



Požadavky na datové zdroje a jejich kvalitu pro zavádění systému predikce stavu dopravy na základě strojového učení

Schválená metodika

2023

T A
Č R

Tento projekt je spolufinancován se státní podporou
Technologické agentury ČR a Ministerstva dopravy
v rámci **Programu DOPRAVA 2020+**.

www.tacr.cz

www.mdcr.cz

Název: Požadavky na datové zdroje a jejich kvalitu pro zavádění systému predikce stavu dopravy na základě strojového učení

Poděkování:



Tato metodika s názvem Požadavky na datové zdroje a jejich kvalitu pro zavádění systému predikce stavu dopravy na základě strojového učení č. projektu CK01000139 vznikla se státní podporou Technologické agentury ČR a Ministerstva dopravy ČR v rámci Programu DOPRAVA 2020+.

Vydavatel:

Vysoká škola báňská – Technická univerzita Ostrava
Centrum dopravního výzkumu, v.v.i.
VARS BRNO a.s

Autoři:

Ing. Jan Rozhon, Ph. D
Ing. Igor Mikolášek
Ing. Jan Mynařík
Ing. Zdeněk Uherka
Ing. Zuzana Švédová, Ph.D.
Ing. Martin Bambušek
Ing. Jakub Šafařík, Ph.D
Ing. Jiří Novobilský
Mgr. Marek Ščerba

Oponenti:

Ing. Jan Vlčínský TamTam Research
doc. Ing. Zdeněk Lokaj, Ph.D., LL.M.

Metodika schválená:

Schvalující orgán, č.j. xxx

Název nakladatelství:

Centrum dopravního výzkumu, v.v.i.

Místo vydání:

Brno

Rok vydání:

2023

ISBN XXX-XX-XXXXX-XX-X

Dedikace

Tato metodika byla vypracována jako výstup projektu CK01000139 „Systém pro predikce vývoje dynamiky dopravních proudů založené na hluboké neuronové síti“ z Programu Technologické agentury ČR programu „DOPRAVA 2020+“. (Koordinátor: VŠB – Technická univerzita v Ostravě; partneři projektu: Centrum dopravního výzkumu, v.v.i., VARS BRNO a.s.). Projekt byl řešen od začátku roku 2020 do konce roku 2023.

Komu je metodika určena – působnost

Metodika se týká činností a aktérů souvisejících s přípravou dat pro možnost využití neuronové sítě pro predikci vývoje dopravního proudu. Je tedy určena pro správce pozemních komunikací, kteří disponují daty a informacemi a budou mít zájem využívat tyto nástroje, dále také pro soukromé subjekty, které disponují daty ze silniční dopravy a mají zájem na přípravě dat pro použití v modelu hluboké neuronové sítě.

OBSAH

1. Úvod	6
1.1. Cíl metodiky	6
2. Obecný popis metody strojového učení	8
2.1. Doplnující informace	9
3. Sběr dat	10
3.1. Kategorie dat.....	10
3.2. Statická data – Úseky silniční sítě a jejich fyzické vlastnosti	12
3.2.1. Segmenty – GN, TMC	12
3.2.2. Počty jízdních pruhů.....	12
3.2.3. Zatáčkovitost / křivolakost	12
3.2.4. Přítomnost exitu	13
3.2.5. Návrhové kapacity segmentů.....	13
3.3. Využívání pozemních komunikací – Intenzity dopravy z detektorů	13
3.3.1. Doporučení na zpracování dat	14
3.4. Využívání pozemních komunikací – FCD data.....	15
3.4.1. Doporučení na zpracování dat	16
3.5. NDIC události	16
3.5.1. Doporučení na zpracování dat	17
3.6. Meteodata	17
3.6.1. Aladin	18
3.6.2. Doporučení na zpracování dat	18
3.6.3. Snímky z meteoradarů	18
3.6.4. Doporučení na zpracování dat	19
3.6.5. Data z meteostanic	19
3.6.6. Doporučení na zpracování dat	20
4. Příprava dat pro strojové učení	21
4.1. Nástroje a techniky	21

4.1.1.	Software pro skriptování a statistickou analýzu	21
4.1.2.	Databázové platformy	22
4.2.	Úplnost dat a konzistence.....	24
4.2.1.	Shromažďování dat z více zdrojů	24
4.2.2.	Využívání dat v reálném čase	24
4.2.3.	Definice rozsahu a cíle sběru dat.....	25
4.2.4.	Zavedení standardizovaných metod sběru dat.....	25
4.2.5.	Monitorování činností sběru dat	25
5.	Základní obsahově-strukturální požadavky na dostupnost dat	25
5.1.	Požadavky na časovou granularitu dat	26
5.2.	Požadavky na prostorovou granularitu ukazatelů.....	27
5.2.1.	Požadavky k příčné granularitě systému	28
5.2.2.	Požadavky k podélné granularitě.....	28
5.2.3.	Požadavky ke struktuře dopravního proudu.....	28
6.	Formát a struktura dat	29
6.1.	Formát a struktura dat vstupních dat	29
6.2.	Formát a struktura dat vstupujících do neuronové sítě	29
6.3.	Formát výstupních dat.....	31
6.4.	Předzpracování dat	31
7.	Požadavky na dokumentaci datových zdrojů	31
7.1.	Dokumentace prostorové složky (geografického referencování)	31
8.	Kontrola a validace dat.....	32
8.1.	Postup pro validaci dat	32
8.1.1.	Ověření dat.....	32
8.1.2.	Profilování dat	32
8.1.3.	Čištění dat.....	33
8.1.4.	Standardizace dat.....	33
8.1.5.	Transformace dat	33
8.1.6.	Integrace dat	33

9. Faktory vstupující do modelů plynulosti dopravního proudu založených na strojovém učení.....	33
9.1. Faktory	33
9.2. Složitost datového modelu průběhu dopravních toků na silniční síti	34
9.2.1. Charakter pokrytí (segmentový, liniový a síťový model)	35
9.2.2. Časová granularita	35
9.2.3. Příčná prostorová granularita silniční sítě	35
9.2.4. Podélná prostorová granularita silniční sítě	36
9.2.5. Rozlišení dopravních druhů a kategorií vozidel	37
10. Srovnání novosti postupů	38
11. Ekonomické aspekty	38
12. Seznam použitých zkratk	39
13. Seznam použité související literatury	40
Obrázek 1: Mapa implementovaných detektorů na komunikaci D1.	15
Obrázek 2: Příklad části dat z meteostanic	20
Obrázek 3: Číselník meteostanic včetně informace o poloze	20
Tabulka 1: Vstupní parametry pro neuronovou síť a formát jejich interní implementace.....	30

1. Úvod

Hlavním předmětem projektu CK01000139 je zhotovení systému pro predikci chování dopravního proudu založené na hluboké neuronové síti. Predikce vývoje chování se provádí pro různé typy komunikací ve variantních situacích a v různých časových intervalech. Je založena na dostupných vstupech včetně obsáhlé databáze historických dat. Část týkající se přípravy dat pro možnost trénování modelů neuronové sítě je právě tématem této metodiky.

Vycházíme z toho, že **predikce vývoje intenzity dopravy** je nezbytnou součástí systémů řízení dopravy. Tato predikce má klíčový význam pro optimalizaci využití sítě pozemních komunikací, snižování dopravních kongescí a zvyšování bezpečnosti na silnicích [1]. V nedávné době se ukázalo, že metody strojového učení, zejména modely založené na konvolučních sítích, rekurentních neuronových sítích a hybridních modelech, poskytují slibné výsledky při predikci dopravního proudu.¹

Metoda strojového učení je oblast umělé inteligence, která se zabývá vývojem algoritmů a modelů, které dokážou analyzovat a interpretovat data a následně se naučit rozpoznávat a předpovídat určité vzorce a chování. Podstatné však je, že přesnost těchto modelů do značné míry závisí **na kvalitě a dostupnosti dat** použitých k jejich trénování. Stejně tomu tak je v dopravě.

Mezi typické úkoly strojového učení patří klasifikace, regrese, shlukování, detekce anomálií, predikce a doporučování [10].

Využití **neuronových sítí pro predikci vývoje dopravy** má několik výhod. Jednou z nich je schopnost sítě zpracovávat velké množství vstupních dat a najít v nich často i netušené **vztahy a souvislosti**. Model vytvořený neuronovou sítí mapuje jednotlivé vstupní parametry zastoupené v trénovací sadě na odpovídající výstupy. Model pak má určitou formu asociativní paměti, vysokou robustnost a toleranci k chybám. Výhodou je schopnost adaptovat se na nové podmínky, jako jsou například **změny** v dopravních vzorech, a vytvářet přesnější predikce na základě aktuálních dat.

Samotný model následně dokáže predikovat různé druhy dopravních událostí, jako jsou zpoždění, nehody, přetížení, uzavírky silnic a další.

1.1. Cíl metodiky

V rámci projektu bylo testováno několik modelů založených na různých technologiích od klasifikátorů založených na k-NN, rozhodovacích stromech a boostingu, po rozdílné typy i kombinace neuronových sítí (dopředné, konvoluční, rekurentní a jejich kombinace). Z výsledků testování jsou identifikovány postupy **pro úpravu stávajících, ale i nově vznikajících datových**

¹ Boukerche, A., & Wang, J. (2020). Machine Learning-based traffic prediction models for Intelligent Transportation Systems. *Computer Networks*, 181. <https://doi.org/10.1016/j.comnet.2020.107530>

sad. Dále metodika obsahuje **doporučení týkající se sledování kvality dat** tak, aby bylo možné hledat vztahy a souvislosti mezi nimi a využívat je [2].

Metodika má tyto základní cíle:

- Popsat informační základnu o procesu přípravy dat pro potřeby strojového učení.
- Poskytnout soubor obecných požadavků na kvalitu dat, které bude možné využívat pro jiné metody strojového učení než ty, jež byly použity v rámci modelu vzniklého v projektu.
- Prokázat, že predikce a prognózy, které budou založeny na těchto datech, budou mít očekávanou a konzistentní úroveň kvality.

2. Obecný popis metody strojového učení

Tato kapitola popisuje jednotlivé typy neuronových sítí a konstrukci datového modelu založeného na vstupních datech. Nejprve se zaměříme na popis jednotlivých neuronových sítí.

Architektura neuronové sítě je často určována na základě vlastností vstupních dat a požadovaných výstupů, přičemž může zahrnovat různé typy vrstev (jako jsou konvoluční, rekurentní nebo plně propojené vrstvy), které lze kombinovat do různých síťových struktur. Některé z nejčastěji používaných vrstev a architektur pro predikci dopravních proudů zahrnují:

- **Plně propojená neuronová síť**

Plně propojené neuronové sítě (nebo také vícevrstvé perceptronové sítě) jsou v kontextu dopravy vhodné jen pro některé úlohy. Využití plně propojených neuronových sítí pro predikci dopravy může být účinné v případech, kdy je k dispozici **dostatek trénovacích dat** a kdy jsou k dispozici jasně definované vstupy a výstupy. Například plně propojené neuronové sítě mohou být využity pro predikci průjezdových dob na základě historických dat o dopravě, o kalendářových událostech a dalších faktorech ovlivňujících dopravu.

Nicméně, pro složitější úlohy, jako je predikce zpoždění v reálném čase na základě aktuálního provozu na silnicích, mohou být potřeba sofistikovanější neuronové sítě, jako jsou **rekurentní neuronové sítě (RNN)** nebo **konvoluční neuronové sítě (CNN)**, které jsou schopny zpracovat sekvenční data a prostorové informace o dopravě [5].

- **Konvoluční neuronová síť**

Konvoluční neuronové sítě (CNN) jsou v kontextu predikce dopravy velmi užitečné pro zpracování prostorových dat, jako jsou snímky z kamerových systémů, a také pro zpracování sekvencí dat, jako jsou historická data o dopravě. Mezi konkrétní případy užití patří například:

- Predikce kongesce na silnicích na základě aktuálních informací o dopravě, které jsou získávány pomocí kamerových systémů. Tyto informace jsou zpracovány pomocí konvolučních vrstev, které jsou schopny rozpoznat vzorce v obrazech a identifikovat oblasti s vysokou hustotou provozu.
- Predikce průjezdových dob: Konvoluční neuronové sítě mohou být využity pro predikci průjezdových dob na základě kombinace historických dat a aktuálních informací o dopravě z dopravních detektorů a senzorů. Tyto informace jsou zpracovány pomocí

konvolučních a rekurentních vrstev, které jsou schopny zpracovat sekvenční data a prostorové informace o dopravě.

Využití konvolučních neuronových sítí v kontextu predikce dopravy umožňuje přesné a rychlé zpracování velkého množství dat a poskytuje cenné informace pro optimalizaci dopravního systému a zlepšení plynulosti a bezpečnosti provozu na silnicích [6].

- **Rekurentní neuronová síť**

Rekurentní neuronové sítě (dále jen RNN) jsou velmi užitečné pro predikci dopravy, protože jsou schopny zpracovávat sekvenční data a pamatovat si informace z předchozích kroků. V kontextu predikce dopravy mohou být rekurentní neuronové sítě využity pro několik úkolů, jako je predikce průjezdových dob, predikce hustoty provozu a identifikace anomálií v chování dopravy [7].

Jeden z nejčastějších způsobů, jak použít rekurentní neuronové sítě pro predikci dopravy, je využití Long Short-Term Memory (dále jen LSTM) sítí. LSTM sítě jsou speciálním typem rekurentní neuronové sítě, která umožňuje zpracování dlouhodobých závislostí v sekvencích dat a umožňuje lépe pracovat s různými časovými prodlevami.

Model pro predikci vývoje dynamiky dopravních proudů založený na hluboké neuronové síti může být naimplementován v jazyce Python a využívá mj. frameworků Tensorflow, numpy, pandas, dataframe, atd. Výsledný model kombinuje různé typy uvedených vrstev do formy hluboké neuronové sítě [8].

2.1. Doplnující informace

Jak již bylo zmíněno, predikce stavu dopravy pomocí strojového učení spočívá v **konstrukci datového modelu** založeného na **vstupních datech**, který je vytvořen s ohledem na různé faktory. Faktory je nezbytné řešit zejména proto, že představují relativně nezávislé jevy, které ovlivňují plynulost dopravního proudu. Nyní je vhodné obecně nastínit, jak je možné tyto faktory aproximovat výběrem vhodných ukazatelů z dostupných dopravních dat. Ukazatelem se rozumí jedna měřená nebo klasifikovaná veličina. Aby byl ukazatel pro predikční model použitelný jakožto vstupní faktor, musím mít vhodnou parametrizaci datové věty. Datová věta každého ukazatele musí pro potřeby zpracování společně jinými ukazateli v rámci datového modelu strojového učení obsahovat vhodný a navzájem kompatibilní časový a prostorový popis každého záznamu.

Vzhledem k tomu, že dopravní data vnikají pomocí různých technologií a technik, jsou dodávána různými výrobci, řídí se různými standardy či jinou doménově specifickou praxí, je vhodné dbát

na to, aby dopravní data bylo možné v podobě ukazatelů v metodicky standardizované struktuře parametrů použít i napříč technologiemi a technikami pořízení i jejich vyhodnocování.

Ze zkušeností získaných při pilotním provozu vzniklého modelu jsou navržena metodická pravidla, která se týkají následujících částí:

- **Sběr dat** - *rozmanitost dat*
- **Příprava dat** - *čištění a předzpracování*
- **Extrakce příznaků** - *transformace dat na formát, který neuronová síť může lépe zpracovat pro případ predikce dopravního proudu*

Jednotlivé části jsou v kapitolách metodiky dále popsány podrobněji.

3. Sběr dat

Tato kapitola nejprve uvádí strukturovaně výčet jednotlivých kategorií dopravních dat, které v případě dostupnosti je možné vybrat a následně připravit pro vstup do modelu neuronové sítě. V dalších podkapitolách se metodika věnuje popisu charakteristik jednotlivých datových sad s ohledem na dané téma a popisu jejich zpracování, který je potřeba provést před vstupem do modelu. Pro budoucí přípravu datových sad a definování požadavků na jejich kvalitu se jedná o podstatné doporučení, které vychází z provedených analýz současných dostupných databází. Na základě tohoto zkoumání je nutno vybrat referenční vzorek datových sad, který je vhodný pro cvičení vzniklého modelu neuronové sítě. Tento vzorek je vybrán také s ohledem na předpoklad, že se stane v budoucnu referenčním vzorkem pro další budoucí modely neuronové sítě.

3.1. Kategorie dat

Dopravní data lze dělit na statická, která se jednou za určitý časový úsek aktualizují a dynamická data, která se aktualizují v řádech milisekund. V následujícím seznamu jsou tato data uspořádána do uvedených kategorií. Toto uspořádání vychází z TP 172 (v době vzniku metodiky se jedná o návrh znění z roku 2023)

Pozemní komunikace – Úseky silniční sítě a jejich fyzické vlastnosti

- Úseky silniční sítě a jejich fyzické vlastnosti – informace obsahuje údaje o identifikaci (název, číslo), průběhu (geometrii, včetně stoupání a klesání), šířce, počtu jízdních pruhů, sklonu a typ povrchu pozemní komunikace.
- Klasifikace pozemních komunikací;
- Objekty na pozemní komunikaci;
- Křižovatky;

Pozemní komunikace – Součásti, příslušenství a vybavení

- Poloha a další údaje o silničních meteostanicích;
- Poloha a další údaje o dopravních detektorech;

Stav pozemních komunikací

- Uzavírky pozemních komunikací;
- Uzavírky jízdních pruhů;
- Práce na pozemní komunikaci;
- Dočasná opatření v oblasti řízení provozu (temporary traffic management measures);
- Uzavírky mostů a tunelů;
- Nehody a mimořádné události;
- Špatný stav pozemní komunikace;
- Povětrnostní podmínky ovlivňující povrch pozemní komunikace a viditelnost;
- Překážka provozu;
- Čekací doby;
- Nadrozměrný náklad;
- Nebezpečný náklad;
- Sjízdnost komunikací (zimní zpravodajství).

Využívání pozemních komunikací v reálném čase

- Stupeň provozu (level of service);
- Rychlost provozu;
- Poloha a délka kolon vozidel;
- Cestovní časy (travel time);

- Čekací doba na hraničních přechodech;
- Dostupnost volných parkovacích míst na odpočívkách.

Ostatní informace

- Data a informace z/do dopravních zařízení;
- Kalendářové události.

3.2. Statická data – Úseky silniční sítě a jejich fyzické vlastnosti

Statickými daty se v kontextu tohoto dokumentu rozumí data, která se v čase typicky nemění. V zásadě jde o data použitá k popisu silniční sítě a také k mapování ostatních dat.

3.2.1. Segmenty – GN, TMC

Segmenty jsou určeny k rozdělení sledované silniční sítě/komunikace na menší jednotky a pro mapování ostatních dat na tyto části. Primárně lze (pro potřeby predikcí) použít systém TMC. Jeho segmenty mají typickou délku jednotek kilometrů, což je pro potřebu predikcí přiměřené. Oproti tomu Global Network (dále jen GN) obsahuje segmenty kratší, obvykle na jeden segment TMC připadají nižší desítky GN segmentů. Nicméně v rámci Global Network lze k segmentům dohledat z našeho pohledu užitečné údaje, využili jsme proto, že je známo přesné mapování TMC -> GN a pro segmenty TMC jsme sestavili další data.

3.2.2. Počty jízdních pruhů

Počet jízdních pruhů, respektive změna počtu pruhů v podélném průběhu komunikace je veličina, která může ovlivnit dopravní proud. Pokud není počet pruhů v segmentu konstantní, je vhodné uvažovat s nejmenší hodnotou. Přídavné pruhy, vyjma stoupacího (odbočovací, připojovací), nemají na kapacitu vliv.

Z důvodu neexistence datových podkladů bohužel není možné zohlednit dočasná omezení počtu a šířky jízdních pruhů z důvodu dopravních uzavírek, neboť aktuální Centrální evidence uzavírek nedisponuje potřebnými strojově čitelnými záznamy v potřebné kvalitě. Toto omezení lze zpětně považovat za jeden z nejzásadnějších identifikovaných datových nedostatků. Obdobné lze uvést také k omezením v důsledku dopravních nehod. Počet jízdních pruhů, respektive kapacita komunikace, je v kombinaci s intenzitou dopravního proudu (stupeň vytížení) pochopitelně klíčovým prediktorem vzniku kongescí a snížení cestovní rychlosti.

3.2.3. Zatáčkovitost / křivolakost

Tento údaj se zpočátku jevil jako další slibný kandidát. Segment je v rámci GN zařazen do jedné z 6 kategorií zatáčkovitosti. Na dálniční síti jsou však prakticky všechny GN segmenty v nejnižší kategorii 0-1 %. Způsob výpočtu není v dokumentaci GN popsán. V případě zájmu či potřeby je možné spočítat konkrétní hodnotu křivolakosti (průměrná úhlová změna v gradech na kilometr)

dle TP 188 a pracovat s ní. S ohledem na celkově malou zatáčkovitost či křivolakost dálnice však mají oba parametry význam spíše v případě predikcí na komunikacích nižších tříd.

3.2.4. Přítomnost exitu

Existence exitu s rozlišením, zda jde o exit nájezdový či sjezdový (popř. zda segment obsahuje oba) na daném TMC segmentu může být dalším ze vstupů do modelů strojového učení. Podklady pro tento údaj jsou získány z TMC tabulek, kde pro každý bod jsou dostupné příznaky, zda lze v daném bodě najet či sjet, a to zvlášť v pozitivním a zvlášť v negativním směru. Směr (+/-) je vlastnost TMC segmentu. Na základě těchto údajů vznikla datová sada se dvěma údaji – nájezdový segment ano/ne a sjezdový segment ano/ne.

3.2.5. Návrhové kapacity segmentů

Kapacita segmentu, tedy počet vozidel, jež může segmentem projet za určitý čas (hodinu), je důležitým údajem. Tento údaj není dostupný v současných datových sadách. Z tohoto důvodu může být vypočten na základě dokumentu TP-188 („Technické podmínky“ - „Posuzování kapacity pozemních komunikací“, Ministerstvo dopravy, 08/2018). Do výpočtu vstupují podélný sklon, délka stoupání a podíl pomalých vozidel. Podíl pomalých vozidel může být uvažován s ohledem na jeho proměnlivost v průběhu dne střední hodnotou 15 %. Podélný sklon a délka stoupání jsou uvažovány v souladu s TP 188 následujícím způsobem:

Z Global Network byl vybrán průměrný sklon segmentu v daném směru. Úseky se sklonem $< 2\%$ jsou spojeny dohromady bez ohledu na reálný sklon. Pokud by byl sklon cca 2-3 %, GN segmenty mohou být zkombinovány (sečtena délka, zprůměrován sklon) do úseku s minimální délkou 500 m. Pokud za sebou následovalo více úseků se sklonem $> 3\%$, opět se stejným způsobem zkombinovaly.

Dle délky úseku, jeho průměrného sklonu a podílu pomalých vozidel lze ve výše uvedených tabulkách najít odhad kapacity daného úseku. V tabulce lze intrapolovat i extrapolovat, jak je uvedeno v TP. V rámci jednoho TMC segmentu takto může vzniknout více úseků s „konstantním“ sklonem, kapacita celého TMC segmentu je rovna nejmenší hodnotě.

V úsecích blízko velkých měst s velkým podílem pravidelně dojíždějících (u Prahy po křižovatku s D3, okolo Brna mezi Kývalkou a Rohlenkou, od Klimkovic po státní hranici) je vynásobena kapacita koeficientem polohy 1,06.

Získané údaje mohou být použity jako vstup do modelů strojového učení v kombinaci s intenzitou k výpočtu stupně vytížení.

3.3. Využívání pozemních komunikací – Intenzity dopravy z detektorů

Intenzita dopravy na segmentu je důležitý údaj přímo ovlivňující možnost vzniku kongesce (v kombinaci s kapacitou). Na sledovaných segmentech silniční síť se nachází stovky detektorů poskytujících informace o intenzitě dopravy. V České republice však ve většině případů nejsou data bohužel dostupná online. Jsou sbírána a zpracovávána v dávkách (v různých časových

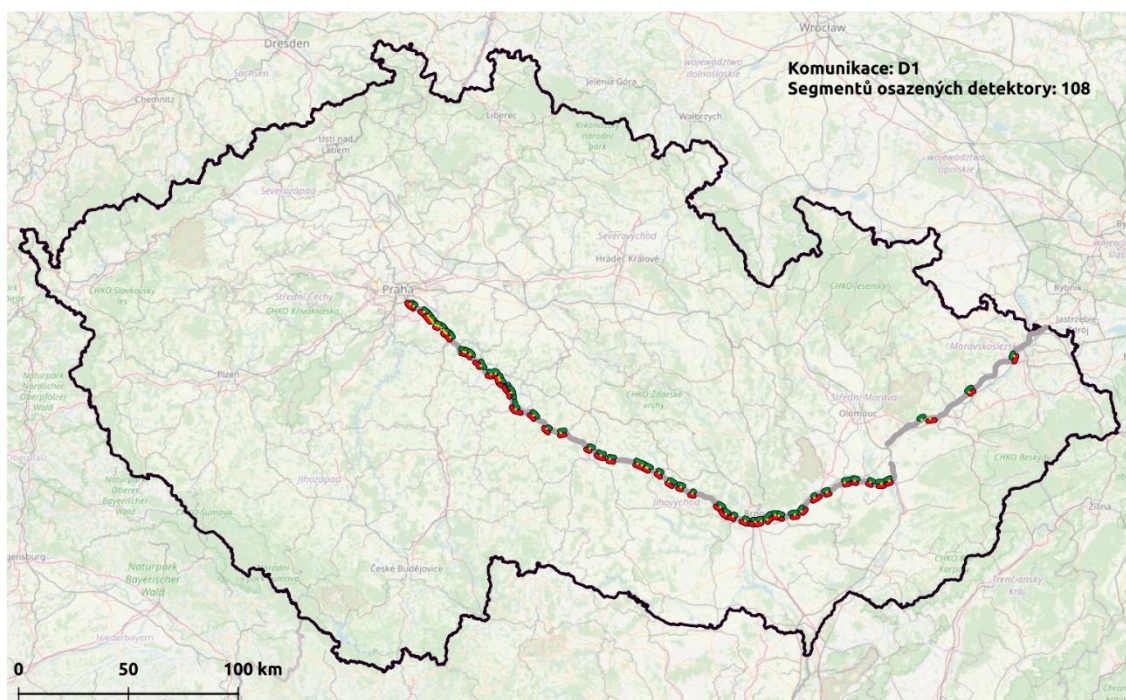
agregacích), což poněkud limituje jejich použití. Pokrytí dopravní sítě vč. dálnic je také značně omezené a nesystematické, což dále omezuje dostupnost dat, ačkoliv toto omezení lze, zejména v případě dálnic, částečně eliminovat interpolací a využitím naměřené intenzity pro celý mezikřižovatkový úsek.

3.3.1. Doporučení na zpracování dat

Data intenzit jsou aktuálně dostupná primárně ve dvou agregacích: hodinová a pětiminutová. Hodinová agregace je pro účely krátkodobých předpovědí příliš hrubý údaj. Údaje s agregací 5 minut jsou uchovávány pouze po omezenou dobu, což limituje možnosti pro natrénování modelů strojového učení. Dalším limitujícím faktorem je, že ne všechny segmenty jsou osazeny detektory. Zde je možno využít toho, že některé TMC segmenty na sledované síti jsou tzv. mezilehlé, tedy se na nich nenachází žádný sjezd ani nájezd. V rámci zpracování dat lze proto rozšířit zdroj dat intenzit z měřených segmentů i na po nich následující segmenty, pokud na nich není nájezdový segment. (Sjezdové segmenty jsou v rámci použitých TMC segmentů vždy na konci segmentu).

Samotná data intenzit obsahují (v závislosti na zdroji, který je poskytuje) zpravidla časový údaj začátku a konce měření, id detektoru, id smyčky (pruhu), kategorii vozidel, počet vozidel, jejich průměrnou rychlost a obsazenost. V databázi detektorů jsou dohledány jednotlivé detektory a metadata k nim a jsou následně namapovány na segmenty. Intenzita je sečtena ze segmentu a směru. Co se týká kategorií vozidel, s ohledem na dostupné kategorie v dalších zdrojích dat (viz oddíl věnovaný FCD datům) jsou namapovány kategorie poskytované detektory intenzit na dvě cílové kategorie, a to na kategorii 1. osobní a kategorii 2. nákladní vozidla. Pro jednotlivé kategorie je nasčítána intenzita.

Takto přiřazená a upravená data pak vstupují do procesu strojového učení.



- VD2 - Aktuální vypočtená rychlost dopravního proudu (km/h)
- VD3 - Aktuální čas průjezdu definovaného TMC segmentu (s)
- VD4 - Aktuální zpoždění na definovaném segmentu (s)
- VD5 - Typická průměrná rychlost při volném průjezdu segmentu (km/h)
- VD6 - Doba volného průjezdu (s)
- VD7_1 - Lokalizace kolony na segmentu
- VD7_2 - Délka kolony (m)
- VD7_3 - Vzdálenost čela kolony od počátku segmentu (m)
- VD8 - Míra spolehlivosti
- VD9 - Reakční doby systému (s)
- VD10 - Parametr Level of Service (LoS)

Obrázek 1: Mapa implementovaných detektorů na komunikaci D1.

3.4. Využívání pozemních komunikací – FCD data

Datová sada Floating Car Data (FCD) jsou data z flotilových vozidel (FVS) poskytovaná Ředitelstvím silnic a dálnic (ŘSD) a poskytují informace o dopravě až s minutovou granularitou. Data jsou modelována na základě pohybu flotilových vozidel, jejichž palubní jednotky minimálně jednou za minutu zasílají základní údaje o jízdě (pozice, směr, rychlost). Data jsou zpracována a vypočtena pro daný TMC segment a čas (minutu). Vypočtená data zahrnují např. aktuální

rychlost na segmentu, dobu průjezdu, přítomnost/nepřítomnost kolony apod. Konkrétní atributy lze vidět níže v seznamu, který nezahrnuje již zmíněné, tj. lokaci a čas.

Datová sada je užitečným zdrojem pro dopravní výzkum a plánování. Může být použita k analýze aktuálního stavu dopravy, předpovídání budoucí dopravy a identifikaci problémových oblastí. Datová sada může být také použita k vývoji nových dopravních řešení, jako jsou inteligentní dopravní systémy (ITS).

Datová sada může být poskytnuta formou dávky (archivu) dat pro celou ČR a pro maximální dostupné období. Datová sada je také dostupná v reálném čase a je blíže popsána v Národním registru dopravních informací: https://registr.dopravniinfo.cz/cs/sources/cz-ndic_ddr-fcd/.

3.4.1. Doporučení na zpracování dat

Predikce lze realizovat pomocí metricky Level-of-Service (LoS) ve formě podílu aktuální rychlosti na typické rychlosti při volném průjezdu, přičemž výsledná hodnota je omezena na interval mezi 0 a 1 včetně. Kromě okamžité hodnoty lze využít také historické průměry za různá období mapující typické chování dopravy na segmentu v odlišných časových horizontech (např. minulý týden či za celý minulý rok).

Datová sada FCD obsahuje i aktuální počet osobních a nákladních plovoucích (flotilových) vozidel na segmentu, takže je nutné srovnat tyto počty s naměřenými intenzitami z jiných datových sad, aby se ověřilo, jestli tyto „pseudointenzity“ jsou informací s dobrou důvěryhodností modelující skutečné intenzity dopravy. Pokud by se to podařilo ověřit, tak vzhledem k dostupnosti dat FCD by se jednalo o použitelná data.

Bohužel je možné, že počty flotilových vozidel na segmentu korelují s intenzitami dopravy naměřenými jinými prostředky pouze minimálně, často jsou navíc uváděné hodnoty zjevně vadné (řádově nižší či vyšší, než by odpovídalo reálnému provozu).

3.5. NDIC události

ŘSD poskytuje datovou sadu „Běžné dopravní informace“, která obsahuje lokalizované informace o nehodách, údržbách, významných událostech ve formě datových vět obsahujících časovou a prostorovou adresaci informace, konkrétní typ události (včetně číselného kódu)

a další atributy. Bližší informace lze opět nalézt v Národním registru dopravních informací: https://registr.dopravniinfo.cz/cs/sources/cz-ndic_ddr-common/.

3.5.1. Doporučení na zpracování dat

Základní zpracování spočívá v kategorizaci typu událostí za pomoci jejich číselných kódů do několika základních skupin. Známá množina možných číselných kódů událostí je rozdělena do těchto vyjmenovaných skupin:

- dlouhodobé práce (např. opravy),
- krátkodobé práce (údržba),
- meteo,
- nehody,
- ostatní.

Dále probíhá namapování událostí na TMC segmenty za pomoci stávající lokalizace uvedené v samotných datech (čísla postižených GlobalNetwork segmentů jsou příliš podrobná a je potřeba je převést na cílovou prostorovou identifikaci pomocí TMC segmentů).

Události je dále potřeba agregovat na TMC segment a minutu. Výsledná projekce zjednodušuje vstupní datovou sadu a sadu informací pro každý segment a minutu ve sledovaném čase na počet (v čase) platných událostí podle jednotlivých typů uvedených ve výčtu výše. A to tak, že pro segment X a minutu M je dostupná sada 5 přirozených čísel.

Tato sada je využitelná pouze pro specifické případy užití. Pro krátkodobé predikce vývoje není vhodná, neboť informace se do ní propíší až ex post po vzniku dané události, a tedy není možné v krátkodobé predikci reagovat na změnu dopravní situace, kterou vyvolaly [3].

3.6. Meteodata

Meteorologická data lze získat z více zdrojů, přičemž je nutno rozlišit, zda jde o skutečná (naměřená) data, nebo jde o předpovědní data. Do první jmenované skupiny patří data z meteostanic a radaru, do druhé pak data předpovědního modelu Aladin. Předpovědi jsou z principu vydávány s předstihem, a tedy je lze obecně využít pro online predikce. Co se týká reálných dat, zde nyní ani u jednoho výše zmíněného zdroje neexistuje možnost jejich odběru v reálném čase, jejich použití pro online předpovědi je tedy sníženo. Zde vidíme potenciál pro jejich větší využití v momentě, kdy by byla dostupná v reálném čase. Na základě dostupných dat

se však význam meteorologických informací v modelech ukázal jako velmi malý, přičemž obdobné závěry v osobních diskusích potvrzovali i kolegové ze zahraničí.

Obecně se jako nejperspektivnější ukázalo využití modelu Aladin, v první řadě díky dopřednému charakteru předpovědí, a dále také díky relativně vysoké úspěšnosti předpovědí.

3.6.1. Aladin

Aladin je předpovědní model atmosferických procesů, určený pro krátkodobé předpovědi. Předpovědi jsou vydávány typicky 4x denně, v 00, 06, 12 a 18 hod. Předpovědi jsou vydávány v krocích po jedné hodině, jejich platnost je pro hlavní termíny (0 a 12 hod UTC) na 54 hodin dopředu, pro vedlejší termíny (6 a 18 hod UTC) pak 24 hodin. Co se týká rozlišení, načítaná data z modelu Aladin mají v horizontálním směru rozlišení 4,7 km – myšleno vzdálenost mezi uzlovými body, pro které se počítají předpovědi. (Data pro novější období jsou pak dostupná v rozlišení 1 km).

Struktura a formát vstupních souborů

Data z předpovědního modelu Aladin jsou dodávána ve standardním meteorologickém formátu GRIB. V rámci souboru je uložena definice bodů (GPS souřadnice), ke kterým jsou pak v souboru uloženy příslušné hodnoty, takto jsou pokryty všechny parametry v souboru obsažené. První soubor v datové sadě s příponou _00 obsahuje pouze omezené množství atributů a není pro naše účely využitelný. Další soubory v daném dni pak dodržují jmenné schéma:

ALAD4cross_yyyymmddhh_ss.grb, kde

- yyyymmdd je den, pro který byla předpověď vydána
- hh je hodina vydání předpovědi: 00, 06, 12 nebo 18
- ss je step – konkrétní hodina, pro kterou předpověď platí

3.6.2. Doporučení na zpracování dat

Pro potřeby využití v predikčním nástroji je potřeba namapovat lokace jednotlivých předpovědí – tedy uzlové body – na TMC segmenty silniční sítě. Všechny dodávané soubory obsahují stejné uzlové body (například krok 4,7 km). Je možno vyhotovit pomocnou datovou sadu sestávající z polygonů okolo uzlových bodů a pomocí namapování atributů (id TMC segmentu) na polygon identifikovaný souřadnicemi vytvořit pomocný soubor s mapováním TMC segment – uzlový bod. Toto mapování je pak použito při načtení konkrétních údajů z datové sady Aladin pro zájmové segmenty.

3.6.3. Snímky z meteoradaru

Radarová data poskytovaná ČHMÚ jsou obrazová data z meteostanic, zobrazující hodnotu odrazivosti zachycenou radarem, jež je možno namapovat na hustotu srážek. Dostupná data jsou poskytována v časovém rozlišení po 5 minutách, referencovaném v UTC času.

Struktura a formát vstupních souborů

Ke každému časovému intervalu je vygenerovaný jeden png (obrazový) soubor s rozlišením 810x610 pixelů (příklad radarových dat viz obrázek 1). Soubor obsahuje oblast s obrysovou mapou ČR, dále pak histogramy hodnot na x a y souřadnici, datum a čas měření, a doplňující informace (copyright). Data jsou v obrázku zakódovaná pomocí barevné stupnice, kdy barva obrazového bodu v daném místě obrázku (uvnitř mapy) odpovídá určité intenzitě srážek, resp. jeho rozsahu, viz obrázek 2. Pro účely mapování těchto informací na TMC segmenty komunikace je třeba vytvořit vztah mezi body obrázku a souřadnicemi, a to v odpovídajícím kartografickém zobrazení (souřadnicovém systému). Každý pixel radarových dat pak odpovídá čtverci plochy mapy, který může obsahovat jeden či více segmentů zájmových komunikací. Výstupem zpracování každého radarového obrázku pro daný čas je pak seznam TMC segmentů a číselné hodnoty radarových údajů, získaných přepočtem barevné hodnoty pixelu na hodnotu číselnou. Radarový snímek hustoty srážek (resp. odrazivosti).

3.6.4. Doporučení na zpracování dat

Pro účely importu je třeba namapovat jednotlivé pixely obrázku na území ČR a identifikovat, na kterých pixelech se nachází informace o srážkách pro sledované segmenty sítě. Metodou georeferencování, kdy je využito obrysové mapy na radarových snímcích, je vytvořen modifikovaný obrázek a určeny kontrolní body. Následně je vytvořen pomocný obrázek, kde je v RGB hodnotě barvy zakódovaná souřadnice každého pixelu a na tento obrázek je aplikována transformace se stejnými kontrolními body, jako v případě původního radarového snímku. Posledním krokem pak je položení mapy se segmenty sítě přes popsání vrstvy a mapováním atributů jsou zjištěny RGB hodnoty v místě jednotlivých segmentů; z nich pak jsou zpětně dekodovány souřadnice pixelů odpovídající jednotlivým segmentům.

Všechny radarové snímky mají stejnou podobu (velikost/rozlišení), takže tyto údaje lze využít opakovaně při načtení všech zpracovávaných radarových snímků.

3.6.5. Data z meteostanic

Meteostanice instalované na silniční síti (SMS) poskytují v pravidelných intervalech reálná meteorologická data, jako je vlhkost vzduchu, typ a intenzita srážek, rosný bod, viditelnost, směr a rychlost větru a další. Vybavenost jednotlivých stanic konkrétními senzory se navzájem liší, stejně jako interval poskytování (zaznamenávání) hodnot – typicky se jedná o interval á 5 minut.

Struktura a formát vstupních souborů

Data z meteostanic je dodána v souborech s oddělenými hodnotami (csv formát), jeden soubor pro každou stanici (soubory jsou pojmenovány číslem meteostanice). Samotný číselník

Požadavky na datové zdroje a jejich kvalitu pro zavádění systému predikce stavu dopravy na základě strojového učení

meteostanic je k dispozici většinou rovněž jako soubor s oddělenými hodnotami, resp. Excel. Meteostanice jsou lokalizovány pomocí GPS souřadnic v EPSG:4326.

3.6.6. Doporučení na zpracování dat

Prvotním krokem je namapování meteostanic na jednotlivé TMC segmenty silniční sítě, toto mapování (číslo stanice -> segment) se následně používá při načítání dílčích dat ze souborů meteodat a k jejich dalšímu zpracování.

Stamp	AirTemp	DewPoint	RelHum	PrecipType	PrecipIntens	PrecipSum	SnowHeight	Visibility	WindSpeed	WindDir	WindSpeedMax	WindDirMax	Rai
1. 9. 2019 0:02:00	22,1	14,4	62 C		0	0	0	2000	1,7	167	3,3	210 -	
1. 9. 2019 0:08:00	21,9	14,5	63 C		0	0	0	2000	1,5	171	2,9	210 -	
1. 9. 2019 0:14:00	21,9	14,5	63 C		0	0	0	2000	1,5	167	2,9	212 -	
1. 9. 2019 0:20:00	21,8	14,5	63 C		0	0	0	2000	1,5	164	3,3	212 -	
1. 9. 2019 0:26:00	21,8	14,5	63 C		0	0	0	2000	1,6	174	2,7	223 -	
1. 9. 2019 0:32:00	21,5	14,5	64 C		0	0	0	2000	1,6	182	2,9	234 -	
1. 9. 2019 0:38:00	21,6	14,5	64 C		0	0	0	2000	1,8	171	3,3	234 -	
1. 9. 2019 0:44:00	21,4	14,5	65 C		0	0	0	2000	2	172	3,8	226 -	
1. 9. 2019 0:50:00	21,1	14,5	65 C		0	0	0	2000	1,7	162	3,8	244 -	

Obrázek 2: Příklad části dat z meteostanic

Id	Name	Latitude	Longitude	Elevation	Road	Stationing	ShortName
1	D4 Voznice 24,5	49,81784722	14,22175	377	D4	24,5	SC
14	D8 Nová Ves 1,0	50,19024722	14,44362222	287	D8	1	SC
15	D8 Nová Ves 17,0	50,30626389	14,32359167	175	D8	17	SC
16	D8 Nová Ves 29,0	50,39324722	14,22996111	226	D8	29	UL
17	D8 Nová Ves 36,5 most	50,44115556	14,15988333	167	D8	36,5	UL
18	D8 Nová Ves 37,3	50,44253333	14,15035	167	D8	37,3	UL
19	D8 Nová Ves 43,9	50,48011111	14,08390833	207	D8	43,9	UL
20	D0 Poříčany 62,5	50,08074167	14,59324444	241	D0	62,5	PH
21	D0 Poříčany 62,4 - most	50,08164167	14,59342778	241	D0	62,4	PH
22	D11 Poříčany 0,0	50,09886111	14,60086667	237	D11	0	SC
23	D11 Poříčany 3,0	50,10905278	14,63856389	268	D11	3	SC

Obrázek 3: Číselník meteostanic včetně informace o poloze

4. Příprava dat pro strojové učení

Tato kapitola stručně představuje proces přípravy dat, který vychází z postupů v rámci učení neuronové sítě. Pro získání přehledu jsou uživateli metodiky nejprve představeny některé důležité softwarové nástroje pro správu a analýzu dopravních dat. V dalších podkapitolách jsou popsány úpravy dat, využitelné pro jednotlivé datové sady a mohou se stát předlohou pro postup přípravy v budoucnu.

4.1. Nástroje a techniky

Tato podkapitola poskytuje kompaktní přehled základních softwarových nástrojů a metod určených pro správu a analýzu dopravních dat. Specificky se tato podkapitola zaměřuje na populární skriptovací a analytický software, systémy pro správu databází, běžné značkovací jazyky a nástroje pro zpracování textových a webových dat. V této části je kladen zvláštní důraz na bezplatné a/nebo open source nástroje, neboť tyto volby poskytují nejekonomičtější, nejvíce propojitelné a obecně nejvíce dostupné možnosti pro práci s alternativními dopravními daty.

4.1.1. Software pro skriptování a statistickou analýzu

V oblasti programovacích a skriptovacích jazyků a analytických nástrojů je dostupná rozsáhlá škála možností. Výběr optimální alternativy pro konkrétní aplikaci závisí na mnoha faktorech, včetně úrovně odbornosti vývojového týmu nebo jednotlivce, poměru mezi časem vývoje a provádění a rozsahu funkcí, které jsou aktuálně nabízeny analytickými knihovnami. V rámci této sekce jsou uvedeny některé oblíbené možnosti pro zpracování dat a provádění statistických analýz, s důrazem na nástroje, jejichž zdrojový kód je otevřený.

A. Python

Python je populární objektově orientovaný interpretovaný skriptovací jazyk, který se široce uplatňuje ve statistických výpočtech, strojovém učení, vývoji webových stránek a dalších aplikacích. Python byl navržen tak, aby usnadňoval rychlý vývoj a lidsky čitelný kód. Je dynamicky typovaný, relativně snadno se učí a je velmi flexibilní. Python je otevřený a bezplatný (i pro komerční aplikace) a prostřednictvím balíčků třetích stran podporuje širokou škálu nástrojů pro zpracování, analýzu a vizualizaci dat. Ve srovnání s jazykem R je Python spíše jazykem pro všeobecné použití a může být vhodnější, pokud se výkon stává problémem nebo pokud je třeba kód pro zpracování/analýzu dat integrovat do většího softwarového rámce. Mezi oblíbené balíčky pro analýzu dat a statistiku patří Pandas, Scikit-learn, Statsmodels a SciPy.

B. R

R je interpretovaný jazyk s otevřeným zdrojovým kódem, který se zaměřuje na statistickou analýzu a vizualizaci. Je široce používán při dolování dat (data mining), statistice a obecných

úlohách analýzy a vizualizace dat a většina jeho základních funkcí a balíčků třetích stran je vyvinuta statistickou komunitou. Jazyk R byl původně vyvinut pro analýzu dat a statistické modelování a je podporován velkou a aktivní komunitou vývojářů. Ve srovnání s jazykem Python nabízí R větší rozsah funkcí, pokud jde o nové statistické metody a nástroje pro analýzu specifické pro danou oblast a vizualizaci dat. Python je však flexibilnější a schopnější pro programování pro obecné účely. Mezi oblíbené balíčky R pro obecnou analýzu dat patří *dplyr*, *stats*, *forecast* a *ggplot2*. (zdroj: <https://dplyr.tidyverse.org/>)

C. MatLab

MatLab je komerční matematické výpočetní prostředí a interpretovaný programovací jazyk. Často je srovnáván s prostředím R kvůli jeho bohatým funkcím pro numerické a statistické výpočty, i když pro obecné programovací úlohy je v některých případech vhodnější prostředí MatLab. Balíčky v MatLabu jsou často vyspělejší a robustnější než jejich protějšky v R, ale software je poměrně drahý a každý další balíček představuje další licenční náklady. Výhodou komerčního produktu je však lepší dokumentace a podpora oproti open source alternativám. Jako univerzální programovací jazyk není MatLab tak flexibilní, interoperabilní a v některých případech ani tak rychlý jako Python. R i Python mohou dosahovat numerického výpočetního výkonu srovnatelného s MatLabem, i když ten se výrazně liší v závislosti na tom, jaké nástroje/balíčky jsou použity, a na schopnosti uživatele paralelizovat a optimalizovat kód.

4.1.2. Databázové platformy

A. MySQL

MySQL je velmi oblíbený software pro správu relačních databází (RDBMS), který spravuje společnost Oracle a který je k dispozici v komerční i bezplatné verzi (pod licencí General Public License). Komunitní verze s otevřeným zdrojovým kódem neobsahuje podnikové funkce, jako je šifrování nebo sofistikovaná správa vláken, avšak poskytuje snadno použitelné a poměrně výkonné řešení pro různé databázové aplikace. Aktuální verze podporují různé typy geoprostorových dat a prostorové indexování. Vzhledem k tomu, že nastavení a používání je poměrně jednoduché, doporučuje se komunitní edice MySQL pro ty, kteří se seznamují s používáním RDBMS, a obecně spíše pro transakční než analytické aplikace. Pro složitější databáze nebo databáze vyžadující sofistikovanější analytické funkce může být nutné přejít na výkonnější řešení, jako je PostgreSQL.

B. PostgreSQL

PostgreSQL patří mezi nejvýkonnější dostupné otevřené relační databáze, ačkoli není tak populární jako MySQL. Je široce využíván v komerčních aplikacích. PostgreSQL podporuje geoprostorové datové typy a indexování pomocí rozšíření PostGIS a nabízí vynikající analytické dotazovací schopnosti ve srovnání s jinými otevřenými alternativami (např. MySQL). PostgreSQL je doporučován jako relační databáze obecného použití a je vhodný zejména pro složitější databáze nebo ty, které vyžadují pokročilejší analytické funkce, přestože je trochu složitější na nastavení a používání. Navíc díky geoprostorovým schopnostem dostupným v PostGIS je

PostgreSQL považován za jedno z nejlepších řešení relačních databází pro správu a analýzu prostorových dat.

C. Komerční systémy řízení báze dat (DBMS)

K dispozici je řada velmi výkonných komerčních DBMS, z nichž některé poskytují velmi užitečné nástroje a značné výkonnostní výhody oproti open source alternativám. Například Microsoft SQL Server poskytuje vynikající nástroje pro verzování, sofistikovaný multithreading a samozřejmě vynikající podporu dalších produktů Microsoft, jako jsou cloudové služby Azure. Mezi další komerční možnosti pro podnikové prostředí patří Oracle RDBMS, IBM DB2 a Sybase ASE. Obecně jsou komerční DBMS poměrně drahé a pro mnoho uživatelů nemusí poskytovat významné výhody. Microsoft Access je levnější a uživatelsky přívětivou variantou, která se často používá v podnikovém prostředí, ale má řadu omezení (především limit velikosti databáze 2 GB), kvůli nimž je méně použitelný pro práci s velkými soubory dat. Komerční DBMS mají výhodu podpory poskytované výrobcem, ačkoli alternativy s otevřeným zdrojovým kódem mají také často velmi aktivní a nápomocnou uživatelskou základnu.

D. Správa dat NoSQL

NoSQL neboli "nejen SQL" je široký pojem zahrnující celou řadu databázových technologií. Obecně lze říci, že datové systémy NoSQL jsou navrženy tak, aby podporovaly správu a analýzu velmi rozsáhlých dat, datových toků s vysokou rychlostí a/nebo dat, která se nevejdou do konvenčního relačního modelu. Velká část současné generace distribuovaného softwaru a softwaru pro zpracování a správu dat NoSQL je navržena tak, aby umožňovala databázím spíše distribuovat než škálovat. To znamená, že namísto nákupu malého počtu velmi výkonných serverů lze zátěž ukládání a zpracování distribuovat na velký počet méně výkonných počítačů. Takové systémy jsou navrženy pro redundanci a snadnou obnovu, takže je možné pokračovat v normálním provozu i při ztrátě jednoho nebo dokonce několika strojů v síti.

Výhody datového systému NoSQL se liší podle aplikace a softwaru, ale obecně zahrnují rychlé škálování, vysokou datovou propustnost, vestavěnou redundanci a obnovu a cenové výhody spojené s možností postupného škálování namísto velkých počátečních investic do hardwaru. Tyto systémy však mají řadu vlastností, které mohou být v závislosti na aplikaci více či méně problematické. Datové systémy NoSQL například často kladou menší důraz na správu dat a ukládají data v nestrukturované podobě nebo ve formě proměnlivých schémat, což může vést k problémům s kvalitou a konzistencí.

E. Práce s webovými daty

Oblíbené datové formáty, se kterými se setkáváme při práci s webovými daty jsou XML, JSON, HTML a text v přirozeném jazyce. XML a JSON jsou lidsky i strojově čitelné formáty používané pro kódování objektů. Je důležité správné zadání adresy URL požadavku s požadovanými parametry pro načtení dat z kanálu RSS nebo rozhraní API a následného rozboru načtených dat, aby byla strojově čitelná. K načítání dat z rozhraní API lze použít knihovnu *requests* jazyka

Python. Data JSON lze zpracovat pomocí balíčku json jazyka Python, zatímco data XML lze analyzovat pomocí balíčku BeautifulSoup jazyka Python. V případech, kdy webové stránky neposkytují kvalitní rozhraní, lze k získání zdrojového kódu HTML webové stránky a získání požadovaného obsahu z něj použít screen scraping. Pro screen scraping lze použít balíček BeautifulSoup jazyka Python. Získávání relevantních informací ze špatně strukturovaných webových stránek pomocí screen scrapingu však může být poměrně obtížné. V takových případech lze použít regulární výrazy, které proces extrakce požadovaného obsahu z webové stránky usnadní.

4.2. Úplnost dat a konzistence

Úplnost dat spočívá v dostupnosti všech nezbytných datových prvků nutných pro trénování a testování strojového učícího modelu. To konkrétně zahrnuje poskytování úplných souborů dopravních a dalších relevantních dat, jež jsou určeny jako vstupy do modelu. Výstupy modelu by měly být použitelné pro různé aplikace.

Absence nebo nedostatečnost dat může vážně negativně ovlivnit přesnost prognóz a spolehlivost modelů. To má potenciálně závažné důsledky pro plánování a řízení dopravy. V případě, že chybí důležitá data, model nemusí správně zachytit skutečné vzory dopravního chování. To zase může vést k nepřesným předpovědím dopravních kongescí, zpoždění, a dokonce i k dopravním nehodám. Výsledkem mohou být nesprávná rozhodnutí a neefektivní strategie pro řízení dopravy.

Konzistence dat souvisí s trvalou stabilitou dat v čase. Je klíčové zajistit, že data, která jsou shromážděna a použita k trénování strojového učícího modelu, zůstanou v průběhu času stabilní. To je nezbytné pro to, aby model mohl na základě historických dat poskytovat přesné prognózy. Dosáhnout toho vyžaduje zavedení **standardizovaných metod** a protokolů pro sběr dat, které zaručí, že shromážděná data si udrží svou konzistenci.

4.2.1. Shromažďování dat z více zdrojů

Sběr dat z více zdrojů, jako jsou dopravní detektory, dopravní kamery, systém FCD a další, může pomoci zajistit úplnost dat. Použitím více zdrojů lze chybějící nebo neúplné údaje z jednoho zdroje nahradit jinými zdroji [4]. To může vést k přesnějším a spolehlivějším predikcím. Nahrazující zdroje dat by však měly poskytovat stejný typ dat v obdobné kvalitě a pokud možno stejné granularitě (agregaci), v opačném případě je zástupnost původního zdroje značně omezená.

4.2.2. Využívání dat v reálném čase

Data v reálném čase mohou poskytnout úplnější, a hlavně aktuální obraz o dopravním toku než predikce založené na historických datech. Data v reálném čase mohou zachytit změny v dopravním toku, jako jsou nehody, uzavírky silnic a dopravní zácpy, které mohou ovlivnit přesnost předpovědí dopravního proudu. Proto by se měla data v reálném čase používat vždy, když je to možné. Tím však není omezeno současné použití historických dat pro trénování

modelů, případně i pro samotnou predikci, zejména v podobě průměrných historických průběhů dané veličiny v případech, kde to dává smysl (LoS, intenzity apod.).

4.2.3. Definice rozsahu a cíle sběru dat

Prvním krokem k zajištění úplnosti údajů je jasné vymezení rozsahu a cíle sběru dat. To zahrnuje určení typu požadovaných dat, zdrojů, četnosti sběru dat a konkrétních veličin, které mají být měřeny. Obecně platí, že četnost a granularita dat by měla být co největší s ohledem na zanedbatelné náklady pro uchovávání dat v porovnání s přínosy jejich potenciálního využití. V případě specifické potřeby případu užití je vždy možné data agregovat, opačný postup však možný není a zvyšování kvality a spolehlivosti modelů vyžaduje kvalitní a detailní data.

4.2.4. Zavedení standardizovaných metod sběru dat

Standardizované metody sběru dat zajišťují, že data jsou shromažďována konzistentním a systematickým způsobem. Tím se snižuje riziko chyb a nesrovnalostí, které mohou vzniknout při použití různých metod sběru dat. Standardizované metody mohou zahrnovat používání šablon či datových formátů pro sběr dat, či jasné a specifické postupy pro zadávání dat pro osoby provádějící jejich sběr. Důležitým aspektem je rovněž jednotné nastavení používání metadat.

4.2.5. Monitorování činností sběru dat

Pravidelné monitorování činností sběru dat a jejich kvality může pomoci odhalit případné chyby nebo nesrovnalosti v údajích. Monitoring by měl rovněž zahrnovat pravidelné kontroly formulářů pro sběr dat. Současně by mělo být zajištěno, aby systémy sběru dat dodržovaly standardizované metody, a byly zajištěny pravidelné návštěvy na místě, aby se ověřila přesnost shromažďovaných údajů. Pravidelný pasport a kontrola na místě je vždy prevencí k nechtěným technologickým výpadkům. Časová posloupnost sběru dat je klíčovým aspektem zachování časových řad.

5. Základní obsahově-strukturální požadavky na dostupnost dat

Následující kapitola popisuje požadavky na přípravu datových sad, které jsou následně považovány za faktory, jež přímo ovlivňují funkci datového modelu predikce stavu dopravy pomocí strojového učení. Pro modelování predikce stavu dopravy na silniční síti je možné v ČR využít několik datových skladů. Nejvýznamnějším zdrojem je Silniční databanka ŘSD a regionální datové sklady a data dostupná přes volně přístupné portály [3].

Při tvorbě predikčních dopravních modelů, založených na strojovém učení je nutné splnit několik požadavků. Obecně platí, že je žádoucí data uchovávat v nejjemnější dostupné granularitě dané primárním zdrojem dat. V případě potřeby je vždy možné data agregovat do větších celků (v časové i prostorové dimenzi), avšak opačný proces není možný. Hodnota dat pro využití ve výzkumu a vývoji v oblasti dopravy přitom s ohledem na dramatický růst datových kapacit v posledním desetiletí významně přesahuje náklady na archivaci a poskytování

historických dat. Nedostatečná dostupnost, kvalita, spolehlivost a objem historických dat se přitom ukazuje jako jedna z největších překážek při vývoji nových inteligentních dopravních systému a výzkumu v oblasti dopravy [9].

5.1. Požadavky na časovou granularitu dat

Data musí být shromažďována v přiměřeném časovém rámci a s přiměřenou četností. To je důležité, protože vzorce dopravy se mohou v čase výrazně měnit a pro přesnou předpověď dopravního toku je nezbytné tyto změny zachytit.

Aby byly splněny časové požadavky, měly by zdroje dat poskytovat pravidelně shromažďované a aktualizované informace, nejlépe v reálném čase. Frekvence sběru dat by měla být dostatečně vysoká, aby zachytila změny dopravních vzorců, k nimž dochází v průběhu dne, týdne a roku. Údaje by měly být opatřeny časovou známkou, aby bylo zřejmé, kdy byly shromážděny, a měly by být uloženy ve formátu, který umožňuje snadný a efektivní přístup [12].

Kromě toho by údaje měly pokrývat dostatečný časový rámec, aby bylo možné vyvinout prediktivní modely, které jsou dostatečně robustní pro různé dopravní scénáře. Pro zachycení změn v dopravním toku mohou být například zapotřebí údaje shromážděné v různých ročních obdobích nebo za různých povětrnostních podmínek. A konečně, časové požadavky by měly zohledňovat také kvalitu dat, která se může v průběhu času zhoršovat v důsledku různých faktorů, jako jsou poruchy senzorů a jiných zařízení. Proto by měly být zdroje dat průběžně monitorovány, aby byla zajištěna jejich kvalita [11].

V případě časové granularity, je třeba aby pro jednotlivé datové zdroje šlo určit:

- **Elementární časově agregační okna (časový krok)** – doba, pro jakou je měřená charakteristika agregována, např. 5 minut. V případě stavových veličin (veličin bez kumulace v čase) se jedná o agregační okna, pro něž naměřená, externím modelem stanovená, či jiným výpočtem určená hodnota vztažená k časovému oknu představuje reprezentativní centrální hodnotu (např. rychlost dopravního proudu). V případě časově kumulativních veličin se pak jedná o kumulativní součet za elementární časově agregační okno – např. počet vozidel za 5 minut. Řada aplikací v ITS vč. některých detailních predikčních modelů však vyžaduje pro efektivní fungování spíše minutovou agregaci. Proto je doporučeno využívat pro uchovávání dynamických dopravních dat primárně tuto. Kratší agregační interval také pozitivně ovlivňuje časovou latenci dat, zejména těch poskytovaných v reálném čase – aktuální stav veličiny může být aktualizován v kratších časových intervalech a lépe tak reflektovat skutečnost.
- **Časová latence dat či časový předstih – časová dostupnost datových sad ve srovnání s ideální časovou dostupností dat v reálném čase** – doba dostupnosti datových záznamů na vstupu pro zpracování v systému predikce dopravy od okamžiku jejich vytvoření systémem vytvářejícím tyto datové záznamy. Např. v případě detektorů, které nejsou

připojeny on-line, je technická latence dána mezi okamžikem vzniku dat na detektoru a dobou, kdy jsou data ze servisního vozidla, které data manuálně stáhlo předána do systému zpracování – v praxi se může jednat až o období několika měsíců. I v případě dat dostupných v reálním čase však dochází ke zpoždění daném zpracováním a výměnou dat a zejména délkou agregačního intervalu. Data z předpovědí mají naopak časové určení dat v předstihu proti reálnému času.

- **Časové určení dat** – časový okamžik, pro nějž jsou data určena či považována za reprezentativní. Např. v případě předpovědi počasí je třeba vědět, zda se jedná o předpověď stavu na hodinu dopředu či na 24 hodin dopředu. Časové určení dat může být popsáno jak časem (např. + 1 h), tak rozpětím času (intervalem +1 h až +2 h) od okamžiku vzniku dat.
- **Časový rozsah dostupných historických dat** – v rámci trénování modelu je pro urychlení vybudování neuronové sítě vhodné začít pracovat s větším balíkem historických dat. V případě, že jsou historická data dostupná, je třeba uvést v jakém rozsahu (např. relativně - rok zpět) nebo absolutně pomocí datumu. Pro další vývoj predikčních modelů, ale i pro další výzkum a vývoj v oblasti dopravy je nezbytné uchovávat veškerá dostupná data v datových skladech a poskytovat je, nejlépe formou open source dat, případně na základě bezplatné licence všem subjektům za účelem výzkumu a vývoje, neboť široká báze kvalitních a dostupných dat je nezbytná pro libovolný výzkum.

5.2. Požadavky na prostorovou granularitu ukazatelů

Prostorové rozlišení určuje úroveň detailů, které lze v datech zachytit a analyzovat. V kontextu předpovídání dopravních toků na základě strojového učení pomocí neuronových sítí znamená vyšší prostorové rozlišení, že lze získat podrobnější informace o dopravních podmínkách, jako jsou dopravní zácpy, nehody a uzavírky silnic, což může zvýšit přesnost předpovědí dopravních toků. Nižší prostorové rozlišení nemusí tyto podrobnosti zachytit, což vede k méně přesným předpovědím. Proto je pro úspěch predikce dopravního proudu na základě strojového učení zásadní výběr zdrojů dat s vhodným prostorovým rozlišením. V případě prostorové granularity je třeba u všech datových zdrojů vstupujících do modelu neuronové sítě řešit následující možné vlastnosti:

5.2.1. Požadavky k příčné granularitě systému

- Datový zdroj je v rozlišení na jízdní pruhy a současně je skladebný na jízdní pás (např. detektory umístěné na všechny jízdní pruhy v jízdním pásu).
- Datový zdroj je v rozlišení na jízdní pásy – např. průměrná rychlost počítaná společně pro všechny jízdní pruhy.
- Datový zdroj nemá směrové rozlišení na jízdní pás, např. data o aktuální meteorologické situaci lze vztahovat k segmentům sítě bez ohledu na směr jízdy.

5.2.2. Požadavky k podélné granularitě

- Datový zdroj je bodové povahy ve vztahu k linii, např. data z detektorů (nejsou-li v párech úsekového měření) jsou umístěna v bodech na liniích a příslušejí k směrově určenému jízdnímu pruhu.
- Datový zdroj je liniové povahy, např. data z FCD jsou prostorově členěna do podoby liniových segmentů prostorově a směrově určených dvěma sousedními body TMC tabulek či příslušností ke Global Network segmentu.
- Datový zdroj je plošné povahy, např. data ze srážkového radaru v podobě rastru či proměnné vystupující z modelu předpovědi počasí v rozlišení na čtverce různé velikosti.

5.2.3. Požadavky ke struktuře dopravního proudu

- Datový zdroj rozlišuje kategorie vozidel, např. nákladní a osobní vozidla, případně rozlišuje jiné kategorie vozidel (například intenzita dopravního proudu vyjádřená v RPDI dle Celostátního sčítání dopravy).
- Datový zdroj nerozlišuje kategorie vozidel (např. rychlost dopravního proudu z FCD systému, který nerozlišuje kategorie vozidel).
- Datový zdroj je nezávislý na kategorizaci vozidel (např. počasí).

6. Formát a struktura dat

6.1. Formát a struktura dat vstupních dat

Data poskytovaná různými subjekty jsou dostupná zpravidla v různé struktuře či formátech. Mohou to být textové soubory s oddělenými hodnotami („csv“ soubory) - data z automatických sčítačů dopravy nebo meteostanic), může jít o obrázky (radarová data) nebo databáze. Pro další zpracování v rámci systému před vstupem do neuronové sítě je potřeba vstupní data předzpracovat, což zahrnuje:

- sjednocení jednotek rychlosti, času aj.,
- georeferencování v souladu s použitým systémem adresace,
- sjednocení agregací,
- doplnění chybějících hodnot (aproximace).

Konkrétněji jsou tyto činnosti popsány v kapitole 8.

S ohledem na výše uvedené není nutné ani možné, aby všechna vstupní data byla ve stejném formátu. Podstatné je, aby data byla v dostatečné kvalitě a rozsahu [13].

6.2. Formát a struktura dat vstupujících do neuronové sítě

Systém počítá se vstupy ve formátu numerických polí knihovny Numpy jazyka Python. Tato pole musí mít tabelární formát s tím, že systém si interně provede změnu dimenzionality a struktury dle zvolené architektury a jejích parametrů. Sloupce takto poskytnutých tabelárních dat musí obsahovat informace uvedené v tabulce níže.

Hodnoty, které jsou v této formě uloženy vycházejí z formátu a obsahu dat popsaných v předchozí podkapitole, avšak také respektují, resp. musí respektovat způsob, jakým jsou interpretovány uvnitř algoritmů strojového učení – v našem případě konvoluční neuronové sítě. To vede k nutnosti surová data předzpracovat i v případě, kdy již splňují formát popsaný v tabulce níže.

Jako příklad si můžeme uvést počty osobních automobilů na daném segmentu a v daném čase. Ačkoliv se jedná o numerický údaj, který může být předán v očekávaném formátu, hodnoty, které jsou v tomto případě předávány jsou omezeny až velmi vysokými čísly v řádu tisíců (pozn. bavíme se v teoretické rovině). Toto by vedlo k situaci, kdy dané okamžiky/segmenty by měly na rozhodování neuronové sítě výrazně nadvážený vliv a výstupy modelu by tak zkreslily. Z tohoto důvodu je třeba provést předzpracování vstupních dat, které plynou z dostupných datových zdrojů tak, aby zohledňovala interní fungování neuronové sítě. Pro náš příklad to znamená například využití nelineární funkce tak, aby počet automobilů byl namapován na rozsah $<0,1>$.

Konkrétní metody využitě při tvorbě modelu ke standardizaci a normalizaci vstupních dat jsou uvedeny v podkapitole 6.4.

Parametr	Očekávaný datový typ
Časová známka	Float32
ID segmentu	String (13 znaků)
FCD osobní automobily	Float16
FCD nákladní automobily	Float16
intenzita osobní automobily (smyčky)	Float16
intenzita nákladní automobily (smyčky)	Float16
NDIC nehody	Float16
NDIC meteo	Float16
NDIC krátkodobá údržba	Float16
NDIC dlouhodobá údržba	Float16
ALADIN teplota v bodě aproximujícím segment	Float16
ALADIN dešťové srážky v bodě aproximujícím segment	Float16
ALADIN sněhové srážky v bodě aproximujícím segment	Float16
ALADIN síla větru – zonální složka – v bodě aproximujícím segment	Float16
ALADIN síla větru – medicínální složka – v bodě aproximujícím segment	Float16
LoS – level of service	Float16

Tabulka 1: Vstupní parametry pro neuronovou síť a formát jejich interní implementace.

6.3. Formát výstupních dat

Výstupní data jsou prezentována jako vektor predikovaných hodnot LoS (float16) opět ve formátu Numpy pole.

6.4. Předzpracování dat

Jak již bylo uvedeno výše, inherentní vlastnosti algoritmů strojového učení vedou k preferenci dat s omezeným oborem hodnot v intervalu $<-1;1>$. Z tohoto důvodu je očekáváno předzpracování pomocí některé z normalizačních/standardizačních metod. V zásadě je možné využít kteroukoliv z dostupných metod, které povedou k tomuto cíli, nicméně v případě potřeby **využití předtrénovaného modelu** je potřeba zohlednit předzpracování.

Ty jsou:

- nelineární aktivační funkce pro počty vozů (neexistuje relevantní maximální hodnota),
- min-max normalizace pro LoS a meteorologické údaje (existuje relevantní maximální hodnota),
- využití tzv. dummy nebo sparse vektorů pro kategorická data.

7. Požadavky na dokumentaci datových zdrojů

Datové zdroje vstupující do systému predikce stavu dopravy musí být dostatečně popsány metadaty tak, aby je bylo možné jednoznačně interpretovat s ohledem na dimenze popsané v kapitole 5. V podkapitole je popsáno, o jaké podklady se jedná.

7.1. Dokumentace prostorové složky (geografického referencování)

Z hlediska geografického referencování datových sad je třeba zajistit jednoznačný popis dat tak, aby je bylo možné interpretovat vhodným způsobem v prostoru – zohlednit jejich vztah ke komunikaci s ohledem na směr jízdy, udat jejich určení podélné a příčné složky polohy na komunikaci. V případě, že se jedná o prostorové referencování pomocí jiného standardizovaného či obecněji používaného či naopak doménově specifického způsobu

referencování, je třeba přiložit k datovým sadám jako metadata i příslušné datové sady, které jsou součástí takového způsobu referencování, může se jednat např. o tyto podklady:

- TMC tabulky v příslušné verzi kompatibilní s datovými sadami, případně jejich odvozeniny, např. segmentace na bázi sekvencí sousedních bodů TMC tabulek.
- Sčítací segmenty – segmenty používané ve sčítání dopravy nebo bodovou vrstvu míst, ve kterých bylo prováděno sčítání a k nimž jsou vztažena data ze sčítání dopravy (sčítací profily).
- Mapové podklady, jejichž ID jsou používány pro prostorovou identifikaci jevů, např. mapové podklady Global Network v příslušné verzi kompatibilní s datovými sadami.
- Mapové podklady typu gridu, např. gridy používané v oblasti předpovědi počasí.

Tyto pomocné datové sady určené k referencování musí být rovněž dokumentovány tak, aby jejich prostorová interpretace byla možná a jednoznačná. Musí být zároveň zachována dostupnost starších verzí referenčních systémů z důvodu zajištění zpětné kompatibility a použitelnosti starších datových sad, které je využívaly.

V případě referencování nezávislého na předdefinovaných setech, např. dynamického referencování pomocí OpenLR či dynamického referencování pomocí Traces v případě C-ITS zpráv je třeba ve specifikaci způsobu referencování uvést způsob referencování např. odkazem na příslušnou normu.

Predikce stavu dopravy na základě strojového učení je založena na mnoha faktorech, které ovlivňují plynulost dopravního proudu. Tyto faktory je třeba zohlednit a sbírat data z různých zdrojů, aby predikční model byl co nejúčinnější a přesný.

8. Kontrola a validace dat

Dalším důležitým krokem v přípravě dat je kontrola a validace. Jedná se o zajištění přesnosti a spolehlivosti a zahrnuje to především kontrolu chybějících údajů, odlehlých hodnot a nesrovnalostí. Validační kontroly lze provádět pomocí statistických metod, nástrojů pro vizualizaci dat nebo ruční kontroly.

8.1. Postup pro validaci dat

8.1.1. Ověření dat

Tento krok zahrnuje potvrzení, že přijatá data jsou úplná, přesná a ve správném formátu. To lze provést kontrolou záhlaví dat a zajištěním, že hodnoty jsou v očekávaném rozsahu. Jakékoli zjištěné nesrovnalosti by měly být nahlášeny zdroji dat k opravě.

8.1.2. Profilování dat

Tento krok zahrnuje analýzu dat s cílem identifikovat případné chybné vzorce nebo odlehlé hodnoty. To lze provést zkoumáním rozložení hodnot dat, identifikací chybějících nebo

duplicitních údajů a hledáním nesrovnalostí mezi různými soubory dat. Tento proces pomáhá identifikovat potenciální problémy s kvalitou dat, které mohou ovlivnit přesnost predikčního modelu.

8.1.3. Čištění dat

Tento krok zahrnuje opravu nebo odstranění případných chyb, nekonzistencí nebo chybějících údajů. To lze provést ručně nebo pomocí automatizovaných nástrojů, jako je software pro čištění dat. Cílem je zajistit, aby data byla konzistentní a spolehlivá.

8.1.4. Standardizace dat

Tento krok zahrnuje zajištění standardizovaného formátu dat. Jedná se například o používání jednotných měrných jednotek nebo standardizaci formátů data a času. To usnadňuje porovnávání a analýzu dat z různých zdrojů. Je přitom možné společný základní standard dle možností konkrétního zdroje případně rozšířit o další parametry dostupné pouze pro daný typ zdroje. Vždy je přitom ale vhodné zachovat nezávislé základní standardizované parametry. Např. pokud umí všechny datové zdroje rozlišit nákladní a osobní vozidla, rozlišení v těchto kategoriích může být společným standardem, který obsahují všechny datové sady ze všech zdrojů, a pokud některý zdroj umí rozlišit různé typy nákladních vozidel, může datovou sadu rozšířit o dalších několik parametrů, které udávají hodnoty navíc také v rozlišení na více kategorií.

8.1.5. Transformace dat

Tento krok zahrnuje transformaci dat do formátu, který lze snadno analyzovat algoritmy strojového učení. To může zahrnovat normalizaci dat nebo jejich převod na specifický typ dat, například na číselnou nebo kategoriální proměnnou.

8.1.6. Integrace dat

Tento krok zahrnuje kombinaci dat z více zdrojů, aby se vytvořil komplexnější soubor dat. To může pomoci doplnit chybějící údaje a poskytnout přesnější reprezentaci dopravního toku.

9. Faktory vstupující do modelů plynulosti dopravního proudu založených na strojovém učení

Pro efektivní práci při vytváření modelu pro predikci stavu dopravy je potřeba zohlednit řadu faktorů. Tyto faktory jsou definovány na základě znalostí více variant modelů predikce plynulosti dopravy na bázi strojového učení.

9.1. Faktory

Jedná se o následující faktory, **kte**ré mohou ovlivňovat predikci plynulosti dopravního proudu:

1. Faktor plynulosti dopravního proudu – vytvořený výběrem z dostupných ukazatelů plynulosti dopravního proudu – rychlost, doba jízdy, zpoždění, level-of-service

- dopravního proudu počítaný jako procento okamžité rychlosti na rychlosti volného průjezdu.
2. Faktor dopravní poptávky – vytvořený výběrem z dostupných ukazatelů intenzity dopravního proudu (počet vozidel za jednotku času).
 3. Faktor skladby dopravního proudu – vytvořený výběrem z dostupných ukazatelů skladby dopravního proudu.
 4. Faktory možného snížení plynulosti dopravního proudu.
 - i. zásahy v důsledku lidské činnosti plánované – práce, omezení, údržba, ...
 - ii. zásahy v důsledku lidské činnosti neplánované – nehody a další incidenty
 - iii. zásahy přírody – množství srážek a jejich forma (déšť, sníh), vítr, riziko námrazy, snížená viditelnost, aj. (metedata)
 5. Faktor časových cyklů s možným vlivem na dopravu – kalendáře zohledňující den týdne (po, út, ..., ne), měsíc, státní svátky a dny sousedící se státními svátky, dny prázdnin (pro základní a střední školy), dny výuky vysokých škol.
 6. Faktor statických charakteristik analytických segmentů silniční sítě – sklon, křivolakost, nadmořská výška, alternativní trasy, počet a šířka jízdních pruhů, či dalších, které ovlivňují kapacitu pozemní komunikace.

V rámci této metodiky jsou uvažovány jen krátkodobé predikce plynulosti dopravy pomocí sekvencí stavů plynulosti podmíněné předchozími stavy a statickými veličinami. Dlouhodobé predikce vývoje dopravy nejsou uvažovány. Globální faktory vedoucí ke změnám mobility a mobilního chování jako např. cenové šoky paliv, restrikce mobility v důsledku snahy o zamezení šíření epidemie Covid19 nevstupují do modelu predikce stavu dopravy přímo, ale jen zprostředkovaně přes výše uvedené faktory.

9.2. Složitost datového modelu průběhu dopravních toků na silniční síti

Komplexitu datového modelu je možné rozdělit do 6 oblastí:

- Rozsah pokrytí – individuální segment či profil, soubor vybraných profilů či segmentů vybrané komunikace, celá síť komunikací s dostatečným objemem dopravy.
- Základní prostorový element – segment či profil, navzájem se ovlivňující segmenty v linii, navzájem se ovlivňující segmenty v síti segmentů v liniích.
- Charakter pokrytí indikuje, zda je dopravní infrastruktura v modelu vnímána jako soubor modelovaných profilů, segmentů seřazených do linií či propojená síť segmentů.
- Časová granularita.
- Prostorová granularita.
 - Podélná prostorová granularita síť indikuje, jak je řešena v modelu poloha jevů či jejich délka na komunikaci reprezentované geografickou linií.
 - Příčná prostorová granularita síť indikuje, zda je budovaný model v rozlišení jízdních pásů nebo jízdních pruhů.
- Kategorizace vozidel v dopravním proudu indikuje, zda je dopravní proud v modelu pojímán jako jeden celek nebo je možné jej dělit dle kategorizace vozidel. Navrhovaný

model může dosahovat různé úrovně složitosti. Navrhovaný model musí být sladěn s tím, jaké úrovně komplexity podporují jednotlivé faktory datového modelu.

9.2.1. Charakter pokrytí (segmentový, liniový a síťový model)

Nejzákladnějším modelem je model samostatných segmentů představujících izolované kusy linií komunikací bez jejich vazby na jiné segmenty. Rozšířením tohoto modelu je model segmentů s vazbou na sousední segmenty – segment předcházející a segment navazující (případně i více úrovní návaznosti). Další úrovní je model samostatných linií vybraných komunikací – sekvence segmentů v linii bez jejich vazby na jiné navazující a křížící komunikace. Nejkomplexnějším modelem je model linií propojených do sítí. Síťový model může mít hierarchickou síť propojených komunikací různé úrovně (dálnice, silnice 1. třídy, hlavní městské průjezdní komunikace).

9.2.2. Časová granularita

Data musí být shromažďována v přiměřeném časovém rámci a s přiměřenou četností. To je důležité, protože vzorce dopravy se mohou v čase výrazně měnit a pro přesnou předpověď dopravního toku je nezbytné tyto změny zachytit. Zvolená časová granularita musí odpovídat plánovanému účelu modelu a současně reflektovat dostupnost dat v požadované granularitě.

9.2.3. Příčná prostorová granularita silniční sítě

Základním modelem je model směrově rozlišených jízdních pásů – tj. jeden směrově rozlišený jízdní pás bez ohledu na počet jízdních pruhů v daném směru je reprezentován jednou linií.

Detailnějším modelem (modelem příští generace) bude model jízdních pruhů členěných v rámci směrově rozlišených jízdních pásů. Očekávat kódování datových sad v důsledném rozlišení na jízdní pruhy standardizovaným způsobem se jeví v okamžiku tvorby tohoto dokumentu jako poměrně předčasné. Nicméně data z některých stávajících či potenciálních zdrojů již mohou být v rozlišení do jízdních pruhů.

Příčná granularita do úrovně jízdních pruhů je systémově ošetřena v rámci C-ITS norem, např. v případě C-ITS zpráv CAM a DENM v nepovinném („optional“) datovém elementu LanePosition. Tento systém je tedy relativním topologickým systémem počítajícím jízdní pruhy z vnější strany cesty pro daný směr.

Zárodky možnosti určit příčnou prostorovou granularitu jsou již ve starších systémech prostorového referencování dopravních událostí, např. ve velmi staré normě ALERT-C určené původně pro systémy na bázi RDS-TMC, ale pouze volitelně a velmi výběrovým způsobem v podobě možnosti volby vhodného kódu s upřesněním do úrovně jízdních pruhů z mnoha možných kódů nevyžadujících upřesnění do úrovně jízdních pruhů dle uvážení operátora

zadávacího informaci vybíranou ze standardizovaného EventListu (např. volbou kódu „levý jízdní pruh uzavřen“).

Rovněž možnost vytvoření datového modelu v rozlišení na jízdní pruhy v podobě vhodného mapového podkladu není zcela dořešeno a standardizováno. V navigačních mapách jsou sice komunikace se směrově oddělenými jízdními pásy odlišeny od ostatních komunikací (tzv. dual carriageway), kdy každý jízdní pás je reprezentován samostatnou geometrií, nicméně samostatnou geometrií nedisponují jednotlivé jízdní pruhy tvořící jeden jízdní pás. Vícepruhové uspořádání na komunikaci je řešeno zejména před křižovatkou pro lepší navigaci vozidla do vhodných jízdních pruhů, mimo křižovátku systémově na úrovni samostatné geometrie středové linie jízdního pruhu řešeno není. Lze očekávat, že v souvislosti s dalším rozvojem pokročilých asistenčních systémů řidiče ADAS a systémů autonomního řízení vozidel může dojít k vytvoření a zavedení nových digitálních modelů silniční sítě v plném rozlišení na jízdní pruhy.

9.2.4. Podélná prostorová granularita silniční sítě

Prostorová granularita podélná může být odvozena zejména od standardů a zvyklostí v oblasti lokalizace dopravních nebo infrastrukturních fenoménů na silniční síti případně může budovat vlastní systém segmentace či pracovat s dynamickým modelem bez segmentace. V rámci ČR se nabízejí pro vytvoření modelu následující možnosti podélné granularity:

- Podélná granularita využívající dostupných segmentací
 - Sčítací segmenty – sčítací segmenty jsou používány k referencování dat z Celostátního sčítání dopravy. Zatímco na dálnicích mají sčítací segmenty topologii ve struktuře z jednoho zobecněného nájezdu/exitu do druhého zobecněného nájezdu/exitu, mimo dálnice jsou segmenty spíše delší, ukončené jen ve významnějších sídlech a větších křižovatkách.
 - TMC segmenty – TMC segmenty jsou na dálnicích zpravidla kratší než úseky exit-exit, mimo dálnice nepokrývají všechny křižovatky a mohou být z hlediska délky velmi různorodé – od několika kilometrů až po několik stovek metrů. Obecně jsou však kratší než sčítací segmenty. Pokrývají jen dopravně významné komunikace.
 - Segmenty digitálního vektorového modelu silniční sítě Global Network – segmenty jsou výrazně kratší než TMC segmenty a sčítací segmenty. Jsou ukončeny tam, kde to vyžaduje topologický a datový model routovatelné mapové vrstvy silniční sítě, tj. pokrývají celou veřejnou silniční síť (známou k datu poslední aktualizace). Nicméně tato datová sada je proprietární – touto datovou sadou disponuje v sektorové licenci ŘSD.
- Vlastní segmentaci vytvořenou na míru systému predikce dopravy – tvůrce systému predikce stavu dopravy může s ohledem na vývoj modelovacích nástrojů a omezení dostupných segmentací dospět k potřebě vytvořit si vlastní systém segmentace.
- Systém bez segmentace – tvůrce systému predikce stavu dopravy může s ohledem na vývoj modelovacích nástrojů dospět k systému, který pracuje bez segmentace, případně

s nějakým interně-analytickým systémem dynamické tvorby segmentace např. v kroku 100 m.

- Systémy dynamického referencování pro liniové jevy – např. metoda OpenLR, která byla vytvořena jako systém popisu dopravních událostí a na dopravní síti nezávislá na mapovém podkladu konkrétního výrobce. V rámci tohoto systému jsou dopravní jevy prostorově referencovány pomocí sekvencí geografických souřadnic (x,y) a směrů lomených čar, které tyto sekvence bodů v geografických souřadnic propojují. Tento systém umožňuje jak referencování liniových jevů v celém jejich průběhu, tak referencování bodových jevů včetně jejich směrovosti ve vztahu k silniční síti. Systém dynamického referencování pro liniové jevy tak může být vhodným převodním rozhraním prostorového referencování pro modelovací systém bez segmentace.
- Systémy dynamického referencování bodových jevů v rámci C-ITS – metoda „traces“ – sekvence GPS poloh ve vhodné, normou předepsané kadenci časoprostorových rozestupů mezi jednotlivými body, kterými se vozidlo může v rámci dopravní sítě ze všech možných směrů dojet do koncového bodu sekvence, jenž představuje místo události. Tento systém neřeší liniové jevy – řeší pouze lokalitu, kde se řidič či vozidlo setká s dopravním jevem, nikoliv kde pak dopravní jev po najetí do něj pomine. Systém dynamického referencování pro bodové jevy pomocí „traces“ tak může být vhodným převodním rozhraním prostorového referencování pro datové zdroje bodového charakteru.

9.2.5. Rozlišení dopravních druhů a kategorií vozidel

Komplexní dopravní modely mohou zahrnovat více dopravních druhů, které se vzájemně ovlivňují a interagují jak přímo (železniční přejezd, tramvajová linka, autobusová doprava), tak nepřímo (přepravní konkurence – rozšíření přepravních kapacit, a naopak přetahování o přepravované osoby či zboží). Pro modely strojového učení pro krátkodobou predikci cestovních rychlostí, je však významnější míra rozlišení různých kategorií vozidel. Některé modely nemusí vozidla rozlišovat vůbec, avšak s ohledem na zásadně odlišné charakteristiky a chování osobních a nákladních vozidel se jedná o značné zjednodušení, které nelze běžně doporučit, což je současně vhodné brát v potaz při vytváření nových datových zdrojů. Naopak rozlišení a přesnost modelu je dále možné, přinejmenším teoreticky, dále zvyšovat podrobnějším členěním vozových kategorií, avšak zde již v praxi často narážíme na omezení daná rostoucí komplexitou modelu a praktické možnosti sběru dat. Například Celostátní sčítání dopravy prováděné manuálně rozlišuje až 12 kategorií vozidel, avšak jeho výsledky jsou zatíženy velkou chybou a je prováděno pouze krátkodobě po dobu několika hodin jednou za pět let na vybraných místech silniční sítě. Řada z těchto kategorií však má minimální praktické využití – jednak obsahuje zpravidla jen několik jednotek vozidel a jednak se svými dopady a významem liší jen minimálně od některé jiné kategorie. Automatické detektory a dopravní sčítače naopak nejsou schopny rozlišit tolik kategorií, avšak jsou schopné provádět kontinuální sčítání. V závislosti na požadovaných vlastnostech, přesnosti, či počtu sčítaných kategorií vozidel jimi lze

obsadit poměrně značnou část dopravní sítě, provádět kontinuální měření a v kombinaci s jinými datovými zdroji vytvářet také dopravní modely odhadující intenzity na zbývajících úsecích dopravní sítě. S ohledem na výše uvedené, zejména praktickou využitelnost pro většinu aplikací a potenciální dostupnost dat lze zhodnotit, že optimální kompromis pro kontinuálně sčítaná data je někde mezi 4-8 kategoriemi, které jsou děleny na základě délky vozidel, která má přímý vliv na hustotu dopravního proudu.

10. Srovnání novosti postupů

Provozování řídicích a informačních dopravních systémů je ve znamení převratných změn a nutného hledání nových přístupů a způsobů jejich integrace do stávajících procesů. Především dostupnost datových zdrojů a analytika nad daty zaznamenala v posledním období bezprecedentní rozvoj. Nové přístupy v řízení dopravy a poskytování dopravních informací (CAD) si tak mnohem více než dříve vyžadují integrovaný přístup, kdy je nutné propojovat informace o aktuálním stavu dopravy, předpovědi počasí, stavu vozovky, uzavírkách pro podporu efektivního rozhodování operátorů dopravy a zejména pro ovlivňování chování řidičů.

V metodice jsou řešeny postupy pro úpravu současných dat tak, aby byly připraveny ke komplementaritě s modelovacími nástroji. Především se jedná o predikci vývoje dopravy na páteřních komunikacích, který by zahrnoval takovou šíři proměnných hodnot (včetně počasí, uzavírek, nehod).

11. Ekonomické aspekty

Ekonomické aspekty certifikované metodiky lze diskutovat jednak z pohledu nákladů nezbytných na zavedení postupů popsaných v metodice a z pohledu ekonomických přínosů pro cílové uživatele. Z hlediska nákladů nezbytných na zavedení postupů popsaných v metodice je nezbytné získat přístup do SW modelu, a zvládnout práci s datovými sadami jejichž vlastnosti jsou v metodice popsány. Zde lze uvést, že jelikož není u cílových uživatelů požadováno datové sady vlastnit, je cena přístupu pouze zlomkem nákladů.

Případné náklady jsou spojeny zejména se školením a přípravou odpovědného pracovníka, který bude schopen připravovat datové sady a provádět kontrolu a úkony spojené s nimi.

Dalším, neméně důležitým, avšak obtížně vyčíslitelným, ekonomickým efektem je možná úspora nejenom v oblasti interních nákladů dopravy ale především v oblasti externích nákladů dopravy spočívajících ve snížení dojezdového času díky možnosti predikce dynamiky dopravního proudu.

12. Seznam použitých zkratk

Alert -C	lokační referenční systém; rozlišujeme lokační tabulky a číselníky události formou jednoznačných kódů	
ADAS	Advanced Driver Assistance System	
API	Application Program Interface	
BO	back office	
CEN	European Committee for Standardisation – Evropský výbor pro normalizaci	
C-ITS	Cooperative Intelligent Transport systém – kooperativní inteligentní dopravní systémy	
CNN	Konvoluční neuronové sítě	
ČHMÚ	Český hydrometeorologický ústav	
ČR	Česká republika	
DBMS	Komerční systémy řízení báze dat	
FCD	plovoucí vozidla	
FVS	flotilová vozidla	
GN	Global Network	
GPS	Global Positioning System	
HW	hardware	
I2I	Infrastructure-to-Infrastructure; vzájemná komunikace prvků infrastruktury	
ICT	informační a komunikační technologie	
ITS	inteligentní dopravní systém	
LoS	Level of Service	
LSTM	Long Short-Term Memory	
MD	Ministerstvo dopravy	
NDIC	Národní dopravní informační centrum	RNN rekurentní neuronová síť
ŘSD	Ředitelství silnic a dálnic	
SMS	silniční meteorologické stanice	
SW	software	
TMC	Traffic Message Control	

TP	Technické podmínky
UTC	koordinovaný světový čas

13. Seznam použité související literatury

- [1] Boukerche, A., & Wang, J. (2020). Machine Learning-based traffic prediction models for Intelligent Transportation Systems. *Computer Networks*, 181.
 - [2] Podklad pro vytvoření Strategického plánu ITS, 44903, Centrum dopravního výzkumu, 2020.
 - [3] Metodika poskytování dat do NDIC, v rámci projektu: Konverze stávajícího systému NDIC do otevřené architektury, TA04031524 TamTam Research s.r.o., Praha 2019.
 - [4] Li, X., & Zhang, Y. (2018). "Handling Missing Data in Traffic Prediction: A Comparative Study." *IEEE Transactions on Intelligent Transportation Systems*, 20(4), 1502-1516.
 - [5] Smith B.L., Williams B.M., Oswald R.K. Comparison of parametric and nonparametric models for traffic flow forecasting *Transp. Res. C*, 10 (4) (2002), pp. 303-321
 - [6] Oh S., Byon Y.-J., Jang K., Yeo H. Short-term travel-time prediction on highway: a review of the data-driven approach *Transp. Rev.*, 35 (1) (2015), pp. 4-32
 - [7] Lv Y., Duan Y., Kang W., Li Z., Wang F.-Y., et al. Traffic flow prediction with big data: A deep learning approach *IEEE Trans. Intell. Transp. Syst.*, 16 (2) (2015), pp. 865-873
 - [8] Yu B., Yin H., Zhu Z. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting Lang J. (Ed.), *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, ijcai.org (2018), pp. 3634-3640
 - [9] Smith, J. (2020). "Data Sources and Quality in Traffic Prediction." *Journal of Transportation Research*, 45(2), 201-215.
 - [10] C. Thay, V. Visoottiviseth, S. Mongkolluksamee, P2P traffic classification for residential network, in: 2015 International Computer Science and Engineering [Conference, ICSEC, IEEE, 2015, pp. 1–6.
 - [11] Sun, L., Zhang, D., & Sun, Q. (2020). "Traffic Flow Data Quality Analysis and Missing Data Imputation Using Long Short-Term Memory Recurrent Neural Networks." *IEEE Transactions on Intelligent Transportation Systems*, 21(11), 4789-4798.
 - [12] ČSN CEN TS 16157-1: Inteligentní dopravní systémy – Specifikace výměnného formátu
 - [13] DATEX II pro řízení dopravy a dopravní informace - Část 1: Obecný rámec a architektura
- Nařízení Komise v přenesené pravomoci (EU) č. 886/2013 ze dne 15. května 2013, kterým se doplňuje směrnice Evropského parlamentu a Rady 2010/40/EU, pokud jde o údaje a postupy pro poskytování minimálních univerzálních informací o dopravním provozu

souvisejících s bezpečností silničního provozu uživatelům, pokud možno bezplatně [online]. 2013 [cit. 2015-12-20]. Dostupný z WWW: <http://eur-lex.europa.eu/legal-content/CS/TXT/PDF/?uri=CELEX:32013R0886&from=CS>

Název: Požadavky na datové zdroje a jejich kvalitu pro zavádění systému predikce stavu dopravy na základě strojového učení

Autoři: Ing. Jan Rozhon, Ph.D.
Ing. Igor Mikolášek
Ing. Jan Mynařík
Ing. Zdeněk Uherka
Ing. Zuzana Švédová, Ph.D.
Ing. Martin Bambušek
Ing. Jakub Šafařík, Ph.D.
Ing. Jiří Novobilský
Mgr. Marek Ščerba

Oponenti: Ing. Jan Vlčínský, TamTam Research
doc. Ing. Zdeněk Lokaj, Ph.D., LL.M.

Fotografie:

Ilustrace:

Odborná korektura:

Jazyková korektura:

Grafická úprava a příprava tiskových podkladů:

Vydalo: Centrum dopravního výzkumu, v. v. i., Líšeňská 33a Brno, Česká republika

Rok, místo a číslo vydání: 2023, Brno, 1. vydání

Tisk:

ISBN XXX-XX-XXXXX-XX-X