ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE

PROVOZNĚ EKONOMICKÁ FAKULTA

Využití data miningových metod při prediktivním modelování

Autor: Ing. Václav Ulrych

Školitel: Prof. Ing. Vladimír Brabenec, CSc., Katedra statistiky

Praha 2006

Obsah

1 Úvod

Cílem většiny dnešních firem je individuální přístup ke každému zákazníkovi. Pochopení jeho přání a potřeb umožňuje podniku vytvářet na trhu takové podmínky, že pro zákazníka je jednodušší a výhodnější obchodovat právě s ním a nikoliv s konkurencí. V malých firmách a živnostech je individuální vztah se zákazníkem vytvářen všímáním si jeho potřeb, pamatováním si jeho preferencí a učením se z posledních vzájemných interakcí tak, aby byl klient při příštím kontaktu s firmou obsloužen lépe. Ve velkých firmách je četnost kontaktů zákazníka s jejími zaměstnanci mnohem nižší, což znamená obtížnou realizaci těchto činností. Intuici maloobchodníka, který pozná své zákazníky podle tváře, hlasu, zná je jménem a pamatuje si jejich zvyky, rovněž nenahradí ani anonymní osoba v call-centru či pokaždé jiný zaměstnanec na přepážce obchodního zastoupení.

Jakým způsobem tedy zabezpečit, aby i velká firma mohla disponovat detailní znalostí konkrétních klientů a tuto znalost využívat k péči a rozvoji vzájemných vztahů? Prvním krokem je zaznamenávání chování zákazníků. Tyto procesy jsou již dnes plně automatizované. Každý z nás v průběhu života vytváří konstantní proud transakčních nahrávek, jako je výběr hotovosti z bankomatu, použití klubové karty při platbě zboží v supermarketu, realizace telefonního hovoru či zapříčinění vzniku pojistné události. Shromažďování záznamů o chování zákazníků představuje pro firmy příležitost k učení. Podniky tak mohou na základě minulého chování klientů předvídat jeho budoucí vývoj. Učení však vyžaduje mnohem více než pouhé shromažďování dat. Mnoho firem shromažďuje data pouze pro účely zabezpečení provozních úkolů, jako je evidence zásob, účetnictví apod. Po splnění účelu jsou data zpravidla smazána nebo v lepším případě uložena na zálohovací pásky. Pro zavedení učení je však nutné, aby firma uměla nehomogenní data z různých zdrojů (interních i externích) užitečně využívat i pro účely sledování jejich vnitřních závislostí a odlišného chování některých jejich podmnožin.

Za účelem uchovávání různých dat v konzistentním formátu, sledování jejich změn v čase, vytváření jejich předmětově orientovaných podmnožin apod. budují firmy datové sklady. Datový sklad tak umožňuje podniku pamatovat si to, co poznal o svých klientech. Aby se však mohl poučit z této paměti, musí být data dále analyzována, pochopena a přeměněna na informace a znalosti. To je úkolem data miningu.

Tak jako datový sklad poskytuje podniku paměť, data mining poskytuje podniku inteligenci, která prochází záznamy paměti, vymýšlí pravidla, přichází s novými myšlenkami k vyzkoušení a provádí předpovědi budoucnosti. Za tímto účelem využívá široké palety algoritmů, od statistických metod až po složité algoritmy umělé inteligence. Ačkoliv řada technik data miningu je již známa i několik desetiletí, jejich masové využití pro komerční účely teprve v posledních letech je důsledkem souběhu několika faktorů.

Prvním z nich je skutečnost, že s rozšířením technologií jako jsou kreditní a bankovní karty, pokladny se snímacím zařízením v supermarketech, home shopping, elektronický transfer peněz, elektronické zpracování objednávek apod. začaly být ve všech oblastech ekonomiky produkovány dosud nevídané objemy dat. Techniky data miningu jsou vhodné právě pro takto rozsáhlé objemy. Druhým faktorem je rozvoj data warehousingu, který umožňuje transformovat a uchovávat data z různých zdrojů ve společném formátu a s jednotnou definicí datových struktur. Techniky data miningu jsou rovněž výpočetně velmi náročné. Výrazný pokles cen hardwaru umožnil společnostem za přiměřené náklady značně zkrátit dobu zpracování svých úloh. Čtvrtým činitelem, který podporuje používání technik data miningu, je stále ostřejší konkurenční prostředí na trhu. Klíčem k získání konkurenční výhody se stávají informace. Data mining a jeho techniky jsou pak nástrojem pro získání klíčových informací z nepřehledných dat. Posledním faktorem je skutečnost, že na trhu je k dispozici cenově dostupný a uživatelsky příjemný software pro komerční využití data miningové metodologie.

V této disertační práci bude použití technik data miningu prezentováno na příkladu z prostředí firmy Česká pojišťovna, a.s. Práce bude zaměřena na segmentaci jejích klientů a na procesy, které souvisí s prováděním data miningových analýz, zejména na přípravu dat pro modelování. Práce se bude rovněž zabývat sběrem dat, ETL procesy, data warehousingem a využitím výsledků data miningu v obchodní praxi.

2 Současný stav problematiky a její řešení v odborné literatuře

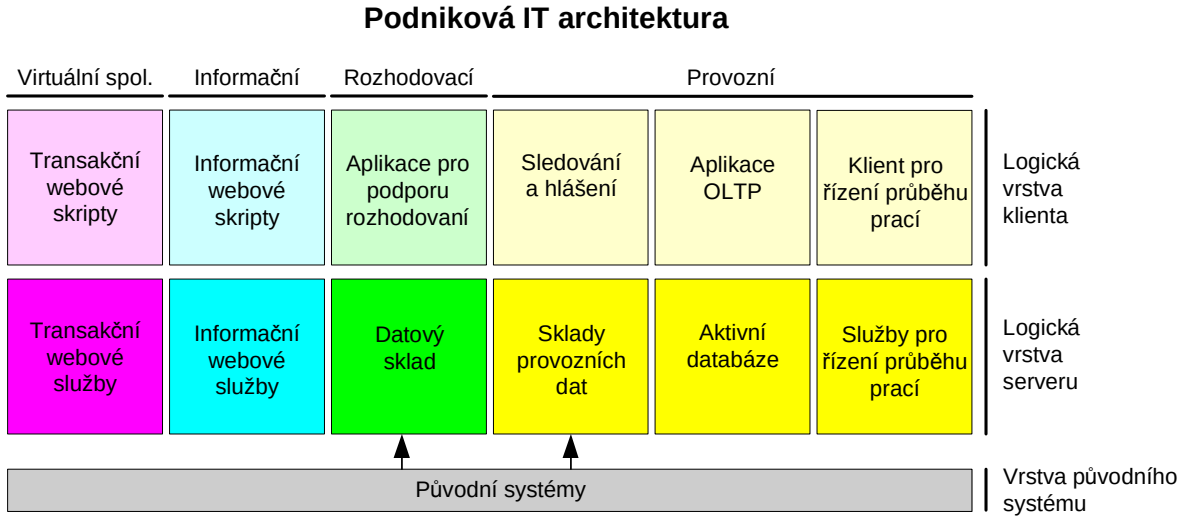
2.1 Obchodní potřeby podniku a jejich požadavky na podnikovou IT architekturu

Od zavedení počítačů v šedesátých letech minulého století byly téměř všechny provozní činnosti firem plně automatizovány. Tato automatizace změnila nejen povahu obchodu, ale i způsob života lidí. Kreditní karty, bankomaty, obchodování prostřednictvím Internetu, GSM služby a kluby věrných zákazníků jsou jen několika málo příklady toho, jak automatizace otevřela nové trhy a revolučně změnila ty stávající.

Prostřednictvím automatizace firmy efektivně uspokojují své provozní potřeby. Se stále ostřejším konkurenčním prostředím však roste význam i dalších potřeb, zejména rozhodovacích. Humphries [16] rozlišuje z obchodního pohledu následující typy podnikových potřeb:

- potřeby provozní,
- potřeby rozhodovací,
- potřeby informační,
- potřeby virtuální společnosti.

Na obrázku 1 je znázorněn přehled technologií, které podporují jednotlivé obchodní potřeby podniku.



Obr. 1 Podniková IT architektura – technologie, které podporují jednotlivé obchodní potřeby podniku [16]

Technologie používané k zabezpečení provozních potřeb podporují hladké provádění a soustavné zdokonalování každodenních operací, identifikaci a opravu chyb pomocí hlášených výjimek řízení průběhu prací (work flow) a celkové sledování provozu. Informace získané o obchodu z provozního pohledu jsou použity k dokončení či optimalizaci obchodního procesu.

Technologie používané pro podporu rozhodování a dlouhodobé plánování zásobují řídící pracovníky pohledy na podniková data z více dimenzí a s různou úrovní podrobnosti. Tyto technologie rovněž umožňují analyzovat trendy a objevovat skrytá pravidla uvnitř dat.

Technologie používané k zabezpečení informačních potřeb podniku umožňují okamžitě zpřístupnit informace velkému počtu uživatelů. Příkladem mohou být vnitropodnikové normy, organizační schémata, firemní tiskopisy a formuláře, školící materiály apod.

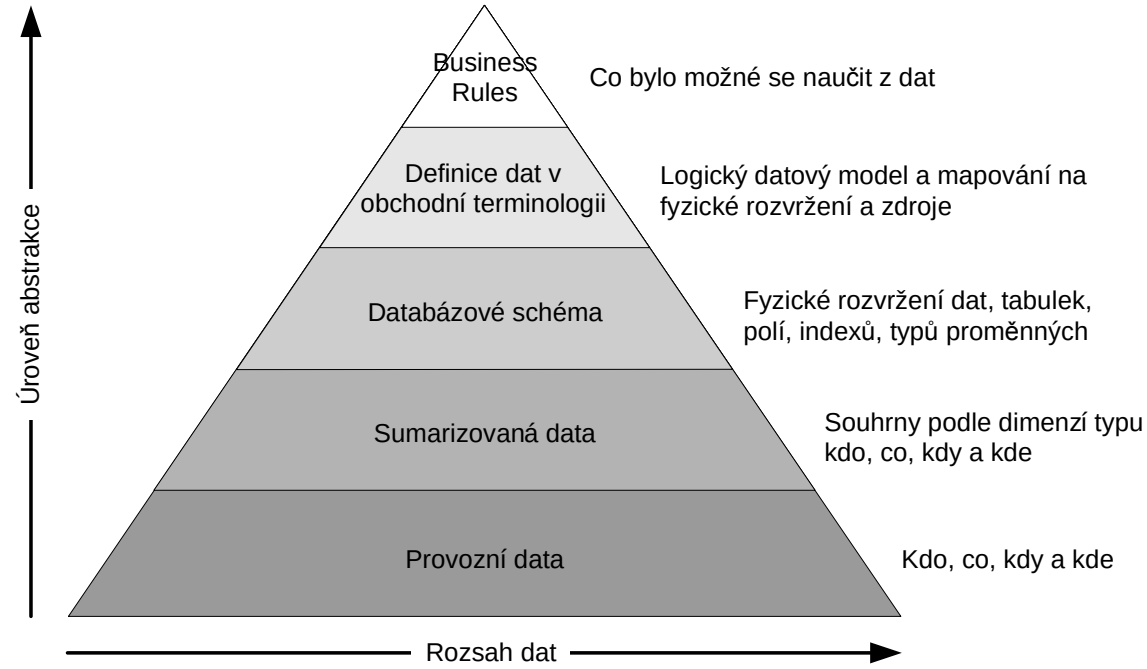
Technologie, které umožňují vytvářet strategická spojení s klíčovými dodavateli a zákazníky jsou technologiemi virtuální společnosti.

Rozdělení technologií do skupin podle typů obchodních potřeb (viz. obrázek 1) není vždy úplně striktní. Inmon [19] definuje např. sklad provozních dat jako součást systému pro podporu rozhodování (Decision Support System). Podle této definice plní sklad provozních dat funkci podpory taktických rozhodovacích procesů.

2.2 Architektura podnikových dat

V důsledku automatizace procesů vznikly ve společnostech rozsáhlé objemy dat, které jsou distribuovány do velkého množství systémů. Jejich využití pro účely získání správných informací ve správný čas je velmi obtížné. Proces spojování a uchovávání heterogenních dat ze všech částí organizace pro účely podpory rozhodování se nazývá data warehousing. Hledání informací prostřednictvím analýz vnitřních závislostí dat je pak úkolem data miningu.

Data, která jsou jádrem data warehousingu i data miningu, mohou být popsána mnoha různými způsoby. Velmi užitečný je např. pohled ve smyslu abstrakce, který je uveden na obrázku 2.



Obr. 2 Hierarchie podnikových dat. Čím více jsou data abstraktní, tím menší je jejich objem [3]

Provozní data jsou nejzákladnější formou firemních dat. Každé přepracování pojistné smlouvy, každá pojistná událost, každá bankovní transakce, každý telefonní hovor jsou zaznamenávány do nějakého provozního systému. Množství dat shromažďované těmito systémy může být enormní. Provozní data se rovněž velmi často mění, což je důsledkem přizpůsobování se provozních systémů obchodním požadavkům firmy.

Sumarizovaná data jsou první úrovní abstrakce podnikových dat. Jsou odvozena z dat provozních a představují nejčastější formu, se kterou se setkávají obchodní uživatelé. Dimenze, které se používají pro sumarizaci nebo agregaci (např. produkt či region), zpravidla představují oblasti, podle kterých je hodnocen obchodní výkon firmy.

Další úrovní abstrakce je fyzické rozvržení provozních a sumarizovaných dat (tj. názvy tabulek a polí, typy proměnných apod.). Fyzické rozvržení informuje především technické uživatele o tom, jaká data jsou k dispozici a co je možné z nich získat.

Nad úrovní fyzického rozvržení se nachází úroveň, která definuje data v obchodní terminologii. Jejím jádrem je logický datový model, který popisuje vlastnosti entit a vztahy mezi nimi (tj. vlastnosti zákazníků, produktové

hierarchie apod.). Cílem logického modelu je informovat především obchodní uživatele o obsahu databáze, o časové a místní dostupnosti reportů, způsobu jejich použití a způsobu jejich získávání.

Nejvyšší úrovní abstrakce jsou tzv. obchodní pravidla (Business Rules). Tato pravidla nepopisují struktury dat ani jejich vztahy, nýbrž odpovídají na otázky typu: proč tyto vztahy existují, jak jsou tyto vztahy aplikovány apod. Business rules mají úzký vztah k data miningu. Výsledkem některých jeho technik (např. analýzy nákupního koše nebo rozhodovacích stromů) jsou jednoznačná pravidla, která by mohla být považována za business rules.

2.3 Datový sklad

Inmon [18] definuje datový sklad jako kolekci sjednocených, předmětově orientovaných databází navržených za účelem poskytovat informace požadované pro rozhodování.

Datový sklad obsahuje data z mnoha podnikových provozních systémů, přičemž může být plněn i externími daty (např.: daty z komerčních databází, které obsahují informace o potenciálních zákaznících a které podnik nakupuje od jiných firem). Tato data pak transformuje podle svých možností do podobné logiky, čímž nabízí svým uživatelům sjednocený pohled na celopodniková data.

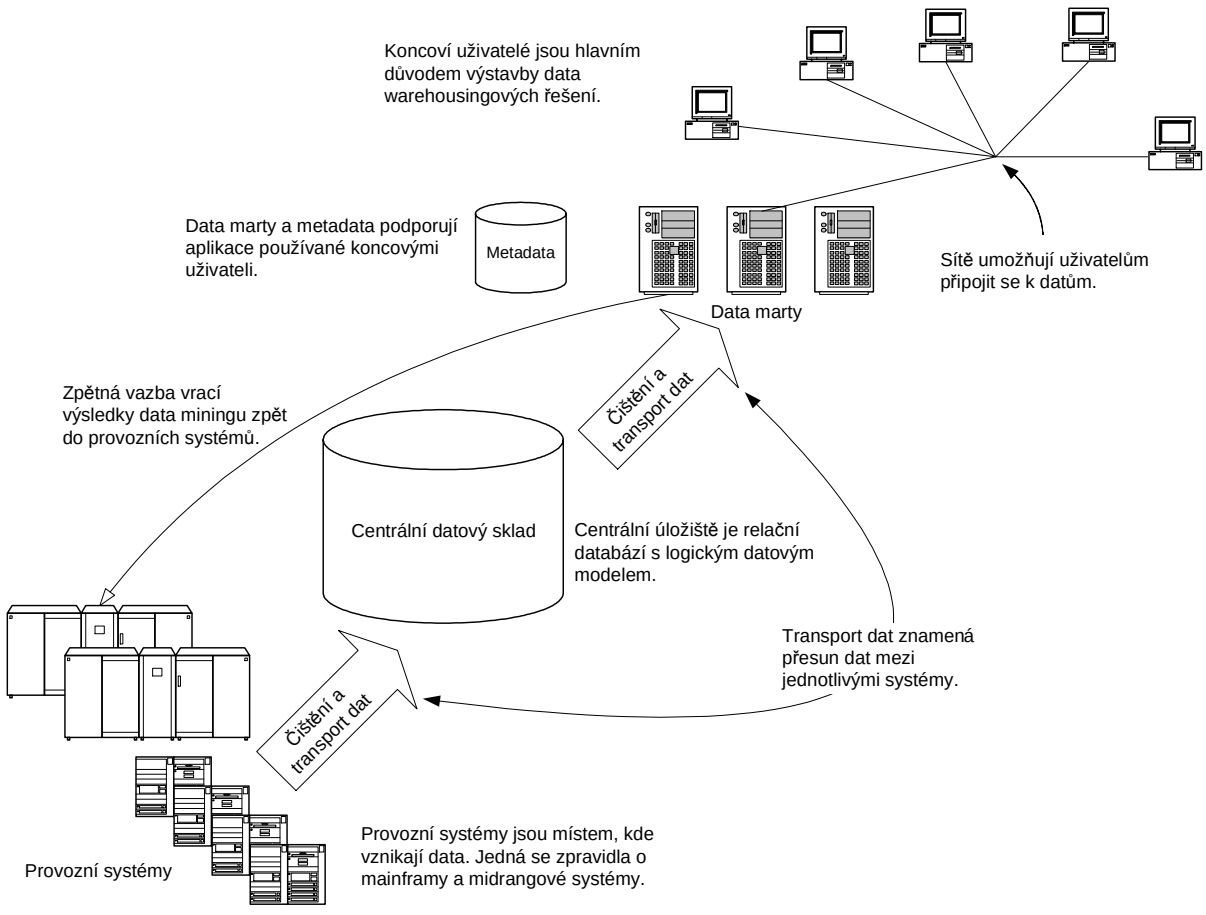
Datový sklad je předmětově orientovaný, tzn. umožňuje pohledy na celopodnikové subjekty, jako jsou zákazníci, prodeje, výše předepsaného pojistného, výše vyplacených plnění včetně rezerv na vzniklé pojistné události apod. Jádrem datového skladu je soubor relačních databází, které byly navrženy tak, aby byly mnohostranně použitelné a snadno rozšiřitelné v případě potřeby. Datový sklad obsahuje historická data, přičemž každý záznam je relevantní pro daný časový okamžik. To umožňuje vytvářet sestavy zaměřené na sledování trendů a periodických změn uchovávaných veličin. Data uložená v datovém skladu mají různou míru podrobnosti, tzn. mohou být jak atomická, tak i sumarizovaná.

Datový sklad zajišťuje přístup k jednotným celopodnikovým datům pro uživatele z řad obchodních analytiků a managementu. Centralizace dat eliminuje polemiky o jejich správnosti. V případech, kdy jsou data podobného charakteru spravována ve více systémech najednou, vykazuje zpravidla každý z nich jiná čísla

(ať již vlivem odlišné metodiky zpracování, nebo vlivem chyb). Datové sklady jsou používány pro zaznamenávání minulosti. To znamená, že tato funkcionalita není vyžadována od provozních systémů, které mohou zůstat specializované pouze na obsluhu obchodu. To umožňuje oddělit analytické a transakční zpracování dat.

V datovém skladu jsou pro analytické účely, ale i reporting, vytvářeny specializované množiny dat, tzv. datová tržiště (Data Marts). Jejich cílem je vyhovět specializovaným požadavkům různých částí organizace. V případě pojišťovny jsou vytvářeny specializované data marty např. pro útvar pojistné matematiky, útvar zajištění, útvar marketingu apod.

Na obrázku 3 je znázorněn proces toku dat z provozních systémů přes datový sklad ke koncovým uživatelům.



Obr. 3 Proces toku dat z provozních systémů přes datový sklad ke koncovým uživatelům [3]

2.4 Zdrojové systémy

Hlavními zdrojovými systémy pro datový sklad jsou podnikové provozní systémy, které z pohledu abstrakce produkují nejnižší úroveň dat (viz. obrázek 2). Tyto systémy byly navrženy pro provozní použití a nikoliv pro podporu rozhodování. V řadě firem jsou provozovány na širokém okruhu hardwaru, nejčastěji mainframech a midrangových systémech. Také software, který je používán pro jejich obsluhu, je velmi různorodý, často vytvořený na míru s vysokým stupněm přizpůsobení požadavkům uživatelů. Úkolem provozních systémů je správa a zpracování dat, nikoliv jejich sdílení. Proto snahy o jejich využití pro podporu rozhodování pro ně představují zátěž a ohrožují jejich základní funkcionalitu, kterou je obsluha obchodu.

Firmy zpravidla používají velké množství provozních systémů. V případě pojišťovny jsou zaměřeny např. na správu smluv životního nebo neživotního pojištění, likvidaci pojistných událostí, správu obchodníků a jejich provizí apod. Bohužel, většina z těchto systémů je obtížně přizpůsobitelná novým obchodním požadavkům firem. Každá jejich větší změna je tak pro firmy stejně náročná a nákladná jako jejich úplné nahrazení novými systémy. Kořeny těchto problémů sahají až do období konce sedmdesátých a počátku osmdesátých let minulého století. Náklady na diskové kapacity byly tehdy velmi vysoké. Podniky, s cílem ušetřit za každou cenu diskový prostor, vytvářely mnohdy i vědomě neohebné systémy, poněvadž se domnívaly, že dříve než nastanou větší změny v jejich obchodních požadavcích, budou staré technologie nahrazeny novými, modernějšími. Opak byl však pravdou.

Pro ekonomiku devadesátých let minulého století byly charakteristické změny, jejichž četnost roste exponenciálně dodnes. Spolu s tržními změnami se mění i obchodní požadavky firem, které je nutné implementovat do provozních systémů. Staré systémy nejsou pro tyto účely vhodné. Při jejich náhradě novými systémy se však objevují problémy, kdy se určitou část dat původního systému nepodaří migrovat do nového. Oba poté žijí určitou dobu vedle sebe, ačkoliv zabezpečují podobnou nebo dokonce stejnou provozní činnost. Provozní systémy jsou tedy nevhodné pro použití pro podporu rozhodování. Nalezení informace v často i desítkách provozních systémů je nereálné. Řešením se proto stává vybudování datového skladu.

Vedle provozních systémů mohou být zdrojem datového skladu také externí databáze (např. různé komerční databáze potenciálních zákazníků nakupované od jiných firem, databáze adresních bodů GIS apod.).

2.5 ETL procesy

Nejpracnější a zároveň nejnákladnější částí implementace data warehousingového řešení jsou tzv. ETL procesy. Jejich hlavním úkolem je přečtení dat ze zdrojových systémů (Extraction), čištění a transformace dat do nových struktur (Transformation) a nahrání dat do datového skladu (Load).

2.5.1 Čtení dat ze zdrojových systémů

Pro čtení dat ze zdrojových systémů je rozhodujícím kritériem skutečnost, aby se podařilo v požadovaném čase načíst všechna data. To nemusí být vždy jednoduchou záležitostí. Je-li např. zdrojový systém provozován v režimu kritické dostupnosti, představuje jeho masivní čtení nepřiměřenou zátěž pro další transakce. Polášek [31] doporučuje v těchto případech využít serveru, který poskytne online zálohu pro případ výpadku systému (Hot Stand-By Server), implementovat triggery na primární systém, takže se budou číst pouze modifikace dat z triggerových tabulek a ne data sama, navrhnout a sestavit systém pro rychlou extrakci dat do textových souborů a ty poté dále zpracovávat.

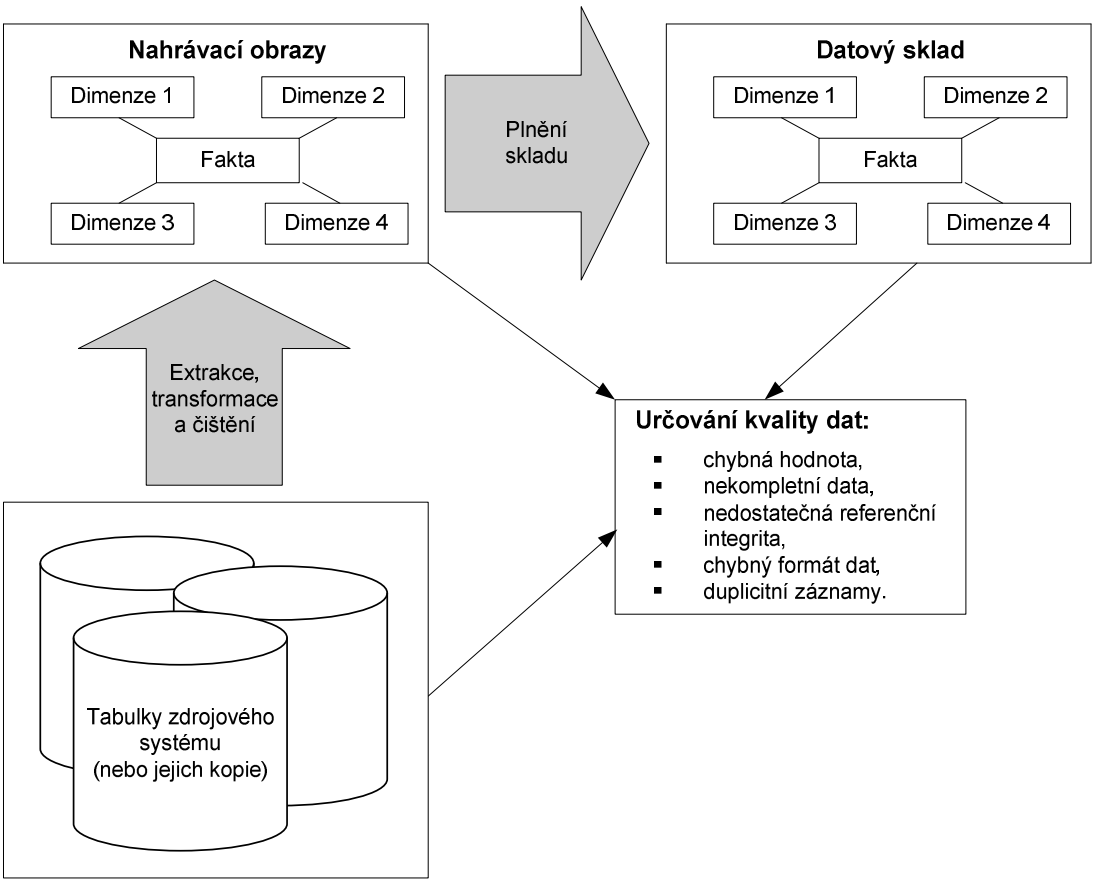
2.5.2 Transformace a čištění dat

Druhým úkolem ETL procesů je transformace a čištění dat. Tento proces však není pouze technickou záležitostí. Před spuštěním transformačních a čistících mechanismů je nutné stanovit pravidla, jakým způsobem bude zacházeno s datovými nekonzistentnostmi. Tato pravidla musí definovat především obchodní uživatelé. Proces transformace a čištění je tedy nejobtížnější částí vrstvy ETL. Vyžaduje mnohostrannou komunikaci napříč podnikem, přičemž je velmi složité nalézt řešení, které by vyhovovalo všem stranám [28]. Chyby v datech je vhodné rovněž reportovat a usilovat o jejich nápravu nejlépe přímo ve zdrojových systémech.

Cílem čištění dat je dosažení inherentní a pragmatické informační kvality. Inherentní informační kvalitou rozumíme stupeň, s jakým data přesně zobrazují objekty reálného světa [10]. Jestliže např. datum narození klienta nabývá hodnoty

8. května 1980, je tato hodnota inherentně kvalitní. Inherentní kvalita tedy znamená, že data nabývají formálně správných hodnot. Vedle toho pragmatickou informační kvalitou rozumíme stupeň užitečnosti dat při podpoře podnikových procesů [10]. Data ve firemním datovém skladu nemají sami o sobě žádnou aktuální hodnotu. Teprve v případě jejich použití v rámci činnosti, která podniku přináší prospěch, je realizována jejich potenciální hodnota. Data, která jsou inherentně kvalitní, tj. validní, nemusí dosahovat pragmatické kvality, ať již z důvodu jejich neaktuálnosti nebo nadbytečnosti. Taková data potom pro firmu představují zbytečné náklady. Kvalitně zpracovaná vrstva ETL tedy na jedné straně zvyšuje potenciální hodnotu firemních dat, na druhé straně rovněž redukuje náklady spojené s jejich údržbou.

Pro efektivní řízení kvality dat je nutné, aby data a jejich kvalita byly neustále monitorovány. Určování kvality dat může být v rámci datového skladu prováděno v různých okamžicích na různé back-end prvky. Z obrázku 4 je patrné, že kontrola kvality dat může být zaměřena na data ve zdrojových systémech, nahrávací obrazy či přímo na datový sklad.



Obr. 4 Určování kvality dat na back-end straně datového skladu [16]

Cílem transformací dat je dosažení sjednoceného pohledu na celopodniková data. Transformace a čištění dat jsou zpravidla prováděny v tzv. staging area. Tato vrstva může zároveň sloužit i jako prostor pro dočasné umístění kopií provozních tabulek po jejich přečtení ze zdrojových systémů a tabulek transformovaných dat před jejich nahráním do datového skladu.

2.5.3 Nahrávání dat

Poslední částí vrstvy ETL je nahrání dat do datového skladu. Tento proces je již pouze technickou záležitostí podobnou čtení dat ze zdrojových systémů. Data v datovém skladu nejsou aktualizována častěji než jednou za 24 hodin. Samotná frekvence aktualizace závisí na výpočetním výkonu používaného hardwaru. Pokud uživatelé vyžadují okamžitá data pro účely provozního sledování, není pro ně datový sklad vhodným řešením. Těmto požadavkům naopak dobře vyhoví sklady provozních dat (ODS). Datový sklad je dostupný koncovým uživatelům během pracovního dne. Jeho plnění probíhá v noci nebo o víkendech.

2.6 Centrální datový sklad

Základní vrstvou datového skladu je centrální datové úložiště, jehož hlavními rysy jsou [3]:

- rozšiřitelný hardware,
- relační databáze a systém pro její řízení (RDBMS),
- logický datový model.

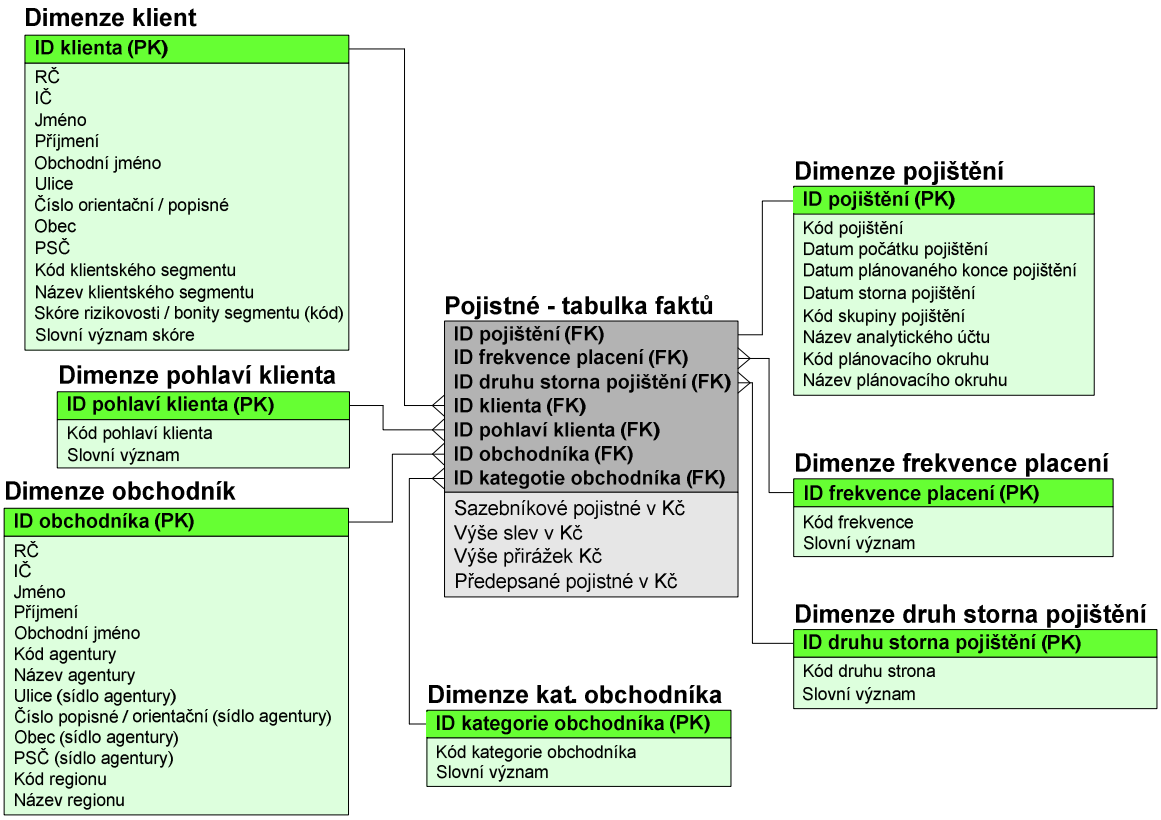
S výrazným rozvojem technologií SMP (Symmetric Multiprocessor), MPP (Massively Parallel Processor) a jejich kombinací, může hardware používaný v datovém skladu růst prakticky bez omezení. Tyto technologie podporují více uživatelů a umožňují rozšiřovat hardware přidáním více disků, více pamětí, více procesorů apod. To vše při využití již realizovaných investic do technologií, vzdělání a podpory provozu. Pro datové sklady, které mohou obsahovat stovky gigabytů nebo až terabyty dat, jsou tyto technologie předpokladem efektivního provozu.

Relační databáze a systémy pro jejich řízení se vyvinuly do podoby, která bezproblémově využívá rozšiřitelný hardware při zpracování strojově náročných operací jako je načítání dat, vytváření indexů, zálohování databáze či zpracování

složitých dotazů. Prakticky každý RDBMS spolupracuje přes konektor typu ODBC s nástroji používanými koncovými uživateli.

Centrální datový sklad je tvořen relační databází, která nemusí vždy striktně oddělovat fakta a dimenze do zvláštních tabulek. Rozdělení dat na fakta a dimenze však zkracuje dobu odezvy na dotazy a zjednodušuje přístup k datům. Fakty rozumíme spojité proměnné (např. výši přijatých nebo vyplacených peněz), dimenzemi pak kategoriální proměnné (např. pohlaví, frekvenci placení apod.). V případě, že jsou všechny tabulky dimenzí navázány přímo na tabulku faktů, hovoříme o schématu hvězdy.

Na obrázku 5 je uveden příklad schématu hvězdy pro pojistné placené klientem na pojištění.

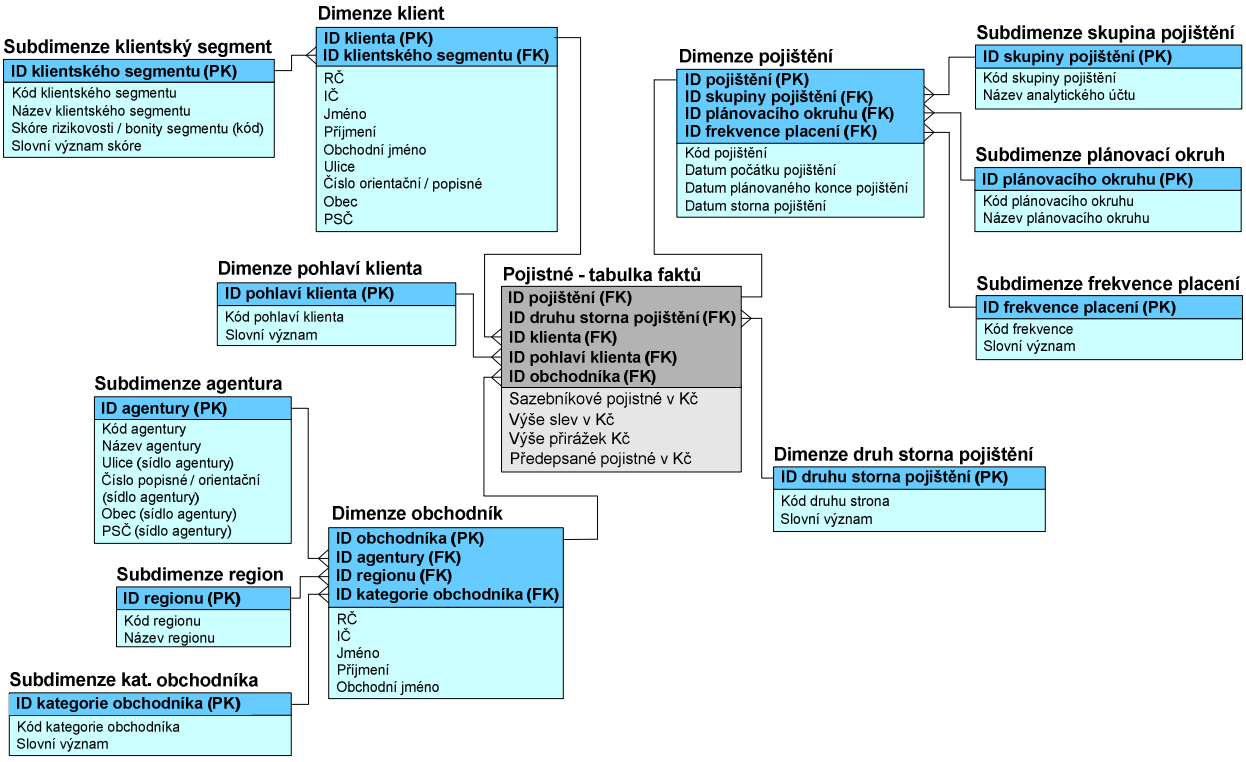


Obr. 5 Dimenzionální model způsobu uložení dat v centrálním datovém skladu – schéma hvězdy

Jak je patrné z obrázku 5, tabulka faktů obsahuje výši různých typů pojistného včetně slev a přirážek. Dimenzemi jsou klient, pohlaví klienta, obchodník, kategorie obchodníka, pojištění, frekvence placení pojistného a druh storna pojištění. Vlastnosti dimenzí pohlaví klienta, kategorie obchodníka, frekvence placení pojistného a druh

storna pojištění by mohly být evidovány jako vlastnosti dimenzí klient, obchodník a pojištění. Vzhledem k tomu, že se očekává vysoká četnost dotazů na výši pojistného podle pohlaví, je tato vlastnost evidována v samostatné dimenzi. Totéž platí i pro dimenze kategorie obchodníka, frekvence placení pojistného a druh storna pojištění.

V případě, že jsou na jednotlivé tabulky dimenzí navázány další tabulky subdimenzí, hovoříme o schématu vločky. Jeho příklad je uveden na obrázku 6.



Obr. 6 Dimenzionální model způsobu uložení dat v centrálním datovém skladu – schéma vločky

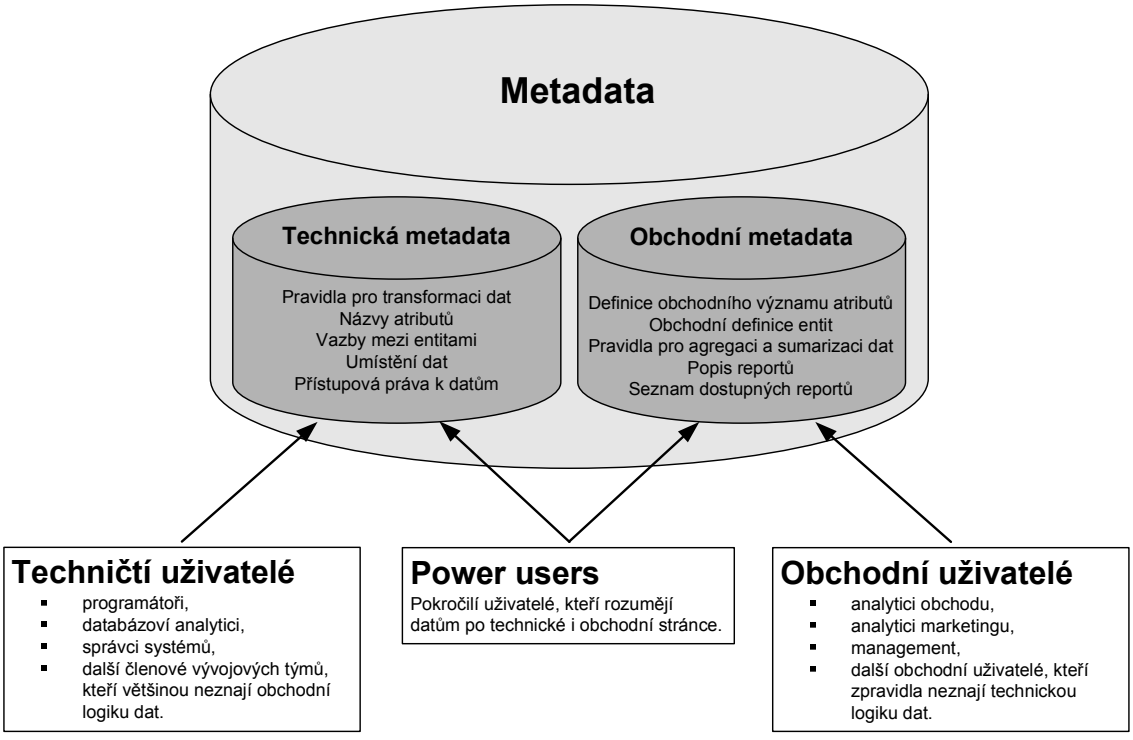
Na tabulku faktů, která opět obsahuje výši různých typů pojistného včetně slev a přirážek, jsou přímo navázány dimenze klient, pohlaví klienta, obchodník, pojištění a druh storna pojištění. Na dimenzi klient je dále navázána subdimenze klientský segment, která představuje vyšší stupeň agregace této dimenze. Totéž platí i pro subdimenze agentura a region ve vztahu k dimenzi obchodník a subdimenze skupina pojištění a plánovací okruh ve vztahu k dimenzi pojištění. Vedle hierarchií dané dimenze mohou být jako subdimenze vyčleněny i další vlastnosti. U dimenze pojištění je to např. subdimenze frekvence placení a u dimenze obchodník subdimenze kategorie obchodníka. Obecně se doporučuje vyčlenit jako subdimenze ty vlastnosti, které se mohou v čase často měnit. Dimenze pohlaví klienta a druh storna pojištění jsou přímo

navázány na tabulku faktů, neboť se opět očekává vysoká četnost dotazů na výši pojistného přes tyto dimenze.

Vzhledem k jednoduchosti dotazů představuje dimenzionální model uložení dat lepší řešení než klasická relační databáze. Do centrálního datového skladu mohou být data pouze vkládána, nesmějí být mazána ani měněna.

Posledním klíčovým rysem centrálního datového skladu je logický datový model. Jeho hlavním úkolem je sdělit obsah databáze obchodním uživatelům (např. analytikům obchodu, marketingu apod.), kteří většinou nemají technickou povahu. Logický datový model je často zaměňován s fyzickým rozvržením, tj. technickými metadaty. Ta jsou naopak určena pro technické uživatele z řad programátorů, databázových analytiků apod. Vedle různých skupin uživatelů spočívá rozdíl fyzického rozvržení a datového modelu také v tom, že fyzické rozvržení je pouze technickou aplikací logického datového modelu.

Na obrázku 7 jsou znázorněny příklady technických a obchodních metadat včetně jejich uživatelů.



Obr. 7 Technická a obchodní metadata a jejich uživatelé [26]

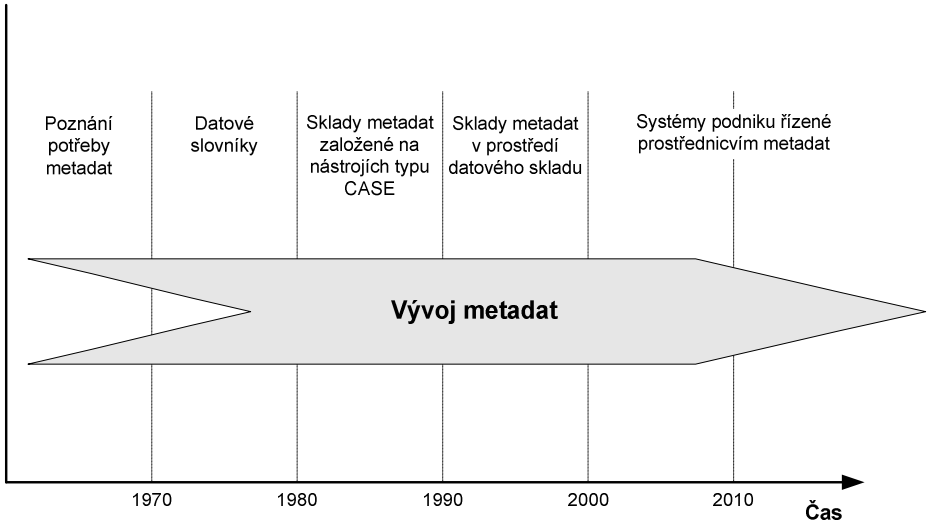
2.7 Metadata

Nejčastěji uváděná a zároveň nejjednodušší definice vymezuje metadata jako data o datech. Tato definice však zdaleka nepokrývá úplný rozsah metadat. Marco [26] definuje metadata následovně:

- Metadata jsou všechna fyzická data obsažená v softwaru a jiných médiích včetně znalostí zaměstnanců, která mají svůj původ uvnitř nebo vně organizace. Metadaty se rovněž rozumí informace o fyzických datech, technických a obchodních procesech, pravidlech, omezeních a strukturách dat používaných v dané firmě.

Metadata tedy umožňují uživatelům datového skladu snadno a rychle lokalizovat správné informace, stopovat data až k jejich internímu nebo externímu zdroji, sledovat historický vývoj vlastností dat a procesů a odpovídat na otázky typu, proč existují vztahy mezi daty a jak jsou tyto vztahy aplikovány. Ve smyslu výše uvedené definice metadata představují nejvyšší tři vrstvy v modelu abstrakce, který je uveden na obrázku 2. Samotný koncept metadat není nový. Jeho počátky sahají až do první poloviny sedmdesátých let minulého století. Zcela nová je však role metadat v systémech pro podporu rozhodování. Metadata zde zachycují historické změny v datech, takže firmy mohou analyzovat trendy, které ovlivňují jejich obchodní rozhodnutí.

Na obrázku 8 je znázorněn vývoj metadat od vzniku jejich konceptu až po současnost.



Obr. 8 Vývoj metadat [26]

Za účelem evidence a analýzy metadat budují firmy centralizované sklady metadat (Metadata Repositories). Tyto sklady uchovávají různé typy metadat z různých zdrojů. Jednotlivé zdroje mohou mít rovněž svůj vlastní sklad, avšak jejich vzájemná propojitelnost je velmi omezená a někdy i nemožná.

V tabulce 1 je uveden přehled nejčastějších zdrojů metadat včetně typů metadat, které obsahují.

Tab. 1 Zdroje metadat včetně typů metadat, které obsahují [26]

Zdroj metadat	Typ metadat
Nástroje / procesy ETL	Pravidla pro transformaci dat
	Skutečnosti, na kterých závisí zpracování úlohy
	Statistiky vytíženosti systému pro podporu rozhodování
	Statistiky procesu nahrávání dat do systému pro podporu rozhodování
	Generace dat vytvořené v průběhu transformace zdrojových dat
Nástroje pro datové modelování	Logický a fyzický datový model
	Obchodní a technické definice entit
	Obchodní a technické definice atributů
	Umístění dat
Reportingové nástroje	Přístupová práva uživatelů
	Čas vytvoření reportu
	Obchodní definice entit
	Obchodní definice atributů
	Definice klíčových obchodních měr
Nástroje pro kontrolu kvality dat	Statistiky kvality dat
	Výsledky datového auditu
Nakoupené aplikace	Logický a fyzický datový model
	Datové slovníky
Dokumenty	Definice obchodní politiky
	Obchodní definice entit
	Obchodní definice atributů
	Definice klíčových obchodních měr
	Osoby zodpovědné za věcnou správnost obchodních a technických definic
Zaměstnanci	Definice obchodní politiky
	Obchodní definice entit
	Obchodní definice atributů
	Osoby zodpovědné za věcnou správnost obchodních a technických definic

Základním rysem skladu metadat je tzv. metamodel, nebo-li fyzické rozvržení způsobu uložení metadat (relační nebo objektové). Cílem dobře navrženého metamodelu je umožnit snadné sdílení metadat pocházejících z různých zdrojů včetně podpory součinnosti nástrojů, které metadata vytvářejí.

2.8 Datová tržiště

Datová tržiště (Data Marts) jsou specializované množiny dat vytvářené v datovém skladu za účelem zabezpečení rychlého a snadného přístupu k datům pro obchodní uživatele. Jejich cílem je uspokojit prostřednictvím reportů, ad hoc dotazů či pokročilých analýz specifické požadavky uživatelů z řad marketingu, lidských zdrojů, risk managementu apod. Zdrojem dat pro data marty může být centrální datový sklad, nebo přímo provozní systémy [29].

Data marty mohou být implementovány mnoha různými způsoby. Jedním z nich je například vytvoření dynamických pohledů (Views) na základní tabulky datového skladu. I když tyto pohledy kombinují data z více tabulek, nezabírají žádné diskové kapacity a je možné je snadno a rychle změnit. Nevýhodou je vysoká investice do hardwaru pro zabezpečení rychlého zpracování dotazů.

Další možností implementace data martů je vytvoření dynamických pohledů na předpřipravené tabulky. Aktuálnost dat v těchto tabulkách je shodná s daty v centrálním datovém skladu. Předpřipravené optimalizované tabulky umožňují zkrátit dobu odezvy, ale zároveň představují i větší nároky na diskové kapacity.

Velmi populární jsou data marty, které uživatelům zpřístupňují data přes OLAP (Online Analytic Processing). Jejich cílem je umožnit dynamické zobrazení faktů v různých dimenzích. Způsob uložení dat v těchto data martech je buď dimenzionální (schéma hvězdy, schéma vločky), multidimenzionální nebo hybridní (kombinace dimenzionálního a multidimenzionálního modelu). Podle způsobu uložení rozlišujeme i typy přístupu k datům. Pro dimenzionální model je používán ROLAP (Relational OLAP), pro multidimenzionální model MOLAP (Multidimensional OLAP) a pro hybridní model HOLAP (Hybrid OLAP) [2].

Základní vrstva datového skladu může být tvořena kolekcí data martů. Tento koncept architektury centrálního datového skladu souvisí s dimenzionálním modelem uložení dat. Za účelem zjednodušení přístupu k datům a využití diskových kapacit,

může být základní vrstva datového skladu tvořena souborem hvězd či vloček, které mohou být zároveň i data marty [22].

2.9 Uživatelé datového skladu

Koncoví uživatelé jsou hlavním důvodem výstavby datového skladu. Mezi tyto uživatele patří analytici, vývojáři aplikací a obchodní uživatelé. Analytici zpravidla hledají informace skryté uvnitř dat s použitím pokročilých nástrojů. Cílem jejich práce je na základě přirozeného chování dat identifikovat oblasti, na které by se měly zaměřit podnikové procesy (např. identifikace bonitních nebo rizikových klientských segmentů, identifikace nestandardního průběhu prací v rámci work flow apod.). Výsledky těchto analýz se vrací zpět do provozních systémů. Příkladem může být skóring klienta, který ovlivňuje míru jeho zvýhodnění při sjednání nového obchodu.

Vývojáři aplikací potřebují, aby firemní data byla relativně stabilní, aby obsahovala validní hodnoty a aby byl dobře interpretovatelný jejich význam. Tyto předpoklady splňuje datový sklad. Aplikace, které jsou vytvářeny nad daty datového skladu, zahrnují EIS (Executive Information System), OLAP a další specifické aplikace vytvářené dle požadavků útvaru obchodu, risk managementu apod.

Obchodní uživatelé jsou konečnými příjemci informací odvozených z dat. Jejich potřeby řídí vývoj aplikací, ale i architektury datového skladu. Obchodní uživatelé rovněž určují priority implementace jednotlivých změn. Vedle standardních reportů z podnikových aplikací vyžadují také ad hoc výstupy podle svých specifických potřeb.

2.10 Sklad provozních dat

Součástí podnikového systému pro podporu rozhodování (Decision Support System) je dle Inmona [19] také sklad provozních dat (Operational Data Store). Sklad provozních dat je určen pro podporu taktického rozhodování a sledování provozu. Může být i jedním ze zdrojů pro datový sklad. Architektura skladu provozních dat je velmi podobná architektuře datového skladu. I přes tuto podobnost se sklad provozních dat liší od datového skladu v následujících rysech [19]:

- frekvence jeho aktualizace je blízká online,
- neobsahuje historická data.

Důvodem těchto odlišností je samotný účel skladu provozních dat, tj. informovat uživatele o okamžitém stavu provozu.

2.11 Data mining a jeho vztah k data warehousingu

Data miningem rozumíme disciplínu, která se začala bouřlivě rozvíjet na počátku devadesátých let minulého století. Tehdy došlo k postupnému propojení technik a technologií, které se do té doby vyvíjely zcela nezávisle na sobě. Databázové technologie představují již několik desetiletí osvědčený prostředek, jak uchovávat rozsáhlé objemy dat a tato data filtrovat pomocí speciálních dotazovacích jazyků (např. SQL). Vedle toho existuje již několik desetiletí řada matematicko statistických metod, které jsou osvědčeným prostředkem pro analýzu a modelování závislostí v datech. Spojením těchto technik a technologií vznikl na počátku devadesátých let minulého století nový obor nazývaný Data mining.

Existuje velké množství definic data miningu. Jako příklad lze uvést následující:

- Data mining je netriviální proces zjišťování platných, neznámých, potenciálně užitečných a snadno pochopitelných informací z dat [12].
- Data mining je zkoumání a analýza rozsáhlých objemů dat, která je prováděna s použitím automatizovaných a poloautomatizovaných prostředků za účelem objevení významných závislostí a pravidel uvnitř skrytých [3].

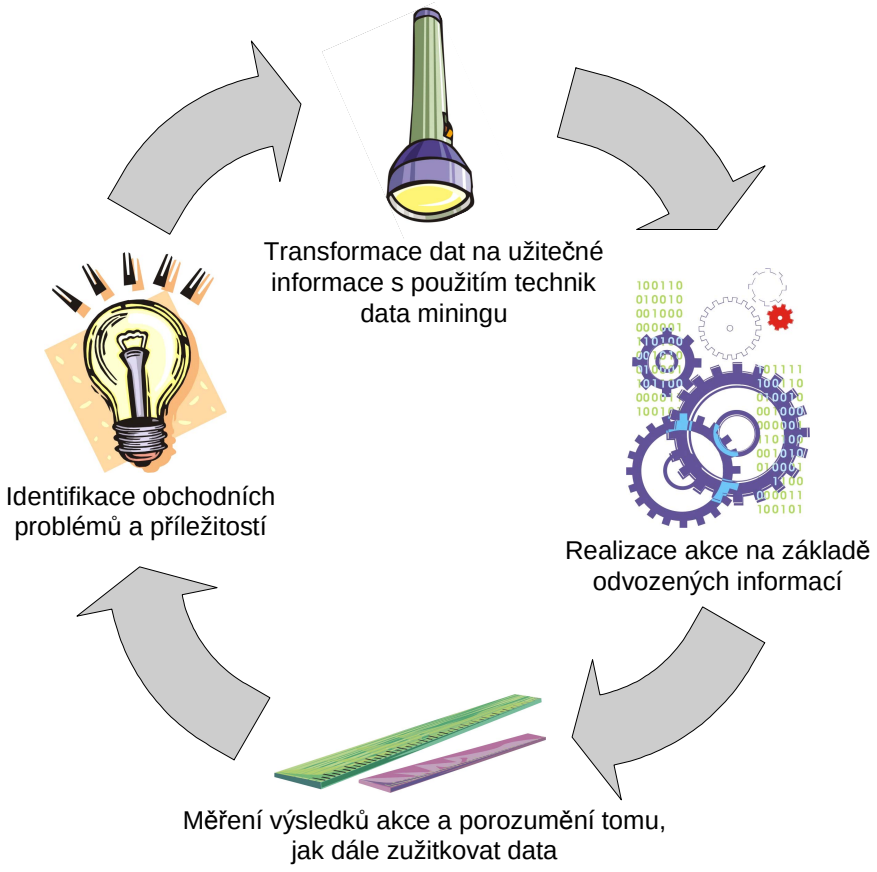
Rozvoj data miningu úzce souvisí s rozvojem data warehousingu. Jeho cílem je poskytnout uživatelům prostředky, které podpoří jejich kreativitu a zvýší úroveň chápání dat. Berry a Linoff [3] uvádějí následující příklady toho, jak se data mining a data warehousing vzájemně podporují:

- Užitečnost technik data miningu se projevuje při jejich aplikaci na rozsáhlé objemy dat, které jsou zpravidla uloženy v datovém skladu.
- Výsledky data miningových analýz jsou velmi závislé na kvalitě vstupních dat. Jejich aplikace na čistá a konzistentní data datového skladu zvyšuje návratnost investic vložených do čistících a kontrolních mechanismů.

- Datový sklad umožňuje testování hypotéz a usnadňuje měření efektů akcí, které byly realizovány na základě výsledků data miningových analýz.
- Rozšiřitelný hardware a systémy pro řízení relačních databází mohou rovněž podporovat techniky data miningu, čímž se zvyšuje návratnost investic vložených do těchto technologií.

2.12 Jak funguje data mining

Hlavním přínosem data miningu v procesu podpory rozhodování je nalezení závislostí a vazeb skrytých uvnitř gigabytů či terabytů dat. Samotná identifikace závislostí a vazeb však k dosažení úplného efektu použití data miningu nestačí. Pro firmy je velmi důležité, aby uměly proměnit informace odvozené z dat v akci a tuto akci následně v hodnotu. Berry a Linoff [3] nazývají tento proces virtuálním cyklem data miningu (viz. obrázek 9).



Obr. 9 Fáze virtuálního cyklu data miningu [3]

Z obrázku 9 je patrné, že virtuální cyklus data miningu se skládá ze čtyřech fází:

1. identifikace obchodního problému, resp. příležitosti,
2. použití technik data miningu k transformaci dat na užitečné informace,
3. realizace obchodní akce na základě odvozených informací,
4. měření výsledků obchodní akce a snaha o identifikaci nového problému nebo příležitosti.

2.12.1 Identifikace obchodních problémů a příležitostí

Identifikace obchodních problémů a příležitostí je fází, která probíhá ve všech částech organizace. Jejím cílem je nalézt oblasti, ve kterých informace odvozené z dat usnadní práci zaměstnancům a poskytnou firmě hodnotu. Existuje několik různých přístupů k této fázi.

Data v provozních systémech, popř. v datovém skladu, zachycují informace o obchodních procesech firmy. Jejich řízení závisí na analýze těchto dat, což zpravidla vede i k rozpoznání potenciálních problémů či příležitostí. Mezi činnosti, které jsou prováděny během řízení obchodních procesů, patří např.:

- plánování marketingových kampaní pro nové výrobky nebo služby,
- ocenění stávajících výrobků nebo služeb,
- plánování aktivit zaměřených na zamezení odchodu bonitních zákazníků ke konkurenci apod.

Častým podnětem k použití data miningu může být i běžné sledování provozu. Toto sledování generuje otázky typu:

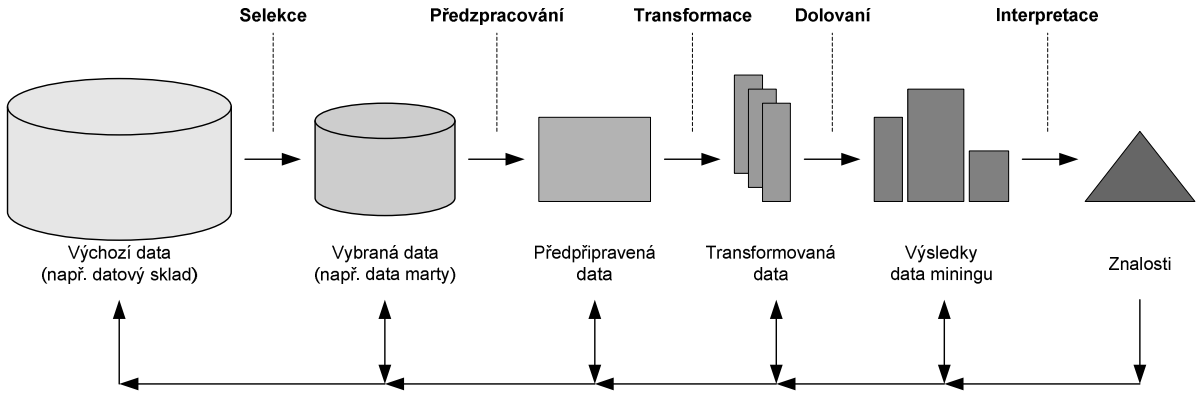
- Proč v regionu Jižní Čechy zaostává prodej životního pojištění za regionem Severní Morava?
- Je doba vyřízení pojistné události delší než dva měsíce příčinou odchodu zákazníků ke konkurenci?
- Proč doba zpracování pojistných smluv pro pojištění staveb a domácností výrazně převyšuje dobu zpracování smluv životního pojištění (pozn.: dobou zpracování pojistné smlouvy se rozumí doba, která uplyne od sepsání jejího návrhu do akceptace pojišťovnou)?

Identifikaci cenných oblastí z hlediska obchodu lze rovněž realizovat na základě rozhovorů analytiků s klíčovými představiteli různých útvarů firmy. Tento přístup se doporučuje zejména v ranných fázích implementace data miningu do podnikových procesů. Většina zaměstnanců firmy totiž neví, jak data mining funguje. Je tedy nutné vysvětlit jim jeho základní principy a zdůraznit jeho příznivé dopady na podnikové procesy. Během těchto rozhovorů mohou být současně identifikovány i potenciální obchodní problémy, resp. příležitosti.

2.12.2 Odvození informací

Cílem druhé fáze virtuálního cyklu je odvodit s použitím technik data miningu užitečné informace pro realizaci obchodní akce. Podnětem pro tuto činnost je obchodní problém, resp. příležitost, identifikovaný v první fázi cyklu. Rozhodujícím kritériem pro použití technik data miningu je kvalita a dostupnost vstupních dat. V této souvislosti hraje významnou roli datový sklad, který obsahuje sjednocená, předmětově orientovaná, neměnná, historická data z mnoha interních a externích zdrojů firmy.

Na obrázku 10 je znázorněn průběh druhé fáze virtuálního cyklu, od přípravy datových zdrojů až po znalosti získané z odvozených informací.



Obr. 10 Proces dolování znalostí z databází [12]

2.12.3 Realizace akce na základě odvozených informací

Třetí fází virtuálního cyklu data miningu je implementace odvozených informací do obchodních procesů firmy. Zde je nutné věnovat zvýšenou pozornost zejména způsobu samotné implementace. Zavedení výsledků data miningu do obchodních procesů podniku je zpravidla spojeno s realizací celého souboru činností, např.:

- Při zavádění nové strategie pro podporu prodeje je důležité, vedle nasazení různých stimulů, shromažďovat i data potřebná k pochopení účinnosti této strategie.
- Při zavádění nového produktu, který byl navržen pro specifický segment klientů, je vhodné nejprve otestovat odezvu na tento produkt jeho nabídnutím malé skupině zákazníků. Výsledky testu je pak možné využít při plánování marketingových kampaní zaměřených na celý segment.
- Jestliže klient často využívá služeb call-centra, je vhodné zařadit tuto informaci do jeho profilu. Klientům, kteří byli na základě data miningové analýzy klasifikováni jako bonitní, lze pak nabídnout speciální telefonní číslo, které bude zvýhodněno při čekání ve frontě hovorů.

2.12.4 Měření výsledků akce

Měření výsledků implementace data miningu do obchodních procesů firmy je náplní čtvrté fáze virtuálního cyklu. Reakce obchodu na použití data miningu je zachycena v datech provozních systémů. Po jejich načtení do systému pro podporu rozhodování mohou uživatelé měřit úspěšnost data miningových analýz. Během tohoto měření jsou zpravidla identifikovány nové obchodní příležitosti a virtuální cyklus znovu opakován.

Jakmile uživatelé poznají klady vyplývající z použití technik data miningu, budou požadovat jejich nasazení ve stále větší míře. Výsledky své práce budou sdělovat i ostatním útvarům firmy, čímž podpoří vzájemnou komunikaci a snadnější implementaci data miningu do většiny podnikových procesů.

2.12.5 Data miningové metodologie

S postupem doby se začaly v oblasti data miningu uplatňovat metodologie, které představují konkrétní implementaci jednotlivých fází virtuálního cyklu. Mezi tyto metodologie patří např. SEMMA firmy SAS Institute Inc. nebo CRISP – DM vlastněná firmami NCR Corporation, SPSS Inc., Daimler Chrysler AG a OHRA.

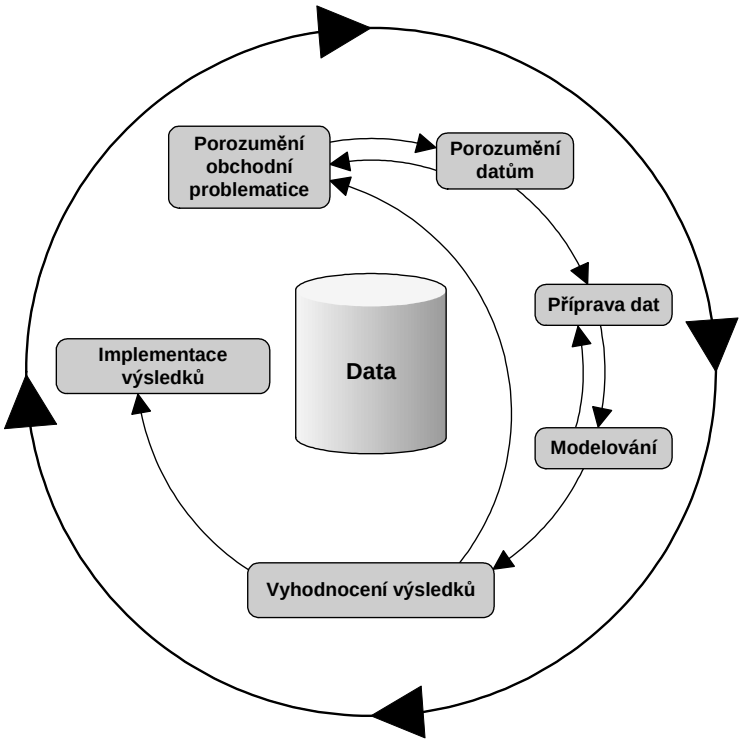
Název metodologie SEMMA vznikl složením počátečních písmen z názvů činností, které jsou postupně prováděny v rámci jejího nasazení. Jedná se o tyto činnosti [39]:

1. *Sample* – vytvoření vzorku z rozsáhlého datového souboru, rozdělení dat na vývojový, validační a testovací soubor.
2. *Explore* – prozkoumání dat za účelem jejich pochopení a získání nápadů pro jejich analýzu, tj. grafická vizualizace dat, výpočet popisných statistik, identifikace významných proměnných apod.
3. *Modify* – příprava dat pro modelování, tj. vytvoření nových a transformace stávajících proměnných, identifikace odlehlých pozorování, nahrazení chybějících hodnot apod.
4. *Model* – vytvoření prediktivního modelu s použitím regresní analýzy, rozhodovacích stromů, neuronových sítí apod.
5. *Assess* – porovnání vytvořených modelů a jejich variant.

Aplikace metodologie SEMMA odpovídá realizaci druhé fáze virtuálního cyklu, tj. transformaci dat na užitečné informace. Úplný průběh virtuálního cyklu je realizován s použitím metodologie CRISP – DM (Cross-Industry Standard Proces for Data Mining). Její vývoj byl zahájen jako projekt Evropské komise s cílem vytvořit standardní postup pro realizaci data miningových projektů. Tento postup se skládá ze šesti základních etap [17]:

1. *Porozumění obchodní problematice* – tato etapa zahrnuje identifikaci obchodních příležitostí (resp. problémů), jejich definici v podobě úlohy data miningu, stanovení cílů projektu včetně způsobu jejich dosažení a analýzu dostupnosti zdrojů. V této fázi je rovněž velmi důležité nastavit kritéria pro hodnocení výsledků celého projektu.
2. *Porozumění datům* – tato etapa začíná shromážděním surových dat potřebných pro analýzu a pokračuje jejich průzkumem za účelem objevení zajímavých podmnožin, posouzení kvality dat apod.
3. *Příprava dat* – cílem této etapy je vytvořit ze surových dat finální soubor, který bude použit pro danou analýzu. Činnosti, které jsou v této fázi realizovány, zahrnují selekci tabulek, polí a záznamů, transformaci a čištění dat.

4. *Modelování* – tato etapa začíná výběrem správného modelovacího algoritmu a pokračuje jeho aplikací na analyzovaná data. Pro vyřešení jednoho data miningového úkolu lze zpravidla použít více technik. Doporučuje se tedy vytvořit více modelů s použitím různých algoritmů a poté porovnat jejich výsledky. Některé modelovací techniky mohou mít speciální požadavky na vlastnosti dat. To si může vyžádat návrat do fáze jejich přípravy, tj. do etapy č. 3. Model je zpravidla vytvářen nad vývojovými daty, jeho správnost pak ověřována s použitím testovacích dat.
5. *Vyhodnocení výsledků* – v této etapě máme k dispozici model (resp. modely), který se z pohledu datové analýzy jeví jako kvalitní. Cílem je posoudit jeho správnost také z pohledu obchodního. Pokud je výsledný model označen jako schopný uspokojit obchodní potřeby, následuje důkladná revize celé data miningové úlohy, během které se zjišťuje, zda nebyl přehlédnut nějaký důležitý úkol.
6. *Implementace výsledků do obchodních procesů* – v této etapě jsou zaváděny výsledky data miningové analýzy do obchodních procesů firmy včetně stanovení plánu kontroly a údržby vytvořeného modelu.



Obr. 11 Metodologie CRISP – DM

Na obrázku 11 je znázorněn životní cyklus projektu data miningu dle metodologie CRISP – DM. Posloupnost jednotlivých fází není striktně vymezena. V závislosti na výstupu každé z nich se rozhoduje o postupu k další fázi, nebo o návratu k předchozí. Šipky ve schématu určují nejdůležitější a zároveň nejčastější závislosti mezi fázemi. Vnější kružnice symbolizuje virtuální cyklus data miningu. Po implementaci a měření výsledků jsou identifikovány nové obchodní příležitosti, které vedou k realizaci další data miningové úlohy.

2.13 Oblasti použití data miningu

Techniky data miningu lze použít prakticky ve všech oblastech lidské činnosti. Jedním z prvních jejich uživatelů byla federální vláda Spojených států. Techniky data miningu pomáhaly např. při vyhodnocení tisíců zpráv předložených agenty FBI během vyšetřování bombového útoku na Oklahoma City v roce 1995 [11]. Své uplatnění mají techniky data miningu také v medicíně. Zaznamenáváním a uchováváním lékařských zápisů vzniká databáze, ze které lze s jejich pomocí zjistit informace o rozšířenosti chorob, účinnosti léků, úrovni péče daného zdravotnického zařízení apod. Např. firma Merck-Medco Managed Care, kterou vlastní farmaceutická firma Merck & CO., Inc. zjišťuje závislosti mezi nemocemi a výsledky léčby po aplikaci různých léků. Při této činnosti využívá technik data miningu, jež jí pomáhají určit, které léky jsou vhodné pro daný typ pacientů [9].

Supermarkety získávají prostřednictvím snímacích zařízení na pokladnách velké množství transakčních nahrávek. Pomocí technik data miningu lze z těchto dat zjistit nákupní chování zákazníků, tj. kolik peněz v průměru utratí během jedné návštěvy, jaké zboží se dobře prodává společně, a je proto vhodné jej umístit na regálu vedle sebe apod. Většina supermarketů dnes používá klubové karty, které jim umožňují spojit si transakční data s konkrétním zákazníkem. Ten poskytne obchodníkovi osobní údaje, za které obdrží např. slevy při nákupu s použitím karty. Chování individuálního zákazníka však spíše než supermarket zajímá přímo výrobce zboží. Bez výslovného souhlasu klienta není zpravidla možné, aby si firmy mezi sebou vzájemně sdílely osobní data (pozn.: v ČR jsou podmínky pro nakládání s osobními údaji upraveny zejména zákonem č. 101/2000 Sb. O ochraně osobních údajů a o změně některých zákonů). Výrobce si však obvykle zaplatí cílenou reklamní kampaň, kterou provede supermarket

dle jeho požadavků, takže on sám nemá přímý přístup ke klientským datům. Např. firma Coca Cola si tak může objednat u firmy Rewe, světového provozovatele supermarketů (v České republice např. Billa nebo Penny Market), reklamní kampaň zaměřenou na zákazníky často kupující její nápoje. Obdobně mohou pojišťovny nebo banky prodávat reklamní prostor na obálkách dopisů s předpisem pojistného nebo s výpisem z účtu. Firmy, které shromažďují velké objemy transakčních dat, tak zastávají roli, kterou Berry a Linoff [3] nazývají informační zprostředkovatel.

Česká pojišťovna, a.s. je firma, která se specializuje na nabídku životního a neživotního pojištění. Zároveň je hlavním členem finanční skupiny České pojišťovny, do které patří i další finanční instituce, např. Penzijní fond České pojišťovny, a.s., Česká pojišťovna Zdraví, a.s., ČP Invest investiční společnost, a.s. atd. V rámci nabídky produktů České pojišťovny, ale i celé finanční skupiny, která již v současné době obsluhuje více než pět miliónů klientů, se nabízí řada příležitostí pro realizaci křížového prodeje (cross sell). Techniky data miningu, jež umožňují analyzovat např. životní cyklus zákazníka, tak mohou pomoci při identifikaci produktů, které klienti v dané životní etapě právě potřebují.

Techniky data miningu mohou dále pomoci během vyřizování záručních reklamací. S jejich použitím lze např. identifikovat ty reklamace, které nevyžadují detailní zkoumání týmem expertů. U těchto reklamací vznikl nárok klienta na vrácení peněz oprávněně. Jejich automatické vyřízení tak firmě sníží náklady spojené s prací posudkového týmu. Obdobným způsobem by se mohlo postupovat i při vyřizování některých typů pojistných událostí.

Data mining je rovněž používán jako podpora v úsilí firem udržet dobré zákazníky. Problém odchodu klientů ke konkurenci je typický zejména v odvětvích, ve kterých změna dodavatele nepředstavuje příliš vysoké náklady. Zároveň konkurence poskytuje zákazníkům stimuly, které podporují jejich odchod. Pro firmy je mnohem nákladnější přilákat nové klienty, než udržet ty stávající. Avšak i stimuly, které jsou poskytnuty současným zákazníkům za účelem jejich podpory, mohou být pro podnik velmi drahé. Techniky data miningu umožňují rozpoznat klienty, kterým je nutné poskytnout stimul, aby zůstali, dále klienty, kteří zůstanou i bez stimulu, a klienty, které je vhodné nechat odejít [36].

Výše uvedené příklady jsou jen zlomkem možností, které poskytuje data mining. Jejich uplatnění v praxi však není jednoduché. Prvním krokem, který umožní jeho nasazení, je podpora ze strany uživatelů jeho výsledků. Jakmile se dostaví první úspěchy, bude data mining používán stále častěji.

2.14 Typy úloh řešených s použitím technik data miningu

Mezi základní typy úloh, které umožňují řešit techniky data miningu, patří [3]:

- klasifikace hodnocených jednotek (Classification),
- odhad hodnot proměnných (Estimation),
- predikce budoucího chování hodnocených jednotek (Prediction),
- analýza asociací, seskupování na základě podobnosti (Affinity Grouping),
- shlukování (Clustering),
- popis hodnocených jednotek (Description).

2.14.1 Klasifikace

Pojem klasifikace je možné chápat ve dvou různých rovinách. Lukasová a Šarmanová [25] nazývají klasifikací rozklad množiny hodnocených jednotek, který vede k vytvoření systému tříd. Této definici vyhovují metody shlukování, jež rozdělují heterogenní populaci do homogenních podsouborů. Výsledkem shlukovacích algoritmů je tedy klasifikační systém, tj. popsané třídy jednotek.

Naproti tomu Berry a Linoff [3] definují klasifikaci jako proces zkoumání vlastností nových jednotek hodnoceného souboru za účelem jejich zařazení do předem definovaných tříd. Pro úlohu klasifikace jsou tedy charakteristické dobře definované třídy a vývojový soubor, který obsahuje již klasifikované jednotky. Na základě vytvořeného modelu jsou neklasifikované případy zařazovány do jednotlivých tříd. Výstupem klasifikačního modelu jsou diskrétní hodnoty. Příkladem technik, které mohou být použity pro klasifikaci, jsou rozhodovací stromy nebo diskriminační analýza.

V kapitole 2.14 je pod pojmem klasifikace chápána činnost dle definice Berryho a Linoffa. Ve zbytku disertační práce se pak klasifikací rozumí vytvoření systému tříd.

2.14.2 Odhad hodnot proměnných

Dalším typem úlohy, který je možné řešit s použitím technik data miningu, je odhad hodnot proměnných. V tomto případě je k dispozici sada vstupních proměnných, na základě kterých jsou odhadovány hodnoty neznámé spojité proměnné. Techniky používané pro tento typ úlohy, mohou být rovněž použity i pro úlohu klasifikace. Např. pojišťovna, která ke konci každého roku zasílá svým klientům potvrzení o výši zaplaceného pojistného na životní pojištění, si přeje prodat reklamní prostor na obálkách svých dopisů výrobci lyžařského vybavení. Potvrzení o zaplacení pojistného používá pojistník při uplatnění nároku na snížení základu daně z příjmu při ročním zúčtování. Pojišťovna může za účelem rozdělení klientů na lyžaře a nelyžaře sestrojit klasifikační model zohledňující např. skutečnost, zda si klient sjednává v zimních měsících cestovní pojištění. Výstupem tohoto modelu budou diskrétní hodnoty, lyžař – nelyžař. Druhou možností je odhadnout pro každého klienta skóre, např. z intervalu od 0 do 1, které bude vyjadřovat jeho sklon k lyžování. V tomto případě bude výstupem modelu spojitá proměnná. Stanovení hraniční hodnoty, do které je klient považován za nelyžaře a po jejímž dosažení je již lyžařem, je opět úlohou klasifikace. Technikou, která je velmi vhodná pro odhad hodnot neznámé spojité proměnné, je např. neuronová síť.

2.14.3 Predikce budoucího chování

Predikce budoucího chování hodnocených jednotek (např. zákazníků) je podobným úkolem jako klasifikace nebo odhad. Rozdíl spočívá v tom, že jednotky jsou klasifikovány na základě nějakého budoucího chování nebo odhadované budoucí hodnoty. Předpokladem pro použití technik klasifikace nebo odhadu v úloze predikce, je existence vývojového souboru, ve kterém jsou hodnoty predikované proměnné již známy spolu s historickými hodnotami všech proměnných. Historická data jsou používána ke konstrukci modelů, které vysvětlují současné chování jednotek. Jestliže však do těchto modelů vstupují současné hodnoty, výsledkem je predikce budoucího chování. Mezi techniky vhodné pro úlohu predikce patří např. rozhodovací stromy nebo neuronové sítě.

2.14.4 Analýza asociací, seskupování na základě podobnosti

Seskupování jednotek na základě podobnosti (Affinity Grouping) je zvláštní formou shlukování. Jejím cílem je nalézt skupiny položek, které se společně vyskytují v dané transakci (resp. nákupním koši). Výstupem této úlohy je pravděpodobnost, s jakou jsou různé produkty prodávány společně. Tento výstup lze rovněž vyjádřit slovně jako pravidlo. Seskupování jednotek na základě příbuznosti slouží jako podpora pro plánování rozmístění výrobků na regálech supermarketů, k identifikaci cross - sellingových příležitostí a návrhu balíčků produktů nebo služeb. Technikou, která je používána pro řešení tohoto typu úloh, je analýza nákupního koše.

2.14.5 Shlukování

Cílem úlohy shlukování je segmentace heterogenní populace do homogenních podskupin (shluků). Shlukovací algoritmy nedisponují předem definovanými třídami, do kterých by mohly rozdělit hodnocené jednotky. Zároveň nemají k dispozici ani soubor s již zařazenými případy. V souladu s definicí klasifikace Lukasové a Šarmanové [25] vede právě shlukování k vytvoření klasifikačního systému, tj. popsaných tříd jednotek (shluků).

Shlukování lze použít např. pro segmentaci klientů firmy za účelem nalezení bonitních nebo rizikových podskupin. Výstupy shlukovacích algoritmů jsou rovněž často používány i jako vstup pro další data miningové úlohy.

2.14.6 Popis

V některých případech je data mining používán pouze k popisu procesů a dat uložených uvnitř složité databáze. Cílem této úlohy je zlepšit chápání jednotek hodnoceného souboru za účelem snadnějšího vysvětlení jejich chování. Mezi techniky používané pro popis patří průzkumová analýza, která zahrnuje výpočet základních popisných statistik, testy rozdělení proměnných, vizualizaci dat formou grafů atd.

2.15 Techniky data miningu

Obecně je možné rozlišit dva základní přístupy k úloze data miningu. Prvním z nich je testování statistických hypotéz (Hypothesis Testing). Jeho cílem je potvrdit nebo vyvrátit určitý předpoklad, který je vyjádřen formou tvrzení jako nulová hypotéza. V průběhu testování jsou analyzována data, která byla shromážděna pozorováním, nebo

vytvořena realizací experimentu, jako je např. testovací direct mail. Hypotéza, která je definována před vlastní analýzou dat, představuje navrhované vysvětlení pozorovaného jevu.

Druhým přístupem k úloze data miningu je objevování znalostí (Knowledge Discovery). Na rozdíl od testování statistických hypotéz je v rámci tohoto přístupu nejprve prováděna analýza dat a poté na základě jejích výsledků odvozeny dosud neznámé skutečnosti. Objevování znalostí může být buď přímé nebo nepřímé. Úkolem přímého přístupu je vysvětlit hodnoty určité proměnné (např. příjmu, odezvy, věku apod.) na základě hodnot ostatních proměnných. V rámci použití tohoto přístupu jsou řešeny úlohy odhadu hodnot cílové proměnné či predikce chování jednotek. Úkolem nepřímého přístupu je identifikovat významné vlastnosti dat. Jeho použití není zaměřeno na vysvětlení hodnot konkrétní závislé proměnné. Mezi techniky, které jsou vhodné pro nepřímé objevování znalostí, patří např. analýza nákupního koše nebo shluková analýza.

Berry a Linoff [3] rozlišují v rámci přímého a nepřímého přístupu sedm základních technik data miningu:

1. rozhodovací stromy (Decision Trees),
2. neuronové sítě (Neural Networks),
3. genetické algoritmy (Genetic Algorithms),
4. analýzu spojitostí (Link Analysis),
5. analýzu nákupního koše (Market Basket Analysis),
6. dedukci na základě paměti (Memory – Based Reasoning),
7. detekci shluků (Cluster Detection).

Oba přístupy, přímý i nepřímý, je možné vzájemně kombinovat. Pomocí nepřímého přístupu lze rozpoznat závislosti uvnitř dat a pomocí přímého tyto závislosti vysvětlit. Přímé a nepřímé objevování znalostí lze rovněž kombinovat s testováním statistických hypotéz. Např. v této disertační práci bude za účelem segmentace klientů nejprve provedena shluková analýza, která reprezentuje nepřímý přístup. Následně bude provedena analýza kategoriálních dat, v rámci níž bude testována nulová hypotéza o nezávislosti kategoriálních proměnných.

V následujícím textu je dále uveden podrobný přehled technik přímého a nepřímého přístupu dle Berryho a Linoffa [3].

2.15.1 Rozhodovací stromy

Rozhodovací stromy jsou používány pro řešení úloh přímého objevování znalostí, zejména pak pro klasifikaci. Jejich cílem je rozdělit záznamy vývojového souboru do několika podsouborů tak, aby se maximalizovaly rozdíly v závislé proměnné. Každý podsoubor je možné popsat jednoduchým pravidlem, které se týká jednoho nebo více polí [6].

Hlavní výhodou rozhodovacích stromů je jejich snadná interpretace. Jednoznačná pravidla umožňují snadné vyhodnocení výsledků na základě identifikace klíčových atributů procesu. Tato vlastnost rozhodovacích stromů je užitečná i např. v rámci kontroly kvality dat. Nekonzistentnosti, které se v nich mohou objevit, jsou zřejmé z jednoznačných pravidel. Pravidla produkovaná rozhodovacími stromy mohou být rovněž vyjádřena logickými podmínkami dotazovacího jazyka (např. SQL) a poté přímo aplikována na nové záznamy databáze.

2.15.2 Neuronové sítě

Neuronové sítě mohou být použity pro řešení úloh přímého i nepřímého objevování znalostí. V rámci přímého přístupu jsou rozpoznávány vzory v datech vývojového souboru, které jsou generalizovány za účelem klasifikace, odhadu či predikce. Použití neuronových sítí si lze představit jako proces získávání informací a poučení se z každé zkušenosti. V rámci nepřímého přístupu jsou neuronové sítě používány pro segmentaci vytvářením tzv. Kohonenových map [5].

Hlavní výhodou neuronových sítí je široká oblast jejich použití a dostupnost nástrojů, které podporují jejich nasazení. Nevýhodou je obtížná interpretace výsledků a vysoká citlivost na formát a kvalitu vstupních dat. Modely neuronových sítí rovněž velmi rychle zastarávají.

2.15.3 Genetické algoritmy

Genetické algoritmy patří mezi techniky přímého objevování znalostí. Jejich cílem je nalézt optimální sadu parametrů, které popisují prediktivní funkci. Na počátku realizace genetického algoritmu je náhodně zvolena první generace modelů, u kterých je testována schopnost dosáhnout cíle úlohy. Mohou být přitom použity i modely vytvořené pomocí jiných technik data miningu. Selekcí, mutací a křížením operátorů (např. náhodnou záměnou znamének, záměnou proměnných apod.) jsou vytvářeny další

generace, přičemž jejich kvalita je opět testována. Podle zákonů evolučního procesu „přežije“ pouze nejlepší řešení.

Výhodou genetických algoritmů je snadná aplikovatelnost a interpretace jejich výsledků. Nevýhodou je pak jejich výpočetní náročnost a nízká dostupnost v softwarových nástrojích.

2.15.4 Analýza spojitostí

Analýza spojitostí patří mezi techniky nepřímého objevování znalostí. Jejím cílem je sledovat vztahy mezi hodnocenými jednotkami za účelem konstrukce modelů založených na informacích skrytých uvnitř těchto vztahů. Analýza spojitostí je aplikací teorie grafů. Typickou oblastí pro její použití jsou telekomunikace. Každý telefonní hovor spojuje zákazníka s jinou osobou (potenciálním dalším zákazníkem). Tato informace může být základem úspěšné marketingové kampaně spočívající např. v nabídce zvýhodněného programu pro volání účastníků z jedné rodiny apod.

Mezi výhody analýzy spojitostí patří snadná identifikace vazeb mezi jednotkami a možnost jejich vizualizace. Informace obsažená ve vazbě může být rovněž uložena do nového atributu a ten dále využíván jinou technikou data miningu. Nevýhodou analýzy spojitostí je její omezená použitelnost pro úzký okruh aplikačních oblastí a nízká dostupnost v softwarových nástrojích. Používání analýzy spojitostí je rovněž velmi náročné na výpočetní výkon.

2.15.5 Analýza nákupního koše

Analýza nákupního koše patří mezi techniky nepřímého objevování znalostí. Jejím cílem je identifikovat položky, které se zpravidla prodávají spolu, a přiřadit jim pravděpodobnost společného výskytu v rámci jedné transakce. Tento výsledek je možné interpretovat rovněž slovně formou pravidla. Typickým výstupem analýzy nákupního koše je např. pravidlo, že zákazník, který kupuje bonboniéru, kupuje i láhev vína, nebo kupuje-li barvu, kupuje i štětec na natírání, ale ne naopak. Analýza nákupního koše má největší uplatnění v oborech, ve kterých se vyskytují anonymní transakce. Výstupy analýzy nákupního koše slouží např. k plánování uspořádání obchodů, omezení speciální nabídky pouze na jeden výrobek ze souboru společně prodávaných, svazování výrobků atd.

Výhodou analýzy nákupního koše je jednoduchý a snadno pochopitelný výpočetní algoritmus. Snadno pochopitelné a interpretovatelné jsou i jeho výsledky. Výstupy analýzy nákupního koše jsou nejvíce reprezentativní, jestliže všechny položky v něm mají přibližně stejnou četnost výskytu. Pokud nikoliv, je nutné provádět modifikace dat, které však mohou vést ke ztrátě některých informací.

2.15.6 Dedukce na základě paměti

Dedukce na základě paměti je technikou, která umožňuje řešit úlohy přímého objevování znalostí. Jejím cílem je s použitím známých vzorových situací provádět klasifikaci nebo predikci chování nových, dosud neznámých případů. Za tímto účelem hledá v souboru známých situací nejbližší sousedy nového případu, kombinuje jejich hodnoty a na základě výsledku přiřadí nové situaci klasifikační nebo predikční hodnotu. Klíčovými prvky této techniky jsou funkce vzdálenosti, která je používána k nalezení nejbližších sousedů, a kombinační funkce, která kombinuje jejich hodnoty za účelem provedení klasifikace nebo predikce. Příkladem použití dedukce na základě paměti může být systém pro likvidaci pojistných událostí, který na základě způsobu jejich vyřízení v minulosti rozhoduje, zda nová pojistná událost bude vyplacena ihned, nebo zda vyžaduje podrobné zkoumání týmem expertů.

Hlavní výhodou této techniky je skutečnost, že její aplikace nevyžaduje modifikaci zdrojových dat. Velkou výhodou je rovněž schopnost modelu učit se z nových situací. Funkce vzdálenosti i kombinační funkce jsou tedy velmi stabilní při zavádění změn do množiny známých dat.

2.15.7 Detekce shluků

Cílem detekce shluků je rozdělit soubor hodnocených jednotek do stejnorodých podsouborů. Vzhledem k tomu, že jsou hledány podobnosti v datech, které nejsou předem známy, jedná se o techniku nepřímého objevování znalostí. Pro detekci shluků jsou používány geometrické metody, statistické metody a neuronové sítě. Shlukování je možné provádět i jako součást jiných data miningových analýz. Jeho aplikace tak může pomoci počátečnímu pochopení analyzovaných dat.

Výhodou detekce shluků je snadná aplikovatelnost jednotlivých metod. Nevýhodou pak může být jejich citlivost na nastavení počátečních parametrů a někdy také obtížná interpretace výsledků.

2.15.8 Další techniky data miningu

Při realizaci data miningových analýz je možné využívat řady dalších převážně statistických metod. Mezi tyto metody patří např. jednoduché a vícenásobné regresní modely, diskriminační analýza, analýza hlavních komponent, faktorová analýza, parametrické a neparametrické testy. Pro účely prvotního prozkoumání analyzovaných souborů je vhodné provádět průzkumovou analýzu, v rámci které jsou zjišťována rozdělení jednotlivých proměnných, počítány základní popisné statistiky, prováděna vizualizace dat pomocí grafů (např. histogramy, krabicové grafy) atd. Volba metody závisí na charakteru řešené úlohy a typu analyzovaných dat.

2.16 Shluková analýza

Tato disertační práce bude zaměřena segmentaci klientů České pojišťovny, a.s. s použitím statistických metod. Za tímto účelem budou použity zejména metody shlukové analýzy. Shluková analýza patří mezi vícerozměrné statistické metody a z pohledu data miningu se jedná o nepřímé objevování znalostí. Jejím cílem je rozdělit soubor statistických jednotek podle většího počtu statistických znaků tvořících nedílnou informaci do stejnorodých podsouborů (shluků).

Existuje velké množství definic shlukové analýzy. Jako příklad lze uvést následující:

- Shluková analýza je metoda, která se zabývá tříděním pozorování do skupin tak, aby stupeň přirozené asociace členů téže skupiny byl co nejvyšší a členů různých skupin co nejnižší [1].
- Shluková analýza je proces třídění statistických jednotek do podsouborů, které mají význam v souvislosti se zvláštním problémem. Jednotky jsou tedy organizovány do takové reprezentace, která dobře charakterizuje populaci, z níž byl analyzovaný vzorek vybrán [21].
- Shluková analýza je metoda, která se zabývá vyšetřováním podobnosti vícerozměrných objektů (tj. objektů, u nichž je změřeno větší množství proměnných) a jejich uspořádáním do tříd neboli shluků. Je vhodná zejména tam, kde objekty projevují přirozenou tendenci se seskupovat [27].

Shlukovou analýzu je možné aplikovat na následující typy úloh [33]:

- klasifikace jednotek ve smyslu vytvoření systému tříd (např. segmentace klientů firmy, segmentace výrobků apod.),
- výběr reprezentativního vzorku dat z rozsáhlého datového souboru,
- snížení dimenze úlohy vyjádřením několika vlastností pozorování pomocí jedné proměnné – příslušností ke shluku.

2.16.1 Typy metod shlukové analýzy

Metody shlukové analýzy je možné rozdělit podle způsobu, jakým provádějí shlukování, na metody hierarchické a nehierarchické. Hierarchické metody vedou k hierarchické klasifikaci, tj. k rozkladu jednotek do tříd, jež mohou být dále děleny. V případě, že vycházíme z množiny všech jednotek jako celku a jejím postupným dělením až na jednotlivé objekty vytváříme systém shluků, hovoříme o divizivním shlukování. V opačném případě, kdy jsou seskupovány jednotky s nejmenší vzdáleností až na úroveň jediného shluku, hovoříme o shlukování aglomerativním. Výhodou hierarchických metod je skutečnost, že v průběhu shlukování nevyžadují informaci o optimálním počtu shluků. Tento počet je určován až dodatečně. Vzhledem k tomu, že hierarchické shlukovací algoritmy jsou zpravidla založeny na výpočtu matice vzdáleností mezi všemi jednotkami, není jejich použití vhodné pro velké objemy dat.

V rámci nehierarchického přístupu rozlišujeme metody optimalizační a metody analýzy módů. Cílem optimalizačních metod je nalézt takový rozklad množiny jednotek, který je optimální podle vhodně zvoleného kritéria optimality. Pro optimalizační algoritmy je typické, že začínají stanovením počátečního rozkladu jednotek do k shluků. Tento rozklad je pak postupně zlepšován, přičemž počet shluků zůstává buď zachován, nebo se mění v závislosti na řídících parametrech. Z tohoto pohledu dělíme algoritmy optimálního rozkladu na algoritmy, které zachovávají počet shluků a algoritmy, které tento počet mění. Problémem při použití optimalizačních metod je nalezení vhodného počtu shluků, jež budou tvořit optimální rozklad, a dále nalezení optimálního rozkladu. Hledání optima zvoleného kritéria vede zpravidla k nalezení pouze lokálního extrému. Výhodou optimalizačních metod je možnost jejich použití při shlukování velkých objemů dat.

Metody označované jako analýza módů vycházejí z chápání znaků jednotek jako náhodných veličin, které jsou buď spojité nebo diskrétní. Shluky jsou hledány v blízkosti módu frekvenční funkce příslušného náhodného vektoru.

2.16.2 Standardizace proměnných

Cílem standardizace je zabezpečit, aby všechny proměnné, které vstupují do shlukovacích algoritmů, měly stejný vliv na výsledek. U proměnných s větším rozptylem je tento vliv větší než u proměnných s menším rozptylem. Provedení standardizace je tedy vhodné zejména v případech, kdy jsou jednotlivé proměnné evidovány v různých jednotkách měření. Standardizací jsou jejich hodnoty převáděny do tvaru, ve kterém mají aritmetický průměr roven nule a směrodatnou odchylku rovnu jedné. Standardizace je prováděna pouze pro spojité proměnné.

Kategoriální proměnné mohou představovat dichotomické znaky, jejichž stavy nabývají hodnot „pravda“ nebo „nepravda“. Stav „pravda“ je nejčastěji označován jedničkou a stav „nepravda“ nulou. V případě, že proměnné obsahují více kategorií, jsou obvykle převáděny na soustavu dichotomických znaků. Tzn., že pro každou kategorii je vytvořen nový znak, přičemž jeho stavy nabývají hodnot jedna, je-li výrok pravdivý, nebo nula, je-li nepravdivý. Např. původní proměnnou Frekvence placení je možné převést na soustavu znaků Pojistné je placeno měsíčně, Pojistné je placeno ročně apod.

Některé shlukovací algoritmy předpokládají, že zkoumané jednotky jsou popsány pouze dichotomickými znaky. Pro tyto algoritmy je nutné převést do dichotomické podoby i spojité proměnné, např. jejich rozdělením do kategorií dle intervalů. Každý interval je poté použit jako nový znak, který nabývá hodnot nula nebo jedna.

2.16.3 Přístupy k hodnocení podobnostních vztahů

Vzájemná podobnost zkoumaných jednotek a její kvantitativní vyjádření je jedním ze základních problémů shlukové analýzy. V některých případech je způsob hodnocení podobnosti vázán přímo na shlukovací metodu. Mezi základní typy měr pro hodnocení podobnostních vztahů patří [25]:

1. koeficienty asociace,

2. koeficient korelace,
3. metriky.

Koeficienty asociace jsou určeny pro jednotky charakterizované výhradně dichotomickými znaky. Asociaci dvou jednotek lze popsat pomocí asociační tabulky, jež obsahuje počty znaků, u kterých nastala pozitivní nebo negativní shoda nebo u kterých shoda nenastala. Tyto četnosti jsou použity k výpočtu hodnot koeficientu asociace, jež tvoří matici měr podobnosti. Hodnoty matice se pohybují v intervalu <0, 1> nebo <-1, 1> podle typu koeficientu. Mezi koeficienty asociace používané ve shlukové analýze patří např. [25]:

- Jaccardův koeficient,
- Diceův koeficient,
- Sokalův a Michenerův koeficient,
- Rogersův a Tanimotoův koeficient,
- Russellův a Raoův koeficient,
- Hamannův koeficient.

Další mírou pro hodnocení podobnosti je korelační koeficient. Jeho použití vede k vytvoření matice korelací, neboli matice měr podobnosti, která má na hlavní diagonále samé jedničky. Hodnoty této matice se pohybují v intervalu <- 1, 1>.

K hodnocení podobnostních vztahů jsou rovněž používány metriky, které vycházejí z geometrického modelu matice dat. S pomocí metrik je možné měřit vzdálenost mezi jednotkami. Jejich použití tedy vede k odvození matice měr vzdálenosti, která má na hlavní diagonále samé nuly. Mezi metriky patří např. [14]:

- euklidovká vzdálenost,
- Čebyševova vzdálenost,
- Hemmingova vzdálenost,
- Minkowského vzdálenost.

2.16.4 Metody hierarchického shlukování

Hierarchické shlukovací algoritmy je možné rozdělit na aglomerativní a divizivní. Na počátku aglomerativního shlukování tvoří každý objekt jednoprvkový shluk. Na základě vyhodnocení podobnostních vztahů mezi shluky podle stanovené míry dochází k jejich postupnému slučování. Míra podobnosti mezi shluky je zpravidla definována na základě vzájemné podobnosti jednotek, které tyto shluky tvoří.

Algoritmus končí sloučením všech jednotek do jediného shluku. Průběh aglomerativního shlukování je možné znázornit také graficky prostřednictvím dendogramu. Ten zaznamenává ve dvou vzájemně kolmých směrech monotónní posloupnost shlukovacích hladin a pořadová čísla jednotek seřazená tak, aby bylo možné znázornit posloupnost slučování. Mezi metody aglomerativního shlukování patří např. [14]:

- metoda nejbližšího souseda,
- metoda nejvzdálenějšího souseda,
- Sokalova–Sneathova metoda,
- centroidní metoda,
- Wardova metoda.

Divizivní metody vytvářejí hierarchický systém rozkladů množiny jednotek postupným dělením existujících shluků. Na počátku divizivního shlukování tvoří všechny jednotky jeden shluk, který je postupně dělen na menší shluky. Algoritmus končí rozdělením všech existujících shluků na jednotlivé objekty. V procesu divizivního shlukování je každý existující shluk zpravidla rozdělen na dva další shluky, tj. při rozkladu množiny *n* jednotek existuje celkem 2^{n-1}- 1 možností. Tento postup je tedy vhodný pouze pro rozklad malého počtu jednotek. Proto Lukasová a Šarmanová [25] považují za nejvhodnější divizivní algoritmus MacNaughton-Smithovu metodu z důvodu její aplikovatelnosti na rozsáhlejší soubory dat.

2.16.5 Metody nehierarchického shlukování

Před aplikací optimalizačních metod je nutné stanovit, v jakém smyslu má být rozklad množiny jednotek optimální. Optimálního rozkladu je dosaženo tehdy, jestliže funkcionál kvality rozkladu, neboli kritérium optimality, nabývá extrémní hodnoty. Funkcionál kvality rozkladu by měl vyjadřovat některou z následujících vlastností shluků tvořících rozklad [25]:

1. vzájemnou podobnost jednotek uvnitř shluku,
2. míru separace shluků,
3. homogenitu rozložení jednotek uvnitř shluků,
4. rovnoměrnost rozložení jednotek do různých shluků, příp. kombinace vlastností 1 až 4.

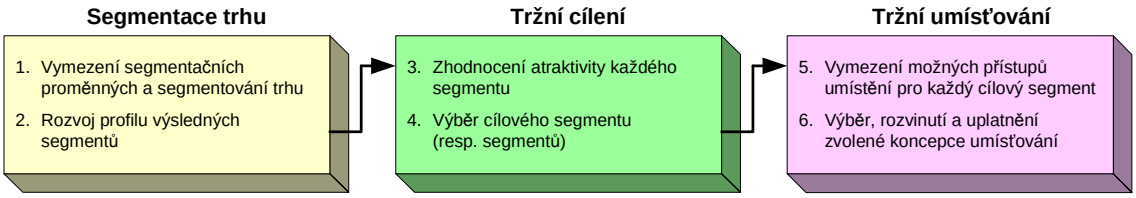
Mezi optimalizační metody patří např. [25]:

- Forgyova a Janceyova metoda,
- MacQueenova k-průměrová metoda,
- MacQueenova a Wishartova metoda,
- MacQueenova metoda se dvěma parametry.

2.17 Obecný přístup k segmentaci trhu

Z pohledu marketingu je pod pojmem trh chápána množina všech skutečných a potenciálních kupců určitého výrobku nebo služby [24]. Firmy se postupně přesvědčily, že nemohou uspokojit všechny zákazníky. Zákazníků je příliš mnoho, jsou prostorově rozptýleni a liší se ve svých kupních požadavcích a praktikách. Firmy, místo toho aby soutěžily s konkurencí celého trhu, se snaží identifikovat pouze jeho nejatraktivnější části ,segmenty, které by mohly nejlépe obsloužit.

Srdce strategického marketingu tvoří tzv. STP marketing (Segmenting – Targeting – Positioning), tj. segmentace, cílení a umísťování [23]. Segmentace trhu je proces, ve kterém jsou kupující rozděleni do skupin, jež mohou požadovat specifické výrobky nebo marketingové mixy. Tržní cílení je proces měření atraktivity jednotlivých segmentů a výběr jednoho nebo několika z nich pro podnikání. Tržní umísťování je pak uplatnění nabídky firmy na cílovém trhu. Jednotlivé etapy STP marketingu jsou znázorněny na obrázku 12.



Obr. 12 Tržní segmentace, zacílení a umístění [23]

Cílený marketing umožňuje společnostem lépe rozpoznávat tržní příležitosti. Stále častěji má charakter mikromarketingu, tzn. že jednotlivé marketingové programy jsou přizpůsobeny skupinám zákazníků na lokální bázi (např. regionu). Nejvyšší formou cíleného marketingu je pak marketing na míru, kde jsou výrobky nebo služby přizpůsobeny potřebám a přáním konkrétního zákazníka.

Segmentace trhu je podle Kotlera [23] obvykle prováděna v následujících krocích:

1. fáze dotazování,
2. fáze analyzování,
3. fáze profilace.

Ve fázi dotazování provádí výzkumník neformální rozhovory se skupinami zákazníků s cílem porozumět jejich motivacím, postojům a chování. Tento výzkum obvykle zpracovávají specializované výzkumné společnosti. Výsledkem jsou formalizované studie významných tržních segmentů, s jejichž použitím firmy snadno identifikují potenciální tržní příležitosti.

Tato disertační práce bude zaměřena na segmentaci klientů České pojišťovny. Zdrojem dat pro analýzu bude firemní datový sklad. Tzn., že data nepochází z výzkumu, nýbrž byla shromážděna v rámci dosavadních obchodních aktivit firmy.

Ve fázi analyzování je obvykle prováděna shluková analýza. Jejím cílem je vytvořit vnitřně homogenní, ale navzájem velmi odlišné shluky klientů.

Ve třetí fázi je vymezen profil každého segmentu ve smyslu zdůraznění odlišností v postojích, chování a demografických a psychografických rysech jeho členů. Segmenty je možné pojmenovat podle dominantních charakteristik.

2.18 Segmentační proměnné pro spotřebitelské trhy

Kotler [23] definuje trh spotřebitelů jako trh jednotlivců nebo domácností, které nakupují zboží nebo služby pro osobní spotřebu. Segmentační proměnné, které jsou obvykle používány pro třídění spotřebitelů, je možné rozdělit do dvou základních skupin. První z nich umožňuje formovat segmenty podle charakteristik klientů. Jedná se o jejich geografické, demografické a psychografické rysy. Druhou skupinu pak tvoří proměnné, které charakterizují chování klienta.

2.18.1 Segmentace klientů na základě jejich geografických, demografických a psychografických charakteristik

S použitím geografických charakteristik lze rozdělit trh na různé územní jednotky (např. státy, regiony nebo okresy). Firma se může rozhodnout, že bude působit pouze v jedné, nebo v několika málo, nebo ve všech oblastech. Podle lokality je zpravidla nutné přizpůsobit se místním proměnám potřeb a preferencí zákazníků. Např. Česká Pojišťovna, a. s. poskytuje své služby vedle České republiky také na Slovenku nebo v Rusku. Mezi geografické proměnné by mohli patřit např. velikost regionu, sklon oblasti k výskytu pojistného nebezpečí (např. povodně nebo zemětřesení) apod.

Druhou skupinu charakteristik tvoří demografické proměnné. Jedná se např. o věk klienta a etapu jeho životního cyklu. Požadavky a schopnosti spotřebitele se s věkem mění. O důchodovou pojistku budou mít zájem většinou klienti středního věku, zatímco o kapitálové životní pojištění mladé páry, které tak mohou financovat bydlení pořízené s použitím hypotéčního úvěru. Kotler [23] však poukazuje na možná úskalí věku jako identifikátoru fáze životního cyklu. Např. zatímco někteří 35letí zakládají rodinu, jiní jsou již v tomto věku prarodiči. Mezi další demografické proměnné patří např. pohlaví klienta, jeho příjem, velikost rodiny, povolání, vzdělání, národnost apod.

Třetí skupinu charakteristik klientů tvoří jejich psychografické rysy. Do této skupiny patří např. sociální třída, životní styl, osobnost klienta apod.

2.18.2 Segmentace na základě charakteristik chování klientů

Mezi segmentační proměnné charakterizující chování klientů, patří např. užitek, který zákazníci očekávají od daného výrobku nebo služby. U pojištění odpovědnosti z provozu motorového vozidla lze např. očekávat skupinu řidičů, kteří jsou pojištěni pouze z důvodu splnění zákonné povinnosti. Jiná skupina si však může navíc ještě sjednat havarijní pojištění, pojištění asistence, pojištění zavazadel apod.

Dalšími segmentačními proměnnými mohou být např. uživatelský status či stupeň používání výrobku nebo služby. Z pohledu uživatelského statusu lze segmentovat trh na neuživatele, bývalé uživatele, potenciální uživatele, uživatele poprvé a pravidelné uživatele [23]. Firmy s velkým tržním podílem mají zpravidla zájem o potenciální uživatele, které se snaží proměnit na uživatele skutečné. Naopak menší

firmy se většinou snaží přesvědčit uživatele výrobků konkurenčních firem, aby přešli na jejich značku.

Z hlediska stupně používání výrobku nebo služby je možné rozlišit lehké, střední a silné uživatele [23]. Silní uživatelé často tvoří malé procento trhu, ale zároveň představují velký podíl na celkové spotřebě, tj. i na příjmech firmy za daný výrobek nebo službu. Např. v oblasti cestovního pojištění mohou být silnými uživateli lidé, kteří často cestují za obchodem do zahraničí.

Mezi segmentační proměnné, které charakterizují chování zákazníka, patří rovněž status jeho věrnosti. Kotler [23] rozlišuje v této souvislosti následující čtyři skupiny klientů:

1. kmenoví příznivci – zákazníci, kteří vždy kupují tutéž značku,
2. slabí příznivci – zákazníci, kteří jsou věrni dvěma nebo třem značkám,
3. proměnliví příznivci – zákazníci, kteří přecházejí z jedné značky na jinou,
4. nestálí zákazníci – zákazníci, kteří nejsou věrni žádné značce.

V případě pojišťovny by bylo možné hodnotit věrnost zákazníka na základě doby trvání jeho smluvního vztahu. Tuto dobu lze např. počítat jako dobu platnosti jeho nejstarší aktivní smlouvy. Analýza odlivu zákazníků je jednou z úloh data miningu. Problém odchodu zákazníků je typický zejména pro odvětví, ve kterých změna dodavatele není příliš nákladná a konkurence zároveň masivně láká klienty na svou stranu. Pozorováním zákazníků, kteří opouštějí značku, může firma identifikovat slabé stránky svého marketingu.

Mezi další charakteristiky chování klientů patří např. jejich připravenost ke koupi, postoje a příležitosti.

2.19 Segmentační proměnné pro obchodní trhy

Pod pojmem obchodní trhy chápe Kotler [23] trhy průmyslové, trhy zprostředkovatelů, vládní trhy a trhy nevýdělečných organizací a zahraniční trhy. Průmyslové trhy tvoří organizace, které nakupují zboží a služby za účelem výroby jiných výrobků nebo služeb, jež dále prodávají. Zprostředkovatelské trhy tvoří společnosti, které nakupují zboží a služby pouze za účelem jejich dalšího prodeje. Vládní trhy a trhy nevýdělečných organizací tvoří agentury, které nakupují zboží a

služby za účelem zabezpečení veřejných nebo charitativních činností. Zahraniční trhy tvoří kupující, kteří se nacházejí mimo území daného státu a mezi než patří zahraniční spotřebitelé, výrobci, zprostředkovatelé a vlády. Z pohledu výše uvedeného členění prodává Česká pojišťovna, a.s. své služby na všech typech trhů, od spotřebitelských až po zahraniční.

Podobně jako u spotřebitelských trhů patří i u obchodních mezi segmentační proměnné např. očekávaný užitek a stupeň užívání výrobku, geografické faktory apod. Obchodní trhy však mají i svá vlastní specifika. Segmentaci zákazníků lze provádět i podle dalších kritérií, jako jsou např.:

- demografické proměnné (tj. velikost firmy, odvětví zákazníka apod.),
- kritéria provozu (tj. zákazníkem používaná technologie, schopnosti zákazníka z hlediska jeho požadavků na množství dodávaných služeb apod.),
- nákupní přístupy (tj. organizace nákupu zákazníkem – centralizovaná, decentralizovaná, nákupní kritéria – preference kvality, nízké ceny, servisu apod.),
- faktory situace (tj. naléhavost nákupu, specifické aplikace výrobku, velikost objednávky apod.),
- osobní charakteristiky nákupčího (tj. jeho postoj k riziku, věrnost apod.).

2.20 Hodnocení tržních segmentů

Při hodnocení tržních segmentů je nutné zohlednit následující tři faktory [23]:

1. velikost a růst segmentu,
2. strukturální přitažlivost segmentu,
3. cíle a zdroje firmy.

2.20.1 Velikost a růst segmentu

Pro každou společnost je optimální velikost tržního segmentu jiná. Velké firmy se zpravidla soustředí na obsluhu segmentů s velkým tržním potenciálem a přehlíží malé segmenty. Naproti tomu malé firmy dávají přednost malým segmentům, neboť velké vyžadují značné zdroje.

Růst segmentu je obecně požadovaná vlastnost, poněvadž firmy mají zájem o rostoucí prodeje a zisky. Rostoucí segment však na druhé straně přitahuje i konkurenci, která často jeho ziskovost sníží.

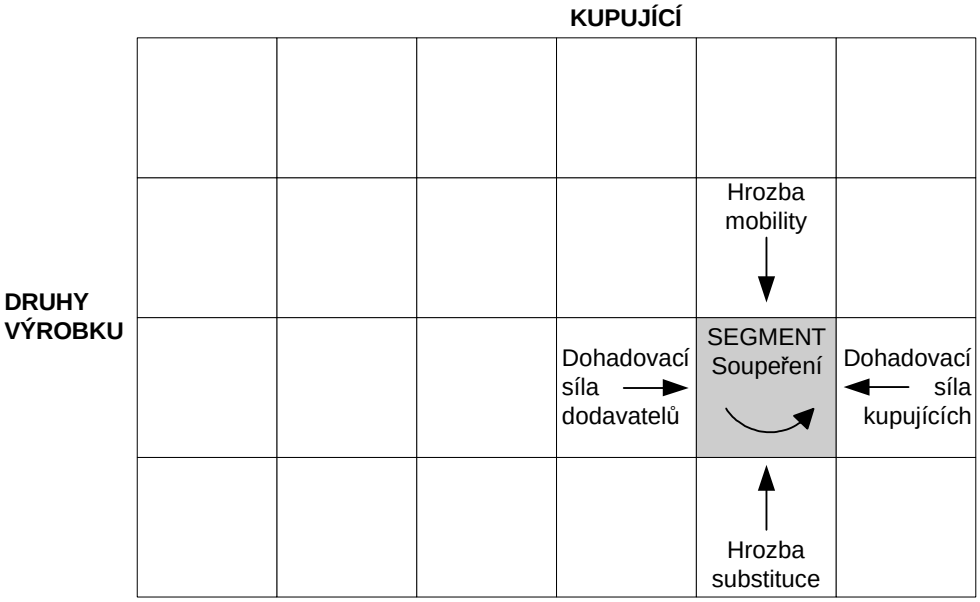
2.20.2 Strukturální přitažlivost segmentu

Porter [32] rozlišuje mezi odvětvovou a tržní segmentací. Cílem tržní segmentace je rozpoznat rozdíly v potřebách a chování zákazníků za účelem jejich obsluhy odlišnými prodejními programy, které odpovídají možnostem podniku. Segmentace trhu je tedy úzce spjata s marketingovou činností. Naproti tomu úkolem segmentace odvětví je spojit chování zákazníků s chováním nákladů, a to jak výrobních, tak i těch, které jsou vynakládány na jejich obsluhu. Umožňuje tak ukázat rozdíly ve strukturální přitažlivosti jednotlivých segmentů a konflikty, které vznikají, jestliže firma obsluhuje segmentů několik. Toto širší pojetí segmentace může být pro podnik základem k vytvoření a udržení konkurenční výhody.

Porter [32] dále definuje odvětví jako trh, na kterém jsou kupujícím prodávány podobné nebo spolu úzce související výrobky. K segmentaci odvětví tedy používá dimenze druh výrobku a kupující, přičemž kupující jsou podrobněji charakterizováni jejich typem, distribuční cestou obsluhy a geografickým sídlem. Strukturální přitažlivost každého segmentu je dána působením pěti dynamických faktorů:

1. dohadovací silou dodavatelů,
2. hrozbou mobility (tj. hrozbou vstupu nových konkurentů do segmentu),
3. dohadovací sílou kupujících,
4. hrozbou substituce novými výrobky,
5. soupeřením se stávající konkurencí v segmentu.

Působení těchto pěti dynamických sil v rámci daného segmentu je znázorněno na obrázku 13.



Obr. 13 Rozdíly mezi segmenty v pěti dynamických faktorech [32]

Analýza pěti dynamických faktorů na úrovni segmentu je odlišná od analýzy na úrovni celého odvětví. Např. potenciální konkurenty představují jak firmy, které působí v jiných segmentech, tak i ty, které dosud v odvětví nepůsobí. Substituty daného výrobku nebo služby mohou být jeho jiné druhy z daného odvětví, ale i výrobky pocházející z jiných odvětví. Soupeření v segmentu se účastní jak firmy, které se zaměřily výhradně na tento segment, tak i ty, které současně působí také v jiných segmentech. Dohadovací síla kupujících i dodavatelů bývá obvykle specifická pro daný segment, ale může být ovlivněna i nákupy kupujících v jiných segmentech nebo dodávkami dodavatelů do jiných segmentů. Strukturální analýza daného segmentu je tedy silně ovlivněna situací v jiných segmentech.

Aby firma dokázala dostatečně posoudit všechny síly, které ovlivňují nebo mohou ovlivnit její činnost v rámci segmentu, je vhodné dále provést i širší analýzu vnějšího prostředí podniku (tj. analýzu odvětví a analýzu globálního prostředí). Tuto analýzu lze podle Hrona [15] rozdělit do následujících subanalýz:

- STEP analýza neboli analýza globálního prostředí,
- analýza ekonomických charakteristik odvětví,
- analýza hybných sil odvětví,
- strukturální analýza na úrovni odvětví neboli analýza konkurence,
- strategické mapy,

- analýza konkurentů,
- analýza atraktivity odvětví.

2.20.3 Cíle a zdroje firmy

V případě, že je segment pro firmu dostatečně velký, má rostoucí potenciál a přitažlivou strukturu, musí podnik dále posuzovat vlastní zdroje a cíle. Ve vztahu k segmentu, ale i k celému odvětví, je vhodné provést analýzu vnitřního prostředí podniku. Do této analýzy lze podle Hrona [15] zahrnout následující subanalýzy:

- evaluaci dosavadní strategie,
- analýzu výsledků v jednotlivých funkcionálních oblastech,
- analýzu exponovanosti,
- portfolio analýzu,
- SPACE analýzu,
- klíčové faktory úspěchu,
- analýzu konkurenceschopnosti.

Na základě posouzení podmínek vnějšího prostředí, analýzy vnitřních zdrojů podniku a zájmových skupin, je možné identifikovat slabé a silné stránky firmy v tzv. SWOT analýze. Posouzení slabých a silných stránek vede k vytvoření strategických alternativ, které firma může použít ve vztahu k segmentu a odvětví jako celku.

2.21 Výběr vhodného segmentu pro podnikání

Ačkoliv volba strategie ve vztahu k odvětvovým nebo tržním segmentům závisí na konkrétních podmínkách vnějšího a vnitřního prostředí firmy, obecně je možné rozlišit pět základních přístupů k výběru cílového trhu [23]:

1. soustředění se pouze na jeden segment,
2. výběrová specializace,
3. výrobková specializace,
4. tržní specializace,
5. pokrytí celého trhu.

Nejjednodušším případem volby cílového trhu je soustředění se pouze na jeden segment, který přirozeně odpovídá možnostem firmy, nebo který je východiskem pro další expanzi. Prostřednictvím soustředného marketingu firma získává silnou pozici v segmentu a může dosáhnout značných provozních úspor v důsledku specializace výroby, distribuce a propagace. Na druhé straně působení pouze v jednom segmentu představuje pro podnik nadprůměrné riziko. Daný segment totiž může ztratit svou výnosnost nebo může být obsazen nebezpečným konkurentem.

Výběrová specializace znamená, že firma působí současně ve více segmentech, z nichž každý je pro ni přitažlivý. Vícesegmentová koncentrace je oproti zaměření pouze na jeden segment pro podnik výhodnější zejména z důvodu nižšího rizika. Jestliže některý ze segmentů přestane být atraktivní, lze nadále vydělávat v ostatních.

V případě, že se firma rozhodne pro výrobkovou specializaci, zaměří se na jeden výrobek, který prodává více segmentům. Prostřednictvím této strategie získává dobrou pověst v dané oblasti, ale zároveň je vystavena nebezpečí, že v případě objevení nové substituční technologie ztratí své tržní pozice.

Tržní specializace znamená, že se firma soustředí na uspokojování různých potřeb určité tržní skupiny. Podnik tak získává dobré jméno a stává se výhradním dodavatelem všech nových výrobků, které by jeho zákazníci mohli pravděpodobně potřebovat. Riziko existuje v případě, že tržní segment náhle omezí své nákupy.

V případě, že se podnik rozhodne pokrýt celý trh, snaží se o uspokojení všech zákaznických skupin. Tento typ strategie si mohou dovolit pouze velké firmy. Obsluhu celého trhu lze provádět dvěma základními způsoby, diferencovaným nebo nediferencovaným marketingem.

Nediferencovaný marketing přehlíží rozdíly v segmentech a uplatňuje na celém trhu pouze jeden typ nabídky. Zaměřuje se přitom na společné vlastnosti zákazníků, nikoliv na jejich odlišnosti. Firma se spoléhá na hromadnou distribuci a hromadnou reklamu. Výhodnou nediferencovaného marketingu jsou především nízké náklady. Nevýhodou pak to, že firmy vytvářejí nabídku pouze pro největší segmenty, o které vzájemně soupeří, což vede k neuspokojení menších segmentů. V důsledku silné konkurence jsou velké segmenty rovněž méně ziskové.

V případě diferencovaného marketingu působí firmy ve většině segmentů, ve kterých používají různé marketingové programy. Oproti nediferencovanému přístupu přináší diferencovaný marketing větší celkový prodej, což je důsledkem nabídky širšího sortimentu a větší rozmanitosti distribučních cest. Na druhé straně však vede i k vyšším nákladům. V některých případech firmy zjišťují, že „přesegmentovaly“ svůj trh, a proto se snaží o uplatnění jednoho výrobku ve více segmentech. Tento postup nazývá Kotler [23] zpětnou segmentací.

2.22 Základní informace o České pojišťovně, a.s.

Použití technik data miningu bude v této disertační práci prezentováno na příkladu z prostředí České pojišťovny, a.s. V následujícím textu jsou proto uvedeny základní informace o této firmě.

Česká pojišťovna je univerzální pojišťovací instituce, jejíž tradice sahá až do roku 1827. Svým klientům nabízí prakticky všechny druhy pojištění (životní, majetkové i odpovědnostní). K tomu využívá více než 700 obchodních míst po celé České republice. Své aktivity však rozvíjí i v zahraničí, např. na Slovensku nebo v Rusku. Česká pojišťovna patří mezi nejúspěšnější společnosti v České republice, kde v konkurenci více než čtyřiceti pojistitelů obhospodařuje téměř čtyřicet procent pojistného trhu. V roce 2004 se umístila na devátém místě v žebříčku nejúspěšnějších českých firem Czech Top 100 [13]. V roce 2005 pak obsadila první místo v hlavní kategorii ankety Pojišťovna roku 2005, kterou vyhlašuje Asociace českých pojišťovacích makléřů. Současně se Česká pojišťovna stala zlatou Pojišťovnou roku 2005 také v prestižní soutěži MasterCard Banka roku. Její základní kapitál činil v roce 2005 2,98 miliardy korun. Prostředky klientů, které tato společnost spravuje, přesahují 80 miliard korun [38].

Datový sklad společnosti zpracovává a uchovává data z více než dvaceti provozních systémů a externích zdrojů. Jejich objem se v současné době pohybuje okolo 1,5 terabytů. Datový sklad podporuje rozhodovací procesy na všech úrovních řízení firmy.

Jedním z dlouhodobých cílů České pojišťovny je udržení a posílení jejího tržního podílu [38]. Data mining a jeho techniky se tak mohou stát účinným nástrojem, který podpoří jeho dosažení.

3 Cíle disertační práce

Tato disertační práce bude zaměřena na segmentaci klientů České pojišťovny, a.s. provedenou na základě výsledků data miningové analýzy. Řešení bude zasazeno do širšího rámce procesů, které s realizací data miningových analýz souvisí. Zvýšená pozornost bude věnována zejména přípravě dat pro modelování a dále kontrole a hodnocení jejich kvality. Vedle problematiky vstupních dat se práce bude zabývat také využitím výsledků data miningu v obchodní praxi.

Realizace data miningových analýz často vyžaduje kombinaci různých metodických postupů. Z tohoto pohledu bude disertační práce kombinovat různé statistické metody za účelem nalezení vhodného postupu pro klientskou segmentaci. Vzhledem k tomu, že kvalita výsledků data miningových analýz závisí na kvalitě vstupních dat stejně jako na metodách pro jejich analýzu, budou v práci navrženy metodické postupy pro jejich čištění, kontrolu a hodnocení kvality. Cíle této práce jsou proto následující:

1. S použitím statistických metod segmentovat klienty České pojišťovny, a. s. za účelem nalezení vhodného postupu pro klientskou segmentaci a vymezení tržních segmentů, které firma obsluhuje.
2. Profilovat jednotlivé klientské segmenty a navrhnout strategie dalšího přístupu k nim.
3. Navrhnout algoritmus pro vytvoření jednoznačné identifikace klientů a algoritmus pro jejich deduplikaci. Tyto algoritmy budou navrženy v metodické části práce, přičemž jejich cílem je vytvořit jednotný pohled na klienty a zvýšit tak významnost výsledků statistických metod.
4. Vytvořit funkční návrh obecného systému kontrol a vyhodnocení kvality dat v datovém skladu. Tento systém bude navržen v diskusi, která se bude věnovat obecným principům řízení kvality firemních dat.

Za účelem splnění výše uvedených cílů budou použita data České pojišťovny, a. s. Z důvodu ochrany informací je podmínkou pro jejich použití jejich modifikace a dále výběr a zpracování pouze vzorku klientů. Závěry disertační práce tedy nelze považovat za obraz skutečné obchodní situace v České pojišťovně. Algoritmy modifikace dat a výběru vzorku klientů nesmí být v disertační práci zveřejněny.

4 Metodika zpracování

4.1 Příprava dat pro statistickou analýzu

Pro účely statistické analýzy budou použita data, která Česká pojišťovna shromažďuje v souvislosti se správou pojistných smluv neživotního a životního pojištění. Zájmovou skupinou klientů jsou ekonomické subjekty (firmy), které na aktuálně platných pojistných smlouvách vystupují v roli pojistníka. Pojistníkem se rozumí osoba, která platí pojistné za pojištění. Vedle interních dat vstupují do analýzy také externí data, která Česká pojišťovna kupuje a která obsahují veřejné informace o firmách (např. z Registru ekonomických subjektů). Centrálním úložištěm interních i externích dat je firemní datový sklad.

V datovém skladu jsou data uložena v relačním databázovém modelu. Před provedením statistické analýzy budou transformována do ploché tabulky (flat table), která bude sloužit jako její vstup. Pro účely této disertační práce bude plochá tabulka vytvořena přímo nad daty základní vrstvy. V případech, kdy je k dispozici data mart obsahující všechna relevantní data, může být vytvořena i nad ním. Tvorba ploché tabulky pro statistickou analýzu odpovídá na obrázku 10 činnostem Předzpracování a Transformace. V ploché tabulce budou všechny informace o klientovi včetně informací z jeho pojistných smluv koncentrovány do jednoho záznamu. Unikátním identifikátorem každého záznamu bude validní identifikační číslo (IČ).

V této kapitole jsou popsány algoritmy, podle kterých budou data datového skladu transformována do finálního souboru. Jejich cílem je ověřit kvalitu vstupních dat a provést jejich transformaci. Z pohledu metodologie SEMMA se jedná o činnosti z etap Explore a Modify. Z pohledu metodologie CRISP – DM pak o činnosti z etap Porozumění datům a Příprava dat.

4.1.1 Identifikace klientů a deduplikace záznamů o nich

Obecným problémem práce s klientskými daty je způsob identifikace unikátního záznamu o klientovi ve firemní databázi. Klientská data často obsahují duplicity, které mohou mít různou příčinu. Např. klient se vyskytuje v databázi tolikrát, kolik má smluv. Další příčinou mohou být chyby způsobené špatným zadáním dat o klientovi do provozního systému nebo procesy následné manipulace s daty.

Pro účely této disertační práce bude identifikátorem záznamu klienta identifikační číslo (IČ). Ve vstupní databázi datového skladu jsou s tímto identifikátorem spojeny následující problémy:

1. Duplicitní záznamy pro jedno IČ.
2. IČ může být nevalidní.
3. IČ je validní, ale patří jinému klientovi.

Při tvorbě ploché tabulky tedy bude ověřena formální správnost IČ. Dále bude ověřena jeho věcná správnost. Např. bude posouzena použitelnost validního IČ ve spojení s různými kombinacemi obchodního jména pro statistickou analýzu. V prvním případě bude zkoumána jeho inherentní informační kvalita, neboli validita, v druhém pak jeho pragmatická informační kvalita, neboli užitečnost při podpoře podnikových procesů. Na vstupní data tedy budou aplikovány algoritmy, které ověří validitu IČ a které dále vytvoří identifikaci všech klientských záznamů pomocí věcně a formálně správných IČ. Interní data o klientech budou dále spojena s externími daty za účelem jejich rozšíření a v některých případech i náhrady. Např. obchodní jméno a adresa klienta budou převzaty z externích dat, která mají svůj původ v Registru ekonomických subjektů, a lze tedy u nich předpokládat větší správnost. Po spojení interních a externích dat budou záznamy o klientech deduplikovány. Ačkoliv jednotlivé algoritmy nezabezpečí stoprocentní vyčištění klientských dat, jejich snahou je zvýšit jejich vypovídající schopnost.

Ověření kvality klientských dat a jejich čištění by měly řešit ETL procesy, které zabezpečují načítání dat ze zdrojových systémů do datového skladu. U dat, která jsou vstupem pro plochou tabulku v této disertační práci, tomu zatím tak není. Proto byly algoritmy transformace klientských dat přesunuty do vrstvy tvorby této tabulky.

4.1.1.1 Ověření validity IČ

Validita IČ v interních i externích datech bude ověřena podle následujícího algoritmu:

1. Vypočítat součet $n*C_1 + (n-1)*C_2 + (n-2)*C_3 + \ldots\ldots + 2*C_{n-1}$, kde n je počet číslic daného IČ a C_i je hodnota i-té číslice, přičemž i nabývá hodnot od 1 do n – 1.

2. Vypočítat zbytek celočíselného dělení součtu z bodu 1 modulo 11.
3. Vypočítat kontrolní (poslední) číslici IČ následovně:
    a. Je-li rozdíl 11 – výsledek bodu 2 < 10, potom kontrolní číslice = tento rozdíl.
    b. Je-li rozdíl 11 – výsledek bodu 2 = 10, potom kontrolní číslice = 0.
    c. Je-li rozdíl 11 – výsledek bodu 2 = 11, potom kontrolní číslice = 1.
4. IČ vyhovuje ověření, je-li jeho skutečná poslední číslice rovna vypočtené (viz. bod 3).

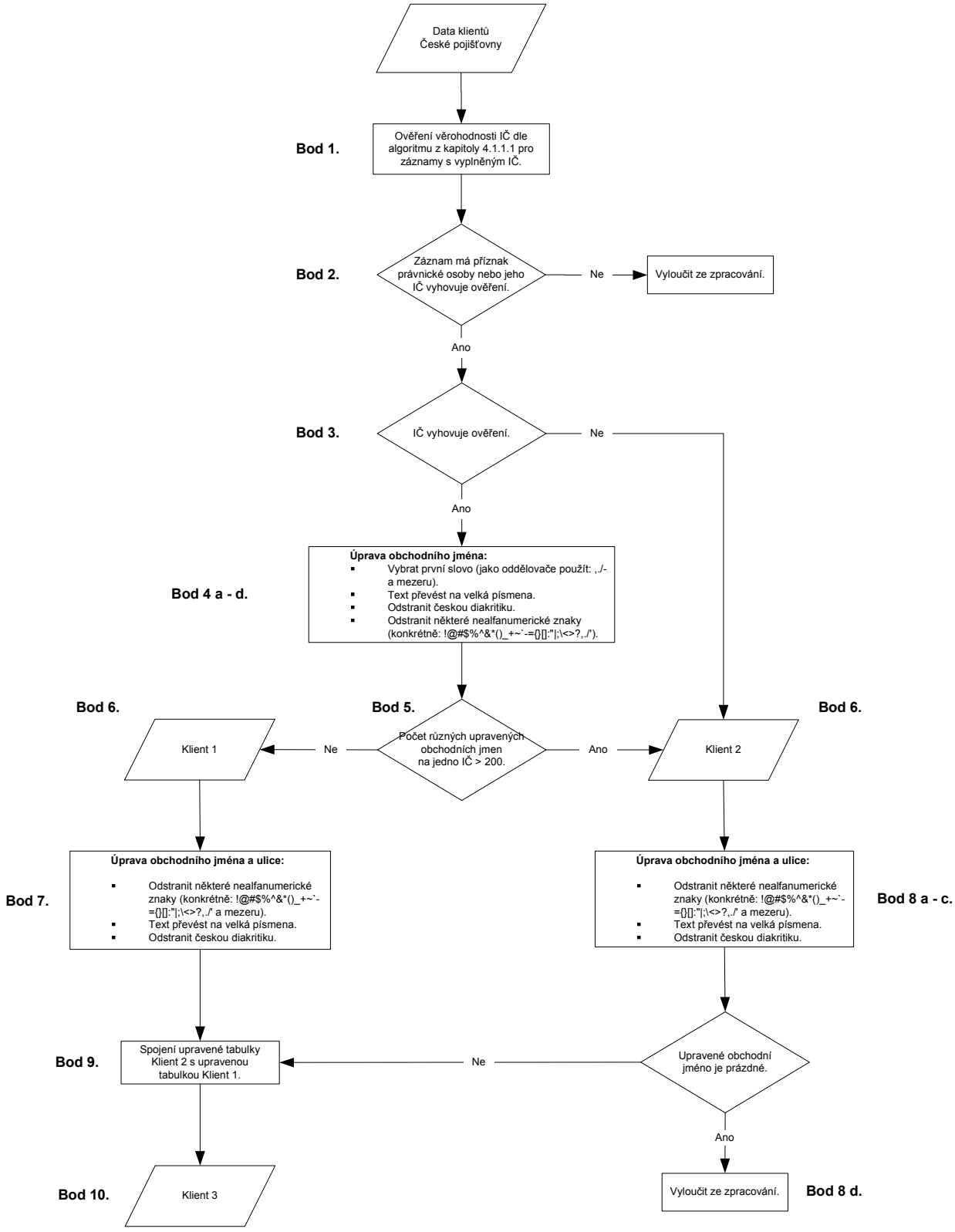
4.1.1.2 Identifikace klientských záznamů validním IČ

Výběr záznamů o firemních klientech z databáze datového skladu a vytvoření jejich identifikace pomocí validního IČ bude provedeno podle algoritmu, který je uveden níže. Tento algoritmus je graficky znázorněn také na obrázku 14 formou vývojového diagramu.

1. U všech klientských záznamů (bez ohledu na roli na pojistné smlouvě a její aktuální platnost či neplatnost), které mají vyplněné IČ, provést ověření jeho věrohodnosti podle algoritmu z kapitoly 4.1.1.1.
2. Ze vstupní databáze datového skladu vybrat všechny klientské záznamy, které se vztahují k ekonomickým subjektům (opět bez ohledu na roli na pojistné smlouvě a její platnost či neplatnost). Konkrétně jsou vybrány všechny věty, na kterých je uveden příznak, že se jedná o právnickou osobu, nebo věty, jejichž IČ vyhovuje ověření. Příznak právnické osoby je předáván do datového skladu z některých provozních systémů.
3. Záznamy vybrané v bodu 2 rozdělit do dvou skupin – na záznamy, jejichž IČ vyhovuje ověření, a na záznamy, jejichž IČ ověření nevyhovuje.
4. U záznamů, jejichž IČ vyhovuje ověření, provést následující úpravu obchodního jména firmy:
    a. Vybrat první slovo (jako oddělovače použít znaky: ,./- a mezeru).
    b. Text převést na velká písmena.
    c. Odstranit českou diakritiku.

   d. Odstranit některé nealfanumerické znaky (konkrétně: !@#$%^&*()_+~`-={}[]:"|;\<>?,./‘).

5. Upravené záznamy z výstupu bodu 4 opět rozdělit do dvou skupin – na záznamy, kde počet různých upravených obchodních jmen na jedno IČ je menší nebo roven 200, a záznamy, kde počet různých upravených obchodních jmen na jedno IČ je větší než 200. U první skupiny se předpokládá, že se jedná o záznamy, které se vztahují k jedné firmě. Obchodní jméno totiž většinou obsahuje různé tvary jednoho názvu firmy nebo jména jejích zástupců. Hranice 200 záznamů je výrazně překročena u druhé skupiny. Zde se v obchodním jméně často vyskytují názvy různých firem. Proto budou IČ z této skupiny dále považována za nevalidní, aby po deduplikaci klientských dat nedošlo k výraznému zkreslení výsledků statistické analýzy.
6. Skupina záznamů, pro které je počet různých upravených obchodních jmen na jedno IČ menší nebo roven 200, tvoří tabulku, která má ve vývojovém diagramu na obrázku 14 název Klient 1. Skupina záznamů, pro které je počet různých upravených obchodních jmen na jedno IČ větší než 200, je dále sloučena se skupinou záznamů s nevalidním IČ (viz. bod 3 výše). Tato druhá skupina tvoří tabulku, která má ve vývojovém diagramu název Klient 2.
7. U záznamů z tabulky Klient 1 provést následující úpravy obchodního jména a ulice:
   a. Odstranit některé nealfanumerické znaky (konkrétně: !@#$%^&*()_+~`-={}[]:"|;\<>?,./‘ a mezeru).
   b. Text převést na velká písmena.
   c. Odstranit českou diakritiku.
8. U záznamů z tabulky Klient 2 provést následující úpravy:
   a. Z obchodního jména a ulice odstranit některé nealfanumerické znaky (konkrétně: !@#$%^&*()_+~`-={}[]:"|;\<>?,./‘ a mezeru).
   b. Text obchodního jména a ulice převést na velká písmena.
   c. Z obchodního jména a ulice odstranit českou diakritiku.
   d. Ze zpracování vyloučit všechny věty, kde je upravené obchodní jméno prázdné.



Obr. 14 Algoritmus vytvoření identifikace klientských záznamů pomocí validních IČ

9. Výstup bodu 8 (upravenou tabulku Klient 2) spojit s výstupem bodu 7 (upravenou tabulkou Klient 1) přes upravené obchodní jméno, upravenou ulici, číslo orientační/popisné a PSČ s tím, že pokud se pro záznam z tabulky Klient 2 nalezne více vět s validním IČ v tabulce Klient 1, převezme věta z tabulky Klient 2 IČ z první věty tabulky Klient 1.
10. Výstup bodu 9 sloučit pod sebe s tabulkou Klient 1. Výsledkem jsou záznamy o firemních klientech České pojišťovny identifikované neunikátním ale validním IČ. Tato tabulka má ve vývojovém diagramu označení Klient 3.

4.1.1.3 Rozšíření interních klientských dat o externí data

Interní data o klientech, která jsou identifikována validním IČ, budou dále spojena s externími daty za účelem jejich rozšíření a náhrady hodnot některých polí. Spojení s externími daty bude provedeno podle algoritmu, který je popsán níže a který je znázorněn formou vývojového diagramu také na obrázku 15.

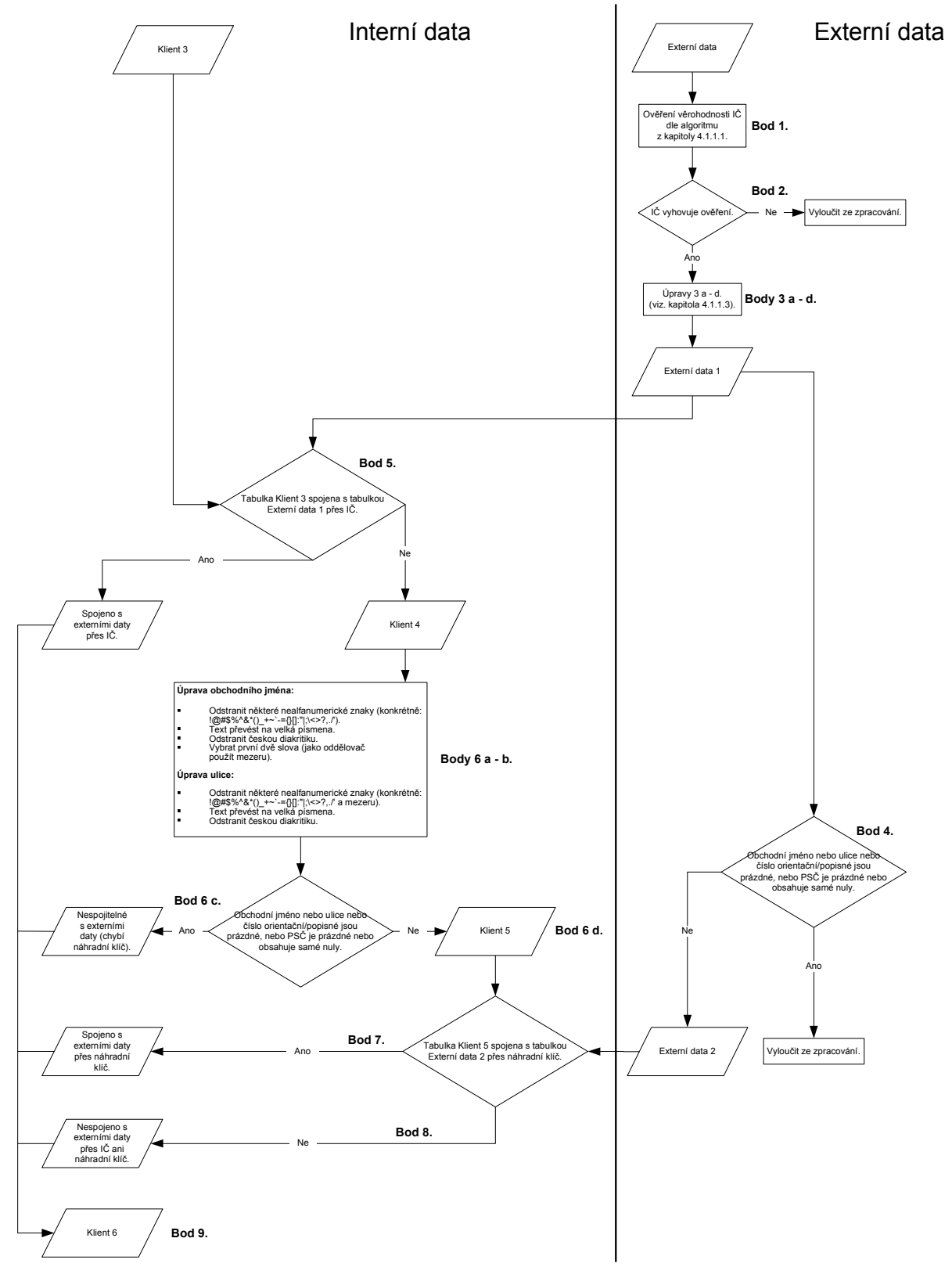
1. U všech záznamů firem v externích datech ověřit věrohodnost IČ podle algoritmu z kapitoly 4.1.1.1.
2. Z externích dat vybrat pouze ty záznamy, jejichž IČ vyhovuje ověření. Ačkoliv externí data obsahují informace z Registru ekonomických subjektů, byla v nich identifikována nevalidní IČ.
3. Ve výstupu bodu 2 provést následující úpravy:
    a. Upravit obchodní jméno:
        i. Odstranit některé nealfanumerické znaky (konkrétně: !@#$%^&*()_+~`-={}[]:"|;\<>?,./‘).
        ii. Text převést na velká písmena.
        iii. Odstranit českou diakritiku.
        iv. Vybrat první dvě slova (jako oddělovač použít mezeru).
    b. Vyčlenit z názvu ulice číslo domu tak, že pokud poslední slovo začíná číslem (oddělovač mezera), je prohlášeno za číslo domu a zbylá slova za název ulice.

c. Upravit název ulice odvozený v bodu 3b.:
    i. Odstranit některé nealfanumerické znaky (konkrétně: !@#$%^&*()_+~`-={}[]:"|;\<>?,./‘ a mezeru).
    ii. Text převést na velká písmena.
    iii. Odstranit českou diakritiku.
    iv. Odstranit samé nuly.
d. Upravit číslo domu odvozené v bodu 3b. tak, aby bylo získáno číslo orientační/popisné:
    i. Řetězec znaků převést na velká písmena.
    ii. Odstranit samé nuly.
    iii. Odstranit lomítka na začátku a na konci řetězce.
    iv. Zaměnit čísla před a za lomítkem, jestliže lomítko je v čísle jediné a jestliže číslo za lomítkem začíná cifrou a obsahuje i písmena a zároveň číslo před lomítkem je tvořeno pouze ciframi. Záměnu provést i v případě, jsou-li obě čísla tvořena pouze ciframi, ale první z nich (před lomítkem) je větší než druhé (za lomítkem).

Externí data upravená v bodech 3 a – d. tvoří tabulku, která má ve vývojovém diagramu na obrázku 15 název Externí data 1.

4. Z výstupu bodu 3 vyloučit ty záznamy, které mají prázdné obchodní jméno nebo ulici nebo číslo orientační/popisné, nebo jejichž PSČ je prázdné nebo obsahuje samé nuly. Výsledkem bude tabulka, která má ve vývojovém diagramu název Externí data 2.
5. Tabulku Klient 3 (viz. výstup bodu 10 kapitoly 4.1.1.2) spojit s tabulkou Externí data 1 přes IČ. Věty, které se přes tento klíč spojily, tvoří tabulku, která má ve vývojovém diagramu název Spojeno s externími daty přes IČ. Obchodní jména a adresy sídel firem jsou ve výsledku převzaty z tabulky Externí data 1. Věty z tabulky Klient 3, které se nespojily s tabulkou Externí data 1, tvoří samostatnou tabulku s názvem Klient 4.

6. V tabulce Klient 4 provést následující úpravy:
    a. Upravit obchodní jméno:
        i. Odstranit některé nealfanumerické znaky (konkrétně: !@#$%^&*()_+~`-={}[]:"|;\<>?,./‘).
        ii. Text převést na velká písmena.
        iii. Odstranit českou diakritiku.
        iv. Vybrat první dvě slova (jako oddělovač použít mezeru).
    b. Upravit ulici:
        i. Odstranit některé nealfanumerické znaky (konkrétně: !@#$%^&*()_+~`-={}[]:"|;\<>?,./‘ a mezeru).
        ii. Text převést na velká písmena.
        iii. Odstranit českou diakritiku.
    c. Z tabulky vyloučit věty, které mají prázdné obchodní jméno nebo ulici nebo číslo orientační/popisné, nebo jejichž PSČ je prázdné nebo obsahuje samé nuly. Tyto vyloučené věty tvoří samostatnou tabulku, která má ve vývojovém diagramu název Nespojitelné s externími daty (chybí náhradní klíč). Obchodní jméno a adresa sídla firmy tvoří náhradní klíč pro spojení interních a externích dat. Při nevyplnění těchto informací není možné data spojit.
    d. Zbylé věty z výstupu bodu 6 c. tvoří tabulku Klient 5.
7. Tabulku Klient 5 spojit s tabulkou Externí data 2 přes upravené obchodní jméno, upravený název ulice, číslo orientační/popisné a PSČ, přičemž pokud bude pro danou větu z tabulky Klient 5 nalezeno více vět v tabulce Externí data 2, potom se vybere první z těchto vět. Záznamy z tabulky Klient 5, které se podařilo spojit s tabulkou Externí data 2 tvoří tabulku, která má ve vývojovém diagramu název Spojeno s externími daty přes náhradní klíč. IČ, obchodní jména a adresy sídel firem jsou v této tabulce převzaty z externích dat.



Obr. 15 Algoritmus rozšíření interních klientských dat o externí data

8. Zbylé záznamy z tabulky Klient 5, které se nepodařilo spojit s tabulkou Externí data 2, tvoří tabulku, která má ve vývojovém diagramu název Nespojeno s externími daty přes IČ ani náhradní klíč.
9. Tabulky Spojeno s externími daty přes IČ, Spojeno s externími daty přes náhradní klíč, Nespojeno s externími daty přes IČ ani náhradní klíč a Nespojitelné s externími daty (chybí náhradní klíč) sloučit pod sebe. Vznikne tabulka Klient 6, která obsahuje všechny záznamy o firmách vybrané z interních dat a rozšířené o externí data.

4.1.1.4 Deduplikace klientských dat

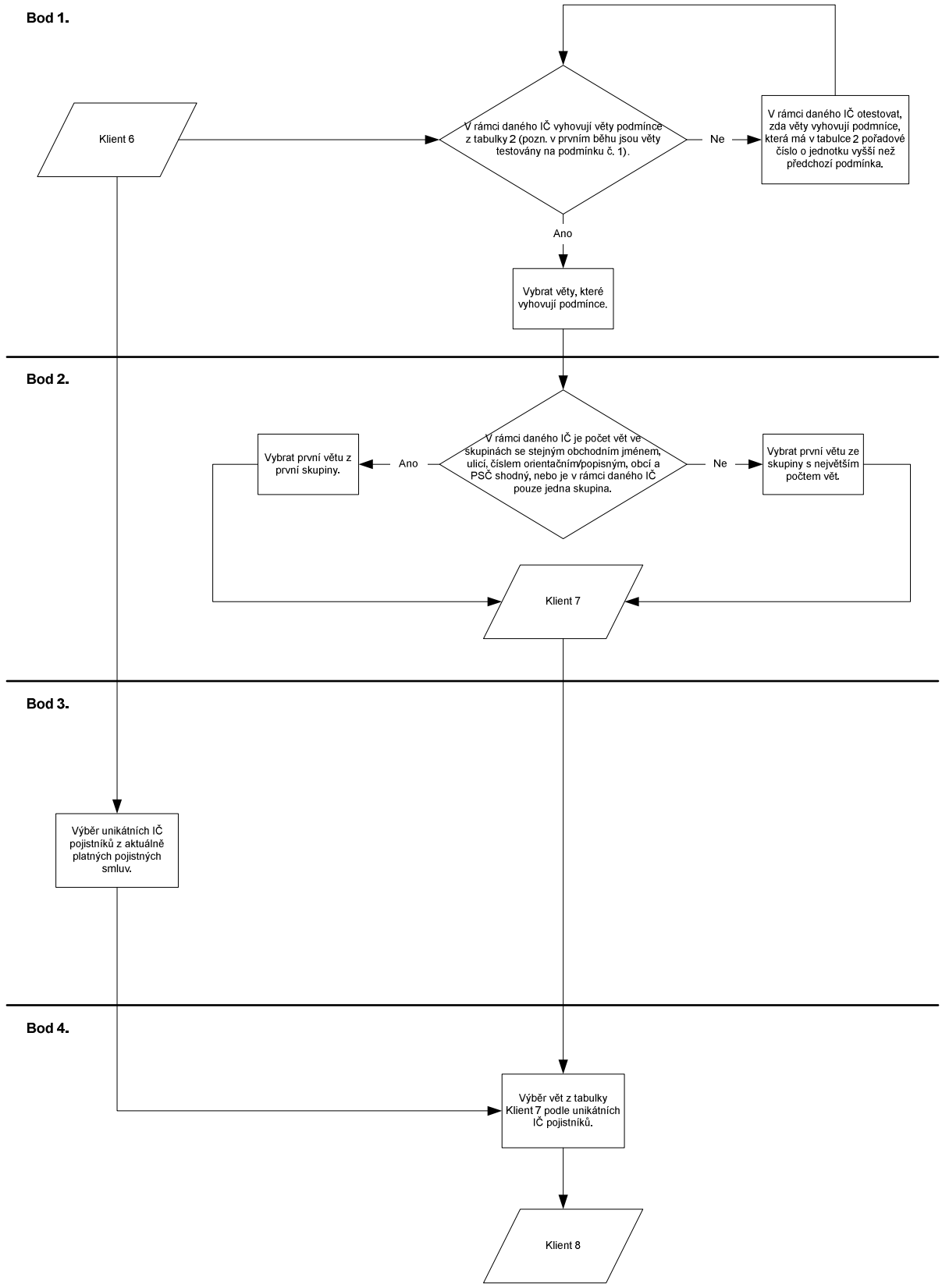
Dalším krokem v přípravě klientských dat bude jejich deduplikace a výběr pouze těch záznamů, které se vztahují k roli pojistníka na aktuálně platných pojistných smlouvách. Deduplikace tabulky Klient 6 (viz. výstup bodu 9 kapitoly 4.1.1.3) bude provedena podle algoritmu, který je uveden níže a který je znázorněn formou vývojového diagramu na obrázku 16.

1. Z tabulky Klient 6 vybrat pro dané IČ věty podle naplněnosti položek Obchodní jméno, Ulice, Číslo orientační/popisné, Obec a PSČ. Tento výběr řídit podle podmínek z tabulky 2. Např. jestliže v tabulce Klient 6 není nalezena pro dané IČ žádná věta s vyplněnými atributy viz. výše (tj. není splněna podmínka č. 1), potom pro toto IČ vybrat všechny věty, které splňují podmínku č. 2. Tj. vybrat věty, které nemusí mít naplněné pole Obec, ale musí mít naplněný zbytek z povinných položek. Pokud v rámci daného IČ nesplňuje tuto podmínku opět žádná věta, potom věty testovat na podmínku č. 3 atd., dokud pro toto IČ není vybrán alespoň jeden záznam.

Tab. 2 Řídící podmínky pro výběr záznamů z tabulky Klient 6 dle naplněnosti položek

Podmínka číslo	Obchodní jméno	Ulice	Číslo orientační/popisné	Obec	PSČ
1	1	1	1	1	1
2	1	1	1	0	1
3	1	1	0	0	1
4	1	0	0	0	1
5	1	0	0	0	0
6	0	0	0	0	0

Legenda: 1 – položka musí být naplněna, 0 – položka nemusí být naplněna (PSČ nemusí být naplněné nebo může obsahovat samé nuly).



Obr. 16 Algoritmus deduplikace klientských dat a výběr pojistníků

2. Z výstupu bodu 1 vybrat první záznam ze skupiny, která obsahuje v rámci daného IČ největší počet vět se stejným obchodním jménem, ulicí, číslem orientačním/popisným, obcí a PSČ. Jestliže v rámci daného IČ bude těchto skupin několik a počet záznamů v nich bude stejný, potom vybrat první větu z první skupiny. Stejně postupovat i v případě, je-li v rámci daného IČ pouze jedna skupina. Vybrané věty budou tvořit tabulku Klient 7, ve které bude IČ jedinečným identifikátorem záznamu.
3. Z tabulky Klient 6 vybrat všechny záznamy, které se vztahují k roli pojistníka na aktuálně platných pojistných smlouvách, a z výsledku vybrat všechna unikátní IČ.
4. Podle IČ z výstupu bodu 3 vybrat záznamy z tabulky Klient 7. Výsledkem bude tabulka Klient 8, která bude obsahovat pojistníky z aktuálně platných pojistných smluv, jež budou identifikováni validním a unikátním IČ.

4.1.1.5 Výběr vzorku klientů a modifikace obsahu dat

Předposledním krokem při transformaci klientských dat do ploché tabulky bude výběr jejich vzorku. Z původních 1 580 076 interních záznamů o ekonomických subjektech, které byly vybrány ze vstupní databáze datového skladu (viz. bod 2 kapitoly 4.1.1.2), vzniklo v tabulce Klient 8 celkem 210 741 záznamů. Z této tabulky bude dále vybrán vzorek 100 000 firem, který bude pro účely statistické analýzy považován za kompletní populaci firemních klientů. Výběr a zpracování pouze vzorku klientů je podmínkou České pojišťovny pro použití jejích dat v této disertační práci. Algoritmus výběru vzorku nesmí být v práci zveřejněn. Vzorek tvoří základ ploché tabulky, ve které budou dále odvozeny nové proměnné (např. budou doplněny některé informace z pojistných smluv klientů). Počet záznamů tabulky se již ale měnit nebude.

Výběr vzorku firem pro plochou tabulku není realizací etapy Sample dle metodologie SEMMA. Výsledek vzorkování je totiž dále považován za základní soubor firemních klientů. Statistická analýza bude provedena nad celým tímto souborem, přičemž jejím cílem je vytvoření systému tříd (segmentů), do kterých budou rozděleny všechny záznamy. Na základě obchodní profilace segmentů bude možné stanovit pravidla, podle kterých budou aktuální klienti firmy do segmentů zařazováni.

Posledním krokem při transformaci klientských dat bude modifikace jejich obsahu. Tato změna bude provedena jak v základním souboru 100 000 klientů, tak i v ostatních datech, ze kterých budou odvozeny nové proměnné pro plochou tabulku. Modifikace obsahu dat je další podmínkou České pojišťovny pro jejich použití v této disertační práci. Algoritmus, podle kterého bude provedena, nesmí být v práci opět zveřejněn. Se změněnými daty bude dále zacházeno, jakoby odrážela reálné obchodní procesy České pojišťovny.

4.1.2 Odvození nových proměnných

V základním souboru, který obsahuje data klientů po modifikaci (viz. výstup kapitoly 4.1.1.5), budou dále odvozeny některé spojité a kategoriální proměnné. V některých případech budou ke klientům doplněny nové informace odvozené z dat mimo základní soubor (např. z pojistných smluv). Zároveň však budou odvozeny i nové proměnné z dat základního souboru. Výsledkem bude plochá tabulka s názvem Segmentace 1, jejíž struktura je popsána v tabulce 5.

V prvním případě (viz. výše) budou v ploché tabulce drženy pouze nové proměnné. V původních datech jsou totiž informace evidovány v jiném detailu než na dimenzi IČ (např. jeden klient může mít v České pojišťovně několik pojistných smluv apod.). V případě, že proměnné budou odvozeny z dat základního souboru, budou v ploché tabulce drženy jak nové, tak i původní proměnné. Obsahují totiž informace, které jsou v detailu na dimenzi IČ. Zároveň může v průběhu statistické analýzy vzniknout potřeba odvození dalších proměnných. Z tohoto důvodu doporučuje Parr Rud [29] zachovávat v souborech původní proměnné, pokud je to možné.

4.1.2.1 Odvození spojitých proměnných

V tabulce Segmentace 1 budou odvozeny následující spojité proměnné:

1. *Suma očekávaného ročního předpisového pojistného v Kč* ze všech aktuálně platných pojistných smluv klienta (viz. položka R_predpis v tabulce 5). Očekávaným ročním předpisovým pojistným se rozumí pojistné, které se očekává od daného klienta k zaplacení za daný rok.
2. *Index storen pro neplacení* na daného klienta (viz. položka Pomer v tabulce 5). Tato proměnná bude vypočtena podle následujícího vzorce:

$$Index\ storen\ pro\ neplacení = \frac{Pojistné\ ze\ storen\ pro\ neplacení}{Celkové\ pojistné}, \quad (4.1)$$

kde *Pojistné ze storen pro neplacení* je rovno sumě očekávaného ročního předpisového pojistného (v Kč) ze všech smluv klienta, jež byly stornovány pro neplacení v období aktuální datum minus dva roky. V případě, že klient nemá žádnou takovou smlouvu, je Pojistné ze storen pro neplacení rovno nule. *Celkové pojistné* je rovno sumě očekávaného ročního předpisového pojistného (v Kč) ze všech smluv klienta, které v období aktuální datum minus dva roky platily alespoň jeden den.

3. *Stáří firmy v letech* (viz. položka Stari v tabulce 5). Tato proměnná bude vypočtena jako rozdíl aktuálního roku a roku vzniku firmy odvozeného z data vzniku. Datum vzniku firmy má svůj původ v externích datech a je součástí základního souboru klientů i ploché tabulky (viz. položka Dat_vzn v tabulce 5). Při výpočtu stáří firmy není zohledněn datum ukončení podnikatelské činnosti v případě, že firma zanikla (viz. položka Dat_ukon v tabulce 5). Z pohledu České pojišťovny se totiž jeví jako aktivní i ty firmy, které mají tento datum vyplněn a který je menší nebo roven aktuálnímu datu. Všechny firmy zařazené do základního souboru klientů mají aktuálně platnou alespoň jednu pojistnou smlouvu. Vůči České pojišťovně tedy všechny vystupují svým jménem jako aktivní klienti.
4. *Základní jmění společnosti v Kč*, *Suma očekávaného ročního předpisového pojistného v Kč* a *Stáří firmy v letech* po substituci hodnot menších než 1. percentil a hodnot větších než 99. percentil dané proměnné (viz. položky Jmeni_subst, R_predpis_subst a Stari_subst v tabulce 5). Substituce bude provedena tak, že každá hodnota proměnné menší než 1. percentil, bude nahrazena hodnotou 1. percentilu a každá hodnota větší než 99. percentil hodnotou 99. percentilu. Cílem substituce je zmírnit vliv odlehlých pozorování na výsledky statistické analýzy a jejich interpretaci. Ze základních popisných statistik vypočtených před a po provedení substituce (viz. tabulka 3) je možné zjistit její efekt na stabilizaci vstupních proměnných. Např. u všech proměnných došlo ke snížení hodnot aritmetického průměru (Avg) a směrodatné odchylky (Std). Výpočet popisných statistik byl proveden v systému SAS

pomocí procedury UNIVARIATE, která za tímto účelem použila následující vzorce [34]:

Aritmetický průměr

$$\bar{x} = \frac{\sum x_i}{n}, \qquad (4.2)$$

kde n je počet neprázdných hodnot dané proměnné a x_i je její i-tá neprázdná hodnota, přičemž i je z intervalu 1 až n.

Směrodatná odchylka

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}, \qquad (4.3)$$

kde n je počet neprázdných hodnot dané proměnné, x_i je její i-tá neprázdná hodnota, přičemž i je z intervalu 1 až n a $\bar{x}$ je její aritmetický průměr.

Šikmost

$$\frac{1}{n}\sum\left(\frac{x_i - \bar{x}}{s}\right)^3, \qquad (4.4)$$

kde n je počet neprázdných hodnot dané proměnné, x_i je její i-tá neprázdná hodnota, přičemž i je z intervalu 1 až n, $\bar{x}$ je její aritmetický průměr a s směrodatná odchylka.

Špičatost

$$\frac{1}{n}\sum\left(\frac{x_i - \bar{x}}{s}\right)^4 - 3, \qquad (4.5)$$

kde n je počet neprázdných hodnot dané proměnné, x_i je její i-tá neprázdná hodnota, přičemž i je z intervalu 1 až n, $\bar{x}$ je její aritmetický průměr a s směrodatná odchylka.

Kvantily

$$y = \frac{1}{2}(x_i + x_{i+1}), \text{ je-li q = 0,}$$
$$y = x_{i+1}, \text{ je-li q > 0,} \qquad (4.6)$$

kde x_i je i-tá hodnota dané proměnné, přičemž i je celočíselná část součinu np. V něm n představuje počet neprázdných hodnot proměnné a p hodnotu požadovaného kvantilu dělenou 100. Desetinnou část součinu představuje znak q.

Počet neprázdných hodnot pro každou proměnnou je uveden v tabulce 3 v položce N. Nad neprázdnými hodnotami byla provedena jak substituce, tak i výpočet popisných statistik. Substituce vedla ke snížení šikmosti a špičatosti nových proměnných – viz. položky Skew a Kurt v tabulce 3. Přesto jejich hodnoty zůstávají daleko od nuly, která je charakteristická pro normální rozdělení. Minimální a maximální hodnoty nových proměnných jsou rovny hodnotám 1. a 99. percentilu původních proměnných.

Tab. 3 Základní popisné statistiky proměnných Základní jmění společnosti v Kč, Suma očekávaného ročního předpisového pojistného v Kč a Stáří firmy v letech – stav před a po provedení substituce hodnot

Proměnná	N	Avg	Std	Skew	Kurt	Min	P1	P5
	Počet neprázdných hodnot	Aritmetický průměr	Směrodatná odchylka	Šikmost	Špičatost	Minimum	1. percentil	5. percentil
Jmeni	87 116	5 904 660,98	210 620 607,85	100,60	12 456,04	0	0	0
Jmeni_subst	87 116	1 170 890,35	7 316 202,29	7,55	57,88	0	0	0
R_predpis	100 000	45 379,56	1 555 174,79	269,40	78 435,34	23,00	216,00	802,00
R_predpis_subst	100 000	30 324,03	63 483,30	4,71	25,01	216,00	216,00	802,00
Stari	97 956	11,77	4,68	0,74	6,76	1,00	2,00	3,00
Stari_subst	97 956	11,74	4,47	0,10	1,90	2,00	2,00	3,00

Proměnná	P10	Q1	Median	Q3	P90	P95	P99	Max
	10.percentil	Spodní kvartil	Medián	Horní kvartil	90. percentil	95. percentil	99. percentil	Maximum
Jmeni	0	0	0	100 000,00	200 000,00	1 050 000,00	63 150 000,00	32 208 990 000,00
Jmeni_subst	0	0	0	100 000,00	200 000,00	1 050 000,00	63 150 000,00	63 150 000,00
R_predpis	1 840,00	5 130,00	11 116,50	26 248,00	62 891,00	121 504,00	455 108,00	462 011 646,00
R_predpis_subst	1 840,00	5 130,00	11 116,50	26 248,00	62 891,00	121 504,00	455 108,00	455 108,00
Stari	5,00	9,00	13,00	15,00	16,00	16,00	30,00	61,00
Stari_subst	5,00	9,00	13,00	15,00	16,00	16,00	30,00	30,00

Legenda: Měřeno nad kombinací dat České pojišťovny a Registru ekonomických subjektů k 30.6.2006. Modře jsou označeny proměnné po substituci.

Odlehlá pozorování mohou být datovými chybami. Zároveň však mohou vyjadřovat neobvyklé, ale správné chování zkoumaných jednotek (v našem případě firem). Nejlepším způsobem jejich identifikace a rozlišení, zda se jedná o datové chyby nebo o neobvyklé chování, je jejich individuální posouzení. To však může být při zpracování velkého objemu dat výpočetně i časově náročné. Proto Parr Rud [30] doporučuje v těchto případech použít metodu substituce, která je obecná a kterou lze snadno a rychle implementovat.

Substituce hodnot nebude provedena u spojité proměnné *Index storen pro neplacení* (viz. položka Pomer v tabulce 5). Její hodnoty se totiž pohybují pouze

v rozmezí od 0 do 1. Základní popisné statistiky této proměnné jsou uvedeny v tabulce 4.

Tab. 4 Základní popisné statistiky proměnné Index storen pro neplacení (Pomer)

Proměnná	N Počet neprázdných hodnot	Avg Aritmetický průměr	Std Směrodatná odchylka	Skew Šikmost	Kurt Špičatost	Min Minimum	P1 1. percentil	P5 5. percentil
Pomer	100 000	0,1166	0,2671	2,1392	2,8901	0	0	0

Proměnná	P10 10. percentil	Q1 Spodní kvartil	Median Medián	Q3 Horní kvartil	P90 90. percentil	P95 95. percentil	P99 99. percentil	Max Maximum
Pomer	0	0	0	0	0,7237	0,8015	0,9213	0,9972

Legenda: Měřeno nad kombinací dat České pojišťovny a Registru ekonomických subjektů k 30.6.2006.

5. *Základní jmění společnosti*, *Suma očekávaného ročního předpisového pojistného*, *Stáří firmy* a *Index storen pro neplacení* po provedení standardizace (viz. položky Jmeni_std, R_predpis_std, Stari_std a Pomer_std v tabulce 5). Standardizace bude provedena nad proměnnými, které vznikly substitucí popsanou v bodě 4 (viz. výše). Dále bude provedena nad proměnnou Pomer, která v bodě 4 transformována nebyla. Standardizací budou proměnné transformovány tak, že jejich aritmetický průměr bude roven nule a směrodatná odchylka rovna jedné.

   V této disertační práci bude v rámci statistické analýzy provedena shluková analýza a analýza kategoriálních dat. Do shlukové analýzy budou vstupovat pouze spojité proměnné. Cílem jejich standardizace je zabezpečit, aby všechny měly stejný vliv na výsledek shlukování. U proměnných s větším rozptylem je tento vliv větší než u proměnných s menším rozptylem. Standardizace je proto doporučována zejména tehdy, jsou-li jednotlivé proměnné evidovány v různých jednotkách měření [35]. Pro účely této disertační práce bude standardizace provedena v systému SAS pomocí procedury STANDARD, která transformuje proměnné podle následujícího vzorce [34]:

$$x_i^{'} = \frac{S\,(x_i - \bar{x})}{s} + M, \qquad (4.7)$$

   kde $x^{'}_i$ je standardizovaná hodnota proměnné, přičemž i je z intervalu 1 až n, S je požadovaná směrodatná odchylka po standardizaci, x_i je hodnota proměnné před

standardizací (*i* je opět z intervalu 1 až n), $\bar{x}$ je aritmetický průměr a *s* směrodatná odchylka proměnné před standardizací a *M* je požadovaný aritmetický průměr po standardizaci.

Standardizace eliminuje rozdíly v rozptylech proměnných, čímž přispívá k jejich vyrovnanému vlivu na výsledek shlukování. Substituce odlehlých hodnot, která stabilizuje rozptyl každé proměnné ještě před standardizací (viz. bod 4 výše), tento efekt podporuje. Standardizované proměnné budou vstupovat pouze do shlukovacího algoritmu. Vzhledem k tomu, že nemají ekonomickou interpretaci, nebude nad nimi prováděna profilace obchodních segmentů. Za tímto účelem budou použity proměnné po substituci, které ekonomickou interpretaci mají a lze na jejich základě definovat přesnější pravidla pro zařazování firem do segmentů než na základě proměnných bez substituce (viz. rozdíly v popisných statistikách v tabulce 3).

4.1.2.2 Odvození kategoriálních proměnných

V tabulce Segmentace 1 budou dále odvozeny následující kategoriální proměnné:

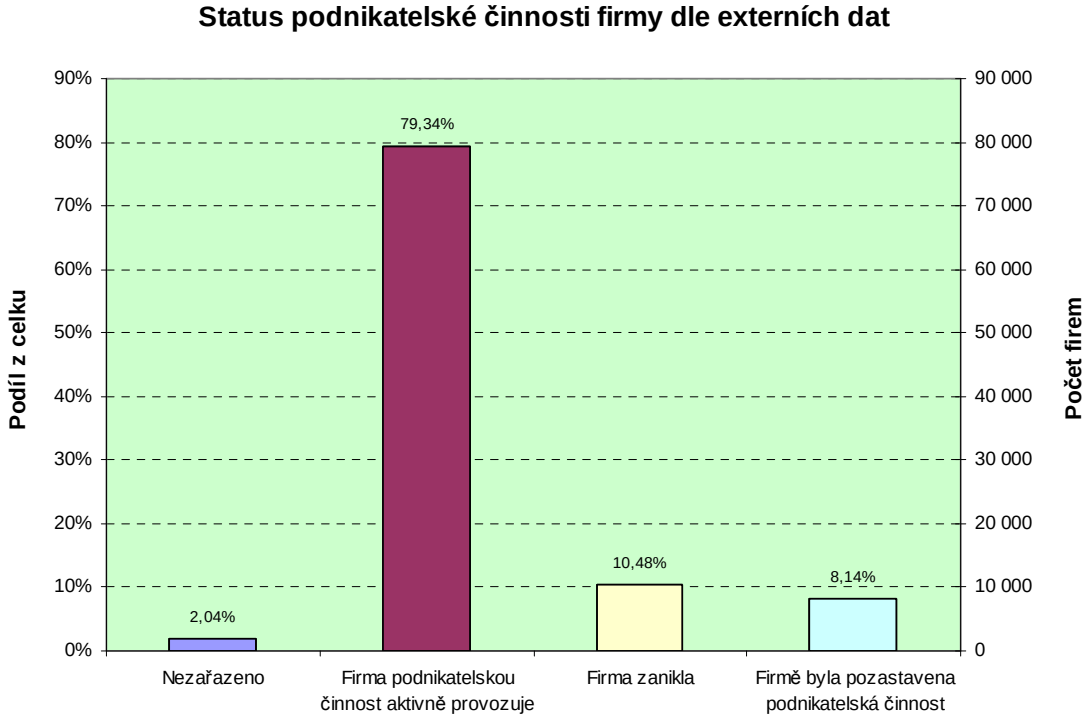
1. *Status podnikatelské činnosti firmy dle externích dat* (viz. položka Aktiv v tabulce 5). Tato proměnná bude odvozena z položek Aktivni a Dat_ukon (viz. tabulka 5), jež mají svůj původ v externích klientských datech. Položka Aktivni obsahuje dvě kategorie: Firmě byla pozastavena podnikatelská činnost (kód 0) a Firma podnikatelskou činnost aktivně provozuje (kód 1). Položka Dat_ukon obsahuje datum ukončení podnikatelské činnosti firmy. Porovnáním obou položek vznikne nová proměnná, která bude obsahovat stejné kategorie jako položka Aktivni a navíc kategorii Firma zanikla (s kódem 2).

   Algoritmus odvození nové proměnné bude následující:

   a. Nabývá-li položka Aktivni hodnot 0 nebo 1 a je-li datum ukončení podnikatelské činnosti (Dat_ukon) menší nebo roven aktuálnímu datu, potom hodnotu nové proměnné nastavit na 2 (Firma zanikla).

b. V ostatních případech jsou hodnoty nové proměnné rovny hodnotám položky Aktivni (tj. 0, v případě, že firmě byla pozastavena podnikatelská činnost, nebo 1, v případě, že podnikatelskou činnost aktivně provozuje).

Na obrázku 17 je znázorněn histogram zastoupení jednotlivých kategorií proměnné Status podnikatelské činnosti firmy dle externích dat (Aktiv) na celkovém počtu firem.

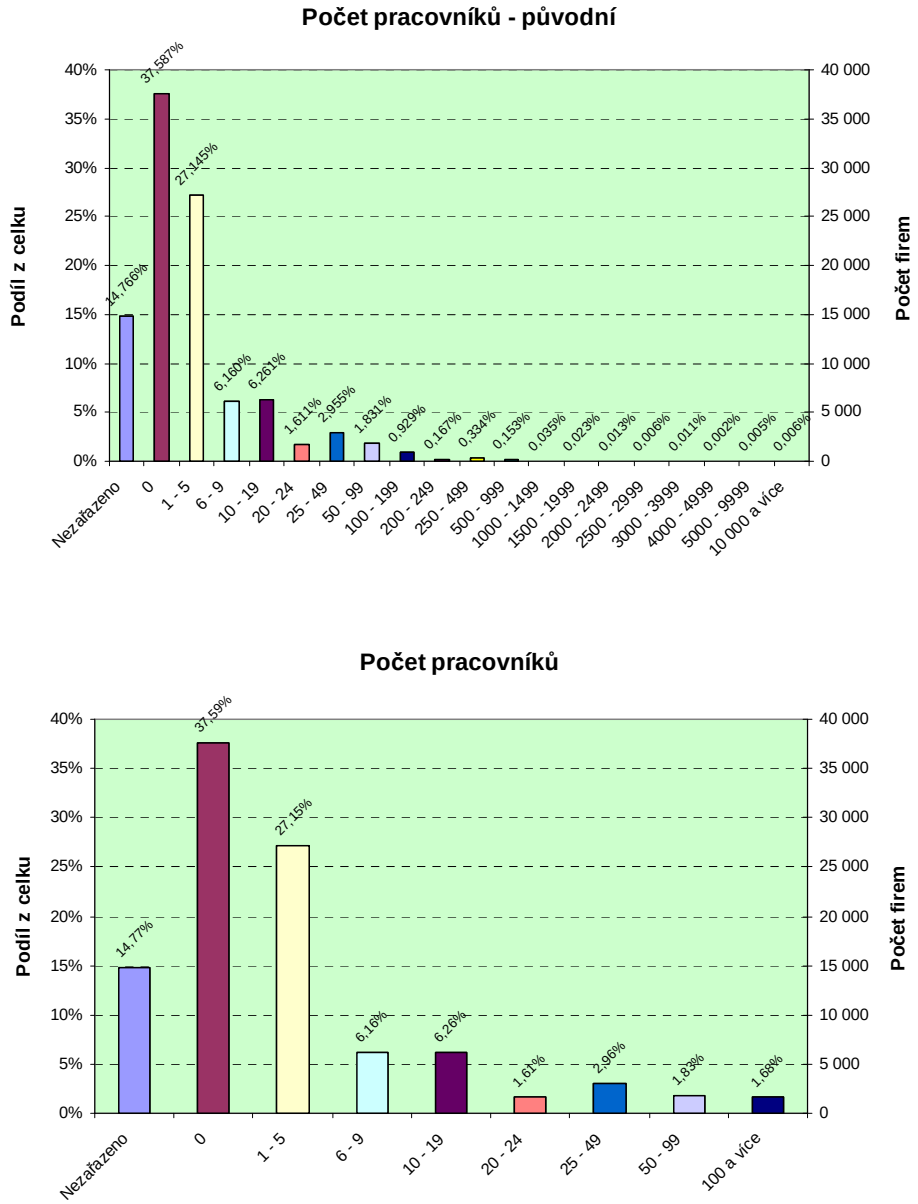


Obr. 17 Histogram proměnné Status podnikatelské činnosti firmy dle externích dat (Aktiv) – stav k 30.6.2006 měřený nad kombinací dat České pojišťovny a Registru ekonomických subjektů

Z pohledu externích dat lze tedy rozlišit mezi třemi stavy podnikatelské činnosti klientů. Z hlediska interních dat je tento stav pouze jeden. Všechny firmy mají u České pojišťovny alespoň jednu aktuálně platnou pojistnou smlouvu, na které vystupují v roli pojistníka (tj. platí pojistné). Vůči pojišťovně se tedy prezentují jako aktivně podnikající klienti, kteří mají sjednané podnikatelské pojištění.

Informace o statusu podnikatelské činnosti dle interních a externích dat budou posuzovány při profilaci obchodních segmentů. Cílem bude stanovení obchodního potenciálu segmentů a strategie dalšího přístupu k nim.

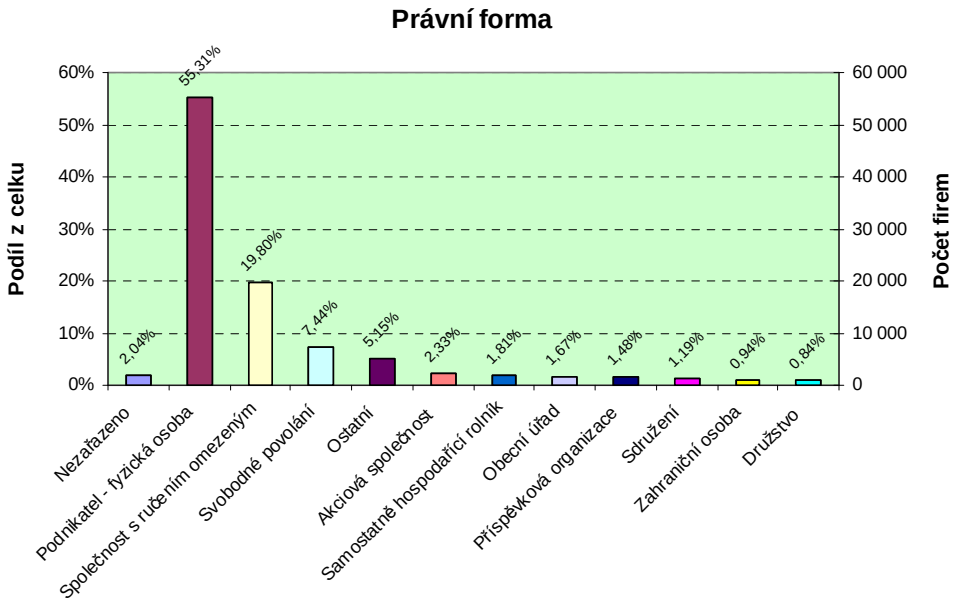
2. *Počet pracovníků* (viz. položka Pocprac v tabulce 5). Tato proměnná bude odvozena z položky Pocprac_ext (viz. tabulka 5), která má svůj původ v externích klientských datech. Původní proměnná obsahuje kategorie, jež představují intervaly s počty zaměstnanců dle metodiky OECD. Pro novou proměnnou budou některé kategorie z důvodu malého počtu zkoumaných jednotek v nich sloučeny. Na obrázku 18 jsou uvedeny histogramy, ze kterých je zřejmé zastoupení jednotlivých kategorií v původní a nové proměnné.



Obr. 18 Histogramy proměnných Počet pracovníků – původní (Pocprac_ext) a Počet pracovníků (Pocprac) – stav k 30.6.2006 měřený nad kombinací dat České pojišťovny a Registru ekonomických subjektů

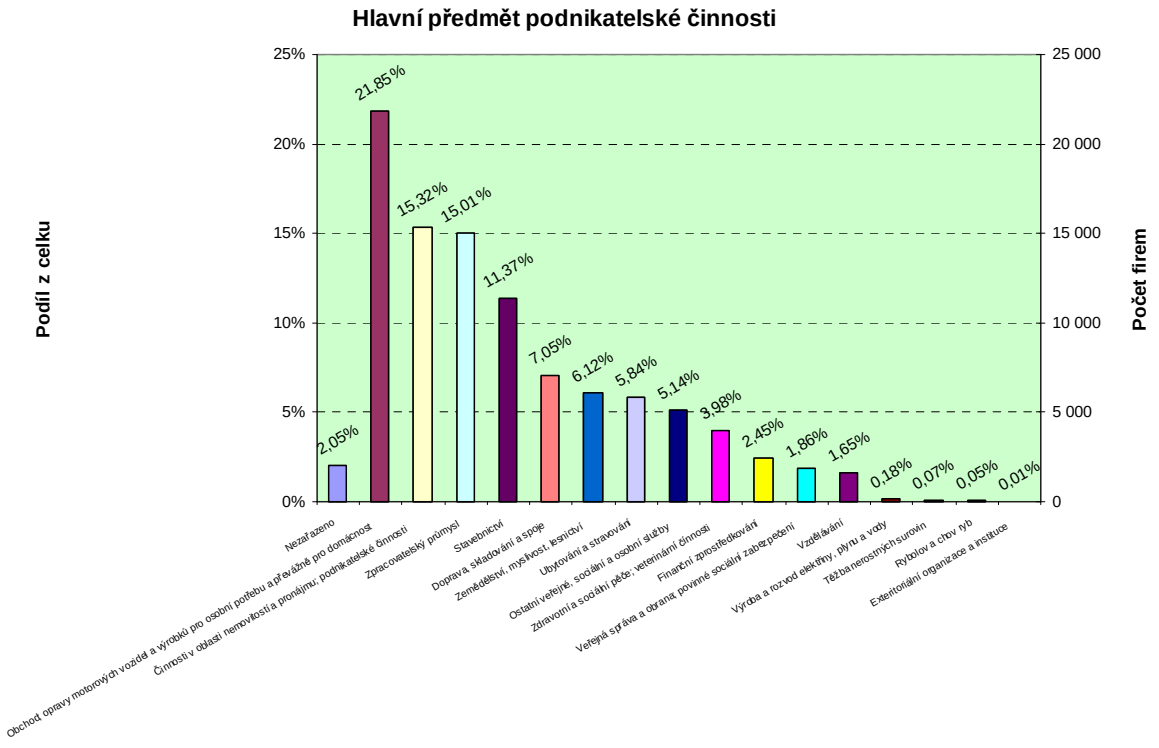
Pro novou proměnnou je vytvořena kategorie 100 a více, do níž jsou zařazeny všechny firmy, které mají 100 a více zaměstnanců. V původní proměnné jsou tyto firmy rozděleny do 12-ti kategorií, ačkoliv představují pouze 1,68 % z celkového počtu. Zbytek souboru (98,32 %) je rozdělen do 8-mi kategorií (pozn.: prázdné hodnoty jsou chápány jako kategorie Nezařazeno). Nová proměnná tedy obsahuje po sloučení firem se 100 a více zaměstnanci do jedné skupiny pouze 9 kategorií.

3. *Právní forma* (viz. položka Pravfor v tabulce 5). Tato proměnná obsahuje informace o právních formách firem. Bude odvozena z položky Pravfor_ext (viz. tabulka 5), která má svůj původ v externích klientských datech. Původní proměnná vychází z číselníku Českého statistického úřadu Právní forma organizace [7] a v ploché tabulce nabývá celkem 42 různých hodnot (pozn. prázdné hodnoty jsou chápány jako kategorie Nezařazeno). 11 z nich, které reprezentují 94,85 % firem, bude ponecháno také v nové proměnné. Zbytek souboru (5,15 %) bude zařazen do nové kategorie Ostatní. Na obrázku 19 je znázorněn histogram zastoupení jednotlivých kategorií proměnné Právní forma (Pravfor) na celkovém počtu firem. Histogram původní proměnné není uveden z důvodu velkého počtu kategorií.



Obr. 19 Histogram proměnné Právní forma (Pravfor) - stav k 30.6.2006 měřený nad kombinací dat České pojišťovny a Registru ekonomických subjektů

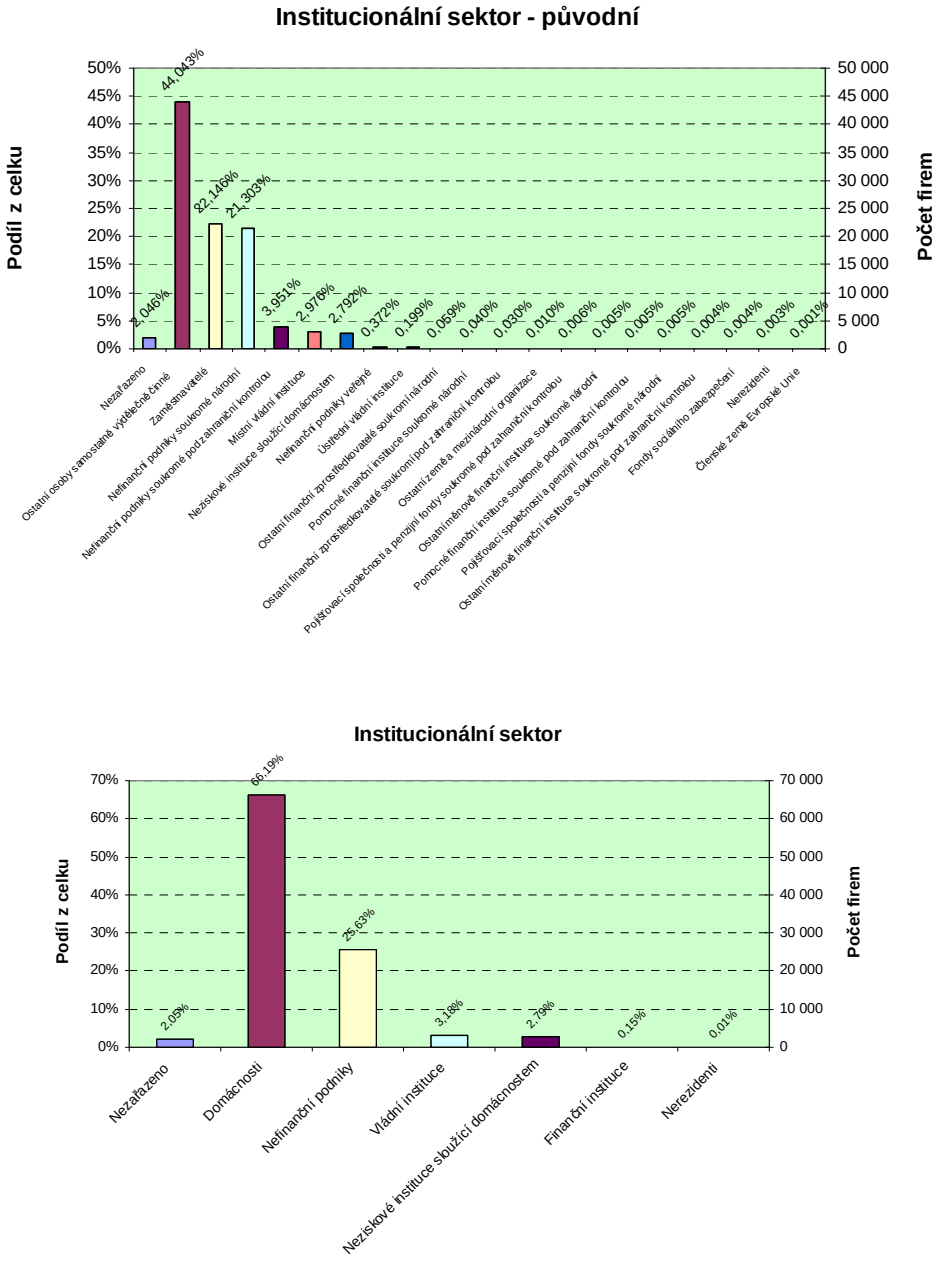
4. *Hlavní předmět podnikatelské činnosti* (viz. položka Okec_sekce v tabulce 5). Tato proměnná obsahuje hlavní předměty podnikatelských činností firem dle klasifikace OKEČ. Bude odvozena z položky Okec (viz. tabulka 5), která má svůj původ v externích klientských datech a jejíž hodnoty lze uspořádat do hierarchie. Původní proměnná obsahuje celkem 732 různých hodnot (pozn. prázdné hodnoty jsou chápány jako kategorie Nezařazeno), které reprezentují různé úrovně třídění. Pro novou proměnnou bude použita nejvyšší úroveň, tzv. sekce, čímž bude redukován počet kategorií na 17. Na obrázku 20 je znázorněn histogram zastoupení jednotlivých kategorií nové proměnné na celkovém počtu firem. Histogram pro původní proměnnou není uveden z důvodu velkého počtu kategorií.



Obr. 20 Histogram proměnné Hlavní předmět podnikatelské činnosti (Okec_sekce) – stav k 30.6.2006 měřený nad kombinací dat České pojišťovny a Registru ekonomických subjektů

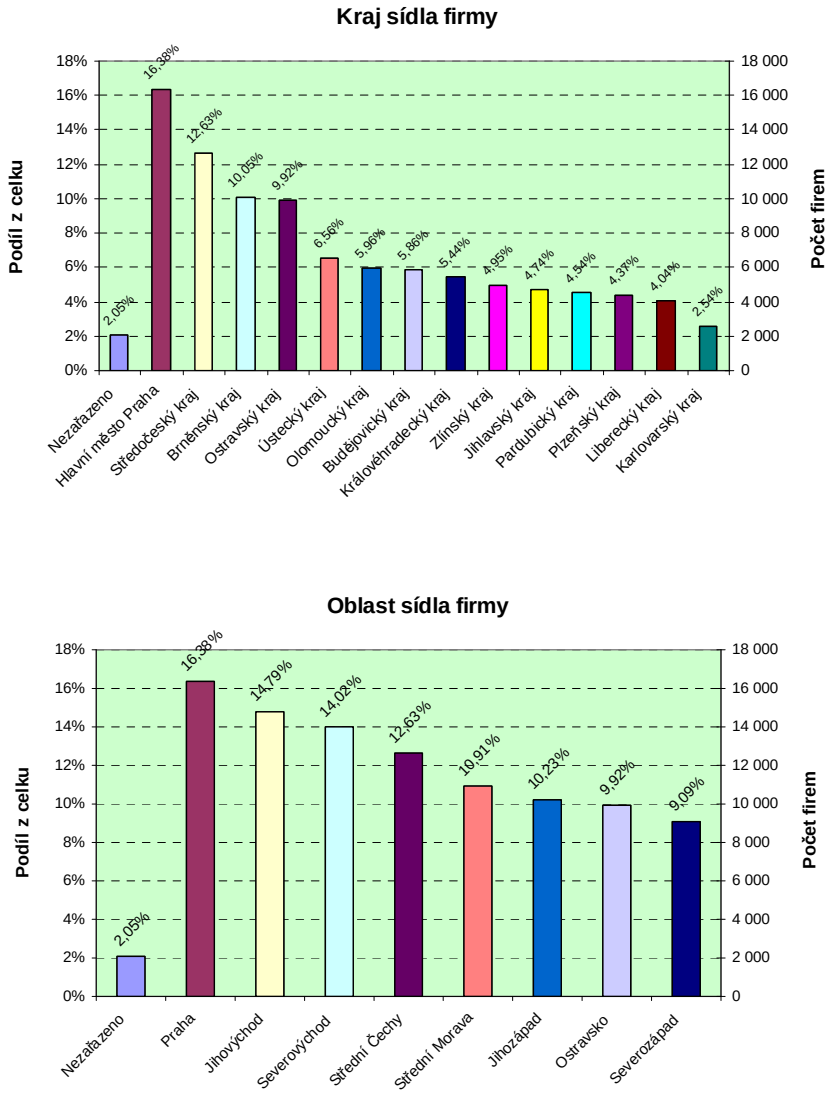
5. *Institucionální sektor* (viz. položka Isektor v tabulce 5). Tato proměnná obsahuje informace o institucionálních formách firem podle typu jejich podnikatelské činnosti. Bude odvozena z položky Isektor_ext (viz. tabulka 5), která má svůj původ v externích klientských datech. Původní proměnná vychází z číselníku Českého statistického úřadu institucionálních sektorů a subsektorů [7], jehož hodnoty lze uspořádat do hierarchie. Obsahuje celkem 21 různých hodnot (pozn. prázdné

hodnoty jsou chápány jako kategorie Nezařazeno), které reprezentují různé úrovně třídění. V nové proměnné bude pro třídění firem použita nejvyšší úroveň (u zahraničních subjektů) a druhá nejvyšší úroveň (u domácích subjektů). Tím bude redukován počet kategorií na 7. Na obrázku 21 jsou znázorněny histogramy, ze kterých je zřejmé zastoupení jednotlivých kategorií v původní a nové proměnné.



Obr. 21 Histogramy proměnných Institucionální sektor – původní (Isektor_ext) a Institucionální sektor (Isektor) - stav k 30.6.2006 měřený nad kombinací dat České pojišťovny a Registru ekonomických subjektů

6. *Oblast sídla firmy* (viz. položka Oblast v tabulce 5). Tato proměnná obsahuje územní členění České republiky dle oblastí, které vychází z klasifikace NUTS (úroveň třídění NUTS 2). Bude odvozena z položky Kraj (viz. tabulka 5), jejíž hodnoty představují jednotlivé kraje (úroveň třídění NUTS 3). Původní proměnná pochází z externích klientských dat a obsahuje celkem 15 kategorií (pozn.: prázdné hodnoty jsou chápány jako kategorie Nezařazeno). V nové proměnné budou kraje sloučeny do oblastí, čímž bude počet kategorií redukován na 9. Na obrázku 22 jsou znázorněny histogramy, ze kterých je zřejmé zastoupení jednotlivých kategorií v původní i nové proměnné.



Obr. 22 Histogramy proměnných Kraj sídla firmy (Kraj) a Oblast sídla firmy (Oblast) - stav k 30.6.2006 měřený nad kombinací dat České pojišťovny a Registru ekonomických subjektů

Tab. 5 Struktura tabulky Segmentace 1 – seznam proměnných a popis jejich typu, původu a významu

	Název proměnné	Typ	Význam	Původ
	IC	Kategoriální	Identifikační číslo (IČ)	Interní a externí data ID řádku tabulky
	Obchjm	Kategoriální	Obchodní jméno firmy	Interní a externí data
	Ulice	Kategoriální	Ulice	Interní a externí data
	Adrcislo	Kategoriální	Číslo orientační/popisné	Interní a externí data
	Obec	Kategoriální	Obec	Interní a externí data
	Psc	Kategoriální	PSČ	Interní a externí data
	R_predpis	Spojitá	Suma očekávaného ročního předpisového pojistného v Kč	Odvozená položka
	R_predpis_subst	Spojitá	Suma očekávaného ročního předpisového pojistného v Kč po substituci	Odvozená položka
	R_predpis_std	Spojitá	Suma očekávaného ročního předpisového pojistného po standardizaci	Odvozená položka
	Stari	Spojitá	Stáří firmy v letech	Odvozená položka
	Stari_subst	Spojitá	Stáří firmy v letech po substituci	Odvozená položka
	Stari_std	Spojitá	Stáří firmy po standardizaci	Odvozená položka
	Dat_vzn	Datum	Datum vzniku firmy	Externí data
	Jmeni	Spojitá	Základní jmění společnosti v Kč	Externí data
	Jmeni_subst	Spojitá	Základní jmění společnosti v Kč po substituci	Odvozená položka
	Jmeni_std	Spojitá	Základní jmění společnosti po standardizaci	Odvozená položka
	Pomer	Spojitá	Index storen pro neplacení	Odvozená položka
	Pomer_std	Spojitá	Index storen pro neplacení po standardizaci	Odvozená položka
	Aktiv	Kategoriální	Status podnikatelské činnosti firmy dle externích dat	Odvozená položka
	Aktivni	Kategoriální	Status podnikatelské činnosti firmy - původní	Externí data
	Dat_ukon	Datum	Datum ukončení podnikatelské činnosti firmy	Externí data
	Pocprac	Kategoriální	Počet pracovníků	Odvozená položka
	Pocprac_ext	Kategoriální	Počet pracovníků - původní	Externí data
	Pravfor	Kategoriální	Právní forma	Odvozená položka
	Pravfor_ext	Kategoriální	Právní forma – původní	Externí data
	Isektor	Kategoriální	Institucionální sektor	Odvozená položka
	Isektor_ext	Kategoriální	Institucionální sektor - původní	Externí data
	Okec_sekce	Kategoriální	Hlavní předmět podnikatelské činnosti	Odvozená položka
	Okec	Kategoriální	Hlavní předmět podnikatelské činnosti - původní	Externí data
	Oblast	Kategoriální	Oblast sídla firmy	Odvozená položka
	Kraj	Kategoriální	Kraj sídla firmy	Externí data

4.2 Statistická analýza

V této disertační práci bude provedena shluková analýza, v rámci které budou firmy rozděleny do několika stejnorodých skupin na základě spojitých proměnných. Příslušnost ke skupině bude vyjádřena novou kategoriální proměnnou Shluk. V návaznosti na shlukovou analýzu bude provedena také analýza kategoriálních dat, ve které budou hodnoceny vztahy mezi novou proměnnou a ostatními kategoriálními proměnnými z ploché tabulky. Cílem obou analýz bude rozdělení firem do obchodních segmentů.

Algoritmy popsané v této kapitole budou použity k odvození informací z ploché tabulky včetně hodnocení jejich průkaznosti. Z pohledu metodologie SEMMA se jedná o realizaci činností z etap Model a částečně i Assess. Z pohledu metodologie CRISP-DM pak o činnosti z etapy Modelování.

4.2.1 Provedení shlukové analýzy

Shluková analýza bude provedena v systému SAS pomocí procedury FASTCLUS, jejíž autoři se inspirovali přímo MacQueenovým k-průměrovým algoritmem [35]. Tato procedura je vhodná pro zpracování velkých datových souborů. Provádí nehierarchické shlukování, při kterém třídí zkoumané jednotky tak, že každá z nich je zařazena právě do jednoho shluku. Optimálního rozkladu jednotek do shluků je dosaženo splněním podmínky kritéria optimality. Do procedury FASTCLUS mohou vstupovat pouze spojité proměnné.

Během shlukování provádí procedura FASTCLUS následující kroky:

1. Výběr počátečních středů shluků (tzv. cluster seeds).
2. Vytvoření dočasných shluků přiřazením všech zkoumaných jednotek k nejbližším středům a jejich přepočtení (tento krok provádí procedura volitelně).
3. Vytvoření finálních shluků přiřazením všech zkoumaných jednotek k nejbližším středům.

Způsob provedení těchto kroků je závislý na nastavení procedury. V následujícím textu je popsáno nastavení, které bude použito pro účely této disertační práce.

4.2.1.1 Výběr počátečních středů shluků

Procedura FASTCLUS bude použita pro dva typy úloh. Prvním z nich bude tzv. předběžné shlukování, jehož cílem je identifikace vhodných počátečních středů pro hlavní shlukovou analýzu. Tento krok je doporučován zejména tehdy, je-li vstupní datový soubor velký nebo obsahuje-li odlehlá pozorování [35]. Druhým typem úlohy bude hlavní shluková analýza, v rámci které budou firmy rozděleny do finálních shluků. Předběžná i hlavní analýza začne stanovením počátečních středů, jimiž mohou být pouze kompletní vektory hodnot. Souřadnice těchto vektorů budou tvořeny hodnotami proměnných Jmeni_std, R_predpis_std, Stari_std a Pomer_std (viz. tabulka 5).

Pro předběžnou analýzu se doporučuje použít velký počet počátečních středů (např. 20 až 100), přičemž přesný počet není důležitý [35]. Pro účely této disertační práce bude z tabulky Segmentace 1 vybráno 50 středů podle standardního algoritmu procedury FASTCLUS [35]. Tento algoritmus testuje vztahy mezi potenciálními středy, přičemž kritériem pro jejich výběr je dosažení alespoň minimální vzdálenosti mezi nimi. V našem případě bude použito standardní nastavení s minimální vzdáleností rovnou 0. Shluková analýza bude dále pokračovat přiřazením všech firem z tabulky Segmentace 1 k nejbližším středům, přičemž nebude proveden jejich přepočet. Výsledné shluky budou vytvořeny na základě jedné iterace (tj. na základě jednoho kola přiřazování). Splnění podmínky optimalizačního kritéria nebude testováno. Optimálnost rozkladu jednotek do shluků je totiž posuzována na základě relativních změn v polohách středů po jejich přepočtu (viz. kapitola 4.2.1.3). Vztahy mezi jednotkami a středy shluků budou posuzovány na základě euklidovské vzdálenosti (viz. kapitola 4.2.1.2).

V rámci předběžné shlukové analýzy vytvoří procedura FASTCLUS tabulku Segmentace 2, která bude obsahovat následující proměnné:

1. *Číslo shluku* (Shluk). Tato proměnná obsahuje označení výsledných shluků hodnotami od 1 do 50.
2. *Počet jednotek uvnitř shluku* (_Freq_). Tato proměnná obsahuje počty firem zařazených do jednotlivých shluků.

3. *Maximální vzdálenost od středu shluku* (_Radius_). Tato proměnná obsahuje euklidovské vzdálenosti mezi středy shluků a jednotkami, které jsou od nich nejdále.
4. *Vzdálenost od středu nejbližšího shluku* (_Gap_). Tato proměnná obsahuje euklidovské vzdálenosti mezi středy shluků, které jsou si nejblíže.
5. *Číslo nejbližšího shluku* (_Near_). Tato proměnná obsahuje čísla nejbližších shluků.
6. *Střední kvadratická směrodatná odchylka* (_Rmsstd_). Hodnoty této proměnné jsou vypočteny s použitím vzorců 4.8 a 4.9 [35]:

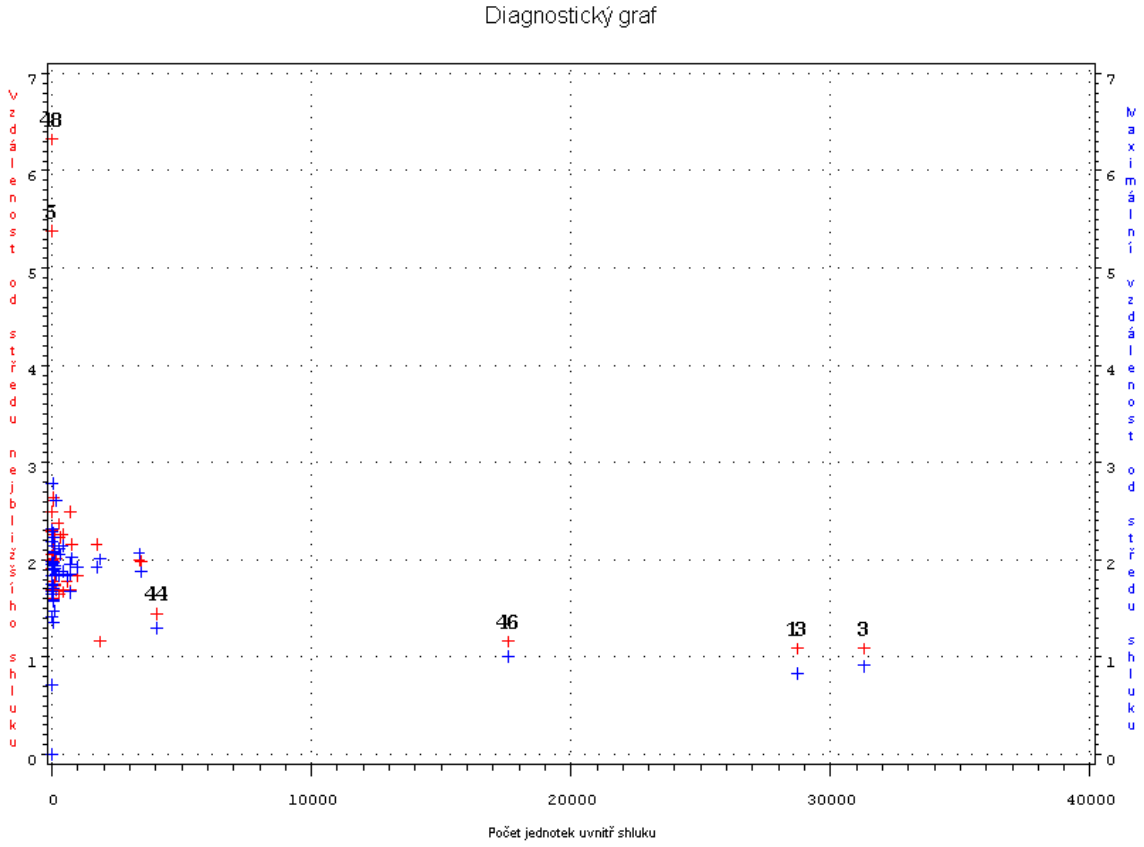
$$RMSSTD = \sqrt{\frac{W_K}{v(N_K - 1)}}, \qquad (4.8)$$

$$\text{přičemž } W_K = \sum_{i \in C_K} \left\| x_i - \bar{x}_K \right\|^2. \qquad (4.9)$$

   V těchto vzorcích odpovídá $|| x_i - \bar{x_K}||$ euklidovské vzdálenosti mezi i-tým řádkovým vektorem x_i a vektorem průměrů $\bar{x_K}$ ze shluku C_k (i je z intervalu 1 až N_k). N_k představuje počet řádkových vektorů v k-tém shluku a v počet proměnných, jejichž hodnoty tvoří souřadnice těchto vektorů.
7. *Vnitroshlukové průměry vstupních proměnných* (Avg_jmeni_std, Avg_r_predpis_std, Avg_stari_std a Avg_pomer_std). Hodnoty těchto proměnných tvoří dohromady vektory průměrů, ze kterých budou vybrány počáteční středy pro hlavní shlukovou analýzu. Nejedná se přitom o středy shluků z předběžné analýzy. Těmi jsou stále počáteční cluster seeds, poněvadž nebyl proveden jejich přepočet.

Tabulka Segmentace 2 je uvedena v příloze 1. Jednotlivé shluky z této tabulky budou zobrazeny v diagnostickém grafu podle proměnných *Počet jednotek uvnitř shluku* (_Freq_), *Maximální vzdálenost od středu shluku* (_Radius_) a *Vzdálenost středu od nejbližšího shluku* (_Gap_). S pomocí tohoto grafu budou z tabulky vybrány nejvhodnější vektory průměrů, které budou dále použity jako počáteční středy v hlavní shlukové analýze. Dobře separované a početné shluky lze očekávat v pravé horní části grafu, která reprezentuje oblast s vysokými hodnotami proměnných _Freq_ a _Gap_. Potenciálně dobré shluky se objevují také v pravé dolní části [35]. V levé části grafu lze

díky nízkým hodnotám proměnné _Freq_ očekávat shluky s malým počtem členů. Jejich vektory průměrů nejsou vhodné pro další použití v hlavní analýze. Shluky v levé horní části často představují odlehlá pozorování, která se jeví jako samostatné shluky. Na základě porovnání hodnot proměnných _Gap_ a _Radius_ je možné dále posoudit vnitřní vyrovnanost jednotlivých shluků. Maximální vzdálenost jednotky od středu (_Radius_) je u vyrovnaných shluků menší než vzdálenost středu od nejbližšího shluku (_Gap_). Diagnostický graf vytvořený pro účely této disertační práce je uveden na obrázku 23.



Obr. 23 Diagnostický graf pro určení počátečních středů v hlavní shlukové analýze

Jako potenciálně dobré se jeví shluky č. 3, 13, 44 a 46. Tyto shluky dosahují vysokých hodnot v proměnné _Freq_, což znamená, že obsahují velké množství firem. Střední kvadratické směrodatné odchylky jsou nízké. Zároveň hodnoty proměnné _Radius_ jsou menší než hodnoty proměnné _Gap_. Shluky lze tedy považovat za početné a vnitřně vyrovnané. Jejich vektory průměrů budou použity jako počáteční středy v hlavní shlukové analýze. Záznamy z tabulky Segmentace 2, které se k těmto shlukům vztahují, jsou uvedeny v tabulce 6.

V levé horní části diagnostického grafu jsou dále patrná odlehlá pozorování (shluky č. 5 a 48), která se nepodařilo identifikovat a transformovat při přípravě tabulky Segmentace 1. Jedná se pravděpodobně o vícenásobné extrémy. Jejich řádkové vektory nebudou zařazeny do shluků v prvním kroku hlavní analýzy (viz. kapitola 4.2.1.2).

Tab. 6 Věty z tabulky Segmentace 2 – vektory průměrů a další charakteristiky shluků č. 3, 13, 44 a 46

Shluk	_Freq_	_Radius_	_Gap_	_Near_	_Rmsstd_
Číslo shluku	Počet jednotek uvnitř shluku	Maximální vzdálenost od středu shluku	Vzdálenost od středu nejbližšího shluku	Číslo nejbližšího shluku	Střední kvadratická směrodatná odchylka
3	31 306	0,91180310	1,08760854	13	0,18351745
13	28 724	0,83049707	1,08760854	3	0,28068248
44	4 044	1,30118400	1,44531272	3	0,15325592
46	17 597	1,00667598	1,16230945	49	0,26475783

Shluk	Avg_jmeni_std	Avg_r_predpis_std	Avg_stari_std	Avg_pomer_std
Číslo shluku	Vnitroshlukový průměr proměnné Jmeni_std	Vnitroshlukový průměr proměnné R_predpis_std	Vnitroshlukový průměr proměnné Stari_std	Vnitroshlukový průměr proměnné Pomer_std
3	-0,13317274	-0,23709700	0,57373674	-0,43086793
13	-0,15457487	-0,28595519	0,09794853	0,54569377
44	-0,00003661	1,12063279	0,09918745	-0,38010968
46	-0,12757006	-0,28050601	-1,44795319	-0,40979002

4.2.1.2 Hlavní shluková analýza

Hlavní shluková analýza bude opět provedena nad tabulkou Segmentace 1. V prvním kroku nebudou do shluků zařazena pozorování, jejichž vzdálenosti ke středům budou větší než stanovená maximální mez. Cílem je zabránit zejména odlehlým pozorováním zkreslení výsledků shlukové analýzy. Tuto mez se doporučuje stanovit tak, aby byla blízká hodnotám proměnných _Radius_ a _Gap_ větších shluků z předběžné analýzy [35]. V našem případě bude nastavena na hodnotu 1,2. První krok hlavní analýzy bude proveden ve dvaceti iteracích, přičemž jako počáteční středy budou použity vektory průměrů vybrané z tabulky Segmentace 2 (viz. tabulka 6). Po každém kole přiřazení všech jednotek budou vektory průměrů přepočteny. Vztahy mezi řádkovými vektory a vektory průměrů budou posuzovány na základě euklidovské vzdálenosti. Do zpracování budou vstupovat i vektory, kterým chybí některá ze souřadnic. Vzdálenost bude počítána podle vzorce 4.10 [35]:

$$\sqrt{\frac{n}{m}\sum (x_i - s_i)^2}, \qquad (4.10)$$

kde n je celkový počet souřadnic v řádkovém vektoru, m je počet nechybějících souřadnic, x_i je i-tá souřadnice řádkového vektoru a s_i je i-tá souřadnice vektoru průměrů. Suma je přitom uvažována pouze pro nechybějící souřadnice.

Výsledkem prvního kroku hlavní analýzy budou tabulky Segmentace 3 a Segmentace 4. Tabulka Segmentace 3 představuje původní plochou tabulku rozšířenou o nové proměnné Číslo shluku a Vzdálenost jednotky od středu shluku (viz. položky Shluk a Distance v tabulce 7). Procedura FASTCLUS identifikovala v prvním kroku celkem čtyři shluky. Pro záznamy, které byly do nich zařazeny, nabývá proměnná Číslo shluku celočíselných hodnot od 1 do 4. Do těchto shluků však nebyla zařazena pozorování, jejichž vzdálenosti k nejbližším středům jsou větší než stanovená mez (včetně odlehlých pozorování). Proměnná Číslo shluku v těchto případech nabývá celočíselných hodnot od –4 do –1 podle toho, k jakému shluku má daný záznam nejblíže.

Na základě tabulky Segmentace 3 budou vypočteny vnitroshlukové popisné statistiky pro proměnné Jmeni_subst, R_predpis_subst, Stari_subst a Pomer. Tento výpočet bude proveden pomocí procedury UNIVARIATE podle vzorců 4.2 až 4.6. Výsledky budou použity pro profilaci obchodních segmentů dle spojitých proměnných. Na základě statistik, při jejichž výpočtu byl zmírněn vliv odlehlých pozorování (jejich transformací nebo vynecháním), lze totiž stanovit průkaznější pravidla pro zařazování firem do segmentů než při použití kompletních a netransformovaných dat. Vnitroshlukové statistiky vypočtené nad tabulkou Segmentace 3 jsou uvedeny v tabulkách 14, 15, 16 a 17 v kapitole 5.1.

Tabulka Segmentace 4 má stejnou strukturu jako tabulka Segmentace 2. Obsahuje vektory průměrů pro shluky vypočtené v prvním kroku hlavní analýzy. Tyto vektory budou použity v druhém kroku jako počáteční středy. Tabulka Segmentace 4 je uvedena v příloze 2.

V druhém kroku hlavní analýzy budou do shluků zařazeny všechny firmy z tabulky Segmentace 1. Výsledkem bude tabulka Segmentace 5, jejíž struktura je shodná s tabulkou Segmentace 3. Proměnná Shluk v ní bude nabývat celočíselných hodnot od 1 do 4 pro všechny záznamy. Druhý krok hlavní analýzy bude proveden v jedné iteraci bez přepočtení středů shluků. Na základě tabulky Segmentace 5 budou opět vypočteny vnitroshlukové popisné statistiky, tentokrát však pro proměnné Jmeni, R_predpis, Stari a Pomer (viz. tabulka 7). Výsledky získané nad netransformovanými a kompletními daty budou při profilaci obchodních segmentů konfrontovány s výsledky vypočtenými z tabulky Segmentace 3. Vnitroshlukové statistiky vypočtené nad tabulkou Segmentace 5 jsou opět uvedeny v tabulkách 14, 15, 16 a 17 v kapitole 5.1.

Tab. 7 Struktura tabulek Segmentace 3 a Segmentace 5 - seznam proměnných a popis jejich typu, původu a významu

	Název proměnné	Typ	Význam	Původ
	IC	Kategoriální	Identifikační číslo (IČ)	Interní a externí data ID řádku tabulky
	Obchjm	Kategoriální	Obchodní jméno firmy	Interní a externí data
	Ulice	Kategoriální	Ulice	Interní a externí data
	Adrcislo	Kategoriální	Číslo orientační/popisné	Interní a externí data
	Obec	Kategoriální	Obec	Interní a externí data
	Psc	Kategoriální	PSČ	Interní a externí data
	R_predpis	Spojitá	Suma očekávaného ročního předpisového pojistného v Kč	Odvozená položka
	R_predpis_subst	Spojitá	Suma očekávaného ročního předpisového pojistného v Kč po substituci	Odvozená položka
	R_predpis_std	Spojitá	Suma očekávaného ročního předpisového pojistného po standardizaci	Odvozená položka
	Stari	Spojitá	Stáří firmy v letech	Odvozená položka
	Stari_subst	Spojitá	Stáří firmy v letech po substituci	Odvozená položka
	Stari_std	Spojitá	Stáří firmy po standardizaci	Odvozená položka
	Dat_vzn	Datum	Datum vzniku firmy	Externí data
	Jmeni	Spojitá	Základní jmění společnosti v Kč	Externí data
	Jmeni_subst	Spojitá	Základní jmění společnosti v Kč po substituci	Odvozená položka
	Jmeni_std	Spojitá	Základní jmění společnosti po standardizaci	Odvozená položka
	Pomer	Spojitá	Index storen pro neplacení	Odvozená položka
	Pomer_std	Spojitá	Index storen pro neplacení po standardizaci	Odvozená položka
	Aktiv	Kategoriální	Status podnikatelské činnosti firmy dle externích dat	Odvozená položka
	Aktivni	Kategoriální	Status podnikatelské činnosti firmy - původní	Externí data
	Dat_ukon	Datum	Datum ukončení podnikatelské činnosti firmy	Externí data
	Pocprac	Kategoriální	Počet pracovníků	Odvozená položka
	Pocprac_ext	Kategoriální	Počet pracovníků - původní	Externí data
	Pravfor	Kategoriální	Právní forma	Odvozená položka
	Pravfor_ext	Kategoriální	Právní forma - původní	Externí data
	Isektor	Kategoriální	Institucionální sektor	Odvozená položka
	Isektor_ext	Kategoriální	Institucionální sektor - původní	Externí data
	Okec_sekce	Kategoriální	Hlavní předmět podnikatelské činnosti	Odvozená položka
	Okec	Kategoriální	Hlavní předmět podnikatelské činnosti - původní	Externí data
	Oblast	Kategoriální	Oblast sídla firmy	Odvozená položka
	Kraj	Kategoriální	Kraj sídla firmy	Externí data
	Shluk	Kategoriální	Číslo shluku	Odvozená položka
	Distance	Spojitá	Vzdálenost jednotky od středu shluku	Odvozená položka

Legenda: Modře jsou vyznačeny proměnné přidané do ploché tabulky v hlavní shlukové analýze.

4.2.1.3 Hodnocení výsledků shlukové analýzy

Rozbor výsledků předběžné analýzy je proveden na základě statistik z tabulky Segmentace 2 a diagnostického grafu, který je uveden na obrázku 23 (viz. kapitola 4.2.1.1). Optimálnost rozkladu jednotek, která je posuzována na základě relativních změn středů shluků po jejich přepočtu, není hodnocena. Předběžná analýza byla totiž provedena v jedné iteraci bez přepočtu středů shluků.

Optimálnost rozkladu bude hodnocena až v prvním kroku hlavní analýzy. Na základě nastavení procedury FASTCLUS bude kvalitního rozkladu jednotek do shluků dosaženo, jestliže maximální relativní změna v polohách středů bude menší nebo rovna hodnotě konvergence. Ta bude pro účely této disertační práce nastavena na 0,0001 (tj. bude blízká 0). Minimální změny v polohách středů znamenají homogenní rozložení jednotek uvnitř shluků. Změny poloh středů budou vypočteny v každé iteraci jako euklidovské vzdálenosti mezi původními a novými středy dělené tzv. škálovacím faktorem. Tím se rozumí průměrná absolutní odchylka mezi původními středy. V tabulce 8 je uvedena historie relativních změn středů shluků pro jednotlivé iterace.

Tab. 8 Historie relativních změn středů shluků v jednotlivých iteracích pro první krok hlavní shlukové analýzy

Číslo iterace	Shluk 1	Shluk 2	Shluk 3	Shluk 4
1	0,427600	0,263200	0,243000	0,285900
2	0,029900	0,543400	0,153300	0,045100
3	0,011900	0,368700	0,091600	0,013200
4	0,004300	0,306300	0,058200	0,004400
5	0,003560	1,104900	0,042700	0,002820
6	0,014900	1,301600	0,036200	0,007890
7	0,015400	0,537100	0,029700	0,001010
8	0,001620	0,074400	0,022400	0,000318
9	0,001070	0,010300	0,016200	0,000085
10	0,000524	0,000916	0,012000	0
11	0,000726	0	0,011200	0,000099
12	0,000594	0	0,006480	0,000006
13	0,000279	0	0,004150	0,000120
14	0,000166	0	0,002150	0
15	0,000079	0	0,002910	0
16	0,000157	0	0,001650	0
17	0,000064	0	0,001010	0,000006
18	0,000053	0	0,001230	0,000086
19	0,000032	0	0,000279	0
20	0	0	0	0

Jak je patrné z této tabulky, v devatenácté iteraci je maximální relativní změna (u shluku č. 3) stále větší než hodnota konvergence. Ve dvacáté iteraci jsou již všechny změny rovny 0, čili bylo dosaženo optimálního rozkladu. Následovat bude hodnocení výsledku podle statistik, které charakterizují jednotlivé shluky. Tyto statistiky jsou uvedeny v tabulce 9 a jsou rovněž součástí tabulky Segmentace 4, která je uvedena v příloze 2.

Tab. 9 Charakteristiky shluků po prvním kroku hlavní analýzy

Shluk	_Freq_	_Radius_	_Gap_	_Near_	_Rmsstd_
Číslo shluku	Počet jednotek uvnitř shluku	Maximální vzdálenost od středu shluku	Vzdálenost od středu nejbližšího shluku	Číslo nejbližšího shluku	Střední kvadratická směrodatná odchylka
1	49 100	1,1000	1,1807	3	0,2315
2	11 720	1,1994	3,0155	1	0,2367
3	5 544	1,0992	1,1807	1	0,3667
4	25 208	1,1999	1,6334	1	0,3064

Jak je patrné z tabulky 9 nejvíce firem obsahuje shluk č. 1. Naopak nejméně firem obsahuje shluk č. 3. Do shluků bylo v prvním kroku hlavní analýzy zařazeno celkem 91 572 firem. Tzn., že 8 428 záznamů z tabulky Segmentace 1 nevyhovělo podmínce maximální vzdálenosti pro přiřazení jednotek k nejbližším středům. Střední kvadratické směrodatné odchylky (_Rmsstd_) jsou pro všechny shluky nízké. Zároveň hodnoty proměnné _Radius_ jsou menší než hodnoty proměnné _Gap_. Výsledné shluky lze tedy považovat za vnitřně vyrovnané.

Výsledek prvního kroku hlavní analýzy bude dále hodnocen na základě porovnání celkových a vnitroshlukových směrodatných odchylek vstupních proměnných (viz. položky Totalstd a Whithinstd v tabulce 10). Celková směrodatná odchylka bude pro každou proměnnou vypočtena podle vzorce 4.3. Aritmetický průměr použitý v tomto vzorci bude vypočten ze všech neprázdných hodnot dané proměnné. Podle vzorce 4.3 bude dále pro každou proměnnou vypočtena také vnitroshluková směrodatná odchylka. V tomto případě budou ve vzorci použity vnitroshlukové průměry. Tzn., že pro danou proměnnou budou od jejích hodnot odečteny průměry shluků, do kterých byly tyto hodnoty, coby souřadnice řádkových vektorů, zařazeny. Čím větší jsou rozdíly mezi celkovými a vnitroshlukovými odchylkami, tím průkaznější výsledky model poskytuje.

V tabulce 10 jsou uvedeny celkové a vnitroshlukové směrodatné odchylky vstupních proměnných, jež byly vypočteny v prvním kroku hlavní analýzy. Jak je patrné z této tabulky, mezi odchylkami jsou pro všechny proměnné velké rozdíly. To potvrzuje kvalitní rozklad zkoumaných jednotek do shluků.

Tab. 10 Směrodatné odchylky vstupních proměnných po prvním kroku hlavní analýzy

Proměnná	Totalstd	Withinstd	Rozdíl
	Celková směrodatná odchylka	Vnitroshluková směrodatná odchylka	
Jmeni_std	0,8811	0,0801	0,8010
R_predpis_std	0,7694	0,1436	0,6258
Stari_std	0,8945	0,4170	0,4775
Pomer_std	1,0069	0,1994	0,8075

V druhém kroku hlavní analýzy budou do shluků zařazeny všechny záznamy z tabulky Segmentace 1 (včetně těch, které v prvním kroku nevyhověly podmínce maximální vzdálenosti pro přiřazení k nejbližším středům). Tento krok bude proveden v jedné iteraci, přičemž středy shluků nebudou přepočteny. Splnění podmínky optimalizačního kritéria tedy nebude testováno.

V tabulce 11 jsou uvedeny statistiky, které charakterizují výsledné shluky po přiřazení všech jednotek. Oproti výsledkům z prvního kroku hlavní analýzy převyšují hodnoty proměnné _Radius_ výrazně hodnoty proměnné _Gap_. Zároveň došlo ke zvýšení hodnot středních kvadratických směrodatných odchylek (_Rmsstd_). Nejvíce firem stále obsahuje shluk č. 1 a nejméně shluk č. 3.

Tab. 11 Charakteristiky shluků po druhém kroku hlavní analýzy

Shluk	_Freq_	_Radius_	_Gap_	_Near_	_Rmsstd_
Číslo shluku	Počet jednotek uvnitř shluku	Maximální vzdálenost od středu shluku	Vzdálenost od středu nejbližšího shluku	Číslo nejbližšího shluku	Střední kvadratická směrodatná odchylka
1	50 729	9,4011	1,7069	4	0,5011
2	12 733	8,7756	2,9370	1	0,4784
3	10 072	11,0114	2,6872	1	1,7414
4	26 466	8,8162	1,7069	1	0,5755

V tabulce 12 jsou dále uvedeny celkové a vnitroshlukové směrodatné odchylky vstupních proměnných. V porovnání s tabulkou 10 se jejich rozdíly snížily.

Tab. 12 Směrodatné odchylky vstupních proměnných po druhém kroku hlavní analýzy

Proměnná	Totalstd	Withinstd	Rozdíl
	Celková směrodatná odchylka	Vnitroshluková směrodatná odchylka	
Jmeni_std	1	0,9430	0,0570
R_predpis_std	1	0,6906	0,3094
Stari_std	1	0,6088	0,3912
Pomer_std	1	0,5844	0,4156

Použití všech zkoumaných jednotek vede ke zhoršení vlastností shluků. Rozklad jednotek však lze stále považovat za optimální. Cílem shlukové analýzy v této disertační práci je zařadit do shluků všechny jednotky. To bylo provedeno v druhém kroku hlavní analýzy jejich přiřazením k optimálním středům z prvního kroku.

4.2.2 Provedení analýzy kategoriálních dat

Analýza kategoriálních dat bude provedena v systému SAS pomocí procedury FREQ. V rámci této analýzy bude nad tabulkou Segmentace 5 testována nezávislost proměnné Shluk a proměnných Pravfor, Isektor, Okec_sekce, Pocprac, Oblast a Aktiv (viz. tabulka 7). V případě zjištění závislosti bude dále měřena i její intenzita (těsnost). Za účelem provedení testů budou zkoumané jednotky nejprve uspořádány do kontingenčních tabulek (viz. kapitola 5.2). Jednotlivé buňky v těchto tabulkách obsahují počty firem, které splňují podmínku výskytu dané kombinace hodnot testovaných proměnných.

Nezávislost proměnných bude testována s použitím Pearsonova chí-kvadrát testu, přičemž hodnota testovacího kritéria bude vypočtena pomocí následujících vzorců [35]:

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - e_{ij})^2}{e_{ij}}, \qquad (4.11)$$

$$\text{přičemž } e_{ij} = \frac{n_{i.}\, n_{.j}}{n}. \qquad (4.12)$$

V těchto vzorcích představuje n_{ij} skutečnou četnost v i-tém řádku a j-tém sloupci kontingenční tabulky. Znak e_{ij} reprezentuje teoretické četnosti, znak $n_{i.}$ řádkové součty, znak $n_{.j}$ sloupcové součty a znak n celkový součet hodnot v kontingenční tabulce.

Pearsonův chí-kvadrát test nelze použít, jestliže více než 20 % teoretických četností e_{ij} je menší než 5, nebo alespoň jedna z nich je menší než 1. Teoretickou

četností se rozumí počet zkoumaných jednotek, který je očekáván v i-tém řádku a j-tém sloupci kontingenční tabulky. Je odvozena z pravděpodobnosti výskytu dané kombinace hodnot testovaných proměnných. V našem případě byla ve všech prováděných testech splněna podmínka pro použití kritéria chí-kvadrát.

Pomocí Pearsonova chí-kvadrát testu bude ověřena platnost nulové hypotézy o nezávislosti testovaných proměnných. Je-li hodnota statistiky chí-kvadrát větší než kritická hodnota z teoretického chí-kvadrát rozdělení pro (r-1)(c-1) stupňů volnosti, kde *r* je počet řádků a *c* počet sloupců v kontingenční tabulce, potom nulovou hypotézu zamítáme. Ve výstupech procedury FREQ je hodnota statistiky chí-kvadrát spojena s pravděpodobností, resp. hladinou významnosti, *p* (viz. položka P_pchi v tabulce 13). Bude-li v našich testech tato pravděpodobnost menší než 0,05, nulovou hypotézu o nezávislosti proměnných zamítneme. Výsledné statistiky spojené s kritériem chí-kvadrát jsou uvedeny v tabulce 13. Hodnoty proměnné P_pchi jsou pro všechny kombinace testovaných proměnných menší než 0,0001, čili jsou menší než 0,05. Ve všech případech tedy zamítáme nulovou hypotézu a proměnné považujeme za vzájemně závislé. V položce N jsou uvedeny počty firem s naplněnými hodnotami obou proměnných. V případě, že není naplněna alespoň jedna z nich, je daná firma vyloučena z testu závislosti. Počty všech těchto firem jsou uvedeny v tabulce 13 v položce Nmiss. Položka _Pchi_ obsahuje hodnoty kritéria chí-kvadrát a položka Df_pchi použité stupně volnosti.

Tab. 13 Statistiky kritéria chí-kvadrát pro jednotlivé kombinace testovaných proměnných

Testované proměnné	N	Nmiss	_Pchi_	Df_pchi	P_pchi	_Contgy_
	Počet firem s naplněnými hodnotami	Počet firem s alespoň jednou nenaplněnou hodnotou	Kritérium chí-kvadrát	Stupně volnosti	Pravděpodobnost p	Koeficient kontingence
Shluk x Aktiv	97 956	2 044	67 030,9351	6	< 0,0001	0,6374
Shluk x Pocprac	85 234	14 766	20 867,7440	21	< 0,0001	0,4435
Shluk x Pravfor	97 956	2 044	22 518,5464	30	< 0,0001	0,4323
Shluk x Isektor	97 951	2 049	16 273,1817	15	< 0,0001	0,3774
Shluk x Okec_sekce	97 954	2 046	6 737,8706	45	< 0,0001	0,2537
Shluk x Oblast	97 954	2 046	561,8956	21	< 0,0001	0,0755

Intenzita závislosti testovaných proměnných bude měřena pomocí Pearsonova koeficientu kontingence. Tato statistika bude vypočtena podle následujícího vzorce [35]:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}, \qquad (4.13)$$

kde χ^2 je hodnota kritéria chí-kvadrát a n celkový součet hodnot v kontingenční tabulce. Pearsonův koeficient kontingence nabývá hodnot z následujícího rozmezí:

$$0 \leq C \leq \sqrt{\frac{m-1}{m}}, \qquad (4.14)$$

kde m je minimum z počtu řádků a počtu sloupců v kontingenční tabulce. Čím více se hodnota koeficientu kontingence blíží horní mezi z výše uvedeného intervalu, tím je závislost mezi proměnnými těsnější.

Hodnoty Pearsonova koeficientu kontingence jsou pro jednotlivé kombinace testovaných proměnných uvedeny v tabulce 13 v položce _Contgy_. Jak je patrné z této tabulky, nejtěsnější závislost je mezi proměnnými Shluk a Aktiv s hodnotou 0,6374 a Shluk a Pocprac s hodnotou 0,4435. Nejmenší těsnost závislosti je mezi proměnnými Shluk a Oblast s hodnotou 0,0755. V tomto případě je hodnota koeficientu kontingence velmi blízká 0, čili těsnost závislosti je malá.

5 Rozbor výsledků

V této kapitole bude proveden rozbor výsledků statistické analýzy z obchodního pohledu. Dle metodologie CRISP-DM se jedná o realizaci činností z etapy Vyhodnocení výsledků.

5.1 Výsledky shlukové analýzy

V rámci shlukové analýzy byly na základě spojitých proměnných Jmeni_std, R_predpis_std, Stari_std a Pomer_std identifikovány čtyři skupiny klientů, právnických osob, České pojišťovny. V tabulkách 14, 15, 16 a 17 jsou uvedeny popisné statistiky vypočtené nad nestandardizovanými proměnnými pro každý shluk. Vždy jsou uvedeny výsledky pro proměnné po provedení substituce odlehlých hodnot a bez jednotek, které v prvním kroku hlavní analýzy nevyhověly podmínce maximální vzdálenosti pro přiřazení k nejbližšímu středu (tj. pro proměnné Jmeni_subst, R_predpis_subst, Stari_subst a Pomer). Na základě těchto výsledků bude provedena profilace obchodních segmentů. Pro porovnání jsou v tabulkách rovněž uvedeny výsledky vypočtené nad kompletními a netransformovanými daty (pro proměnné Jmeni, R_predpis, Stari a Pomer).

V tabulce 14 jsou pro jednotlivé shluky uvedeny popisné statistiky proměnných Základní jmění společnosti v Kč po substituci (Jmeni_subst) a Základní jmění společnosti v Kč (Jmeni). Jak je patrné z této tabulky, všechny shluky obsahují firmy s nulovým základním jměním. Jedná se o ekonomické subjekty, které nemají povinnost základní jmění vytvářet. Tyto firmy převažují ve shlucích č. 1 a 2, přičemž v prvním jsou obsaženy minimálně ze 75-ti % a ve druhém minimálně z 99-ti % (viz. hodnoty horního kvartilu proměnné Jmeni_subst pro první shluk a 99. percentilu proměnné Jmeni_subst pro druhý shluk). Firmy se základním jměním jsou zastoupeny převážně ve shlucích č. 3 a 4. Ve shluku č. 3 činí průměrná hodnota základního jmění 431 180,22 Kč. Ve shluku č. 4 je tato hodnota rovna 158 410,11 Kč. Z průměrných hodnot ale i jednotlivých kvantilů proměnné Jmení_subst je tedy zřejmé, že shluk č. 3 obsahuje firmy s vyšším základním jměním než shluk č. 4. Popisné statistiky se pro všechny shluky výrazně liší při použití kompletních a netransformovaných dat (viz. hodnoty proměnné Jmeni).

Tab. 14 Popisné statistiky pro proměnné Základní jmění společnosti v Kč po substituci (Jmeni_subst) a Základní jmění společnosti v Kč (Jmeni)

Shluk	Proměnná	N	Avg	Std	Skew	Kurt	Min	P1	P5
		Počet neprázdných hodnot	Aritmetický průměr	Směrodatná odchylka	Šikmost	Špičatost	Minimum	1. percentil	5. percentil
1	Jmeni	43 540	1 203 378,69	82 175 386,33	157,64	27 719,38	0	0	0
1	Jmeni_subst	42 969	78 868,30	513 452,51	11,63	160,06	0	0	0
2	Jmeni	12 408	70 751,05	2 785 517,12	56,34	3 461,00	0	0	0
2	Jmeni_subst	11 653	3 858,83	86 978,00	34,87	1 365,59	0	0	0
3	Jmeni	8 269	50 603 118,76	641 181 050,22	34,87	1 462,94	0	0	0
3	Jmeni_subst	4 611	431 180,22	1 243 469,89	4,69	24,80	0	0	0
4	Jmeni	22 899	1 863 848,64	81 652 153,89	129,80	18 253,27	0	0	0
4	Jmeni_subst	21 926	158 410,11	652 678,38	8,65	89,64	0	0	0

Shluk	Proměnná	P10	Q1	Median	Q3	P90	P95	P99	Max
		10.percentil	Spodní kvartil	Medián	Horní kvartil	90. percentil	95.percentil	99.percentil	Maximum
1	Jmeni	0	0	0	0	100 000,00	200 000,00	6 000 000,00	15 200 000 000,00
1	Jmeni_subst	0	0	0	0	100 000,00	140 000,00	2 000 000,00	10 050 000,00
2	Jmeni	0	0	0	0	0	0	100 000,00	191 000 000,00
2	Jmeni_subst	0	0	0	0	0	0	0	4 000 000,00
3	Jmeni	0	0	100 000,00	2 001 000,00	50 000 000,00	126 912 000,00	797 923 000,00	32 208 990 000,00
3	Jmeni_subst	0	0	100 000,00	135 000,00	1 000 000,00	2 400 000,00	7 000 000,00	10 600 000,00
4	Jmeni	0	0	0	100 000,00	200 000,00	1 000 000,00	20 000 000,00	11 664 000 000,00
4	Jmeni_subst	0	0	0	100 000,00	200 000,00	500 000,00	3 100 000,00	10 100 000,00

Legenda: Měřeno nad kombinací dat České pojišťovny a Registru ekonomických subjektů k 30.6.2006.

V tabulce 15 jsou dále pro jednotlivé shluky uvedeny popisné statistiky proměnných Suma očekávaného ročního předpisového pojistného v Kč po substituci (R_predpis_subst) a Suma očekávaného ročního předpisového pojistného v Kč (R_predpis). Z tabulky je patrné, že mezi shluky č. 1, 2 a 4 nejsou výrazné rozdíly. Průměrné hodnoty očekávaného ročního předpisového pojistného se u těchto shluků pohybují v rozmezí 13 000 až 16 000 Kč (viz. proměnná R_predpis_subst). Oproti zbytku firem se výrazně liší shluk č. 3 s průměrným předpisem 87 789,50 Kč. Použití kompletních a netransformovaných dat nevede u shluků č. 1, 2 a 4 k výrazné změně hodnot popisných statistik (viz. proměnná R_predpis). Výrazná změna je patrná pouze u shluku č. 3, kde je např. průměrný předpis téměř čtyřnásobný (316 873,34 Kč).

Tab. 15 Popisné statistiky pro proměnné Očekávané roční předpisové pojistné v Kč po substituci (R_predpis_subst) a Očekávané roční předpisové pojistné v Kč (R_predpis)

Shluk	Proměnná	N	Avg	Std	Skew	Kurt	Min	P1	P5
		Počet neprázdných hodnot	Aritmetický průměr	Směrodatná odchylka	Šikmost	Špičatost	Minimum	1. percentil	5. percentil
1	R_predpis	50 729	13 180,15	11 642,21	1,18	0,71	39,00	216,00	586,00
1	R_predpis_subst	49 100	13 029,21	11 532,56	1,18	0,72	216,00	216,00	570,00
2	R_predpis	12 733	17 237,12	27 023,11	4,62	30,27	56,00	216,00	285,00
2	R_predpis_subst	11 720	14 337,43	16 331,12	2,22	5,75	216,00	216,00	285,00
3	R_predpis	10 072	316 873,34	4 891 637,30	85,79	7 934,26	29 172,00	50 082,00	53 582,00
3	R_predpis_subst	5 544	87 789,50	31 154,72	0,86	-0,21	47 574,00	50 652,00	52 646,00
4	R_predpis	26 466	17 317,01	20 193,03	2,82	11,56	23,00	223,00	1 000,00
4	R_predpis_subst	25 208	15 891,20	16 488,62	1,96	4,35	216,00	223,00	1 000,00

Shluk	Proměnná	P10	Q1	Median	Q3	P90	P95	P99	Max
		10.percentil	Spodní kvartil	Medián	Horní kvartil	90. percentil	95.percentil	99.percentil	Maximum
1	R_predpis	1 500,00	4 193,00	9 027,00	19 210,00	30 943,00	38 234,00	47 541,00	58089,00
1	R_predpis_subst	1 500,00	4 099,00	9 027,00	18 995,50	30 616,50	37 875,00	47 296,50	51 918,00
2	R_predpis	1 000,00	3 814,00	8 756,00	19 784,00	38 454,00	59 986,00	141 687,00	334 092,00
2	R_predpis_subst	1 000,00	3 814,00	8 640,00	18 657,50	34 673,00	49 039,00	81 146,00	101 616,00
3	R_predpis	57 246,00	71 310,00	113 488,00	216 000,00	452 457,00	763 012,00	2 315 156,00	462 011 646,00
3	R_predpis_subst	54 668,00	61 990,00	78 877,00	107 933,00	137 381,00	151 823,00	168 050,00	175 109,00
4	R_predpis	1 910,00	5 000,00	9 900,00	22 401,00	40 356,00	56 860,00	98 664,00	201 919,00
4	R_predpis_subst	1 860,00	4 793,00	9 578,50	21 538,50	37 282,00	51 500,00	78 750,00	104 522,00

Legenda: Měřeno nad kombinací dat České pojišťovny a Registru ekonomických subjektů k 30.6.2006.

V tabulce 16 jsou dále pro jednotlivé shluky uvedeny popisné statistiky proměnných Stáří firmy v letech po substituci (Stari_subst) a Staří firmy v letech (Stari). Jak je patrné z této tabulky, mezi shluky č. 1, 2 a 3 nejsou výrazné rozdíly. Průměrné hodnoty stáří se u těchto shluků pohybují v rozmezí 13 až 16 let (viz. proměnná Stari_subst). Výrazně mladší jsou firmy ze shluku č. 4 s průměrným stářím 6 let. Při použití kompletních a netransformovaných dat nedojde u žádného ze shluků k výrazné změně hodnot popisných statistik (viz. proměnná Stari).

Tab. 16 Popisné statistiky pro proměnné Stáří firmy v letech po substituci (Stari_subst) a Stáří firmy v letech (Stari)

Shluk	Proměnná	N	Avg	Std	Skew	Kurt	Min	P1	P5
		Počet neprázdných hodnot	Aritmetický průměr	Směrodatná odchylka	Šikmost	Špičatost	Minimum	1. percentil	5. percentil
1	Stari	49 284	14	3,16	4,91	42,76	10	10	10
1	Stari_subst	47 655	13	1,76	-0,39	-0,97	10	10	10
2	Stari	12 648	15	1,97	0,73	32,08	1	7	12
2	Stari_subst	11 707	16	0,89	-2,98	11,57	10	12	14
3	Stari	9 970	13	4,58	2,40	13,44	1	3	7
3	Stari_subst	5 496	13	2,10	-0,59	-0,59	7	8	9
4	Stari	26 054	6	2,33	-0,31	-1,09	1	1	2
4	Stari_subst	24 898	6	2,28	-0,28	-1,19	2	2	2

Shluk	Proměnná	P10	Q1	Median	Q3	P90	P95	P99	Max
		10.percentil	Spodní kvartil	Medián	Horní kvartil	90. percentil	95.percentil	99.percentil	Maximum
1	Stari	11	12	14	15	15	16	33	61
1	Stari_subst	11	12	14	15	15	16	16	19
2	Stari	14	15	16	16	16	16	17	33
2	Stari_subst	15	16	16	16	16	16	16	21
3	Stari	9	11	13	15	16	16	33	61
3	Stari_subst	10	12	14	15	15	16	16	18
4	Stari	3	4	6	8	9	9	9	9
4	Stari_subst	3	4	6	8	9	9	9	9

Legenda: Měřeno nad kombinací dat České pojišťovny a Registru ekonomických subjektů k 30.6.2006.

V tabulce 17 jsou dále pro jednotlivé shluky uvedeny popisné statistiky proměnné Index storen pro neplacení (Pomer). Pro tuto proměnnou nebyla v kapitole 4.1.2.1 provedena substituce odlehlých hodnot. Popisné statistiky jsou tedy vypočteny vždy nad netransformovanými daty. Avšak vzhledem k tomu, že v prvním kroku hlavní analýzy nebyly všechny firmy zařazeny do shluků, jsou popisné statistiky vypočteny nad nekompletními a kompletními daty. Proměnná Pomer, jejíž hodnoty se pohybují od 0 do 1, bude použita pro hodnocení platební morálky klientů. Čím více se její hodnota blíží 1, tím větší je podíl předpisového pojistného ze smluv stornovaných pro neplacení na celkovém pojistném a tím horší je i platební morálka klienta.

Z tabulky 17 je zřejmé, že firmy ze shluků č. 1, 3 a 4 nemají problémy s placením pojistného. Hodnoty proměnné Pomer - nekompletní data jsou pro tyto firmy rovny 0 nebo jsou 0 blízké. Naopak pro shluk č. 2 jsou hodnoty proměnné Pomer – nekompletní data blízko 1. Firmy z tohoto shluku tedy mají špatnou platební morálku. Při použití kompletních dat nedojde u žádného ze shluků k výrazné změně hodnot popisných statistik (viz. proměnná Pomer - kompletní data).

Tab. 17 Popisné statistiky pro proměnnou Index storen pro neplacení (Pomer)

Shluk	Proměnná	N	Avg	Std	Skew	Kurt	Min	P1	P5
		Počet neprázdných hodnot	Aritmetický průměr	Směrodatná odchylka	Šikmost	Špičatost	Minimum	1.percentil	5. percentil
1	Pomer - kompletní data	50 729	0,0104	0,0497	5,6311	32,9585	0	0	0
1	Pomer - nekompletní data	49 100	0,0075	0,0385	6,1712	40,3566	0	0	0
2	Pomer - kompletní data	12 733	0,7864	0,1140	-1,3505	2,6052	0,2520	0,4089	0,5000
2	Pomer - nekompletní data	11 720	0,8004	0,0918	-0,9028	1,9260	0,4288	0,4890	0,6908
3	Pomer - kompletní data	10 072	0,0404	0,1025	4,1970	22,9729	0	0	0
3	Pomer - nekompletní data	5 544	0,0234	0,0581	2,9661	8,7911	0	0	0
4	Pomer - kompletní data	26 466	0,0270	0,0825	3,9168	17,2961	0	0	0
4	Pomer - nekompletní data	25 208	0,0170	0,0525	3,4221	11,9289	0	0	0

Shluk	Proměnná	P10	Q1	Median	Q3	P90	P95	P99	Max
		10. percentil	Spodní kvartil	Medián	Horní kvartil	90. percentil	95. percentil	99. percentil	Maximum
1	Pomer - kompletní data	0	0	0	0	0	0,0570	0,3150	0,4434
1	Pomer - nekompletní data	0	0	0	0	0	0,0213	0,2490	0,3803
2	Pomer - kompletní data	0,7123	0,7237	0,8015	0,8570	0,9213	0,9213	0,9357	0,9972
2	Pomer - nekompletní data	0,7237	0,7523	0,8015	0,8570	0,9213	0,9213	0,9289	0,9972
3	Pomer - kompletní data	0	0	0	0,0213	0,1327	0,2465	0,4205	0,9213
3	Pomer - nekompletní data	0	0	0	0	0,0822	0,1601	0,2667	0,3764
4	Pomer - kompletní data	0	0	0	0	0,0942	0,1846	0,4421	0,7474
4	Pomer - nekompletní data	0	0	0	0	0,0725	0,1514	0,2566	0,3889

Legenda: Měřeno nad kombinací dat České pojišťovny a Registru ekonomických subjektů k 30.6.2006.

5.2 Výsledky analýzy kategoriálních dat

V rámci analýzy kategoriálních dat byly zkoumány vztahy mezi novou proměnnou Shluk (odvozenou během shlukové analýzy) a dalšími kategoriálními proměnnými z ploché tabulky. Za tímto účelem byly testované proměnné uspořádány do kontingenčních tabulek.

Kontingenční tabulka pro proměnné Právní forma (Pravfor) a Shluk je uvedena v tabulce 18 a zároveň také v příloze 3 (pozn.: v tabulce 18 nejsou uvedeny teoretické četnosti a procentická vyjádření skutečných četností na celkovém počtu testovaných firem).

Tab. 18 Kontingenční tabulka pro proměnné Právní forma (Pravfor) a Shluk

Pravfor	**Shluk**				
Skutečná četnost Procento řádku Procento sloupce	1	2	3	4	Součet
Společnost s ručením omezeným	6 705	240	4 239	8 620	19 804
	33,86	1,21	21,40	43,53	
	13,60	1,90	42,52	33,09	
Akciová společnost	351	19	1 235	721	2 326
	15,09	0,82	53,10	31,00	
	0,71	0,15	12,39	2,77	
Podnikatel - fyzická osoba	30 193	10 807	2 145	12 162	55 307
	54,59	19,54	3,88	21,99	
	61,26	85,44	21,51	46,68	
Samostatně hospodařící rolník	1 011	554	89	152	1 806
	55,98	30,68	4,93	8,42	
	2,05	4,38	0,89	0,58	
Svobodné povolání	4 936	785	295	1 421	7 437
	66,37	10,56	3,97	19,11	
	10,02	6,21	2,96	5,45	
Družstvo	332	4	273	234	843
	39,38	0,47	32,38	27,76	
	0,67	0,03	2,74	0,90	
Příspěvková organizace	532	1	266	682	1 481
	35,92	0,07	17,96	46,05	
	1,08	0,01	2,67	2,62	
Zahraniční osoba	340	41	34	527	942
	36,09	4,35	3,61	55,94	
	0,69	0,32	0,34	2,02	
Sdružení	799	6	119	261	1 185
	67,43	0,51	10,04	22,03	
	1,62	0,05	1,19	1,00	
Obecní úřad	1 241	56	372	5	1 674
	74,13	3,35	22,22	0,30	
	2,52	0,44	3,73	0,02	
Ostatní	2 844	135	903	1 269	5 151
	55,21	2,62	17,53	24,64	
	5,77	1,07	9,06	4,87	
Součet	49 284	12 648	9 970	26 054	97 956

Legenda: Měřeno nad kombinací dat České pojišťovny a Registru ekonomických subjektů k 30.6.2006.

Jak je patrné z tabulky, největší podíl z celkového počtu firem zaujímají podnikatelé-fyzické osoby (56,46 %), společnosti s ručením omezeným (20,22 %) a svobodná povolání (7,59 %). Nejvíce podnikatelů-fyzických osob obsahuje shluk č. 1 (54,59 %), který následují shluky č. 2 a 4 (každý přibližně s 20-ti %). V rámci shluku č. 2 reprezentují podnikatelé-fyzické osoby nejvíce firem (85,44 %). Společnosti s ručením omezeným jsou nejvíce zastoupeny ve shluku č. 4 (43,53 %). Více než 50 % společností s ručením omezeným je rozděleno také mezi shluky č. 1 a 3, přičemž ve shluku č. 3 mají tyto firmy největší podíl. Svobodná povolání jsou nejvíce zastoupena ve shluku č. 1 (66,37 %). Následuje shluk č. 3 (19,11 %) a shluk č. 2 (10,56 %). Svobodná povolání představují ve shluku č. 2 druhou nejpočetnější skupinu (6,21 %). Spolu s podnikateli-fyzickými osobami tak tvoří více než 90 % shluku. Shluk č. 2 obsahuje oproti ostatním minimum obchodních společností. Naproti tomu ve shluku č. 3 obchodní společnosti převažují. Podnikatelé-fyzické osoby a svobodná povolání převažují také ve shluku č. 1.

V tabulce 19 je dále uvedena kontingenční tabulka pro proměnné Institucionální sektor (Isektor) a Shluk. Tato tabulka je uvedena také v příloze 4, kde obsahuje navíc teoretické četnosti a procentická vyjádření skutečných četností na celkovém počtu testovaných firem. Z institucionálních sektorů zaujímá největší podíl kategorie Domácnosti (67,57 %). Tato skupina zahrnuje jednotlivce nebo skupiny jednotlivců jako konečné spotřebitele a podnikatele, kteří produkují tržní výrobky a finanční i nefinanční služby za předpokladu, že odpovídající činnosti nejsou od těchto subjektů odděleny. Největší podíl těchto firem je ve shluku č. 1 (56,56 %). Následuje shluk č. 4 s 21,03 % a shluk č. 2 s 18-ti %. Shluk č. 2 je přitom tvořen firmami ze sektoru Domácnosti z 94,21 %. Druhým nejpočetnějším sektorem jsou Nefinanční podniky, které reprezentují 26,16 % firem. Jedná se o společnosti, které jsou tržními výrobci a jejichž základní činnost spočívá v produkci výrobků a poskytování nefinančních služeb (tj. služeb mimo peněžnictví a pojišťovnictví). Rozdělovací a finanční transakce jsou přitom odděleny od transakcí jejich vlastníků. Největší podíl těchto firem je ve shluku č. 4 (40,22 %). Následuje shluk č. 1 s 33,37 % a shluk č. 3 s 23,84 %. Ve shluku č. 3 jsou nefinanční podniky zastoupeny z více než 60-ti %.

Tab. 19 Kontingenční tabulka pro proměnné Institucionální sektor (Isektor) a Shluk

Isektor Skutečná četnost Procento řádku Procento sloupce	Shluk 1	2	3	4	Součet
Nefinanční podniky	8 552	657	6 109	10 308	25 626
	33,37	2,56	23,84	40,22	
	17,35	5,19	61,27	39,57	
Finanční instituce	38	0	68	48	154
	24,68	0,00	44,16	31,17	
	0,08	0,00	0,68	0,18	
Vládní instituce	1 776	58	655	690	3 179
	55,87	1,82	20,60	21,70	
	3,60	0,46	6,57	2,65	
Domácnosti	37 434	11 915	2 918	13 922	66 189
	56,56	18,00	4,41	21,03	
	75,96	94,21	29,27	53,44	
Neziskové instituce	1 476	16	216	1 084	2 792
sloužící domácnostem	52,87	0,57	7,74	38,83	
	3,00	0,13	2,17	4,16	
Nerezidenti	6	1	4	0	11
	54,55	9,09	36,36	0,00	
	0,01	0,01	0,04	0,00	
Součet	49 282	12 647	9 970	26 052	97 951

Legenda: Měřeno nad kombinací dat České pojišťovny a Registru ekonomických subjektů k 30.6.2006.

Kontingenční tabulka pro proměnné Hlavní předmět podnikatelské činnosti (Okec_sekce) a Shluk je uvedena v tabulce 20 a zároveň také v příloze 5 (pozn.: v tabulce 20 nejsou uvedeny teoretické četnosti a procentická vyjádření skutečných četností na celkovém počtu testovaných firem). Největší podíl z celkového počtu zaujímá sekce Obchod; opravy motorových vozidel a výrobků pro osobní potřebu a převážně pro domácnost s 22,31 %. Následují sekce Zpracovatelský průmysl a Činnosti v oblasti nemovitostí a pronájmu a podnikatelské činnosti, každá přibližně s 15-ti %. Nejvíce firem z těchto sekcí je ve shluku č. 1 (v rozmezí 45 až 51 %). Druhé nejvyšší zastoupení je pak vždy ve shluku č. 4. Podíly největších sekcí v rámci jednotlivých shluků jsou při mezishlukovém porovnání téměř stejné. Pouze u zpracovatelského průmyslu převyšuje shluk č. 2 mírně ostatní shluky (s 21,51 %). Zároveň firmy zpracovatelského průmyslu tvoří v tomto shluku nejpočetnější skupinu. Podobně jako shluk č. 2 převyšuje ostatní shluky také shluk č. 4 v sekci Činnosti v oblasti nemovitostí a pronájmu a podnikatelské činnosti (s 20,37 %). V tomto případě se však jedná až o druhou nejpočetnější skupinu v rámci shluku.

Tab. 20 Kontingenční tabulka pro proměnné Hlavní předmět podnikatelské činnosti (Okec_sekce) a Shluk

Okec_sekce Skutečná četnost Procento řádku Procento sloupce	**Shluk** 1	2	3	4	Součet
Zemědělství, myslivost, lesnictví	3 361	602	1 293	868	6 124
	54,88	9,83	21,11	14,17	
	6,82	4,76	12,97	3,33	
Rybolov a chov ryb	24	3	18	3	48
	50,00	6,25	37,50	6,25	
	0,05	0,02	0,18	0,01	
Těžba nerostných surovin	18	10	30	12	70
	25,71	14,29	42,86	17,14	
	0,04	0,08	0,30	0,05	
Zpracovatelský průmysl	7 115	2 721	1 638	3 537	15 011
	47,40	18,13	10,91	23,56	
	14,44	21,51	16,43	13,58	
Výroba a rozvod elektřiny,	58	11	75	39	183
plynu a vody	31,69	6,01	40,98	21,31	
	0,12	0,09	0,75	0,15	
Stavebnictví	5 164	2 079	941	3 186	11 370
	45,42	18,28	8,28	28,02	
	10,48	16,44	9,44	12,23	
Obchod; opravy motorových vozidel a	11 125	2 386	1 962	6 381	21 854
výrobků pro osobní potřebu a převážně	50,91	10,92	8,98	29,20	
pro domácnost	22,57	18,86	19,68	24,49	
Ubytování a stravování	3 101	978	195	1 567	5 841
	53,09	16,74	3,34	26,83	
	6,29	7,73	1,96	6,01	
Doprava, skladování a spoje	3 229	869	1 243	1 708	7 049
	45,81	12,33	17,63	24,23	
	6,55	6,87	12,47	6,56	
Finanční zprostředkování	1 055	361	121	917	2 454
	42,99	14,71	4,93	37,37	
	2,14	2,85	1,21	3,52	
Činnosti v oblasti nemovitostí a	7 044	1 826	1 141	5 307	15 318
pronájmu; podnikatelské činnosti	45,99	11,92	7,45	34,65	
	14,29	14,44	11,44	20,37	
Veřejná správa a obrana; povinné	1 322	63	441	31	1 857
sociální zabezpečení	71,19	3,39	23,75	1,67	
	2,68	0,50	4,42	0,12	
Vzdělávání	630	94	131	792	1 647
	38,25	5,71	7,95	48,09	
	1,28	0,74	1,31	3,04	
Zdravotní a sociální péče;	3 096	116	262	502	3 976
veterinární činnosti	77,87	2,92	6,59	12,63	
	6,28	0,92	2,63	1,93	
Ostatní veřejné, sociální a osobní služby	2 935	528	476	1 202	5 141
	57,09	10,27	9,26	23,38	
	5,96	4,17	4,77	4,61	
Exteritoriální organizace a instituce	7	1	3	0	11
	63,64	9,09	27,27	0,00	
	0,01	0,01	0,03	0,00	
Součet	49 284	12 648	9 970	26 052	97 954

Legenda: Měřeno nad kombinací dat České pojišťovny a Registru ekonomických subjektů k 30.6.2006.

V tabulce 21 je dále uvedena kontingenční tabulka pro proměnné Počet pracovníků (Pocprac) a Shluk. Tato tabulka je uvedena také v příloze 6, kde obsahuje navíc teoretické četnosti a procentická vyjádření skutečných četností na celkovém počtu testovaných firem. Největší podíl z celkového počtu zaujímají firmy o velikosti do 5-ti zaměstnanců. 44,10 % reprezentují subjekty bez zaměstnanců a 31,85 % subjekty s 1 až 5-ti zaměstnanci. Nejvíce firem z obou kategorií obsahuje shluk č. 1 (v obou případech více než 50 %). Následuje shluk č. 4 vždy přibližně s 25-ti %. Shluky č. 1, 2 a 4 jsou tvořeny firmami o velikosti do 5-ti zaměstnanců přibližně z 80-ti %. Naopak ve shluku č. 3 jsou ve větší míře zastoupeny střední a velké podniky. Podíl firem s 25-ti a více zaměstnanci činí v tomto shluku téměř 40 % (13,88 % tvoří firmy s 25-ti až 49-ti zaměstnanci, 10,82 % firmy s 50-ti až 99-ti zaměstnanci a 12,33 % firmy se 100 a více zaměstnanci).

Tab. 21 Kontingenční tabulka pro proměnné Počet pracovníků (Pocprac) a Shluk

Pocprac	**Shluk**				
Skutečná četnost Procento řádku Procento sloupce	1	2	3	4	Součet
0	21 451	6 224	711	9 201	37 587
	57,07	16,56	1,89	24,48	
	50,35	55,17	7,32	42,52	
1 - 5	14 307	3 561	1 987	7 290	27 145
	52,71	13,12	7,32	26,86	
	33,58	31,56	20,47	33,69	
6 - 9	2 546	666	1 087	1 861	6 160
	41,33	10,81	17,65	30,21	
	5,98	5,90	11,20	8,60	
10 - 19	2 239	562	1 700	1 760	6 261
	35,76	8,98	27,15	28,11	
	5,25	4,98	17,51	8,13	
20 - 24	476	104	628	403	1 611
	29,55	6,46	38,98	25,02	
	1,12	0,92	6,47	1,86	
25 - 49	852	105	1 347	651	2 955
	28,83	3,55	45,58	22,03	
	2,00	0,93	13,88	3,01	
50 - 99	473	46	1 050	262	1 831
	25,83	2,51	57,35	14,31	
	1,11	0,41	10,82	1,21	
100 a více	264	14	1 197	209	1 684
	15,68	0,83	71,08	12,41	
	0,62	0,12	12,33	0,97	
Součet	42 608	11 282	9 707	21 637	85 234

Legenda: Měřeno nad kombinací dat České pojišťovny a Registru ekonomických subjektů k 30.6.2006.

Kontingenční tabulka pro proměnné Oblast sídla firmy (Oblast) a Shluk je uvedena v tabulce 22 a zároveň také v příloze 7 (pozn.: v tabulce 22 nejsou uvedeny teoretické četnosti a procentická vyjádření skutečných četností na celkovém počtu testovaných firem). Nejvíce subjektů má své sídlo v Praze (16,72 %) a na jihovýchodě a severovýchodě republiky (v obou případech přibližně 15 %). Zastoupení firem v ostatních oblastech se pohybuje vždy okolo 10-ti %. Ve všech oblastech pochází nejvíce firem ze shluku č. 1 a následně ze shluku č. 4. Podíly oblastí v rámci jednotlivých shluků jsou při mezishlukovém porovnání téměř stejné. Pouze na severovýchodě republiky převyšuje shluk č. 2 mírně ostatní shluky (s 18,30 %). Zároveň firmy z oblasti severovýchodu tvoří v tomto shluku nejpočetnější skupinu. Shluk č. 2 má oproti ostatním menší zastoupení v Praze (12,02 %). Obecně lze rozdíly mezi shluky v jednotlivých oblastech charakterizovat jako malé. Malá těsnost závislosti je zřejmá také z nízké hodnoty koeficientu kontingence (viz. tabulka 13).

Tab. 22 Kontingenční tabulka pro proměnné Oblast sídla firmy (Oblast) a Shluk

Oblast	**Shluk**				
Skutečná četnost					
Procento řádku					
Procento sloupce	1	2	3	4	Součet
Praha	8 225	1 520	1 915	4 715	16 375
	50,23	9,28	11,69	28,79	
	16,69	12,02	19,21	18,10	
Střední Čechy	6 515	1 749	1 292	3 069	12 625
	51,60	13,85	10,23	24,31	
	13,22	13,83	12,96	11,78	
Jihozápad	5 214	1 318	1 058	2 639	10 229
	50,97	12,88	10,34	25,80	
	10,58	10,42	10,61	10,13	
Severozápad	4 386	1 332	914	2 462	9 094
	48,23	14,65	10,05	27,07	
	8,90	10,53	9,17	9,45	
Severovýchod	7 008	2 315	1 319	3 376	14 018
	49,99	16,51	9,41	24,08	
	14,22	18,30	13,23	12,96	
Jihovýchod	7 450	1 795	1 508	4 036	14 789
	50,38	12,14	10,20	27,29	
	15,12	14,19	15,13	15,49	
Střední Morava	5 564	1 482	1 023	2 837	10 906
	51,02	13,59	9,38	26,01	
	11,29	11,72	10,26	10,89	
Ostravsko	4 922	1 137	941	2 918	9 918
	49,63	11,46	9,49	29,42	
	9,99	8,99	9,44	11,20	
Součet	49 284	12 648	9 970	26 052	97 954

Legenda: Měřeno nad kombinací dat České pojišťovny a Registru ekonomických subjektů k 30.6.2006.

V tabulce 23 je dále uvedena kontingenční tabulka pro proměnné Status podnikatelské činnosti firmy dle externích dat (Aktiv) a Shluk. Tato tabulka je uvedena také v příloze 8, kde obsahuje navíc teoretické četnosti a procentická zastoupení skutečných četností na celkovém počtu testovaných firem. 80,99 % subjektů aktivně provozuje podnikatelskou činnost. 10,70 % firem již dle externích dat zaniklo a 8,30 % byla pozastavena podnikatelská činnost. Převážná většina firem, které zanikly, pochází ze shluku č. 2 (92,40 %). Zároveň tyto firmy tvoří ve druhém shluku nepočetnější skupinu. Zbytek kategorií obsahuje nejvíce firem ze shluků č. 1 a 4. Při mezishlukovém porovnání podílů firem s pozastavenou podnikatelskou činností v rámci shluků, převyšují shluky č. 1 a 4 ostatní (shluk č. 1 s 8,94 % a shluk č. 4 s 12,65 %). Ve shlucích č. 1, 3 a 4 je vysoký podíl firem, které aktivně provozují podnikatelskou činnost.

Tab. 23 Kontingenční tabulka pro proměnné Status podnikatelské činnosti firmy dle externích dat (Aktiv) a Shluk

Aktiv Skutečná četnost Procento řádku Procento sloupce	**Shluk** 1	2	3	4	Součet
Firmě byla pozastavena podnikatelská činnost	4 407	220	213	3 295	8 135
	54,17	2,70	2,62	40,50	
	8,94	1,74	2,14	12,65	
Firma podnikatelskou činnost aktivně provozuje	44 234	2 743	9 604	22 758	79 339
	55,75	3,46	12,11	28,68	
	89,75	21,69	96,33	87,35	
Firma zanikla	643	9 685	153	1	10 482
	6,13	92,40	1,46	0,01	
	1,30	76,57	1,53	0,00	
Součet	49 284	12 648	9 970	26 054	97 956

Legenda: Měřeno nad kombinací dat České pojišťovny a Registru ekonomických subjektů k 30.6.2006.

5.3 Profilace segmentů a strategie přístupu k nim

5.3.1 Profily segmentů

V rámci statistické analýzy byli hodnoceni firemní klienti České pojišťovny za účelem jejich segmentace. Výsledkem jsou čtyři skupiny klientů, jejichž obchodní profily jsou popsány v této kapitole. Pro posouzení správnosti výsledků profilace se dále doporučuje vybrat z jednotlivých segmentů vzorky zkoumaných jednotek (v našem případě firem) a prozkoumat, zda jsou pravidla zjištěná v rámci statistické analýzy

v souladu se zdrojovými daty. Proto byla v ploché tabulce zachována identifikace firem pomocí jejich IČ, obchodního jména a adresy (viz. tabulka 7).

5.3.1.1 Segment č. 1

Segment č. 1 obsahuje firmy, které byly v rámci shlukové analýzy zařazeny do prvního shluku. Tyto firmy lze charakterizovat následujícími vlastnostmi:

1. Převážná většina z nich nevytváří základní jmění.
2. Průměrná výše očekávaného ročního předpisového pojistného činí přibližně 13 000 Kč.
3. Průměrné stáří firem je 13 let.
4. Firmy mají velmi dobrou platební morálku.
5. Jedná se o malé firmy do 5-ti zaměstnanců, převážně podnikatele-fyzické osoby a svobodná povolání.
6. Společně s firmami ze segmentu č. 4 mají oproti ostatním zvýšenou tendenci k pozastavení podnikatelské činnosti.

5.3.1.2 Segment č. 2

Segment č. 2 obsahuje firmy, které byly v rámci shlukové analýzy zařazeny do druhého shluku a které lze charakterizovat následovně:

1. Firmy nevytváří základní jmění.
2. Průměrná výše očekávaného ročního předpisového pojistného činí přibližně 14 000 Kč.
3. Průměrné stáří firem je 16 let.
4. Firmy mají špatnou platební morálku. Jejich smlouvy často končí stornem pro neplacení.
5. Jedná se o podnikatele-fyzické osoby a svobodná povolání.
6. Z institucionálních sektorů převládá kategorie Domácnosti, pro kterou je charakteristické, že zahrnuje jednotlivce nebo skupiny jednotlivců, kteří jsou podnikateli i konečnými spotřebiteli a kteří produkují tržní výrobky a finanční i nefinanční služby.

7. Jedná se o malé firmy do 5-ti zaměstnanců s vyšším podílem zpracovatelského průmyslu.
8. Oproti ostatním segmentům mají tyto firmy menší zastoupení v Praze a zároveň větší zastoupeni na severovýchodě republiky (tj. v Libereckém, Královehradeckém a Pardubickém kraji).
9. Převážná většina firem má dle externích dat ukončenou podnikatelskou činnost.

5.3.1.3 Segment č. 3

Segment č. 3 obsahuje firmy, které byly v rámci shlukové analýzy zařazeny do třetího shluku. Tyto firmy lze charakterizovat následujícími vlastnostmi:

1. Oproti segmentům č. 1 a 2 jsou ve větší míře zastoupeny firmy, které vytváří základní jmění. Jeho výše v tomto případě vypovídá o podnikatelské činnosti většího rozsahu.
2. Oproti ostatním segmentům se jedná o subjekty s výrazně vyšším očekávaným ročním předpisovým pojistným (v průměru okolo 88 000 Kč).
3. Průměrné stáří firem je 13 let.
4. Firmy mají velmi dobrou platební morálku.
5. Jedná se převážně o obchodní společnosti.
6. Z hlediska institucionálních sektorů převládají nefinanční podniky, pro které je charakteristická produkce tržních výrobků a poskytování nefinančních služeb. Rozdělovací a finanční transakce jsou přitom odděleny od jejich vlastníků.
7. Podle počtu zaměstnanců jsou oproti ostatním segmentům zastoupeny ve větší míře středně velké a velké podniky.

5.3.1.4 Segment č. 4

Segment č. 4 obsahuje firmy, které byly v rámci shlukové analýzy zařazeny do čtvrtého shluku a které lze charakterizovat následovně:

1. Oproti segmentům č. 1 a 2 jsou ve větší míře zastoupeny firmy, které vytváří základní jmění. Ve srovnání se segmentem č. 3 se však jedná o subjekty s menším rozsahem podnikatelské činnosti.

2. Průměrná výše očekávaného ročního předpisového pojistného činí přibližně 16 000 Kč.
3. Oproti ostatním segmentům se jedná o výrazně mladší firmy s průměrným stářím 6 let.
4. Firmy mají velmi dobrou platební morálku.
5. Jedná se o malé firmy do 5-ti zaměstnanců, většinou podnikatele-fyzické osoby nebo společnosti s ručením omezeným.
6. Společně s firmami ze segmentu č. 1 mají oproti ostatním zvýšenou tendenci k pozastavení podnikatelské činnosti.

5.3.2 Strategie přístupu k segmentům

Na základě obchodní profilace lze považovat segment č. 3 za vysoce bonitní. Obsahuje zavedené firmy, převážně obchodní společnosti, které platí pojišťovně vysoké pojistné. Zároveň tyto firmy mají velmi dobrou platební morálku. Segment č. 3 je však nejméně početný. České pojišťovně lze tedy ve vztahu k němu doporučit následující strategii:

- Udržet stávající klienty a postupně segment zvětšovat.

Naopak jako nejméně bonitní se jeví segment č. 2, který obsahuje podnikatele-fyzické osoby a svobodná povolání. Firmy zařazené do tohoto segmentu mají špatnou platební morálku a jejich pojistné smlouvy často končí stornem pro neplacení. Tato skutečnost není vyvážena ani výší plateb pojistného. Dle externích dat má převážná většina firem z tohoto segmentu ukončenou podnikatelskou činnost, což může mít souvislost s jejich špatnou platební morálkou. České pojišťovně lze tedy ve vztahu k tomuto segmentu doporučit strategii:

- Klientům nenabízet žádná zvýhodnění a v případě, že budou chtít pojišťovnu opustit, nechat je odejít.

Segment č. 1 je nejpočetnější. Obsahuje zavedené firmy, převážně podnikatele-fyzické osoby, které mají velmi dobrou platební morálku. Výše plateb pojistného je srovnatelná se segmenty č. 2 a 4. Firmy z tohoto segmentu mají zvýšenou tendenci k pozastavení podnikatelské činnosti. České pojišťovně lze tedy doporučit následující strategii:

- Firmy podporovat za účelem zvýšení výnosnosti segmentu. Zároveň monitorovat firmy s pozastavenou činností, aby se nestaly členy segmentu č. 2.

Segment č. 4 obsahuje mladší firmy, podnikatele-fyzické osoby a společnosti s ručením omezeným, s velmi dobrou platební morálkou. Výše plateb pojistného je srovnatelná se segmenty č. 1 a 2. Podobně jako u segmentu č. 1 mají i zde firmy zvýšenou tendenci k pozastavení podnikatelské činnosti. České pojišťovně lze tedy doporučit strategii:

- Firmy ze segmentu podporovat tak, aby se postupně staly členy segmentu č. 3 nebo č. 1. Zároveň monitorovat firmy s pozastavenou činností, aby se nestaly členy segmentu č. 2.

Vzhledem k tomu, že ve fázi přípravy statistické analýzy byl ze zdrojových dat vybrán pouze vzorek firem, který byl dále modifikován, nelze výše uvedené závěry považovat za skutečný obraz obchodní situace v České pojišťovně. Výběr pouze vzorku dat a jejich modifikace je podmínkou České pojišťovny pro jejich použití v této disertační práci.

6 Diskuse

Jak již bylo uvedeno dříve, výsledky data miningových analýz jsou kriticky závislé na kvalitě vstupních dat. Parr Rud [29] v této souvislosti potvrzuje platnost známého tvrzení „Garbage in, Garbage out!“ (Odpadky na vstupu = brak na výstupu). Schopnost data miningové úlohy produkovat užitečné informace tak závisí na kvalitních datech stejně jako na výběru správně metody pro jejich analýzu.

6.1 Metodologie TQdM

Řízení kvality firemních dat zahrnuje procesy, které jsou obecné a které podporují nejen data miningové úlohy. English [10] v této souvislosti doporučuje implemntovat ve firmách principy metodologie TQdM (Total Quality data Management), kterými jsou:

1. Všeobecné uznání, že každý zaměstnanec firmy je závislý na informacích od svých spolupracovníků.
2. Přesvědčení, že kvalitní informace poskytují firmě hodnotu. Tzn., že jí umožňují provádět její procesy pořádně a optimálně.
3. Víra ve spokojenost zákazníků, kterým jsou poskytovány správné a přesné informace.
4. Podniková kultura, ve které každý zodpovídá za zdokonalování procesů, jež podporují spokojenost zákazníků a vedou ke snížení nákladů.

Metodologie TQdM vede ke zvyšování úrovně dvou typů podnikových procesů:

1. Obchodních a výrobních, v rámci kterých jsou data vytvářena, aktualizována nebo mazána a dále dolovány, distribuovány nebo prezentovány informace jejich uživatelům.
2. Procesů vývoje informačních systémů, v rámci kterých jsou definovány informace, vyvíjeny a implementovány obchodní procesy, informační systémy, informační architektura a databáze.

Implementace metodologie TQdM spočívá v realizaci následujících kroků:

1. *Posouzení datových definic a informační architektury.* V tomto kroku jsou definovány základní charakteristiky dat a charakteristiky informační

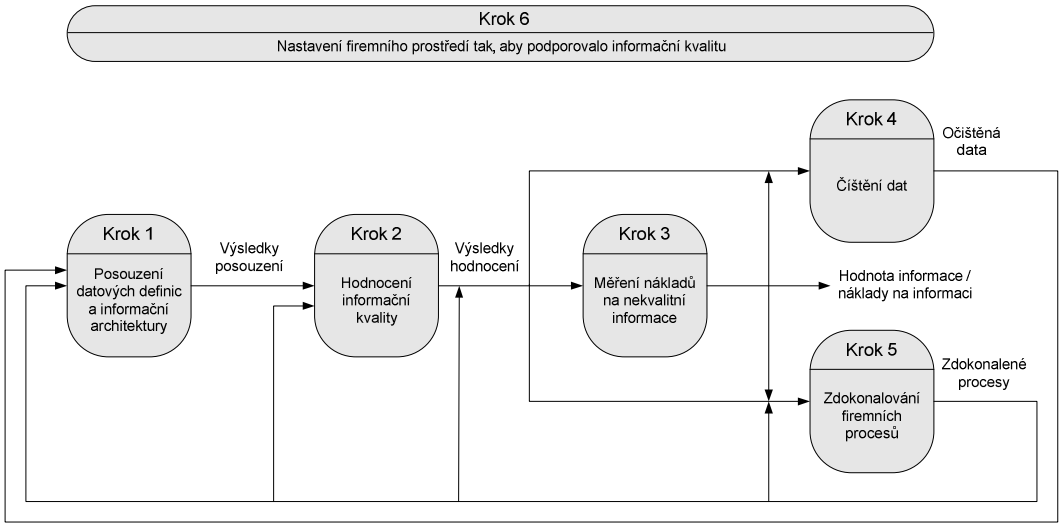
architektury. Jsou identifikovány důležité skupiny informací, jejichž nedostatečná kvalita vede ve firmě k vysokým nákladům nebo má jiné nežádoucí důsledky. Pro každou skupinu jsou také identifikováni klíčoví uživatelé. Dále je posuzována shoda datových definic s datovými standardy a směrnicemi. Informační architektura je srovnávána s nejlepšími podobnými aplikacemi. Nejdůležitější částí je však měření spokojenosti zákazníků s definicemi informačních produktů (tj. s definicemi tvaru informací a způsobů jejich zpřístupňování).

2. *Hodnocení informační kvality.* V tomto kroku jsou identifikovány nebo potvrzeny skupiny informací, jejichž kvalita bude hodnocena. Dále jsou stanoveny charakteristiky, které budou v rámci hodnocení sledovány (např. přesnost, úplnost, shoda s business rules apod.). Pro každou skupinu jsou zdokumentovány všechny procesy a aplikace, které vytvářejí, aktualizují, transformují nebo předávají data, a databáze, ve kterých jsou data uchovávána. Následuje identifikace zdrojů validačních dat, se kterými budou porovnána data z hodnocených procesů a databází. Dále je provedeno samotné hodnocení nad statistickými vzorky a prezentace výsledků.
3. *Měření nákladů na nekvalitní informace.* Tento krok začíná výběrem ukazatelů, na základě kterých je měřen výkon firmy a proti kterým budou náklady na nekvalitní informace postaveny. Jedná se např. o zisk firmy, spokojenost zákazníků apod. Následuje analýza nákladů. V rámci ní je zjišťován stupeň redundance uchovávaných dat a náklady spojené s jejich nedostupností a nepřesností. Dále je prováděna klientská segmentace a výpočet hodnoty životního cyklu zákazníka. Výsledek je použit k vyčíslení hodnoty nevzniklých nebo ztracených příležitostí z důvodu nekvalitních informací. Na základě těchto podkladů je možné stanovit hodnotu kvalitních informací.
4. *Čištění dat.* Tento krok začíná identifikací zdrojů, jejichž data vyžadují opravu. Následuje analýza za účelem objevení různých anomálií a vztahů v datech tak, aby mohla být provedena jejich standardizace ve všech databázích, které uchovávají stejnou informaci. Dále jsou opravovány chybné hodnoty (včetně doplňování chybějících) a duplicity. Mezi datovými chybami jsou zkoumány vztahy za účelem zlepšení procesů, jež mají vliv na informační kvalitu.

Výše uvedené činnosti jsou obecné pro čištění a korekci zdrojových dat. Při přenosu dat ze zdrojových systémů do datového skladu nebo jejich konverzi do nových systémů je dále prováděno jejich mapování do cílové struktury, odvození nových proměnných, sumarizace a kontrola, zda všechny procesy při přenosu proběhly korektně.

5. *Zdokonalování firemních procesů za účelem zlepšení informační kvality.* V tomto kroku jsou nejprve definovány problémy, které je nutné řešit, včetně souvisejících procesů. Následuje ustanovení řešitelského týmu a vývoj plánu řešení. Dále jsou pro účely testů implementovány jednotlivé změny. Po vyhodnocení úspěšnosti je buď provedena jejich redefinice, nebo implementace v celé společnosti.
6. *Nastavení firemního prostředí tak, aby podporovalo informační kvalitu.* Tento krok představuje systémové, organizační a kulturní změny, které z dlouhodobého hlediska podporují zlepšování informační kvality. Její úroveň je ovlivňována podnikovou kulturou a prostředím. Změna obou je důležitá pro odstranění bariér, kterými jsou např. skutečnost, že firmy upřednostňují rychlost provádění procesů nad kvalitou informací, nebo skutečnost, že náklady způsobené nekvalitními informacemi nejsou považovány za zbytečné plýtvání prostředků. Krok 6 vytváří rámec pro všechny předchozí kroky.

Jednotlivé kroky metodologie TQdM a jejich vzájemné vztahy jsou uvedeny na obrázku 24.



Obr. 24 Kroky metodologie TQdM

6.2 Kontroly a vyhodnocení kvality dat v datovém skladu

V této kapitole je uveden funkční návrh systému kontrol a vyhodnocení kvality dat v datovém skladu. Ačkoliv se nejedná o kompletní řešení dle metodologie TQdM, implementace jejích principů je důležitým předpokladem pro fungování systému z dlouhodobého hlediska. Předpokládá se totiž komunikace napříč celou společností. Jestliže některý z útvarů nepřijme tyto principy (např. princip, že každý zaměstnanec je závislý na informacích od svých spolupracovníků), nemůže systém dlouhodobě fungovat a rozvíjet se.

Systém kontrol a vyhodnocení kvality dat bude zaměřen na následující oblasti:

1. Na kontroly kvality dat při jejich přenosu ze zdrojových systémů do datového skladu. V této souvislosti bude testována správnost a úplnost technických parametrů vstupních souborů, obsahová správnost zdrojových dat a správnost transformací do cílových datových struktur.
2. Na vyhodnocení kvality dat a návazné procesy komunikace datového skladu směrem ke zdrojovým systémům a směrem k obchodním uživatelům dat a metodikům.

Systém bude sledovat následující cíle:

1. Postupně zvyšovat úroveň kvality dat v datovém skladu.
2. V maximální možné míře identifikovat chyby dat před jejich zpřístupněním obchodním uživatelům. Tzn. zachytit chyby ještě před základní vrstvou datového skladu nebo v ní.
3. Nastavit procesy oprav chybných dat a informovat obchodní uživatele o chybách a zpožděních, které z oprav vyplývají.

V následujícím textu budou popsány jednotlivé činnosti, které budou v systému realizovány za účelem dosažení těchto cílů.

6.2.1 Správa metadat

Systém kontrol a vyhodnocení kvality dat bude řízen pomocí metadat, jež budou spravována v nástroji, který umožní jejich zadávání a editaci (tzn., že pomocí metadat budou definovány jednotlivé typy kontrol a prováděno vyhodnocení jejich výsledků).

6.2.2 Kontroly technických parametrů vstupních datových souborů

V rámci této činnosti budou kontrolovány technické parametry vstupních datových souborů před jejich načtením do staging area. Důvody pro provedení těchto kontrol jsou následující:

1. Chybu není možné identifikovat po načtení dat do staging area.
2. Chybu je možné identifikovat ve staging area, ale je výhodnější ji odhalit ještě před načtením (např. z časových důvodů).

V rámci tohoto kroku budou prováděny např. tyto kontroly:

1. Porovnání počtu vět vstupního souboru v aktuální dávce dat s počtem vět v předchozí dávce.
2. Porovnání počtu vět vstupního souboru s počtem vět deklarovaným zdrojovým systémem v průvodce.
3. U souborů, které obsahují věty s pevnou délkou, porovnat maximální a minimální délky vět.
4. U souborů, které obsahují věty s oddělovači, zjistit, zda všechny mají stejný počet oddělovačů.

Výše uvedené kontroly předpokládají přenos dat ze zdrojových systémů v textových souborech (ve formátu txt). Přenos však může být proveden i v jiném formátu (např. xml) nebo přímým čtením z provozních databází. Podle způsobu přenosu dat a jejich formátu je vhodné nastavit i kontroly z této skupiny.

Výsledky kontrol technických parametrů budou ukládány do logové tabulky, která bude vždy obsahovat informaci o typu kontroly, název testovaného souboru a výsledek.

6.2.3 Vyhodnocení výsledků kontrol technických parametrů

V rámci této činnosti bude vyhodnocena logová tabulka s výsledky kontrol technických parametrů. Na základě hodnocení bude testované dávce dat přiřazen status a spuštěny procesy uvedené v tabulce 24.

Tab. 24 Status a návazné procesy pro kontroly technických parametrů

Status výsledků kontrol		Následný proces
OK Žádný ze souborů dané dávky neobsahuje definovanou chybu.	1.	Načíst všechny soubory dané dávky do staging area.
ERROR Alespoň jeden ze souborů dané dávky obsahuje definovanou chybu.	1.	Vrátit celou dávku dat zdrojovému systému k opravě. Zároveň zdrojovému systému zpřístupnit informace o chybách a o souborech, ve kterých se tyto chyby vyskytly.
	2.	Informovat uživatele dat o identifikovaných chybách a o předpokládaném termínu zpřístupnění dat opravné dávky. Distribuce hlášení uživatelům bude prováděna automatizovaně pomocí metadat (např. reportem v uživatelské aplikaci, emailem, sms apod.).

6.2.4 Kontroly obsahu zdrojových dat

Kontroly obsahu zdrojových dat budou prováděny v rámci načítání vstupních datových souborů do staging area. V tomto kroku budou prováděny např. tyto kontroly:

1. Kontroly obsahu jednotlivých polí testovaných tabulek:
    a. Kontroly typů proměnných.
    b. Kontroly, zda hodnoty proměnných vyhovují předepsané vstupní masce.
    c. Kontroly, zda hodnoty proměnných odpovídají předpokládaným číselníkovým hodnotám.
    d. Kontroly, zda hodnoty proměnných nejsou mimo předpokládaný rozsah.
    e. Kontroly naplněnosti proměnných.
2. Kontroly povolených kombinací hodnot proměnných v rámci jedné tabulky (testování vnitřní integrity).
3. Kontroly povolených kombinací hodnot proměnných mezi více tabulkami.
4. Kontroly vazeb mezi tabulkami dle klíčových položek (testování referenční integrity).
5. Kontroly duplicit.

Výsledky kontrol obsahu zdrojových dat budou ukládány do logové tabulky, která bude obsahovat pro daný typ kontroly, danou tabulku/tabulky zdrojového systému a testovanou proměnnou/proměnné klíče řádků, ve kterých byly identifikovány chyby včetně chybných hodnot. Každému řádku v logové tabulce bude přiřazeno bodové hodnocení. Hodnoty bodů budou z oboru reálných čísel a jejich výše bude záviset na typu kontroly, tabulce/tabulkách zdrojového systému, testované proměnné/proměnných

a poměru počtu chybných záznamů k jejich celkovému počtu. Tyto hodnoty nebudou počítány během provádění kontrol, ale budou stanoveny před jejich provedením.

6.2.5 Vyhodnocení výsledků kontrol obsahu zdrojových dat

V rámci vyhodnocení bude pro každou tabulku zdrojového systému vypočten součet bodů za všechny výsledky kontrol. Tento součet bude porovnán se stanovenou mezí, přičemž na základě jejího překročení nebo nepřekročení bude hodnocené dávce přiřazen status a spuštěny procesy, které jsou uvedeny v tabulce 25. Hraniční meze budou nastaveny zvlášť pro každou tabulku.

Tab. 25 Status a návazné procesy pro kontroly obsahu zdrojových dat

Status výsledků kontrol		Následný proces
OK Žádná z tabulek zdrojového systému neobsahuje chybu (tj. součet bodů je pro všechny tabulky dané dávky roven 0).	1.	Všechny tabulky transformovat do finální struktury.
WARNING Alespoň pro jednu z tabulek je součet bodů větší než 0 a zároveň menší než stanovená mez.	1.	Všechny tabulky transformovat do finální struktury.
	2.	Informovat zdrojový systém o identifikovaných chybách a požadovat jejich postupnou opravu v rámci dalších dávek dat. Zároveň zdrojovému systému zpřístupnit klíče vět, u kterých byla identifikována chyba, včetně chybných hodnot. Uživatelé dat budou informováni o chybách typu WARNING v rámci vyhodnocení celkové kvality dat pro návazné úlohy (viz. kapitola 6.2.8; status WARNING). Chyby totiž nemusí mít dopad na všechny uživatele.
ERROR Alespoň pro jednu z tabulek je součet bodů větší nebo roven stanovené mezi.	1.	Vrátit celou dávku dat zdrojovému systému a požadovat okamžitou opravu. Zároveň zdrojovému systému zpřístupnit klíče vět, u kterých byla identifikována chyba, včetně chybných hodnot.
	2.	Informovat uživatele dat o identifikovaných chybách a o předpokládaném termínu zpřístupnění dat opravné dávky. Chyby se týkají všech uživatelů. Distribuce hlášení bude prováděna automatizovaně pomocí metadat (např. reportem v uživatelské aplikaci, emailem, sms apod.).

6.2.6 Kontroly obsahu finálních tabulek

Kontroly obsahu finálních tabulek budou zaměřeny na chyby v transformacích dat. Jejich příčinou mohou být např. metodické chyby, které se nepodařilo identifikovat ve fázi jejich vývoje a testování. V této vrstvě lze rovněž odhalit chyby, kdy zdrojové systémy zavedou novou funkcionalitu, a do datového skladu nejsou předány informace o dopadech do stávajících transformací. Data nové funkcionality přitom mohou bez problémů projít přes vrstvu kontrol zdrojových dat.

V tomto kroku budou prováděny např. tyto kontroly:

1. Kontroly obsahu odvozených proměnných:
    a. Kontroly, zda hodnoty proměnných odpovídají předpokládaným číselníkovým hodnotám.
    b. Kontroly, zda hodnoty proměnných nejsou mimo předpokládaný rozsah.
    c. Kontroly naplněnosti proměnných.
2. Kontroly povolených kombinací hodnot odvozených proměnných v rámci jedné tabulky (testování vnitřní integrity).
3. Kontroly povolených kombinací hodnot odvozených proměnných mezi více tabulkami.
4. Kontroly vazeb mezi tabulkami dle klíčových položek (testování referenční integrity).
5. Kontroly duplicit.

Výsledky kontrol budou ukládány do logové tabulky podle stejné logiky jako u kontrol obsahu zdrojových dat. Každému řádku této tabulky bude opět přiřazeno bodové hodnocení.

6.2.7 Vyhodnocení výsledků kontrol obsahu finálních tabulek

V rámci vyhodnocení bude pro každou z finálních tabulek vypočten součet bodů za všechny kontroly. Tento součet bude opět porovnán se stanovenou mezí, přičemž na základě jejího překročení nebo nepřekročení bude hodnocené tabulce přiřazen status a spuštěny procesy, které jsou uvedeny v tabulce 26. Hraniční meze budou nastaveny zvlášť pro každou finální tabulku.

Tab. 26 Status a návazné procesy pro kontroly obsahu finálních tabulek

Status výsledků kontrol		Následný proces
OK Ve finální tabulce nebyla identifikována žádná chyba (tj. součet bodů je roven 0).	1.	Tabulku načíst do základní vrstvy datového skladu.
WARNING Součet bodů je pro finální tabulku větší než 0 a zároveň menší než stanovená mez.	1.	Tabulku načíst do základní vrstvy datového skladu.
	2.	Informovat pracovníka datového skladu odpovědného za ETL proces o identifikovaných chybách (např. automatizovaně odeslaným emailem). Tento pracovník provede ve spolupráci s metodiky a uživateli dat postupnou revizi transformací. Podkladem budou data z logové tabulky. Uživatelé dat budou informováni o chybách typu WARNING v rámci vyhodnocení celkové kvality dat pro návazné úlohy (viz. kapitola 6.2.8; status WARNING). Chyby totiž nemusí mít dopad na všechny uživatele.

Status výsledků kontrol		Následný proces
ERROR Součet bodů je pro finální tabulku větší nebo roven stanovené mezi.	1.	Tabulku do základní vrstvy datového skladu nenačítat.
	2.	Varovat pracovníka datového skladu odpovědného za ETL proces a informovat jej o identifikovaných chybách (např. automatizovaně odeslaným emailem). Tento pracovník provede ve spolupráci s metodiky a uživateli dat okamžitou revizi transformací. Podkladem budou data z logové tabulky.
	3.	Informovat uživatele dat o identifikovaných chybách a o předpokládaném termínu opravy. Chyby se týkají všech uživatelů. Distribuce hlášení bude provedena automatizovaně pomocí metadat (např. reportem v uživatelské aplikaci, emailem, sms apod.).

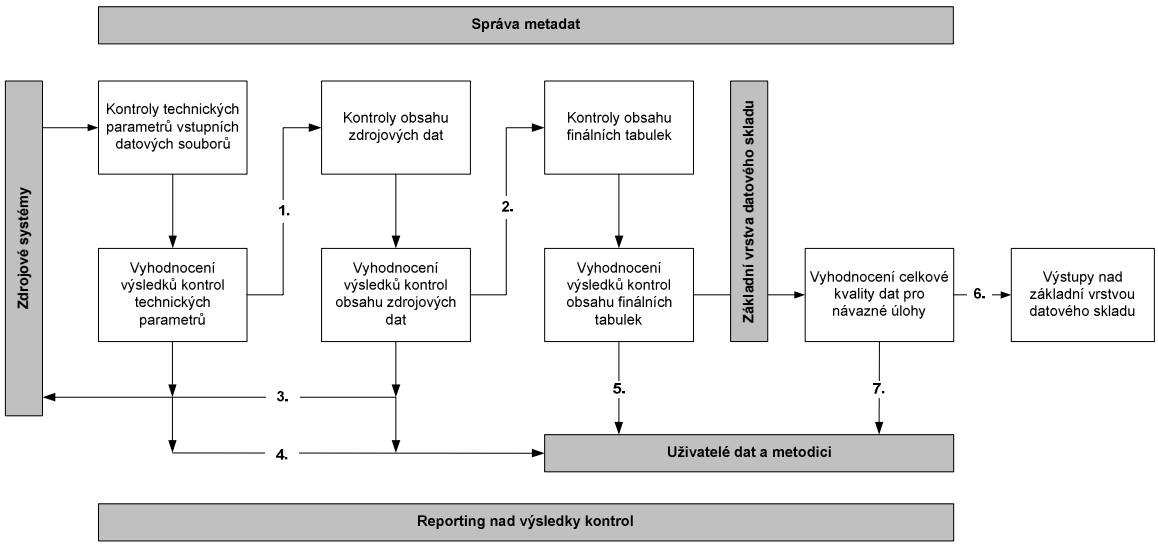
6.2.8 Vyhodnocení celkové kvality dat pro návazné úlohy

V rámci této činnosti bude hodnocena kvalita zdrojových a odvozených dat, jež vstupují do zpracování úloh nad základní vrstvou datového skladu (např. do data martů). Podkladem pro vyhodnocení budou logové tabulky vytvořené v rámci kontrol obsahu zdrojových dat a kontrol finálních tabulek. Pro každou tabulku datového skladu, která vstupuje do zpracování dané úlohy, bude vypočten za všechny používané položky a všechny výsledky kontrol celkový součet bodů. Na základě výsledku budou spuštěny procesy uvedené v tabulce 27. Úlohy budou zpracovány vždy, poněvadž data splňují obecné požadavky na kvalitu, jež byly stanoveny pro základní vrstvu datového skladu.

Tab. 27 Status a návazné procesy pro vyhodnocení celkové kvality dat

Status výsledků kontrol		Následný proces
OK Součet bodů je ve všech tabulkách, které vstupují do zpracování úlohy nad základní vrstvou, roven 0.	1.	Zpracovat danou úlohu.
WARNING Alespoň v jedné z tabulek, které vstupují do zpracování úlohy nad základní vrstvou, je součet bodů větší než 0.	1.	Zpracovat danou úlohu.
	2.	Informovat uživatele dat o chybách, se kterými byla úloha zpracována. Distribuce hlášení uživatelům bude provedena automatizovaně pomocí metadat (např. reportem v uživatelské aplikaci, emailem, sms apod.).

Systém kontrol a vyhodnocení kvality dat datového skladu je dále znázorněn také na obrázku 25. Vedle všech výše popsaných procesů je zde uveden také reporting nad výsledky kontrol pro monitorování kvality dat.



Legenda: 1. Načtení vstupních datových souborů do staging area, 2. Transformace dat do finální struktury, 3. Reklamace chybných dat zdrojovému systému, 4. Informování uživatelů dat o chybách typu ERROR, 5. Revize metodiky transformací, 6. Zpracování úloh nad základní vrstvou (je provedeno vždy – data splňují obecné požadavky na kvalitu pro základní vrstvu datového skladu), 7. Informování uživatelů dat o chybách typu WARNING.

Obr. 25 Systém kontrol a vyhodnocení kvality dat datového skladu

Úspěšná implementace výše uvedeného systému závisí na dobré součinnosti všech útvarů firmy, které jsou do něj zapojeny. Systém nelze považovat za neměnný. Na základě získaných zkušeností je nutné jej permanentně udržovat a rozvíjet.

7 Závěr

Současné tržní prostředí je charakteristické častými změnami a silnou konkurencí v jednotlivých odvětvích. Klíčem k podnikatelskému úspěchu se stávají informace. Za účelem získání většího tržního podílu, zvýšení tržeb nebo snížení nákladů se firmy orientují na správné zákazníky. Snaží se poznat jejich chování, poučit se z něj a připravit se na další vzájemný kontakt. Velkým firmám může při této práci pomoci klientská segmentace, která umožňuje identifikovat bonitní zákazníky, ale i klienty, kteří spíše než k zisku přispívají k vyšším nákladům firmy.

Cíle této disertační práce sledovaly následující úkoly:

1. Ověřit s použitím statistických metod, zda v rozsáhlém souboru firem, klientů České pojišťovny, existují významné skupiny, které vykazují specifické chování.
2. Vytvořit obchodní profily těchto skupin (segmentů) a navrhnout vhodnou strategii dalšího přístupu k nim.
3. Navrhnout algoritmus pro vytvoření jednoznačné identifikace klientů ve vstupních datech a algoritmus pro jejich deduplikaci.
4. Vytvořit funkční návrh obecného sytému kontrol a vyhodnocení kvality dat v datovém skladu.

Při realizaci prvního úkolu se osvědčila kombinace metod shlukové analýzy a analýzy kategoriálních dat. Na základě nehierarchického shlukování byli klienti České pojišťovny rozděleni do čtyřech shluků, přičemž tento rozklad splnil podmínky kritéria optimality. Shluková analýza byla provedena nad spojitými charakteristikami firem a jejím výsledkem je kategoriální proměnná vyjadřující jejich příslušnost ke shlukům. Následovalo provedení analýzy kategoriálních dat. V rámci ní byla testována nezávislost odvozené proměnné a kategoriálních proměnných ze vstupních dat. Ve všech testovaných případech byla prokázána statisticky významná závislost. Použité kategoriální charakteristiky, ale i spojité proměnné, na základě kterých byla provedena shluková analýza, tak mají dobrý potenciál pro formulaci profilů obchodních segmentů.

Tyto profily se podařilo formulovat v rámci druhého úkolu pro všechny čtyři segmenty. Zároveň byly doporučeny strategie dalšího přístupu k nim.

- Segment č. 3 obsahuje vysoce bonitní zákazníky, převážně obchodní společnosti, které pojišťovně platí vysoké pojistné a zároveň mají velmi dobrou platební morálku. Doporučená strategie dalšího přístupu k tomuto segmentu je udržet stávající klienty a snažit se segment rozšířit.
- Segment č. 2 obsahuje zákazníky s nízkou bonitou. Jedná se převážně o podnikatele-fyzické osoby a svobodná povolání. Firmy z tohoto segmentu mají velmi špatnou platební morálku a jejich pojistné smlouvy často končí stornem pro neplacení. Dle externích dat z Registru ekonomických subjektů má většina z nich ukončenou podnikatelskou činnost. Doporučená strategie dalšího přístupu k tomuto segmentu je nenabízet jeho členům žádná zvýhodnění a v případě, že budou chtít pojišťovnu opustit, nechat je odejít.
- Segment č. 1 je nejpočetnější. Obsahuje firmy, převážně podnikatele-fyzické osoby, které mají velmi dobrou platební morálku. Zároveň však tyto firmy mají zvýšenou tendenci k pozastavení podnikatelské činnosti. Doporučenou strategií je podporovat firmy za účelem zvýšení výnosnosti segmentu a monitorovat klienty s pozastavenou činností, aby se nestali členy segmentu č. 2.
- Segment č. 4 obsahuje oproti ostatním mladší firmy, podnikatele-fyzické osoby a společnosti s ručením omezeným. Tyto firmy mají opět velmi dobrou platební morálku. Podobně jako u segmentu č. 1 je i zde zvýšená tendence k pozastavení podnikatelské činnosti. Doporučenou strategií je podporovat firmy tak, aby se postupně staly členy segmentu č. 3 nebo č. 1, a zároveň monitorovat klienty s pozastavenou činností, aby se nestali členy segmentu č. 2.

Použité statistické metody umožnily získat odpovědi na cílové otázky z oblasti klientské segmentace. Identifikace segmentů, vytvoření jejich profilů a návrhy strategií dalšího přístupu k nim tak mohou České pojišťovně pomoci zefektivnit řízení vztahů se zákazníky. Kvalita výsledků segmentace je však vedle použitých statistických metod výrazně ovlivněna také kvalitou vstupních dat. Ve vztahu k zákazníkům je např. velkým problémem nedostatečná identifikace klientů a jejich duplicitní výskyty ve firemních databázích. V této disertační práci bylo použito jako identifikátor klientů identifikační číslo (IČ), se kterým byly spojeny ve vstupních datech následující problémy:

- Výskyty duplicitních záznamů pro jedno IČ.
- IČ bylo nevalidní.
- IČ bylo validní, ale patřilo jinému klientovi.

V rámci třetího úkolu byly proto navrženy algoritmy, které identifikují klientské záznamy věcně a formálně správným IČ a dále provádějí jejich deduplikaci. Ačkoliv tyto algoritmy nezabezpečí stoprocentní vyčištění klientských dat, v této disertační práci výrazně přispěly k eliminaci výše uvedených chyb a zvýšily tak významnost výsledků statistických metod.

Procesy kontroly kvality dat a jejich čištění by měly být realizovány komplexně v rámci ETL procesů datového skladu. Provádění těchto činností v rámci data miningových analýz zvyšuje náročnost a snižuje celkovou efektivitu těchto úloh. Mnohdy jsou tyto činnosti prováděny také duplicitně ve více úlohách. Proto bylo čtvrtým úkolem této disertační práce vytvořit funkční návrh systému kontrol a vyhodnocení kvality dat datového skladu. Při jeho realizaci se podařilo identifikovat klíčové etapy kontrol, navrhnout způsob jejich vyhodnocení a definovat návazné procesy. Mezi klíčové etapy kontrol patří:

- kontroly technických parametrů vstupních datových souborů před jejich nahráním do staging area,
- kontroly obsahu zdrojových dat před jejich transformací do struktury finálních tabulek datového skladu,
- kontroly obsahu finálních tabulek před jejich nahráním do základní vrstvy datového skladu.

Vyhodnocení výsledků kontrol vede ke třem stavům:

- OK. Data neobsahují žádnou chybu.
- WARNING. Data obsahují přípustné chyby.
- ERROR. Data obsahující závažné chyby.

Podle výše uvedených stavů následuje spuštění procesů, jako je např. reklamace dat zdrojovému systému, pokračování ve zpracování dat, informování uživatelů o identifikovaných chybách apod.

S ohledem na výsledky úkolů této disertační práce řešitel doporučuje:

1. Při realizaci úloh zaměřených na segmentaci rozsáhlého souboru statistických jednotek vycházet z nehierarchických metod shlukové analýzy. Dle povahy zkoumaného problému je vhodné kombinovat různé shlukovací algoritmy. Např. v této disertační práci se tento postup osvědčil při realizaci úloh předběžného shlukování a hlavní shlukové analýzy. Shlukovou analýzu je možné dále kombinovat s dalšími statistickými metodami (např. s analýzou kategoriálních dat).
2. Ve všech úlohách věnovat zvýšenou pozornost kvalitě vstupních dat. Při práci se zákazníky, kteří jsou firmami, je možné využít algoritmus pro vytvoření jejich identifikace pomocí věcně a formálně správných IČ a dále algoritmus pro jejich deduplikaci. Tyto algoritmy však nelze považovat za již neměnné. Na základě zkušeností z datové analýzy je nutné je permanentně udržovat a rozvíjet. Zároveň je nutné přesouvat činnosti související s kontrolami a čištěním dat do ETL vrstvy datového skladu.
3. Ve firmách implementovat komplexní systémy pro řízení kvality dat, které umožní řešit nejen důsledky datových defektů, ale i jejich příčiny. V této souvislosti je možné vyjít z navrženého systému kontrol a vyhodnocení kvality dat. Kontrolní mechanismy a na ně navazující procesy totiž podporují nejen výsledky data miningu. Kvalitní data jsou předpokladem úspěchu datových analýz, obchodního reportingu a všech dalších aktivit, které je využívají při podpoře rozhodování. V neposlední řadě pak vedou i k obchodnímu úspěchu celé firmy.

Originální přínos této disertační práce tedy spočívá v návrhu metodických postupů pro segmentaci klientů a pro řízení kvality firemních dat (zejména pro jejich kontrolu, čištění a hodnocení jejich kvality včetně návazných procesů).

8 Seznam odborné literatury

[1] ANDERBERG, M. R. Cluster analysis for aplications. New York: Academic Press, 1973. 359 s.

[2] BERKA, Petr. *Dobývání znalostí z databází*. 1. vyd. Praha: Academia, 2003. 370 s, 1 CD-ROM. ISBN 80-200-1062-9.

[3] BERRY, Michael J. A. – LINOFF, Gordon. *Data mining techniques: for marketing, sales and customer support*. 1. vyd. New York: John Wiley & Sons, Inc., 1997. 454 s. Database management. ISBN 0-471-17980-9.

[4] BERRY, Michael J. A. *Data mining – seven years later, lessons learned*. USA: Business intelligence network, 2004. [cit. 2004-06-27]. Dostupné z: http://www.b-eye-network.com/view/178

[5] BIGUS, J. P. *Data mining with neural networks: solving business problems – from application development to decision support*. New York: McGraw-Hill, 1996. 220 s.

[6] BREIMAN, Leo, aj. *Classification and regression trees*. 1. vyd. Boston: Kluwer Academic Publishers, 1984. 368 s. ISBN 0-412-04841-8.

[7] ČESKÝ STATISTICKÝ ÚŘAD. *Přehled vybraných číselníků z katalogu METIS ČSÚ*. 2006, poslední aktualizace 2006-05-06 [cit. 2006-06-06]. Dostupné z: http://dw.czso.cz/pls/metis/TUCC_zk.KATAL?obr=1024768

[8] DASU, Tamraparni – JOHNSON, Theodore. *Exploratory data mining and data cleaning*. 1. vyd. New York: John Wiley & Sons, Inc., 2003. 228 s. Wiley series in probability and statistics. ISBN 0-471-26851-8.

[9] DOUG, Alexander. *Data mining*. 1997, poslední aktualizace 1997-07-14 [cit. 2006-08-16]. Dostupné z: http://www.eco.utexas.edu/~norman/BUS.FOR/course.mat/Alex/

[10] ENGLISH, Larry P. *Improving data warehouse and business information quality: methods for reducing costs and increasing profits*. 1 vyd. New York: John Wiley & Sons, Inc., 1999. 544 s. Database, data warehousing. ISBN 0-471-25383-9.

[11] ESTIVIL-CASTRO, Vladimir – BRANKOVIC, Ljiljana – DOWE, David. L. *Privacy in data mining*. 1999, poslední aktualizace 2001-05-12 [cit. 2006-05-16]. Dostupné z: http://www.austlii.edu.au/au/journals/PLPR/1999/44.html#Heading6.

[12] FAYYAD, U., aj. *Advances in knowledge discovery and data mining*. Cambridge, Massachusetts: AAAI Press / MIT Press, 1996. 560 s. ISBN 0-262-56097-6.

[13] FREITICHOVÁ, Jarmila. Nejobdivovanější firmy České republiky. In Business World [online]. 2004-03-18. [cit. 2006-06-24]. Stalo se (starší). Dostupné z: http://www.businessworld.cz/bw.nsf/stalo_se/F86490CD584FCBB2C1256E8500362232?OpenDocument&cast=1

[14] HEBÁK, Petr – HUSTOPECKÝ, Jiří. *Vícerozměrné statistické metody s aplikacemi*. 1. vyd. Praha: SNTL, 1987. 456 s.

[15] HRON, Jan – TICHÁ, Ivana – DOHNAL, Jan. *Strategické řízení*. 3. vyd. Praha: Credit, 2000. 274 s. Skriptum. ISBN 80-213-0625-4.

[16] HUMPHRIES, Mark – HAWKINS, Michael W. – DY, Michelle C. *Data warehousing: návrh a implementace*. Z angličtiny přeložil Marek Kocan. 1. vyd. Praha: Computer Press, 2002. 257 s, 1 CD-ROM. Profi, databáze. ISBN 80-7226-560-1.

[17] CHAPMAN, Pete, aj. *CRISP-DM 1.0: Step-by-step data mining guide* [online]. USA: SPSS, Inc., 2000. [cit. 2006-06-24]. Dostupné z: http://www.crisp-dm.org/download.htm.

[18] INMON, William H. *Building the data warehouse*. 3 vyd. New York: John Wiley & Sons, Inc., 2002. 464 s. Database, data warehousing technologies. ISBN 0-471-08130-2.

[19] INMON, William H. *Building the operational data store*. 2 vyd. New York: John Wiley & Sons, Inc., 1999. 336 s. Database. ISBN 0-471-32888-X.

[20] INMON, William H. *Data Quality* [online]. USA: Business intelligence network, 2004. [cit. 2004-06-27]. Dostupné z: http://www.b-eye-network.com/view/188

[21] JAIN, Anil K. – DUBES, Richard C. *Algorithms for clustering data*. 1. vyd. New Jersey: Prentice Hall, 1988. 320 s. Prentice Hall advanced reference series. ISBN 0-13-022278-X.

[22] KIMBALL, Ralf, aj. *The data warehouse lifecycle toolkit: expert methods for designing, developing and deploying data warehouses*. 1. vyd. New York: John Wiley & Sons, Inc., 1998. 800 s, 1 CD-ROM. Database, data warehousing. ISBN 0-471-25547-5.

[23] KOTLER, Philip. *Marketing management: analýza, plánování, realizace a kontrola*. Z angličtiny přeložil Václav Dolanský. 2 vyd. Praha: Victoria Publishing, 1995. 789 s. ISBN 80-85605-08-2.

[24] KOTLER, Philip – ARMSTRONG, Gary. *Marketing: eine Einführung*. Z angličtiny do němčiny přeložil Peter Linnert. 1. vyd. Wien: Service-Fachverlag, 1988. 832 s. ISBN 3-85428-109-9.

[25] LUKASOVÁ, Alena – ŠARMANOVÁ, Jana. *Metody shlukové analýzy*. 1. vyd. Praha: SNTL, 1985. 212 s.

[26] MARCO, David. *Building and managing the meta data repository: a full lifecycle guide*. 1. vyd. New York: John Wiley & Sons, Inc., 2000. 416 s, 1 CD-ROM. Database, data warehousing. ISBN 0-471-35523-2.

[27] MELOUN, Milan – MELITKÝ, Jiří. *Kompendium statistického zpracování dat: metody a řešené úlohy včetně CD*. 1. vyd. Praha: Academia, 2002. 764 s, 1 CD-ROM. ISBN 80-200-1008-4.

[28] MRÁZEK, Jan. ETL: the best-kept secret of success in data warehousing. *DM Review*. 2003, ročník 13, č. 6, s 44.

[29] PARR RUD, Olivia. *Data mining: praktický průvodce dobýváním dat pro efektivní prodej, cílený marketing a podporu zákazníků (CRM)*. Z angličtiny přeložil Ivo Magera a Milan Daněk. 1. vyd. Praha: Computer Press, 2001. 329 s, 1 CD-ROM. Rychle a jistě, databáze. ISBN 80-7226-577-6.

[30] PARR RUD, Olivia. *Introduction to Effective Predictive Modeling, Course notes*. Praha: SAS Institute Inc., 2006. 402 s.

[31] POLÁŠEK, Marek. Jak poznat kvalitní systém business inteligence. *IT system*. 2003, ročník 5, č. 6, s. 53.

[32] PORTER, Michaele E. *Konkurenční výhoda*. Z angličtiny přeložil Vladimír Irgl. 1. vyd. Praha: Victoria Publishing, 1994. 626 s. ISBN 80-85605-12-0.

[33] PYLE, Dorian. *Data preparation for data mining*. 1 vyd. San Francisco: Morgan Kaufmann Publishers, Inc., 1999. 562 s, 1 CD-ROM. ISBN 1-55860-529-0.

[34] SAS Institute Inc. *Sas Procedures Guide, Version 6*. 3. vyd. Cary NC: SAS Institute Inc., 1990. 705 s, ISBN 1-55544-378-8.

[35] SAS Institute Inc. *SAS/STAT User's Guide, Version 6, Volume 1*. 4. vyd. Cary NC: SAS Institute Inc., 1989. 943 s., ISBN 1-55544-376-1.

[36] ŠÁLY, Martin. Chyťte je než utečou ke konkurenci. *Ekonom*. 2003, ročník 47, č. 19, s. 42.

[37] ŠIMŮNEK, Milan. *SQL: kompletní kapesní průvodce*. 1. vyd. Praha: Grada Publishing, 1999. 248 s. ISBN 80-7169-692-7.

[38] *Výroční zpráva 2005*. Praha: Česká pojišťovna, a. s., 2005.

[39] WALSH, Sue, aj. *Applying data mining techniques using Enterprice miner: course notes*. Cary: SAS Institute, Inc., 2002. 300 s.

9 Použité pojmy a zkratky

CASE (Computer Aided Software Engineering) – nástroj pro automatizovanou podporu úloh prováděných při vývoji softwaru. Využívá centrální databázi, která obsahuje všechny technické, organizační a řídící informace nutné pro výstavbu a údržbu informačního systému.

Call-centrum – útvar klientského servisu, který poskytuje zákazníkům služby prostřednictvím bezplatné telefonní linky.

Data Marts – datová tržiště. Specializované podmnožiny statistických jednotek uložených v datovém skladu, které jsou určeny pro analytické účely a tvorbu reportů.

Data mining – „dolování dat“. Netriviální proces zjišťování platných, neznámých, potenciálně užitečných a snadno pochopitelných informací z dat.

Data Warehouse – datový sklad. Systém určený pro podporu rozhodování a plánování, který obsahuje mimo jiné základní soubor statistických jednotek (např. klientů, pojistných smluv, pojistných událostí apod.). Pro hodnoty znaků statistických jednotek udržuje historii jejich vývoje v čase.

Direct Mail – reklama zaslaná poštou potenciálnímu zákazníkovi.

DSS (Decision Support System) – systém pro podporu rozhodování. Jeho hlavní součásti tvoří datový sklad a sklad provozních dat.

EIS (Executive Information System) – informační systém určený převážně pro management firmy, který umožňuje zadávat předpřipravené dotazy na vysoce sumarizovaná data.

FK (Foreing Key) – cizí klíč. Tvoří jej sloupec tabulky, jehož hodnoty jsou rovny hodnotám primárního klíče v jiné tabulce.

GSM (Global System for Mobile Communication) – standard pro mobilní komunikaci. Služby GSM umožňují bezdrátový přenos hovorů a dat.

MPP (Massively Parallel Processor) – hardwarová technologie, která s použitím vysokorychlostní sítě spojuje nezávislé jednotky pamětí, procesorů a disků, jenž mají vlastní lokální sběrnici.

NUTS (La Nomenclature des Unités Territoriales Statistiques) – klasifikace vytvořená statistickým úřadem Evropské komise (Eurostatem) za účelem vymezení jednotné unifikované struktury územních jednotek. Je závazná pro všechny členské státy Evropské unie.

ODS (Operational Data Store) - sklad provozních dat určený pro podporu taktického rozhodování a sledování provozu.

ODBC (Open Database Connectivity) – konektor, který umožňuje zpřístupnit data z různých aplikací různým systémům pro řízení databází.

OECD (Organisation for Economic Co-operation and Development) - Organizace pro hospodářskou spolupráci a rozvoj, která sdružuje třicet ekonomicky nejrozvinutějších států světa. Koordinuje ekonomickou a sociálně-politickou spolupráci členských zemí, zprostředkovává nové investice a prosazuje liberalizaci mezinárodního obchodu.

OKEČ (Odvětvová klasifikace ekonomických činností) – klasifikace Českého statistického úřadu pro kategorizaci činností vykonávaných ekonomickými subjekty.

OLTP (Online Transaction Processing) – označení souboru informačních systémů, které automatizují a sbírají obchodní transakce. Tyto systémy bývají rovněž nazývány jako provozní nebo transakční systémy.

PK (Primary Key) – primární klíč. Sloupec, jehož hodnoty jednoznačně identifikují řádky tabulky.

RDBMS (Relational Database Management System) – systém pro řízení relačních databází. Soubor programů, který umožňuje skladování, modifikaci a výběr dat uložených v databázi.

Risk management – pečlivé a přesné sledování obchodních procesů a produktivity, které umožňuje alokovat firemní zdroje tak, aby bylo minimalizováno riziko neúspěchu a ztráty.

Trigger – příkaz nebo procedura, která je spuštěna při vzniku určité události. Touto událostí může být např. vložení, změna nebo vymazání řádku tabulky.

SQL (Structured Query Language) – dotazovací jazyk určený pro manipulaci s daty uložených v databázi.

Work Flow – tok pracovních činností (např. všechny pracovní úkony, které jsou vykonány od sepsání návrhu pojistné smlouvy po její akceptaci pojišťovnou). Předmětem zájmu firmy je podpora plynulého toku prací a identifikace odchylek od standardu za účelem zajištění nápravy.

Report – výstupní sestava, která může mít buď statickou nebo dynamickou podobu.

View – dynamický pohled na tabulky databáze. Nad jednou databází může být vytvořeno několik různých pohledů pro různé uživatele, přičemž není nutné duplikovat podkladová data.

Statistická jednotka – jednoznačně identifikovatelný objekt (např. klient, pojistná událost apod.). Pro vyjádření statistické jednotky je možné použít také pojem entita.

Statistický znak – měřená vlastnost statistické jednotky. V různém kontextu lze pro jeho vyjádření použít také pojmy: proměnná, pole, sloupec tabulky, vlastnost, vektor apod.

SMP (Symmetric Multiprocessor) – hardwarová technologie, která je založena na centrální sběrnici spojující procesory, paměti a disky.

Záznam – statistická jednotka se všemi svými měřenými znaky. V různém kontextu lze pro vyjádření záznamu použít také pojmy: řádek tabulky, nahrávka apod.

10 Přílohy

Příloha 1 Tabulka Segmentace 2

Příloha 2 Tabulka Segmentace 4

Příloha 3 Kontingenční tabulka pro proměnné Právní forma (Pravfor) a Shluk

Příloha 4 Kontingenční tabulka pro proměnné Institucionální sektor (Isektor) a Shluk

Příloha 5 Kontingenční tabulka pro proměnné Hlavní předmět podnikatelské činnosti (Okec_sekce) a Shluk

Příloha 6 Kontingenční tabulka pro proměnné Počet pracovníků (Pocprac) a Shluk

Příloha 7 Kontingenční tabulka pro proměnné Oblast sídla firmy (Oblast) a Shluk

Příloha 8 Kontingenční tabulka pro proměnné Status podnikatelské činnosti firmy dle externích dat (Aktiv) a Shluk

Příloha 1 Tabulka Segmentace 2

Shluk	_Freq_	_Radius_	_Gap_	_Near_	_Rmsstd_	Avg_jmeni_std	Avg_r_predpis_std	Avg_stari_std	Avg_pomer_std
Číslo shluku	Počet jednotek uvnitř shluku	Maximální vzdálenost od středu shluku	Vzdálenost od středu nejbližšího shluku	Číslo nejbližšího shluku	Střední kvadratická směrodatná odchylka	Vnitroshlukový průměr proměnné Jmeni_std	Vnitroshlukový průměr proměnné R_predpis_std	Vnitroshlukový průměr proměnné Stari_std	Vnitroshlukový průměr proměnné Pomer_std
1	2	0,71978162	2,29883434	4		3,94124479	4,31860166	4,08551915	1,36423282
2	3 447	1,87850477	1,98943215	29	0,27639194	8,40711865	-0,33057514	-0,23732063	-0,42371327
3	31 306	0,91180310	1,08760854	13	0,18351745	-0,13317274	-0,23709700	0,57373674	-0,43086793
4	18	1,41944374	2,29883434	1	0,32549889	5,11083395	3,33031952	4,01559152	-0,34898318
5	1	0	5,38183501	38		3,01132778	-0,27144225	-0,38984949	2,74919507
6	66	1,58786661	2,60702332	41	0,41078790	-0,16003981	3,87698139	0,92522037	2,52650746
7	55	1,89013342	2,14958933	15	0,58561461	4,24767478	1,90065899	-0,69383680	-0,28331072
8	43	2,28785017	2,04982544	25	0,63446329	7,18295298	3,63484052	-0,80382109	-0,18786214
9	313	2,04940544	2,23660670	13	0,41757507	-0,14509797	-0,32406555	-1,48932708	2,12094722
10	50	1,36288065	2,64236497	6	0,21964849	-0,16003981	6,51827668	0,91622748	2,60114659
11	286	2,11339857	1,64150228	30	0,40014133	8,42915535	6,44321638	0,20396709	-0,34602067
12	54	1,97360098	1,60093520	17	0,64461366	5,33792634	5,59446156	0,28145580	-0,37366549
13	28 724	0,83049707	1,08760854	3	0,28068248	-0,15457487	-0,28595519	0,09794853	0,54569377
14	427	1,88234189	2,26491042	23	0,40146133	0,18025381	-0,31722146	3,89655914	-0,18260351
15	83	2,07240802	2,14958933	7	0,54286116	3,04373323	3,09838007	0,61710845	-0,14860444
16	20	1,75213709	2,49623406	34	0,32427774	8,31367260	2,52145458	3,90885986	-0,39466037
17	39	1,73726817	1,60093520	12	0,51867138	6,19259571	6,50357425	-0,71591207	-0,26733940
18	18	1,67271990	1,69205069	40	0,41753948	-0,07514412	5,23354306	-0,69259502	0,50543556
19	576	1,83392771	1,78328092	47	0,37713851	0,04949368	6,52752674	0,22697306	-0,33941605
20	1 008	1,92959499	1,83531973	42	0,48863582	5,82313841	0,69160092	0,58391956	-0,35199201
21	165	2,61003279	2,01503484	44	0,71097864	-0,01921961	1,74473279	2,01146550	-0,26310846
22	133	1,93367546	1,74819155	27	0,47128488	8,43416424	1,74172964	-1,01332642	-0,27623564
23	97	2,08591207	2,26491042	14	0,38137132	1,65541156	1,39003377	4,04436634	-0,31360523
24	250	1,83179144	2,37466892	29	0,51226912	6,64015100	0,15452947	-1,87442070	0,40892627
25	17	1,64831384	2,04982544	8	0,45981595	5,84136359	3,14748365	0,65906503	-0,34397016
26	31	1,90737108	2,00459715	33	0,51920947	3,39234446	6,56065973	-1,60565797	-0,38870907

Tabulka pokračuje

Pokračování tabulky

Shluk	_Freq_	_Radius_	_Gap_	_Near_	_Rmsstd_	Avg_jmeni_std	Avg_r_predpis_std	Avg_stari_std	Avg_pomer_std
Číslo shluku	Počet jednotek uvnitř shluku	Maximální vzdálenost od středu shluku	Vzdálenost od středu nejbližšího shluku	Číslo nejbližšího shluku	Střední kvadratická směrodatná odchylka	Vnitroshlukový průměr proměnné Jmeni_std	Vnitroshlukový průměr proměnné R_predpis_std	Vnitroshlukový průměr proměnné Stari_std	Vnitroshlukový průměr proměnné Pomer_std
27	416	2,14965325	1,67206111	28	0,49558615	8,39909119	2,58451907	0,51654746	-0,34054313
28	76	1,56824260	1,67206111	27	0,26743886	8,47143797	1,51935756	0,61238095	0,94273439
29	80	1,70657328	1,98943215	2	0,37825344	8,35799179	-0,26429145	-0,72712365	1,50271511
30	60	2,22447103	1,64150228	11	0,34137090	8,44289920	6,45035774	-1,43663063	-0,39826019
31	22	1,94902853	2,28362560	15	0,58712068	2,20660358	3,65199478	-1,43037270	-0,27306934
32	15	1,97857767	1,99512818	45	0,43537584	-0,08319407	-0,28447311	2,82427890	2,47228471
33	100	2,14947314	2,00459715	26	0,53730763	2,58467301	6,40950547	0,22208867	-0,33800258
34	712	1,95882427	2,49623406	16	0,51334532	8,03153049	0,04182271	3,85924852	-0,41800214
35	719	1,85342059	1,68687931	40	0,45649330	-0,01569713	3,78880853	0,44436671	-0,36032633
36	1 736	1,92497005	2,16386946	38	0,43397251	3,42234793	-0,21868210	-1,54369011	-0,40474893
37	15	2,31113834	1,90867996	43	0,68336035	1,83165148	4,70343995	3,81011185	-0,25155163
38	764	2,02075153	2,16386946	36	0,35050553	3,70244231	-0,39974792	0,59431722	-0,40898831
39	22	2,18587241	1,83266636	11	0,55434925	8,37781070	4,86485466	1,13390697	-0,34999689
40	184	1,84976935	1,68687931	35	0,44002852	-0,06024728	3,87303115	-1,23968042	-0,33867256
41	114	1,91338271	1,73087930	35	0,56221662	-0,10663414	2,52520797	-0,21954785	0,61445810
42	729	1,66954298	1,83531973	20	0,38117137	7,03958994	-0,02929686	0,76204032	0,80439183
43	29	2,79112954	1,90867996	37	0,60296795	0,86430748	6,33340029	3,59003191	-0,20573819
44	4 044	1,30118400	1,44531272	3	0,15325592	-0,00003661	1,12063279	0,09918745	-0,38010968
45	3 367	2,06914406	1,99512818	32	0,30977737	-0,15978229	0,08004591	0,89726759	2,83064764
46	17 597	1,00667598	1,16230945	49	0,26475783	-0,12757006	-0,28050601	-1,44795319	-0,40979002
47	86	1,47617289	1,78328092	19	0,28969818	-0,03511868	6,59267364	-1,55292495	-0,36492205
48	1	0	6,32407145	39		7,98108100	6,49039601	4,00611745	-0,34111539
49	1 876	2,01033361	1,16230945	46	0,39407223	-0,11616358	0,87957677	-1,41400202	-0,34743290
50	7	1,83908039	2,32463875	37	0,56094731	3,19972460	6,55982652	3,92568456	0,01825035

Příloha 2 Tabulka Segmentace 4

Shluk	_Freq_	_Radius_	_Gap_	_Near_	_Rmsstd_	Avg_jmeni_std	Avg_r_predpis_std	Avg_stari_std	Avg_pomer_std
Číslo shluku	Počet jednotek uvnitř shluku	Maximální vzdálenost od středu shluku	Vzdálenost od středu nejbližšího shluku	Číslo nejbližšího shluku	Střední kvadratická směrodatná odchylka	Vnitroshlukový průměr proměnné Jmeni_std	Vnitroshlukový průměr proměnné R_predpis_std	Vnitroshlukový průměr proměnné Stari_std	Vnitroshlukový průměr proměnné Pomer_std
1	49 100	1,09996018	1,18070649	3	0,23154072	-0,14925992	-0,27242963	0,34422648	-0,40861263
2	11 720	1,19937266	3,01548049	1	0,23667988	-0,15951238	-0,25182245	0,87095288	2,56041955
3	5 544	1,09916618	1,18070649	1	0,36673574	-0,10110517	0,90520153	0,30722826	-0,34890563
4	25 208	1,19991541	1,63341426	1	0,30637506	-0,13838798	-0,22734724	-1,28814071	-0,37299122

Příloha 3 Kontingenční tabulka pro proměnné Právní forma (Pravfor) a Shluk

Pravfor Skutečná četnost Teoretická četnost Procento z celku Procento řádku Procento sloupce	Shluk 1	2	3	4	Součet
Společnost s ručením omezeným	6 705	240	4 239	8 620	19 804
	9 963,86	2 557,08	2 015,66	5 267,40	
	6,84	0,25	4,33	8,80	20,22
	33,86	1,21	21,40	43,53	
	13,60	1,90	42,52	33,09	
Akciová společnost	351	19	1 235	721	2 326
	1 170,27	300,33	236,74	618,66	
	0,36	0,02	1,26	0,74	2,37
	15,09	0,82	53,10	31,00	
	0,71	0,15	12,39	2,77	
Podnikatel - fyzická osoba	30 193	10 807	2 145	12 162	55 307
	27 826,27	7 141,20	5 629,17	14 710,37	
	30,82	11,03	2,19	12,42	56,46
	54,59	19,54	3,88	21,99	
	61,26	85,44	21,51	46,68	
Samostatně hospodařící rolník	1 011	554	89	152	1 806
	908,64	233,19	183,82	480,35	
	1,03	0,57	0,09	0,16	1,84
	55,98	30,68	4,93	8,42	
	2,05	4,38	0,89	0,58	
Svobodné povolání	4 936	785	295	1 421	7 437
	3 741,73	960,26	756,94	1 978,07	
	5,04	0,80	0,30	1,45	7,59
	66,37	10,56	3,97	19,11	
	10,02	6,21	2,96	5,45	
Družstvo	332	4	273	234	843
	424,13	108,85	85,80	224,22	
	0,34	0,00	0,28	0,24	0,86
	39,38	0,47	32,38	27,76	
	0,67	0,03	2,74	0,90	
Příspěvková organizace	532	1	266	682	1 481
	745,13	191,23	150,74	393,91	
	0,54	0,00	0,27	0,70	1,51
	35,92	0,07	17,96	46,05	
	1,08	0,01	2,67	2,62	
Zahraniční osoba	340	41	34	527	942
	473,94	121,63	95,88	250,55	
	0,35	0,04	0,03	0,54	0,96
	36,09	4,35	3,61	55,94	
	0,69	0,32	0,34	2,02	
Sdružení	799	6	119	261	1 185
	596,20	153,01	120,61	315,18	
	0,82	0,01	0,12	0,27	1,21
	67,43	0,51	10,04	22,03	
	1,62	0,05	1,19	1,00	
Obecní úřad	1 241	56	372	5	1 674
	842,23	216,15	170,38	445,24	
	1,27	0,06	0,38	0,01	1,71
	74,13	3,35	22,22	0,30	
	2,52	0,44	3,73	0,02	

Tabulka pokračuje

Pokračování tabulky

Pravfor	Shluk				
Skutečná četnost					
Teoretická četnost					
Procento z celku					
Procento řádku					
Procento sloupce	1	2	3	4	Součet
Ostatní	2 844	135	903	1 269	5 151
	2 591,59	665,09	524,27	1 370,05	
	2,90	0,14	0,92	1,30	5,26
	55,21	2,62	17,53	24,64	
	5,77	1,07	9,06	4,87	
Součet	49 284	12 648	9 970	26 054	97 956
	50,31	12,91	10,18	26,60	100,00

Legenda: Měřeno nad kombinací dat České pojišťovny a Registru ekonomických subjektů k 30.6.2006.

Příloha 4 Kontingenční tabulka pro proměnné Institucionální sektor (Isektor) a Shluk

Isektor Skutečná četnost Teoretická četnost Procento z celku Procento řádku Procento sloupce	Shluk 1	2	3	4	Součet
Nefinanční podniky	8 552	657	6 109	10 308	25 626
	12 893,19	3 308,72	2 608,36	6 815,74	
	8,73	0,67	6,24	10,52	26,16
	33,37	2,56	23,84	40,22	
	17,35	5,19	61,27	39,57	
Finanční instituce	38	0	68	48	154
	77,48	19,88	15,67	40,96	
	0,04	0,00	0,07	0,05	0,16
	24,68	0,00	44,16	31,17	
	0,08	0,00	0,68	0,18	
Vládní instituce	1 776	58	655	690	3 179
	1 599,45	410,46	323,58	845,52	
	1,81	0,06	0,67	0,70	3,25
	55,87	1,82	20,60	21,70	
	3,60	0,46	6,57	2,65	
Domácnosti	37 434	11 915	2 918	13 922	66 189
	33 301,61	8 546,03	6 737,09	17 604,27	
	38,22	12,16	2,98	14,21	67,57
	56,56	18,00	4,41	21,03	
	75,96	94,21	29,27	53,44	
Neziskové instituce	1 476	16	216	1 084	2 792
sloužící domácnostem	1 404,74	360,49	284,19	742,59	
	1,51	0,02	0,22	1,11	2,85
	52,87	0,57	7,74	38,83	
	3,00	0,13	2,17	4,16	
Nerezidenti	6	1	4	0	11
	5,53	1,42	1,12	2,93	
	0,01	0,00	0,00	0,00	0,01
	54,55	9,09	36,36	0,00	
	0,01	0,01	0,04	0,00	
Součet	49 282	12 647	9 970	26 052	97 951
	50,31	12,91	10,18	26,60	100,00

Legenda: Měřeno nad kombinací dat České pojišťovny a Registru ekonomických subjektů k 30.6.2006.

Příloha 5 Kontingenční tabulka pro proměnné Hlavní předmět podnikatelské činnosti (Okec_sekce) a Shluk

Okec_sekce Skutečná četnost Teoretická četnost Procento z celku Procento řádku Procento sloupce	Shluk 1	2	3	4	Součet
Zemědělství, myslivost, lesnictví	3 361	602	1 293	868	6 124
	3 081,19	790,74	623,32	1 628,75	
	3,43	0,61	1,32	0,89	6,25
	54,88	9,83	21,11	14,17	
	6,82	4,76	12,97	3,33	
Rybolov a chov ryb	24	3	18	3	48
	24,15	6,20	4,89	12,77	
	0,02	0,00	0,02	0,00	0,05
	50,00	6,25	37,50	6,25	
	0,05	0,02	0,18	0,01	
Těžba nerostných surovin	18	10	30	12	70
	35,22	9,04	7,12	18,62	
	0,02	0,01	0,03	0,01	0,07
	25,71	14,29	42,86	17,14	
	0,04	0,08	0,30	0,05	
Zpracovatelský průmysl	7 115	2 721	1 638	3 537	15 011
	7 552,55	1 938,25	1 527,86	3 992,35	
	7,26	2,78	1,67	3,61	15,32
	47,40	18,13	10,91	23,56	
	14,44	21,51	16,43	13,58	
Výroba a rozvod elektřiny, plynu a vody	58	11	75	39	183
	92,07	23,63	18,63	48,67	
	0,06	0,01	0,08	0,04	0,19
	31,69	6,01	40,98	21,31	
	0,12	0,09	0,75	0,15	
Stavebnictví	5 164	2 079	941	3 186	11 370
	5 720,63	1 468,12	1 157,27	3 023,98	
	5,27	2,12	0,96	3,25	11,61
	45,42	18,28	8,28	28,02	
	10,48	16,44	9,44	12,23	
Obchod; opravy motorových vozidel a výrobků pro osobní potřebu a převážně pro domácnost	11 125	2 386	1 962	6 381	21 854
	10 995,49	2 821,83	2 224,35	5 812,32	
	11,36	2,44	2,00	6,51	22,31
	50,91	10,92	8,98	29,20	
	22,57	18,86	19,68	24,49	
Ubytování a stravování	3 101	978	195	1 567	5 841
	2 938,81	754,20	594,51	1 553,48	
	3,17	1,00	0,20	1,60	5,96
	53,09	16,74	3,34	26,83	
	6,29	7,73	1,96	6,01	
Doprava, skladování a spoje	3 229	869	1 243	1 708	7 049
	3 546,59	910,18	717,46	1 874,76	
	3,30	0,89	1,27	1,74	7,20
	45,81	12,33	17,63	24,23	
	6,55	6,87	12,47	6,56	
Finanční zprostředkování	1 055	361	121	917	2 454
	1 234,69	316,86	249,77	652,67	
	1,08	0,37	0,12	0,94	2,51
	42,99	14,71	4,93	37,37	
	2,14	2,85	1,21	3,52	

Tabulka pokračuje

Pokračování tabulky

Okec_sekce	Shluk				
Skutečná četnost					
Teoretická četnost					
Procento z celku					
Procento řádku					
Procento sloupce	1	2	3	4	Součet
Činnosti v oblasti nemovitostí a pronájmu; podnikatelské činnosti	7 044	1 826	1 141	5 307	15 318
	7 707,01	1 977,89	1 559,10	4 074,00	
	7,19	1,86	1,16	5,42	15,64
	45,99	11,92	7,45	34,65	
	14,29	14,44	11,44	20,37	
Veřejná správa a obrana; povinné sociální zabezpečení	1 322	63	441	31	1 857
	934,32	239,78	189,01	493,89	
	1,35	0,06	0,45	0,03	1,90
	71,19	3,39	23,75	1,67	
	2,68	0,50	4,42	0,12	
Vzdělávání	630	94	131	792	1 647
	828,66	212,66	167,64	438,04	
	0,64	0,10	0,13	0,81	1,68
	38,25	5,71	7,95	48,09	
	1,28	0,74	1,31	3,04	
Zdravotní a sociální péče; veterinární činnosti	3 096	116	262	502	3 976
	2 000,46	513,39	404,69	1 057,46	
	3,16	0,12	0,27	0,51	4,06
	77,87	2,92	6,59	12,63	
	6,28	0,92	2,63	1,93	
Ostatní veřejné, sociální a osobní služby	2 935	528	476	1 202	5 141
	2 586,61	663,82	523,26	1 367,31	
	3,00	0,54	0,49	1,23	5,25
	57,09	10,27	9,26	23,38	
	5,96	4,17	4,77	4,61	
Exteritoriální organizace a instituce	7	1	3	0	11
	5,53	1,42	1,12	2,93	
	0,01	0,00	0,00	0,00	0,01
	63,64	9,09	27,27	0,00	
	0,01	0,01	0,03	0,00	
Součet	49 284	12 648	9 970	26 052	97 954
	50,31	12,91	10,18	26,60	100,00

Legenda: Měřeno nad kombinací dat České pojišťovny a Registru ekonomických subjektů k 30.6.2006.

Příloha 6 Kontingenční tabulka pro proměnné Počet pracovníků (Pocprac) a Shluk

Pocprac Skutečná četnost Teoretická četnost Procento z celku Procento řádku Procento sloupce	Shluk 1	2	3	4	Součet
0	21 451	6 224	711	9 201	37 587
	18 789,53	4 975,20	4 280,65	9 541,61	
	25,17	7,30	0,83	10,79	44,10
	57,07	16,56	1,89	24,48	
	50,35	55,17	7,32	42,52	
1 - 5	14 307	3 561	1 987	7 290	27 145
	13 569,63	3 593,05	3 091,45	6 890,87	
	16,79	4,18	2,33	8,55	31,85
	52,71	13,12	7,32	26,86	
	33,58	31,56	20,47	33,69	
6 - 9	2 546	666	1 087	1 861	6 160
	3 079,35	815,37	701,54	1 563,74	
	2,99	0,78	1,28	2,18	7,23
	41,33	10,81	17,65	30,21	
	5,98	5,90	11,20	8,60	
10 - 19	2 239	562	1 700	1 760	6 261
	3 129,84	828,74	713,04	1 589,38	
	2,63	0,66	1,99	2,06	7,35
	35,76	8,98	27,15	28,11	
	5,25	4,98	17,51	8,13	
20 - 24	476	104	628	403	1 611
	805,33	213,24	183,47	408,96	
	0,56	0,12	0,74	0,47	1,89
	29,55	6,46	38,98	25,02	
	1,12	0,92	6,47	1,86	
25 - 49	852	105	1 347	651	2 955
	1 477,19	391,14	336,53	750,14	
	1,00	0,12	1,58	0,76	3,47
	28,83	3,55	45,58	22,03	
	2,00	0,93	13,88	3,01	
50 - 99	473	46	1 050	262	1 831
	915,31	242,36	208,53	464,81	
	0,55	0,05	1,23	0,31	2,15
	25,83	2,51	57,35	14,31	
	1,11	0,41	10,82	1,21	
100 a více	264	14	1 197	209	1 684
	841,82	222,90	191,78	427,49	
	0,31	0,02	1,40	0,25	1,98
	15,68	0,83	71,08	12,41	
	0,62	0,12	12,33	0,97	
Součet	42 608	11 282	9 707	21 637	85 234
	49,99	13,24	11,39	25,39	100,00

Legenda: Měřeno nad kombinací dat České pojišťovny a Registru ekonomických subjektů k 30.6.2006.

Příloha 7 Kontingenční tabulka pro proměnné Oblast sídla firmy (Oblast) a Shluk

Oblast Skutečná četnost Teoretická četnost Procento z celku Procento řádku Procento sloupce	Shluk 1	2	3	4	Součet
Praha	8 225	1 520	1 915	4 715	16 375
	8 238,82	2 114,37	1 666,69	4 355,12	
	8,40	1,55	1,95	4,81	16,72
	50,23	9,28	11,69	28,79	
	16,69	12,02	19,21	18,10	
Střední Čechy	6 515	1 749	1 292	3 069	12 625
	6 352,07	1 630,16	1 285,00	3 357,76	
	6,65	1,79	1,32	3,13	12,89
	51,60	13,85	10,23	24,31	
	13,22	13,83	12,96	11,78	
Jihozápad	5 214	1 318	1 058	2 639	10 229
	5 146,56	1 320,79	1 041,13	2 720,52	
	5,32	1,35	1,08	2,69	10,44
	50,97	12,88	10,34	25,80	
	10,58	10,42	10,61	10,13	
Severozápad	4 386	1 332	914	2 462	9 094
	4 575,50	1 174,23	925,61	2 418,65	
	4,48	1,36	0,93	2,51	9,28
	48,23	14,65	10,05	27,07	
	8,90	10,53	9,17	9,45	
Severovýchod	7 008	2 315	1 319	3 376	14 018
	7 052,93	1 810,03	1 426,79	3 728,25	
	7,15	2,36	1,35	3,45	14,31
	49,99	16,51	9,41	24,08	
	14,22	18,30	13,23	12,96	
Jihovýchod	7 450	1 795	1 508	4 036	14 789
	7 440,85	1 909,58	1 505,26	3 933,31	
	7,61	1,83	1,54	4,12	15,10
	50,38	12,14	10,20	27,29	
	15,12	14,19	15,13	15,49	
Střední Morava	5 564	1 482	1 023	2 837	10 906
	5 487,18	1 408,20	1 110,04	2 900,58	
	5,68	1,51	1,04	2,90	11,13
	51,02	13,59	9,38	26,01	
	11,29	11,72	10,26	10,89	
Ostravsko	4 922	1 137	941	2 918	9 918
	4 990,08	1 280,63	1 009,48	2 637,81	
	5,02	1,16	0,96	2,98	10,13
	49,63	11,46	9,49	29,42	
	9,99	8,99	9,44	11,20	
Součet	49 284	12 648	9 970	26 052	97 954
	50,31	12,91	10,18	26,60	100,00

Legenda: Měřeno nad kombinací dat České pojišťovny a Registru ekonomických subjektů k 30.6.2006.

Příloha 8 Kontingenční tabulka pro proměnné Status podnikatelské činnosti firmy dle externích dat (Aktiv) a Shluk

Aktiv	**Shluk**				
Skutečná četnost					
Teoretická četnost					
Procento z celku					
Procento řádku					
Procento sloupce	1	2	3	4	Součet
Firmě byla pozastavena	4 407	220	213	3 295	8 135
podnikatelská činnost	4 092,91	1 050,38	827,98	2 163,72	
	4,50	0,22	0,22	3,36	8,30
	54,17	2,70	2,62	40,50	
	8,94	1,74	2,14	12,65	
Firma podnikatelskou činnost	44 234	2 743	9 604	22 758	79 339
aktivně provozuje	39 917,34	10 244,19	8 075,15	21 102,31	
	45,16	2,80	9,80	23,23	80,99
	55,75	3,46	12,11	28,68	
	89,75	21,69	96,33	87,35	
Firma zanikla	643	9 685	153	1	10 482
	5 273,74	1 353,43	1 066,86	2 787,97	
	0,66	9,89	0,16	0,00	10,70
	6,13	92,40	1,46	0,01	
	1,30	76,57	1,53	0,00	
Součet	49 284	12 648	9 970	26 054	97 956
	50,31	12,91	10,18	26,60	100,00

Legenda: Měřeno nad kombinací dat České pojišťovny a Registru ekonomických subjektů k 30.6.2006.