3.4 Určení vnitřní struktury analýzou vícerozměrných dat

1. Metoda hlavních komponent PCA

Zadání:
Byly provedeny analýzy chladící vody pro 4 odběrové místa. Byly stanoveny parametry - pH, vodivost, celková alkalita, chloridy, vápník, zinek, fosforečnany a železo. Zjistěte jestli půjde rozlišit odběrová místa.

Data : Tabulka č.1

	pH	vodivost µS	cel.alk. mmol/l	Cl mg/l	Ca mg/l	Zn mg/l	PO4 mg/l	Fe mg/l
odběrové místo 1	8,72	1097	6,20	74	168,8	0,465	0,92	0,102
odběrové místo 1	8,77	917	5,30	57	137,9	0,514	1,25	0,162
odběrové místo 1	8,68	974	5,65	67	150,7	0,617	1,26	0,192
odběrové místo 1	8,66	1007	5,40	70	143,4	0,696	1,53	0,215
odběrové místo 1	8,78	1036	5,70	82	155,1	1,008	1,72	0,288
odběrové místo 1	8,80	998	5,60	73	150,2	0,135	1,72	0,187
odběrové místo 1	8,75	995	5,65	75	145,4	0,838	1,54	0,142
odběrové místo 1	8,73	893	5,45	54	138,8	0,729	1,69	0,088
odběrové místo 1	8,76	1042	5,80	82	149,9	1,049	1,77	0,149
odběrové místo 1	8,83	1090	5,95	89	156,0	1,096	1,79	0,160
odběrové místo 1	8,80	978	5,80	87	145,6	0,363	1,04	0,145
odběrové místo 2	8,61	1089	5,15	64	159,6	0,062	1,23	0,020
odběrové místo 2	8,72	1094	5,60	63	170,6	0,083	0,96	0,044
odběrové místo 2	8,74	1135	5,75	68	179,6	0,067	0,91	0,021
odběrové místo 2	8,76	1110	5,85	64	174,5	0,027	0,67	0,029
odběrové místo 2	8,74	1104	5,65	63	170,3	0,028	0,86	0,013
odběrové místo 2	8,64	1039	5,45	49	164,3	0,049	0,99	0,019
odběrové místo 2	8,75	1168	5,45	68	183,1	0,080	0,86	0,101
odběrové místo 3	8,83	1189	5,85	97	168,3	0,060	0,6	0,018
odběrové místo 3	8,81	1177	5,85	96	167,7	0,076	0,68	0,030
odběrové místo 3	8,92	1207	5,75	101	169,0	0,073	0,51	0,015
odběrové místo 3	8,90	1227	5,70	109	172,9	0,060	0,48	0,023
odběrové místo 3	8,99	1213	5,85	97	166,8	0,064	0,48	0,020
odběrové místo 3	8,88	1180	5,80	82	163,0	0,068	0,52	0,011
odběrové místo 3	8,86	1153	5,65	90	161,4	0,083	0,61	0,050
odběrové místo 4	8,79	1122	5,90	87	157,8	0,346	1,19	0,094
odběrové místo 4	8,66	1197	6,10	104	164,2	0,372	1,51	0,165
odběrové místo 4	8,65	1147	5,65	96	149,9	0,656	1,75	0,139
odběrové místo 4	8,76	1119	5,90	94	164,8	0,732	1,48	0,132
odběrové místo 4	8,75	1018	5,65	75	146,5	0,300	1,27	0,065
odběrové místo 4	8,73	1027	5,60	80	143,3	0,245	1,13	0,066
odběrové místo 4	8,70	1001	5,70	75	146,8	0,161	1,33	0,030
odběrové místo 4	8,76	1099	6,10	82	160,0	0,170	1,14	0,034
odběrové místo 4	8,81	1104	6,00	89	156,2	0,317	1,25	0,036
odběrové místo 4	8,74	1059	5,70	82	151,8	0,349	1,41	0,065
odběrové místo 4	8,85	1023	5,85	68	147,3	0,209	1,11	0,030

3.4 Určení vnitřní struktury analýzou vícerozměrných dat
Mária Kalhousová

Program: Statistica

Předzpracování dat

Data se standardizují, což znamená, že se od základních údajů odečte aritmetický průměr a podělí se směrodatnou odchylkou.

Tabulka č.2 – Průměry a směrodatné odchylky pro dané parametry

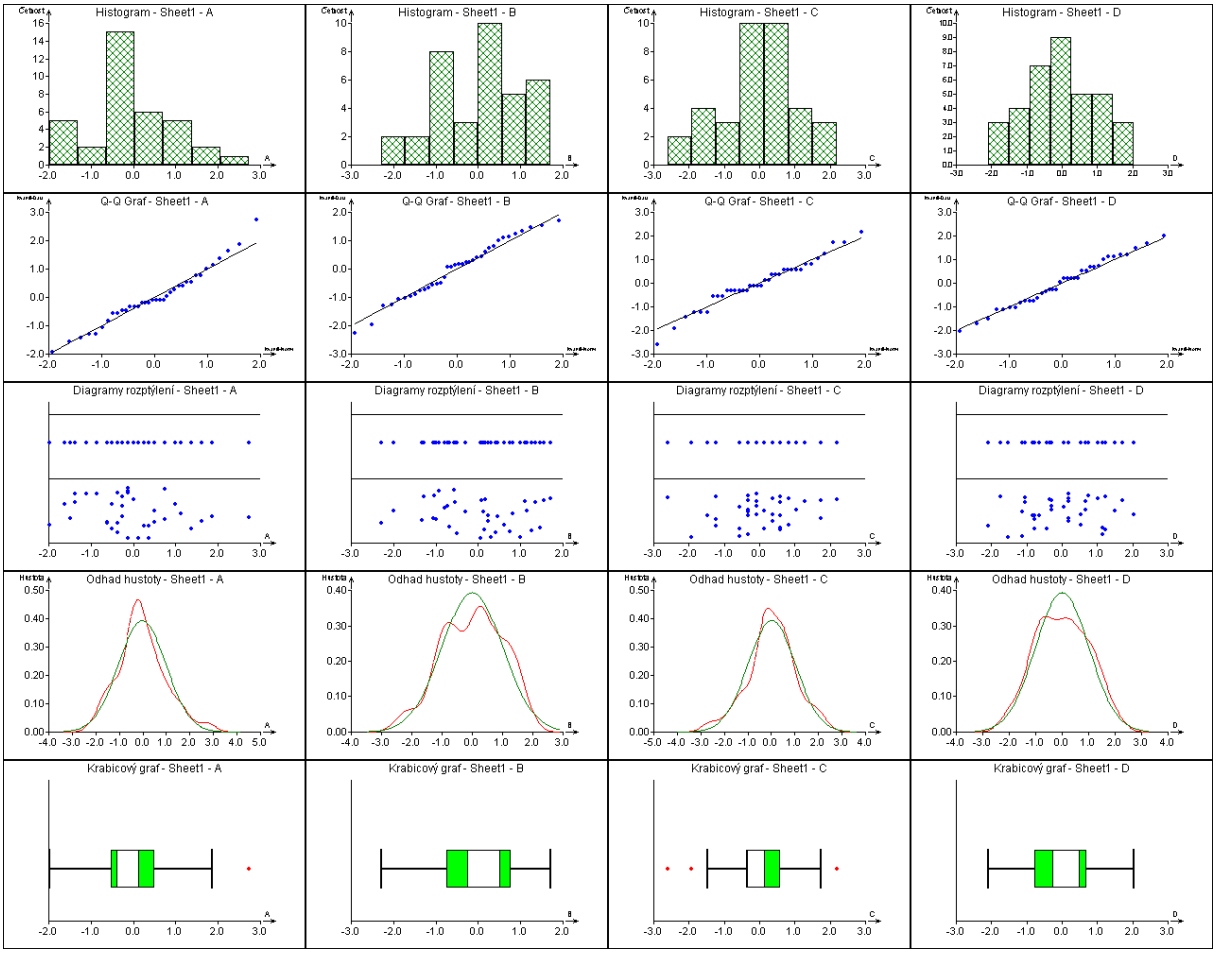
	pH	vodivost µS	cel.alk. mmol/l	Cl mg/l	Ca mg/l	Zn mg/l	PO4 mg/l	Fe mg/l
průměr	8,7675	1084,1	5,7222	79,25	158,38	0,3402	1,1433	0,0861
smodch	0,0805	83,475	0,2184	14,55	11,504	0,3178	0,4044	0,0700

Tabulka č.3 - Standardizované výsledky

	pH	vodivost µS	cel.alk. mmol/l	Cl mg/l	Ca mg/l	Zn mg/l	PO4 mg/l	Fe mg/l
odběrové místo 1	-0,6211	0,1545	2,1978	-0,3608	0,9061	0,3928	-0,5522	0,2270
odběrové místo 1	0,0000	-2,0018	-1,9231	-1,5292	-1,7809	0,5470	0,2638	1,0835
odběrové místo 1	-1,1180	-1,3190	-0,3205	-0,8419	-0,6678	0,8711	0,2885	1,5118
odběrové místo 1	-1,3665	-0,9236	-1,4652	-0,6357	-1,3026	1,1197	0,9561	1,8401
odběrové místo 1	0,1242	-0,5762	-0,0916	0,1890	-0,2852	2,1016	1,4260	2,8822
odběrové místo 1	0,3727	-1,0314	-0,5495	-0,4296	-0,7113	-0,6457	1,4260	1,4404
odběrové místo 1	-0,2484	-1,0674	-0,3205	-0,2921	-1,1287	1,5666	0,9809	0,7980
odběrové místo 1	-0,4969	-2,2893	-1,2363	-1,7354	-1,7026	1,2236	1,3518	0,0271
odběrové místo 1	-0,1242	-0,5043	0,3663	0,1890	-0,7374	2,2306	1,5496	0,8979
odběrové místo 1	0,7453	0,0707	1,0531	0,6701	-0,2070	2,3786	1,5991	1,0550
odběrové místo 1	0,3727	-1,2710	0,3663	0,5326	-1,1113	0,0718	-0,2555	0,8408
odběrové místo 2	-1,9876	0,0587	-2,6099	-1,0481	0,1061	-0,8755	0,2143	-0,9436
odběrové místo 2	-0,6211	0,1186	-0,5495	-1,1168	1,0626	-0,8094	-0,4533	-0,6010
odběrové místo 2	-0,3727	0,6098	0,1374	-0,7732	1,8452	-0,8597	-0,5770	-0,9293
odběrové místo 2	-0,1242	0,3103	0,5952	-1,0481	1,4017	-0,9856	-1,1704	-0,8151
odběrové místo 2	-0,3727	0,2384	-0,3205	-1,1168	1,0365	-0,9825	-0,7006	-1,0435
odběrové místo 2	-1,6149	-0,5403	-1,2363	-2,0790	0,5148	-0,9164	-0,3791	-0,9579
odběrové místo 2	-0,2484	1,0051	-1,2363	-0,7732	2,1496	-0,8188	-0,7006	0,2127
odběrové místo 3	0,7453	1,2567	0,5952	1,2199	0,8626	-0,8818	-1,3435	-0,9722
odběrové místo 3	0,4969	1,1129	0,5952	1,1512	0,8104	-0,8314	-1,1457	-0,8009
odběrové místo 3	1,8634	1,4723	0,1374	1,4948	0,9235	-0,8409	-1,5661	-1,0150
odběrové místo 3	1,6149	1,7119	-0,0916	2,0447	1,2626	-0,8818	-1,6403	-0,9008
odběrové místo 3	2,7329	1,5442	0,5952	1,2199	0,7322	-0,8692	-1,6403	-0,9436
odběrové místo 3	1,3665	1,1488	0,3663	0,1890	0,4017	-0,8566	-1,5413	-1,0721
odběrové místo 3	1,1180	0,8254	-0,3205	0,7388	0,2626	-0,8094	-1,3188	-0,5153
odběrové místo 4	0,2484	0,4540	0,8242	0,5326	-0,0504	0,0183	0,1154	0,1128
odběrové místo 4	-1,3665	1,3525	1,7399	1,7010	0,5061	0,1001	0,9067	1,1263
odběrové místo 4	-1,4907	0,7535	-0,3205	1,1512	-0,7374	0,9939	1,5002	0,7552
odběrové místo 4	-0,1242	0,4181	0,8242	1,0137	0,5583	1,2330	0,8325	0,6552
odběrové místo 4	-0,2484	-0,7919	-0,3205	-0,2921	-1,0330	-0,1265	0,3132	-0,3012
odběrové místo 4	-0,4969	-0,6840	-0,5495	0,0515	-1,3113	-0,2996	-0,0330	-0,2869
odběrové místo 4	-0,8696	-0,9955	-0,0916	-0,2921	-1,0070	-0,5639	0,4616	-0,8009
odběrové místo 4	-0,1242	0,1785	1,7399	0,1890	0,1409	-0,5356	-0,0082	-0,7438
odběrové místo 4	0,4969	0,2384	1,2821	0,6701	-0,1896	-0,0730	0,2638	-0,7152
odběrové místo 4	-0,3727	-0,3007	-0,0916	0,1890	-0,5722	0,0277	0,6594	-0,3012
odběrové místo 4	0,9938	-0,7320	0,5952	-0,7732	-0,9635	-0,4129	-0,0824	-0,8009

Charakter vícerozměrných dat

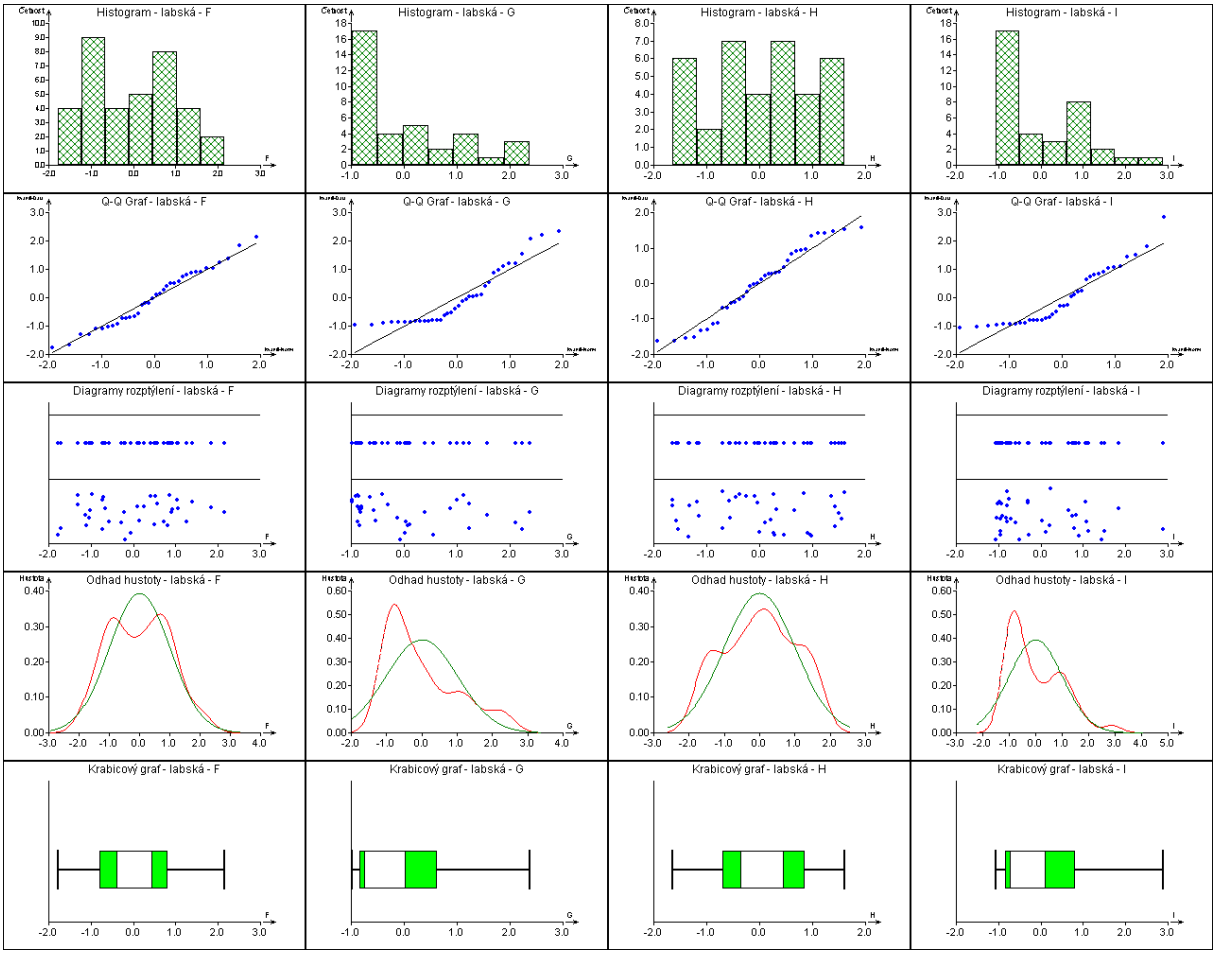
Graf č. 1 – Vybrané diagnostiky průzkumové analýzy dat – Voda – v pořadí (shora) histogram, QQ, diagramy rozptýlení a odhad hustoty pro znaky pH, vodivost, celková alkalita a chloridy (QCExpert)



Klasické parametry :

Název sloupce :	pH	vodivost	Celk.alkalita	Cl
Průměr :	-0,03105556	0,00013611	0,01016389	-5,56E-06
Spodní mez :	-0,3743449	-0,34301406	-0,3329959	-0,34313959
Horní mez :	0,31223379	0,34328628	0,35332368	0,34312848
Rozptyl :	1,02940176	1,02856722	1,02862489	1,02847052
Směr. odchylka :	1,01459438	1,01418303	1,01421146	1,01413536
Šikmost	0,47463149	-0,23533385	-0,24297697	0,01548647
Odchylka od 0 :	Nevýznamná	Nevýznamná	Nevýznamná	Nevýznamná
Špičatost :	3,33646222	2,34996572	3,32077587	2,25781027
Odchylka od 3 :	Nevýznamná	Nevýznamná	Nevýznamná	Nevýznamná
Polosuma	0,37265	-0,2887	-0,20605	-0,01715
Modus :	-0,30041922	0,39463033	-0,28412628	0,34776051
Homogenita :	Přijata	Přijata	Přijata	Přijata
Normalita :	Přijata	Přijata	Přijata	Přijata

Graf č. 2 – Vybrané diagnostiky průzkumové analýzy dat – Voda – v pořadí (shora) histogram, QQ, diagramy rozptýlení a odhad hustoty pro znaky Ca, Zn, PO_4 a Fe (QCExpert)



Tabulka č.4

Klasické parametry :

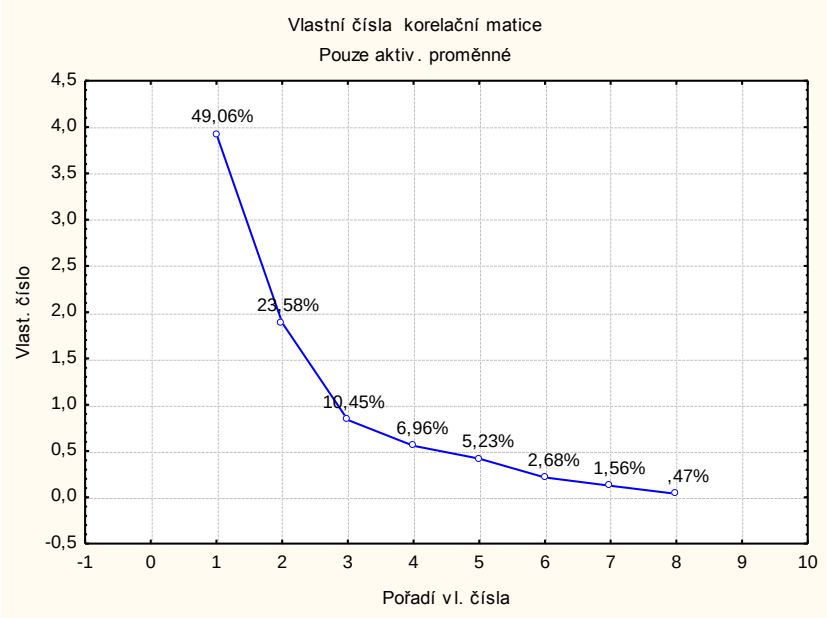
Název sloupce :	Ca	Zn	PO4	Fe
Průměr :	-0,00043611	5,56E-06	1,67E-05	0,00015556
Spodní mez :	-0,34371882	-0,34314915	-0,34313646	-0,34299394
Horní mez :	0,3428466	0,34316026	0,34316979	0,34330505
Rozptyl :	1,02936194	1,02859442	1,02858495	1,02856318
Směr. odchylka :	1,01457476	1,01419644	1,01419177	1,01418104
Šikmost	0,12220892	0,95947098	-0,06349805	0,86996233
Odchylka od 0 :	Nevýznamná	Významná	Nevýznamná	Významná
Špičatost :	2,11781989	2,74430249	1,91475338	2,97332544
Odchylka od 3 :	Nevýznamná	Nevýznamná	Nevýznamná	Nevýznamná
Polosuma	0,18435	0,6965	-0,0206	0,90505
Modus :	0,08136426	-1,030247	0,15497387	-0,87133213
Homogenita :	Přijata	Přijata	Přijata	Přijata
Normalita :	Přijata	Přijata	Přijata	Přijata

3.4 Určení vnitřní struktury analýzou vícerozměrných dat
Mária Kalhousová

Analýza hlavních komponent - PCA

Metoda snižuje počet původních proměnných tím, že vytvoří lineární kombinaci zdrojových proměnných, které vysvětlují největší část jejich variability. První hlavní komponenta je taková kombinace vstupujících proměnných, která má největší rozptyl mezi všemi lineárními kombinacemi. Podobně následuje druhá hlavní komponenta. Pro dostatečné vysvětlení chování zdrojových proměnných požadujeme 85 – 90 % vysvětlené variability. Vstupní data byla při výpočtu standardizována (nemají stejný rozměr).

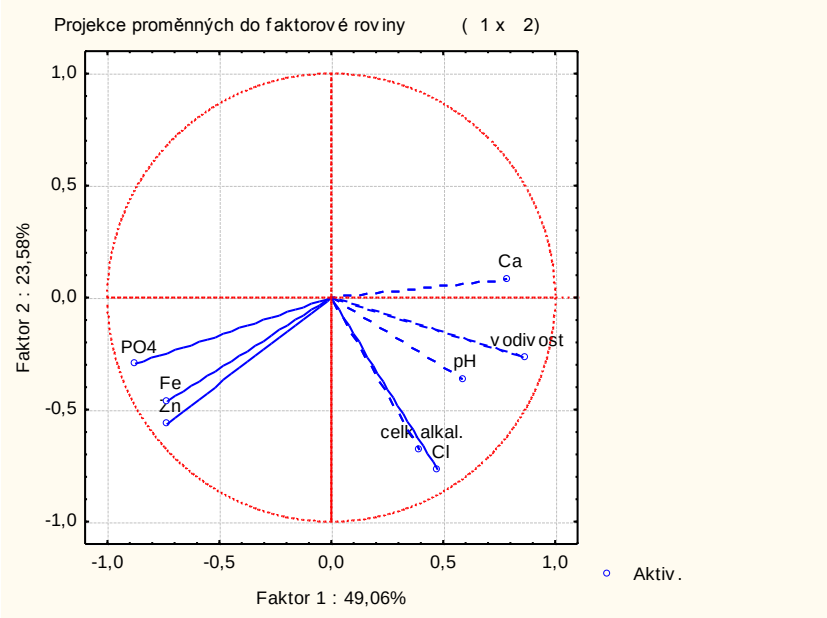
1. Vyšetření indexového grafu úpatí vlastních čísel
z hrany úpatí v tomto diagramu se určí vhodný počet hlavních komponent
Graf.č 3 – Cattelův indexový graf úpatí vlastních čísel



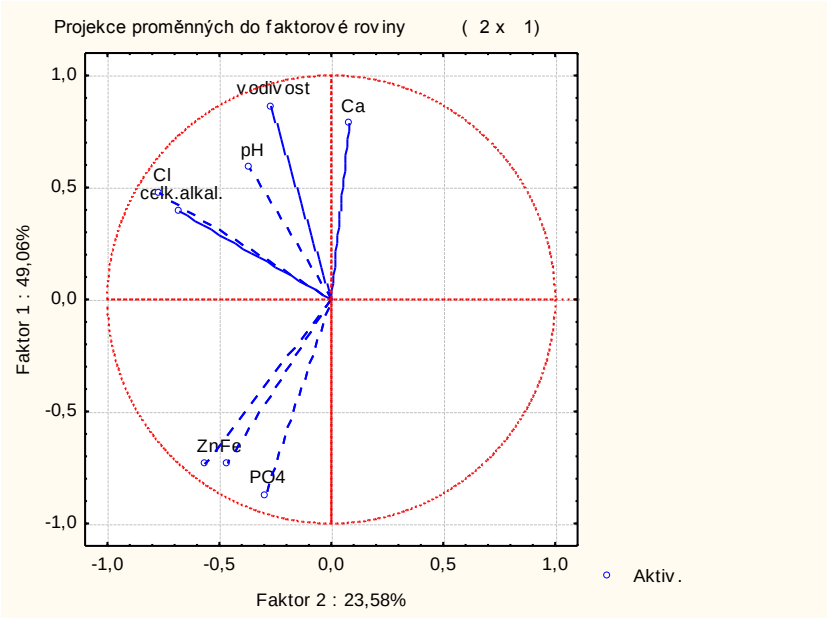
Je patrné že zlom není moc zřetelný. První hlavní komponenta popisuje 49,06% celkového rozptylu, druhá hlavní komponenta popisuje 23,58% celkového rozptylu a třetí hlavní komponenta 10,45%. První a druhá komponenta popisují celkem 72,64. První tři popisují celkem 83,09. Pro dostatečné vysvětlení chování zdrojových proměnných požadujeme 85 – 90 % vysvětlené variability.

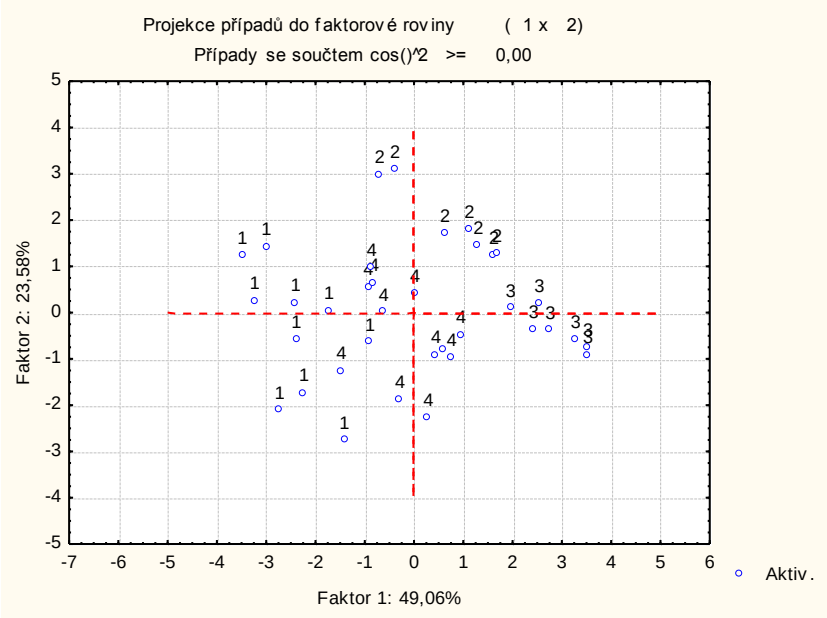
Plot Component Weights – Graf komponentních vah
Zobrazuje komponentní váhy vstupujících proměnných pro 1 – 2 hl. komponenty
Největším přínosem pro danou komponentu mají proměnné, které se na grafu nachází co nejblíže u souřadnice dané komponenty a na číselné ose co nejdále od nuly

Graf č. 4 – Graf komponentních vah – pH(1) a vodivost (2)
Tady vodivost, vápník a pH ,chloridy a celková alkalita. Záporné hodnoty – fosforečnany, železo a zinek.



Graf č.5 – Graf komponentních vah – vodivost a pH





Závěr : PCA se jeví užitečnou pomůckou pro rozlišení odběrových míst. Odběrové místa 2 a 3 jsou dobře rozlišeny. U odběrových míst 1 a 4 bude problém protože se částečně překrývají. V případě předpokladu normálního rozdělení vstupních dat by měly být body rozmístěny v jakémsi pomyslném kruhu – mušinec.Rozmístění neodpovídá normálnímu rozdělení, protože data nejsou ze stejných zdrojů vody.

Tabulka č 5. Korelace faktorů a proměnných (fakt.zátěže) podle korelací

	pH	vodivost	celk.alkal.	Cl	Ca	Zn	PO4	Fe
pH	0,5866	-0,3638	0,6462	-0,1154	0,2673	0,0814	0,1203	-0,0068
vodivost	0,8633	-0,2694	-0,3043	-0,2259	-0,1042	0,0912	0,0150	-0,1381
celk.alkal.	0,3940	-0,6798	-0,1273	0,5904	0,0968	-0,0900	-0,0005	-0,0197
Cl	0,4756	-0,7683	0,0532	-0,2006	-0,3564	-0,0620	-0,0354	0,0919
Ca	0,7851	0,0835	-0,5005	-0,1127	0,3108	0,0660	0,0581	0,0952
Zn	-0,7340	-0,5655	-0,0519	-0,0597	0,1680	0,2682	-0,1872	0,0033
PO4	-0,8738	-0,2981	-0,1832	0,0496	-0,1351	0,1527	0,2646	0,0075
Fe	-0,7315	-0,4622	-0,1412	-0,2919	0,2382	-0,2965	0,0250	-0,0295

Dobrá korelace – pH a vodivost – 0,8633; pH a vápník 0,7851. Korelace pH – zinek, fosforečnany a železo je záporná. Korelace je u celkové alkality a chloridů. Jinak korelace nízká a záporná.

Tabulka č 6 - Korelace

Korelace

	pH	vodivost	celk.alkal.	Cl	Ca	Zn	PO4	Fe
pH	1,0000	0,4162	0,3467	0,5107	0,2146	-0,2073	-0,5202	-0,2758
vodivost	0,4162	1,0000	0,4131	0,6649	0,7944	-0,4484	-0,5986	-0,4455
celk.alkal.	0,3467	0,4131	1,0000	0,5538	0,2720	0,0588	-0,1161	-0,0781
Cl	0,5107	0,6649	0,5538	1,0000	0,1970	0,0250	-0,1762	-0,0119
Ca	0,2146	0,7944	0,2720	0,1970	1,0000	-0,5314	-0,6407	-0,4562
Zn	-0,2073	-0,4484	0,0588	0,0250	-0,5314	1,0000	0,7852	0,7787
PO4	-0,5202	-0,5986	-0,1161	-0,1762	-0,6407	0,7852	1,0000	0,7173
Fe	-0,2758	-0,4455	-0,0781	-0,0119	-0,4562	0,7787	0,7173	1,0000

Korelace pH-chloridy (0,51), pH – vodivost (0,42); vodivost-chloridy (0,66), vodivost – vápnik (0,79); celková alkalita- vodivost (0,41) , celková alkalita – chloridy (0,55); zinek-fosforečnany (0,79) a zinek- železo (0,78).

Tabulka č.7 - Kovariance

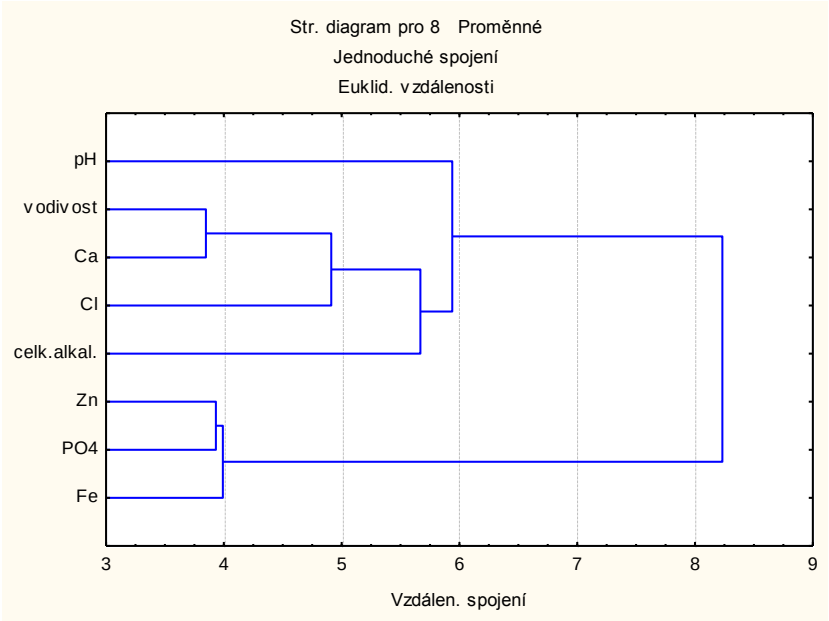
	pH	vodivost	celk.alkal.	Cl	Ca	Zn	PO4	Fe
pH	1,0294	0,4283	0,3567	0,5255	0,2209	-0,2133	-0,5353	-0,2838
vodivost	0,4283	1,0286	0,4249	0,6839	0,8175	-0,4612	-0,6157	-0,4583
celk.alkal.	0,3567	0,4249	1,0286	0,5696	0,2798	0,0605	-0,1194	-0,0803
Cl	0,5255	0,6839	0,5696	1,0285	0,2027	0,0257	-0,1813	-0,0122
Ca	0,2209	0,8175	0,2798	0,2027	1,0294	-0,5468	-0,6592	-0,4694
Zn	-0,2133	-0,4612	0,0605	0,0257	-0,5468	1,0286	0,8076	0,8010
PO4	-0,5353	-0,6157	-0,1194	-0,1813	-0,6592	0,8076	1,0286	0,7378
Fe	-0,2838	-0,4583	-0,0803	-0,0122	-0,4694	0,8010	0,7378	1,0286

Kovariace - pH-chloridy (0,53), pH – vodivost (0,43); vodivost-chloridy (0,68), vodivost – vápnik (0,82); celková alkalita- vodivost (0,43) , celková alkalita – chloridy (0,57); zinek-fosforečnany (0,81) a zinek- železo (0,80).

Shluková analýza – Cluster analysis

Metoda která na základě podobnosti objektů umožňuje rozklad objektů do několika sourodých tříd (shluků). Posuzování podobnosti se provádí podle různých kritérií. Možnosti Statistiky

Graf. č.7 - Horizontální graf hierarchického stromu – Jednoduché spojení Euklidovské vzdál.



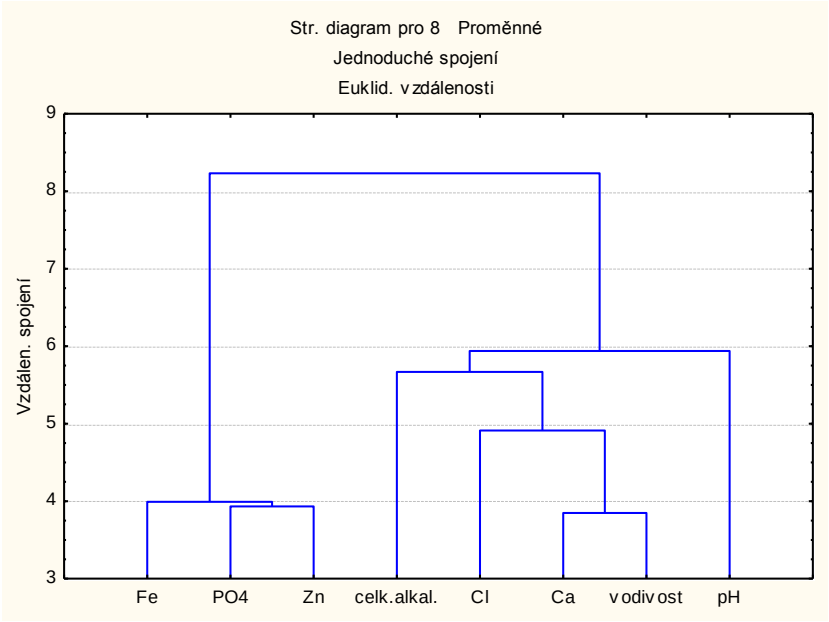
Objekty se seskupili do jediného shluku. Seskupení objektů do shluků znázorníme do dendrogramu . Nejdřív se vytvořilo spojení vodivost a vápník, přidal se chlór a celková alkalita.A potom pH to je jedna část. Zinek a fosforečnany a pak železo vytvořili druhou část. Velice podobné si jsou vodivost - vápník a zinek a fosforečnany.

Tabulka č. 8 – Matice vzdáleností

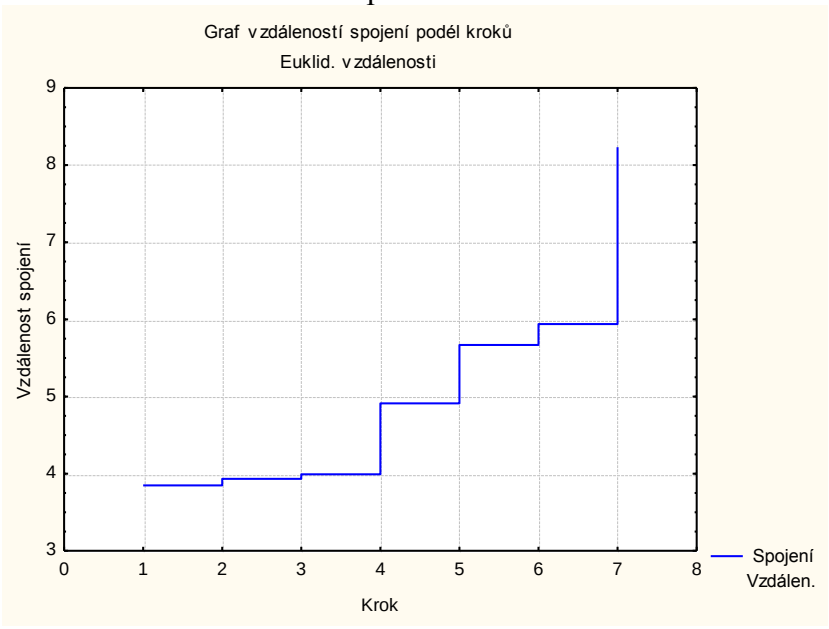
	pH	vodivost	celk.alkal.	Cl	Ca	Zn	PO4	Fe
pH	0,0	6,5	6,86	5,94	7,5	9,3	10,5	9,6
vodivost	6,5	0,0	6,50	4,91	3,8	10,2	10,7	10,2
celk.alkal.	6,9	6,5	0,00	5,67	7,2	8,2	9,0	8,8
Cl	5,9	4,9	5,67	0,00	7,6	8,4	9,2	8,5
Ca	7,5	3,8	7,24	7,60	0,0	10,5	10,9	10,2
Zn	9,3	10,2	8,23	8,38	10,5	0,0	3,9	4,0
PO4	10,5	10,7	8,96	9,20	10,9	3,9	0,0	4,5
Zn	9,6	10,2	8,81	8,54	10,2	4,0	4,5	0,0

Nejkratší vzdálenosti – vodivost – vápník 3,8 a zinek-fosforečnany 3,9.

Graf č.8 – Vertikální třásňový graf

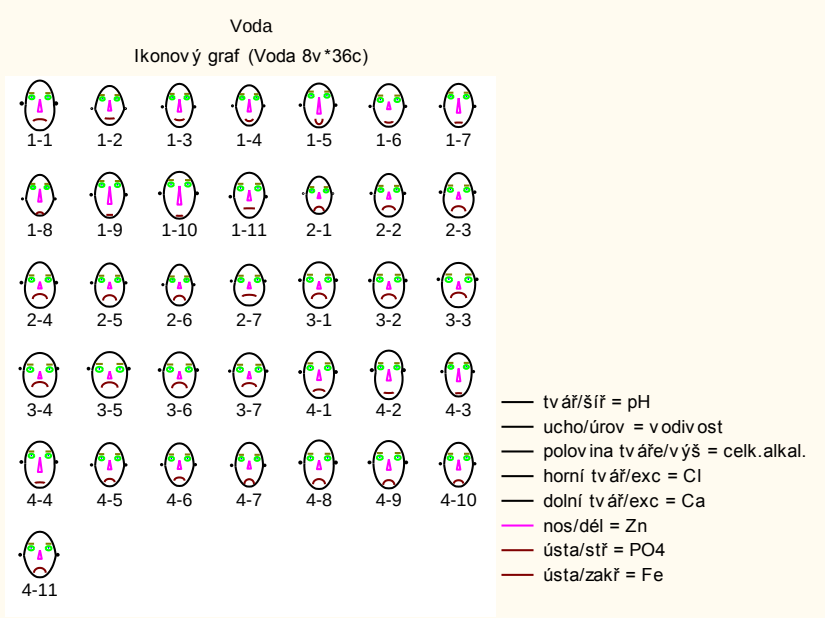


Graf č. 9 – Graf vzdáleností podél kroků

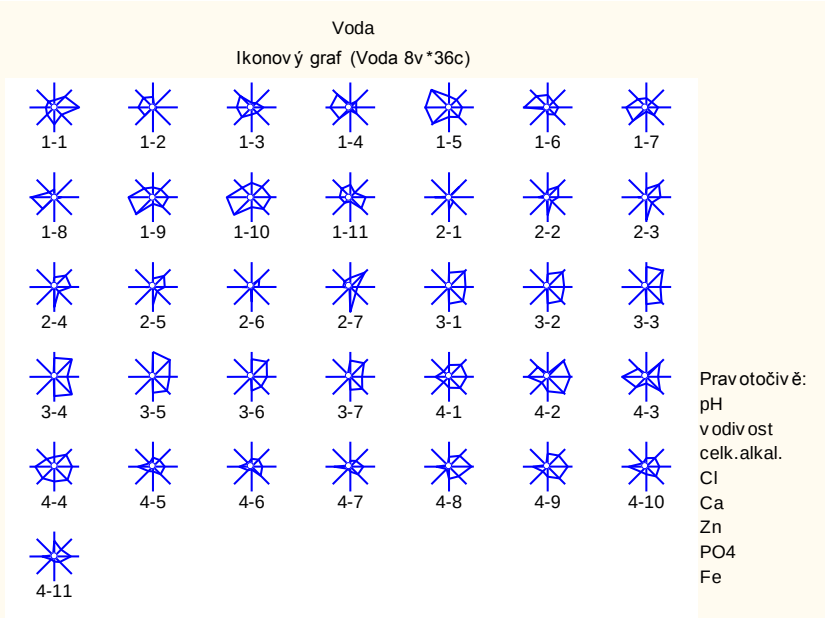


Grafické metody zkoumání podobnosti objektů - slouží k vizuálnímu srovnání různých objektů

Graf č. 10 - Chernffonovy tváře – pro každou chladící vodu.



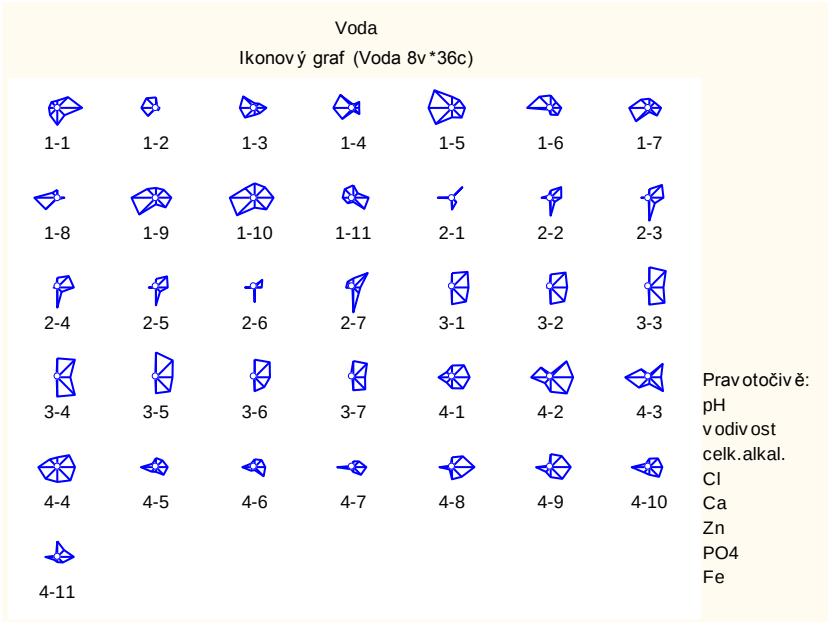
Graf č. 11 - Sun Ray Plot – graf slunečních paprsků
Počet paprsků odpovídá počtu proměnných. Střed každého paprsku představuje průměr odpovídající proměnná a jeho délka 2. n násobek směrodatné odchylky této proměnné, kde n je námi zadané číslo
Legendu k grafu s popisem jednotlivých paprsků poskytuje Plot Key – klíč



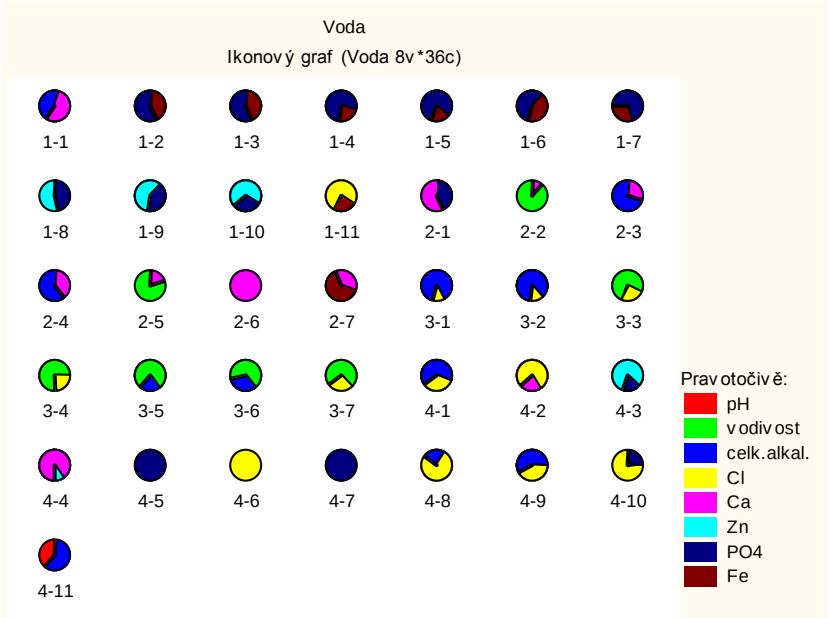
Počet paprsků odpovídá počtu proměnných. Střed každého paprsku představuje průměr odpovídající proměnná a jeho délka 2. n násobek směrodatné odchylky této proměnné, kde n je námi zadané číslo
Legendu k grafu s popisem jednotlivých paprsků poskytuje Plot Key – klíč

Graf č.12 - Star Symbol Plot – hvězdicový graf

Délka paprsku zde představuje relativní velikost hodnoty příslušného objektu. Konce paprsků jsou spojeny čárami. V případě velkého množství objektů je graf nepřehledný.
Klíč popisuje řazení jednotlivých paprsků.



Graf č. 13 - Ikonový graf - výseče



Vidíme že vzorek ze čtvrtého odběrového místa č.6 je úplně atypický a 5 a 7.Vzorky z prvního odběrového místa 2,3,4,5,7,8 a 9 jsou si podobné složením – Zinek a fosforečnany. U druhého odběrového místa je atypický vzorek č. 6. Vzorky z první skupiny jsou si podobné kromě č.1,6,10 a 11. V druhé skupině jsou odlišné 1,4 a 6.

3.4 Určení vnitřní struktury analýzou vícerozměrných dat
Mária Kalhousová

Závěr :
Z vyšetření indexového grafu úpatí vlastních čísel – Catelův indexový graf – jsme určili vhodný počet hlavních komponent. V našem případě tři. První hlavní komponenta nám popisuje 49,06% celkového rozptylu, druhá hl.komponenta popíše 23,58% a třetí 10,45%. První tři popíšou 83,09%. Pro dostatečné vysvětlení chování zdrojových proměnných požadujeme 85-90 % vysvětlené variability. V našem případě je patrné že zlom není moc zřejmý. Čtvrtá komponenta popíše 6,96%, pátá komponenta 5,23, šestá komponenta 2,68%, sedmá komponenta 1,56% a osmá 0,47%. Celkem 99,99 %.
Z grafu komponentních vah jsme určili - souvislost - vodivost , chloridy a vápník ,pH a alkalita. Druhá hlavní komponenta popisuje vztah – vápník – vodivost – pH a první hlavní popisuje vztah – celková alkalita a chloridy. Fosforečnany,zinek a železo mají záporné hodnoty a sestupnou tendenci korelačního vztahu.
Metoda hlavních komponent je užitečná pomůcka pro rozlišení odběrových míst. Odběrová místa 2 a 3 jsou dobře rozlišeny. U odběrových míst 1 a 4 to už není tak jednoznačné protože se částečně překrývají. Pro rozklad objektů do shluků jsem použila shlukovou analýzu.Objekty vytvořili nakonec 1 shluk. Velice podobné jsou shluky vodivost – vápník a zinek – fosforečnany. Pro vizuální zkoumání podobnosti objektů jsem použila ikonové grafy. Jako nejpřehlednější mi připadali výseče. Hodně pomohlo barevné rozlišení jednotlivých parametrů, vytvoření podobných skupin a odlišné objekty byly vidět už na první pohled.