Česká zemědělská univerzita v Praze

Provozně ekonomická fakulta

Katedra sytémového inženýrství

Diplomová práce

Aplikace systémových analýz v Data miningu IS PČR

Bc. Petr Netík

ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE

Katedra systémového inženýrství

Provozně ekonomická fakulta

ZADÁNÍ DIPLOMOVÉ PRÁCE

Netík Petr

Veřejná správa a regionální rozvoj nav.- Hradec

Název práce

Aplikace systémových analýz v Data-miningu IS PČR

Anglický název

Application of System Analysis Tools in Datamining of IS PCR

Cíle práce

V teoretické části diplomové práce se věnujte popisu metod, které se využívají při data miningu. V praktické části vyhledejte vhodné datové soubory a vybranou metodu aplikujte pomocí analytického software.

Metodika

Při vypracování diplomové práce využijte kvantitativní metodu CRISP - DM.

Harmonogram zpracování

01.05.2014 - 31.05.2014: Formulace problému a stanovení cílů
01.06.2014 - 31.07.2014: Studium literatury
01.08.2014 - 31.08.2014: Charakteristika zkoumaného problému
01.09.2014 - 30.09.2014: Vlastní zpracování dat a upřesnění hypotézy řešení
01.10.2014 - 31.10.2014: Analýza výsledků, interpretace a implementace
01.11.2014 - 30.11.2014: Finální zpracování textu

Rozsah textové části

Rozsah práce je stanovený na 60 normovaných stran.

Klíčová slova

data mining, informační systém, policie, databáze

Doporučené zdroje informací

POŽÁR, Josef., Manažerská informatika, Plzeň: Aleš Čeněk, 2010, 357 s. ISBN 978-80-7380-276-9.

BERKA, Petr., Dobývání znalostí z databází, Vyd. 1. Praha: Academia, 2003, 366 s. ISBN 80-200-1062-9.

BÍNOVÁ, Dagmar, Využití vybraných statistických metod při zpracování dat technikami Data mining, [online]. Praha, 2006 [cit. 2014-08-01]. Dostupné z: "http://www.pef.czu.cz/cs/?dl=1&f=12934". Disertační práce. Česká zemědělská univerzita.

Internetové zdroje

Vedoucí práce

Švasta Jaroslav, doc. Ing., CSc.

Termín odevzdání

listopad 2014

Elektronicky schváleno dne 16.10.2014

doc. Ing. Tomáš Šubrt, Ph.D.

Vedoucí katedry

Elektronicky schváleno dne 10.11.2014

Ing. Martin Pelikán, Ph.D.

Děkan fakulty

Čestné prohlášení

Prohlašuji, že svou diplomovou práci "Aplikace systémových analýz v Data miningu IS PČR" jsem vypracoval samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou citovány v práci a uvedeny v seznamu literatury na konci práce. Jako autor uvedené diplomové práce dále prohlašuji, že jsem v souvislosti s jejím vytvořením neporušil autorská práva třetích osob.

V Praze dne 26. listopadu 2014 _____

Poděkování

Rád bych touto cestou poděkoval doc. Ing. Jaroslavovi Švastovi, CSc. za poskytnuté rady, odborné vedení, podnětné návrhy a zpětnou vazbu při konzultacích tohoto textu. Dále děkuji panu Ing. Janu Stolínovi ze Skupiny zpracování dat Policejního prezidia ČR za poskytnutí rad při poznávání dat ze systému ESSK.

Poděkování patří i rodině a nejbližším přátelům. Jim vděčím za podporu při studiích a za jejich trpělivost a toleranci při zpracování mé diplomové práce.

Aplikace systémových analýz v Data-miningu IS PČR

Application of System Analysis Tools in Datamining of IS PCR

Souhrn

Tato diplomová práce seznamuje s technikou data miningu v oblasti zdrojových dat o kriminalitě. Teoretická část představuje metody dolování dat, dobývání znalostí a popíše použitý software. Empirická část práce zkoumá zdrojová data pocházející z Evidenčně statistického systému kriminality. Cílem práce bylo popsat možnosti a limity tohoto typu dat s ohledem na jejich data miningové zpracování pro potřeby predikce a prevence. Ve výzkumu jsou použita data z let 2011 až 2013, která se týkají vykradených rodinných domů na území celé České republiky.

Summary

This thesis introduces the technique of data mining in the area of source data relating to criminality. The theoretical part introduces the methods of data mining, knowledge discovery and describe of the use of software. The empirical part examines the source data from the “Evidenčně statistického systému kriminality“. The aim was to describe the possibilities and limitations of this type of data with regard to the data mining processing of the prediction and prevention needs. The research uses data from the years 2011-2012, which relate burgled houses all over the Czech Republic.

Klíčová slova: data mining, informační systém, policie, databáze, CRISP - DM

Keywords: data mining, information system, police, database, CRISP - DM

Obsah

1 Úvod

Současná společnost produkuje rozsáhlé množství dat, která jsou dále zpracovávána v databázích. Patří mezi ně např. telefonní hovory, nákupy kreditní kartou, návštěva lékaře, připojení k internetu, spáchání dopravního přestupku apod. Subjekty vlastnící tyto informace se je logicky snaží využít. Supermarkety takto získávají informace o odbytu zboží, lékař zjišťuje navýšení pacientů onemocněných chřipkou, Policie České republiky (dále jen policie) získává informace o trestné činnosti. Managerská rozhodnutí jsou často založena právě na interpretaci dat z databází.

Předkládaná diplomová práce se zaobírá využíváním informačních a komunikačních technologií ve zkoumání vybraných druhů kriminality se zaměřením na techniku data mining. Vybrána byla trestná činnost krádeže vloupáním do bytů a rodinných domů.

Hlavním cílem této práce je nalezení nových, inovativních a nečekaných informací, které by vedly ke zlepšení prediktivní činnosti policie, případně až k vytipování případného pachatele či pachatelů u vybrané trestné činnosti, např. společné rysy s případy zaznamenanými v minulosti, což by mohlo mít vliv na způsob vyšetřování případu.

Vedlejším cílem práce je možnost zkoumání tvorby typologie jednotlivých krádeží či vloupání, zlepšení efektivity preventivních opatření či zlepšení sběru dat a následné definování na změnu.

Primárním zdrojem informací o sledovaných druzích kriminality jsou statistiky policie. Policie zpracovává informace o podaných oznámeních o spáchání konkrétních trestních činů, jakožto i přidružené informace typu objasnění skutku apod.

Zájem autora o problematiku je zcela pragmatický, poněvadž je zaměstnán u policie a zabývá se analytickou činností, jejíž součástí je i aplikace metod data mining informačních systémů Policie České republiky (dále jen IS PČR).

1.1 Definice

Pro úvod do problematiky je v této podkapitole uvedeno několik definic pojmu data miningu. První zmínky o data miningu pocházejí z 60. let 20. století a souvisí s rozvojem počítačové techniky. Nejednalo se však o využití ve větší míře. To nastalo až v 90. letech 20. století související s nárůstem dat, především v komerčních organizacích. Pojem data mining neboli dolování dat dosud nebyl jednoznačně definován a názory na výstižnou definici se liší dle pojetí jedlotlivých autorů. Většina autorů se shoduje na názoru, že se jedná o postup, při kterém jsou zpracována data na základě data miningových metod a následné výsledky lze použít při managerském rozhodování.

Data mining je analytická metodologie získávání netriviálních skrytých a potenciálně užitečných informací z dat. Někdy se chápe jako analytická součást dobývání znalostí z databází, jindy se tato dvě označení chápou jako souznačná.[1]

Data mining je proces, který používá různé analytické nástroje pro odhalení ukrytých vzorů a závislostí v datech. Výsledkem je predikční model, který je podkladem pro rozhodování.[2]

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]

---

[1] *Data mining* [online]. [cit. 2014-07-10]. Dostupné z: http://cs.wikipedia.org/wiki/Data_mining

[2] Data mining [online]. [cit. 2014-07-10]. Dostupné z: http://axpsu.fpf.slu.cz/~sos10um/trendy/DM.pdf

[3] Berry, Michael J. A. – Linoff, Gordon. Data mining techniques: for marketing, sales and customer support, 1. vydání, str. 454, ISBN 0-471-17980-9

Dolování dat umožňuje pomocí speciálních algoritmů automaticky objevovat v datech strategické informace. Je to analytická technika pevně spjatá s datovými sklady jako s velmi kvalitním datovým zdrojem pro tyto speciální analýzy.

Dolování dat lze charakterizovat jako proces extrakce relevantních, předem neznámých nebo nedefinovaných informací z velmi rozsáhlých databází. Důležitou vlastností dolování dat je, že se jedná o analýzy odvozené od obsahu dat, nikoli předem specifikované uživatelem nebo implementátorem, a jedná se především o odvozování prediktivních informací, nikoliv pouze desktiptivních. Dolování dat slouží manažerům k objevování skutečností, čímž pomáhají zaměřit jejich pozornost na podstatné faktory podnikání, umožňují testovat hypotézy, odhalují ve stále se zrychlujícím a složitějším obchodním prostředí skryté korelace mezi ekonomickými proměnými apod. Data mining je orientován na praktickou využitelnost výsledků.[4]

1.2 Informační systémy PČR

Policie ČR (dále jen PČR) zpracovává velké množství informací, které při své činnosti získá. PČR může zpracovávat osobní údaje včetně citlivých údajů bez souhlasu osoby, jíž se tyto údaje týkají, pokud je to nezbytné pro plnění jejích úkolů. Shromažďovat údaje o rasovém původu, náboženském nebo politickém přesvědčení, o příslušnosti k zákonem nezakázanému sdružení nebo sexuálním chování lze pouze tehdy, je-li to nezbytné pro účely šetření konkrétního trestného činu nebo správního deliktu.[5] Tyto informace jsou ukládány do databází a je s nimi dále pracováno. Dalším způsobem, jak PČR získává informace je možnost vyžádat si informace od třetí strany. Policie je v souvislosti s odhalováním a šetřením přestupku oprávněna vyžadovat výpis z evidence Rejstříku trestů v případech, ve kterých by předchozí

[4] BÍNOVÁ, Dagmar. Vvyužití vybraných statistických metod při zpracování dat technikami Data mining [online]. Praha, 2006 [cit. 2014-10-28]. Dostupné z: www.pef.czu.cz/cs/?dl=1&f=12934. Disertační práce. Česká zemědělská univerzita v Praze. Vedoucí práce Doc. RNDr. Bohumil Kába, CSc., str. 3

[5] Zákon o Policii České republiky., In: 273/2008 Sb. ve znění pozdější předpisů.

trestní postih mohl vést k posouzení skutku jako trestného činu.[6] PČR může v rozsahu potřebném pro plnění konkrétního úkolu žádat od správce evidence nebo zpracovatele poskytnutí informací z evidence provozované na základě jiného právního předpisu a může v rozsahu potřebném pro plnění konkrétního úkolu žádat od správce evidence nebo zpracovatele poskytnutí informací z databáze evidence údajů o mýtném, základního registru obyvatel, registru silničních vozidel, základního registru právnických osob, podnikajících fyzických osob a orgánů veřejné moci[7] a další. V praxi to znamená, že PČR využívá jako zdroj informací svou vlastní činnost, například údaje o průběhu tresného řízení, údaje o uložené blokové pokutě. Na druhé straně jsou informace poskytované třetí stranou.

Tyto informace, ale i další, které PČR svou činností získá se snaží ukládát pro budoucí potřeby. Schéma policejních databází se živě vyvíjí. Aktuálně má PČR k dispozici cca 70 informačních systémů. Na základě obsahu informací, které obsahují je můžeme rozdělit na evidence, podpůrné a managerské informační systémy, poznatkové fondy a specializované, laboratorní a expertní informační systémy. Každý informační systém má svůj účel, správce a managera. Tyto informace jsou stanoveny vnitřními předpisy PČR – interními akty řízení, které dále upravují další podmínky provozování systémů jako jsou například způsob zpracování osobních údajů, vydávání dat třetím stranám, pro jaké složky policie je systém určen apod. Přílohou interních aktů řízení bývá i metodika pro uživatele. Metodické pokyny jsou směřovány především pro vkladatele dat do systémů a verifikátory, případně mohou popisovat využití prokročilejších funkcí systému.

[6] Zákon o Policii České republiky., In: 273/2008 Sb. ve znění pozdější předpisů.

[7] Zákon o Policii České republiky., In: 273/2008 Sb. ve znění pozdější předpisů.

1.2.1 Evidence

Evidencí má PČR celou řadu. Některé, jako například NTC již slouží pouze jako zdrojová databáze pro novější evidenci KSÚ, ve většině případů se ale jedná o evidence, do kterých jsou data aktivně přidávána uživateli či automatizovaným způsobem. Automatizovaný způsob umožňuje napojení na další informační systémy, jako je například napojení na evidenci osob.

1.2.1.1 Pátrací evidence

Pátrací evidence jsou u PČR tři. První dvě mají podobný název i účel – PATROS a PATRMV. Systém PATROS je určený k pátrání po hledaných a pohřešovaných osobách, po totožnosti osob a totožnosti neznámých mrtvol a kosterních nálezů. PATRMV je určený k pátrání po motorových a přípojných vozidlech.[8] Oba systémy často používají hlídky v terénu pro zjištění, zda jsou osoby či motorová vozidla v pátrání, od čehož se následně odvíjí jejich další činnost. Poněkud odlišný je systém TELEFOTO sloužící ke sdělení obrazových informací s doplňující textovou položkou. Systém funguje na principu nástěnky, kdy zobrazí např. fotografii neznámého muže při neoprávněném výběru z bankomatu. Policisté si tyto záznamy prohlíží a v případě identifikace osoby kontaktují zodpovědného policistu. Touto cestou lze zajistit i propojení případů páchaných v různých částech republiky.

1.2.1.2 IS Kontrola

Jedná se o jednoduchý systém, který je však svým obsahem velice významný. Jsou zde evidovány kontrolované osoby a motorová vozidla včetně informací o místě, čase a datu kontroly.

[8] CHMELÍK, Jan. Rukověť kriminalistiky. Plzeň: Aleš Čeněk, 2005, 192, 193, 203 a 204., ISBN 80-86898-36-9

1.2.1.3 SEUD

Název je zkratkou plného jména systému – systém evidence uměleckých děl. V systému jsou evidovány odcizené, ale i nalezené umělecké předměty. Systém je určen především pro specialisty z řad služby kriminální policie a vyšetřování (dále jen SKPV), kteří se zabývají trestnou činností páchanou v souvislosti s uměleckými předměty.

1.2.1.4 KSU

Systém KSU byl již zmiňován v úvodu kapitoly 1.2.1. Jedná se o centrální systém obsahující kriminalisticky relevantní události. Obsahem i významem dat se jedná o jeden ze základních systémů PČR pomáhající k odhalování a předcházení trestné činnosti. KSU navazuje na zastaralý systém NTC, jehož data převzal. Systému KSU využívají policisté zařazení na SKPV, zejména analytici.

1.2.1.5 Událost

Událost je systém obsahující hlášení o závažných porušeních veřejného pořádku a dalších událostech spočívajících v protiprávním jednání. Jeho role je částečně i statistického charakteru a umožňuje například sledování kriminality v konkrétním regionu.

1.2.1.6 ETŘ

Systém je zkratkou slovního spojení evidence trestního řízení. Svým obsahem se jedná o jeden z největších systému u PČR. Obsahuje veškeré informace k jednotlivým trestním spisům, vyjma pomocných materiálů, a přestupkům. Umožňuje spolupráci mezi jednotlivými útvary, vytěžování informací pomocí analytické nadstavby, zadávání úkolů v rámci spisu či přeposílání spisů mezi útvary. Obsahuje potřebné formuláře pro jednotlivé úkony včetně předepsaného poučení i administrativních náležitostí, systém je přístupný z jakékohokoliv počítače s připojením do služební sítě. Jednotlivá krajská ředitelství policie (dále jen KŘP),

útvary s celostátní působností (dále jen ÚCP) a policejní prezidium mají svá jednotlivá ETŘ, což způsobuje těžkopádnost při data miningu jednotlivých ETŘ, jelikož neexistuje nástroj pro položení jednoho dotazu do všech ETŘ. Systém je náročný na hardware, neboť obsahuje velké množství dat, ke kterému přistupuje velké množství uživatelů. Přístupy do ETŘ mají prakticky všichni policisté.

1.2.2 Managerské a podpůrné systémy

Tyto systémy slouží ke zvýšení efektivity činnosti PČR, především v ekonomické, logistické a personalistické oblasti. Jsou zde zařazeny i statistické systémy, ze kterých je čerpáno v praktické části diplomové práce.

1.2.2.1 EKIS

Ekonomický informační systém není dílem PČR, nýbrž ministerstva vnitra (dále jen MV). V systému jsou zpracovávány finanční, materiální, personální a mzdové účetnictví. Jsou určeny vedoucím pracovníkům, kteří jej využívají např. k přihlášení na školení podřízených, k evidenci přesčasů atd. V ostatních případech jej policisté nevyužívají a systém je určen především pro specializované pracovníky, jako jsou např. mzdové účetní.

1.2.2.2 EDN

Evidence dopravních nehod je rize statistický systém, jejímž obsahem jsou údaje o dopravních nehodách (zraněné či usmrcené nehody, řidiči pod vlivem alkoholu atd.)

1.2.2.3 ESSK

Centrální evidenčně statistický systém kriminality obsahuje informace o počtu trestných činů, počtu pachatelů trestných činů a ve vymezeném rozsahu i informace o obětech trestných činů. Systém čerpá data ze standardizovaných formulářů, především z evidence trestního řízení. Data jsou volně přístupná i veřejnosti na internetových stránkách. Pro PČR má význam především v možnosti predikovat a

analyzovat vývoj konkrétního typu trestné činnosti a na základě výsledků analýz následně směřovat svou činnost k zamezení páchání trestné činnosti.

1.2.3 Poznatkové fondy

Poznatkové fondy obsahují data získaná především činností policistů přímo v terénu, zejména od policistů zařazených na SKPV. Rozsah informací je široký a obsahuje tzv. operativní informace, informace poskytnuté informátory, výsledky prověřování, výskyt osob atd. Tyto informace jsou navzájem propojené vazbami, které popisují vztah mezi jednotlivými objekty. OBRÁZEK Z ANALYSTU. Poznatkové fondy jsou speciální i v možnosti uvést neúplné informace, např. policista zná pouze přezdívku osoby, ale nezná její pravé jméno. Pokud jsou zjištěny další informace k objektu, lze je doplnit. V poznatkových fondech se využívá systém hodnocení validity informace.

1.2.4 Specializované, laboratorní a expertní systémy

Tyto systémy využívají specialisté zařazeni na kriminalistickém ústavu, OKTE a další.

1.2.4.1 AFIS 2000

Jedná se o elektronicky vedenou sbírku obrazců papilárních linií otisků prstů. Systém obsahuje pouze informace o místě a čase, kdy byly otisky sejmuty, identifikaci osoby, které byly sejmuty, kód útvaru, jenž otisky sejmul a důvod vedení osoby v systému. Samotné otisky papilárních linií prstů v systému nejsou uvedeny a pro potřeby jejich zkoumání musí být vyhledány v archivu.

1.2.4.2 FODAGEN

Systém FODAGEN obsahuje fotografické, daktyloskopické a DNA údaje k osobám. Slouží především pro kriminalistické techniky.

1.2.4.3 SIS

Členské země Schengenské dohody vyhotovily Schengenský Informační Systém – SIS. Účelem systému je získávat a evidovat informace týkající se sledovaných osob a věcí. Jeho praktické využití spočívá zejména při pátrání po osobách nebo věcech, udělování víz apod.

2 Teoretická část

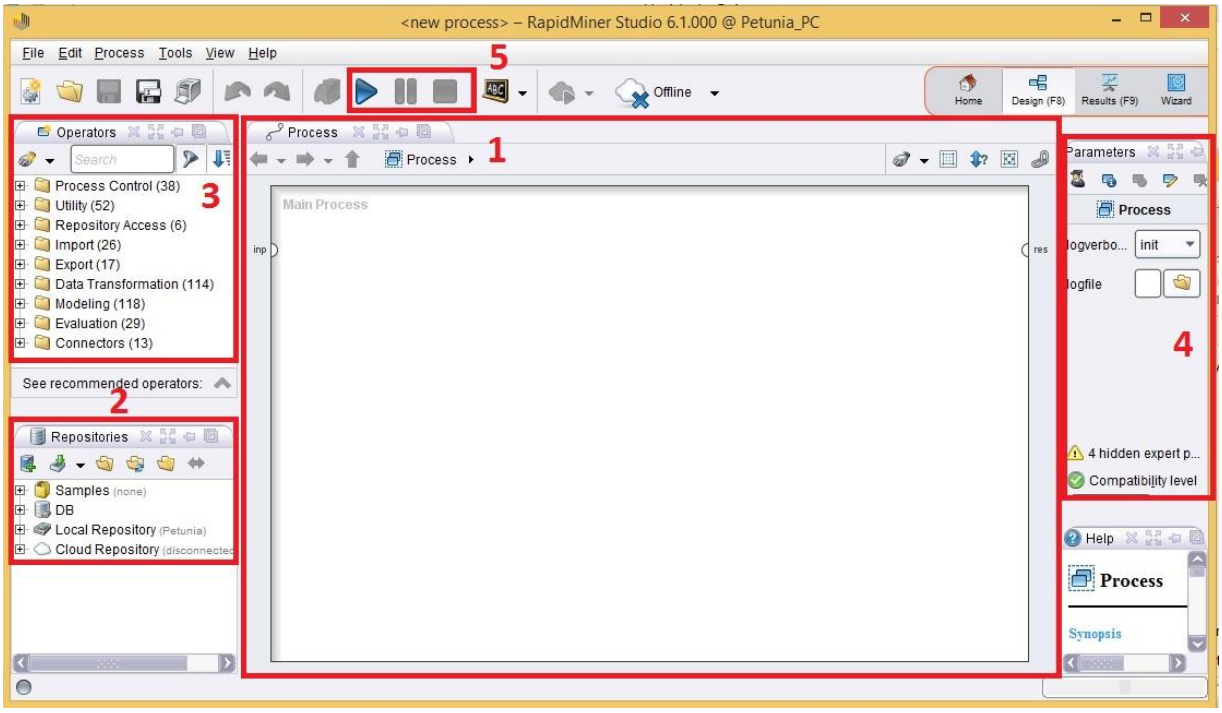
Předkládaná diplomová práce je pojatá jako kvantitativní studie, při níž budou použity vybrané poznatky a metody z oblasti IT. Pozornost je věnována modelování trestné činnosti, konkrétně trestné činnosti krádeže vloupáním do bytů a rodinných domů. Trestnou činnost lze predikovat, záleží však na spoustě faktorů, které jednání pachatelů ovlivní, jako je např. motiv pachatele, počasí, denní doba, roční období zabezpečení objektu apod. Obsáhnout veškeré proměnné, které ovlivňují jednání pachatelů je v prakticky nemožné. Lze pouze s jistou pravděpodobností předpokládat lokalitu a čas spáchání trestného činu a dle toho učinit patřičné kroky k zamezení jeho provedení. Ačkoliv analýze budou podrobena zdrojová data PČR, stále je nutné brát na zřetel, že tyto informace nejsou primárně určeny pro potřeby data miningového zpracování. Výsledky zkoumané metody pak budou pravděpodobně neúplné. Data byla získána ze stránek PČR – www.policie.cz, kde jsou volně dostupná. Ke zpracování těchto dat bude využito software MS Excel a analytický program RapidMiner.

2.1 RapidMiner

Produkt RapidMiner lze získat stažením ze stránek http://rapidminer.com/. Jedná se o software, který je na základě vygenerované licence poskytnutý na 14 dní zdarma. Po uplynutí lhůty je program stále aktivní, ale za zdrojová data už nelze použít databázové zdroje, čímž není dotčena možnost využívat zdrojová data uložená v programu MS Excel. Takto byla získána verze produktu RapidMiner Studio 6.1.000, která byla v empirické části využita pro zpracování dat. Po spuštění programu se zobrazí úvodní obrazovka, která je přiložena jako obrázek č. 1. Hlavní a svou plochou největší část zabírá tzv. Process view. Zde jsou vyobrazeny procesy, které mají být provedeny na zdrojových datech. Pomocí jednoduché metody „drag and drop“ v této části lze celý proces nadefinovat (označeno jako část 1). V levé části programu jsou dvě nejdůležitější části programu. Záložka Repositories (část 2) umožňuje načíst

uložené procesy nebo data. Záložka Operators (část 3) obsahuje přednastavené funkce pro import a export dat, jejich převod, modelování i vyhodnocení. Pravá část programu obsahuje záložku Parametrs (část 4). Tato záložka umožňuje nastavit parametry a kritéria pro jednotlivé operátory. Jakmile je vytvořen celý proces, jeho spuštění se provádí cestou tlačítek Run, Pause a Stop (část 5).

Obr. č. 1 – RapidMiner Studio



2.2 Zdroje informací

Pro maximální pochopení dat, které lze potencionálně využít při vypracování data miningové analyýzy spojenou se schopností predikce kriminality je v této kapitole popsán zdroj dat. Jedná se tedy o data primárně určená pro statistické účely, nejedná se tedy o data určená pro data mining. Data v systému ESSK jsou původem z formuláře o trestném činu. Jedná se o základní dokument, jehož pomocí PČR zaznamenává informace o trestném činu. Takto získaná data obsahují vypovídající hodnotu pro potřeby statistiků, ale neobsahují informace taktického charakteru.

Volná dostupnost dat má za následek jejich častou interpretaci odborné, ale i laické veřejnosti. Především u textů publicistického charakteru dochází k využití dat za účelem potvrzení předem vypracované teorie. Tento postup ve čtenáři často evokuje pocit, že se jedná o věcné, úplné a spolehlivé informace, i když na zkoumaný jev je nahlíženo v úzké rovině souvislostí.

Statistiky o trestné činnosti nevede pouze PČR. Vězeňská služba PČR, Ústavní soud, jednotlivá státní zastupitelství, Probační a mediační služba ČR a další taktéž vedou statistické výstupu. Každá z těchto institucí logicky zpracovává statistiky spojené s jejich činností autonomně, což má logicky za důsledek rozdílné výstupy, které by se ale v určitých oblastech měly shodovat. V ČR neexistuje instituce, která by centrálně řídila statistickou činnost jednotlivých organizací, což by vedlo ke sjednocení pravidel. Nejčastější rozdíly panují při rozhodování, v jaké fázi trestního řízení lze označit osobu za známého pachatele, třídění trestných činů nebo způsob zakomponování změny patřičných právních norem.

Právě na základě informací uvedených v odstavci výše bude zpracovatel vycházet pouze z jednoho zdroje dat. Jelikož jedním z cílů práce je poznání konkrétní trestné činnosti, tímto zdrojem jsou statistické výstupy PČR.

Statistické výstupy PČR mají několik specifik. Zejména je nutné si uvědomit skutečnost, že počet zjištěných trestných činů neodpovídá počtu pachatelů. Je zcela

běžné, že jeden trestný čin spáchá více pachatelů nebo pachatel spáchá několik trestných činů, což právě případ zkoumaných trestných činů. Pakliže pachatel násilím vnikne do bytu nebo rodinného domu a odcizí cennosti, lze předpokládat, že se svým jednáním dopustil vícera trestných činů. Další omezení spočívá v kategorizaci trestných činů. Původně byly jednotlivé kategorie členěny dle trestního zákona, ale s jeho postupnými novelizacemi a nakonec i jeho úplnou rekodifikací, která nabyla účinnosti dne 1. ledna 2010, nové kategorie vznikaly, i zanikaly. Velmi obecné, zato jednoznačné a v praxi často používané je rozdělení na hlavní kategorie:

- Obecná kriminalita
- Hospodářská kriminalita
- Vojenské trestné činy
- Ostatní kriminalita

Kriminalitu lze členit i na základě lokality a časových údajů.

2.3 Cíle výzkumu a jeho vybrané aspekty

Stežejním cílem předkládané práce je otestování vhodnosti získaných dat k jejich využití pro metody data miningu. Získaná data obsahují informace z výše popisovaného systému ESSK.

Druhým cílem práce je navržení takových opatření, která by vedla k zefektivnění budoucí práce s daty. Lze očekávat, že v průběhu práce budou zodpovězeny otázky, které by mohli napomoci při predikci a prevenci zkoumané kriminality. Na základě těchto zjištění může PČR efektivně směřovat svou činnost k zamezení nebo odhalení trestné činnosti. To může mít pozitivní dopad na potenciální oběti, což lze vnímat jako snahu o zlepšení bezpečnostní situace v České republice.

Při samotných analýzách se pak autor zaměří na zjišťování takových informací, které by poukazovaly na systémovost kriminality, vzájemnou provázanost jednotlivých případů či skupin případů. V této situaci je nutné opět připomenout povahu zkoumaných dat. Vzhledem k jejich obecnosti lze zjistit limity, které následně neumožní splnění některého z cílů, především vzájemnou provázanost jednotlivých případů. Z uvedeného vyplývá další úkol práce, který spočívá ve formulaci ideální podobě dat, která mají být využita pro data minigové zpracování.

2.4 Vybrané metodiky data mining

Způsobů, jakým lze pracovat s procesem data mining je několik. Jednotlivé metodiky vznikali v rámci potřeb konkrétních organizací, především z úspěšných projektů těchto organizací. Zmiňované organizace jsou komerčního rázu (firmy SPSS nebo SAS) nebo se jedná o „neutrální“ organizace (Evropského výzkumného projektu).

2.4.1 Metodika 5A

Metodiku 5A nabízí firma SPSS jako svůj pohled na proces dobývání znalostí. Název metodiky je akronymem pro jednotlivé prováděné kroky:

- Assess – posouzení potřeb projektu
- Access – shromáždění potřebných dat
- Analyze – provedení analýz
- Akt – přeměna znalostí na akční znalosti
- Automate – převedení výsledků analýzy do praxe[9]

[9] BERKA, Petr. Dobývání znalostí z databází. Vyd. 1. Praha: Academia, 2003, 22, 23. ISBN 80-200-1062-9.

Prvním krokem je tedy stanovení cílů, příprava na stanovené projekty, vzdělávání personálu, zaměření se na efektivní využití software atd. V druhém kroku je třeba zajistit vhodná data, která budou podrobena analýze. Data mohou pocházet přímo z podniku nebo z veřejně dostupných zdrojů. Třetím krokem je samotná analytická činnost za využití specializovaných nástrojů. Toto je fáze, kdy dochází k samotné přeměně dat na informace a znalosti. Čtvrtý krok mění znalosti nalezené v předchozí fázi na znalosti akční, výsledkem jsou obvykle doplňující otázky. Posledním, pátým krokem je uvedení zjištěných informací do praxe.

2.4.2 Metodika SEMMA

Metodika SEMMA je dílem firmy SAS a její název opět charakterizuje jedtnolivé kroky:

- Sample – vybrání vhodných objektů
- Explore – vizuální exploreace a redukce dat
- Modify – seskupování objektů a hodnot atributů, datové transformace
- Model – analýza dat
- Assess – porovnání modelů a intrepretace

V prvním kroku jsou vybrány vhodné data sety, na kterých bude provedeno samotné modelování. Fáze Explore se zabývá porozuměním dat a dochází k poznání očekávaných i neočekávaných vztahů mezi jednotlivými proměnnými, výsledky jsou vizualizovány. Třetí krok obsahuje metody vedoucí k úpravě dat do formy, která je vhodná pro modelování. Jak vyplývá z názvu čtvrtého kroku, zde dochází k samotnému modelování za využití různých metod s cílem vytvoření modelů, které poskytnou požadovaný výsledek. Při závěrečném kroku dochází k vyhodnocení použitých modelů a zkoumání jejich spolehlivosti a užitečnosti.

2.4.3 CRISP – DM

Pro potřeby diplomové práce byla zvolena metoda CRISP – DM (Cross – Industry Standart Proces for Data Mining). Metodika k CRISP – DM vznikla v rámci Evropského výzkumného projektu. Cílem projektu bylo navrhnout uvinerzální postup (tzv. standardní model procesu dobývání znalostí z databází), který bude použitelný v nejrůznějších komerčních aplikacích.[10] Kromě standardního postupu má CRISP – DM nabízet cestu řešení potencionálních problémů, které se mohou vyskytnout v reálných aplikacích. Na projektu spolupracovali firmy NCR, DaimlerChrysler, ISL a OHRA.

CRISP DM je souhrnná metodologie dobývání znalostí z databází. Její model nabízí návody krok po kroku, úkoly a cíle pro každou část celého procesu. CRISP-DM umožňuje provádět rozsáhlé projekty dobývání znalostí z databází rychleji, efektivněji a méně nákladně prostřednictví osvědčených postupů. Model pomáhá vyhnout se běžných chybám.[11]

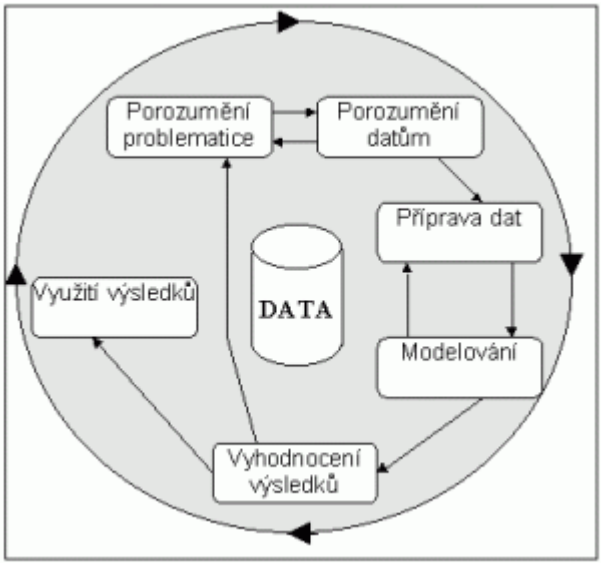
Životní cyklus metodiky je dán cyklem 6 fází, přičemž není pevně stanoveno kterou pořadí jednotlivých částí, viz obrázek č. 1.[12] Na základě výsledku jedné fáze je zvolena další fáze, je tedy nutné jednotlivé kroky neustále vyhodnocovat a zkoumat, k některým fázím cyklu je nutné se i vracet.

[10] BERKA, Petr. Dobývání znalostí z databází. Vyd. 1. Praha: Academia, 2003, 24. ISBN 80-200-1062-9.

[11] Metodologie CRISP-DM: definice. [online]. [cit. 2014-10-28]. Dostupné z: http://www1.osu.cz/studium/dozna/crispdm.htm

[12] BERKA, Petr. Dobývání znalostí z databází. Vyd. 1. Praha: Academia, 2003, s. 24. ISBN 80-200-1062-9.

Obr. č. 2 - Životní cyklus metodiky CRISP - DM



2.4.3.1 Porozumění problematice

Tato zpravidla prvotní fáze se zaměřuje na pochopení cílů úlohy a požadavků na úrovni managerského rozhodování. Na základě vytyčených cílů musí být následně formulovány požadavky na výstupní informace úlohy. Součástí této fáze je i vyhodnocení požadavků na lidské, datové finanční, HW a další zdroje se stanovením předběžného plánu prací.

2.4.3.2 Porozumění datům

Fáze porozumění datům začíná samotným sběrem dat. Následně jsou prováděny činnosti, které umožní získat představu, jaká data jsou k dispozici. Především se jedná o posouzení kvality dat, výběr zajímavých zjištění, zjištění charakteristiky dat atd.

2.4.3.3 Příprava dat

Příprava dat zahrnuje činnosti, které vedou k úpravě dat pro jejich využití při dalších krocích. Nejčastěji se jedná o činnosti selekce, transformace, čištění, formátování a integrování dat. Zpravidla se jedná o nejnáročnější část cyklu.

2.4.3.4 Modelování

Při modelování jsou využity analytické metody – nástroje. Při řešení zadaného úkolu je nutné zvolit vhodnou metodu a postup. Doporučuje se vybrat vícero metod a jednotlivé výsledky konfrontovat. V rámci modelování může řešitel dojít k závěru, že je nutné znovu upravit vstupní data, tedy návrat k fázi úpravy dat.

2.4.3.5 Vyhodnocení výsledků

V této fázi má řešitel k dispozici výsledky, které se zdají relevantní a tyto výsledky je nutné vyhodnotit z pohledu managerů, především vyhodnotit, zda byly splněny všechny cíle formulované při zadání úlohy.

V návaznosti na zjištěné informace lze učinit rozhodnutí, zda opakovat některé fáze, zda v úloze pokračovat nebo projekt ukončit a přesunout se do fáze využití výsledků.

Součástí je i vyhodnocení celého procesu a mělo by být přijato rozhodnutí o způsobu využití zjištěných výsledků.

2.4.3.6 Využití výsledků

Vyhodnocením výsledků řešení úlohy obvykle nekončí. Získané znalosti je nutné upravit do vhodné formy, která bude zákazníkovi nebo nadřízenému předána. Ve výsledku se může jednat o vyhotovení zprávy, která popíše zjištěné informace, ale může obsahovat i návrh na zavedení systému zaručující vyšší efektivitu při zpracování dalších analýz za použití modelu CRISP - DM.

Závěrečným dokumentem je revize celého procesu zisku znalostí. Jednotiví pracovníci, kteří se podíleli na zpracování úlohy by měli na základě individuálních postřehů zaznamenat, jaké fáze měly dobrý či špatný průběh, shrnou zkušenosti, které při řešení nabyly, upozorňují na nebezpečná místa, na zavádějící postupy apod. Cílem závěrečného dokumentu je sdílení zkušeností, které umožní při další úloze pracovat efektivněji.

2.5 Metody dobývání znalostí

Výpočetním jádrem procesu data miningu je použití vhodných analytických metod. Požadovaným vstup jsou data, výstupem pak jsou znalosti.

Všechny používané metody vycházejí z předpokladu, že jednotlivé objekty (příklady, pozorování) lze popsat pomocí charakteristik takových, že objekty patřící k témuž konceptu (do téže třídy) mají podobné charakteristiky (tyto metody bývají proto někdy nazývány učení na základě podobnosti similarity-based learning). Pokud jsou objekty popsány hodnotami atributů, lze je reprezentovat body v n-rozměrném prostoru atributů (příznaků), kde n je počet atributů. Učení na základě podobnosti pak vychází z představy, že objekty představující příklady téhož konceptu vytvářejí jakési shluky v tomto prostoru. Cílem modelování je tedy nalézt vhodnou reprezentaci těchto shluků. Způsob reprezentace znalostí přitom může být značně rozmanitý. Mohou to být reprezentativní příklady-etalony (tak je tomu u metod založených na analogii), mohou to být funkce přiřazené jednotlivým shlukům (to je případ subsymbolických metod), může to být rozdělení prostoru atributů na snadno popsatelné, pravidelné útvary (to je případ metod symbolických).[13]

Následuje výčet nejpoužívanějších analytických metod.

[13] *Metody dobývání znalostí* [online]. [cit. 2014-10-25]. Dostupné z: http://euromise.vse.cz/kdd/index.php?page=metody

2.5.1 Rozhodovací stromy

Způsob reprezentování znalostí v podobě rozhodovacích stromů je dobře znám z řady oblastí. Vzpomeňme jen nejrůznějších „klíčů k určování“ různých živočichů nebo rostlin známých z biologie. Indukce rozhodovacích stromů patří k nejznámějším algoritmům z oblasti symbolických metod strojového učení. Při tvorbě rozhodovacího stromu se postupuje metodou „rozděl a panuj“ (divide and concquer). Trénovací data se postupně rozdělují na menší a menší podmnožiny tak, aby v těchto podmnožinách převládaly příklady jedné třídy.[14]

Použití rozhodovacích stromů pro klasifikaci odpovídá analogii s klíči k určování rostlin nebo živočichů. Od kořene stromu se na základě odpovědí na otázky (umístěné v nelistových uzlech) postupuje příslušnou větví stále hlouběji, až do listového uzlu, který odpovídá zařazení příkladu do třídy.

Rozhodovací stromy jsou vhodné pro úlohy, kde:

- Příklady jsou reprezentovány hodnotami atributů,
- Úkolem je klasifikovat příklady do konečného (malého) počtu tříd,
- Hledaný popis konceptů může být tvořen disjunkcemi,
- Trénovací data mohou být zatížena šumem,
- Trénovací data mohou obsahovat chybějící hodnoty.

Rozhodavací stromy postupně rozdělují prostor atributů nadrovinami rovnoběžnými s osami součadné soustavy. [15]

[14] BERKA, Petr. Dobývání znalostí z databází. Vyd. 1. Praha: Academia, 2003, s. 86. ISBN 80-200-1062-9.

[15] BERKA, Petr. Dobývání znalostí z databází. Vyd. 1. Praha: Academia, 2003, s. 101, 102. ISBN 80-200-1062-9.

2.5.2 Asociační pravidla

V případě asociačních pravidel neni žádný atribut (sloupec tabulky) vyčleněn jako cíl klasifikace. Asociační pravidla hledají „všechny zajímavé“ asociace (implikace, ekvivalence) mezi hodnotami různých atributů. K výše uvedeným (rozhodovacím) pravidlům tak mohou přibýt např. pravidla pro rozhodování vyšetřovatele o návrhu uvalení vazby na obviněného v souladu s trestním řádem.[16]

Jde o hledání vzájemných vazeb (asociací) mezi různými entitami ve zdrojových datech, přičemž není jakákoliv entita upředňostňována.

2.5.3 Rozhodovací pravidla

Zatímco asociační pravidla hledala zajímavé souvislosti mezi hodnotami různých atributů a jejich kombiancí, rozhodovací pravidla se používají stejně jako rozhodovací stromy – pro klasifikaci.[17]

If-then konstrukce nalezneme ve všech programovacích jazycích, používají se i v běžné mluvě. Není tedy divu, že pravidla s touto syntaxí patří, vedle stromů k nejčastěji používaným prostředkům pro reprezentaci znalostí, ať už získaných od expertů, nebo vytvořených automatizovaně z dat.

Jedním z nejznámějších algoritmů pro tvorbu pravidel je algoritmus pokrývání množin pracující metodou „odděl a panuj“. Při pokrývání množin jde totiž o to, nalézt pravidla, která pokrývají příklady téže třídy, a oddělit je od příkladů třídy jiné. Použití těchto pravidel pro rozhodování v úloze je jednoduché. Nalezneme první pravidlo, jehož předpokladům úloha vyhovuje. Závěr tohoto pravidla pak určí, zda vybrat tu či jinou danou úlohu.[18]

---

[16] *POŽÁR, Josef. Manažerská informatika. Plzeň: Aleš Čeněk, 2010, s. 213. ISBN 978-80-7380-276-9.*

[17] BERKA, Petr. Dobývání znalostí z databází. Vyd. 1. Praha: Academia, 2003, s. 139. ISBN 80-200-1062-9.

[18] BERKA, Petr. Dobývání znalostí z databází. Vyd. 1. Praha: Academia, 2003, s. 213. ISBN 80-200-1062-9.

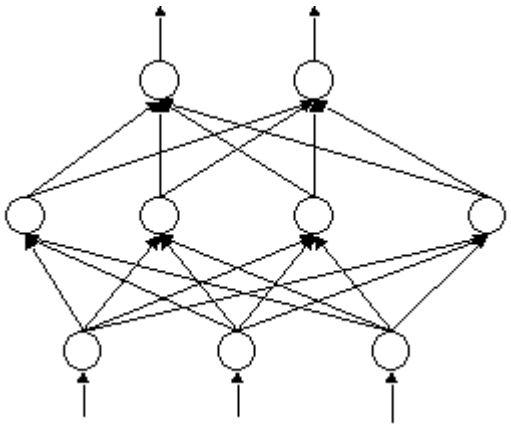
2.5.4 Neuronové sítě

Umělé neuronové sítě vycházejí z analogie s lidským mozkem. Podobně jako mozek jsou tvořeny množstvím navzájem propojených elementů; neuronů. V umělých neuronových sítích je neuron chápán jako buňka, která přijímá podněty od jiných neuronů, které jsou k ní připojeny „na vstupu“. Pokud souhrnný účinek těchto vstupních podnětů překročí určitý práh, neuron se aktivuje a sám začne svým výstupem působit na další neurony. První modely neuronů a neuronových sítí se zkoumaly v rámci umělé inteligence již v 50. letech.

Důležitá (z hlediska dobývání znalostí) je schopnost těchto modelů učit se z příkladů. Na rozdíl od rozhodovacích stromů, nebo rozhodovacích pravidel, kde jsou nalezené znalosti srozumitelné uživateli, v neuronové síti jsou znalosti „rozprostřeny“ v podobě vah jednotlivých vazeb mezi neurony.

Složitější umělé neuronové sítě bývají tvořeny množstvím různě navzájem propojených neuronů. K nejznámějším typům umělých neuronových sítí (používaných pro klasifikaci) patří vícevrstvá síť uvedená na obr. 3.[19]

Obr. č. 3 – *Vícevrstvá neuronová síť*



[19] Metody dobývání znalostí [online]. [cit. 2014-10-25]. Dostupné z: http://euromise.vse.cz/kdd/index.php?page=metody

2.5.4.1 Model jednoho neuronu

Neuron zjednodušeně řečeno, přijímá kladné a záporné podněty od jiných neuronů a ve chvíli, kdy souhrn těchto podnětů překročí daný práh, sám se aktivuje. Výstupní hodnota neuronu je obvykle nějakou nelinerání transformací souhrnu podnětů. Z tohoto pohledu vycházejí matematické modely neuronu.

Důležitou vlastností neuronů je jejich schopnost učit se. Učením se myslí (algoritmus) nastavení vah na základě předložených příkladů tak, aby systém co nejsprávněji zpracovával nebo klasifikoval i neznámé příklady.

Mezi první způsoby učení patří Hebbův zákon z roku 1949. Byl formulován jako model učení na úrovni neuronů v mozku. Vychází z představy, že se posilují ty vazby, které u daného neuronu způsobují jeho aktivaci. V (umělých) neuronech lze toto pravidlo formulovat takto:

Model Adaline používal jiný způsob učení, tzv. gradientní metodu. Zde se vycházelo z požadavku, aby chování sítě bylo co nejvíce podobno celkovému chování učitele, který provádí klasifikace vstupních příkladů trénovací množiny. Zavádí se tedy tzv. střední kvadratická chyba, která má pro n příkladů z trénovací množiny D_{TR} podobu.[20]

2.5.4.2 Model více neuronů

První neuronová síť pochází z roku 1957. Rosenblattův Perceptron byl navržen jako model zrakové soustavy. Perceptron je hierarchický systém tvořený třemi úrovněmi.

První z nich, nazývaná sítnice, slouží k přijímání informace z prostředí. Je tvořena receptory, prvky, jejichž výstup nabývá hodnoty 1 nebo 0 podle toho, zda jsou

[20] BERKA, Petr. Dobývání znalostí z databází. Vyd. 1. Praha: Academia, 2003, s. 157,158, 161. ISBN 80-200-1062-9.

prostředím excitovány nebo ne. Výstupy receptorů jsou (přes náhodně zvolené vazby) přivedeny na asociativní elementy. Asociativní element připomíná adaptivní lineární neuron s tím, že všechny váhy w_j mají pevné hodnoty +1 nebo -1. Asociativní element se aktivuje (vydá hodnotu 1), pokud souhrn všech jeho vstupů překročí zadaný práh. Počet asociativních elementů je řádově desítky tisíc. Výstupy z asociativních elementů jsou náhodně zvolenými vazbami propojeny na reagující elementy, jejichž počet odpovídá počtu tříd, do kterých klasifikujeme. Reagující elementy realizují vážený součet.[21]

Model více neuronů se používá zejména k predikci. Je tvořený více, než jedním neuronem a je typickým představitelem sítě učené učitelem.

Aby neuronová síť správně předpovídala, je nutné ji ale předem „naučit“. To znamená správně nastavit váhy a prahy u jednotlivých neuronů a také určit počet neuronů v jednotlivých skrytých vrstvách. Proces učení probíhá na základě historických údajů o sledované skutečnosti. Neuronová síť v procesu učení postupně prochází každý záznam historických dat. Načte vstupní hodnoty a vypočítá svůj výstup. Tento výstup porovná se skutečnou hodnotou z dat a na základě rozdílu upraví váhy a prahy neuronů. Cílem učení je, aby se výstupy sítě lišily od skutečných hodnot co možná nejméně.[22]

Při učení takové sítě se minimalizuje suma druhých mocnin rozdílů mezi získanými a očekávanými výstupy (známými hodnotami vysvětlované proměnné).[23]

[21] BERKA, Petr. Dobývání znalostí z databází. Vyd. 1. Praha: Academia, 2003, s. 163. ISBN 80-200-1062-9.

[22] PIRKL, David. Neuronové sítě určené pro predikční úlohy. *Data Mining Magazine*, 2003, roč. 1, č. 2, str. 4 – 7. Adastra Corporation

[23] HEBÁK, Petr. *Vícerozměrné statistické metody.* Vyd. 1. Praha: Informatorium, 2004. ISBN 80-7333-025-31.

2.5.5 Statistické metody

Statistika nabízí celou řadu teoreticky dobře prozkoumaných a léty praxe ověřených metod pro analýzu dat. Pro oblast dobývání znalostí z databází mají význam (ať už přímo jako používané metody nebo nepřímo jako zdroj inspirace):

- kontingenční tabulky – pro zjišťování vztahu mezi dvěma kategoriálními veličinami,
- regresní analýza – pro zjišťování funkční závislosti jedné numerické (spojité) veličiny na jiných numerických veličinách,
- diskriminační analýza – pro odlišení příkladů (pozorování) patřících do různých tříd,
- shluková analýza – pro nalezení skupin (shluků) navzájem si podobných příkladů.[24]

2.5.6 Nejbližší soused

V případě nejbližšího souseda jsou koncepty (třídy) reprezentovány svými typickými představiteli. V procesu klasifikace se pak nový příklad zařadí do třídy na základě podobnosti - nejmenší vzdálenosti k reprezentantovi nějaké třídy. Jde tedy o metodu, která vychází ze shlukové analýzy. Klíčovým pojmem je koncept podobnosti, resp. vzdálenosti dvou příkladů.[25]

Protože v případě rozsáhlého souboru by bylo značně časově náročné pro každý nezařazený objekt počítat vzdálenosti od všech zařazených objektů, používá se obvykle pro klasifikaci či předpověď jen výběr (tzv. trénovací množina), případně

[24] Metody dobývání znalostí [online]. [cit. 2014-10-27]. Dostupné z: http://euromise.vse.cz/kdd/index.php?page=metody

[25] Metody dobývání znalostí [online]. [cit. 2014-10-28]. Dostupné z: http://euromise.vse.cz/kdd/index.php?page=metody

mohou být pro klasifikaci využity centroidy vypčítané pro jednotlivé skupiny vytvořené na základě kategorií vysvětlované proměnné. Pro urychlení nalezení k – nejbližších sousedů lze dále využít stromovou strukturu objektů, a to kd – stromy, v nichž jsou nelistové uzly tvořeny proměnnými a listy obsahují seznamy podobných objektů.[26]

2.6 Zdroje dat

Tradičně nejčastější využití počítačů je v oblasti zpracování dat. Neustálý vývoj v oblasti IT vede ke zdokonalování zdrojových dat. Samotná data musí být vyhodnocena, aby měla vypovídající hodnotu pro koncového uživatele. Za tímto účelem byly vytvořeny nástroje – informační systémy nebo programy. Pro jejich správné využití je často nutné zdrojová data upravit do požadovaného tvaru. Nejčastěji jsou informace zpracovány do podoby databáze, statistických zdrojů nebo pro potřeby strojového učení.

2.6.1 Databázové systémy

Původní databázové systémy se skládaly z jednoho souboru. S postupem doby tento soubor nabíral na velikosti i na počtu přístupu uživatelů, kteří využívali v něm obsažená data. V době, kdy nebyl outsourcing, obdobně jako časté zálohování dat, zcela běžná záležitost, měl tento systém jednu zásadní nevýhodu. Pokud některý z uživatelů poškodil zdrojový soubor, byla ztracena veškerá data nebo přinejmenším jejich značná část. Právě tato situace vedla k vytvoření speciálních opatření, které směřovaly k uspokojení potřeb uživatelů. Výsledkem je vznik databázových systémů, včetně řídících procesů.

[26] HEBÁK, Petr. *Vícerozměrné statistické metody.* Vyd. 1. Praha: Informatorium, 2004. ISBN 80-7333-025-31.

2.6.1.1 Relační databáze

Zásadní vývojový stupeň databázových systémů je vznik relačních databází. Tento evoluční stupeň znamená rozdělení jednoho souboru do několika relací (menších tabulek).

Operace nad relační databází vychází z matematické teorie relační algebry, která např. vyvozuje, že tabulka nesmí obsahovat totožné řádky. Jedno pole nebo kombinace více polí záznamu, která slouží pro jednoznačnou identifikaci záznamů tabulky, se nazývá klíč. Spojení záznamů, které jsou uloženy ve dvou různých tabulkách, ale mají stejnou hodnotu klíčů se nazývá relační propojení. Relační propojení umožňuje snadno a operativně kombinovat hodnoty záznamů různých tabulek a tak získávat odpovědi na nejrůznější dotazy.[27]

Relační databáze je tedy tvořena:

- Množinou relací – relace je reprezentována dvourozměrnou tabulkou (řádky odpovídají záznamům, sloupce atributůmm jednotlivé záznamy jsou jednoznačně identifikovány pomocí primárního klíče),
- Operacemi selekce, projekce a spojení pro manipulaci s tabulkami – selekce slouží k výběru záznamů (řádků tabulky), prodejce slouží k výběru atributů (sloupců tabulky) a spojení slouží k propojování tabulek (spojují se řádky se stejnou hodnotou nějakého atributu – obvykle klíče).[28]

V relačních databázích se pro kladení otázek nejprve používali následující dva způsoby. Oba vyvinula v 70. letech firma IBM.

[27] POŽÁR, Josef. *Manažerská informatika.* Plzeň: Aleš Čeněk, 2010, s. 179. ISBN 978-80-7380-276-9.

[28] BERKA, Petr. Dobývání znalostí z databází. Vyd. 1. Praha: Academia, 2003, s. 33. ISBN 80-200-1062-9.

QBE (Query By Example) umožňuje položit dotaz, který je specifikován zadáním vzorku dat do odpovídajících polí. Pro uživatele se jedná o jednoduchý způsob jak položit dotaz. Postačí mu vyplnění předem připraveného formuláře.

SQL (Structured Query Language) je oproti QBE určen uživatelům s většími zkušenostmi. Jedná se o programovací jazyk určený pro práci s daty, ale svými možnostmi je oproti QBR mnohem komplexnější. Uživatel musí znát pravidla syntaxe jazyka, ale velmi detailně strukturu databáze.

Dotazování do databází lze provádět i pomocí EIS (Executive Information Systems). Zmiňované metody QBE a SQL přinášely potřebné výsledky, ovšem nikoliv pro managery.

EIS byl první pokus jak přiblížit dotazování do databáze managerům. Zavádění EIS bylo spojeno se zaváděním osobních počítačů v dané firmě; počítače přestaly být doménou programátorů, objevily se na stolech „prostých“ uživatelů. Základním požadavkem se tedy stalo snadné ovládání.[29]

Výsledkem bylo prostředí, které bylo uchopitelné i pro osoby s menší znalostí dotazovacích nástrojů a samotné databáze. Výsledkem je situace, kdy si uživatel vybral z nabídky dotaz, který doplnil o proměnné a program jej následně vykonal a zobrazil výsledek. Stále byla ale možnost k vytvoření dotazu jazykem SQL. Pokud chtěl uživatel položit „nestandartní“ dotaz, mohl využít i tuto cestu.

Metoda dotazování do databází OLAP (On-Line Analytical Processing) nabídl uživatelům spojení výhod předešlých způsobů dotazování, konkrétně flexibilitu a intuitivní ovládání.

[29] BERKA, Petr. Dobývání znalostí z databází. Vyd. 1. Praha: Academia, 2003, 35. ISBN 80-200-1062-9.

Pro metodu OLAP je typická vizualizace výsledku dotazu. Grafické rozhraní umožňuje uživateli nahlížet na data jak v numerické podobě, tak v podobě nejrůznějších grafů.[30]

Základem OLAP je uspořádání dat do tzv. OLAP krychle. OLAP krychle je způsob organizace dat, který rozšiřuje dvojrozměrně tabulkové uspořádání tak, že každá datová dimenze je uložena v jedné ose kostky. Tím překonává některá omezení relačních databází.[31]

2.6.2 Strojové učení

Jedním z požadavků na označení organizmu za živý organizmus je i schopnost adaptability nebo učení se. Někdy bývá schopnost učit se považována za definici inteligence. Je tedy logické, že se člověk snaží těmito vlastnostmi vybavit i systémy, které vytvořil.

Prvky učení můžeme pod různými názvy nalézt v řadě vědních disciplín; ve statistice se objevuje explorační analýza dat nebo inteligentní analýza dat, v umělé inteligenci se hovoří o metodách rozpoznávání obrazů, či strojového učení nebo automatizovaného získávání znalostí, v kybernetické teorii řízení najdeme adaptivní a učící se systémy, v souvislosti se získáváním znalostí z databází se používá termín dolování z dat.

V zásadě lze rozlišit dva typy učení:

- Učení se znalostem (knowledge acquisition) hledá koncepty, obecné zákonitosti apod.
- Učení se dovednostem (skill refinement) jde o to zdokonalit své schopnosti na základě procvičování nějaké činnosti.

[30] BERKA, Petr. Dobývání znalostí z databází. Vyd. 1. Praha: Academia, 2003, s. 35. ISBN 80-200-1062-9.

[31] OLAP kostka. [online]. [cit. 2014-10-04]. Dostupné z: http://cs.wikipedia.org/wiki/OLAP_kostka.

Během učení si systém vytvoří obecnou reprezentaci jednotlivých typů chování, reps. tříd. Pokud chceme nalezené znalosti používat „ručně“, můžeme tímto krokem skončit. Při automatizovaném používání těchto znalostí se naučenému systému předkládají nové případy a systém se sám rozhoduje.[32]

2.6.3 Statistické databáze

Původním významem statistiky byl pouhý sběr čísel. Název je odvozen z latinského „status“, což lze přeložit jako stát, z čehož vyplývá i oblast dat, kterou se statistika zabývala ve svých ranných fázích vývoje, jednalo se o sběr informací o státu, počtu obyvatel, ekonomice atd. Důvody jsou zcela zřejmé a praktické. Každý vládce chtěl mít přehled o počtu poddaných, kteří mají odevzdávat daně, o počtu bojeschopných vojáků nebo jaký má majetek.

Na počátku 16. století se v Anglii začíná na příkaz lorda kancléře Thomase Cromwella soustavně zaznamenávat narození a úmrtí v církevních matrikách. Věk začínajícího merkantilismu a osvíceného absolutismu urychlil další zkoumání o struktuře obyvatelstva proto, že rychlý růst obyvatelstva byl pokládán za záruku vzkvétajícího hospodářství.[33]

V následujících letech se tato politickoaritmetická statistika změnila na světskou statistiku. Matematici a teoretici nacházeli v běhu života lidské společnosti stále více zákonitostí. Vliv základních teoretických koncepcí pravděpodobnosti na statistiku se začal projevovat stále výrazněji. O rozšíření slova statistika se nejvíce zasloužil německý profesor Gottfried Achenwall, který ho užíval pro označení vědy o tzv. státních pozoruhodnostech, tj. o skutečnostech, určujících sílu tehdejších států.

[32] BERKA, Petr. *Dobývání znalostí z databází*. Vyd. 1. Praha: Academia, 2003, s. 60 a 61. ISBN 80-200-1062-9.

[33] REITEROVÁ, Eva. *Přehled historického vývoje statistiky, její význam v současné době a využití v psychologii* [online]. [cit. 2014-10-06]. Dostupné z: http://publib.upol.cz/~obd/fulltext/psychol8/psychol8-6.pdf. Univerzita Palackého v Olomouci, s. 76.

Šlo o slovní popis státního území, obyvatelstva, armády, zemědělství, obchodu apod., doprovázený číselnými údaji.[34]

Dne 6. března 1897 byl pak zřízen Zemský statistický úřad Království českého, který se stal prvním skutečně statistickým úřadem na území dnešní České republiky. Poprvé byla soustředěna všechna statistická pracoviště, která až do té doby působila v rámci různých ministerstev a dalších institucí. V roce 1909 vyšla první „Statistická příručka království Českého“, další pak následovala v roce 1913. Zemský statistický úřad se v nich snažil podat veřejnosti výbor z nejdůležitějších statistických dat o Čechách, velmi často s několikaletou retrospektivou a v porovnání s obdobnými údaji z Moravy, Slezska a celé monarchie. Devatenácté století bylo dobou prudkého rozvoje průmyslu, což současně kladlo daleko větší nároky na rozsah a kvalitu statistického zjišťování a zpracování statistických dat. Lze říci, že právě tehdy se začala rodit současná tvář statistiky, která je z velké části zjišťováním (makro) ekonomických ukazatelů.[35]

Z historie vyplývá, že statistika je dlouhověký vědní obor. S nástupem výpočetní techniky vzrostla využitelnost a především efektivita při vytěžování dat, s čímž úzce souvisí vznik statistických databází.

Statistika nabízí celou řadu teoreticky dobře prozkoumaných a zdůvodněných a léty praxe ověřených metod pro analýzu dat. Pro oblast dobývání znalostí z databází mají význam:

- Kontingenční tabulky – pro zjišťování vztahu mezi dvěma kategoriálními veličinami,

[34] REITEROVÁ, Eva. *Přehled historického vývoje statistiky, její význam v současné době a využití v psychologii* [online]. [cit. 2014-10-06]. Dostupné z: http://publib.upol.cz/~obd/fulltext/psychol8/psychol8-6.pdf. Univerzita Palackého v Olomouci, s. 76.

[35] Historie statistiky v Čechách do roku 1918. *Český statistický úřad* [online]. 18.1. 2012 [cit. 2014-10-06]. Dostupné z: http://www.czso.cz/csu/redakce.nsf/i/historie_statistiky_v_cechach_do_roku_1918

- Regresní analýza – pro zjišťování funkčních závislostí jedné numerické (spojité) veličiny na jiných numerických veličinách,
- Diskriminační analýza – pro odlišení příkladů (pozorování) patřících do různých tříd,
- Shluková analýza – pro nalezení skupin (shluků) navzájem si podobných příkladů,
- Korelační analýza – pro posouzení, zda je mezi dvěma numerickými veličinami lineární závislost.[36]

[36] BERKA, Petr. *Dobývání znalostí z databází*. Vyd. 1. Praha: Academia, 2003, s. 46. ISBN 80-200-1062-9.

3 Empirická část

Ze zmiňovaných metodologií data miningu byla zvolena CRISP – DM. Tato metodologie nabízí komplexní řešení od přípravy dat, po jejich zpracování až k vyhodnocení výsledků za možnosti opakovat jednotlivé kroky, a to i opakovaně. Obsah empirické části vychází z manuálu CRISP – DM Step-by-Step data mining guide (Chapman, Clilnton, Kerber a kol., 2000). Předepsaným postupem se zpracovatel překládané práce bude řídit a budou zaznamenány veškeré relevantní úkony, které souvisí se zpracování dat a jejich vyhodnocením.

Před samotnou empirickou částí je vhodné upzornit na některé aspekty, které mají vliv na tuto část práce. Mnoho autorů zabývajících se problematikou data miningu upozorňují na zpracování celého procesu ve formě skupinové kooperace. Součástí skupiny by měli být odborníci na jednotlivá témata, jichž se konkrétní témata bezprostředně týkají. Z pohledu diplomové práce tuto častou podmínku nelze splnit, neboť se jedná o samostatnou práci.

Druhý podstatný aspekt se týká povahy dat, se kterými bude autor pracovat. Data byla čerpána z volně dostupného zdroje – z oficiálních statistik PČR. Původně se tedy jedná o data z popisovaného systému ESSK. Jelikož jsou informace veřejně přístupné, neobsahují žádné osobní informace ani jiná data, která by výrazně prospěla při vyhodnocení data miningu v systémech PČR. Získaná data reflektují zejména základní charakteristiku sledované kriminality. Ty lze rozdělit do dvou částí, a to na kvantitativní a kvalitativní. Zcela zásadní kvantitativní charakteristika je stav kriminality. Především jde o údaje popisující počty pachatelů sledované kriminality nebo množství trestných činů. V případě kvalitativní stránky kriminality lze hovořit o její struktuře. Na základě rysů jednotlivých případů se dozvídáme, zda mají jednotlivé případy společné znaky. Kvalitativní znaky kriminality jsou v policejních statistikách v malé míře taktéž obsaženy.

Z kriminalistického hlediska je předpoklad, že žádná trestná činnost nevzniká nahodile a každý případ je ovlivněn různými faktory, který má za následek chování pachatele, poškozených, svědků a dalších zúčastněných osob, ale i zvířat. Tyto faktory však nelze zakomponovat do prováděných zkoumání, neboť jich je celá řada, což může především u predikce trestných činů vést k zavádějící interpretaci výsledných dat.

3.1 Porozumění problematice

Prvotní krok, který umožní pochopit zkoumanou problematiku je vyjasnění toho, jaký výsledek lze očekávat. Je zásadní, aby cíle nebyly od počátku zcela nesplnitelné nebo si navzájem odporovaly. Taktéž je nutné vzít v úvahu veškerá omezení. Požadovaný výstup musí být co nejvíce realistický.

Součástí prvního kroku je sumarizace informací o instituci, jejíž data budou zpracována. Tyto informace musí souviset s vytyčeným úkolem. Data použitá pro potřeby předkládané práce pocházejí z vnitřních zdrojů Policie ČR. Totožná data jsou k dostání cestou internetu, konkrétně na stránkách www.policie.cz. Vnitřní zdroje byly využity, jelikož se jedná o data získaná přímo od zdroje, čímž se snižuje riziko chybovosti.

Druhý krok spočívá v popisu hlavního cíle celého procesu, a to s ohledem na výsledný produkt. Obecně platí, že se vysloví jasně vymezená otázka nebo jako v našem případě několik otázek. Všechny tyto otázky musí směřovat k jednomu cíli, kterým je zisk a popis takových informací, které by potencionálně mohly být použité při prevenci nebo predikci sledované kriminality. Základní otázka tedy zní: Jaké možnosti nabízejí získaná data s ohledem na zpracování metodou data miningu a jsou tato data vhodná i k modelování?

Třetím krokem ve fázi porozumění problematice je vyhodnocení situace. V tomto kroku dochází k upřesnění položených otázek směřující ke konkrétním datům a jejich možnostem.

Data mají povahu několika rozsáhlých excelových tabulek. Konkrétně se jedná o tři soubory obsahující data za tři po sobě jdoucí roky.

V oblasti lidských zdrojů bylo čerpáno z odborné literatury ve všech podstatných oblastech výzkumu (kriminalistika, kriminologie, data mining, analýza dat apod.). Z výše uvedeného výčtu je zřejmé, že se jedná o penzum informací, kterými se zabývá několik vědních oborů. Dle manuálu by měly být jednotlivé problematiky přiřazeny konkrétním specialistům s odpovídajícími znalostmi. Tento krok byl v podstatě vynechán a autor práce se snažil v maximální možné míře získat relevantní informace sám.

Po stránce hardwarového vybavení byl použit běžný notebook s adekvátním výkonem pro prováděné operace, především pro využití software RapidMiner. Další použitý software byly zejména programy sady MS Office – MS Excel a poznámkový blok.

Součástí vyhodnocení situace je i vyhodnocení rizik a omezení, nákladů a přínosů. Typickým příkladem rizika je nejasnost ohledně reálné využitelnosti vybraných dat k dosažení stanovených cílů. Důležité je i vypracování cost and benefits analýzy (dále jen CBA). Podstata CBA spočívá v kvantifikaci veškerých pozitivních i negativní efektů plynoucích z projektu a zdrojů vynaložených na jejich dosažení, převodu těchto nákladů a přínosů na peněžní jednotky a jejich mezičasové agregaci.[37] Přínosy analýzy spočívají v prověření vhodnosti získaných dat k analýze za využití metodiky CRISP – DM, získání informací pro prevenci a predikci sledované kriminality. Největším nákladem je bezesporu časová náročnost celé analýzy. Dalším nákladem je nutnost výkonné počítačové techniky a multioborová povaha analýzy.

Čtvrtý krok obsahuje vymezení cílů, kterých má být analýzou dosaženo. Výsledkem by měl být popis požadovaného výstupu a jeho ověření využitelnosti

[37] Společenská Cost-Benefit Analysis (CBA) [online]. [cit. 2014-10-11]. Dostupné z: http://www.sieber-uchytil.cz/analyza-nakladu-a-prinosu-cba.html

v praxi. Tato část je v případě předkládané práce těžko proveditelná, jelikož není v silách autora zajistit provedení opatření, která by navazovala na výsledky analýzy, tudíž nelze zjistit, k jaké interakci by došlo v reálném světě.

Poslední krok sestává z popisu zamýšleného plánu provedení analýzy, jehož součástí je konkretizace předpokládaných kroků včetně časového harmonogramu. Seznam všech částí je de facto popsán v empirické části diplomové práce. Určení časového harmonogramu trvání jednotlivých částí je značně složité, natož jeho určení předem.

3.2 Porozumění datům

Prvotní krok spočívá v samotném získání dat, se kterými se bude dále pracovat. V tomto případě se jedná o data získaná přímo z oddělení Skupiny zpracování dat. Čerpáno bylo z totožného zdroje, který je umístěn na oficiálním webu PČR. Získaná data tvoří celkem 3 soubory obsahující požadovaná data. Každý soubor obsahuje souhrnné informace za jeden kalendářní rok. Součástí je i 6 souborů obsahující klíč k dekódování statistických záznamů. Takto získaná data už obsahovala pouze informace, které byly nadefinovány při osobním setkání. Součástí získaných dat je i číselník potřebný k dekódování záznamů.

Pro správné pochopení dat je užitečné v této fázi data nahrát a případně podrobit předzpracování. Pokud by se již v této fázi vyskytly problémy, budou zaznamenány. Právě tato část se týká i získaných dat. Po jejich otevření v programu Excel se zobrazil náhled obsahující kódy. Tato situace je zachycena na obrázku č. 3.

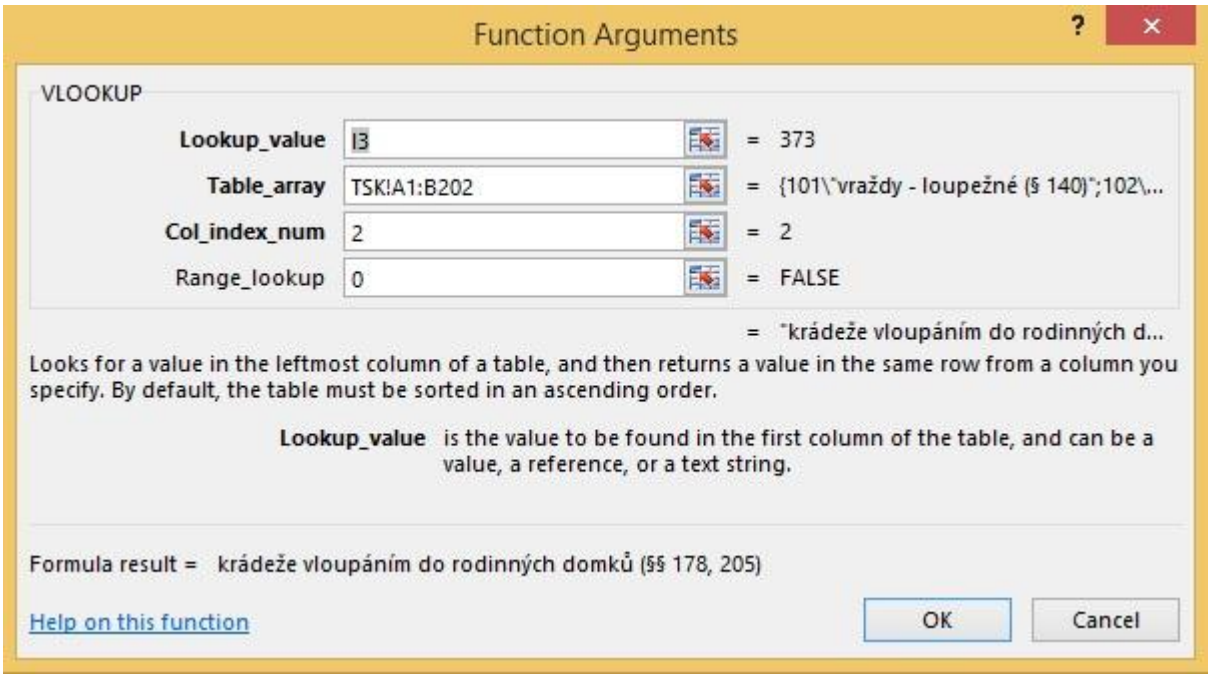
Obr. č. 4 – Pohled na neupravená získaná data

	A	B	C	D	E	F
1	t01_kr	t01_ok	t01_ut	t01_cvs	t01_rok	t01_pc1
2						
3	04	00	00	000000283	11	01
4	04	00	00	000000465	11	01
5	17	00	00	000000506	11	01
6	14	00	00	000000282	11	01
7	02	00	00	000015887	10	02
8	14	00	00	000000525	11	01
9	14	00	00	000000411	11	01
10	03	00	00	000000078	11	01
11	03	00	00	000017695	10	01
12	01	00	00	000000113	11	01
13	02	00	00	000000494	11	01
14	18	00	00	000000316	11	01
15	18	00	00	000000113	11	01

Je evidentní, že získaná data nemají vypovídající hodnotu. Proto muselo dojít k dekódování. Dekódování proběhlo za využití programu MS Excel a jeho funkce VLOOKUP.

Pro správné použití funkce VLOOKUP bylo nutné provést úpravu dat. Neupravená data musela být přeformátována, veškeré hodnoty ve sloupcích musely být převedeny na čísla. Totéž muselo být provedeno se 6 soubory obsahující klíč k dekódování. Zároveň musely být některé sloupce obsahující statistická data sloučeny, aby odpovídaly struktuře dodaných souborů s klíči. Do třech souborů bylo následně vloženo 6 nových listů. Každý list obsahoval klíč k jedné konkrétní části statistických dat. Nyní byly provedeny veškeré úpravy nutné pro spuštění funkce VLOOKUP. Nadefinování zmiňované funkce ukazuje obrázek č. 4.

Obr. č. 5 – Definování funkce VLOOKUP



Funkce VLOOKUP byla na každý soubor s daty použita celkem šestkrát. Výsledný soubor byl ve formátu, kdy sloupec s původními daty střídal sloupec s daty dekódovanými. I tento výsledek vyžadoval úpravu vyčištění dat. Byly tedy odstraněny sloupce s daty, které nebyly dekódovány. Současně byly některé sloupce posunuty tak, aby na sebe co se jejich obsahu týká navazovali. Takto zpracovaná data popisuje obrázek č. 5.

Obr. č. 6 – Výsledek úpravy původních dat

	A	B	C	D	E	F
1	ČTS	stadium_t	druh_TČ_t	kraj_t	okres_t	útvar_t
2	000000283/2011	dokončený tr. čin	trestný čin	ÚSTECKÝ KRAJ	ÚO CHOMUTOV	OOP CHOMUTOV-MĚSTO
3	000000465/2011	dokončený tr. čin	přečin	ÚSTECKÝ KRAJ	ÚO LITOMĚŘICE	OOP ROUDNICE NAD LABEM
4	000000506/2011	dokončený tr. čin	přečin	PARDUBICKÝ KRAJ	ÚO CHRUDIM	OOP SKUTEČ
5	000000282/2011	dokončený tr. čin	přečin	OLOMOUCKÝ KRAJ	ÚO PŘEROV	OOP PŘEROV 2
6	000015887/2010	pokus	přečin	JIHOČESKÝ KRAJ	ÚO ČESKÉ BUDĚJOVICE	OOP BORŠOV NAD VLTAVOU
7	000000525/2011	pokus	přečin	OLOMOUCKÝ KRAJ	ÚO JESENÍK	OOP JAVORNÍK
8	000000411/2011	dokončený tr. čin	přečin	OLOMOUCKÝ KRAJ	ÚO JESENÍK	OOP JAVORNÍK
9	000000078/2011	dokončený tr. čin	přečin	PLZEŇSKÝ KRAJ	ÚO DOMAŽLICE	OOP HOLÝŠOV
10	000017695/2010	dokončený tr. čin	přečin	PLZEŇSKÝ KRAJ	ÚO DOMAŽLICE	OOP DOMAŽLICE
11	000000113/2011	dokončený tr. čin	přečin	STŘEDOČESKÝ KRAJ	ÚO MĚLNÍK	OOP HORNÍ POČÁPLY
12	000000494/2011	dokončený tr. čin	přečin	JIHOČESKÝ KRAJ	ÚO JINDŘICHŮV HRADEC	OOP JINDŘICHŮV HRADEC
13	000000316/2011	dokončený tr. čin	přečin	LIBERECKÝ KRAJ	ÚO JABLONEC NAD NISOU	OOP JABLONEC N/NISOU
14	000000113/2011	dokončený tr. čin	přečin	LIBERECKÝ KRAJ	ÚO JABLONEC NAD NISOU	OOP ŽELEZNÝ BROD
15	000000485/2011	dokončený tr. čin	zločin	JIHOČESKÝ KRAJ	ÚO ČESKÉ BUDĚJOVICE	OOP BORŠOV NAD VLTAVOU

Nyní mají data podobu vhodnou pro orientování se v tom, co jaké hodnoty znamenají. Pro potřeby samotného modelování je však nutné provést další úpravy.

Druhý krok spočívá v popisu dat a jejich vlastností. Popis by měl obsahovat formát dat a jejich objem. Pokud je k dispozici více souborů, je nutné popsat každý zvlášť. Následující popis souborů se vztahuje na soubory po úvodním předzpracování. Předzpracování původní soubory zjednodušilo. Několik sloupců bylo sloučeno a nahrazeno jediným. Na obsahu souboru to nemělo vliv.

Pro potřeby diplomové práce bude dále pracováno se třemi soubory. Všechny soubory byly vytvořeny v programu MS Excel. Níže jsou uvedeny jejich základní vlastnosti:

Vloupani_do_rodinnych_domu_2011[38]: 5257 řádků; 29 sloupců

Vloupani_do_rodinnych_domu_2012[39]: 5479 řádků; 29 sloupců

[38] Příloha č. 1: Vloupani_do_rodinnych_domu_2011

[39] Příloha č. 2: Vloupani_do_rodinnych_domu_2012

Vloupani_do_rodinnych_domu_2013[40]: 6671 řádků; 29 sloupců

Ve výčtu řádků nejsou započteny hlavičky jednotlivých souborů.

Verifikace kvality dat, prozkoumání dat ve smyslu zjištění charakteristiky dat, výběru zajímavých zjištění, chybovosti dat atd. je třetím krokem. Dle manuálu se v této fázi autor zaměřil na rozložení klíčových atributů, na vztahy mezi dvojicemi či malým počtem atributů za použití základních statistických metod. Zjištěné informace musí být zaznamenány, jelikož mohou mít dopad na konečné hypotézy a analýzy.

V této fázi úpravy dat dosud nedošlo k rozlišení dat na základě místa spáchání trestného činu. Získaná data se vztahují na celé území České republiky, což umožní zkoumat sledovanou kriminalitu uceleně.

Kriminalita krádeže vloupáním do bytů rodinných domků je oblast kriminality s nízkou latencí. Důsledkem je, že získaná data poměrně věrohodně popisují reálný stav sledované kriminality. Chybovost dat nebyla zjištěna. Může jít o následek původce zdroje dat, kdy data byla získána přímo od patřičného oddělení PČR nebo bude chybovost objevena v průběhu dalšího pracování s daty. Za chybu by se mohl na první pohled ukazovat sloupec „dat.spách.od“ obsahující datum počátku spáchání trestného činu. Hodnoty v tomto sloupci nabývají datumů i mimo rok, ke kterému se vztahuje konkrétní soubor. Jednotlivé soubory totiž obsahují trestné činy seřazené dle datumu oznámení.

V některých sloupcích chybí hodnoty. Autor předkládané práce tuto situaci taktéž nepovažuje za chybu, neboť např. sloupce „O“ až „T“ označují předmět zájmu pachatele. Každý sloupec označuje jednu konkrétní skupinu věcí (peníze, elektroinstalace, jízdní kola a příslušenství atd.). V případech, kdy nebyl trestný čin dokonaný, nebyl tedy ani zcizen žádný předmět, tudíž u těchto skutků budou řádky „O“ až „T“ prázdné. Taktéž budou prázdné, pokud pachatel odcizil pouze jednu věc

[40] Příloha č. 3: Vloupani_do_rodinnych_domu_2013

nebo více věcí z totožné skupiny věcí. Pak bude zaplněn pouze sloupec „O“ a další budou prázdné.

Soubory obsahující data za jednotlivé roky byly vybrány na základě sloupce „AA“, tedy dle data zahájení úkonů trestního řízení. O zahájení úkonů trestního řízení k objasnění a prověření skutečností důvodně nasvědčujících tomu, že byl spáchán trestný čin, sepíše policejní orgán neprodleně záznam, ve kterém uvede skutkové okolnosti, pro které řízení zahajuje, a způsob, jakým se o nich dověděl.[41] Toto datum neodpovídá datu spáchání předmětného trestného činu, ale vystihuje moment, kdy poškozený podá PČR podnět, oznámí skutečnost o vykradení bytu nebo rodinného domu nebo tuto informaci zjistí PČR svou činností sama. Nahlášení vzniklé situace na stranu PČR v drtivé většině případů nastává až po samotném vykradení, případně v jeho průběhu.

O získaných datech lze říci, že jeden řádek je roven právě jednomu zjištěnému, nahlášenému případu. Z počtu řádků jednotlivých souborů je zřejmé, že počet sledovaných trestných činů má každý rok stoupající tendenci.

V tomto ohledu data z roku 2013 zaznamenaly nejrazantnější nárůst případů. To může být způsobeno vyhlášením Amnestií prezidenta republiky ze dne 1. 1. 2013. V důsledku amnestie bylo z věznic ke dni 4. března 2013 7:00 hod. propuštěno 6 443 vězňů.[42] Toto číslo ovšem není finální. Stále se vedou řízení, která mají určit, zda se konkrétních činů amnestie týká či nikoliv. Amnestie se netýká pouze pravomocně odsouzených, ale i dalších zhruba 14 500 osob podmíněně odsouzených s probačním dohledem a dalších lidí.[43]

---

[41] *Zákon o trestním řízení soudním.* In: 141/1961 Sb. ve znění pozdější předpisů.

[42] *INFORMACE K AMNESTII* [online]. 2.1.2013 [cit. 2014-11-01]. Dostupné z: http://www.vscr.cz/generalni-reditelstvi-19/informacni-servis/amnestie-1681/

[43] Amnestie. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 25.8.2014 [cit. 2014-11-01]. Dostupné z: http://cs.wikipedia.org/wiki/Amnestie

Na základě statistické ročenky kriminality za rok 2013[44], kterou vydalo ministerstvo spravedlnosti je z celkového počtu odsouzených za rok 2011 6,7 % pachatelů označených soudem jako recidivisté. Za rok 2012 je to 6,1 % a v roce 2013 činí recidivisté 5,5 % z celkového počtu odsouzených. Klesající procento recidivistů dle autora není relevantní údaj. Za recidivistu lze v oblasti trestního práva označit osobu, která se opakovaně dopouští páchání trestné činnosti i přes předchozí odsouzení. V ohledu času zde není časové vymezení, kdy ještě osobu při spáchání trestného činu považovat za recidivistu a kdy už nikoliv. Z tohoto důvodu budou procenta recidivistů s postupem času a doplňováním statistik narůstat. Vzhledem k výše uvedenému je bezpředmětné se aktuálně zabývat analýzou recidívy.

Sloupec „B“ obsahuje informace o stadiu trestného činu, ke kterému jsou v ostatních sloupcích další relevantní informace. Po prozkoumání sloupce „B“ lze konstatovat, že ve všech třech zdrojových souborech se nachází pouze dvě hodnoty, konkrétně „dokončený tr. čin“ a „pokus“. Z toho vyplývá, že ve sledovaných letech 2011 až 2013 nebyl dopaden pachatel, který byl ve fázi přípravy trestného činu. Jak je zřejmé ze sloupců „V“ až „Z“, nevětší zastoupení trestných činů má krádež a porušování domovní svobody. V případě trestného činu krádeže je dle § 205 zákona č. 40/2009 Sb. příprava trestná. U trestného činu porušování domovní svobody dle § 179 zákona č. 40/2009 Sb. příprava trestná není.

Druh trestného činu, který je formulován ve sloupci „C“ se dělí do třech kategorií – "Trestný čin“, „přečin“ a „zločin“. Dle trestního zákoníku účinného v době zpracování předkládané práce je trestným činem protiprávní čin, který trestní zákon označuje za trestný a který vykazuje znaky uvedené v takovém zákoně.[45] Trestné činy se dělí na přečiny a zločiny. Přečiny jsou všechny nedbalostní trestné činy a ty

[44] MINISTERSTVO SPRAVEDLNOSTI ČR. *Statistická ročenka kriminality: Příloha 1* [online]. 2013 [cit. 2014-11-01]. Dostupné z: http://cslav.justice.cz/InfoData/servlet/FileServlet?tabulka=ccav_dokument_sestavy&sloupec=obsah_dokumentu_pdf&where=id_dokumentu=743335&typSloupce=pdf&fileName=null

[45] Zákon č. 40/2009 Sb., trestní zákoník. In: 40/2009 Sb. ve znění pozdější předpisů.

úmyslné trestné činy, na něž trestní zákon stanoví trest odnětí svobody s horní hranicí trestní sazby do pěti let. Zločiny jsou všechny trestné činy, které nejsou podle trestního zákona přečiny; zvlášť závažnými zločiny jsou ty úmyslné trestné činy, na něž trestní zákon stanoví trest odnětí svobody s horní hranicí trestní sazby nejméně deset let.[46] Trestní zákoník, ze kterého je citováno nabyl účinnosti dnem 1. ledna 2010. Ovšem dle časové působnosti tohoto zákona se trestnost činu posuzuje podle zákona účinného v době, kdy byl čin spáchán; podle pozdějšího zákona se posuzuje jen tehdy, jestliže to je pro pachatele příznivější. Jestliže se zákon změní během páchání činu, užije se zákona, který je účinný při dokončení jednání, kterým je čin spáchán. Při pozdějších změnách zákona, který je účinný při dokončení jednání, jímž je čin spáchán, se užije zákona nejmírnějšího.[47]

Dřívější právní úprava, konkrétně trestní zákon č. 141/1961 Sb., neobsahovala rozdělení na přečin a zločin, nýbrž pouze pojem trestný čin. Trestné činy spáchané před účinností trestního zákoníku č. 40/2009 Sb., ale na PČR zaevidované v roce 2010 a pozdějším, mají tedy ve sloupci „K“ datum nižší, nežli 1. ledna 2010. Soubor „Vloupani_do_rodinnych_domu_2011.xls“ však obsahuje jeden případ, který i přes výše uvedené má ve sloupci „K“ datum vyšší, než 1. ledna 2010. Tento záznam byl odstraněn, neboť je považován za chybný.

Lokalitu spáchání trestného činu označují sloupce „D“, „E“ a „F“. Lokalita je v tomto souboru dat odvozena od útvaru, který šetří konkrétní případ. Sloupec „D“ zobrazuje kraj, sloupec „E“ okres a nakonec sloupec „F“ obsahuje informace o konkrétním útvaru, který případ šetří.

Dle zákona o Policii České republiky policie působí na území České republiky. Trestní oznámení je možné podat na jakékoliv služebně PČR, která je povinna jej přijmout bez ohledu, zda má předmětný útvar příslušnost tento případ řešit. PČR dále

[46] Zákon č. 40/2009 Sb., trestní zákoník. In: 40/2009 Sb. ve znění pozdější předpisů.

[47] Zákon č. 40/2009 Sb., trestní zákoník. In: 40/2009 Sb. ve znění pozdější předpisů.

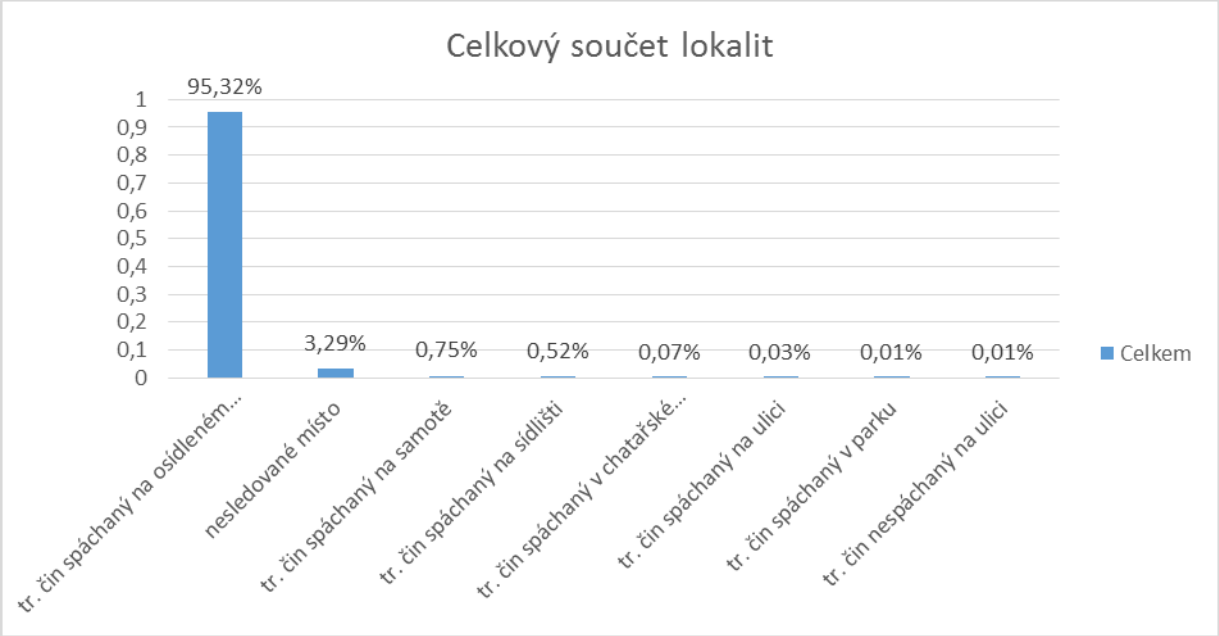
trestní oznámení předá, případně si jej ponechá, místně a věcně příslušnému útvaru PČR k dalšímu postupu. Lze tedy například podat trestní oznámení na neznámého pachatele v Karlových Varech, i když se neznámý pachatel vloupal do bytu v Ostravě. PČR je povinna toto oznámení předat příslušným kolegům do Ostravy, kteří podniknout kroky směřující k odhalení neznámého pachatele.

K porozumění datům bylo v tomto případě využito funkcionality programu MS Excel. Nejprve byla zjištěna v každém ze sledovaných sloupců „D“ až „F“ četnost záznamů jednotlivých lokalit. To bylo učiněno ve všech třech zdrojových souborech.

Místo spáchání trestného činu je uvedeno ve sloupci „G“. Tento sloupec obsahuje informace o místě spáchání trestného činu ve smyslu taktické informace o umístění napadeného objektu. Tyto informace jsou rozděleny do kategorií jako je např. „tr. čin spáchaný na osídleném místě“, „tr. čin spáchaný na samotě“, „nesledované místo“ atp.

Ve sledovaných letech 2011 až 2013 v tomto ohledu nedošlo k výrazné změně. Nejvíce případů bylo spácháno v kategoriích „tr. čin spáchaný na osídleném místě“, „nesledované místo“, „tr. čin spáchaný na samotě“ a „tr. čin spáchaný na sídlišti“. Tyto kategorie činí celkem 99,88 % z celkového počtu zjištěných případů. Největším podílem je zastoupena kategorie „tr. čin spáchaný na osídleném místě“, a to rovných 95,32 %. Následují „nesledované místo“ s 3,29 %, „tr. čin spáchaný na samotě“ s 0,75 % a „tr. čin spáchaný na sídlišti s 0,52 %. Zbývajících 0,12 % z celkového počtu kategorií zabírají kategorie „tr. čin spáchaný v chatařské kolonii“, „tr. čin spáchaný na ulici“, „tr. čin nespáchaný na ulici“ a „tr. čin spáchaný v parku“.

Graf č. 1 – Celkový součet lokalit



Zjištěné informace korespondují s umístěním rodinných domů, které se zpravidla nenacházejí v parku ani v chatařské kolonii. I když jsou rodinné domy umístěny na samotě, značně převyšuje procentuální část domů, které byly napadeny v zástavbě. To může svědčit o drzosti pachatelů, ale i o špatném zabezpečení napadených objektů.

Sloupce „I“ a „J“ obsahují informace o objektu napadení. Jelikož se předkládaná práce zabývá analýzou trestných činů, při kterých dojde k vniknutí do bytů a rodinných domů, bylo nutné tyto sloupce upravit. Sloupec „I“ obsahoval nejen lokality byty a rodinné domy, nýbrž i další objekty jako je např. sklep, garáž, chlév apod. Tyto případy byly ze zdrojových souborů odstraněny. V sloupci „J“ nebyla data, která by byla pro další potřeby upotřebitelná, proto byl celý tento sloupec odstraněn. Po úpravách byly soubory v následujícím tvaru:

Vloupani_do_rodinnych_domu_2011: 4558 řádků; 28 sloupců

Vloupani_do_rodinnych_domu_2012: 4877 řádků; 28 sloupců

Vloupani_do_rodinnych_domu_2013: 6025 řádků; 28 sloupců

Ve výčtu řádků nejsou započteny hlavičky jednotlivých souborů.

Zajímavá odchylka se ukazuje při porovnání dní, ve kterých byly jednotlivé případy nahlášeny. Vzhledem k velkým rozdílům mezi počtem případů v jednotlivých letech, byla zvolena následující metoda porovnání. Za využití funkcí programu MS Excel byla zjištěna četnost nahlášených případů na jeden konkrétní den v týdnu. Ke každému dni byl následně přiřazen číselný atribut pohybující se v celých číslech od 1 do 7. Den s největším počtem nahlášených případů obdržel číselný atribut 1, naopak den s nejmenším počtem nahlášených případů obdržel číselný atribut s hodnotou 7. Takto byly ohodnoceny všechny tři soubory a hodnoty číselných atributů následně sečteny. Výsledná tabulka je uvedena jako obrázek č. 6.

Tab. č. 1 – Četnost dní ohlášení TČ.

DEN V TÝDNU	HODNOTA ATRIBUTU 2011	HODNOTA ATRIBUTU 2012	HODNOTA ATRIBUTU 2013	SOUČET HODNOT ATRIBUTŮ
pondělí	3	2	3	8
úterý	1	5	2	8
středa	4	1	4	9
čtvrtek	5	4	5	14
pátek	2	3	1	6
sobota	6	7	6	19
neděle	7	6	7	20

Z uvedených hodnot vyplývá, že nejčastěji byly skutky nahlášeny v pátek. Následovaly dny pondělí, úterý a středa. Čtvrtek, sobota a neděle už mají výrazně nižší četnost ohlášených trestných činů. Výsledné zjištění lze zobecnit na tvrzení, že kraje pracovního týdne mají vysokou četnost ohlášení, naopak víkendy a dny uprostřed pracovního týdne mají četnost ohlášení výrazně nižší. Státní svátky nebyly v tomto výpočtu zohledněny.

Obdobně bylo postupováno při zpracování četnosti měsíce ohlášení trestného činu. Výsledek je opět zaznamenán jako tabulka:

Tab. č. 2 – Četnost měsíců ohlášení TČ.

MĚSÍC V ROCE	HODNOTA ATRIBUTU 2011	HODNOTA ATRIBUTU 2012	HODNOTA ATRIBUTU 2013	SOUČET HODNOT ATRIBUTŮ
leden	1	7	2	10
únor	6	12	3	21
březen	7	6	1	14
duben	10	9	7	26
květen	11	10	6	27
červen	12	8	9	29
červenec	2	11	10	23
srpen	8	5	11	24
září	9	3	8	20
říjen	5	2	12	19
listopad	4	1	5	10
prosinec	3	4	4	11

Nejvíce zaznamenaných případů bylo v měsících listopad, leden a prosinec. Naopak nejméně případů bylo zaznamenáno v měsících duben, květen a červen. Červenc a srpen mají taktéž nízký počet nahlášených případů. Z výsledků lze odvodit sezónost sledované kriminality, neboť v zimních měsících je výrazně více nahlášených případů, než je tomu v případě jarních, ale i letních měsících.

Dle předpokladů autora práce den s největším počtem nahlášených případů připadá na 1. ledna. Ve sledovaných letech byl tento den pokaždé na pomyslném prvním místě co se týká počtu zjištěných skutků.

Je všeobecně známé, že rok má 365 dní. V důsledku tohoto faktu lze jednoduše spočítat, že v roce 2011 vychází na jeden den průměrně 12,49 zjištěných trestných činů sledované kriminality. V roce 2012 je to 13,36 případů na jeden den a v roce 2013 je to dokonce 16,51. Tabulka č. 2 ukazuje 10 dní v daných letech, ve kterých bylo ohlášeno nejvíce skutků. Dny, které se vyskytly mezi deseti nejčastějšími alespoň ve dvou letech jsou barevně odlišeny. Jedná se o již zmíněný 1. leden, ale i o 19. září, které je ve statistikách v letech 2011 a 2012 vedeno mezi deseti nejfrekventovanějšími.

Tab. č. 3 – Nejčastější dny ohlášení TČ

DNY 2011	ČETNOST
1.1.2011	25
19.1.2011	24
15.7.2011	24
6.10.2011	24
4.1.2011	22
12.2.2011	22
21.12.2011	22
12.8.2011	21
19.9.2011	21
20.9.2011	21

DNY 2012	ČETNOST
1.1.2012	28
6.8.2012	26
28.11.2012	26
4.9.2012	25
19.9.2012	24
15.11.2012	24
26.9.2012	22
14.12.2012	22
21.12.2012	22
5.6.2012	21

DNY 2013	ČETNOST
1.1.2013	53
15.3.2013	34
20.2.2013	33
26.2.2013	33
10.1.2013	30
4.2.2013	28
2.3.2013	27
5.4.2013	27
17.5.2013	27
22.11.2013	27

Sloupce „N“ až „S“ obsahují informace k odcizeným předmětům. Některé případy neobsahují v těchto sloupcích data. Jedná se o případy, kdy nebylo nic odcizeno, nicméně i přes to došlo k trestnému činu.

Vývoj zcizených věcí sledují grafy č. 1 až 3. Bylo vybráno pouze prvních 6 nejčastěji se vyskytujících odcizených věcí. Z povahy dat je zřejmé, že věci nebyly rozepsány jednotlivě, ale byly zařazeny do kategorií. Ve sledovaném období se objekty zájmu pachatelů výrazně nezměnily. Nejčastěji jsou odcizeny věci z kategorií peníze, zlato, zpracované klenoty a šperky, nástroje a nářadí a přenosné počítače, a to uvedeném pořadí. V dalších položkách již dochází k rozdílům. V letech 2011 a 2012 následovaly kategorie fotografické přístroje a kufry, aktovky, kabelky, peněženky a pásky. V roce 2013 došlo ke změně a pachatelé nejčastěji zcizili hodiny a hodinky a fotografické přístroje.

Graf č. 2 – Předměty zájmu za rok 2013

Graf č. 3 – Předměty zájmu za rok 2012

Graf č. 4 – Předměty zájmu za rok 2011

Zájem pachatelů na odcizených věcech lze popsat následovně. Předmětem odcizení jsou nejčastěji věci malých rozměrů, tedy snadno přenositelných. Dalším atributem zcizených věcí je jejich snadné zpeněžení, případně jiné využití, ze kterého má pachatel nebo pachatelé prospěch.

Zhodnocení vybraných aspektů poznání dat

Po získání dat bylo nutné nejprve jejich úpravu ve smyslu jejich dekódování. Nebyly to však jediné úpravy. V průběhu zpracovávání dat vyvstaly nové požadavky na úpravy, v jejichž důsledku byly soubory upraveny do požadovaného tvaru. S takto upravenými soubory se bude pracovat i nadále, nejednalo se o úpravy pouze pro potřeby fáze porozumění datům.

Dle očekávání bylo během této fáze zjištěno několik zajímavých zjištění, které byly popsány. Zcela jistě by se v datech nalezly i jiné zajímavé odchylky nebo souvislosti, nicméně tato fáze je určena především porozumění datům, nikoliv hledáním souvislostí a anomálií.

3.3 Příprava dat

V této fázi již data prošla základními úpravami, předzpracování a jsou připravena k hlavním analýzám. V následujících krocích budou tedy data popsána – jakou mají podobu a jak se dají využít, případně je dále upravit.

Prvním krokem je vhodný výběr dat, která budou použita. Vlastnosti dat musí odpovídat cílům data miningu, musí být kvalitně zpracována a musí odpovídat technickým limitům. V praxi se jedná o zvolení správných sloupců i řádků v jednotlivých souborech.

Sloupce „J“ až „M“ obsahují časové údaje ke konkrétním případům. První dva sloupce obsahují hodinu a datum, kdy byl trestný čin spáchán, případně kdy započalo jeho páchání. Další dva sloupce obsahují hodinu a datum, kdy byl trestný čin dokonán. Pokud není znám datum a čas, kdy byl trestný čin dokonán, jsou tyto pole

prázdné. Nejedná se tedy o chybu, nýbrž konkrétní informace nejsou k dispozici. Tyto časové údaje pochází od oznamovatele trestného činu.

Ačkoliv se čas i datum zdají jako pevně stanovené a tedy jednoznačné, nemusí to být zcela pravdivé. Mnohdy mohou být nepřesné a podle toho je nutné s nimi zacházet při vyhodnocení jakékoliv činnosti s nimi spojené.

Datumové označení bývá zpravidla přesnější, nežli časové. To je z důvodu, že majitel nebo oznamovatel ví, v jaký den opustil objekt. Hodinový údaj už nemusí vědět zcela přesně a může se z jeho strany jednat o odhad, což ale ve zdrojových souborech není zaznamenáno. Zejména hodinový údaj určující konec spáchání trestného činu je matoucí. Tento údaj spíše odpovídá hodině zjištění vniknutí do objektu. Nelze předpokládat, že v tu samou hodinu bylo do rodinného domu vniknuto, ale že k tomu došlo již dříve. Chybí tedy jakási stupnice míry věrohodnosti uvedených hodnot, která by umožnila přesnější vyhodnocení dat.

Jak je zřejmé z kapitoly 3.2. Porozumění datům, některé úpravy dat byly provedeny již v této fázi. Úpravy byly nutné pro správné pochopení dat a zjištění některých zajímavostí. Vzhledem ke stavu dat po těchto úpravách již nebylo zapotřebí dalších úprav a lze přistoupit k fázi modelování.

3.4 Modelování

Pro účel modelování byl vybrán představený produkt s názvem RapidMiner Studio ve verzi 6.1. Po seznámení s produktem bylo nutné nejprve upravená data vložit do databáze. Celkem byly vytvořeny 3 databáze, přičemž každá z databází obsahovala data právě z jednoho souboru. Postup vzniku databází je popsán níže.

Nejprve bylo nutné vyvtvořit prázdné úložiště. Ty byly vytvořeny jednoduše přes kliknutí na pole „Add a connection to a new repository server.“, které se nachází v záložce Repositories, která je na obrázku č. znázorněna jako část 2. Pro udržení

pořádku byly v každém nově vzniklém úložišti vytvořeny dvě složky. Jedna s názvem data, druhá s názvem procesy. Výsledek je znázorněn na obrázku č. 10.

Obr. č. 7 – RapidMiner Studio 6.1, úložiště

Nyní již nic nebrání k tomu, aby byla načtena data do úložiště a tím vznikla databáze. V záložce Repositories je ikona „Import data into an existing repository.“, která slouží k výběru dat a jejich následnému importu. Před samotným importem si musí uživatel dát pozor, aby importoval do správné složky v rámci úložiště. K tomu postačí kliknout nejprve na přílušnou složku a následně na tlačítko importu. Po kliknutí na tuto ikonu následuje výběr souboru obsahují data. V případě předkládané práce se jedná o soubory „Vloupani_do_rodinnych_domu_2011.xls“, „Vloupani_do_rodinnych_domu_2012.xls“ a „Vloupani_do_rodinnych_domu_2013.xls“. Tyto soubory podle totožného postupu naimportovány do jim odpovídajících úložišť a složek.

Při kontrole importovaných dat bylo zjištěno, že došlo k přeformátování datumů. Jelikož se jedná o program vytvářený a vyvíjený americkou společností a jedná se o demoverzi ve které není lokalizace na české podmínky, zobrazuje se

datum ve formátu měsíc/den/rok. Ke změně dat nedošlo, jedná se pouze o změnu zobrazení, viz obrázek č. 11.

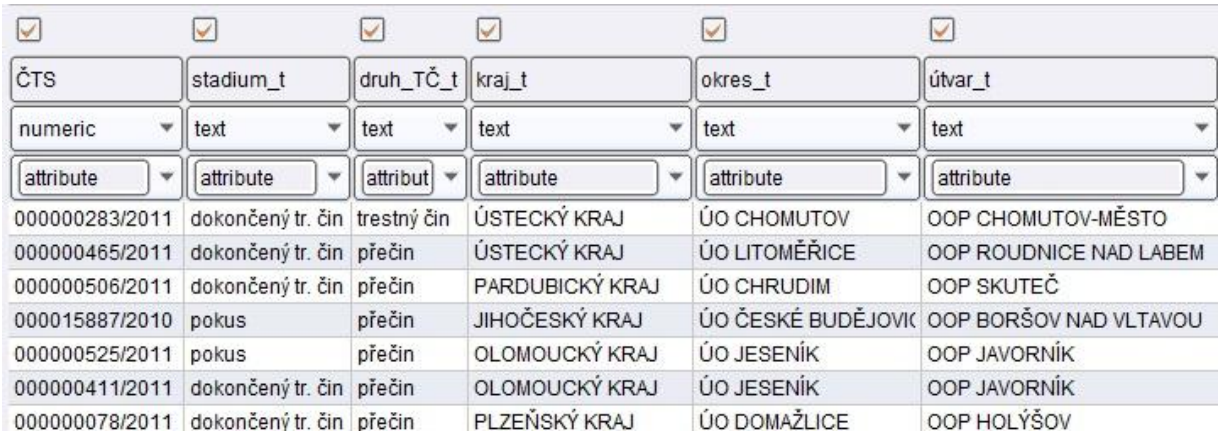
Obr. č. 8 – RapidMiner Studio 6.1, Import dat - změna formátu

Vloupani_do_rodinnych_domu_2011

H	I	J	K	L	M
objekt_obec	objekt1_kon	hod.spách.o	dat.spách.oc	hod.spách.d	dat.spách.dc
soukromý	byty a rodiné	4	11/5/09		
soukromý	byty a rodiné	12	11/20/10	9	1/5/11
soukromý	byty a rodiné	21	1/15/11	8	1/16/11
soukromý	byty a rodiné	7	11/23/10	14	11/23/10
soukromý	byty a rodiné	4	1/8/11		
soukromý	byty a rodiné	18	1/7/11	12	1/9/11
soukromý	byty a rodiné	8	9/5/10	17	9/5/10
soukromý	byty a rodiné	16	12/28/10	15	12/31/10
soukromý	byty a rodiné	19	12/23/10	12	1/2/11
soukromý	byty a rodiné	13	1/11/11	20	1/11/11
soukromý	byty a rodiné	7	1/6/11	15	1/6/11
soukromý	byty a rodiné	9	1/4/11	10	1/4/11

V dalším kroku byla zobrazena tabulka s daty. V záhlaví přibyly dva řádky a zaškrtávací pole. Zaškrtávací pole označuje jaké sloupce má program importovat. První řádek záhlaví označuje názvy sloupců, které převzal ze souboru. Druhý řádek záhlaví označuje typ dat ve sloupci. Ve sloupci „ČTS“ byla změněna hodnota na „numeric“, ve sloupci „škoda_ve_100Kč“ na integer, ve sloupcích „dat.spách.od“, „dat.spách.do“ a „dat.zah.tr.ř.“ byla hodnota změněna na „date“, ve sloupcích „hod.spách.od“ a „hod.spách.do“ byla hodnota změněna na „time“. V poli „dat.ukonč.tr.ř“ bylo ponecháno defaultní nastavení „polynominal“. Ve všech zbývajících sloupcích byly změněny hodnoty na „text“. Třetí řádek záhlaví zobrazuje atributy k jednotlivým slupcům. Tyto hodnoty byly ponechány v defaultním nastavení, tedy na hodnotě „attribute“. Výsledek úprav je pro přehlednost uveden na výřezu obrazovky jako obrázek č. 12.

Obr. č. 9 – RapidMiner Studio 6.1, Import dat – definování sloupců



ČTS	stadium_t	druh_TČ_t	kraj_t	okres_t	útvar_t
numeric	text	text	text	text	text
attribute	attribute	attribut	attribute	attribute	attribute
000000283/2011	dokončený tr. čin	trestný čin	ÚSTECKÝ KRAJ	ÚO CHOMUTOV	OOP CHOMUTOV-MĚSTO
000000465/2011	dokončený tr. čin	přečin	ÚSTECKÝ KRAJ	ÚO LITOMĚŘICE	OOP ROUDNICE NAD LABEM
000000506/2011	dokončený tr. čin	přečin	PARDUBICKÝ KRAJ	ÚO CHRUDIM	OOP SKUTEČ
000015887/2010	pokus	přečin	JIHOČESKÝ KRAJ	ÚO ČESKÉ BUDĚJOVIC	OOP BORŠOV NAD VLTAVOU
000000525/2011	pokus	přečin	OLOMOUCKÝ KRAJ	ÚO JESENÍK	OOP JAVORNÍK
000000411/2011	dokončený tr. čin	přečin	OLOMOUCKÝ KRAJ	ÚO JESENÍK	OOP JAVORNÍK
000000078/2011	dokončený tr. čin	přečin	PLZEŇSKÝ KRAJ	ÚO DOMAŽLICE	OOP HOLÝŠOV

Závěrečný krok importu nabídne kam mají být data uložena a kde vznikne nová databáze. Právě zde budou využity vytvořené složky. V případě dat obsahující případy zjištěné PČR v roce 2011 budou data uložena do úložiště pojmenovaném „2011“ do složky „data“ s pojmenováním databáze „data_2011“. Analogicky s tímto postupem byly vytvořeny i databáze u zbývajících dvou souborů.

Nyní jsou pro potřeby modelování k dispozici tři samostatné databáze. Pro utvočení uceleného pohledu je nutné vytvořit ještě jednu, celkovou databázi. Bohužel software ve verzi a konfiguraci jakou má řešitel úlohy k dispozici nelze sloučit několik databází do jedné, což by výrazně zefektivnilo celý postup. Takto musely být tři zdrojové soubory sloučeny do jednoho souboru a poté dle totožného postupu importovány do databáze pojmenované „11_12_13“.

Pro potřeby modelování byl tento soubor doplněn o jeden sloupec. Tento sloupec obsahuje konvertovaná data zobrazující místo datumu zahájení trestního řízení den v týdnu, který odpovídá datumu. Takto upravený soubor čítá celkem 15460 řádků a 29 sloupců.

Pro ověření dosavadních výsledků, ale i pro ověření, zda byla data správně naimportována byla provedena krátká analýza. Na databázi „11_12_13“ byl vybrán

sloupec „místo_spách._t.“ a byly zvoleny jeho statistické vlastnosti. Výsledek analýzy je v tabulce č. 3.

Obr. č. 10 – RapidMiner Studio 6.1, kontrolní analýza

Index	Nominal value	Absolute count	Fraction
1	tr. čin spáchaný na osídlené	14736	0.953
2	nesledované místo	509	0.033
3	tr. čin spáchaný na samotě	116	0.008
4	tr. čin spáchaný na sídlišti	81	0.005
5	tr. čin spáchaný v chatařské	11	0.001
6	tr. čin spáchaný na ulici	5	0.000
7	tr. čin nespáchaný na ulici	1	0.000
8	tr. čin spáchaný v parku	1	0.000

Tabulka č. 4 zobrazuje výchozí tabulku, ze kterých byl vypočten graf č. 1.

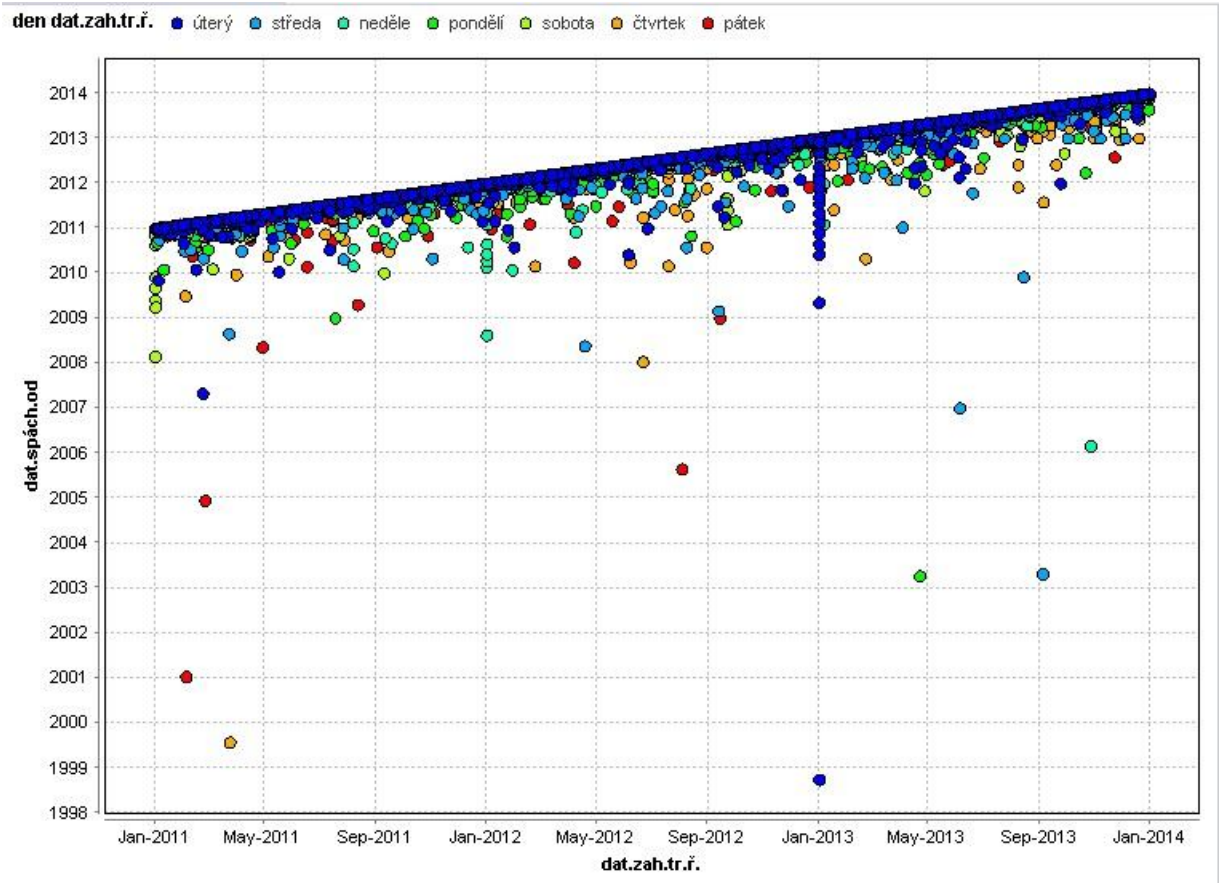
Tabulka č. 4 – Celkový součet lokalit

LOKALITA SPÁCHÁNÍ	ČETNOST
tr. čin spáchaný na osídleném místě	14736
nesledované místo	509
tr. čin spáchaný na samotě	116
tr. čin spáchaný na sídlišti	81
tr. čin spáchaný v chatařské kolonii	11
tr. čin spáchaný na ulici	5
tr. čin spáchaný v parku	1
tr. čin nespáchaný na ulici	1

Výsledky obou měření jsou identické, lze tedy s daty nadále pracovat. Drobná odchylka je u výsledků zaznamenaná programem RapidMiner v posledním poli „Fraction“. Toto pole udává po vynásobení stem procentuální zastoupení jednotlivých položek. Tyto hodnoty jsou odlišné od hodnot uvedených v Grafu č. 1. To bylo zapříčiněno zaokrouhlením na menší počet desetinných míst.

V kapitole 3.2 Porozumění datům je rozvedena okolnost výběru dat na základě datumu, kdy byl trestný čin oznámen nebo zjištěn PČR. Datum zjištění spáchání trestného činu zcela jistě ve všech případech nekoresponduje s datem, kdy byl trestný čin spáchán. Pro zjištění v jaké míře jsou si tato data podobná byla provedena analýza, jejíž výsledek je znázorněn na grafu č. 5.

Graf č. 5 – Analýza závislosti data zahájení trestního řízení na počátečním datu spáchání sktutku

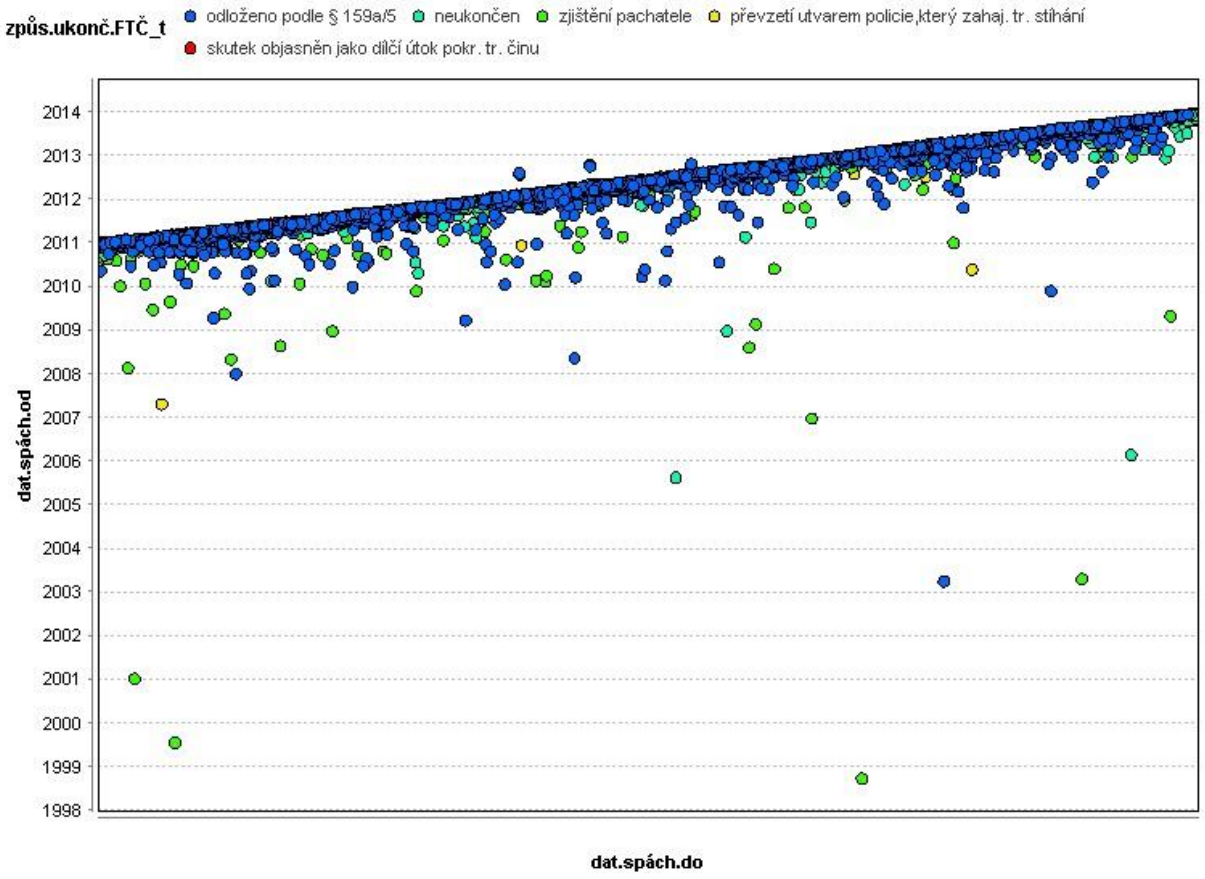


Na osu x byly vneseny hodnoty z pole „dat.zah.tr.ř.“, na osu y hodnoty z pole „dat.spách.od“. Body znázorňující jednotlivé případy jsou barevně odlišeny v závislosti na dni zahájení trestního řízení.

Zobrazené hodnoty ukazují, že v několika ojedinělých případech je rozdíl značný. Dokonce činí několik let. V lednu roku 2011, 2012 a 2013 je zcela zřetelná abnormalita. Ta jen dokazuje předchozí analýzu, kdy vyšlo datum 1. ledna v každém ze sledovaných let s nejvyšším počtem zjištěných případů. Taktéž zcela zjevně ukazuje na skutečnost, že v roce 2013 bylo v den 1. ledna zjištěno dvojnásobné množství případů. Na toto datum připadá i největší výkyv v rozpětí od počátku spáchání trestného činu do data zahájení trestního řízení, kdy se jedná o hodnoty od 1. října 1998 do 1. ledna 2013.

Při zjišťování závislosti objasněnosti na délce trvání trestného činu zobrazil program DataMiner graf č. 6.

Graf č. 6 – Analýza závislosti objasněnosti na délce trvání trestného činu.



Výsledný graf znázorňuje na ose y časový údaj kdy bylo započato páchání trestného činu, osa x znázorňuje časový údaj, kdy bylo páchání trestného činu ukončeno. Dle legendy je pak barevně odlišen způsob ukončení jednotlivých případů.

Trestné činy, které byly páchány po dobu několika let jsou z větší části objasněny, viz zelená barva. Několik tyrkysových bodů označuje případy neukončené, tedy rozpracované a nelze predikovat, zda budou objasněny či nikoliv. Modré body znázorňují případy odložené dle § 159a/5 trestního řádu. To v důsledku znamená případy, které nebyly objasněny.

3.5 Vyhodnocení výsledků

První zjištěné informace, které můžeme považovat za uchopitelné výsledky byly zjištěny ve fázi Porozumění datům, kde byl zjištěn nápad případů sledované kriminality. V této fázi byly dále sledovány abnormality nebo hodnoty vyjímající se standardu.

První zpozorovanou abnormalitou je počet skutků nahlášený v roce 2013. Autor diplomové práce tuto odchylku od běžného růstu sledované kriminality spatřuje jako důsledek vyhlášení Amnestie prezidenta republiky ze dne 1. 1. 2013.

Ve sledovaných letech 2011 až 2013 nebyl dle získaných dat zjištěn jediný pachatel, který byl zjištěn ve fázi přípravy trestného činu, což autor považuje za druhou zachycenou abnormalitu. Jak je již uvedeno výše, příprava je u jednoho z nejčastěji páchaného trestného činu trestná. I přes to není hodnota příprava zaznamenána. Nabízí se tři možné vysvětlení. Policie tuto část statistiky nevede nebo ve sledovaném období opravdu nebyl zjištěn jediný pachatel ve stádiu přípravy trestného činu. Poslední možným vysvětlením je chyba v datech či při dekodovaní dat. Z tohoto důvodu bylo přistoupeno k opětovnému dekodování a nebyla zde nalezena chyba v postupu či špatné manipulaci s daty.

Dále byla zjištěno, že nejvyšší četnost napadených objektů se váže k lokalitě trestný čin spáchaný na osídleném místě. Toto snad ani nelze považovat za abnormalitu, nýbrž za fakt, který odpovídá urbanistickému rozložení rodinných domů v České republice.

Co již zcela jistě lze považovat za abnormalitu, je četnost zjištěné kriminality v jednotlivých dnech v týdnu a v jednotlivých měsících v roce. Zcela jednoznačně je nejčastějším dnem zjištění trestného činu u sledované kriminality pátek. Popisovanou skutečnost mohl způsobit styl života, kdy v pracovních dnech jsou členové rodiny v práci či ve školách a v pátek při návratu zjišťují, že byly vykradeni. V případě nejvíce se vyskytujících měsíců v roce se jedná o listopad, leden a prosinec, což ukazuje na sezónnost sledované kriminality.

Z matematického hlediska se jistě jedná o abnormální rozložení odcizených věcí, nicméně z pohledu pachatele se jedná o zcela logické řešení. Pachatel chce zpravidla místo činu opustit co nejrychleji a s co možná nejhodnotnější kořistí. Odcizené věci navíc chce nejsnadnějším způsobem zpeněžit. K těmto účelům samozřejmě nejvíce odpovídá finanční hotovost, zpracované zlato, šperky, nástroje, nářadí, ale i drobná elektronika. Přesně tyto hodnoty jsou ve sledovaných letech pozorovány jako nejčastěji zcizené.

Průběh fáze porozuměním datům skutečně přineslo několik zajímavých skutečností, které je užitečné zkoumat vzhledem k poznání sledované kriminality. Hlubší využití dat pocházejících ze systému ESSK je však limitováno. Vvyužití k predikci a pokročilejším analýzám se jeví jako poněkud povrchní, jelikož data ze systému ESSK k tomuto nejsou primárně určena a tím pádem chybí další přidané hodnoty. Pro komplexnější analýzy jsou využitelné zejména časové a lokalizační údaje, které byly využity při fázi modelování.

Ve fázi modelování byl použit program analytický nástroj RapidMiner, který svýmu funkcemi umožňuje mnohem širší využití při vyhodnocování dat.

V výsledkem jsou data, ze kterých je zřejmé, že pokud má PČR v rámci sledované kriminality dostatek času, dojde s vysokou pravděpodobností k objasnění případu. Tato analýza zároveň potvrdila výsledky z fáze Porozumění datům, ve které mimo další výsledky vyšel ve všech sledovaných letech den připadající na datum 1. ledna jako dny s nejvyšším počtem zjištěných skutků za jeden konkrétní rok.

3.6 Využití výsledků

Závěrečným produktem analýzy za použití metodologie CRISP – DM je dokument, který bude předán zákazníkovi nebo nadřízenému. Tento krok v předkládané práci nebude vynechán, pouze bude zvolena vhodná forma, tedy shrnutí celého procesu. V předchozích fázích byla data prozkoumána, upravena a nakonec i analyzována. Dle manuálu následuje vytvoření plánů na využití zjištěných výsledků.

Zjištěné závěry jsou výsledkem využití metody CRISP – DM a programu RapidMiner Studio 6.1 na statistických datech pocházejících z ESSK. Zjištěné výsledky lze využít především v oblasti strategických analýz.

4 Závěr

V této diplomové práci se autor zabýval hloubkovou analýzou dat PČR za využití metodiky CRISP – DM. Výsledky ukazují na využitelnost metody CRISP – DM, ale i na využitelnost programu RapidMiner Studio 6.1. Využití metody CRISP – DM je zcela vhodné. Její metodologie, zejména opakování a vracení se k jednotlivým krokům, umožňuje využití jak začínajícím, tak i pokročilým a zkušeným analytikům. Z pohledu využití software RapidMiner Studio 6.1 lze taktéž konstatovat jeho možné využití při data miningu v systémech PČR. Značně pokročilé funkce umožní i velmi komplikované analýzy. Jeho další nespornou výhodou je licenční politika, která umožňuje i využití neplacené verze produktu s omezením využitých zdrojů. Nelze načíst zdroje z databází, nicméně zdroje pocházející z programového prostředí MS Excel načíst lze. Konkrétně byly testovány zdroje s koncovkou „*.xls“, „*.xlsx“ a *.csv“.

V průběhu práce bylo několikrát zmíněno, že data pro tuto práci pocházejí ze statistického systému ESSK, tedy ze systému primárně určeného pro vyhodnocení statistických údajů. Tyto data jsou veřejně přístupná z webových stránek PČR.

Obě zmíněná fakta o datech mají samozřejmě vliv na jejich využití. Z výsledků praktické části je patrné, že obsahují informace upotřebitelné pro data minigové využití. Nicméně možnost vyhotovení analýz s reálným využitím v praxi shledává autor práce jako málo pravděpodobné. Získaná data jsou příliš obecná, vyjma části týkající se datumů a hodin. Například místo spáchání rozdělené na kraje, okresy a útvar, který případ šetří je nedostačující. K řádnému využití je nutné znát přesnou lokalitu spáchání trestného činu. Z těchto podkladů lze například vypracovat tzv. mapy kriminality, které zobrazují četnost výskytu sledované kriminality a zobrazuje ji přímo v mapě, podobně jako u zobrazení teplotních rozdílů v mapě. Získaná data lze využít k data miningu, především v souvislosti s časovými údaji, informacemi o lokalitě, kde byl trestný čin spáchán, předmět zájmu pachatelů a výše škody.

Na základě zaměření práce byla zvolena data obsahující informace vztahující se k sledované kriminalitě. Tyto informace byly dostupné pouze ze systému ESSK. Data byla nejprve dekódována a následně byla zpracována metodikou CRISP – DM.

Uvedené výsledky lze využít při predikci i prevenci. Především v oblasti zjištěných dní a měsíců, kdy byla sledovaná kriminalita nejvyšší. Jedná se pouze o jeden krok, samotné výsledky je nutné podrobit další analýze, která by zdůvodnila právě tyto výsledky a samozřejmě je nutné provést další analýzy, např. vztažené ke konkrétním lokalitám.

Součástí předkládané práce je také příloha, která obsahuje zdrojové soubory, soubory nutné k dekódování. Instalační soubor programu RapidMiner Studio 6.1.000 je k dispozici ke stažení na http://rapidminer.com/.

Seznam použitých pramenů a literatury:

BERKA, Petr. *Dobývání znalostí z databází.* Vyd. 1. Praha: Academia, 2003, ISBN 80-200-1062-9

POŽÁR, Josef. Manažerská informatika. Plzeň: Aleš Čeněk, 2010, ISBN 978-80-7380-276-9

HEBÁK, Petr. *Vícerozměrné statistické metody.* Vyd. 1. Praha: Informatorium, 2004. ISBN 80-7333-025-31.

CHMELÍK, Jan. Rukověť kriminalistiky. Plzeň: Aleš Čeněk, 2005, ISBN 80-86898-36-9

Berry, Michael J. A. – Linoff, Gordon. *Data mining techniques: for marketing, sales and customer support,* 1. vydání, ISBN 0-471-17980-9

PIRKL, David. Neuronové sítě určené pro predikční úlohy. *Data Mining Magazine*, 2003, roč. 1, č. 2, Adastra Corporation

Zákon o *Policii České republiky*., In: 273/2008 Sb. ve znění pozdější předpisů

Zákon o trestním řízení soudním. In: 141/1961 Sb. ve znění pozdější předpisů

Zákon č. 40/2009 Sb., *trestní zákoník*. In: 40/2009 Sb. ve znění pozdější předpisů

BÍNOVÁ, Dagmar. *Vvyužití vybraných statistických metod při zpracování dat technikami Data mining* [online]. Praha, 2006 [cit. 2014-10-28]. Dostupné z: www.pef.czu.cz/cs/?dl=1&f=12934. Disertační práce. Česká zemědělská univerzita v Praze. Vedoucí práce Doc. RNDr. Bohumil Kába, CSc.

Data mining [online]. Dostupné z: http://cs.wikipedia.org/wiki/Data_mining [cit. 2014-07-10].

Data mining [online]. Dostupné z: http://axpsu.fpf.slu.cz/~sos10um/trendy/DM.pdf [cit. 2014-07-10].

Metodologie CRISP-DM: definice. [online]. Dostupné z: http://www1.osu.cz/studium/dozna/crispdm.htm [cit. 2014-10-28].

Metody dobývání znalostí [online]. Dostupné z: http://euromise.vse.cz/kdd/index.php?page=metody [cit. 2014-10-25].

OLAP kostka. [online]. Dostupné z: http://cs.wikipedia.org/wiki/OLAP_kostka. [cit. 2014-10-04].

REITEROVÁ, Eva. *Přehled historického vývoje statistiky, její význam v současné době a využití v psychologii* [online]. Dostupné z: http://publib.upol.cz/~obd/fulltext/psychol8/psychol8-6.pdf. Univerzita Palackého v Olomouci, [cit. 2014-10-06].

Historie statistiky v Čechách do roku 1918. Český statistický úřad [online]. Dostupné z: http://www.czso.cz/csu/redakce.nsf/i/historie_statistiky_v_cechach_do_roku_1918 [cit. 2014-10-06].

Společenská Cost-Benefit Analysis (CBA) [online]. Dostupné z: http://www.sieber-uchytil.cz/analyza-nakladu-a-prinosu-cba.html [cit. 2014-10-11].

INFORMACE K AMNESTII [online]. 2.1.2013 Dostupné z: http://www.vscr.cz/generalni-reditelstvi-19/informacni-servis/amnestie-1681/ [cit. 2014-11-01].

Amnestie. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 25.8.2014 Dostupné z: http://cs.wikipedia.org/wiki/Amnestie [cit. 2014-11-01].

MINISTERSTVO SPRAVEDLNOSTI ČR. *Statistická ročenka kriminality: Příloha 1* [online]. 2013 Dostupné z: http://cslav.justice.cz/InfoData/servlet/FileServlet?tabulka=ccav_dokument_sestavy&sloupec=obsah_dokumentu_pdf&where=id_dokumentu=743335&typSloupce=pdf&fileName=null [cit. 2014-11-01].

Seznam obrázků:

Seznam grafů:

Seznam tabulek: