

**HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI**

Kandidaatintutkielma
Maantiede
Geoinformatiikka

**Matkapuhelinten käytöstä kerättävä mobiili big data ihmisten
liikkumisen tutkimisessa**

Pasi Okkonen

2021

Ohjaaja(t):

Petteri Muukkonen

**HELSINGIN YLIOPISTO
GEOTIETEIDEN JA MAANTIETEEN OSASTO
MAANTIETEEN KANDIOHJELMA**

**PL 64 (Gustaf Hällströmin katu 2)
00014 Helsingin yliopisto**

Tiedekunta/Osasto Fakultet/Sektion – Faculty		Laitos/Institution– Department	
Matemaattis-luonnontieteellinen		Geotieteiden ja maantieteen osasto	
Tekijä/Författare – Author			
Pasi Okkonen			
Työn nimi / Arbetets titel – Title			
Matkapuhelinten käytöstä kerättävä mobiili big data ihmisten liikkumisen tutkimuksessa			
Oppiaine /Läroämne – Subject			
Maantiede			
Työn laji/Arbetets art – Level		Aika/Datum – Month and year	Sivumäärä/ Sidoantal – Number of pages
LuK		12/2021	29
Tiivistelmä/Referat – Abstract			
<p>Tässä kandidaatintutkielmassa tarkastellaan matkapuhelimista kerättävän mobiilin big datan ominaispiirteitä ja saatavuutta ihmisten liikkumisen tutkimisen kontekstissa. Teemaa pohjustetaan aluksi keskeisten termien ja laatukriteerien määrittelyllä sekä tarkastelemalla liikkumisen tutkimisen historiaa ja mobiilin big datan käyttökohteita kyseisen tutkimusalan sisällä. Tutkielmassa eritellään toisistaan kaksi pääasiallista matkapuhelinten mobiilin big datan lähdeä: matkapuhelinverkosta sekä älypuhelinsovelluksista kerätyt sijaintitiedot, sekä tarkastellaan miten erilaiset maantieteellisen datan laatukriteerit täyttyvät näistä eri datalähteistä kerätystä aineistosta. Lopuksi tarkastellaan mobiilin big datan saatavuutta eri lähteistä ja yksityisyysdenuojakysymyksiä, sekä pohditaan minkälaiset ovat näihin datalähteisiin nojaavan tutkimuksen tulevaisuuden näkymät.</p> <p>Mobiili big data on uusi ja pinnalla oleva aineistolähde, jota tutkitaan ja käytetään hyvin monessa eri kontekstissa. Aineisto on kuitenkin jatkuvassa muutoksessa hyvin nopeassa tahdissa ja se on myös vaikeammin hallittavissa kuin perinteiseen aktiiviseen osallistamiseen perustuva tutkimusaineisto. Dataa kerätään ja jaetaan eteenpäin pitkälti yritysten omien intressien mukaan, ja datan laadun vaihdellessa paljon eri datalähteiden välillä, liittyy sen käyttöön paljon epävarmuuksia käytettävyyden ja saatavuuden suhteen. Tämä tutkielma antaa tiivistetyn kokonaiskuvan tähän aiheeseen, joka on noussut erityisen pinnalle viimeisen kahden vuoden aikana covid-19 pandemian aiheuttamien liikkumisrajoitusten myötä.</p>			
Avainsanat – Nyckelord – Keywords			
liikkuminen, liikkuvuus, matkapuhelin, matkapuhelinoperaattori, matkapuhelinverkko, mobiili big data, saatavuus, saavutettavuus, yksityisyysdenuoja, älypuhelinsovellus			
Säilytyspaikka – Förvaringställe – Where deposited			
Elektronisena maantieteen kandiohjelman toimesta			
Muita tietoja – Övriga uppgifter – Additional information			

Sisällys

1. Johdanto	1
2. Keskeiset käsitteet.....	1
2.1 Liikkuminen, liikkuvuus ja saavutettavuus	1
2.2 Big data.....	2
2.3 Datan laadun mittareita.....	4
3. Liikkumisen tutkimisen historiaa ja käyttökohteita.....	5
3.1 Historiaa	5
3.2 Käyttökohteita	7
4. Mobiilin liikkumisdatan kerääminen matkapuhelinten avulla	7
4.1 Matkapuhelinverkkoliikenne	8
4.2 Älypuhelinten mobiilipaikannus	9
5. Mobiilin big datan laatu eri lähteistä	11
5.1 Sijainti	11
5.2 Aika	12
5.3 Ominaisuustiedot.....	14
6. Mobiilin big datan saatavuus	14
6.1 Datan saatavuus ja jakamislustat	14
6.2 Yksityisyydensuoja	17
7. Yhteenveto	20
Lähteet.....	23

1. Johdanto

Ihmiset liikkuvat arjessaan ja ihmisten liikkumista tutkitaan monilla tieteenaloilla, erityisesti maantieteen saralla. Viimeaikainen covid-19 pandemia on vaikuttanut ihmisten liikkuvuuteen ennennäkemättömällä tavalla, kun liikkumista rajoittavat poikkeuslait, palveluiden sulkeminen sekä vapaaehtoinen sosiaalisten kontaktien rajoittaminen on rajoittanut huomattavasti sitä missä ja miten paljon ihmiset liikkuvat (Santamaria ym., 2020). Pandemian leviämisen ja liikkumisrajoitusten vaikutusten ennustaminen sekä tulkitseminen on herättänyt myös paljon keskustelua siitä, miten ihmisten liikkumista voidaan seurata ja mallintaa, sekä minkälaisia teknologisia ja yhteiskunnallisia kysymyksiä siihen liittyy (Poom ym., 2020). Erityisesti mobiilin big datan hyödyntämistä ihmisten liikkumisen mallintamiseen ja seurantaan on tutkittu paljon viime vuosikymmenen aikana, kun saatavilla oleva data on lisääntynyt GNSS-paikannusta hyödyntävien älypuhelimien yleistymisen myötä (Thums ym., 2018). Viimeisten kahden vuoden aikana covid-19-pandemia on entisestään lisännyt tällaisiin datalähteisiin nojaavaa tutkimusta (Chang ym., 2020; Santamaria ym., 2020).

Tässä tutkielmassa tarkastellaan minkälaista mobiiliksi big dataksi luokiteltavaa tietoa tällä hetkellä kerätään matkapuhelimista, mikä on sen saatavuus tutkimuskäyttöön sekä miten datan laatu vaihtelee eri lähteiden välillä ihmisten liikkumisen tutkimisen kontekstissa. Lisäksi tutkielmassa käsitellään minkälaisia käyttötarkoituksia tällä datalla on eri tutkimusaloilla, sekä minkälaisia mahdollisuuksia ja uhkakuvia datan keräämiseen ja käyttöön liittyy ihmisten yksityisyydensuojan osalta.

2. Keskeiset käsitteet

2.1 Liikkuminen, liikkuvuus ja saavutettavuus

Liikkuminen (*movement*) on luonteeltaan spatio-temporaalista toimintaa, missä yksittäinen kohde siirtyy sijainnista toiseen tietyssä aikamääreessä (Kaufmann ym., 2004). Tämä tarkasteltava liikkuva kohde voi olla niin fyysinen esine tai eliö, kuten

liikenneväline tai ihminen, tai abstrakti idea tai konsepti, jonka sijaintia voidaan havainnoida. Jotta tätä liikettä voidaan esittää sekä tilassa että ajassa, esitetään liikettä usein diskreetteinä ajanhetkinä havainnoitujen sijaintien avulla, jotka pystytään yhdistämään tiettyyn yksilöön.

Liikkuvuus (*mobility*) on liikkumiseen liittyvä termi, joka maantieteen kontekstissa viittaa usein kykyyn ja mahdollisuuksiin liikkua paikasta toiseen, eli liikkumisen toteutumisen potentiaaliin (Handy, 2002), joskin sillä voidaan viitata myös tietyn populaation tai alueen sisällä toteutuvaan liikkumiseen erotuksena yksilön liikkumisesta (Last, 2007). Tässä tutkielmassa pidättäydytään ensimmäisessä määritelmässä.

Liikkuvuuteen läheisesti liittyvä termi on **saavutettavuus** (*accessibility*), millä viitataan helppouteen päästä eri sijainteihin ja erilaisten palveluiden piiriin (Handy, 2002; Mayhew, 2015). Erotuksena liikkuvuuden liikkumisen mahdollisuuksiin ja helppouteen pidättäytyvästä määritelmästä, saavutettavuuteen voi vaikuttaa niin fyysiset tekijät kuten matkan pituus ja matka-aika paikasta toiseen, että sosiaaliset tekijät kuten yhteiskuntaluokka, ikä ja tulotaso.

2.2 Big data

Usein käytetty määritelmä kuvailee big dataa kolmen ominaisuuden kautta: 1) datamäärät ovat suuria, 2) datan saatavuus on jatkuvaa ja lähes reaaliaikaista ja 3) dataa saadaan useammasta lähteestä sekä järjestelmällisessä että epäjärjestelmällisessä muodossa (Sagiroglu & Sinanc, 2013). Madden (2012) tarkentaa tätä määritelmää omassa tulkinnassaan ja määrittelee big datan sen mukaan miten dataa pystytään hallinnoimaan ja analysoimaan olemassa olevien työkalujen keinoin. Big data on Maddenin (2012, s. 4) mukaan liian suurta, liian nopeaa ja liian vaikeasti käsiteltävää käytössä oleville datankäsittely- ja hallinnointityökaluille, kuten perinteisille tietokantaohjelmistoille. Hän viittaa erityisesti perinteisten relaatiotietokantojen kyvyttömyyteen käsitellä ja hallinnoida riittävän nopeasti erilaisia tietorakenteita, joista big data usein koostuu.

Big datalle ominaista on, että sitä saadaan useammasta eri lähteestä ja se voi koostua niin järjestelmällisestä kuin epäjärjestelmällisestä datasta (Xie ym., 2018; Yazti & Krishnaswamy, 2014). Näistä ensimmäinen on relaatiotietokannoille suoraan luettavissa olevaa taulukkomuotoista dataa, jossa eri tietotyypit on valmiiksi järjestelty omiin sarakkeisiin. Epästrukturoitu data koostuu vähemmän järjestellystä datasta, kuten ohjelmien ja sensorien lokitiedoista jotka on vaikeammin taulukoitavissa (Sagiroglu & Sinanc, 2013; Yazti & Krishnaswamy, 2014).

Chen (2016) taas erottaa toisistaan kaksi eri datatyyppiä datan keruumenetelmän mukaan. Ensinnäkin small dataksi luokitellaan aineisto, jossa tiedonkeruun kohde on suostunut luovuttamaan aineistoa juuri kyseisen tutkimuksen käyttöön esimerkiksi kantamalla ulkoista GNSS-paikanninta tai vastaamalla kyselyyn. Big dataksi taas luokitellaan aineisto, jossa tutkimuskohteita ei olla aktiivisesti värvätty tutkimukseen eikä dataa tarkoitettu tutkimuskäyttöön, mutta heidän passiivisesti tuotettua dataansa voidaan kuitenkin käyttää tutkimusaineistona. Tämänkaltaista dataa tuottavat esimerkiksi sijaintimerkityt sosiaalisen median päivitykset, maksutiedoista kerätyt sijaintitiedot sekä erinäiset GNSS-paikannusta käyttävät älypuhelinsovellukset.

Mobiili big data on yksi big datan osa-alue, jolla viitataan erilaisista mobiililaitteista kerättyyn dataan ja johon kuuluu olennaisesti laitteen sijaintitiedot. Merkittävien mobiilin big datan lähde on matkapuhelimet, joiden sijainteja kerätään matkapuhelinverkkoliikenteestä sekä älypuhelinsovelluksista, jotka molemmat hyödyntävät puhelimen lähettämiä ja vastaanottamia signaaleja sijainnin määrittämiseksi (Yazti & Krishnaswamy, 2014). Muita mobiilin big datan lähteitä ovat esimerkiksi esineiden internet, jolla viitataan yleisesti erilaisiin automaattista tiedonsiirtoa tukeviin laitteisiin, kulkuneuvoihin kiinnitetyt paikannus- ja navigointilaitteet sekä erilaiset kiinteät infrastruktuuriin upotetut sensorit (Xie ym., 2018; Yazti & Krishnaswamy, 2014). Sijaintitiedon lisäksi mobiiliin big dataan voi kuulua erilaisia ominaisuustietoja näiden mobiililaitteiden käytöstä ja niiden käyttäjistä, joita on voitu kerätä esimerkiksi älypuhelinsovelluksista (Xie ym., 2018).

Tämän tutkielman puitteissa keskitytään matkapuhelimista kerättyyn ja sijaintiin sidottuun mobiiliin big dataan. Kaikkia matkapuhelinten sijaintitietoja ja sijaintiin liitettäviä ominaisuustietoja käsitellään mobiilina big datana, sillä jaottelu aktiivisesti

ja passiivisesti kerättyyn dataan ei ole mielestäni mielekästä. Lähes kaikki matkapuhelimista kerätty sijaintitieto vaatii ainakin periaatteessa kertaluonteisen suostumuksen käyttäjältä, vaikka tämän toteutumisessa onkin havaittu ongelmia ja sen valvonta on hankalaa (Calacci ym., 2019; K. E. Martin, 2015).

2.3 Datan laadun mittareita

Mobiilin big datan laadun mittaamiseen voidaan käyttää montaa erilaista mittaria, sillä muun maantieteellisen tiedon tavoin se koostuu useasta mitattavasta ominaisuudesta. Sinton (1978) määrittelee maantieteelliseen dataan kuuluvaksi kolme eri mitattavaa elementtiä: **ominaisuus**, **sijainti** ja **aika**. Ominaisuus viittaa tässä yhteydessä sijainti- ja aikatietoihin liitettyihin kontekstuaalisiin tietoihin, josta käytetään myös termejä tieto tai muuttuja. Vereginin (1999) mukaan kaikkia näitä kolmea elementtiä voidaan arvioida ainakin neljän eri mittarin avulla: tarkkuus (*accuracy*), erottelukyky (*precision*), yhtenäisyys (*consistency*) ja kattavuus (*completeness*).

Datan **tarkkuus** viittaa siihen, kuinka hyvin datan ilmoittamat arvot vastaavat todellisia arvoja määritellyllä aggregointitasolla (*aggregation level*), eli toisin sanoen kuinka suuri on niiden virheaste (Kuhn, 2012; Veregin, 1999, ss. 178–181). Esimerkiksi sijainnin osalta tämä tarkoittaa sitä, kuinka paljon ja kuinka suuri osa ilmoitetuista mittaustuloksista eroavat todellisista sijainneista (Chrisman, 1991). Aikaulottuvuudessa tarkkuuden voidaan katsoa kertovan kuinka lähelle datan ilmoittamat mittaussijat vastaavat todellisia mittausten ajanhetkiä, ja ominaisuustiedoissa tarkkuus viittaa kategorisen datan osalta tietojen paikkansa pitävyyteen ja numeerisen datan osalta saatujen arvojen virheasteeseen (Veregin, 1999, s. 181).

Datan **erottelukyky**, josta käytetään usein termiä resoluutio, viittaa siihen kuinka pienillä mittayksiköillä sijainteja, ajanhetkiä tai ominaisuustietoja voidaan mitata ja asettaa näin tarkkuudelle alarajan (Chrisman, 1991, s. 166). Sijainnin osalta tämä tarkoittaa esimerkiksi onko sijaintitieto esitettävissä metrien vai kilometrien kokoisina yksiköinä ja aikaulottuvuutta tarkastellessa kuinka usein sijaintitietoja on saatavilla aikayksikköä kohden (Veregin, 1999, ss. 181–182). Ominaisuustiedoissa

erottelukyky kertoo usein mitattavan ilmiön, esimerkiksi lämpötilan tai asukasluvun, merkitsevien lukujen ja desimaalien määrän.

Datan **yhtenäisyys** viittaa datan ristiriitojen puuttumiseen ja on näin datan sisäisen eheyden mittari (Veregin & Hunter, 2000, s. 7). Tämä voi tarkoittaa ristiriitoja sijainnissa ja ajassa, kuten esimerkiksi tilanne, jossa ihminen on havaittu useassa eri sijainnissa samana ajanhetkenä (Veregin, 1999, s. 182). Ominaisuustiedoissa vastaavanlaisia ristiriitoja voi tapahtua, kun yksi ominaisuustieto rajaa toisen ominaisuustiedon mahdollisia arvoja, mutta kyseiset arvot eivät kuitenkaan vastaa näitä asetettuja rajoitteita (Veregin & Hunter, 2000, s. 7). Esimerkiksi suomalaisia kaupunkeja listaavassa aineistossa muiden maiden kaupungit aiheuttaisivat dataan ristiriidan.

Datan **kattavuus** kertoo kuinka suuren osan käytössä oleva data kattaa todellisesta maailmasta tai sen rajatusta alueesta, aikaikkunasta ja attribuuteista (Veregin, 1999, s. 183). Sijaintitiedon osalta tämä tarkoittaa esimerkiksi sitä jos määrittelemme tarkastelualueeksi koko maapallon, datalla, joka on kerätty ainoastaan Suomesta, on heikko tilallinen kattavuus sillä Suomi kattaa maapallon pinta-alasta vain pienen osan. Toisaalta jos tarkastelualueeksi otetaan Suomi, on sen tilallinen kattavuus täydellinen. Samanlaiset kriteerit pätevät myös aikaulottuvuudessa ja ominaisuustiedoissa; data on sitä kattavampi, mitä paremmin se kattaa määritellyn aikaikkunan ja kuinka suuri osa tarkasteltavista ominaisuustiedoista löytyy kerätystä datasta (Joksić & Bajat, 2004, s. 80).

3. Liikkumisen tutkimisen historiaa ja käyttökohteita

3.1 Historiaa

Paikannusteknologian käyttö ihmisten liikkumisen tutkimiseen ja yksityiskohtaisen liikkumisdatan keräämiseen on melko uutta verrattuna siihen, kuinka pitkään eläinten liikkumista on seurattu rengastuksen ja satelliittipaikannuksen avulla (Thums ym., 2018, ss. 1–4). Ennen liikkumisdatan saatavuuden yleistymistä teknologisista välineistä kuten matkapuhelimista, älykelloista ja maksukorteista, ihmisten liikkumista on tutkittu erilaisilla aktiivisen osallistamisen metodeilla, esimerkiksi matkapäiväkirjojen avulla (Kostyniuk & Kitamura, 1982). Passiivisesti tuotettua

dataa ihmisten liikkuvuudesta on saatu muun muassa seuraamalla merkittyjen setelien kiertoa (Brockmann ym., 2006) sekä tarkastelemalla sukunimien alueellisen jakautumisen vaihtelua suhteessa aikaan (Piazza ym., 1987).

Matkapuhelinten sijaintitietoja on kerätty aivan matkapuhelinverkon alkuajoista lähtien, sillä puheluiden ja muiden signaalien yhdistämiseksi verkon ylläpitäjien on tiedettävä mitkä linkkimastot pystyvät kommunikoimaan kunkin puhelimen kanssa tietyllä ajanhetkellä (Bar-Noy & Kessler, 1993). Ensimmäisiä tunnettuja matkapuhelinten sijaintitietoja hyödyntäneitä tutkimuksia on kuitenkin toteutettu niinkin myöhään kuin vuonna 2005, kun sadan yliopisto-opiskelijan liikkumista tarkasteltiin seuraamalla mihin Bluetooth-tukiasemiin puhelimet muodostivat yhteyksiä sekä mihin linkkimastoihin heidän soittamansa ja vastaanottamansa puhelut yhdistettiin (Eagle & Pentland, 2005).

Maailmanlaajuinen satelliittipaikannusjärjestelmä (*Global Navigation Satellite System, GNSS*) tuli ensimmäistä kertaa siviilikäyttöön vuonna 1995, jota ennen se oli ollut lähinnä sotilaskäytössä. Tämän jälkeen satelliittipaikannuksen mahdollistavaa teknologiaa on upotettu moniin eri laitteisiin ja on yleisesti käytössä lähes kaikessa sijaintitietoa tarvitsevilla tehtävillä kuten lentokoneiden ja muiden liikennevälineiden navigoinnissa ja seurannassa sekä infrastruktuurin hallinnassa (Lechner & Baumann, 2000). Ensimmäinen GNSS-paikannusta hyödyntävä matkapuhelin Benefon Esc! saapui markkinoille vuonna 1999 (Robustelli ym., 2019) ja GNSS-teknologian tarkkuuden parantuessa sekä vastaanottimien pienentyessä yhä useampi uusi puhelinmalli alkoi sisältämään GNSS-signaaleja vastaanottavan sirun, joka on nykyään pääasiallinen paikannusmenetelmä älypuhelimissa (Merry & Bettinger, 2019; Robustelli ym., 2019).

Mobiilipaikannuksen ja älypuhelinien yleistyessä laajeni myös matkapuhelinten käyttö pelkästä kommunikaatiovälineestä hyvin monenlaisiin erilaisiin käyttötarkoituksiin erilaisten puhelinsovellusten myötä. Nykyaikana älypuhelinien käyttötarkoitukset ovat lähes rajattomat ja puhelinsovelluksiin syötetyt ja niiden käytöstä kerätyt tiedot voidaan usein liittää puhelinien sijaintitietoihin ollen näin yleinen mobiilin big datan ominaisuustietojen lähde (Xie ym., 2018).

3.2 Käyttökohteita

Tutkimuskäytössä yleisimpiä käyttökohteita mobiilille big datalle ovat erilaisten liikkuvuus- ja saavutettavuusmallien laatiminen datan perusteella ja valmiiden mallien testaaminen tätä empiiristä dataa vasten (Smolak ym., 2020). Liikkuvuus- ja saavutettavuusmalleja voidaan käyttää esimerkiksi tautien leviämisen ennustamiseen (Milusheva, 2020), sosiaalisten verkostojen ja segregaaation analyysiin (Ganjour ym., 2020; Silm ym., 2021), liikennesuunnitteluun (Ghahramani ym., 2020; Järv ym., 2012) sekä mainonnan kohdentamiseen (Ghose ym., 2019).

Ihmisten liikkumista voidaan tutkia sekä sijainti-, että yksilökohtaisesti.

Sijaintikohtaisessa tutkimuksessa tutkitaan staattisen sijainnin läpi kulkevia liikennemääriä, ja yksilökohtaisessa tutkimuksessa ihmisten liikkumista monien eri sijaintien välillä. Sijaintikohtaisessa saavutettavuustutkimuksessa voidaan tutkia esimerkiksi kuinka nopeasti tietystä sijainnista pääsee toiseen sijaintiin, ja yksilökohtaisessa tutkimuksessa kuinka nopeasti eri ihmiset pääsevät tähän sijaintiin, johon vaikuttaa olennaisesti missä ihmiset sijaitsevat eri ajanhetkinä (Järv ym., 2018). Tutkimuksissa on käytetty sekä yksilökohtaista että sijaintikohtaista dataa, joskin yksilökohtaisen datan ollessa henkilökohtaisempaa ja potentiaalisesti yksityisyydensuojalainsäädäntöä rikkovaa, on data usein aggregoitu sijaintitasolle (Xu ym., 2017).

Yksilötasolla pystytään ihmisten liikkumisesta ja siihen vaikuttavista tekijöistä saada huomattavan paljon monipuolisempaa ja yksityiskohtaisempaa tietoa verrattuna sijaintikohtaiseen dataan. Yksilötasolla pystytään muodostamaan esimerkiksi käyttäytymisprofiileja ja tutkimaan vuorovaikutussuhteita ihmisten välillä tai heidän kohtaamiaan terveysriskejä (Finazzi & Paci, 2019; Smolak ym., 2020). Sijaintitasolla voidaan tutkia esimerkiksi ihmisten suosimia alueita turistikohdeissa (Hasnat & Hasan, 2018).

4. Mobiilin liikkumisdatan kerääminen matkapuhelinten avulla

Matkapuhelimien käytöstä kerättävää mobiilia big dataa saadaan pääasiallisesti kahdesta eri lähteestä. Ensinnäkin matkapuhelinoperaattorit pystyvät keräämään matkapuhelimien sijainteja hyödyntämällä linkkimastojen ja matkapuhelinten välistä

puhelinverkkoliikennettä. Toiseksi älypuhelimista sijaintitietoa saadaan hyödyntämällä useita puhelimen vastaanottamia ja lähettämiä signaaleja kuten GPS-, Wi-Fi- ja puhelinverkkosignaaleja, joita mobiilisovellukset lukevat ja lähettävät edelleen sitä hallinnoiville yrityksille.

4.1 Matkapuhelinverkkoliikenne

Matkapuhelinverkkoliikenteestä kerätyn datan voi karkeasti jakaa kahteen eri tyyppiin sen mukaan miten ja milloin datan keräys tapahtuu: **tapahtumapohjaisiin** (*event-driven*) ja **matkapuhelinverkkopohjaisiin** (*network-driven*) datankeruumenetelmiin (Pinelli ym., 2015). Pääasiallinen tapahtumapohjainen datalähde ovat puhelurekisteritiedot (*Call Detail Records, CDR*), joihin kerätään tietoa vastaanotetuista ja lähetetyistä puheluista sekä tekstiviesteistä, ja siihen voi kuulua esimerkiksi tieto mistä liittymästä ja milloin puhelu tai tekstiviesti on vastaanotettu ja lähetetty, kuinka pitkään puhelu kesti sekä minkä linkkimaston tai linkkimastojen kautta puhelu tai tekstiviesti yhdistettiin (Horak, 2007). Puhelurekisteritiedot on alun perin tarkoitettu matkapuhelinoperaattoreiden laskutuksen ja tietoliikennekapasiteetin hallintaan, mutta niiden käyttö myös tutkimuskäyttöön on lisääntynyt viime vuosikymmenten aikana (Zhao ym., 2016). Toinen mahdollinen tapahtumapohjainen datalähde on internettiin yhdistettäessä kerätyt IP-tietorekisterit (*IP Detail Record, IPDR*), joihin kerätään puhelurekisteritietojen tavoin erilaista tietoa laitteen aktiivisesta internetin käytöstä, mukaan lukien linkkimastot, joiden kautta verkkoyhteys muodostettiin (Pinelli ym., 2015; S. Zhang ym., 2020).

Liikkumisen seurantaan puhelu- ja IP-tietorekisterejä voidaan käyttää seuraamalla eri linkkimastoihin yhdistettyjä puheluita, tekstiviestejä ja verkkopyyntöjä, sillä kullakin linkkimastolla on oma kattavuusalueensa minkä sisällä sijaitsevien puhelimien signaaleja kukin masto vastaanottaa ja lähettää eteenpäin. Yksilötason liikkumista voidaan näin seurata tarkastelemalla mihin linkkimastoihin yksittäisen liittymän puhelut on yhdistetty eri ajanhetkinä, ja sijaintitasolla tarkastelemalla miten eri linkkimastojen vastaanottama tietoliikennemäärä ja palvelupyyntöjen määrä on vaihdellut ajassa (S. Zhang ym., 2020).

Matkapuhelinverkkopohjaiset datankeruumenetelmät keräävät tapahtumapohjaisista menetelmistä poiketen sijaintitietoja passiivisesti eikä käyttäjältä vaadita aktiivista puhelimen käyttöä datan keräämiseksi (Ma ym., 2013). Dataa kerätään matkapuhelinverkkopohjaisilla menetelmillä esimerkiksi säännöllisin aikavälein tehtävillä sijaintipäivityksillä, jotka antavat tiedon mihin linkkimastoon matkapuhelin on kulloinkin yhdistettynä, sekä puhelimen vaihtaessa sijaintiaan linkkimastojen palvelualueelta toiselle (S. Zhang ym., 2020).

Matkapuhelinsignaaleihin perustuvassa paikannuksessa pystytään keräämään myös linkkimastojen kantoalueita tarkempaa sijaintitietoa käyttämällä erinäisiä signaalien voimakkuuksia, lähtö- ja saapumisaikoja ja tulokulmia mittaavia menetelmiä (Ghahramani ym., 2020). Yksi paikannustarkkuutta parantava menetelmä on esimerkiksi laskea puhelinsignaalin saapumisaikoja ja voimakkuuksia useassa eri linkkimastossa ja näiden signaalien saapumisaikojen eroja vertaamalla saadaan puhelimen sijainti määritettyä kolmiomittaukseen perustuvalla laskennalla (J. Martin ym., 2009; Shafeeque ym., 2019). Pääasiallisesti matkapuhelinverkosta kerättyjen sijaintitietojen tarkkuus riippuu siitä kuinka paljon linkkimastoja sijaitsee matkapuhelimen kantoalueella sekä minkälainen on matkapuhelinverkon infrastruktuuri ja käytetyt laskentamenetelmät (Ma ym., 2013).

4.2 Älypuhelin mobiilipaikannus

Älypuhelimet voidaan pitkälti jakaa kahteen eri tyyppiin niiden järjestelmäarkkitehtuurin perusteella, sillä näitä kahta järjestelmäarkkitehtuuria käyttävien älypuhelimien yhteenlaskettu markkinaosuus vastaa lähestulkoon koko älypuhelinmarkkinoita. Nämä järjestelmäarkkitehtuurit ovat Android-, sekä iOS-pohjaiset käyttöjärjestelmät (IDC, ei pvm.). Molemmilla järjestelmillä on käytössä omat ohjelmointirajapintansa, jotka mahdollistavat sovellusten kehittämisen käyttöjärjestelmän ympärille ja pääsyn puhelimen vastaanottamiin signaalilähteisiin, joiden avulla myös puhelimen sijainti pystytään määrittämään (Bu-Pasha ym., 2017; Moayeri ym., 2019). Isona erona matkapuhelinverkkojen paikantamismenetelmiin, älypuhelin mobiilipaikannuksessa puhelimen sijainnin laskenta ja tallentaminen tapahtuu itse mobiililaitteessa, joten puhelin tarvitsee vaadittavan teknologian eri

signaalien vastaanottamiseen ja laskentaan, eikä näin ole pääasiallisesti mahdollista vanhemmissa puhelinmalleissa (Shafeeque ym., 2019).

Näiden eri järjestelmäarkkitehtuurien paikantamismenetelmät eroavat hieman toisistaan, mutta pääasiallisesti on tiedossa miten ja mitä signaalilähteitä nämä rajapinnat keräävät ja käyttävät sijainnin määrittämiseen. Tähän dataan kuuluvat muun muassa puhelinverkkosignaalit, GNSS- ja Wi-Fi-signaalit sekä Bluetooth-signaalit (Merry & Bettinger, 2019; Moayeri ym., 2019).

Samalla periaatteella kuin matkapuhelinverkkosignaalien kolmiomittaukseen perustuva paikannus, GNSS-paikannus perustuu GNSS-vastaanottimen mittaamiin signaalien saapumisaikoihin useasta eri GNSS-satelliitista. Kun satelliittien sijainti suhteessa maapalloon tunnetaan, voidaan näiden signaalien lasketuista saapumisajoista laskea vastaanottimen sijainti maan pinnalla (Lechner & Baumann, 2000). Sijainnin tarkkuuteen vaikuttaa myös matkapuhelinverkon linkkimastojen läheisyys, joiden lähettämien signaalien avulla puhelimet pystyvät tarkentamaan sijaintiaan käyttämällä avustetuksi GNSS:ksi kutsuttua paikannusmenetelmää (Bu-Pasha ym., 2017; Ruktanonchai ym., 2018; Shafeeque ym., 2019).

GNSS- ja matkapuhelinverkkosignaalit ovat pääasialliset älypuhelimien sijaintitietojen tarkkuutta määrittävät tekijät ulkotiloissa, mutta sisätiloissa muiden datalähteiden, kuten Wi-Fi- ja Bluetooth-signaalien, merkitys voi olla suurempi (Moayeri ym., 2019). GNSS-signaali ei useimmiten pysty läpäisemään rakennusten ulkorakenteita kovin tehokkaasti, ja GNSS-signaalin tarkkuus voi heikentyä myös esimerkiksi ympäröivien korkeiden rakennusten vuoksi ulkotiloissakin (Robustelli ym., 2019). Jotta puhelimen absoluuttista sijaintia pystyttäisiin mittaamaan lyhyen kantomatkan omaavien Wi-Fi ja Bluetooth signaalien avulla, täytyy tämän staattisen signaalilähteen sijainti olla tunnettu sijainnin laskemiseksi (C. Yang & Shao, 2015). Lisäksi näitä signaalilähteitä käytetään suhteellisessa sijainnin määrittämisessä, minkä avulla saadaan tieto kahden signaalilähteen välisestä etäisyydestä. Tätä periaatetta hyödyntää esimerkiksi Suomessa käytössä oleva Koronavilkku-sovellus (Weiß ym., 2021). Suhteellista paikantamista voidaan käyttää myös puhelimen paikantamiseen rajatulla alueella Wi-Fi- ja Bluetooth -signaaleja vastaanottavien staattisten tukipisteiden avulla, ja esimerkiksi ostoskeskukset pystyvät keräämään

tietoa ihmisten liikkumisesta kauppakeskuksen sisällä ja tarkkailla missä kaupoissa ihmiset käyvät ja viettävät aikaa (Oosterlinck ym., 2017).

5. Mobiilin big datan laatu eri lähteistä

Älypuhelinsovelluksista sekä matkapuhelinverkkoliikenteestä kerätyn datan laadussa on huomattavia eroja eri laatukriteereissä niin tilassa, ajassa kuin ominaisuustiedoissa riippuen esimerkiksi maantieteellisestä sijainnista ja käytetystä teknologiasta (Ghahramani ym., 2020; Ruktanonchai ym., 2018). Datan monipuolisuuden vuoksi seuraavat katsaukset datan laatuun ovat vahvasti yleistettyjä ja keskittyvät lähinnä potentiaaliseen datan laatuun eri lähteistä. Datan saatavuuteen ja yksityisyydensuojaan keskittyvät kappaleet tarkentavat minkälaista dataa on usein saatavilla tutkimuskäyttöön.

5.1 Sijainti

Spatiaalinen tarkkuus ja resoluutio vaihtelevat puhelinverkkoliikenteestä kerätyn sijaintitiedon osalta riippuen kunkin alueen linkkimastoverkon tiheydestä, matkapuhelinverkon infrastruktuurista ja käytetystä paikantamismenetelmästä (Ma ym., 2013; Zhao ym., 2016). Huonoimmillaan tarkkuus on kunkin linkkimaston palvelualue, mikä on usein puhelurekisteritiedoista kerätyn datan resoluutio, mutta erilaisia matkapuhelinverkkopohjaisia menetelmiä hyödyntämällä sijainnin tarkkuus voi olla huomattavasti tarkempi (Ghahramani ym., 2020; Shafeeque ym., 2019). Esimerkiksi Telian tarjoama Crowd Insights palvelu lupaa noin 200 metrin tarkkuuden linkkimastoverkon tiheydestä riippuen (Telia Company & Grön, ei pvm.) ja Man (2013) teettämässä tapahtumapohjaista matkapuhelinverkkodataa hyödyntäneessä tutkimuksessa keskimääräinen spatiaalinen tarkkuus oli 320 metriä.

Älypuhelinsovelluksista kerättyjen sijaintitietojen tarkkuuteen vaikuttaa muun muassa Wi-Fi-yhteyden saatavuus, signaalia heijastavien rakennusten läheisyys sekä käytetty puhelinmalli (Merry & Bettinger, 2019). Yleisesti useita signaalilähteitä hyödyntävistä älypuhelimista pystytään keräämään spatiaalisesti parempi tarkkuuksista dataa verrattuna puhelurekisteritietoihin tai muihin vain matkapuhelinverkkosignaaleja hyödyntäviin menetelmiin (Ghahramani ym., 2020). Monet käytössä olevat GPS-vastaanottimet ilmoittavat niiden tarkkuuden olevan

noin 3-15 metriä mikäli vastaanottimella on yhteys riittävän moneen satelliittiin (Shafeeque ym., 2019), mutta uusinta satelliittipaikannusteknologiaa hyödyntävissä älypuhelimissa päästään jo 1-2 metrin paikannustarkkuuksiin (Robustelli ym., 2019).

Tilallisen kattavuuden osalta sekä puhelinverkosta että älypuhelinsovelluksista pystytään dataa keräämään käytännössä kaikkialta maailmasta. Matkapuhelinyhteys oli vuoden 2019 alussa käytössä 67 prosentilla maailman populaatiosta ja vuoden 2020 lopulla mobiili internetyhteys oli käytössä yli puolella väestöstä (GSM Association, ei pvm.; Kemp, 2019). Koska matkapuhelinoperaattorit toimivat usein tietyillä maantieteellisillä alueilla ja samoilla alueilla voi olla toiminnassa monia matkapuhelinoperaattoreita, täydellisen kattavuuden saamiseksi on data kuitenkin usein hankittava usealta eri yritykseltä ja älypuhelinsovellusten sekä eri käyttöjärjestelmien käyttöasteet voivat vaihdella eri maiden osalta (Ruktanonchai ym., 2018).

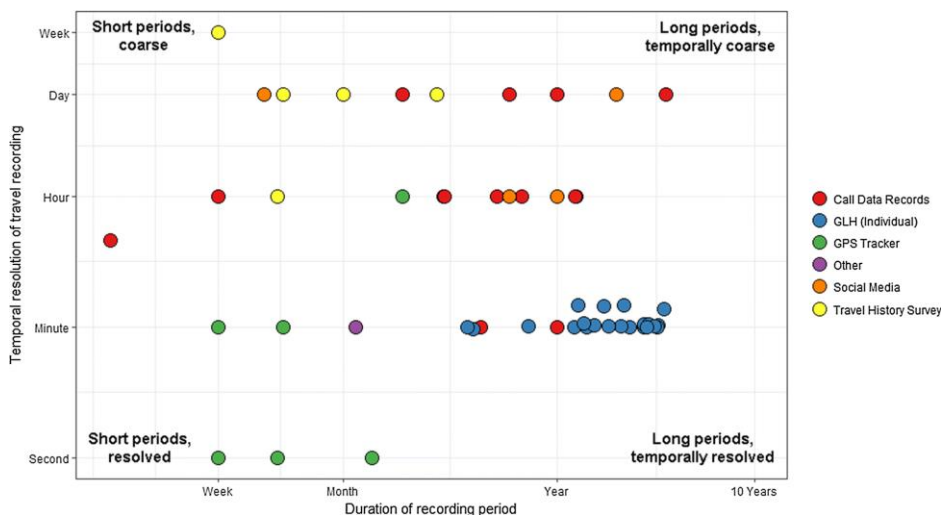
5.2 Aika

Temporaalinen resoluutio on suure, joka vaihtelee paljon eri datalähteiden välillä. Esimerkiksi sosiaalisen median sovellukset tuottavat sijaintitietoja useimmiten vain silloin, kun käyttäjä aktiivisesti tekee sijaintimerkityn päivityksen palveluun, ja näiden päivitysten vaihteluväli voi olla suurta niin yksittäisellä käyttäjällä kuin myös vertailtaessa useiden käyttäjien päivitystiheyttä keskenään (Jurdak ym., 2015). Sama pätee myös puhelurekisteritietoihin, sillä ihmisten soittamien ja vastaanottamien puheluiden määrä voi vaihdella huomattavasti riippuen sijainnista, henkilöstä ja tarkasteltavasta ajanjaksosta (Zhao ym., 2016). Esimerkiksi Ruktanonchain (2018) (kuva 1) teettämän kirjallisuuskatsauksen mukaan puhelurekisteritietojen temporaalinen resoluutio ihmisten liikkumista tarkastelleissa tutkimuksissa vaihteli minuuteista päiviin, ja sosiaalisen median päivityksistä saadun datan tunneista päiviin.

Älypuhelinsovellusten keräämä data on temporaaliselta resoluutioltaan usein paljon puhelurekisteritietoja tiheämpää. Esimerkiksi Google kerää Location History -palveluunsa puhelimen sijaintitietoja Ruktanonchain (2018) tutkimuksen mukaan noin minuutin välein (kuva 1). Sijaintitiedon keräyksen aikaväli kuitenkin riippuu sovelluksesta ja käyttäjän asettamista yksityisyysasetuksista ja jotkin sovellukset

keräävät sijaintitietoja ainoastaan sovelluksen ollessa aktiivisessa käytössä (Calacci ym., 2019). Myös matkapuhelinoperaattoreiden matkapuhelinverkkopohjaisilla menetelmillä kerätyn liikkumisdatan temporaalinen resoluutio on puhelurekisteritietoja parempi ja sijaintitietoja saadaan tasaisin väliajoin tehtävillä päivityksillä, tai kun puhelin siirtyy mastojen kantoalueelta toiselle (Pinelli ym., 2015; Willberg ym., 2021). Matkapuhelinverkkopohjaisilla menetelmillä puhelimen sijaintia pystytään näin periaatteessa tarkastelemaan reaaliajassa ainakin linkkimastojen kantoalueiden välillä.

Suomessa puhelinoperaattorit ovat velvoitettuja säilyttämään puhelurekisteritietoja vuoden ajan järjestelmissään (Liikenne- ja viestintäministeriö, 2017), mutta esimerkiksi Ruktanonchain (2018) (kuva 1) teettämässä kirjallisuuskatsauksessa puhelurekisteritietojen temporaalinen kattavuus oli joissain tapauksissa suurempi. Muuta matkapuhelinverkkoliikenteestä kerättyä dataa ei välttämättä säilytetä matkapuhelinoperaattoreiden toimesta pidempiä aikoja ellei tähän ole operaattorin puolesta erityistä insentiiviä (Ghahramani ym., 2020). Siihen kuinka pitkään dataa saa säilyttää vaikuttaa myös lainsäädäntö, ja esimerkiksi GDPR määrittelee henkilökohtaisen datan säilytysajaksi lyhyimmän mahdollisen ajan mikä on tarpeellista datan prosessoimiseksi sen käyttötarkoitusta varten (Bu-Pasha ym., 2017; Gruschka ym., 2018).



Kuva 1 Eri datalähteiden temporaalinen resoluutio ja kattavuus kirjallisuuskatsauksen perusteella (Ruktanonchai ym., 2018, s. 8).

5.3 Ominaisuustiedot

Ominaisuustietojen osalta älypuhelinsovelluksista voidaan potentiaalisesti kerätä huomattavan paljon erilaisia käyttäjätietoja riippuen käyttäjän suostumuksesta tietojen jakamiseen. Esimerkiksi sosiaalisen median alustoilla ihmiset jakavat alustoille usein henkilötietojaan, kuten ikänsä, sukupuolensa ja ammattinsa (Toivonen ym., 2019). Lisäksi ihmisten tekemät teksti- ja kuvapäivitykset sekä ystäväverkostonsa voivat paljastaa tietoja heidän sosiaalisesta piiristään ja mieltymyksistään (Calacci ym., 2019).

Sosiaalisen median sisältöanalyysin avulla voidaan potentiaalisesti kerätä myös sijaintitietoja tarkastelemalla tekstissä ja kuvissa esiintyvää sijaintiin liitettävää informaatiota (Middleton ym., 2018). Sisältöanalyysia voidaan käyttää näin lisäksi paikantamisen tukena tai antamaan lisää kontekstuaalista sijaintitietoa, kuten mitkä ja minkä tyyppiset alueet esiintyvät ihmisten viesteissä ja kuvissa. Näiden eri tietolähteiden hyödyntäminen on ottanut suuria harppauksia viime vuosien aikana kun sisältöanalyysissä on alettu hyödyntämään koneoppimismenetelmiä tekstin ja visuaalisen aineiston analysointiin (Middleton ym., 2018; Toivonen ym., 2019).

Myös eri älypuhelinsovellusten käyttöä seuraamalla pystytään keräämään paljon kontekstuaalista tietoa ihmisistä, ja näitä tietoja kyetään usein yhdistämään puhelimen sijaintitietoihin (Xie ym., 2018). Matkapuhelinverkosta kerätystä datassa ei usein ole saatavilla yhtä tarkkoja ominaisuustietoja, joskin matkapuhelinverkosta kerättyihin sijaintitietoihin pystytään mahdollisesti liittämään kolmannen osapuolen dataa mikäli puhelimen tunnistetiedot ovat yhdistettävissä (Telia Company, ei pvm.). Lisäksi puhelurekisteritiedoista voidaan esimerkiksi päätellä ihmisten sosiaalisia verkostoja tarkastelemalla mitkä liittymät ovat yhteydessä toisiinsa (Järv ym., 2021).

6. Mobiilin big datan saatavuus

6.1 Datan saatavuus ja jakamisolustat

Kuten aiemmissa kappaleissa todettiin, sijaintitietoja matkapuhelimista keräävät pääasiallisesti matkapuhelinoperaattorit sekä älypuhelimien sovelluksia tarjoavat ja niitä hallinnoivat yritykset. Se kenelle nämä sijaintitiedot ovat saatavilla, mitä kautta ja missä muodossa riippuu pitkälti sitä keräävän tahon datapolitiikasta sekä paikallisesta yksityisyydensuojalainsäädännöstä. Esimerkiksi Euroopassa

matkapuhelimista kerätyn sijaintitiedon kerääminen ja saatavuus on monilta osin vaikeutunut GDPR-lainsäädännön myötä (Gruschka ym., 2018). Yritykset voivat käyttää liikkumisdataa sisäisesti omiin analyyseihinsä, myydä sitä eteenpäin joko erikseen tai yhdessä muiden datalähteiden kanssa (K. E. Martin, 2015) ja joskus tutkijat saavat dataa suoraan käyttöönsä dataa kerääviltä yrityksiltä, matkapuhelinoperaattoreilta ja sosiaalisen median alustoilta (Apple, ei pvm.; Bengtsson ym., 2015; Google, ei pvm.; Järv ym., 2012; Jurdak ym., 2015; X-Mode, ei pvm.).

Matkapuhelinoperaattoreiden keräämän datan saatavuus vaihtelee paljon operaattoreiden datapolitiikan että alueellisen lainsäädännön mukaan. Esimerkiksi EU:n sisällä kaikkien operaattoreiden on lain mukaan kerättävä ja tallennettava puhelurekisteritietoja määrääjäksi, mutta tämä aineisto on harvoin saatavilla sellaisenaan tutkijoille (Liikenne- ja viestintäministeriö, 2017). Suomessa puhelurekisteritiedot ovat pyynnöstä saatavilla lähinnä viranomaisille rikosepäilyjen selvittämiseen ja hätätilanteisiin reagoimiseen (Liikenne- ja viestintäministeriö, 2017), mutta operaattorit ovat joissain maissa, esimerkiksi Virossa (Järv ym., 2012) ja Kiinassa (Zhao ym., 2016) olleet aktiivisesti tekemässä yhteistyötä tutkijoiden kanssa jakaen anonymisoituja puhelurekisteritietoja tutkimuskäyttöön.

Puhelurekisteritietojen jakamisen sijaan esimerkiksi matkapuhelinoperaattori Telia on aloittanut myymään Suomessa verkkoliikenteestään kerättyä aggregoitua ja anonymisoitua liikkumisdataa (Telia Company, ei pvm.). Koska matkapuhelimien sijainnin laskenta matkapuhelinverkkosignaalien avulla vaatii kuitenkin laskentatehoa sekä tallennuskapasiteettia, ei kaikilla matkapuhelinverkkojen ylläpitäjillä ole välttämättä kapasiteettia tarkkojen sijaintien laskentaan ja tallentamiseen eikä sijaintihistoriaa välttämättä tallenneta ellei operaattorilla ole sille erityistä käyttöä (Ghahramani ym., 2020; Pinelli ym., 2015).

Suoraviivaisin liikkumisdatan lähde älypuhelinsovelluksista ovat saatavuuden osalta sosiaalisen median palvelut, kuten Twitter, joissa ihmiset voivat tehdä julkisia sijaintimerkittyjä päivityksiä. Päivitysten ollessa usein julkisesti kaikkien saatavilla ovat yksityisyydensuojaan liittyvät ongelmallisuudet vähäisempiä kuin monissa muissa mobiilin big datan lähteissä, vaikkakin myös sosiaalisen median alustoilla

sijaintitiedon saatavuutta on rajoitettu monilla alustoilla juuri käyttäjien yksityisyydensuojaan vedoten (Hu & Wang, 2020; Tenkanen ym., 2017). Myös eri ominaisuustietojen saatavuus on riippuvaista sosiaalisen median alustasta ja käyttäjien halukkuudesta jakaa tietoja palveluun (Toivonen ym., 2019). Useilla sosiaalisen median palveluilla, kuten Twitterillä, on lisäksi käytössä erityinen ohjelmointirajapinta, jonka avulla on mahdollista ladata ja suodattaa päivityksiä sijainnin ja ajanjakson mukaan tehden sosiaalisesta mediasta helposti käytettävän ja saatavilla olevan datalähteen (Gengec, ei pvm.; Gulnerman & Karaman, 2020). Datan saatavuuden ollessa kuitenkin riippuvaista yrityksen ylläpitämästä ohjelmointirajapinnasta, voi datan saatavuus ja käyttöoikeudet muuttua nopeasti ilman ennakoilmoitusta (Toivonen ym., 2019).

Muista kuin julkisista sosiaalisen median älypuhelinsovelluksista kerätyn datan saatavuus on enemmän riippuvaista yritysten datanjakopolitiikasta, yksityisyydensuojalainsäädännöstä ja käyttäjien suostumuksesta. Data voidaan saada käyttöön suoraan älypuhelinsovellusta ylläpitävältä yritykseltä ja sijaintitietoa keräävät usein myös kolmannen osapuolen yritykset – niin kutsutut datameklarit (*data broker*) – useista eri sovelluksista ja muista datalähteistä, joita myydään isoissa datapaketeissa esimerkiksi markkinointitarkoituksiin (Calacci ym., 2019; Datarade, ei pvm.; X-Mode, ei pvm.). Sijaintitietojen kerääminen ja jälleenmyynti älypuhelinsovelluksista on huomattavan yleistä ja esimerkiksi vuonna 2014 kolmanneksella suosituimmista Android-sovelluksista ja puolella iOS-sovelluksista oli pääsy puhelimen sijaintitietoihin, ja noin puolet kaikista Android- ja iOS-sovelluksista kerätyistä sijaintitiedoista oli verkkomainostajien saatavilla (Bu-Pasha ym., 2017; de Montjoye ym., 2013). Covid-19-pandemian aikaan Google ja Apple ovat myös tarjonneet keräämäänsä aggregoitua liikkumisdataa vapaasti ladattavaksi, jotta tutkijat ja muut tahot pystyisivät reagoimaan paremmin pandemian aiheuttamiin ja vaatimiin muutoksiin ihmisten liikkuvuudessa (Apple, ei pvm.; Google, ei pvm.).

Myös erityisesti tutkimuskäyttöön tarkoitettuja älypuhelinsovelluksia on kehitetty ja käytetty vastaamaan rajattuihin tutkimuskysymyksiin esimerkiksi terveystieteissä. Tutkimuskäyttöön tarkoitetuissa sovelluksissa etuna on datan suora saatavuus sekä muokkausmahdollisuudet tutkimuskysymyksen tarpeisiin, joka mahdollistaa esimerkiksi yksityiskohtaisten ominaisuustietojen keräämisen. Näiden aktiivisen

osallistamisen vaatimuksen vuoksi otoskoot jäävät kuitenkin usein pieniksi verrattuna passiivisesti kerättyihin mobiilin big datan lähteisiin (Freifeld ym., 2010; Gupta ym., 2021; Sottile ym., 2021).

6.2 Yksityisyydensuoja

Tarkan liikkumisdatan saatavuuteen vaikuttaa datan omistusoikeuksien ja yritysten datanjakopolitiikan lisäksi paljon yksityisyydensuojakysymykset. Yksilökohtainen liikkumisdata on potentiaalisesti arkaluonteista tietoa, jonka avulla voidaan tulkita esimerkiksi ihmisten poliittisia, uskonnollisia tai seksuaalisia suuntauksia ja muita henkilökohtaisia tietoja tarkastelemalla sijainteja missä hän on käynyt. Tämän vuoksi esimerkiksi EU:n yksityisyydensuojalainsäädäntö velvoittaa että matkapuhelimista saatavilla oleva mobiili big data aggregoidaan ja anonymisoidaan sellaiselle tasolle, että yksittäisiä henkilöitä ei pystytä datasta tunnistamaan (Bu-Pasha ym., 2017).

Datan anonymisointi ja aggregointi tarkoittaa esimerkiksi spatiaalisen ja temporaalisen kattavuuden ja resoluution pienentämistä sekä henkilökohtaisten tunnistetietojen poistamista aineistosta. Henkilökohtaisten tunnistetietojen, kuten puhelinnumeron tai nimen sijaan yksittäisen puhelimen liikeradat voidaan yhdistää satunnaisesti luodulla tunnisteella mistä puhelimen omistajan henkilöllisyys ei käy ilmi (Calacci ym., 2019). Lisäksi yksittäisten puhelinten tunnistetiedot voidaan poistaa aineistosta kokonaan, jolloin puhelinten liikeratojen sijaan aineisto sisältää vain paikkojen tai alueiden käyttäjien määrän eri ajanhetkinä (Calacci ym., 2019; Ma ym., 2013; Xu ym., 2017).

Liikkumisdatan anonymisointi ei ole kuitenkaan yksinkertaista, sillä vaikka matkapuhelimen liikkumisdata anonymisoitaisiin niin, ettei henkilön tunnistetietoja olisi datassa saatavilla, liikeratojen ollessa riittävän yksityiskohtaisia on data usein liitettävissä yksittäisiin henkilöihin (de Montjoye ym., 2013). Ihmisten liikkumistottumusten on havaittu olevan hyvin yksilöllisiä ja tarkoilla koordinaatti- ja tunnistetiedoilla liikeradat pystytään lähes poikkeuksetta yhdistämään yksittäisiin henkilöihin ja naapurustotasollekin aggregoidussa aineistossa jo neljän havaintopisteen avulla voidaan liikeradat yhdistää yksittäiseen henkilöön 95% varmuudella (Calacci ym., 2019). Myös kun liikeratojen tunnistetiedot poistetaan aineistosta kokonaan, pystytään yksittäiset liikeradat seulomaan aineistosta huomattavalla varmuudella ja todennäköisesti yhdistämään myös yksittäiseen

henkilöön hyödyntämällä ihmisten vähäistä liikkuvuutta yöaikaan, päiväsajan liikkumisen jatkuvuutta ja ihmisten liikeratojen yksilöllisyyttä, sekä soveltamalla matemaattisia ennustemalleja liikeratojen luomiseksi (Xu ym., 2017).

Toinen yksityisyydensuojalle ongelmallinen mobiilin big datan piirre on eri datalähteiden yhdistäminen. Jokaisella matkapuhelimella on useita uniikkeja tunnisteita, esimerkiksi IMEI-koodi (*International Mobile Equipment Identity*), jonka avulla useasta eri lähteestä kerätty data voidaan yhdistää tiettyyn puhelimeen (Bu-Pasha ym., 2017). Näin puhelinten sijaintitiedoista tulee vielä helpommin tunnistettavia, henkilökohtaisia ja potentiaalisesti arkaluonteisia, mikäli sovelluksilla on pääsy tähän tunnisteseen ja sovelluksiin syötetyt henkilötiedot tai niiden käyttötiedot pystytään yhdistämään sijaintitietoihin. Myös itse sijaintitiedot voivat tarkentua huomattavasti ja heikentää näin anonymiteettiä, mikäli useammasta sovelluksesta tai muista mobiilin big datan lähteistä saadut sijaintitiedot pystytään yhdistämään yksittäiseen matkapuhelimeen (Bu-Pasha ym., 2017).

Koska yleisesti käytetyt aggregointimenetelmät, kuten spatiaalisen ja temporaalisen resoluution heikentäminen, ovat huomattavan tehottomia tapoja anonymisoida aineistoa, ihmisten anonymiteetin säilyttämiseksi data joudutaan aggregoimaan sellaiselle tasolle, että sen käytettävyys tietyissä tutkimuskysymyksissä heikentyy (Calacci ym., 2019; Xu ym., 2017). Lisäksi datan ollessa valmiiksi aggregoitua, tutkijat eivät pysty tekemään datalle omia spatiaalisia ja kategorisia aggregointejaan, eivätkä validoimaan datan käytettävyttä (Poom ym., 2020). Tämän johdosta liikkumisdatan anonymisointimenetelmät, missä pyritään säilyttämään mahdollisimman tehokkaasti alkuperäisen datan käytettävyys tutkimukseen yksityisyydensuojaa kunnioittaen, ovat keränneet viime vuosina paljon tutkimustietoa (Smolak ym., 2020; P. Yang ym., 2011; Y. Zhang ym., 2016).

Suurilta osin yksityisyydensuojakysymysten vuoksi tarkemman aggregointitason tutkimuksissa joudutaan usein käyttämään aktiivisempaa osallistamista, jossa käyttäjät vapaaehtoisesti luovuttavat henkilötietojaan ja tarkkaa liikkumisdataansa tutkimukseen. Esimerkiksi Googlen Location History -palvelu tallentaa käyttäjien liikkumista hyvin tarkalla spatiaalisella, että temporaalisella resoluutiolla ja kattavuudella (Ruktanonchai ym., 2018) (kuva 1), mutta data ei ole vapaasti

ladattavissa ja käytettävissä tutkimuskäyttöön ja yksittäisiä käyttäjiä on pyydetty manuaalisesti jakamaan heidän datansa tutkijoille.

Saatavilla oleva data kuitenkin ei aina vastaa sitä dataa mitä on alun perin kerätty ja mitä yrityksellä on sen sisäisessä tai viranomaiskäytössä. Niin kuin aiemmin todettiin, matkapuhelinoperaattoreilla on useissa maissa velvollisuus kerätä puhelurekisteritietoja lainsäädännöllisistä syistä ja dataa kerätään myös verkkoliikenteen hallintaa ja laskutusta varten, mutta tätä dataa ei useissa maissa ole laillista jakaa eteenpäin julkisesti tai tutkimuskäyttöön (Liikenne- ja viestintäministeriö, 2017; Teng & Chou, 2007). Esimerkiksi GDPR kuitenkin asettaa ehtoja myös yritysten sisäisesti hallinnoituun dataan, jossa periaatteena on henkilökohtaisen datan määrän, säilöntäajan ja käytön minimalisointi sovelluksen toiminnallisuuksien puitteissa (Bu-Pasha ym., 2017; Gruschka ym., 2018). Telian mukaan Crowd Insights -palvelun liikkumisanalyseissäkin käytetty aineisto automaattisesti anonymisoidaan niin, että kukaan yksittäinen henkilö ei pääse käsiksi yksilökohtaiseen liikkumistietoon ja käyttäjät voivat halutessaan lisäksi olla antamatta liikkumistietojaan palvelun käyttöön (Telia Company & Grön, ei pvm.).

Periaatteessa yritysten on mahdollista kerätä ja jakaa liikkumisdataa myös salassa ilman ihmisten suostumusta (Bu-Pasha ym., 2017; Momen ym., 2019). Tämä on esimerkiksi EU:n sisällä lainvastaista, mutta monet älypuhelinsovellukset eivät välttämättä kerro selvästi mihin ihmisten sijaintitietoja tullaan käyttämään ja jakamaan tai osa tiedosta on vaikeasti löydettävissä yksityisyydensuojalausunnoissa (Calacci ym., 2019). Lisäksi vaikka käyttäjiltä oltaisiin kysytty suostumus datan keräykseen ja käyttöön, eivät käyttäjät ole kuitenkaan usein tietoisia mihin heidän dataansa käytetään ja jaetaan (Jones ym., 2019; Weller & Kinder-Kurlanda, 2016). GDPR-lainsäädäntö on EU:n sisällä asettanut tiukempia säädöksiä yrityksille kertoa tarkasti asiakkaille mitä dataa heistä kerätään ja mihin sitä käytetään, mikä on vähentänyt älypuhelinsovellusten kysymiä sijainti- ja ominaisuustietoja. Ei ole kuitenkaan täyttä varmuutta onko tämä vähentänyt tietojen keräämistä vai vain muuttanut tapaa miten yritykset kysyvät hyväksyntää tietojen keräämiseksi (Momen ym., 2019).

7. Yhteenveto

Mobiilin big datan osalta voidaan todeta, että suurimmat esteet sen hyödyntämisessä ihmisten liikkumisen tutkimukseen ei ole niinkään teknologiset rajoitteet, sillä älypuhelimien kattavuus väestössä on todella hyvä suurella osalla maapalloa ja niihin upotettu teknologia mahdollistaa tarkan liikkumisdatan keräämisen hyvällä spatiaalisella, että temporaalisella resoluutiolla. Matkapuhelinten, joista ei löydy älypuhelinominaisuuksia kuten internetyhteyttä tai GNSS-paikannusta, kattavuus on vielä älypuhelimienkin suurempi ja hyödyntämällä kehittyneitä matkapuhelinverkkopohjaisia paikantamismenetelmiä päästään puhelinverkkoliikenteestä kerätyllä datalla hyvään spatiaaliseen ja temporaaliseen resoluutioon etenkin alueilla, jossa linkkimastojen väliset etäisyydet ovat pieniä (S. Zhang ym., 2020). Isommat kysymykset liittyvät siihen, kuinka tarkkaa liikkumisdataa saadaan sitä kerääviltä yrityksiltä tutkimuskäyttöön, sekä miten ihmisten yksityisyydensuoja pystytään turvaamaan.

Matkapuhelinten liikkumisdatan ollessa valtaosin yksityistä tai yritysomisteista tietoa, on sen saatavuus tutkimuskäyttöön riippuvaista ihmisten ja yritysten halukkuudesta jakaa dataa eteenpäin. Poomin ym. (2020) mukaan tutkijat usein joutuvat tyytymään yritysten tuottamiin aggregoituihin datasetteihin, joista puuttuvat tarvittavat tiedot alkuperäisestä kerätyistä datasta tai niiden spatiaalinen ja temporaalinen resoluutio on liian huono että niiden perusteella voisi tehdä luotettavia tieteellisiä johtopäätöksiä. Tutkijat pääsevät harvoin käsiksi matkapuhelinoperaattoreiden keräämään tarkkaan dataan, vaikkakin yritykset ajoittain tekevät tietyissä maissa tutkimusyhteistyötä tutkijoiden ja viranomaisten kanssa. Älypuhelinsovelluksista kerätyn mobiilin big datan saatavuus on vieläkin epävarmempaa, sillä sovelluskehittäjiä tai useista lähteistä dataa kerääviä datameklareita ei usein ole velvoitettu jakamaan samalla tavalla dataa yleishyödylliseen käyttöön niin kuin matkapuhelinoperaattorit, joiden käyttämät linkkimastot on esimerkiksi Yhdysvalloissa rakennettu julkisella avustuksella (Calacci ym., 2019). Kun datalla on huomattavaa markkina-arvoa esimerkiksi markkinoinnissa, on datan omistavilla yrityksillä myös vähän insentiiviä antaa kerättyä dataa vapaasti käytettäväksi.

Datan keruuseen ja käyttöön liittyy myös huomattavia yksityisyydensuojaan liittyviä kysymyksiä. Tarkka liikkumisdata mahdollistaa liikeratojen yhdistämisen yksittäisiin henkilöihin sekä henkilökohtaisten ja mahdollisten arkaluontoisten liikkumistietojen saatavuuden, ja monet käytetyt anonymisointi ja aggregointimenetelmät ovat riittämättömiä anonymiteetin takaamiseksi. Lisäksi ihmisten hyväksyntä ja tietoisuus siitä, mitä dataa heistä kerätään ja minne sitä jaetaan vaihtelee, eikä datan keräämisestä kieltäytyminen usein ole käytännössä vaihtoehto matkapuhelimien sekä älypuhelimien paikannusominaisuuksien tultua yhä suuremmaksi osaksi ihmisten elämää.

Jotta datan saatavuuteen ja yksityisyydensuojaan liittyvät kysymykset saadaan ratkaistua, Poom ym. (2020, s. 4) ovat pohtineet että yritysten ja tutkijoiden välille tarvitaan yhteistyöverkostoja, jossa yritysten keräämää raakaa liikkumisdataa voitaisiin käyttää tutkimukseen turvatuissa ja yksityisyydensuojaa kunnioittavissa olosuhteissa. Tämä kuitenkin vaatisi todennäköisesti lainsäädännöllisiä muutoksia, joka velvoittaisi yrityksiä jakamaan dataansa eteenpäin tutkijoille. Toisekseen mitkä virastot ja tutkimuslaitokset olisivat oikeutettuja käyttämään dataa jää vielä avoimeksi.

Mikäli tutkijoiden ja yritysten välistä yhteistyötä ei saada rakennettua ja lainsäädännöllisiä kysymyksiä ratkaistua, joutuvat tutkijat todennäköisesti myös tulevaisuudessa tyytymään joko pienen kattavuuden ja resoluution omaaviin datalähteisiin, kuten sosiaalisen median päivityksiin, tai yritysten mielivaltaisesti tarjoamaan aggregoituun dataan. Tarkempaa dataa tarvitsevat tutkimukset on tällöin toteutettava aktiivisen osallistamisen keinoilla, mutta otoskoot jäävät tällöin huomattavasti passiivista datankeruuta hyödyntäviä lähteitä pienemmiksi. Tämä on kuitenkin todennäköisesti myös tulevaisuudessa ainoa tapa saada liitettyä yksilökohtaisia ominaisuustietoja liikkumisdataan.

Jos yksityisyydensuojaan ja datan saatavuuteen liittyvät kysymykset saadaan ratkaistua, on matkapuhelinten liikkumisdatalla huomattavasti käyttämätöntä potentiaalia monilla eri tieteenaloilla. Esimerkiksi sosiaali- ja terveystieteissä on kysyntää datalle, jonka avulla pystyttäisiin tarkastelemaan alueellisia terveysriskejä ja sosiaalista segregatiota hyvällä resoluutiolla ja kattavuudella (Matthews & Yang,

2013; Ruktanonchai ym., 2018), johon pystytään liittämään myös sosiaalista ja demografista ominaisuusdataa ja mobiili big data voi tulevaisuudessa olla korvaamaton apuväline tämänkaltaisen tutkimuksen toteuttamisessa.

Lähteet

- Apple. (ei pvm.). *Mobility Trends Reports*. Noudettu 15. joulukuuta 2021, osoitteesta <https://covid19.apple.com/mobility>
- Bar-Noy, A., & Kessler, I. (1993). Tracking Mobile Users in Wireless Communications Networks. *IEEE Transactions on Information Theory*, 39(6), 1877–1886. <https://doi.org/10.1109/18.265497>
- Bengtsson, L., Gaudart, J., Lu, X., Moore, S., Wetter, E., Sallah, K., Rebaudet, S., & Piarroux, R. (2015). Using Mobile Phone Data to Predict the Spatial Spread of Cholera. *Scientific Reports*, 5, 8923. <https://doi.org/10.1038/srep08923>
- Brockmann, D., Hufnagel, L., & Geisel, T. (2006). The scaling laws of human travel. *Nature (London)*, 439(7075), 462–465. <https://doi.org/10.1038/nature04292>
- Bu-Pasha, S., Alén-Savikko, A., Mäkinen, J., Guinness, R., & Korpisaari, P. (2017). EU Law Perspectives on Location Data Privacy in Smartphones and Informed Consent for Transparency. *European Data Protection Law Review*, 2(3), 312–323. <https://doi.org/10.21552/edpl/2016/3/7>
- Calacci, D., Berke, A., Larson, K., & Pentland, A. (2019). The tradeoff between the utility and risk of location data and implications for public good. *arXiv:1905.09350v2*.
- Chang, S., Pierson, E., Koh, P. W., Gerarding, J., Redbird, B., Grusky, D., & Leskovec, J. (2020). Mobility network models of COVID-19 explain inequities and inform reopening. *Nature*, 589, 82–87. <https://doi.org/10.1038/s41586-020-2923-3>
- Chen, C., Ma, J., Susilo, Y., Liu, Y., & Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C-Emerging Technologies*, 68, 285–299. <https://doi.org/10.1016/j.trc.2016.04.005>
- Chrisman, N. (1991). The error component in spatial data. *Geographical Information Systems: Principles and Applications*, 1(12), 165–174. https://www.researchgate.net/publication/243624498_The_error_component_in_spatial_data
- Datarade. (ei pvm.). *Mobile Location Data*. Noudettu 16. joulukuuta 2021, osoitteesta <https://datarade.ai/data-categories/mobile-location-data>
- de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the Crowd: The privacy bounds of human mobility. *Scientific reports*, 3(1), 1376. <https://doi.org/10.1038/srep01376>
- Eagle, N., & Pentland, A. (2005). Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, 10(4), 255–268. <https://doi.org/10.1007/s00779-005-0046-3>
- Finazzi, F., & Paci, L. (2019). Quantifying personal exposure to air pollution from smartphone-based location data. *Biometrics*, 75, 1356–1366. <https://doi.org/10.1111/biom.13100>

- Freifeld, C. C., Chunara, R., Mekaru, S. R., Chan, E. H., Kass-Hout, T., Ayala Iacucci, A., & Brownstein, J. S. (2010). Health in Action Participatory Epidemiology: Use of Mobile Phones for Community-Based Health Reporting. *PLoS Medicine*, 7(12), 1–5. <https://doi.org/10.1371/journal.pmed.1000376>
- Ganjour, O., Widmer, E. D., Viry, G., Gauthier, J.-A., Kaufmann, V., & Drevon, G. (2020). Understanding the Reconstruction of Personal Networks Through Residential Trajectories. *Migration Letters*, 17(5), 621–638. <https://doi.org/10.33182/ml.v17i5.694>
- Gengec, N. (ei pvm.). *Geo Tweets Downloader*. Noudettu 17. joulukuuta 2021, osoitteesta <https://github.com/nagellette/geo-tweet-downloader/>
- Ghahramani, M., Zhou, M., & Wang, G. (2020). Urban sensing based on mobile phone data: approaches, applications, and challenges. *IEEE/CAA journal of automatica sinica*, 7(3), 627–637. <https://doi.org/10.1109/JAS.2020.1003120>
- Ghose, A., Li, B., & Liu, S. (2019). Mobile Targeting Using Customer Trajectory Patterns. *Management Science*, 65(11), 5027–5049. <https://doi.org/10.1287/mnsc.2018.3188>
- Google. (ei pvm.). *COVID-19 Community mobility reports*. Noudettu 16. joulukuuta 2021, osoitteesta <https://www.google.com/covid19/mobility/>
- Gruschka, N., Mavroeidis, V., Vishi, K., & Jensen, M. (2018). Privacy Issues and Data Protection in Big Data: A Case Study Analysis under GDPR. *Proceedings - 2018 IEEE International Conference on Big Data (Big Data)*, 5027–5033. <https://doi.org/10.1109/BigData.2018.8622621>
- GSM Association. (ei pvm.). *The State of Mobile Internet Connectivity 2021*. Noudettu 16. joulukuuta 2021, osoitteesta <https://www.gsma.com/r/somic/#regions>
- Gulnerman, A. G., & Karaman, H. (2020). Spatial Reliability Assessment of Social Media Mining Techniques with Regard to Disaster Domain-Based Filtering. *Isprs International Journal of Geo-Information*, 9(4), 245. <https://doi.org/10.3390/ijgi9040245>
- Gupta, S., Sharma, S., Gohil, R., & Sachdeva, S. (2021). Epicollect5: A Free, Fully Customizable Mobile-based Application for Data Collection in Clinical Research. *Journal of Postgraduate Medicine, Education and Research*, 54(4), 248–251. <https://doi.org/10.5005/jp-journals-10028-1375>
- Handy, S. (2002). *Accessibility-vs. Mobility-Enhancing Strategies for Addressing Automobile Dependence in the US*. UC Davis: Institute of Transportation Studies. <https://escholarship.org/uc/item/5kn4s4pb>
- Hasnat, M. M., & Hasan, S. (2018). Identifying tourists and analyzing spatial patterns of their destinations from location-based social media data. *Transportation Research Part C: Emerging Technologies*, 96(January), 38–54. <https://doi.org/10.1016/j.trc.2018.09.006>
- Horak, R. (2007). Voice Communications Systems: KTS, PBX, Centrex, and ACD. Teoksessa *Telecommunications and Data Communications Handbook* (ss. 101–153). <https://doi.org/https://doi.org/10.1002/9780470127247.ch3>

- Hu, Y., & Wang, R.-Q. (2020). Understanding the removal of precise geotagging in tweets. *Nature Human Behaviour*, 4, 1219–1221.
<https://doi.org/10.1038/s41562-020-00949-x>
- IDC. (ei pvm.). *Smartphone Market Share*. Noudettu 17. joulukuuta 2021, osoitteesta
<https://www.idc.com/promo/smartphone-market-share>
- Järv, O., Ahas, R., Saluveer, E., Derudder, B., & Witlox, F. (2012). Mobile Phones in a Traffic Flow: A Geographical Perspective to Evening Rush Hour Traffic Analysis Using Call Detail Records. *PLoS ONE*, 7(11).
<https://doi.org/10.1371/journal.pone.0049171>
- Järv, O., Tenkanen, H., Salonen, M., Ahas, R., & Toivonen, T. (2018). Dynamic cities: Location-based accessibility modelling as a function of time. *Applied Geography*, 95(May), 101–110. <https://doi.org/10.1016/j.apgeog.2018.04.009>
- Järv, O., Tominga, A., Müürisepp, K., & Silm, S. (2021). The impact of COVID-19 on daily lives of transnational people based on smartphone data: Estonians in Finland. *Journal of Location Based Services*, 1–29.
<https://doi.org/10.1080/17489725.2021.1887526>
- Joksić, D., & Bajat, B. (2004). Elements of spatial data quality as information technology support for sustainable development planning. *Spatium*, 11, 77–83.
<https://doi.org/10.2298/SPAT0411077J>
- Jones, K. H., Daniels, H., Heys, S., & Ford, D. V. (2019). Public Views on Using Mobile Phone Call Detail Records in Health Research: Qualitative Study. *JMIR mHealth and uHealth*, 7(1). <https://doi.org/10.2196/11730>
- Jurdak, R., Zhao, K., Liu, J., AbouJaoude, M., Cameron, M., & Newth, D. (2015). Understanding Human Mobility from Twitter. *PLOS ONE*, 10(7), e0131469.
<https://doi.org/10.1371/journal.pone.0131469>
- Kaufmann, V., Bergman, M., & Joye, D. (2004). Motility: Mobility as Capital. *International Journal of Urban and Regional Research*, 28(4), 745–756.
<https://doi.org/10.1111/j.0309-1317.2004.00549.x>
- Kemp, S. (2019). *Digital 2019: Global Digital Overview*.
<https://datareportal.com/reports/digital-2019-global-digital-overview>
- Kostyniuk, L., & Kitamura, R. (1982). Life Cycle and Household Time-Space Paths: Empirical Investigation. *Transportation Research Record*, 879, 28–37.
- Kuhn, W. (2012). Core concepts of spatial information for transdisciplinary research. *International Journal of Geographical Information Science*, 26(12), 2267–2276.
<https://doi.org/10.1080/13658816.2012.722637>
- Last, J. (2007). *A Dictionary of Public Health*. Oxford University Press.
<https://doi.org/10.1093/acref/9780195160901.001.0001>
- Lechner, W., & Baumann, S. (2000). Global navigation satellite systems. *Computers and Electronics in Agriculture*, 25(1), 67–85.
[https://doi.org/https://doi.org/10.1016/S0168-1699\(99\)00056-3](https://doi.org/https://doi.org/10.1016/S0168-1699(99)00056-3)
- Liikenne- ja viestintäministeriö. (2017). Selvitys sähköisen viestinnän välitystietojen säilytysvelvollisuudesta. Teoksessa *RAPORTIT JA SELVITYKSET 9/2017*.

<http://urn.fi/URN:ISBN:978-952-243-519-4>

- Ma, J., Li, H., Yuan, F., & Bauer, T. (2013). Deriving Operational Origin-Destination Matrices From Large Scale Mobile Phone Data. *International Journal of Transportation Science and Technology*, 2(3), 183–204. <https://doi.org/10.1260/2046-0430.2.3.183>
- Madden, S. (2012). From Databases to Big Data. Teoksessa *IEEE Internet Computing* (Vsk. 16, Numero 3, ss. 4–6). IEEE. <https://doi.org/10.1109/MIC.2012.50>
- Martin, J., Panario, M., & Alberto, J. (2009). Fast-Trilateration + Cell-ID Method for Mobile Phones Location Estimation. *Reunión de Procesamiento de Información y Control*.
- Martin, K. E. (2015). Ethical Issues in the Big Data Industry. *MIS Quarterly Executive*, 14(2). <https://aisel.aisnet.org/misqe/vol14/iss2/4>
- Matthews, S. A., & Yang, T.-C. (2013). Spatial Polygamy and Contextual Exposures (SPACES): Promoting Activity Space Approaches in Research on Place And Health. *American Behavioral Scientist*, 57(8), 1057–1081. <https://doi.org/10.1177/0002764213487345>
- Mayhew, S. (2015). *A Dictionary of Geography*. Oxford University Press. <https://www.oxfordreference.com/view/10.1093/acref/9780199680856.001.0001/acref-9780199680856-e-21>
- Merry, K., & Bettinger, P. (2019). Smartphone GPS accuracy study in an urban environment. *PloS one*, 14(7), e0219890. <https://doi.org/10.1371/journal.pone.0219890>
- Middleton, S., Kordopatis-Zilos, G., Papadopoulos, S., & Kompatsiaris, Y. (2018). Location Extraction from Social Media. *ACM transactions on information systems*, 36(4), 1–27. <https://doi.org/10.1145/3202662>
- Milusheva, S. (2020). Managing the spread of disease with mobile phone data. *Journal of Development Economics*, 147, 102559. <https://doi.org/10.1016/j.jdeveco.2020.102559>
- Moayeri, N., Li, C., & Shi, L. (2019). Indoor Localization Accuracy of Major Smartphone Location Apps. - 2019 *IEEE Wireless Communications and Networking Conference (WCNC)*, 1–8. <https://doi.org/10.1109/WCNC.2019.8885639>
- Momen, N., Hatamian, M., & Fritsch, L. (2019). Did App Privacy Improve after the GDPR? *IEEE Security and Privacy*, 17(6), 10–20. <https://doi.org/10.1109/MSEC.2019.2938445>
- Oosterlinck, D., Benoit, D. F., Baecke, P., & Van de Weghe, N. (2017). Bluetooth tracking of humans in an indoor environment: An application to shopping mall visits. *Applied Geography*, 78, 55–65. <https://doi.org/https://doi.org/10.1016/j.apgeog.2016.11.005>
- Piazza, A., Rendine, S., Zei, G., Moroni, A., & Cavalli-Sforza, L. (1987). Migration rates of human populations from surname distributions. *Nature*, 329(6141), 714–716. <https://doi.org/10.1038/329714a0>

- Pinelli, F., Lorenzo, G. Di, & Calabrese, F. (2015). Comparing Urban Sensing Applications Using Event and Network-Driven Mobile Phone Location Data. - *2015 16th IEEE International Conference on Mobile Data Management, 1*, 219–226. <https://doi.org/10.1109/MDM.2015.33>
- Poom, A., Järv, O., Zook, M., & Toivonen, T. (2020). COVID-19 is spatial: Ensuring that mobile Big Data is used for social good. *Big Data & Society*, *7*(2), 2053951720952088. <https://doi.org/10.1177/2053951720952088>
- Robustelli, U., Baiocchi, V., & Pugliano, G. (2019). Assessment of dual frequency GNSS observations from a Xiaomi Mi 8 android smartphone and positioning performance analysis. *Electronics (Switzerland)*, *8*(1). <https://doi.org/10.3390/electronics8010091>
- Ruktanonchai, N. W., Ruktanonchai, C. W., Floyd, J. R., & Tatem, A. J. (2018). Using Google Location History data to quantify fine-scale human mobility. *International Journal of Health Geographics*, *17*(1), 28. <https://doi.org/10.1186/s12942-018-0150-z>
- Sagioglu, S., & Sinanc, D. (2013). Big data: A review. - *2013 International Conference on Collaboration Technologies and Systems (CTS)*, 42–47. <https://doi.org/10.1109/CTS.2013.6567202>
- Santamaria, C., Sermi, F., Spyrtos, S., Iacus, S. M., Annunziato, A., Tarchi, D., & Vespe, M. (2020). Measuring the impact of COVID-19 confinement measures on human mobility using mobile positioning data. A European regional analysis. *Safety Science*, *132*, 104925. <https://doi.org/10.1016/j.ssci.2020.104925>
- Shafeeque, S., Meedin, G., & Ratnayake, H. (2019). Locating the Position of a Cell Phone User Using GSM Signals. Teoksessa J. Hemanth, T. Silva, & A. Karunananda (Toim.), *Artificial Intelligence* (ss. 49–63). Springer Singapore. <https://docplayer.net/192736676-Locating-the-position-of-a-cell-phone-user-using-gsm-signals.html>
- Silm, S., Mooses, V., Puura, A., Masso, A., Tominga, A., & Saluveer, E. (2021). The Relationship between Ethno-Linguistic Composition of Social Networks and Activity Space: A Study Using Mobile Phone Data. *Social Inclusion*, *9*(2), 192–207. <https://doi.org/10.17645/SI.V9I2.3839>
- Sinton, D. (1978). The inherent structure of information as a constraint to analysis mapped thematic data as a case study. Teoksessa G. Dutton (Toim.), *Harvard Papers on Geographic Information Systems* (Vsk. 6). Addison-Wesley.
- Smolak, K., Rohm, W., Knop, K., & Sila-Nowicka, K. (2020). Population mobility modelling for mobility data simulation. *Computers Environment and Urban Systems*, *84*, 101526. <https://doi.org/10.1016/j.compenvurbsys.2020.101526>
- Sottile, E., Giacchetti, T., Tuveri, G., Piras, F., Calli, D., Concas, V., Zamberlan, L., Meloni, I., & Carrese, S. (2021). An innovative GPS smartphone based strategy for university mobility management: A case study at the University of RomaTre, Italy. *Research in Transportation Economics*, *85*, 100926. <https://doi.org/10.1016/j.retrec.2020.100926>
- Telia Company. (ei pvm.). *Telia Crown Insights - Henkilötietojen käsittely*. Noudettu

17. joulukuuta 2021, osoitteesta
<https://www.telia.fi/yrityksille/palvelut/teknologiat/crowd-insights/henkilötietojen-kasittely>
- Telia Company, & Grön, N. (ei pvm.). *Yksityisyys edellä*. Noudettu 17. joulukuuta 2021, osoitteesta <https://www.telia.fi/yrityksille/artikkelit/artikkeli/yksityisyys-edella>
- Teng, W.-G., & Chou, M.-C. (2007). Mining communities of acquainted mobile users on call detail records. *Proceedings of the 2007 ACM symposium on applied computing*, 957–958. <https://doi.org/10.1145/1244002.1244212>
- Tenkanen, H., Di Minin, E., Heikinheimo, V., Hausmann, A., Herbst, M., Kajala, L., & Toivonen, T. (2017). Instagram, Flickr, or Twitter: Assessing the usability of social media data for visitor monitoring in protected areas. *Scientific Reports*, 7(1). <https://doi.org/10.1038/s41598-017-18007-4>
- Thums, M., Fernández-Gracia, J., Sequeira, A. M. M., Eguíluz, V. M., Duarte, C. M., & Meekan, M. G. (2018). How Big Data Fast Tracked Human Mobility Research and the Lessons for Animal Movement Ecology. *Frontiers in Marine Science*, 5, 21. <https://www.frontiersin.org/article/10.3389/fmars.2018.00021>
- Toivonen, T., Heikinheimo, V., Fink, C., Hausmann, A., Hiippala, T., Järv, O., Tenkanen, H., & Di Minin, E. (2019). Social media data for conservation science: A methodological overview. *Biological Conservation*, 233, 298–315. <https://doi.org/10.1016/j.biocon.2019.01.023>
- Veregin, H. (1999). Data quality parameters. Teoksessa *Geographical information systems 1* (ss. 177–189).
- Veregin, H., & Hunter, G. (2000). Unit 100 - Data Quality Measurement and Assessment. Teoksessa M. Goodchild & K. Kemp (Toim.), *Core Curriculum in Geographic Information Science* (ss. 1–20). UC Santa Barbara: National Center for Geographic Information and Analysis. <https://escholarship.org/uc/item/2p0262jp#author>
- Weiß, J.-P., Esdar, M., & Hübner, U. (2021). Analyzing the Essential Attributes of Nationally Issued COVID-19 Contact Tracing Apps: Open-Source Intelligence Approach and Content Analysis. *JMIR Mhealth Uhealth*, 9(3). <https://doi.org/10.2196/27232>
- Weller, K., & Kinder-Kurlanda, K. E. (2016). A Manifesto for Data Sharing in Social Media Research. *WebSci '16: Proceedings of the 8th ACM Conference on Web Science*, 166–172. <https://doi.org/10.1145/2908131.2908172>
- Willberg, E., Järv, O., Väisänen, T., Toivonen, T., & Manley, E. (2021). *Escaping from Cities during the COVID-19 Crisis: Using Mobile Phone Data to Trace Mobility in Finland*. <https://doi.org/10.3390/ijgi10020103>
- X-Mode. (ei pvm.). *X-Mode Offers Premium Data for COVID-19 Research*. Noudettu 15. joulukuuta 2021, osoitteesta <https://xmode.io/covid-19/>
- Xie, J., Song, Z., Li, Y., & Ma, Z. (2018). *Mobile big data analysis with machine learning*. <https://arxiv.org/abs/1808.00803>
- Xu, F., Tu, Z., Li, Y., Zhang, P., Fu, X., & Jin, D. (2017). Trajectory recovery from

- Ash: User privacy is not preserved in aggregated mobility data. *26th International World Wide Web Conference, WWW 2017*, 1241–1250. <https://doi.org/10.1145/3038912.3052620>
- Yang, C., & Shao, H. (2015). WiFi-Based Indoor Positioning. *IEEE Communications Magazine*, *53*(3), 150–157. <https://doi.org/10.1109/MCOM.2015.7060497>
- Yang, P., Cao, Z., Dong, X., & Zia, T. (2011). An Efficient Privacy Preserving Data Aggregation Scheme with Constant Communication Overheads for Wireless Sensor Networks. *IEEE Communications Letters*, *15*(11), 1205–1207. <https://doi.org/10.1109/LCOMM.2011.092911.111598>
- Yazti, D. Z., & Krishnaswamy, S. (2014). Mobile Big Data Analytics: Research, Practice, and Opportunities. *IEEE 15th International Conference on Mobile Data Management*, *1*, 1–2. <https://doi.org/10.1109/MDM.2014.73>
- Zhang, S., Yang, Y., Zhen, F., & Lobsang, T. (2020). Exploring Temporal Activity Patterns of Urban Areas Using Aggregated Network-driven Mobile Phone Data: A Case Study of Wuhu, China. *Chinese Geographical Science*, *30*(4), 695–709. <https://doi.org/10.1007/s11769-020-1130-3>
- Zhang, Y., Chen, Q., & Zhong, S. (2016). Privacy-Preserving Data Aggregation in Mobile Phone Sensing. *IEEE Transactions on Information Forensics and Security*, *11*(5), 980–992. <https://doi.org/10.1109/TIFS.2016.2515513>
- Zhao, Z., Shaw, S.-L., Xu, Y., Lu, F., Chen, J., & Yin, L. (2016). Understanding the bias of call detail records in human mobility research. *International Journal of Geographical Information Science*, *30*(9), 1738–1762. <https://doi.org/10.1080/13658816.2015.1137298>