanja. Matrika je simetrična in ima na diagonali ničle, zato bomo pisali samo njen zgornji trikotnik. Za naš ilustrativni primer bomo uporabili kvadrat Evklidske razdalje.

	A	B	C	D
A		1	5	17
B			2	10
C				4
D				

2 METODE RAZVRŠČANJA

Kratko bomo predstavili dve skupini metod razvrščanja, hierarhične metode združevanja in optimizacijske metode.

2.1 HIERARHIČNE METODE ZDRUŽEVANJA

Pri hierarhičnih metodah združevanja je postopek naslednji:

- Vsaka enota je skupina.
- V vsakem koraku združevanja se poišče med seboj najbližji skupini in se ju združi v novo skupino. Staro matriko razdalj nadomesti nova matrika razdalj. Ta korak se ponavlja, dokler niso vse enote združene v eno skupino.

Različne metode združevanja izračunajo razdaljo med novima skupinama na različne načine. Poglejmo postopek združevanja pri minimalni in maksimalni metodi.

Minimalna metoda (Single Linkage, Nearest Neighbor)

Najmanjša razdalja je med enotama A in B, ti dve enoti se združita v skupino [A,B], nivo združevanja je 1. Razdalja med že nastalo skupino [A,B] in C ter med [A,B] in D se izračuna po *pravilu najbližjega soseda= minimalna razdalja.*

d([A,B], C)=min (5; 2) = 2
d([A,B], D)=min (17; 10) = 10

	[A,B]	C	D
[A,B]		2	10
C			4
D			

Najmanjša razdalja je med skupinama [A,B] in C, ki se združita v novo skupino [[A,B], C] . Raven združevanja je 2. Nova matrika razdalj je:

	[[A,B],C]	D
[[A,B],C]		4
D		

D se pridruži skupini [[A,B], C]. Raven združevanja je 4.

Maksimalna metoda (Complete Linkage, Furtest Neighbor)

Najmanjša je razdalja med enotama A in B, ti dve enoti se združita v skupino [A,B], nivo združevanja je 1. Razdalja med že nastalo skupino [A,B] in C ter med [A,B] in D se izračuna po *pravilu najbolj oddaljenega soseda= maksimalna razdalja.*

d([A,B], C)=max(5; 2) = 5
d([A,B], D)=max(17; 10) = 17

	[A,B]	C	D
[A,B]		5	17
C			4
D			

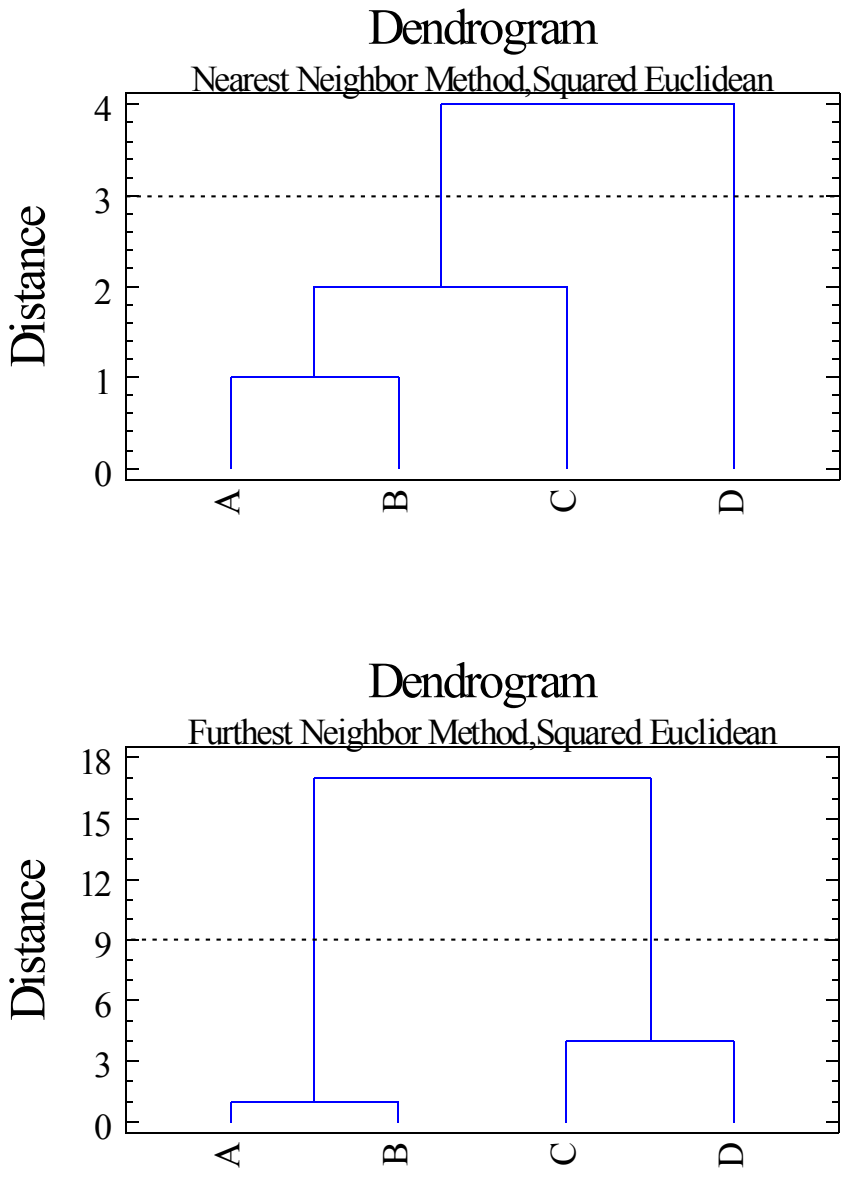
Najmanjša razdalja je med C in D, ki se združita v novo skupino [C, D]. Raven združevanja je 4. Nova matrika razdalj ima obliko:

	[A,B]	[C,D]
[A,B]		17
[C,D]		

Skupini [A,B] ter [C,D] se združita v eno skupino na nivoju 17.

V veliko pomoč pri analizi rezultatov je grafični prikazi združevanja, ki se imenuje *dendrogram*, saj omogoča vpogled v hierarhično strukturo in pomaga pri določanju

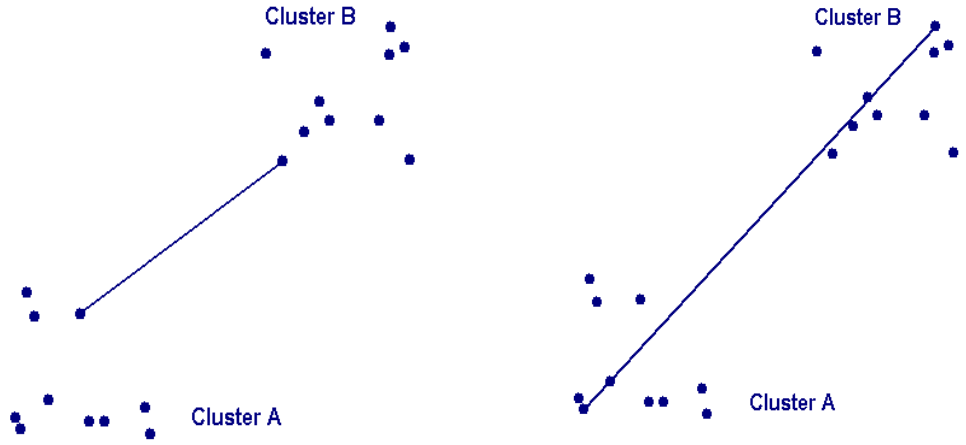
smiselnega števila skupin. Kjer je »skok« pri združevanju največji, drevo prerežemo in iz dendrograma odčitamo dobljene skupine. Poglejmo dendrogram za minimalno in za maksimalno metodo, njuna oblika je različna. Za minimalno metodo je največji skok med 2 in 4. Rez na tem intervalu generira skupini [[A,B],C,] ter [D]. Maksimalna metoda ima največji skok med 4 in 17, dobljeni skupini sta [A,B] ter [C,D].



Slika 2. Dendrogram za minimalno metodo (zgoraj) in za maksimalno metodo (spodaj).
Figure 2: Dendrogram obtained with Minimum method (above) and Maximum method (below).

Nivoji združevanja skupin pri različnih metodah so različni. Primerjava načina izračuna razdalje med skupino A in skupino B pri minimalni in maksimalni je

grafično razvidna iz slike 3. Minimalna metoda izračuna razdaljo med skupinama kot razdaljo med najbližjima dvema enotama, ki sta v teh dveh skupinah, maksimalna metoda pa med najbolj oddaljenima enotama.



Slika 3. Primerjava načina izračuna razdalje med skupino A in skupino B pri minimalni in maksimalni metodi.
Figure 3. Distance between Cluster A and Cluster B obtained with Minimum Method (left) and Maximum Method (right).

V skupini hierarhičnih metod združevanja je še nekaj metod, ki razdaljo med skupinama izračunajo na različne »kompromisne« načine. *Metoda težišč (Centroid method, Gower method)* računa razdaljo med skupinama glede na položaj njunih težišč; *povprečna metoda (Group Average method)* kot povprečje razdalj med vsemi enotami v skupini, itd.

Posebna metoda v teh skupini je *Wardova metoda.* Razdalja med skupinama se vrednoti z »izgubo informacije«, ki jo povzroča združevanje dveh skupin v novo skupino. Pri tem se poveča vsota kvadriranih odklonov znotraj skupin (VKO), to je količina, ki jo poznamo iz analize variance. Za skupino S se ta količina izračuna kot vsota kvadriranih odklonov od povprečja skupine S:

$$VKO(S) = \sum_{i \in S} (x_i - \bar{x}_S)^2$$

Ward je kot kriterij za vrednotenje posamezne razvrstitve upošteval *VKO* za pripadajočo razvrstitev; če je spremenljivk več, se vrednosti za *VKO* po spremenljivkah seštejejo. Za ilustracijo izračunajmo vrednosti tega kriterija za razvrstitev, v kateri sta A in B združeni v skupino [A, B], enoti C in D pa v skupino [C, D]. Povprečje skupine [A, B] za X je 1,5 in za Y je 1, povprečje skupine [C, D] za X je 4 in za Y je 2.

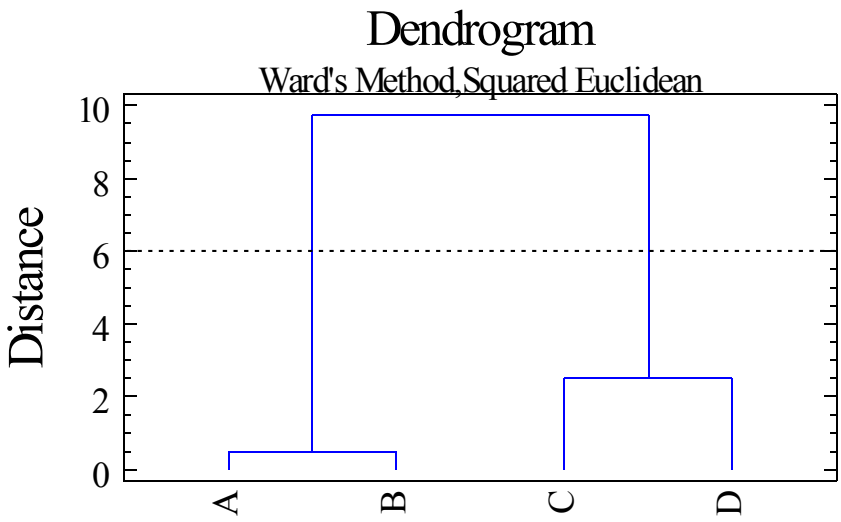
Za X:

$VKO([A,B]) + VKO([C,D]) = (1-1{,}5)^2 + (2-1{,}5)^2 + (3-4)^2 + (5-4)^2 = 2{,}5$

Za Y:

$VKO([A,B]) + VKO([C,D]) = (1-1)^2 + (1-1)^2 + (2-2)^2 + (2-2)^2 = 0$

Za predstavljeni ilustrativni primer velja, da se A in B prva združita v skupino na nivoju 0,5; nato se C in D združita na nivoju 2,5; vse enote se združijo v eno skupino na nivoju 9,75. Rezultati so razvidni iz dendrograma. Najbolj smiselno je dendrogram razrezati tako, da dobimo skupini [A,B] ter [C,D].



Slika 4. Dendrogram za Wardovo metodo.
Figure 4. Dendrogram obtained with Ward's method.

V splošnem minimalna metoda generira zelo razvlečene skupine, pogosto je težko določiti smiselno število skupin. Maksimalna metoda generira sferične skupine. Slabost metode težišč in povprečne metode je možno nemonotono združevanje; nivo združevanja skupin na npr. tretjem koraku je nižje kot na drugem koraku. Posledica tega je »zveriženo« drevo, ki ne daje jasne slike hierarhije.Wardova metoda teži k skupinam, ki imajo primerljivo variabilnost.

Hierarhične metode imajo izjemno lastnost vizualizacije rezultatov v obliki dendrograma, ki je zlahka razumljiv. Njihova slabost pa je dejstvo, da so primerne za razvrščanje do nekaj 100 enot in da dobljene razvrstitve niso nujno optimalne. Najbolj pogosto uporabljene metode so minimalna, maksimalna in Wardova, ki lahko dajejo različne rešitve.

2.2 OPTIMIZACIJSKE METODE

V tej skupini je več metod, najbolj pogosto uporabljeni sta *metoda voditeljev* (*K-means*) in *metoda prestavljanj (Relocation method).* V vsakem koraku se izračuna določeno kriterijsko funkcijo in poskuša z drugačno razvrstitvijo doseči, da se zmanjša vrednost kriterijske funkcije. Optimizacijske metode izhajajo iz vnaprej določenega števila skupin *k*, to število mora določiti uporabnik. Na kratko bomo opisali metodo voditeljev.

Metoda voditeljev (K-means)

Računalnik izbere *k* točk slučajno, te točke so začetni voditelji. Vsako enoto doda najbližjemu voditelju. Ko so vse enote dodane in s tem narejene nove skupine, se izračuna težišča novo-nastalih skupin, ki postanejo novi voditelji.Ta postopek se ponavlja, dokler se voditelji ne premikajo več. Tedaj se izračuna vrednost iste kriterijske funkcije kot pri Wardovi metodi. Dobljena razvrstitev je odvisna od začetne pozicije voditeljev, zato se ta postopek velikokrat ponavlja z različnimi začetnimi konfiguracijami voditeljev. Kot "najboljšo" razvrstitev se vzame tisto, ki ima najmanjšo vrednost kriterijske funkcije.

Wardova metoda in metoda voditeljev se dopolnjujeta; če uporabljamo kvadrat Evklidske razdalje, optimirata isto kriterijsko funkcijo. Najbolje je, da najprej uporabimo Wardovo metodo, iz dendrograma ugotovimo najbolj smiselno število skupin in nato uporabimo metodo voditeljev.

3 PRIMER UPORABE METOD RAZVRŠČANJA ENOT V SKUPINE

Iščemo razvrstitev za 19 meteoroloških postaj v Sloveniji, ki so opisane s 4 klimatskimi spremenljivkami: povprečna letna maksimalna temperatura zraka (*temp_pmax*), povprečna letna minimalna temperatura zraka (*temp_pmin*), povprečna letna količina padavin (*padavine*) in povprečno število dni s snežno odejo ob 7. uri na leto (*dni_sneg*) (Vir: Agencija RS za okolje, http://www.arso.gov.si/podro~cja/vreme_in_podnebje/napovedi_in_podatki/podneb_30_tabele.html). Podatki veljajo za 30-letno obdobje, od leta 1961 do leta 1990, in so prikazani v tabeli 1.

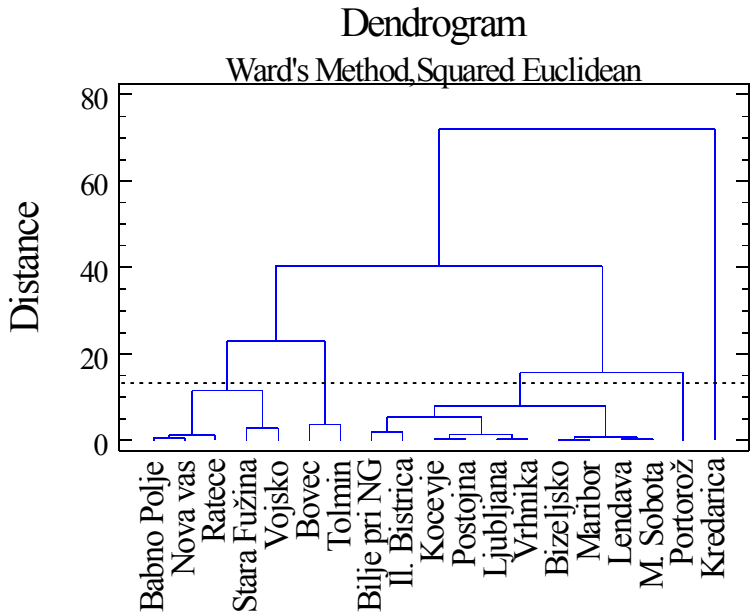
Tabela 1. Klimatski podatki za 19 meteoroloških postaj.
Table 1. Climate data for 19 meteorological stations.

Ime postaje (Name of the station)	temp_pmax (°C)	temp_pmin (°C)	padavine (mm)	dni_sneg
Babno Polje	12,1	0,1	1662	91,3
Bilje pri Novi Gorici	17,9	6,2	1456	1,8
Bizeljsko	15,2	4,9	1059	42,1
Bovec	14,5	4,5	2733	62,6
Ilirska Bistrica	15,5	4,4	1448	19,8
Kočevje	14,0	3,3	1523	73,9
Kredarica	1,2	-4,2	1993	265,1
Lendava	15,2	5,3	805	36,1
Ljubljana	14,8	5,5	1393	64,9
Maribor	14,7	5,2	1045	58,6
Murska Sobota	14,5	4,1	814	46,7
Nova vas na Blokah	12,3	1,6	1472	94,6
Portorož	16,6	10,8	1046	3,0
Postojna	13,4	3,9	1578	47,0
Rateče	11,9	0,7	1563	132,2
Stara Fužina	13,6	2,8	2333	109,7
Tolmin	16,4	5,8	2246	20,1
Vojsko	9,7	3,0	2456	136,5
Vrhnika	14,6	4,9	1594	63,9

Klimatske spremenljivke so bile najprej standardizirane, nato pa je sledilo razvrščanje meteoroloških postaj z uporabo hierarhične Wardove metode in optimizacijske metode voditeljev.

Wardova metoda

Razdaljo med enotami smo merili s kvadratom Evklidske razdalje. Drevo združevanja z Wardovo metodo prikazuje slika 5. Ta kaže, da ena skupina (v njej je le Kredarica) močno odstopa od vseh ostalih postaj, le-te pa se razvrstijo v dve večji skupini. V prvi je 11 postaj (Portorož, Bizeljsko, Maribor, Lendava, Murska Sobota, Ljubljana, Vrhnika, Kočevje, Postojna, Bilje pri Novi Gorici in Ilirska Bistrica), v drugi pa 7 (Bovec, Tolmin, Rateče, Babno Polje, Nova vas na Blokah, Stara Fužina in Vojsko).



Slika 5. Dendrogram za 19 meteoroloških postaj v Sloveniji – Wardova metoda.
Figure 5. Dendrogram for 19 meteorological stations in Slovenia – Ward's method.

Z metodološkega stališča bi bila najbolj zanimiva razvrstitev meteoroloških postaj v 2 ali 3 skupine, kjer je skok v drevesu združevanja največji. Vendar se pokaže kot vsebinsko bolj zanimiva razvrstitev v 5 skupin, prikazanih v tabeli 2 (glej rez dendrograma, prikazan na sliki 5)

Tabela 2. Razvrstitev meteoroloških postaj v 5 skupin – Wardova metoda.
Table 2. Five clusters of meteorological stations – Ward's method.

Skupina 1	Skupina 2	Skupina 3	Skupina 4	Skupina 5
Kredarica.	Portorož.	Bizeljsko, Maribor, Lendava, Murska Sobota, Ljubljana, Vrhnika, Kočevje, Postojna, Bilje pri NG, Ilirska Bistrica.	Bovec, Tolmin.	Rateče, Babno Polje, Nova vas na Blokah, Stara Fužina, Vojsko.

V skupino 1 je razvrščena Kredarica. Med vsemi postajami ima najnižji povprečni letni temperaturi ter najdaljšo snežno odejo, zanjo pa je značilna tudi relativno velika količina padavin.

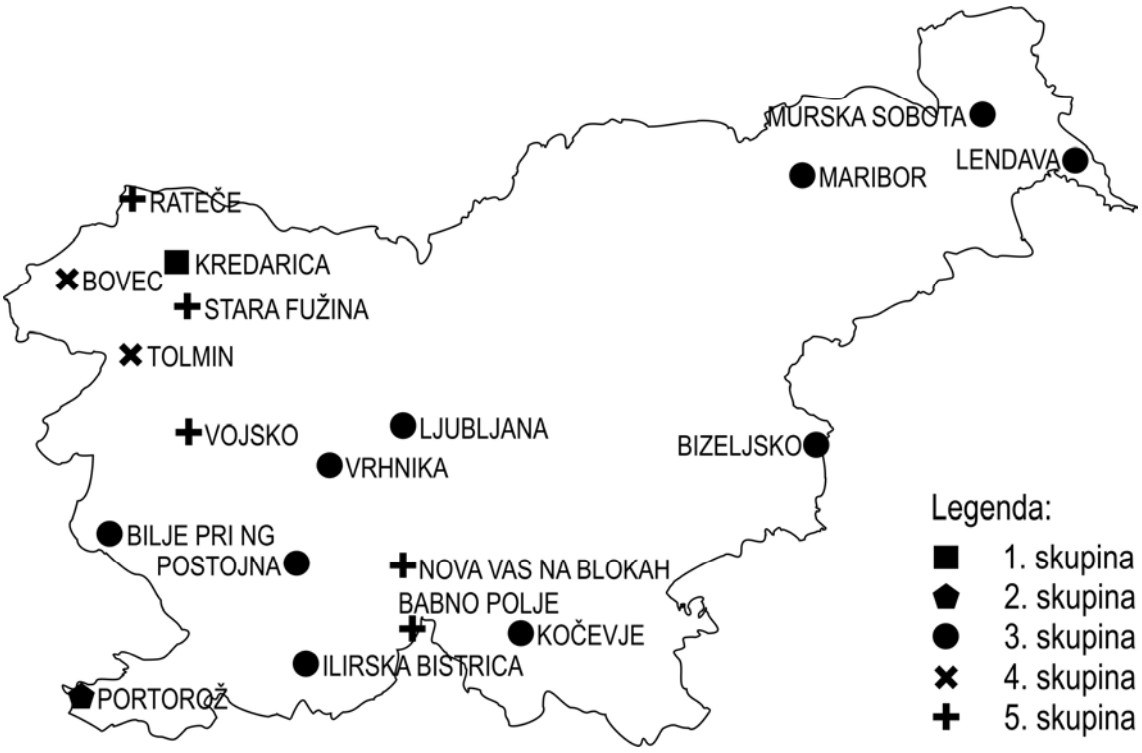
V skupino 2 je razvrščen Portorož, ki ima glede na ostale postaje najvišjo povprečno letno minimalno temperaturo, visoko povprečno letno maksimalno temperaturo, majhno količino padavin in je skoraj brez snežne odeje – v povprečju le 3 dni na leto.

V skupino 3 so razvrščeni Bizeljsko, Maribor, Lendava, Murska Sobota, Ljubljana, Vrhnika, Kočevje, Postojna, Bilje pri Novi Gorici in Ilirska Bistrica. Njihove klimatske značilnosti so podobne značilnostim 2. skupine, le da so manj izrazite. Zanje sta značilni visoki povprečni letni temperaturi, relativno skromna količina padavin ter majhno število dni s snežno odejo na leto.

V skupino 4 sta razvrščena Bovec in Tolmin. Najmočneje ju označuje velika količina padavin. Sicer sta zanju značilni tudi srednji povprečni letni temperaturi ter majhno število dni s snežno odejo na leto.

V skupino 5 pa so razvrščeni Rateče, Babno polje, Nova vas na Blokah, Stara Fužina in Vojsko. Z izjemo prve skupine je to skupina z najnižjimi povprečnimi letnimi temperaturami in najdaljšo snežno odejo, zanje pa so značilne tudi obilne padavine.

Prostorski prikaz razvrstitve, ki je bila dobljena z Wardovo metodo, je prikazan na sliki 6.



Slika 6. Kartogram za 19 meteoroloških postaj razvrščenih v 5 skupin – Wardova metoda.
Figure 6. Map of 19 meteorological stations clustered in 5 groups – Ward's method.

Metoda voditeljev (K-means)

Metoda voditeljev izhaja iz vnaprej določenega števila skupin, v katere je potrebno razvrstiti meteorološke postaje. V skladu z rezanjem dendrograma po Wardovi metodi smo privzeli 5 skupin. Postopek razvrščanja z izbrano metodo je bil ponovljen 1000 krat, vsakokrat z naključno začetno konfiguracijo voditeljev. Najboljša končna razvrstitev meteoroloških postaj v 5 skupin z metodo voditeljev je podana v tabeli 3.

Tabela 3. Razvrstitev meteoroloških postaj v 5 skupin – metoda voditeljev.
Table 3. Five clusters of meteorological stations – K-means method.

Skupina 1	Skupina 2	Skupina 3	Skupina 4	Skupina 5
Kredarica.	Bilje pri NG, Portorož.	Bizeljsko, Maribor, Lendava, Murska Sobota, Ljubljana, Vrhnika, Kočevje, Postojna, Ilirska Bistrica.	Bovec, Tolmin.	Rateče, Babno Polje, Nova vas na Blokah, Stara Fužina, Vojsko.

Razvrstitev meteoroloških postaj z metodo voditeljev je podobna razvrstitvi z Wardovo metodo. Edina razlika med njima se tiče razvrstitve postaje Bilje pri Novi

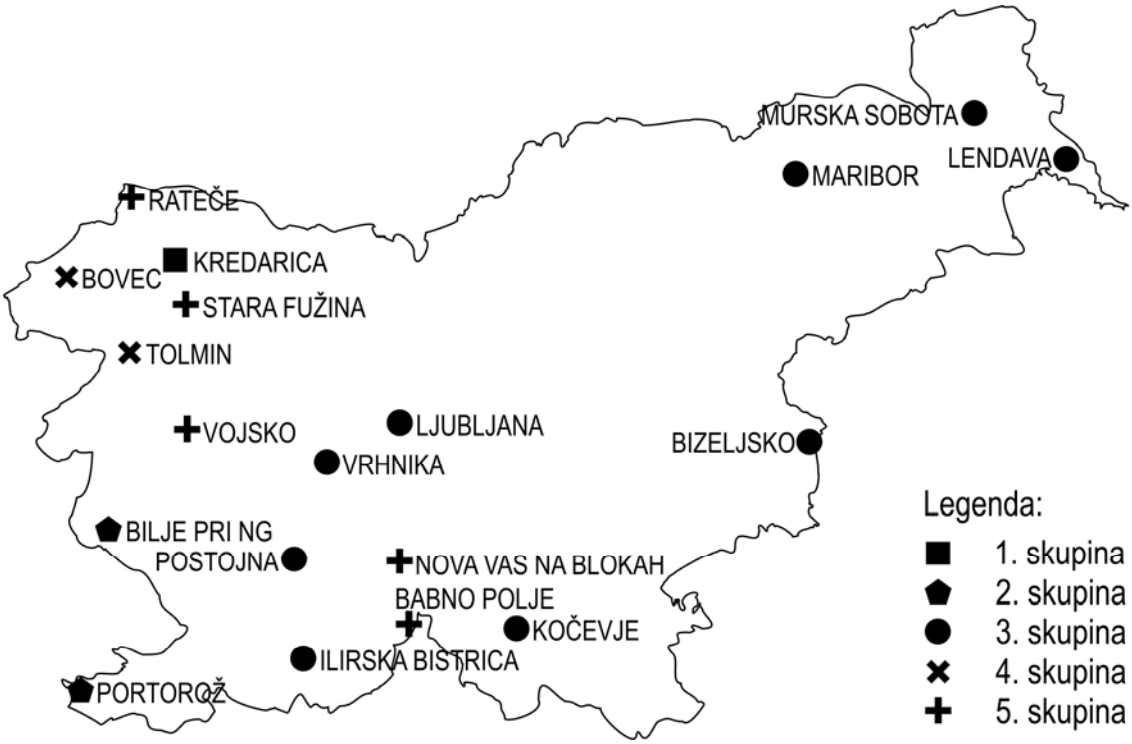
Gorici, ki je z Wardovo metodo razvrščena v 3. skupino, z metodo voditeljev pa v 2. skupino.

Razlog je mogoče iskati v značilnosti hierarhičnih metod. Enota, ki je uvrščena v neko skupino na določenem koraku, v vseh naslednjih korakih v tej (pod)skupini ostane. Hierarhične metode ne omogočajo kasnejših popravkov, saj se hierarhija ohranja. Tako je bila z Wardovo metodo postaja Bilje pri Novi Gorici razvrščena v skupino skupaj z Ilirsko Bistrico, s katero imata podobne značilnosti, predvsem količino padavin.

Analiza istih podatkov z optimizacijsko metodo voditeljev pa pokaže, da bi bila bojša razvrstitev Bilja pri Novi Gorici v 2. skupino, skupaj s Portorožem. Večkratno razvrščanje z metodo voditeljev z različnim številom ponovitev postopka je vsakokrat pokazalo isto razvrstitev, po čemer sklepamo, da je razvrstitev stabilna. Pogled v značilnosti Bilja pri Novi Gorici in Portoroža pokaže, da imata postaji več skupnih značilnosti kot Bilje pri Novi Gorici in Ilirska Bistrica. Napram ostalim postajam imata obe najvišji povprečni letni temperaturi ter najmanjše povprečno število dni s snežno odejo na leto. Sklepamo, da je razvrstitev z metodo voditeljev boljša od razvrstitve z Wardovo metodo.

Oceno stabilnosti razvrstitve enot s hierarhičnimi metodami je zato priporočljivo preveriti na ta način, da podatke analiziramo še z drugimi optimizacijskimi metodami razvrščanja v skupine.

Prostorski prikaz z metodo voditeljev razvrščenih meteoroloških postaj je prikazan na sliki 7.



Slika 7. Kartogram za 19 meteoroloških postaj razvrščenih v 5 skupin – metoda voditeljev.

Figure 7. Map of 19 meteorological stations clustered in 5 groups – K-means method.

4 LITERATURA

Chatfield C., Collins A. J., 1980: Introduction to Multivariate Analysis. Chapman and Hall, New York, 246 str.

Daly F., Hand D. J., Jones M. C., Lunn A. D., McConway K. J., 1995: Elements of Statistics. Addison-Wesley Publishing Company, 682 str.

Ferligoj A., 1989: Razvrščanje v skupine: teorija in uporaba v družboslovju. Fakulteta za sociologijo, politične vede in novinarstvo, Raziskovalni inštitut, Ljubljana, 182 str.

Ferligoj A., 2002: Cluster Analysis. School of Biometrics, BIOSTAT 2002, Cavtat, Croatia

Krzanowski W. J., 1988: Principles of Multivariate Analysis, A User's Perspective. Clarendon Press, Oxford, 563 str.

A Tutorial on Clustering Algorithms. http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/AppletKM.html. (20.8.2006)