Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra informatiky a výpočetní techniky

Diplomová práce

Analýza diagnostických hlášení drážních vozidel

Plzeň, 2017 Bc. Daniel Špale

Prohlášení

Předkládám tímto k posouzení a obhajobě diplomovou práci zpracovanou na závěr studia na Fakultě aplikovaných věd Západočeské univerzity v Plzni.

Prohlašuji, že jsem diplomovou práci vypracoval samostatně a výhradně s použitím odborné literatury a pramenů, jejichž úplný seznam je její součástí.

V Plzni dne 28. 06. 2017 .....

vlastnoruční podpis

Poděkování

Tímto bych chtěl poděkovat Ing. Pavlu Novému, Ph.D. a Ing. Ondřeji Borusíkovi, Ph.D. za vstřícnost, trpělivost a odborné rady při konzultacích. Děkuji i své rodině a přátelům za veškerou podporu a energii, bez které by tato práce nikdy neměla šanci býti dokončena.

Abstract

The aim of this thesis is to create a methodology for analysing diagnostic messages of railway vehicles. Its first part defines the different types of messages sent by the vehicles. Moreover, an analysis of both the database, where the messages are stored, and its disadvantages is performed. That concludes in a draft for optimal adjustments on the one hand and adjustments for the purpose of this thesis on the other hand. Furthermore, a methodology based on an ability to separate all messages regularly sent (process data) according to single drives is elaborated, their similarity is calculated and all drives are grouped corresponding to their track. This procedure is implemented as part of a reporting system which summarizes all important operational data and allows its user to perform statistical analyses that represent the last part of this thesis.

Keywords: diagnostic messages, alarms, process data, database, energy consumption, statistical analysis, railway vehicles

Abstrakt

Cílem této práce je sestavit obecnou metodiku pro analýzu diagnostických hlášení kolejových vozidel. První část práce se zabývá rozborem samotných hlášení a definicí jejich různých typů. Zároveň je provedena analýza databáze, kde se hlášení ukládají, jejích nedostatků a návrh jak optimální varianty úpravy, tak varianty úpravy pro účely realizace této práce. Dále je vytvořena metodika, která je založena na rozdělení všech pravidelně zasílaných hlášení (procesních dat) podle jednotlivých jízd, na což navazuje určení podobnosti jízd a provedení seskupení do tratí. Tento postup je implementován jako součást reportovacího systému, jenž shrnuje všechny důležité provozní údaje a zároveň umožňuje na nich provádět statistické analýzy, kterým se věnuje poslední část této práce.

Klíčová slova: Diagnostická hlášení, alarmy, provozní data, databáze, spotřeba energie, statistická analýza, kolejová vozidla

Obsah

Seznam obrázků

Seznam tabulek

Přehled zkratek

ŠTRN	ŠKODA Transportation, a.s.
UIC	Mezinárodní železniční unie
SŽDC	Správa železniční dopravní cesty
DC	Direct Current – stejnosměrný proud
AC	Alternate Current – střídavý proud
ČD	České Dráhy, a.s.
APU	Auxiliary power unit – pomocné pohony
KW-test	Kruskal-Wallisův test
EDF	empirická distribuční funkce
MDF	modelová distribuční funkce
GUID	generated unique identification number

1 Úvod

Nezávisle na tom, zda se jedná o výrobce automobilů, kolejových nebo jiných vozidel, diagnostická hlášení jsou jak pro daného výrobce, tak i pro koncového provozovatele velmi cennými údaji. Umožňují totiž jak sledování technických parametrů, tak také případných poruch v době provozu. V případě sledování poruch může být diagnostický systém výraznou podporou, jelikož dokáže poskytnout informace o tom, kdy a za jakých okolností k poruše došlo, což může být klíčem nejen k odstranění poruchy, ale i k prevenci jejího vzniku do budoucna.

Společnost ŠKODA Transportation a.s. vybavuje některá svá kolejová vozidla on-line monitorovacím systémem, který sleduje nejen mimořádné události, ale také eviduje průběžný stav vozidel. Skutečnost, že přenos probíhá dálkově, je výhodný opět z hlediska údržby, protože pokud nastane porucha, pak se servisní tým může dopředu na poruchu připravit, a náprava i opětovné plné zprovoznění vozidla může být výrazně operativnější a rychlejší. On-line monitoring ale též umožňuje vyhodnocení chování vozidla v průběhu jeho jízdy a případně tak odkrýt nesrovnalosti související přímo s tratí, její charakteristikou či profilem. Mezi průběžně zasílané údaje patří například také spotřeba a rekuperace energie vozidla.

Údaje, které zasílá diagnostický systém jednotlivých vozidel, se ukládají do databáze, která v závislosti na typu vozidla eviduje hlášení i několik měsíců zpět a obsahuje jak průběžné zasílané technické parametry vozidel, tak i případné poruchy či jiné mimořádné události.

Obecným cílem této práce je provést analýzy údajů uložených v databázi. Pro další rozbor těchto informací je nejprve nutné provést úpravu databáze a případně i způsobu zasílání hlášení a jejich uložení tak, aby na jejich základě bylo možné realizovat další analýzy.

Práce se dále věnuje oběma oblastem diagnostických hlášení, které jsme naznačili již výše, tj. hlášením týkajících se poruch na jedné straně a na straně druhé pak průběžně zasílaným hlášením, speciálně pak spotřebě a rekuperaci energie jednotlivých vozidel.

V rámci poruchovosti vozidel jsou provedeny statistické analýzy za účelem zjištění, zda existuje souvislost mezi místem či úsekem na trati a vznikem poruchy, nebo zjištění, zda zasílané údaje neobsahují příčinnou vazbu s tratí a tedy původem vlastností je vozidlo samotné.

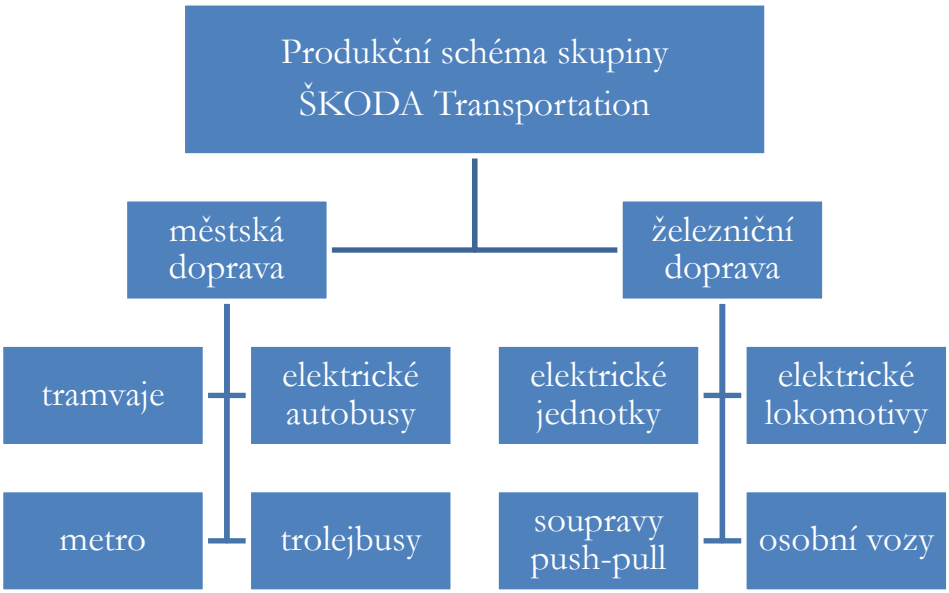
Pro spotřeby energie pak práce posuzuje s využitím statistických analýz, zda vozidla mají stejnou spotřebu, nebo zda v závislosti na směru jízdy na trati nebo trati samotné mají spotřebu různou.

2 Analýza struktury diagnostických hlášení a způsob jejich uložení

2.1 Obecné informace

Společnost ŠKODA Transportation, a.s. (ŠTRN) ve spolupráci s dceřinými společnostmi vyvíjí a vyrábí vozidla pro městskou a železniční dopravu. Obecný přehled produktů představuje Obrázek 2.1. Dodávky probíhají vždy na základě konkrétní zakázky, která přesně určuje, kolik vozidel bude vyrobeno a jaké technické parametry vozidla mají splňovat – mimo rozchodu[1] se to týká například i trakčního vybavení vozidel, které musí být přizpůsobeno trati a jejímu systému trakční soustavy, na které se budou později nasazovat.

Pro dodávky v České republice je toto například zcela zásadní, jelikož máme aktuálně na tratích v tomto ohledu čtyři různé systémy, dva se stejnosměrným (DC) a dva se střídavým napájením[2] (AC). Příslušnou mapu s rozlišením typů trakcí dává k dispozici Správa železniční dopravní cesty (SŽDC) (1). V rámci projektu se formálně zavede pro všechna vozidla společné tovární a obchodní označení.



Obrázek 2.1: Produkční schéma ŠTRN

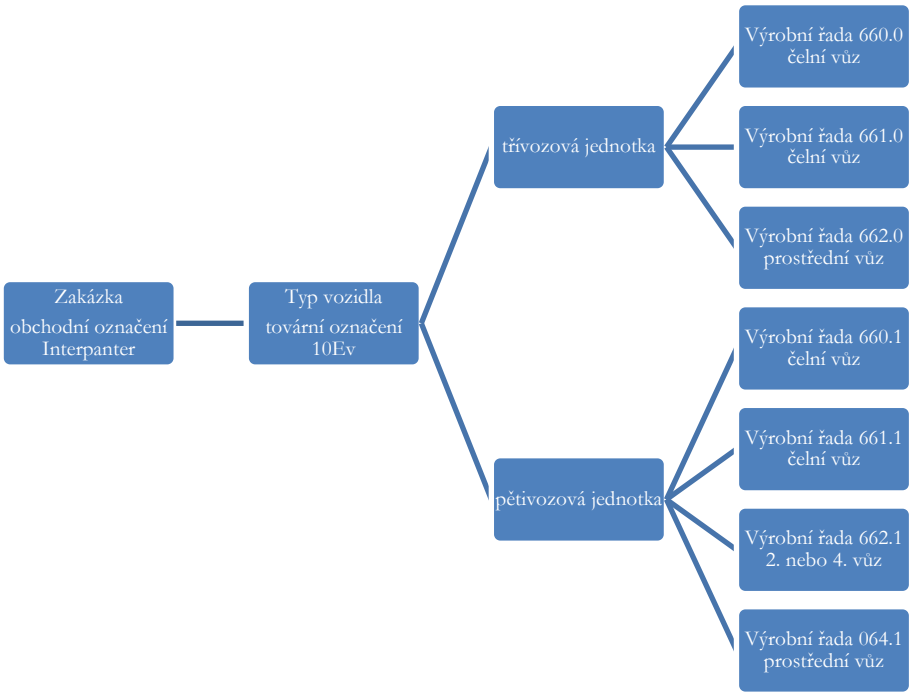
Ověření později použité metodiky pro určení podobnosti jízd v kapitole 3 a analýzy spotřeby energie a alarmů budou dále provedeny na elektrických jednotkách.

---

1 Rozchod kolejí odpovídá vzdálenosti mezi vnitřními hranami kolejnic

2 Samotné vozidlo může být tzv. vícesystémové, tj. jeho trakční vybavení umožňuje jezdit po tratích s odlišným systémem trakční soustavy

Většinou se na základě zakázky vyrobí s ní spojený zcela nový typ vozidla[3], který se může dále dělit do různých výrobních řad. Toto názorně představuje Obrázek 2.2 na příkladu železničních elektrických jednotek InterPanter.



Obrázek 2.2: Rozdělení zakázky na vozidla InterPanter

Dále Tabulka 2.1 uvádí základní technické parametry zmíněných vozidel (2), (3).

Ve výše uvedených dokumentech je též uvedeno, že elektrické jednotky jsou vybavené systémem tzv. **rekuperačního brždění**, který umožňuje přeměnu kinetické na elektrickou energii. Ta se dá následně vracet zpět do napájecího systému. V současné době jsou systémy rekuperace energie důležitým požadavkem v tendrech, jelikož snižují spotřebu elektrické energie.

Po převzetí vozidel je jejich nasazení v režii provozovatele. Trať, po které vozidla jezdí, se může v průběhu času změnit, například České Dráhy, a.s. (ČD) mění pro některé tratě jízdní řád a tím i nasazení vozidel každý půlrok.

Jen některé typy vozidel jsou vybavené **diagnostickým systémem**, který zasílá pravidelná hlášení o stavu vozidla. Obvykle se instaluje do produktů železniční dopravy tak, aby provozovatel, respektive jeho servis měl možnost se předem připravit na opravu poruchy, která nastala během jízdy. Naopak vozidla pro městskou dopravu diagnostický systém ve standardní výbavě nemají, jelikož se do depa vracejí mnohem častěji, a tedy s předstihem zaslaná informace o poruše nemá až takovou hodnotu. Jednotlivá hlášení rozebereme podrobněji v dalším odstavci.

V rámci této práce se tudíž budeme soustředit na vozidla železniční dopravy.

[3] Různé typy vozidel plynou z jednotlivých realizovaných projektů, tudíž tramvaje 15T dodávané do Prahy jsou jiný typem vozidla, než tramvaje 30T, které jezdí v Bratislavě

	Třívozová jednotka	Pětivozová jednotka
Obchodní název	InterPanter	
Tovární označení	10Ev	
Rok výroby	2015 - 2016	
Rozchod	1 435 mm	
Délka soupravy	79 400 mm	132 400 mm
Výška vozu	4 260 mm	
Šířka vozu	2 820 mm	
Jmenovité napětí troleje	3kV DC /25kV 50Hz	
Výkon (*pč. motorů*výkon jednoho motoru*)	6 x 340 kW	8 x 340 kW
Maximální rychlost	160 km/h	160 km/h
Max. zrychlení	1,1 m/s²	1,1 m/s²
celkem dodaných souprav	4	10

Tabulka 2.1: Základní specifikace vozidel InterPanter

2.2 Vymezení pojmů a obsahu hlášení

Pojem „diagnostická hlášení“ popisuje v rámci provozu vozidel ŠTRN dva pojmy:

i. Alarmy,
ii. Procesní data.

Jedná se o dva různé druhy zpráv, které posílá **diagnostický systém** vozidla na **přijímací server ŠTRN**, který tato hlášení zpracuje a pošle dál do **databáze**. V době realizace této práce se celkově evidovalo 452 vozidel deseti různých typů.

Obrázek 2.3: Přenos dat z vozidla do databáze

Alarmy odpovídají **mimořádným hlášením** a tudíž případu, kdy diagnostika vozidla detekuje problém či nesrovnalost oproti obvyklému provozu. Jedná se tedy o událostní systém sběru dat.

Tyto události se v softwaru vozidla definují během vývoje – každý alarm tedy odpovídá **definovanému odchýlení od normálního stavu**. Vozidla například mohou hlásit poruchu jednoho ze zařízení, výpadek jističe a další události, jako vypnutí jednoho z motorů. Každý alarm má své **vlastní identifikační číslo označené jako „ŠKODA Alarm ID“** a pro něj specifické údaje, které vozidlo společně s hlášením o alarmu zasílá. Například pro alarm „Nadproud troleje při stání“ vozidlo zašle údaje o proudu troleje a dále pro každý sběrač proudu, zda se nachází ve zvednutém stavu či nikoli. Tyto dodatečné údaje se označují

jako **proměnné.** Identifikační čísla alarmů obecně nejsou pro všechny typy vozidel stejná, tj. **jeden konkrétní alarm může mít v rámci všech typů vozidel různá označení**. Stává se to například na základě požadavku zákazníka, který pro „svůj" typ vozidla přímo předepisuje identifikační čísla alarmů.

V neposlední řadě se **alarmy rozlišují dle jejich typu**. Pokud diagnostika vozidla detekuje alarm, pošle v rámci daného hlášení příznak „vznik", naopak v případě, kdy již nejsou splněny podmínky daného alarmu, vozidlo opět pošle zprávu s příznakem „zánik". Tato konkrétní informace má podobu osmibitového čísla, kde každý bit zastupuje jeden z možných příznaků alarmu. Mimo vznik a zánik jsou dalšími příznaky například informace o tom, jestli se alarm objevil v režimu údržby nebo třeba aktivního odstavení[4]. Teoreticky může vozidlo v rámci diagnostického hlášení zaslat libovolnou kombinaci všech osmi příznaků, což by znamenalo celkově 256 možných typů alarmu. V praxi je některé z těchto kombinací možné vyloučit (například vznik a zánik alarmu zároveň).

Procesní data jsou **pravidelně** podávaná hlášení o stavu a základních provozních údajích vozidla. Jejich základem jsou též **proměnné, které mohou obsahovat jiné informace než u alarmů** - standardně se zasílají informace o **aktuální rychlosti**, **kódu řidiče** i o tom, jestli se dané vozidlo nachází v režimu aktivního odstavení. Konkrétní proměnné, které vozidlo zasílá, se určují v softwaru, který je napříč všemi typy vozidel různý. Ani v rámci jednoho typu vozidel navíc není zaručeno, že všechna vozidla zasílají stejné údaje, což nastává v případě odlišností ve verzi nahraného softwaru.

Obecně pod kategorii procesních dat patří také hlášení o spotřebované a rekuperované energii, kterým se budeme věnovat více v dalších částech této práce.

Na rozdíl od alarmů, kde se hlášení zašle pouze v případě, kdy dojde k určité nepředvídané události, **procesní data se zasílají pravidelně,** a to v případě některých typů vozidel **každých 1,5 sekundy**. **Diagnostika nezasílá nic pouze v době, kdy je vozidlo zcela vypnuté**, což může u některých typů vozidel znamenat méně než pět hodin denně. Příslušné tabulky v databázi obsahují kolem 20 miliónů záznamů s historií pouze několik měsíců – starší údaje se v ŠTRN archivují v textových souborech.

Pro oba typy diagnostických hlášení pak dále platí, že se mimo jiné vždy zasílá mezinárodně unikátní identifikační číslo vozidla v rámci mezinárodní železniční unie (UIC), které se označuje jako **UIC číslo vozidla**, čas odeslání hlášení systémem vozidla a přijetí serverem ŠTRN a GPS poloha vozidla[5].

I přesto, že obsahově posílají vozidla různé údaje, tak z hlediska struktury lze stanovit obecnou podobu diagnostických hlášení. S každým jednotlivým hlášením vozidlo zasílá tyto data (viz Obrázek 2.4):

i. **Alarm:** UIC číslo vozidla, číslo řídícího počítače, čas odeslání hlášení systémem, GPS souřadnice, SKODA_Alarm_ID, typ alarmu, slovní popis

[4] V tomto režimu se nachází vozidlo v případě, že je odstavené v zapnutém stavu se zvednutým sběračem.

[5] Přesnost pět míst za desetinnou čarou

alarmu, k danému alarmu přidružené proměnné, čas přijetí hlášení serverem, Alarm_GUID, Alarm_UID (k posledním dvěma údajům více v další části),

ii. **Procesní data:** UIC číslo vozidla, číslo řídícího počítače, čas odeslání hlášení systémem, GPS souřadnice, aktuální rychlost, kód řidiče, stav odstavení vozidla, k hlášení přidružené proměnné, čas přijetí hlášení serverem, kód depa, Proc_GUID.

2.3 Analýza původní databáze

2.3.1 Popis a význam jednotlivých tabulek

V ŠTRN se **pro účely ukládání alarmů a procesních dat** využívala relační databáze s podobou, jako ji představuje Obrázek 2.4, kde písmenem F je označen cizí a písmenem P primární klíč dané tabulky.

Oba typy diagnostických hlášení se vkládají do tabulek Alarm a ProcData, přičemž pro každý typ vozidla existuje **separátní dvojice těchto tabulek**, jak také Obrázek 2.4 naznačuje. Právě tam se ukládají údaje, o kterých pojednává předchozí část této kapitoly.

U většiny označení atributů se jedná o anglický ekvivalent k tomu českému a tedy „neintuitivní" je pouze atribut Data, který obsahuje proměnné (viz nahoře). Ty se do databáze ukládají jako textový řetězec „proměnná 1=hodnota1; proměnná 2=hodnota2;…", kde označení jednotlivých proměnných je uvedeno ve zkratce, aby byl textový řetězec co nejkratší. Tato skutečnost je velký problém pro pozdější práci s daty, k čemuž se vrátíme v odstavci 2.3.2.

Dalším důležitým údajem jsou „Alarm_GUID" u alarmů, respektive Proc_GUID u procesních dat – jedná se o číslo, které se generuje na úrovni přijímacího serveru na základě dat pro každou jednotlivou zprávu[6]. Účelem tohoto čísla je zamezit uložení duplicitních záznamů, které na úrovni databáze není ošetřené (nebyl definován primární klíč). Může se totiž stát, že vozidlo z nějakého důvodu odešle jedno konkrétní hlášení více než jedenkrát. Přijímací server by pak měl na základě určení GUID předat hlášení do databáze pouze jednou.

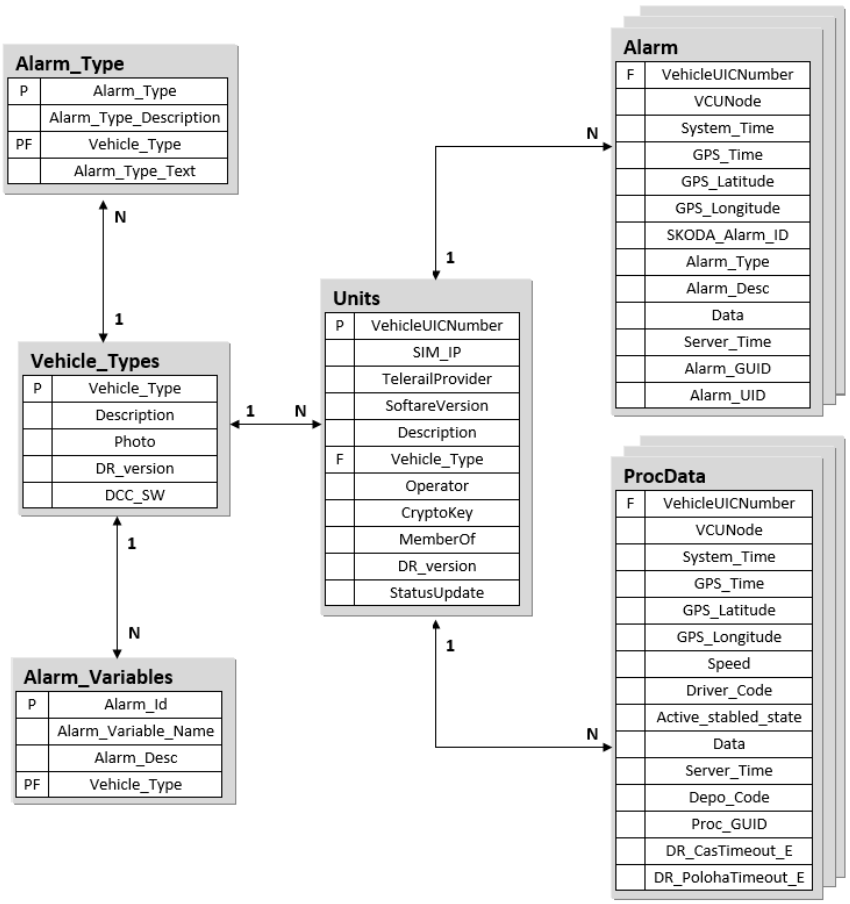
Dvojice tabulek k alarmům a procesním datům daného typu vozidla je napojena na tabulky „Units", kde jsou evidovaná všechna vozidla, ke kterým se diagnostická hlášení ukládají. Primárním klíčem této tabulky je UIC číslo vozidla, ale evidují se zde další údaje, jako nahraná verze softwaru[7], provozovatel a typ vozidla. Pro pozdější analýzy diagnostických hlášení však tyto údaje nebudeme potřebovat.

Tabulka „Vehicle_Types" s primárním klíčem „Vehicle_Type" obsahuje interní označení typu vozidel ŠTRN (Description) a také údaje o aktuální verzi SW vozidla i diagnostického počítače. Pro účely analýz budou pro nás více relevantní tabulky, které jsou na tuto tabulku napojené.

---

[6] Generuje se na základě všech dat, nejen proměnných

[7] SW vozidla se přehrává vzdáleně, čili se z těchto údajů dá poznat, zda dané vozidlo již má aktuální verzi, či nikoli



Obrázek 2.4: Relační model původní databáze

V tabulce „Alarm_Type“ se evidují slovní popis a příslušný symbol k číselnému označení, které jsou součástí diagnostických hlášení a jsou uvedeny v tabulkách „Alarm“. Pro databázi navíc platí, že původní osmibitové číslo definující typ alarmu se do ní ukládá jako příslušné číslo desítkové soustavy. V původní databázi nebyla označení jednotlivých typů alarmů jednotná, tj. jeden konkrétní typ alarmu měl pro dva různé typy vozidel různé významy. Proto součástí primárního klíče a zároveň i klíčem cizím této tabulky je „Vehice_Type“.

Součástí relačního modelu původní databáze je také tabulka „Alarm_Variables“, kde je ke dvojici identifikačního čísla alarmu a typu vozidla přiřazena jedna proměnná a její slovní popis.

2.3.2 Rozbor z hlediska pozdějších analýz

V obou případech diagnostických hlášení, tj. jak pro tabulky Alarm, tak i ProcData, máme v tabulce **duplicitní záznamy**, takže se některá hlášení shodují. Důležité tedy bude při pozdějších úpravách zajistit jedinečnost každého záznamu.

Dále obě tabulky obsahují **přebytečné sloupce obsahující pouze hodnotu NULL**, tj. některá vozidla nezasílají údaje, které by tyto sloupce naplňovaly. Těmito atributy jsou v obou případech „GPS_Time“ a u tabulek ProcData „Depo_code“, „Proc_GUID“, „DR_CasTimeout_E“ a „DR_PolohaTimeout_E“.

Do tabulek Alarm se ukládá identifikační číslo alarmu společně s jeho slovním popisem (SKODA_Alarm_ID a Alarm_Desc) – v podstatě se tedy **jedna informace ukládá dvakrát** v různých podobách, máme funkční závislost mezi zmíněnými dvěma atributy a tabulka v tomto ohledu neodpovídá třetí normální formě.

V našem rozboru se dále pozastavíme nad tabulkou „Alarm_Type" – z hlediska architektury databáze by zcela určitě měl být atribut „Alarm_Type" v tabulkách Alarm cizím klíčem, který bude odkazovat na tabulku „Alarm_Type". Navíc však vadí, že **hodnoty typů alarmů nejsou pro všechna vozidla jednotné**, což by bylo lepší z pohledu přehlednosti a způsobu práce s daty.

Největší problémem je však výše zmíněná skutečnost, že **proměnné** se do databáze **ukládají v podobě textového řetězce**, což znemožňuje jednoduché dotazování nad příslušnými tabulkami v databázi a získání proměnných včetně jejich hodnot. V současném stavu je proto **nutné provést rozsáhlé předzpracování dat**, než s nimi začneme pracovat. Úpravou architektury databáze se toto pokusíme eliminovat.

Dalším výraznějším nedostatkem je, že **pro každý typ vozidla máme separátní tabulku pro alarmy a procesní data**. Sotva bychom chtěli provést úpravu v architektuře databáze, musíme ji provést pro každý typ vozidla zvlášť. Pokud bychom chtěli přidat nový typ vozidla, což odpovídá situaci nové zakázky, a tedy jde o běžnou záležitost, musíme přidávat několik nových tabulek a celá databáze se tím stává nepřehlednou a komplikovanou.

2.3.3 Úpravná opatření stávající databáze – optimální varianta

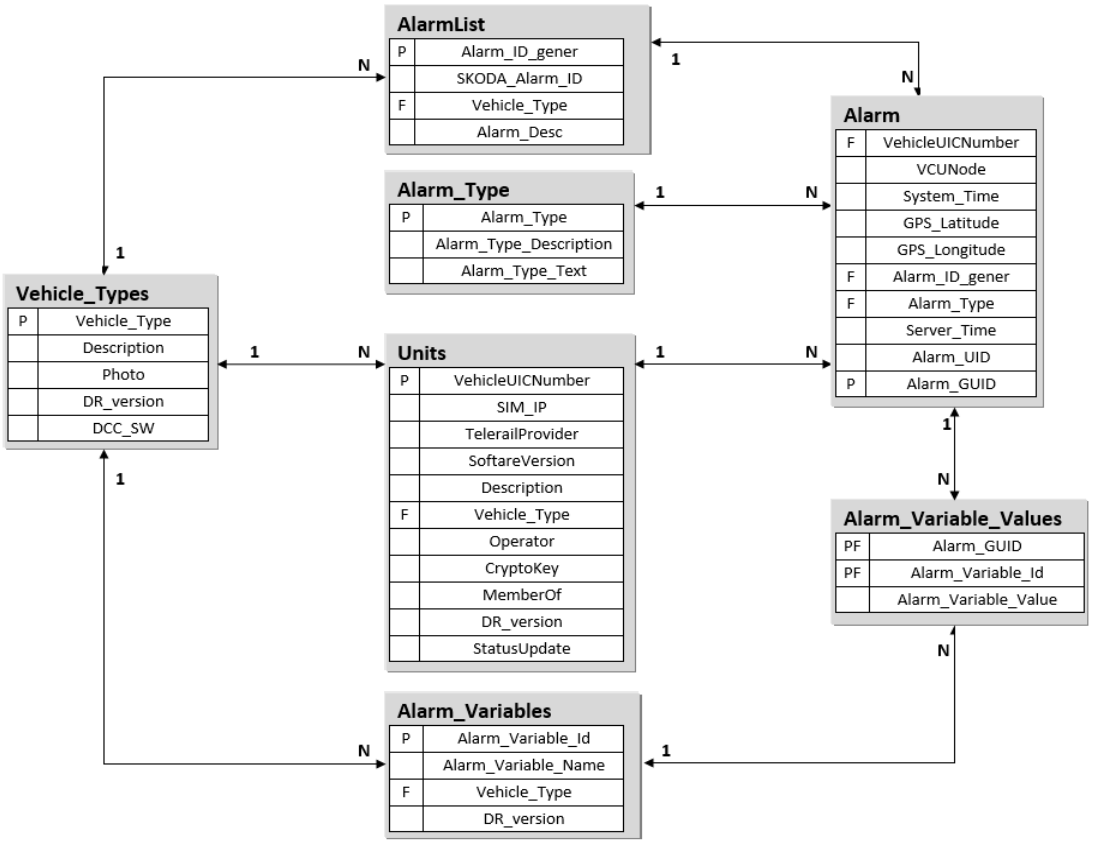
Pro zachování přehlednosti rozdělíme relační model na dvě části, kde Obrázek 2.5 znázorňuje úpravy pro tabulku Alarm a Obrázek 2.6 to samé pro procesní data. Tabulky „Units" a Vehicle_Types" jsou pro celý relační model společné a ve srovnání se svojí původní podobou neprošly žádnou úpravou.

Jak v případě tabulky pro procesní data, tak pro alarmy je naším hlavním záměrem udělat společnou tabulku pro všechny typy vozidel a navíc ulehčit práci s proměnnými, respektive umožnit jejich jednoduché čtení z databáze pomocí SQL-dotazů. Navíc pro oba typy hlášení řešíme menší obtíže, které byly zmíněny v odstavci 2.3.2.

Řešení problému s proměnnými v textovém řetězci bude v obou případech podobné. Proměnné začneme vnímat jako samostatnou entitu - založíme tabulku „ProcData_Variables" a oživíme „Alarm_Variables". Také tyto tabulky budou společné pro všechny typy vozidel – budou zde tedy evidovány veškeré proměnné, které se v databázi vyskytují. Právě proto je atributem obou tabulek také typ vozidla, a abychom ošetřili případné pozdní změny ve významu proměnných, je zahrnuta i verze softwaru, pro kterou dané označení proměnné platí.

Mezi entitou proměnných a příslušným typem hlášení máme relace M:N, kterou vyřešíme rozkladovou tabulkou, kde každému hlášení (identifikovaným primárním klíčem) a obdobným způsobem každé proměnné bude přiřazena jediná hodnota. Tím zcela nahradíme atribut „Data" pro obě varianty.

V případě alarmů potřebujeme vyřešit funkční závislost atributů „SKODA_Alarm_ID“ a „Alarm_Desc“. Tuto dvojici vyčleníme do separátní tabulky „AlarmList“. K tomu, aby navíc bylo možné evidovat všechny alarmy všech typů vozidel, přidáme „Vehicle_Type“, a jako primární klíč bude sloužit automaticky se generující číslo „Alarm_ID_gener“ (tj. při přidání nového řádku do tabulky databáze se toto unikátní číslo

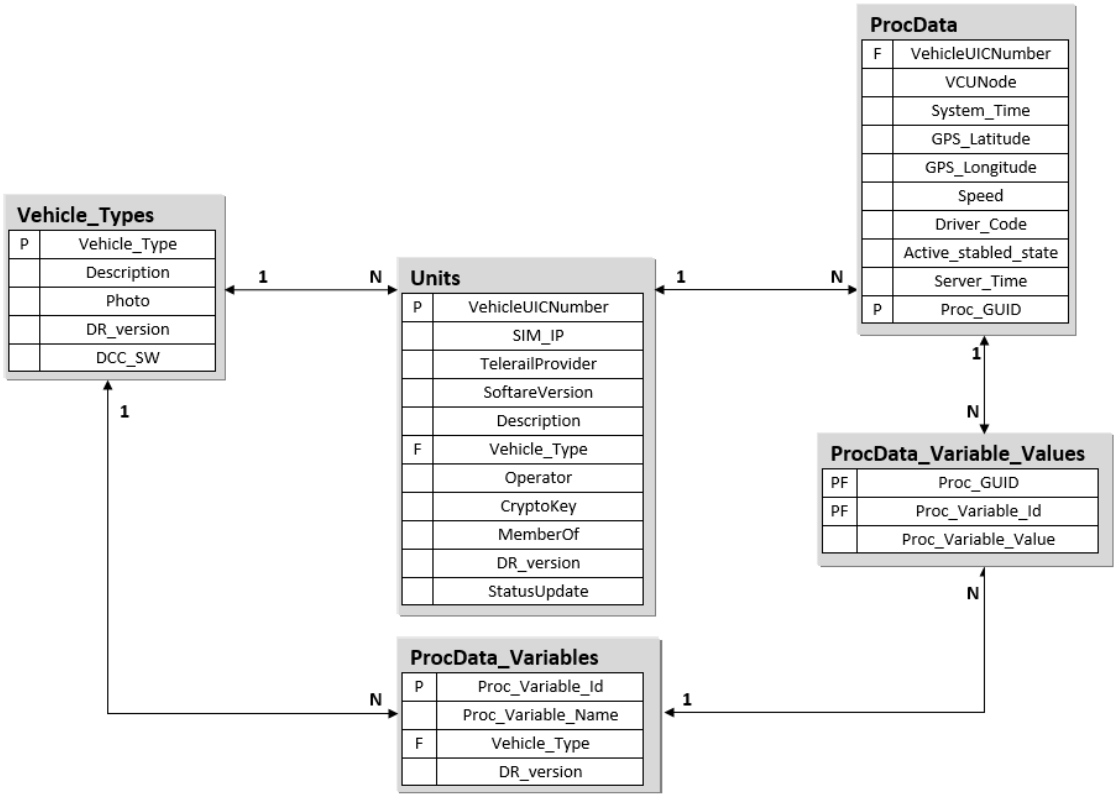


Obrázek 2.5: Relační model po úpravě – část alarm

vygeneruje samostatně). V tabulce „Alarm_Type“ sjednotíme značení typů alarmů a přímo ji „napojíme“ na tabulku „Alarm“.

Jako primární klíč tabulky „Alarm“ zvolíme „Alarm_GUID“ a abychom zabránili duplicitním záznamům, definujeme čtveřici „VehicleUICNumber“, „System_Time“, „Alarm_Type“ a z tabulky „AlarmList“ převzaté „Alarm_ID_gener“ jako unikátní index. Atribut „GPS_Time“ z relace odstraníme.

Úprava tabulky pro procesní data bude jednodušší. Duplicitám zde zabráníme podobným způsobem jako u alarmů s tím rozdílem, že jako primární klíč definujeme „Proc_GUID“ a unikátní index bude na atributech „VehicleUICNumber“ a „System_Time“. Nepotřebné atributy „GPS_Time“, „Depo_Code“, „DR_CasTimeout_E“, „DR_PolohaTimeout_E“ z relace odstraníme.



Obrázek 2.6: Relační model po úpravě – část procesních dat

2.3.4 Úpravná opatření stávající databáze pro účely této práce

Modifikace, které byly uvedeny v odstavci 2.3.3, vyžadují velký zásah do databáze. Proto byly v současném stádiu realizovány jen úpravy, které byly nutné pro realizaci této diplomové práce.

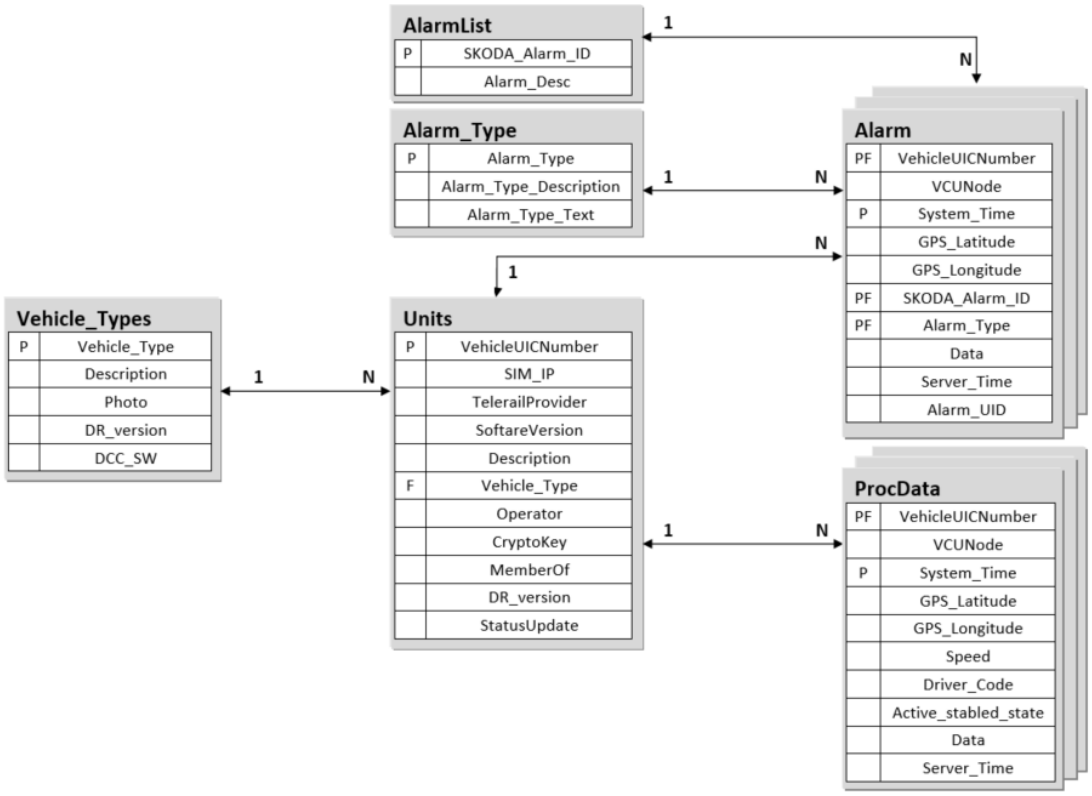
i. tabulky Alarm:
   - duplicitní záznamy způsobené tím, že nefunguje ošetření na úrovni serveru pomocí generování GUID
     - ⇨ **vyřešíme definováním primárního klíče a odstraníme z tabulky „Alarm_GUID“**,
   - atribut „GPS_Time“ je pro všechny záznamy bez hodnoty (je „NULL“)
     - ⇨ **atribut z relace odstraníme**,
   - funkční závislost mezi atributy „SKODA Alarm ID“ a „Alarm_Desc“
     - ⇨ **z tabulek Alarm odstraníme „Alarm_Desc“ a vytvoříme pro každý typ vozidla tabulku „AlarmList“, kde budeme ukládat dvojice výše uvedených atributů**,

ii. tabulky ProcData:
   - duplicitní záznamy (viz výše)
     - ⇨ **stejné řešení, jako pro tabulky Alarm**,

- atributy „GPS_Time“, „Depo_Code“, „DR_CasTimeout_E“, „DR_PolohaTimeout_E“ jsou u všech záznamů „NULL“
  - ⇨ **atributy z relace odstraníme**,

iii. tabulka Alarm_Type:

- značení typu alarmů není pro všechny typy vozidel jednotné
  - ⇨ **sjednotíme značení typů alarmu a z tabulky odstraníme sloupec „Vehicle_Type“**



Obrázek 2.7: Relační model databáze pro účely diplomové práce po úpravě

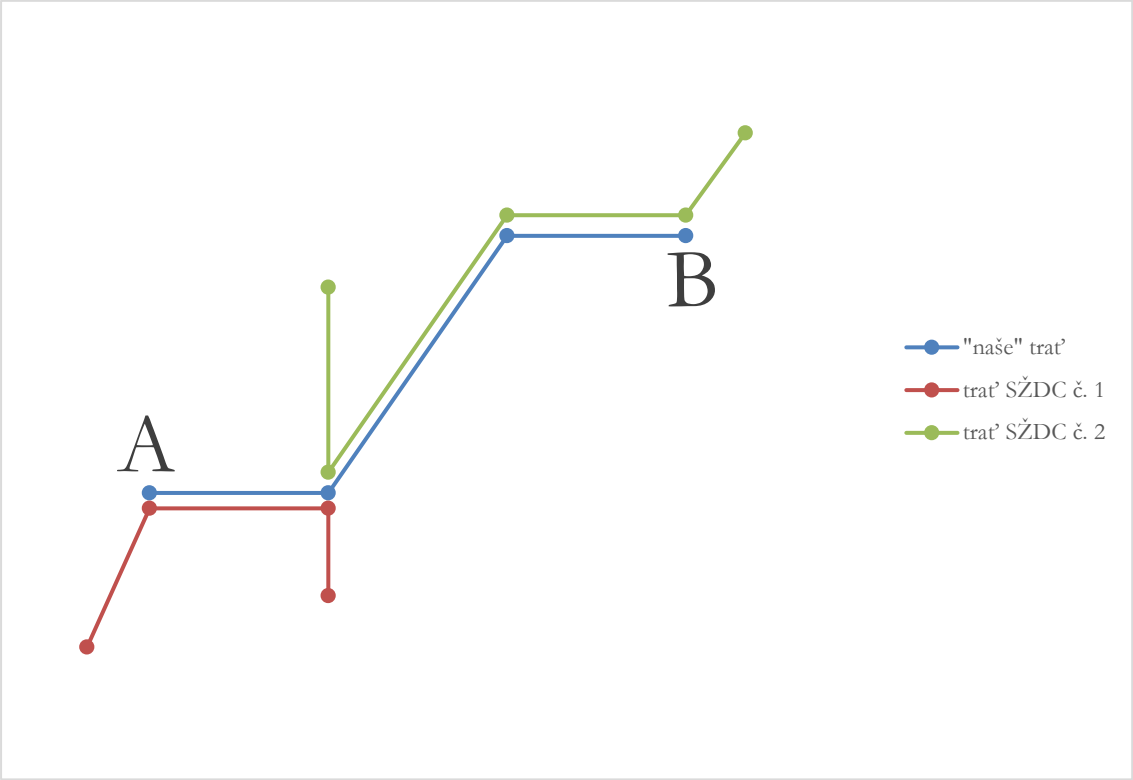
Po úpravě bude mít relační model databáze takovou podobu, jako ji představuje Obrázek 2.7. Před úpravou bylo provedeno odstranění duplicit a nutné pročištění dat.

Jako primární klíče jsme v případě tabulky Alarm zvolili čtveřici UIC čísla vozidla, čas odeslání alarmu systémem, identifikační číslo alarmu a jeho typ, což odpovídá tomu, že se konkrétní čtveřice těchto údajů nemůže v databázi vyskytovat dvakrát. Obdobně to platí pro tabulky ProcData a dvojici UIC čísla vozidla a čas odeslání hlášení systémem.

3 Určení podobnosti jízd a klasifikace podle tratí

3.1 Úvod do problematiky

Délka v kilometrech, výškový profil či počet zastávek jsou pro každou trať specifické. Tyto faktory mohou mít výrazný vliv na spotřebu energie vozidel, ale také například vznik alarmu může souviset s konkrétním místem či úsekem na trati. Zároveň ale platí, že z databáze získáme pouze jednotlivá hlášení[8], která podávají zprávu o **okamžitém stavu** vozidla v konkrétní čas na konkrétním místě[9]. Jinak řečeno, **jednotlivá hlášení neobsahují informaci o tom, kam vozidlo v dané chvíli jelo, na které trati se nachází a ani zastávky, kterými projelo nebo zrovna projíždí.**



Obrázek 3.1: Ilustrace rozdílu tratě SŽDC a tratě v našem chápání z místa A do B

Chceme-li tedy získat **souhrnné statistiky z celé jízdy**, tak je nutné jednotlivá hlášení seskupit. Nejdříve objasníme tři důležité pojmy, které budou dále používány:

i. **Trať podle SŽDC:** pro každý úsek železniční cesty v České republice eviduje SŽDC unikátní identifikační číslo, přičemž takto očíslované tratě se nepřekrývají v žádném úseku – protínají se maximálně v jednotlivých zastávkách – **dále jen „trať SŽDC“**.

ii. **Trať v našem chápání:** konkrétní úseky, po kterých jezdí sledovaná vozidla, nemusí plně odpovídat tratím podle SŽDC – naopak je běžné, že

[8] Zde jsou tím míněna převážně procesní data

[9] Místo je určené na základě zaslaných GPS souřadnic

tratě SŽDC vozidla kombinují a také je neprojedou celé. Proto **pojem trať v této práci** popisuje úsek od jedné koncové zastávky ke druhé – **není tedy ekvivalentní s pojmem trať SŽDC** (pro ilustraci viz Obrázek 3.1). Samotným pojmem „trať" navíc **nerozlišujeme směr**, kterým vozidlo po této trati jede – **dále jen „trať"**.

iii. **Jízda:** takto je označen přesun vozidla z místa A do místa B a všechna hlášení, která k danému časovému období provozu patří. Rozlišovat budeme pro každou jízdu i její směr jako „tam" (z A do B) a „zpět" (z B do A). Z principu tedy k jedné trati můžeme přiřadit několik jízd, ale každá jízda se může uskutečnit pouze na jedné konkrétní trati.

Cílem v této kapitole je **pro offline vyhodnocení**[10] postupně ze **všech hlášení**, která získáme v jejich „hrubé" podobě z databáze, **určit jednotlivé jízdy**. Na jejich základě následně bude možné sbírat informace o **technických parametrech při provozu vozidla na trati**.

Při zvolení postupu jsme museli brát v potaz, že **nebyly k dispozici** základní informace o jednotlivých tratích, jakými jsou zastávky a jejich GPS souřadnice. Právě proto byl vytvořen postup, kterým našeho cíle dosáhneme nezávisle na tom, po které trati se vozidlo pohybovalo.

Jednotlivé kroky, které pro to jsou potřebné, znázorňuje Obrázek 3.2.

Obrázek 3.2: Postup při určení podobnosti jízd a klasifikace do tratí

Nejprve se pokusíme pomocí vhodných kritérií rozdělit všechna hlášení do skupin tak, aby každá tato skupina odpovídala jedné konkrétní jízdě. Tomuto postupu se věnuje podkapitola 3.4. a 3.5.

Poté bude naší snahou porovnáním jízd provést jejich seskupení, aby ve výsledku všechny jízdy v jedné skupině odpovídaly jedné trati. Důležité je, že v tomto kroku nás nebude zajímat konkrétní trať, kde se vozidlo pohybovalo, tudíž seskupení jízd provedeme čistě na základě **vzájemného porovnání jednotlivých jízd** – neodpovídá to tedy postupu, kdy bychom brali jízdu po jízdě a každou jízdu nezávisle na dalších se pokoušeli přiřadit k nějaké trati.

Posledním krokem je přiřadit skupiny jízd trati, na které se uskutečnily. Jinak řečeno naší snahou bude jízdy přiřadit konkrétnímu úseku na síti tratí. Určení podobnosti jízd a seskupení podle tratí se věnuje podkapitola 3.6.

[10] rozdělení všech hlášení do jízd neprobíhá online v databázi, ale offline na vybraném souboru z databáze

Obsahem této kapitoly je vysvětlení zvoleného postupu při seskupení a kritéria, na kterých je celý postup založen. Nejdříve ale v podkapitolách 3.2 a 3.3 zavedeme postupy, které budeme při pozdějších výpočtech využívat.

3.2 Proměnné v alarmech a procesních datech

Hlavním problémem je, že pro oba typy hlášení se **proměnné** nachází v **textovém řetězci** a po získání dat z databáze musíme provést další kroky, abychom se dostali k jednotlivým hodnotám proměnných. Při exportu dat se uloží příslušný soubor ve formátu csv, kde **každý jednotlivý řádek představuje jedno konkrétní zaslané hlášení**. K dalším úpravám použijeme tabulkový procesor Microsoft Excel.

Jak bylo výše uvedeno, proměnné můžou být v textovém řetězci v závislosti na typu vozidla a například i na verzi softwaru na různých pozicích. Zároveň ale platí, že u všech hlášení napříč všemi typy vozidel je dodržena struktura „proměnná 1=hodnota1; proměnná 2=hodnota2;…“, tudíž:

i. Najdeme-li hledanou proměnnou, tak její příslušná hodnota je v textovém řetězci napravo od ní, oddělená rovnítkem,
ii. Jednotlivé dvojice proměnná-hodnota jsou oddělené středníkem.

Na sloupec data použijeme funkci „Text do sloupců“ a jako oddělovače zvolíme středník a rovnítko. Po této úpravě se v nově vzniklých sloupcích budou střídat proměnné, které jsou v buňce vpravo od nich následovány jejich příslušnou hodnotou v daném diagnostickém hlášení. Z původního jednoho sloupce „data“[11] nám tímto způsobem vznikne počet sloupců odpovídající dvojnásobnému počtu proměnných v daném hlášení.

Pro každý řádek jsme textový řetězec data upravili do podoby, jako ji naznačuje Tabulka 3.1.

Vehicle UIC number	**…**					
…	…	Proměnná 1	Hodnota 1	Proměnná 2	Hodnota 2	…
…	…	Proměnná 2	Hodnota 3	Proměnná 4	Hodnota 5	…
…	…	…	…	…	…	…

Tabulka 3.1: Struktura proměnných po aplikaci funkce Text do sloupců

Tato podoba nám ale stále nevyhovuje, jelikož se proměnné a jejich hodnoty napříč všemi záznamy v tabulce mohou nacházet na různých pozicích a práce s nimi by byla značně složitá. Ideálně bychom ve výsledku chtěli mít strukturu podobnou tomu, co naznačuje Tabulka 3.2 – pro námi zvolenou proměnnou mít separátní sloupec, ve kterém budou uvedeny hodnoty příslušné k hlášení v daném řádku.

[11] Zde tím je míněn atribut relací Alarm a ProcData v databázi

K dosažení našeho cíle budeme využívat dvě funkce programu Microsoft Excel:

i. INDEX,
ii. POZVYHLEDAT.

INDEX vrátí hodnotu z buňky podle zadaných čísel řádku a sloupce ze zadané oblasti. POZVYHLEDAT najde hodnotu v zadané oblasti buněk a vrátí její relativní hodnotu, tj. číslo pozice hledané hodnoty. Kombinací těchto dvou funkcí získáme podobu, jakou má Tabulka 3.2.

Vehicle UIC number	**…**	**Proměnná 1**	**Proměnná 2**	**…**	**Proměnná 4**	
…	…	Hodnota 1	Hodnota 2	…	Hodnota 6	…
…	…	-	Hodnota 3	…	Hodnota 5	…
…	…	…	…	…	…	…

Tabulka 3.2: Struktura proměnných po všech úpravách

V tabulce, kterou jsme získali úpravou pomocí funkce Text do sloupců, vytvoříme nový sloupec, jehož nadpisem bude konkrétní proměnná. Hodnoty položek budou určeny vzorcem typu

INDEX($F$2:$XA$929;ŘÁDEK()-1;POZVYHLEDAT($A$1;F2:XA2;0)+1),

kde v daném příkladu je

$F$2:$XA$929	celá tabulka po úpravě funkcí Text do sloupců,
$A$1	odkaz na buňku s proměnnou, jejíž hodnoty chceme získat,
F2:XA2	řádek s údaji konkrétního hlášení, ze kterého vybereme hodnotu proměnné.

Princip spočívá v tom, že **vyhledáme pozici hledané proměnné v daném hlášení,** a jelikož se **příslušná hodnota nachází v buňce vpravo** od ní, přičteme k návratové hodnotě funkce POZVYHLEDAT jedničku. Výsledné číslo z toho součtu využijeme ve funkci INDEX, která pomocí tohoto odkazu opět hodnotu proměnné najde a vypíše jako výsledek celého vzorce.

3.3 Výpočet vzdálenosti GPS souřadnic

Vozidla standardně v rámci procesních dat zasílají stav tachometru v celých kilometrech. Pro účel této práce by toto krokování mohlo však být až příliš hrubé a proto nás budou zajímat vzdálenosti mezi jednotlivými hlášeními, tj. mezi dvojicemi GPS souřadnic.

Zde budeme využívat tzv. haversine vzorec, jehož pomocí určíme vzdálenost mezi dvěma body na sféře (4).

Platí:

$$d = r * haversin^{-1}(h) = 2r * arcsin(\sqrt{h}) \tag{3.1}$$

kde

$h = haversin\left(\frac{d}{r}\right)$,

$haversin$ funkce haversine ($haversin(\theta) = sin^2\left(\frac{\theta}{2}\right)$),

d vzdálenost mezi oběma body,

r poloměr sféry.

Z výše uvedeného se dá dále odvodit:

$$d = 2\,r * arcsin\left(\sqrt{sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + cos(\phi_1)\,cos(\phi_2)\,sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right), \tag{3.2}$$

kde

ϕ_1, λ_1 zeměpisná šířka a délka bodu 1,

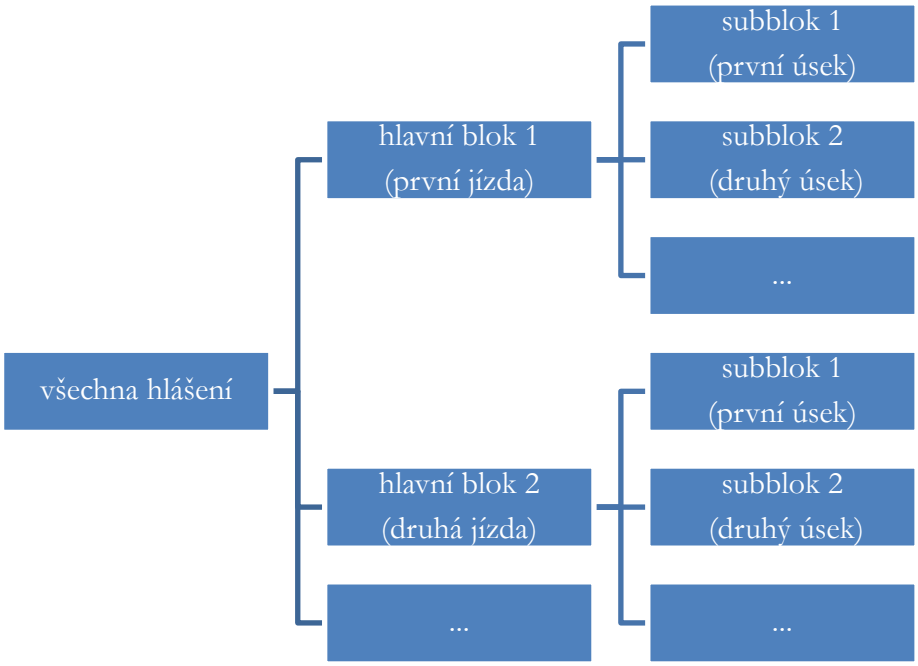
ϕ_2, λ_2 zeměpisná šířka a délka bodu 2.

I přes lehce elipsoidní tvar zeměkoule budeme ve výpočtech Zemi považovat pro zjednodušení za kouli a jako poloměr uvažovat konstantní hodnotu 6371 km.

Bylo ověřeno, že takto dopočtené vzdálenosti mezi jednotlivými hlášeními odpovídají poté v celkovém součtu počtu ujetých kilometrů (dle stavu tachometru), které vozidlo zasílá v rámci procesních dat.

3.4 Rozdělení hlášení do bloků

Naším cílem je k úloze o spotřebách energie a dalším přistupovat tak, aby **závěry byly co nejobecněji aplikovatelné**. Proto bude programové řešení koncipované tak, aby jeho vstupem nebyla pouze skupina hlášení příslušející konkrétní trati, ale naopak **hlášení sesbíraná přes delší časový horizont**, která by následně byla rozdělena do jednotlivých jízd, které by se poté seskupily podle tratí – princip naznačuje Obrázek 3.3.



Obrázek 3.3: Schéma rozdělení všech hlášení do bloků a subbloků

Základem bude správné rozdělení hlášení do jednotlivých jízd nebo jinými slovy seskupení hlášení, která patří do jedné jízdy. Všechna hlášení patřící **k jedné jízdě** budou tvořit jednu skupinu, které budeme říkat **hlavní blok**.

Jako kritérium pro rozlišení hlavních bloků budeme využívat časový rozestup mezi jednotlivými hlášeními a to především z toho důvodu, že GPS souřadnice zastávek jsme měli k dispozici až v konečné fázi zpracování této práce, kdy celý zde popsaný postup již byl vytvořen.

Definujme si, co konkrétně budeme chápat pod pojmem časový rozestup mezi hlášeními. Připomeňme, že vozidlo zasílá procesní data vždy s určitým časovým rozestupem. O tomto rozestupu platí následující:

i. **Doba mezi hlášeními** vlivem vnějších podmínek provozu nebo třeba aktuální verzí softwaru **není vždy konstantní**,
ii. Procesní data zasílá vozidlo standardně i v případě, že **stojí na zastávce či konečné stanici (čekající na obrátku)** nebo je **v režimu aktivního odstavení například v depu.**

Budeme-li pracovat s pojmem **„časový rozestup po sobě jdoucích hlášení“**, budeme tím chápat vždy takové Δt, které od sebe dělí dvě hlášení, při kterých vozidlo mělo

nenulovou rychlost. Jedním z hlavních bodů předzpracování dat, než začneme s analýzami, bude tím pádem odstranění všech hlášení, kdy vozidlo stojí, a tedy má nulovou rychlost – k tomu více v podkapitole 4.2.

Pokud v pročištěném souboru, tj. takovém, kde hlášení s nulovou rychlostí byla odstraněna, bude mezi **dvěma po sobě jdoucími hlášeními velký časový rozestup**, bude mezi těmito hlášeními stanovena hranice dvou hlavních bloků. Jak přesně ale tento „práh" zvolit? Časové rozestupy budou během regulérní jízdy velmi malé, naopak pokud vozidlo stojí v depu, tak dvě hlášení může dělit i několik hodin. „Správný" práh tedy bude ležet někde mezi těmito dvěma oblastmi. K této otázce se vrátíme ke konci této podkapitoly v rámci krátké analýzy časových rozestupů (viz Obrázek 3.4 a dále).

Samotné rozdělení hlášení do hlavních bloků nám ale nebude stačit, a to z následujících důvodů:

i. **Redukce rozsahu dat:** v závislosti na délce jízdy může do jednoho hlavního bloku spadat i několik set jednotlivých hlášení,
ii. **Sledování parametrů po úsecích:** pokud bychom například chtěli provádět detailnější analýzy spotřeby energie v závislosti na profilu trati, budeme potřebovat shrnutí důležitých hodnot po úsecích určité délky.

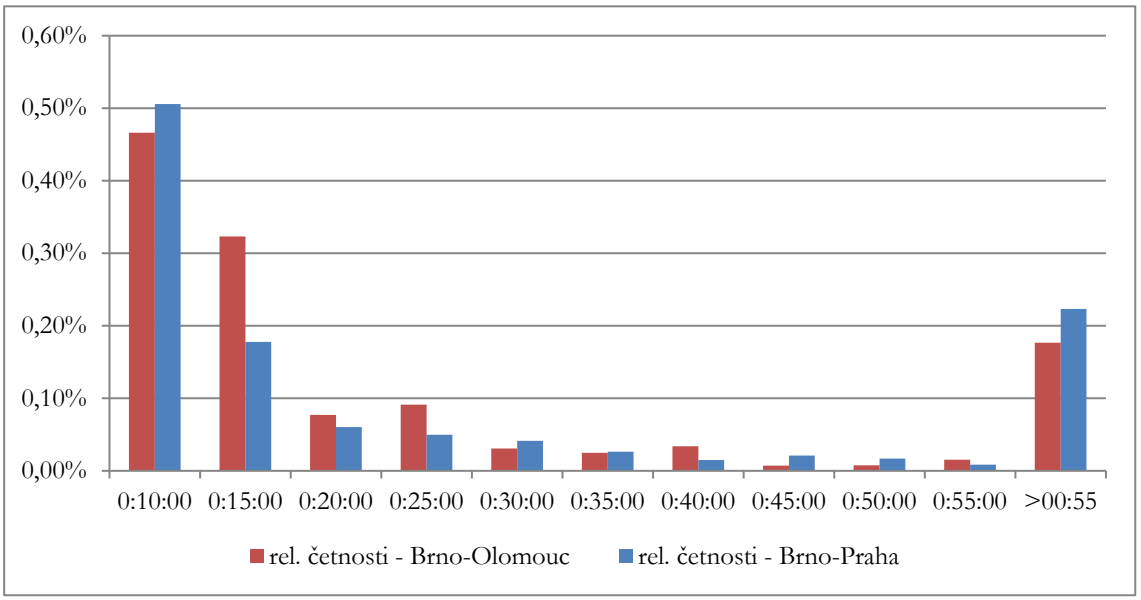
Zavedeme si proto další pojem – **subblok**. Každý jednotlivý **subblok** představuje úsek **v rámci jednoho hlavního bloku**, ve kterém bychom chtěli spotřeby energií sečíst.

Jako **kritérium pro rozdělení do subbloků** bychom mohli opět využívat časové hledisko a tímto způsobem sledovat parametry vždy po určitém časovém intervalu. Vzhledem k tomu, že po celou jízdu **vozidlo nemá konstantní rychlost**, docházelo by k tomu, že **subbloky by měly** z části i **výrazně odlišné délky úseků** (odpovídá to například případu, kdy v jednom subbloku by byla obsažena zastávka a rychlosti by byly nižší. Naopak v tom dalším subbloku by jelo vozidlo maximální rychlostí a ujelo by během stejného časového rozpětí výrazně větší počet kilometrů). Takto získaná data by proto nebyla zcela vypovídající – vždy nás více zajímá spotřeba podle ujetých kilometrů, než spotřeba po časovém úseku.

Proto budeme **hlášení do subbloků rozdělovat podle počtu ujetých kilometrů**. Postupovat budeme tak, že mezi hlášeními vypočteme vzdálenost GPS souřadnic dle kapitoly 3.3 a vzdálenosti budeme postupně sčítat. Sčítanec (a s ním spojené hlášení), jehož přičtením celková suma přesáhne zvolenou maximální délku jednoho úseku, zařadíme do dalšího subbloku.

Číslo hlavního bloku a subbloku nám dá číslo samotného bloku – například hlavní blok s číslem 2 a subblok s číslem 10 určuje blok 2.10.

Zvolený přístup rozdělení hlášení do hlavních bloků má **velkou výhodu** v jednoduchosti svého provedení a navíc zohlední i ty případy, kdy časový prostoj mezi jízdami je velký.



Obrázek 3.4: Histogram rozdělení časových rozestupů mezi hlášeními

V praxi se ale může stát, že pouze časová hranice pro rozlišení hlavních bloků nestačí. Obrázek 3.4 představuje **histogram relativních četností hodnot časových rozestupů mezi po sobě jedoucími hlášeními**, které byly sesbírány pro šest konkrétních vozidel, která se během jednoho měsíce pohybovala **téměř výhradně** na trati Brno-Olomouc, a další tři vozidla, pro která platí totéž, co v případě trati Brno-Praha.

U všech vozidel byla odstraněna ta hlášení, která uvádí rychlost vozidla jako nulovou. Popisky vodorovné osy odpovídají horní hranici jednotlivých tříd v minutách. Z obrázku jsme vynechali první třídu, ve které by byly četnosti hodnot menší než pět minut. Ta tvoří naprostou většinu ze všech hlášení (v obou případech bezmála 99 %) a odpovídá nejkratším časových rozestupům, kdy vozidlo je v pohybu na trati a pravidelně zasílá hlášení. Poslední třída obsahuje všechna Δt, která jsou větší než 55 minut a klasickým příkladem takovéhoto případu je odstavení vozidla v depu.

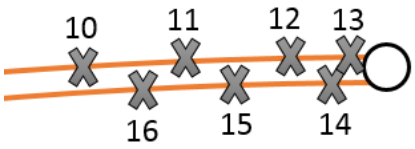
Budeme-li na základě toho, co uvádí Obrázek 3.4, **uvažovat mez pro hlavní bloky například zmíněných 30 minut**, tak ta sice rozpozná případy, kdy bylo vozidlo odstaveno na delší dobu v depu, ale pokud se pohybuje na jedné trati a po jízdě „tam“ bez větší prodlevy se začne vracet „zpět“, může se stát, že tyto **dvě jízdy budou součástí jednoho hlavního bloku** a získaná data pro spotřebu energie apod. budou zkreslená.

Pokud bychom **časovou mez pro hlavní bloky volili menší**, například 10 nebo 15 minut, tak se v listu „bloky“ velmi pravděpodobně zaregistruje více případů „rychlého návratu“, který byl popsán v předchozím odstavci, ale na druhou stranu dojde k nesprávnému rozdělení do hlavních bloků i v případě, kdy se vozidlo zdrží na jedné ze zastávek. Ve výsledku by se tedy mohlo stát, že **jedna jízda bude rozdělená do několika hlavních bloků**.

Popsané dva problémy nás vedou k tomu, že je nutné po rozdělení hlášení dle časového měřítka provést dodatečně kontrolu hlavních bloků.

Jednodušší pro nás je dodatečně hlavní bloky rozdělit než dva k sobě patřící hlavní bloky dodatečně spojovat – ošetříme tedy výskyt prvního problému. Proto také zvolíme jako práh, kterým od sebe oddělíme hlavní bloky, větší, aby pokud možno docházelo pouze k výskytu prvního problému. Podle hodnot a pro dálkové soupravy, které představuje Obrázek 3.4, toto splňuje práh s hodnotou 30 minut.

3.5 Kontrola hlavních bloků



Obrázek 3.5: Dvě jízdy v jednom hlavním bloku - detail

Jak tedy zajistíme dodatečně rozpoznání toho, že jsme nesprávně do jednoho hlavního bloku „vložili“ dvě jízdy? Obrázek 3.5 nám znázorňuje situaci, kdy máme dvě jízdy v jednom hlavním bloku a byly vykresleny poslední souřadnice z každého subbloku s jeho číslem. Na první pohled je evidentní, že během 13. subbloku dojelo vozidlo do konečné stanice a následně se začalo vracet – čili správně by skupina subbloků počínaje číslem 14 a výše měla patřit do dalšího hlavního bloku.

Vyhodnocení toho, co je z grafického znázornění zřejmé, je ale nutné nějakým způsobem vyjádřit analyticky. Vraťme se proto k situaci, kterou představuje Obrázek 3.5. **Vizuálně** poznáme souřadnice dvou jízd na stejné trati podle toho, že **si jsou vzájemně blízké**. Tuto myšlenku převedeme do čísel ve smyslu **vzájemných vzdáleností**. Konkrétně se na obrázku zaměříme na vzdálenosti **prvního a posledního**, posléze **druhého a předposledního subbloku**. Pokud by jízda po subbloku 13 pokračovala víceméně dál ve směru rovně, očekávali bychom mezi subbloky 10 a 16 vzdálenost okolo 60 kilometrů a mezi subbloky 11 a 15 okolo 40 kilometrů[12] - toto zde evidentně platit nebude a **obě vzdálenosti budou daleko menší**.

Přesně na této skutečnosti založíme **kontrolu hlavních bloků.** Označíme-li SB_t jako subblok na t-té pozici daného hlavního bloku, $d(SB_t, SB_{t-1})$ jako vzdálenost posledních souřadnic subbloků SB_t a SB_{t-1} a ε bude námi volená mez vzdálenosti, tak platí, že pokud je splněno

$$(d(SB_{t+3}, SB_{t-3}) \leq \varepsilon) \wedge (d(SB_{t+2}, SB_{t-2}) \leq \varepsilon), \tag{3.3}$$

tak naším vyhodnocením bude, že SB_t bude subblokem, v jehož průběhu se vozidlo začíná vracet po stejné trati, a SB_{t+1} zařadíme do nového hlavního bloku.

[12] V tomto příkladu jako maximální délku subbloků volíme 10 kilometrů

Hodnotu $d(SB_{t+1}, SB_{t-1})$ zkoumat nebudeme, jelikož vozidlo od kroku $t-1$ do $t+1$ nestihne ujet takovou vzdálenost, podle které by se dalo rozpoznat, zda se již během této doby vracelo, nebo jelo naopak po trati dál.

Implementace postupu při kontrole hlavních bloků je provedena v makru „kontrolaJizdPreproc“ a využívá se v sešitu „tvorbaReportu.xlsb“. Jednotlivé kroky kódu jsou součástí preprocessingu dat v podkapitole 4.2, kde si je představíme blíže.

3.6 Koncept podobnosti jízd a jejich klasifikace do tratí

3.6.1 Obecný postup

Pokud bychom chtěli porovnávat spotřeby energie a alarmy na jednotlivých tratích, narazíme na zásadní problém – **nelze bez dodatečných informací jednoznačně říct, kdy se konkrétní vozidlo bude pohybovat na jaké konkrétní trati.**

Samozřejmě platí, že se vozidla mohou pohybovat pouze na takových tratích, které odpovídají trakčnímu vybavení vozidla – výběr se tím však nezúží v dostatečné míře. Provozovatel může nasazovat vozidla dle potřeby a tím pádem se nemůžeme spolehnout na to, že by námi sledované tratě vozidla projížděla například vždy v pondělí, středu a pátek mezi desátou a dvanáctou. Shrnuto: **pouze časová informace z hlášení je pro přiřazení ke konkrétní trati nedostačující.**

Dále máme možnost využít **informaci o poloze vozidla**, jelikož součástí každého hlášení jsou i GPS souřadnice. I zde se ale musíme vypořádat s několika problémy. Budeme-li uvažovat, že máme skupinu souřadnic pro několik jízd určitého vozidla, souřadnice všech tratí, a budeme chtít určit trať, které přísluší, tak platí následující:

i. **Jeden pár souřadnic nám nedává dostatečnou informaci** – tratě se mohou různě prolínat (viz (1)),
ii. V závislosti na algoritmickém řešení problému by **porovnání každého jednotlivého páru souřadnic** z jízd se souřadnicemi z tratí nebo jiných jízd mohlo být **z hlediska náročnosti výpočtu neúnosné** (velké objemy dat z jízd oproti velkému počtu souřadnic z tratí).

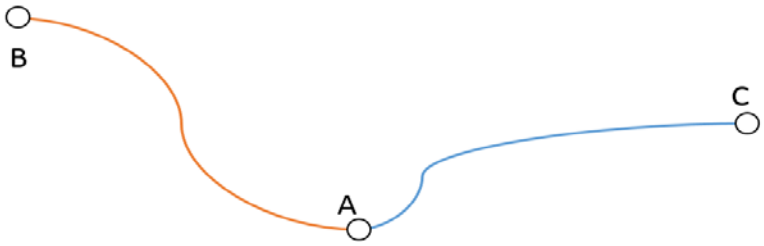
Abychom se vyhnuli oběma problémům, musíme učinit určitý **kompromis**. Ten bude spočívat v tom, že z každé jízdy budeme počítat jen s jistou **podmnožinou souřadnic**, abychom snížili počet porovnání a tím i čas potřebný pro provedení výpočtu. S tímto rozhodnutím se ale dostáváme k otázce pro další postup naprosto zásadní – podle jakého kritéria budeme vybírat souřadnice pro zmíněnou podmnožinu?

Cílem je, aby souřadnice dokázaly vhodně reprezentovat trať. K tomu zcela jistě není zapotřebí všech souřadnic, které vozidlo zasílá v intervalu několika minut či vteřin – postačí nám jen jejich určitá část. Na tomto místě je důležité si uvědomit, že nemůžeme vybrat libovolnou podmnožinu, jelikož nám například prvních deset hodnot velmi pravděpodobně nesplní podmínku vhodné reprezentace trati.

Budeme tedy chtít vybrat souřadnice tak, abychom **významně zredukovali rozsah párů**, se kterými budeme pracovat, ale zároveň by měly mít **souřadnice mezi sebou co nejpravidelnější rozestupy**, abychom získali za daných podmínek co nejkvalitnější představu o tom, kudy vozidlo jelo.

Toho se pokusíme dosáhnout **vhodným rozdělením všech hlášení do bloků a subbloků**, kde **bloky** by měly odpovídat **jednotlivým jízdám** a **subbloky** určité ujeté vzdálenosti, **například desetikilometrovému úseku na trati** (viz kapitola 3.4). Naší podmnožinou souřadnic, kterou využijeme pro klasifikaci jízd do tratí, bude vždy **jeden vybraný pár z jednoho subbloku** – nejjednodušeji první nebo poslední, abychom se pravidelnosti rozestupů přiblížili co nejvíce.

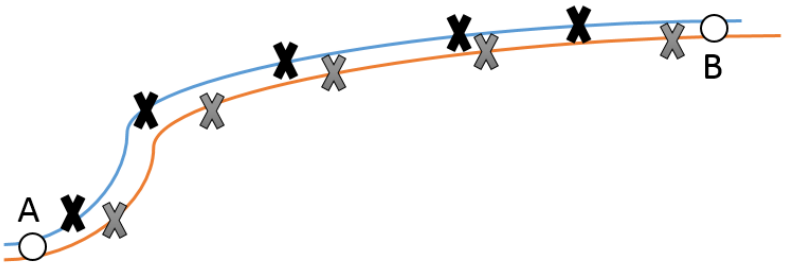
Způsobem dělení všech hlášení do bloků a subbloků se budeme podrobněji zabývat v kapitole 4, která je věnovaná analýze spotřeby energie.



Obrázek 3.6: Různé tratě s podobnou délkou

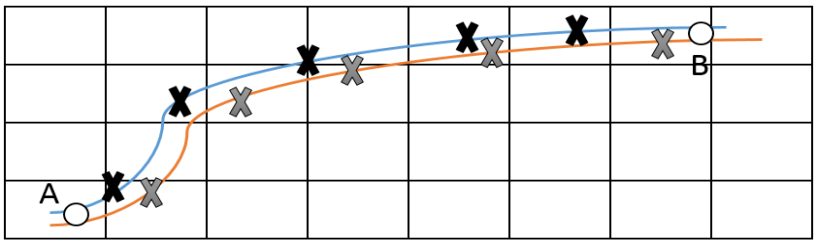
Jak budeme **jednotlivé jízdy mezi sebou porovnávat?** Jednou možností je **porovnávat součet celkově ujetých kilometrů** jednotlivých jízd a za podobné považovat ty s podobnou délkou. Toto kritérium nám však **nemůže stačit**, jelikož může dojít na situaci, kterou znázorňuje Obrázek 3.6.

Nemůžeme se ani spoléhat na to, že pro jízdy na stejné trati budou souřadnice sobě velmi blízké – vozidla sice zasílají hlášení vcelku pravidelně, ale **ne ze stejných míst nebo ve stejný čas**. Proto nebudeme porovnávat jednotlivé hodnoty souřadnic v jejich původním stavu, tj. s pěti místy za desetinnou čárkou, ale pokusíme se **souřadnice sdružovat k určitému okolí**. Ideálně by pak mělo dojít na situaci, kdy se souřadnice z jednotlivých subbloků různých jízd **dostanou do stejného okolí** – to by nám dovolilo jejich **podobnost určit pomocí „počtu shodných okolí“.**



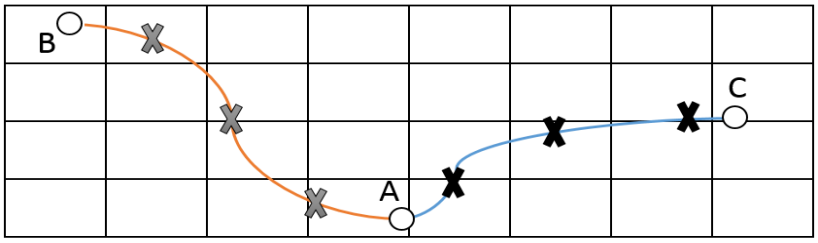
Obrázek 3.7: Dvě jízdy s vybranou podmnožinou souřadnic

Princip ilustrujeme na dvou obrázcích. Obrázek 3.7 obsahuje dvě jízdy, které pro lepší přehlednost jsou od sebe lehce odsazené. Je zřejmé, že vybrané podmnožiny GPS souřadnic popisují stejnou trať, a to přestože si některé dvojice jsou z hlediska vzájemné vzdálenosti bližší, naopak jiné vzdálenější. **Jak ale toto zjištění kvantifikovat?** Oblast, kterou jízdy probíhají, se pokusíme rozdělit do zmíněných okolí nebo úseků.



Obrázek 3.8: Dvě jízdy v mřížce

Obrázek 3.8 obsahuje stejné jízdy jako Obrázek 3.7 s tím rozdílem, že jsme je **promítli na mřížku.** Celou úlohu si tím dokážeme zjednodušit, jelikož nemusíme počítat vzdálenosti jednotlivých souřadnic a podle nich se poté rozhodovat, zda jízdy patří ke stejné trati či nikoli. Naopak budeme postupovat tak, že pro **každou souřadnici určíme, které buňce v mřížce přísluší**. Poté srovnáme buňky, kterými jízdy prochází, a dopočteme **počet těch buněk, kterými prochází obě jízdy**. Pokud se jízdy konaly na stejné trati, shodnost buněk by měla být větší, pokud se vozidlo pohybovalo během dvou jízd na odlišných tratích, tak bude shodnost buněk menší či téměř nulová (viz Obrázek 3.9).

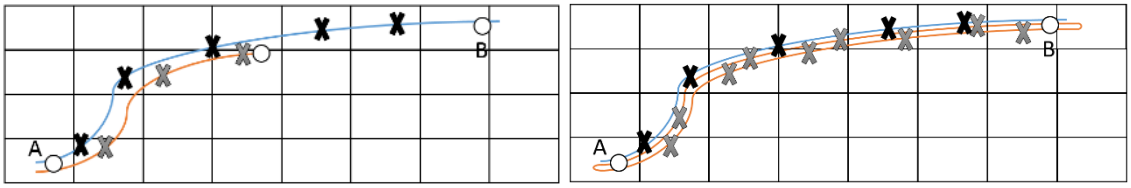


Obrázek 3.9: Dvě jízdy na odlišných tratích v mřížce

Postup založený na posouzení míry shodnosti buněk nám pomůže **od sebe odlišit tratě**, jejichž průnik je velmi malý. Mohou však nastat situace, kdy nám tato metodika vyhodnotí jízdy jako zdánlivě podobné, ve skutečnosti tomu tak ale nebude.

Právě takovéto případy ilustruje Obrázek 3.10. Vlevo je **jedna jízda výrazně kratší**, což může v reálu odpovídat jízdě z nebo do depa, a vpravo naopak je uveden příklad **jízdy tam i zpět**. Shodnost buněk bude obzvlášť v druhém případě velmi vysoká – takové vyhodnocení by bylo z hlediska dalších výpočtů velmi nevhodné: například v rámci srovnání spotřeby energie na jedné konkrétní trati v rámci jedné jízdy by mohlo docházet k silnému zkreslení výsledků.

Abychom se vyhnuli těmto případným chybám, tak jako **kritérium pro podobnost jízd** budeme mimo **postup se shodnými buňkami** využívat také **srovnání jejich délky**, tj. pokud jízdy budou vyhodnoceny jako shodné, pokud počet shodných buněk v mřížce, kterými jízdy prochází, přesáhne zvolenou mez a zároveň délka jízd bude s určitou tolerancí shodná. Zároveň při zpracování budeme využívat postup v kapitole 3.5, tj. kontroly hlavních bloků.



Obrázek 3.10: Dvě jízdy na stejné trati – různé délky

3.6.2 Realizace na datech

Budeme předpokládat, že máme data vhodně rozdělená do hlavních bloků a subbloků, takže **všechna hlášení a GPS souřadnice patřící do jednoho hlavního bloku patří k jedné jízdě** a jednotlivé ujeté vzdálenosti v rámci jednoho subbloku jsou přibližně konstantní. Vzhledem k těmto skutečnostem a taky proto, abychom zmenšili rozsah dat, se kterými budeme dál pracovat, z každého subbloku vybereme z časového hlediska poslední zaslané hlášení a tím i poslední pár souřadnic.

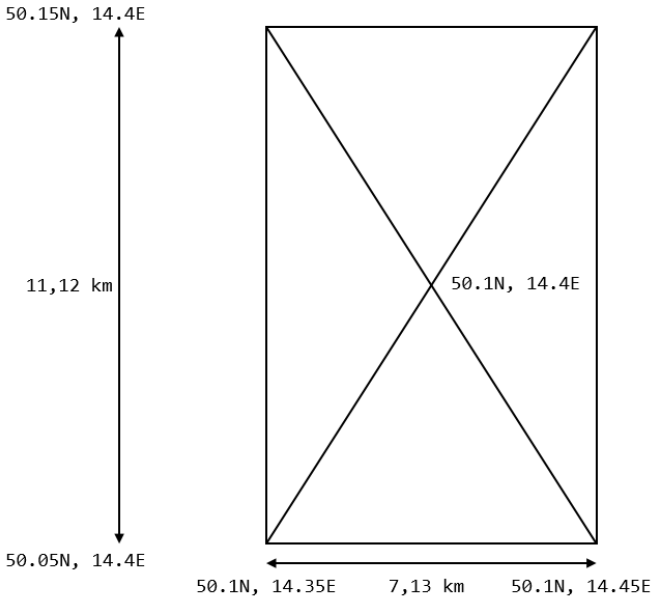
Ze **stejné délky subbloků plyne**, že i **vzdálenosti mezi takto vybranými páry souřadnic by měly být přibližně konstantní**, čímž jsme splnili jednu z našich podmínek z předchozí kapitoly. Počet celkově ujetých kilometrů během dané jízdy určíme však na základě všech hlášení patřících do daného hlavního bloku, jelikož by jinak mohlo dojít ke ztrátě na přesnosti.

Naším dalším krokem bude **stanovení mřížky**, na kterou souřadnice promítneme. Důležité je, aby **jednotlivé buňky měly stejnou velikost** a zároveň pozdější přiřazení souřadnic k patřičné buňce bylo co nejjednodušší a co nejméně výpočtově náročné. Taktéž budeme chtít zachovat tvar mřížky, takže buňky by měly mít tvar obdélníku.

Místo toho, abychom začali hledáním hranic jednotlivých buněk, pokusme se úlohu zjednodušit tím, že najdeme **vhodné středy buněk**, které splní námi požadované vlastnosti:

i. **Stejná velikost buněk:** splníme tím, že sousední středy buněk nebo obdélníku budou mít od sebe stejnou vzdálenost. Poté bychom na jejich základě určili délku a šířku buněk,
ii. **Malá náročnost výpočtu:** určení buňky, do které souřadnice patří, by znamenalo porovnávat každou složku souřadnice s dvěma hodnotami, pokud by samotné buňky byly určeny na základě složitějšího vzorce či postupu. Naším cílem bude vytvořit jednodušší postup pomocí středů buněk, tj. nebudeme porovnávat složky souřadnic s hranicemi buňky, ale z hodnoty složek souřadnice rovnou střed buňky, do které bude patřit.

Obě tato kritéria budou splněna, pokud středem buněk bude **každá dvojice souřadnic s přesně jedním místem za desetinnou čárkou**. U **první podmínk**y to je zřejmé – elipsoidním tvarem zeměkoule sice je dáno to, že se vzdálenosti dvou od sebe nominálně stejně odsazených souřadnic budou lišit, pokud je umístíme na různých místech, ale tyto rozdíly nejsou natolik velké, aby později dokázaly ovlivnit výsledky určení podobnosti jízd výraznějším způsobem.



Obrázek 3.11: Ilustrace parametrů jedné buňky

Druhá podmínka je splněna taktéž, jelikož nám k tomu, abychom z hodnot složek GPS souřadnic získali příslušný střed buňky, stačí pouhé zaokrouhlení jejich hodnot na desetiny.

hlavní blok	GPS_Latitude	GPS_Longitude	GPS_Latitude (zaokrouhl.)	GPS_Longitude (zaokrouhl.)
1	49,21083	16,64409	49,2	16,6
1	49,13523	16,60621	49,1	16,6
1	49,03439	16,59126	49	16,6
1	48,93017	16,64893	48,9	16,6
1	48,89222	16,77293	48,9	16,8
1	48,83854	16,85287	48,8	16,9
1	48,75858	16,89787	48,8	16,9
1	48,80231	17,02844	48,8	17
1	48,85656	17,12275	48,9	17,1
2	48,84926	17,10724	48,8	17,1
2	48,7877	16,98352	48,8	17
2	48,77802	16,91287	48,8	16,9
2	48,85105	16,83905	48,9	16,8
2	48,91288	16,70554	48,9	16,7
2	48,97694	16,60014	49	16,6
2	49,08657	16,60537	49,1	16,6
2	49,1837	16,60745	49,2	16,6
3	49,11319	16,60584	49,1	16,6
3	49,05772	16,59818	49,1	16,6
3	48,9381	16,63534	48,9	16,6

nastavit počet míst pro zaokrouhlení

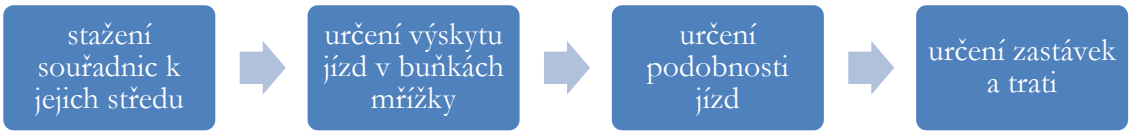
nastavit mez pro hlavní bloky

provést přepočet podobnosti jízd

Obrázek 3.12: Ukázka z listu „výběr bloků“

V kapitole 4 se budeme věnovat spotřebám energie a mimo jiné také tvorbě reportů, na kterých založíme analýzy. **Určení podobnosti jízd je důležitou součástí tvorby celého reportu.**

Při jeho tvorbě je vygenerován list „výběr bloků“ (viz Obrázek 3.12), kde má **uživatel možnost změnit počet míst za desetinnou čarou**, na které se zaokrouhlí souřadnice (standardně v této práci budeme pracovat se zaokrouhlením na desetiny) a také může **nastavit horní mez pro hlavní bloky**, tj. kolik subbloků hlavní blok (respektive jízda) musí mít, aby byl zahrnut do výpočtů podobnosti jízd.



Obrázek 3.13: Shrnutí postupu při určení podobnosti jízd

Stažení souřadnic k jejich středu bude také **prvním krokem v rámci určení podobnosti jízd** a je provedeno právě na listu „výběr bloků“ v reportech ke spotřebám energie. Po nastavení parametrů a stisknutí tlačítka pro přepočet podobnosti jízd je spuštěno makro „srovnavaniJizd“, které stejně jako všechna ostatní makra vytvořená v rámci této práce je součástí souboru „PERSONAL_DP“. Výpočty jsou z velké části založeny na kontingenčních tabulkách. Nejprve je vytvořen dočasný list „tempe“ a do něj je vložena první kontingenční tabulka, která se vztahuje na data z listu „výběr bloků“. Jako **popisky řádků nastavíme kombinaci zaokrouhlených GPS souřadnic** a ve sloupcích budou naopak **čísla vybraných hlavních bloků (tj. těch, jejichž počet subbloků přesahuje námi volenou mez)**. Ukázku z tohoto listu představuje Obrázek 3.14.

Počet z hlavní blok		hlavní blok (výběr)							
GPS_Latitude (z	GPS_Longitude (zaokrouhl.)	3	4	7	8	12	13	14	15
⊟48,8	16,8								
	16,9	2	2	2	2	1	1	2	1
	17	1		1	1	1	1	1	1
	17,1		1			1			1
⊟48,9	16,6	1	1						
	16,7			1	1	1	1	1	1
	16,8	1	1	1	1	1	1		1
	17,1	1			1		1	1	
	17,2	1	1	1		1			1
	17,3				1	1	1	1	
⊟49	16,6		1	1	1	1	1	2	1
	17,3	1	1	1	1		1	1	1
	17,4			1		1			
⊟49,1	16,6	2	1	1	1	1	1	1	1

Obrázek 3.14: Ukázka z listu „tempe“

Hodnoty v tabulce poté udávají četnosti zaokrouhlených hodnot souřadnic v dané jízdě. Pokud se vrátíme k interpretaci pomocí mřížky a buněk, tak **každý řádek v této tabulce představuje jednu konkrétní buňku**, kterou identifikujeme přes její střed, a **zmíněné četnosti odpovídají počtu souřadnic, které do dané buňky spadají.**

Dalším krokem je **pro každou kombinaci dvou jízd určit počet shodných výskytů souřadnic v buňkách**. V rámci makra „srovnavaniJizd“ jsou proto vytvořeny další dva listy – „tempe2“ a „podobnostJizd“. **Druhý jmenovaný list** bude obsahovat tabulku či matici o velikosti $n \times n$, kde n je počet hlavních bloků zahrnutých do výpočtů. Popisky řádků a sloupců odpovídají číslům hlavních bloků (respektive jízd) a **obsahem této tabulky**

bude číselné vyhodnocení podobnosti jízd, tj. prvky této matice a_{ij} představují podobnost jízd na i-tém řádku a j-tém sloupci. Jelikož každou dvojici jízd budeme porovnávat jen jednou, abychom nenavyšovali zbytečně náročnost výpočtu, a také nebudeme počítat podobnost dvou stejných jízd (tj. vylučujeme prvky matice, kde $i = j$), budeme celkově počítat $\binom{n}{2}$ podobností, a po ukončení všech výpočtů bude mít tabulka na listu „podobnostJizd" tvar horní trojúhelníkové matice bez diagonály. Obrázek 3.16 představuje část této tabulky po provedení výpočtů.

Jak ale určíme jednotlivé prvky a_{ij}? Právě zde hraje roli list „tempe2", kde je kontingenční tabulka vztahující se na data z listu „tempe", pomocí níž dopočteme, do jaké **míry prochází jízdy shodnými buňkami**. Poté, co makro „srovnavaniJizd" vytvoří listy „tempe2" a „podobnostJizd", na druhém jmenovaném listu vybere novou kombinaci čísel dvou hlavních bloků, přepne na list „tempe2" a použije tato dvě čísla jako parametry pro popisky řádků a sloupců. Výsledkem je tabulka, jakou ji představuje Obrázek 3.15. Interpretovat ji je třeba tímto způsobem: obě jízdy mají osm společných buněk, kde se souřadnice k těmto jízdám vyskytly jen jednou. Dále existují dva případy, kde první jízda prošla jednou buňkou dvakrát, zatímco druhá jen jednou a jeden případ, kde obě jízdy jednou konkrétní buňkou prošly dvakrát.

K určení počtu buněk, které projely obě uvažované jízdy, si vytvoříme pomocnou tabulku, která bude mít stejné rozměry, jako má tabulka, kterou znázorňuje Obrázek 3.15 – v našem příkladu tedy 2x2. Její hodnoty dopočteme podle vzorce $\min(u_r, v_s)$, kde u_r je hodnota popisku r-tého řádku a v_s popiska s-tého sloupce kontingenční tabulky. V našem příkladu to znamená pro hodnotu druhého řádku a prvního sloupce pomocné tabulky:

$$\min(u_2, v_1) = 1$$

Znamená to, že do celkového počtu shodných buněk bude každý případ, kde jedna jízda prošla určitou buňkou dvakrát a druhá jízda jen jednou, započten pouze jednou.

Počet z buněk	Popisky sloupců		
Popisky řádků	**1**	**2**	**Celkový součet**
1	8		8
2	2	1	3
Celkový součet	**10**	**1**	**11**

Obrázek 3.15: Ukázka z listu „tempe2"

Obě tabulky, tj. kontingenční i pomocné, se převedou na vektory a jejich skalárním součinem získáme počet společných výskytů souřadnic v buňkách, což bude celé číslo z uzavřeného intervalu omezeného zdola nulou a shora minimem z počtu subbloků obou jízd. Označíme-li počet společných výskytů jako x_{ij}, bude platit:

$$x_{ij} \in \langle 0, \min(m_i, m_j) \rangle \cap \mathbb{Z},$$

kde m_i a m_j je počet subbloků z jízdy na i-tém řádku, respektive j-tém sloupci tabulky na listu „podobnostJizd".

Předtím, než tuto hodnotu evidujeme v tabulce na listu „podobnostJizd“, provedeme úpravu. x_{ij} **nemusí mít nutně vypovídací hodnotu**, což by nastalo v případě, kdyby se sledované vozidlo pohybovalo na tratích výrazně odlišných délek (jedna by měla například okolo 15 subbloků a druhá okolo 30). Došlo by tak na případ, který jsme popsali v předchozí kapitole (viz Obrázek 3.10). x_{ij} je nutné brát v kontextu s počtem subbloků příslušných jízd, na základě kterého můžeme provést úsudek o podobnosti. Proto budou prvky a_{ij} vypočteny dle následujícího vztahu:

$$a_{ij} = \frac{x_{ij}}{\max(m_i, m_j)} \tag{3.3}$$

Srovnáme-li toto vyjádření se zápisem intervalu možných hodnot x_{ij} výše, zjistíme, že platí:

$$\max(m_i, m_j) \geq x_{ij} \tag{3.4}$$

Důsledkem je, že platí $a_{ij} \in \langle 0,1 \rangle$, což znamená, že a_{ij} **můžeme chápat jako procentuální podobnost jízd** na i-tém řádku a j-tém sloupci tabulky z listu „podobnostJizd“. Maximum hodnot m_i a m_j bereme z toho důvodu, abychom zabránili nepřesnému vyhodnocení vysoké podobnosti jízd v situaci, jakou představuje Obrázek 3.10, respektive jeho levá část – dvě jízdy různé délky, které se bez ohledu na konkrétní důvod vyskytují na stejné trati. Zvolením $\max(m_i, m_j)$ namísto $\min(m_i, m_j)$ do jmenovatele tuto potencionálně až 100% podobnost uměle snížíme a tím pádem a_{ij} získává na vypovídací hodnotě.

	celkem subbloků	4	7	8	12
3	17	0,7059	0,7059	0,6667	0,6471
4	17		0,7647	0,7778	0,7059
7	17			0,7778	0,7647
8	18				0,7222
12	17				

Obrázek 3.16. Ukázka list „podobnostJizd“ po provedení výpočtů

Výpočet této podobnosti pomocí kontingenční tabulky na listu „tempe2“ provedeme tolikrát, kolik je prvků v tabulce na listu „podobnostJizd“ - tedy $\binom{n}{2}$-krát. Makro poté listy „tempe“ a „tempe2“ smaže, vloží do listu „podobnostJizd“ tlačítko „nastavit mez podobnosti a vytvořit soupis tratí“, zavolá další makro „podobnHighlightSoupis“ a poté skončí.

Zmíněné spuštěné makro si od uživatele přes dialogové okno vyžádá zadání meze podobnosti. Na základě této hodnoty se poté provedou dva úkony:

i. V tabulce na listu „podobnostJizd“ se označí zeleně prvky a_{ij} přesahující zvolenou mez,
ii. Vytvoří se list „soupis tratí“, kde se na základě zvolené meze sepíšou jízdy do tratí, určí se jejich směr a zastávky, kterými trať prochází.

Druhou část rozebereme trochu více do detailu. **Sepsání jízd do tratí** se provádí **na základě hodnot prvků** a_{ij}, které se postupně po řádcích všechny projdou.

Při započetí procházení každého řádku proběhne kontrola, zda hlavní blok na řádku i je již v soupisu tratí obsažen. Pokud tomu tak není, je vytvořen **nový typ tratě** a hlavní blok na i-tém řádku bude takzvanou **referenční jízdou ve směru tam** (v tabulce je označen jako „ref"). Při procházení prvků a_{ij} v i-tém řádku platí, že pokud je **hodnota prvku větší než zvolená mez a zároveň bude délka jízd s určitou tolerancí stejná** (standardně budeme uvažovat 10 %)[13], připíše se hlavní blok pod indexem j na listu „soupis tratí" do právě toho typu tratě, ve kterém je obsažen hlavní blok pod indexem i.

Pokračováním je **určení směru jednotlivých jízd**. Procentuální podobnost jízd a_{ij} nevypovídá nic o směru jednotlivých jízd. Jak bylo uvedeno v předchozím odstavci, tak první jízda v každém typu tratě je uvažovaná jako referenční ve směru tam. Směr ostatních jízd určíme pomocí **vzdáleností GPS souřadnic z prvních subbloků** každé z dvojic uvažovaných jízd – pokud budou menší, než námi volená mez, tak bude jízda vyhodnocena jako ve směru „tam", v opačném případě bude vyhodnocení „zpět".

V poslední řadě provedeme **výpis zastávek patřících k danému typu tratě**. Toto založíme opět na výpočtu vzdáleností GPS souřadnic – pro každý typ tratě vybereme prvních pět jízd ve směru „tam" a v každé z nich projdeme každou jednotlivou souřadnici, kterou budeme porovnávat se seznamem zastávek a jejich GPS souřadnic. **Vzdálenost zastávky a souřadnice z jízdy**, která bude menší, než námi volená mez (budeme uvažovat standardně 2,5 km) bude znamenat to, že danou zastávku přidáme do seznamu zastávek k danému typu tratě a k tomu určíme ujetou vzdálenost do této zastávky.

Jelikož se v praxi může stát, že se **zastávky nerozpoznají v jejich správném pořadí**, provedeme na konci seřazení zastávek podle ujeté vzdálenosti na trati. Zamezíme tím situaci, kdy například během procházení první jízdy jsou rozpoznány všechny zastávky až na druhou ve „správném pořadí", kterou by se podařilo identifikovat až při procházení druhé jízdy a na konci by byla nesprávně evidována až na konci seznamu.

[13] Tím jsou splněny obě podmínky z posledního odstavce kapitoly 3.4.1

4 Analýza spotřeby a rekuperace energie

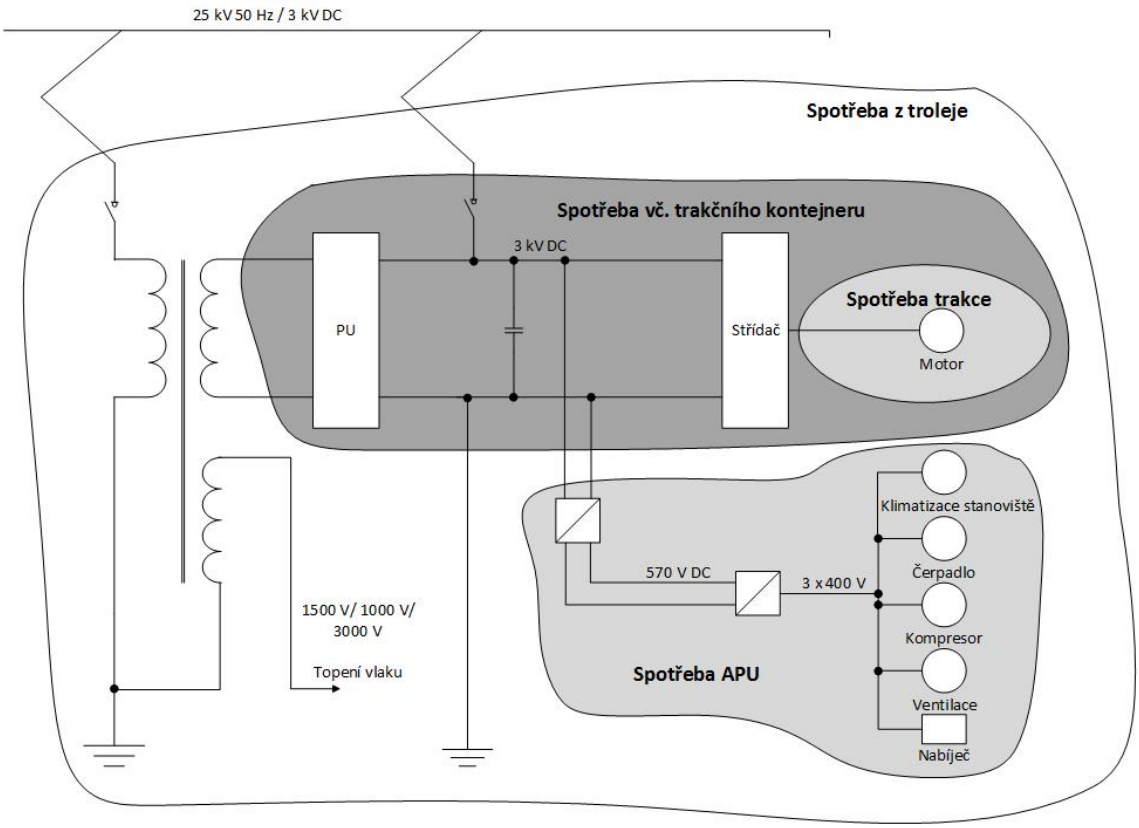
V předchozích kapitolách byly představeny nevýhody současného uložení dat, v čem přesně tyto nevýhody spočívají a také způsoby řešení dostupnosti dat v námi požadované podobě. V této kapitole dojde k reálné aplikaci těchto postupů.

Nejprve jsou uvedeny **pojmy spojené se spotřebou energie**, poté provedeme **preprocessing dat a jeho implementaci** v rámci této práce vytvořeném reportovacím systému. Výstup tohoto systému (reporty) využijeme v poslední části této kapitoly, kdy údaje o **spotřebě energie** z těchto reportů podrobíme **statistickým analýzám** a na konkrétních vozidlech zjistíme, zda mají stejnou spotřebu energie.

Postup pro vytvoření reportu a provedení analýz je uveden v uživatelském manuále, který je k této práci přiložen. Je obecně aplikovatelný na všechna vozidla.

4.1 Obecné informace o spotřebě energie

Konkrétní způsob zapojení jednotlivých spotřebičů elektrické energie je v rámci všech typů vozidel různý. Obrázek 4.1 představuje proto obecné a zjednodušené schéma elektrického obvodu, ze kterého je patrné zapojení pohonu a dalších spotřebičů v elektrické lokomotivě. Pro elektrické jednotky je schéma podobné – jediným rozdílem je zapojení topení ve vlaku, ke kterému se ještě vrátíme níže.



Obrázek 4.1: Zjednodušené schéma elektrického obvodu pohonu a APU elektrické lokomotivy

Obvod odpovídá vozidlu, které je **dvousystémové**, tudíž dokáže využít napájecí systém DC a AC. Pokud jezdí na trati se systémem AC, tak se proud vede do další soustavy přes transformátor a pulzní usměrňovač (PU), který přemění střídavý proud na stejnosměrný 3 kV DC. V případě, že vozidlo jezdí po trati se systémem DC, pak odebraná elektrická energie jde přímo do tohoto meziobvodu. Vyhlazovací kondenzátor zajišťuje co nejrovnoměrnější průběh proudu do další soustavy.

Do motoru jde proud přes střídač, který převede stejnosměrný proud na střídavý. Ostatní zařízení jsou obsloužena přes další obvod – proud vede přes dva měniče, kde se postupně z 3kV DC mění na 570 V DC a třikrát 400 V. Pak už následují nejdůležitější spotřebiče – konkrétně klimatizace stanoviště[14], čerpadlo, kompresor, ventilace i nabíječ baterie. Tato zařízení se také označují jako **pomocné pohony** (anglicky auxiliary power unit **(APU)**).

Jak bylo zmíněno, tak Obrázek 4.1 znázorňuje obvod pro elektrickou lokomotivu. V případě elektrické jednotky se mění pouze způsob zapojení topení ve vlaku, kde v meziobvodu 3 kV DC přibyde měnič na 570 V DC, který bude napájet topení a klimatizaci ve vlaku. Stále však platí, že topení nebude součástí APU.

Pro oba druhy vozidla musíme předpokládat **ztráty elektrické energie**, tj. proud dodaný do sítě nemusí odpovídat energii spotřebované jednotlivými spotřebiči. Konkrétní hodnoty jsou závislé na jejich konstrukci a charakteristice – obecně ale hrají ztráty transformátoru a PU důležitou roli, mohou se pohybovat v hodnotách několika set kilowatt.

Na vozidlech ŠTRN se sledují tyto hodnoty:

i. Spotřeba trakce (sítě AC i DC),
ii. Spotřeba včetně trakčního kontejneru (pouze sítě AC),
iii. Spotřeba APU (sítě AC i DC),
iv. Spotřeba z troleje (sítě AC i DC),
v. Spotřeba v režimu aktivního odstavení (sítě AC i DC),
vi. Rekuperace z troleje (sítě AC i DC).

Jednotlivé pojmy z části uvádí i Obrázek 4.1. **Celkově odebraná elektrická energie** se označuje jako **spotřeba z troleje** – analyticky to platí pro **rekuperaci z troleje**. Jako dílčí spotřeby se sleduje **spotřeba včetně trakčního kontejneru** (na základě softwarového řešení vozidel pouze na sítích se systémem AC), která zahrnuje **PU, střídač a motory** a **spotřeba APU**. **Spotřeba motorů** se sleduje samostatně jako **spotřeba trakce** a tedy pokud budeme chtít analyzovat spotřebu na pohon vozidla, bude nás zajímat tato hodnota. Nezávisle na již řečeném se sleduje **spotřeba v režimu aktivního odstavení**. Obecně platí, že se spotřeby energie sledují v kilowatthodinách.

4.2 Preprocessing dat

Údaje o spotřebě energie jsou součástí každého hlášení v rámci procesních dat, tj. v **pravidelných intervalech** se eviduje aktuální stav celkové spotřeby jednotlivých oblastí a spotřebičů, které jsme si představili v předchozím odstavci.

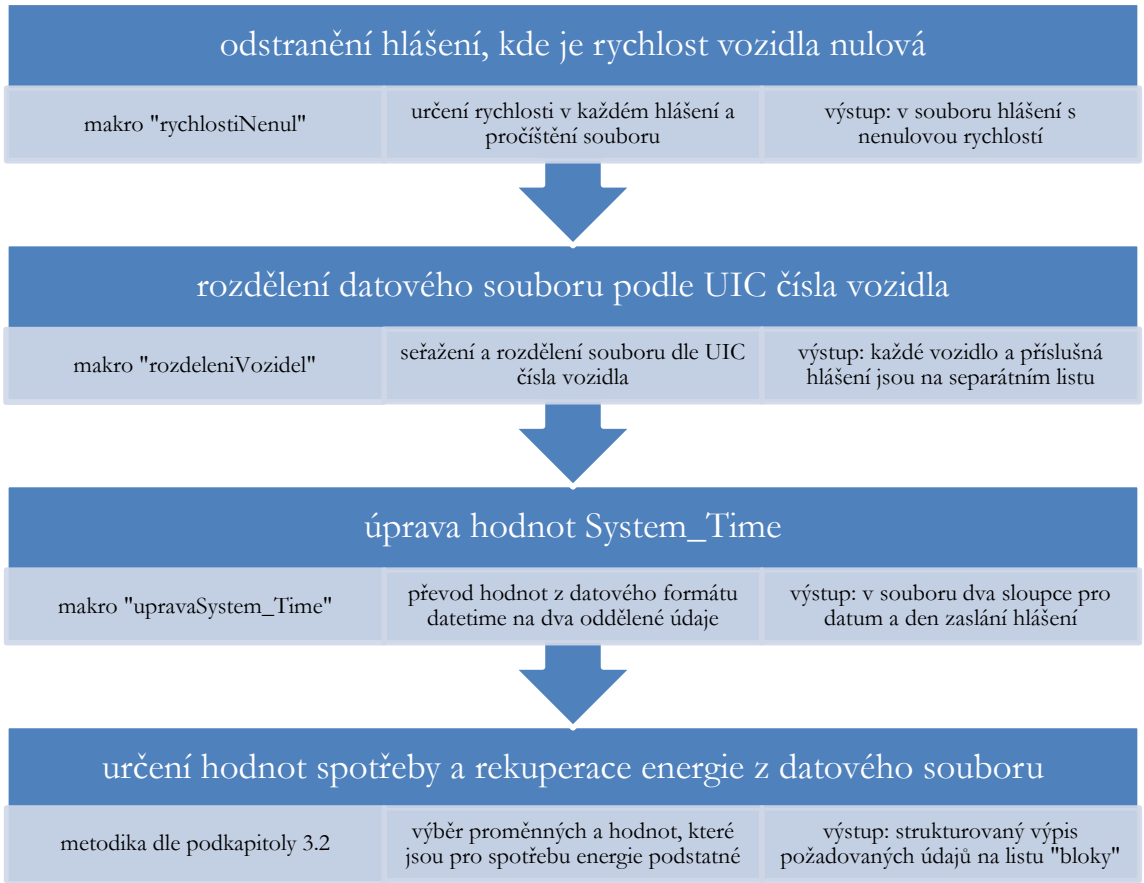
[14] Jako stanoviště se označuje kabina strojvedoucího

Obecným cílem preprocessingu dat v případě údajů spotřeby energie je postupnými kroky z jednotlivých hlášení procesních dat určit spotřebu a rekuperaci energie vozidla pro každou jednotlivou jízdu.

Jelikož vozidlo zasílá údaje do procesních dat, i když není v provozu na trati, **velká část hlášení je pro nás nepodstatná**. Jinak řečeno, je nutné ze všech hlášení vybrat ta, při jejichž zaslání je vozidlo v pohybu a hodnoty spotřeby energie se mění. Přitom je **frekvence zasílání jednotlivých hlášení** pro nás až **moc vysoká**, jelikož se během té doby sledované hodnoty buď „nestihnou změnit" vůbec anebo se přírůstky pohybují v řádech jednotek kilowatthodin.

Nutné pročištění dat bude tedy také součástí **předzpracování** zvoleného datového souboru. Avšak vzhledem k podobě zaslaných dat a s ní spojenými problémy (viz kapitola 2) rozdělíme preprocessing na dvě části:

i. Zpracování „hrubé podoby" souboru z databáze do podoby jednoduché tabulky,
ii. Na základě hodnot z tabulky rozdělení hlášení do bloků a určení spotřeby a rekuperace energie pro každou jízdu.



Obrázek 4.2: Schéma postupu při první fázi předzpracování

První část předzpracování obsahuje následující kroky:

i. Odebrání takových hlášení, ve kterých je rychlost vozidla udána jako nulová – makro „rychlostNenul“,
ii. Rozdělení datového souboru podle UIC čísla vozidla[15] - makro „rozdeleniVozidel“,
iii. Úprava hodnot System_Time, které jsou ve formátu datetime a budeme je chtít rozdělit na dvě samostatné hodnoty data a času hlášení,
iv. Vybrat z jednotlivých hlášení údaje o spotřebě energie,

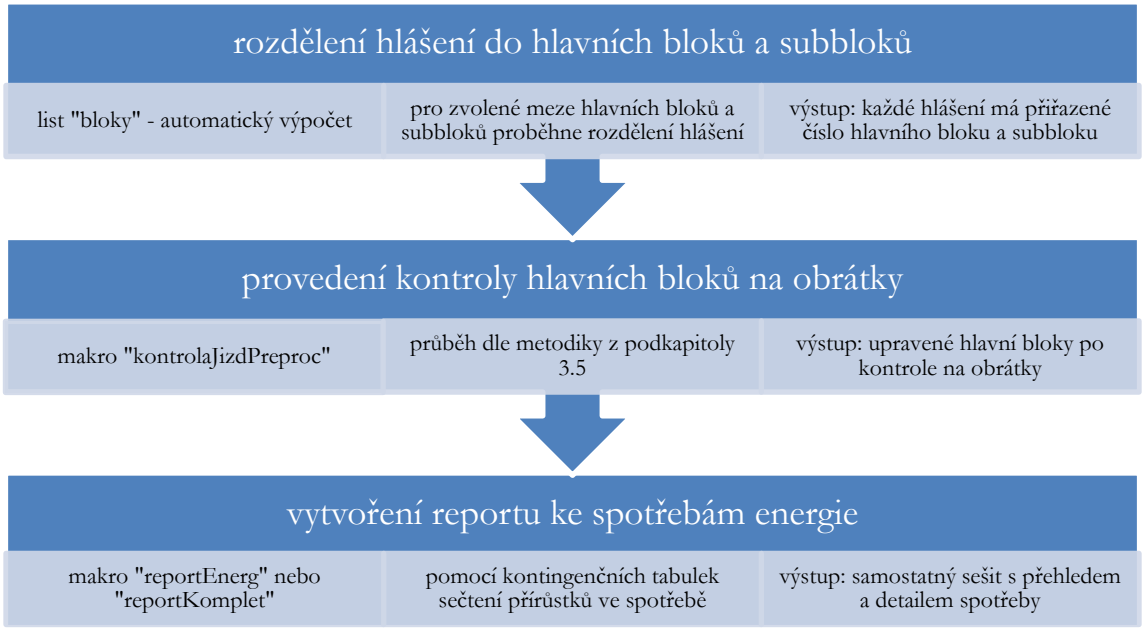
Implementaci jednotlivých kroků provedeme opět pomocí **softwaru Microsoft Excel a jazyku VBA**. V něm napsaná makra postupně zpracují soubor získaný z databáze ve formátu csv až do konečně formy v podobě přehledného reportu o dvou základních listech s akumulovanými hodnotami spotřeby energie.

Obrázek 4.2 znázorňuje postup a je návodem při první fázi předzpracování dat. U každého kroku je uveden název makra, stručný popis toho, co makro provádí a výstup z provedených úkonů.

Základním sešitem pro předzpracování dat ze souboru z databáze do finální podoby reportů je sešit „tvorbaReportu.xlsb“, který je nutné otevřít stejně jako soubor „PERSONAL_DP“, který obsahuje veškerá makra využitá v této práci.

Určení rychlosti a aktuálního stavu spotřeby energie z každého hlášení provedeme na základě postupu, který je uveden v části 3.2.

Tento postup využijeme konkrétně v prvním makru „rychlostiNenul“, které spustíme přímo v csv souboru z databáze. Ze všech hlášení se vyfiltrují ta, kde je rychlost vozidla větší než nula, a ta se převedou na list „korig“ v sešitu „tvorbaReportu.xlsb“.



Obrázek 4.3: Schéma postupu při druhé fázi předzpracování

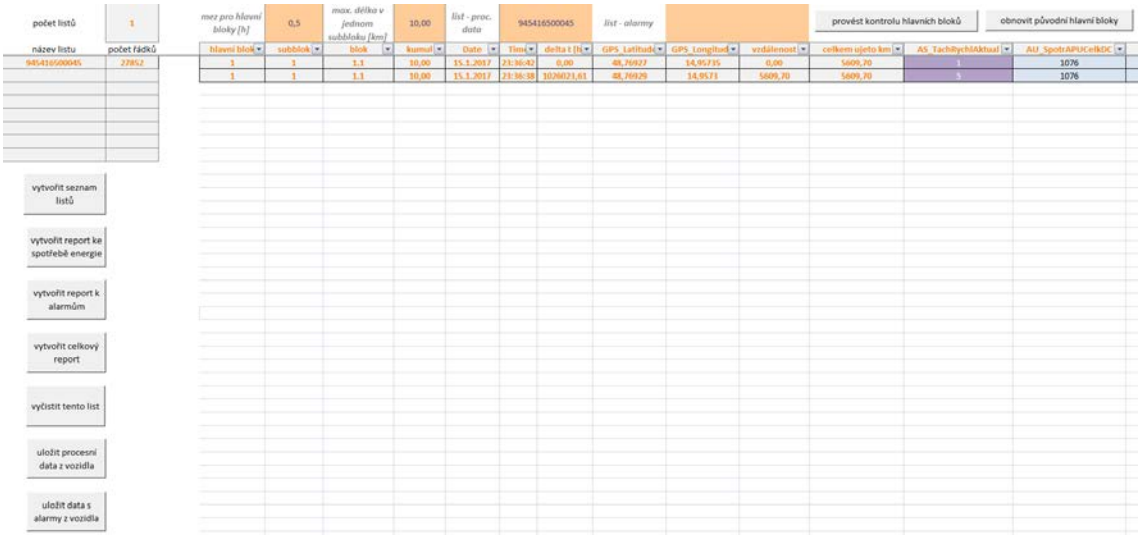
[15] V případě elektrických jednotek (vícevozidlové soupravy) můžeme v SQL dotazu specifikovat UIC číslo všech vozidel v soupravě. Takto získaný datový soubor je poté nutné rozdělit dle jednotlivých vozidel

Na tomto listu je automaticky spuštěno další makro „rozdeleniVozidel“, které nám ošetří případ, že v datovém souboru bude více vozidel. Pro každé z nich je vytvořen zvláštní list, který nese označení UIC číslo v něm obsaženého vozidla. Princip fungování tohoto makra je založen na jednoduchém seřazení souboru právě podle UIC čísla vozidla a postupném převedení jednotlivých takto vzniklých segmentů na separátní list. Na závěr jsou hlášení listu seřazena dle data a času zaslání sestupně, tj. „nejstarší hlášení“ budou uvedena na posledních řádcích souboru a hodnoty Systém_Time, které jsou zprvu v klasickém „datetime“-formátu, jsou rozštěpeny na dva separátní údaje data a času zaslání.

Tímto je ukončena první fáze předzpracování. V druhé fázi provedeme tyto kroky:

i. Rozdělení hlášení do bloků (viz podkapitola 3.4),
ii. Provedení kontroly a případné opravy hlavních bloků (viz podkapitola 3.5),
iii. Vytvoření reportu ke spotřebě energie na základě rozdělení hlášení v blocích.

Všechny tyto kroky se odehrávají na centrálním listu daného sešitu s názvem „bloky“, který představuje Obrázek 4.4. Shrnutí a návod pro druhou fázi předzpracování uvádí Obrázek 4.3.



Obrázek 4.4: Ukázka listu „bloky“ souboru „tvorbaReportu.xlsb“

Veškeré požadavky pro druhou část předzpracování může uživatel provádět na tomto listu pomocí ovládacích tlačítek, která spustí s nimi spojené makro. Ta si jednotlivě projdeme a vysvětlíme si princip, na kterém je založen jejich kód:

i. Vytvořit seznam listů – makro „seznam_listu“,
ii. Provést kontrolu hlavních bloků – makro „kontrolaJizdPreproc“,
iii. Obnovit původní hlavní bloky – makro „puvodStav“,
iv. Vytvořit report ke spotřebě energie – makro „reportEnerg“,
v. Vytvořit report k alarmům – makro „reportAlarm“,
vi. Vytvořit celkový report – makro „reportKomplet“,
vii. Vyčištění listu „bloky“ – makro „vycistit“,

viii. Uložit procesní data z vozidla – makro „ulozitProcData“,
ix. Uložit data s alarmy z vozidla – makro „ulozitAlarmy“.

Při tvorbě **seznamu listů** je procházen každý jednotlivý list před listem „bloky“, název a počet řádků je zapsán do tabulky v levém horním rohu listu „bloky“. Z důvodu přehlednosti a zajištění co nejefektivnějšího chodu všech maker je dovoleno v sešitu zachovat jen osm listů s daty.

Zároveň je mimo makra z výše uvedeného seznamu listem „bloky“ spojeno makro „Worksheet_Change“ a „prevodHodnot“. Ta jsou spojena se **změnou buňky** vpravo od buňky „list – proc. data“, kde si uživatel vybere po vytvoření seznamu listů ten **list s procesními daty**, na jehož základě budou vytvořeny další reporty. Pokud v této buňce dojde ke změně její hodnoty (například při vytvoření nového seznamu listů), pak je postupně spuštěno makro „prevodHodnot“, které přečte název listu a jeho počet řádků, se kterým má pracovat, a přes vzorce **převede z každého hlášení datum, čas, GPS souřadnice, aktuální rychlost a všechny hodnoty související se spotřebou a rekuperací energie.**

Souběžně s převedením hodnot se do každého řádku vloží i **vzorce**, které určí **pro každé jednotlivé hlášení příslušný hlavní blok i subblok**[16]. Časovou hranici pro určení hlavních bloků i maximální délku jednoho subbloku si může uživatel nastavit v prvním řádku tohoto listu. Pokud například celkově ujetá vzdálenost v hlavním bloku činí 200 km a námi volená délka úseku bude 10 km, budeme očekávat, že v rámci hlavního bloku nám vznikne okolo 20 subbloků.

Mimo již řečené jsou na základě postupu popsaného v kapitole 3.3 určeny **vzdálenosti mezi jednotlivými hlášeními a také celkový počet ujetých kilometrů** (sloupce „vzdálenost“ a „celkem ujeto km“). Jednotlivé vzdálenosti se využívají ve sloupci **„kumul“**, který tyto hodnoty kumuluje a odčítá od uživatelem volené meze ujetých kilometrů pro jeden subblok. Pro hlášení, ve kterém by tato kumulovaná hodnota byla menší než nula, to znamená zařazení do dalšího subbloku a hodnota „kumul“ se vrátí na úroveň zmíněné meze.

Po převodu hodnot z hlášení na list „bloky“ je doporučeno **provést kontrolu hlavních bloků** pomocí příslušného tlačítka a makra. Důvody, které nás k tomu vedou, jsme si představili v podkapitole 3.5.

Celé této problematice se věnuje **kód makra „kontrolaJizdPreproc“**. Po jeho spuštění je v sešitu „tvorbaReportu.xlsb“ vytvořen dočasný list „tempe“, na který se převedou hodnoty všech hlavních bloků a subbloků, tudíž bude mít dočasná tabulka nejprve stejný počet řádků, jako tabulka na listu „bloky“. Následně provedeme odstranění duplicit, jehož výsledkem bude, že každý blok[17] se bude vyskytovat právě jednou. Ke každému bloku přiřadíme dle posledního hlášení tomuto bloku odpovídající pár souřadnic. Makro poté prochází každý jednotlivý hlavní blok a jeho subbloky.

[16] Přiřazení čísla hlavního bloku a subbloku se provádí dle postupu v kapitole 3.3.

[17] Blok ve smyslu kombinace čísla hlavního bloku a subbloku

Následně se v makru zkoumá v podkapitole 3.5. uvedený vztah:

$$(d(SB_{t+3}, SB_{t-3}) \leq \varepsilon) \wedge (d(SB_{t+2}, SB_{t-2}) \leq \varepsilon), \quad (4.1)$$

Narazí-li makro na případ, kdy jsou splněny podmínky tohoto vztahu, pak upraví patřičné hodnoty hlavních bloků v dočasné tabulce a v celkovém přehledu všech hlášení na listu „bloky".

Pokud by se po provedené kontrole hlavních bloků chtěl uživatel vrátit k jejich původní podobě, tj. k rozdělení hlavních bloků před kontrolou, má též **možnost pomocí tlačítka a makra původní hlavní bloky obnovit**.

Dalším krokem je vytvoření reportu ke spotřebě energie. Po použití příslušného tlačítka je spuštěno makro, které postupně vytvoří nový soubor a v něm dva listy:

i. Přehled spotřeby v hlavních blocích,
ii. Detail spotřeby v každém subbloku.

V obou případech pracuje makro na bázi kontingenčních tabulek, které pro přehled spotřeby sečtou spotřeby a rekuperace energie stejně jako ujetou vzdálenost v hlavních blocích, pro detail spotřeby naopak v subblocích.

Pro přehled spotřeby se navíc v kontingenční tabulce vyberou datum a čas prvního a posledního hlášení v každém hlavním bloku, ze kterých je určena doba provozu vozidla v hodinách. Navíc se pro každý blok uvádí počet jeho subbloků. Na listu přehled spotřeby je tedy po vytvoření reportu **pro každý hlavní blok** mimo spotřeby a rekuperace energie uveden **počet jeho subbloků, součet ujetých kilometrů, doba provozu, čas i den začátku a konce provozu**.

V případě detailu spotřeby určí makro souřadnice posledního hlášení v každém subbloku, stav ujetých kilometrů a průměrnou rychlost v daném subbloku. Všechny tyto informace jsou uvedeny na listu „detail spotřeba" ke každému subbloku mimo společné aspekty zmíněné v předchozích odstavcích.

Po vytvoření zmíněných dvou listů přehledu a detailu spotřeby je do sešitu s reportem převeden též list „mainStations", s jehož pomocí se později určí zastávky při určení podobnosti jízd a klasifikace do tratí. Navíc je vytvořen list „výběr bloků", ve kterém uživatel má možnost nastavit základní parametry pro výpočty podobnosti (detailní postup viz kapitola 3.6).

Postup při vytvoření reportu k alarmům rozebereme v kapitole 5.1 v rámci preprocessingu dat. Vytvoření celkového reportu je poté pouze spojením reportu o spotřebě energie a reportu k alarmům.

Zvolí-li uživatel tlačítko **„vyčistit tento list"**, smaže s ním spojené makro všechna hlášení z listu „bloky", který je poté připraven pro zobrazení hlášení jiného vozidla, respektive listu. Toto pročištění listu je prováděno automaticky, pokud zvolíme ze seznamu vpravo od buňky „list – proc. data" jiné vozidlo. Změna hodnoty této buňky spustí makro pro převod hodnot ze zvoleného listu (viz odstavec k makrům „Worksheet_Change" a „prevodHodnot" výše).

Poslední dvě tlačítka k **uložení procesních dat či alarmu z vozidla** spustí makra, která dle obsahu buňky „list – proc. data“, respektive „list – alarmy“ odeberou patřičný list ze sešitu „tvorbaReportu.xlsb“ a uloží ho do nového souboru. V případě procesních dat nese název jména odebraného listu s koncovkou „…_data.xlsb“, v případě alarmů je koncovka „…_alarm_data.xlsb“.

4.3 Statistická analýza spotřeby energie

4.3.1 Teoretický základ

Cílem je porovnat spotřeby různých skupin a z nich vyvodit závěr o tom, jestli panuje obecná shodnost nebo zda naopak jedna ze skupin se jeví jako odlišná.

Námi stanovený cíl nás z hlediska statistiky vede do oblasti postupů pro jednoduché třídění, jehož klasickým představitelem je ANOVA. Jelikož ale na jedné straně u jednotlivých výběrů[18] **nemusí být splněn předpoklad o shodě rozptylů** a na straně druhé nás bude především zajímat případ, kdy se **jednotlivá rozdělení liší ve své střední hodnotě**[19], přejdeme na neparametrickou alternativu postupu ANOVA, kterou je **Kruskal-Wallisův test** (5).

Ten testuje hypotézu H_0 o tom, že *„všechna rozdělení jsou stejná“*, oproti alternativě, že *„alespoň jedna skupina má jiné rozdělení“*. Metodiku jsme převzali z (6). Předpokladem testu je, aby data byla realizací **k nezávislých náhodných výběrů**, což je v našem případě splněno vzhledem k charakteru dat. Dále musí mít **rozdělení pozorování spojitou distribuční funkci**, což splňujeme skutečností, že budeme využívat průměrné spotřeby na jeden kilometr (viz níže).

Testová statistika bude mít následující tvar:

$$KW = \frac{12}{n(n+1)} \sum_{j=1}^{k} n_j \left(\bar{R}_j - \frac{n+1}{2} \right)^2, \tag{4.2}$$

kde

n_j je počet dat v j-té skupině,

n je celkový počet dat ze všech výběrů a tedy platí $\sum_j n_j$,

$\bar{R}_j$ je průměrné pořadí dat v j-té skupině.

Při platnosti H_0 o shodě všech k rozdělení má tato statistika asymptoticky χ^2 rozdělení s $v = k - 1$ stupni volnosti (zkráceně budeme značit $\chi^2(v)$). Pokud je hodnota testového kritéria KW větší než $(1-\alpha)\%$ kvantil rozdělení $\chi^2(v)$, zamítáme H_0 na přibližné hladině významnosti α. P-hodnotu testu počítáme jako $1 - F_v(KW)$, kde $F_v(\chi^2)$ je distribuční funkce rozdělení $\chi^2(v)$.

[18] V našem případě je pojem vozidlo a skupina totožný. Jako výběr chápeme všechny jízdy jednoho vozidla.

[19] Lze také říct, že se liší posunem.

Pokud výsledkem testu bude to, že **nezamítáme nulovou hypotézu**, pak testování končíme výsledkem, že všechna rozdělení jsou stejná. V případě **zamítnutí H_0 se budeme ptát, které rozdělení tuto neshodu způsobilo**.

K tomu využijeme základní vlastnosti KW statistiky. Ta zkoumá, jak mnoho se průměrná pořadí liší od celkového průměru pořadí $\frac{n+1}{2}$. Toto vyjadřují hodnoty sčítanců KW-statistiky $n_j\left(\bar{R}_j - \frac{n+1}{2}\right)^2$. **Skupinu j, pro niž bude mít sčítanec ve srovnání s ostatními největší hodnotu**, vyhodnotíme tak, že jeví **největší známky odlišnosti** od ostatních skupin.

Pro takové vozidlo budeme hledat **extrémní hodnoty**, tj. hodnoty výrazně odlišné od ostatních, abychom zjistili, zda tyto hodnoty mají vliv na výsledek testu. Vyhodnocení provedeme pomocí **boxplotů**, **kvantilů příslušného modelového normálního rozdělení**, a **Dixonova testu**, který jsme převzali z (7).

Boxploty využijeme k vizualizaci variability dat. Každý boxplot bude zobrazovat horní kvartil q_3 a dolní kvartil q_1 (horní a dolní hranice samotného „boxu“) a medián (horizontála uvnitř „boxu“) stejně jako vousy. Ty zobrazují největší, respektive nejmenší hodnotu, která je menší než $q_3 + 1{,}5(q_3 - q_1)$, respektive větší než $q_1 - 1{,}5(q_3 - q_1)$. Hodnoty, které jsou větší než horní vous nebo menší než dolní vous, budou vyhodnoceny jako potencionálně extrémní[20].

Dále pro každý výběr vypočteme průměr a výběrovou směrodatnou odchylku a pomocí těchto dvou charakteristik určíme **modelové normální rozdělení**. Hodnoty, které budou menší než 5% nebo větší než 95% kvantil, tak opět vyhodnotíme jako potencionálně extrémní.

Dixonovým testem poté hodnoty podezřelé jako extrémní v tomto smyslu otestujeme. Předpokladem tohoto testu je, že data splňují podmínky náhodného výběru a bez zkoumané extrémní hodnoty pochází z normálního rozdělení. První podmínku splňujeme charakterem dat a druhou vždy ověříme po odstranění extrémních hodnot dvěma testy normality – Jarque-Bera a Lilliefors testem, jejichž předpokladem pro pozdější testování normality je, že data pochází z náhodného výběru spojité veličiny (postupy pro testy uvádí zdroj (8) a (9)). Pro seřazený soubor $x_{(1)}, x_{(2)}, \dots, x_{(n_i-1)}, x_{(n_i)}$ mají testové statistiky tvar:

$$\begin{aligned} TN_u &= \frac{x_{(n_i)} - x_{(n_i-2)}}{x_{(n_i)} - x_{(3)}}, \text{ pokud testujeme hodnotu } x_{(n_i)}, \\ TN_l &= \frac{x_{(3)} - x_{(1)}}{x_{(n_i-2)} - x_{(1)}}, \text{ pokud testujeme hodnotu } x_{(1)}. \end{aligned} \tag{4.3}$$

Kritické hodnoty jsou uvedeny ve speciálních tabulkách, které uvádí zdroj (10). **Překročí-li hodnota testové statistiky** příslušnou kritickou hodnotu, zamítáme **nulovou hypotézu** o tom, že **zkoumaná hodnota se od ostatních hodnot v souboru významně neliší**, a tuto hodnotu ze souboru vyjmeme.

[20] Takové hodnoty se také označují jako odlehlá pozorování

Bude-li ve **výběru několik hodnot**, které budou podezřelé z toho, že jsou extrémní, Dixonův test **použijeme postupně od nejméně k nejvíc odlehlé hodnotě**. Budeme-li pro nějakou ze zkoumaných hodnot zamítat nulovou hypotézu, **zároveň s ní vyloučíme hodnoty**, které jsou od zbytku souboru ještě vzdálenější než zkoumaná hodnota.

Pokud po odstranění extrémních hodnot zůstane výsledek testu nezměněn, vozidlo z dalšího testování vyloučíme. Celý Kruskal-Wallisův (KW) test provedeme znovu na ostatních skupinách a v případě zamítnutí nulové hypotézy o shodě zkoumaných rozdělení opět vyloučíme skupinu s největší hodnotou zmíněného sčítance. Postup budeme opakovat až do chvíle, kdy nulovou hypotézu KW-testu již zamítat nebudeme.

4.3.2 Praktické provedení

Celý výše uvedený test byl v rámci této práce naprogramován též v jazyce VBA tak, abychom mohli v rámci vyhodnocení používat stejný software, ve kterém postupně tvoříme reporty. Konkrétně test je součástí makra „KruskalWallis“. Soubor, ve kterém se **analýzy příslušné podkapitoly** nachází, bude uveden vždy **na začátku podkapitoly**. Zároveň vždy platí, že na listu „Dixon“ příslušného souboru se nachází výpočty k určení extrémních hodnot (přístup pomocí kvantilů normálního rozdělení a Dixonův test). U každého vozidla jsou na tomto listu i vypočtené hodnoty příslušné testové statistiky a opět červeně jsou označené ty, které překračují příslušnou kritickou hodnotu z tabulky zdroje (10) pro hladinu významnosti $\alpha = 10\ \%$.

Pokud v některém výběru Dixonův test indikuje extrémní hodnotu, následně **otestujeme výběr bez této hodnoty**, případně hodnot, na normalitu (viz výše Jarque-Bera a Lilliefors test). Tyto testy byly provedeny pomocí softwaru MATLAB, konkrétně v souboru „DP_normalita.m“. Hladina významnosti pro každý ze dvou testů byla volena jako $\alpha = 5\ \%$, jelikož **se zde setkáváme s problémem násobného testování hypotéz** (více viz kapitola 5.2.1). V souborech s výpočty pro KW-test jsou také na listu „Dixon“ uvedeny pouze testové statistiky a kritické hodnoty k zmíněným dvěma testům. Jejich podobu znázorňuje na konkrétním příkladu Tabulka 4.1.

testy normality po odebrání extrémů	
Lilliefors	
test. stat	0,1287
krit. hodnota	0,1614
Lilliefors OK	
Jarque-Bera	
test. stat	2,9436
krit. hodnota	4,3588
Jarque-Bera OK	

Tabulka 4.1: Podoba tabulek s informacemi z testů normality

Pro každý list, kde je proveden KW-test, jsou pro nás důležité první tři sloupce, kam vložíme vstupní data, na základě kterých chceme test provést. **V prvním sloupci** uvedeme **vysvětlující proměnnou**, což odpovídá označení skupiny – v našem případě to bude vždy číslo vozidla.

Do **druhého a třetího sloupce** vložíme **spotřeby sledovaných skupin** na střídavé a stejnosměrné síti, pokud se vozidla pohybovalo na trati s oběma systémy. V případě, kdy se vozidlo pohybovalo na síti pouze jednoho typu[21], tak stačí vložit data pouze do jednoho sloupce. Makro je totiž koncipováno tak, že rozpozná, zda uživatel vložil data pro oba typy sítí nebo jen pro jeden, a ve výstupu to také zohlední.

Data jsou postupně zpracována tak, že je **KW-test proveden pro všechny sady dat**, takže pokud máme data pro oba typy tratí, tak se vyhodnotí na sobě nezávisle spotřeby na trati DC a AC a následně i součet těchto spotřeb – ve výsledku tedy KW-test provedeme třikrát.

Analýzu shodnosti spotřeb energie provedeme exemplárně na **čelních vozech dálkových jednotek** – konkrétně budeme zkoumat trať Brno-Olomouc v jejích jednotlivých směrech a typech trakčního systému.

Pro výpočty jsme vybrali ta vozidla, která měla v obou směrech **za sledované období alespoň 20 nebo více jízd**, abychom dokázali co nejlépe eliminovat krátkodobé výkyvy související například s počasím, obsazeností vlaku apod. Testovat budeme na hladině významnosti $\alpha = 10\ \%$.

Pro přehlednost bude **každá podkapitola** věnovaná **jednomu konkrétnímu směru jízdy jednoho typu trakce**. Směry mezi sebou srovnávat nebudeme, jelikož profil tratě může mít na hodnoty spotřeb velký vliv a tedy pokud trať není víceméně rovinatá, tak budou spotřeby v jednotlivých směrech z principu odlišné i bez toho, aby na tento rozdíl mělo vliv vozidlo samotné.

Dále platí, že zkoumaná vozidla jsou z technického a konstrukčního hlediska stejná, a tedy má smysl zkoumat, zda je rozdělení spotřeb u všech vozidel stejné. Porovnávat budeme **průměrné spotřeby na jeden kilometr**, tj. hodnoty spotřeb trakce vydělené počtem ujetých kilometrů při dané jízdě. Takto spotřebu i-tého vozidla budeme chápat jako realizaci náhodné veličiny X_i (například spotřeba na jeden kilometr na síti DC).

4.3.2.1 Trať Brno-Olomouc ve směru tam – síť DC

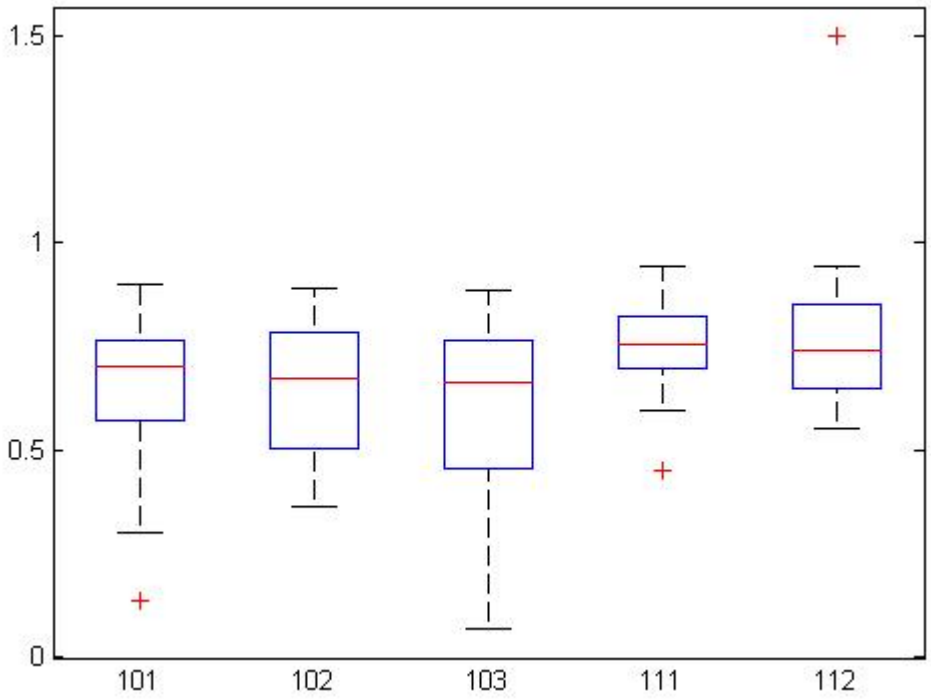
Všechny výpočty k této části se nachází v souboru „Energie_Brno_Olomouc_DCtam.xlsx“

vozidlo	průměry	směrodatné odchylky	počet jízd
101	0,65	0,18	31
102	0,64	0,16	40
103	0,58	0,24	37
112	0,77	0,19	23
111	0,76	0,11	28

Tabulka 4.2: Základní charakteristiky – Brno-Olomouc tam – síť DC

[21] Typ ve smyslu AC/DC

Naše požadavky[22] na této trati splnilo pro sledované období pět vozidel, přičemž platí, že vozidla s čísly 111 a 112 jsou čelní vozidla třívozové soupravy, zatímco zmíněná tři další vozidla (jmenovitě 101, 102 a 103) jsou čelními vozy pětivozové soupravy.



Obrázek 4.5: Boxplot spotřeby DC trať Brno-Olomouc – směr tam

Boxploty k původním výběrům včetně všech hodnot uvádí Obrázek 4.5. Výsledky Kruskal-Wallisova testu pro všechna vozidla uvádí Tabulka 4.3 – v souboru s výpočty se nachází na listu „DC - vš. vozidla".

Jak je evidentní podle toho, co uvádí Tabulka 4.3, tak zamítáme na hladině významnosti $\alpha = 10\,\%$ nulovou hypotézu H_0 o tom, že průměrné spotřeby všech vozidel na síti DC jsou stejně rozdělené. Vozidlo s číslem 111 má ve výpočtu KW-statistiky sčítanec s největší hodnotou (viz list „DC - všechna vozidla"), ale než ho vyloučíme, budeme hledat extrémní hodnoty. Jak již napovídá boxplot (Obrázek 4.5) a nakonec i Dixonův test, v tomto případě dojde k vyloučení nejmenší hodnoty, jak ale ukazuje Tabulka 4.3 v její pravé části, tak nedochází ke kvalitativní změně výsledku testu a vozidlo 111 tedy vyloučíme.

	všechny hodnoty	**bez extrémn. hodnot**
KW-statistika	17,5403	19,1902
krit. hodnota	7,7794	
p-hodnota	0,0015	0,0007

Tabulka 4.3: Výsledky KW-testu – srovnání všech vozidel trať Brno-Olomouc – směr tam síť DC

[22] Počet jízd v jednom směru větší než 20

Tabulka 4.4 představuje výsledky KW-testu pro zbylá čtyři vozidla. KW-statistika opět spadá do kritického oboru a zamítáme H_0 o shodě rozdělení zkoumaných vozidel. Sčítanec s největší hodnotou zde má vozidlo 112. I v tomto případě již na základě boxplotu nalezneme jednu extrémní hodnotu, po jejímž odstranění testovou statistiku přepočítáme. Ani zde však extrémní hodnota nemá vliv na kvalitu výsledku a opět H_0 zamítáme - vozidlo 112 ze souboru odstraníme.

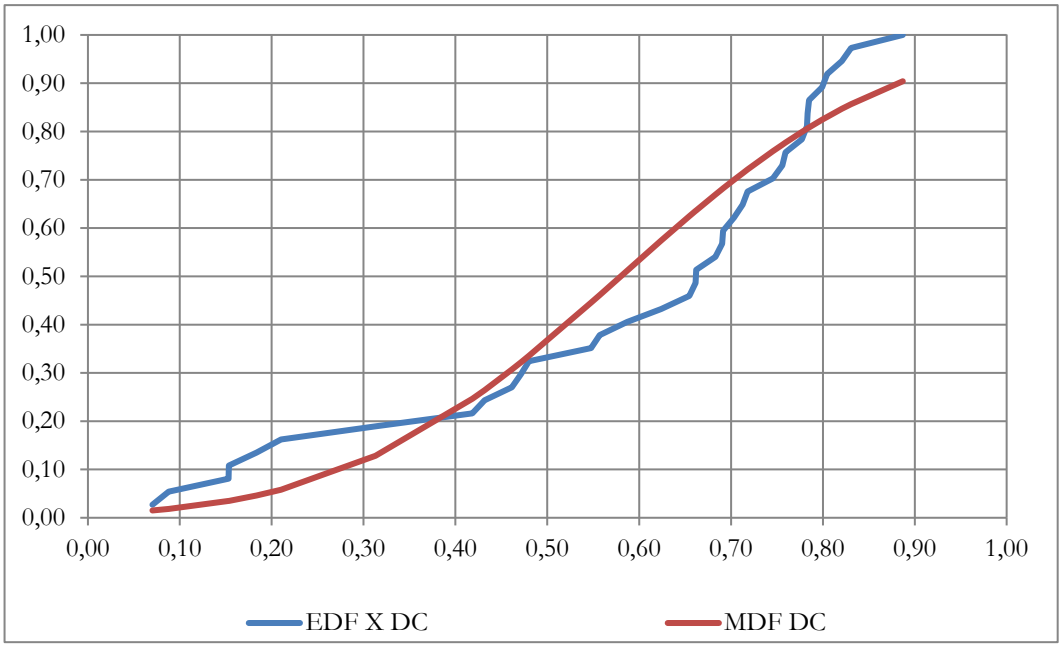
	všechny hodnoty	**bez extrémn. hodnot**
KW-statistika	10,0675	8,6345
krit. hodnota	6,2514	
p-hodnota	0,0180	0,0346

Tabulka 4.4: Výsledky KW-testu – vozidla 101,102,103, 112 trať Brno-Olomouc – směr tam síť DC

Pro zbylá tři vozidla již na dané hladině významnosti nezamítáme H_0 o shodě rozdělení průměrných spotřeb na jeden kilometr. Zároveň jsme otestovali shodu dvou vyřazených vozidel 111 a 112 a ani zde H_0 o shodě rozdělení nezamítáme. Konkrétní hodnoty uvádí Tabulka 4.5 a výpočty se nacházejí na listech „DC – 3 vozidla“ a „DC – 2 odstr. vozidla“.

	101, 102, 103	**111,112**
KW-statistika	1,5155	0,5531
krit. hodnota	4,6052	2,7055
p-hodnota	0,4687	0,4570

Tabulka 4.5: Výsledky KW-testu – hodnocení dvou skupin vozidel trať Brno-Olomouc – směr tam síť DC



Obrázek 4.6: Empirická a modelová distribuční funkce pro výběr vozidla 103

Podobně jako jsme zkoumali extrémní hodnoty u těch vozidel, která jsme později ze zkoumání vyloučili, tak tyto hodnoty také budeme hledat u trojice vozidel 101, 102, 103. Cílem je zjistit, zda odstraněním extrémních hodnot bude shoda rozdělení průměrných spotřeb porušena, tj. jestli dojde ke kvalitativní změně výsledku.

Z podoby modelových distribučních funkcí normálního rozdělení a podoby boxplotů soudíme, že jak výběr dat pro vozidlo 101, tak i pro vozidlo 103 obsahuje několik velmi nízkých hodnot, které otestujeme Dixonovým testem.

Připomeňme, že pro testové statistiky, které překračují příslušnou kritickou hodnotu, zamítáme nulovou hypotézu o tom, že zkoumaná hodnota se od ostatních hodnot v souboru významně neliší, a přijmeme alternativní hypotézu, že se významně liší.

Podle Dixonova testu však pouze výběr z vozidla 103 obsahuje extrémní hodnoty – zde odstraníme souboru vozidla 103 šest nejnižších hodnot. Obrázek 4.6 uvádí pro toto vozidlo empirickou (EDF) a modelovou distribuční funkci (MDF), ze kterých je patrná vysoká četnost výskytu velmi nízkých hodnot v tomto výběru.

Pro datové soubory vozidel 101, 102, 103, ze kterých jsme odstranili extrémní hodnoty, provedeme znovu Kruskal-Wallisův test. Jak ale představuje Tabulka 4.6, tak se výsledek testu oproti případu, kdy jsme zahrnuli všechna pozorování, nezměnil.

	bez extrémn. hodnot
KW-statistika	1,8176
krit. hodnota	4,6052
p-hodnota	0,4030

Tabulka 4.6: Výsledky KW-testu – vozidla 101, 102, 103 bez extrémních hodnot – směr tam síť DC

Závěrem této části tedy je, že spotřeba všech pěti vozidel na síti DC ve směru „tam" není stejně rozdělená. Při dalším zkoumání jsme však došli k závěru, že stejně rozdělená bude spotřeba čelních vozů pětivozové soupravy (vozidla 101, 102, 103) a spotřeba čelních vozů třívozové soupravy (vozidla 111 a 112).

4.3.2.2 Trať Brno-Olomouc ve směru tam – síť AC

vozidlo	**průměry**	**směrodatné odchylky**	**počet jízd**
101	0,91	0,21	31
102	0,87	0,16	40
103	0,79	0,23	37
112	0,80	0,17	23
111	0,94	0,37	28

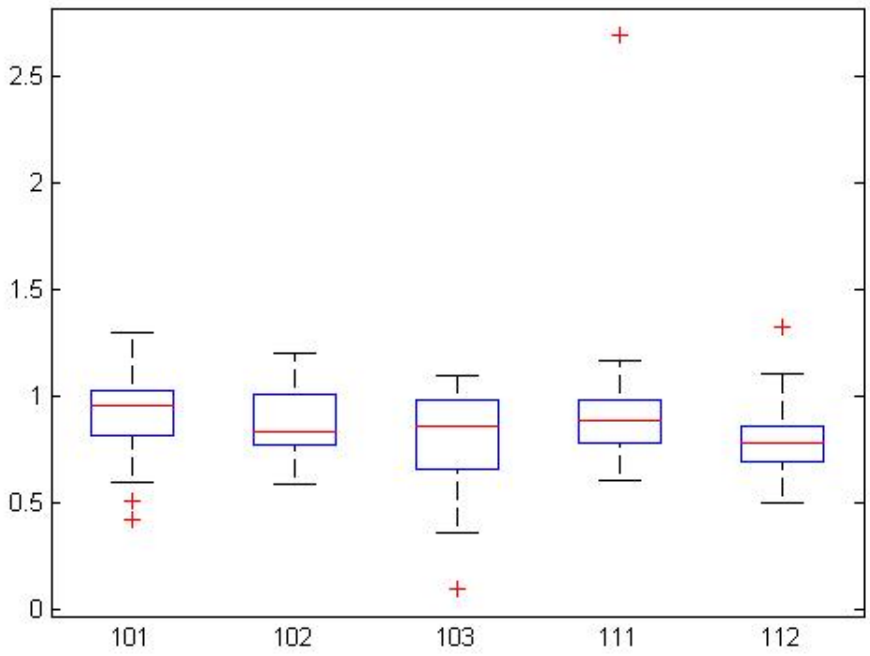
Tabulka 4.7: Základní charakteristiky – Brno-Olomouc tam – síť AC

Všechny výpočty k této části se nachází v souboru „Energie_Brno_Olomouc_ACtam.xlsx". Boxploty k původním výběrům včetně všech hodnot uvádí Obrázek 4.7. Výsledky Kruskal-Wallisova testu pro všechna vozidla uvádí Tabulka 4.8.

	všechny hodnoty	bez extrémn. hodnot
KW-statistika	8,8025	11,0966
krit. hodnota	7,7794	
p-hodnota	0,0662	0,0255

Tabulka 4.8: Výsledky KW-testu – srovnání všech vozidel trať Brno-Olomouc – směr tam síť AC

Dle výsledků, které uvádí Tabulka 4.8 (v souboru s výpočty list „AC – vš. vozidla"), i na síti AC ve stejném směru zamítáme nulovou hypotézu o shodě rozdělení všech vozidel. Ze všech vozidel má největší hodnotu sčítance vozidlo 112 a v jeho výběru tedy budeme hledat extrémní hodnoty. Zatímco boxplot označil pouze největší hodnotu z výběru jako extrémní, postup přes MDF a Dixonův test však mimo ní dále označí i druhou největší a nejmenší hodnotu jako extrémní. Po ověření normality upraveného výběru tyto tři hodnoty odebereme. Jak ale představuje Tabulka 4.8 ve své pravé části, výsledek se touto úpravou kvalitativně nezmění a vozidlo 112 z dalšího testování vyloučíme. Detailně viz list „AC – vš. vozidla (2)".



Obrázek 4.7: Boxplot spotřeby AC trať Brno-Olomouc – směr tam

Na zbývajících vozidlech provedeme znovu KW-test. Jak plyne z výsledků z listu „AC – 4 vozidla", které znázorňuje Tabulka 4.9, nulovou hypotézu o shodě rozdělení již nezamítáme a pro soubory včetně extrémních hodnot platí, že průměrná spotřeba na kilometr testovaných vozidel je stejně rozdělená. Prověříme ale, zda zmíněné extrémní hodnoty mají na kvalitu výsledku vliv.

	všechny hodnoty	bez extrémn. hodnot
KW-statistika	4,5897	3,7837
krit. hodnota	6,2514	
p-hodnota	0,2044	0,2858

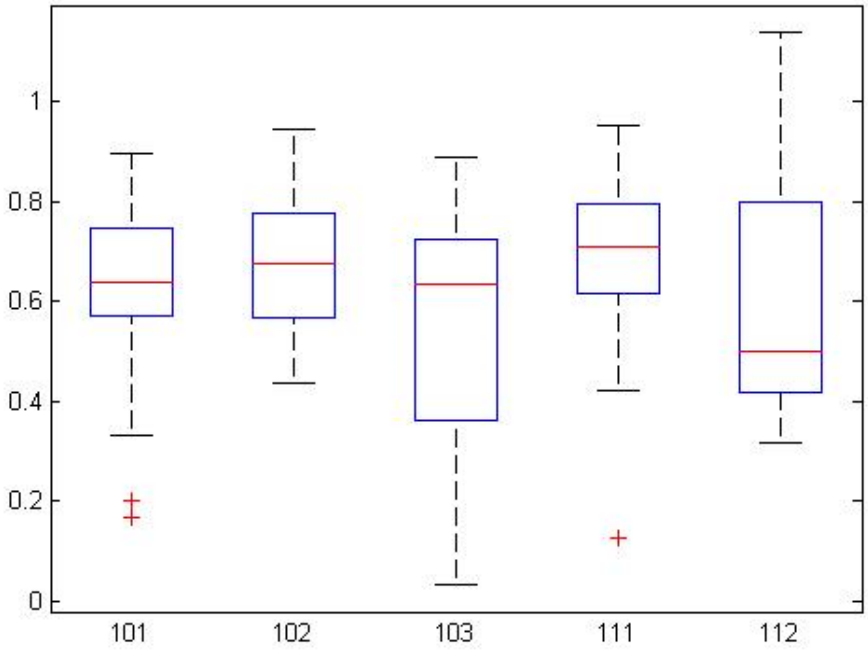
Tabulka 4.9: Výsledky KW-testu – vozidla 101, 102, 103, 111 včetně extrémních hodnot – směr tam síť AC

Podle Dixonova testu ze souboru vyloučíme ty hodnoty, které již boxplot označil za odlehlá pozorování, tj. nejmenší hodnotu výběru vozidla 103 a největší z výběru vozidla 111. Jejich vyloučení však nemá na kvalitu výsledku vliv (viz Tabulka 4.9 a list „AC – 4 vozidla bez extr.“).

Závěrem této části tedy je, že ani na síti AC není spotřeba všech pěti vozidel ve směru „tam“ stejně rozdělená. Avšak zde pouze spotřeba vozidla 112 se jevila jako jinak rozdělená.

4.3.2.3 Trať Brno-Olomouc ve směru zpět – síť DC

Všechny výpočty k této části se nachází v souboru „Energie_Brno_Olomouc_DCzpet.xlsx“.



Obrázek 4.8: Boxplot spotřeby DC trať Brno-Olomouc – směr zpět

Boxploty k původním výběrům včetně všech hodnot uvádí Obrázek 4.8. Výsledky Kruskal-Wallisova testu pro všechna vozidla uvádí Tabulka 4.11 (v souboru s výpočty viz list „DC – vš. vozidla“).

vozidlo	průměry	směrodatné odchylky	počet jízd
101	0,64	0,15	43
102	0,68	0,14	41
103	0,55	0,25	35
112	0,59	0,24	23
111	0,68	0,17	28

Tabulka 4.10: Základní charakteristiky – Brno-Olomouc zpět – síť DC

Jak ve směru „tam", tak i ve směru „zpět" zamítáme H_0 o shodě rozdělení spotřeb všech vozidel a budeme se ptát, které vozidlo toto způsobuje. Největší sčítanec v KW-statistice má v tomto případě vozidlo 103, v jehož výběru budeme hledat extrémní hodnoty. Na boxplotu k tomuto vozidlo je nápadné, že dolní vous je poměrně dlouhý, což svědčí o zvýšeném výskytu velmi malých hodnot.

Totéž potvrzuje i Dixonův test, podle kterého vyloučíme devět nejmenších hodnot. Poté provedeme s upraveným výběrem KW-test znovu – výsledky uvádí Tabulka 4.11 ve své pravé části a list „DC - vš. vozidla (103 bez ext.)" v souboru s výpočty.

	všechny hodnoty	bez extrémn. hodnot
KW-statistika	8,2810	5,2712
krit. hodnota	7,7794	
p-hodnota	0,0818	0,2606

Tabulka 4.11: Výsledky KW-testu – srovnání všech vozidel trať Brno-Olomouc – směr zpět síť DC

Z ní je však patrné, že po úpravě zmíněného výběru dochází ke kvalitativní změně výsledku – KW-test nyní již označil spotřeby všech vozidel jako stejně rozdělené.

I po odebrání extrémních hodnot z daných výběrů se výsledek KW-testu z předchozího kroku nemění – přitom nehraje roli, zda napřed odebereme extrémní hodnoty z výběru vozidla 101 nebo vozidla 111. Výsledkem vždy zůstává, že H_0 o shodě rozdělení spotřeb všech pěti vozidel nezamítáme. Konkrétní hodnoty testové statistiky i kritické hodnoty představuje Tabulka 4.12 a výpočty se nachází na listu „DC - vš. vozidla bez ext".

	bez extrémn. hodnot
KW-statistika	5,8779
krit. hodnota	7,7794
p-hodnota	0,2085

Tabulka 4.12: Výsledky KW-testu – vozidla 101, 102, 103 bez extrémních hodnot – směr zpět síť DC

Budeme se tedy ptát, zda případné extrémní hodnoty v dalších výběrech nemohou opět způsobit kvalitativní změnu výsledku. Jak lze pozorovat na boxplotu, tak výběry z vozidel 101 a 111 obsahují velmi nízké hodnoty, které jsou výrazně vzdálené od zbytku patřičného výběru. Tuto potvrzuje i Dixonův test, podle kterého ve výběru vozidla 101 vyloučíme tři nejnižší a u vozidla 111 nejnižší hodnotu.

Závěrem je, že pro síť DC ve směru zpět má 5 sledovaných vozidel stejnou rozdělenou spotřebu.

4.3.2.4 Trať Brno-Olomouc ve směru zpět – síť AC

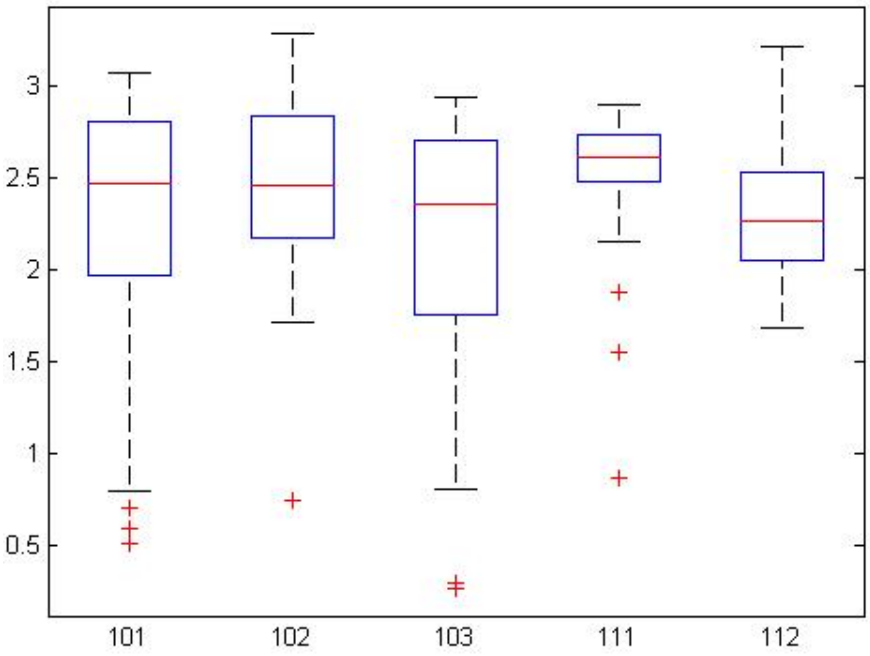
Všechny výpočty k této části se nachází v souboru „Energie_Brno_Olomouc_ACzpet.xlsx“.

Boxploty k původním výběrům včetně všech hodnot uvádí Obrázek 4.9. Ty naznačují, že v daných výběrech se potkáme s větším výskytem extrémních hodnot. O to důležitější bude zkoumání vlivu těchto hodnot na celkový výsledek.

vozidlo	průměry	směrodatné odchylky	počet jízd
101	2,22	0,74	43
102	2,47	0,49	41
103	2,13	0,77	35
112	2,31	0,34	23
111	2,50	0,44	28

Tabulka 4.13: Základní charakteristiky – Brno-Olomouc zpět – síť AC

Výsledky Kruskal-Wallisova testu pro všechna vozidla uvádí Tabulka 4.14 (v souboru s výpočty viz list „AC – vš. vozidla“).



Obrázek 4.9: Boxplot spotřeby AC trať Brno-Olomouc – směr zpět

	Všechny hodnoty
KW-statistika	7,4008
krit. hodnota	7,7794
p-hodnota	0,1162

Tabulka 4.14: Výsledky KW-testu – srovnání všech vozidel trať Brno-Olomouc – směr zpět síť AC

Jak plyne z výsledků, které představuje Tabulka 4.14, tak pro výběry včetně extrémních hodnot nezamítáme H_0 o shodě rozdělení spotřeby všech vozidel. Avšak vzhledem k vysokému počtu odlehlých pozorování očistíme všechny výběry o tyto extrémní hodnoty. Poté provedeme KW-test znovu, abychom zjistili, zda dojde ke změně jeho výsledku.

Pomocí Dixonova testu jsme v tomto případě našli v každém z výběrů alespoň jedno odlehlé pozorování, většinou i více. Pro přehlednost uvádí Tabulka 4.15 základní charakteristiky výběrů po odstranění extrémních hodnot. Již na těchto údajích je znatelný vliv, který odebrané hodnoty na „polohu" celého výběru měly.

vozidlo	průměry	směrodatné odchylky	počet jízd
101	2,50	0,38	36
102	2,51	0,41	40
103	2,43	0,40	29
112	2,27	0,29	22
111	2,65	0,15	24

Tabulka 4.15: Základní charakteristiky po odstranění extrémních hodnot Brno-Olomouc zpět – síť AC

V souboru s výpočty je na listech „AC – vš. vozidla (2)" - „AC – vš. vozidla (6)" uveden KW-test vždy po úpravě jednoho z výběrů, přičemž na posledním z nich je KW-test s výběry, které již všechny byly očištěny o odlehlá pozorování. Výsledky znázorňuje Tabulka 4.16.

	bez extrémn. hodnot
KW-statistika	14,4760
krit. hodnota	7,7794
p-hodnota	0,0059

Tabulka 4.16: Výsledky KW-testu – srovnání všech vozidel trať Brno-Olomouc bez extrémních hodnot – směr zpět síť AC

Je evidentní, že extrémní hodnoty měly na výsledek testu zásadní vliv. Nyní zamítáme H_0 o shodě rozdělení spotřeb všech vozidel a budeme hledat to vozidlo, které má největší sčítanec KW-statistiky. V tomto případě to je vozidlo s číslem 112, které z dalšího testování odstraníme.

	bez extrémn. hodnot
KW-statistika	3,4693
krit. hodnota	6,2514
p-hodnota	0,3248

Tabulka 4.17: Výsledky KW-testu – srovnání vozidel 101,102,103,111 trať Brno-Olomouc bez extrémních hodnot – směr zpět síť AC

Tabulka 4.17 představuje výsledky KW-testu poté, co jsme odstranili vozidlo 112. Z nich plyne, že již nezamítáme H_0 o shodě spotřeb testovaných vozidel. Jelikož všechny výběry již byly očištěny o odlehlá pozorování, testování na tomto místě končí a naším závěrem je, že vozidla 101, 102, 103 a 111 mají na síti AC stejně rozdělenou spotřebu.

4.3.2.5 Shrnutí výsledků

V případě trakčního systému AC vyšel pro oba směry stejný výsledek, tj. vozidla 101, 102, 103 a 111 mají stejnou spotřebu[23] - vozidlo 112 vždy během testování ze shody „vypadlo“. Je otázkou, co přesně tento výsledek způsobuje.

Na síti s trakčním systémem DC již výsledky vyšly různé – v směru tam se skupina testovaných pěti vozidel rozdělila na dvě podskupiny, které měly stejnou spotřebu, zatímco ve směru zpět měla všechna vozidla stejnou spotřebu.

Předmětem dalšího zkoumání by proto mělo být zjištění konkrétního důvodu pro případné rozdíly, například proč má pro trakční systém AC pouze vozidlo 112 jinou spotřebu než ostatní. Jelikož se jedná o jedno specifické vozidlo, které v obou směrech jeví známky jiné spotřeby, tak zde může být souvislost například s obsazením vlaku a tedy obdobím provozu daného vozidla.

Druhou otázkou je, proč výsledky na trakčním systému DC vychází tak „různorodě“. Vzhledem k této nekonzistenci výsledků v porovnání obou směrů zde může být souvislost se stylem jízdy řidiče, ale také již výše zmíněná obsazenost vlaku může hrát roli.

[23] Zde již volně řečeno – korektně „stejně rozdělenou spotřebu“

5 Analýza vzniku alarmů

Hlášení obsahující **alarmy** jsou pro **zpracování jednodušší než data procesní**. Je to tím, že pro nás důležitá data z této kategorie, především identifikační číslo alarmu, mají v tabulce z databáze zvláštní sloupec, a tedy není nutné provádět v tomto ohledu předzpracování dat.

Hlášení z alarmů však **nejsou přímo provázána s procesními daty**, tj. v každém hlášení z alarmu jsou uvedeny pro daný alarm důležité údaje, avšak například neznáme počet kilometrů, které vozidlo do daného hlášení alarmu ujelo, a tím pádem nedokážeme jednoduše určit přesný úsek na trati, kde alarm vznikl. Sice máme k dispozici GPS souřadnice alarmu, ale jen na jejich základě nelze jednoduše dopočítat příslušné číslo kilometru na trati.

Preprocessing dat v této kapitole se tedy bude věnovat především **propojení procesních dat a alarmů**, které je nezbytné pro provedení na to navazujících analýz. Implementaci z této části **spojíme s reportovacím systémem ke spotřebám energie** tak, abychom ve výsledku dokázali vytvořit celkový report o alarmech a spotřebě energie. Na závěr i v této kapitole provedeme statistické analýzy – zde nás bude zajímat, zda vznik alarmu má souvislost s konkrétním místem či úsekem na trati.

5.1 Preprocessing dat

Cílem preproccesingu dat z alarmů je především **vyjádřit místo jejich vzniku číslem kilometru na trati**, kde se vozidlo pohybuje a také počet kilometrů, které vozidlo do alarmu ujelo za sledované období celkově[24].

Mimo to bude cílem předzpracování také odstranění těch alarmů, které vznikly během odstavení v depu nebo během servisních opatření. Jinými slovy, **druhým cílem je ze všech alarmů vybrat pouze ty, které přímo souvisejí s jízdami nebo provozem vozidla na trati.**

K řešení obou cílů můžeme přistupovat dvěma způsoby:

i. Využití informace o GPS souřadnicích,
ii. Využití časového údajů zaslaného v rámci hlášení.

GPS souřadnice nám udávají přesné místo vzniku alarmu, čili s každým alarmem bychom si dokázali v rámci vizualizace například na mapě s tratěmi udělat obrázek o rozložení alarmů na trati. Pokud bychom ale chtěli **identifikovat úsek na trati**, kde se vyskytuje větší množství alarmů, tak to jen na základě GPS souřadnic není bez další informace možné provést. Museli bychom mít přehled nebo tabulku a dále mapovací funkci, která by **dvoudimensionální informaci** (souřadnice) dokázala převést na **jednu hodnotu ujetého kilometru na trati**. Vzhledem k tomu, že taková evidence neexistuje, je první cíl přístupem přes GPS souřadnice jen složitě dosažitelný.

Naopak **druhého cíle** by se dalo dosáhnout celkem jednoduchým způsobem: zmapovali bychom depa, kde mohou být vozidla odstavena, a získali jejich GPS souřadnice. Ze všech alarmů bychom pak odstranili pak ty, které se objevily v blízkosti depa – aparát na

[24] Na základě jednoho hlášení určíme tedy dvě hodnoty počtu ujetých kilometrů

určení vzdálenosti mezi dvojicí GPS souřadnic k dispozici máme a tudíž by i výpočetní náročnost nemusela být velká.

Budeme-li naopak chtít využít **přístupu využívajícího čas** zaslání daného hlášení, tak obou našich cílů dosáhneme pomocí procesních dat. Jak plyne z předchozí kapitoly, tak **po první fázi předzpracování procesních dat** dokážeme na listu „bloky" sešitu „tvorbaReportu.xlsb" (viz Obrázek 4.4) **pro každé hlášení stanovit**, **kolik kilometrů vozidlo do daného času ujelo** a **které jízdě** (číslu hlavního bloku) toto **hlášení přísluší**. Na základě této informace pak je již jednoduché určit i počet kilometrů, který vozidlo ujelo na příslušné trati.

Našeho **prvního cíle**, tj. vyjádření místa vzniku alarmu jako celkového počtu ujetých kilometrů za sledované období a počtu kilometrů, které vozidlo ujelo na trati, po které se v dané chvíli pohybuje, dosáhneme v případě, že se nám podaří **promítnout časy alarmů do časů procesních dat**. Jinými slovy využijeme toho, že pokud známe čas, tak dokážeme jednoznačně stanovit, kolik kilometrů vozidlo do tohoto času ujelo jak celkově, tak i na trati, na které se vozidlo zrovna pohybuje.

Druhého cíle dosáhneme tím, že budeme zkoumat hlavní blok, do kterého hlášení dle času svého zaslání spadá[25]. Pokud **alarm dle svého času spadne mezi dva hlavní bloky**, což odpovídá případu, že se **alarm uskutečnil mezi dvěma jízdami**, tak jej pro další výpočty uvažovat nebudeme, jelikož takovéto **hlášení nesouvisí přímo s provozem vozidla** na trati a mohlo vzniknout v době odstavení vozidla v depo apod.

Zařazení alarmů do procesních dat ilustrujeme na následujícím příkladu. Budeme chtít zařadit alarm, jehož parametry uvádí Tabulka 5.1. Cílem je určení, kolik kilometrů vozidlo ujelo, než daný alarm nastal.

Datum	Čas	SKODA Alarm ID
31.1.2017	**21:47:22**	**123**

Tabulka 5.1: Alarm 123 před zařazením do procesních dat

V **naprosté většině případů se čas alarmu nekryje se žádným časem z hlášení procesních dat**, což je hlavní problém, který zde musíme řešit. Proto pro každý alarm najdeme **první hlášení v procesních datech**, které leží **časově za**, a **poslední hlášení**, které leží **časově před daným alarmem**. Odpovídající dvě hlášení z procesních dat představuje Tabulka 5.2.

Hlavní blok	Subblok	Datum	Čas	GPS šířka	GPS délka	Celkem ujeto km	Aktuální rychlost
113	**15**	**31.1.2017**	**21:47:32**	**49,9712**	**16,37762**	**14394,97**	**1**
113	**15**	**31.1.2017**	**21:47:16**	**49,97126**	**16,37682**	**14394,91**	**28**

Tabulka 5.2: Úryvek procesní data pro zařazení alarmu

[25] Připomeňme, že hlavní bloky se tvoří na základě hlášení procesních dat, ze kterých byla vyjmuta hlášení s nulovou rychlostí

Spojíme-li hlášení z procesních dat a alarmů do jednoho datového souboru, pak jednoduchým **seřazením nejprve podle data a poté podle času** zařadíme alarm do procesních dat. Hodnotu ujetého kilometru a rychlost pro daný alarm určíme jako průměr z patřičných hodnot hlášení procesních dat ležící časově přímo před a za alarmem. V našem příkladu tedy dojdeme k výsledku, který znázorňuje Tabulka 5.3.

Hlavní blok	Subblok	Datum	Čas	GPS šířka	GPS délka	Celkem ujeto km	Aktuální rychlost
113	15	31.1.2017	21:47:32	49,9712	16,37762	14394,97	1
113	**15**	**31.1.2017**	**21:47:22**	**-**	**-**	**14394,94**	**15**
113	15	31.1.2017	21:47:16	49,97126	16,37682	14394,91	28

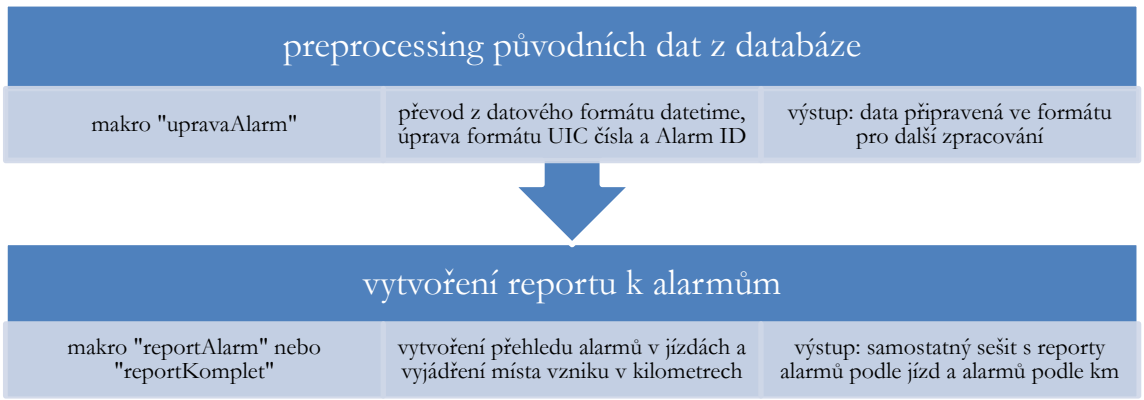
Tabulka 5.3: Procesní data včetně zařazení alarmu 123

Zároveň se zařazením alarmů mezi procesní data mimo počet ujetých kilometrů **určí i číslo hlavního bloku a subbloku**, do kterých alarm patří. Především určené číslo hlavního bloku je **velmi důležité** – pokud procesní data, mezi která je alarm zařazen, jsou ve stejném hlavním bloku, pak i alarm bude do tohoto hlavního bloku patřit.

Pokud však procesní data, mezi která bychom chtěli alarm zařadit, již **nejsou ve stejném hlavním bloku**, pak to znamená, že vozidlo v čase mezi těmito hlášeními provedlo obrátku nebo bylo odstavené v depu – jsou to tedy případy, které **nesouvisí přímo s provozem vozidla na trati**, a takový alarm nezahrneme do dalšího vyhodnocení.

Na základě výše uvedených důvodů proto **využijeme přístupu předzpracování alarmů přes časový údaj** zaslání jednotlivých hlášení alarmů. Shrnutí postupu při předzpracování a návod k němu představuje Obrázek 5.1.

Při realizace je důležité, abychom **pro konkrétní vozidlo měli k dispozici procesní data a alarmy pro stejné období**. Oba listy vložíme do sešitu „tvorbaReportu.xlsb“ a po vytvoření seznamu listů je nutné správně vybrat listy v buňkách „list – proc. data“ a „list – alarmy“. Následně můžeme pomocí příslušného tlačítka vytvořit celkový report, jehož součástí je mimo část ke spotřebám a rekuperaci energie též samostatný report k alarmům, nebo tu je i možnost vytvořit pouze report k alarmům.



Obrázek 5.1: Schéma postupu při předzpracování alarmů

Ať už v celkovém nebo samostatném reportu jsou vytvořeny dva listy:

i. „alarmy v blocích - …“ a číslo vozidla,
ii. „alarmy podle km - …“ a číslo vozidla.

První ze zmíněných listů je v podstatě kontingenční tabulka, která dává informaci o **četnostech konkrétních alarmů v jednotlivých jízdách**, tj. hodnoty řádků odpovídají číslu hlavního bloku (jízdy) a každý sloupec obsahuje jeden konkrétní alarm. Navíc pokud je report o alarmech součástí celkového reportu, tak po vytvoření soupisu tratí je do této tabulky zahrnuta i informace o číslu tratě i směru dané jízdy.

Vytvoření **druhého z listů** předchází provedení zařazení či **promítnutí alarmů do procesních dat**. V praxi makro, které vytváří listy s reporty k alarmům, na dočasný list vykopíruje nejprve všechna procesní data a pod ně všechna data k alarmům. Poté, jak již bylo naznačeno v příkladu výše, ve dvou fázích provede seřazení celého datového souboru, tj. procesní data a alarmy v jednom – nejprve seřadí všechna hlášení podle data a poté podle času.

Tímto postupem jsme **zařadili alarmy mezi procesní data**, určíme příslušnost každého z alarmů k hlavnímu bloku a počet ujetých kilometrů celkový a na dané trati. Pokud po zařazení spadá některý z alarmů **mezi dva různé hlavní bloky**, tak číslo jeho hlavního bloku bude nula. Po provedení všech kroků pak nulou označené alarmy odstraníme a zbytek je převeden do reportu.

Výsledkem je, jak naznačeno výše, že **pro každý alarm** máme určen **celkový počet ujetých kilometrů za sledované období** stejně jako **ujeté kilometry v dané jízdě**.

5.2 Statistická analýza závislosti vzniku alarmu s úsekem na trati

Údaje z celkového reportu využijeme ke zkoumání souvislosti vzniku alarmu s určitým místem na trati, tj. chceme zjistit, zda **pravděpodobnost vzniku alarmu je na některém z úseků vyšší**, než u ostatních.

5.2.1 Teoretický základ

Otázka, zda při projetí určitého úseku vozidlo nahlásí alarm, má jako odpověď buď „ano, v tomto úseku vznikl alarm“ nebo „ne, vozidlo projelo úsek bez toho, aby nahlásilo alarm“. Ze statistického hlediska se tedy jedná o alternativní rozdělení, kde jako „úspěch“ budeme považovat vznik alarmu a jako „neúspěch“ opačný případ.

Každý úsek i na sledované trati bude „zastoupen“ vlastní náhodnou veličinou z alternativního rozdělení

$$X_i \sim A(p_i),$$

kde p_i je pravděpodobnost vzniku alespoň jednoho alarmu na sledované trati během jedné jízdy a posléze $1 - p_i$ je pravděpodobnost, že na úseku i během jízdy žádný alarm nenastane.

Než začneme zkoumat párově jednotlivé úseky, tak nejprve otestujeme, zda pravděpodobnost vzniku alarmu na všech úsecích je stejná. **První fázi analýzy** tvoří **test shody alternativních rozdělení** (6).

Předpokládejme, že jsme sledovanou trať rozdělili do I úseků – platí tedy $i \in \{1,2,\dots,I\}$. Pak budeme testovat hypotézu:

$$H_0: p_1 = p_2 = \cdots = p_I, \tag{5.1}$$

kde $\boldsymbol{p_1, p_2, \dots p_I}$ odpovídají výše zavedeným parametrům alternativních rozdělení a tedy **pravděpodobnosti, že na $\boldsymbol{i}$-tém úseku vozidlo nahlásí alespoň jeden alarm**. Jinými slovy **testujeme hypotézu, zda pravděpodobnost vzniku alarmu je na všech úsecích stejná.**

Alternativní hypotézou $\boldsymbol{H_1}$ pak je, že alespoň jeden úsek i a s ním spojené p_i od ostatních liší, tj. **pravděpodobnost vzniku alarmu v tomto úseku bude buď větší, nebo menší než u ostatních úseků.**

Pro provedení tohoto testu je nutné sestavit kontingenční tabulku, která bude mít dva řádky a I sloupců, a tedy každý sloupec reprezentuje jeden úsek. Označme x_i jako počet jízd, v rámci kterých pro daný úsek i došlo k alarmu a n jako počet celkově absolvovaných jízd.

Do prvního řádku naší kontingenční tabulky uvedeme pro každý úsek patřičnou hodnotu x_i, a do druhého řádku počet jízd, kdy vozidlo žádný alarm nenahlásilo – tedy $n - x_i$. Součet hodnot v každém sloupci po sestavení tabulky by na konci měl odpovídat celkovému počtu jízd.

Hypotézu $\boldsymbol{H_0}$ otestujeme testem nezávislosti v kontingenční tabulce, kterou vytvoříme dle postupu v předchozím odstavci. Hladinu významnosti, na které budeme test nezávislosti provádět, označíme jako α_1. Důležité je, aby pro dopočtené očekávané četnosti o_{ij} platilo $o_{ij} \geq 5$.

V případě, že $\boldsymbol{H_0}$ zamítneme, tak evidentně má smysl se ptát, **který úsek nebo úseky má nebo mají vyšší pravděpodobnost vzniku alarmu**. Pak postoupíme k **druhé fázi testování**, kde budeme jednotlivé úseky na trati proti sobě zkoumat po párech. Nulová hypotéza v těchto testech bude mít tvar:

$$H_0: p_i = p_j; i, j \in \{1,2,\dots,I\}, i \neq j, \tag{5.2}$$

Oproti tomu testujeme oboustrannou alternativu, tj. $p_i \neq p_j$. Jedná se tedy o test shody parametrů dvou alternativních rozdělení.

Testová statistika má následující podobu (11):

$$z = \frac{\widehat{p_i} - \widehat{p_j}}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \tag{5.3}$$

kde

$\widehat{p_i}, \widehat{p_j}$ jsou odhady parametrů p_i a p_j a platí $\widehat{p_i} = \frac{x_i}{n_i}$,

$\hat{p} = \frac{n_i\widehat{p_i} + n_j\widehat{p_j}}{n_i + n_j}$.

Jelikož v našem případě platí $n_1 = n_2 = \cdots = n_I = n$, tak se podoba vzorce pro výpočet testové statistiky navíc zjednoduší. Zdroj (11) navíc doporučuje, aby pro každé i platilo $\min\{n, n - x_i\} \geq 5$.

V případě platnosti H_0 má z-statistika aproximativně normované normální rozdělení. Testujeme-li na hladině významnosti α_2, tak má kritický obor podobu $\left(-\infty, u_{\frac{\alpha_2}{2}}\right) \cup \left(u_{1-\frac{\alpha_2}{2}}, \infty\right)$ a p-hodnotu testu určíme jako $2(1 - F(|z|))$, kde $F(x)$ zde je distribuční funkce normovaného normálního rozdělení.

Párové testy jednotlivých dvojic úseků provádíme z hlediska celkového souboru opakovaně a tím pádem se setkáváme s **problémem násobného testování hypotéz** a postupného navyšování pravděpodobnosti, že při porovnání dvojic odhalíme statisticky významný rozdíl tam, kde ve skutečnosti není.

Tento problém řeší například **Bonferroniho procedura** (12). Uvažujme α_1 jako hladinu významnosti, na které jsme prováděli test shody parametrů alternativních rozdělení (test se všemi úseky). Při pozdějším párovém testování bychom zamítli nulovou hypotézu o shodě parametrů p_i a p_j, pokud pro patřičnou p-hodnotu p^* bude platit $p^* < \frac{\alpha_1}{m}$, kde m je celkový počet provedených testů.

Tento **postup je však velmi konzervativní**, tj. Bonferroniho korekce může považovat více dvojic za shodné, než jich ve skutečnosti shodných je. **Proto budeme využívat Holm-Bonferroniho postup** při korekci hladiny významnosti (13), který je založen na Bonferroniho korekci, ale není tak konzervativní.

Na základě výsledků jednotlivých testů i konkrétních hodnot reálných dat vyhodnotíme, zda pro **zkoumaný alarm existuje jeden nebo více úseků, které by byly „více rizikové" než ostatní.**

Pro **zvolený alarm** provedeme **obě fáze testu nezávisle na sobě pro oba směry**. Přitom platí, že hodnoty ujetých kilometrů na trati **pro jízdy ve směru zpět musíme transformovat**. Myšlenku ilustrujeme na jednoduchém příkladu, kdy vozidlo jede z místa A do místa B, což budeme považovat za směr „tam". Problémem je, že počty ujetých kilometrů počítáme jako ujetou vzdálenost od počátečního bodu, což pro jízdy ve směru „tam" je okolí místa A, naopak pro jízdy ve směru „zpět" okolí místa B.

Chceme-li srovnávat výsledky obou směrů, pak musíme nejprve provést zmíněnou transformaci dat pro jízdy ve směru „zpět“ tak, aby i tyto počtu ujetých kilometrů odpovídaly vzdálenosti od místa A.

5.2.2 Praktické provedení

Výše představená metodika byla pro účel této práce naprogramována v jazyce VBA a je součástí maker „shodaAlternRozd“ a „analyzaUsekyTrat“ – první z nich provádí všechny nutné kroky v rámci první fáze, tj. testu shody alternativních rozdělení, zatímco druhé provádí totéž pro druhou fázi - párové porovnání jednotlivých dvojic úseků a vyhodnocení statistické významnosti testů s korekcí hladiny významnosti podle Holm-Bonferroniho postupu.

V rámci této práce jsme pro praktické představení výše uvedené metodiky vybrali dvě vozidla, která **jezdí na trati Brno-Olomouc** a navíc jejich **nejčetnější alarm je stejný**. Pokusíme se tedy zjistit, zda **pro tento alarm existuje úsek**, kde se vyskytuje **častěji než na jiných úsecích**, srovnáme přitom, zda se pro oba směry jízdy chová stejně a výsledky každého z vozidel porovnáme.

Pro obě vozidla zkoumáme alarm s číslem 811008000. Označme pak první ze zkoumaných vozidel číslem 1001 a druhé 1002. V předchozím odstavci zmíněné makro „shodaAlternRozd“ spouštíme v listu „alarmy podle km…“, kde do příslušných buněk zadáme číslo alarmu a tratě, kterou chceme zkoumat.

Makro si navíc od uživatele vyžádá **směr referenční tratě**. Účel tohoto údaje jen ten, abychom mohli srovnávat mezi sebou vozidla, která sice jezdí po stejné trati, ale při vytvoření soupisu tratí není referenční směr stejný. Tento případ nastává v našem příkladu, kdy pro vozidlo 1001 a typ tratě 1 je „referenční směr tam“ Brno-Olomouc, zatímco u vozidla 1002 tomu je právě naopak.

Volit si můžeme i **délku jednoho úseku**, pro který platí, že čím větší bude, tím spíše budou splněny předpoklady testů[26], ale zato rozdělení tratě do úseků bude hrubější, při menší délce úseku bude platit opak. Stejně jako u spotřeb energie i **zde uživatel zadá hladinu významnosti testu**.

Po spuštění makra je postupně vytvořen pro každý směr jízdy na sledované trati list „rizik“. trať *číslo zkoumané tratě*…“, na kterém se provádí veškeré výpočty.

Pro náš příklad jsme zvolili **délku úseku 20 km**, abychom **splnili předpoklad o očekávaných četnostech** při testu nezávislosti (viz předchozí podkapitola) – při délce 10 km byly již menší než pět. Testovat budeme ve všech případech na hladině významnosti $\alpha = 10\ \%$.

Podívejme se nejprve na vyhodnocení pro vozidlo 1001, které provedeme pečlivě, zatímco u ostatních jen představíme výsledky.

[26] Především se to týká požadavku na počty pozorování. Pro test shody alternativních rozdělení má volba délky jednoho úseku pak vliv na hodnoty očekávaných četností.

	20	40	60	80	100	120	140	160	180	200	součty
1	14	19	16	7	15	15	15	15	6	3	125
0	24	19	22	31	23	23	23	23	32	35	255
součty	38	38	38	38	38	38	38	38	38	38	380

Tabulka 5.4: Skutečné četnosti – vozidlo 1001 směr tam

Tabulka 5.4 představuje skutečné četnosti výskytu alespoň jednoho alarmu během jízdy na daném úseku (řádek označen jedničkou) oproti počtu jízd, kde na daném úseku žádný alarm nevznikl (řádek označen nulou). Popisky sloupců odpovídají horní mezi úseku, například první sloupec značí úsek 0-20km. Podle postupu při testu nezávislosti vytvoříme kontingenční tabulku pro očekávané četnosti, výsledek představuje Tabulka 5.5.

	20	40	60	80	100	120	140	160	180	200	součty
1	12,5	12,5	12,5	12,5	12,5	12,5	12,5	12,5	12,5	12,5	125
0	25,5	25,5	25,5	25,5	25,5	25,5	25,5	25,5	25,5	25,5	255
součty	38	38	38	38	38	38	38	38	38	38	380

Tabulka 5.5: Očekávané četnosti – vozidlo 1001 směr tam

Test shody alternativních rozdělení pro směr „tam“ má ve výsledku **p-hodnotu** (viz Tabulka 5.6) **výrazně menší než hladina významnosti** a nulovou hypotézu o shodě zamítáme. Znamená to tedy, že **alespoň jeden úsek rovnost parametrů pravděpodobnosti vzniku alarmu** v daném směru výrazně porušuje.

vyhodnocení	
test. krit	29,1482
p-hodn.	<0,001

Tabulka 5.6: Vyhodnocení testu shody alternativních rozdělení – vozidlo 1001 směr tam před úpravou

Postoupíme k další fázi a **provedeme testy o shodě parametrů dvou alternativních rozdělení** pro jednotlivé dvojice úseků.

	20	40	60	80	100	120	140	160	180	200
20		0,2472	0,6388	0,0726	0,8133	0,8133	0,8133	0,8133	0,0372	
40	0,2472		0,4899	0,0037	0,3561	0,3561	0,3561	0,3561	0,0015	
60	0,6388	0,4899		0,0246	0,8154	0,8154	0,8154	0,8154	0,0114	
80	0,0726	0,0037	0,0246		0,0430	0,0430	0,0430	0,0430	0,7607	
100	0,8133	0,3561	0,8154	0,0430		1,0000	1,0000	1,0000	0,0210	
120	0,8133	0,3561	0,8154	0,0430	1,0000		1,0000	1,0000	0,0210	
140	0,8133	0,3561	0,8154	0,0430	1,0000	1,0000		1,0000	0,0210	
160	0,8133	0,3561	0,8154	0,0430	1,0000	1,0000	1,0000		0,0210	
180	0,0372	0,0015	0,0114	0,7607	0,0210	0,0210	0,0210	0,0210		
200										

Tabulka 5.7: P-hodnoty k párovým testům dvojic úseků – vozidlo 1001 směr tam

Výstup, který generuje makro „analyzaUsekyTrat“, představuje Tabulka 5.7. **Zelená barva buňky** značí p-hodnoty takových testů, jejichž **nulovou hypotézu** po Holm-Bonferroniho korekci hladiny významnosti **nezamítáme**. **Prázdné zůstaly buňky pro testy**, u kterých **minimálně jedna ze skupin nesplnila požadavek o alespoň pěti pozorováních**. Takové úseky musely být z **druhé fáze testování zcela vyloučeny**. Jelikož ale hledáme úseky s nejvyšším rizikem vznikem alarmu, tak nám toto vyloučení konečný výsledek stejně neovlivní. **P-hodnoty**, jejichž hodnota ve výsledné tabulce jsou **rovny jedné**, vznikly testováním takové dvojice, kde **oba úseky měly stejný počet jízd, při kterých vznikl alarm**. Testová statistika je pak rovná nule, což způsobuje **„nesmyslné“ p-hodnoty rovné jedné**. **Výsledek testu**, tj. nezamítnutí nulové hypotézy o shodě testovaných parametrů, však **zůstane neovlivněn**.

Evidentní tedy je, že **úsek 180-200km**, který byl z druhé fáze testování vyloučen, bude mít **nižší pravděpodobnost vzniku alarmu**. Jak naznačují nízké p-hodnoty, které uvádí Tabulka 5.7, a skutečné četnosti a Tabulka 5.4, tak navíc i na **druhém krajním úseku 160-180km** je menší pravděpodobnost vzniku alarmu. Hypotézu o shodě parametrů sice při porovnání s ostatními úseky nikdy nezamítáme, ale oproti ostatním dvojícím jsou p-hodnoty v řádku či sloupci „180“ nižší.

vyhodnocení	
test. krit	8,9215
p-hodn.	0,2583

Tabulka 5.8: Vyhodnocení testu shody alternativních rozdělení vozidlo 1001 směr tam po úpravě

Pokud **tyto dva úseky odstraníme** z celého vyhodnocení (tedy i první fáze analýzy), tak **p-hodnota** k testu shody alternativních rozdělení, kterou představuje Tabulka 5.8, je již **vyšší než hladina významnosti** a nulovou hypotézu k tomuto testu nezamítáme. V tuto chvíli jsme ověřili, že úseky na trati mimo dva již zmíněné úseky mají stejnou pravděpodobnost vzniku alarmu, a naším závěrem je, že se **sledovaný alarm ve směru tam zjevně neváže na konkrétní úsek na trati**, ale naopak vzniká nezávisle na poloze vozidla.

	20	40	60	80	100	120	140	160	180	200	součty
1	4	12	16	17	17	17	17	17	14	4	135
0	38	30	26	25	25	25	25	25	28	38	285
součty	42	42	42	42	42	42	42	42	42	42	420

Tabulka 5.9: Skutečné četnosti – vozidlo 1001 směr zpět

Ve směru „zpět“ vypadá situace velmi podobně. Při počátečním testu všech úseků je **p-hodnota testu také menší než hladina významnosti**. Výsledek znázorňuje Tabulka 5.10.

vyhodnocení	
test. krit	27,3450
p-hodn.	0,0012

Tabulka 5.10: Vyhodnocení testu shody alternativních rozdělení vozidlo 1001 směr zpět před úpravou

Oproti ostatním ale mají velmi **malé skutečné četnosti krajní úseky 0-20km a 180-200km**, které vzhledem k jejich hodnotám ani nezahrnujeme do párového testování. Provedeme tedy totéž, co v předchozím případě, a dva uvedené úseky odstraníme. P-hodnota testu shody alternativních rozdělení pak dosahuje velmi vysokých hodnot (viz Tabulka 5.11) a náš **závěr ze zkoumání tohoto úseku bude stejný, jako v případě směru tam**.

vyhodnocení	
test. krit	2,5191
p-hodn.	0,9256

Tabulka 5.11: Vyhodnocení testu shody alternativních rozdělení vozidlo 1001 směr zpět po úpravě

Přejdeme k vozidlu 1002. Jak již bylo řečeno výše, tak zde musíme při vytvoření analýzy rizikovosti úseku volit jako referenční směr „zpět".

Vyhodnocení – směr tam		Vyhodnocení – směr zpět	
test. krit	17,3115	test. krit	41,9797
p-hodn.	0,0441	p-hodn.	<0,001

Tabulka 5.12: Vyhodnocení testu shody alternativních rozdělení vozidlo 1002 oba směry před úpravou

Pro **oba směry však budou závěry stejné, jako pro vozidlo 1001**. Počáteční test pro všechny úseky má pro oba směry ve výsledku p-hodnotu nižší než je hladina významnosti a zamítneme nulovou hypotézu o shodě. Shrnutí představuje Tabulka 5.12.

Opět ale **krajní úseky mají výrazně nižší skutečné četnosti než ostatní** – pro směr tam[27] to jsou úseky 0-20km a 20-40km a pro směr zpět naopak úseky 160-180km a 180-200km. Totéž potvrzují i párové testy úseků. **Po odstranění zmíněných úseků** jsou **p-hodnoty testu shody alternativních rozdělení výrazně vyšší** než hladina významnosti (viz Tabulka 5.13), tudíž **i pro vozidlo 1002** je pro oba směry jízdy závěrem, že **sledovaný alarm zřejmě nesouvisí s konkrétním úsekem na trati**.

[27] Směr tam po provedené transformaci hodnot ujetých kilometrů podle referenčního směru

Vyhodnocení – směr tam	
test. krit	1,7816
p-hodn.	0,9709

Vyhodnocení – směr zpět	
test. krit	4,4091
p-hodn.	0,7316

Tabulka 5.13: Vyhodnocení testu shody alternativních rozdělení – vozidlo 1002 oba směry po úpravě

Naším celkovým závěrem tedy je, že **testy na obou dvou vozidlech nesvědčí o tom**, že by **vznik alarmu 811008000 souvisel s konkrétním úsekem na trati**.

6 Závěr

Oproti původním předpokladům, kdy se největší část této práce měla věnovat analýzám diagnostických hlášení obecně, se větší část práce zaměřuje na zpracování dat. Souvisí to především s tím, že uspořádání původní databáze neumožňovalo po získání dat z databáze pomocí SQL-dotazů okamžitě bez dalších úprav na těchto datech provádět analýzy. Způsob ukládání důležitých údajů do textového řetězce je pak hlavním důvodem, proč je nutné provést rozsáhlé předzpracování většiny údajů.

Dalším důvodem, proč zpracování dat hraje podstatnou roli v rozsahu práce, jsou okolnosti kolem jednotlivých jízd. Vzhledem k tomu, že provozovatel vozidel může nasazení variabilně měnit, bylo nutné vytvořit systém, který dokáže určit podobnost jednotlivých jízd. Jednou z hlavních výhod vytvořeného konceptu je skutečnost, že je obecně aplikovatelný na jakékoli tratě, tj. nejen na ty, ze kterých jsme získali data pro tuto práci. Neméně podstatné je, že přehled o souřadnicích jednotlivých zastávek jsme obdrželi až v pozdějším průběhu této práce, a proto metodika pro klasifikaci tratí tyto informace využívá pouze v rámci doplnění našeho původního postupu a naopak na nich není založená, jak by odpovídalo intuitivnímu předpokladu. Příslušná kapitola 3 o určení podobnosti jízd proto také tvoří nezanedbatelnou část práce.

S použitím jazyka VBA byl vytvořen celý reportovací systém jak pro alarmy, tak pro spotřeby energie, který dokáže překlenout nedostatky a problémy spojené s původní podobou uložení dat. Jako vstupní data lze tedy použít původní „hrubou“ podobu dat za jakkoli dlouhé sledované období. Makra vytvořena v rámci této práce vytvoří přehledné reporty, kde pro spotřeby energie máme obecný přehled o tom, jaká byla spotřeba a rekuperace při jednotlivých jízdách, a pro alarmy dokážeme vyjádřit jak jejich rozložení v jízdách, tak i rozložení v závislosti na průběhu počtu ujetých kilometrů.

Součástí zpracování dat je také již zmíněná metoda určení podobnosti jízd, která je robustní vůči případu, kdy provozovatel vozidel změní jejich nasazení, a tím i zastávky, které vozidla projíždějí.

V rámci analýz jsme zjistili rozdíl ve spotřebách některých vozidel, která se pohybují na stejné trati, což může být pro výrobce velmi cenná informace. Pokud by se v rámci průzkumů navazujících na tuto práci tyto rozdíly měly potvrdit, tak by se z nich daly vyvodit patřičné důsledky podle toho, jaký by byl konkrétní důvod pro rozdíl ve spotřebách.

V případě alarmů je naopak zajímavé zjištění, že sledovaný alarm zřejmě s tratí nesouvisí a v úvahu připadá možnost, že by příčinou mohla být vozidla samotná. I zde se nabízí možnost na tato zkoumání navázat dalšími analýzami, zda například zjištěné výsledky platí i pro ostatní vozidla a tedy by se jednalo o systematicky se vyskytující problém.

Reference

1. Správa železniční dopravní cesty. *Železniční mapy ČR.* **[Online] 2012. [Citace: 7. leden 2017.] http://www.szdc.cz/o-nas/zeleznicni-mapy-cr.html.**

2. ŠKODA Transportation a.s. **Katalog elektrických jednotek.** *Elektrická jednotka InterPanter.* **[Online] 2016. [Citace: 28. leden 2017.] http://www.skoda.cz/cs/produkty/elektricke-jednotky/jednopodlazni-elektricka-jednotka/Contents.3/0/B793B3A52DAC25FAD92E7B97B1B617E9/resource.pdf.**

3. —. Produktový list. *Elektrická jednotka InterPanter.* **[Online] 2016. [Citace: 28. leden 2017.] http://www.skoda.cz/cs/produkty/elektricke-jednotky/elektricka-jednotka-interpanter/Contents.3/0/C4218E60BF6A40089836174CC15C5BBF/resource.pdf.**

4. Ludvíček, Pavel. *Pokročilé metody řízení trajektorie modelu stanice v prostředí OPNET Modeler.* **Brno : Fakulta elektrotechniky a komunikačních technologií, Vysoké učení technické v Brně, 2011.**

5. Blatná, Dagmar. *Neparametrické metody - Testy založené na pořádkových a pořadových statistikách.* **Fakulta informatiky a statistiky : Vysoká škola ekonomická v Praze, 1996.**

6. Reif, Jiří. *Metody matematické statistiky.* **Fakulta aplikovaných věd : Západočeská univerzita v Plzni, 2004.**

7. Dixon, Wilfrid. **Processing Data for Outliers.** *Biometrics.* **1953.**

8. Jarque, Carlos a Bera, Anil K. **A Test for Normality of Observations and Regression Residuals.** *International Statistical Review.* **1987, Sv. 55, 2.**

9. Lilliefors, Hubert. **On the Kolmogorov-Smirnov test for normality with mean and variance unknown.** *Journal of the American Statistical Association.* **1967, Sv. 62.**

10. P. Verma, Surendra a Quiroz-Ruiz, Alfredo. **Critical values for six Dixon tests for outliers in normal samples up to sizes 100, and applications in science and engineering.** *Revista Mexicana de Ciencias Geológicas.* **2006.**

11. Utts, Jessica a Heckard, Robert. *Mind on Statistics.* **místo neznámé : CENGAGE Learning, 2013. 978-1-285-46318-6.**

12. Dubjaková, Eva. **Metody mnohonásobného porovnání pro jednoduché třídění.** *Diplomová práce.* **Brno : Masarykova univerzita, 2009.**

13. Holm, Sture. **A simple sequentially rejective multiple test procedure.** *Scandinavian Journal of Statistics.* **6, 1979.**

Přílohy na CD

i. DP_A15N0013P_Daniel_Spale.pdf (i ve zdrojovém formátu *.docx) – elektronická verze textu této práce
ii. PERSONAL_DP.xlsb – osahuje makra vytvořená v této práci.
iii. seznamModulu.xlsx – seznam maker a modulů, ve kterých se nachází.
iv. soubor tvorbaReportu.xlsb – soubor pro tvorbu reportů.
v. DP_manual_reporty_analyzy.pdf (i ve zdrojovém formátu *.docx) – uživatelský manuál pro reportovací systém a statistické analýzy.
vi. Soubor 170101_2016,2017,2018.xlsb (procesní data tří vozidel) – na něm lze vyzkoušet funkčnost maker pro preprocessing procesních dat.
vii. Složka Analyza_Energie. Obsahuje:
    - Složka data – upravené datové soubory pro vložení do sešitu tvorbaReportu.xlsb, na základě kterých lze vytvořit reporty pro spotřebu energie.
    - Složka reporty – vytvořené reporty s daty, na kterých jsou provedeny analýzy spotřeby energie.
    - Složka analýzy – soubory s výsledky analýz spotřeby energie (podkapitola 4.3.2).
    - Složka normalita_MATLAB – skript pro testy normality upravených výběrů.
    - Soubor Energie_analyza_vzor.xlsx – obsahuje vzorový list „Kruskal-Wallis“, ve kterém lze provádět příslušné testy
viii. Složka Analyza_Alarm. Obsahuje:
    - Složka data – upravené datové soubory pro vložení do sešitu tvorbaReportu.xlsb, na základě kterých lze vytvořit celkové reporty pro spotřebu energie a alarmy.
    - Složka reporty – vytvořené reporty s daty, ve kterých jsou provedeny analýzy vzniku alarmů.
ix. Složka Poster – obsahuje poster ve formátu *.pub i *.pdf