Master's thesis in Geography

Geoinformatics

Twitter as an Indicator of Sports Activities in the Helsinki Metropolitan Area

Sonja Koivisto

2021

Supervisors: Petteri Muukkonen and Pengyuan Liu

Master's Programme in Geography

Faculty of Science

Tiedekunta – Fakultet – Faculty	Osasto – Institution – Department
Faculty of Science	Department of Geosciences and Geography

Tekijä – Författare – Author

Sonja Koivisto

Tutkielman otsikko – Avhandlings titel – Title of thesis

Twitter as an indicator of sports activities in the Helsinki Metropolitan Area

Koulutusohjelma ja opintosuunta – Utbildningsprogram och studieinriktning – Programme and study track

Master's programme in geography, Geoinformatics

Tutkielman taso – Avhandlings nivå – Level of the thesis	Aika – Datum – Date	Sivumäärä – Sidoantal – Number of pages
Master's thesis, 30 credits	9.8.2021	86 + 15 appendixes

Tiivistelmä – Referat – Abstract

Being physically active is one of the key aspects of health. Thus, equal opportunities for exercising in different places is one important factor of environmental justice and segregation prevention. Currently, there are no openly available scientific studies about actual physical activities in different parts of the Helsinki Metropolitan Area other than sports barometers. In the lack of comprehensive official data sources, user-generated data, like social media, may be used as a proxy for measuring the levels and geographical distribution of sports activities. In this thesis, I aim to assess 1) how Twitter tweets could be used as an indicator of sports activities, 2) how the sports tweets are distributed spatially and 3) which socio-economic factors can predict the number of sports tweets.

For recognizing the tweets related to sports, out of 38.5 million tweets, I used Named Entity Matching with a list of sports-related keywords in Finnish, English and Estonian. Due to the spatial nature of my study, I needed tweets that contain a geotag, meaning that the tweet is attached to coordinates that indicate a location. However, only about 1% of tweets contain a geotag, and since 2019 Twitter doesn't support precise geotagging anymore with some exceptions. Therefore, I implemented geoparsing methods to search for location names in the text and transform them to coordinates if the mentioned place was within the study area. After that, I aggregated the posts to postal code areas and used statistical and spatial methods to measure spatial autocorrelation and correlation with different socio-economic variables to examine the spatial patterns and socio-economic factors that affect the tweeting about sports.

My results show that the sports tweets are concentrated mainly in the center of Helsinki, where the population is also concentrated. The distribution of the sports tweets exhibits local clusters like Tapiola, Leppävaara, Tikkurila and Pasila besides the largest cluster in the center of Helsinki. Sports-wise mapping of the tweets reveals that for example racket sport and skiing tweets are heavily concentrated around the corresponding facilities. Statistical analyses indicate that the number of tweets per inhabitant does not correlate with the education level or the amount of average income in the postal code area. The factors that predict the number of tweets per inhabitant are number of sports facilities per inhabitant, employment, and percentage of children (0-14 years old) in the postal code area. Keys to a successful study when analyzing Twitter data are geoparsing, having enough data, and a good language model to process it. Despite the promising results of this study, Twitter as indicator of physical activity should be studied more to better understand the kind of bias it inherently has before basing real-life decisions on Twitter research.

Avainsanat – Nyckelord – Keywords

Content analysis, GIS, geoinformatics, geoparsing, social media, spatial accessibility, sports, Twitter

Säilytyspaikka – Förvaringställe – Where deposited

University of Helsinki electronic theses library E-thesis/HELDA

Muita tietoja – Övriga uppgifter – Additional information

This thesis is made under YLLI-research project (Equality in suburban physical activity environments)

Tiedekunta – Fakultet – Faculty	Osasto – Institution – Department
Matemaattis-luonnontieteellinen tiedekunta	Geotieteiden ja maantieteen osasto

Tekijä – Författare – Author

Sonja Koivisto

Tutkielman otsikko – Avhandlings titel – Title of thesis

Twitter liikunta- ja urheiluaktiivisuuden indikaattorina pääkaupunkiseudulla

Koulutusohjelma ja opintosuunta – Utbildningsprogram och studieinriktning – Programme and study track

Maantiede, geoinformatiikka

Tutkielman taso – Avhandlings nivå – Level of the thesis	Aika – Datum – Date	Sivumäärä – Sidoantal – Number of pages
Maisterin tutkielma, 30 opintopistettä	9.8.2021	86 + 15 liitteitä

Tiivistelmä – Referat – Abstract

Fyysinen aktiivisuus vaikuttaa vahvasti yksilön terveyteen ja hyvinvointiin. Alueellisen eriytymisen ehkäisyn ja ympäristöllisen tasa-arvon kannalta on tärkeää, että eri alueiden asukkailla on yhtäläiset mahdollisuudet harrastaa liikuntaa. Avoimesti saatavilla olevia kattavia tutkimuksia ihmisten fyysisestä aktiivisuudesta eri puolilla pääkaupunkiseutua ei juurikaan ole tehty, paikallisia liikuntabarometrejä lukuun ottamatta. Virallisten ja kattavien tietolähteiden puutteessa käyttäjien itse tuottamaa dataa, kuten sosiaalisen median dataa, voidaan mahdollisesti käyttää fyysisen aktiivisuuden arviointiin. Tässä tutkielmassa pyrin vastaamaan kysymyksiin: 1) kuinka Twitter-dataa voidaan käyttää indikaattorina liikunnallisen aktiivisuuden arviointiin, 2) miten liikunta-aiheistet twiitit ovat jakautuneet pääkaupunkiseudulla ja 3) mitkä sosio-ekonomiset tekijät selittävät twiittien lukumäärää alueella.

Liikunta-aiheisten twiittien keräämiseen hyödynsin hakua urheiluun ja liikuntaan liittyvien avainsanalistojen avulla. Haetut avainsanat sisälsivät suomen-, englannin- ja vironkielisiä termejä. Tutkimuksen alueellisen luonteen takia tarvitsin geotägättyjä twiittejä, joihin on liitetty tieto paikan koordinaateista. Vain alle 1 % twiiteistä sisältää geotägin, joten hyödynsin geoparsing-tekniikkaa tuottaakseni lisää paikkaan sidottua aineistoa. Geoparsing tarkoittaa paikan nimien tunnistamista tekstistä ja niiden muuttamista koordinaateiksi. Yhdistin geotägätyt ja geoparsing-tekniikalla sijoitetut twiitit ja ryhmitin datan postinumeroalueittain. Postinumeroalueittain ryhmitetystä datasta tein spatiaalisia ja tilastollisia analyysejä mitatakseni spatiaalista autokorrelaatiota sekä korrelaatiota eri sosio-ekonomisten muuttujien kanssa.

Tulokseni osoittavat, että urheilu- ja liikunta-aiheiset twiitit keskittyvät pääasiassa Helsingin keskustaan, mihin myös väestö on keskittynyt. Helsingin keskustan lisäksi on nähtävissä paikallisempia klustereita Tapiolassa, Leppävaarassa, Tikkurilassa ja Pasilassa. Twiittien urheilulajittainen tarkastelu paljastaa mailapeli- ja hiihtotwiittien keskittyneen voimakkaasti vastaavien urheilupaikkojen ympärille. Tilastoanalyysit osoittavat, että postinumeroalueen tuloilla ja koulutustasolla ei ole yhteyttä alueella havaittuun urheilutwiittien määrään. Parhaiten urheilutwiittien määrää ennustaa liikuntapaikkojen määrä, työllisyystaso ja lasten (0–14-vuotiaat) osuus väestöstä. Avaimia onnistuneeseen vastaavaan Twitter-tutkimukseen ovat geoparsing, riittävä datan määrä ja tarpeeksi hyvä kielimalli. Tämän tutkimuksen lupaavista tuloksista huolimatta Twitteriä fyysisen aktiivisuuden indikaattorina tulee tutkia lisää kartoittamalla tarkemmin sosiaalisen median sisäsyntyisiä vinoumia ennen kuin Twitter-tutkimusten tuloksia voidaan soveltaa oikean elämän ratkaisuihin.

Avainsanat – Nyckelord – Keywords

Sisällönanalyysi, GIS, geoinformatiikka, geoparsing, sosiaalinen media, maantieteellinen saavutettavuus, urheilu, liikunta, Twitter

Säilytyspaikka – Förvaringställe – Where deposited

Helsingin yliopiston verkkokirjasto E-thesis/HELDA

Muita tietoja – Övriga uppgifter – Additional information

Tämä tutkielma on tehty yhteistyössä YLLI-tutkimusprojektin (Yhdenvertainen liikunnallinen lähiö) kanssa

Table of Contents

List of figures

List of Tables

List of abbreviations

AIC	Aikeke Information Critetion
AICc	Corrected Aikeke Information Critetion
API	Application Programming Interface
COVID-19	Coronavirus disease 2019
GDPR	General Data Protection Regulation
DGL	Digital Geography Lab, University of Helsinki
LIPAS	Liikuntapaikat.fi, Finnish online map service for sports facilities
LISA	Local Indicators of Spatial Autocorrelation
MAUP	Modifiable Areal Unit Problem
NEM	Named Entity Matching
NER	Named Entity Recognition
NLP	Natural Language Processing
OLS	Ordinary Least Squares
PAAVO	Postal Code Area Statistics by Statistics Finland
TPSN	Territory, Place, Scale, Network (framework)
VIF	Variance Inflation Factor
WFS	Web Feature Services
WMS	Web Map Services
YLLI- project	Equality in suburban physical activity environments- research project

1. Introduction

Many studies have concluded that the spatial accessibility of sports facilities plays a significant role in maintaining an active lifestyle (e.g., Asefi & Ghanbarpour Nosrati, 2020; Kajosaari & Laatikainen, 2020). Having an active lifestyle has positive impacts on both individuals and the society. Active lifestyle reduces the risk of having obesity-related diseases and therefore increases the length and quality of individual's life, as well as reduces the health care cost of the society and increases productivity by longer careers with less sick leaves (Karusisi et al., 2013; Vasankari et al., 2018). As sedentary lifestyle is becoming more prevalent, the rates of obesity are rising and inactivity is listed as the fourth most common cause of death globally (Borodulin et al., 2016; Kohl et al., 2012; Lundqvist et al., 2018). The Finnish government has acknowledged the importance of the topic by setting an agenda for encouraging physical activities in the government programme (Finnish Government, 2019).

Despite the importance of the topic, up to my knowledge, there are no openly available scientific studies about actual physical activities in different parts of the Helsinki Metropolitan Area, other than sports barometers. Only a few sports facilities collect official visitation statistics (mainly swimming halls and national parks), so it can be a challenge to estimate and compare the overall physical activity rate in different neighborhoods. User-generated data, like social media data, could be used as a proxy for physical activity levels around the city, due to the lack of first-hand data (Roberts et al., 2017). When using social media data, it is key to acknowledge a certain bias that comes from the fact that social media users do not represent the population as a whole and share only selected things in social media (Graham et al., 2014).

Using social media data as an indicator of physical activities has been previously studied by for example Roberts et al. (2017). In the Helsinki area, Twitter and other social media data has been used to assess the use of green space by Heikinheimo et al. (2020). I will continue along the same lines, and study **1) whether Twitter data is suitable for assessing physical activities in the Helsinki Metropolitan Area and what kind of methods could be used**. Furthermore, I will **2) examine what kind of physical activity patterns Twitter data reveals**. Where are the hotspots for sports-related tweets generally and are there any sport-specific hotspots? From where do people tweet most about running, for example? To see behind the numbers and

distribution, I will also **3) investigate the relationship of sports-related tweets to socio-economic factors of neighborhoods**. Which socio-economic factors impact the number of sports-related tweets in an area and why?

To collect the sports-related tweets, I will use Named Entity Matching (NEM) to see if the words in the tweet match to any of the sports-related keywords. Due to the spatial nature of this study, I will collect geo-referenced tweets of two kinds: geotagged and geoparsed. Geotagged tweets have location attached to them by users and geoparsed tweets contain a place name which is converted to coordinates by the geoparsing process (see detailed introduction in section 4.2.). Further analyses I will perform with the combination of sports-related geotagged and geoparsed tweets.

It is hard to assess whether the sports-related tweets give a realistic representation of physical activity because there is no official data to validate it with. Therefore, I will conduct a Facebook survey about physical activity and social media to use as validation data. The survey is carried out in a social media platform and therefore it does not eliminate the bias of social media users. However, Facebook has a larger user base than Twitter throughout different age groups, which means the results may be more representative of the entire population (Kohvakka & Saarenmaa, 2019).

This thesis is conducted as a part of YLLI- project (short for Yhdenvertainen liikunnallinen lähiö), Equality in suburban physical activity environments in English. YLLI project aims to prevent segregation between neighbourhoods and encourage equal chances for physical activities in different parts of the cities Helsinki and Jyväskylä (University of Jyväskylä, 2020).

2. Background

The environment plays a key role when it comes to healthy and active lifestyles. This spatial context of the study setting calls for a framework that broadly accounts for the spatial nature of the phenomenon. Therefore, I have chosen to utilize the Territory, Place, Scale, Network (TPSN) framework for understanding socio-spatial multidimensionalism, introduced by Jessop et al.

(2008). Jessop et al. (2008) argue that most socio-spatial research considers only one of these dimensions and therefore creates the misleading idea of one-dimensionalism. In reality, all these aspects are present, and they can only be separated in theory. In this framework, territory is defined as the dimension that creates inside-outside divides and constructs borders that define how the space is organized and governed. In this study, the territories are municipalities and postal code areas. The second dimension, place, is defined as particular and singular areal differentiations that create identities. In this context, places could be sports facilities and other physical activity environments where sports and physical activities can be practiced. Scale refers to the hierarchization of spatial structures and social relationships. Scales are visible in my study as the different levels of access to sports facilities and usage of social media platforms by professional athletes, coaches of sports clubs, members of sports clubs or regular individuals. The fourth dimension, network, is conceived as interconnectivity and interdependence. In the context of this study, network may be understood as the socio-spatial networks of people, places, and activities.

2.1. Promoting healthy lifestyles

The importance of healthy lifestyles is recognized in Sustainable Development Goals as the third goal to "*Ensure healthy lives and promote well-being for all at all ages*" (UN, n.d.). The equality and sustainability aspects come into play in the eleventh goal to "*Make cities and human settlements inclusive, safe, resilient and sustainable*" (UN, n.d.). As this may seem to concern more developing countries, it is a good reminder that wellbeing and equitable access to facilities supporting wellbeing are very fundamental in building healthy and sustainable societies. Good nutrition, exercising habits and access to healthcare are all building blocks of a healthy lifestyle. During the 21st century, inactivity has become the fourth most common cause for death globally (Kohl et al., 2012). The accessibility to sports facilities and to other physical activity environments has been proven to have an influence in physical activity levels and thus affect the broader wellbeing of population (Asefi & Ghanbarpour Nosrati, 2020; Kajosaari & Laatikainen, 2020).

2.1.1. Importance of promoting healthy lifestyles in Finland

Sedentary lifestyle and non-physical work are on the rise globally, and Finland is no exception to this (Borodulin et al., 2016). While physically demanding work is declining, exercising during one's free time is on the rise but so is the percentage of overweight people (Borodulin et al., 2016; Helldán & Helakorpi, 2015). However, over 70% of adults reported that they exercise in their free time (Borodulin et al., 2018). Most popular sports amongst adults are walking, biking, gym, and running (Kuntoliikuntaliitto, 2010). Nonetheless, only half of the adult population (30+ years) reach the activity goal of 2.5 hours of moderate or 1h 15mins of heavy cardio training weekly (Borodulin et al., 2018). According to Bennie et al. (2017), only 10% of Finnish adults reach the requirements of the guidelines for overall physical activity, including sufficient cardio training, muscle-strengthening, stretching and balance exercises (Bennie et al., 2017).

Consequently, combined with excess calorie intake, a large proportion of the population is gaining weight. According to FinTerveys research (Lundqvist et al., 2018), 75% of Finnish men and 66% of women are overweight (BMI > 25) and 25% of both are obese (BMI > 30). Prevention of obesity is key since it is hard to permanently lose weight once it has been gained (Lundqvist et al., 2018). Close attention should be paid to children's increasing levels of physical inactivity, screen time and weight gain (Jääskeläinen et al., 2020).

Physical inactivity and sedentary lifestyle increase the risk of cardiovascular diseases, some cancers and type II diabetes (Lavie et al., 2019). Studies also recognize a link between physical inactivity and depression (Leavitt, 2008). These diseases bear a cost to society as they reduce productivity, increase need for sick leave and health care. The cost estimates of physical inactivity cover a wide range. The UKK-institute (the Research and Expert Center of Health and Exercise) claims the annual cost of inactivity in Finland to be between 3.2 and 7.5 billion euros (2018). This includes direct health care costs (0,6 billion), loss of income tax (1.4 – 2.8 billion), productivity costs (0.9 – 3.8 billion) and many other costs (Vasankari et al., 2018). Consequently, the prevention of inactivity could save significant costs from both the government and individuals, increase the productivity of society, and improve the wellbeing of individuals by preventing illnesses.

The current government programme by prime minister Sanna Marin has identified three objectives related to physical activity and sports (Finnish Government, 2019). Under section 3.7.1. objective 6 states that "*A physically more active lifestyle will be encouraged for all population groups*" (Finnish Government, 2019). The implementation measures include setting up a physical activity programme and its evaluation unit, sports policy coordination unit and launching "Finland on the Move" programme. Objective 7 promises that "Conditions for outdoor and daily activity will improve" by creating more neighborhood outdoor places and a national strategy for recreational use of nature (Finnish Government, 2019). Objective 8 aims to improve conditions for club and elite sports by providing more funding and focusing on inclusivity and gender equality aspects (Finnish Government, 2019). Having these objectives in the government's programme shows that the importance of the topic is recognized on a national level, and it is relevant in today's Finland.

2.1.2. How can the urban environment encourage physical activities?

European Healthy Cities programme by WHO recognizes that healthy choices and active lifestyle can be encouraged by interventions to the built environment (de Leeuw, 2017). These interventions include investing in sports facilities, building adequate cycling infrastructure and converting roads to pedestrian streets. European Healthy Cities project aims to construct cities that are physical, social and cultural environments which enable and drive health and well-being for all (de Leeuw, 2017).

As mentioned previously, urban dwellers use public open and green spaces for working out (Chacón-Borrego et al., 2018; Kajosaari & Laatikainen, 2020). It would be beneficial to create such urban spaces that encourage and enable active lifestyle and provide many places for exercising. An investment in creating such cities would pay itself back as reduction of the high costs of inactivity (see section 2.1.1.) provided that previously inactive people were also encouraged for more active lifestyle.

For example, favoring bicycles and pedestrians in traffic and constructing safe routes for them can result in an increase in physical activity and simultaneously reduce the traffic emissions (Haustein et al., 2020; McCormack & Shiell, 2011). Generally, reducing barriers and increasing

walkability in urban environments result in higher physical activity levels (Gharaveis, 2020; McCormack & Shiell, 2011). According to a study conducted by the City of Helsinki, 39% of trips were done by walking and 9% by biking in 2019. Walking is the most popular mode of transport, and its popularity has been increasing since 2015, which may be credited to the walking promotion programme (City of Helsinki, 2020a; Norppa, 2020). The share of biking trips has been steadily around 10% since 2010 despite growing investments to the biking infrastructure and implementation of a common bike sharing system for Helsinki and Espoo (City of Helsinki, 2020b). According to the biking promotion programme, Helsinki wants to double the share of trips made by bike by 2035 and has made investments of 20 million euros for biking infrastructure in 2020 to achieve that (City of Helsinki, 2020b). It has been estimated that every euro invested in biking will produce worth of eightfold benefits in the form of time savings and reduced general health cost (Helsingin kaupunkisuunnitteluvirasto, 2014).

Researchers have found that higher exposure to green areas correlates with better mental and physical health (Cole et al., 2017; D'Alessandro et al., 2015). In a study about jogging environments in Paris, Karusisi et al. (2012) found that people are 29% more likely to jog in environments perceived as pleasant, like in proximity to lakes and parks. Especially in densely built large cities like Paris, green areas might be even greater attraction than in medium-sized coastal cities like Helsinki. While studying the activity spaces in Helsinki, Hasanzadeh et al. (2021) found that 86% of trips by young adults (25–40 years) done outside of their own neighborhood are to areas that are greener than their own neighborhood (Lilius et al., 2021). This would indicate that green areas are a strong attraction, but this is not a direct indicator that sports and physical activities are done in the green areas.

Besides physical environment, social environment is another factor affecting physical activity. High social cohesion of a neighborhood increases the likelihood of jogging there rather than in the surrounding neighborhoods (Karusisi et al., 2012). Overall, green environments and open spaces encourage people to exercise in the neighborhood (Kajosaari & Laatikainen, 2020; Karusisi et al., 2012). Hence, it would be important that different neighborhoods of the Metropolitan Area have similar access to sports facilities and greenery and consequently, the inhabitants have equal chances to lead an active and healthy life. This concept is called spatial equity which, according to Talen and Anselin (1998), requires the consideration of needs, fairness and justice in the distribution of public goods and services (Talen & Anselin, 1998).

2.1.3. Predictors of physical activity

There has been little research about spatial issues in the physical condition of people – especially in different parts of Finland, let alone the Metropolitan area. In a research addressing health behavior, Helldán and Helakorpi (2015) found that the socio-economic background of a person is more relevant to the habits in physical activity than home location. Factors that typically predict high physical activity in adult age are competing in sports during childhood, positive experience with physical education during school years and having sports as a hobby (Mäkinen, 2011). In France, Karusisi et al. (2013) have found that highly educated people might give higher priority to health and therefore exercise more because they are more aware of the health benefits of exercising. Meanwhile, low education level or professional status of one's parents predicted low activity levels in the adult age. For men, physically heavy work and for women mentally stressful jobs reduce the amount of free time exercising (Mäkinen, 2011). The differences in physical activity between Finnish provinces are small but in the comparison of overweight people, Uusimaa, where the Helsinki region is located, stands out with lower values (Helldán & Helakorpi, 2015). This might be explained by the generally higher level of education in the region.

The time used for sports varies according to the population and professional group. Managers and people in leading positions use the most amount of time for sports per week (4h 22min) while farmers (2h 4min) and stay-at-home parents (2h 50min) use the least (Suomi et al., 2012). According to the same study, 70% of men and 63% of women feel that they can do as much sports as they want. The most common barriers for doing sports are lack of time, illness or injury, work or studies and life situation. About 10% of respondents name long distances and high expenses as barriers for doing sports, but trend has been declining (Suomi et al., 2012).

Recently, there has been an increasing amount of research pointing out the connection between the physical activity levels of population and the surrounding environment. Accessibility to sports facilities, urban green spaces and other open public spaces where exercising is possible is an important factor encouraging for more active lifestyle (eg. Asefi & Ghanbarpour Nosrati, 2020; Gharaveis, 2020; Kajosaari & Laatikainen, 2020).

2.2. Equal spatial accessibility to sports facilities and physical activity environments

2.2.1. How does spatial accessibility to sports facilities affect physical activity?

Accessibility can generally be defined as "*the ability to reach desired goods, services, activities and destinations*" (Litman, 2010, p.1). The broader concept of accessibility includes physical, cultural, social and financial aspects but spatial accessibility focuses on geographical distance and the time it takes to cover it (Karusisi et al., 2013). Many studies have addressed the relationship between spatial accessibility to sports places and the amount of sports practice (Higgs et al., 2015; Kajosaari & Laatikainen, 2020; Karusisi et al., 2013; Shrestha et al., 2019; Sterdt et al., 2014). However, the findings have been inconsistent. A review study by Sterdt et al. (2014) suggests that some of the contradicting findings can be explained with diverging methodologies and imprecisions in the studies that rely on people self-reporting activities.

The connection with frequency of physical activity and spatial accessibility of sports facilities has been found to vary between sports. This makes sense, since some physical activities can be done anywhere while some are more location specific. Karusisi et al. (2013) found the strongest positive connection between spatial accessibility and swimming compared to other sports. They speculate that this might be because swimming always requires a pool (or natural water body), but other sports can also be done elsewhere like in public parks. In team sports, the accessibility for other team members needs to be considered and come to a compromise.

Shrestha et al. (2019) conducted a study about spatial accessibility to sports facilities taken into account the selective daily mobility bias and came to the same conclusion with Karusisi et al. (2013). According to Shrestha et al. (2019), the spatial accessibility to sports places from home or work does not affect the amount of sport practice. They highlight that the entire activity space, excluding sports places, must be accounted for while assessing the spatial accessibility. This means considering the accessibility from all frequently visited places like grocery store, library and children's school instead of just assessing the spatial accessibility from home location. Then the true spatial accessibility can be counted while selective mobility bias

has been accounted for. This way, swimming is still the only sport where the proximity of the facility increases the activity (Shrestha et al., 2019).

Kajosaari and Laatikainen (2020) studied people whose physical leisure time activities are on moderate to vigorous level in Helsinki. They had divided sports places into four categories: indoor and outdoor sports facilities, public green space and built public open space. Out of these categories, built public open space is most frequently used for exercising, followed by public green space. These are often closer to people's homes than official sports facilities. According to their findings, the mean distance from home to a sports place is 3.2 kilometers. The mean distance to public spaces and green areas is shorter than the mean distance to indoor (4.7 km) or outdoor facilities (3.9 km). This indicates that physically active people travel to different parts of the city to do sports and is in line with findings of Shrestha et al. (2019) and Karusisi et al. (2013).

One interesting finding from Kajosaari and Laatikainen (2020) was that people with lower activity levels are more likely to exercise in close proximity to their homes. Thus, to activate the parts of the population with lower activity levels, it is important that sports facilities and green spaces are available in the same neighborhood. All in all, 60% of sports activities take place within 1600 m radius from home, according to Kajosaari and Laatikainen (2020). One explanation might be that physically active people choose to live closer to sports facilities (McCormack & Shiell, 2011). In their review study, Sterdt et al. (2014) found out that proximity to sports facilities increases the physical activity amongst children and adolescents. It is important to note that children and elderly might not have an easy access to a car, so the definition of accessibility varies between age and income groups.

Promoting an active and healthy lifestyle by city planning is not always straightforward. In order to have a societal effect for better health, the facilities and parks should be easily accessible to all, especially for citizens whose current activity level does not reach the recommended level. Equal access to urban green areas has been studied quite a lot from urban planning and environmental justice perspectives (e.g., Kabisch & Haase, 2014; Rigolon et al., 2018). Some greening efforts have even led to "green gentrification" where new parks and green areas have raised the housing prices in an area and therefore displaced the poorer residents and changed the social environment of a place (Blok, 2020; Cole et al., 2017; Gould & Lewis, 2012). This further emphasizes the inequities within the city and fails to provide the desired outcome

like improving equal access to green areas or providing health benefits to city dwellers in a more vulnerable position.

Many research groups in different parts of the world have studied equal access to sports facilities (Asefi & Ghanbarpour Nosrati, 2020; Billaudeau et al., 2011; Langford et al., 2018; Reimers et al., 2014; Shen et al., 2020). The results vary by location, sport, age, and gender. For example, Asefi and Nosrati found unfair access to sports facilities in Isfahan city in Iran (2020), but Billadeau et al. (2011) did not find connection between the neighborhood income and accessibility or quality of sports facilities in the greater Paris region (2011). In China, equal access to sports facilities seems to depend on the sport. Shen et al. (2020) found that in Nanning city the most equally accessible sports are those that are popular amongst all age groups and relatively cheap to build, like basketball courts and table tennis facilities. Most inequitable access was found to be to swimming pools and fitness centers (Shen et al., 2020). In Wales, the public sports facilities were more easily accessed from lower income areas and higher income areas had better access to privately-owned facilities (Higgs et al., 2015). Accessibility also depends on the transport network and mode available to the user. This might vary by age and gender. Reimers et al. (2014) found that in rural areas of Germany young women would exercise more if gyms were more accessible. Of course, this depends a lot on the regional planning strategies, the service providers, sports culture, and the market demand.

2.2.2. Accessibility to sports facilities in Finnish context

In Finland, it is municipalities' responsibility to organize sufficient sports facilities and services for the citizens according to the Sports Act (*390/*2015). The Sports Act promotes health and well-being and supports sports on all levels from competitive to hobbyist (Bergsgard et al., 2019). The sports facilities seem to be relatively accessible, taken into account that Finland is a sparsely populated country with long distances. Kotavaara and Rusanen (2016) investigated the accessibility of sports facilities in Finland and concluded that the facilities are located close to population agglomerations and thus easily accessible to most. Ball fields and indoor sports halls are the most accessible since half of the population can reach one within 1,3 kilometers and 90% of people can reach one within 5,5 kilometers. Swimming halls, ice-hockey halls and track and

field facilities are somewhat less accessible, as half of the population can reach one within 4,5 kilometers from home (Kotavaara & Rusanen, 2016).

The accessibility varies a lot between municipalities; the best accessibility is in large cities and in southern Finland where distances are smaller. Entire Finland has almost 40 000 sports facilities, of which the majority, 70%, is owned by municipalities (LIPAS, 2021). The rest are privately-owned, owned by sports clubs or by trusts. Private ownership is common in sports like tennis, horse-riding, golf and bowling, which are considered more expensive (Bergsgard et al., 2019).

In addition to the spatial measures of accessibility, it is also important to account for financial, physical, and cultural measures of accessibility. Even though a facility is spatially accessible to one, it can be financially unaffordable, physically impossible to reach in case of disability or culturally unfit. One example of the last-mentioned is swimming pools in case of religious constraints. Experiences of marginalization in sports in Finland have been recognized amongst immigrants and people with an immigrant background, disabled people, obese people and youth from peripheral areas (Armila, 2020; Eriksson et al., 2020; Harjunen, 2020; Rannikko & Armila, 2020; Seppänen et al., 2020). Good spatial accessibility does not help marginalized groups if the facilities are not accessible on other levels. So, accessibility needs to improve on all levels which may require changes in practices, atmosphere, and culture.

2.3. Applying social media data to activities' research

2.3.1. Research opportunities and challenges with use of social media data

Social media data has increasing popularity among researchers as a source of data and a study object itself. It may be characterized as one type of "big data" since it has high volume, velocity and variety (Kitchin, 2013). As social media data is readily and freely available in large quantity, using it reduces the time and financial resources needed in data collection (Roberts et al., 2017). The disadvantage is the bias of the data, since it does not represent all categories of population (Graham et al., 2014). Still, the readily available social media data has been used for

research covering a wide range of topics including health and diseases (e.g., Alotaibi et al., 2020; Bornmann et al., 2020; Osakwe et al., 2020), natural disaster management (e.g., de Bruijn et al., 2017; Middleton et al., 2014, 2018), wildlife conservation (e.g., Toivonen et al., 2019), popularity of athletes (e.g., Chmait et al., 2020), and green area use and park visitation (e.g., Hamstead et al., 2018; Heikinheimo et al., 2017, 2020). Generally, social media data provides a larger sample than what researchers would be able to collect in the scope of their research. However, the data may be more heterogenous than what would be data collected, for instance, in the context of a structured survey (Roberts et al., 2017).

Despite the large sample size, social media data has its flaws and challenges. Heikinheimo et al. (2020) identify the main challenges to be access to the data, ethical use of the data and representativeness. Access to the data can be limited and change over time since social media data is owned by private companies. Recently, many popular social media platforms like Facebook and Instagram have disabled or restricted the access to their data due to increasing privacy concerns (Mancosu & Vegetti, 2020; McCrow-Young, 2020). Twitter still keeps their data available for research purposes. However, it is only possible to gain access to a percentage of public tweets. Furthermore, how the data retrieval works internally remains unclear (Boyd & Crawford, 2012).

Because social media research is a relatively new field, laws and regulations lag behind as research and platforms advance at fast pace. Currently, all studies conducted using social media data need to comply with GDPR (General Data Protection Regulation) and abstain from giving away any individual's identity (Mancosu & Vegetti, 2020). In other words, data needs to be aggregated to general level and it needs to be kept safely. On ethical concerns pertaining to the use of social media data, the data is user-generated and the users will not be aware or give their consent that their data is used for a research (Boyd & Crawford, 2012; Lansley et al., 2020). Boyd and Crawford (2012) also question that does the public availability make the data suitable or ethical instrument of investigation and research. When posts and texts are taken out of context, they can be misunderstood and misanalysed (Boyd & Crawford, 2012; Lansley et al., 2020).

Representativeness is one challenge with social media data that I will further address in section 2.3.2. Certain ethnic and age groups are usually not present in social media platforms and the user group varies by the platform (Pew Research Center, 2019). For example, elderly people

use less internet and might not have the technology and skills required to use social media platforms. According to Statistics Finland, 31% of 75+ population use internet and only 5% use social media (2016). For those between 65 and 74, the figure for social media use is 21%, and for age group between 55 and 64 years 39%, while the younger age groups have figures close to 90%. A steady decline in social media use is seen with the older age groups (Figure 1). It can be also seen that Facebook and Instagram are the most popular platforms while much smaller portion of population uses Twitter. Since Twitter has less users, the users will not probably be as representative of the entire population as in other platforms (Kohvakka & Saarenmaa, 2019).

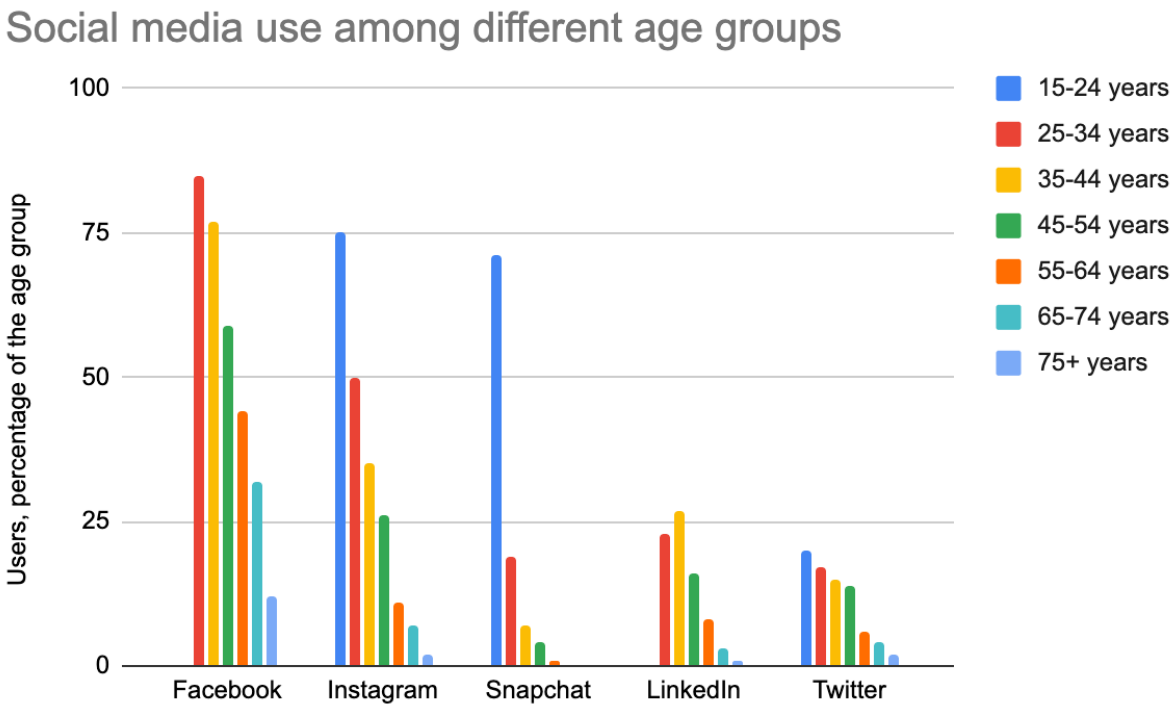


Figure 1. Social media usage quite uncommon amongst elderly (Adapted from data: Statistics Finland, 2017).

2.3.2. Challenges with data representativeness

Challenges with the representativeness of social media data warrants discussion, because many factors can make it skewed towards certain population groups. Large quantities of (social media) data is often mistakenly assumed better data just because of its huge volume. However, the size of the data is not an indicator of the validity nor of the representativeness (Boyd &

Crawford, 2012). Furthermore, social media data does not inherently contain any demographic variables. Therefore, evaluating its representativeness and suitability for research purpose can be a hard task itself (Lansley et al., 2020). Many factors affect the representativeness of the data. For example, how big percentage of population is using the platform and who are not represented there at all? For example, how large is the proportion of population using the platform, and who are not represented? Usually social media data is biased towards young, educated and technologically skilled people, leaving out older generations and economically disadvantaged people (Q. Huang & Wong, 2016).

Even if a person uses the platform, they might be left out of the data. Twitter estimates that 40% of its users are "listeners" who are not creating content but following the ongoing discussion (Twitter, 2011). Similarly, only 10% of users generate a total of 80% of the tweets in the United States (Hughes & Wojcik, 2019a). This kind of super users are more likely to be women although most of Twitter users in general are men (Hughes & Wojcik, 2019b). In addition, it is good to keep in mind that a Twitter account does not always correspond to a person. One person can have multiple accounts, multiple people can use shared account, and some accounts are companies or even bots (Boyd & Crawford, 2012).

Furthermore, many choose to protect their profile and share their tweets only with their followers, and hence they are not publicly available (Boyd & Crawford, 2012; Heikinheimo et al., 2020a). Even fewer people, about 1% of Twitter users, are willing to geotag their posts which further limits the spatial research conducted using Twitter data (Q. Huang & Wong, 2016; Sloan et al., 2013). In spatial research with Twitter data, the sample is restricted to those who openly tweet and share their location and biased towards the people who do that the most. These are some reasons why combining multiple data sources would probably bring more complete coverage and understanding of the human behavior because all the platforms have slightly different user groups (Heikinheimo et al., 2020a; Tenkanen et al., 2017).

2.3.3. Challenges with geotagging

Geotagging means attaching the geographical information of the place to a social media post. Geotags can be precise, which means that the exact coordinates of the user's location are

attached, or more coarse point-of-interests for instance a name of a city or country (Hochmair et al., 2018). In Twitter, it is also possible to use these both simultaneously. This can, however, be confusing as the point-of-interest does not have to match the coordinates. This indicates that the tweet is talking about a different place than from where it was posted. Since 2019, Twitter does not support precise geotagging anymore except for pictures, and posts shared via third party app (Hu & Wang, 2020). This is primarily to protect the user's privacy. Even before the change, less than 1% of all tweets had a precise geotag (Q. Huang & Wong, 2016; Sloan et al., 2013). Hence, using only the geotagged tweets disqualifies the majority of the posts. For all the posts, geotagging a point-of-interest like Helsinki or Finland is still supported. For neighborhood-level analysis these top-level geotags do not bring much extra value and are usually removed (Shelton et al., 2014). Besides choosing the right level of specificity for geotags, spatial precision and accuracy of the geotags are data quality issues that need to be taken into the account while doing research with user generated data.

Geotagging behavior varies with user's preferences and language (B. Huang & Carley, 2019). Therefore, it is important to acknowledge the bias i.e., who are the people who geotag their posts and how are they different from all Twitter users. Although, if the research area is linguistically homogenous, then the geotagging behavior varying by the linguistic background does not skew the results further but of course yields geotagging behavior varying by the linguistic background does not skew the results further but of course gives the best results where a larger percentage of tweets are geotagged. Twitter users who geotag their tweets are found to have different types of profiles and different tweeting behaviors than the users who do not geotag their tweets (Karami et al., 2021). Since only a fraction of the data is precisely geotagged, this fraction fails to be fully representative

However, it is estimated that approximately 10% of the tweets mention a recognizable place in the text (MacEachren et al., 2011). These location names could be used to retrieve more spatial information by geoparsing the names into coordinates (Middleton et al., 2018). On one hand, geoparsing introduces more location uncertainty as it is not known if the user is talking about their location or another place from the other side of the world (Hu & Wang, 2020). On the other hand, even if the location mentioned in the tweet is not the user's current location, it can be seen which places people are talking about. Still, commonly the place names mentioned are quite general similar to the use of the points-of-interest, e.g., Helsinki or Finland. Reverse geocoding

such general place names would just create artificial hotspots to certain central locations that has been attributed to represent the city (i.e. the central railway station) even though the real tweeting locations would be scattered around the city (Hiippala et al., 2020).

The ethics of location extraction need to be considered as well. In Twitter, geotagging is turned off by default and the user can choose to turn it on (Twitter, n.d.). In case someone has made a conscious decision to not share their location with the tweet, one must consider the implications of extracting location information from the tweet text. Still, it can be assumed that the user had no intention to hide the location if they have written it in a public tweet.

2.3.4. Dealing with linguistic diversity in social media research

One challenge in doing research with social media data is dealing with different languages and multilingualism. Language detection from short texts like tweets can be challenging as tweets might mix different languages and contain a lot of abbreviations (Carter et al., 2011; Graham et al., 2014). Automatic language detection algorithms have been developed, but still, they are not 100% accurate. Furthermore, language models for processing the data are usually not trained with social media data but with more formal type of text. Therefore, they are not as performant with social media data which may contain informal language, slang words, abbreviations, type errors and hashtags (Carter et al., 2011). For optimal performance, one should train their own language model with social media data which is very similar to the data used in the study (Middleton et al., 2018). For this thesis, training my own language model is out of the scope. Hence, I use ready-made stanza language models that have been trained with social media, blogs and emails for English (ewt) and news, blogs and fiction for Finnish (tdt) (Stanford NLP Group, 2020).

In social media and internet in general, English is a lingua franca. Meaning that many people post in English, although it would not be their first language (C. Lee, 2016). This could be because they have followers from different backgrounds and English has the widest reach of audience. In this study's dataset, English was almost as common as Finnish, although majority of people speak Finnish as their first language (see Figure 2 and Figure 3).

With respect to different languages, Hiippala et al. (2020) have studied the linguistic landscape of Twitter in Finland and they found that Helsinki and Espoo have the highest number of languages detected (71–91 languages). This observation is expected because the Helsinki Metropolitan Area also has a high number of immigrants, tourists, and business travelers on the national scale. Furthermore, it is located on the coast which is home to majority of the Swedish-speaking population and Helsinki harbor works as a gateway to Estonia (Hiippala et al., 2020).

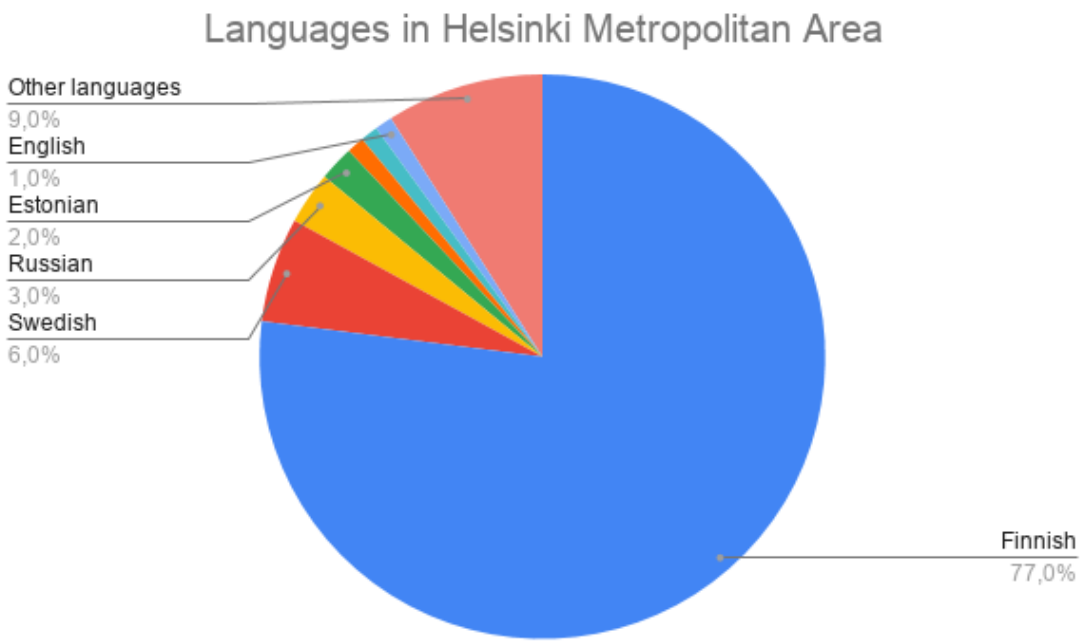


Figure 2. Inhabitants' first languages in Helsinki Metropolitan Area (Statistics Finland, 2019).

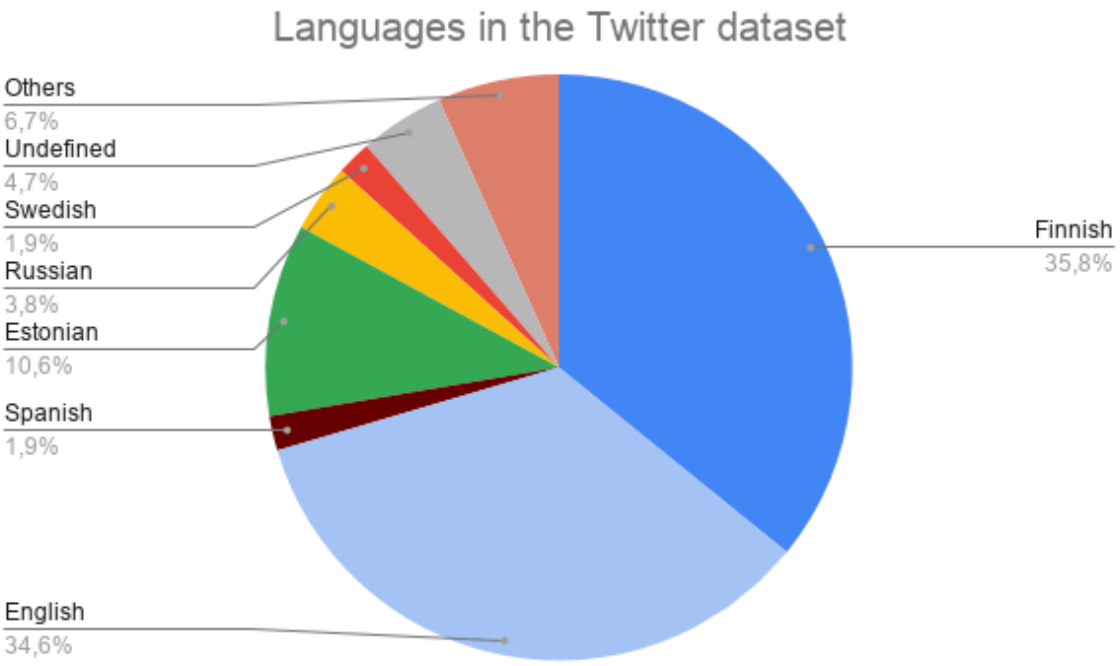


Figure 3. Languages in the Twitter dataset. Spatial coverage of the data is Finland and Estonia.

Most social media studies are conducted with comprising one language, most commonly English. Majority of the language models are unilingual or perform the best when only one language is analyzed. In multilingual studies where post content is analyzed the data is usually filtered with the same keywords in multiple languages (Sass et al., 2020), with a common hashtag (Viguria et al., 2020) or with placenames (Scholz & Jeznik, 2020). I am going to use similar approach as (Sass et al., 2020) and use keywords in three different languages for filtering out the target content. Areas with the most linguistically diverse landscape, such as areas with high percentage of people with immigrant backgrounds, might capture lower number of tweets when the smaller minority languages are left out.

2.3.5. Previous usage of Twitter data in sports and physical activity related studies

In research about sports and physical activities, Twitter data has been used for instance to identify activities undertaken in parks and green areas (Heikinheimo et al., 2020a; Roberts et al., 2017) and to assess the temporal and spatial patterns of fitness and exercising (Torres & Vaca, 2017). In the realm of sports, tweets have also been used to investigate Twitter engagement and how it affects sports events like match visitation in tennis Grand slam tournament (Chmait et al., 2020) as well as mobility of fans in sports events (Xin & MacEachren, 2020). All the above-mentioned studies about sports activities adopted only geotagged tweets. Both Heikinheimo et al. (2020) and Roberts et al. (2017) first performed a fine spatial selection to obtain only the tweets from green areas and parks and then manually categorized them by activity. Torres and Vaca (2017) had a broader spatial extent – the entire Ecuador. They chose the tweets that had been shared from sports application (e.g., Endomondo, Nike+) and then further categorized them manually by sport using hashtags.

Roberts et al. (2017) found out that over 60% of the tweets were related to an organized sports event. This brings into question that not all sports-related tweets indicate that the users would be physically active. It might as well be that the tweeter is a spectator in the sports event. This brings in inaccuracies and questions number of tweets as a qualified measure of physical activity. It would be expected that some sports stadiums and venues may have hotspots related to professional sports events that attract lots of spectators. However, this might not be an issue because, even if tweeting from a sports event, someone is doing the sport and being physically active.

From the fact that Twitter data has been successfully used for similar studies before, one could hypothesize that the data is at least somewhat suitable for this kind of study. Then again, Heikinheimo et al. (2020a) did comparison between geotagged Twitter and Instagram data in Helsinki and found out that the majority of Twitter data (74%) was reposted from Instagram. This indicates that Instagram would be a better platform for research, and it also has more monthly active users (Pew Research Center, 2019). Heikinheimo et al. (2020a) also found that Instagram had 19 times more posts than Twitter and 21 times more users. This would strongly speak for the advantage of Instagram but as their data is not available, Twitter is a more viable

option. Instagram might be more common platform for sports posts as Twitter is usually more discussion-based, evolving around topical and professional conversation whereas Instagram evolves more around free time topics, and every post contains a photo. Instagram might be the platform for posting about user's own physical activities and Twitter the platform for tweeting as a spectator in sport competitions.

3. Data

3.1. Study area

The study area, depicted in Figure 4, encompasses the Helsinki Metropolitan area including the cities of Helsinki, Espoo, Vantaa and Kauniainen. I chose this particular study area as it is the largest population center in Finland, with roughly 1 200 000 inhabitants (Statistics Finland, 2019). Therefore, Helsinki is expected to have the most sports-related Twitter posts. Although, working with a larger data set does not necessarily mean that the data would be more valid, representative or insightful, the risk of having completely random data is lower and level of aggregation also plays an essential role (Zelenkauskaite & Bucy, 2016).

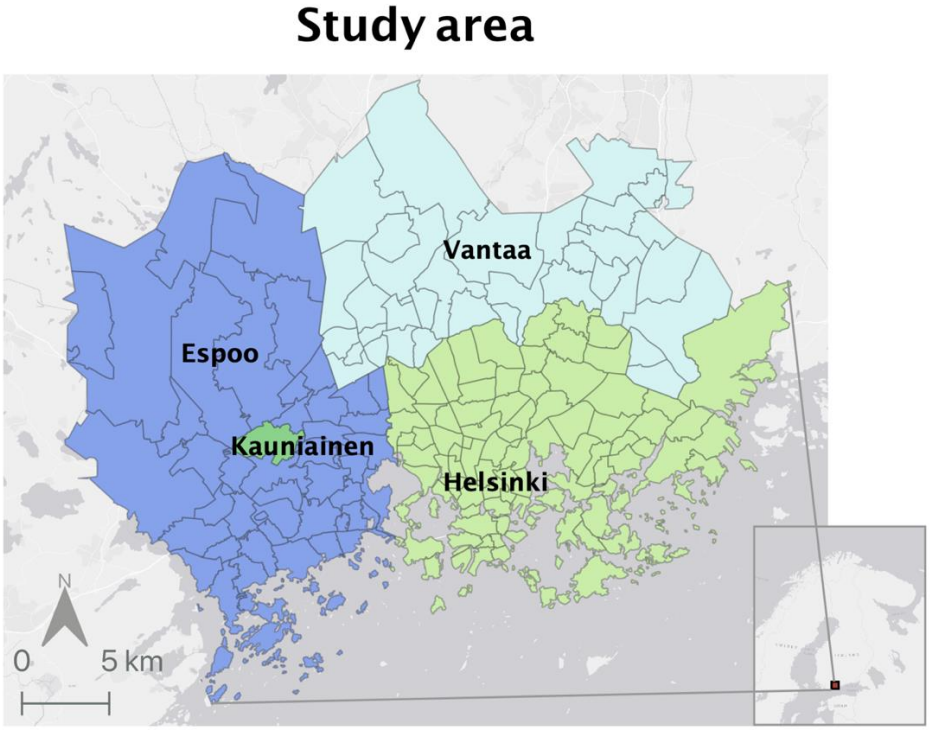


Figure 4. Map of the study area.

As the metropolitan area is the largest urban agglomeration in Finland, there are also more possibilities for segregation and a larger socio-economic gap to grow between the areas (Karhula et al., 2021; Andersen et al., 2016). I am aggregating the tweets to postal code areas (n= 168) and then comparing the disparities between those and with socio-economic indicators of the postal code areas. Postal code areas are natural units of aggregation as they follow the neighborhood borders in most cases. Another reason for my decision is the availability of open access socio-economic data by postal code areas, which makes the analysis meaningful and practical.

3.2. GIS data sets

My main data sources are Twitter posts collected by Digital Geography Lab, GeoNames gazetteer, LIPAS sports facilities database, boundaries of postal code areas, PAAVO socio-economic indicators by postal code areas and Facebook survey results, conducted in the scope of this thesis. All data is openly accessible except for the tweets and survey results. Table 1 offers an overview of the data used in this thesis; more precise data descriptions are provided in the following subchapters.

Table 1. Data used in this research

Data	Producer	Description	Spatial extent	Format	Open data
Twitter posts	User generated, collected by Digital Geography Lab	Tweets from 2006 to 2020, collected with getting user histories of people who have geotagged themselves in Finland or Estonia	Mainly Finland and Estonia, global outliers	Table in PostGIS database	No
GeoNames gazetteer	Crowd sourced	Place names and the corresponding coordinates	Global	TXT-files	Yes
LIPAS sports facilities	University of Jyväskylä	Sports facilities which are in use (points, lines, and polygons)	Finland	Available via an API in GeoJSON, WFS or WMS formats	Yes
Postal code areas	HSY in cooperation with cities of Helsinki, Vantaa, and Espoo	Postal code areas in Helsinki Metropolitan Area	Helsinki Metropolitan Area	Available in shapefile, excel, tab, MapInfo, KML, WFS and WMS	Yes

PAAVO open data by postal code areas	Statistics Finland	Socio-economic indicators such as population, education, and income	Finland	PxWeb and CSV-file	Yes
Survey results	User generated, collected by the author	Facebook questionnaire about sports activities and social media	Helsinki Metropolitan Area	CSV-file	No

3.2.1. Twitter data

Twitter is a free microblogging platform enabling sharing posts up to 280 characters, links to other websites and media content including photos, videos, and audios. Currently, Twitter has 353 million monthly active users and 500 million tweets per day globally (Statista, 2020). Twitter has been profiling itself as a platform for discussion about news, opinions, and professional matters. However, Twitter is not the biggest social media platform. In 2020, Instagram had three times more and Facebook 7.5 times more monthly active users than Twitter (Figure 5). As mentioned before, the other social media platforms might be more fit for purpose for this analysis but unfortunately, they are unavailable as they have closed their Application Programming Interfaces (APIs) lately.

The Twitter data used in this study has been collected by Digital Geography Lab (DGL), University of Helsinki. All the data are maintained in encrypted form in the Enhanced Security Research Database of the University of Helsinki. The data includes 38.5 million tweets with a spatial focus on Finland and Estonia. Out of these tweets, 2 million (5.3%) are geotagged, which is a larger percent than generally. This is due to the collection method described hereafter.

The data set was collected from May 2019 to April 2020 and covers tweets starting from year 2006 until the end of the collection period. The collection was executed by Python script with the help of tweepy Python library from the Twitter API. Spatial selection of the data was performed by first combining two global geotagged Twitter datasets (one from DGL, the other as courtesy of Matthew Zook, University of Kentucky) and retrieving unique users who have geotagged themselves in Finland or Estonia. Then, the past tweets of these users were collected with user_timeline() function from Twitter API. A maximum of 3 200 most recent tweets were collected per person as Twitter had it defined as the limit at the time. Thus, the dataset is concentrated in Finland and Estonia but contains global outliers (from user histories). In

addition, some obvious bots were cleaned from the data set by Digital Geography Lab. The spatial extent of the data also covers Estonia since it was originally collected for a research project that investigates cross border mobility between Finland and Estonia.

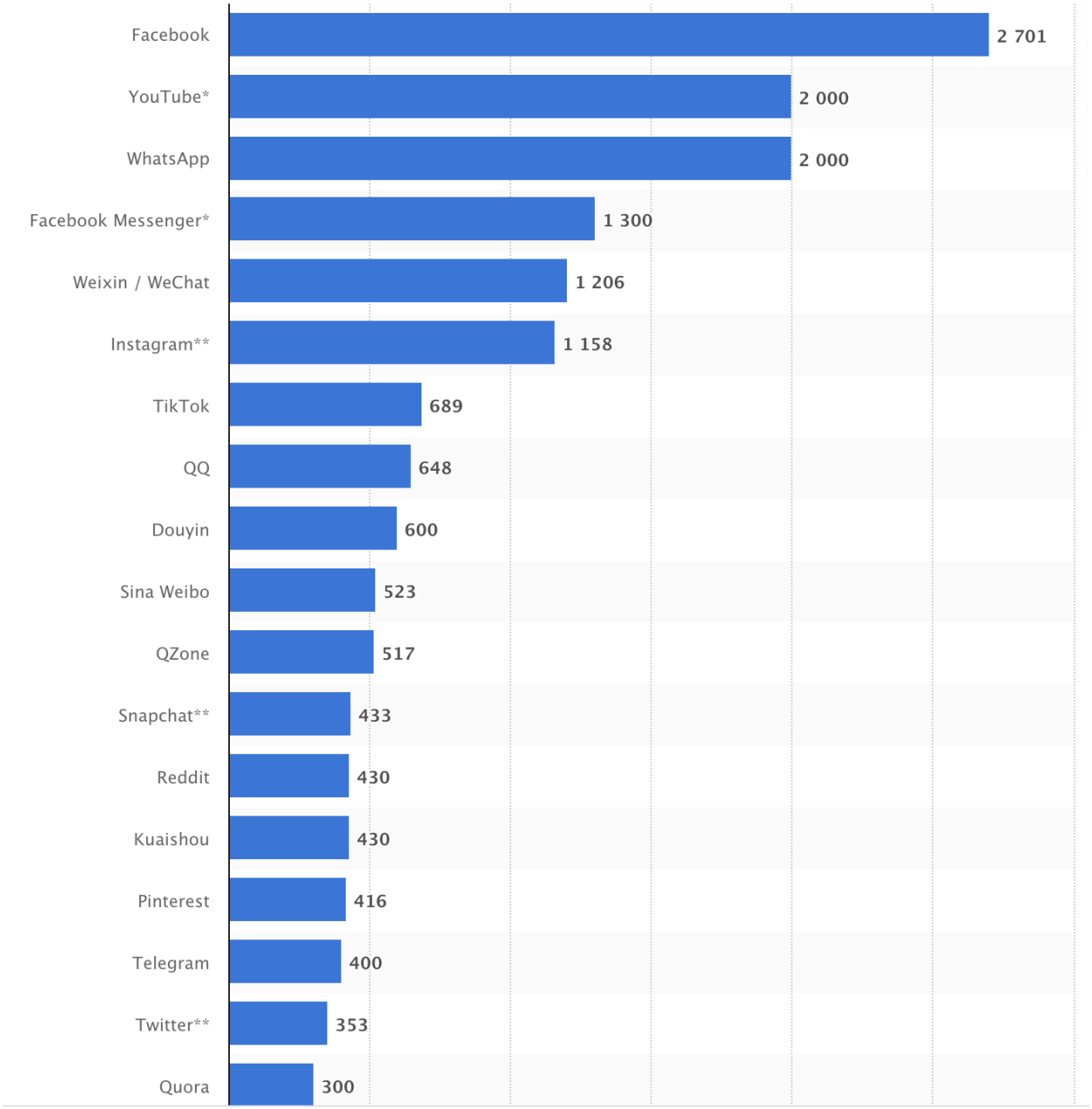


Figure 5. Social media monthly active users by platform (Statista 2020).

3.2.2. GeoNames toponym gazetteer

GeoNames is an open-source toponym gazetteer which holds globally over 25 million place names and their corresponding coordinates (GeoNames, n.d.). GeoNames uses crowdsourcing to update its data, meaning that people can freely add missing locations or fix location coordinates to be more precise. Moderators verify the changes before adding them to the official database. GeoNames also receives data from official sources including National Geospatial Agency world gazetteer and the U.S. Geological Survey of Geographical Names (Gelernter, 2013). GeoNames is used by many open-source geoparsing services (Gritta et al., 2020).

When the spatial extent is limited to Helsinki Metropolitan Area, GeoNames recognizes 1197 places (Figure 6). For every place, there is a primary name and alternative names. In the case of Helsinki, the primary name is most often the name in Finnish. Alternative names can include translations in Swedish, English, and other languages as well as the Finnish name without special characters (ä, ö). For this analysis, I have chosen to only use the primary name for the sake of clarity. The alternative names could be easily mixed with other words that are not place names. In some cases, like small islands, the primary name has been registered in Swedish in the gazetteer which causes inconsistency and might affect the results.

Most of these places in the gazetteer are names of neighborhoods, although some smaller places like islands, hotels and famous venues are also present. From specific sport venues, for example Olympic stadium, Helsinki ice hall and Velodrome are included in the data, but many are also absent like Töölö sports hall (Kisahalli) and Vantaa Energy Arena. The gazetteer does not recognize commercial sports places like gyms, which will affect the results. For example, it would associate the gym Forever Matinkylä with the general coordinates of Matinkylä neighborhood instead of the precise coordinates of the Forever-chain gym.

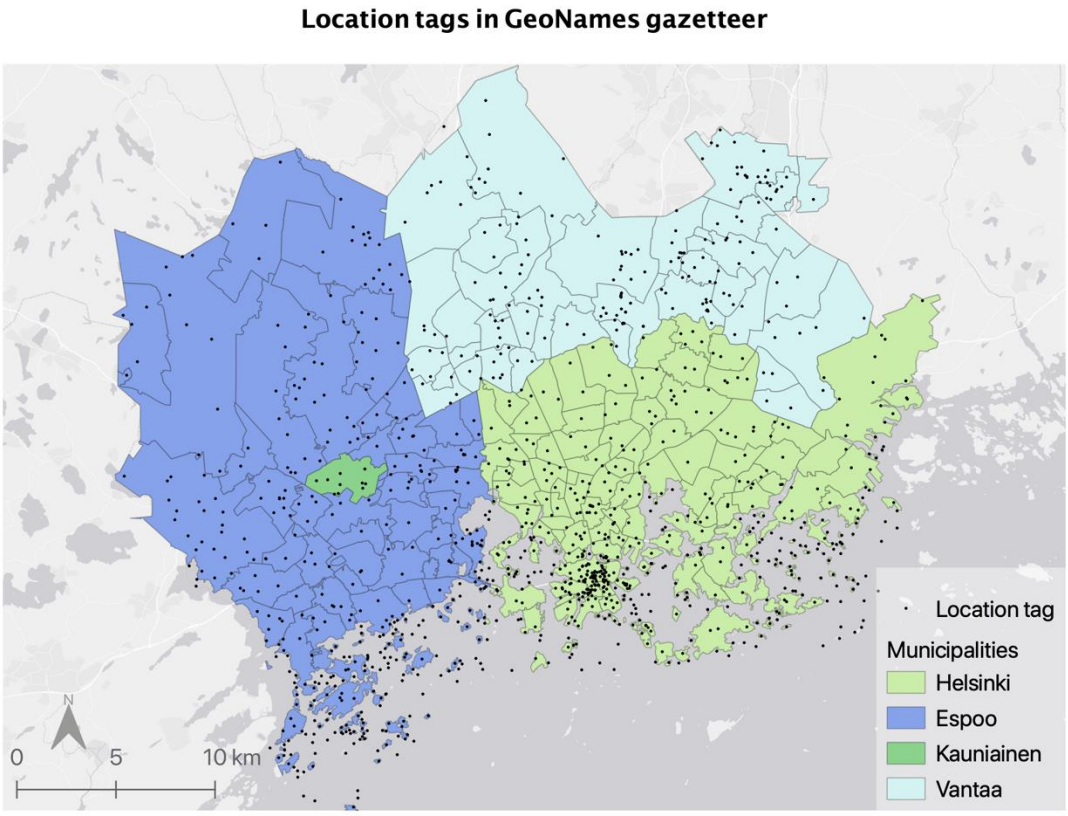


Figure 6. GeoNames gazetteer has 1197 named location tags in Helsinki Metropolitan area.

3.2.3. LIPAS sports facilities database

LIPAS database contains all built sports facilities in Finland. LIPAS includes three types of data; point data for sports facilities like swimming halls, line data for walking, running, skiing and biking routes and polygon data for recreational areas like national parks (University of Jyväskylä, 2020a). In order to be recorded in the database, the sports facilities need to be publicly accessible, well maintained and equipped. Some facilities which have different functions during winter and summer can be recorded multiple times in the database to capture the different use cases. Some of the facilities are specific services like information placards or outdoor fireplaces inside a larger recreational area (University of Jyväskylä, 2020a). In the database, the sports facilities are divided to 8 different categories and further to subcategories based on their function. Every sports facility holds basic information, like name and ownership type of the facility and many also have sports-specific attributes, like the surface type of a

football field. Altogether, there are almost 40 000 sports facilities recorded in the database in early 2021 (LIPAS, 2021).

The database is created and maintained by University of Jyväskylä, the Faculty of Sport and Health Sciences, and funded by the Ministry of Culture and Education (University of Jyväskylä, 2020b). Facility data is kept up to date by municipalities sports workers and private companies who can upload information about new services as they emerge. LIPAS data is openly available (with Creative Commons 4.0- license) and can be accessed via REST API, WMS (Web Map Service) or WFS (Web Feature Service) (University of Jyväskylä, 2020c). Data exploration, queries and downloading are also possible in an online map service (Figure 7).

In this study, I retrieve the sports facilities in Helsinki Metropolitan area from LIPAS database to assess whether the proximity to sports facilities increases the amount of sports-related Twitter posts.

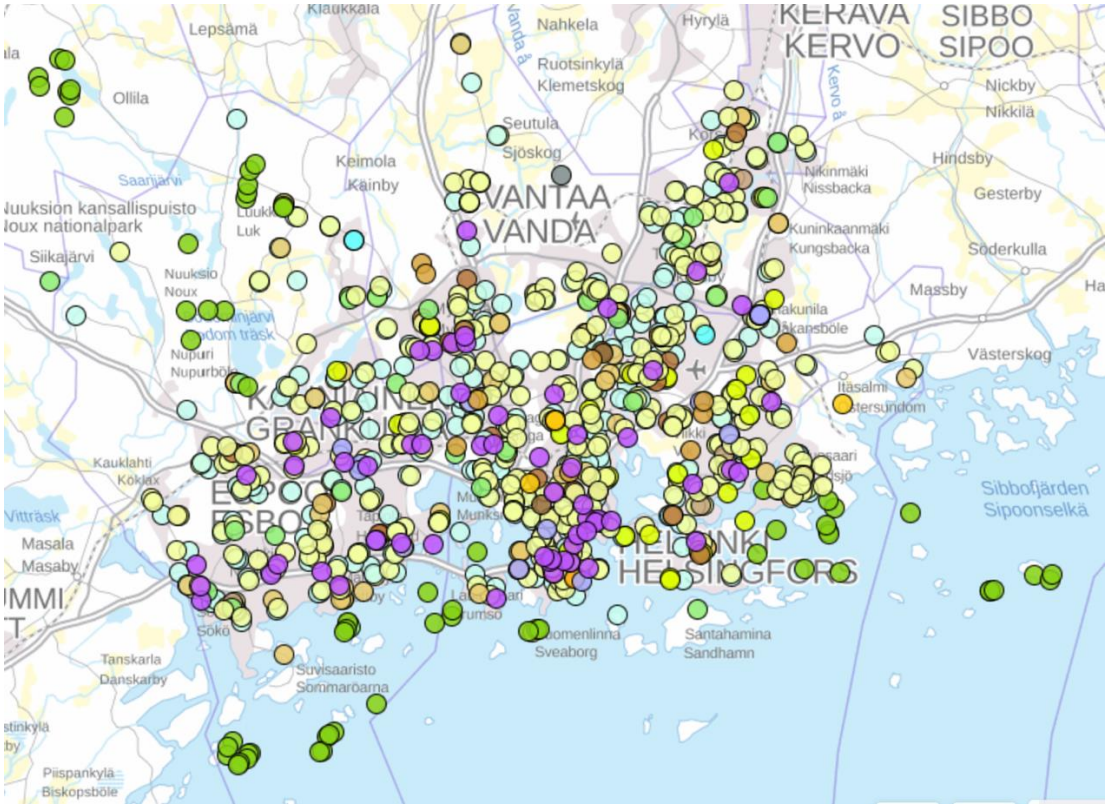


Figure 7. Example of sports facilities in Helsinki area viewed in screenshot from the LIPAS online map service (https://www.lipas.fi/liikuntapaikat).

3.2.4. PAAVO database – socio-economic data by postal code areas

PAAVO Postal code area statistics, collected by Statistics Finland, cover a wide range of statistics related to population and education structure, housing and households, disposable income, and workplaces. All of the above mentioned are aggregated to postal code areas. Most statistics record the actual number of people belonging to a certain group (e.g., 30–34-years-old or holding higher level university degree) and then the percentage of population in that group can be calculated by diving by total population in that area. For example, disposable income in an area is presented in median and mean (euros) and the number of people belonging to each income group (low, middle or high income) (Statistics Finland, 2020).

For this thesis, I used the income and education statistic from PAAVO database. I chose inhabitants' median disposable income for economic indicator (data from 2017, published in 2020). For educational indicator, I chose the number of people holding an academic degree (data from 2018, published in 2020). To make that comparable between postal code areas, I divided it by the number of people over 18 to obtain the share of people over 18 having academic education. The database covers entire Finland, but I only used data for my area-of-interest, the Helsinki Metropolitan Area. In order to make the data spatial, I combined it with the geographical information of the postal codes, which is produced by Helsinki Region Environmental Services (HSY) in cooperation with the cities of Helsinki, Espoo and Vantaa. The data can be retrieved from Helsinki Region Infoshare website in many different formats. Joining of the data was possible based on a common key, postal code numbers. PAAVO postal code data can be retrieved from either Paikkatietoikkuna online map service or PxWeb database by Statistics Finland (Statistics Finland, 2020).

3.2.5. Facebook survey on sports and social media use

I designed a questionnaire to collect more qualitative and quantitative information about the physical activity habits and social media usage related to sports. One of the aims of this survey is also to validate the results from Twitter data and to assess the bias in it. The survey was conducted with Google forms and posted to 28 local neighborhood Facebook groups from Helsinki Metropolitan area. This included 18 groups in Helsinki with 209 000 members altogether, 6 groups in Espoo with 23 000 members, and 6 groups in Vantaa with 40 000

members. The survey was open for two weeks from 15.2.2021 to 1.3.2021. The questionnaire was available in Finnish and English.

The survey consisted of three parts: physical activity, social media usage and background information (Appendix C). Sports activity section included three questions inquiring about the frequency of exercising, exercising that requires facilities and exercising without facilities. There were also two open questions about which sports facilities are frequently used by the respondent and in which neighborhoods the respondent practices sports that don't require facilities. The social media section asked about how often the respondent posts about sports to social media, to which social media platform, in how many percentages of the posts the respondent is the one being active (and not e.g., a spectator or fan). An open question about whether the respondent is more likely to posts about certain sports, and a multiple-choice question which had listed factors that might encourage to posts (like breaking personal best), were also included. Background information inquired about age (in 10-year intervals), gender, city of residence, postal code area, education (4 levels) and annual income (in 15 000-euro intervals). None of the questions were mandatory. See the entire questionnaire in Appendix C or https://forms.gle/vC4eTt7RC2mG8Y9S8.

This survey is used to validate the results from Twitter and to assess the bias of the Twitter data. It is acknowledged that making a survey in Facebook still includes certain social media bias in it which can result in underrepresentation of those who do not use social media. However, Facebook has a wider reach than Twitter as it is the largest social media platform measured by monthly active users (Figure 5). Less biased validation data could be retrieved by making phone interviews that cover all sections of population, in the right ratios. Regardless, I chose the social media survey option due to its cost effectiveness and relatively wide reach. The survey was also easy to execute even in a COVID-19 pandemic situation when interviews with people (at least in person) pose a health hazard.

4. Methods

Main methods in this thesis are those of natural language processing and spatial statistics. In the workflow, I first filtered the GeoNames gazetteer to contain the place names in my research area only. Next, I designed a Python script to lemmatize 38.5 million tweets and used keyword matching to capture the sports-related tweets. I divided those to geotagged ones, that are ready for further analyses, and non-geotagged ones that didn't contain geographical information. Thereafter, I geoparsed the tweets that mentioned place names in the Metropolitan area using Named Entity Matching and the GeoNames gazetteer. For further analyses, I added socio-economic data, sports facilities data and Facebook survey results by postal code areas and performed statistical analyses with those. Figure 8 demonstrates the full workflow of methods and analyses in this thesis.

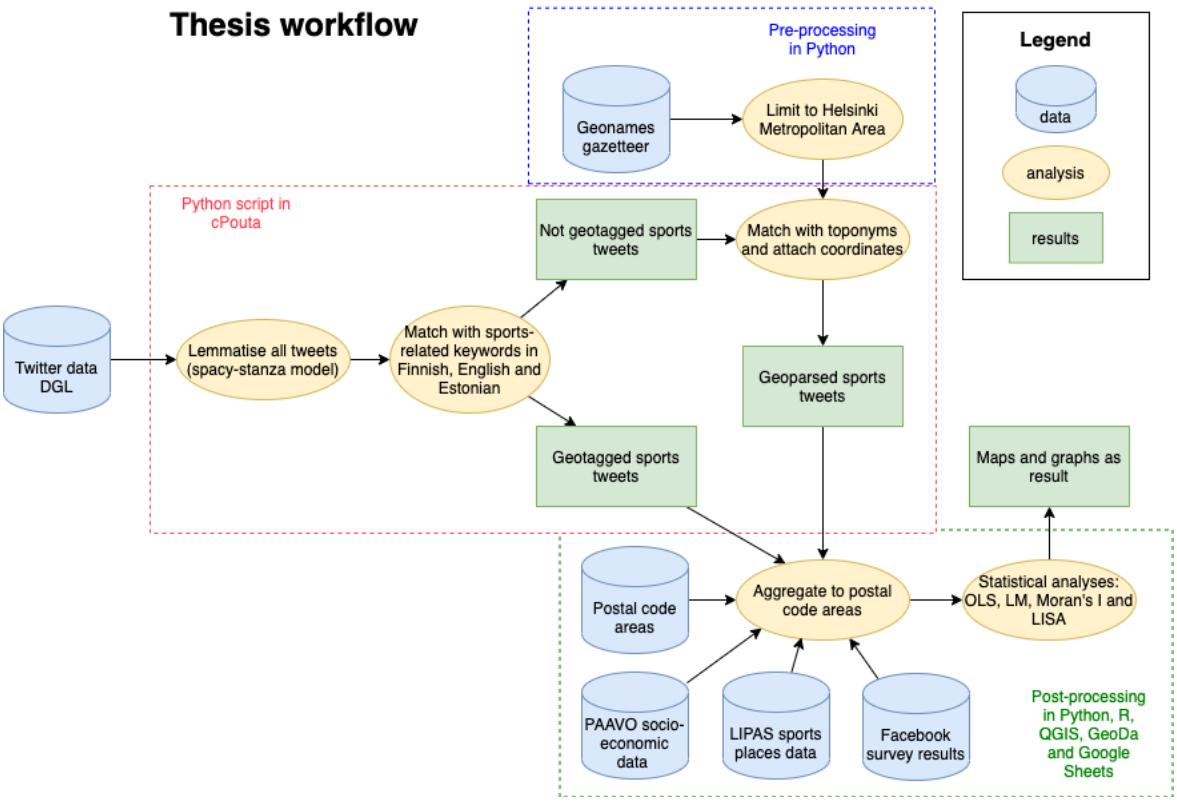


Figure 8. Overview and workflow of methods and analyses.

4.1. Content analysis of the tweets

4.1.1. Lemmatizing and keyword matching

As one of the first procedures for the data, the desired content needs to be filtered from all the data (see Figure 8). In this case, this means gathering the tweets related to sports and concentrating on analyzing those. The method for this content analysis needs to be scalable as the data includes totally around 38.5 million tweets. I chose to lemmatize all the words in the tweets and then search for matches to certain sports-related keywords in English, Finnish and Estonian (Appendix 1). Prior to this, the tweets were divided to different data frames according to the automatically detected language by fastText, including only English, Finnish and Estonian tweets which rules out about 20% of the data (Figure 3). Despite that such a process may create a bias towards the majority languages and flatten the linguistic landscape as minority languages are left out, the inclusion of other languages is out of the scope of this thesis. Spacy-stanza language models were applied for English, Finnish and Estonian data and their lemmatizing functionalities were used to retrieve the lemmas. Lemma means the basic form of a word, so lemmatizing works as follows:

"Loving my new running shoes, they make me fly" → "loving my new run shoe, they make I fly"

This is extremely useful in Finnish, because there are many case endings instead of prepositions and postpositions (cf. In Helsinki - Helsingissä) as Finnish is a morphologically complex language. After these case endings are removed, it is straightforward to search for matches with the keywords (Korenius et al., 2004). The Spacy-stanza language model used for English (ewt) was trained with social media, blogs and emails, for Finnish (tdt) with news, blogs and fiction, and for Estonian (edt) with fiction, news and scientific texts (Stanford NLP Group, 2020).

4.1.2. Content analysis in similar studies

Content analysis in Twitter is usually done by limiting the search to one or couple relevant keywords and hashtags. Examples of these are studies conducted with keywords

"alzheimer" (Cheng et al., 2018), "#eatingdisorder" (Viguria et al., 2020) and "hookah" (Allem et al., 2018). All the aforementioned have a very specific topic, restricted time frame and they consider only English tweets. On the contrary, I would like to capture all tweets about different types of physical activities over a span of a decade, considering multiple languages. To retrieve sports-related posts, other researchers have used manual screening (Roberts et al., 2017) and gathering retweets from sport applications (Torres & Vaca, 2017) as methods. Manual screening was not feasible for me due to the volume of the data. I also tried to avoid any further limitation of the sample by ruling out those who do not own sports tracking devices and connect them to Twitter. Therefore, my approach is different from the ones previously mentioned in the literature. I am trying to capture all tweets related to a relatively broad topic with multiple languages.

4.2. Geoparsing

4.2.1. The meaning and purpose of geoparsing

To tackle the previously mentioned issue of limited amount of geotagged data, I aim to geoparse the tweets that mention place names in order to retrieve more spatial data. Geoparsing means extracting location names, also known as toponyms, from text and converting them to geographical coordinates (e.g., de Bruijn et al., 2017; Gritta et al., 2020). The process of geoparsing can be broken down to two parts: geotagging and geocoding (Figure 9). Geotagging refers to finding the toponyms and geocoding refers to connecting the toponyms to the corresponding coordinates (Gritta et al., 2018).

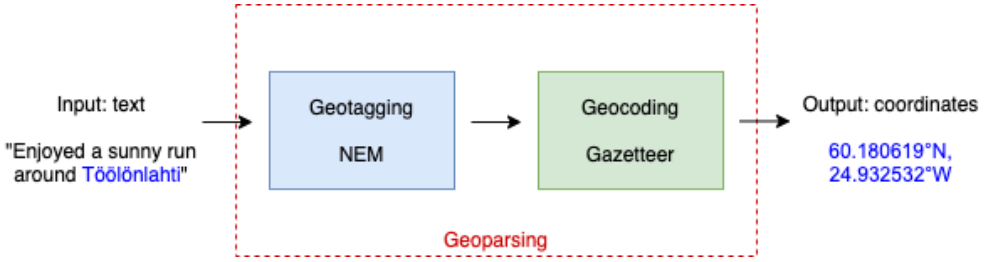


Figure 9. Geoparsing workflow in this thesis. Adapted from Gritta et al, (2018).

4.2.2. Geotagging

I used Named Entity Matching (NEM) approach to find the toponyms (see Figure 8). NEM sees if the words in the tweets match with the pre-existing list of location names stored in either a gazetteer or geospatial database (de Bruijn et al., 2017; Middleton et al., 2018). Since lemmatizing was performed already in the thematic analysis part, the basic forms of the words are easily available for use also in geotagging phase. If any of the lemmas in a tweet match to a placename in the gazetteer, this tweet can be geocoded to attach coordinates to it. Lemmatizing and matching approach only works for one word place names, so for example "*Talin siirtolapuutarha*" cannot be geocoded because it consists of two words and therefore two lemmas.

Alternative for NEM is to use Named Entity Recognition (NER) which tags all toponyms in text without the need to match them to a location in gazetteer or database. NER utilizes Natural Language Processing to analyze the grammar and linguistic properties of the text (Middleton et al., 2018). The upside of this method would be that it is not restricted to a list of toponyms and that it can analyze the sentence structure to differentiate with for example similar names of location and people (de Bruijn et al., 2017; Gritta et al., 2020). However, SpaCy's NER functionality is not yet available for Finnish and the capability to recognize Finnish placenames is not advanced for the English version either. Therefore, in this thesis, I am using NEM for the task of toponym recognition.

4.2.3. Geocoding

For the geocoding part, I chose to use gazetteer approach. As a gazetteer, I am using GeoNames, which contains globally 25 million placenames and corresponding coordinates (GeoNames, n.d.). GeoNames is used by many open source geoparsing services (Gritta et al., 2020). When the spatial extent is limited to Helsinki Metropolitan area, the gazetteer holds 1197 unique toponyms, mostly on neighborhood level. A larger local gazetteer including very specific place names on building level would of course yield more accurate results. For example, "Käpylä football field" could be geocoded to the exact location of the football field instead of geocoding it to the coordinates of "Käpylä". If any of the place names match to a lemma in a tweet, the coordinates of that place name are attached to that tweet. The matching was performed in a way

that does not distinguish between capital and small letters since many do not bother with capital letters in social media text. All analysis processes were done in Python and the code is visible in GitHub.

4.2.4. Unsolved challenges in geoparsing

There are many challenges in geoparsing. One of them is dealing with metonyms, figures of speech where one concept substitutes for another. Metonyms for place names are common in situations where the place name refers for example to the government of the place or a national sports team (Gelernter & Balaji, 2013; Gritta et al., 2018). Examples include "Helsinki goes to lockdown" or "Finland beat Sweden in the finals". Here place names are representing actors in the area although location themselves do not have any agency of their own (Gritta et al., 2018). Similar challenges are posed by demonyms, when place name refers to the inhabitants of the place (Gritta et al., 2020). There can also be confusions in between languages and place names or people's names and place names (de Bruijn et al., 2017). For example, *Tapanila* is a neighborhood in Helsinki but also a common last name. In Named Entity Matching, these challenges are disregarded as the method does not account for the words' purpose in the sentence like Named Entity Recognition would.

Other challenges include that the use of a place name does not necessarily indicate the person's location at the moment of posting. A tweet can for example address a topical issue on the other side of the world or be a throwback to a different time. While a tweet can also mention multiple different place names, one of them must be chosen as representative. For example: "*Enjoyed my commute from Hakaniemi to Haukilahti by bike.*". In this case, both Hakaniemi and Haukilahti are relevant locations since they are origin and destination. I solved this in a simple and straight-forward manner by allocating the coordinates according to the first toponym mentioned. For more accurate and sophisticated results, the degree or generality could be measured and the most specific one would be picked.

It is expected that many tweets mention general place names like *Helsinki*. This will unproportionally increase the number of tweets geocoded to a certain location which has the coordinates for Helsinki. To avoid the creation of artificial hotspots because of this, I removed

the tweets that had Helsinki, Espoo, or Vantaa as their only location-indicative word. In addition, I also removed obvious bots that increased the tweet count in certain areas.

4.2.5. Manual corrections to geocoded data

General toponyms like Helsinki, Espoo, Vantaa and Uusimaa were removed in post-processing. These would create artificial hotspots to the data as they are too high-level to be included. Furthermore, a bike counter bot tweeting every day biker amounts from Baana cycling route was removed.

I manually inspected some of the posts to address the issues related to geoparsing mentioned in section 4.2.4. To find potentially incorrectly geoparsed tweets, I inspected the place names that have multiple meanings and the areas which had very high ratio of sports related tweets to population or percentage of sports related tweets of all geotagged tweets. As a result, I found over 1400 tweets that were geoparsed on the wrong basis and I cleaned those from the data. Table 2 presents the number of posts that were removed and mentions what other meanings the place name had when it was not used in the place context. An example tweet incorrectly geoparsed to place name *Kilo* would be "*After dropping 6 kilos running feels much lighter*". NER can be a tool to explore in my future studies which may prevent this kind of errors since it would recognize from the context that kilo in this case is not a geographical location. Many tweets that were from sports and leisure center Vierumäki in Heinola were incorrectly geoparsed to Vierumäki in Korso.

Table 2. Tweets cleaned from the geoparsed dataset.

Place name	Count not related to the place	Other meanings
Kera	327	"With" in old fashioned Finnish, Python library Keras
Vierumäki	291	Sports and leisure center in Heinola (outside study area)
Kilo	224	Short for "kilogram"
Melkki	189	Incorrectly lemmatized to "melkein"("almost" in Finnish)
Reuna	141	"Edge" in Finnish
Kallio	74	"Rock" in Finnish, last name
Virkamies	55	"Public officer" in Finnish

Ruusuvuori	34	Last name
Saarijärvi	28	Municipality, last name
Tapola	25	Last name
Uusimaa	23	Province
Kalmari	15	Last name
Metsola	13	Last name
Syväoja	5	Last name

4.3. Data analyses

4.3.1. Data aggregation and normalization

For further analysis, I aggregated the sports tweets (n = 20 599) to postal code areas. For the social media data, it is important to choose the aggregation level carefully to protect the users' privacy and to have meaningful analysis outcomes (de Andrade et al., 2021; Zelenkauskaite & Bucy, 2016). Postal code areas correspond to neighborhood borders, so they are a natural unit of aggregation. Many people seek for services primarily within their neighborhood and have a certain neighborhood identity (Paananen, 2020). Additionally, many socio-economic variables are available by postal code areas. In the same manner, I aggregated the sports facilities and survey results also by post code areas. Aggregation was done with point-in-polygon method in QGIS programme (version 3.16.3).

Postal code areas can also be problematic as aggregation units. When assessing how area-based attributes impact individual activity patterns, two inherent problems arise (Kwan, 2012). Firstly, zoning (irregular shapes) and varying sizes of postal code areas can affect the analysis outcome due to modifiable areal unit problem (MAUP) (Fotheringham & Wong, 1991; Openshaw, 1984). Secondly, uncertain geographical context problem (UGCoP) questions the suitability of pre-defined units of analysis for capturing the patterns of human activities and therefore the analysis can produce different outcomes with different units of aggregation (Kwan, 2012).

Since postal code areas are of varying size and population, it is recommended that absolute numbers are not represented on them because they are not comparable (Grubesic, 2008). Thus, data normalization process is needed to enable the comparability of numbers across the postal code areas (Foster, 2019). Figure 10. Figure 11Figure 12Figures 10-12 showcase how different

data normalizations produce different analysis outcomes. The normalizations I tried for sports related tweets are: 1) number of sports related tweets divided by number of all geotagged tweets, 2) number of sports related tweets per 1000 inhabitants of age 13 or above (Twitter's lower age limit, (Twitter, n.d.-a)), and 3) number of sports tweets per square kilometer.

Each of the normalization methods have their own challenges. Normalization per area is substantially affected by the population density in the area. Normalization per geotagged tweets can be problematic because the sports related tweets also contain geoparsed tweets, so the number would not strictly represent the percentage of sports related tweets. In both normalizations per inhabitant and geocoded tweets, areas with very small number of total tweets or population can easily be misrepresented as hotspot if for example 3 out of 6 tweets are sports related. Slightly inaccurate geocoding can further emphasize this problem. For example, "Herttoniemi" is geocoded inside Roihupelto industrial area which artificially elevates the score of the industrial area which has only few inhabitants.

4.3.2. Measures of spatial autocorrelation

Tobler's (1970) first law of geography, "*everything is related to everything else, but near things are more related than distant things*" summarizes the idea of spatial dependence and spatial autocorrelation. Spatial autocorrelation means that a variable has similar values in places that are close to each other (Moran, 1950). This phenomenon is also known as clustering, and it can be statistically measured by Moran's Index. Global Moran's Index takes a value between -1 and 1 based on whether the variable is perfectly dispersed (-1), random (0) or perfectly clustered (1) (Moran, 1950).

While Global Moran's I indicates the strength of spatial autocorrelation, Local Moran's I (also known as Local Indicators of Spatial Autocorrelation, LISA) reveals where the clustering occurs (Anselin, 1995). LISA produces a map which exhibits the different clusters (high-high and low-low) and outliers (high-low and low-high). For instance, low-low cluster means that the particular area has a low value (e.g., number of sports tweets) and the surrounding areas have low values also. Similarly, a high-low outlier means that the particular area has a high value, but

the surrounding areas have low values. Both Global Moran's I and LISA can also be performed as bivariate, meaning that they measure the covariance of two variables spatially.

In order to measure spatial autocorrelation, neighborhood connectivity has to be defined. The connectivity is defined by creating spatial weights (W), which indicates which areas are considered neighbors. I use Queen contiguity (order = 1) for neighborhood definition, meaning that if the postal code areas share an edge (border) or a vertex (corner), they are considered neighbors (Foster, 2019). Suomenlinna was removed from the data as a neighborless observation. Same spatial weights are used in LISA, Bivariate LISA and OLS analyses. Analysis of spatial autocorrelation was executed in GeoDa open-source software (version 1.16.0.16).

4.3.3. Ordinary Least-Squares Regression

For the aggregated and normalized variables, I searched for relationships between the number of tweets and explanatory variables with different statistical analyses. I used Ordinary least-squares (OLS) regression, which is a generalized linear modelling technique. OLS takes one dependent (response) variable which can be predicted with one or multiple independent (explanatory) variables (Hutcheson & Sofroniou, 1999). OLS calculates a straight trend line for the data by minimizing sum of the squared errors (i.e., distances to the trend line). Errors are the differences between the observed and predicted values of response variable. OLS returns values for the intercept (alpha), slope (beta), R^2 and p-value. R^2 depicts how much of the variation of response variable can be explained by the explanatory variables (Hutcheson & Sofroniou, 1999). The bigger the R^2 value is, the better the model explains variations in the response variable (maximum R^2 being 1). The p-value measures the statistical significance of the model.

The fit of a model is often assessed with Akaike's Information Criterion (AIC), (Akaike, 1974). In this thesis, I used the second-order AIC (AICc), which gives a relative number as an output and therefore enables the comparison between different models. The smaller the AICc, the better fit the model is.

Socio-economic variables often exhibit multicollinearity, meaning that the independent variables correlate with each other. This violates one of the assumptions of OLS and thus impacts the reliability of the model (Tabachnick & Fidell, 2007). Multicollinearity can be measured with Variance Inflation Factor (VIF), which takes values from 1 upwards. A VIF score of 1 indicates no multicollinearity, 5 indicates moderate multicollinearity and 10 strong

multicollinearity (Dodge, 2008). Variables with too high VIF scores can be eliminated from the model to improve the model's performance and reliability. The analysis was carried out in R Studio (version 1.4.1106) with the packages foreign, spdep, spatialreg, car, AICcmodavg and corrplot. The script is available in GitHub.

4.3.4. Lagrange Multiplier tests

The ordinary least-squares (OLS) regression is a very simple model, but not always the most fitting. It assumes, among other things, normality, linearity and homoscedasticity of the errors (Tabachnick & Fidell, 2007). Spatial dependence and heterogeneity of the variables can cause some of the aforementioned assumptions to fail. Therefore, I performed Lagrange Multiplier test to assess the fitness of OLS and to observe whether some spatial statistical models would perform better. Lagrange Multiplier tests assess whether data exhibits spatial dependence or spatial heterogeneity and thus does not fulfill the assumptions for OLS regression (Anselin, 1988). OLS regression and Lagrange Multiplier test were performed in R; the script is available in GitHub.

4.4. Processing survey answers

The aim of the Facebook survey on sports and social media use was to validate the Twitter results and assess the inherent bias in it. The survey included questions about frequency of doing sports, social media behavior related to sports and background information. The collection method is described more in detail in section 3.2.5. After data collection, I converted many of the answers into numerical format to calculate correlations. I coded the "how often" questions to have values from 1 (rarely) to 6 (many times a day). As these are ordinal scale variables, Spearman's rank-based correlation coefficients can be calculated after numerating them (Vogt & Johnson, 2015). I also coded the four levels of education from 1 to 4 and used the mean of every income class and age group. Most of the data numerating was carried out in Google Sheets and some afterwards in Python.

To compare the physical activity levels in different post code areas, I aggregated the data to postal code areas and compared with the number of sports facilities in an area. However, not all

respondents wanted to disclose the postal code area where they live, and I only obtained responses from 80 out of 168 postal code areas. Correlations are calculated in Python with SciPy library, see the code in GitHub.

5. Results

5.1. Number and distribution of sports tweets

Out of 2 030 499 geotagged tweets, only 59 550 (2,93%) are in English, Finnish, or Estonian and related to sports with the keywords in Appendix 1. Out of those, 16 946 are geotagged in the study area. Out of these tweets, 12 096 are in English, 4 690 in Finnish and 160 in Estonian. The original data contains 36 457 267 non-geotagged tweets. 19 412 of these are sports related and geoparsed to the study area. 8 319 of these are in English, 10 983 in Finnish and 110 in Estonian. This makes 36 358 sports related tweets in the study area in total. After removing those geoparsed to city geotags, 25 243 tweets remain. After cleaning a bot and incorrectly geoparsed tweets, as mentioned in section 4.2.5., the final data consists of 20 599 tweets of which 13 746 are geotagged and 6 853 geoparsed.

Maps A–C in Figure 10 (A: Heat map, B: Point map and C: Theme map) present the absolute number of tweets. Maps D–F present different methods of normalizing the data: number of sports tweets divided by number all geotagged tweets (D), sports tweets per square kilometer (E) and sports tweets per 1000 inhabitants (F). The normalization per inhabitant (in map F in Figure 10 and in all following figures) takes into account only those over 13 years old because that is the age limit for Twitter users (Twitter, n.d.-a). As can be seen from the figure, the method of normalization has a profound effect on the patterns that are revealed (compare maps D–F in Figure 10). Normalization of the data is required to make the postal code areas comparable, since they vary in area and population (Foster, 2019). In the spatial autocorrelation analysis, I present the results of each normalization method as well as absolute numbers for the sake of comparison. For all maps A–F in Figure 10, I have used Natural Breaks as classification method and classified the areas into 5 classes.

Maps A –C in Figure 10 show that most sports tweets are located in the peninsula of Helsinki where the most people also live. That is also where geotagged tweets are generally concentrated, hence the percentage of sports related tweets is not necessarily high in the Helsinki peninsula (map D in Figure 10). The percentage of sports tweets is the highest in the post code areas that contain some important sports facilities. For instance, Nupuri-Nuuksio has Nuuksio National Park, Maunula-Suursuo has Pirkkola sports park, and Myllypuro has Liikuntamylly, Pallomylly and Arena Center. Tweets are the densest in Helsinki peninsula and surrounding areas, in the most populous areas (map E in Figure 10). Tweets per 1000 inhabitants show quite similar patterns as the absolute number of tweets, although some areas with small number of inhabitants stand out, such as Petikko in western Vantaa (61 inhabitants, 13+ years old) and Veromiehenkylä in mid-Vantaa (414 inhabitants, 13+ years old) (map F in Figure 10).

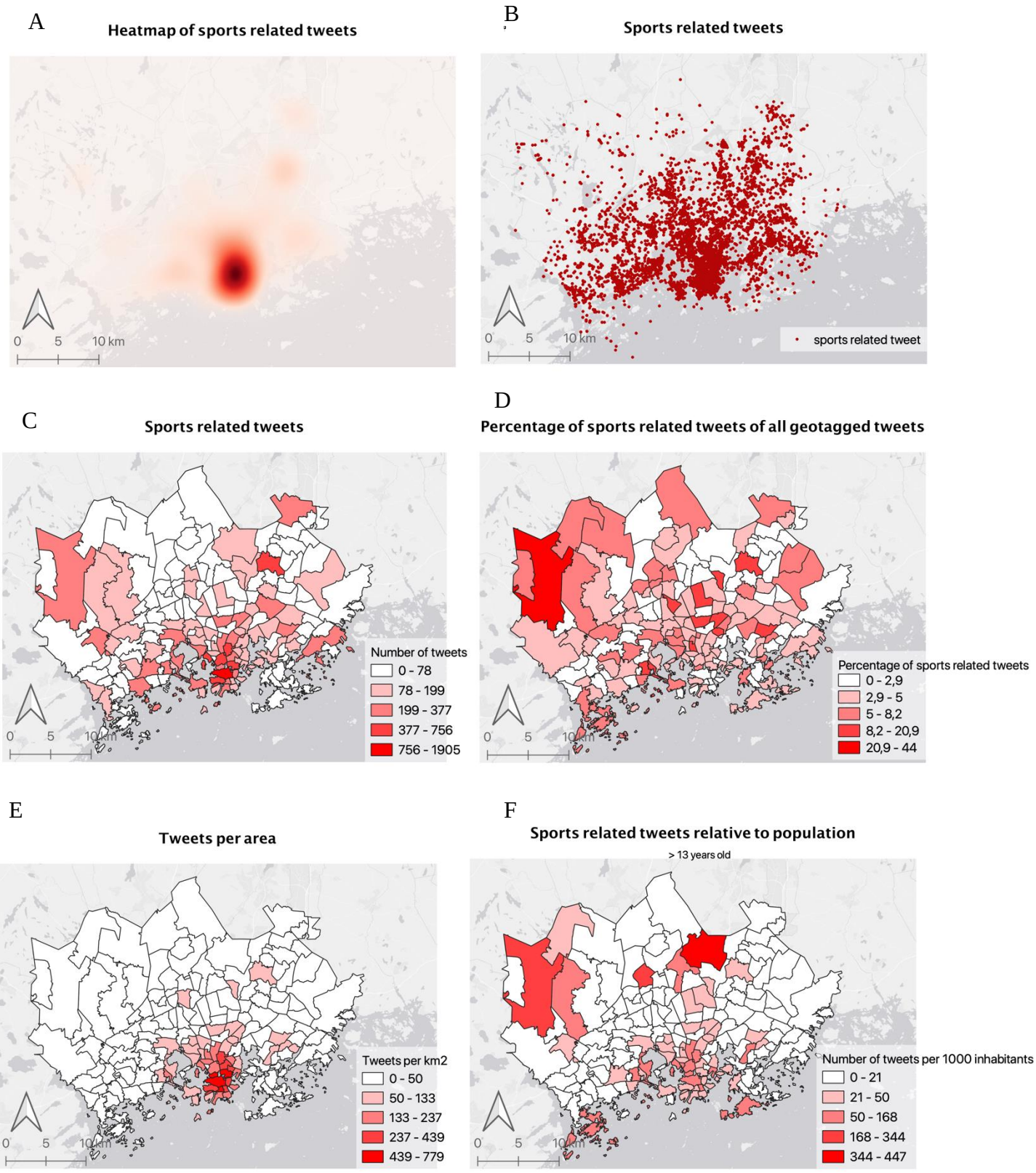


Figure 10. Collection of maps showing distributions of sports related tweets in the study area.

5.2. Spatial autocorrelation and clustering

5.2.1. Moran's Index

Spatial autocorrelation is often measured by Moran's Index (henceforward Moran's I) which takes values from –1.0 (strongly dispersed) to 1.0 (strongly clustered). Figure 11 demonstrates the spatial autocorrelation of sports related tweets with absolute numbers and three aforementioned normalization methods.

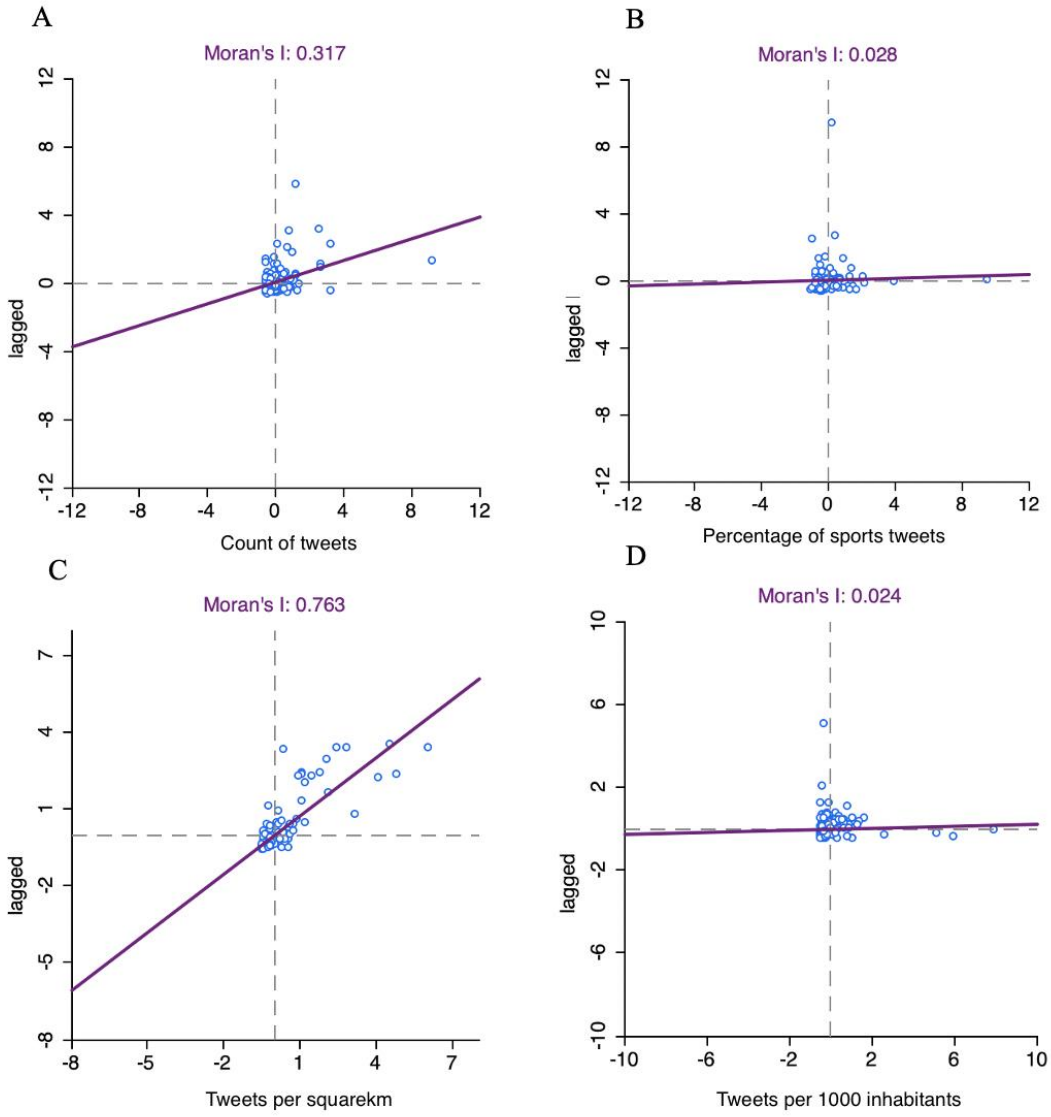


Figure 11. Moran's I for tweets in postal code areas (A), percentage of sports related tweets of all geotagged tweets (B), tweets per square kilometer (C) and tweets per population (D: only 13+ years old)..

Figure 11A shows that the clustering of absolute number of sports tweets per postal code area is moderate (0.297). Percentage of sport tweets (Figure 11B) and sports tweets relative to population (Figure 11D) do not exhibit any clustering but seem to be randomly distributed. However, the Moran's I of tweets per square kilometer (Figure 11C) shows strong clustering (0.762). This can be explained by the fact that the peninsula of Helsinki is tightly populated while outskirts of Vantaa and Espoo are more sparsely populated. Therefore, sports related tweets follow the distribution of population, because when looked at relative to the population, the distribution does not exhibit clustering.

5.2.2. Local indicators of spatial autocorrelation

Local indicator of spatial autocorrelation (also known as Local Moran's I, LISA) is used to examine hot spots, cold spots and outliers. As can be seen from Figure 12A and 12C, absolute number of sports tweets and sports tweets per square kilometer are clustered in the center of Helsinki and the coldest spots are in northern Vantaa and Espoo. For the absolute number of tweets, some low-high outliers are found in the center, including Jätkäsaari and Meilahti hospital region. Low-high outlier means that a certain neighborhood (e.g., Meilahti hospital region) has low number of sports tweets while the surrounding neighborhoods have high number of sports tweets.

For the percentage of sports tweets and sports tweets relative to population (Figures 12B and 12D), no clear patterns of clustering were found. The identified clusters are mostly significant only on $p < 0.05$ level and most areas are statistically insignificant. Such a pattern makes sense because the Moran's I charts (see Figure 11) already show that the distribution pattern is almost random – no real clustering is expected to be present.

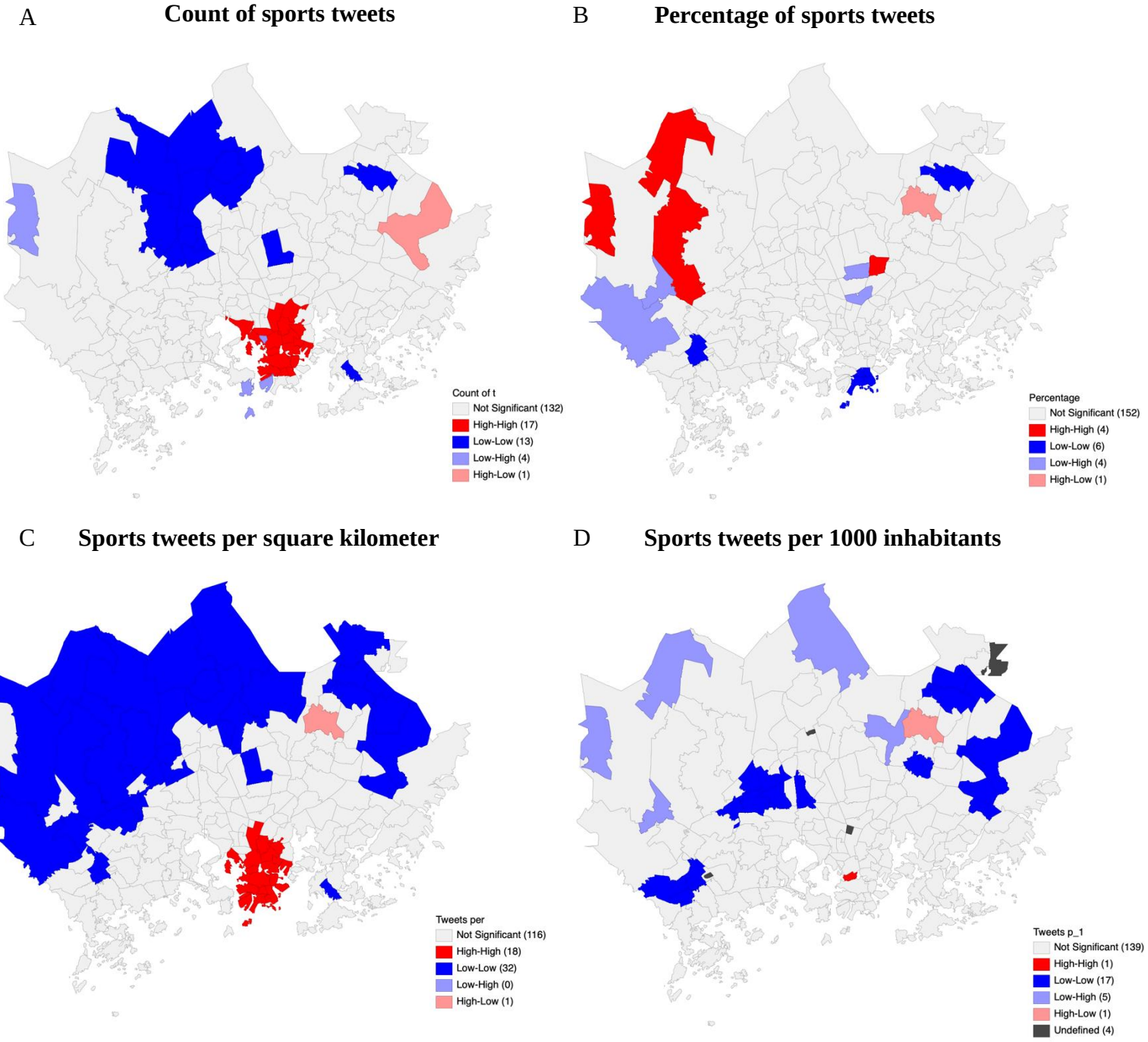
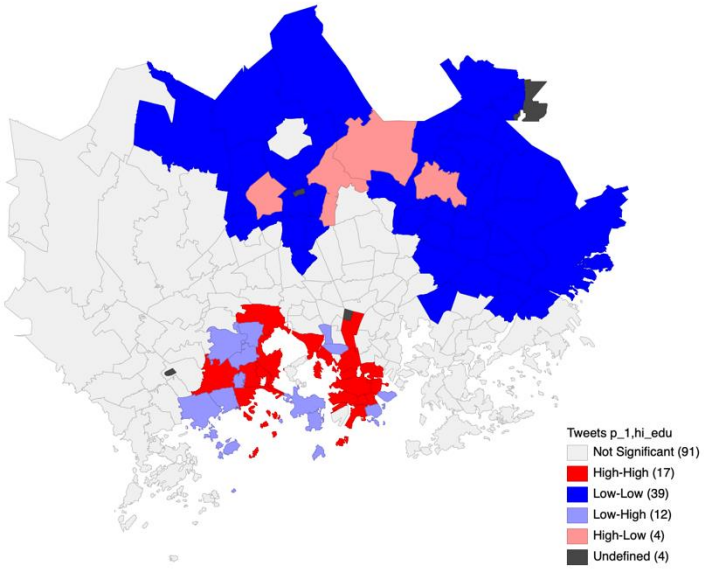


Figure 12. LISA analysis for number of sports related tweets (A), percentage of sports tweets (B), sports tweets per square kilometer (C) and sport tweets per 100 inhabitants (D).

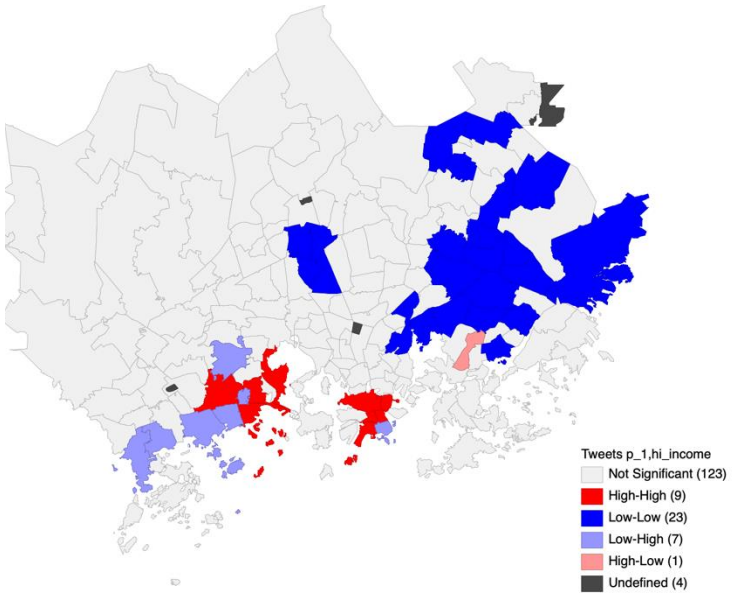
5.2.3. Bivariate LISA

Bivariate LISA is meant to explore common patterns of clustering between two different variables. In this thesis, I used bivariate LISA to search for common patterns within the number of sports tweets and socio-economic variables as well as the number of sports facilities.

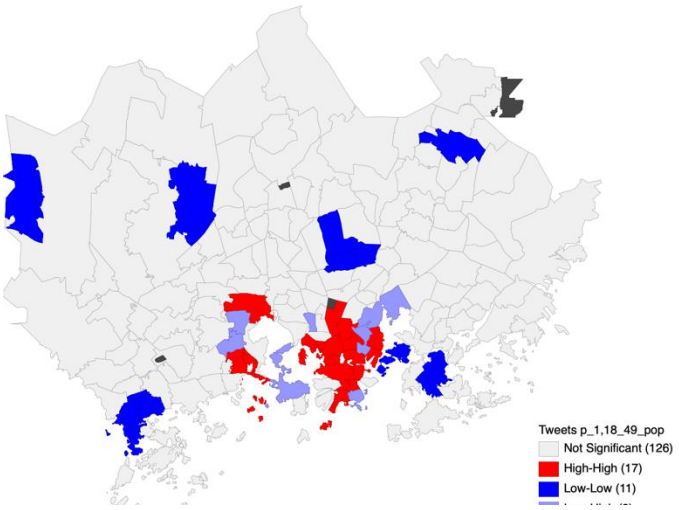
A **Sports tweets and academic degree**



B **Sports tweets and high income**



C **Sports tweets and share of 18 – 49 years olds**



D **Sports tweets and sports facilities**

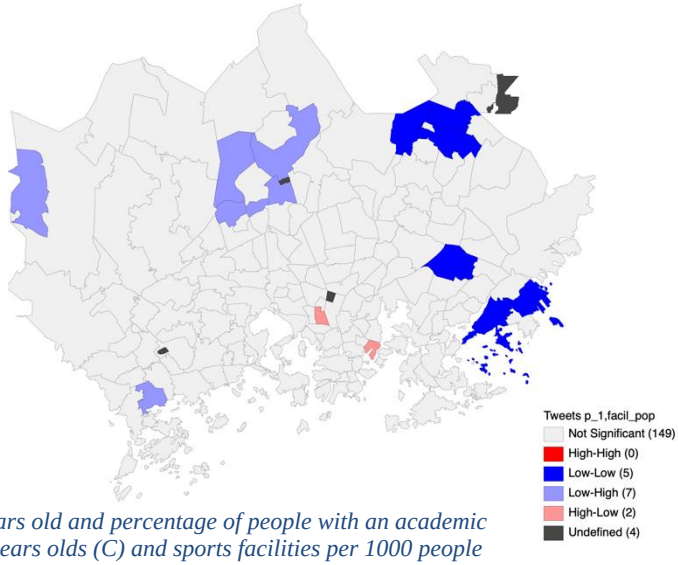


Figure 13. Bivariate LISA with sports tweets per 1000 inhabitants over 13 years old and percentage of people with an academic degree (A), percentage of people with high income (B), percentage of 18-49 years olds (C) and sports facilities per 1000 people (D).

The Moran's I is low (< 0.1) for all bivariate tests with sports tweets and socio-economic indicators, which means that there is no major clustering. However, certain patterns which can be seen in LISA analysis of the sports tweets are repeated in many of the bivariate LISA analysis. The same juxtaposition between the peninsula of Helsinki and the outskirts of Vantaa and Espoo are visible in both univariate and bivariate LISA.

When comparing the number of sports tweets per inhabitant to the percentage of inhabitants who have academic degrees, the same Helsinki peninsula – outskirts of Metropolitan area pattern appears (Figure 13A). Center of Helsinki and areas around Laajalahti are part of a high-high cluster, meaning that there are many sports tweets and many people holding academic degrees. On the contrary, the surrounding areas like Lauttasaari, Haukilahti, Westend and Mankkaa have low numbers of sports tweets and plenty degree holders. Western and Eastern Vantaa as well as north-eastern Helsinki form a low-low cluster, meaning that there are few sports tweets per person and few academic degree holders. However, some neighborhoods in mid-Vantaa such as Veromiehenkylä, Tikkurila, and Petikko stand out with their high number of sports tweets per person.

A similar pattern can be seen in the bivariate analysis with percentage of people with high income, although fewer postal code areas belong to significant clusters (Figure 13B). Central Helsinki and western Espoo have high number of sports tweets per 1000 inhabitants and high proportion of high-income earners. Eastern Helsinki and Vantaa belong to a low-low cluster having fewer sports tweets per person and a smaller share of high-income earners. Figure 13C shows that Helsinki peninsula is a high-high cluster also in sports tweets per person and proportion of those between 18 and 49. I chose this age group because the group contains the most active tweeters, see Figure 1 (Statista, 2020).

Sports tweets per inhabitant comparison to sports facilities per inhabitant show a different kind of spatial pattern. There are few significant clusters, see Figure 13D. Kontula, Vuosaari and northwest Vantaa are part of a cluster where number of sports tweets per 1000 inhabitants is low and number of sports facilities per 1000 inhabitant is also low. This might be because Kontula and Vuosaari are the most populous postal code areas. To conclude, the lack of clustering could be a result of successful positive discrimination and sports equity politics in the Metropolitan area.

5.3. Statistical prediction

5.3.1. Variable preselection

For socio-economic analysis, I chose 16 variables from PAAVO postal code database and one variable measuring the sports facilities from LIPAS, see Table 3. PAAVO variables concern inhabitants' income and education level, their employment status, age structure and housing type.

Table 3. All independent variables considered for the socio-economic analyses.

Variable	Description	Year	Source
Sports facilities (number of sport facilities / 1000 inhabitants)	Sports facilities aggregated to postal code areas and divided by 1000 inhabitants	2020	LIPAS (2020)
High education (%)	Percentage of people over 18 years old who hold higher or lower academic degree	2018	Statistics Finland (2020)
Low education (%)	Percentage of people over 18 years old who have only basic level education	2018	Statistics Finland (2020)
High income (%)	Percentage of people belonging to the highest income class	2018	Statistics Finland (2020)
Low income (%)	Percentage of people belonging to the lowest income class	2018	Statistics Finland (2020)
Median income (%)	Median income of inhabitant	2018	Statistics Finland (2020)
Children (%)	Percentage of 0-14-years-olds of all inhabitants	2018	Statistics Finland (2020)
Students (%)	Percentage of students of all inhabitants	2018	Statistics Finland (2020)
Employed (%)	Percentage of employed of all inhabitants	2018	Statistics Finland (2020)

Unemployed (%)	Percentage of unemployed of all inhabitants	2018	Statistics Finland (2020)
Pensioners (%)	Percentage of pensioners of all inhabitants	2018	Statistics Finland (2020)
Household size (people)	Average household size	2018	Statistics Finland (2020)
Kid households (%)	Percentage of households that have at least one 0–17-year-old	2018	Statistics Finland (2020)
Adult households (%)	Percentage of households where all members are 18 – 64-years-old	2018	Statistics Finland (2020)
Pensioner households (%)	Percentage of households that have at least one over 64-year-old	2018	Statistics Finland (2020)
Homeowners (%)	Percentage of households who own their homes	2018	Statistics Finland (2020)
Renters (%)	Percentage of households who live in a rented home	2018	Statistics Finland (2020)

Many of these variables are defined to measure a certain socio-economic aspect of the society. For example, to assess the inhabitants' financial conditions, income is classified into three different levels (high, median and low), and thus these variables are correlated with each other (Figure 14). With variable preselection, I aim to choose the most suitable variables for Ordinary Least Squares regression. I based the variable preselection on a correlation matrix with 0.05 significance level, to pick the most suitable variables for OLS (Figure 14). The matrix shows that the variables that correlate positively with sports tweets per 1000 inhabitants over 13 years old are sports facilities, employment, percentage of rental homes, and percentage of households with only adult (18-64 years) population. The variables that correlate negatively are percentage of homeowners, percentage of children, household size, percentage of households with at least one underaged. Therefore, in the initial regression analysis, I entered these as independent variables. Crosses in the matrix mean that the variables are not significantly correlated on 0.05 confidence level.

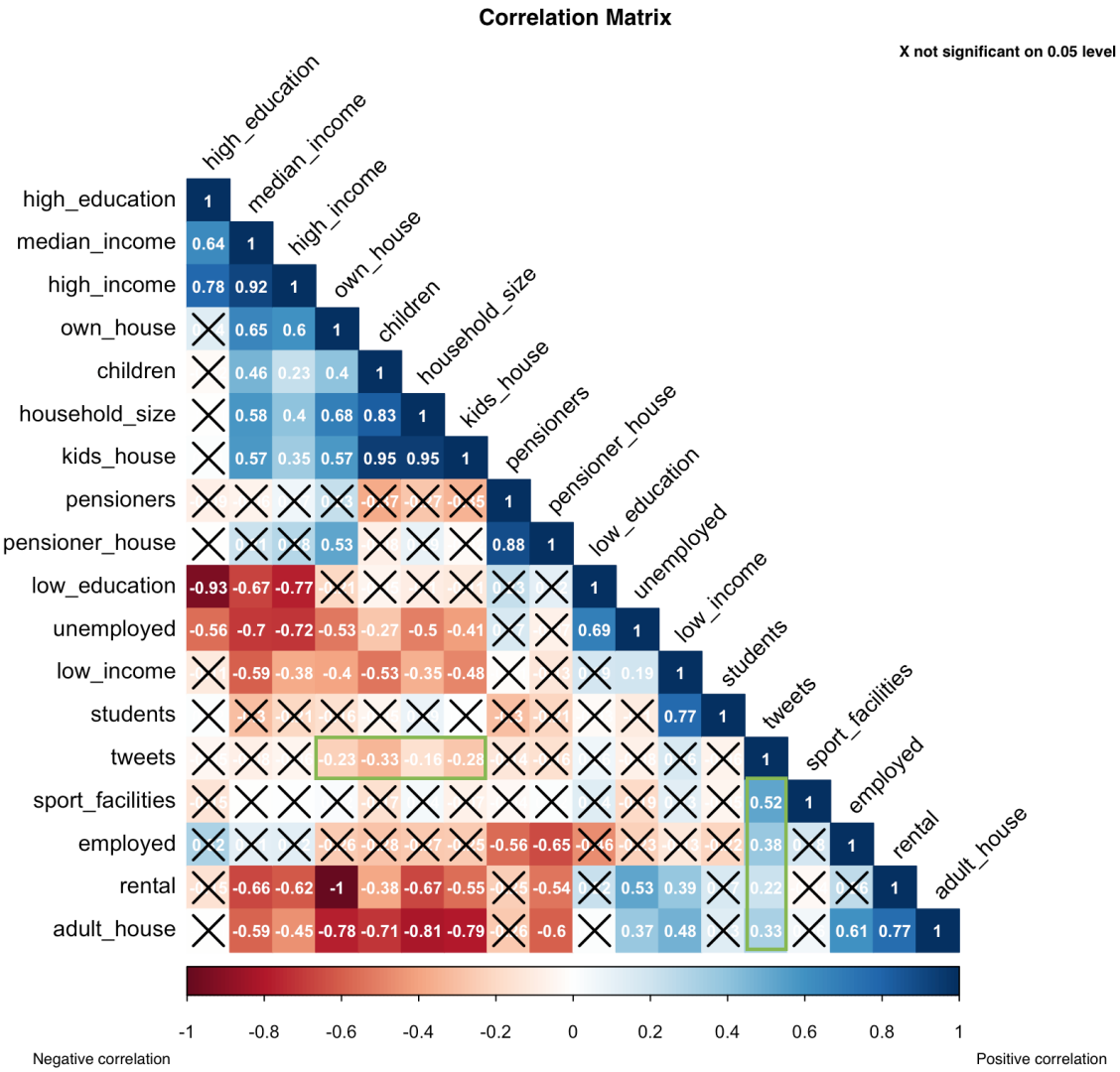


Figure 14. Correlation matrix between all variables. Significant correlation with sports related tweets per 1000 inhabitants marked with green rectangles.

5.3.2. Ordinary Least Squares

Ordinary Least Squares (OLS) is a linear regression that measures the relationship of a dependent and independent variable or variables. In other words, a dependent variable is predicted with independent variables. In this study, the dependent variable is the number of sports tweets per 1000 inhabitants aged over 13. The independent variables are sports facilities, employment, percentage of rental homes, percentage of households with only adult (18-64 years) population, percentage of homeowners, percentage of children, household size, and percentage of households with at least one minor.

Many of these variables measure the same thing with different classifications and therefore multicollinearity, i.e., correlation between the independent variables, is expected to be present. Multicollinearity can be measured with a VIF-score, which should not exceed 5 or the p-values cannot be trusted (James et al., 2013). On the first run, percentage of rental homes and percentage of owned homes have very high (> 500) VIF-scores since they exhibit perfect negative correlation (Figure 14). Percentage of households with underaged people also shows a VIF-score of over 50. To produce a reliable model, I eliminated variables with VIF-scores exceeding 5 until no severe multicollinearity was present.

Thereafter, I started removing the variables with the smallest absolute t-values (highest p-values) one by one until AICc could not obtain lower values without adjusted R^2 also decreasing. This process eliminates the unnecessary variables, aiming to simplify and optimize the model. Table 4 illustrates that the final model contains only three variables with R^2 (0.378) and AICc (1684.9)

Number of sports facilities per inhabitant, employment and percentage of children together explain 38% of the variation in the number of sports tweets per 1000 inhabitants that are over 13 years old. Number of sports facilities is statistically most significant variable (highest t-value) but has smaller impact than employment or share of children (smaller estimate). Employment has the largest impact, although almost equal with the percentage of children. The more sports facilities per inhabitant and the higher the employment in a postal code area, the more sports tweets there is estimated to be according to the model. On the contrary, the higher the percentage of children is, the smaller the estimated number of sports tweets is.

I reproduced the analysis with the geotagged tweets only. The OLS model became similar, but it included percentage of rental homes in addition to variables in the original model: sports facilities per inhabitant, employment rate and percentage of children. R^2 was 0.360, so a bit inferior to the model with also the geoparsed tweets, which makes sense since that one has more data to make more accurate predictions.

Table 4. Final OLS model. AICc 1684.9, adjusted R^2 0.378 and p-value <0.001.

Variable	Estimate	Standard Error	t-value	p-value	VIF-score
Intercept	-62.011	36.947	-1.678	0.095	-
Sports facilities	0.765	0.108	7.058	< 0.001	1.050
Employed	2.356	0.637	3.698	< 0.001	1.110
Children	-2.177	0.744	-2.926	0.004	1.105

5.3.3. Lagrange Multiplier tests

Table 6 shows that the Lagrange Multiplier test indicates that OLS is the best fitting model for the analysis, rather than conducting analyses that account for the spatiality of the variables. This can be read from the p-values which are not significant (p-values > 0.05). This is in line with the Moran's I (Figure 11), which shows that the number of sports tweets per 1000 inhabitant that are over 13 years old does not exhibit spatial clustering.

Table 5. Lagrange Multiplier test results with final OLS model.

Test	Value	Degrees of Freedom	p-value
LM Error	0.00157	1	0.968
LM Lag	0.00404	1	0.949

5.4. Hotspot areas by sport categories

To map the distribution of different sports tweets, I used the lemmatized tweet text and coded all keywords related to a sport or sports with a number. For example, all tweets mentioning keywords related to biking (bicycle, bike, biking, cycling, pyöräillä, pyörä, pyöräily, pyöräileminen, jalgratas, jalgrattasõit, rattasõit) were coded as number 3. In total, 2360 tweets contained multiple sports related keywords. In these cases, I chose to map according to the first word. The

same sport was expressed multiple languages or different forms in many of these cases (e.g., *What a good run! #juoksu #running*).

Table 6. Number of tweets per sport and the hotspot areas for each sport.

Class number	Sport(s)	Number of tweets	Percentage of tweets (%)	Hot spot areas
0	Walking, Hiking	2776	13.31	Center[1], Nuuksio, Oulunkylä, Kallio
1	General (sport, workout, gym, sweat)	4062	19.47	Ruoholahti, Center, Tapiola, Leppävaara, Tikkurila
2	Running, jogging	3628	17.39	Center, Pirkkola, Munkkiniemi (Meilahden liikuntakeskus)
3	Biking	1570	7.53	Center, Tammisto, Kallio, Pasila
4	Swimming	537	2.57	Töölö, Center, Vuosaari, Tapiola
5	Skiing	772	3.70	Oittaa, Paloheinä, Tikkurila, Hakunila
6	Skating, ice-hockey	1486	7.12	Pasila, Taka-Töölö, Center
7	Basketball	422	2.02	Lauttasaari, Vallila, Tapiola, Center
8	Football	1139	5.46	Töölö, Tikkurila, Matinkylä
9	Floorball	1744	8.36	Itä-Pasila, Tikkurila, Korso
10	Volleyball, beach volley	98	0.47	Pasila, Töölö, Tikkurila
11	Tennis, badminton, squash, table tennis	499	2.39	Tali (Munkkivuori), Puotinharju (Smash Center Myllypuro)
12	Dance	1367	6.55	Kallio, Center, Eira
13	Yoga	287	1.38	Eira, Center, Kallio, Herttoniemi
14	Sailing, kayaking, canoeing, rowing	419	2.00	Nuuksio, Center, Suomenlinna, coastline

[1] 00100 – Keskusta and Etu-Töölö postal code area.

After coding the sports related words to numbers, I calculated the number and share of tweets per sport, see Figure 15. Categories general sport, running and walking together form half of the tweets. General is the biggest category (containing key words sport, workout, train, training, gym, sweat, sweating), holding almost a fifth of all sports related tweets. Other most prevalent sports are floorball, biking, skating and ice hockey, dance, and football in this order. The sports with lots of tweets roughly compare with the sports with most hobbyist, illustrated in

Figure 15. Skiing and swimming constitute smaller share of tweets than hobbyists. Team sports including floorball, football and ice hockey seem to be overrepresented in the tweets. This might be because the aforementioned sports are popular spectator sports, and the age structure of the hobbyist is skewed towards the young generations that are well represented in Twitter space. The floorball community and clubs seem to be very active in Twitter, tweeting about every match and mentioning the location while this is not the case for football and ice hockey. The number of hobbyists in different sports can be outdated to some extent, since the research is over 10 years old.

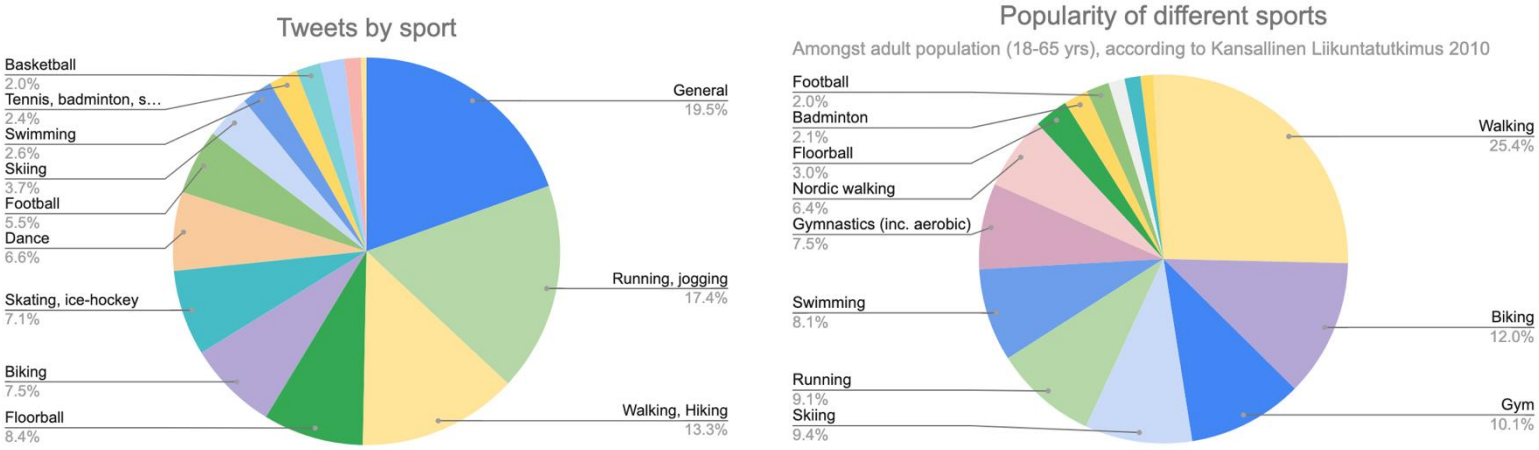


Figure 15. Sports tweets by sport compared to popularity of different sports among adult population according to Kansallinen Liikuntatutkimus.

Some common patterns are seen in many different sports, but some sports also exhibit their own distribution. Most sports have a good portion of the tweets clustered in the Center (00100, Keskusta, Etu-Töölö), see Appendix D and Table 7. Besides the Center, Tikkurila, Leppävaara and Tapiola are common local hotspot areas in Vantaa and Espoo, see Figure 16. All these places serve as local CBDs and have a good collection of sports facilities. Tikkurila has sports park Trio which includes ice halls, swimming hall, football fields (inside and outside), volleyball, floorball, basketball. Leppävaara sports park has track and field stadium, football fields, facilities for floorball, basketball, tennis, skiing, ice hockey during winter and outdoor gym. Tapiola sports park encompasses various sports facilities including but not limited to ice hall, football fields, facilities for basketball, floorball, volleyball, tennis, badminton. Brahe Park

in Kallio also stands out, having outdoor ice facilities in the winter and football facility in the summer.

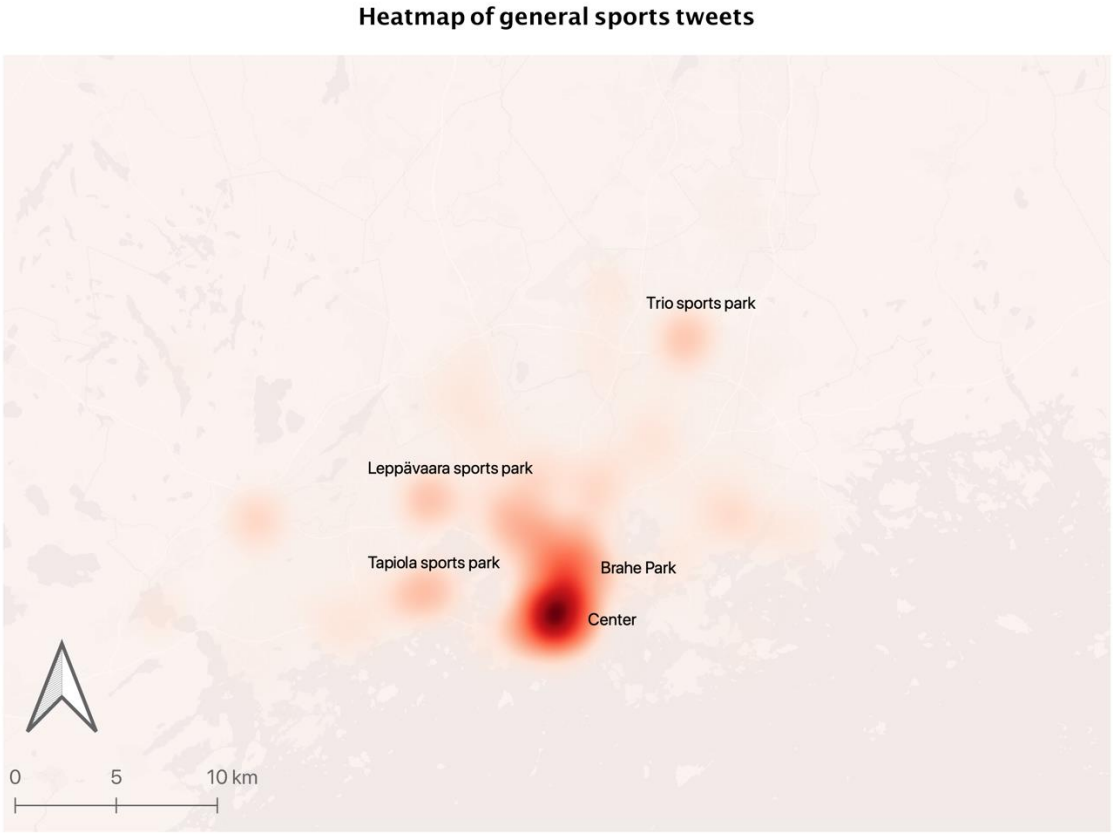


Figure 16. Distribution of general sports tweets are clustered in the center and have local hotspots.

Some sports are more concentrated in the Helsinki city center than others. Dance and yoga tweets are clustered in the center while team sports (basketball, football, floorball, volleyball) and racket sports have more regional hotspots scattered around the area. Skiing, racket sports and water sports show interesting clusters away from the center and around the facilities, see Figure 17.

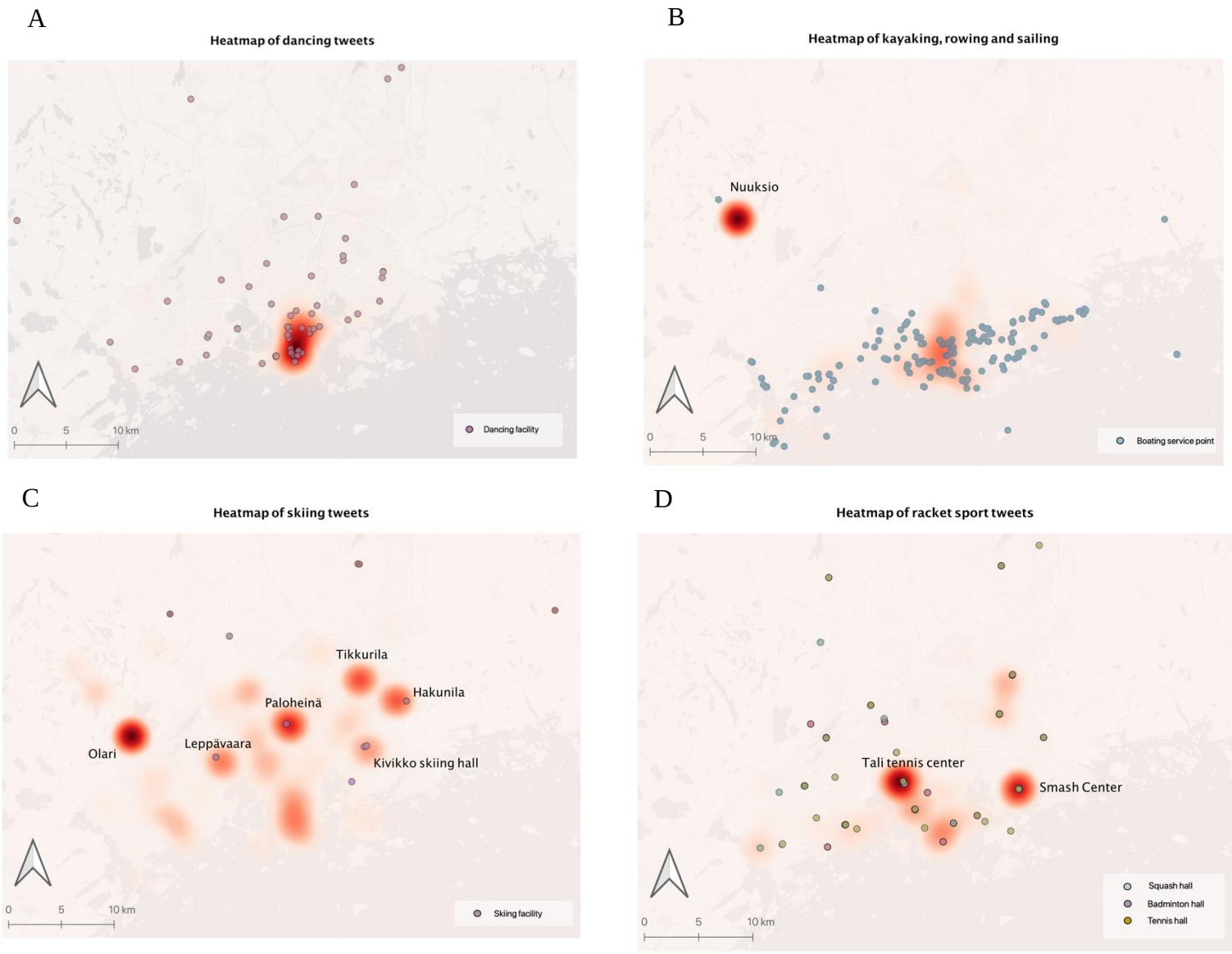


Figure 17. Heatmaps of water sports, skiing and racket sport tweets show clear clustering around the respective facilities while dancing tweets are clustered in the center.

Dancing tweets do not seem to follow the distribution of the facilities as some other sports do (Figure 17A). Of course, there are many dance studios clustered in the center too and these might be the most popular ones. Besides dancing as a sport, my keyword search would have captured tweets that talk about dancing in nightclubs, which are also concentrated in the center. Kayaking, rowing, and sailing tweets show distribution along the coastline and a strong hotspot in Nuuksio (Figure 17B). Nuuksio has at least four firms that provide kayaking and canoeing trips on the lake. Customers, especially tourists, and the firms themselves have apparently been very active tweeting about the experience, maybe for advertisement purposes or not. Reasonably, the water

sport tweets are mostly around waterbodies. Skiing and racket sports tweets are both clustered around the facilities. Skiing heatmap exhibits the most popular places to go ski in the Metropolitan area: Olari, Paloheinä, Hakunila, Kivikko, Leppävaara, and Tikkurila (Figure 17C). While Olari and Tikkurila do not have official skiing facilities, they have good skiing track networks going around the area. The heatmap of racket sports shows two clear hotspots: Tali tennis center and Smash Center Myllypuro (Figure 17D). Tali tennis center in the largest in Europe with 33 tennis courts and 4 badminton courts. Smash center Myllypuro has 14 tennis courts, 12 badminton courts, 5 squash courts, 4 padel courts and 3 table tennis tables. Other smaller racket sport facilities have local hotspots.

5.5. Survey results

5.5.1. Demographics of respondents

The Facebook survey got 344 responses from 77 out of 168 postal code areas, see Figure 18. The missing areas include the neighborhoods in central Helsinki, coastal eastern Helsinki and neighborhoods in northern Espoo and Vantaa. Neighborhoods with most responses include northern and southern Haaga, Lauttasaari and Pähkinärinne. A large majority of the respondents were female (80%). Figure 19 reveals that 30–49 years old have produced the most answers (52%). People under 20 years old are the most underrepresented age group. The demographics of this survey differ from the demographics in the Metropolitan area, at least in terms of age and gender, which is important to keep in mind when interpreting the results in this survey.

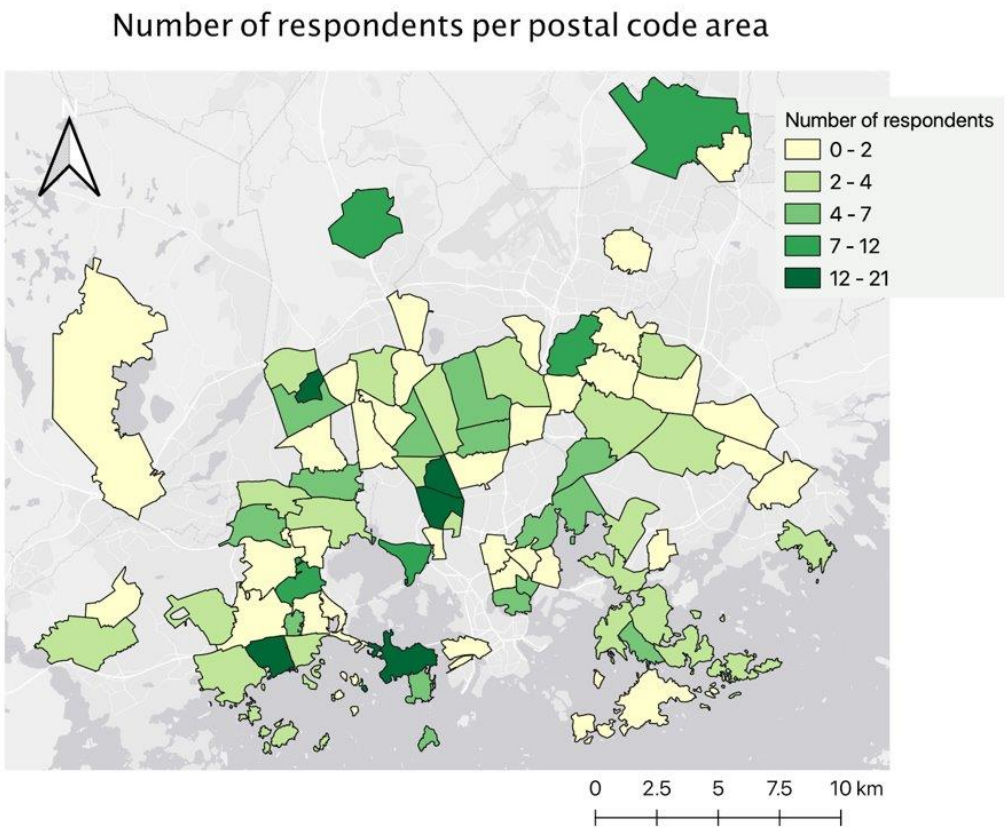


Figure 18. Number of respondents to the survey per postal code area.

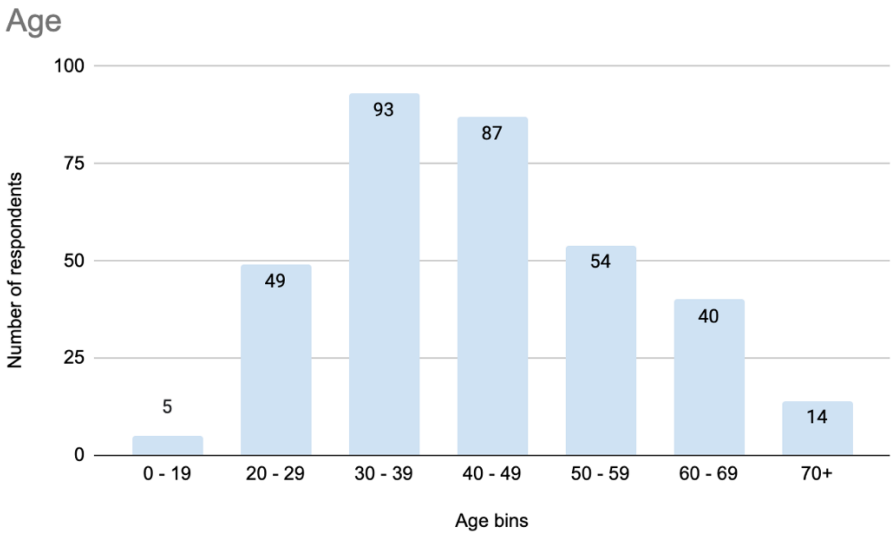


Figure 19. Age distribution of survey respondents.

5.5.2. Physical activity

Over 40% of the survey respondent answered that they do sports activities (including going for a walk or biking to work) daily, see Table 7. Activities that do not require facilities are undertaken more frequently than activities that require facilities, see Table 7. Still, 60% of the respondents do sports that require facilities at least weekly. However, the number of people who rarely do sports that require facilities (30,7%) is much larger than the number of people who rarely do sports that don't require facilities (2,6%).

Table 7. Facebook survey answers: frequency of doing sports, sport with and without facilities.

Frequency	**Sports (%)**	**Sports without facilities (%)**	**Sports with facilities (%)**
Many times a day	8.2	7.2	1.5
Daily	32.7	31.9	1.5
Many times a week	44.6	36.8	29.2
Weekly	10.5	15.8	27.8
Couple of times a month	1.7	5.3	9.4
Rarely	2.0	2.6	30.7

Figure 20. 40 most frequented sports facilities mentioned by survey respondents, produced with Monkeylearn AI.

Figure 21. 40 most frequented neighborhoods to do sports without facilities mentioned by survey respondents, produced with Monkeylearn AI.

Most common sports facilities mentioned by survey respondents are delineated in a word cloud in Figure 20. Swimming halls (uimahalli) around the city seem to be the most important

type of sports facilities followed by sports parks which provide facilities for many sports. The most popular sports facilities according to this survey are Leppävaara swimming hall (n = 23), Mäkelänrinne swimming hall (16) and Pirkkola swimming hall (12).

Figure 21 illustrates the neighborhoods where the respondents do informal sports, such as running, walking, biking, or home exercise in a word cloud. Haaga (n = 37), Keskuspuisto (35), and Tapiola (26) were among the most frequently mentioned locations. This is strongly affected by where the respondents live. Haaga (southern and northern) has the most respondents, and with Keskuspuisto (central park) being in the immediate proximity to Haaga, the result is quite expected.

I mapped the average frequency of sports activities by neighborhoods, see Figure 22. All areas with the average of "many times a day" (like Nupuri-Nuuksio) have only one respondent. From the areas that have over five respondents, the most active are Vattuniemi, Pähkinärinne and Kivistö.

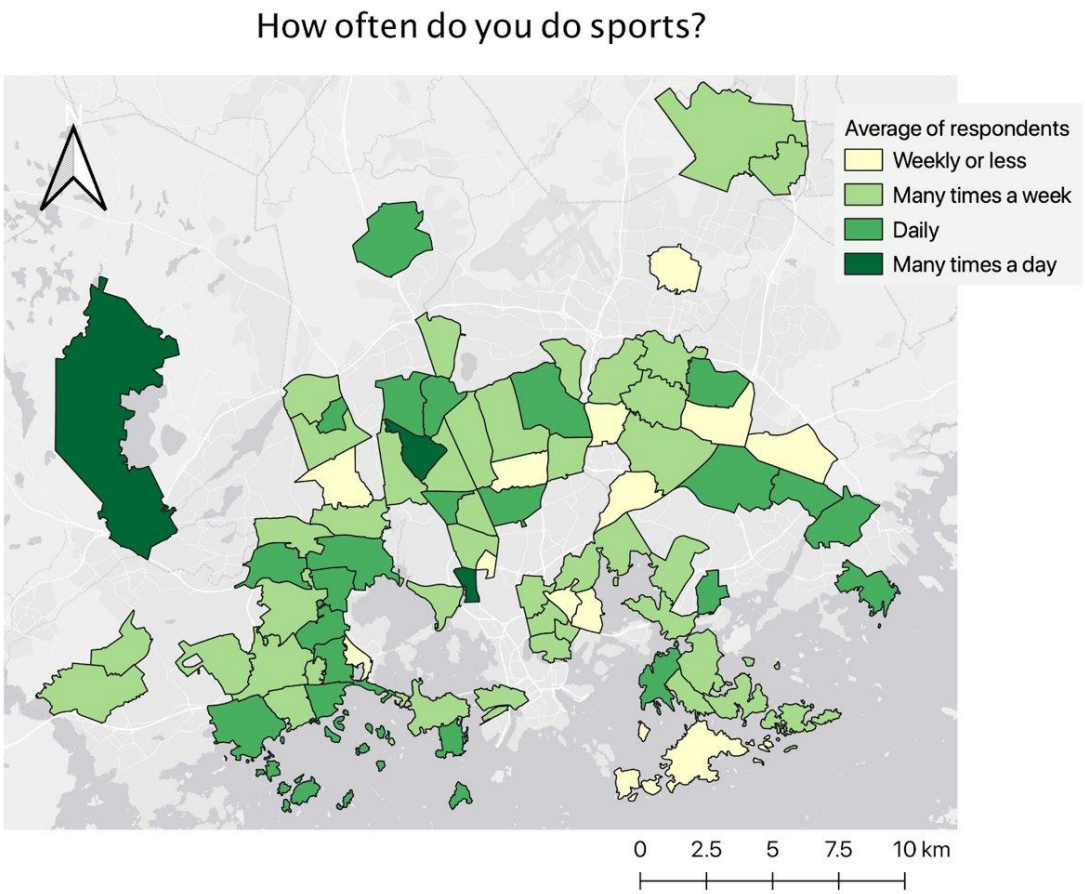


Figure 22. Average frequency of sports activities by neighborhoods.

5.5.3. Social media and sports

Majority of the respondents, 59%, never or rarely post about sports to social media, see Figure 23. However, when respondents post, the most popular social media platforms are Facebook and Instagram, see Figure 24. Although, Facebook is probably overrepresented in here since the survey was conducted on Facebook, and therefore it reaches the people who are active on Facebook neighborhood groups. A vast majority of the sports posts are about people themselves doing sports rather than posting about professional sports where they did not take part in. 42% of the respondents answered that none of their sports related posts concern situations where they are not doing sports themselves. Altogether, people estimated that 20% of the posts are from situations where they are not being active, like from sports competitions where they are as spectators of fans.

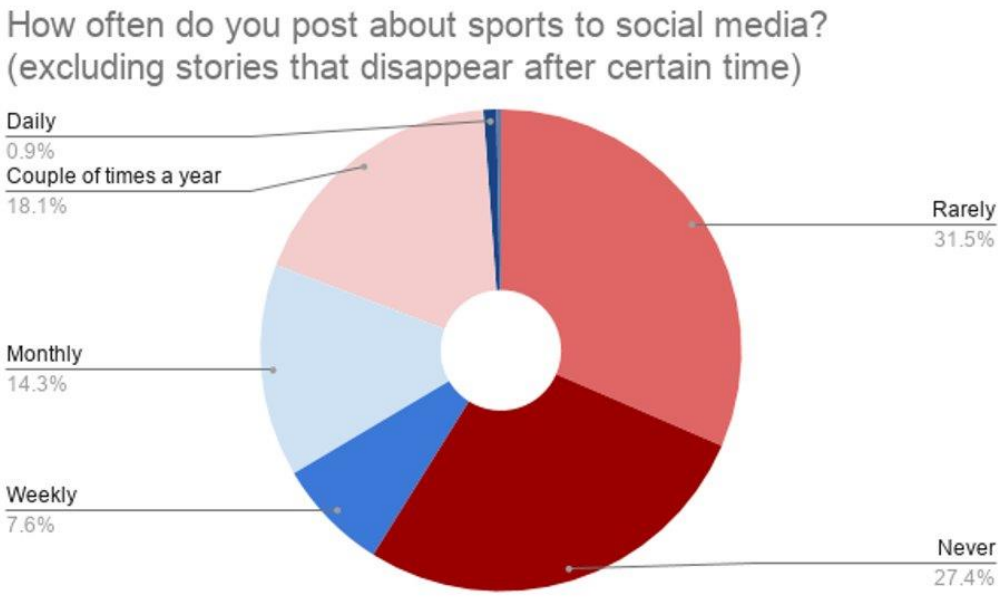


Figure 23. Frequency of posting about sports to social media.

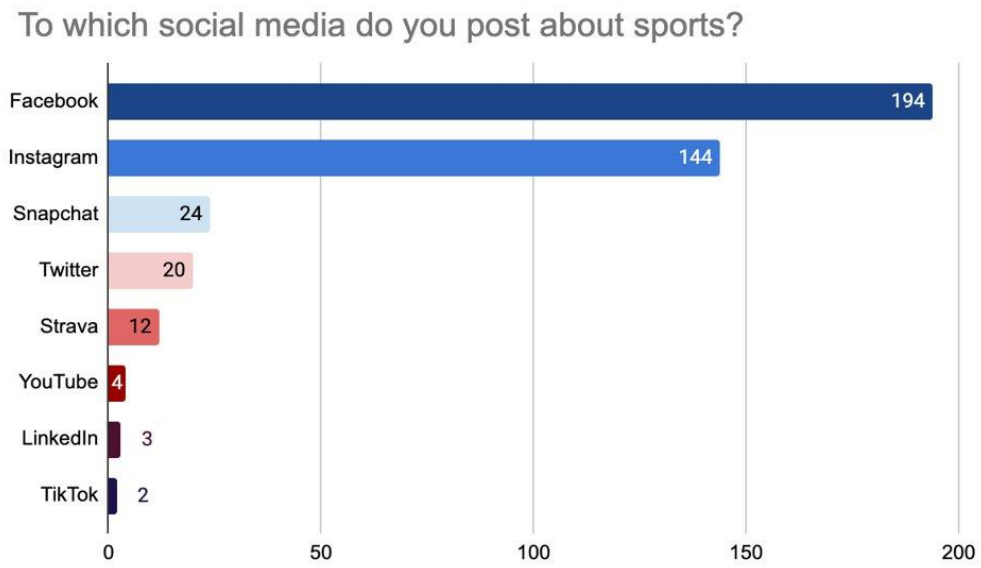


Figure 24. Social medias where the respondents post about sports.

People have different incentives to post about sports. Natural phenomena, like sunset, got the most answers in this multiple-choice question (Figure 25). Other factors that had been named as options are in the order of popularity: the sport itself, who the sports are done with, competitions or events, sports results, and trendiness of the sport. Other factors that emerged multiple times in the "other, what" field are a good feeling (n=14) and kids (n=4).

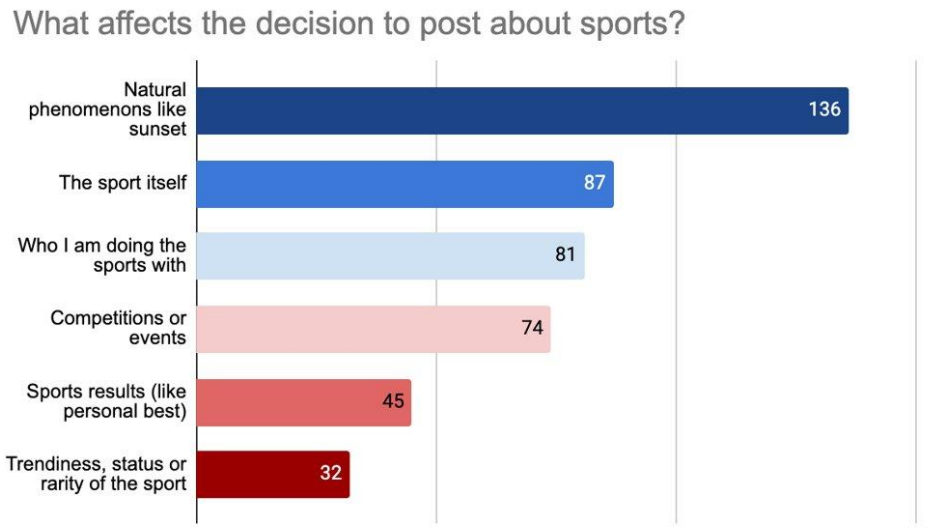


Figure 25. Factors that affcet the decision to post about sports.

Figure 26 illustrates the answers to question “Are you more likely to post about certain sports? If yes, which?” in Finnish. Besides “yes” (7%) and “no” (12%) answers, cross-country skiing (16%) was mentioned the most frequently, followed by running (11%, including also jogging and trail running), football (10%), and cycling (8%). Some of the popularity of skiing could be explained by the time when the survey was carried out. It was a snowy February when all the ski tracks where open and in good condition after a year with almost no snow in the Metropolitan Area.

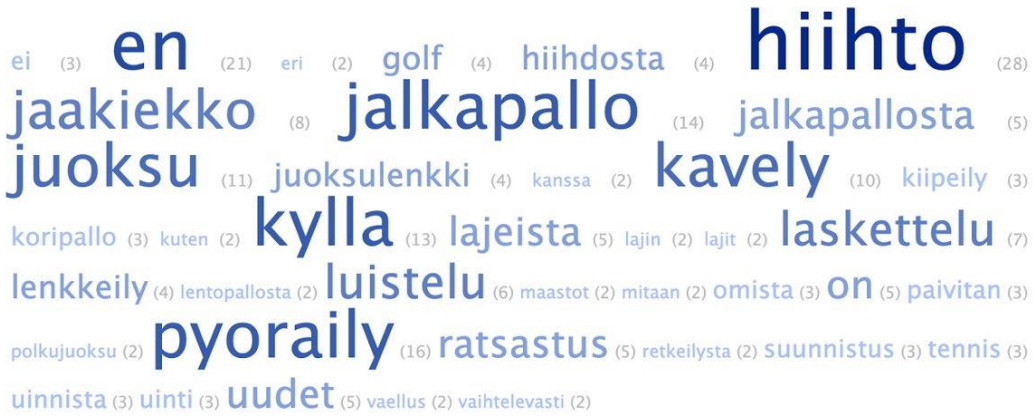


Figure 26. Word cloud of the answers to a question: Are you more likely to post about certain sports? If yes, which?

Spearman correlation coefficients between variables in the data were weak in general. However, statistically significant correlation was found between frequency of doing sports and frequency of posting to social media about sports ($r= 0.21$, $p< 0.001$). Income has a similar positive effect on posting about sports ($r= 0.21$, $p<0.001$). Doing sports with facilities also increased the probability of posting to social media about sports ($r=0.19$, $p< 0.001$). Other correlations found in the data indicate that both education and income increase the frequency of sports practiced with facilities ($r= 0.18$, $p= 0.001$ and $r= 0.13$, $p= 0.021$) but have no effect to the sports done without facilities. Age does not affect the frequency of doing sports or the frequency of posting sports related content. The number of sports facilities (absolute nor per person) in a post code area had no effect the frequency of sports activities (with or without facilities) practiced in the area nor the sports related posts according to this survey.

5.5.4. Validation of Twitter findings with the survey findings

As mentioned in the Introduction, the results of this survey are meant to validate the results from Twitter analysis. Therefore, in this section, I compare the results of these two analyses and assess where the survey results may and may not be sufficient validation for Twitter analyses. Besides reflecting results from two different social media platforms, Twitter and Facebook, the Twitter results are aggregated to postal code areas whereas the Facebook survey results can also be interpreted on individual level. Some patterns may be visible on individual level but faded out on neighborhood level.

The survey does not provide enough data to assess and validate the frequency of sports activities in different parts of the Metropolitan area. This is because less than half of the postal code areas had any respondents and most of the areas with respondents included less than five responses, which cannot be considered representative of the area. The lack of data is probably the reason why no statistically significant correlation could be seen between any of the variables when the survey data was aggregated to postal code areas. Therefore, the survey data can not verify the positive relationship between number of sports facilities per person in an area and frequency of sports practice.

However, the number of responses is sufficient to explore correlations on individual level. Contrary to the Twitter results, survey results find a correlation between both education and income and frequency of sports practice with facilities. High income also increases the probability of posting to social media about sports. In the Twitter analysis, it seemed that income and education of the postal code area have no correlation with the number of tweets per inhabitant in the area. It may well be that this relationship is only visible on the individual level. On the Twitter analysis, it was found that percentage of children and employment rate also affect the number of tweets. The survey did not include questions of either since it was done beforehand. However, on an individual level, income and employment are often heavily linked.

Sports mentioned in tweets (Figure 15) and the sports that people are likely to post about (Figure 26), exhibit mostly similar patterns. Skiing is overrepresented in the Facebook survey, probably due to seasonal reasons. Floorball and dance, on the contrary, are underrepresented in the Facebook survey. The survey also recorded important information that many people rarely or never share their sports activities in social media (Figure 23) and that from the ones who share,

only 8% choose Twitter as a platform compared to 77% who share on Facebook and 57% on Instagram (Figure 24).

6. Discussion

Four major findings emerged from this study were that 1) sports tweets have similar distribution as population, 2) sports facilities per inhabitant, employment rate, and proportion of children are the best predictors of the number of sports tweets per inhabitant, 3) geoparsed tweets produced valuable extra data to the point that it changed the analysis outcome and 4) before applying social media research on physical activities, the connection between social media posts and real-life physical activity needs to be better established by a large-scale comparative study despite the promising results of this study.

6.1. Where do people tweet about sports?

Sports related tweets are distributed largely at the same places where people are living, except for a few outliers with more special environments for sports. This is proved by having moderate spatial autocorrelation (Moran's I: 0.317) in the absolute number of sports tweets but random spatial distribution (Moran's I: 0.024) in the sports tweets per 1000 inhabitants that are over 13 years old. The largest cluster of tweets is in the center of Helsinki. Nuuksio, Paloheinä and Tikkurila are examples of local hotspots, probably because they have sports places that attract people also from further away. Nuuksio is a national park that has many hiking trails, kayaking and swimming options in the summer and skiing during winter. Paloheinä provides good facilities for skiing, running, golf and an ice hall. Tikkurila has Trio Sports Park which includes ice halls, swimming hall, football fields, volleyball, floorball, and basketball.

The Ordinary Least Squares regression showed that the number of sports facilities per 1000 inhabitants is significantly correlated with the number of tweets per inhabitant. The same can be interpreted from the heatmaps of tweets which often have hotspots where the respective sports facilities are. Most sports have the largest accumulation of tweets in the center of Helsinki, although I have removed all tweets that were geoparsed to the coordinates of Helsinki. The

hotspot in the center probably consists of tweets geotagged to the general coordinates of Helsinki, although they might be posted from different sides of Helsinki in reality. Facebook-survey answers fail to show statistically significant correlation with the number of sports facilities and sports frequency, but this might be due to too low number of respondents and many postal code areas were not present in the survey data at all.

Interestingly, some sports are more clustered in the center of Helsinki than others. Skiing and racket sports are less clustered in the center and more clustered around the respective facilities. Naturally, the importance of facilities and therefore the spatial accessibility to them varies by sport as Karusisi et al. (2013) have concluded. Although, Karusisi et al. (2013) found that for swimming, the importance of facilities is greater than for any other sport, for example racket sports. In my analysis, tweets of both racket sports and swimming were concentrated around facilities but racket sports more than swimming. My analysis does not measure from how far people travel to these facilities, but the importance of facilities can be seen in to what extent the tweeting locations align with the facility locations. However, when dealing with social media data, factors such as the precision of the geotags can play a part in the analysis outcome. There can also be cultural and environmental differences, since some of the hotspots of the swimming tweets are from places with natural water bodies like Nuuksion Pitkäjärvi and Kivenlahti, but in Paris where Karusisi et al.'s (2013) study was conducted there does not exist many places to go swim outside swimming halls.

6.2. Which factors affect tweeting about sports?

The correlation matrix (Figure 14) showed that none of the variables that measure income or education correlates with the number of sports tweets per 1000 inhabitants that are over 13 years old. Besides, none of the socio-economic variables correlate with the number of sports facilities per 1000 inhabitants. Both facts demonstrate that the Helsinki Metropolitan area is not spatially segregated in regards of sports on the postal code area level. This could be due to successful mixed housing policy meaning that each area should have a balanced mix of different housing and house tenure types as well as enough suitable housing for families, students, elderly, disabled, and other population groups (City of Helsinki, 2020). Individual sports activity

frequency can vary according to income level or education, but these patterns are not reflected on postal code level. In a study conducted in Paris, Karusisi et al. (2013) found that having higher education promotes doing more sports even when the financial and spatial aspects are controlled for. The survey results present relatively weak (r=0.21) connection between income level and posting frequency about sports. Although, total sports activities affected the frequency of posting to social media about sports just as much as income.

The number of sports facilities per 1000 inhabitants show significant correlation with the sports tweets per 1000 inhabitants that are over 13 years old. Many previous studies have presented that most of the sports activities are done in proximity to home location (Kajosaari & Laatikainen, 2020) and that in some sports, accessibility to facilities can encourage sports activities (Karusisi et al., 2012, 2013). According to my findings, this connection was also found on neighborhood level in the Metropolitan area. This supports the claim that people seek for sports services inside their own postal code area. On an individual level, the connection between sports facilities and sports frequency was not observed but this might be due to insufficient data.

Employment had the strongest effect on sports tweets according to my Ordinary Least Squares model (Table 4). On a contradicting note, Suomi et al. (2012) found that 27% of their survey respondents experienced work or studies as barrier for doing sports. Although, the same survey found that stay-at-home parents exercise 49 minutes less weekly than employed people on average. The people who are not employed (unemployed, retired, students) might be preoccupied by searching for jobs or other duties like parenting or have health problems that make sports difficult. 31% of people stated that an illness or injury is a barrier for doing sports (Suomi et al., 2012).

A third factor affecting the sports tweets in the final OLS model was percentage of children in the area. Also, other variables that reflect the number of children like percentage of households with underaged people and household size correlated strongly negatively with the number of tweets. It might be that after having children, the time for sports and social media decreases or that the content of social media posts involves less sports. According to Suomi et al. (2012), parents under 30 years old exercise on average 25 minutes less weekly than their counterparts who do not have children. As barriers for sports, 36% mentioned lack of time and 19% stated family situation. One explanatory factor could be that people with larger families in

Helsinki Metropolitan area often belong to language minorities which were not included in this analysis and therefore they might not be fully represented (City of Helsinki, 2021).

It should be borne in mind that sports tweets are not a complete nor unproblematic representation of sports activities. For example, the survey answers show that certain factors encourage people to post about certain kind of sports or situations. Therefore, some postal code areas could have a distortedly large proportion of tweets if they have rarer sports facilities where people come from afar and feel like the place are worth posting about. According to the survey results, skiing, running, and football are sports that incentivize most people to post to social media (Figure 27). When comparing to the number of hobbyists, floorball, skating and ice hockey, dance, and racket sports seem to be overrepresented in the tweets. Furthermore, places where organized sports events are held, could be overrepresented if the spectators are posting actively.

Due to lack of comprehensive spatial study about sports activities in Helsinki, the sports tweets may be used as a proxy (secondary source of information) to assess the overall sports activity. The maps of different sports demonstrate that many sports tweets are tightly clustered around the facilities for respective sport, which indicates that the data quality and geolocation precision is sufficient for at least approximation analysis. One possible application is to explore which facilities of the sports are most popular and build more of that kind of facilities. Although one cannot forget about suitable sports facilities for also elderly and other population groups, which are most likely underrepresented in the Twitter data.

The ordinary least squares regression suggests that number of sports facilities per inhabitant, employment and percentage of children predict the number of sports tweets. As the survey results show, exercising frequency and posting frequency have a connection. Thus, the results may be cautiously generalized to concern physical activity. This information can be applied when allocating resources to encourage exercise and build a healthier society. More facilities and sports clubs could be funded in postal code areas with lower employment rate, higher percentage of children and less sports facilities per inhabitant. Free after school sports groups could work as well. However, the connections between physical activity and areal employment rate and percentage of children calls for further individual level research to find and tackle the root causes why areas with a larger share of children and non-employed tweet less

about sports and what kind of connections and barriers for sports related to these areal variables exist on an individual level.

6.3. How can Twitter data be used in sports activities research?

6.3.1. Data requirements and preparation

Quite abundant amount of Twitter data is needed to conduct a research with a limited thematic and spatial scope. In this study, I started off with 38.5 million tweets and after cropping and cleaning the data to fit both the thematic and the spatial scope, the data was narrowed down to 20 599 tweets. It may be discussed whether this is sufficient amount of data for this kind of analysis. Surely, it would be beneficial to have some more data, which would mean collecting and processing perhaps hundreds of millions of tweets. The data preparation needs lemmatization and matching at the minimum, and possibly running a more sophisticated language model to reach more reliable results. This requires some processing power, maybe a virtual machine to externalize the workload, and a well optimized code.

6.3.2. Geoparsing efforts

In this section, I assess what was the significance of the geoparsed tweets to my analysis outcome and quality compared to performing the analysis only with geotagged tweets. I performed the geoparsing in order to acquire more sports related georeferenced content. In the end, geoparsing increased my final data surprisingly little. From my final cleaned sports tweets data, 67% were geotagged (n= 13 746) and 33% geoparsed (n= 6 853). If I had not removed the tweets which were geoparsed to city coordinates, the percentage of geoparsed tweets would have been 57% (n = 17 968).

2.93% of the geotagged tweets were captured with the sports-themed keyword matching. Provided that the same percentage of non-geotagged tweets would be sports related, it indicates that 1.82% of all sports related tweets mentioned a geocodable toponym in the Metropolitan area. According to MacEachren et al. (2011), 10% of the tweets mention a toponym which is possible to geocode. Now close to 2%, seems like a large percentage considering the limited spatial scope

of this study. If the names and nick names of the common sports facilities were included in the gazetteer used, the geoparsed would have probably captured many more tweets.

Liu et al. (2021) performed a similar study with the same data, keywords and spatial scope using a bilingual neural network-based language model with Google geocoding API and captured triple the number of tweets. This highlights the importance of a good language model and not being constrained to the certain place names in a gazetteer but using a larger database for those. This approach would also be able to avoid the kind of geoparsing mistakes that mistake a last name or general noun for a place name (documented in Table 2).

Despite the relatively naïve geoparsing and keyword matching methods I used, I managed to collect more valid sports tweets. Having more tweets generally improves the quality and reliability of the analyses. Moreover, the users who post geotagged tweets have certain kind of profiles and therefore may not be representative of all twitter users as counter to how geotagging users are often generalized in literature (Karami et al., 2021). Hence, having the addition of geoparsed tweets, I was probably able to reach a more diverse group of tweeters who have not shared their location.

6.3.3. Limitations of this study

When using user-generated data, like Twitter data, one of the key epistemological limitations is lack of information on who produced the data, and hence, how representative it is of all the users on the platform (Karami et al., 2021; Ruths & Pfeffer, 2014). The data may be biased by for example gender, age, income, geotag, educational background or the studied platform. Thus, it is important to acknowledge the possible bias during the analysis and when interpreting the results.

As mentioned in the section 6.3.2, the kind of language model in use affects the results and their quality to a great extent. With neural network-based language model, Liu et al. (2021) were able to capture triple the number of sports related tweets from the same data. Also, (transformer based) Named Entity Recognition would have probably avoided many misclassifications if all the language models were trained with Finnish toponyms in the sentences. However, I opted for using simple Named Entity Matching approach in the interest of

time and scope of my thesis. Despite NEM being a suboptimal model, it fulfilled its purpose by producing enough geoparsed tweets, and the errors were contained by manual inspection.

Data aggregation and normalization may also cause bias and alter the analysis outcome (Figures 10 and 11) (Foster, 2019). Data aggregation to postal code areas may raise issues like Modifiable Areal Unit Problem and Uncertain Geographical Context problem due to irregular shapes and different sizes and populations of the areas (Fotheringham & Wong, 1991; Kwan, 2012; Openshaw, 1984). The bias of varying shapes and sizes can be eliminated by aggregation to grid cells. Although, the size of grid cells would also affect the analysis outcome and the cells would have different amount of population. Therefore, I normalized the data with population and kept postal code areas as units of aggregation. Most of the locations in GeoNames refer to a neighborhood, so the postal code areas that follow the neighborhood borders are more meaningful aggregation units than arbitrary grid cells.

Performing the socio-economic analysis on postal code level has its shortcomings also. The analysis inherently assumes that the people who live in that postal code area have produced the sports-related tweets in the area. Even though most sports activities are practiced within a mile radius from the home location (Kajosaari & Laatikainen, 2020), some areas are clear exceptions. For example, to the Nuuksio National Park, people come from all over the Metropolitan Area, Finland and even from abroad. Thus, the comparison of the number of sports tweets and the socio-economic variables of the postal code area does not make sense in the case of Nuuksio since majority of the tweets are probably generated by people living elsewhere.

6.3.4. Recommendations for future research

For future research, a similar kind of study with improved language model to catch more tweets would be of good use. More efficient and accurate geoparsing can be executed with transformer Named Entity Recognition or Neural Network based model and Google Geocoding API or Open Street Map database (Liu et al., 2021; Middleton et al., 2018).

As mentioned in the literature review, Twitter may not be the best platform for sports activities research since it focuses on topical and professional discussion. Therefore, it would be beneficial to capture the patterns in other social media and sports platforms as well. A mix of different social media platforms will most probably provide the most accurate representation of

reality since all social media platforms have their own demographics (Tenkanen et al., 2017). A comparison between number of posts and number of users would also be helpful to identify if a large proportion of the data is created by very active "super users" which may cause biases (Heikinheimo et al., 2020b).

Before applying the results of social media-based sports activity research on a large scale to society, it would be crucial to have more studies validating the connection between physical activity patterns in social media and in real life. This calls for studies that would compare spatially self-reported or measured sports activity and patterns of sports activity related posts in social media. After verifying this linkage, investigating, and defining the biases in it, social media research would become more reliable to the point that policy suggestions and location allocation of sports facilities could be done based on it. This would be on the condition that social media data turns out to be a reliable indicator of real-life sports activity patterns.

7. Conclusions

This thesis aimed to identify where people tweet about sports in Helsinki Metropolitan Area, which socio-economic factors affect the number of tweets and how Twitter data can be used in physical activity research. I found that the sports-related tweets follow similar distribution as population. The largest hotspot is the center of Helsinki, and smaller local hotspots are Tapiola, Leppävaara, Tikkurila and Pasila. Some sports, like skiing and racket sports, are distributed around the respective facilities, while some, like dance, are not.

I aggregated the sports-related tweets to postal code areas and studied the correlations to socio-economic variables in postal code areas. Number of sports facilities per inhabitant, employment rate and proportion of children in the area proved out to be the best predictors of the number of sports tweets per inhabitant in Ordinary Least Squares regression analysis. The sports tweets did not exhibit significant correlation with income or education variables on postal code area level. Although, on an individual level, income and education showed a weak correlation with physical activity that requires facilities and income predicts sports-related social media activity as much as doing sports according to the Facebook survey I conducted.

My results seem promising, suggesting that Twitter data can be used as a proxy for physical activities in the lack of official data. The crux to account for in similar studies is

sufficient data, because restricting both spatial and thematic scope reduces the original data significantly. Furthermore, geoparsing the tweets that did not contain geotag yielded more of valuable data and probably captured the activity patterns of different kind of Twitter users. Even better geoparsing results can be attained with more sophisticated language models like transformer-based Named Entity Recognition or Neural-Network-based model. However, before making policy suggestions or other real-life decisions based on social media research, large-scale comparative studies between social media data and measured or self-reported physical activity are called for. To avoid a situation whereby a society would be built to meet just the most vocal social media user's needs, it is important to identify the biases in social media data and to acknowledge which population groups are left out or overrepresented.

Acknowledgements

I would like to express my gratitude to all members and funders of YLLI and Digital Geography Lab research groups. Special thanks belong to my supervisors Petteri Muukkonen and Pengyuan Liu for commenting and supporting all along the way. I would also like to thank Tuomas Väisänen for helping me put up a virtual computing environment, Kerli Müürisepp for the Estonian translation of sports related keywords, and Maaria Koivisto for invaluable comments on readability and scientific writing.

References

Sports Act, 1:5 (2015) (testimony of 390).

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Allem, J. P., Dharmapuri, L., Leventhal, A. M., Unger, J. B., & Cruz, B. (2018). Hookah-related posts to twitter from 2017 to 2018: Thematic analysis. *Journal of Medical Internet Research*, *20*(11), 1–7. https://doi.org/10.2196/11669

Alotaibi, S., Mehmood, R., Katib, I., Rana, O., & Albeshri, A. (2020). Sehaa: A big data analytics tool for healthcare symptoms and diseases detection using twitter, apache spark, and machine learning. *Applied Sciences (Switzerland)*, *10*(4). https://doi.org/10.3390/app10041398

Anselin, L. (1988). Lagrange Multiplier Test Diagnostics for Spatial Dependence and Spatial Heterogeneity. *Geographical Analysis*, *20*(1), 1–17. https://doi.org/10.1111/j.1538-4632.1988.tb00159.x

Anselin, L. (1995). Local Indicators of Spatial Association—LISA. *Geographical Analysis*, *27*(2), 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

Armila, P. (2020). Syrjäseudun nuoret ja liikuntaharrastukset: Tanssii mummojen kanssa. *Eriarvoisuuden Kasvot Liikunnassa*, 142–160.

Asefi, A., & Ghanbarpour Nosrati, A. (2020). The spatial justice in the distribution of built outdoor sports facilities. *Journal of Facilities Management*, *18*(2), 159–178. https://doi.org/10.1108/JFM-09-2019-0051

Bennie, J. A., Pedisic, Z., Suni, J. H., Tokola, K., Husu, P., Biddle, S. J. H., & Vasankari, T. (2017). Self-reported health-enhancing physical activity recommendation adherence among 64,380 finnish adults. *Scandinavian Journal of Medicine and Science in Sports*, *27*(12), 1842–1853. https://doi.org/10.1111/sms.12863

Bergsgard, N. A., Borodulin, K., Fahlen, J., Høyer-Kruse, J., & Iversen, E. B. (2019). National structures for building and managing sport facilities: a comparative analysis of the Nordic countries. *Sport in Society*, *22*(4), 525–539. https://doi.org/10.1080/17430437.2017.1389023

Billaudeau, N., Oppert, J. M., Simon, C., Charreire, H., Casey, R., Salze, P., Badariotti, D., Banos, A., Weber, C., & Chaix, B. (2011). Investigating disparities in spatial accessibility to

and characteristics of sport facilities: Direction, strength, and spatial scale of associations with area income. *Health and Place*, *17*(1), 114–121. https://doi.org/10.1016/j.healthplace.2010.09.004

Blok, A. (2020). Urban green gentrification in an unequal world of climate change. *Urban Studies*, *57*(14), 2803–2816. https://doi.org/10.1177/0042098019891050

Bornmann, L., Haunschild, R., & Patel, V. M. (2020). Are papers addressing certain diseases perceived where these diseases are prevalent? The proposal to use Twitter data as social-spatial sensors. *PLoS ONE*, *15*(11 November), 1–22. https://doi.org/10.1371/journal.pone.0242550

Borodulin, K., Jousilahti, P., Mäki-Opas, T., Männistö, S., Valkeinen, H., & Wennman, H. (2018). Fyysinen aktiivisuus ja istuminen. In P. Koponen, K. Borodulin, A. Lundqvist, K. Sääksjärvi, & S. Koskinen (Eds.), *Terveys, toimintakyky ja hyvinvointi Suomessa FinTerveys 2017 -tutkimus* (pp. 38–42). National Institute for Health and Welfare (THL). http://www.julkari.fi/bitstream/handle/10024/90832/Rap068_2012_netti.pdf?sequence=1

Borodulin, Katja, Harald, K., Jousilahti, P., Laatikainen, T., Männistö, S., & Vartiainen, E. (2016). Time trends in physical activity from 1982 to 2012 in Finland. *Scandinavian Journal of Medicine and Science in Sports*, *26*(1), 93–100. https://doi.org/10.1111/sms.12401

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication and Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Carter, S., Tsagkias, M., Weerkamp, W., & Expansion, I. Q. (2011). Semi-Supervised Priors for Microblog Language Identification. *Month*. http://wouter.weerkamp.com/downloads/dir2011-lid.pdf

Chacón-Borrego, F., Corral-Pernía, J. A., Martínez-Martínez, A., & Castañeda-Vázquez, C. (2018). Usage behaviour of public spaces associated with sport and recreational activities. *Sustainability (Switzerland)*, *10*(7), 1–9. https://doi.org/10.3390/su10072377

Cheng, T. Y. M., Liu, L., & Woo, B. K. P. (2018). Analyzing twitter as a platform for Alzheimer-related dementia awareness: Thematic analyses of tweets. *JMIR Aging*, *1*(2), 1–7. https://doi.org/10.2196/11542

Chmait, N., Westerbeek, H., Eime, R., Robertson, S., Sellitto, C., & Reid, M. (2020). Tennis

influencers: The player effect on social media engagement and demand for tournament attendance. *Telematics and Informatics*, *50*(November 2019), 101381. https://doi.org/10.1016/j.tele.2020.101381

Cole, H. V. S., Lamarca, M. G., Connolly, J. J. T., & Anguelovski, I. (2017). Are green cities healthy and equitable? Unpacking the relationship between health, green space and gentrification. *Journal of Epidemiology and Community Health*, *71*(11), 1118–1121. https://doi.org/10.1136/jech-2017-209201

D'Alessandro, D., Buffoli, M., Capasso, L., Fara, G. M., Rebecchi, A., & Capolongo, S. (2015). Green areas and public health: Improving wellbeing and physical activity in the urban context. *Epidemiologia e Prevenzione*, *39*(4), 8–13.

de Andrade, S. C., Restrepo-Estrada, C., Nunes, L. H., Rodriguez, C. A. M., Estrella, J. C., Delbem, A. C. B., & Porto de Albuquerque, J. (2021). A multicriteria optimization framework for the definition of the spatial granularity of urban social media analytics. *International Journal of Geographical Information Science*, *35*(1), 43–62. https://doi.org/10.1080/13658816.2020.1755039

de Bruijn, J., de Moel, H., Jongman, B., Wagemaker, J., & Aerts, J. C. J. H. (2017). TAGGS: Grouping Tweets to Improve Global Geotagging for Disaster Response. *Natural Hazards and Earth System Sciences Discussions*, 1–22. https://doi.org/10.5194/nhess-2017-203

de Leeuw, E. (2017). From Urban Projects to Healthy City Policies. In E. de Leeuw & J. Simos (Eds.), *Healthy Cities* (pp. 407–438). Springer.

Dodge, Y. (2008). *The Concise Encyclopedia of Statistics*. Springer.

Eriksson, S., Saukkonen, E., Mietola, R., & Katsui, H. (2020). Vaikeimmin vammaisten nuorten liikunnan harrastaminen ja eriarvoinen osallisuus. *Eriarvoisuuden Kasvot Liikunnassa*, 94–114.

Foster, M. (2019). Statistical Mapping (Enumeration, Normalization, Classification). *The Geographic Information Science & Technology Body of Knowledge*, *2nd Quarte*. https://doi.org/10.22224/gistbok/2019.2.2

Fotheringham, A. S., & Wong, D. W. S. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment & Planning A*, *23*(7), 1025–1044. https://doi.org/10.1068/a231025

Gelernter, J. (2013). *Cross-lingual geo-parsing for non-structured data*.

Gelernter, J., & Balaji, S. (2013). An algorithm for local geoparsing of microtext. *GeoInformatica*, *17*(4), 635–667. https://doi.org/10.1007/s10707-012-0173-8

GeoNames. (n.d.). *GeoNames*. Retrieved March 11, 2021, from https://www.geonames.org/

Gharaveis, A. (2020). A systematic framework for understanding environmental design influences on physical activity in the elderly population: A review of literature. *Facilities*, *38*(9–10), 625–649. https://doi.org/10.1108/F-08-2018-0094

Gould, K., & Lewis, T. (2012). The environmental injustice of green gentrification. In *Green gentrification* (Issue January). Routledge. http://books.google.com/books?hl=en&lr=&id=S-YNs4ih1FUC&oi=fnd&pg=PA113&dq=The+Environmental+Injustice+of+Green+Gentrification&ots=OXhzOVCkOq&sig=xcipOgjk6xV8hz8obFs0Lz7-cs0%5Cnhttp://books.google.com/books?hl=en&lr=&id=S-YNs4ih1FUC&oi=fnd&pg=PA113&dq=The

Government, F. (2019). *3.7.1 Youth, culture and sport*. Government Programme. https://valtioneuvosto.fi/en/marin/government-programme/youth-culture-and-sport

Graham, M., Hale, S. A., & Gaffney, D. (2014). Where in the World Are You? Geolocation and Language Identification in Twitter. *Professional Geographer*, *66*(4), 568–578. https://doi.org/10.1080/00330124.2014.907699

Gritta, M., Pilehvar, M. T., & Collier, N. (2020). A pragmatic guide to geoparsing evaluation: Toponyms, Named Entity Recognition and pragmatics. *Language Resources and Evaluation*, *54*(3), 683–712. https://doi.org/10.1007/s10579-019-09475-3

Gritta, M., Pilehvar, M. T., Limsopatham, N., & Collier, N. (2018). What's missing in geographical parsing? *Language Resources and Evaluation*, *52*(2), 603–623. https://doi.org/10.1007/s10579-017-9385-8

Group, S. N. (2020). *Available Models & Languages*. http://cmip-pcmdi.llnl.gov/cmip5/terms.html

Grubesic, T. H. (2008). Zip codes and spatial analysis: Problems and prospects. *Socio-Economic Planning Sciences*, *42*(2), 129–149. https://doi.org/10.1016/j.seps.2006.09.001

Hamstead, Z. A., Fisher, D., Ilieva, R. T., Wood, S. A., McPhearson, T., & Kremer, P. (2018). Geolocated social media as a rapid indicator of park visitation and equitable park access. *Computers, Environment and Urban Systems*, *72*(February), 38–50. https://doi.org/10.1016/j.compenvurbsys.2018.01.007

Harjunen, H. (2020). Lihavien ihmisten liikunta ja sen esteet . *Eriarvoisuuden Kasvot Liikunnassa*, 50–68.

Haustein, S., Koglin, T., Nielsen, T. A. S., & Svensson, Å. (2020). A comparison of cycling cultures in Stockholm and Copenhagen. *International Journal of Sustainable Transportation*, *14*(4), 280–293. https://doi.org/10.1080/15568318.2018.1547463

Heikinheimo, V., Minin, E. Di, Tenkanen, H., Hausmann, A., Erkkonen, J., & Toivonen, T. (2017). User-generated geographic information for visitor monitoring in a national park: A comparison of social media data and visitor survey. *ISPRS International Journal of Geo-Information*, *6*(3). https://doi.org/10.3390/ijgi6030085

Heikinheimo, V., Tenkanen, H., Bergroth, C., Järv, O., Hiippala, T., & Toivonen, T. (2020a). Understanding the use of urban green spaces from user-generated geographic information. *Landscape and Urban Planning*, *201*(May), 103845. https://doi.org/10.1016/j.landurbplan.2020.103845

Heikinheimo, V., Tenkanen, H., Bergroth, C., Järv, O., Hiippala, T., & Toivonen, T. (2020b). Understanding the use of urban green spaces from user-generated geographic information. *Landscape and Urban Planning*, *201*(January), 103845. https://doi.org/10.1016/j.landurbplan.2020.103845

Helldán, Anni, & Helakorpi, S. (2015). *Suomalaisen aikuisväestön terveyskäyttäytyminen ja terveys, kevät 2014* (A. Helldán & S. Helakorpi (Eds.)). National Institute for Health and Welfare.

Helsingin kaupunkisuunnitteluvirasto. (2014). Pyöräilyn hyödyt ja kustannukset Helsingissä. In *Helsingin kaupunkisuunnitteluviraston liikennesuunnitteluosaston selvityksiä 2014:5*. http://www.hel.fi/hel2/ksv/julkaisut/los_2014-5.pdf

Helsinki, C. of. (2021). *Syntyvyys vaihtelee kieliryhmittäin _ Ulkomaalaistaustaiset Helsingissä*. https://ulkomaalaistaustaisethelsingissa.fi/fi/syntyvyys

Helsinki, Ci. of. (2020). *Asumisen ja siihen liittyvän maankäytön toteutusohjelma 2020*. https://www.hel.fi/static/kanslia/Julkaisut/Kotikaupunkina-Helsinki/2020/Asumisen_ja_maankayton_ohjelma_2020.pdf

Helsinkiläisten liikkumistottumukset 2019. (2020).

Higgs, G., Langford, M., & Norman, P. (2015). Accessibility to sport facilities in Wales: A GIS-based analysis of socio-economic variations in provision. *Geoforum*, *62*, 105–120.

https://doi.org/10.1016/j.geoforum.2015.04.010

Hiippala, T., Väisänen, T., Toivonen, T., & Järv, O. (2020). Mapping the languages of Twitter in Finland : Richness and diversity in space and time. *Neuphilologische Mitteilungen*, *121*(1), 12–44. https://tuhat.helsinki.fi/ws/portalfiles/portal/157585891/hiippalaetal2020_nm.pdf

Hochmair, H. H., Juhász, L., & Cvetojevic, S. (2018). Data quality of points of interest in selected mapping and social media platforms. *Lecture Notes in Geoinformation and Cartography*, *208669*, 293–313. https://doi.org/10.1007/978-3-319-71470-7_15

Hu, Y., & Wang, R. Q. (2020). Understanding the removal of precise geotagging in tweets. *Nature Human Behaviour*, *4*(December), 2019–2021. https://doi.org/10.1038/s41562-020-00949-x

Huang, B., & Carley, K. M. (2019). A large-scale empirical study of geotagging behavior on twitter. *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2019*, 365–373. https://doi.org/10.1145/3341161.3342870

Huang, Q., & Wong, D. W. S. (2016). Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us? *International Journal of Geographical Information Science*, *30*(9), 1873–1898. https://doi.org/10.1080/13658816.2016.1145225

Hughes, A., & Wojcik, S. (2019a, April 24). How Americans use Twitter. *Pew Research Center*.

Hughes, A., & Wojcik, S. (2019b, April 24). Sizing up twitter. *Pew Research Center*. https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/

Hutcheson, G. D., & Sofroniou, N. (1999). Ordinary Least-Squares Regression. In *The Multivariate Social Scientist: Introductory Statistics Using Generalized Linear Models* (1., pp. 55–112). SAGE Publications Ltd.

Jääskeläinen, S., Mäki, P., Mölläri, K., & Mäntymaa, P. (2020). *Lasten ja nuorten ylipaino ja lihavuus 2018 Joka neljäs poika ja lähes joka viides tyttö on ylipainoinen tai lihava*. http://www.julkari.fi/bitstream/handle/10024/138015/Tilastoraportti_lasten_nuorten_lihavuus_20190417_lopullinen_PDF.pdf?sequence=2&isAllowed=y

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *Springer Texts in Statistics An Introduction to Statistical Learning - with Applications in R*.

Jessop, B., Brenner, N., & Jones, M. S. (2008). Theorizing sociospatial relations. *Environment and Planning D: Society and Space*, *26*(3), 389–401. https://doi.org/10.1068/d9107

Jyväskylä, U. of. (2020a). *Lipas-järjestelmän esittely — Liikuntatieteellinen tiedekunta*. https://www.jyu.fi/sport/fi/yhteistyo/lipas-liikuntapaikat.fi/esittely-2

Jyväskylä, U. of. (2020b). *Lipas Liikuntapaikat*. https://www.jyu.fi/sport/fi/yhteistyo/lipas-liikuntapaikat.fi

Jyväskylä, U. of. (2020c). *Lipas tarjoaa avointa dataa liikuntapaikoista*. https://www.jyu.fi/sport/fi/yhteistyo/lipas-liikuntapaikat.fi/rajapinnat-ja-ladattavat-aineistot/lipas_avoindata_kuvaus.pdf

Jyväskylä, U. of. (2020d). *Maantieteilijät ja liikuntatieteilijät tutkimaan liikunnan yhdenvertaisuutta lähiöissä — Jyväskylän yliopisto*. https://www.jyu.fi/fi/ajankohtaista/arkisto/2020/06/maantieteilijat-ja-liikuntatieteilijat-tutkimaan-liikunnan-yhdenvertaisuutta-lahioissa

Kabisch, N., & Haase, D. (2014). Green justice or just green? Provision of urban green spaces in Berlin, Germany. *Landscape and Urban Planning*, *122*, 129–139. https://doi.org/10.1016/j.landurbplan.2013.11.016

Kajosaari, A., & Laatikainen, T. E. (2020). Adults' leisure-time physical activity and the neighborhood built environment: A contextual perspective. *International Journal of Health Geographics*, *19*(1). https://doi.org/10.1186/s12942-020-00227-z

Karami, A., Kadari, R. R., Panati, L., Nooli, S. P., Bheemreddy, H., & Bozorgi, P. (2021). Analysis of Geotagging Behavior: Do Geotagged Users Represent the Twitter Population? *ISPRS International Journal of Geo-Information*, *10*(6), 373. https://doi.org/10.3390/ijgi10060373

Karhula, A., McMullin, P., Sutela, E., Ala-Mantila, S., & Ruonavaara, H. (2021). Rural-Urban Migration Pathways and Residential Segregation in the Helsinki Region. *Finnish Yearbook of Population Research*, *55*(2020), 1–24. https://doi.org/10.23979/fypr.96011

Karusisi, N., Bean, K., Oppert, J. M., Pannier, B., & Chaix, B. (2012). Multiple dimensions of residential environments, neighborhood experiences, and jogging behavior in the RECORD Study. *Preventive Medicine*, *55*(1), 50–55. https://doi.org/10.1016/j.ypmed.2012.04.018

Karusisi, N., Thomas, F., Méline, J., & Chaix, B. (2013). Spatial accessibility to specific sport facilities and corresponding sport practice: The RECORD Study. *International Journal of Behavioral Nutrition and Physical Activity*, *10*, 1–10. https://doi.org/10.1186/1479-5868-10-48

Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography*, *3*(3), 262–267. https://doi.org/10.1177/2043820613513388

Kohl, H. W. 3rd, Craig, C. L., Lambert, E. V., Inoue, S., Alkandari, J. R., Leetongin, G., & Kahlmeier, S. (2012). The pandemic of physical inactivity: global action for public health. *Lancet (London, England)*, *380*(9838), 294–305. https://doi.org/10.1016/S0140-6736(12)60898-8

Kohvakka, R., & Saarenmaa, K. (2019). *WhatsApp suosituin – some on suomalaisten arkea iän mukaan vaihdellen*. Statistics Finland. https://www.tilastokeskus.fi/tietotrendit/artikkelit/2019/whatsapp-suosituin-some-on-suomalaisten-arkea-ian-mukaan-vaihdellen/

Korenius, T., Laurikkala, J., Järvelin, K., & Juhola, M. (2004). Stemming and lemmatization in the clustering of finnish text documents. *International Conference on Information and Knowledge Management, Proceedings*, 625–633. https://doi.org/10.1145/1031171.1031285

Kotavaara, O., & Rusanen, J. (2016). *Liikuntapaikkojen saavutettavuus paikkatietoperusteisessa tarkastelussa: Liikuntapaikkojen saavutettavuusindeksi (LINDA) -hankkeen loppuraportti* (Issue January 2016).

Kuntoliikuntaliitto, S. (2010). *Kansallinen Liikuntatutkimus*.

Kwan, M. (2012). The Uncertain Geographic Context Problem. *Annals of the Association of American Geographers*, *102*(January 2011), 958–968. https://doi.org/10.1080/00045608.2012.687349

Langford, M., Higgs, G., & Radcliffe, J. (2018). The application of network-based GIS tools to investigate spatial variations in the provision of sporting facilities. *Annals of Leisure Research*, *21*(2), 178–198. https://doi.org/10.1080/11745398.2016.1272059

Lansley, G., de Smith, M. J., Goodchild, M. F., & Longley, P. A. (2020). Big Data and Geospatial Analysis. In and L. P. A. de Smith M J, Goodchild M F (Ed.), *Geospatial Analysis: A comprehensive guide to principles, techniques and software tools* (6th editio). The Winchelsea Press.

Lavie, C. J., Ozemek, C., Carbone, S., Katzmarzyk, P. T., & Blair, S. N. (2019). Sedentary Behavior, Exercise, and Cardiovascular Health. *Circulation Research*, *124*(5), 799–815. https://doi.org/10.1161/CIRCRESAHA.118.312669

Leavitt, M. O. (2008). 2008 Physical Activity. *Health (San Francisco)*.

Lee, C. (2016). Multilingual resources and practices in digital communication. In A. Georgakopoulou & T. Spilioti (Eds.), *The Routledge Handbook of Language and Digital Communication.* (pp. 118–132).

Lilius, J., Ramezani, S., Rinne, T., Hasanzadeh, K., & Kytt, M. (2021). *Centricity and multi-locality of activity spaces : The varying ways young and old adults use neighborhoods and extra-neighborhood spaces in Helsinki Metropolitan Area*. *110*(December 2020). https://doi.org/10.1016/j.cities.2020.103062

LIPAS. (2021). *LIPAS - Liikuntapaikat*. https://www.lipas.fi/liikuntapaikat

Litman, T. A. (2010). Accessibility. In K. Button, H. Vega, & P. Nijkamp (Eds.), *A Dictionary of Transport Analysis* (pp. 1–3). Edward Elgar Publishing Limited. https://doi.org/https://doi.org/

Liu, P., Koivisto, S., Hiippala, T., Van der Lijn, C., Väisänen, T., Nurmi, M., Toivonen, T., Vehkakoski, K., Pyykönen, J., Virmasalo, I., Simula, M., Hasanen, E., Salmikangas, A.-K., & Muukkonen, P. (2021). Extracting Locations from Sports and Exercise Related Social Media Messages using a Neural Network-based Bilingual Toponym Recognition Model. *Submitted for Peer Review*.

Lundqvist, A., Männistö, S., Jousilahti, P., Kaartinen, N., Mäki, P., & Borodulin, K. (2018). Terveys, toimintakyky ja hyvinvointi Suomessa - FinTerveys 2017 -tutkimus. Terveyden ja hyvinvoinnin laitos (THL), Raportti 4/2018. In K. S. and S. K. P. Koponen, K. Borodulin, A. Lundqvist (Ed.), *Terveyden ja hyvinvoinnin laitos* (pp. 38–41). Terveyden ja hyvinvoinnin laitos. http://www.julkari.fi/handle/10024/136223%0Ahttp://www.julkari.fi/bitstream/handle/10024/90832/Rap068_2012_netti.pdf?sequence=1

MacEachren, A. M., Jaiswal, A., Robinson, A. C., Pezanowski, S., Savelyev, A., Mitra, P., Zhang, X., & Blanford, J. (2011). SensePlace2: GeoTwitter analytics support for situational awareness. *VAST 2011 - IEEE Conference on Visual Analytics Science and Technology 2011, Proceedings, October*, 181–190. https://doi.org/10.1109/VAST.2011.6102456

Mäkinen, T. (2011). Liikunnan sosioekonomisia eroja selittävät tekijät aikuisilla. In P. Husu, O. Paronen, J. Suni, & T. Vasankari (Eds.), *Suomalaisten fyysinen aktiivisuus ja kunto 2010* (pp. 53–60). Ministry of Education and Culture.

Mancosu, M., & Vegetti, F. (2020). What You Can Scrape and What Is Right to Scrape: A Proposal for a Tool to Collect Public Facebook Data. *Social Media and Society*, *6*(3). https://doi.org/10.1177/2056305120940703

McCormack, G. R., & Shiell, A. (2011). In search of causality: a systematic review of the relationship between the built environment and physical activity among adults. *International Journal of Behavioral Nutrition and Physical Activity*, *8*(125), 557–560. https://doi.org/10.1016/j.ics.2007.02.011

McCrow-Young, A. (2020). Approaching Instagram data: reflections on accessing, archiving and anonymising visual social media. *Communication Research and Practice*, 1–14. https://doi.org/10.1080/22041451.2020.1847820

Middleton, S. E., Kordopatis-Zilos, G., Papadopoulos, S., & Kompatsiaris, Y. (2018). Location extraction from social media: Geoparsing, location disambiguation, and geotagging. *ACM Transactions on Information Systems*, *36*(4). https://doi.org/10.1145/3202662

Middleton, S. E., Middleton, L., & Modafferi, S. (2014). Real-time crisis mapping of natural disasters using social media. *IEEE Intelligent Systems*, *29*(2), 9–17. https://doi.org/10.1109/MIS.2013.126

Moran, P. (1950). Notes on Continuous Stochastic Phenomena Published by : Biometrika Trust Stable URL : http://www.jstor.org/stable/2332142. *Biometrika*, *37*(1), 17–23.

Norppa, M. (2020). *Helsingin kävelyn edistämisohjelma : tutkimuskatsaus*.

Openshaw, S. (1984). *The modifiable areal unit problem*. Geo Books.

Osakwe, Z. T., Ikhapoh, I., Arora, B. K., & Bubu, O. M. (2020). Identifying public concerns and reactions during the COVID-19 pandemic on Twitter: A text-mining analysis. *Public Health Nursing*, *19*(June), 1–7. https://doi.org/10.1111/phn.12843

Paananen, V. (2020, October 17). Asuinalueesta on tullut helsinkiläisen keskiluokan käyntikortti. *Helsingin Sanomat*. https://www.hs.fi/kaupunki/art-2000006672549.html

Paavo postal code area statistics 2020. (2020). http://www.stat.fi/static/media/uploads/tup/paavo/paavo2020_pitkakuvaus_en.pdf

Pyöräliikenteen kehittämisohjelma 2020–2025. (2020).

Rannikko, A., & Armila, P. (2020). Avoimuuden paradoksi: Vammaiset nuoret nuorisokulttuurisen liikunnan kentillä. In J. Kokkonen & K. Kauravaara (Eds.), *Eriarvoisuuden kasvot liikunnassa* (pp. 30–50). Liikuntatieteellinen seura.

Reimers, A. K., Wagner, M., Alvanides, S., Steinmayr, A., Reiner, M., Schmidt, S., & Woll, A. (2014). Proximity to sports facilities and sports participation for adolescents in Germany. *PLoS ONE*, *9*(3), 1–7. https://doi.org/10.1371/journal.pone.0093059

Rigolon, A., Browning, M., Lee, K., & Shin, S. (2018). Access to Urban Green Space in Cities of the Global South: A Systematic Literature Review. *Urban Science*, *2*(3), 67. https://doi.org/10.3390/urbansci2030067

Roberts, H., Sadler, J., & Chapman, L. (2017). Using Twitter to investigate seasonal variation in physical activity in urban green space. *Geo: Geography and Environment*, *4*(2). https://doi.org/10.1002/geo2.41

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, *346*(6213), 1063–1064. https://doi.org/10.1126/science.346.6213.1063

Sass, C. A. B., Pimentel, T. C., Aleixo, M. G. B., Dantas, T. M., Cyrino Oliveira, F. L., de Freitas, M. Q., da Cruz, A. G., & Esmerino, E. A. (2020). Exploring social media data to understand consumers' perception of eggs: A multilingual study using Twitter. *Journal of Sensory Studies*, *May 2019*, 1–9. https://doi.org/10.1111/joss.12607

Scholz, J., & Jeznik, J. (2020). Evaluating Geo-Tagged Twitter Data to Analyze Tourist Flows in Styria, Austria. *ISPRS International Journal of Geo-Information*, *9*(11), 681. https://doi.org/10.3390/ijgi9110681

Seppänen, A., Lilja, E., Mäki-Opas, J., & Wennman, H. (2020). Ulkomailla syntyneiden naisten vapaa-ajan liikunta. *Eriarvoisuuden Kasvot Liikunnassa*, 210–229.

Shelton, T., Poorthuis, A., Graham, M., & Zook, M. (2014). Mapping the data shadows of Hurricane Sandy: Uncovering the sociospatial dimensions of "big data." *Geoforum*, *52*, 167–179. https://doi.org/10.1016/j.geoforum.2014.01.006

Shen, J., Cheng, J., Huang, W., & Zeng, F. (2020). An exploration of spatial and social inequalities of urban sports facilities in Nanning City, China. *Sustainability (Switzerland)*, *12*(11). https://doi.org/10.3390/su12114353

Shrestha, S., Kestens, Y., Thomas, F., El Aarbaoui, T., & Chaix, B. (2019). Spatial access to sport facilities from the multiple places visited and sport practice: Assessing and correcting biases related to selective daily mobility. *Social Science and Medicine*, *236*(February 2018), 112406. https://doi.org/10.1016/j.socscimed.2019.112406

Skifter Andersen, H., Andersson, R., Wessel, T., & Vilkama, K. (2016). The impact of housing

policies and housing markets on ethnic spatial segregation: comparing the capital cities of four Nordic welfare states. *International Journal of Housing Policy*, *16*(1), 1–30. https://doi.org/10.1080/14616718.2015.1110375

Sloan, L., Morgan, J., Housley, W., Williams, M., Edwards, A., Burnap, P., & Rana, O. (2013). Knowing the Tweeters: Deriving Sociologically Relevant Demographics from Twitter. *Sociological Research Online*, *18*(3), 74–84. https://doi.org/10.5153/sro.3001

Statista. (2020). *Most used social media 2020 | Statista*. Statista.Com. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Statistics Finland. (2019). *Tunnuslukuja väestöstä alueittain, 1990-2019*. http://pxnet2.stat.fi/PXWeb/pxweb/fi/StatFin/StatFin__ene__ehk/statfin_ehk_pxt_002.px/

Sterdt, E., Liersch, S., & Walter, U. (2014). Correlates of physical activity of children and adolescents: A systematic review of reviews. *Health Education Journal*, *73*(1), 72–89. https://doi.org/10.1177/0017896912469578

Suomi, K., Sjöholm, K., Matilainen, P., Glan, V., Nuutinen, L., Myllylä, S., Pavelka, B., Vettenranta, J., Vehkakoski, K., & Lee, A. (2012). *LIIKUNTAPAIKKAPALVELUT JA VÄESTÖN TASA-ARVO - Seurantatutkimus liikuntapaikkapalveluiden muutoksista 1998–2009*.

Tabachnick, B. G., & Fidell, L. S. (2007). Using multivariate statistics, 5th ed. In *Using multivariate statistics, 5th ed.* Allyn & Bacon/Pearson Education.

Talen, E., & Anselin, L. (1998). Assessing Spatial Equity: An Evaluation of Measures of Accessibility to Public Playgrounds. *Environment and Planning A*, *30*, 595–613.

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), 1–11. https://doi.org/10.1038/s41598-017-18007-4

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*(November 2018), 298–315. https://doi.org/10.1016/j.biocon.2019.01.023

Torres, J., & Vaca, C. (2017). E-Health and fitness in Ecuador: A social media based analysis.

2017 4th International Conference on EDemocracy and EGovernment, ICEDEG 2017, 132–139. https://doi.org/10.1109/ICEDEG.2017.7962523

Twitter. (n.d.-a). *About account restoration*. Retrieved June 4, 2021, from https://help.twitter.com/en/managing-your-account/account-restoration

Twitter. (n.d.-b). *Tweet location FAQs _ Twitter Help*. https://help.twitter.com/en/safety-and-security/tweet-location-settings

Twitter. (2011). *One hundred million voices*. Twitter Blog. http://blog.twitter.com/2011/09/one-hundred-million-voices.html

UN. (n.d.). *The United Nations Sustainable Development Goals*. Retrieved November 20, 2020, from https://sdgs.un.org/goals

Vasankari, T., Kolu, P., Kari, J., Pehkonen, J., Havas, E., Tammelin, T., Jalava, J., Koski, H., Pihlainen, K., Kyröläinen, H., Santtila, M., Sievänen, H., Raitanen, J., & Kari, T. (2018). *Costs of physical activity are increasing – the societal costs of physical inactivity and poor physical fitness*. http://tietokayttoon.fi/documents/10616/6354562/31-2018-Liikkumattomuuden+lasku+kasvaa.pdf/3dde40cf-25c0-4b5d-bab4-6c0ec8325e35?version=1.0

Viguria, I., Alvarez-Mon, M. A., Llavero-Valero, M., del Barco, A. A., Ortuño, F., & Alvarez-Mon, M. (2020). Eating disorder awareness campaigns: Thematic and quantitative analysis using twitter. *Journal of Medical Internet Research*, *22*(7), 1–11. https://doi.org/10.2196/17626

Vogt, W., & Johnson, R. (2015). Correlation and Regression Analysis. *Correlation and Regression Analysis*. https://doi.org/10.4135/9781446286104

Xin, Y., & MacEachren, A. M. (2020). Characterizing traveling fans: a workflow for event-oriented travel pattern analysis using Twitter data. *International Journal of Geographical Information Science*, *34*(12), 2497–2516. https://doi.org/10.1080/13658816.2020.1770259

Zelenkauskaite, A., & Bucy, E. (2016). A scholarly divide: Social media, Big Data, and unattainable scholarship. *First Monday*, *21*(5), 1. http://10.0.20.90/fm.v21i5.6358%0Ahttp://search.ebscohost.com/login.aspx?direct=true&db=llf&AN=115285750&site=ehost-live

Appendix 1. Keywords used to retrieve sports related tweets

English	Finnish	Estonian
run, running, jog, jogging	juosta, juoksu, juokseminen, lenkkeillä, lenkki, lenkkeily	jooksmine, jooksma, jooks, sörkimine, sörkima, sörk, sörksjooks
walk, walking	kävellä, kävely, käveleminen	kõndimine, kõnd, kõndima, jalutama, jalutus, jalutamine,
hike, hiking, trek, trekking	patikoida, patikointi, patikoiminen,	matk, matkamine, matkama,
bicycle, bike, biking, cycling	pyöräillä, pyörä, pyöräily, pyöräileminen	jalgratas, jalgrattasõit, rattasõit
exercise, exercising, workout, training, sport, sporting	urheilla, treenata, treenaaminen, treeni, urheilu, liikunta	treening, treenima, võimlema, võimlemine, sportima, sportimine, trenn, sport
sweat, sweating	hiki, hikoilla	higi, higistama, higistamine
ski, skiing	hiihtää, hiihto, hiihtäminen	suusatama, suusatamine
skate, skating	luistella, luisteleminen, luistelu	uisutamine, uisutama
ice-hockey, hockey	jääkiekko, lätkä, hockey	jäähoki, hoki
football	jalkapallo, futis	jalgpall, jalka
basketball	koripallo, koris	korvpall, koss
floorball	salibandy, sähly	saalihoki
tennis, badminton, squash, tabletennis	tennis, sulkapallo, squash, kössi, pingis, pöytätennis	tennis, sulgpall, bädminton, squash, seinatennis, lauatennis
volley, volleyball, beach volley	lentopallo, lentis	volle, rannavolle, võrkpall, rannavõrkpall
swim, swimming	uida, uinti, uiminen	ujuma, ujumine
sail, sailing, kayak, kayaking, canoe, canoeing, rowing	purjehtia, purjehdus, kajakki, meloa, melonta, soutaa, soutaminen	purjetama, purjetamine, meresüst, kajakisõit, kanuutama, kanuutamine, kanuusõit, sõudmine, sõudma, aerutama, aerutamine
dance, dancing	tanssia, tanssi, tanssiminen	tants, tantsimine, tantsima

yoga	jooga	jooga, joogatama,
gym	kuntosali	jõusaal, võimla, spordihall, spordisaal

Estonian translation by Kerli Müürisepp.

Appendix 2. Survey form and questions

Sports activities and social media use

Tämä kyselylomake kartoittaa pääkaupunkiseutulaisten urheilutottumuksia ja urheiluun liittyvää sosiaalisen median käyttöä. Kysely on tehty Pro gradu -tutkielmaani varten, joka käsittelee Twitter-dataa urheiluun liittyvänä indikaattorina pääkaupunkiseudulla. Kyselyn tuloksia käytetään yhtenä keinona Twitter datan validoimiseen ja sen puutteiden kartoittamiseen. Teen gradututkielmaa Helsingin yliopiston maantieteen laitokselle, erikoistumislinjana geoinformatiikka. Kaikki vastaukset ovat anonyymejä ja niitä käytetään vain tutkimukseen. Mitään vastauksiasi ei liitetä sinun henkilötietoihisi. Lämmin kiitos kyselyyn vastaamisesta! Jos teillä herää kysymyksiä kyselyyn liittyen, voitte olla minuun yhteydessä sähköpostilla: sonja.koivisto@helsinki.fi.

This survey is made to assess people's sports activities and social media usage related to sports mainly in Helsinki Metropolitan area. The survey is for my master's Thesis where I use Twitter data as an indicator of sports activities. The responses will be used as one means to validate the Twitter data and assess its biases. The thesis will be done for Geography department of University of Helsinki, having a specialisation in geoinformatics. All responses are anonymous and used strictly for research purposes. None of the information you have provided can be retraced back to you. Warm thank you for answering the questionnaire! If you have any questions related to the survey, feel free to contact me by email: sonja.koivisto@helsinki.fi

Which language would you like to use?

- Suomi
- English

Sports activities

If your sports activities have significantly changed during the pandemic, answer as you would in normal situation (before or after covid).

How often do you do sports activities (including going for a walk or biking to work)?

- Many times a day
- Daily
- Many times a week
- Weekly
- Couple of times a month
- Rarely

How often do you do sports activities that require facilities (like go to swimming hall, play basketball or do horse-riding)?

- Many times a day
- Daily
- Many times a week
- Weekly
- Couple of times a month
- Rarely

Name the sports facilities you use (eg. Myyrmäki swimming hall, Lauttasaari sports field, Kivikko frisbeegolf track) If there are many, name 5 most frequently used.

[Open question]

How often do you do sports activities that DO NOT require facilities (like go for a run, walk, biking or home workout)?

- Many times a day
- Daily
- Many times a week
- Weekly
- Couple of times a month
- Rarely

Where do you run, bike, walk or do home workout? (name of the neighbourhood(s))

[Open questions]

Social media use related to sports

How often do you post about sports to social media? (excluding stories that disappear after certain time)

- Daily
- Weekly
- Monthly
- Couple of times a year
- Rarely
- Never (Skips straight to the next section)

Think about all sports-related posts you've posted to social media. How big percentage of them are about events in which you did not part take as an athlete? (E.g. Congrats to new ice-hockey world champions. Go Finland!)

0 – 100%

To which social media do you post about sports? You can choose multiple.

- Instagram
- Facebook
- Twitter
- Snapchat
- TikTok
- LinkedIn
- Tumblr
- YouTube
- Flickr
- Nike
- Strave
- Other, what?

Are there some sports that you more likely post about than others? E.g. running or trying out extreme sports

[Open question]

What affects the decision to post about sports? You can choose multiple.

- The sport itself
- Trendiness, status or rarity of the sport
- The venue / environment / location where you do the sports
- Who you are doing the sports with
- Natural phenomena like sunset
- Sports results (like personal best)
- Competitions or events
- Other, what?

Background

Age

- 0–19
- 20–29
- 30–39

- 40–49
- 50–59
- 60–69
- 70+

Gender

- Female
- Male
- Other
- I prefer not to say

Place of residence

- Helsinki
- Espoo
- Vantaa
- Kauniainen
- Other in Finland
- Outside Finland

Postal code area (if living in Helsinki Metropolitan Area)

[Open question]

Education

- Primary or Secondary school
- Matriculation examination or vocational school
- Bachelor's degree or equivalent
- Master's degree or equivalent

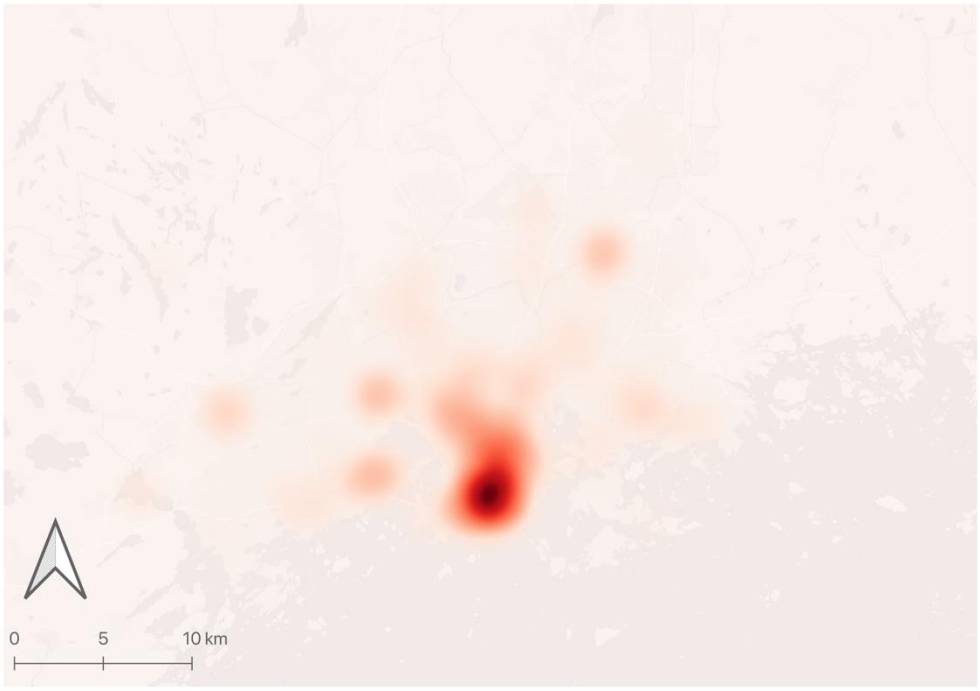
Annual income (netto in euros)

- < 15 000
- 15 000 - 30 000
- 30 000 - 45 000
- 45 000 - 60 000
- 60 000 +
- I prefer not to say

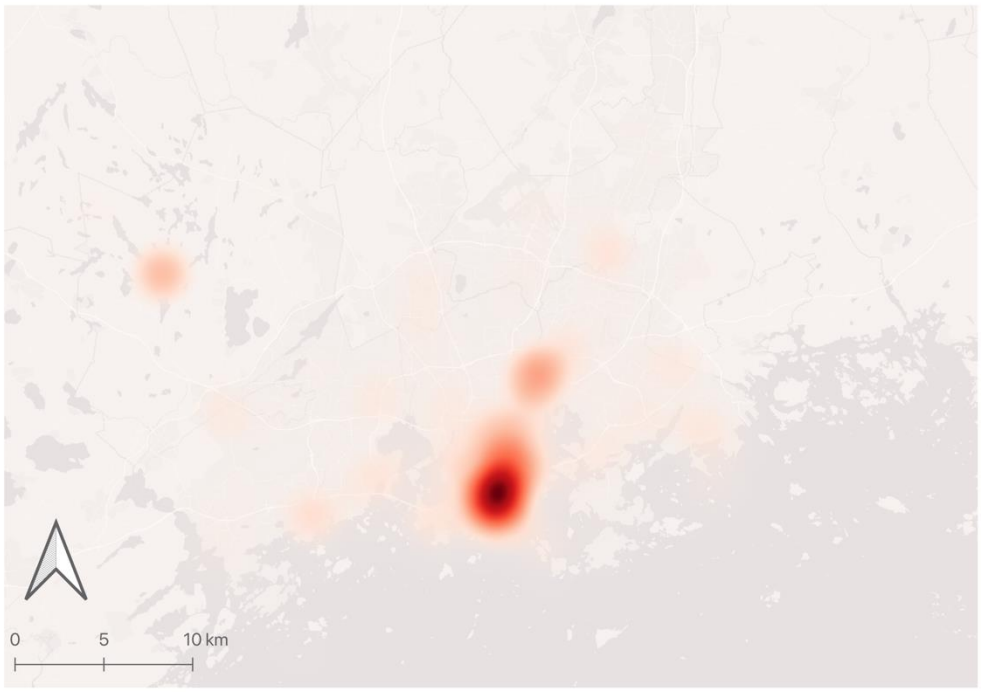
Thank you for your responses!

Appendix 3. Heatmaps of tweets by sport

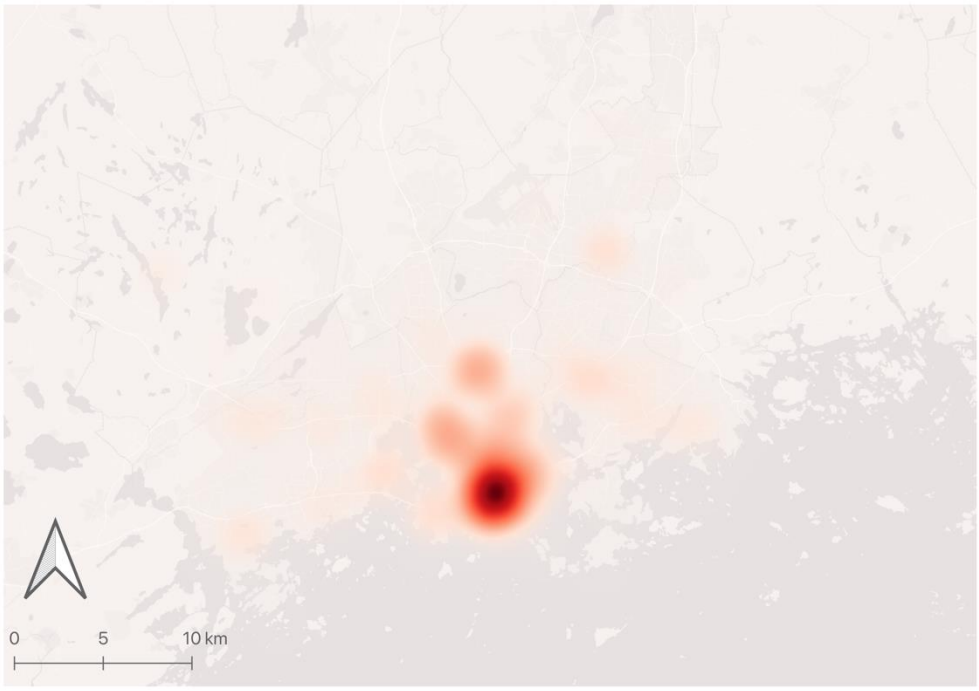
Heatmap of general sports tweets



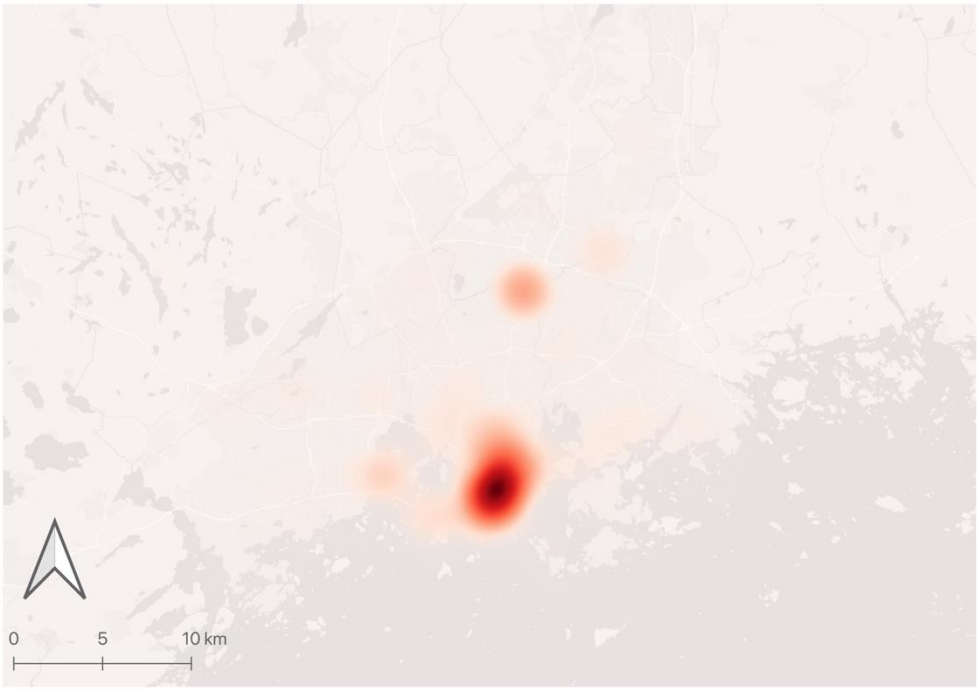
Heatmap of walking and hiking tweets



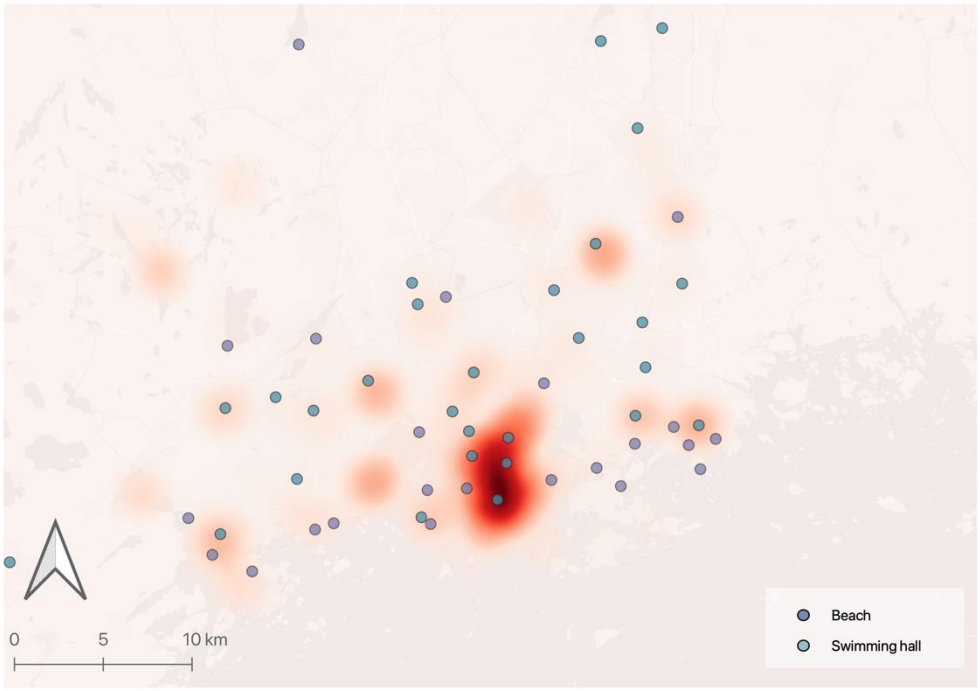
Heatmap of running tweets



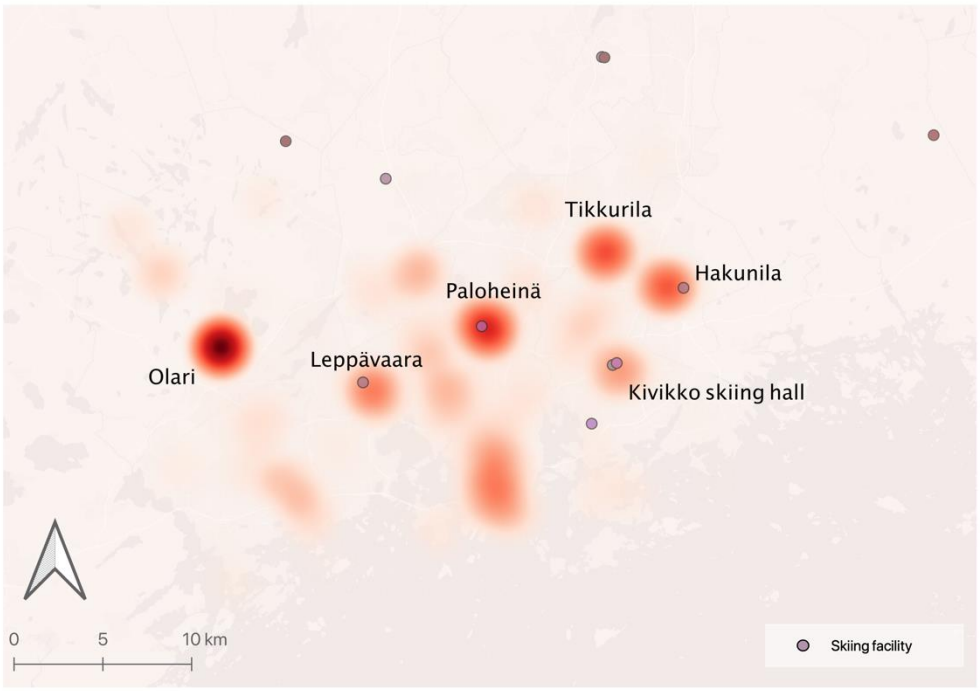
Heatmap of biking tweets



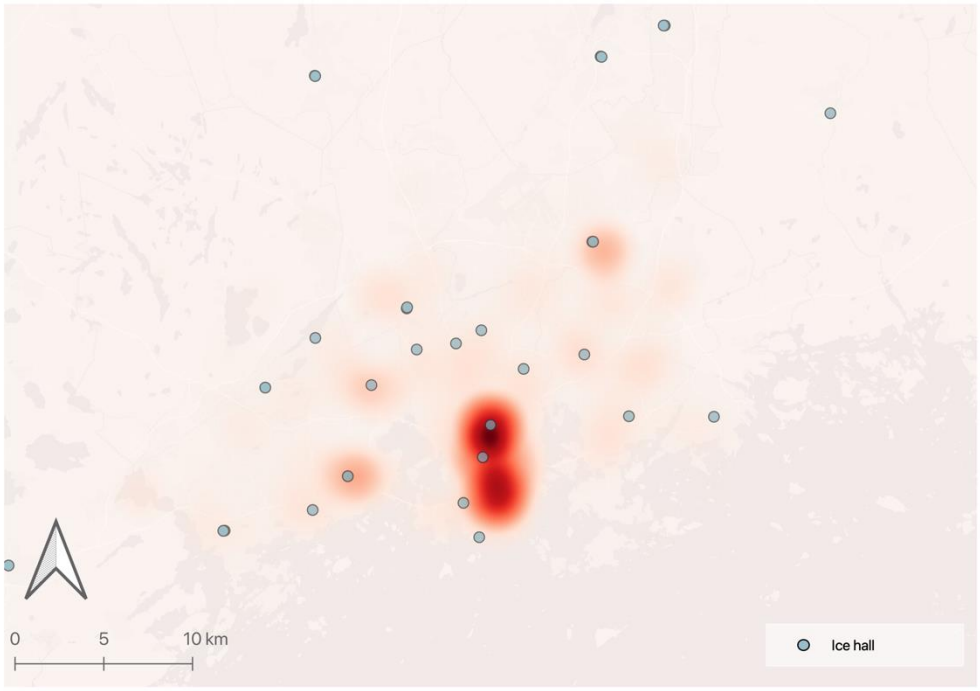
Heatmap of swimming



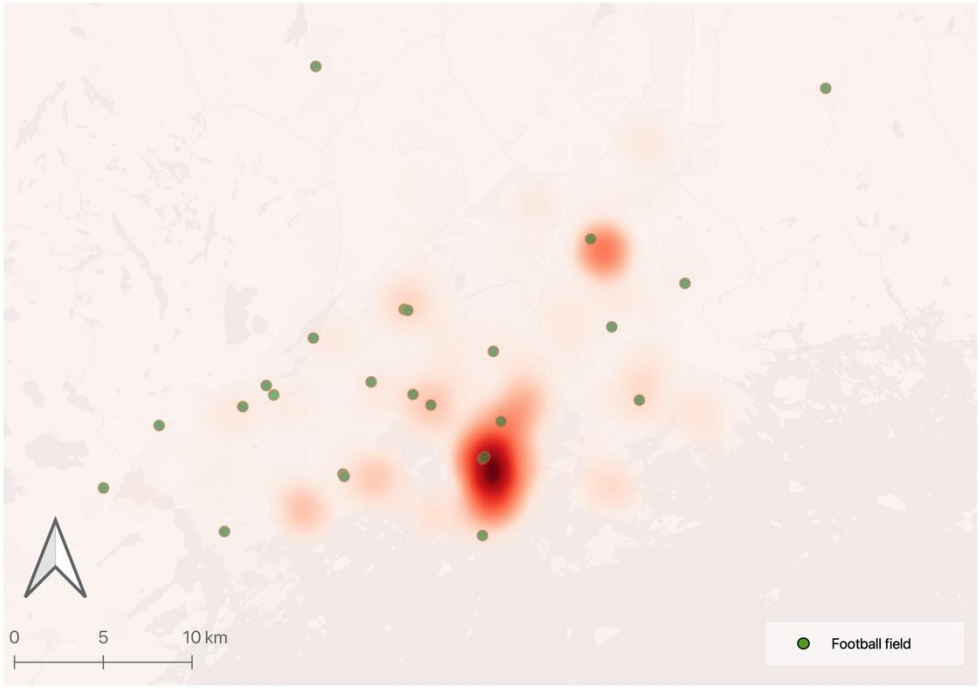
Heatmap of skiing tweets



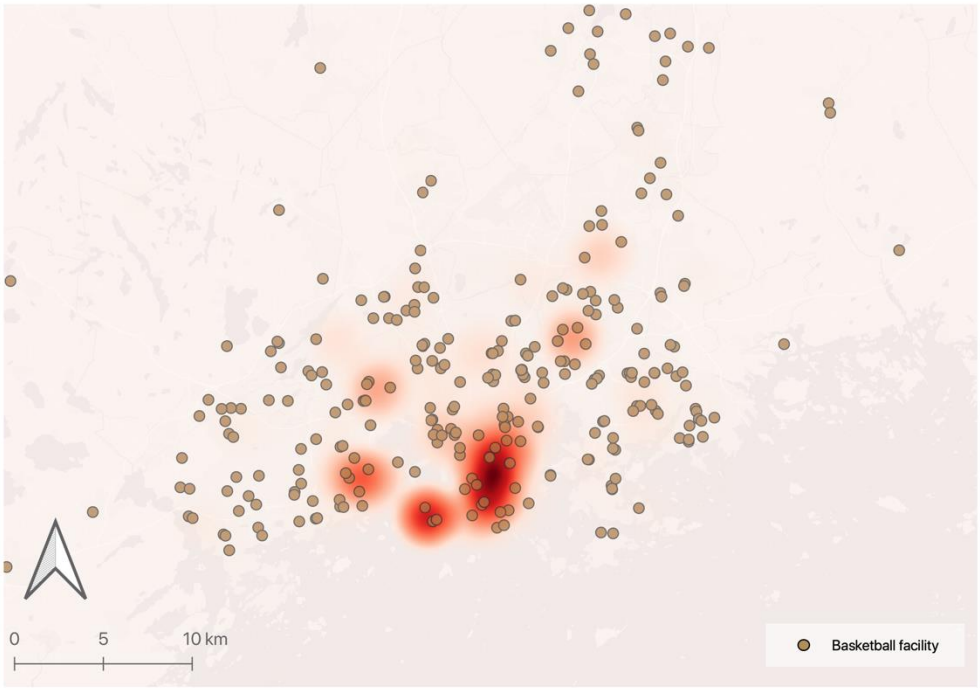
Heatmap of skating and ice hockey tweets



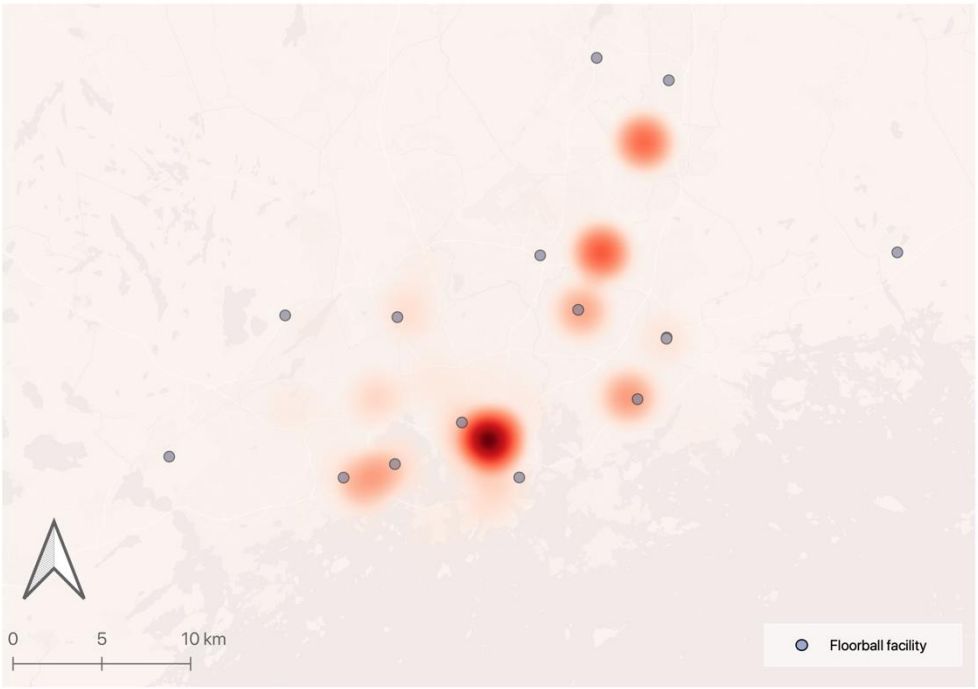
Heatmap of football tweets



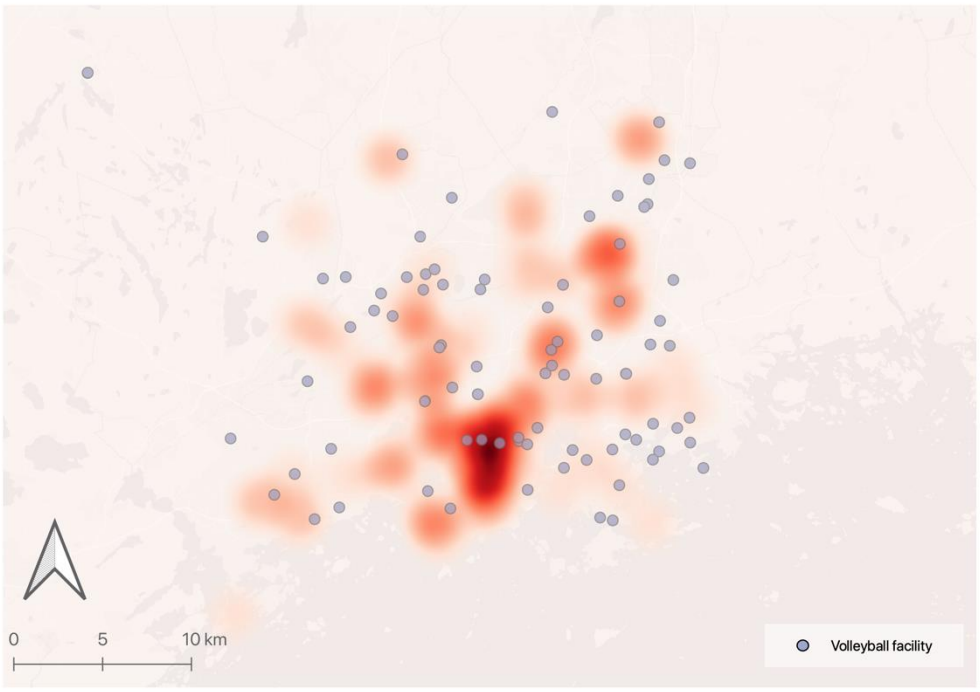
Heatmap of basketball tweets



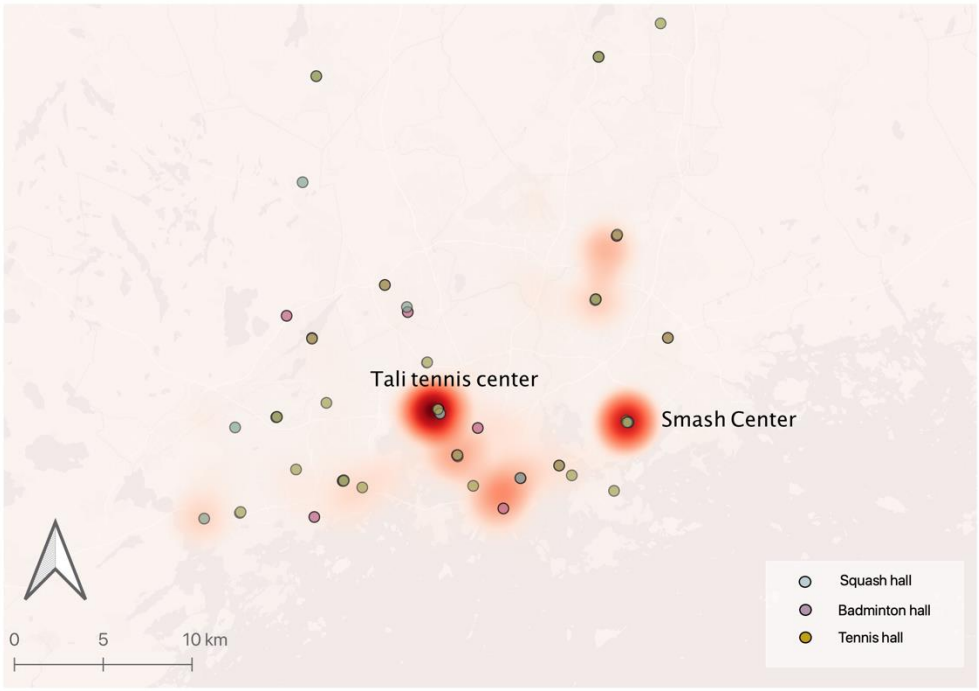
Heatmap of floorball tweets



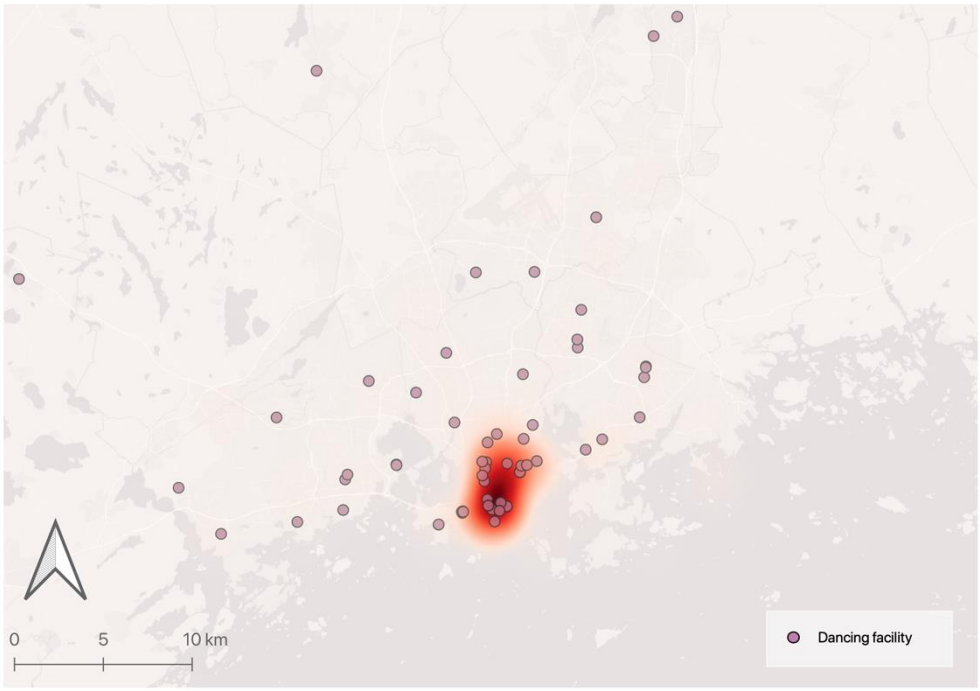
Heatmap of volleyball tweets



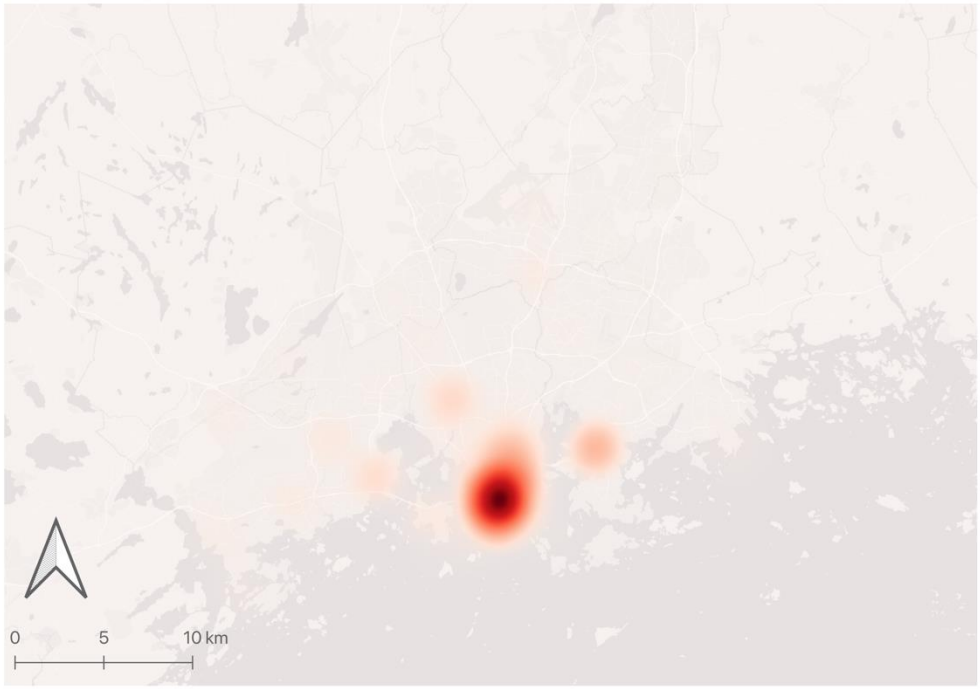
Heatmap of racket sport tweets



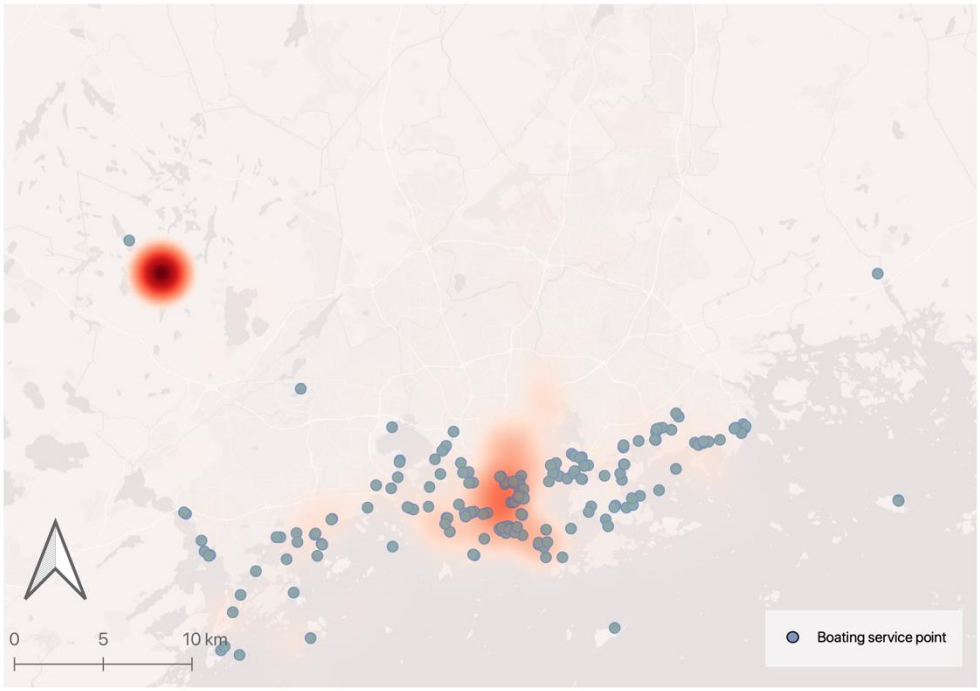
Heatmap of dancing tweets



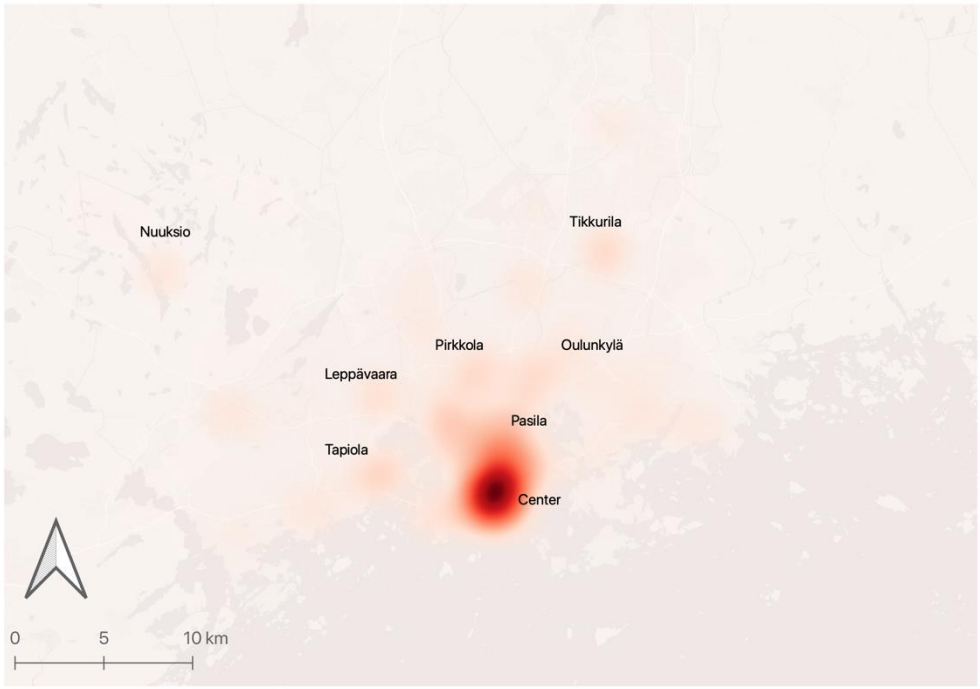
Heatmap of yoga tweets



Heatmap of kayaking, rowing and sailing tweets



Heatmap of tweets about sports that do not require facilitiies



Heatmap of tweets about sports that require facilities