TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Georgi Suikanen 178070 IAMB

MOBIILTELEFONISIDE ANDMETE KASUTAMINE PIIRKIIRUSE ÜLETAMISE HINDAMISEKS

Magistritöö

Juhendaja: Ants Torim
PhD

Tallinn 2020

Autorideklaratsioon

Kinnitan, et olen koostanud antud lõputöö iseseisvalt ning seda ei ole kellegi teise poolt varem kaitsmisele esitatud. Kõik töö koostamisel kasutatud teiste autorite tööd, olulised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on töös viidatud.

Autor: Georgi Suikanen

06.01.2020

Annotatsioon

Käesolevas lõputöös uurib töö autor, kas mobiiltelefoniside (CDR) andmed sobivad piirkiiruse ületamise hindamiseks. Selleks kasutab autor andmeid, mis avalikustati *Data for Development* väljakutse raames ja mis kirjeldavad mobiiltelefonivõrgu kasutamist Senegalis 2013. aastal. Andmete põhjal hindab töö autor klientide liikumiskiirust ja valideerib saadud tulemusi *Google Maps Distance Matrix API* abil.

Lõputöö teema on hästi aktuaalne just sellepärast, et viimaste aastate trend näitab arenevate riikide autostumist. Suurema autode arvuga aga kahjuks tihti kaasneb ka suurem liiklusõnnetuste arv, ja paljud arenevad riigid vajavad lahendusi, millega nad saavad muuta liiklust ohutumaks ilma suurte investeeringute tegemata. CDR andmete kasutamine piirkiiruse hindamiseks võib olla hea lahendus, kuna CDR andmete kogumiseks ja töötlemiseks ei ole vaja suuri investeeringuid infrastruktuuri ja riistvarasse. CDR andmete analüüsimine ei vaja suuri arvutusvõimsusi ja saadud tulemuste tõlgendamine on lihtne kaartide abil.

Lähenemine, mis on kasutatud käesolevas lõputöös, aitab riigil hinnata, mis teelõikudel ületatakse kõige sagedamini piirkiirust ning vastavalt sellele planeerida investeeringuid teevõrgustikku ja liikluspolitsei tööd. Sarnase lähenemise kasutamine aitab muuta liiklust ohutumaks ja päästa inimeste elusid.

Lõputöö on kirjutatud eesti keeles ning sisaldab teksti 45 leheküljel, 6 peatükki, 23 joonist, 7 tabelit.

Abstract

Using Call Detail Record Data for Assessing Speeding

In the following Master's thesis author analyzes whether Call Detail Record (CDR) Data can be used for assessing speeding. For this purpose, author uses data that was published in Data for Development challenge and which describes using of Orange S.A mobile network in Senegal in 2013. Based on this data author of the thesis calculates driving speed of mobile network customers. The results are then validated using Google Maps Distance Matrix API.

The topic of the analysis is very actual because of long-term trend of increasing number of vehicles in developing countries. Higher traffic often causes higher mortality in car accidents. So, it is very important for developing countries to find cheap and effective actions that make driving safer with minimal investments. CDR data can be a good option because collecting and analyzing CDR data does not require big investments in infrastructure or hardware. CDR data is collected by mobile network operators and can be analyzed even without huge computational power. The results of CDR data analysis can be interpreted in an easy way by using maps.

Methods that are used in the following Master's thesis help country to find road sections where the maximum speed limit is exceeded the most often. According to that country can better plan investments in the road network and traffic police work. Using a similar approach can make traffic safer and help save lives.

The thesis is in Estonian and contains 45 pages of text, 6 chapters, 23 figures, 7 tables.

Lühendite ja mõistete sõnastik

ASIRT	*Association for Safe International Road Travel*
CDR	*Call Detail Record*, andmekirje mis tekib telefoni andmeside vahendusel
D4D	*Data for Development*, Orange S.A. korraldatud väljakutse CDR andmetega
GDPR	*General Data Protection Regulation*, isikuandmete kaitse seadus
GPS	*Global Positioning System*, globaalne sateliitnavigatsiooni süsteem
IMEI	International Mobile Equipment Identity, rahvusvaheline mobiiltelefoni seadme kood
NMEA	*National Marine Electronics Association*
LTE	*Long Term Evolution,* telekommunikatsiooni standart, mis võimaldab tõsta mobiilside kiirust kuni 10 korda võrreldes 3G-ga
VoIP	*Voice Over Internet Protocol*, telefoniside interneti vahendusel

Sisukord

Jooniste loetelu

Tabelite loetelu

1 Sissejuhatus

Käesoleva lõputöö teemaks on mobiiltelefoniside andmete kasutamine piirkiiruse ületamise hindamiseks. See teema on saanud väga aktuaalseks kuna tänapäeval peaaegu iga inimene kasutab mobiiltelefoni ja mobiiltelefonide kaudu kogutavate andmete hulk pidevalt kasvab. Muuhulgas koguvad mobiilsideoperaatorid informatsiooni kõnede kohta, mis on tähtis arvete koostamiseks. See andmestik sisaldab lisaks helistaja ja vastaja numbritele ja kõne toimumise ajale ka infot lähimate antennide kohta.

1.1 Uuringu taust

Kuna andmete kogumine, salvestamine ja töötlemine läheb ajas üha soodsamaks, siis üha rohkem pööratakse tähelepanu valdkondadele, kus kombineerides erinevaid andmeid või kasutades andmeid „loomingulisemalt“ saab teha analüüse, mis muidu maksavad oluliselt rohkem. Üheks näiteks on ka käesolev teema, mille eesmärk on hinnata, kuivõrd hästi sobib CDR andmestik piirkiiruse ületamise hindamiseks. Juhul, kui selline lähenemine õigustab ennast, siis see võib olla kasutatud hindamaks, kuivõrd õigustatud on kiiruskaamerate paigaldamine ühel või teisel teelõigul. Samuti võimaldab selline lähenemine hinnata autojuhtide sõidustiili ja -kultuuri.

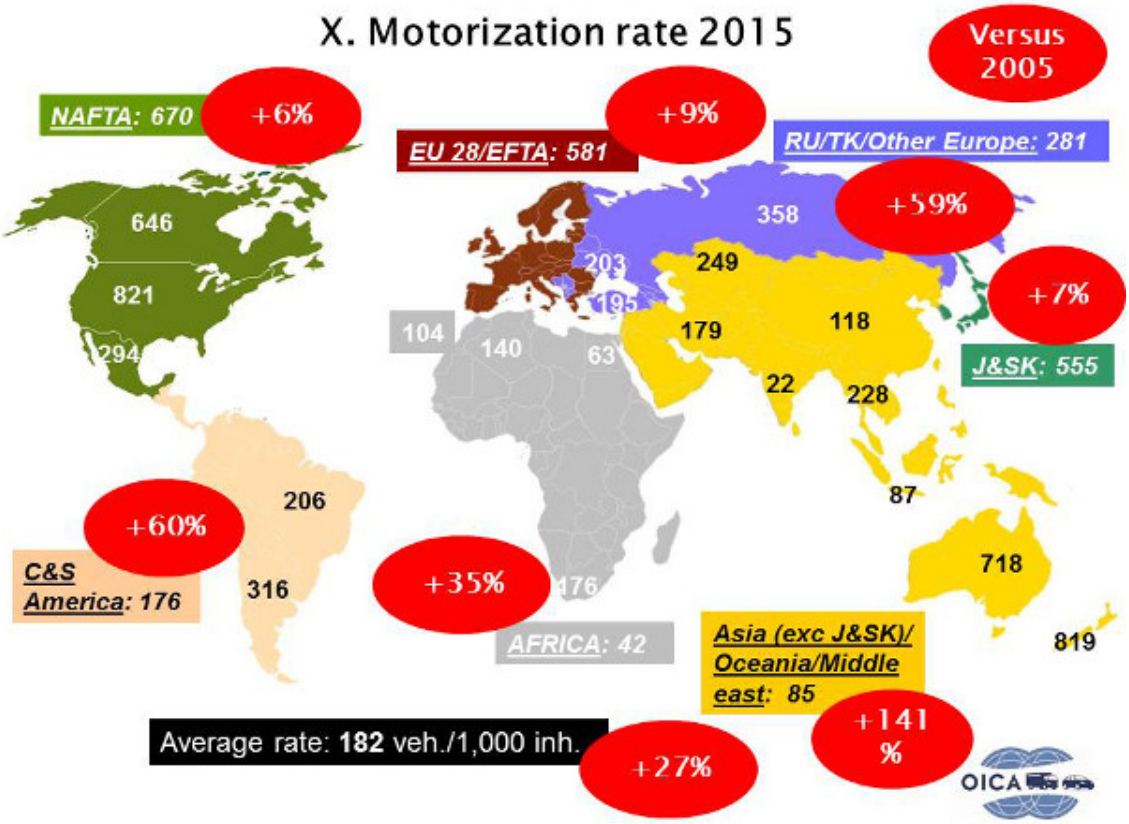
Alates aastast 2011 Ericsson avalikustab igal aastal uuringu „Ericsson Mobility Report“, mis on olnud peamine andmeallikas mobiiltelefonide maailma trendidest ja tulevikust. Selle andmetel oli 2019. aasta esimese kvartali lõpu seisuga maailmas umbes 7,9 miljardit mobiilset lairibaühenduse abonenti. Kusjuures kõige kõrgem elanike kaetus mobiilside teenusega on Kesk- ja Ida-Euroopas (141%) ja Lääne-Euroopas (122%). Isegi arenevates riikides on paljudel elanikel olemas tänapäeval mobiilside. Näiteks, India regioonis on 85% elanikkonnast kaetud mobiilsideoperaatorite poolt, Aafrikas on see näitaja 81%. Kusjuures iga aastaga kasvab ka LTE 4G ühendusega abonentide arv, ja 2019. aastal on see jõudnud 47%-ni. (Ericsson Mobility Report, 2019)

Kahtlemata, see fakt, et 104% maailma elanikkonnast on juba kaetud mobiilsidevõrkude teenuseoperaatoritega, on veidi petlik. Kindlasti on nende abonentide

seas ka mitteaktiivseid kliente, samuti mõnedel klientidel on olemas mitu lepingut mobiilsideoperaatoritega. Kuid maailm liigub selles suunas, et üha rohkem inimesi saavad kasutada mobiiltelefoni ja üha rohkemate inimeste kohta saab CDR andmete põhjal teha järeldusi nende liikumisharjumuste kohta.

ASIRT andmetel hukkub igal aastal maailmas umbes 1,25 miljonit inimest liiklusõnnetustes. Lisaks sellele igal aastal kuni 20-50 miljonit inimest saavad liiklusõnnetustes vigastada. Kuni 90% nendest liiklusõnnetuste ohvreid saavad surma või vigastada just arengumaades, kus on ainult 50% kõikidest transpordivahenditest. Liiklusõnnetused maksavad arengumaadele kuni 65 miljardit USD aastas. (ASIRT Road Safety Facts 2019) Muidugi ei ole kõik liiklusõnnetused seotud ainult kiiruse ületamisega, kuid sellega võitlus aitab parandada autojuhtide kultuuri ja leevendada ka teisi probleeme (joobes auto juhtimine, turvavöö kasutamine jne).

Lisaks mobiiltelefoni kasutuse ja liiklusõnnetuste statistikale on tähtis pöörata tähelepanu autostumise statistikale. Vaatamata sellele, et kuni 90% liiklusõnnetustes saadud vigastustest ja surmadest leiab aset just arengumaades, hetkel ainult 50% kõikidest liiklusvahenditest on kasutusel arengumaadel. (ASIRT Road Safety Facts 2019) Kuid joonisel 1 on näha, et arenevate riikide autostumine toimub oluliselt kiiremates tempodes kui arenenud riikides. Näiteks, ainuüksi Aafrikas on kümne aastaga (2005-2015) autode arv 1000 inimeste kohta kasvanud 35%. Sellest veel kiiremat kasvu on näidanud Venemaa (kasv 59%), Lõuna-Ameerika (60%) ning Aasia ja Okeaania regioon (+141%). Võrreldes Euroopa (9%) ja Põhja-Ameerika (6%) tulemusega on see väga märgatav kasv, ja võib eeldada et see trend jätkub. Seoses sellega võib liikluse turvalisuse teema muutuda veel aktuaalsemaks lähitulevikus, ja seda eelkõige arenevates riikides.



Joonis 1. Autostumise tase erinevates maailma regioonides

Allikas: www.oica.net

Statistika näitab, et liikluse ohutuse tagamine on väga aktuaalne teema, ja seda eriti arengumaades. Kuna autode arv arenevates riikides kasvab keskmisest oluliselt kiiremini ja hetkel liikluse kultuur ja infrastruktuur ei taga soovitud tasemel autoliikluse ohutust, siis peavad arenevad maad pöörama rohkem tähelepanu sellele teemale.

See näitab käesoleva lõputöö teema aktuaalsust. Kuna isegi arenevates riikides elanikkonna kaetus mobiiltelefonisidega on päris kõrge (üle 80%), ja mobiilside operaatorid koguvad oma tööprotsessides CDR andmeid, siis kavatseb lõputöö autor kontrollida, kas CDR andmed sobivad selleks, et leida potentsiaalselt ohtlikud teelõigud, kus inimesed ületavad piirkiirust kõige sagedamini. Kui selline lähenemine ennast õigustab, siis saab sarnast algoritmi ja lähenemist kasutada selleks, et teadlikumalt investeerida täiendavad summad liiklusohutusse (kiiruskaamerate, patrullide ja muude vahendite näol).

1.2 Uuringu probleem ja küsimused

Käesoleva lõputöö keskseks uuringu teemaks on CDR andmete kasutamine piirkiiruse ületamise hindamiseks. Peamine probleem, mida soovib lõputöö autor käesolevas uuringus lahendada, seisneb selles, et puudub teadmine, kas ja kui hästi sobivad CDR andmed piirkiiruse ületamise hindamiseks. Selleks, et lahendada antud probleemi, tuleb vastata järgmistele küsimustele:

- Mis on CDR andmete standart ja mis valdkondades on CDR andmed kasutatud?
- Kui täpselt saab määrata kliendi liikumiskiirust punktist A punkti B?

Selleks, et vastata kõikidele küsimustele, plaanib lõputöö autor kasutada kombineeritud uuringumeetodit. Kui välja töötatud lähenemine piirkiiruse ületamise hindamiseks osutub praktikas teostatavaks, siis võib see avaldada suurt mõju riigi sotsiaalmajanduslikule seisule.

1.3 Lõputöö struktuur

Käesolev lõputöö koosneb kokku kuuest peatükist:

1. Sissejuhatus. Sissejuhatuses kirjeldab lõputöö autor probleemi, mida antud lõputöö peab lahendama, räägib uuringu taustast ja püstitab küsimused, millele peab vastama teema uurimisel;
2. Teoreetilised alused. Selles peatükis annab töö autor detailsemat ülevaadet CDR andmetest ja nende kasutusvaldkondadest, põhinedes eelmistele uuringutele;
3. Varasemad uuringud. Varasemate uuringute all kirjeldab lõputöö autor uuringuid, mis on viidud läbi CDR andmeid kasutades. Uuringute kirjeldamisel pöörab lõputöö autor tähelepanu hüpoteesidele, mis varasemate uuringute raames püstitati, meetoditele, mis olid kasutatud uuringu läbiviimisel ja järeldustele, milleni on igas uuringus jõutud;

4. Metoodika. Metoodika peatükis kirjeldab töö autor käesolevas analüüsis kasutatud andmete struktuuri, lisaks sellele ka meetodid, mida ta kasutab andmete analüüsimisel;

5. Analüüs. Selles peatükis kirjeldab autor andmete analüüsimise käiku ja tulemusi, samuti valideerib töö autor saadud tulemusi;

6. Kokkuvõte. Kokkuvõttes annab töö autor ülevaadet kogu lõputööle, kirjeldades lõputöö probleemi, küsimusi, läbiviidud analüüsi ja selle tulemusi.

Selline mitmekülgne lähenemine aitab autoril uurida lõputöö teemat erinevate nurkade alt ning vastata käesolevas peatükis püstitatud küsimustele.

2 Teoreetilised alused

Käesolevas peatükis kirjutab lõputöö autor lahti CDR andmete olemusest, nende andmete kasutusvaldkondadest ja sellest, kuidas saab CDR andmeid kasutada klientide liikuvuse hindamiseks. Samuti annab töö autor ülevaadet teistest meetoditest, mida saab kasutada inimeste liikuvuse hindamiseks.

2.1 CDR andmete standart

CDR (ingl. *Call Detail Record*) on rahvusvaheline formaat, mis kirjeldab kõnede andmestikku, mida kasutavad mobiilsideoperaatorid oma igapäevategevuses. See sisaldab nii sisenevate, kui ka väljuvate kõnede infot ning on kasutatud eelkõige telefoniarvete koostamiseks. Lisaks kõnedele on sama standardi alusel kirjeldatud kõik saadetavad SMS-id. (Horak)

Lisaks mobiiltelefonivõrkudele on CDR kasutatud ka teistes valdkondades, näiteks VoIP (ingl. *Voice over Internet Protocol*), siis CDR andmestikul puudub rangelt kirjeldatud standard. Kuid üldjuhul sisaldab CDR fail järgmist infot (Forte *et al* 2010):

- kirje unikaalne number;
- helistaja telefoni number;
- kõne vastuvõtja telefoni number;
- telefoni IMEI kood;
- helistaja ja vastuvõtja antenni info;
- kõne tüüp (kas kõne või SMS);
- kõne kuupäev ja kellaaeg;
- kõne pikkus.

Nagu on eelnevalt juba öeldud, siis sõltuvalt mobiilsideoperaatori süsteemist võib CDR andmestik sisaldada ka muud infot, nagu näiteks andmed kõne eest maksva kliendi kohta, kes katkestas kõne (kas helistaja või vastuvõtja), võimalikud veakoodid, kõnet

alustanud ja katkestanud antenni kood jne. (Petersen 2000) Käesoleva lõputöö raames ei ole teised võimalikud väljad nii tähtsad, kuna neid ei saa kuidagi kasutada inimese positsioneerimise täpsustamiseks.

2.2 CDR andmete kasutusvaldkonnad

Nagu on öeldud peatükis 2.1, siis sisaldavad CDR andmed muuhulgas ka infot, mida võib käsitleda isikuandmetena (eelkõige helistaja ja vastuvõtja telefoninumbrid). Lisaks sellele saab kasutada ka muud infot erahuvides nii, et telefonivõrgu klientide huvid saavad kahjustada. See seab mõningaid piiranguid CDR andmete kättesaadavusele ja kasutamisele. Eriti aktuaalseks on see muutunud seoses sellega, et mõned uuringud on tõestanud, et ka anonüümsetest andmetest on võimalik saada kätte klientide info kätte. Üks näide sellest on kirjeldatud The Guardian artiklis, kus NYC taksofirma anonüümsetest andmetest on suudetud teha järeldusi paljude autojuhtide koduaadressitest ja sissetulekutest. (The Guardian)

Samuti seab piiranguid ka 2018. aasta 25. mail jõustunud isikuandmete kaitse seadus (GDPR). See seab piiranguid isikuandmete hoidmisele ja töötlemisele, mis teeb paljude analüüside läbiviimist oluliselt keerulisemaks. (Core DNA) Tänu suurele meediakajastusele on inimesed muutunud ka teadlikumaks oma isikuandmete kaitsmise teemas, mistõttu ei ole väga paljud ettevõtted valmis jagama CDR andmeid välispartneritega.

Kuid vaatamata piirangutele on olemas ka ettevõtted, mis leiavad võimalusi CDR andmete kättesaamiseks. Need töötavad CDR andmetega ja analüüsivad neid, ja turule on juba praeguseks hetkeks tekkinud mitu näidet (Mörner, 2017):

- Motionlogic (Saksamaa), www.motionlogic.de
- Mezuro (Holland), www.mezuro.com
- Teralytics (Šveits), www.teralytics.ch
- Airsage (USA), www.airsage.com

Lisaks CDR andmete otsesele kasutusele (arvete esitamine mobiilkõnede ja SMS-ide eest) on võimalik lähtudes CDR andmete olemuselt kasutada neid ka inimeste

liikumisharjumuste hindamiseks. Kuna CDR andmed kogutakse nö „taustal“ ja iga kord, kui inimene teeb või võtab vastu kõne või saadab SMS-i, jääb maha märge, kus ja millal see toimus, siis sobivad CDR andmed väga hästi liikumisharjumuste ja liikuvuse hindamiseks.

Üks näide CDR andmete kasutamiseks pärineb Liberiast: ITU (ingl. *International Telecommunication Union*) viis 2017. aastal uuringu, mille eesmärgiks on hinnata CDR andmete sobivust inimeste linnadevahelise liikluse hindamiseks. See teema on olnud eriti populaarne seoses Ebola levikuga. Uuringu läbiviijad jõudsid järelduseni, et CDR andmed sobivad hästi inimeste liikumise trajektooride ja võimaliku epideemia leviku ennustamiseks. (ITU 2017)

Detailsemat ülevaadet varasematest uuringutest annab töö autor 3. peatükis, kuid juba nendest näidetest on näha, et CDR andmed on populaarsed just tänu sellele, et nende kogumine on suhteliselt soodne (andmed kogutakse automaatselt) ja hõlmab palju isikuid. Muidugi on inimesi, kelle liikumisharjumusi ei saa selle meetodiga hinnata kuna need ei kasuta mobiiltelefoni või mingil muul põhjusel, kuid ilmselt ei ole ühtegi meetodit mis oleks sama efektiivne nii madala kulu juures.

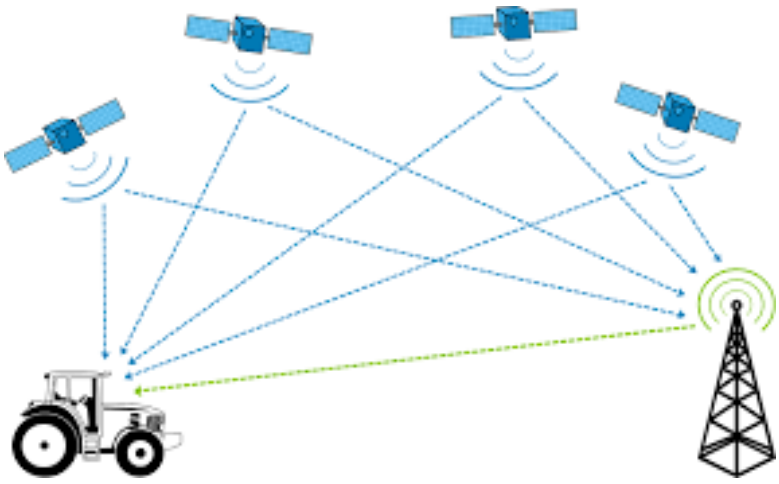
2.3 Inimeste liikumisharjumuste hindamine

Käesoleva lõputöö raames keskendub töö autor just CDR andmetele ning ei kasuta teisi andmete tüüpe liikuvuse hindamiseks. Käesolevas peatükis loetleb aga autor teisi tehnoloogiaid, mida saab kasutada inimeste liikuvuse hindamiseks. Käesoleva ja sarnaste uuringute kontekstis saab neid tehnoloogiaid kasutada näiteks tulemuste täpsustamiseks ja valideerimiseks.

2.3.1 GPS

Kui rääkida inimeste liikumiste jälgimisest, siis ei saa jätta mainimata GPS-i, mis on levinud ja populaarne kogu maailmas. See süsteem on välja töötatud 1960-ndatel aastatel Ameerika Ühendriikide sõjaväe jaoks. Süsteem koosneb 24 satelliidist, mis ringlevad ümber maakera umbes 19 300 km kõrgusel. Kõik satelliidid liiguvad niimoodi, et igas maailma punktis on otseses nähtavuses vähemalt 4 satelliiti, mida on vaja et määrata seadme asukohta (vt. joonis 2). Iga satelliit edastab andmed oma hetke asukoha ja orbiidi kohta koos täpse kellaajaga. GPS vastuvõtja (telefon või muu seade) saab

andmeid neljast või rohkemast satelliidist ning leiab nende põhjal välja oma asukohta. (TechTerms)



Joonis 2. GPS põhimõte
Allikas: https://www.maxpixel.net/Scheme-Navigator-Satellite-Gps-1826792

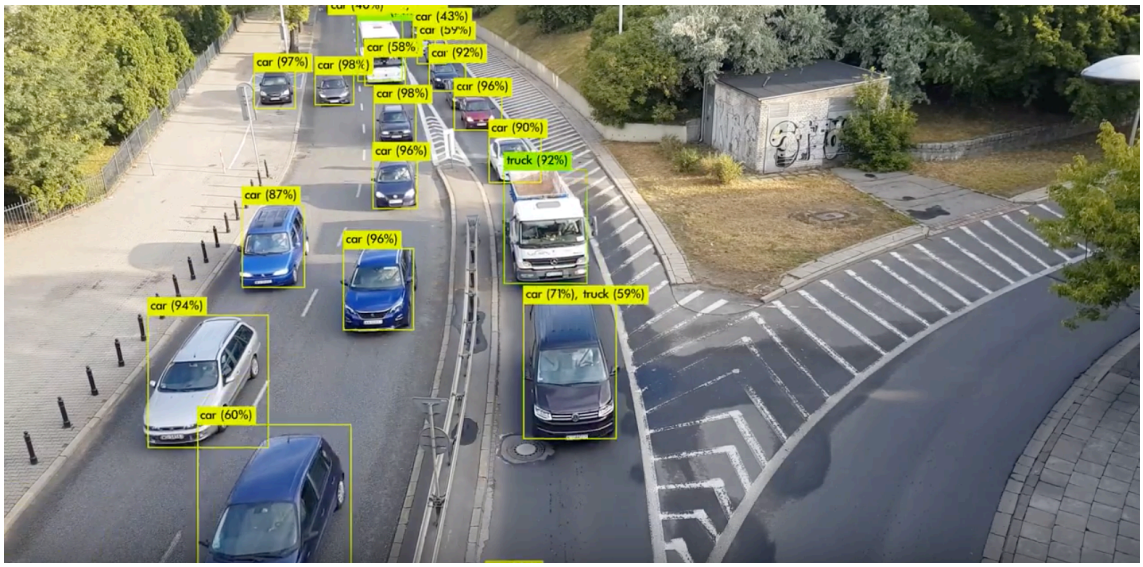
Meedias on läbi käinud ettepanekuid kasutada GPS seadmeid autojuhtide piirkiiruse ületamise jälgimiseks. Idee seisnes selles, et autosse on võimalik paigaldada väike GPS seade, mis kontrollib pidevalt auto sõidukiirust ja võrdleb seda lubatud piirkiirusega. Kui seade märgab, et piirkiirus on ületatud tahtlikult (üle 20 sekundit peale hoiatust), siis saab autojuht trahvi kiiruse ületamise eest. (Eesti Päevaleht, 2009) Hoolimata sellest, et see aitaks tõsta turvalisust teedel, loobuti sellest ideest, kuna selle süsteemi kasutuselevõtt on suhteliselt kulukas. Lisaks sellele kerkisid üles küsimused GPS täpsuse ja turvalisuse kohta.

Käesoleva lõputöö kontekstis aga GPS on hea võimalus kontrollida saadud tulemuste täpsust ning hinnata selle sobilikkust. Selleks on võimalik kasutada tarkvara, mis salvestab seadme asukohta NMEA formaadis. Sellest failist on võimalik lugeda lisaks asukohale ka kiirust sõlmedes, mida saab kasutada tulemuste valideerimisel. (GPS World) Samas aga ei sobi GPS käesoleva lõputöö ülesande lahendamiseks, kuna andmete kogumine toimub kasutaja seadmes. Seega paljude inimeste liikumisharjumuste hindamiseks on tarvis jagada neile eraldiseisvad seadmed, mis koguvad ja edastavad infot nende kiiruse ja liikumisteede kohta. See on ilmselt peaaegu võimatu, ehkki selline lahendus oleks täpsem. CDR andmeid omakorda kogutakse keskselt mobiilsideoperaatorite poolt.

2.3.2 Andurid ja kaamerad

Veel üks viis inimeste liikumisharjumuste jälgimiseks on kaamerad ja andurid. Targa linna osana, kasutatakse neid selleks, et hinnata erinevate tänavate koormust, liikluse koosseisu jne. Üks sarnane projekt on ette võetud ka Tallinnas, kus Tallinna Tehnikaülikool koostöös Thinnect OÜ-ga arendavad tehnoloogiat, mille abil on võimalik jälgida linnaõhu kvaliteeti ja mõõta liiklusvoogusid. Tegemist on sensoritega, mis suudavad muuhulgas arvestada sellega, mitu jalakäijat ja autot läbib sellest ning mis suunas. Tehnoloogiat saab kasutada linnaplaneerimises transpordivoogude ennustamiseks ja liikluse turvalisuse parandamiseks. (TalTech 2019)

Üheks kaasaegse linna lahutamatuks osaks on saanud ka liikluskaamerad. Kasutades pilti kaameratest ja masinõppe algoritmi, on võimalik analüüsida transpordivoogu ning saada aru, mis transpordivahendid läbivad konkreetset teelõiku. Pythoni ja OpenCV abil on võimalik saada suhteliselt detailset ülevaadet liiklusest (vt joonis 3). (GeeksforGeeks)



Joonis 3. Transpordivoo analüüs liikluskaamera ja masinõppe abil
Allikas: Karol Majek, www.youtube.com

Nii sensorite, kui ka liikluskaamerate abil on võimalik analüüsida transpordivoogusid, autode kiirust ja transpordivahendite tüüpe. Kuid need on oluliselt kallimad võrreldes CDR andmete kasutamisega. Seoses sellega kavatseb töö autor hinnata CDR andmete sobilikkust potentsiaalselt ohtlike teelõikude leidmiseks.

3 Varasemad uuringud

Vaatamata piirangutele, mis on seotud isikuandmete jagamise ja analüüsimisega, CDR andmete kasutamine uuringute läbiviimiseks on väga populaarne kogu maailmas. Eelkõige on need uuringud seotud inimeste liikuvuste hindamisega. Käesolevas peatükis annab lõputöö autor ülevaadet varasematest uuringutest, mis on hiljuti läbi viidud.

NetMob on peamine konverents, mis on pühendatud mobiilsete andmete teaduslikule analüüsile. 2017. aastal osales konverentsil üle 250 inimese 30 erinevast riigist. Konverentsil presenteeritakse uuringuid, mis on tehtud mobiiltelefoni andmete põhjal, kaasa arvatud CDR, mobiiltelefoni asukoha-, wifi, äppide poolt genereeritud, Facebooki ja Twitteri andmete põhjal. 2019. aastal toimus NetMob konverents juba kuuendat korda ja see toimus Oxfordi ülikoolis. NetMob on peamine konverents, kus presenteeritakse uuringuid, mis lahendavad ühiskondlikke, sotsiaalseid, tööstuslikke ja urbanistlikke probleeme mobiilsete andmete abil. (NetMob 2019)

3.1 Data for Development väljakutse, Côte d'Ivoire 2013

Orange S.A. avalikustas 2013. aastal väljakutset „Data for Development (D4D)“, mille eesmärgiks on lahendada ühiskonna arenguküsimusi, aidates Côte d'Ivoire'i elanike sotsiaalmajanduslikule arengule ja heaolule. Antud andmestiku vaatlusperioodiks on 5 kuud (01.12.2011 – 28.04.2012) ja see koosnes 4 erinevast andmebaasist(Blondel *et al*):

- Tunnipõhine liiklus antennide vahel;
- 50 000 kliendi individuaalsed liikumistrajektoorid kahenädalase perioodi sees, koos antennide andmetega;
- 500 000 kliendi individuaalsed liikumistrajektoorid kogu vaatlusperioodi sees, koos regioonide andmetega;
- 5 000 kliendi suhtlusgraafiku näidis.

Lisaks eeltoodule pakkusid väljakutse korraldajad kasutada ka teisi andmeid, pakkudes selleks 32 avalikku andmestikku, mis muuhulgas sisaldas World Trade Organization, World Bank Data, United Nations Data, The World Factbook ja teisi andmeallikaid. Lisaks sellele said kõik osalejad ka andmete detailset kirjeldust ja näiteid andmete kasutamisest. Kokku valis välja 7-liikmeline kommitee 74 uuringut, mis said avalikustatud ja hinnatud selle projekti raames. Kõik uuringud on grupeeritud valdkondade kaupa (Blondel *et al*):

- Sotsiaalne ja majanduslik areng;
- Andmekaeve;
- Mobiilsus ja transport;
- Tervis ja epideemiad.

Lõputöö autor toob välja teemad just sellepärast, et need on ühtlasi ka populaarsed uuringute läbiviimisel CDR andmete põhjal. D4D väljakutse võitjaks osutus terviseprobleemidele pühendatud uuring, kuid lisaks võitjale kuulutati välja veel kaks uuringut: parim visualiseerimine ja parim arendus. (UN Global Pulse, 2013)

3.1.1 Parim uuring: *"Exploiting Cellular Data for Disease Containment and Information Campaigns Strategies in Country-Wide Epidemics"*

D4D väljakutse võitjaks osutus uuring, mis on läbi viidud Birmingham ülikooli arvutiteaduse teaduskonna andmeteadlaste poolt. Uuringu autorid töötasid välja mudeli, mis näitab kuidas haigused levivad riigi territooriumil epideemiate puhul. Selleks on kasutatud CDR andmetest välja arvutatav inimeste liikumisharjumuste mudel. Seejärel teadlased simuleerisid erinevaid stsenaariume ja hindasid, kuidas saab haiguste levikut tõkestada erinevate mehhanismide abil, tuginedes selles inimeste liikuvuse ja sotsiaalsete sidemete infole. (Lima *et al*)

Uuringu järelduseks on see, et liikumiste tõkestamine ei aita kaasa haiguste leviku peatamisel. Kuid samaaegselt selgus, et üks-ühele teavituskampaania telefoni teel võib osutuda väga tõhusaks vastumeetmeks haiguste epideemiate puhul. Lisaks sellele leiavad autorid, et välja töötatud mudel võib saada veel täpsemaks kui kasutada ajast sõltuvad maatriksid. Lisaks sellele on erinevates inimgruppides erinevad kontakti määrad, ja töö

autorid näevad et võimalusel tuleks kasutada ka seda infot, kuna haiguse levik erinevates inimgruppides võib olla erinev. Edasise arenguna nähakse ka võimalust katsetada hübriidmudeleid, nagu näiteks samaaegsed osalised liikumispiirangud ja üks-ühele teavituskampaaniad. (Lima *et al*)

3.1.2 Parim visualiseerimine: *"Exploration and Analysis of Massive Mobile Phone Data: A Layered Visual Analytics Approach"*

Selle rahvusvahelise koostööprojekti tulemusena on sündinud interaktiivse kasutajaliidesega tarkvara (vt joonis 4), mis suudab töötada läbi 2,5 miljardit andmerida. See lahendus võimaldab kasutajal otsida mustreid kõnede toimumise ja SMS-ide saatmise andmestikust ja kasutada neid edasistes uuringutes. Üks peamisi väljakutseid, mida pidi lahendama, on andmete hulk: arhiveerimata kujul andmed võtavad 33,5 GB, mis on ilmselgelt liiga palju operatiivmälus hoidmiseks. (S. van den Elzen *et al*)

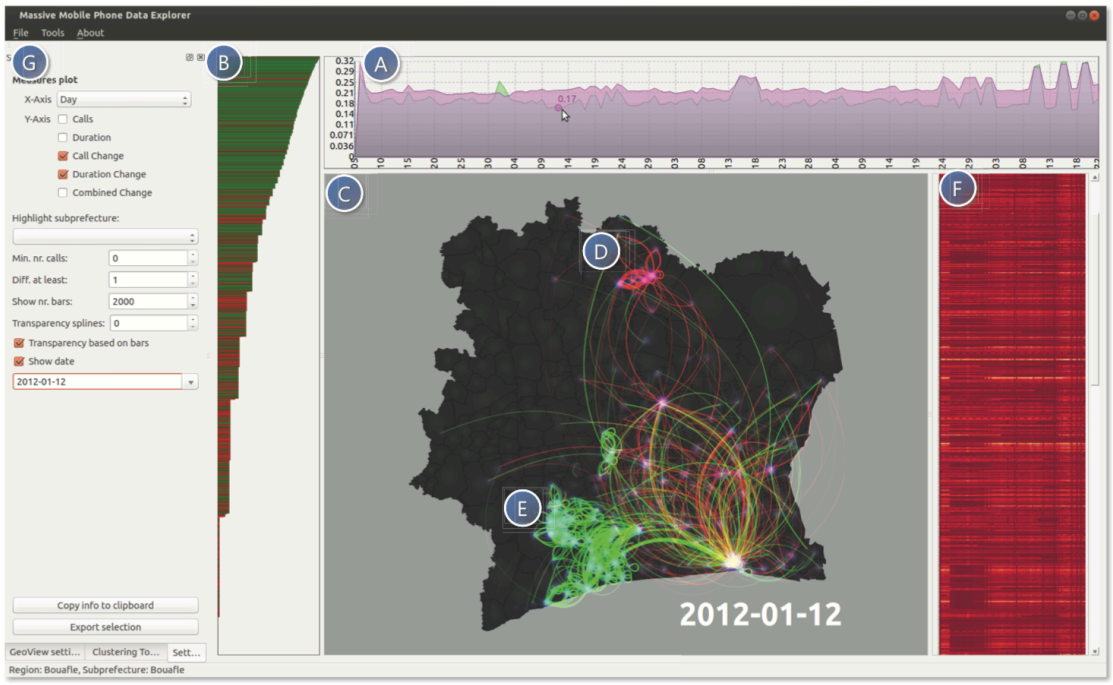


Fig. 2. Graphical user interface: providing high level overview of different measures such as the number of calls and call change over time in the measure overview (A). Individual contributions of communication channels to the measure at selected time interval are shown in the measure contribution view (B) and in the geospatial view (C) revealing localized call change behavior: large decrease (D) and large increase (E) of the number of calls. The matrix view (F) provides a high level overview of call behavior over time for the individual towers. Different settings and algorithmic support are offered by according controls (G).

Joonis 4. Kihiline visuaalne analüüs – kasutajaliides

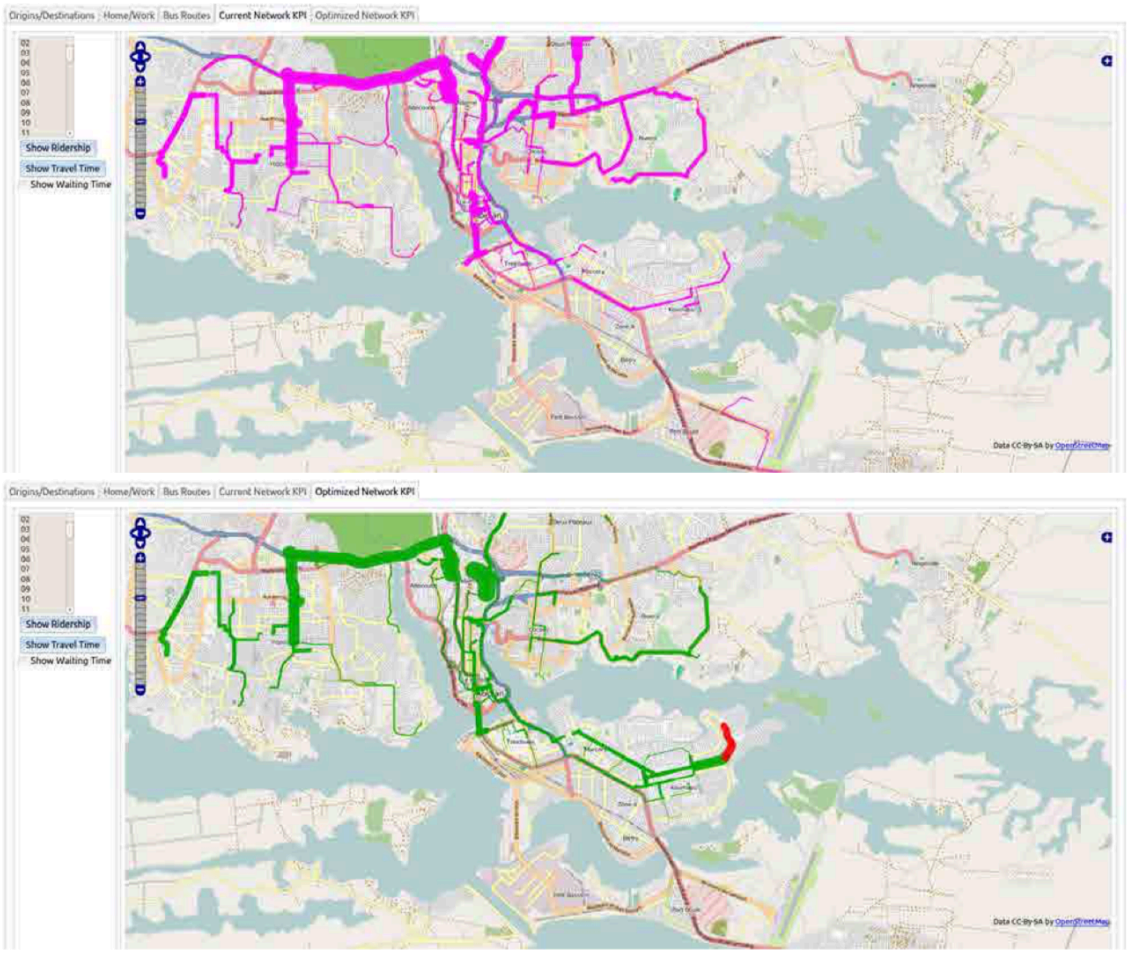
Allikas: S. van den Elzen *et al*

Uuringu läbiviimisel meeskond keskendus kõnede toimumise mustritele ja leidis, et ühe riigi raames kõnede arv suureneb ja väheneb sõltuvalt erinevatest sündmustest.

Andmeteadlased leidsid, et on kaks tüüpi sündmusi, mida saab tuvastada mobiiltelefoni andmete muutuse abil. Esimeses grupid on sündmused, mis on seotud ilmastikuga (näiteks, tugevad vihmad, mis mõjutavad eelkõige riigi kakao piirkondi). Teiseks on ühiskondlikult olulised sündmused (näiteks, Uue Aasta tähistamine). Ühiskondlikult oluliste sündmuste tuvastamine võimaldab riigil vajadusel varem sekkuda ja tegeleda probleemide lahendamisega. (UN Global Pulse, 2013)

3.1.3 Parim arendus: *"AllAboard: a System for Exploring Urban Mobility and Optimizing Public Transport Using Cellphone Data"*

Selle uuringu autorid lähtusid hüpoteesist et CDR andmeid on võimalik kasutada et paremini planeerida linnalogistikat. Kuna mobiiltelefoni kasutajate arv on võrreldav kogu elanikkonna suurusega, siis kasutades mobiiltelefonide poolt genereeritud CDR andmeid on võimalik suhteliselt täpselt hinnata inimeste liikumisharjumusi. Seda teadmist otsustasid kasutada oma uuringus IBM Research andmeteadlased Dublinist. Selleks nad kasutasid D4D väljakutse raames avalikustatud andmestikku nr 2, mis koosnes 50 000 inimese andmetest. Andmestiku struktuuriks on olnud <User ID, Day, Time, Antenna>, ehk see andmebaas näitab mis hetkel keegi 50 000 inimesest on saanud või saatnud SMS-i, teinud või võtnud vastu kõne ning kus on see toimunud (antenni koordinaatide täpsusega). (Berlingerio *et al*)



Joonis 5. Côte d'Ivoire ühistranspordi skeem (üleval) ja parandatud liiklusvood (all)
Allikas: Berlingerio *et al*

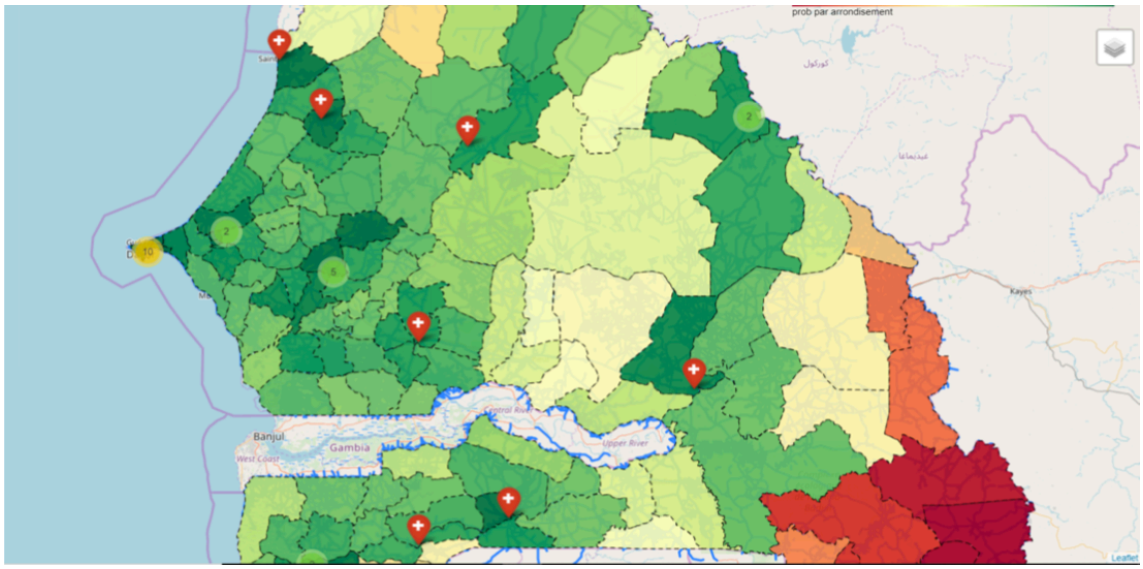
Uuringu käigus andmeteadlased võrdlesid inimeste liikumisharjumusi Côte d'Ivoire'i linnatranspordi (SOTRA) võrgustikuga. Nad arvutasid välja, kuidas oleks võimalik panna ühistransport liikuma nii, et elanikel kuluks vähem aega punktist A punkti B jõudmiseks. Uuringu tulemus on näidatud joonisel 5. Tulemuste hindamisel jõudsid Dublini teadlased järelduseni, et suhteliselt madalate ümberkorralduskuludega on võimalik vähendada keskmise linnaelaniku transiidi aega kuni 10,2%. Uuringu autorid toovad välja ka sarnaste uuringute piiranguid, milleks on (1) mobiilside operaatorite turuosad ja võimalik erinevus sihtgruppide vahel, (2) võimalus et kasutajate valim ei ole juhuslik, (3) kõnede tariifid võivad mõjutada klientide kõnede tegemist ja (4) igal inimesel võib olla kaasas ka mitu seadet ja sellise inimese käitumismuster saab seeläbi suurema osakaalu. Nende piirangutega peab kindlasti arvestama sarnaste uuringute läbiviimisel. (Berlingerio *et al*)

3.2 Data for Development väljakutse, Senegal 2014

Peale esimese D4D väljakutse lõppemist, avalikustas Orange S.A. 2014. aastal teist korda D4D väljakutset, mille raames jagati andmeteadlastega Senegali CDR andmeid. 2015. aastal järjekordselt toimuval NetMob konverentsil tutvustati ka 20 uuringut, mis on läbi viidud Senegali CDR andmete põhjal. Sellel korral jagati kõik uurimistööd 6 erinevasse kategooriasse: (1) põllumajandus, (2) energia, (3) tervis, (4) riiklik statistika, (5) transport/linnaelu ja (6) muu. (NetMob 2015) Teemade valik annab järjekordselt kinnitust sellele, et CDR andmete kasutusvaldkond on väga lai ja mitmekesine. Käesolevas peatükis aga keskendub töö autor kahele analüüsile, mis on teostatud samade andmete põhjal kuid üks nendest on teostatud 2019. aastal, väljaspool D4D väljakutset. Põhjuseks, miks võtab töö autor ette just neid uuringuid, on see, et lõputöö autor kavatseb kasutada oma töös samu andmeid ja sarnast lähenemist.

3.2.1 Early public health emergency anticipating with non health data per se

Uuringu autorid lähtusid ÜRO säästva arengu eesmärkidest nr 3 (tervisliku elu tagamine ja heaolu edendamine igas vanuses) ja nr 10 (ebavõrdsuse vähendamine riikide sees ja riikide vahel). Uuringu eesmärkideks on (1) disainida lahendus anonüümsete CDR andmete kasutamiseks rahvastiku tervise küsimustes, (2) tõendada selle lähenemise sobilikkust insulti ja meningiidi analüüsimisel ja (3) disainida visuaalsed raportid rahvatervise eest vastutavate töötajate jaoks. (Semassel *et al*)



Joonis 6. Ellujäämise tõenäosus insuldi puhul erinevates regioonides

Allikas: Semassel *et al*

Uuringu eesmärkide täitmiseks on kasutatud D4D väljakutse raames avalikustatud Senegali CDR andmed. Autorid lähtusid reeglist, et insuldi puhul inimene peaks saama arstlikku abi maksimaalselt 3 tunni jooksul. Lähtudes sellest arvutasid nad iga regiooni puhul, kui kaua läheb keskmiselt elanikul aega, et jõuda lähima haiglani ja vastavalt sellele arvutasid välja tõenäosus taastuda peale insulti. Tulemused on toodud joonisel 6. (Semassel *et al*)

Meningiidi puhul lähtusid töö autorid sellest, et kõige kõrgema riskiga on need kohad, kus koguneb kõige rohkem inimesi. Kasutades CDR andmeid, suutsid autorid leida need regioonid ja identifitseerida teoreetiliselt ohtlikud regioonid, kus meningiidi levik oleks kõige lihtsam. Uuringu tulemuseks on töötatud välja kaardid, mis näitavad, kuhu on kõige otstarbekam rajada uued haiglad. (Semassel *et al*)

3.2.2 Where do we Develop? Discovering Regions for Urban Investment in Senegal

Linnastumine ehk linnade suuruse kasv on kahtlemata viimaste aastate trend. Kuna üha rohkem ja rohkem inimesi kolib linnadesse, siis juba tänapäeval suuremad linnad seisavad silmitsi piiratud ressursside, industrialiseerimise ja vähearenenud infrastruktuuri probleemidega. Seoses sellega üha rohkem eeldatakse, et riik teeb investeeringuid väiksematesse linnadesse, et meelitada uusi elanikke sinna, vähendades selle kaudu suurimate linnade koormust. Uuringu autorid töötasid välja meetodeid CDR andmete kasutamiseks linnade planeerimisel. (Doran *et al*)

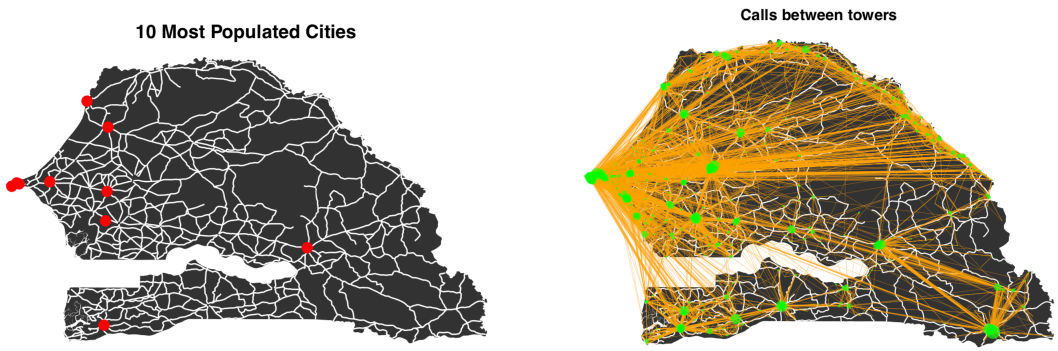


Figure 6: Spatial comparison of the most populated cities in Senegal and Pagerank centrality of calling towers

Joonis 7. 10 kõrgelt asustatud linna ja linnade tsentraalsus (CDR andmete põhjal)
Allikas: Doran *et al*

Uuringu autorid arvutasid välja erinevate Senegali regioonide jaoks näitajad, mis aitavad paremini hinnata linnade atraktiivsust ja populaarsust. Muuhulgas määrati iga linna kohta selle linna elanike kõnede arvu, linna iseseisvust, linna elanike suhtlust teiste linnade elanikega ja tsentraalsust. Senegali 10 kõige tihedamini asustatud linna ja suuremate linnade tsentraalsus on näidatud joonisel 7. (Doran *et al*)

Uuringu tulemusena grupeerisid andmeteadlased kõik Senegali linnad 4 gruppi: (1) Dakar ja lähiümbrus, (2) keskmised linnad – 9 linna, mis on suhteliselt sõltumatud ja on suure tsentraalsusega, (3) väiksed kohad – 27 linna, mis sõltuvad teistest linnadest ja (4) kesk-väiksed kohad, mille hulka kuulub suurem enamus linnu. Kõige suuremat arengupotentsiaali nähakse just teises grupis, näiteks sellistes linnades nagu Thies, St Louis, Mbour ja Ziguinchor. (Doran *et al*)

3.3 Järeldused varasematest uuringutest

Peale tutvumist eeltoodud uuringutega, jõudis lõputöö autor kahe järelduseni, mis võivad olla vajalikud käesoleva lõputöö raames:

- CDR andmete kasutusvaldkond on väga lai, kuna tegemist on laialt levinud andmetega, siis neid saab kasutada paljude probleemide lahendamiseks;
- CDR andmed ei ole nii täpsed nagu näiteks GPS, näiteks madala asustusega regioonides tänu sellele, et mobiilsideantennide võrgustik ei ole väga tihe, toimub asukoha määramine CDR andmete põhjal kuni 2-3 km veapiiriga.

Lõputöö autor arvestab nende järeldustega analüüsi läbiviimisel ja kindlasti pöörab tähelepanu ka analüüsi tulemuste valideerimisele, et hinnata CDR andmete eripäradest tulenevat viga.

4 Metoodika

Käesolevas peatükis kirjeldab töö autor andmeid, mida kavatseb kasutada käesolevas lõputöös. Eelkõige kirjeldab töö autor algandmete struktuuri ja sisu ja seejärel kirjeldab andmete ettevalmistamise protsessi. Samuti annab autor käesolevas peatükis esmase hinnangu andmete sobilikkusele analüüsi läbiviimiseks.

4.1 Andmete kirjeldus

Käesolevas lõputöös kasutab töö autor andmeid, mis on jagatud 2014. aasta D4D väljakutse raames Orange S.A. poolt. Teise väljakutse raames jagati CDR andmeid Senegalist ja see on korraldatud koostöös UN Global Pulse, Gates Foundation, World Economic Forum, MIT ja teiste organisatsioonidega. Kõik osalised said kasutada oma analüüsides 5 erinevat andmestikku, mis on ette valmistatud Orange S.A. poolt. (UN Global Pulse)

Käesolevas lõputöös keskendub töö autor kahele andmestikule, milleks on antennide andmed ning kõnede ja SMS-ide andmed. Esimese andmestiku põhjal saab töö autor leida iga antenni asukoha (koordinaadid kaardil) ja teist andmestikku kasutab autor inimeste liikumisteede hindamiseks.

Tabel 1. Antennide andmed

timestamp	count	site_id	arr_id	lon	lat
2013-01-19 17:40:00	71	165	4	-17.449754	14.679568
2013-01-14 18:10:00	28	97	1	-17.460410	14.729414
2013-01-08 18:00:00	79	235	4	-17.436481	14.672241
2013-01-08 19:30:00	31	281	3	-17.426557	14.736299
2013-01-20 10:50:00	15	474	10	-17.232490	14.690125

Allikas: D4D väljakutse, 2014

Esimene andmestik näitab mobiilside antennide koormust erinevatel ajahetkedel. Andmete näidis on toodud tabelis 1. Need andmed on vajalikud just selleks, et saada aru, kus asuvad erinevad antennid ja kui koormatud need on. Iga antenni kohta on teada lisaks selle koordinaatidele (pikkuse ja laiuse kraadid) ka regioon, kuhu antenn kuulub. Lühidalt saab andmete struktuuri kirjeldada niimoodi (Montjoye *et al*):

- timestamp – kuupäev ja kellaaeg, mille kohta edasine info käib;
- count – telefonide arv, mis antud hetkel antenni kasutavad;
- site_id – antenni kood;
- arr_id – antenni regiooni kood;
- lon, lat – antenni asukoha pikkuse ja laiuse kraadid.

Andmete ettevalmistamise etapil kustutas töö autor veerud „*timestamp*“ ja „*count*“, kuna tulevases analüüsis kavatseb ta kasutada neid andmeid ainult antenni asukoha määramiseks. Peale duplikaatide eemaldamist jäi andmestikku 1 495 unikaalset antenni. Antennide võrgustik on kirjeldatud täpsemalt peatükis 4.2. Siinkohal peab mainima, et D4D väljakutse raames on avalikustatud andmestik 1 666 antenni andmetega, kuid käesolevas lõputöös on kasutatud ainult kahe nädala andmed (7. jaanuarist 20. jaanuarini 2013), seega osa antennidest, mis ei olnud aktiivsed sel perioodil, et ole käesolevas lõputöös kajastatud.

Tabel 2. Kõnede ja SMS-ide andmed

user_id	timestamp	site_id
1	2013-01-07 13:10:00	461
1	2013-01-07 17:20:00	454
1	2013-01-07 17:30:00	454
1	2013-01-07 18:40:00	327
1	2013-01-07 20:30:00	323

Allikas: D4D väljakutse, 2014

Teises andmestikus (vt tabel 2) on toodud info mobiilsideoperaatori klientide kõnede info. Kõik andmed on toodud anonüümsel kujul ja tabeli koosneb kolmest veerust (Montjoye *et al*):

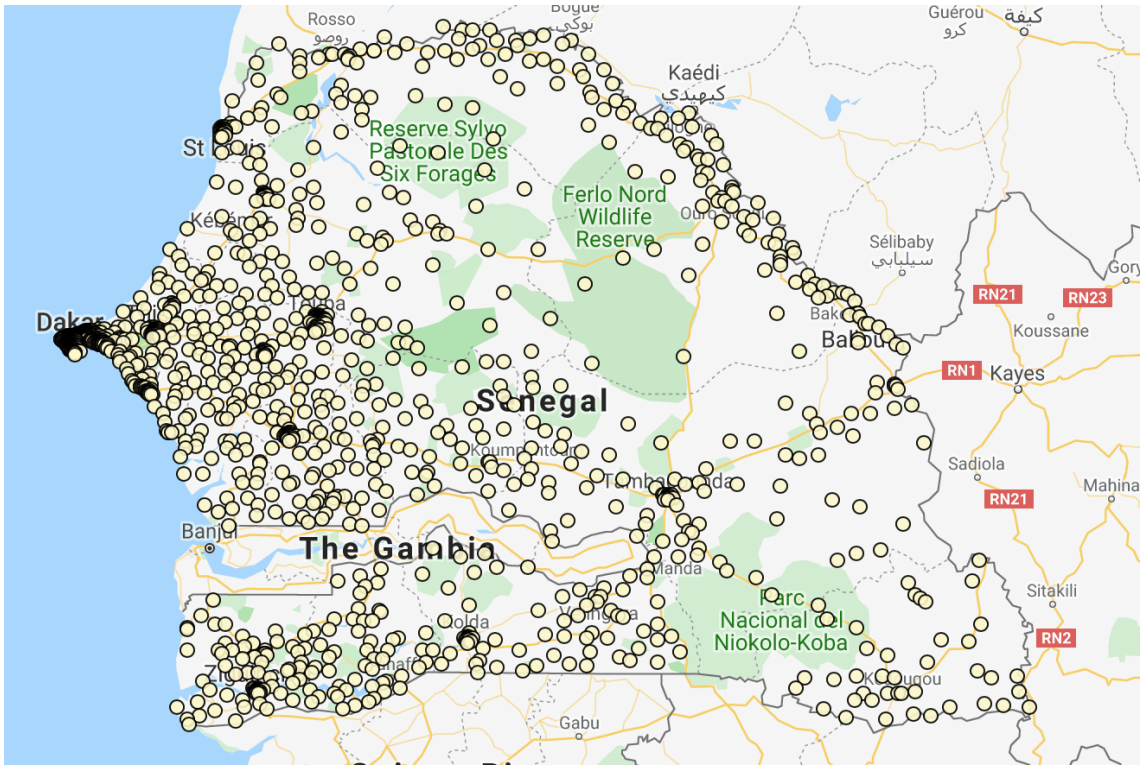
- user_id – mobiiltelefoni kasutaja unikaalne kood;
- timestamp – kuupäev ja kellaaeg, millal toimus kõne;
- site_id – antenni kood, mille levis asus klient kõne tegemise hetkel.

Käesolevas analüüsis kasutab töö autor ainult osa avalikustatud andmestikust. Kõik edaspidi toodud numbrid pärinevad perioodist 07.01.2013 – 20.01.2013. Põhjuseks on eelkõige andmete maht: teises andmetabelis on kokku 44 403 074 kirjet, suuremate andmehulkade töötlemine tavalise arvuti abil võib osutuda problemaatiliseks. Samas aga annab 2-nädalane periood piisavalt hea ülevaate CDR andmete sobilikkusest piirkiiruse ületamise hindamiseks.

4.2 Antennide asukohad

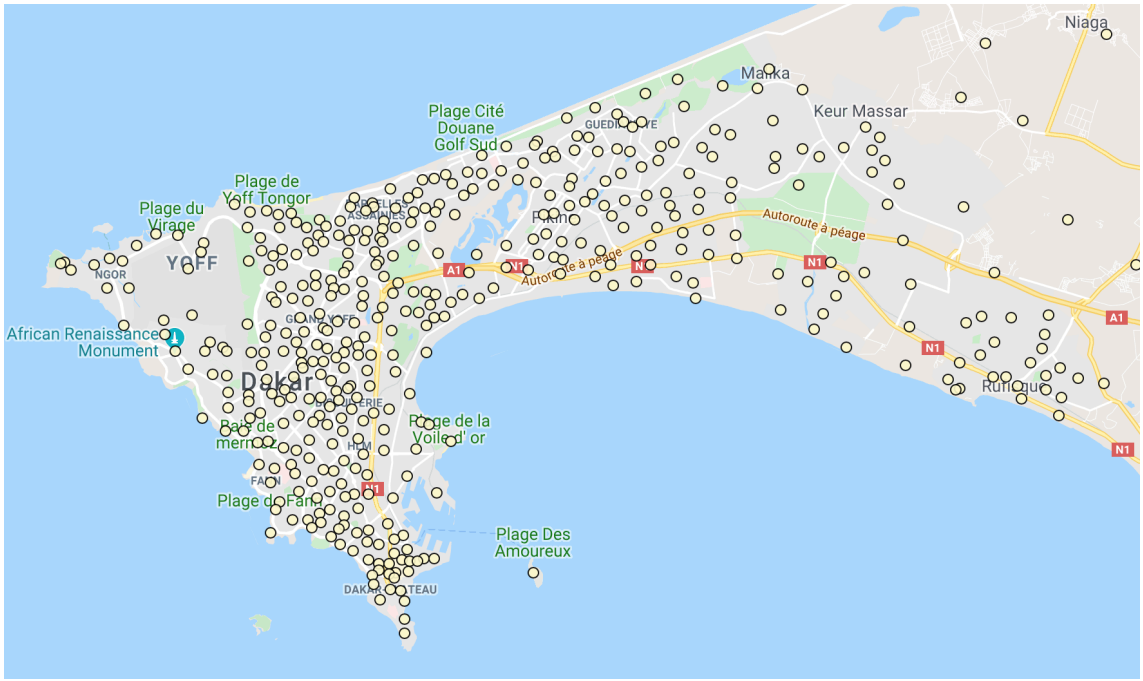
Lõputöö peatükis 4.1 kirjeldatud andmestikust on võimalik tuletada kõikide antennide asukohad. Selleks kustutas töö autor veerud „timestamp“ ja „count“, ja jättis alles ainult unikaalsed kirjed, mis kirjeldavad antennide võrgustikku. Iga antenni kohta on teada regiooni, kuhu see antenn kuulub ja samuti selle koordinaadid.

Kasutades maps.co keskkonda, lisas töö autor kõikide antennide asukohad kaardile. Täpsemalt saab vaadata antennide jaotust joonisel 8. Nagu on näha, siis antennid ei ole jaotatud riigi territooriumile ühtlaselt ja teatud regioonides on selgelt näha, et vahemaad antennide vahel on liiga pikad. Kõige tihedamini on paigaldatud antennid Dakaris, mis on Senegali pealinn, ja selle lähiümbruses. Antennide asukohad Dakaris ja selle lähiümbruses on toodud välja joonisel 9. Käesoleva lõputöö raames vahemaad antennide vahel on väga tähtsad, kuna mida tihedamalt on paigutatud antennid, seda täpsemini saab määrata kliendi asukohta.



Joonis 8. Mobiilvõrgu antennide asukohad, Senegal

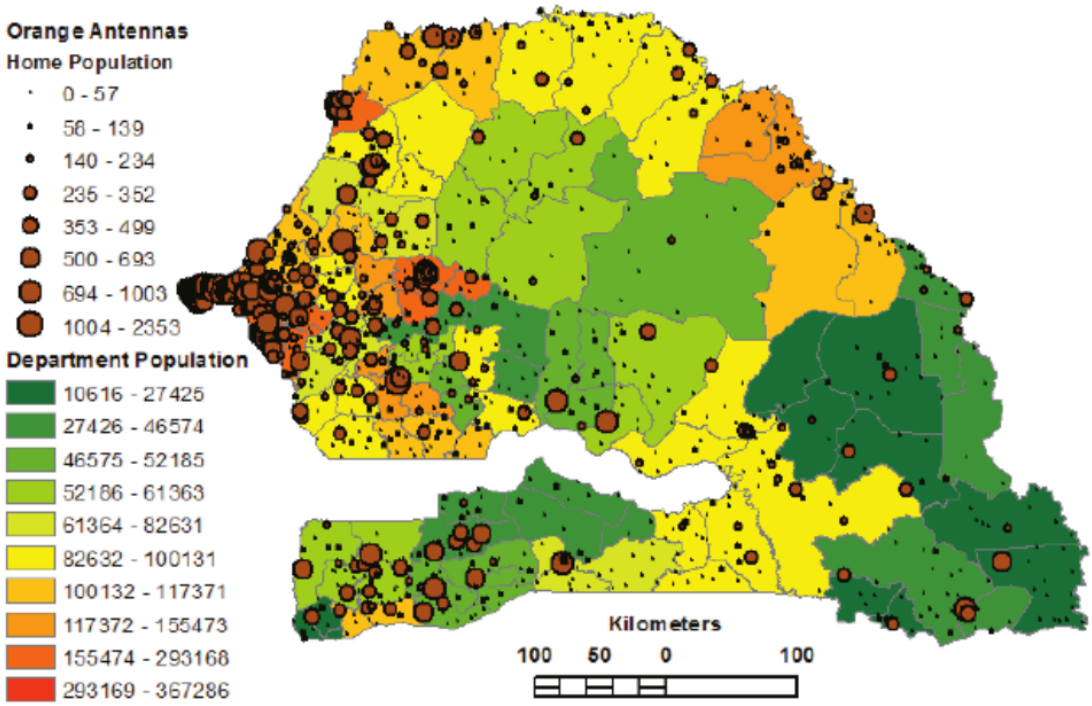
Allikas: koostatud autori poolt maps.co keskkonnas



Joonis 9. Mobiilvõrgu antennide asukohad, Dakar ja selle lähiümbrus

Allikas: koostatud autori poolt maps.co keskkonnas

Senegalis elab kokku 16,5 miljonit inimest, tervest elanikkonnast elab 49% inimesi linnades. Dakaris ja selle lähiümbruses elab peaaegu 2,5 miljonit inimest. (Worldometers) Seega kui arvestada demograafia andmetega, siis on igati loogiline, et ka mobiiltelefoniside antennid on paigutatud just niimoodi. Antennide vahemaid ja nende sobilikkust analüüsi läbiviimiseks hindab töö autor peatükis 5.1.



Joonis 10. Senegali rahvastiku võrdlus antennide asukohtadega
Allikas: Vogel *et al*

Samad andmed on kasutatud ka Grand Valley State University poolt läbi viidud uuringus „*Mining Mobile Datasets to Enable the Fine-Grained Stochastic Simulation of Ebola Diffusion*“. Uuringu autorid uurisid inimeste liikumisharjumusi ja hindasid erinevad meetodid epideemiate leviku tõkestamiseks ja selle tagajärgede leevendamiseks. Joonisel 10 on näidatud võrdlus Senegali rahvastiku Orange S.A. antennide asukohtadega. (Vogel *et al*) Nagu joonisel on näha, suurem osa Senegali elanikkonnast elab riigi lääne osas. Teine regioon, mis on asustatud suhteliselt tihedalt, on Senegali põhja osa, kus on samuti suhteliselt palju antenne (antennide asukohad joonisel 8 ja elanikkonna jaotus joonisel 10).

4.3 Kasutatud lahendused

Käesoleva lõputöö raames kasutab autor peamiselt analüüsi läbiviimiseks Anaconda Navigator paketti, mis sisaldab Jupyter Notebookí, kus toimub koodi kirjutamine. Lisaks sellele sisaldab Anaconda Python 3.6 ja selle populaarsemaid teeke, millest kasutab autor Pandas, SQLite3, Scipy ja teisi. (Anaconda koduleht)

Andmete hoidmiseks on kasutatud SQLite, mille eeliseks on selle lihtsus, mis on kombineeritud võimalusega teha keerulisemaid SQL päringuid, mis on ilmselgelt võimatu kui andmeid hoitakse .csv failides. Samuti SQLite eeliseks võrreldes teiste andmebaasidega on andmebaasi hoidmine ühes failis. Analüüsis kasutatud andmestikku on võimalik salvestada ühte faili, mida on lihtne jagada ja salvestada. (SQLite koduleht)

```
In [17]: import googlemaps

         gmaps = googlemaps.Client(key = API_key)
         x = (14.747767,-17.508766)
         y = (14.854418,-15.881534)
         gmaps.distance_matrix(x, y, mode='driving')["rows"][0]["elements"][0]

Out[17]: {'distance': {'text': '195 km', 'value': 194953},
          'duration': {'text': '2 hours 34 mins', 'value': 9264},
          'status': 'OK'}
```

Joonis 11. Google maps Distance matrix API

Allikas: koostatud autori poolt

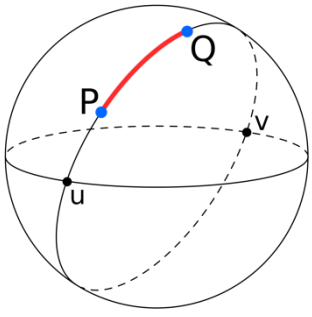
Esmase distantsi mõõtmise jaoks kasutab töö autor lihtsustatud mudelit, kus kaugus kahe punkti vahel on arvutatud linnulennult mõõdetuna. Selline lähenemine õigustab ennast esmaseks filtreerimiseks ja hindamiseks, kuid tulemuste täpsustamiseks kasutab töö autor *Google Maps Distance matrix API*-d, mille päring ja päringu tulemused on toodud joonisel 11. Päringu tulemusena tagastab Google Maps lisaks distantsile ja reisi pikkusele ka staatuse, lähte- ja sihtkoha andmeid, kuid käesoleva analüüsi raames huvitab töö autorit eelkõige distants ja eeldatav reisile kuluv aeg kahe punkti vahel. (Google Maps Platform)

5 Analüüs

Käesolevas peatükis kirjeldab töö autor andmete töötlemise ja analüüsi läbiviimise protsessi. Peale analüüsi tulemusi saamist hindab töö autor ka nende tulemuste valiidsust ja võrdleb saadud tulemusi varasemate uuringute tulemustega. Peatükis 5.4 toob autor välja järeldused uuringust.

5.1 Andmete puhastamine ja ettevalmistamine

Nagu ka kõikide teiste andmetega, vajab ka D4D väljakutse raames avalikustatud andmestik puhastamist ja ettevalmistamist. Peatükis 4.2 kirjutas juba töö autor, et antennid ei ole paigaldatud Senegalis ühtlaselt, vaid teatud piirkondades kipuvad vahemaad antennide vahel olla lubamatult suured. Selleks, et saada täpsemaid tulemusi, on vajalik elimineerida edasisest analüüsist need antennid, mis asuvad teistest väga kaugel. Selleks kasutab töö autor funktsiooni, mis mõõdab ortodroomi (ingl. *Great-circle distance*) ehk kaugust kahe punkti vahel. Ortodroom on lähim tee kahe maakeral asuva punkti vahel (vt joonis 12). Üldjuhul on tegemist kaarega, mille pikkus sõltub kera diameetrist ja punktide koordinaatidest. Kaardil joonistatuna ei pruugi aga ortodroom olla sirge joon (Tartu Ülikool)



Joonis 12. Ortodroom (punane joon PQ)
Allikas: Wikipedia, *Great-circle distance*

Iga antenni jaoks otsis lõputöö autor, kui kaugel asub järgmine lähim antenn. Selle tulemusena antennide andmestik on täiendatud veergudega „closest_site_id“ ehk lähima

antenni number ja „distance“ ehk kaugus lähima antennini (linnulennult mõõdetuna). Täpne andmestiku struktuur on näidatud tabelis 3. Kokku on andmestikus 1 495 antenni ja keskmine distants lähima antennini on 4,89 km. Edaspidises analüüsis kasutab autor ainult nende antennide andmed, mis on paigaldatud teiste antennide lähedal. See tähendab, et kõik antennid, mille lähim naaber antenn asub kaugemal kui 2 km, on edaspidisest analüüsist välistatud. Kokku on andmestikus sobivaid antenne 689 (46% kõikidest antennidest). Nende antennide puhul on keskmine distants lähima antennini oluliselt väiksem: ainult 0,59 km, see tuleneb sellest, et enamus nendest antennidest asuvad linnades.

Tabel 3. Antennide andmestik peale töötlemist

	site_id	arr_id	lon	lat	closest_site_id	distance
1427338	1	2	-17.525142	14.746832	2	0.107514
887018	2	2	-17.524360	14.747434	1	0.107514
264555	3	2	-17.522576	14.745198	2	0.314136
2127011	4	2	-17.516398	14.746730	6	0.498396
1465465	5	2	-17.512870	14.740658	9	0.627559

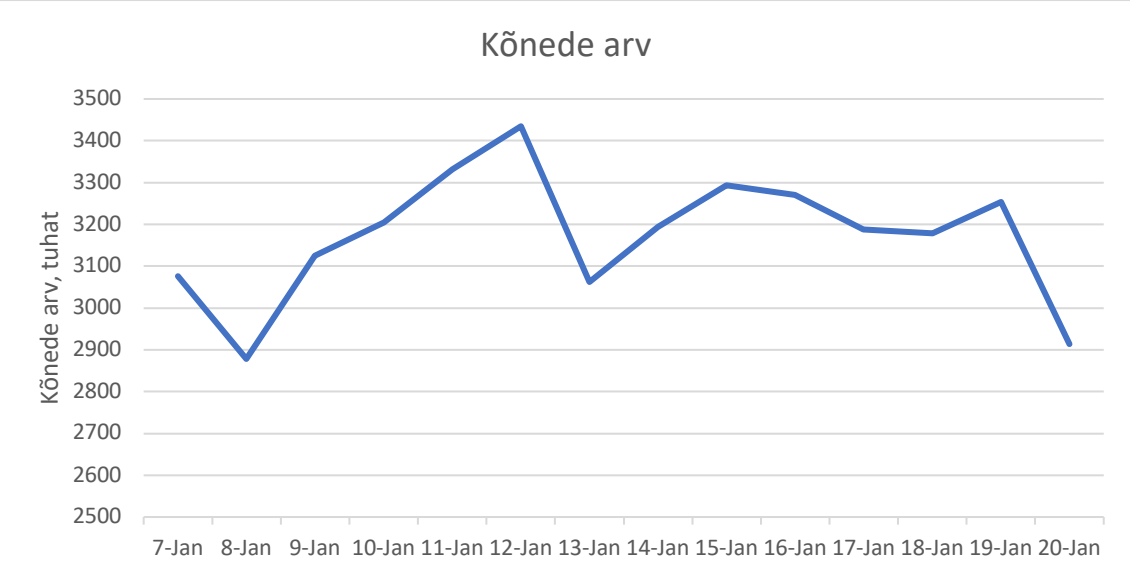
Allikas: koostatud autori poolt

Käesoleval sammul peab mainima, et D4D väljakutse raames avalikustatud andmestiku kohta on teada, et antennide asukohad ei pruugi olla õiged. Mobiilsideoperaatori puhul info antennide võrgustiku kohta on ärisaladus. Kuna Orange S.A. soovis hoida antennide asukoha andmed konfidentsiaalsena, siis asukoha koordinaadid on andmestikus muudetud. (Montjoye et al) Kahjuks ei saa nende andmete täpsust kuidagi tõsta, seega edaspidi autor kasutab just neid andmeid, mis on antud ette, kuid arvestab sellega asukoha määramisel: suure tõenäosusega ei saa asukohta määrata täpsemini kui 1 km veapiiriga.

Nagu on kirjeldatud peatükis 4.1, teises andmestikus on salvestatud kõnede info. Info on esitatud klientide tasemel ja võimaldab jälgida iga inimese liikumistrajektoori läbi päeva. See andmestik on juba enne avalikustamist ette valmistatud ja puhastatud. Andmestikust on eemaldatud kliendid, kes (1) kasutasid mobiilsidet väga vähe, ehk vähem kui 75% kahenädalasest perioodist ja (2) kasutasid mobiilsidet liiga aktiivselt ehk

kelle kohta on süsteemis rohkem kui 1000 kirjet nädalas. Põhjuseks on asjaolu, et esimest tüüpi klientide kohta ei ole piisavalt andmeid, et nendest midagi järeldada. Teises kliendigrupis on aga suure tõenäosusega arvutid või telefonid, mida kasutab suur hulk inimesi. (Montjoye *et al*)

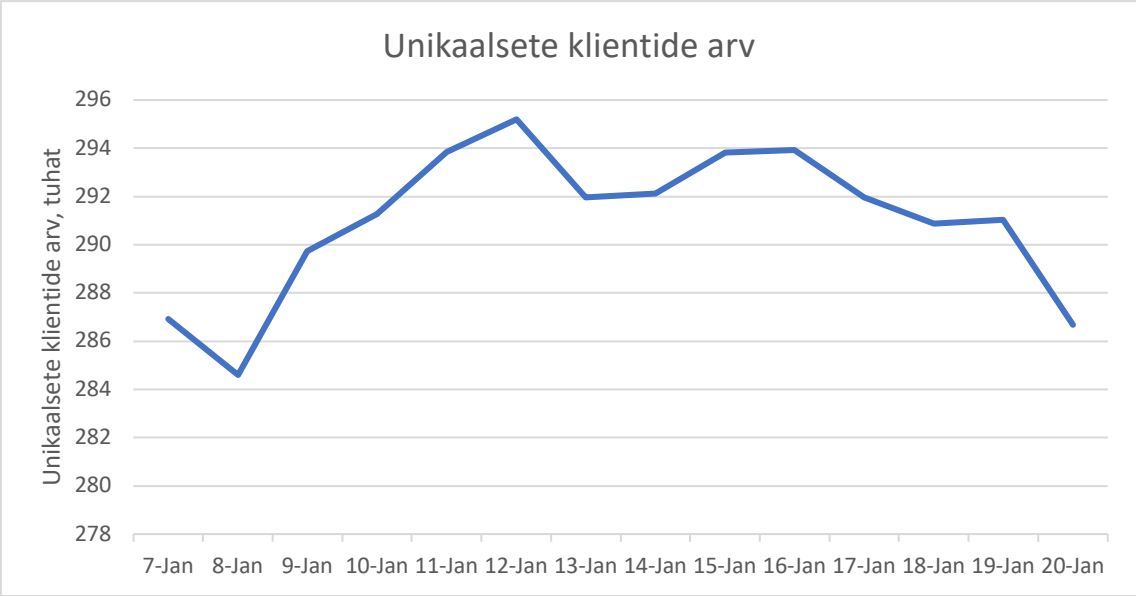
Kogu andmete puhastamise ja ettevalmistamise protsess on viidud läbi Orange S.A. Prantsusmaa esinduses. Kuna CDR andmed pärinevad Orange enda tütarfirmast ja antennide andmestiku omanikuks on teine ettevõte (Sonatel), siis eelkõige need andmestikud ühendati Orange andmekeskuses (ingl. *Orange Data Center*) ja seejärel puhastati ja valmistati ette andmete laboratooriumis (ingl. *Orange Labs*). (Smoreda)



Joonis 13. Kõnede arv päevade kaupa
Allikas: koostatud autori poolt

Kui hinnata andmete terviklikkust, siis on näha, et need on jaotatud väga ühtlaselt. Andmestikus on esindatud 44 miljonit kõnet ja SMS-i, mis on tehtud 2-nädalase perioodi sees (7. jaanuarist 20. jaanuarini 2013). Joonisel 13 on näha, et kõnede arv jaguneb päevade vahel loomulikult: kõige aktiivsem päev on 12. jaanuar ja kõige vaiksemad on 8. ja 20. jaanuar. Iga päeva kohta on andmestikus 2,9...3,4 miljonit kirjet.

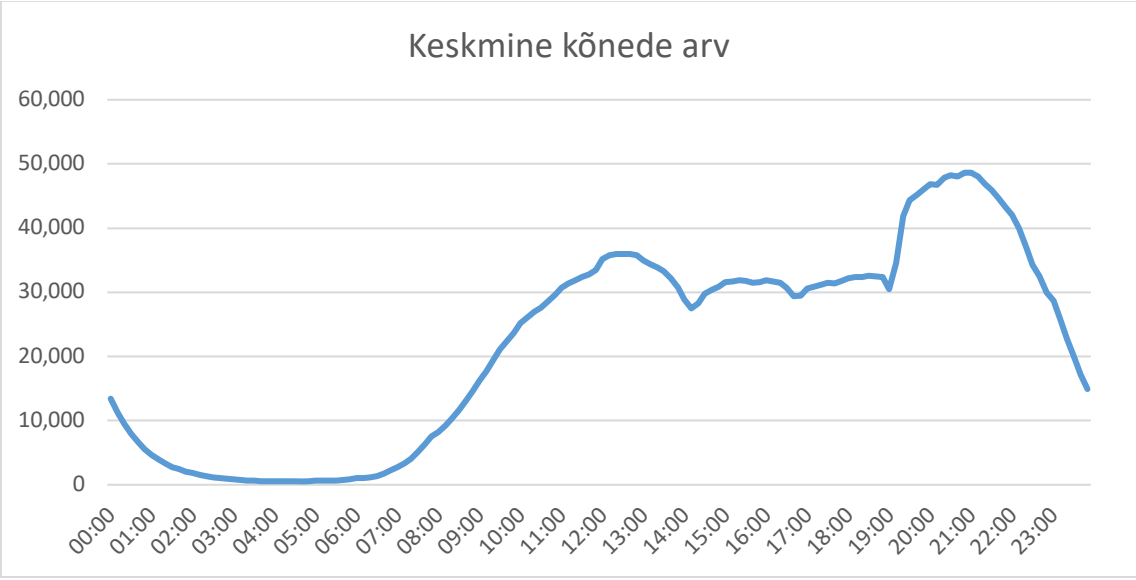
Aktiivsete klientide arv on ka suhteliselt stabiilne erinevatel päevadel. Nagu on joonisel 14 näha, vaadeldava perioodi igal päeval kasutas mobiiltelefonisidet 284...295 tuhat unikaalset klienti. Nendest kahest graafikust saab järeldada, et andmestiku kvaliteet on suhteliselt kõrge.



Joonis 14. Unikaalsete klientide arv päevade kaupa

Allikas: koostatud autori poolt

Lisaks eeltoodule vaatas töö autor üle ka kõnede kellaajalist statistikat. Keskmine kõnede arv igas 10-minutilises ajavahemikus läbi päeva on toodud joonisel 15. Üldiselt saab öelda, et andmed on korrektsed, kuna öisel ajal (01:00-07:00) on kõnede arv nullilähedane ja kõige aktiivsemad ajavahemikud on 12:00-13:00 ja 20:00-21:00.



Joonis 15. Keskmine kõnede arv igas 10-minutilises intervallis

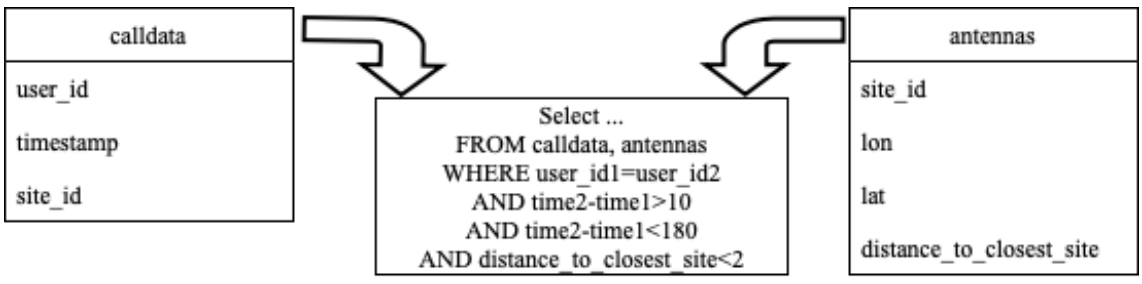
Allikas: koostatud autori poolt

Andmete kvaliteedi hindamisel saab kasutada 6-dimensionilist andmete kvaliteedi mudelit. Antud lähenemist saab kasutada eelkõige andmete majandamise hindamiseks operatiivsetes süsteemides, kuid iseenesest sobib selline lähenemine ka andmeteaduse projektide raames. Nimetatud mudeli järgi andmete kvaliteet koosneb kuuest komponendist: (1) täielikkus, (2) kokku sobivus, (3) vastavus etteantud formaadile, (4) täpsus, (5) terviklikkus ja (6) kättesaadavus ja õigeaegsus. (Smartbridge) Eelnimetatud punktidest etteantud andmetega analüüsides ei ole relevantne ainult 6. punkt. Andmete konfidentsiaalsuse tõttu osa andmetest on ettevalmistamise faasis muudetud, mis vähendab andmete täpsust (punkt 4), kuid saab öelda, et teistes aspektides on andmed piisavalt kõrge kvaliteediga.

5.2 Inimeste teekonna ja keskmise kiiruse hindamine

Selleks, et hinnata inimeste liikumistrajektoori ja kiirust, kasutab lõputöö autor kõnede ja SMS-ide andmestikku. Selleks otsib ta selliseid kirjete paare, kus on vahe kahe kirje ajatemplite vahel 20-170 minutit. Kuna kõik andmed on toodud 10-minutilise täpsusega, siis iga ajavahemiku täpsuseks on +/- 5 minutit. Samuti liiga pikad ajavahemikud ei pruugi näidata õiget pilti, kuna pikematel distantsidel teevad inimesed suurema tõenäosusega peatusi ja keskmine kiirus langeb. Samuti pikemate distantside puhul on suurem tõenäosus, et inimene liigub mõne ettearvamatu trajektoori kaudu.

Liikumistrajektoori väljendab töö autor kirjena, kus on toodud välja kliendi kood, asukoht hetkel *time1*, asukoht hetkel *time2* ja kasutab tabelite ühendamiseks SQL-i. Joonisel 16 on näidatud skeem, kuidas on lähteandmetest saadud liikumiste info kliendi tasemel. Nagu on juba eelnevalt mainitud, analüüsis on kasutatud ainult nende antennide info, mille lähimad naaber antennid asuvad maksimaalselt 2 km kaugusel.



Joonis 16. Kõnede andmete ettevalmistamine

Allikas: koostatud autori poolt

Peale tabelite ühendamist sai töö autor andmestikku, kus on näha iga kliendi liikumised 2-nädalase perioodi sees (tabel 4). Selles andmetabelis on kokku 43 963 853 kirjet. Kuid kui süveneda sellesse tabelisse, siis on näha, et kasutatud lähenemisel on üks probleem. Nimelt, sattusid andmestikku ka kirjed, kus *time1* ja *time2* ei ole järjestikused. Ehk, lisaks paaridele *time1-time2* ja *time2-time3* on sattunud andmestikku ka kirjed *time1-time3* ja *time1-time4*.

Tabel 4. Klientide liikumiste andmestik (puhastamata)

user_id	time1	site1	lat1	lon1	time2	site2	lat2	lon2	timediff
3568	2013-01-15 17:30:00	1	14.746832	-17.525142	2013-01-15 18:40:00	3	14.745198	-17.522576	70.0
3568	2013-01-15 17:30:00	1	14.746832	-17.525142	2013-01-15 18:40:00	6	14.748411	-17.512103	70.0
3568	2013-01-15 17:30:00	1	14.746832	-17.525142	2013-01-15 18:50:00	2	14.747434	-17.524360	80.0
3568	2013-01-15 17:30:00	1	14.746832	-17.525142	2013-01-15 19:00:00	5	14.740658	-17.512870	90.0
3568	2013-01-15 17:30:00	1	14.746832	-17.525142	2013-01-15 20:10:00	182	14.736088	-17.444960	160.0

Allikas: koostatud autori poolt

Selleks, et analüüsi tulemused oleksid võimalikud täpsed, tuleb välistada analüüsist nii nimetatud liigsed liikumised, ehk need kirjed, kus kaks ajatemplit ei ole tegelikult järjestikud. Selleks kasutab töö autor Pandas teegi funktsiooni drop_duplicates, jättes alles ainult esimese kirje, kus user_id, time1 ja site1 on unikaalsed. (Pandas koduleht) See tagab, et sama liikumine ei lähe teist korda arvesse. Peale seda jääb andmestikus ainult 12 087 898 kirjet, millega töötab lõputöö autor edasi. Puhastatud andmestiku näide on toodud Tabelis 5.

Tabel 5. Klientide liikumiste andmestik (puhastatud)

user_id	time1	site1	lat1	lon1	time2	site2	lat2	lon2	timediff
3568	2013-01-15 17:30:00	1	14.746832	-17.525142	2013-01-15 18:40:00	3	14.745198	-17.522576	70.0
3568	2013-01-15 17:40:00	1	14.746832	-17.525142	2013-01-15 18:40:00	3	14.745198	-17.522576	60.0
3568	2013-01-15 18:20:00	1	14.746832	-17.525142	2013-01-15 18:40:00	3	14.745198	-17.522576	20.0
5811	2013-01-08 16:00:00	1	14.746832	-17.525142	2013-01-08 17:00:00	2	14.747434	-17.524360	60.0
5811	2013-01-08 16:10:00	1	14.746832	-17.525142	2013-01-08 17:00:00	2	14.747434	-17.524360	50.0

Allikas: koostatud autori poolt

Saadud tulemuste töötlemiseks kasutab töö autor sarnast meetodit sellega, mis on kirjeldatud peatükis 5.1. Kauguse mõõtmiseks kahe asukoha vahel kasutab autor ortodroomi. Kiiruse arvutamine toimub lähtudes ajast, mis on möödunud kahe ajatempli

vahel ja saadud distantsist. Peab arvestama sellega, et distantsi mõõtmine toimub linnulennult, seega peab arvestama, et reaalne kiirus maanteel on üldjuhul kõrgem.

Kui vaadata saadud kiiruse tulemustele otsa, siis on lihtne märgata, et keskmine kiirus andmestikus on väga madal. Tegelikult 75% juhtumitest on liikluskiirus alla 3 km/h, mis on igati loogiline, kuna enamuse ajast inimesed veedavad kodus või tööl, mitte liikudes. Samuti analüüsi tulemusena on tulnud välja ka mõned statistilised vead, kus keskmiseks kiiruseks on 200+ km/h. Puhastamata andmete statistiline jaotus on toodud välja tabelis 6.

Tabel 6. Puhastamata klientide liikumiste andmestiku statistiline jaotus

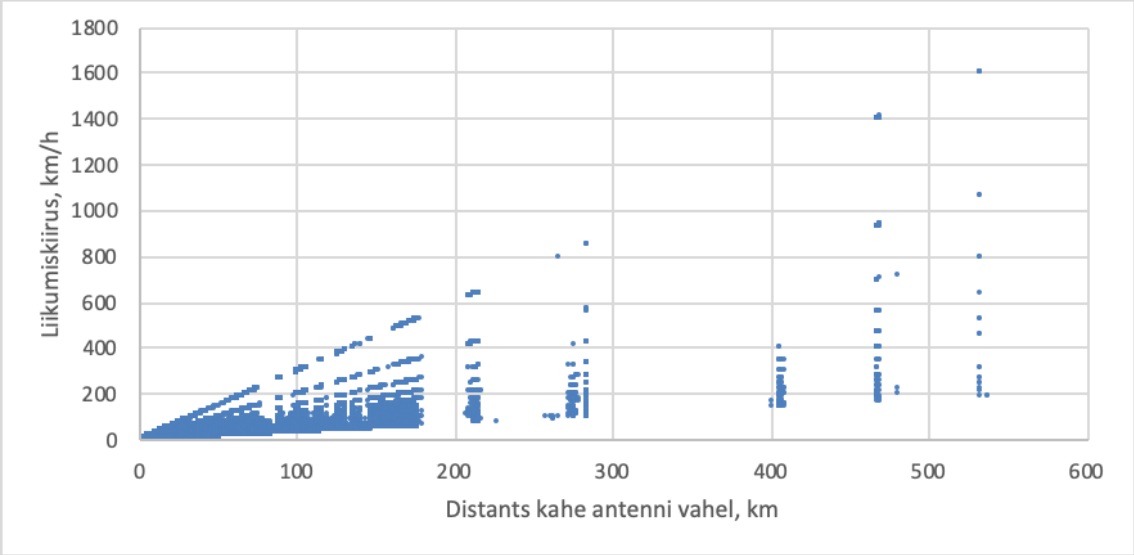
	user_id	timediff	distance	speed
count	12 087 900	12 087 900	12 087 900	12 087 900
mean	161 634	64,38	2,58	3,32
std	91 082	43,60	8,37	12,01
min	1	20,00	0,11	0,04
25%	78 810	30,00	0,59	0,62
50%	160 707	50,00	1,01	1,35
75%	243 569	90,00	2,12	2,98
max	320 000	170,00	537,77	1 598,55

Allikas: koostatud autori poolt

Tabelis 6 on näha, et maksimaalne liikumiskiirus andmestikus on 1 599 km/h, mis on ilmselgelt võimatu, kuna isegi reisilennukid liikuvad oluliselt madalama kiirusega. Kuna tegemist on statistilise veaga, siis otsustas töö autor hinnata, kui palju sarnaseid vigu leidub terves andmetabelis. Selleks autor võttis ette kõik andmeread, kus liikumiskiiruseks on vähemalt 10 km/h. Kokku on selliseid kirjeid 654 810, mis on piisavalt suur andmehulk. Kiiruste jaotus sõltuvalt distantsist kahe antenni vahel on näidatud joonisel 17.

Joonise peale vaadates võib tekkida tunne, et andmestikus on suhteliselt palju vigaseid andmeid, kuid see ei ole päriselt nii. Autori arvates kiirused kuni 100 km/h on piisavalt usaldusväärsed. Kirjeid, kus keskmine liikumiskiirus on üle 100 km/h, on kogu andmestikus 24 861, mis on 3,8% andmestikust. Lõputöö autor eeldab, et suur osa nendest näidetest on tingitud liikumisest lennutranspordiga. Üle 500 km/h liikluskiirus on fikseeritud 412 juhtumil, mis on 0,06% kogu andmestikust. Need on autori hinnangul statistilised vead. Üldjuhul nendel juhtudel on tegemist kirjetega, kus üks ja sama klient

oli alguses Dakaris ja 20-30 minuti pärast Toubas (190 km Dakarist) või Bakelis (670 km Dakarist). Selline viga võib olla tingitud näiteks vigadest klientide kodeerimisest (kui kaks klienti, kes elavad erinevates linnades, on saanud sama user_id andmete ettevalmistamise etapis).



Joonis 17. Kiiruse jaotus sõltuvalt distantsist

Allikas: koostatud autori poolt

Selleks, et saada usaldusväärseid tulemusi, kasutab töö autor edaspidi ainult neid näiteid, kus inimese kiirus on linnulennult mõõdetuna olnud 50...100 km/h. Kuna reaalne distants on üldjuhul veidi pikem, kui see, mis on saadud ortodroomi abil, siis need juhtumid, kus kiiruseks on 50...100 km/h, on potentsiaalsed kiiruse ületamised.

Tabel 7. Puhastatud klientide liikumiste andmestiku statistiline kirjeldus

	timediff	distance	speed
count	24915.000000	24915.000000	24915.000000
mean	75.212523	81.384081	65.657676
std	39.517515	41.691870	12.283013
min	20.000000	16.667766	50.000085
25%	40.000000	39.228094	55.656724
50%	80.000000	101.741744	61.763764
75%	100.000000	102.654165	74.896937
max	170.000000	282.700068	99.931506

Allikas: koostatud autori poolt

Tabelis 7 on toodud välja saadud andmestiku statistiline kirjeldus. Andmete kirjeldamiseks on kasutatud Pandas teeki sisse ehitatud funktsioon *describe()*. (Pandas koduleht) Nagu on tabelist näha, esineb andmestikus 24 915 „kahtlast“ liikumistrajektoori, mida kavatseb lõputöö autor täpsemini uurida. Vaatamata sellele, et andmestikus leidub ka lühemaid distantse, siis keskmine distants, mida on läbitud ette antud ajaga, on üle 80 km, mis viitab sellele, et üldjuhul on tegemist linnadevahelise liiklusega. Reaalses maailmas on sisendandmed oluliselt suurema granulaarsusega (antennid asuvad tihedamini ja iga kirje juures on olemas korrektne ajatempel, mis näitab aja sekundi täpsusega). Seega võib eeldada, et reaalsete andmete põhjal on võimalik leida kahtlaseid teelõike ka linna sees. Kuid käesoleva lõputöö raames, just andmete eripärade tõttu, keskendub töö autor just linnadevahelisele liiklusele.

5.3 Tulemuste valideerimine

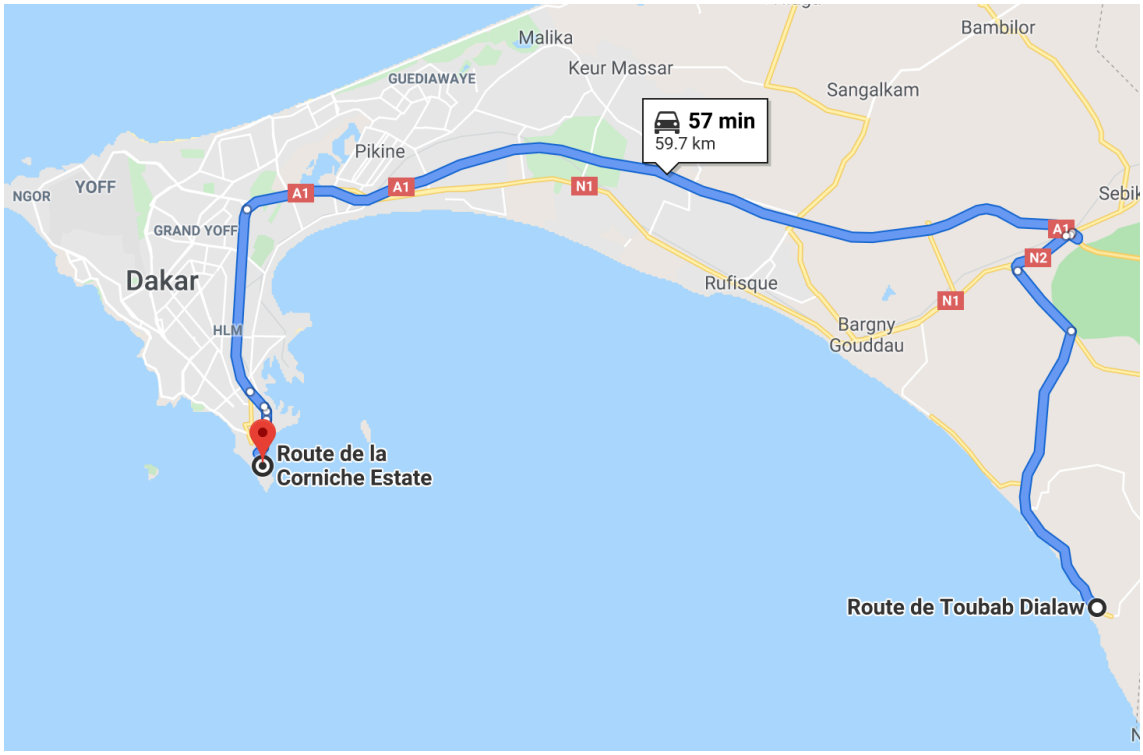
Nagu on mainitud peatükis 4.3, saadud tulemuste valideerimiseks kasutab töö autor *Google Maps Distance Matrix API*-d. See võimaldab kontrollida, mis on reaalne kaugus kahe koordinaatpunkti vahel ja kui palju aega läheb, et jõuda punktist A punkti B.

Andmete täiendamiseks leidis töö autor kõik lähte- ja sihtkohtade paarid, mis leidusid andmestikus. Selle põhjuseks on see, et Google API ühendus on tasuline ja töö autor soovis minimeerida korduvate päringute saatmist. Kokku leidus 4 638 unikaalset antennide paari ja iga sellise paari jaoks leidis töö autor, kui pikk distants nende vahel on ja kui kiiresti jõuab autoga liikudes inimene punktist A punkti B.

Lisas 2 on toodud välja kõige sagedamini esinevad marsruudid, ehk siis need, mis esinevad valimis 100 ja rohkem korda kahenädalase perioodi sees. Kui vaadata nendele andmetele otsa, siis on lihtne märgata kahte probleemi:

- Kuna aja määramine toimub 10-minutilise täpsusega (esialgse andmestiku eripära), siis sõidu pikkust saab määrata +/- 10 minutilise täpsusega. Mis tähendab, et just selle andmestiku puhul lühemad sõidud annavad oluliselt suurema vea (näiteks, 20-minutlise sõidu puhul, mille distants on 21 km, saab väita et kliendi kiirus oli vahemikus 43...128 km/h, mis ei anna lõputöö kontekstis eriti mingit kasu). Seega analüüsis on võimalik ainult kasutada võimalike kiiruste vahemikku.

- Isegi juhul, kui kasutada vahemiku alumist piiri (ehk, eeldada, et enamus inimesi ei ületa piirkiirust), siis tundub, et kiiruse ületajaid on suhteliselt palju. Selle põhjuseks võib olla eelkõige see, et eelmisel sammul oli valitud kiiruse vahemik 50...100 km/h (linnulennult mõõdetuna), mida nimetas töö autor „potentsiaalseks kiiruse ületamiseks“. Kuid lisaks andmete ettevalmistamisele, mõjutab tulemusi ka andmete kvaliteet ja konteksti tundmine. Autor leidis 2 juhtumit, mis kirjeldavad antud probleemi, täpsemalt joonistel 18-19.

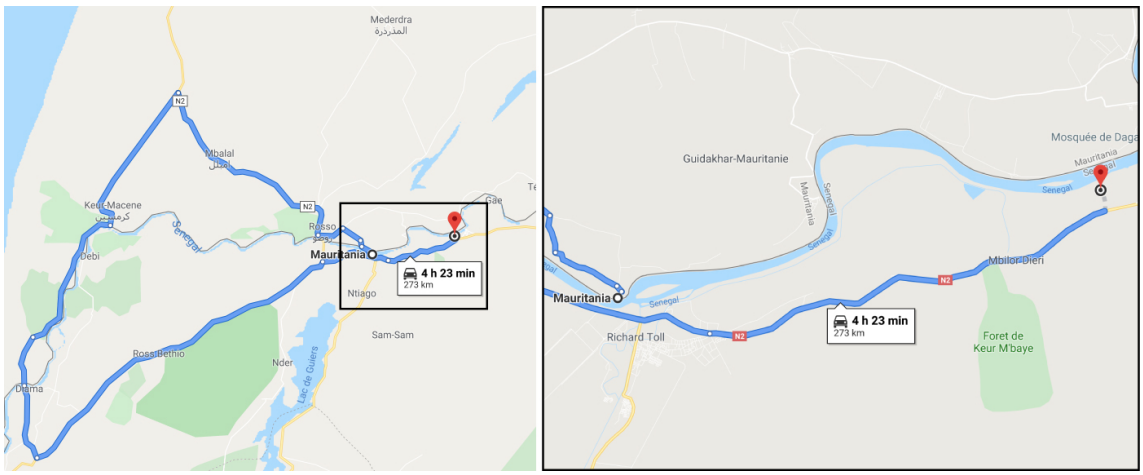


Joonis 18. 485 (14.60897, -17.15001) -> 241 (14.65547, -17.43353)

Allikas: koostatud autori poolt

Joonisel 18 on esimene näide ühest erijuhtumist: lähte- ja sihtkoha vahel on 59,7 km, samas kui linnulennult mõõdetuna on distants ainult 30,9 km. Andmestikus leidub 9 juhtumit, kus inimene on liikunud nende antennide vahel ja keskmiselt on sellele kulunud 22 minutit. Arvestades sellega, et aja määramisel esineb viga +/- 10 minutit, eeldatav kiirus sellel teelõigul on 111-293 km/h, mis ilmselgelt ei ole võimalik, kuna tegemist on Senegali suurima linnaga – Dakar. Antud juhul saab tervest mõistusest lähtudes väita, et tegemist on inimestega, kes on liikunud paadiga kahe antenni teeninduspiirkondade vahel. Ehk, sellisel juhul on väga tähtis konteksti tundmine.

Teine näide kirjeldab andmete kvaliteedi probleemi ja on kujutatud joonisel 19, kus kliendid on liikunud antenni 1185 ja 1122 vahel. Kuna üks antennidest asub teisel pool jõge, siis lähtuvalt Google Maps andmetest vahemaa lähte- ja sihtkoha vahel on 273 km. Kui lähtuda tervest mõistusest, siis on selge, et inimene oli tegelikult teisel pool jõge ja nii öelda õige vahemaa on 19,9 km.



Joonis 19. 1185 (16.50834, -15.53094) -> 1122 (16.47206, -15.70038)
Allikas: koostatud autori poolt

Lisaks kahele eeltoodud näitele peab analüüsi läbiviimisel arvestama ka ühistranspordi liiklusega. Kahjuks käesoleva lõputöö autor ei leidnud piisavalt infot Senegali ühistranspordi kohta, kuid sarnase lähenemise kasutamisel peab arvestama sellega, kuna ühistranspordi kiirus võib oluliselt erineda auto lubatust kiirusest. Näiteks, Eesti kontekstis sõit Tallinnast Tartusse võtab keskmiselt aega 2 tundi ja 15 minutit, samas kui Elron ekspress rong jõuab Tallinnast Tartusse vähem kui 2 tunniga. (Elron sõiduplaan)

5.3.1 Tulemuste statistiline jaotus

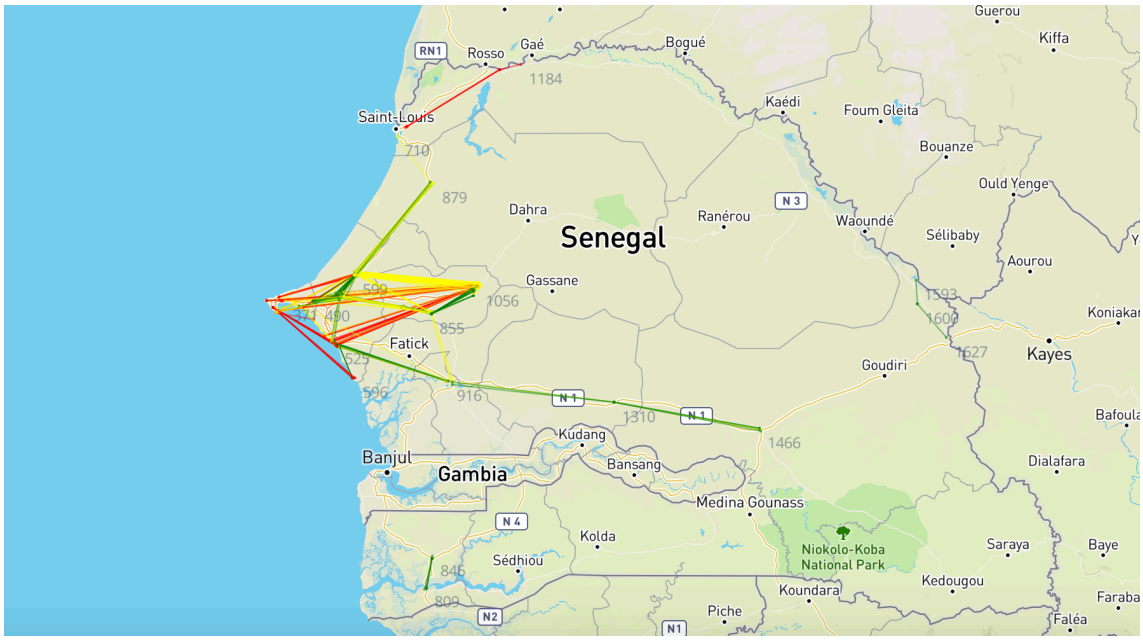
Arvestades kõikide asjaoludega, mis on loetletud punktis 5.3, käesoleva lõputöö autor otsustas, et korrektsete tulemuste saamiseks on vaja rakendada mitu täiendavat filtrit. Eeldused ja täiendavad piirangud, mida kavatseb kasutada autor, tagavad tulemuste usaldusväärsust:

- Edaspidistes arvutustes läheneb töö autor minimaalsest kiirusest (peatükis 5.3 kirjutas töö autor, et andmete eripärade tõttu liikumiskiirust on võimalik määrata ainult vahemikuna);

- Kuna andmestikus on olnud marsruute, mida läbiti kogu vaadeldava perioodi sees ainult üksikuid kordi, siis lõputöö autor otsustas filtreerida need välja ja arvestada ainult nende marsruutidega, mida on läbitud 10 ja rohkem korda kahenädalase perioodi sees.

Peale eeltoodud tingimuste rakendamist sai lõputöö autor andmestikku 355 marsruudiga, mida läbiti vähemalt 10 korda ja kus on tõenäosus, et kiirust ületatakse. Kirjeldatud marsruudid on lisatud joonisele 20. Nagu on jooniselt näha, kõik marsruudid on märgitud kolme erineva värviga. Erinevad värvid on kasutatud, et näidata keskmist kiirust, millega inimesed liiguvad: kuni 60 km/h on märgitud rohelise, 60-80 km/h kollase ja üle 80 km/h punase värviga. See võimaldab lihtsalt näha, mis linnade vahel suurema tõenäosusega ületatakse piirkiirust ja milliste linnade vahel kiirust ületatakse harvemini.

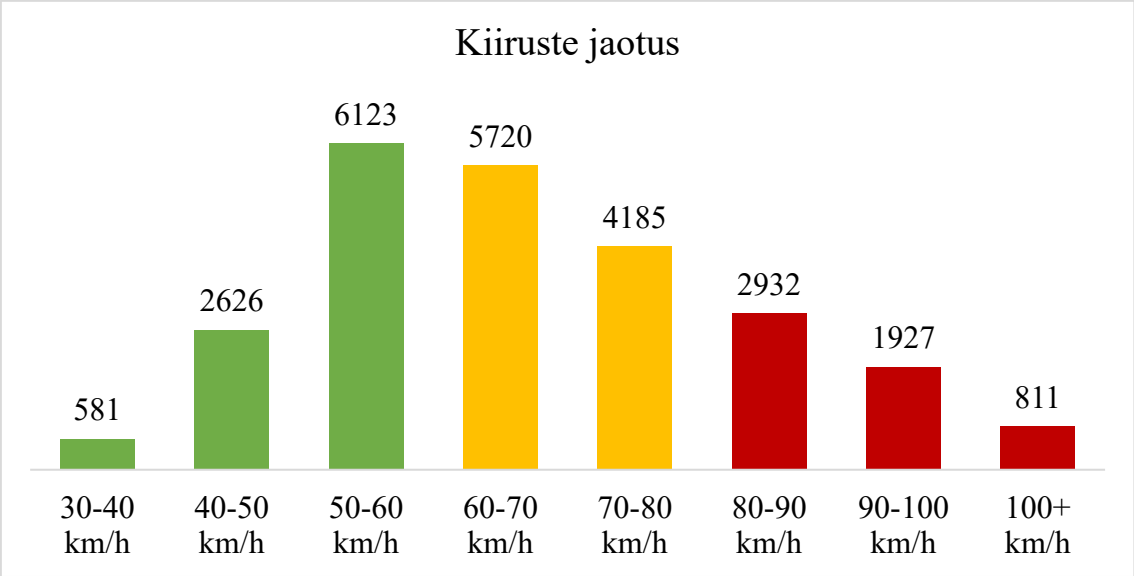
Põhjus, miks on kasutatud just sellised värvitoonid, on maksimaalselt lubatud kiiruses Senegalis. Nii nagu Eestis, Senegalis on lubatud liikuda kiirusega 50 km/h linnas ja 90 km/h väljaspool linna. (Rhinocarhire) Seega juhul, kui kliendi minimaalne kiirus on olnud 80+ km/h, tekib väga suur kahtlus, et sellel teelõigul on maksimaalselt lubatud kiirus ületatud.



Joonis 20. Keskmine kiirus kõige sagedasematel marsruutidel

Allikas: koostatud autori poolt

Nagu on joonisel näha, Senegalis on teatud teelõigud, kus keskmine liikumise kiirus viitab asjaolule, et sellel maanteel ületatakse suhteliselt tihti piirkiirust. Üks selline teelõik on näiteks Dakari ja Touba ning teine Touba ja Saly vahel. Kuid on märkimisväärne, et kiirust ületatakse ka Dakarist eemal, näiteks St Louis ja Richard Toll vahel. Kuid antud joonisest on näha, et kuna käesolevas analüüsis on kasutatud ainult kahe nädala andmed, siis väiksemates kohtades ei kogunenud piisavalt palju „kahtlaseid" sõite ja näiteks Senegali Idapoolse teevõrgustiku kohta ei ole piisavalt andmeid järelduste tegemiseks.



Joonis 21. Kiiruste jaotus

Allikas: koostatud autori poolt

Lisaks sellele vaatas töö autor ka üksiksõite ja üritas saada aru, kuidas jagunevad liikumiskiirused üksiksõitude tasemel. Joonisel 21 on näidatud kiiruste jaotus (aluseks on võetud potentsiaalsed kiiruse ületamised). Nagu on joonisel näha, suhteliselt kindlalt saab öelda, et esialgses andmestikus on kiiruse ületamisi, ja kirjeldatud meetod võimaldab neid tuvastada.

5.3.2 Võrdlus varasemate uuringutega

Peatükis 3.2 tõi lõputöö autor välja 2 näidet analüüsidest, mis on viidud läbi samade andmete põhjal (D4D väljakutse 2014 raames avalikustatud CDR andmestik Senegalist). Lisaks juba kirjeldatud analüüsidele kirjutab töö autor ühest analüüsist, mis on viidud läbi 2016. aastal Rainer Kujala, Talayeh Aledavood ja Jari Saramäki poolt.

Uuringu pealkirjaks on „*Estimation and monitoring of city-to-city travel times using call detail records*“. Põhjuseks, miks see uuring on oluline antud lõputöö raames, on asjaolu et antud analüüsis üritasid autorid kasutada sarnast lähenemist selleks, et arvutada välja CDR andmete põhjal, kui kaua võtab aega sõit Senegali linnade vahel.

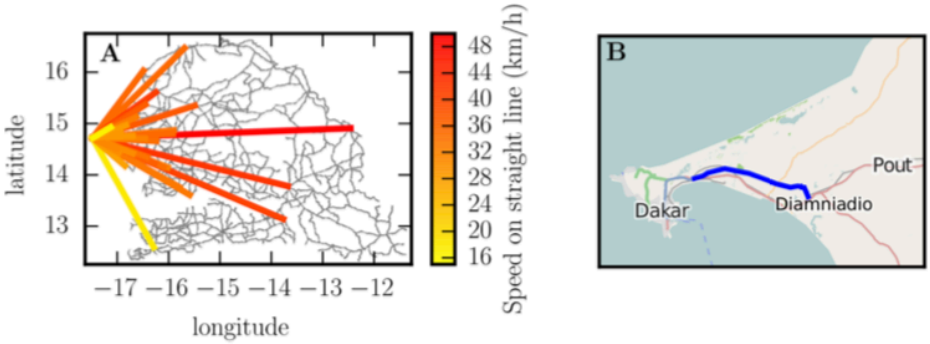
Uuringu autorid üritasid leida viisi reisile kuluva aja arvutamiseks, mis oleks kasutatav arenevates riikides ja oleks (1) soodne, (2) infrastruktuurist mittesõltuv ja (3) täpne. Selleks kasutasid autorid CDR andmeid. Autorid kinnitavad, et kuna CDR andmeid kogutakse eelkõige arveldamise jaoks, siis andmed ühe inimese liikumiste kohta võib olla lünklik, kuid terve andmebaasi põhjal saab teha järeldusi inimeste liikumiskiiruste kohta. Meetod, mida töötasid välja R. Kujala ja tema kolleegid, ei võimalda jälgida reaalajas inimeste liikumisi (nagu oleks võimaldanud näiteks GPS-il põhinev meetod), kuid selle abil saab määrata muutused linnadevahelise liikluse kiirustes, mis aitab hinnata teevõrgustikus tehtavate muudatuste efekti.

Uuringu autorid alustasid analüüsi andmete puhastamisest. Esimese sammuna autorid välistasid andmed, mis on seotud mitteaktiivsete antennidega. Andmestikust eemaldati antennid, mis ei ole seotud vähemalt ühe kirjega igal päeval vaadeldava perioodi sees. Samuti kõik antennid grupeeriti linnade kaupa ja jäeti alles ainult need kirjed, kus inimene on liikunud ühest linnast teisse. Kokku analüüsis kasutati 1093 antenni info 62 Senegali linnast.

Analüüsi läbiviimisel uuringu autorid pöörasid tähelepanu nii keskmisele, kui ka minimaalsele reisile kuluvale ajale. Kuna inimeste telefoni kasutusviisid on erinevad ja erinevate inimeste puhul aeg kohale jõudmise ja kõne tegemise vahel võib olla pikk, siis minimaalne reisile kuluv aeg on tähtis et saada aru, kui kiiresti on võimalik jõuda ühest linnast teisse. Autorid kasutasid oma analüüsis ainult need andmeread, kus inimese kiirus linnulennult mõõdetuna on olnud alla 100 km/h. Sellega said välistatud kõik linnadevahelised liikumised õhu- või teise transpordiga, mis on autodest oluliselt kiiremad. Autorid jõudsid järelduseni et usaldusväärsete tulemuste saavutamiseks on vajalik minimaalselt 1000 ja soovitavalt vähemalt 10 000 kirjet iga linnapaari jaoks.

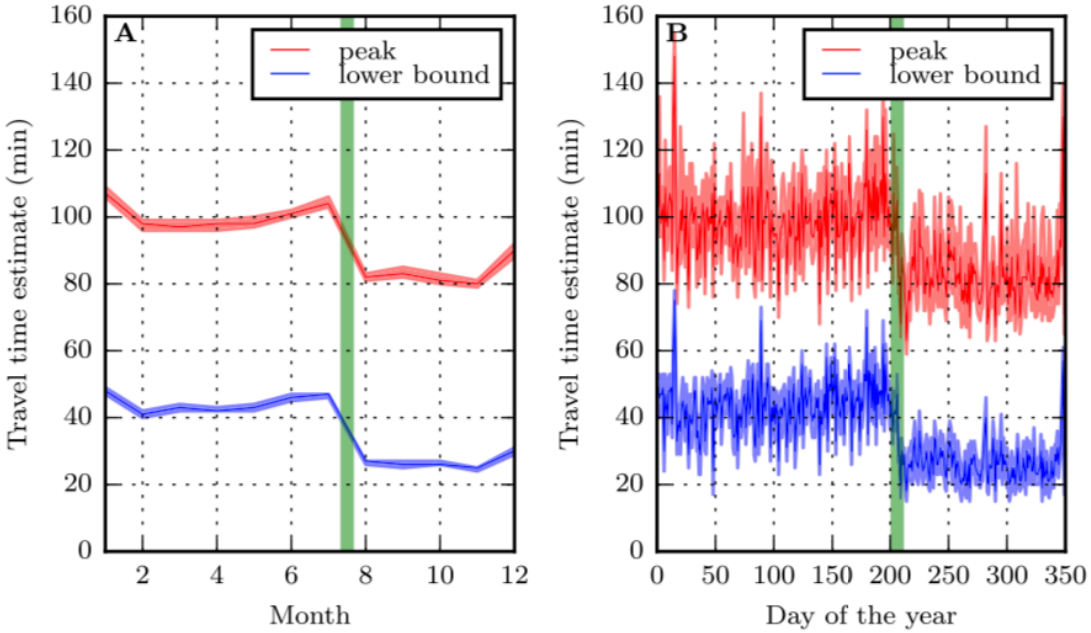
Peale tulemuste võrdlust *Google Maps API* numbritega, jõudsid uuringu autorid järelduseni, et üldjuhul kehtib järgmine reegel: keskmine reisile kuluv aeg on üldjuhul pikem kui *Google Maps* oma. Minimaalne reisile kulunud aeg on samuti veidi kõrgem,

kuid on oluliselt lähedasem *Google Maps* tulemusele. Autorid järeldasid, et sellel on kaks põhjust: (1) *Google Maps* hindab üle kiirust, millega on võimalik Senegali maanteedel liikuda ja (2) *Google Maps* ei arvesta peatustega, mida inimesed üldjuhul teevad liikudes pikemale distantsile. Lõputöö autori hinnangul võib erinevust tulemustes põhjendada veel üks asjaolu: CDR andmed genereeritakse siis, kui klient teeb kõne, ehk suure tõenäosusega kohalejõudmise ja kõne tegemise hetkede vahel on teatud vaheaeg, mida uuringu autorid ei ole arvutuste sisse arvestanud. Saadud keskmine kiirus Senegali linnade vahel on näidatud joonisel 22.



Joonis 22. Liikluse kiirused Senegali linnade vahel

Allikas: R. Kujala *et al*



Joonis 23. Reisile kuluva aja muutus peale maantee avamist

Allikas: R. Kujala *et al*

Lisaks keskmise reisile kuluva aja arvutamisele on välja töötatud mudelil ka teisi kasutusvaldkondi. Näiteks, võimaldab mudel hinnata erinevate uuenduste mõju keskmisele liikumiskiirusele. Joonisel 23 on näidatud, kui palju on muutunud keskmine reisile kuluv aeg peale uue maantee avamist suvel 2013. Eraldi on näidatud nii kuu, kui ka päeva vaade ja on selgelt näha, et (1) maantee avamine mõjutas väga positiivselt keskmist liikumiskiirust ja reisile kuluvat aega ja (2) reisile kulunud aeg võib erinevatel päevadel erineda kuni 2 korda, kuid peale maantee avamist isegi kõige pikemad reisiajad jäävad alla maantee avamise-eelset keskmist aega.

Uuring „*Estimation and monitoring of city-to-city travel times using call detail records*“ on väga tähtis käesoleva lõputöö kontekstis, kuna selles on kasutatud samad andmed ja uuringu autorid keskendusid just linnadevahelisele liiklusele. Kuna käesoleva lõputöö autor samuti töötab eelkõige linnadevahelise liiklusega, siis on hea võrrelda mõlemas uuringus saadud tulemusi. R. Kujala ja tema kolleegid on jõudnud järelduseni, et CDR andmete põhjal välja arvutatud reisi aeg on üldjuhul pikem kui *Google Maps* eeldus. Kuna käesolevas lõputöös keskendus töö autor just kiiruse ületamistele, siis on loogiline, et välja filtreeritud andmete puhul on pilt teistsugune: peatükis 5.3 viidatud andmete puhul on keskmine liikumiskiirus üldjuhul kõrgem kui *Google Maps* eeldatud kiirus. Sellest saab järeldada, et lõputöö autor on teinud õigeid eeldusi potentsiaalsete kiiruse ületamiste välja filtreerimiseks. Samuti viidatud uuringust sai lõputöö autor kinnitust, et tulemuste valideerimine *Google Maps* andmete abil õigustab ennast.

Veel üks analüüs, mida soovib lõputöö autor mainida, on viidud läbi Hiinas kolme Pekingi ülikooli koostöös. Uuringus „*Travel Distance Characteristics Analysis Using Call Detail Record Data*“ kasutasid autorid 27 604 mobiilside antennide ja mitme miljoni Pekingi elaniku CDR andmeid. See uuring on mingis mõttes unikaalne just sellepärast, et selles on kasutatud selline lähenemine, kus iga antenni jaoks on täpselt määratud piirkond, kus klient asub. Riigi või linna jagamine tsoonideks aitab paremini näha suurt pilti ja saada aru, kuidas inimesed päriselt liiguvad erinevate tsoonide vahel. Analüüsi käigus jõudsid autorid järelduseni, et CDR andmed sobivad inimeste liikumisharjumuste ja -kiiruse hindamiseks paremini, kui GPS andmed just nende kogumise lihtsuse ja soodsuse tõttu. (Wang *et al*)

5.3.3 Eksperthinnang

Selleks, et saada hinnangut lõputöös kasutatud lähenemisele, kohtus lõputöö autor Telia Eesti AS esindajaga. Autor viis läbi intervjuud Ain Õiglase, mobiilsideteenuste omanikuga Telias, kellele tutvustas enne intervjuu läbiviimist metoodikat, mis on kasutatud lõputöös. Intervjuu täispikkuses on leitav käesoleva lõputöö lisas 1.

Intervjuu raames rääkis ekspert sellest, kas ja kuidas on CDR andmeid kasutatud Eestis ja sellest, kuidas need andmed genereeritakse. Selgus, et lisaks SMS sõnumitele ja kõnedele genereeritakse andmed siis, kui inimene kasutab mobiilset interneti. Samuti rääkis Ain sellest, kui täpselt saab määrata inimese asukoha CDR andmete abil ja kui tihedalt on paigaldatud mobiilside antennid Eesti kontekstis. Ekspert tõi välja mitu takistust, mis teevad CDR andmete kasutamist piirkiiruse ületamise hindamiseks raskendatuks:

- Isegi asukohas, kus lähiümbruses asub palju antenne, on väga raske määrata täpselt inimese asukoha; mobiilsidevõrgu funktsioneerimise eripäradest lähtuvalt ei saa määrata inimese asukoha täpsemalt kui 0,1-1,0 km veapiiriga;

- Seoses asukoha määramise täpsusega saab CDR andmeid kasutada ainult selleks, et leida kiiruse ületamise statistikat pikematel teelõikudel (hinnanguliselt alates paarikümnest kilomeetrist);

- Pikemate distantside puhul keskmise kiiruse määramine läheb täpsemaks, kuid selline lähenemine ei anna infot konkreetsete teelõikude kohta, kus inimesed piirkiirust ületavad.

Samuti andis Telia ekspert soovitusi selle kohta, kuidas saaks vajadusel sarnast analüüsi viia läbi Eesti kontekstis ja kuidas oleks võimalik saada sellest võimalikult paremat tulemust:

- Lisaks kõnede ja SMS sõnumite andmetele on soovitatav kasutada ka mobiilse interneti andmeid, kuna sellisel juhul andmed on salvestatud tihedamini;

- Selleks, et parandada analüüsi täpsust, on vajalik, et analüütikul oleks ligipääs kogu antennide võrgustiku andmetele, mis on rangelt konfidentsiaalne info.

Lähtudes intervjuu vastustest, järeldas töö autor, et ideaalses olukorras peaks CDR andmete analüüs olema läbi viidud mobiilsideoperaatori poolt. See vähendab probleeme, mis on seotud GDPR nõuete ja konfidentsiaalse info hoidmisega. Samuti aitab see saavutada täpsemaid tulemusi andmete analüüsimisel ja tõlgendamisel. Samuti mainis intervjuus Telia esindaja, et mitte kõik mobiilsideoperaatorid ei ole võimelised lihtsal ja kiirel moel saada kätte CDR andmed (kuna andmete eksportimise funktsionaalsus ei pruugi olla ettevõtte süsteemis arendatud). See võib tähendada, et sarnaste analüüside läbiviimisel peab hindama väga tõsiselt andmete kättesaadavust.

5.4 Järeldused

Käesoleva lõputöö raames autor jõudis järelduseni, et CDR andmete abil on võimalik avastada teelõigud, kus ületatakse kõige sagedamini piirkiirust. Isegi vaatamata suhteliselt halvale lähteandmete kvaliteedile (ajatemplid 10-minutilise täpsusega, periood 2 nädalat, piiratud klientide arv), sai lõputöö autor usaldusväärseid tulemusi, mida riik saab kasutada liiklusohutuse tagamiseks. Kasutades sarnast lähenemist, on võimalik paremini planeerida investeeringuid liiklusohutusse. Selleks, et parandada sarnaste analüüsi tulemuste täpsust, soovitab töö autor:

- Kasutada lisaks SMS ja kõnede andmetele ka mobiilse interneti CDR andmeid, kuna sellisel juhul paraneb andmekirjete tihedus ja esineb vähem lünkasid andmetes;
- Parandada andmete kvaliteeti: mobiiltelefoni operaatorid salvestavad CDR andmed sekundi täpsusega, ja selline täpsusaste tagab, et kiiruse arvutamisel esineb oluliselt väiksem viga;
- Võimalusel kasutada terve antennide võrgustiku infot, kuna see aitab paremini määrata inimese asukohta;
- Võimalusel kasutada mitme kuu või isegi aasta andmed, kuna liikumisharjumused aasta sees võivad erineda, samuti aitab see märgata muudatusi peaaegu reaalajas;
- Kasutada ka ühistranspordi (eelkõige rongide) andmeid, kuna rongid võivad liikuda sõiduautodest kiiremini ja anda valepositiivseid signaale

- Üksikantennide asemel kasutada tsoone (linnad või linnaosad), kuna see aitab koondada kokku andmed erinevatelt antennidelt ja teha järeldusi suurema andmehulga põhjal, mis omakorda aitab tõsta analüüsi tulemuste täpsust.

Käesoleva lõputöö raames on autor mitu korda rõhutanud, et analüüsi läbiviimisel peab tegema eeldusi ja lihtsustusi. Need on seotud väga tihti just asjaoluga, et D4D väljakutse korraldajad tegid kõike selleks, et tagada Orange S.A. klientide anonüümsust ja Sonatel antennide võrgustiku andmed saladuses. Seoses sellega ei ole võimalik arvutada välja aega paremini kui +/- 10 min täpsusega. Selleks, et tulevikus vältida sarnaseid olukordi, kus on vaja otsida tasakaalu andmete konfidentsiaalsuse hoidmise ja analüüsi kasulikkuse vahel, näeb töö autor kahte varianti. Esimeseks lahenduseks oleks analüüsi läbiviimine mobiilsideoperaatori poolt. Selle lahenduse eeliseks on muuhulgas see, et mobiilsideoperaatori töötajatel on lihtsam tõlgendada CDR ja antennide andmeid. Teiseks lahenduseks oleks kolmanda osapoole kasutamine, kes tagaks andmete anonüümsust.

Üks selline lahendus on näiteks Opal, mis eksisteerib just selleks, et lahendada andmete jagamise probleemi suurandmete analüüsi projektides. Ettevõte toimib niimoodi, et kõik kliendi poolt edastatud anonüümsed andmed hoitakse Opali serverites. Kõik analüütikud, kes on huvitatud andmetega töötamises, saavad andmete näidist koos kirjeldusega algoritmide välja töötamiseks. Peale koodi kirjutamist jagavad analüütikud oma koodi Opaliga ja analüüsi viiakse jällegi läbi Opali serverites, väljastades ainult kokkuvõtet analüüsi tulemustest. (Opal koduleht) Autori arvates on selline lahendus väga mugav, kuna võimaldab analüütikutel töötada eelnevalt töötlemata andmetega ja tagab klientidele andmete konfidentsiaalsust.

Peale D4D väljakutse läbiviimist avalikustasid Sonatel ja Orange raporti selle tulemuste kokku võtmiseks. Raportist saab lugeda välja, et D4D väljakutses on osalenud kokku 260 uurimislaborit, millest ainult 11 asuvad Senegalis. 40% kõikidest esitatud uuringutest viidi läbi just transpordi ja urbanismi teemal. (Sonatel ja Orange raport) See näitab suurt huvi CDR andmete vastu rahvusvahelisel tasandil. Paljud andmeteadlased ja analüütikud näevad potentsiaali CDR andmete kasutamises transpordivõrgustiku planeerimisel ja optimeerimisel, mis on otseselt seotud ka käesoleva lõputöö teemaga. Seoses sellega usub autor, et CDR andmete kasutamine teevõrgustiku optimeerimiseks ja turvalisemaks muutmises muutub rohkem levinuks nii arenevates, kui ka arenenud riikides.

6 Kokkuvõte

Käesoleva lõputöö uuringu teemaks on CDR andmete kasutamine piirkiiruse ületamise hindamiseks. Lõputöö raames soovis autor saada aru, millistes valdkondades ja milliste probleemide lahendamiseks on ajalooliselt kasutatud CDR andmeid ning kas need sobivad piirkiiruse ületamise hindamiseks. Selleks, et vastata nendele küsimustele, tutvus töö autor uurimistöödega, mis on tehtud *Data for Development* (D4D) 2012 ja 2013 väljakutsete raames, analüüsis D4D 2013 raames avalikustatud Senegali andmeid ja tegi intervjuud Telia Eesti AS eksperdiga, et saada eksperthinnangut kasutatud lähenemisele.

Analüüsi läbiviimisel pidi töö autor lahendama mitu probleemi, mis olid seotud andmete kvaliteedi ja tõlgendamisega. Sarnane probleem esineb peaaegu kõikide analüüside läbiviimisel, ja nagu ka teistes andmeanalüüsi projektides, pidi töö autor tegelema andmete puhastamise ja ettevalmistamisega. Analüüsi raames lõputöö autor töötas läbi 12 087 900 liikumise kirjet, filtreeris välja statistilised vead ja identifitseeris potentsiaalsed kiiruse ületamised. Peale seda valideeris töö autor tulemusi *Google Maps Distance Matrix API* abil ja sai kätte andmekirjed, mis viitavad kiiruse ületamisele. Nende abil hindas töö autor, kui palju kiiruse ületamisi kokku on ja mis teedel kõige rohkem kiirust ületatakse. Analüüsi tulemuseks on kaart, kus on näidatud teed, mida läbiti vaadeldavas perioodis vähemalt 10 korda ja kus on kahtlusi lubatud kiiruse ületamises.

Peamine probleem, millega pidi töö autor kokku puutuma, on see, et lähteandmetes on näidatud aeg ainult 10-minutilise täpsusega, mis teeb kiiruse arvutamise keerulisemaks. Lisaks sellele internetis leidub vähe infot ühistranspordist Senegalis. Kuid isegi vaatamata sellele sai lõputöö autor usaldusväärseid tulemusi, mis viitavad teatud problemaatilisematele teelõikudele, kus kiirust ületatakse keskmisest rohkem. Seoses sellega usub töö autor, et lõputöös kasutatud lähenemine on elujõuline ja see võimaldab avastada teelõike, kus inimesed ületavad kiirust keskmisest sagedamini. Selle lähenemise rakendamine reaalelus aitab riigil paremini planeerida liikluspolitsei tööd ja investeeringuid teevõrgustikku. Teades, mis teelõikudel ületatakse kõige sagedamini

liikluskiirust, saab riik suunata ressursse, et muuta liikluse ohutumaks ja perspektiivis päästa inimeste elusid. Seega autori hinnangul on käesolev lõputöö oluline eelkõige sotsiaalmajanduslikust aspektist.

Kokkuvõtteks, autori hinnangul käesolev lõputöö vastab püstitatud küsimustele ja annab piisavalt head ja mitmekülgset ülevaadet sellest, kas ja mis tingimustel saab kasutada CDR andmeid piirkiiruse ületamise hindamiseks. Autori arvates näidatud meetodil on mitu olulist eelist: (1) CDR andmete kogumine ei nõua täiendavaid investeeringuid riistvaradesse ega infrastruktuuri, (2) andmete analüüsimine on piisavalt lihtne ja seda võib teostada isegi tavalise lauaarvuti abil ilma suurte investeeringuteta ja (3) analüüsi tulemusi on lihtne tõlgendada ja teha otsuseid investeeringuteks teevõrgustikku.

Kasutatud kirjandus

Anaconda koduleht (04.12.2019): https://www.anaconda.com/why-anaconda/

ASIRT Road Safety Facts. Link raporti allalaadimiseks (09.11.2019): https://www.asirt.org/safe-travel/road-safety-facts/

M. Berlingerio, F. Calabrese, G. D. Lorenzo, R. Nair, F. Pinelli, M. L. Sbodio, 2013. *AllAboard: a System for Exploring Urban Mobility and Optimizing Public Transport Using Cellphone Data*

V. Blondel, N. de Cordes, A. Decuyper, P. Deville, J. Raguenez, Z. Smoreda, 2013. *Mobile Phone Data for Development. Analysis of mobile phone datasets for the development of Ivory Coast*

Core DNA. GDPR Explained In 5 Minutes: Everything You Need to Know (14.11.2019): https://www.coredna.com/blogs/general-data-protection-regulation

D. Doran, A. Fox, V. Mendiratta, 2015. *Where do we Develop? Discovering Regions for Urban Investment in Senegal*

Eesti Päevaleht, 10.11.2009. Uus seade hoiaks kihutajad ohjes: Eesti IT-firmad soovivad kõik autod varustada kiirust kontrolliva süsteemiga.

Elron Idasuuna Sõiduplaan alates 13. Detsembrist 2019 (24.12.2019): http://elron.ee/wp-content/uploads/2019/11/Idasuuna-sõiduplaan-al-13.12-6.pdf

S. van den Elzen, J. Blaas, D. Holten, J.-K. Buenen, J. J. van Wijk, R. Spousta, A. Miao, S. Sala, S. Chan, 2013. *Exploration and Analysis of Massive Mobile Phone Data: A Layered Visual Analytics approach*

Ericsson Mobility Report. June 2019. Link raporti allalaadimiseks (09.11.2019): https://www.ericsson.com/49d1d9/assets/local/mobility-report/documents/2019/ericsson-mobility-report-june-2019.pdf

D. Forte, A. de Donno, *Handbook of Digital Forensics and Investigation*, 2010

GeeksforGeeks, *OpenCV Python program for Vehicle detection in a Video frame* (17.11.2019): https://www.geeksforgeeks.org/opencv-python-program-vehicle-detection-video-frame/

Google Maps Platform. Distance Matrix API (01.01.2019): https://developers.google.com/maps/documentation/distance-matrix/start

GPS World, *What Exactly is GPS NMEA Data (17.11.2019): https://www.gpsworld.com/what-exactly-is-gps-nmea-data/*

R. Horak. *Telecommunications and Data Communications* Handbook, pp.111-113

ITU 2017. *Call Detail Record (CDR) Analysis: Republic of Liberia*

R. Kujala, T. Aledavood, J. Saramäki, 2016. *Estimation and monitoring of city-to-city travel times using call detail records*

A. Lima, M. De Domenico, V. Pejovic, M. Musolesi, 2013. *Exploiting Cellular Data for Disease Containment and Information Campaigns Strategies in Country-Wide Epidemics*

Y. de Montjoye, Z. Smoreda, R. Trinquart, C. Ziemlicki, V. D. Blondel, 2014. *D4D-Senegal: The Second Mobile Phone Data for Development Challenge*

M. von Mörner, 2017. *Application of Call Detail Records - Chances and Obstacles*

NetMob koduleht (23.11.2019): https://netmob.org

NetMob 2015. *Data for Development Challenge Senegal. Book of Abstracts: Scientific papers*

OICA International Organization of Motor Vehicle Manufacturers veebileht (09.11.2019): http://www.oica.net/category/vehicles-in-use/

OPAL koduleht (01.01.2020): https://www.opalproject.org/general-overview

Pandas koduleht (15.12.2019): https://pandas.pydata.org

K. Petersen, *Business Telecom Systems: A Guide to Choosing the Best Technologies and Services*, 2000

Rhinocarhire – Guide to Driving in Senegal (02.01.2020): https://www.rhinocarhire.com/Drive-Smart-Blog/Drive-Smart-Senegal.aspx#/searchcars

SciPy.org koduleht, *Distance calculations (scipy.spatial.distance.cdist)*. (04.12.2019): https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.cdist.html

I. E. Semassel, M. Kachroudi, R. Azzi, S. B. Yahia, G. Diallo, 2019. *Early public health emergency anticipating with non health data per se*

Smartbridge: *Data Done Right: 6 Dimensions of Data Quality* (02.01.2020): https://smartbridge.com/data-done-right-6-dimensions-of-data-quality/

Z. Smoreda SENSE (Orange Labs, Paris) 2017. *Mobile phone data for migration analysis: new possibilities, risks and difficulties*

Sonatel ja Orange raport, 2015. *Orange Data for Development Challenge in Senegal*

SqLite koduleht. *Appropriate Uses For SQLite* (01.01.2019): https://www.sqlite.org/whentouse.html

TalTech, Tõhus koostöö: 900 anduri võrk kogub andmeid linnaõhu ja liikluse kohta (17.11.2019): https://www.ttu.ee/ttu-uudised/uudised/instituudid/tarkvarateaduse-instituut-2/tohus-koostoo-900-anduri-vork-kogub-andmeid-linnaohu-ja-liikluse-kohta/

Tartu Ülikool, Geograafia osakond. Projektsiooni omadused (08.12.2019): http://www.geo.ut.ee/kooligeo/EGCD/opik/juts/karto/proj_omad.html

TechTerms, *GPS* (17.11.2019): https://techterms.com/definition/gps

The Guardian, 2014. *New York taxi details can be extracted from anonymised data, researchers say* (14.11.2019): https://www.theguardian.com/technology/2014/jun/27/new-york-taxi-details-anonymised-data-researchers-warn

UN Global Pulse, *Winning Research from the Data 4 Development Challenge* (06.05.2013): https://www.unglobalpulse.org/D4D-Winning-Research

UN Global Pulse, *The Second "Data for Development" (D4D) Challenge in Africa* (09.06.2014): https://www.unglobalpulse.org/D4D-Challenge-Senegal

N.Vogel, C. Theisen, J.P. Leidig, J. Scripps, D.H. Graham, G. Wolffe, 2015. *Mining Mobile Datasets to Enable the Fine-Grained Stochastic Simulation of Ebola Diffusion*

X. Wang, H. Dong, Y. Zhou, K. Liu, L. Jia, Y. Qin, 2017. *Travel distance characteristics analysis using call detail record data*

Wikipedia, *Great-circle distance* (08.12.2019): https://en.wikipedia.org/wiki/Great-circle_distance

Worldometers, *Senegal population* (04.12.2019): https://www.worldometers.info/world-population/senegal-population/

Lisa 1. Intervjuu Telia Eesti AS esindajaga

Intervjuu läbi viidud: 10.12.2019
Küsimustele vastas: Ain Õiglane, mobiilsideteenuste omanik Telia Eesti AS-is
Enne intervjuu läbiviimist tutvustas autor lõputöös kasutatud metoodikat ja intervjuu läbiviimise eesmärgiks on saada hinnangut selle meetodi usaldusväärsusele.

Kas oskate öelda, kas ja millistes analüüsides on kasutatud Eestis CDR andmeid?

Eestis on tehtud rände uuringuid, mille abil on võimalik saada piirkondade tasemel aru, kus inimesed elavad ja kus nad töötavad. Oleme Telias ka nende analüüside läbiviimiseks andnud välja anonüümsel kujul andmeid.

Kui täpselt saab määrata inimese asukoha CDR andmete abil?

Peab arvestama, et mobiilside tugijaamad asuvad päris tihedalt linnades, aga maapiirkondades ühe tugijaama teeninduspiirkond võib olla kilomeetrites. Sõltuvalt regioonist võib ühe tugijaama teeninduspiirkond olla isegi 5 kilomeetrit või üle selle.

Samuti ei saa alati väita, et kui inimene teeb telefonikõne siis tema telefon on ühendatud just lähima antenni külge. Tugijaamal on teatud suunad, milles signaal levib, samuti eksisteerivad erinevad takistused hoonete ja maastiku näol, mis mõjutavad signaali tugevust. Seega võib juhtuda, et kõige tugevam signaal tuleb näiteks majade vahelt antennilt, mis asub tegelikult kaugemal.

Kas on mingi viis, kuidas asukoha määramist teha täpsemaks?

Selleks, et määrata inimese asukoha suurema täpsusega, peab analüütikul olema täielik andmebaas konkreetse mobiilsideoperaatori raadiovõrgust: mis on iga konkreetse masti täpne asukoht, selle suund ja eeldatav tööpiirkond. See on rangelt konfidentsiaalne info.

Kuivõrd hästi sobivad Teie arvates CDR andmed inimese liikluskiiruse määramiseks?

Eelkõige peab arvestama, et kui inimene teeb kõne, siis CDR andmestikus registreeritakse kõne algus ja lõpp. Ehk, need ei võimalda jälgida, kuidas liikus inimene kõne jooksul.

Kuna asukoha määramine toimub teatud veapiiriga, siis näiteks Tallinn-Saue sõidu puhul ei pruugi selline lähenemine anda korrektseid tulemusi, kuna vahemaa ei ole väga pikk. Ehk, siin peab kindlasti mõtlema, kui pikk vahemaa peab olema, et see mudel ennast õigustaks.

Milliste piirangutega peab arvestama sellise analüüsi läbiviimisel?

Minu arust selle lõputöö kontekstis tuleb mõelda, kuidas neid andmeid saab kasutada? Näiteks, kui panna Teie lähenemist Eesti tingimustesse. Kui me saame teada, et inimene on jõudnud Tallinnast Tartusse liiga kiiresti, siis see ei anna infot konkreetse teelõigu kohta, kuhu panna täiendav liikluskaamera või millega peab tegelema.

Kindlasti peab arvestama GDPR-i nõuetega, ehk selle analüüsi läbiviimiseks peavad andmed olema kindlasti anonüümsed. Samuti peab veenduma, et nendest andmetest ei saa *reverce engineering* meetodi abil luua seoseid konkreetsete inimestega. Sellistes projektides on juristide sõnadel väga suur kaal. Seadus näeb ette, mis tingimustel meie kui mobiilsideoperaator tohime CDR andmeid kasutada ja milleks mitte. Kindlasti ei saa me neid kasutada kommertseesmärkidel.

Kas Teil on soovitusi, kuidas saaks sarnast analüüsi teha täpsemaks?

Mina soovitaksin kasutada selleks andmeside andmeid. Inimesed teevad kõnesid ja saadavad sõnumeid aeg-ajalt, aga andmeside on telefonis kogu aeg püsti. Muidugi jääb alati asukoha määramine ebatäpseks, kuid nendelt andmetelt saaks tulla suurema granulaarsusega info, kuna andmed tulevad tihedamini.

Lisa 2. Kõige sagedasemad liikumistrajektorid

Lähtekoht	Sihtkoht	N	Keskmine aeg (CDR)	Vahemaa (google)	Eeldatud aeg (Google)	Reaalne kiirus (vahemik)	Lubatud kiirus
1057 (14.881638, -15.874619)	604 (14.948716, -16.818662)	2507	92	120	114	70 - 88	63
604 (14.948716, -16.818662)	1057 (14.881638, -15.874619)	1581	92	121	118	71 - 89	61
1057 (14.881638, -15.874619)	607 (14.966171, -16.81996)	1022	92	121	118	71 - 88	62
607 (14.966171, -16.81996)	1057 (14.881638, -15.874619)	832	92	123	123	73 - 91	60
1055 (14.861432, -15.873642)	599 (14.939203, -16.827819)	400	91	119	108	70 - 88	66
599 (14.939203, -16.827819)	1055 (14.861432, -15.873642)	331	93	122	110	71 - 88	67
570 (14.811445, -16.912631)	609 (14.953608, -16.806563)	220	20	21	28	43 - 128	45
609 (14.953608, -16.806563)	570 (14.811445, -16.912631)	194	20	21	28	43 - 128	46
1058 (14.875379, -15.864323)	599 (14.939203, -16.827819)	147	95	121	115	69 - 86	63
599 (14.939203, -16.827819)	1058 (14.875379, -15.864323)	145	93	122	119	71 - 88	61
599 (14.939203, -16.827819)	1069 (14.857921, -15.856952)	143	91	124	113	74 - 92	66
1069 (14.857921, -15.856952)	599 (14.939203, -16.827819)	138	92	123	110	72 - 90	67
570 (14.811445, -16.912631)	599 (14.939203, -16.827819)	115	20	18	22	36 - 109	50
1073 (14.866089, -15.852217)	599 (14.939203, -16.827819)	108	93	124	115	72 - 89	65
599 (14.939203, -16.827819)	1043 (14.866479, -15.87807)	107	91	120	111	71 - 89	64
599 (14.939203, -16.827819)	1073 (14.866089, -15.852217)	106	93	126	117	73 - 91	64
484 (14.749488, -17.145683)	529 (14.801999, -16.971878)	104	20	26	40	52 - 157	39
604 (14.948716, -16.818662)	1019 (14.882788, -15.897334)	101	82	119	112	78 - 100	64