TECHNICKÁ UNIVERZITA V LIBERCI
Ekonomická fakulta

Využití data miningových postupů v prevenci kriminality

Diplomová práce

Studijní program: N0688A140016 Systémové inženýrství a informatika
Autor práce: **Bc. Daniel Rytina**
Vedoucí práce: Ing. Marián Lamr, Ph.D.
Ústav mechatroniky a technické informatiky

Liberec 2022

TECHNICKÁ UNIVERZITA V LIBERCI
Ekonomická fakulta

Zadání diplomové práce

Využití data miningových postupů v prevenci kriminality

Jméno a příjmení: **Bc. Daniel Rytina**
Osobní číslo: E20000312
Studijní program: N0688A140016 Systémové inženýrství a informatika
Zadávající katedra: Katedra informatiky
Akademický rok: **2021/2022**

Zásady pro vypracování:

1. Rešerše využívání dat v prevenci kriminality ve světě.
2. Algoritmy, postupy a nástroje využitelné k prevenci kriminality.
3. Analýza vhodných veřejně dostupných datových zdrojů vedoucích k rozšíření datové matice o kriminalitě.
4. Příprava datové matice, analýza dat a realizace případové studie.
5. Zhodnocení výsledků a formulace doporučení.

Rozsah grafických prací:
Rozsah pracovní zprávy: 65 normostran
Forma zpracování práce: tištěná/elektronická
Jazyk práce: Čeština

Seznam odborné literatury:

- WITTEN, I. H. a I. H. WITTEN, ed., 2017. *Data mining: practical machine learning tools and techniques*. Fourth Edition. Amsterdam: Elsevier. ISBN 978-0-12-804291-5.
- PETR, Pavel, 2014. *Metody Data Miningu*. Pardubice: Univerzita Pardubice. ISBN 978-80-7395-872-5.
- KEYVANPOUR, Mohammad Reza, Mostafa JAVIDEH a Mohammad Reza EBRAHIMI, 2011. *Detecting and investigating crime by means of data mining: a general crime matching framework*. Procedia Computer Science [online]. 3, 872–880. ISSN 18770509. Dostupné z: doi:10.1016/j.procs.2010.12.143
- WENDLER, Tilo a Sren GRTTRUP, 2016. *Data mining with SPSS modeler: theory, exercises and solutions*. New York, NY: Springer Berlin Heidelberg. ISBN 978-3-319-28707-2.
- PROQUEST, 2021. *Databáze článků ProQuest* [online]. Ann Arbor, MI, USA: ProQuest. [cit. 2021-09-26]. Dostupné z: http://knihovna.tul.cz

Konzultant: Ing. Pavel Tyl, odborný asistent FM TUL

Vedoucí práce: Ing. Marián Lamr, Ph.D.
Ústav mechatroniky a technické informatiky

Datum zadání práce: 1. listopadu 2021
Předpokládaný termín odevzdání: 31. srpna 2023

L.S.

doc. Ing. Aleš Kocourek, Ph.D.
děkan

Ing. Petr Weinlich, Ph.D.
vedoucí katedry

V Liberci dne 1. listopadu 2021

Prohlášení

Prohlašuji, že svou diplomovou práci jsem vypracoval samostatně jako původní dílo s použitím uvedené literatury a na základě konzultací s vedoucím mé diplomové práce a konzultantem.

Jsem si vědom toho, že na mou diplomovou práci se plně vztahuje zákon č. 121/2000 Sb., o právu autorském, zejména § 60 – školní dílo.

Beru na vědomí, že Technická univerzita v Liberci nezasahuje do mých autorských práv užitím mé diplomové práce pro vnitřní potřebu Technické univerzity v Liberci.

Užiji-li diplomovou práci nebo poskytnu-li licenci k jejímu využití, jsem si vědom povinnosti informovat o této skutečnosti Technickou univerzitu v Liberci; v tomto případě má Technická univerzita v Liberci právo ode mne požadovat úhradu nákladů, které vynaložila na vytvoření díla, až do jejich skutečné výše.

Současně čestně prohlašuji, že text elektronické podoby práce vložený do IS/STAG se shoduje s textem tištěné podoby práce.

Beru na vědomí, že má diplomová práce bude zveřejněna Technickou univerzitou v Liberci v souladu s § 47b zákona č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších předpisů.

Jsem si vědom následků, které podle zákona o vysokých školách mohou vyplývat z porušení tohoto prohlášení.

6. května 2022 Bc. Daniel Rytina

Anotace

Tato práce se zabývá využitím data miningových postupů v oblasti prevence kriminality. v první části práce krátce představuje data mining jako vědní obor. Následuje část, která se věnuje oblasti prevence kriminality v několika vybraných zahraničních státech. Jsou zde popsány postupy a algoritmy, které se v různých státech světa používají za účelem prevence kriminality. Vedle postupů jsou také zmíněny softwarové nástroje, které policejní orgány těchto států využívají. Nalezené algoritmy a softwarové nástroje jsou poté představeny blíže. Další část se věnuje zkoumanému datasetu, který pochází z reálného prostředí České republiky, konkrétně z okolí Liberce. Nejprve je věnována pozornost způsobům rozšíření základního datasetu, poté se autor práce věnuje ověřování některých předpokládaných anomálií. Ke konci práce je využito některých popsaných algoritmů k nalezení možných hlubších vztahů.

Klíčová slova: data mining, kriminalita, deskripce, algoritmus, analýza

Annotation

This thesis deals with the use of data mining techniques in the field of crime prevention. The first part of the thesis briefly introduces data mining as a scientific discipline. This is followed by a section that examines the field of crime prevention in a few selected states. a larger section is devoted to foreign countries. The procedures and algorithms used in different countries around the world for crime prevention are described. In addition to the procedures, the software tools used by the police authorities of these countries are also mentioned. The algorithms and software tools found are then presented in more detail. The next section is devoted to the dataset under study, which comes from the real environment of the Czech Republic, specifically from the surroundings of Liberec. First, attention is paid to the ways of extending the basic dataset, then a significant part of the paper is devoted to the verification of some assumed anomalies. Towards the end of the paper some of the described algorithms are used to find possible deeper relationships.

Keywords: data mining, crime, description, algorithm, analysis

Poděkování

Zde bych rád poděkoval panu inspektorovi Bc. Lukáši Porubovi za cenné rady, věcné připomínky a nápady, které vedly k obohacení mé diplomové práce. V neposlední řadě mu velice děkuji za projevenou ochotu a pomoc při zpracování této práce.

Také bych rád poděkoval panu Ing. Bc. Mariánu Lamrovi, Ph.D. za vedení této práce.

OBSAH

Seznam obrázků

Seznam grafů

Seznam tabulek

Seznam zkratek

CAU	Census Area Unit
CCIS	The Crime Criminal Information System
CHI	Crime Harm Index
CRISP-DM	Cross-Industry Standard Process for Data mining
CSV	Comma-Separate Values
DBSCAN	Density-based Spatical Clustering of Applications with Noise
FBI	Federal Bureau of Investigation
GA	Genetic Algorithm
GNU	General Public License
GPS	Global Positioning Systém
HDBSCAN	Hierarchical Density-based Spatical Clustering of Applications with Noise
KNN	K-Nearest Neighbor
MAE	Mean Absolute Error
MP	Městská policie
NCRB	The National Crime Records Burreau
OSM	Open Street Maps
PBF	Protocolbuffer Binary Format
PČR	Policie České republiky
PLI	Priority Location Index
PRECOBS	Pre Crime Observation Systém
TČ	trestná činnost
XML	Extensible Markup Language

Úvod

Vše, co člověk vidí, slyší, cítí nebo ochutná je nějakým druhem informací. Člověk, potažmo lidský mozek, jakožto nejvýkonnější počítač na světě, tyto informace zpracovává velmi houževnatě. Sám o sobě je schopný nalézt a pochopit mnoho dodatečných informací z těch primárních. Počítač je to sice nejvýkonnější, ale ne dokonalý, a proto je potřeba mu v lecčem pomoci. Informací je někdy až příliš velké množství, aby na jejich zpracování a následné vyvození výsledků stačil mozek sám. Za tímto účelem je třeba informace v první řadě zaznamenat. Tyto zaznamenané informace se dají nazývat data.

Pokud jsou mozku předložena data bez dokonalého vysvětlení, pravděpodobně si z nich nic sám neodvodí. Obzvlášť, pokud je dat velký objem. v tomto ohledu může pomoci právě výpočetní technika. Vědní obor, který se zabývá využitím výpočetní techniky v oblasti zpracování dat, se nazývá data mining. Jde o širokou oblast, která nabízí celou řadu způsobů a postupů, jak data zpracovat, pochopit a využít.

Jednou z oblastí, ve které se data mining začíná také využívat, je právě oblast kriminality. Kriminalita je nezákonná a amorální oblast lidského jednání s cílem poškodit druhou osobu, případně obohatit jedince na úkor druhého jedince. Pro civilizovanou společnost je tedy žádoucí kriminalitu omezit na co nejnižší možnou míru. Jedním ze způsobů, jak toho dosáhnout, je kvalitní prevence. Aby se něčemu mohlo předcházet, je nejprve potřeba dokonale pochopit příčiny. Právě zde nastupuje data mining.

V různých státech světa je přístup k prevenci kriminality odlišný, včetně přístupu v České republice. Úkolem této práce je seznámit zevrubně s použitím data miningových metod v prevenci kriminality v několika vybraných státech. Zjistit, co se v rámci prevence používá a jakým způsobem. Nabyté znalosti jsou pak použity na reálných datech z České republiky, konkrétně z oblasti Liberce. S odborníky byly diskutovány některé možné případy, které je z dat možné zjistit.

Jedním z hlavních cílů práce je poskytnout náhled na data z různých pohledů. Data jsou nejprve obohacena o další dostupná data, která by mohla vést k lepšímu vysvětlení vztahů v datech. Jsou zkoumány spojitosti z hlediska zeměpisné polohy, času či počasí. Některé zjednodušující teze byly po diskusi s odborníky vyvráceny, a proto nebyly dále zkoumány (vliv veřejného osvětlení není plošný, ale ovlivňuje pouze určité druhy kriminality, vandalismus není vázán jen na rezidenční oblasti apod.), avšak naopak byly doporučeny

a prozkoumány jiné (např. vliv pietních svátků na protiprávní činnost na hřbitovech apod.). Z analýz vzešly výsledky, které jsou popsány a prezentovány pomocí grafů. Po deskriptivní analýze je provedena analýza pomocí složitějších data miningových algoritmů, pomocí kterých jsou zkoumány hlubší vztahy. Použity jsou například algoritmy shlukové analýzy, či asociačních pravidel.

Dalším hlavním cílem této práce je využít studiem nabyté zkušenosti autora s dostupným data miningovým softwarem a tyto teoretické znalosti využít v praktickém případě. Znalosti autora a jeho schopnost práce s daným softwarem jsou prohloubeny a jdou nad rámec výuky.

1 Úvod do problematiky data miningu

Data mining je obor na pomezí počítačových věd a matematiky, konkrétně matematické statistiky. Jeho hlavním úkolem je získat informace, které nejsou na první pohled vidět, ze zdánlivě nesouvisejících velkých objemů dat. Tato data mohou být více či méně rozsáhlá. Jejich charakter může být různý a jejich podrobnost také. Těmto datům se obecně říká *dataset* neboli sada dat, na kterých má být provedena data miningová úloha. Pojem data mining, tedy „*dolování dat*“ je však jakýmsi nadneseným výrazem. Správným překladem není „*dolování dat*“, ale spíše „*dolování z dat*“. Úlohou data miningu totiž není dolování, (získávání) dat jako takových, ale zjišťování vztahů a vzorců, které data mezi sebou mají. Mezi hlavní cíle data miningu patří tyto spojitosti odhalit, správně popsat a výsledky interpretovat.

Postupem času, kdy data mining začal získávat na významu, bylo potřeba ucelit základní postup řešení úloh, ze kterého by se mělo vycházet. Existovalo několik variant, avšak nejrozšířenějším se stal postup CRISP-DM (Martinez-Plumed et al. 2021).

Jednotlivé fáze postupu CRISP-DM (Martinez-Plumed et al. 2021):

- Porozumění problematice (Business understanding) – v této fázi dochází ke stanovení charakteru a cílů projektu, v jehož rámci má být data miningová úloha řešena. Jsou jasně stanoveny požadavky na výsledky, které má řešení poskytnout. Vedle těchto požadavků jsou také hodnocena rizika a přínosy řešení. v neposlední řadě se berou ohledy na náklady s řešením spojené. Dále je nutno stanovit základní informace o daném odvětví, které jsou potřeba k pochopení problému, který má být řešen.

- Porozumění datům (Data understanding) – Tato fáze se skládá z několika dílčích kroků, které je nutné provést, aby vůbec bylo možné provádět budoucí analýzy. Základním krokem je samotný sběr vstupních dat. Někdy bývá tento krok již proveden zadavatelem projektu, někdy je součástí postupu. Následuje popis dat, zevrubné a povrchové nahlédnutí na charakter dat, tedy typicky na datové typy, hodnoty jednotlivých polí v datasetu apod. Jako další je zkoumání dat. Tento krok zajistí základní přehled o vlastnostech informací, jež jsou v datech obsaženy. Typicky je využito deskriptivních technik, například frekvence výskytu, modus, medián, maximum a minimum. Po zkoumání je třeba ověřit kvalitu dat.

- Příprava dat (Data preparation) – Jedná se o časově nejnáročnější fázi. Tato fáze zahrnuje veškeré přípravy datasetu pro budoucí analýzy. v rámci této fáze jsou například vybírány takové atributy, které mají nejlepší vztah vzhledem k cílům řešení úlohy. Dále je pak velmi důležité čištění dat od chyb, chybějících údajů a nerelevantních záznamů, které nevedou k vylepšení řešení. Mezi další činnosti patří například integrace dat a vytváření odvozených atributů. Odvozené atributy jsou vytvářeny kombinací některých již existujících atributů, například může být odvozeným atributem doba letu jako rozdíl mezi časem odletu a časem příletu. Výstupem této fáze je upravený a ucelený dataset, který je podle požadavků připraven pro následné analýzy.

- Modelování (Modelling) – Fáze, ve které probíhá výběr vhodného nástroje pro řešení úlohy. Po výběru musí proběhnout nastavení parametrů nástroje a jeho kalibrace, aby nastavení bylo optimální a poskytlo relevantní výsledky.

- Vyhodnocení výsledků (Evaluation) – Úkolem této fáze je zprostředkovat informace zjištěné v rámci modelování a analýz. Je nutné zvolit vhodné metody, jak výsledky interpretovat a poskytnout ucelené a kompletní informace jako základ k rozhodnutím, která mají být založena na výsledcích řešení zadané úlohy.

- Využití výsledků (Deployment) – Zde dochází ke shrnutí procesu řešení data miningové úlohy. Jsou zde definovány způsoby, jak mohou být výsledky řešení využity zadavatelem a jak mají být výsledky chápány. Typicky se jedná o závěrečnou zprávu či report.

Existuje několik základních podmínek, které je nutné splnit a pochopit ještě před započetím řešení data miningové úlohy. Tato fáze se nazývá pre-processing. Pomocí data miningových postupů lze odhalit pouze takové vzorce a vztahy, které v datech skutečně jsou. Data je potřeba seskládat v celistvý dataset. Dataset by měl být dostatečně velký, aby mohly být vzorce a vztahy odhaleny. Na druhou stranu musí být ale jen tak velký, aby mohla být úloha vyřešena v přijatelné době. Zpracování příliš velkého datasetu zabere mnohdy i násobně delší dobu, ale nevede k výrazně přesnějším výsledkům. Data je také potřeba očistit od šumů a duplicit.

Data miningové metody mohou řešit celou řadu různých úloh. Dnešní doba je velmi bohatá na data, což velmi napomáhá růstu tohoto vědního oboru. Postupem času se vyvinuly typické

úlohy, jež lze řešit. Tyto úlohy se dají rozdělit do šesti základních kategorií (Chapman et al. 2000):

- Popis dat a sumarizace – Popis dat a jejich sumarizace se zaměřuje na popis základní charakteristiky dat, typicky v jejich jednoduché formě. To poskytuje obecný přehled o datech a jejich struktuře. Mnohdy je deskripce dat dostatečnou analýzou, která nevyžaduje použití dalších, složitějších technik. Většinou je ale popis dat použit současně s jinými technikami, typicky jako první fáze rozsáhlejšího řešení. Kromě prvotního zjištění základních informací o konkrétních datech také hraje tato technika velkou roli v prezentaci konečných výsledků řešení. Výstupy složitějších technik, jako je clustering[1], je třeba správně a detailně popsat. Technika popisu a sumarizace dat je nedílnou součástí fáze porozumění datům ze zmíněného CRISP-DM algoritmu.

- Segmentace, clustering – Segmentace rozděluje data do zajímavých a smysluplných celků či skupin. Všechny prvky dané skupiny spolu sdílí nějaké společné charakteristiky. Rozdělení do skupin pomáhá k identifikaci všech prvků, tedy odhalení jejich příslušnosti k nějaké skupině. Vedle této funkce může segmentace pomáhat například při rozdělení dat v řešení úlohy s velkým objemem informací, které jsou na první pohled velmi provázané. Segmentace pomáhá sjednotit velké objemy dat do menších sobě podobných celků, které se pak analyzují snadněji. Mezi nejběžnější algoritmy patří techniky clustering, neuronové sítě apod.

- Klasifikace – Předpokladem této techniky je sada objektů, které jsou charakterizovány určitými vlastnostmi, díky kterým spadají do jistých skupin. Každý prvek má svůj třídní popisek (*class label*), což je symbolická informace. Úkolem této techniky je sestavit klasifikační model – klasifikátor, který přiřadí správné třídní popisky k předtím neznámým, a tedy nepopsaným, objektům. Klasifikátor je vytvořen na základě předem známých vlastností u existujících objektů, které určitě spadají do zmíněných tříd/kategorií. Klasifikátor je tzv. naučen na trénovací množině dat, která je předem správně popsána. Příkladem může být identifikace, zda zvíře je pes či kočka. Trénovací sada obsahuje obrázky psů, které jsou popsány popiskem „pes“ a koček s popiskem „kočka“. Klasifikátor se z obrázků naučí některé stěžejní

[1] Clustering je metoda třídění dat do shluků (*cluster* = anglicky shluk).

vlastnosti a podle nich s určitou chybou třídí nové obrázky. Mezi hlavní algoritmy klasifikátorů patři diskriminační analýza, rozhodovací stromy, neuronové sítě apod.

- Predikce – Technika predikce slouží hlavně k předpovědi hodnoty nějakého atributu. Na rozdíl od klasifikace jsou tyto predikované atributy většinou kvantitativní. Na základě dalších předem známých informací se model snaží předpovědět hodnotu atributu pro danou kombinaci konkrétního záznamu. Jako předem známé informace slouží hlavně atributy předešlých záznamů, například k predikci vývoje ceny ropy je potřeba znát dosavadní průběh vývoje.

- Analýza závislostí – Technika analýzy závislostí slouží k odhalení pravděpodobných závislostí určitých dat na kombinaci jiných dat v datové sadě. Prvky, které určují danou závislost, se nazývají antecedenty. Prvky, jejichž závislost je tvořena antecedenty, se nazývají consequenty („*konsekventy*“). Typickým příkladem data miningové úlohy, která je řešena pomocí analýzy závislostí, je analýza nákupního košíku zákazníka. Pomocí této analýzy se dají odhalit vztahy mezi různými druhy nakupovaného zboží, například zda se dá prohlásit, že člověk, kupující mléko a mouku, koupí také cukr spíše než člověk, který koupí víno a banán.

2 Data mining a predikce kriminality

Kriminalita neboli zločinné chování, jde paralelně s člověkem již od pravěku. Jedná se o chování, které je tzv. společensky nepřípustné a narušuje běžný společenský chod. Toto chování je zločinné, a tedy je jeho jediným účelem způsobit škodu jedné strany na druhé. Škody mohou být různého druhu – majetkové, duševní, na zdraví apod. Obor, který se více zabývá výzkumem kriminality, se nazývá kriminologie.

V rámci morálního kodexu každého člověka je se v zásadě kriminálnímu jednání vyhnout, pokud ho k tomu nedovedou zvláštní doprovázející okolnosti a situace. Je jen velmi málo lidí, kteří se dopouštějí zločinného chování tzv. „z rozmaru“, neboli z vlastního dobrovolného rozhodnutí bez zjevné další příčiny. Z toho plyne, že většina lidí se snaží kriminalitě vyhnout. Nejen jejímu páchání, ale hlavně jejímu dopadu na ně samotné. Odtud vyplývá fakt, že lidem je přirozené podnikat takové kroky, které sníží pravděpodobnost výskytu zločinného chování.

Metody, jejichž úkolem je zamezit trestné činnosti, zažily bohatý vývoj. Díky moderní době je lidstvo ještě o kousek blíž k zásadnímu snížení kriminality. Výkonné vyspělé počítače a bohatý obor matematické statistiky a analýzy totiž nabízí zásadní pomocnou ruku ve směru predikce, a tedy i prevence, kriminality. Velká datová úložiště poskytují možnost ukládat velké objemy dat. Velké objemy dat mohou poskytovat na první pohled neznámé informace. Obecně se dá říct, že čím více dat je shromážděno, tím přesnější informace mohou poskytnout. Tato data mohou mj. odhalit vzorce chování. Svět jsou data. Každý spáchaný trestný čin generuje také určité množství dat jako svou charakteristiku – místo, čas, typ zločinu, nadmořskou výšku, aktuální počasí, věk a pohlaví pachatele apod. Všechna tato data, pokud jsou zaznamenána, mohou poskytnout klíč k predikci takové kriminality.

Predikce je jedna ze základních úloh data miningu. Využívá znalosti předem zjištěných informací, na jejichž základě je vytvořen model, který s určitou pravděpodobností odhaduje (predikuje) závislou skutečnost. Závislou skutečností je v tomto případě myšlen trestný čin. Ze zmíněných informací o trestném činu, tedy dat, lze pomocí určitých metody vytvořit modely, které mohou vést k predikci výskytu trestného činu. Tato predikce má samozřejmě jistou míru chyby a nikdy se nedá považovat na naprosto bezchybnou.

Predikce kriminality by měla hlavně sloužit k její prevenci. To zahrnuje snížení pravděpodobnosti výskytu na co nejnižší možné minimum. Snížení počtu trestných činů

vede k většímu pocitu bezpečí člověka, a tedy k nižší míře stresu z běžného života. Mezi další důvody, proč podnikat kroky v prevenci kriminality, je snížení nákladů z neefektivního využívání ochranných prostředků, ať už policejních hlídek, či kamerového systému apod. Efektivní rozmístění hlídek PČR či MP snižuje riziko efektu „příležitost dělá zloděje“, tedy předchází trestným činům, tak, že policisté hlídkují na místech s předpovězeným vyšším výskytem trestné činnosti spíše než na místech s nižším výskytem. Omezí se chyby prvního a druhého typu, tedy moment, kdy byla spáchána trestná činnost a nebyl u ní policista, nebo naopak, kam byl poslán policista, ale nebyla spáchána trestná činnost.

2.1 Predikce kriminality v zahraničí

K predikci kriminality je využívaná celá řada různých metod, algoritmů a softwarových nástrojů, které se více či méně liší v různých zemích světa.

2.1.1 Německo

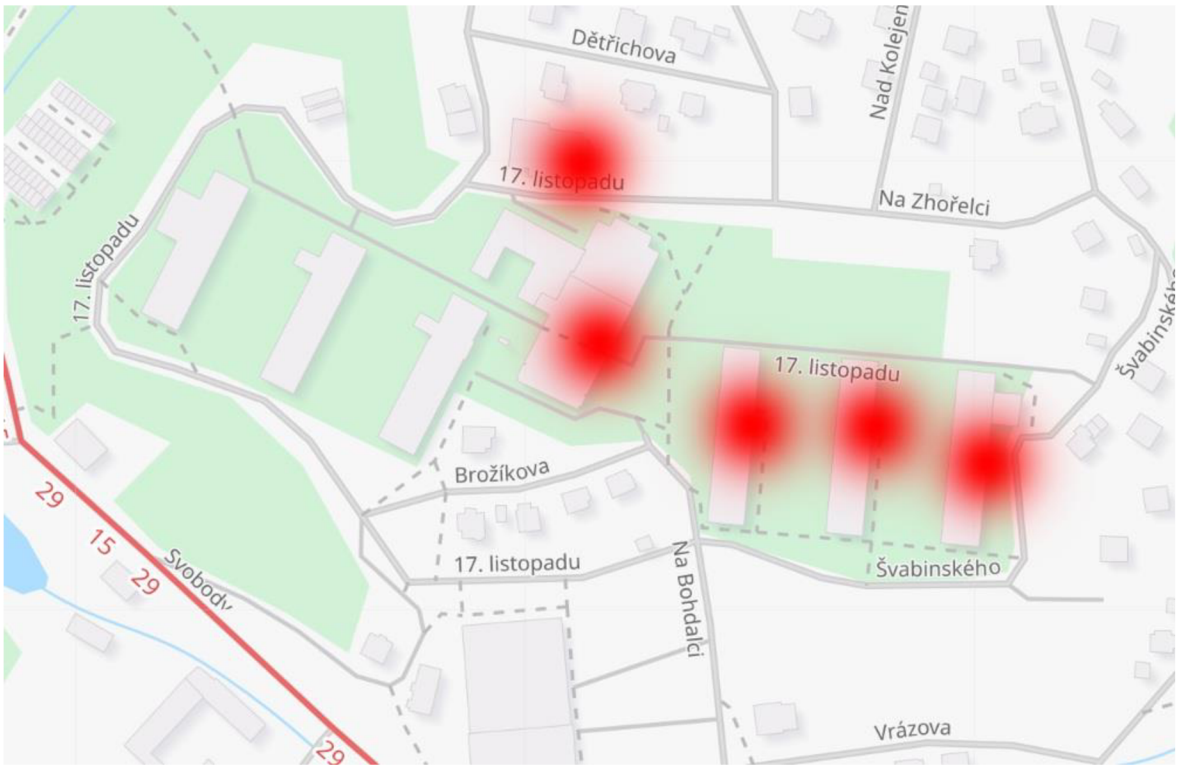
Jedním z prvních protagonistů prevence kriminality je Německo. Německo, jakožto velmi konzervativní a pokrokový stát, podniká kroky ve snížení počtu trestných činů kvalitním systémem ochranných složek. Policie v Německu je jedna z nejlépe organizovaných na světě. Prevence kriminality s ohledem na *big data* a moderní metody data miningu je v Německu stále v rané fázi. Doposud nejsou příliš aktivní v prevenci ani predikci kriminality.

O trestných činech neukládají mnoho informací. Většinou se jedná jen o místo a čas nebo čin a čas. Jak již bylo řečeno, více dat znamená více informací, tedy více nalezených spojitostí mezi nimi. Bez dostatečného počtu informací se nedají efektivně zjišťovat charaktery trestných činů a vytvářet modely k jejich predikci, potažmo prevenci. Německá policie při ukládání dat o trestných činech nespojuje činy s žádnými osobními daty, např. věkem, váhou, příjmem, majetkem, vzděláním apod. Toto jsou však zásadní vlastnosti, které mohou pomoci k navýšení přesnosti modelů (Bauer 2019).

Data miningové metody se u německé policie omezují spíše jen na statistické vyhledávání než na predikci. Doposud nepracují příliš aktivně s umělou inteligencí. Jedním z nejefektivnějších nástrojů, které jejich policie používá, jsou tzv. *heat mapy*.

Heat mapa je clusterová metoda, která generuje matici sloupců a řádků. Jednotlivé buňky této matice jsou zbarvené podle míry pravděpodobnosti výskytu zkoumaného jevu (Wilkinson a Friendly 2009). Pomocí heat mapy se dá zobrazit pravděpodobnost výskytu

určitého jevu ve 2D prostoru. Heat mapou se zobrazuje například nejsledovanější oblast webové stránky, výška nově napadlého sněhu apod.



Obr. 1 – Heat mapa výskytu studentů TUL ve večerních hodinách

Mezi činy, na které se německá policie hlavně zaměřuje, jsou krádeže a loupeže. Jen velmi zřídka se zabývají jiným typem trestného činu. Jednou z hlavních teorií, které v tomto ohledu testují, je tzv. *near-repeate hypothesis*. Tato teorie vychází z předpokladu, že budovy či obydlí, které se nacházejí v blízkosti již vykradené budovy či obydlí, mají vyšší riziko být vykradené také (Egbert 2019).

Jako hlavní nástroj využívá německá policie převážně komerční software PRECOBS Enterprise.

2.1.2 Indie

Indie je jedním z největších států na světě co do počtu obyvatel. S velkým počtem obyvatel také roste počet spáchaných trestných činů. Indická policie si tento fakt velmi dobře uvědomuje, a proto v tomto ohledu aktivně vyvíjí snahu, jejímž cílem je snížit míru kriminality. v rámci zlepšení služeb k odhalení a omezení trestné činnosti indická policie spolu s úřady vyvíjí počítačový systém, který jim bude v tomto ohledu velkou pomocí. Automatizace systému pro sběr dat o trestných činech přináší velké pozitivum ve formě širokého záběru dat o kriminalitě.

V roce 1986 byl v tomto ohledu založen Národní úřad pro sběr dat o kriminalitě NCRB (NCRB 2018). Policie spolu s vědci na poli kriminality a odborníky na kriminologii využívá data z NCRB ke statistickým analýzám, aby společně snížili míru kriminality. Aby ale data mohla být efektivně využita, je potřeba zvolit správnou metodu. k tomu využívají metody data miningu. Ten hraje zásadní roli v popisu profilů trestných činů a také předpovědi.

Indické NCRB je součástí poměrně zdařilého a funkčního e-governmentu. v rámci tohoto řízení byly vybudovány i další elektronické systémy, které mají za úkol digitalizovat v podstatě veškeré dění ve státě. Jsou jimi například *The Crime Criminal Information System* (CCIS), *Prison Statistics, Jail Management* a další (Gupta a Chandra 2006).

Pro zlepšení přístupu do databáze trestných činů byla vyvinuta aplikace *Crime Analysis Tool*, které se skládá ze 3 modulů:

- *Data Extraction Module* – modul pro stahování dat z databáze,
- *Crime Analysis Module* – modul pro základní analýzu získaných dat,
- *Data information and Comparison Module* – modul pro porovnání dat.

Výstupem prvního modulu jsou datasety, které odpovídají vlastnostem, které si uživatel nastaví. Tato data jsou určena pro další možnou analýzu mimo tuto aplikaci. Základními výstupy druhého modulu jsou *hot spoty* trestných činů, oblasti s vysokou mírou kriminality a oblasti s nízkou mírou kriminality. Výstupem posledního modulu jsou popisné charakteristiky dat, jako například grafy.

Někteří autoři upozorňují (Institute of Electrical and Electronics Engineers 2012), že na míru kriminality mají přímý vliv některé základní lidské charakteristiky, jako jsou věk, pohlaví, příjmy, zda člověk pracuje či ne, úroveň a typ vzdělání apod. Tyto faktory se mohou u trestného činu zjistit a zaznamenat. Na jejich základě se dá s určitou pravděpodobností predikovat, jak je daný jedinec náchylný k trestné činnosti. Tyto faktory však nejsou efektivní ke geografické predikci.

V současné době používají v Indii k prevenci kriminality algoritmy *Multiple Linear Regression*, *K-Means Clustering* a *Asociační pravidla* (Singh et al. 2018).

2.1.3 Rusko

Studie (Ingilevich a Ivanov 2018) se věnuje predikci kriminality v oblasti Petrohradu. Ruská policie, podobně jako ta v ČR, sbírá data o spáchaných trestných činech. Místní instituce Ministerstva pro vnitřní události Ruské federace, má na starosti sběr těchto dat od policie a jejich následné zpracování. Data, která jejich policie zaznamenává, mají stejnou strukturu, jako ta u PČR. Záznamy obsahují datum a čas trestného činu, polohu v podobě zeměpisné šířky a délky (GPS souřadnice) a také popis typu trestného činu (zda se jedná o krádež, výtržnictví apod.).

Jako hlavní faktory, jež mají vliv na kriminalitu, byly vysledovány výše HDP, hustota zalidnění, míra nezaměstnanosti a počet lidí bez domova. Vedle těchto faktorů existují také geografické faktory, které mají zásadní vliv. Těchto faktorů je celá řada a je potřeba je nejprve řádně posoudit, aby výsledný prediktivní model nebyl příliš komplexní a nezabíral zbytečně široké oblasti (pokud algoritmus označí celé město, je pravděpodobnost predikce 100 %).

Datům je potřeba nejprve dobře porozumět, a tak je důležité z nich odstranit šum. Riskem u velkého objemu dat je, že budou obsahovat šum. Šumem se myslí odlehlé nebo nesouvisející hodnoty, které jen zvyšují paměťovou náročnost, ale nijak nepřispívají kvalitě modelu. Jednou z nejefektivnějších metod pro odstranění šumu je clustering. Clustering rozdělí data do shluků vzájemně podobných záznamů. Tyto vytvořené shluky je pak možné dále blíže zkoumat a s jejich pomocí je možné nalézt jejich hlavní ovlivňující faktory. Studie v zásadě objevila osm základních faktorů. Tyto faktory však způsobují vysokou komplexnost, proto je potřeba vybrat ty podstatnější. Pomocí testu X^2 *(„chí kvadrát“*, pozn. autora) byly vybrány jako zásadní faktory hustota obyvatel dané oblasti, pozice barů, škol a kostelů.

Na datech byly aplikovány celkem tři metody – lineární regrese, logistická regrese a gradient boosting. Jejich výsledky byly porovnány pomocí Mean Absolute Error (MAE)[2] a také pomocí R^2 koeficientu determinace[3]. Ze studie vyplývá, že logistická regrese není příliš vhodná, zatímco lineární regrese a gradient boosting jsou poměrně přesné metody a vhodné k predikci kriminality. Lineární regrese bohužel predikuje také záporné hodnoty, což u počtu

[2] Mean Absolute error je jednotka chyby mezi dvěma párovými pozorováními stejného jevu (Willmott a Matsuura 2005).
[3] Koeficient determinace je číslo, které vyjadřuje kvalitu regresního modelu, tedy jak dobře model popisuje zkoumanou problematiku (Yin a Fan 2001).

trestných činů není relevantní ani logické. Jako nejlepší metoda se tedy jeví gradient boosting.

2.1.4 USA, Chicago

Spojené státy Americké se dlouhodobě potýkají s vysokou mírou kriminality. Obecně nejhorším městem, dle slov samotné FBI, je město Chicago, které bylo označeno za „*Hlavní město zločinu*“. To zapříčinilo vznik několika různých studií, které mají napomoci ke snížení míry kriminality právě na území Chicaga.

Data jsou zdejší policií sbírána různými způsoby a také mají různé formy. Tato data je potřeba převést do relevantní formy – tedy hlavně na záznamy se strukturou typ činu, jeho souřadnice a čas. To slouží jako základ ke geografické analýze a clusteringu. Vstupní data pro analýzy se skládají celkem ze tří částí. První část obsahuje oblasti (městské části, čtvrtě) se souřadnicemi. Další část obsahuje data o trestných činech spáchaných v těchto oblastech a třetí část obsahuje informace o populaci, tedy hustotě zalidnění každé oblasti, věku, vzdělání apod.

Mezi hlavní používané nástroje ve studiích (Alqahtani et al. 2019) jsou softwarové nástroje WEKA a SaTScan. Nástroj WEKA je používán hlavně pro clusteringovou metodu K-means a nástroj SaTScan slouží pro geografickou analýzu. Pomocí softwaru WEKA je aplikován algoritmus K-means clustering, který slouží také k detekci shluků a ověření kvality detekce. Ověření probíhá pomocí výpočtu reziduálního součtu čtverců. Tím se rozumí součet druhých mocnin rozdílu mezi modelovanou funkcí a skutečnými hodnotami, tedy jak dobře model popisuje skutečná data. Výpočet reziduálního součtu čtverců může ukázat na ty clustery, které mohou obsahovat odlehlé hodnoty. Příliš velká hodnota tohoto součtu znamená, že data jsou geograficky velmi rozptýlená a s největší pravděpodobností jsou spíše náhodně rozmístěné, než aby mezi nimi existoval vztah. Odlehlé hodnoty je možné odhalit například pomocí mezikvartilového rozpětí. Hlavním účelem SaTScan je vyhledávání clusterů, tedy shluků a jejich rozdělení podle vyšší či nižší míry kriminality, resp. rizika trestného činu. Výstupem tohoto nástroje je seznam shluků, seřazený sestupně podle míry rizika.

2.1.5 Nový Zéland a Velká Británie

Na Novém Zélandu používají data mining v prevenci kriminality k určení geografické oblasti s nejvyšším rizikem trestné činnosti s ohledem na její dopad na společnost. k určení dopadu trestného činu používají CHI – Crime Harm Index (Curtis-Ham a Walton 2018). Jedná se o celostní metriku, která napomáhá určení závažnosti činů vzhledem k jejich podstatnému okolí.

CHI na Novém Zélandu určují podobně jako ve Velké Británii, dle modelu Univerzity v Cambridge. k výpočtu váhy daného trestného činu je nejprve nutné daný trestný čin převést na jeho ekvivalent doby strávené ve vězení. U trestných činů, které projdou soudním řízením a pachatel je skutečně uvězněn, je brán průměr doby strávené za mřížemi. U těch, u kterých nedojde k uvěznění, je výpočet aproximován. Jedná se například o veřejně prospěšné práce, škody na majetku atp. Dle Cambridge je CHI patnáctým percentilem rozdělení u daného trestného činu. Novozélandské CHI je komplexnější. Rozdělují trestné činy do více specifických kategorií, zatímco britské je více obecné.

Na Novém Zélandu slouží určení CHI jako předcházející etapa pro výpočet PLI (Priority Location Index), což je hlavní měřítko deskripce a predikce. Toto číslo určuje míru kriminality v dané oblasti. Slouží k určení oblasti s nejvyšší sociální nevýhodou a nejvyšším výskytem trestných činů s vysokým sociálním dopadem. Jedná se o socio-demografický ukazatel, který se skládá celkem z osmi demografických ukazatelů a tří proměnných, vlastních trestným činům. Jako výchozí oblasti (geografické šablony) používají vlastní rozdělení do komunit, tzv. CAU (Census Area Unit), což se dá připodobnit českým sídlištím, obvykle čítající 3000–5000 obyvatel (Curtis-Ham a Walton 2017).

Trestnou činnost ovlivňuje celá řada neuvažovaných faktorů a značný vliv na šum v datech má i fakt, že mnoho trestných činů může být spácháno pachateli ze vzdálených oblastí, což uměle zvedá dopad v jiné oblasti.

K výpočtům zmíněných indexů a odhalení rizikových oblastí je použita nemalá řada socio-demografických ukazatelů. Ze všech komunit má 60 % rozdíl mezi CHI a PLI do jednoho decilu, 81 % má rozdíl do dvou decilů. To potvrzuje vztah mezi sociální úrovní oblasti a závažností zde spáchaných trestných činů. z toho plyne, že CHI je možným ukazatelem pro identifikaci rizikové oblasti a může sloužit k efektivnější distribuci sil ochranných složek.

2.2 Predikce kriminality v České republice

Data mining a s ním spojená prevence kriminality je na území ČR zatím pouze spíše teoretickým pojmem. Většina příslušníků je k tomuto přístupu skeptická. Avšak oddělení informatiky, které se této problematice věnuje, vyvíjí snahu v tomto směru. Policie používá software na serveru *kriminalita.policie.cz* (GIS PČR 2022). Tento software má celkem 3 vrstvy. První vrstva je *veřejná* – dostupná všem, kdo stránku navštíví. Poskytuje základní přehled o kriminalitě na daném území. Druhou vrstvou je vrstva *obce* – tato vrstva poskytuje přesnější náhled na kriminalitu, včetně geografických informací. Data, která jsou zkoumána v této práci, pochází právě z vrstvy obce. Tato vrstva je veřejnosti nedostupná a je využívána městskou policií příslušné obce. Třetí vrstva je *státní*. Ta je přístupná pouze státní policii a obsahuje veškeré informace o trestných činech. Navíc, oproti předchozím dvěma vrstvám, obsahuje i základní moduly pro statistiku, zahrnuje i jiné faktory jako například počasí. Obsahuje také predikční modul, který určuje pravděpodobnost výskytu nějakého deliktu v dané oblasti.

Vzhledem k dosavadní skepsi je toto odvětví na území ČR zatím nepříliš rozvinuté, avšak v poslední době je na vzestupu, včetně nových projektů, které mají být v tomto ohledu do budoucna nápomocné.

3 Software

V této kapitole jsou představeny některé softwarové nástroje, které jsou v různých státech na světě používány za účelem predikce a prevence kriminality.

3.1 PRECOBS

PRECOBS je zkratka *Pre Crime Observation System*. Jedná se o softwarový nástroj, který využívá některých data miningových algoritmů ke zjištění charakteru a podstaty trestných činů, v návaznosti na dříve sesbíraná data.

Software se využívá hlavně k prevenci nepříliš závažných trestných činů, typicky vloupání. Vloupání spadají do kategorie zločinů blízkého opakování. Charakteristickým rysem těchto trestných činů je, že se výhradně nevyskytují v čase ani prostoru samostatně. Z toho plyne předpoklad, že se v dané oblasti stanou vícekrát a během krátké doby, většinou do 72 hodin.

Software využívá dříve nabytých znalostí o trestných činech k výpočtu pravděpodobnosti výskytu takového zločinného řetězce. Mezi základní charakteristiky činů patří místo zločinu, způsob, jakým byl spáchán, odcizené věci a také datum a čas spáchání. Vedle těchto hlavních spouštěčů využívá také tzv. anti-spouštěče, což jsou okolnosti, které naopak naznačují, že se jedná spíše o spontánní čin. Tyto rysy mívá většinou vandalství.

Software analyzuje výše zmíněná data a následně předpovídá budoucí možné trestné činy v poloměru do 250 m od původního místa spáchání a v časovém rozmezí 24 až 72 hodin od času spáchání. Tato předpověď je zobrazena na mapě, včetně časových a prostorových informací. Přesnost softwaru PRECOBS je okolo 80 %.

Do budoucna se uvažuje, že by PRECOBS mohl sbírat i více dát z různých zdrojů (např. Twitter), aby předpověď byla efektivnější a rychlejší.

PRECOBS není běžně dostupným softwarovým nástrojem. Je vyvíjen v Institutu pro predikční techniky (Institut für musterbasierte Prognostechnik), který sídlí v Německu (Institute für musterbasierte Prognostechnik 2022).

3.2 WEKA

WEKA je jednodušší software, který obsahuje celou řadu algoritmů pro řešení data miningových úloh. Jeho aktuální verze WEKA 3 obsahuje nástroje pro přípravu dat, klasifikaci, regrese, clustering, asociační pravidla a vizualizaci. Tento software podporuje i tzv. deep learning, což jsou algoritmy strojového učení, které se skládají z mnoha vrstev. Jedná se o jednodušší typ programu, do kterého je nahrán soubor obsahující data. s daty jsou pak prováděny jednotlivé operace a výpočty. Obsahuje pouze základní způsoby vizualizace a poskytuje pouze jednoduché akce, které je možné s daty provádět. Jednotlivé výpočty se nedají zřetězit, to znamená, že výstup jednoho algoritmu se nedá použít jako vstup pro další algoritmus bez toho, aniž by musel být předchozí výstup uložen do samostatného souboru.

WEKA je vyvíjen na univerzitě Waikato na Novém Zélandu v rámci projektu přiblížení metod strojového učení veřejnosti a zpřístupnění jejich používání v různých odvětvích, jako je např. prevence kriminality. Je vyvíjen pod GNU General Public License[4] a je publikován jako open-source[5] program (University of Waikato 2018).

3.3 KNIME

KNIME (*Konstanz Information Miner*) je velmi sofistikovaný a komplexní nástroj, který umožňuje odbornou práci v rámci analýzy dat. Umožňuje práci od základních jednoduchých datových analýz až po aplikaci složitějších algoritmů strojového učení a umělé inteligence, jako jsou např. neuronové sítě apod. KNIME má, na rozdíl od WEKA, přívětivé uživatelské rozhraní.

Program je schopen pracovat s celou řadou vstupních typů souborů: od textových (CSV, PDF, XML atd.) přes nestrukturované datové typy až po časové řady. Umožňuje také práci s daty, která nejsou uložena lokálně, ale v nějaké databázi.

S daty je možné provádět několik základních operací, jako je transformace, filtrování, čištění. Dále pak je možné data velmi dobře vizualizovat pomocí přehledných grafů. Vedle zpracování je také možné aplikovat několik již zmíněných algoritmů data miningu.

[4] GNU General Public License je licence pro svobodný software. Software pod touto licencí je volně použitelný v jiných dílech, ta však nemohou mít jinou licenci, např. nemohou být komerčně využita.
[5] Open-source program je takový program, ke kterému jeho tvůrce poskytuje kompletní zdrojové kódy včetně dokumentace.

S daty se pracuje v tzv. pracovním workflow. Jednotlivé kroky workflow jsou reprezentovány jednotlivými komponentami, které jsou spojeny. Komponenty pak představují konkrétní operace. Ukázka vybraných komponent je na obrázku (Obr. 2).

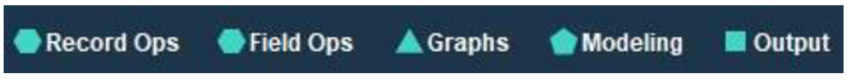
Obr. 2 – Ukázka vybraných komponent z programu KNIME

KNIME Analytics Platform je open-source program, poskytovaný zdarma. Je vyvíjen na švýcarské univerzitě Konstanz pod ochranou GNU General Public License, podobně jako WEKA. Nabízí širokou škálu zdrojů a návodů, jak s programem pracovat. Program má také velmi širokou komunitu a svá vlastní diskusní fóra (University of Konstanz 2022).

3.4 SPSS Modeler

SPSS Modeler je program vyvíjen celosvětově známou společností v oblasti informatiky – IBM. Program jako takový je velmi komplexní a nabízí širokou škálu možností práce s daty.

Jako vstupy s daty mohou pro Modeler být různé typy, podobně jako u KNIME, např. textové, databázové, strukturované a nestrukturované apod. v rámci programu se pracuje s tzv. datovými proudy neboli streamy. Datový stream prochází skrze uzly, které reprezentují jednotlivé operace, které jsou s daty prováděny. Uzly jsou rozděleny do čtyř kategorií, kdy každá má vlastní tvar ikony (Obr. 3).



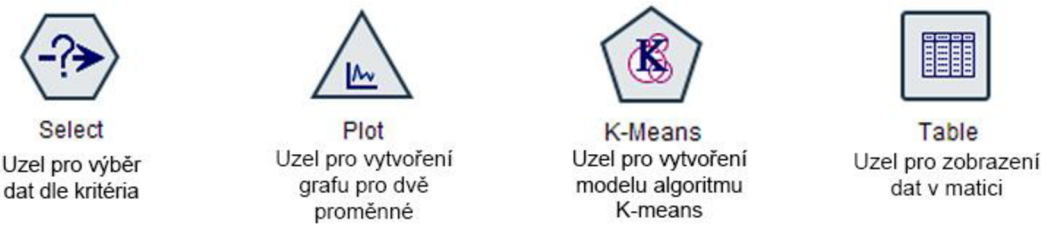
Obr. 3 – Kategorie uzlů a vzhled jejich ikon

Jednotlivé kategorií obsahují několik uzlů a mají různý účel:

- Record a Field operations – Jedná se o operace se záznamy (řádky datové matice) a proměnnými (sloupce datové matice). Patří sem např. funkce Select (výběr), Sort (řazení), Filter (třídění) apod.
- Graphs – Uzly s grafy. Data je pomocí tohoto uzlu možné zobrazit prostřednictvím několika neběžnějších grafů.

- Modeling – Uzly sloužící k modelování. Tyto uzly jsou jednotlivé data miningové algoritmy. Jsou propojené s upraveným zdrojem dat, na jejichž základě jsou modely naučeny a připraveny k použití na konkrétní aplikaci. Typicky je potřeba je naučit na trénovací sadě a model pak ověřit na testovací sadě. Modeler obsahuje širokou škálu algoritmů, od lineární regrese, přes K-means až po neuronové sítě.
- Output – Uzly v této kategorii slouží k náhledu na data jiným způsobem než pomocí grafů. Jedná se o zobrazení v tabulce, datové matici apod.

Software IBM SPSS Modeler je uzavřený program s placeným přístupem. Je ale velice komplexní, nabízí celou řadu silných nástrojů pro práci s daty a má rozsáhlou dokumentaci. z těchto důvodů byl vybrán jako hlavní program pro práci s daty v rámci této diplomové práce. Příklady jednotlivých uzlů jsou zobrazeny na obrázku (Obr. 4).



Obr. 4 – Příklady uzlů z programu SPSS Modeler

4 Algoritmy

K získávání informací z velkého objemu dat se v rámci data miningu využívá celá řada různě náročných metod. Data mining je chápán jako proces hledání vzorců v datech. Vzorec v tomto případě znamená určitý platný a vysvětlený vztah. v této kapitole jsou představeny některé používané metody data miningu.

4.1 Statistical finding

Jednou z nejjednodušších metod data miningu je obecná statistika. Tato metoda neslouží primárně pro získávání nových složitých vzorců, které se v datech skrývají, ale spíše k popisu dat a jejich vlastností. Typicky je tato metoda součástí druhé fáze postupu řešení, tedy porozumění datům. Dataset se v případě statistiky nazývá statistickým souborem, což je konečná množina zkoumaných dat. Velikost této množiny, tedy počet prvků, které množinu tvoří, se nazývá rozsah výběru.

V rámci obecné statistiky se zjišťuje několik základních parametrů a vlastností daného výběru:

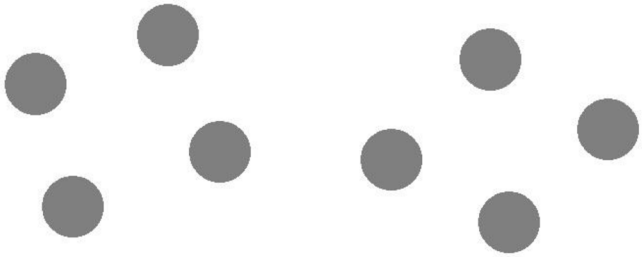
- Četnost – Četnost jevu označuje absolutní počet výskytů daného jevu. Příklady: počet stisků jednotlivých kláves na klávesnici, počet jednotlivých známek na vysvědčení v celé třídě, počet aut, která projela na dané komunikaci apod.
- Aritmetický průměr – Značí se $\bar{x}$ („*x s pruhem*“). Aritmetický průměr je číslo, které ukazuje, jaká hodnota statistického znaku připadá rovnoměrně každému prvku statistického souboru, tedy jaká hodnota je průměrná napříč celým souborem. Příklady: průměr známek na vysvědčení, průměrná hmotnost dětí ve třídě apod.
- Modus – Značí se $\hat{x}$. Modus označuje takovou hodnotu statistického znaku, která se v celém souboru vyskytuje nejčastěji. Příklady: druh vozidla, který lidé kupují nejčastěji, nejprodávanější velikost bot nebo oblečení apod. Oproti průměru se jedná vždy o skutečnou hodnotu ze souboru. Průměr může být vyjádřen i číslem, které v souboru neexistuje (např. známka 1,8 není platná známka, velikost bot 45,68 není platná velikost).

- Medián – Značí se $x_{0,5}$ nebo $\tilde{x}$. Medián označuje takovou hodnotu statistického znaku, která dělí celý soubor na dvě přesné poloviny, tedy 50 % všech hodnot je menší než medián a 50 % je větších. Výhodou mediánu oproti průměru je, že je robustní vůči odlehlým hodnotám.

4.2 K-Means clustering

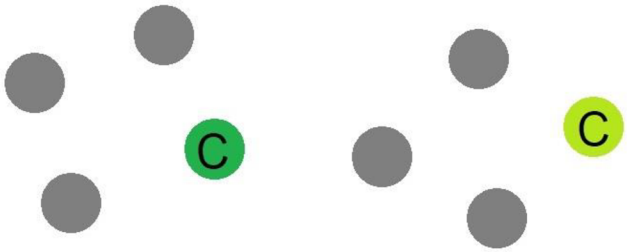
K-means je jednou z metod shlukové analýzy. Data miningovou úlohou je zjistit vztahy a vzorce mezi daty, které nejsou na první pohled patrné. Shluková analýza pomáhá odhalit vzorce mezi daty pomocí roztřídění do shluků. Shluk je chápán jako množina bodů s co nejvíce vzájemně podobnými prvky. Zároveň platí, že shluky by měly být vytvořeny tak, aby prvky jednotlivých shluků byly co nejvíce odlišné od prvků z jiného shluku.

Algoritmus zpravidla náhodně zvolí tzv. centroid. Centroid je takový bod každého shluku, ke kterému mají všechny ostatní prvky daného shluku nejmenší vzdálenost. Centroid je tedy těžištěm shluku. Úkolem je minimalizovat vzdálenost mezi jednotlivými prvky v daném shluku a centroidem, aby shluk co nejlépe popisoval danou množinu prvků. Jedná se o minimalizační metodu.



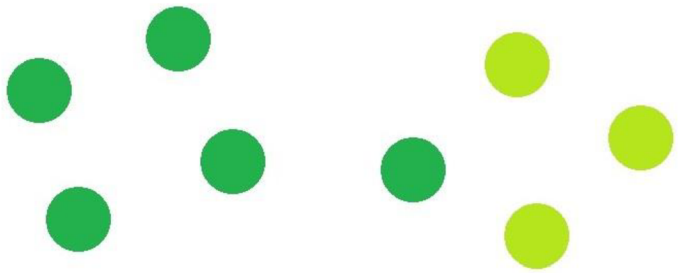
Obr. 5 – Prvky datového souboru rozmístěné v prostoru

K-means je iterační metoda, proto je její výpočet prováděn postupně. Na začátku je v základní množině dat (Obr. 5) centroidem zvolen náhodně jeden prvek (Obr. 6).



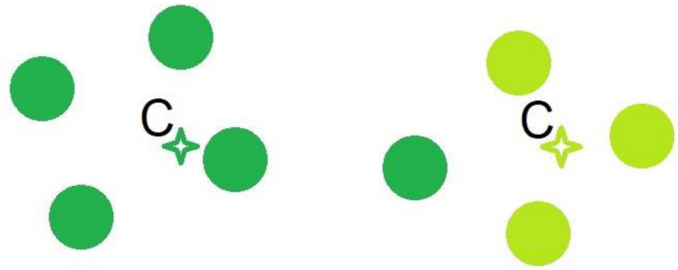
Obr. 6 – Zvolení centroidů mezi existujícími daty

Pro všechny centroidy jsou spočteny vzdálenosti ke všem bodům. Body jsou pak dle vzdálenosti přiřazeni jednotlivým centroidům. Tím vzniknou primární shluky (Obr. 7).



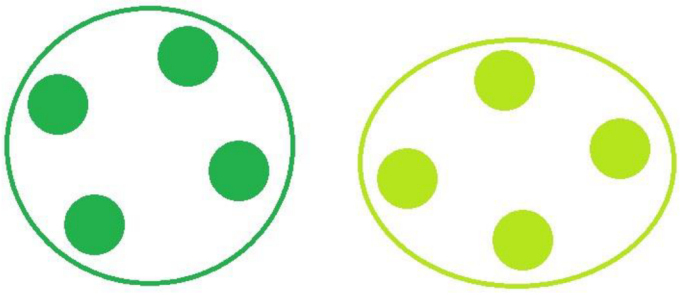
Obr. 7 – Primární shluky dat před dalším výpočtem

Pro všechny shluky je následně přepočítán nový centroid. Tyto nové vypočítané centroidy již nejsou konkrétní prvky z datového souboru. Po výpočtu mají nové centroidy zpravidla polohu mezi jednotlivými prvky (Obr. 8).



Obr. 8 – Nové vypočtené centroidy shluků

Nové těžiště je pak použito pro další výpočet, kterým se data opět roztřídí do nových shluků (Obr. 9). Algoritmus je takto opakován, dokud se poloha těžišť neustálí.



Obr. 9 – Finální podoba shluků s ustálenými těžišti

Počet konečných shluků může být definován na začátku na základě odhadu. Některé programy (SPSS Modeler) mají navíc funkcionalitu, že pomocí doplňujících dat v záznamech jsou schopny identifikovat počet potřebných shluků, ve kterých jsou si prvky podobné na základě jiných vlastností, než je jejich vzájemná poloha. Algoritmus má celkem tři základní podmínky úspěšného ukončení (Sharma 2019):

- těžiště nově vzniklých shluků už se nemění (nemění svou polohu),
- jednotlivé prvky zůstávají stále ve stejných shlucích (nemění shlukovou příslušnost),
- maximální počet iterací byl dosažen.

K-means v prevenci kriminality lze použít pro hledání shluků trestných činů, které mají společné vlastnosti, hlavně geografické. Lze zohlednit například typ činu nebo časové rozmezí, kdy byly trestné činy spáchány.

4.3 K-Nearest Neighbor

K-Nearest Neighbor (zkráceně KNN), nebo také K-Nejbližší soused, je další z metod shlukové analýzy. Jedná se o algoritmus strojového učení, který je řízený. Dělí se na dvě základní úlohy – klasifikační a regresní. Jelikož se jedná o algoritmus strojového učení, dochází zde k predikci na základě naučení na trénovací množině. Ta se skládá z prediktorů a predikovaných vlastností.

Klasifikační úloha má za úkol na základě prediktorů přiřadit příslušnost prvkům z datasetu k určité skupině či kategorii sobě nejbližších sousedů. Výstup této úlohy je diskrétní. To znamená, že na základě množiny prediktorů algoritmus predikuje příslušnost ke konkrétní skupině. Např. je možné predikovat, zda má daný člověk rád ananas na pizze (závislá proměnná, 1 ano, 0 ne) na základě věku dotyčného člověka (nezávislá proměnná). To předpokládá, že daná skupina lidí přibližně stejného věku má podobná rozhodnutí, např. lidé ve věku 20-30 let mají ananas spíše rádi a lidé ve věku 50-60 let spíše ne.

Regresní úloha je podobná klasifikační úloze, avšak algoritmus zde přiřazuje konkrétní (spojitou) hodnotu predikovanému znaku na základě odhalení sousedství, které vyplývá ze sady prediktorů, např. predikce hmotnosti člověka (predikovaná, závislá proměnná) v závislosti na výšce. Hmotnost je pak predikovaná na základě podobnosti lidí s podobnou výškou.

KNN algoritmus předpokládá, že vzájemně podobné prvky jsou zároveň i sobě blízké (nejsou výrazně daleko). KNN algoritmus vychází ze zvoleného čísla K, které označuje počet nejbližších prvků nově přidanému prvku.

Postup algoritmu je následující. Nejprve je nutné zvolit optimální číslo k sousedů. Pro každý prvek v datasetu je spočtena vzdálenost s novým dotazovaným prvkem (ten, u kterého má být predikována vlastnost). Tyto vzdálenosti jsou spolu s indexy spojeny do jedné řady

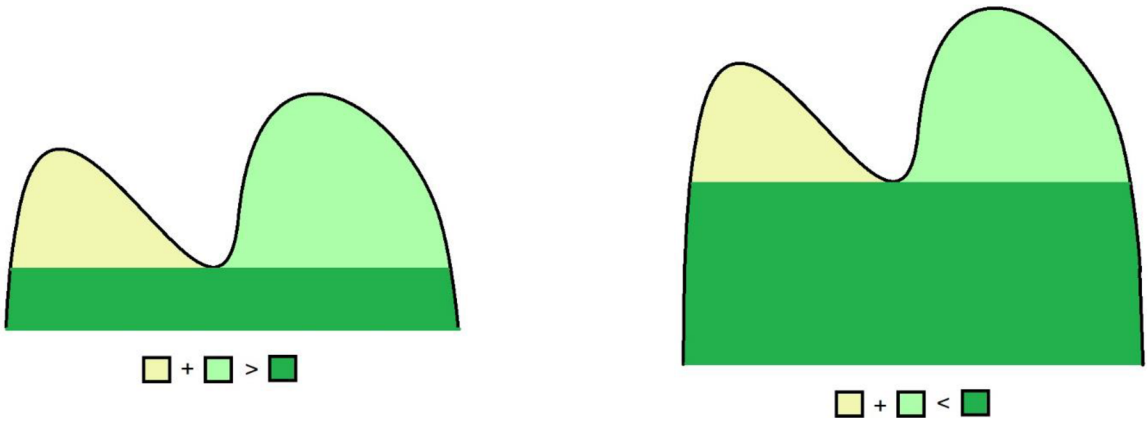
a následně seřazeny vzestupně. z této řady jsou pak zkoumány vlastnosti prvků od 1 do *K*. Vzhledem k postavení řady se jedná o k nejbližších sousedů zkoumaného prvku. v případě klasifikační úlohy je pak predikované vlastnosti (proměnné) přiřazen modus hodnoty této proměnné u k sousedů. v případě regresní úlohy je predikované proměnné přiřazen medián této proměnné u k sousedů (Harrison 2019).

4.4 HDBSCAN

Algoritmus HDBSCAN (*Hierarchical Density-based Spatical Clustering of Applications with Noise*) je čelenem rodiny algoritmů shlukové analýzy založených na hustotě.

Algoritmus vyhledává shluky na základě hustoty. Běžným způsobem výpočtu hustoty je pomocí tzv. core distance. Ta označuje vzdálenost daného bodu k jeho *k*-tému sousedu. Zvolené *k* označuje pořadí nejbližšího souseda. Čím kratší je tato vzdálenost k např. desátému sousedu napříč celým datasetem, tím větší je hustota bodů v okolí daného bodu.

Vyhledávání shluků probíhá na základě zvolené hladiny (úrovně) hustoty. Tuto úroveň je možné nastavit globálně. Lze si ji představit například jako hladinu moře a shluky jako ostrovy. Čím více se zvedá hladina moře, tím více je jednotlivých ostrovů, přičemž některé mohou zmizet pod hladinou. Naopak, čím nižší hladina je, tím větší ostrovy se tvoří a objevují se i nepoměrně malé ostrovy. To může být problém v případě, že je použit na datech, které mají proměnlivou hustotu. Práce s daty o kriminalitě je právě takový případ. Tento problém řeší právě převod klasického DBSCAN na HDBSCAN, který je hierarchický. Tento algoritmus totiž rekurzivně hledá shluky bez ovlivnění již nalezených shluků. Zároveň algoritmus také rozhoduje, zda je vhodné jednotlivé vrcholy dále rozdělit, či ponechat v jednom shluku. Jinými slovy rozeznává, zda se jedná o dva vrcholky jedné hory, či o dvě různé hory. Situace je znázorněna na následujícím obrázku (Obr. 10):



Obr. 10 – Ilustrace rozhodnutí o shluku pomocí HDBSCAN

Algoritmus (Obr. 10) o rozdělení do shluků rozhoduje na základě plochy (případně objemu). Pokud součet ploch pod křivkou potenciálních vrcholů (světle zelené) je větší než obsah plochy jejich společné základny, pak jsou vrcholy jednotlivými horami. Pokud je obsah plochy základny větší, pak se jedná pouze o vrcholy jedné hory (shluku).

Výhodou algoritmů založených na hustotě je bezpochyby jejich možnost detekce šumu. Algoritmy K-means nebo K-NN nutně zařadí úplně všechna data do shluků. Oproti těmto algoritmům umí HDBSCAN zařadit data do tzv. „žádného shluku." Tím označuje ta data, která logicky nezapadají do žádného shluku, a tedy se pravděpodobně jedná o šum v datech. To napovídá, že daný záznam je pravděpodobně náhodně rozmístěn (Berba 2022).

4.5 Genetic algorithm

Genetic algorithm (GA) spadá do oblasti evolučních algoritmů, což jsou algoritmy strojového učení a umělé inteligence na bázi biologických procesů. Podobnost s biologickými procesy je v podobě mutací a postupné zpřesňování algoritmu, aby co nejlépe odpovídal konkrétní aplikaci a co nejlépe plnil daný účel.

Genetické algoritmy slouží k rozpoznávání vzorců mezi daty nebo k jejich optimalizaci. v základu se genetické algoritmy k rozpoznávání vzorců používají dvěma různými způsoby. Prvním způsobem je přímý výběr vzorců a šablon na konkrétních datech. Druhým způsobem je využití schopností optimalizace pro výpočet vah jednotlivých proměnných (vlastností dat) a jejich příprava do následný klasifikátor, např KNN. Tato možnost upravuje váhy jednotlivých proměnných optimálně tak, aby následná klasifikace byla co nejpřesnější. Váhy přiřazuje dle jejich relevance vzhledem k výsledku klasifikace (Verma a Verma 2012).

Existují tři obecné úlohy, ve kterých mohou genetické algoritmy najít uplatnění:

1. Optimalizace – Pro konkrétní řešený a jasně definovaný problém mohou genetické algoritmy na základě vstupních dat sestrojit optimální řešení, tedy optimální hodnoty proměnných tak, aby výsledek byl co nejoptimálnější (nejlépe odpovídal zadaným vlastnostem). Může se jednat např. o maximalizaci zisku výroby atp.

2. Predikce – Genetické algoritmy mohou předpřipravit půdu pro následující data miningové metody a algoritmy, a tím jim usnadnit práci a zároveň zpřesnit výsledky. Mohou rozhodnout o relevanci jednotlivých proměnných. Rozlišují, které proměnné jsou vhodné k predikci dané vlastnosti, a které jsou naopak redundantní.

3. Simulace – v některých případech je konkrétní definice řešeného problému složitá, např. existuje mnoho vstupních proměnných a neexistuje jedno konkrétní optimální řešení. v takovém případě mohou genetické algoritmy nasimulovat jak optimum, tak průběh optimálních řešení a průběžně měnit strategii tak, aby bylo optima dosaženo.

V prevenci kriminality pomocí data miningu je možné genetické algoritmy použít jako optimalizační algoritmus. Pokud existuje několik separovaných datasetů s různými informacemi, jejichž vzájemná korelace je zkoumána, pak je možné nejprve odděleně v těchto datasetech aplikovat algoritmy pro zjištění asociačních pravidel. Tato asociační pravidla je pak možné dále zkoumat pomocí genetických algoritmů.

4.6 Lineární regrese

Lineární regrese spadá do oblasti statistiky. Jedná se o metodu predikce, která se snaží data proložit vhodným způsobem (matematickou funkcí), který nejlépe popisuje chování dat. Pokud je nalezena taková funkce, je možné dle této funkce předpovídat následný vývoj určitého statistického znaku.

Regresí existuje celá řada. v tomto případě se jedná o lineární regresi, tedy hledaná matematická funkce je právě lineární funkce. Základní předpis lineární funkce:

$$y = ax + b \tag{1}$$

kde:

a … směrnice lineární funkce

b … posun po ose y

y … závislá proměnná

x … nezávislá proměnná

V software *SPSS Modeler* k výpočtu lineární regrese slouží uzel *Linear Regression*. Vhodnost vypočteného modelu ukazuje vypočtený koeficient determinace (R^2). Koeficient determinace vyjadřuje kvalitu regresního modelu. Obecně vyjadřuje, jaké procento variability dat daný model vysvětluje (Rao a Machek 1978).

Pomocí lineární regrese je možné sestavit přímku, dle které se daná data přibližně chovají. Dle předpisu této přímky se tak dá predikovat (odhadnout), jakým způsobem se budou data chovat v dosud neznámých podmínkách. Př. pokud existují data vývoje počtu trestných činů v konkrétní oblasti v letech 2012-2019, je možné s pomocí lineární regrese odhadnout, kolik trestných činů bude v dané oblasti spácháno v roce 2020.

4.7 Gradient boosting

Gradient boosting spadá do oblasti strojového učení. Je jedním z nejsilnějších nástrojů pro predikci a klasifikaci. Základním principem gradient boostingu je spojování několika tzv. slabých učňů (*weak learner*) do soustavy jednoho silného učně. Hlavním cílem tohoto algoritmu je minimalizovat ztrátu modelu, tedy snížit podíl těch dat, které model neumí pokrýt, na co nejmenší možné minimum.

Obsahuje tři základní elementy:

- ztrátovou funkci, která má být optimalizována;
- jeden či více slabých učňů;
- aditivní model, který přidá dalšího slabého učně pro lepší popis a snížení ztráty modelu.

Ztrátová funkce se liší dle charakteru řešeného problému. Například pro regresní modely je tato funkce reziduální součet čtverců.

Slabý učeň je algoritmus strojového učení nižšího řádu. Typicky se jedná o rozhodovací strom. Rozhodovací stromy vypočítají dělící hodnotu a dle ní rozdělují data do dvou skupin na dané vrstvě. Při gradient boostingu je důležité omezování, aby slabý učeň zůstal slabým.

Jedná se o iterační metodu, která po každé iteraci oddělí snadno klasifikovaná data a přidá nový rozhodovací strom, který vypočte nové parametry. Tento strom pak slouží k dodefinování zbylé množiny dat a je přidán do celkového modelu. Model jako celek je pak vylepšen a je schopný pokrýt více různých pozorování (Boehmke a Greenwell 2019).

4.8 Asociační pravidla

Algoritmus asociačních pravidel umožňuje získat závislost některých vlastností záznamů na jeho jiných vlastnostech. Jinými slovy, jeden konkrétní znak je tvořen příslušnou kombinací jiných znaků. Znaky, které jsou tvůrci závislosti, tedy jí předcházejí, se nazývají antecedenty (*antecedents*) neboli předpoklady. Znaky, které jsou touto závislostí tvořeny, se nazývají konsekventy (*consequents*) neboli závěry či následky.

Asociační pravidlo bývá psáno ve tvaru $X, Y \rightarrow Z$, kde X a Y jsou antecedenty dané závislosti a z je konsekventem. v obecné řeči tento zápis znamená, že pokud platí X a zároveň Y, pak

má daný prvek i vlastnost *Z*. Tento algoritmus se v praxi používá například v úloze analýzy nákupního košíku. Existují záznamy o nákupních košících věrných zákazníků. Tyto košíky mohou obsahovat různé typy zboží a úkolem asociačních pravidel je odhalit kombinace druhů zboží, které zákazníci kupují. Například, pokud koupí květiny a parfém, pravděpodobně koupí lahev vína spíše než člověk, který koupil mouku a cukr.

Asociační pravidla mají několik charakteristik, z nichž nejdůležitější jsou dvě – podpora (*support*) a spolehlivost (*confidence*) (García et al. 2007).

Podpora je číslo, které vyjadřuje poměr počtu záznamů, které splňují jak předpoklad, tak závěr, ku počtu všech záznamů. Vyjádřeno rovnicí (2):

$$Support = \frac{a}{a + b + c + d} \tag{2}$$

kde:

a … počet záznamů, které splňují předpoklad i závěr

b … počet záznamů, které splňují předpoklad, ale ne závěr

c … počet záznamů, které splňují závěr, ale nesplňují předpoklad

d … počet záznamů, které nesplňují ani předpoklad, ani závěr.

Spolehlivost je číslo, které vyjadřuje poměr počtu záznamů, které splňují předpoklad, ku počtu všech záznamů, které splňují závěr. Vyjádřeno rovnicí (3):

$$Confidence = \frac{a}{a + c} \tag{3}$$

kde

a … počet záznamů, které splňují předpoklad i závěr

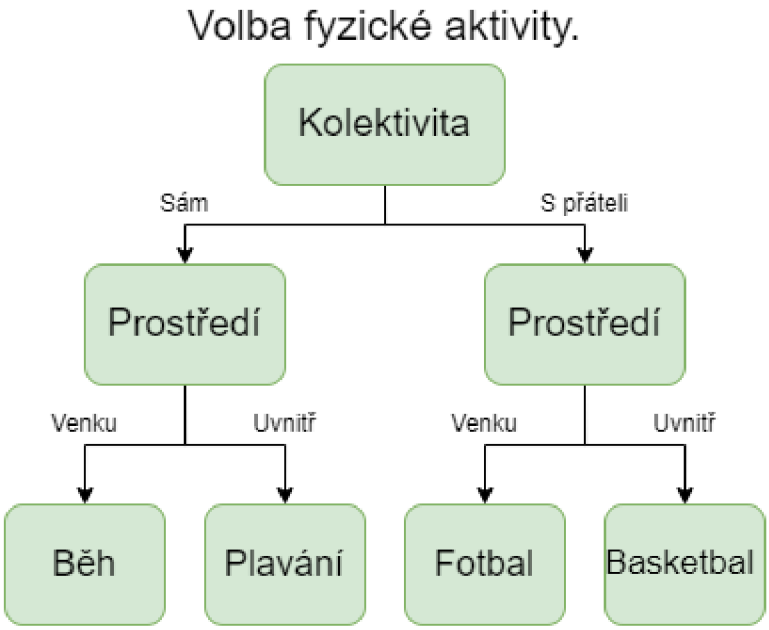
c … počet záznamů, které splňují závěr, ale nesplňují předpoklad.

4.9 Rozhodovací stromy

Rozhodovací stromy jsou algoritmy sloužící ke klasifikaci záznamů. Model rozhodovacího stromu je naučen na trénovací sadě záznamu, že které vydestiluje pravidla pro rozhodování. Pomocí tohoto modelu pak můžou být nová data zařazena do příslušných skupin dle nalezených pravidel.

Rozhodovací stromy pracují na principu „rozděl a panuj“. Každý strom se skládá z alespoň jednoho rozhodovacího rozvětvení. Tomuto rozvětvení se říká vrstva. Každá vrstva představuje jedno konkrétní pravidlo s jeho vlastní prahovou hodnotou pro rozhodnutí. Pro

inicializaci modelu je nutné určit cílovou proměnnou (*target field*). U algoritmu C5.0 může být cílová proměnná pouze nominální či kategoriální. Model při učení nejprve vybere proměnnou (sloupec), která poskytne maximální informační zisk. Pak model rozhodne o prahové hodnotě dané proměnné, dle které bude základní vzorek rozdělen do dvou skupin. Strom pracuje rekurzivně, a tak dělení pokračuje na již rozděleném vzorku.



Obr. 11 – Příklad rozhodovacího stromu

Na obrázku (Obr. 11) je zobrazen příklad jednoduchého rozhodovacího stromu, jehož úkolem je určit vhodnou fyzickou aktivitu dle zadaných kritérií. Rozhodovací proměnné mohou být jakéhokoli druhu, ne pouze nominální či kategoriální, jako cílová proměnná. Strom může růst teoreticky donekonečna, dokud nenajde definici, tedy sekvenci pravidel, pro každou podmnožinu cílové proměnné. Omezení stromu může být zadáno před modelováním jako maximální počet rozhodovacích úrovní (Yobero 2018).

Dalšími druhy rozhodovacích stromů jsou například C&RT, CHAID či MARS.

5 Dataset

Dataset, který je zpracovaný v rámci této diplomové práce, je složen ze skutečných záznamů o kriminalitě na území České republiky, konkrétně na území okolo města Liberce. Záznamy byly pořízeny příslušníky městské policie. Data jsou ověřená a pravdivá.

5.1 Základní dataset

Dataset obsahuje záznamy v období od začátku roku 2013 do konce roku 2021, konkrétně od 1. ledna 2013 do 22. prosince 2021. Celkem obsahuje pět polí:

- *id* – Jedná se o unikátní číselné označení konkrétního incidentu,
- *x* – číselná hodnota, která odpovídá zeměpisné délce,
- *y* – číselná hodnota, která odpovídá zeměpisné šířce,
- *date* – datum a čas spáchání incidentu,
- *type* – druh incidentu a rodina, do které daný druh spadá.

5.1.1 Příprava dat

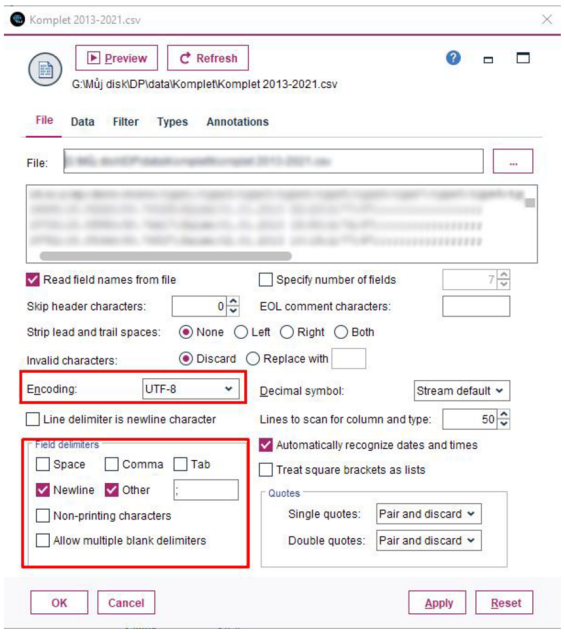
Dataset byl poskytnut v datovém formátu „.csv“, nebyl však správně strukturovaný. Jednotlivé záznamy byly ve formě:
id,xx.xxxxx,yy.yyyyy,RRRR-MM-DDTHH:MM:SS.SSSS+0Z.00,“typ“

Z této formy lze data, pomocí programu Microsoft Excel, snadno převést tak, aby sloupce byly odděleny a hodnoty napříč sloupci a řádky také. SPSS Modeler obsahuje podobné funkce, avšak pro jednodušší následnou práci s datasetem byl využit MS Excel.

Hromadnou úpravou musel projít sloupec se záznamem času. Formát času, který byl v datasetu, nebyl rozpoznaný jako platný časový formát. Pomlčky byly nahrazeny lomítkem, místo písmena „T“ byly vloženy mezery a byla odstraněna část, která určuje časové pásmo. Po této úpravě už je časový formát SPSS Modelerem rozpoznán jako spojitá veličina a je připraven k analýze.

Dále pak bylo nutné upravit sloupec s označením typu incidentu. Každý incident je označený několika kódy, kdy každý kód odpovídá konkrétnímu typu. Všechny kódy byly uvedené v jedné buňce, pročež s nimi nebylo možné pracovat odděleně. Před započetím analýz bylo

nutné jednotlivé kódy oddělit. Po této úpravě bylo možné dataset nahrát do programu SPSS Modeler.



Obr. 12 – Nastavení vstupního uzlu SPSS Modeler

Na obrázku (Obr. 12) je možné vidět nastavení uzlu pro vstup datového souboru. Kódování je nutné nastavit na UTF-8, aby souhlasilo s kódováním datasetu. Dále je nutné označit, kde se v datech nachází nový řádek (zaškrtnuté políčko *Newline*) a jakým způsobem jsou vzájemně odděleny sloupce (v tomto případě středníkem). Od tohoto momentu je datový soubor dostupný pro všechny úpravy.

Upraven musel být i sloupec „*typ*“. Tento sloupec obsahoval sloučená označení deliktu v jednom textovém řetězci. Každý delikt má své označení neboli tag. Tagy jsou definovány tak, jak je vymezuje zákon. Tento tag jednoznačně určuje podstatu daného deliktu. Tyto tagy mají hodnotu řazenou vzestupně dle závažnosti daného činu. Například tag *A* je násilná trestná činnost, *B* jsou požáry až po *Z* – přestupky. Každý delikt může být označen několika různými tagy podle své podstaty. Záznam tedy může obsahovat více tagů najednou. Všechna tato označení jsou sloučena v jednom řetězci jako hodnota sloupce *typ*. Například, pokud byla spáchána trestná činnost násilná zároveň s požárem, bude mít záznam o tomto deliktu ve sloupci typ hodnotu „*A, B*“. Každý delikt může být označen několika tagy. Situace je znázorněna v následujících tabulkách.

Tabulka 1 – Ukázka záznamu o deliktu včetně označení typu v původní tabulce

id	…	typ
1234	…	A, B, Z
5678	…	A, D

Jak je znázorněno v tabulce (Tabulka 1), každý delikt může mít více označení. Počet tagů se pohybuje mezi 2 až 19 tagy. V tomto formátu nebylo možné s označeními nijak pracovat, a proto musela být jednotlivá označení od sebe oddělena. První způsob oddělení je naznačen v následující tabulce (Tabulka 2).

Tabulka 2 – Ukázka záznamu o deliktu včetně označení v upravené tabulce

id	…	typ_1	typ_2	typ_3	…
1234	…	A	B	Z	…
5678	…	A	Z	-	…

Tímto způsobem je možné od sebe oddělit jednotlivá označení. Problém je ale různý počet tagů pro jednotlivé delikty. Kdyby bylo potřeba filtrovat například všechny přestupky (pro tento případ označení *Z*), není možné určit, ve kterém sloupci se má daná hodnota nacházet. Pokud by byla data filtrovaná podle sloupce *typ_2*, jako přestupek by byl označen pouze delikt s *id = 5678* a delikt s *id=1234* by byl vyřazený. Naopak, pokud by bylo filtrováno podle sloupce *typ_3*, situace by byla opačná. Z tohoto důvodu bylo nutné data dále upravit, aby bylo možné data filtrovat bez opomenutí všech tagů, kterými je označen. Situace je znázorněna v následující tabulce (Tabulka 3).

Tabulka 3 – Ukázka záznamu o deliktu včetně označení po transponování

id	…	A	B	F	Z	…
1234	…	1	1	0	1	…
5678	…	1	0	0	1	…

Data byla transponována. Ze všech hodnot, kterými může být delikt označen, byly vytvořeny sloupce. Pro datovou reprezentaci byla pomocí těchto sloupců vytvořena matice sousednosti. Každý řádek (záznam) s deliktem byl tedy rozšířen o všechny možné tagy (pro tento modelový případ *A, B, F a Z*) a hodnota sloupce byla nastavena na *1* pokud daný delikt

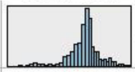
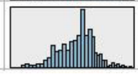
obsahoval příslušné označení a hodnotou *0* pokud daným tagem označen nebyl. Po této úpravě bylo možné s daty efektivně a správně pracovat. Pokud bylo například nutné filtrovat všechny přestupky jako v předchozím případě, stačí nastavit podmínku $Z > 0$, pak budou vybrány všechny řádky matice, kde Z je různé od *0* a tedy se jedná o všechny přestupky. Výběr je pak nezávislý na všech ostatních označeních deliktu. Data byla restrukturalizována pomocí programu MS Excel.

Jednotlivé tagy mají hodnotu 1 až 102 podle typu daného deliktu. Všechny tyto tagy byly v jednom textovém řetězci v rámci jednoho sloupce u každého záznamu o deliktu. Jednotlivé tagy od sebe byl odděleny čárkami. MS Excel nabízí možnost rozdělit textové řetězce do jednotlivých sloupců, pokud je textový řetězec konkrétním způsobem strukturovaný. V tomto případě byly jednotlivé tagy rozděleny pomocí oddělovače čárka. Do matice byly přidány sloupce, v jejichž hlavičce byly hodnoty 1 až 102, které reprezentují tagy deliktů. Pro každý sloupec byla definována funkce IF (když). Ta zkoumá, jestli na řádku v nově vzniklých sloupcích po rozdělení tagů existuje práva hodnota z hlavičky konkrétního sloupce. Pokud existuje, daný delikt má onen tag a hodnota sloupce je rovna 1, pokud tag z daného sloupce neobsahuje, hodnota je rovna 0 (viz. Tabulka 3).

Celý takto upravený řádek je pak spojen do jednoho řetězce, který obsahuje pouze číslice 1 a 0. Například vzniklý řetězec 10001101 má tagy 1, 5, 6 a 8. Jednotlivé tagy jsou seřazeny vzestupně podle závažnosti. Úkolem této restrukturalizace bylo mj. také určit dominantní typ deliktu čili nejzávažnější tag. Hodnotu nejzávažnějšího tagu je možné získat podle pozice první číslice 1 ve vzniklém řetězci. Funkce Najít vrátí pozici prvního výskytu daného znaku v řetězci. Pozice 1 v řetězci odpovídá danému tagu.

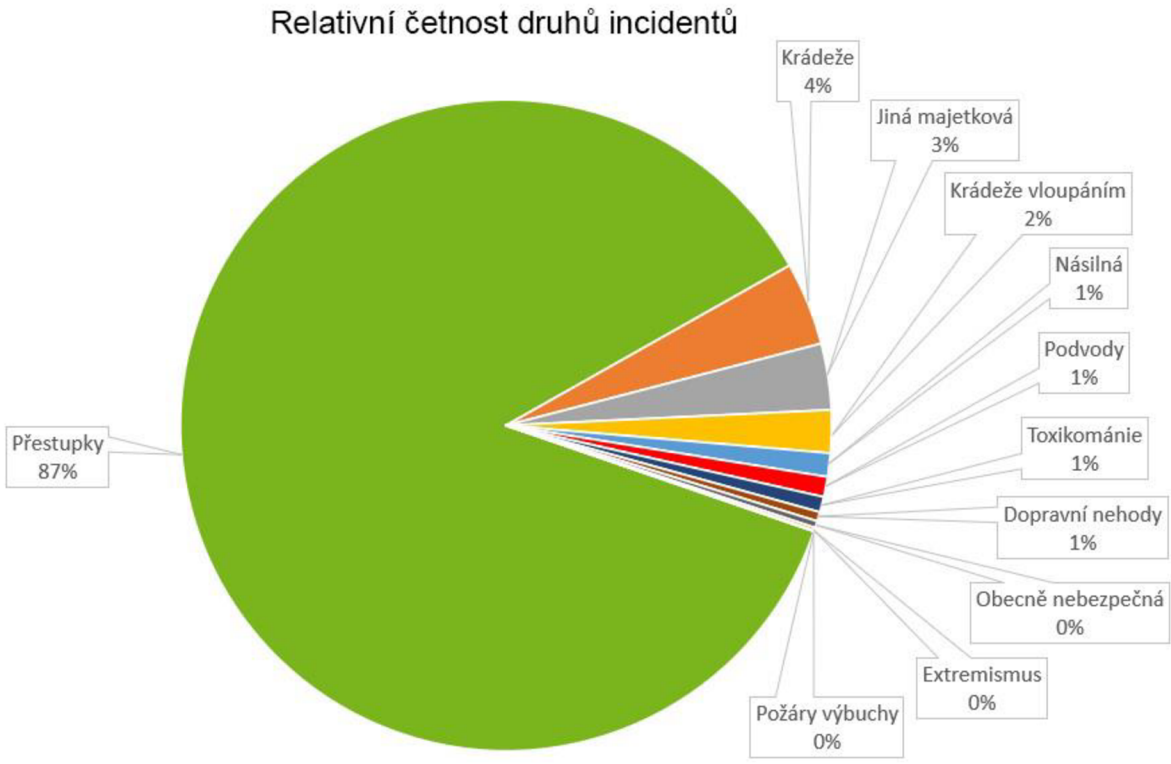
5.1.2 Porozumění datům

Prvním krokem je základní analýza dat. Tu poskytne uzel *Data Audit*, který slouží k zobrazení několika základních charakteristik záznamů, např. jaké jsou datové typy jednotlivých sloupců (proměnných), a také slouží k odhalení vlastností datasetu, např. počet chybějících hodnot aj.

Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
x		Continuous	14.960	15.132	15.056	0.019	-0.008	--	109112
y		Continuous	50.708	50.823	50.763	0.015	0.018	--	109112

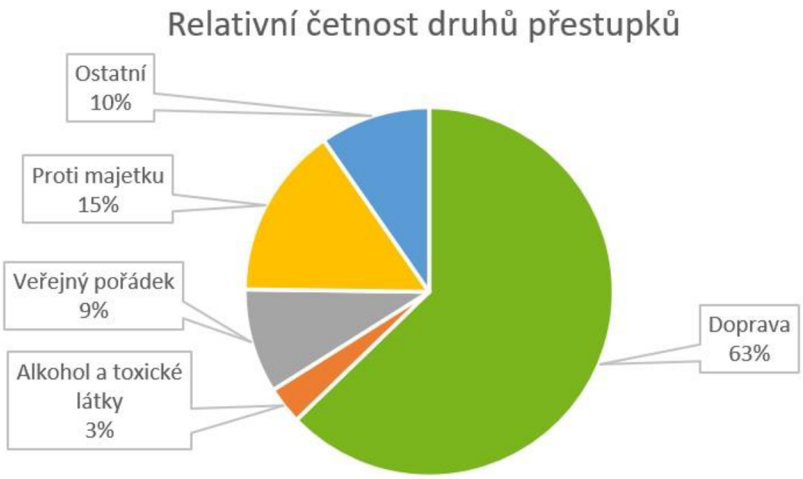
Obr. 13 – Výstup z uzlu Data Audit

Výstup z uzlu *Data Audit* (Obr. 13) zobrazuje tabulku s několika sloupci. Sloupec *Field* označuje, ke které proměnné z datasetu se vážou. *Sample Graph* ukazuje hrubý náhled na vzhled grafu této dané proměnné. Jedná se o graf četností. Sloupec *Measurement* označuje, o jaký typ proměnné se jedná, v tomto případě se jedná o proměnnou spojitou (*Continuous*). Uzel *Data Audit* ukazuje informace o proměnných *x* a *y*, tedy o zeměpisné délce a šířce. Sloupce *Min* a *Max* uvádějí minimální a maximální hodnoty ze všech, které se v datasetu objevují. Z toho vyplývá, že zkoumaná oblast se nachází mezi 14,96 ° až 15,132 ° východní délky a 50,708 ° až 50,823 ° severní šířky. Dle měření má oblast 12,1 km od východu na západ a 12,78 km od severu k jihu. Sloupec *Mean* je pro průměr. Sloupec *Std. Dev* odpovídá směrodatné odchylce Normálního rozdělení a *Skewness* ukazuje šikmost daného rozdělení. Tyto dva sloupce však v tomto případě nejsou relevantní. Důležitý je sloupec *Valid*, který ukazuje celkový počet záznamů v datasetu. Ten činí celkem 109 112 záznamů.



Graf 1 – Graf relativních četností jednotlivých druhů incidentů

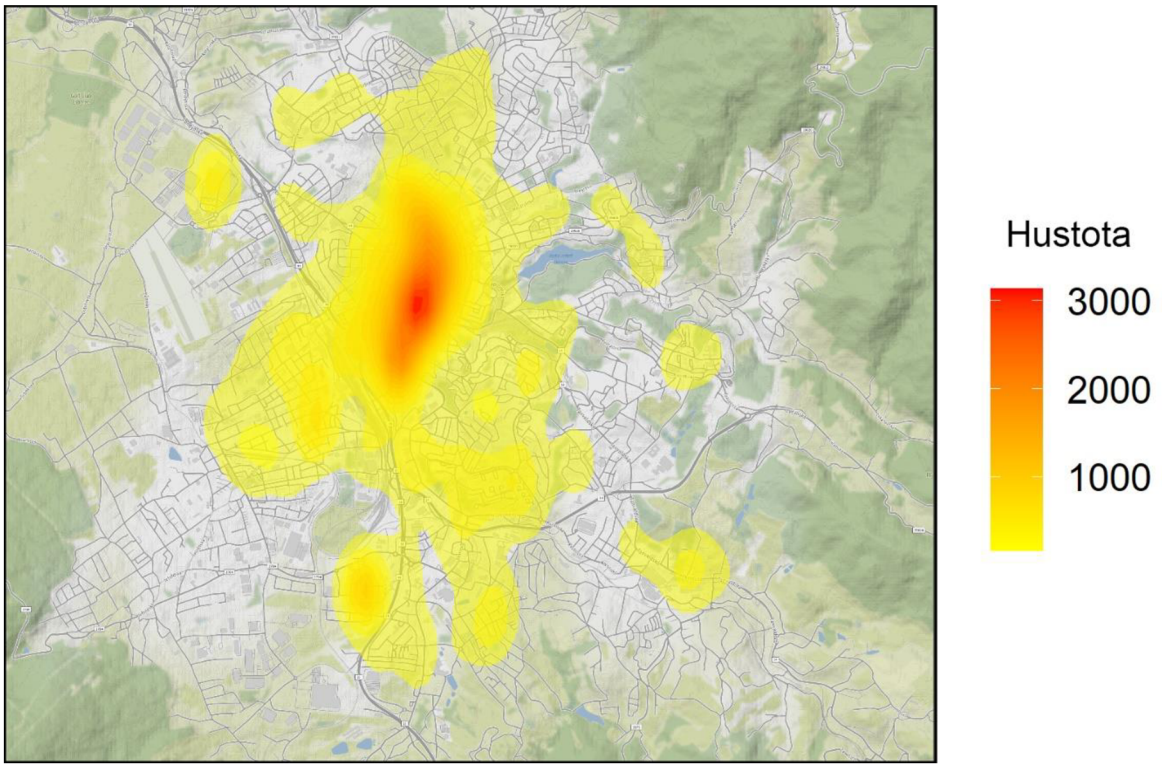
V grafu (Graf 1) je možné vidět rozložení jednotlivých druhů incidentů v celém datasetu. Typem činu, který se vyskytuje nejčastěji, je čin spadající do kategorie *Přestupky*. Celkem tvoří 87 % všech incidentů. Kategorie *Přestupky* má své jemnější dělení: přestupek v oblasti dopravy, přestupek s přítomností alkoholu či jiných omamných látek, přestupek proti veřejnému pořádku, přestupek proti majetku, ostatní přestupky (bez specifického zařazení). Rozdělení jednotlivých druhů přestupků je zobrazeno v grafu (Graf 2).



Graf 2 – Graf relativních četností druhů přestupků

Důležitým aspektem, který může hrát velkou roli v prevenci, případně predikci, incidentů, je čas. z hlediska času mohou (ale nemusí) být incidenty náhodně rozmístěné. To je však vhodné ověřit.

Záznamy o přestupcích jsou bohužel zaznamenány jen mělce. U jednotlivých přestupků nelze určit, o jaký konkrétní přestupek šlo. Pod označení přestupek proti veřejnému pořádku se řadí například rušení nočního klidu či poškození turistické značky apod. Z tohoto důvodu byly na doporučení odborníky vyřazeny z datasetu.



Obr. 14 – Heat mapa trestných činů v oblasti města Liberce (období 2013-2019)

Heat mapa (Obr. 14) je výstupem uzlu Heatmap v programu SPSS Modeler. Poskytuje prvotní náhled na situaci ohledně rozložení trestných činů ve zvolené zeměpisné oblasti. Barva označuje (dle stupnice vpravo) hustotu deliktů v dané oblasti. Poměrně zřetelně hustá oblast je v centru města. Vedle tohoto ohniska je pak zvýrazněno několik dalších oblastí. Vlevo nahoře je menší shluk v okolí nákupního centra NC Géčko, vlevo dole okolo nákupního centra OC Nisa. Vpravo jsou zvýrazněné oblasti sídlišť – vpravo nahoře Starý Harcov (okolí univerzitních kolejí), pod ní sídliště Kunratická a vpravo dole městská část Vratislavice nad Nisou. Nevybarvené oblasti mají v poměru ke zvýrazněným oblastem výrazně nižší hustotu deliktů.

5.2 Rozšíření datasetu

Základní dataset poskytuje jen velmi málo informací, které by mohly vést k efektivní predikci kriminality. Jak napovídají zmíněné publikace na toto téma v rešerši, je více než žádoucí základní dataset vhodným způsobem rozšířit o doplňující data, která by mohla vést k detailnějšímu porozumění vztahů mezi jednotlivými daty v základním datasetu. Data mohou být různého druhu. V tomto případě bohužel nejsou známé žádné doplňující informace o pachateli, demografických aspektech nebo ekonomicko-zeměpisných aspektech. Je tedy nutné vycházet pouze z dat veřejně dostupných a téměř výlučně geografických.

5.2.1 OpenStreetMaps

Jedním z veřejně dostupných zdrojů s (nejen) geografickými údaji je webová stránka *OpenStreetMaps* (Allan 2022). Díky této webové stránce je možné nahlížet do záznamů globální databáze *open source* projektu *OpenStreetMap*, jehož hlavním cílem je shromáždit veškeré dostupné geografické informace ze všech zemí, které jsou ochotné se projektu zúčastnit. Zmíněná webová stránka je ale výhradně prostředkem pro zobrazení a má jen velmi omezené možnosti exportu dat. Z předchozí kapitoly je známé, že zkoumaná oblast se nachází mezi souřadnicemi 14,960 ° a 15,132 ° východní délky a 50,708 ° a 50,823 ° severní šířky. Takto velká oblast (cca 150 km^2) je pro export ze zmíněné stránky příliš velká. Jsou však dostupné další nástroje, které poskytují kompletní databázový soubor se všemi geografickými údaji, jako jsou uzly, cesty i relace. Webová stránka (OSMF Data Working Group 2022) nabízí možnost stažení kompletní databáze pro celý svět. Ta je bohužel pro účely této práce zbytečně velká (112 GB) a její zpracování by bylo příliš časově náročné. Několik států však spravuje vlastní datovou oblast dané databáze, a tak poskytuje

databázový soubor, který již obsahuje pouze data z daného státu. Není poté nutné stahovat soubor obsahující celý svět, ale stačí jen soubor příslušného státu. V České republice takový soubor poskytuje, spravuje a pravidelně aktualizuje instituce Vysoké učení technické v Brně (VÚT Brno 2022). Velikost tohoto souboru je pouze 770 MB.

Ačkoliv je soubor dostatečně malý na to, aby práce s ním nebyla příliš časově náročná, obsahuje stále příliš velkou oblast (území celé ČR). Proto je nutné data dále extrahovat. Soubor má koncovku „*pbf*[6]“, což znamená, že je pro práci s ním potřeba speciální program. K tomuto účelu slouží program Osmosis (Migurski 2020), který je založen, podobně jako celý projekt, na principu open source. Program je určen pro uživatele operačního systému Linux, avšak existuje i verze pro uživatele operačního systému Windows. Program se instaluje pomocí balíčku v souboru s příponou „*bat*“. Program funguje pouze v textovém rozhraní příkazové řádky jako terminál pro dotazy do databáze. Ta je však uložena lokálně ve zmíněném PBF souboru.

```
D:\TUL\Studium\EF - SI\DP\Osmosis\bin>osmosis --read-pbf CR_big.osm.pbf --bounding-box
 left=14.96 right=15.132 top=50.823 bottom=50.708 --write-xml liberec.xml
Led 17, 2022 5:48:01 ODP. org.openstreetmap.osmosis.core.Osmosis run
INFO: Osmosis Version 0.48.3
Led 17, 2022 5:48:02 ODP. org.openstreetmap.osmosis.core.Osmosis run
INFO: Preparing pipeline.
Led 17, 2022 5:48:02 ODP. org.openstreetmap.osmosis.core.Osmosis run
INFO: Launching pipeline execution.
Led 17, 2022 5:48:02 ODP. org.openstreetmap.osmosis.core.Osmosis run
INFO: Pipeline executing, waiting for completion.
Led 17, 2022 5:49:49 ODP. org.openstreetmap.osmosis.core.Osmosis run
INFO: Pipeline complete.
Led 17, 2022 5:49:49 ODP. org.openstreetmap.osmosis.core.Osmosis run
INFO: Total execution time: 107834 milliseconds.
```

Obr. 15 – Výstup z programu Osmosis

Na obrázku (Obr. 15) je možné vidět zadání příkazu pro extrakci dané oblasti ze souboru, který obsahuje informace o celé České republiky. Program se spustí příkazem, který začíná „osmosis“, následuje část příkazu, která udává, co je úkolem programu – „--read-pbf“ znamená, že má program otevřít soubor typu PBF. Poté následuje název daného souboru „CR_big.osm.pbf“, za ním následuje část příkazu „--bounding-box“. Pomocí tohoto příkazu dojde k extrakci určité oblasti, která je definována parametry geografických souřadnic, a sice západní, východní, severní a jižní hranice oblasti. Poslední část příkazu „--write-xml“ určuje, co se má s výstupem stát. v tomto případě jde o zápis do XML souboru, jehož název je „liberec.xml“. s tímto souborem je již možné dále pracovat. Veškeré uzly, cesty a vztahy

[6] PBF – Protocol Binary Format, datový formát, který je alternativou ke XML (Extensible Markup Language) (OSMF Data Working Group 2021)

mezi nimi z této oblasti jsou přehledně uložené v jednom malém souboru o konečné velikosti 80 MB.

Soubor je kompaktní a dobře strukturovaný, neobsahuje však strukturu, která by byla dobře zpracovatelná do takové míry, aby bylo jednoduché z něj extrahovat potřebná data (například všechny autobusové zastávky se souřadnicemi atd.). Náhled XML souboru je vidět na obrázku (Obr. 16).

```
<?xml version='1.0' encoding='UTF-8'?>
<osm version="0.6" generator="Osmosis 0.48.3">
  <bounds minlon="14.96000" minlat="50.70800" maxlon="15.13200" maxlat="50.82300" origin="0.48.3"/>
  <node id="73280136" version="3" timestamp="2011-09-22T11:44:28Z" uid="0" user="" lat="50.8181121" lon="15.0365272">
    <tag k="source" v="HELP SERVICE - REMOTE SENSING spol. s r.o. http://www.bnhelp.cz"/>
  </node>
  <node id="73280474" version="6" timestamp="2014-12-18T21:52:07Z" uid="0" user="" lat="50.7798737" lon="15.0304471">
    <tag k="source" v="HELP SERVICE - REMOTE SENSING spol. s r.o. http://www.bnhelp.cz"/>
  </node>
```

Obr. 16 – Náhled XML souboru s daty

Z tohoto důvodu bylo potřeba napsat jednoduchý program, který umožní pohodlně extrahovat data zvoleného typu s vlastnostmi, které jsou žádoucí pro doplnění základního datasetu. Ukázka kódu je vidět na obrázku (Obr. 17).

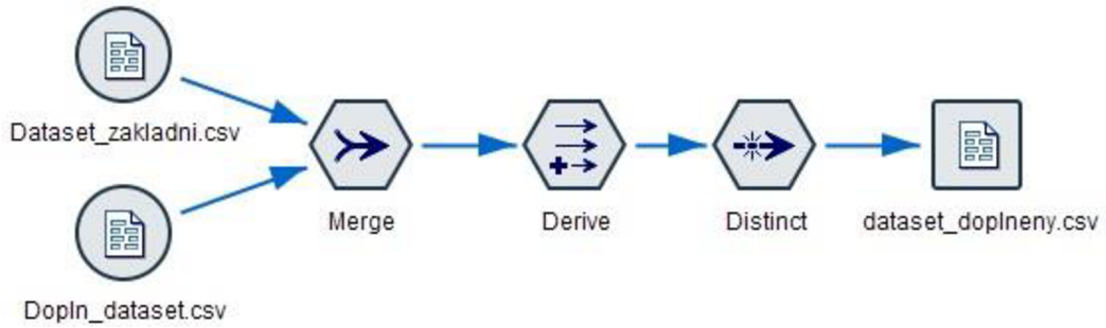
```
Počet odkazů: 0
class Program
{
    Počet odkazů: 0
    static void Main(string[] args)
    {
        ReadRecords();
        Console.ReadKey();
    }
    Počet odkazů: 1
    static void ReadRecords()
    {
        XmlDocument doc = new XmlDocument();
        int idnum = 0;
        doc.Load("D:\\TUL\\Studium\\EF - SI\\DP\\Mapa\\liberec.xml");
        XmlNodeList nodes = doc.SelectNodes("nodes/node");
        foreach (XmlNode node in nodes)
        {
            if (node.Name == "node")
            {
                if (node.HasChildNodes)
                {

                    foreach (XmlNode cnodes in node.ChildNodes)
                    {
                        if (cnodes.Attributes[0].Value == "highway" && cnodes.Attributes[1].Value == "bus_stop")
                        {
                            foreach (XmlNode cnodes2 in node.ChildNodes)
                            {
                                if (cnodes2.Attributes[0].Value == "name")
                                {
                                    idnum++;
                                    Console.WriteLine(idnum + " " + node.Attributes[0].Value + " " + node.Attributes[5].Value + " " +
                                        node.Attributes[6].Value + " " + cnodes2.Attributes[1].Value);
                                }
                            }
                        }
                    }
```

Obr. 17 – Ukázka kódu vytvořeného programu pro extrakci dat

Program byl následně doplněn o možnost exportu extrahovaných dat do datového souboru typu CSV, který je možné připojit k základnímu datasetu (základní dataset je také ve formátu CSV). S datovým formátem tohoto typu umí pracovat většina data miningových softwarů.

Ze zmíněného souboru se všemi daty z dané oblasti bylo pomocí tohoto programu extrahováno celkem osm doplňujících datasetů – zastávky, restaurační zařízení, hřbitovy, bankomaty, obchody, policie (policejní stanice), pošty a benzinky. Všechny doplňující datasety obsahují informace o poloze ve stejném formátu jako základní data, tedy *x a y* zeměpisné souřadnice. K souřadnicím je většinou připojen popis ve formě názvu. Datasety s restauracemi a obchody obsahují ještě informaci o typu. Obchody mohou být například s elektronikou nebo potravinami. Restaurační zařízení mohou být restaurace, kavárny, hospody a bary.

Tato data byla použita tak, aby ke každému záznamu o deliktu byl přiřazen právě jeden takový záznam. Jinými slovy, aby každému deliktu byl přiřazen nejbližší bod z daného datasetu, tedy nejbližší zastávka, nejbližší bankomat apod.



Obr. 18 – Stream pro nalezení nejbližšího prvku danému deliktu

Stream (Obr. 18) se skládá celkem z šesti uzlů. První uzly (nejvíce vlevo) jsou uzly, které nahrají datasety do prostředí programu. Uzel *Merge* slouží ke spojení dvou či více datasetů do jednoho na základě zvolených pravidel. Řádky nové matice jsou tvořeny průnikem všech deliktů z původního datasetu a všech možností z doplňujícího datasetu. Situace a princip spojení datasetů jsou znázorněny v následujících tabulkách. Tabulky (Tabulka 4, Tabulka 5) představují modelové původní datasety před spojením. Tabulka (Tabulka 6) zobrazuje výsledný dataset po spojení uzlem *Merge*.

Tabulka 4 – Modelový dataset 1

Název	Hodnota
A	AA
B	BB

Tabulka 5 – Modelový dataset 2

Název	Hodnota
C	CC
D	DD

Tabulka 6 – Modelová tabulka jako výstup uzlu Merge

1_Název	1_Hodnota	2_Název	2_Hodnota
A	AA	C	CC
A	AA	D	DD
B	BB	C	CC
B	BB	D	DD

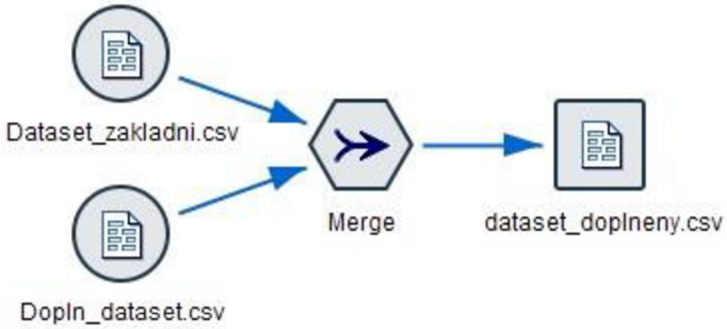
Pokud jeden dataset má *m* řádků a druhý dataset má *n* řádků, pak výsledná datová matice má $m \times n$ řádků.

Uzel *Derive* slouží k vytvoření nového sloupce v datové matici pomocí výpočtu z již existujících sloupců. V tomto případě je přidán sloupec *vzdalenost*. Vzdálenost je vypočtena jako vzdálenost dvou bodů mezi pozicí deliktu a pozicí příslušného bodu (bankomat, obchod apod). Pro reálnou hodnotu v měrných jednotkách je nutný přepočet. Hodnota sloupce *vzdalenost 0,001160666* přibližně odpovídá vzdálenosti *100 m*.

Uzel *Distinct* slouží k vyřazení řádků datové matice, pokud spadají do nějaké skupiny. Uzel je nastaven tak, aby data seskupil podle všech sloupců, kromě vypočtené vzdálenosti. Ta pak slouží k seřazení vzniklé skupiny vzestupně. S nastavením „*include only the first record in each group*“ vynechá všechny řádky kromě prvního v dané skupině. To znamená, že je ponechán pouze řádek s nejnižší hodnotou vzdálenosti včetně informace, k jakému bodu je tato vzdálenost nejmenší. Postupně jsou takto přidány všechny doplňující datasety, čímž se zjistí, k jakému bodu (zastávce, obchodu apod.) má daný delikt nejblíže.

5.2.2 Hydrometeorologické údaje

Dalším způsobem, jak rozšířit dataset, je přidat informace o hydrometeorologických podmínkách, které panovaly v den, kdy k deliktu došlo. Server *www.chmi.cz* (ČHMÚ 2022) poskytuje záznamy o počasí pro konkrétní stanici zpětně až do roku 1961. Pro účely této práce byly vybrány záznamy o průměrné teplotě vzduchu, tlaku, relativní vlhkosti, úhrnu srážen a počtu hodin slunečního svitu na území Liberce.



Obr. 19 – Spojení datasetu s delikty a počasím

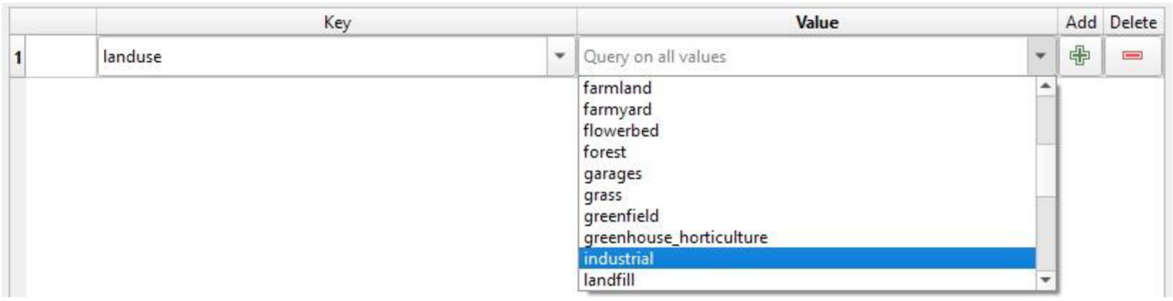
V tomto případě (Obr. 19) je uzel *Merge* použit jiným způsobem. Data s počasím obsahují datum, stejně jako dataset s delikty. Spojení proběhlo na principu „*inner join*" – spojení na základě shodné hodnoty. Pro každý řádek s deliktem je vyhledán takový řádek v datasetu s počasím, kde datum deliktu odpovídá datu s počasím. Následně jsou data o počasí připojena k deliktu. Výsledný počet řádků se nezmění, počet sloupců je navýšen o počet přidaných proměnných, v tomto případě o pět sloupců.

5.2.3 QGIS

Dalším možným způsobem, jak rozšířit dataset, je využít schopností a možností programu QGIS, který spadá do rodiny programů GIS. Tento program obecně slouží k práci s mapami a mapovými vrstvami. Program jako takový je schopný pracovat s rozsáhlou škálou dat. Velmi dobré a užitečné je spojení s výše zmíněným OpenStreetMaps. Na rozdíl od předchozího případu extrakce jednotlivých záznamů o bodech umí program QGIS z databáze vyextrahovat také oblasti, tzv. polygony. Těmito polygony jsou popsány plošné geografické útvary. Jedná se typicky o louky, pole, mokřady apod.

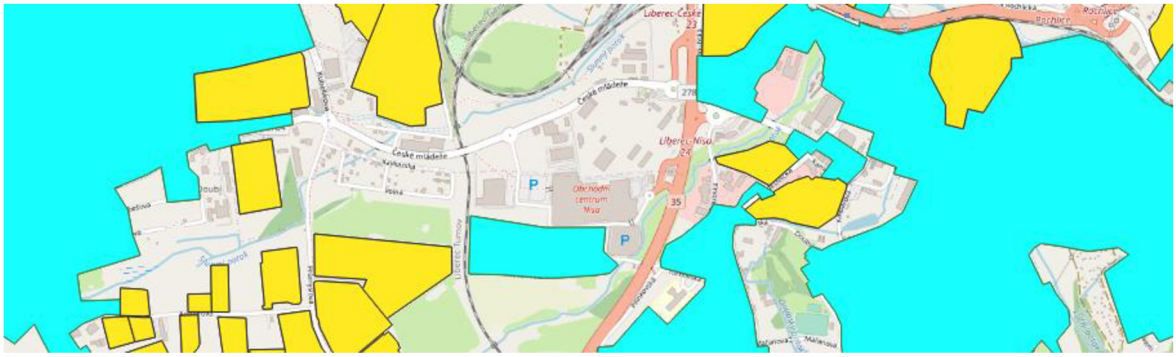
Do programu QGIS je pro práci s OSM nejprve nutné doinstalovat doplněk, který je schopný zacházet s OSM databází. Tento doplněk se jmenuje QuickOSM. Hlavní funkcí doplňku jsou dotazy do databáze ve formě žádostí o poskytnutí nějaké konkrétní informace. Data v OSM databázi jsou buď bodové, nebo plošné. Všechny záznamy mají své atributy a jejich hodnoty.

Vyhledávání v databázi pak probíhá na základě atributu (klíče) a hodnoty daného atributu. V tomto případě poslouží k rozšíření datasetu informace, zda byl daný trestný čin spáchán v rámci rezidenční čtvrti nebo např. v průmyslové zóně. Atribut, který se týká tohoto třídění, se nazývá *landuse*, neboli typ užití dané plochy. Každý atribut má vlastní seznam možných hodnot daného atributu. Způsob, jakým se volí parametry daného dotazu, je zobrazen na následujícím obrázku (Obr. 20):



Obr. 20 – Volba parametrů dotazu v QuickOSM

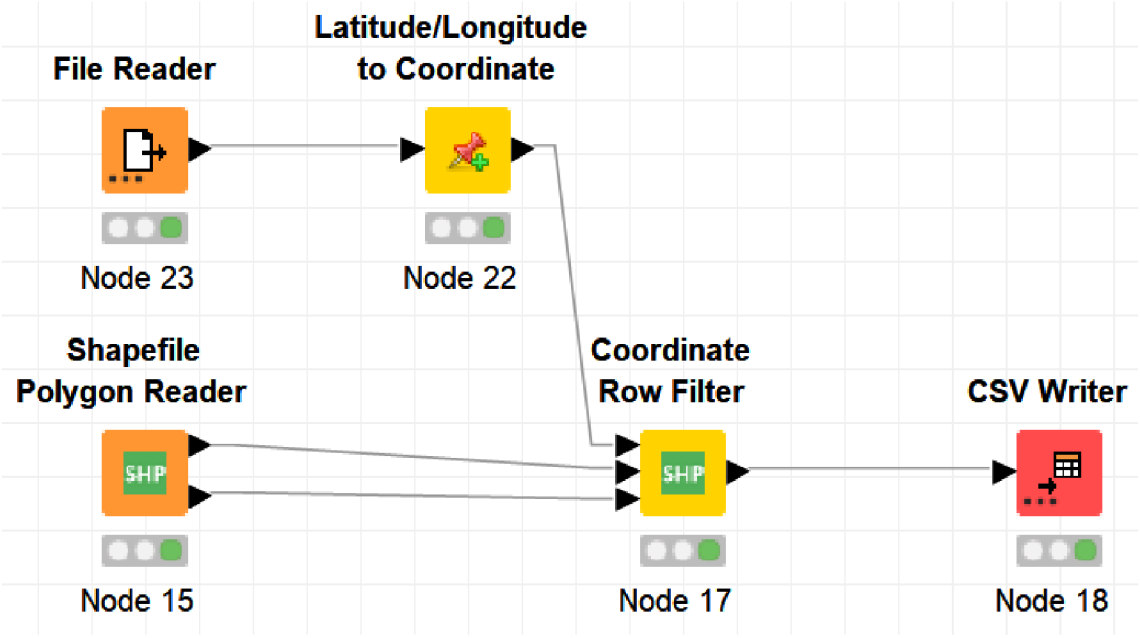
V rámci jednoho dotazu je možné použít více vnořených dotazů. V tomto případě byl zvolen klíč *landuse* a jeho hodnota *residential*, která označuje rezidenční oblast, a *industrial*, která označuje průmyslovou zónu. Výsledkem dotazu je pak nová vrstva nad původní mapou, která obsahuje polygony (mnohoúhelníky), které vymezují dané oblasti. Ukázka mapy po spuštění dotazu je v následujícím obrázku (Obr. 21):



Obr. 21 – Ukázka výstupu dotazu QuickOSM

Na obrázku (Obr. 21) jsou zobrazeny polygony obou typů oblastí. Žluté polygony označují průmyslové oblasti, tyrkysové polygony označují rezidenční čtvrti. Nevyznačené oblasti jsou jiného typu využití, např. škola, les, obecná plocha apod. Informace o polygonech, tedy jejich polohy a vymezení oblasti, je pak možné exportovat do souboru typu *shapefile*, se kterým je možné dále pracovat v jiných programech, jako například software KNIME. Na základě takto exportovaných polygonů lze ke každému záznamu doplnit informaci, zda byl

delikt spáchán v rezidenční oblasti, průmyslové zóně, či jiné další. V tomto případě se jedná pouze o tyto dva typy oblasti. K přiřazení oblasti každému záznamu byl použit software KNIME s využitím rozšíření Palladian a ESRI Shapefiles.

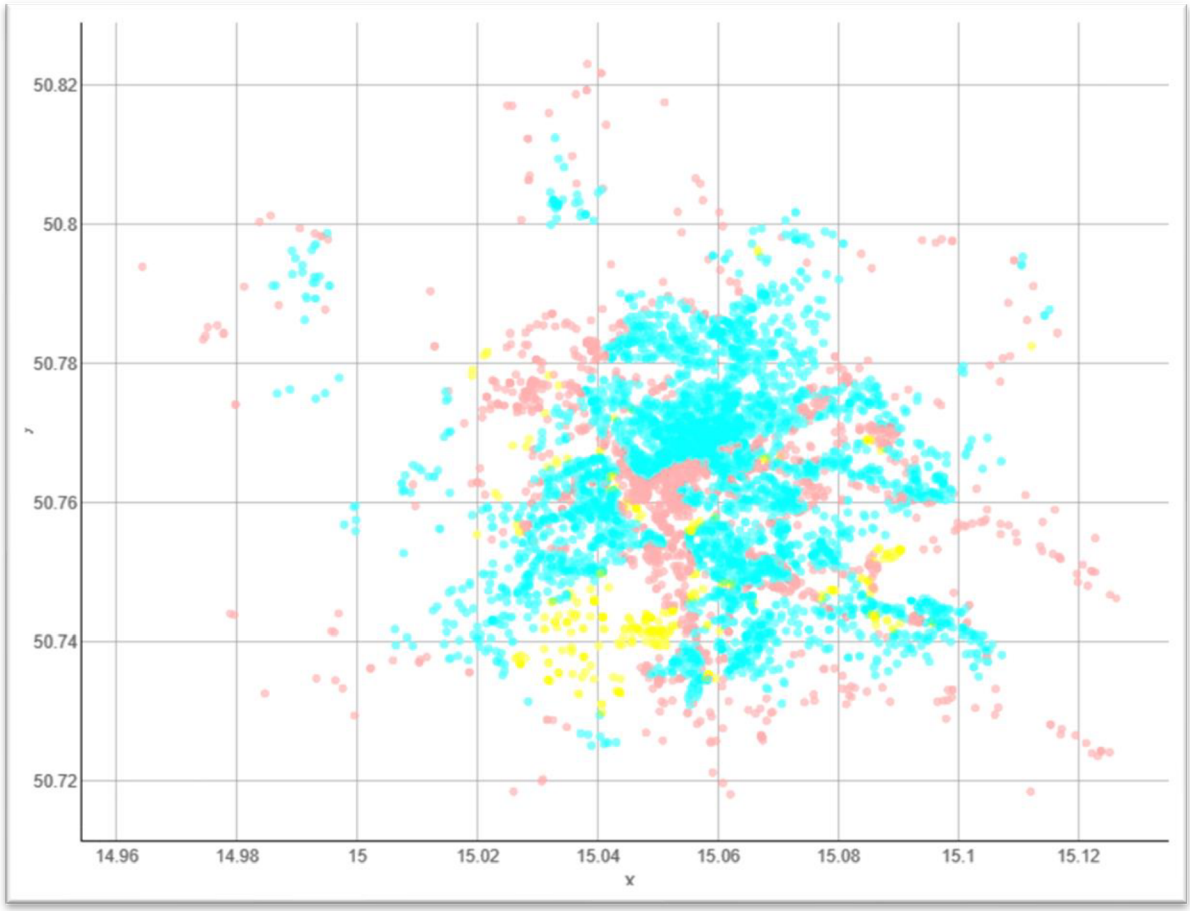


Obr. 22 – Workflow z programu KNIME – využití polygonů

Za tímto účelem byl sestaven workflow (Obr. 22). Workflow má dva vstupní uzly. První uzlem je *File Reader*, který slouží k nahrání datasetu (v tomto případě dataset se záznamy o trestných činech). Tento dataset je zdrojem geografických značek – každá pro jeden trestný čin. Druhým vstupním uzlem je uzel *Shapefile Polygon Reader*. Úkolem tohoto uzlu je načíst a zpracovat soubor ve formátu *shapefile*, který obsahuje geometrické údaje o polygonech daných oblastí. Tento soubor je výstupem z dříve zmíněného programu QGIS. Za uzlem *File Reader* následuje uzel *Latitude/Longitude to Coordinate*. Tento uzel převádí data ze sloupců *x* a *y*, což jsou z pohledu programu pouze čísla, na geografickou polohu s konkrétní zeměpisnou šířkou a délkou. Tento krok je nutný pro správnou práci s geografickými daty. Data po této úpravě spolu se zpracovanými polygony putují do uzlu *Coordinate Row Filter*. S pomocí tohoto uzlu je možné vyfiltrovat, které trestné činy svou polohou spadají do oblasti daného polygonu. Jinými slovy zachová pouze ty záznamy, které byly spáchány na území dané oblasti. Pomocí uzlu CSV Writer je pak vytvořen nový csv soubor, který již obsahuje pouze tyto trestné činy.

Postup je opakován dvakrát – jednou pro polygony rezidenčních oblastí a jednou pro průmyslové zóny. Pomocí MS Excel jsou pak oba soubory spojeny a každému činu jsou

přidány příznaky *residential* (0 nebo 1) a *industrial* (0 nebo 1) podle toho, do které oblasti spadají. Na základě těchto značek je vytvořen další sloupec *zone_type*, který má hodnotu „*prumyslova zona*“ pokud *industrial* = 1, „*rezidencni oblast*“ pokud *residential* = 1. Pokud jsou oba parametry rovny 0, pak má sloupec *zone_type* hodnotu „*jina*“. Tato informace může poskytnout náhled na to, jaká oblast přitahuje jaký typ trestné činnosti.



Obr. 23 – Vyobrazení trestných činů, zabarvených podle oblasti

Na mapě (Obr. 23) jsou zobrazeny trestné činy včetně zabarvení podle příslušnosti do dané zóny. Tyrkysová je opět rezidenční oblast, žlutá je průmyslová zóna a růžová reprezentuje všechny ostatní typy oblastí.

5.2.4 Další možnosti

Vedle výše zmíněných zdrojů informací existuje celá řada dalších, které by mohly poskytnout nové informace pro rozšíření původního datasetu. Jedním z nich je server *data.gov.cz* (Ministerstvo vnitra České republiky 2022). Na tomto je možné najít různé datové sady z mnohých českých úřadů a ministerstev.

Dalším možným zdrojem je webová aplikace na principu GIS systému – *marushka.liberec.cz* (Geovap 2008). Tato aplikace poskytuje náhled do všeobecných datových sad města Liberce, jako jsou například urbanistické studie, rozdělení do městských částí, síť veřejného osvětlení apod. Základní nevýhodou aplikace je ale fakt, že se z ní nedá exportovat žádný druh dat. Z tohoto důvodu není možné data použít pro rozšíření datasetu.

Na technické univerzitě v Liberci je provozován vlastní výzkumný server, který poskytuje otevřené datové sady. Server lze nalézt na adrese opendata.tul.cz (Fakulta přírodovědně-humanitní a pedagogická TUL 2021). Tento server bohužel obsahuje zatím jen 19 datových sad, ze kterých ani jedna neposlouží účelu této práce.

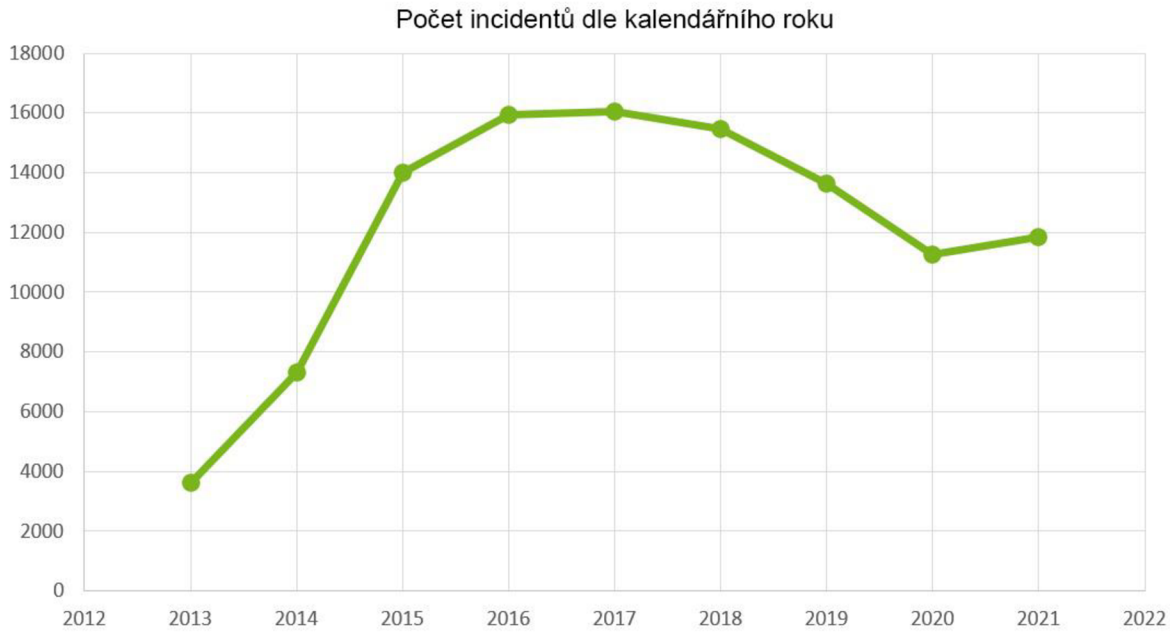
Existuje více dalších možných zdrojů, jako je například server *data.mpsv.cz* (Ministerstvo práce a sociálních věcí 2022) nebo server *data.cssz.cz* (Česká správa sociálního zabezpečení 2022), případně *www.czso.cz* (Český statistický úřad 2022). Bohužel, stejně jako předchozí případy, neposkytují data relevantní vzhledem k případu této práce. Pokud už obsahují data, která by byla relevantní, bohužel jsou lokalizována pouze na úroveň krajů, maximálně okresů. Vzhledem k faktu, že data, kterými se tato práce zabývá, se nachází pouze v oblasti jednoho okresu (Liberec), nejsou taková data přínosná.

6 Případová studie

Tato kapitola se věnuje analýzám rozšířeného datasetu a některým zajímavým případům, které byly odborníky doporučeny k prozkoumání.

6.1 Čas

Vyobrazení na časové ose může poskytnout základní náhled na vývoj trestné činnosti. Z této analýzy je možné zjistit, jaký je trend trestné činnosti na daném území.

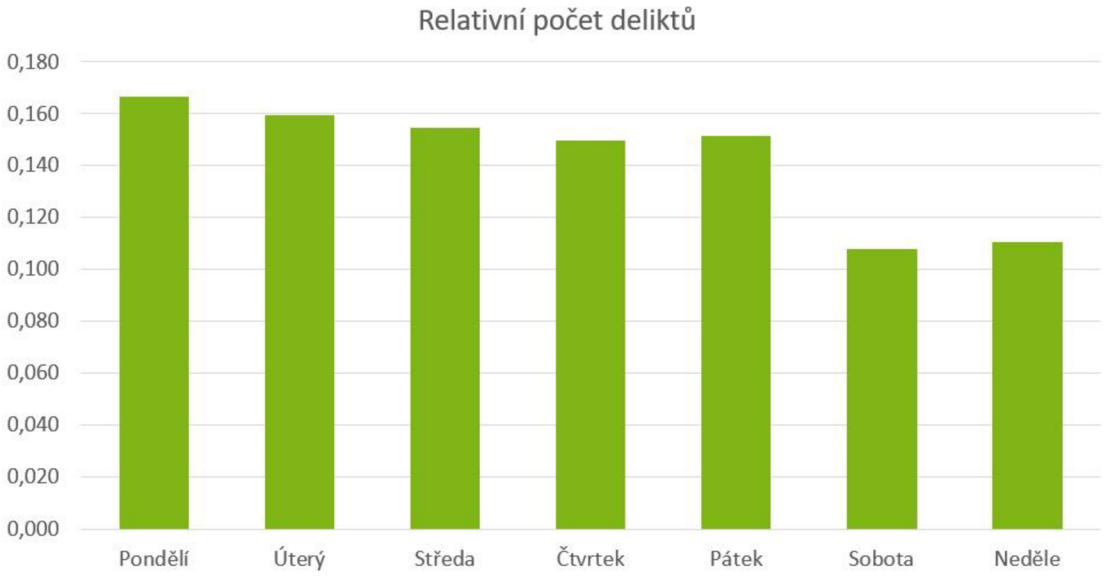


Graf 3 – Počet deliktů v jednotlivých letech 2013-2019

Z grafu (Graf 3) je patrné, že ve zkoumaném intervalu vrcholila kriminalita v roce 2017. Od tohoto roku trend klesá. V letech 2013 a 2014 je počet deliktů výrazně nižší než v ostatních letech. To je způsobeno nižší mírou zadávání záznamů o deliktech do systému. Reálný počet deliktů takto výrazně nižší nebyl, jen o nich nejsou záznamy v systému, ze kterého dataset pochází.

Zajímavou anomálií jsou prudké poklesy kriminality v letech 2020 a 2021. To je způsobeno dvěma příčinami. Obě příčiny jsou v úzkém vztahu s propuknutím pandemie viru COVID - 19. První je, že volný pohyb obyvatel byl značně omezen, což znamená méně aut na vozovkách, méně lidí ve městech atd. Důsledkem tohoto bylo snížení počtu příležitostí ke spáchání trestného činu. Druhou příčinou bylo také vyhlášení nouzového stavu. Pokud je vyhlášen nouzový stav, všechny trestné sazby jsou přísnější. Nutno také nezapomínat na fakt, že mnoho provozoven bylo uzavřeno a v určitých hodinách byl zákaz vycházení. To

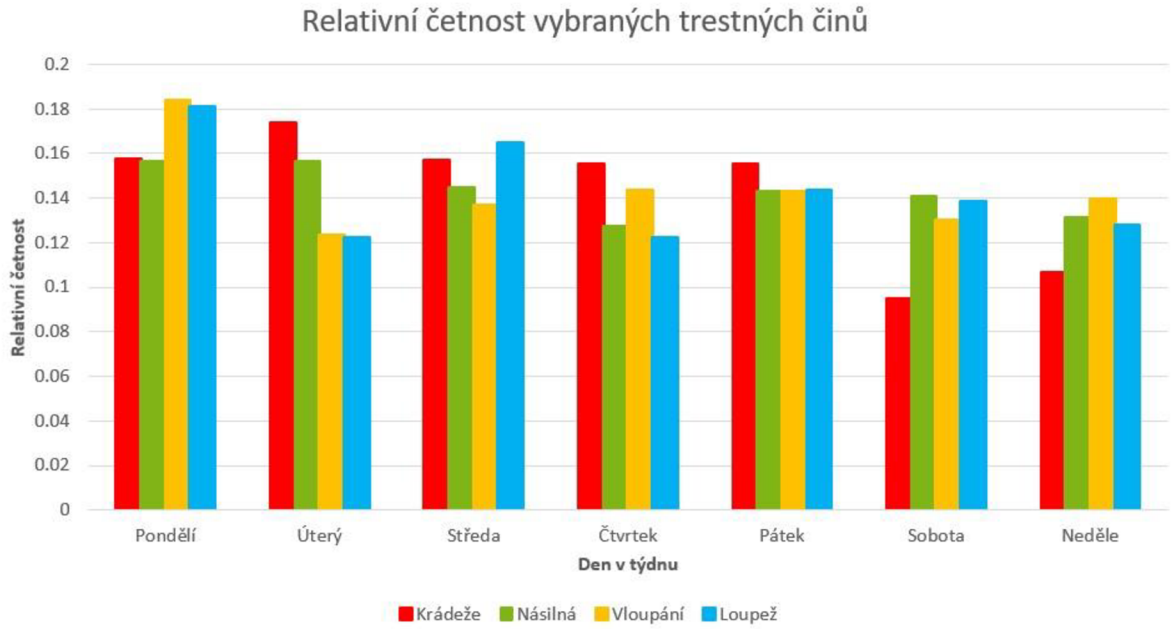
způsobilo do jisté míry zastrašení delikventů a následně vedlo ke snížení míry kriminality. z tohoto důvodu byly z datasetu vyřazeny všechny záznamy, které se staly v letech 2020 a 2021.



Graf 4 – Relativní počet deliktů v závislosti na dni v týdnu

Z pohledu času se také nabízí analýza závislosti počtu deliktů na dni v týdnu. To je zobrazeno na grafu (Graf 4). Z grafu je zřetelné, že nejvyšší počet deliktů se stává v pondělí. Naopak, o víkendu je počet deliktů výrazně nižší.

V návaznosti na předchozí graf je zajímavé prozkoumat vývoj této závislosti u jednotlivých vybraných typů deliktů (Graf 5).



Graf 5 – Graf relativních četností vybraných trestných činů dle dne v týdnu

Na grafu (Graf 5) je možné vidět několik zajímavých anomálií. Krádeže jsou běžné ve všední den, zatímco o víkendu je jich výrazně méně. Násilná trestná činnost je naopak poměrně konstantní během celého týdne. Zajímavou anomálií je také výrazný počet vloupání v pondělí. Oproti dalším dním je počet vloupání v pondělí vyšší o více než 30 %. Loupež má tyto anomálie dvě, a sice v pondělí a ve středu.

Hypotéza

Graf (Graf 4) napovídá, že rozdělení počtu trestných činů v týdnu může být závislé na tom, zda je všední den či víkend. Je vhodné ověřit, zda je tato hypotéza i statisticky významná. K jejímu ověření byl použit program STATGRAPHICS 18.

Nejprve je nutné data rozdělit do výběrů. Jedna proměnná je kategoriální (všední den nebo víkend) a druhá proměnná kardinální (počet trestných činů v daném dni). Z důvodu zpřesnění výpočtů byla data rozdělena podle dní v letech 2016 až 2019. Hodnoty jsou zapsány v následující tabulce (Tabulka 7). Počty trestných činů o víkendu odpovídají počtům trestných činů spáchaných v sobotu nebo v neděli. Počty trestných činů ve všedních dnech odpovídají počtům trestných činů spáchaných ve dnech pondělí až pátek.

Tabulka 7 – Rozdělení počtu trestných činů víkend/všední den

Víkend	350	290	287	238	302	297	306	290												
Všední den	456	449	435	421	437	409	451	369	439	362	442	370	427	394	351	393	468	399	367	338

Nulová hypotéza zní, jestli je průměrný počet trestných činů spáchaných o víkendu stejný jako průměrný počet trestných činů spáchaných v pracovních dnech. Jelikož se jedná o dva nezávislé výběry a je testována významnost rozdílu dvou průměrů, hypotéza je ověřena pomocí dvouvýběrového t-testu.

Nejprve je otestována shoda rozptylů použitých výběrů. Shoda je ověřena pomocí F-testu.

Tabulka 8 – Směrodatná odchylka a rozptyl výběrů

	Víkend	Všední den
Průměr	295,0	408,85
Výběrová směrodatná odchylka	30,6175	39,1021
Výběrový rozptyl	937,429	1528,98

V tabulce (Tabulka 8) jsou vypsány charakteristiky obou výběrů. Poměr mezi výběrovými rozptyly je 937,429 / 1528,98 = 0,613109. Tato hodnota je také hodnotou testové statistiky F. Interval spolehlivosti pro poměr rozptylů je [0,20962; 2,74851] a *P-value* = 0,52314. Pakliže *P-value* ≥ 0,05 (hladina významnosti), nezamítáme nulovou hypotézu o shodě rozptylů. To zde platí, předpoklad t-testu je splněn.

Dalším předpokladem t-testu je, že oba výběry pocházejí z normálního rozdělení. K ověření, zda jednotlivé výběry mohou pocházet z normálního rozdělení, je použit Shapiro-Wilkův test, který má testovou statistiku *W*.

Hypotéza, že data pocházejí z normálního rozdělení, je zamítnuta, pokud hodnota testové statistiky *W* je menší, než tabelovaná hodnota pro danou hladinu významnosti (0,05). V následující tabulce (Tabulka 9) jsou zobrazeny výsledky testu pro oba výběry.

Tabulka 9 – Výsledky Shapiro-Wilkova testu normality

	W Statistika	P-Value
Víkend	0,886561	0,217324
Všední den	0,941989	0,26139

Vzhledem k tomu, že P-value je u obou výběrů větší, než hladina významnosti (0,05), nulová hypotéza není zamítnuta ani v jednom případě. Rozdělení obou použitých výběrů je možné považovat za normální.

Oba předpoklady t-testu jsou splněny a je tedy možné ho použít k ověření hypotézy. Hypotéza předpokládá, že střední hodnota počtu trestných činů spáchaných o víkendu se neliší od střední hodnoty počtu trestných činů spáchaných v pracovních dnech. To je nulová hypotéza. Jako alternativní hypotéza je zvolena ta, že o víkendu je tento průměr nižší než v pracovní dny.

H0 ... $\mu_1 = \mu_2$

H1 ... $\mu_1 < \mu_2$

Kritický obor je v tomto případě $t \leq t_\alpha(n_1 + n_2 - 2)$ tedy $t \leq t_{0,05}(8 + 20 - 2)$. Testová statistika *t* musí být menší nebo rovna hodnotě –1,706. Vypočtená *t* statistika je rovna –7,35359.

Jelikož $-7{,}35359 \leq -1{,}706$ a vypočtená P-value je $4{,}135 \times 10^{-8} < 0{,}05$, tudíž je na hladině významnosti 0,05 zamítnuta hypotéza o shodě středních hodnot a je přijata alternativní hypotéza. O víkendu je spácháno v průměru méně trestných činů než ve všedních dnech.

6.2 Vztah dat k rozšíření

Tato kapitola se zabývá vztahem původních dat k rozšiřujícím informacím, které byly doplněny v kapitole 5.2.

6.2.1 Zeměpisná poloha a vzdálenosti

S pomocí OpenStreetMaps byly získány informace o určitých místech na mapě, jako jsou zastávky, bankomaty, restaurační zařízení, obchody apod. Vzájemná poloha těchto míst a trestných činů by mohla vysvětlit podstatu některého druhu trestné činnosti. Z polohy jednotlivých záznamů a všech zmíněných míst byly spočteny vzdálenosti každého činu k nejbližšímu zástupci dané kategorie. Základní myšlenka byla taková, že některý typ trestné činnosti bude například blíže k bankomatům (krádeže) nebo k restauracím (násilná trestná činnost). Za tímto účelem slouží v SPSS Modeleru uzel *Means*, který porovná průměry několika vybraných proměnných.

Field	Krádeže	Krádeže vloupáním	Majetková TČ	Násilná TČ	Toxikománie	Importance
vzdalenost_zastavka_m	159.9074783855	157.3531772841	158.2377015367	160.2070857780	151.87359087...	0.4110416987 Unimportant
vzdalenost_shop_m	215.2242697012	217.9300619246	214.5099910171	222.4704197226	202.04913302...	0.2926014864 Unimportant
vzdalenost_rest_m	324.5635588881	339.1879308306	331.2742072498	353.5694353928	333.53084613...	0.7272024921 Unimportant
vzdalenost_posta_m	593.9466723766	584.3797216169	582.7601656042	598.1629265796	591.06243478...	0.2116831647 Unimportant
vzdalenost_hrbitov_m	1052.5286374206	1068.6824682309	1043.2725722712	1055.3568816...	1058.5373223...	0.2783126685 Unimportant
vzdalenost_fuel_m	603.4860666461	588.0612845260	601.8189126393	612.4977615579	577.30509480...	0.6685612955 Unimportant
vzdalenost_bankomat_m	577.5377826812	586.3227037097	578.3806872408	590.2068788351	557.28079103...	0.2158302544 Unimportant

Obr. 24 – Tabulka porovnání vzdáleností dle typu trestné činnosti

V tabulce (Obr. 24) je možné vidět porovnání průměrných vzdáleností k jednotlivým místům. Vybráno bylo pouze pět nejčetnějších skupin trestných činů. Jednotlivé hodnoty jsou uvedeny v metrech [m]. Jak je z tabulky možné vidět, průměrná vzdálenost všech typů trestné činnosti ke všem vybraným místům (bankomat, pošta, …) se vzájemně výrazně neliší. Jinými slovy, žádná z testovaných proměnných nevykazuje tak velké rozdíly v průměrných hodnotách, aby se daná proměnná dala prohlásit za definiční pro druh trestné

činnosti, resp. druh trestné činnosti není možný hodnotou dané proměnné rozlišit. Rozšíření datasetu v tomto směru tedy nebylo příliš přínosné s ohledem na lepší vysvětlení jednotlivých druhů trestné činnosti.

6.2.2 Počasí a hydrometeorologické údaje

Některé studie (Reichhoff 2017) naznačují, že by mohl existovat vztah některých druhů trestné činnosti k počasí, zejména k teplotě. Dataset byl rozšířen o údaje o počasí, které panovaly v den spáchání daného trestného činu. Jedná se o teplotu vzduchu, úhrn srážek, relativní vlhkost vzduchu, dobu slunečního svitu a tlak.

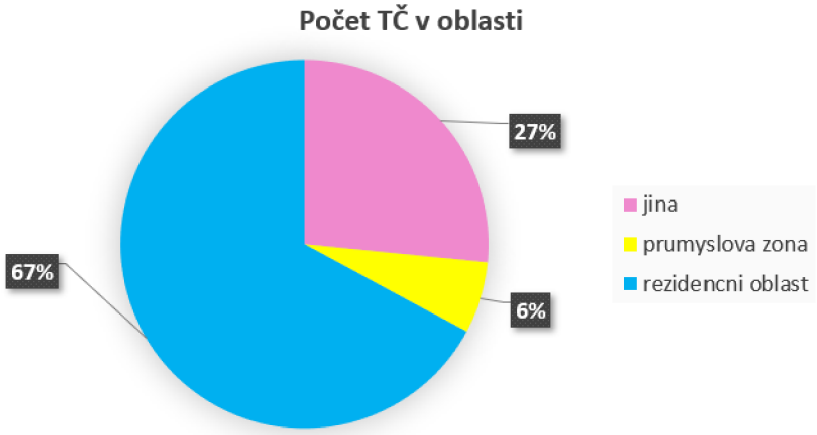
Field	Krádeže	Krádeže vloup.	Majetková TČ	Násilná TČ	Toxikománie	Importance
teplota [°C]	9.0532200603	9.0755820896	8.9750861079	9.1756097561	8.8658064516	0.0749243608 Unimportant
tlak [hPa]	969.0478213209	969.1976716418	969.4475315729	969.1752613240	968.7791935484	0.7857301310 Unimportant
uhrn_srazek [mm]	2.1246094820	1.9511641791	2.0548794489	2.3398954704	2.2243548387	0.6809854817 Unimportant
vlhkost [%]	73.0465881063	72.8507462687	73.1833141982	73.0069686411	72.9064516129	0.0449998175 Unimportant
slunce [h]	4.4904357358	4.4341492537	4.4966322235	4.6321428571	4.5124193548	0.1670961370 Unimportant

Obr. 25 – Tabulka porovnání údajů o počasí dle typu trestné činnost

Ve výstupu z uzlu *Means* (Obr. 25) v programu SPSS Modeler je vidět porovnání průměrných hodnot proměnných, které nesou údaje o počasí. Průměry těchto hodnot jsou rozděleny dle typu trestné činnosti. Sloupec *Importance* ukazuje hodnotu vypočtené statistiky. Podobně jako v předchozím případě se vzdálenostmi byly všechny proměnné označeny jako nedůležité. Rozdíly v průměrech všech proměnných nejsou dostatečně velké, aby popisovaly jednotlivé druhy trestné činnosti a vzájemně je od sebe odlišovaly. Z tohoto důvodu nebylo toto rozšíření datasetu přínosné.

6.2.3 Oblasti a zóny

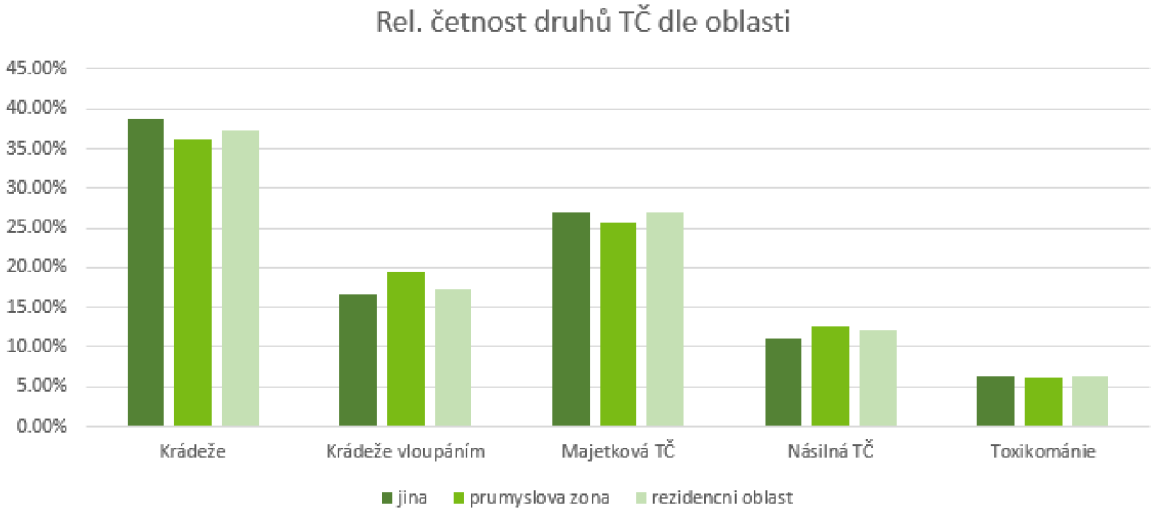
Jednotlivé záznamy byly obohaceny o informaci, zda byly spáchány v průmyslové, rezidenční nebo jiné oblasti. To může například podat informaci, jestli každá z těchto zón přitahuje více nějaký konkrétní typ trestné činnosti. Nejprve je vhodné zobrazit rozložení v daných zónách. Rozložení je zobrazeno v následujícím grafu (Graf 6).



Graf 6 – Rozložení trestných činů dle oblasti

Z grafu (Graf 6) je patrné, že většina (67 %) trestných činů byla spáchána v rezidenční oblasti, zatímco mizivé minimum (6 %) bylo spácháno v průmyslových oblastech.

Základní myšlenka tohoto postupu je odhalit, zda nějaký typ oblasti přitahuje nějaký druh trestné činnosti více než jiné oblasti. K tomu může posloužit zobrazení relativních četností jednotlivých typů trestné činnosti podle typu oblasti. Toto rozložení je zobrazeno v následujícím grafu (Graf 7).



Graf 7 – Rozdělení relativních četností dle typu oblasti

Z grafu je možné vidět, že v rámci jednotlivých oblastí jsou všechny druhy trestné činnosti spíše rovnoměrně rozložené. Jediným rozdílem, který i tak není příliš signifikantní, je počet krádeží vloupáním. Z grafu vyplývá, že relativní počet tohoto druhu trestné činnosti je lehce zvýšený v oblastech průmyslových zón. Žádný z rozdílů ale není příliš markantní natolik, aby se závislost dala prokázat. V tomto směru tedy rozšíření datasetu o oblasti nebyl příliš přínosný.

6.3 Konkrétní případy

Tato kapitola se zabývá konkrétními případy, které byly odborníkem doporučeny k prozkoumání.

6.3.1 Hřbitovy a jejich okolí

Většina deliktů je závislá nejen na čase, ale také na jejich zeměpisné poloze. V tomto případě se jedná o průzkum deliktů v okolí hřbitovů v návaznosti na den v roce. Základní teze je, zda se počet deliktů v okolí hřbitovů nějak zásadně mění během roku. Hřbitovy jsou většinou odlehlá, špatně osvětlená a málo hlídaná místa. Taková místa kriminalitu přitahují. V určitých obdobích roku ale přitahují pozornost i větší části obyvatel, jako například v době Dušiček (2.11.) či v období Vánoc (23.-26.12.). V této době mnoho lidí navštěvuje hroby svých blízkých. To by teoreticky mohlo mít za následek zvýšení počtu deliktů v dané oblasti.

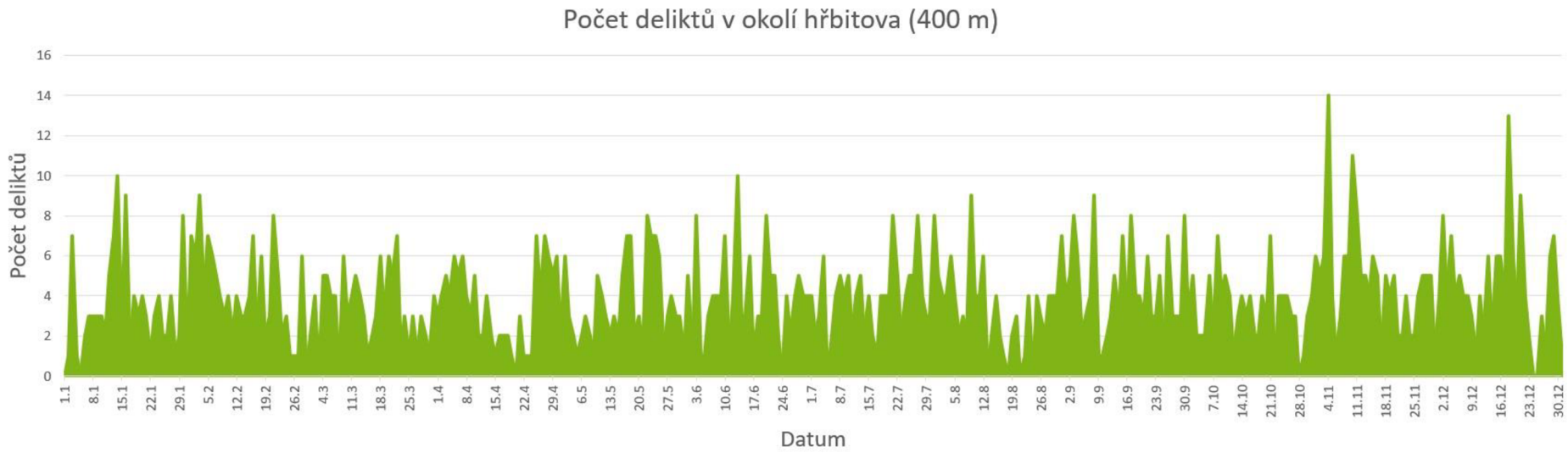
Graf (Graf 8) na následující straně zobrazuje časovou osu dnů během roku a počet deliktů v oblasti 400 m od hřbitova. Tato vzdálenost je zvolena, protože souřadnice hřbitova je vždy umístěna uprostřed hřbitova. Hřbitovy jsou různě velké, vzdálenost 400 m zabezpečí, že se jedná i o okolí hřbitova, ne pouze o jeho vlastní oblast.

Ačkoliv je počet deliktů v roce proměnlivý, ukazuje zřetelné špičky v období 29.10. až 5.11. a pak v období 17.12. až 24.12., tedy období těsně před Vánočními svátky. To podporuje původní tezi. Nejvyšší hodnoty počtu deliktů jsou 4.11. (14 deliktů) a 18.12. (13 deliktů).

Odlehlost hodnot lze ověřit pomocí mediánových souřadnic. Ty se vypočítají následovně (4):

$$x_{madi} = \frac{(x_i - x_{0,5})}{(1{,}483 \times MAD)} \tag{4}$$

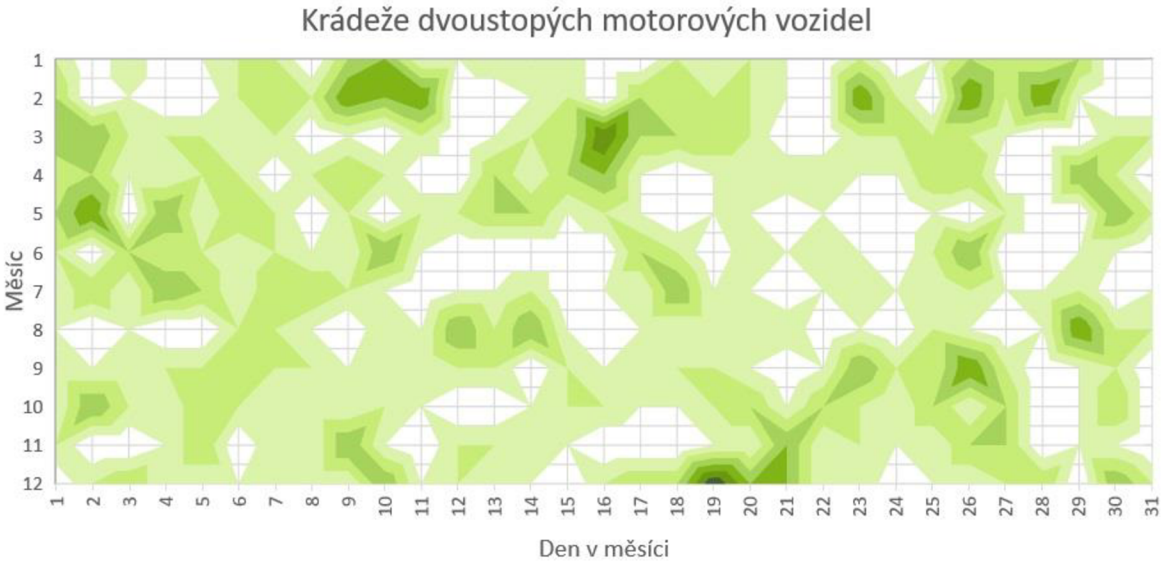
Přičemž *MAD* je medián absolutních hodnot odchylek od mediánu. Pokud je hodnota x_{madi} větší než 3, pak je daná hodnota odlehlá. Pro hodnotu 13 je x_{madi} rovna 3,03439, pro hodnotu 14 je x_{madi} rovna 3,371544. Obě tato čísla jsou větší, než 3. Dané hodnoty lze považovat za odlehlé. Počet trestných činů spáchaných v období pietních svátků je oproti zbytku roku skutečně zvýšený.



Graf 8 – Časová osa počtu deliktů v okolí hřbitovů v závislosti na dni v roce

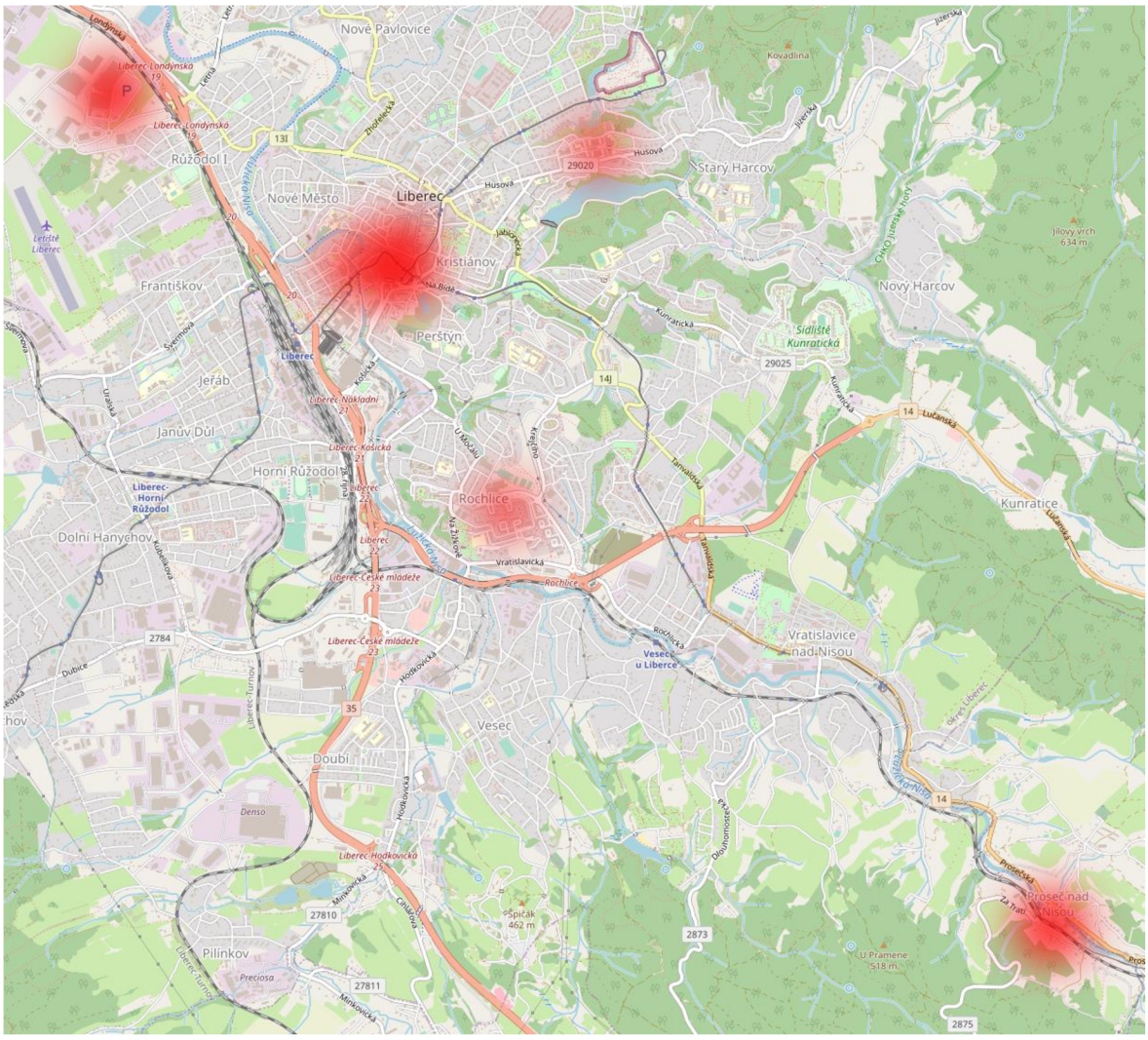
6.3.2 Krádeže vozidel

Jedním z nejběžněji rozšířených druhů krádeží jsou krádeže vozidel. Vozidla mohou být kradena z několika důvodů, ať už je to pro osobní potřebu zloděje, pro nouzové řešení v dané situaci, na další přeprodej, na zakázku či na náhradní díly. V zásadě se rozdělují na dva základní typy – krádeže motorových a nemotorových vozidel. Motorová vozidla se dále dělí na dvoustopá a jednostopá. Dvoustopými motorovými vozidly se myslí osobní automobily, nákladní vozy apod. Jednostopými motorovými se myslí převážně motocykly. Mezi nemotorová vozidla se řadí jízdní kola. Jednotlivé druhy krádeží mohou být závislé na různých veličinách. Jedním takovým faktorem může být například den v měsíci. Náhled na situaci může poskytnout následující graf (Graf 9):



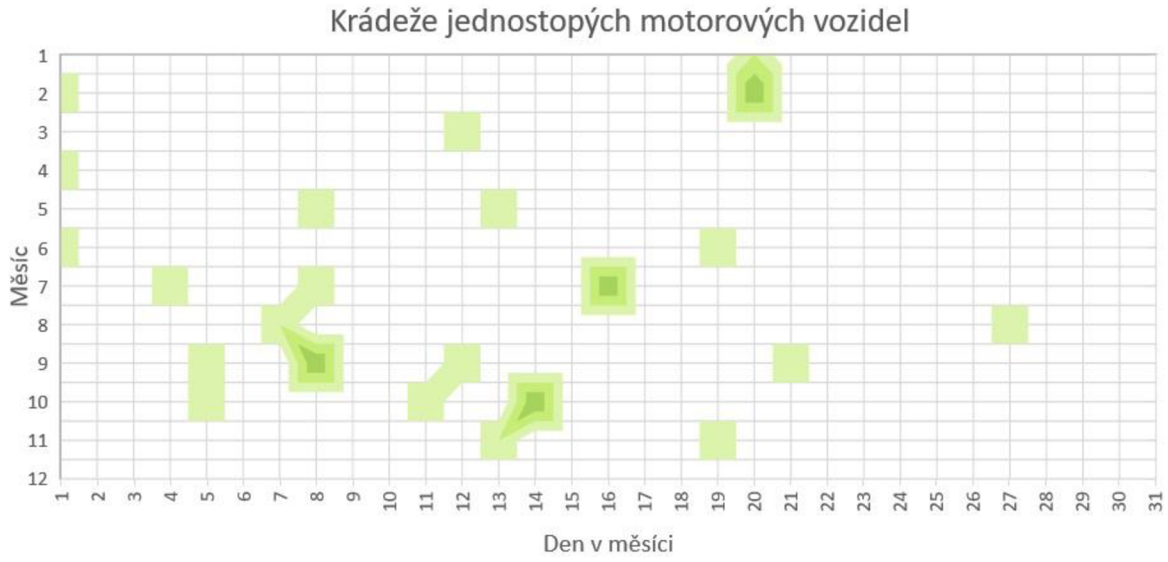
Graf 9 – Mapa hustoty krádeží dvoustopých motorových vozidel

Na grafu (Graf 9) je zobrazeno rozložení krádeží dvoustopých motorových vozidel napříč měsíci a dny v daném měsíci. Čím tmavší je barva, tím vyšší je počet krádeží daného typu. Z grafu vyplývá, že nejméně kradena jsou vozidla v době letních prázdnin (6. a 7. měsíc v roce). To může být způsobeno dvěma faktory. Během prázdnin jsou lidé na dovolených, tím pádem je méně aut v provozu. Druhým faktorem je pak fakt, že je déle světlo. Zajímavou výchylkou je 19. prosinec, tedy těsně před Vánoci. V týdnu před Vánoci je obecně vyšší počet vyjetých automobilů, což zvyšuje šanci ke krádeži. Doplňujícím faktorem je všeobecný shon a stres lidí, což může vést ke zvýšení nepozornosti a nedbalosti například na uzamčení vozidla.



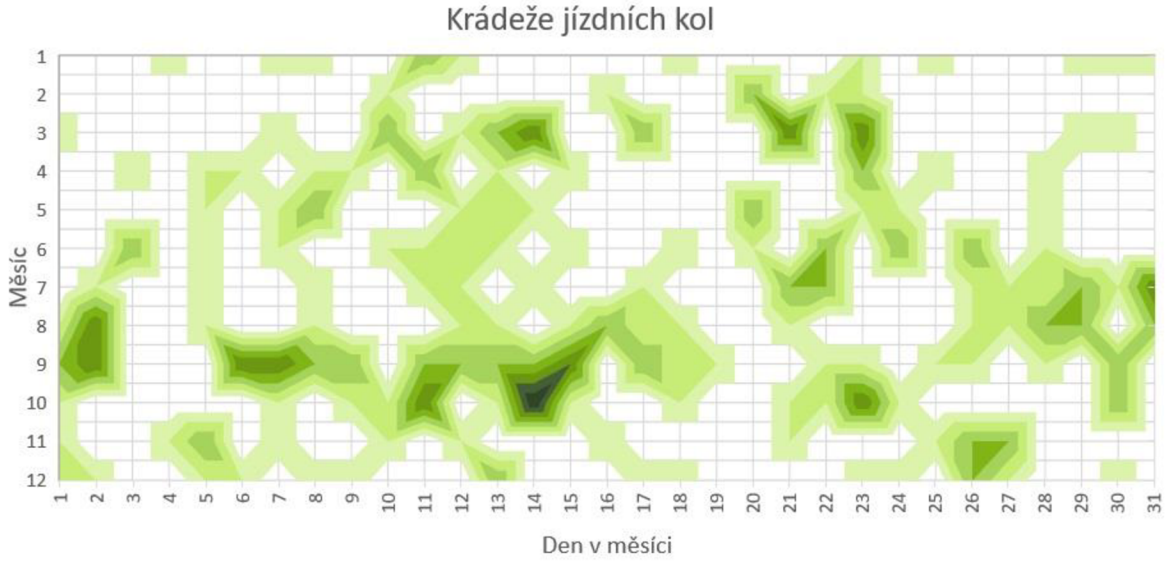
Obr. 26 – Heat mapa krádeží dvoustopých motorových vozidel (období 2013-2019)

Před Vánoci je počet automobilů v provozu zvýšený hlavně z důvodu předvánočních nákupů. Z tohoto důvodu by se větší počet krádeží automobilů mohl vyskytovat v okolí nákupních center. Heat mapa (Obr. 26) ale tuto tezi, vzhledem k rozmístění krádeží, vyvrací. Zvýšená četnost je v oblasti centra města a pak okolí nákupního centra Géčko. Vedle těchto míst se ale krádeže vyskytují i v oblastech, které s nákupy nesouvisejí.



Graf 10 – Mapa hustoty krádeží jednostopých motorových vozidel

Na grafu (Graf 10) je zobrazena mapa krádeží jednostopých motorových vozidel. Z grafu je patrné, že krádeže tohoto typu jsou méně četné. Krádeže těchto vozidel jsou spíše ojedinělé.



Graf 11 – Mapa hustoty krádeží jízdních kol

Grafu (Graf 11) zobrazuje mapu krádeží jízdních kol. Čím tmavší je barva, tím četnější jsou krádeže. Tmavší souvislá oblast je v době od 10. do 24. března. Toto období bývá po zimě první, kdy je více slunečno a tepleji. Počet vyjetých kol je tedy zvýšen a s tím i počet krádeží. Další velmi tmavá oblast je pak na podzim v září a říjnu. Nejtmavší oblast je okolo 14. října. V této době bývají v roce poslední slunné a teplé dny, kdy lidé chtějí před zimou ještě vyjet na kolo. To znamená opět vyšší počet vyjetých kol a s tím spojených krádeží. Graf (Graf 12) pak zobrazuje průměrnou teplotu v měsíci říjnu.

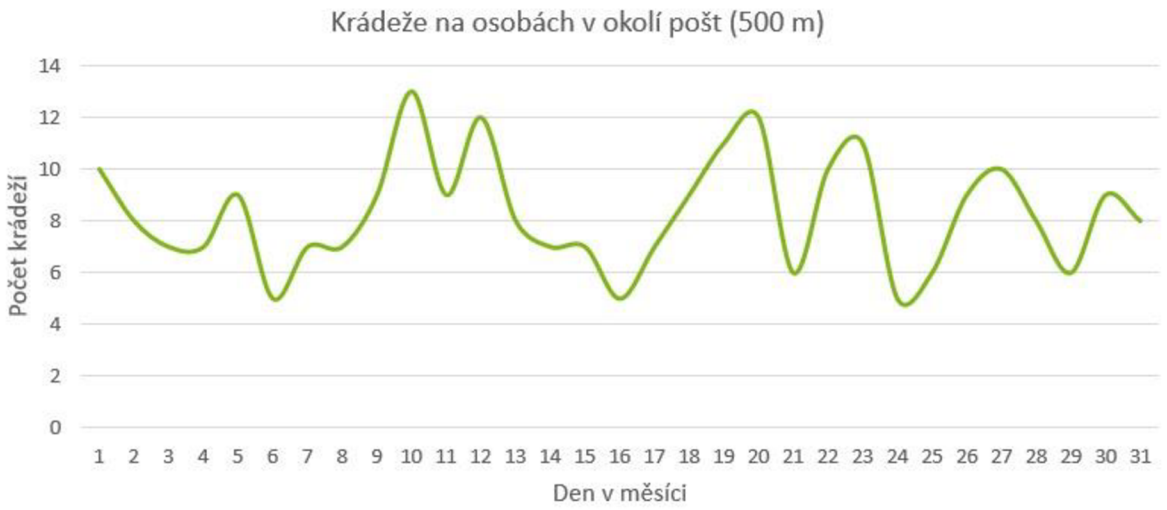
Graf 12 – Průměrná denní teplota v měsíci říjnu

Z grafu (Graf 12) je vidět, že okolo 14. října je skutečně v průměru nejvyšší teplota. Od tohoto data začíná teplota postupně klesat. Jedná se tedy opravdu o jedny z posledních teplých dní v roce.

6.3.3 Krádeže na osobách v okolí pošt

Jednou z nejohroženějších skupin jsou lidé v důchodovém věku. Vzhledem k jejich zvyklostem jich většina není zvyklá vlastnit a používat bankovní účty. Tím, že nevlastní bankovní účet (většina), jsou odsouzeni k výplatě důchodu v hotovosti. Důchod ve formě hotovosti bývá zasílán na poštu, kam si pro něj dotyčná osoba musí dojít. Vedle pošt může být také zaslána do sociálního zařízení dlouhodobého pobytu. Tento případ ale nebude zkoumán. Člověk v důchodovém věku, který obdrží jistý obnos peněz v hotovosti, se může stát potenciální obětí krádeže v době, kdy peníze vyzvedává. Z tohoto důvodu je vhodné prozkoumat výskyt trestných činů krádeží na osobě v blízkém okolí pošt (do 500 m).

Důchod je v České republice vyplácen v sudých dnech měsíce v období od 2. dne v měsíci do 24. dne v měsíci. Většina jich je ale vyplácena v období mezi 10. až 15. dnem měsíce. Pokud výplata připadá na sobotu, je vyplacena v pátek. Pokud připadá na neděli, je vyplacena až v pondělí. Vzhledem k otevírací době pošt a podstatě výplaty důchodu byly tedy dny sobota a neděle vyřazeny.



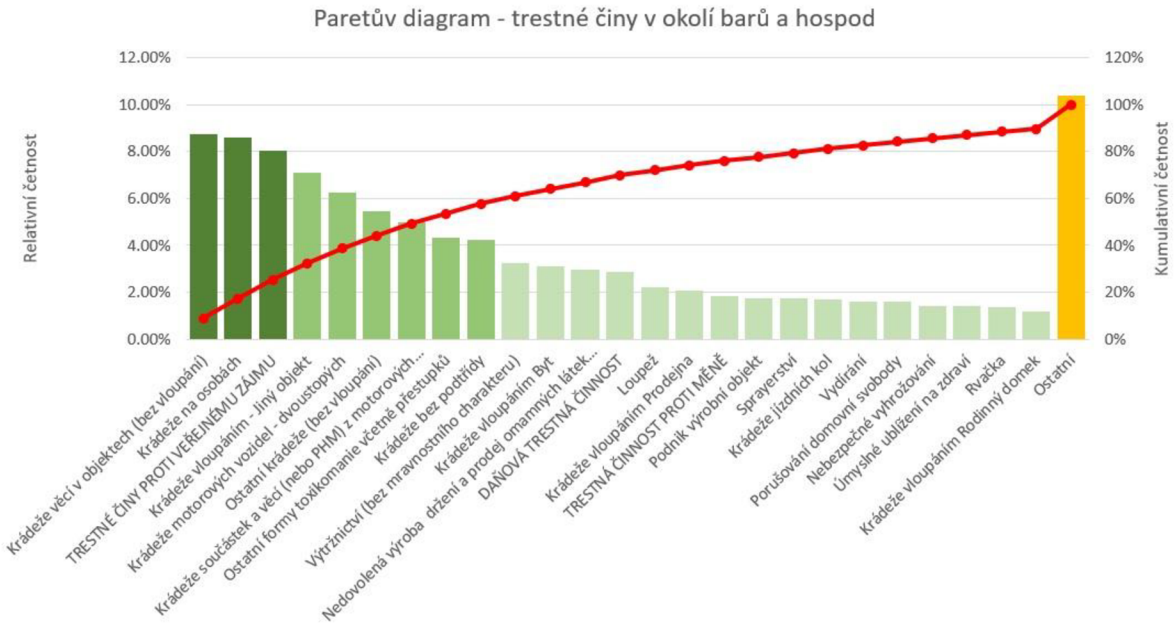
Graf 13 – Graf závislosti počtu krádeží na osobách v okolí pošt (500 m) a dni v měsíci

Z grafu (Graf 13) je vidět, že v období mezi 10. až 13. dnem v měsíci je počet krádeží na osobách zvýšený. Původní myšlenku o zvýšeném počtu krádeží z důvodu vyzvedávání důchodů to ale nepotvrzuje. V druhé polovině měsíce je počet krádeží velmi podobný. Vhodným doplněním dat by bylo zahrnutí podstaty krádeže – jestli byla peněžní nebo materiální. Rozšíření by mohlo pomoci identifikovat, zda je krádež spojena s důchodem či nikoliv. Dalším vhodným doplněním by mohla být také částka, která byla v případě peněžní krádeže odcizena. Pomocí částky by se dalo rozhodnout, zda šlo o peněžní obnos menšího charakteru nebo o nárazovou velkou částku, která by byla odlehlou hodnotou.

6.3.4 Bary a hospody

Bary a hospody jsou bezpochyby zařízení, která svou podstatou trestné činy přitahují. Hlavním důvodem náchylnosti ke spáchání trestných činů je jednoznačně prodej a konzumace alkoholických nápojů. Prodej a konzumace alkoholu mohou být podpořeny také užíváním drog a jiných omamných látek. Tato kombinace obecně snižuje pud sebezáchovy a boří sociální bariéry, takže i jinak stabilní jedinec může inklinovat ke spáchání trestného činu. Restauračních zařízení typu bar je v Liberci celkem 17 a hospod je celkem 35.

Je vhodné prozkoumat, jaký typ trestné činnosti bary a hospody přitahují. Z datasetu byly vyfiltrovány ty případy, které mají jako nejbližší restaurační zařízení právě bar nebo hospodu. Z nich pak byly vybrány ty, které jsou do vzdálenosti 100 m od daného zařízení.

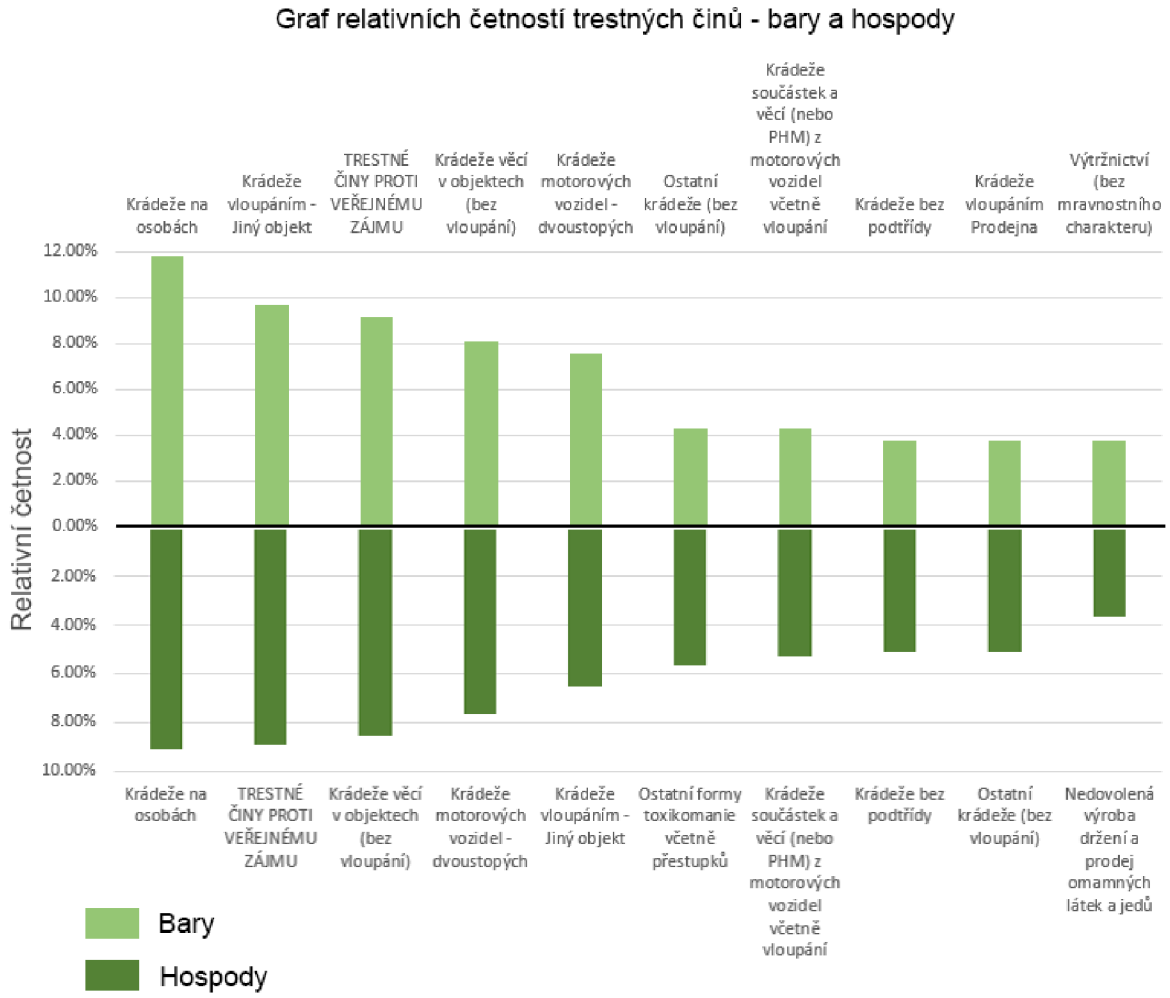


Graf 14 – Paretův diagram trestných činů v okolí hospod a barů

Na Paretově diagramu (Graf 14) jsou vyobrazeny relativní četnosti trestných činů, které byly spáchány do vzdálenosti 100 m od baru či hospody. Kumulativní křivka (červená) pomáhá rozhodnout, které činy jsou z hlediska četnosti nejběžnější. Činy jsou rozděleny do tří skupin. Skupina A (tmavě zelené) jsou činy, jejichž relativní četnost je >7 %. Tyto činy se v okolí barů a hospod vyskytují nejčastěji. Další skupina B (středně zelené) jsou činy, jejichž relativní četnost se pohybuje mezi 3 % a 7 %. Třetí skupina C (světle zelené) mají relativní četnost od 1 % do 3 %. Činy, které mají relativní četnost pod 1 %, jsou zahrnuty jako ostatní, neboť jsou spíše ojedinělé.

Nejdůležitější jsou skupiny A a B. Z diagramu (Graf 14) je patrné, že se jedná převážně o různé druhy krádeží. Mezi krádežemi jsou dvě výjimky – ostatní formy toxikománie včetně přestupků a pak trestné činy proti veřejnému pořádku.

Rozložení jednotlivých trestných činů se ale může lišit podle toho, jestli byly spáchány v okolí baru nebo hospody. Z toho může vyplynout, že jeden typ zařízení přitahuje konkrétní typ deliktů spíše, než druhý a naopak. Situace je znázorněna na následujícím grafu (Graf 15).



Graf 15 – Graf relativních četností TČ dle barů a hospod zvlášť

Z grafu (Graf 15) je patrné, že u obou typů restauračních zařízení dominuje trestný čin krádeže na osobách. V ostatních trestných činech se už poměrně různí. Bary (světle zelená barva, nahoře) přitahují kromě krádeží obecně spíše činnost rušivou. Oproti hospodám se zde objevuje výtržnictví (bez mravnostního charakteru). Naopak hospody kromě běžných trestných činů přitahují také trestnou činnost neobvyklou, která je spojena s užíváním a držením drog. O tom vypovídají trestné činy *ostatní formy toxikománie včetně přestupků a* také *nedovolená výroba, držení a prodej omamných látek a jedů*. To může být následkem několika různých faktorů. Je nutné podotknout, že v centru města jsou hospody a bary poměrně blízko u sebe a jednotlivé oblasti v okolí těchto restauračních zařízení se překrývají, což může způsobit zkreslení výsledků.

V rámci této analýzy by bylo možné označit nejhorší bar a hospodu z hlediska trestné činnosti. To by však bylo v rozporu s ochranou osobních údajů a zároveň mohlo poškodit dobré jméno daného zařízení.

6.4 Hledání ohnisek (hotspotů) kriminality

V rámci této kapitoly je provedena data miningová úloha. Tato úloha se zaměřuje na odhalení ohnisek, tzv. hotspotů, kriminality v oblasti města Liberce a jeho blízkého okolí.

6.4.1 Porozumění problému

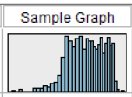
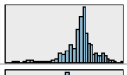
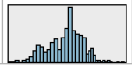
Hledání hotspotů, tedy ohnisek kriminality, je jedna z data miningových úloh, která může policii pomoci k optimalizaci rozmístění hlídek. Cílem je odhalit taková místa, kolem kterých je četnost výskytu trestné činnosti vyšší. To může položit základ k analýze a definici charakteru daných míst. Na základě zjištěných informací je pak možné sestavit optimálnější mapu pro kontrolu kriminality.

Ke hledání hotspotů výskytu trestných činů lze použít několik různých algoritmů. Trestné činy jsou ve svém charakteru geografická data. K analýze geografických dat se používají dva typy algoritmů – algoritmy založené na mřížce a algoritmy založené na hustotě. Právě algoritmům založených na hustotě se věnuje tato kapitola. Jako dva zástupci tohoto druhu byly vybrány algoritmy DBSCAN a jeho vylepšená verze HDBSCAN. Cílem je odhalit vhodnost těchto algoritmů, a hlavně jejich optimální nastavení parametrů, které poskytne nejlepší výsledky při hledání shluků kriminality.

K ověření úspěšnosti či vhodnosti daného nastavení neexistuje vhodná metrika, která by mohla poskytnout nezávislou a robustní informaci. Proto je vhodnost ověřena pouze pomocí srovnání nalezených shluků s mapou. Jako vhodné nastavení bude považováno takové, které dle úsudku autora vytvoří nejvíce smysluplné výsledky vzhledem k pohledu na mapu.

6.4.2 Porozumění datům

Nejprve je nutné získat základní přehled o datasetu. Dataset je převzatý po předchozích úpravách a některé jeho charakteristiky jsou představeny v předchozích kapitolách. Hledání ohnisek je geografická úloha. Stěžejními proměnnými pro tuto úlohu jsou tedy pozice trestných činů ve formě zeměpisných souřadnic *x* a *y*. Základní charakteristiky vstupních proměnných jsou zobrazeny v následujícím obrázku (Obr. 27).

Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
1_id		Continuous	119767	14920332	9284747.12033...	2287281.2067...	-0.2163634726	--	11219
x		Continuous	14.9542400000	15.1263500000	15.0571860790	0.0173220745	0.1226307856	--	11219
y		Continuous	50.7181500000	50.8229800000	50.7623157162	0.0136427883	-0.0577358746	--	11219

Obr. 27 – Charakteristika proměnných x a y

Sloupce *Min* a *Max* ukazují minimální a maximální hodnotu dané proměnné. V tomto případě je oblast vymezena mezi 14,96 ° až 15,132 ° východní délky a 50,708 ° až 50,823 ° severní šířky. Sloupec *Mean* ukazuje průměrnou hodnotu proměnných. Hodnoty v tomto sloupci lze označit za střed oblasti.

Jak již bylo zjištěno v kapitole 6.2, rozšiřující údaje nebyly přínosné, a proto byly přidané proměnné z datasetu odstraněny. Dataset má po této úpravě pouze 8 proměnných. Jejich názvy a vlastnosti jsou popsány v následující tabulce (Tabulka 10).

Tabulka 10 – Proměnné datasetu

Název proměnné	Typ	Popis
1_id	Spojitá	Unikátní identifikátor trestného činu
x	Spojitá	Hodnota souřadnice – východní délka
y	Spojitá	Hodnota souřadnice – severní šířka
date	Spojitá	Datum záznamu spáchání trestného činu
dominant_type	Spojitá	Nejzávažnější typ trestného činu daného záznamu
dominant_type_group	Spojitá	Rodina, do které spadá nejzávažnější typ trestného činu daného záznamu
dtg_name	Nominální	Název rodiny, do které spadá nejzávažnější typ trestného činu daného záznamu
label	Nominální	Slovní popis nejzávažnějšího typu trestného činu daného záznamu

6.4.3 Příprava dat

V datasetu se po úpravách v předchozích kapitolách nachází 11 219 záznamů o deliktech. Data jsou podle typu trestné činnosti rozdělena do skupin. Zastoupení jednotlivých skupin je zobrazeno v následujícím grafu (Graf 16).

Value	Proportion	%	Count
Krádeže		32.53	3649
Majetková TČ		23.29	2613
Krádeže vloupáním		14.93	1675
Násilná TČ		10.23	1148
Podvody		6.89	773
Toxikománie		5.53	620
Dopravní nehody		3.39	380
Obecně nebezpečná		1.52	171
Extremismus		1.32	148
Požáry a výbuchy		0.37	42

Graf 16 – Rozložení druhů trestné činnosti

Trestné činy ze skupiny podvody jsou svou podstatou z hlediska geografické polohy obtížně predikovatelné a zpravidla nejsou vázané na polohu. Proto jsou z datasetu vyřazeny. Dopravní nehody jsou ovlivňovány celou řadou faktorů a jejich výzkum není předmětem této práce, a proto jsou z datasetu také vyřazeny. Obecně nebezpečná trestná činnost, extremismus a požáry a výbuchy jsou v datasetu spíše výjimečné a jejich četnost je velice nízká. O jejich rozmístění se dá předpokládat, že je spíše náhodné. Z tohoto důvodu jsou z datasetu také vyřazeny. Celkový počet záznamů v novém datasetu je 9 705.

Odborníci upozornili, že někdy dojde k pořízení záznamu o deliktu až na místě policejní služebny. Zároveň se může stát, že je daný trestný čin nahlášen na policejní služebně. V obou těchto případech je k záznamu připojena zeměpisná poloha dané služebny. Tento fakt může zkoumání závislosti výskytu trestných činů na zeměpisné poloze zkreslit. Takové záznamy je tedy nutné z datasetu také vyřadit. Vyřazeny jsou tedy záznamy, u nichž je proměnná *vzdalenost_police* menší než 0,000232, což odpovídá přibližně vzdálenosti 20 m. Celkový počet takových záznamů je 470. Po jejich odebrání má nový dataset celkem 9 235 záznamů.

6.4.4 Modelování

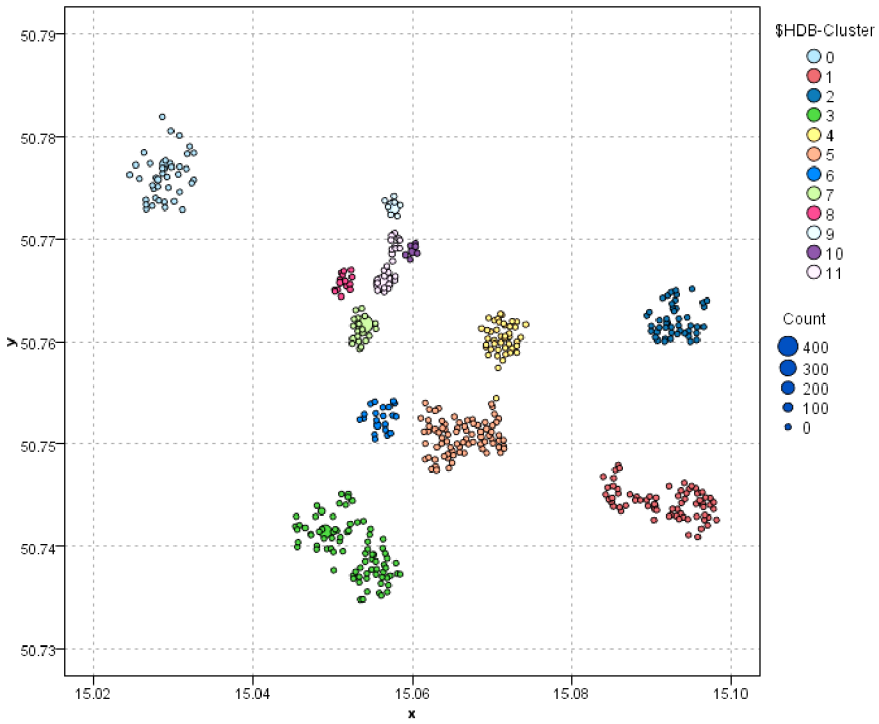
V této kapitole jsou analyzovány oba výše zmíněné algoritmy. Nejprve je zkoumán algoritmus HDBSCAN v prostředí programu SPSS Modeler a poté algoritmus DBSCAN v prostředí programu KNIME. U obou je provedeno několik testů s různým nastavením s cílem odhalit to nejvhodnější pro zvolený dataset.

HDBSCAN

Algoritmus HDBSCAN je, založen na hustotě. V základu má dva parametry, které jsou stěžejní pro tvorbu shluků. Tyto parametry jsou *Min Cluster Size* a *Min Samples*. Oba parametry jsou zkoumány zvlášť. Parametr Min Cluster Size v podstatě označuje minimální počet prvků ve shluku, který je považován za shluk. Jinými slovy, v tomto případě znamená

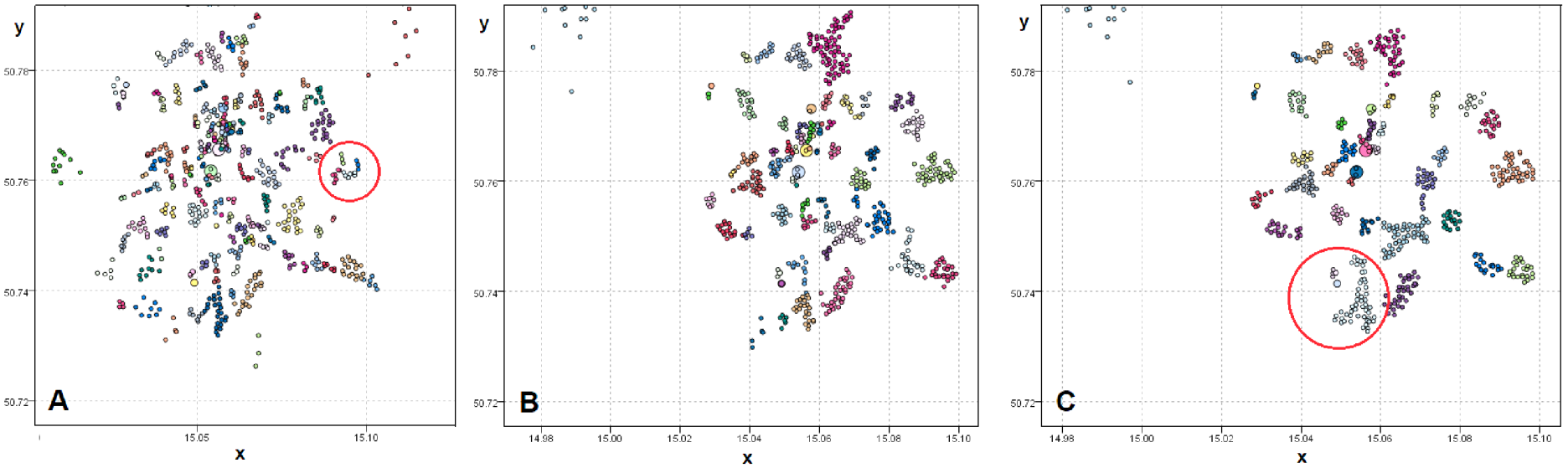
Min Cluster Size minimální počet prvků, od kterého je daný shluk považován za hotspot. Parametr Min Samples je číslo, které označuje konzervativnost algoritmu. Čím vyšší je toto číslo, tím konzervativnější je rozhodnutí, zda jde o shluk nebo ne. To znamená, že čím vyšší je toto číslo, tím více dat je prohlášeno za šum. Se zvětšujícím se číslem Min Samples roste hustota nalezených shluků.

Nejprve je testováno nastavení hodnoty Min Cluster Size, tedy optimální číslo, u kterého lze předpokládat, že nalezne hotspoty kriminality. Parametr Min Samples byl tedy ponechán roven 0 (nejméně konzervativní) a měněn byl parametr Min Cluster Size. Nejprve byly zkoumány hodnoty 100 a 200. Jejich výsledkem ale byly příliš velké shluky (jednotlivé činy byly v rámci shluků velmi daleko od sebe). Některé shluky se takto rozkládají na ploše s poloměrem i větším, než 1 km. Situace je znázorněna v následujícím obrázku (Obr. 28).



Obr. 28 – Přehled clusterů HDBSCAN při Min Cluster Size = 100

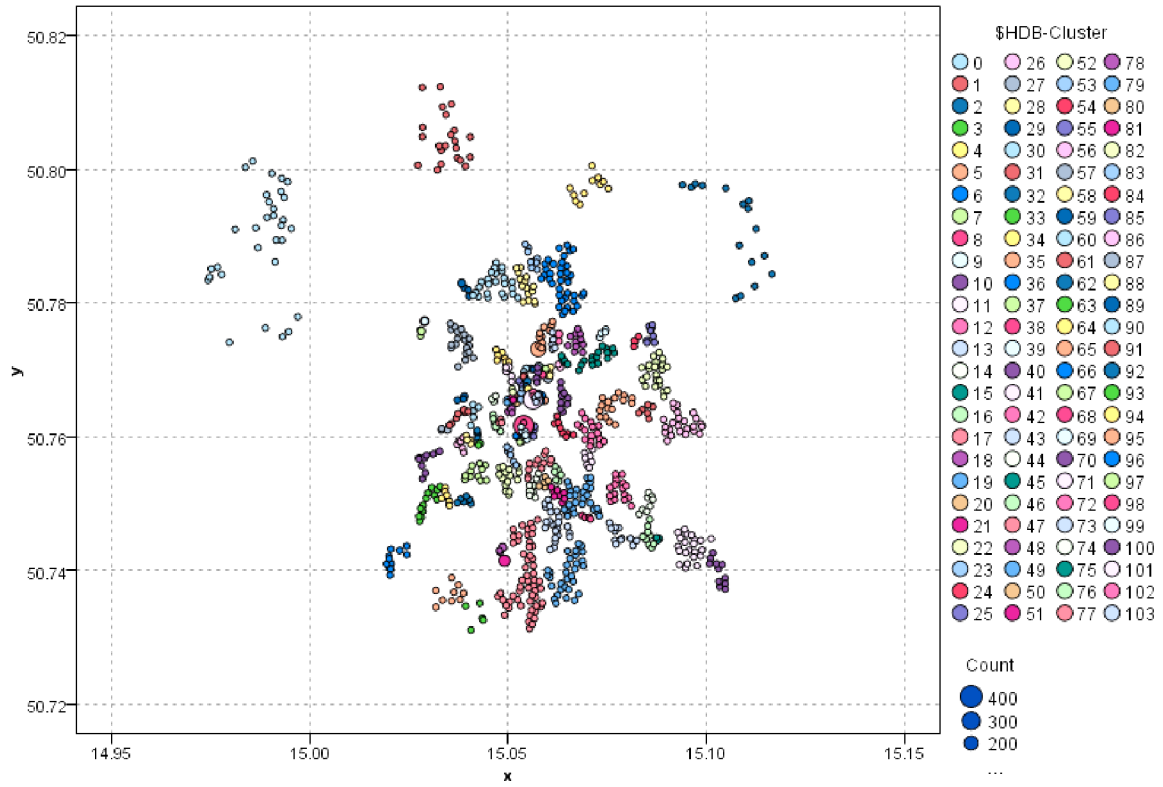
Bylo tedy odhaleno, že 100 není nejlepší hodnota. Dále proto byly postupně vyzkoušeny hodnoty 10, 20 a 30. Výstup z algoritmu pro všechna nastavení je zobrazen v následujícím obrázku (Obr. 29).



Obr. 29 – Přehled clusterů HDBSCAN při Min Cluster Size = 10, 20, 30

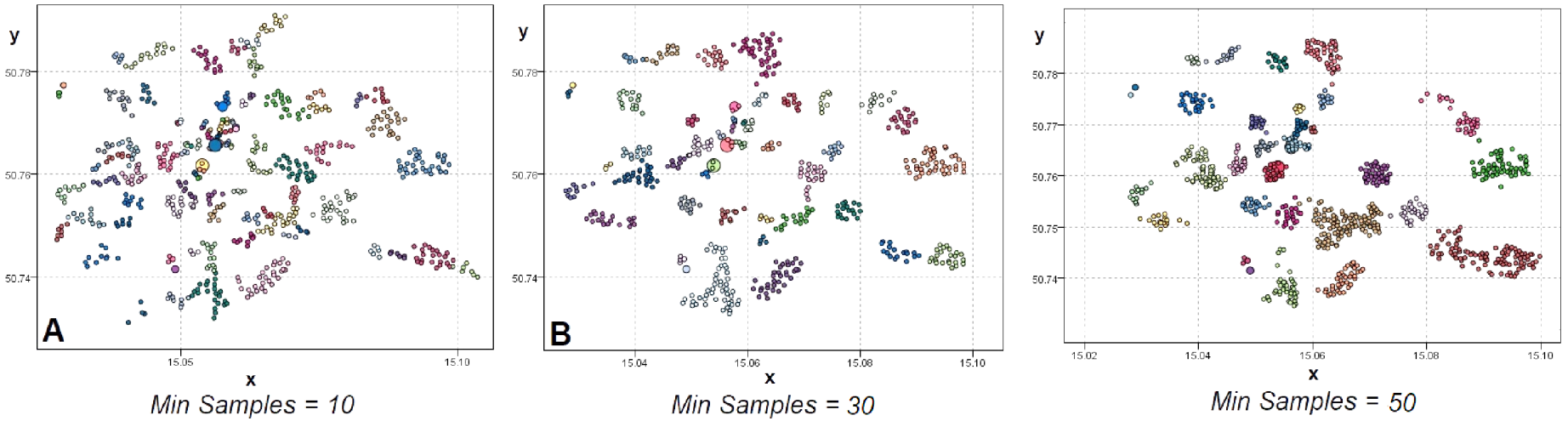
Z jednotlivých testů vyšlo najevo, že Min Cluster Size roven 10 je malá hodnota, protože vytvoří příliš jemné shluky. Důvod je zobrazen v obrázku (Obr. 29), část A v červeném kroužku. Označená část je jedno sídliště. Jako shluky jsou tedy označeny tři různé panelové domy v rámci jednoho sídliště. Opačná situace je znázorněna v části C v červeném kroužku. Hodnota Min Cluster Size rovna 30 naopak vytváří příliš velké shluky. Daná oblast má rozlohu více než 1 km. Z tohoto důvodu se jeví Min Cluster Size = 20 jako nejvhodnější hodnota. Parametr Min Samples tedy bude testován při této hodnotě.

V případě testování parametru Min Samples byly nejprve provedeny testy s velmi malými hodnotami (Min Samples = 1). Ty vedly k příliš jemnému rozdělení shluků. Naopak velké hodnoty (Min Samples = 100) způsobyly sloučení shluků do velkých, které spolu vzájemně úzce sousedí. Výsledek je možné vidět na obrázku (Obr. 30).



Obr. 30 – Přehled clusterů HDBSCAN při Min Cluster Size = 20, Min Samples = 100

Nastavení pak bylo upraveno a otestovány byly postupně hodnoty Min Samples = 10, 30, a 50. Výsledky těchto testů jsou pak zobrazeny na dalším obrázku (Obr. 31).



Obr. 31 – Přehled clusterů HDBSCAN při Min Cluster Size = 20, Min Samples = 10, 30, 50

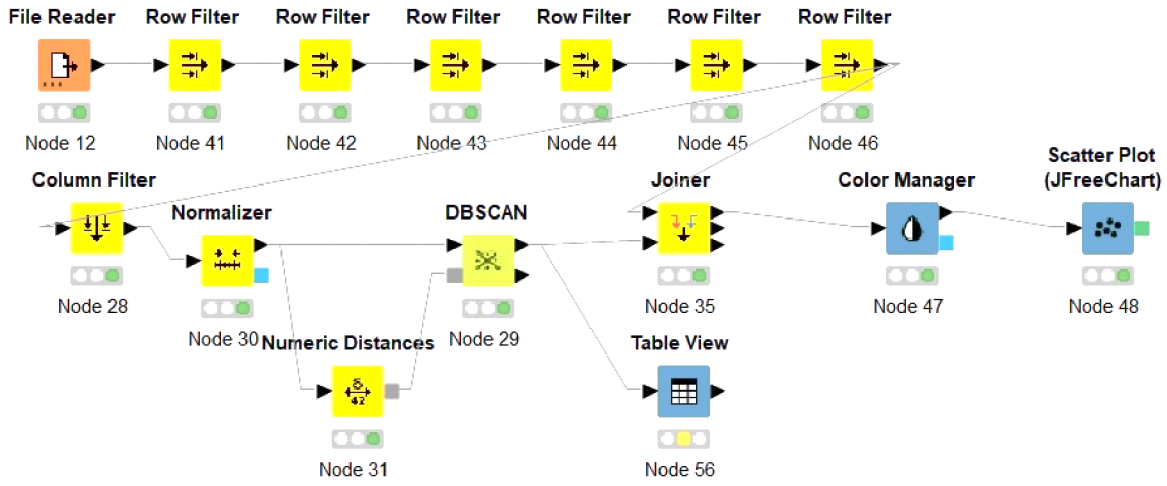
Jak je vidět z obrázku (Obr. 31), hodnota Min Samples = 10 vytváří příliš jemné shluky v rámci jednoho viditelného shluku. Naopak hodnota Min Samples = 50 slučuje některé shluky do příliš velkých shluků, které nepoukazují na konkrétní oblast.

Ze zjištěných testů vyplývá, že optimálním nastavením uzlu HDBSCAN v SPSS Modeler pro tento případ je Min Cluster Size = 20 a Min Samples = 30.

DBSCAN

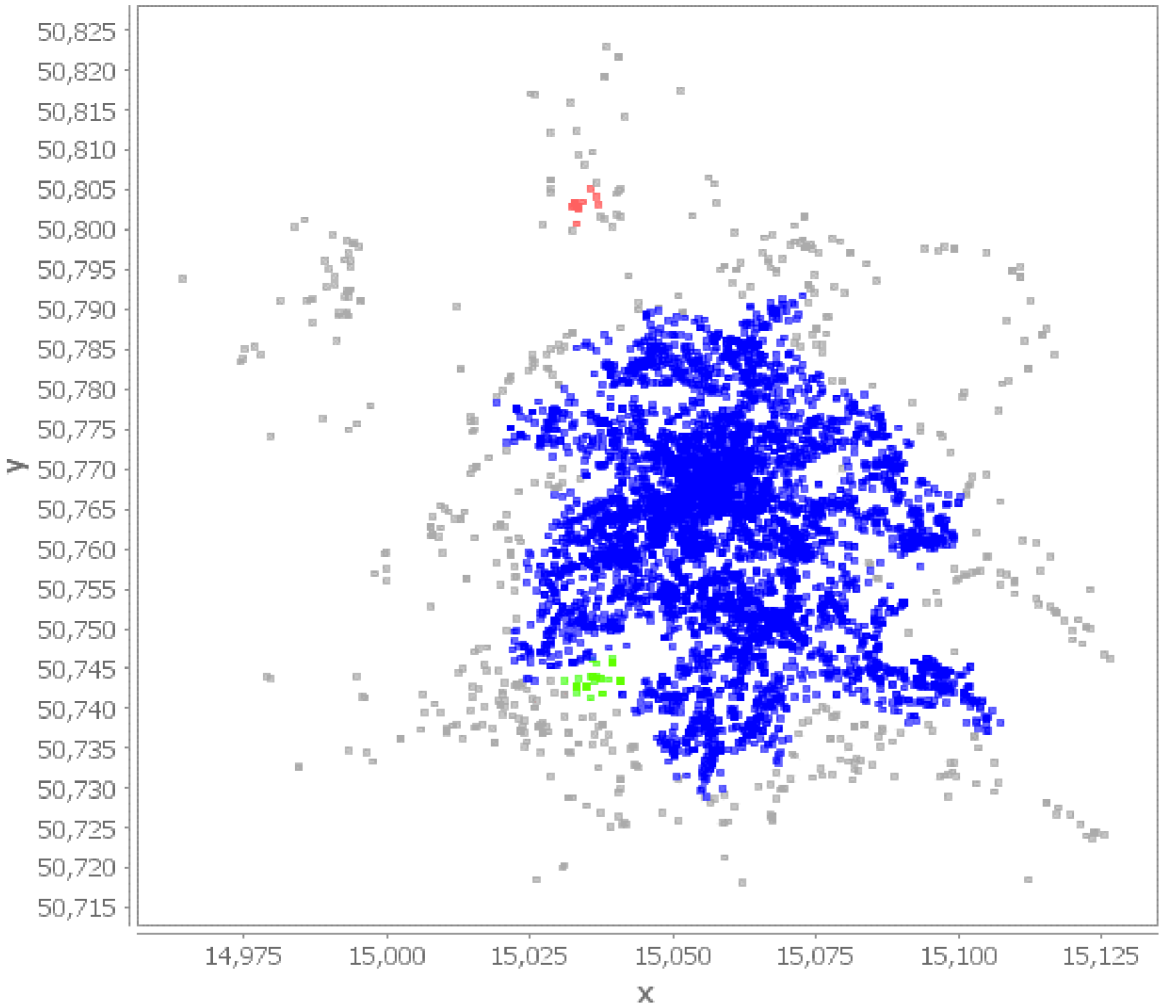
DBSCAN je jedním ze zástupců algoritmů založených na hustotě. Je to jeden z mnoha algoritmů, které jsou svou podstatou vhodné k analýze geografických dat. Oproti HDBSCAN vyžaduje vyšší úroveň znalosti daného problému. DBSCAN má také dva základní parametry – *Minimum Points* a *Epsilon*. Minimum Points je parametr, který určuje minimální počet prvků ve shluku, aby byl vytvořen shluk. Epsilon označuje vypočtenou vzdálenost mezi prvky, které tvoří shluk. Prakticky to znamená, že tzv. core point se stává ten bod, který má do vzdálenosti Epsilon alespoň Minimum Points bodů. Hraniční bod shluku, tzv. border point, je takový, který je do Epsilon vzdálenosti od core point, avšak nemá Minimum Points v Epsilon vzdálenosti od sebe. Všechny ostatní body jsou považovány za šum nebo odlehlé hodnoty. Epsilon je v tomto případě počítán jako Euklidovská vzdálenost (vzdálenost dvou bodů v souřadném systému). Pokud je Epsilon nastavené příliš velké, budou všechny body v datasetu zařazeny do jednoho shluku. Pokud bude naopak Epsilon příliš malé, nebude vytvořen žádný shluk a všechny body budou označeny jako šum.

Na obrázku (Obr. 32) je vidět workflow z programu KNIME s využitím komponenty DBSCAN.



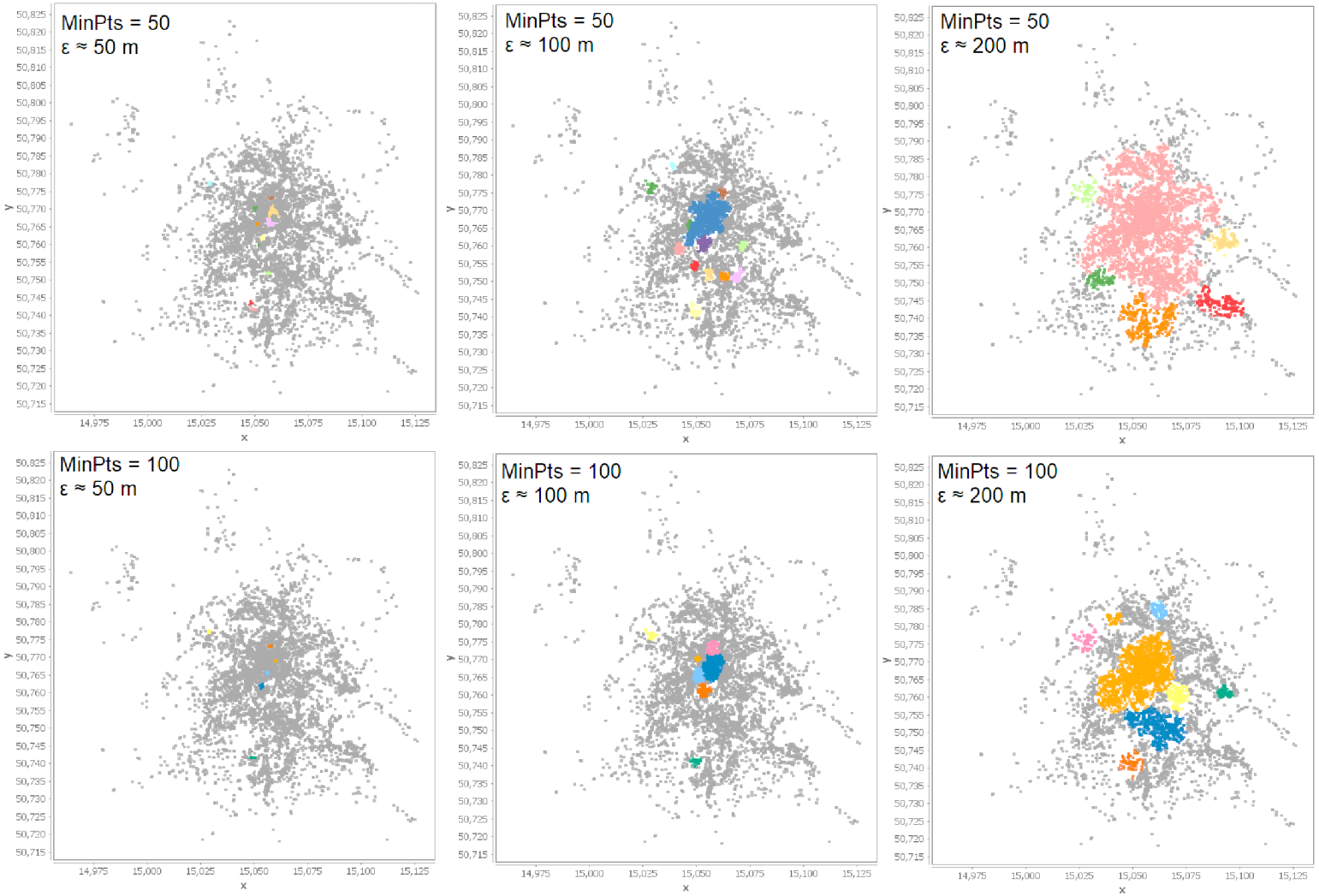
Obr. 32 – Workflow z programu KNIME s využitím DBSCAN

Na začátku workflow jsou provedeny úpravy popsané v části „Příprava dat“. První stěžejní komponentou je Normalizer. Komponenta DBSCAN totiž pracuje s normalizovanými daty (zeměpisná poloha je přemapovaná na rozsah 0.0 až 1.0). Komponenta *Numeric Distances* pak vypočítá Euklidovské vzdálenosti mezi všemi prvky a poskytne je jako matici. Tato data putují do DBSCAN. Kvůli normalizaci je následně nutné přepočítat i vzdálenost. Jako výchozí je počítána vzdálenost 100 m, pro niž odpovídá Epsilon = 0,00794. Jako výchozí nastavení parametrů zde bylo využito zjištění o velikosti shluku z případu s HDBSCAN. Minimum points je tedy nastaven na 20 a je hledáno optimální Epsilon. První test proběhl s parametry Epsilon = 0,007945 a MinPts = 20. Bohužel, pro toto nastavení bylo zjištěno, že není nalezen žádný shluk. V druhém případě bylo Epsilon = 0,01589, což odpovídá vzdálenosti 200 m. Výstup z komponenty je vidět v následujícím obrázku (Obr. 33).



Obr. 33 – Přehled shluků DBSCAN pro MinPts = 20, Vzdálenost 200 m

Z tohoto zjištění vyplývá, že DBSCAN potřebuje trochu jiné nastavení než HDBSCAN. Proto byla vyzkoušena různá nastavení.



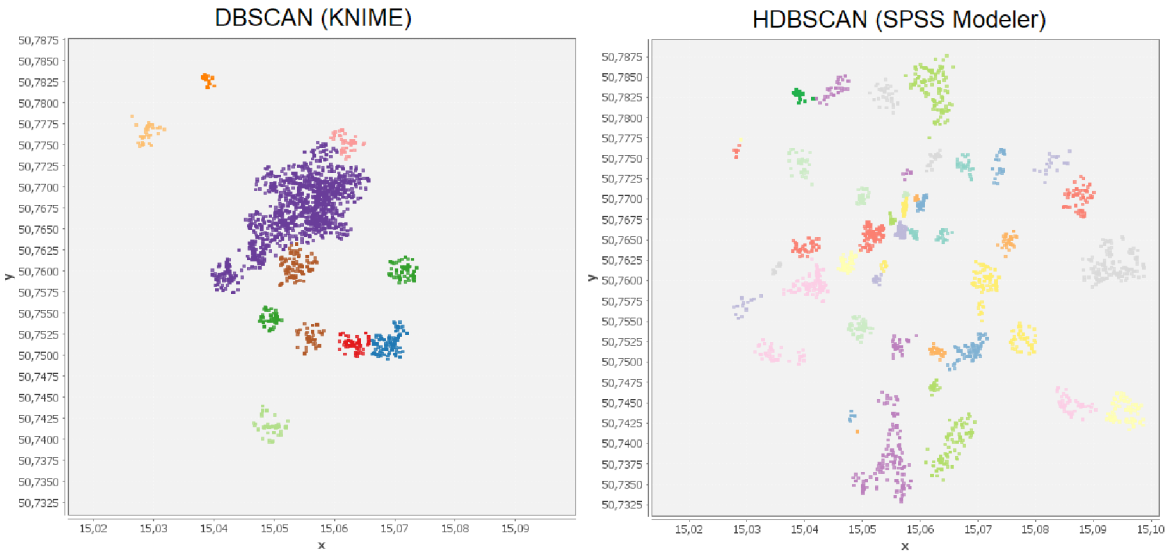
Obr. 34 – Přehled shluků z DBSCAN při různém nastavení

Na obrázku (Obr. 34) je možné vidět výsledky několika různých nastavení algoritmu DBSCAN. Testy proběhly s nastavením o počtu prvků shluku (řádky; 50 a 100 prvků) a pro různé vzdálenosti (sloupce; 50, 100 a 200 m). Z obrázků plyne, že nejslibnější nastavení jsou pro MinPts = 50 a vzdálenost 50 a 100 m. Nastavení 50 m by odpovídalo, pokud by cílem bylo odhalit ohniskovou budovu kriminality. Nastavení 100 m je vhodnější spíše pro identifikaci oblasti se zvýšenou kriminalitou.

6.4.5 Evaluace

Parametry modelů bývají v praxi často voleny odborníky na danou problematiku. Volbě parametrů také přispívá zkušenost nejen s modelováním, ale hlavně s reálnou situací. V tomto případě byly parametry voleny náhodně, pouze s ohledem na charakter vytvořených shluků. Cílem bylo také porovnat dva použité algoritmy.

V následujícím obrázku (Obr. 35) jsou porovnány výstupy dvou nalezených nejslibnějších nastavení obou algoritmů. Některé shluky nalezly oba algoritmy shodně, v drtivé většině se ale liší.



Obr. 35 – Porovnání výstupů algoritmů DBSCAN a HDBSCAN

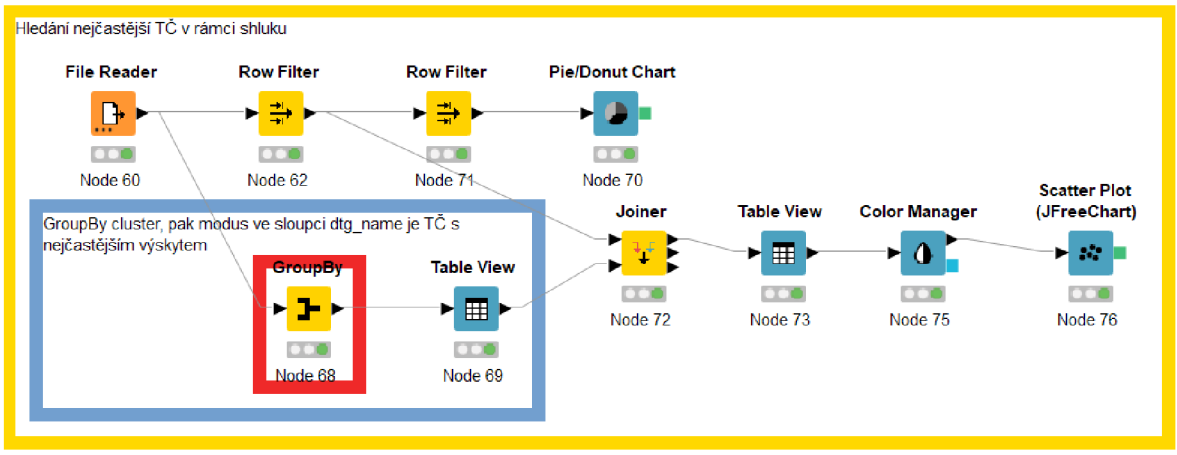
Algoritmus DBSCAN vyžaduje dobrou znalost problematiky, hlavně přehled o vzdálenostech v modelovaném prostoru. DBSCAN má oproti HDBSCAN velikou nevýhodu v tom, že shluky jsou pomocí fixních parametrů hledané napříč celou oblastí. To mj. způsobuje, že se změnou parametrů se některé shluky postupně buď objevují, nebo naopak již objevené mizí (podobně, jako když se zvedá hladina moře – ostrovy se objevují a mizí). To je hlavní výhodou například v problémech, u kterých jsou jasné špičky.

V případě hledání hotspotů kriminality, kdy je hustota proměnlivá a špičky jsou různě vysoké, je vhodnější algoritmus HDBSCAN díky svému charakteru hierarchie. Algoritmus totiž po nalezení zřetelných shluků vyhledává shluky i v oblastech s nižší hustotou, než má hlavní shluk. Vyhledává je do doby, dokud splňují požadavky parametrů. Pakliže je nesplňuje, označuje prvky za šum. Tato vlastnost je velikou výhodou právě při zkoumání dat o kriminalitě, která je páchána ve městě, kde je různá hustota osídlení, různá hustota veřejných komunikací, pohybu lidí aj. Vzhledem k proměnlivé hustotě těchto faktorů je skoro nutné předpokládat, že i výskyt trestné činnosti bude mít proměnlivou hustotu. Tento fakt algoritmus HDBSCAN právě zohledňuje.

6.4.6 Využití výsledků

Výsledky obou algoritmů je možné využít k identifikaci míst se zvýšeným počtem spáchaných trestných činů, a tak odhalit ohniska kriminality. V takovém případě je možné podniknout kroky k jejímu snížení. Vedle tohoto případu je ale vhodné nalezené shluky podrobit hlubší analýze, která by mohla vést k lepšímu pochopení trestné činnosti.

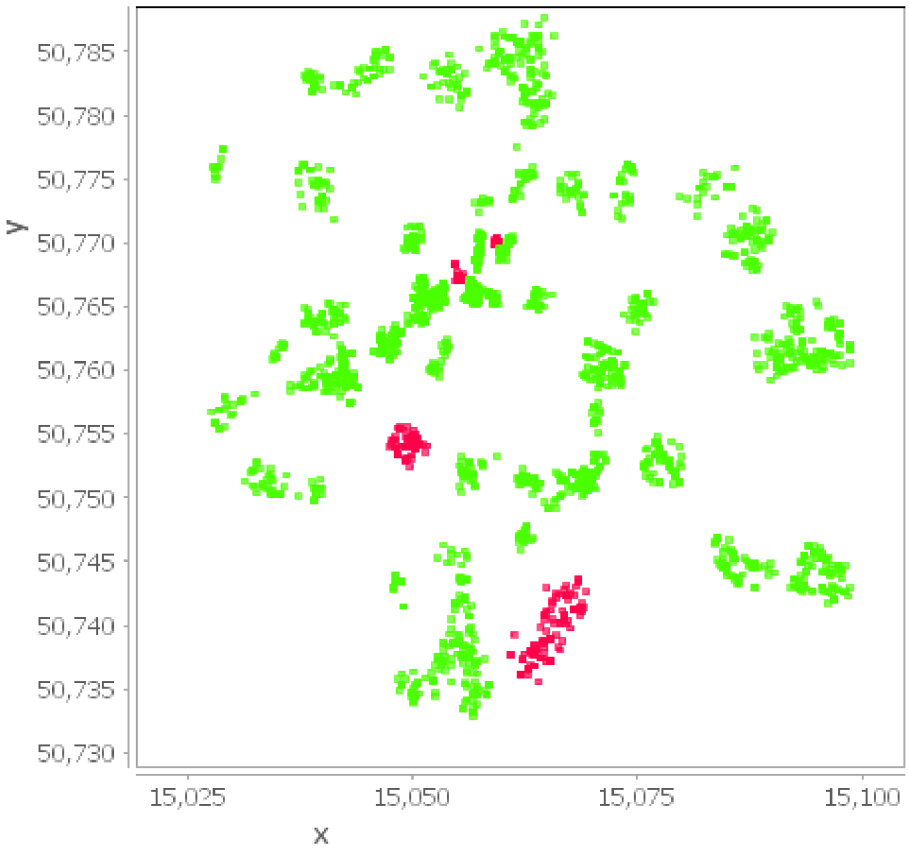
Jedním ze způsobů, jak provést hlubší analýzu vytvořených shluků, je zjistit, čím je daný shluk typický, resp. výjimečný oproti ostatním. Takovou charakteristikou může být například odhalení nejčastěji se vyskytující trestné činnosti pro každý shluk. Toto hledání bylo uskutečněno v prostředí programu KNIME. Workflow je zobrazen v následujícím obrázku (Obr. 36).



Obr. 36 – Workflow s určením nejčastější TČ v rámci shluků

Na začátku workflow je nutné nahrát soubor s přiřazenými shluky z předchozí kapitoly. Stěžejní je spodní větev ve workflow. Po nahrání komponentou *File Reader* putují data do komponenty *GroupBy* (na Obr. 36 červeně označena). Tato komponenta slouží k seskupení

dat dle zvoleného sloupce a k následné agregaci seskupených dat. Nejprve jsou data seskupena podle sloupce *$HDBCluster*. Komponenta seskupí data do řádků podle hodnot v daném sloupci. V tomto případě vzniklo 50 shluků (viz. Obr. 35). Seskupením tedy vzniklo 50 řádků. Po seskupení přijde agregace. Komponenta nabízí několik způsobů agregace. Nejvhodnějším způsobem je agregace podle jiného sloupce, v tomto případě podle sloupce *dtg_name*. Tento sloupec může obsahovat pouze pět hodnot (dle typu TČ) – krádeže, krádeže vloupáním, majetková TČ, násilná TČ a toxikománie. Data byla seskupena podle shluků a agregována podle sloupce *dtg_name*. V rámci agregace je možné použít několik způsobů v závislosti na hodnotách ve sloupce, dle kterého se agreguje. V tomto případě nejčastější trestnou činností v každém shluku je právě modus sloupce *dtg_name* pro každý shluk. Jinými slovy, pro každý řádek nové matice (shluk) je nalezen modus hodnot ve sloupci *dtg_name*. Komponentou *Joiner* je pak spojena nově vzniklá matice s původní maticí, kde je ke každému záznamu o trestné činnosti přidána hodnota zjištěného modu. Průnikem je zde příslušnost záznamu ke clusteru z jedné tabulky a název clusteru z nové tabulky. Záznamy jsou pak zobrazeny v obrázku (Obr. 37). Shluky jsou zabarvené podle modálního druhu trestné činnosti. Jako modální byly odhaleny pouze dva – krádeže (zelené) a majetková trestná činnost (červená).



Obr. 37 – Shluky se zabarvením podle nejčastější trestné činnosti

7 Doporučení

V rámci případové studie jsou ověřeny či vyvráceny některé možné závislosti, které v datech mohou existovat. Nejprve jsou data analyzována z hlediska času. Počet trestných činů je výrazně nižší v letech 2013 a 2014. To je způsobeno menším počtem záznamů v systému.

V roce 2020 propukla pandemie viru COVID-19, což způsobilo značný pokles kriminality a z tohoto důvodu byly všechny záznamy od data 1. ledna 2020 z analýz vyřazeny. Dále pak bylo prozkoumáno rozložení trestných činů v týdnu. Bylo zjištěno, že počet trestných činů spáchaných o víkendu je menší, než počet činů spáchaných v pracovní dny. Pomocí statistické hypotézy je možné ověřit, zda je tento rozdíl statisticky významný. Na hladině významnosti 5 % byla zamítnuta hypotéza o shodě průměrů a je přijata alternativní hypotéza. O víkendu je spácháno průměrně méně trestných činů než ve všední dny.

Další část práce se věnuje ověřování či vyvracení některých možných případů. První takový případ je teze, jestli je v okolí hřbitovů zvýšená kriminalita v době pro hřbitov významných dnů v roce. V grafu (Graf 8) jsou zřetelné špičky v počtu trestných činů ve vzdálenosti do 400 m od hřbitova v období na začátku listopadu a ke konci prosince. To podporuje fakt, že je kriminalita v okolí hřbitovů je skutečně zvýšena v období, kdy je hřbitov navštěvován frekventovaněji. Z tohoto důvodu je vhodné oblasti v blízkém okolí hřbitovů kontrolovat frekventovaněji.

Dalším případem je zjištění možné závislosti mezi datem a krádežemi vozidel. Bylo zjištěno, že motorová vozidla dvoustopá jsou nejvíce kradena v týdnu před Vánoci, což může být způsobeno zvýšenou řidičskou aktivitou v době předvánočních nákupů. Krádeže jednostopých motorových vozidel jsou spíše ojedinělé a žádné konkrétní vztahy mezi jejich krádežemi a datem nebyly nalezeny. U krádeží jízdních kol bylo zjištěno, že počet krádeží je zvýšen v období března a pak v období září a října. Březen je doba, kdy začínají být první hezké dny, a tak lidé začínají po zimě více jezdit na kole. Se zvýšeným počtem vyjetých jízdních kol se zvedá i počet jejich krádeží. Naopak na podzim jsou v říjnu poslední dny, kdy je před zimou hezké počasí. Platí stejný případ, jako v březnu. Zvýšený počet kol má za následek zvýšený počet jejich krádeží. Na základě tohoto zjištění by bylo vhodné zvýšit počet hlídek. Toto doporučení je ale vhodné k diskusi. Pro policistu je téměř nemožné zjistit, zda krádež probíhá či proběhla, proto přímý vliv na kriminalitu není možný. Účel hlídek v této době je spíše odstrašující. Tím se myslí nepřímo potenciálního pachatele odradit.

Další myšlenkou, která byla prozkoumána, jsou krádeže na osobách v okolí pošt v době důchodů. Lidé v důchodovém věku jsou pro pachatele snadnější cíle. Mnoho starších lidí nepoužívá bankovní účty, a proto si pro své důchody musí chodit na poštu. Důchody nejvíce chodívají v době mezi 10. a 15. dnem měsíce. Byla prozkoumána oblast ve vzdálenosti do 500 m od všech pošt. Pokud je teze správná, měl by v této době být počet krádeží na osobě zvýšený. Graf (Graf 13) skutečně naznačuje zvýšený počet krádeží v tomto období. Teze ale není jednoznačně potvrzena, neboť v druhé polovině měsíce se objevují velké výkyvy. Doporučení, které vzešlo z této analýzy, je přidání informace, jestli daná krádež byla materiální nebo peněžní. Peněžní krádeže budou s krádeží důchodů souviset, materiální nikoliv. To by mohlo vést k izolaci krádeží důchodů. Dále by bylo vhodné získat skutečná data ohledně výběru důchodů na poštách. Mohlo by tak být například odhaleno, že hlavní dobou pro výběr důchodů není období mezi 10. a 15. dnem měsíce. Vedle informace o typu krádeže (peněžní nebo materiální), by vhodným doplněním byl i věk oběti. To by mohlo vést k lepší izolaci deliktů, v nichž jsou obětmi např. senioři. Obecně v této problematice figuruje více faktorů, které by zmíněná doporučení mohla omezit.

Existují oblasti, které kriminalitu přitahují více než jiné. Z tohoto důvodu byly prozkoumány oblasti v blízkosti barů a hospod. Tato restaurační zařízení, vzhledem ke svému charakteru, kriminalitu logicky přitahují. Konzumace alkoholu a omamných látek má na kriminalitu zásadní vliv. Pomocí Paretova diagramu (Graf 14) byly identifikovány nejběžnější typy deliktů, které byly spáchány do 100 m od baru či hospody. Je zřejmé, že nejběžnějším typem je nějaký druh krádeže. Typy trestných činů se nepatrně odlišují podle toho, jestli byly spáchány v blízkosti baru nebo hospody (Graf 15). Ve většině jsou si výčty podobné. Rozdílem je, že bary přitahují více výtržnosti, zatímco hospody přitahují trestné činy spojené s drogami, jako je toxikománie či nedovolené držení omamných látek.

Dalším doporučením, které by mohlo vést k větší úspěšnosti data miningu v prevenci kriminality, je přesnější časový údaj záznamů. S odborníky bylo konzultováno, že jednotlivé záznamy nemají nutně přesný časový údaj ohledně hodin. Pokud bude časový údaj přesnější, může to vést k lepší identifikaci určitých trestných činů (den/noc, odpoledne/dopoledne apod.).

Dále bylo zjištěno, že popis přestupků není dostatečně jemný. Pod označení přestupek proti veřejnému pořádku se řadí například rušení nočního klidu či poškození turistické značky.

Z důvodu nedostatečné hloubky popisu musely být přestupky z datasetu vyřazeny. Pokud budou přestupky popsány detailněji, bude možné je lépe analyzovat, případně predikovat.

Dalším možným zlepšením je zápis typu krádeže v případě krádeže na osobě. Je vhodné oddělit, jestli se jedná o materiální či peněžní krádež. To by mohlo vést k lepší identifikaci například v případě krádeží důchodů v okolí pošt.

Jako největší možný přínos autor hodnotí propojení s dalšími neveřejnými daty po vzoru zahraničních zkušeností. Vzhledem k nedávnému sčítání lidu jsou nyní dostupná různá data, jako je například náboženství, příjem, počet osob v obydlí, etnika apod. v návaznosti na adresu bydliště. Na základě těchto dat by mohly vzniknout speciální mapy. Tyto mapy by pomocí zmíněných dat mohly poskytovat informace o oblastech. Takto vzniklá data by se dala proložit se záznamy o trestných činech. Důsledkem by mohlo být lepší vysvětlení některých trestných činů. Tato data bohužel nejsou veřejně dostupná, a proto nemohla být zohledněna v této práci.

Závěr

Práce se zabývala výzkumem přístupu k prevenci kriminality. V první části práce jsou představeny přístupy, které jsou aplikovány v zahraničí. Vedle zahraničí je také přidán náhled na data mining a prevenci kriminality v oblasti ČR. Na základě tohoto přehledu vznikl seznam několika vybraných algoritmů, které jsou v různých státech světa používané. Některé z těchto postupů byly dále použity v práci k analýze poskytnutého datasetu.

Dataset, který je analyzovaný v této práci, pochází z prostředí České republiky, konkrétně z oblasti města Liberce. Tato data byla sbírána od roku 2013 až do současnosti, přičemž pro účely této práce data končí 31. prosincem roku 2021. Dataset tvoří celkem 109 112 záznamů o trestných činech a přestupcích ze zmíněného období. Data pochází z oblasti 14,96 ° až 15,132 ° východní délky a 50,708 ° až 50,823 ° severní šířky. Jedná se tedy o čtverec s přibližnými rozměry 12×12 km. Vzhledem k ovlivnění dat v letech 2020 a 2021, které bylo způsobeno pandemií koronaviru COVID-19, jsou data z těchto let vynechána a data končí 31. prosincem roku 2019. Data byla upravena a restrukturalizována tak, aby práce s nimi byla jednodušší a mohly být aplikovány některé postupy. Bylo zjištěno, že 87 % všech záznamů jsou přestupky. Vzhledem k nedostatečně jemnému označení přestupků nejsou vhodné k analýzám, a proto byly z datasetu vyřazeny.

Dále byly prozkoumány další možnosti rozšíření datasetu. Drtivá většina dat, které jsou veřejně dostupné a použitelné, jsou vztažena pouze na oblasti okresů. Celý poskytnutý dataset pochází z jednoho okresu. Taková data k lepšímu pochopení trestné činnosti ne úrovni města nepomohou. Data byla rozšířena dvěma veřejně dostupnými způsoby. Prvním způsobem jsou data z OpenStreetMaps. Tato celosvětová databáze poskytuje geografické údaje téměř o všem. Z této databáze byly získány údaje o zastávkách MHD, bankomatech, policejních stanicích, poštách, obchodech, restauračních zařízeních, hřbitovech a čerpacích stanicích. Tato data byla k původnímu datasetu přidána. Ke každému trestnému činu byl přidán jeden nejbližší prvek z každé kategorie včetně vzdálenosti k němu. Za účelem tohoto rozšíření datasetu autor práce naprogramoval miniaplikaci v programu C#. Tato aplikace z databázového souboru vyexportuje záznamy spadajících do zmíněných skupin. Druhým způsobem jsou hydrometeorologické údaje. Ty byly získány z portálu ČHMÚ. K datům byly přidány na základě průniku datumu spáchání trestného činu. Bohužel, bylo zjištěno, že toto obohacení dat nebylo příliš přínosné a nepomohlo odhalit nové vysvětlující charakteristiky.

Jedním z hlavních přínosů této práce bylo obohacení autorových zkušeností v oblasti zpracování dat. Autor si mohl vyzkoušet, jaké to je zpracovat skutečná data, která pocházejí z reálného prostředí. Vyzkoušel si celý proces přípravy dat. Oproti modelovým datům je to náročná práce. Dále si také autor ověřil teorii. Příprava a zpracování dat je v celém postupu ten nejzdlouhavější proces. Dále se autor zdokonalil v používání programu SPSS Modeler, v jehož rámci se autor snažil získat nové informace z dat pomocí deskripce, a také se pokusil získat další informace pomocí některých modelovacích algoritmů. Zároveň autor získal zkušenosti a dalšími programy, jako KNIME nebo QGis.

Část práce se také věnoval použití algoritmů DBSCAN a HDBSCAN. Prozkoumán byl nejen způsob použití, ale také bylo do značné míry prověřeno možné nastavení parametrů obou algoritmů tak, aby výstupem byly kvalitní shluky. Na základě výstupů z těchto algoritmů by mohly být provedeny další analýzy jednotlivých nalezených shluků.

Do budoucna má data mining ve spojení s kriminalitou své uplatnění. Tato oblast je stále rozvíjena a měla by být podrobena dalším studiím. Během tvorby této diplomové práce byla odhalena některá doporučení, která by do budoucna mohla zkvalitnit modely a pochopení trestné činnosti.

Prevence kriminality pracuje s tzv. plovoucím úzkým místem. Odborníci správně argumentovali, že celková úspěšnost prevence kriminality je složitá záležitost. Pokud pomocí data miningu bude identifikována konkrétní oblast s danou trestnou činností, podnikne policie příslušné kroky ke snížení kriminality pomocí opatření, jako je například vyšší počet hlídek. V dané oblasti, která byla odhalena algoritmem, dojde ke snížení kriminality. Následkem ale bude, že bude jiná kriminalita růst v jiné oblasti. Tento cyklus se bude dále opakovat. Výsledkem tedy může být pouze celkové snížení kriminality. Snížení bude ale pouze na minimální úroveň, nikoliv nulovou.

Citovaná literatura

ALLAN, Andy, 2022. OpenStreetMap. *OpenStreetMap* [online] [vid. 2022-01-27]. Dostupné z: https://www.openstreetmap.org/

ALQAHTANI, Ayidh, Ajwani GARIMA a Ahmad ALAIAD, 2019. Crime Analysis in Chicago City. In: *2019 10th International Conference on Information and Communication Systems (ICICS)*: *2019 10th International Conference on Information and Communication Systems (ICICS)* [online]. Irbid, Jordan: IEEE, s. 166–172 [vid. 2022-02-06]. ISBN 978-1-72810-045-6. Dostupné z: doi:10.1109/IACS.2019.8809142

BAUER, Vanessa, 2019. Predictive Policing in Germany. Opportunities and challenges of data-analytical forecasting technology in order to prevent crime. [online]. [vid. 2021-12-23]. Dostupné z: doi:10.13140/RG.2.2.35642.95688

BERBA, 2022. *A gentle introduction to HDBSCAN and density-based clustering | by Pepe Berba | Towards Data Science* [online] [vid. 2022-05-16]. Dostupné z: https://towardsdatascience.com/a-gentle-introduction-to-hdbscan-and-density-based-clustering-5fd79329c1e8

BOEHMKE, Brad a Brandon M. GREENWELL, 2019. *Hands-on machine learning with R*. Boca Raton: CRC Press. Chapman & Hall/CRC the R series. ISBN 978-1-138-49568-5.

CURTIS-HAM, Sophie a Darren WALTON, 2017. Mapping crime harm and priority locations in New Zealand: A comparison of spatial analysis methods. *Applied Geography* [online]. **86**, 245–254. ISSN 01436228. Dostupné z: doi:10.1016/j.apgeog.2017.06.008

CURTIS-HAM, Sophie a Darren WALTON, 2018. The New Zealand Crime Harm Index: Quantifying Harm Using Sentencing Data. *Policing: A Journal of Policy and Practice* [online]. **12**(4), 455–467. ISSN 1752-4512, 1752-4520. Dostupné z: doi:10.1093/police/pax050

ČESKÁ SPRÁVA SOCIÁLNÍHO ZABEZPEČENÍ, 2022. *Základní informace - Česká správa sociálního zabezpečení - otevřená data* [online] [vid. 2022-05-25]. Dostupné z: https://data.cssz.cz/

ČESKÝ STATISTICKÝ ÚŘAD, 2022. Český statistický úřad. *Český statistický úřad* [online] [vid. 2022-05-25]. Dostupné z: https://www.czso.cz/csu/czso/domov

ČHMÚ, 2022. *Portál ČHMÚ : Historická data : Počasí : Denní data : Data ze stanic sítě RBCN* [online] [vid. 2022-04-03]. Dostupné z: https://www.chmi.cz/historicka-data/pocasi/denni-data/data-ze-stanic-site-RBCN#

EGBERT, Simon, 2019. Predictive Policing and the Platformization of Police Work. *Surveillance & Society* [online]. **17**(1/2), 83–88. ISSN 1477-7487. Dostupné z: doi:10.24908/ss.v17i1/2.12920

FAKULTA PŘÍRODOVĚDNĚ-HUMANITNÍ A PEDAGOGICKÁ TUL, 2021. *Otevřená data TUL* [online] [vid. 2022-05-25]. Dostupné z: https://opendata.tul.cz/

GARCÍA, Enrique, Cristóbal ROMERO, Sebastián VENTURA a Toon CALDERS, 2007. Drawbacks and solutions of applying association rule mining in learning management systems. 10.

GEOVAP, 2008. *Marushka* [online] [vid. 2022-05-25]. Dostupné z: https://marushkapub.liberec.cz/

GIS PČR, 2022. *Mapa kriminality* [online] [vid. 2022-04-02]. Dostupné z: https://kriminalita.policie.cz/

GUPTA, Manish a B. CHANDRA, 2006. *Crime Data Mining for Indian Police Information System*. B.m.: Indian Institute of Technology Delhi.

HARRISON, Onel, 2019. Machine Learning Basics with the K-Nearest Neighbors Algorithm. *Medium* [online] [vid. 2022-02-17]. Dostupné z: https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

CHAPMAN, Pete, Julian CLINTON, Randy KERBER, Thomas KHABAZA, Thomas REINARTZ, Colin SHEARER a Rüdiger WIRTH, 2000. *CRISP-DM 1.0: Step-by-step data mining guide.* [online]. 2000. B.m.: CRISP-DM Consortium. Dostupné z: https://www.the-modeling-agency.com/crisp-dm.pdf

INGILEVICH, Varvara a Sergey IVANOV, 2018. Crime rate prediction in the urban environment using social factors. *Procedia Computer Science* [online]. **136**, 472–478. ISSN 18770509. Dostupné z: doi:10.1016/j.procs.2018.08.261

INSTITUTE FÜR MUSTERBASIERTE PROGNOSTECHNIK, 2022. Institute für musterbasierte Prognostechnik. *LogObject AG* [online] [vid. 2022-03-06]. Dostupné z: https://logobject.com/en/solutions/precobs-predictive-policing/

INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, ed., 2012. *2012 International Conference on Advances in Engineering, Science and Management (ICAESM 2012): Nagapattinam, Tamil Nadu, India, 30 - 31 March 2012*. Piscataway, NJ: IEEE. ISBN 978-81-909042-2-3.

MARTINEZ-PLUMED, Fernando, Lidia CONTRERAS-OCHANDO, Cesar FERRI, Jose HERNANDEZ-ORALLO, Meelis KULL, Nicolas LACHICHE, Maria Jose RAMIREZ-QUINTANA a Peter FLACH, 2021. CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering* [online]. **33**(8), 3048–3061. ISSN 1041-4347, 1558-2191, 2326-3865. Dostupné z: doi:10.1109/TKDE.2019.2962680

MIGURSKI, Michal, 2020. *Osmosis* [online]. Java. B.m.: OpenStreetMap on GitHub [vid. 2022-01-27]. Dostupné z: https://github.com/openstreetmap/osmosis

MINISTERSTVO PRÁCE A SOCIÁLNÍCH VĚCÍ, 2022. *Datové sady* [online] [vid. 2022-05-25]. Dostupné z: https://data.mpsv.cz/web/data

MINISTERSTVO VNITRA ČESKÉ REPUBLIKY, 2022. *Datové sady - Národní katalog otevřených dat (NKOD)* [online] [vid. 2022-05-25]. Dostupné z: https://data.gov.cz/datov%C3%A9-sady

NCRB, 2018. *National Crime Records Bureau* [online] [vid. 2022-01-05]. Dostupné z: https://ncrb.gov.in/en

OSMF DATA WORKING GROUP, 2021. *PBF Format – OpenStreetMap Wiki* [online] [vid. 2022-01-27]. Dostupné z: https://wiki.openstreetmap.org/wiki/PBF_Format

OSMF DATA WORKING GROUP, 2022. Planet OSM. *Planet OSM* [online] [vid. 2022-01-27]. Dostupné z: https://planet.openstreetmap.org/

RAO, C. Radhakrishna (Calyamp a Josef MACHEK, 1978. Lineární metody statistické indukce a jejich aplikace : Linear Statistical Inference and Its Applications (Orig.) [online]. [vid. 2022-02-22]. Dostupné z: https://is.muni.cz/publication/136304/en/Linearni-metody-statisticke-indukce-a-jejich-aplikace-Linear-Statistical-Inference-and-Its-Applications-Orig/Radhakrishna-Calyamp-Rao-Machek

REICHHOFF, Martin, 2017. The Effect of Weather on Crime: An Investigation of Weather and Annual Crime Rates. 41.

SHARMA, Pulkit, 2019. *K Means Clustering | K Means Clustering Algorithm in Python* [online] [vid. 2022-02-17]. Dostupné z: https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/

SINGH, Neetu, Chengappa BELLATHANDA KAVERAPPA a Jehan D. JOSHI, 2018. Data Mining for Prevention of Crimes. In: Sakae YAMAMOTO a Hirohiko MORI, ed. *Human Interface and the Management of Information. Interaction, Visualization, and Analytics* [online]. Cham: Springer International Publishing, Lecture Notes in Computer Science, s. 705–717 [vid. 2022-01-05]. ISBN 978-3-319-92042-9. Dostupné z: doi:10.1007/978-3-319-92043-6_55

UNIVERSITY OF KONSTANZ, 2022. : Open for innovation. *KNIME* [online] [vid. 2022-03-06]. Dostupné z: https://www.knime.com/open-for-innovation-0

UNIVERSITY OF WAIKATO, 2018. *Weka 3 - Data Mining with Open Source Machine Learning Software in Java* [online] [vid. 2022-03-06]. Dostupné z: https://www.cs.waikato.ac.nz/~ml/weka/

VERMA, Gunjan a Vineeta VERMA, 2012. Role and Applications of Genetic Algorithm in Data Mining. *International Journal of Computer Applications* [online]. **48**, 5–8. Dostupné z: doi:10.5120/7438-0267

VÚT BRNO, 2022. Extract: Czech Republic. *Extract: Czech Republic* [online] [vid. 2022-01-27]. Dostupné z: https://osm.fit.vutbr.cz/extracts/czech_republic/

WILKINSON, Leland a Michael FRIENDLY, 2009. The History of the Cluster Heat Map. *The American Statistician* [online]. **63**(2), 179–184. ISSN 0003-1305, 1537-2731. Dostupné z: doi:10.1198/tas.2009.0033

WILLMOTT, Cj a K MATSUURA, 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research* [online]. **30**, 79–82. ISSN 0936-577X, 1616-1572. Dostupné z: doi:10.3354/cr030079

YIN, Ping a Xitao FAN, 2001. Estimating R^2 Shrinkage in Multiple Regression: A Comparison of Different Analytical Methods. *The Journal of Experimental Education* [online]. **69**(2), 203–224. ISSN 0022-0973, 1940-0683. Dostupné z: doi:10.1080/00220970109600656

YOBERO, Czar, 2018. *RPubs - C5.0 Decision Tree Algorithm* [online] [vid. 2022-04-09]. Dostupné z: https://rpubs.com/cyobero/C50