Agrovoc descriptors: statistical methods; statistical data; methods; experimentation; case studies; climatology; meteorology

Agris category code: U10, P40

COBISS koda 1.02

Metoda glavnih komponent: osnove in primer

Katarina KOŠMELJ[1]

Prispelo 15. maja 2007; sprejeto 20. junija 2007.
Received May 15, 2007; accepted June 20, 2007.

POVZETEK

V članku predstavljamo osnove metode glavnih komponent in njeno uporabo na enostavnem primeru. Izračuni so bili narejeni s programom SPSS.

Ključne besede: soodvisnost, metoda glavnih komponent

ABSTRACT

PRINCIPAL COMPONENT ANALYSIS: THEORY AND ILLUSTRATION

The paper presents the essential elements of the principal component analysis. We illustrate its use on a simple example. Calculations were done with the SPSS program.

Key words: interdependence, principal component analysis

1. UVOD

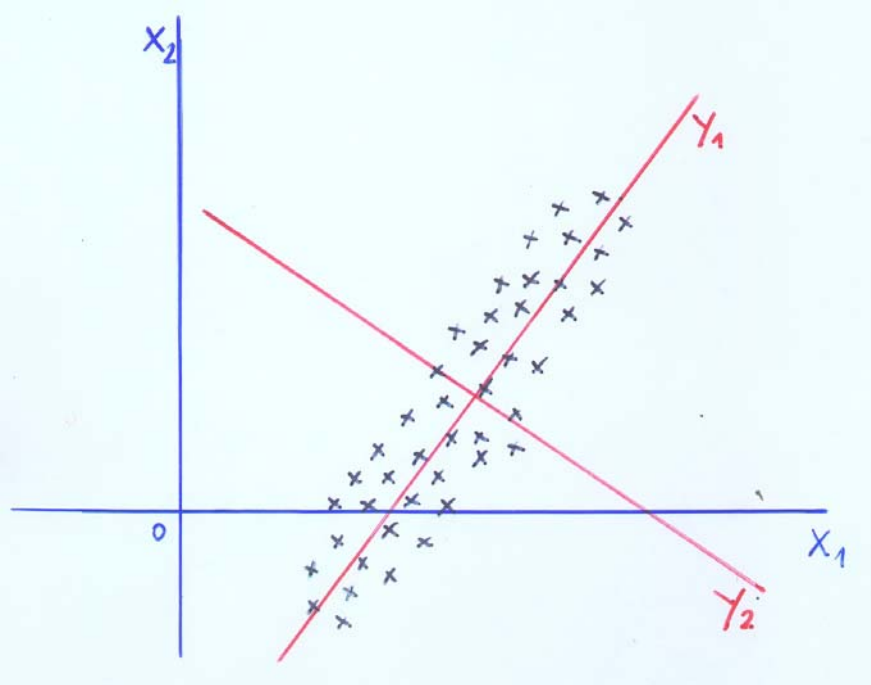
Metoda glavnih komponent (angl. *principal component analysis,* PCA) je statistična tehnika, ki analizira medsebojno soodvisnost spremenljivk z namenom, da se število spremenljivk zmanjša. Pri tem osnovni nabor spremenljivk preslikamo v množico novih spremenljivk, ki jih imenujemo *glavne komponente*.

Glavnih komponent je toliko, kolikor je osnovnih spremenljivk in so med seboj neodvisne (pravokotne). Glavne komponente se izražajo kot linearna kombinacija osnovnih spremenljivk in ohranjajo njihovo skupno variabilnost. Prva glavna komponenta je določena tako, da pojasni kar se da velik del celotne variance osnovnih spremenljivk. Druga glavna komponenta je določena tako, da je neodvisna od prve in pojasni kar se da velik del še nepojasnjene variance. Tretja glavna komponenta je neodvisna od prve in od druge glavne komponente in pojasni kar se da velik del še nepojasnjene variance, itd.

[1] Prof. Ph. D., Jamnikarjeva 101, SI-1111 Ljubljana, P.O. Box 2995,
E-mail: katarina.kosmelj@bf.uni-lj.si

Zaporedne glavne komponente so urejene po padajoči velikosti variance. Če so osnovne spremenljivke dovolj povezane, pojasnijo »pozne« glavne komponente majhen delež celotne variance in jih lahko zanemarimo. Bolj ko so izhodiščne spremenljivke med seboj povezane, bolj uspešna bo redukcija. Kot mero povezanosti uporabimo koeficient kovariance oz. korelacije, pri tem pa mora veljati, da je povezanost med spremenljivkami linearna.

Na Sliki 1 prikazujemo primer v dvorazsežnem prostoru izhodiščnih spremenljivk X_1 in X_2 ter pripadajoči glavni komponenti Y_1 in Y_2. Ker je povezanost med X_1 in X_2 velika, lahko Y_1 uspešno nadomesti obe izhodiščni spremenljivki X_1 in X_2. Dvorazsežni prostor reduciramo v enorazsežnega, pri tem pa je izguba informacije minimalna.



Slika 1. Primer v dvorazsežnem prostoru: X_1 in X_2 sta izhodiščni spremenljivki, podatki so grafično prikazani kot točke. Y_1 in Y_2 sta dobljeni glavni komponenti. Dvorazsežni prostor lahko reduciramo v enorazsežnega, ki ga določa Y_1 (slika povzeta po Ferligoj, A.).

Figure 1. X_1 and X_2 are the original variables, the data are represented by points. Y_1 and Y_2 are the corresponding principal components. Two-dimensional space can be reduced to the one-dimensional space defined by Y_1 (Figure by Ferligoj, A.).

2. MATEMATIČNO OZADJE

2.1 Ideja metode glavnih komponent

Imamo p spremenljivk $X_1, X_2, ..., X_p$, ki jih zapišemo v matriko $\mathbf{X}$,

$\mathbf{X} = \left[X_1, X_2, \ldots, X_p\right]$.

Naš namen je najti novo množico spremenljivk $Y_1, Y_2, ..., Y_p$, ki jih zapišemo kot linearno kombinacijo osnovnih spremenljivk

$$Y_1 = a_{11}X_1 + a_{12}X_2 + ... + a_{1p}X_p = \mathbf{X}\mathbf{a}_1$$
$$Y_2 = a_{21}X_1 + a_{22}X_2 + ... + a_{2p}X_p = \mathbf{X}\mathbf{a}_2$$
…
$$Y_p = a_{p1}X_1 + a_{p2}X_2 + ... + a_{pp}X_p = \mathbf{X}\mathbf{a}_p$$

Pri tem velja: $Var(Y_i) = Var(\mathbf{X}\mathbf{a}_i) = \mathbf{a}_i^{\mathrm{T}}\ \mathbf{\Sigma}\ \mathbf{a}_i$, $Cov(Y_i, Y_k) = \mathbf{a}_i^{\mathrm{T}}\ \mathbf{\Sigma}\ \mathbf{a}_k$,
pri čemer je $\mathbf{\Sigma}$ variančno-kovariančna matrika med osnovnimi spremenljivkami $X_1, X_2, ..., X_p$.

Prvo glavno komponento želimo določiti tako, da je linearna kombinacija osnovnih spremenljivk: $Y_1 = \mathbf{X}\mathbf{a}_1$, ki maksimizira $Var(Y_1)$ pri dodatnem pogoju $\mathbf{a}_1^{\mathrm{T}}\mathbf{a}_1 = 1$ (opomba: ta pogoj določa, da je problem enolično rešljiv). Druga glavna komponenta je linearna kombinacija $Y_2 = \mathbf{X}\mathbf{a}_2$, ki maksimizira $Var(Y_2)$ pri pogoju $\mathbf{a}_2^T\mathbf{a}_2 = 1$ in je pravokotna na prvo glavno komponento: $Cov(Y_1, Y_2) = 0$. Ta postopek nadaljujemo do zadnje, to je *p*-te, komponente.

Koeficiente linearnih kombinacij zapišemo v matriko **A.** Kako dobiti koeficiente matrike **A** ? Navajamo izrek brez dokaza.

Izrek
Glavne komponente $Y_1, Y_2, ..., Y_p$ se izražajo kot linearna kombinacija izhodiščnih spremenljivk: $Y_i = \mathbf{X}\mathbf{a}_i$, pri čemer velja:

- vektorji $\mathbf{a}_i$ so lastni vektorji matrike $\mathbf{\Sigma}$. Rešujemo torej sistem: $|\mathbf{\Sigma} - \lambda\mathbf{I}| = 0$;
- varianca posamezne glavne komponente je enaka pripadajoči lastni vrednosti: $Var(Y_i) = Var(\mathbf{X}\mathbf{a}_i) = \mathbf{a}_i^{\mathrm{T}}\mathbf{\Sigma}\mathbf{a}_i = \mathbf{a}_i^{\mathrm{T}}\lambda\mathbf{I}\mathbf{a}_i = \lambda_i$;
- zaporedne lastne vrednosti so urejene po velikosti: $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$.

Posledica
Vsota varianc osnovnih spremenljivk je enaka vsoti varianc glavnih komponent, kar je $\sum_{i=1}^{p}\lambda_i$. Delež celotne variance, ki ga pojasnjuje *i*-ta glavna komponenta, je $\lambda_i \Big/ \sum_{i=1}^{p}\lambda_i$. Skupni vpliv prvih *m* glavnih komponent, *m* < *p*, je $\sum_{i=1}^{m}\lambda_i \Big/ \sum_{i=1}^{p}\lambda_i$. Ta vrednost prestavlja ključno mero uspešnosti določanja glavnih komponent.

2.2 Izračun glavnih komponent iz korelacijske matrike

Na določanje glavnih komponent najbolj vpliva tista osnovna spremenljivka, ki ima največjo varianco. Če so osnovne spremenljivke merjene v različnih merskih enotah oz. če imajo iste merske enote in različen velikostni red vrednosti, je smiselno vpliv spremenljivk izenačiti. V takih primerih osnovne spremenljivke standardiziramo in s tem dosežemo, da imajo enaka povprečja in enake variance (povprečje 0 in varianca 1). Standardizirane spremenljivke označimo $Z_1, Z_2, ..., Z_p$.

V postopku določanja glavnih komponent se namesto variančno-kovariančne matrike $\mathbf{\Sigma}$ uporabi korelacijska matrika $\boldsymbol{\rho}$. Torej gre v tem primeru za izračun lastnih vektorjev in lastnih vrednosti korelacijske matrike. Odstotek pojasnjene celotne variance s posamezno glavno komponento je $100 \cdot \frac{\lambda_i}{p}$, saj je skupna varianca v tem primeru enaka številu spremenljivk *p*, skupni vpliv prvih *m* komponent pa je $\sum_{i=1}^{m} \lambda_i \Big/ p$.

Zavedati se moramo, da so rezultati, ki jih dobimo iz variančno-kovariančne matrike oz. iz korelacijske matrike, različni. Zato je treba vsakič premisliti, katera od matrik je bolj primerno izhodišče za analizo.

2.3 Reskalirani lastni vektorji

Lastne vektorje interpretiramo v smislu koeficientov pri multipli regresiji: a_{ij} predstavlja vpliv X_j na Y_i ob upoštevanju vseh ostalih osnovnih spremenljivk. Obrazložitev v smislu bivariatne analize imajo vektorji $\mathbf{c}_i$, ki jih dobimo tako, da lastne vektorje množimo z ustrezno konstanto: $\mathbf{c}_i = \mathbf{a}_i \sqrt{\lambda_i}$ (*angl. rescaled eigenvectors*). Za posamezni reskalirani lastni vektor $\mathbf{c}_i$ velja, da je vsota kvadratov njegovih komponent enaka pripadajoči lastni vrednosti:

$$\mathbf{c}_i^T \mathbf{c}_i = \lambda_i \mathbf{a}_i^{\mathrm{T}} \mathbf{a}_i = \lambda_i .$$

Te vektorje zložimo v matriko **C** (angl. *component matrix*), vsak stolpec predstavlja posamezni reskalirani lastni vektor $\mathbf{c}_i$. Vsak element matrike **C** predstavlja korelacijo med standardizirano osnovno spremenljivko in glavno komponento:

$$c_{ij} = \rho(Z_i, Y_j) .$$

Posledica: iz vrednosti korelacijskih koeficientov v tej matriki se pogosto da najti vsebinsko obrazložitev glavnih komponent. Poudarek posvečamo velikim korelacijskim koeficientom.

3. GLAVNE KOMPONENTE NA PODATKIH

V praksi imamo vzorec velikosti *n*, na vsaki enoti zberemo podatke za *p* spremenljivk $X_1, X_2, ..., X_p$. Za *i*-to spremenljivko podatke uredimo v vektor, $X_i = [x_{i1}, x_{i2}, \ldots, x_{in}]^T$. Matrika podatkov je dimenzije *n* x *p*. Iz podatkov izračunamo vzorčno variančno-kovariančno matriko **S** oz. vzorčno korelacijsko matriko **R** ter njene lastne vrednosti in lastne vektorje.

3.1 Ali je metoda glavnih komponent za podatke smiselna ?

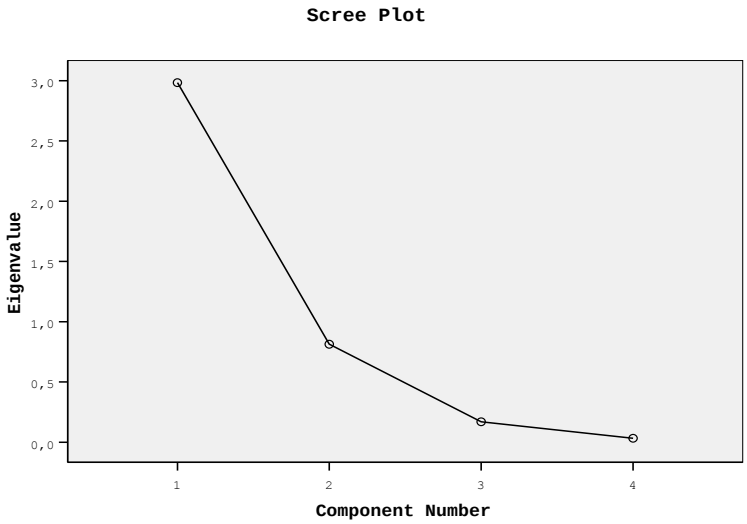
Ali je povezanost spremenljivk dovolj velika, da je smiselno nadomestiti izhodiščne spremenljivke z glavnimi komponentami? Če lahko privzamemo, da so naši podatki vzorec iz populacije ter večrazsežno normalno porazdelitev osnovnih spremenljivk, smemo uporabiti določene statistične teste. Ničelna domneva pravi, da je populacijska variančno-kovariančna matrika $\boldsymbol{\Sigma}$ diagonalna oz. da je populacijska korelacijska matrika $\boldsymbol{\rho}$ enaka identični matriki. Metoda glavnih komponent je smiselna, če H_0 zavrnemo pri dovolj majhnem tveganju. Najbolj znan test v te namene je Bartlettov test, ki temelji na hi-kvadrat statistiki. Dejstvo je, da je omenjena predpostavka v praksi redko izpolnjena, posledično je uporabnost tega testa relativno majhna.

Druga mera, ki služi v isti namen, je Kaiser-Meyer-Olkin-ova mera (KMO). Njena definicija temelji na vrednostih korelacijskih koeficientov in parcialnih korelacijskih koeficientov (Hutcheson, 1999) in vrednoti, ali bi spremenljivke lahko združili v skupine in posamezno skupino spremenljivk nadomestili z glavno komponento. Vrednost KMO mere je med 0 in 1, vrednosti blizu 1 kažejo, da bo redukcija uspešna, vrednosti pod 0,6 pa nakazujejo, da gre za nekoreliranost spremenljivk in neprimernost uporabe te metode.

3.2 Koliko glavnih komponent?

Ko izračunamo lastne vrednosti in lastne vektorje iz variančno-kovariančne oz. iz korelacijske matrike, se odločimo za redukcijo in privzamemo prvih *m, m* < *p* glavnih komponent. Določanje števila *m* je do določene mere subjektivno, opiramo se na različne hevristične postopke. Navajamo nekatere:

- vnaprej določen prag: npr. izbrano število glavnih komponent naj pojasni 80 % skupne variabilnosti osnovnih spremenljivk;
- po grafu »scree plot«, ki prikazuje velikost lastne vrednosti glede na njeno zaporedno mesto. Lokacija »kolena« nakazuje število potrebnih komponent: do kolena vrednosti »strmo« padajo in pripadajoče glavne komponente upoštevamo, od kolena dalje so spremembe manjše in pripadajoče glavne komponente ne upoštevamo več. Na Sliki 2 je »koleno« pri drugi glavni komponenti, kar nakazuje, da sta dve glavni komponenti dovolj;
- Kaiser-jevo pravilo: $m = Max(j, \lambda_j \geq 1)$. Število glavnih komponent je enako številu lastnih vrednosti, katerih vrednost je vsaj ena. Za primer na Sliki 2 bi po tem pravilu zadoščala ena glavna komponenta.



Slika 2. »Scree plot« uporabljamo za določanje potrebnega števila glavnih komponent.

Figure 2. »Scree plot« is an exploratory plot used to determine the relevant number of principal components.

4. ILUSTRACIJA

Za 19 meteoroloških postaj v Sloveniji imamo podatke za 4 klimatske spremenljivke: povprečna letna maksimalna temperatura zraka (X_1=*temp_pmax*), povprečna letna minimalna temperatura zraka (X_2=*temp_pmin*), povprečna letna količina padavin (X_3=*padavine*) in povprečno število dni s snežno odejo na leto (X_4=*dni_sneg*). Podatki veljajo za 30-letno obdobje od leta 1961 do leta 1990 in so prikazani v Tabeli 1 (Košmelj, Breskvar Žaucer, 2006).

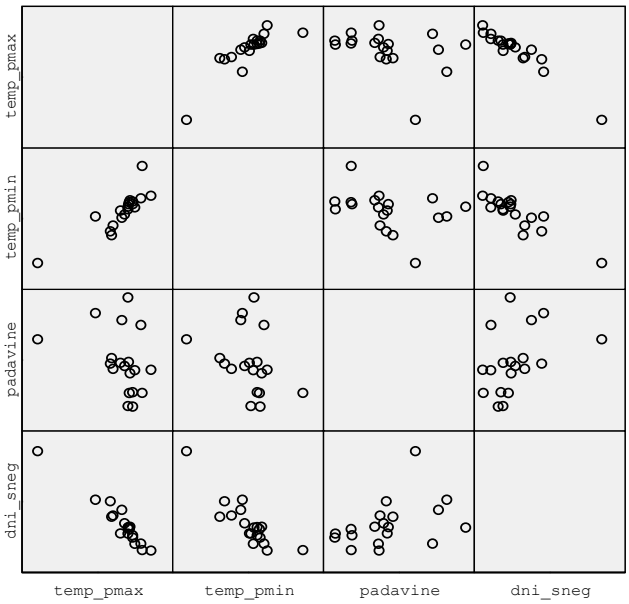
Tabela 1. Klimatski podatki za 19 meteoroloških postaj.
Table 1. Climate data for 19 meteorological stations.

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

Imamo 4 dimenzionalni prostor, v njem pa 19 enot (postaj). Ugotoviti želimo, ali lahko obravnavane 4 spremenljivke nadomestimo z manjšim številom glavnih komponent in grafično prikažemo podatke v manj dimenzionalnem prostoru. Metodo glavnih komponent uporabimo na standardiziranih spremenljivkah zaradi njihove različne narave.

Poglejmo preliminarne izpise, ki smo jih dobili s programom SPSS in so na Sliki 3 in v Tabeli 2. Kaže se močna pozitivna korelacija med *temp_pmax* in *temp_pmin* ter njuna posamična negativna korelacijo z *dni_sneg*. Spremenljivka *padavine* je statistično značilno korelirana le z *dni_sneg* (p=0,044), ostali dve korelaciji sta mejno statistično značilni. Točka, ki na razsevnih grafikonih »štrli ven« (nizke temperature, veliko število dni s snegom), pripada Kredarici.

Bartlettov test kaže, da ničelno domnevo, ki pravi, da je korelacijska matrika enaka identiteti, zavrnemo brez tveganja (p=0,0000), torej je uporaba metode glavnih komponent utemeljena. Tudi vrednost KMO mere potrjuje to trditev.



Slika 3. Matrika razsevnih grafikonov za 4 osnovne spremenljivke.
Figure 3. Scatterplot matrix for the 4 original variables.

Tabela 2. Preliminarna analiza: korelacijska matrika med obravnavanimi spremenljivkami ter testi za ugotavljanje smiselnosti uporabe metode glavnih komponent.
Table 2. Preliminary analysis: correlation matrix on the variables under study and tests for adequacy of PCA.

		temp_pmax	temp_pmin	padavine	dni_sneg
Correlation	temp_pmax	1,000	,853	-,331	-,963
	temp_pmin	,853	1,000	-,328	-,873
	padavine	-,331	-,328	1,000	,403
	dni_sneg	-,963	-,873	,403	1,000
Sig. (1-tailed)	temp_pmax		,000	,083	,000
	temp_pmin	,000		,085	,000
	padavine	,083	,085		,044
	dni_sneg	,000	,000	,044	

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,730
Bartlett's Test of Sphericity	Approx. Chi-Square	68,007
	df	6
	Sig.	,000

Rezultati metode glavnih komponent so v Tabeli 3. Prva lastna vrednost je 2,98, kar predstavlja 75 % celotne variabilnosti, druga lastna vrednost 0,81 pojasnjuje dodatnih 20 % celotne variabilnosti, tretja lastna vrednost 0,17 pojasnjuje še nadaljnje 4 %, četrta lastna vrednost je zanemarljiva v vseh pogledih.

Tabela 3. Lastne vrednosti, odstotek pojasnjene variance ter % celotne pojasnjene variance za 4 glavne komponente.
Table 3. Eigenvalues, % of variance explained and % of total variance explained for 4 principal components.

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,984	74,589	74,589	2,984	74,589	74,589
2	,813	20,333	94,922	,813	20,333	94,922
3	,170	4,252	99,174	,170	4,252	99,174
4	,033	,826	100,000	,033	,826	100,000

SPSS program ne posreduje informacije o lastnih vektorjih. V Tabeli 4 prikazujemo matriko **C**, vsote kvadratov ilustrirajo teorijo. V celicah matrike so korelacijski koeficienti, krepki tisk označuje pomembne korelacijske koeficiente.

Tabela 4. Matrika **C**.
Table 4. Component matrix **C**.

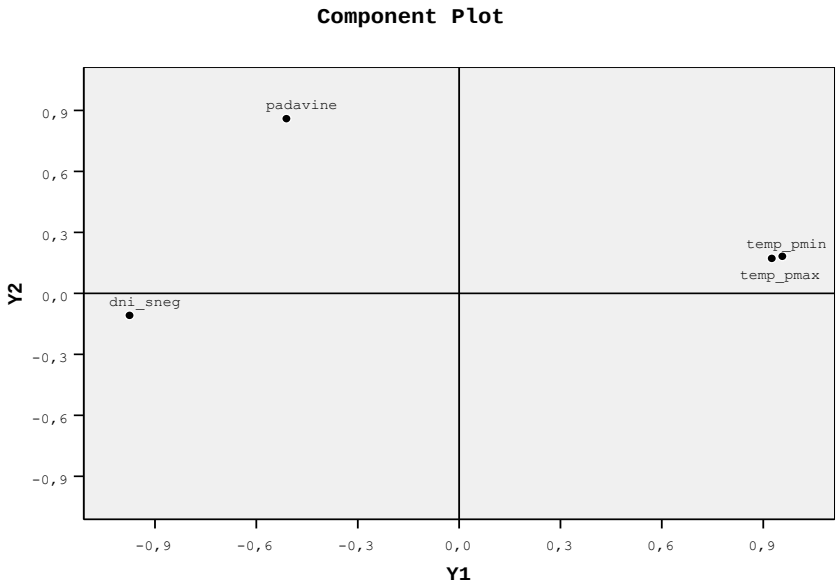
Matrika C	Y_1	Y_2	Y_3	Y_4
temp_pmax	**0,957**	0,183	-0,193	0,120
temp_pmin	**0,925**	0,172	0,340	0,013
padavine	-0,511	**0,859**	-0,001	-0,011
dni_sneg	**-0,975**	-0,108	0,137	0,136
vsota kvadratov	$2{,}984=\lambda_1$	$0{,}813=\lambda_2$	$0{,}171=\lambda_3$	$0{,}033=\lambda_4$

Sedaj se odločimo za potrebno število glavnih komponent. Slika 2 nakazuje 2 glavni komponenti, z njima bomo 4 dimenzionalni prostor zmanjšali v dvodimenzionalnega in ohranili 95 % izhodiščne informacije. Program poženemo znova in zahtevamo 2 glavni komponenti. Dodatno zahtevamo, da se glavni komponenti shranita v datoteko podatkov ter sliko, ki kaže položaj originalnih standardiziranih spremenljivk v prostoru prvih dveh glavnih komponent.

Posvetimo se matriki **C** za prvi dve glavni komponenti (Tabela 4). Glede na vrednosti koeficientov korelacije med standardiziranimi osnovnimi spremenljivkami in glavnimi komponentami bomo prvo glavno komponento poimenovali »indikator temperature« (njene vrednosti so pozitivne za višje vrednosti temperature in negativne za višje število dni s snegom), drugo glavno komponento pa »padavine«. Slika 4 vizualizira položaj osnovnih standardiziranih spremenljivk v prostoru prvih dveh glavnih komponent in dopolnjuje zgornje ugotovitve.

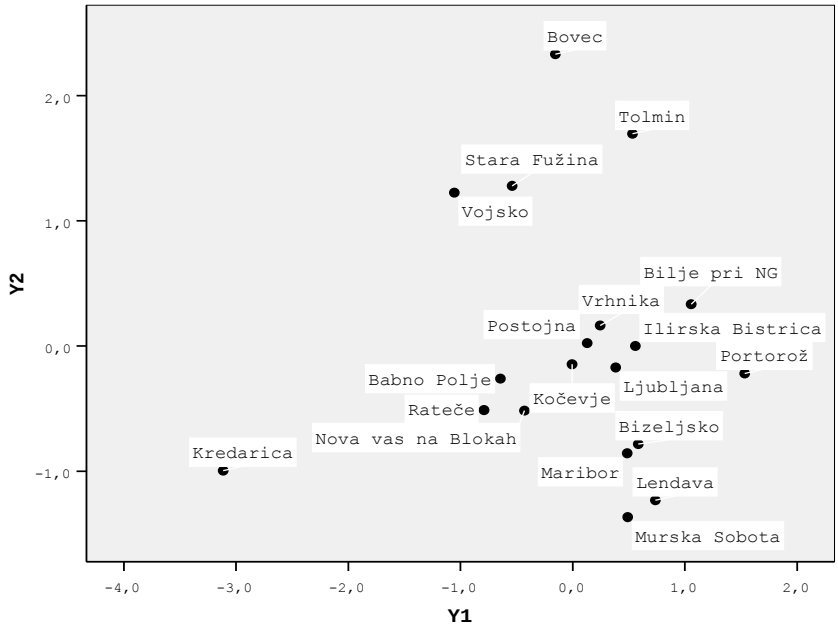
Redukcija 4-dimenzionalnega prostora v dvodimenzionalni prostor omogoča tudi grafično predstavitev postaj in interpretacijo njihove lege glede na pomen glavnih komponent. Narišimo razsevni grafikon postaj na shranjenih glavnih komponentah ter sliko dopolnimo z imeni postaj (Slika 5). Iz slike ugotovimo, da ima npr. Murska Sobota visok indikator temperature in malo padavin, Bovec srednjo vrednost

indikatorja temperature in veliko padavin, Kredarica nizke temperature in relativno malo padavin, itd.



Slika 4. Položaj osnovnih standardiziranih spremenljivk v prostoru prvih dveh glavnih komponent (SPSS: *Rotation/Display/Loadings Plot*).

Figure 4. Original standardized variables in two-dimensional space of principal components (SPSS: *Rotation/Display/Loadings Plot*).



Slika 5. Predstavitev meteoroloških postaj v dvorazsežnem prostoru glavnih komponent.

Figure 5. Representation of stations in two-dimensional space of principal components.

Analizo bomo naredili še enkrat tako, da bomo izločili postajo Kredarica, ki je potencialni osamelec. Analizirali bomo njen vpliv na rezultate. Preliminarna analiza je v Tabeli 5 in kaže, da se korelacijski koeficienti le malo spremenijo (primerjaj s Tabelo 2). Posledično se tudi rezultati metode glavnih komponent le malenkostno spremenijo (primerjaj Tabelo 4 in Tabelo 6). Slika 6 pravzaprav kaže podobno situacijo kot Slika 5, vendar bolj podrobno. Iz tega lahko sklepamo, da Kredarica glede na obravnavane spremenljivke izstopa, vendar ni osamelec.

Tabela 5. Preliminarna analiza: korelacijska matrika med obravnavanimi spremenljivkami. Izločena Kredarica.
Table 5. Preliminary analysis: correlation matrix on the variables under study. Kredarica eliminated from the analysis.

Correlation Matrix

		temp_pmax	temp_pmin	padavine	dni_sneg
Correlation	temp_pmax	1,000	,754	-,346	-,916
	temp_pmin	,754	1,000	-,284	-,771
	padavine	-,346	-,284	1,000	,423
	dni_sneg	-,916	-,771	,423	1,000
Sig. (1-tailed)	temp_pmax		,000	,080	,000
	temp_pmin	,000		,127	,000
	padavine	,080	,127		,040
	dni_sneg	,000	,000	,040	

Tabela 6. Lastne vrednosti, odstotek pojasnjene variance ter % celotne pojasnjene variance 4 in za 2 glavne komponente ter matrika **C**. Izločena Kredarica.
Table 6. Eigenvalues, % of variance explained and % of total variance explained for 4 and 2 principal components and component matrix **C**. Kredarica eliminated from the analysis.

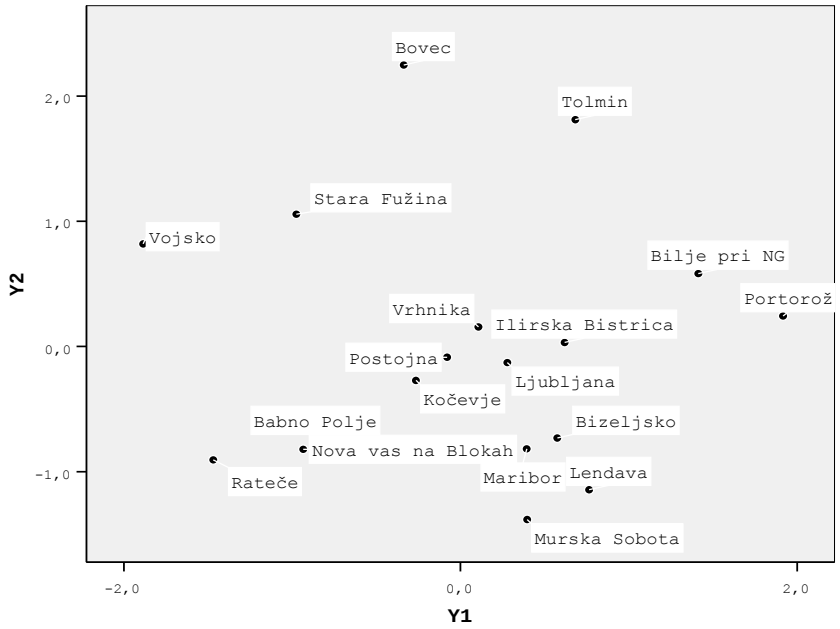
Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,833	70,837	70,837	2,833	70,837	70,837
2	,807	20,180	91,017	,807	20,180	91,017
3	,280	7,006	98,024			
4	,079	1,976	100,000			

Extraction Method: Principal Component Analysis.

Component Matrix

	Component	
	1	2
temp_pmax	,937	,169
temp_pmin	,870	,241
padavine	-,532	,845
dni_sneg	-,957	-,085



Slika 6. Predstavitev meteoroloških postaj v dvorazsežnem prostoru glavnih komponent. Kredarica izločena.

Figure 6. Representation of stations in the two-dimensional space of principal components. Kredarica eliminated from the analysis.

4 SKLEPI

Metoda glavnih komponent se uporablja za analizo soodvisnosti osnovnih spremenljivk s ciljem, da se zmanjša prostorska zahtevnost problema. V statističnih analizah služi predvsem za pregledovanje in raziskovanje večdimenzionalnih podatkov. Če se metoda izkaže za uspešno, dobimo globlji vpogled v izhodiščne večrazsežne podatke. Pogosto se zgodi, da je v prvih dveh/treh glavnih komponentah shranjena večina informacije (variance). V takih primerih lahko grafični prikazi v prostoru glavnih komponent razkrijejo »novosti« ter generirajo nove hipoteze o podatkih, ki jih analiziramo.

Če se redukcija izkaže kot uspešna, lahko delamo nadaljnje analize na manjšem številu glavnih komponent. To se še posebej izkaže za koristno pri določenih metodah (npr. multipla regresija, diskriminantna analiza), če:

- imamo veliko število spremenljivk za relativno majhno število enot. Če lahko osnovne spremenljivke uspešno nadomestimo z manj glavnimi komponentami, ta problem odpade;
- so osnovne spremenljivke visoko korelirane (problem multikolinearnosti). Tedaj pride do numeričnih problemov, ki se jim na ta način izognemo.

Metoda glavnih komponent nima predpostavk o verjetnostni porazdelitvi podatkov. Vendar pa izračun kovariance oz. korelacija zahteva, da so izhodiščne spremenljivke številske z vsaj ordinalno mersko lestvico. Zavedati se moramo tudi, da imajo lahko osamelci velik vpliv na izračun korelacijskih koeficientov in posledično na rezultate, zato je priporočljivo v preliminarni analizi narisati matriko razsevnih grafikonov in

narediti analizo z osamelci in brez njih. Če lahko privzamemo večrazsežno normalno porazdelitev, omogoča statistično sklepanje korak naprej.

Zahvala

Zahvaljujem se dr. Damijani Kastelec za koristne pripombe in komentarje.

5 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, Anuška. Multivariatna analiza. Metoda glavnih komponent. http://vlado.fmf.uni-lj.si/vlado/podstat/Mva.htm (14.05.2007)

Johnson R. A., Wichern D.W.2002. Applied Multivariate Statistical Analysis. Prentice Hall, 767 str.

Košmelj K., Breskvar Žaucer, L. (2006). Metode za razvrščanje enot v skupine; osnove in primer = Methods for cluster analysis; introduction and a case study. *Acta agric.* 87, str. 299-310.

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

Rencher A. C., 1995. Methods of Multivariate Analysis. John Wiley & Sons, 563 str.

DODATEK /APPENDIX

Analizo glavnih komponent smo naredili s programom SPSS, ki ima metodo glavnih komponent vgrajeno v sklop programa **Analyze/Data Reduction/Factor.** Pri analizi rezultatov, ki jih SPSS v tem primeru posreduje, moramo biti pazljivi, kajti program je v osnovi namenjen faktorski analizi in ne metodi glavnih komponent. Posledično vsi rezultati, ki jih dobimo pri metodi glavnih komponent, niso relevantni (npr. komunalitete).

```
/* prva analiza: vse 4 glavne komponente
FACTOR
  /VARIABLES temp_pmax temp_pmin padavine dni_sneg  /MISSING LISTWISE
 /ANALYSIS temp_pmax temp_pmin padavine dni_sneg
  /PRINT UNIVARIATE INITIAL CORRELATION SIG KMO EXTRACTION
  /PLOT EIGEN
  /CRITERIA FACTORS(4) ITERATE(25)
  /EXTRACTION PC
  /ROTATION NOROTATE
  /METHOD=CORRELATION .

/*druga analiza: zahtevamo dve glavni komponenti
FACTOR
  /VARIABLES temp_pmax temp_pmin padavine dni_sneg  /MISSING LISTWISE
 /ANALYSIS temp_pmax temp_pmin padavine dni_sneg
  /PRINT UNIVARIATE INITIAL CORRELATION SIG KMO EXTRACTION
  /PLOT EIGEN ROTATION
  /CRITERIA FACTORS(2) ITERATE(25)
  /EXTRACTION PC
  /ROTATION NOROTATE
  /SAVE REG(ALL)
  /METHOD=CORRELATION .
```