VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV ELEKTROENERGETIKY

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF ELECTRICAL POWER ENGINEERING

Možnosti využití metod vícerozměrné statistické analýzy dat při hodnocení spolehlivosti distribučních sítí

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE **LUKÁŠ GESCHWINDER**
AUTHOR

BRNO 2009

VYSOKÉ UČENÍ
TECHNICKÉ V BRNĚ
Fakulta elektrotechniky
a komunikačních technologií
Ústav elektroenergetiky

Diplomová práce

magisterský navazující studijní obor
Elektroenergetika

Student: Geschwinder Lukáš, Bc. ***ID:*** 83792
Ročník: 2 ***Akademický rok:*** 2008/09

NÁZEV TÉMATU:

Možnosti využití metod vícerozměrné statistické analýzy dat při hodnocení spolehlivosti distribučních sítí

POKYNY PRO VYPRACOVÁNÍ:

1. Přehled metod vícerozměrné statistické anylýzy dat použitelných pro vybrané analýzy spolehlivosti distribuční sítě
2. Aplikace vybrané metody na zadaný vzorek dat
3. Zhodnocení získaných výsledků

DOPORUČENÁ LITERATURA:

podle pokynů vedoucího

Termín zadání: 9.2.2009 ***Termín odevzdání:*** 25.5.2009

Vedoucí práce: doc. Ing. Vladimír Blažek, CSc.
Konzultanti diplomové práce:

doc. Ing. Petr Toman, Ph.D.
předseda oborové rady

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ Fakulta elektrotechniky a komunikačních technologií Ústav elektroenergetiky 1

UPOZORNĚNÍ:

Autor diplomové práce nesmí při vytváření diplomové práce porušit autorská práva třetích osob, zejména nesmí zasahovat nedovoleným způsobem do cizích autorských práv osobnostních a musí si být plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení § 152 trestního zákona č. 140/1961 Sb.

>>Vložit licenční smlouvu<<

Bibliografická citace práce:

GESCHWINDER, L. Možnosti využití metod vícerozměrné statistické analýzy dat při hodnocení spolehlivosti distribučních sítí. Diplomová práce. Brno: Ústav elektroenergetiky FEKT VUT v Brně, 2009, 65 stran. Vedoucí diplomové práce doc. Ing. Vladimír Blažek, CSc.

Prohlašuji, že jsem svou **diplomovou práci** vypracoval samostatně a použil jsem pouze podklady (literaturu, projekty, SW atd.) uvedené v přiloženém seznamu.

Zároveň bych na tomto místě chtěl poděkovat vedoucímu diplomové práce doc. Ing. Vladimíru Blažkovi, CSc. a Ing. Petru Skalovi, Ph.D. z EGÚ Brno za cenné rady, poskytnutou literaturu a připomínky k mé práci a panu Mgr. Břetislavu Fajmonovi, Ph.D. z UMAT a svým rodičům za podporu během celé doby mého studia.

……

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav elektroenergetiky

Diplomová práce

Možnosti využití metod vícerozměrné statistické analýzy dat při hodnocení spolehlivosti distribučních sítí

Lukáš Geschwinder

vedoucí: doc. Ing. Vladimír Blažek, CSc.

Ústav elektroenergetiky, FEKT VUT v Brně, 2009

Brno

BRNO UNIVERSITY OF TECHNOLOGY

Faculty of Electrical Engineering and Communication

Department of Electrical Power Engineering

Master's Thesis

Possibilities of using multi - dimensional statistical analyses methods when evaluating reliability of distribution networks

by

Lukáš Geschwinder

Supervisor: doc. Ing. Vladimír Blažek, CSc.

Brno University of Technology, 2009

Brno

ABSTRAKT

Účelem je zhodnotit využití metod vícerozměrné statistické analýzy dat jako pomůcky při simulacích spolehlivosti distribuční sítě. Zvolené metody jsou shluková analýza (CLU) a metody hlavních komponent (PCA). CLU slouží k rozdělení objektů na základě jejich znaků a výpočtu vzdálenosti mezi objekty do skupin, jejichž vlastnosti by měly být podobné. Výstupem je možné odhalit skrytou strukturu v datech. PCA slouží k nalezení struktury ve znacích vícerozměrné matice dat. Znaky zde představují jednotlivé veličiny popisující daný objekt. PCA využívá rozložení původní matice dat na matici strukturní a šumovou. Jedná se o transformaci původní matice dat do nového souřadnicového systému hlavních komponent. Nové souřadnice se nazývají skóre. Hlavní komponenty tvoří ortogonální systém nových souřadnic. Distribuční síť z hlediska spolehlivosti může být charakterizována řadou statistických veličin. Spolehlivostní ukazatele mohou být: počet přerušení, doba přerušení. Integrální ukazatele spolehlivosti mohou být: ukazatel průměrné systémové četnosti přerušení (SAIFI) a ukazatel průměrné systémové doby trvání přerušení (SAIDI). V závěru je porovnání provedené simulace SAIFI podle negativně-binomického rozdělení a poskytnutých hodnot od distribuční společnosti. Je proveden pokus o popis závislostí znaků a rozdělení vývodů.

KLÍČOVÁ SLOVA: CLU; shluková analýza; PCA; metoda hlavních komponent; distribuční síť

ABSTRACT

The aim of this study is evaluation of using multi-dimensional statistical analyses methods as a tool for simulations of reliability of distribution network. Prefered methods are a cluster analysis (CLU) and a principal component analysis (PCA). CLU is used for a division of objects on the basis of their signs and a calculation of the distance between objects into groups whose characteristics should be similar. The readout can reveal a secret structure in data. PCA is used for a location of a structure in signs of multi-dimensional matrix data. Signs present separate quantities describing the given object. PCA uses a dissolution of a primary matrix data to structural and noise matrix data. It concerns the transformation of primary matrix data into new grid system of principal components. New conversion data are called a score. Principal components generating orthogonal system of new position. Distribution network from the aspect of reliability can be characterized by a number of new statistical quantities. Reliability indicators might be: interruption numbers, interruption time. Integral reliability indicators might be: system average interruption frequency index (SAIFI) and system average interruption duration index (SAIDI). In conclusion, there is a comparison of performed SAIFI simulation according to negatively binomial division and provided values from a distribution company. It is performed a test at description of sign dependences and outlet divisions.

KEY WORDS: CLU; cluster analysis; PCA; principal component analysis; distribution network

OBSAH

SEZNAM OBRÁZKŮ 12

SEZNAM TABULEK 14

SEZNAM SYMBOLŮ A ZKRATEK 15

1 ÚVOD 16

2 DATA A JEJICH ZÁKLADNÍ POPIS 16

3 ANALÝZA HLAVNÍCH KOMPONENT (PCA) 20

4 SHLUKOVÁ ANALÝZA (CLU) 22

4.1 HIERARCHICKÉ POSTUPY SHLUKOVÁNÍ 22

4.1.1 METODY METRIKY SHLUKOVÁNÍ 23

4.1.2 DENDROGRAM 23

4.2 NEHIERARCHICKÉ SHLUKOVÁNÍ 23

4.3 INTERPRETACE VÝSLEDKŮ 24

5 PŘÍKLAD 25

5.1 POPISNÉ CHARAKTERISTIKY VÍCEROZMĚRNÝCH NÁHODNÝCH VELIČIN A EXPLORATORNÍ ANALÝZA DAT 25

5.2 SHLUKOVÁ ANALÝZA 28

5.3 ANALÝZA HLAVNÍCH KOMPONENT 33

6 APLIKACE PCA 39

7 ROZDĚLENÍ VÝVODŮ PODLE CLU A DÉLEK 49

7.1 POSOUZENÍ VLIVNÝCH BODŮ 49

7.2 NEHIERARCHICKÉ SHLUKOVÁNÍ 51

7.3 HIERARCHICKÉ SHLUKOVÁNÍ 52

7.4 ROZDĚLENÍ PODLE DÉLEK 57

8 POSOUZENÍ ROZDĚLENÍ VÝVODŮ PODLE CLU 59

9 ZÁVĚR 64

POUŽITÁ LITERATURA 65

PŘÍLOHA A ROZDĚLENÍ VÝVODŮ NEHIERARCHICKOU SHLUKOVACÍ METODOU K-MEANS NA ČTYŘI SKUPINY 66

PŘÍLOHA B ROZDĚLENÍ VÝVODŮ HIERARCHICKOU SHLUKOVACÍ METODOU 68

PŘÍLOHA C ZÁKLADNÍ POPISNÉ STATISTIKY SKUPIN VÝVODŮ 70

PŘÍLOHA D ROZPTYLOVÝ DIAGRAM KOMPONENTNÍHO SKÓRE – DATA PO SLOUPCOVÉM CENTROVÁNÍ 73

PŘÍLOHA E ROZPTYLOVÝ DIAGRAM KOMPONENTNÍHO SKÓRE – DATA PO Z-SCORE 74

SEZNAM OBRÁZKŮ

Obr. 2-1 Příklad negativně – binomického a Poissonova rozdělení 19

Obr. 2-2 Vliv parametru r při p = 0,2 a parametru p při r = 1,2 na NBi 19

Obr. 5-1 Maticový graf pro Tab. 5-1 26

Obr. 5-2 Ikonový graf - Chernoffovy tváře 27

Obr. 5-3 Krabicový graf s mediánem 28

Obr. 5-4 Dendrogram s provedením Z-score 30

Obr. 5-5 Dendrogram proměnných se Z-score 31

Obr. 5-6 Dendrogram bez provedení Z-score 31

Obr. 5-7 Dendrogram proměnných bez provedení Z-score 32

Obr. 5-8 Graf vyjádření mezí rozdělených objektů do jednotlivých skupin podle dendrogramu s provedením Z-score v závislosti na vybraném znaku 32

Obr. 5-9 Cattelův indexový graf úpatí vlastních čísel 33

Obr. 5-10 Krabicový graf pro Tab. 5-12 34

Obr. 5-11 Rozptylový diagram komponentního skóre 35

Obr. 5-12 Rozptylový diagram komponentního skóre pro komponenty 1 a 3 35

Obr. 5-13 Dvojný graf (biplot) 36

Obr. 5-14 Indexový graf úpatí celkového reziduálního rozptylu (modrá čára) 38

Obr. 6-1 Maticový graf 40

Obr. 6-2 Krabicový graf pro Tab. 6-2 40

Obr. 6-3 Cattelův indexový graf úpatí vlastních čísel (čárkovaná čára) a indexový graf úpatí celkového reziduálního rozptylu (plná čára) pro sloupcově standardizovaná data 42

Obr. 6-4 Cattelův indexový graf úpatí vlastních čísel (čárkovaná čára) a indexový graf úpatí celkového reziduálního rozptylu (plná čára) pro data po autoškálování 42

Obr. 6-5 Rozptylový diagram komponentního skóre pro data sloupcově centrovaná (vlevo) a po provedení Z-score (vpravo) 43

Obr. 6-6 Rozptylový diagram komponentního skóre při použití Z-score pro 1. a 3. hlavní komponentu 43

Obr. 6-7Krabicový graf hlavních komponent pro skóre s daty sloupcově centrovanými (vlevo) a po autoškálování (vpravo) 43

Obr. 6-8 Rozptylový diagram komponentního skóre po roztřídění metodou K-means na 5 shluků pro data sloupcově centrovaná (vlevo) a po provedení Z-score (vpravo) 44

Obr. 6-9 Rozptylový diagram komponentního skóre po roztřídění metodou K-means na 4 shluky pro data sloupcově centrovaná (vlevo) a po provedení Z-score (vpravo) 44

Obr. 6-10 Dvojný graf pro data sloupcově centrovaná 45

Obr. 6-11 Dvojný graf pro data po provedení Z-score 45

Obr. 6-12 Rozptylový diagram komponentního skóre pro reprezentaci shluků vytvořených hierarchickým shlukováním, data po sloupcovém centrování, eukleidovská vzdálenost, metoda shlukování average (vlevo) a weighted, (vpravo) 46

Obr. 6-13 Rozptylový diagram komponentního skóre pro reprezentaci shluků vytvořených hierarchickým shlukováním, data po Z-score, eukleidovská vzdálenost, metoda shlukování average (vlevo) a weighted (vpravo) 46

Obr. 6-14 Dvojné grafy pro jednotlivé oblasti, data po sloupcovém centrování 47

Obr. 6-15 Dvojné grafy pro jednotlivé oblasti, data po Z-score 48

Obr. 7-1 Ikonový graf – hvězdy 49

Obr. 7-2 Dvojný graf bez vlivných bodů 50

Obr. 7-3 Příklad dendrogramu po odstranění vlivných bodů 51

Obr. 7-4 Krabicový graf pro sloupcově centrovaná data s výše uvedenými znaky 51

Obr. 7-5 Zobrazení jednotlivých skupin vývodů při použití sloupcového centrování po rozdělení podle K-means na čtyři skupiny 52

Obr. 7-6 Dendrogram proměnných, data po sloupcové standardizaci 53

Obr. 7-7 Dendrogram vývodů, data sloupcově centrovaná 53

Obr. 7-8 Dendrogram vývodů, data sloupcově centrovaná 53

Obr. 7-9 Zobrazení jednotlivých skupin vývodů při použití sloupcového centrování po rozdělení hierarchickým shlukováním, shlukovací metoda weighted, eukleidovská vzdálenost 54

Obr. 7-10 Zobrazení jednotlivých skupin vývodů při použití sloupcového centrování po rozdělení hierarchickým shlukováním, shlukovací metoda average, eukleidovská vzdálenost 55

Obr. 7-11 Dendrogram proměnných, data po Z-score 55

Obr. 7-12 Dendrogram vývodů, data po Z-score 56

Obr. 7-13 Dendrogram proměnných, data po Z-score 56

Obr. 7-14 Dendrogram vývodů, data po Z-score 56

Obr. 7-15 Zobrazení skupin ve znaku Suma poruch při třídění pomocí délek vývodů 57

Obr. 7-16 Rozptylový diagram komponentního skóre pro data sloupcově centrovaná (vlevo) a data autoškálovaná (vpravo), barvy odpovídají třídění podle délek 57

Obr. 7-17 Rozptylový diagram komponentního skóre pro data sloupcově centrovaná, barvy představují oblasti s vývody 58

Obr. 8-1 Závislost chyby SAIFI na počtu vývodů ve skupině 61

Obr. 8-2 Rozdílné tvary histogramů proložené negativně – binomickým rozdělením 62

Obr. 8-3 Rozdělení K-means, 4 skupiny, rozložení četností poruch, parametry r a p odpovídají příslušným hodnotám v Tab. 8-3 63

SEZNAM TABULEK

Tab. 5-1 Tabulka příkladu 25

Tab. 5-2 Základní popisné statistiky 25

Tab. 5-3 Korelační matice 26

Tab. 5-4 Kovarianční matice (n) 27

Tab. 5-5 Korigovaná kovarianční matice (n-1) 27

Tab. 5-6 Tabulka příkladu po provedení Z-score 28

Tab. 5-7 Eukleidovské vzdálenosti 29

Tab. 5-8 Tab. 5-7 po přepočtu 29

Tab. 5-9 Tab. 5-8 po přepočtu 29

Tab. 5-10 Rozvrh slučování 30

Tab. 5-11 Pro Cattelův indexový graf úpatí vlastních čísel, vlastní čísla z korelační matice 33

*Tab. 5-12 Komponentní skóre, matice **T*** 34

*Tab. 5-13 Vlastní vektory hlavních komponent, matice **P*** 36

Tab. 5-14 Komponentní váhy jednotlivých znaků, korelace mezi znaky a komponentami 36

*Tab. 5-15 Reziduum pro hlavní komponenty 1 a 2, matice **E*** 37

*Tab. 5-16 Reziduum pro hlavní komponenty 1 a 3, matice **E*** 37

Tab. 6-1 Kódy poruch, EGÚ Brno 39

Tab. 6-2 Základní popisné statistiky, počet případů je 321 39

Tab. 6-3 Základní popisné statistiky, počet případů je 321 41

Tab. 6-4 Korelační matice původní matice dat 41

Tab. 6-5 Vlastní čísla korelační matice a vyšetření potřebných hlavních komponent při sloupcovém centrování dat 41

Tab. 6-6 Vlastní čísla korelační matice a vyšetření potřebných hlavních komponent při autoškálování dat 42

Obr. 6-15 Dvojné grafy pro jednotlivé oblasti, data po Z-score 48

Tab. 7-1 Vlastní čísla korelační matice a vyšetření potřebných hlavních komponent (s vlivnými body) po provedení Z-score 49

Tab. 7-2 Korelační matice bez vlivných bodů 50

Tab. 7-3 Porovnání kofenetických korelačních koeficientů pro různé metody shlukování 50

Tab. 8-1 Přehled chyb simulace oproti dodaným datům 59

Tab. 8-2 Rozdělení vývodů podle délek 59

Tab. 8-3 Rozdělení vývodů shlukovacími metodami 60

SEZNAM SYMBOLŮ A ZKRATEK

$\boldsymbol{A}$	matice
$\boldsymbol{C}$	kovarianční matice
cc	kofenetický korelační koeficient
c_{il}	střední hodnota i-tého znaku v patřičném shluku
$D(x)$	rozptyl náhodné veličiny
$\boldsymbol{E}$	matice reziduí
$E(X)$	střední hodnota veličiny X
$F(x)$	distribuční funkce
$f(x)$	hustota pravděpodobnosti
k	počet shluků
m	počet sloupců (znaků, proměnných)
N	obor přirozených čísel
N^0	obor přirozených čísel včetně nuly
n	počet řádků (objektů)
n_k	počet objektů v k-tém shluku
n_u	počet přerušení v u-tém uzlu za dané časové období
O_c	celkový počet odběrných míst
O_u	počet odběrných míst v u-tém uzlu
P	pravděpodobnost
$\boldsymbol{P}$	matice komponentních vah
p	počet hlavních komponent; parametr negativně – binomického rozdělení
p_{ij}	projekce souřadnice délek i-tého bodu do j-té komponenty
$\boldsymbol{R}$	korelační matice
r	odhad korelačního koeficientu; parametr negativně – binomického rozdělení
$\boldsymbol{S}$	výběrová korigovaná kovarianční matice, odhad
$\boldsymbol{S}_o$	odhad kovarianční matice
s^2	odhad rozptylu
$\boldsymbol{T}$	matice komponentního skóre
$\boldsymbol{v}$	vlastní vektor
$\mathrm{WSS_K}$	kritérium věrohodnosti proložení mezishlukové sumy čtverců pro standardizovaná data
$\boldsymbol{X}$	zdrojová matice experimentálních dat (n;m)
X	náhodná veličina
$\boldsymbol{x}$	vektor naměřených hodnot
x_j	hodnoty diskrétního rozdělení
$\tilde{x}$	medián
$\hat{x}$	modus
y_{ij}	data po standardizaci
$\delta(x)$	chyba veličiny x
δ_{ijl}	chybějící hodnota
λ	vlastní číslo
$\boldsymbol{\mu}$	odhad střední hodnoty náhodné veličiny, vektor; výběrový průměr
ρ	korelační koeficient
σ	směrodatná odchylka
CLU	Cluster Analysis
CV	poruchy způsobené cizími vlivy
DTS	distribuční transformátor
N	nezjištěné poruchy
NBi	negativně – binomické rozdělení
PCA	Principial Component Analysis
PU	poruchy způsobené provozem a údržbou
SAIFI	ukazatel průměrné systémové četnosti přerušení
SVD	Singular Value Decomposition
VV	poruchy způsobené vnějšími vlivy
ZP	poruchy související se započetím provozu

1 ÚVOD

Metody vícerozměrné analýzy dat nám mohou odhalit na první pohled ne zrovna zřejmé struktury ve vícerozměrných datech. Závěr může být při nehodnotných datech pouze popisem výsledků získaných z programu. Analýza je takřka nemyslitelná bez použití výpočetní techniky. Soubory dat jsou velké a vícerozměrné, tj. obsahují pro každý objekt více než jednu popisovanou vlastnost. Podle mého názoru má smysl hodnocení spolehlivosti sítě v dnešní době pouze z ekonomického hlediska a to při penalizaci za nedodávku a za náklady spojené s opravami. Spolehlivost je vlastnost prvku nebo systému vykonávat svou činnost ve stanovených mezích za daných normálních provozních podmínek. Informaci o spolehlivosti distribuční sítě nám může přinést řada ukazatelů, jako je počet přerušení za rok, délka přerušení v minutách a integrální ukazatele SAIFI (ukazatel průměrné systémové četnosti přerušení) a SAIDI (ukazatel průměrné systémové doby trvání přerušení), které zveřejňuje Energetický regulační úřad za celou síť pro jednotlivé distribuční společnosti. Vést kabel je jedno řešení, ale určitě nákladnější i při poruchách. Proč nechtít nalézt strukturu i ve zdánlivě nesouvisejících znacích. Při tomto se musí dávat obzvlášť pozor, aby nedošlo k chybnému výkladu příčinné souvislosti. Úkolem je nalézt takové rozdělení vývodů, které by svými počty poruch a dalšími ukazateli odpovídalo simulacím. Dnes je vývod charakterizován měrnou poruchovostí, což je počet poruch na 100 kilometrů za rok.

2 DATA A JEJICH ZÁKLADNÍ POPIS

Rozdělení pravděpodobnosti je charakterizováno distribuční funkcí v bodě x $F(x)$, která určuje pravděpodobnost toho, že náhodná veličina X bude mít hodnotu menší nebo rovnu x. Diskrétní rozdělení lze popsat též pravděpodobnostní funkcí $P(X = x)$ a spojité hustotou pravděpodobnosti $f(x)$, což je derivace distribuční funkce.

Naměřená data jsou většinou rozsáhlá, proto chceme tento soubor dat popsat několika čísly zvanými charakteristiky náhodné veličiny. Pro popis dat se zavádějí různé charakteristiky, které více či méně vystihují naměřená data. Naměřená data se posuzují podle ukazatelů zvaných statistiky. Statistiky umožňují provést závěr o rozdělení pravděpodobnosti náhodných veličin a jejich příslušných charakteristik. Experimentální data mají někdy daleko do teoretického rozdělení. Asi nejpopulárnější charakteristika je střední hodnota (očekávaná hodnota, matematická naděje) veličiny X, značí se $E(X)$. Protějškem této charakteristiky je statistika výběrový průměr daná vztahem pro vícerozměrná data

$$\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \,, \tag{2.1}$$

kde n je počet případů, s rostoucím n se výběrový průměr přibližuje střední hodnotě rozdělení a $\boldsymbol{x}$ je vektor hodnot. Tento vztah platí pro symetrické rozdělení podle bodu μ, což je střední hodnota. Ten je velmi náchylný na odlehlé hodnoty. Obecně je střední hodnota dána pro diskrétní rozdělení vztahem

$$E(X) = \sum_{x_j} x_j P(X = x_j) \text{ pro } j \in N \tag{2.2}$$

a pro spojité rozdělení

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx \,. \tag{2.3}$$

Dalšími charakteristikami jsou kvantily, které jsou jednoznačně určeny pro spojité a rostoucí distribuční funkce. Kvantil je číslo, pro nějž platí, v pravděpodobnosti $0 < P < 1$,

$$F(x_P) \le P \wedge F(x_P + 0) \ge P \,. \tag{2.4}$$

Kvantil $x_{0,5}$ se nazývá medián (označení $\tilde{x}$), který je definován hodnotou vzatou ze středu seřazených naměřených hodnot podle velikosti a stanovuje, že 50% hodnot v souboru dat leží nad i pod mediánem. Dalšími kvantily jsou horní ($x_{0,75}$) a dolní kvartil ($x_{0,25}$). Mezi nimi je oblast nazývaná mezikvartilové rozpětí, které obsahuje právě 50% hodnot.

Charakteristika polohy určuje posunutí distribuční funkce o konstantu. Rozdělení dvou náhodných veličin posunutých o konstantu se od sebe liší pouze polohou. Modus ($\hat{x}$) je jednou z charakteristik polohy a je to hodnota s největší četností. Jinými jsou z jiného pohledu na charakteristiku polohy také medián a statistika aritmetický průměr. Pro symetrické rozdělení se tyto tři veličiny sobě rovnají.

Histogram vyjadřuje graficky četnost výskytu daných měření v souboru dat. Pro spojitou veličinu by se měl blížit teoretickému normálnímu rozdělení.

Obsahují-li data odlehlé hodnoty, zkreslují průměr, z toho důvodu se zavádí veličina rozptyl (variance, druhý centrální moment) $D(X)$, která udává, jak moc se hodnoty v souboru dat liší od průměru. Odhad rozptylu (výběrový rozptyl) se určí ze vztahu

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{2.5}$$

nebo také

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2 . \tag{2.6}$$

Druhá odmocnina z výběrového rozptylu je výběrová směrodatná odchylka. Obecně je vztah pro rozptyl

$$D(X) = E\{[X - E(X)]^2\}. \tag{2.7}$$

Mírou lineární závislosti mezi dvěma proměnnými je kovariance $\mathrm{cov}(x_1; x_2)$. Kovariance záleží na měřítku veličin, mezi kterými se provádí. Uvádí se pro dvě proměnné. Kladná hodnota kovariance značí přímou vazbu mezi proměnnými a záporná nepřímou vazbu. Je to střední hodnota ze součinu odchylek veličin x_1 a x_2 od jejích středních hodnot. Kovariance je definována např. vztahem pomocí středních hodnot

$$\mathrm{cov}(x_1, x_2) = E(x_1 x_2) - E(x_1)E(x_2). \tag{2.8}$$

Pro m znaků dostaneme kovarianční matici, která je definována

$$\boldsymbol{C} = \begin{bmatrix} D(x_1) & \mathrm{cov}(x_2, x_1) & \cdots & \mathrm{cov}(x_m, x_1) \\ \mathrm{cov}(x_1, x_2) & D(x_2) & \cdots & \mathrm{cov}(x_m, x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(x_1, x_m) & \mathrm{cov}(x_2, x_m) & \cdots & D(x_m) \end{bmatrix}. \tag{2.9}$$

Matice je symetrická a prvky na diagonále jsou rovny rozptylu.

Odhad kovarianční matice

$$\boldsymbol{S}_o = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T . \tag{2.10}$$

Výběrová korigovaná kovarianční matice

$$\boldsymbol{S} = \frac{n}{n-1}\boldsymbol{S}_o \tag{2.11}$$

Standardizací vzhledem k součinu rozptylů získáme veličinu zvanou korelační koeficient. Korelační koeficient se nemění po provedení standardizace dat. Je roven střední hodnotě součinu proměnných po standardizaci.

Koeficient korelace

$$\rho(x, y) = \frac{\mathrm{cov}(x, y)}{\sigma(x)\sigma(y)}, \tag{2.12}$$

kde σ je směrodatná odchylka. Tento koeficient není možno určit, proto se zavádí jeho odhad.

Odhad koeficientu korelace mezi dvěma proměnnými

$$r_{x1x2} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_{i1}x_{i2} - \frac{1}{n}\sum_{i=1}^{n} x_{i1} \cdot \frac{1}{n}\sum_{i=1}^{n} x_{i2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{i1}-\mu_1)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{i2}-\mu_2)^2}} \tag{2.13}$$

a zřejmě platí

$$r_{x1x2} = \frac{\frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_{i1}x_{i2} - \frac{1}{n}\sum_{i=1}^{n} x_{i1} \cdot \frac{1}{n}\sum_{i=1}^{n} x_{i2}\right)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{i1}-\mu_1)^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{i2}-\mu_2)^2}}. \tag{2.14}$$

Korelační koeficient může nabývat hodnot $r \in \langle -1;1 \rangle$.

Korelace zkoumá míru lineární závislosti. Je-li korelační koeficient roven nule, neznamená to ještě, že mezi veličinami nemůže být jiná závislost, např. kvadratická. Jde pouze o statistickou hodnotu a jako taková nemůže být výkladem příčin, ale jen nástrojem.

Vybočující body lze vizuálně identifikovat jen u jednorozměrných dat, kde se jedná o odlehlé body v grafu. Lze je vidět např. v krabicovém grafu, který zobrazuje předem zvolené statistické charakteristiky. Odstraněním některých bodů může dojít ke zborcení struktury dat. U vícerozměrných dat můžeme vybočující objekty nalézt např. rozptylovým grafem (zobrazení vícerozměrných dat ve dvou rozměrech a to závislostí znaků mezi sebou navzájem, lze jím i zobrazit míru korelace) nebo symbolovými diagramy (tváře, polygony, křivky).

Symbolové diagramy v sobě zahrnují pro jednotlivé objekty všechny znaky různě graficky vyjádřené tzv. symboly, např. u tváří je třeba pro jeden znak symbolem šířka hlavy, pro druhý je symbolem tvar protažení nosu atd. Vlastnosti objektů se posuzují na základě viditelných rozdílů mezi reprezentujícími obrázky.

Standardizace dat se provádí, jsou-li data mezi jednotlivými znaky řádově odlišná, ale to neplatí vždy. Pokud jsou data určitým způsobem přirozeně spjata, standardizace se nedoporučuje. Jako ve všech případech by se mělo vycházet ze zadané úlohy a jejího charakteru. Použitím dojde k odstranění vlivu hodnot na jednotkách, proměnné se v jednotné stupnici lépe porovnávají. Zobecněný pojem je škálování, které zahrnuje změnu počátku stupnice a odstranění vlivu rozdílného měřítka. Technikou sloupcového centrování je nové centrum stupnice v nule místo v průměru původního sloupce před centrováním.

$$y_{ij} = x_{ij} - \mu_j \tag{2.15}$$

Sloupcová standardizace – původní hodnoty ve sloupci jsou poděleny směrodatnou odchylkou sloupce, pomocí vztahu (2.6).

Autoškálování (studentizace, Z-score) je kombinací předchozích dvou technik. Získáme data s průměrem pro každý sloupec rovným nule a směrodatnou odchylkou rovnou jedné. A další způsoby.

Vztah pro Z-score, směrodatná odchylka vypočítána podle vztahu (2.6).

$$y_{ij} = \frac{x_{ij} - \mu_j}{s_j} \tag{2.16}$$

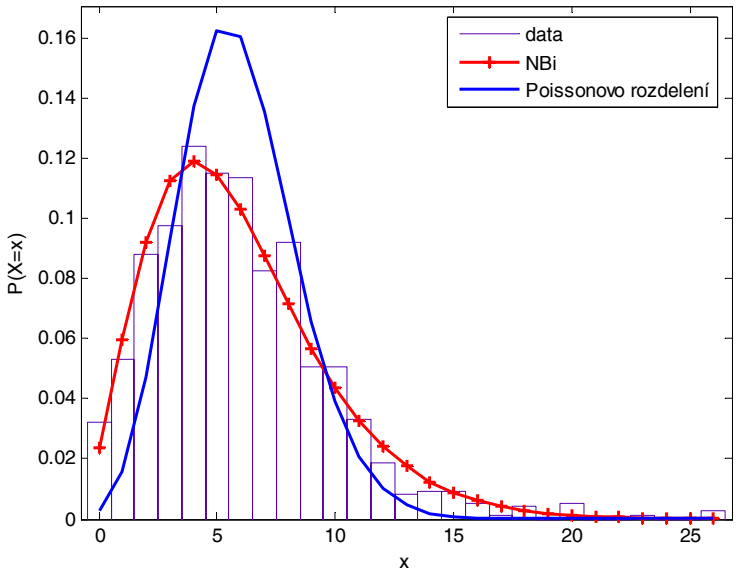
Jelikož předložené data obsahují počet poruch, což jsou diskrétní hodnoty, je pro rozdělení pravděpodobnosti možno použít jedno z diskrétních rozdělení, a to Poissonovo nebo negativně – binomické. Negativně - binomické rozdělení pravděpodobnosti (NBi) je definováno pomocí rozdělení gama vztahem

$$P(X=x)=\frac{\Gamma(r+x)}{x!\Gamma(r)}p^r(1-p)^x, \tag{2.17}$$

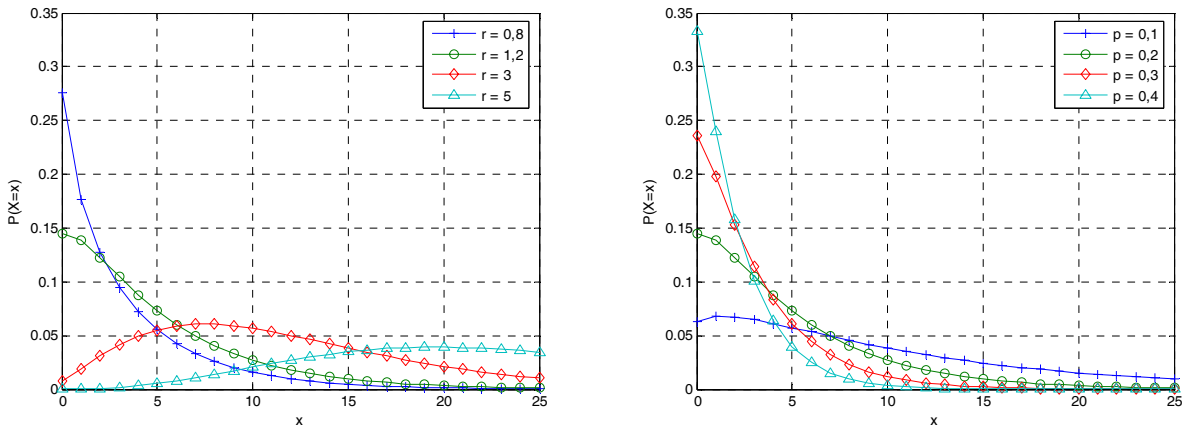
kde parametry r a p mohou nabývat následujících hodnot $r>0$; $p\in(0;1)$ a $x\in N^0$.

Je-li r přirozené číslo, NBi se nazývá Pascalovo rozdělení a je dáno pomocí kombinačního čísla vztahem

$$P(X=x)=\begin{cases}\binom{x+r-1}{r-1}p^r(1-p)^x & x\in N^0\\ 0 & \text{jinak}\end{cases}. \tag{2.18}$$



Obr. 2-1 Příklad negativně – binomického a Poissonova rozdělení



Obr. 2-2 Vliv parametru r při p = 0,2 a parametru p při r = 1,2 na NBi

3 ANALÝZA HLAVNÍCH KOMPONENT (PCA)

Metoda byla vytvořena pro psychologické účely, kdy se věřilo ve spojení IQ a kriminality s fyziologickými charakteristikami a sociálním prostředím. Slouží k zjednodušení popisu vzájemně korelovaných znaků, metoda rozloží původní zdrojovou matici $\boldsymbol{X}$ na matici šumovou a matici strukturní. Je vhodná pro případy, kdy sloupce matice $\boldsymbol{X}$ jsou silně korelovány. Vytvoří se komponenty (1, 2, …, m), které tvoří ortogonální systém nových souřadnic. Hlavních komponent je min(m; n). První hlavní komponenta má obsahovat co největší část z rozptylu dat a druhá hlavní komponenta naopak neobsahovat největší část rozptylu v první. Všechny hlavní komponenty jsou navzájem nekorelované, to vyplývá z toho, že jsou ortogonální. První hlavní komponenta má největší rozptyl, poslední nejmenší. Hlavní komponenty jsou vhodně škálované vektory v prostoru znaků, alespoň centrováním. Čím je rozptyl znaku větší, tím je více informace o variabilitě v původních datech obsaženo. Má-li nějaký znak malý rozptyl, nemůže ve značné míře přispívat k rozlišení mezi objekty.

Hlavní myšlenka metody spočívá v lineární transformaci souřadnicového systému matice $\boldsymbol{X}$ do nového souřadnicového systému hlavních komponent. Metoda slouží ke snížení dimenze úlohy, avšak nemá vliv na původní polohu a počet znaků. Jakýkoliv znak může přispívat do více než jedné komponenty. Každý objekt je vyjádřen v novém souřadnicovém systému podle souřadnic zvaných skóre.

První hlavní komponenta

$$y_1 = \sum_{j=1}^{m} v_{1j} x_1 = \boldsymbol{v}_1^T \boldsymbol{x} \tag{3.1}$$

podmínka pro určení vlastního vektoru $\boldsymbol{v}_2$.

$$\operatorname{cov}(y_1; y_2) = 0 \tag{3.2}$$

$$\boldsymbol{v}_j^T \boldsymbol{v}_j = \lambda_j \tag{3.3}$$

Rozptyl je maximální pro vektor koeficientů $\boldsymbol{v}_1$, tj. má vyjadřovat největší proměnlivost.

$$D(y_1) = \boldsymbol{v}_1^T \boldsymbol{R} \boldsymbol{v}_1, \tag{3.4}$$

kde místo korelační matice $\boldsymbol{R}$ může být i matice kovarianční.

Součet rozptylů ze sloupců matice $\boldsymbol{X}$ je roven součtu rozptylů hlavních komponent, je možné usuzovat na míru proměnlivosti v datech hlavních komponent v tzv. Cattelově indexovém grafu úpatí vlastních čísel. Je-li součet míry proměnlivosti prvních hlavních komponent větší než 80 % - 90 % lze brát v úvahu jen tyto. Rozdíl mezi souřadnicemi objektů v původních proměnných a v hlavních komponentách se nazývá chybou modelu PCA.

Vlastní číslo λ je definováno

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}, \tag{3.5}$$

kde $\boldsymbol{A}$ je čtvercová matice řádu n.

Normalizační podmínka $\boldsymbol{v}_j^T \boldsymbol{v}_j = \lambda_j$ pro analýzu podle korelační matice a $\boldsymbol{v}_j^T \boldsymbol{v}_j = 1$ pro matici kovariance.

Nejprve se provede sloupcové centrování matice $\boldsymbol{X}$. V metodě se rozděluje matice dat $\boldsymbol{X}$ na matici strukturní a šumovou

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T + \boldsymbol{E} = \text{struktura} + \text{šum}, \tag{3.6}$$

kde $\boldsymbol{T}$ je matice komponentního skóre, $\boldsymbol{E}$ je matice reziduí a $\boldsymbol{P}$ je matice komponentních vah. Matice $\boldsymbol{P}$ v sobě nese informaci, v jakém vztahu jsou hlavní komponenty s původními osami. Souvisí se směrovými kosiny mezi původními osami a hlavními komponentami. $\boldsymbol{E}$ nám ukazuje tzv. těsnost proložení, tj. jak dobře model znázorňuje proložení objekty. Uživatel musí zvolit počet hlavních komponent tak, aby šumová matice nebyla větší než experimentální chyba dat.

$$\boldsymbol{X} = \boldsymbol{t}_1 \boldsymbol{p}_1^T + \boldsymbol{t}_2 \boldsymbol{p}_2^T + \ldots + \boldsymbol{t}_A \boldsymbol{p}_A^T + \boldsymbol{E}\,, \tag{3.7}$$

kde součiny $\boldsymbol{tp}^T$ mají hodnost 1 a určí se pomocí SVD (Singular Value Decomposition) algoritmu (MATLAB) a vyčíslují se postupně, až je zobrazen dostatečný počet hlavních komponent. SVD převede matici $\boldsymbol{A}$ na ortogonální matice $\boldsymbol{U}$ a $\boldsymbol{V}$ a matici diagonální $\boldsymbol{S}$.

$$\boldsymbol{A} = \boldsymbol{U}.\boldsymbol{S}.\boldsymbol{V}^T \tag{3.8}$$

Cattelův indexový graf úpatí vlastních čísel

Slouží k určení, kolik hlavních komponent je nutno použít, aby bylo dosaženo nejlepšího proložení původních dat. Vyjadřuje relativní velikost vlastních čísel v závislosti na indexu. Zlomové místo nastává v úpatí, kde za ním jsou již komponenty s malým rozlišením dat, tedy malým rozptylem a vyjadřuje počet potřebných hlavních komponent. Průměrná hodnota vlastních čísel je rovna průměrné hodnotě rozptylu původních znaků. Z tohoto tvrzení se vyloučí hlavní komponenty, jejichž vlastní čísla jsou pod průměrem. Uvažujeme-li při analýze kovarianční matici $\left(\sum_{i=1}^{m} \lambda_i = \mathrm{tr}\boldsymbol{S}\right)$, pro korelační matici je její stopa rovna m a průměr je jedna, proto stačí brát v úvahu jen vlastí čísla korelační matice větší než jedna. Na základě simulací se doporučuje volit vlastní čísla větší než 0,7.

Graf komponentních vah (zátěží)

Zobrazuje komponentní zátěže (matice $\boldsymbol{P}$) pro různé kombinace hlavních komponent a znázorňuje vzdálenosti mezi znaky, krátká vzdálenost znamená silnou korelaci. Podle průmětu spojovací úsečky počátku a bodu reprezentujícím znak do os nového souřadnicového systému, můžeme vidět, jak jednotlivé znaky přispívají svou proměnlivostí do hlavní komponenty. Znaky s malou důležitostí leží blízko počátku. Znaky s úhlem 90° nejsou korelované s kolmou komponentou. Leží-li znaky blízko sebe, vypovídá to o jejich silné kovarianci.

Rozptylový diagram komponentního skóre

Zobrazuje komponentní skóre (matice $\boldsymbol{T}$) pro zvolenou kombinaci hlavních komponent pro jednotlivé objekty. V diagramu lze nalézt objekty odlehlé a tvořící skupiny, to jsou ty, jež leží blízko u sebe. Rovnoměrné rozložení objektů na ploše se považuje za ideální. Pro odlehlé body se někdy doporučuje použít rozptylový diagram ze dvou posledních hlavních komponent. Pro analýzu více než dvou komponent současně se berou v úvahu jen průměty do roviny. Umístění objektů blízko počátku jsou typickými představiteli, daleko jsou extrémy. Shluky objektů by měli být dobře oddělené, jinak lze očekávat podobnost s blízkým shlukem. Pokud neodpovídá rozložení objektů po celé ploše, může to ukazovat na nedobrý model, obvykle se nachází v datech silně odlehlý objekt. To má za následek shromáždění ostatních objektů na jedno malé místo a nelze je od sebe dobře rozlišit. Po odstranění by mělo dojít k rozptýlení namačkaných bodů reprezentující zbylé objekty.

Dvojný graf

Je zobrazením předchozích dvou grafů současně. U grafu komponentních vah dojde ke změně měřítka a to na měřítko komponentního skóre (tak to dělá Matlab) podle vztahu

$$\frac{p_{i1}^2}{\lambda_1} + \frac{p_{i2}^2}{\lambda_2} = 1\,, \tag{3.9}$$

kde p_{ij} jsou projekce souřadnice délek i-tého bodu do j-té komponenty.

Můžeme jím zdůvodnit umístění objektu pomocí znaků umístěnými v grafu podle matice komponentních vah. Nachází-li se objekt v blízkosti znaku, má to význam ten, že velký podíl informace pro umístění tohoto objektu pochází z tohoto znaku. To umožňuje interpretaci rozmístění znaků a objektů pomocí vzájemné polohy v nových souřadnicích.

4 SHLUKOVÁ ANALÝZA (CLU)

Shluková analýza slouží ke klasifikaci objektů do shluků, sjednocuje podobné objekty podle podobnosti na základě zvolených znaků tak, aby si byly objekty podobné ve shluku a dostatečně nepodobné mimo tento shluk. Zvolit tyto znaky je zásadní úkol. Měly by být voleny tak, aby rozlišovaly objekty. Takto klasifikované objekty by měly mít stejné vlastnosti, které nejsou na první pohled zřejmé ze získané matice dat $\boldsymbol{X}$. Numerická taxonomie se využívá v biologii ke klasifikaci rostlin a živočichů. CLU nerozlišuje významnost znaků, ale pouze odlišení shluků.

Jsou formulovány tři hlavní cíle CLU:

- popis systematiky
- zjednodušení dat – shluky jsou představovány jejich vlastnostmi
- identifikace vztahů

Odlehlé objekty značně zkreslují analýzu, stejně tak i nevýznamné znaky, a můžou vést k tomu, že nebude nalezena odpovídající struktura daných objektů. Ale na druhé straně odstraněním těchto objektů může dojít ke ztrátě původní struktury dat. Analýza nerozlišuje významné a nevýznamné proměnné.

Pro vyjádření podobnosti mezi objekty se používá různých metrik, které vyjadřují geometrické rozložení objektů. Souřadnice objektů vyjadřují jednotlivé znaky. Stejně jako na vše má vliv nestejné měřítko měřených veličin a škálováním se změní vytvořené pořadí podobností. Na míru podobnosti má velký vliv rozptyl znaku.

Eukleidovská metrika – přímá vzdálenost mezi body, zobrazuje v rovině kružnici

$$d(\boldsymbol{x}_k;\boldsymbol{x}_g)=\sqrt{\sum_{j=1}^{m}\left(x_{jk}-x_{jg}\right)^2}\ . \tag{4.1}$$

Metrika městských bloků (Hammingova metrika, manhattanská vzdálenost) – vzdálenost mezi objekty je dána součtem velikosti dvou navzájem kolmých úseček vycházejících z bodů reprezentující objekty.

$$d(\boldsymbol{x}_k;\boldsymbol{x}_g)=\sum_{j=1}^{m}\left|x_k-x_g\right| \tag{4.2}$$

Zobecněním předchozích dvou metrik dostaneme Minkovského metriku danou vztahem

$$d(\boldsymbol{x}_k;\boldsymbol{x}_g)=\sqrt[z]{\sum_{j=1}^{m}\left(x_{jk}-x_{jg}\right)^z}\ , \tag{4.3}$$

která s rostoucím z zdůrazňuje rozdíly mezi vzdálenými objekty.

Mahalanobisova vzdálenost - zobrazuje elipsu, uvažuje lineární závislost mezi znaky pomocí kovarianční matice, tato vzdálenost je statistickou mírou, umožňující určit vzdálenost v systému, jehož osy nemusí být ortogonální. Když jsou znaky korelovány, zdá se, že tato metrika je nejlepší, protože váží všechny znaky stejně. Dokáže kompenzovat vliv multikolinearity (značí silnou korelovatelnost mezi sloupci, které mohou být popsány ostatními sloupci, tento jev je třeba z dat odstranit vyloučením příslušného znaku).

$$d(\boldsymbol{x}_k;\boldsymbol{x}_g)=\sqrt{\left(\boldsymbol{x}_k-\boldsymbol{x}_g\right)^T\boldsymbol{C}^{-1}\left(\boldsymbol{x}_k-\boldsymbol{x}_g\right)} \tag{4.4}$$

4.1 Hierarchické postupy shlukování

U aglomeračního shlukování se shluky tvoří postupně, z matice vzdáleností se vezme nejmenší vzdálenost mezi dvěma objekty. Přesněji řečeno, většina metod pracuje na principu nepodobnosti než podobnosti. Ty vytvoří první shluk, poté se vypočte nová matice vzdáleností s prvním shlukem vcelku a jsou vynechány objekty z prvního shluku. To se opakuje, dokud není

vytvořen jeden shluk se všemi objekty nebo není naplněn zadaný počet shluků. Divizní shlukování je opačné, z velkého shluku jsou dělením získány shluky o jednom prvku.

4.1.1 Metody metriky shlukování

Průměrová metoda (weighted) – vzdálenost mezi shluky je průměrem všech objektů v jednotlivých shlucích.

Metoda skupinového průměru (group average)

Centroidní metoda zvaná také metoda těžiště (centroid) – vzdálenost shluků je určena jako eukleidovská vzdálenost mezi jejich těžišti nebo čtvercem této vzdálenosti. Souřadnice těžiště shluku jsou vyjádřen průměrnými hodnotami znaků pro všechny objekty v odpovídajícím shluku. Pro každý krok shlukování se počítá nové těžiště. Je zde menší ovlivnění odlehlými body.

Metoda nejbližšího souseda (single) – nachází se minimální vzdálenost mezi objekty z matice vzdáleností a vytvoří se jeden shluk. Další shluk je vytvořen přidáním dalšího nejbližšího objektu. Dochází k jevu nabalování a ve shlucích může nastat případ, že i při největší blízkosti objektů nepůjde o nejbližší shluky.

Metoda nejvzdálenějšího souseda (complete) – ta hledá oproti metodě nejbližšího souseda maximální vzdálenost mezi objekty.

Mediánová metoda (median) – vylepšená metoda těžiště, odstraňuje významnosti různě velkých shluků, které jim metoda těžiště dává.

Wardova metoda – základ tvoří čtverec euklidovské vzdálenosti, která má největší rozptýlení mezi shluky.

Rozdíl mezi jednotlivými shlukovacími metodami je v tom, jaký vytvářejí prostor mezi vznikajícími shluky. Metoda nejbližšího souseda prostor zužuje, metoda průměrné vzdálenosti, mediánová a těžiště prostor zachovávají a Wardova s metodou nejvzdálenějšího souseda prostor rozšiřují. Dlouhodobým pozorováním vychází nejlépe metoda průměrné vzdálenosti.

4.1.2 Dendrogram

Zobrazením CLU je dendrogram podobnosti objektů a dendrogram podobnosti proměnných. V diagramu je zanesena vzdálenost mezi jednotlivými shluky. Řezem dendrogramem se určují prvky v jednotlivých shlucích. Počet shluků je dán počtem řezu čar vzdáleností. Míra věrohodnosti dendrogramu je provázena určením koeficientu kofenetické korelace *cc*. Jde o Pearsonův korelační koeficient mezi skutečnou a predikovanou vzdáleností určenou z dendrogramu, čím je hodnota koeficientu bližší jedné, tím je proložení lepší. Nebo je míra věrohodnosti dána delta kritériem, které měří stupeň přetvoření struktury dat, kde se zase hodnota koeficientu musí blížit nule. Interpretace dendrogramu je jednoduchá, objekty spojené nízko jsou si velmi podobné a objekty spojené vysoko značí nepodobnost.

Problémy: volba výpočtu vzdálenosti mezi objekty, volba shlukovací procedury.

4.2 Nehierarchické shlukování

Průnik podmnožin množiny dat $\boldsymbol{X}$ je prázdná množina, na rozdíl od předešlé metody shlukování, jak je patrné na dendrogramu. Nehierarchické postupy shlukování mají výhodu oproti hierarchickým v tom, že na základě teorie lze vybrat zárodky shluků a tím je méně náchylná na odlehlé a nevýznamné body. Existuje celá řada metod, se kterou pracuje Matlab 7 ve svém Toolboxu je K-means.

K-means (shlukování metodou nejbližších těžišť) [6]

Počet shluků k musí být předem zadán, objekt je zařazen do shluku právě tehdy, když je jeho vzdálenost nejmenší od těžiště. Jedná se o iterační proces. Metoda vybere lokální optimální řešení pro daný soubor dat a je to takové uspořádání, kdy nedojde ke snížení mezishlukové sumy čtverců WSS_K. Pro předem známé těžiště shluků se provádí shlukování pomocí hierarchické

metody, kde se určí počet shluků a počáteční těžiště shluků. Vytvoří se prostorové shluky nalezením souboru těžišť shluků s přiřazením objektu do shluku, tím se určí nové shluky a děj se opakuje. Nejsou-li známy těžiště shluků, musí se použít jiných metod. Třeba vybíraní objektů, jež mají velké vzdálenosti mezi sebou, jejichž počet je počtem shluků. Objekt se přiřadí k těžišti, bude-li jeho vzdálenost větší než k ostatním těžištím a nahradí ho, bude-li jeho vzdálenost větší než vzdálenost mezi dvěma nejbližšími těžišti. Těžiště, které je bližší k objektu, je pak vyměněno. Objekt se dosadí na místo těžiště, když vzdálenost z objektu k těžišti je větší než nejmenší vzdálenost mezi těžištěm a všemi ostatními těžišti. Znovu se vymění těžiště nejtěsnější k němu. Počáteční přiblížení ovlivňuje konečné uspořádání shluků. Proto algoritmus pro každý pokus zcela náhodně přiřazuje každý objekt jednomu shluku. Toto uspořádaní je pak optimalizováno. Start procesu z rozličných náhodných uspořádání velmi zvýší pravděpodobnost nalezení nejlepšího řešení, globálního optima počtu shluků.

Kritérium věrohodnosti proložení mezishlukové sumy čtverců pro standardizovaná data

$$WSS_K = \frac{nm}{nm - m} \sum_{l=1}^{k} \sum_{i=1}^{m} \sum_{j=1}^{n_k} \left(1 - \delta_{ijl}\right)\left(x_{ij} - c_{il}\right)^2 , \tag{4.5}$$

kde n je počet objektů, m počet znaků, k počet shluků, n_k je počet objektů v k-tém shluku, x_{ij} jsou data, c_{il} je střední hodnota i-tého znaku v patřičném shluku a δ_{ijl} je chybějící hodnota.

Metoda je určena pro metrická data. Konverguje k řešení po konečném počtu kroků a řešení může být více, záleží na počátečních podmínkách.

4.3 Interpretace výsledků

Problémem je určení počtu shluků. Neexistuje žádné kritérium či statistický test, jejž by spolehlivě odhalily konečný počet shluků. Zavádí se např. vyšetření měr podobnosti mezi shluky. Ale kolik shluků utvořit, to závisí na řešiteli, který by měl uznat za vhodné, co je pro daný problém dostatečným kritériem na základě teoretických znalostí. Spornou otázkou jsou shluky o jednom či malém počtu objektů vzhledem k ostatním shlukům. Tyto objekty se neprojevily vybočujícími ani při předběžné analýze. Jak už bylo uvedeno výše, při odstraňování objektů je třeba mít na paměti, že i tento objekt může být reprezentativním vzorkem. Po odstranění objektů je třeba provést analýzu znovu.

Konečným výsledkem analýzy je pojmenování shluků na základě jejich vlastností a srovnání vlastností shluků z analýzy a předpověďmi z teorie. Pro zjištění správnosti shluků se provádí rozdělení dat a analýza se dělá odděleně. Poté zbývá jen určit shodu s předpokládanými výsledky z celého souboru dat.

5 PŘÍKLAD

Tab. 5-1 Tabulka příkladu

Index	x_1	x_2	x_3	x_1^2	x_2^2	x_3^2	$x_1.x_2$	$x_1.x_3$	$x_2.x_3$	$(x_1-\mu)^2$	$(x_2-\mu)^2$	$(x_3-\mu)^2$
1	5	2	3	25	4	9	10	15	6	1,96	0,36	1
2	20	7	5	400	49	25	140	100	35	184,96	19,36	1
3	3	2	2	9	4	4	6	6	4	11,56	0,36	4
4	8	6	2	64	36	4	48	16	12	2,56	11,56	4
5	12	1	1	144	1	1	12	12	1	31,36	2,56	9
6	1	1	2	1	1	4	1	2	2	29,16	2,56	4
7	2	2	2	4	4	4	4	4	4	19,36	0,36	4
8	4	2	12	16	4	144	8	48	24	5,76	0,36	64
9	7	3	8	49	9	64	21	56	24	0,36	0,16	16
10	2	0	3	4	0	9	0	6	0	19,36	6,76	1
suma	64	26	40	716	112	268	250	265	112	306,4	44,4	108

Data v tabulce Tab. 5-1 jsou v kardinální škále, tj. škála, v níž je zavedena vzdálenost a lze provádět základní matematické operace, a nestrukturovaná (neobsahují znak, který je závislý na ostatních proměnných). Jelikož jsou data řádově shodná, není nutno provádět standardizaci.

5.1 Popisné charakteristiky vícerozměrných náhodných veličin a exploratorní analýza dat

Předpokládá se normální rozdělení hodnot pro jednotlivé znaky.
Výpočet průměrné hodnoty proměnných x_1 a x_2.

$$\mu_1 = \frac{1}{n}\sum_{i=1}^{n} x_{i1} = \frac{1}{10}64 = 6{,}4$$

$$\mu_2 = \frac{1}{n}\sum_{i=1}^{n} x_{i2} = \frac{1}{10}26 = 2{,}6$$

Výpočet rozptylu pro proměnnou x_1.

$$s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{1i} - \mu_1\right)^2 = \frac{1}{10-1}306{,}4 = 34{,}0\overline{4}$$

Tab. 5-2 Základní popisné statistiky

znak	průměr	medián	minimum	maximum	rozptyl	sm. odch.
x_1	6,4	4,5	1	20	34,0444	5,8348
x_2	2,6	2,0	0	7	4,9333	2,2211
x_3	4,0	2,5	1	12	12,0000	3,4641

Výpočet korelační matice, korelace mezi proměnnými x_1 a x_2.

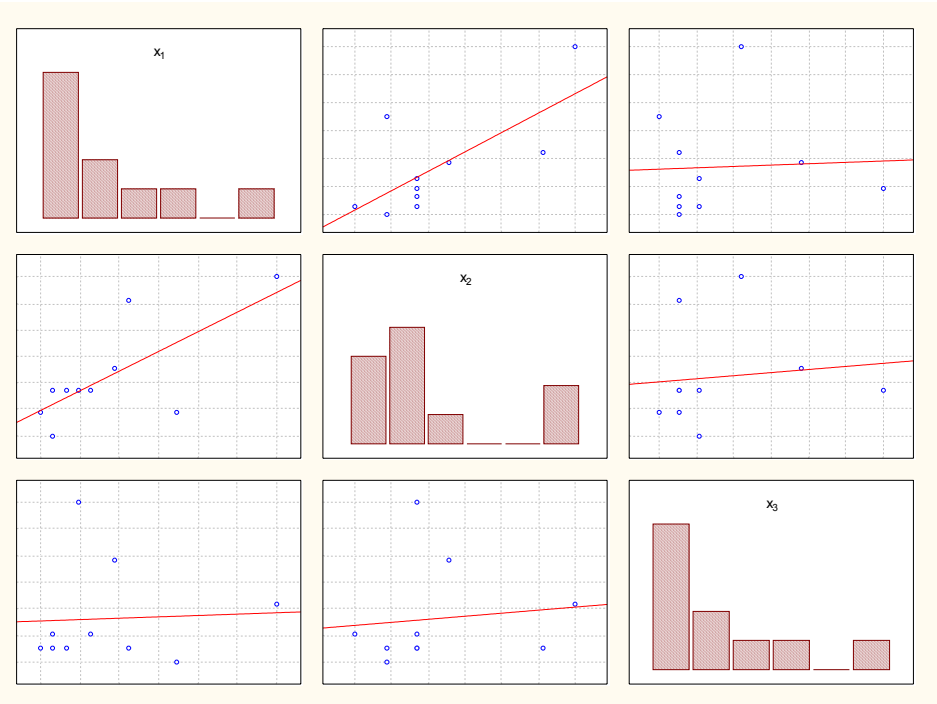
$$r_{x1x2} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_{i1}x_{i2} - \frac{1}{n}\sum_{i=1}^{n} x_{i1}\cdot\frac{1}{n}\sum_{i=1}^{n} x_{i2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{i1}-\mu_1\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{i2}-\mu_2\right)^2}} = \frac{\frac{1}{10}\sum_{i=1}^{10} x_{i1}x_{i2} - \frac{1}{10}\sum_{i=1}^{10} x_{i1}\cdot\frac{1}{10}\sum_{i=1}^{10} x_{i2}}{\sqrt{\frac{1}{10}\sum_{i=1}^{10}\left(x_{i1}-6{,}4\right)^2}\sqrt{\frac{1}{10}\sum_{i=1}^{10}\left(x_{i2}-2{,}6\right)^2}} =$$

$$= \frac{\frac{1}{10}250 - \frac{1}{10}64\cdot\frac{1}{10}26}{\sqrt{\frac{1}{10}306{,}4}\sqrt{\frac{1}{10}44{,}4}} \doteq 0{,}7168$$

Tab. 5-3 Korelační matice

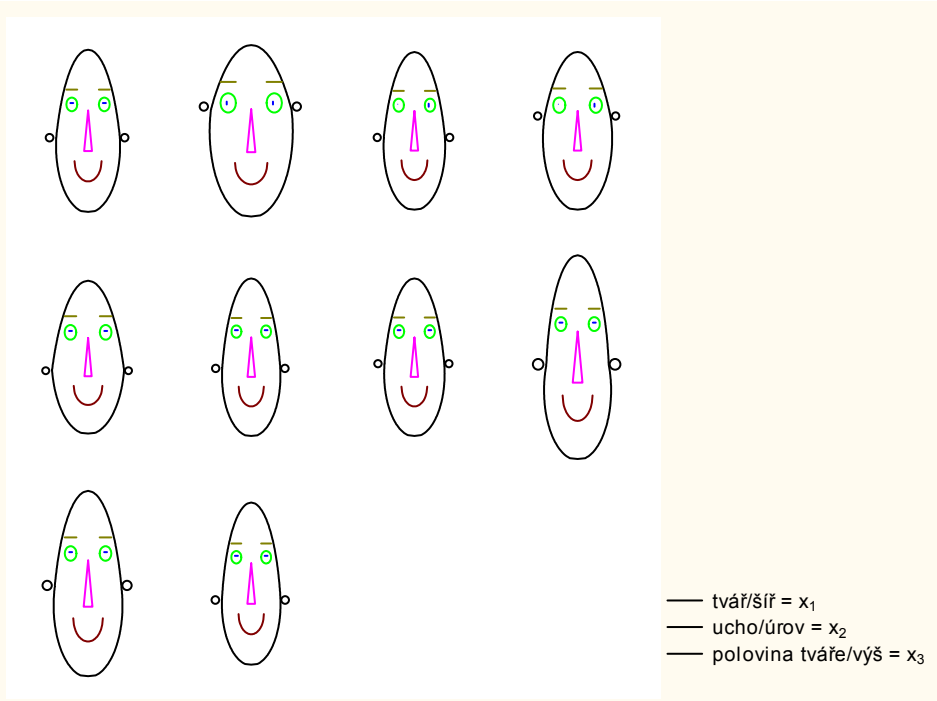
znak	x_1	x_2	x_3
x_1	1	0,7168	0,0495
x_2	0,7168	1	0,1155
x_3	0,0495	0,1155	1

Tab. 5-3 ukazuje na dobrou lineární závislost znaků x_1 a x_2, zatímco znaky x_1 a x_3 nevykazují lineární závislost. Pro malou hodnotu korelace není třeba provádět PCA. Nějaká korelace mezi jednotlivými sloupci se nalezne vždy. Pro odhalení odlehlých bodů může posloužit maticový graf.



Obr. 5-1 Maticový graf pro Tab. 5-1

Maticový graf vyjadřuje prakticky předešlou tabulku. Červená čára má vyjadřovat lineární proložení experimentálních dat. Z histogramů na diagonále je vidět, že rozdělení není podobné binomickému, ani normálnímu a ani jinému rozdělení. Dalším prostředkem, z něhož se dají zjistit vybočující body, jsou např. Chernoffovy tváře.



Obr. 5-2 Ikonový graf - Chernoffovy tváře

Z obrázku není patrno, že by nějaké objekty byly vybočující a nereprezentativní u předloženého vzorku dat. Snad se nejvíce liší jen tvář na osmé pozici.

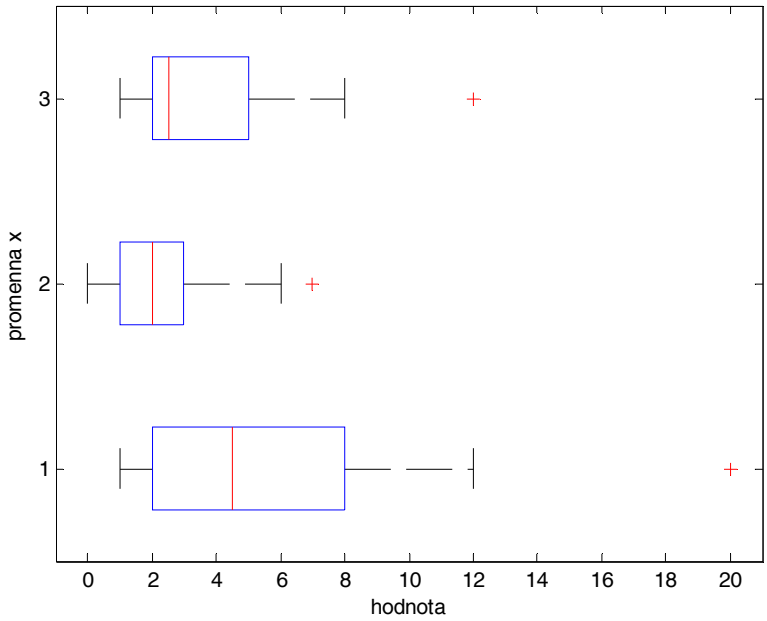
Tab. 5-4 Kovarianční matice (n)

znak	x_1	x_2	x_3
x_1	30,64	8,36	0,90
x_2	8,36	4,44	0,80
x_3	0,90	0,80	10,80

Tab. 5-5 Korigovaná kovarianční matice (n-1)

znak	x_1	x_2	x_3
x_1	34,0444	9,2889	1,0000
x_2	9,2889	4,9333	0,8889
x_3	1,0000	0,8889	12,0000

Krabicový graf pro Tab. 5-1 je z mediánu, z 25% a 75% kvartilu s rozsahem neodlehlých bodů. Krabicový graf znázorňuje proměnlivost v datech, čím delší „krabice“, tím větší proměnlivost.



Obr. 5-3 Krabicový graf s mediánem

5.2 Shluková analýza

Slouží k určení struktury v objektech.

<u>Hierarchické shlukování</u>

Provedeme Z-score (autoškálování) původní matice dat.

$$s_1 = \sqrt{\frac{\sum_{i=1}^{n}(x_{i1} - \mu_1)^2}{n-1}} = \sqrt{\frac{306{,}4}{10-1}} \doteq 5{,}8348$$

$$y_{11} = \frac{x_{11} - \mu_1}{s_1} = \frac{5 - 6{,}4}{5{,}8348} \doteq -0{,}2399$$

Tab. 5-6 Tabulka příkladu po provedení Z-score

Index	x_1	x_2	x_3
1	-0,2399	-0,2701	-0,2887
2	2,3309	1,9810	0,2887
3	-0,5827	-0,2701	-0,5774
4	0,2742	1,5308	-0,5774
5	0,9598	-0,7204	-0,8660
6	-0,9255	-0,7204	-0,5774
7	-0,7541	-0,2701	-0,5774
8	-0,4113	-0,2701	2,3094
9	0,1028	0,1801	1,1547
10	-0,7541	-1,1706	-0,2887

Výpočet vzdáleností (eukleidovské) pro dendrogram.

$$d_{12} = \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + (x_{13} - x_{23})^2}$$
$$d_{12} = \sqrt{(-0{,}239941 - 2{,}330858)^2 + (-0{,}270135 - 1{,}980991)^2 + (-0{,}288675 - 0{,}288675)^2}$$
$$d_{12} \doteq 3{,}4655$$

Tab. 5-7 Eukleidovské vzdálenosti

	1	2	3	4	5	6	7	8	9	10
1	0,00	3,47	0,45	1,89	1,41	0,87	0,59	2,60	1,55	1,04
2	3,47	0,00	3,78	2,28	3,24	4,32	3,92	4,08	2,99	4,45
3	0,45	3,78	0,00	1,99	1,63	0,57	0,17	2,89	1,92	0,96
4	1,89	2,28	1,99	0,00	2,37	2,55	2,07	3,47	2,20	2,90
5	1,41	3,24	1,63	2,37	0,00	1,91	1,80	3,49	2,37	1,86
6	0,87	4,32	0,57	2,55	1,91	0,00	0,48	2,97	2,21	0,56
7	0,59	3,92	0,17	2,07	1,80	0,48	0,00	2,91	1,98	0,95
8	2,60	4,08	2,89	3,47	3,49	2,97	2,91	0,00	1,34	2,77
9	1,55	2,99	1,92	2,20	2,37	2,21	1,98	1,34	0,00	2,15
10	1,04	4,45	0,96	2,90	1,86	0,56	0,95	2,77	2,15	0,00

Vybereme nejmenší vzdálenost mezi dvěma objekty, to vytvoří shluk 3-7 (jejich vzdálenost je 0,17), přepočítáním pomocí hierarchického shlukování průměrovou metodou dostaneme tabulku Tab. 5-8.

$$d_{(3,7)1} = 0{,}5(d_{1,3} + d_{1,7}) = 0{,}5(0{,}45 + 0{,}59) = 0{,}52$$
$$d_{(3,7)2} = 0{,}5(d_{2,3} + d_{2,7}) = 0{,}5(3{,}78 + 3{,}92) = 3{,}85$$
$$d_{(3,7)4} = 0{,}5(d_{4,3} + d_{4,7}) = 0{,}5(1{,}99 + 2{,}07) = 2{,}03$$
$$\vdots$$

Tab. 5-8 Tab. 5-7 po přepočtu

	1	2	3,7	4	5	6	8	9	10
1	0,00	3,85	0,52	2,03	1,72	0,53	2,90	1,95	0,96
2	3,85	0,00	3,78	2,28	3,24	4,32	4,08	2,99	4,45
3,7	0,52	3,78	0,00	1,99	1,63	0,57	2,89	1,92	0,96
4	2,03	2,28	1,99	0,00	2,37	2,55	3,47	2,20	2,90
5	1,72	3,24	1,63	2,37	0,00	1,91	3,49	2,37	1,86
6	0,53	4,32	0,57	2,55	1,91	0,00	2,97	2,21	0,56
8	2,90	4,08	2,89	3,47	3,49	2,97	0,00	1,34	2,77
9	1,95	2,99	1,92	2,20	2,37	2,21	1,34	0,00	2,15
10	0,96	4,45	0,96	2,90	1,86	0,56	2,77	2,15	0,00

Nejmenší vzdálenost je nyní mezi objekty (3,7) a 1 (jejich vzdálenost je 0,52), ty vytvoří další shluk.

Tab. 5-9 Tab. 5-8 po přepočtu

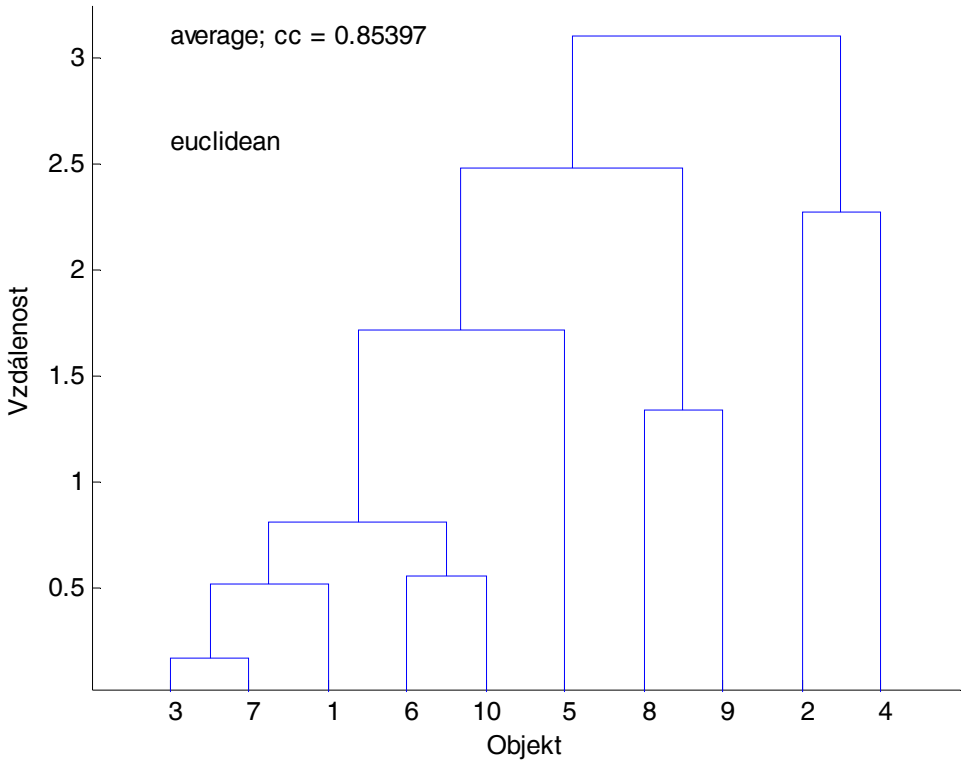
	2	3,7,1	4	5	6	8	9	10
2	0,00	3,82	2,01	1,68	0,55	2,90	1,94	0,96
3,7,1	3,82	0,00	1,99	1,63	0,57	2,89	1,92	0,96
4	2,01	1,99	0,00	2,37	2,55	3,47	2,20	2,90
5	1,68	1,63	2,37	0,00	1,91	3,49	2,37	1,86
6	0,55	0,57	2,55	1,91	0,00	2,97	2,21	0,56
8	2,90	2,89	3,47	3,49	2,97	0,00	1,34	2,77
9	1,94	1,92	2,20	2,37	2,21	1,34	0,00	2,15
10	0,96	0,96	2,90	1,86	0,56	2,77	2,15	0,00

Další přepočítávání je zřejmé. Rozvrh slučování je v následující tabulce.

Tab. 5-10 Rozvrh slučování

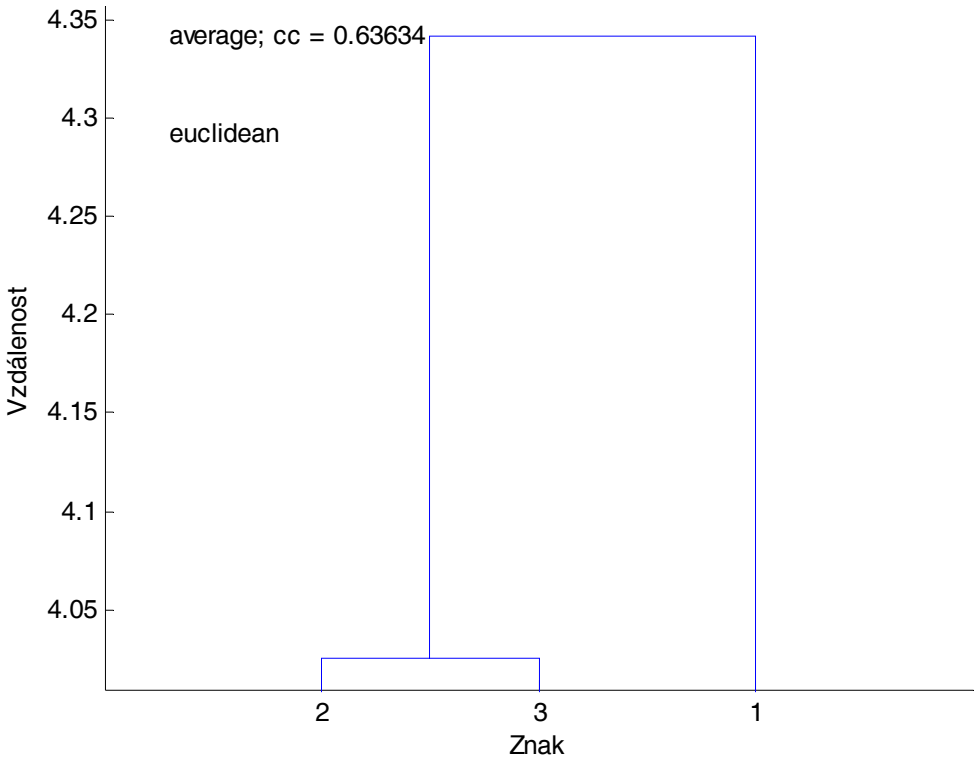
vzdálenost	1	2	3	4	5	6	7	8	9	10
0,1713866	P_3	P_7								
0,5188963	P_1	P_3	P_7							
0,5616132	P_6	P_10								
0,8100975	P_1	P_3	P_7	P_6	P_10					
1,341789	P_8	P_9								
1,720867	P_1	P_3	P_7	P_6	P_10	P_5				
2,276503	P_2	P_4								
2,484368	P_1	P_3	P_7	P_6	P_10	P_5	P_8	P_9		
3,106986	P_1	P_3	P_7	P_6	P_10	P_5	P_8	P_9	P_2	P_4

Pro předchozí výpočty a metodu shlukování byl vytvořen následující dendrogram (Obr. 5-4). Metoda průměrová (average), Eukleidovská metrika a původní data po Z-score.

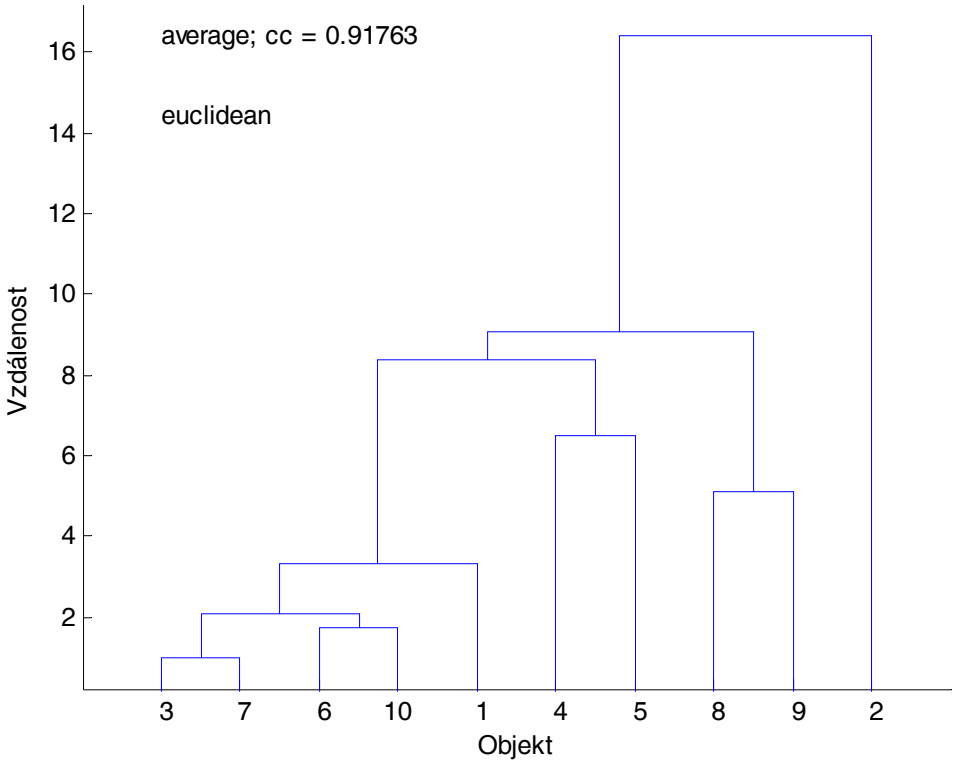


Obr. 5-4 Dendrogram s provedením Z-score

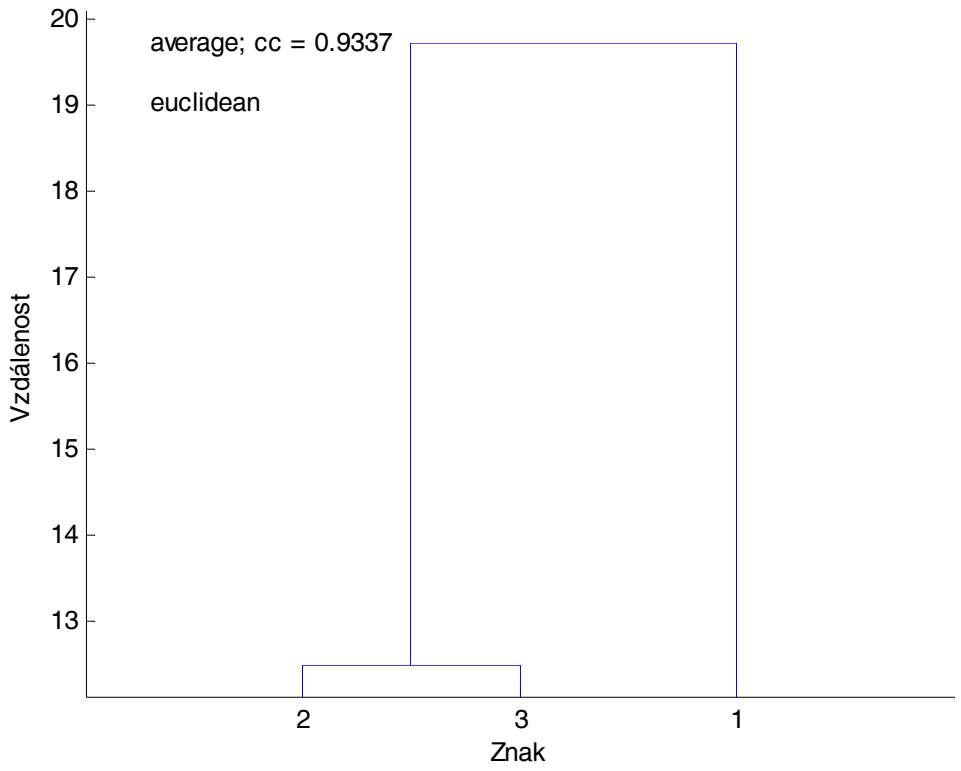
Dendrogram proměnných byl vytvořen v Matlabu 7 tak, že matice dat pro analýzu byla transponována.



Obr. 5-5 Dendrogram proměnných se Z-score

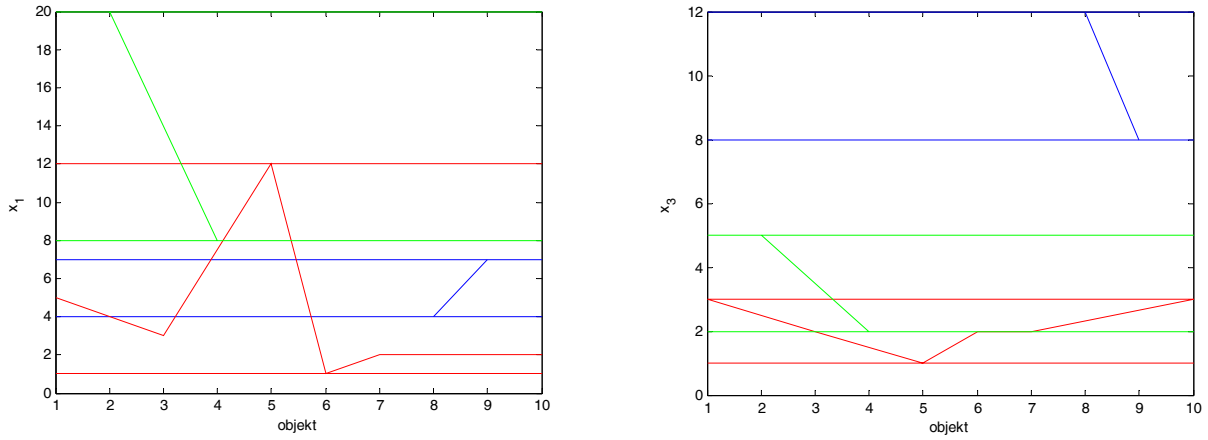


Obr. 5-6 Dendrogram bez provedení Z-score



Obr. 5-7 Dendrogram proměnných bez provedení Z-score

Horizontálním řezem dendrogramem na Obr. 5-4 v úrovni vzdálenosti spojení asi 2,4 nalezneme tři shluky, první obsahuje objekty č. 2 a 4, druhý má objekty 8 a 9, třetímu náleží prvky 1, 3, 5, 6, 7 a 10. Zde je vidět rozdíl při škálování dat. Ale na druhou stranu neškálovaná data mají lepší proložení modelu, jak je možné poznat z kofenetického korelačního koeficientu *cc*. Dendrogram proměnných odhaluje podobnost znaků 2 a 3. Vzniklé shluky musíme dále dát do souvislostí s tabulkou dat.



Obr. 5-8 Graf vyjádření mezí rozdělených objektů do jednotlivých skupin podle dendrogramu s provedením Z-score v závislosti na vybraném znaku

Z grafu můžeme usuzovat, že je-li více znaků ve zkoumaném souboru dat, nemůžeme rozhodnout na rozdělení objektů podle jednoho ze znaků. I pro různé rozdělení se grafy pro jednotlivé skupiny značně překrývají.

Nehierarchické shlukování

Použití metody shlukování K-means, byly podle programu Matlab 7, při zadání počtu shluků rovno 3, zvolené metrice eukleidovská vzdálenost a zvolené skupině inicializující polohu středu shluku vykonávaném na předběžném seskupování na náhodném 10% vzorku z matice dat (cluster), objekty rozděleny po dvou iteracích takto: 8 a 9; 2 a 5; 1, 3, 4, 6, 7 a 10.

5.3 Analýza hlavních komponent

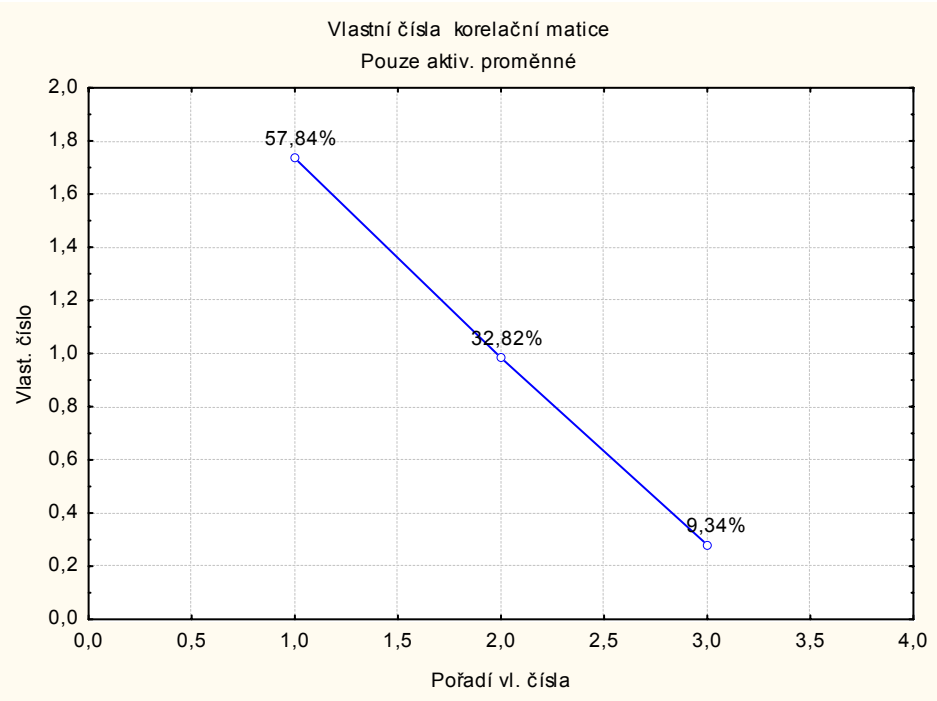
Slouží k nalezení struktury ve znacích. Analýza byla provedena na základě korelační matice pro směrodatnou odchylku dle vzorce (2.6) a s daty po autoškálování. Z korelační matice zjistíme vlastní čísla a procento celkového rozptylu. Součet vlastních čísel se u korelační matice rovná počtu znaku *m* neboli stopě této matice. Stejná procenta se získají i z kovarianční matice a to stejným postupem přes stopu matice a jednotlivá vlastní čísla.

$$\%\,\text{celkového rozptylu} = \frac{\lambda_i}{tr\boldsymbol{R}} \cdot 100 \tag{5.1}$$

$$\%\,\text{celkového rozptylu} = \frac{\lambda_1}{tr\boldsymbol{R}} \cdot 100 = \frac{1{,}7353057}{3} \cdot 100 \doteq 57{,}844\%$$

Tab. 5-11 Pro Cattelův indexový graf úpatí vlastních čísel, vlastní čísla z korelační matice

index	vlastní číslo	% celk. rozptylu	kumulativní vlastní číslo	kumulativní procento
1	1,7353057	57,844	1,7353057	57,844
2	0,9845197	32,817	2,7198255	90,661
3	0,2801745	9,339	3,0000000	100,000



Obr. 5-9 Cattelův indexový graf úpatí vlastních čísel

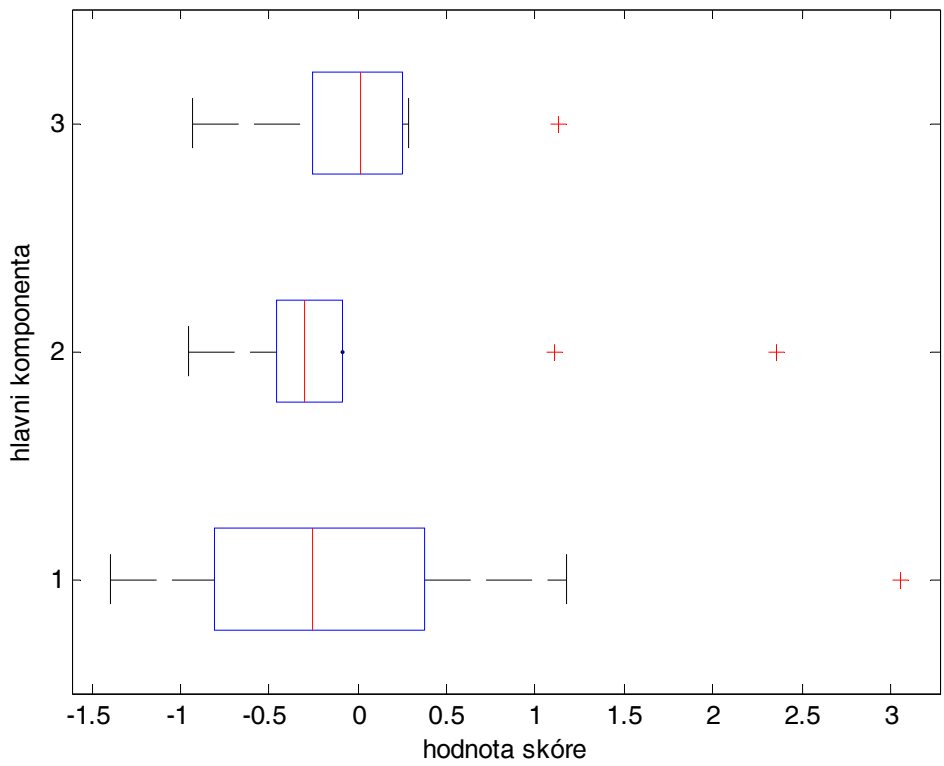
Obr. 5-9 určí kolika procenty je popsána proměnlivost v datech při použití prvních dvou komponent, je to 90,66 %. Hladina, při které volíme počet hlavních komponent, by neměla být nižší než 80 %.

Z korigované korelační matice (n-1) je získáno komponentní skóre.

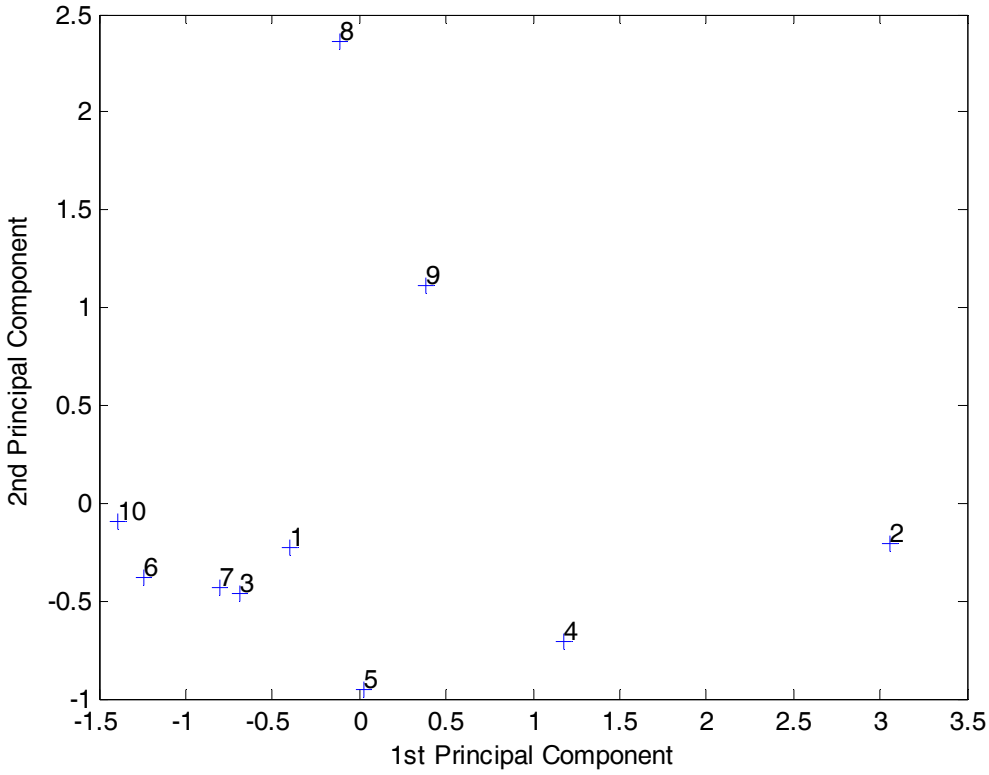
*Tab. 5-12 Komponentní skóre, matice **T***

	x_1	x_2	x_3
1	-0,4016	-0,2292	0,0043
2	3,0552	-0,2105	0,2496
3	-0,6851	-0,4597	-0,2552
4	1,1743	-0,7110	-0,9313
5	0,0252	-0,9579	1,1277
6	-1,2393	-0,3767	-0,1764
7	-0,8042	-0,4328	-0,3755
8	-0,1127	2,3579	0,0545
9	0,3792	1,1100	0,0202
10	-1,3909	-0,0901	0,2822

Z těchto hodnot sestrojíme graf komponentních vah pro různé kombinace proměnných, ale pro popis daných dat stačí užít jen dvě, jak je uvedeno výše. Na osu x se vynáší ta hlavní komponenta, která vykazuje největší hodnotu rozptýlení, ve většině případů se jedná o první hlavní komponentu.

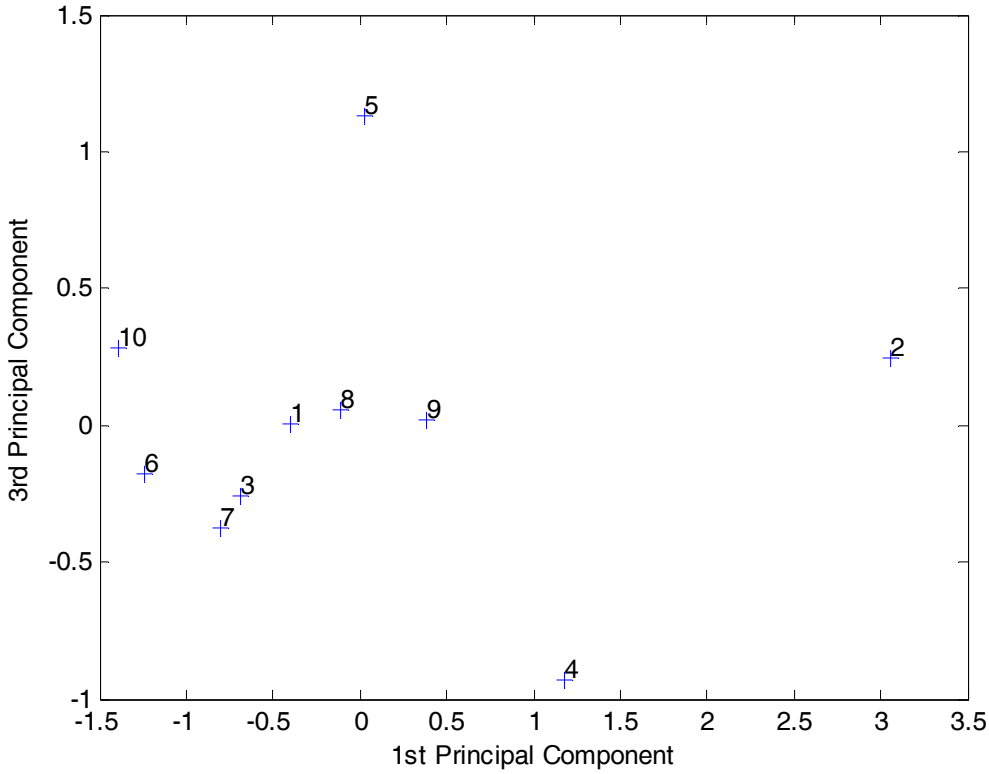


Obr. 5-10 Krabicový graf pro Tab. 5-12



Obr. 5-11 Rozptylový diagram komponentního skóre

Graf komponentních vah můžeme porovnat s dendrogramem na Obr. 5-4, kde se objevují stejné shluky objektů. Dá se říct, že objekty jsou rozloženy v rovině grafu relativně rovnoměrně.



Obr. 5-12 Rozptylový diagram komponentního skóre pro komponenty 1 a 3

Obr. 5-12 vypovídá, že použitím hlavních komponent jedna, s proměnlivostí v datech určených z vlastních čísel je 57,84 %, a tři, s proměnlivostí 9,34 %, nesouhlasí v plném rozsahu modelu vypočteném v CLU. Už jen to, že podle dendrogramu na Obr. 5-6 by měly být objekty 8 a 9 dále od sebe a objekty 6 a 10 blíže sobě, čemuž tak v Obr. 5-12 není. Ale je tomu tak při uvažování průmětu vzdálenosti mezi těmito objekty do první hlavní komponenty. Ano, průmět vzdáleností do první hlavní komponenty je pro oba grafy stejný.

*Tab. 5-13 Vlastní vektory hlavních komponent, matice **P***

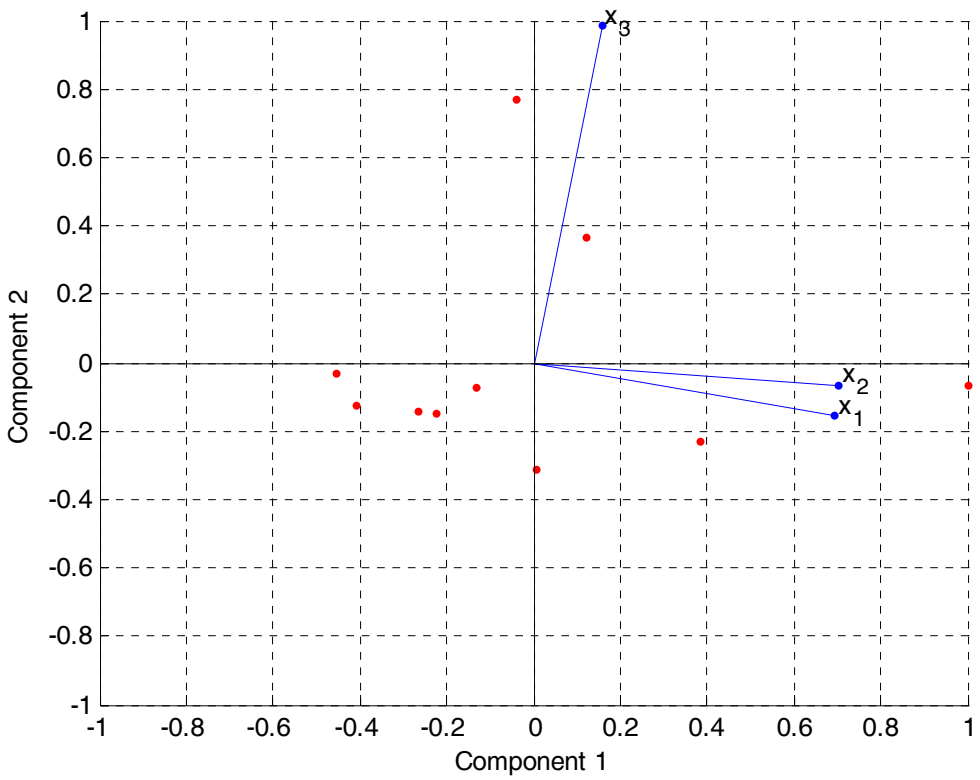
	komp. 1	komp. 2	komp. 3
x_1	0,69475	-0,15744	0,70181
x_2	0,7019	-0,064619	-0,70934
x_3	0,15703	0,98541	0,065609

Vlastní vektory dávají přehled o relativním zastoupení původních znaků v hlavních komponentách. Sloupce v ***P*** jsou ortogonální.

Tab. 5-14 Komponentní váhy jednotlivých znaků, korelace mezi znaky a komponentami

	komp. 1	komp. 2	komp. 3
x_1	0,9152	-0,1562	0,3715
x_2	0,9246	-0,0641	-0,3755
x_3	0,2069	0,9778	0,0347

Z těchto hodnot je sestrojen dvojný graf pro proměnné a projekce proměnných do roviny komponent.



Obr. 5-13 Dvojný graf (biplot)

Rozbor dvojného grafu nám ukazuje na kovarianci mezi znaky x_1 a x_2, které přispívají nejvíce do první komponenty. Stejně tak i Tab. 5-14 odhaluje silnou korelovatelnost mezi první komponentou a znaky x_1 a x_2, jejichž úhel mezi úsečkami, spojující počátek s bodem reprezentující znak, je malý. Naopak velký úhel ukazuje na malou korelaci. Slabá korelace všech znaků je s třetí komponentou. Objekty charakterizované znaky x_1 a x_2 jsou ve spodní části grafu. Objekty 8 a 9 jsou určeny znakem x_3, obsahují podle Tab. 5-1 největší hodnoty z x_3 oproti ostatním objektům.

Výpočet reziduí a určení těsnosti proložení modelu hlavních komponent se provede podle vztahu (3.7).

$$\boldsymbol{X} = \boldsymbol{t}_1 \boldsymbol{p}_1^T + \boldsymbol{t}_2 \boldsymbol{p}_2^T + \boldsymbol{t}_3 \boldsymbol{p}_3^T + \boldsymbol{E}$$

Matice $\boldsymbol{X}$ je zde Tab. 5-1 po provedení Z-score. Reziduum pro analýzu prvních dvou komponent je v Tab. 5-15 $\left(\boldsymbol{E} = zscore(\boldsymbol{X}) - \boldsymbol{t}_1 \boldsymbol{p}_1^T - \boldsymbol{t}_2 \boldsymbol{p}_2^T\right)$ a pro zvolené hlavní komponenty 1 a 3 je v Tab. 5-16 $\left(\boldsymbol{E} = zscore(\boldsymbol{X}) - \boldsymbol{t}_1 \boldsymbol{p}_1^T - \boldsymbol{t}_3 \boldsymbol{p}_3^T\right)$. Sice to neodpovídá teoriím uvedeným v seznamu použité literatury, ale jak bude níže uvedeno, je zde porovnáno proložení modelu v závislosti na použitých hlavních komponentách.

*Tab. 5-15 Reziduum pro hlavní komponenty 1 a 2, matice **E***

objekt	x_1	x_2	x_3
1	0,00301	-0,0030	0,00028
2	0,17514	-0,1770	0,01637
3	-0,17911	0,1810	-0,01674
4	-0,65357	0,6606	-0,06110
5	0,79145	-0,8000	0,07399
6	-0,12381	0,1251	-0,01157
7	-0,26352	0,2664	-0,02464
8	0,03822	-0,0386	0,00357
9	0,01416	-0,0143	0,00132
10	0,19803	-0,2002	0,01851

*Tab. 5-16 Reziduum pro hlavní komponenty 1 a 3, matice **E***

objekt	x_1	x_2	x_3
1	0,03609	0,01481	-0,22589
2	0,03314	0,01360	-0,20743
3	0,07238	0,02971	-0,45303
4	0,11194	0,04595	-0,70065
5	0,15081	0,06190	-0,94397
6	0,05930	0,02434	-0,37118
7	0,06813	0,02796	-0,42644
8	-0,37122	-0,15237	2,32350
9	-0,17476	-0,07173	1,09380
10	0,01419	0,00582	-0,08879

Z předchozích dvou tabulek je zřejmé, že hodnoty reziduí jsou jak malá, tak velká. V potaz musíme vzít i tabulku dat po Z-score. Z průměru každého znaku se určí počátek dat společný všem komponentám. Centrováním zdrojové matice dat $\boldsymbol{X}$ dostaneme matici reziduí $\boldsymbol{E}_0$, která je důležitá při posouzení relativní velikosti $\boldsymbol{E}$ a představuje, že reziduový rozptyl bude 100 % a modelem objasněný rozptyl 0 %. Můžeme obdržet rezidua každého objektu při sčítání řádků nebo rezidua znaků při sčítání sloupců. Rozptyl reziduí objektu je vzdálenost mezi ním a modelem. Z reziduí lze odhalit i vybočující body. Vztah pro čtverec reziduí objektu

$$e_i^2 = \sum_{k=1}^{m} e_{ik}^2 \ . \tag{5.2}$$

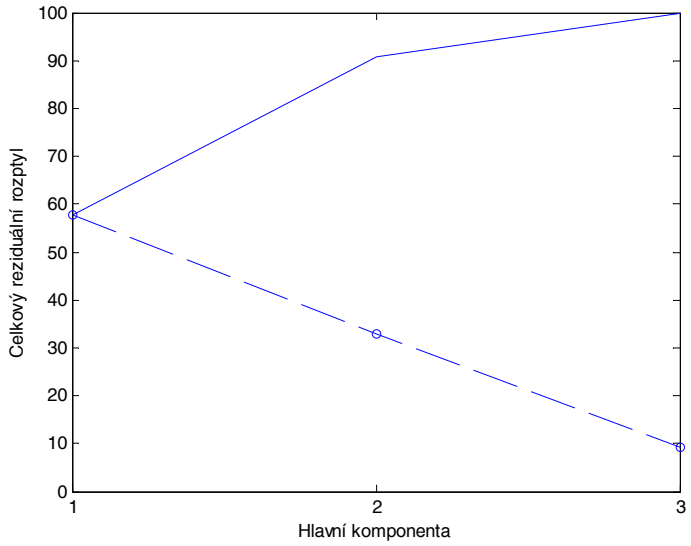
Celkový čtverec reziduí všech objektů

$$e_{tot}^2 = \sum_{i=1}^{n} e_i^2 \ . \tag{5.3}$$

Celkový rozptyl reziduí je dán

$$\frac{e_{tot}^2}{pn}, \tag{5.4}$$

kde p je počet hlavních komponent. Porovnáním tabulek Tab. 5-15 a Tab. 5-16 vychází lepší proložení modelu podle předchozích vztahů pro první tabulku s hlavními komponentami 1 a 2.



Obr. 5-14 Indexový graf úpatí celkového reziduálního rozptylu (modrá čára)

Indexový graf úpatí celkového reziduálního rozptylu vznikne z postupného vyčíslování matic reziduí, jak se přidávají další komponenty a výpočtem e_{tot}^2. Sloupce vyjadřují totéž, co Cattelův indexový graf úpatí vlastních čísel a také určuje optimální počet použitých hlavních komponent.

6 APLIKACE PCA

Zde je uveden pokus o zobecnění vlivu znaků na jednotlivé vývody. Rozptylové diagramy komponentního skóre jsou vyobrazeny v barvách, které odpovídají jednotlivému třízení podle hierarchického aglomeračního shlukování a K-means uvedené v další kapitole. Tzv. vlivné body nebyly odstraněny, neboť potom je model horší. Na začátek stojí za to uvést, jak vlastně použité programy zpracovávají PCA. Jak bylo uvedeno v předchozích kapitolách, je nutné data před provedením PCA standardizovat minimálně sloupcovým centrováním. Matlab 7 provádí sloupcové centrování. Jsou-li data už centrována nebo autoškálována, provedením sloupcového centrování se hodnoty nezmění. Matlab vychází z korelační matice, kdežto ve Statistice máme na výběr analýzu provádět z korelační nebo kovarianční matice a to buď s n-1 nebo jen n výpočtem směrodatné odchylky. Máme-li už data autoškálována, kovarianční matice je rovna korelační. Program Statistica 7 CZ provádí autoškálování. Z toho plyne, že jsou-li data už sloupcově centrována, analyzovaná matice dat změní svoje hodnoty na autoškálované původní nestandardizované matice dat.

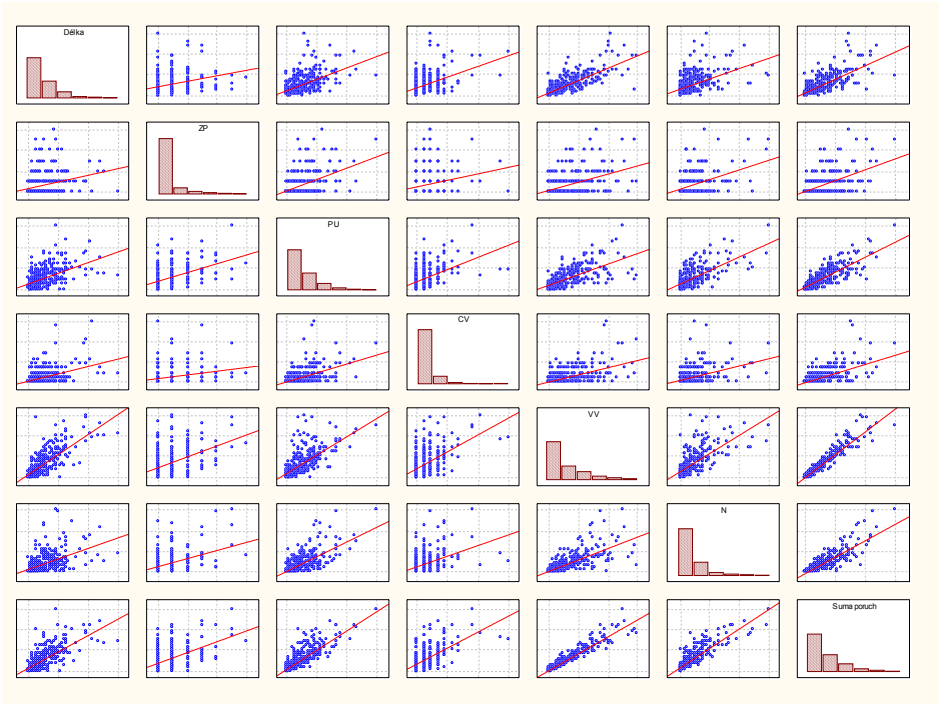
Ještě před vlastní analýzou rozebereme předložená data. Aplikace metod vícerozměrné analýzy dat na data charakterizovanými znaky: Délka, poruchy Související se započetím provozu (ZP), poruchy vzniklé provozem a údržbou - Provoz a údržba (PU), poruchy jež způsobily Cizí vlivy (CV) a Vnější vlivy (VV), poruchy nezjištěné Nezjištěno (N) a součet poruch za 11 let Suma poruch. Znaky byly tak zvoleny, protože chceme najít závislosti a vlastnosti vývodů na počtu poruch a analyzovaná matice dat musí tato data obsahovat. Matice má 321 řádků, tj. objektů. Byly vybrány jen ty vývody, u kterých nechyběla data.

Tab. 6-1 Kódy poruch, EGÚ Brno

kód nový	Související se započetím provozu (ZP)	Provoz a údržba (PU)	Cizí vlivy (CV)	Vnější vlivy (VV)	Nezjištěno (N)
kódy poruch	26 - vada na projektu 28 - vadná montáž 29 - vadný pracovní postup 31 - výrobní vada	3 - koroze 4 - koroze uložením v zemi 8 - netěsnost 10 - opotřebení materiálu 11 - ostatní provozní vlivy 14 - porucha DT 16 - úroveň NN (VVN) 17 - porucha u zákazníka 18 - jiný RD 19 - ovládací a sign. obvody 20 - řídící systémy 21 - přechodový odpor 22 - přetížení 23 - spínací přepětí 24 - stárnutí izolace 25 - uvnitř PTE 27 - vadná manipulace 32 - zatečení vody 33 - znečištění 34 - zvýšená vlhkost 36 - defekt bleskojistky 39 - průraz izolátoru 40 - snížení izolačního stavu 41 - stáří materiálu	2 - cizí zásah 13 - pohyb půdy, důlní vlivy 15 - mimo SME 35 - živočich	1 - bouře, atmosférické vlivy 5 - mráz pod 10°C 6 - námraza 7 - neořezaná vegetace 12 - pád větve (stromu) 30 - vítr 37 - déšť 38 - sníh 42 - teplota nad 30°C 43 - mlha	9 - nezjištěno

Tab. 6-2 Základní popisné statistiky, počet případů je 321

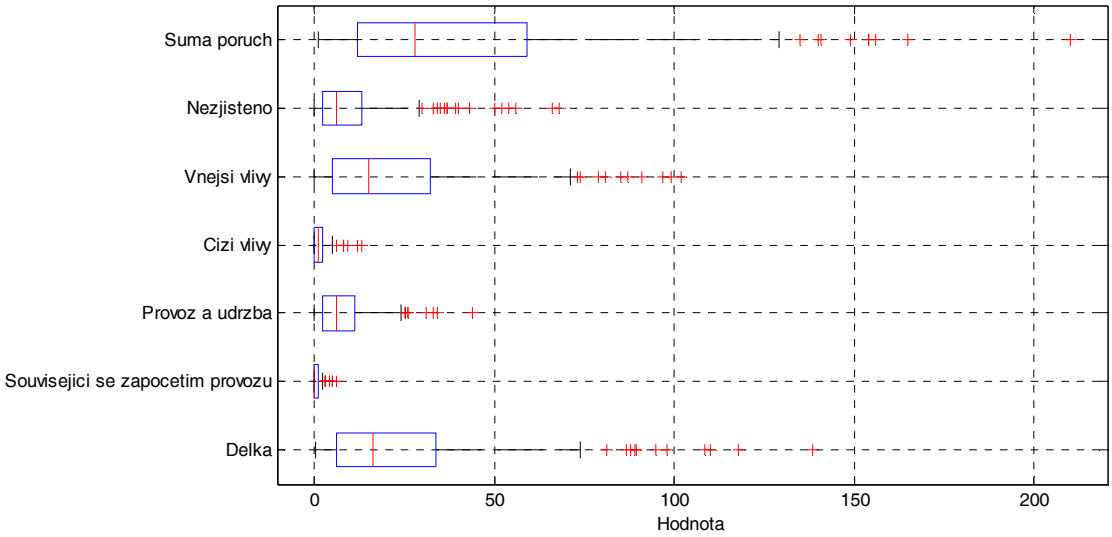
Znak	Průměr	Medián	Minimum	Maximum	Spodní kvartil	Horní kvartil	Kvantil 10%	Kvantil 90%	Směrodatná odchylka
Délka vývodu v km	23,1825	16	0,04	138,60	6,187	33,649	2,295	52,000	22,4673
Související se započetím provozu	0,6573	0	0,00	6,00	0,000	1,000	0,000	2,000	1,0670
Provoz a údržba	7,4424	6	0,00	44,00	2,000	11,000	1,000	16,000	6,8486
Cizí vlivy	1,1371	1	0,00	13,00	0,000	2,000	0,000	3,000	1,7464
Vnější vlivy	21,2773	15	0,00	102,00	5,000	32,000	1,000	50,000	21,1238
Nezjištěno	9,3925	6	0,00	68,00	2,000	13,000	0,000	22,000	10,9497
Suma poruch za 11 let	39,9065	28	1,00	210,00	12,000	59,000	3,000	93,000	36,6755



Obr. 6-1 Maticový graf

Základní představu o poskytnutých datech přinesou základní popisné charakteristiky, kde vystižnějším údajem o střední hodnotě bude medián, vzhledem k tomu, jak bylo uvedeno dříve, jsou data jednotlivých znaků blízká negativně-binomickému rozdělení. To neplatí pro znak Délka, který může nabývat všech kladných hodnot, tedy není diskrétní. Jeho rozdělení se blíží exponenciálnímu.

Úsečkou uvnitř krabice v každém zobrazovaném krabicovém grafu se bude znázorňovat medián daného znaku, krabice je tvořena 25% (spodním) až 75% (horním) kvartilem a graf bude zahrnovat i rozpětí neodlehlých bodů a odlehlé body.



Obr. 6-2 Krabicový graf pro Tab. 6-2

Mohou se analyzovat data i pomocí jiných znaků, které jsou k dispozici.

Tab. 6-3 Základní popisné statistiky, počet případů je 321

Znak	Průměr	Medián	Minimum	Maximum	Spodní kvartil	Horní kvartil	Kvantil 10%	Kvantil 90%	Směrodatná odchylka
Velkoodběratelé	8,0903	6,000	0,000	44,000	3,000	11,000	1,000	18,000	7,5308
Maloodběratelé v tis.	1,5608	1,176	0,000	9,567	0,385	2,399	0,001	3,487	1,4629
Délka v km	23,1825	16,000	0,040	138,600	6,187	33,649	2,295	52,000	22,4673
Délka kmene v km	8,1310	6,431	0,000	40,100	3,305	11,480	1,575	16,900	6,3237
Délka větví v km	15,0515	8,216	0,000	98,500	1,800	21,795	0,164	37,012	18,0193
Počet DTS	28,4299	22,000	1,000	217,000	9,000	41,000	2,000	63,000	26,4170
Zatížení v A	74,7994	68,300	0,000	256,000	28,000	115,100	0,200	151,600	56,1816
Suma poruch za 11 let	39,9065	28,000	1,000	210,000	12,000	59,000	3,000	93,000	36,6755

Tab. 6-4 Korelační matice původní matice dat

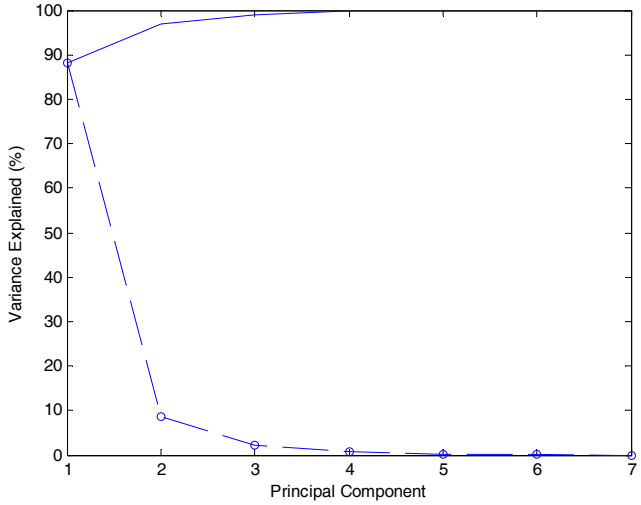
Znak	Délka	ZP	PU	CV	VV	N	Suma poruch
Délka	1	0,29231	0,52920	0,42926	0,76191	0,50472	0,71728
ZP	0,29231	1	0,48267	0,23827	0,46067	0,42989	0,52424
PU	0,52920	0,48267	1	0,50179	0,67153	0,68193	0,81504
CV	0,42926	0,23827	0,50179	1	0,53400	0,42501	0,58270
VV	0,76191	0,46067	0,67153	0,53400	1	0,70235	0,94988
N	0,50472	0,42989	0,68193	0,42501	0,70235	1	0,86317
Suma poruch	0,71728	0,52424	0,81504	0,58270	0,94988	0,86317	1

Vyšetříme počet potřebných hlavních komponent pro přijatelné vysvětlení proměnlivosti původních dat. Vychází se z vlastních čísel korelační nebo kovarianční matice.

Tab. 6-5 Vlastní čísla korelační matice a vyšetření potřebných hlavních komponent při sloupcovém centrování dat

hlavní komponenta	vlastní číslo	% celkového rozptylu	kumulativní vlastní číslo	kumulativní procento
1	4,5673	88,1021	4,5673	88,1021
2	0,8034	8,6949	5,3707	96,7970
3	0,6152	2,2006	5,9859	98,9976
4	0,5078	0,8659	6,4936	99,8635
5	0,3169	0,0979	6,8105	99,9614
6	0,1895	0,0386	7,0000	100,0000
7	0,0000	0,0000	7,0000	100,0000

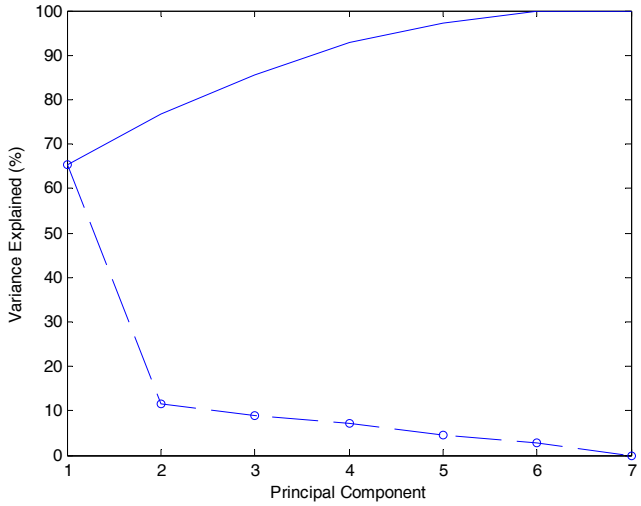
Procento celkového rozptylu je určeno z vlastních čísel kovarianční matice a nemá vliv, jestli použijeme vzorec s n-1 nebo jen s n prvky, rozhodující je stopa kovarianční matice.



Obr. 6-3 Cattelův indexový graf úpatí vlastních čísel (čárkovaná čára) a indexový graf úpatí celkového reziduálního rozptylu (plná čára) pro sloupcově standardizovaná data

Tab. 6-6 Vlastní čísla korelační matice a vyšetření potřebných hlavních komponent při autoškálování dat

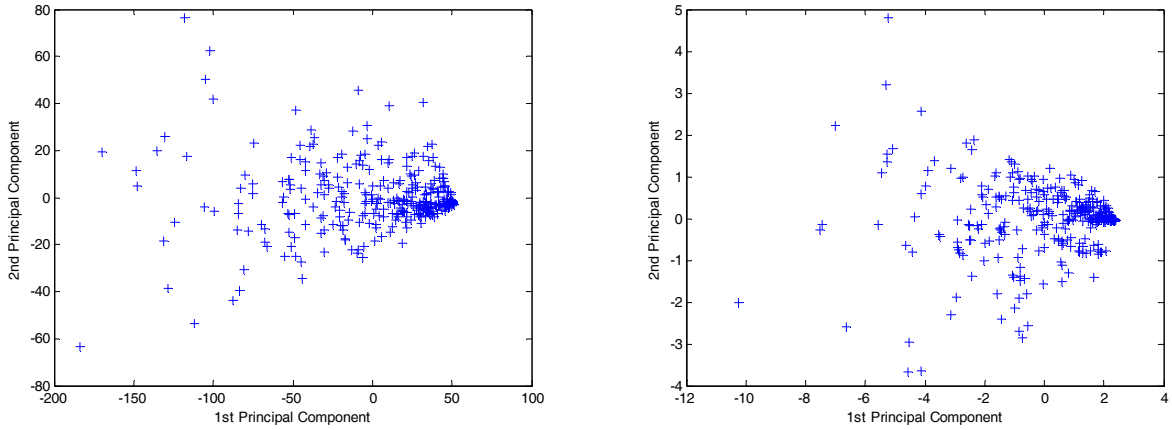
hlavní komponenta	vlastní číslo	% celkového rozptylu	kumulativní vlastní číslo	kumulativní procento
1	4,5673	65,2473	4,5673	65,2473
2	0,8034	11,4764	5,3707	76,7237
3	0,6152	8,7886	5,9859	85,5123
4	0,5078	7,2539	6,4936	92,7662
5	0,3169	4,5264	6,8105	97,2926
6	0,1895	2,7074	7,0000	100,0000
7	0,0000	0,0000	7,0000	100,0000



Obr. 6-4 Cattelův indexový graf úpatí vlastních čísel (čárkovaná čára) a indexový graf úpatí celkového reziduálního rozptylu (plná čára) pro data po autoškálování

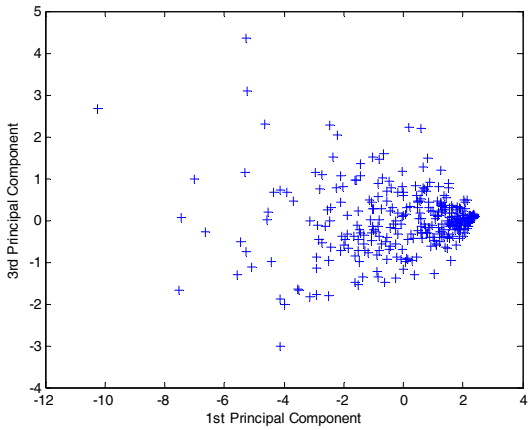
Odečtením z předchozích dvou obrázků nebo tabulek zjistíme, že nejméně hlavních komponent pro objasnění proměnlivosti v původní matici dat je při použití sloupcové standardizace, kdy už první dvě hlavní komponenty obsahují 96,797 % rozptylu, zatímco pro dosažení kýžených 90 % proměnlivosti je při užití autoškálování až u čtyř hlavních komponent. Podle dříve zmíněného je přijatelného modelu dosaženo i při 80 %. Proto by u Z-score měly být

zobrazeny i další souřadnice, aby bylo v modelu obsaženo alespoň 80 % proměnlivosti původních dat. Přidávání dalších hlavních komponent přináší i to, že do modelu je zahrnuto více šumu.

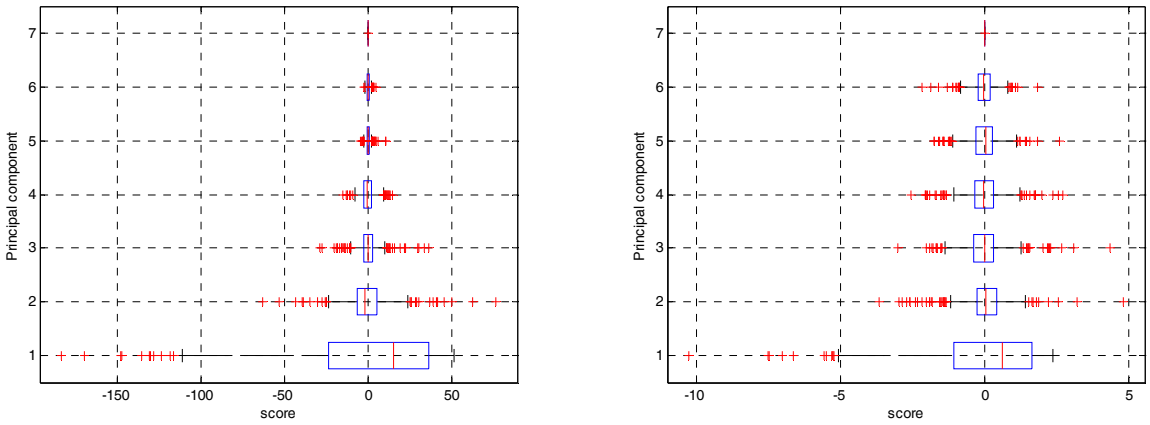


Obr. 6-5 Rozptylový diagram komponentního skóre pro data sloupcově centrovaná (vlevo) a po provedení Z-score (vpravo)

Podrobnější rozptylové diagramy komponentního skóre s příslušnými indexy vývodů jsou uvedeny v příloze D (Obr. 6-5 vlevo) a v příloze E (Obr. 6-5 vpravo).



Obr. 6-6 Rozptylový diagram komponentního skóre při použití Z-score pro 1. a 3. hlavní komponentu



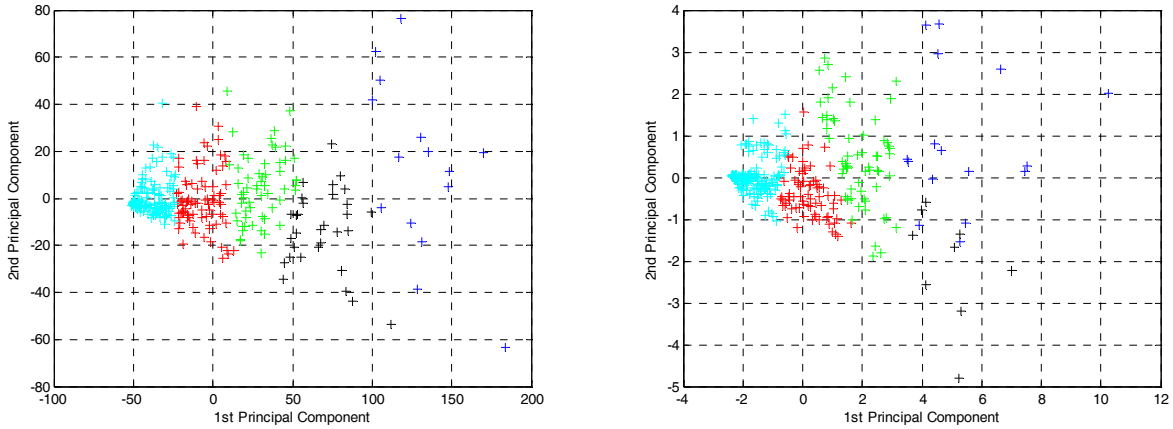
Obr. 6-7Krabicový graf hlavních komponent pro skóre s daty sloupcově centrovanými (vlevo) a po autoškálování (vpravo)

Pro přehlednost v jednotlivých komponentách pro sloupcově centrovaná a autoškálovaná data byl vytvořen Obr. 6-7, krabicový graf pro jednotlivé komponenty. Obr. 6-7 má souvislost s Obr. 6-5.

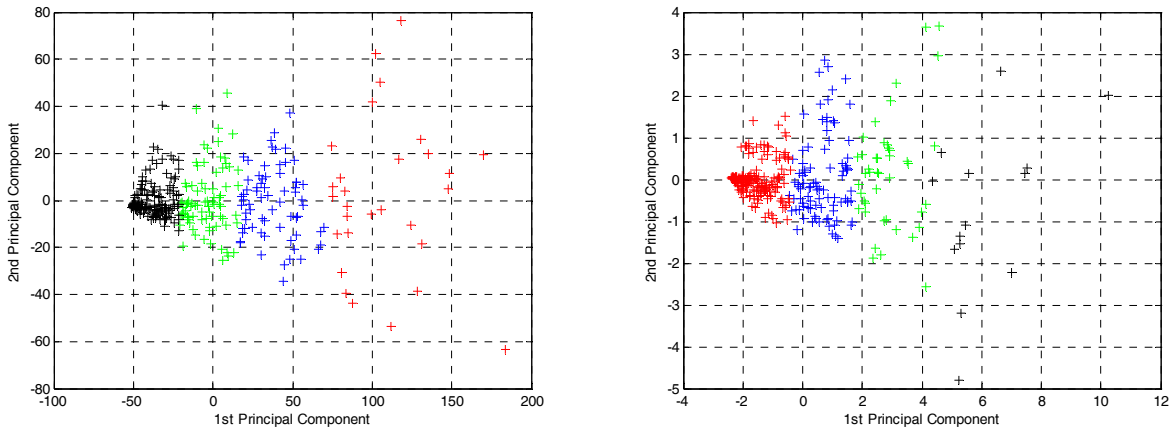
Dále zobrazíme jednotlivé objekty v rozptylovém diagramu komponentního skóre pro první a druhou komponentu, tj. v nových souřadnicích. Následuje zobrazení skupin vývodů pro různé metody shlukování, aby bylo vidět, kde se skupiny nacházejí i ve dvojném grafu a uvést je v souvislost se znaky.

Jak budou rozděleny vývody podle K-means, bude zmíněno v následujícím odstavci, ale nebude už zmíněno v kapitole o použití CLU. Hierarchické aglomerační shlukování je provedeno metodami average (skupinový průměr) a weighted (vážený průměr vzdáleností mezi shluky) při užití eukleidovské vzdálenosti v programu Matlab.

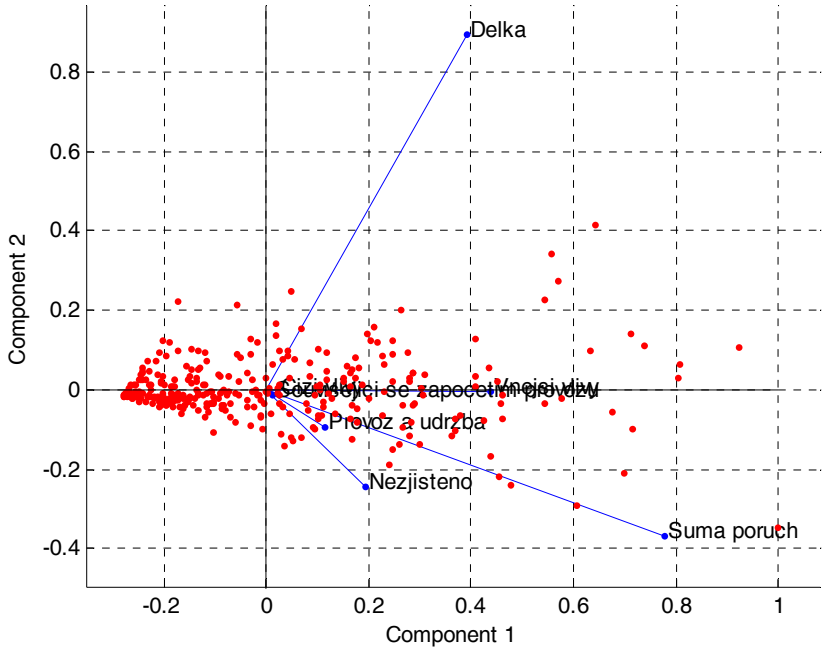
Použití nehierarchické shlukovací metody K-means rozdělí vývody na tolik skupin, kolik je zvoleno. V případě zvolení čtyř skupin vzniknou shluky při datech sloupcově centrovaných (při použití programu Statistica – startovací metoda pro počáteční středy shluků uspořádá vzdálenosti a vybere pozorování v konstantních intervalech a Eukleidovské metriky) o 143, 81, 68 a 29 vývodech. Za stejných podmínek, ale po provedení Z-score budou velikosti shluků následující: 174, 90, 43 a 14. Pro 5 shluků v programu Statistica při počátečních středech shluků zvolených z prvních pěti pozorování a při sloupcovém centrování původní matice dat jsou počty vývodů ve skupinách 134, 83, 58, 31 a 15 a pro Z-score jsou velikosti shluků vývodů 160, 74, 62, 16 a 9. Následující čtyři obrázky spolu souvisí.



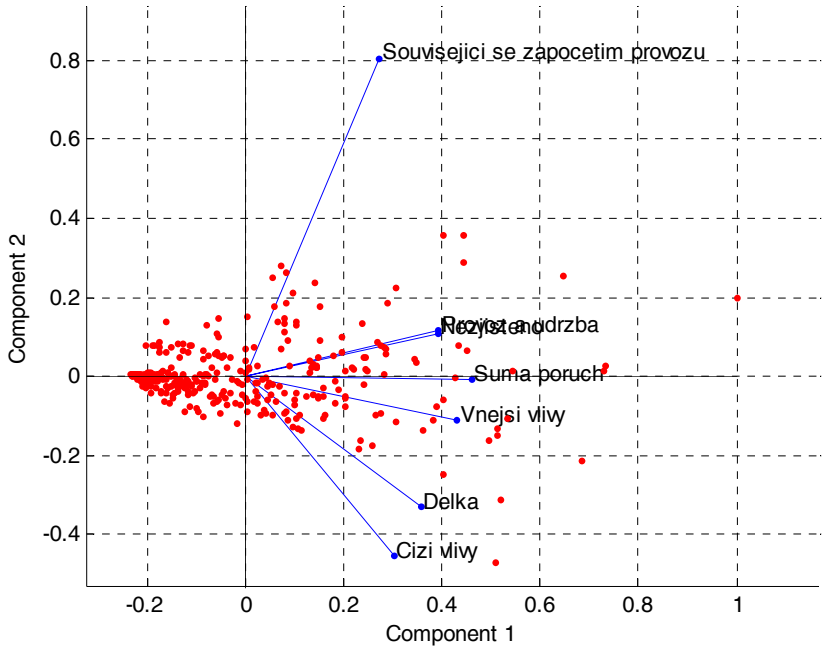
Obr. 6-8 Rozptylový diagram komponentního skóre po roztřídění metodou K-means na 5 shluků pro data sloupcově centrovaná (vlevo) a po provedení Z-score (vpravo)



Obr. 6-9 Rozptylový diagram komponentního skóre po roztřídění metodou K-means na 4 shluky pro data sloupcově centrovaná (vlevo) a po provedení Z-score (vpravo)



Obr. 6-10 Dvojný graf pro data sloupcově centrovaná



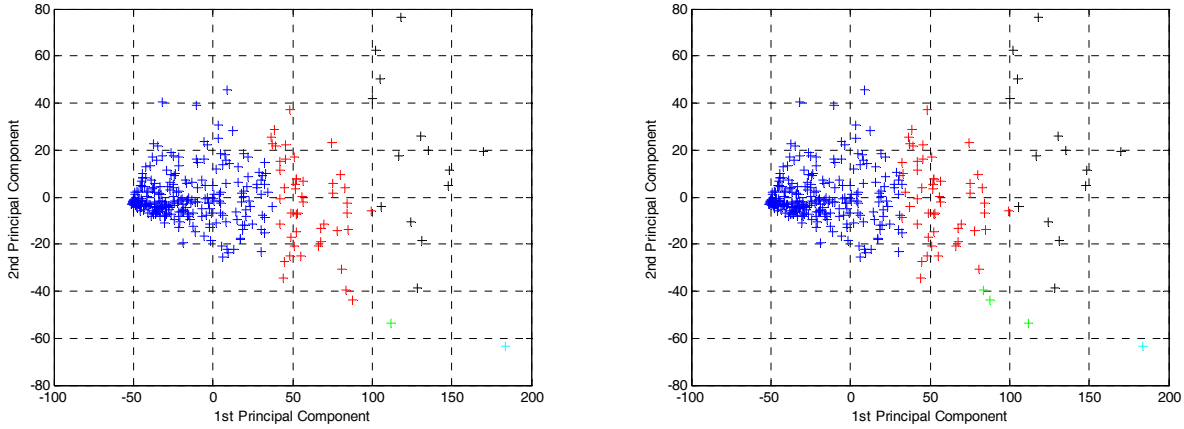
Obr. 6-11 Dvojný graf pro data po provedení Z-score

Sloupcově centrovaná data dávají podobný výsledek jako data nestandardizovaná. Z Obr. 6-10 a Obr. 6-11 je zřejmé, že rozdíl mezi sloupcově centrovanými daty a daty autoškálovanými je značný. Všechny následující obrázky včetně Obr. 6-8, Obr. 6-9 a přílohy D a E zobrazující rozptylový diagram komponentního skóre jsou vytvořeny tak, že matice skóre byla vynásobena -1 pro všechny sloupcově centrovaná data a pro data po Z-score se musí vynásobit pouze první sloupec matice skóre číslem -1. Tím bude korespondovat zobrazení s dvojným grafem na Obr. 6-10 a Obr. 6-11, které vytvořil Matlab. a dají se snáze určit vlastnosti jednotlivých skupin vývodů.

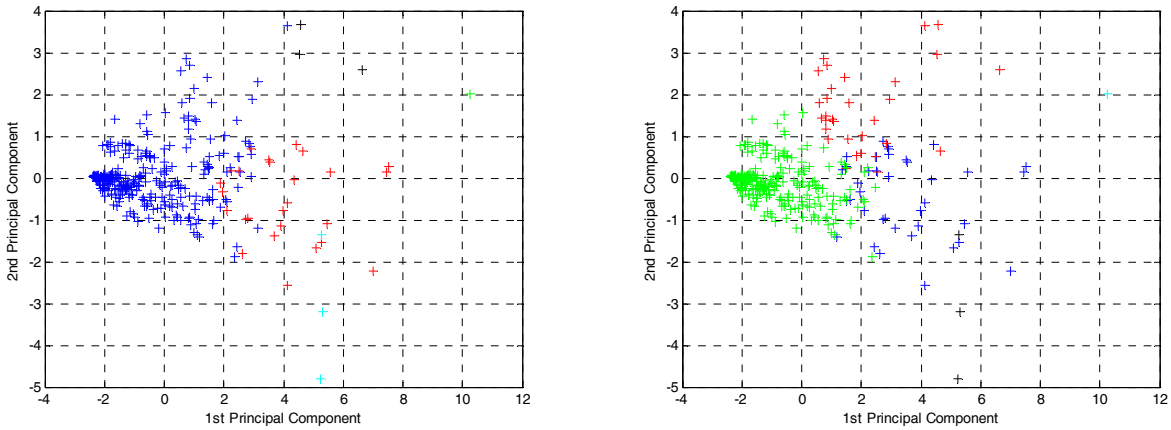
Z dvojných grafů a korelační matice zjistíme, že můžeme zmenšit počet znaků, neboť některé spolu silně korelují. Podle Obr. 6-10 znaky CV a ZP nejsou vůbec významné.

Na základě dvojných grafů a grafů komponentního skóre se shluky odlišenými barvami můžeme říct, že červený shluk na Obr. 6-9 vlevo a černý shluk na témže obrázku vpravo jsou vývody s velkým počtem poruch za 11 let. Přičemž ve vertikále jsou závislé na jiném znaku. Odlehlý objekt je v levém obrázku v pravém dolním rohu, vývod s indexem 202, a v pravém obrázku v pravém horním rohu, vývod s indexem 202. Podrobnější obrázek komponentního skóre s indexy jednotlivých vývodů je uveden v příloze D. Nejvíce odlehlé proto, že jsou nejvíce vzdálené na první hlavní komponentě. Pohled na znak Suma poruch a první hlavní komponentu vede k domněnce, že první hlavní komponenta určuje počet poruch za 11 let, ale čím větší délka, tím se zároveň umístění vývodu v rozptylovém diagramu komponentního skóre posouvá doprava. Čím méně poruch za 11 let a větší délka, tím je vývod výše a nalevo u dat po sloupcové standardizaci. Při autoškálování původní matice dat je tomu jinak.

Vývody umístěné vpravo nahoře u předchozích dvou rozptylových diagramů komponentního skóre (v obrázku vlevo, data po sloupcovém centrování) jsou delší než dole vpravo, ale všechny vývody vpravo mají hodně poruch za 11 let. Vývody, které mají málo poruch způsobených provozem a údržbou, leží ve čtvrtém kvadrantu. Vnějšími vlivy téměř neovlivněné vývody se nacházejí úplně vlevo na rozptylovém grafu komponentního skóre.



Obr. 6-12 Rozptylový diagram komponentního skóre pro reprezentaci shluků vytvořených hierarchickým shlukováním, data po sloupcovém centrování, eukleidovská vzdálenost, metoda shlukování average (vlevo) a weighted, (vpravo)

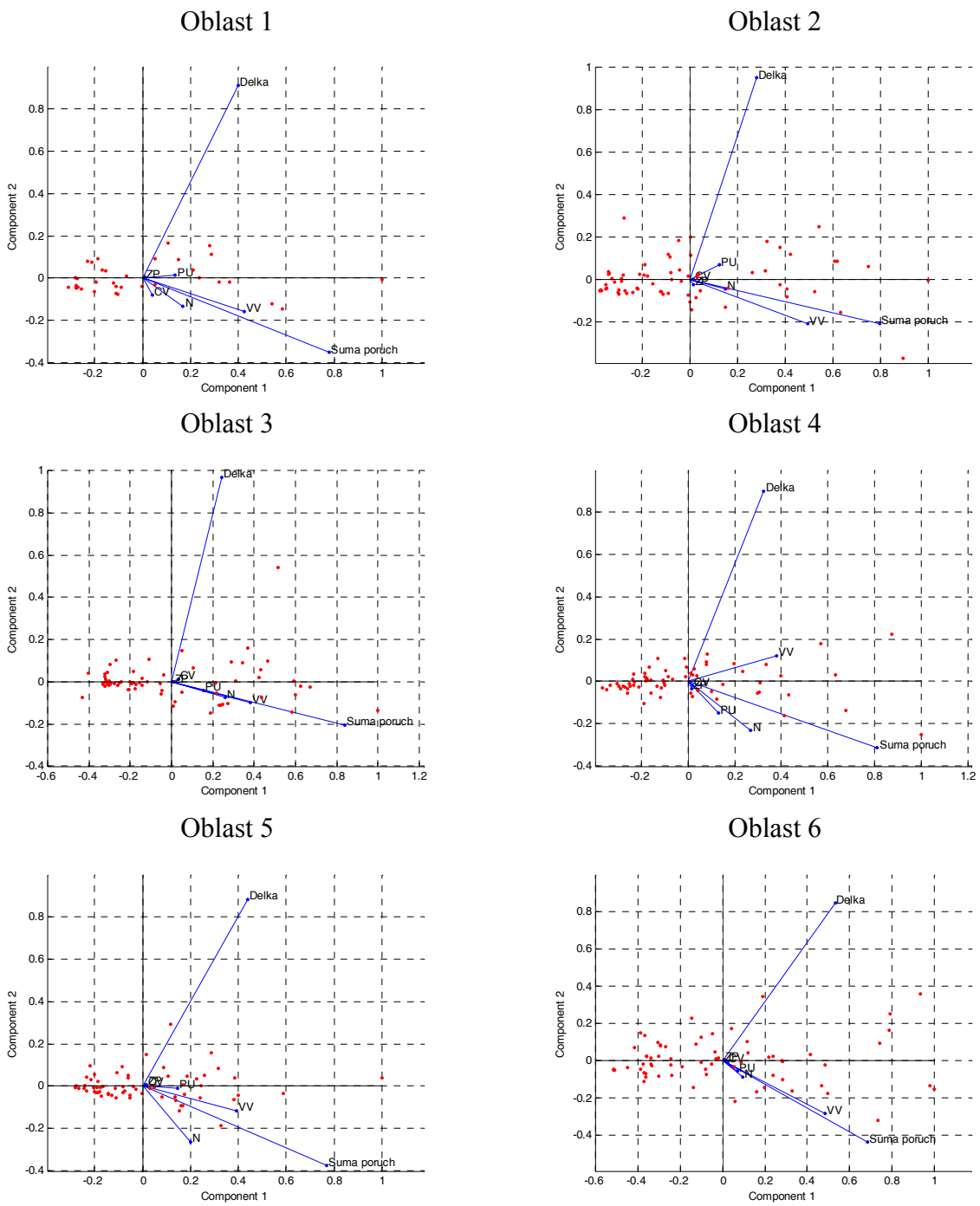


Obr. 6-13 Rozptylový diagram komponentního skóre pro reprezentaci shluků vytvořených hierarchickým shlukováním, data po Z-score, eukleidovská vzdálenost, metoda shlukování average (vlevo) a weighted (vpravo)

Výše dva uvedené obrázky se opět mohou dát do souvislostí s Obr. 6-10 a Obr. 6-11. Hned je očividné, že hierarchické a nehierarchické shlukování dává jiné výsledky a to i při spojení

s dvojným grafem, kde na Obr. 6-13 vlevo jsou azurovou barvou označeny vývody s malým výskytem poruch nezjištěných, kdežto černá ukazuje vývody s větším počtem nezjištěných poruch. Jinak se ještě liší v cizích vlivech a poruchami souvisejícími se započetím provozu a to tak, kam směřuje graf komponentních vah, který zobrazuje znaky v nových ortogonálních souřadnicích hlavních komponent.

Porovnání modelu analýzy PCA pro celou síť a mezi jednotlivými oblastmi. Svým způsobem se chce ověřit, jestli model platí i na jednotlivé oblasti sítě. Vzhledem k tomu, že oblasti obsahují v porovnání s celou sítí třeba jen desetinu všech dat, se nejspíš dobré shody nedosáhne.

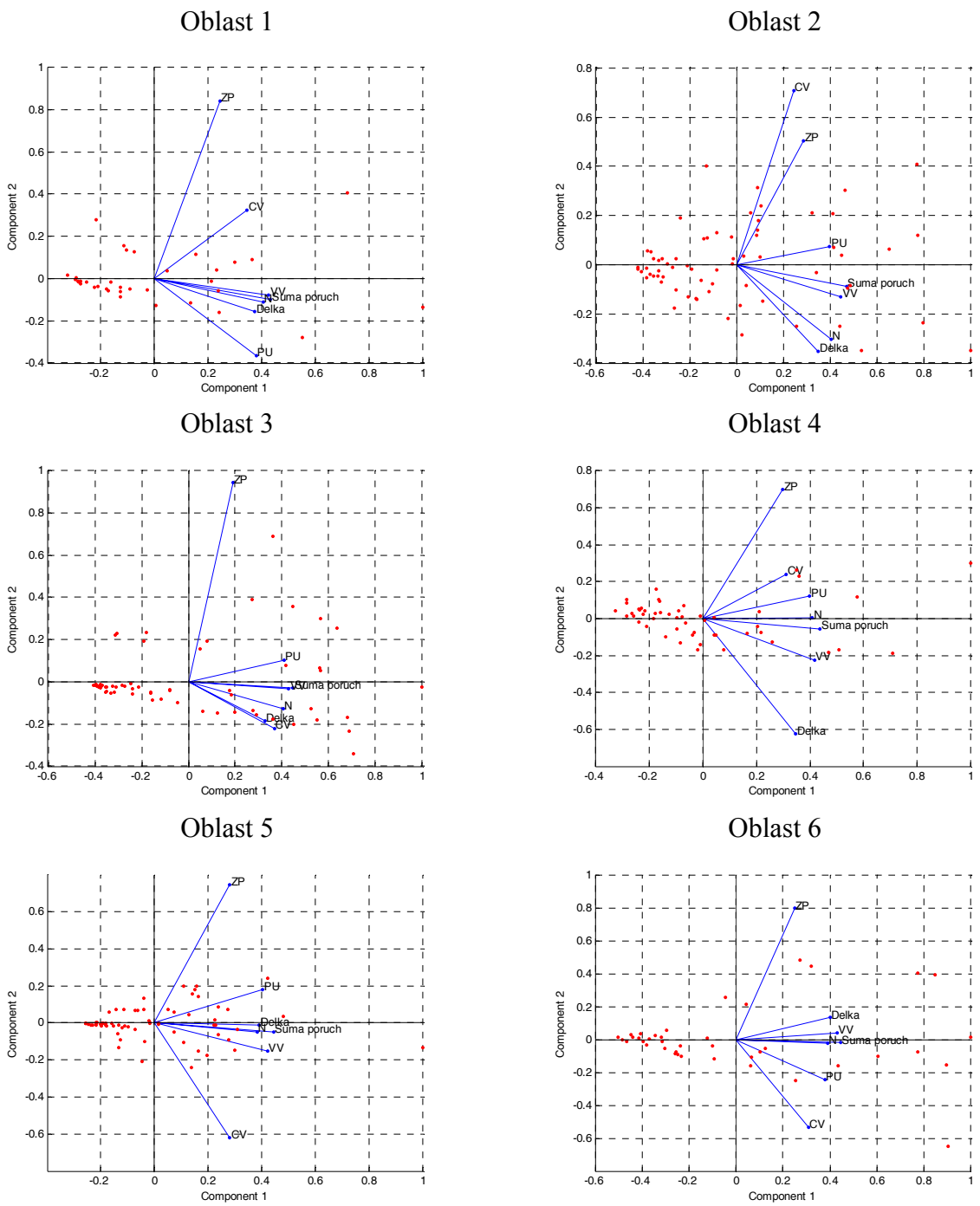


Obr. 6-14 Dvojné grafy pro jednotlivé oblasti, data po sloupcovém centrování

Znaky Délka a Suma poruch podle Obr. 6-14 prakticky nezměnily polohu v žádné oblasti. V oblasti 4 znak Vnější vlivy nekoreluje tak úspěšně se znakem Suma poruch jako u ostatních

oblastí, ale zato znak Nezjištěno je více úměrný Sumě poruch než u ostatních oblastí. Za významnější znak pro analýzu lze ještě považovat znak Nezjištěno.

U dat autoškálovaných je situace jiná, neboť tato standardizace zajistí každému znaku podobnou váhu. Znak Délka zde získal jinou podobu a je zřejmé, že přispívá více k první hlavní komponentě, než u dat sloupcově centrovaných. Znaky CV a ZP tvoří významně druhou hlavní komponentu, kdežto u sloupcového centrování nehrály žádnou roli. V žádném zobrazeném dvojném grafu nenalezneme zápornou korelaci dvou znaků. Můžeme říct, že největší podíl na počtu poruch za 11 let mají poruchy způsobené vnějšími vlivy a nezjištěné poruchy, do kterých jsou zahrnuty i krátkodobé výpadky.

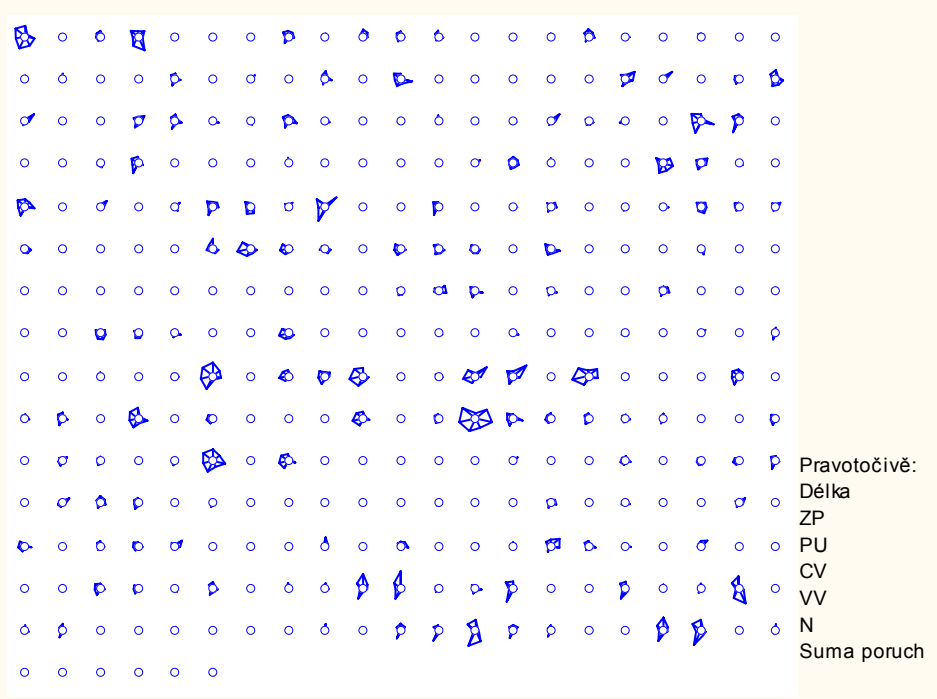


Obr. 6-15 Dvojné grafy pro jednotlivé oblasti, data po Z-score

7 ROZDĚLENÍ VÝVODŮ PODLE CLU A DÉLEK

7.1 Posouzení vlivných bodů

Odhalíme odlehlé body pomocí ikonového grafu hvězdy na Obr. 7-1. Data před analýzou byla standardizována aplikací Z-score (autoškálování) nebo sloupcovým centrováním. Použijeme-li při sloupcovém centrování místo aritmetického průměru medián, výsledný dendrogram se nezmění. Názvy metod aglomeračního hierarchického shlukování budou odpovídat názvům v programu Matlab.



Obr. 7-1 Ikonový graf – hvězdy

Odstraněné objekty na základě ikonového grafu jsou vývody s indexy 1, 4, 61, 81, 93, 174, 178, 181, 182, 184, 202, 216, 283, 284, 293, 307, 312 a 313. Tím se možná dosáhne lepšího modelu původních dat. Při pohledu na Přílohu A nebo B zjistíme, že z dat byly převážně vyjmuty objekty s velkým součtem poruch za 11 let.

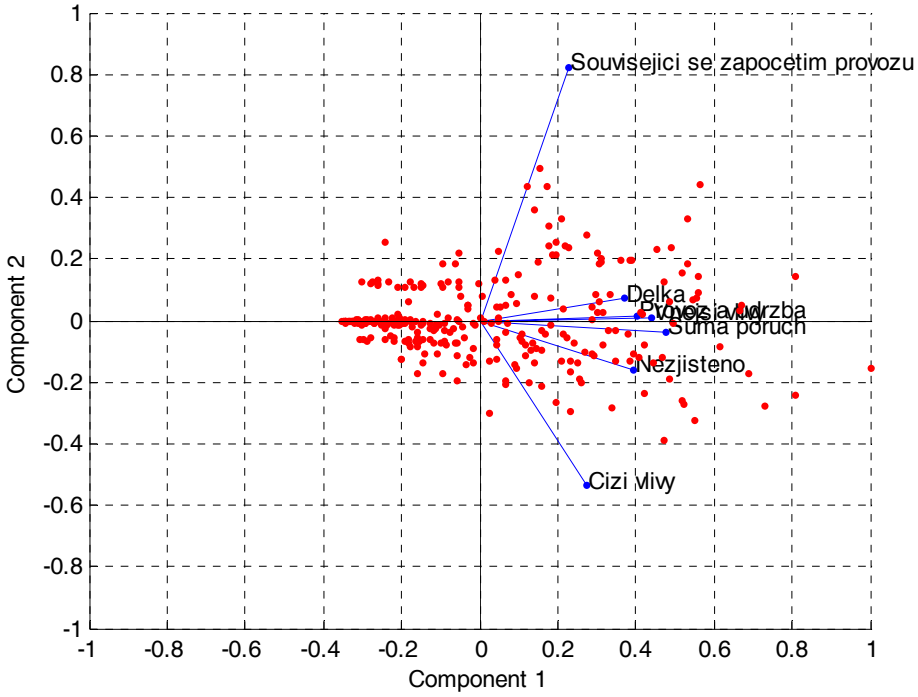
Tab. 7-1 Vlastní čísla korelační matice a vyšetření potřebných hlavních komponent (s vlivnými body) po provedení Z-score

index	vlastní číslo	% celk. rozptylu	kumulativní vlastní číslo	kumulativní procento
1	4,5673	65,2473	4,5673	65,2473
2	0,8034	11,4764	5,3707	76,7237
3	0,6152	8,7886	5,9859	85,5123
4	0,5078	7,2539	6,4936	92,7662
5	0,3168	4,5264	6,8105	97,2926
6	0,1895	2,7074	7,0000	100,0000
7	0,0000	0,0000	0,0000	100,0000

Z této tabulky odhalíme počet potřebných hlavních komponent, při uvažování dvou je objasněnost rozptylu v původních datech cca 76,72%.

Tab. 7-2 Korelační matice bez vlivných bodů

Znak	Délka	ZP	PU	CV	VV	N	Suma poruch
Délka	1	0,28938	0,53112	0,29458	0,72715	0,49217	0,70142
ZP	0,28938	1	0,40439	0,13623	0,35203	0,25472	0,40078
PU	0,53112	0,40439	1	0,47306	0,64787	0,60817	0,79710
CV	0,29458	0,13623	0,47306	1	0,39255	0,39241	0,49066
VV	0,72715	0,35203	0,64787	0,39255	1	0,67955	0,94650
N	0,49217	0,25472	0,60817	0,39241	0,67955	1	0,84288
Suma poruch	0,70142	0,40078	0,79710	0,49066	0,94650	0,84288	1



Obr. 7-2 Dvojný graf bez vlivných bodů

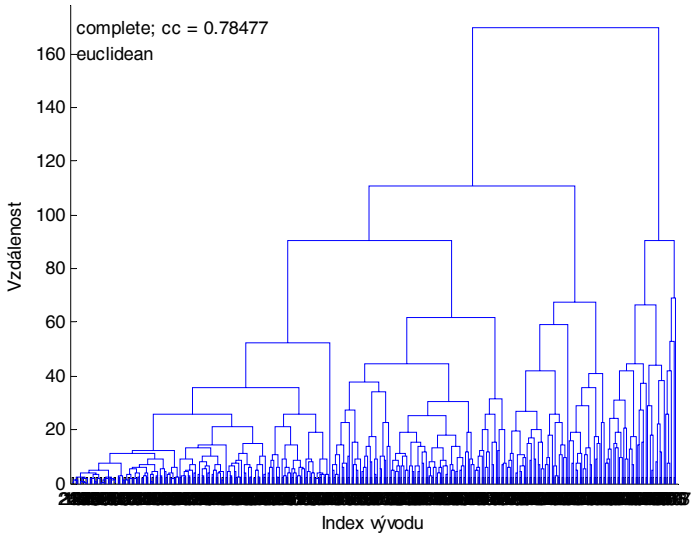
Tab. 7-3 Porovnání kofenetických korelačních koeficientů pro různé metody shlukování

Metoda (Matlab)	Sloupcové centrování		Z-score	
	Bez vlivných bodů	Původní data	Bez vlivných bodů	Původní data
single	0,75630	0,83206	0,82255	0,88129
complete	0,78477	0,83403	0,82064	0,82779
average	0,82163	0,87924	0,83161	0,86222
weighted	0,80112	0,86270	0,83203	0,78764
ward	0,66993	0,60010	0,64136	0,54530
centroid	0,81453	0,84932	0,82796	0,88559
median	0,74644	0,81782	0,74662	0,69790

V Tab. 7-3 je zvolena Eukleidovská metrika. Odstraněním výše uvedených vlivných bodů nedojde u CLU k lepší těsnosti proložení, čili kofenetický korelační koeficient *cc* nebude větší; $cc \in \langle 0;1 \rangle$. Porovnáno s *cc* uvedené v každém obrázku vyobrazující dendrogram v následujících částech kapitoly.

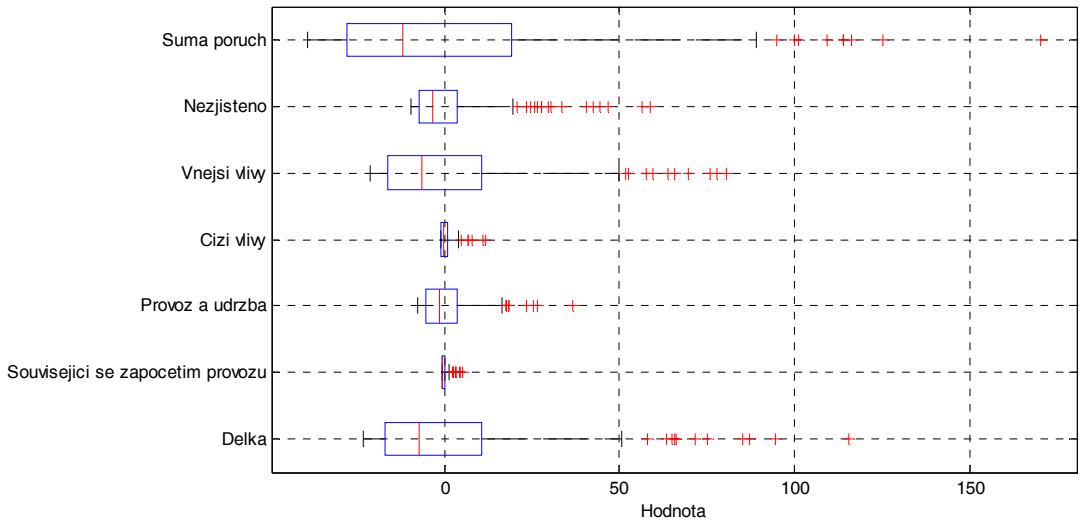
Z Obr. 7-2 vidíme, že rozprostření bodů reprezentující objekty je celkem rovnoměrné a svědčí o dobrém modelu původních dat analýzou hlavních komponent. Tento obrázek by měl být porovnáván s dvojným grafem na Obr. 6-11. Lze usuzovat na velkou závislosti objektů na znacích PU, VV, N, Délka a Suma poruch. Při autoškálovaných datech se korelační matice rovná kovarianční (směrodatná odchylka je rovna jedné). Velká korelovatelnost značí lineární závislost,

a ta je mezi znaky VV a Suma poruch. Můžeme vidět, že velká část objektů je popsána znakem Suma poruch, který se asi bude nejvíce blížit první hlavní komponentě. Také můžeme říct, že nejvíce poruch na vedení je způsobeno vnějšími vlivy. Velká část z počtu poruch je neurčených. Délka více či méně koreluje také s počtem vnějších vlivů. Korelace délky vývodů a sumy poruch za 11 let je také významná.



Obr. 7-3 Příklad dendrogramu po odstranění vlivných bodů

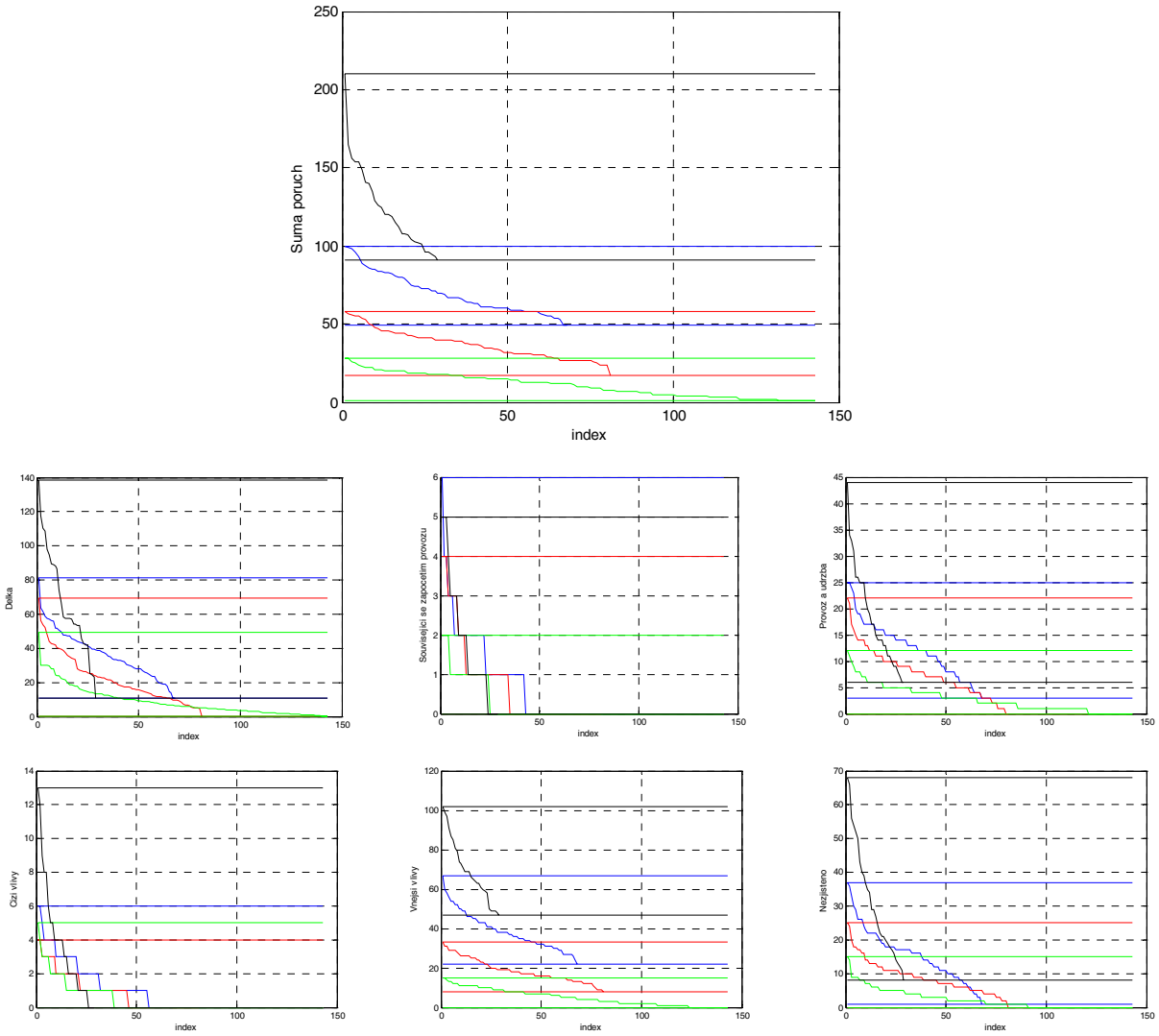
Na Obr. 7-3 je použito sloupcového centrování, rozdělení je rovnoměrnější, než jak tomu bude u dendrogramů z původní matice dat po standardizaci uvedených dále. Na druhou stranu se u tohoto rozdělení vývodů zhoršilo kritérium *cc*.



Obr. 7-4 Krabicový graf pro sloupcově centrovaná data s výše uvedenými znaky

7.2 Nehierarchické shlukování

Rozdělení vývodů pomocí K-means bylo zmíněno v minulé kapitole. Kompletní tabulky rozdělení na čtyři shluky a jejich základní popisné statistiky jsou uvedeny v příloze A a C. Počet vývodů v jednotlivých skupinách po rozdělení programem Statistica při výše zmíněné startovací metodě je 143, 81, 68 a 29. Program Matlab při třídění podle zvolené startovací metody vyhodnotil velikost a obsazení shluků po každém spuštění jinak, ale řešení se pro zvolenou startovací metodu opakovaně zobrazovalo. Program Statistica byl v tomto ohledu stabilní a dával pro zvolenou startovací metodu po každém spuštění stejný výsledek.



Obr. 7-5 Zobrazení jednotlivých skupin vývodů při použití sloupcového centrování po rozdělení podle K-means na čtyři skupiny

Předchozí obrázek ukazuje, jak jsou hodnoty jednotlivých znaků po roztřídění podle metody K-means při datech sloupcově centrovaných rozděleny mezi skupiny vývodů rozlišených barvami. Jako nejdůležitější parametr je zde vybrán znak Suma poruch, který má největší proměnlivost v datech, což bylo vyvozeno z Obr. 7-4. Obr. 7-5 vznikl seřazením hodnot pro každý znak v jednotlivých skupinách vývodů.

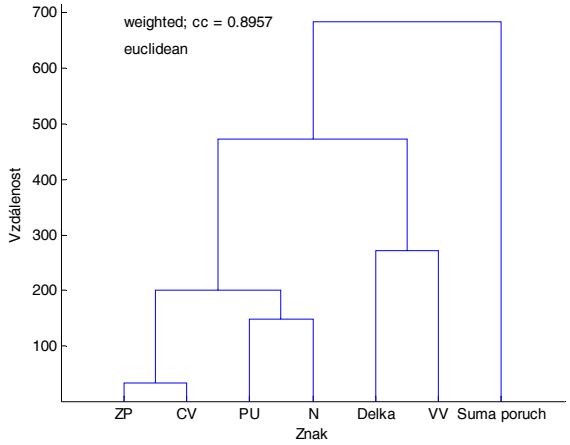
Obr. 7-5 můžeme použít jako pomůcku ke krabicovému grafu, v níž vidíme i kolik obsahuje skupina vývodů. A navíc z tohoto obrázku je i patrný rozdíl při uvažování více charakteristických znaků objektů oproti tomu, když se bere v úvahu pouze jeden znak.

7.3 Hierarchické shlukování

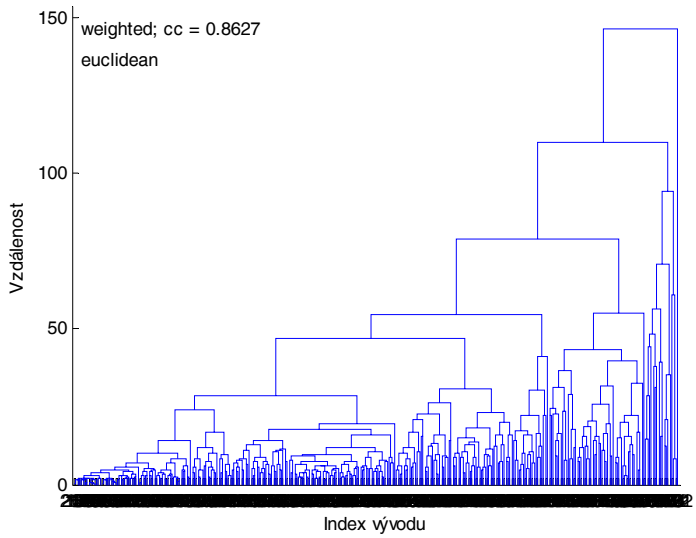
Slovo weighted v Obr. 7-6 znamená, že byla použita metoda shlukování s váženým průměrem vzdáleností mezi shluky, euclidean informuje, jak byla vypočítána vzdálenost. Dendrogram proměnných, pro metriku shlukování metodou průměrovou, zvanou také group average, (v Matlabu average), vytvoří stejné shluky podobných znaků a jeho kofenetický korelační koeficient *cc* je jen o pár tisícin větší.

U všech vyobrazených dendrogramů vývodů nemůže být vidět index vývodu. Příklad rozdělení do shluků je uveden v příloze B, kde jsou data před analýzou sloupcově centrovaná, je

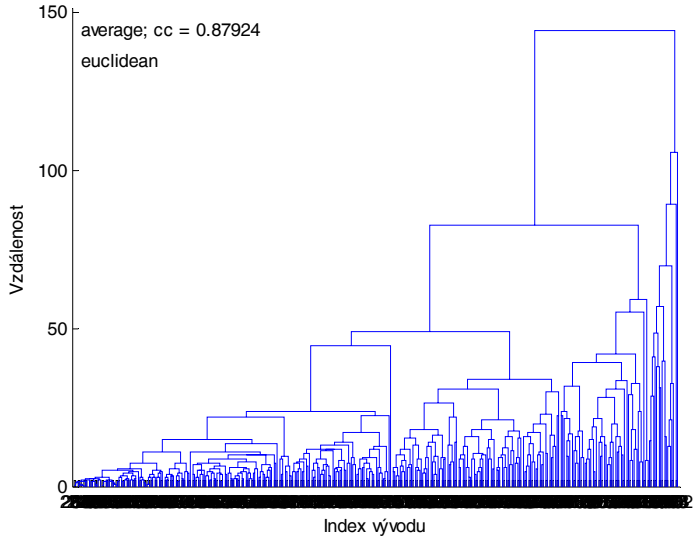
metoda shlukování weighted a metrika eukleidovská. Počet shluků je pět a odpovídají řezu ve vzdálenosti cca 75 na Obr. 7-7.



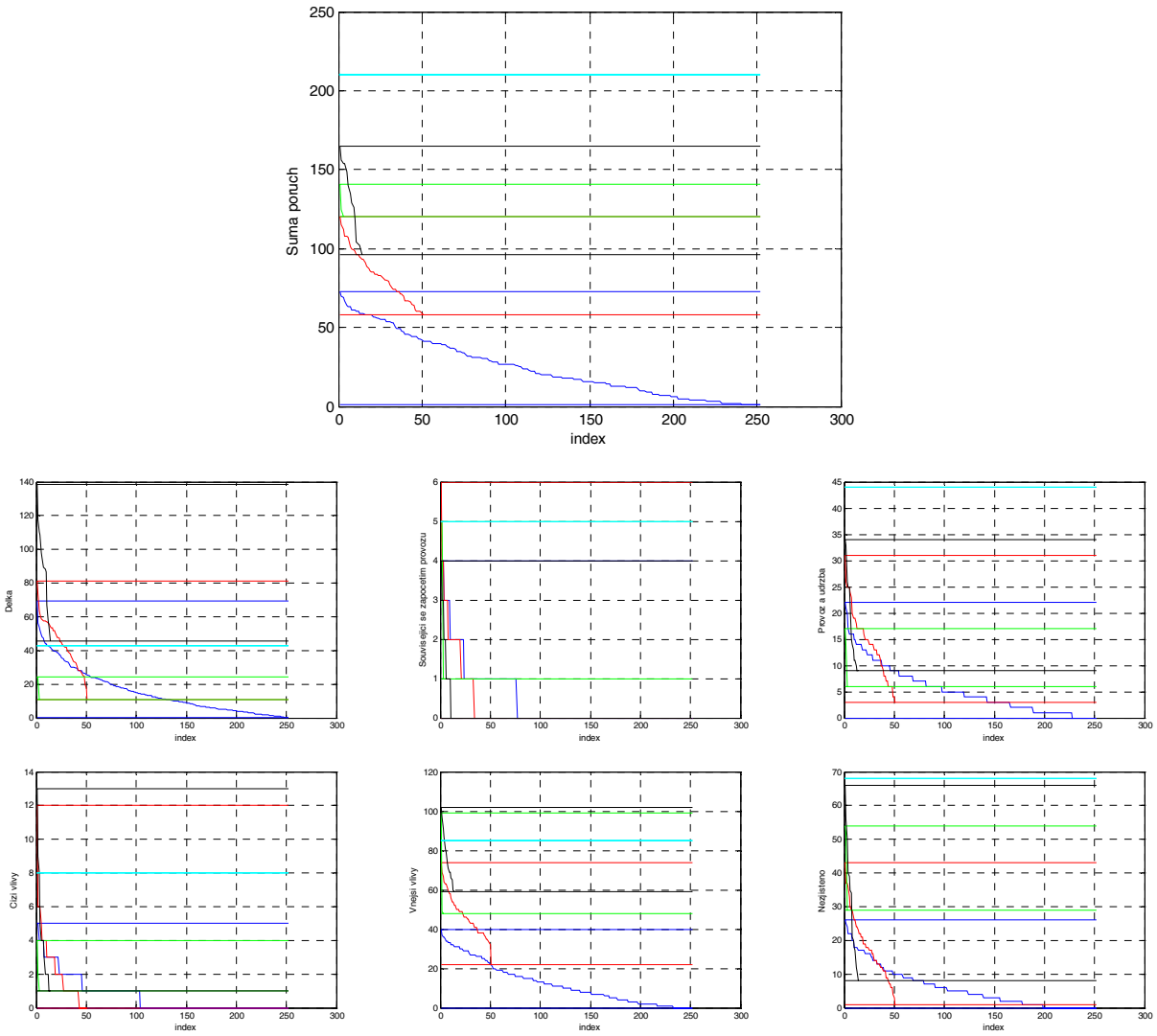
Obr. 7-6 Dendrogram proměnných, data po sloupcové standardizaci



Obr. 7-7 Dendrogram vývodů, data sloupcově centrovaná

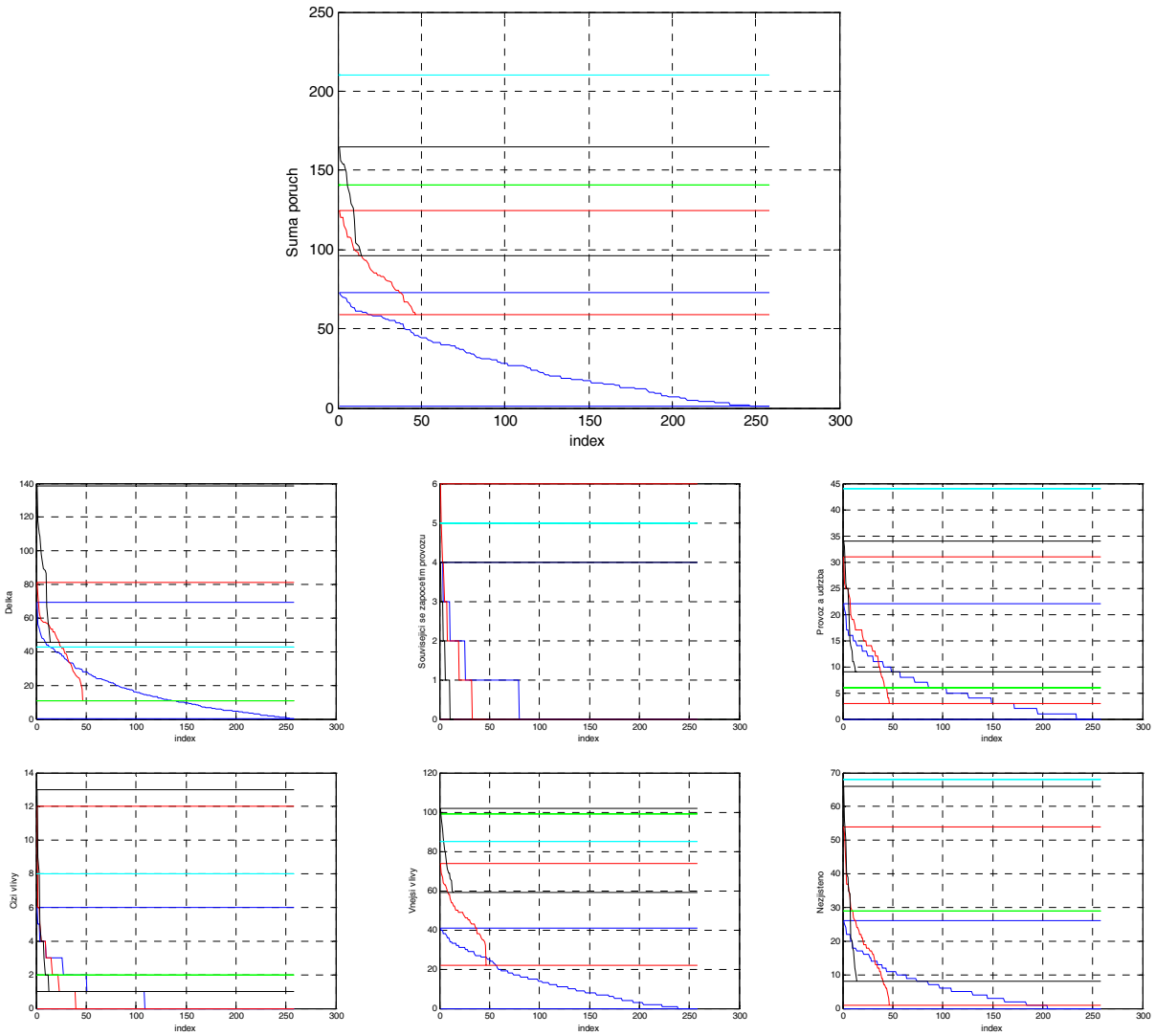


Obr. 7-8 Dendrogram vývodů, data sloupcově centrovaná

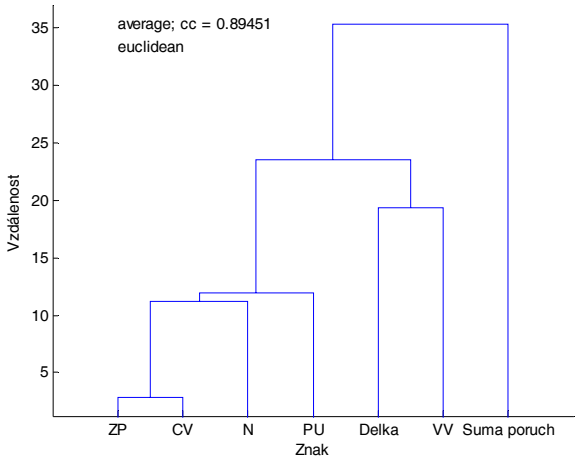


Obr. 7-9 Zobrazení jednotlivých skupin vývodů při použití sloupcového centrování po rozdělení hierarchickým shlukováním, shlukovací metoda weighted, eukleidovská vzdálenost

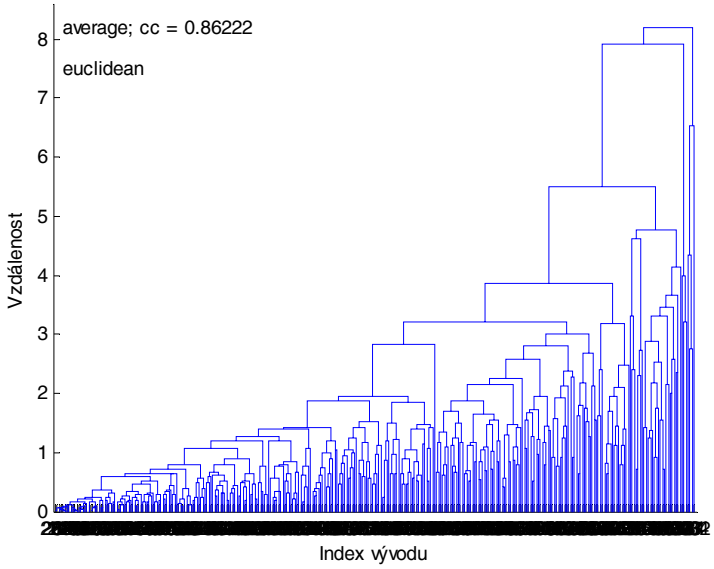
Tyto obrázky pro hierarchické shlukovací metody budou obsahovat pět shluků vývodů, neboť jak je patrné z dendrogramu, vyskytují se shluky, které obsahují velmi málo objektů. Jak bylo ukázáno dříve, odstraněním těchto vývodů na základě ikonového grafu nastalo horší proložení původních dat modelem vytvořeným jak shlukovacími metodami, tak i metodou hlavních komponent.



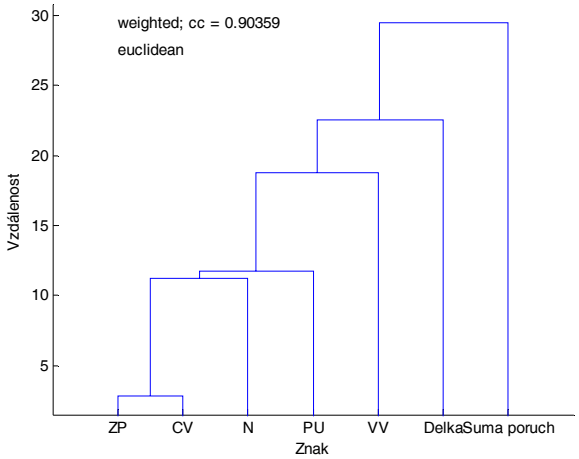
Obr. 7-10 Zobrazení jednotlivých skupin vývodů při použití sloupcového centrování po rozdělení hierarchickým shlukováním, shlukovací metoda average, eukleidovská vzdálenost



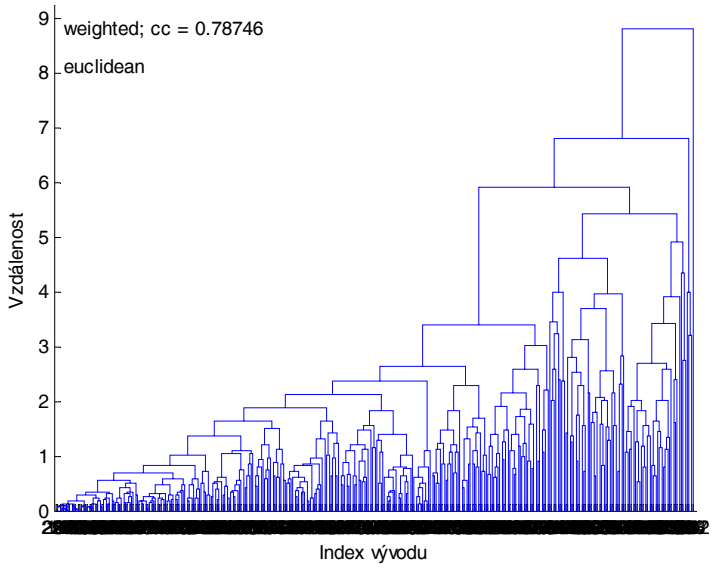
Obr. 7-11 Dendrogram proměnných, data po Z-score



Obr. 7-12 Dendrogram vývodů, data po Z-score



Obr. 7-13 Dendrogram proměnných, data po Z-score



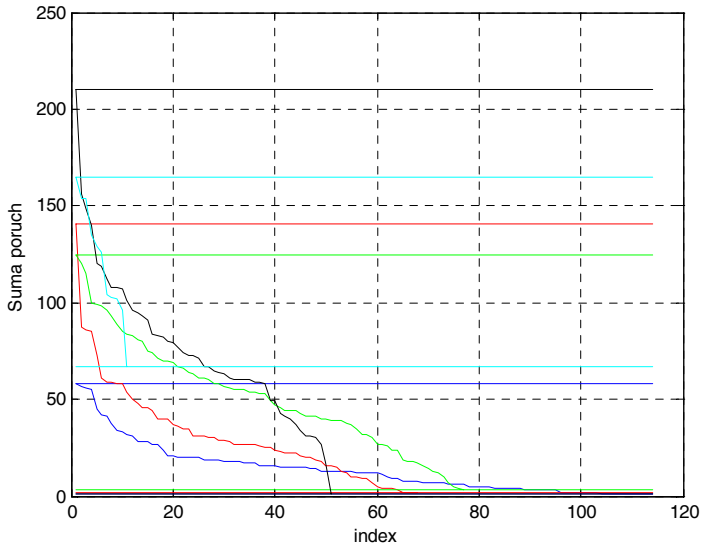
Obr. 7-14 Dendrogram vývodů, data po Z-score

Z kofenetického korelačního koeficientu *cc* bylo zjištěno, že data je výhodnější sloupcově centrovat. Při sloupcovém centrování se analýza chová, jako by byla použita nestandardizovaná data. Neprokázalo se, že by musela být nutně autoškálována, i když na první pohled na krabicový graf na Obr. 6-2 to vypadá, že velké rozdíly mezi znaky jsou. Zdá se, že data jsou mezi sebou nějakým způsobem svázána, v takových případech se standardizace nedoporučuje a zásah Z-score je velice významný.

Rozdělení vývodů se liší v intervalu velikostí hodnot jednotlivých znaků. Např. je-li podobný interval u znaku Délka v odlišných skupinách vývodů, budou různé intervaly ve znaku CV. Rozdělením se neukázala souvislost mezi vývody a oblastí, do které patří. Může se z toho vyvodit, že podmínky provozu vedení v jednotlivých oblastech jsou prakticky shodné.

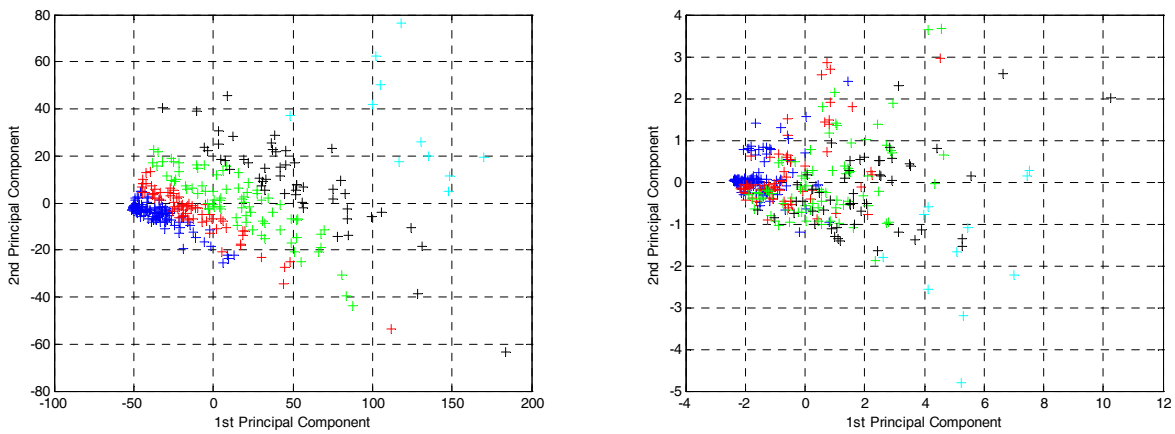
7.4 Rozdělení podle délek

Dělení vývodů podle délek se děje v závislosti na provádění simulací poruch na vývodu s použitím měrné poruchovosti vedení udávané v počtu poruch na 100 km za rok. Rozdělení vývodů podle jejich délky bylo vytvořeno v intervalech $\langle 0;10\rangle$; $(10;20\rangle$; $(20;40\rangle$; $(40;80\rangle$ a nad 80 km. Nebylo užito shlukovacích metod.



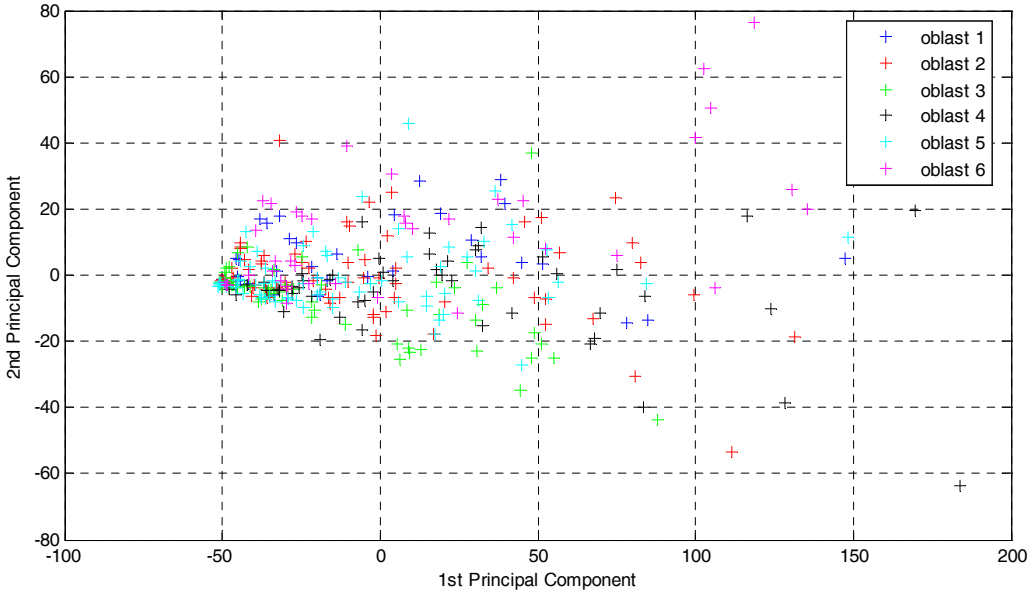
Obr. 7-15 Zobrazení skupin ve znaku Suma poruch při třídění pomocí délek vývodů

V Obr. 7-15 se skupiny vývodů ve zvoleném znaku značně kříží, při použití shlukovacích metod se křížení čar u významných znaků prakticky odstraní.



Obr. 7-16 Rozptylový diagram komponentního skóre pro data sloupcově centrovaná (vlevo) a data autoškálovaná (vpravo), barvy odpovídají třídění podle délek

Barvy na Obr. 7-16 jsou následující: modrá - $\langle 0;10 \rangle$; červená - $(10;20\rangle$; zelená - $(20;40\rangle$; černá - $(40;80\rangle$ a azurová – vývody nad 80 km délky. Všimneme-li si, jak jde spojnice středu a souřadnic znaku Délka ve dvojném grafu na Obr. 6-10, můžeme konstatovat, že sklon úsečky téměř koresponduje s tím, jak jsou zešikma vybarveny skupiny vývodů na Obr. 7-16 vlevo. Tedy délka vývodů roste s úsečkou mezi počátkem dvojného grafu a bodem určeným komponentními váhami pro první dvě hlavní komponenty znaku Délka. Bohužel na Obr. 7-16 vpravo se data hodně překrývají, ale tendence uvedená v předchozí větě je zachována. Právě na tomto můžeme ověřit, jestli model dat pomocí PCA se alespoň zčásti shoduje s předloženými daty. Podle tohoto a dvojného grafu zjistíme, že znak Délka svou šikmostí přispívá do první i druhé hlavní komponenty.



Obr. 7-17 Rozptylový diagram komponentního skóre pro data sloupcově centrovaná, barvy představují oblasti s vývody

Číslo oblasti je totožné s první číslicí u čísla vývodu.

8 POSOUZENÍ ROZDĚLENÍ VÝVODŮ PODLE CLU

Ukazatel průměrné systémové četnosti přerušení

$$SAIFI = \frac{\sum_u n_u O_u}{O_c}, \tag{8.1}$$

kde n_u je počet přerušení v u-tém uzlu za dané časové období, O_u je počet odběrných míst v u-tém uzlu a O_c celkový počet odběrných míst v síti.

Metodika sestavení Tab. 8-1 je taková, že bylo provedeno roztřídění do skupin podobných vývodů pomocí metod uvedených v prvním sloupci a za podmínek a stejnými znaky zmíněných v předešlých dvou kapitolách. Pro každou skupinu vývodů byla provedena simulace za 10 000 let pomocí negativně - binomického rozdělení (NBi). Tyto skupiny byly sloučeny do jedné tabulky a vypočítáno SAIFI za každý rok. Z těchto SAIFI byl proveden průměr a medián. Porovnání je provedeno vzhledem k původním netříděným datům, u nichž bylo SAIFI určeno stejným způsobem pro všech 321 vývodů. Původní zde značí dodaná data počtů poruch za jednotlivé roky v období 11 let.

Tab. 8-1 Přehled chyb simulace oproti dodaným datům.

metoda rozdělení	průměr			průměr SAIFI			medián SAIFI		
	původní	simulace	chyba	původní	simulace	chyba	původní	simulace	chyba
celá tabulka	3,40923	3,40759	0,00048	4,75964	3,40822	0,28393	4,79067	3,39328	0,29169
rozdělení podle délek		3,40715	0,00061		3,50325	0,26397		3,49728	0,26998
4 sk., K-means, sl. cen.		3,40997	-0,00022		3,20500	0,32663		3,19963	0,33211
4 sk., K-means, z-score		3,40877	0,00013		3,61010	0,24152		3,60377	0,24775
4 sk., K-means		3,40920	0,00001		2,91900	0,38672		2,91082	0,39240
CLU, weighted, z-score		3,40974	-0,00015		3,50246	0,26413		3,49466	0,27053
CLU, weighted, sl. cen.		3,40819	0,00031		3,51071	0,26240		3,50259	0,26887
CLU, average		3,40867	0,00016		3,51620	0,26125		3,50990	0,26735
3 sk., K-means		3,41094	-0,00050		3,42945	0,27947		3,42216	0,28566
5 sk., K-means, sl. cen.		3,40857	0,00019		3,21212	0,32513		3,20369	0,33126

Pozn.: Chyba SAIFI je brána vzhledem k prvnímu řádku tabulky k původním datům.

Chyba průměru musí být malá, neboť je dáno takové rozdělení (v tomto případě NBi), které by mělo co nejlépe kopírovat data týkající se určité skupiny vývodů.

Příklad výpočtu chyby průměru SAIFI pro 4 sk., K-means, sl. cen.:

$$\delta(SAIFI) = \frac{SAIFIp - SAIFIs}{SAIFIp} = \frac{4{,}75964 - 3{,}205}{4{,}75964} \doteq 0{,}32663\,. \tag{8.2}$$

Sestrojení grafu na Obr. 8-1.

Tab. 8-2 Rozdělení vývodů podle délek

délky v km	počet vývodů ve skupině	r	p	průměr poruch skupiny za rok	SAIFI data	SAIFI simulace	δ(SAIFI)
nad 80	11	4,17611	0,28399	10,529	6,03652	5,78230	0,04211
40 - 80	51	1,99661	0,23571	6,474	5,03872	4,98367	0,01093
10 - 20	68	1,04287	0,28245	2,663	2,33583	2,17995	0,06673
20 - 40	77	1,67952	0,27978	4,323	3,96879	3,51302	0,11484
do 10	114	0,78725	0,40046	1,179	1,40913	1,17878	0,16347

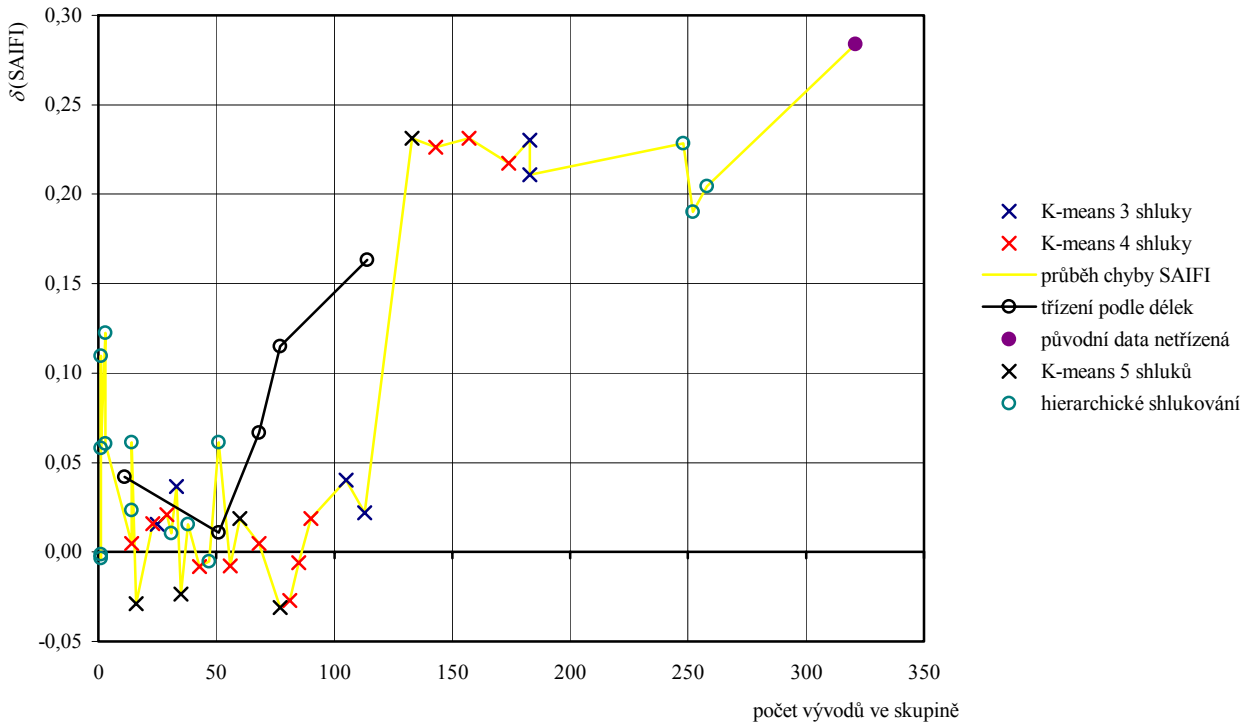
SAIFI data je průměr SAIFI za 11 let a SAIFI simulace je průměr SAIFI za 10 000 let. Chyba SAIFI je určena jako v příkladu výpočtu, tj. vztah (8.2), kde se váže k původním datům, ale zde je vypočítána chyba vzhledem k SAIFI data. Parametry r a p náleží k negativně – binomickému

rozdělení. Průměr poruch skupiny za rok je za sledované období jedenácti let. V Tab. 8-3 je počet vývodů ve skupině dán shlukovací metodou uvedenou v posledním sloupci (vysvětlivky pro metodu jsou pod tabulkou). Znaky pro shlukování jsou Délka, Poruchy související se započetím provozu, Poruchy provozem a údržbou, poruchy způsobené Cizími vlivy, Vnějšími vlivy, Nezjištěné poruchy a Suma poruch za 11 let.

Tab. 8-3 Rozdělení vývodů shlukovacími metodami

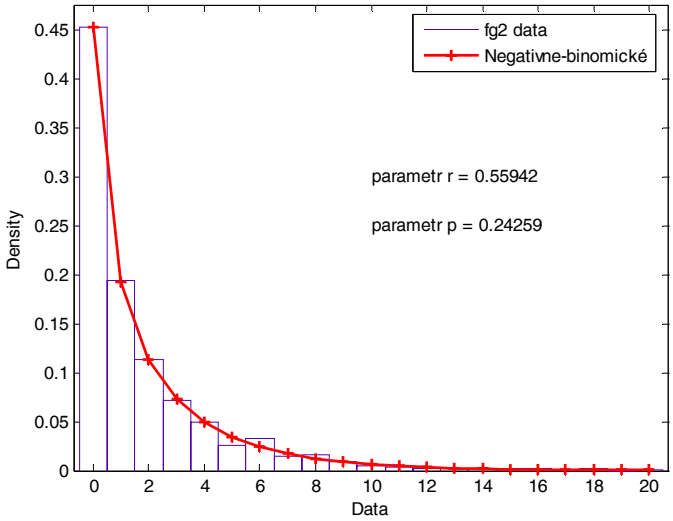
počet vývodů ve skupině	r	p	průměr poruch skupiny za rok	SAIFI data	SAIFI simulace	δ(SAIFI)	metoda
1	24,04280	0,702500	10,182	10,18182	10,1958	-0,00137	9_HSa
1	2,93860	0,142824	17,636	17,63636	17,6970	-0,00344	9_HSa
1	2,93860	0,142824	17,636	4,51200	4,01725	0,10965	9ZHSw
1	2,93860	0,142824	17,636	4,69960	4,42666	0,05808	9SHSw
3	4,64920	0,310308	10,333	3,40912	2,99171	0,12244	9ZHSw
3	4,37373	0,298620	10,273	4,06912	3,82284	0,06052	9SHSw
14	3,81665	0,240319	12,065	12,14045	12,08124	0,00488	9Z4
14	3,90955	0,256958	11,305	11,57502	11,30465	0,02336	9_HSa
14	3,90955	0,256958	11,305	4,71182	4,42268	0,06136	9SHSw
16	3,72616	0,242459	11,642	6,87470	7,07337	-0,02890	9S5
23	3,86681	0,259821	11,016	11,18263	11,00373	0,01600	9_4
25	4,10172	0,274983	10,815	10,99001	10,82031	0,01544	9Z3
29	3,66880	0,258510	10,524	8,40037	8,22657	0,02069	9S4
31	3,67327	0,387248	5,812	10,43124	10,32209	0,01046	9ZHSw
33	3,58640	0,259446	10,237	10,62311	10,23258	0,03676	9_3
35	4,36654	0,355482	7,917	5,45435	5,58249	-0,02349	9S5
38	3,47469	0,283542	8,780	17,64138	17,36900	0,01544	9ZHSw
43	4,13627	0,353568	7,562	7,49758	7,55908	-0,00820	9Z4
47	4,18012	0,361203	7,393	7,35354	7,39213	-0,00525	9_HSa
51	4,16877	0,371769	7,045	4,06269	3,81407	0,06120	9SHSw
56	4,36196	0,398898	6,573	6,52599	6,57753	-0,00790	9_4
60	4,32059	0,467505	4,921	5,01441	4,92032	0,01876	9S5
68	4,36287	0,424460	5,916	5,95592	5,92862	0,00458	9S4
77	3,02792	0,508894	2,922	2,83185	2,92016	-0,03118	9S5
81	3,36990	0,515850	3,163	3,07736	3,16095	-0,02716	9S4
85	3,44725	0,486357	3,641	3,61380	3,63508	-0,00589	9_4
90	3,77437	0,464603	4,349	4,43544	4,35282	0,01863	9Z4
105	3,87221	0,436997	4,989	5,20172	4,99206	0,04031	9_3
113	3,50056	0,401246	5,224	5,34386	5,22690	0,02189	9Z3
133	0,84618	0,504064	0,833	1,08363	0,83298	0,23131	9S5
143	0,88909	0,493050	0,914	1,17954	0,91278	0,22616	9S4
157	0,90254	0,464352	1,041	1,35545	1,04187	0,23135	9_4
174	0,89222	0,426426	1,200	1,53053	1,19809	0,21721	9Z4
183	0,92670	0,421528	1,272	1,65784	1,27634	0,23012	9_3
183	0,89270	0,411402	1,277	1,61661	1,27581	0,21081	9Z3
248	0,85046	0,283936	2,145	2,77765	2,14280	0,22856	9ZHSw
252	0,92742	0,306676	2,097	2,58970	2,09764	0,19001	9SHSw
258	0,92203	0,297838	2,174	2,72608	2,16917	0,20429	9_HSa
321	0,78595	0,187345	3,409	4,75964	3,40822	0,28393	původní

Z - provedení Z-score před analýzou
S - data pro analýzu sloupcově centrována
_ - data bez škálování
HS – hierarchické shlukování s výpočtem vzdálenosti pomocí Eukleidovské metriky
a – average, w – weighted – metody metriky shlukování
3, 4, 5 - odpovídá počtu skupin, na které se třídilo za užití metody K-means

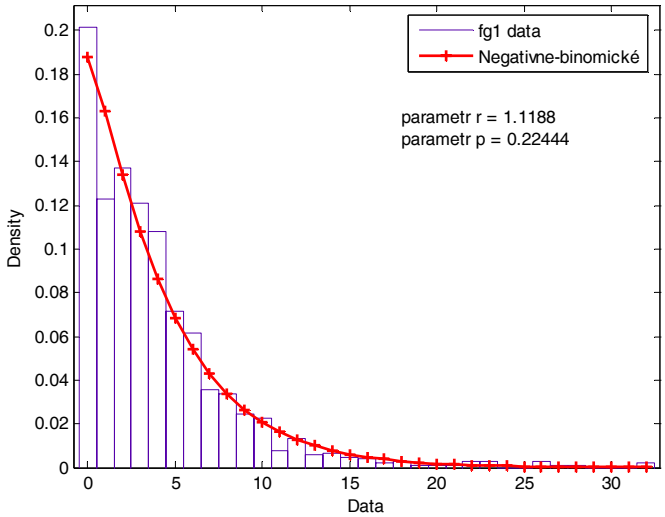


Obr. 8-1 Závislost chyby SAIFI na počtu vývodů ve skupině

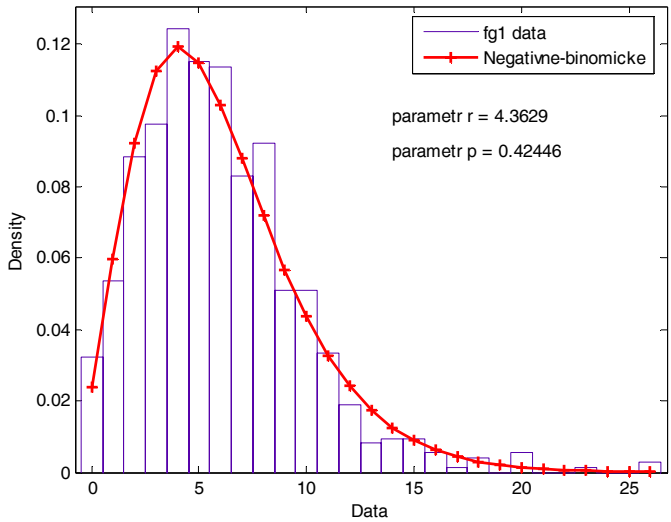
Ve skutečnosti je chyba SAIFI dána parametrem *r*, kterým je dáno negativně-binomické rozdělení. Je-li *r* menší než jedna, bude chyba větší než 20 %. Skupiny s velkým počtem vývodů obsahují vývody s převážně nulovým počtem poruch za rok. Toto dělá parametr *r* menším než jedna. Velký vliv má i průběh náhradního negativně-binomického rozdělení v histogramech na Obr. 8-2. Když je strmost mezi 0 poruchami za rok a 1 poruchou za rok větší než mezi 1 a 2, bude chyba větší než 20 %. Bude-li strmost mezi 0 a 1 menší než mezi 1 a 2, chyba bude do 5 %. Následující obrázek poslouží k objasnění chyby SAIFI při různé hodnotě parametru *r*.



a)



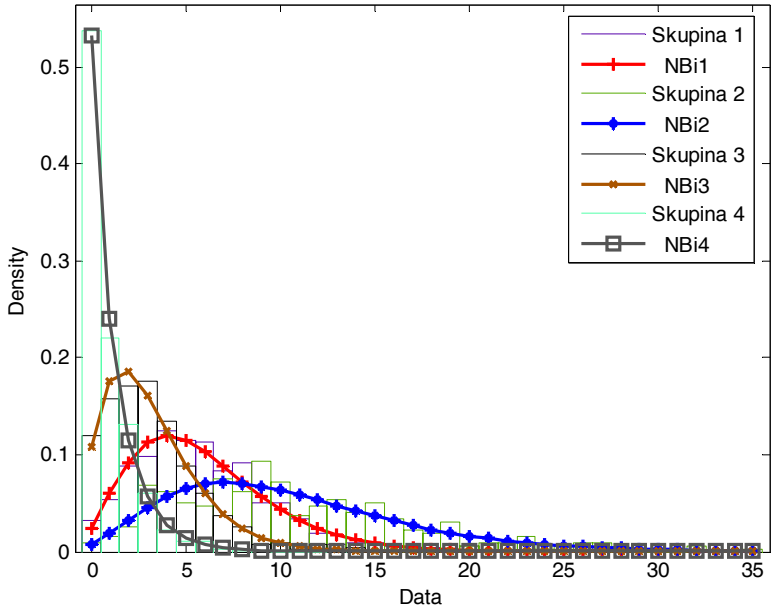
b)



c)

Obr. 8-2 Rozdílné tvary histogramů proložené negativně – binomickým rozdělením

Chyba SAIFI u prvního histogramu na Obr. 8-2 *a)* je až 35,6 % (122 vývodů ve skupině), u druhého cca 2,97 % (97 vývodů) a u třetího přibližně 0,458 % (68 vývodů). Na následujícím obrázku je zachyceno, jak vypadají četnosti a průběhy NBi rozdělení pro všechny vytvořené shluky pomocí CLU. Charakteristické pro distribuční síť je, že největší skupina vývodů má nejvíce výskytů s nulovou poruchou za rok. Četnost je vztažena k jednotlivým počtům vývodů ve skupině.



Obr. 8-3 Rozdělení K-means, 4 skupiny, rozložení četností poruch, parametry r a p odpovídají příslušným hodnotám v Tab. 8-3

9 ZÁVĚR

Posouzení vlivů zvolených znaků (Délka, ZP, PU, CV, VV, N, Suma poruch) na skupiny vývodů vytvořených hierarchickými a nehierarchickými shlukovacími metodami za pomoci analýzy hlavních komponent ukázalo, že použití sloupcového centrování a autoškálování na původní data vede k rozdílným diagramům, ale v podstatě jsou shluky vývodů popsány stejnými znaky. Nenašla se souvislost mezi počtem poruch na vývodu a oblastí, do které patří. Dobře se ukázalo, jak metody shlukování přiřadily vývody do skupin na základě podobných vlastností a intervalu hodnot patřičných znaků. Nejlépe je to vidět na Obr. 7-5, Obr. 7-9, Obr. 7-10, kdežto na Obr. 7-15, kde bylo třídění podle délek, se charakteristiky i u nejvýznamnějšího znaku značně kříží. Charakter skupin vývodů je ukázán v přílohách na dvou metodách shlukování. Asi nejcharakterističtějším znakem každého vývodu byl při analýze počet poruch za 11 let. Tento znak silně koreluje se znakem Vnější vlivy.

Byly provedeny i analýzy s jinými znaky než výše uvedenými, a to s ohledem na zkoumané vlastnosti sítě jako SAIFI, počet poruch za rok, celkový počet poruch za 11 let, doby přerušení, počet odběratelů na vývodu. U posledních dvou znaků je již Z-score plně opodstatněné.

Porovnáním kofenetických korelačních koeficientů *cc* a rozptylových diagramů komponentního skóre zjistíme, že vlivné body zjištěné z ikonového grafu jsou součástí modelu. Přesto, že některé tyto body tvoří shluky o malém počtu vývodů ve skupině, jsou pro daný soubor dat charakteristickými. Můžeme také říct, že čím byl *cc* vyšší, tím byly shluky vytvořeny tak, že jeden shluk obsahoval více než 250 vývodů z 321 a zbylé shluky obsahovaly i po jednom vývodu, a přesto odstraněním těchto shluků o jednom vývodu se lepšího rozdělení nedosáhlo.

Předložený soubor dat obsahoval 13 212 poruch na 368 vývodech, z toho bylo kompletních dat k analýze pro 321 vývodů. Data byla před analýzou buď sloupcově centrována, nebo autoškálována (Z-score). Lepších výsledků bylo dosaženo (tj. kritéria těsnosti proložení dat modelem bylo příznivější) za použití sloupcového centrování a to nejspíš z toho důvodu, že Z-score se lépe využije pro data velmi různorodá. Analýza za pomocí sloupcového centrování je téměř totožná, jako by data nebyla standardizována vůbec. Jak již bylo uvedeno dříve, je-li závislost mezi znaky významná, standardizace se nedoporučuje.

Důležitou otázkou je, jestli můžeme použít zmiňované metody, které byly vyvinuty na datech s normálním rozdělením. Délka vývodů má rozdělení exponenciální a počty poruch zase negativně – binomické.

Chyby SAIFI byly takřka stejné, ať byly použity pro analýzu jakékoliv znaky. Z mnoha simulací pro výpočet SAIFI za 10 000 let s porovnáním za 11 let a porovnávání s různými hodnotami parametru *r* pro negativně – binomické rozdělení, vycházely chyby podobně pro přibližně stejný parametr *r*. Na druhou stranu průměry a mediány z původních ročních poruch za 11 let pro vývod a za 10 000 let byly prakticky shodné, neboť pro proložení histogramu rozdělením bylo vybráno to, které nejvíce kopírovalo data. Jakmile se však dostalo na výpočet SAIFI, výsledek zdaleka nebyl tak optimistický. Můžeme si všimnout, že počet odběratelů není v analyzovaných znacích i přesto, že chceme tento znak vývodů zohlednit později. Ale i po jeho zařazení do analýzy byl výsledek úplně stejný jako u předchozích simulací.

Celé to ztroskotalo na chybě SAIFI, která byla až 25% (viz. Obr. 8-1, kde chyba prudce vzrostla od shluku obsahujícího cca 120 vývodů), kdy byla v histogramu proloženém rozdělením jen část klesající, podobně jako na Obr. 8-2 *a)*, což je skupina vývodů, kde jsou převážně zastoupeny vývody bez poruchy za rok. Ale malé chyby bylo dosaženo i v histogramu podobnému Obr. 8-2 *b)*. Současně byly pro simulace použity modely s rozdělením vedení podle délek vývodů. Všechny simulace dopadly obdobně. Chyba SAIFI závisí na parametru *r*.

Z těchto výroků je těžké usuzovat na výhodnost využití, ale jako pomůcka pro popis vlastností skupin vývodů by metody mohly být použity. Osobně si myslím, že existují možnosti využití vícerozměrné statistické analýzy dat při hodnocení spolehlivosti distribučních sítí.

POUŽITÁ LITERATURA

[1] FAJMON, B., RŮŽIČKOVÁ, I. Matematika 3. Elektronické skriptum VUT Brno, Brno 2004

[2] LIKEŠ, J., MACHEK, J. Počet pravděpodobnosti. SNTL, Praha 1981

[3] MELOUN, M. Přednáška Postup analýzy shluků. [citace 2008-31-10]. Dostupné z http://meloun.upce.cz/separaty/204slidy.pdf

[4] MELOUN, M., MILITKÝ, J. Kompendium statistického zpracování dat. ACADEMIA, Praha 2005

[5] MELOUN, M., MILITKÝ, J. Statistické zpracování experimentálních dat. PLUS, Praha 1994

[6] MELOUN, M., MILITKÝ, J, HILL, M. Počítačová analýza vícerozměrných dat v příkladech. ACADEMIA, Praha 2005, 1. vydání

[7] NĚMEC, P. Seminární práce Statistická analýza vícerozměrných dat. ÚKZÚZ Brno, Brno 2003, [citace 2008-31-10]. http://meloun.upce.cz/licencni/semprace/nemec/nemec1.pdf

[8] Program MATLAB Version 7.0.1.24704 (R14) Service Pack 1, 2004

[9] Program STATISTICA Verze 7 CZ

Příloha A Rozdělení vývodů nehierarchickou shlukovací metodou K-means na čtyři skupiny

Počáteční středy shluků uspořádány podle vzdálenosti a vybrány pozorování v konstantních intervalech, Statistica.

Skupina vývodů, černá barva v Obr. 6-9 vlevo.

Index	Vývod	Délka	ZP	PU	CV	VV	N	Suma poruch
7	101205	12,100	0	9	1	4	1	15
9	102206	22,200	0	5	0	7	4	16
13	103031	3,600	0	2	0	3	0	5
14	103038	3,900	0	3	0	5	0	8
18	103123	6,000	1	4	3	2	7	17
21	103196	3,200	0	5	0	0	1	6
22	103201	27,900	0	5	2	2	1	10
24	104207	10,000	0	0	0	2	1	3
25	104903	5,071	0	4	0	1	0	5
27	105028	16,200	1	3	2	14	2	22
28	105052	10,000	2	1	0	0	0	3
29	105142	21,500	0	3	0	10	0	13
31	106050	24,200	0	4	0	2	1	7
33	106145	23,900	0	1	0	3	0	4
34	107122	4,000	0	3	0	3	0	6
35	107188	2,200	0	0	0	1	0	1
37	201101	18,090	0	2	0	10	4	16
44	202271	8,919	0	4	1	10	2	17
45	202275	6,047	0	2	0	11	2	15
49	203192	14,401	0	4	3	3	0	10
52	203270	17,695	1	5	0	7	9	22
55	204277	6,536	0	0	0	2	0	2
63	206280	13,248	0	0	0	2	0	2
64	206288	3,090	0	0	0	7	2	9
65	206290	4,089	0	3	0	1	0	4
68	207066	12,450	0	2	0	4	3	9
69	207258	2,313	0	0	0	0	1	1
70	207259	24,109	0	4	0	8	6	18
71	207265	49,418	0	0	0	0	1	1
72	207267	3,538	0	1	1	0	2	4
73	207268	10,792	0	3	0	1	1	5
74	207269	13,643	1	2	0	0	0	3
76	209115	17,722	2	5	1	7	5	20
79	209262	11,070	0	1	1	6	1	9
80	210015	8,510	1	6	0	7	5	19
83	210148	2,000	0	12	0	0	0	12
84	210149	1,967	0	1	0	0	0	1
86	210260	4,800	0	5	0	4	1	10
88	210263	9,132	0	4	0	1	2	7
89	210264	3,826	2	3	4	2	5	16
94	211151	5,779	1	1	0	4	7	13
97	211289	15,000	0	0	1	1	0	2
98	212081	13,176	0	11	2	11	1	25
100	212285	2,000	1	1	0	0	0	2
109	302129	3,050	1	2	0	0	1	4
110	302137	2,200	1	3	0	1	0	5
115	302184	5,100	0	3	2	6	8	19
119	303094	6,105	0	7	0	7	3	17
121	303157	3,188	0	6	3	12	7	28
122	303161	4,665	1	8	0	7	0	16
123	303162	2,295	0	1	0	1	0	2
124	303167	8,035	0	10	5	8	5	28
125	303168	1,165	0	2	0	1	0	3
126	303169	1,480	0	1	0	3	0	4

Index	Vývod	Délka	ZP	PU	CV	VV	N	Suma poruch
154	309164	3,460	0	0	1	0	1	2
156	309190	11,400	0	0	0	2	0	2
157	311140	2,240	0	1	0	2	3	6
158	311141	0,980	0	2	0	1	1	4
160	312014	4,805	1	3	1	7	2	14
162	313187	6,030	0	1	0	0	0	1
163	315175	1,755	0	1	0	2	0	3
166	401043	7,265	2	6	1	4	5	18
167	401044	6,846	0	4	0	7	5	16
170	401318	9,038	1	1	0	11	6	19
172	401331	13,125	1	2	0	13	4	20
183	403016	0,100	0	2	0	2	0	4
187	403302	4,325	0	5	0	5	5	15
189	404030	0,645	1	3	0	0	3	7
192	404107	12,496	0	1	0	11	9	21
194	404304	12,322	1	1	2	3	15	22
196	405106	10,001	0	8	1	9	8	26
197	405295	8,092	0	2	1	8	9	20
198	405296	7,178	0	1	2	5	4	12
208	409309	4,921	0	1	0	7	5	13
209	409315	3,701	1	1	0	4	2	8
211	409956	1,738	0	2	1	8	9	20
219	501227	4,263	1	4	0	5	5	15
221	501231	6,742	1	0	0	12	0	13
222	501238	9,943	0	2	0	2	3	7
223	501243	4,165	0	0	0	2	3	5
225	502155	2,064	0	4	0	1	1	6
226	502156	1,289	0	1	0	0	0	1
228	502216	1,691	0	2	0	11	3	16
232	503026	6,214	0	1	0	9	14	24
236	505225	5,806	0	3	2	2	4	11
238	506047	4,768	0	1	1	15	4	21
239	506048	3,443	0	2	0	3	8	13
240	506109	10,250	0	0	0	8	2	10
241	506239	18,517	0	0	0	2	0	2
242	506248	14,252	0	3	0	4	0	7
245	506252	8,544	0	5	0	6	4	15
247	506256	2,621	0	4	0	6	2	12
249	507055	0,936	0	1	0	0	0	1
250	507059	5,172	0	3	0	11	5	19
252	507224	9,949	0	5	0	15	1	21
254	508247	0,879	0	4	0	4	0	8
258	509222	0,055	0	1	0	0	0	1
260	510144	3,624	0	0	0	5	0	5
262	510219	10,121	0	2	0	1	1	4
264	510244	0,113	0	1	0	0	0	1
266	510289	28,040	1	4	0	8	6	19
270	511221	6,187	0	4	0	6	2	12
272	511229	0,508	0	1	0	2	0	3
273	511234	5,469	1	5	0	8	6	20
274	511240	2,918	0	7	1	4	3	15
278	512062	21,759	0	3	3	10	2	18
288	601356	11,600	0	5	0	13	5	23
289	601357	15,900	0	3	0	10	3	16

Skupina vývodů, zelená barva v Obr. 6-9 vlevo.

Index	Vývod	Délka	ZP	PU	CV	VV	N	Suma poruch
2	101052	20,800	0	7	2	21	7	37
5	101203	14,700	1	3	2	12	7	25
6	101204	9,600	0	6	1	15	5	27
10	102207	56,100	2	12	0	17	10	41
15	103075	24,100	0	4	1	14	8	27
17	103089	9,600	0	10	0	12	5	27
19	103136	25,300	1	10	1	24	7	43
20	103194	15,100	0	5	0	15	7	27
23	103224	42,300	0	9	0	20	8	37
36	201100	9,248	1	6	2	15	7	31
39	201275	19,991	4	7	1	19	6	37
40	201284	11,536	0	1	0	20	10	31
43	202103	17,291	4	10	0	27	5	46
48	203179	35,899	0	16	1	16	6	39
51	203261	23,084	1	11	2	14	4	32
53	204021	10,951	0	2	0	15	12	29
54	204057	48,174	1	7	0	19	7	34
56	205020	26,216	0	9	2	16	8	35
57	205033	12,417	4	11	1	25	5	46
58	205262	26,449	1	14	3	23	3	44
59	205272	25,545	0	1	0	18	25	44
60	205278	33,204	0	2	2	16	7	27
66	207020	23,916	1	8	3	16	13	41
75	209114	34,053	0	6	1	15	4	26
78	209195	43,863	0	8	0	13	9	30
87	210261	18,577	3	10	3	11	8	35
92	211027	4,765	2	8	0	26	9	45
95	211153	11,270	0	2	1	19	5	27
101	212287	9,469	0	6	0	29	6	41
102	212292	11,682	0	14	2	17	11	44
107	301088	20,013	0	7	3	19	10	39
108	301105	27,375	1	8	0	17	6	32
114	302176	10,570	0	14	4	12	24	54
129	303177	5,180	0	10	1	13	4	28
137	306039	15,880	0	8	3	28	11	50
138	307002	4,929	3	22	1	15	16	57
141	307074	5,291	0	17	1	29	8	55
151	309067	6,680	0	12	4	31	11	58
152	309110	5,665	0	15	1	26	14	56
159	312013	13,113	0	11	0	12	4	27
161	312184	6,713	1	14	1	12	10	38

Index	Vývod	Délka	ZP	PU	CV	VV	N	Suma poruch
164	401016	26,712	0	7	2	15	17	41
165	401023	37,514	0	6	0	13	11	30
169	401090	11,012	1	3	0	26	6	36
171	401320	28,220	1	7	1	19	10	38
173	402001	25,332	0	9	0	17	18	44
175	402029	18,042	1	7	1	20	11	40
179	402056	7,318	0	5	0	16	13	34
180	402294	15,235	0	13	1	16	10	40
185	403019	0,100	1	8	0	8	16	33
186	403298	6,172	0	10	3	18	11	42
200	406007	17,269	0	9	2	9	9	29
207	409080	39,309	0	5	0	33	8	46
213	410319	35,032	0	5	2	29	13	49
214	410324	16,950	0	0	0	17	10	27
217	501141	12,095	1	9	1	9	8	28
220	501228	17,218	0	1	0	18	12	31
224	502004	11,534	2	3	0	9	13	27
229	502217	23,070	0	4	3	29	17	53
233	503054	21,703	3	7	0	25	20	55
237	506005	13,965	0	3	1	26	5	35
243	506249	21,650	0	3	0	16	5	24
244	506250	22,607	1	6	0	14	3	24
248	507051	22,880	1	11	0	10	18	40
251	507202	17,603	3	12	1	25	7	48
259	509246	8,272	1	8	1	16	6	32
261	510210	69,332	1	7	0	19	4	31
263	510241	33,853	2	21	0	15	9	47
265	510255	16,025	1	9	0	18	2	30
269	511213	21,610	0	13	0	22	6	41
271	511228	39,795	3	4	0	19	14	40
275	511242	11,428	1	5	1	13	11	31
280	512064	13,511	0	5	0	17	2	24
281	512104	41,970	1	5	1	18	2	27
282	601035	40,400	0	9	1	24	9	43
285	601078	15,800	0	1	1	26	12	40
292	602345	38,500	1	4	0	31	3	39
295	603053	42,700	0	10	2	22	6	40
303	604085	53,900	0	0	0	17	0	17
315	605339	52,000	0	4	0	23	4	31
321	605353	16,000	0	4	2	13	11	30

Skupina vývodů, modrá barva v Obr. 6-9 vlevo.

Index	Vývod	Délka	ZP	PU	CV	VV	N	Suma poruch
3	101055	45,500	1	8	3	29	18	59
8	102008	45,300	2	17	2	46	12	79
11	102208	41,000	1	11	1	34	16	63
12	103005	47,300	0	14	3	27	5	49
16	103079	58,600	2	17	1	35	9	64
26	105006	42,800	0	14	2	45	12	73
30	106006	63,600	1	17	1	37	4	60
38	201102	33,087	4	19	3	49	6	81
41	202070	23,479	0	8	1	33	16	58
46	202276	36,708	3	11	3	41	8	66
47	202283	55,737	1	17	2	50	3	73
50	203209	51,667	2	24	1	43	13	83
77	209116	56,945	2	15	3	35	16	71
82	210022	39,278	3	14	2	38	17	74
90	211011	34,881	2	15	1	60	18	96
91	211022	27,915	0	10	6	52	17	85
96	211273	33,649	1	3	0	58	18	80
99	212102	10,827	2	16	0	37	4	59
103	301013	28,000	2	15	6	33	14	70
104	301067	29,199	0	12	3	30	15	60
105	301074	19,185	2	8	0	32	16	58
106	301079	27,875	1	16	4	27	8	56
111	302159	81,273	0	15	4	22	26	67
113	302174	10,975	1	12	3	38	33	87
116	303003	19,742	0	14	3	43	26	86
117	303018	33,274	1	17	4	38	11	71
118	303054	23,866	0	12	4	31	22	69
120	303099	24,689	0	25	4	41	19	89
139	307013	26,678	0	19	1	43	21	84
144	307121	37,337	2	16	1	32	10	61
150	309060	14,008	2	14	5	35	17	73
155	309170	25,900	0	16	4	36	37	93
168	401083	47,705	0	6	0	40	12	58
176	402040	49,755	1	12	1	34	36	84

Index	Vývod	Délka	ZP	PU	CV	VV	N	Suma poruch
177	402042	29,894	2	10	2	54	30	98
182	403010	38,879	6	18	2	54	19	99
190	404034	33,275	1	11	1	24	17	54
191	404041	47,614	0	13	0	46	18	77
195	405012	23,416	1	8	1	34	26	70
201	406330	35,481	1	6	3	31	14	55
203	408007	32,724	1	25	0	50	24	100
204	408023	43,347	0	7	0	31	22	60
205	409010	42,515	0	6	3	38	13	60
206	409077	31,935	0	11	0	27	17	55
210	409321	28,481	0	8	1	44	22	75
212	410076	30,545	2	4	0	33	18	57
215	501032	33,053	0	6	1	30	16	53
227	502214	38,662	1	13	1	27	22	64
230	503003	19,529	1	9	0	27	22	59
231	503011	40,549	0	9	0	56	17	82
234	503143	61,508	2	14	1	32	11	60
235	504233	44,685	0	7	2	38	14	61
246	506253	25,768	2	15	0	31	10	58
253	508058	17,666	0	17	1	38	29	85
255	509051	39,398	0	11	2	30	11	54
256	509210	29,546	1	13	1	32	20	67
257	509212	14,724	3	13	2	26	17	61
267	511025	50,750	4	23	1	41	11	80
268	511104	39,510	1	20	1	29	10	61
276	512046	37,914	1	12	1	46	23	83
277	512061	19,169	0	5	0	36	17	58
279	512063	51,724	1	13	3	43	7	67
286	601354	20,974	0	16	2	36	9	63
296	603092	56,100	0	6	1	45	7	59
305	604091	57,600	2	4	1	52	6	65
306	604093	43,800	0	4	2	67	1	74
308	604096	47,200	2	6	1	49	9	67
309	604098	44,200	1	4	1	40	3	49

Skupina vývodů, červená barva v Obr. 6-9 vlevo.

Index	Vývod	Délka	ZP	PU	CV	VV	N	Suma poruch
1	101028	86,600	1	25	8	81	39	154
4	101200	41,900	3	13	12	63	21	112
32	106071	42,100	0	31	4	49	24	108
42	202081	73,691	1	19	5	50	16	91
61	206069	57,644	1	34	1	79	34	149
62	206273	58,211	2	6	0	74	13	95
67	207021	57,335	1	14	0	64	22	101
81	210018	25,085	3	26	6	61	19	115
85	210160	57,212	2	25	1	63	29	120
93	211057	10,965	5	6	2	99	29	141
112	302163	22,940	1	17	4	49	54	125
174	402009	110,200	3	22	3	87	50	165
178	402045	66,941	0	15	4	65	56	140
181	403001	24,426	5	14	1	48	52	120
184	403017	45,621	4	25	2	59	66	156

Index	Vývod	Délka	ZP	PU	CV	VV	N	Suma poruch
188	404024	53,498	2	8	1	58	27	96
193	404299	88,033	1	26	4	60	35	126
199	405297	54,440	1	17	0	47	43	108
202	407016	42,948	5	44	8	85	68	210
216	501065	94,792	3	33	5	73	40	154
218	501223	57,396	1	20	0	49	37	107
283	601072	117,800	2	11	2	69	12	96
284	601073	138,600	0	9	2	71	21	103
287	601355	54,300	0	10	1	91	17	119
290	602068	53,500	0	7	3	69	14	93
293	602347	98,100	0	14	13	66	11	104
307	604095	89,400	1	9	9	102	8	129
312	604359	108,500	2	11	2	69	18	102
313	605087	89,100	1	13	4	97	20	135

Příloha B Rozdělení vývodů hierarchickou shlukovací metodou

Metoda shlukování weighted, metrika eukleidovská vzdálenost.

Skupina vývodů, azurová barva v Obr. 6-12 vpravo.

Index	Vývod	Délka	ZP	PU	CV	VV	N	Suma poruch
202	407016	42,948	5	44	8	85	68	210

Skupina vývodů, zelená barva v Obr. 6-12 vpravo.

Index	Vývod	Délka	ZP	PU	CV	VV	N	Suma poruch
93	211057	10,965	5	6	2	99	29	141
112	302163	22,94	1	17	4	49	54	125
181	403001	24,426	5	14	1	48	52	120

Skupina vývodů, černá barva v Obr. 6-12 vpravo.

Index	Vývod	Délka	ZP	PU	CV	VV	N	Suma poruch
1	101028	86,6	1	25	8	81	39	154
61	206069	57,644	1	34	1	79	34	149
174	402009	110,2	3	22	3	87	50	165
178	402045	66,941	0	15	4	65	56	140
184	403017	45,621	4	25	2	59	66	156
193	404299	88,033	1	26	4	60	35	126
216	501065	94,792	3	33	5	73	40	154
283	601072	117,8	2	11	2	69	12	96

Index	Vývod	Délka	ZP	PU	CV	VV	N	Suma poruch
283	601072	117,8	2	11	2	69	12	96
284	601073	138,6	0	9	2	71	21	103
287	601355	54,3	0	10	1	91	17	119
293	602347	98,1	0	14	13	66	11	104
307	604095	89,4	1	9	9	102	8	129
312	604359	108,5	2	11	2	69	18	102
313	605087	89,1	1	13	4	97	20	135

Skupina vývodů, červená barva v Obr. 6-12 vpravo.

Index	Vývod	Délka	ZP	PU	CV	VV	N	Suma poruch
4	101200	41,9	3	13	12	63	21	112
8	102008	45,3	2	17	2	46	12	79
16	103079	58,6	2	17	1	35	9	64
26	105006	42,8	0	14	2	45	12	73
30	106006	63,6	1	17	1	37	4	60
32	106071	42,1	0	31	4	49	24	108
38	201102	33,087	4	19	3	49	6	81
42	202081	73,691	1	19	5	50	16	91
46	202276	36,708	3	11	3	41	8	66
47	202283	55,737	1	17	2	50	3	73
50	203209	51,667	2	24	1	43	13	83
62	206273	58,211	2	6	0	74	13	95
67	207021	57,335	1	14	0	64	22	101
77	209116	56,945	2	15	3	35	16	71
81	210018	25,085	3	26	6	61	19	115
82	210022	39,278	3	14	2	38	17	74
85	210160	57,212	2	25	1	63	29	120
90	211011	34,881	2	15	1	60	18	96
91	211022	27,915	0	10	6	52	17	85
96	211273	33,649	1	3	0	58	18	80
103	301013	28	2	15	6	33	14	70
111	302159	81,273	0	15	4	22	26	67
113	302174	10,975	1	12	3	38	33	87
116	303003	19,742	0	14	3	43	26	86
117	303018	33,274	1	17	4	38	11	71
120	303099	24,689	0	25	4	41	19	89

Index	Vývod	Délka	ZP	PU	CV	VV	N	Suma poruch
139	307013	26,678	0	19	1	43	21	84
155	309170	25,9	0	16	4	36	37	93
168	401083	47,705	0	6	0	40	12	58
176	402040	49,755	1	12	1	34	36	84
177	402042	29,894	2	10	2	54	30	98
182	403010	38,879	6	18	2	54	19	99
188	404024	53,498	2	8	1	58	27	96
191	404041	47,614	0	13	0	46	18	77
199	405297	54,44	1	17	0	47	43	108
203	408007	32,724	1	25	0	50	24	100
205	409010	42,515	0	6	3	38	13	60
210	409321	28,481	0	8	1	44	22	75
218	501223	57,396	1	20	0	49	37	107
231	503011	40,549	0	9	0	56	17	82
234	503143	61,508	2	14	1	32	11	60
235	504233	44,685	0	7	2	38	14	61
253	508058	17,666	0	17	1	38	29	85
267	511025	50,75	4	23	1	41	11	80
276	512046	37,914	1	12	1	46	23	83
279	512063	51,724	1	13	3	43	7	67
290	602068	53,5	0	7	3	69	14	93
296	603092	56,1	0	6	1	45	7	59
305	604091	57,6	2	4	1	52	6	65
306	604093	43,8	0	4	2	67	1	74
308	604096	47,2	2	6	1	49	9	67

Skupina vývodů, modrá barva v Obr. 6-12 vpravo.

Index	Vývod	Délka	ZP	PU	CV	VV	N	Suma poruch
2	101052	20,8	0	7	2	21	7	37
3	101055	45,5	1	8	3	29	18	59
5	101203	14,7	1	3	2	12	7	25
6	101204	9,6	0	6	1	15	5	27
7	101205	12,1	0	9	1	4	1	15
9	102206	22,2	0	5	0	7	4	16
10	102207	56,1	2	12	0	17	10	41
11	102208	41	1	11	1	34	16	63
12	103005	47,3	0	14	3	27	5	49
13	103031	3,6	0	2	0	3	0	5
14	103038	3,9	0	3	0	5	0	8
15	103075	24,1	0	4	1	14	8	27
17	103089	9,6	0	10	0	12	5	27
18	103123	6	1	4	3	2	7	17
19	103136	25,3	1	10	1	24	7	43
20	103194	15,1	0	5	0	15	7	27
21	103196	3,2	0	5	0	0	1	6
22	103201	27,9	0	5	2	2	1	10
23	103224	42,3	0	9	0	20	8	37
24	104207	10	0	0	0	2	1	3
25	104903	5,071	0	4	0	1	0	5
27	105028	16,2	1	3	2	14	2	22
28	105052	10	2	1	0	0	0	3
29	105142	21,5	0	3	0	10	0	13
31	106050	24,2	0	4	0	2	1	7
33	106145	23,9	0	1	0	3	0	4
34	107122	4	0	3	0	3	0	6
35	107188	2,2	0	0	0	1	0	1
36	201100	9,248	1	6	2	15	7	31
37	201101	18,09	0	2	0	10	4	16
39	201275	19,991	4	7	1	19	6	37
40	201284	11,536	0	1	0	20	10	31
41	202070	23,479	0	8	1	33	16	58
43	202103	17,291	4	10	0	27	5	46
44	202271	8,919	0	4	1	10	2	17
45	202275	6,047	0	2	0	11	2	15
48	203179	35,899	0	16	1	16	6	39
49	203192	14,401	0	4	3	3	0	10
51	203261	23,084	1	11	2	14	4	32
52	203270	17,695	1	5	0	7	9	22
53	204021	10,951	0	2	0	15	12	29
54	204057	48,174	1	7	0	19	7	34
55	204277	6,536	0	0	0	2	0	2
56	205020	26,216	0	9	2	16	8	35
57	205033	12,417	4	11	1	25	5	46
58	205262	26,449	1	14	3	23	3	44
59	205272	25,545	0	1	0	18	25	44
60	205278	33,204	0	2	2	16	7	27
63	206280	13,248	0	0	0	2	0	2
64	206288	3,09	0	0	0	7	2	9
65	206290	4,089	0	3	0	1	0	4
66	207020	23,916	1	8	3	16	13	41
68	207066	12,45	0	2	0	4	3	9
69	207258	2,313	0	0	0	0	1	1
92	211027	4,765	2	8	0	26	9	45
94	211151	5,779	1	1	0	4	7	13
95	211153	11,27	0	2	1	19	5	27
97	211289	15	0	0	1	1	0	2
98	212081	13,176	0	11	2	11	1	25
99	212102	10,827	2	16	0	37	4	59
100	212285	2	1	1	0	0	0	2
101	212287	9,469	0	6	0	29	6	41
102	212292	11,682	0	14	2	17	11	44
104	301067	29,199	0	12	3	30	15	60
105	301074	19,185	2	8	0	32	16	58
106	301079	27,875	1	16	4	27	8	56

Index	Vývod	Délka	ZP	PU	CV	VV	N	Suma poruch
159	312013	13,113	0	11	0	12	4	27
160	312014	4,805	1	3	1	7	2	14
161	312184	6,713	1	14	1	12	10	38
162	313187	6,03	0	1	0	0	0	1
163	315175	1,755	0	1	0	2	0	3
164	401016	26,712	0	7	2	15	17	41
165	401023	37,514	0	6	0	13	11	30
166	401043	7,265	2	6	1	4	5	18
167	401044	6,846	0	4	0	7	5	16
169	401090	11,012	1	3	0	26	6	36
170	401318	9,038	1	1	0	11	6	19
171	401320	28,22	1	7	1	19	10	38
172	401331	13,125	1	2	0	13	4	20
173	402001	25,332	0	9	0	17	18	44
175	402029	18,042	1	7	1	20	11	40
179	402056	7,318	0	5	0	16	13	34
180	402294	15,235	0	13	1	16	10	40
183	403016	0,1	0	2	0	2	0	4
185	403019	0,1	1	8	0	8	16	33
186	403298	6,172	0	10	3	18	11	42
187	403302	4,325	0	5	0	5	5	15
189	404030	0,645	1	3	0	0	3	7
190	404034	33,275	1	11	1	24	17	54
192	404107	12,496	0	1	0	11	9	21
194	404304	12,322	1	1	2	3	15	22
195	405012	23,416	1	8	1	34	26	70
196	405106	10,001	0	8	1	9	8	26
197	405295	8,092	0	2	1	8	9	20
198	405296	7,178	0	1	2	5	4	12
200	406007	17,269	0	9	2	9	9	29
201	406330	35,481	1	6	3	31	14	55
204	408023	43,347	0	7	0	31	22	60
206	409077	31,935	0	11	0	27	17	55
207	409080	39,309	0	5	0	33	8	46
208	409309	4,921	0	1	0	7	5	13
209	409315	3,701	1	1	0	4	2	8
211	409956	1,738	0	2	1	8	9	20
212	410076	30,545	2	4	0	33	18	57
213	410319	35,032	0	5	2	29	13	49
214	410324	16,95	0	0	0	17	10	27
215	501032	33,053	0	6	1	30	16	53
217	501141	12,095	1	9	1	9	8	28
219	501227	4,263	1	4	0	5	5	15
220	501228	17,218	0	1	0	18	12	31
221	501231	6,742	1	0	0	12	0	13
222	501238	9,943	0	2	0	2	3	7
223	501243	4,165	0	0	0	2	3	5
224	502004	11,534	2	3	0	9	13	27
225	502155	2,064	0	4	0	1	1	6
226	502156	1,289	0	1	0	0	0	1
227	502214	38,662	1	13	1	27	22	64
228	502216	1,691	0	2	0	11	3	16
229	502217	23,07	0	4	3	29	17	53
230	503003	19,529	1	9	0	27	22	59
250	507059	5,172	0	3	0	11	5	19
251	507202	17,603	3	12	1	25	7	48
252	507224	9,949	0	5	0	15	1	21
254	508247	0,879	0	4	0	4	0	8
255	509051	39,398	0	11	2	30	11	54
256	509210	29,546	1	13	1	32	20	67
257	509212	14,724	3	13	2	26	17	61
258	509222	0,055	0	1	0	0	0	1
259	509246	8,272	1	8	1	16	6	32
260	510144	3,624	0	0	0	5	0	5
261	510210	69,332	1	7	0	19	4	31
262	510219	10,121	0	2	0	1	1	4

Pokračování

Index	Vývod	Délka	ZP	PU	CV	VV	N	Suma poruch
107	301088	20,013	0	7	3	19	10	39
108	301105	27,375	1	8	0	17	6	32
109	302129	3,05	1	2	0	0	1	4
110	302137	2,2	1	3	0	1	0	5
114	302176	10,57	0	14	4	12	24	54
115	302184	5,1	0	3	2	6	8	19
118	303054	23,866	0	12	4	31	22	69
119	303094	6,105	0	7	0	7	3	17
121	303157	3,188	0	6	3	12	7	28
122	303161	4,665	1	8	0	7	0	16
123	303162	2,295	0	1	0	1	0	2
124	303167	8,035	0	10	5	8	5	28
125	303168	1,165	0	2	0	1	0	3
126	303169	1,48	0	1	0	3	0	4
127	303171	3,965	0	3	0	3	3	9
128	303172	1,02	0	1	0	0	0	1
129	303177	5,18	0	10	1	13	4	28
130	303179	8,636	0	7	1	7	3	18
131	303182	3,255	0	3	0	5	0	8
132	304167	7,02	0	1	1	0	0	2
133	304177	4,763	0	1	0	0	0	1
134	305146	4,975	0	8	0	5	0	13
135	305173	2,035	0	0	0	7	0	7
136	305197	0,47	0	0	0	1	0	1
137	306039	15,88	0	8	3	28	11	50
138	307002	4,929	3	22	1	15	16	57
140	307014	7,51	0	6	1	12	1	20
141	307074	5,291	0	17	1	29	8	55
142	307117	1,26	0	4	0	7	2	13
143	307119	2,16	0	0	0	2	0	2
144	307121	37,337	2	16	1	32	10	61
145	307181	5,145	0	2	1	9	2	14
146	308060	18,811	0	6	1	9	3	19
147	308163	14,281	0	1	0	3	0	4
148	308174	10,757	0	5	0	8	0	13
149	308875	0,04	0	1	0	1	0	2
150	309060	14,008	2	14	5	35	17	73
151	309067	6,68	0	12	4	31	11	58
152	309110	5,665	0	15	1	26	14	56
153	309126	1,3	0	5	0	3	5	13
154	309164	3,46	0	0	1	0	1	2
156	309190	11,4	0	0	0	2	0	2
157	311140	2,24	0	1	0	2	3	6
158	311141	0,98	0	2	0	1	1	4

Index	Vývod	Délka	ZP	PU	CV	VV	N	Suma poruch
263	510241	33,853	2	21	0	15	9	47
264	510244	0,113	0	1	0	0	0	1
265	510255	16,025	1	9	0	18	2	30
266	510289	28,04	1	4	0	8	6	19
268	511104	39,51	1	20	1	29	10	61
269	511213	21,61	0	13	0	22	6	41
270	511221	6,187	0	4	0	6	2	12
271	511228	39,795	3	4	0	19	14	40
272	511229	0,508	0	1	0	2	0	3
273	511234	5,469	1	5	0	8	6	20
274	511240	2,918	0	7	1	4	3	15
275	511242	11,428	1	5	1	13	11	31
277	512061	19,169	0	5	0	36	17	58
278	512062	21,759	0	3	3	10	2	18
280	512064	13,511	0	5	0	17	2	24
281	512104	41,97	1	5	1	18	2	27
282	601035	40,4	0	9	1	24	9	43
285	601078	15,8	0	1	1	26	12	40
286	601354	20,974	0	16	2	36	9	63
288	601356	11,6	0	5	0	13	5	23
289	601357	15,9	0	3	0	10	3	16
291	602096	29,6	0	2	0	10	2	14
292	602345	38,5	1	4	0	31	3	39
294	602348	19,9	0	0	0	4	0	4
295	603053	42,7	0	10	2	22	6	40
297	603318	15,2	0	6	1	9	2	18
298	603338	1	0	1	0	2	0	3
299	603360	6,7	0	0	0	5	2	7
300	604036	30	0	1	0	9	2	12
301	604037	9,7	0	1	0	11	0	12
302	604084	8,6	0	5	1	8	4	18
303	604085	53,9	0	0	0	17	0	17
304	604086	30	0	0	0	4	1	5
309	604098	44,2	1	4	1	40	3	49
310	604352	4,1	0	6	1	8	5	20
311	604358	30	0	0	1	1	1	3
314	605097	30	0	2	2	11	2	17
315	605339	52	0	4	0	23	4	31
316	605340	6,8	0	6	0	11	1	18
317	605341	1,5	0	0	0	2	0	2
318	605342	1,1	0	2	0	1	0	3
319	605343	8,6	0	1	0	11	3	15
320	605344	13,1	0	1	1	10	0	12
321	605353	16	0	4	2	13	11	30

Příloha C Základní popisné statistiky skupin vývodů

K-means, 4 shluky vývodů, počáteční středy shluků uspořádány podle vzdálenosti a vybrány pozorování v konstantních intervalech, Statistica

Počet vývodů ve skupině je 143

Znak	Průměr	Medián	Minimum	Maximum	Spodní kvartil	Horní kvartil	Kvantil 10%	Kvantil 90%	Směrodatná odchylka
Délka vývodu v km	8,6726	6	0,04	49,42	3,060	11,975	1,241	20,220	8,2084
Související se započetím provozu	0,1958	0	0,00	2,00	0,000	0,000	0,000	1,000	0,4636
Provoz a údržba	2,7552	2	0,00	12,00	1,000	4,000	0,000	6,000	2,4530
Cizí vlivy	0,4266	0	0,00	5,00	0,000	1,000	0,000	1,200	0,8680
Vnější vlivy	4,9580	4	0,00	15,00	1,250	8,000	0,000	11,000	4,0348
Nezjištěno	2,3007	1	0,00	15,00	0,000	4,000	0,000	6,000	2,8310
Suma poruch za 11 let	10,6360	10	1,00	28,00	4,000	16,750	2,000	20,000	7,2719

Počet vývodů ve skupině je 81

Znak	Průměr	Medián	Minimum	Maximum	Spodní kvartil	Horní kvartil	Kvantil 10%	Kvantil 90%	Směrodatná odchylka
Délka vývodu v km	21,9780	18	0,10	69,33	11,508	27,586	6,700	42,102	13,9470
Související se započetím provozu	0,7037	0	0,00	4,00	0,000	1,000	0,000	2,400	1,0659
Provoz a údržba	7,7160	7	0,00	22,00	4,750	10,000	2,000	14,000	4,4923
Cizí vlivy	0,9506	1	0,00	4,00	0,000	2,000	0,000	3,000	1,0828
Vnější vlivy	18,5560	17	8,00	33,00	15,000	23,000	12,000	27,400	5,8864
Nezjištěno	8,9383	8	0,00	25,00	5,750	11,000	4,000	16,000	4,8460
Suma poruch za 11 let	36,8640	37	17,00	58,00	29,750	43,000	27,000	49,400	9,1320

Počet vývodů ve skupině je 68

Znak	Průměr	Medián	Minimum	Maximum	Spodní kvartil	Horní kvartil	Kvantil 10%	Kvantil 90%	Směrodatná odchylka
Délka vývodu v km	36,7630	36	10,83	81,27	27,276	45,400	19,288	55,991	13,8750
Související se započetím provozu	1,1029	1	0,00	6,00	0,000	2,000	0,000	2,000	1,1988
Provoz a údržba	12,4410	13	3,00	25,00	8,000	16,000	6,000	18,700	5,2615
Cizí vlivy	1,7647	1	0,00	6,00	1,000	3,000	0,000	4,000	1,4874
Vnější vlivy	38,5440	37	22,00	67,00	31,500	44,500	27,000	52,000	9,4497
Nezjištěno	15,6910	16	1,00	37,00	10,000	19,500	6,000	26,000	7,7887
Suma poruch za 11 let	69,5440	67	49,00	100,00	59,000	80,000	55,300	86,700	12,9620

Počet vývodů ve skupině je 29

Znak	Průměr	Medián	Minimum	Maximum	Spodní kvartil	Horní kvartil	Kvantil 10%	Kvantil 90%	Směrodatná odchylka
Délka vývodu v km	66,2510	57	10,97	138,60	44,953	89,175	24,690	109,520	31,0220
Související se započetím provozu	1,7586	1	0,00	5,00	1,000	3,000	0,000	4,600	1,5505
Provoz a údržba	18,0690	15	6,00	44,00	10,750	25,000	7,400	32,200	9,5690
Cizí vlivy	3,6897	3	0,00	13,00	1,000	5,000	0,000	8,600	3,4652
Vnější vlivy	68,8620	66	47,00	102,00	58,750	79,500	49,000	94,600	15,9230
Nezjištěno	30,8620	27	8,00	68,00	17,750	40,750	12,400	55,200	16,9150
Suma poruch za 11 let	123,2400	119	91,00	210,00	102,750	140,250	95,400	155,200	27,1250

Hierarchické shlukování, 5 shluků, metoda shlukování average, Eukleidovská metrika.

Počet vývodů ve skupině je 1

Znak	Průměr	Medián	Minimum	Maximum	Spodní kvartil	Horní kvartil	Kvantil 10%	Kvantil 90%	Směrodatná odchylka
Délka vývodu v km	42,9480	43	42,95	42,95	42,948	42,948	42,948	42,948	0,0000
Související se započetím provozu	5,0000	5	5,00	5,00	5,000	5,000	5,000	5,000	0,0000
Provoz a údržba	44,0000	44	44,00	44,00	44,000	44,000	44,000	44,000	0,0000
Cizí vlivy	8,0000	8	8,00	8,00	8,000	8,000	8,000	8,000	0,0000
Vnější vlivy	85,0000	85	85,00	85,00	85,000	85,000	85,000	85,000	0,0000
Nezjištěno	68,0000	68	68,00	68,00	68,000	68,000	68,000	68,000	0,0000
Suma poruch za 11 let	210,0000	210	210,00	210,00	210,000	210,000	210,000	210,000	0,0000

Počet vývodů ve skupině je 3

Znak	Průměr	Medián	Minimum	Maximum	Spodní kvartil	Horní kvartil	Kvantil 10%	Kvantil 90%	Směrodatná odchylka
Délka vývodu v km	19,4440	23	10,97	24,43	13,959	24,055	10,965	24,426	7,3802
Související se započetím provozu	3,6667	5	1,00	5,00	2,000	5,000	1,000	5,000	2,3094
Provoz a údržba	12,3330	14	6,00	17,00	8,000	16,250	6,000	17,000	5,6862
Cizí vlivy	2,3333	2	1,00	4,00	1,250	3,500	1,000	4,000	1,5275
Vnější vlivy	65,3330	49	48,00	99,00	48,250	86,500	48,000	99,000	29,1600
Nezjištěno	45,0000	52	29,00	54,00	34,750	53,500	29,000	54,000	13,8920
Suma poruch za 11 let	128,6700	125	120,00	141,00	121,250	137,000	120,000	141,000	10,9700

Počet vývodů ve skupině je 14

Znak	Průměr	Medián	Minimum	Maximum	Spodní kvartil	Horní kvartil	Kvantil 10%	Kvantil 90%	Směrodatná odchylka
Délka vývodu v km	88,9740	89	45,62	138,60	66,941	108,500	53,432	119,880	26,0150
Související se započetím provozu	1,3571	1	0,00	4,00	0,000	2,000	0,000	3,100	1,2774
Provoz a údržba	18,3570	15	9,00	34,00	11,000	25,000	9,000	33,100	8,9064
Cizí vlivy	4,2857	4	1,00	13,00	2,000	5,000	1,000	9,400	3,4736
Vnější vlivy	76,3570	72	59,00	102,00	66,000	87,000	59,900	97,500	13,5680
Nezjištěno	30,5000	28	8,00	66,00	17,000	40,000	10,700	57,000	18,0880
Suma poruch za 11 let	130,8600	132	96,00	165,00	104,000	154,000	101,400	156,900	23,2240

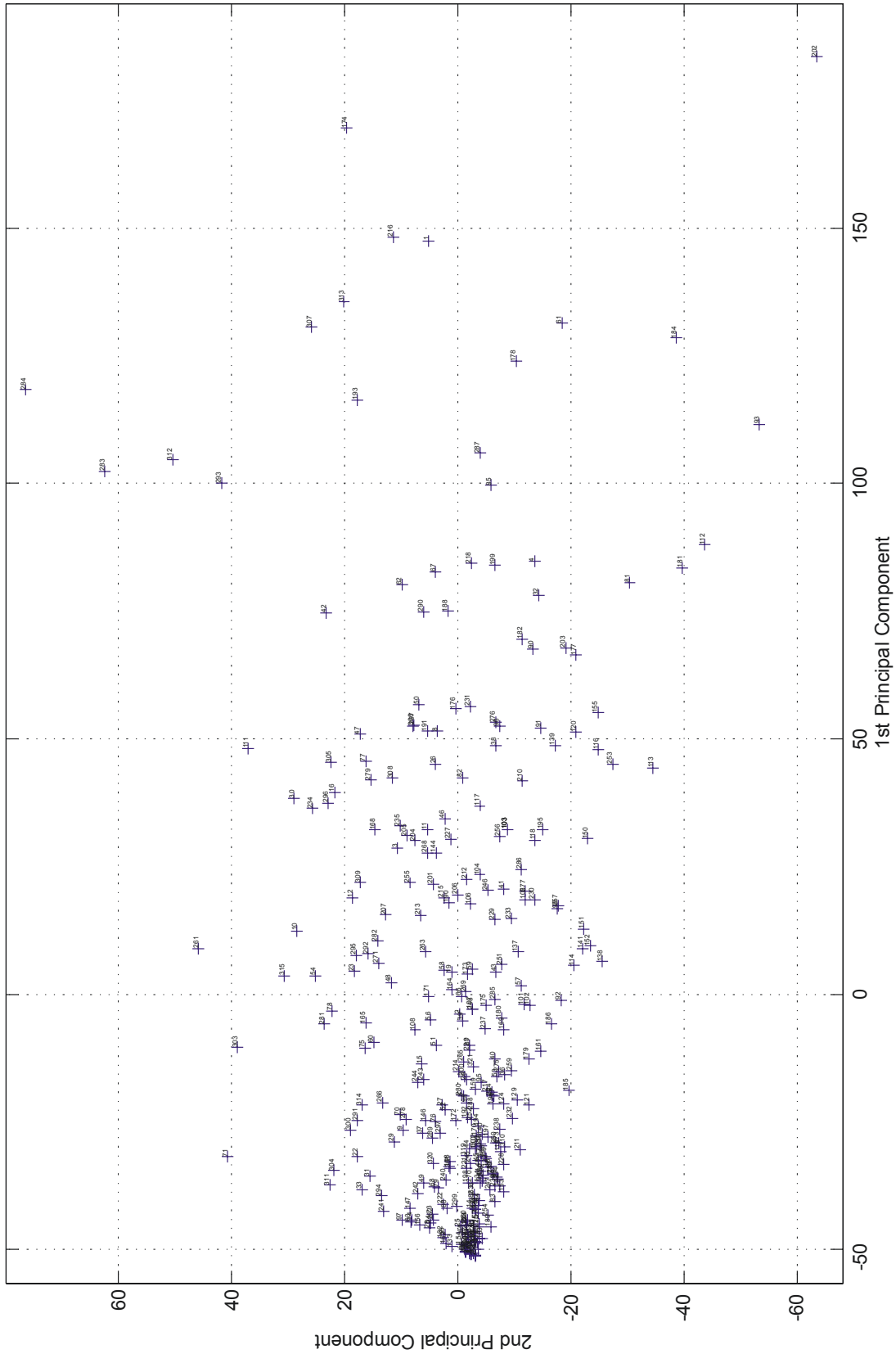
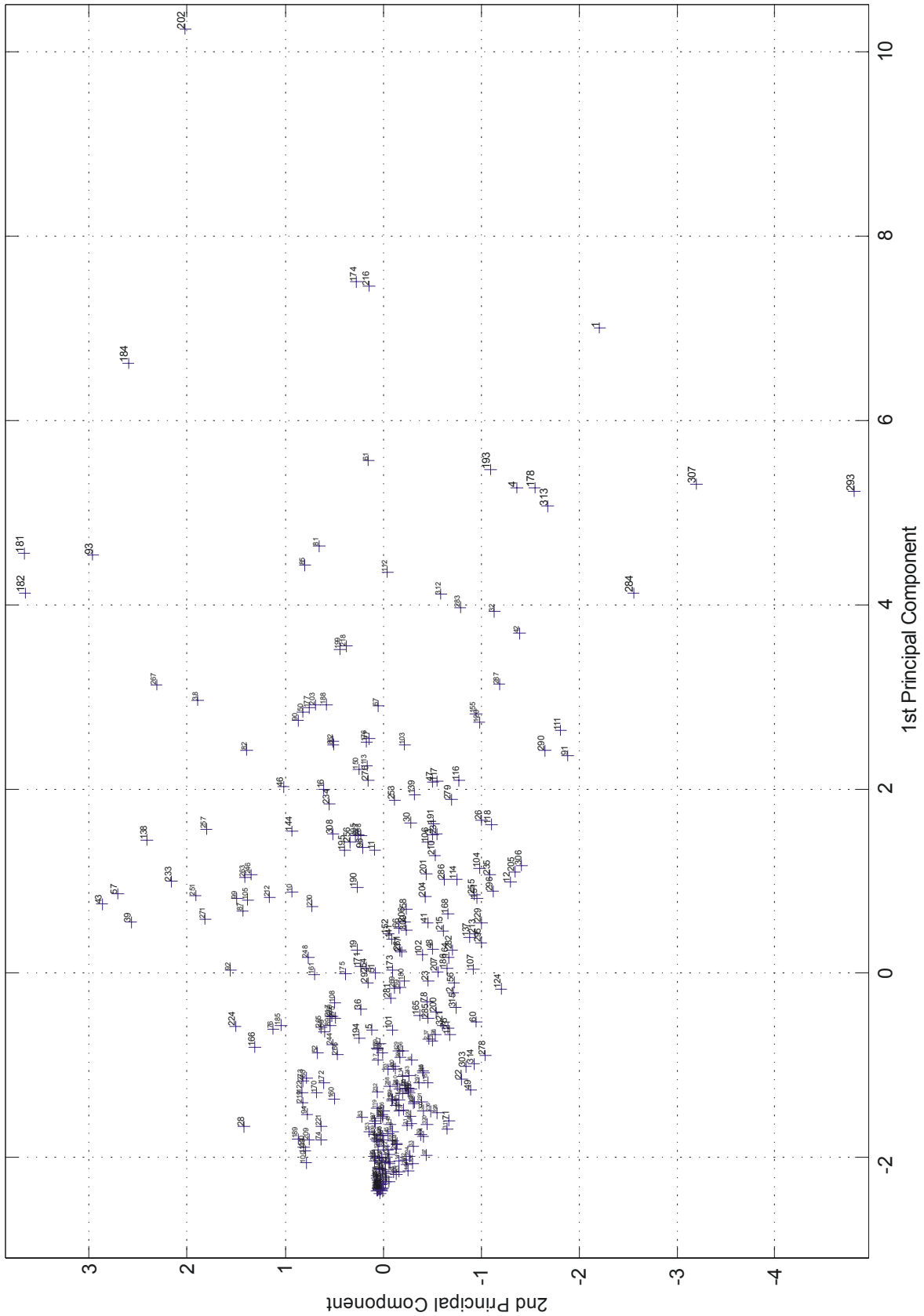
Počet vývodů ve skupině je 51

Znak	Průměr	Medián	Minimum	Maximum	Spodní kvartil	Horní kvartil	Kvantil 10%	Kvantil 90%	Směrodatná odchylka
Délka vývodu v km	43,7670	44	10,98	81,27	33,134	55,413	25,574	58,367	14,4690
Související se započetím provozu	1,2745	1	0,00	6,00	0,000	2,000	0,000	3,000	1,3127
Provoz a údržba	14,2160	14	3,00	31,00	9,250	17,000	6,000	24,400	6,4259
Cizí vlivy	2,1765	2	0,00	12,00	1,000	3,000	0,000	4,400	2,1698
Vnější vlivy	47,0000	46	22,00	74,00	38,000	53,500	35,000	63,000	10,7180
Nezjištěno	17,9220	17	1,00	43,00	11,250	23,750	6,600	31,200	9,5202
Suma poruch za 11 let	82,5880	82	58,00	120,00	70,250	94,500	60,600	107,400	16,0840

Počet vývodů ve skupině je 252

Znak	Průměr	Medián	Minimum	Maximum	Spodní kvartil	Horní kvartil	Kvantil 10%	Kvantil 90%	Směrodatná odchylka
Délka vývodu v km	15,3280	11	0,04	69,33	4,925	22,975	2,000	35,606	13,2350
Související se započetím provozu	0,4405	0	0,00	4,00	0,000	1,000	0,000	1,000	0,8036
Provoz a údržba	5,2619	4	0,00	22,00	1,000	8,000	0,700	12,000	4,5668
Cizí vlivy	0,7103	0	0,00	5,00	0,000	1,000	0,000	2,000	1,0709
Vnější vlivy	12,2340	10	0,00	40,00	3,000	18,000	1,000	29,000	10,2250
Nezjištěno	5,8373	4	0,00	26,00	1,000	9,000	0,000	16,000	5,9293
Suma poruch za 11 let	24,4840	20	1,00	73,00	8,000	39,000	3,000	55,000	18,8330

Příloha D Rozptylový diagram komponentního skóre – data po sloupcovém centrování



Příloha E Rozptylový diagram komponentního skóre – data po Z-score