POSKUS DOLOČITVE KLIME S FAKTORSKO ANALIZO IN TAKSONOMIJO

AN ATTEMPT TO DETERMINE THE CLIMATE BY USING FACTOR ANALYSIS AND TAXONOMY

551.582.2

JOŽE ROŠKAR

Meteorološki zavod SRS, Ljubljana

SUMMARY

The idea of this attempt was to determine the climatic regions, by taking into consideration all the data available at a certain measurement point. Weather parameters, observed or measured at a certain station, represent a vector, where the number of coordinates is the same as the number of parameters which are being measured. As the subject in question is the climate, every weather parameter is a time function itself, which means that we are deling with a vector, the coordinates of which are time functions. In this attempt, standard climatological derived data (mean values, etc.) were taken as coordinates. We are trying to solve the problem by comparing several vectors of the kind, each of them representing one measurement point. Such points are usually called stations. Vectors can be represented as points in a space, the dimensions of which are the number of coordinates, thus the number of considered values of weather parameters. The basis for classification are Euclidean distances, measured among separate points. The smaller is the distance between the two points, the greater climatological accordance exists between the stations, represented by the two points. In our attempt, the number of coordinates was reduced by using factor analysis. The methods of taxonomy were used for classification.

POVZETEK

Ideja poskusa je določitev klimatskih področij, če pri tem upoštevamo vse podatke, s katerimi razpolagamo na neki merilni točki. Parametri vremena, opazovani ali merjeni na neki postaji, predstavljajo vektor s toliko koordinatami kolikor parametrov vremena merimo. Ker govorimo o klimi, je vsak parameter vremena zase časovna funkcija, kar pomeni, da imamo opraviti z vektorjem, katerega koordinate so časovne funkcije. V tem poskusu smo kot koordinate upoštevali standardne klimatološke izvedene podatke (povprečne vrednosti itd.). Problem poskušamo rešiti

tako, da primerjamo več takih vektorjev, kjer vsak predstavlja eno merilno točko. Take točke ponavadi imenujemo kar postaje. Vektorje lahko postavimo kot točke v prostoru, katerega dimenzija je število koordinat, torej število upoštevanih vrednosti parametrov vremena. Evklidske razdalje, izračunane med posameznimi točkami so osnova za klasifikacijo. Čim manjša je razdalja med dvema točkama, tem bolj se postaji, ki ju točki predstavljata, klimatološko ujemata. V našem poskusu smo število koordinat zmanjšali s pomočjo faktorske analize. Za klasifikacijo pa smo uporabili metode taksonomije.

UVOD

Klimo na nekem področju lahko definiramo kot skupek vremenskih dogajanj v prizemni plasti nad istim področjem čez daljši čas. Klasična klimatologija jo opisuje s preprostimi izvedenimi vrednost parametrov vremena, največkrat ka z aritmetičnim popreċjem (npr. temperatura) ali poprečnimi vsotami (npr. padavine). S takimi podatki so poskušali klasificirati geografska področja v tako imenovane klimatske rejone glede na vsak opazovani parameter vremena zase. Z rezultati, ponavadi s kartami, so potem poskušali bolj ali manj uspešno dobiti sintezo. Pri takem načinu dela pa naredimo dve očitni napaki. Najprej se nam izgubijo karakteristike časa, saj je vsak parameter vremena, ki ga opazujemo ali merimo, časovna funkcija, ki je samo z aritmetičnim poprečjem ali podobnim pokazalcem ne moremo prikazati. Druga napaka, ki jo naredimo, je v tem, da ne upoštevamo soodvisnosti parametrov vremena, ki jih upoštevamo pri prikazu klime nekega področja. Parametri, ki opisujejo stanje atmosfere, so nedvomno med seboj odvisni. Zaradi tega tudi klime ne moremo opisovati, ne da bi upoštevali medsebojno odvisnost parametrov vremena.

Zaradi tega poskušamo v pričujočem članku obravnavati objektivno opisovanje klime na osnovi čim več razpoložljivih parametrov vremena hkrati, pri čemer upoštevamo njihovo medsebojno odvisnost. V tem poskusu nismo upoštevali časa, kot je to primer v analizi časovnih vrst (markovske verige, spektralna analiza časovnih vrst), ampak smo čas površno zajeli s podatki, kot na primer: število dni z maksimalno temperaturo večjo od 25^{o}C na časovno enoto itd. Zaradi tega bomo vsekakor dobili nekoliko popačeno sliko. Upoštevali pa smo medsebojno odvisnost parametrov vremena, ki smo jih upoštevali. Ne pričakujemo že kar uporabnih rezultatov; poskus želi prikazati objektivno metodo za prikaz klimatskih razmer pri obravnavanju klime. Dobiti želimo objektivni prikaz klimatskih razmer za določeno področje, kjer naj bi bili subjektivni vplivi popolnoma izključeni. Pokazala se bo še ena ugodnost. Ves čas analize bo prisotna metrika, kar pomeni, da lahko rezultate med seboj primerjamo v smislu metrike, torej merimo.

Za poskus oziroma prikaz metode smo izbrali izvedene meteorološke podatke za poletni čas, to je za mesece od maja do septembra za petletno obdobje 1971-1975. Obravnavali smo 50 enakomerno porazdeljenih postaj po celi Sloveniji. Vsekakor rezultati ne bodo dali tega, kot če bi parametre obravnavali kot časovne funkcije, pa tudi petletno obdobje je prekratko za neki določen rezultat; vsekakor pa je vidna uporabnost metode.

OPIS METODE

Denimo, da v nekem kraju merimo N parametrov vremena, oziroma da lahko iz meritev dobimo čez daljši čas N izvedenih vrednosti parametrov vremena. Tako merilno točno lahko definiramo kot vektor z N koordinatami. Zapišemo ga lahko

$$X = (x_1, x_2, \ldots, x_N),$$

kjer predstavlja vsaka od koordinat neko izvedeno vrednost parametra vremena. Denimo še, da imamo M krajev, kjer merimo. Tedaj lahko za neki kraj zapišemo vektor

$$X_i = (x_1^i, x_2^i, \ldots, x_N^i); \tag{1}$$

pri tem predpostavljamo, da imamo v vseh krajih enako število izvedenih vrednosti parametrov vremena. Predpostavimo še, da predstavlja vsak od opisanih vektorjev stanje atmosfere v prizemni plasti za daljši čas, torej opisuje klimo v izbrani točki. Grupirati želimo tiste kraje, ki imajo podobno klimo in tako določiti klimatska področja. Med informacije, zajete v vektorju (1) lahko štejemo tudi podatke o geografski legi kraja, kjer merimo.

Vektorji, definirani z izrazom (1) naj predstavljajo točke v N-dimenzionalnem prostoru. Denimo, da v tem prostoru lahko definiramo evklidsko razdaljo med vektorjema X_i in X_j:

$$d_{ij}^2 = \sum_{k=1}^{N} (x_k^j - x_k^i)^2 \tag{2}$$

Rečemo lahko, da sta si kraja, ki ju predstavljata vektorja X_i in X_j klimatsko tem bolj podobna, kolikor manjša je razdalja, definirana z izrazom (2). Koordinate M vektorjev sestavljajo matriko velikosti MxN, razdalje med njimi pa simetrično matriko velikosti MxM. Problem določitve klimatskih področij smo tako prevedli na problem klasifikacije, kjer je vhodni podatek matrika razdalj.

Poglejmo bliže matriko, ki jo sestavljajo koordinate vektorjev, definiranih z izrazom (1). Vrstice v tej matriki, imenujmo jo A, predstavljajo posamezne vektorje, stolpci pa izbrane izvedene vrednosti parametrov vremena. Rečemo lahko, da so vrstice posamezna opazovanja v našem poskusu, stolpci pa posamezne slučajne spremenljivke opazovanega slučajnega vektorja. Kadar je velikost matrike A velika, torej kadar obravnavamo večje število izvedenih vrednosti parametrov vremena na večjem številu mernih točk, je smiselno, da število stolpcev reduciramo in tako zmanjšamo velikost matrike A. Več slučajnih spremenljivk, ki jih nastopa v poskusu, ima lahko namreč neke skupne lastnosti. Statistična metoda, ki jo lahko uporabimo za redukcijo večjega števila spremenljivk v manjše število faktorjev, je faktorska analiza /2/. Faktor je neodvisen izvor variacije. Na ta način razdelimo karakteristike oziroma informacije, ki jih vsebujejo spremenljivke, na neodvisne izvore variacije, ki jih je manj ali toliko kot osnovnih spremenljivk. Faktorska analiza razdeli torej spremenljivke na skupine medsebojno nekoreliranih spremenljivk. Ker so slučajne spremenljivke, ki jih študiramo v našem problemu, zelo verjetno medsebojno korelirane, jih lahko reduciramo v manjše število spremenljivk oziroma faktorjev. Če se izrazimo matematično, je vsak faktor neka linearna kombinacija osnovnih spremenljivk.

Denimo sedaj, da imamo $n < N$ faktorjev. Število faktorjev bomo določili tako, da vsi faktorji zajemajo najmanj 95% celotne variance. Opazovane slučajne vektorje, definirane z izrazom (1), projiciramo v prostor, ki ga razpenjajo izbrani faktorji. To je n-dimenzionalen prostor, kjer lahko definiramo kvadrat razdalje analogno razdalji, definirani z izrazom (2):

$$\rho_{ij}^2 = \sum_{k=1}^{n} (f_k^j - f_k^i)^2 \tag{3}$$

ρ_{ij} je potem evklidska razdalja med projekcijama vektorjev X_i in X_j. Tako dobljeno matriko razdalj (imenujmo jo D) smo uporabili kot vhodni podatek za klasifikacijo. Uporabili smo taksonomsko metodo oziroma t.im. "clustering tehniko" /3/. Pri tem se je zopet potrdila potreba po redukciji koordinat. Matrika razdalj, ki jo dobimo s prvotnimi vektorji, definiranimi z izrazom (1), je namreč singularna in zato neuporabna za klasifikacijo. Le-to smo izvedli po treh strategijah in sicer po maksimalni, minimalni in UPGMA strategiji. Rezultate klasifikacije smo predstavili z dendrogramom /3/, ki je v bistvu neke vrste "drevo", narisano v smislu razdalj matrike D. Podobnost lahko merimo z merilom, priloženim k dendrogramom, le da je odnos obraten, ker smo kot "podobnostno" matriko v klasifikaciji uporabili "nepodobnostno" matriko (disimillarity matrix).

PRAKTIČNI PRIMER

Po opisani metodi smo poskusili klasificirati 50 krajev v Sloveniji. Upo-

števali smo standardne izvedene vrednosti parametrov vremena za čas od 1. maja do 30. septembra, dobljene iz podatkov petletnega niza 1971-1975. Koordinate vektorja, definiranega z izrazom (1), predstavljajo v našem primeru podatki, opisani v tabeli 1. Iz pregleda izvedenih vrednosti, ki smo jih uporabili v poskusu, se vidi, da predstavljajo skupine spremenljivk isti parameter vremena. Če pogledamo npr. spremenljivke od zaporedne številke 4 do zaporedne številke 10, opisujejo vse te izvedene vrednosti temperaturo. Torej je dejansko smiselno obravnavane spremenljivke reducirati. Za redukcijo v smislu faktorske analize smo uporabili

Tabela 1 Spisek parametrov vremena, ki smo jih upoštevali pri obravnavi problema.

Table 1 List of weather parameters, taken into consideration when dealing with the problem.

1. nadmorska višina kraja;
2. geografska širina kraja;
3. geografska dolžina kraja;
4. poprečna temperatura, izračunana iz klimatološkega poprečna ((T_7 + T_{14} + 2 T_{21})/4);
5. poprečna maksimalna temperatura;
6. poprečna minimalna temperatura;
7. absolutna maksimalna temperatura;
8. absolutna minimalna temperatura;
9. število dni z maksimalno temperaturo, večjo ali enako 25°C;
10. število dni z maksimalno temperaturo, večjo ali enako 30°C;
11. poprečni pritisk vodne pare;
12. poprečna jakost vetra v Bufforih;
13. odstotek pojavljanja vetra iz smeri N;
14. odstotek pojavljanja vetra iz smeri NE;
15. odstotek pojavljanja vetra iz smeri E;
16. odstotek pojavljanja vetra iz smeri SE;
17. odstotek pojavljanja vetra iz smeri S;
18. odstotek pojavljanja vetra iz smeri SW;
19. odstotek pojavljanja vetra iz smeri W;
20. odstotek pojavljanja vetra iz smeri NW;
21. število dni z jakostjo vetra, večjo ali enako 6 po Bufforu;
22. poprečna dnevna oblačnost v desetinkah;
23. število dni, ko je poprečna dnevna oblačnost manjša od 2/10;
24. število dni, ko je poprečna dnevna oblačnost večja od 8/10;
25. poprečna poletna vsota padavin;
26. število dni s padavinami, večjimi ali enakimi 0,1 mm;
27. število dni s padavinami, večjimi ali enakimi 1,0 mm;
28. število dni s padavinami, večjimi ali enakimi 10,0 mm;
29. število dni s točo;
30. število dni z nevihto;
31. število dni z meglo.

metodo glavnih komponent /2/. Stolpci matrike A, ki predstavlja izvedene vrednosti parametrov vremena, so tedaj vrednosti slučajne spremenljivke, na katerih smo uporabili metodo glavnih komponent. Pogoj za uporabo te metode je, da so spremenljivke normalno porazdeljene. V našem poskusu imamo navidezno opraviti s tremi vrstami porazdelitev, in sicer z normalno, logaritemsko-normalno in enakomerno porazdelitvijo. Na srečo lahko zadnji dve s transformacijami prevedemo na normalne porazdelitve in nato uporabimo metodo glavnih komponent.

Oglejmo si najprej transformacijo, s katero dobimo iz enakomerne porazdelitve normalno. To je t.im. Erf funkcija, oziroma invezni integral normalne porazdelitve:

$$P(x \leq x_i) = y_i = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x_i} e^{-\frac{1}{2}t^2} \, dt \,. \tag{4}$$

Če je slučajna spremenljivka

$$Y = \{y_1, y_2, \ldots, y_n\}$$

porazdeljena enakomerno, potem je slučajna spremenljivka

$$X = \{x_1, x_2, \ldots, x_n\}$$

dobljena s transformacijo (4), porazdeljena normalno s porazdelitvijo $N(0,1)$.

Poglejmo še logaritemsko-normalno porazdelitev. Če je X slučajna spremenljivka in je $Y = \ln X$ normalna porazdelitev, potem rečemo, da je X logaritemsko-normalno porazdeljena slučajna spremenljivka. V našem primeru smo uporabili naslednji transformaciji

$$Y = \ln\left(\frac{x+a}{b}\right) \tag{5}$$

in

$$Y = \ln\left(1 - \frac{x+a}{b}\right), \tag{6}$$

odvisno od tega, ali je originalna porazdelitev levo ali desno asimetrična. V tabeli 2 lahko vidimo, kakšen tip transformacij smo izvedli za posamezne spremenljivke. Zaporedne številke v tabeli 2 se ujemajo s tistimi v tabeli 1. Ilustracije transformacij lahko vidimo na slikah 1, 2 in 3; prikazane so le tipične transformacije.

Na tako transformiranih spremenljivkah smo uporabili metodo glavnih komponent. V ta namen smo izračunali korelacijsko matriko. Lastni vektorji, ki pripadajo lastnim vrednostim korelacijske matrike, so iskani

faktorji, ki pripadajo lastnim vrednostim korelacijske matrike, so iskani faktorji, ki razpenjajo prostor, v katerega smo preslikali vektorje, definirane z izrazom (1). Koordinate smo seveda transformirali v skladu s tabelo 2. Tabela 3 prikazuje uporabljene faktorje.

Tabela 2 Oblike transformacij, ki smo jih naredili na osnovnih spremenljivkah pziroma parametrih vremena z ozirom na tabelo 1.

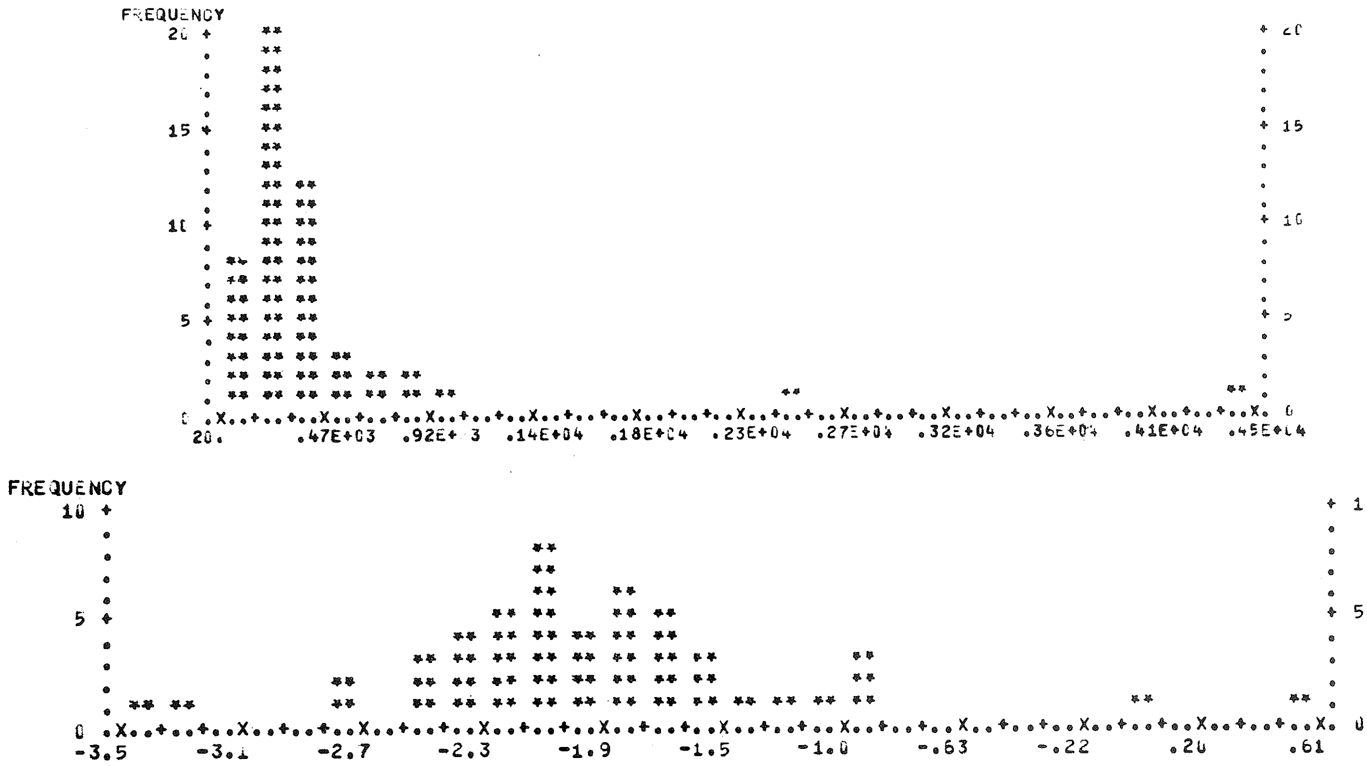
Table 2 Shapes of transformations, carried out at basic variables resp. weather parameters, with regard to Table 1.

1. $y = \ln(x)$	12. $y = \text{erf}(x)$	23. $y = \ln(x)$
2. $y = \text{erf}(x)$	13. $y = \text{erf}(\ln(x))$	24. $y = \ln(1-x)$
3. $y = \text{erf}(\text{erf}(x))$	14. $y = \text{erf}(\ln(x))$	25. $y = \text{erf}(x)$
4. $y = \ln(1-x)$	15. $y = \text{erf}(\ln(x))$	26. $y = x$
5. $y = \ln(1-x)$	16. $y = \text{erf}(\ln(x))$	27. $y = \ln(x)$
6. $y = \ln(1-x)$	17. $y = \text{erf}(\ln(x))$	28. $y = \ln(x)$
7. $y = \ln(1-x)$	18. $y = \ln(x)$	29. $y = \ln(\ln(x))$
8. $y = x$	19. $y = \text{erf}(\ln(x))$	30. $y = \ln(x)$
9. $y = \text{erf}(x)$	20. $y = \text{erf}(\ln(x))$	31. $y = \text{erf}(\ln(x))$
10. $y = \text{erf}(x)$	21. $y = \text{erf}(\ln(x))$	
11. $y = \ln(1-x)$	22. $y = \ln(1-x)$	

Tabela 3 Pregled lastnih vrednosti korelacijske matrike oziroma upoštevanih faktorjev ter pripadajočih varianc.

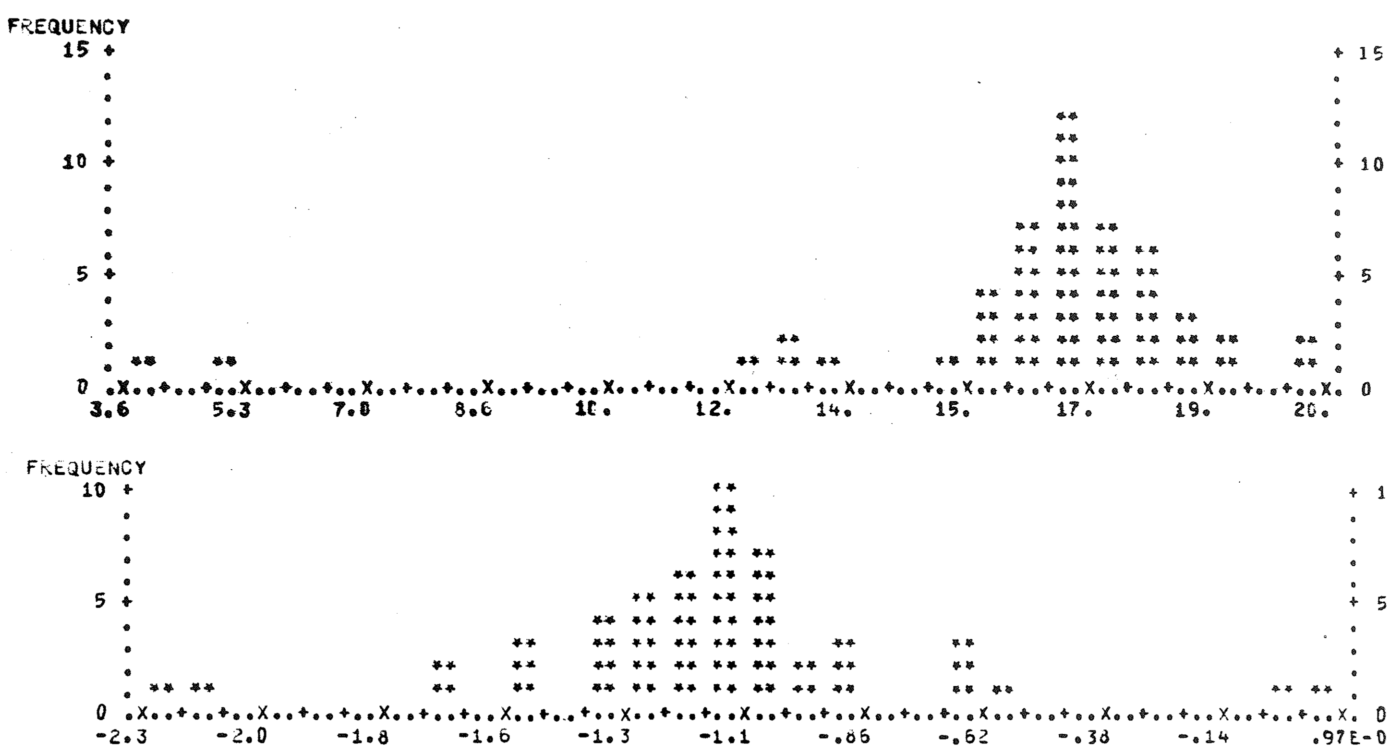
Table 3 Summary of eigen values of correlation matrix resp. of factors taken into consideration and related variances.

zap. št. faktorja	faktorska varianca	procent faktorske variance	procent celotne variance
1.	9.5717	32.3	30.9
2	3.1902	10.8	10.3
3	2.8297	9.6	9.1
4	2.5073	8.5	8.1
5	2.2214	7.5	7.2
6	1.7450	5.9	5.6
7	1.5096	5.1	4.9
8	1.3029	4.4	4.2
9	1.0311	3.5	3.3
10	0.72414	2.4	2.3
11	0.64353	2.2	2.1
12	0.60304	2.0	1.9
13	0.51965	1.8	1.7
14	0.48729	1.6	1.6
15	0.39122	1.3	1.3
16	0.33184	1.1	1.1
skupaj	29.609	100.0	95.5



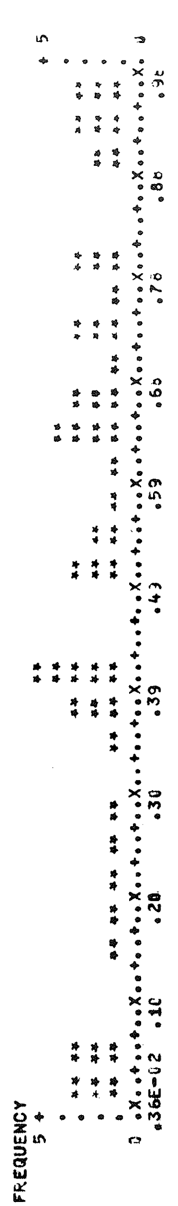
Slika 1 Histogram za spremenljivko 1, to je nadmorsko višino
zgoraj - pred transformacijo
spodaj - po transformaciji tipa y = ln (x)

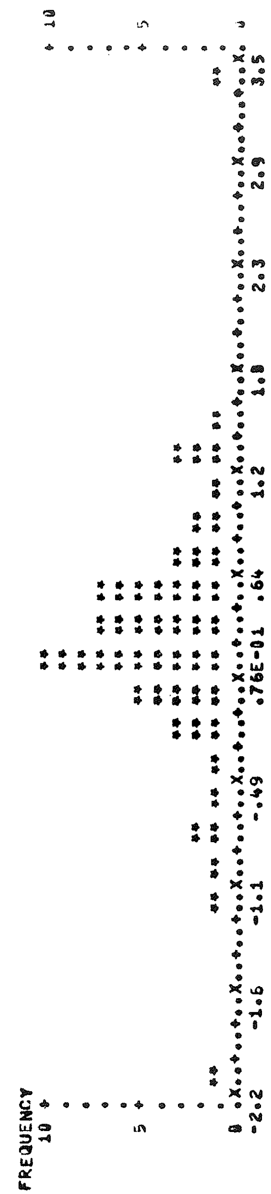
Fig. 1 Histogram for variable 1, i.e. height above-sea level
above - before transformation
below - after transformation of type y = ln (x)



Slika 2 Histogram za spremenljivko 4, to je poprečno temperaturo
zgoraj - pred transformacijo
spodaj - po transformaciji tipa y = ln (1 - x)

Fig. 2 Histogram for variable 4, i.e. mean temperature
above - before transformation
below - after transformation of type y = ln (1 - x)





Slika 3 Histogram za spremenljivko 9, to je število dni z maksimalno temperaturo večjo ali enako 25°C

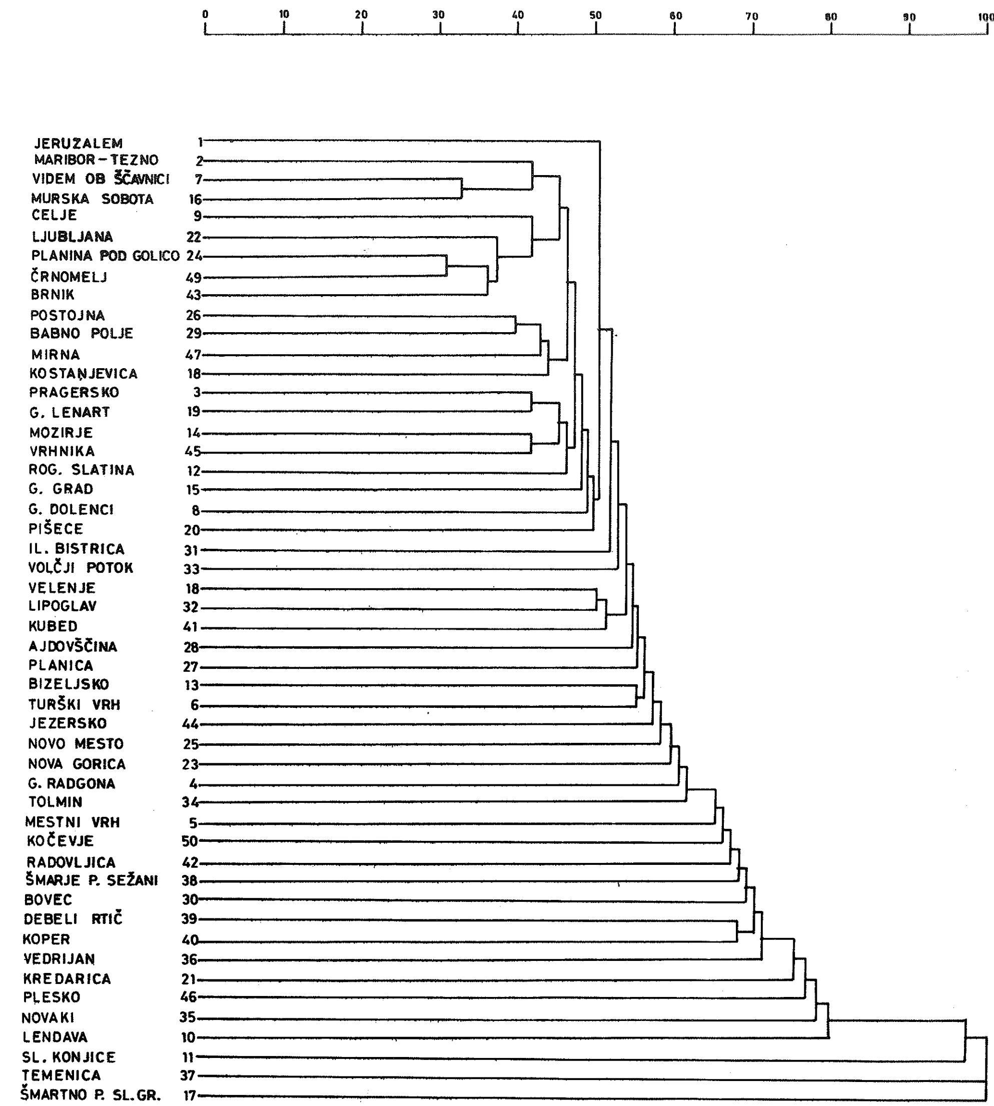
zgoraj - pred transformacijo

spodaj - po transformaciji tipa y = erf (x)

Fig. 3 Histogram for variable 9, i.e. number of days with maximum temperature higher or equal to 25°C

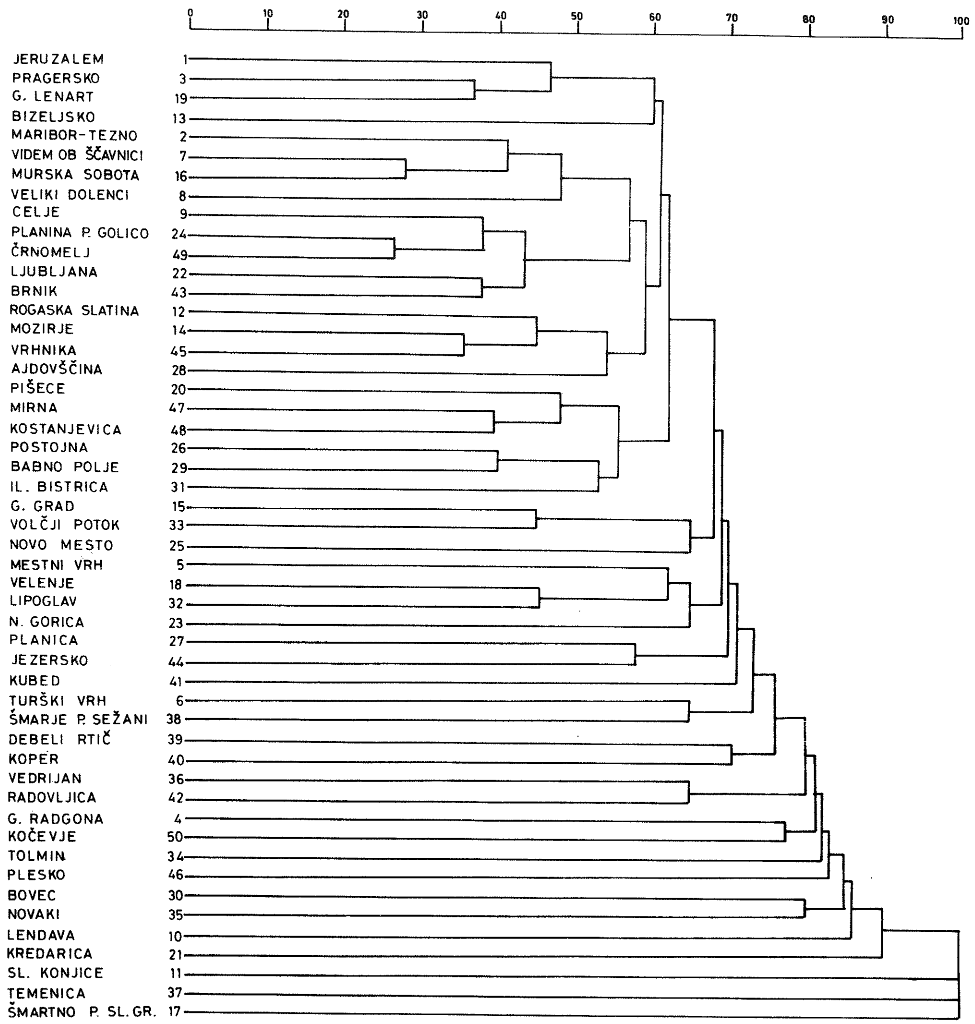
above - before transformation

below - after transformation of type y = erf (x)



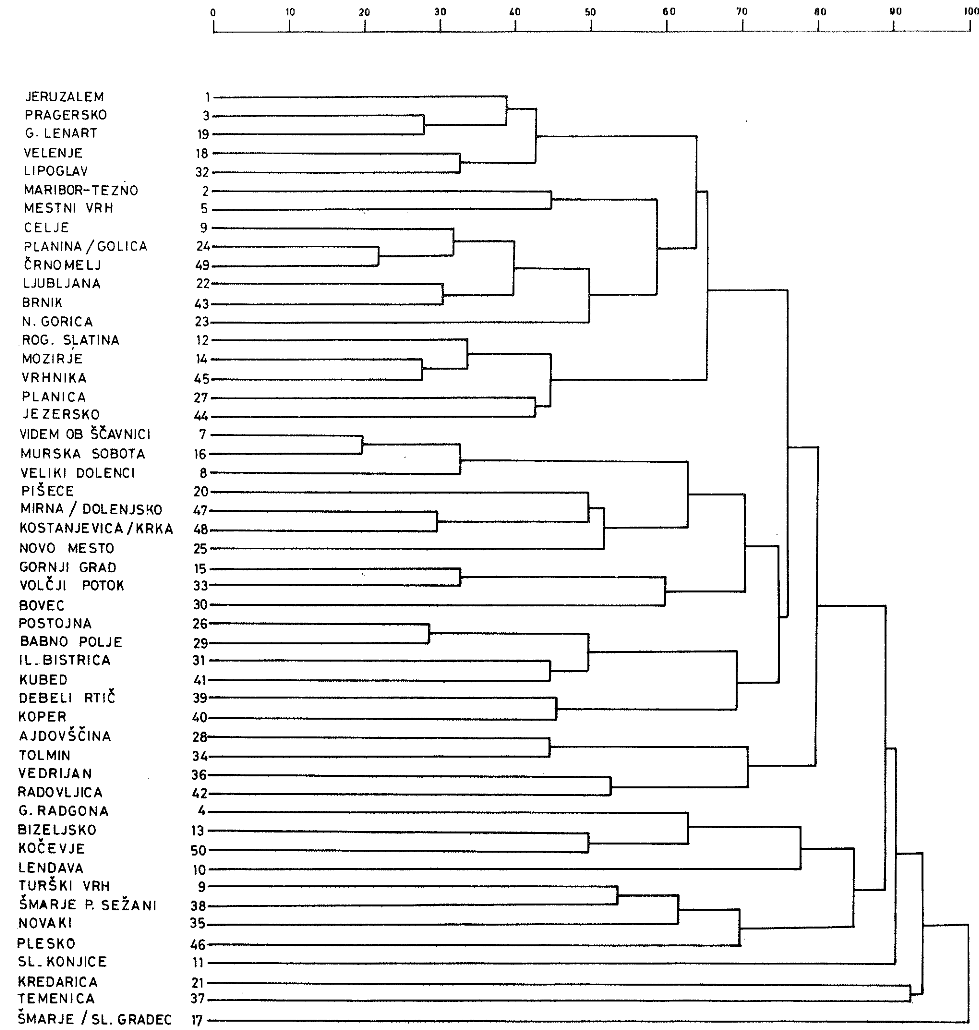
Slika 4 Dendrogram za minimalno strategijo.

Fig. 4 Dendrogram for minimum strategy.



Slika 5 Dendogram za UPGMA strategijo.

Fig. 5 Dendogram for UPGMA strategy.



Slika 6 Dendogram za maksimalno strategijo.

Fig. 6 Dendogram for maximum strategy.

Izračunali smo matriko razdalj, in za posamezne strategije klasifikacije dobili dendograme, predstavljene na slikah 4, 5 in 6. Iz teh slik vidimo, da je merilo podobnosti glede na klimatske spremenljivke oziroma na parametre vremena pri tvorjenju klimatskih področij precej drugačno, kakor smo bili navajeni. Težko bi namreč govorili o klimatskih področjih, saj se po opisani metodi izkažejo za podobne postaje kraji, ki geografsko ležijo relativno precej narazen. Za ilustracijo poglejmo dendogram na sliki 6, kjer smo uporabili maksimalno strategijo. Izberimo si mero podobnosti m = 65 in poglejmo klastre, ki jih dobimo:

1. klaster sestavljajo naslednje postaje:
   Jeruzalem, Pragersko, Gornji Lenart, Velenje, Lipoglav, Maribor-Tezno, Mestni vrh pri Ptuju, Celje, Planina pod Golico, Črnomelj, Ljubljana, Brnik, Nova Gorica;
2. klaster sestavljajo naslednje postaje:
   Rogaška Slatina, Mozirje, Vrhnika, Planica, Jezersko;
3. klaster sestavljajo postaje:
   Videm ob Ščavnici, Murska Sobota, Veliki Dolenci, Pišece, Mirna na Dolenjskem, Kostanjevica ob Krki, Novo mesto;
4. klaster sestavljajo postaje:
   Gornji grad, Volčji potok, Bovec;
5. klaster sestavljajo postaje:
   Postojna, Babno polje, Ilirska Bistrica, Kubed;
6. klaster sestavljata postaji:
   Debeli rtič in Koper;
7. klaster sestavljata postaji:
   Ajdovščina in Tolmin;
8. klaster sestavljata postaji:
   Vedrijan in Radovljica;
9. klaster sestavljajo postaje:
   Gornja Radgona, Bizeljsko in Kočevje;
10. klaster sestavljajo postaje:
    Turški vrh pri Zavrču, Šmarje pri Sežani in Novaki;
11. naslednje postaje so klastri vsaka zase:
    Lendava, Plesko pri Hrastniku, Slovenske Konjice, Kredarica, Temnica in Šmartno pri Slovenj Gradcu.

Rezultate lahko vzamemo le kot ilustracijo opisane metode. Za uporabne rezultate je potrebno še precej dela, tako pri izbiri elementov, ki jih vzamemo kot koordinate vektorjev, kakor tudi pri tolmačenju rezultatov samih. V to se nismo spuščali, ker smo imeli prekratek časovni niz podatkov. Nismo pa upoštevali tudi lege kraja oziroma postaje.

Praktični del naloge je izračunan na računalniku CDC CYBER 172. Statistični del smo naredili s sistemskim paketom STATJOB, programe za transformacijo spremenljivk je napisal avtor prispevka.

ZAHVALA

Ob tej priliki se najlepše zahvaljujem mgr. Veljku Boletu za izčrpne nasvete pri obravnavanju tega problema. Prav tako sem dolžan zahvalo dr. Juretu Zupanu, ki mi je odstopil programe za klasifikacijo.

LITERATURA

/1/ Alexander M. Mood, Franklin A. Graybill: Introduction to the theory of statistics, 1963 by the McGraw-Hill Company, New York.

/2/ William W. Cooley, Paul R. Lohnes: Multivariate data analysis, John Wiley & sons, 1971, New York.

/3/ Peter H.A. Sneath, Robert R. Sokal: Numerical taxonomy, 1973 by W.H. Freeman and Company, San Francisco.

/4/ Arhiv Meteorološkega zavoda SR Slovenije.