České vysoké učení technické v Praze
Fakulta elektrotechnická
Katedra telekomunikační techniky

Využití signalizačních dat z mobilní sítě

Use of Signalling Data from a Mobile Network

Diplomová práce

Jakub Staněk

Studijní program:	Komunikace, multimédia a elektronika
Studijní obor:	Sítě elektronických komunikací
Vedoucí práce:	Ing. Robert Bešťák, Ph.D.

Praha 2016

Čestné prohlášení

Prohlašuji, že jsem zadanou diplomovou práci zpracoval sám s přispěním vedoucího práce a konzultanta a používal jsem pouze literaturu v práci uvedenou. Dále prohlašuji, že nemám námitek proti půjčování nebo zveřejňování mé diplomové práce nebo její části se souhlasem katedry.

České vysoké učení technické v Praze
Fakulta elektrotechnická

katedra telekomunikační techniky

ZADÁNÍ DIPLOMOVÉ PRÁCE

Student: **Bc. Jakub Staněk**

Studijní program: Komunikace, multimédia a elektronika
Obor: Sítě elektronických komunikací

Název tématu: **Využití signalizačních dat z mobilní sítě**

Pokyny pro vypracování:

S využitím registru osob a registru územní identifikace, adres a nemovitostí navrhněte způsob zpracování signalizačních dat z mobilní sítě za účelem stanovení počtu osob v prostoru a čase v administrativně náročně členěném území.

Seznam odborné literatury:

[1] Walke, B. H.: *Mobile Radio Networks, Networking and Protocols*. John Wiley & Sons, Stuttgart 1999. ISBN: 0-471-97595-8.
[2] EMC Education Services: *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. Wiley 1st edition, January 2015. ISBN: 1-118-87613-X.
[3] Baesens, B: *Analytics in a Big Data World: The Essential Guide to Data Science and its Applications*. Wiley; 1 edition. April 2014. ASIN: B00JR5LAC6.

Vedoucí: Ing. Robert Bešťák, Ph.D.

Platnost zadání: do konce letního semestru 2016/2017

prof. Ing. Boris Šimák, CSc.
vedoucí katedry

prof. Ing. Pavel Ripka, CSc.
děkan

V Praze dne 21. 12. 2015

Anotace

Signalizační data v mobilních sítích obsahují informace o umístění uživatelů v buňkách. Většina je získána procedurou aktualizace polohy. Po zpracování nabytých dat lze obyvatelstvo klasifikovat a určit vývoj počtu obyvatel v jednotlivých skupinách v různých časových oknech. Také je možné stanovit trajektorie pohybu obyvatel mezi územními celky. Z důvodů velkého objemu dat je k jejich analýze nutné použít speciálních nástrojů. Tato práce je řešena pomocí otevřeného aplikačního rámce Hadoop.

Klíčová slova: Aktualizace polohy, velká data, Hadoop

Summary

Signalling data in a mobile network provide information on users' location in cells. Most of them are acquired via a Location Update procedure. After processing the data it is possible to classify the population and determine a number of inhabitants in these classes using various time windows. It is also possible to determine the trajectories of paths, which people take between territorial units. Due to big data volumes it is necessary to use special tools for their analysis. In this thesis the open-source software framework Hadoop is used.

Index Terms: Location Update, Big Data, Hadoop

Poděkování

Rád bych poděkoval vedoucímu práce Ing. Robertu Bešťákovi, Ph.D. za cenné odborné i metodické rady při zpracování této diplomové práce i za další pomoc přesahující její rámec.

Obsah

1 Úvod **1**

2 Mobilní sítě **3**

2.1 Buňková rádiová síť 3

2.2 Architektura GSM sítí 5

2.2.1 Mobilní stanice 5

2.2.2 Systém základnových stanic 7

2.2.3 Síťový spojovací subsystém 8

2.2.4 Operační subsystém 9

2.3 Správa mobility 9

2.3.1 Skupiny buněk 10

2.3.2 Aktualizace polohy mobilní stanice 10

3 Administrativní členění ČR **14**

3.1 Evropské systémy územního členění 14

3.2 Soustava územních prvků ČR 14

3.2.1 Obec 15

3.2.2 Základní sídelní jednotka 16

3.3 Sčítání lidu, domů a bytů 18

4 Analýza velkých dat **20**

4.1 Procesní model 20

4.2 Technické prostředky 21

4.3 Platforma Hadoop 22

4.3.1 Komponenta MapReduce 22

4.3.2 Komponenta HDFS 22

4.3.3 Další komponenty 23

5 Zpracování dat z mobilní sítě **24**

5.1 Struktura vstupních dat 24

5.2 Definice pojmů 25

5.3 Počet obyvatel v územních prvcích 25

5.3.1 Výběr dat ze SLDB 25

5.3.2 Analýza dat z mobilní sítě 26

5.3.3 Výsledky 27

5.4 Trajektorie pohybu obyvatel 35

5.4.1 Výběr dat ze SLDB 35

5.4.2 Analýza dat z mobilní sítě 35

5.4.3 Výsledky 36

6 Závěr **42**

Seznam zkratek **44**

Literatura **46**

1 Úvod

Tržby evropských mobilních operátorů za poskytované mobilní služby v posledních letech klesají. To je především způsobeno snižováním spotřeby a zlevňováním hlasových služeb a služeb s přidanou hodnotou. Obvykle jsou označovány zkratkou VAS (*Value-Added Service*). Jejich příkladem jsou textové zprávy SMS (*Short Message Service*) a multimediální zprávy MMS (*Multimedia Message Service*).

Ztráta zájmu o tyto služby je dána jejich postupným nahrazováním službami datovými. I přes soustředění na perspektivní odvětví poskytování mobilního připojení k Internetu, nejsou mobilní operátoři schopni úbytek příjmů z klasických mobilních služeb zcela nahradit. Proto hledají nové příležitosti příjmů.

Jednou z možností je prodej informací nabytých z lokalizačních dat jejich zákazníků. Ta jsou získávána od samého počátku budování mobilních sítí, protože jsou klíčová pro zajištění mobility uživatelů v síti. Až nyní je výpočetní technika, která je schopná zpracovat velké množství dat ze signalizace mobilních sítí, natolik finančně dostupná, aby byl nákup informací ze zpracovaných lokalizačních dat pro obchodní společnosti a veřejnoprávní právnické osoby dostatečně ekonomicky zajímavý.

Práce je rozdělena do šesti kapitol. Ve druhé kapitole jsou vysvětleny principy buňkových rádiových sítí. Dále je popsána architektura druhé generace mobilních sítí s důrazem na části potřebné ke správě mobility, především procesu aktualizace polohy. Ten je základním zdrojem lokalizačních dat.

Lokalizace uživatelů v buňkách sítě není tolik zajímavá, jako jejich umístění v územních prvcích. Proto se další kapitola zabývá administrativním členěním České republiky – evropskými systémy územního členění a soustavou územních prvků z registru sčítacích obvodů a budov. Využil jsem některých ukazatelů z výsledků Sčítání lidu, domů a bytů Českého statistického úřadu pro porovnaní s výslednými hodnotami.

Velké objemy signalizačních dat není možno zpracovávat běžnými postupy a nástroji. Ve čtvrté kapitole je popsán možný způsob analýzy a potřebné technické prostředky. Také je zde sekce o sadě projektů pro práci s velkými daty Hadoop, který jsem využil pro získání výsledků. Jsou popsány jeho hlavní komponenty – výpočetní aplikační rámec, souborový systém a další.

Pátá kapitola obsahuje informace o mém způsobu zpracování dat z mobilní sítě. Je zde popsána struktura vstupních dat, kterých jsem měl k dispozici. Kapitola pokračuje popisem dvou analýz, které jsem zpracoval. První ukazuje vývoj počtu různě klasifikovaných obyvatel v územních prvcích. Druhá se zabývá cestami obyvatel absolvovanými v jednom dni. V obou případech je popsán výběr dat ze Sčítání lidu, domů a bytů. Dále

je vysvětleno řešení daného problému. Na konci je výčet, zobrazení a popis některých výsledků. V poslední kapitole jsou výsledky práce shrnuty a navrženy možnosti budoucího pokračovaní.

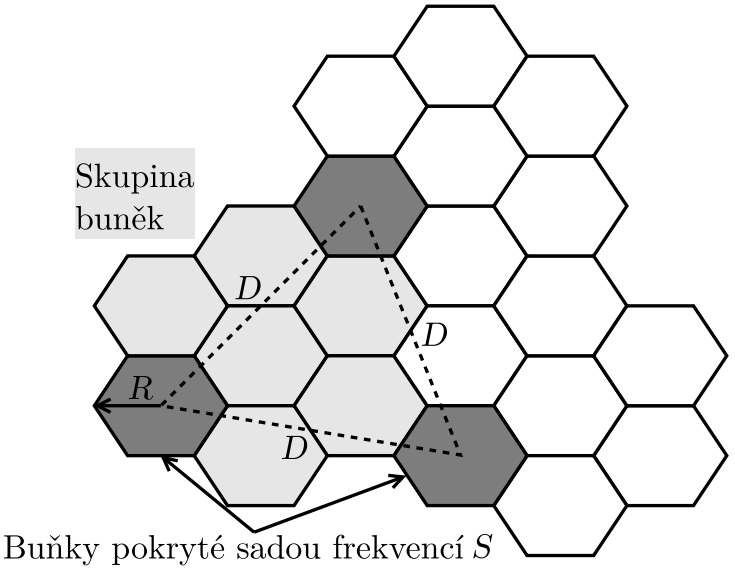
Praktickou pasáž této diplomové práce jsem řešil jako dílčí část dlouhodobého projektu, na kterém spolupracují České vysoké učení technické v Praze, MODATA a.s., VŠB – Technická univerzita Ostrava a Vysoká škola ekonomická v Praze s využitím infrastruktury Národního superpočítačového centra IT4Innovations. Více informací lze nalézt na stránkách projektu `http://modata.vsb.cz`.

2 Mobilní sítě

Mobilní sítě využívají omezené velikosti šířky kmitočtového pásma. Například digitální mobilní systémy druhé generace – GSM (*Global System for Mobile Communications*) mají v základním standardu P-GSM-900 alokováno frekvenční pásmo o šířce 25 MHz pro každý směr komunikace. Aby byl tento frekvenční příděl efektivněji využit, jsou mobilní sítě stavěny na principu buňkových rádiových sítí.

2.1 Buňková rádiová síť

V buňkových sítích je území pokryté signálem rozděleno do buněk. Při modelování sítí tohoto typu se často využívá zjednodušení tvaru buněk na šestiúhelník se základnovou stanicí umístěnou uprostřed [1]. Tvar šestiúhelníku je nejblíže ideálnímu pokrytí všesměrovou anténou – kruhu bez nepraktických překryvů. Každá buňka využívá sady frekvencí, která musí být odlišná od sady frekvencí sousední buňky. Navíc pokud není splněna podmínka, že každá frekvence ze sady frekvencí je znovu použita až ve vzdálenosti D (viz obrázek 2.1), může dojít k rušení ovlivňující kvalitu nabízených služeb síťovým operátorem [2]. Při přechodu mobilní stanice do sousední buňky je potřeba zajistit rychlou a automatickou změnu vysílací a přijímací frekvence (kanálu). Tento proces se nazývá *handover*.



Obrázek 2.1: Zjednodušený tvar buněk, skupina ($k = 7$), vzdálenosti D a R

Aby bylo přidělené kmitočtové pásmo co nejefektivněji využito a podmínka vzdálenosti znovupoužitelné frekvence byla dodržena, jsou buňky rozmístěny v prostoru ve skupinách. Každá taková skupina (*cluster*) má rozmístěné buňky (a odpovídající frekvence) podle stejného vzorce.

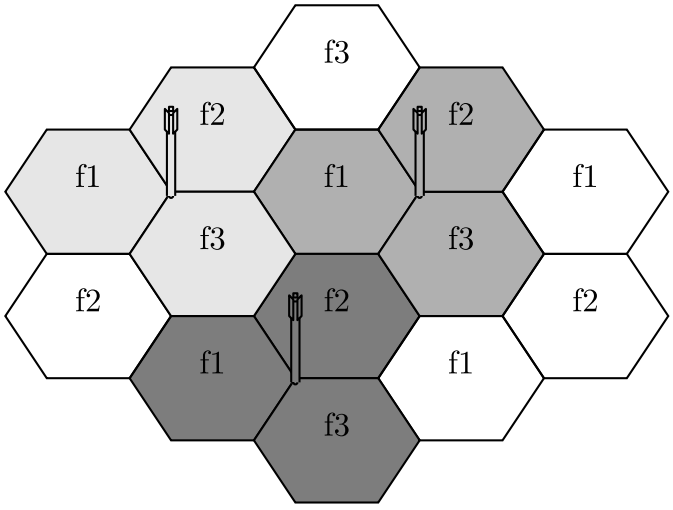
Velikost skupiny buněk je charakterizována počtem buněk k a poloměrem buňky R. Z šestiúhelníkového modelu buněk lze geometricky odvodit vzorec pro výpočet vzdálenosti znovuvyužití frekvence D [3]:

$$D = R\sqrt{3k} \tag{2.1}$$

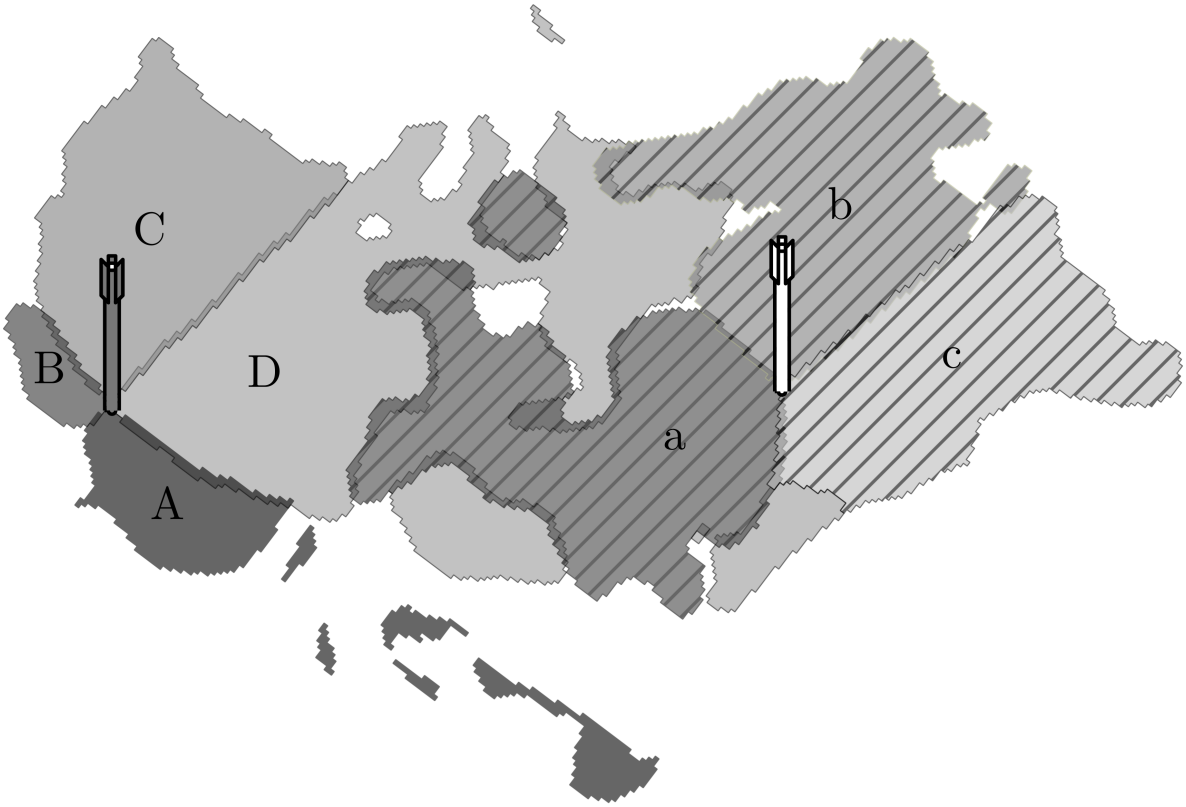
Na obrázku 2.1 jsou tyto veličiny zobrazeny. Ze vzorce 2.1 vyplývá, že čím je velikost skupiny (hodnota R a k) větší, tím je větší vzdálenost znovuvyužití frekvence D. S nárůstem počtu buněk k se zmenšuje počet dostupných kanálů a tím i počet obsluhovatelných mobilních stanic. Ve skupině buněk je typicky využit celý frekvenční rozsah přidělený síťovému operátorovi. Žádný kmitočet není použit více než jednou.

Velikost buněk se nepřímo úměrně přizpůsobuje místní hustotě provozu. Použití menšího poloměru buněk má za následek potřebu většího množství buněk a tedy i síťové infrastruktury k pokrytí určitého území. Každou buňku lze typicky rozdělit na tři nebo šest částí (sektorů). Všesměrové antény jsou nahrazeny sektorovými s vyřazovacím úhlem 120° (tři sektory) nebo 60° (šest sektorů). Na obrázku 2.2 je příklad rozdělení skupiny buněk ($k = 3$) na tři sektory s využitím třísektorové základnové stanice.

Na obrázku 2.3 je vidět reálné pokrytí území. Šedá čtyřsektorová základnová stanice spravuje buňky bez šrafování – A, B, C a D. Třísektorová bílá základnová stanice spravuje buňky šrafované plnými čarami – a, b, c. Na hranicích buněk lze pozorovat jejich částečný překryv. Velikost buněk je různorodá, nespojitost je způsobena terénním reliéfem a zástavbou.



Obrázek 2.2: Sektorizace skupiny buněk ($k = 3$) a znovuvyužití frekvencí f1, f2, f3



Obrázek 2.3: Reálný tvar buněk v mobilní síti

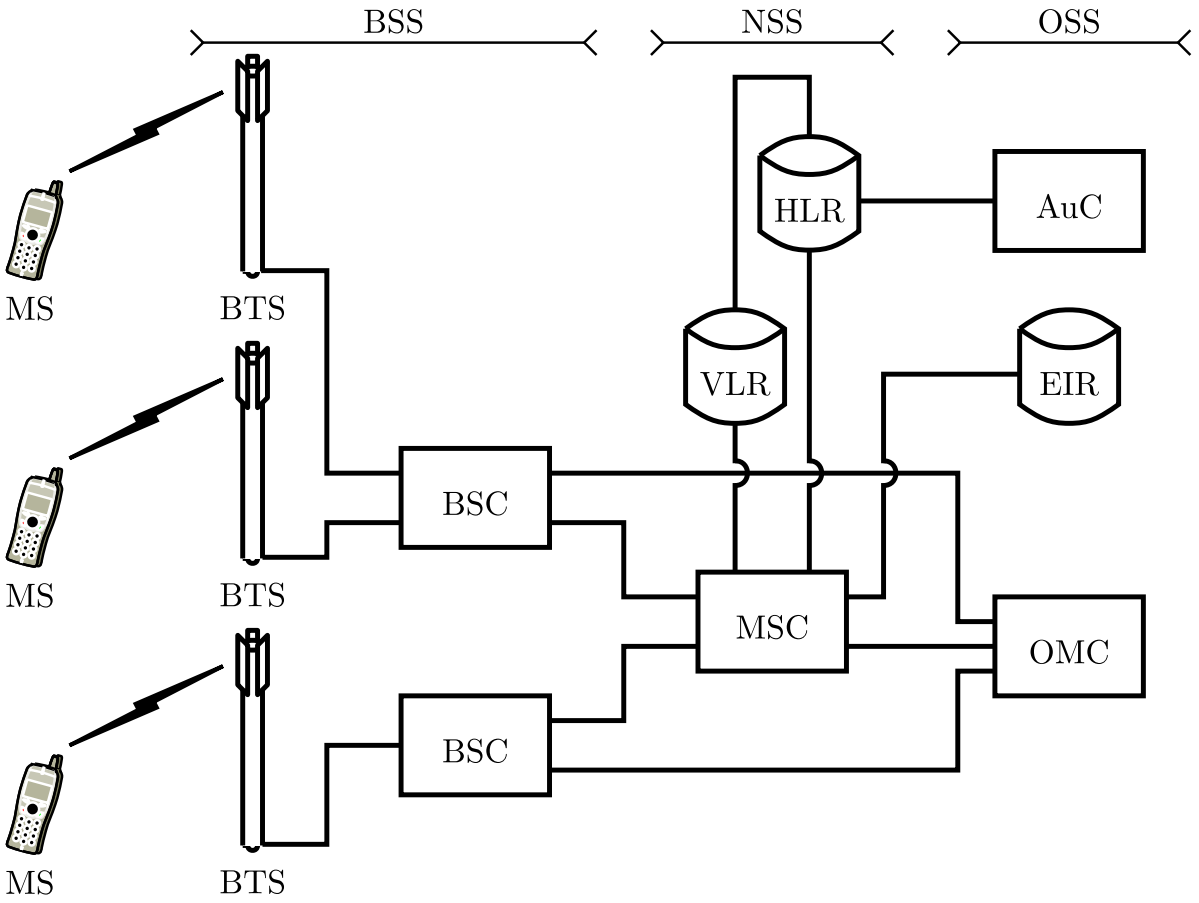
2.2 Architektura GSM sítí

2.2.1 Mobilní stanice

Mobilní stanice (MS – *Mobile Station*) je zařízení, které uživatel používá k bezdrátovému přístupu k síti. Součástí každé mobilní stanice je karta SIM (*Subsriber Identity Module*). Na SIM kartě jsou uloženy následující neměnitelné údaje [1] [3] [4]:

- typ SIM karty,
- sériové číslo SIM karty,
- seznam podporovaných služeb,
- IMSI (*International Mobile Subscriber Identity*),
- PIN (*Personal Identity Number*) – čtyř až osmimístný uživatelsky měnitelný kód pro aktivaci SIM karty,
- PUK (*PIN Unblocking Key*) – osmimístný neměnitelný kód pro odblokování SIM karty v případě opakovaného chybného zadání PIN,
- autentizační algoritmus A3 a autentizační klíč K_i,
- algoritmus A8 generující klíč K_c pro hlasovou komunikaci

a následující proměnné údaje:



Obrázek 2.4: Architektura GSM sítí [1]

- šifrovací klíč K_c a související sekvenční číslo,
- seznam používaných nosných frekvencí v síti,
- seznam blokovaných sítí,
- TMSI (*Temporary Mobile Subscriber Identity*),
- LAI (*Location Area Identitification*),
- parametr pro časování lokalizačních procedur.

Dále zde může mít uživatel uloženy kontakty, zprávy SMS (*Short Message Service*) nebo informace o účtování.

SIM karta je přenositelná mezi stanicemi. Díky tomu lze rozlišit pohyb uživatele od mobilního zařízení. K identifikaci uživatelů a mobilních stanic slouží čísla [1] [3] [5]:

- IMEI (*International Mobile station Equipment Identity*),
- IMSI (*International Mobile Subscriber Identity*),
- TMSI (*Temporary Mobile Subscriber Identity*),
- LAI (*Location Area Identitification*),
- MSISDN (*Mobile Station international ISDN number*),
- MSRN (*Mobile Station Roaming Number*).

IMEI Patnáctimístné číslo IMEI je unikátní identifikátor mobilního zařízení. Je hierarchicky členěno. Začíná osmimístným kódem TAC (*Type Allocation Code*), podle kterého lze identifikovat výrobce a typ mobilního zařízení. Všechna čísla IMEI zařízení registrovaných v síti jsou uložena v EIR (*Equipment Identity Register*).

IMSI IMSI je celosvětově unikátní identifikátor uživatele uložený na SIM kartě. Číslo může být až patnáct číslic dlouhé. Obsahuje kód země (MCC – *Mobile Country Code*) a kód mobilní sítě (MNC – *Mobile Network Code*).

TMSI TMSI je číslo lokálního charakteru přidělované jednotkou VLR (*Visitor Location Register*). Mobilní operátor si může definovat vlastní formát čísla, nesmí však přesáhnout délku 32 bitů. Používá se pro utajení uživatele při možném poslechu komunikace na rádiovém rozhraní. Unikátní je pouze na území obsluhované VLR. Dvojici TMSI + LAI lze použít jako unikátní identifikátor uživatele místo IMSI.

LAI Skupina základnových stanic sdílející jedno BSC (*Base Station Controller*) pokrývá území nazývané *location area* (LA). Každé takové území je unikátně identifikováno číslem LAI. To se periodicky vysílá všem mobilním stanicím po BCCH (*Broadcast Control Channel*). Pokud MS zjistí změnu LAI, zažádá o aktualizaci své polohy v HLR (*Home Location Register*) a VLR. LAI je složeno z kódu země (MCC), kódu mobilní sítě (MNC) a kódu území (LAC – *Location Area Code*).

MSISDN Mobilní číslo MSISDN je přidělováno uživateli mobilní sítě. Každá služba může mít vlastní číslo. Je uloženo v HLR a na rozdíl od IMSI není utajováno. Skládá se z kódu země (CC – *Country Code*), kódu mobilní sítě (NDC – *National Destination Code*) a čísla uživatele.

MSRN MSRN je dočasné číslo přidělované VLR takovým způsobem, aby bylo zřejmé MSC (*Mobile Switching Center*), pod kterým se MS nachází. Může být uloženo v HLR nebo vyžádáno v případě směrování služby k mobilní stanici. Je složeno ze stejných částí jako MSISDN.

2.2.2 Systém základnových stanic

Systém základnových stanic (BSS – *Base Station Subsystem*) je funkčně zodpovědný za domluvu protokolů a přenos signálů přes rádiové rozhraní.

Základnová stanice

Základnová stanice (BTS – *Base Transceiver Station*) spravuje rozhraní mezi mobilními stanicemi a sítí. Její hlavní funkcí je vysílání a příjem rádiových signálů, jejich

kódování, dekódování (pomocí TRAU – *Transcoding and Rate Adaptation Unit*) a zabezpečení.

Každá základnová stanice periodicky vysílá svůj BSIC (*Base Station Identity Code*) složený z [3]:

- NCC (*Network Color Code*) – tříbitový kód sítě,
- BCC (*BTS Color Code*) – tříbitový kód BTS.

Sousední sítě musí mít rozdílné NCC a sousední BTS rozdílné BCC.

Každá buňka je označena pomocí CI (*Cell Identity*), které je unikátní uvnitř LA. Maximální délka CI je 16 bitů. První či poslední číslice může označovat sektor. Skupina LAI + CI se nazývá CGI (*Cell Global Identification*) a je celosvětově unikátní identifikací buňky.

Ovladač základnové stanice

Řídící stanice (BSC – *Base Station Controller*) se stará o přidělování a uvolňování rádiových komunikačních kanálů. Také udržuje spojení v případě *handoveru*, řídí úrovně vysílacích výkonů a signalizaci do MSC. Dále časuje vysílání závisle na vzdálenosti MS od BTS tak, aby se vešlo do přiděleného časového okna (*timeslot*) – tento princip se nazývá *timing advance* [6].

Jedno BSC obvykle ovládá desítky základnových stanic. Skupiny BTS pod jedním BSC jsou logicky seskupovány do LA. Pod jedním BSC může být více LA.

2.2.3 Síťový spojovací subsystém

Síťový spojovací subsystém (NSS – *Network and Switching Subsystem*) spojuje buňkovou rádiovou síť s ostatními sítěmi. Sestavuje, směruje a řídí hovory mezi operátory. Spravuje informace o uživatelích a jejich zařízeních v síti.

Ústředna veřejné mobilní sítě

Ústředna veřejné mobilní sítě (MSC – *Mobile Switching Centre*) je základním prvkem NSS. Spravuje všechny spojení s přepojováním okruhů. To je součástí protokolu řízení hovoru (*Call Control Protocol*), který se stará o [6]:

- registraci mobilních stanic při jejich zapnutí,
- sestavení a směrování volání mezi účastníky,
- přeposílání SMS zpráv.

MSC také zajišťuje mobilitu uživatelů – jejich autentizaci, aktualizaci polohy (*location update*) a předání spojení (*handover*).

Domovský registr

Domovský registr (HLR – *Home Location Registr*) je databáze uživatelů. Jsou zde uloženy trvalé i dočasné informace. U každého záznamu (uživatele) je uvedeno IMSI, MSISDN, odebírané služby, typ zařízení, autentizační data, MSRN, adresa VLR, adresa MSC a LA. Obvykle je v celé síti pouze jeden HLR [3].

Návštěvnický registr

Návštěvnický registr (VLR – *Visitor Location Registr*) je databáze uživatelů, kteří jsou obsluhováni přidruženým MSC. Z důvodu snížení množství signalizačních dat v síti je zde mj. kopie záznamů z HLR. U každého záznamu je uvedeno IMSI, MSISDN, typ zařízení, autentizační data, MSRN, TMSI a LAI. Obvykle je u každé MSC jeden VLR [3]. Při připojení uživatele pod jiné MSC je záznam HLR zkopírován do nového VLR a následně vymazán ve starém.

2.2.4 Operační subsystém

Operační subsystém (OSS – *Operation Subsystem*) je část GSM systému umožnující [1]:

- správu uživatelů – autentizace uživatele podle informací z HLR a poskytnutí dohodnutých služeb,
- správu uživatelských dat – bezpečnostní uživatelská data jsou uložena v AuC (*Authentication Centre*),
- účtování,
- obsluhu a údržbu sítě – standardizováno ITU-T (*International Telecommunication Union – Telecommunication standardization sector*) jako TMN (*Telecommunications Management Network*): řízení obchodní činnosti, řízení služeb, řízení sítě a řízení síťových prvků,
- správu mobilních zařízení.

Součástí OSS je OMC (*Operation and Maintenance Centre*), AuC (*Authentication Centre*) a EIR (*Equipment Identity Register*). OMC ovládá a monitoruje ostatní síťové prvky pomocí TMN, aby byla zajištěna co nejlepší funkčnost sítě. V AuC jsou uložena uživatelská data použita k ochraně uživatelovy totožnosti a zabezpečení jeho komunikace. Je zde uložena kopie autentizačního klíče K_i, který je použit v algoritmu A3. EIR je databáze obsahující seznamy povolených a blokovaných IMEI. Seznam blokovaných zařízení je sdílen mezi mobilními operátory a obsahuje IMEI odcizených zařízení.

2.3 Správa mobility

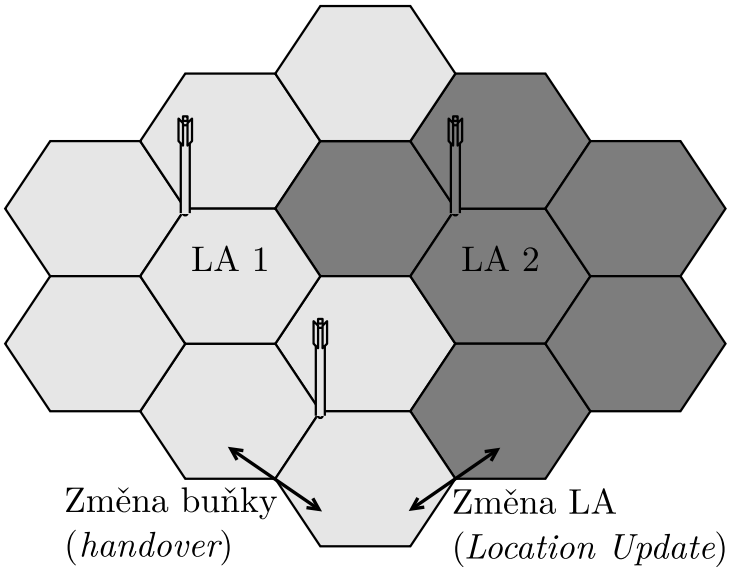
Správa mobility (MM – *Mobility Management*) je soubor procedur na třetí vrstvě architektury signalizačních protokolů. Veškerá činnost MM je založena na spolupráci MS

a MSC. Správa mobility je nutná pro poskytnutí všech služeb mobilního operátora volně se pohybujícímu uživateli. Mezi činnosti MM patří [1] [3]:

- přiřazení TMSI – TMSI je platné pouze na území obsluhované VLR a napomáhá utajení uživatelovy identity,
- identifikace MS – MS je vyzvána, aby síti sdělila své IMSI, TMSI nebo IMEI,
- autentizace MS,
- připojení a odpojení IMSI (*IMSI attach*, *IMSI detach*) – v případě zapnutí nebo vypnutí mobilního zařízení, anebo připojení nebo odpojení SIM karty,
- aktualizace polohy (*location update*) – periodicky nebo při změně přijímaného LAI na kanálu BCCH.

2.3.1 Skupiny buněk

Buňky jsou seskupovány do oblastí nazývaných LA (*Location Area*) jako v příkladu na obrázku 2.5. Kód, podle kterého jsou identifikovány – LAI, je periodicky vysílán na kanálu BCCH. Vysílaný kód je porovnáván s LAI uloženým v MS. V případě rozdílných hodnot stanice pozná, že přešla do buňky patřící do jiné LA a informuje o tom síť pomocí procedury aktualizace polohy.



Obrázek 2.5: Seskupování buněk do LA

Díky tomu, že se mobilní stanice nemusí informovat síť při změně buňky ale až při změně LA, je sníženo množství přenášených signalizačních zpráv, což má za následek snížení spotřeby energie MS [6]. Velikost oblastí závisí na volbě mobilního operátora. V jedné LA se typicky nachází desítky až stovky buněk.

2.3.2 Aktualizace polohy mobilní stanice

Pro využívání služeb poskytovaných mobilní sítí se musí každá mobilní stanice v síti nejprve registrovat. Registraci lze provést nejen v domácí síti (v síti operátora, se kterým

má uživatel uzavřenou smlouvu), ale i v cizích sítích, pokud s nimi má domácí operátor uzavřenou dohodu. K registraci dochází pouze při změně sítě, tj. když žádný VLR v síti zatím nepřiřadil mobilní stanici TMSI. To je MS přiřazeno pomocí procedury *Location Update* (LU) [3].

Registraci zahajuje MS posláním zprávy *Location Update Request*, ve které uvádí své IMSI a LAI [7]. MSC vyzve VLR zprávou *Update Location Area*, aby MS zaregistroval. K tomu VLR potřebuje mobilní stanici identifikovat a autentizovat. Kontaktuje proto HLR, který dále komunikuje s AuC. Úspěšnou autentizací pokračuje přiřazení MSRN uživateli. To je společně s LAI uloženo v HLR. Nově vygenerované TMSI je v šifrované podobě zasláno MS. VLR potvrzuje úspěšnou registraci zprávou *LocUpdate Accept*. MS potvrzuje přijetí TMSI pomocí zprávy *TMSI Reallocation Complete*. K aktualizaci polohy touto procedurou dochází v případě, kdy se MS ocitne v nové oblasti LA beze změny VLR [3].

Periodický LU

K aktualizaci polohy dochází i periodicky po vypršení intervalu vysílaném na kanálu BCCH. Tento interval je uveden v osmibitovém parametru *T3212* s jednotkou 6 minut. Jeho maximální hodnota je tedy $255 \cdot 6$ minut $= 1\,530$ minut $= 25,5$ hodin. Jeho doporučená hodnota je 180 minut [8]. Může být nastaven různě podle lokality (malý/velký provoz), času (den/noc, všední dny / víkend), generace sítě (větší u 2. generace) atd.

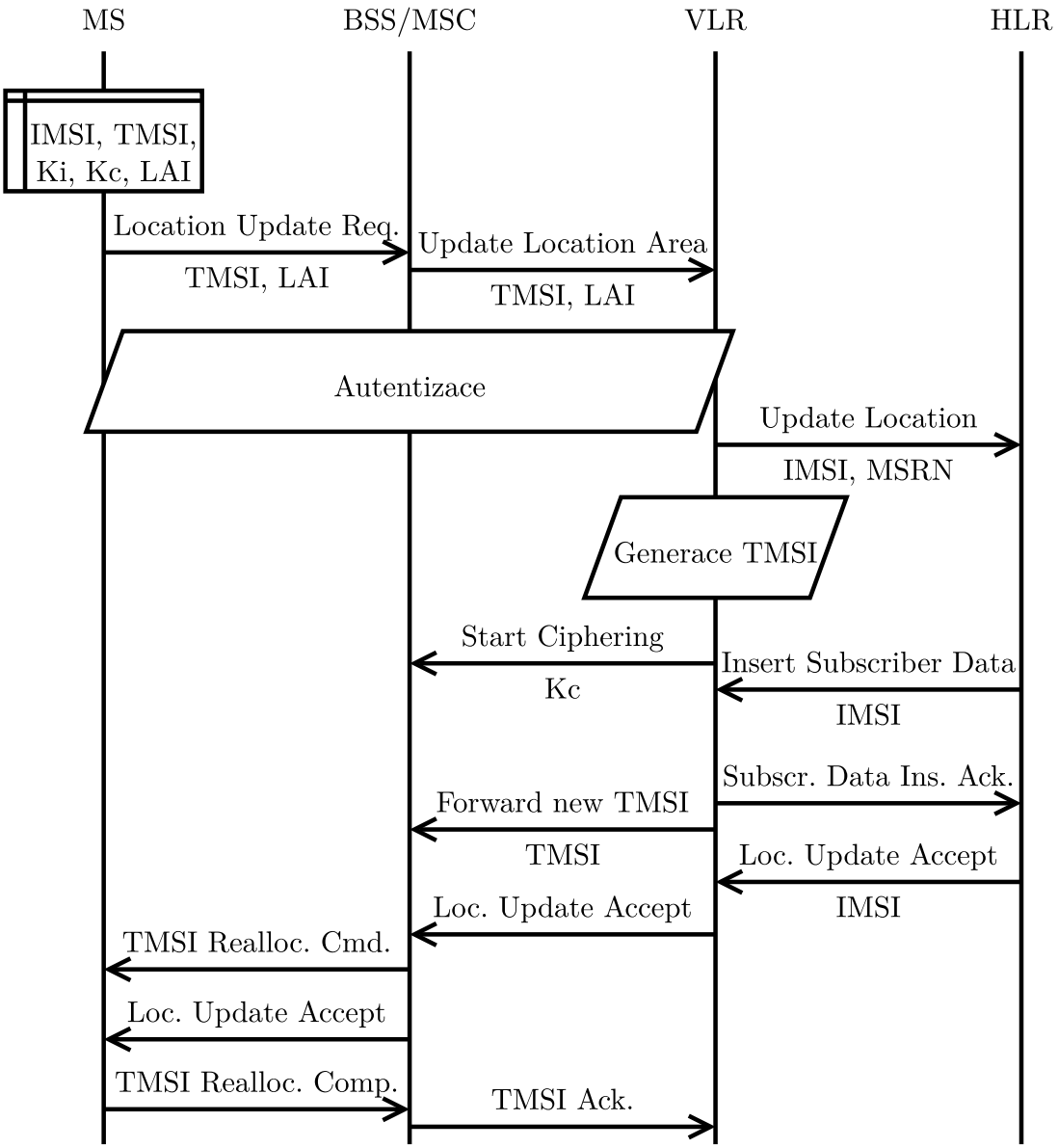
Na rozdíl od registrace MS zná své TMSI, které je v těchto případech použito v proceduře místo IMSI. TMSI je uloženo v non-volatilní paměti SIM karty, proto si ho mobilní stanice pamatuje i po vypnutí zařízení. Na obrázku 2.6 je zjednodušeně znázorněn sled zpráv při periodickém typu procedury *Location Update*.[1]

Registraci opět zahajuje MS zprávou *Location Update Request* pro MSC, která obsahuje její současný TMSI a LAI [3]. Tyto dva identifikátory jsou dále přeposlány do VLR. Po autentizaci mobilní stanice VLR informuje HLR zprávou *Update Location*, ve které je odpovídající IMSI a MSRN. Po vygenerování nového TMSI návštěvnickým registrem se kanál zabezpečí využitím klíče Kc pro šifrování a nové TMSI je přes MSC přeposláno k MS. Z HLR postupně až do MS je aktualizace polohy potvrzena zprávou *Location Update Accept*. Proceduru ukončuje MS zprávou *TMSI Reallocation Complete* a z MSC do VLR je poslána zpráva *TMSI Ack*. Výkonem této série zpráv dochází nejen k aktualizaci polohy, ale i TMSI mobilní stanice.

Normální LU

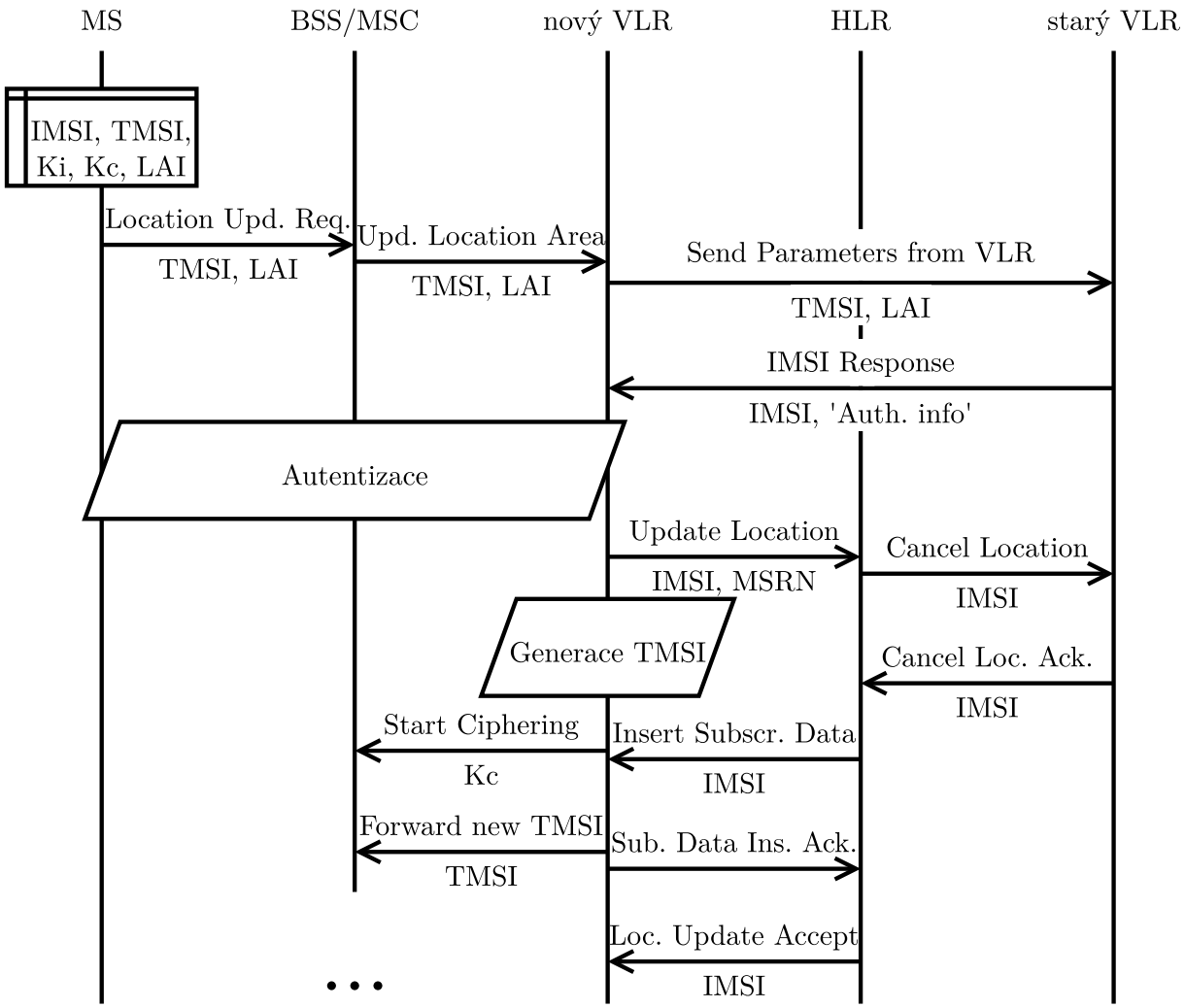
V případě, že je aktualizace polohy MS vyvolána změnou LA i návštěvnického registru, dochází k normálnímu typu *Location Update*. Proceduru znovu zahajuje MS vysláním zprávy *Location Update Request* k MSC. To pošle TMSI a LAI z předchozí zprávy pomocí *Update Location Area* do nového VLR.

[1] Názvy zpráv jsou pro lepší názornost a vyhledatelnost ponechány v anglickém jazyce se snahou o kompatibilitu se standardy skupiny 3GPP.



Obrázek 2.6: Zjednodušený sled zpráv periodické procedury *Location Update* [3]

Na rozdíl od periodického *Location Update* vyžaduje nový VLR po starém VLR identifikační informace a data pro zabezpečení komunikace dané MS zprávou *Send Parameters from VLR*. V ní je uvedeno současné TMSI stanice a LAI. Požadovaná data jsou poslána ve zprávě *IMSI Response*. Poté, co je možné vygenerovat nové TMSI, zruší HLR záznam ve starém VLR zprávou *Cancel Location* [3]. Smazání záznamu je potvrzeno zprávou *Cancel Location Ack*. Poté již následují stejné zprávy jako u periodického typu LU (viz obrázek 2.7).



Obrázek 2.7: Zjednodušený sled zpráv normální procedury *Location Update* [3]

3 Administrativní členění ČR

3.1 Evropské systémy územního členění

Systém administrativního členění České republiky je od 1. ledna 2008 založen na standardu statistického úřadu Evropské unie (Eurostat) – NUTS (z francouzského *Nomenclature des unités territoriales statistiques*) a LAU (*Local Administrative Unit*) [9].

Podle klasifikace NUTS je území každé členské země Evropské unie rozděleno Eurostatem do tří úrovní: NUTS1 (území), NUTS2 (oblast) a NUTS3 (kraj). Každá část území je označena unikátním kódem závislým na úrovni, kterým je pak jednoznačně identifikována v rámci celé Evropské unie. Jak je vidět v tabulce 3.1, kódování je hierarchicky členěno. Z kódu NUTS3 lze vyčíst kód NUTS2 i NUTS1. Podle poslední číslice kódu NUTS3 lze určit, zdali se NUTS2 dělí na jeden či více celků NUTS3.

Detailnější dělení územních celků v zemích Evropské unie se řídí standardem LAU. Jeho součástí jsou dvě úrovně – LAU1 (okresy), která se v osmi státech nepoužívá (je rovna NUTS3), a LAU2 (obce), která je definována ve všech dvaceti osmi státech [11]. Tento systém je určen především pro potřeby statistiky regionů. Ukázka rozdělení České republiky na celky LAU1 je na obrázku 3.1.

V České republice kódy úrovně LAU1 logicky navazují na kódy NUTS3. Například kód LAU1 okresu Beroun je CZ0202, kód NUTS3 Středočeského kraje je CZ020. Kódy LAU2 mají formát šesticiferných čísel s počáteční číslicí pět. Například kód LAU2 obce Beroun je 531057. V roce 2015 bylo v České republice evidováno 77 okresů a 6 253 obcí [11].

3.2 Soustava územních prvků ČR

„Soustava územních prvků a územně evidenčních jednotek vychází ze zákona o územním členění státu, který stanovuje základní členění státu na kraje, okresy a obce a vojenské újezdy (viz zákon č. 320/2002, část 114., čl. CXIV). Statuty hlavního města Prahy a některých statutárních měst stanovují členění na městské obvody nebo městské části. Na základě rozhodnutí jednotlivých obcí se pak mění členění obcí na části obce. Na územní členění státu navazuje technické členění dle katastrálního zákona, a dále sídelní a statistické členění dle zákona o státní statistické službě a dle zákona o sčítání lidu, domů a bytů [12].“

Kód	NUTS1 (území)	NUTS2 (oblast)	NUTS3 (kraj)
CZ0	Česká republika		
CZ01		Praha	
CZ010			Hl. m. Praha
CZ02		Střední Čechy	
CZ020			Středočeský kraj
CZ03		Jihozápad	
CZ031			Jihočeský kraj
CZ032			Plzeňský kraj
CZ04		Severozápad	
CZ041			Karlovarský kraj
CZ042			Ústecký kraj
CZ05		Severovýchod	
CZ051			Liberecký kraj
CZ052			Královéhradecký kraj
CZ053			Pardubický kraj
CZ06		Jihovýchod	
CZ063			Vysočina
CZ064			Jihomoravský kraj
CZ07		Střední Morava	
CZ071			Olomoucký kraj
CZ072			Zlínský kraj
CZ08		Moravskoslezsko	
CZ080			Moravskoslezský kraj

Tabulka 3.1: Administrativní členění NUTS [9]

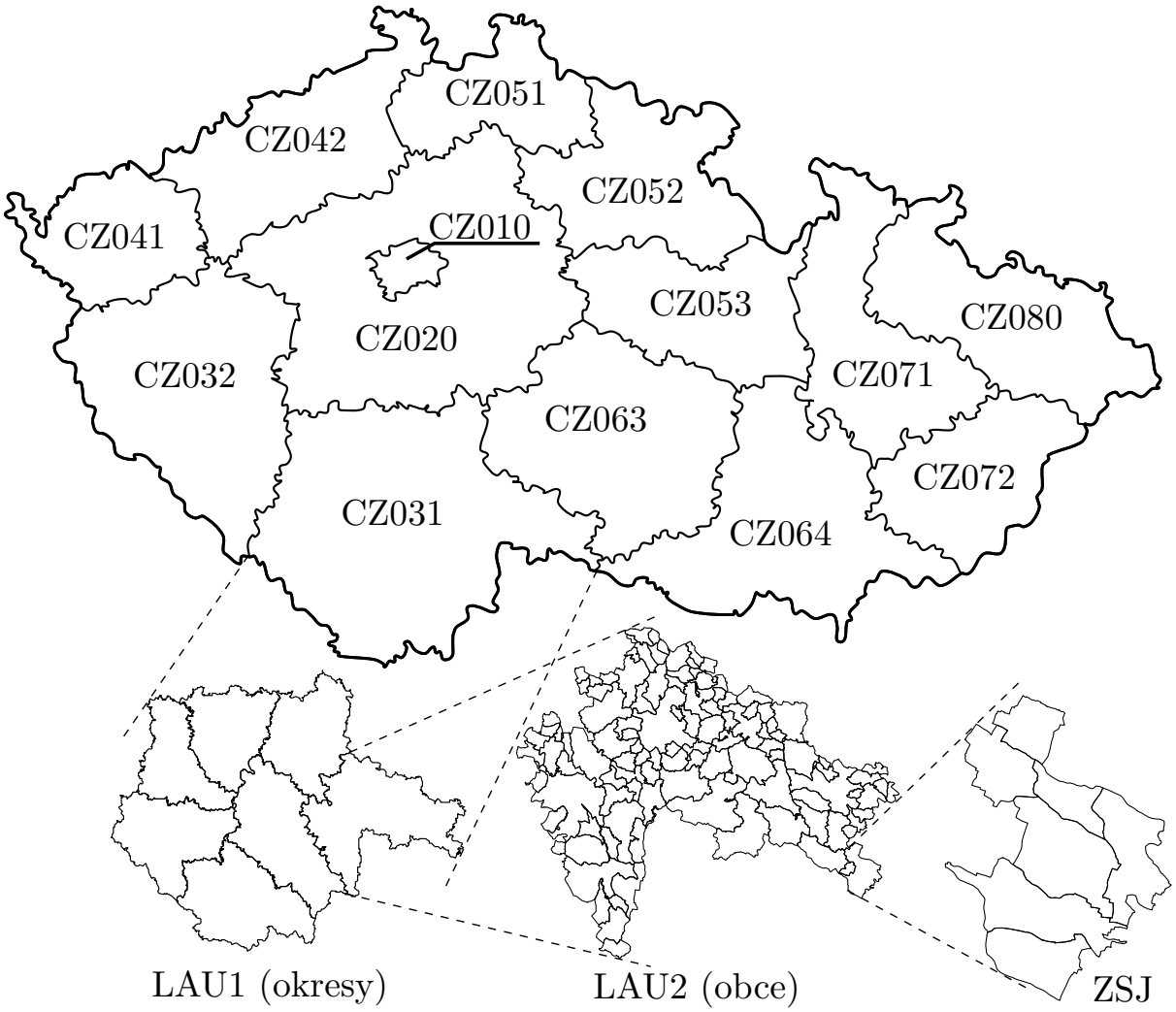
Na obrázku 3.2 je zobrazeno schéma soustavy územních prvků a územně evidenčních jednotek využitých v registru sčítacích obvodů a budov Českého statistického úřadu.

3.2.1 Obec

„Obec je základním územním samosprávným společenstvím občanů. Tvoří územní celek, který je vymezen hranicí území obce [13].“ Definice obce a pravidla jejich vzniku jsou určena zákonem o obcích č. 128/2000 Sb.

Vojenský újezd je v soustavě územních prvků (viz obrázek 3.2) na stejné úrovni jako obec. Vojenský újezd je část území sloužící k výcviku ozbrojených sil a k zajištění obrany státu.

Každý městys, město nebo statutární město je obcí. Za obec se také považuje Hlavní město Praha. Obec je určena názvem a číselným kódem. Název obce nemusí být unikátní ani v rámci stejného okresu (přímo nadřazeného územního prvku). Správcem názvů je



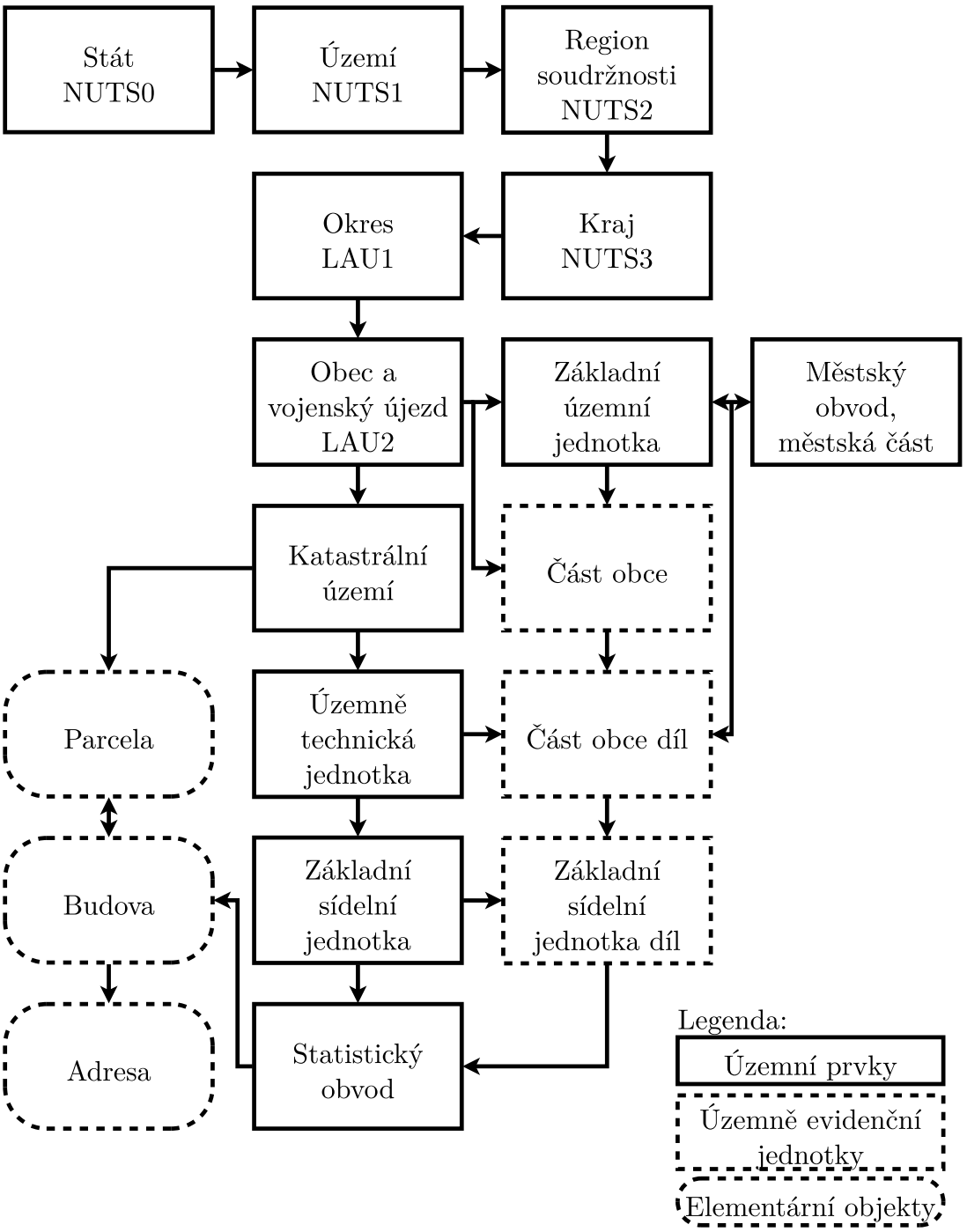
Obrázek 3.1: Mapa ČR – administrativní členění NUTS3 (kraje) [10]; dělení Jihočeského kraje na LAU1 (okresy), okresu Jindřichův Hradec na LAU2 (obce) a obce Dešná na ZSJ

Ministerstvo vnitra ČR a jejich garantem je Český statistický úřad. Kód obce je unikátní na celém území České republiky. Nabývá hodnot v oboru 500011–599999 [13]. Jejich správcem i garantem je Český statistický úřad.

K 1. lednu 2015 bylo v České republice 6 253 obcí [11]. V systému územního dělení LAU jsou obce na úrovni LAU2.

3.2.2 Základní sídelní jednotka

Základní sídelní jednotka je územní celek určený pro prostorovou identifikaci a sledování sociálně ekonomických a územně technických jevů přímo vázaných na osídlení, zejména výsledků Sčítání lidu, domů a bytů [14]. Každá obec je bezezbytku rozdělena soustavou základních sídelních jednotek.



Obrázek 3.2: Schéma soustavy územních prvků a územně evidenčních jednotek aplikované v registru sčítacích obvodů a budov [12]

Každá základní sídelní jednotka má definován tzv. charakter jako ohodnocení převažujícího způsobu využití území a struktury zástavby. Charakterem základních sídelních jednotek může být [15]:

- obytná plocha v kompaktní zástavbě,
- odloučená obytná plocha včetně přilehlých zemědělských ploch,

- průmyslový areál,
- dopravní areál,
- areál občanské vybavenosti,
- ostatní účelová plocha,
- rezervní plocha,
- rekreační plocha,
- zemědělská plocha,
- lesní plocha,
- venkovská smíšená lokalita,
- venkovská lokalita bez zástavby.

Základní sídelní jednotka je definována v zákoně č. 230/2006 Sb. Jejich garantem a správcem včetně kódů je od 10. března 2014 Český statistický úřad. Každá základní sídelní jednotka je určena názvem, který je měnlivý a unikátní v rámci obce, do které patří, a kódem, který je unikátní a neměnný na území celé České republiky. Kódy ZSJ (*základních sídelní jednotek*) mohou nabývat až šesticiferných hodnot v oboru 0 až 199999 a 300010 až 399999 [14]. K 7. květnu 2016 bylo v České republice evidováno 22 455 ZSJ [16].

Územní prvek		Počet
NUTS1	Území	1
NUTS2	Oblast	8
NUTS3	Kraj	14
LAU1	Okres	77
LAU2	Obec	6 253
	ZSJ	22 455

Tabulka 3.2: Počty územních prvků v České republice

3.3 Sčítání lidu, domů a bytů

Sčítání lidu, domů a bytů (SLDB) je proces, kterým se dotazníkovou metodou získávají informace o obyvatelstvu a jejich nemovitostech [17]. Sčítání se provádí pravidelně každých deset let. Poslední se uskutečnilo v březnu 2011 a rozhodným okamžikem byla půlnoc z 25. na 26. března. Celý proces byl zajištěn Českým statistickým úřadem.

Formuláře bylo možné pohodlně vyplnit a odeslat elektronicky na Internetu. Sčítání poprvé probíhalo ve všech zemích evropské unie ve stejný rok najednou. Náklady jsou přibližně 250 Kč na osobu [18].

Určité formy sčítání lidu byly na našem území zaznamenány již ve středověku. První moderní (v rozsahu a kvalitě informací) sčítání zde proběhlo v roce 1869 (v Rakousko-Uhersku) a od té doby bylo vynecháno pouze jednou z důvodu probíhající druhé světové války [18].

Data ze SLDB jsem využil především k porovnání a ke kontrole vypočítaných dat z mobilní sítě. Mezi nejzajímavější patří informace o počtu obyvatelstva na úrovni *Základní sídelní jednotka díl* v soustavě územních prvků a o počtu obyvatelstva vyjíždějících do zaměstnaní a škol na úrovni *Základní sídelní jednotka díl*. Také jsem použil dokument obsahující počty obyvatel vyjíždějících do zaměstnání a škol mezi jednotlivými územními prvky na úrovni obcí.

Naměřená data se mohou od SLDB lišit z důvodu jeho staří. Sčítací listy rozlišovaly mezi místem trvalého pobytu a bydlištěm. I přesto si myslím, že ne všichni uvedli místo svého typického pobytu jako své bydliště. Mnoho lidí může často i výjimečně přespávat u svých partnerů, přátel či někde jinde. Problematika počtu lidí dojíždějících do zaměstnání a škol je na tom podobně. Své vyjížďky nemuseli všichni ve sčítacích formulářích správně uvést. Mnoho občanů nejezdí denně pracovat na stejné místo.

Ukazatel	Počet
Obyvatelstvo celkem	10 463 560
Obyvatelstvo muži	5 109 766
Obyvatelstvo ženy	5 326 794
Obydlené domy	1 800 075
Obydlené byty	4 104 635

Tabulka 3.3: Základní výsledky Sčítání lidu, domů a bytů 2011 pro celou Českou republiku [19]

4 Analýza velkých dat

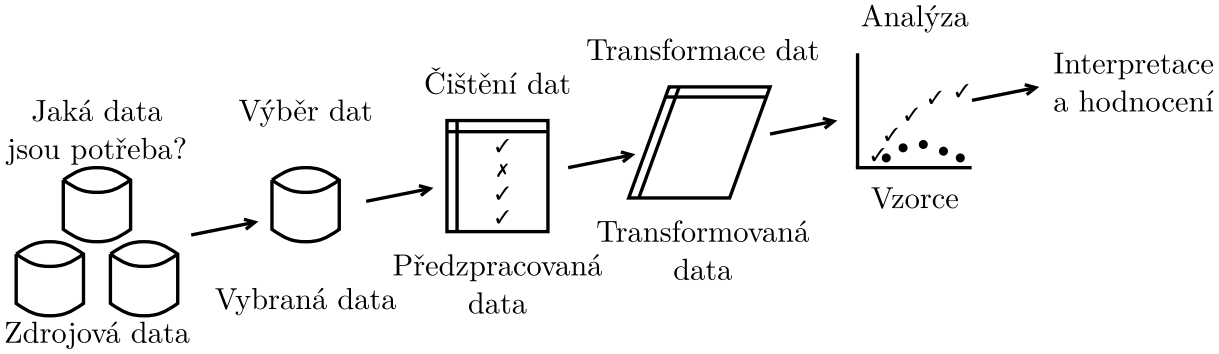
Analýza informací z velkých dat (*big data* – někde přeloženo do češtiny jako veledata) přináší řadu výzev z hlediska [20]:

- objemu dat – řádově jednotky terabytů až petabytů,
- různorodosti dat – různé struktury (surový text → logy → tabulky),
- platnosti dat – data mají určitou hodnotu pouze po omezený časový interval.

Tyto tři problémy, jimiž jsou velká data obvykle charakterizována, představují překážku při použití tradičních analytických nástrojů.

4.1 Procesní model

Pro efektivní analýzu nejen velkých dat je vhodné dodržovat určitý procesní model [21]. Příklad takového modelu je uveden na obrázku 4.1. Prvně je potřeba definice obchodního problému, který bude analýzou řešen. Dále se identifikují všechna data, která mohou mít podíl na řešení problému. Tato data se pročistí např. od duplicit, extrémů nebo chybějících hodnot. Poté se data mohou transformovat – definice skupin a jejich agregace (např. geografická agregace), změna kódování dat atd. Volba analytického modelu má poslední dopad na výstupní data, která jsou vhodná k interpretaci a hodnocení člověkem.



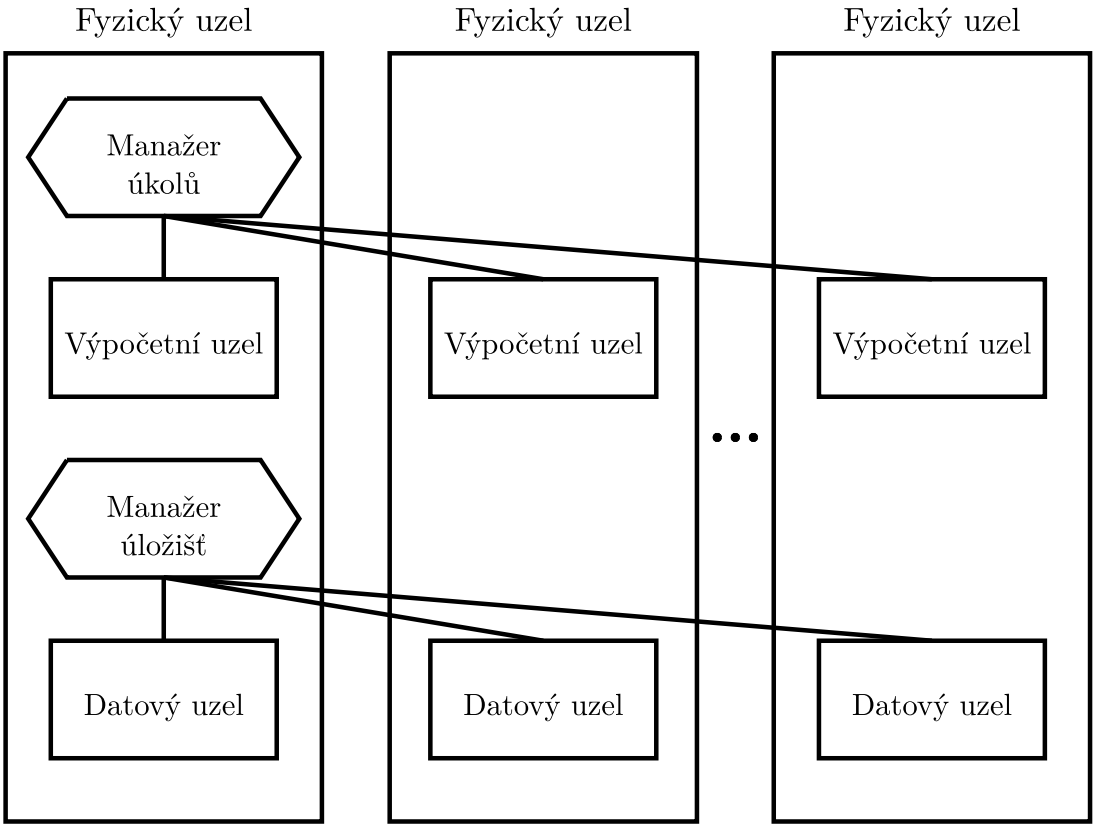
Obrázek 4.1: Model analýzy dat [21]

4.2 Technické prostředky

Analýza velkých dat má speciální požadavky na použitý hardware i software. Mezi nejčastější patří [22]:

- **Výpočetní platforma**, která je přímo optimalizována pro rozsáhlou analýzu. Často se skládá z několika vícejádrových výpočetních uzlů, které jsou připojeny vysokorychlostní linkou k paměti a diskovým polím.
- **Škálovatelné úložiště** k úschově velkého objemu dat.
- **Prostředí pro správu dat** využívající vysokého stupně paralelismu k přístupu do úložiště.
- **Prostředí pro vývoj aplikací**, které umožňuje psaní, spouštění (včetně pravidelného spouštění podle časového plánu) a testování aplikací.
- **Nástroje pro analytické modely**, které jsou lehce sestaveny a použity.
- **Nástroje pro dohled a správu** těchto prostředků.

Typická architektura platformy pro velká data na obrázku 4.2 rozděluje správu výpočetních prostředků od správy úložišť. Hlavní manažer úkolů (*job manager*) dohlíží na skupinu výpočetních uzlů a přiděluje jim výpočetní úkoly. Nezávisle na něm pracuje hlavní manažer úložišť (*storage manager*), který dohlíží na skupinu datových uzlů a rozděluje mezi ně datové sady [22].



Obrázek 4.2: Architektura platformy pro velká data [22]

4.3 Platforma Hadoop

Apache Hadoop je sadou otevřených (*open-source*) projektů, které spolu tvoří aplikační rámec (*framework*) umožnující práci s velkými daty. Mezi jeho hlavní části patří distribuovaný výpočetní aplikační rámec MapReduce a distribuovaný souborový systém HDFS (*Hadoop Distributed File System*). Jeho alternativami jsou např. BigQuery, Cluster Map Reduce, High Performance Computing Cluster nebo Hydra.

4.3.1 Komponenta MapReduce

MapReduce se používá k výkonu série operací na distribuovaných datových sadách. Data, na která je nahlíženo v párech klíč-hodnota, jsou zpracována ve dvou fázích – mapování (*map*) a redukce (*reduce*). Obecně se úloha MapReduce vykoná v těchto krocích [20]:

1. Vstupní data jsou rozdělena do mnoha částí, které jsou jednotlivě přiřazovány mapovacím úlohám (*map tasks*).
2. Mapovací úlohy jsou distribuovány mezi výpočetní uzly.
3. Každá mapovací úloha zpracuje přidělené dvojice klíč-hodnota. Výstupem jsou nové páry klíč-hodnota.
4. Nové páry jsou seřazeny podle klíče a rozděleny do částí, jejichž počet je roven počtu redukčních úloh (*reduce tasks*).
5. Redukční úlohy jsou rozděleny mezi uzly, kde zpracují přidělená data do výstupních párů klíč-hodnota.
6. Výstupní páry jsou zapsány do HDFS.

Hadoop ve své druhé verzi uvedl nový systém pro správu prostředků zvaný YARN (*Yet Another Resource Negotiator*). Jeho největší výhodou je spouštění více MapReduce aplikací najednou, které navíc mají větší pravomoc při přidělování prostředků. Aplikace jsou si také lépe vědomy rozdělení prostředků a dat mezi uzly. Díky tomu je omezen tok dat mezi uzly, které jsou tímto efektivněji využity [22].

4.3.2 Komponenta HDFS

HDFS je souborový systém optimalizovaný pro práci s velkými soubory (v průměru většími než 500 MB) [20]. Přístup k datům se řídí modelem „Zapiš jednou, čti často“ (*Write Once, Read Often*). Kvůli tomu nelze soubory měnit jinak než přidáním nových dat na konec souboru.

Každý soubor je rozdělen na bloky, které jsou rozděleny mezi uzly. Velikost bloků je definovatelná, bývá více než tisícinásobně vyšší ve srovnání s běžnými souborovými systémy [20]. Pro ochranu dat v případě poruchy hardwaru jsou vytvářeny kopie bloků, jejichž počet lze také definovat. Poruchu lze identifikovat ztrátou „heartbeat“ komunikace. Integrita dat je zajištěna pomocí kontrolních součtů. Je možné zvýšit výkon systému procesem, který migruje bloky dat k uzlům, kde je po nich zvýšená poptávka [22].

4.3.3 Další komponenty

Jak bylo popsáno v kapitole 4.2 – pro efektivní práci s velkými daty MapReduce a HDFS nestačí. Hadoop proto nabízí další komponenty, které doplňují a zjednodušují jeho základní funkcionalitu.

Jednou z komponent je centralizovaná služba **Zookeeper** pro udržování konfigurací, jmen, poskytování synchronizace a skupinových služeb. **HBase** je prostředí pro přístup k datům ve velkých tabulkách se sloupcovým rozložením. Tento typ datového rozložení podporuje kompresi.

Pro práci se strukturovanými databázovými tabulkami existuje **Hive**. Ten umožňuje organizovat data na HDFS do tabulkových struktur. Pro dotazování do těchto tabulek je vyvíjen jazyk HiveQL, který je podobný běžnému SQL (*Structured Query Language*). Hive umožňuje nativní přístup do MapReduce modelu, díky kterému lze použít vlastních mapovacích nebo redukčních funkcí uvnitř HiveQL dotazů [22].

Pig je projekt pro zjednodušení vývoje programů využívajících modelu MapReduce. Součástí Pigu je programovací jazyk Pig Latin, jehož funkce a operátory jsou programátorem použity na datové sady logicky podobně jako u SQL. Pig samostatně provádí analýzu kódu, na jejíž základě identifikuje příležitosti pro optimalizaci programu.

Nevýhodou modelu MapReduce je jeho časová náročnost v případě jednoduchých úloh. **Impala** firmy Cloudera využívá HiveQL jako programovací rozhraní, ale MapReduce nahrazuje vlastním řešením. Proto je pro jednoduché dotazy do tabulek několikanásobně rychlejší než Hive [20].

Často se může hodit zřetězení několika akcí či programů do jednoho spustitelného balíku. K tomu je vyvíjen systém Apache **Oozie**. Podporuje MapReduce aplikace, Pig, Hive, Java programy, Shell skripty a další. Série akcí, která se nazývá *workflow*, je zapsána jazykem XML (*eXtensible Markup Language*) a nemůže obsahovat žádný cyklus.

Pro zjednodušení a zpříjemnění práce s výše popsanými nástroji slouží webové grafické uživatelské rozhraní **Hue**. Pomocí Hue je možné procházet HDFS a HBase, spravovat uživatelské účty, spouštět a kontrolovat MapReduce úlohy. Také lze spouštět, kontrolovat i psát dotazy pro Hive, Pig nebo Impalu. Pomocí jednoduchého grafického rozhraní je možné sestavovat i spravovat *Oozie workflow*.

5 Zpracování dat z mobilní sítě

Pro svou činnost jsem měl k dispozici otevřenou (*open-source*) distribuci platformy Hadoop firmy Cloudera – CDH (*Cloudera's Distribution including Apache Hadoop*) ve verzi 5.4.7. Ta obsahuje všechny užitečné komponenty platformy pro velká data popsané v kapitole 4.3.3.

Zpracoval jsem dva pohledy na stanovení počtu osob v prostoru a čase v administrativně náročně členěném území. První pohled (viz kapitola 5.3) popisuje vývoj počtu osob v jednotlivých územních celcích (ZSJ a obcích) podle dne v týdnu. Druhý pohled (viz kapitola 5.4) popisuje trajektorie pohybu osob – počet osob pohybujících se mezi jednotlivými územními celky (ZSJ a obcemi). Obě analýzy jsem provedl na základě praktických zkušeností s interpretací výsledků referenčních projektů, které se zabývaly zpracováním lokalizačních a provozních dat mobilní telekomunikační sítě.

5.1 Struktura vstupních dat

Signalizační data lze transformovat do souborů, ve kterých je uveden identifikátor uživatele, datum a čas události, název (identifikátor) buňky a LAC. Identifikátor uživatele může být výstupem hašovací funkce, která má na vstupu nějaký identifikátor uživatele v síti a datum. Z toho vyplývá, že je platný a unikátní pouze v časovém intervalu 00:00:00 až 23:59:59, poté dochází k jeho změně. Kvůli tomu nelze přesný pohyb konkrétního uživatele sledovat v intervalu delším než jeden den a označení uživatelů je tím zcela anonymizováno. Mobilní stanice uživatele vyprodukuje typicky 100 až 1000 záznamů denně. Více než polovina z nich je získána navzdory pasivitě MS (záznamy jsou např. vygenerovány procedurou *Location Update*). Další mohou být následkem nějaké aktivity MS (např. hovoru, SMS apod.).

Z těchto informací mohou být vytvořeny vektory pohybů uživatelů, které udávají posloupnosti párů buňka-čas pro jednotlivé uživatele za 24 hodin. Poté lze tyto vektory mapovat na územní celky České republiky – vektory buněk jsou nahrazeny vektory ZSJ.

Tato data lze snadno agregovat do souboru, který udává vývoj počtu obyvatel v jednotlivých základních sídelních jednotkách. Může být vhodné rozřadit uživatele do ZSJ po hodinách. Také lze uživatele klasifikovat ve vztahu k danému území např. na bydlící, vyjíždějící, dojíždějící apod.

5.2 Definice pojmů

Pro přehlednost zde uvádím definice názvosloví použitého v následujících kapitolách 5.3 a 5.4.

Bydlící Osoba bydlí v ZSJ, ve které byla v měřený den evidována v nočních hodinách 00:00:00–04:10:00 a také ve večerních hodinách 19:50:00–23:59:59. Nejsou tedy započítáni lide s nestandardní pracovní dobou, která se kryje s jedním ze zmíněných časových intervalů.

Nevyjíždějící Za nevyjíždějící jsou považováni lidé, kteří po celý den neopustili místo svého bydliště na déle než 30 minut.

Vyjíždějící Vyjíždějící obyvatelé se vypočítají rozdílem bydlících a nevyjíždějících.

Dojíždějící Osoby dojíždějí do ZSJ, ve které nebydlí a stráví zde déle než 5 hodin za celý den.

Průměr (Po–Pá) Průměr všech naměřených hodnot ve dnech, které se v týdnu nacházejí mezi dvěma krajními dny (včetně) určenými typem průměru (např. pondělí, úterý, středa, čtvrtek, pátek).

Cesta Série ZSJ, kterými člověk projede včetně výchozí a konečné ZSJ (viz cíl cesty). Při sčítání unikátních cest se započítávají všechny stejné cesty (stejné série ZSJ) osoby pouze jednou.

Cíl (cesty) Poslední ZSJ na cestě, ve které osoba zůstala déle než 30 minut. Dělí se do čtyř kategorií podle doby v něm strávené:

- 30–60 minut,
- 60–180 minut,
- 180–300 minut,
- 300+ minut (více než 300 minut).

5.3 Počet obyvatel v územních prvcích

5.3.1 Výběr dat ze SLDB

Pro jednodušší kontrolu výsledných dat jsem zahrnul počty obyvatel získané při Sčítání lidu, domů a bytů 2011 přímo do výstupních souborů. Abych toho dosáhl, musel jsem

tabulky Českého statistického úřadu předem zpracovat do formátu vhodného pro připojení k výstupům.

Z více než tisíce tabulek s výsledky SLDB dostupných na webových stránkách Českého statistického úřadu jsem použil dvě – „Tab. 111 *Obyvatelstvo podle pohlaví a podle druhu pobytu, státního občanství, způsobu bydlení, národnosti a náboženské víry*“ a „Tab. 115 *Vyjíždějící do zaměstnání a škol*“. Z tabulky 111 jsem získal počty obyvatel bydlících v jednotlivých územních celcích a z tabulky 115 jsem použil počty vyjíždějících obyvatel z uvedených územních celků.

Všechny hodnoty v těchto tabulkách byly uvedeny pro územní prvek *Základní sídelní jednotka díl*. Soubory byly ve formátu programu Microsoft Excel, proto jsem je pro další práci převedl na textové soubory s položkami oddělenými středníky a odstranil nadbytečné zástupné znaky nových řádků.

Pro výběr a agregaci dat jsem napsal skript v programovacím jazyce Perl. Na vstupu skriptu se definují kódy ZSJ, LAU2 (obcí), LAU1 (okresů) nebo NUTS3 (krajů). Druhým vstupem je seznam ukazatelů, o které je na výstupu zájem, definovaných Českým statistickým úřadem. Podle formátu kódů územních jednotek (viz kapitoly 3.1, 3.2.2 a 3.2.1) skript určí typ územního prvku a provede odpovídající agregaci hodnot vybraných ukazatelů. Výstupem skriptu je textový soubor s hlavičkou a položkami oddělenými středníky. K tomu se generuje informační textový soubor s užitečnými statistikami.

5.3.2 Analýza dat z mobilní sítě

Vstupní data jsem rozdělil do dvou částí. V první části jsou všechny informace platné bez ohledu na výběr dat pocházejících z mobilní sítě. Jsou zde číselníky územních prvků, které slouží k dohledání názvů územních jednotek podle jejich kódu. Z těchto číselníků je také možné určit hierarchii území, jež je potřebná pro agregaci územních prvků. Dále jsou zde číselníky obsahující kód území, počet obyvatel ze SLDB a kategorii území. Kategorizace území se provádí na základě počtu obyvatel. Další číselník vyjadřuje množství vyjížděk ve formátu: kód území, počet vyjíždějících obyvatel do zaměstnání ze SLDB a počet vyjíždějících obyvatel do škol ze SLDB. Oba druhy číselníků s daty Českého statistického úřadu jsou vytvořeny pro každou úroveň územního členění (ZSJ, LAU2 atd.) zvlášť.

V první části vstupních dat se rovněž nachází číselník, který přiřazuje dny v týdnu vybraným průměrům, které jsem použil ve výstupech. Konkrétně je zahrnuto těchto pět průměrů:

- pondělí až pátek,
- úterý až čtvrtek,
- sobota až neděle,
- pondělí až neděle,
- pátek až pondělí.

Do druhé části vstupních dat jsem zařadil soubory s informacemi platnými pouze pro konkrétní čas a území. Je zde soubor s počty uživatelů v ZSJ vytvořený z dat mobilního operátora ve formátu: identifikátor data, typ uživatele, kód ZSJ a dalších dvacet čtyři

hodnot, které udávají počty uživatelů v rozpětí jedné hodiny.

V této části je dále číselník přiřazující identifikátorům dat reálné datum a den v týdnu. Také je zde soubor, ve kterém je výpis kódů ZSJ. Používá se pro výběr (filtraci) sídelních jednotek z území definovaného výběrem dat z mobilní sítě.

Pomocí HiveQL jsem z každého souboru vytvořil tabulku – tj. definoval jeho strukturu. Díky tomu, že Impala vychází z HiveQL, jsem tyto tabulky mohl použít při řešení oběma nástroji. Dokonce i samotné dotazy jsou mezi Impalou a HiveQL z velké části zaměnitelné.

V případě řešení nástrojem Impala jsem napsal Shell skript, který se spouští z lokálního počítače. Skript předává lokálně uložené dotazy programu *impala-shell*, který zajistí jejich vykonání na vzdáleném Hadoop serveru a výsledky ukládá zpět na lokání disk. Ty mají formát textových souborů s položkami oddělenými středníky. Shell skript nakonec ve výsledcích nahradí desetinné tečky za čárky a na začátek souboru vloží znak *Byte order mark*. Jedná se o trojici bytů `0xEF`, `0xBB` a `0xBF`. Některé programy operačního systému Microsoft Windows (např. Microsoft Excel) tento znak potřebují k identifikaci kódování UTF-8.

Impala neumí parametrizaci dotazů. Definici datové sady a filtru území jsem provedl nahrazením části kódů Unixovým programem *sed*. Největší výhodou řešení v Impale je malá časová náročnost. Vykonání celé série dvaceti sedmi dotazů trvá přibližně devět minut.

Hive je pokročilejší nástroj než Impala. Datovou sadu a filtr území jsem volil pomocí dvou parametrů, kterým se přiřazuje hodnota až při volání dotazu. Výstup dotazů se ukládá do souborů (tabulky) na HDFS. Pro kopírovaní souborů na lokální úložiště jsem vytvořil Shell skript. Ten měl za úkol spojit výstupní soubory do jednoho, přidat záhlaví, přepsat desetinné tečky na čárky, vložit *Byte order mark* a výsledek správně pojmenovat.

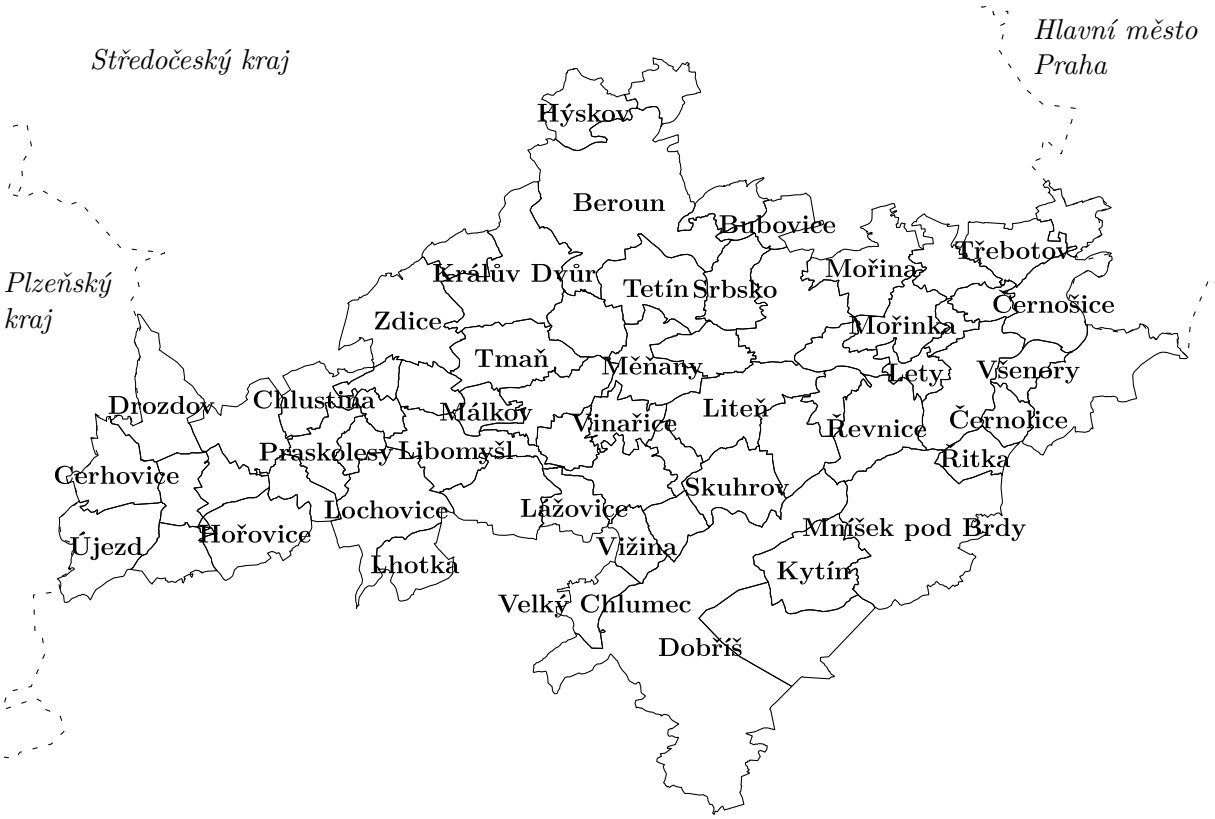
Část dotazů jsem původně zkoušel napsat v Pig Latin s použitím vlastní uživatelsky definované funkce napsané v jazyce Python. Toto řešení bylo paměťově i časově náročné, proto jsem u výsledného řešení zůstal pouze u HiveQL.

Všechny HiveQL dotazy a Shell skripty jsem spojil do jednoho *Oozie workflow*. Výhodou je jednoduché spouštění, definice parametrů a sdílení s ostatními uživateli v grafickém rozhraní Hue. Průběh celého *workflow* (dvaceti devíti akcí) trvá přibližně čtyřicet minut. Při zvyšování množství vstupních dat se budou časy průběhů obou řešení vyrovnávat.

5.3.3 Výsledky

K dispozici jsem měl výsledky referenčního projektu, který interpretoval data z mobilní sítě z patnácti po sobě jdoucích dní na území o rozloze necelých 18 000 km^2 (necelá čtvrtina České republiky). Výsledky jsem zpracoval pro skupinu obcí jihozápadně od Prahy o celkové rozloze 550 km^2. Výčet těch největších je na obrázku 5.1. Všechny výsledky jsou odhadem absolutního počtu obyvatelstva. Jsou násobeny koeficientem, který se snaží reflektovat podíl mobilního operátora na trhu.

V tabulkách 5.1, 5.2 a 5.3 jsem uvedl naměřené počty bydlících, vyjíždějících a dojíždějících občanů pro deset obcí s nejvyšším počtem obyvatelstva. Výběr obsahuje pět obcí



Obrázek 5.1: Zpracované území

kategorie G (5 000 až 19 999 obyvatel) a pět obcí kategorie F (2 000 až 4 999 obyvatel). Tyto kategorie jsou definovány ČSÚ (*Českým statistickým úřadem*). V každé tabulce je pro srovnání uvedena hodnota získaná ze Sčítání lidu, domů a bytů ČSÚ.

Naměřená data jsou minimálně o tři roky mladší než ta ze SLDB, proto mohou být mezi nimi rozdíly. Sčítací listy rozlišovaly mezi místem trvalého pobytu a bydlištěm. I přesto si myslím, že ne všichni uvedli místo svého typického pobytu jako své bydliště. Mnoho lidí může často i výjimečně přespávat u svých partnerů, přátel či zcela někde jinde. Problematika počtu lidí dojíždějících do zaměstnání a škol je na tom podobně. Své vyjížďky nemuseli všichni ve sčítacích formulářích správně uvést. Mnoho občanů nejezdí pracovat denně na stejné místo. Naměřená data jsou přepočtena na základě celorepublikového tržního podílu operátora, který ale není mezi obcemi a ZSJ konstantní.

V tabulce 5.1 jsou zapsány průměrné počty bydlících obyvatel ve vybraných dnech. Pro názornější srovnání s daty ČSÚ jsem do tabulky také uvedl své výsledky v poměru s hodnotami ČSÚ. Na posledním řádku tabulky je uvedena průměrná odchylka od hodnot ČSÚ k hodnocení daného průměru. Hodnotám získaným ze SLDB nejlépe odpovídá průměr měření ve všedních dnech, nejméně pak průměr měření o víkendu. O sobotách a nedělích bylo naměřeno o 15 % méně obyvatel než ve zbytku týdne. Předpokládám, že většina lidí z této ztráty tráví víkendy rekreací např. na svých chatách.

Při srovnání jednotlivých průměrů s hodnotami ČSÚ bydlícího obyvatelstva v celém

Kód obce	Název obce	ČSÚ	Po–Pá	Út–Čt	So–Ne	Pá–Po	Po–Ne
531057	Beroun	18819	19646	21700	14989	15521	18404
		100 %	104 %	115 %	80 %	82 %	98 %
540111	Dobříš	8672	8338	9068	6725	6893	7908
		100 %	96 %	105 %	78 %	79 %	91 %
531189	Hořovice	6951	6598	7149	5717	5675	6363
		100 %	95 %	103 %	82 %	82 %	92 %
533203	Králův Dvůr	6861	8743	9490	7039	7238	8289
		100 %	127 %	138 %	103 %	105 %	121 %
539139	Černošice	6849	5529	5891	4934	4915	5370
		100 %	81 %	86 %	72 %	72 %	78 %
540765	Mníšek pod Brdy	4632	5405	5781	4770	4758	5235
		100 %	117 %	125 %	103 %	103 %	113 %
532011	Zdice	4099	4619	5032	3778	3837	4395
		100 %	113 %	123 %	92 %	94 %	107 %
539198	Dobřichovice	3410	2925	3095	2667	2647	2856
		100 %	86 %	91 %	78 %	78 %	84 %
539643	Řevnice	3217	2185	2288	2004	2005	2137
		100 %	68 %	71 %	62 %	62 %	66 %
532029	Žebrák	2173	2526	2709	2146	2176	2425
		100 %	116 %	125 %	99 %	100 %	112 %
Průměrná odchylka od hodnot ČSÚ			+0 %	+8 %	−15 %	−14 %	−4 %

Tabulka 5.1: Průměrné počty bydlících obyvatel

zpracovaném území na sloupcovém grafu v obrázku 5.2 je patrné, že celotýdenní průměr nejlépe odpovídá výsledkům SLDB. V pracovních dnech je v průměru naměřeno přibližně o 5 000 obyvatel více a o víkendu o 12 000 méně. Průměrný počet bydlících obyvatel o víkendu je téměř stejný jako v průměru od pátku do pondělí. To by mohlo znamenat, že lidé odjíždějí na víkend už v pátek odpoledne a vrací se až v pondělí ráno.

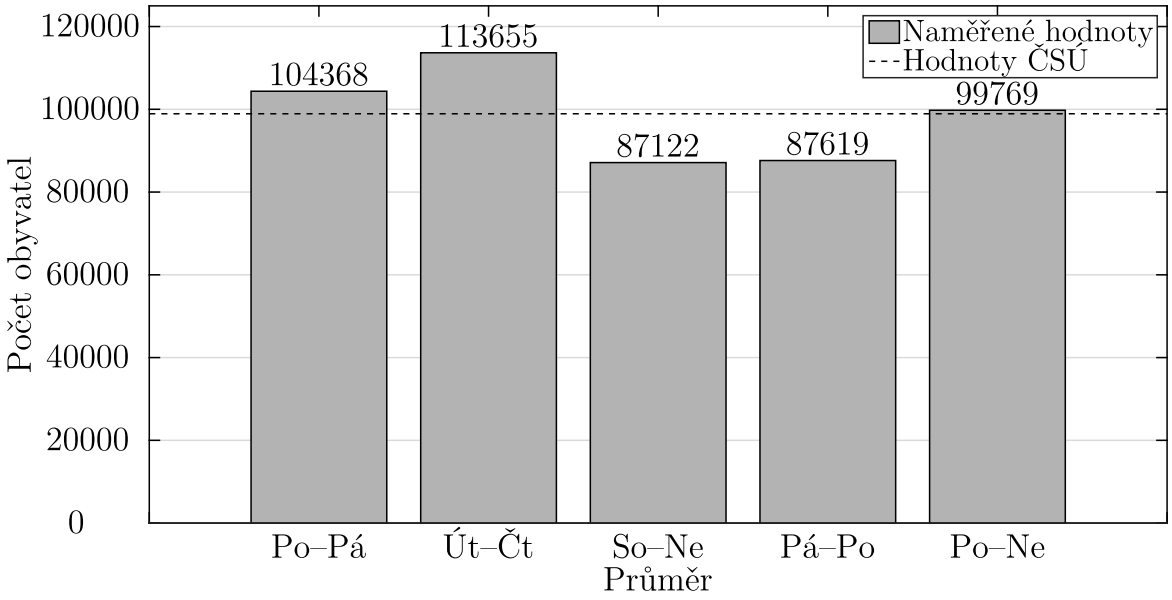
Hodnoty jsem získal pomocí dotazu, který lze vyjádřit rovnicí 5.1:

$$X_{průměr} = \frac{\sum_{j=0}^{J-1} \sum_{n=0}^{N-1} a_j x_{jn}}{\sum_{j=0}^{J-1} a_j}, \tag{5.1}$$

kde $X_{průměr}$ je počet bydlících obyvatel daného průměru, J je celkový počet dní, N je celkový počet ZSJ, x_{jn} je počet bydlících obyvatel ve dni j a ZSJ n a

$$a_j = \begin{cases} 1 & \text{pokud je den } j \text{ součástí průměru,} \\ 0 & \text{jinak.} \end{cases} \tag{5.2}$$

Stejným způsobem jako v předchozím případě jsou v tabulce 5.2 uvedeny absolutní



Obrázek 5.2: Průměrné počty bydlících obyvatel v celém zpracovaném území

a relativní počty vyjíždějících obyvatel. V celém týdnu bylo v průměru naměřeno o 34 % více vyjíždějících obyvatel, než uvádí ČSÚ. Naopak o víkendu byl počet vyjíždějících osob méně než poloviční ve srovnání se všedními dny.

Největší odchylka (+70 %) byla naměřena pro dny úterý, středa a čtvrtek. Podle mého názoru za tento velký rozdíl může neúplné vyplnění sčítacích formulářů. Část o vyjíždění do zaměstnání a škol vyplnilo méně než 20 % obyvatel České republiky. Do zbylé sumy patří lidé, kteří nikam nevyjíždějí, vyjíždějí jinam než do zaměstnání a škol nebo tuto skutečnost neuvedli.

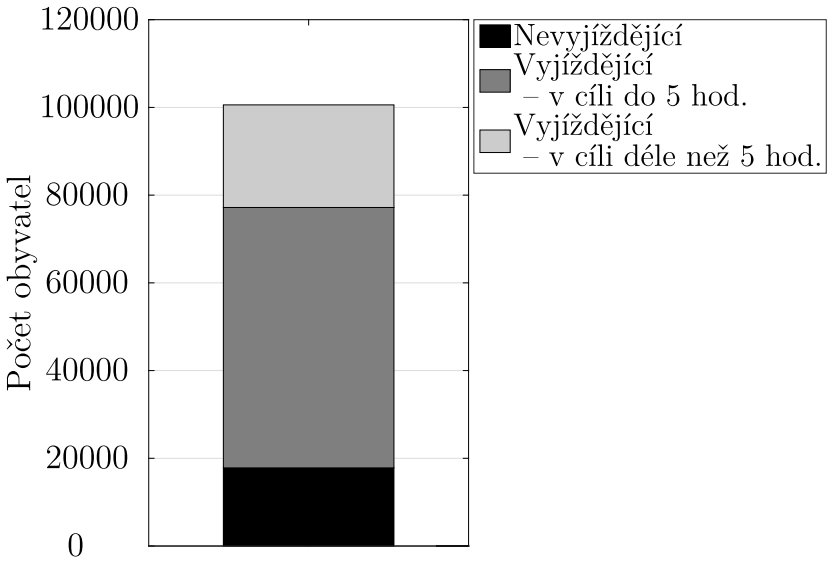
Stejného formátu se drží také tabulka 5.3, ve které jsou ukázány absolutní a relativní počty dojíždějících obyvatel. Průměrná odchylka od hodnot SLDB pro víkend byla +20 %. Tento průměr je především ovlivněn Černošicemi, do kterých o sobotách a nedělích přijelo průměrně 757 obyvatel denně místo 239 podle ČSÚ. Pravděpodobně se tu přes víkend konala nějaká akce, protože průměry ostatních dnů jsou zde nižší. Také je možné, že sem na víkend jezdí lidé z Prahy a okolních obcí. V obci Řevnice byl rovněž naměřen vyšší počet dojíždějících lidí o víkendu než ve všední dny.

Skladba obyvatelstva bydlícího ve zpracovaném území je ukázána na obrázku 5.3. 18 % všech obyvatel v území neopouští své domovské místo na déle než 30 minut. 59 % obyvatel ze svého domova vycestuje, ale v cíli se nezdrží déle než 5 hodin. Zbylých 23 % bydlících osob je v cíli své vyjížďky déle než 5 hodin denně. Pravděpodobně vyjeli do svých zaměstnání a škol.

Na obrázku 5.4 je graf vývoje počtu vyjíždějících obyvatel celého zpracovaného území a dojíždějících do tohoto území během dne v průměru od pondělí do neděle. Na obrázku 5.5 je stejný vývoj pro soboty a neděle. Je vidět, že o víkendu jsou hodnoty obou druhů obyvatel výrazně vyrovnanější než ve zbytku týdne. Maximum dojíždějících osob je v obou případech naměřeno ve dvanácté hodině. Nejvíce vyjíždějících obyvatel v území je v brz-

Kód obce	Název obce	ČSÚ	Po–Pá	Út–Čt	So–Ne	Pá–Po	Po–Ne
531057	Beroun	5325	8589	9530	3787	5365	7308
		100 %	161 %	179 %	71 %	101 %	137 %
540111	Dobříš	2340	3768	4093	1589	2393	3187
		100 %	161 %	175 %	68 %	102 %	136 %
531189	Hořovice	1908	2835	3054	1361	1906	2442
		100 %	149 %	160 %	71 %	100 %	128 %
533203	Králův Dvůr	2055	4169	4532	1827	2680	3544
		100 %	203 %	221 %	89 %	130 %	172 %
539139	Černošice	2291	2467	2586	1220	1739	2135
		100 %	108 %	113 %	53 %	76 %	93 %
540765	Mníšek pod Brdy	1354	2282	2409	790	1424	1884
		100 %	169 %	178 %	58 %	105 %	139 %
532011	Zdice	1071	2058	2240	938	1339	1759
		100 %	192 %	209 %	88 %	125 %	164 %
539198	Dobřichovice	1029	1336	1391	647	943	1152
		100 %	130 %	135 %	63 %	92 %	112 %
539643	Řevnice	873	1065	1096	510	760	917
		100 %	122 %	126 %	58 %	87 %	105 %
532029	Žebrák	626	1156	1251	453	721	968
		100 %	185 %	200 %	72 %	115 %	155 %
Průměrná odchylka od hodnot ČSÚ			+58 %	+70 %	−31 %	+3 %	+34 %

Tabulka 5.2: Průměrné počty vyjíždějících obyvatel



Obrázek 5.3: Skladba obyvatelstva bydlícího ve zpracovaném území

Kód obce	Název obce	ČSÚ	Po–Pá	Út–Čt	So–Ne	Pá–Po	Po–Ne
531057	Beroun	2588	4499	4939	2078	2903	3853
		100 %	174 %	191 %	80 %	112 %	149 %
540111	Dobříš	1043	2405	2635	891	1447	2001
		100 %	231 %	253 %	85 %	139 %	192 %
531189	Hořovice	1623	1682	1845	614	1005	1397
		100 %	104 %	114 %	38 %	62 %	86 %
533203	Králův Dvůr	805	1704	1860	876	1154	1483
		100 %	212 %	231 %	109 %	143 %	184 %
539139	Černošice	239	619	640	757	670	656
		100 %	259 %	268 %	317 %	280 %	274 %
540765	Mníšek pod Brdy	343	1006	1079	573	726	891
		100 %	293 %	315 %	167 %	212 %	260 %
532011	Zdice	755	1148	1268	520	730	981
		100 %	152 %	168 %	69 %	97 %	130 %
539198	Dobřichovice	228	414	418	351	379	397
		100 %	182 %	183 %	154 %	166 %	174 %
539643	Řevnice	199	300	343	319	272	305
		100 %	151 %	172 %	160 %	137 %	153 %
532029	Žebrák	1345	1660	1738	434	979	1333
		100 %	123 %	129 %	32 %	73 %	99 %
Průměrná odchylka od hodnot ČSÚ			+88 %	+102 %	+21 %	+42 %	+70 %

Tabulka 5.3: Průměrné počty dojíždějících obyvatel

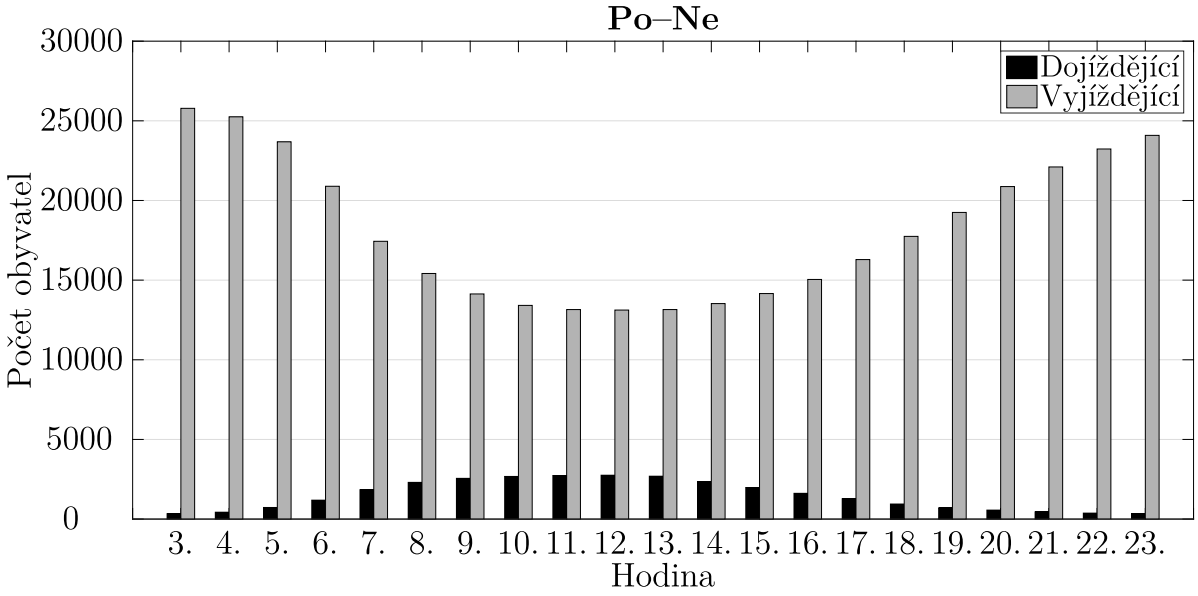
kých ranních hodinách, ve kterých postupně odjíždějí např. do zaměstnání, a ve večerních hodinách, kdy se vrací zpět do místa svého bydliště. Minimální počet vyjíždějících osob se o víkendu přesune z dvanácté hodiny na čtrnáctou. To by mohlo nasvědčovat tomu, že o víkendu více lidí vyjíždí až po obědě.

Obyvatel dojíždějících do zpracovaného území od rána postupně přibývá až do dvanácté hodiny. V té začne postupný pokles, který vydrží až do večera. Podobný průběh obou grafů může být částečně způsoben lidmi pracujícími o víkendu a těmi, kteří o víkendu podnikli nějaký výlet či se v měřeném území rekreují. Mnoho lidí jezdí o víkendech na nákupy.

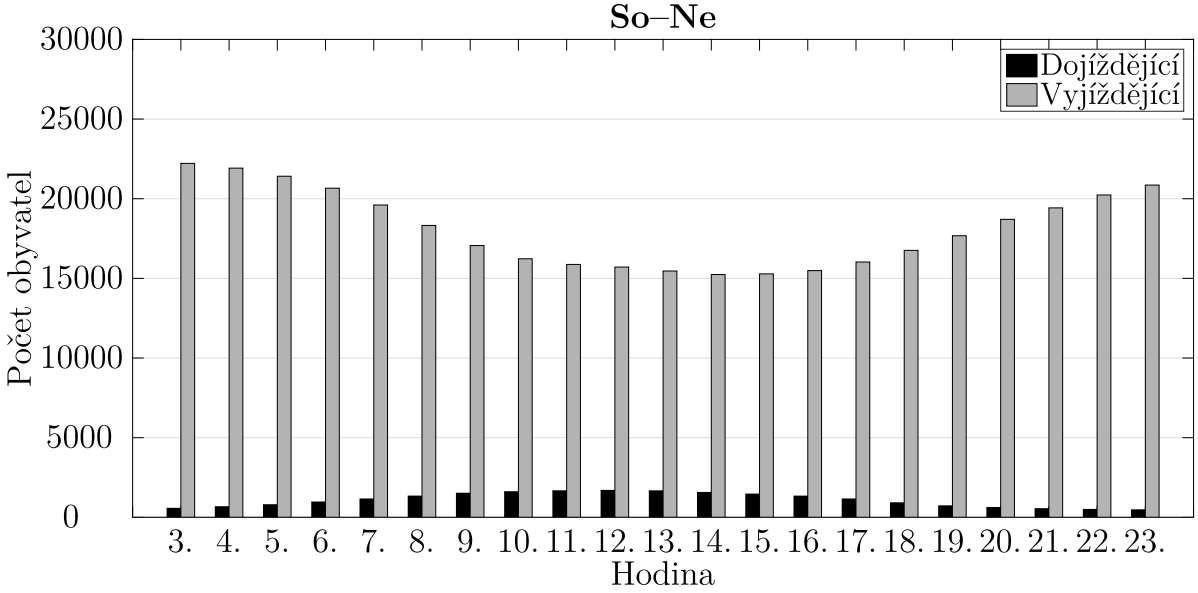
Tyto hodnoty jsem získal pomocí dotazu, který lze přepsat do rovnice 5.3:

$$Y_{i,\text{průměr}} = \frac{\sum_{j=0}^{J-1} \sum_{n=0}^{N-1} a_j y_{ijn}}{\sum_{j=0}^{J-1} a_j}, \tag{5.3}$$

kde $Y_{i,\text{průměr}}$ je počet vyjíždějících/dojíždějících daného průměru v hodině i, J je celkový počet dní, N je celkový počet ZSJ, y_{ijn} je počet vyjíždějících/dojíždějících obyvatel ve dni j a ZSJ n a a_j znovu nabývá hodnot podle rovnice 5.2.



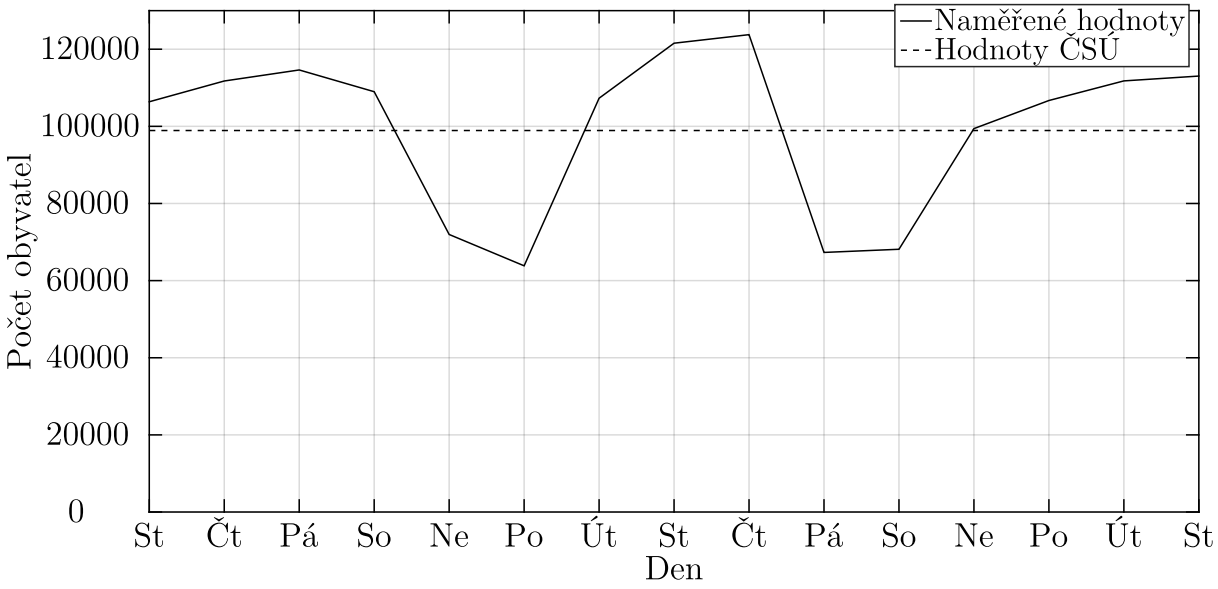
Obrázek 5.4: Denní vývoj průměrného počtu dojíždějících a vyjíždějících obyvatel zpracovaného území v celém týdnu



Obrázek 5.5: Denní vývoj průměrného počtu dojíždějících a vyjíždějících obyvatel zpracovaného území o víkendu

Graf na obrázku 5.6 zobrazuje vývoj počtu bydlících obyvatel v celém zpracovaném území během všech měřených dní. Je zajímavé, že výrazný dvoudenní pokles nemá periodu rovnou sedmi dnům a ani v jednom případě nebyl naměřen v obou po sobě jdoucích víkendových dnech. V prvním případě pokles nastal v neděli a pondělí, ve druhém případě v pátek a sobotu. Maximální počet bydlících obyvatel byl v mobilní síti zaznamenám ve čtvrtek, ve kterém hodnota po přepočtu nabyla necelých 124 000 osob. Minimum 64 000

bydlících osob bylo získáno první měřené pondělí. Z toho vyplývá, že počet lidí bydlících v tomto území může během týdne výrazně klesnout – téměř na polovinu své maximální hodnoty. V měřené dny nebyly žádné státní svátky ani prázdniny, které by mohly měření ovlivnit. Z důvodu malého počtu měřených dní nemohu tento vývoj přesně vysvětlit.



Obrázek 5.6: Vývoj počtu bydlících obyvatel zpracovaného území v měřených dnech

5.4 Trajektorie pohybu obyvatel

5.4.1 Výběr dat ze SLDB

Pro kontrolu výsledných dat jsem použil tabulku *Dojížďkové proudy SLDB 2011: obec versus obec*. Tabulka obsahuje tyto ukazatele na úrovni obcí:

- ppracel – počet osob vyjíždějících do zaměstnání celkem,
- ppraden – počet osob vyjíždějících denně do zaměstnání,
- pskocel – počet vyjíždějících žáků, studentů a učňů celkem,
- pskoden – počet denně vyjíždějících žáků, studentů a učňů.

Soubor byl stejně jako v předchozím případě ve formátu programu Microsoft Excel, proto jsem ho pro další práci převedl na textový soubor s položkami oddělenými středníky.

Pro výběr dat jsem znovu napsal skript v jazyce Perl. Pokud je na vstupu skriptu seznam kódů obcí, vypíší se všechny nenulové kombinace vyjížděk a dojížděk pro dané obce. Druhou možností je seznam kódů výjezdových obcí a dojezdových obcí. V tomto případě se vypíší všechny kombinace dojezdů z výjezdových do dojezdových obcí. Také je možno definovat různé skupiny obcí, které se výstupu objeví jako samostatná území vyjížděk i dojížděk. Výběr ukazatelů je také možný. Výstupem skriptu je textový soubor s hlavičkou a položkami oddělenými středníky. Dále se k němu vygeneruje informační textový soubor s užitečnými statistikami.

5.4.2 Analýza dat z mobilní sítě

Do první části vstupních dat jsem opět zařadil soubory, které jsou platné bez ohledu na data pocházející z mobilní sítě. V tomto případě stačí číselník územních prvků, který má stejnou funkci jako v předchozí úloze v kapitole 5.3.

V druhé části jsem znovu použil interpretovaná data pocházející z mobilní sítě, která byla v jiném formátu než ta použitá pro analýzu v kapitole 5.3. Soubory obsahovaly identifikátory uživatelů a cest, jejich klasifikace, časy začátků a konců cest a série kódů ZSJ. Tyto řady ZSJ nabývají různých délek, protože cesta uživatele může procházet různým počtem ZSJ.

Pro snadnou definici struktury souboru pomocí tabulky jsem napsal skript v Perlu, který tento soubor rozdělí na dva. První výstupní soubor obsahuje identifikátory uživatelů a cest, jejich klasifikaci i časy začátků a konců cesty z původního souboru. Díky identifikátorům uživatelů a cest je každá cesta jednoznačně určena v celém souboru. Proto jsou zapsány i v druhém výstupním souboru, který mj. obsahuje série ZSJ. Všechny jsou již ale vypsány pouze v jednom sloupci.

Všechny dotazy jsem napsal pro nástroj Impala. S jejich pomocí jsem zpracoval tyto statistiky:

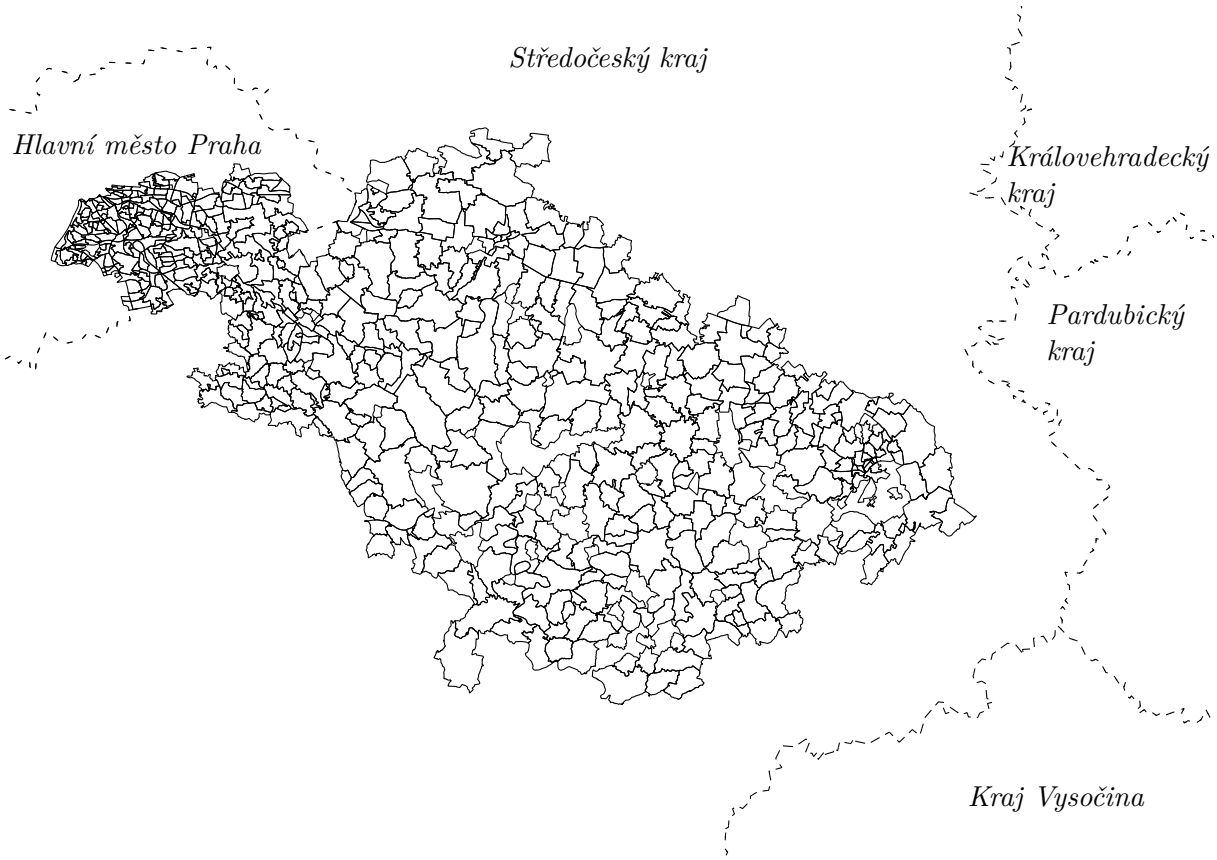
- Počty všech cest a počty unikátních uživatelů cestujících mezi dvěma obcemi nebo ZSJ, ve kterých uživatel strávil více než pět hodin.
- Počty uživatelů rozdělených do kategoriích podle doby strávené v konkrétní ZSJ

nebo obci.
- Počty všech cest a počty unikátních uživatelů rozdělených podle hodin, ve kterých končí v konkrétních ZSJ.

Pro spuštění všech dotazů požívám stejný Shell skript jako v předchozí úloze. Ten předává lokálně uložené dotazy programu *impala-shell*, který zajistí jejich vykonání na vzdáleném Hadoop serveru a výsledky ve formě textových souborů ukládá zpět na lokání disk. Skript nakonec ve výsledcích nahradí desetinné tečky za čárky a na začátek souboru vloží znak *Byte order mark*.

5.4.3 Výsledky

K dispozici jsem měl výsledky jiného referenčního projektu než toho v kapitole 5.3. Proto jsem musel pracovat s jiným územím. Data se vztahovala na jednu středu v území o celkové rozloze přibližně 1 340 km^2. Území, které je znázorněno na obrázku 5.7, pokrývá jihovýchod Prahy a dále se rozprostírá jihovýchodním směrem. Všechny výsledky jsou znovu pouze odhadem absolutního počtu obyvatelstva. Jsou násobeny koeficientem, který se snaží reflektovat podíl mobilního operátora na trhu.

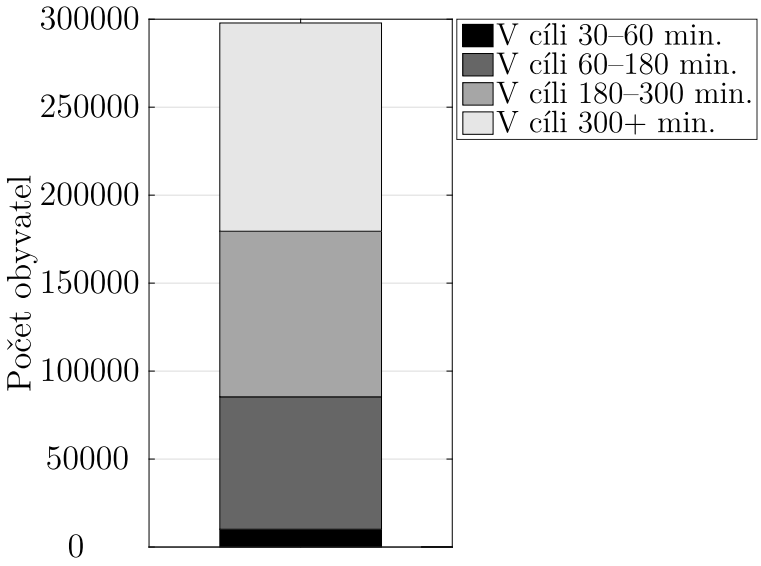


Obrázek 5.7: Zpracované území

Každý uživatel, který projel tímto územím a opustil svou domácí pozici na déle než 30 minut, uskutečnil v průměru 5,5 cest, v jejichž cíli pobyl déle než půl hodiny. Ve dvou

z těchto cest strávil v cíli 30–60 minut, u dalších dvou 60–180 minut a u jedné 180 až 300 minut. Každý druhý uživatel učinil v průměru jednu cestu, na jejíž cíli strávil déle než 5 hodin.

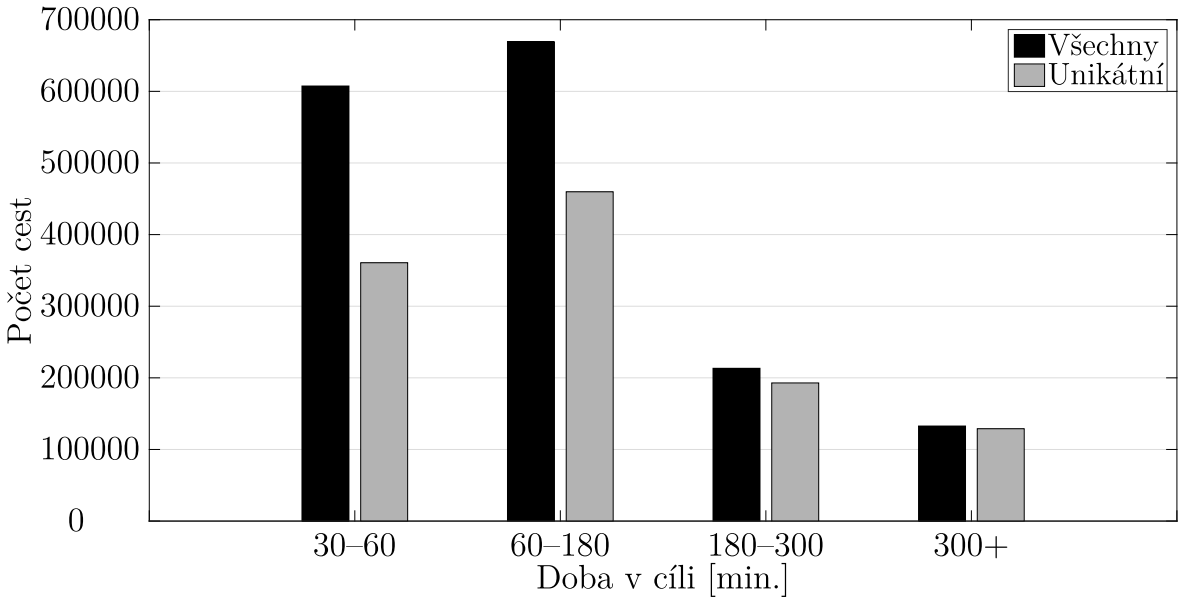
Na obrázku 5.8 je složení přepočítaného obyvatelstva rozděleného podle maximální doby strávené v cíli ze všech jejich cest. Z necelých 300 000 cestujících osob pouze 3 % nestráví v cíli alespoň jedné ze svých cest déle než jednu hodinu. Čtvrtina z počtu obyvatel stráví v cíli 60–180 minut, necelá třetina cestujících 180–300 minut a zbylých 40 % déle než 300 minut.



Obrázek 5.8: Skladba obyvatelstva podle maximální doby strávené v cílech

Dohromady jsem měl po přepočtu popsáno přes 1 600 000 cest, jejichž složení je v grafu na obrázku 5.9. Ve sloupcích označených jako unikátní se všechny cesty stejné osoby, dojezdové ZSJ a stejného typu započítaly pouze jednou. Největší množství cest – přibližně 670 000 spadá do kategorie 60–180 minut. Asi 30 % z těchto cest bylo absolvováno více než jednou. Z naměřených hodnot jsem vypočítal přes 600 000 cest s cílem do 60 minut. Takto krátké vyjížďky jsou lidmi opakovány nejčastěji. Unikátních bylo pouze necelých 60 %. Přes 210 000 cest bylo v kategorii 180–300 minut a necelých 130 000 v kategorii 300+ minut. Cesty obsahující cíle s delší dobou pobytu byly opakovány méně než z 10 %.

Pro srovnání naměřených vyjížděk s počty osob vyjíždějících do zaměstnání a škol ze SLDB jsem vybral obce Bečváry a Zásmuky. Ty jsou sousedními obcemi okresu Kolín ve Středočeském kraji, které se nacházejí přibližně ve středu zpracovaného území. V tabulce 5.4 je dvanáct nejčastějších vyjížďkových proudů začínajících v jedné z těchto dvou obcí. Jsou vybrány pouze proudy, v jejichž obcích dojížďky lidé zůstali déle než 300 minut. Ve sloupci *První cesta* jsou sečteny pouze první cesty osob ve dne, v jejichž cíli strávily déle než 300 minut. Tyto cesty považuji za vyjížďky do zaměstnání a škol, a proto je srovnávám s údaji od ČSÚ. SLDB ale neobsahuje cesty v rámci stejné obce, proto chybí i v tabulce.



Obrázek 5.9: Skladba cest podle doby strávené v cílech

Obec vyjížďky	Obec dojížďky	Kód obce doj.	ČSÚ	První cesta[1]	Všichni[2]
Zásmuky	Dolní Chvatliny	533297	0	23	133
Bečváry	Suchdol	534439	3	3	77
Zásmuky	Zásmuky	533921		7	70
Bečváry	Kořenice	533408	8	10	57
Bečváry	Dolní Chvatliny	533297	0	13	53
Zásmuky	Horní Kruty	533327	0	13	43
Zásmuky	Malotice	533513	0	3	37
Zásmuky	Bečváry	533181	11	7	37
Zásmuky	Drahobudice	564681	0	13	37
Bečváry	Bečváry	533181		10	33
Bečváry	Zásmuky	533921	19	13	30
Zásmuky	Kouřim	533424	5	7	27

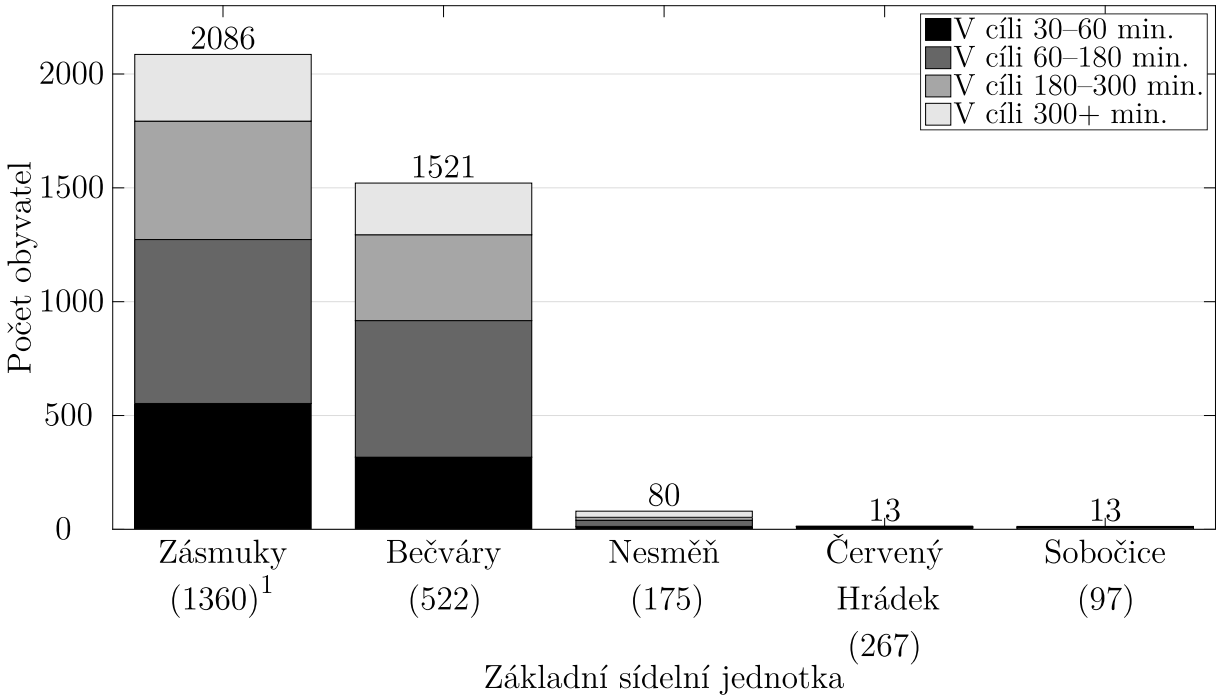
[1] Suma pouze prvních cest obyvatel ve dne s cílem 300+ minut.
[2] Suma všech cest ve dne s cílem 300+ minut.

Tabulka 5.4: Cíl a počet vyjíždějících obyvatel z obcí Zásmuky a Bečváry

Lidé ze Zásmuk nejčastěji vyjíždějí do sousedních Dolních Chvatlin, ale tento proud se ve výsledcích SLDB nenachází. Po přepočtu bylo naměřeno 133 cestujících osob, ale jen u necelé pětiny to byla jejich první cesta. Z Bečvár nejvíce osob – 77 vyjíždí do sousední obce Suchdol, ale pouze u třech to byla jejich první cesta. Tato hodnota se shoduje s daty ČSÚ. Pokud pro danou dvojici obcí existuje záznam v SLDB, je jeho rozdíl s naměřenou a přepočtenou hodnotou ve sloupci *První cesta* malý.

Na obrázku 5.10 je skladba obyvatelstva přijíždějícího do základních sídelních jednotek

v obcích Zásmuky a Běčváry. ZSJ Zásmuky (1360 obyvatel)[1], Nesměň (175 obyvatel) a Sobočice (97 obyvatel) patří do obce Zásmuky. Bečváry (522 obyvatel) a Červený Hrádek (267 obyvatel) jsou v obci Bečváry. Do zbylých šesti ZSJ v těchto obcích nebyly zaznamenány žádné dojížďky. Ačkoli je v ZSJ Bečváry evidováno méně než polovina obyvatel Zásmuk, má přibližně tři čtvrtiny všech dojezdů co Zásmuky. Složení dojezdů je ale téměř stejné. Přes pětinu obyvatel stráví v ZSJ 30 až 60 minut, více než třetina 60–180 minut, necelá čtvrtina 180–300 minut a zbylá část je zde déle než 300 minut.



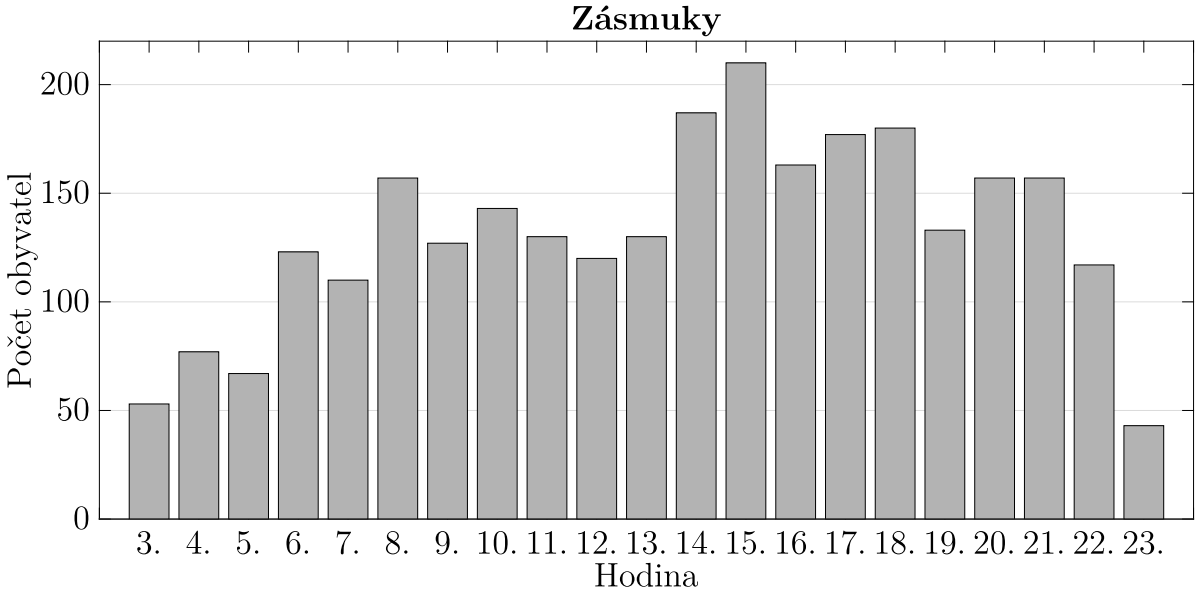
Obrázek 5.10: Skladba obyvatelstva podle doby strávené v ZSJ obcí Zásmuky a Bečváry; v závorkách jsou počty bydlících obyvatel v ZSJ dle ČSÚ

Podle dat ze SLDB bydlí v Červeném Hrádku přibližně polovina obyvatel ve srovnání se sousedními Bečváry. Naměřeno bylo pouze 13 dojíždějících obyvatel do Červeného Hrádku. Je možné, že někteří dojíždějící do této ZSJ byli započítáni mezi osoby dojíždějící do Bečvárů.

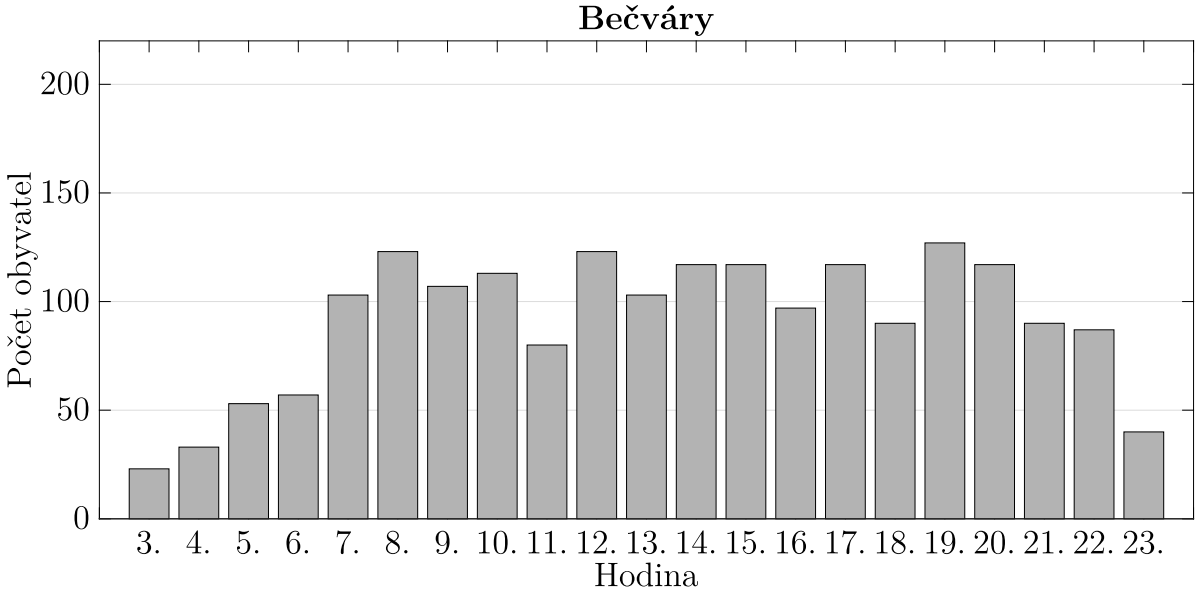
V ZSJ Nesměň je skladba jiná pravděpodobně proto, že je zde naměřeno málo záznamů. V Červeném Hrádku a Sobočicích bylo po přepočtu naměřeno pouhých 13 přijíždějících obyvatel. Nejsou v nich ani zastoupeny všechny kategorie strávené doby.

Dále jsem zpracoval statistiku počtu přijíždějících obyvatel v jednotlivých hodinách do ZSJ Zásmuky (viz obrázek 5.11) a ZSJ Bečváry (viz obrázek 5.12). U Zásmuk je vidět nárůst počtu přijíždějících obyvatel v šesté hodině ranní a u Bečvárů o hodinu později. Po této hodině se v Bečvárech drží hodnota příchozích kolem 100 obyvatel za hodinu až do večerních hodin. Oproti tomu v Zásmukách se po čtrnácté hodině ukazuje výrazný nárůst. Hodiny menších poklesů hodnot během dne se u obou ZSJ liší. Ve dvacáté třetí hodině je v obou grafech zaznamenána výrazná redukce přijíždějících obyvatel.

[1]Počty bydlících obyvatel v ZSJ dle ČSÚ.



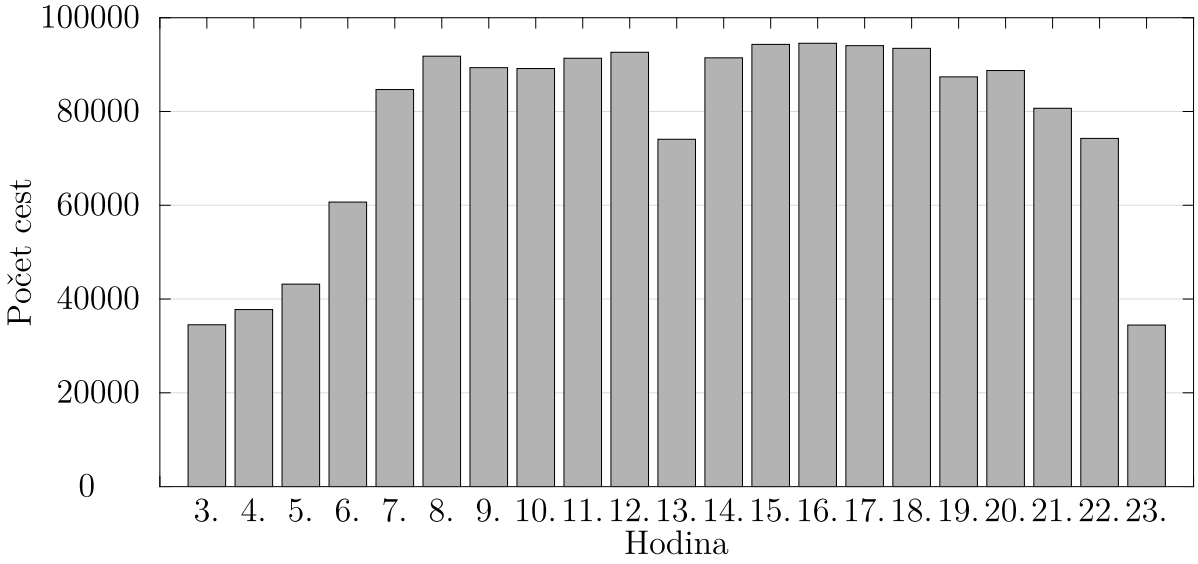
Obrázek 5.11: Počty obyvatel s cílem cesty v ZSJ Zásmuky v měřenou středu



Obrázek 5.12: Počty obyvatel s cílem cesty v ZSJ Bečváry v měřenou středu

Pro srovnání s grafy v obrázcích 5.11 a 5.12 je na obrázku 5.13 denní vývoj přepočtených sum všech naměřených cest v celém území. Zde může být pozorován nárůst počtu cest ráno v šesté hodině následovaný dalším nárůstem během sedmé hodiny. Hodnoty se drží nad hranicí 80 000 až do dvacáté první hodiny. Jedinou výjimkou je pokles o 20 % během třinácté hodiny.

Je zajímavé, že počet cílů cest je od sedmé do dvacáté hodiny relativně neměnný. Ráno mezi sedmou a devátou hodinou, kdy lidé přijíždějí do zaměstnání, stejně tak odpoledne mezi patnáctou a osmnáctou hodinou, kdy se vrací domů, bych očekával lokální maxima



Obrázek 5.13: Denní vývoj počtu cílů cest ve zpracovaném území v měřenou středu

grafu. Pozoruhodný je také relativně velký počet končících cest v nočních a brzkých ranních hodinách. 37 % maximálního denního počtu cest skončí během třetí hodiny ranní nebo dvacáté třetí hodiny večer. Ještě větší počty byly naměřeny ve čtvrté a páté hodině. To může být částečně způsobeno lidmi s nestandardní pracovní dobou.

Dotaz pro získání hodnot v obrázcích 5.11 a 5.12 lze přepsat na rovnici 5.4. Rovnice 5.5 odpovídá případu na obrázku 5.13 :

$$Z_{in} = \sum_{k=0}^{K_i-1} b_{ikn}, \tag{5.4}$$

$$Z_i = \sum_{n=0}^{N-1} \sum_{k=0}^{K_i-1} b_{ikn}, \tag{5.5}$$

kde Z_{in} je počet končících cest (obyvatel) v hodině i a ZSJ n, K_i je celkový počet cest probíhajících v hodině i, N je celkový počet ZSJ a b_{ikn} je záznam poslední ZSJ n na cestě k v hodině i, pro který platí:

$$b_{ikn} = \begin{cases} 1 & \text{když je časová známka záznamu z intervalu } \langle i; i+1) \text{ a} \\ 0 & \text{jinak.} \end{cases} \tag{5.6}$$

6 Závěr

Signalizační data v mobilní síti nemusí sloužit pouze pro zajištění služeb mobilního operátora, ale mohou být využity k získání cenných informací o uživatelích sítě. K těmto informacím se lze dostat pouze po vhodném zpracování signalizačních dat a jejich následnou analýzou. Především z důvodu velkého objemu těchto dat není možné použít klasických výpočetních a analytických nástrojů. Naštěstí již existuje mnoho platforem a nástrojů pro práci s velkými daty.

K dispozici jsem měl výpočetní platformu skládající se z několika uzlů včetně velkého úložného prostoru. Pro správu dat, vývoj aplikací, použití analytických modelů a správu těchto prostředků zde byla instalována distribuce aplikačního rámce Hadoop společnosti Cloudera. Ta obsahuje rozmanitou sadu komponent pro práci s velkými daty, jejichž většinu jsem při řešení využil.

V této práci jsem se zabýval zpracováním lokalizačních dat ze signalizace mobilní sítě. Každá mobilní stanice vyprodukuje typicky 100 až 1000 záznamů denně. Většina z nich pochází ze zpráv procedury *Location Update*. Ta je vykonávána periodicky nebo při každém přepojení do nové LA.

Poloha uživatele v síti (tj. buňka, ve které se nachází) je převedena na umístění v územním celku České republiky. Územní členění se na nejvyšších úrovních řídí evropskými systémy NUTS a LAU. Výsledky jsem zpracoval na úrovni obcí, které odpovídají standardu LAU2, a základních sídelních jednotek.

Pro kontrolu a porovnání mých výstupních hodnot jsem využil data z výsledků Sčítání lidu, domů a bytů zajištěných Českým statistickým úřadem v roce 2011. Pro jejich efektivní výběr jsem napsal jednoduše použitelné skripty v programovacím jazyce Perl. Některá data byla bohužel dostupná pouze na úrovni obcí.

Z dostupných předzpracovaných dat jsem zpracoval dvě analýzy. Produktem první analýzy byl vývoj počtu různých skupin obyvatel v územních prvcích. Obyvatelé byli rozděleni do skupin: bydlící, vyjíždějící, dojíždějící a případně nevyjíždějící. Definoval jsem několik průměrů podle zařazených dní v týdnu a porovnal je s hodnotami ČSÚ. Nejvíce se jim blížil celotýdenní průměr bydlících osob celého zpracovaného území (jihozápadně od Prahy). O víkendu bylo naměřeno v průměru nejméně bydlících a naopak nejvíce bylo v průměru zachyceno v úterý, středu a čtvrtek.

18 % bydlících obyvatel ve 24 hodinách neopustí svou domácí pozici a proto jsou klasifikováni jako nevyjíždějící. Dalších 59 % vycestuje, ale v cíli své cesty stráví méně než 5 hodin. Zbylých 23 % v cíli zůstane déle než 5 hodin. Ve vývoji počtu bydlících obyvatel v patnácti po sobě jdoucích měřených dnech jsem nezpozoroval žádnou periodu

či vzorec. Kvůli malému množství měřených dní jsem bohužel nebyl schopen analýzou vyvodit přesné a konkrétní závěry.

Ve druhé analýze jsem se zabýval trajektoriemi pohybu obyvatel v jednom dni. Majorita zaznamenaných cest byla kratšího charakteru. Lidé strávili v cíli většiny cest 60 až 180 minut. Cest s cílem, ve kterém osoby zůstaly 30–60 minut, nebylo o moc méně. Ale při srovnání pouze nejdelších cest jednotlivých obyvatel bylo největší množství – 40 % delších než 300 minut.

V budoucnu by bylo dobré ověřit funkčnost vytvořených postupů pro ještě větší množství vstupní dat (rozlehlejší území a delší časový interval). Především by bylo zajímavé porovnat mezi sebou obě řešení v nástrojích Hive a Impala a ověřit domněnku, že s přibývajícími daty bude výkon nástroje Impala klesat rychleji než u Hive. Dále by bylo vhodné se zamyslet, jaké další užitečné informace lze získat z lokalizačních dat nebo i jiných signalizačních dat z mobilní sítě.

Seznam zkratek

AuC	Authentication Centre. 9, 11
BCC	BTS Color Code. 8
BCCH	Broadcast Control Channel. 7, 10, 11
BSC	Base Station Controller. 7, 8
BSIC	Base transceiver Station Identity Code. 8
BSS	Base Station Subsystem. 7
BTS	Base Transceiver Station. 7, 8
CC	Country Code. 7
CDH	Cloudera's Distribution including Apache Hadoop. 24
CGI	Cell Global Identification. 8
CI	Cell Identity. 8
ČSÚ	Český statistický úřad. 28, 30, 37–39, 42
EIR	Equipment Identity Register. 7, 9
GSM	Global System for Mobile Communications. 3, 9
HDFS	Hadoop Distributed File System. 22, 23, 27
HLR	Home Location Register. 7, 9, 11, 12
IMEI	International Mobile station Equipment Identity. 6, 7, 9, 10
IMSI	International Mobile Subscriber Identity. 5–7, 9–12
ITU-T	International Telecommunication Union – Telecommunication standardization sector. 9
LA	Location Area. 7–11, 42
LAC	Location Area Code. 7, 24
LAI	Location Area Identitification. 6–12
LAU	Local Administrative Unit. 14, 16, 26, 42
LU	Location Update. 11, 12
MCC	Mobile Country Code. 7
MM	Mobility Management. 9, 10
MMS	Multimedia Message Service. 1

MNC Mobile Network Code. 7
MS Mobile Station. 5, 7–12, 24
MSC Mobile Switching Center. 7–11
MSISDN Mobile Station international ISDN number. 6, 7, 9
MSRN Mobile Station Roaming Number. 6, 7, 9, 11

NCC Network Color Code. 8
NDC National Destination Code. 7
NSS Network and Switching Subsystem. 8
NUTS Nomenclature des unités territoriales statistiques. 14–16, 26, 42

OMC Operation and Maintenance Centre. 9
OSS Operation Subsystem. 9

PIN Personal Identity Number. 5
PUK PIN Unblocking Key. 5

SIM Subsriber Identity Module. 5–7, 10, 11
SLDB Sčítání lidu, domů a bytů. 18, 19, 26, 28–30, 35, 37–39
SMS Short Message Service. 1, 6, 8
SQL Structured Query Language. 23

TAC Type Approval Code. 7
TMN Telecommunications Management Network. 9
TMSI Temporary Mobile Subscriber Identity. 6, 7, 9–12
TRAU Transcoding and Rate Adaptation Unit. 8

VAS Value-Added Service. 1
VLR Visitor Location Register. 7, 9–12

XML eXtensible Markup Language. 23

YARN Yet Another Resource Negotiator. 22

ZSJ Základní sídelní jednotka. 16, 18, 24–29, 32, 35–37, 39–41

Literatura

[1] WALKE, Bernhard. *Mobile Radio Networks: Networking and Protocols*. New York: John Wiley & Sons, c1999. ISBN 0471975958.

[2] MISHRA, Ajay R. *Fundamentals of Cellular Network Planning and Optimisation 2G/2.5G/3G... Evolution to 4G*. Chichester: John Wiley & Sons, 2004. ISBN 9780470862681.

[3] EBERSPÄCHER, Jorg, Hans-Joerg VOGEL, Christian BETTSTETTER. *GSM: Architecture, Protocols and Services*. 3rd ed., U.K.: Wiley, 2008. ISBN 9780470741726

[4] VANĚK, Tomáš. *Zabezpečení komunikace v mobilních sítích*. [přednáška]. Praha: ČVUT FEL, 6. 1. 2016.

[5] 3GPP TS 23.003 *Numbering, addressing and identification*. 13. vyd. Sitges: ETSI, prosinec 2015. Dostupné z: `http://www.etsi.org/deliver/etsi_ts/123000_123099/123003/13.04.00_60/ts_123003v130400p.pdf`

[6] SAUTER, Martin. *From GSM to LTE-Advanced: An Introduction to Mobile Networks and Mobile Broadband (Revised 2nd Edition)*. Chichester, England: Wiley, 2014. ISBN 9781118861936.

[7] 3GPP TS 23.012 *Location management procedures*. 13. vyd. Sitges: ETSI, prosinec 2015. Dostupné z: `http://www.etsi.org/deliver/etsi_ts/123000_123099/123012/13.00.00_60/ts_123012v130000p.pdf`

[8] *T312* [online]. DeFire BD. [vid. 21. 5. 2016]. Dostupné z: `http://gsm-optimization.blogspot.cz/2012/04/t3212.html`

[9] *Administrativní členění NUTS - Česko* [online]. Centrum pro regionální rozvoj České republiky. [vid. 5. 3. 2016]. Dostupné z: `http://www.risy.cz/cs/administrativni-cleneni-nuts-cesko`

[10] *File:CZ-NUTS3.svg* [online]. Lukáš Mižoch. [vid. 5. 3. 2016]. Dostupné z: `https://commons.wikimedia.org/wiki/File:CZ-NUTS3.svg`

[11] *Local Administrative Units (LAU)* [online]. Eurostat. [vid. 5. 3. 2016]. Dostupné z: `http://ec.europa.eu/eurostat/web/nuts/local-administrative-units`

[12] *Soustava prvků* [online]. Český statistický úřad. [vid. 5. 3. 2016]. Dostupné z: `https://www.czso.cz/csu/rso/soustava_prvku`

[13] *Obec a vojenský újezd* [online]. Český statistický úřad. [vid. 5. 3. 2016]. Dostupné z: https://www.czso.cz/csu/rso/obec_rso

[14] *Základní sídelní jednotka* [online]. Český statistický úřad. [vid. 5. 3. 2016]. Dostupné z: https://www.czso.cz/csu/rso/zsj_rso

[15] *Charaktery základních sídelních jednotek* [online]. Český úřad zeměměřický a katastrální. [vid. 5. 3. 2016]. Dostupné z: http://www.cuzk.cz/Uvod/Produkty-a-sluzby/RUIAN/2-Poskytovani-udaju-RUIAN-ISUI-VDP/Ciselniky-ISUI/Dalsi-atributy.aspx#CE_CHARAKTER_ZSJ

[16] *Základní sídelní jednotky* [online]. Český úřad zeměměřický a katastrální. [vid. 7. 5. 2016]. Dostupné z: http://www.cuzk.cz/Uvod/Produkty-a-sluzby/RUIAN/2-Poskytovani-udaju-RUIAN-ISUI-VDP/Ciselniky-ISUI/Nizsi-uzemni-prvky-a-uzemne-evidencni-jednotky/Zakladni-sidelni-jednotky.aspx

[17] *Informace o sčítání* [online]. Český statistický úřad. [vid. 6. 3. 2016]. Dostupné z: https://www.czso.cz/csu/sldb/o_scitani

[18] *Základní informace o sčítání* [online]. Český statistický úřad. [vid. 6. 3. 2016]. Dostupné z: https://www.czso.cz/csu/sldb/zakladni_informace_o_scitani

[19] *Základní výsledky – ČR* [online]. Český statistický úřad. [vid. 6. 3. 2016]. Dostupné z: https://vdb.czso.cz/vdbvo2/faces/cs/shortUrl?su=55f5cb04

[20] DEROOS, Dirk, Paul ZIKOPOLOUS, Rafael COSS. *Hadoop For Dummies*. Hoboken, New Jersey: John Wiley & Sons, 2014. ISBN 9781118705032.

[21] BAESENS, Bart. *Analytics in a Big Data World: The Essential Guide to Data Science and its Applications*. Hoboken, New Jersey: John Wiley & Sons, 2014. ISBN 9781118892701.

[22] LOSHIN, David. *Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph*. Amsterdam: Elsevier, Morgan Kaufmann, 2013. ISBN 9780124173194.