TECHNICKÁ UNIVERZITA V LIBERCI
Ekonomická fakulta

Včasné varování před zvýšeným rizikem vzniku dopravní nehody s využitím data miningu

Disertační práce

Studijní program: P6209 – Systémové inženýrství a informatika
Studijní obor: 6209V003 – Ekonomická informatika

Autor práce: **Ing. Bc. Marián Lamr**
Vedoucí práce: doc. Ing. Jan Skrbek, Dr.

Liberec 2018

TECHNICAL UNIVERSITY OF LIBEREC
Faculty of Economics

Early Warning of the Increased Risk of Traffic Accidents Based on the Application of Data Mining

Dissertation

Study programme: P6209 – System Engineering and Informatics
Study branch: 6209V003 – Managerial Informatics

Author: **Ing. Bc. Marián Lamr**
Supervisor: doc. Ing. Jan Skrbek, Dr.

Liberec 2018

Prohlášení

Byl jsem seznámen s tím, že na mou disertační práci se plně vztahuje zákon č. 121/2000 Sb., o právu autorském, zejména § 60 – školní dílo.

Beru na vědomí, že Technická univerzita v Liberci (TUL) nezasahuje do mých autorských práv užitím mé disertační práce pro vnitřní potřebu TUL.

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

Disertační práci jsem vypracoval samostatně s použitím uvedené literatury a na základě konzultací s vedoucím mé disertační práce a konzultantem.

Současně čestně prohlašuji, že tištěná verze práce se shoduje s elektronickou verzí, vloženou do IS STAG.

Datum:

Podpis:

Anotace a klíčová slova

V poslední dekádě dochází v České republice každoročně k nárůstu počtu dopravních nehod. Obdobný trend prezentují i statistiky počtu lehkých i těžkých zranění vzniklých na silničních komunikacích. Navzdory očekávání neklesá významněji ani výše materiálních škod. V reakci na tento trend policie pravidelně realizuje represivní kroky vůči řidičům, porušujícím pravidla silničního provozu, které však ne vždy splňují původní očekávání. Následující text se snaží přispět ke zvýšení bezpečnosti silničního provozu formou prevence vzniku dopravních nehod díky využití progresivních ICT prostředků.

Předložená disertační práce se zabývá především včasným varováním řidičů před zvýšeným rizikem vzniku dopravní nehody s využitím data miningu. Nejprve je v práci popisován aktuální stav preventivních prvků a systémů zvyšujících bezpečnost účastníků silničního provozu v ČR. Dále jsou analyzovány a vyhodnocovány nejvhodnější data miningové metody, algoritmy a přístupy s ohledem na potřeby řešení realizovaného v rámci disertační práce. Stěžejním tématem disertační práce je konceptuální návrh systému včasného varování před zvýšeným rizikem dopravní nehody. Navrhovaný systém by měl v reálném čase a místě predikovat riziko vzniku dopravní nehody na základě modelů vytvořených pomocí data miningových algoritmů a poskytovat adekvátní upozornění pro řidiče. Součástí práce je implementace databáze dopravních nehod a vyhodnocení možností využití těchto dat pro predikční účely. V práci je dále řešeno hledání skrytých závislostí v datech o dopravních nehodách, které je nedílnou součástí řídicí části navrhovaného řešení. V neposlední řadě jsou v práci představeny modelové situace a ekonomické aspekty a přínosy celého řešení.

Klíčová slova: včasné varování, data mining, predikce, systém, dopravní nehody, data, analýza, shluky

Annotation and keywords

Recently, the number of traffic accidents has been regularly increasing. This unflattering phenomenon unevitably resulted in a growing amount of ligh and havy injuries. Despite the anticipation, material demages are not declining either. In response to this trend, the police is regularly implementing repressive actions against drivers, but these do not always meet initial expectations. This doctoral thesis deals with the use of modern ICT tools to prevent the occurrence of traffic accidents.

This thesis deals with the possibility to generate early warnings in order to inform the drivers about the increased risk of traffic accidents using data mining techniques. The first section describes the current state of existing features and systems increasing the safety of road users in the Czech Republic. Further, the most suitable data mining methods, algorithms and approaches are analyzed and evaluated with regard to the needs and specifications the solution proposed in this dissertation. The main objective of the thesis is the conceptual design of an early warning system informing the drivers about the increased risk of a traffic accident. The proposed system should predict the risk of accidents in real-time and on the basis of the models created using data mining algorithms, and provide adequate warnings to the driver. The implementation of the traffic accidents database and the evaluation of the possibilities of using this data for prediction purposes represents a significant part of the dissertation. The examination of the hidden dependencies in traffic accidents data is described as an integral proces of the proposed solution. Last but not least, the thesis presents model situations, economic aspects and benefits of the potential implemention of such concept.

Keywords: early warning, data mining, prediction, system, traffic accidents, data, analysis, clusters

Annotations et mots clés

Ces dernières années, les accidents de la route se sont multipliés. Cette tendance peu flatteuse concerne également le nombre de blessures légères et lourdes induites. Malgré les attentes, le montant des dégâts matériels ne diminue pas de manière significative. Face à cette tendance, la police planifie régulièrement des actions répressives contre les conducteurs, mais celles-ci ne répondent pas toujours aux attentes initiales. Cette thèse traite de l'utilisation des outils TIC modernes pour prévenir les accidents de la circulation.

La thèse porte sur l'alerte précoce des conducteurs contre le risque accru d'accident de la route en utilisant l'exploration de données. On décrit d'abord l'état actuel des éléments et des systèmes de sécurité qui augmentent la sécurité des usagers de la route en République tchèque. En outre, les données les plus appropriées de la méthode d'extraction, des algorithmes et des approches sont analysées et évaluées en fonction des besoins de la solution réalisée dans le mémoire. Le sujet principal de la thèse est la conception d'un système d'alerte précoce contre le risque accru d'accident de la circulation. Le système proposé devrait prévoir le risque d'accident en temps réel et sur la base de modèles créés à l'aide d'algorithmes d'exploration de données et fournir des avertissements adéquats au conducteur. Une partie du travail consiste à mettre en place une base de données sur les accidents de la circulation et à évaluer les possibilités d'utiliser ces données à des fins de prévision. La thèse traite également de la recherche de dépendances cachées dans les données sur les accidents de la route, qui font partie intégrante de la partie contrôle de la solution proposée. Enfin, la thèse présente les situations modèles et les aspects économiques ainsi que les avantages de la solution globale.

Mots-clés: alerte précoce, exploration de données, prédiction, système, accidents de la route, données, analyse, nuage de données

Poděkování

Touto cestou bych chtěl poděkovat za podporu celé své rodině a vedoucímu disertační práce doc. Ing. Janu Skrbkovi, Dr.za odborné vedení, ochotu vždy pomoci a za cenné rady nejen při zpracování disertační práce ale i během celého studia.

Obsah

Seznam zkratek

ABS - Anti-lock Brake System

ACC - Autonomous Cruise Control

API - Application Programming Interface

ASR - Anti-Slip Regulation

CRISP-DM - Cross-Industry Standard Process for Data Mining

CARMA - Continuous Association Rule Mining Algorithm

CURL - Client for URLs

DBSCAN - Density-Based Spatial Clustering of Applications with Noise

DENCLUE - DENsity-based CLUstEring

ECALL – Emergency Call

ELKI - Environment for Developing KDD-Applications Supported by Index-Structures

ESP - Electronic Stability Program,

FTP - File Transfer Protocol

HeERO – Harmonised eCall European Pilot

HTTP - Hypertext Transfer Protocol

KNIME - Konstanz Information Miner

MinPts- Minimum of Points

LTE - Long Term Evolution

OPTICS - Ordering points to identify the clustering structure

PČR – Policie české republiky

RDS-TMC - Radio Data System - Traffic Message Channel

SEMMA - Sample, Explore, Modify, Model, and Assess

SVM -Ssupport Vector Machine

WEKA - Waikato Environment for Knowledge Analysis

Seznam obrázků

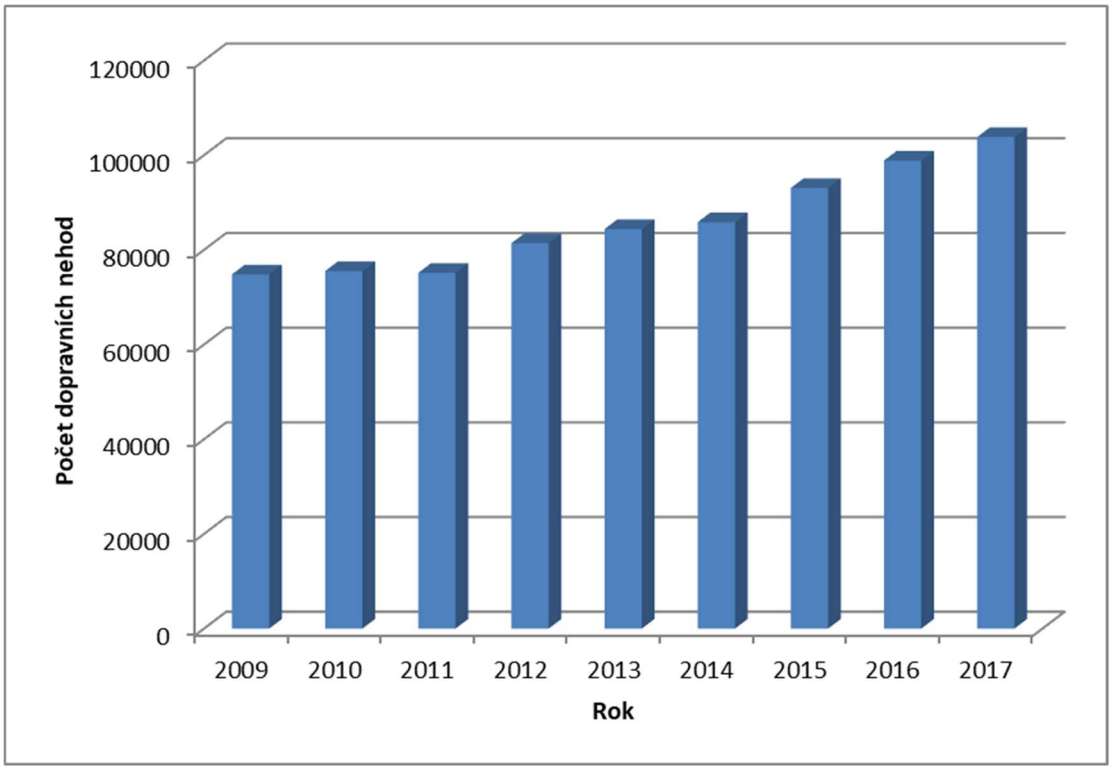
Seznam grafů a tabulek

Úvod

Dopravní nehody jsou bohužel každodenní a neoddělitelnou součástí silničního provozu. I přesto, že se policie snaží různými způsoby zvýšit bezpečnost silničního provozu, počty dopravních nehod, ztráty na životech, majetku a počty zranění se nedaří snížit podle očekávání.

Z aktuálních statistik dopravní nehodovosti vyplývá, že od roku 2011 dochází k neustálému nárůstu dopravních situací končících materiálními škodami či zdravotní újmou. Je důležité připomenout, že v roce 2009 se zvýšila hranice pro povinné nahlášení nehody na částku 100 000 Kč. Ve srovnání s daty před rokem 2008 by se mohlo zdát, že v roce 2009 došlo k dramatickému poklesu počtu nehod, nicméně v tomto kontextu by porovnání neměla dostatečnou vypovídající hodnotu. V Grafu 1 je pro relevantní porovnání zobrazen vývoj počtu dopravních nehod od roku 2009 do roku 2017

Graf 1: Vývoj počtu dopravních nehod

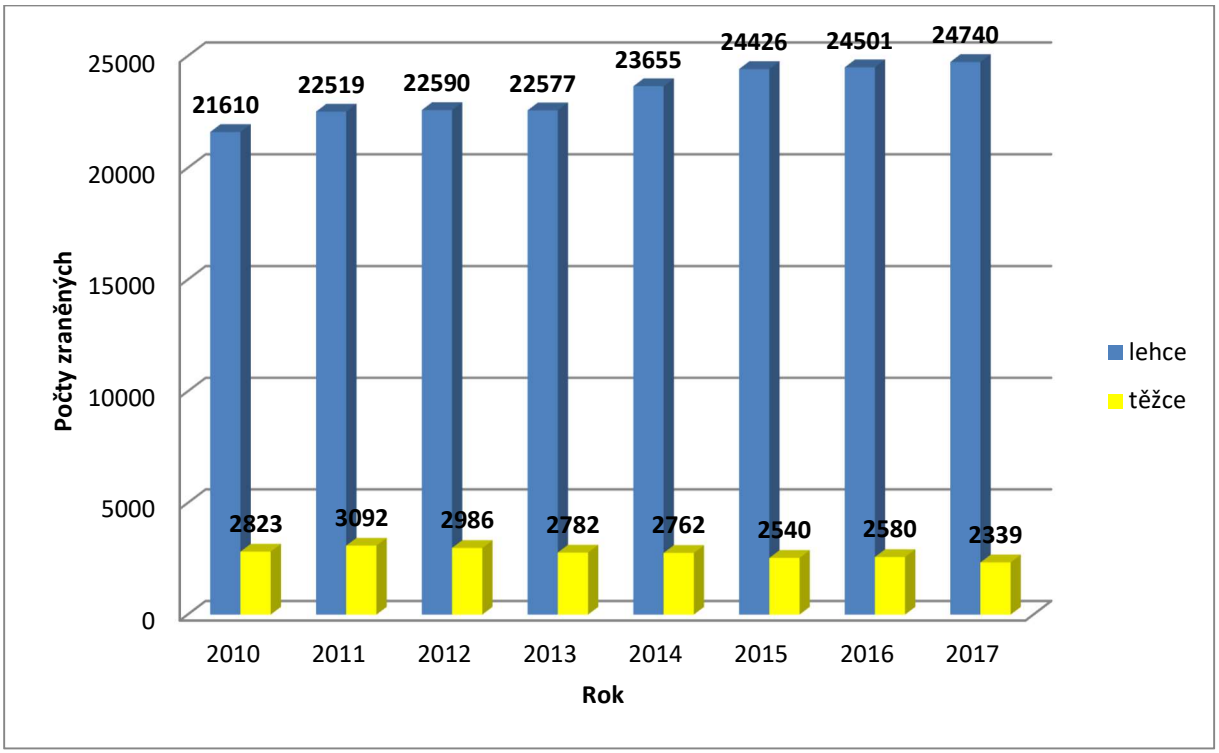


Zdroj: vlastní zpracování, data: Policie ČR

Vzhledem k již zmiňovanému růstu počtu dopravních nehod se policie opakovaně snaží hledat způsoby zvýšení bezpečnosti provozu. Mezi tato opatření lze zařadit zejména zvýšení počtu silničních kontrol, zvýšeni pokut za dopravní přestupky, zavedení bodového systému a řadu regionálních či celostátních dopravně bezpečnostních akcí. Ministerstvo dopravy plánuje další

zvyšování sankcí za určité druhy přestupků i přesto, že se ukazuje, že sankcionování řidičů pomáhá pouze krátkodobě. („Ministerstvo dopravy chce zvýšit pokuty za rychlost. Až trojnásobně", 2015). I přes všechna zmiňovaná opatření neklesají ani počty lehce a těžce zraněných osob. Vývoj počtu lehkých a těžkých zranění od roku 2010 do roku 2017 je zobrazen na grafu 2.

Graf 2: Počty lehce a těžce zraněných v letech 2010 až 2017

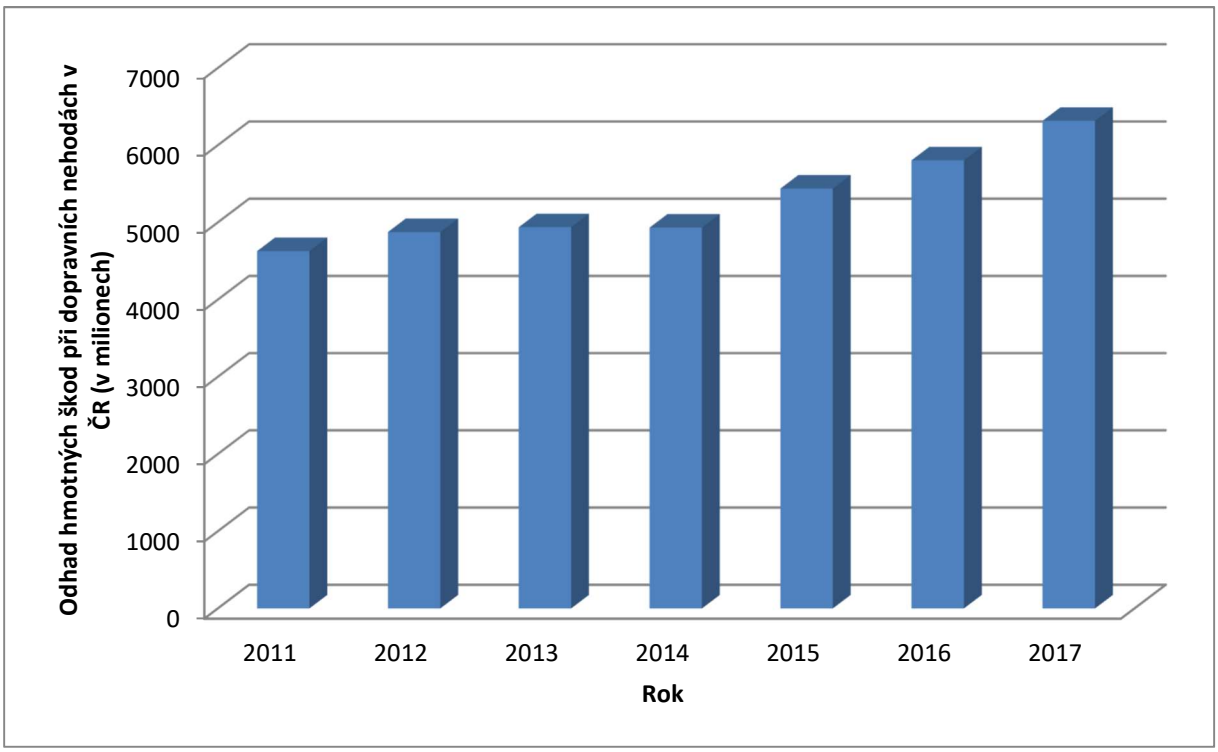


Zdroj: vlastní zpracování, data: Policie ČR

S rostoucím počtem dopravních nehod souvisí i zvyšování hmotných škod. Trend posledních let v této oblasti lze vidět v Grafu 3, kde jsou zobrazeny odhady škod při dopravních nehodách podle policie ČR od roku 2011 do roku 2017.

I z toho důvodu je třeba hledat nové přístupy předcházení dopravním nehodám. Jedním z možných způsobů, který by měl vést ke zlepšení situace na silnicích, je cesta využívání moderních technologií, které by umožnily zvýšit bezpečnost dopravy díky lepší informovanosti řidičů o potenciálních nebezpečích na trase.

Graf 3: Hmotné škody při dopravních nehodách 2011-2017



Zdroj: vlastní zpracování, Policie ČR

Cílem této práce je vytvořit konceptuální návrh systému umožňujícího v reálném čase a místě predikovat riziko dopravní nehody. Systém je založený na predikčních modelech, které jsou vytvářeny pomocí data miningových technik a nástrojů využívajících především algoritmy shlukové analýzy. V následujících kapitolách budou diskutovány východiska a předpoklady pro existenci systému včasného varování před zvýšeným rizikem dopravní nehody. Nejprve bude provedena rešerše aktuálního stavu technických zařízení a systémů zvyšujících bezpečnost účastníků dopravního provozu. Dále bude věnována pozornost samotnému oboru data mining jako takovému. Budou diskutovány vybrané postupy, programové vybavení a algoritmy při řešení dané problematiky. V další kapitole se čtenář dozví, jak je v současné době možné získávat, zpracovávat a uchovávat data o dopravních nehodách, na jejichž základě je budován systém včasného varování. Následně bude představen princip systému včasného varování s modelovými úlohami jeho využití. V práci jsou dále řešeny možnosti hledání skrytých závislostí v datech o dopravních nehodách, což je nedílnou součástí řídicí části navrhovaného řešení. V neposlední řadě jsou v práci představeny modelové situace užití navrhovaného řešení, ekonomické aspekty a přínosy celého řešení.

1 Aktuální stav systémů zvyšujících bezpečnost dopravy v ČR

Zvyšování bezpečnosti účastníků silničního provozu je možné provádět pomocí celé řady prvků. V této kapitole nebude věnována pozornost preventivním opatřením typu dopravně bezpečnostních akcí prováděných policií, či zásahům do bodového systému. Tato kapitola je věnována vybraným typickým aktivním a pasivním prvkům umístěným v běžných automobilech dnešní doby, asistenčním systémům a systémům, které by měly předcházet dopravním nehodám, nebo umožňují zmírnit dopad dopravní nehody.

1.1 Pasivní prvky

Nejprve budou popisovány vlastnosti vozu, zařízení a systémy, které ke své funkci nepoužívají žádný externí zdroj informací a jsou integrovány ve vozidlech. Daná zařízení ke své činnosti buď nepotřebují žádná čidla, anebo využívají informace, které jim dodávají pouze čidla dostupná v daném automobilu.

Prvky pasivní bezpečnosti lze definovat jako konstrukční zařízení, která jsou aktivována až ve chvílích, kdy nastane nebezpečná situace nebo dopravní nehoda. Cílem těchto zařízení je tedy minimalizovat následky dopravní nehody. Řeší tedy situaci ex post. Nejběžnějšími prvky pasivní bezpečnosti v dnešních automobilech jsou především zádržné systémy (bezpečnostní pás, předpínač bezpečnostního pásu, dětská sedačka), opěrka hlavy, bezpečná konstrukce karoserie a airbag (Vlk, 2006).

1.1.1 Zádržné systémy

Zádržný systém slouží ke snížení rizika poranění pasažéra v případě náhlého snížení rychlosti vozidla omezením dopředného pohybu uživatele. Zádržné systémy eliminují nežádoucí pohyb posádky vozidla během nárazu. Mohou mít různá provedení a konstrukční řešení, přičemž nejznámější jsou tříbodové a dvoubodové bezpečnostní pásy a dětské autosedačky.

Jedním ze základních prvků pasivní bezpečnosti je bezpečnostní pás, jehož úkolem je snížení rychlosti nárazu hlavy a hrudníku. Udržení dopředného posunutí cestujícího v rámci volného prostoru v interiéru vozidla zabraňuje poranění o vnitřní vybavení vozidla. Pasažéři všech osobních a nákladních vozidel včetně autobusů jsou dle zákona o silničním provozu povinni se před jízdou připoutat. Při nedodržení této povinnosti je totiž ohrožen nejen ten, kdo se

nepřipoutal, ale i ostatní pasažéři vozidla (Vlk, 2006), ("Aktivní a pasivní prvky bezpečnosti motorových vozidel", 2015).

1.1.2 Airbag

Stejně jako bezpečnostní pásy, mají i airbagy za úkol chránit cestujícího před nárazem do vybavení interiéru vozidla, přičemž airbagy většinou chrání konkrétní část těla. Aby byl účinek airbagu co nejefektivnější, je nutné používání bezpečnostních pásů. Hlavní součástí systému airbagů je silný textilní vak, který je složen a uložen v modulu na sloupku řízení řidiče a na přístrojové desce pro cestujícího. Když snímače na palubě detekují čelní nehodu, která překračuje nastavenou prahovou hodnotu, rozbušky rozvinou airbagy. Vysokotlaké chemicky vyráběné plyny vytlačují vak z modulu a dostatečně rychle ho nafouknou tak, že je včas na správném místě před cestujícím (Evans, 2004).

1.1.3 Hlavová opěrka

Hlavová opěrka patří mezi standardní vybavení všech moderních automobilů a je důležitým prvkem pasivní ochrany pasažérů v automobilu. Správné nastavení opěrky snižuje riziko úrazu krční páteře a zamezuje trvalým následkům nehody. Při nárazech zezadu může dojít k poranění krční páteře. Barnsley (1994) definuje poranění krční páteře jako poranění jednoho nebo více elementů krční páteře, které vznikají z inerciálních sil působících na hlavu v průběhu nehody motorového vozidla, které vede k vnímání bolesti krku. Správné nastavení hlavové opěrky je však důležité i při čelním nárazu, kdy po zachycení těla z airbagu a bezpečnostních pásů se tělo vrátí zpět do sedadla. U některých novějších automobilů je nastavení hlavové opěrky řešeno aktivním opěrným systémem, který může zmírnit následky uvedených situací.

1.2 Aktivní prvky

Aktivní prvky bezpečnosti ve vozidle se na rozdíl od prvků pasivní bezpečnosti snaží dopravním nehodám a krizovým situacím předcházet. Jedná se především o nejrůznější technická zařízení, systémy a vlastnosti vozu využívající data z čidel ve vozidle. Základními prvky aktivní ochrany jsou kvalitní brzdy a přesné řízení. V dnešních vozidlech však existuje celá řada sofistikovaných systémů a řešení jako například elektronický protiblokovací systém ABS (Anti-lock Braking Systems), protiprokluzový systém ASR (Antriebs-Schlupf-Regelung) a elektronický stabilizační systém ESP (Electronic Stability Program) a další.

1.2.1 Protiblokovací systém

V situacích, kdy je řidič nucen prudce brzdit automobil, může dojít k zablokování kol. Tyto krizové situace nastávají především na kluzké vozovce (voda, námraza). Zmiňovaným situacím lze předejít použitím elektronických protiblokovacích systémů (ABS). ABS pracuje na principu měření otáček na kolech vozidla pomocí snímačů. Signály ze snímačů přicházejí do řídicí jednotky, která je zpracovává a v případě, že z nich vyhodnotí nebezpečí zablokování kol, následně aktivuje elektropneumatické řídicí ventily příslušného kola a tím dojde ke snížení brzdného účinku (Vlk, 2006).

1.2.2 Protiprokluzový systém

Především v situacích, kdy kola na jedné straně automobilu mají jinou přilnavost k vozovce než kola na straně druhé, je užitečný protiprokluzový systém ASR. Protiprokluzových systémů existuje celá řada a různé automobilky tato elektronická zařízení nazývají jinak. Princip těchto systémů spočívá v kontrole prokluzu hnacích kol pomocí čidel, která využívá i ABS a optimalizaci přenosu točivého momentu. Pokud je dostatečná adheze a zatížení nápravy, začnou kola po sešlápnutí pedálu přenášet krouticí moment na vozovku a vozidlo začne zrychlovat. V případě, že je přenášený krouticí moment vyšší, než jaký je kolo schopno přenést na vozovku, začne kolo prokluzovat. Pokud se náhle zvýší otáčky jednoho z hnacích kol, systém ASR dá pokyn řídicí jednotce a ta buď prostřednictvím elektromagnetického ventilu a elektropneumatického řídicího ventilu kolo přibrzdí anebo sníží výkon motoru. V tu chvíli může kolo, které se pohybuje na vozovce s vyšším součinitelem adheze, přenášet hnací moment na vozovku (Vlk, 2006).

1.2.3 Elektronický stabilizační systém

Variant elektronického stabilizačního systému existuje celá řada stejně jako u implementací prokluzovacích systémů. Všechny varianty elektronického stabilizačního systému však pracují na stejném principu. Systém ESP umožňuje využití jízdních vlastností automobilu až na samou hranici fyzikálních zákonů. Systém ESP je rozšířením ABS a ASR, využívá například snímače natočení volantu, snímače otáček všech kol, snímače podélného a příčného zrychlení a další. Na rozdíl od ABS a ASR je systém ESP schopen regulovat skluz pneumatiky v příčném směru. Systém ESP je užitečný především v hraničních situacích a dokáže eliminovat nedotáčivost vozidla, přetáčivost vozidla, či je přínosem při vyhýbacím manévru, kdy řidič uhýbá předmětu na vozovce a prudce trhne volantem. Ve chvíli, kdy řídicí jednotka vyhodnotí konkrétní situaci jako nebezpečnou, automaticky dokáže regulovat účinek brzd jednotlivých kol a hnací moment

motoru. Například nedotáčivý či přetáčivý pohyb vozidla systém eliminuje přibrzděním příslušných kol a snížením točivého momentu na hodnotu odpovídající dané situaci. Každý nově homologovaný vůz musí být od roku 2011 vybaven ESP, přičemž všechna nově prodaná auta od roku 2014 musí mít ESP, i když byla homologována před rokem 2011 (Vlk, 2006), ("Aktivní a pasivní prvky bezpečnosti motorových vozidel", 2015).

1.2.4 Adaptivní tempomat

Dalším milníkem ve vývoji asistenčních systémů se stal adaptivní tempomat ACC (Adaptive Cruise Control), který detailně popisuje (Winner a kol., 2009, Winner 2012). Částečně automatizované řízení bylo umožněno díky implementaci elektronického ovládání brzd a užití dříve velice drahé radarové technologie, která se postupem času zlevnila. Když byl v roce 1999 zaveden adaptivní tempomat, byly jeho možnosti omezené a zpočátku použitelné pouze při rychlostech vyšších než 30 km/h (Jones, 2001). Současné systémy s automatickou převodovkou však mají schopnost využívat tyto vlastnosti při nižších rychlostech a například automaticky následovat další vozidla v dopravní zácpě.

1.2.5 Systémy prevence srážky

Systémy prevence čelní srážky jsou dalším významným krokem ke zvýšení bezpečnosti provozu. Implementací těchto systémů do nových modelů automobilů se zabývá řada automobilek. Systémy předcházení kolizím, které používají levné verze snímačů lidar s nízkým dosahem a s nízkým rozlišením, se v současné době používají pro situace, kdy automobily jedou nízkou rychlostí. Příkladem těchto systémů můžou být: "City Safety" (City Safety, 2014) a "City Stop" (Euro NCAP Advanced Reward 2011, 2011). Oba systémy, zavedené kolem roku 2010, pomáhají předcházet poškození karoserie, což je ekonomicky velmi výhodné. Pro pokročilé aplikace (např. při vyšších rychlostech) je však malý detekční rozsah levných lidarových systémů silně omezujícím faktorem. V letech 2003-2006 byly zavedeny systémy snižování kolizí na dlouhé vzdálenosti založené na radarové technologii původně zavedené s adaptivním tempomatem. Prostřednictvím stoupající úrovně výstrah je řidič upozorněn na hrozící kolizi. Pokud řidič nereaguje, vozidlo aktivně brzdí, aby zmírnilo závažnost havárie, jakmile se srážce již nelze vyhnout (Maurer, 2012). Takové systémy byly zkoumány v rámci projektu EUROPEAN PREVENT (2004-2008) a mohou se ukázat jako zvláště účinné u větších vozidel, jako jsou nákladní automobily, které kvůli své omezené jízdní dynamice vyžadující brzdění dříve než menší vozidla.

Nejnovější třída asistenčních systémů volí a ovládá trajektorie mimo aktuální požadavek řidiče. Krátkodobým cílem je automatizovat jízdu ve vybraných situacích. Jako příklad byly v nedávné době zavedeny asistenční systémy pro dopravní zácpy založené na radarech a 3D kamerových systémech. Spojením podélného a bočního ovládání jsou tyto systémy navrženy pro automatickou jízdu za nízké rychlosti na přetížených dálnicích. Maximální rychlost při jízdě je stále nízká (30 km/h) a omezuje se pouze na zastavení a rozjíždění vozidla, ale tato funkce může nakonec vést směrem k plně automatizované jízdě na dálnici (Bengler, 2014).

1.2.6 Systémy využívající Car2Car a Car2X komunikace

Internet dosud hrál ve vozidlech pouze okrajovou roli. Doposud bylo používání datových spojení omezeno především na informační a navigační podporu. V budoucnu spolu s vývojem v oblasti informačních technologií lze očekávat nové možnosti asistence při řízení. "Řidičská kancelář" je jistě zajímavý koncept pro manažery a podnikatele. Lze předpokládat, že mobilní kancelář a autonomní řidičský balíček budou v budoucnu podléhat intenzivním požadavkům na osobní automobily. Další možné případy využití pro datovou komunikaci zahrnují přidělení parkovacích míst před vlastním příjezdem vozidla na parkoviště nebo komunikace v intermodální dopravě.

Začleněním všech účastníků provozu určité oblasti do společné sítě může být dosaženo nové fáze asistence při řízení, založené na výrazně lepší kvalitě a množství informací o místní dopravní situaci. Světelná zařízení by mohla být nahrazena bezdrátovými přístupovými body, které by řídily vozidla na křižovatce. Zatímco tento přístup by byl účinnější u automatizovaných vozidel než u klasických vozidel, měl by přinést pozitivní výsledky bez ohledu na to, kdo řídí vozidlo. Vozidla vybavená čidly a komunikačními zařízeními v2v by mohla rozšířit svůj obzor prostřednictvím družicového snímání.

V případě těchto systémů mohou vozidla sdělovat zpomalení směrem dozadu nebo informovat o přítomnosti vozidel ve svém slepém místě. Dále mohou připojené vozy vyjednávat o směru jízdy ve prospěch celkového dopravního toku a bezpečnosti.

Při pohledu na obrovský potenciál zvyšování bezpečnosti a efektivity jízdy s ohledem na energetickou, časovou a dopravní infrastrukturu se tyto sítě pravděpodobně brzy stanou skutečností (Bengler, 2014).

1.3 Telematické systémy v ČR

1.3.1 RDS-TMC a informační tabule

RDS-TMC (Radio Data System - Traffic Message Channel) je kanál poskytující přenos předzpracovaných dopravních informací do vozidla, kde se tyto údaje dále zpracují a poskytnou řidiči. RDS-TMC je součástí dopravně informačního systému a vždy je pevně spjat s konkrétní rozhlasovou stanicí, a proto se vysílaná data mohou stanici od stanice lišit. RDS-TMC mohou využívat navigace i mobilní telefony s TMC dekodérem. K označení pozic objektů reálného světa slouží tzv. lokalizační tabulky. Každý řádek v lokalizační tabulce je pevně spjat s konkrétní geografickou entitou (křižovatka, silnice, významný objekt atd.). Nevýhoda tohoto systému tedy spočívá v kvalitě (přesnosti) lokalizačních tabulek, a může se stát, že událost bude zobrazena v jiném místě na komunikaci, než kde ve skutečnosti je. Často se také stává, že informace, které RDS-TMC poskytuje jsou neaktuální a jsou zobrazeny s určitou časovou prodlevou. Služba může mít problémy s přijetím zpráv v místech se slabým FM signálem. I když RDS-TMC informuje o problémech na komunikaci, které se již staly, lze jej zařadit k aktivním systémům, jelikož se snaží dalším nehodám předcházet. S obdobně nastíněnými problémy je nutné počítat u dalšího prostředku pro šíření informací pro řidiče, a to u informačních tabulí, se kterými se nejčastěji setkáváme na dálnicích. Informace, které jsou na těchto tabulích řidičům poskytovány, jsou získávány z jednotného systému dopravních informací. Projekt jednotného systému dopravních informací je společným dílem Ministerstva dopravy, Ředitelství silnic a dálnic a několika dalších organizací (Lamr, 2015a), (Lamr, 2015c).

1.3.2 eCall

Systém eCall řeší především následky dopravní nehody, a proto patří mezi pasivní systémy. Jde o službu navrženou pro umožnění rychlého pohotovostního zásahu v případě dopravní nehody, a to kdekoli na území EU. Cílem systému eCall je zvýšit bezpečnost dopravy v Evropě a snížit počet úmrtí zapříčiněných dopravními nehodami, a také zamezit zraněním a ztrátě na majetku s nimi spojených (European Commission MEMO 13/547, 2013).

Tento projekt spolufinancovaný Evropskou unií předpokládá, že po plném nasazení dokáže v EU každý rok zachránit 2500 životů. eCall by měl urychlit dobu reakce při mimořádné události o 40 % v městských oblastech a o 50 % na venkově. Pilotní testování a zavádění systému eCall v Evropě provádí HeERO consorcium. Do projektu HeERO (Harmonised eCall European Pilot) je zapojena i Česká republika. Podle odhadů stojí ročně dopravní nehody okolo 160 miliard eur, v případě plného nasazení systému eCall by se mělo ušetřit až 20 miliard eur

ročně (ECall deployment - Publication by the European Commission of the Delegated Regulation No 305/2013, 2013).

Každý automobil vybavený eCall zařízením může navázat nouzové spojení s linkou 112 a to buď automaticky při nehodě, anebo manuálně pomocí tlačítka v automobilu. Dle posledních informací Evropská komise schválila, že systém eCall se aktivuje až v případě, kdy senzory v automobilu detekují nehodu. V tu chvíli se kromě hlasového spojení s pracovníky pohotovostní centrum odešle také tzv. minimální soubor dat (MSD), který je standardizován Evropskou komisí pro standardizaci. V následujícím seznamu atributů je uveden i jejich význam.

Minimální soubor dat:

- *Identifikátor zprávy*: verze ve formátu MSD (pozdější verze budou zpětně kompatibilní)
- *Aktivace*: zdali byl eCall požadavek realizován manuálně či automaticky vygenerován
- *Typ požadavku*: zdali byl eCall požadavek v rámci opravdové pohotovostní situace anebo jen testovací
- *Typ vozidla*: osobní automobil, mistní a dálkové autobusy, lehké užitkové vozidla, těžká mašinérie, motorky
- *Identifikační číslo vozidla (VIN)*
- *Typ úložiště paliva ve vozidle*: Jde o důležitou informaci vzhledem k riziku požáru a problémy se zdrojem napájení (např. Benzínová nádrž, Dieselová nádrž, Stlačený zemní plyn (CNG), atd.)
- *Časové razítko*: Časové razítko události
- *Pozice vozidla*: určena palubním systémem v čase vygenerování zprávy – jde o poslední známou pozici vozidla (zeměpisná šířka a délka)
- *Spolehlivost pozice*: tato část bude nastavena na „Nízkou spolehlivost pozice“ v případě že pozice není v mezích +/-150m s 95% spolehlivostí
- Směr: užitečné pro určení směru vozu v okamžiku události
- Nedávná pozice vozidla (Nepovinné): pozice vozidla v (n-1) a (n-2)
- Počet pasažérů (Nepovinné): počet připoutaných pásů
- Další nepovinná data (Nepovinné): v některých případech mohou být v MSD obsaženy další data (dle úvahy výrobce). Tyto data mají na svém začátku identifikační tag (identifikace typu a struktury). Tyto data budou registrována a spravována. PSAP bude mít volný přístup k registru těchto dat

(Ecall becomes a reality!, 2015).

2 Data mining: metodologie, big data, základní přístupy a typické úlohy

V současném moderním světě shromažďuje většina společností data o svých zákaznících. Tato data obsahují mnohdy i velice citlivé a osobní informace, které lze využívat pro nejrůznější účely. Obrovské datové základny ale nebudují jen velké korporace, ale například i státní organizace. Vrátíme-li se v úvahách o významu data miningu k velkým korporacím, lze spojení data miningu a dat o zákaznících využít například v oblasti marketingu, což může být důležitým faktorem pro úspěch firmy v případě, že chce vítězit v konkurenčním boji. Jestliže v dobách minulých znamenal data mining konkurenční výhodu, dnes se pomalu stává nepostradatelnou nutností a nachází se ve všech komerčních i nekomerčních sférách a nejrůznějších oborech.

I samotný vznik data miningu byl podmíněn tím, že organizace shromažďovaly s vynaložením značných prostředků velké množství dat, a právě data mining umožnil zhodnotit tyto prostředky.

Data mining můžeme chápat jako cílené prozkoumávání údajů, kdy na začátku často čerpáme z velkých objemů dat. Postupným zaměřováním se na konkrétní cíl se objem dat zužuje až po fázi modelování, která pracuje již s významně menší datovou množinou. Gartner group definuje data mining jak proces objevování významných netriviálních závislostí, vzorů a trendů prozkoumáváním velkých objemů dat pomocí algoritmů pro odhalování pravidel a pomocí matematických a statistických algoritmů.

Data mining využívá celá řada oborů. Kromě marketingových úloh a úloh zaměřených na zákazníka můžeme na data mining narazit například ve finančním sektoru při skórování žádostí o úvěr či při detekci podvodného chování. Data mining může řešit také medicínské problémy jako je diagnostika chorob či genové inženýrství. S data miningem se lze setkat též v personalistice, školství, státní sféře, logistice, či ve výrobě při predikci selhání strojů. Data mining je rovněž hojně využíván při analýze sociálních sítí. Za speciální druhy data miningu lze označit text mining či web mining, kde v případě text miningu jde o transformaci volného textu na strukturovaná data a v případě web miningu o transformaci záznamů generovaných internetovými servery (logy, cookies). Data mining je využíván také při zpracování obrazu (rozpoznávání osob a předmětů).

Ať už je obor, na který aplikujeme data miningové nástroje a postupy, jakýkoliv, vždy je při řešení data miningových úloh kladen důraz na konkrétní nasazení do praxe. V praxi není podstatné, jaké modely při řešení problému použijeme, je však nutné, aby řešení jako celek fungovalo a přineslo něco užitečného.

Díky možnostem, které s sebou přináší data miningové postupy při zpracovávání velikých objemů dat, se jeví využití DM pro hledání skrytých závislostí u nehod, které tvoří v nějakém konkrétním místě shluk, jako ideální prostředek.

2.1 Data mining a s ním související pojmy

Obor analýzy dat se velmi rychle rozšiřuje, a to jak v rozsahu možností její aplikace, tak i v počtu organizací používajících pokročilou analytiku. Z toho důvodu vzniká značný překryv a nesrovnalosti v oblasti definic pojmu data mining. Termín data mining má pro jednotlivé jedince a organizace různý význam. Pro širší veřejnost má data mining obecnější a poněkud nejasný pejorativní význam - prohledávání velkých zdrojů (často osobních) dat za účelem najít něco vzbuzující náš zájem. Příkladem odlišnosti chápání pojmu data mining může být velká poradenská firma, která má „oddělení data miningu“, jehož hlavním cílem je však pouze tvorba grafů z historických dat za účelem nalézt obecné trendy. Aby to nebylo příliš jednoduché, za pokročilejší prediktivní modely této velké nejmenované korporace je zodpovědné „oddělení pokročilé analytiky.“ Další termíny, které organizace používají, jsou například: prediktivní analytika, prediktivní modelování a strojové učení.

Data mining se tedy nachází na soutoku oborů statistiky a strojového učení (také známo pod pojmem umělá inteligence). Řada technik pro analýzu dat a vytváření modelů již dlouho existuje ve světě statistiky: lineární regrese, logistická regrese, diskriminační analýza a analýza hlavních komponent. Avšak základní principy tradičních statistiky (výpočty jsou složité a data nedostačující) se nevztahují na data mining, kde je pro potřeby modelování přebytek dat i výpočetních zdrojů. Další zásadní rozdíl mezi statistikou a strojovým učením je, že statistika se zaměřuje na vytváření závěrů z výběrového vzorku populace. Na druhou stranu, strojové učení se zaměřuje na predikci jednotlivých záznamů. Například: předpovídaná poptávka osoby „X“ s daným cenovým vzrůstem o 1 korunu, je 1 políčko, kdežto pro osobu „Y“ to jsou 3 políčka. Důraz, který tradiční statistika klade na vytváření závěrů (určení, zdali zajímavý výsledek mohl vzniknout v našem vzorku jen náhodou) v data miningu chybí (Shmueli, 2018).

V porovnání se statistikou, se data mining zabývá velkými soubory dat flexibilně, což znemožňuje nastavit přísné hranice pro danou otázku, kterou se zabýváme, což inference (vytváření závěrů) vyžaduje. V data miningu na rozdíl od statistiky hrozí tzv. „přeučení modelu“, které nastává ve chvíli, když model sedí tak přesně na dostupný vzorek dat, že popisuje nejen pouhé strukturální charakteristiky daných dat, ale i náhodné zvláštnosti. Použijeme-li výrazy z inženýrství, model sedí na signál ale i na šum (Shmueli, 2018).

2.2 Big data

Pojmy data mining a big data jsou spolu velmi úzce spojeny. Big data (lze se setkat i s českým překladem objemná data) jsou dosti relativní pojem. Dnešní data jsou v porovnání s minulostí skutečně objemná. Výzva, kterou big data představují, je často charakterizována čtyřmi „V“, volume (objem), velocity (rychlost), variety (rozmanitost), veracity (věrohodnost). Objemem rozumíme množství dat. Rychlost odkazuje na rychlost toku, tedy na rychlost, kterou jsou data generována a měněna. Rozmanitost znamená, že jsou generovány různé typy dat (měna, datum, čísla, text, atd.). Věrohodnost se vztahuje ke skutečnosti, že data jsou generována organicky rozšiřovanými procesy. Například miliony lidí se registrují v nejrůznějších online službách či v online obchodech pro bezplatné stahování aplikací. Taková data nejsou často transformována do vhodné podoby pro modelování a neprochází kontrolou kvality, která se vztahuje na data sbíraná za účelem statistických studií.

Jak ve své knize uvádí (Shmueli, 2018) většina velkých organizací využívá big data, protože většina rutinních firemních procesů v současnosti generuje data, která lze ukládat a případně i analyzovat. Rozměr této problematiky lze vizualizovat porovnáním dat v tradiční statistické analýze databáze řetězce Walmartu (řekněme, že máme k dispozici 15 proměnných a 5000 záznamů). Představíme-li tradiční statistické studie jako reprezentaci velikosti tečky za větou, tak zákaznická databáze Walmartu má velikost fotbalového stadionu. A to pravděpodobně ani nezahrnuje další data spojená se společností Walmart. Například data ze sociálních médií existují ve formě nestrukturovaného textu, který lze jen velmi těžko objektivizovat. Online seznamovací portál OKCupid aplikuje statistické modely na svá data za účelem předpovězení, jaké typy zpráv nejpravděpodobněji zapříčiní, že na zprávu bude odpovězeno. Norský mobilní operátor Telenor byl schopen snížit odchod zákazníků o 37 procent díky využití modelů predikujících, jaký typ zákazníků má nejvyšší náchylnost k odchodu a následným zaměřením své pozornosti právě na tyto klienty. Pojišťovací agentura Allstate ztrojnásobila přesnost

predikce odpovědnosti za úrazy způsobené automobily tím, že zapojila více informací o typu vozidla. Výše zmíněné příklady pocházejí z (Siegel, 2016).

2.3 Metodologie v data miningu

S postupem času začaly v data miningu vznikat metodologie, který byly následně všeobecně aplikovány na různé data miningové projekty. Zpočátku byly metodologie utvářeny především velkými společnostmi zabývajícími se tímto oborem. Příkladem takovéhoto postupu jsou např. metodologie SEMMA či 5A, které vytvořily společnosti SAS resp. SPSS. Tyto metodologie si byly velmi podobné a v podstatě řešily stejný problém.

2.3.1 Metodologie CRISP-DM

V roce 1996 byl evropskou komisí financován grant, který měl zajistit vytvoření jednotné univerzální metodologie pro data mining. Díky tomuto grantu Tak vznikla v roce 1999 první verze volně dostupné metodologie CRISP DM (Cross-Industry Standard Process for Data Mining). Tuto metodologii zaštiťovalo zaštiťovalo konsorcium společností SPSS, NCR, DAIMLER CHRYSLER a OHRA.

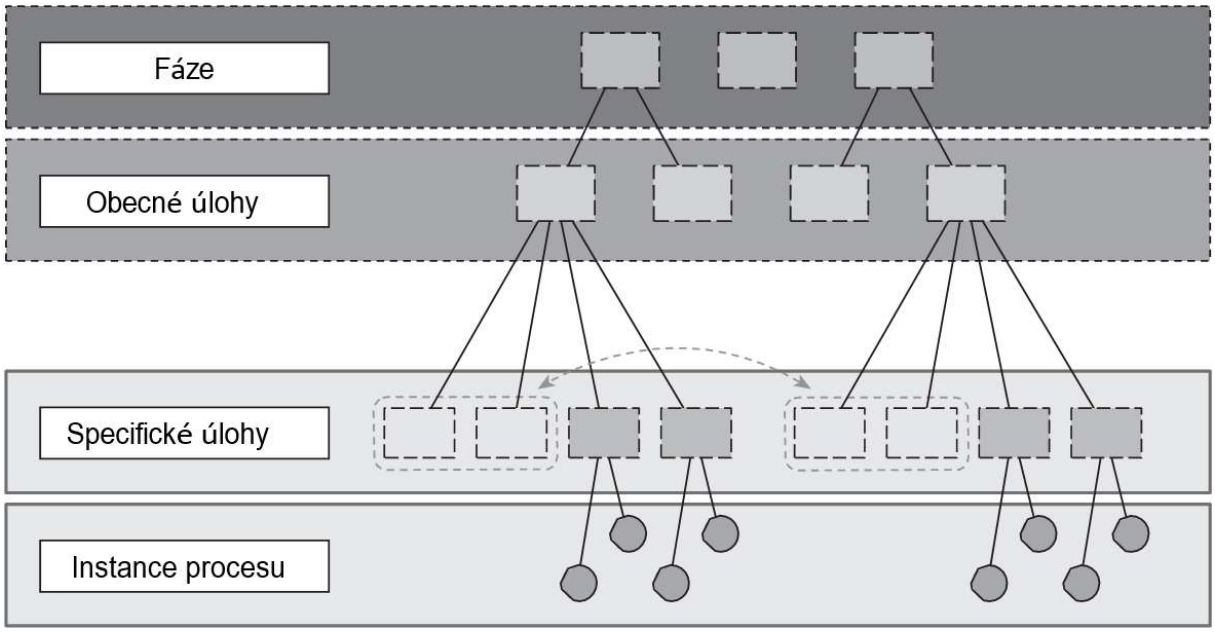
Metodologie CRISP-DM je popisována jako hierarchicky-procesový model, který se skládá ze souboru úkonů popisovaných ve čtyřech úrovních abstrakce (od obecných až po specifické): fáze, obecný úkon, specializovaný úkon a instance procesu (viz Obrázek 1).

Z hlediska nejvyšší úrovně je proces data miningu organizován do několika fází a každá fáze se skládá z několika druho-úrovňových obecných úkonů. Druhová úroveň se nazývá obecná, protože má byt dostatečně obecná, aby pokryla všechny možné situace v data miningu. Tyto obecné úkony by měly mít maximální stabilitu a celistvost. Celistvostí se myslí to, že by měly pokrýt celý proces data miningu včetně všech možných uplatnění. Stabilním se označuje takový model, který by měl být validní i pro zatím nepředpokládané vývoje, jako například nové modelovací techniky. Třetí úroveň, tzv. úroveň specializovaných úkonů, je úroveň popisující, jak by měli probíhat činnosti obecných úkonů v určitých specifických situacích. Tzn. například na druhé úrovni se může nacházet obecný úkon jménem "vyčisti data". Třetí úroveň nám říká, jak se tento úkon liší v různých situacích, například při čistění numerických hodnot oproti čistění kategorických hodnot, nebo například zdali je typ problému shlukové nebo prediktivní modelovaní (Chapman, 2000).

Popisování fází a úkonů jako diskrétních kroků vykonaných ve specifickém pořadí, představuje idealizovanou sekvenci událostí. V praxi může být mnoho úkonů vykonáno v různém pořadí a často je také třeba se opakovaně vracet k předchozím úkonům a opakovat určité činnosti. Procesový model CRISP-DM se nepokouší o zachycení všech těchto možných cest přes proces data miningu, protože by to vyžadovalo nesmírně komplexní procesový model.

Čtvrtá úroveň, tzv. instance procesu, je záznam činností, rozhodnutí a výsledků dané aktuální aplikace data miningu. Instance procesu je organizována dle úkonů definovaných v rámci vyšších úrovní, ale reprezentuje to, co se doopravdy událo v určitém nasazení, nepopisuje tedy obecné události (Chapman, 2000), (Petr, 2010).

Obrázek 1: Čtyř-úrovňový rozbor metodologie CRISP-DM



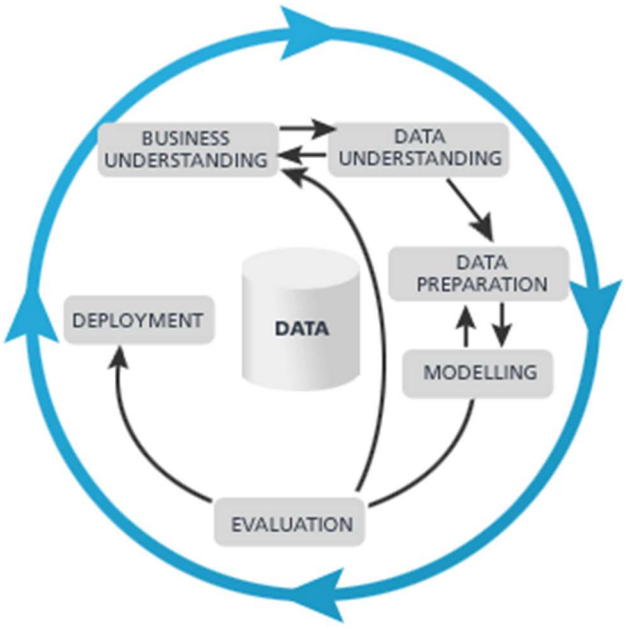
Zdroj: vlastní zpracování dle (Chapman, 2000).

Z horizontálního hlediska metodologie CRISP-DM rozlišuje mezi referenčním modelem a uživatelskou příručkou. Referenční model představuje rychlý pohled na fáze, úkony a jejich výstupy, a také popisuje co by data miningový projekt měl obsahovat. Na druhé straně, uživatelská příručka udává detailnější tipy a rady pro každou fázi a každý úkon v ní obsažený, a vyobrazuje, jak by data miningový projekt měl být vykonán.

Životní cyklus daného data miningového projektu se skládá z šesti fází zobrazených na Obrázku 2. Sekvence fází není pevně daná. Procházení a vracení se k fázím je vždy potřebné. Výsledek každé fáze určuje, která fáze nebo který úkon je nutné vykonat dále. Šipky znázorňují

ty nejdůležitější a nejčastější vztahy mezi fázemi. Vnější kruh na Obrázku 2 symbolizuje cyklickou povahu samotného data miningu. Data mining nekončí ve chvíli, kdy je nasazeno řešení. Znalosti získané při procesu a z implementovaného řešení mohou vyvolat další, často více specifické otázky ohledně řešeného problému. Další následující data miningové procesy čerpají ze zkušeností z těch dřívějších.

Obrázek 2: Schéma CRISP DM



Zdroj: (CRISP DM: Data Mining Session 2, 2013)

První dvě fáze metodologie jsou analytické. První fáze se zaměřuje na definici problému a rozplánování celého projektu. Jsou rovněž stanovena kritéria úspěšnosti, tj. očekávání přínosů celého projektu.

Ve druhé fázi se snažíme porozumět datům, provádíme jejich analýzu, dostupnost a hodnotíme jejich kvalitu. Tato fáze začíná prvotním sběrem dat a poté pokračuje aktivitami, které umožní seznámit se s těmito daty, identifikovat problémy s kvalitou dat, objevit první poznatky ohledně těchto dat a také odhalit zajímavé podsoubory pro utvoření hypotéz ohledné skrytých informací.

Ve třetí fázi *Data Preparation* se vytváří modelovací matice (tabulka), která je dále využívána pro DM modely. Jde o časově nejnáročnější fázi celého projektu. Fáze přípravy dat obsahuje všechny aktivity, které jsou potřeba ke konstrukci konečné datové matice. Úkony přípravy dat jsou většinou vykonávány několikrát za sebou a v žádném specifickém pořadí. Tyto úkony

zahrnují, výběr tabulek, záznamů a vlastností, a zároveň i transformaci a čištění dat pro modelovací nástroje.

Dále následuje modelování, což je vytváření predikčních modelů. V případě používání sofistikovaných sw řešení je nutné modely jen správně nastavit. V této fázi je vybráno několik různých modelovacích technik a následně jsou aplikovány. Jejich parametry jsou nastavovány na optimální hodnoty.

Po vytváření modelů následuje evaluace, což je kontrola a vyhodnocení celého vlastního řešení. Odhadují se budoucí přínosy, hodnotí se splnění počátečních kritérií, která byla stanovena ve fázi Business *Understanding*. V této fázi projektu již existuje vytvořený model (či modely) které, z hlediska analýzy dat, jsou vysoce kvalitní. Před tím, než se přejde ke konečnému nasazeni modelu je důležité, aby byl model důkladně zhodnocen a byly vyhodnoceny kroky, které byly vykonány při jeho tvorbě. Měli bychom se ujistit, že daný model adekvátně dosahuje cílů, které byly stanoveny na samotném počátku projektu. Klíčovým cílem je určit, zda-li existuje nějaký důležitý problém, který nebyl brán v potaz. Na konci teto fáze by mělo existovat rozhodnutí ohledně využití výsledků data miningu.

Data miningový projekt většinou nekončí vytvořením modelu. I v případě, že účelem modelu je zvýšit znalost daných dat, tyto získané znalosti budou muset být zorganizovány a prezentovány ve formě v jaké jim zákazník bude rozumět a bude je moci využít. To často zahrnuje aplikaci "živých" modelů do rozhodovacích procesů dané organizace. Tedy například personalizace webových stránek v reálnem čase nebo budovaní marketingových databází. Záleží na požadavcích, fáze nasazení může být buď jen jednoduchá generace reportů, anebo na druhou stranu, komplexní implementace opakovatelných data miningových procesů do celé firmy.

Standardizace a otevřenost metodologie CRISP DM představují značnou výhodu, a proto lze tuto metodologii považovat za nejvhodnější rámec pro zpracování data miningové části této práce. hovoří jasně pro to, aby celý projekt byl v data miningové části zpracováván metodologií CRISP DM.

2.4 Základní přístupy a typické úlohy v data miningu

2.4.1 Klasifikace a predikce

Klasifikace je nejspíše tou nejzákladnější formou datové analýzy. Klasifikace může řešit celou řadu úloh a problémů z různých oborů a oblastí života. V následujícím odstavci budou v několika větách nastíněny vybrané případy, typické pro úlohu klasifikace.

Příjemce nabídky na ni může buďto odpovědět, nebo naopak. Žadatel o půjčku ji buďto může včas splatit, splatit ji pozdě, anebo vstoupit do insolvence. Platba kreditní kartou může být standardní či podvodná. Paket dat cestující po síti může být buďto neškodný nebo představovat hrozbu. Houba může být jedovatá, či jedlá. Zákazník chce odejít, nebo je loajální. Nemocný člověk může být uzdraven, stále nemocen anebo může umřít.

Proces klasifikace se skládá ze dvou základních kroků. Prvním krokem je učení, při němž je tvořen klasifikační model schopný klasifikovat data pomocí trénovacích dat. Trénovací data představují vzorky historických dat, u kterých známe výsledek klasifikace. Druhým krokem klasifikace je použití vytvořeného modelu pro klasifikaci nových dat, která potřebujeme rozřadit do tříd.

Predikcí bývá často označován proces, při kterém se určují dodatečné, nově vygenerované či chybějící proměnné. Samotná klasifikace je někdy označována jako predikce. Do predikčních metod bývá často zařazována i regrese. Regrese je podobná klasifikaci, ale s tím rozdílem, že se snažíme předpovědět hodnotu číselné reálné proměnné.

Nejčastějšími algoritmy používanými pro predikci či klasifikaci jsou nejrůznější druhy rozhodovacích stromů. Rozhodovací stromy mohou být regresní či klasifikační v závislosti na typu cílové proměnné. V případě, že je cílová proměnná kvantitativní, používáme regresní stromy. Pro predikci kvalitativní cílové proměnné používáme klasifikační stromy. Pro klasifikaci či predikci lze použít i statistické metody jako lineární regrese či Bayesovkou klasifikaci. Pro klasifikování či predikování cílové proměnné lze však využít i Neuronové sítě, či některý z algoritmů shlukové analýzy.

2.4.2 Analýza vztahů

I když využívání asociačních pravidel nedosahuje tak dobrých výsledků při predikci jako v případě použití například rozhodovacích stromů či neuronových sítí, jsou velmi oblíbena pro svoji snadnou čitelnost. Velké databáze zákaznických transakcí představují učebnicový příklad

pro aplikaci analytiky asociací ohledně zakoupených produktů. Pomocí analytiky asociací jsou vyhodnocovány například produkty, které se k sobě hodí, a ty jsou následně nabídnuty zákazníkovi. Asociační pravidla, jsou navržena k tomu, aby našla právě takové obecné množiny asociací mezi položkami ve velkých databázích. Tato pravidla lze poté využít rozmanitými způsoby. Například samoobsluhy mohou využít tyto informace pro product placement. Asociační pravidla lze však využít i v jiných úlohách. Právě asociační pravidla hrají v navrhovaném systému včasného varování důležitou roli, a z tohoto důvodu jim bude věnována samostatná kapitola.

2.4.3 Seskupování

Myšlenka seskupování se týká takových úloh, kdy se dají jednotlivé záznamy slučovat do relativně homogenních skupin. Případy v jedné skupině jsou si podobné a případy z různých skupin se odlišují. Může se jednat například o zákazníky, výrobky, respondenty, státy či dopravní nehody, které se snažíme seskupovat pro účely systému včasného varování. Při seskupování se nejčastěji používají algoritmy shlukové analýzy. Algoritmů shlukové analýzy existuje veliké množství a jejich použití je odlišné pro různé typy úloh. Jelikož je shluková analýza velmi důležitou součástí navrhovaného řešení, bude i vybraným vhodným algoritmům věnována samostatná kapitola.

2.4.4 Analýza časových řad

Časové řady tvoří hodnoty, které jsou pozorované, zaznamenávané nebo shromažďované chronologicky v čase. Čas nemusí být samozřejmě jedinou nezávislou proměnnou, na které jsou hodnoty časové řady závislé. Časové řady mohu řešit například úlohy vývoje počtu zaměstnanců, kteří opustili zaměstnání, počty automobilů projíždějících daným úsekem, vývoj úrokové míry, vývoj množství srážek, teploty vody či výšky hladiny. Hlavní úlohou analýzy časových řad je pochopení základního mechanizmu, podle kterého jsou pozorované hodnoty generovány, a především predikce budoucích hodnot (Madsen, 2008).

2.4.5 Detekce anomálií

Anomálií nazýváme jakoukoliv odchylku od normálu. Detekování anomálií je zjišťování takových vzorů v datech, které neodpovídají očekávanému chování. Detekce anomálií nachází uplatnění v celé řadě odvětví. V bankovnictví lze pomocí detekce anomálií odhalovat bankovní podvody, v průmyslu lze pomocí anomálií detekovat chyby v produktech, či ve zdravotnictví. Způsoby detekce anomálií se odvíjí od charakteru dat. Pro detekci anomálií se častěji používají metody učení bez učitele.

2.5 Použité DM nástroje

Vytváření modelů pro systém včasného varování před vysokým rizikem dopravní nehody je možné realizovat několika cestami. Jednou z možností je využití nejrůznějších softwarových produktů, více či méně univerzálních, které jsou vhodné pro naše potřeby. Cílem této kapitoly je ukázat možnosti vybraných data miningových nástrojů pro hledání shluků dopravních nehod s podobnými vlastnostmi. Nástroje budou porovnávány vzhledem k existenci jednotlivých potřebných algoritmů, rozšiřitelnosti softwarového nástroje a možnostem vizualizace výsledků v mapovém podkladu.

Pro porovnání možností při hledání shluků nehod s podobnými vlastnostmi byly vybrány následující softwarové nástroje či utility: IBM SPSS Modeler, KNIME, RapidMiner, Weka, Orange, ELKI. Tyto nástroje byly vybrány, jelikož se jedná o nejběžnější a nejoblíbenější data miningové nástroje mezi uživateli.

Pro výzkum a vytváření prvotních modelů byl použit DM nástroj IBM SPSS modeler a open source nástroj KNIME. Postupně byly vyzkoušeny nejběžnější algoritmy shlukové analýzy, které by měly umožnit vytvořit mapu nebezpečných úseků na vozovkách.

Každý z těchto nástrojů má vzhledem k vyhledávání shluků v geografických datech a jejich následné reprezentaci pomocí mapových podkladů, své pozitivní, ale také negativní vlastnosti.

2.5.1 IBM SPSS Modeler

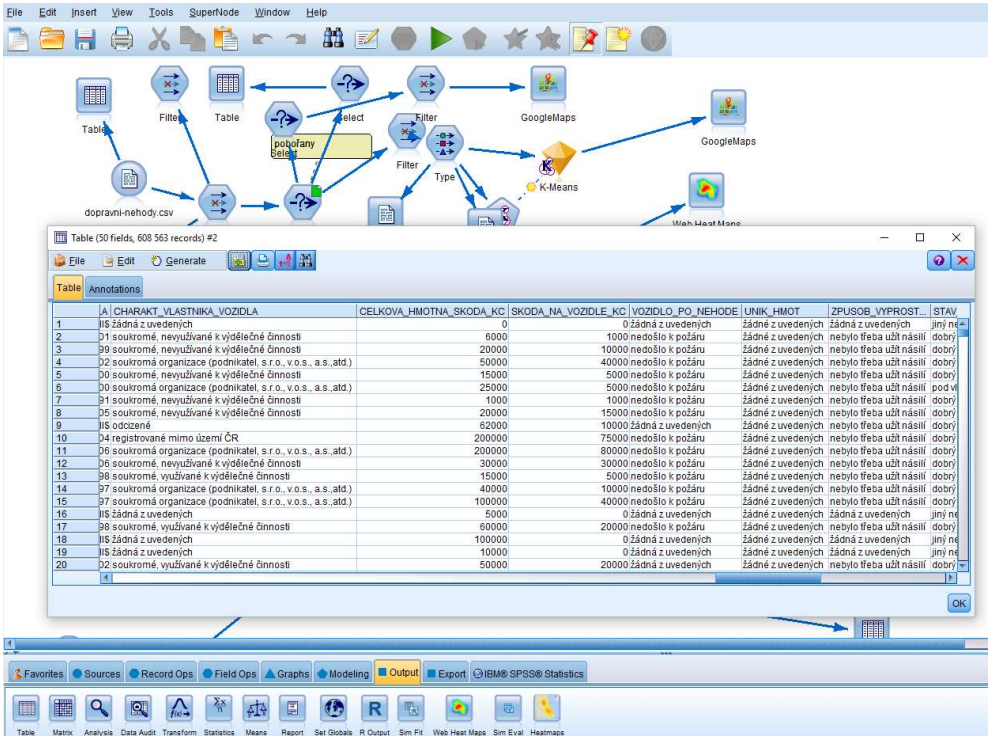
IMB SPSS Modeler (dále uváděn jako Modeler) je komerční komplexní data miningový nástroj umožňující uživateli řešit data miningové projekty nejrůznějšího charakteru (IBM SPSS Modeler, 2016). IBM SPSS Modeler (Obrázek 3) patří mezi nejznámější a nejpoužívanější komerční data miningové nástroje. Velkou předností tohoto nástroje je intuitivní ovládání, které umožňuje rychle se orientovat v datech, čímž značně urychluje fázi porozumění datům a jejich přípravu pro modelování. Modeler podporuje všechny fáze jednotné, otevřené data miningové metodologie CRISP-DM (Cross industry standart proces for data mining). Data miningový projekt v Modeleru se skládá z tzv. streamů, které jsou tvořeny uzly propojenými pomocí tzv. datových toků.

Silnou stránkou Modeleru je jeho přímočarost a uživatelská přívětivost především ve fázi přípravy dat. Modeler disponuje celou řadou uzlů určených pro načítání, přípravu, a analýzu dat. To umožnuje uživateli velmi rychle porozumět datům a velice efektivně s nimi

manipulovat. Modeler dále disponuje řadou algoritmů pro superviozované i nesupervizované učení. Samozřejmostí jsou i uzly umožňující evaluaci modelů i export výsledků. Nejnovější verze Modeleru přinášejí i možnosti stahování rozšiřujících balíčků přímo z prostředí aplikace.

S ohledem na potřeby pro dolování informací z dat o dopravních nehodách lze IMB SPSS Modeler použít s výhodou především ve fázi přípravy datové matice. V přípravě a analýze dat o dopravních nehodách Modeler vyčnívá nad ostatními vybranými open source projekty svojí přímočarostí. To, co lze v Modeleru udělat jednoduše pomocí jednoho uzlu je v případě použití open source nástroje KNIME složitější a dané kroky je nutné udělat v rámci několika uzlů. Modeler je dále možné využít pro hledání asociačních pravidel, jelikož disponuje uzly jako je Apriori, Carma a Sequence, které generování asociačních pravidel umožnují.

Obrázek 3: Prostředí IBM SPSS Modeler



Zdroj: vlastní zpracování

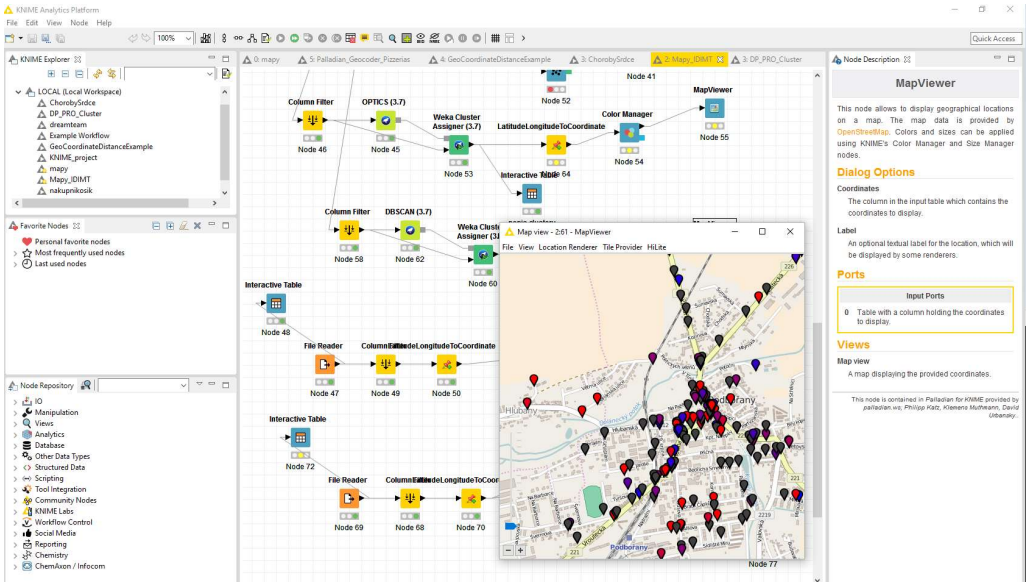
Většina modelovacích algoritmů je v modeleru tzv. Ful-proof. Modeler umí automaticky upravovat proměnné pro různé typy rozhodovacích stromů (např. kategorizace číselných proměnných). Dále Modeler disponuje celou řadou propracovaných klasifikačních algoritmů (např. CHAID, C&RT, C5.0, lineární i logistická regrese, SVM a řadou dalších). Modeler také obsahuje několik uzlů pro segmentaci (K-means, kohonen mapy, TwoStep, Anomaly). Bohužel žádný z těchto uzlů není vhodný pro hledání shluků v geografických datech. Modeler umožňuje

vytvářet vlastní uzly pomocí jazyka R či Python a je dále rozšiřitelný o další algoritmy, avšak k dispozici jich v současné chvíli není tolik jako na jiných platformách. V tuto chvíli jsou k dispozici uzly umožňující vizualizovat shluky nehod v mapě. Vizualizace výsledků v mapě by v případě Modeleru mohla být samostatnou kapitolou. V případě SPSS Modeleru verze 18 je nejprve třeba naistalovat opravný patch. Bez patche nefunguje korektně hub pomocí, kterého je možné stahovat do SPSS Modeleru další rozšiřující balíčky. Dále je nutné mít v počítači nainstalován jazyk R. Korektně by měla fungovat pro vykreslování do mapy pomocí uzlu googlemaps verze jazyka R 3.1.1. V neposlední řadě je nutné ještě nainstalovat tzv. R Essentials pro danou verzi Modeleru.

2.5.2 KNIME

Druhým vybraným nástrojem je KNIME (Obrázek 4). Jde o open source projekt, jehož využití je v data miningových úlohách rovněž široké. Stejně jako u Modeleru je data minigový projekt složen ze série tzv. uzlů a datových toků, které je spojují. I když jde v případě KNIME o vizuální programování není rozhodně vytváření streamů tak přímočaré a uživatelsky přívětivé. Slabou stránkou KNIME je fáze přípravy dat. To, co lze v SPSS Modeleru provést pomocí jednoho uzlu, je nutné v KNIME udělat pomocí několika uzlů. KNIME je velmi snadno rozšiřitelný o další uzly a má širokou uživatelskou komunitu, která vytváří velké množství rozšíření. Je však nutné podotknout, že algoritmy vytvořené komunitou uživatelů nejsou vždy zcela odladěny. Může se tedy stát, že díky neoptimalizovaným algoritmům je při velkém množství dat určitá úloha velmi časově náročná.

Obrázek 4: Prostředí KNIME



Zdroj: vlastní zpracování

Velikou nevýhodou KNIME je závislost některých rozšíření nebo algoritmů na datovém typu vstupních parametrů nebo přesném názvu určitého sloupce. Porozumění datům a jejich příprava pro modelování není zdaleka tak uživatelsky přívětivá jako v modeleru. Vlastní algoritmy lze vytvářet pomocí jazyka JAVA, Python či jazyka R. V nástroji KNIME je dostupná také celá řada algoritmů vytvořených původně pro projekt Weka.

Pro hledání shluků dopravních nehod vytvořených na základě GPS souřadnic je v KNIME k dispozici uzel DBSCAN a dva uzly (DBSCAN a OPTICS) importované z projektu Weka. Naimplementovaná varianta algoritmu DBSCAN však není optimalizována a hledání shluků ve velkých datech je časově náročné. I v KNIME nalezneme uzel pro hledání asociačních pravidel, které by měly sloužit pro hledání typických vlastnosti některých shluků.

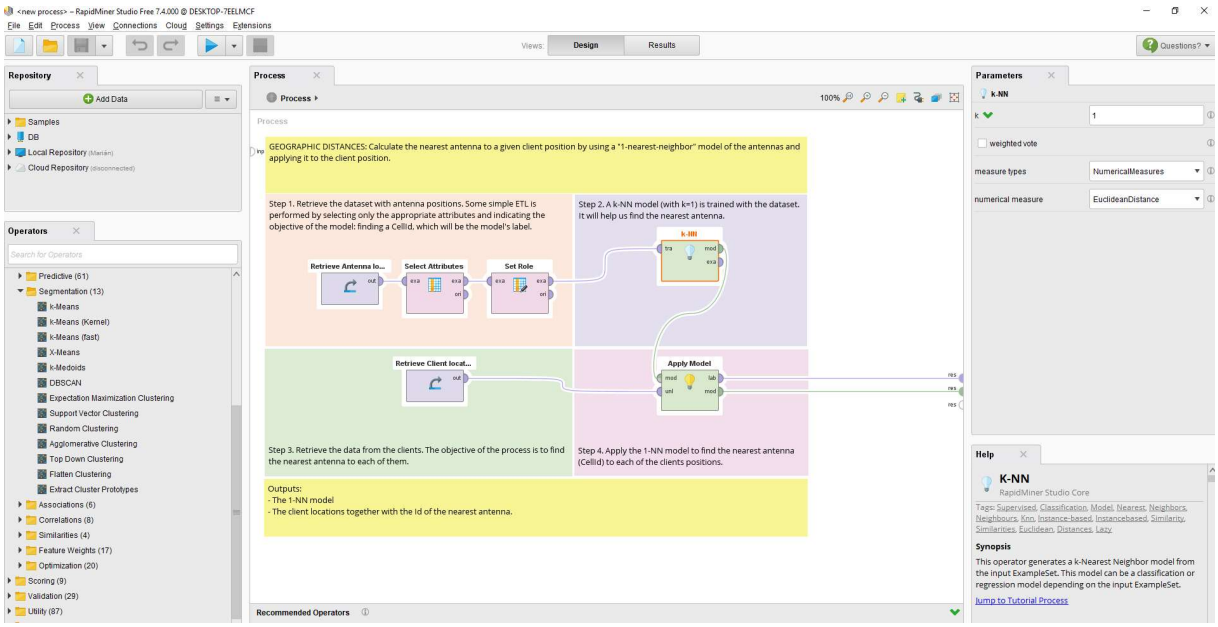
Pro vizualizaci výsledků v mapovém podkladu je nutné do KNIME naimportovat balíček Palladian, který obsahuje kromě jiných uzlů umožňujících manipulaci s geografickýmmi daty uzel s názvem MapViewer.

2.5.3 RapidMiner studio

Třetím vybraným data miningovým nástrojem je RapidMiner Studio, jehož prostředí je vidět na obrázku 5. Jde o komerční projekt, který ale existuje i ve free verzi s omezením na 10 000 záznamů (RapidMiner, 2017). Jeho uživatelské prostředí je vcelku přívětivé. Relativně povedeně je zde implementována fáze přípravy a klasické statistické analýzy dat. Projekty jsou realizovány v rámci tzv. procesu a logika práce s programem je trochu jiná než v předchozích dvou. RapidMiner disponuje celou řadou modelovacích algoritmů a uzlů pro export a import dat. Další balíčky či uzly je možné zdarma stáhnout či zakoupit v obchodě zabudovaném přímo v programu. Takto lze do RapidMineru naimportovat velké množství algoritmů z projektu Weka. RapidMiner umožnuje také vytváření skriptů pomocí jazyka R a Python.

Stejně jako KNIME, obsahuje i RapidMiner algoritmy vhodné pro hledání shluků podobných dopravních nehod. Pro vytváření shluků na základě geografické polohy je k dispozici optimalizovaný algoritmus DBSCAN. Neoptimalizovaná verze DBSCANu pro výkon je k dispozici po importu balíčku rozšiřujícího RapidMiner o algoritmy projektu Weka. I v Rapidmineru je k dispozici uzel pro generování asociačních pravidel. Vyhledávání asociačních pravidel ve shlucích vytvořených pomocí DBSCANu nelze realizovat tak jednoduše jako v KNIME či SPSS Modeleru. V tuto chvíli bohužel neexistuje žádné rozšíření pro RapidMiner umožňující vizualizovat vytvořené shluky v mapovém podkladu.

Obrázek 5: Prostředí RapidMiner Studio



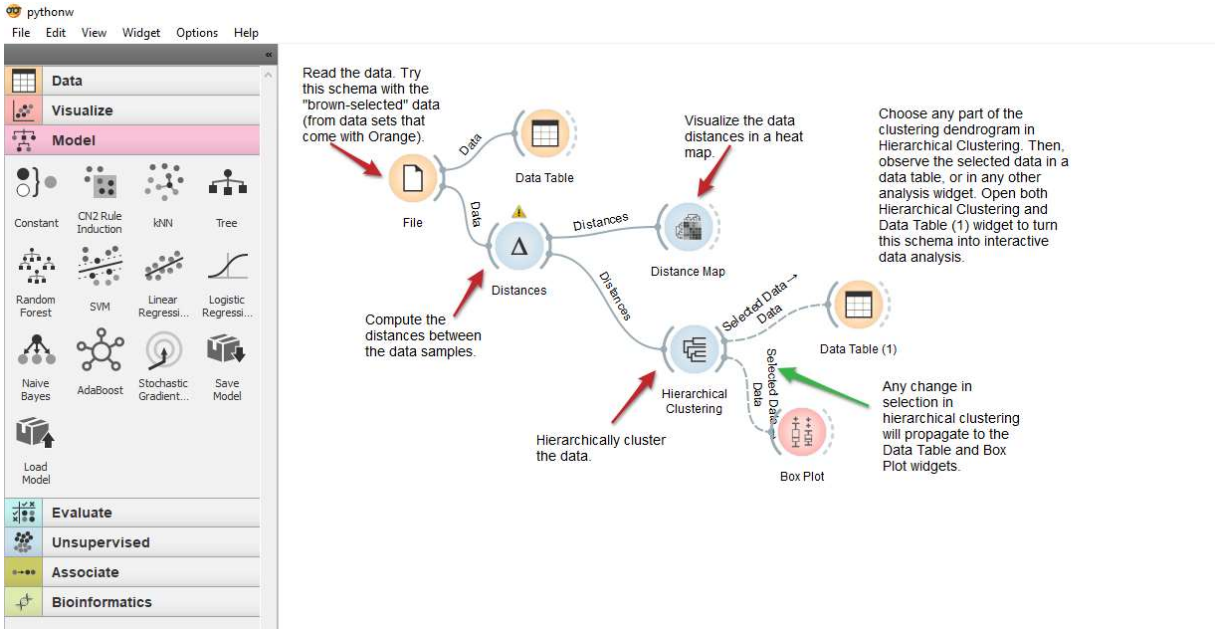
Zdroj: vlastní zpracování

2.5.4 Orange

Dalším data miningovým nástrojem, který je oblíbený zejména v bio informatice je Orange (obrázek 6). Orange je open source data miningový nástroj určený pro vizualizaci a datovou analýzu pro začátečníky i experty (Orange, 2018). V Orange je možné programovat jak vizuálně, tak i pomocí skriptů napsaných v jazyce Python. Uživatelské prostředí je přívětivé. Postup a logika při budování grafického programu je podobný jako v SPSS Modeleru či KNIME. Orange nedisponuje tolika možnostmi při přípravě dat jako SPSS Modeler. Uzlů pro vytváření prediktivních modelů zde není také mnoho. Orange je možné rozšířit pomocí dalších rozšiřujících balíčků, které je možné doinstalovat přímo z prostředí programu. Nabídka rozšiřujících balíčků pro Orange však není zdaleka tak široká jako pro KNIME či RapidMiner. Je to způsobeno především tím, že komunita vývojářů pro tento software není tak početná jako v případě již zmíněných konkurenčních programů.

Aktuálně není v Orange k dispozici žádný interní algoritmus nebo rozšiřující balíček vhodný pro hledání shluků na základě GPS souřadnic. V případě, že bychom měli již shluky vytvořené z jiného nástroje, je možné data načíst do Orange a dále hledat ve shlucích pomocí uzlu Association Rules. Uzel Association Rules je k dispozici poté, co si uživatel stáhne rozšiřující balíček s názvem Associate. Aktuálně také není možné provádět v rámci Orange žádnou vizualizaci do mapových podkladů.

Obrázek 6: Prostředí nástroje Orange



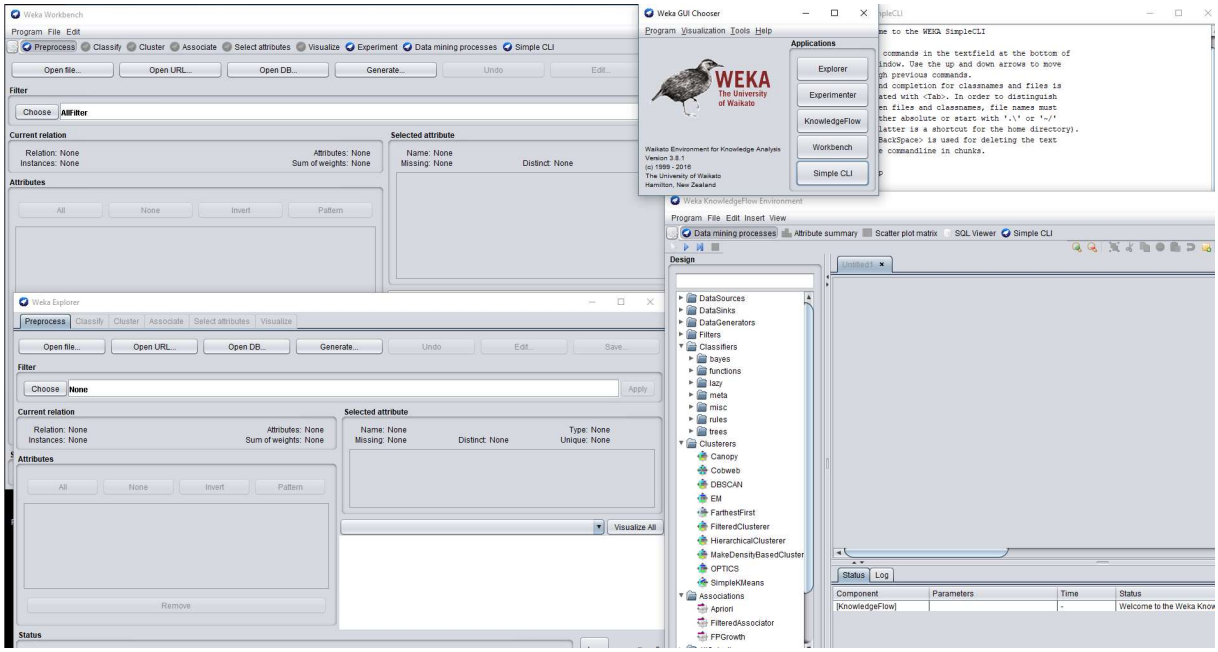
Zdroj: vlastní zpracování

2.5.5 Weka

Weka je software s otevřeným kódem vyvíjený na University of Wakaito. Weka sdružuje celou řadu algoritmů strojového učení. Na rozdíl od ostatních porovnávaných aplikací má Weka několik uživatelských rozhraní (aktuálně 5 viz obrázek 7), mezi kterými si při startu programu uživatel volí (WEKA, 2018). Jedním typem je tzv. KnowledgeFlow, což je rozhraní, které svým vzhledem a logikou připomíná rozhraní SPSS Modeleru, KNIME či Orange. Pro načítání dat Weka disponuje několika uzly umožňujícími načítat data z řady typů souborů.

V některých případech je však nastavení datasource uzlů striktní a nelze například nastavit jednoduše kódování zdrojového souboru pomocí GUI. I z hlediska přípravy datové matice nepatří Weka mezi nejlepší mezi vybranými data miningovými nástroji. Weka však disponuje širokou paletou klasifikačních, shlukovacích a dalších algoritmů. Pravdou však je, že některé algoritmy nejsou odladěny a v porovnání s jinými implementacemi stejných algoritmů jsou pomalejší. Algoritmy strojového učení lze ve Weka volat i pomocí dostupného API (Application Programming Interface). Pro grafickou analýzu dat jsou v určitých ohledech vybaveny jiné nástroje také lépe. Weka lze rozšířit o celou řadu dalších algoritmů stažením ze zabudovaného Package manageru.

Obrázek 7: Prostředí softwaru WEKA



Zdroj: vlastní zpracování

Weka s ohledem na možnosti které poskytuje pro naše účely, není jednoduše použitelná. Například již při načítání dat o dopravních nehodách ze souboru CSV nelze jednoduše zvolit kódování souboru a je nutné jej „natvrdo“ nastavit v inicializačním souboru. Uživatelsky nepříliš povedené je i načítání souboru s nehodami ve formátu MS Excelu. Soubor nebylo možné vůbec načíst při počtu sloupců větším než 2. Počet řádků byl cca 9500. Pro vyhledávání clusterů na základě GPS souřadnic obsahuje Weka algoritmy DBSCAN a OPTICS, obě verze algoritmů však nejsou optimalizovány pro rychlost výpočtů. Pro vyhledávání asociačních pravidel je ve Weka dostupný algortmus APRIORI. Vizualizace výsledků do mapového podkladu není v tuto chvíli podporována.

2.5.6 ELKI

ELKI představuje open source data miningový software napsaný v jazyce JAVA. ELKI se zaměřuje především na algoritmy učení bez učitele. Jsou zde dostupné především algoritmy využívané při shlukové analýze a při detekci anomálií. Elki se zaměřuje na vysoký výkon algoritmů a škálovatelnost (ELKI, 2018). Nástroj ELKI není v žádném případě uživatelsky přívětivý, programování neprobíhá vizuálně. ELKI není určen pro přípravu dat, analýzu dat a další vizualizace.

Pro hledání shluků na základě GPS souřadnic jsou v ELKI k dispozici algoritmy DBSCAN a OPTICS. Nelze však pracovat s celou datovou maticí a ani ji nelze modifikovat. Vyhledávání

asociačních pravidel je možné pomocí algoritmu APRIORI. Je opět ale nutné mít data připravena pomocí jiného nástroje. Vizualizace dat v mapových podkladech nepřichází vůbec v úvahu.

2.5.7 Vyhodnocení možností DM nástrojů pro využití v systému včasného varování

Analýza vybraných data miningových nástrojů ukázala, že pro potřeby vyhledávání shluků dopravních nehod s podobnými vlastnostmi není žádný z uvedených nástrojů plně vybaven. Ideální nástroj pro přípravu datové matice je dle našeho názoru IBM SPSS Modeler. Pro vyhledávání shluků na základě geografických dat jsou důležité optimalizované algoritmy jako DBSCAN či OPTICS. Ty jsou k dispozici pouze v nástroji ELKI, který ale nelze použít na žádnou další námi vyžadovanou činnost. Optimalizovanou verzi DBSCANu lze nalézt ještě v nástroji RapidMiner studio. Neoptimalizované verze se nacházejí v softwaru Weka a KNIME. Algoritmy generující asociační pravidla lze nalézt ve všech nástrojích. Vizualizace dat do mapových podkladů je snadno realizovatelná v nástroji KNIME a při doinstalování potřebného balíčku je dostupná i v IBM SPSS Modeler. Většina vybraných nástrojů je dále rozšiřitelná o další balíčky a algoritmy. Všechny aplikace umožňují vytváření vlastních skriptů ve vybraných jazycích jako je JAVA, jazyk R či Python. Hodně možností při psaní vlastních skriptů poskytuje KNIME. Shrnující porovnání jednotlivých nástrojů z hlediska jednotlivých potřeb popsaných v této kapitole je znázorněno v tabulce 1. Každý nástroj byl bodově ohodnocen pomocí sedmi kritérií. Každému kritériu byla přidělena určitá bodová maximální hranice dle jeho důležitosti

Jelikož v žádném nástroji nelze proces detekce shluků podobných vlastností zcela automatizovat, je nejlepší cestou implementace vlastního uzlu. I když například nástroj KNIME obsahuje jak uzly pro detekci shluků v geo datech a uzel pro hledání asociačních pravidel, nelze v prostředí KNIME automaticky a hromadně ve vyhledaných shlucích následně hledat asociační pravidla. Nejschůdnějším řešením je implementace vlastního upraveného algoritmu APRIORI do aplikace KNIME. Řešení, které za tímto účelem vzniklo, může využívat pro vyhledávání shluků vestavěné algoritmy DBSCAN a OPTICS a následně automaticky prohledávat všechny shluky a detekovat asociační pravidla v jednotlivých shlucích.

Tabulka 1: Porovnání DM nástrojů z hlediska potřeb systému včasného varování

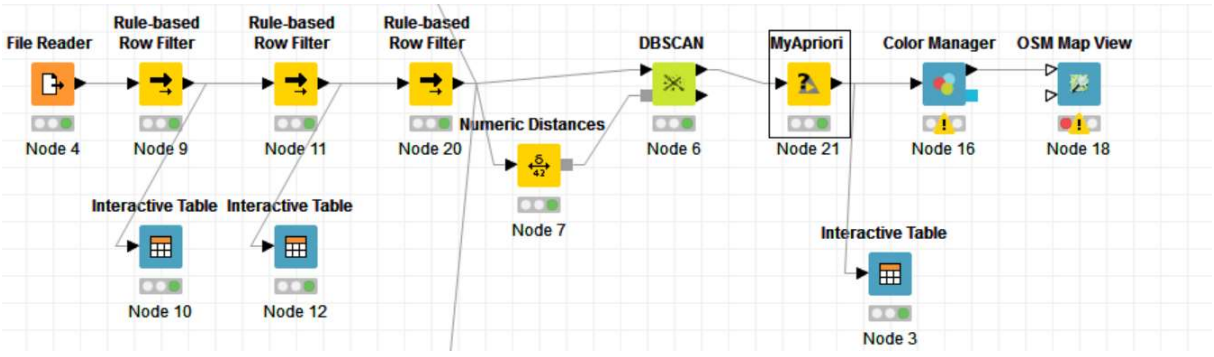
Kritérium (body)	IBM SPSS Modeler	KNIME	Rapid Miner	Orange	Weka	ELKI
User experience (0-3)	3	1	2	2	3	0
Možnosti přípravy dat (0-5)	5	3	3	2	2	0
Dostupnost a odladěnost algoritmů shlukové analýzy vhodných pro geospatial data (0-5)	0	3	4	0	3	5
Dostupnost a odladěnost algoritmu APRIORI (0-5)	5	3	3	3	3	5
Existence algoritmu hromadně detekujícího asociační pravidla pro jednotlivé shluky (0-5)	0	0	0	0	0	0
Možnosti rozšiřitelnosti nástroje (0-5)	3	5	4	3	4	1
Vizualizace do mapového podkladu (0-5)	3	5	0	0	0	0
Celkem	19	20	16	10	15	11

Zdroj: vlastní zpracování

Na obrázku 8 je vidět zjednodušený Workflow z prostředí KNIME realizující detekci shluků na základě hustoty, jejich následné testování pomocí vlastního algoritmu Apriori a následnou vizualizaci pomocí uzlu OSM Map View.

Upravený uzel Apriori umožnuje kromě volby základních charakteristik asociačních pravidel i volbu proměnné, podle které se budou pravidla grupovat.

Obrázek 8: Aplikace upraveného algoritmu DBSCAN a OPTICS



Zdroj: vlastní zpracování

3 Vybrané algoritmy data miningu

V následujících podkapitolách jsou detailněji rozebrány vybrané algoritmy, které je vhodné využít v řídící části navrhovaného systému včasného varování. Pozornost je věnována nejprve generování asociačních pravidel pomocí algoritmu Apriori, poté následují kapitoly zabývající se algoritmy shlukové analýzy.

3.1 Asociační pravidla a algoritmus Apriori

Koncept systému včasného varování využívá pro detekci specifických shluků dopravních nehod asociační pravidla, a proto bude v této kapitole tato oblast získávání pravidel popisujících specifičnost shluků nehod podrobněji popsána.

V běžném životě lidé často využívají konstrukty typu IF-THEN, aniž by si to často uvědomovali. Tyto zmíněné konstrukty lze proto najít ve všech běžných programovacích jazycích. Nelze se tedy podivovat, že společně s rozhodovacími stromy patří asociační pravidla k nejčastěji používaným prostředkům pro reprezentaci znalostí. Počátkem 90. let byl termín asociační pravidla zpopularizován v souvislosti s analýzou nákupního koše. Při analýze nákupního košíku se zjišťuje, jaké zboží si zákazníci supermarketu nakupují společně. Jde tedy o hledání skrytých vzájemných vazeb mezi položkami či dráhy zboží v obchodě. Žádné zboží nemusí být v tomto případě na straně závěru pravidla. Pro vyhledávání asociačních pravidel je všeobecně nejpoužívanější algoritmus Apriori (Berka, 2005).

Asociační pravidla zapisujeme ve formě implikací ve tvaru:

$$X \leftarrow A \,\&\, B \,\&\, C$$

přičemž A, B, a C jsou předpoklady a X je závěr.

Asociační pravidla jsou oblíbená především díky své snadné čitelnosti i pro laiky.

Vyhledávání frekventovaných množin a asociačních pravidel se většinou odehrává nad tzv. transakčními daty. Transakční data či databáze má tvar dvojice identifikátoru transakce, označovaného jako TID a vektoru, který obsahuje položky dané transakce (Berka, 2005).

Příklad rozdílu transakčních a tabulkových dat, tak jak je chápe například program IBM SPSS Modeler je vidět na obrázku 9.

Obrázek 9: Transakční a tabulková data

Transakční data	
TID	**Nákup**
1	Jam
2	Mléko
3	Jam
3	Chléb
4	Jam
4	Chléb
4	Mléko

Tabulková data			
TID	**Jam**	**Chléb**	**Mléko**
1	T	F	F
2	F	F	T
3	T	T	F
4	T	T	T

Zdroj: vlastní zpracování

Z formálního hlediska lze definovat následující pojmy:

Nechť $I = \{I_1, I_2 \ldots I_m\}$ je množina položek

D je množina transakcí, přičemž každá transakce představuje množinu položek

A je množina položek, T obsahuje A právě tehdy, když A je podmnožinou T

Asociačním pravidlem nazveme implikaci $A \rightarrow B$

Množina položek je tzv. frekventovaná, když splňuje podmínku minimální podpory.

Podpora (anglicky support) vyjadřuje procentuální zastoupení transakcí, které obsahují $A \cup B$. Míru podpory můžeme tedy vyjádřit jako:

$$Podpora\ (A \rightarrow B) = P(A \cup B)$$

Spolehlivost (anglicky confidence) je míra, která určuje sílu implikace pravidla. Spolehlivost lze definovat jako:

$$Spolehlivost\ (A \rightarrow B) = P(B|A) = \frac{podpora\ (A \cup B)}{podpora(A)}$$

Spolehlivost vyjadřuje poměr transakcí obsahujících $A \rightarrow B$ k počtu transakcí obsahujících A

Silná asociační pravidla jsou taková, která splňují podmínku předem stanovené minimální podpory i spolehlivosti

K testování specifičnosti jednotlivých shluků je teoreticky možné opět využít metod shlukové analýzy, ale pro snazší interpretaci nalezených typických vlastností shluku je možné využít i algoritmy jako je Apriori. Apriori je algoritmus umožňující vyhledávání frekventovaných množin a následně tzv. asociačních pravidel. Algoritmus APRITORI byl definován v roce 1994 (Argwall,1994).

Algoritmus Apriori vytváří postupně množiny L_1, L_2…L_n tak, že pokaždé z množiny L_{k-1} vygeneruje množinu L_k. Základním kamenem algoritmu je tzv. apriori vlastnost. Apriori vlastnost lze definovat následujícím způsobem: každá neprázdná podmnožina frekventované množiny je opět frekventovanou množinou, přičemž frekventovaná množina je množina položek splňujících stanovený práh minimální podpory.

Algoritmus Apriori má dva základní kroky:

1. vygenerování množiny kandidátů s využitím Apriori vlastnosti frekventovaných množin

2. ořezání množiny kandidátů na množiny, které jsou frekventovanými množinami (tento krok vyžaduje v každém cyklu algoritmu jednou projít celou databázi)

Jednotlivé položky frekventovaných množin musí mít uspořádání (matematické či lexikografické) Tím je dosaženo efektivního porovnávání při spojování 2 frekventovaných množin v nadmnožinu.

Pseudokód algoritmu Apriori:

```
Vstupní parametry: transakční databáze D, minimální podpora min_supp
Výstupy: množina frekventovaných množin L
Metoda:
L1 = nalezni frekventovane_1_polozky(D);
for (k = 2; Lk-1 ≠0 ;; k++)
      Ck = generuj_mnoziny(Lk-1);
      foreach t ∈ Dk
            foreach c ∈ Ck
                  if c ⊆ t then c.count++;
      Lk = {c ∈Ck| c.count ≥ minsupp}
return L = ⋃_{i=1}^{k} Li

function generuj_mnoziny(Lk-1)
      foreach l1 ∈ Lk-1
```

```
            foreach l₂ ∈ Lₖ₋₁
            if (l₁[1]=l₂[1]) ∧ l₁[2]=l₂[2]) ∧ … ∧ l1[k−2]=l2[k−2])
            ∧l1[k−1]<l2[k−1])
             then
c = l1 x l2;
if obsahuje frekventovane_podmnoziny(c, Lₖ₋₁)
        then
      Ck = Ck [ {c};
return Ck;

function obsahuje_frekventovane_podmnoziny(c, Lₖ₋₁)
foreach(k – 1)-podmnozina s z c
      if s ∉ Lk−1 then
      return false;
return true;
```

Pseudokód algoritmu Apriori byl zpracován dle (Šebek, 2010).

Algoritmus Apriori tedy postupně vytváří dvoupoložkové množiny, tří položkové množiny, čtyř položkové množiny atd. Následně jsou z frekventovaných množin $l \in Lk$ generována asociační pravidla ve tvaru:

$$s \Rightarrow (l - s)$$

Všechna vygenerovaná asociační pravidla musí splňovat také následující podmínku:

$$Spolehlivost\ (\mathrm{A} \rightarrow \mathrm{B}) = \mathrm{P}(\mathrm{B}|\mathrm{A}) = \frac{podpora\ (A \cup B)}{podpora(A)}$$

3.2 Shluková analýza

Jedním ze způsobů, jak vyhledávat nehody s podobnými vlastnostmi, je shluková analýza. Jak samotný název napovídá, jde v případě této metody využívané nejen v data miningu o vytváření shluků ve vstupních datech. Shluk je možné definovat jako množinu objektů mající podobné vlastnosti.

Jak uvádí (Han, 2011), shluková analýza, nebo jednoduše řečeno shlukování, je proces rozdělování souborů data objektů (nebo poznatků) do podsouborů. Každý podsoubor je takovým shlukem, aby si byly objekty uvnitř jednoho shluku vzájemně podobné, a také aby se nepodobaly objektům v ostatních shlucích. Soubor shluků vytvořených shlukovou analýzou se může označovat jako "shlukování". V rámci tohoto kontextu mohou různé shlukovací metody generovat různé "shlukování" ze stejného souboru dat. Výše zmíněný proces rozdělování není prováděn manuálně, nýbrž shlukovacími algoritmy.

Shluková analýza je ve velké míře používána pro mnoho různých aplikaci, jako je například business intelligence, rozpoznávání útvarů v obrázcích, hledání na webu, biologie a také v zabezpečovacích systémech. V business intelligence se shlukování dá využít pro organizaci velkého počtu zákazníků do takových skupin, kde zákazníci v jednotlivých skupinách navzájem sdílí velmi podobné charakteristiky. Toto umožňuje vývoj byznys strategií zacílených na vylepšení managementu vztahů k zákazníkovi.

Představme si consultingovou společnost s velkým počtem projektů. Pro zlepšení organizace projektů je možno aplikovat shlukování, které rozdělí projekty do kategorií na základě vzájemné podobnosti, tudíž bude možné efektivně provádět audity a diagnostiku (pro vylepšení vydávání projektů a výsledků) projektů. Pro rozpoznávání obrázků může být shlukování využito jako způsob objevování shluků či "podtříd" v systémech rozpoznávání ručně psaných písmen. Řekněme, že máme soubor dat obsahující ručně napsané číslice, kde každá číslice nese popisek 1,2,3 atd. Je třeba ale vzíti v potaz, že způsob, jakým různí lidé píší stejnou číslici, se může značně lišit. Vezmeme-li například číslici 2, někteří lidé ji třeba píší tak, že levé dolní části dělají spirálu, zatímco jiní lidé ne. Můžeme tedy využít shlukování pro zjištění podtříd číslice "2", kde každá reprezentuje určitou variaci ve způsobu, jakým je číslice napsána.

Aplikace několika modelů založených na těchto podtřídách může zlepšit celkovou přesnost rozpoznávání. Shlukování se také s oblibou využívá při hledání na webu. Například hledání na základě klíčových slov často způsobuje nalezení velkého počtu výsledků (tedy stránek relevantních k našemu hledání) z důvodu existence extrémně velkého množství webových stránek. Shlukování se tedy dá využít pro organizaci výsledků hledání do skupin a prezentaci výsledků stručným a snadno využitelným způsobem. Navíc byly vyvinuty shlukovací techniky, které rozdělují dokumenty dle témat, která jsou běžně používaná při získávaní informací. Jako data miningová funkce, shluková analýza může být využita jako samostatný nástroj pro získání poznatků ohledně distribuce dat, pozorování charakteristiky každého shluku anebo na zaměření se na konkrétní soubor shluků za účelem bližší analýzy.

Alternativně může shluková analýza sloužit jako předzpracovaní pro následující algoritmy, jako je například charakterizace, selekce podsouborů atributů, anebo klasifikace, které by potom operovaly s detekovanými shluky a jejich vybranými atributy či vlastnostmi.

Jelikož shluky jsou sbírky datových objektů, které jsou si v rámci jednotlivých shluků vzájemně podobné a zároveň se nepodobají objektům v ostatních shlucích, shluk datových objektů lze tedy brát jako implicitní třídu. V tomto ohledu se poté občas shlukování označuje jako

"automatická klasifikace". Zásadní rozdíl je ten, že shlukování dokáže automaticky detekovat daná seskupení, což je značná výhoda shlukové analýzy. Shlukování se někdy též označuje jako "segmentace dat" v rámci některých aplikací, protože shlukování dokáže rozdělit velké soubory dat do skupin na základě jejich podobnosti. Shlukování lze také využít pro detekci anomálií, v případech, kde výkyvy (hodnoty které jsou "daleko" od kteréhokoliv shluku) mohou pro nás být zajímavější než běžné případy. Možné aplikace pro detekci anomálií zahrnují např. detekci podvodů spojených s kreditními kartami a sledování trestné činnosti v internetovém prodeji, tedy výjimečné případy transakcí kreditními kartami jako jsou třeba velmi drahé a nepravidelné platby.

Obor shlukování dat je velmi intenzivně vyvíjen, výzkumné obory, které se na tomto vývoji podílejí, zahrnují kupříkladu data mining, statistiku, strojové učení, technologii prostorových databází, získávání informací, hledání na webu, biologii, marketing a spousta dalších oblastí, kde jej lze aplikovat. V poslední době se ze shlukové analýzy stalo žhavé téma v oblasti výzkumu data miningu, a to díky obrovskému množství dat sesbíraných v různých databázích. Jako odvětví statistiky byla shluková analytika podrobně prostudována, a to s hlavním zaměřením na shlukovou analýzu založenou na vzdálenosti. Nástroje pro shlukovou analýzu založené na metodě k-means, metod k-medoidů a mnoha dalších metod, byly zabudovány do mnoha software balíčků a systémů pro statistickou analytiku jako například, SPSS, KNIME, SAS a dalších. V rámci strojového učení je klasifikace známá jako případ učení s učitelem, jelikož jsou vždy udány popisné informace, tzn. učící algoritmus je řízen tím, že je mu dodáno členství třídy pro každou tréninkovou dvojici. Shlukování je spíše známé jako učení bez učitele, ve chvíli, kdy popisné informace nejsou udány.

V rámci data miningu bylo zaměřeno úsilí na hledání metod pro výkonné a efektivní metody shlukové analýzy ve velkých databázích. Současná aktivní témata výzkumu se zaměřují na škálovatelnost metod shlukování, efektivitu metod shlukování komplexních tvarů (např. nekonvexní) a typy dat (např. text, grafy a obrázky), vysoko-dimenzionální shlukovací techniky (např. shlukování objektů s tisíci vlastnostmi), a metody pro shlukování smíšených číselných a nominálních dat ve velkých databázích.

Většina algoritmů shlukové analýzy se velice často snaží reprezentovat vlastnosti objektů (atributy) pomocí čísel, a proto je nutné kategoriální a dichotomické atributy transformovat vhodným způsobem na čísla. Tento převod je nutné provést u většiny atributů dat získaných od ministerstva dopravy, jelikož většina atributů jsou kategoriální data. Algoritmy shlukové

analýzy, kterými disponuje software IBM SPSS Modeler, již v sobě tyto metody obsahují. Naopak KNIME většinou obsahuje jen algoritmy bez předzpracování dat.

Atributy vstupující do shlukové analýzy by měly být také standardizované tak, aby například atribut s většími hodnotami lehce nedominoval nad atributy ostatními. Taková dominance některých atributů nad ostatními by výrazně zkreslila vyhodnocování podobnosti. Postupu zabraňujícímu zmiňovaným problémům říkáme standardizace atributů a jeden ze způsobů lze popsat například vztahem:

$$y^i_j = \frac{x^i_j - \bar{x}_j}{s_j}$$

y^i_j nazýváme standardizovaný j-tý atribut i-tého objektu.

x^i_j je j-tý atribut o původní hodnotě, vztahující se k i-tému objektu

$\bar{x}_j$ je střední hodnotou j-tého atributu pro všechny objekty:

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x^i_j$$

s_j je směrodatnou odchylkou pro atribut j přes všechny objekty:

$$s_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x^i_j - \bar{x}_j\right)^2}$$

Stěžejním tématem shlukové analýzy jsou však způsoby hodnocení podobnosti atributů, způsobů vyjadřujících podobnost je velké množství a neexistuje žádný jediný ideální, hodící se na všechny typy úloh. Zmínit můžeme například hodnocení pomocí koeficientů asociace objektů či nejrůznějších metrik jako Eukleidovská, Manhatanská či Chebyshevova.

3.2.1 Obecný pohled na základní shlukovací metody

V literatuře se objevuje mnoho shlukovacích algoritmů. Je tedy složité provést přesné rozdělení shlukovacích metod, protože tyto kategorie se mohou navzájem překrývat, a jedna metoda tedy může mít vlastnosti, které patří do několika kategorií. Navzdory tomu je však užitečné prezentovat relativně organizovaný náhled na metody shlukování. Obecně řečeno, ty hlavní ze základních shlukovacích metod lze klasifikovat do následujících kategorií, které budou probrány následně. Rozdělovací metody: Je-li zadán soubor n objektů, rozdělovací metoda vytvoří k oddílu data, přičemž každý oddíl reprezentuje shluk a platí že $k \leq n$. Což znamená, že rozdělí data do k skupin takových, aby každá skupina obsahovala aspoň jeden objekt. Jinak

řečeno, rozdělovací metody provádí jedno-úrovňové oddělování s datovými soubory. Základní oddělovací metody typicky využívají exkluzivní separace shluků, tj. každý objekt musí patřit do právě jedné skupiny.

Většina rozdělovacích metod je založena na vzdálenosti. Je-li zadána proměnná k - počet shluků, které je třeba vytvořit, oddělovací metoda vytvoří prvotní rozdělení. Poté použije iterativní přemisťovací techniku, která se pokusí o zlepšení daného rozdělení přesunem objektů z jedné skupiny do druhé. Obecným kritériem pro dobré oddělení je to, že objekty ve stejném shluku musí být "blízké" nebo podobné mezi sebou navzájem, zatímco objekty v ostatních shlucích jsou "vzdálené" nebo velmi rozdílné. Existují různé typy ostatních kritérií pro posouzení kvality oddílů. Tradiční rozdělovací metody mohou být rozšířeny pro účely shlukování dílčího prostoru dat, namísto toho, aby prohledávaly celý prostor dat. Toto se hodí v případech, kde existuje mnoho atributů a dat je nedostatek.

Dosažení globální optimálnosti pří shlukování založeném na oddílech je často velmi náročné na výpočetní sílu, a případně dokonce vyžaduje vyčerpávající enumeraci všech možných oddílů. Namísto toho, většina aplikací využívá populární heuristické metody, jako jsou například hladové algoritmy, např. metoda nejbližších středů a shlukování k-medoidů, které postupně zlepšují kvalitu shlukování a přibližují se lokálnímu optimu. Tyto heuristické shlukovací metody fungují velmi dobře pří hledání kulových shluků v malých až středně velkých databázích. K nalezení shluků komplexních tvarů a pří práci s velmi velkými soubory dat, metody založené na oddělování je třeba rozšířit.

Hierarchické metody vytváří hierarchické rozklady daných souborů datových objektů. Hierarchické metody lze klasifikovat buď jako aglomerační nebo divizní, záleží na tom, jakým způsobem je jejich hierarchický rozklad zformován. Aglomerativní přístup, který je také označován jako přístup "zdola-nahoru", začíná tím, že každý objekt tvoří svou vlastní skupinu. A tedy poté postupně sjednocuje objekty či skupiny, které si jsou navzájem blízké, dokud nejsou všechny skupiny sjednoceny v jednu (nejvyšší úroveň dané hierarchie), nebo dokud není splněna terminační podmínka.

Divizní přístup, také označován jako přístup "shora-dolu", začíná tím, že všechny objekty jsou ve stejném shluku. V každé následující iteraci je poté shluk rozdělen do menších shluků, dokud není každý objekt ve svém vlastním shluku, nebo dokud není splněna terminační podmínka. Hierarchické shlukovací metody mohou být založeny buď na vzdálenosti, hustotě nebo posloupnosti. Různá rozšíření hierarchických metod pracují také se shlukováním v

podprostorech. Největší nevýhodou hierarchických metod je to, že jakmile je krok (sjednocení či rozdělení) proveden, už nikdy ho nelze vzít zpět. Tento tvrdohlavý přístup je užitečný, protože vede k nižším výpočetním nárokům, díky tomu, že se nemusí zaobírat kombinatorickým počtem různých možností. Takovéto techniky nemohou opravit špatná rozhodnutí, avšak již byly navrženy metody pro zlepšení kvality hierarchického shlukování.

Většina oddělovacích metod shlukuje objekty dle vzdálenosti mezi nimi. Takové metody mohou najít pouze kulové shluky a mají potíže s odhalováním shluků libovolných tvarů. Jsou však i jiné shlukovací metody, které byly vyvinuty, aby pracovaly na základě hustoty. Jejich obecný princip je takový, že postupně rozšiřují daný shluk, dokud hustota (počet objektů či datových bodů) v daném "sousedství" neklesne pod definovanou hladinu. Například pro každý datový bod uvnitř daného shluku, "sousedství" o daném průměru musí obsahovat alespoň dané minimum bodů. Takové metody lze využít pro odfiltrování šumu či odlehlých hodnot a objevení shluků libovolného tvaru. Metody založené na hustotě dokážou rozdělit soubor objektů do několika exkluzivních shluků, nebo do hierarchie shluků. Většinou metody založené na hustotě pracují pouze s exkluzivními shluky, a ne tedy s "neostrými/měkkými" shluky.

Metody založené na mřížce diskretizují objektový prostor do diskrétního počtu buněk a formují tedy strukturu mřížky. Všechny shlukovací operace jsou tedy prováděny na dané struktuře mřížky. Hlavní výhodou tohoto přístupu je veliká rychlost výpočtů, které jsou typicky nezávislé na počtu datových objektů a závisí pouze na počtu buněk v každé ose daného diskrétního prostoru.

Využití mřížek bývá často velmi výkonné při řešení mnoha prostorových problémů v data miningu, tedy včetně shlukování. Metody založené na mřížce lze integrovat s jinými shlukovacími metodami, jako například metody založené na hustotě a hierarchické metody. Některé shlukovací algoritmy integrují principy z několika shlukovacích metod najednou, takže je někdy složité klasifikovat daný algoritmus jako exkluzivně patřící pouze do jedné kategorie shlukovacích metod. Některé aplikace mohou míti kritéria pro shlukování, které vyžadují integraci několika shlukovacích technik najednou (Řezanková, 209), (Han, 2011).

Přehled shlukovacích metod

Rozdělovací metody:

- Nacházejí vzájemně vylučující se kulové shluky
- Principiálně jsou založeny na základě vzájemné vzdálenosti bodů
- Mohou používat nejbližší středy nebo medoidy pro reprezentaci středu shluku
- Jsou efektivní v případech malých až středně velkých souborů dat

Hierarchické metody:

- Shlukování je provedeno přes hierarchický rozklad (několika-úrovňový)
- Nedokáží opravit chybné sloučení či rozdělení
- Mohou zapojovat další techniky jako např. mikro-shlukování či brát v potaz "propojení" objektů

Metody založené na hustotě:

- Dokáží nalézt shluky libovolných tvarů
- Shluky jsou oblasti s velkou hustotou objektů v prostoru, které jsou odděleny oblastmi o nízké hustotě (šumem)
- Hustota shluku: Každý bod musí mít ve svém "sousedství" alespoň dané minimum ostatních bodů
- Mohou odfiltrovat šum

Mřížkové metody:

- Využívají mřížkové struktury dat o několika rozlišeních
- Jsou pro ně typické rychlé výpočty (typicky nezávislé na počtu datových objektů, však závislé na velikosti mřížky)

Pro hledání shluků v geografických datech budou v této práci používány také metody založené na hustotě, a proto budou v této kapitole blíže popsány algoritmy DBSCAN, OPTICS a DENCLUE. Větší pozornost bude také věnována základnímu rozdělovacímu algoritmu K – means, který je v praktické části práce porovnáván s algoritmy založenými na hustotě.

3.2.2 Algoritmus K-means

Algoritmus K-means patří mezi nejpoužívanější algoritmy shlukové analýzy a je implementován prakticky ve všech data miningových nástrojích. K-means metoda představuje techniku vycházející z centroidového rozdělení objektů. K-means využívá centroid clusteru C_i, jako reprezentaci takového klastru, kdy centroid představuje jeho středový bod. Centroid může být definován různými způsoby, například průměr nebo medoid objektů přiřazených ke klastru. Předpokládejme, že množina dat *D*, obsahuje *n* počet objektů nacházejících se v euklidovském prostoru. Metody založené na rozdělování objektů distribuují objekty z množiny *D* do klastrů o počtu *k*, pro něž platí $C_1, \ldots, C_k$, tj. $C_i \subset D$ a $C_i \cap C_j = \emptyset$ pro $(1 \leqq i, j \leqq k)$. Vhodně zvolená funkce je následně použita k posouzení kvality rozdělení.

Rozdíl mezi objektem p ∈ Ci a ci zástupcem shluku lze změřit pomocí dist(p, ci), kde dist (x, y) je euklidovská vzdálenost mezi dvěma body x a y.

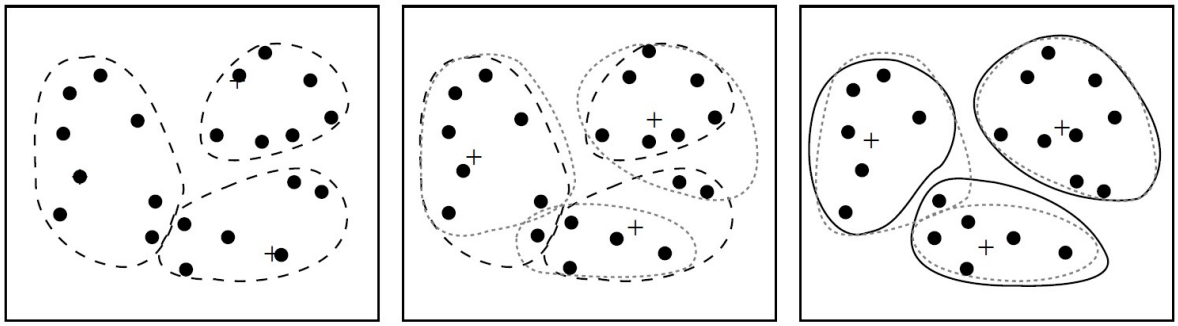
Algoritmus K-means definuje centroid klastru jako střední hodnotu bodů v klastru. Algoritmus nejprve náhodně vybere *k* počet objektů v množině *D*, z nichž každý zpočátku reprezentuje průměr nebo střed klastru. Zbývající objekty jsou přiřazeny k takovému klastru, kterému jsou nejvíce podobné na základě Euklidovské vzdálenosti mezi objektem a klastrovým průměrem. Kmenový algoritmus pak iterativně zlepšuje variaci uvnitř klastru.

Pro každý klastr je vypočítán nový průměr pomocí objektů přiřazených ke klastru v předchozí iteraci. Všechny objekty jsou následně znovu přiděleny pomocí aktualizovaných průměrů jako nové středy klastrů. Iterace pokračují, dokud není přiřazení stabilní. To znamená, že klastry vytvořené v aktuálním kole jsou stejné jako seskupení vytvořená v předchozím kole. Metoda k-means je graficky shrnuta na v následujícím pseudokódu:

Pseudokód algoritmu K-means:

```
Vstup:
k: počet klastrů,
D: soubor dat obsahující n objektů.
Výstup: Sada klastrů k.
libovolné označení k objektů z D za počáteční středy klastrů;
repeat
opětovné přiřazení objektů ke klastrům, jimž jsou objekty nejvíce
podobné na základě střední hodnoty objektů v klastru.
aktualizace středů klastrů, tj. výpočet střední hodnoty objektů pro
každý cluster;
until opakovat postup, dokud se shlukování neustálí
```

Obrázek 10: Příklad shlukování objektů pomocí algoritmu K-means



Zdroj: vlastní zpracování dle (Han, 2011)

Podle výše zmíněného algoritmu si libovolně vybíráme tři objekty jako tři počáteční centra clusterů, kde centra klastrů jsou označena znakem „+“. Každý objekt je na základě nejbližší střední hodnoty přiřazen ke klastru. Takováto distribuce vytváří siluety obklopené tečkovanými křivkami, jak je znázorněno na obrázku 10 (a). Dále jsou centra klastrů aktualizována. To znamená, že průměrná hodnota každého klastru je přepočítána na základě aktuálních objektů ve shluku. Pomocí nových středových hodnot klastrů jsou objekty opět přerozděleny. Tato redistribuce tvoří nové siluety obklopené přerušovanými křivkami (obrázek 10 (b)).

Tento proces je iterován, což vede k obrázku 10 (c). Proces iterativního opětovného přiřazení objektů ke klastrům za účelem zlepšení rozdělení je označován jako iterativní přemístění. Pokud nedojde k žádnému opětovnému přerozdělení objektů, proces končí.

Avšak u metody k-means hrozí riziko, že nebude nalezeno globální optimum. K-means mnohdy končí u nalezení lokálního optima. Výsledky mohou záviset na počátečním náhodném výběru středů klastrů. V praxi bývá zvykem nechat algoritmus běžet vícekrát s různými počátečními středy klastrů. Časová složitost algoritmu k-means je označována jako O (nkt), kde n je celkový počet objektů, k představuje počet klastrů a t počet iterací. Metoda je tudíž i relativně škálovatelná a efektivní při zpracování velkých souborů dat. Existuje několik variant metody k-means. Ty se mohou lišit výběrem počátečních středových hodnot, výpočtem rozdílnosti a strategiemi výpočtu klastrových průměrů.

3.2.3 Algoritmus DBSCAN

Jako nejvhodnější algoritmy pro hledání shluků v geografických datech se jeví algoritmy DBSCAN a OPTICS či DENCLUE patřící do skupiny algoritmů založených na hustotě. Metody založené na hustotě umožňují hledat shluky libovolného tvaru. Výhodou těchto metod je, že na rozdíl od metod založených na rozdělování umí nalézt i shluky jiného než kulovitého tvaru. Shluky algoritmů založených na hustotě jsou chápány jako oblasti s velkou hustotou a jsou od sebe odděleny šumem, tj. oblastmi s hustotou nízkou. To znamená, že některé objekty nemusí být přiřazeny do žádného shluku.

Pro vytváření shluků dopravních nehod byl jako první vybrán algoritmus DBSCAN (Density-Based Spatial Clustering of Applications with Noise), který lze použít pro jeho dobrou použitelnost v prostorových datech. Abychom mohli naznačit princip fungování algoritmu, je nejprve nutné uvést a definovat alespoň některé pojmy.

Vstupními parametry algoritmu DBSCAN je datová množina D, ve které budeme hledat shluky, velikost epsilon-okolí.

Dále definujme d(p; q), což je vzdálenost mezi body p a q. Okolí objektu p určené poloměrem epsilon lze definovat jako $N_{eps}(p) = \{q \in D \mid d(p,q) \leq \varepsilon\}$.

Jádrem nazýváme objekt p v případě, že jeho okolí obsahuje alespoň zadaný počet objektů MinPts, formálně to lze zapsat jako $|N_{\varepsilon}(p)| \geq MinPts$.

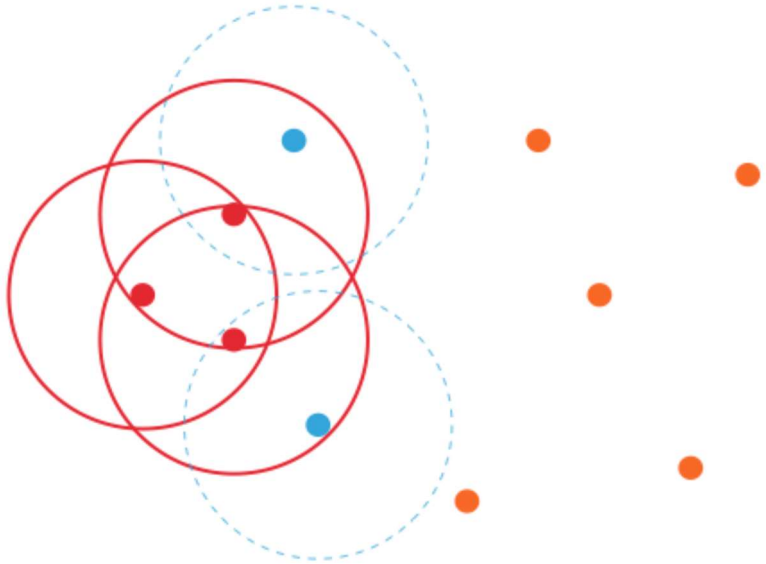
Ve chvíli, kdy řekneme o bodu p, že je přímo dosažitelný na základě hustoty z objektu q, tak to znamená, že bod p je v ϵ-okolí bodu q a současně je bod q jádrovým bodem.

Bod p je dosažitelný na základě hustoty z bodu q, jestliže můžeme říci, že existuje posloupnost x_i.. x_n taková, že bod x_i je přímo dosažitelný na základě hustoty z bodu X_{i+1}. Také musí platit podmínk, že i musí být větší nebo rovno jedné a menší nebo rovno n a jednotlivé body xi musí patřit do datové množiny D.

O bodu p můžeme říci, že je spojený na základě hustoty s bodem q v množině D, jestliže existuje bod o, který je dosažitelný na základě hustoty s body p i q.

Pro vyjasnění významu zmiňovaných pojmů je dále uveden obrázek 11. V obrázku je definováno minPts=4, epsilon je kružnoce bodu, přičemž červené body jsou označeny jako jádrové. Body, které jsou obarveny modrou barvou nejsou jádrové, ale patří do shluku. Šum je v obrázku vyznačen barvou oranžovou.

Obrázek 11: Ilustrace základních pojmů algoritmu DBSCAN



Zdroj: převzato z (Kováčik, 2016)

Metoda DBSCAN pracuje tak, že začne procházet vstupní datovou množinu a pro každý objekt v množině zkontroluje jeho ε-okolí. V případě, že ε-okolí obsahuje alespoň tolik objektů jako je zadaná hodnota MinPts, vytvoří nový shluk a následně iterativně hledá objekty, které jsou přímo dosažitelné na základě hustoty z jádra shluku. Pokud již nemůže k shluku připojit další

objekty, pokračuje v procházení datové množiny a postup opakuje pro všechny dosud nepřiřazené objekty.

Mezi silné stránky algoritmu DBSCAN patří to, že nepožaduje na začátku informaci týkající se počtu požadovaných shluků. DBSAN bere v úvahu existenci šumu, umí najít shluky libovolných tvarů, velikostí a dokáže dokonce určit i shluky, které se dokonale obklopují. Díky těmto vlastnostem dokáže nalézt shluky, které nemohou být nalezeny jinými algoritmy jako např. K-means. Pro vytváření shluků požaduje na počátku pouze dva vstupní parametry a většinou také není citlivý na pořadí bodů v databázi. (Ester, 1996), (Řezanková, 2009)

Algoritmus je bohužel značně citlivý na oba vstupní parametry ε a MinPts. Při zmenšení parametru ε je většinou třeba změnit i parametr MinPts. Geometrický tvar nalezených shluků je ovlivněn použitím různých vzdálenostních funkcí. Nejběžnější používanou metrikou je eukleidovská vzdálenost, která není příliš vhodná pro vysoce dimenzionální data, pro která se taktéž obtížně definuje i hustota. Algoritmus neoperuje příliš dobře s datovými množinami s měnící se hustotou.

Ilustrovat princip algoritmu DBSCAN je možné pomocí následujícího pseudokódu:

```
DBSCAN(D, eps, MinPts)
  C = 0
        for each unvisited point P in dataset D
        mark P as visited
        N = getNeighbors (P, eps)
        if sizeof(N) < MinPts
          mark P as NOISE
        else
          C = next cluster
          expandCluster(P, N, C, eps, MinPts)

expandCluster(P, N, C, eps, MinPts)
  add P to cluster C
  for each point P' in N
        if P' is not visited
          mark P' as visited
          N' = getNeighbors(P', eps)
          if sizeof(N') >= MinPts
                N = N joined with N'
        if P' is not yet member of any cluster
          add P' to cluster C

getNeighbors(P, eps)
  return all points within P's eps-neighborhood (including P)
```

Zdroj: („GitHub Pseudoceode“, 2011)

3.2.4 Algoritmus OPTICS

Dalším vhodným algoritmem pro vyhledávání shluků dopravních nehod je algoritmus OPTICS (Ordering Points to Identify the Clustering Structure), který je velice podobný algoritmu DBSCAN.

Aby mohl algoritmus DBSCAN detekovat optimální shluky, je mu třeba zadat vstupní parametry ε (maximální průměr kruhu sousedství) a MinPts (minimální počet bodů které jsou třeba v sousedství zásadního objektu), což ale zatěžuje uživatele odpovědností za výběr hodnot parametrů. Tento problém se vyskytuje u mnoha dalších shlukovačích algoritmů. Parametry ε a MinPts jsou většinou zadávané empiricky a jejich hodnoty se určují složitě, a to hlavně u reálných vysoce dimenzionálních dat.

Většina podobných algoritmů je velmi citlivá na hodnoty těchto parametrů – malé změny v nastavení parametrů mohou vést k velmi rozdílně vypadajícím shlukům. Navíc, reálné vysoce dimenzionální soubory dat mají velmi často nerovnoměrně rozložená data, což má za následek že jejich implicitní strukturu shlukování nelze moc dobře charakterizovat jedním souborem globálních parametrů hustoty (Řezanková, 2009).

Jak upozorňuje (Han, 2011), je třeba vzít v potaz, že shluky založené na hustotě jsou z hlediska limitní hodnoty sousedství monotónní. Což znamená, že v případě algoritmu DBSCAN se zadanou hodnotou MinPts a dvěma limitními hodnotami sousedství, $\varepsilon 1 < \varepsilon 2$, shluk C vycházející z hodnot ε1 a MinPts musí být podsoubor shluku C´ vycházející z hodnot ε2 a MinPts. To znamená, že jestliže jsou dva objekty ve shluku založeném na konkrétní hodnotě hustoty, musí být také ve shluku s nižšími požadavky na hodnotu hustoty.

Za účelem vypořádání se s problémem používání jednoho souboru globálních parametrů při shlukování, byla navržena shlukovací metoda OPTICS. Algoritmus OPTICS explicitně nevytváří shlukování souboru dat. Na místo toho je výstupem algoritmu seřazení shluků, což je lineární seznam všech objektů, které analyzujeme a reprezentuje danou strukturu shlukování založenou na hustotě. Objekty patřící do shluků s větší hustotou jsou v seznamu blíže k sobě. Toto seřazení je ekvivalentní se shlukováním založeném na hustotě, které bychom získali s využitím velké šíře nastavení parametrů.

Algoritmem OPTICS není po uživateli vyžadováno zadání specifické limitní hodnoty hustoty. Seřazený seznam se dá využít k získání základních informací o shlukování (například středy shluků, shluky libovolných tvarů), odvodit implicitní strukturu shlukování a také vizualizovat

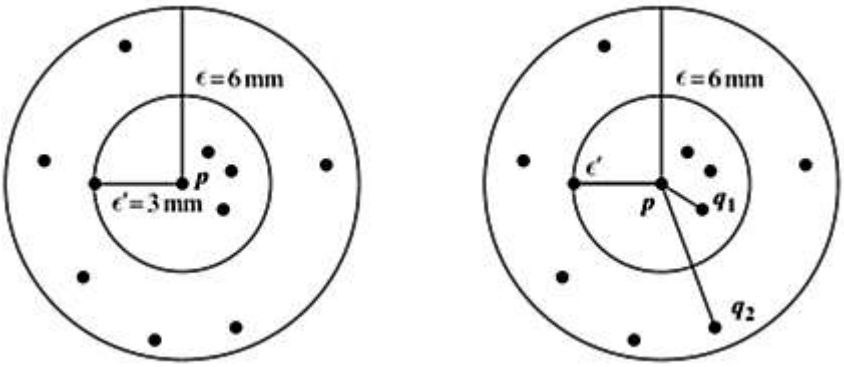
shlukování. Pro vytvoření všech možných shlukování najednou jsou objekty zpracovávány v přesně daném pořadí. Ze seznamu jsou nejprve vybrány objekty s nejmenší hodnotou ε, tudíž shluky s vyšší hustotou budou vytvořeny jako první. Vzhledem k tomuto konceptu, vyžaduje algoritmus OPTICS dvě zásadní informace pro každý objekt:

Vzdálenost od jádra objektu p je nejmenší hodnota ε′, taková, aby ε′- okolí objektu p obsahovalo alespoň MinPts objektů. Tudíž ε′ je minimální hodnota vzdálenosti taková, aby platilo, že p je jádrovým bodem. V případě, že p není jádrovým bodem vzhledem k hodnotám ε a MinPts poté je jádrová vzdálenost nedefinována.

Dosažitelná vzdálenost k objektu p z objektu q je minimální hodnota průměru kruhu, aby byl objekt p hustotně dosažitelný z objektu q. Jak nám udává definice hustotní dosažitelnosti, q musí být jádrový bod a p musí být v sousedství objektu q. Proto dosažitelná vzdálenost objektu q k objektu p je max{jádrová vzdálenost (q), vzdálenost (p,q)}. Není-li q jádrovým bodem vzhledem k ε a MinPts, potom dosažitelná vzdálenost od objektu p k objektu q není definována.

Objekt p může být přímo dosažitelný z několika jádrových bodů najednou. A proto p může mít několik dosažitelných vzdáleností vůči různým jádrovým bodům. Zvláštní zajímavostí je dosažitelná vzdálenost bodu p s nejnižší hodnotou, jelikož udává nejkratší cestu, kterou je objekt p k shluku o vysoké hustotě.

Obrázek 12: Terminologie algoritmu OPTICS



Jádrová vzdálenost p Dosažitelná vzdálenost (p, q_1) = ε′ = 3mm

Zdroj: zpracováno dle (Ankerst,1999)

Obrázek 12 zobrazuje principy jádrové vzdálenosti a dosažitelné vzdálenosti. Řekněme, že ε =6 mm a MinPts=5. Jádrová vzdálenost p je vzdálenost, ε′, od bodu p, do této vzdálenosti se vejdou

čtyři nejbližší body. Dosažitelná vzdálenost ke q_1 z p je jádrová vzdálenost objektu p (tzn. ε′= 3 mm), protože tato vzdálenost je větší než euklidiovská vzdálenost z p do q_1. Dosažitelná vzdálenost objektu q_2 vůči objektu p je euklidiovská vzdálenost z bodu p do bodu q_2. Jelikož tato vzdálenost je větší než jádrová vzdálenost bodu p.

OPTICS vypočítává způsob řazení všech objektů v dané databázi a pro každý objekt v databázi ukládá jádrovou vzdálenost a vhodnou dosažitelnou vzdálenost. Algoritmus OPTICS spravuje seznam zvaný „OrderSeeds“ za účelem vygenerování výsledného seřazení. Objekty v seznamu „OrderSeeds“ jsou seřazeny podle hodnot jejich dosažitelné vzdálenosti od jejich konkrétních nejbližších jádrových objektů, tedy podle nejmenší hodnoty dosažitelné vzdálenosti pro každý objekt.

Objekt p je objekt, se kterým algoritmus OPTICS právě pracuje v dané iteraci – na začátku si jako p zvolí náhodně ze vstupní databáze. Následně získává ε-okolí objektu p, určuje jádrovou vzdálenost a nastavuje hodnotu dosažitelné vzdálenosti na nedefinováno. Tento současný objekt p je potom zapsán na výstupu.

V případě, že p není jádrovým objektem, se algoritmus OPTICS jednoduše přesune na další objekt v seznamu „OrderSeeds“ (anebo ze vstupní databáze, je-li seznam „OrderSeeds“ prázdný). Pokud je p jádrovým objektem, tak pro každý objekt q v ε-okolí bodu p algoritmus OPTICS aktualizuje hodnotu dosažitelné vzdálenosti z bodu p a vloží objekt q do seznamu OrderSeeds, v případě že objekt q ještě zpracován nebyl. Iterace poté pokračuje, dokud nejsou zpracovány všechny vstupy a dokud seznam „OrderSeeds“ není prázdný.

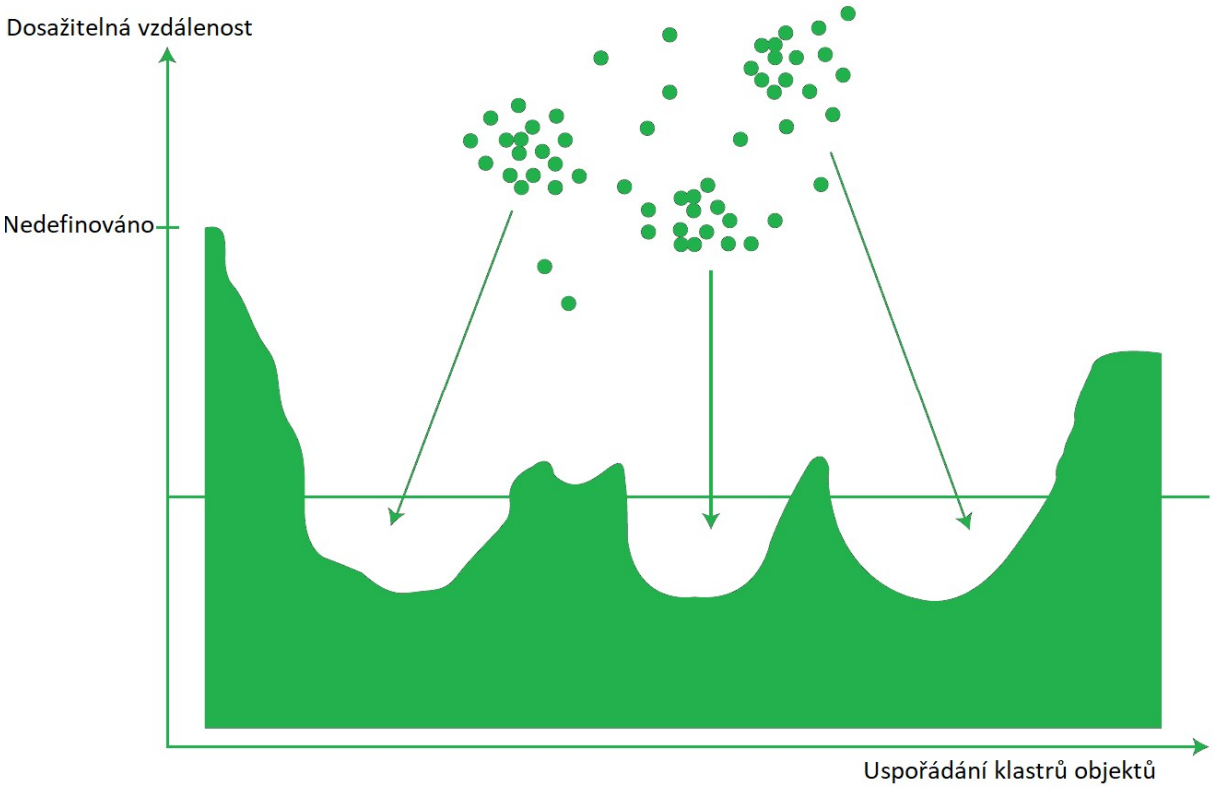
Pro lepší pochopení struktury shlukování souboru dat lze graficky vizualizovat seřazení shlukování tohoto souboru. Příkladem nám může být Obrázek 13, kde nalezneme nákres dosažitelnosti pro jednoduchý 2D datový soubor, který slouží jako obecný náhled na to, jak jsou data shlukována a strukturována.

Tyto datové objekty jsou nakresleny ve shlukovaném pořadí (vodorovná osa), společně s jejich příslušejícími hodnotami dosažitelné vzdálenosti (svislá osa). Tři Gaussovi „hrboly“ v nákresu reprezentují tři konkrétní shluky v datovém souboru. V pozdější době byly také vyvinuty metody za účelem zobrazení struktur shlukování vysoce dimenzionálních dat na různých úrovních detailnosti.

Struktura algoritmu OPTICS se velice podobá struktuře algoritmu DBSCAN. Jako následek jsou oba algoritmy stejně časově náročné. V případě že je využito prostorových indexů tak je

časová náročnost udána výrazem O(nlogn), kde n je počet objektů ke zpracování. V opačném případě je dána výrazem O(n2).

Obrázek 13: Uspořádání klastrů objektů pro algoritmus OPTICS



Zdroj: vlastní zpracování dle (Ankerst, 1999)

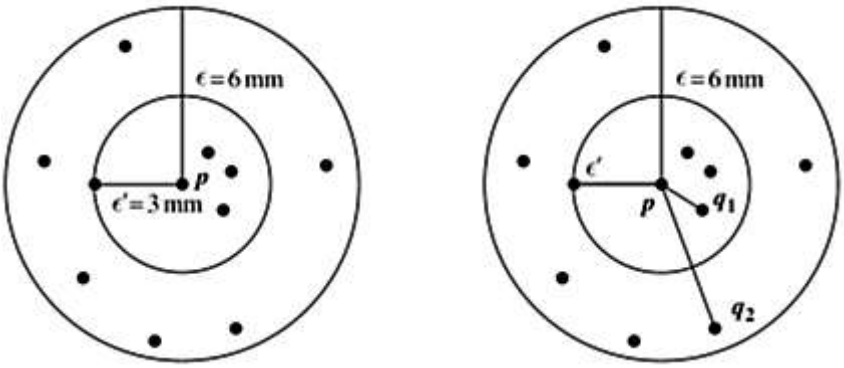
3.2.5 Algoritmus DENCLUE

Správný odhad hustoty představuje stěžejní problém všech shlukovacích metod pracujících s touto proměnou. DENCLUE (DENsity-based CLUstEring) představuje shlukovací metodu vycházející ze sady funkcí distribuce hustoty.

V pravděpodobnostní statistice odhadování hustoty spočívá v nalezení nepozorovatelné základní funkce pravděpodobné hustoty na základě souboru pozorovaných dat. V kontextu shlukování na základě hustoty tato funkce představuje skutečnou distribuci celku, tedy všech možných objektů, které analyzujeme. Námi pozorovaná data jsou tudíž považována za náhodný vzorek z tohoto celku.

Jak ve své knize upozorňuje (Han, 2011), v případě algoritmů DBSCAN a OPTICS je hustota vypočítávána na základě počtu objektů v okolí, které je definováno průměrem o velikosti parametru ε. Tyto odhady hustoty jsou často velmi citlivé na hodnotu této veličiny. Obrázek 14 demonstruje, do jaké míry, se může hustota změnit i při nepatrné úpravě sledovaného radiusu. Zvětšení poloměru sousedství mírně od ε1 do ε2 vede k mnohem vyšší hustotě.

Obrázek 14: Demonstrace jemné změny hustoty algoritmů DBSCAN a OPTICS



Zdroj: vlastní zpracování dle: (Han, 2011),

Tomuto problému lze předejít využitím neparametrického statistického postupu označovaného jako jádrový odhad hustoty (KDE - kernel density estimation). Princip stojící za touto metodou je jednoduchý. Pozorovaný objekt je považován za indikátor vysoce pravděpodobné hustoty okolí. Pravděpodobná hustota v bodě se odvíjí od vzdálenosti mezi tímto bodem a pozorovanými objekty. Za předpokladu, že x1,…, xn je nezávislým a identicky distribuovaným vzorkem náhodné proměnné f lze funkci jádrového odhadu pravděpodobné hustoty vyjádřit následovně:

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{\boldsymbol{x}-\boldsymbol{x_i}}{h}\right),$$

kde K() je jádro a h je šířka pásma sloužící jako vyhlazovací parametr. Jádro lze brát jako funkci, která modeluje vliv vzorkového bodu na jeho okolí. Technicky vzato, jádro K() je nezáporná integrovatelná funkce reálných čísel, která musí splňovat oba následující požadavky: $\int_{-\infty}^{\infty} K(u)du = 1$ a současně K(−u) = K(u) pro všechny hodnoty u. Často používaným jádrem je klasická Gaussova křivka se středovou hodnotou nula a hodnotou variance 1:

$$K\left(\frac{\boldsymbol{x}-\boldsymbol{x_i}}{h}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(\boldsymbol{x}-\boldsymbol{x_i})^2}{2h^2}}$$

Za účelem odhadu hustoty na základě daného souboru objektů určených ke shlukování algoritmus DENCLUE využívá jako své jádro Gaussovu funkci. Bod x^* je označován jako hustotní atraktor v případě, kdy je lokálním maximem odhadované funkce hustoty. Jako způsob předejití nálezu triviálních lokálních maxim, DENCLUE využívá šumovou hranici ξ a pracuje pouze z takovými atraktory x^*, pro něž platí, že $\hat{f}(x^*) \geq \xi$. Netriviální atraktory představují středy těchto shluků. Analyzované objekty jsou řazeny ke shlukům pomocí atraktorů na základě hustoty prostřednictvím krokového gradientnímu algoritmu. Při výpočtu pro objekt x gradientní algoritmus tedy začíná objektem x a je veden gradientem odhadnuté funkce hustoty. Hustotní atraktor pro x lze vyjádřit jako:

$$\boldsymbol{x^0 = x}$$

$$\boldsymbol{x^{j+1} = x^j + \delta \frac{\nabla \hat{f}(x^j)}{|\nabla \hat{f}(x^j)|}},$$

kde δ je parametr pro řízení rychlosti konvergence a

$$\nabla \hat{f}(\boldsymbol{x}) = \frac{1}{h^{d+2} n \sum_{i=1}^{n} K\left(\frac{\boldsymbol{x} - \boldsymbol{x_i}}{h}\right)(\boldsymbol{x_i} - \boldsymbol{x})}.$$

Gradientní procedura se zastaví na kroku k >0 v případě, že $\hat{f}(x^{k+1}) < \hat{f}(x^k)$, a přiřadí x k atraktoru $x^* = x^k$. Objekt x je outlinerem (výkyvem) či šumem v případě, že v průběhu gradientního algoritmu konverguje k lokálnímu maximu x^* s $\hat{f}(x^*) < \xi$.

Shluk v případě algoritmu DENCLUE představuje soubor atraktorů založených na hustotě X a soubor vstupních objektů C, přičemž platí, že každý objekt v C je přiřazen k atraktoru v X, a zároveň existuje cesta mezi každým párem atraktorů, u nichž je hustota vyšší než ξ. Díky využití několika atraktorů spojených cestami najednou může algoritmus DENCLUE nalézt shluky libovolných tvarů.

Algoritmus DENCLUE má několik výhod. Lze ho považovat za zobecnění několika dobře známých shlukovacích metod, jako jsou například jedno-spojové přístupy nebo DBSCAN. Navíc je algoritmus DENCLUE invariantní vůči šumu. Metoda KDE dokáže efektivně snížit vliv šumu tím, že rovnoměrně rozloží šum mezi jednotlivá vstupní data.

3.2.6 Algoritmus TwoStep (provedení v IBM SPSS Modeler)

Pro hledání shluků bez ohledu na geografickou polohu ale přes většinu ostatních atributů je v práci používán algoritmus TwoStep. Proto mu bude v této kapitole též věnována pozornost.

Algoritmus TwoStep se může použít k seskupování datové sady do odlišných skupin, pokud uživatel předem neví, jaké jsou tyto skupiny na začátku. Stejně jako u uzlů Kohonen a uzlů K-Means, modely TwoStep Cluster nepoužívají cílové pole. Namísto toho, aby se TwoStep pokusil předpovědět výsledek, pokusí se TwoStep odhalit vzory v sadě vstupních polí.

Data jsou seskupena tak, že záznamy v rámci skupiny nebo clusteru mají tendenci být navzájem podobné, ale záznamy v různých skupinách jsou odlišné. Algoritmus TwoStep je založen na dvoustupňové klastrovací metodě.

První krok jednou projde set dat, přičemž tyto vstupní data zkomprimuje do zvladatelného setu podklastrů. Druhý krok využívá metodu hierarchického klastrování pro postupné sloučení subclusterů do větších a větších clusterů, aniž by vyžadoval další průchod dat.

Hierarchické shlukování má tu výhodu, že nevyžaduje předběžnou volbu počtu klastrů. Mnoho hierarchických metod shlukování začíná s jednotlivými záznamy jako počáteční klastry a sloučí je rekurzivně k vytváření stále větších klastrů.

Ačkoliv takové přístupy často selhávají při větším množství dat, prvotní předklastrování TwoStepu zajišťuje rychle hierarchické shlukování i pro velké datové množiny.

Výsledný model závisí do jisté míry na pořadí tréninkových dat. Změna uspořádání dat a přestavování modelu může vést k jinému modelu konečného klastru.

Chceme-li použít model TwoStep, potřebujete jedno nebo více polí s rolí nastavenou na hodnotu „Input“. Políčka s rolí nastavenou na „Target“, „Both“ nebo „None“ jsou ignorována. Algoritmus TwoStep nepracuje s chybějícími hodnotami. Záznamy s chybějícími hodnotami pro kterékoli vstupní pole jsou při vytváření modelu ignorovány.

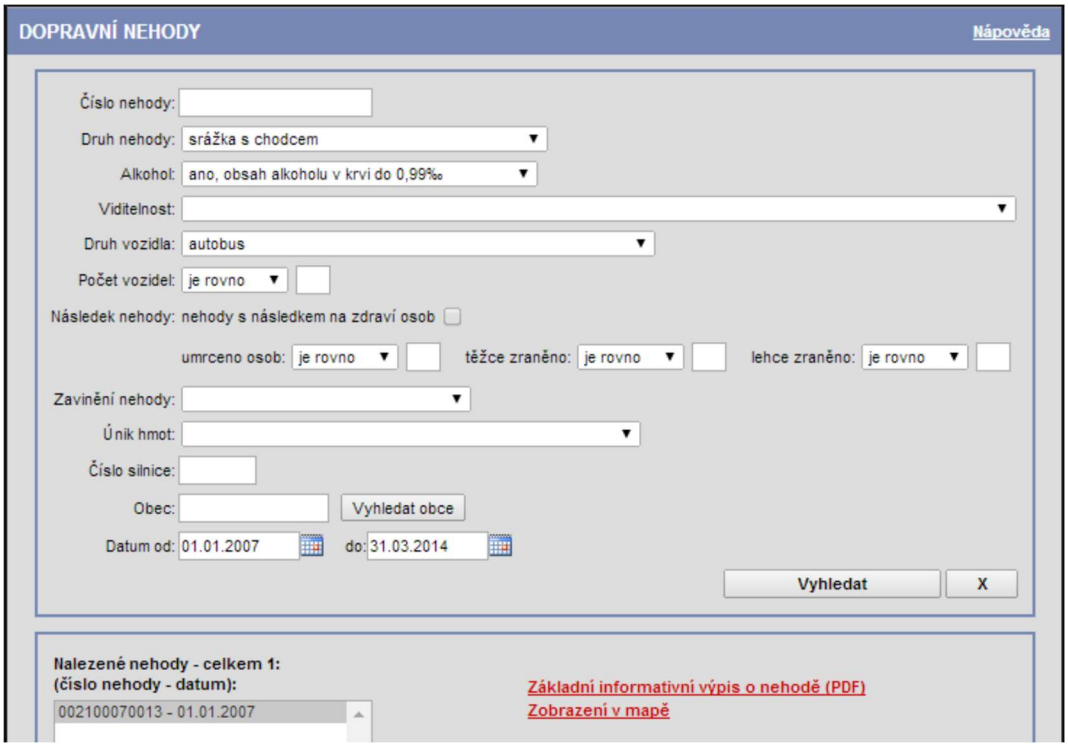
Silné stránky algoritmu TwoStep jsou, že dokáže pracovat s různými typy polí a je schopen efektivně zpracovávat velké datové sady. Má také možnost vyzkoušet několik řešení clusterů a vybrat to nejlepší, takže uživatel nemusí vědět, kolik klastrů je třeba od počátku požadovat. Uzel TwoStep může být nastaven tak, aby automaticky vyloučil outlinery nebo extrémně neobvyklé případy, které mohou nepříznivě ovlivnit výsledky modelování (“TwoStep Cluster Node”, 2012).

4 Zdroje dat o dopravních nehodách a alternativní možnosti jejich využití

4.1 Informace o dopravních nehodách

Ministerstvo dopravy poskytuje od roku 2006 v rámci projektu Jednotná dopravní mapa databázi dopravních nehod na území ČR. Aplikace (obrázek 15) je dostupná na webu *http://www.jdvm.cz*

Obrázek 15: Aplikace Policie ČR pro vyhledávání dopravních nehod



Zdroj: (Základní informativní výpis o nehodě, 2014)

Díky tomuto projektu je možné na webu nehody vyhledávat, ke každé nehodě získat detailní výpis a nechat nehodu zobrazit na mapě. Policie ČR také pravidelně vytváří statistiky nehodovosti dostupné z webu. Tyto statistiky jsou však pouze jednoduché promítnutí jednotlivých parametrů nehod do tabulek a grafů. Většinou jde o zobrazení hodnot zkoumané veličiny v daném období a porovnání hodnot s předchozím obdobím.

Jedná se například o počty nehod v jednotlivých dnech, počty nehod dle viníků a zavinění nehod, počet nehod dle druhu nehody v jednotlivých měsících, hlavní příčiny nehod, počty nehod v jednotlivých krajích v určitém období a další. Tato data jsou dostupná ve formátu pdf

a dávají základní přehled o nehodách. Neposkytují však možnost hlubšího data miningového výzkumu pro zjišťování kombinace aspektů dopravní nehody.

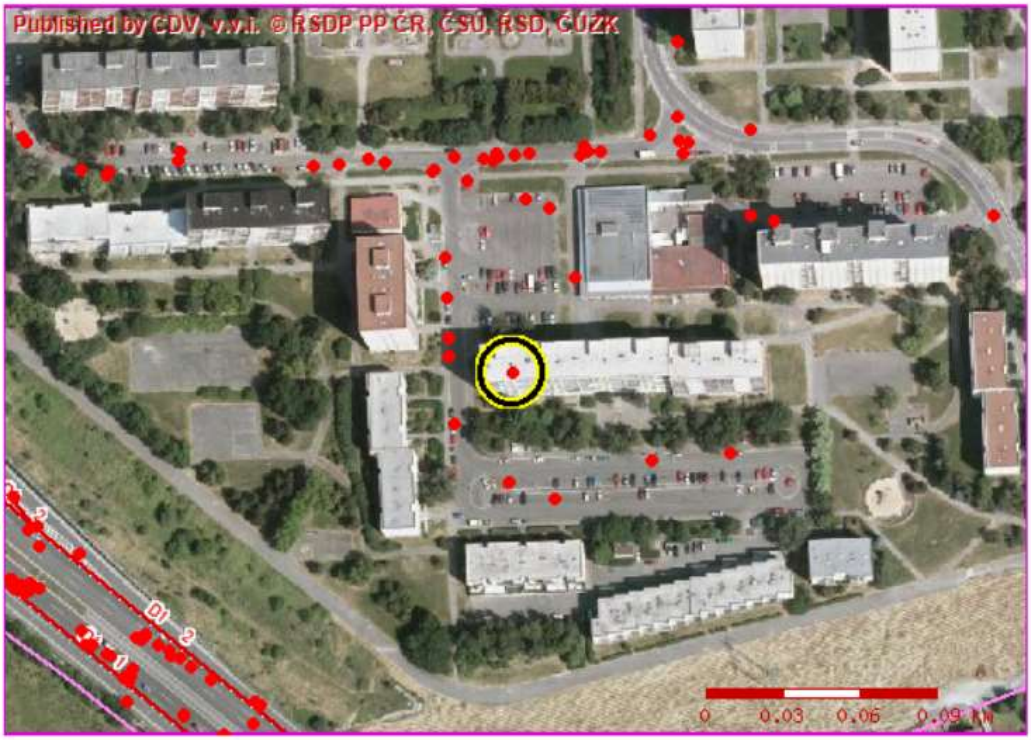
V ČR vznikl projekt opendata.cz, který by měl poskytovat nejrůznější data pro zkoumání. Nicméně je tento projekt v raném stádiu a dostupné jsou pouze údaje o veřejných zakázkách, rejstříky škol, výsledky voleb v roce 2006 a 2010 či informace o hospodaření obcí.

Data na ministerstvu dopravy jsou dostupná však také pouze ve formátu pdf. Náhledy na formát dat a dostupné atributy dopravní nehody jsou vidět na obrázcích 16 a 17.

Obrázek 16: Základní informativní výpis o nehodě 1. část

002100070013

Lokalita nehody	Praha (Hlavní město Praha)
Datum nehody	01.01.2007
Den v týdnu	pondělí
Čas nehody	19:00



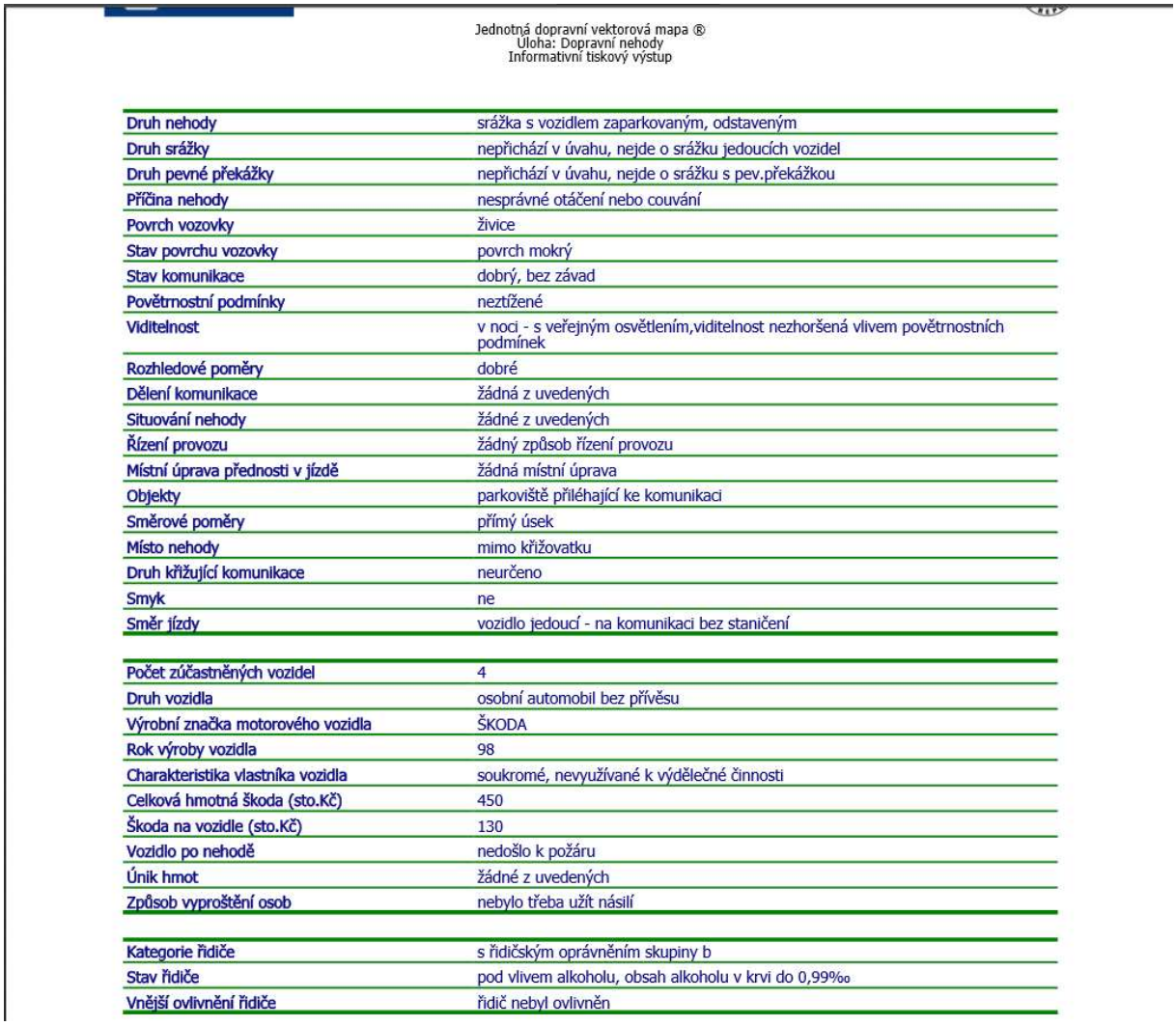
Druh pozemní komunikace	komunikace účelová - ostatní (parkoviště apod.)
Číslo pozemní komunikace	0

Zavinění nehody	řidičem motorového vozidla
Alkohol	ano, obsah alkoholu v krvi do 0,99‰
Usmrceno osob (do 24 hodin od nehody)	0
Těžce zraněno osob	0

Zdroj: (Základní informativní výpis o nehodě, 2014)

Existuje možnost, jak převést data o dopravních nehodách dostupná na stránkách ministerstva dopravy do podoby, ve které bude možné s daty dále pracovat. V ideálním případě by bylo možné spojit se s ministerstvem dopravy a policií. Tyto organizace by mohly data poskytovat v lepším formátu pro modelování, například v csv souboru. Pro modelování je k dispozici v tuto chvíli zhruba 700 000 záznamů o dopravních nehodách, přičemž každý záznam obsahuje 44 atributů. Otázkou je však jejich kvalita a možnosti při přípravě pro modelování.

Obrázek 17: Základní informativní výpis o nehodě 2. část



Jednotná dopravní vektorová mapa ®
Úloha: Dopravní nehody
Informativní tiskový výstup

Druh nehody	srážka s vozidlem zaparkovaným, odstaveným
Druh srážky	nepřichází v úvahu, nejde o srážku jedoucích vozidel
Druh pevné překážky	nepřichází v úvahu, nejde o srážku s pev.překážkou
Příčina nehody	nesprávné otáčení nebo couvání
Povrch vozovky	živice
Stav povrchu vozovky	povrch mokrý
Stav komunikace	dobrý, bez závad
Povětrnostní podmínky	neztížené
Viditelnost	v noci - s veřejným osvětlením,viditelnost nezhoršená vlivem povětrnostních podmínek
Rozhledové poměry	dobré
Dělení komunikace	žádná z uvedených
Situování nehody	žádné z uvedených
Řízení provozu	žádný způsob řízení provozu
Místní úprava přednosti v jízdě	žádná místní úprava
Objekty	parkoviště přiléhající ke komunikaci
Směrové poměry	přímý úsek
Místo nehody	mimo křižovatku
Druh křižující komunikace	neurčeno
Smyk	ne
Směr jízdy	vozidlo jedoucí - na komunikaci bez staničení

Počet zúčastněných vozidel	4
Druh vozidla	osobní automobil bez přívěsu
Výrobní značka motorového vozidla	ŠKODA
Rok výroby vozidla	98
Charakteristika vlastníka vozidla	soukromé, nevyužívané k výdělečné činnosti
Celková hmotná škoda (sto.Kč)	450
Škoda na vozidle (sto.Kč)	130
Vozidlo po nehodě	nedošlo k požáru
Únik hmot	žádné z uvedených
Způsob vyproštění osob	nebylo třeba užít násilí

Kategorie řidiče	s řidičským oprávněním skupiny b
Stav řidiče	pod vlivem alkoholu, obsah alkoholu v krvi do 0,99‰
Vnější ovlivnění řidiče	řidič nebyl ovlivněn

Zdroj: (Základní informativní výpis o nehodě, 2014)

4.2 Alternativní možnosti využití databáze dopravních nehod

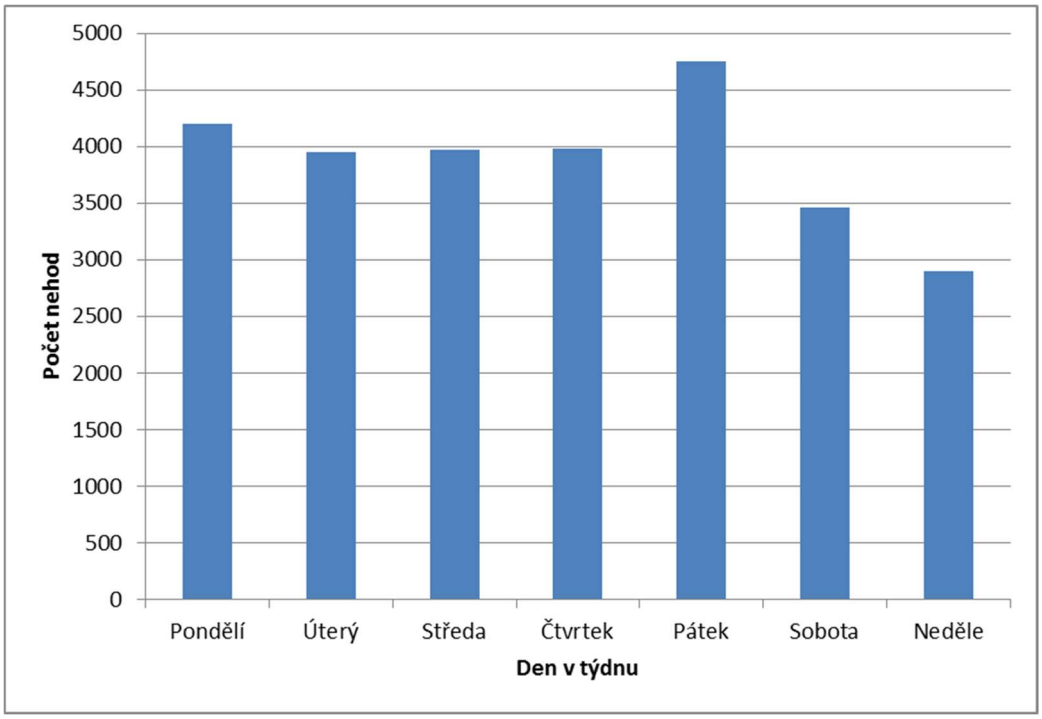
Na první pohled se nabízí možnost v těchto datech zkoumat nejčastější příčiny tragických dopravních nehod, tedy nalezení skupin typických nehod a následná lokalizace výskytu těchto nehod dle okresů a denní doby, pro navržení vhodného rozmístění preventivních policejních hlídek.

Dále se mohou pomocí data miningových technik řešit například nehody způsobené vysokou rychlostí, přičemž cílem by mohlo být vhodné rozmístění radarů měřících rychlost vozidel. Možné je také nalézt nejrizikovější křižovatky, kruhové objezdy a železniční přejezdy s ohledem na příčiny nehod.

Jako další problém k řešení se nabízí rizikovost řidičů pro pojišťovny, přičemž cílem by bylo lépe rozdělit řidiče žádající o pojištění do skupin dle jejich rizikovosti.

Jako malou ukázku z dat, která máme k dispozici, jsou na grafu 4 znázorněny počty nehod podle dne v týdnu a na grafu 5 počty smrtelných dopravních nehod podle měsíců. Oba grafy jsou pouze orientační a ve skutečnosti se čísla mohou lišit od skutečnosti, jelikož v tuto chvíli nejsou k dispozici všechna data. Grafy ukazují především převládající trendy a týkají se nehodovosti v Libereckém kraji.

Graf 4: Trend počtu nehod podle dne v týdnu (data z let 2007-2013, Liberecký kraj)

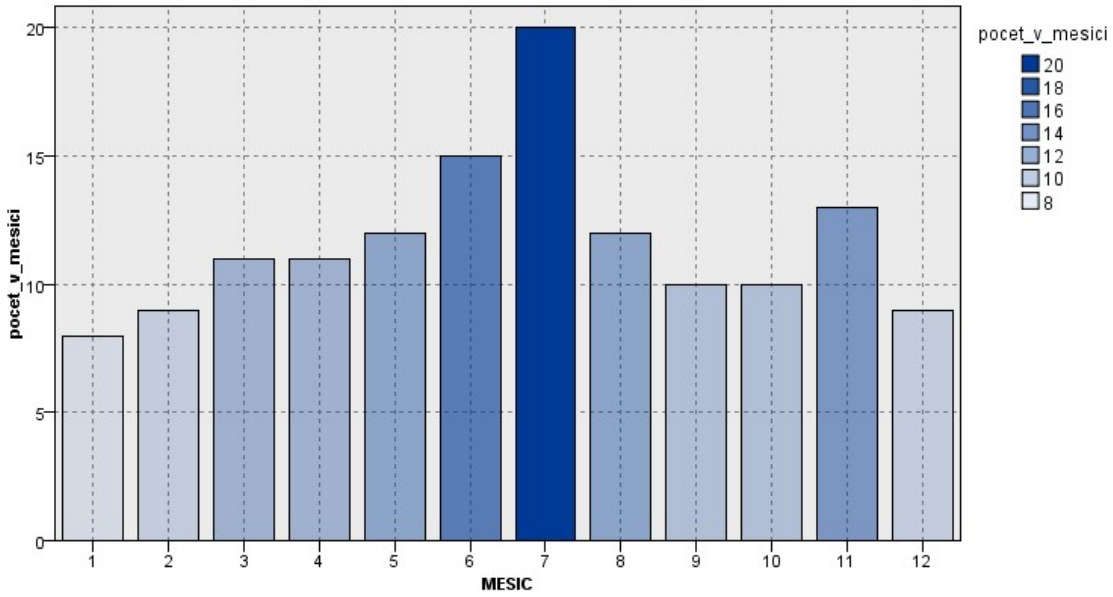


Zdroj: vlastní; data PČR

Z Grafu 4 vyplývá, že nejvíce dopravních nehod se stává v pátek a tento den je tedy z hlediska rizikovosti nejdůležitější. Druhým nejrizikovějším dnem je pondělí, ostatní všední dny jsou na tom přibližně stejně. Překvapivě nejméně dopravních nehod se naopak stává v neděli. Kdybychom dále zkoumali příčiny dopravních nehod a důvody, proč se nejméně nehod stává v neděli, došli bychom pravděpodobně k názoru, že v neděli se na vozovkách pohybují spíše osobní auta s rodinami, tedy že nehody se týkají především automobilů soukromých nevyužívaných k výdělečné činnosti.

Graf 5 ilustruje vývoj počtu smrtelných dopravních nehod v jednotlivých měsících. I když čísla opět neodpovídají přesně realitě, jasně ukazují, že nejvíce smrtelných dopravních nehod se stává s velkým náskokem v červenci. Druhým nejvíce rizikovým měsícem je červen.

Graf 5: Trend počtu smrtelných nehod podle měsíce (2007-2013, Liberecký kraj)



Zdroj: vlastní; data PČR

5 Cíle práce, metody a další směřování disertační práce

Hlavním cílem této práce je vytvořit konceptuální návrh systému umožňujícího v reálném čase a místě predikovat riziko dopravní nehody. Systém je založený na využívání predikčních modelů, které jsou vytvářeny pomocí data miningových technik a nástrojů využívajících především algoritmy shlukové analýzy a asociačních pravidel. Dílčími cíli práce jsou analýza současného stavu systémů zvyšujících bezpečnost účastníků silničního provozu, analýza využitelnosti technik dobývání znalostí z databází, určení vhodných algoritmů a postupů v souvislosti s realizací daného systému, návrh a implementace vlastní databáze dopravních nehod. Mezi další výsledky práce patří systémový návrh daného řešení, pohled na systém včasného varování z perspektivy obecné teorie systémů, a vytvoření modelů schopných identifikovat nebezpečná místa na silničních komunikacích.

5.1 Metody a fáze výzkumu

Při zpracovávání práce je vycházeno z rešerše dostupných publikací v databázích, knižních zdrojů a zdrojů dostupných online. V první fázi byla provedena rešerše stávajících systémů zvyšujících bezpečnost účastníků silničního provozu. V současné době existuje řada systémů zvyšujících bezpečnost v dopravě, ale doposud nebyla realizována myšlenka systému včasného varování popsaného v této disertační práci.

V druhé fázi bylo nutné získat data, za pomoci kterých by bylo možné postavit systém včasného varování. Jednou variantou je simulace dat užívaných pro predikci. Od této možnosti bylo prozatím upuštěno. Aby byla práce postavena na skutečných datech, byla provedena analýza současných možností pro získání a zpracování relevantních dat o dopravních nehodách. Možnosti při získávání dat o dopravních nehodách jsou popsány v kapitole 4. Následně byla prováděna rozsáhlá analýza a příprava dat pomocí data miningových nástrojů (kapitola 7).

Výběr vhodných nástrojů a algoritmů je uveden v kapitolách 2 a 3. Za tímto účelem byla provedena rešerše a využita forma analytického zhodnocení, což umožnuje výběr algoritmů a nástrojů pro vyhledávání shluků a následné dolování informací s využitím asociačních pravidel. Na základě analytického zhodnocení vhodných data miningových nástrojů jsou syntetizovány poznatky a vytvořena doporučení pro optimální řešení zkoumané problematiky.

Momentálně je vyřešen návrh tzv. řídicí a uživatelské části systému včetně popisu funkcí jejich jednotlivých částí. V budoucnosti je plánováno další rozšíření funkcí implementovaného systému. V tomto kontextu je nutné se zabývat automatizovanou distribucí dat o dopravních nehodách a dat o hustotě dopravy na jednotlivých komunikacích do systému včasného varování. V době vzniku této práce byla k dispozici pouze data o sčítání dopravy z roku 2010. Jelikož se sčítání dopravy provádí většinou jednou za 5 let, je nutné zajistit aktualizaci a analýzu těchto nových dat.

Na základě analýzy datové matice lze vytvářet doporučení například pro Policii ČR nejen z hlediska prevence, ale také doporučení pro změnu struktury dat, která policie sbírá. V době vzniku předem definované struktury sbíraných dat, kterou policie používá, nebylo počítáno s tím, že by data šla využít pro predikci rizika vzniku nehody v reálném čase a místě. V případě, že by policie realizovala tyto změny, došlo by jistě k významnému kvalitativnímu skoku v predikci rizika vzniku nehody. Zvětšení datové matice o další atributy a zpřesnění hodnot, jež atributy nabývají, by mohlo přinést také zavedení systému eCall, jehož jednotky v nově vyrobených automobilech zasílají řadu dalších dat do tísňového centra v blízkosti nehody. Výzkum se zabýval, a i nadále se bude zabývat návrhem propojení systému eCall se systémem včasného varování před zvýšeným rizikem dopravní nehody popisovaným v této práci.

V neposlední řadě byl v rámci této doktorské disertační práce vytvořen systémový návrh konceptu systému včasného varování tak, aby řešení bylo co nejuniverzálnější a dále rozšiřitelné.

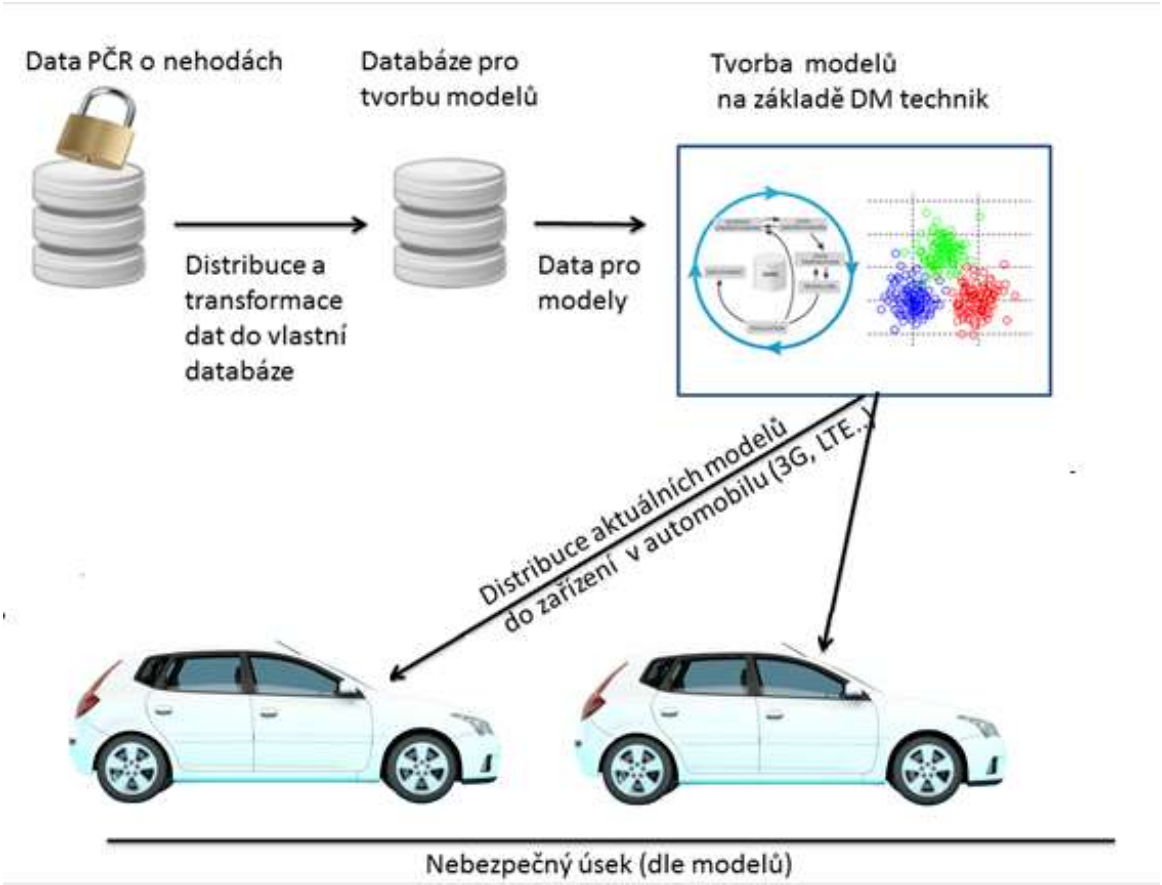
Navrhované postupy, techniky a samotný princip systému včasného varování je možné využít i na zcela odlišné typy úloh s typem dat stejného charakteru týkajících se však zcela jiné oblasti zájmu. Příkladem využití navrhovaného řešení může být vyhledávání skrytých závislostí v datech týkajících se kriminální činnosti.

Budoucí vývoj navrhovaného systému včasného varování by se měl zabývat především samotnou implementací jednotlivých bloků řídicí a uživatelské části. Z hlediska uživatelské části, je dále vyvíjena mobilní aplikace realizující samotné varování. Dále je počítáno s dalším vývojem vlastní aplikace umožňující provádět hromadnou a automatizovanou realizaci činností řídicí části systému

6 Princip systému včasného varování před dopravní nehodou

Ve zkratce lze konceptuální návrh systému včasného varování před dopravní nehodou popsat jako komplexní systém využívající vytvořených modelů předpovídajících na základě aktuální polohy a dalších atributů riziko nehody v reálném čase. Systém včasného varování před vysokým rizikem dopravní nehody se skládá ze dvou základních částí. První řídicí část zajišťuje shromažďování, zpracovávání a distribuci dat o dopravních nehodách. Druhým úkolem řídicí části je tvorba a distribuce predikčních modelů. Uživatelská část systému v reálném čase vyhodnocuje situaci a vhodným způsobem informuje řidiče o vysokém riziku dopravní nehody. Schéma vystihující nejdůležitější části systému je znázorněno na obrázku 18.

Obrázek 18: Princip systému

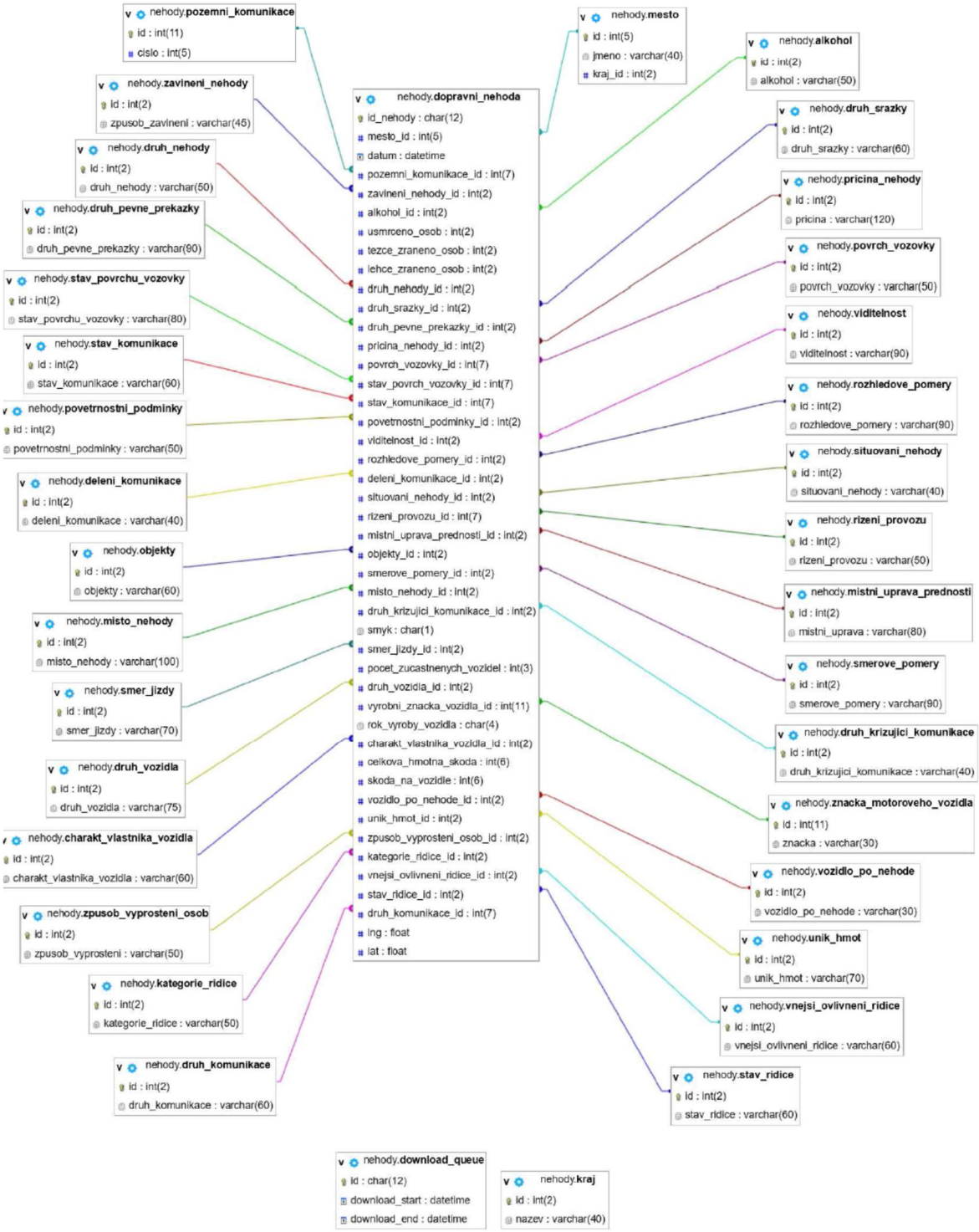


Zdroj: vlastní zpracování

6.1 Řídicí část

Řídicí část sytému slouží pro získávání a zpracovávání heterogenních dat o dopravních nehodách od Policie České republiky, resp. Ministerstva dopravy.

Obrázek 19: Konceptuální návrh vlastní databáze dopravních nehod (MySQL)



Zdroj: vlastní zpracování

Získávání dat z jednotné dopravní mapy PČR je v tuto chvíli možné pomocí skriptů napsaných v linuxového programu cURL, což je nástroj pro přenos dat po protokolech jako HTTP, FTP a dalších. Takto získaná data je třeba ještě pomocí dalších skriptů importovat do databáze pro

pozdější zpracovávání. Součástí řídicího systému je tedy databáze dopravních nehod sloužící nejen pro potřeby data miningu. Prozatím jsou data ukládána do relační databáze Mysql běžící na linuxovém stroji. Za účelem ukládání dat do databáze byl navržen vlastní konceptuální model databáze, respektující okolnosti zaznamenávané Policií ČR při dopravní nehodě. Aktuální konceptuální model databáze je zobrazen na obrázku 19. Pomocí vhodných skriptů jsou získaná data transformována do potřebné podoby a následně ukládána do databáze.

Zmiňovaná databáze slouží pro potřeby vývoje při vytváření modelů popisujících aspekty nehod. Ve druhé fázi je databáze využívána pro vytvoření komplexních modelů předpovídající riziko nehody v daném čase a místě. Hledáním skrytých závislostí v datech o dopravních nehodách a vytvářením predikčních modelů, což je nedílnou součástí řídicího systému, se zabývá kapitola 7.

6.2 Uživatelská část

Uživatelská část systému předpokládá využívání predikčních modelů ve speciálním zařízení automobilu (klient), které v reálném čase vyhodnocuje riziko nehody v závislosti na čase, aktuální poloze, stavu vozovky, stavu automobilu a počasí a dalších atributů reflektujících aktuální situaci. V dnešních automobilech je většina těchto informací běžně dostupná z čidel, kterými vozidla disponují. Zařízení v automobilu totiž porovnává aktuální situaci s výsledkem predikce a v případě vysoké míry podobnosti aktuální situace s predikcí řidiče upozorní. Řidič vozidla je upozorněn v případě, že se v daném místě za obdobných podmínek stalo více podobných dopravních nehod. V praxi se však setkáme s nebezpečnými místy, která nebudou specifická (např. časem, povětrnostními podmínkami a dalšími atributy), taková místa nazýváme jako obecné shluky a jsou vytvořeny pouze na základě častého výskytu dopravních nehod v dané lokalitě.

Po hardwarové stránce může být uživatelská část realizována velmi obdobným způsobem jako zařízení typu tablet, či mobilní telefon. Nejdůležitějším faktorem u těchto zařízení je samotný výkon mikroprocesoru a velikost paměti. Velikost paměti a výkon procesoru jsou důležitými parametry z hlediska vytváření rozhodování a varování řidiče na základě modelů. Druhým faktorem, který je důležitý především z hlediska bezpečnosti, je dostatečná velikost a čitelnost dotykové obrazovky ovládající uživatelskou část systému v automobilu. Dalším prvkem uživatelské části zvyšujícím uživatelský komfort a s tím velmi úzce související bezpečnost provozu celého systému je hlasové ovládání, které by mělo umožňovat jednoduchými příkazy

měnit základní uživatelská nastavení. Propojení zařízení realizujícího uživatelskou část systému včasného varování s automobilem by mělo být realizováno pomocí standartních komunikačních protokolů používaných pro diagnostiku vozů.

Jednou z variant implementace uživatelské části popisovaného systému včasného varování je také její zabudování přímo do infotainmentu vozidla. To by však vyžadovalo spolupráci jednotlivých výrobců automobilů. Následující podkapitola se zabývá návrhem vícezdrojové predikční mobilní aplikace, která může být také variantou realizace uživatelské části systému včasného varování.

6.2.1 Predikční vícezdrojová mobilní aplikace

Pro varování řidiče koncept popisované mobilní aplikace využívá více datových zdrojů, technologií a data miningové postupy pro vytváření predikčních modelů.

Koncept aplikace předpokládá využívání více informačních zdrojů k dosažení maximální věrohodnosti, dostupnosti a aktuálnosti informací pro řidiče. Datové zdroje a systémy distribuce informací pro koncept predikční aplikace jsou předpokládány Ecall (teoreticky i ve spojení se systémem RADIO HELP), Národní dopravní informační centrum (NDIC), RDS-TMC, a řídící část navrhovaného systému včasného varování.

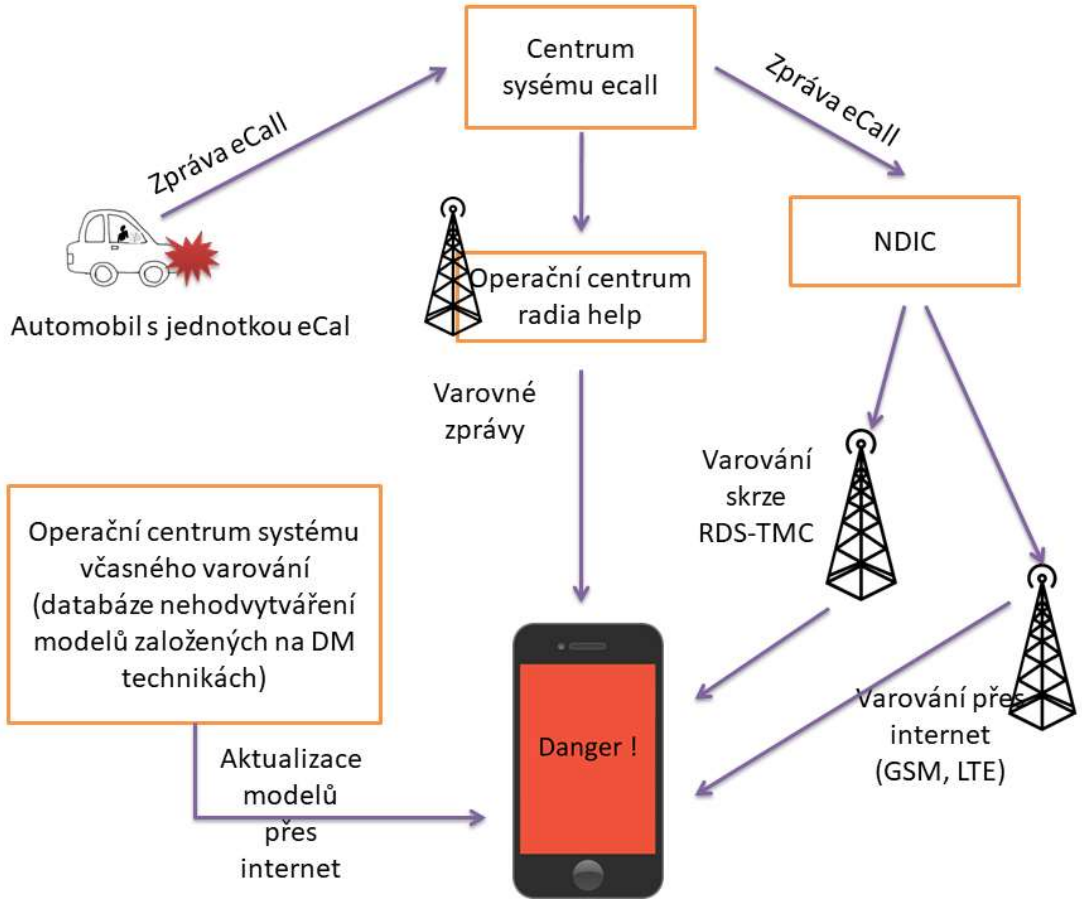
V případě nebezpečí je varování řidiče realizováno jak vizuálně na displeji telefonu, tak pomocí audio sekvencí. Použití zvukových hlášení například pomocí syntetizovaného hlasu umožňuje řidiči maximálně se věnovat řízení vozu, aniž by byl nucen odvrátit svoji pozornost od sledování situace na silnici. Pro hlasové ovládání aplikace by mělo být možné využít systém rozpoznávání hlasu, který je k dispozici ve většině telefonů s operačním systémem android.

Aplikace by měla umožnovat varovat řidiče nejen před existující nebezpečnou situací na trase (nehoda na silnici, zácpa atd.), ale také při vysoké pravděpodobnosti nehody v daném místě a čase. V principu tedy mohou nastat dva typy varování: varování před existujícím problémem a varování před nebezpečnou situací dle predikce. Získávání a distribuce informací o nehodách je znázorněna na obrázku číslo 20 společně s distribucí predikčních modelů do zařízení.

Varování řidiče před existujícím problémem

Předpokladem pro tuto situaci je, že na trase se stala např. nehoda, a je třeba varovat řidiče. Do aplikace proudí data o problému na trase pomocí 3 nezávislých kanálů.

Obrázek 20: Přenos a distribuce informací do vícezdrojové aplikace



Zdroj: vlastní zpracování

Prvním kanálem je přenos dat z národního informačního centra přes internet. V tomto případě by šlo využít i získávání dat ze serverů zprostředkovávajících tyto informace z NDIC jako je například doprava.idnes.cz

Druhým kanálem je RDS-TMC. Zdrojem dat pro RDS-TMC je taktéž NDIC. Zpráva přenášená pomocí RDS-TMC má stručnější charakter, nicméně v případě nedostupnosti signálu mobilní sítě umožňuje tento kanál přenést alespoň základní varování pomocí rozhlasového vysílání.

Nejrychleji by měla informace o problému na trase být doručena prostřednictvím automatického systému přenosu nouzových zpráv využívající princip radia-help. Použitím této technologie by byla zajištěna vyšší spolehlivost z hlediska doručení a vyšší přesnost určení polohy, jelikož by byly krom jiných informací přenášeny GPS souřadnice, které jsou součástí MSD systému eCall.

Jelikož data o nehodách jsou distribuována, až třemi kanály jistota doručení důležité informace je vysoká. Vhodný algoritmus aplikace by musel zajistit, aby se informace ze zmíněných kanálů neduplikovala a nebyla tak mylně považována za tři různé nehody. I pro tento účel by šel využít jeden z algoritmů strojového učení, vhodný pro textminingové účely, například algoritmus SVM - support vector machine. Princip a další vlastnosti algoritmu SVM popisuje ve svém článku (Ben-Hur, 2002).

Varování řidiče před nebezpečím dle predikčních modelů

Tento druh varování má řidiče upozorňovat především na riziková místa na komunikacích s ohledem na další atributy reflektující reálnou situaci.

Pro predikci nebezpečného místa v reálném čase a místě jsou používány modely vytvořené pomocí data miningových technik a na základě historických dat o nehodách z databáze dopravních nehod. Aplikace v automobilu porovnává výsledek predikce s aktuální situací a ve chvíli kdy podobnost překročí předem stanovenou hladinu je uživatel upozorněn.

Využití predikčních modelů v aplikaci vyžaduje pro svoji správnou činnost také řídicí část, kde je pravidelně aktualizována databáze o nehodách a jsou aktualizovány predikční modely. Řídicí část má stejné funkce jako kdyby byla propojena s uživatelskou částí realizovanou pomocí speciálního zařízení: importuje a zpracovává heterogenní data, udržuje a spravuje zmíněnou databázi a připravuje modelovací matici pro potřeby data miningu. Pomocí data miningových technik a algoritmů (např. shluková analýza) jsou vytvářeny modely, jejichž aktuální verze je přenášena do mobilních zařízení v automobilu. Aktualizace modelů pro aplikaci jsou přenášeny pomocí internetového připojení z aktualizačních serverů.

Požadavky na hardware a software

Pro základní funkci aplikace lze použít běžný smartphone disponující datovým připojením a GPS modulem. Pro příjem emergency zpráv od Národního dopravního centra lze využít internet. Pro příjem hlášení eCall do aplikace je použit princip RADIA HELP, a proto by mobilní telefon musel být vybaven čipem realizujícím příjem tohoto druhu dat. Dalším alternativním kanálem pro příjem emergency zpráv od národního dopravního centra je RDS-TMC. Pro příjem RDS-TMC zpráv je nutné, aby zařízení bylo vybaveno RDS TMC dekodérem. Implementace RDS do mobilního telefonu již byla vyřešena.

6.3 Spolupráce řídicí a uživatelské části

V principu může být fungování uživatelské části realizováno třemi způsoby. Veškerá komunikace predikčního zařízení v automobilu s řídicí částí by měla probíhat přes wifi či LTE sítě čtvrté generace. (Lamr, 2015a)

První přístup (varianta A) předpokládá, že již v řídicí části budou vybudovány a do mapového podkladu přidány tzv. heat mapy, které budou představovat jakási „horká místa“, kde se vyskytlo více dopravních nehod podobného charakteru. Analýza takových shluků by měla ukázat, zda jsou i pro takové shluky nějaká další společná kritéria jako je například roční období, resp. měsíc, denní doba, teplota, či povětrnostní podmínky. V případě existence vlivu dalších ovlivňujících aspektů na dané místo nehody, bude takovýto shluk reprezentován jako jeden bod se specifickými vlastnostmi. V případě, že nebude prokázána specifická charakteristika pro dané místo, bude takový shluk označen všeobecně jako „místo častých dopravních nehod“. Zařízení v automobilu v reálném čase kontroluje pohyb vozidla a v případě, že se vozidlo bude blížit k nějakému shluku reprezentujícímu častější výskyt nehod označené jako „místo častých dopravních nehod“ bude řidič pouze upozorněn, že projíždí místem častých dopravních nehod. Pokud se bude vozidlo blížit ke shluku, který bude charakteristický specifickými podmínkami, zařízení v automobilu porovná realitu se specifickými podmínkami blížícího se shluku a řidiče upozorní na vysoké riziko nehody vzhledem k tomu, že se blíží k místu, kde se za podobných podmínek stalo vícero nehod. S tímto přístupem by bylo nutné pravidelně aktualizovat celý mapový podklad s vygenerovanými shluky do uživatelských zařízení v automobilu, zařízení by však pro korektní fungování nemuselo být stále online.

Druhý možný přístup (varianta B) využívá řídicí část spíše jako výkonnou databázi shluků dopravních nehod a více operací je přesunuto na klienta v uživatelské části. Analýzou dat v řídicí části by byly vyřazeny nevýznamné atributy nepotřebné k predikcím, a tak by byla zredukována datová náročnost při přenosu informací o nehodách ke klientovi. Shluky v řídicí části by byly vytvářeny pouze na základě počtu výskytů dopravních nehod a jejich geografických souřadnic. Zařízení v automobilu by stejně jako v předchozím přístupu monitorovalo pohyb automobilu, ale ze serveru by stahovalo s předstihem data o nehodách z míst, do kterých se automobil blíží. Každý takto stažený shluk by byl testován z hlediska ročního období (měsíce), data, času, povětrnostních podmínek či příčiny nehody (např.: nedodržení bezpečné vzdálenosti za vozidlem). Takový přístup by odstranil nutnost pravidelné masivní aktualizace mapových podkladů klienta, jelikož by se v zařízení tato data neudržovala,

ale stahovala se vždy jen data na krátkodobé použití. Nevýhodou je nutnost stálého připojení klienta k internetu, což by mohlo znamenat v případě nedostupnosti datového připojení určitá úskalí. Tím, že by se všechny shluky z databáze testovaly na svoji specifičnost (např.: místo s nehodami za deště, místo s nehodami večer) online v zařízení by vzrostly i hardwarové nároky na klienta.

Třetí možný přístup (varianta C) by mohl být kombinací prvních dvou. Stejně jako v případě varianty A by v řídicí části byly v databázi uloženy nejen shluky všeobecné vytvořené na základě počtu nehod v určitém místě, ale byly by zde i shluky specifické. Stahování dat do klienta by probíhalo obdobně jako v případě varianty B, tedy stahovala by se vždy data potřebná pro aktuální potřebu predikce v daném místě na trase vozidla. Varianta C se v tuto chvíli jeví jako nejoptimálnější varianta pro reálné použití.

6.4 Systémový návrh řešení a pohled z hlediska obecné teorie systémů

Systém lze definovat jako účelově uspořádanou množinu prvků a množinu vazeb mezi nimi, s dynamickým chováním, které společně určují vlastnosti celku. V rámci dekompozice daného systému je možné vyčlenit podsystém. Podsystémem rozumíme podmnožinu systémových prvků a vazeb, která je z nějakého důvodu vyčleněna ze systému a je chápana jako nový systém nebo jako prvek. Každý systém tedy obsahuje prvky a vazby, které tvoří základ struktury systému.

6.4.1 Místo systému v hierarchii dopravně bezpečnostních orgánů

Z hlediska provozování systému a jeho dalšího vývoje by měl systém včasného varování před vysokým rizikem dopravní nehody patřit (být podsystémem) pod vyšší autoritu jako je například Ministerstvo dopravy, Ministerstvo vnitra či Policie ČR. Vzhledem k tomu, že nedílnou součástí systému jsou data o dopravních nehodách, kterými disponuje Ministerstvo dopravy, se jeví jako logické, aby systém z větší části spravovalo Ministerstvo dopravy. To však veškerá data dostává od Policie ČR, která data o každé dopravní nehodě (od roku 2008 o každé nehodě nad 100 000 Kč) zaznamenává. Policie by do systému včasného varování zasahovala nejen poskytováním dat o nehodách, ale také je zodpovědná za jejich kvalitu, formát a zjišťované atributy při šetření dopravní nehody a není jednoduše možné z vlivu na systém včasného varování před vysokým rizikem dopravní nehody policii vynechat. Grafické vyjádření výše popsaného je zobrazeno na obrázku 21. Ministerstvo dopravy však provozuje i další systémy, které mají zvýšit bezpečnost na komunikacích. Mezi nejvýznamnější telematické

systémy, které mají zvyšovat bezpečnost na silničních komunikacích v ČR, patří „systém informačních tabulí“, a systém šíření dopravních informací pomocí technologie RDS-TMC. Systém včasného varování před vysokým rizikem dopravní nehody se od zmíněných systémů liší především tím, že se snaží dopravním nehodám aktivně předcházet, kdežto zmíněné systémy řeší situaci ve chvíli, kdy se nehoda stane. Existovat vedle sebe mohou paralelně všechny zmíněné systémy.

Obrázek 21: Systém včasného varování jeho místo v systémové hierarchii dopravně bezpečnostních orgánů

Zdroj: vlastní zpracování

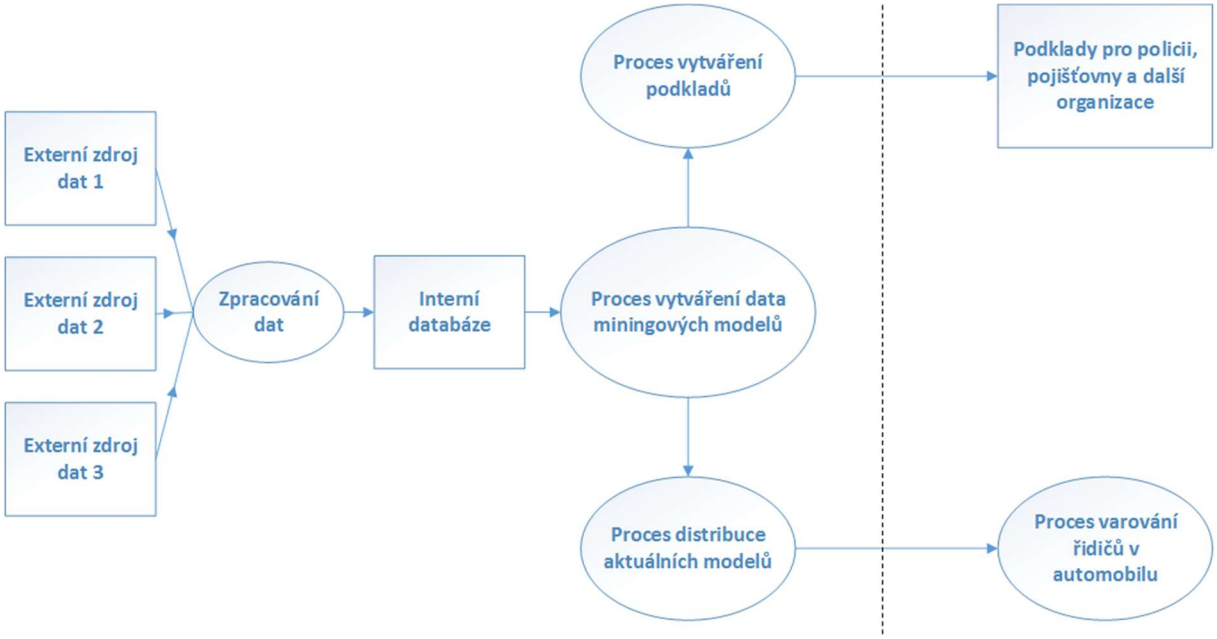
6.4.2 Prvky a vazby systému včasného varování

Jak již bylo řečeno, systém včasného varování před zvýšeným rizikem dopravní nehody podléhá určitým suprasystémům, ale na druhou stranu i on se skládá z určitých prvků a vazeb mezi nimi, přičemž některé procesy lze definovat jako podsystémy systému včasného varování. Hranice systému jsou na jedné straně tvořeny vstupy v podobě externích zdrojů dat a na straně druhé končí varováním řidiče v reálném čase a místě.

Pomyslná hranice mezi řídicí částí a uživatelskou částí systému je na obrázku 22 znázorněna čárkovanou čárou. Data z externích zdrojů v systému včasného varování reprezentují například databázi dopravních nehod Policie ČR, která však nemusí být jediným datovým zdrojem navrhovaného systému. Pro vytváření predikčních modelů je k dispozici v tuto chvíli více jak

700 000 záznamů o dopravních nehodách na území ČR, přičemž každý záznam obsahuje 44 atributů s číselnými a kategoriálními daty.

Obrázek 22: Prvky systému včasného varování



Zdroj: vlastní zpracování

Díky různorodosti potenciálních dalších datových zdrojů je nutné data zpracovat (proces zpracování dat) a uložit do interní databáze systému. Jak již bylo řečeno, data o dopravních nehodách od policie nemusí být jediným datovým zdrojem, který systém využívá.

Dalším příkladem externích zdrojů dat může být databáze výsledků celostátního sčítání dopravy. Základní výsledky sčítání dopravy včetně metodiky sběru dat lze nalézt na webové prezentaci projektu (Prezentace výsledků sčítání dopravy 2010, 2011). Jednotlivé komunikace jsou rozděleny na úseky a na těchto úsecích jsou počítány četnosti různých druhu dopravních prostředků (osobní automobily, autobusy, nákladní vozy atd..) a také je k dispozici již vypočtená relativní hustota provozu v daném místě. Data však nejsou jednoduše přístupná a je nutné je spojit s jednotlivými záznamy o dopravních nehodách.

Stěžejním je pro systém včasného varování subsystém vytváření data miningových modelů, které budou sloužit pro predikci nebezpečných míst za daných okolností. I když by bylo vhodné co nejvíce tento subsystém automatizovat, aby charakter co nejvíce odpovídal systému tvrdému, nelze to v případě vytváření modelů zcela realizovat a vždy bude nutné, aby do procesu vstupoval lidský faktor. V procesu vytváření data minigových modelů je třeba načítat

z databáze informace o dopravních nehodách a vytvořit tzv. modelovací matici (data denormalizovat). Data v modelovací matici procházejí analýzou například pomocí data miningového nástroje a následně jsou v datech pomocí algoritmů hledány vzory sloužící pro budoucí predikci.

Vytvořené modely je třeba následně distribuovat do uživatelské části systému, což zajišťuje proces distribuce aktuálních modelů. Přenos informací by měl být možný pomocí několika alternativních přenosových kanálů.

Proces varování řidičů v automobilu (uživatelská část systému) předpokládá využívání predikčních modelů ve speciálním zařízení automobilu (klient), které v reálném čase vyhodnocuje riziko nehody v závislosti na čase, aktuální poloze, stavu vozovky, stavu automobilu a počasí a dalších atributů reflektujících aktuální situaci.

6.4.3 Vstupy a výstupy systému

Na základě pozorování organických systémů zjistil (Bertalanfy, 1969), že všechny systémy jsou otevřené, jelikož nemohou existovat bez výměny hmoty a energie s okolním prostředím. Výměna energie a hmoty probíhá dvěma směry. Směrem z prostředí do systému plynou toky, které nazýváme vstupy a toky směřující ze systému ven do okolí nazýváme výstupy. Díky procesům, které zpracovávají, přetváří vstupy nebo si z nich odebírají určité složky nutné pro existenci systému, dochází k odlišnosti mezi vstupy a výstupy. Jinými slovy rozumíme vstupem systému množinu vazeb či proměnných, jejichž prostřednictvím se uskutečňuje působení okolí na systém. Výstupem systému rozumíme množinu vazeb či proměnných, jejichž prostřednictvím se uskutečňuje působení systému na jeho okolí.

Chceme-li definovat nejvýznamnější vstupy vstupující do systému včasného varování, měli bychom mluvit nejprve o mimořádných událostech, nehodách či dalších okolnostech (Obrázek 23), které jsou svým charakterem spíše nepředvídatelné a daly by se přirovnat k měkkým systémům. Tyto mimořádné události se stávají však za určitých okolností, jejichž podstatu lze hrubě popsat pomocí určitých atributů.

Dalším vstupem do systému včasného varování je Policie ČR, která vytváří záznam o dopravní nehodě a záleží na několika okolnostech, určujících kvalitu záznamu o dopravní nehodě. Významnou roli v tu chvíli má především policista, který zaznamená důležité okolnosti, za kterých se nehoda stala. Dále však záleží i na pravidlech a vyšších autoritách Policie ČR určujících, které atributy se mají zaznamenat. Zvýšení počtu dnes zaznamenávaných atributů

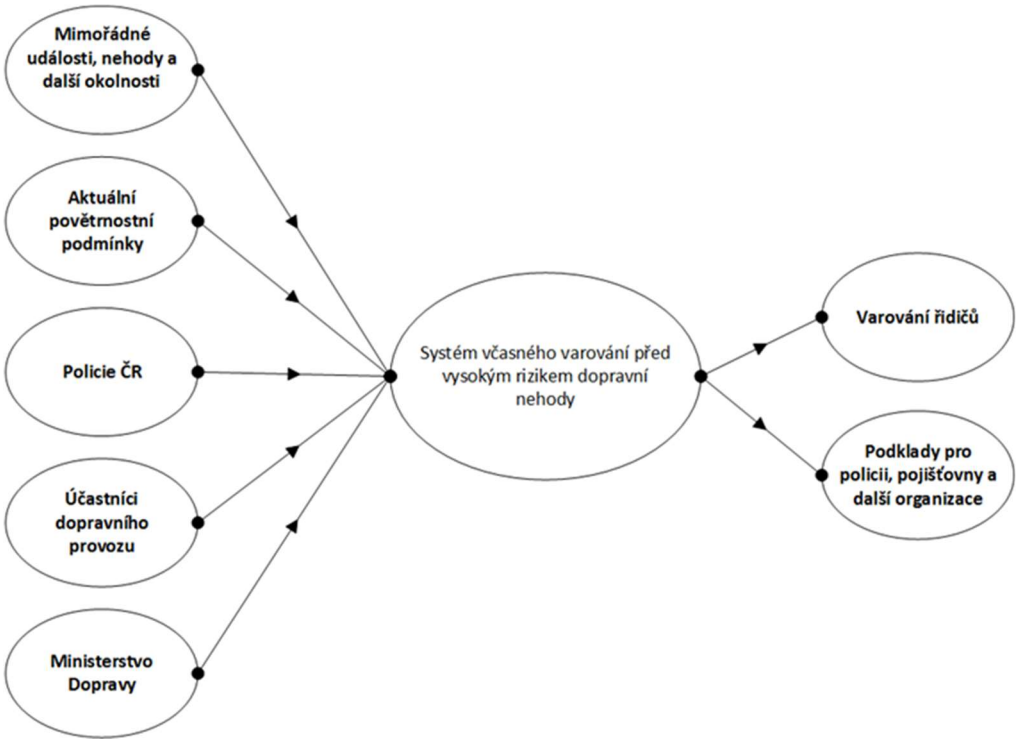
a přesné vytipování jejich důležitosti by zvýšilo přesnost predikčních modelů. Společně s policií do systému vstupuje i Ministerstvo dopravy jako předpokládaný provozovatel systému a zprostředkovatel dat.

Povětrnostní podmínky, další okolnosti a účastníci dopravního provozu vstupují do systému nejen z hlediska historických dat, na kterých se systém učí, ale vstupují do systému aktivně v reálném čase ve chvíli, kdy systém (zařízení v automobilu, které je součástí uživatelské části) vyhodnocuje aktuální situaci a porovnává ji s výsledkem predikce.

Mezi výstupy systému patří podklady pro policii, pojišťovny a další organizace, které by těmto subjektům měly usnadnit rozhodování v adekvátních situacích. Tyto výstupy mohou být současně vstupy pro jiné systémy.

Hlavním výstupem systému včasného varování by však mělo být varování řidiče před potenciálně vysokým nebezpečím na trase.

Obrázek 23: Vstupy a výstupy systému včasného varování



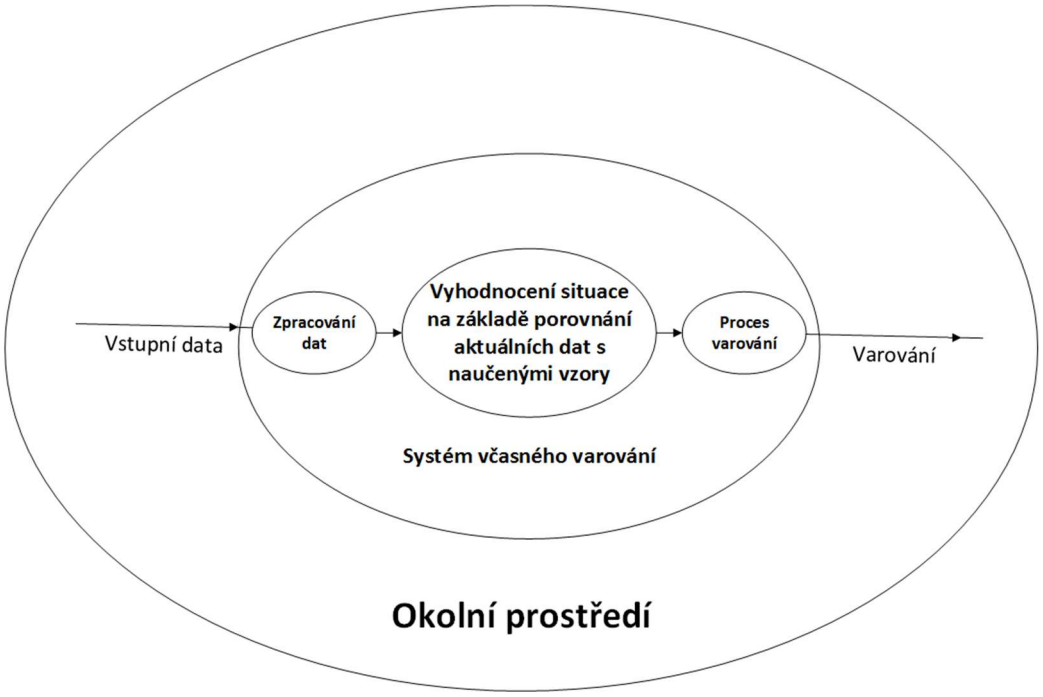
Zdroj: vlastní zpracování

6.4.4 Hledání izomorfizmu

Vzájemná korespondence mezi objekty, která zachovává i vztahy mezi těmito objekty, se nazývá izomorfií. Pokusíme-li se hledat obdobný systém srovnatelný se systémem včasného varování před vysokým rizikem dopravních nehod, nacházíme ve světě například systémy varování před tsunami. Obdobný systém jako v dokumentu popisovaný by ovšem s úplně jiným typem dat mohl fungovat při vyhledávání vzorů v datech při teroristických útocích a včasným varováním lidí blížících se do nebezpečného místa.

Budeme-li dále pokračovat v hledání isomorfismu najdeme podobné systémy i v říši některých savců. Zobecněné schéma, nad kterým bude popisována korespondence mezi systémem včasného varování a systémem včasného varování před rizikem u vybraných druhů savců je uvedeno na obrázku 24. Podobné chování jako u systému včasného varování můžeme pozorovat například u surikat, svišťů, i kozorožců. Více o hledání izomorfizmů v zobecněném schématu systému včasného varování je uvedeno v (Lamr, 2017a).

Obrázek 24: Zobecněné schéma systému včasného varování



Zdroj: vlastní zpracování

6.4.5 Cíle systému a pohled na systém včasného varování z hlediska tvrdých a měkkých systémů

Chování systémů směřuje vždy k určitému cíli nebo cílům. Cílů systému může být více, a to na všech úrovních systémové hierarchie. Každý systém je ovlivněn hned několika záměry. Je ovlivněn záměrem, kterého chce dosáhnout suprasystém, záměrem, kterým se řídí části daného systému a v neposlední řadě je ovlivněn záměrem, o jehož naplnění se snaží systém sám o sobě.

Cílem systému včasného varování je zvýšení bezpečnosti silničního provozu. Nadřízeným systémem, pod který by systém včasného varování měl patřit, je Ministerstvo dopravy v kooperaci s Policií ČR, jejichž cílem je také zvýšení bezpečnosti silničního provozu a s tím souvisejících dalších aspektů jako je snížení počtu úmrtí, lehkých i těžkých zranění. Cíle jednotlivých podsystémů (např. proces zpracování dat, či proces vytváření modelů) systému včasného varování byly popsány již výše.

V obecné teorii systémů rozlišujeme systémy na tvrdé a měkké. Každý z těchto dvou typů vyžaduje jiný analytický přístup a umožňuje odlišnou úroveň modelování současného či budoucího chování. Úlohy, které probíhají v měkkých systémech, nelze snadno strukturovat a automatizovat, jelikož na prvky systému působí mnoho faktorů, z nichž některé nejsou kvantifikovatelné. Jejich řízení vyžaduje časté zásahy člověka. Procesy probíhající v tvrdých systémech řeší dobře strukturované úlohy, které lze snadno algoritmizovat. Ve chvíli, kdy jsou systémy tohoto typu jednou navrženy, se řídí především automaticky a nevyžadují větší zásahy ze strany člověka. Každý systém inklinuje jen k jedné z uvedených dvou kategorií, avšak zmiňované přístupy jsou doplňkové a můžou tedy být aplikovány na jeden systém zároveň. Systém včasného varování patří spíše do kategorie tvrdých systémů, jelikož jsme schopni vyčíslit kvalitu výstupů, procesy, které uvnitř probíhají, se dají dobře strukturovat a úlohy lze snadno algoritmizovat. I když princip navrhovaného systému v podstatě nevyžaduje časté obměny či zásahy člověka, lze v tomto systému nalézt i charakteristiky měkkých systémů. Některé části systému jako například proces vytváření predikčních modelů, či příprava dat a datové matice (v případě, že se formát dat změní) vyžadují zásahy ze strany člověka. Ze své podstaty je i samotná dopravní nehoda vždy souhrnem faktorů, z nichž spousta není kvantifikovatelných.

6.4.6 Zpětná vazba a systém včasného varování

Zpětnou vazbou můžeme rozumět funkce, které usměrňují systém, při nichž je určitá veličina trvale sledována a porovnávána s veličinou řídicí a ovlivňována tak, aby se přibližovala

k veličině nastavené. Dle Bouldingovy klasifikace lze systém včasného varování označit jako kybernetický systém, a proto je zde možné vysledovat zpětnou vazbu jak v rámci určitých subsystémů systému včasného varování, tak i systému jako celku. Příkladem zpětné vazby v subsystému vytváření modelů může být nutnost evaluace modelů vzniklých na základě dat a v případě potřeby zpětná úprava modelovací matice. Dalším příkladem zpětné vazby systému včasného varování jako celku je nutnost pravidelného vyhodnocování kvality predikčních modelů v případě, že se změní hustota dopravy na určitých místech v rámci sledovaných komunikací. Například nový obchvat města může znamenat menší hustotu dopravy v okolí. Proto je nutné aktualizovat pravidelně nejen informace o hustotě dopravy ale následně i vygenerovat nové predikční modely respektující dané okolnosti.

6.5 Ekonomické zhodnocení systému včasného varování

V souvislosti se zvyšujícím se počtem dopravních nehod je nutné zamyslet se nad ekonomicko-společenskými důsledky souvisejícími nejen s hmotnými škodami, které jenom v roce 2017 dosáhly 6, 316 miliardy Kč, ale především nelze opomenout újmy na lidských životech. Zatímco materiální škody je velmi snadné vyčíslit, v případě lidských životů lze uvést pouze přibližné hodnoty vyjádřené pomocí současných plnění poskytovaných českými pojišťovnami. Přikloníme-li se k tomuto pragmatickému hledisku, dostaneme se k částce 10 000 000 Kč. Odborníci Nejvyššího soudu a lékaři z 1. lékařské fakulty Univerzity Karlovy se pokusili vyčíslit hodnotu lidského zdraví. Došli k přesné částce 10 051 200 korun (Zdravý člověk má cenu 10 milionů korun, stanovili experti, 2014). Nový občanský zákoník tuto hodnotu zvýšil na zhruba 10-20 mil. Kč (Vláda: Cena lidského života je 120 mil. Kč, 2016). Tato částka se odvíjí od průměrné hrubé měsíční nominální mzdy a od stanovených doplňkových kritérií, jež zahrnují nejen lékařem vyřčenou diagnózu, ale také to, jakým způsobem poranění jedince ovlivní v budoucím životě. Hodnotí se například to, zda bude jeden vyřazen kvůli nehodě ze společenského života, jakým způsobem se změní či zatíží jeho rodinné vztahy, či zda přijde o možnost dále se vzdělávat či řídit auto. Výše pojistného plnění se tudíž liší případ od případu. Orientační pojistná plnění za nejčastější dopravní nehodou způsobená zranění lze nalézt v tabulce 2.

Nejčastější a také nejbezpečnější důsledek nárazu při autonehodách bývá zranění hlavy. Ke zraněním hlavy dochází při více než 70% dopravních nehod a je také nejčastější příčinou smrti. Mezi další častá zranění patří také poranění hrudníku či páteře, poranění kostí a zlomeniny, popáleniny a šokové stavy.

Tabulka 2: Pojistná plnění vybraných zranění

Povrchní poranění nosu	1 256 Kč	Otřes mozku - lehký	5 026 Kč
Povrchní poranění ucha	1 256 Kč	Otřes mozku - těžký	15 077 Kč
Povrchní poranění rtu a dutiny ústní	1 256 Kč	Rozdrcení obličeje	150 768 Kč
Mnohočetná povrchní poranění hlavy	3 769 Kč	Poranění krční tepny	37 692 Kč
Mnohočetné rány hlavy max.	20 102 Kč	Zlomenina žebra	5 026 Kč
Zlomenina lebeční spodiny	37 692 Kč	Popálenina na 0,25 % - méně než 1 % povrchu těla II st	2 513 Kč
Ztráta zubu	5 026 Kč	Popálenina na 5 % - méně než 10 % povrchu těla III st.	75384 Kč
Traumatická ruptura hrudní meziobratlové ploténky	25 128 Kč	Zlomenina jiné zápěstní kosti - bez dislokace	8 795 Kč

Zdroj: vlastní zpracování dle (Zdravý člověk má cenu 10 milionů korun, stanovili experti, 2014)

Ekonomická náročnost samotné implementace systému včasného varování do automobilů závisí na ekonomické náročnosti řídicí a uživatelské části systému.

V řídící i uživatelské části je kromě vývoje softwaru třeba počítat i s náklady na hardwarové části. Vzhledem k tomu, že vývoj aplikace řídící části a vývoj aplikace jedné z variant uživatelské části systému včasného varování je tvořen v rámci disertační práce, a bude dále řešen v rámci výzkumu, je nutné podotknout, že náklady by připadly pouze na pořízení dostatečně robustního hardwaru.

V řídící části systému je nutné disponovat serverem, na kterém bude probíhat konverze heterogenních dat do databáze, dále serverem, kde bude prováděno vytváření modelů a servery zajišťujícími distribuci dat do klientských zařízení.

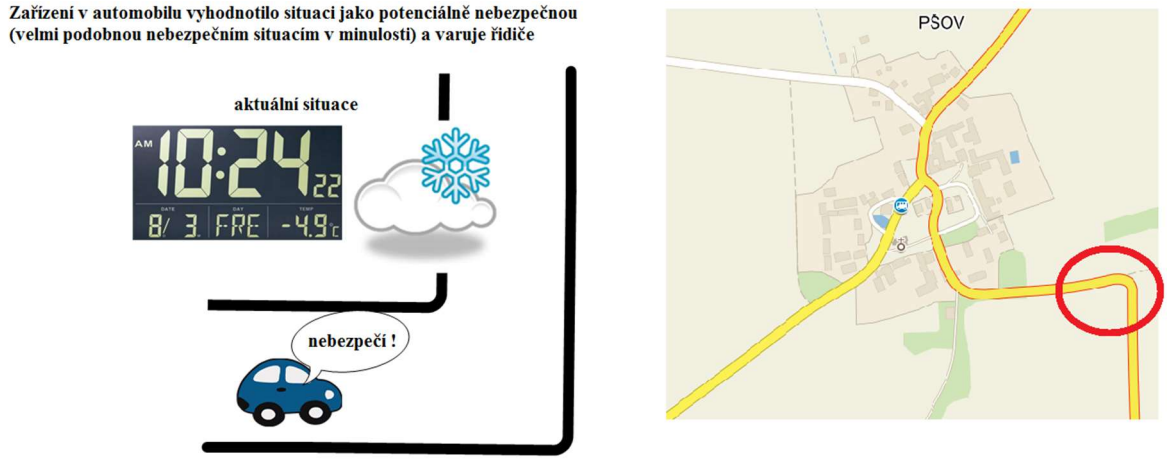
Uživatelská část může být realizována třemi variantami. První varianta předpokládá vývoj speciálního zařízení v automobilu, které by však mělo být ze své podstaty velmi podobné zařízení typu tablet a bylo by třeba vyřešit především implementaci konektoru pro spojení s automobilem. Druhá varianta předpokládá řešení uživatelské části formou mobilní aplikace, která je vyvíjena v rámci výzkumu a pro uživatele by byla dostupná volně ke stažení. Uživatel by byl nucen pořídit si, dle konkrétního typu automobilu, konektor umožňující komunikaci

automobilu s telefonem prostřednictvím bluetooth. Třetí varianta předpokládá implementaci uživatelské části systému do infotainmentu automobilů. Náklady s tím spojené by připadly na úkor jednotlivých výrobců automobilů.

6.6 Modelové situace

Princip systému včasného varování před dopravní nehodou lze ilustrovat na zatáčce, která je označována Centrem dopravního výzkumu ČR jako nejnebezpečnější zatáčka z celé sítě silnic první třídy a dálnic v ČR. Danou situaci ilustruje obrázek 25.

Obrázek 25: Ilustrace fungování systému v nebezpečném místě



Zdroj: vlastní zpracování

Tato zatáčka se nachází se na silnici 1. třídy s číslem 27 u obce Pšov na Podbořansku. Silnici, která je významným tahem mezi Ústeckým krajem a Plzní, ve velké míře využívají kamiony. Často se při projíždění ostré zatáčky dostávají minimálně částí vozidla do protisměru a ohrožují tak řidiče, kteří přijíždějí ve směru od Pšova. Mnohá auta jedou ještě pár desítek metrů před záhybem silnice výrazně vyšší rychlostí, než je zde povolená třicítka. Pokud je pak silnice mokrá, nebo je na ní dokonce sníh či náledí, auto se snadno dostane do smyku a havárie je téměř neodvratná. Je evidentní, že i přes několikanásobné značení nebezpečného místa řidiči toto značení podceňují. Riziko nehody se v tomto místě rapidně zvyšuje v případě sněhové pokrývky nebo náledí. Ve chvíli, kdy bude automobil vybaven systémem včasného varování před vysokým nebezpečím nehody, který porovnává aktuální povětrnostní situaci s výsledkem predikčních modelů, je možné řidiče varovat pouze v situacích, které jsou opravdu nebezpečné. Věříme, že v případě, že budou řidiči dostávat pouze adekvátní informace, které jsou z hlediska

bezpečnosti opravdu zásadní, budou mít v tento systém důvěru a přizpůsobí styl jízdy v případě potřeby. (Lamr, 2016c)

Pro ilustraci vyhledávání shluků jsme z databáze dopravních nehod vybrali geografickou oblast spadající dle policie ČR pod město Podbořany. Tato množina obsahuje 430 záznamů o dopravních nehodách za období let 2007-2013.

Oblast nebezpečné zatáčky je označena algoritmem DBSACN jako Cluster 55 a obsahuje 71 případů. Podívejme se nyní detailněji na tuto množinu. Nejvíce nehod se stává mezi druhou a třetí hodinou odpoledne. Většina nehod, které se zde odehrály, naštěstí není tragických, není zde výskyt smrtelného zranění, a jen ve dvou případech došlo k těžkému zranění. Zhruba ve třiceti procentech nehod však dochází k lehkým zraněním, přičemž ve většině případů jde o zranění pouze jedné osoby.

Zajímavé je, že celá polovina ze všech nehod se tu stává během června, července a srpna. Budeme-li dále zkoumat tuto skupinu nehod, zjistíme, že v 90 procentech nehod bylo jako příčina nehody označeno nepřizpůsobení rychlosti dopravně technickému stavu vozovky. V 63 procentech případů nehod situaci zkomplikovaly povětrnostní podmínky v podobě deště. Ve více jak třech čtvrtinách případů byl atribut *Stav povrchu vozovky* označen jako mokrý. U 55 procent případů byla Policií ČR označena viditelnost jako snížená právě vlivem povětrnostních podmínek.

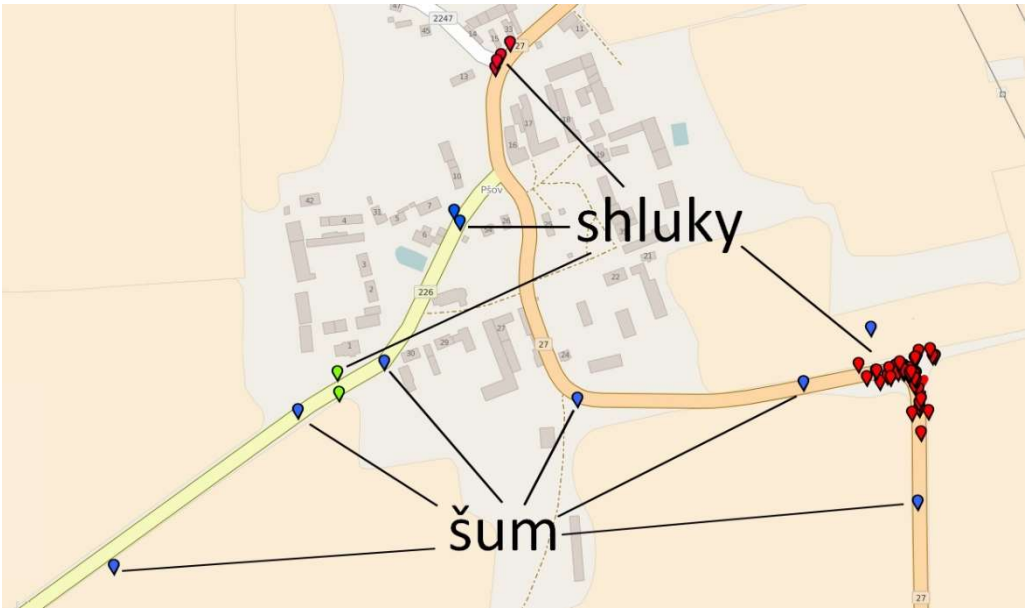
Porovnat výsledky algoritmů založených na hustotě (DBSCAN) a algoritmu K-means lze v obrázcích 26 a 27. Na obrázku 26 jsou shluky vytvořeny pomocí algoritmu DBSCAN v software KNIME. Shluky byly vytvářeny na datové matici obsahující 431 záznamů o dopravních nehodách v okolí "nejnebezpečnější" zatáčky v ČR, která se nachází u obce Pšov. V případě algoritmu DBSCAN se jako nejoptimálnější volba jeho parametrů s ohledem na počet záznamů jeví eps=0,025 a MinPts=2. Při daném nastavení nalezl algoritmus v dané oblasti 87 clusterů, resp. 86 clusterů a posledním clusterem byl šum. Jako šum jsou ve většině případů označeny odlehlé hodnoty, což je v pořádku. V následujících obrázcích jsou jednotlivé záznamy o dopravních nehodách patřící do stejného clusteru vyznačeny stejnou barvou.

Pro srovnání jsou v obrázku 27 zobrazeny shluky dopravních nehod vytvořené pomocí shlukovacího algoritmu K-Means. Abychom dosáhli stejného počtu shluků jako v předchozím případě, nastavili jsme k=87. A lgoritmus K-means však nepočítá s šumem a shluky které vytvořil, nejsou určeny tak dobře jako pomocí algoritmu DBSCAN. Z nehod v inkriminované

zatáčce, které zjevně patří do jednoho stejného shluku, vytvořil shluků hned několik. V místě zmiňované zatáčky s větší koncentrací dopravních nehod tedy algoritmus selhal. V oblastech s menší koncentrací nehod algoritmus vytvořil některé shluky dle očekávání. Dále jsou v obrázku 27 vyznačena dvě místa, která algoritmus Kmeans označil jako jeden cluster, i když spolu evidentně nesouvisí. Hlavní nevýhodou algoritmu K-means v případě použití pro hledání shluků v geografických datech je nutnost zadat předpokládaný počet shluků. (Lamr, 2016c)

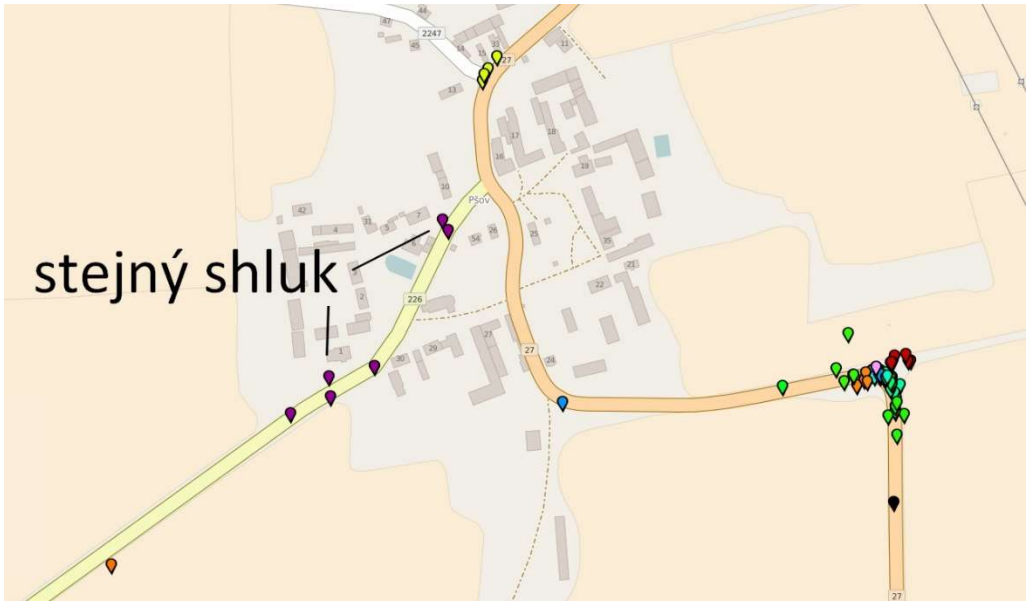
Systém včasného varování by měl být užitečný v případech, kdy řidič jede danou trasou poprvé. V situaci, kdy jede řidič takovým místem prvně, může situaci na první pohled vyhodnotit zdánlivě za méně zákeřnou a díky nepřiměřené rychlosti, kterou vjíždí do zatáčky, nezvládne řízení. Ve chvíli, kdy bude řidič přijíždějící do daného místa včas upozorněn na vysoké riziko nebezpečí nehody, by mohlo dojít ke snížení počtu podobných nehod. (Lamr, 2016c)

Obrázek 26: Shluky nehod v okolí nejnebezpečnější zatáčky v ČR (DBSACN eps=0,025, MinPts=2)



Zdroj: vlastní zpracování

Obrázek 27: Shluky nehod v okolí nejnebezpečnější zátáčky v ČR (K-means k=87)



Zdroj: vlastní zpracování

7 Hledání skrytých závislostí v datech o dopravních nehodách pomocí DM nástrojů jako součást řídícího části systému včasného varování

V následujících podkapitolách bude pozornost věnována hledání skrytých závislostí v datech o dopravních nehodách. Nejprve bude rozebírána detekce míst (shluků), kde se často stávají dopravní nehody a následně budou v těchto jednotlivých shlucích hledána asociační pravidla. Tuto úlohu budeme v textu označovat jako „**úloha 1**“. Druhou úlohou (v textu označována jako „**úloha 2**“), kterou je též možné řešit, je hledání asociačních pravidel bez předchozí detekce shluků určených na základě polohy nehody. Tyto dvě samostatné úlohy budou popisovány jako dva samostatné data miningové projekty, u kterých budou komentovány jednotlivé fáze CRISP-DM, které byly doposud realizovány.

7.1 Úloha 1

7.1.1 Porozumění problému

V datech o dopravních nehodách chceme vyhledávat nejprve shluky nehod detekované pouze pomocí GPS souřadnic. Jednotlivě nalezené shluky budou následně testovány tak, že v nich budou hledány asociační pravidla s vysokou mírou podpory a spolehlivosti. Cílem úlohy je tedy hledání zajímavých asociačních pravidel v rámci jednotlivých „skupin“ nehod, které se staly na jednom místě.

7.1.2 Porozumění datům

Porozumění a příprava dat byla prováděna především v programu IBM SPSS Modeler. Každý záznam o dopravní nehodě se skládá ze 44 atributů popisujících okolnosti nehody. Většina atributů obsahuje kategoriální data (tabulka 3). Dostupná jsou například data o místě nehody (GPS souřadnice, obec, kraj), datu a času. V databázi též nalezneme informace o povětrnostních podmínkách, viditelnosti, stavu vozovky a příčině nehody. Dalšími dostupnými atributy jsou informace o viníkovi nehody, stavu řidiče vozidla, počtu raněných a usmrcených. V neposlední řadě jsou k dispozici údaje o škodě na vozidle i celkové škodě vzniklé při nehodě. V rámci analýzy datového souboru byly atributy rozděleny do několika skupin. Atributy v jednotlivých skupinách mají vždy určitý společný jmenovatel.

První skupinou jsou atributy určující místo nehody. Do této skupiny lze zařadit například atributy: *GPS X*, *GPS Y*, *Obec* a *Kraj*. Nejdůležitější v této skupině jsou prvně dva jmenované atributy určující přesně místo nehody.

Další skupinou jsou atributy určující čas a datum nehody. Jedná se o následující atributy: *Datum nehody*, *Měsíc*, *Rok*, *Den nehody* (např. pondělí, úterý…), *Den v měsíci* (1.,2.,3…), *Čas nehody*. Některé atributy v této skupině jsou redundantní (např. rok, den nehody…) a lze je odvodit z jiných atributů. V rámci jedné datové matice však jejich existence není na závadu a hodí se například pro filtrování nehod podle roku či dne v týdnu.

Třetí skupinu tvoří atributy týkající se komunikace, na které se nehoda stala. Tyto atributy lze použít pro upřesnění polohy nehody, a hlavně pro lepší rozdělení nehod s různou hustotou nehodovosti, což je důležité pro algoritmy vyhledávající shluky. Do této skupiny patří následující atributy: *Druh pozemní komunikace*, *Směrové poměry*, *Číslo pozemní komunikace*, *Dělení komunikace*, *Místo dopravní nehody*, *Situování nehody na komunikaci*, *Řízení provozu v době nehody*, *Místní úprava přednosti v jízdě*. Detailněji budou tyto atributy rozebrány v kapitole o přípravě datové matice.

Do čtvrté skupiny jsme zařadili atributy týkající účastníků nehody. V souboru jsou informace o počtu lehce zraněných, těžce zraněných a usmrcených lidí, dále jsou zde atributy jako: *Způsob vyproštění osob z vozidla*, *Vlastník vozidla*, *Kategorie řidiče*, *Stav řidiče*, *Vnější ovlivnění řidiče*, *Alkohol u viníka nehody*. Tato skupina atributů může být využita pro nejrůznější filtrování nehod (např. nehody způsobené alkoholem, drogami) avšak pro účel predikce nebezpečných situací se nehodí.

V páté skupině najdeme atributy týkající se vozidla: *Počet zúčastněných vozidel*, *Druh vozidla*, *Výrobní značka motorového vozidla*, *Vozidlo po nehodě*, *Rok výroby vozidla*, *Únik provozních hmot*, *Škoda na vozidle*, *Celková hmotná škoda*. V této skupině se opět vyskytují atributy, pomocí kterých můžeme datovou sadu všemožně filtrovat, avšak pro hledání shluků přichází v úvahu pouze dva atributy. Prvním je *Počet zúčastněných vozidel*, pomocí kterého bychom mohli detekovat například hromadné nehody. Druhým atributem, který lze použít pro modelování je *Druh vozidla*, který určuje, zda se jedná o osobní automobil či nákladní vůz.

Tabulka 3: Ukázka struktury záznamů o dopravní nehodě

ID	2100070013	2100070019	2100070022
Obec	Praha	Praha	Praha
Kraj	Hlavní město Praha	Hlavní město Praha	Hlavní město Praha
Datum nehody	01.01.2007	02.01.2007	02.01.2007
Cas	19:00:00	7:30:00	8:30:00
Den	pondělí	úterý	úterý
Druh pozemní komunikace	komunikace účelová - ostatní (par	uzel (křižovatka sledovaná ve vybı	komunikace místní
Číslo pozemní komunikace	0	0	0
Zavinění nehody	řidičem motorového vozidla	řidičem motorového vozidla	řidičem motorového vozidla
Alkohol u viníka nehody	ano, obsah alkoholu v krvi do 0,99	ne	ne
Usmrceno osob (počet)	0	0	0
Těžce zraněno osob (počet)	0	0	0
Lehce zraněno osob (počet)	0	0	0
Druh nehody	srážka s vozidlem zaparkovaným,	srážka s pevnou překážkou	srážka s jedoucím nekolejovým vo
Druh srážky jedoucích vozidel	nepřichází v úvahu, nejde o srážku	nepřichází v úvahu, nejde o srážku	boční
Druh pevné překážky	nepřichází v úvahu, nejde o srážku	odrazník, patník, sloupek, dopr.zn	nepřichází v úvahu, nejde o srážku
Hlavní příčiny nehody	nesprávné otáčení nebo couvání	řidič se plně nevěnoval řízení vozi	při přejíždění z jednoho pruhu do
Druh povrchu vozovky	živice	živice	živice
Stav povrchu vozovky v době nehody	povrch mokrý	povrch suchý, neznečistěný	povrch suchý, neznečistěný
Stav komunikace	dobrý, bez závad	dobrý, bez závad	dobrý, bez závad
Povětrnostní podmínky v době nehody	neztížené	neztížené	neztížené
Viditelnost	v noci - s veřejným osvětlením,vid	ve dne, viditelnost nezhoršená vli	ve dne, viditelnost nezhoršená vli
Rozhledové poměry	dobré	dobré	dobré
Dělení komunikace	žádná z uvedených	žádná z uvedených	třípruhová
Situování nehody na komunikaci	žádné z uvedených	na jízdním pruhu	na jízdním pruhu
Řízení provozu v době nehody	žádný způsob řízení provozu	místní úprava (vyplní se pol. 24)	žádný způsob řízení provozu
Místní úprava přednosti v jízdě	žádná místní úprava	přednost vyznačena dopravními zr	přednost nevyznačena - vyplývá z
Specifické objekty v místě nehody	parkoviště přiléhající ke komunika	žádné nebo žádné z uvedených	v blízkosti přechodu pro chodce (d
Směrové poměry	přímý úsek	křižovatka průsečná - čtyřramenná	přímý úsek
Místo dopravní nehody	mimo křižovatku	na křižovatce, uvnitř hranic křižova	mimo křižovatku
Druh křižující komunikace	neurčeno	neurčeno	neurčeno
Smyk	ne	ne	ne
Směr jízdy nebo postavení vozidla	vozidlo jedoucí - na komunikaci be	zachycuje směr jízdy křižovatkou	vozidlo jedoucí - na komunikaci be
Počet zúčastněných vozidel	4	1	2
Druh vozidla	osobní automobil bez přívěsu	osobní automobil bez přívěsu	osobní automobil bez přívěsu
Výrobní značka motorového vozidla	ŠKODA	NISSAN	ŠKODA
Rok výroby vozidla	98	4	5
Vlastník vozidla	soukromé, nevyužívané k výděleč	soukromé, nevyužívané k výděleč	registrované mimo území ČR
Celková hmotná škoda (100 Kč)	450	400	400
Škoda na vozidle (100 Kč)	130	400	300
Vozidlo po nehodě	nedošlo k požáru	nedošlo k požáru	nedošlo k požáru
Únik provozních, přepravovaných hmot	žádné z uvedených	žádné z uvedených	žádné z uvedených
Způsob vyproštění osob z vozidla	nebylo třeba užít násilí	nebylo třeba užít násilí	nebylo třeba užít násilí
Kategorie řidiče	s řidičským oprávněním skupiny b	s řidičským oprávněním skupiny b	s řidičským oprávněním skupiny b
Stav řidiče	pod vlivem alkoholu, obsah alkoh	dobrý -žádné nepříznivé okolnosti	dobrý -žádné nepříznivé okolnosti
Vnější ovlivnění řidiče	řidič nebyl ovlivněn	řidič nebyl ovlivněn	řidič nebyl ovlivněn
lng	14.515	14.408	14.422
lat	50.022	50.037	50.077

Zdroj: vlastní zpracování

Velmi významná z hlediska využití pro tvorbu predikčních modelů je skupina atributů popisující počasí a jeho vliv v místě nehody. Do této skupiny patří následující atributy: *Stav povrchu vozovky v době nehody*, *Povětrnostní podmínky v době nehody* a *Viditelnost*. Nejdůležitějším atributem v této kategorii jsou *Povětrnostní podmínky*. Tento atribut může nabývat následujících hodnot: déšť, jiné ztížené, mlha, na počátku deště nebo slabý déšť, nárazový vítr (boční, vichřice apod.), neztížené podmínky, sněžení, tvoří se námraza či náledí.

Atribut *Viditelnost* nabývá sedmi kategoriálních hodnot, které v podstatě říkají, zda byla viditelnost zhoršena povětrnostními podmínkami s ohledem na denní dobu.

Poslední sedmou skupinu tvoří atributy popisující příčinu nehody a její další důležité okolnosti. Do této skupiny byly zařazeny následující atributy: *Zavinění nehody*, *Druh nehody*, *Druh srážky jedoucích vozidel*, *Druh pevné překážky*, *Hlavní příčiny nehody*, *Stav komunikace*, *Rozhledové poměry*, *Specifické objekty v místě nehody*, *Smyk*, *Druh povrchu vozovky*, *Směr jízdy* nebo *Postavení vozidla*. Tato skupina atributů je důležitá nejen pro efektivní selekci nehod podle zmíněných atributů, ale především pro svůj predikční potenciál. Předpokládáme, že především tato skupina atributů bude použita pro hledání asociačních pravidel upřesňujících vlastnosti specifických shluků nehod. Specifickým shlukem nazýváme shluk nehod vykazující s vysokou konfidencí určité specifické vlastnosti. Obecným shlukem nazýváme takové shluky, u nichž nelze najít asociační pravidla s vysokou konfidencí. Více o návrhu rozdělení shluků na obecné a specifické lze najít v (Lamr, 2015a).

I když se může na první pohled zdát, že důležité budou v této skupině především atributy *Hlavní příčina nehody* a *Zavinění nehody*, skutečnost je trochu jiná. Většina atributů v této skupině má svůj určitý zásadní význam pro predikci, avšak některé informace ze zmíněných atributů se částečně duplikují a při hledání asociačních pravidel bude nutné k tomu přihlédnout a datovou matici dodatečně vhodně upravit. Ukázka struktury datové matice z hlediska dostupných atributů a hodnot, kterých mohou nabývat, je zobrazena v tabulce 1. Pro názornost jsou vybrány pouze tři případy dopravních nehod.

7.1.3 Příprava dat

Jelikož je příprava dat tou časově nejnáročnější úlohou, byla i tato fáze prováděna z velké části v uživatelsky přívětivém a robustním nástroji IBM SPSS modeler.

Jak bylo řečeno již výše, motivací pro práci s daty o dopravních nehodách je pro nás realizace systému včasného varování před vysokým rizikem dopravní nehody. Součástí tohoto systému je hledání shluků dopravních nehod vytvořených na základě údajů o místě nehody a jejich následné testování na případné specifické vlastnosti. Jelikož mohou být některé algoritmy určené pro vytváření shluků v geospatial datech citlivé na změny hustoty výskytu jednotlivých záznamů, je důležité na tuto věc pamatovat a adekvátně se s tím vypořádat, například při přípravě datové matice. V této kapitole se pokusíme nastínit návrh řešení tohoto problému.

Jedním z možných řešení, jak do hledání shluků zavést informaci o hustotě dopravy, je připojení do datové matice informací o intenzitách dopravy, které vytváří Ředitelství silnic a dálnic ČR. Ředitelství silnic a dálnic provádí přibližně každých 5 let sčítání dopravy. Poslední sčítání dopravy s dostupnými výsledky se uskutečnilo v roce 2010. Sčítání plánované na rok 2015 bylo odloženo a bylo prováděno až v loňském roce 2016. Základní výsledky sčítání z roku 2010 jsou veřejně dostupné na webu scitani2010.rsd.cz.

Výsledky celostátního sčítání dopravy 2010 poskytují informace o intenzitách automobilové dopravy na dálniční a silniční síti ČR v roce 2010 a navazují na výsledky z předchozích sčítání v roce 2005. Na dálnicích jsou intenzity dopravy stanoveny zejména pomocí údajů z automatických detektorů dopravy. Na silnicích jsou intenzity dopravy stanoveny z výsledků ručních průzkumů a pomocí přepočtových koeficientů variací intenzit dopravy. Oproti předchozím Celostátním sčítáním dopravy (2005 a starším) byly koeficienty zpřesněny a více diferencovány podle charakteru provozu na komunikaci. Uváděné hodnoty jsou ročním průměrem denních intenzit dopravy ve vozidlech za 24 hodin (Prezentace výsledků sčítání dopravy 2010, 2011).

V průběhu roku 2017 probíhaly pokusy o spojení dat z databáze dopravních nehod s daty z celostátního sčítání dopravy 2010. Hlavním problémem bylo hledání atributu, či kombinace atributů pro snadné spojení těchto datových sad. Jedním z atributů, který lze použít pro spojení, je číslo pozemní komunikace, avšak tento atribut sám o sobě je pro spojení nepoužitelný. Komunikace bývají totiž dále rozděleny na menší úseky a v těchto úsecích jsou také prováděna měření intenzit dopravy. Z tohoto problému vyplývá naše doporučení pro sběr informací o dopravních nehodách pro policii, která by měla ukládat do databáze i informace o úseku komunikace, na kterém se nehoda stala.

Příprava dat o dopravních nehodách

Alternativou či doplněním v případě nedostupnosti informací o intenzitách dopravy v nebezpečných místech může být využití stávajících informací upřesňujících polohu nehody uložených v databázi dopravních nehod. Pro dělení nehod z hlediska místa výskytu nehody je z databáze nehod policie ČR možné využívat atributy námi označované v kapitole 7.1.3 jako skupina tři. Seznam těchto atributů společně s jejich možnými hodnotami lze vidět v tabulce 4.

Tabulka 4: Atributy upřesňující polohu nehody a jejich možné hodnoty

Název atributu	Možné hodnoty	
Druh pozemní komunikace	Komunikace místní silnice 1. třídy silnice 2. třídy silnice 3. třídy dálnice	komunikace účelová (parkoviště), komunikace sledovaná komunikace účelová (polní a lesní cesty) uzel (křižovatka sledovaná ve vybraných městech)
Směrové poměry	kruhový objezd křižovatka pěti a víceramenná křižovatka průsečná - čtyřramenná křižovatka styková - tříramenná	přímý úsek přímý úsek po projetí zatáčkou (do vzdálenosti cca 100 m od optického konce zatáčky) zatáčka
Číslo pozemní komunikace	0-94823	
Dělení komunikace	čtyřpruhová s dělící čarou čtyřpruhová s dělícím pásem dvoupruhová	rychlostní komunikace třípruhová vícepruhová žádná z uvedených
Místo dopravní nehody	-mimo křižovatku -mimo zónu 11-19 a 22-28 -na křižovatce, uvnitř hranic křižovatky definovaných pro systém evidence nehod (zóna 9) -na křižovatce, jedná-li se o křížení silnic 3.tř.,místních, účelových komunikací -na vjezdové nebo výjezdové části větve při mimoúrovňovém křížení	-uvnitř zóny 1 předmětné křižovatky -uvnitř zóny 2 předmětné křižovatky -uvnitř zóny 3 předmětné křižovatky -uvnitř zóny 4 předmětné křižovatky -uvnitř zóny 5 předmětné křižovatky -uvnitř zóny 6 předmětné křižovatky -uvnitř zóny 7 předmětné křižovatky -uvnitř zóny 8 předmětné křižovatky
Situování nehody na silnici	mimo komunikaci na chodníku nebo ostrůvku na jízdním pruhu na kolejích tramvaje na krajnici	na odbočovacím, připojovacím pruhu na odstavném pruhu na pruhu pro pomalá vozidla na stezce pro cyklisty žádné z uvedených
Řízení provozu v době nehody	místní úprava (vyplní se pol. 24) policistou nebo jiným orgánem	světelným signalizačním zařízením žádný způsob řízení provozu
Místní úprava přednosti v jízdě	přednost nevyznačena - vyplývá z pravidel přednost vyznačena dopravními značkami přednost vyznačena přenosnými	dopravními značkami nebo zařízením světelná signalizace mimo provoz světelná signalizace, přerušovaná žlutá žádná místní úprava
Druh křižující komunikace	místní komunikace neurčeno silnice 1.třídy	silnice 2.třídy silnice 3.třídy účelová komunikace větev mimoúrovňové křižovatky

Zdroj: vlastní zpracování

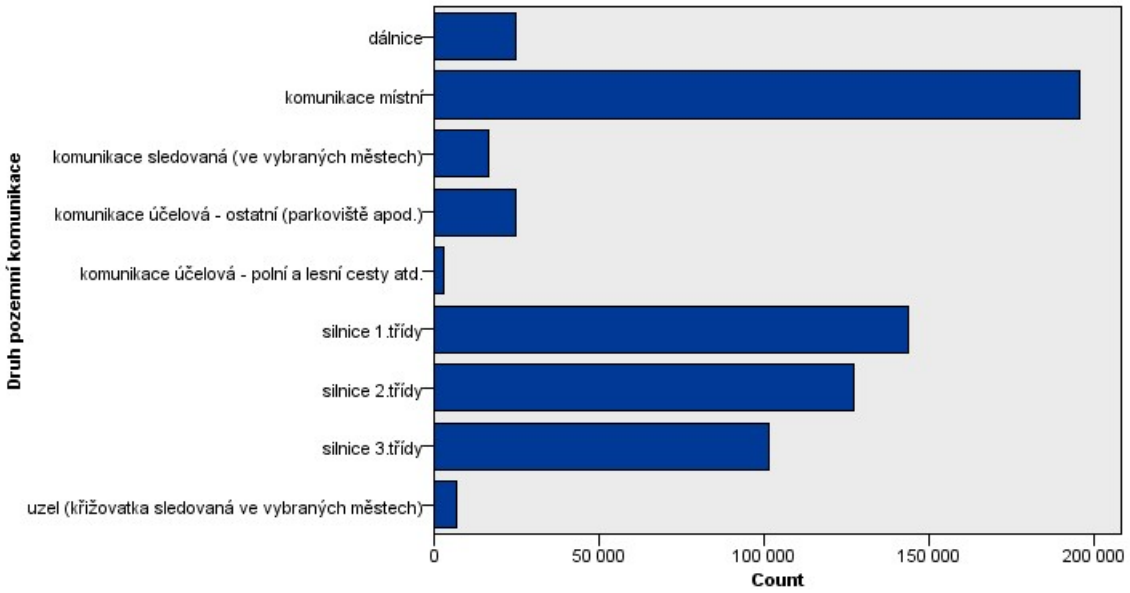
Pro lepší vytváření shluků pomocí algoritmu DBSCAN je optimální od sebe oddělit různé typy nehod z hlediska místa nehody. Lepšího výsledku při vytváření shluků nehod bude dosaženo, budou-li od sebe odděleny místa s různou hustotou nehodovosti. Místem nehody je myšleno, zda se nehoda stala například ve městě, komunikaci první či jiné třídy, nebo na dálnici. Toto rozdělení je důležité pro přesnější vytváření shluků nehod v místech s rozdílnou hustotou nehod. Jiná hustota výskytu nehod bude zřejmě ve městech, jiná bude na komunikaci třetí třídy

a jiná bude na dálnici. Vzdálenost jednotlivých nehod patřících do jednoho shluku je ve městě dozajista kratší nežli vzdálenost jednotlivých nehod patřících do stejného shluku na dálnici. Na dálnicích jsou nebezpečná místa (shluky) rozlohou větší.

V následujícím textu bude detailněji rozebrán význam jednotlivých atributů pro vytváření shluků nehod. Dále budou prezentovány autorova doporučení pro úpravu jednotlivých kategorií hodnot v rámci jednotlivých atributů. Nastíněny budou také doporučení pro úpravu celé datové matice (např. slučování stávajících atributů, či odvozování nových atributů z již existujících).

Nejdůležitějším atributem z hlediska rozdělení záznamů o nehodě podle hustoty výskytu nehod je atribut *Druh pozemní komunikace*. Atribut může nabývat devíti kategoriálních hodnot viz Tabulka 4. Rozložení četností nehod podle druhu komunikace je zobrazeno na grafu 6. Nejvíce nehod (více než 30 %) se stává na místních komunikacích, což jsou většinou komunikace ve městech a obcích. Dalšími početnými skupinami jsou silnice první (22 %), druhé (20 %) a třetí třídy (16 %). Za účelem zmenšení počtu kategorií bychom doporučili sjednotit kategorie Dálnice s kategorií Silnice první třídy vzhledem k jejich podobné intenzitě výskytu dopravních nehod. Dále doporučujeme sloučení kategorií komunikace sledovaná a kategorie uzel s kategorií místní komunikace.

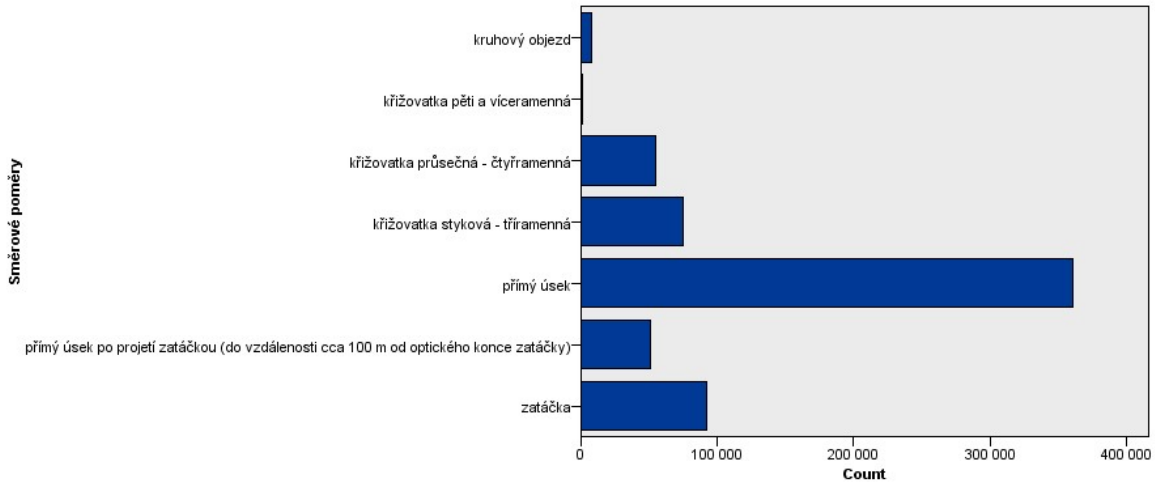
Graf 6: Rozdělení nehod podle druhu komunikace



Zdroj: vlastní zpracování

Velmi důležitý je pro správné rozdělení záznamů o nehodě atribut *Směrové poměry*. Tento atribut přesně určuje, zda se nehoda stala na přímém úseku, v křižovatce či v zatáčce. Více jak 56 % nehod se stává na přímém úseku, přičemž dalších 8 % nehod tvoří nehody na přímém úseku po projetí zatáčkou do vzdálenosti 100 metrů od zatáčky. Sedm kategorií tohoto atributu (viz Tabulka 2) navrhujeme sloučit do kategorií tří. První novou kategorií budou tvořit všechny typy křižovatek včetně kruhových objezdů. Další kategorií budou přímé úseky a poslední kategorií budou zatáčky. Rozdělení této skupiny atributů podle počtu nehod v databázi je vidět v grafu 7.

Graf 7: Směrové poměry



Zdroj: vlastní zpracování

Atribut *Číslo pozemní komunikace* je číselná proměnná, a nabývá hodnot od 0 do 94823. Tento atribut jsme zařadili do skupiny středně důležitých atributů pro upřesnění polohy nehody. Využití tohoto atributu je možné především při případném spojování dat o dopravních nehodách s daty z celostátního sčítání dopravy, kde jej lze použít jako součást klíče pro spojování těchto dvou datových sad. Důležité je však podotknout, že číselná hodnota atributu pozemní komunikace může nabývat stejných hodnot pro různé druhy komunikací (například číslo komunikace s hodnotou „1“ může existovat jak u dálnice, tak i u silnice první třídy atd.).

Další atribut, který jsme zařadili do skupiny středně důležitých, je *Dělení komunikace*. Tento atribut může nabývat jedné ze sedmi možných kategoriálních hodnot určujících počet jízdních pruhů dané komunikace, kde se stala dopravní nehoda. Téměř 80 % všech nehod se však stalo na dvoupruhové komunikaci, zhruba 7 % nehod se dle policie stalo na čtyř pruhové komunikaci

a zhruba 8 % nehod bylo přiřazeno do kategorie žádná z uvedených. Využití tohoto atributu shledáváme především při tvorbě nového kategoriálního atributu, který by vznikl kombinací tohoto s atributem *Druh pozemní komunikace*. Díky teoretické možnosti vytvoření nového atributu respektujícího aspekty atributu dělení komunikace a aspekty atributu *Druh pozemní komunikace* budeme schopni lépe detekovat komunikace s vyšší hustotou dopravy, neboť se dá předpokládat, že například komunikace první třídy, která bude čtyřpruhová, bude mít i větší hustotu výskytu dopravních nehod než komunikace první třídy dvoupruhová.

Atribut *Místo dopravní nehody* může nabývat jedné ze třinácti kategoriálních hodnot, které určují, zda se nehoda stala mimo křižovatku či uvnitř jedné ze zón křižovatky, jak je definuje Policie ČR. Více než 78 % nehod je označeno hodnotou mimo křižovatku, cca 13 % nehod je označenou hodnotou „zóna 9“. Všechny možné hodnoty atributu jsou opět uvedeny v Tabulce 4. Pro naše účely tento atribut nemá větší význam, jelikož jej můžeme nahradit atributem směrové poměry, který poskytuje z našeho pohledu téměř stejné informace.

Taktéž atribut *Situování nehody na silnici* jsme označili jako málo důležitý. Může nabývat jedné z deseti kategoriálních hodnot, které určují, zda se nehoda stala například mimo komunikaci, na chodníku nebo ostrůvku, na jízdním pruhu či na kolejích tramvaje. U většiny nehod (80 %) je uvedena hodnota „na jízdním pruhu“. Dalších 10 % nehod je označeno hodnotou mimo komunikaci, 4 % nehod se staly na krajnici a zbylých 8 kategorií tvoří jen nepatrné procento vzhledem k ostatním.

Do poslední trojice atributů námi označené jako „nedůležitá skupina“ patří atributy *Řízení provozu v době nehody*, *Místní úprava přednosti v jízdě*, *Druh křižující komunikace*. Atribut *Řízení provozu v době nehody* rozlišuje 4 kategoriální hodnoty a lze podle něj určit, zda byla doprava řízena například místní úpravou, policistou, světelným zařízením či nebyla řízena žádným způsobem. Přibližně 72 % nehod bylo označeno jako žádný způsob řízení a 25 % nehod bylo označeno hodnotou místní úprava řízení. Možné hodnoty, kterých mohou nabývat poslední dva zmíněné atributy lze opět nalézt v Tabulce 2. Pro naše účely jsou však tato data téměř nedůležitá. Pro přehlednější orientaci o důležitosti jednotlivých atributů upřesňujících polohu byla vytvořena *Tabulka 3*, ve které jsou i stručně popsána autorova doporučení pro přípravu dat.

Detailní rozbor dat o dopravních nehodách a zjišťování možností využití jednotlivých atributů je velice důležitou součástí projektu *Systém včasného varování před zvýšeným nebezpečím dopravní nehody*. Nutná příprava těchto dat do podoby vhodné pro vytváření modelů má velký

vliv na kvalitu získaných výsledků při hledání shluků dopravních nehod podle místa nehody a na vytváření modelů sloužících pro predikci rizika nehody v reálném čase a místě. Příprava dat je bezesporu velmi časově náročnou fází každého data miningového projektu. Skupina atributů, která byla vybrána z databáze o nehodách pro detailnější rozbor (tabulka 5), obsahuje několik důležitých atributů, které nám pomohou lépe identifikovat shluky nehod v oblastech s různou intenzitou dopravy. (Lamr, 2016d)

Tabulka 5: Atributy upřesňující polohu nehody a jejich důležitost pro vytváření shluků

Typ atributu	Důležitost	Poznámka, doporučení
Druh pozemní komunikace	4 -nejdůležitější	Vytvořit novou proměnnou, určit nové kategorie, za tímto účelem některé stávající kategorie vhodně sloučit
Směrové poměry	3 - velmi důležitá	Přesně určuje, zda byla nehoda na přímém úseku, v křižovatce, či zatáčce
Číslo pozemní komunikace	2 - středně důležitá	Určuje číslo komunikace (0 -94823), pozor např. č. „1“ může existovat u druhu komunikace „dálnice“ i „silnice 1. třídy“ případné použití tohoto atributu má smysl pouze s atributem „druh pozemní komunikace“
Dělení komunikace	2 - středně důležitá	Použít pro vytváření nové proměnné společně s atributem „druh pozemní komunikace“
Místo dopravní nehody	1 - středně důležitá	78 % nehod je mimo křižovatku, 13 % na křižovatce, uvnitř hranic křižovatky Doporučení: sloučit všechny úrovně křižovatky do jedné kategorie. Lepší dělení poskytuje atribut „směrové poměry“
Situování nehody na silnici	1 - málo důležitá	80 % nehod se stalo na jízdním pruhu, 10 % mimo komunikaci
Řízení provozu v době nehody	0 - nejméně důležitá	Nezahrnovat do datové matice pro modelování - není vhodné pro dělení do skupin
Místní úprava přednosti v jízdě	0 - nejméně důležitá	Nezahrnovat do datové matice pro modelování - není vhodné pro dělení do skupin
Druh křižující komunikace	0 - nejméně důležitá	Nezahrnovat do tvorby kategorií (80 % tvoří kategorie „neurčeno“)

Zdroj: vlastní zpracování

Získávání a příprava dat z celostátního sčítání dopravy

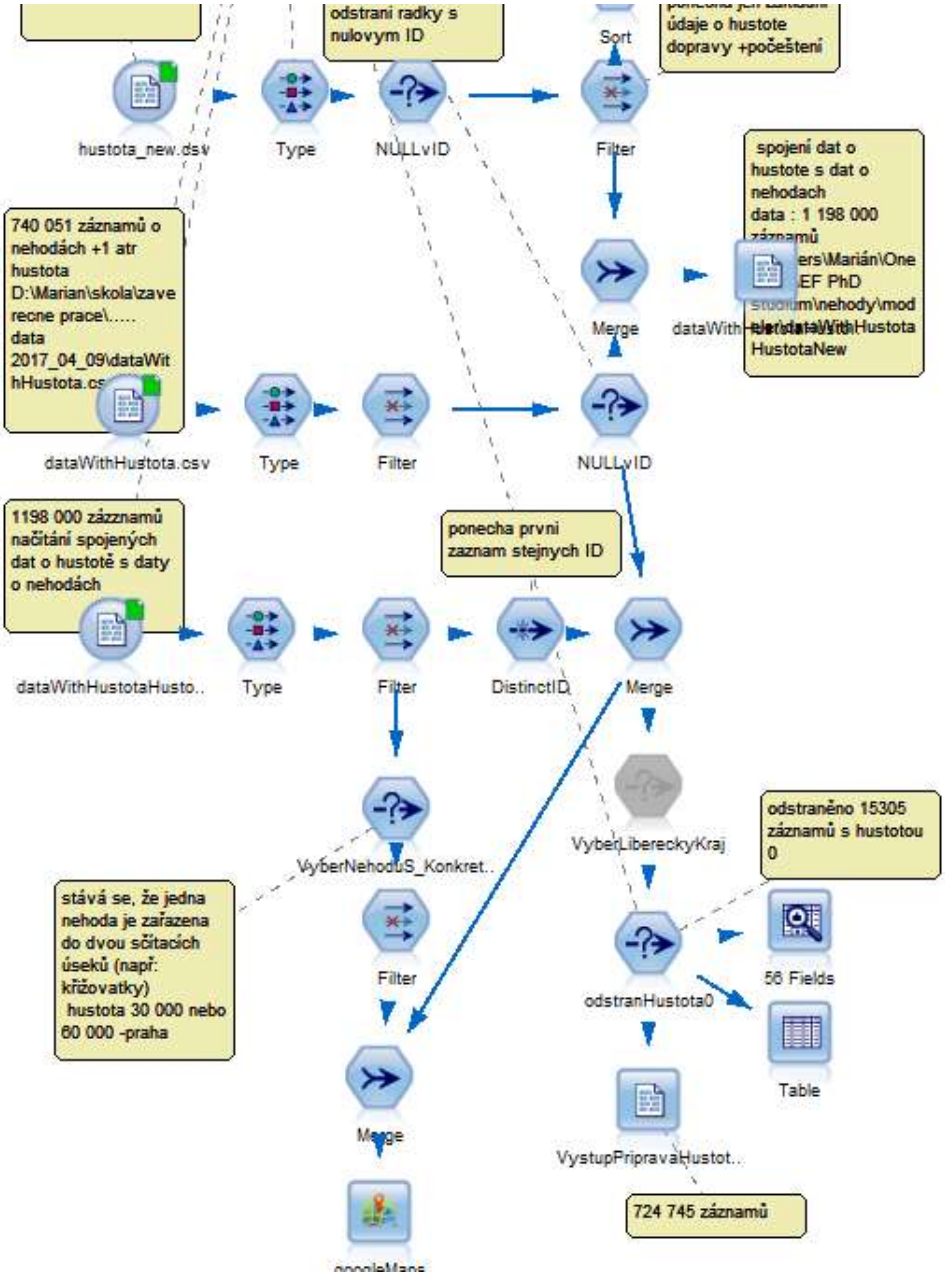
Jak bylo zmiňováno na začátku této kapitoly, v průběhu roku 2017 byla řešena dostupnost a distribuce dat o hustotě dopravy, kterými disponuje Ředitelství silnic a dálnic. Jelikož není možné najít jednoznačný identifikátor, který by propojil jednotlivé nehody s informacemi o hustotě dopravy v jednotlivých úsecích měřených komunikací, bylo nutné se s tímto problémem nějak vypořádat. Za tímto účelem je nutné vytvořit skripty, které za pomocí GPS souřadnic jednotlivých nehod zajišťují přiřazování jednotlivých ID nehod ke sčítacím úsekům.

Zjednodušeně řečeno se skript na serveru s informacemi o hustotě dopravy snaží najít pro každou nehodu v databázi nejbližší sčítací úsek v určitém námi stanoveném okolí. Pokud se zadaným okolím pozice nehody nelze najít žádný sčítací úsek, okolí se zvětší a úsek se vyhledává znovu.

Problém tohoto řešení je stanovení velikosti okolí, ve kterém se vyhledává. Při malé hodnotě okolí je potřeba více dotazů na server s informacemi o sčítání dopravy a tím je server více zatížen. Přesnost určení správného sčítacího úseku je však vysoká. S větším okolím hledání sčítacího úseku je vytížení serveru malé, avšak přesnost hledání sčítacího úseku je menší. V tomto případě se může stát, že zejména u nehod, které se stanou v blízkosti u křižovatek, jsou k jedné nehodě přiřazeny např. dvě informace o hustotě dopravy.

Část jednoho ze streamů zajišťujících část přípravy datové matice je znázorněn na Obrázku 28. V tomto streamu se načítají data o dopravních nehodách a data o hustotě dopravy. V datech jsou odstraněny záznamy s null hodnotami v ID a jsou přejmenovány špatně zakódované názvy proměnných. V nově vzniklé datové matici jsou dostupné informace jak o nehodách, tak i data o parametrech hustoty dopravy na komunikacích. Jelikož může být nehoda přiřazena do více sčítacích úseků, existuje více záznamů se stejným ID, datová matice v prvním výstupu obsahuje 1 198 000 záznamů. V další části streamu jsou takto upravená data opět načtena a je proveden DISTINCT nad položkou ID. Pro kontrolu jsou v datech ponechána jen data o hustotě a následně jsou spojena s původní maticí dat o dopravních nehodách. Předposledním krokem v tomto streamu je odstranění záznamů s NULL hodnotami ve sloupci hustota. Jedná se o 15 305 záznamů u kterých nemáme tento údaj k dispozici. Výstupní matice v tomto streamu tedy obsahuje 724 745 záznamů a 59 sloupců.

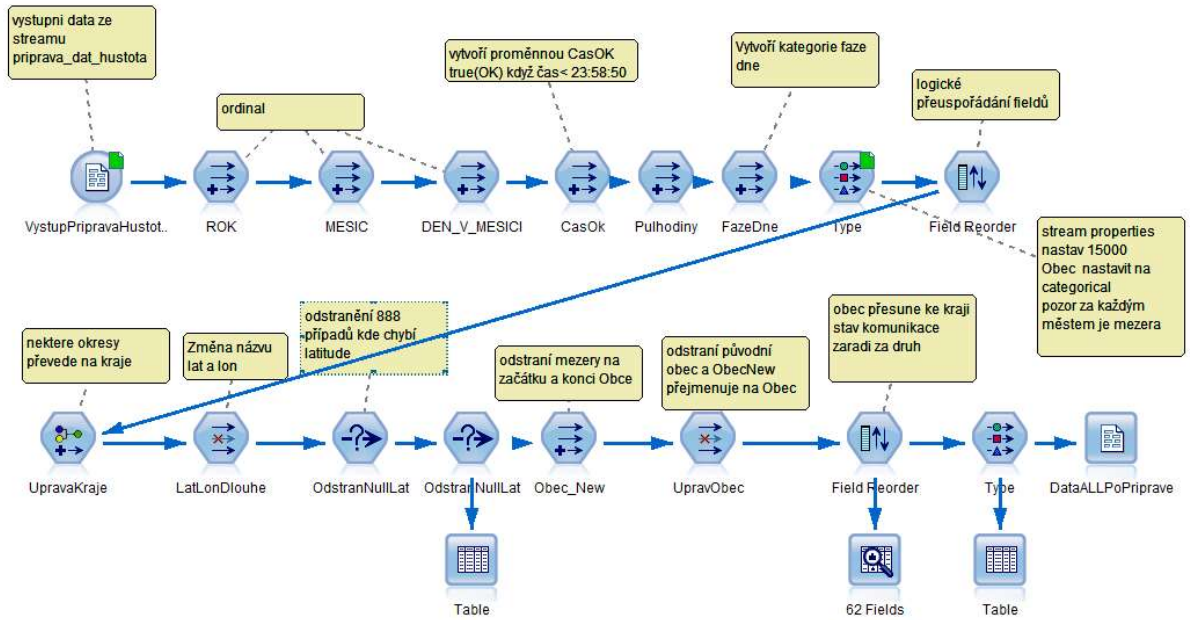
Obrázek 28: Ukázka streamu přípravy dat



Zdroj: vlastní zpracování

Další část přípravy dat, kde jsou derivovány některé nové proměnné a čištěna data, je zobrazena na Obrázku 29.

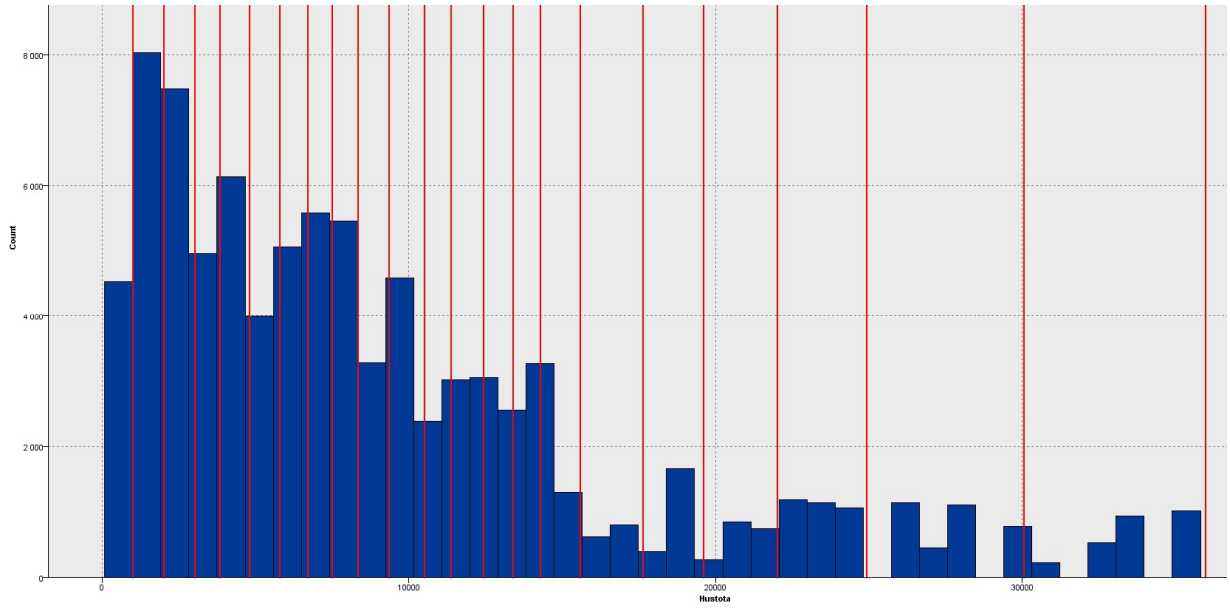
Obrázek 29: Ukázka streamu části přípravy dat



Zdroj: vlastní zpracování

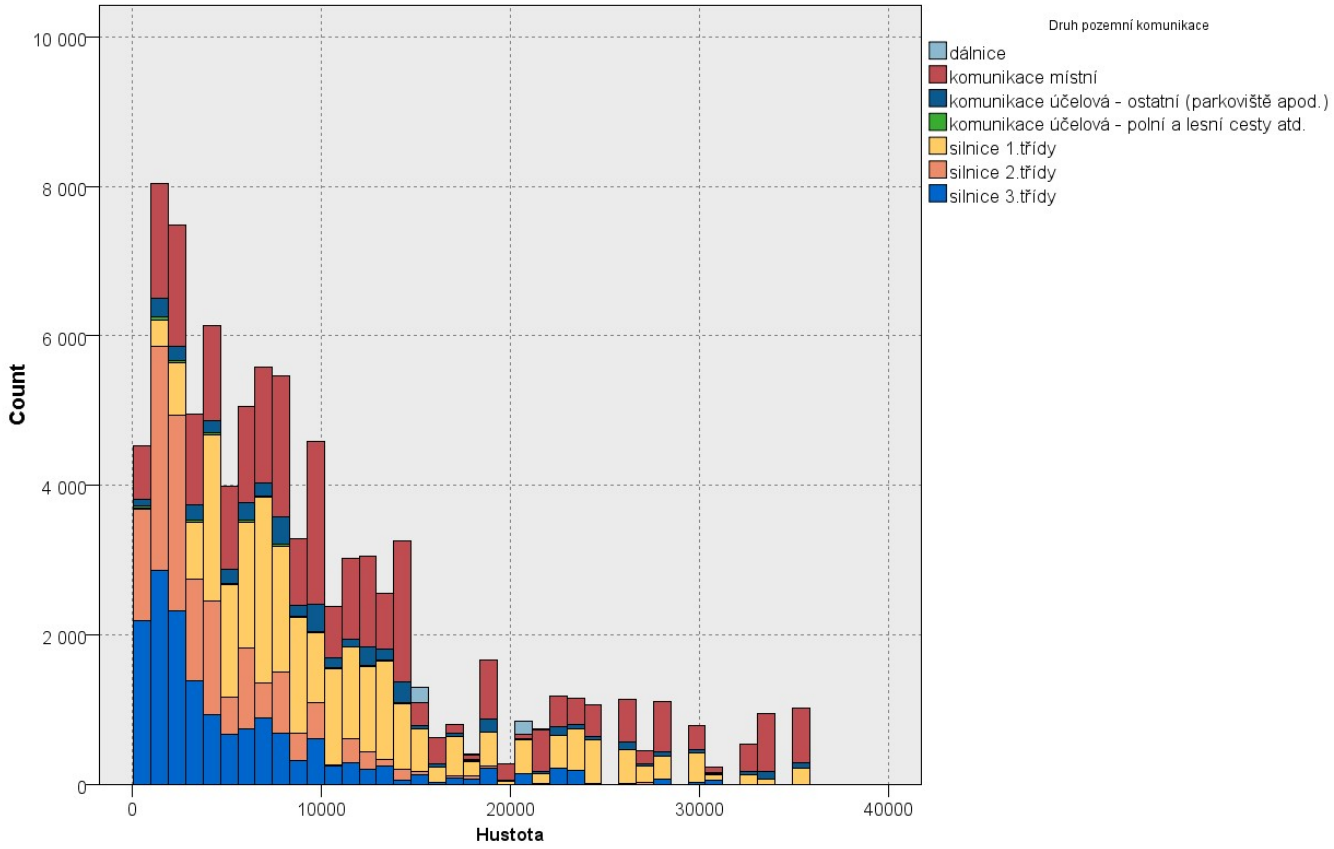
Na grafech 8 a 9 je zobrazen histogram hustoty dopravy pro data z Královehradeckého a Libereckého kraje pro všechny typy komunikací. Z grafů je vidět rozložení této veličiny a především to, že největší skupinu tvoří nehody s hustotou dopravy za 24 hodin do 10 000 automobilů.

Graf 8: Počet nehod podle hustoty dopravy za 24 hodin



Zdroj: vlastní zpracování

Graf 9: Počet nehod podle hustoty dopravy za 24 hodin v závislosti na typu komunikace

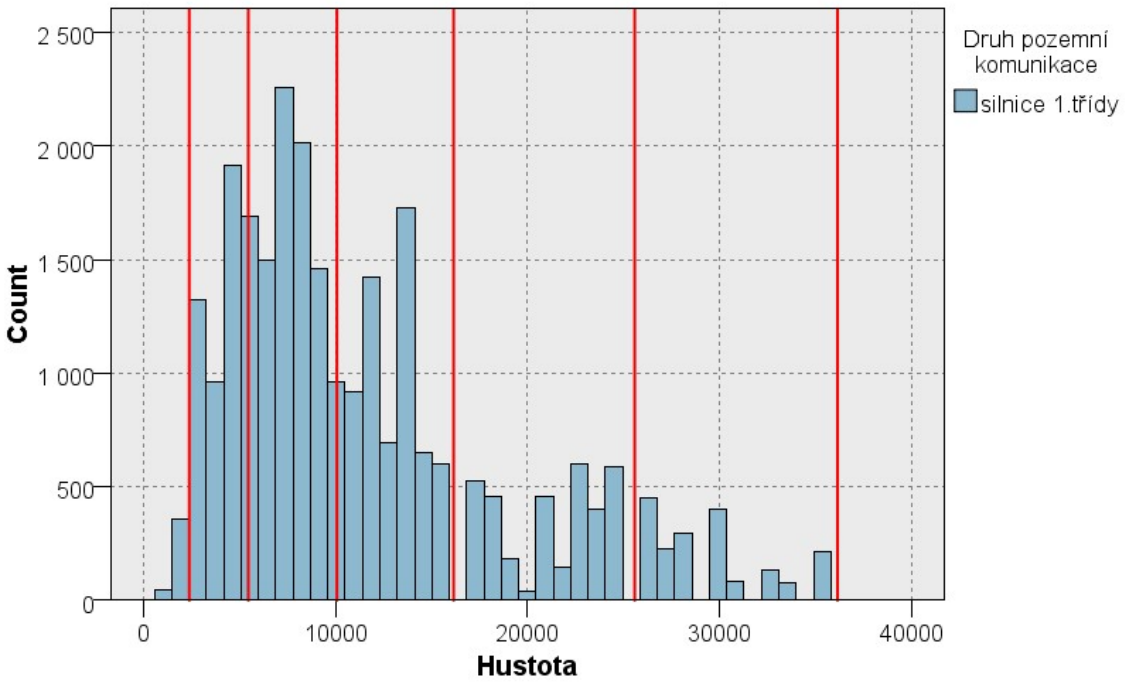


Zdroj: vlastní zpracování

Největší skupinu nehod s hustotou dopravy do 5 tisíc tvoří nehody, které se staly na komunikacích 2. a 3. třídy. Z grafu 10 je velmi dobře vidět, že na komunikacích 1. třídy se nejvíce pohybuje hustota dopravy mezi 5 až 15 tisíci vozidly za den. Zajímavá je kategorie nehod, které se staly na místní komunikaci. Jedná se především o komunikace ve městech. Nemalý podíl nehod se stává právě na těchto komunikacích bez ohledu na hustotu dopravy. Však největší podíl má tato kategorie na nehodách v místech s vysokou hustotou dopravy.

Vybereme-li pouze silnice 1. třídy, lze v grafu 10 intuitivně vytvořit kategorie pro novou proměnnou hustota, podle které bychom mohli rozdělit nehody tak, že shluky budeme vyhledávat v jednotlivých nově vzniklých kategoriích.

Graf 10: Počet nehod podle hustoty dopravy za 24 hodin (komunikace 1. třídy)



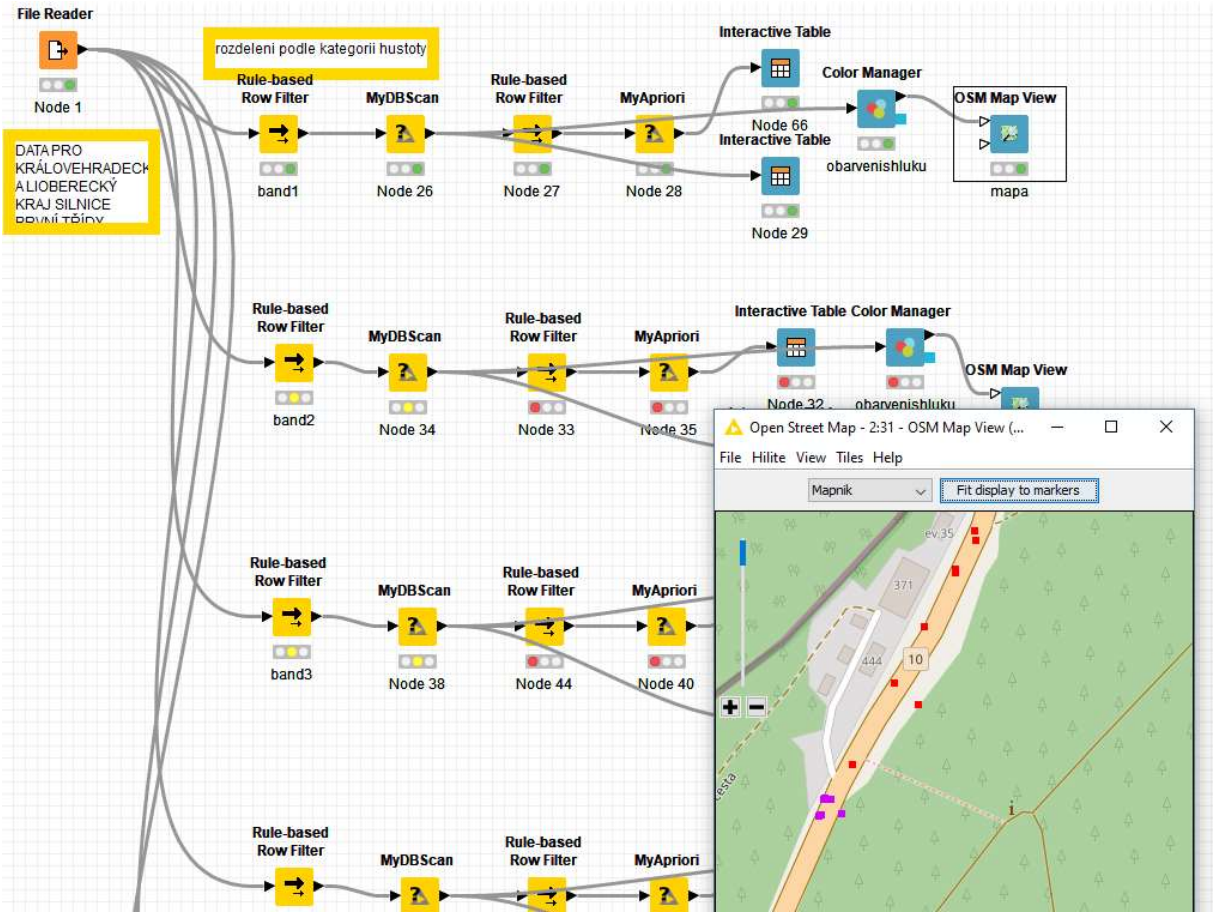
Zdroj: vlastní zpracování

7.1.4 Modelování

Fáze modelování není realizována v prostředí IBM SPSS Modeler, jelikož tento software nedisponuje algoritmy specializovanými pro vyhledávání shluků v geospatial datech. Tato fáze je řešena především v software KNIME, který disponuje algoritmy typ DBSCAN či OPTICS. Navíc je v tomto programu možné jednodušeji vytvořit vlastní specifický uzel, který bude vyhledávat asociační pravidla v jednotlivých shlucích.

Naším cílem při modelování je nejprve vyhledat místa častých dopravních nehod. To znamená místa, kde je výskyt počtu nehod zvýšený oproti okolí. Každý záznam o nehodě obsahuje kromě jiného atributy určující geografickou polohu nehody (GPS X a GPS Y souřadnice). Pro vyhledávání shluků v geografických datech je velmi často používán algoritmus DBSCAN (Density-Based Spatial Clustering of Applications with Noise),) či algoritmus OPTICS. Prostředí KNIME obsahuje základní neoptimalizovanou verzi algoritmu DBSCAN a algoritmu OPTICS. Na obrázku 30 je znázorněn výřez WorkFlow z aplikace KNIME realizující vyhledávání shluků nehod založených na GPS souřadnicích pomocí vlastního uzlu DBSCAN. V obrázku je vidět také možnost vizualizace do mapového podkladu.

Obrázek 30: Ilustrace Workflow v KNIME realizující shlukování nehod a vizualizaci shluků



Zdroj: vlastní zpracování

V tomto konkrétním WorkFlow je vyhledávání shluků a následné vyhledávání asociačních pravidel ve shlucích prováděno na datech z Královehradeckého a Libereckého kraje. Výběr je dále omezen na komunikace 1. třídy (cca 89 000 záznamů). Podle analýzy rozložení výskytu nehod podle hustoty dopravy provedené v IBM SPSS Modeler byla vytvořena nová proměnná, která nám umožní dále zvláště vyhledávat shluky na datech s různou hustotou dopravy. Datová sada je rozdělena do šesti skupin s nehodami patřícími do míst s různou hustotou dopravy. To je důležité především proto, že můžeme zvláště nastavit parametry algoritmu DBSCAN pro data s různou hustotou dopravy. Tím je dosaženo větší přesnosti při vytváření shluků. Ve workFlow je dále možné pomocí upraveného algoritmu APRIORI vyhledávat **hromadně** asociační pravidla v jednotlivých shlucích. Pro ilustraci je v tabulce 6 zobrazena část tabulky s asociačními pravidly.

Tabulka 6: Tabulka s asociačními pravidly

Row ID	Group	Prepoklady	Zaver	Support	Confide...
Row 0	16	[neztížené[Povětrnostní podmínky v době nehody]]	[dobrý, bez závad[Stav komunikace]]	71.429	7
Row 1	16	[dobrý, bez závad[Stav komunikace]]	[neztížené[Povětrnostní podmínky v...	100	5
Row 2	16	[ve dne, viditelnost nezhoršená vlivem povětrnos...	[dobrý, bez závad[Stav komunikace]]	71.429	7
Row 3	16	[dobrý, bez závad[Stav komunikace]]	[ve dne, viditelnost nezhoršená vliv...	100	5
Row 4	16	[neztížené[Povětrnostní podmínky v době nehody]]	[ve dne, viditelnost nezhoršená vliv...	100	5
Row 5	16	[ve dne, viditelnost nezhoršená vlivem povětrnos...	[neztížené[Povětrnostní podmínky v...	100	5
Row 6	16	[ano[Smyk]]	[dobrý, bez závad[Stav komunikace]]	71.429	7
Row 7	16	[dobrý, bez závad[Stav komunikace]]	[ano[Smyk]]	100	5
Row 8	16	[neztížené[Povětrnostní podmínky v době nehody...	[dobrý, bez závad[Stav komunikace]]	71.429	7
Row 9	16	[neztížené[Povětrnostní podmínky v době nehody...	[ve dne, viditelnost nezhoršená vliv...	100	5
Row 10	16	[dobrý, bez závad[Stav komunikace], ve dne, vid...	[neztížené[Povětrnostní podmínky v...	100	5
Row 11	16	[ve dne, viditelnost nezhoršená vlivem povětrnos...	[neztížené[Povětrnostní podmínky v...	100	5
Row 12	16	[neztížené[Povětrnostní podmínky v době nehody]]	[dobrý, bez závad[Stav komunikace...	100	5
Row 13	16	[dobrý, bez závad[Stav komunikace]]	[neztížené[Povětrnostní podmínky v...	100	5
Row 14	104	[neztížené[Povětrnostní podmínky v době nehody]]	[ne[Smyk]]	100	5
Row 15	104	[ne[Smyk]]	[neztížené[Povětrnostní podmínky v...	100	5
Row 16	52	[jiný (neuvedený) stav nebo závada komunikace[...	[nepř. rychlosti stavu vozovky (nál...	71.429	7
Row 17	52	[nepř. rychlosti stavu vozovky (náledí, výtluky, bl...	[jiný (neuvedený) stav nebo závad...	71.429	7
Row 18	52	[jiný (neuvedený) stav nebo závada komunikace[...	[ano[Smyk]]	63.636	11
Row 19	52	[ano[Smyk]]	[jiný (neuvedený) stav nebo závad...	100	7
Row 20	52	[neztížené[Povětrnostní podmínky v době nehody]]	[ano[Smyk]]	54.545	11
Row 21	52	[ano[Smyk]]	[neztížené[Povětrnostní podmínky v...	100	6
Row 22	52	[nepř. rychlosti stavu vozovky (náledí, výtluky, bl...	[ano[Smyk]]	63.636	11

Zdroj: vlastní zpracování

7.2 Úloha 2

7.2.1 Porozumění problému

Cílem úlohy dvě je hledání shluků nehod s podobnými vlastnostmi bez zapojení informací o poloze. Při hledání podobných typů nehod budou zapojeny všechny atributy popisující aspekty dopravní nehody. Bude nutné vynechat atributy, které mezi sebou korelují, nebo se k predikci nehodí. Předpokládáme, že takto vyhledané shluky by mohly být užitečné například ve chvíli budování nových komunikací. Místa označená jako problematická by se vyznačovala určitými parametry a při budování nové komunikace by na ně byl brán zřetel. Největší nevýhodou takto navrhovaného řešení je však fakt, že nedokáže identifikovat shluky, které nejsou ničím typické. Takto nebudou odhalena místa častých dopravních nehod, u kterých nemáme v datové matici zachyceny všechny příčiny a okolnosti nehody. Policie při vyplňování záznamu o dopravní nehodě nezachytí veškeré okolnosti, za kterých byla nehoda způsobena. Jelikož nebylo dosaženo v této úloze uspokojivých výsledků (viz kapitola 5.2.4), pokusili jsme se cíle této úlohy definovat jinak.

7.2.2 Porozumění datům

Fáze porozumění datům je v úloze 2 prakticky stejná jako v úloze 1 a tak jí nebude v této úloze věnována pozornost. Porozumění datům probíhalo rovněž v software IBM SPSS Modeler. Pro

analýzu dat byly využívány uzly pro tvorbu grafů, Data audit, Feature selection, Statistics a další.

7.2.3 Příprava dat

I fáze přípravy dat má velmi podobný charakter v této úloze jako v úloze 1. Pro Modelování byla nejprve vybrána data pro Královehradecký a Liberecký kraj a omezena na komunikace první třídy. Po analýze dat byly z datové matice odstraněny následující atributy, které se pro vyhledávání shluků nebezpečných míst nehodí, nebo korelují s jinou vhodnější proměnnou pro predikci. Seznam vynechaných proměnných je následující.

Vynechané proměnné

- Datum nehody
- Cas
- Pulhodiny
- CasOk
- Rok
- Den v mesici
- Kraj
- Obec
- Druh pozemni komunikace
- Cislo pozemni komunikace
- Druh povrchu vozovky
- Výrobní značka
- Vozidlo po nehodě
- Rok výroby vozidla
- Únik provozních, přepravních hmot
- Celková škoda
- Způsob vyproštění osob z vozidla
- Vlastník vozidla
- Stav řidiče
- Alkohol u viníka nehody
- Longitude
- Latitude
- všechny atributy týkající se hustoty dopravy kromě atributu kategorie hustoty

Množina byla pomocí uzlu Partition rozdělena na testovací a trénovací v poměru 50:50. Dále bylo ve fázi přípravy dat využíváno i uzlu Auto data preparation.

7.2.4 Modelování

Při modelování bylo využíváno uzlu pro automatické vytváření shluků Auto Cluster, který zkouší vyhledávat shluky pomocí algoritmů K-means, Two Step a Kohonen.

Náhled na popisné charakteristiky výsledku uzlu automatické segmentace je znázorněn na obrázku 15. Totožných výsledků bylo dosaženo i bez použití uzlu pro automatickou přípravu

dat. Je to způsobeno především tím, že algoritmy jako Two Step (viz kapitola 3.2.6) v Modeleru jsou vybaveny automatickým předzpracováním dat.

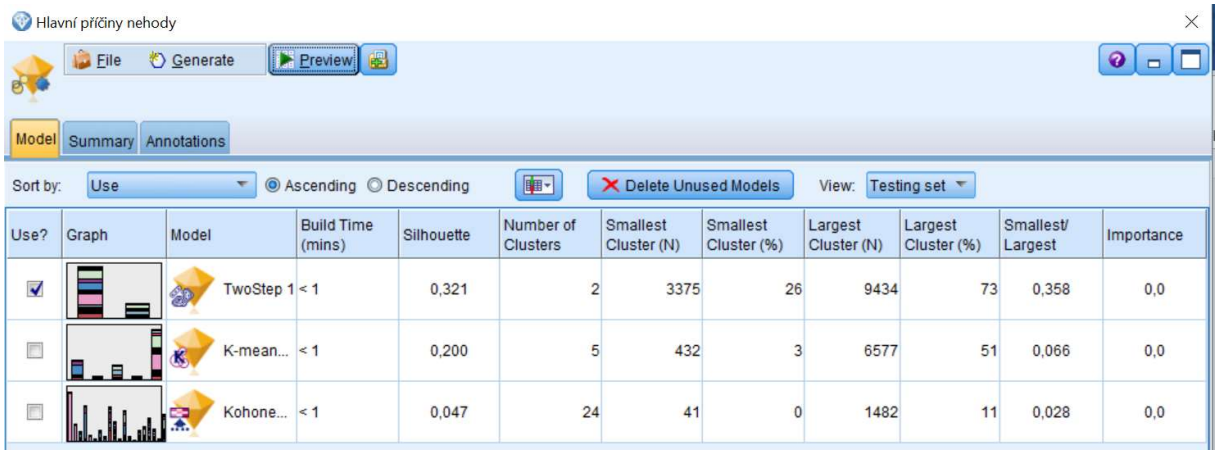
Při použití algoritmu TwoStep odděleně na stejná data bylo dosaženo stejných výsledků, algoritmus detekoval pouze dva shluky, přičemž kvalita shluků je nedostačující.

Jako vstupní parametry algoritus TwoStep použil následující atributy:

- FazeDne
- MESIC
- Den
- Směrové poměry
- Dělení komunikace
- Místo dopravní nehody
- Situování nehody na komunikaci
- Řízení provozu v době nehody
- Místní úprava přednosti v jízdě
- Druh křižující komunikace
- Usmrceno osob (počet)
- Těžce zraněno osob (počet)
- Lehce zraněno osob (počet)
- Zavinění nehody
- Druh nehody
- Druh srážky jedoucích vozidel
- Druh pevné překážky
- Hlavní příčiny nehody
- Stav komunikace
- Rozhledové poměry
- Specifické objekty v místě nehody
- Smyk
- Směr jízdy nebo postavení vozidla
- Stav povrchu vozovky v době nehody
- Povětrnostní podmínky v době nehody
- Viditelnost
- Počet zúčastněných vozidel
- Druh vozidla
- Výrobní značka motorového vozidla
- Škoda na vozidle (100 Kč)
- Kategorie řidiče
- Vnější ovlivnění řidiče
- KatHustotyProKom1Tr

Detailnější informace o Modelu, který byl vybudován pomocí algoritmu TwoStep jsou zobrazeny na obrázku 31 a 32 a v grafu 11.

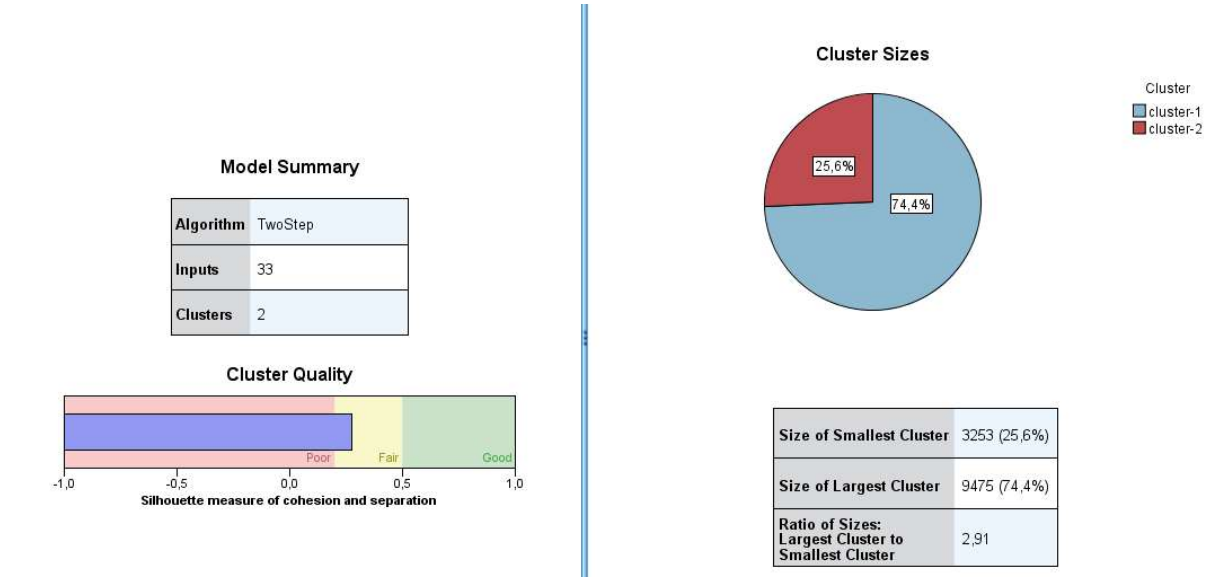
Obrázek 31: Výsledky automatického klasifikátoru



Use?	Graph	Model	Build Time (mins)	Silhouette	Number of Clusters	Smallest Cluster (N)	Smallest Cluster (%)	Largest Cluster (N)	Largest Cluster (%)	Smallest/ Largest	Importance
☑		TwoStep 1	< 1	0,321	2	3375	26	9434	73	0,358	0,0
☐		K-mean...	< 1	0,200	5	432	3	6577	51	0,066	0,0
☐		Kohone...	< 1	0,047	24	41	0	1482	11	0,028	0,0

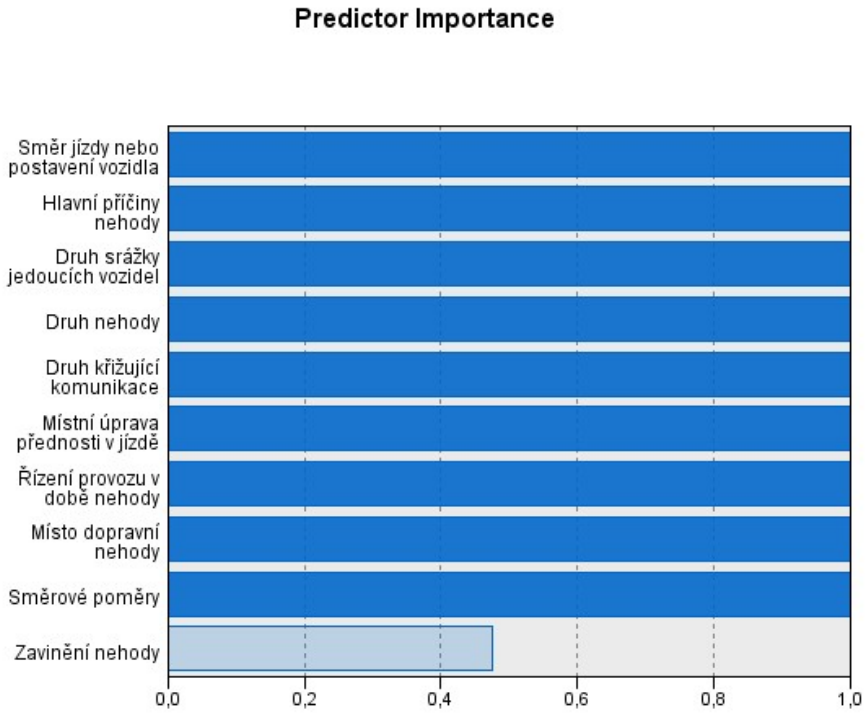
Zdroj: vlastní zpracování

Obrázek 32: Shrnutí modelu TwoStep



Zdroj: vlastní zpracování

Graf 11: Význam prediktorů



Zdroj: vlastní zpracování

Tyto první nepříznivé výsledky úlohy 2 naznačují, že hledání nebezpečných míst podobných vlastností bez informací o geografické poloze pravděpodobně není správnou cestou, kterou by bylo vhodné se vydávat. Hlavním důvodem jsou data, která máme v tuto chvíli k dispozici. I přes velký počet atributů popisujících aspekty dopravní nehody nelze vždy přesné okolnosti, za kterých se nehoda stala, z dat vyčíst. Příkladem, na kterém lze situaci ilustrovat, může být například dopravní omezení (rekonstrukce povrchu vozovky, rekonstrukce mostu či uzavírka), díky kterému se v určitém místě po určitou dobu stávají nehody častěji. V záznamu o dopravní nehodě může však být jako hlavní příčina nehody uvedeno například nedodržení bezpečné vzdálenosti a další. I tyto situace jdou však teoreticky detekovat a z dat odfiltrovat a v rámci výzkumu bude na tomto přístupu dále pracováno.

Nad daty, která máme k dispozici, jde však realizovat několik dalších typů data miningových úloh. Například by bylo možné z dat odhadovat místa nehod se stejnou hlavní příčinu nehody, snažit se odhalit místa tragických dopravních nehod či nehod způsobených zvěří a další.

Závěr

Se zvyšujícím se počtem nehod se policie opakovaně snaží zvrátit nepříznivý stav na dnešních silničních komunikacích pomocí represivních zásahů. S vědomím tohoto nepříznivého trendu by se měla společnost snažit přicházet především s novými nápady řešícími nastíněné problémy. Myslím si, že příznivé výsledky by měla přinášet především systémová řešení a využívání nových informačních a telekomunikačních technologií. V dnešní době disponujeme prostředky, které nám pomáhají budovat rychle obrovské datové základny. Využívání dat o dopravních nehodách společně s algoritmy pro získávání užitečných vzorů je dle mého názoru jednou z cest, jak by se silniční komunikace mohly stát bezpečnějším místem pro řidiče.

Datová základna využívaná v rámci této práce disponuje obrovským množstvím dat, která se dají využívat pro hledání skrytých vzorů a trendů. Databáze dopravních nehod, kterou vytváří Policie České republiky, však ne vždy postihuje všechny okolnosti, za kterých se nehoda stala. Tento fakt stěžuje detekci nebezpečných míst na vozovkách, která jsou něčím specifická.

Hledání shluků dopravních nehod je nedílnou součástí vývoje systému včasného varování před zvýšeným rizikem dopravní nehody. Po důkladné analýze byly vybrány a v práci popsány nejvhodnější metody a postupy shlukové analýzy pro řešenou problematiku. Shlukovací algoritmy založené na hustotě (DBSCAN, OPTICS) přinášejí dobré výsledky při vytváření shluků dopravních nehod založených na geografických datech, jelikož předpokládají existenci šumu (nehod, které nepatří do žádného shluku). Pro získávání lepších výsledků při detekci shluků je třeba do datové matice o nehodách získat informaci o hustotě dopravy. Tak je možné vyhledávat shluky odděleně v místech s rozdílnou hustotou dopravy. Sběr informací o hustotě dopravy provádí každých 5 let Ředitelství silnic a dálnic. Reálně je ale problém datové matice spojit, jelikož neexistuje jednoznačný klíč, pomocí kterého je to možné provést. V rámci této práce byl navržen způsob, který spojení obou databází řeší.

Hledání skrytých závislostí je možné provádět pomocí řady data miningových nástrojů. V rámci této práce byla provedena analýza nejvhodnějších DM nástrojů s ohledem na specifika řešené problematiky. Pro řešení byly využívány především IBM SPSS Modeler a software KNIME. Pro porozumění datům a jejich přípravu je efektivnější využít IBM SPSS Modeler. Při hledání skrytých závislostí byly používány oba nástroje. Výhodou open source nástroje KNIME je snadnější implementace vlastního algoritmu, čehož bylo při práci využito.

V rámci práce jsou testovány dva přístupy k detekci nebezpečných míst.

První přístup spočívá v prvotní detekci míst častých dopravních nehod. Vstupními parametry pro algoritmy, které jsou přímo určené na vyhledávání shluků v geo datech, jsou GPS souřadnice nehod. Odděleně je možné vyhledávat shluky v místech s velmi odlišnou hustotou dopravy díky informacím o počtu vozidel projíždějících daným místem za 24 hodin a typu silniční komunikace. Druhým krokem tohoto přístupu je detekce relevantních asociačních pravidel v rámci každého detekovaného shluku. Za tímto účelem je možné využít vlastního uzlu například v prostředí KNIME, který hromadně prochází jednotlivé shluky a hledá v nich asociační pravidla.

Druhý přístup řešený v rámci této práce je vyhledávání shluků podobných dopravních nehod s využitím většiny atributů popisujících aspekty dopravní nehody bez využívání GPS souřadnic. Tento přístup by mohl být využíván v případě budování nových komunikací, tak aby se předcházelo vzniku míst, která jsou z hlediska svého charakteru nebezpečná. V tomto případě je nutné důkladně analyzovat data tak, aby do algoritmů nevstupovaly atributy, které spolu korelují. Pro vyhledávání podobných nehod bylo využíváno algoritmů automatické klasifikace (IBM SPSS Modeler), který dokáže kombinovat několik shlukovačích algoritmů najednou. Samostatně bylo testováno vyhledávání skupin pomocí algoritmu TwoStep, který nevyžaduje na vstupu pevný počet shluků, jelikož částečně využívá hierarchického shlukování. Při vytváření modelů s tímto přístupem nebylo dosaženo prozatím uspokojivých výsledků.

Zásadním problémem druhého přístupu detekce podobných dopravních nehod je to, že data o dopravních nehodách nepostihují všechny aspekty, za kterých se nehoda stala. Tak nemohou být odhalena místa častých dopravních nehod, u kterých nejsou dostatečné informace popisující okolnosti nehody.

První přístup dovoluje vyhledat i místa, která jsou nebezpečná počtem výskytu dopravních nehod relativně k hustotě dopravy v daném místě. Taková místa nemusí být na první pohled ničím typická (nemáme o nich prozatím dostatečné informace). V konceptu systému včasného varování definujeme tato místa jako obecné shluky.

Modely založené na datech o dopravních nehodách a algoritmech pro odhalování skrytých závislostí by mohly najít využití i v autonomních automobilech.

V práci je definována řídicí a uživatelská část systému včasného varování před zvýšeným rizikem dopravní nehody. Diskutovány jsou i možnosti kooperace mezi řídicí a uživatelskou částí systému.

Nastíněný princip systému včasného varování by mohl přispět k prevenci častých dopravních nehod i předcházet tragickým nehodám. Navrhovaný systém by měl být užitečný především v situacích, kdy řidič jede konkrétní lokalitou poprvé. Hlavní výhodou navrhovaného systému včasného varování by mělo být především to, že se nehodám snaží předcházet, tj. neřeší situaci ve chvíli, kdy nehoda nastala, ale zaměřuje se na to, aby k nehodě vůbec nedošlo.

Navrhované postupy, techniky a samotný princip systému včasného varování je možné využít díky zobecnění, které je v práci taktéž prezentováno i na zdánlivě zcela odlišné typy úloh s jiným typem dat. Příkladem využití navrhovaného řešení může být vyhledávání skrytých závislostí v datech týkajících se kriminální činnosti. Budoucí vývoj navrhovaného systému včasného varování by se měl zabývat především samotnou implementací jednotlivých bloků řídicí a uživatelské části. Z hlediska uživatelské části, je dále vyvíjena mobilní aplikace realizující samotné varování. Dále se také počítá s dalším vývojem a prací na vylepšování vlastní aplikace umocňující provádět hromadnou a automatizovanou realizaci činností řídicí části systému.

Seznam použité literatury

Citace

Aktivní a pasivní prvky bezpečnosti motorových vozidel, 2015. Observatoř bezpečnosti silničního provozu [online]. Praha: Centrum dopravního výzkumu [cit. 2017-09-12]. Dostupné z: http://www.czrso.cz/clanky/aktivni-a-pasivni-prvky-bezpecnosti-motorovych-vozidel/

AGRAWAL, Rakesh a Ramakrishnan SRIKANT, 1994. Fast Algorithms for Mining Association Rules. In: Proceeding VLDB '94 Proceedings of the 20th International Conference on Very Large Data Bases [online]. San Francisco: Morgan Kaufmann Publishers, s. 487-499 [cit. 2017-04-29]. ISBN 1-55860-153-8.

ANKERST, Mihael, Markus M BREUNIG, Hans-Peter KRIEGEL a Jorg SANDER, 1999. OPTICS: Ordering Points To Identify the Clustering Structure. In: Proceedings of the 1999 ACM SIGMOD international conference on Management of dat [online]. Philadelphia: ACM [cit. 2017-04-29]. ISBN 1-58113-084-8. Dostupné z: http://www.dbs.ifi.lmu.de/Publikationen/Papers/OPTICS.pdf

BARNSLEY, Les, Susan LORD a Nikolai BOGDUK, 1994. Whiplash injury. Pain [online]. 58(3), 283-307 [cit. 2017-09-14]. DOI: 10.1016/0304-3959(94)90123-6. ISSN 0304-3959. Dostupné z: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00006396-199409000-00001

BENGLER, Klaus, Klaus DIETMAYER, Berthold FARBER, Markus MAURER, Christoph STILLER a Hermann WINNER, 2014. Three Decades of Driver Assistance Systems: Review and Future Perspectives. IEEE Intelligent Transportation Systems Magazine [online]. 6(4), 6-22 [cit. 2017-09-14]. DOI: 10.1109/MITS.2014.2336271. ISSN 1939-1390. Dostupné z: http://ieeexplore.ieee.org/document/6936444/

BEN-HUR, Asa, David HORN, Hava SIEGELMANN a Vladimir VAPNIK, 2002. Support vector clustering. The Journal of Machine Learning Research [online]. 2002, 2002(2), 125-137 [cit. 2018-09-01]. Dostupné z: http://delivery.acm.org/10.1145/950000/944807/2-125-horn.pdf?ip=147.230.232.216&id=944807&acc=OPEN&key=D6C3EEB3AD96C931%2E981AA4EB460CF8D4%2E4D4702B0C3E38B35%2E6D218144511F3437&__acm__=1537872235_3ff7c5a148e61c4fe4eb51ba6a104e5e

BERKA, Petr, 2005. *Dobývání znalostí z databází*. Praha: Academia. ISBN 8020010629.

BERTALANFFY, Ludwig, 1968. General system theory: foundations, development, applications. New York: George Braziller.

City Safety, 2014. Volvo Cars Support [online]. Praha: Volvo Cars Support [cit. 2017-09-14]. Dostupné z:
http://support.volvocars.com/cz/cars/Pages/owners-manual.aspx?mc=v526&my=2016&sw=15w46&article=dedd207c897617ddc0a801514bfcc7b4

CRISP DM: Data Mining Session 2, 2013. STATSOFT [online]. Palo Alto: Statsoft [cit. 2016-03-13]. Dostupné z: http://www.statsoft.com/support/blog/entryid/540/crisp-data-mining-session-2

DBSCAN pseudocode, 2011. GITHUB [online]. Kanwar Bhajneek [cit. 2017-08-01]. Dostupné z: https://github.com/KanwarBhajneek/DBSCAN

Dopravní nehody a jejich následky v roce 2017, 2018. Centrum služeb pro silniční dopravu [online]. Praha: Centrum služeb pro silniční dopravu [cit. 2018-09-19]. Dostupné z: https://www.cspsd.cz/707-dopravni-nehody-a-jejich-nasledky-v-roce-2017

Ecall becomes a reality!, 2015. European Emergency Number Association [online]. Brusel: European Emergency Number Association [cit. 2017-09-12]. Dostupné z: http://www.eena.org/press-releases/ecall-becomes-a-reality#.Wbg2A8hJZaS

ECall deployment - Publication by the European Commission of the Delegated Regulation No 305/2013, 2013. Harmonised eCall European Pilot - HeERO [online]. Brusel: Evropská komise [cit. 2017-09-12]. Dostupné z: http://www.heero-pilot.eu/view/en/media/news/20130415.html

ELKI [online], 2018. Germany: Ludwig Maximilian University of Munich [cit. 2018-07-29]. Dostupné z: https://elki-project.github.io/

ESTER, Martin, Hans KRIEGEL, Jorg SANDER a Xiaowei XU, 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In: PROCEEDINGS OF THE SECOND INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING [online]. AAAI Press, s. 1-6 [cit. 2017-04-29]. ISBN 978-1-57735-004-0.
Dostupné z: http://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf

Euro NCAP Advanced Reward 2011, 2011. Euro NCAP [online]. Leuven: Euro NCAP [cit. 2017-09-14]. Dostupné z: https://www.euroncap.com/en/ratings-rewards/euro-ncap-advanced-rewards/2011-ford-active-city-stop/

European Commission MEMO 13/547, 2013. European Commission [online]. Brusel: European Commission [cit. 2015-09-12]. Dostupné z: http://europa.eu/rapid/press-release_MEMO-13-547_en.htm

EVANS, Leonard, 2004. Traffic Safety. 1. Michigan: Science Serving Society. ISBN 978-0975487105.

HAN, Jiawei. a Micheline. KAMBER, c2011. Data mining: concepts and techniques. 3rd ed. Burlington, MA: Elsevier. ISBN 9780123814791.

IBM SPSS Modeler, 2016. IBM [online]. New York: IBM Corporation [cit. 2017-04-29]. Dostupné z: http://www-03.ibm.com/software/products/cs/spss-modeler

JONES, W.D., 2001. Keeping cars from crashing. IEEE Spectrum [online]. 38(9), 40-45 [cit. 2017-09-13]. DOI: 10.1109/6.946636. ISSN 00189235. Dostupné z: http://ieeexplore.ieee.org/document/946636/

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]. 1. CRISP-DM consortium: [cit. 2018-09-15]. Dostupné z: https://www.the-modeling-agency.com/crisp-dm.pdf

KNIME, 2017. KNIME [online]. Konstanz: University of Konstanz [cit. 2017-04-29]. Dostupné z: https://www.knime.org/

KOVÁČIK, Milan, 2016. Hledání shluků podobných dopravních nehod. Liberec. Diplomová práce. Technická univerzita v Liberci. Vedoucí práce Marián Lamr.

MADSEN, Henrik, 2008. Time series analysis. 1. Boca Raton: Chapman & Hall/CRC. Texts in statistical science. ISBN 978-1-4200-5967-0.

MAURER, Markus, 2012. Forward Collision Warning and Avoidance. MAURER, Markus. Handbook of Intelligent Vehicles. 1. London: Springer London, s. 657-687. DOI: 10.1007/978-0-85729-085-4_25. ISBN 978-0-85729-084-7. Dostupné také z: http://link.springer.com/10.1007/978-0-85729-085-4_25

Ministerstvo dopravy chce zvýšit pokuty za rychlost, 2015. ČTK. DENÍK.CZ. DENÍK.CZ [online]. ČTK, ČTK. Praha: VLTAVA-LABE-PRESS, 17.8.2015, s. 2 [cit. 2015-11-13]. Dostupné z: http://www.denik.cz/automoto-denik/ministerstvo-dopravy-chce-zvysit-pokuty-za-rychlost-az-trojnasobne-20150817.html

Orange [online], 2018. Ljubljana: University of Ljubljana [cit. 2018-08-29]. Dostupné z: https://orange.biolab.si/

PETR, Pavel, 2010-. *Data Mining*. Vyd. 3. Pardubice: Univerzita Pardubice. ISBN 9788073953256.

PETR, Pavel, 2014. Metody Data Miningu. Pardubice: Univerzita Pardubice. ISBN 9788073958732.

Prezentace výsledků sčítání dopravy 2010, 2011. ŘSD [online]. Praha: ŘSD [cit. 2016-09-20]. Dostupné z: http://scitani2010.rsd.cz/pages/informations/default.aspx

RapidMiner, 2017. RapidMiner [online]. Boston [cit. 2017-04-29]. Dostupné z: https://rapidminer.com/

ŘEZANKOVÁ, Hana, Dušan HÚSEK a Václav SNÁŠEL, 2009. Shluková analýza dat. 2., rozš. vyd. Praha: Professional Publishing, 218 s. ISBN 9788086946818.

Statistika nehodovosti, 2014. POLICIE ČR [online]. Praha: [cit. 2015-11-13]. Dostupné z: http://www.policie.cz/clanek/statistika-nehodovosti-900835.aspx

ŠEBEK, Michal, 2010. Dolování asociačních pravidel z databází a datových skladů. Brno. Akademická práce. Vysokého učení technického v Brně.

SIEGEL, Eric, 2016. Predictive analytics: the power to predict who will click, buy, lie, or die. Revised and Updated Edition. Hoboken, New Jersey: Wiley. ISBN 9781119145677.

SHMUELI, Galit, Peter C BRUCE, Inbal YAHAV, Nitin R PATEL a Kenneth C LICHTENDAHL, 2018. Data mining for business analytics: concepts, techniques, and applications in R. 1. Hoboken, New Jersey: John Wiley. ISBN 9781118879368.

Statistika nehodovosti, 2018. Statistika nehodovosti - Policie České republiky [online]. Praha: Policie ČR [cit. 2018-09-21]. Dostupné z: Statistika nehodovosti na pozemních komunikacích v ČR 2016, 2017. Autoklub ČR [online]. Praha: Autoklub ČR [cit. 2017-09-08]. Dostupné z: http://www.autoklub.cz/dokument/12022-statistika-nehodovosti-za-rok-2016.html

Statistika nehodovosti na pozemních komunikacích v ČR 2015, 2016. Autoklub ČR [online]. Praha: Autoklub ČR [cit. 2017-09-08]. Dostupné z: http://www.autoklub.cz/dokument/9698-statistika-nehodovosti-za-rok-2015.html

Statistika nehodovosti na pozemních komunikacích v ČR 2016, 2017. Autoklub ČR [online]. Praha: Autoklub ČR [cit. 2017-09-08]. Dostupné z: http://www.autoklub.cz/dokument/12022-statistika-nehodovosti-za-rok-2016.html

Toto je nejnebezpečnější zatáčka v zemi, 2016. PECÁK, Radek. Aktuálně.cz [online]. ČR: Economia [cit. 2016-02-18]. Dostupné z: http://zpravy.aktualne.cz/ekonomika/auto/toto-je-nejnebezpecnejsi-zatacka-v-cr-pred-rizikem-ted-u-pso/r~2977752cc9bc11e593630025900fea04/

TwoStep Cluster Node, 2012. IBM® IBM Knowledge Center [online]. USA: IBM [cit. 2017-08-31]. Dostupné z: https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/clusternode_general.htm

V srpnu zemřelo na silnicích 70 lidí, o 12 víc než loni, 2015. Aktuálně.cz [online]. ČTK. Praha: Economia, 1.9.2015 [cit. 2015-11-13]. Dostupné z: http://zpravy.aktualne.cz/domaci/v-srpnu-zemrelo-na-silnicich-70-lidi-o-12-vice-nez-loni/r~9d183314509e11e5adcb0025900fea04/

Vláda: Cena lidského života je 120 mil. Kč, 2016. Banky cz [online]. Praha: Top-in.cz, a.s [cit. 2018-09-01]. Dostupné z: https://www.banky.cz/clanky/vlada-cena-lidskeho-zivota-je-120-mil-korun/

VLK, František, 2006. Automobilová elektronika: systémy řízení podvozku a komfortní systémy. 1. Brno: Prof.Ing.František Vlk,DrSc., nakladatelství a vydavatelství, 308 s. ISBN 80-239-7062-3.

WEKA [online], 2018. Waikato: University of Waikato [cit. 2018-07-29]. Dostupné z: http://www.cs.waikato.ac.nz/ml/weka/

WINNER, Hermann, Bernd DANNER a Joachim STEINLE, 2009. Adaptive Cruise Control. Handbuch Fahrerassistenzsysteme [online]. 1. Wiesbaden: Vieweg+Teubner, s. 478. DOI: 10.1007/978-3-8348-9977-4_33. ISBN 978-3-8348-0287-3. Dostupné také z: http://www.springerlink.com/index/10.1007/978-3-8348-9977-4_33

WINNER, Hermann, 2012. Adaptive Cruise Control. Handbook of Intelligent Vehicles [online]. 1. London: Springer London, s. 613. DOI: 10.1007/978-0-85729-085-4_24. ISBN 978-0-85729-084-7. Dostupné také z: http://link.springer.com/10.1007/978-0-85729-085-4_24

Základní informativní výpis o nehodě, 2014. POLICIE ČR. Jednotná dopravní vektorová mapa [online]. Praha: PČR [cit. 2015-09-27]. Dostupné z: http://pcr.jdvm.cz/pcr/Reports.aspx?S_Type=01&S_LID=41aa962a-f5bb-4e2b-953d-c56b6ba94b63&S_IdNehoda=002100070013

Zdravý člověk má cenu 10 milionů korun, stanovili experti, 2014. Idnes.cz [online]. Praha: Mafra [cit. 2018-09-01]. Dostupné z: https://zpravy.idnes.cz/cena-lidskeho-zivota-je-10-milionu-d4b-/domaci.aspx?c=A140414_131414_domaci_hv

Bibliografie

BABCOCK, Brian, Shivnath BABU, Mayur DATAR, Rajeev MOTWANI a Jennifer WIDOM, 2002. Models and issues in data stream systems. In: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems - PODS '02 [online]. New York, New York, USA: ACM Press, 2002, s. 1- [cit. 2018-09-19]. DOI: 10.1145/543613.543615. ISBN 1581135076. Dostupné z: http://portal.acm.org/citation.cfm?doid=543613.543615

CIOS, Krzysztof a Roman SWINIARSKI, 2007. Data mining: a knowledge discovery approach. 1. New York: Springer Sience+Business Media. ISBN 978-0-387-33333-5.

CORMODE, Graham a Marios HADJIELEFTHERIOU, 2010. Methods for finding frequent items in data streams. The VLDB Journal. 19(1), 3-20. DOI: 10.1007/s00778-009-0172-z. ISSN 0949-877X. Dostupné také z: https://doi.org/10.1007/s00778-009-0172-z

MAYER-SCHÖNBERGER, Viktor a Kenneth CUKIER, 2014. *Big Data*. Brno: Computer Press. ISBN 9788025141199.

HOFMANN, Markus a Ralf. KLINKENBERG, 2013. RapidMiner: Data Mining Use Cases and Business Analytics Applications. 1. Florida: Taylor & Francis Group. ISBN 9781482205497.

MANKU, Gurmeet a Rajeev MOTWANI, 2002. Approximate frequency counts over data streams. In: Proceeding VLDB '02 Proceedings of the 28th international conference on Very Large Data Bases [online]. China: VLDB Endowment, s. 346-357 [cit. 2018-09-19]. Dostupné z: https://dl.acm.org/citation.cfm?id=1287400

SUBBA RAO, T., Suhasini SUBBA RAO a C. Radhakrishna RAO, 2012. Time series analysis: methods and applications. London: North Holland. Handbook of statistics, v. 30. ISBN 978-0-444-53858-1.

Publikace autora související s tématem disertační práce

LAMR, Marián a Jan SKRBEK, 2017a. A Systems Approach to Designing a Traffic Collision Avoidance Early Warning System. Journal of Systemics, Cybernetics and Informatics [online]. Orlando, USA, 15(Volume 15 - Number 3), 55-59 [cit. 2017-09-13]. ISSN 1690-4524. Dostupné z: http://www.iiisci.org/journal/sci/Contents.asp?Previous=ISS1703 (100 %)

LAMR, Marián a Jan SKRBEK, 2017b. Using Data Mining Tools for Retrieving Information from Databases of Traffic Accidents. In: IDIMT-2017 Digitalization in Management, Society and Economy: 25th Interdisciplinary Information Management Talks,. Austria: Trauner Verlag, s. 391-400. ISBN 978-3-99062-119-6. (100 %)

LAMR, Marián a Robin DVOŘÁK, 2016a. Application for visualization and analysis of traffic accident information. In: Proceedings of the 28th International Business Information Management Association Conference – Vision 2020: Innovation Management, Development Sustainability, and Competitive Economic Growth. Madrid: International Business Information Management Association, s. 1703-1709. ISBN 978-0-9860419-8-3. (70 %)

LAMR, Marián a Jan SKRBEK, 2016b. Real-time approaches to improving traffic safety. In: 2016 5th Mediterranean Conference on Embedded Computing (MECO) [online]. Bar: IEEE, s. 452-455 [cit. 2017-09-14]. DOI: 10.1109/MECO.2016.7525690. ISBN 978-1-5090-2222-9. Dostupné z: http://ieeexplore.ieee.org/document/7525690/ (100 %)

LAMR, Marián a Jan SKRBEK, 2016c. Searching for Traffic Accident Clusters to Increase Road Traffic Safety. In: IDIMT-2016 Information Technology, Society and Economy - Strategic Cross-Influence - 24th Interdisciplinary Information Management Talks. Linz, Österreich: Johannes Kepler Universitat Linz., s. 425-432. ISBN 978-3-99033-869-8. (100 %)

LAMR, Marián a Jan SKRBEK, 2016d. Traffic Data and Possibilities of their Utilization for Safer Traffic. In: Proceedings of the International Conference: Liberec Informatics Forum 2016. Liberec: Technical University of Liberec, s. 61-73. ISBN 978-80-7494-303-4. (100 %)

LAMR, Marián a Jan SKRBEK, 2015a. Advanced Approaches to Traffic Accident Prevention. In: IBIMA 2015 - Proceedings of the 26th International Business Information Management Association Conference. Madrid: International Business Information Management Association, s. 10. ISBN 978-0-9860419-5-2. (100 %)

SKRBEK, Jan a Marián LAMR, 2015b. Increasing effectiveness of early warning through smart ICT. In: Information and Communication Technology (ICoICT), 2015 3rd International Conference. 2015. Indonesia: IEEE, s. 160-165. ISBN 978-1-4799-7751-2. (100 %)

LAMR, Marián, David KUBÁT a Jan SKRBEK, 2015c. New Approaches to Smart Solutions for eliminating Car Accidents. In: Proceedings of the 12th International Conference Liberec Economic Forum 2015. 1. Liberec: Vysokoškolský podnik, spol., s. 392-401. (50 %)
ISBN 978-80-7494-225-9.

LAMR, Marián a Jan SKRBEK, 2015d. System Approach to Increasing Safety of Road Traffic. In: System approaches‘15 Interaction of soft and hard systems. Prague: University of Economics, Prague Publishing Oeconomica, 54 – 58. ISBN 978-80-245-2125-1. (100 %)

LAMR, Marián a Jan SKRBEK, 2015e. The Options for Actively Increasing Road Safety. In: IDIMT 2015: Information Technology and Society – Interaction and Interdependence – 23rd Interdisciplinary Information Management Talks. Linz, Österreich: Johannes Kepler Universitat Linz, 487 – 494. ISBN 978-399033395-2. (100 %)

Ostatní publikace autora

ANTLOVÁ, Klára, Petra RYDVALOVÁ a Marián LAMR, 2017. Motivation in the students´ start–ups. In: DIMT-2017 Digitalization in Management, Society and Economy: 25th Interdisciplinary Information Management Talks. Austria: Trauner Verlag, s. 79-84. ISBN 978-3-99062-119-6. (30 %)

CÍSAŘOVÁ, Klára, Marián LAMR, Přemysl SVOBODA a Pavel TYL, 2016. Attitudes of students to the lecture streaming and new elements of e-learning portal. In: Proceedings of the 28th International Business Information Management Association Conference – Vision 2020: Innovation Management, Development Sustainability, and Competitive Economic Growth. Seville: International Business Information Management Association, s. 2080-2087. ISBN 978-0-9860419-8-3. (25 %)

CÍSAŘOVÁ, Klára, Marián LAMR, Jan LOUFEK a Pavel TYL, 2015. Stačí technické prvky a nové technologie, aby byly nastavené rovné šance na vzdělávání pro studenty s SVP. In: Sborník příspěvků z X. ročníku z konference „vysokoškolské studium bez bariér". 1. Liberec: Vysokoškolský podnik, spol., s.r.o 59-69. ISBN 978-80-7494-228-0. (25 %)

HYBLEROVÁ, Šárka, Martina ČERNÍKOVÁ, Olga MALÍKOVÁ, Klára CÍSAŘOVÁ a Marián LAMR, 2017. Indicating Financial Health of Czech Companies with the Support of Modern Methods of Multidimensional Data Processing. In: SGEM 2017: 4th international multidisciplinary scientific conference on social sciences & arts, Albena, 24-30 August 2017 : [recenzovaný zborník]. Book 3 Science and Society, Vol. 5. Sofia: STEF92 Technology, 353 – 360. ISBN 978-619-7408-15-7. ISSN 2367-5659. (10 %)

LAMR, Marián, Klára CÍSAŘOVÁ a Jana VITVAROVÁ, 2015. Advanced Learning Space as an Asset for Students with Disabilities. Turkish Online Journal of Educational Technology [online]. 14(2), 10-14 [cit. 2015-06-03]. ISSN 2146 - 7242. Dostupné z: http://www.tojet.net/volumes/v14i2.pdf (50 %)

LAMR, Marián, Pavel TYL a Alena GREGOVÁ, 2014. Nabídka nových sw řešení na elearningovém portále ALS na TUL v Liberci. In: Sborník příspěvků z VIII. ročníku mezinárodní konference "Vysokoškolské studium bez bariér". 1. Liberec: Vysokoškolský podnik, spol., s.r.o, s. 58-66. ISBN 978-80-7494-066-8. (33 %)

LOUFEK, Jan, Marián LAMR a Klára CÍSAŘOVÁ, 2014. Využití richmedií pro studenty s poruchou zraku. In: Sborník příspěvků ze IX. ročníku konference „Vysokoškolské studium bez bariér“. Liberec: Vysokoškolský podnik spol., s.r.o, s. 36-44. ISBN 978-80-7494-169-6. (33 %)

RYDVALOVÁ, Petra, Klára ANTLOVÁ a Marián LAMR, 2017. Vztah studentů k podnikání: celosvětový průzkum GUESSS 2016 z pohledu studentů českých vysokých škol. Liberec: Technická univerzita v Liberci. ISBN 978-80-7494-379-9. (33 %)