TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Priit Tiganik 175756IDDR

Telefonirakenduse ajalis-ruumilise kasutuskäitumise profileerimine

Diplomitöö

Juhendaja: Priit Järv

PhD

Tallinn 2021

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: Priit Tiganik

17.05.2021

Annotatsioon

Käesoleva lõputöö eesmärgiks on telefonirakenduse ajalis-ruumiliste kasutusmustrite välja töötamine ja analüüs. Selleks analüüsitakse varasemaid sarnase valdkonna uurimistöösid ning selle alusel sünteesitakse kolm erinevat võimalust mustrite arvutamiseks: reeglite, kasutaja aktiivsuse ja linna piirkondade vahel liikumise alusel. Töö käigus antakse ülevaade andmete töötlemisest ja kasutusmustrite arvutamisest, nende omapäradest ning analüüsitakse, kas ja kuidas saab nende arvutamist alustada telefonirakenduse ettevõtte andmeplatvormil.

Lõputöö on kirjutatud eesti keeles ning sisaldab teksti 29 leheküljel, 5 peatükki, 11 joonist, 1 tabelit.

Abstract

Phone App Time and Geographical Usage Pattern Profiling

Given thesis has the goal of developing usage patterns and analyse them for spatio-temporal data generated by phone app. To achieve this previous works in similar field have been analysed and based on that 3 new candidates for usage patterns are developed: rule based, user activity based and based on movement between different areas of the city. An overview is given about the process of processing the data and how the usage patterns are calculated, how they compare to each other and it is analysed if and how it would be possible to start periodic usage pattern calculation on a data platform.

The thesis is in Estonian and contains 29 pages of text, 5 chapters, 11 figures, 1 table.

Lühendite ja mõistete sõnastik

ajalis-ruumilised andmed	*Spatio-temporal data*. Andmepunktid, millel on ruumilised (näiteks geograafilised koordinaadid) ja ajalised mõõtmed.
ETL	*Extract, transform, load*; andmete hankimine, transformeerimine ja salvestamine andmebaasis.
heksagon	Antud töös on linn jaotatud sama suurteks umbes 250 meetrise läbimõõduga kuusnurkadeks. Sessiooni alguse või lõpu koordinaadid on ümardatud selle heksagoni tasemele. Antud töös kasutatud h3 implementatsiooni.
kasutusmuster	Antud töös sarnasel individuaalsel eesmärgil tehtud ajalis-ruumiliste liikumiste kogum.
klasterdamine	Sarnaste objektide kogumite leidmine nende objektide tunnuste alusel. Vt ka klasteranalüüs.
sessioon	Antud töös kasutaja soov liikuda telefonirakenduse abil kahe punkti vahel. Sessiooni kohta on teada alguse ja lõpu koordinaadid ning alguse ja lõpu kellaaeg.

Sisukord

Jooniste loetelu

Tabelite loetelu

1 Sissejuhatus

Käesoleva uurimistöö eesmärk on uurida linnasisest liikumist pakkuva telefonirakenduse ajalis-ruumilisest kasutamisest tekkivaid kasutusmustreid ning neid kasutusmustreid sarnasuse alusel grupeerida.

Nutitelefonid koos nendele loodud rakendustega on inimeste igapäevase suhtluse, meelelahutuse ning info kogumise ja töötlemise asendamatud kaaslased. Olles pidevalt võrguühenduses ja pakkudes võimalust kasutaja asukoht GPSiga täpselt määrata, on üheks nutitelefonidele loodud teenuseks reaalajas asukohapõhiseid andmeid kasutavad rakendused.

Ajalis-ruumilisi andmeid (*spatio-temporal data*) ja inimeste käitumist on palju uuritud ka varem, kuid enamasti on nende tööde eesmärgiks olnud inimeste kui massi [1], [2] või vahel ka kui indiviidi [3]–[6] liikumise ennustamine tulevikus. See võib olla kasulik erinevate avalike teenuste (transpordivõrgustik, ühistransport, linnaruumi planeerimine jne) analüüsil ja arendamisel, kuid võib pakkuda vähe infot individuaalse kasutaja liikumisharjumuste kohta. See info omakorda võib olla aga oluline ettevõtetele, kes saavad selle teadmise alusel oma kasutajatele paremaid aeg-ruumilisi teenuseid pakkuda.

Antud töös kasutatakse linnasisese autotranspordi rakendusest saadavaid andmeid ning liikumise ennustamise asemel on eesmärgiks mõista, kuidas või mis eesmärkidel erinevad kasutajad seda rakendust enda igapäevaelu liikumisvajaduse rahuldamiseks kasutavad. Neid käitumismustreid saab ettevõte kasutada paljude enda valdkondade, nagu suhtlus kasutajatega, rakenduse disain, hinnastamine, turundus, kampaaniad jne., arendamisel.

Andmete, millega tööd tehakse, algühik on sessioon ehk rakenduse avamine ja sihtkoha sisestamine. Sessiooni kohta on teada selle alustanud kasutaja ID, algus- ja lõpp-punkti koordinaadid ning sessiooni alustamise ja lõpu aeg. Iga sessiooni kohta on ettevõttel kasutada ka muud metainfot (näiteks nagu rakenduse enda kasutamine, mõned väljad lisainfot kasutaja profiililt, täpsem info sessiooni kohta jne), kuid antud projekti puhul sellega ei arvestata. Sessiooni tasemele keskendumine muudab antud töö erinevaks varasematest, kus tihti on alamaks tasemeks kasutaja. Antud ülesande puhul saab selleks

olema sessioon ja **hakatakse otsima ja grupeerima kasutaja sarnaseid sessioone**. Näiteks liiguvad inimesed linnas ringi väga paljudel erinevatel eesmärkidel: perioodiliselt töö või kooli ja kodu vahet, tööpäeva sees tööasju ajades, õhtuti trennides või meelelahutust nautides, nädalavahetusel kauplemas või vaba aega veetmas jne. Neid sarnaseid liikumisvajadusi nimetatakse antud töös **kasutusmustriteks**. Samal inimesel võib olla mitu erinevat kasutusmustrit. Lisaks võivad kasutusmustrid ja vajadus antud rakenduse järele aja jooksul ka muutuda, kui näiteks muutub kasutaja elukoht, töö sisu vms.

Erinevaid andmete grupeerimise meetodeid on palju, kuid **ärilistest nõuetest** tulenevalt on oluline see, et need grupid oleksid inimesele mõistetavad ning meetod tagaks selle, et andmete gruppi määramine oleks ajas muutumatu. Lisaks peab olema võimalik nende kasutusmustrite arvutamine ettevõtte andmeplatvormil. See seab omakorda piiranguid grupeerimise meetodi keerukusele ja kulukusele: lahendus peab olema välja töötatav mõistliku kuluga, toetatav praeguse andmeplatvormi poolt, olema teostatav lühikese aja jooksul, arvutatav perioodiliselt väga suure hulga ja ajas kasvavate andmete pealt ning olema tulevikus laiendatav. Käesoleva töö eesmärgiks pole optimaalseima kasutusmustri meetodi välja valimine, vaid ainult nendest ülevaate ja võrdluse loomine.

Töö on jaotatud järgmiselt. Järgnevas, teises peatükis käiakse üle varasem asjakohane uurimistöö ajalis-ruumiliste andmetega ning tutvustatakse peamisi selliste andmete grupeerimise meetodeid, mis võiksid aidata uurimisprobleemi lahendada. Lisaks kirjeldatakse lühidalt metoodikat, millega andmeid tulemuse saavutamiseks töödeldakse.

Töö kolmandas peatükis kirjeldatakse detailsemalt kogu protsessi andmete pärimisest kuni nende arvutamiseni andmeplatvormil. Lisaks protsessi kirjeldamisele antakse põgus ülevaade andmetest enestest.

Neljandas peatükis antakse esiteks täpsem ülevaade ärilistest nõuetest, võrreldakse ja analüüsitakse grupeerimise meetodeid erinevate ärinõuete vaatevinklist ning tehakse soovitus, millist meetodit võiks ettevõte kasutama hakata.

Viimane, kokkuvõttev osa annab kogu töö tulemustest ülevaate.

2 Varasem kirjandus ja metoodika

Järgmises alampeatükis kirjeldatakse peamiseid andmeallikaid ajalis-ruumilistele uurimistöödele ning antakse ülevaade, milliseid peamiseid probleeme nendega on lahendatud. Teises alampeatükis antakse ülevaade ajalis-ruumiliste andmete kasutamises kasutusmustrite või kasutajate, kui mustreid pole eraldi vaadeldud, grupeerimise meetoditest ja tööde tulemustest erinevate kasutusmustrite defineerimisel. Kolmandas alampeatükis kirjeldatakse antud töö andmetest lähtuvalt meetodeid nende töötlemiseks.

2.1 Varasem uurimistöö ajalis-ruumiliste andmetega

Inimeste asukoha ajalis-ruumiliste (liikumise) andmete kasutamine erinevatel eesmärkidel on huvi pakkunud paljude valdkondade teadlastele. Tänu mobiilside tehnoloogia ja nutitelefonide arengule ja levikule on nende andmete allikad viimaste aastakümnete jooksul muutunud palju mitmekesisemaks. Näiteks koos mobiiltelefonide tekke ja levikuga tekkis võimalus kasutada telefoni umbkaudset lokaliseerimist mobiilimastide asukoha järgi [3], [7], [8]. Need andmed katavad kasutaja terve päeva aktiivsuse ja koos mobiiltelefonide väga laia levikuga katavad need andmed ka suure enamuse populatsioonist, mistõttu on see andmete kogumise meetod jätkuvalt teatud probleemide lahendamisel asendamatu.

Peamine selliste mobiilside asukoha andmete piirang on nende täpsus. See võib ideaalsetes tingimustes olla kümnetes meetrites, halbades tingimustes sadades või enamates meetrites [9]. Andmete täpsust saab tunduvalt suurendada GPSi abil, ning seda tehnoloogiat on hakanud ära kasutama paljud teenused, mille kasutamisel salvestatakse kasutaja asukoht ja/või liikumistrajektoor. Transpordiga seotud näidetest on peamised näited kirjanduses linnade rattajagamise süsteemide andmete kasutamine [1], [2], kuid leidub ka töid autode jagamise süsteemidest [10]. Samasse kategooriasse võiks lugeda ka telefonirakendused, kus aja ja ruumi info on teisese tähendusega. Näiteks erinevate sotsiaalvõrgustike *check-in*-id [4], [5].

Kolmanda kategooria võimalikest andmetest moodustavad telefonirakendused, mille funktsioon poleks ilma ajakohase ja väga täpse asukohata võimalik. Näiteks erinevad navigatsiooni või sõidujagamise rakendused. Käesolevas uurimistöös kasutatakse antud viisil tekkinud andmeid.

Teise ja kolmanda kategooria andmete kasutamise piiranguks uurimistöödes on see, et need tekivad ainult juhul, kui kasutaja antud rakendust kasutab. Seda piirangut vähendab erinevate andmeallikate kombineerimine. Näiteks mobiilside asukoha ja sotsiaalvõrgustikest saadavate asukohtade alusel on võimalik täpselt modelleerida, kus inimesed liiguvad ja mida nad parasjagu teevad [5].

Osaliselt tingituna kasutatavatest andmetest on olnud erinevad ka uurimisküsimused, mida ajalis-ruumiliste andmetega on lahendatud. Alljärgnevalt on autor toonud mõned näited, mis on otseselt või kaudsemalt antud uurimisprobleemiga seotud:

- populatsiooni asukoha ennustamine või kirjeldamine ajas ja ruumis [1], [2];
- indiviidi asukoha ennustamine ajas ja ruumis [4], [5];
- indiviidi kasutusgrupi ennustamine trajektooride põhjal [3], [6];
- kasutajate grupeerimine sarnasuse alusel [8], [11], [12];
- visualiseerimine [7].

Eelnevast kirjanduse ülevaatest tuleneb, et vaatamata tööde eesmärgile on peamiseks analüüsitavaks üksuseks olnud kõige madalamal tasemel indiviid või kasutaja. Kasutaja käitumist ajas ja ruumis on küll kirjeldatud ja ennustatud tema asukohtadele/teekondadele kvalitatiivse info lisamisega, kuid autor ei leidnud antud uurimistöö probleemile sarnaseid uuringuid. Sellegipoolest on varasematest töödest kasu võimalike grupeerimise meetodite arendamisel. Järgnevas jaotises antakse sellest ülevaade.

2.2 Grupeerimise meetodite ülevaade

Antud lõike esimeses alapeatükis vaadeldakse varasemaid ajalis-ruumiliste andmete grupeerimise viise reeglite alusel ning teises alapeatükis kirjeldatakse samade andmete grupeerimist klasterdamise meetodite abil.

2.2.1 Ajatsoonid reeglite alusel

Mitmed autorid [1], [2] on rattajagamise süsteemide puhul täheldanud esiteks väga erinevat päevasisest sõitude jaotust antud transpordimeetodi kasutamisel tööpäevadel ja nädalavahetusel, ning teiseks, kuidas kasutus erineb alguse ja sihtkoha liikide (elamispiirkond, kontorite piirkond, puhkealad jne) ja päevade lõikes. Eriti ilmekalt tuleb erinev kasutus välja Brisbane'i rattajagamise süsteemi uurimisel [1], kus on tööpäevadel selgelt näha inimeste suuremat liikumist hommikuti tööle ja õhtuti töölt koju ning nädalavahetusel näiteks suuremat liikumist parkide läheduses ning vähem töö- ja kodupiirkondade vahel. Kellaaega ja nädalapäeva, sealjuures ka asukoha populaarsust nende lõikes, on kasutatud näiteks kas kasutaja järgmise *check-in*-i ennustamiseks [4]. Samas uurimuses [4] leiti ka, et ajaliste näitajate ennustusvõime oli mõnevõrra madalam kui näiteks asukohtade kategooriate või kasutaja varasema ajalooga seotud muutujad.

Mõnevõrra sarnaselt [4] tööle võib kasutaja sessioonid ja algus-/sihtkohad jaotada ka populaarsuse järgi, eristades näiteks kõige populaarsemad sihtpunktid, alguspunktid või punktid vaatamata sellele, kas tegemist oli alguse või lõpuga. Sarnaselt saaks leida ka populaarseimad teekonnad ja teha seda sarnaselt punktidega ehk vaadelda teekonda kui kas suunatud või suunamata graafi küljena. Eelduslikult näitab punkti või teekonna populaarsus selle olulisust kasutaja jaoks ning seetõttu võiks tegemist olla ka tähtsuse järjestamisega. Erineva tähtsusega punktid ja teekonnad võivad, kuigi ei pea, omama kasutaja jaoks erinevat sisu. Samas ei anna see meetod lisainfot sessiooni eesmärgi kohta.

Edasi on võimalik kahte eelnevat veel kombineerida ning leida näiteks kõige populaarsemad teekonnad mingil ajaperioodil. Ühe näitena võiksime leida kõige populaarsema teekonna hommikuse tipptunni ajal ja eeldada, et tegemist on sessiooniga (kontori)tööle jõudmiseks. Eeldades teadvat üldiseid mustreid inimeste liikumises on seeläbi võimalik sessioonile kvalitatiivne sisu omandada. Küsimuseks jääb muidugi, kui täpsed saavad intuitiivsed eeldused olla: on ka olemas ju öised vahetused ning ehk kasutatakse antud rakendust suhteliselt rohkem just ajal, mil alternatiivid nagu ühistransport, kasutatavad ei ole.

2.2.2 Erinevad klasterdamise meetodid

Sarnaste gruppide automaatse tuvastamise peamine metoodika on klasterdamine. Sellel on palju eriliike [13], kuid kuna nende põhimõte on enamasti „lähedaste" andmepunktide

sidumine, siis geograafilised andmed sobivad neile väga hästi. Seda on ka palju tehtud, kuid sellisel juhul on sisendandmetena kasutatud kogu populatsiooni. Üheks heaks näiteks on [2] töö, kus, kasutades linna rattajagamise andmeid, tuvastati klasterdamise abil piirkonnad, mis olid omavahel sõitudega sarnaselt ühendatud. Sarnast metoodikat saaks kasutada ka antud uurimisprobleemi lahendamiseks, kuid veidi teise külje alt: arvutada linna erinevatele piirkondadele rakenduse kasutamise ajaloo põhjal erinevad meetrikad (näiteks sessioonide arv, hommikuste sessioonide osakaal kõikidest sessioonidest, nädalavahetuse sessioonide osakaal jne) ning nende alusel klasterdada linn erinevateks piirkondadeks. Kui klastrid tunduvad loogilised, saab edasi uurida kasutaja sessioone nende piirkondade vahel.

Teine võimalus oleks meetodit kasutada analüüsitavate sessioonide peal. Tulemuseks võib sellisel juhul olla just see, mida uurimisprobleem lahendada üritab: erinevad sessioonid on grupeeritud sarnastesse gruppidesse. Selle lahenduse negatiivseks küljeks võib olla küll selle interpreteeritavus: vastavalt andmete muutumisele ajas saame ilmselt igal arvutusel erinevad klastrid ja need ei pruugi olla võrreldavad eelnevate perioodide arvutustega. Niisiis seda meetodit töös ei kasutata.

Klasterdamise meetodeid on kasutatud ka aja ja sageduse andmete grupeerimisel. See võib olla alternatiiv eelmises jaotises kirjeldatud reeglite alusel nädala ajaperioodideks jaotamisele. Hea näide selleks on *spectral bi-clusteringi* metoodika rakendamine ühe sõidujagamisplatvormi autojuhtide erinevate aktiivsuse gruppide kirjeldamisel [11]. Viidatud töös kirjeldati meetodi tulemusena autojuhte, kes olid aktiivsed pigem tööpäevade tööajal (ilmselt oli sellisel juhul rakenduse teenistus nende peamine tuluallikas) või õhtuti ja öösiti (ilmselt teeniti lisatulu pärast enda päevatööd). Samas, see meetod töötab kasutaja tasemel. See tähendab seda, et meetod ühendab sarnase kasutusmustriga kasutajad. Kuna aga töö eesmärgiks on kasutaja **erinevate** kasutusmustrite tuvastamine (eeldus on, et samal kasutajal võib olla erinevaid kasutusmustreid), ei saa ka seda meetodit probleemi lahendamiseks rakendada.

2.3 Andmete töötlemise ja grupeerimise protsess

Ettevõtte analüütikaks ja raporteerimiseks kasutatavaid andmeid hoitakse Amazoni Redshifti PostgreSQLil põhinevas andmebaasis [14]. Redshifti PostgreSQLi implementatsioon on selle andmebaasi keele kohandatud versioon, niisiis pole selles

võimalik kasutada kogu PostgreSQLi funktsionaalsust. Õnneks ei ole küll andmete päring liiga keeruline, kuid see võib olla üheks põhjuseks, miks võiks suurema osa andmete töötlemisest teha andmete pärimisele järgnevas analüütika etapis.

Analüütika etapis kasutatakse Pythoni programmeerimiskeelt. Sellele keelele on loodud hulgaliselt lisapakette andmetabelitega töötamiseks, mis muudab selle keele antud ülesandeks eriti sobivaks. Antud töös on peamised kasutatavad paketid numpy [15], pandas [16], scikit-learn [17]. Olemas on ka eelnevas peatükis toodud grupeerimismeetodite implementatsioon selles keeles.

Peale andmete kättesaamist andmebaasist on esmane ülesanne nendest hea ülevaate saamine. See aitab teha esialgse töötluse ja selle tulemusena näiteks eemaldada või parandada kahtlased või mitte kasutatavad andmed. Selles etapis tuleb lähema vaatluse alla võtta ka geograafilised koordinaadid ja neid „ümardada“, ehk nende täpsust vähendada. Vastasel korral jääks sessiooni alguse ja lõpu punkt väga palju sõltuma konkreetse hetke GPSi täpsusest või sellest, kus parasjagu kasutaja sessiooni alustades või lõpetades seisis. Täpsuse vähendamine vähendab ka tunduvalt võimalust, et nende andmete põhjal oleks võimalik kasutaja personaliseerida. Koordinaatide täpsuse vähendamiseks on mitmeid võimalusi [18]. Üks võimalus selleks oleks koordinaate sisuliselt ümardada, mille tulemusena jaotataks punktid ruumis ruudukujulistesse piirkondadesse. See on lihtne ja efektiivne viis, kuid antud töös on kasutatud koordinaatide teisendamist heksagonidesse, kuna see töötab paremini just ruumis liikumise analüüsimiseks [19]: selle puhul on kõik naaber-heksagonid samal kaugusel. Selleks kasutatakse h3 [20], [21] paketti. h3 paketis on juba sisse ehitatud kasulikud funktsioonid heksagonide naabrite leidmiseks, mis võib uurimistöös kasulikuks osutuda

Puhaste andmetega edasi liikudes arvutatakse juurde hulk üldisemaid lisamuutujaid. Näiteks sessiooni tund, kuupäev, nädalapäev. Seejärel saab asuda arvutama grupeerimismeetodeid. Seda on lähemalt kirjeldatud alapeatükis 3.3.

Visualiseerimine on andmeanalüüsis oluline osa ning seda tehakse nii andmetest ülevaate saamiseks kui ka üksiku kasutaja käitumise mõistmiseks.

Ettevõtte peamine andmete pärimiseks, töötlemiseks ja salvestamiseks (ETL, *extract transform load*) kasutatav tööriist on Pythoni keeles kirjutatud ja kasutatav vabavaraline

Airflow [22]. See on ehitatud Pythonis, niisiis on võrdlemisi lihtne analüütikast implementatsiooni edasi liikuda.

3 Ülevaade andmetest ja kasutusmustrite leidmine

Kui töö eelmises peatükis anti ülevaade, kuidas ja mida töö eesmärkide saavutamiseks oleks vaja teha, siis antud peatükis käiakse arendusprotsess täpsemalt läbi ning antakse ülevaade ka andmete struktuurist ning etteruttavalt ka analüüsitakse neid. Alustatakse andmete pärimise ja töötlemise kirjeldamisest ning seejärel antakse ülevaade andmete jaotumisest. Kolmandas alapeatükis kirjeldatakse kasutusgruppide arvutamist ning viimases, neljandas osas kirjeldatakse arvutust Airflow platvormil.

3.1 Andmete pärimine ja töötlemine

Andmeid, mida kasutatakse, on pärit Tallinnast. Perioodiks, mille jooksul kasutajate käitumist jälgitakse, on ajavahemik 2020-08-31 kuni 2020-09-27. See on 4 nädalat kasutamist. Kuigi autor teab varasemast kogemusest, et enamus liikumismustreid on tihti nädalase perioodilisusega (ehk mustrid korduvad igal nädalal, tulenevalt töö- ja puhkepäevade tsüklist), siis rohkemate nädalate kaasamine aitab kasutaja kohta koguda rohkem andmeid ning loodetavasti lisada ka kasutajaid, kelle kasutusmuster on juhuslik või igakuine (nt sõltudes palgapäevast). Iga sessiooni kohta päritakse sellest perioodist kasutaja ID, algus- ja lõpp-punkti koordinaadid ning sessiooni alustamise ja lõpu kuupäev ja kellaaeg. Päringus arvutatakse lisaks sessiooni alustamise tund, kuupäev, nädalapäev.

Esialgsete andmete puhul oli märgata, et üpris tihti on samal kasutajal järgmise sessiooni algus koheselt peale eelmise lõppu. Kuna selliseid näiteid oli palju, siis pidi see viitama pigem rakenduse sisemisele loogikale kui kasutaja vajadusele järgmine sessioon teha. Näiteks luuakse uus sessioon, kui kasutaja peale sihtkohta jõudmist rakenduse uuesti avab, näiteks lõpliku hinna uurimiseks või tagasiside jätmiseks. Seetõttu eemaldati andmetest sellised topelt-sessioonid, mis olid tekkinud vähem kui 30 minuti peale eelmise sessiooni lõppu.

Kui andmetes polnud sessioonil lõpp-punkti määratud, siis ka need sessioonid eemaldati. Eeldus on, et kui lõpp-punkti pole sisestatud, siis pole kasutajal olnud tegelikku huvi liikuda. Andmetena on kasutatud küll sessioonid, kuid sessioon ei tähenda tegelikult, kas kasutaja ka tegelikult liikus või mitte. Antud töö puhul on eeldatud, et kui sihtkoht on sisestatud, on kasutajal olnud soov siiski ka liikuda, aga ehk polnud rakenduse poolt

pakutud ooteaeg, hind või mõni muu tingimus sobilik ning kasutaja leidis alternatiivse ja sobilikuma viisi soovitud sihtkohta jõudmiseks.

Järgmine samm andmete töötlemises on heksagonide arvutamine. Autori poolt valitud implementatsioonis [21, p. 3] on maakera jaotatud erineva resolutsiooniga heksagonide tasemeteks. Autor on valinud resolutsiooni number 9, mille puhul on iga heksagoni läbimõõt umbes 350 meetrit ja pindala 0.1 km2. See tagab andmete piisava de-personaliseerimise ja samas grupeerib suurel enamusel samad tegelikud algus- ja sihtkohad samasse heksagoni. Antud implementatsiooni puhul saab iga heksagon oma unikaalse ID, millest on omakorda võimalik soovi korral genereerida kuue tipu koordinaatide nimekiri. Andmete mahu tõttu muutub implementatsioon aeglaseks, kuid seda aitab kiirendada funktsiooni vektoriseerimine.

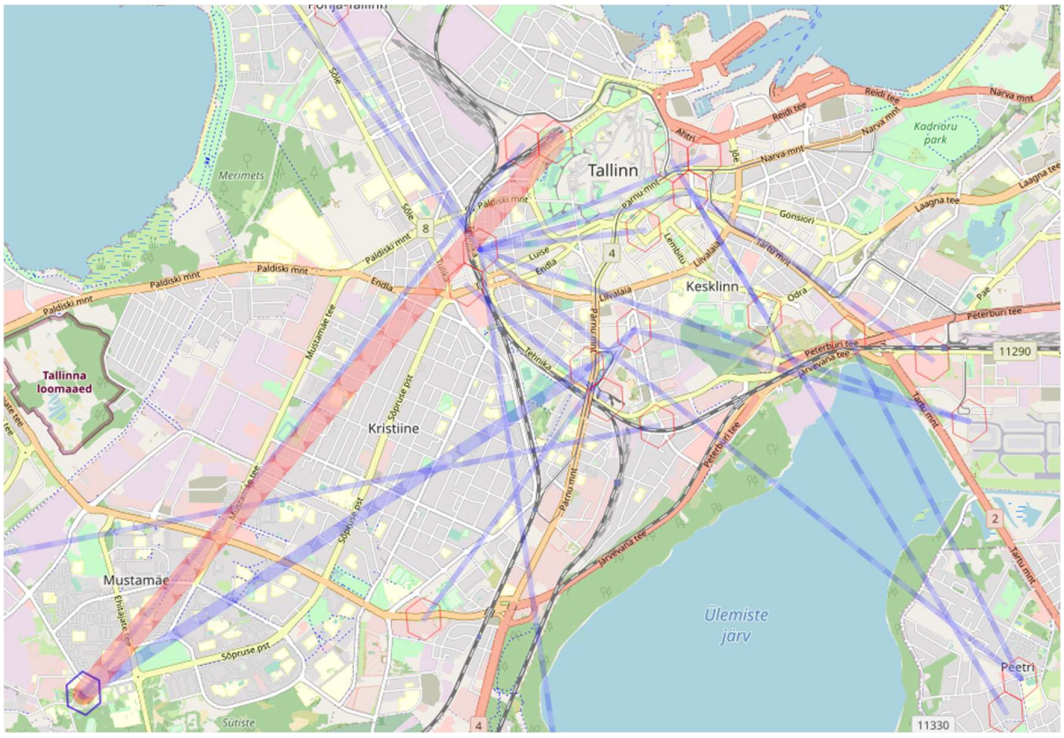
Lisaks ainult teekonna alguse ja lõpu punktidele saab andmetest välja joonistada ka sirgjoonelised teekonnad. Sisuliselt on see siis kahe tipuga suunatud graaf, kus tippudeks on vastloodud heksagonid. Lisavõimalusena luuakse ka suunamata kahetipulised graafid ehk teekonnad, tänu millele on hõlbus leida, kui tihti kasutaja kahe punkti vahel kokku, mitte ainult ühes suunas, liigub. Selleks luuakse kaks lisavälja, millest esimene on suunamata kahetipulise graafi esimene punkt ja selle väärtuseks saab lihtsalt madalama ID-ga heksagon ühe teekonna heksagonidest. Kahetipulise graafi teiseks punktiks saab järgi jäänud ehk kõrgema ID-ga heksagon.

Järgneval Joonis 1 antakse ülevaade töödeldud andmetest ühe rakenduse kasutaja (kelleks on autor) pikema aja kasutamise ajaloost Tallinnas.

Joonisel on ära märgitud algus- ja sihtkohtade heksagonid ja on näha, kuidas see lahendus suudab andmeid anonümiseerida ja grupeerida. Näiteks algavad/lõppevad (joonisel on kujutatud suunamata kahetipulised graafid) kõik IT Kolledži kasutused samast punktis: heksagoni keskpunktist. Samas on näiteks autori kasutus Balti Jaama ja Viru Väljaku lähedal jaotunud mitme heksagoni vahel, mis viitab kohati suurema heksagoni resolutsiooni vajadusele. Sinise värviga heksagon on kasutaja kõige populaarsem sihtkoht. Selle arvutamist kirjeldatakse alapeatükis 3.3.2.

Lisaks heksagonidele on joonisele kantud ka teekonnad. Nagu mainitud, vaadeldakse hetkel teekondasid suunamata kahetipuliste graafidena. Joonisel on laiema joonega tähistatud teekonnad, millel on tehtud rohkem sessioone. Antud juhul on selliseid märgata

2 tükki ning mõlema üks tipp on IT Kolledžiga seotud. Punaselt värvitud joonega on tähistatud populaarseim teekond. Selle arvutamist kirjeldatakse alapeatükis 3.3.2.



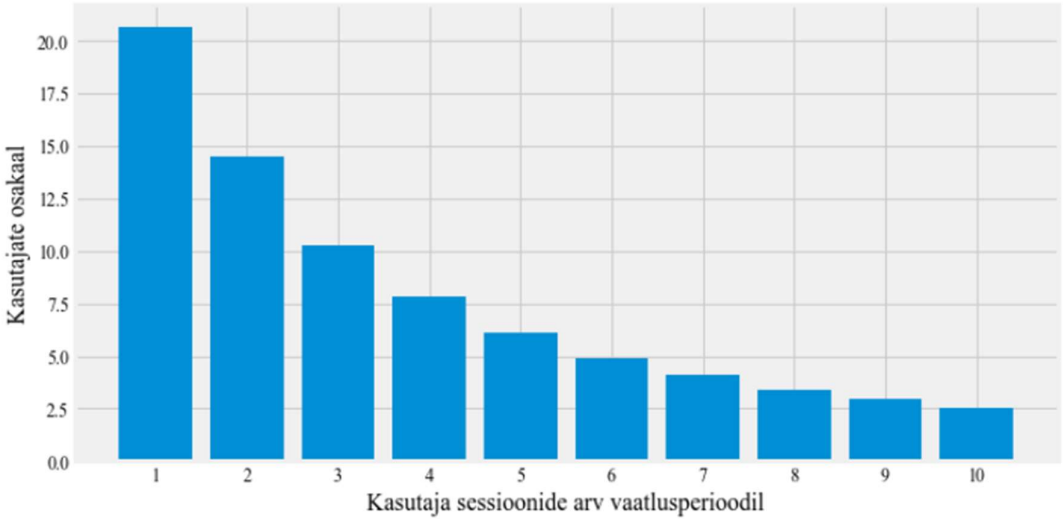
Joonis 1. Rakenduse ühe kasutaja kasutusmuster ruumis.

3.2 Andmete ja kasutajate ülevaade

Järgnevatelt joonistelt on sihilikult eemaldatud absoluutskaalad ja asendatud need suhteliste väärtustega. Ülevaadetes on ainult kasutajad, kes tegid vaatlusperioodi jooksul vähemalt ühe tingimustele vastava sessiooni, st jaotuste arvutusse pole lisatud mitteaktiivseid kasutajaid.

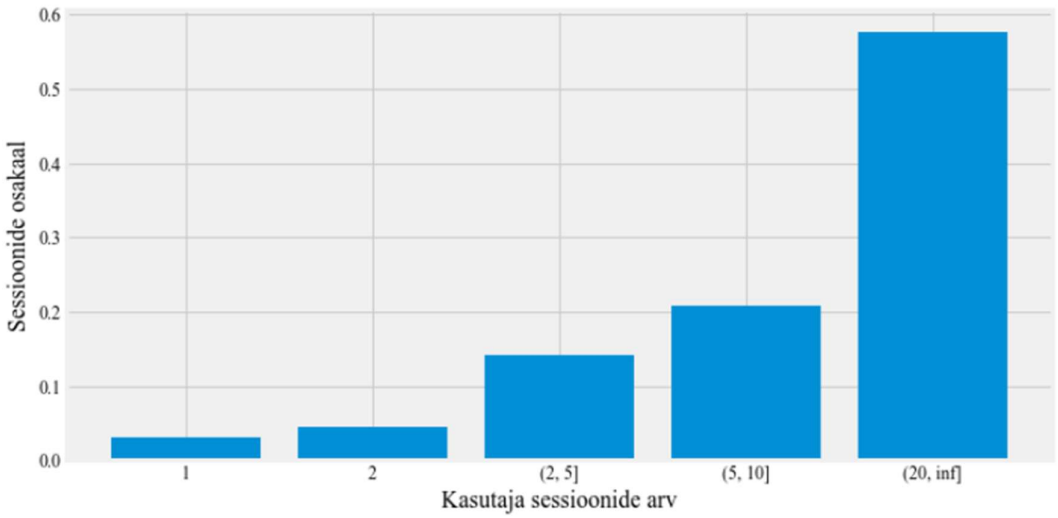
Kasutajate kasutussagedus on vaatlusperioodil erinev ning kasutuse sagedust kujutab Joonis 2 toodud histogramm. Joonisel on maksimaalseks väärtuseks valitud 10 sessiooni vaatlusperioodi ajal, kuid see ei ole maksimaalne sessioonide arv ühe kasutaja kohta. Sellegipoolest on jooniselt näha, et kõige suurem osa kasutajatest, veidi üle 20% aktiivsetest kasutajatest, on vaatlusperioodi jooksul teinud vaid ühe sessiooni. Koos kahe sessiooni tegijatega on madalama aktiivsusega kasutajate osakaal suisa 35%. Selge on see, et sellise väga pika parempoolse sabaga jaotuse (tuntud ka kui 80-20 reegel, *power-law*, Pareto jaotus) puhul võib kasutusaktiivsus ise olla üheks kasutusmustriks või peab vähemalt madalama kasutusaktiivsusega kasutajate puhul nende aktiivsust

kasutusmustrite loomisel arvesse võtma. Näiteks eemaldama need kasutajad valimist või määrama neile eraldi kasutusmustri, kuna ilmselt mingit erilist mustrit ühest või kahest sessioonist välja ei joonistu. Lõplik toiming sõltub ka kasutusmustri arvutamise metoodikast.



Joonis 2. Kasutajate aktiivsuse jaotus vaatlusperioodil.

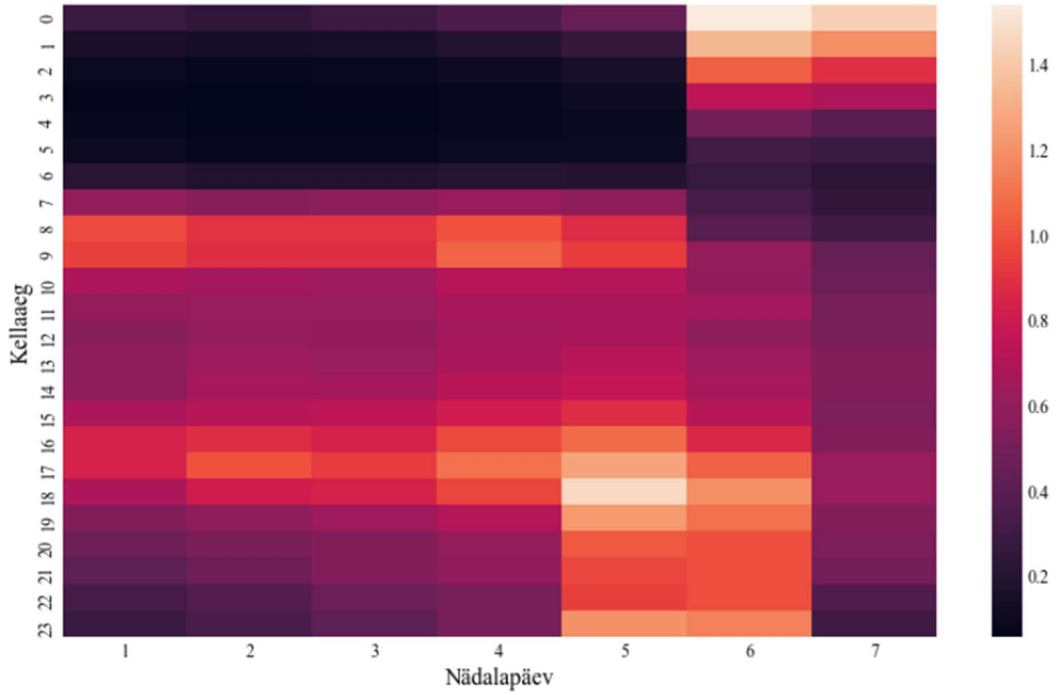
Joonis 3 on kujutatud kõikide sessioonide jaotust vastavalt sellele, kui palju neid sessioone teinud kasutaja perioodil sessioone tegi. Siit on selgelt näha, et kuigi kasutajaid, kes teevad vähe sessioone, on väga palju, panustavad nad kõikide sessioonide koguarvu marginaalselt.



Joonis 3. Kõikide sessioonide osakaal vastavalt kasutaja sessioonide arvule.

Rakenduse kasutus ei ole ajas ühtlane. Nagu varem mainitud, sõltub inimeste liikumine peamiselt iganädalasest töö ja puhkeaja tsüklist. See on selgelt nähtav Joonis 4, kus on

toodud vaatlusperioodi sessioonide jaotus päevade ja kellaaegade lõikes. See joonis ütleb palju, kuid ilmselt kõige erksamalt on näha tume ala tööpäevade öötundidel, mis öötundidel vastu laupäeva ja pühapäeva asendub väga kõrge aktiivsusega. Lisaks on tööpäeviti märgata selget hommikuse ja õhtuse tipptunni rütmi.



Joonis 4. Sessioonide jaotus kellaaja ja nädalapäeva lõikes. Heledam värv tähendab kõrgemat osakaalu antud sessioonidest. Väärtus 1.4 tähendab, et 1.4% nädala sessioonidest tehti sellel tunnil.

3.3 Kasutusmustrite arvutamine

Järgnevalt antakse täpsem ülevaade metoodika peatükis 2.2 välja toodud grupeerimismetoodikate arvutamisest.

3.3.1 Ajatsoonid

Vastavalt eelmise peatüki Joonis 4. Sessioonide jaotus kellaaja ja nädalapäeva lõikes. Heledam värv tähendab kõrgemat osakaalu antud sessioonidest. Väärtus 1.4 tähendab, et 1.4% nädala sessioonidest tehti sellel tunnil.Joonis 4 grupeeritakse sessioonid nende päeva ja kellaaja alusel. Seda on kujutatud järgneval Joonis 5. Gruppide definitsioonid hoitakse lihtsad, kuid samas üritatakse võimalikult hästi grupeerida tehtud sessioonid nende eeldatava otstarbe järgi. Näiteks tööpäeva hommikute ja õhtute ajal liiguvad inimesed tööle või koju ning linnaliiklus on sellel ajal tippkoormuse all. Tippkoormuse all pole linnaliiklus ilmselt aga joonisel märgitud pidude ajal, kuid antud rakenduse

kasutusstatistikast tundub see olevat jällegi suurima nõudlusega periood nädalas. Lisaks on eristatud veel kolm perioodi, millal tehtud sessioonid võiksid olla sarnased: tööpäeva päevane aeg töö- või ootamatute muude sessioonide tegemiseks, tööpäevade öine aeg, kui nõudlus on väga madal, ning nädalavahetuse päevane aeg, mil ilmselt kasutatakse rakendust tööpäevadest erineval eesmärgil, näiteks puhkeaja veetmiseks.

Joonis 5. Sessioonide ajatsoonidesse määramine.

3.3.2 Populaarseimad heksagonid ja teekonnad

Tuginedes andmete ülevaatele otsustas autor leida ainult ühe peamise heksagoni ja teekonna. Eelnevast oli näha, et suurel osal kasutajatest on vaatlusperioodil tehtud vähe sessioone, mistõttu oleks kasutajate arv, kellele saaks leida teise või kolmanda järgu tähtsusega heksagoni või teekonna, äärmiselt väike. Peamiste heksagonide määramisel on samaaegselt võetud arvesse nii alguse kui lõpu punktid ning peamise teekonna puhul pole arvestatud teekonna suunaga. Peamiste heksagonide ja teekondade arvutamise reeglite loomisel on autor soovinud kindel olla, et reegel tagaks suurema tõenäosusega kasutajale olulised punktid või teekonnad. Seetõttu on reeglites lisaks arvulistele muutujatele ka tingimused osakaaludele, näiteks kui kasutaja kastutab kümmet erinevat heksagoni, kuid ühegi arv või osakaal pole piisav, ei määratagi kasutajale peamise heksagoni tunnust.

Lisaks otsustas autor rakendada järgmiseid reegleid peamise heksagoni defineerimisel.

Heksagon on peamine heksagon, kui:

- kasutaja on seda kasutanud kõige rohkem, ja
- heksagoni on kasutatud üle kuue korra, või
- kui peamist heksagoni on kasutatud alla kuue korra, siis kui antud heksagoni osakaal kõikidest sessioonidest on ≥50%.

Teekond on peamine teekond, kui:

- kasutaja on seda kasutanud kõige rohkem, ja
- teekonda on kasutatud üle viie korra, või
- kui kasutajal on kokku ≥5 teekonda ning peamise teekonna osakaal nendest on ≥40%, või
- kui kasutaja teekondade arv on vahemikus [2,5) ning populaarseima teekonna osakaal on ≥60%.

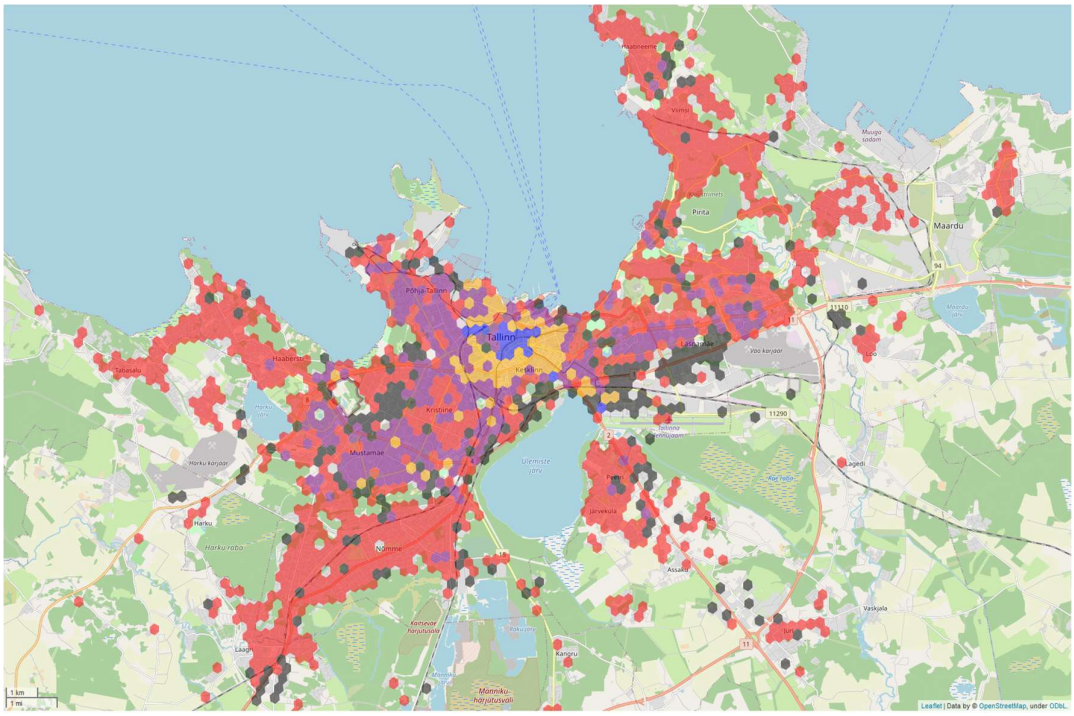
Nende reeglite alusel võib tulemuseks olla, et kasutajale pole võimalik peamist heksagoni või teekonda leida.

3.3.3 Heksagonide klasterdamine

Klasterdamise tarvis arvutas autor üle saja erineva tunnuse, mis kirjeldavad iga heksagoni kasutust alguspunti ja sihtkohana läbi erinevate ajaperioodide. Itereerides erinevaid kombinatsioone nendest jäid valikusse 36 sobivamat. Autor otsis klasterdamise tulemust, mis ei sõltuks nii palju heksagoni asukohast (eemaldati kõik teekonna pikkusega otseselt või kaudselt seotud tunnused. Nende olemasolu korral eristaks klasterdamine muidu kindlasti linna lähiasulad nagu Tabasalu, Laagri, Viimsi jne), omaks mingisugust infot heksagoni alguse ja lähtekohaks olemise kohta (kasutati eraldi tunnuseid, mis lugesid kokku heksagoni rolli alguse ja sihtkohana), omaks infot ajatsoonide kohta (jäeti sisse erinevates ajatsoonides tehtud sessioonide osakaalud kõikidest sessioonidest), kuid samas ei sõltuks ülemäära palju sessioonide koguarvust (sessioonide koguarvu kasutati, aga sellega oli seotud vaid kolm muutujat) ning omaks kvalitatiivset infot heksagoni olulisuse kohta kasutajale (lisati sihtarvud, kui paljud sessioonid lõppevad ka reaalse liikumisega). Peale tunnuste väljavalimist ja enne klasterdamist skaleeriti nende suurused enne samasse vahemikku.

Klasterdamise meetodina langes valik K-keskmiste kasuks. Autor proovis lisaks läbi meetodeid nagu DBSCAN, *AffinityPropagation* ja *AgglomerativeClustering*, kuid K-keskmiste tulemus tundus autorile, kes Tallinna geograafiaga mõnevõrra kursis, täiesti piisaval heal tasemel. Klastrite arvuks valiti visuaalset „küünarnuki meetodit“ (*elbow method*) kasutades 5. See on ka piisav arv klastreid, mida on veel võimalik hallata ja millega saab tööd edasi teha.

Joonis 6 on toodud erinevate heksagonide klasterdamise tulemus. Lisaks on Joonis 7 täpselt sama info, aga keskendutud on Mustamäele.



Joonis 6. Tallinna heksagonide K-keskmiste klasterdamismeetodi tulemus.

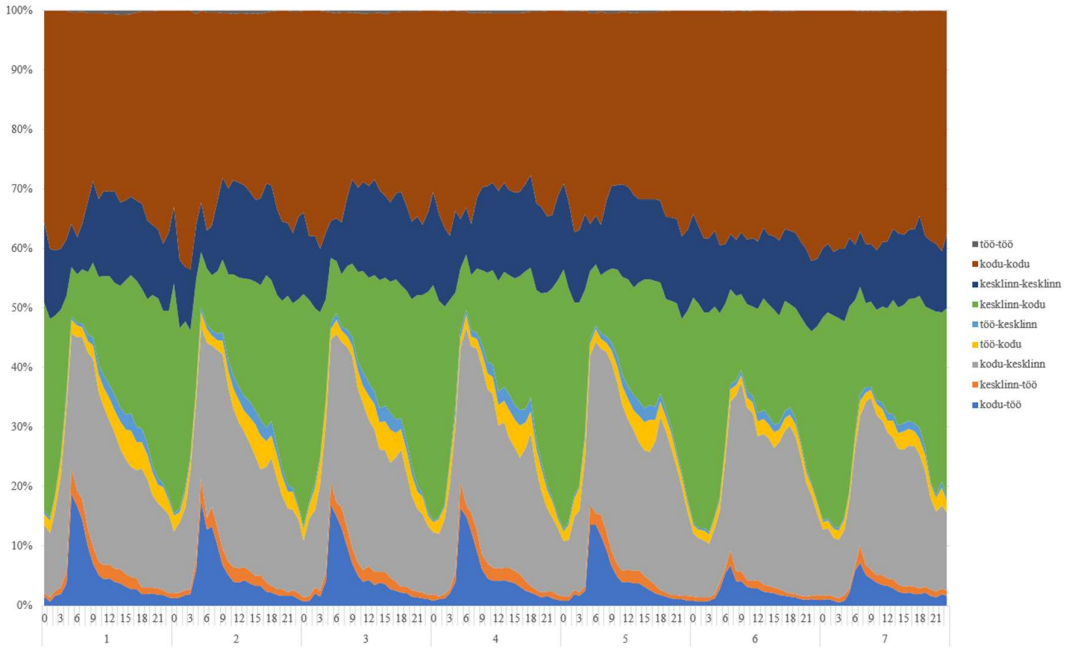
Joonis 7. Tallinna heksagonide K-keskmiste klasterdamismeetodi tulemus Mustamäel.

Intuitiivselt võib järeldada, mida erinevad klastrid tähendavad:

- punane (klaster 0, heksagonide arv 1013) – madalama tihedusega elupiirkonnad;
- sinine (1, 17) – suurima tihedusega äri ja meelelahutuse piirkonnad;
- lilla (2, 266) – suurema tihedusega elupiirkonnad;
- must (3, 231) – töökohad;
- oranž (4, 70) – elu, äri ja meelelahutuse tihedalt asustatud segu-piirkonnad.

Uurimistöö eesmärgiks on uurida erinevat liiki sessioone, mistõttu ei ole liialt oluline erineva asustustiheduse või sessioonide aktiivsusega elamise või muu tegevuse piirkonnad. Seetõttu on asjakohane grupeerida kokku lillad ja punased ning sinised ja kollased heksagonid. Selle tulemusel jääb alles kolm piirkonda: elamis-, töötamis- ja meelelahutuse piirkond. Nende piirkondade vahel liikumist saab omakorda juba sessiooni eesmärgiks kokku liita. Näiteks liikumine kodu ja töö vahel, või liikumine kesklinna ja kodu vahel. Neid liikumisi võibki pidada kasutusmustriteks.

Joonis 8 annab ülevaate nende mustrite suhtelisest kasutamisest nädala jooksul. On selgesti näha, kuidas tööpäeva hommikuti domineerivad liikumised tööle ja kesklinna, ning õhtuse tipptunni ajal liikumine vastassuunas. 30%-40% sessioonidest on määratud sama liiki piirkondade vahel: kesklinn-kesklinn ja kodu-kodu. Nädalavahetuse õhtusel ajal saaks seda seletada tuttavate külastamisega, kuid kuna nende liikumiste osakaal on kõrge igal ajal, võib see viidata ka klasterdamise ja heksagonide meetodi täpsuse piirangutele: see ei ole võimeline piisavalt täpselt tuvastama, mis kasutajat sessiooni lõpp-punktis ootab. Heksagon on piisavalt suur, et selles võib lisaks elamispinnale olla ka erinevaid töökohti ja meelelahutust.



Joonis 8. Klastrite kasutusmustrite nädalase kasutuse ülevaade.

3.4 Kasutusmustrite arvutamise implementeerimine

Ettevõte kasutab enda andmete töötlemise platvormina vabavaralist Airflow-d. See on skaleeritav sõltuvate töövoogude programmeerimise tööriist koos nende perioodilise käivitamise ja monitoorimisega. Ettevõttes kasutatakse seda kõige suuremal määral andmeaita andmete kogumisel, ehk sisuliselt on Airflow abil ülesse ehitatud ETL (*extract, transform, load*) protsess. Ka kasutusmustrite perioodiline arvutamine ning salvestamine on ETL protsess. Airflow võimaldab töövoo käivitamist ka mineviku andmete pealt, mis on ka antud ülesande puhul aruandluse ehitamiseks vajalik.

Andmete struktuur lõplikus andmebaasis peab olema paindlik muudatustele kasutusmustrite definitsioonis (vt jaotis „4.1 Ärilised nõuded“), kuna tulevikus võib tekkida soov nende definitsiooni muuta või defineerida uusi kasutusmustreid uute metodoloogiate alusel. Struktuur võiks lubada ka mitmete kasutusmustrite liikide salvestamist. Seetõttu on mõistlikum tulemused salvestada pigem „pikalt“ kui „laialt“. See tähendab seda, et iga kasutaja võimalikud ühe perioodi kasutusmustritest iga kasutusmustri kasutamine on eraldi real, mitte eraldi veerus. Sellisel juhul on võimalik väiksema ajakuluga mustreid juurde lisada, aga ka näiteks kasutusest ja arvutamisest eemaldada. Loodava tabeli etl_user_scenario struktuur on seetõttu alljärgnev:

```
CREATE TABLE etl_user_scenario (
	created timestamp NULL,
	user_id int4 NULL,
	city_id int2 NULL,
	period_start timestamp NULL,
	period_end timestamp NULL,
	period_length_days int2 NULL,
	type varchar(60) NULL,
	name varchar(60) NULL,
	session_with_destination_count int4 NULL,
	session_with_ride_count int4 NULL
);
```

Joonis 9. Programmikood kasutusmustrite tabeli loomiseks.

Loodavas tabelis kirjeldatakse iga kasutusmustri kohta ajaperioodi, millal see kasutaja kohta on arvutatud, mis tüüpi see on (nt kas ajatsoon, klasterdamine vms), mis on kasutusmustri nimi ning kasutaja aktiivsus selle mustriga antud perioodis. Aktiivsuse alusel on antud andmetest võimalik lihtsa vaevaga aruandeid ehitada.

4 Analüüs

Antud peatüki esimeses alapeatükis kirjeldatakse lähemalt ärinõudeid, millele kasutusmustrid ja nende arvutamine peavad vastama. Teises alapeatükis antakse kasutusmustrite kasutamisest üldisem ülevaade. Kolmandas alapeatükis analüüsitakse kasutusmustreid lähtuvalt nende sisulistest ärinõuetest. Neljandas alapeatükis hinnatakse mustrite arvutamise kulukust ning viimases alapeatükis pakutakse ideid edasiseks arenduseks.

4.1 Ärilised nõuded

Ärilised nõuded võib jaotada kaheks: nõuded kasutusmustritele ning nende arvutusele.

Eelnevas peatükis on juba mõningal määral kasutusmustrite eneste nõudeid puudutatud, kuid sellegipoolest tasub need ühes kohas kokku võtta. Esiteks, et kasutusmustrid oleksid ettevõtte erinevate valdkondade poolt kasutatavad, peavad need olema selgelt defineeritud ning arusaadavad. Ei ole mõeldav, et klasterdamise kasutusmustri tulemuseks oleksid abstraktsed numbrid vahemikus 1-5. Teiseks, kasutusmustrid peaksid olema ajas stabiilsed. See tähendab seda, et kasutusmuster peab kirjeldama täpselt sama kasutusmustrit poole aasta või ühe aasta pärast olenemata sellest, kuidas rakenduse üldine kasutamine on muutunud. See võimaldab jälgida kasutaja(te) kasutuse muutust ajas ning sellele adekvaatselt reageerida. Kolmandaks, kasutusmustrite arvutamine peab olema võimalik väga vähese järelevalvega. Kuna telefonirakendust saab kasutada erinevates linnades, pole mõeldav, et iga linna arvutuses tuleb näiteks igal nädalal klastrid ümber nimetada vastavalt sellele, kuidas algoritm parasjagu sessioone või heksagone grupeeris.

Kasutusmustrite arvutamisel tuleb arvestada järgmiste piirangutega. Esiteks peab arvutamine olema ettevõtte andmeplatvormil võimalik. Näiteks, kui tegemist on keerukama grupeerimise mudeliga, võiks sellest olla olemas implementatsioon Pythonis. Eelduslikult ei saa see nõue liiga suureks probleemiks. Teiseks, arvutus peab saama tehtud mõistliku ressursiga. Kuigi ettevõtte andmeplatvorm on skaleeritav, ehk keerukuse või andmemahu kasvades on võimalik arvutusvõimsust juurde lisada, on sellel oma hind ning ideaalsel juhul peaks seetõttu eelistama meetodit, mis suudaks ka kesise arvutusvõimsusega kõik linnad ja kasutajad kasutusmustritega paari tunni jooksul katta.

Tuleb arvesse võtta ka ajas suurenevat andmemahtu. Kolmandaks, nagu eelnevalt mainitud, meetod peab olema kasutatav perioodiliselt, tõenäoliselt iganädalaselt. See tähendab, et selle jooksutamine peab olema automaatne ning mitte nõudma inimressurssi.

4.2 Kasutusmustrite ülevaade

Eelmises peatükis kirjeldati kokku kolme erinevat viisi kasutusmustrite arvutamiseks: peamised heksagonid ja teekonnad, sessioonide ajatsoonid ning liikumine linna piirkondade vahel. Esimese grupi võib omakorda tinglikult jaotada kaheks, kuna peamise heksagoni sessioon võib samaaegselt olla ka peamise teekonna osa ja vastupidi. Samas ajatsoonid ja linna piirkondade vaheline liikumine jaotab kõik sessioonid eraldiseisvatesse mittekattuvatesse alamgruppidesse. Kõige lihtsam ülevaade kasutusmustritest ja nende liikidest on sessioonide jaotus nende lõikes. See ülevaade on toodud järgnevas Tabel 1.

Tabel 1. Kasutusmustrite osakaal sessioonidest.

Grupp	**Kasutusmuster**	**Osakaal sessioonidest, %**
peamine	peamine heksagon	54,86
peamine	peamine teekond	13,90
ajatsoon	tööpäev õhtu	18,25
	pidu	18,06
	tööpäev hommik	17,20
	nädalavahetus päev	16,87
	tööpäev päev	16,78
	öö või muu	12,84
liikumine piirkondade vahel	kodu-kodu	33,84
	kesklinn-kodu	22,24
	kodu-kesklinn	21,55
	kesklinn-kesklinn	12,98
	kodu-töö	3,45
	töö-kodu	2,80
	kesklinn-töö	1,52
	töö-kesklinn	1,36
	töö-töö	0,25

Eelnevast tabelist on näha, kuidas nii ajatsooni kui piirkondade vahelise liikumise kasutusmustrites on suurema ja väiksema aktiivsusega mustreid, kusjuures osad liikumise mustrid on lausa marginaalsed oma alla 2% osakaaludega. Samas näiteks peamine heksagon katab 55% kõikidest sessioonidest. Võib eeldada, et peamine heksagon on inimese kodu, niisiis üle poole sessioonidest toimub kodu ja mingi sihtkoha vahel ning 45% sessioonidest toimub muude heksagonide osalusel.

4.3 Grupeerimise meetodite võrdlus

Antud alapeatükis arutletakse kasutusmustrite sisulistest ärinõuetest lähtuvalt kolmel teemal: kasutusmustrite selge definitsioon, nende stabiilsus ajas ning perioodiline arvutamine ilma järelevalveta.

4.3.1 Selge definitsioon ja arusaadavus

Kõik kolm kasutusmustrite arvutamise meetodit kirjeldavad kasutajate käitumist erinevate külgede alt ning seetõttu peab ka definitsiooni ja arusaadavust nende puhul eraldi analüüsima.

Esiteks, ajatsoonidesse jaotamine. Kui võtta seda meetodit kui ainult ajalise mõõtme alusel jaotatud sessioonide kogumeid, siis võib lugeda seda täielikult selgelt defineerituks ja arusaadavaks. Probleeme hakkaks valmistama küll see, kui hakkaksime ajatsoonidele nimetusi panema. Autor on seda ise küll näiteks juba kategooria „pidu“ puhul teinud, kuid on selge, et kuigi ilmselt sellel ajal suur osa sõite tehakse just vaba aja veetmise eesmärgil, ei saa kõik nendel tundidel tehtud sõidud selle eesmärgiga seotud olla. Seetõttu oleks võinud autor loomulikult ajatsooni täieliku selguse eesmärgil ümber nimetada, näiteks nimetusega „reede ja laupäeva õhtud ja ööd“. Samas antud andmete puhul tundus see ebamõistlik ning vaadates Joonis 4 toodud sessioonide sagedust, paistab see periood siiski väga selgelt teistest silma ning autori arvates pole kahtlust, et ajatsoon ja sõidu eesmärk just üpris suure tõenäosusega seda kirjeldavad. Küll peaks selle nimetuse ja küllap ka ajatsoonide piirid erinevate ettevõtte teiste linnade puhul üle vaatama ja võimalik, et ehk ka piire vastavalt linna liikumise dünaamikale muutma. Näiteks võib olla võimalik, et tudengilinnas Tartus suureneb sessioonide arv juba neljapäeva õhtul. Mõnes teises linnas võib olla mõistlik rakendada teisi tööaegasid jne. Sellegipoolest on see kasutusmustrite

grupp kõige selgemalt defineeritud ja seetõttu ilmselt ka kõige kergemini praktikas ettevõtte toodete arendamisel kasutatav.

Peamiste teekondade ja heksagonide puhul sõltub kasutusmustri definitsioonide arusaadavus nende arvutamiseks kasutatud tingimustest. Kui neid mustreid praktikas kasutusele võtta, tuleks alati arvestada ka reegleid, mille alusel on need määratud. Sarnaselt ajatsoonidega võib olla vajalik need reeglid erinevates linnades üle vaadata: ehk on mingi reegel liiga karm või vastupidi, määrab peamised teekonnad ja heksagonid liiga kergekäeliselt.

Piirkondade vahel liikumise täpsuse piirangutest oli varasemalt ka juttu ning ilmselt saavad need piirangud peamisteks probleemideks definitsioonidega töötamisel. Samas on jälle piirkondade vahel liikumise kasutusmustrid intuitiivselt üpris arusaadavad, kuid sarnaselt peamiste heksagonide ja teekondadega on nende kasutamisel ilmselt alati vajalik ära tuua ka linna jaotus erinevate piirkondade vahel. Sellest saaks kasutusmustri kasutaja omakorda järeldada, et nende mustrite gruppide täpsus ei ole täielik ja et neid tuleks seetõttu ettevaatlikkusega kasutada.

Definitsioonide ja arusaadavusega tegelemisel tuleb arvestada ka andmestiku eripära: uurimistöös on kasutada ainult üks osa kasutaja liikumisvajadusest, kus rakenduse poolt pakutav lahendus kasutajale parasjagu sobilik tundub. On ilmselge, et rakendust ei kasutata kõikide liikumisvajaduste rahuldamiseks. Näiteks lühemad distantsid tehakse ilmselt jalutades, mõne sobilikuma teekonna puhul on ühistransport mõistlikum jne. Lisaks on muidugi kasutajate eneste eelistused erinevad. Mõni eelistab näiteks ühe kilomeetri läbimiseks see vahemaa jalutada, kuid teine eelistab selle autoga sõita, kolmandal pole võimalust autot kasutada, kuid ka jalutamine ei sobi, niisiis jääb valikusse ühistransport või takso. See kasutajate erinevate transpordiviiside valik ja eelistus lisab lisadimensiooni kasutusmustrite arendamisse. Näiteks võime mõne kasutaja puhul näha, et tema peamine teekond on liikumine reede õhtuti kodu ja kesklinna vahel. Sellest ei saa me aga järeldada, et antud kasutajal poleks liikumisvajadust ka näiteks tööpäeva hommikuti ja ettevõte ei peaks seetõttu antud kasutajat selle ajaperioodi potentsiaalsete kasutajate hulgast eemaldama. Ehk hoopis vastupidi: kasutamata kasutusmuster näitab potentsiaalset vajadust, mida rakenduse kasutamine pole veel suutnud katta.

4.3.2 Stabiilsus ajas

Valitud kasutusgruppide grupeerimise meetodid peavad enamuses tänu nende arvutamise meetodile juba tagama nende stabiilsuse ajas. Mõned eeldused iga meetodi kasutuse kohta siiski on.

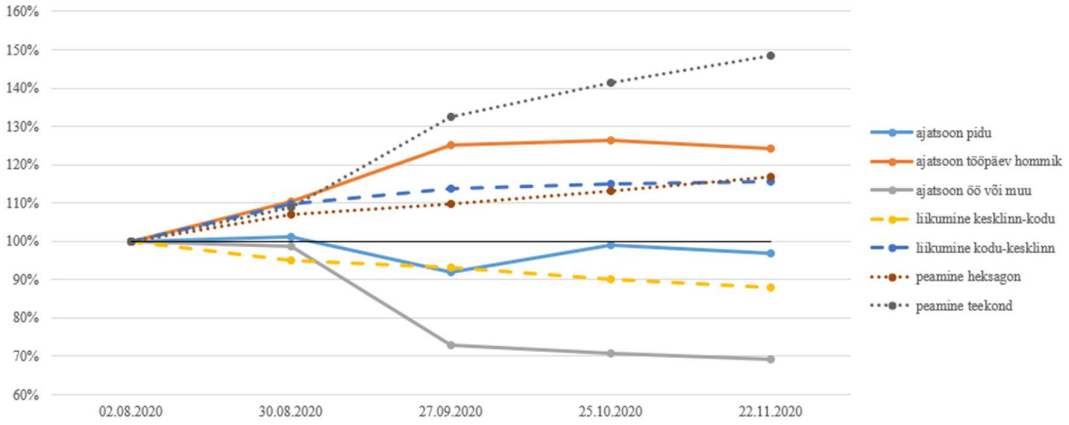
Ilmselt võivad kõige enam muutuda ajas peamiste heksagonide ja teekondade arvutused, kuna need sõltuvad kõige enam kasutaja konkreetse perioodi kasutuskäitumisest. Näiteks, kui üldiselt on kasutaja peamine teekond kodu ja töö vahel, siis puhkuse saabudes võib saada oluliseks mõni muu teekond, mille eesmärk kasutaja poolt vaadates on hoopis erinev kui eelnevalt.

Kuna ajatsoonid olid defineeritud ainult kellaaja ja nädalapäeva alusel, ei saa nende gruppide sisu ajas kuidagi varieeruda. Muidugi, kui üritada jällegi erinevatele ajatsoonidele nimetusi anda, võime tulemuseks saada ajas mõningal määral muutuvad tulemused. Kui võtta näiteks jällegi suvine puhkuste periood, siis võivad õhtutel tehtavad töölt koju tulemise sessioonid asenduda sessioonidega vaba aja veetmiseks.

Klasterdamise tulemusena saadud piirkonnad linnas sõltuvad väga palju sellest, mis alusandmed parasjagu kasutada on. Ühe näitena arvutas autor välja täpselt samade muutujatega klastrid, kuid kasutas septembri asemel sama aasta augusti andmeid. Tulemuses oli üpriski erinev. Näiteks oli näha, et enam ei olnud mudel suuteline eristama töökohtasid, kuid samas eristas see eraldi grupina linnalähedased elamispiirkonnad. Seetõttu, et saada ajas stabiilseid ja võrreldavaid piirkondi, tuleks erinevatel perioodidel kasutada eelnevalt defineeritud ja salvestatud piirkondade nimekirja. See muudab antud kasutusmustri arvutamise töövoo kiiremaks, kuid samas nõuab esialgsete piirkondade üle vaatamist ja võimalik, et ka kohatist korrigeerimist ettevõtte töötaja poolt. Järgnevas, kuluefektiivsuse arvutamise alapeatükis on sellegipoolest sisse arvestatud ka klastrite arvutamise aeg.

Kirjeldamaks kasutajate kasutusmustrite kasutamise muutust, on järgneval Joonis 10 ära toodud erinevate kasutusmustrite osakaal erinevatel perioodidel. Joonise valikusse on jäetud suurema mahu ja olulisema muutusega kasutusmustrid. Joonisel on kasutusmustrite kasutamine arvutatud joonisel toodud kuupäevale eelneva nelja nädala jooksul. Näiteks kuupäeva 02.08.2020 puhul on olnud vaatluse all kasutajate aktiivsus ajavahemikus 6.07.2020-02.08.2020. Kogu vaatlusperioodi pikkus on niisiis 6.07.2020-

22.11.2020. Et ilmestada suhtelist kasutusmustrite muutust ajas, on esimese perioodi väärtuseks määratud 100% ning järgnevate perioodide väärtused on suhtelised erinevused sellest.



Joonis 10. Valitud kasutusmustrite osakaalu suhteline muutus ajas. 02.08.2020 osakaal on 100%.

Selle aja sees on toimunud selge nihe kasutusmustrite kasutamises. Kõige selgemalt hakkab silma „ajatsoon öö või muu“ osakaalu suhteline vähenemine ja püsimine madalal tasemel alates septembrist. Mäletatavasti oli peamiselt tegemist öise ajaga tööpäevade sees. Ilmselt suvisel puhkuste ajal on võimalus inimestel enda igapäevast elurütmi hilisemaks nihutada. Seda väidet ilmestab ka suhteliselt madalam „ajatsoon tööpäev hommik“ osakaal suvisel ajal. Selle kasutusmustri kasutatavus aga koos septembri saabumisega kasvab. Sügise saabumisega kasvavad ka peamiste heksagonide ja teekondade osakaalud, kuid erinevalt „ajatsoon tööpäev hommik“ kasutusmustrist saavutavad need enda suhtelise maksimumi alles novembri lõpuks.

4.3.3 Arvutamise järelevalvevajadus

Antud ärinõude puhul oli oluline, et arvutus töötaks võimalikult väikese järelevalvega. Kõikide arvutuste puhul on mingi järelevalve vajalik, kuna nii linnad kui ühiskonnad on aeglases, kuid pidevas muutumises: inimeste töö- ja elukorraldus muutub vastavalt ühiskonna ja miks mitte, tehnoloogia arengule, linnasid arendatakse, ehitatakse juurde täiesti uusi elu- ja tööpiirkondasid jne. Nendel aeglaste, kuid järjepidevate muutustega peaks kõikide antud uurimistöös kirjeldatud, kuid ka võimalike tulevaste kasutusmustrite loomisel arvestama ja ilmselt perioodiliselt neid üle kontrollima. Järgnevalt vaadeldakse kasutusmustrite järelevalvevajadust pigem iganädalase perioodilise arvutamise vaatevinklist ning jäetakse tähelepanuta aeglaselt muutvad linnad.

Ajatsoonide arvutus, nagu varasemalt mainitud, on kõige lihtsam, ning kui ajatsoonid on linnale defineeritud, ei saa selle arvutamisega probleeme olla. Eeldus on, et linnale on määratud selle konkreetse linna kasutusmustritele sobivad ajatsoonid.

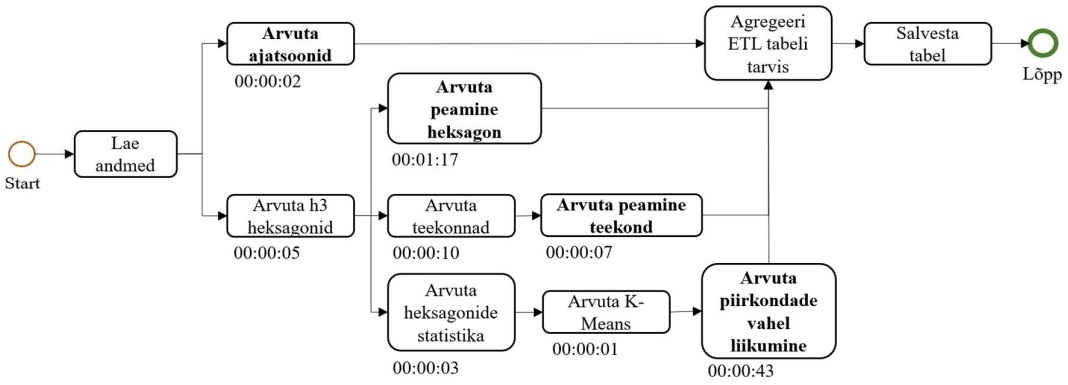
Peamiste heksagonide ja teekondade puhul peab kasutaja arvestama nende mustrite arvutusloogikaga: igal uuel arvutusel on võimalus saada samale kasutajale teistsugune peamine heksagon või teekond. Arvutuse järelevalvevajadust see omadus ei suurenda.

Piirkondade vahel liikumise puhul on suurim mõju mustritele linna piirkondadeks jaotamisel. Nagu selle kasutusmustri arvutuse juures juba kirjeldatud, ei ole mõistlik seda perioodiliselt üle arvutada: piirkondade definitsioon hakkaks sellisel juhul väga suurel määral perioodist perioodi varieeruma. Lahenduseks oleks sellisel juhul piirkondade määramine igale linnale üks kord ja nende piirkondade kasutamine igal järgneval piirkondade vahel liikumise arvutamisel. Sellega kaob ära vajadus perioodilisele järelevalvele, kuid samas muutub oluliseks, et määratud piirkonnad kajastaksid kasutajate liikumist nendes heksagonides võimalikult hästi. Eelduslikult on selle meetodi eeltöö iga uue linna lisamisel seetõttu kõige mahukam.

4.4 Kasutusmustrite arvutamise kuluefektiivsus

Käesolevas jaotises analüüsitakse kasutusmustrite arvutamise ajakulu arvutamist ning kirjeldatakse tulemusi uurimistöös kasutusel olevate andmete pealt.

Eelnevalt on välja arendatud neli erinevat kasutusmustrit. Kahel neist (ajatsoonid ja liikumine piirkondade vahel) on oma alamklassid, kaks on kasutajapõhised. Kõigi nende arvutamiseks on vaja teha erinevaid samme, kusjuures osad sammud on vajalikud mitme kasutusmustri arvutamiseks ja/või hilisemaks kombineerimiseks lõpliku tabeli tarvis. Järgneval Joonis 11 on ära toodud peamised erinevate mustrite arvutamise sammud alates andmete pärimisest kuni andmebaasi salvestamiseni koos sammu arvutuse ajakuluga. Rasvases kirjas on ära märgitud kasutusmustrite eneste arvutamine.



Joonis 11. Kasutusmustrite arvutamise töövoog ja ajakulu.

Nagu eelnevalt jooniselt Joonis 11 on näha, pole näiteks ajatsoonide arvutamiseks vaja erilisi lisaarvutusi üldse teha: sessiooni alguse aeg on algandmetes olemas ning selle alusel on väga lihtne sessioonid ajatsoonidesse määrata. Samas ülejäänud kasutusmustrite eelduseks on h3 heksagonide olemasolu. Peamise teekonna arvutamiseks on vaja lisaks arvutada kõik teekonnad ning kõige mahukam protsess on piirkondade vahel liikumise arvutamisega, mille tarvis on vaja arvutada heksagonide statistika ning seejärel selle alusel heksagonid klasterdada. Lõpliku ETL tabeli kuju saamiseks on vaja kõik kasutusmustrid välja arvatud ajatsoonid veel sessioonidega ühendada.

Niisiis, et hinnata ajakulu kasutusmustrite arvutamiseks tuleb nelja kasutusmustri kogukulu saamiseks kokku liita neile eelnevate ja järgnevate vajalike sammude ajakulu.

Kasutusmustrite töövoos sõltub andmete laadimise ja salvestamise sammud andmebaasi koormusest. Samas on need sammud ka sellised, mis on ühised kõigi kasutusmustrite jaoks. Seetõttu on mõistlik neid sammusid praeguses arvutuses, mille eesmärgiks on peamiselt kasutusmustrite kuluhinnangute võrdlemine, mitte arvestada. Lisaks pole kulu puhul arvestatud ka sammuga „Agregeeri ETL tabeli tarvis“, kuna selles sammus pole enam võimalik liiga lihtsalt eristada ühe kasutusmustri mõju. Lõplik ajakulu kasutusmustrite kohta eraldi ja kokku on toodud järgnevas nimekirjas:

- ajatsoonid – 2 (sekundit);
- peamine heksagon – 83;
- peamine teekond – 23;
- piirkondade vahel liikumine – 52;
- kõik kasutusmustrid korraga – 148.

Tulemustest on näha suhteliselt väga suured erinevused erinevate kasutusmustrite arvutamisel. Lühima ja pikima arvutuse vahe on 40-kordne. Autor annab, aru, et ilmselt on tema implementatsioonis palju ruumi arvutuse optimeerimiseks, kuid sellegipoolest demonstreerib arvutus suhtelist erinevust kulukuses, millega ettevõte saab arvestada ja suurema hulga andmetega läbi proovida.

Kokkuvõtteks, ühe linna arvutamise kulu ei ole märkimisväärne. Kuna kõik arvutused on lineaarse keerukusega, kuluks näiteks 100 korda suurema andmehulga arvutuse peale autori poolt kasutatud ressursil järjest arvutades kokku 4 ja pool tundi. Positiivse küljena saab arvutusi käivitada ka paralleelselt ning vajadusel arvutusvõimsust lisada. Lisaks on antud hetkel piirkondade vahel liikumise kuluarvutus tehtud koos klastrite arvutamisega (ainukene mitte-lineaarne komponent, kuigi samas linnade suurused ehk heksagonide arv ei kasva koos seal tehtavate sessioonide arvuga), kuid nagu mainitud, peavad klastrid olema pigem varem valmis arvutatud ja ettevõtte töötaja poolt üle kontrollitud. Niisiis ei ole vaja ega isegi mõistlik piirkondade klastreid perioodiliselt üle arvutada, ja see muudab ka piirkondade vahel liikumise arvutuse kiiremaks ning lineaarseks.

Lisaks suuremale andmemahule peab ettevõte veel testima arvutust ka koos andmete laadimise ja salvestamisega. Optimeerimise poole pealt võib proovida osade meetrikate arvutamist otse SQL-andmebaasis. Näiteks saab kindlasti ajatsoonide arvutuse andmebaasi üle viia, kuigi see oleks väike võit. Kui ettevõttel oleks võimalus h3 heksagone arvutada andmebaasis, oleks ilmselt võimalik ka peamiste heksagonide ja teekondade leidmine muuta selliselt, et andmeid andmebaasist väljaspoole ei liigutatagi. Lisaks on need arvutused andmebaasis ilmselt rohkem optimeeritud kui Pythonis tehes. Viimast väidet pole küll testitud.

4.5 Võimalikud grupeerimise edasiarendused

Antud töö eesmärgiks pole olnud parima meetodi välja valimine, vaid ainult kandidaatide arendamine ja analüüs. Seda tööd saab ettevõte kerge vaevaga jätkata, kuna võimalike uute kasutusmustrite lihtne lisamine süsteemi oli üks ärinõuetest.

Näiteks oli antud andmetega ja antud linnas optimaalne kasutada K-keskmiste klasterdamise meetodit linna piirkondade tuvastamisel. Klasterdamisel kasutati ühte valikut heksagonide põhistest tunnustest, millest suure osa moodustas seal tehtavate sessioonide arv ning info selle kohta, kunas need sessioonid on tehtud. Tänu sellele sai autori arvates tulemuseks klastrid, mis töö probleemi paremini lahendasid. Aga samas võiks kaasata ka teisi tunnuseid, nagu sessioonide pikkuse ajas või ruumis, ehk sisuliselt geograafilise komponendi, mis eristaks kaugemal asuvad kasutajad. Võib rohkem katsetada ka teiste klasterdamise meetoditega.

Klasterdamise tulemusel saadud piirkondade vahelisest liikumisest oli näha, et 30%-40% sessioonidest toimub teekonnal kodu-kodu ja eeldati, et järelikult pole praeguse lahenduse täpsus eri liiki sihtkohtade määramisel liiga suur. Seda oleks potentsiaalselt võimalik parandada, kui ettevõtte kasutaks täpset sihtkoha aadressi (võrdluseks praeguse umbes 350m läbimõõduga heksagonile) ning määraks selle aadressi kasutusotstarbe avalikke kaardiandmeid kasutades: kas seal asub elumaja, kontorid, tööstus, meelelahutus, transport vms. Varasemalt toodi küll välja, et ka sellisel juhul ei oleks lihtne tuvastada näiteks seda, kas kasutaja läheb kaubanduskeskusesse teenindajana tööle või sisseoste tegema. Võimalik, et selle saaks küll omakorda tuvastada konkreetse teekonna populaarsuse ja kellaaegade järgi: kui kasutaja ei kasuta seda sihtkohta perioodiliselt või on sinna ja tagasi sessiooni (kui on olemas tagasi või kuskile edasi viiv sessioon) vahe vähem kui tavaline tööpäeva või vahetuse pikkus siis on ilmselt tegu mitte tööga seotud teekonnaga. Kaubanduskeskuste vms teenindusasutuste puhul oleks võimalik avalikest andmetest hankida külastajatele lahti oleku ajad: kui asukohta või asukohast tehti sessioon väljaspool seal asuva peamise asutuse avatud tunde siis on ilmselt tegemist sessiooniga tööle minemiseks.

5 Kokkuvõte

Käesolevas töös uuriti erinevaid võimalusi ühe ettevõtte telefonirakenduse ajalis-ruumilise kasutusmustrite sarnasuse alusel grupeerimiseks ning analüüsiti, kuidas loodud mustrid rakenduse kasutust kirjeldavad ning kui mahukas on nende arvutamine ettevõtte andmeplatvormil.

Uurimistöö koostamiseks kasutati rakenduse kasutamise andmeid Tallinnas. Alustati andmete ülevaate ja esmase töötlemise ülevaatega. Edasi arendati kolm erinevat viisi mustrite arvutamiseks ning tuvastati kokku 17 erinevat mustrit. Esimene liik mustreid oli rakenduse kasutus jaotatud reeglite alusel ajatsoonidesse. Seeläbi on võimalik selgelt tuvastada rakenduse erineva kasutusaktiivsusega ajatsoonid. Teine meetod vaatles kasutajapõhist aktiivsust ja iga kasutaja kohta leiti tema peamine heksagon (alguse ja lõpu punktid olid eelnevalt jaotatud heksagonidesse) ning peamine teekond. Need kasutusmustrid väljendavad ühe heksagoni ja teekonna olulisust konkreetsele kasutajale, kuid ei ütle midagi nende sisu kohta. Kolmanda viisina grupeeriti linna heksagonid klasterdamise abil kolme erinevasse kategooriasse (ainult elamispiirkonnad, ainult tööpiirkonnad, kesklinna töö ja elupiirkonnad segamini) ja seejärel jaotati rakenduse kasutus liikumiseks nende eri liiki piirkondade vahel.

Analüüsides kasutusmustrite ärinõuetele vastavust ja kasutuse muutust ajas oli näha, et need kirjeldasid hästi kasutajate kasutuskäitumise muutust suve- ja sügiskuudel, mis annab ettevõttele kindluse soovi korral neid mustreid enda töös kasutama hakata. Iga kasutusmustrite arvutamise meetodi juures oli sellegipoolest välja toodud hulk kitsendusi, millega nende kasutamisel peaks arvestama. Lisaks pakuti ideid kasutusmustrite edasiseks arendamiseks.

Kasutatud kirjandus

[1] I. Mateo-Babiano, R. Bean, J. Corcoran, and D. Pojani, „How does our natural and built environment affect the use of bicycle sharing?“, *Transportation Research Part A: Policy and Practice*, vol. 94, pp. 295–307, Dec. 2016, doi: 10.1016/j.tra.2016.09.015.

[2] M. Z. Austwick, O. O'Brien, E. Strano, and M. Viana, „The Structure of Spatial Networks and Communities in Bicycle Sharing Systems“, *PLOS ONE*, vol. 8, no. 9, p. e74685, Jun. 2013, doi: 10.1371/journal.pone.0074685.

[3] K. Puntumapon and W. Pattara-atikom, „Classification of Cellular Phone Mobility using Naive Bayes Model“, in *VTC Spring 2008 - IEEE Vehicular Technology Conference*, May 2008, pp. 3021–3025, doi: 10.1109/VETECS.2008.324.

[4] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo, „Mining User Mobility Features for Next Place Prediction in Location-Based Services“, in *2012 IEEE 12th International Conference on Data Mining*, Dec. 2012, pp. 1038–1043, doi: 10.1109/ICDM.2012.113.

[5] E. Cho, S. A. Myers, and J. Leskovec, „Friendship and mobility: user movement in location-based social networks“, in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, Aug. 2011, pp. 1082–1090, doi: 10.1145/2020408.2020579.

[6] Q. Gao, F. Zhang, F. Yao, A. Li, L. Mei, and F. Zhou, „Adversarial Mobility Learning for Human Trajectory Classification“, *IEEE Access*, vol. 8, pp. 20563–20576, 2020, doi: 10.1109/ACCESS.2020.2968935.

[7] H. Senaratne *et al.*, „Urban mobility analysis with mobile network data: a visual analytics approach“, *IEEE Transactions on Intelligent Transportation Systems*, May 2018, doi: 10.1109/TITS.2017.2727281.

[8] M. Yan, S. Li, C. A. Chan, Y. Shen, and Y. Yu, „Mobility Prediction Using a Weighted Markov Model Based on Mobile User Classification“, *Sensors*, vol. 21, no. 5, Art. no. 5, Jan. 2021, doi: 10.3390/s21051740.

[9] E. Trevisani and A. Vitaletti, „Cell-ID location technique, limits and benefits: an experimental study“, in *Sixth IEEE Workshop on Mobile Computing Systems and Applications*, 2004, pp. 51–60, doi: 10.1109/MCSA.2004.9.

[10] Z. Sun, Y. Wang, H. Zhou, J. Jiao, and R. E. Overstreet, „Travel behaviours, user characteristics, and social-economic impacts of shared transportation: a comprehensive review“, *International Journal of Logistics Research and Applications*, vol. 24, no. 1, pp. 51–78, Jan. 2021, doi: 10.1080/13675567.2019.1663162.

[11] Q. Feng, „Gleaning Insights from Uber's Partner Activity Matrix with Genomic Biclustering and Machine Learning“, *Uber Engineering Blog*, Dec. 07, 2017. [Online]. Loetud aadressil: https://eng.uber.com/activity-matrix Kasutatud: 17.03.2021.

[12] R. Urtasun, „Deep Spectral Clustering Learning“, *Uber Engineering Blog*, Aug. 01, 2017. [Online]. Loetud aadressil: https://eng.uber.com/research/deep-spectral-clustering-learning Kasutatud: 17.03.2021.

[13] G. W. Milligan and M. C. Cooper, „Methodology Review: Clustering Methods“, *Applied Psychological Measurement*, vol. 11, no. 4, pp. 329–354, Dec. 1987, doi: 10.1177/014662168701100401.
[14] Amazon Redshift - Cloud Data Warehouse - Amazon Web Services, *Amazon Web Services, Inc.* [Online]. Loetud aadressil: https://aws.amazon.com/redshift Kasutatud: 19.04.2021.
[15] NumPy. [Online]. Loetud aadressil: https://numpy.org Kasutatud: 19.04.2021.
[16] pandas - Python Data Analysis Library. [Online]. Loetud aadressil: https://pandas.pydata.org Kasutatud: 19.04.2021.
[17] scikit-learn: machine learning in Python. [Online]. Loetud aadressil: https://scikit-learn.org/stable/ Kasutatud 19.04.2021 (accessed Apr. 19, 2021).
[18] Why hexagons?—ArcGIS Pro | Documentation. [Online]. Loetud aadressil: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/h-whyhexagons.htm Kasutatud: 03.04.2021.
[19] C. P. D. Birch, S. P. Oom, and J. A. Beecham, „Rectangular and hexagonal grids used for observation, experiment and simulation in ecology“, *Ecological Modelling*, vol. 206, no. 3, pp. 347–359, Aug. 2007, doi: 10.1016/j.ecolmodel.2007.03.041.
[20] uber/h3-py, Apr. 01, 2021. [Online]. Loetud aadressil: https://github.com/uber/h3-py Kasutatud: 03.04.2021.
[21] H3 Documentation. [Online]. Loetud aadressil: https://h3geo.org/docs Kasutatud: 03.04.2021.
[22] Apache Airflow Documentation - Airflow Documentation. [Online]. Loetud aadressil: https://airflow.apache.org/docs/apache-airflow/stable/index.html Kasutatud: 17.03.2021.

Lisa 1 – Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks[1]

Mina, Priit Tiganik

1. Annan Tallinna Tehnikaülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Telefonirakenduse ajalis-ruumilise kasutuskäitumise profileerimine“, mille juhendaja on Priit Järv
   1.1. reprodutseerimiseks lõputöö säilitamise ja elektroonse avaldamise eesmärgil, sh Tallinna Tehnikaülikooli raamatukogu digikogusse lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
   1.2. üldsusele kättesaadavaks tegemiseks Tallinna Tehnikaülikooli veebikeskkonna kaudu, sealhulgas Tallinna Tehnikaülikooli raamatukogu digikogu kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.
2. Olen teadlik, et käesoleva lihtlitsentsi punktis 1 nimetatud õigused jäävad alles ka autorile.
3. Kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest ning muudest õigusaktidest tulenevaid õigusi.

17.05.2021

1 Lihtlitsents ei kehti juurdepääsupiirangu kehtivuse ajal vastavalt üliõpilase taotlusele lõputööle juurdepääsupiirangu kehtestamiseks, mis on allkirjastatud teaduskonna dekaani poolt, välja arvatud ülikooli õigus lõputööd reprodutseerida üksnes säilitamise eesmärgil. Kui lõputöö on loonud kaks või enam isikut oma ühise loomingulise tegevusega ning lõputöö kaas- või ühisautor(id) ei ole andnud lõputööd kaitsvale üliõpilasele kindlaksmääratud tähtajaks nõusolekut lõputöö reprodutseerimiseks ja avalikustamiseks vastavalt lihtlitsentsi punktidele 1.1. ja 1.2, siis lihtlitsents nimetatud tähtaja jooksul ei kehti.