

Vícerozměrná analýza

Úloha:

Provedení vícerozměrné statistické analýzy (metoda hlavních komponent, faktorová analýza, shluková analýza,...) na datech vodárenské laboratoře.

Zadání:

Na 32 vzorcích pitné vody, u kterých bylo provedeno 17 stanovení parametrů kvality, proveďte vícerozměrnou statistickou analýzu.

28 vzorků bylo odebráno na úpravách vody VKM a v přilehlé vodovodní síti (ze 3 zdrojů) a 4 vzorky nespádají pod správu VKM (zakázky od cizích organizací). Shlukovou analýzou ověřte toto rozdělení.

Data: pro velký počet parametrů rozděleny do dvou tabulek:

Číslo vzorku	Parametr kvality								
	barva mgPt/l	zákal ZF	vodivost μS/cm	pH -	VL mg/l	CHSK mg/l	tvrdost mmol/l	alkalita mmol/l	Fe mg/l
1	0,00	0,00	518	7,50	164	0,4	3,33	5,28	0,09
2	0,60	0,00	518	7,42	416	0,5	3,37	5,24	0,03
3	0,00	0,00	518	7,47	448	0,4	3,38	5,26	0,06
4	0,30	1,70	824	7,40	366	0,2	4,58	5,62	0,26
5	3,45	1,44	863	7,95	654	1,4	4,25	6,32	0,42
6	12,53	1,49	520	7,67	269	4,4	2,81	2,65	0,04
7	7,87	0,00	525	8,54	446	3,6	2,70	2,80	0,10
8	0,45	0,00	793	7,36	641	0,6	4,70	5,70	0,12
9	0,92	0,76	822	7,45	616	0,6	4,40	5,40	0,33
10	3,92	1,78	590	7,73	423	2,9	2,70	2,80	0,55
11	2,46	0,28	653	6,22	599	0,5	2,50	1,60	0,03
12	3,56	4,43	1934	7,16	1767	1,2	9,70	10,75	0,29
13	2,53	0,35	2031	7,51	1694	1,2	9,50	10,75	0,15
14	0,00	0,00	539	7,30	384	0,4	3,10	5,50	0,04
15	0,35	3,03	619	7,58	391	0,2	3,20	4,90	0,09
16	2,19	1,40	719	8,08	512	0,4	3,60	4,80	0,20
17	4,13	1,02	788	7,95	567	1,2	3,80	5,30	0,01
18	1,96	0,54	642	7,67	427	0,8	3,47	5,70	0,09
19	11,71	1,86	556	7,68	376	3,1	2,75	2,85	0,03
20	5,06	0,00	556	7,48	390	2,8	2,63	2,50	0,07
21	8,64	0,93	556	7,87	372	2,7	2,48	2,95	0,13
22	1,91	0,00	757	8,10	362	1,9	3,45	5,00	0,07
23	2,28	0,00	872	8,31	608	1,2	4,21	6,40	0,02
24	0,00	0,00	142	6,30	69	0,1	0,60	1,15	0,02
25	4,60	0,00	565	7,47	395	2,4	2,62	2,26	0,16
26	2,80	0,90	828	8,10	420	1,3	4,19	5,66	0,19
27	0,50	0,10	584	7,70	398	0,3	3,28	5,30	0,03
28	2,00	0,00	589	7,62	415	0,5	3,30	5,50	0,01
29	3,40	1,50	553	7,42	397	2,5	2,73	2,22	0,63

30	0,00	0,00	726	7,21	578	0,2	4,27	5,02	0,08
31	1,36	1,29	581	7,34	381	0,2	3,37	5,17	0,25
32	1,08	0,00	581	7,38	410	0,1	3,33	5,15	0,07

2. část tabulky s daty:

Číslo vzorku	Parametr kvality							
	Mn	amon.iont y	chloridy	sírany	dusičnany	saponáty	hum.látky	absorbance
	mg/l	mg/l	mg/l	mg/l	mg/l	mg/l	mg/l	-
1	0,02	0,06	16,3	109,5	10,3	0,06	0,00	0,01
2	0,05	0,19	17,6	75,9	10,4	0,05	0,09	0,01
3	0,06	0,13	18,4	86,4	10,3	0,00	0,00	0,01
4	0,20	0,17	34,5	168,6	32,7	0,05	0,00	0,02
5	0,03	0,14	16,7	96,1	4,2	0,00	0,75	0,06
6	0,03	0,21	18,1	72,1	9,4	0,09	2,40	0,13
7	0,05	0,50	22,7	72,1	8,7	0,00	1,91	0,10
8	0,02	0,08	36,5	168,1	31,1	0,00	0,39	0,04
9	0,27	0,02	30,1	170,5	31,5	0,02	0,25	0,05
10	0,05	0,13	26,2	132,1	9,4	0,00	1,04	0,14
11	0,03	0,00	51,0	168,1	36,4	0,01	2,30	0,05
12	0,46	0,18	93,6	614,8	35,1	0,02	0,41	0,07
13	0,24	0,13	103,0	629,2	35,3	0,09	0,32	0,08
14	0,01	0,00	13,8	57,6	11,3	0,00	0,10	0,04
15	0,04	0,00	15,6	57,6	10,7	0,00	0,08	0,04
16	0,02	0,03	23,4	160,9	3,7	0,00	0,20	0,06
17	0,00	0,00	16,7	194,5	4,7	0,00	0,40	0,06
18	0,00	0,28	21,3	91,2	12,5	0,05	2,80	0,01
19	0,12	0,50	27,3	134,5	7,93	0,08	3,30	0,11
20	0,04	0,45	24,8	163,3	7,66	0,01	0,90	0,06
21	0,06	0,39	25,2	146,5	7,72	0,05	0,40	0,08
22	0,06	0,35	22,7	173,8	3,04	0,00	1,10	0,03
23	0,05	0,47	18,4	211,3	2,84	0,11	4,60	0,03
24	0,00	0,03	8,5	4,8	5,46	0,00	0,00	0,01
25	0,07	0,39	24,4	160,4	7,60	0,00	0,69	0,09
26	0,09	0,46	17,6	215,2	4,20	0,00	0,50	0,06
27	0,04	0,24	17,7	64,4	8,80	0,00	0,75	0,10
28	0,01	0,20	15,2	59,6	9,50	0,00	1,69	0,02
29	0,04	0,02	25,9	151,3	7,21	0,02	0,58	0,09
30	0,06	0,10	28,0	140,7	21,55	0,00	0,00	0,01
31	0,05	0,04	14,2	68,2	8,68	0,01	0,40	0,01
32	0,06	0,05	13,8	91,2	9,59	0,05	0,00	0,00

Program: STATGRAPHICS - Multivariate Methods

Output z PC: VODA.asf, Statgraf.doc

Řešení:

Korelační analýza

Při zpracování veškerých dat (17 proměnných - parametrů kvality - pro 32 vzorků) se pomocí programu Statgraphics spočítá korelační matice (17x17), která obsahuje korelační koeficienty pro uvedený počet vzorků.

Korelační koeficienty jsou mírou lineární závislosti mezi dvěma náhodnými veličinami.

Pokud se korelační koeficient blíží 1 $\Rightarrow$ mezi danými proměnnými je silný korelační vztah.

Záporné znaménko u korelačního koeficientu znamená sestupnou tendenci korelačního vztahu.

Závěr:

Z korelační matice vyplývá, že silný korelační vztah je mezi proměnnými (parametry kvality):

barva - CHSK

barva - absorbance

vodivost - tvrdost

vodivost - alkalita

vodivost - chloridy

vodivost - sírany

vodivost - veškeré látky (VL)

tvrdost - alkalita

tvrdost - chloridy

tvrdost - sírany

chloridy - sírany

chloridy - veškeré látky (VL).

Analýza hlavních komponent - Principal Components Analysis - **PCA**

Poskytuje odpověď na otázku, kterou není schopná zodpovědět korelační analýza, a to: proč spolu některé z proměnných souvisí. Pomáhá nalézt společné latentní proměnné - v PCA nazývané hlavními komponentami, které ve velké míře popisují a vystihují variabilitu v původních datech.

První tabulka, kterou Statgraf poskytl při analýze veškerých dat, je tabulka procentuálního podílu jednotlivých hlavních komponent na vysvětlení variability dat (Cumulative Percentage).

Závěr (vyvozen z tabulky): 2 komponenty vysvětlují 63,11 %

5 komponent vysvětluje 86,79 %

8 komponent 95,63 %

12 komponent 99,35 % variability původních dat.

Vzhledem k tomu, že se pro „dostatečně dobré“ vysvětlení variability původních dat vyžaduje kolem 85 - 90 %, postačuje brát v úvahu prvních 5 hlavních komponent.

Další nabídkou výstupu PCA jsou grafy:

Plot Component Weights (graf komponentních vah), který zobrazuje komponentní váhy vstupujících proměnných *pro první dvě hlavní komponenty*.

Největším přínosem pro danou komponentu mají proměnné, které se na grafu nachází co nejbližší u souřadnice dané komponenty a na číselné ose co nejdále od nuly.

Závěr:

Z grafu je patrné, že největší přínos do první komponenty mají proměnné: vodivost, tvrdost, chloridy, sírany, menší alkalita, dusičnany, mangan.

Druhá komponenta již není tak jednoznačná, ale nejvíce se na ní podílí barva, CHSK, absorbance, amoniakální ionty.

Z uvedeného by se nechalo vyvodit asi to, že první komponenta představuje jakousi „zasolenost“ vzorku, druhá spíš zahrnuje indikátory organického znečištění.

Scatterplot (bodový diagram): zobrazuje komponentní hodnoty prvních dvou hlavních komponent u jednotlivých objektů - v našem případě u 32 vzorků vody.

V případě předpokladu normálního rozdělení vstupních dat, by měly být body rozmístěny v jakémsi pomyslném kruhu („mušinec“).

Rozmístění bodů neodpovídá normálnímu rozdělení, jelikož data nejsou z jednoho zdroje vody.

Biplot (dvojný graf) - kombinuje dva předchozí grafy.

Přímky spojující se v bodě (0,0) představují původní proměnné a jejich příspěvky k prvním dvěma hlavním komponentám. Body představují komponentní skóre jednotlivých objektů. Viz přiložený graf.

Faktorová analýza - FA

Podobně jako PCA patří k metodám redukce počtu proměnných. Faktorová analýza předpokládá, že každou vstupující proměnnou lze vyjádřit jako lineární funkci nevelkého počtu společných latentních proměnných - faktorů a jediného chybového faktoru.

První tabulka, kterou Statgraf poskytuje při analýze veškerých dat, tedy 17 proměnných pro 32 vzorků vody, je tabulka komunalit jednotlivých proměnných (čtverce korelačních faktorů), charakteristických čísel a procent variability vysvětlených jednotlivými faktory zvlášť, nebo v posledním sloupci kumulativně.

Závěr vyvozen z tabulky:

Největší komunalitu mají proměnné: vodivost, tvrdost, alkalita, chloridy, sírany, veškeré látky.

Podobně jako u PCA, k popsání alespoň 85 % variability dat je zapotřebí 5 faktorů, které kumulativně popisují 86,8 % variability dat.

Dalším výstupem FA je odhad matice faktorových vah pro již zmíněných 5 faktorů. Její jednotlivé prvky představují kovariance mezi jednotlivými proměnnými (v řádcích) a společnými faktory (v sloupcích).

Závěr vyvozen z tabulky:

1. faktor je silně korelován proměnnými: vodivost, tvrdost, alkalita, chloridy, sírany, dusičnany, veškeré látky, mangan.
2. faktor proměnnými: barva, CHSK, amonné ionty, absorbance.
3. faktor proměnnými: Fe, zákal, saponáty.
4. faktor hlavně proměnnou pH.
5. faktor nejvíce korelován proměnnou saponáty.

Dalším výstupem FA je tabulka odhadů komunalit odpovídající výše spomínané faktorové matici. Oproti prvním odhadům komunality u jednotlivých proměnných se výrazně zvýšila komunalita u pH a saponátů a snížila u barvy a dusičnanů.

Významnou nabídkou FA je rotace faktorů - jde o optimální natočení souřadnic tak, aby se získalo co nejvíce faktorově čistých proměnných.

Při rotaci Varimax je maximalizován rozptyl čtverců faktorových vah tak, aby se faktorové váhy co nejvíce blížily nule a jedničce.

Podobně jako PCA poskytuje FA grafy faktorových zátěží Plot Factor Weights - souřadnice tvoří první 2 společné faktory. K práci jsou přiloženy grafy faktorových zátěží před a po rotaci k vizuálnímu porovnání. Graf faktorových zátěží před rotací je velmi podobný grafu komponentních vah v PCA.

Shluková analýza - Cluster analysis

Shluková analýza patří mezi metody zabývající se zkoumáním podobnosti objektů u nichž je změřeno větší množství proměnných.

Posuzování podobnosti se provádí podle různých kritérií. Program Statgraphics nabízí několik možností: metoda průměrová (Average), centroidní (Centroid), metoda nejbližšího souseda (Nearest) a metoda nejvzdálenějšího souseda (Furthest). Poslední jmenovaná metoda počítá vzdálenost dvou shluků jako maximum z možných mezishlukových vzdáleností objektů. Jelikož tvoří poměrně kompaktní (nepříliš „rozlehlé“) shluky, byla tato metoda použita i v této úloze.

Závěr:

Pokud byly pro posuzování vzaty v úvahu všechny proměnné (17), rozdělení do 6 shluků (zadáno), které by měly představovat zdroje pitné vody, jak je uvedeno v zadání, nedopadlo podle očekávání. Jeden ze shluků byl příliš rozlehlý - obsahoval 20 vzorků - na úkor ostatních. Obdobně to dopadlo v případě, kdy byly vzaty v úvahu všechny parametry kvality korelující první faktor (viz FA).

Správného rozdělení bylo dosaženo při posuzování *podle 3 parametrů kvality silně korelujících 1. faktor: alkalita, sírany a chloridy*

Výsledky:

1. zdroj vody: vzorky č. 6, 7, 10, 19, 20, 21, 25, 29 byly seskupeny ve shluku č.3
 2. zdroj vody: vzorky č. 1, 2, 3, 5, 14, 15, 18, 27, 28, 31, 32 byly seskupeny ve shluku č.2
 3. zdroj vody: vzorky č. 4, 8, 9, 16, 17, 22, 23, 26, 30 byly seskupeny ve shluku č.1
- cizí vzorek č.11 tvoří shluk 4
cizí vzorky č. 12 a 13 tvoří shluk 5
cizí vzorek č. 24 tvoří shluk 6, což odpovídá skutečnosti.

Grafické metody zkoumání podobnosti objektů

Slouží k vizuálnímu srovnání různých vícerozměrných objektů.

Sun Ray Plot (graf slunečních paprsků)

Počet paprsků odpovídá počtu proměnných. Střed každého paprsku představuje průměr odpovídající proměnné a jeho délka 2.n násobek směrodatné odchylky této proměnné, kde n je námi zadané číslo (zpravidla se zadává číslo 3).

Legendu k grafu s popisem jednotlivých paprsků poskytuje Plot Key (klíč). (Přiložen ke grafu).

Star Symbol Plot (hvězdicový graf)

Délka paprsku zde představuje relativní velikost hodnoty proměnné u příslušného objektu. Konce paprsků jsou spojeny čarami. V případě velkého množství objektů je graf nepřehledný. Plot Key (klíč) popisuje řazení jednotlivých paprsků.

Závěr:

Podle hvězdicového grafu bych rozdělila vzorky do těchto shluků:

1. shluk: 1, 2, 3, 14, 15, 27, 28, 31, 32, s otazníkem 5 a 18
2. shluk: 10, 19, 20, 21, 25, 29, s velkým otazníkem 6, 7 (spíš bych je zařadila do zvláštního shluku)
3. shluk: 4, 8, 9, 30, s otazníkem 16, 17, 22, 23, 26
4. shluk: 11
5. shluk: 12, 13

6. shluk: 24

Výsledky jsou téměř stejné jako po provedení shlukové analýzy. Podle hvězdicového grafu, kde bylo použito více srovnávacích parametrů bych ale vytvořila spíš více shluků, což se u shlukové analýzy nepotvrdilo - čím více srovnávacích parametrů, tím méně kompaktní shluky.

Název souboru:	STATGRAF
Adresář:	E:\Pom\vicerozm
Šablona:	D:\Program Files\Microsoft Office\Sablony\Normal.dot
Název:	Vícerozměrná analýza
Předmět:	
Autor:	Ing. Hosnedlová
Klíčová slova:	
Komentáře:	
Datum vytvoření:	11.09.96 14:17
Číslo revize:	20
Poslední uložení:	16.09.96 14:15
Uložil:	Ing. Hosnedlová
Celková doba úprav:	507 min.
Poslední tisk:	15.09.00 10:00
Jako poslední úplný tisk	
Počet stránek:	6
Počet slov:	1 751 (přibližně)
Počet znaků:	9 985 (přibližně)