Diversity of places and people

Using big data to understand languages and activities across geographical space

Tuomas Väisänen

Academic dissertation
to be presented for public examination,
with the permission of the Faculty of Science of the University of Helsinki,
in the Festive hall (Juhlasali) of Language Center (Kielikeskus), University of Helsinki,
on the 10th of November 2023 at 12 o'clock.

UNIVERSITY OF HELSINKI

Department of Geosciences and Geography A111
Helsinki, Finland
2023

Photo credits: Front cover: The Author
Back cover: Mimosa Kyyhkynen

Author: Tuomas Väisänen
Department of Geosciences and Geography
P.O. Box 64, FIN-00014
University of Helsinki
tuomas.vaisanen@helsinki.fi

Supervisors: Associate professor Tuomo Hiippala
Department of Languages
University of Helsinki

Docent, senior researcher Olle Järv
Department of Geosciences and Geography
University of Helsinki

Professor Tuuli Toivonen
Department of Geosciences and Geography
University of Helsinki

Pre-examiners: Professor Daniel Arribas-Bel
Department of Geography and Planning
University of Liverpool

Associate professor Leticia Serrano-Estrada
Department of Building Sciences and Urbanism
University of Alicante

Opponent: Associate professor Grant McKenzie
Department of Geography
McGill University

ISSN-L 1798-7911
ISSN 1798-7911 (print)
ISBN 978-951-51-9246-2 (paperback)
ISBN 978-951-51-9247-9 (PDF)
Available online https://ethesis.helsinki.fi/

Abstract

Urban populations are becoming highly diverse, or "super-diverse", through increasing globalization and international mobility. Super-diversity implies diversity that occurs across multiple variables such as language, ethnicity, religion, gender, age, country of origin, mobility, employment, and housing career. Language provides a useful, but underexplored perspective to super-diversity, as languages mediate every social interaction in urban areas and constitute a central part of individual and group identity. Superdiverse populations increase the diversity in urban areas also through their activities related to leisure, work, and everyday errands, all of which also vary across geographical space and time. I use the term urban diversity, by which I mean the diversity which emerges from the presence of super-diverse populations and their activities. Urban diversity exhibits spatio-temporal variation due to people's everyday mobility and the change of their residential areas. In this thesis, I concentrate two variables of urban diversity: languages and activities. Social media and population registers capture information about urban diversity, such as languages and activities, to different degrees. To better understand spatio-temporal urban diversity, the use of several sources of data and interdisciplinary approaches are necessary, as this understanding enables urban planners and decision-makers to support social cohesion and social sustainability in our cities.

In this thesis, I explore urban diversity from the perspectives of languages and activities using social media and population register data, focusing on Finland and especially the Helsinki Metropolitan Area (HMA). As urban diversity is a complex phenomenon, my work draws conceptually and methodologically on several fields of research: geographic information science, urban geography, and research on urban multilingualism, which also covers linguistic landscapes research. This thesis is strongly methodological in nature. I use diversity metrics originally developed in the fields of ecology and information science to assess the diversity of languages in population registers and social media content across Finland and the Helsinki Metropolitan Area. I apply computer vision techniques to extract information on activities from visual social media content. Finally, I use techniques from spatial analysis and statistics to examine the spatio-temporality of urban diversity across geographical scales from the national to local level.

I report the results of my research results in four articles. Article I explores the spatio-temporal diversity and richness of languages used by Finnish Twitter users from regional and user-based perspectives. The article shows how language use and linguistic diversity on Finnish Twitter varies across Finland, and characterizes the diversity of the linguistic repertoires of the users. Article II shows how to extract information on activities and visual preferences with several computer vision techniques from Flickr photographs taken in Finnish national parks. The focus on visual social media content circumvents challenges arising from multilingual and textually limited content. The article shows how the activities and landscape preferences of domestic and international visitors in the parks differ across the parks. Article III examines the variation in linguistic diversity in the Helsinki Metropolitan Area from population registers and social media during 2015. The article demonstrates how linguistic diversity derived from first language information in the population register and the linguistic repertoires of social media users can be used to assess where encountering a language other than one's own is likely. The article also explores what the background characteristics influencing linguistic diversity are. Article IV examines the spatio-temporal patterns of linguistic diversity in residential areas of the HMA between 1987–2019 and the integration of two sizeable local minority groups, Somali and Estonian speakers, to the Finnish society from the perspective of languages. The article reveals that while linguistic diversity is rising in all neighbourhoods, speakers of Somali and Estonian are exposed to it differently. The article also shows how linguistic diversity has changed in terms of the languages that constitute it and the locations where it is concentrated. Finally, article IV demonstrates that linguistic diversity in moderately diverse neighbourhoods is more likely to change, whereas monolingual and multilingual neighbourhoods are highly likely to remain as they are.

My results show that urban diversity is a spatio-temporal phenomenon and language is a useful variable for bringing out spatio-temporal patterns in urban diversity. My data shows the HMA has diversified rapidly from a monolingual area to a multilingual one. Moreover, the languages spoken in the HMA and the locations of multilingual neighbourhoods have changed, and changes in linguistic diversity in residential areas are influenced by their spatial surroundings. Furthermore, my results show that social media data reveals a more diverse spatio-temporal linguistic view of the HMA compared to what population registers demonstrate. Such a dynamic view provides more understanding of where and when urban populations encounter diversity. These results emphasize the importance of understanding the emerging spatio-temporal and social patterns of urban diversity, which provide vital information for policies fighting segregation, social tensions and social polarization. My work demonstrates the value of combining several sources of data, analysing them using interdisciplinary methods, while drawing conceptually on several fields of study to better understand urban diversity. As urbanization continues globally, and is accelerated by the climate crisis and increasing global instability, it draws more people into cities, interdisciplinary approaches to examining diversity in urban areas have become necessary for supporting inclusive, socially sustainable, and resilient urban futures.

Tiivistelmä

Kaupunkiväestöt muuttuvat alati monimuotoisemmiksi, tai “supermonimuotoisiksi”, kasvavan kansainvälisten muuttoliikkeiden ja globalisaation myötä. Supermonimuotoisuus näkyy väestön kielten, etnisyyksien, uskontojen, sukupuolen, iän, lähtömaiden, työllisyyden, asuntouran ja laillisen statuksen kirjon kautta. Näistä muuttujista erityisesti kieli muodostaa käyttökelpoisen, mutta vähän tutkitun näkökulman supermonimuotoisuuteen. Kielet ovat kaupungeissa tapahtuvan sosiaalisen kanssakäymisen keskiössä ja muodostavat yksilö- ja ryhmäidentiteetin keskeisen osan. Supermonimuotoinen väestö lisää kaupunkien monimuotoisuutta myös vapaa-aikaan, työskentelyyn ja arkipäivän askareisiin liittyvien aktiviteettien kautta, jotka ilmenevät eri tavoin maantieteellisesti ja ajassa. Käytän tässä työssä termiä kaupunkidiversiteetti, jolla kuvaan supermonimuotoisen väestön ja heidän aktiviteettiensa kautta kaupunkitilaan syntyvää spatiotemporaalista monimuotoisuutta. Kaupunkidiversiteetin maantieteellinen ja ajallinen vaihtelu syntyy ihmisten jokapäiväisen liikkumisen ja heidän asuinalueiden muutoksen kautta. Sosiaalisen median aineistot ja väestötietorekisteri tarjoavat mahdollisuuden tutkia kaupunkidiversiteettiä monipuolisesti, sillä ne kuvaavat kaupunkidiversiteettiä eri lähtökohdista ja eri tavoin. Kaupunkidiversiteetin parempi maantieteellinen ja ajallinen ymmärtäminen vaatii usean eri aineistolähteen käyttöä ja poikkitieteellisiä lähestymistapoja, koska ne mahdollistavat kaupunkialueiden sosiaalisen yhteenkuuluvuuden ja kestävyyden tukemisen kaupunkisuunnittelussa.

Tässä väitöskirjassa tutkin kaupunkidiversiteettiä kielten ja aktiviteettien näkökulmasta hyödyntämällä sosiaalisen median ja väestötietorekisterin aineistoja keskittyen Suomen ja erityisesti pääkaupunkiseudun tarkasteluun. Kaupunkidiversiteetti on monitahoinen ilmiö, jonka vuoksi työni ammentaa metodologisesti ja käsitteellisesti usealta tieteenalalta: geoinformatiikasta, kaupunkimaantieteestä ja kaupunkien monikielisyyden tutkimuksesta, johon myös kielimaisemallinen tutkimus kuuluu. Analysoin väestörekistereissä ja sosiaalisen median aineistoissa olevan kielitiedon rikkautta ja kirjoa ekologiassa ja informaatiotieteissä kehitetyillä mittariluvuilla arvioidakseni kielidiversiteetin maantieteellistä vaihtelua läpi Suomen ja pääkaupunkiseudun. Sovellan konenäkötekniikoita aktiviteettien tunnistamiseen visuaalisesta sosiaalisen median sisällöstä. Näiden lisäksi käytän spatiaalisen analyysin ja statistiikan menetelmiä kaupunkidiversiteetin spatiotemporaalisten piirteiden tutkimiseen niin kansallisella kuin myös paikallisella mittakaavatasolla.

Esitän tutkimukseni tulokset neljän artikkelin kautta. Artikkeli I tutkii Suomessa asuvien Twitter-käyttäjien käyttämien kielten diversiteettiä ja rikkautta aluetasolla ja käyttäjäkohtaisesti. Artikkeli osoittaa kuinka Twitter-käyttäjien kielten käyttö ja kielidiversiteetti vaihtelevat alueellisesti läpi Suomen. Artikkeli myös kuvaa käyttäjien henkilökohtaisia kielirepertuaareja. Artikkeli II näyttää miten aktiviteetteihin ja visuaalisiin mieltymyksiin liittyvää tietoa pystyy louhimaan konenäkömenetelmin Flickr-valokuvista. Tulokset demonstroivat kuinka visuaaliseen sisältöön keskittymällä voidaan välttää tekstuaalisen sisällön monikielisyydestä tai rajallisuudesta kumpuavat haasteet. Artikkelin tulokset osoittavat myös kuinka eri väestöryhmien aktiviteetit tai mieltymykset eroavat toisistaan. Artikkeli III selvittää pääkaupunkiseudun kielidiversiteetissä ilmeneviä muutoksia sosiaalisen median ja väestörekisteritietojen avulla vuodelta 2015. Artikkeli osoittaa miten eri tietolähteiden kielitietoa voi hyödyntää selvittääkseen missä ja mihin vuorokaudenaikaan kielidiversiteetin kohtaaminen on todennäköisintä pääkaupunkiseudulla. Artikkelin tuloksen kuvaavat myös mitkä sosioekonomiset ja ympäristötekijät vaikuttavat sosiaalisen median kielidiversiteettiin. Artikkeli IV selvittää kielidiversiteetin spatiotemporaalisia muutoksia pääkaupunkiseudun asuinalueilla vuosina 1987–2019, sekä kahden paikallisesti merkittävän kieliryhmän, somalin- ja vironkielisten, kotoutumista suomalaiseen yhteiskuntaan kielten näkökulmasta. Tulokset osoittavat miten pääkaupunkiseudun kielidiversiteetti on muuttunut alueellisesti ja sisällöllisesti. Artikkeli IV näyttää myös, miten kielidiversiteetti muuttuu herkimmin keskitasoisesti monikielisillä alueilla, kun taas yksikieliset ja erittäin monikieliset alueet eivät ole alttiita kielidiversiteetin muutoksille.

Väitöskirjani tulokset antavat uudenlaisen näkökulman kaupunkidiversiteetin spatiotemporaalisuuteen. Tulokset myös osoittavat kielen olevan hyödyllinen kaupunki-

diversiteetin spatiotemporaalisuuden ilmi tuomiseen. Käyttämäni aineistot osoittavat pääkaupunkiseudun nopean muutoksen yksikielisestä alueesta monikieliseksi. Lisäksi pääkaupunkiseudulla puhutut kielet ja monikielisten asuinalueiden sijainnit ovat muuttuneet, sekä asuinalueiden maantieteellinen ympäristö vaikuttaa asuinalueiden kielidiversiteetin muutoksiin. Sen lisäksi sosiaalisen median analyyseihini liittyvät tulokset paljastavat vieläkin kirjavamman urbaanin kielimaiseman kuin mitä rekisteriaineiston tarjoama näkökulma demonstroi. Tämä sosiaalisen median aineistojen tarjoama dynaaminen näkökulma tarjoaa mahdollisuuden parantaa ymmärrystämme siitä, missä ja milloin kaupunkiväestöt kohtaavat monimuotoisuutta. Nämä tulokset korostavat uusien kaupunkidiversiteetistä kumpuavin spatiotemporaalisten ja sosiaalisten rakenteiden ymmärtämisen tärkeyttä tietolähteinä kaupunkisuunnittelulle, joka pyrkii vähentämään segregaatiota, sosiaalisia jännitteitä ja polarisaatiota. Työni myös demonstroi usean aineistolähteen, poikkitieteellisten menetelmien, sekä usealta tieteenalalta yhteen sovitettujen käsitteiden muodostaman lähestymistavan arvoa kaupunkidiversiteetin tutkimukselle. Jatkuva kaupungistuminen houkuttelee väestöä alati kasvavissa määrin kaupunkeihin, jonka vuoksi poikkitieteellisestä lähestymistavasta kaupunkialueiden monimuotoisuuden tutkimiseen on tullut välttämätöntä sosiaalisesti kestävien ja yhteenkuuluvuutta edistävien kaupunkien tulevaisuuden takaamiseksi.

Contents

List of original contributions

This thesis is based on the following publications, which are referred to in the text by their Roman numerals.

Article I: 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://doi.org/10.51814/nm.99996

Article II: Väisänen, T., Heikinheimo, V., Hiippala, T., & Toivonen, T. (2021). Exploring human–nature interactions in national parks with social media photographs and computer vision. *Conservation Biology*, *35*(2), 424–436. https://doi.org/10.1111/cobi.13704

Article III: Väisänen, T., Järv, O., Toivonen, T., & Hiippala, T. (2022). Mapping urban linguistic diversity with social media and population register data. *Computers, Environment and Urban Systems*, *97*. https://doi.org/10.1016/j.compenvurbsys.2022.101857

Article IV: Väisänen, T., Järv, O., Toivonen, T., & Hiippala, T. (2023). Capturing urban diversity through languages: Long-term changes in multilingual residential neighbourhoods in the Helsinki Metropolitan Area. *Population, Space and Place.* https://doi.org/10.1002/psp.2717

Author's contribution

Table 1: The contribution of the author to each article. TV: Tuomas Väisänen (The author), TH: Tuomo Hiippala, OJ: Olle Järv, TT: Tuuli Toivonen, VH: Vuokko Heikinheimo.

	Article I	**Article II**	**Article III**	**Article IV**
Original idea	TH,TV,OJ,TT	TV,VH,TH,TT	TV,OJ,TT,TH	TV,OJ,TH
Study design	TH,TV,OJ,TT	TV,VH,TH,TT	TV,TH	TV,TH,OJ
Data collection	TV,OJ,TH	TV,VH	TV,OJ,TT,TH	TV,OJ,TT,TH
Analysis	TV,TH,OJ	TV,VH	TV	TV
Visualization	TV	TV	TV	TV
Manuscript preparation	TH,TV,OJ,TT	TV,VH,TH,TT	TV,OJ,TT,TH	TV,OJ,TT,TH

Acknowledgments

How can I even start this section? A PhD is not an endeavour of a single person, but a team effort, as this section will make very clear. Me being here would not be possible without the support of many amazing people within and without academia. First and foremost, I want to thank my supervisors, **Tuuli Toivonen**, **Tuomo Hiippala**, and **Olle Järv**, who have guided me through the process. They have been supernaturally patient with my slow writing pace and progress, especially Tuomo, in whose research project I did most of my PhD in. Tuomo, thank you for your tireless guidance, countless insightful comments, keeping my English writing in the academic ballpark, and all the conversations on science, music, and computers. Olle, thank you for always taking the time to read my manuscripts with thoroughness and providing highly important guidance and reassurances in more difficult times, but also not forgetting to include humour into our work. Tuuli, thank you for nearly everything: taking me onboard the Digital Geography Lab boat, being the Digital Geography Lab captain and navigating any rough waters with such confidence, teaching me not to stress about the difficult things, giving me responsibilities in teaching and lab-wide matters, but most of all for always being ready to sit down and discuss whatever the topic may be. If it was not for my supervisors, none of this would have been possible, I am deeply grateful. Thank you.

I want to thank **Grant McKenzie** for agreeing to be my opponent. I originally found your work on combining big data, spatial analysis, and natural language processing already when doing my Master's Thesis and afterwards got the chance to go to AAG 2019 in Washington, D.C., where I came to say hi. Since then, we've exchanged some messages on social media, and the idea of asking you to be my opponent started brewing in the back of my mind. I am really glad you agreed to be my opponent, thank you. I look forward to our discussion!

I would like to thank my pre-examiners, **Leticia Serrano-Estrada** and **Daniel Arribas-Bel**, for examining my PhD thesis and their encouraging words. Furthermore, I want to thank my thesis committee members, **Venla Bernelius** and **Katja Vilkama**, for being highly supportive, always willing to listen, and giving me insightful comments throughout the years.

There are so many colleagues to thank that I am afraid to lose track. **Elias**, if it wasn't for you referring a teaching opportunity in the USP programme to me back in early 2017, I think I would not be here today writing these words to a doctoral dissertation. That referral led me directly to my course assistantship in Tuuli's GIS course, research assistant positions, and the eventual PhD position in the group culminating in my likely opportunity to eventually get a cool sword and a fancy hat. Thanks for being a great colleague and an amazing friend since the very beginning of our studies. Our conversations have been a lifesaver more times than once, you are wicked smart, you have a great sense of humour, and it's been inspiring to work alongside you. How about that game of billiards with the Triumvirate? I think Chaplin survived the pandemic! Another gargantuan thanks goes to **Kerli**. You, Elias, and I started our PhD's roughly at the same time, and it has been a privilege to share the journey with such a bright, helpful, and joyous colleague. Thank you for the peer support and great discussions. *Doctors 2023, let's go!*

A massive thanks to all the wonderful past and present colleagues from Digital Geography Lab and beyond including **Chris, Vuokko, Joel, Håvard, Tatu, Johanna, Janika, Age, Oleksandr, Hanna-Mari, Robert, Matti, Aina, Om, Emil, Charlotte, Henna, Petteri, Henrikki, Enrico, Anna, Gonzalo, Ago, Laura, Roope, Rosa, Bryan, Daniel, Jeison,** and **Sławek**. One could not ask for better colleagues, thank you!

I thank the **Emil Aaltonen Foundation** for funding and supporting the MAPHEL project and Tuomo for having me as a PhD researcher in the project. I also want to thank **KONE Foundation** for funding and supporting the SoMeCon project, in which I did the first leg of my PhD journey, and **the city of Helsinki** for funding 3 months of my work in 2019. Additionally, I want to thank **Lähiöohjelma 2020–2022** for funding the YLLI project and **Petteri Muukkonen** for having me on board that project as well. Furthermore, I thank the Doctoral Programme in Interdisciplinary Environmental Sciences (DENVI) and **Karen Sims-Huopaniemi** for providing assistance in course selection and all administrative matters, and helping me get the credits registered in time before summer 2023. I thank professor **Petri Pellikka** for helping me with getting some remaining credits at the end of my PhD studies. Additionally, I want to thank **Ian Dobson** and **Martina Kurtén** for checking the language of my manuscripts and this synopsis.

I would also like to thank the folks in our department's coffee rooms (both Kumpula and Porthania) for providing much needed momentary escape from the stresses of academia. Thank you **Arttu, Tuuli, Heli, Panu, Johanna, Heidi, Mareike, Seppo, Janne, Veronica, Oskar, Mathew, Yu-Shan, Ian,** and **Lauri**.

I also want to thank my dear friends **Tatu**, **Tuomas**, and **Anton**. You have kept me sane by making sure there is enough skateboarding, disc golfing, fishing, beer festivals, and whisky game nights in my life. *Kaikesta ollaan selvitty, eikä mihinkään olla kuoltu. Tästä on erittäin hyvä jatkaa! Kiitos!*

Then **Heli** and **Teuvo**, my parents. I can not express the amount of gratitude I have for always supporting me no matter what. You gave me such a great childhood, got me interested in science, and instilled in me the idea that hard work always pays off one way or another. I can not believe, that me flipping over some rocks in our backyard to study the earthworms underneath, and staying up late to look at the Moon through dad's telescope have led me here. You know how many twists and turns there have been along the way, and it has been your support and belief in me that has given my much of the courage and strength to go through all of it. Also, I want to thank **Samu**, my brother, who has been there for the full journey and always being supportive and ready to talk about science, beers, games, or whatever our brains come up with. *Olette parhaita, kiitos kaikesta.*

Most importantly, my family. **Mimosa**, you are my compass, without which I would be hopelessly lost. Thank you for all the years of support and love. You have seen all of it. Me at my lowest and at my highest point, but you have always been there and believed in me and given me the extra energy to push on through whatever obstacle there was. You have made sure that when I am exhausted, you are there with open arms and a cup of coffee, but you have also been there raising a glass when I have been successful. I am forever grateful for your support. *Rakastan sinua.*

Maca, you are such a lovable and silly dog. Thank you for being the weird little creature you are. *Woof, woof!*

Renée, you are my Sun, Moon, North star, and the centre of my universe. You have given me so much strength, determination, and pure happiness by being there. You remind me that work and academia are not the most important things in the world, you are. I miss you every second I am away from you and can not wait to come home to be with you, hear about your day, invent new games, and go on our adventures. At the end of it all, it is you who is driving me forward, to become a better researcher, a better father, a better person. *Olet kaikista rakkain. Tämä väitöskirja on omistettu sinulle.*

1 Introduction

Cities are becoming increasingly diverse at an unprecedented scale due to accelerating growth of urbanization, and international migration and mobility. In 2022, 56% of the world's population lives in urban areas, and by 2050 this number is expected to be nearly 70% (United Nations, 2022). Not only are cities becoming more populous, but the populations living in cities are becoming more diverse. Contemporary urban populations can be characterized as being "super-diverse" (Vertovec, 2007), that is, they are diverse across multiple variables such as language, ethnicity, religion, gender, age, country of origin, mobility and access to the labour market and housing. Each of these demographic variables unfold in distinct spatio-temporal patterns across population groups and urban space, but also create complex and intersectional local and regional configurations that are difficult to quantify, visualize, and capture with conventional methods and data sets (Vertovec et al., 2022). The increasing diversity of population in urban areas poses a challenge for sustainable development goals pertaining to social issues in cities, but also an opportunity to increase social cohesion, resilience, well-being, and a sense of unity (Chriost & Thomas, 2008; Schroedler et al., 2023; United Nations, 2022; Wessendorf, 2014). In this work, I focus on the diversity of urban populations from the perspective of languages and activities. Moreover, I focus on "urban diversity" instead of super-diversity, because activities are not included in the original super-diversity variables, but activities of diverse populations arguably play a part in making an area diverse.

In the Finnish context, the rapidly increasing diversity is especially true in the Helsinki Metropolitan Area (HMA). In recent decades, cultural and linguistic diversity has increased in the HMA considerably (City of Helsinki, 2022; Dhalmann, 2013; Saukkonen, 2021b; Vaattovaara & Joutsiniemi, 2018), and the HMA has become what Pisarevskaya et al. (2022) describe as a "New Diverse" city. In a "New Diverse" city, the recent high increase in the diversity of population is caused by immigrants making up a relatively small, but growing, proportion of the total population, and who come from a wide variety of backgrounds. This increase of diversity in the HMA challenges and puts political pressure on urban planners and decision-makers in the region to evaluate and adjust their policies regarding housing, education, languages, and immigrant integration (Kraus, 2011; Saukkonen, 2021b), but also necessitates improving the understanding about the dynamic and structural spatio-temporal patterns to enable urban planning and decision-making to support sustainability and social cohesion (Kandt & Batty, 2020; United Nations, 2022; Vaattovaara & Joutsiniemi, 2018; Vertovec et al., 2022).

Simultaneously with the diversifying population, the amount of data about people and places generated by public entities, corporations, various sensors, Internet-of-Things (IoT) devices, mobile devices carried by private individuals, or big data, is increasing rapidly (Goodchild, 2013; Kitchin, 2013; Lansley et al., 2018). These novel data sources provide continuous information on people and places that was previously scarce or difficult to obtain, enabling finer-grained analyses of urban diversity. Big data is traditionally described through the concept of four V's: volume, velocity, variety, and veracity, which describe the large amount of data being generated at fast speeds, in a variety of formats and with varying levels of noisiness and certainty (Goodchild, 2013; Kitchin & McArdle, 2016). Accompanying this explosion of big data, tools and methods to analyse the vast volumes of data are becoming more efficient and accessible (Toivonen et al., 2019). Much of this big data contains geographical and temporal information (Goodchild, 2016), which has spawned concepts like digital twins and smart city (Batty, 2018), but also entire fields of study like urban informatics and city science (Batty, 2012; Shi et al., 2022; Singleton et al., 2018). Out of the varieties of big data, social media data in particular has received wide attention from geographers over the past ten years (Martí, Serrano-Estrada, et al., 2019; Martin & Schuurman, 2020; Toivonen et al., 2019). The sheer variety of data types and sources of big data necessitates interdisciplinarity to become the core for geographical analysis on big data (Kitchin, 2013).

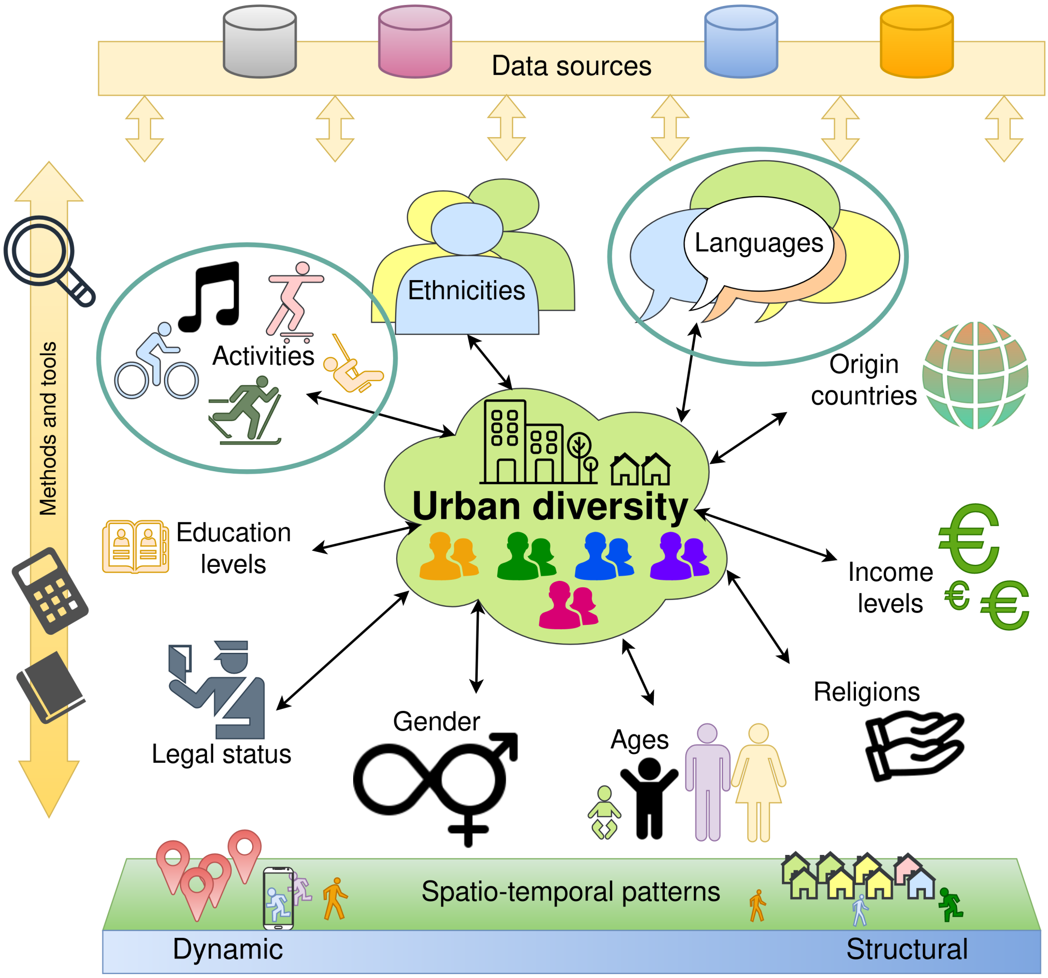

Figure 1: A framework for conceptualizing urban diversity from the point of view of geographical research. In addition to variables proposed by Vertovec (2007), there is an added variable about activities, which together constitute urban diversity. This thesis studies urban diversity from the perspective of languages and activities using traditional and big data sources. Big data sources reveal spatio-temporal patterns that reflect the dynamic side of urban diversity, whereas traditional data sources reflect the structural side.

The increased diversity in urban areas has led to a growing need to examine and understand the spatio-temporal configurations of urban diversity from novel perspectives (Vertovec, 2007), and the simultaneous proliferation of big data provides a data source for such endeavours (Arribas-Bel, 2014; Martí, Serrano-Estrada, et al., 2019; Vertovec et al., 2022). The framework of this study combines various variables of urban diversity and sources of data, but also how these unfold as particular dynamic and structural spatio-temporal patterns, which can be studied with various data, tools, and methods (Figure 1). The various data sources, such as official statistics and social media data, capture different aspects of urban diversity to varying degrees and at various scales. For instance, mobile phone data might reflect the presence of population across urban space as it captures the daily mobility of people fairly accurately and can indicate which population groups are present in the same area at the same time, but can not capture what people are doing there nor why they are there (Ahas et al., 2010; Toivonen et al., 2019). To understand reasons why people spend time in one area over the other, additional data sources, like social media data, are needed (Di Minin et al., 2015; Heikinheimo et al., 2020; Müürisepp et al., 2022). This underscores the necessity of using multiple data sources and methods (Martí, Serrano-Estrada, et al., 2019; Toivonen et al., 2019), so that findings from one big data source can be contextualized with or even verified by another big data or traditional data source (Arribas-Bel, 2014; Kandt & Batty, 2020; Tenkanen et al., 2017). However, gaining a comprehensive view into urban diversity is difficult, as the number of variables, scales, and sources of data that need to be visualized simultaneously might make any such visualizations uninformative or impossible to do (Vertovec et al., 2022). This reality has guided my choices throughout this work, which is why I am not attempting to combine multiple urban diversity variables into a "one-stop-shop" variable that would describe the diversity in its entirety, as that be antithetical to the intersectional nature of urban diversity (Vertovec, 2007). Instead, I am advocating for a focus on one or a couple of variables at a time and selecting variables that make most sense in the geographical and socio-spatial context of the analysis.

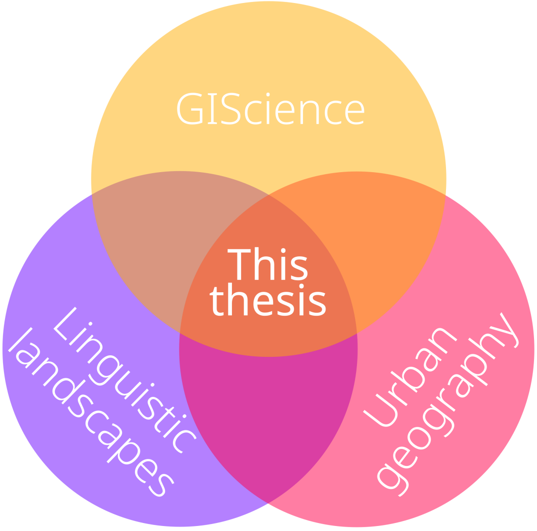

Figure 2: This thesis is an interdisciplinary work and draws from the research traditions of geographic information science (GIScience), urban geography and linguistic landscapes.

In this thesis, I am focusing on two perspectives on urban diversity: languages and activities. I focus on these for two reasons. First, languages are an often underexplored variable for understanding urban diversity (El Ayadi, 2021; Gorter, 2006; Johnston et al., 2021; Valentine et al., 2008). Not only do languages mediate all social interactions in urban areas and enable sharing of information in various social, cultural, political and communal contexts, but their connection to individual and group identities, and value to social cohesion and sustainable development are well known (Saarikivi & Toivanen, 2015; Schroedler et al., 2023; Tabouret-Keller, 2017; Zenker, 2018). Furthermore, the traditional variables used to analyse diverse areas and populations, country of origin and ethnicity, can be monolithic and obscure a large degree of additional diversity within (Johnston et al., 2021; Vertovec, 2007). Regardless, languages are often overlooked in describing the diversity of population groups or areas (Chriost & Thomas, 2008; El Ayadi, 2021; Johnston et al., 2021). Second, the activities various population groups engage in across geographical space also contribute to diversity. The mere co-presence of people of various backgrounds might not reveal intergroup contacts. Understanding which activities are taking place, when, where, and by whom thus provides much needed contextual information about what is happening across the geographical space (Di Minin et al., 2015; Heikinheimo et al., 2022; Toivonen et al., 2019). As a result, to examine a complex and multivariate issue like urban diversity (Figure 1) interdisciplinary approaches are necessary (Kitchin, 2013; Vertovec, 2007). In this work I draw methodologically from spatial analysis, machine learning, natural language processing, and biodiversity assessment

(Figure 4), and conceptually from the fields of geographical information science (GIScience), urban geography and linguistic landscapes (Figure 2). I use both traditional and big data sources, such as population registers and social media data, to understand spatio-temporal patterns in the diversity of languages at several spatio-temporal scales ranging from local to regional, and daily to several decades long.

1.1 Objectives

By examining languages and activities using social media and population register data from Finland and the Helsinki Metropolitan Area, I aim to pursue several objectives in this thesis to advance the study of urban diversity in the field of geography. The objectives of this thesis are as follows:

1. Reveal the urban diversity of the Helsinki Metropolitan Area and its dynamism using social media and population register data.
2. Explore the potential of applying both traditional and novel sources of data with interdisciplinary methods to the study of urban diversity through languages and activities.
3. Advance the methodological framework for studying linguistic diversity and activities in GIScience.

I address these objectives with this synopsis and four distinct articles, which constitute the thesis. The articles address the first two objectives in varying degrees, but the third objective is addressed by each article equally.

Article I addresses objective 2 and 3 by analysing the spatio-temporal linguistic diversity and richness of Finnish Twitter users from locational and user-based perspectives. Using geotagged and non-geotagged Twitter content from Finnish Twitter users, Article I identifies the languages used by Finnish Twitter and quantifies the diversity of languages with metrics originally developed in ecology and information sciences, to explore the spatio-temporal patterns of linguistic diversity and language use across Finland.

Article II addresses objective 2 and 3 by applying computer vision techniques on visual social media content to understand differences in activities and visual preferences between domestic and international visitors to Finnish national parks from geotagged Flickr content. To detect the activities and visual preferences, Article II uses openly-available and off-the-shelf computer vision models trained to perform object detection, instance segmentation, and image classification. The article shows that identifying user activities based on the photographs they have shared is a useful way to circumvent challenges arising from poor textual content, how activities vary across user groups and geographical regions, and how using several distinct computer vision techniques together provides complementary perspectives to understanding differences in activities and visual preferences between visitors. Even though the user data is from Finnish national parks, this article illustrates the feasibility of using computer vision techniques to enrich spatial social media content with information on activities and visual preferences derived from the visual content.

Article III addresses objectives 1, 2, and 3 through examining the variations between linguistic diversity in the Helsinki Metropolitan Area from population registers and social media during 2015, and exploring which background variables have an effect on linguistic diversity from social media. Article III uses geotagged Twitter and Instagram, population register, and mobile phone data to examine the spatio-temporal patterns of linguistic diversity in the HMA. The results show that everyday mobility and language use on social media reveals a more diverse and dynamic linguistic landscape compared to the more monolingual linguistic landscape from the population register data, and how various background variables influence spatio-temporal patterns of linguistic diversity from social media with spatial and non-spatial regression analyses.

Article IV addresses objectives 1, 2 and 3 by exploring the changing spatio-temporal patterns of linguistic diversity in the HMA with population register data from 1987 until 2019. It focuses on two locally important language groups: Somali and Estonian speakers, and examines their integration into Finnish society from the perspective of languages and several socio-economic variables. The article also explores the spatio-temporal stability of linguistic diversity across different types of neighbourhoods to understand how the likelihood and types of changes in linguistic diversity differ across neighbourhoods and spatial contexts. From the point of view of this thesis, Article IV provides a contextual backdrop for observations from social media in Articles I and III as it explores the diversifying population in the HMA and by extension in the Finnish society.

1.2 Positionality

I approach these objectives from the perspective of geography, and more specifically through the methodological and conceptual lenses of GIScience and urban geography. These analyses enable examining various spatial and social phenomena at multiple spatio-temporal scales, statistical testing of patterns, and modelling of relationships. However, this approach captures only one side of the phenomenon. For instance, the "emotions, values, beliefs and opinions; the varied, contextual, rational and irrational ways in which people interact and make sense of the world" (Kitchin, 2013, p. 265) are not captured in the register or social media data nor by the methods I use. Moreover, my focus on analysing spatio-temporal patterns in urban diversity from the perspectives of languages and activities constitute separate perspectives on a multivariate issue (Figure 1). Diversity in urban areas is inherently complex and intersectional (Putnam, 2007; Vertovec, 2007), and thus an interdisciplinary approach to questions about the topic is necessary. I believe the interdisciplinarity of this work strengthens it, as the work has been conducted from the perspective of several disciplines (Figure 2) with a diverse selection of methods and sources of data (see Section 3). My work demonstrates useful methods for analysing the spatio-temporal patterns of urban diversity on the scales of neighbourhoods, cities and regions, and by using large datasets, whereas for more ethnographical perspectives the reader is advised to seek out the work of Pienimäki et al. (2023).

2 Background

2.1 Urban diversity

Diversity plays a major role in making cities socially vibrant, and appealing places to live as well as nurturing innovation hubs, that boost economic development (Florida, 2003; Putnam, 2007). The diversity of urban populations has traditionally been assessed through the lenses of ethnicity, country of origin, and socio-economic status, however the increasing international mobility of people, things, and information over the past few decades (OECD, 2018; Sheller & Urry, 2006; United Nations, 2022) pose a challenge to the traditional monolithic descriptors of population (Johnston et al., 2021; Putnam, 2007; Vertovec, 2007). By coining the term "super-diversity", Vertovec (2007) made the claim that contemporary urban societies are not merely diverse, but super-diverse, due to increasing numbers of migrants and refugees from diverse backgrounds, and that the long-standing variables such as country of origin obscure much of the additional, potentially more meaningful, diversity within.

Vertovec (2007) called for more attention to be placed on other variables as well, such as language, gender, age, religion, and immigration status to examine the emerging patterns of socio-cultural diversity. Through examining and acknowledging this new diversity, urban policies and planning can better meet "the needs and conditions of immigrants, ethnic minorities and the wider population of which they are inherently part" (Vertovec, 2007,

p. 1050). This acknowledgment and meeting of needs will enable urban policies and planning to better support social cohesion and sustainability, by also addressing how the in-group is defined (Putnam, 2007). Consequently, the concept of super-diversity has been adopted by numerous fields of research, from sociology and linguistics to gender studies and geography (Beebeejaun, 2022; Blommaert, 2014; Cadier & Mar-Molinero, 2014; Lazar, 2022; Lehtonen, 2016; López Peláez et al., 2022; Vertovec, 2019). However, the way in which researchers have used or adapted the concept of super-diversity varies considerably (Vertovec, 2019), and its most fruitful use has been identified to be the examination of new social complexities (Vertovec, 2019; Vertovec et al., 2022). Furthermore, much of the work focusing on super-diversity has not considered the role of activities, whereas the role of people's activities in urban diversity is well-established in research on urban morphology (Crooks et al., 2015; Kang et al., 2021; Niu & Silva, 2021; Sayyar & Marcus, 2011). However, super-diversity and the diversity of people's activities have not been brought together to understand the spatio-temporality of urban diversity, as I aim to do in this thesis.

Urban diversity is an intensely local experience and unfolds as varying spatio-temporal patterns across urban space (OECD, 2018; Syrett & Sepulveda, 2012; United Nations, 2022; Vertovec et al., 2022). Exposure to diversity is thus largely seen as a useful way to reduce tensions between population groups belonging to various social, cultural, and ethnic strata, because exposure can increase understanding, dialogue, and generate a sense of community in highly diverse urban areas (Amin, 2002; Chriost & Thomas, 2008; El Ayadi, 2021; Powers et al., 2022; Syrett & Sepulveda, 2012; Valentine et al., 2008; Wessendorf, 2014; Ye, 2019). Social mixing policies are considered a policy tool that supports social cohesion and integration in a society through exposure to diversity, and simultaneously mitigates neighbourhood segregation and marginalization (Bolt et al., 2010; Fincher et al., 2014; Syrett & Sepulveda, 2012). That said, solely passing encounters with diversity or co-presence in the same residential neighbourhood might not be sufficient for generating understanding or a meaningful dialogue with 'others' (Amin, 2002; Blommaert, 2014; Fincher et al., 2014; Valentine et al., 2008; Ye, 2019), although encounters in communal places like parks, courtyards, or corner stores, and in more formal spaces with organized activities are more likely to lead to positive outcomes (Ho et al., 2021; Hoekstra & Pinkster, 2019; Pienimäki et al., 2023; Powers et al., 2022; Valentine et al., 2008; Wessendorf, 2014). To summarize, we need more knowledge about the spatio-temporal patterns of urban diversity across various types of places, such as home, work, and third places, to understand its impacts on social sustainability and cohesion (Putnam, 2007; Vertovec et al., 2022; Wessendorf & Farrer, 2021).

Increasing diversity can also pose a challenge to urban areas (Amin, 2002; Chriost & Thomas, 2008; Fincher et al., 2014; Vertovec, 2019; Ye, 2017). Increasing diversity can also result in socio-culturally homogenous neighbourhoods due to the concentration of 'others' and the local (majority) population moving out (Hoekstra & Pinkster, 2019; Matejskova & Leitner, 2011; Putnam, 2007; Vilkama et al., 2013). Recent studies from Europe, however, challenge these claims and indicate that cities are more likely to have increasingly socially and culturally mixed neighbourhoods despite growing population diversity due to immigration (Catney et al., 2023; Zwiers et al., 2018). Moreover, some studies indicate increasing ethnic diversity causes segregation and lack of trust between residents initially, but can after a while materialize as increased well-being and social cohesion (Pisarevskaya et al., 2022; Putnam, 2007). Complicating matters further, policies aimed to support diversity can simultaneously and inadvertently reinforce differences while also celebrating them (Fincher et al., 2014; Hewidy & Lilius, 2022; Matejskova & Leitner, 2011). Much depends on the socio-spatial contexts whether encounters with diversity occur and how sustained these contacts are (El Ayadi, 2021; Matejskova & Leitner, 2011; Valentine et al., 2008; Wessendorf, 2014). Furthermore, there are some critiques towards super-diversity as a concept as well. For instance, Beebeejaun (2022) claims the concept obscures and preserves social and racial hierarchies as it views population groups through what separates them, instead of how they are similar.

2.2 The role of languages in urban diversity

All forms of interaction in the city are mediated by language. Language has a key role in social interactions as it may be used to include or exclude individuals (El Ayadi, 2021; Putnam, 2007), but also as one of the main components of and a method to express individual and group identity (Alexander et al., 2007; Järv et al., 2015; Saarikivi & Toivanen, 2015; Tabouret-Keller, 2017; Zenker, 2018), which also makes it a potentially powerful marker of urban diversity, especially when compared to country of origin and country of birth. To exemplify, an origin country merely indicates where the individual migrated from, and the country of birth where they were born. These are monolithic background descriptors that aggregate potentially more meaningful information away and disregard whether the individual was already an immigrant in the origin country (Abascal & Baldassarri, 2015; Anniste & Tammaru, 2014; Catney et al., 2023; Gesthuizen et al., 2009; Johnston et al., 2021; Kandylis et al., 2012; Vertovec, 2007; Zwiers et al., 2018). For example, a single country can have speakers of numerous languages and each individual language can signify widely different cultures and social statuses (Christopher, 2004; Kumar, 2019; Mustapha, 2014). Similarly, a focus on ethnicity can obscure potentially more relevant background information. For instance, the Kurds constitute an ethnic group whose members speak many languages, only some of which are interrelated, and are spread out across multiple countries. That said, languages can also be equally monolithic and reductive descriptors. Depending on how information on language is recorded, it might not indicate the individual's competency in the language, the varied ways the individual might use the language in the real world, if the person is multilingual, or how the society values competencies in different languages (Artamonova & Androutsopoulos, 2019; Latomaa, 2012; Saukkonen, 2016; Schroedler et al., 2023).

The HMA has implemented some policies to deal with increasingly diverse and multilingual populations in recent decades. The social mix housing policy adopted by the city of Helsinki is largely seen as successful in reducing residential segregation in the city (Torpan et al., 2022), even though there are some challenges, such as the continuing concentration of low-income and ethnic minority households to eastern Helsinki (Hyötyläinen, 2019; Vilkama, 2011). In this regard, calls for more detailed information on the diverse populations in the HMA has been raised by Saukkonen (2021b). The approach of Helsinki to increasing multilingualism has also garnered some attention from researchers (Kraus, 2011; Nuolijärvi, 2015). For instance, Kraus (2011) describes this as "integrative multilingualism" whereby education services stress the need for immigrants to acquire Finnish skills, while also supporting the first language skills of the immigrant children. This is done by providing two hours of language instruction weekly. However, Nuolijärvi (2015) points out that these two hours are likely not enough for actual development and maintenance of these languages if they are not used daily. In fact, policies that safeguard minority cultures and languages have a long political and institutional history in Finland and Helsinki (Nuolijärvi, 2015; Saukkonen, 2018). In this regard, Finland is considered to be one of the most multicultural-friendly countries in Europe in terms of policies that safeguard cultural and language rights of minorities (Saukkonen, 2018, 2021a). However, these policies stand at odds with the general Finnish public as there is a "widely shared idea of the Finnish nation and society, or Finnish national identity" as an ethnically and culturally homogeneous Finland (Saukkonen, 2018, p. 69).

The increasing diversity of population groups present in urban areas has led to an intensified multilingualization of the society (Roberts, 2010; Schroedler et al., 2023) and increasing interest towards language from researchers (Gorter, 2006). As the role of language is central in social interactions (Artamonova & Androutsopoulos, 2019; El Ayadi, 2021; Putnam, 2007; Valentine et al., 2008) and constitutes a central component of individual and group identity (Chriost & Thomas, 2008; Saarikivi & Toivanen, 2015; Segrott, 2001; Tabouret-Keller, 2017; Zenker, 2018), it is a good variable to use when studying urban diversity. Although language has gained some attention in the field (Farber et al., 2012; Kellert & Matlis, 2022; Müürisepp et al., 2022; Väisänen et al., 2022), its potential is largely underexplored in GIScience and geography. Calls to use language as an important trait for characterizing diversity of populations and to explore its spatio-temporal patterns (e.g. Johnston et al., 2021; Vertovec, 2007) has also generated interest in a subfield of sociolinguistics, linguistic landscape studies (Blommaert & Maly, 2019; Vertovec, 2019).

2.3 Linking linguistic landscapes and geography

Linguistic landscapes is a subfield of sociolinguistics concerned with examining the presence of languages in public space, and their social, cultural and political implications (Blackwood, 2015; Landry & Bourhis, 1997). Connecting the study of linguistic landscapes to the study of super-diversity has been highlighted as being a natural next step to examine the contemporary socio-cultural patterns of urban areas (Vertovec, 2019).

In the seminal study by Landry and Bourhis (1997), linguistic landscapes were originally concerned in documenting the languages visible in public and commercial signs in urban space to examine their "symbolic" and "informational" functions. These functions inform passers-by which language communities are present in the area, who the area caters to, but also about the vitality of a language group in demographic and political terms, and whether the use of one's own language is encouraged (Blackwood, 2015; Landry & Bourhis, 1997). The field has since broadened to become more interdisciplinary and thus includes more diverse geographical settings, methods, sources of data, types of language use, languages in virtual spaces, and has in some cases moved beyond language into semiotics (Barni & Bagna, 2015; Biró, 2018; Blackwood, 2015; Gorter & Cenoz, 2015; Hiippala et al., 2023; S.-Y. Hong, 2020; Ivkovic & Lotherington, 2009; Lazar, 2022; Moshnikov, 2016; Peukert, 2013). Not only have massive digital data sets, such as Google Street View images (S.-Y. Hong, 2020), social media content (Biró, 2018; Hiippala et al., 2019; Koskinen, 2013), and other digital sources (Chun, 2014; Ivkovic & Lotherington, 2009; Keles et al., 2020; Moshnikov, 2016, 2022) become more common sources of data enabling the coverage of large urban areas initially doubted by Blackwood (2015), but also the methods have diversified from quantitative to more qualitative such as in-depth interviews and ethnographies (Blommaert & Maly, 2019; Lazar, 2022; Pienimäki et al., 2023). Furthermore, linguistic landscape studies have long been focused on providing analyses depicting the situation at some frozen point in time and recently more emphasis on recognizing the spatio-temporal dynamism of the linguistic landscape has been called for (Blommaert & Maly, 2019; El Ayadi, 2021; Pennycook & Otsuji, 2015).

Linguistic landscapes are linked to geography in several ways. First, just the name linguistic landscape is linked to a central concept in geography, landscape. Research in human geography has explored the relationship between landscapes and people extensively (Cosgrove, 1985; Jones, 2003; Sauer, 1925; Schein, 1997; Wylie, 2009). Second, and following from the first, the implicit relationship between geography and linguistic landscapes have been acknowledged several times in linguistic landscape studies (Aboelezz, 2015; El Ayadi, 2021; Gorter, 2006; Hiippala et al., 2019; Leeman & Modan, 2009). To exemplify, Leeman and Modan (2009) pointed out that by focusing linguistic landscape research on language policies, the field has largely disregarded that languages in public space are impacted by urban planning policies and the socio-spatial practices in the area, which are studied predominantly by geographers. Furthermore, El Ayadi (2021) calls for applying concepts from human and urban geography to better conceptualize the linguistic landscape and broaden its scope, but also to support geographers in their research of linguistic diversity. More broadly, Derungs et al. (2020) shows that even dialectometry, a field focused on studying the geographical distribution of languages and dialects, does not use methods from spatial analysis and statistics as widely as it could, reifying a much earlier call for more methodological and conceptual cooperation between linguists and geographers (Trudgill, 1974). Finally, as urban diversity is encountered and experienced on the local level (Amin, 2002; Syrett & Sepulveda, 2012; United Nations, 2022), languages and linguistic diversity unfold as interesting local patterns (Vertovec, 2007) and reveal larger patterns of diversity underneath (Gorter, 2006), the application of geography and spatial methods in studying linguistic diversity at various geographical and temporal scales is a natural next step (Derungs et al., 2020).

2.4 Big data in geographical research

Big data has become a popular source of data in geography and GIScience in particular (Ash et al., 2018; Boyd & Crawford, 2012; Goodchild, 2013; Janowicz et al., 2015; Kitchin, 2013; Lansley et al., 2018; Singleton & Arribas-Bel, 2021; Tasse et al., 2017). Big data refers to the data that is continuously generated by people, their mobile devices, sensors in the environment, and IoT (Internet of Things) devices (Arribas-Bel et al., 2015; Batty, 2012; Kitchin, 2013; Miller & Goodchild, 2015). For example, this data can consist of mobile phone call records, social media content, travel card and customer card data, and smartphone application use data (Järv et al., 2014; Lansley et al., 2018). Big data is commonly distinguished from traditional data through the concept of three Vs: volume, velocity, and variety (Goodchild, 2013; Kitchin, 2013), that correspond to the amount of data being generated, the speed of data generation, and the variety of data types. Some scholars have added additional Vs to the original three, such as variability, value, and veracity, which correspond to the variability in meaning or structure of the data, the value generated by the data, and the data's reliability or truth value (Goodchild, 2016; Lansley et al., 2018; S. Li et al., 2016). Some have also described the era of big data through availability of data: previously research data was rare and difficult to obtain, but with the ubiquity of computers, mobile devices, sensors, IoT devices, and people who use them, data has become more abundant and relatively easy to obtain (Poorthuis & Zook, 2017; Townsend, 2013). As much of big data is georeferenced or has a spatial component, there have been numerous discussions on the role of big data in geography (Goodchild, 2013; Kitchin, 2013; Lansley et al., 2018) and even proposals of a new geographic subfield: geographic data science (Singleton & Arribas-Bel, 2021).

2.4.1 Strengths of using big data in research

Big data has brought about a new era of geographical urban research (Arribas-Bel et al., 2015; Batty, 2012; Kandt & Batty, 2020; Kitchin, 2013; Martin & Schuurman, 2020; Miller & Goodchild, 2015; Poorthuis et al., 2021; Shelton et al., 2015), as it provides information on people and places not available in more traditional data sources at an unprecedented scale and resolution (Kitchin, 2013; Lansley et al., 2018; Shelton et al., 2015). Unlike traditional sources of data that have been created specifically for research or governmental purposes, big data has no such agenda, but it comes about "accidentally, as a byproduct" (Arribas-Bel, 2014, p. 45) of other activities, like social media use. Big data thus enables urban research to be performed outside of their traditional "spatial and temporal boundaries" (Müürisepp et al., 2022, p. 11). The usefulness of big data for studying and understanding cities, particularly "as a complementary alternative to those already in wide use (such as population censuses or surveys)" (Arribas-Bel, 2014, p. 45) is recognized quite widely (Arribas-Bel et al., 2015; Ilieva & McPhearson, 2018; Martí, Serrano-Estrada, et al., 2019; Niu & Silva, 2020). Moreover, the amount of big data generated about urban areas and urban populations by people, devices, and sensors has led to the concepts of smart cities and digital twins (Batty, 2012, 2018). The concepts of smart cities and digital twins are used to describe how urban planners and city governments can leverage the data for modelling urban processes, making decisions, governing, and drafting policies (Batty, 2012, 2018; Kandt & Batty, 2020; Shi et al., 2022; Townsend, 2013).

Geotagged social media data is a unique source of in situ information on human presence and activities in urban and natural areas, and is increasingly used in geographic research (Ilieva & McPhearson, 2018; Kitchin, 2013; Lopez et al., 2019; Martí, Serrano-Estrada, et al., 2019; Martin & Schuurman, 2020; Miller, 2020; Miller & Goodchild, 2015; Tenkanen et al., 2017; Toivonen et al., 2019). Social media content typically consists of either textual or visual content or both, which can be georeferenced either through coordinate or place information (Di Minin et al., 2015; Toivonen et al., 2019), or by deriving the location from place names in the content with geoparsing (Leppämäki, 2022; Middleton et al., 2018). Particularly among geographers, social media data has received a lot of attention and several overviews have been written on the applicability, prospects and challenges of using social media data in urban studies (Arribas-Bel, 2014; Ilieva & McPhearson, 2018; Martí, Serrano-Estrada, et al., 2019; Martin & Schuurman, 2020), conservation science (Di Minin et al., 2015; Toivonen et al., 2019), environmental research (Lopez et al., 2019), and geospatial analysis (Niu & Silva, 2020; Owuor & Hochmair, 2020).

Through georeferenced social media content, the connections between digital activities to the physical location have been used to understand population dynamics in cities (Arribas-Bel et al., 2015; Heikinheimo et al., 2020; Hochman & Manovich, 2013; Silva et al., 2014). For example, social media data has been used to study where people spend their free time (Adelfio et al., 2020; Heikinheimo et al., 2020; Martí et al., 2017; Silva et al., 2014), how they perceive different areas (Boy & Uitermark, 2017; Jenkins et al., 2016; Martí et al., 2020; Vasquez-Henriquez et al., 2020; Zukin et al., 2017), and how neighbourhoods interact (Cvetojevic & Hochmair, 2021; García-Palomares et al., 2018; Martí, García-Mayor, & Serrano-Estrada, 2019; Shelton et al., 2015). This explicit connection of digital content to a real location through coordinate information has also led to theorizations about the intertwined digital and physical spaces, such as "double space" (Kellerman, 2014) or "hybrid space" (de Souza e Silva, 2006), which assume digital activities reflect actual activities in physical spaces, and through which researchers study how urban and digital spaces are affected by each other (Croitoru et al., 2015; Ferreira & Vale, 2021; Rose, 2022).

Mobile phone data is another source of big data used widely to study the mobility and activities of various population groups in urban areas (Ahas et al., 2010; Bergroth et al., 2022; Järv et al., 2021; Müürisepp et al., 2023). The drawback with mobile phone data compared to social media data is the information is largely based on the presence of a mobile device in a spatial unit, often without detailed contextual information about the users or their activities (Ahas et al., 2010; Silm, Jauhiainen, et al., 2021). Some contextual information can be at least partially derived from contract details, land use, or depending on the type of mobile phone data, the applications used (Farber et al., 2015; Järv et al., 2014; Müürisepp et al., 2022; Silm, Mooses, et al., 2021).

Analysing the vast amount and variety of data available from social media platforms has become possible through recent developments in data science and machine learning regarding textual and visual content analysis (Singleton & Arribas-Bel, 2021). This move from labour-intensive manual classification to processing millions of records automatically have enabled extensive analyses of content on multiple geographical and temporal scales (Martin & Schuurman, 2020; Toivonen et al., 2019). Techniques from computational linguistics and natural language processing have also been used to analyse spatio-temporal patterns from georeferences social media data. There is a sizeable body of geographical research on modelling spatio-temporal topics (Fu et al., 2018; L. Hong et al., 2012; Lansley & Longley, 2016), sentiments (Cao et al., 2018; Gruebner et al., 2018; Yan et al., 2020) and lexical change (Eisenstein et al., 2014; Grieve et al., 2018; Y. Huang et al., 2016) from the textual content. Simultaneously, computer vision techniques have improved rapidly during the past ten years and moved from more simple tasks to increasingly complex tasks (Chai et al., 2021; Vanky & Le, 2023), and are applied across various urban and natural contexts (Biljecki & Ito, 2021; Ghermandi et al., 2022; Kruse et al., 2021; Wilkins et al., 2022).

As social media data is unstructured and noisy, language information and other metadata can enrich social media data with information about ethnic, cultural, linguistic identity of the user and what they are doing (Dunn & Adams, 2020; Herdağdelen, 2013). Peukert (2013) introduced the idea of estimating linguistic diversity in urban areas using common measures from ecology and information sciences. Subsequent work has applied these measures to analysing diversity at various spatial scales, ranging from specific locations (Hiippala et al., 2019) to entire cities (Bereitschaft & Cammack, 2015; Jiang et al., 2022) and countries (Dunn & Adams, 2020; Hiippala et al., 2020).

2.4.2 Challenges of using big data in research

The strength of big data and other novel data sources, is that they provide information not available from more traditional sources on a massive scale (Boy & Uitermark, 2016; Lansley et al., 2018), however they are not free of problems. One main issue in big data is representativeness of the data (Crampton et al., 2013; Hargittai, 2020; Heikinheimo et al., 2022; Ilieva & McPhearson, 2018; Lansley et al., 2018; Toivonen et al., 2019). Social media data often lacks background information, such as age, gender, and ethnicity, on the users (Ilieva & McPhearson, 2018). Furthermore, as users are free to use whichever social media platform they want for whatever purpose they want, there is a systemic bias stemming from which population groups are present on various platforms and to what degree (Hargittai, 2020). Additionally, the social media users are more likely to reflect young, more well-off people that have access to technology and a good level of internet skills (Ash et al., 2018; Dunn et al., 2020; Hargittai, 2020; Koiranen et al., 2020; Manikonda et al., 2016; Robinson, Schulz, Blank, et al., 2020; Robinson, Schulz, Dunn, et al., 2020). As a result, there likely are population groups missing from these data sets, and they likely represent more disadvantaged population segments (Boyd & Crawford, 2012; Hargittai, 2020; Robinson, Schulz, Blank, et al., 2020).

Spatial and temporal coverage are also not uniform in big data, and is compounded by the representativeness issue. There is an abundance of data from densely populated urban areas, whereas more remote rural and natural areas are very sparse in data coverage, and similar differences exist between high-income and low-income countries (Boyd & Crawford, 2012; Di Minin et al., 2015; Ilieva & McPhearson, 2018; Lansley et al., 2018; Levin et al., 2015; Zook et al., 2017). This is a natural product of big data being produced by people, their devices and various sensors, which are concentrated in urban areas and likely owned by economically more well-off people. However, it also creates new types of digital divides between people and areas that are rich or poor in big data, and in the areas where big data is prevalent, the data might be generated by a small subset of the population making inferences from such data tenuous (Boyd & Crawford, 2012)

Another challenge is access to data (Boyd & Crawford, 2012; Lansley et al., 2018; Toivonen et al., 2019). Due to the scale and resolution of big data, much of it is not released to the public in order to safeguard commercial value, individual privacy, or strategic importance (Hargittai, 2020; Kitchin, 2013; Lansley et al., 2018). In case access is granted, there might be problems related to how the provided data was collected or sampled, but also strict conditions and limitations on how to use the data (Lansley et al., 2018; Toivonen et al., 2019). For instance, several APIs (Application Programming Interfaces) that can provide data access to researchers might provide only a portion of data and with strict guidelines and limits to how much data can be queried within a given timespan (Lansley et al., 2018; Poorthuis & Zook, 2017). If a researcher wants a highly specific data set, the cost of acquiring it might be prohibitively high (Boyd & Crawford, 2012; Poorthuis & Zook, 2017).

Due to these representation and coverage issues, the choice of data sources and careful consideration of how these issues affect the analysis are imperative in order not to perpetuate existing social and spatial inequalities (Martí, Serrano-Estrada, et al., 2019; Owuor & Hochmair, 2020; Tenkanen et al., 2017; Toivonen et al., 2019). Explicit discussion on the limitations and biases of the data is a cornerstone of the scientific process, and this discussion is increasingly important with big data sources (Goodchild, 2013; Kitchin, 2013; Müürisepp et al., 2022). Moreover, as social media data can contain highly accurate information on the individuals, responsible and ethical collecting, analysing, and storing of data, even if the data was publicly available is necessary (Di Minin et al., 2021; Zook et al., 2017). Furthermore, providing open access to the analysis methods and, if possible, data supports responsible, ethical, and open research (Holbrook, 2019; Nelson et al., 2022; Zook et al., 2017).

Many scholars have also pointed out that big data should be used alongside other data sources, especially if the research aims to describe phenomena and produce changes in the real world (Goodchild, 2016; Lansley et al., 2018; Martí et al., 2021; Shelton, 2017). Kandt and Batty (2020) stress the need to "triangulate" findings from big data with more established and reliable data sources. Echoing this stance, Lansley et al. (2018, p. 551) advise researchers to consider big data as a "by-product, or 'exhaust' from a process that does not have re-use of data for research purposes at its heart". Similarly, Tasse and Hong (2017, p. 256) points out that geotagged content on social media should be considered as "postcards, not ticket stubs", due to self-censorship and other social media platform-specific digital behaviours. Finally, if there is data available from traditional sources on the same topic as from big data sources, emphasizing the traditional data source over big data would make sense as traditional data sources have well-established and rigorous sampling frameworks which ensure the representativeness and reliability of these data sets (Goodchild, 2013; Hargittai, 2020).

Finally, recent developments regarding access to big data from social media companies are indicating changes to how social media data can or will be used in future research. During the past few years, the availability and accuracy of social media data has reduced due to several controversies related to privacy of users, data protection laws, and changes in management (Brembs et al., 2023; Bruns, 2019; Freelon, 2018; Maurer, 2020). For instance, the Cambridge Analytica scandal caused several social media companies to limit or cut off access to their data (Bruns, 2019; Freelon, 2018). As a consequence, there have been fears over social media companies allowing only research that presents the companies in positive light will get access to the data, and some scholars have called for more research using techniques like web scraping to extract digital geospatial information from social media platforms and other websites (Brenning & Henn, 2023; Freelon, 2018). Furthermore, there has been a general tendency among social media platforms to move away from geotagging with GPS coordinates to place-tagging with points-of-interest data (S.-Y. Hong, 2020; Kruspe et al., 2021; Maurer, 2020; Tasse & Hong, 2017), which reduces the spatial granularity of social media data as more individual posts will share identical geographical coordinates, e.g., the centroid of an administrative area.

2.5 Previous research on languages and activities using social media data

2.5.1 Research using language information to study spatio-temporal patterns of population

Social media is inherently multilingual (Coats, 2019b; Eleta & Golbeck, 2014; Hiippala et al., 2019; L. Hong et al., 2011; Magdy et al., 2014, 2016; Mocanu et al., 2013), and is considered a rich resource on spatio-temporal linguistic information (Herdağdelen, 2013). Language has been shown to affect what is shared on social media (Androutsopoulos, 2014; Artamonova & Androutsopoulos, 2019; Weerkamp et al., 2011), formation of social ties (Eleta & Golbeck, 2014; Takhteyev et al., 2012), and the likelihood of using geotags (B. Huang & Carley, 2019; Magdy et al., 2014). Despite the wide use of social media data and the methodological advancements to identify languages from textual content (Barman et al., 2014; Hiippala et al., 2019; Jauhiainen et al., 2019; Lui & Baldwin, 2012; Pratap et al., 2023; Zubiaga et al., 2016), geographical research on social media has rarely used information on languages. Many geographical studies focusing on analysing the textual content on social media are concentrated on English content, and either discard content in other languages from the analysis (see e.g., Cao et al., 2018; Fu et al., 2018; Gruebner et al., 2018; Jenkins et al., 2016; Karami et al., 2021; Lansley & Longley, 2016) or do not mention the language of the content at all (see e.g., G. Andrienko et al., 2013; Chapple et al., 2021; Croitoru et al., 2015; Crooks et al., 2013; Martin & Schuurman, 2017; Yan et al., 2020). This body of work thus inadvertently compounds the representation issues of social media by focusing on content in just one language. Nevertheless, the focus on English is not surprising. English is the global lingua franca and widely used on social media (Coats, 2019b; Mocanu et al., 2013). Moreover, English is well-resourced in terms of language technology needed for analysing large volumes of social media content (Del Gratta et al., 2021). However, the dominance of English in social media content varies across the globe (Mocanu et al., 2013).

Language has not gone entirely without attention from geographers examining spatio-temporal patterns from big data sources. For instance, Graham and Zook (2013) explored languages in Google Maps content in Canada, Israel, Spain, and Belgium, and found connections between language visibility, power relations, and segregation. Heikinheimo et al. showed that languages on social media can be used to understand who are using urban green spaces (2020), and to understand who social media users represent (2022). Kellert and Matlis (2022) found variation in formal and informal use of Spanish on Twitter is connected to the socio-spatial context. Järv et al. (2015) compared the activity spaces of Estonian and Russian speakers in Tallinn, Estonia and found differences in the spatial extent and shape of the activity spaces, which was followed up by Silm, Mooses, et al. (2021), who found the differences disappear when the social networks of individuals are more interethnic. Adelfio et al. (2020) examined the social media user activities in several Swedish neighbourhoods and the neighbourhood with most Twitter activity to have more multilingual geotagged content.

More attention on languages has come from the field dialectometry. Dialectometry is a subdiscipline of linguistics, which studies the geographical distribution and spread of languages, dialects, and linguistic features computationally (Donoso & Sanchez, 2017; Wieling & Nerbonne, 2015; Wieling et al., 2011). The field draws methodologically from GIScience and spatial statistics (Grieve et al., 2019; Grieve et al., 2017; Wieling & Nerbonne, 2015). However, the prime focus of the field is on the spatial diffusion of phonology, morphology and lexical innovation (Donoso & Sanchez, 2017; Grieve et al., 2018; Wieling & Nerbonne, 2015), and not on the spatial distribution and diversity of language use or language speakers. For example, Donoso and Sanchez (2017) explored the geographical variation of Spanish dialects from geotagged Twitter posts, and Eisenstein et al. (2014) studied the geographical diffusion of several words in the USA from geotagged Twitter data. Nevertheless, and given the methodological overlap, geographical information science and dialectometry could benefit from each other.

2.5.2 Research on identifying activities from social media data

Social media has been shown to be a rich resource for study of activities in the real world (Martí, Serrano-Estrada, et al., 2019; Toivonen et al., 2019). Activities of social media users across geographical space have been studied by extracting information either from the textual or visual content, or by connecting the location of the geotagged content to land use or a POI (García-Palomares et al., 2018; Lopez et al., 2019; Niu & Silva, 2021; Shen & Karimi, 2016; Toivonen et al., 2019). Using textual content, topic modelling of social media content has been popular in trying to understand how discussion topics vary across geographical space and how they might imply the activities of social media users at various locations (Crooks et al., 2015; Fu et al., 2018; Hasan & Ukkusuri, 2014; Lansley & Longley, 2016; Martí et al., 2021; Martin & Schuurman, 2017). Some studies have combined topic modelling and sentiment analysis to understand whether certain topics are perceived more negatively or positively (Cao et al., 2018; Gruebner et al., 2018; Hausmann et al., 2020; Yan et al., 2020). Also, activities related to cross-border mobility of social media users have been studied through geotagged content shared consistently between two or more countries (Aagesen et al., 2022; Järv et al., 2022).

Using visual content to derive activities, which is the approach we take in Article II, many studies have used computer vision techniques to identify activities or objects commonly associated with particular activities. As photographs, such as those on social media, are a visual medium, they are embedded with information about what is being captured in the photo, but also implicit information on the values and preferences of the photographer (Moldez & Gomez, 2022). Consequently, social media photographs have become a popular data source in research to understand activities and preferences of social media users (Hausmann et al., 2018; Toivonen et al., 2019). For instance, Chen et al. (2020) used Flickr photographs to study the popularity and spatial distribution of landscape characteristics in London and found seasonality to affect activities detected from the images, Ghermandi et al. (2022) explored which type of human-nature interactions occur in urban green spaces, and Kruse et al. (2021) identified areas where children are more likely to start playing to inform where safety interventions ought to be directed to. Also, street view imagery has become a popular data source for understanding cities (Biljecki & Ito, 2021). Recently, Vanky and Le (2023) pointed out that many descriptions of activities provided by computer vision techniques do not perform well in complex urban environments and recommends both fine-tuning training data with local data and a posteriori manual verification.

From the land use perspective, the activities of social media users have also been modelled based on the land use surrounding the geotagged content (N. Andrienko et al., 2016; Cao et al., 2018; García-Palomares et al., 2018; Heikinheimo et al., 2020) or based on information on points-of-interest from the area (Calafiore et al., 2021; Martí et al., 2021; Niu & Silva, 2021). For example, Niu and Silva (2023) enriched Twitter content with user-specific metadata about gender and age and connected the geotagged content to information on land-use data to infer potential activities, and found it to perform well in densely populated areas where gender differences between activities could be distinguished. Similar land-use methods have been used by Kang et al. (2021) alongside mobile phone data to study activity diversity, which they derive from the primary function of surrounding the land use (e.g., commercial, residential, transport etc.). However, the assumption that the land use actually describes the activities of social media users, while partially true in some cases, might break down in closer inspection.

3 Data and methods

3.1 Study areas and temporal scales

The works in this thesis have a diverse range of spatio-temporal scales. The study areas in are located in Finland and span spatial scales from national and regional levels (Figure 3a) down to the level of neighbourhoods (Figure 3b). The analyses in this thesis focus on temporal scales ranging from times of day and weeks to decades.

Article I presents analysis of regional and municipal level differences in the linguistic diversity of Twitter content posted by users from Finland. It explores variations in linguistic diversity and language use across Finland, the linguistic repertoires of Finnish Twitter users, and the home locations of multi- and monolingual Finnish Twitter users. The temporal perspective in Article I is focused on changes in various Finnish regions, municipalities and cities on a weekly level. Article II focuses geographically on national parks located in all four of the general landscape regions of Finland: Lapland fells, Eastern hills, Forests & lakes, and Archipelago. The analysis focus in Article II is on the 20 most popular national parks based on the availability of Flickr photographs (Figure 3a).

Articles III and IV zoom into the Helsinki Metropolitan Area and focus on linguistic diversity on the spatial scale of 250-metre spatial grids that cover the area of Helsinki Metropolitan Area (Figure 3b). Temporally, Article III is focusing on changes across times of day, while Article IV is focusing on annual changes between 1987-2019. The Helsinki Metropolitan Area is the main national hub for political, economic, scientific, and cultural activities. It consists

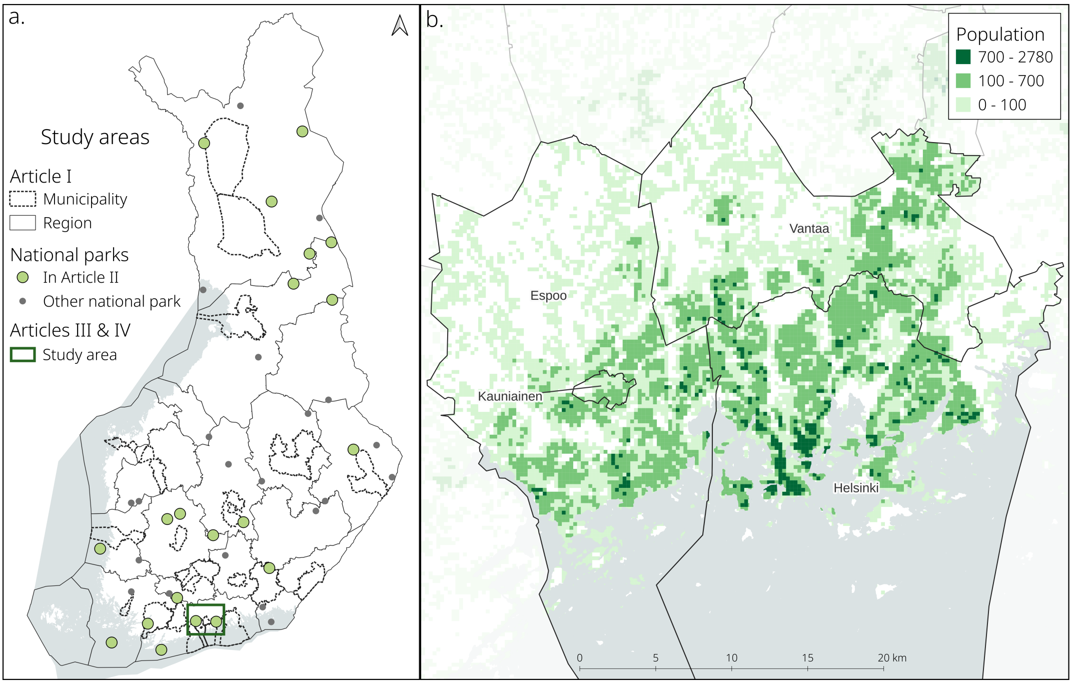

Figure 3: The study areas in this thesis. Articles I and II are focused on Finnish national parks and various regions across Finland (a.), whereas Articles III and IV are focused on the Helsinki Metropolitan Area (b.). The map for Articles III and IV (b.) shows the population density of the Helsinki Metropolitan Area in 2019 and is adapted from Article IV.

of Helsinki (pop. 659 000), the capital of Finland, and three surrounding municipalities, Espoo (pop. 297 000), Vantaa (pop. 239 000), and Kauniainen (pop. 10 000).

3.2 Sources of data

Geographically referenced data are the backbone of all analyses of this thesis. The data can be roughly divided into two categories: user-generated big data and official demographic statistics. More specifically, the sources of data used in the thesis are social media platforms, population registers, and mobile phone operators. Mobile phone and social media data reveal more about the dynamic side of urban diversity, as they are continuously generated by people and their devices during their everyday mobility, whereas population register data reveals more about the structural side of urban diversity, as it is updated annually and based on home locations (Figure 1). Table 2 shows the distribution of data sources by each articles in this dissertation.

Table 2: The main sources of data used in the articles that comprise this dissertation.

Source of data	**Article I**	**Article II**	**Article III**	**Article IV**
		Big data		
Twitter	-	X	X	-
Instagram	-	-	X	-
Flickr	X	-	-	-
Mobile phone data	-	-	X	-
		Traditional data		
Population register	-	-	X	X
Statistical grid database	-	-	X	X

The social media data contains geotagged and non-geotagged content from three platforms: Twitter, Flickr, and Instagram. This data was collected through their Application Programming Interfaces (APIs) using tools written in Python. All social media data were collected by DGL data contributors (2022), and in the case of Article III by the author with a purpose-built Python tool (Väisänen, Heikinheimo, et al., 2021). The mobile phone data used in Article III was provided by the Elisa Oyj mobile network operator and covers the period from October 2017 to January 2018. The data reflects the dynamic population present in 250-metre grid cells in the HMA (see Bergroth et al. (2022) for a detailed description of the mobile phone data).

Twitter is a social media platform focused on microblogging, that is, sharing short written posts, or Tweets, with the maximum length of 280 characters. Tweets can also be accompanied by images and video, but also location information in the form of accurate GPS coordinates until 2015 or a place tag linked to a point of interest (POI) after 2015 (Hu & Wang, 2020; Maurer, 2020; Tasse & Hong, 2017). Articles II and III use different types of Twitter data: Article II is based on both Twitter user timeline data and geotagged Twitter posts, which then is based on a combination of Twitter data from two separate datasets: one collected from Finland between 2015-2019 with the public free API by DGL data contributors (2022) and another between 2009-2013 by Poorthuis and Zook (2017). The user timeline Twitter data contains 3,200 most recent Tweets from users identified to have their home in Finland. Article III data is based on a Twitter data covering the year 2015, which was queried by the author in 2021 using the Academic API with a purpose-built Python tool (Väisänen, Heikinheimo, et al., 2021). Twitter data has been shown to represent population groups that are numerous and more wealthy (Dunn & Adams, 2020). Recently, Twitter opened the full archive to academic researchers for free through the Academic API, with a download quota of 10 million tweets per month (Tornes & Trujillo, 2021). However, the continuity of Twitter data's usefulness for academic research has been recently questioned due to changes in management and data sharing practices. In late 2022, Twitter was bought by Elon Musk, which likely affects research using data from Twitter (Brembs et al., 2023).

Instagram is a social media platform focused on sharing photographs, and is owned by Meta/Facebook. The photographs are often accompanied by textual information in the form of captions. A singular Instagram post can contain numerous photographs, but only one caption. The post can also be enriched with location information, which is nowadays mostly based on place tagging based on POIs, but Article III used data from 2015 where accurate GPS coordinates still were available and the dominant type of location information. The data used in Article III was collected from the Instagram API in early 2016 before Instagram changes to the public API and the eventual shutdown of the public API due to the Cambridge Analytica scandal (Bruns, 2019; Freelon, 2018), but also to comply with GDPR and CCPA legislation (Owuor & Hochmair, 2020).

Flickr is a platform where the sharing of photographs is the key activity, somewhat similar to Instagram. In contrast to Instagram, Flickr is highly popular among nature photographers, making it a good source of data for research on human-nature interactions (Di Minin et al., 2015; Toivonen et al., 2019). In terms of user numbers, Flickr is the least popular platform used in this thesis. A post on Flickr contains the photograph, a caption, and metadata about the geographical location, the time taken, the time shared, and the camera settings. Unlike with Twitter and Instagram, the geotagged locations on Flickr still reflect the accurate GPS coordinates of where photographs were taken, making it a good source of spatially fine-grained content (Hochmair et al., 2018).

The population register is an individual-level database maintained by Statistics Finland, which contains socio-economic information on every individual living in Finland from 1987 onward on an annual basis, although some variables have records since the 1530s (DVV, 2023). The variables contain individual-level information on education, income, employment, marital status, home location, and cultural background of every individual residing in Finland either temporarily or permanently. The articles in this thesis mostly focus on information on individuals' first language, which is recorded in the register as ISO-639-1 language codes from 1999 onward. Pre-1999 records were harmonized to use the ISO-639-1 scheme by using observations from 1999 to replace the values for individuals between 1987-1998. Due to this harmonization, some data loss occurred, as people who had moved abroad or passed away before 1999 are not present in the data from 1999. Unlike censuses conducted at five or ten-year intervals, the register is updated continuously as people are born, pass away, or move, so its accuracy in representing the demographic situation and distribution in Finland is fairly high. The spatial format of the population register is a spatial square grid that consists of 250-metre by 250-metre square grid cells. As the information in the population register is highly accurate and sensitive, access to it is not open for everyone. All outputs from the environment are vetted by officials working for Statistics Finland to ensure privacy laws are followed.

The Statistical Grid Database is a national standard that covers the area of Finland in rectangular grids at two spatial scales: 250-metre and 1-kilometre grids, of which we use the 250-metre grids. This database is based on information from the population register, but is a separate data product. As it is a national standard, the mobile phone data uses it to provide the percentage of population present in each 250-metre grid cell at hourly intervals (Bergroth et al., 2022), and was provided to us by the largest mobile network operator in Finland. The grid database is used in Article III to examine the socio-economic and physical environment of each neighbourhood, and in Article II to capture an estimate of the dynamic population presence in the HMA from the mobile phone data as an independent variable.

3.2.1 Data collection

Social media data was collected from Twitter, Instagram, and Flickr with methods developed by the DGL data contributors (2022), and in Article I's case, the data was enriched with another Twitter dataset formed from the full data archive (Poorthuis & Zook, 2017). Data collection from social media platforms is rather similar regardless of the social media platform. After registering as a developer and receiving access credentials to use the APIs, the data is collected by Python scripts that send requests in the form of queries (e.g. all geotagged posts within a 1-kilometre radius from the city centre of Helsinki) to the API. The API then responds with a JSON (JavaScript Object Notation) payload containing the queried data. The JSON response is then parsed to contain only the relevant data, and the resulting dataset is saved either locally to disk or to a database server running a PostgreSQL database. The majority of the social media data in Articles I-III is geotagged data, however, in Article I, we also collected the 3200 most recent Tweets from users, most of which were not geotagged.

I developed *tweetsearcher*, a Python tool for downloading Tweets (Väisänen, Hiippala, et al., 2021), as a dedicated response to Twitter's decision to provide free and full access to the Twitter archive of tweets for academic researchers in their then-new second version of the API (Tornes & Trujillo, 2021). This tool collects geotagged content from Twitter API using country codes or bounding boxes, but also non-geotagged content from user-specific timelines. I used tweetsearcher in Article III to download a comprehensive dataset of all geotagged tweets from Finland posted in 2015, as previously collected data by DGL data contributors (2022) were collected using Twitter's public API which provided access to approximately 2 % of all tweets. The future usability of tweetsearcher is uncertain due to recent developments at Twitter regarding their APIs (Brembs et al., 2023).

3.3 Methods

The methods in this thesis are interdisciplinary and form one of the main contributions of the work. Each article of this thesis draws from several fields of study that use different methods (Figure 4). More specifically, this work combines methods from the fields of spatial analysis, computational linguistics, ecology, and machine learning in both data preprocessing and analysis proper (see Table 3).

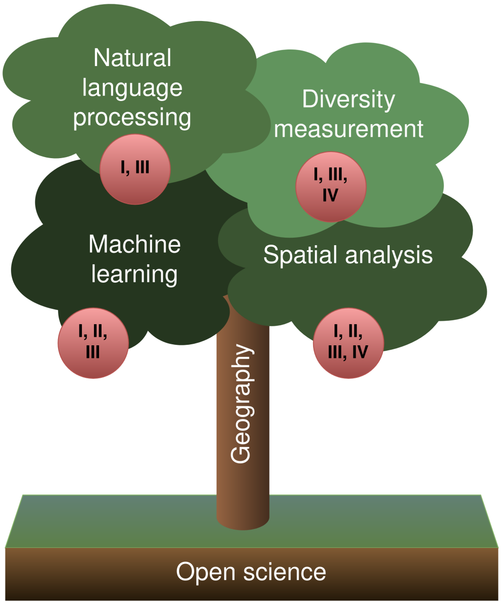

Figure 4: This thesis draws from various fields of study with different methods. Geography as the basis forms the tree trunk, whereas the branches represent different fields of study and the fruits indicate the articles that have benefited from the corresponding branch. However, the principles of open research provide the soil without which much of the work in this thesis would not have been possible.

The articles in this thesis have two methodological backbones. First, spatial analysis methods are the main analysis methods for all articles in this thesis and ground the work firmly within geography. Second, diversity measurement techniques originally developed in ecology and information sciences constitute the other methodological backbone as they are used in three of the four articles (Articles I, III, and IV). The tweetsearcher data collection tool (Väisänen, Hiippala, et al., 2021) can be considered as a methodological contribution of this thesis, albeit a supporting contribution.

Table 3: The main analysis methods used in the articles that comprise this dissertation.

Method	**Article I**	**Article II**	**Article III**	**Article IV**
Data preprocessing				
Social media data collection	X	X	X	-
Automatic language identification	X	-	X	-
User classification	X	X	-	-
Home detection	X	-	-	-
Analysis methods				
Diversity metrics	X	-	X	X
Computer vision	-	X	-	-
Dimensionality reduction	-	X	-	-
Spatial clustering	-	-	X	X
Linear regression	-	-	X	-
Spatial regression	-	-	X	-
Discrete Markov chains	-	-	X	X
Matrix similarity metrics	-	-	-	X

3.3.1 Automatic language identification

Language identification is the task of identifying the language a document or a part of a document is written in with computational tools (Jauhiainen et al., 2019). Automatic language identification was used to enrich the social media data in Articles I and III with information on languages used by Twitter and Instagram users. The language identification performed with fastText (Joulin et al., 2016; Joulin et al., 2017) as it had been recognized to perform well on social media texts (Hiippala et al., 2019). We performed the identification on the sentence level to capture the use of multiple languages and the linguistic repertoires of social media users. As social media texts are free-form in style and length, identification of the language can be challenging (Barman et al., 2014; Carter et al., 2013). I used threshold values for inclusion, where the post had to be more than 12 characters long and the identification confidence higher than 70% following Hiippala et al. (2019). In the case of Twitter, there is information on the language of the tweet identified by Twitter included in the API response, but there is little documentation on how the languages are identified, and no confidence scores provided, which is why we have disregarded this information.

3.3.2 User classification

Classifying users were integral parts of the analyses in Articles I and II. In Article I, we classified Finnish Twitter users into three groups based on the diversity of their language use on Twitter with metrics developed in ecology and information sciences (see Section 3.3.4 for more) following Holloway et al. (2012). This enabled us to understand where the more monolingual and multilingual Finnish Twitter users are from. In Article II, we manually classified the Flickr users with geotagged content from Finnish national parks into domestic and international visitors, and further between male and female users. This classification was performed by myself and Vuokko Heikinheimo using the user profile information of all users in our dataset.

3.3.3 Home detection

We detected the home locations of Finnish Twitter users in Article I using a home detection algorithm developed by Massinen (2019). The method uses geotagged tweets and counts the number of weeks spent within given spatial boundaries, municipalities and regions in the case of Article I (see Figure 3a), and the municipality with the most user weeks is then assigned as the likely home location of the user. In case of a tie, we added 0.5 to the user count for both municipalities to reflect multiple home locations. We used both municipalities and regions to balance the number of Twitter users and content from each type of administrative area: a focus on municipalities alone would have rendered many rural municipalities void of Twitter content, whereas a focus on regions would have omitted finer scale spatio-temporal patterns. In the end, we chose 25 municipalities with the highest user counts to complement the 19 regions in Finland. If the user was not predicted to reside within one of the 25 municipalities, their home location was assigned to the region where the predicted home municipality is located.

3.3.4 Diversity metrics

Diversity metrics are computational tools commonly used in ecology to describe the species-level properties of a sample (Morris et al., 2014). These metrics are widely used in ecology (Magurran & Henderson, 2010; Morris et al., 2014; Sherwin & Prat i Fornells, 2019), but also in segregation studies (Holloway et al., 2012; Massey & Denton, 1988; Reardon & Firebaugh, 2002). Following Peukert (2013), we use these metrics to assess the diversity of languages of spatial units and social media users. Assessing the diversity of languages with various metrics have an integral role throughout this work and constitute the basis of analysis in

Articles I, III, and IV. The language information is based on language codes adhering to the ISO-639-1 standard in social media data and in the population register. The input for the metric calculation is a count vector, where each value represents the count of observations for a particular language in the sample, e.g. all languages used in a certain area or by a certain user. I used the Scikit-bio (2020) Python library to calculate alpha diversity metrics for areas and social media users. I will briefly present the three main metrics (Shannon entropy, Simpson diversity and unique observations) used in this work, the descriptions of the remaining metrics are presented in the Article I.

Shannon entropy is a widely-used metric originating from information sciences, and it describes the amount of information required to describe the identities of individuals in a given sample (Magurran and McGill, 2011, p. 56; Morris et al., 2014). The metric is sensitive both to rare and abundant observations, and thus provides a well-rounded metric for estimating diversity (Magurran & Henderson, 2010; Morris et al., 2014). It is bottom-bounded at 0, representing total lack of diversity (e.g. complete monolingualism), but there is no top bound, although values rarely exceed 4.5 (Ortiz-Burgos, 2016). Shannon entropy has a rather long tradition of use in segregation studies as well (Massey & Denton, 1988; Reardon & Firebaugh, 2002), but it has also been used to describe diversity of languages in virtual and urban environments (Coats, 2019b; Hiippala et al., 2019; Jiang et al., 2022; Peukert, 2013). **Simpson's diversity** describes the probability that two randomly chosen samples are not members of the same group (Morris et al., 2014; Peukert, 2013). As it is a probability distribution, it ranges between 0 and 1, with higher values indicating higher diversity. **Unique observations** is the simplest metric as it reports the number of unique languages regardless of the number of speakers, and is the most commonly applied due to its intuitiveness, despite its sensitivity to sample size (Morris et al., 2014).

3.3.5 Computer vision techniques

Computer vision methods have an integral role in Article II, where we used several pre-trained computer vision models to perform instance-level object detection, scene classification, and semantic clustering on Flickr photographs from Finnish national parks to understand differences in activities and preferences between domestic and international visitors. Instance-level object detection detects the instances of objects in images and can thus provide information on the presence of objects commonly associated with activities, e.g., bicycles. For this we used Mask R-CNN (He et al., 2017) trained on MS COCO (Lin et al., 2014). Scene classification is a computer vision task, where the aim is to classify an image to a class representing a scene, such as "tundra" or "forest path". We used VGG16 (Simonyan & Zisserman, 2015) trained on Places365 (Zhou et al., 2018) to perform scene classification. Finally, semantic clustering (Table 3) is done in two parts. First, the images are passed through a ResNeXt101 model (S. Xie et al., 2017) pre-trained for the task of object detection with ImageNet (Russakovsky et al., 2015). However, the final layer in the neural network that performs the classification is removed, and thus the output from the model is a 2054-dimensional vector of the features in the image instead of a classification like "backpack" with a confidence score. Second, this high dimensional feature vector is then used as an input for dimensionality reduction, after which the clustering becomes possible.

3.3.6 Dimensionality reduction

Dimensionality reduction is a technique that enables visualization and preprocessing of high-dimensional data (McInnes et al., 2020). As the final part of semantic clustering, we performed dimensionality reduction with UMAP (Uniform Manifold Approximation and Projection for Dimensionality Reduction, see McInnes et al. (2020)), which is an unsupervised machine learning technique. UMAP attempts to preserve the local and global structure of the high dimensional data, and can thus provide an overview of the semantic structure in the data. The dimensions of the 2054-dimensional feature vectors are reduced to two dimensions, enabling more intuitive visualizations and interpretations of the data. The resulting two-dimensional data can be plotted as a splatter plot to show how semantically similar content, such as pictures of skiing, clusters together (see Figure 2 in Article II). Semantic clustering thus allows researchers to gain a quick overview of the activities and visual themes in photographic content.

3.3.7 Spatial and statistical methods

Methods from spatial statistics have a paramount role in Articles III and IV. I performed univariate Local Moran's *I* analyses (Anselin, 1995) in Article IV and the bivariate version in III to identify statistically significant areas of high and low linguistic diversity at various points in time. Local Moran's *I* is a local indicator of spatial association, which in turn is a decomposition of the Global Moran's *I*, a metric of spatial autocorrelation (Anselin, 1995). The univariate local Moran's *I* indicates "the extent of significant spatial clustering of similar values" of a single variable around an observation (Anselin, 1995, p. 94), whereas the bivariate version assesses the match between two variables in geographical and attribute spaces (Anselin et al., 2002). In both cases, I used k-NN (k=8) spatial weights matrices to formalize the spatial relationships between the rectangular 250-metre grid cells. In Article III, the bivariate cluster analyses were performed across times of day based on values of Simpson's diversity and Shannon entropy to detect spatio-temporal patterns of multi- and monolingual hotspots, where language contacts are likely to occur.

I performed both spatial and aspatial linear regression analysis in Article III across the times of day to understand how various socio-economic and environmental variables affect spatio-temporal linguistic diversity at different times and whether spatial interaction plays a part in linguistic diversity. For the spatial regression, I used the Spatial Lag Model (SLM) regression analysis, which uses the spatial lag of linguistic diversity in neighbouring locations to assess the strength of spatial interaction (Anselin, 2003). I used a k-NN (k=8) spatial weights matrix to formalize the spatial relationships between the grid cells. For the aspatial regression, we performed an Ordinary Least Squares (OLS) regression analysis across the times of day.

In Article IV, I performed a spatially explicit Markov Chain analysis (Rey, 2001) to explore the spatio-temporal dynamics of linguistic diversity in residential areas. This analysis shows how likely linguistic diversity is to change across residential neighbourhoods of various population groups in the HMA, and how these probabilities change when a grid cell is surrounded by grid cells with differing levels of linguistic diversity. The outputs are probability matrices, which I then compared with Jensen-Shannon distances to understand differences in the change probabilities between neighbourhoods. I used k-NN (k=8) spatial weights matrices to formalize the spatial relationships between the statistical grid cells here as well. Furthermore, I assessed how the spatial concentration of various population groups has changed in the HMA between 1987-2019 with the Delta concentration index (Massey & Denton, 1988).

3.3.8 Software and scripts used in this thesis

The software used by researchers to perform their analyses has an integral role in the research, however, software development is an underappreciated and often unrewarded side of academic work (Arribas-Bel et al., 2021; Merow et al., 2023). I want to encourage citing the tools, especially the free and open-source tools I have used to perform the analyses in this dissertation. Without open availability of such tools, the works in this thesis would look markedly different or not have been even started in the first place.

I performed all analyses with Pandas (Reback et al., 2021), GeoPandas (Jordahl et al., 2021), PySAL (Rey & Anselin, 2010), Scikit-learn (Pedregosa et al., 2011), Scikit-bio (Scikit-bio, 2020), Statsmodels (Seabold & Perktold, 2010), NumPy (Harris et al., 2020), and NLTK (Bird et al., 2009) Python libraries, and GeoDA (Anselin et al., 2006). I visualized all maps with QGIS (QGIS Development Team, 2021), and all graphs and plots with Seaborn (Waskom, 2021) and matplotlib (Hunter, 2007) Python libraries.

In accordance with the principles of open science (UNESCO, 2021), the Python scripts I have written to perform all analyses of every article, besides those done in GeoDA in Article III, are freely and openly available at the following Zenodo repositories:

- Article I: https://doi.org/10.5281/zenodo.4279402
- Article II: https://doi.org/10.5281/zenodo.4282145
- Article III: https://doi.org/10.5281/zenodo.8054821
- Article IV: https://doi.org/10.5281/zenodo.8054946

4 Results and discussion

As the articles that constitute this thesis present their results in detail, in this section I will stay on a more general level and summarize the broader results and implications of this thesis. I will do so through eight claims. For each claim, I present my results supporting the claim, how the results fit in with existing literature, what considerations there might be, and give suggestions for future research.

4.1 The Helsinki Metropolitan Area is becoming linguistically diverse

My results show how the number of spoken languages in the HMA has doubled in the past 35 years. During the same period, the number of languages and linguistic diversity has nearly quadrupled across the neighbourhoods (IV). Furthermore, the number of speakers per each language, except Finnish and Swedish, have increased considerably. For instance, Russian speakers have increased from 915 speakers in 1987 to nearly 34 000 in 2019. My results further reveal large changes in the composition of languages and speakers. In the late 1980s the most common first languages were mostly European languages, whereas in 2019 the composition of languages had gained a more global character with several languages from Africa, the Middle East, and Asia present. While the linguistic diversity overall and across neighbourhoods in the HMA has more than tripled, the relative differences between neighbourhoods have started to reduce.

Moreover, language use across the times-of-day on social media platforms reveal a more diverse linguistic landscape than the population register initially indicated (III). This result shows how social media data that reflects everyday mobility and language use can complement the residential perspective on linguistic diversity by providing a perspective based on language use on social media across times-of-day. As social media users can draw on their full linguistic repertoires and through geotags and timestamps, the location and time of each language use can be mapped. This reflects real-life language use to some extent, as each social media user has the potential to choose a different language depending on the context and intended audience.

Previous research in linguistics has observed that Helsinki has always been multilingual, which contrasts to my findings (IV). However, this observation is not based on register data, but on actual language use and historical documents (Lehtonen, 2016; Nuolijärvi, 2015; Tervonen, 2014). Language use on social media (III) seems to reflect the earlier observations of linguists. The population registry shows the HMA has changed from a monolingual urban area to an increasingly diverse one during the past three decades (IV). Regardless of this change, the HMA remains the least diverse of the Nordic capital city regions (Karlsdottir et al., 2018). Furthermore, much of previous research on multilingualism in the HMA has focused on specific situations and population groups, with a focus on a single year instead of many (Kraus, 2011; Lehtonen, 2016; Nuolijärvi, 2015). Purely a continuous time-series analysis thus addresses a gap in knowledge (IV), but it also illustrates the underexplored nature of linguistic diversity in the HMA. Reflecting the finding about the increasing global character of the popular first languages in the HMA and the languages used on social media (III, IV), Lehtonen (2016) found the "new" slang used by youths to have more influence from Somali and Arabic compared to the "old" slang influenced by Swedish and Russian. Similarly, the composition of languages in social media for the HMA (III) reflect the observations by Hiippala et al. (2019) about social media content being dominated by English and the locally dominant language, Finnish. Finally, these changes in composition of languages and the diversity are in line with the prediction that by 2030 one in four inhabitants in the HMA will speak a language other than Finnish or Swedish as their first language (City of Helsinki, 2022). However, this does not mean the inhabitants would not speak Finnish or Swedish as a second language.

My results showed the relative differences in linguistic diversity are lessening between the neighbourhoods in the HMA (IV). Research has shown that Somali speakers are much more marginalized compared to Estonian speakers, however, the marginalization lessens with time, although prejudice hinders integration (Dhalmann, 2013; Kemppainen et al., 2022; Skovgaard Nielsen et al., 2015), which is also reflected in my results (IV). Estonians have been integrating into society quicker, partly due to less prejudice and fewer socio-cultural barriers (Anniste & Tammaru, 2014). However, the geographical, linguistic, and cultural proximity have rendered this integration only partial as many Estonian speakers live transnational lives in Estonia and Finland (Anniste et al., 2017; Torpan et al., 2022), which is likely represented in the relative differences in linguistic diversity between neighbourhoods with Estonian and native language speakers (IV).

As the HMA is becoming increasingly diverse in terms of languages, more research is needed to better understand the connection of languages to broad social issues like segregation, education and employment trajectories, and housing careers, to name a few (Andersson et al., 2017; City of Helsinki, 2022; Latomaa, 2012; Lehtonen, 2016; Torpan et al., 2022). To this end, information regarding multilingualism and urban diversity in the HMA and Finland has to become more comprehensive. The underlying register data needs to be improved. First, moving from the ISO-639-1 standard to ISO-639-3 would increase the coverage of languages exponentially, from 184 to 7893. To exemplify, the language code for unknown languages describes nearly 8000 inhabitants in the HMA in 2019 and only one Sámi language, Northern Sámi, has a code, whereas speakers of nearly all Sámi languages are likely present in Finland, if not the HMA. It is not known if speakers of other Sámi languages are included under the "se" language code or under the unknown category. Second, including information on the bi- or multilingual individuals and the corresponding skill level would improve the data immensely and provide an improved basis for future research into urban diversity, segregation, integration and migration in Finland (Latomaa, 2012). These deficiencies in the register data have been known about for over a decade (Latomaa, 2012) and reiterated more recently by Saukkonen (2016; 2021b), but at the time of writing they have not been addressed.

Finally, as the increasing linguistic, ethnic, and cultural diversity in the HMA and Finland has been identified in previous research (Kraus, 2011; Lehtonen, 2016; Nuolijärvi, 2015; Saukkonen, 2018, 2021b; Tervonen, 2014; Vilkama, 2010; Vilkama et al., 2013), some have consequently called for societal discussions on what being "Finnish" constitutes (Lehtonen, 2016; Saukkonen, 2018; Tervonen, 2014). My work does not directly contribute to this discussion, but given the trends and degree of linguistic diversity (I, III, IV), questions about Finland's status as either a bilingual and monocultural or a multilingual and multicultural country can be explored.

4.2 Linguistic diversity varies across space and time in the Helsinki Metropolitan Area

My results reveal changing spatio-temporal patterns of linguistic diversity in the HMA from several perspectives (III, IV). The linguistic diversity in social media content reveals a spatio-temporal distribution that stands in stark contrast to that in the population register data (III). Furthermore, linguistically diverse areas are located in largely different areas between the data sources, indicating the locations where exposure to linguistic diversity happens change throughout the day. Second, there are two spatio-temporal patterns in linguistic diversity between 1987 and 2019 (IV). Firstly, the locations where linguistic diversity is concentrated geographically have completely changed. In the late 1980s, these concentrations were in areas with a large proportion of Swedish speakers or with several foreign embassies, whereas by 2019 the concentrations have shifted to highly multilingual suburban neighbourhoods. Secondly, the probability that linguistic diversity in a grid cell changes in the HMA is generally rather low, but the geographical context of the grid cells influence the change probabilities, especially for moderately diverse grid cells. To exemplify, if a moderately diverse grid cell is neighboured by highly diverse grid cells, the moderately diverse grid cell is more likely to turn into a highly diverse grid cell, whereas, grid cells that are either highly monolingual or multilingual are not likely to change, but to remain as they are. This spatio-temporal stability of highly monolingual and multilingual neighbourhoods likely reflects varying moving patterns of the inhabitants, and potentially spatial polarization.

Previous research on the spatio-temporal patterns of linguistic diversity in the HMA is sparse, but the patterns reflect similar findings about socio-economic changes from the HMA and elsewhere. For instance, some suburban neighbourhoods in the HMA have a larger share of disadvantaged population groups because the neighbourhoods have not fully recovered from a recession in the 1990s and the areas have a lot of social housing, which often are the only options for refugees and immigrants from low-income level countries (Dhalmann & Vilkama, 2009; Kauppinen, 2002; Tornes & Trujillo, 2021; Vaattovaara & Bernelius, 2010; Vilkama, 2010; Vilkama et al., 2013). The changing locations of linguistically diverse areas and socio-economic analyses (III, IV) indicate the high linguistic diversity in these areas is likely connected to these broader patterns. The temporal stability of linguistic diversity in residential areas (IV) are similar to general demographic stability of neighbourhoods (Zwiers et al., 2018), but also the moving patterns and neighbourhood preferences (Dhalmann, 2013; Torpan et al., 2022; Vilkama et al., 2013). To exemplify, the stability of linguistic diversity in Somali neighbourhoods shows a strong divergent pattern, meaning that the neighbourhoods are more likely to become either increasingly monolingual or multilingual (IV). This pattern resembles the observation of "white flight" moving patterns identified in Helsinki (Vilkama et al., 2013). The finding of Somali speakers being increasingly exposed to linguistic diversity, but decreasingly to Finnish and Swedish speakers (IV), supports the observation of "white flight" as well. However, the moving patterns are not purely dictated by increased diversity, but a mix of socio-economic issues is at play (Catney, 2016; Kemppainen et al., 2022; Skovgaard Nielsen et al., 2015; Torpan et al., 2022; Vilkama et al., 2013), as is evident in the finding that the linguistic diversity is increasing in all neighbourhoods regardless of the language group (see Figure 4 in IV).

There are some considerations to these findings. First, the register data is spatially based on home locations, which can over-emphasize the linguistic diversity in densely built areas and exaggerate the lack of diversity in sparsely populated areas. Second, as the diversity metrics only consider the language observations within the current grid cell, the low linguistic diversity in less densely populated grid cells might obscure a more multilingual reality, as different languages might be present in neighbouring grid cells. This could be mitigated by alternating the grid cell sizes, or by using focal statistics to enrich a grid cell with information from neighbouring cells. Third, linguistic diversity in social media content is influenced by social media user behaviour, whereby content depicting an exciting and prosperous life is overrepresented and more mundane everyday content is underrepresented (Boy & Uitermark, 2017; Hausmann et al., 2018; B. Huang & Carley, 2019; Koiranen et al., 2020; Manikonda et al., 2016; Martí et al., 2021). Moreover, the languages one uses on social media might not reflect language use in everyday face-to-face interactions, as the intended audiences and socio-spatial contexts affect language choice (Androutsopoulos, 2014; Artamonova & Androutsopoulos, 2019; El Ayadi, 2021; Latomaa, 2012). Fourth, the demographic representativeness of social media users is influenced by platform bias and the digital divide (Robinson, Schulz, Blank, et al., 2020; Robinson, Schulz, Dunn, et al., 2020) and while representativeness can be improved (Heikinheimo et al., 2022; Toivonen et al., 2019) it can not be considered to be representative of more than the users (Hargittai, 2020). The perspective into linguistic diversity afforded by social media data should thus be seen as a complementary perspective and not the complete picture (Arribas-Bel, 2014) that likely represents more socio-economically advantaged people (Hargittai, 2020). Finally, the accuracy of the location information on social media posts has been reducing, as many platforms are moving away from GPS coordinates to points-of-interest information (Hu & Wang, 2020; Maurer, 2020; Tasse & Hong, 2017). This will reduce the spatial granularity of social media content and complicate analyses that require operating on the scale of buildings, streets, and city blocks.

Future work examining spatio-temporality of linguistic diversity in the HMA has several possibilities. First, it is not known whether intergroup language contacts are more likely in linguistically diverse areas identified from social media data or population registers. The link between linguistic diversity on geotagged social media and in the real world needs to be established, which is why actual interactions between individuals of different language groups should not be conflated with interaction on social media. This could be done with ethnographic field work following Pienimäki et al. (2023), or with a more traditional survey study with some public participatory GIS elements included. Second, identifying which elements in the spatio-temporal context of the language contact are most likely to lead to positive outcomes is important (El Ayadi, 2021; Powers et al., 2022; Wessendorf, 2018; Wessendorf & Farrer, 2021; Ye, 2016), as it would give some indication on how these contacts can be supported by urban planners and local NGOs. Finally, identifying whether there are linguistically diverse areas that lack multilingual information and support on how to access public services is important, as populations in these areas are likely under an increased risk of segregation.

4.3 Spatio-temporal inspection provides new perspectives to the study of linguistic diversity

My results show that spatio-temporal approaches to linguistic diversity provide more fine-grained understanding of how linguistic diversity unfolds on regional (I) and local scales (III, IV). First, my work on spatio-temporality of linguistic diversity in social media content addresses a knowledge gap. Some studies have examined temporal changes of linguistic diversity of social media content from one location (Hiippala et al., 2019), or on the level of a city (Jiang et al., 2022), but without a spatial perspective. Other studies have shown that geography affects which languages are used on social media, but the dominant focus has been on country-level differences (Coats, 2019a, 2019b; Graham et al., 2014; L. Hong et al., 2011; Weerkamp et al., 2011) and, to my knowledge, little to no research interest has been given to more regional perspectives (I).

Not only does the linguistic diversity of Finnish Twitter vary regionally and seasonally across Finland, but it also varies depending on the user's home location (I). The most linguistically diverse areas and Twitter users in Finland are in coastal and urban areas, whereas more rural and remote regions are linguistically less diverse. These geographical variations can partly be an effect of the varying popularity level of social media between urban and rural areas in Finland (Koiranen et al., 2020). Furthermore, the higher diversity in coastal regions

(I) coincides with locations of Swedish-speaking communities (Sjöholm, 2004). Urban areas have been shown to influence linguistic content on social media (Grieve et al., 2019; Grieve et al., 2018; B. Huang & Carley, 2019), but these studies focus on lexical innovation and variation among English speakers. Seasonality also affects linguistic diversity across various spatial contexts (Hiippala et al., 2019; Jiang et al., 2022; Mocanu et al., 2013), reflecting the results about temporal variations in linguistic diversity across regions, especially those associated with tourism (I). Furthermore, the presence of Swedish, Russian, and Estonian Twitter content increased in border regions of each nation, indicating the influence of geographical proximity of other states and transnational lifestyles of the users (Hedberg, 2007; Järv et al., 2021; Silm et al., 2020) on the content, reflecting similar findings of the most common languages on Twitter elsewhere (Coats, 2019b; Magdy et al., 2014; Mocanu et al., 2013).

Second, my analyses of linguistic diversity with social media and population register data in the HMA provide both a dynamic and structural view of linguistic diversity (III, IV). Linguistic diversity differs starkly between social media and population register data, indicating the places, where encountering diverse populations is more likely, change throughout the day. The importance of encountering diversity for social cohesion and a sense of community is recognized widely (Powers et al., 2022; Wessendorf, 2014; Ye, 2019), which is why using several data sources (III) to understand where and when linguistically diverse populations are likely to encounter one another is crucial. As exposure to diversity is more likely to happen during the daily mobility (Moya-Gómez et al., 2021; Müürisepp et al., 2023), big data sources which capture daily mobility and language use become important for understanding urban diversity. Encounters with diversity in urban areas can be more likely to occur for more wealthy people (Farber et al., 2012), which echoes my results about the high spatial concentration of Somali speakers, their shorter commute distances, and their neighbourhoods being generally more socio-economically deprived (IV). However, in contrast to the observation of Farber et al. (2012), Somali speakers in the HMA are more exposed to linguistic diversity than any other population group (IV), although the approach in Article IV only considers residential areas, not exposure during daily mobility.

Third, my regression analyses (III) indicate the neighbouring socio-economic and built environment are associated with higher linguistic diversity on social media, which indicates that the well-known effect of the socio-spatial context on language use and choice (El Ayadi, 2021; Valentine et al., 2008) also affects virtual language use to some extent. Similar observations have been made by Kellert and Matlis (2022) about the influence of the spatio-temporal and social context on where formal and informal Spanish is used in Buenos Aires varied. Although not related to linguistic diversity, García-Palomares et al. (2018) also performed regression analyses to exploring how land use affects Twitter activity in Madrid and found similar results on the varying influence of the built environment to the amount of content across times of day, e.g., the increasing importance of commercial areas during evening and night (Tables 4 and 5 in Article IV).

Fourth, my analysis on the changes of linguistic diversity in the HMA with population register data between 1987–2019 (IV) addresses two needs identified by other researchers. The first need regards a call for more temporally continuous approaches to examining long-term changes across neighbourhoods (Casarin et al., 2023; Catney, 2016; Zwiers et al., 2018). This call was made as urban policies impact and changes in neighbourhoods happen gradually over long time periods, and the majority of analyses used data from only two years separated by several years (Catney, 2016; Zwiers et al., 2018). The second need is a more implicit one, emerging from my understanding of literature regarding multilingualism in the HMA, as previous studies used information on languages and their number of speakers from a singular year (Lehtonen, 2016; Nuolijärvi, 2015). Exploring how linguistic diversity developed annually between 1987–2019 thus addresses this “continuity” gap in knowledge to some extent.

Future work examining spatio-temporality of linguistic diversity has several possibilities.

The connection of language use on social media platforms to the geographical locations of language communities has been studied on the level of countries (Coats, 2019a, 2019b; Graham et al., 2014; L. Hong et al., 2011; Magdy et al., 2014; Weerkamp et al., 2011), but much work remains on more regional and local scales. Another possible direction would be to assess the influence of the socio-spatial context on language use on social media following Kellert and Matlis (2022), but focusing on language choice instead of changes between more formal and informal use of one language, and whether the level of diversity in surrounding language use influences language choice across different types of areas.

4.4 Languages provide a view into urban diversity

My results show how information on languages and language use can be used to characterize and describe the spatio-temporal patterns of urban diversity using a variety of data sources (I, III, IV). Information on languages used on social media provides a way to better understand which population groups the users represent (I, III). The linguistic diversity examined through register data is based on first-language information and connected to the home locations of inhabitants. As such it reveals the locations of potential language users, but also where social contacts between language groups are likely to happen when not at work, school or in third places (III, IV). Whereas social media data, which is more connected to leisure and everyday activities, may capture language contacts that may occur in work places, schools, and third places, such as parks, restaurants, and event venues (III). Information on first languages of individuals from population registers can be used to assess and examine the more structural side of urban diversity, but also as a lens to examine the integration of immigrant groups to a host society (IV).

Previous literature has identified language as a potent marker of individual and group identity. Language is thus useful in examining urban diversity, but it has been sparingly used in research on social media data (Dunn et al., 2020; Hiippala et al., 2019). Some studies exploring social media user demographics have identified demographic characteristics of the users using first and last names of social media users (Coats, 2019a; Longley & Adnan, 2016; Longley et al., 2015; Luo et al., 2016). However, the approach is limited as only half of the social media users used a name suitable for identification (Longley & Adnan, 2016; Longley et al., 2015). Another way is to detect the place of residence of the users by using information on the content, social network, and spatio-temporal patterns of social media posts (Heikinheimo et al., 2022). Some have connected the detected place of residence with official demographic statistics of the neighbourhood to assign the user as a likely member of some group (Bernabeu-Bautista et al., 2021; Chapple et al., 2021; L. Li et al., 2013). However, this approach has a risk stemming from ecological fallacy, as an "average" person from a neighbourhood might not exist. Identifying the place of residence (I) and the languages of the user (I, III) can thus provide a good basis for studying urban diversity with social media data.

In research about urban populations not related to big data, language is used more frequently, although much remains unexplored (Chriost & Thomas, 2008; El Ayadi, 2021; Johnston et al., 2001). Language has been found to explain the level of socio-spatial segregation and isolation better than socio-economic variables, ethnic heritage, and origin countries in some geographical contexts (Christopher, 2004; Johnston et al., 2001; Johnston et al., 2021). This might be reflected in the HMA given the differences across socio-economic variables and spatial concentration between speakers of Finnish, Swedish, Somali, and Estonian (IV). Elsewhere, language has not been the explicit focus, but used as one of several variables examined (Catney, 2016; Pisarevskaya et al., 2022) or languages have just been assumed to correlate with other variables of super-diversity (Connor, 2014; Pisarevskaya et al., 2022). However, languages should not be overlooked, as language can reveal additional diversity within the more traditional variables of ethnicity or origin country (Johnston et al., 2021; Vertovec, 2007).

Information on languages from social media and population registers can be used to identify likely areas where contacts between language groups occur (III, IV). For example, study site selection in research on places of encounter and linguistic landscapes is often informed by statistical demographic information on some neighbourhood or local knowledge of the researchers (Blackwood, 2015; Dirksmeier et al., 2014; Hiippala et al., 2023; Hoekstra & Pinkster, 2019; Soukup, 2020; Wessendorf, 2014), but analyses on the spatio-temporality of languages in population registers and social media can support and expand site selection (III, IV). The importance of the places of encounter are known for establishing social cohesion and intercultural understanding in highly diverse neighbourhoods (Ho et al., 2021; Matejskova & Leitner, 2011; Wessendorf, 2014, 2018; Ye, 2019). Information on these places could be leveraged by urban planners and policy-makers in policies that aim to increase social cohesion, resilience, and a sense of community (Fincher et al., 2014). As encounters with "the other" are more likely to lead to positive outcomes in shared, communal, and formal spaces depending on the socio-spatial context (Ho et al., 2021; Hoekstra & Pinkster, 2019; Pienimäki et al., 2023; Powers et al., 2022; Valentine et al., 2008; Wessendorf, 2014), analyses identifying the locations of clusters of linguistic diversity (III, IV) can be used as supporting data in selecting study sites that fall into the three types of places. Moreover, the cluster locations can be enriched with information on points-of-interest to understand in which places these encounters are likely taking place (Psyllidis et al., 2022). Finally, these data-driven approaches for site selection in linguistic landscapes research can help research in the field move beyond typical study sites, like main streets, into unexplored areas (Hiippala et al., 2023).

Using information on languages to understand urban diversity comes with several considerations. First, languages and multilingualism in the real world are somewhat resistant to quantification (Pennycook & Otsuji, 2015, p. 47). Fixed categories are poor in describing fluid and fuzzy entities like languages, which are continuously evolving. Furthermore, an individual speaking in, e.g. Finnish can draw on other languages to add emphasis or alternative meanings, as is commonly done in the Helsinki slang (Lehtonen, 2016). Also, the way languages are spoken and what the act of speaking in a certain language means socio-culturally varies across time and space (Lehtonen, 2016). Furthermore, the language use on social media platforms can also be resistant to categorization as users can draw on multiple languages and use creative spellings (Baldwin et al., 2013; Grieve et al., 2017; Hiippala et al., 2019), but also incorporate multimodal elements, such as emojis, smileys, hashtags, and animated GIFs, that can alter the meaning of the text considerably and do not constitute an identifiable language. The methods and data that I have used cannot capture this kind of multilingualism.

Second, the quantification of languages into a score calculated with a diversity metric can provide a one-sided view on linguistic diversity. Not only are languages fuzzy entities, but different metrics emphasize different aspects of diversity, such as abundant, moderate or rare observations. This underscores the importance of understanding what elements each metric is sensitive to, as naive approaches can perpetuate socio-spatial disadvantages. For instance, as Shannon entropy is sensitive to both rare and abundant observations, whereas moderate observations do not affect the score as much (Magurran & Henderson, 2010; Morris et al., 2014; Peukert, 2013). As diversity metrics are commonly used in ecology, the recommended approach has been to use of two or more metrics simultaneously (Magurran & Henderson, 2010; Morris et al., 2014; Ortiz-Burgos, 2016; Peukert, 2013), which has been my approach throughout this thesis (I, III, IV).

Third, the automatic language identification model used in the analyses (I, III), fastText (Bojanowski et al., 2017; Joulin et al., 2016), is limited to 176 languages and is also based on the ISO-639-1 standard. As a result, it can not identify all languages potentially present on social media (I, III). To exemplify, fastText is not trained on content in Sámi languages and thus cannot identify them. Furthermore, identification of very short texts by automatic methods is challenging, as posts with only a few words or just one have less "contextual" information for the model to identify the language accurately (Barman et al., 2014; Hiippala et al., 2019). In these cases, especially orthographically identical words such as “Helsinki” or “Pizza” are written identically in numerous languages and without other contextualizing information, fastText essentially guesses the language. To reduce this, only posts where the textual content was longer than 12 characters were analysed (I, III). Finally, the multilingual nature of social media text poses a challenge for automatic language identification, as use of multiple languages can occur multiple times in a sentence or a post (Barman et al., 2014). Language identification performed on the sentence level (I, III) tackles this to some degree, but does not capture the use of multiple languages that occur mid-sentence, or code-switching (Barman et al., 2014), whereby some of the diversity in the content is likely lost.

Future work focusing on the role of languages in spatio-temporal urban diversity could focus on several directions. First, exploring the role of languages in explaining socio-spatial disadvantage compared to other variables such as socio-economic status, gender, and age to improve understanding of the role of languages in socio-spatial segregations. Second, work using language information on social media should use techniques that can identify more languages, such as HeLI-OTS (Jauhiainen et al., 2022) or MMS (Pratap et al., 2023), to improve the coverage of previously unaccounted groups of users (Hargittai, 2020). Third, using fine-grained and highly accurate spatial data will be paramount for understanding spatial complexities of urban diversity, as extreme disadvantage and social deprivation can be concentrated in very small geographical areas or to very specific population groups. Moreover, analyses connecting urban diversity to topics like travel-time accessibility of urban amenities and to urban parks and natural areas would provide more understanding about concentration of disadvantage and knowledge urban planners need to support social sustainability (Willberg, Fink, et al., 2023; Willberg, Poom, et al., 2023). Furthermore, calls for explorations of linguistic diversity outside the “traditional” areas, such as metropolises and urban areas, have been made by some scholars, as increasing diversity and globalization also influence more peripheral areas (Catney, 2016; Wang et al., 2014).

4.5 Computer vision can support analyses of diversity when language is not useful

My results show how off-the-shelf computer vision models are capable of extracting information on activities from photographic content on social media, but also how the use of several computer vision techniques concurrently provides complementary perspectives to automatic analysis of visual content (II). This approach is highly relevant with Flickr data, as the textual content on the platform is limited compared to other social media platforms. To exemplify, many posts have no textual content, the textual content is a technical description (e.g. a filename or camera settings), or the same textual content is repeated for several posts. While the models used in Article II (He et al., 2017; Zhou et al., 2018) were not specifically trained to detect activities, but more general tasks like object detection and image classification, the detected contextual information, such as the presence of skis or bicycles in the photographs are implicit signs of activities. Furthermore, the detection of the landscapes in the photographs also provide information on what landscape values the social media users have (van Zanten et al., 2016). Additionally, using the high-dimensional feature vectors, that is, the numerical output from the models before classifying the image to belong in some category, provides information that can be used to semantically cluster the information to reveal the “activity landscape” of the content (II). This approach enables rapid overviews of the visual content and can provide understanding. Furthermore, by classifying the users into foreign and national visitors to the national parks, the differences in activities and visual preferences of the two user groups become distinguishable. Even though the Flickr content is from national parks, there is no apparent reason why the techniques would not be transferrable to urban contexts.

Similar research using computer vision and social media photographs elsewhere has focused on mapping of cultural ecosystem services, landscape preferences, and monitoring of visitors (Chen et al., 2020; Huai et al., 2022; Mouttaki et al., 2022; Santos Vieira et al., 2021; X. P. Song et al., 2020; Staab et al., 2021; Winder et al., 2022). Most of the body of work focuses on national and urban parks, and only Chen et al. (2020) focused primarily on urban areas. Regardless, activity detection is generally a by-product of these broader goals. In the context of cultural ecosystem services, activities are often allocated to two categories: leisure and sport recreation, without more in-depth analyses (Huai et al., 2022; Santos Vieira et al., 2021). Staab et al. (2021) detected visitors from trail camera content, but also their likely activities by detecting recreational equipment carried by visitors. Similarly, Y. Song et al. (2022) derived likely activities of Instagram users in urban parks by detecting objects commonly associated with specific activities, whereas Chen et al. (2020) detected the activities based on the scenes identified in Flickr photographs. These studies have reached similar conclusions (II) about the applicability of computer vision techniques in monitoring the activities in natural areas as a way to complement more labour-intensive manual monitoring field work. Similar, but reversed, gender imbalance was observed in Instagram data, with young women being the majority of users (Y. Song et al., 2022). Several studies have also observed that domestic and foreign visitors have somewhat differing visual preferences and content, with foreign visitors focusing more on content they consider exotic (Huai et al., 2022; Santos Vieira et al., 2021; X. P. Song et al., 2020). Reflecting the findings regarding the prevalence of orienteering (II), the influence of Flickr super-users on the analysis, even when mitigated, has been recognized to affect the results (Winder et al., 2022). Much of the research using computer vision techniques on social media content relies on classification and detection techniques (Ghermandi et al., 2020; Mouttaki et al., 2022; Richards & Tunçer, 2018; Santos Vieira et al., 2021; X. P. Song et al., 2020; Winder et al., 2022), and the call for using several methods in the analysis made in Article II was reiterated by Ghermandi et al. (2022). Finally, there appears not to be work using similar methods to semantic clustering.

There are some considerations to using visual content from social media and computer vision models. First, what is shared on social media has a bias towards content that shows something exclusive, exotic, popular or “cool” especially in visual content (Boy & Uitermark, 2017; Hausmann et al., 2018; Hochman & Manovich, 2013; Manikonda et al., 2016; Tenkanen et al., 2017). Activities inferred from social media photographs thus likely reflect middle- or upper-class lifestyles (Boy & Uitermark, 2017; Hu et al., 2014), as much of the shared content is planned to fit in certain “aesthetic and lifestyle ideals” (Boy & Uitermark, 2017, p. 622). Furthermore, the user classification (II) revealed most content to be generated by male users. Second, the computer vision models are openly available off-the-shelf models, which have been pretrained on massive datasets and replicate the biases in these datasets. As COCO and ImageNet are databases of images from the internet, they replicate gender, cultural, and social biases, because the original images uploaded to the internet contained the biases in the first place (Denton et al., 2021; Lin et al., 2014; Mitchell et al., 2020; Prabhu & Birhane, 2021; Shankar et al., 2017; Steed & Caliskan, 2021; Winder et al., 2022). Finally, social media content is often multimodal with textual, visual and audio content, whereby a more comprehensive understanding of the content could be achieved by analysing all modalities of the content (Toivonen et al., 2019).

Future work on using computer vision techniques should focus on finding relevant areas where to apply these techniques, considering the ethics of using computer vision in urban and natural contexts, and examining the biases that extend from the training data to the final outputs. The strengths and weaknesses of these techniques need to be mapped and understood in order not to perpetuate inequalities and disadvantages, especially given the popularity of smart city and digital twin initiatives across the world, where machine learning methods are integrated into decision and policy pipelines (Biljecki & Ito, 2021; Ibrahim et al., 2020; Vanky & Le, 2023). In terms of research on urban areas using social media data, multimodal models that incorporate both textual and visual content to extract activities and semiotics from the content (Koylu et al., 2019; Lucas et al., 2022; You et al., 2016;

Zhao et al., 2023) would be a natural next step. Another option would be to complement the semantic clustering we performed in Article II by leveraging multilingual or language-agnostic language models like BERT (Devlin et al., 2019) and LASER (Artetxe & Schwenk, 2018), to cluster the semantically similar textual content regardless of the language used in the posts (George & Sumathy, 2023; Q. Xie et al., 2020).

4.6 Diverse data sources are crucial for studying urban diversity

My results show how information on languages and activities from a wide array of data sources can be used to examine urban diversity from several perspectives and on different spatial scales (I, II, III, IV). The analysis based purely on textual content from Finnish Twitter shows how the diversity of languages used on social media platforms varies across Finland spatially, but also per user home locations (I). The challenges of poor textual content and other challenges arising from textual content can be circumvented by focusing on the visual content to explore how activities and visual preferences differ between population groups (II). Combining social media data sources provides a more balanced view of dynamic urban diversity, but also demonstrates how findings from social media data analyses can be contextualized with more established sources of data (III). Finally, the strength of focusing on highly detailed longitudinal register data (IV) enables the investigation of temporal changes in urban diversity at a high level of detail from a structural perspective (Figure 1). As the language information is derived from a national population register, it can be easily combined with other official data sources describing the demographic and socio-economic composition of the residential neighbourhood. More generally, big data sources provide information not available from more traditional sources (I, II, III), such as the spatio-temporality of language use, the degree of multilingualism across multiple spatio-temporal scales, spatio-temporal activities and visual preferences of population groups. However, this does not negate the importance of traditional data sources (III, IV) as these data are not as subject to issues regarding bias and representativeness.

All data sources used in this thesis provide varying perspectives on urban diversity, which is essentially a prerequisite for studying complex social phenomena (Vertovec et al., 2022). Previous research has recognized the biases and representation issues of social media and have proposed approaches to mitigate these issues (Heikinheimo et al., 2022; Martí, Serrano-Estrada, et al., 2019; Martin & Schuurman, 2020; Toivonen et al., 2019). For instance, Hargittai (2020) stresses that as social media data are more likely to reflect the views of people in a higher socio-economic status with good internet skills, which likely affects the diversity of languages and activities based on social media content (I, II, III). There are three common ways to deal with bias from social media: first, combining social media data sources with each other, second, complementing social media sources with traditional data sources, and third, being open about the limitations of the data (Crampton et al., 2013; Hargittai, 2020; Hausmann et al., 2018; Heikinheimo et al., 2017; Herdağdelen, 2013; Kandt & Batty, 2020; Martí, Serrano-Estrada, et al., 2019; Tenkanen et al., 2017; Tu et al., 2017). These approaches are equally important, as there are platform-level differences in the user base and type of content (Arribas-Bel, 2014; Boy & Uitermark, 2017; Hargittai, 2020; Hu et al., 2014; Lansley & Longley, 2016; Longley & Adnan, 2016; Manikonda et al., 2016; Tenkanen et al., 2017). I have adopted these approaches to varying degrees throughout this thesis (I, II, III). Combining traditional and big data sources has been found to be useful in contextualizing social media findings, reducing the possibility of erroneous conclusions, and identifying emerging phenomena (Chapple et al., 2021; Kandt & Batty, 2020; Lansley et al., 2018; Martí, García-Mayor, & Serrano-Estrada, 2019; Martí et al., 2021). In fact, the importance of traditional data will likely increase because of the prevalence of big data (Kandt & Batty, 2020), which is why examining Article III and IV together will likely provide a more holistic understanding of linguistic diversity in the HMA. The perspective into the diversity of languages and activities afforded by social media data on its own (I, II) should thus be seen as a complementary perspective and not the complete picture (Arribas-Bel, 2014; Tenkanen et al., 2017).

As different sources capture different aspects of people, combining data sources can result in highly detailed information on individuals. This raises questions of ethics and equity regarding the analysis, data, methods, and presentation of the results in any analysis. Protecting the privacy of individuals whose data is analysed, especially if the topic is sensitive, while adhering to local privacy laws is paramount (Di Minin et al., 2015; Zook et al., 2017). Acknowledging this, researchers need to carefully consider all analysis steps to reduce potential harm the research may cause (Di Minin et al., 2021; Nelson et al., 2022; Toivonen et al., 2019; Zook et al., 2017). This was a challenge when combining individual-level population register data with data on languages used in geotagged social media content (III). Following the data minimization and pseudonymization principles (Di Minin et al., 2021; Zook et al., 2017), the social media data combined with population register data contained a pseudonymized identifier, information on the language of the post, coordinates and local time. The rest of social media analyses either aggregated information on languages and activities to coarser spatial scales (I, II, III) or grouped information on the user level by variables such as low, moderate, and high diversity (II). With register data (III, IV), all analyses were performed in a secure virtual environment and the outputs were vetted by trained professionals at Statistics Finland before export.

Future works should develop frameworks for combining various data sources for research on urban diversity. Using several social media platforms reduces platform-specific bias and increases the coverage of the population included in the analysis (Hausmann et al., 2018; Owuor & Hochmair, 2020; Tenkanen et al., 2017), however, how to best combine them with more established traditional data sources remains a challenge. As big data and traditional data sources have both strengths and weaknesses, mapping them and communicating about them openly and clearly supports the selection of appropriate data sources for various contexts by researchers and decision-makers (Arribas-Bel, 2014; Janowicz et al., 2015; Kandt & Batty, 2020).

4.7 Research on urban diversity requires open science and interdisciplinarity

Contributing to open research and interdisciplinarity in geography is a common meta-theme throughout this thesis. First, all Python scripts are published openly alongside each publication. These scripts are fully commented and document the analysis workflows from beginning to end, supporting transparency and replicability of each article. Open availability also enables their application across various spatio-temporal contexts and fields of research. Second, each article is published as an open-access article, which ensures better accessibility to the work. Third, by using free and open-source tools to perform the analyses in every article of this thesis, this work demonstrates the value and impact of open source tools for contributing to research on urban diversity. Every article of this thesis would have looked starkly different or not happened at all, if I had had to rely on proprietary software and tools. Finally, each article of this dissertation is inherently interdisciplinary, as they draw on and contribute to GIScience, urban geography, and research on urban multilingualism and linguistic landscapes, while using an interdisciplinary selection of methods from natural language processing, machine learning, ecology, and conservation science.

Interdisciplinarity and openness are necessary in research on urban diversity using big data (Kitchin, 2013; Martin & Schuurman, 2020; Nelson et al., 2022; Vertovec, 2019; Vertovec et al., 2022). Knowledge and understanding about spatio-temporality of diverse urban populations derived from several perspectives is highly important information for supporting social cohesion and resilience (Nelson et al., 2022; Vertovec et al., 2022), as has been argued in each article. Furthermore, democratizing access to science and tools of science, especially research relating to highly diverse populations, is also a question of ethics, social sustainability, and justice (Nelson et al., 2022; UNESCO, 2021). This is why using interdisciplinary methods, publishing open-access articles, and sharing the analysis scripts openly are highly important, which is what I have done with each article (see Section 3.3.8). Moreover, openness enables others to reproduce, build upon, critically evaluate, transfer, and potentially improve the methods (Arribas-Bel et al., 2021; Owuor & Hochmair, 2020).

Open research is not a perfect solution. Publishing open-access articles is a costly endeavour for researchers (Beall, 2012) and places many universities and institutes with less monetary resources at a disadvantage (Kwon, 2022; Smith et al., 2021). The open-access model can paradoxically lead to less diversity, because research performed in rich countries become more widely available, while the high quality and relevant research from poorer countries remains less accessible behind paywalls (Smith et al., 2021) or is published in questionable journals (Taşkın et al., 2023). As Nelson et al. (2022) point out, greater access does not automatically translate to greater equity, as it empowers those with the skills and knowledge to use the open tools and data. Furthermore, providing open access to the research data can be ethically and legally problematic due to sensitive and personally-identifiable information or trade secrets (Di Minin et al., 2021; Martin & Schuurman, 2020; Toivonen et al., 2019). Finally, given the rising popularity of open research (Grahe et al., 2020; Holbrook, 2019; Inkpen et al., 2021), predatory publishers are focusing on open-access publishing models, which can challenge the credibility of open research (Beall, 2012).

Interdisciplinarity also poses some challenges. Contemporary urban diversity is highly complex (Vertovec, 2007; Vertovec et al., 2022; Wessendorf, 2014), and understanding such phenomena "requires a combination of approaches" (Casadevall & Fang, 2014, p. 1357), which increase the difficulty of performing the necessary analyses. Simultaneously, big data has transformed GIScience into an inherently interdisciplinary field (Kitchin, 2013; Martin & Schuurman, 2020; Nelson et al., 2022). This requires geographic research performed with big data to find a balance between "machine learning algorithms, postpositivist subjectivity of user-generated content" (Martin & Schuurman, 2020, p. 1340) and the theories that inform geographic research (Kandt & Batty, 2020; Kitchin & McArdle, 2016). Balancing between requirements of interdisciplinarity and field-specific expertise is difficult for individual researchers, but necessary to address the complex challenges contemporary human civilization faces (Casadevall & Fang, 2014; Nelson et al., 2022; Simon & Graybill, 2010; Ye, 2019).

4.8 More information on the diversity of people and places is needed

This thesis shows how linguistic diversity is increasing at multiple geographical scales in Finland and the HMA, but also provides a methodological framework for studying urban diversity. Given the complexity of urban diversity, a diverse selection of data sources and methods are necessary to extract meaningful information from big data and traditional data sources (I, II, II, IV). The information extracted helps to generate knowledge and understanding about the more short-term and dynamic processes (I, II, III) and the more structural long-term processes (III, IV) associated with urban diversity. Traditional data sources, like the population register data (III, IV), offer a perspective to long-term and structural changes in urban diversity through annual information on residential demographics and the built environment. This perspective is necessary as it provides the "canvas" on which everyday mobility and activities take place, but also points to potentially important residential areas where encounters with "the other" are more likely. Big data sources like social media data can then complement this information by addressing the diversity of languages and activities that emerge from the mobility and activities of people (I, II, III).

Accounting for urban diversity is recognized as an important endeavour that supports social sustainability, well-being, and cohesion (Fincher et al., 2014; Schroedler et al., 2023). As the local matters in urban diversity (OECD, 2018; Syrett & Sepulveda, 2012; United Nations, 2022), it is imperative for cities to keep up with the emerging spatio-temporal patterns of urban diversity (I, II, III, IV) to support the development of socially sustainable, resilient and thriving urban areas (El Ayadi, 2021; Fincher et al., 2014; Gorter, 2006). Frequent, even fleeting, exposure to diversity can increase social cohesion and improve the sense of community as interaction between diverse people can generate understanding, while reducing tensions and prejudices (Chriost & Thomas, 2008; Powers et al., 2022; Wessendorf, 2014). However, diversity can also lead to increased tensions and conflicts due to the lack of a common language, unusual use of public space, and existing prejudices (Hoekstra & Pinkster, 2019; Matejskova & Leitner, 2011; Pienimäki et al., 2023; Powers et al., 2022; Valentine et al., 2008). Being able to support the positive effects of urban diversity, research on urban diversity from various perspectives and a wide selection of methods and sources of data is necessary (I, II, III, IV). Luckily, the increased amount and availability of geospatial big data on urban diversity enables new approaches to understanding urban phenomena (Arribas-Bel, 2014; Arribas-Bel et al., 2015; Batty, 2010; Martí, Serrano-Estrada, et al., 2019; Martin & Schuurman, 2020) that traditional data sources can not map (Batty, 2010; Martin & Schuurman, 2020), but which are also needed to ground the findings (Hargittai, 2020; Kandt & Batty, 2020).

Future work on urban diversity should focus on exploring which selections of variables of urban diversity in which geographical context capture the phenomenon best (Vertovec et al., 2022). Gaining increasingly detailed information on the diversity of urban populations and their local patterns are essential in ensuring social sustainability and cohesion in highly diverse urban environments (Syrett & Sepulveda, 2012; United Nations, 2022), which is why more detailed data about diverse populations and their interactions are needed. Work on urban diversity could also be extended to examining how accessibility (Willberg, Fink, et al., 2023), exposure to environment (Torkko et al., 2023; Willberg, Poom, et al., 2023), and use of urban green areas (Heikinheimo et al., 2020; Powers et al., 2022; X. P. Song et al., 2020) vary across super-diverse population segments.

Given the complexity, broadness, intersectionality, and the increasing rate of urban diversity, the need for more research and interdisciplinary approaches is evident. The need is compounded by the climate crisis, the increasing level of urbanization, and global instability, which drive more people into cities (United Nations, 2022). My thesis has contributed towards this aim by examining urban diversity in the Helsinki Metropolitan Area and by advancing the adoption of an open and interdisciplinary methodological framework in studying urban diversity. While being rooted firmly in geography, my thesis builds especially from geographical information science and reaches out to linguistic landscapes and urban geography. This thesis represents one small step towards improving the understanding of urban diversity and supporting social sustainability of urban areas, much more work from diverse and interdisciplinary perspectives is needed for there to be a leap.

References

Aagesen, H. W., Järv, O., & Gerber, P. (2022). The effect of COVID-19 on cross-border mobilities of people and functional border regions: The Nordic case study from Twitter data. *Geografiska Annaler: Series B, Human Geography*, 1–23. https://doi.org/10.1080/04353684.2022.2101135

Abascal, M., & Baldassarri, D. (2015). Love Thy Neighbor? Ethnoracial Diversity and Trust Reexamined. *American Journal of Sociology*, *121*(3), 722–782. https://doi.org/10.1086/683144

Aboelezz, M. (2015). The Geosemiotics of Tahrir Square: A study of the relationship between discourse and space. *Journal of Language and Politics*, *13*(4), 599–622. https://doi.org/10.1075/jlp.13.4.02abo

Adelfio, M., Serrano-Estrada, L., Martí-Ciriquián, P., Kain, J.-H., & Stenberg, J. (2020). Social Activity in Gothenburg's Intermediate City: Mapping Third Places through Social Media Data. *Applied Spatial Analysis and Policy*, (13), 985–1017. https://doi.org/10.1007/s12061-020-09338-3

Ahas, R., Silm, S., Järv, O., Saluveer, E., & Tiru, M. (2010). Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. *Journal of Urban Technology*, *17*(1), 3–27. https://doi.org/10.1080/10630731003597306

Alexander, C., Edwards, R., & Temple, B. (2007). Contesting Cultural Communities: Language, Ethnicity and Citizenship in Britain. *Journal of Ethnic and Migration Studies*, *33*(5), 783–800. https://doi.org/10.1080/13691830701359223

Amin, A. (2002). Ethnicity and the Multicultural City: Living with Diversity. *Environment and Planning A: Economy and Space*, *34*(6), 959–980. https://doi.org/10.1068/a3537

Andersson, R., Brattbakk, I., & Vaattovaara, M. (2017). Natives' opinions on ethnic residential segregation and neighbourhood diversity in Helsinki, Oslo and Stockholm. *Housing Studies*, *32*(4), 491–516. https://doi.org/10.1080/02673037.2016.1219332

Andrienko, G., Andrienko, N., Bosch, H., Ertl, T., Fuchs, G., Jankowski, P., & Thom, D. (2013). Thematic Patterns in Georeferenced Tweets through Space-Time Visual Analytics. *Computing in Science & Engineering*, *15*(3), 72–82. https://doi.org/10.1109/MCSE.2013.70

Andrienko, N., Andrienko, G., Fuchs, G., & Jankowski, P. (2016). Scalable and privacy-respectful interactive discovery of place semantics from human mobility traces. *Information Visualization*, *15*(2), 117–153. https://doi.org/10.1177/1473871615581216

Androutsopoulos, J. (2014). Moments of sharing: Entextualization and linguistic repertoires in social networking. *Journal of Pragmatics*, *73*, 4–18. https://doi.org/10.1016/j.pragma.2014.07.013

Anniste, K., Pukkonen, L., & Paas, T. (2017). Towards incomplete migration: Estonian migration to Finland. *Trames. Journal of the Humanities and Social Sciences*, *21*(2), 97–114. https://doi.org/10.3176/tr.2017.2.01

Anniste, K., & Tammaru, T. (2014). Ethnic differences in integration levels and return migration intentions: A study of Estonian migrants in Finland. *Demographic Research*, *30*(13), 377–412. https://doi.org/10.4054/DemRes.2014.30.13

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

Anselin, L. (2003). Spatial Econometrics. In B. H. Baltagi (Ed.), *A Companion to Theoretical Econometrics* (pp. 310–330). Blackwell Publishing Ltd. https://doi.org/10.1002/9780470996249.ch15

Anselin, L., Syabri, I., & Kho, Y. (2006). GeoDa: An Introduction to Spatial Data Analysis. *Geographical Analysis*, *38*(1), 5–22. https://doi.org/10.1111/j.0016-7363.2005.00671.x

Anselin, L., Syabri, I., & Smirnov, O. (2002). Visualizing Multivariate Spatial Correlation with Dynamically Linked Windows. *New Tools for Spatial Data Analysis: Proceedings of the Specialist Meeting*, 1–20.

Arribas-Bel, D., Alvanides, S., Batty, M., Crooks, A., See, L., & Wolf, L. (2021). Urban data/code: A new EP-B section. *Environment and Planning B: Urban Analytics and City Science*, *48*(9), 2517–2519. https://doi.org/10.1177/23998083211059670

Arribas-Bel, D. (2014). Accidental, open and everywhere: Emerging data sources for the understanding of cities. *Applied Geography*, *49*, 45–53. https://doi.org/10.1016/j.apgeog.2013.09.012

Arribas-Bel, D., Kourtit, K., Nijkamp, P., & Steenbruggen, J. (2015). Cyber Cities: Social Media as a Tool for Understanding Cities. *Applied Spatial Analysis and Policy*, *8*, 231–247. https://doi.org/10.1007/s12061-015-9154-2

Artamonova, O., & Androutsopoulos, J. (2019). Smartphone-Based Language Practices among Refugees: Mediational Repertoires in Two Families. *Journal für Medienlinguistik*, *2*(2), 60–89. https://doi.org/10.21248/jfml.2019.14

Artetxe, M., & Schwenk, H. (2018). Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, *7*, 597–610. https://doi.org/10.1162/tacl_a_00288

Ash, J., Kitchin, R., & Leszczynski, A. (2018). Digital turn, digital geographies? *Progress in Human Geography*, *42*(1), 25–43. https://doi.org/10.1177/0309132516664800

Baldwin, T., Cook, P., Lui, M., MacKinlay, A., & Wang, L. (2013). How Noisy Social Media Text, How Diffrnt Social Media Sources? *Proceedings of the 6th International Joint Conference on Natural Language Processing*, 356–364.

Barman, U., Das, A., Wagner, J., & Foster, J. (2014). Code Mixing: A Challenge for Language Identification in the Language of Social Media. *Proceedings of the First Workshop on Computational Approaches to Code Switching*, 13–23. https://doi.org/10.3115/v1/W14-3902

Barni, M., & Bagna, C. (2015). The critical turn in LL. *Linguistic Landscape*, *1*(1-2), 6–18. https://doi.org/10.1075/ll.1.1-2.01bar

Batty, M. (2010). The Pulse of the City. *Environment and Planning B: Planning and Design*, *37*(4), 575–577. https://doi.org/10.1068/b3704ed

Batty, M. (2012). Smart Cities, Big Data. *Environment and Planning B: Planning and Design*, *39*(2), 191–193. https://doi.org/10.1068/b3902ed

Batty, M. (2018). Digital twins. *Environment and Planning B: Urban Analytics and City Science*, *45*(5). https://doi.org/10.1177/2399808318796416

Beall, J. (2012). Predatory publishers are corrupting open access. *Nature*, *489*(7415), 179–179. https://doi.org/10.1038/489179a

Beebeejaun, Y. (2022). Whose diversity? Race, space and the European city. *Journal of Urban Affairs*. https://doi.org/10.1080/07352166.2022.2075269

Bereitschaft, B., & Cammack, R. (2015). Neighborhood diversity and the creative class in Chicago. *Applied Geography*, *63*, 166–183. https://doi.org/10.1016/j.apgeog.2015.06.020

Bergroth, C., Järv, O., Tenkanen, H., Manninen, M., & Toivonen, T. (2022). A 24-hour population distribution dataset based on mobile phone data from Helsinki Metropolitan Area, Finland. *Scientific Data*, *9*(1). https://doi.org/10.1038/s41597-021-01113-4

Bernabeu-Bautista, Á., Serrano-Estrada, L., Perez-Sanchez, V. R., & Martí, P. (2021). The Geography of Social Media Data in Urban Areas: Representativeness and Complementarity. *ISPRS International Journal of Geo-Information*, *10*(11), 1–28. https://doi.org/10.3390/ijgi10110747

Biljecki, F., & Ito, K. (2021). Street view imagery in urban analytics and GIS: A review. *Landscape and Urban Planning*, *215*, 1–20. https://doi.org/10.1016/j.landurbplan.2021.104217

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Biró, E. (2018). More Than a Facebook Share: Exploring Virtual Linguistic Landscape. *Acta Universitatis Sapientiae, Philologica*, *10*(2), 181–192. https://doi.org/10.2478/ausp-2018-0022

Blackwood, R. (2015). LL explorations and methodological challenges: Analysing France's regional languages. *Linguistic Landscape*, *1*(1-2), 38–53. https://doi.org/10.1075/ll.1.1-2.03bla

Blommaert, J. (2014). Infrastructures of superdiversity: Conviviality and language in an Antwerp neighborhood. *European Journal of Cultural Studies*, *17*(4), 431–451. https://doi.org/10.1177/1367549413510421

Blommaert, J., & Maly, I. (2019). Invisible Lines in the Online-Offline Linguistic Landscape. *Tilburg Papers in Culture Studies*, *223*, 1–9.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. https://doi.org/10.1162/tacl_a_00051

Bolt, G., Phillips, D., & Van Kempen, R. (2010). Housing Policy, (De)segregation and Social Mixing: An International Perspective. *Housing Studies*, *25*(2), 129–135. https://doi.org/10.1080/02673030903564838

Boy, J. D., & Uitermark, J. (2016). How to Study the City on Instagram. *PLoS ONE*, *11*(6). https://doi.org/10.1371/journal.pone.0158161

Boy, J. D., & Uitermark, J. (2017). Reassembling the city through Instagram. *Transactions of the Institute of British Geographers*, *42*(4), 612–624. https://doi.org/10.1111/tran.12185

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

Brembs, B., Lenardic, A., & Chan, L. (2023). Mastodon: A move to publicly owned scholarly knowledge. *Nature*, *614*(7949), 624–624. https://doi.org/10.1038/d41586-023-00486-3

Brenning, A., & Henn, S. (2023). Web scraping: A promising tool for geographic data acquisition.

Bruns, A. (2019). After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Information Communication and Society*, *22*(11), 1544–1566. https://doi.org/10.1080/1369118X.2019.1637447

Cadier, L., & Mar-Molinero, C. (2014). Negotiating networks of communication in a superdiverse environment: Urban multilingualism in the City of Southampton. *Multilingua*, *33*(5-6), 505–524. https://doi.org/10.1515/multi-2014-0026

Calafiore, A., Palmer, G., Comber, S., Arribas-Bel, D., & Singleton, A. (2021). A geographic data science framework for the functional and contextual analysis of human dynamics within global cities. *Computers, Environment and Urban Systems*, *85*. https://doi.org/10.1016/j.compenvurbsys.2020.101539

Cao, X., Macnaughton, P., Deng, Z., Yin, J., Zhang, X., & Allen, J. G. (2018). Using Twitter to better understand the spatiotemporal patterns of public sentiment: A case study in Massachusetts, USA. *International Journal of Environmental Research and Public Health*, *15*(2). https://doi.org/10.3390/ijerph15020250

Carter, S., Weerkamp, W., & Tsagkias, M. (2013). Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation*, *47*(1), 195–215. https://doi.org/10.1007/s10579-012-9195-y

Casadevall, A., & Fang, F. C. (2014). Specialized science. *Infection and Immunity*, *82*(4), 1355–1360. https://doi.org/10.1128/IAI.01530-13

Casarin, G., MacLeavy, J., & Manley, D. (2023). Rethinking urban utopianism: The fallacy of social mix in the 15-minute city. *Urban Studies*. https://doi.org/10.1177/00420980231169174

Catney, G. (2016). The Changing Geographies of Ethnic Diversity in England and Wales, 1991–2011. *Population, Space and Place*, *22*(8), 750–765. https://doi.org/10.1002/psp.1954

Catney, G., Lloyd, C. D., Ellis, M., Wright, R., Finney, N., Jivraj, S., & Manley, D. (2023). Ethnic diversification and neighbourhood mixing: A rapid response analysis of the 2021 Census of England and Wales. *The Geographical Journal*, *189*(1), 63–77. https://doi.org/10.1111/geoj.12507

Chai, J., Zeng, H., Li, A., & Ngai, E. W. T. (2021). Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, *6*, 100134. https://doi.org/10.1016/j.mlwa.2021.100134

Chapple, K., Poorthuis, A., Zook, M., & Phillips, E. (2021). Monitoring streets through tweets: Using user-generated geographic information to predict gentrification and displacement. *Environment and Planning B: Urban Analytics and City Science*. https://doi.org/10.1177/23998083211025309

Chen, M., Arribas-Bel, D., & Singleton, A. (2020). Quantifying the Characteristics of the Local Urban Environment through Geotagged Flickr Photographs and Image Recognition. *ISPRS International Journal of Geo-Information*, *9*(4), 264. https://doi.org/10.3390/ijgi9040264

Chriost, D. M. G., & Thomas, H. (2008). Linguistic Diversity and the City: Some Reflections, and a Research Agenda. *International Planning Studies*, *13*(1), 1–11. https://doi.org/10.1080/13563470801969624

Christopher, A. J. (2004). Linguistic segregation in urban South Africa, 1996. *Geoforum*, *35*(2), 145–156. https://doi.org/10.1016/j.geoforum.2003.08.007

Chun, C. W. (2014). Mobilities of a linguistic landscape at Los Angeles City Hall Park. *Journal of Language and Politics*, *13*(4), 653–674. https://doi.org/10.1075/jlp.13.1.04chu

City of Helsinki. (2022). Helsingin seudun väestö äidinkielen mukaan 31.12.1999- sekä v. 2009-2018 laaditut väestöennusteet.

Coats, S. (2019a). Language choice and gender in a Nordic social media corpus. *Nordic Journal of Linguistics*, *42*(1), 31–55. https://doi.org/10.1017/S0332586519000039

Coats, S. (2019b). Online Language Ecology: Twitter in Europe. In E. Stemle & C. Wigham (Eds.), *Building Computer-Mediated Communication Corpora for sociolinguistic Analysis* (pp. 73–96). Clermont-Ferrand: Presses universitaires Blaise Pascal.

Connor, P. (2014). Quantifying immigrant diversity in Europe. *Ethnic and Racial Studies*, *37*(11), 2055–2070. https://doi.org/10.1080/01419870.2013.809131

Cosgrove, D. (1985). Prospect, Perspective and the Evolution of the Landscape Idea. *Transactions of the Institute of British Geographers*, *10*(1), 45. https://doi.org/10.2307/622249

Crampton, J. W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M. W., & Zook, M. (2013). Beyond the geotag: Situating 'big data' and leveraging the potential of the geoweb. *Cartography and Geographic Information Science*, *40*(2), 130–139. https://doi.org/10.1080/15230406.2013.777137

Croitoru, A., Wayant, N., Crooks, A., Radzikowski, J., & Stefanidis, A. (2015). Linking cyber and physical spaces through community detection and clustering in social media feeds. *Computers, Environment and Urban Systems*, *53*, 47–64. https://doi.org/10.1016/j.compenvurbsys.2014.11.002

Crooks, A., Croitoru, A., Stefanidis, A., & Radzikowski, J. (2013). #Earthquake: Twitter as a Distributed Sensor System. *Transactions in GIS*, *17*(1), 124–147. https://doi.org/10.1111/j.1467-9671.2012.01359.x

Crooks, A., Pfoser, D., Jenkins, A., Croitoru, A., Stefanidis, A., Smith, D., Karagiorgou, S., Efentakis, A., & Lamprianidis, G. (2015). Crowdsourcing urban form and function. *International Journal of Geographical Information Science*, *29*(5), 720–741. https://doi.org/10.1080/13658816.2014.977905

Cvetojevic, S., & Hochmair, H. H. (2021). Modeling interurban mentioning relationships in the U.S. Twitter network using geo-hashtags. *Computers, Environment and Urban Systems*, *87*, 101621. https://doi.org/10.1016/j.compenvurbsys.2021.101621

de Souza e Silva, A. (2006). From Cyber to Hybrid: Mobile Technologies as Interfaces of Hybrid Spaces. *Space and Culture*, *9*(3), 261–278. https://doi.org/10.1177/1206331206289022

Del Gratta, R., Goggi, S., Pardelli, G., & Calzolari, N. (2021). The LRE Map: What does it tell us about the last decade of our field? *Language Resources and Evaluation*, *55*(1), 259–283. https://doi.org/10.1007/s10579-020-09520-6

Denton, E., Hanna, A., Amironesei, R., Smart, A., & Nicole, H. (2021). On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society*, *8*(2). https://doi.org/10.1177/20539517211035955

Derungs, C., Sieber, C., Glaser, E., & Weibel, R. (2020). Dialect borders – political regions are better predictors than economy or religion. *Digital Scholarship in the Humanities*, *35*(2), 276–295. https://doi.org/10.1093/llc/fqz037

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *1*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

DGL data contributors. (2022). DGL Data Library. https://doi.org/10.5281/zenodo.6425396

Dhalmann, H. (2013). Explaining Ethnic Residential Preferences—The Case of Somalis and Russians in the Helsinki Metropolitan Area. *Housing Studies*, *28*(3), 389–408. https://doi.org/10.1080/02673037.2013.759178

Dhalmann, H., & Vilkama, K. (2009). Housing policy and the ethnic mix in Helsinki, Finland: Perceptions of city officials and Somali immigrants. *Journal of Housing and the Built Environment*, *24*(4), 423–439. https://doi.org/10.1007/s10901-009-9159-8

Di Minin, E., Fink, C., Hausmann, A., Kremer, J., & Kulkarni, R. (2021). How to address data privacy concerns when using social media data in conservation science. *Conservation Biology*, *35*(2), 437–446. https://doi.org/10.1111/cobi.13708

Di Minin, E., Tenkanen, H., & Toivonen, T. (2015). Prospects and challenges for social media data in conservation science. *Frontiers in Environmental Science*, *3*. https://doi.org/10.3389/fenvs.2015.00063

Dirksmeier, P., Helbrecht, I., & Mackrodt, U. (2014). Situational places: Rethinking geographies of intercultural interaction in super-diverse urban space. *Geografiska Annaler: Series B, Human Geography*, *96*(4), 299–312. https://doi.org/10.1111/geob.12053

Donoso, G., & Sanchez, D. (2017). Dialectometric analysis of language variation in Twitter. *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, 16–25. https://doi.org/10.18653/v1/w17-1202

Dunn, J., & Adams, B. (2020). Mapping Languages and Demographics with Georeferenced Corpora. *Proceedings of The 15th International Conference on GeoComputation*, 1–16.

Dunn, J., Coupe, T., & Adams, B. (2020). Measuring Linguistic Diversity During COVID-19. *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, 1–10. https://doi.org/10.18653/v1/2020.nlpcss-1.1

DVV. (2023). Population Information System.

Eisenstein, J., O'Connor, B., Smith, N. A., & Xing, E. P. (2014). Diffusion of Lexical Change in Social Media. *PLoS ONE*, *9*(11), e113114. https://doi.org/10.1371/journal.pone.0113114

El Ayadi, N. (2021). Linguistic sound walks: Setting out ways to explore the relationship between linguistic soundscapes and experiences of social diversity. *Social & Cultural Geography*, *23*(2), 227–249. https://doi.org/10.1080/14649365.2019.1707861

Eleta, I., & Golbeck, J. (2014). Multilingual use of Twitter: Social networks at the language frontier. *Computers in Human Behavior*, *41*, 424–432. https://doi.org/10.1016/j.chb.2014.05.005

Farber, S., O'Kelly, M., Miller, H. J., & Neutens, T. (2015). Measuring segregation using patterns of daily travel behavior: A social interaction based model of exposure. *Journal of Transport Geography*, *49*, 26–38. https://doi.org/10.1016/j.jtrangeo.2015.10.009

Farber, S., Páez, A., & Morency, C. (2012). Activity Spaces and the Measurement of Clustering and Exposure: A Case Study of Linguistic Groups in Montreal: *Environment and Planning A: Economy and Space*, *44*(2), 315–332. https://doi.org/10.1068/a44203

Ferreira, D., & Vale, M. (2021). From cyberspace to cyberspatialities? *Fennia - International Journal of Geography*, *199*(1), 113–117. https://doi.org/10.11143/fennia.100343

Fincher, R., Iveson, K., Leitner, H., & Preston, V. (2014). Planning in the multicultural city: Celebrating diversity or reinforcing difference? *Progress in Planning*, *92*, 1–55. https://doi.org/10.1016/j.progress.2013.04.001

Florida, R. (2003). Cities and the Creative Class. *City & Community*, *2*(1), 3–19. https://doi.org/10.1111/1540-6040.00034

Freelon, D. (2018). Computational Research in the Post-API Age. *Political Communication*, *35*(4), 665–668. https://doi.org/10.1080/10584609.2018.1477506

Fu, C., McKenzie, G., Frias-Martinez, V., & Stewart, K. (2018). Identifying spatiotemporal urban activities through linguistic signatures. *Computers, Environment and Urban Systems*, *72*, 25–37. https://doi.org/10.1016/j.compenvurbsys.2018.07.003

García-Palomares, J. C., Salas-Olmedo, M. H., Moya-Gómez, B., Condeço-Melhorado, A., & Gutiérrez, J. (2018). City dynamics through Twitter: Relationships between land use and spatiotemporal demographics. *Cities*, *72*, 310–319. https://doi.org/10.1016/j.cities.2017.09.007

George, L., & Sumathy, P. (2023). An integrated clustering and BERT framework for improved topic modeling. *International Journal of Information Technology*, *15*(4), 2187–2195. https://doi.org/10.1007/s41870-023-01268-w

Gesthuizen, M., Van Der Meer, T., & Scheepers, P. (2009). Ethnic Diversity and Social Capital in Europe: Tests of Putnam's Thesis in European Countries. *Scandinavian Political Studies*, *32*(2), 121–142. https://doi.org/10.1111/j.1467-9477.2008.00217.x

Ghermandi, A., Depietri, Y., & Sinclair, M. (2022). In the AI of the beholder: A comparative analysis of computer vision-assisted characterizations of human-nature interactions in urban green spaces. *Landscape and Urban Planning*, *217*. https://doi.org/10.1016/j.landurbplan.2021.104261

Ghermandi, A., Sinclair, M., Fichtman, E., & Gish, M. (2020). Novel insights on intensity and typology of direct human-nature interactions in protected areas through passive crowdsourcing. *Global Environmental Change*, *65*. https://doi.org/10.1016/j.gloenvcha.2020.102189

Goodchild, M. F. (2013). The quality of big (geo)data. *Dialogues in Human Geography*, *3*(3), 280–284. https://doi.org/10.1177/2043820613513392

Goodchild, M. F. (2016). GIS in the Era of Big Data. *Cybergeo: European Journal of Geography.*

Gorter, D. (2006). Further Possibilities for Linguistic Landscape Research. In D. Gorter (Ed.), *Linguistic Landscape: A New Approach to Multilingualism* (pp. 81–89). Multilingual Matters.

Gorter, D., & Cenoz, J. (2015). Translanguaging and linguistic landscapes. *Linguistic Landscape*, *1*(1-2), 54–74. https://doi.org/10.1075/ll.1.1-2.04gor

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

Graham, M., & Zook, M. (2013). Augmented Realities and Uneven Geographies: Exploring the Geolinguistic Contours of the Web. *Environment and Planning A: Economy and Space*, *45*(1), 77–99. https://doi.org/10.1068/a44674

Grahe, J. E., Cuccolo, K., Leighton, D. C., & Cramblet Alvarez, L. D. (2020). Open Science Promotes Diverse, Just, and Sustainable Research and Educational Outcomes. *Psychology Learning & Teaching*, *19*(1), 5–20. https://doi.org/10.1177/1475725719869164

Grieve, J., Montgomery, C., Nini, A., Murakami, A., & Guo, D. (2019). Mapping Lexical Dialect Variation in British English Using Twitter. *Frontiers in Artificial Intelligence*, *2*, 1–18. https://doi.org/10.3389/frai.2019.00011

Grieve, J., Nini, A., & Guo, D. (2017). Analyzing lexical emergence in Modern American English online. *English Language and Linguistics*, *21*(1), 99–127. https://doi.org/10.1017/S1360674316000113

Grieve, J., Nini, A., & Guo, D. (2018). Mapping Lexical Innovation on American Social Media. *Journal of English Linguistics*, *46*(4), 293–319. https://doi.org/10.1177/0075424218793191

Gruebner, O., Lowe, S. R., Sykora, M., Shankardass, K., Subramanian, S. V., & Galea, S. (2018). Spatio-temporal distribution of negative emotions in New York city after a natural disaster as seen in social media. *International Journal of Environmental Research and Public Health*, *15*(10). https://doi.org/10.3390/ijerph15102275

Hargittai, E. (2020). Potential Biases in Big Data: Omitted Voices on Social Media. *Social Science Computer Review*, *38*(1), 10–24. https://doi.org/10.1177/0894439318788322

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hasan, S., & Ukkusuri, S. V. (2014). Urban activity pattern classification using topic models from online geo-location data. *Transportation Research Part C: Emerging Technologies*, *44*, 363–381. https://doi.org/10.1016/j.trc.2014.04.003

Hausmann, A., Toivonen, T., Fink, C., Heikinheimo, V., Kulkarni, R., Tenkanen, H., & Di Minin, E. (2020). Understanding sentiment of national park visitors from social media data. *People and Nature*, *2*(3), 750–760. https://doi.org/10.1002/pan3.10130

Hausmann, A., Toivonen, T., Slotow, R., Tenkanen, H., Moilanen, A., Heikinheimo, V., & Di Minin, E. (2018). Social Media Data Can Be Used to Understand Tourists' Preferences for Nature-Based Experiences in Protected Areas. *Conservation Letters*, *11*(1), e12343. https://doi.org/10.1111/conl.12343

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2980–2988. https://doi.org/10.1109/ICCV.2017.322

Hedberg, C. (2007). Direction Sweden: Migration fields and cognitive distances of Finland Swedes. *Population, Space and Place*, *13*(6), 455–470. https://doi.org/10.1002/psp.462

Heikinheimo, V., Järv, O., Tenkanen, H., Hiippala, T., & Toivonen, T. (2022). Detecting country of residence from social media data: A comparison of methods. *International Journal of Geographical Information Science*, 1–22. https://doi.org/10.1080/13658816.2022.2044484

Heikinheimo, V., Minin, E. D., 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), 85. https://doi.org/10.3390/ijgi6030085

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

Herdağdelen, A. (2013). Twitter n-gram corpus with demographic metadata. *Language Resources and Evaluation*, *47*(4), 1127–1147. https://doi.org/10.1007/s10579-013-9227-2

Hewidy, H., & Lilius, J. (2022). In the blind spot: Ethnic retailing in Helsinki and the spontaneous placemaking of abandoned spaces. *European Planning Studies*, *30*(8), 1493–1513. https://doi.org/10.1080/09654313.2021.1932763

Hiippala, T., Hausmann, A., Tenkanen, H., & Toivonen, T. (2019). Exploring the linguistic landscape of geotagged social media content in urban environments. *Digital Scholarship in the Humanities*, *34*(2), 290–309. https://doi.org/10.1093/llc/fqy049

Hiippala, T., Väisänen, T., & Pienimäki, H.-M. (2023). A multimodal approach to physical and virtual linguistic landscapes across different spatial scales. In S. Henricson, V. Syrjälä, C. Bagna, & M. Bellinzona (Eds.), *Sociolinguistic Variation in Urban Linguistic Landscapes*. The Finnish Literature Society.

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://doi.org/10.51814/nm.99996

Ho, E. L. E., Liew, J. A., Zhou, G., Chiu, T. Y., Yeoh, B. S. A., & Huang, S. (2021). Shared spaces and “throwntogetherness” in later life: A qualitative GIS study of non-migrant and migrant older adults in Singapore. *Geoforum*, *124*, 132–143. https://doi.org/10.1016/j.geoforum.2021.05.014

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

Hochman, N., & Manovich, L. (2013). Zooming into an Instagram City: Reading the local through social media. *First Monday*, *18*(7). https://doi.org/10.5210/fm.v18i7.4711

Hoekstra, M. S., & Pinkster, F. M. (2019). ‘We want to be there for everyone’: Imagined spaces of encounter and the politics of place in a super-diverse neighbourhood. *Social & Cultural Geography*, *20*(2), 222–241. https://doi.org/10.1080/14649365.2017.1356362

Holbrook, J. B. (2019). Open Science, Open Access, and the Democratization of Knowledge. *Issues in Science and Technology*, *35*(3), 26–28.

Holloway, S. R., Wright, R., & Ellis, M. (2012). The Racially Fragmented City? Neighborhood Racial Segregation and Diversity Jointly Considered. *The Professional Geographer*, *64*(1), 63–82. https://doi.org/10.1080/00330124.2011.585080

Hong, L., Ahmed, A., Gurumurthy, S., Smola, A., & Tsioutsiouliklis, K. (2012). Discovering Geographical Topics in The Twitter Stream. *Proceedings of the 21st International Conference on World Wide Web - WWW '12*, 769–778. https://doi.org/10.1145/2187836.2187940

Hong, L., Convertino, G., & Chi, E. H. (2011). Language Matters in Twitter: A Large Scale Study. In L. A. Adamic, R. Baeza-Yates, & S. Counts (Eds.), *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* (pp. 518–521). The AAAI Press.

Hong, S.-Y. (2020). Linguistic Landscapes on Street-Level Images. *ISPRS International Journal of Geo-Information*, *9*(1), 57. https://doi.org/10.3390/ijgi9010057

Hu, Y., & Wang, R. Q. (2020). Understanding the removal of precise geotagging in tweets. *Nature Human Behaviour*, *4*, 1219–1221. https://doi.org/10.1038/s41562-020-00949-x

Hu, Y., Manikonda, L., & Kambhampati, S. (2014). What we instagram: 8th International Conference on Weblogs and Social Media, ICWSM 2014. *Proceedings of the 8th In-*

ternational Conference on Weblogs and Social Media, ICWSM 2014, 595–598. https://doi.org/10.1609/icwsm.v8i1.14578

Huai, S., Chen, F., Liu, S., Canters, F., & Van de Voorde, T. (2022). Using social media photos and computer vision to assess cultural ecosystem services and landscape features in urban parks. *Ecosystem Services*, *57*. https://doi.org/10.1016/j.ecoser.2022.101475

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*, 365–373. https://doi.org/10.1145/3341161.3342870

Huang, Y., Guo, D., Kasakoff, A., & Grieve, J. (2016). Understanding U.S. regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems*, *59*, 244–255. https://doi.org/10.1016/j.compenvurbsys.2015.12.003

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Hyötyläinen, M. (2019). *Divided by Policy: Urban Inequality in Finland* (Doctoral dissertation). Helsingin yliopisto.

Ibrahim, M. R., Haworth, J., & Cheng, T. (2020). Understanding cities with machine eyes: A review of deep computer vision in urban analytics. *Cities*, *96*, 102481. https://doi.org/10.1016/j.cities.2019.102481

Ilieva, R. T., & McPhearson, T. (2018). Social-media data for urban sustainability. *Nature Sustainability*, *1*, 553–565. https://doi.org/10.1038/s41893-018-0153-6

Inkpen, R., Gauci, R., & Gibson, A. (2021). The values of open data. *Area*, *53*(2), 240–246. https://doi.org/10.1111/area.12682

Ivkovic, D., & Lotherington, H. (2009). Multilingualism in cyberspace: Conceptualising the virtual linguistic landscape. *International Journal of Multilingualism*, *6*(1), 17–36. https://doi.org/10.1080/14790710802582436

Janowicz, K., Van Harmelen, F., Hendler, J. A., & Hitzler, P. (2015). Why the Data Train Needs Semantic Rails. *AI Magazine*, *36*(1), 5. https://doi.org/10.1609/aimag.v36i1.2560

Järv, O., Aagesen, H. W., Väisänen, T., & Massinen, S. (2022). Revealing mobilities of people to understand cross-border regions: Insights from Luxembourg using social media data. *European Planning Studies*, 1–22. https://doi.org/10.1080/09654313.2022.2108312

Järv, O., Ahas, R., & Witlox, F. (2014). Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records. *Transportation Research Part C: Emerging Technologies*, *38*, 122–135. https://doi.org/10.1016/j.trc.2013.11.003

Järv, O., Müürisepp, K., Ahas, R., Derudder, B., & Witlox, F. (2015). Ethnic differences in activity spaces as a characteristic of segregation: A study based on mobile phone usage in Tallinn, Estonia. *Urban Studies*, *52*(14), 2680–2698. https://doi.org/10.1177/0042098014550459

Järv, O., Tominga, A., Müürisepp, K., & Silm, S. (2021). The impact of COVID-19 on daily lives of transnational people based on smartphone data: Estonians in Finland. *Journal of Location Based Services*, *15*(3), 169–197. https://doi.org/10.1080/17489725.2021.1887526

Jauhiainen, T., Jauhiainen, H., & Lindén, K. (2022). HeLI-OTS, Off-the-shelf Language Identifier for Text. *Proceedings of the 13th Conference on Language Resources and Evaluation*, 3912–3922. https://doi.org/10.5281/zenodo.4780897

Jauhiainen, T., Lui, M., Zampieri, M., Baldwin, T., & Lindén, K. (2019). Automatic Language Identification in Texts: A Survey. *Journal of Artificial Intelligence Research*, *65*(1), 675–782. https://doi.org/10.1613/jair.1.11675

Jenkins, A., Croitoru, A., Crooks, A. T., & Stefanidis, A. (2016). Crowdsourcing a Collective Sense of Place. *PLoS ONE*, *11*(4). https://doi.org/10.1371/journal.pone.0152932

Jiang, R., Luo, Q., & Yang, G. (2022). Exploring the geo virtual linguistic landscape of Dublin urban areas: Before and during the COVID-19 outbreak. *International Journal of Multilingualism*, 1–21. https://doi.org/10.1080/14790718.2022.2096615

Johnston, R., Forrest, J., & Poulsen, M. (2001). The Geography of an EthniCity: Residential Segregation of Birthplace and Language Groups in Sydney, 1996. *Housing Studies*, *16*(5), 569–594. https://doi.org/10.1080/02673030120080062

Johnston, R., Forrest, J., & Siciliano, F. (2021). Exploring the residential segregation of Chinese languages and language groups of the Indian subcontinent in Sydney. *Geographical Research*, *59*(4), 554–563. https://doi.org/10.1111/1745-5871.12479

Jones, M. (2003). The Concept of Cultural Landscape: Discourse and Narratives. In H. Palang & G. Fry (Eds.), *Landscape Interfaces: Cultural Heritage in Changing Landscapes* (pp. 21–51). Springer Netherlands. https://doi.org/10.1007/978-94-017-0189-1_3

Jordahl, K., den Bossche, J. V., Fleischmann, M., McBride, J., Wasserman, J., Gerard, J., Badaracco, A. G., Snow, A. D., Tratner, J., Perry, M., Farmer, C., Hjelle, G. A., Cochran, M., Gillies, S., Culbertson, L., Bartos, M., Caria, G., Eubank, N., sangarshanan, . . . abonte. (2021). Geopandas/geopandas: V0.9.0. https://doi.org/10.5281/zenodo.4569086

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). FastText.zip: Compressing text classification models.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*, *2*, 427–431. https://doi.org/10.18653/v1/e17-2068

Kandt, J., & Batty, M. (2020). Smart cities, big data and urban policy: Towards urban analytics for the long run. *Cities*, *109*, 102992. https://doi.org/10.1016/j.cities.2020.102992

Kandylis, G., Maloutas, T., & Sayas, J. (2012). Immigration, inequality and diversity: Socioethnic hierarchy and spatial organization in Athens, Greece. *European Urban and Regional Studies*, *19*(3), 267–286. https://doi.org/10.1177/0969776412441109

Kang, C., Fan, D., & Jiao, H. (2021). Validating activity, time, and space diversity as essential components of urban vitality. *Environment and Planning B: Urban Analytics and City Science*, *48*(5), 1180–1197. https://doi.org/10.1177/2399808320919771

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

Karlsdottir, A., Rispling, L., Norlén, G., Randall, L., Gassen, N. S., Heleniak, T., Peurell, E., Rehn-Mendoza, N., & Lagercrantz, H. (2018). *State of the Nordic Region 2018: Immigration and integration edition* (A. Karlsdottir, L. Rispling, G. Norlén, & L. Randall, Eds.). Nordic Council of Ministers. https://doi.org/10.6027/ANP2018-742

Kauppinen, T. (2002). The beginning of immigrant settlement in the Helsinki metropolitan area and the role of social housing. *Journal of Housing and the Built Environment*, *17*(2), 173–197. https://doi.org/10.1023/A:1015645008211

Keles, U., Yazan, B., & Giles, A. (2020). Turkish-English bilingual content in the virtual linguistic landscape of a university in Turkey: Exclusive de facto language policies. *International Multilingual Research Journal*, *14*(1), 1–19. https://doi.org/10.1080/19313152.2019.1611341

Kellerman, A. (2014). The Satisfaction of Human Needs in Physical and Virtual Spaces. *The Professional Geographer*, *66*(4), 538–546. https://doi.org/10.1080/00330124.2013.848760

Kellert, O., & Matlis, N. (2022). Geolocation of multiple sociolinguistic markers in Buenos Aires. *PLOS ONE*, *17*(9). https://doi.org/10.1371/journal.pone.0274114

Kemppainen, L., Kemppainen, T., Rask, S., Saukkonen, P., & Kuusio, H. (2022). Transnational Activities and Identifications – A population-based study on three immigrant groups in Finland. *Migration and Development*, *11*(3), 762–782. https://doi.org/10.1080/21632324.2020.1830563

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

Kitchin, R., & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society*, *3*(1), 1–10. https://doi.org/10.1177/2053951716631130

Koiranen, I., Keipi, T., Koivula, A., & Räsänen, P. (2020). Changing patterns of social media use? A population-level study of Finland. *Universal Access in the Information Society*, *19*(3), 603–617. https://doi.org/10.1007/s10209-019-00654-1

Koskinen, K. (2013). Social media and the institutional illusions of EU communication. *International Journal of Applied Linguistics (United Kingdom)*, *23*(1), 80–92. https://doi.org/10.1111/ijal.12018

Koylu, C., Larson, R., Dietrich, B. J., & Lee, K.-P. (2019). CarSenToGram: Geovisual text analytics for exploring spatiotemporal variation in public discourse on Twitter. *Cartography and Geographic Information Science*, *46*(1), 57–71. https://doi.org/10.1080/15230406.2018.1510343

Kraus, P. A. (2011). The Multilingual City: *The cases of Helsinki and Barcelona. Nordic Journal of Migration Research*, *1*(1), 25–36. https://doi.org/10.2478/v10202-011-0004-2

Kruse, J., Kang, Y., Liu, Y.-N., Zhang, F., & Gao, S. (2021). Places for play: Understanding human perception of playability in cities using street view images and deep learning. *Computers, Environment and Urban Systems*, *90*. https://doi.org/10.1016/j.compenvurbsys.2021.101693

Kruspe, A., Häberle, M., Hoffmann, E. J., Rode-Hasinger, S., Abdulahhad, K., & Zhu, X. X. (2021). Changes in Twitter geolocations: Insights and suggestions for future usage. *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, 212–221. https://doi.org/10.18653/v1/2021.wnut-1.24

Kumar, D. (2019). Language Discrimination in Indian Higher Education. In P. Singh (Ed.), *Contouring Exclusion: Manifestations and Implication* (pp. 149–169). Lokmitra Publication.

Kwon, D. (2022). Open-access publishing fees deter researchers in the global south. *Nature*. https://doi.org/10.1038/d41586-022-00342-w

Landry, R., & Bourhis, R. Y. (1997). Linguistic Landscape and Ethnolinguistic Vitality. *Journal of Language and Social Psychology*, *16*(1), 23–49. https://doi.org/10.1177/0261927X970161002

Lansley, G., & Longley, P. A. (2016). The geography of Twitter topics in London. *Computers, Environment and Urban Systems*, *58*, 85–96. https://doi.org/10.1016/j.compenvurbsys.2016.04.002

Lansley, G., Smith, M., Goodchild, M., & Longley, P. (2018). Big Data and Geospatial Analysis. In *Geospatial Analysis: A comprehensive guide to principles, techniques and software tools* (pp. 547–570). The Winchelsea Press.

Latomaa, S. (2012). Kielitilasto maahanmuuttajien väestöosuuden mittarina. *Yhteiskuntapolitiikka*, *77*(5), 525–534.

Lazar, M. (2022). Semiotic timescapes. *Language in Society*, *51*(5), 735–748. https://doi.org/10.1017/S0047404522000641

Leeman, J., & Modan, G. (2009). Commodified language in Chinatown: A contextualized approach to linguistic landscape. *Journal of Sociolinguistics*, *13*(3), 332–362. https://doi.org/10.1111/j.1467-9841.2009.00409.x

Lehtonen, H. (2016). What's Up Helsinki?: Linguistic Diversity Among Suburban Adolescents. In R. Toivanen & J. Saarikivi (Eds.), *Linguistic Genocide or Superdiversity?: New and Old Language Diversities* (pp. 65–90). Multilingual Matters.

Leppämäki, T. (2022). *Developing a Finnish geoparser for extracting location information from unstructured texts* (MSc). University of Helsinki. Helsinki, Finland.

Levin, N., Kark, S., & Crandall, D. (2015). Where have all the people gone? Enhancing global conservation using night lights and social media. *Ecological Applications*, *25*(8), 2153–2167. https://doi.org/10.1890/15-0113.1

Li, L., Goodchild, M. F., & Xu, B. (2013). Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science*, *40*(2), 61–77. https://doi.org/10.1080/15230406.2013.777139

Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., Pettit, C., Jiang, B., Haworth, J., Stein, A., & Cheng, T. (2016). Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, *115*, 119–133. https://doi.org/10.1016/j.isprsjprs.2015.10.012

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer Vision – ECCV 2014* (pp. 740–755). Springer International Publishing. https://doi.org/10.1007/978-3-319-10602-1_48

Longley, P. A., & Adnan, M. (2016). Geo-temporal Twitter demographics. *International Journal of Geographical Information Science*, *30*(2), 369–389. https://doi.org/10.1080/13658816.2015.1089441

Longley, P. A., Adnan, M., & Lansley, G. (2015). The Geotemporal Demographics of Twitter Usage. *Environment and Planning A: Economy and Space*, *47*(2), 465–484. https://doi.org/10.1068/a130122p

Lopez, B. E., Magliocca, N. R., & Crooks, A. T. (2019). Challenges and Opportunities of Social Media Data for Socio-Environmental Systems Research. *Land*, *8*(7). https://doi.org/10.3390/land8070107

López Peláez, A., Aguilar-Tablada, M. V., Erro-Garcés, A., & Pérez-García, R. M. (2022). Superdiversity and social policies in a complex society: Social challenges in the 21st century. *Current Sociology*, *70*(2), 166–192. https://doi.org/10.1177/0011392120983344

Lucas, L., Tomás, D., & Garcia-Rodriguez, J. (2022). Detecting and locating trending places using multimodal social network data. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-022-14296-8

Lui, M., & Baldwin, T. (2012). Langid.py: An Off-the-shelf Language Identification Tool. *Proceedings of the ACL 2012 System Demonstrations*, 25–30.

Luo, F., Cao, G., Mulligan, K., & Li, X. (2016). Explore spatiotemporal and demographic characteristics of human mobility via Twitter: A case study of Chicago. *Applied Geography*, *70*, 11–25. https://doi.org/10.1016/j.apgeog.2016.03.001

Magdy, A., Ghanem, T. M., Musleh, M., & Mokbel, M. F. (2014). Exploiting Geo-tagged Tweets to Understand Localized Language Diversity. *Proceedings of Workshop on Managing and Mining Enriched Geo-Spatial Data - GeoRich'14*, 1–6. https://doi.org/10.1145/2619112.2619114

Magdy, A., Ghanem, T. M., Musleh, M., & Mokbel, M. F. (2016). Understanding Language Diversity in Local Twitter Communities. *Proceedings of the 27th ACM Conference on Hypertext and Social Media*, 331–332. https://doi.org/10.1145/2914586.2914612

Magurran, A. E., & Henderson, P. A. (2010). Temporal turnover and the maintenance of diversity in ecological assemblages. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1558), 3611–3620. https://doi.org/10.1098/rstb.2010.0285

Magurran, A. E., & McGill, B. J. (2011). *Biological Diversity: Frontiers in Measurement and Assessment*. Oxford University Press.

Manikonda, L., Meduri, V. V., & Kambhampati, S. (2016). Tweeting the Mind and Instagramming the Heart: Exploring Differentiated Content Sharing on Social Media. *Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016*, 639–642. https://doi.org/10.48550/arXiv.1603.02718

Martí, P., García-Mayor, C., Nolasco-Cirugeda, A., & Serrano-Estrada, L. (2020). Green infrastructure planning: Unveiling meaningful spaces through Foursquare users' preferences. *Land Use Policy*, *97*, 104641. https://doi.org/10.1016/j.landusepol.2020.104641

Martí, P., García-Mayor, C., & Serrano-Estrada, L. (2019). Identifying opportunity places for urban regeneration through LBSNs. *Cities*, *90*, 191–206. https://doi.org/10.1016/j.cities.2019.02.001

Martí, P., Serrano Estrada, L., & Aboutorabi, M. (2021). Culturally Diverse Street-Level Urban Activities through the Lens of Digital Footprints. *Sustainability*, *132*. https://doi.org/10.3390/su132011141

Martí, P., Serrano-Estrada, L., & Nolasco-Cirugeda, A. (2017). Using locative social media and urban cartographies to identify and locate successful urban plazas. *Cities*, *64*, 66–78. https://doi.org/10.1016/j.cities.2017.02.007

Martí, P., Serrano-Estrada, L., & Nolasco-Cirugeda, A. (2019). Social Media data: Challenges, opportunities and limitations in urban studies. *Computers, Environment and Urban Systems*, *74*, 161–174. https://doi.org/10.1016/j.compenvurbsys.2018.11.001

Martin, M. E., & Schuurman, N. (2017). Area-Based Topic Modeling and Visualization of Social Media for Qualitative GIS. *Annals of the American Association of Geographers*, *107*(5), 1028–1039. https://doi.org/10.1080/24694452.2017.1293499

Martin, M. E., & Schuurman, N. (2020). Social Media Big Data Acquisition and Analysis for Qualitative GIScience: Challenges and Opportunities. *Annals of the American Association of Geographers*, *110*(5), 1335–1352. https://doi.org/10.1080/24694452.2019.1696664

Massey, D. S., & Denton, N. A. (1988). The Dimensions of Residential Segregation. *Social Forces*, *67*(2), 281–315. https://doi.org/10.2307/2579183

Massinen, S. (2019). *Modeling Cross-Border Mobility Using Geotagged Twitter in the Greater Region of Luxembourg* (MSc). University of Helsinki. Helsinki, Finland.

Matejskova, T., & Leitner, H. (2011). Urban encounters with difference: The contact hypothesis and immigrant integration projects in eastern Berlin. *Social & Cultural Geography*, *12*(7), 717–741. https://doi.org/10.1080/14649365.2011.610234

Maurer, S. M. (2020). Evolving Approaches to Place Tagging in Social Media. *Journal of Urban Planning and Development*, *146*(3). https://doi.org/10.1061/(ASCE)UP.1943-5444.0000583

McInnes, L., Healy, J., & Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. https://doi.org/10.48550/arXiv.1802.03426

Merow, C., Boyle, B., Enquist, B. J., Feng, X., Kass, J. M., Maitner, B. S., McGill, B., Owens, H., Park, D. S., Paz, A., Pinilla-Buitrago, G. E., Urban, M. C., Varela, S., & Wilson, A. M. (2023). Better incentives are needed to reward academic software development. *Nature Ecology & Evolution*, *7*, 626–627. https://doi.org/10.1038/s41559-023-02008-w

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), 1–27. https://doi.org/10.1145/3202662

Miller, H. J. (2020). Geographic information science III: GIScience, fast and slow – Why faster geographic information is not always smarter. *Progress in Human Geography*, *44*(1), 129–138. https://doi.org/10.1177/0309132518799596

Miller, H. J., & Goodchild, M. F. (2015). Data-driven geography. *GeoJournal*, *80*(4), 449–461. https://doi.org/10.1007/s10708-014-9602-6

Mitchell, M., Baker, D., Moorosi, N., Denton, E., Hutchinson, B., Hanna, A., Gebru, T., & Morgenstern, J. (2020). Diversity and Inclusion Metrics in Subset Selection. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 117–123. https://doi.org/10.1145/3375627.3375832

Mocanu, D., Baronchelli, A., Perra, N., Gonçalves, B., Zhang, Q., & Vespignani, A. (2013). The Twitter of Babel: Mapping World Languages through Microblogging Platforms. *PLoS ONE*, *8*(4). https://doi.org/10.1371/journal.pone.0061981

Moldez, C., & Gomez, D. (2022). Looking at the bigger picture: A semiotic analysis on online news photographs. *International Journal of Research Studies in Education*, *11*(3). https://doi.org/10.5861/ijrse.2022.115

Morris, E. K., Caruso, T., Buscot, F., Fischer, M., Hancock, C., Maier, T. S., Meiners, T., Müller, C., Obermaier, E., Prati, D., Socher, S. A., Sonnemann, I., Wäschke, N., Wubet, T., Wurst, S., & Rillig, M. C. (2014). Choosing and using diversity indices: Insights for ecological applications from the German Biodiversity Exploratories. *Ecology and Evolution*, *4*(18), 3514–3524. https://doi.org/10.1002/ece3.1155

Moshnikov, I. (2016). Karjalankieliset verkkosivut virtuaalisena kielimaisemana. *Lähivõrdlusi. Lähivertailuja*, *26*, 282–310. https://doi.org/10.5128/LV26.09

Moshnikov, I. (2022). Use of the Karelian language online: Websites in Karelian. *Suomen soveltavan kielitieteen yhdistyksen julkaisuja*, *79*, 192–216. https://doi.org/10.30661/afinlavk.113920

Mouttaki, I., Bagdanavičiūtė, I., Maanan, M., Erraiss, M., Rhinane, H., & Maanan, M. (2022). Classifying and Mapping Cultural Ecosystem Services Using Artificial Intelligence and Social Media Data. *Wetlands*, *42*(7), 86. https://doi.org/10.1007/s13157-022-01616-9

Moya-Gómez, B., Stepniak, M., García-Palomares, J. C., Frías-Martínez, E., & Gutiérrez, J. (2021). Exploring night and day socio-spatial segregation based on mobile phone data: The case of Medellin (Colombia). *Computers, Environment and Urban Systems*, *89*. https://doi.org/10.1016/j.compenvurbsys.2021.101675

Mustapha, A. S. (2014). Linguistic Hegemony of the English Language in Nigeria. *Íkala, revista de lenguaje y cultura*, *19*(1), 57–71. https://doi.org/10.17533/udea.ikala.15315

Müürisepp, K., Järv, O., Sjöblom, F., Toger, M., & Östh, J. (2023). Segregation and the pandemic: The dynamics of daytime social diversity during COVID-19 in Greater Stockholm. *Applied Geography*, *154*. https://doi.org/10.1016/j.apgeog.2023.102926

Müürisepp, K., Järv, O., Tammaru, T., & Toivonen, T. (2022). Activity spaces and big data sources in segregation research: A methodological review. *Frontiers in Sustainable Cities*, *4*. https://doi.org/10.3389/frsc.2022.861640

Nelson, T., Goodchild, M., & Wright, D. (2022). Accelerating ethics, empathy, and equity in geographic information science. *Proceedings of the National Academy of Sciences*, *119*(19). https://doi.org/10.1073/pnas.2119967119

Niu, H., & Silva, E. A. (2020). Crowdsourced Data Mining for Urban Activity: Review of Data Sources, Applications, and Methods. *Journal of Urban Planning and Development*, *146*(2). https://doi.org/10.1061/(ASCE)UP.1943-5444.0000566

Niu, H., & Silva, E. A. (2021). Delineating urban functional use from points of interest data with neural network embedding: A case study in Greater London. *Computers, Environment and Urban Systems*, *88*. https://doi.org/10.1016/j.compenvurbsys.2021.101651

Niu, H., & Silva, E. A. (2023). Understanding temporal and spatial patterns of urban activities across demographic groups through geotagged social media data. *Computers, Environment and Urban Systems*, *100*, 101934. https://doi.org/10.1016/j.compenvurbsys.2022.101934

Nuolijärvi, P. (2015). Helsinki as a Multilingual City. In E. Boix-Fuister (Ed.), *Urban Diversities and Language Policies in Medium-Sized Linguistic Communities* (pp. 67–84). Multilingual Matters.

OECD. (2018). *Divided cities: Understanding Intra-urban Inequalities*.

Ortiz-Burgos, S. (2016). Shannon-Weaver Diversity Index. In M. J. Kennish (Ed.), *Encyclopedia of Estuaries* (pp. 572–573). Springer Netherlands. https://doi.org/10.1007/978-94-017-8801-4_233

Owuor, I., & Hochmair, H. H. (2020). An Overview of Social Media Apps and their Potential Role in Geospatial Research. *ISPRS International Journal of Geo-Information*, *9*(9), 526. https://doi.org/10.3390/ijgi9090526

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. https://doi.org/10.48550/arXiv.1201.0490

Pennycook, A., & Otsuji, E. (2015). *Metrolingualism: Language in the City* (1st ed.). Routledge.

Peukert, H. (2013). Measuring language diversity in urban ecosystems. In J. Duarte & I. Goglin (Eds.), *Linguistic Superdiversity in Urban Areas: Research approaches* (pp. 75–96). Benjamins. https://doi.org/10.1075/hsld.2.06peu

Pienimäki, H.-M., Väisänen, T., & Hiippala, T. (2023). Making sense of linguistic diversity in Helsinki, Finland: The timespace of affects in the linguistic landscape. *Journal of Sociolinguistics*, 1–19. https://doi.org/10.1111/josl.12633

Pisarevskaya, A., Scholten, P., & Kaşlı, Z. (2022). Classifying the Diversity of Urban Diversities: An Inductive Analysis of European Cities. *Journal of International Migration and Integration*, *23*(2), 655–677. https://doi.org/10.1007/s12134-021-00851-z

Poorthuis, A., Shelton, T., & Zook, M. (2021). Changing neighborhoods, shifting connections: Mapping relational geographies of gentrification using social media data. *Urban Geography*, 1–24. https://doi.org/10.1080/02723638.2021.1888016

Poorthuis, A., & Zook, M. (2017). Making Big Data Small: Strategies to Expand Urban and Geographical Research Using Social Media. *Journal of Urban Technology*, *24*(4), 115–135. https://doi.org/10.1080/10630732.2017.1335153

Powers, S. L., Webster, N., Agans, J. P., Graefe, A. R., & Mowen, A. J. (2022). The power of parks: How interracial contact in urban parks can support prejudice reduction, interracial trust, and civic engagement for social justice. *Cities*, *131*, 104032. https://doi.org/10.1016/j.cities.2022.104032

Prabhu, V. U., & Birhane, A. (2021). Large image datasets: A pyrrhic win for computer vision? *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1536–1546. https://doi.org/10.1109/WACV48630.2021.00158

Pratap, V., Tjandra, A., Shi, B., Babu, P. T. A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Baevski, M. F.-Z. A., Adi, Y., Zhang, X., Hsu, W.-N., Conneau, A., & Auli, M. (2023). Scaling Speech Technology to 1,000+ Languages. https://doi.org/10.48550/arXiv.2305.13516

Psyllidis, A., Gao, S., Hu, Y., Kim, E.-K., McKenzie, G., Purves, R., Yuan, M., & Andris, C. (2022). Points of Interest (POI): A commentary on the state of the art, challenges, and prospects for the future. *Computational Urban Science*, *2*(1), 20. https://doi.org/10.1007/s43762-022-00047-w

Putnam, R. D. (2007). E Pluribus Unum: Diversity and Community in the Twenty-first Century The 2006 Johan Skytte Prize Lecture. *Scandinavian Political Studies*, *30*(2), 137–174. https://doi.org/10.1111/j.1467-9477.2007.00176.x

QGIS Development Team. (2021). QGIS Geographic Information System.

Reardon, S. F., & Firebaugh, G. (2002). Measures of Multigroup Segregation. *Sociological Methodology*, *32*, 33–67. https://doi.org/10.1111/1467-9531.00110

Reback, J., McKinney, W., jbrockmendel, den Bossche, J. V., Augspurger, T., Cloud, P., Hawkins, S., gfyoung, Sinhrks, Roeschke, M., Klein, A., Petersen, T., Tratner, J., She, C., Ayd, W., Naveh, S., patrick, Garcia, M., Schendel, J., . . . h-vetinari. (2021). Pandas-dev/pandas: Pandas 1.2.4. https://doi.org/10.5281/zenodo.4681666

Rey, S. J. (2001). Spatial Empirics for Economic Growth and Convergence. *Geographical Analysis*, *33*(3), 195–214. https://doi.org/10.1111/j.1538-4632.2001.tb00444.x

Rey, S. J., & Anselin, L. (2010). PySAL: A Python Library of Spatial Analytical Methods. In M. M. Fischer & A. Getis (Eds.), *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications* (pp. 175–193). Springer. https://doi.org/10.1007/978-3-642-03647-7_11

Richards, D. R., & Tunçer, B. (2018). Using image recognition to automate assessment of cultural ecosystem services from social media photographs. *Ecosystem Services*, *31*, 318–325. https://doi.org/10.1016/j.ecoser.2017.09.004

Roberts, C. (2010). Language Socialization in the Workplace. *Annual Review of Applied Linguistics*, *30*, 211–227. https://doi.org/10.1017/S0267190510000127

Robinson, L., Schulz, J., Blank, G., Ragnedda, M., Ono, H., Hogan, B., Mesch, G. S., Cotten, S. R., Kretchmer, S. B., Hale, T. M., Drabowicz, T., Yan, P., Wellman, B., Harper, M.-G., Quan-Haase, A., Dunn, H. S., Casilli, A. A., Tubaro, P., Carvath, R., . . . Khilnani, A. (2020). Digital inequalities 2.0: Legacy inequalities in the information age. *First Monday*, *25*(7). https://doi.org/10.5210/fm.v25i7.10842

Robinson, L., Schulz, J., Dunn, H. S., Casilli, A. A., Tubaro, P., Carvath, R., Chen, W., Wiest, J. B., Dodel, M., Stern, M. J., Ball, C., Huang, K.-T., Blank, G., Ragnedda, M., Ono, H., Hogan, B., Mesch, G. S., Cotten, S. R., Kretchmer, S. B., . . . Khilnani, A. (2020). Digital inequalities 3.0: Emergent inequalities in the information age. *First Monday*, *25*(7). https://doi.org/10.5210/fm.v25i7.10844

Rose, G. (2022). Introduction: Seeing The City Digitally. In G. Rose (Ed.), *Seeing the City Digitally: Processing Urban Space and Time* (pp. 9–33). Amsterdam University Press.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*, 211–252. https://doi.org/10.1007/s11263-015-0816-y

Saarikivi, J., & Toivanen, R. (2015). Change and Maintenance of Plurilingualism in the Russian Federation and the European Union. In H. F. Marten, M. Rießler, J. Saarikivi, & R. Toivanen (Eds.), *Cultural and Linguistic Minorities in the Russian Federation and the European Union: Comparative Studies on Equality and Diversity* (pp. 3–29). Springer International Publishing. https://doi.org/10.1007/978-3-319-10455-3_1

Santos Vieira, F. A., Vinhas Santos, D. T., Bragagnolo, C., Campos-Silva, J. V., Henriques Correia, R. A., Jepson, P., Mendes Malhado, A. C., & Ladle, R. J. (2021). Social media data reveals multiple cultural services along the 8.500 kilometers of Brazilian coastline. *Ocean & Coastal Management*, *214*. https://doi.org/10.1016/j.ocecoaman.2021.105918

Sauer, C. (1925). The morphology of landscape. In J. Leighly (Ed.), *Land and life* (pp. 315–350). University of California Press.

Saukkonen, P. (2016). Monikulttuurisuuden tilastointi kaipaa uudistamista.

Saukkonen, P. (2018). Multiculturalism and nationalism in Finland. *Finnish Journal of Social Research*, *11*, 65–73. https://doi.org/10.51815/fjsr.110781

Saukkonen, P. (2021a). Cultural policy and cultural diversity. In A. Koivunen, J. Ojala, & J. Holmén (Eds.), *The Nordic Economic, Social and Political Model* (pp. 177–194). Routledge.

Saukkonen, P. (2021b). *Ulkomaalaistaustaiset pääkaupunkiseudulla: asuminen, työllisyys ja tulot* (tech. rep. 2021(1)). Helsingin kaupunki. Helsinki.

Sayyar, S. S., & Marcus, L. (2011). Urban diversity and how to measure it – an operational definition of classes and scales. *Proceedings of the 18th International Seminar on Urban Form*, 1–15.

Schein, R. H. (1997). The Place of Landscape: A Conceptual Framework for Interpreting an American Scene. *Annals of the Association of American Geographers*, *87*(4), 660–680. https://doi.org/10.1111/1467-8306.00072

Schroedler, T., Chik, A., & Benson, P. (2023). The value of multilingualism for sustainable development: A case study of languages in Australia. *International Multilingual Research Journal*, 1–15. https://doi.org/10.1080/19313152.2023.2208509

Scikit-bio. (2020). Scikit-bio: A Bioinformatics Library for Data Scientists, Students, and Developers.

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Scientific Computing Conference* (pp. 92–96). https://doi.org/10.25080/MAJORA-92BF1922-011

Segrott, J. (2001). Language, geography and identity: The case of the Welsh in London. *Social & Cultural Geography*, *2*(3), 281–296. https://doi.org/10.1080/14649360120073860

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., & Sculley, D. (2017). No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World. https://doi.org/10.48550/arXiv.1711.08536

Sheller, M., & Urry, J. (2006). The New Mobilities Paradigm. *Environment and Planning A: Economy and Space*, *38*(2), 207–226. https://doi.org/10.1068/a37268

Shelton, T. (2017). The urban geographical imagination in the age of Big Data. *Big Data & Society*, *4*(1). https://doi.org/10.1177/2053951716665129

Shelton, T., Poorthuis, A., & Zook, M. (2015). Social media and the city: Rethinking urban socio-spatial inequality using user-generated geographic information. *Landscape and Urban Planning*, *142*, 198–211. https://doi.org/10.1016/j.landurbplan.2015.02.020

Shen, Y., & Karimi, K. (2016). Urban function connectivity: Characterisation of functional urban streets with social media check-in data. *Cities*, *55*, 9–21. https://doi.org/10.1016/j.cities.2016.03.013

Sherwin, W. B., & Prat i Fornells, N. (2019). The Introduction of Entropy and Information Methods to Ecology by Ramon Margalef. *Entropy*, *21*(8), 794. https://doi.org/10.3390/e21080794

Shi, W., Goodchild, M., Batty, M., Li, Q., Liu, X., & Zhang, A. (2022). Prospective for urban informatics. *Urban Informatics*, *1*. https://doi.org/10.1007/s44212-022-00006-0

Silm, S., Järv, O., & Masso, A. (2020). Tracing human mobilities through mobile phones. In M. Büscher, N. Grauslund Kristersen, M. Freudendal-Pedersen, & S. Kesselring (Eds.), *Handbook of Research Methods and Applications for Mobilities* (pp. 182–192). Edward Elgar Publishing. https://doi.org/10.4337/9781788115469.00025

Silm, S., Jauhiainen, J. S., Raun, J., & Tiru, M. (2021). Temporary population mobilities between Estonia and Finland based on mobile phone data and the emergence of a cross-border region. *European Planning Studies*, *29*(4), 699–719. https://doi.org/10.1080/09654313.2020.1774514

Silm, S., Mooses, V., Puura, A., Masso, A., Tominga, A., & Saluveer, E. (2021). The Relationship between Ethno-Linguistic Composition of Social Networks and Activity Space: A Study Using Mobile Phone Data. *Social Inclusion*, *9*(2), 192–207. https://doi.org/10.17645/si.v9i2.3839

Silva, T. H., Melo, P. O. S. V. D., Almeida, J. M., & Loureiro, A. A. F. (2014). Large-scale study of city dynamics and urban social behavior using participatory sensing. *IEEE Wireless Communications*, *21*(1), 42–51. https://doi.org/10.1109/MWC.2014.6757896

Simon, G. L., & Graybill, J. K. (2010). Geography in interdisciplinarity: Towards a third conversation. *Geoforum*, *41*(3), 356–363. https://doi.org/10.1016/j.geoforum.2009.11.012

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. https://doi.org/10.48550/arXiv.1409.1556

Singleton, A., & Arribas-Bel, D. (2021). Geographic Data Science. *Geographical Analysis*, *53*(1), 61–75. https://doi.org/10.1111/gean.12194

Singleton, A., Spielman, S., & Folch, D. (2018). *Urban Analytics*. SAGE Publications Ltd.

Sjöholm, K. (2004). Swedish, Finnish, English? Finland's Swedes in a changing world. *Journal of Curriculum Studies*, *36*(6), 637–644. https://doi.org/10.1080/0022027042000186600

Skovgaard Nielsen, R., Holmqvist, E., Dhalmann, H., & Søholt, S. (2015). The Interaction of Local Context and Cultural Background: Somalis' Perceived Possibilities in Nordic Capitals' Housing Markets. *Housing Studies*, *30*(3), 433–452. https://doi.org/10.1080/02673037.2014.973386

Smith, A. C., Merz, L., Borden, J. B., Gulick, C. K., Kshirsagar, A. R., & Bruna, E. M. (2021). Assessing the effect of article processing charges on the geographic diversity of authors using Elsevier's "Mirror Journal" system. *Quantitative Science Studies*, *2*(4), 1123–1143. https://doi.org/10.1162/qss_a_00157

Song, X. P., Richards, D. R., & Tan, P. Y. (2020). Using social media user attributes to understand human–environment interactions at urban parks. *Scientific Reports*, *10*(1), 808. https://doi.org/10.1038/s41598-020-57864-4

Song, Y., Ning, H., Ye, X., Chandana, D., & Wang, S. (2022). Analyze the usage of urban greenways through social media images and computer vision. *Environment and Planning B: Urban Analytics and City Science*, *49*(6), 1682–1696. https://doi.org/10.1177/23998083211064624

Soukup, B. (2020). Survey area selection in Variationist Linguistic Landscape Study (VaLLS): A report from Vienna, Austria. *Linguistic Landscape*, *6*(1), 52–79. https://doi.org/10.1075/ll.00017.sou

Staab, J., Udas, E., Mayer, M., Taubenböck, H., & Job, H. (2021). Comparing established visitor monitoring approaches with triggered trail camera images and machine learning based computer vision. *Journal of Outdoor Recreation and Tourism*, *35*. https://doi.org/10.1016/j.jort.2021.100387

Steed, R., & Caliskan, A. (2021). Image Representations Learned With Unsupervised Pre-Training Contain Human-like Biases. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 701–713. https://doi.org/10.1145/3442188.3445932

Syrett, S., & Sepulveda, L. (2012). Urban governance and economic development in the diverse city. *European Urban and Regional Studies*, *19*(3), 238–253. https://doi.org/10.1177/0969776411430287

Tabouret-Keller, A. (2017). Language and Identity. In F. Coulmas (Ed.), *The Handbook of Sociolinguistics* (2nd ed., pp. 315–326). John Wiley & Sons, Ltd.

Takhteyev, Y., Gruzd, A., & Wellman, B. (2012). Geography of Twitter networks. *Social Networks*, *34*(1), 73–81. https://doi.org/10.1016/j.socnet.2011.05.006

Taşkın, Z., Krawczyk, F., & Kulczycki, E. (2023). Are papers published in predatory journals worthless? A geopolitical dimension revealed by content-based analysis of citations. *Quantitative Science Studies*, *4*(1), 44–67. https://doi.org/10.1162/qss_a_00242

Tasse, D., & Hong, J. I. (2017). Using User-Generated Content to Understand Cities. In P. Thakuriah, N. Tilahun, & M. Zellner (Eds.), *Seeing Cities Through Big Data: Research, Methods and Applications in Urban Informatics* (1st ed., pp. 49–64). Springer. https://doi.org/10.1007/978-3-319-40902-3_3

Tasse, D., Liu, Z., Sciuto, A., & Hong, J. I. (2017). State of the Geotags: Motivations and Recent Changes. *The Proceedings of the Eleventh International AAAI Conference on Web and Social Media*, *11*, 250–259. https://doi.org/10.1609/icwsm.v11i1.14872

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), 17615. https://doi.org/10.1038/s41598-017-18007-4

Tervonen, M. (2014). Historiankirjoitus ja myytti yhden kulttuurin Suomesta. In P. Markkola, H. Snellman, & A.-C. Östman (Eds.), *Kotiseutu ja kansakunta: miten suomalaista historiaa on rakennettu* (pp. 137–162). Suomalaisen Kirjallisuuden Seura.

Toivonen, T., Heikinheimo, V., Fink, C., Hausmann, A., Hiippala, T., Järv, O., Tenkanen, H., & Di Minin, E. (2019). Social media data for conservation science: A methodological overview. *Biological Conservation*, *233*, 298–315. https://doi.org/10.1016/j.biocon.2019.01.023

Torkko, J., Poom, A., Willberg, E., & Toivonen, T. (2023). How to best map greenery from a human perspective?: Comparing computational measurements with human perception. *Frontiers in Sustainable Cities*, *5*. https://doi.org/10.3389/frsc.2023.1160995

Tornes, A., & Trujillo, L. (2021). Enabling the future of academic research with the Twitter API.

Torpan, K., Sinitsyna, A., Kährik, A., Kauppinen, T. M., & Tammaru, T. (2022). Overlap of migrants' housing and neighbourhood mobility. *Housing Studies*, *37*(8), 1396–1421. https://doi.org/10.1080/02673037.2020.1849574

Townsend, A. M. (2013). *Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia* (1st ed.). W. W. Norton & Company.

Trudgill, P. (1974). Linguistic Change and Diffusion: Description and Explanation in Sociolinguistic Dialect Geography. *Language in Society*, *3*(2), 215–246. https://doi.org/10.1017/S0047404500004358

Tu, W., Cao, J., Yue, Y., Shaw, S.-L., Zhou, M., Wang, Z., Chang, X., Xu, Y., & Li, Q. (2017). Coupling mobile phone and social media data: A new approach to understanding urban functions and diurnal patterns. *International Journal of Geographical Information Science*, *31*(12), 2331–2358. https://doi.org/10.1080/13658816.2017.1356464

UNESCO. (2021). *UNESCO Recommendation on Open Science* (tech. rep.). UNESCO Digital Library.

United Nations. (2022). *World Cities Report 2022: Envisaging the Future of Cities* (tech. rep.). United Nations Human Settlements Programme. Nairobi, Kenya.

Vaattovaara, M., & Bernelius, V. (2010). Alueellinen eriytyminen Helsingin metropolialueella. In M. Rimpelä & V. Bernelius (Eds.), *Peruskoulujen oppimistulokset ja oppilaiden hyvinvointi eriytyvällä Helsingin seudulla, MetrOP-tutkimus 2010–2013: Mitä tiedettiin tutkimuksen käynnistyessä keväällä 2010?* (pp. 13–19). Yliopistopaino.

Vaattovaara, M., & Joutsiniemi, A. (2018). Kohti monimuotoistuvan kaupungin ymmärrystä ja kehityksen ohjaamista. *Tieteessä tapahtuu*, *36*(6), 19–23.

Väisänen, T., Heikinheimo, V., Hiippala, T., & Toivonen, T. (2021). Exploring human–nature interactions in national parks with social media photographs and computer vision. *Conservation Biology*, *35*(2), 424–436. https://doi.org/10.1111/cobi.13704

Väisänen, T., Hiippala, T., Järv, O., & Tuuli Toivonen. (2021). Tweetsearcher. https://doi.org/10.5281/ZENODO.4723336

Väisänen, T., Järv, O., Toivonen, T., & Hiippala, T. (2022). Mapping urban linguistic diversity with social media and population register data. *Computers, Environment and Urban Systems*, *97*. https://doi.org/10.1016/j.compenvurbsys.2022.101857

Väisänen, T., Järv, O., Toivonen, T., & Hiippala, T. (2023). Capturing urban diversity through languages: Long-term changes in multilingual residential neighbourhoods in the Helsinki Metropolitan Area. *Population, Space and Place*. https://doi.org/10.1002/psp.2717

Valentine, G., Sporton, D., & Bang Nielsen, K. (2008). Language use on the move: Sites of encounter, identities and belonging. *Transactions of the Institute of British Geographers*, *33*(3), 376–387. https://doi.org/10.1111/j.1475-5661.2008.00308.x

van Zanten, B. T., Van Berkel, D. B., Meentemeyer, R. K., Smith, J. W., Tieskens, K. F., & Verburg, P. H. (2016). Continental-scale quantification of landscape values using social media data. *Proceedings of the National Academy of Sciences*, *113*(46), 12974–12979. https://doi.org/10.1073/pnas.1614158113

Vanky, A., & Le, R. (2023). Urban-semantic computer vision: A framework for contextual understanding of people in urban spaces. *AI & Society*, *38*(3), 1193–1207. https://doi.org/10.1007/s00146-022-01625-6

Vasquez-Henriquez, P., Graells-Garrido, E., & Caro, D. (2020). Tweets on the go: Gender differences in transport perception and its discussion on social media. *Sustainability (Switzerland)*, *12*(13). https://doi.org/10.3390/su12135405

Vertovec, S. (2007). Super-diversity and its implications. *Ethnic and Racial Studies*, *30*(6), 1024–1054. https://doi.org/10.1080/01419870701599465

Vertovec, S. (2019). Talking around super-diversity. *Ethnic and Racial Studies*, *42*(1), 125–139. https://doi.org/10.1080/01419870.2017.1406128

Vertovec, S., Hiebert, D., Spoonley, P., & Gamlen, A. (2022). Visualizing superdiversity and "seeing" urban socio-economic complexity. *Urban Geography*, 1–22. https://doi.org/10.1080/02723638.2022.2151753

Vilkama, K. (2010). Kaupungin laidalla : kantaväestön ja maahanmuuttajataustaisten alueellinen eriytyminen Helsingissä. *Terra*, *122*(4), 183–200.

Vilkama, K. (2011). Yhteinen kaupunki, eriytyvät kaupunginosat? Kantaväestön ja maahanmuuttajataustaisten asukkaiden alueellinen eriytyminen ja muuttoliike pääkaupunkiseudulla. *Tutkimuksia*, *2*, 283.

Vilkama, K., Vaattovaara, M., & Dhalmann, H. (2013). Kantaväestön pakoa? Miksi Maahanmuuttajakeskittymistä muutetaan pois? *Yhteiskuntapolitiikka*, *78*, 485–497.

Wang, X., Spotti, M., Juffermans, K., Cornips, L., Kroon, S., & Blommaert, J. (2014). Globalization in the margins: Toward a re-evalution of language and mobility. *Applied Linguistics Review*, *5*(1), 23–44. https://doi.org/10.1515/applirev-2014-0002

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60). https://doi.org/10.21105/joss.03021

Weerkamp, W., Carter, S., & Tsagkias, M. (2011). How people use twitter in different languages. *Proceedings of the ACM Web Science 2011*.

Wessendorf, S. (2014). 'Being open, but sometimes closed'. Conviviality in a super-diverse London neighbourhood. *European Journal of Cultural Studies*, *17*(4), 392–405. https://doi.org/10.1177/1367549413510415

Wessendorf, S. (2018). 'All the people speak bad English': Communicating across differences in a super-diverse context. In A. Creese & A. Blackledge (Eds.), *The Routledge Handbook of Language and Superdiversity* (1st ed., pp. 57–70). Routledge.

Wessendorf, S., & Farrer, J. (2021). Commonplace and out-of-place diversities in London and Tokyo: Migrant-run eateries as intercultural third places. *Comparative Migration Studies*, *9*(1), 1–17. https://doi.org/10.1186/s40878-021-00235-3

Wieling, M., & Nerbonne, J. (2015). Advances in Dialectometry. *Annual Review of Linguistics*, *1*, 243–264. https://doi.org/10.1146/annurev-linguist-030514-124930

Wieling, M., Nerbonne, J., & Baayen, R. H. (2011). Quantitative Social Dialectology: Explaining Linguistic Variation Geographically and Socially. *PLoS ONE*, *6*(9), e23613. https://doi.org/10.1371/journal.pone.0023613

Wilkins, E. J., Van Berkel, D., Zhang, H., Dorning, M. A., Beck, S. M., & Smith, J. W. (2022). Promises and pitfalls of using computer vision to make inferences about landscape preferences: Evidence from an urban-proximate park system. *Landscape and Urban Planning*, *219*. https://doi.org/10.1016/j.landurbplan.2021.104315

Willberg, E., Fink, C., & Toivonen, T. (2023). The 15-minute city for all? – Measuring individual and temporal variations in walking accessibility. *Journal of Transport Geography*, *106*(1032521). https://doi.org/10.1016/j.jtrangeo.2022.103521

Willberg, E., Poom, A., Helle, J., & Toivonen, T. (2023). Cyclists' exposure to air pollution, noise, and greenery: A population-level spatial analysis approach. *International Journal of Health Geographics*, *22*(1). https://doi.org/10.1186/s12942-023-00326-7

Winder, S. G., Lee, H., Seo, B., Lia, E. H., & Wood, S. A. (2022). An open-source image classifier for characterizing recreational activities across landscapes. *People and Nature*, *4*(5), 1249–1262. https://doi.org/10.1002/pan3.10382

Wylie, J. (2009). Landscape, Absence and the Geographies of Love. *Transactions of the Institute of British Geographers*, *34*(3), 275–289.

Xie, Q., Zhang, X., Ding, Y., & Song, M. (2020). Monolingual and multilingual topic analysis using LDA and BERT embeddings. *Journal of Informetrics*, *14*(3). https://doi.org/10.1016/j.joi.2020.101055

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated Residual Transformations for Deep Neural Networks. *The Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5987–5995. https://doi.org/10.1109/CVPR.2017.634

Yan, Y., Chen, J., & Wang, Z. (2020). Mining public sentiments and perspectives from geotagged social media data for appraising the post-earthquake recovery of tourism destinations. *Applied Geography*, *123*. https://doi.org/10.1016/j.apgeog.2020.102306

Ye, J. (2016). The ambivalence of familiarity: Understanding breathable diversity through fleeting encounters in Singapore's Jurong West. *Area*, *48*(1), 77–83. https://doi.org/10.1111/area.12237

Ye, J. (2017). Contours of urban diversity and coexistence. *Geography Compass*, *11*(9), 1–8. https://doi.org/10.1111/gec3.12327

Ye, J. (2019). Re-orienting geographies of urban diversity and coexistence: Analyzing inclusion and difference in public space. *Progress in Human Geography*, *43*(3), 478–495. https://doi.org/10.1177/0309132518768405

You, Q., Cao, L., Jin, H., & Luo, J. (2016). Robust Visual-Textual Sentiment Analysis: When Attention meets Tree-structured Recursive Neural Networks. *Proceedings of the 24th ACM International Conference on Multimedia*, 1008–1017. https://doi.org/10.1145/2964284.2964288

Zenker, O. (2018). Language and Identity. In H. Callan (Ed.), *The International Encyclopedia of Anthropology*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118924396.wbiea2271

Zhao, B., Zhang, A., Watson, B., Kearney, G., & Dale, I. (2023). A Review of Vision-Language Models and their Performance on the Hateful Memes Challenge. https://doi.org/10.48550/arXiv.2305.06159

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2018). Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1452–1464. https://doi.org/10.1109/TPAMI.2017.2723009

Zook, M., Barocas, S., Boyd, D., Crawford, K., Keller, E., Gangadharan, S. P., Goodman, A., Hollander, R., Koenig, B. A., Metcalf, J., Narayanan, A., Nelson, A., & Pasquale, F. (2017). Ten simple rules for responsible big data research. *PLOS Computational Biology*, *13*(3). https://doi.org/10.1371/journal.pcbi.1005399

Zubiaga, A., Vicente, I. S., Gamallo, P., Pichel, J. R., Alegria, I., Aranberri, N., Ezeiza, A., & Fresno, V. (2016). TweetLID: A benchmark for tweet language identification. *Language Resources and Evaluation*, *50*(4), 729–766. https://doi.org/10.1007/s10579-015-9317-4

Zukin, S., Lindeman, S., & Hurson, L. (2017). The omnivore's neighborhood? Online restaurant reviews, race, and gentrification. *Journal of Consumer Culture*, *17*(3), 459–479. https://doi.org/10.1177/1469540515611203

Zwiers, M., van Ham, M., & Manley, D. (2018). Trajectories of ethnic neighbourhood change: Spatial patterns of increasing ethnic diversity. *Population, Space and Place*, *24*(2), 1–11. https://doi.org/10.1002/psp.2094