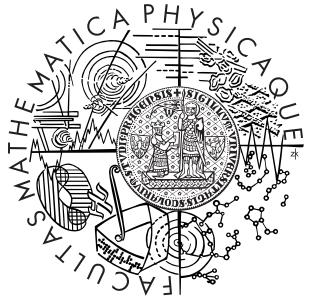
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE



Jan Martínek

Shlukování segmentů pojistných smluv do rizikově homogenních skupin

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: Ing. Pavel Zimmermann

Studijní program: Matematika
Studijní obor: Finanční a pojistná matematika

2010

Poděkování:

Děkuji vedoucímu mé práce Ing. Pavlu Zimmermannovi za odborné vedení a cenné rady při zpracovávání diplomové práce.

Prohlašuji, že jsem svou diplomovou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

V Praze dne 20.7. 2010 Jan Martínek

Obsah

Název práce: Shlukování segmentů pojistných smluv do rizikově homogenních skupin
Autor: Jan Martínek
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: Ing. Pavel Zimmermann
e-mail vedoucího: zimmerp@vse.cz
Abstrakt: V předložené práci studujeme skupiny homogenních pojistných smluv dle definice CEIOPS a pro které počítáme rizikové charakteristiky. První část práce se zabývá kvantifikací těchto rizik. Jako hlavní kategorie studujeme upisovací a rezervové riziko. V druhé části práce tyto skupiny analyzujeme na základě rizikových parametrů a hledáme rizikově homogenní shluky napříč všemi skupinami. Na závěr hledáme charakteristické znaky těchto shluků pro lepší pochopení výsledku shlukové analýzy.
Klíčová slova: Shluková analýza, Upisovací riziko, Riziko rezerv, Modelování škod

Title: Insurance contracts clustering in to risk homogenous groups
Author: Jan Martínek
Department: Department of Probability and Mathematical Statistics
Supervisor: Ing. Pavel Zimmermann
Supervisor's e-mail address: zimmerp@vse.cz
Abstract: In the present work we study classes of homogenous policy contracts by CEIOPS definition and its specified risk characteristics. First part of the thesis study the risk measures and the methods used to measure these risks. As the main risk categories we study underwriting and reserve risk. In the second part of the thesis we analyse these classes by its risk characteristics and cluster them in homogenous groups. At the end we outline the characteristic features of each group for better understanding the result of presented cluster analysis.
Keywords: Cluster analysis, Underwriting Risk, Reserve Risk, Loss modelling

Úvod

V předložené práci analyzujeme skupiny homogenních pojistných smluv dle definice CEIOPS[1] na jejichž úrovni počítáme rizikové charakteristiky. První část práce obsahuje popis a metody, které se používají ke kvantifikaci rizik těchto skupin. Jako hlavní rizikové kategorie studujeme upisovací a riziko rezerv.

Upisovacím rizikem se myslí výše plnění ze škod vzniklých z pojistných smluv upsaných během předcházejícího upisovacího roku (*underwriting year*). V pojišťovnách je kladen vysoký důraz na správný odhad parametrů distribučních funkcí škod, důležitých pro správný výpočet kapitálu a solventnosti. Rizikové charakteristiky jsou následně podrobeny další analýze pro potřeby zajištění, stanovení pojistného, investiční strategii a jako zpětná vazba pro produkt management. Dále se zabýváme metodami odhadu rezervy a rizika rezerv. V druhé části práce studujeme metody vícerozměrné statistiky k určení optimálního počtu shluků na základě rizikových parametrů stanovených pro jednotlivé segmenty pojistných smluv. Na základě stanovených shluků a segmentů pojistných smluv příslušných do těchto jednotlivých shluků se zabýváme měřením rizika v těchto shlucích a stanovení celkového rizikového profilu pojišťovny.

Cílem práce je nalezení vhodného počtu rizikově homogenních shluků, dle předem stanovené kriteriální funkce, pro homogenní segmenty pojistných smuv dle jejich rizikových profilů a následné nalezení charakteristických znaků těchto shluků pro lepší pochopení analyzoveného portfolia.

[1] Committee of European Insurance and Occupational Pensions Supervisors

Kapitola 1

Rizikové charakteristiky

V rámci shlukování pojistného kmene do rizikově homogenních skupin je nutné si ujasnit rizikové parametry, jaké riziko v pojistném kmeni sledujeme a měříme. CEIOPS[1] definuje rizikově homogenní skupinu jako soubor pojistných závazků, které jsou společně řízeny a které mají společné rizikové charakteristiky ve smyslu upisovací politiky, likvidace škod, rizikového profilu pojištenců, vlastnosti produktů (záruk) a struktury nákladů. Rizika v každé skupině by měla být přiměřeně podobná, aby byl možný správný a odpovídající výpočet technických rezerv. S tím je také spojen správný a odpovídající výpočet kapitálu kryjícím solventnost.

Pro účely diplomové práce se zaměříme na základní jednotku pojistného kmene, to jest segmenty pojistných smluv kryjících stejný druh rizika (např. pojištění karavanů, pojištění domácnosti, pojištění proti únosu atd.). Tyto segmenty pojistných smluv považujeme za rizikově homogenní dle definice CEIOPS. Základní rizika na úrovni těchto skupin jsou upisovací riziko (*Underwrting risk*) a riziko nepostačitelnosti rezerv (*Reserve risk*), riziko defaultu zajišťovny, riziko poklesu cen aktiv aj.

V předložené práci se zaměříme na první dvě rizikové kategorie - upisovací riziko a riziko nepostačitelnosti rezerv.

[1] Committee of European Insurance and Occupational Pensions Supervisors

1.1 Upisovací riziko

Upisovací riziko měří riziko ztráty z pojistek upsaných v předchozím upisovacím roce. Tento druh rizika se odhaduje na základě dat z minulých období, klade se vysoký důraz na výběr distribuční funkce škod a odhad jejích parametrů. Jedním z hlavních předpokladů paremeterizace upisovacího rizika je dostatečné množství kvalitních dat, vypovídajících o složení analyzovaného portfolia.

Pro odhad parametrů distribuční funkce rozdělení škod se v praxi využívají zejména následující metody - momentová metoda, metoda maximální věrohodnosti a kvantilová metoda. Při nedostatku dat nezbývá než provést parameterizaci na základě expertního odhadu (tato metoda je ovšem ovlivněna mnoha faktory a nelze ji dobře matematicky popsat).

1) Momentová metoda

Výpočet odhadů metodou maximální věrohodnosti a nebo metodou nejmenších čtverců může být v některých případech komplikovaný a vést k náročným algoritmům. Z těchto důvodů se v praxi používá metoda momentů. Tento postup sice nezaručuje optimalitu odhadu, ale lze jím odvodit odhady parametrů poměrně snadno. Princip momentové metody spočívá v odhadu momentů distribuční funkce *F(x)* empirickými odhady a řešením příslušných rovnic pro analytický výpočet momentů.

2) Metoda maximální věrohodnosti

Předpokládejme, že rozdělení škod $\mathbf{X}$ je absolutně spojité s distribuční funkcí *F(x)* závislou na neznámém vektoru parametrů Θ.

Na základě náhodného výběru $X_1, X_2, ..., X_n$ vytvoříme věrohodnostní funkci

$$l(\Theta) = f(X_1)f(X_2)...f(X_n), \tag{1.1}$$

vezmeme logaritmus věrohodnostní funkce

$$L(\Theta) = \sum_{i=1}^{n} lnf(X_i) \tag{1.2}$$

a hledáme maximum věrohodnostní funkce. Sestavíme soubor rovnic

$$\frac{\delta L}{\delta \theta_i} = 0, \tag{1.3}$$

hledáme řešení pro všechna θ_i, viz [5].

3) Kvantilová metoda

Předpokládejme, že distribuční funkce rozdělení škod je $F(x, \theta)$. Pro p-tý kvantil u_p platí $F(u_p, \theta) = p$, kde $p \epsilon (0, 1)$ je námi zvolené. Nechť $X_{(p)}$ je p-tý výběrový kvantil ve výběru o rozsahu n. Odhad parametru θ spočívá v nahrazení u_p výběrovým kvantilem $X_{(p)}$ a poté řešením rovnice

$$F(X_{(p)}, \theta) = p. \tag{1.4}$$

Pokud je parametr θ k-rozměrný potom využijeme k vhodně zvolených výběrových kvantilů $X_{(p_1)}, X_{(p_2)}, ..., X_{(p_k)}$ a řešíme rovnice

$$F(X_{(p_j)}, \theta) = p_j, \tag{1.5}$$

kde *j=1...k*. Více o jednotlivých metodách odhadu parametrů se lze dočíst v [5].

1.1.1 Dělení škod

Škody zpravidla dělíme na malé, velké a katastrofické. Dělení škod nám umožnuje lepší odhad rizikových parametrů a zároveň je důležité pro následné modelování zajištění v případě, že byl sjednán zajistný program.

Malé škody
Modelujeme jako škody pod určitou předem stanovenou hranicí. Zpravidla se hranice pro malé škody volí menší než priorita prvopojistitele XL-zajištění, případně pokud lze z dat vypozorovat skok mezi jednotlivými škodami, tak se jako hranice volí právě tento skok. Výše uvedený předpoklad nám dovoluje malé škody modelovat v agregaci, jako součet malých škod za určité období. Často se pro modelování malých škod volí distribuční funkce logaritmicko-normálního rozdělení. „*Předpokládejme, že náhodná veličina vzniká jako součin efektů původních příčin, příčiny jsou nezávislé náhodné veličiny, příčin je velké množství. Potom má náhodná veličina logaritmicko-normální*

rozdělení", viz [5]. V praxi se pro modelování malých škod v agregaci ještě používá Gamma rozdělení, jakožto rozdělení součetu exponenciálně rozdělených náhodných veličin.

Věta:
Předpokládejme, že náhodná veličina X_i, $i = 1...n$ je exponenciálně rozdělená se střední hodnotou θ.

$$P[X_1 + X_2 + ...X_n < x] = P[Y < x], \text{ kde } Y = X_1 + X_2 + ... + X_n \quad (1.6)$$

Potom má náhodná veličina $\mathbf{Y}$ gamma rozdělení $Gamma(n, \frac{1}{\theta})$, viz [5].

Velké škody
Velké škody definujeme jako škody nad určitou předem stanovenou hranicí. Tyto škody modelujeme na individuální bázi, proto z dat odhadujeme frekvenci, se kterou jednotlivé škody nastávají a také výši jednotlivých škod. Dále je třeba správně stanovit maximální možnou ztrátu pro pojišťovnu (pojistnou částku). Při modelování pojistného plnění (škody pro pojišťovnu), modelované plnění nesmí přesáhnout tuto částku. Je nutné upozornit, že katastrofické škody (např. povodně, vichřice, hurikán, zemětřesení aj.) jsou vyjmuty a modelovány zvlášť. Pro modelování frekvence vysokých škod se používají distribuční funkce diskrétních náhodných veličin (Binomické, Poissonovo, Negativně binomické rozdělení). Nejpoužívanější a od regulátorů trhu doporučené je použití Negativně binomického rozdělení, jelikož zohledňuje i chybu odhadu parametru λ Poissonova rozdělení, viz [6].

Věta: Předpokládejme, že náhodná veličina $\mathbf{N}$ má Poissonovo rozdělení se střední hodnotou $\theta = \vartheta.\eta$, kde ϑ je konstanta a η je gamma rozdělená náhodná veličina s hustotou

$$Gamma(h, \frac{1}{h}) = \frac{h^h}{\Gamma} y^{h-1} e^{-hy}, y \geq 0, h > 0. \quad (1.7)$$

Potom náhodná veličina $\mathbf{N}$ má negativně binomické rozdělení s pravděpodobnostní funkcí

$$P(N = n) = EP(N = n|\theta) = E\frac{(\vartheta\eta)^n}{n!} e^{-\vartheta\eta} = \frac{\vartheta^n}{n} \frac{h^h}{\Gamma(n)} \int_0^\infty z^{n+h-1} e^{-z} dz =$$

$$\binom{n+h-1}{n} \Big(\frac{h}{\vartheta+h}\Big)^h \Big(\frac{\vartheta}{\vartheta+h}\Big)^n \quad (1.8)$$

$$EN = \vartheta \qquad VarN = \vartheta + \frac{\vartheta^2}{h}$$

Výše jednotlivých škod se většinou modeluje pomocí distribučních funkcí subexponenciálního typu.

Definice: Rozdělení nezáporné náhodné veličiny je subexponenciálního typu, když pro jeho distribuční funkci F(x) platí

$$\lim_{x\to\infty} \frac{1 - F^{n*}(x)}{1 - F(x)} = n, n = 2, 3, ..., \tag{1.9}$$

kde $n*$ je n-tá konvoluce distribuční funkce *F(x)*.

Z hlediska parameterizace upisovacího rizika je důležitý správný výběr distribuční funkce a odhad jejích parametrů. Dále je vhodné zpětné vyhodnocení kvality výběru a odhadu proložením odhadnutou distribuční fukci původními daty. Pro tento účel se nejčastější používá Q-Q[2] graf nebo P-P[3] graf, kde graficky sledujeme odchýlení odhadnuté distribuční funkce od empirické distribuční funkce reálných dat.

Pro ilustraci uvedeme nejčastější distribuční funkce používané pro modelování výše vysokých škod.

(i) Paretovo rozdělení

$$F(x) = 1 - (\frac{\beta}{x})^{\alpha}, x > \beta, \alpha > 0, \beta > 0 \tag{1.10}$$

(ii) Logaritmicko-normální rozdělení

$$F(x) = \phi(\frac{ln(x) - \mu}{\sigma}), \mu \epsilon R, \sigma \epsilon R \tag{1.11}$$

(iii) Zobecněné Paretovo rozdělení

$$F(x) = 1 - (1 + \frac{\epsilon(x - \mu)}{\sigma})^{\frac{-1}{\epsilon}}, x \geq \mu, x \leq \mu - \frac{\sigma}{\epsilon}, \epsilon < 0, \sigma > 0 \tag{1.12}$$

[2]Kvantil - Kvantil

[3]Percentil - Percentil

(iv) Weibullovo rozdělení

$$F(x) = 1 - e^{-(\frac{x}{\lambda})^k}, k > 0, \lambda > 0 \tag{1.13}$$

(v) Burr rozdělení

$$F(x) = 1 - (1 + x^c)^{-k}, c > 0, k > 0 \tag{1.14}$$

Výše uvedené distribuční funkce se používají i pro modelování výše katastrofických škod. Více o jednotlivých typech distribučních funkcí se můžete dočíst v [5].

Katastrofické škody

Katastrofické škody definujeme jako škody při kterých utrpí škodu velké množství pojistných smluv, s periodou návatu větší než předem stanovený počet let. Pro modelování katastrofických škod se zpravidla postupuje po jednotlivých předem definovaných scénářích, při kterých by naše portfolio utrpělo škodu. Pro každý scénář jsou stanoveny různé periody návratu a odhady celkové škody v danných periodách návratu. Klademe si otázku, jak vysoká škoda může nastat jednou za x let. Předpokládáme, že ztráta při nejnižší periodě návratu stanovuje minimální ztrátu pokud katastrofa nastane. Poté odhadujeme parametry distribuční funkce celkové škody.

Příklad: *Scénář povodeň*
Pro modelování frekvence použijeme Poissonovo rozdělení s parametrem λ a pro modelování velikosti škody Paretovo rozdělení. Expertním odhadem vzhledem k profilu našeho portfolia byly odhadnuty následující výše škody pro periodu návratu 20 a 200 let (označíme je jako perioda návratu jedna PN_1, perioda návratu dvě PN_2). Očekáváme, že jednou za 20 let nastane povodeň, která způsobí škodu ve výši 15 mil a jednou za 200 let nastane povodeň, která způsobí škodu za 100 mil.

Tabulka 1.1 Odhad výše škody pro jednotlivé periody návratu

Perioda návratu	Předpokládaná výše zráty
20	15 mil
200	100 mil

Dále stanovíme maximální možné pojistné plnění, jako součet pojistných částek v pojistném kmeni nebo jiným expertním odhadem. V našem případě je stanovena maximální zráta na 150 mil.

Odhad parametrů distribuční funkce celkové škody provedeme kvantilovou metodou

$$1 - \frac{1}{PN_1} = 1 - (\frac{\beta}{Loss_{PN_1}})^{\alpha}, \tag{1.15}$$

$$1 - \frac{1}{PN_2} = 1 - (\frac{\beta}{Loss_{PN_2}})^{\alpha}, \tag{1.16}$$

kde PN_1 je perioda návratu jedna a $Loss_{PN_1}$ je výše škody při periodě návratu jedna. Můžeme to chápat jako, jak velká škoda se stane jednou za PN_1 let. Pro odhad parametru Paretova rozdělení α dostáváme

$$\alpha = \frac{ln\frac{PN_2}{PN_1}}{ln\frac{Loss_{PN_2}}{Loss_{PN_1}}} \tag{1.17}$$

V našem příkladě budeme modelovat celkové pojistné plnění pro scénář *povodeň* s frekvencí danou Poissonovým rozdělením s parametrem $\lambda = \frac{1}{20}$ a výši plnění budeme modelovat Paretovým rozdělením s parametrem $\alpha = 1.21$, β (dolní hranicí) ve výši 15 mil a stropem pojistného plnění ve výši 150 mil (předpokládáme, že pojistné plnění nemůže být vyšší než celkový součet pojistných částek v kmeni pro pojistné smlouvy vystavené riziku povodně).

1.2 Riziko nepostačitelnosti rezerv

Riziko nepostačitelnosti rezerv je jednou z důležitých rizikových charakteristik v pojišťovně, jakožto rizika stanovení správného odhadu výše prostředků pro budoucí plnění. Každý druh pojištění je specifický co se týče upisovací politiky, pojistných podmínek a krycího rizika. V rámci rizika rezerv se jako vhodné dělení často používá dělení na pojištění s dlouhým, nebo krátkým koncem. Tento druh dělení určuje s jakou rychlostí jsou všechny škody z daného upisovacího roku nahlášeny, případně s jakým zpožděním jsou škody dohlašovány.

Rizikem rezervy, jak již bylo naznačeno, je myšlen správný odhad výše

IBNR[4] rezervy. K výpočtu výše IBNR (IBNR + IBNER[5]) se využívá celá řada metod a technik např. *Mack model, Bootstrap, Over-Dispersed-Poisson.* Pro účely diplomové práce popíšeme *Mack* model, který byl využit ve výpočetní sekci ke stanovení rezervového rizika.

1.2.1 Mack model

Máme vývojový trojúhelník $\triangle = \{X_{ij}\}$ kumulativních škod.

Pro účely popisu modelu označíme:
X_{ij} celkovou výši škod, vzniklých v roce i a uhrazených do konce roku $i+j$
c_s je kumulativní vývojový faktor mezi jednotlivými lety odhadovaný vztahem

$$\widehat{c}_s = \frac{\sum_{j=1}^{t-s-1} X_{j,s+1}}{\sum_{j=1}^{t-s-1} X_{j,s}}.$$

Kumulativní trojúhelník škod $\triangle$ doplníme na čtverec počínaje od diagonály násobením odhadnutými vývojovými faktory $\widehat{c}_s$ jako

$$\widehat{X}_{i,j} = X_{i,t-i}\widehat{c}_s \cdots \widehat{c}_{j-1} \; i = 2, ..., t \; j = t-j+1, ..., t-1.$$

Odhad výše rezervy pro rok i je dán jako

$$R_i = X_{i,\infty} - X_{i,t-i} = \widehat{X}_{i,\infty} - X_{i,t-i}. \tag{1.18}$$

Předpokládáme, že vývoj škod je po t-letech ukončen tedy $X_{i,\infty} = X_{i,t}$.

Mackovou metodou dostáváme odhad střední kvadratické chyby rezervy pro rok i jako

$$\begin{aligned} mse\widehat{R}_i &= \widehat{X}_{i\infty}^2 \sum_{j=t-i}^{t-2} \frac{\widehat{\sigma}_j^2}{\widehat{c}_j^2}\left(\frac{1}{\widehat{X}_{ij}} + \frac{1}{\sum_{s=1}^{t-j-1} X_{ij}}\right) = \\ &= \widehat{X}_{i\infty}^2 \sum_{j=t-i}^{t-2} \frac{\widehat{\sigma}_j^2}{\widehat{c}_j^2}\frac{1}{\widehat{X}_{ij}} + \widehat{X}_{i\infty}^2 \sum_{j=t-i}^{t-2} \frac{\widehat{\sigma}_j^2}{\widehat{c}_j^2}\frac{1}{\sum_{s=1}^{t-j-1} X_{ij}}, \end{aligned} \tag{1.19}$$

kde $\widehat{\sigma}^2$ je

$$\widehat{\sigma}^2 = \frac{1}{t-j-1} \sum_{s=1}^{t-j-1} X_{s,j}\left(\frac{X_{s,j+1}}{X_{s,j}} - \widehat{c}_j\right)^2. \tag{1.20}$$

[4]Incurred But Not yet Reported; Škody, které se již staly a ještě nebyly nahlášeny.

[5]Incurred But Not Enough Reported; Škody, které nebyly dostatečně nahlášeny a tudíž se očekává změna jejich výše.

První sčítanec v (1.19) představuje riziko procesu a druhý sčítanec riziko odhadu parametrů. Více se o Mackově modelu můžete dočíst v [7].

Riziko odhadu parametrů

$$ParameterRisk_{it} = \widehat{X}_{it}^2 \sum_{j=t+1-i}^{t-1} \frac{\widehat{\sigma}_j^2}{\widehat{c}_j^2} \frac{1}{\sum_{s=1}^{t-j} X_{sj}} \tag{1.21}$$

pro $i = 2, ..., t$.

Rekurzivní formule rizika odhadu parametrů uvedená v [11], je dána jako

$\text{ParameterRisk}_{ik} = \widehat{X}_{ik}^2 \sum_{j=t+1-i}^{k-1} \frac{\widehat{\sigma}_j^2}{\widehat{c}_j^2} \frac{1}{\sum_{s=1}^{t-j} X_{sj}} = \widehat{c}_{k-1}^2 \widehat{X}_{i,k-1}^2 (\sum_{j=t+1-i}^{k-2} \frac{\widehat{\sigma}_j^2}{\widehat{c}_j^2} \frac{1}{\sum_{s=1}^{t-j} X_{sj}} + \frac{\widehat{\sigma}_{k-1}^2}{\widehat{c}_{k-1}^2} \frac{1}{\sum_{s=1}^{t-k-1} X_{s,k-1}}) = \widehat{c}_{k-1}^2 \widehat{X}_{i,k-1}^2 \sum_{j=t+1-i}^{k-2} \frac{\widehat{\sigma}_j^2}{\widehat{c}_j^2} \frac{1}{\sum_{s=1}^{t-j} X_{sj}} + \widehat{X}_{i,k-1}^2 \frac{\widehat{\sigma}_{k-1}^2}{\sum_{s=1}^{t-k-1} X_{s,k-1}} = \widehat{c}_{k-1}^2 ParameterRisk_{i,k-1} + \widehat{X}_{i,k-1}^2 Var(\widehat{c}_{k-1})$

a dostáváme

$$ParameterRisk_{ik} = \begin{cases} \widehat{c}_{k-1}^{\,2} ParameterRisk_{i,k-1} + \widehat{X}_{i,k-1} \widehat{Var(c_{k-1})} + \\ + \widehat{Var(c_{k-1})} ParameterRisk_{i,k-1} & \text{pro } k > t+2-i \\ X_{i,t+1-i} \widehat{Var(c_{k-1})} & \text{pro } k = t+2-i \end{cases}$$

Riziko procesu

$$ProcesRisk_{it} = \widehat{X}_{it}^2 \sum_{j=t+1-i}^{t-1} \frac{\widehat{\sigma}_j^2}{\widehat{c}_j^2} \frac{1}{\widehat{X}_{ij}} \tag{1.22}$$

pro $i = 2, ..., t$.

Uvedeme ještě rekurzivní formuly výpočtu rizika procesu

$$ProcessRisk_{ik} = \begin{cases} \widehat{c}_{k-1}^2 ProcessRisk_{i,k-1} + \widehat{X}_{i,k-1} \widehat{\sigma}_{k-1}^2 & \text{pro } k > t+2-i \\ X_{i,t+1-i} \widehat{\sigma}_{k-1}^2 & \text{pro } k = t+2-i \end{cases}$$

jak je uvedeno v [11].

Celková rezerva

Celková rezerva je dána jako součet odhadů rezerv pro jednotlivé roky, tedy $\widehat{R} = \widehat{R}_2 + ... + \widehat{R}_t$ a odhad roptylu celkové rezervy je

$$\widehat{mseR} = \sum_{i=2}^{t}(mse\widehat{R}_i + \widehat{X}_{i\infty}(\sum_{j=i+1}^{t}\widehat{X}_{j\infty})\sum_{k=t-i}^{t-2}\frac{2\widehat{\sigma}_k^2/\widehat{c}_k^2}{\sum_{s=1}^{t-k-1}X_{sk}}, \tag{1.23}$$

viz [7].

1.2.2 Modelování výše škody a rezervy

Modely stanovení výše rezervy stanoví $R_2, ..., R_t$ rezervu v jednotlivých letech a celkovou rezervu R. Dále dostáváme $X_{2,t}...X_{t,t}$ celkovou kumulativní výši škod a $VarX_{2,t}...VarX_{t,t}$ rozptyl celkové kumulativní výše škod v jednotlivých letech.

K modelování výše škody v daném roce j pro $j = 2, ..., t$ použijeme logaritmicko-normální rozdělení. V roce j je střední hodnota log-normálního rozdělení $EX_j = X_{j,t}$ a rozptyl $VarX_j = VarX_{j,t}$, případně $CoV_j = \frac{\sqrt{VarX_j}}{EX_j}$. Z odhadu parametrů momentovou metodou pro $LN(\mu, \sigma)$ dostáváme

$$\mu_j = ln(EX) - \frac{1}{2}ln(1 + \frac{Var(x)}{(EX)^2}), \tag{1.24}$$

$$\sigma_j = \sqrt{ln(1 + \frac{Var(x)}{(EX)^2})}. \tag{1.25}$$

Celkovou kumulativní výši škody $X_{j,t}$ v danném roce j modelujeme pomocí logaritmicko-normálního rozdělení $LN(\mu_j, \sigma_j)$ a výše rezervy v roce j je dána jako

$$R_j = X_{j,t} - X_{j,t-j}, \tag{1.26}$$

kde $X_{j,t-j}$ je diagonální hodnota vývojového trojúhelníhu příslušného roku j.

Další možností je modelovat přímo rezervu pomocí logaritmicko-normálního rozdělení, jelikož modely stanovení výše rezervy nám dávají rozpty rezervy

v jednotlivých letech dle (1.19). Odhad parametrů logaritmicko-normálního rozdělení je analogický jako v předchozím případě viz. (1.23) (1.24). Celková ztáta v roce j je pak dána jako

$$X_{j,t} = R_j + X_{j,t-j}, \tag{1.27}$$

kde $X_{j,t-j}$ je diagonální hodnota vývojového trojúhelníhu příslušného roku j.

V praxi se používají oba postupy, nelze přesně určit, který z těchto postupů je lepší (modelovat celkovou ztrátu nebo rezervu) i když se více upřednostňuje postup modelování rezervy, jelikož při modelování celkové ztráty může v určitých simulacích vyjít záporná rezerva.

Riziko nepostačitelnosti rezerv můžeme vyjádřit jako koeficient variace definovaný jako $\frac{\sqrt{VarX_j}}{EX_j}$, v druhém případě $\frac{\sqrt{VarR_j}}{ER_j}$.

Výpočet kapitálu

Každá pojišťovna musí mít dostatečný kapitál, který i v málo pravděpodobné situaci zajistí převoditelnost aktiv a závazků na třetí stranu. Od regulátorů trhu je pak stanovena hladina spolehlivosti na které je každá pojišťovna nucena držet vlastní kapitál. Ve Velké Británii je od FSA[6] stanovena hladina významnosti na úroveň 99.5 %.

Požadovaný kapitál na hladině spolehlivosti 99.5% je následně odhadnut jako rozdíl 99.5 % kvantilu rozdělení rezervy a nejlepšího odhadu celkové rezervy, tedy

$$ReserveCapital_{99,5} = R^{\{99.5\}} - ER. \tag{1.28}$$

[6]Financial Services Authority - regulátor trhu ve Velké Británii

Kapitola 2

Shluková analýza

Cílem shlukové analýzy je nalézt skupiny podobných objektů dle předem stanovených kriérií. Uplatnění shlukové analýzy je důležité zejména tam, kde můžeme vypozorovat společné charakteristiky pro různé objekty a tudíž je seskupovat do shluků. Následně je důležité najít vhodnou interpretaci pro vzniklé shluky (najít vhodnou charakteristiku daných shluků). Je nutné poznamenat, že občas se nám může stát, že potřebujeme shlukovat proměnné. V případě, že popis třídy je zastoupen vektorem proměnných, je vhodné předem udělat analýzu těchto proměnných, pokud je mezi skupinou proměnných silný korelační vztah, pak je dobré tuto skupinu nahradit pouze jedním parametrem a snížit tím rozměr úlohy.

Máme k dispozici datovou matici $\mathbf{X}$ typu n x p kde n je počet objektů a p je počet proměnných charakterizujících objekt. Naším úkolem je najít mezi všemi rozklady *S(k)* (rozklad možných n objektů do k shluků) ten nejlepší. Spolehlivou metodou nalezení nejlepšího rozkladu by bylo projít všechny možnosti a poté vybrat mezi všemi rozklady ten nejlepší. Tento přístup je možné aplikovat na malý počet objektů, ale v praxi máme většinou velký počet objektů a aplikovat metodu všech rozkladů by nebylo v našich silách. Počet způsobů, kterými lze n objektů rozdělit do k shluků udává Stirlingovo číslu 2. druhu, jak je uvedeno v [4]

$$S^{(k)} = \frac{1}{k!}\sum_{h=0}^{k}(-1)^{k-1}\binom{k}{h}h^n \tag{2.1}$$

a počet všech rozkladů je dán jako

$$S = \sum_{k=1}^{n} S^{(k)}. \tag{2.2}$$

Proto se v praxi používají různé algoritmy pro nalezení aspoň jednoho lokálního extrému kriteriální funkce rozkladu. Naším cílem je najít takový rozklad, kde uvnitř shluku jsou si objekty co nejvíce podobné a objekty z různých shluků jsou si podobné co nejméně. Existují různé míry podobností. Většinou požadujeme, aby nabývaly hodnot od 0 pro maximální rozdílnost do 1 pro totožnost objektů.

2.1 Kvalita rozkladu a míry vzdáleností

Jednou z důležitých otázek je, do jaké míry je použitý algoritmus shlukové analýzy kvalitní. Shluková analýza nám rozdělila soubor dat do k shluků a v každém objektu je n_k objektů. Pro tento účel se používají matice vnitroshlukové variability

$$E = \sum_{h=1}^{k} \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)(x_{hi} - \bar{x}_h)^T \tag{2.3}$$

a matice mezishlukové variability

$$B = \sum_{h=1}^{k} n_h (\bar{x}_h - \bar{x})(\bar{x}_h - \bar{x})^T. \tag{2.4}$$

Součet těchto matic nám dává matici celkové variability

$$T = \sum_{h=1}^{k} \sum_{i=1}^{n_h} (x_{hi} - \bar{x})(x_{hi} - \bar{x})^T, \tag{2.5}$$

kde x_{hi} je vektor pozorování u i-tého objektu v h-tém shluku, $\bar{x}_h$ je vektor průměrů pro h-tý shluk a $\bar{x}$ je vektor průměrů pro celý soubor dat, viz [4].

Maxima vzdálenosti mezi kompaktními shluky dosáhneme minimalizací Wardova kriteria *G1*

$$G1 = st\mathbf{E} = \sum_{h=1}^{k} \sum_{i=1}^{n_h} \sum_{j=1}^{p} (x_{hij} - \bar{x}_{hj})^2, \tag{2.6}$$

kde p je počet proměnných charakterizujících objekt, viz [4]. Chceme dosáhnout minima celkového součtu čtverců odchylek všech hodnot od příslušných shlukových průměrů. „*Kromě Wardova kritéria lze použít i výše uvedené monotóní funkce, jelikož matice* $\mathbf{T}$ *je pro všechny rozklady stejná, znamená minimalizace stopového kritéria* $\mathbf{E}$ *totéž co maximalizace stopového kritéria* $\mathbf{B}$. *Pokud chceme dosáhnout nezávislosti na užitých měřitelných jednotkách, lze použít minimalizaci determinantu matice vnitroshlukové variability* $|\mathbf{E}|$, viz [4]."

Míry vzdáleností

Po provedení výběru proměnných charaterizujících vlastnosti objektu, musíme rozhodnout o způsobu hodnocení vzdáleností nebo podobností objektů. Jako první se provádí výpočet příslušných měr pro všechny páry objektů. Výsledkem je symetrická čtvercová matice n x n, s nulami na diagonále, pokud se jedná o matici měr vzdáleností, nebo s jedničkami na diagonále pokud se jedná o matici podobnosti.

K měření vzdáleností mezi objekty charakterizovaných hodnotami kvantitativních proměnných můžeme použít míry vzdáleností uvedené v tabulce 2.1 .

Tabulka 2.1. Míry vzdáleností

Název míry	Vzorec	Poznámka
Hemmingova vzdálenost	$D_H(\mathbf{x}_i, \mathbf{x}'_i) = \sum_{j=1}^{p} \lvert x_{ij} - x_{i'j} \rvert$	
Euklidovská vzdálenost	$D_E(\mathbf{x}_i, \mathbf{x}'_i) = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{i'j})^2}$	
Čebyševova vzdálenost	$\mathrm{D}_C(\mathbf{x}_i, \mathbf{x}'_i) = \max_j \lvert x_{ij} - x_{i'j} \rvert$	
Minkowského metrika	$\mathrm{L}_{ii'}^{(m)} = \sqrt[m]{\sum_{j=1}^{p} \lvert x_{ij} - x_{i'j} \rvert^m}$	$\mathrm{D}_C = \lim_{m\to\infty} L^{(m)}$

Častou nevýhodou uvedených měr v tabulce 2.1 je značná závislost na měřících jednotkách, která někdy brání smysluplnému pořízení součtu pro různé proměnné. Další nevýhodou je silná korelovanost některých proměnných, jejichž součet pak má signifikantní vliv na konečný výsledek. Jednou z možností jak se vyhnout nežádoucímu vlivu vysokých hodnot některých proměnných je jejich transformace. Vliv měřících jednotek se dá odstranit tak, že

hodnoty proměnných normujeme.

Další možností odstranění vlivu měřících jednotek je přisouzení vah jednotlivým proměnným. Uvažujeme vektor koeficientů (vah) c_j, $j = 1...n$, přisuzujících každé proměnné váhu. Uvažujeme diagonální matici $\mathbf{C}$ s koeficienty c_j na diagonále, potom výslednou transformaci můžeme vyjádřit jako $u_i = C^T x_i$, i= 1,2,3,...,n.

K odstranění vlivu měřících jednotek lze využít inverzní korelační matice jako matici vah, tedy $\mathbf{C}\mathbf{C}^T = \mathbf{S}^{-1}$. Tímto odstraníme kromě závislosti na měřících jednotkách také nevýhodu nadměrného vlivu korelovaných proměnných a dostáváme *Mahalanobisovu vzdálenost*, jak je uvedeno v [4].

$$D_E^2(\mathbf{u}_i, \mathbf{u}_{i'}) = (\mathbf{u}_i - \mathbf{u}_{i'})^T \mathbf{S}^{-1} (\mathbf{u}_i - \mathbf{u}_{i'}). \tag{2.7}$$

Pro úplnost ještě uvedeme míru pro kvantitativní data známou jako *Lanceyova-Williamsova vzdálenost*, viz [4]

$$D_{LW}(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_{j=1}^{p} \frac{|x_{ij} - x_{i'j}|}{|x_{ij}| + |x_{i'j}|}. \tag{2.8}$$

Z používaných měr podobností uvedeme *Jaccardův koeficient*, viz [4]

$$A_J(\mathbf{x}_i, \mathbf{x}_{i'}) = \frac{\sum_{j=1}^{p} x_{ij} x_{i'j}}{\sum_{j=1}^{p} x_{ij}^2 + \sum_{j=1}^{p} x_{i'j}^2 - \sum_{j=1}^{p} x_{ij} x_{i'j}} \tag{2.9}$$

Míry pro alternativní data

Pro určení prvku matice vzdálenosti nebo podobnosti pro *i*-tý a *i'*-tý objekt zjišťujeme shodu či neshodu hodnot u p alternativních proměnných. Protože jsou přípustné pouze dvě hodnoty 0 a 1, dostáváme následující tabulku četností:

a u obou objektů shodně 1
b u *i*-tého objektu 0, u *i'*-tého objektu 1
c u *i*-tého objektu 1, u *i'*-tého objektu 0
d u obou objektů shodně 0

Platí $a+b+c+d=p$

Z měr podobnosti pro alternativní data uvedeme *Sokalův-Michenerův* koeficient asociace, čili koeficient prosté shody jako

$$A_{SM}(\mathbf{x}_i, \mathbf{x}_{i'}) = \frac{a+d}{p}, \tag{2.10}$$

viz [9].

Podíl shodně pozitivních výsledků *Russelův Raovův-koeficient* definovaný jako

$$A_{SM}(\mathbf{x}_i, \mathbf{x}_{i'}) = \frac{a}{p}, \tag{2.11}$$

viz [9].

Podíl shodných výsledků při vyloučení shodně negativních výsledků. *Jaccardův koeficient* pro alternativní data definovaný jako (pro kvantitativní data je definován (2.9)

$$A_J(\mathbf{x}_i, \mathbf{x}_{i'}) = \frac{a}{p+b+c}. \tag{2.12}$$

Pro alternativní data definujeme korelační koeficient jak je uveden v [9]

$$A_r(\mathbf{x}_i, \mathbf{x}_{i'}) = r_{ii'} = \frac{ad+bc}{\sqrt{(a+b)+(c+d)+(a+c)+(b+d)}}. \tag{2.13}$$

Rogersův a Tanimotův koeficient asociace

$$A_r(\mathbf{x}_i, \mathbf{x}_{i'}) = r_{ii'} = \frac{a+d}{a+2b+2c+d}, \tag{2.14}$$

a Sörensenův koeficient asociace

$$A_r(\mathbf{x}_i, \mathbf{x}_{i'}) = r_{ii'} = \frac{2a}{2a+b+c}. \tag{2.15}$$

viz [9].
V některých připadech se pracuje se smíšenými daty (jisté charakteristické znaky jsou nominální a některé znaky jsou kvantitativní)

Growerův koeficient podobnosti, jak je uveden v [9], je definován následovně

$$S_{G_{ij}} = \frac{\sum_{k=1}^{m} w_{ijk} S_{ijk}}{\sum_{k=1}^{m} w_{ijk}}, \tag{2.16}$$

kde *i, j* jsou indexy dvou objektů charakterizovaných proměnnými *k=1,...,m.* Váha proměnné w_{ijk} může nabýt hodnoty 1 nebo 0 podle toho, zda lze srovnávat hodnoty proměnné k u objektů i a j, a S_{ijk} označuje hodnotu pro k-tou proměnnou. Pro všechny měření uvažujeme jen případy, kdy jsou hodnoty proměnné x_{ik} a x_{ij} známé. Pro nominální data se provádí libovolné číslování začínající od 1, tedy w_{ijk}=1. Pokud nejsou hodnoty nějaké proměnné známé, je automaticky w_{ijk}=0. Platí, že

1. Pro nominální proměnné je S_{ijk}=0 pro $x_{ik} \neq x_{jk}$ a S_{ijk}=1 pro $x_{ik} = x_{jk}$

2. Pro alternativní proměnné je $S_{ijk} = w_{ijk} = 1$ pro $x_{ik} = x_{jk} = 1$ resp. $x_{ik} = x_{jk} = 0$. Pokud je $x_{ik} \neq x_{jk}$, je $S_{ijk} = 0$ a $w_{ijk} = 1$.
V případě, že $x_{ik} = x_{jk} = 0$ a negativní shoda nemá smysl, je $S_{ijk} = \mathrm{w}_{ijk} = 0$.

3. Pro kvantitativní data počítáme

$$S_{ijk} = 1 - |x_{ik} - x_{jk}|R_k, \tag{2.17}$$

kde R_k je rozdíl mezi maximální a minimální hodnotou proměnné u všech objektů jak je uvedeno v[9]

2.2 Hierarchické shlukování

Jedním z nejpoužívanějších postupů uplatňovaných v rámci shlukové analýzy patří vytváření hierarchické posloupnosti rozkladu. Jednou z možností hierarchického shlukování je aglomerativní hierarchický postup, který provádíme v následujících krocích:

1. Vypočítáme matici měr vzdáleností **D**

2. Začneme proces rozkladu $S^{(n)}$. Vytvoříme n shluků a do každého shluku přiřadíme jeden objekt.

3. Najdeme dva shluky (k-tý a i-tý), jejichž vzdálenost je minimální. Následně spojíme tyto dva shluky v nový shluk h.

4. V matici odstraníme tyto dva shluky a nahradíme je novým shlukem h.

5. Poznamenáme pořadí cyklu v=1,2,...,n-1, identifikaci spojených objektů k, i a hladinu spojení $d_1 = D_{ki}$.

6. Pokud nejsou všechny objekty v jednom shluku $S^{(1)}$, pokračujeme krokem **3**.

Méně užívaný je opačný, divizivní hierarchický postup, kdy vychazíme z jediného shluku $S^{(1)}$ a v každém kroku rozštěpíme nějaký shluk na dva. Proces končí rozdělením všech n objektů do n shluků.

Při dané volbě proměnných a dané matici vzdáleností se budou výsledky lišit dle metody výpočtu vzdáleností. Jednou ze základních metod aglomerativních postupů je metoda nejbližšího souseda, kde ve *3*-tím kroku nahradíme *k*-tý a *i*-tý shluk novým shlukem g a do matice vzdáleností zapíšeme vzdálenost shluku g od ostatních shluků. V *m*-tém cyklu zapíšeme celkem *n-m-1* vzdáleností dle

$$D_{gg'} = min(D_{g'h}, D_{g'h'}). \tag{2.18}$$

Uvedený postup lze využít k vytvoření hierarchické posloupnosti rozkladů a následné sestrojení dendrogramu.

2.2.1 Přehled aglomerativních postupů

Pro hierarchické shlukování je nutné definovat aglomerativní postup. V této sekci uvedeme přehled těchto postupů:

I. Metoda nejbližšího souseda

Tato metoda byla popsána na konci předchozího odstavce. Kritériem pro spojování shluků je minimum ze všech mezishlukových vzdáleností objektů. Nevýhodou této metody je, že i značně vzdálené objekty se mohou sejít v jednom shluku, pokud dostatečný počet objektů/shluků mezi těmito objekty vytvoří pomyslný most.

II. Metoda nejvzdálenějšího souseda

Jde o podobnou metodu jako metoda nejbližšího souseda, kromě toho, že kritérium je postaveno na maximální vzdálenosti ze všech možných mezishlukových vzdáleností objektů.

$$D_{gg'} = max(D_{g'h}, D_{g'h'}). \tag{2.19}$$

Nejdelší vzdálenost mezi objekty v každém shluku představuje nejmenší pomyslnou kouli (pro tří dimenzionální prostor), která obklopuje všechny objekty v jednotlivých shlucích. Této metodě se také říká *metoda úplného propojení*, protože všechny objekty ve shluku jsou propojeny každý s každým při maximální vzdálenosti, čili minimální podobnosti.

III. Metoda průměrné vzdálenosti

Kritériem vzniku shluků je průměrná vzálenost všech objektů v jednom shluku ke všem objektům ve druhém shluku. Při přepočtu matice vzdáleností použijeme

$$D_{gg'} = \frac{n_h D_{g'h} + n_{h'} D_{g'h'}}{n_{h'} + n_h}. \tag{2.20}$$

Tato metoda často vede ke stejným výsledkům jako metoda nejbližšího souseda, viz [4].

IV. Metoda těžiště

V této metodě jde o vzdálenost dvou těžišť shluků vyjádřených euklidovskou vzdáleností nebo čtvercem euklidovské vzdálenosti. Těžiště shluků má souřadnice odpovídající průměrným hodnotám objektů pro jednotlivé proměnné, jak je uvedeno v [9]. Tedy

$$D_E(\bar{x}_h, \bar{x}_{h'}) = \sum_{j=1}^{p} (\bar{x}_{hj} - \bar{x}_{hj'})^2 \tag{2.21}$$

a po každém kroku shlukování provedeme přepočet těžišť, dle

$$D_{g'g} = \frac{1}{n_h + n_{h'}} (n_h D_{g'h} + n_{h'} D_{g'h'} - \frac{n_h n_{h'}}{n_h + n_{h'}} D_{hh'}), \tag{2.22}$$

, viz [4].

V. Wardova metoda

Principem této metody není optimalizace vzdáleností mezi shluky, ale maximalizace heterogenity shluků dle Wardova kritéria (2.6), tedy podle kritéria minima přírůstku vnitroskupinového součtu čtverců odchylek objektů od těžiště shluků. V každém kroku se pro všechny dvojice odchylek spočítá přírůstek součtu čtverců odchylek, vzniklý jejich sloučením a pak se spojí ty shluky, kterým odpovídá minimální hodnota tohoto přírůstku. Předpokládejme, že máme m objektů, které jsou charakterizovány p proměnnými. Tudíž máme k dispozici matici m x p s prvky x_{ij}. Vnitroshluková variabilita je pak dána jako

$$E = \sum_{j=1}^{p} \sum_{i=1}^{m} (x_{ij} - \widehat{x}_j)^2 \tag{2.23}$$

a přírůstek celkového vnitroskupinového součtu čtverců odhylek je dán jako (viz. [9])

$$\Delta G1 = \sum_{i=1}^{g} \sum_{j=1}^{p} (x_{gij} - \widehat{x}_{gj})^2 - \sum_{i=1}^{h} \sum_{j=1}^{p} (x_{hij} - \widehat{x}_{hj})^2 - \sum_{i=1}^{h'} \sum_{j=1}^{p} (x_{h'ij} - \widehat{x}_{h'j})^2 \tag{2.24}$$

Přírůstek je vyjádřen jako součet čtverců v nově vznikajícím shluku, zmenšený o součty čtverců v obou zanikajících shlucích. Tento výraz můžeme zjednodušit na (viz. [9])

$$\Delta G1 = \frac{n_h n_{h'}}{n_h + n_{h'}} \sum_{j=1}^{p} (\widehat{x}_{hj} - \widehat{x}_{h'j})^2. \tag{2.25}$$

$\Delta G1$ roste s rostoucí velikostí shluků a pro pevné $n_h + n_{h'}$ je maximální při shlucích shodné velikosti $n_h = n_{h'}$. Jelikož minimalizujeme kritérium $\Delta G1$, má Wardova metoda tendenci kombinovat shluky s malým počtem objektů a vytvářet shluky shodné velikosti.

2.3 Optimalizační algoritmy

Jak již bylo zmíněno v úvodu této kapitoly, cílem shlukové analýzy je vytvořit kompaktní a dobře separované shluky. Naším úkolem je dosáhnout extrému nějakého funkcionálu, obvykle Wardova kritéria (2.6). Rozklad, který tuto podmínku splní, považujeme za optimální. Spolehlivou cestou

by bylo probrání všech možných variant rozkladu s výpočtem (2.6). V reálných úlohách je však možných variant rozkladu příliš mnoho. V této kapitole si ukážeme různé optimalizační algoritmy, které zaručují nalezení aspoň jednoho lokálního extrému funkcionálu rozkladu (např. minimalizace vzdáleností mezi objekty v jednotlivých shlucích).

2.3.1 Metoda k-průměrů

Princip metody k-průměrů spočívá v rozdělení n objektů charakterizovaných m proměnnými do k shluků, tak, aby mezishluková suma čtverců byla minimální a počet shluků k je předem dán. Jelikož je počet možných uspořádání veliký, nelze očekávat jediné nejlepší řešení. Algoritmus metody k-průměrů spíše najde optimum lokální než globální. Tento algoritmus pracuje iterativně, startuje vždy z jiného počátečního uspořádání a nakonec vybere vhodné řešení ze všech možných dosažených uspořádání shluků.

a) Klasifikace objektů do shluků, když těžiště shluků jsou předem známa.

Jelikož metoda požaduje uživatelem předem zadaný počet shluků, je třeba nejprve aplikovat hierarchickou analýzu shluků na náhodný výběr objektů a určit tak počet shluků před vlastním použitím metody k-průměrů na všechny objekty. Hierarchickou analýzou se také určí počáteční těžiště shluků pro následnou metodu k-průměrů. Zadáváme počet shluků, jež mají být nalezeny, a pak jsou vytvořeny prostorové shluky a to nalezením souboru těžisť shluků tak, že každý objekt je přiřazen do jednoho shluku a následně jsou určeny nové shluky a celý proces se opakuje, jak je uvedeno v [9].

b) Klasifikace objektů do shluků, když těžiště shluků nejsou předem známa.

Algoritmus pracuje tak, že prvních k objektů v datech (kde k je požadovaný počet shluků) je vybráno jako dočasná těžiště. V následujících krocích nahradí objekt těžiště, když jeho nejmenší vzdálenost k těžišti bude větší než vzdálenost mezi dvěma nejbližšími těžišti. Těžiště, které je bližší k objektu, je pak vyměněno. Objekt se dosadí na místo, když vzdálenost 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 uspořádání je pak optimalizováno. Start procesu z rozličných náhodných uspořádání velmi zvýší pravděpodobnost na nalezení nejlepšího řešení a optimálního počtu shluků, viz [9].

Uvedeme ještě kritérium věrohodnosti (těsnosti proložení). Předpokládejme n objektů rozdělených do k shluků. Pak k-tý shluk obsahuje n_k objektů. Každý objekt je popsán m proměnnými. Kritérium těsnosti proložení, dle *mezishlukové sumy čtverců* je definované jako (viz [9])

$$E_k = \frac{nm}{nm - m} \sum_{l=1}^{k} \sum_{i=1}^{m} \sum_{j=1}^{n_k} (x_{ij} - \widehat{x}_{il})^2 \tag{2.26}$$

a procento variace PV_k je definováno jako

$$PV_k = \frac{E_k}{E_1}, \tag{2.27}$$

které udává sumu čtverců pro danný počet shluků, ve formě procenta, kdyby vůbec nedošlo ke shlukování.

2.3.2 Metoda optimálních středů (medoidů)

Medoid je optimální střed shluku, tedy takový střední objekt, pro který platí, že průměrná vzdálenost k ostatním objektům v tomto shluku je minimální. Požadujeme-li k shluků, bude existovat také k medoidů, čili každý shluk má vlastní medoid. Po nalezení medoidů jsou data klasifikována do shluků vždy okolo nejbližšího medoidu.

a) Späthova metoda

Metoda minimalizuje účelovou funkci přemísťováním objektů z jednoho shluku do druhého. Začíná se u počátečního uspořádání shluků, algoritmus pak najde lokální minimum přesouváním objektů ze shluku do shluku. Jakmile se nepřemístí už žádný objekt, proces končí. Lokální minimum však nemusí být globálním. Aby program obešel toto omezení, zopakuje se několikrát hledání vždy z jiného startovacího uspořádání a nejlepší uspořádání shluků je nakonec bráno za výsledné. Jako účelová funkce se používá celková vzdálenost

mezi všemi objekty ve shlucích dle

$$D = \sum_{l=1}^{k} \sum_{i=1}^{c_k} \sum_{j=1}^{c_k} d_{ij} \tag{2.28}$$

kde k je celkový počet shluků, d_{ij} představuje vzdálenost mezi i-tým a j-tým objektem a c_k udává soubor všech objektů ve shluku l.[9]

b) PAM metoda *(Partition Around Medoids)*

Minimalizuje celkovou vzdálenost D (2.28), kde proces postupuje následovně:

1. Nalezne se reprezentativní soubor k objektů a označí se jako medoidy.

2. Přiřadíme ostatní objekty do shluků, dle nejkratší vzdálenosti k některému z medoidů. A spočítáme celkovou vzdálenost D

3. Pro každý shluk, vyměníme medoid a některý z objektů v daném shluku a spočítáme celkovou vzdálenost D.

4. Vybereme uspořádání s nejmenší hodnotou kritéria D.

5. Opakujeme kroky 2-5 dokud nedojde k žádné změně medoidů. Celá iterace končí, jakmile změny nezpůsobí další snížení kritéria D.

Silueta

Jde o statistické kritérium, které poskytuje informaci o dobrém a špatném shluku. Hodnota siluety s objektu i se určí následujícím způsobem:

1. Objekt i je ve shluku A a má průměrnou vzdálenost a ke všem objektům ve svém shluku. Je-li ve shluku A jediný objekt, potom $a=0$

2. Sousední shluk B obsahuje objekty, které jsou nejbližší k objektu i ve shluku A a b je průměrná vzdálenost mezi objektem i a všemi objekty ve shluku B.

3. Silueta s objektu i se vyčíslí tak, že pokud shluk A obsahuje pouze

jeden objekt je *s=0*. Když $a < b$, je $s = 1 - \frac{a}{b}$. Když $a > b$, je $s = \frac{b}{a} - 1$. Když $a = b$, je $s = 0$.

Silueta se vypočítá pro každý objekt. Hodnota siluety nabývá hodnot od -1 do 1 a je mírou úspěšné klasifikace do shluků při porovnávání vzdáleností uvnitř shluku A se všemi vzdálenostmi objektů od nejbližšího souseda B, viz [9]

Kvalita klasifikace se hodnotí následovně:

1. Je-li s blízko 1, objekt i je dobře klasifikován do shluku A, protože jeho vzdálenosti k ostatním objektům v tomto shluku jsou podstatně kratší než vzdálenosti k objektům nejbližšího sousedního shluku B.

2. Je-li s blízko hodnoty 0, objekt i se nachází uprostřed mezi shluky A a B a byl přiřazen do shluku A čistě náhodně.

3. Je-li s blízko hodnotě -1, objekt i je špatně klasifikován, Vzdálenosti k ostatním objektům ve svém shluku jsou větší než vzdálenosti k objektům nejbližšího shluku B.

Toto statistické kritérium se dá využít i na určení optimálního počtu shluků. Relativně přehledným kritériem je průměrná silueta s počítaná přes všechny objekty. Tato hodnota určuje, jak těsně je proloženo navržené shlukové uspořádání analyzovanými daty. Optimální počet shluků je dán maximalizací průměrné siluety. Označíme maximální hodnotu průměrné siluety všech shluků k jako *MaxS* a rozlišujeme typy shlukových uspořádání dle (viz [9])

Tabulka 2.1 Typy shlukových uspořádání dle siluety

MaxS	Kvalita uspořádání
od 0,71 do 1	silná a dobrá struktura
od 0,51 do 0,7	přijtelná struktura
od 0,26 do 0,5	slabá struktura; je třeba najít lepší strukturu
od -1 do 0,25	naprosto nevhodná struktura

Po úspěšném nalezení počtu shluků by měla následovat diskriminační analýza, která statisticky testuje, jak dobře byly objekty roztříděny do shluků. Testování se provádí pomocí *Wilkovy statistiky* Λ, tento přístup popsán ve

třetí kapitole.

2.3.3 Fuzzy shlukování

Fuzzy shlukování, na rozdíl od předchozích shlukovacích metod, umožňuje shlukování jednoho objektu do více shluků. Předpokládejme, že máme k shluků a budeme definovat soubor tzv. účastí $m_{i1}, m_{i1}, ..., m_{ik}$. Kde m_{ik} představuje pravděpodobnost, že objekt i je klasifikován do j-tého shluku. V předchozích metodách byla vždy právě jedna tato účast rovna 1 a zbytek byl roven 0, čili došlo k přiřazení každého objektu do právě jednoho shluku dle [9].

Ve fuzzy shlukování je přítomnost objektu rozdělena do všech shluků. Účast m_{ij} se dá interpretovat jako pravděpodobnost a musí tudíž musí platit, že $0 \leq m_{ij} \leq 1 \; \forall \; i,j$ a $\sum_{k=1}^{j} m_{ik} = 1 \; \forall \; i$. Tento postup uvažující jednotlivé pravděpodobnosti účasti m_{ij} nazýváme fuzifikací shlukové konfigurace. Výhodou je, že objekt nemusí být zařazen pouze do jediného specifického shluku. Nevýhodou ovšem je, že se objevuje mnohem více informací, které musí být posléze analyzovány. Fuzzy algoritmus minimalizuje účelovou funkci f, jak je uvedeno v [4], která je funkcí neznámých účastí ve shluku a dále funkcí vzdáleností

$$f = \sum_{h=1}^{k} \frac{\sum_{i=1}^{n} \sum_{i'=1}^{n} m_{ih}^2 m_{i'h}^2 D_{ii'}}{2 \sum_{i'=1}^{n} m_{i'h}^2}, \tag{2.29}$$

kde nepodobnosti/vzdálenosti mezi objekty i a i' $D_{ii'}$ jsou známé a m_{ih} jsou neznámé účasti objektu i v h-tém shluku. Ohodnocení fuzzy shlukování do k shluků se měří Dunnovým rozdělovacím koeficientem, který představuje míru, jak těsně padne vysledné fuzzy shlukování na odpovídající pevné shluky. Za pevné shluky budeme považovat klasifikaci každého objektu do shluku, v nemž má uvažovaný shluk největší m_{ij}. Pro n objektů a k shluků se vyjádří Dunnův rozdělovací koeficient jako (viz [9])

$$F_k = \sum_{j=1}^{k} \sum_{i=1}^{n} \frac{m_{ij}^2}{n}, \tag{2.30}$$

který leží v intervalu $\langle \frac{1}{k}, 1 \rangle$.

Rozlišujeme dvě extrémní situace, jak jsou popsány v [9]:

a) Úplné fuzzy shlukování

$$u_{ik} = \frac{1}{k}; F_k = nk\frac{1}{nk^2} = \frac{1}{k} \tag{2.31}$$

Všechny objekty jsou shodně ve všech shlucích.

b) Pevné shlukování

$$u_{ik} \text{ nabývají hodnot pouze 0 nebo 1}; F_k = \frac{n}{n} = 1 \tag{2.32}$$

Každý objekt je jednoznačně přiřazen do právě jednoho shluku.

Normovaný tvar Dunnova koeficientu, který je vždy v intervalu $\langle 0, 1\rangle$, je

$$F'_k = \frac{kF_k - 1}{k - 1} \tag{2.33}$$

Další rozdělovací koeficient představuje Kaufmanův rozdělovací koeficient definovaný jako (viz [9])

$$D_k = \sum_{l=1}^{k}\sum_{i=1}^{n}\frac{(h_{ij} - m_{ij})^2}{n}, \tag{2.34}$$

kde h_{ij} a m_{ij} jsou pravděpodobnosti, že objekt i je zařazen do shluku k či j. Kaufmanův koeficient leží v intervalu $\langle 0$ (pevné shluky),$1 - \frac{1}{k}$ (úplné fuzzy) $\rangle$

Normovaná verze Kaufmanova koeficientu má tvar

$$D'_k = \frac{D_k}{1 - \frac{1}{k}}, \tag{2.35}$$

„*Oba normované koeficienty F'_k a D'_k poskytují dobrou indikaci optimálního počtu shluků. Celočíselná hodnota k by měla být volena tak, že F'_k bude nabývat velkých hodnot a D'_k malých hodnot*, viz [9]".

Kapitola 3

Diskriminační analýza

Diskriminační analýza patří mezi metody zkoumání vztahu mezi skupinou p vzájemně nezávislých znaků *diskriminátory* a jednou kvalitativní závislou proměnnou (hledaným výstupem). V nejjednodušším případě si tento výstup můžeme představit jako binární proměnnou, nabývající hodnoty 0, pokud je sledovaný objekt v prvním shluku a 1, pokud je v druhém shluku. Každý objekt patří právě do jednoho z nich a shluky jsou vzájemně odlišené. Cílem je nalézt predikční model umožňující zařadit objekty do shluků.

3.1 Kanonická diskriminační analýza

Základem kanonické diskriminační analýzy je nalézt takovou lineární kombinaci p sledovaných proměnných, tedy $\mathbf{Y}= \mathbf{v}^T\mathbf{x}$, kde $\mathbf{v}^T = [\mathrm{v}_1,v_2, \ldots ,v_p]$ je vektor parametrů, tak aby tato lineární kombinace byla co nejlepší z pohledu rozlišení uvažovaných p shluků, v tom smyslu, aby vnitroskupinová variabilita byla co nejmenší a meziskupinová variabilita co největší.

Celková variabilita původních proměnných je vyjádřena čtvercovou maticí $\mathbf{T}$ typu p x p vyjádřená dle (2.5). Matice $\mathbf{T}$ se dá vyjádřit jako součet matice vnitroshlukové variability $\mathbf{E}$ (2.3) a matice mezishlukové variability $\mathbf{B}$ (2.4). Míry meziskupinové a vnitroskupinové variability pro novou veličinu $\mathbf{Y}$, jsou dány jako (viz [4]):

$$Q_B(Y) = v^T B v \qquad \text{a} \qquad Q_E(Y) = v^T E v \tag{3.1}$$

Největší meziskupinové a nejmenší vnitroskupinové variability dosáhneme

při maximání hodnotě podílu

$$\lambda = \frac{Q_B(Y)}{Q_E(Y)} = \frac{v^T B v}{v^T E v} \tag{3.2}$$

známém jako Fischerovo diskriminační kritérium, viz [4].

Pro nalezení vektoru **v** Fischerovo diskriminační kritérium zderivujeme a parciální derivace položíme rovny 0. Dostáváme tedy

$$\frac{\delta\lambda}{\delta v} = \frac{(v^T B^T + v^T B)v^T E v - (v^T E^T + v^T E)v^T B v}{(v^T E v)^2} = (B - \lambda E)v$$

Jelikož matice **E** a **B** jsou symetrické. Uvedenou rovnici ještě přepíšeme do tvaru

$$(BE^{-1} - \lambda I)v = 0 \text{ , kde } |E^{-1}| \neq 0. \tag{3.3}$$

Netriviální řešení dostáváme pouze v případě, když charakteristický polynom $|BE^{-1} - \lambda|$ je roven 0. Čili celý výpočet se dá zjednodušit na výpočet charakteristických čísel l_k matice $\mathbf{BE^{-1}}$. Charakteristická čísla uspořádáme sestupně a vypočítáme pro ně příslušné ortogonální charakteristické vektory v_1, v_2,...,v_N. Kde charakteristický vektor v_1 příslušný největšímu vlastnímu číslu stanovuje poměr mezi prvky, který maximalizuje Fischerovo diskriminační kritérium. Maximum diskriminačního kritéria je pak dáno velikostí maximálního vlastního čísla a tedy určuje meziskupinovou variabilitu veličiny $x.v_1$. *Můžeme to chápat jako projekci bodů reprezentující jednotlivé p-rozměrné objekty na přímku, pokud danné objekty dělíme do dvou skupin*, viz [4]. V případě dělení do *K* skupin potřebujeme zobrazení do *(K-1)* rozměrného prostoru a *(K-1)* diskriminantů. Kde další kanonickou proměnnou $v_r^T x, r = 2, 3, ..., N$ získáme při použití charakteristického vektoru příslušnému dalšímu vlastnímu číslu.

Výhodné je postačit si s co nejmenším počtem diskriminantů, hlavně z hlediska zobrazení do prostoru. K tomuto účelu je vhodné určit poměr $\frac{l_{(k)}}{\sum l}$, který vyjadřuje nakolik se *r*-tý diskriminant podílí na odlišení jednotlivých skupin a celkovou variabilitu *r*-tého diskriminantu vyjadřuje prvek $1 + l_{(r)}$. Pak *r*-té diskriminační skóre je definováno jako $D_r^s = v_r^T x$.

Přičteme-li ke každému skóre konstantu w_r, která je definována jako

$$w_r = -v_r^T \bar{x} = -\sum_{j=1}^{p} v_{jr} \bar{x_j}, \tag{3.4}$$

kde $\bar{x_j}$ je průměr všech hodnot *j*-té veličiny, *j=1, 2, ..., p*, potom je průměrné diskriminační skóre jednolivých diskriminantů nulové. Potom pro *i*-tou jednotku, *i=1, 2, ..., n_k*, v *k*-té skupině, *k=1, 2, ..., k*, je *r*-té diskriminační skóre definováno jako

$$y_{kir} = w_r + \sum_{j=1}^{p} v_{jr} x_{kij} \tag{3.5}$$

Koeficienty v_{jr} vektoru r-té kanonické proměnné můžeme chápat jako vliv původní X_j té proměnné na *r*-tou kanonickou proměnnou Y_r. Vhodné je vektor v_r normovat, pro lepší pochopení vlivu původních proměnných *r*-tou kanonickou proměnnou. Definujeme matici **F** jako diagonální matici s odmocninami diagonálních prvků matice **E**, potom normovaný vektor v_r^* je definován jako

$$v_r^* = \frac{1}{\sqrt{n-k}} K v_r \tag{3.6}$$

Dále je vhodné určit korelační koeficienty mezi kanonickou proměnnou a původními proměnnými. Potom pro *r*-tý diskriminant dostáváme

$$r_r = \frac{1}{\sqrt{n-k}} F^{-1} E v_r. \tag{3.7}$$

V dalším kroku je potřeba určit, které diskriminanty jsou jsou vhodné k odlišení jednotlivých shluků. Vyslovíme hypotézu, že žádný disktriminant není vhodný pro odlišení skupin. Což znamená, že všechna vlastní čísla jsou nulová. Testovou statistiku můžeme založit na testu o shodě vektorů středních hodnot na Wilksově statistice $\Lambda = \frac{|E|}{|E+B|}$, kdy počet diskriminantů je menší než dva. V případě, že je počet diskriminantů roven dva a více využijeme Bartlettovu statistiku a jako testové kritérium volíme (viz [4]).

$$V = c(-ln\Lambda), \tag{3.8}$$

kde $c = n - 1 - (p + K)/2$, má přibližně $\chi_{p(K-1)}$. Zamítnutí hypotézy $\lambda_{(1)} = \lambda_{(2)} = ... = \lambda_{(N)} = 0$, znamená, že aspoň jedno z vlastních čísel je větší než 0. Dále můžeme navázat testováním hypotézy, pro další vlastní čísla $\lambda_{(r+1)}, \lambda_{(3)}, ..., \lambda_{(N)} = 0$, kdy použijeme testové kritérium (viz [4]).

$$V = (n - 1 - \frac{p+k}{2}) \sum_{i=r}^{R} ln(1 + l_i), \tag{3.9}$$

jež má přibližně $\chi_{(p-r+1)(K-r)}$ rozdělení.

Kapitola 4

Výpočetní část

V této části provedeme výpočty na reálných datech. Historická data obsahují celkovou škodu (celková škoda, kterou pojištěný utrpěl), celkovou výši plnění (celková výše plnění, které bylo pojištěnému vyplaceno v důsledku pojistné události), rok vzniku události, rok nahlášení události a popis události. Data je nejprve nutno roztřídit dle námi stanovených homogenních segmentů pojistných smluv a poté odhadnout parametry distribučních funkcí. Parametry distribučních funkcí byly odhadnuty jednak za pomoci historických dat (pro segmenty, které mají dostatečnou historii), v případech, kde není dostatek historických dat se parametry odhadují na základě expertních odhadů oddělení pojistné matematiky a upisovacího oddělení. Pro účely diplomové práce předpokládáme, že máme již k dispozici odhady parametrů distribučních funkcí pro malé, velké a katastrofické škody pro každý segment pojistných smluv. Dále máme k dispozici pro každý segment pojistných smluv kumulativní vývojový trojúhelník škod na jehož základě byla odhadnuta celková rezerva a variabilita odhadu v jednotlivých letech. Pro účely diplomové práce předpokládáme, že již máme k dispozici odhady rezerv a příslušnou variabilitu těchto odhadů. Automatické použití výpočetních metod může často vést k chybnému odhadu, proto se v praxi často používají expertní odhady, které verifikují a případně upraví příslušné odhady.

V první části této kapitoly vypočítáme rizikové charakteristiky upisovacího rizika jednotlivých segmentů na jejichž úrovni byly odhadnuty parametry distribučních funkcí škod a ukážeme si metody odhadu rezervy a variability pomocí *Mack* nebo *Bootstrap* metody. A následně provedeme výpočet rizika rezerv (*Reserve Risk*) pro jednotlivé segmenty pojistných smluv.

V druhé části této kapitoly na segmenty pojistných smluv, charakterizovaných vypočítanými rizikovými parametry, aplikujeme shlukovací metody. Dále určíme vhodné počty shluků, charakteristické znaky jednotlivých shluků a jejich význam pro pojišťovnu. Následně v této kapitole provedeme srovnání použitých metod.

4.1 Výpočet rizikových parametrů

Pro každý homogenní segment pojistných smluv dle definice CEIOPS máme k dispozici odhady parametrů distribučních funkcí škod a kumulativní vývojový trojúhelník škod.

4.1.1 Upisovací riziko

Jak bylo již uvedeno v teoretické sekci, upisovací riziko dělíme na riziko ztráty z malých, velkých, katastrofických škod.

Pro každý segment pojistných smluv máme k dispozici následující parametry:

1) Parametry distribuční funkce rozdělení malých škod μ, σ. Předpokládáme, že malé škody jsou ve všech skupinách log-normálně rozdělené.

2) Předpokládáme, že frekvence vysokých škod je pro všechny skupiny modelována pomocí negativně binomického rozdělení. Výše jednotlivých škod je poté modelována jedním z následujících rozdělení (Paretovo, posunuté Log-normální, Zobecněné Paretovo).

3) Frekvence katastrofických škod je modelována pomocí Poissonova rozdělení a výše škod jednotivých scénářů pomocí Paretova rozdělení.

Pro každý segment máme k dispozici následující tabulku s odhadnutými parametry pro upisovací riziko (*underwriting risk*):

Tabulka 4.1. Příklad odhadnutých parametrů distribučních funkcí upisovacího rizika pro jeden segment

Název	Parametr	Hodnota
Malé škody		
Logaritmicko-normální rozdělení	μ	-1.644
	σ	0.5104
Velké škody		
- *frekvence*		
Negativně-binomické	p	0.277
	n	1
- *Výše škody*		
Log-normální rozdělení	μ	14.47
	σ	1.20
	treshold	750 000
	Maximální pojistná částka	50 000 000
- *Katastrofické škody*		
Selhání jedince	Čas Návratu 1	25 let
	Škoda 1	1.3 mil
	Čas Návratu 2	250 let
	Škoda 2	24.1 mil
	Maximální ztráta	48.2 mil
Vysoká inflace	Čas Návratu 1	25 let
	Škoda 1	1 mil
	Čas Návratu 2	250 let
	Škoda 2	3 mil
	Maximální ztráta	6 mil
Ekonomická recese	Perioda návratu 1	25 let
	Škoda$_{PN_1}$	0,01 mil
	Perioda návratu 2	250 let
	Škoda$_{PN_1}$	80 mil
	Maximální ztráta	117 mil
Přijaté pojistné		35.292 mil

Malé škody

Jak již bylo uvedeno malé škody modelujeme agregovaně pomocí logaritmicko-normálního rozdělení. Výše malých škod je simulována pomocí kvantilové metody, kde kvantilová funkce logaritmicko-normálního rozdělení je

$$X_{Att} = e^{\Phi(P)^{-1}.\sigma+\mu}, \tag{4.1}$$

kde $P \sim R(0,1)$.

Velké škody

Frekvence

Frekvenci velkých škod modelujeme pomocí negativně binomického rozdělení. Moderní software (např. Igloo od společnosti EMB [3]) nám dovuluje modelovat toto rozdělení přímo a výstupem je počet škod v daném období. Pokud tento software k dispozici nemáme, můžeme počet událostí simulovat pomocí jednotlivých dob mezi událostmi.

Uvažujeme, že škody mají Poissonovo rozdělení s parametrem λ. Potom doba mezi jednolivými událostmi je exponenciálně rozdělená se střední hodnotou $\frac{1}{\lambda}$. Pro Poissonovo rozdělení konjugovaný systém tvoří systém gamma rozdělení. Je-li apriorní rozdělení Poissonovo rozdělení, pak aposteriorní rozdělení je negativně binomické:

$$\lambda \sim Gamma(\frac{1}{h}, h).EN;\ h = \frac{EN}{VarN - EN}, \tag{4.2}$$

kde $N \sim NegBin(n,p)$. V případě negativně-binomického rozdělení má čas mezi událostmi exponenciální rozdělení s parametrem
$\Theta \sim Gamma(\frac{VarN-EN}{EN}, \frac{EN}{VarN-EN}) \cdot EN$.

Simulujeme čas mezi jednotlivými událostmi t_i a čas i-té události je dán jako

$$T_i = \sum_{k=1}^{i} t_k, \tag{4.3}$$

počet událostí v danné periodě je dán jako

$$N = min\{i; T_i \leq 1\}. \tag{4.4}$$

Výše škody a pojistného plnění

V průběhu parameterizace byly odhadnuty parametry distribuční funkce velých škod. Musíme ovšem také brát v potaz maximální pojistnou částku (MPC), která stanovuje horní mez pojistného plnění. Výši pojistného plnění modelujeme pomocí kvantilové funkce:

I. Posunuté a useknuté Logaritmicko-normální rozdělení

$$X_{LLSev} = Max(Treshold, Min(MPC, e^{\Phi(P)^{-1}.\sigma+\mu})), \tag{4.5}$$

kde $P \sim R(0,1)$ $\mu \epsilon R, \sigma \epsilon R, 0 < Treshold < MPC$.

II. Paretovo rozdělení

$$X_{LLSev} = Min(MPC, \frac{Treshold}{(1-P)^{\frac{1}{\alpha}}}), \tag{4.6}$$

kde $P \sim R(0,1)$, $\alpha > 0, 0 < Treshold < MPC$.

III. Zobecněné-Paretovo rozdělení

$$X_{LLSev} = Min(MPC, Treshold + \frac{scale(P^{-shape}-1)}{shape})), \tag{4.7}$$

kde $P \sim R(0,1)$, $scale > 0, 0 < Treshold < MPC$.

Celkové pojistné plnění je dáno jako

$$S = \sum_{k=1}^{N} X_k. \tag{4.8}$$

Předpokládáme-li nezávislost mezi výší pojistného plnění a frekvencí potom střední hodnota pojistného plnění je dána jako

$$ES = EXEN, \tag{4.9}$$

celkový rozptyl je dán dle (viz [8])

$$VarS = EVar(S|N) + VarE(S|N) = EN.VarX + VarN.(EX)^2. \tag{4.10}$$

Katastrofické škody

Frekvence

Frekvence katastrofických škod je modelována pomocí Poissonova rozdělení s parametrem $\lambda = \frac{1}{PeriodaNavratu_1}$. Frekvenci simulujeme podobně jako v případě velkých škod pomocí doby mezi událostmi. V případě škod, jejiž frekvence se modeluje pomocí Poissonova rozdělení, je doba mezi událostmi exponenciálně rozdělena se střední hodnotou $PeriodaNavratu_1$. Počet událostí určen stejně jako v (4.4).

Výše škody a pojistného plnění

Výši škody modelujeme pomocí kvantilové funkce Paretova rozdělení s parametry α odhadnutým dle (1.17), $Treshold = Skoda_{PN_1}$ a pojistné plnění je zhora omezeno maximální možnou ztátou. Pojistné plnění je pak dáno jako

$$X_{Catsev} = Min(MaxLoss, \frac{Treshold}{(1-P)^{\frac{1}{\alpha}}}), \tag{4.11}$$

kde $P \sim R(0,1)$, $\alpha > 0, 0 < Treshold < MPC$.

Celkové pojistné plnění je dáno jako

$$S = \sum_{k=1}^{N} X_k. \tag{4.12}$$

Předpokládáme nezávislost mezi výší pojistného plnění a frekvencí potom střední hodnota pojistného plnění je dána jako

$$ES = EXEN, \tag{4.13}$$

předpokláname, že frekvence má Poissonovo rozdělení, potom celkový rozptyl je dán jako

$$VarS = EN.VarX + VarN.(EX)^2 = EN.EX^2. \tag{4.14}$$

Výpočet

Každý segment modelujeme na bázi škodního poměru *(Loss Ratio)*, který je definován jako

Definice: Škodní poměr je poměr celkové výše škod pro pojišťovnu (pojistného plnění) vzniklých v daném roce k zaslouženému pojistnému připadajícímu na tento rok.

Modelování pomocí škodního poměru je výhodné zejména kvůli nezávislosti na výši pojistného (objemu) v jednotlivých segmentech a díky tomu nám umožňuje lépe analyzovat rizikovost danné skupiny. Z parametrů distribučních funkcí pro každý segment dostáváme následující výstup škodního poměru pro upisovací riziko:

Tabulka 4.2 Výsledek pro segment definovaný dle parametrů v tabulce 4.1

Charakteristiky / Percentil	Malé Škody	Malé a Velké Škody	Malé, velké a Katastrofické Škody
Průměrný škodní poměr	22 %	56 %	59%
Směrodatná odchylka	12%	50%	55%
Koeficient Variace	55%	88%	92%
1%	6%	8%	8%
5%	8%	12%	13%
10%	10%	15%	16%
20%	13%	21%	22%
25%	14%	23%	25%
30%	15 %	26%	27%
40%	17%	32%	34%
50%	19%	40%	42%
60%	22%	49%	52%
70%	25%	63%	66%
75%	27%	71%	75%
80%	30%	82%	86%
90%	37%	117%	124%

ad) Tabulka 4.2

Charakteristiky / Percentil	Malé Škody	Malé a Velké Škody	Malé, velké a Katastrofické Škody
95%	45%	156%	166%
96%	47%	169%	181%
99%	63%	245%	278%
99.5%	73%	283%	344%
99.6%	75%	295%	353%
99.9%	100%	368%	418%

Výpočet a simulace byla provedena v softwaru *Wolfram Mathematica 6* [12].

4.1.2 Riziko rezerv

Pro každý segment máme k dispozici kumulativní vývojový trojúhleník škod $\{X_{ik}\}$ a provedeme výpočet rizikových parametrů pomocí *Mack* metody, jak byl popsán v kapitole 1.2. Uvedeme jeden konkrétní příklad výpočtu odhadu rezerv pomocí *Mack* a *Bootstrap* metody, následně provedeme srovnání těchto přístupů.

Mack metoda

Tabulka 4.3 Kumulativní vývojový trojúhleník škod pro jeden segment:

Rok	1	2	3	4	5	6	7
2001	357 848	1 124 788	1 735 330	2 218 270	2 745 596	3 319 994	3 466 336
2002	352 118	1 236 139	2 170 033	3 353 322	3 799 067	4 120 063	4 647 867
2003	290 507	1 292 306	2 218 525	3 235 179	3 985 995	4 132 918	4 628 910
2004	310 608	1 418 858	2 195 047	3 757 447	4 029 929	4 381 982	4 588 268
2005	443 160	1 136 350	2 128 333	2 897 821	3 402 672	3 873 311	
2006	396 132	1 333 217	2 180 715	2 985 752	3 691 712		
2007	440 832	1 288 463	2 419 861	3 483 130			
2008	359 480	1 421 128	2 864 498				
2009	376 686	1 363 294					
2010	344 014						

ad) Tabulka 4.3

Rok	8	9	10
2001	3606286	3 833 515	3 901 463
2002	4914039	5 339 085	
2003	4909315		
2004			
2005			
2006			
2007			
2008			
2009			
2010			

Metodou *Chain ladder* (viz [8]) odhadneme vývojové koeficienty a následně vypočítáme příslušné kumulativní vývojové koeficienty. Vývojové koeficienty představují nejlepší odhad nárůstu ve vývoji v letech metodou nejmenších čtverců.

Tabulka 4.4 Vývojové koeficienty pro kumulativní trojúhelník uvedený v tabulce 4.3

k	Vývojové koeficienty	Kumulativní koeficienty
1	3.4906	14.4466
2	1.1747	4.1387
3	1.4574	2.3686
4	1.1739	1.6252
5	1.1038	1.3845
6	1.0863	1.2543
7	1.0539	1.1547
8	1.0766	1.0956
9	1.0177	1.0177

Dále dle rovnic (1.18), (1.21) a (1.22) spočítáme rezervu, riziko odhadu parametru a riziko procesu pro jednotlivé roky. Celkem výsledek výpočtu Mackovou metodou můžeme shrnout do následující tabulky

Tabulka 4.5 Výsledek odhadu rezervy a rizika pomocí *Mack* metody

Rok	Rezerva	Celková škoda v danném roce	Riziko procesu (Směrodatná odhylka)	Riziko odhadu parametru (Směrodatná odhylka)	CoV Rezervy	Celková škoda CoV
2001	0	3 901 463	0	0	0%	0%
2002	94 634	5 433 719	48 832	57 628	80%	2%
2003	469 511	5 378 826	90 524	81 338	26%	2%
2004	709 638	5 297 906	102 622	85 464	19%	2%
2005	984 889	4 858 200	277 880	128 078	27%	5%
2006	1 419 459	5 111 171	366 582	185 867	29%	8%
2007	2 177 641	5 660 771	500 202	248 023	26%	11%
2008	3 920 301	6 784 799	785 741	385 759	22%	15%
2009	4 278 972	5 642 266	895 570	375 893	23%	14%
2010	4 625 811	4 969 825	1 284 882	455 270	29%	24%

Z tabulky 4.5. dostáváme, že celková rezerva je $18,680,856$ (jako součet rezerv v jednotlivých letech) s koeficientem variace 13.1% dle (1.23). Celková škoda byla Mackovým modelem odhadnuta na $53,038,946$ s koeficientem variace $4,6\%$. Celý výpočet Mackovou metodou je možné nalézt v přiloženém souboru *Mack.xls*

Bootstrap

Pro srovnání provedeme výpočet rezervového rizika metodou *bootstrap* na stejný trojúhelník. Metoda *bootstap* je poměrně jednoduchá, její myšlenka spočívá v tom, že vývojový koeficient mezi danými roky je náhodná veličina, jejíž napozorovanou hodnotu replikujeme mezi jednotlivými roky. Nyní si ukážeme metodu *bootstrap* na vývojovém trojúhelníku uvedeném v tabulce 4.3.

Máme kumulativní vývojový trojúhelník škod $\{X_{ij}^c\}$ $i,j = 1,...,t$, příslušný nekumulativní trojúhelník $\{Y_{ij}\}$, kde $Y_{ij} = X_{i,j+1} - X_{i,j}$ a dle metody *chain ladder* vypočítáme odhady vývojových koeficientů c_k viz. kapitola 1.2.

V dalším kroku spočítáme vyrovnaný kumulativní vývojový troúhelník škod

jako

$$X_{ij}^{Fit} = \frac{X_{it}}{\prod_{k=1}^{t-j} f_{t-k}} \tag{4.15}$$

a příslušný nekumulativní vývojový trojúhelník Y_{ij}^{Fit}.

Nyní stanovíme *Paersonova residua* pro každou položku nekumulativního trojúhelníku Y_{ij}^{Fit} jako

$$PaRes_{ij} = \frac{Y_{ij} - Y_{ij}^{Fit}}{\sqrt{Y_{ij}^{Fit}}} \tag{4.16}$$

Nyní dle hlavní myšlenky metody *bootstrap* náhodně zaměníme pozice residuií a získáme nový trohúhelník $PaRes_{ij}^{resha}$. Z nových residuí vypočítáme nový nekumulativní trojúhelník škod jako

$$Y_{ij}^{resha} = PaRes_{ij}^{resha}\sqrt{Y_{ij}^{Fit}} + Y_{ij}^{Fit}. \tag{4.17}$$

Následuje výpočet kumulativního trojúhelníku škod X_{ij}^{resha}, odhad vývojových koeficientů $c(k)^{resha}$ a výpočet rezervy pro každý upisovací rok. Po provedení *1000* simulací dostáváme (*1000*-krát náhodně zaměníme pozice residuí) *1000* vývojových trojúhelníků a *1000* různých výpočtů rezervy ze kterých vypočítáme odpovídající odhady výše rezerv a koeficientů variace.

Tabulka 4.6 Výsledek odhadu rezervy pomocí metody *Bootstrap*

rok	Rezerva	Celková škoda v danném roce	CoV Rezervy	CoV Škody
2001	0	3 901 463	0 %	0 %
2002	96 641	5 433 633	84.9%	1.6%
2003	475 643	5 382 496	31.3%	2.8 %
2004	723 263	5 294 363	24.6%	3.3 %
2005	998 386	4 863 863	20.3%	4.2 %
2006	1 432327	5 111 250	18.5%	5.1 %
2007	2 211 074	5 686 963	16.8%	6.6 %
2008	3 966 204	6 810 931	16.2%	9.8 %
2009	4 341 007	5 656 475	21.5%	16.7 %
2010	4 746 462	4 973 119	41.7 %	40.1 %

Celková rezerva byla metodou *bootstrap* stanovena na hodnotu $18,991,007$ s koeficientem variace 15%. Celková škoda byla odhadnuta na hodnotu $53,114,555$ s koeficientem variace 5.4%.

Srovnání metod

Pokud se podíváme na celkové výsledky získané pomocí *Mack* metody a metody *Bootstrap*, tak *Bootstrap* nám dává o něco vyšší odhad rezervy a celkové škody v danných letech s vyšší mírou chyby odhadu.

Tabulka 4.7 Srovnání metod vypočtu *Mack* a *Bootstrap* na stejný trojúhelník

rok	Bootstrap - Rezerva	Mack - Rezerva	Bootstrap CoV Rezervy	Mack CoV Rezervy
2001	0	0	0 %	0 %
2002	96 641	94 634	84.9%	80%
2003	475 643	469 511	31.3%	26 %
2004	723 263	709 638	24.6%	19 %
2005	998 386	984 889	20.3%	27 %
2006	1 432 327	1 419 459	18.5%	29 %
2007	2 211 074	2 177 641	16.8%	26 %
2008	3 966 204	3 920 301	16.2%	22 %
2009	4 341 007	4 278 972	21.5%	23 %
2010	4 746 462	4 625 811	41.7 %	29 %

V tomto konkrétním případě není mezi výsledky obou metod významný rozdíl (vyjma koeficientu variace v roce 2010, kde metoda *Bootstrap* vykazuje mnohem větší chybu odhadu). Nicméně výsledky obou metod se mohou za určitých okolností lišt mnohem významněji, zejména pokud nějaký vývojový koeficient výrazně převyšuje ostatní koeficienty, pak dochází k „rolování" tohoto koeficientu trojúhelníkem a tím k zvyšování chyby odhadu. Tomuto vlivu se dá zabránit pokud tento koeficient nepřemisťujeme s ostatními a ponecháváme ho na svém místě po celou dobu výpočtu. Obecně se obě metody musí používat velice obezřetně, protože jejich automatické užití, bez zpětné kontroly může vést k chybným závěrům.

4.2 Shlukování

V předcházející kapitole jsme ukázali jaká rizika sledujeme a metody výpočtu těchto rizik. Celkově máme 47 individualních rizikově homogenních

segmentů pojistných smluv dle definice CEIOPS ve kterých jsme vypočítali následující hodnoty rizikových charakteristik.

Tabulka 4.8 Odhady rizikových charakteristik pro jednolivé segmenty

Segment	Att Mean LR	Att SD	Large Mean LR	Large SD	Threat Mean LR	Threat SD	Res CoV
AD	36,4 %	14 %	4,6 %	10,8 %	8,1 %	31,1 %	14,9 %
BD	31,9 %	20,2 %	40,2 %	22,8 %	9,5 %	28,6 %	19,6 %
CE	38,1 %	8,5 %	21,6 %	12,1 %	5,7 %	16,2 %	14,1 %
CG	13,8 %	26,3 %	31 %	40,7 %	23,7 %	58,6 %	20 %
CW	16,8 %	2,6 %	35,2 %	24,6 %	11,4 %	23,3 %	17 %
CY	35,8 %	19,8 %	47,2 %	62,2 %	1,9 %	6,5 %	0 %
DJ	23,5 %	6,4 %	24,4 %	39,1 %	18,3 %	117,1 %	18,1 %
EC	6,8 %	2,6 %	35,6 %	26,9 %	14,8 %	28,6 %	14,5 %
FS	26,3 %	21,9 %	25,1 %	28,5 %	10,8 %	25,2 %	43,5 %
FW	21,4 %	11,7 %	32,4 %	46,1 %	3 %	20,1 %	10,9 %
GN	10,4 %	11,3 %	46,3 %	42,4 %	2,7 %	10,8 %	97,2 %
HX	11 %	6,1 %	14,7 %	8,7 %	20,4 %	30,1 %	8,7 %
IF	66,1 %	17,9 %	0 %	0 %	9,3 %	27,7 %	0 %
IN	31,4 %	12,4 %	35,9 %	21 %	8,1 %	21,3 %	19,6 %
JC	39,3 %	6,5 %	30,7 %	17,7 %	8,5 %	40,4 %	15 %
JI	20 %	8,4 %	23,8 %	25,9 %	29,8 %	73,2 %	20 %
KF	6,3 %	2,7 %	14,4 %	35,7 %	26,8 %	123,9 %	9,4 %
KI	78,3 %	38,6 %	0 %	0 %	3,8 %	16 %	0 %
KV	43 %	11,3 %	30,6 %	23,7 %	7,8 %	18,3 %	16,1 %
LB	38,3 %	6,5 %	20,4 %	16,9 %	12,1 %	29,5 %	27,9 %
LD	0 %	0 %	74,5 %	73,2 %	24,4 %	126 %	34,2 %
LJ	1 %	0,9 %	16,9 %	27,9 %	21,9 %	100,2 %	127,8 %
ML	40,2 %	5,4 %	5,5 %	3,4 %	17,6 %	32,4 %	9,7 %
MR	65,9 %	23,3 %	0 %	0 %	0 %	0 %	15,1 %
MU	52,2 %	10,4 %	14,9 %	9,3 %	6,6 %	12 %	22,5 %
NA	74 %	15 %	0 %	0 %	4,6 %	14,1 %	9,9 %
NY	74,4 %	36,7 %	0 %	0 %	18 %	80,9 %	7,3 %
OB	49 %	18,9 %	9,4 %	23,9 %	18,1 %	70,3 %	0 %
OE	42,8 %	13 %	17,7 %	11,8 %	8,6 %	19,1 %	8,7 %
QB	21,2 %	10,7 %	16,7 %	17,9 %	19,9 %	86,7 %	0 %

ad) Tabulka 4.8

Segment	Att Mean LR	Att SD	Large Mean LR	Large SD	Threat Mean LR	Threat SD	Res CoV
QE	2,1 %	3,2 %	9,8 %	19,8 %	24 %	79,3 %	33,7 %
QM	17 %	7,8 %	24,8 %	17,1 %	31,4 %	49 %	47,1 %
RM	34,9 %	8,5 %	26,5 %	13,3 %	8,4 %	22,5 %	13,8 %
SW	2,9 %	1,2 %	14,2 %	25,4 %	27,2 %	103,3 %	35 %
TF	54,5 %	12,5 %	10,6 %	4,7 %	5,8 %	17,1 %	21,8 %
TW	74,4 %	36,7 %	0 %	0 %	7 %	18,9 %	12,1 %
UE	15,9 %	7 %	40,8 %	26,2 %	11,2 %	19,8 %	28,4 %
UN	18,6 %	12,2 %	41,8 %	36,8 %	9,3 %	35,8 %	29,9 %
VP	45,2 %	7,8 %	15,9 %	9 %	7,6 %	13,1 %	14,6 %
WD	64,9 %	6,6 %	11 %	16,8 %	1,5 %	7,4 %	44,9 %
WN	43,3 %	9,1 %	37,7 %	19,2 %	1,4 %	12,6 %	16,4 %
WW	64,5 %	20,5 %	0 %	0 %	5,9 %	42,2 %	20,3 %
XH	44,9 %	9,6 %	31,8 %	20,3 %	5,8 %	13,5 %	16,9 %
XS	37,8 %	14,1 %	33,5 %	18,5 %	2,6 %	16,2 %	10,4 %
XX	31,9 %	8 %	40,3 %	40 %	10,8 %	38,8 %	16,9 %
YK	6,9 %	3,5 %	5,4 %	4,9 %	34,1 %	43,9 %	20,6 %
ZA	33 %	5,8 %	28,4 %	14,6 %	6,3 %	21,1 %	12,3 %

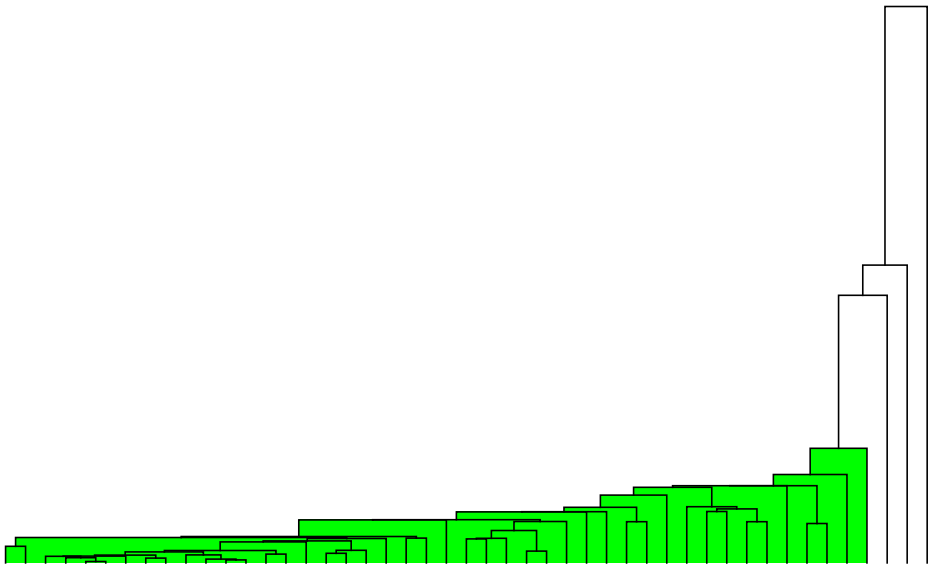
Všechny proměnné jsou kvantitativní.

4.2.1 Hierarchické shlukování

V této sekci provedeme analýzu studovaného portfolia pomocí metod hierarchického shlukování. Jednotlivé metody hierarchického shlukování byly popsány v teoterické sekci a výpočet byl proveden v softwaru *Wolfram Mathematica 6* [12].

Metoda nejbližšího souseda

Obrázek 4.1 Dendrogram portfolia metodou nejbližšího souseda

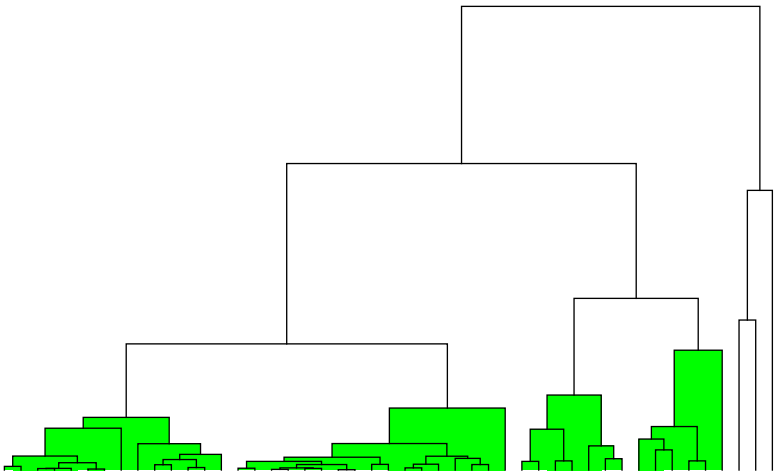


Poznámka: pro větší přehlednost doporučuji dendrogram v přiloženém souboru shlukovani.nb

Z dendrogramu je patrné, že segmenty *GN, LD* a *LJ* se shlukují až na vysoké hladině a jsou tudíž značně vzdálené od ostatních segmentů, jak je patrné i z dalších, v této práci uvedených, shlukovacích metod. A proto se jim budeme při podrobné analýze věnovat zvlášť.

Metoda nejvzdálenejšího souseda

Obrázek 4.2 Dendrogram portfolia metodou nejvzdálenejšího souseda



Poznámka: pro větší přehlednost doporučuji dendrogram v přiloženém souboru shlukovani.nb

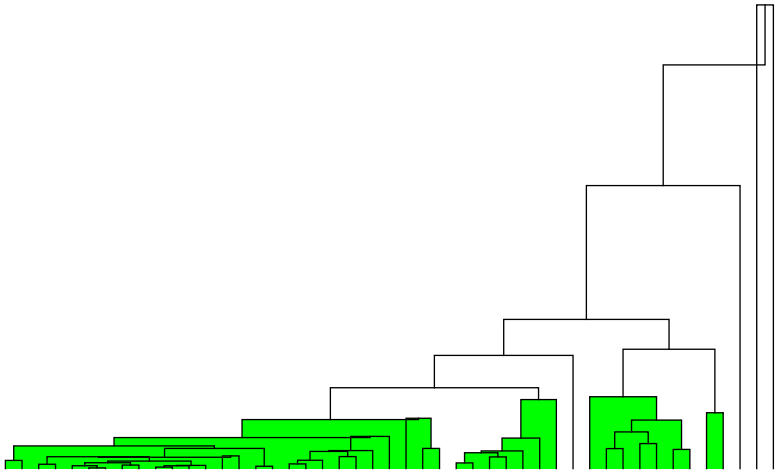
Jako vhodný počet shluků se jeví čtyři, pokud vyloučíme příliš vzdálené segmenty *GN, LD* a *LJ*, kterým se budeme věnovat zvlášť. Následující tabulka znázorňuje počty segmentů v jednotlivých shlucích a středy jednotlivých shluků.

Tabulka 4.9 Středy jednotlivých shluků vypočítaných metodou *nejvzdálenějšího souseda*

Shluk	Počet segmentů ve shluku	Att Mean	Att SD	LL Mean	LL SD	Threat Mean	Threat SD	Reserve CoV
1	14	57%	16%	7%	6%	7%	19%	15%
2	17	31%	10%	35%	27%	8%	23%	17%
3	7	18%	7%	16%	27%	23%	93%	17%
4	6	25%	17%	17%	17%	23%	48%	25%
5	1	10%	11%	46%	42%	3%	11%	97%
6	1	1%	1%	17%	28%	22%	100%	128%
7	1	0%	0%	74%	73%	24%	126%	34%

Metoda těžiště

Obrázek 4.3 Dendrogram portfolia metodou těžiště



Poznámka: pro větší přehlednost doporučuji dendrogram v přiloženém souboru shlukovani.nb

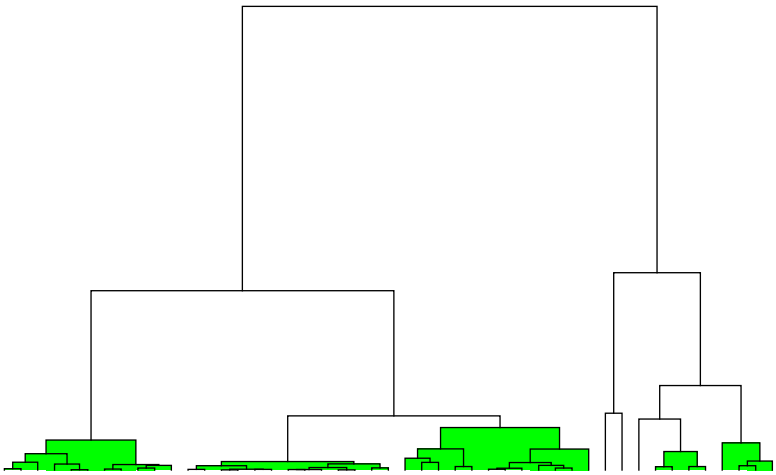
Následující tabulka znázorňuje počty segmentů v jednotlivých shlucích a středy jednotlivých shluků vypočítaných metodou *těžiště*.

Tabulka 4.10 Středy jednotlivých shluků vypočítaných metodou *těžiště*

Shluk	Počet segmentů ve shluku	Att Mean	Att SD	LL Mean	LL SD	Threat Mean	Threat SD	Reserve CoV
1	27	32%	10%	26%	19%	11%	25%	19%
2	7	70%	23%	2%	2%	5%	18%	15%
3	1	36%	20%	47%	62%	2%	6%	0%
4	7	13%	8%	19%	29%	24%	92%	19%
5	2	62%	28%	5%	12%	18%	76%	4%
6	1	10%	11%	46%	42%	3%	11%	97%
7	1	0%	0%	74%	73%	24%	126%	34%
8	1	1%	1%	17%	28%	22%	100%	128%

Wardova metoda

Obrázek 4.4 Dendrogram portfolia vypočítaný metodou *Ward*



Poznámka: pro větší přehlednost doporučuji dendrogram v přiloženém souboru shlukovani.nb

Následující tabulka znázorňuje počty segmentů v jednotlivých shlucích a středy jednotlivých shluků získaných Wardovou metodou.

Tabulka 4.11 Středy jednotlivých shluků vypočítaných metodou *Ward*

Shluk	Počet segmentů ve shluku	Att Mean	Att SD	LL Mean	LL SD	Threat Mean	Threat SD	Reserve CoV
1	11	61%	18%	4%	4%	6%	20%	16%
2	13	39%	10%	29%	17%	7%	21%	16%
3	12	19%	11%	31%	30%	15%	32%	21%
4	4	9%	3%	16%	30%	24%	106%	24%
5	4	41%	19%	12%	17%	21%	78%	7%
6	1	1%	1%	17%	28%	22%	100%	128%
7	1	10%	11%	46%	42%	3%	11%	97%
8	1	0%	0%	74%	73%	24%	126%	34%

Z dendrogramu je vidět, že jako vhodný počet shluků se jeví pět. Rozložení segmentů je při tomto počtu shluků relativně rovnoměrné (což je naprosto v

souladu s metodou minimalizace Wardova kritéria, která má tendenci vytvářet shluky o stejné velikosti, viz kapitola 2.2.1), pokud sloučíme shluky číslo 4 a 5 (analyzujeme je zvlášť, jestli jsou mezi shluky určité podobné znaky). Relativně malý počet shluků nám dovoluje provést hlubší analýzu.

Srovnání metod hierarchického shlukování

Metoda nejbližšího souseda neposkytuje příliš dobrý výsledek, neboť vytvořila „ most" mezi jednotlivými segmenty a shlukuje na nízké hladině i poměrně vzdálené segmenty. Vytváří postupně jeden velký shluk, který postupně „ absorbuje" další jednotlivé segmenty.

Metoda nejvzdálenějšího souseda rozlišuje segmenty na relativně nízké hladině, pokud vyloučíme zcela odlehlé objekty *GN, LD, LJ* potom metoda nejvzdálenějšího souseda dává přijatelnou strukturu shluků.

Metoda těžiště se nejeví jako vhodná slukovací metoda pro danné portfolio, jelikož vytváří jeden relativně velký shluk, který shlukuje i poměrně vzdálené segmenty.

Wardova metoda má tendenci vytvářet shluky stejné velikosti jako v tomto případě. Nicméně tato metoda se jeví jako vhodná z hlediska vyrovnanosti počtů segmentů v jednotlivých shlucích a poskytuje dostetečný prostor pro následující hlubší analýzu těchto shluků. Lze vybrat charakteristické segmenty pro jednotlivé shluky a následně provést diskriminační analýzu.

Klasifikace

V rámci klasifikace a výpočtu rizikového profilu jednotlivých shluků a posléze celého portfolia budeme analyzovat pouze shluky získané Wardovou metodou. Středy jednotlivých shluků znázorňuje tabulka 4.11.

Wardova metoda rozdělila segmenty do osmi rizikových shluků z toho tři shluky obsahují pouze jeden segment a proto se jimi budeme zabývat zvlášť. Za povšimnutí stojí zejména shluky 4 a 5 s vysokou míru volatility u katastrofických škod. Segmenty v této kategorii jsou pro pojišťovnu vysoce rizikové, jelikož škody sice nenastávají s vysokou četností, ale jejich dopad může být pro dannou pojišťovnu zničující (příklad WTC 9/11, kdy mnoho

amerických a evropských pojišťoven zkrachovalo v důsledku katastrofické škody). Shluky 2, 3 jsou z hlediska rizika kategorizovány jako mírně až spíše riziková. Shluk 1 považujeme za málo rizikový, většina škod je klasifikována jako malé škody a u ostatních kateogrií je škodní poměr a směrodatná odchylka nízká.

Ještě nám zbývá analyzovat odlehlé segmenty v předcházející tabulce uvedené jako shluky 6, 7, 8. Shluk 6 řadíme mezi středně rizikové ovšem s velkou mírou volatility u historických škod, tento segment je složen ze skupin smluv s dlouhým koncem, tedy škody jsou dohlašovány s několika letým zpožděním. Shluky 7 a 8 jsou klasifikovány jako velmi rizikové s vysokou mírou volatility u katastrofických škod. Shluk 7 je převážně ovlivněn velkými škodami, kde rezervové riziko je značné a v důsledku toho je pojišťovna nucena držet velký kapitál kryjící riziko rezerv (typicky k tomu dochází u pojištění s dlouhým koncem, např. *Casualty*, zanedbaní stavebních postupů a následné zřícení domu po několika letech). Shluk 8 se vyznačuje vysokou mírou volatility u velkých a katastrofických škod, což z něj činí vysoce rizikový shluk.

Rizikové míry ve shlucích

Další otázkou je jak stanovit celkové riziko jednotlivých shluků a jakou míru rizika k tomu použít. Jako vhodná míra rizika se jeví výpočet TVaR *(Tail Value-at-Risk)* pro jednotlivé shluky a jejich příspěvek do celkové TVaR portfolia. TVaR je koherentní mírou rizika tudíž můžeme jednotlivé shluky porovnávat pomocí této míry rizika.

Celkový TVaR pro naše portfolio vypočítáme následovně, pro každý segment máme simulaci celkové škody, výpočet simulace celkové škody pro jednotlivé segmenty byl proveden v [3]. Mezi simulacemi celkové škody jednotlivých segmentů aplikujeme *Iman-Conover* algoritmus, jak je popsan v [10], abychom dostali požadovanou korelační strukturu mezi jednolivými segmenty a následně vypočítáme TVaR na hladině spolehlivosti 99.5% pro celé portfolio. Dále ještě vypočítáme příspěvek jednotlivých segmentů do celkové TVaR, jako průměrnou hodnotu škod ze stejné simulace jako škoda celkového portfolia uvažovaná při výpočtu celkové hodnoty TVaR ku celkové výši TVaR pro celé portfolio (Výpočet celkové hodnoty TVaR a jednotlivých příspěvků včetně použité korelační matice lze najít v přiloženém souboru *TVaR.xls*).

Budeme měřit následující rizika:

$TVaR$ jednotlivých shluků na hladině spolehlivosti 99,5% definujeme jako

$$TVaR^{shluk} = E(X_{shluk}|(\sum_{segment \epsilon shluk} X_{segment}) > VaR^{shluk}_{99.5\%}), \tag{4.18}$$

kde X_{shluk} je součet škod jednolivých segmentů.

Příspěvek shluku do celkového TVaR portfolia je pak dán jako

$$ContributionToTotal^{shluk} = \sum_{segment \epsilon shluk} Contribution_{segment} \tag{4.19}$$

Tabulka 4.12 Výpočet rizikových měr pro portfolio, při použití shlukovací metody *Ward*:

Shluk	TVaR (v mil.)	Pojistné (v mil.)	Contribution to Total
1	1 724	445	0.13
2	3 450	752	0.27
3	5 961	999	0.48
4	438	45	0.03
5	572	71	0.02
6	391	36	0.02
7	955	129	0.04
8	733	75	0.04
Celkem portfolio	12 028	2549	1

Z tabulky je vidět, že celkové riziko společnosti je nevíce ovlivěno druhým a třetím shlukem, což znamená středně rizikové portfolio, jelikož jsou tyto shluky převážně ovlivněny vysokými škodami. Nutné je ovšem dodat, že v tomto výpočtu jsme zcela ignorovali zajištění, které jako takové vede ke snížení TVaR a tím snížení celkové rizikovosti.

4.2.2 Metoda k-průměrů

V této sekci provedeme analýzu danného portfolia metodou k-průměrů a následně provedeme výpočet kapitálu a rizikových měr v jednotlivých shlucích. Jako první se zaměříme na určení vhodného počtu shluků.

Tabulka 4.13 Vyčíslení minimálního počtu iterací

Číslo iterace	Počet shluků	Procento variace	Změna procenta
2	2	69.99	30.01
9	3	51.63	18.36
11	4	43.05	8.58
16	5	35.92	7.13
21	6	29.47	6.45

Procento variace udává procento celkové sumy čtverců pro danný počet shluků, kdyby vůbec nedošlo ke shlukování. Pro jeden shluk je tento ukazatel roven 1 a pro maximalní roztřídění (každý segment tvoří vlastní shluk) je tento ukazatel roven 0. Jako vhodný počet shluků se v tomto případě počet čtyři, jelikož křivka změny procenta sumy čtverců se v tomto bodě láme, tudíž konvergence minimalizace kriteriální funkce významně klesla (tzn. příspěvek dalšího shluku do minimalizace kriterialní funkce klesá). Relativně malý počet shluků poskytuje prostor pro hlubší analýzu jednotlivých shluků a nalezení jejich charakteristických znaků.

Tabulka 4.14 Těžiště jednotlivých shluků - střední hodnoty proměnných v jednotlivých shlucích

Proměnná	Shluk 1	Shluk 2	Shluk 3	Shluk 4
AttMean	58%	31 %	12 %	3 %
AttSD	18%	10 %	7 %	4 %
LLMean	6%	33%	17%	45 %
LLSD	6%	25%	23%	47%
Threat Mean	8 %	8%	25%	16%
Threat SD	26%	23%	76%	79%
ReserveCoV	13 %	18%	21%	86%
Počet objektů ve shluku	15	19	10	3

Tabulka 4.15 Výsledek shlukování metodou *k-průměrů*:

Shluk	Počet segmentů ve shluku	Segmenty ve shluku
1	15	AD, IF, KI, ML, MR, MU, NA, NY, OB, OE, TF, TW, VP, WD, WW
2	19	BD, CE, CW, CY, EC, FS, FW, IN, JC, KV, LB, RM, UE, UN, WN, XH, XS, XX, ZA
3	10	CG, DJ, HX, JI, KF, QB, QE, QM, SW, YK
4	3	GN, LD, LJ

Metoda k-průměrů rozdělila naše portfolio do tří početně dobře zastoupených shluků a odhalila odlehlé segmenty *GN, LD, LJ*, kterým přiřadila samostatný shluk.

Rizikové charakteristiky v jednotlivých shlucích dostáváme dle vztahů (4.18) a (4.19).

Tabulka 4.16 Rizikové charakteristiky v jednotlivých shlucích vypočítaných metodou *k-průměrů*

Shluk	TVaR (v mil.)	Pojistné (v mil.)	Contribution to Total
1	2 156	559	0.17
2	5 376	1 182	0.4
3	4 934	658	0.39
4	901	151	0.06
Celkem portfolio	12 028	2 549	1

4.2.3 Metoda optimálních středů - medoidů

V této sekci provedeme analýzu portfolia metodou medoidů popsané v kapitole 2. V první řadě se zaměříme na určení vhodného počtu shluků. K tomuto účelu provedeme výpočet minimální průměrné vzdálenosti pro různé počty shluků včetně příslušného výpočtu siluety[1]. Poté analyzujeme výsledky shlukování pro ruzné počty shluků a na základě hodnoty nalezené minimální

[1]Silueta - statistické kritérium popsané v kapitole 2.3.2

průměrné vzdálenosti a hodnoty siluety vybereme vhodný počet shluků.

Tabulka 4.17 Výpočet siluety jednolivé počty shluků

Počet shluků	Nalezená minimální průměrná vzdálenost	Silueta s
2	106	0,307
3	58	0,265
4	40	0,204
5	29	0,193
6	22	0,185

Z této tabulky plyne, že optimální počet shluků dle siluety je roven 2, neboť maximální hodnota siluety indikuje nejlepší počet shluků, jak je uvedeno v [9]. Ovšem nesmíme zapomínat, že naším cílem je minimalizace vzdáleností mezi segmenty uvnitř shluku. Z tohoto hlediska zvolíme jako optimální počet shluků 3, jelikož mezi 2 a 3 shluky dojde k výraznému snížení nalezené minimální průměrné vzdálenosti.

Provedli jsme analýzu našeho portfolia metodou medoidů pro tři shluky. Následující tabulka znázorňuje medoidy jednolivých shluků.

Tabulka 4.18 Medoidy portfolia

Proměnná	shluk 1	shluk 2	shluk 3
AttMean	54%	20%	31%
AttSD	13%	8%	12%
LLMean	11%	24%	36%
LLSD	5%	26%	21%
Threat Mean	6%	30%	8%
Threat SD	17%	73%	21%
ReserveCoV	22%	20%	20%
Medoid	TF	JI	IN

Dále ještě uvedeme podrobný výsledek shlukování jednotlivých segmentů a včetně kvality jejich přiřazení do shluků.

Tabulka 4.19 Rozdělení objektů do jednolivých shluků

Segment	Index nejbližšího shluku	Index nejbližšího souseda	Průměrná vzdálenost k němu	Průměrná vzdálenost souseda	Silueta	Čárový diagram
NA	1	3	20.64	35.77	0.423	|IIIIIIIIIIIIIIIIIII
TW	1	3	22.35	38.35	0.4171	|IIIIIIIIIIIIIIIIIII
IF	1	3	19.95	33.96	0.4125	|IIIIIIIIIIIIIIIIIII
KI	1	3	25.36	41.49	0.3888	|IIIIIIIIIIIIIIIII
WW	1	3	21.28	34.06	0.3752	|IIIIIIIIIIIIIIIII
MR	1	3	23.18	35.66	0.35	|IIIIIIIIIIIIIII
TF	1	3	18.16	23.91	0.2403	|IIIIIIIIIII
NY	1	3	38.78	50.52	0.2324	|IIIIIIIIIII
MU	1	3	19.8	21.83	0.0931	|IIIII
WD	1	3	28.67	31.56	0.0916	|IIIII
ML	1	3	22.3	24.27	0.081	|IIII
OB	1	3	33.75	35.25	0.0425	|II
AD	1	3	21.54	21.3	-0.0113	|
VP	1	3	20.43	19.31	-0.0546	|
OE	1	3	20.58	17.66	-0.142	|
SW	2	3	33.16	51.33	0.354	|IIIIIIIIIIIIIIII
KF	2	3	40.82	60.95	0.3303	|IIIIIIIIIIIIIII
DJ	2	3	37.81	54.45	0.3056	|IIIIIIIIIIIII
LD	2	3	55.39	71.27	0.2228	|IIIIIIIIII
QE	2	3	31.66	40.66	0.2212	|IIIIIIIIII
LJ	2	3	62.35	78.03	0.2009	|IIIIIIIII
QB	2	3	36.61	40.43	0.0947	|IIIII
JI	2	3	31.35	33.05	0.0514	|III
CG	2	3	36.36	29.01	-0.2021	|
QM	2	3	36.82	29.06	-0.2109	|
GN	2	3	63.85	49.11	-0.2308	|
YK	2	3	41.73	31.41	-0.2474	|
CW	3	1	15.14	34.67	0.5631	|IIIIIIIIIIIIIIIIIIIIIIIIII
IN	3	1	12.75	28.93	0.5593	|IIIIIIIIIIIIIIIIIIIIIIIIII
UE	3	1	16.93	37.45	0.5479	|IIIIIIIIIIIIIIIIIIIIIIIII

ad) Tabulka 4.19

Segment	Index nejbližšího shluku	Index nejbližšího souseda	Průměrná vzdálenost k němu	Průměrná vzdálenost souseda	Silueta	Čárový diagram
UN	3	1	19.13	39.29	0.5131	|IIIIIIIIIIIIIIIIIIIIIIIII
BD	3	1	14.84	30.42	0.5123	|IIIIIIIIIIIIIIIIIIIIIIIII
EC	3	1	19.18	39.3	0.5118	|IIIIIIIIIIIIIIIIIIIIIIIII
FW	3	1	19.07	37.56	0.4924	|IIIIIIIIIIIIIIIIIIIIIII
XX	3	1	18.29	35.6	0.4863	|IIIIIIIIIIIIIIIIIIIII
ZA	3	1	13.81	25.35	0.4551	|IIIIIIIIIIIIIIIIIII
KV	3	1	14.17	25.15	0.4366	|IIIIIIIIIIIIIIIIIII
XS	3	1	14.69	25.91	0.433	|IIIIIIIIIIIIIIIIIII
WN	3	1	16.05	27.62	0.4188	|IIIIIIIIIIIIIIIIII
RM	3	1	14.15	23.61	0.4005	|IIIIIIIIIIIIIIIIII
XH	3	1	15.52	24.92	0.3774	|IIIIIIIIIIIIIIIII
FS	3	1	21.15	33.42	0.3671	|IIIIIIIIIIIIIIIII
JC	3	1	16.84	26.48	0.364	|IIIIIIIIIIIIIIIII
CY	3	1	30.28	45.81	0.339	|IIIIIIIIIIIIIIII
LB	3	1	17.05	24.02	0.2901	|IIIIIIIIIIIIII
HX	3	1	23.05	32.38	0.2882	|IIIIIIIIIIIII
CE	3	1	16	21.97	0.2716	|IIIIIIIIIIIII

Výstup z programu NCSS.

Index nejbližšího shluku značí pořadové číslo shluku, do kterého byl tento segment zařazen.
Index nejbližšího souseda značí pořadové číslo nejbližšího shluku vůči vlastnímu shluku, do kterého byl segment zařazen.
Průměrná vyzdálenost k němu značí průměrnou vzdálenost tohoto segmentu k ostatním segmentům umístěným ve stejném shluku, jde o veličinu a ve výpočtu *siluety* (viz. kapitola 2).
Průměrná vzdálenost souseda značí průměrnou vzdálenost tohoto segmentu vůči objektům v nejbližším sousedním shluku, jde o veličinu b ve výpočtu *siluety*.

Z předchozí tabulky vyplývá, že segmenty byly přiřazeny do jednotlivých

shluků na hranici věrohodnosti. Některé objekty byly zařazeny do jednotlivých shluků dokonce pod hranicí věrohodnosti, v tomto případě je na nás jestli se rozhodneme změnit strukturu shlukování - použití více shluků, nebo případně takové segmenty z analýzy vyloučit úplně. Naším cílem je přiřazení všech segmentů do jednotlivých shluků, proto tomto případě se jako vhodné jeví určení charakteristických segmentů pro jednotlivé shluky a provedení diskriminační analýzy, tento postup je uveden v kapitole 4.2.5.

Rizikové charakteristiky jednotlivých shluků dostáváme dle vztahů (4.18) a (4.19).

Tabulka 4.20 Rizikové charakteristiky jednotlivých shluků vypočítaných metodou *medoidů*

Shluk	TVaR (v mil.)	Pojistné (v mil.)	Contribution to Total
1	2 156	559	0.17
2	5 348	750	0.43
3	5 572	1 242	0.42
Celkem portfolio	12 028	2 549	1

4.2.4 Fuzzy shlukování

Fuzzy shlukování se od předchozích shlukovacích metod liší zejména tím, že přiřazuje každému segmentu pravděpodobnost příslušnosti do jednotlivých shluků oproti předchozím metodám, které přiřazovaly segmenty jednoznačně do příslušných shluků. U každého segmentu určíme pravděpodobnost příslušnosti do jednotlivých shluků a na základě této informace se rozhodneme do jakého shluku tento segment náleží, případně nám dovoluje analyzovat do jaké míry danné segmenty charakterizují příslušný shluk.

Jako první se opět zaměříme na určení vhodného počtu shluků, k tomuto účelu spočítáme pro jednotlivé počty shluků průměrnou siluetu, Dunnův a Kaufmanův rozdělovací koeficient.

Tabulka 4.21 Výsledek výpočtu siluety, Dunnova a Kaufmanova koeficientu

Počet shluků	Průměrná silueta	F(U)	$F_c(U)$	D(U)	$D_c(U)$
2	0.340989	0.5368	0.0737	0.3027	0.6055
3	0.256444	0.3804	0.0707	0.4864	0.7295
4	0.11719	0.2907	0.0542	0.6172	0.8229
5	0.113878	0.2233	0.0291	0.683	0.8538
6	-0.026218	0.1957	0.0349	0.7385	0.8862

Jak bylo uvedeno v teoretické sekci vhodný počet shluků by měl dosahovat minimální hodnoty Dunnova koeficientu $F_c(U)$ a maximální hodnoty Kaufmanova rozdělovacího koeficientu $D_c(U)$. V našem případě se jeví jako vhodný počet shluků dva nebo tři. Pro další analýzu jsme zvolili jako vhodný počet shluků tři. Tento počet jsme zvolili z následujícího důvodu. Velikost siluety pro tři shluky je ještě přijatelná a mezi dvěma a třemi shluky dojde k významnému nárůstu velikosti normovaného Kaufmanova rozdělovacího koeficientu $D_c(U)$.

Provedeme analýzu portfolia pro tři shluky.

Tabulka 4.22 Středy jednotlivých shluků pro Fuzzy shlukováním.

Proměnná	Shluk 1	Shluk 2	Shluk 3
AttMean	31%	3%	45%
AttSD	12%	1%	8%
LLMean	36%	14%	16%
LLSD	21%	25%	9%
Threat Mean	8%	27%	8%
ThreatSD	21%	103%	13%
ReserveCoV	20%	35%	15%
Střed shluku	IN	SW	VP
Počet segmentů ve shluku	20	13	14

Z tabulky je patrné, že shluky jsou dělené dle převažujícího rizika. V prvním shluku, převažují segmenty primárně ovlivněné velkými škodami, druhý shluk se vyznačuje převahou katastrofických škod a ve třetím shluku převažují segmenty z velkou mírou malých škod.

Tabulka 4.23 Procentuální zastoupení jednotlivých segmentů v danných shlucích.

Segment	Shluk 1	Shluk 2	Shluk 3
AD	40.3%	16.32%	43.39%
BD	44.29%	16.77%	38.94%
CE	44.45%	9.31%	46.24%
CG	26.5%	48.76%	24.75%
CW	43.15%	19.43%	37.42%
CY	38.55%	24.52%	36.93%
DJ	20.66%	59.13%	20.21%
EC	38.58%	27.17%	34.25%
FS	39.93%	22.77%	37.3%
FW	40.87%	22.07%	37.06%
GN	34.26%	32.47%	33.27%
HX	37.88%	25.78%	36.34%
IF	37.73%	19.73%	42.54%
IN	46.96%	12.27%	40.77%
JC	42.24%	18.3%	39.46%
JI	19.48%	61.89%	18.62%
KF	21.45%	57.53%	21.02%
KI	37.47%	21.37%	41.16%
KV	46.07%	10.22%	43.72%
LB	43.29%	14.39%	42.32%
LD	26.52%	47.71%	25.77%
LJ	28.8%	42.81%	28.39%
ML	39.02%	18.83%	42.15%
MR	38.89%	17.69%	43.42%
MU	41.32%	11.45%	47.23%
NA	38.33%	18.3%	43.36%
NY	30.99%	36.91%	32.1%
OB	29.07%	41.55%	29.38%
OE	42.81%	10.07%	47.11%
QB	21.36%	57.71%	20.93%
QE	19.76%	61.12%	19.12%
QM	29.7%	42.08%	28.22%
RM	45.97%	9.69%	44.34%
SW	19.07%	62.35%	18.58%
TF	40.41%	12.52%	47.07%

ad) Tabulka 4.23

Segment	Shluk 1	Shluk 2	Shluk 3
TW	37.71%	20.36%	41.93%
UE	41.96%	21.11%	36.92%
UN	37.79%	28.82%	33.39%
VP	42.23%	10.31%	47.46%
WD	39.35%	18.82%	41.82%
WN	44.97%	11.87%	43.17%
WW	36.59%	22.97%	40.45%
XH	45.27%	10.22%	44.51%
XS	45.73%	10.67%	43.6%
XX	38.85%	26.45%	34.71%
YK	31.07%	38.52%	30.41%
ZA	46.25%	10.46%	43.29%

Jelikož není žádný segment jednoznačně přiřazen do jedinného shluku, je potřeba upravit výpočet TVaR v jednotlivých shlucích na základě procentuálního zastoupení segmentů v jednotlivých shlucích. Pro tento účel uvedeme dva odlišné výpočty pro rozdělení škod mezi jednotlivé shluky.

I. Poměrné zastoupení segmentů ve shlucích

Malé, historické a katastrofické škody rozdělíme proporcionálně mezi jednotlivé shluky. Velké škody modelujeme na individuální bázi, pro tento druh škod tudíž můžeme využít přístup přiřazení jednotlivých škod do shluků dle následujícího klíče.

1. Pro každou škodu simulujeme náhodnotu veličinu **T** rovnoměrně rozdělenou na intervalu (0,1)

2. Pokud je $0<T<P[Segment=1]$ potom je škoda přiřazena do prvního shluku, $P[Segment=1]<T<P[Segment=1]+P[Segment=2]$ potom je přiřazena do druhého shluku atd.

V případě katastrofických škod musíme rozlišovat zda-li se jedná o jednu událost, která významně ovlivní pouze jeden segment, nebo zda-li škoda nastává ve více segmentech. Pro účely výše zmíněného postupu ovšem potřebujeme mít k dispozici informaci o jednotlivých katastrofách a jejich vliv na portfolio. Z tohoto důvodu uvedeme další metodu, která je založena na

proporcionálním rozdělení celkové škody jednotlivých segmentů do shluků. Touto metodou dostáváme následující tabulku rizikových charakteristik pro jednotlivé shluky

Tabulka 4.24 Rizikové charakteristiky pro jednotlivé shluky

Shluk	TVaR (v mil.)	Pojistné (v mil.)	Contribution to Total
1	4 380	959	0.36
2	3 401	649	0.29
3	4 289	942	0.37
Celkem portfolio	12 028	2 549	1

II. Jednoznačné přiřazení

Předpokládáme, že každý segment je přiřazen jednoznačně do toho shluku pro který má největší pravděpodobnostní zastoupení. Celkem dostáváme následující tabulku rizikových charakteristik pro jednotlivé shluky

Tabulka 4.25 Rizikové charakteristiky pro jednotlivé shluky

Shluk	TVaR (v mil.)	Pojistné (v mil.)	Contribution to Total
1	5 435	1 212	0.4
2	5 322	693	0.43
3	2 487	645	0.19
Celkem portfolio	12028	2549	1

Druhý přístup zvýšil rozdíly mezi jednotlivými shluky, ale ignoruje fakt, že objekty jsou přiřazeny napříč shluky.

4.2.5 Srovnání shlukovacích metod

Ke srovnání shlukovacích metod uvedených výše využijeme výsledky výpočtu rizikových charakteristik.

Následující tabulka[2] znázorňuje výsledky příspěvku do celkové hodnoty TVaR portfolia, pro jednotlivé metody.

Tabulka 4.26 Rizikové charakteristiky pro jednotlivé shlukovací metody a jejich shluky

Shluk	K-průměrů	Metoda medoidů[2]	Fuzzy-jednoznačné[2]	Fuzzy - % zastoupení[2]
1	0.17	0.17	0.19	0.36
2	0.4	0.42	0.4	0.37
3	0.39	0.43	0.43	0.29
4	0.06			

Uvedené metody shlukují jednotlivé segmenty do shluků s podobnou struktutou rizikových charakteristik. Fuzzy shlukování s percentuálním zastoupením zmenšuje rozdíly mezi rizikovými charakteristikami shluků, což je výsledek, který jsme očekávali. Srovnání námi použitých shlukovacích metod ukázalo, že celkové riziko našeho portfolia je řízeno segmenty ovlivněnými velkými škodami. Toto zjištění je důležité pro výběr vhodného typu zajištného programu.

4.2.6 Shlukovací funkce

Jedna z dalších možností jak shlukovat segmenty do rizikových shluků vychází z předchozích výpočtů medoidů jednotlivých shluků.

I. Kapitál na pojistné

Princip této metody spočívá ve výpočtu kapitálu na pojistné na hladině spolehlivosti 99.5% pro jednotlivé segmenty a následné přiřazení segmentů do shluků dle nejmenší vzdálenosti ke předem stanoveným medoidům jednotlivých shluků. Kapitál na pojistné je definován jako:

$$RizikoPojistne^{segment} = \frac{Kapital_{99.5}^{segment}}{Pojistne^{segment}}. \tag{4.20}$$

Přiřazení do jednotlivých shluků je pak definováno jako

$$CisloShluku^{i} = min\{j; |(RizikoPojistne^{medoid^{j}} - RiskOnPrem^{segment})|\}, \tag{4.21}$$

[2]Přerovnali jsme shluky dle převažujícího rizika. První shluk je vždy s převahou malých škod, druhý s převahou velkých škod a třetí s převahou katastrofických škod

čili na základě nejmenší euklidovské vzdálenosti k jednotlivým medoidům.

K výpočtu využijeme medoidy určené fuzzy shlukováním. Výpočtem dle (4.21) dostáváme následující strukturu segmentů v jednotlivých shlucích:

Tabulka 4.27 Segmenty v jednotlivých shlucích

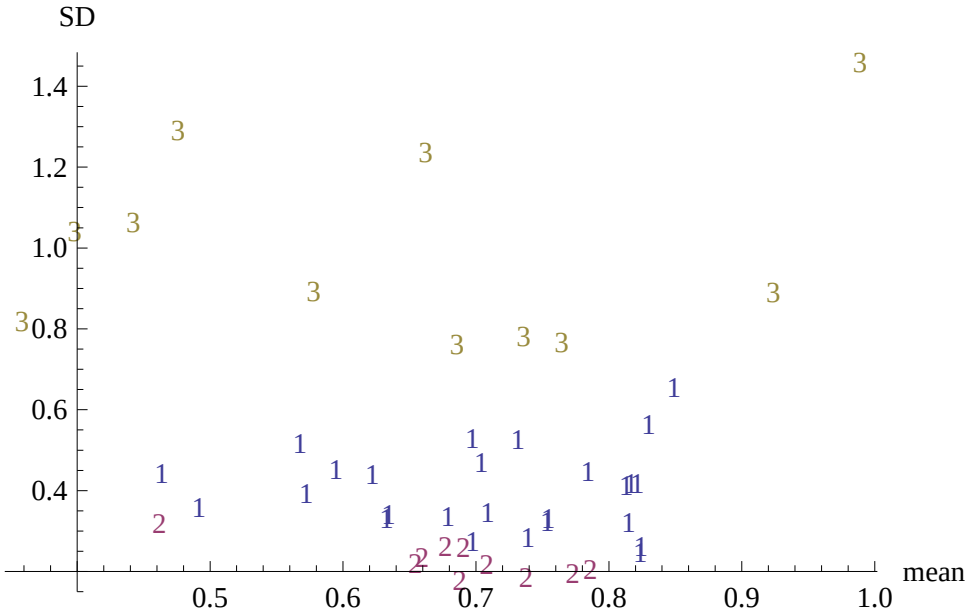
Shluk	Medoid	Segmenty ve shluku
1	IN	BD, CW, CY, EC, FS, GN, IF, IN, JC, LB, RM, UE, UN, XX,
2	VP	CE, FW, KI, KV, MR, MU, NA, OE, TF, TW, VP, WD, WN, WW, XH, XS, ZA,
3	SW	CG, DJ, HX, JI, KF, LD, LJ, ML, NY, OB, QB, QE, QM, SW, YK,

Kde první skupina shlukuje středně rizikové segmenty kolem kapitálu na pojisté $IN = 3.26$. Druhou skupinu považujeme za méně rizikovou shlukuje objekty okolo objektu $VP = 3.22$ a třetí skupina je jako taková považována za vysoce rizikovou, jelikož objekty v této skupině mají nejblíže k segmentu s rizikovým parametrem $SW = 8.48$. Segmenty IN a VP mají téměř podobnou hladinu kapitálu na pojistné a vytváří pomyslnou hranici mezi málo a středně rizikovými segmenty.

II. Statistika celkové ztráty

Tato metoda vychází z výpočtu střední hodnoty a směrodatné odchylky celkové ztráty z upisovacího rizika (*Underwriting risk*) dělené celkovým přijatým pojistným segmentu. Cílem této metody je stanovení rizikovosti jednotlivých segmentů pro účely dalšího upisování těchto rizik a případně pro účely upravení výše pojistného a rizikových přirážek. Metoda rozřazuje segmenty do shluků na základě nejmenší vzdálenosti k předem stanovených medoidům těchto shluků. K výpočtu využijeme medoidy vypočítané při *Fuzzy* shlukování. Výsledné rozdělení segmentů mezi jednotlivé shluky si ukážeme na grafu *(Pro výpočet vzdálenosti od jednotlivých medoidů byla použita euklidovská vzdálenost)*

Obrázek 4.5 Rozložení shluky vypočítaných pomocí statistiky celkové ztráty



Tabulka 4.28 Segmenty v jednotlivých shlucích

Shluk	Medoid	Segmenty ve shluku
1	IN	AD, BD, CW, CY, EC, FS, FW, GN, IF, IN, JC, KI, KV, LB, ML, QM, RM, TW, UE, UN, WN, WW, XH, XS, XX, YK,
2	VP	CE, HX, MR, MU, NA, OE, TF, VP, WD, ZA
3	SW	CG, DJ, JI, KF, LD, LJ, NY, OB, QB, QE, SW

První shluk *IN* je s převahou velkých škod a je pro pojišťovnu středně riziková. Druhý shluk *VP* je s převahou malých škod a nízkou směrodatnou odchylkou a tento shluk je pro nás málo rizikový. Třetí shluk *SW* je s převahou katastrofických a velkých škod s vysokou směrodatnou odchylkou a jako takový je pro pojišťovnu vysoce rizikový.

Diskriminační analýza

Pro účely diskriminační analýzy využijeme výpočet ze shlukování metodou medoidů, jelikož výpočet siluety pro některé segmenty ukázal jejich

nevhodné přiřazení do shluků. Pro každý shluk vybereme určitý počet referenčních segmentů na kterých provedeme diskriminační analýzu. Pro jednotlivé shluky jsme určili následující referenční segmenty:

Tabulka 4.28 Referenční segmenty v jednotlivých shlucích

Shluk	Referenční segmenty ve shluku
1	IF, KI, MR, NA, NY, TF, TW, WW
2	DJ, KF, LD, LJ, QE, SW
3	RM, WN, XS, KV, ZA, XX, FW, EC, BD, UN, UE, IN, CW

Vypočítáme matici vnitroskupinové variability **E** dle (2.3) a matici meziskupinové variability **B** (2.4) a dostáváme matici $\mathbf{BE^{-1}}$.

$$\mathbf{BE^{-1}} = \begin{pmatrix} 8.9 & 1 & -0.6 & 2.3 & 12.7 & -4.7 & 0.3 \\ 3.3 & 0.3 & -0.3 & 0.9 & 4.6 & -1.7 & 0.1 \\ -4.8 & 0.4 & 2.1 & -2 & -6.1 & 1.1 & -0.5 \\ -5.3 & -0.5 & 0.5 & -1.4 & -7.4 & 2.7 & -0.2 \\ -1.7 & -0.7 & -0.8 & 0 & -2.8 & 1.7 & 0.1 \\ -7.5 & -3.8 & -4.8 & 0.4 & -12.6 & 8.4 & 0.7 \\ -3.6 & -1.1 & -1.1 & -0.3 & -5.5 & 3 & 0.1 \end{pmatrix}$$

Dále vypočítáme vlastní čísla matice $\mathbf{BE^{-1}}$ a příslušné vlastní vektory

l_1=11.2278 v_1=(-0.489, -0.175, 0.157, 0.278, 0.149, 0.728, 0.275)
l_2=4.505 v_2=(0.435, 0.173, -0.551, -0.285, 0.0874, 0.618, 0.0776)

konstanty w_r dostáváme dle (3.4) $w_1 = -0.29$ a $w_2 = -0,285$. Dále dle (3.5) vypočítáme diskriminační skóre pro středy jednotlivých shluků definovaných jako vektor průměrných hodnot v jednotlivých shlucích $\bar{x}_k = [\bar{x}_{k,1}, \bar{x}_{k,2}, ..., \bar{x}_{k,7}]$ pro k=1, 2, 3 a dále vypočítáme diskriminační skóre pro jednolivé segmenty.

Na základě minimální euklidovské vzdálenosti k jednolivým středům přiřadíme jednotlivé segmenty pojistných smluv do shluků.

Tabulka 4.29 Výpočet diskriminačního skóre pro jednolivé segmenty a přiřazení do jednotlivých shluků

Segment	R1	R2	Shluk
$\bar{x}_1$	-0,431	0,232	1
$\bar{x}_2$	1,052	0,222	2
$\bar{x}_3$	-0,086	-0,246	3
OE	-0,286	-0,075	3
VP	-0,328	-0,089	3
AD	-0,176	0,053	1
OB	0,057	0,292	1
ML	-0,19	0,082	1
WD	-0,375	-0,017	1
MU	-0,356	-0,051	1
NY	-0,082	0,624	1
TF	-0,356	0,03	1
MR	-0,612	0,055	1
WW	-0,27	0,314	1
KI	-0,619	0,225	1
IF	-0,429	0,213	1
TW	-0,537	0,235	1
NA	-0,541	0,163	1
YK	0,119	0,024	3
GN	0,18	-0,451	3
QM	0,233	-0,015	3
CG	0,276	-0,067	3
JI	0,339	0,106	3
QB	0,325	0,236	3
LJ	0,921	0,286	2
QE	0,471	0,157	2
LD	1,079	-0,077	2
DJ	0,66	0,336	2
KF	0,764	0,362	2
SW	0,676	0,269	2
CE	-0,258	-0,142	3
HX	-0,034	-0,121	3
LB	-0,1	-0,052	3
CY	-0,203	-0,49	3
JC	-0,048	-0,054	3

ad Tabulka 4.29

Segment	R1	R2	Shluk
FS	-0,019	-0,153	3
XH	-0,267	-0,204	3
RM	-0,183	-0,145	3
WN	-0,266	-0,252	3
XS	-0,245	-0,222	3
KV	-0,217	-0,182	3
ZA	-0,18	-0,183	3
XX	0,059	-0,205	3
FW	-0,055	-0,346	3
EC	0,073	-0,322	3
BD	-0,079	-0,197	3
UN	0,122	-0,265	3
UE	-0,004	-0,349	3
IN	-0,129	-0,23	3
CW	-0,02	-0,304	3

V dalším kroku dostáváme dle (3.7) korelační koeficienty mezi kanonickými proměnnými a původními proměnnými. Korelační koeficienty určují důležitost jetnolivých proměnných pro rozdělení do jednotlivých shluků.

Tabulka 4.30 Korelační koeficienty pro jednotlivé proměnné

Proměnná	r1	r2
Att mean	-0,061	0,048
Att SD	-0,015	0,046
LL Mean	0,065	-0,053
LL SD	0,072	-0,044
Threat mean	0,073	0,036
Threat SD	0,096	0,055
Reserve CoV	0,036	-0,006

jak plyne z této tabulky, tak žádná proměnná nemá významný vliv na rozdělení segmentů do jednotlivých shluků. Celkem malý vliv na přiřazení do jednotlivých shluků má proměnná *Reserve CoV* a *Att SD*.

Pokud jde o potřebu obou kanonických proměnných pro odlišení jedno-

litvých shluků, pak provedeme Wilksův test, kdy testujeme hypotézu o nulových hodnotách obou vlastních čísel. Pro první vlastní číslo je Bartletovo testové kritérium dle (3.8)

$$V = [47 - 1 - (7 + 3)/2.(ln(1 + 11.2278) + ln(1 + 4.505))] = 172.584$$

Přibližné rozdělení je chí kvadrát s 14-ti stupni volnosti a na hladině spolehlivosti 5% vede k zamítnutí hypotézy o nulové hodnotě prvního vlastního čísla.

Pro druhé vlastní číslo je hodnota Bartletova testového kritéria dána dle

$$V = [47 - 1 - (7 + 3)/2.ln(1 + 4.505)] = 69.93$$

Přibližné rozdělení je chí kvadrát s 6-ti stupni volnosti a na hladině spolehlivosti 5% vede k zamítnutí hypotézy o nulové hodnotě druhého vlastního čísla.

Po provedení diskriminační analýzy jsme přeřadili některé nesprávně zařazené segmenty po shlukování metodou medoidů. Provedeme výpočet rizikových charakteristik v jednotlivých skupinách a provedeme srovnání s původním rozdělením segmentů.

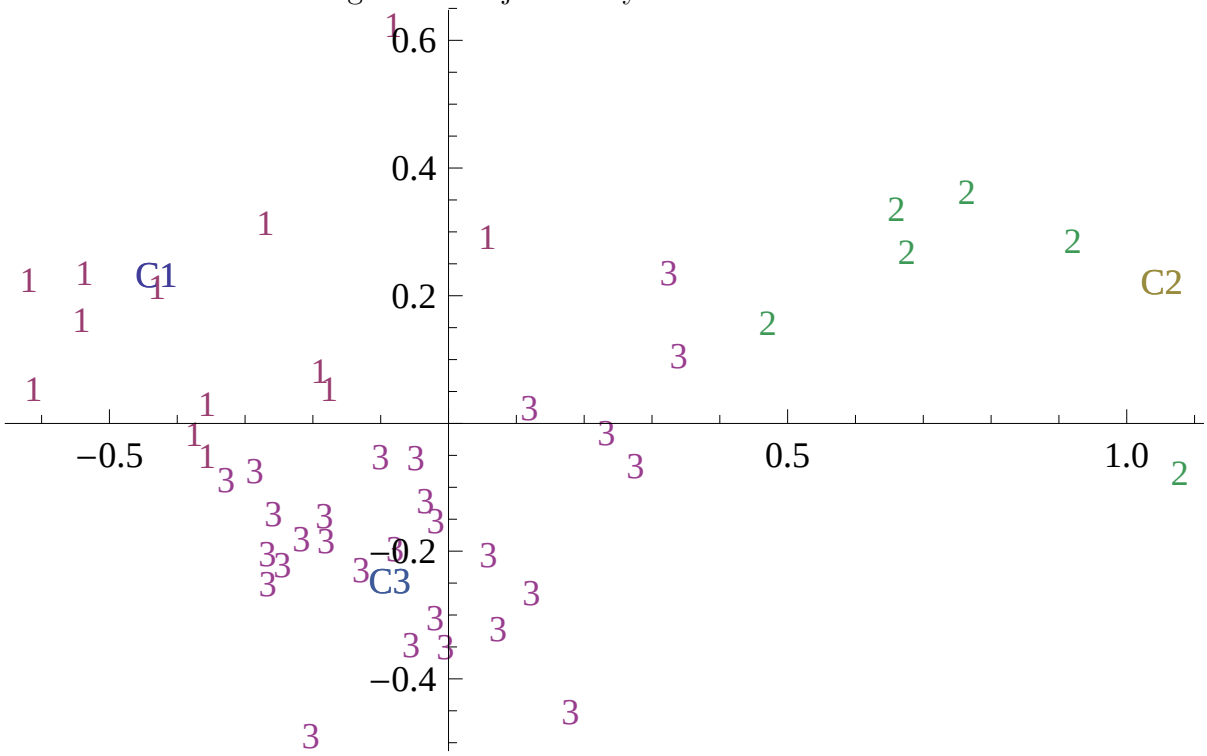
Tabulka 4.31 Srovnání rizikových charakteristik před a po diskriminační analýze

Shluk	TVaR Old	Contribution to Total Old	TVaR New	Contribution to Total New
1	2 156	0.17	1 851	0.14
2	5 348	0.43	1 219	0.08
3	5 572	0.42	9 555	0.79

Diskriminační analýza odhalila odlehlou vysoce rizikovou skupinu (převážně s katastrofickými škodami), ale na druhou stranu z důvodu malého počtu segmentů v této skupině je její vliv na celkovou rizikovost portfolia menší.

Následující graf znázorňuje centroidy jednolitvých shluků a rozdělení segmentů do jednotlivých shluků.

Obrázek 4.6 Rozdělení segmentů do jednotlivých shluků



Slukovací funkce

Následující shlukovací funkce byla použita pro přiřazení segmentů do jednotlivých shluků při diskriminační analýze.

Shluk=$min\{i; D_i\}$, kde D_i, i=1, 2, 3 je definováno jako

$$D_1 = \sqrt{(-0.431 - x^T v_1 + w_1)^2 + (0,232 - x^T v_2 + w_2)^2}$$
$$D_2 = \sqrt{(1.052 - x^T v_1 + w_1)^2 + (0,222 - x^T v_2 + w_2)^2}$$
$$D_3 = \sqrt{(-0.086 - x^T v_1 + w_1)^2 + (-0,246 - x^T v_2 + w_2)^2}.$$

Zajišťění

Ve výpočetní sekci jsme všechny výsledky uváděli v *hrubé* výši, tj. bez zajištění. Shlukovací metody rozdělily naše portfolio do shluků dle převažujícího rizika - malé škody, velké škody a katastrofické škody. Výše uvedenou ana-

lýzu portfolia lze dále využít pro potřeby zajištění, tedy výběru vhodného zajištění dannou skupinu pojistných smuv, případně pro celý rizikový shluk. V případě malých škod, se jako vhodné zajištění jeví zajištění proporcionální (kvótové, případně lze ještě uvažovat kvótové zajištění s *loss coridorem*, tedy pokud se škodní poměr dostane do určitého pásma, tak zajistitel zaplatí určitý předem dohodnutý poměr navíc). Neproporciální typ zajištění na idividuální škodní bázy (XoL zajištění) v tomto případě nemá vůbec smysl, proto se také povětšinou malé škody mohou modelovat v kumulaci.

V případě převažujících velkých škod se jako rozumná volba zajištění jeví zajištění škodního nadměrku (pokud škoda přesáhne předem stanovenou hranici, tak zajistitel platí škody nad touho hranicí do předem určeného limitu).

V případě katastrofických škod se v praxi většinou volí zajištění *Stop-Loss*[1], případně se pro potřeby zajištění volí tzv. ART *Alternative Risk Transfer* kontrakty. Mezi nejběžnější ART kontrakty patří takzvané katastrofické dluhopisy *Cat bonds*, kdy cedent platí kupón a v případě předem specifikované katastrofické události (např. tornádo, kdy celková tržní ztráta bude větší než 100 mil USD) obdrží nominální hodnotu dluhopisu.

Volba vhodného zajištění může významně snížit kapitálový požadavek a omezit rizikovost našeho portfolia. Na druhou stranu nevhodná volba zajištění nejenom, že nesníží kapitálový požadavek, ale navíc ještě snižuje profitabilitu portfolia (například pokud bychom uzavřeli XoL *Excess of Loss* zajištění s příliš vysokou prioritou na skupinu smluv s významnou převahou malých škod). V zajištění platí stejná logika jako v pojištění, tedy zajištění bychom měli uzavřít v případě, kdy případá škoda může významně ovlivnit chod pojišťovny. Více o zajištění a metodách výpočtu je možné se dočíst v [1] [2]. Výše uvedená analýza výsledků pro účely zajištění je pouze naznačení další aplikace výsledků.

Závěr

Cílem diplomové práce bylo shrnout metody výpočtu rizikových charakteristik, které se sledují a měří v pojišťovnictví na úrovni homogenních skupin pojistných smluv. Jádro textu spočívá v představení shlukovacích metod vícerozměrné statistiky a jejich následná aplikace na soubor homogenních segmentů pojistných smluv za účelem analýzy celkového rizika daného portfolia, dle přiřazení jednotlivých segmentů do shluků s odlišným rizikovým profilem.

V teoretické části diplomové práce byly popsány a rozebrány metody výpočtu rizikových charakteristik, zaměřili jsme se přitom na odhad upisovacího rizika a rizika rezerv. Pro odhad parametrů a následné modelování upisovacího rizika jsme rozdělili škody do tří kategorií - malé, velké a katastrofické. Každou z těchto kategorií upisovacího rizika jsme uvedli zvlášť, včetně popisu způsobu modelování těchto kategorií a jejich významu z hlediska zajištění. Pro účely odhadu parametrů rizika rezerv jsme popsali dvě metody a to metody *Mack* a *Bootstrap* s poukázáním na fakt, že obě metody nejsou ideální a nemůžou být používány automaticky, jelikož v některých případech je jejich použití naprosto nevhodné. V druhé části teoretické sekce jsme shrnuli používané shlukovací metody a charakteristické aspekty těchto metod. Dostatek prostoru byl přitom věnován problematice určení vhodného počtu shluků na bázi *průměrné siluety, procenta variace, Dunnova a Kaufmanova rozdělovacího koeficientu*. Následně jsme se zaměřili na různé metody shlukování objektů do shluků, včetně výpočtu *siluety* jakožto statistického testu určující vhodnost zařazení jednotlivých objektů do příslušných shluků.

Výsledkem praktické části diplomové práce jsou rizikové charakteristiky vypočítané na reálných datech pro jednotlivé homogenní segmenty pojistných smluv a následné aplikování shlukovacích metod popsaných v teoretické části. V první fázi jsme aplikovali metody hierarchického shlukování, za účelem získání představy o struktuře portfolia a vhodného počtu shluků. Výstupem shlukovacích metod je roztřídění objektů do jednotlivých shluků dle rizikových parametrů včetně výpočtu *siluety*. V další fázi jsme vybrali charakteristické segmenty pro jednotlivé shluky a aplikovali jsme diskriminační analýzu za účelem rozdělení všech segmentů do jednotlivých shluků, jelikož výpočet siluety pro některé segmenty indikoval naprosto nevhodné přiřazení. Dále jsme se zaměřili na určení rizikových charakteristik jednotlivých shluků.

Jako hlavní rizikovou charakteristiku jednotlivých shluků v práci uvádíme výpočet TVaR a příspěvek jednotlivých shluků do TVaR celého portfolia. Pro účely modelování a následného stanovení celkové TVaR jsme stanovili korelace mezi jednotlivými segmenty pojistných smluv. Problematika odhadu korelačních koeficientů je poměrně náročná, zejména při nedostatku dat a její zpracování by mohlo být skutečnou výzvou pro budoucí diplomanty, jelikož výše korelačních koeficientů může významně ovlivnit celkovou rizikovost portfolia. Jako hlavní přínost práce vidím v možnosti využití těchto metod k obsáhlé analýze portfolia pojišťovny a pro lepší pochopení, která složka rizika ovlivňuje celkovou výši kapitálu a zároveň zodpovídá otázku, který shluk/segment pojistných smluv představuje pro pojišťovnu potencionální riziko a výši tohoto rizika. Případně jdou uvedené metody využít ke srovnání rizikovosti portfolií mezi pojišťovnami.

Literatura

[1] Cipra Tomáš: *Zajištění v pojišťovnictví a jeho matematické aspekty*, Robust 2004.

[2] Cipra Tomáš: *Zajištění a přenos rizik v pojišťovnictví*, Grada, Praha, 2004.

[3] EMB Consultancy LLP: EMB Igloo Extreme, Version 4

[4] Hebak Petr: *Vícerozměrné statistické metody*, Informatorium, spol. s.r.o., 2007.

[5] Hurt Jan: *Teorie spolehlivosti*, Státní pedagogické nakladatelství, Praha, 1984.

[6] Lloyds: *ICA 2009 Minimum Standards and Guidance*, 2009.

[7] Mack Thomas: *Distribution-Free Calculation of the Standard Error of Chain Ladder Reserve Estimates*, ASTIN Bulletin 23, 1993.

[8] Mandl Petr, Mazurová Lucie: *Matematické základy neživotního pojištění*, MATFYZPRESS, Praha 1999.

[9] Meloun Milan, Militký Jiří, Hill Martin: *Počítačová analýza v příkladech*, Academia, Praha 2005.

[10] Mildenhall J. Stephen: *Correlation and Aggregate Loss Distribution With An Emphasis On The Iman-Conover Method* , CAS Working Party on Correlation, 2005.

[11] Murphy Daniel: *Chain Ledder Reserve Risk Esimators*, CAS E-Forum 2007.

[12] Wolfram Research, Inc.: Mathematica, Version 6.0, Champaign, 2007