Mapping urban linguistic diversity with social media and population register data

Väisänen, Tuomas Lauri Aleksanteri

2022-10

Väisänen , T L A , Järv , O , Toivonen , T & Hiippala , T 2022 , ' Mapping urban linguistic diversity with social media and population register data ' , Computers, Environment and Urban Systems , vol. 97 , 101857 . https://doi.org/10.1016/j.compenvurbsys.2022.101857

http://hdl.handle.net/10138/346612
https://doi.org/10.1016/j.compenvurbsys.2022.101857

Computers, Environment and Urban Systems 97 (2022) 101857

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Mapping urban linguistic diversity with social media and population register data

Tuomas Väisänen[a,c,*], Olle Järv[a,c], Tuuli Toivonen[a,c], Tuomo Hiippala[a,b,c]

[a] *Digital Geography Lab, Department of Geosciences and Geography, University of Helsinki, Finland*
[b] *Department of Languages, University of Helsinki, Finland*
[c] *Helsinki Institutes of Sustainability Science & Urban and Regional Studies, University of Helsinki, Finland*

ARTICLE INFO

Keywords:
Urban diversity
Linguistic diversity
Spatial analysis
Segregation
Social media data
Register data

ABSTRACT

Globalization, urbanization and international mobility have led to increasingly diverse urban populations. Compared to traditional traits for measuring urban diversity, such as ethnicity and country of origin, the role of language remains underexplored in understanding diversity, interactions between different groups and socio-spatial segregation. In this article, we analyse language use in the Helsinki Metropolitan Area by combining individual-level register data, socio-economic grid database, mobile phone and social media data to understand spatio-temporal patterns of linguistic diversity better. We measured linguistic diversity using metrics developed in the fields of ecology and information theory, and performed spatial clustering and regression analyses to explore the spatio-temporal patterns of linguistic diversity. We found spatial and temporal differences between register and social media data, show that linguistic diversity is influenced by the physical and socio-economic environment, and identified areas where different linguistic groups are likely to interact. Our results provide insights for urban planning and understanding urban diversity through linguistic information. As global urbanization, international migration and refugee flows and climate change drive diverse populations into cities, understanding urban diversity and its implications for urban planning and sustainability become increasingly important.

1. Introduction

Urban populations are becoming increasingly diverse due to urbanization, globalization and human mobility. In addition to their original inhabitants, cities are prominent destinations for immigrants who seek work and education, refugees who seek shelter and a better life, and tourists who wish to experience foreign places and cultures. As a consequence, urban populations are becoming super-diverse, and this diversity extends beyond traits such as ethnicity and country of origin, which have traditionally been used for understanding urban diversity (Vertovec, 2007). Recently, information about language use has been proposed as a useful trait for understanding urban diversity, interactions between different groups (Chriost & Thomas, 2008; Peukert, 2013) and socio-spatial segregation (Järv, Masso, Silm, & Ahas, 2020).

In this article, we pursued this idea further by examining urban diversity through language use. To do so, we combined individual-level register data, socio-economic grid database, social media and mobile phone data to study linguistic diversity in the Helsinki Metropolitan Area, Finland. We used automatic language identification to detect languages used on social media platforms and quantified this information using measures developed in the fields of ecology and information sciences. We analysed the spatial distribution of linguistic diversity using methods from spatial statistics and explored the relationship between social media and register data using spatial regression. Our analyses revealed spatial and temporal differences in linguistic diversity between register and social media data, and factors that affect linguistic diversity. Our results emphasize the value of linguistic information for understanding urban diversity in the field of geoinformatics and beyond.

2. Related work

Urban populations are becoming super-diverse in terms of language, ethnicity, religion, gender, age, country of origin, mobility and access to the labour market and housing. Language can be a particularly useful

* Corresponding author at: Digital Geography Lab, Department of Geosciences and Geography, Faculty of Sciences, University of Helsinki, PO Box 64, Gustaf Hällströminkatu 2, FI-00014, Finland.
E-mail address: tuomas.vaisanen@helsinki.fi (T. Väisänen).

https://doi.org/10.1016/j.compenvurbsys.2022.101857
Received 29 April 2022; Received in revised form 23 June 2022; Accepted 14 July 2022
Available online 1 August 2022

variable for describing super-diverse populations, because it can reveal “interesting local configurations” among subpopulations in urban areas (Vertovec, 2007, p. 1033). Language also plays a crucial role in the formation of identities among individuals and groups (Chriost & Thomas, 2008; de Vries, 1990; Valentine, Sporton, & Bang Nielsen, 2008), and can thus complement ethnicity, country of origin and other traditional measures for understanding urban diversity.

To exemplify, a single country of origin may be home to speakers of multiple languages, and thus the first language can be a more important marker of identity than the country of origin (de Vries, 1990; Vertovec, 2007). However, although the first language can be a strong indicator of cultural identity, such identities are continuously shaped by language use (Pennycook & Otsuji, 2015). In multinational and immigrant families, the first language(s) may only be used at home, whereas other languages may be used as the main language(s) of everyday life outside home (de Vries, 1990). Language can also be used to include or exclude individuals from social interactions, to avoid standing out in public spaces, and to signal membership in particular subpopulations (Cadier & Mar-Molinero, 2014; Chriost & Thomas, 2008; El Ayadi, 2021).

Previously, the relationship between linguistic and urban diversity has been mainly explored from qualitative perspectives in the fields of linguistics, and cultural geography (Vertovec, 2019). Previous research has examined spatio-temporal characteristics of individual and group identities and their relation to language Chríost & Aitchison, 1998; El Ayadi, 2021; Segrott, 2001), and how the language of toponyms reflects spatial patterns of power relationships (Kearns & Berg, 2002; Wanjiru & Matsubara, 2017). It has also been suggested that encountering diverse languages in everyday life can foster a sense of belonging and social cohesion, as speakers of different languages work together, live in the same neighbourhoods and form relationships (Cadier & Mar-Molinero, 2014; Hoekstra & Pinkster, 2019).

Linguistic diversity can also be a source of conflict, as the choice of language can be an inclusionary or exclusionary act in different social and spatial contexts (El Ayadi, 2021; Hoekstra & Pinkster, 2019; Valentine et al., 2008). Whether encounters with urban linguistic diversity foster a sense of community and social cohesion depends on the place in which they occur. Public spaces may encourage groups to avoid conflict with each other by being civil, whereas more personal interactions are likely to take place in communal and private spaces (Wessendorf, 2014). However, these places can also become territorialized and exclusionary despite contrary intentions (Hoekstra & Pinkster, 2019).

Exposure to linguistic diversity is increasing, because current social networks extend from physical locations – as exemplified by home, school, and work – to virtual environments, such as social media platforms (Cadier & Mar-Molinero, 2014; Hoekstra & Pinkster, 2019). These platforms are inherently multilingual (Graham, Hale, & Gaffney, 2014; Hong, Convertino, & Chi, 2011; Mocanu et al., 2013), and the choice of language is largely determined by the participants and the context of communication (Androutsopoulos, 2014; Artamonova & Androutsopoulos, 2019; Weerkamp, Carter, & Tsagkias, 2011). Language choice is also a key factor in the formation of social ties on social media (Eleta & Golbeck, 2014; Takhteyev, Gruzd, & Wellman, 2012) and affects the probability of geotagging a post (Huang & Carley, 2019; Magdy, Ghanem, Musleh, & Mokbel, 2014). English is often the lingua franca of social media, but its dominance varies between platforms and countries (Baldwin, Cook, Lui, MacKinlay, & Wang, 2013; Hiippala, Hausmann, Tenkanen, & Toivonen, 2019; Hiippala, Väisänen, Toivonen, & Järv, 2020). The importance of language extends to other digital platforms as well. Quinn (2016), for example, found that in South America, OpenStreetMap contributors who spoke local languages mapped locally important features, such as corner stores, schools and health clinics, whereas contributors from outside South America who spoke English influenced the mapping of poorer and rural areas.

Social media and other digital platforms can provide new sources of data for studying linguistic and urban diversity, as the platforms provide a rich source of data on language use that is combined with geographical information. For example, these data have been used to understand the spatio-temporal dynamics of topics, geodemographics and social activities in urban environments (Adelfio, Serrano-Estrada, Martí-Ciriquián, Kain, & Stenberg, 2020; Lansley & Longley, 2016; Longley, Adnan, & Lansley, 2015). Unfortunately, previous geographical research has largely ignored the multilingualism of social media platforms by excluding or ignoring content in languages other than English (see, for example, Fu, McKenzie, Frias-Martinez, & Stewart, 2018; Karami et al., 2021; Koylu, Larson, Dietrich, & Lee, 2019; Lansley & Longley, 2016). This is not surprising, because English is well-resourced in terms of language technology needed for analysing large volumes of social media content (Del Gratta, Goggi, Pardelli, & Calzolari, 2021). However, limiting analysis to English can lead to biased outcomes if English content is taken as representative of social media as a whole.

Finally, the high volume of social media data requires quantifying information about linguistic diversity. Peukert (2013) introduced the idea of estimating linguistic diversity in urban areas using common measures from ecology and information sciences, such as Shannon entropy and Simpson diversity, which are commonly combined for a balanced estimation of diversity (Morris et al., 2014). Subsequent work has successfully applied these measures to analysing diversity at various spatial scales, ranging from specific locations (Hiippala et al., 2019) to entire cities (Bereitschaft & Cammack, 2015) and countries (Hiippala et al., 2020).

3. Data

3.1. Study area

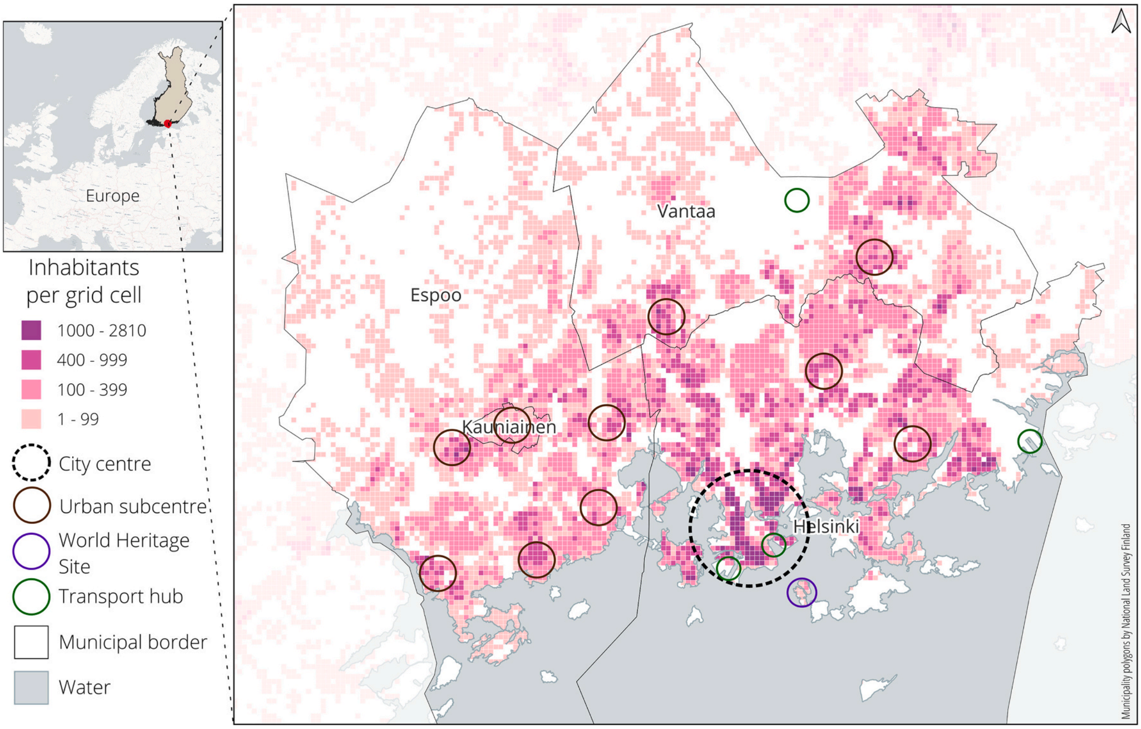
The Helsinki Metropolitan Area (HMA) consists of Helsinki (pop. 659,000), the capital of Finland, and three surrounding municipalities, Espoo (pop. 297,000), Vantaa (pop. 239,000), and Kauniainen (pop. 10,000). In addition to the official languages, Finnish and Swedish, the most common languages are Russian, Estonian, and Somali (Table 2). Finland has a high percentage of internet (89%), smartphone (80%), and social media users (61% overall; around 90% for ages 16–34; Kohvakka, Melkas, & Tarkoma, 2018). Fig. 1 shows a population density map for HMA with major urban centres, transport hubs and a UNESCO World Heritage Site.

3.2. Register and mobile phone data

The Finnish Population Information System is a register that contains roughly 140 variables for every resident of Finland. The variables in the registry, which are updated annually, cover diverse socio-economic, cultural, marital and geographical information, including the self-reported first language of each individual as an ISO-639-2 language code. However, whether the individual is bi- or multilingual is not recorded in the register (Latomaa, 2012). The spatial format of the register is based on a 250 by 250-m square grid, which is a national standard.

Each grid cell has a unique identifier, which is used to connect the home location of individuals in the register to a grid cell. In other words, the register aggregates information about the home location of each individual into 250-m cells to preserve privacy. As such, the register provides a spatially accurate but temporally coarse view of individual language users that reside in the Helsinki Metropolitan Area. Because the data are updated yearly, the register can be argued to provide a static view of languages spoken in the study area and reflective of linguistic diversity only from the perspective of first languages, not everyday language use, and in private spaces. We used register data from the year 2015 to ensure comparability with social media data (see Section 3.3).

We complemented the individual-level register data with a database of socio-economic variables from Statistics Finland to study whether these variables affect linguistic diversity. The spatial format of this database follows the same national standard as the register. The



Municipality polygons by National Land Survey Finland

Fig. 1. A population density map of the Helsinki Metropolitan Area with key locations.

database contains variables that describe the population, built environment and job structure in each grid cell, such as education level, annual income and employment status, proportion of non-residential buildings and rental dwellings. Finally, we complement the grid-level statistics with mobile phone data that describe the dynamic population in the study area (Bergroth, Järv, Tenkanen, Manninen, & Toivonen, 2022). The data, which provides the percentage of population present in each 250-m grid cell at hourly intervals, was provided by the largest mobile network operator in Finland.

3.3. Social media data

We collected social media content from two platforms, Instagram and Twitter, which were geotagged to locations within the study area in 2015. We studied data from 2015 for three reasons. First, Instagram closed its Application Programming Interface (API) after the Cambridge Analytica scandal in 2018 (Bruns, 2019). Thus, we only have access to legacy data collected via the Instagram API in 2014–2016 by collected by the Authors (2022, link removed for double-blind review). Second, after 2015 Twitter switched from using geotags with GPS coordinates to points-of-interest defined at the levels of locations, cities and countries (Hu & Wang, 2020; Tasse & Hong, 2017). Finally, our aim is to illustrate how linguistic diversity can be examined with comprehensive social media data, and the use of legacy data does not impede this aim. We collected all tweets from Twitter's full archive via the API for academic research in 2021 using a Python tool developed by Väisänen et al. (2021).

To pre-process the social media data, we first removed non-human users ('bots') using an existing list of Twitter users identified as bots. We then filtered out Twitter and Instagram users with the word 'bot' in their usernames and removed tweets with similar content posted at regular intervals. We also identified and removed cross-platform posts, as exemplified by Instagram posts that were simultaneously shared on Twitter. Finally, we removed posts with city- and country-level geotags from both Twitter and Instagram data, because they aggregate multiple observations under a single geographical point, which distorts the spatial distribution of the data. On Twitter, for example, the geotag for Helsinki is located in the middle of an in-land bay.

We combined the pre-processed Instagram and Twitter data into a single dataset, because social media users can use different social media platforms for different communicative purposes (Manikonda, Meduri, & Kambhampati, 2016), whereas we are concerned with linguistic diversity stemming from everyday language use in general. The social media data reflects everyday mobilities of Twitter and Instagram users present in the HMA, including inhabitants but also commuters and visitors from outside the HMA, in addition to foreign tourists. Thus, social media provides a more dynamic perspective of the spatio-temporalities of everyday language use. The combined dataset contains unique identifiers for social media posts and users, time stamps, spatial coordinates and the linguistic content of the original post. We also used the time stamps to place each observation into five temporal categories that correspond to morning (06:00–10:00), noon (10:00–14:00), afternoon (14:00–18:00), evening (18:00–22:00) and night (22:00–06:00).

To enable comparisons of register and social media data, we aggregated the social media dataset to the 250-m spatial grid used by the register data. This also allowed us to contextualize the language observations from social media data with the official register data. However, the spatial accuracy of social media data may be affected by a bad GPS signal, a mixture of GPS and point-of-interest (POI) geotags in the data, and whether the posts are uploaded to the platform immediately or afterwards (Cvetojevic, Juhasz, & Hochmair, 2016; Huang & Carley, 2019; Poblete, Garcia, Mendoza, & Jaimes, 2011). To mitigate issues related to spatial accuracy, we used queen contiguity to aggregate observations from the social media data for each grid cell: in addition to the original observations, each grid cell also inherits the observations from its immediate neighbouring cells.

4. Methods

4.1. Automatic language identification

We used automatic language detection to detect the languages used in each social media post at sentence level (Jauhiainen, Lui, Zampieri, Baldwin, & Lindén, 2019). We first used the Punkt tokenizer (Kiss & Strunk, 2006) to segment each post into orthographic sentences to capture sentence-level switches between different languages. Following Hiippala et al. (2019, 2020), we removed hashtags and emojis and filtered out sentences with less than eight characters before language identification. We then used a pre-trained language identification model capable of detecting 176 languages to detect the language of each sentence, and discarded sentences whose language was detected with less than 70% confidence. The model was trained using the fastText algorithm (Bojanowski, Grave, Joulin, & Mikolov, 2017). The final dataset contained approximately 555,000 posts, with roughly 800,000 sentences posted across different times of day (Table 1).

4.2. Measures of diversity

To assess linguistic diversity in social media and register data, we used two diversity metrics, Simpson diversity and Shannon entropy, originally developed in the fields of ecology and information science. These metrics are commonly used in assessing species diversity in ecology and generally considered to provide a good overview of diversity (Magurran & Henderson, 2010; Morris et al., 2014). We calculated the linguistic diversity per grid cell based on language observations from social media and register data within the grid cell.

Simpson diversity describes the probability that two randomly chosen samples do not belong to the same group (Morris et al., 2014). This measure is sensitive to abundant observations and is considered to be “one of the most meaningful and robust diversity measures available” (Magurran, 2013, p. 115). The values for Simpson diversity range from 0 to 1. Higher values indicate higher probability that the two samples are not from the same group, which implies higher diversity.

Shannon entropy is a widely-used information-theoretic measure for estimating the amount of information needed to describe the identities of individuals in a system (Magurran & McGill, 2011, p. 56; Morris et al., 2014). This measure is sensitive to both rare and abundant species (Magurran & Henderson, 2010; Morris et al., 2014) and ranges from 0 (only one species present) to higher values as diversity increases. The values for Shannon entropy are often scaled to a range from 0 to 1 to improve their interpretability (Magurran, 2013, pp. 107–108). We scaled Shannon entropy values to this range, in which 0 corresponds to low and 1 corresponds to maximum linguistic diversity.

4.3. Spatio-temporal analyses

We used bivariate local Moran’s *I* with Shannon entropy and Simpson diversity as independent variables to identify clusters of low and high linguistic diversity. Bivariate local Moran’s *I* (Anselin, Syabri, & Smirnov, 2002) is an extension of local Moran’s *I*, which indicates “the extent of significant spatial clustering of similar values” of a single variable around an observation (Anselin, 1995, p. 94). Bivariate Local Moran’s *I* constructs a statistic that assesses the match between two variables in geographical and attribute spaces (Anselin et al., 2002). We performed the analysis with 9999 permutations and a pseudo *p*-value threshold of 0.001, and used *k*-nearest neighbours with a value of 8 for defining spatial neighbours. We applied the same method to the register data as a whole and to social media data at different times of day, to identify statistically significant clusters with low or high linguistic diversity in both datasets.

For social media data, we then analysed if the linguistic diversity of each cluster remained low or high at different times of day. We counted how many times a given grid cluster was present over different times of day, as defined in Section 3.3, and placed each cluster into three categories that describe the cluster’s *temporal stability*. Low stability includes clusters that were present once or twice a day, whereas moderate stability describes clusters that were present thrice. High stability, in turn, covers clusters that were present four or five times a day.

4.4. Regression analyses

We used aspatial ordinary least squares (OLS) and spatial lag model (SLM) regression to analyse the effect of various socio-economic and demographic variables on the linguistic diversity of social media data, as everyday language use is influenced by the socio-spatial context (El Ayadi, 2021; Valentine et al., 2008). We used scaled Shannon entropy as the dependent variable in both regression analyses, as the measure is sensitive to both rare and abundant languages. We initially chose 18 variables (details on the variables are provided in Appendix B), including socio-economic variables from the grid database that have been connected to urban diversity in previous research (e.g. owner-occupancy and income level in Abascal & Baldassarri, 2015; Bereitschaft & Cammack, 2015); the proportion of dynamic population present in the cluster from mobile phone data; the number of social media posts per grid cell to control its influence on linguistic diversity; and the diversity metrics calculated for each cell using the register data. The same variables were used in the SLM regression, except SLM additionally includes the spatial lag of Shannon entropy as an independent variable, with queen contiguity as the parametrization of spatial dependence.

We removed intercorrelated variables using a variance inflation factor (VIF) test with a threshold value of 5.0, which is generally considered to indicate problems with multicollinearity (Dormann et al., 2013; James, Witten, Hastie, & Tibshirani, 2013, pp. 102–103). Next, we identified statistically significant ($p < 0.05$) variables that affect linguistic diversity using backwards stepwise OLS regression. Details on the variables and their stepwise elimination are provided in Appendix B.

5. Results

5.1. Languages in register and social media data

To acquire an overview of the languages recorded in register and social media data, we first extracted information about languages in both datasets. The register data contains 143 unique languages self-reported by inhabitants, whereas the language identification model

Table 1
Sentences, posts and users in Twitter and Instagram data across times of day.

	Sentences		Posts		Users	
Time of day	Instagram	Twitter	Instagram	Twitter	Instagram	Twitter
Morning	52,545	23,771	33,755	16,988	17,129	3461
Noon	133,626	41,170	87,690	30,449	34,442	5523
Afternoon	166,621	43,054	113,659	32,438	43,665	6083
Evening	154,903	50,550	108,744	37,218	40,780	5406
Night	95,479	36,865	66,649	27,786	31,183	3917
Total	603,174	195,410	410,497	144,879	90,277	11,503

detected 99 unique languages in the social media data. Table 2 lists the ten most common languages for both register and social media data. Whereas the register data highlights the presence of the Swedish-speaking minority and largest immigrant groups who speak Russian, Estonian and Somali (Kraus, 2011), the social media data emphasizes the role of English as the lingua franca of social media platforms in Finland and the Nordic countries (Coats, 2019; Hiippala et al., 2020). Furthermore, languages such as French, Japanese, Korean and Spanish reveal the presence of tourists in the social media data (Hiippala et al., 2019).

5.2. *Measuring linguistic diversity*

To explore differences in linguistic diversity between register and social media data, we first calculated Shannon entropy and Simpson for the languages observed in each grid cell. For social media data, we also calculated both diversity indices across different times of the day. We then estimated the distribution of both indices using kernel density estimation (KDE; Fig. 2).

In Fig. 2a, the KDEs for Shannon entropy show that the register data are less diverse than social media data. The KDEs for both datasets follow a similar pattern, with a smaller peak for zero values, which represent monolingual grid cells. Most grid cells, however, are multilingual, and the peaks of the density estimations suggest that social media data are approximately twice as diverse as register data. Given that the register data feature 143 unique languages compared to 99 languages detected in social media data, this indicates that the grid cells feature a more diverse mix of languages when viewed through social media data. Furthermore, in terms of temporal change, the linguistic diversity of social media increases towards afternoon and evening (Table 3).

For Simpson diversity in Fig. 2b, the KDEs follow a pattern similar to Shannon entropy due to the presence of monolingual grid cells, but the difference between the two datasets becomes more pronounced. Because Simpson diversity represents the probability that two random observations drawn from a single grid cell do not belong to the same language, this reinforces the previous view that the social media data are indeed more diverse. For the largest number of cells, which corresponds to the highest peak, there is approximately a 50% chance that two observations represent different languages. The mean values for Simpson diversity also increase towards the evening and peak at night (Table 3).

5.3. *Spatial distribution of linguistic diversity*

To understand the spatial distribution of linguistic diversity, we calculated the diversity indices for each 250-metre grid cell for both register and social media data. Fig. 3 shows the spatial distribution of values for Shannon entropy and Simpson diversity across the study area. The spatial distribution of both indices reveals considerable contrasts between the datasets, which complement the initial findings in Fig. 2 by providing a spatial perspective to linguistic diversity.

Both Shannon entropy and Simpson diversity show a sharp contrast between register and social media data. The register data are dominated by cells with low linguistic diversity, apart from a few urban subcentres that feature cells with moderate to high linguistic diversity. The social media data, in contrast, is characterized by cells with moderate linguistic diversity, whereas cells with the highest linguistic diversity can be found in the Helsinki city centre, transport and tourism destinations. However, these locations do not stand out in the register data, which shows the impact of everyday mobilities and tourism on linguistic diversity on social media (cf. Hiippala et al., 2019). For the Helsinki city centre, the observations for register and social media data mirror each other. In contrast, the urban subcentres, which appear to be as linguistically diverse based on register data, are also diverse according to social media data.

Table 2
The ten most common languages in both register and social media data. The *Sentences* column gives the total number of sentences in the dataset for each language. The *Users* column gives the number of users who wrote at least 50% of their posts in the given language and posted at least 10 times.

Register	Residents	Social media	Sentences	Users
Finnish	879,011	Finnish	417,531	8831
Swedish	63,903	English	233,881	4441
Russian	28,404	Russian	32,375	735
Estonian	23,169	Swedish	11,242	220
Somali	11,735	Japanese	5961	143
English	8616	Korean	2874	71
Arabic	7562	Spanish	2736	45
Chinese	5753	Portuguese	2111	33
Kurdish	4830	German	2160	28
Albanian	4252	Turkish	2453	19

5.4. *Clusters of low and high linguistic diversity*

5.4.1. *Register data*

To identify statistically significant clusters of high and low linguistic diversity in the register data, we used Shannon entropy and Simpson diversity as input variables for a bivariate local Moran's *I* analysis. We used 9999 permutations and a pseudo *p*-value threshold of 0.001 for the clusters.

Fig. 4 reveals that based on the register data, linguistic diversity – as measured by Shannon entropy and Simpson diversity – is not statistically significant in the city centre and most urban subcentres. Statistically significant clusters with high linguistic diversity can mainly be found in densely populated areas, whereas clusters with low linguistic diversity can be mainly found in peripheral areas with low population density (see Fig. 1). One notable area with high linguistic diversity is the city of Kauniainen, which has a sizeable Swedish-speaking population. The outlier cells are particularly interesting, as the high-low clusters are likely to feature rare languages surrounded by abundant languages, as Shannon entropy is sensitive to both. Conversely, the low-high clusters are likely to have fewer rare languages, which are surrounded by a diversity of abundant languages. These cells are mainly found close to clusters of high linguistic diversity, which may indicate sites for potential language contact.

5.4.2. *Social media data*

For social media data, we applied bivariate local Moran's *I* to observations for Shannon entropy and Simpson diversity in each grid cell across different times of day, to understand whether the spatial distribution of linguistic diversity changes over time (Fig. 5). In contrast to the register data, which can be treated as the presence of potential language users, social media data allows diversity to be examined in terms of everyday language use and daily mobilities. In contrast to the register data, Fig. 5 shows that the Helsinki city centre, several transport hubs and the Suomenlinna World Heritage Site remain linguistically diverse throughout the day, whereas linguistic diversity in many urban subcentres is not statistically significant. Interestingly, the social media data are similar to the register data in terms of the distribution of low-diversity cells, which are also located in peripheral areas with low population density. Notably, the size of the high-diversity cluster in the city centre varies across the times of the day. This illustrates how daily mobilities, such as commuting, work, recreation and other activities that involve changes in the spatial concentration of population, affect linguistic diversity.

5.4.3. *Temporality of linguistic diversity*

To understand the temporal variation of linguistic diversity for social media data better, we combined the cells with high linguistic diversity in both social media and register data in Fig. 6. For social media data, we categorized each cell according to its temporal stability, as described in

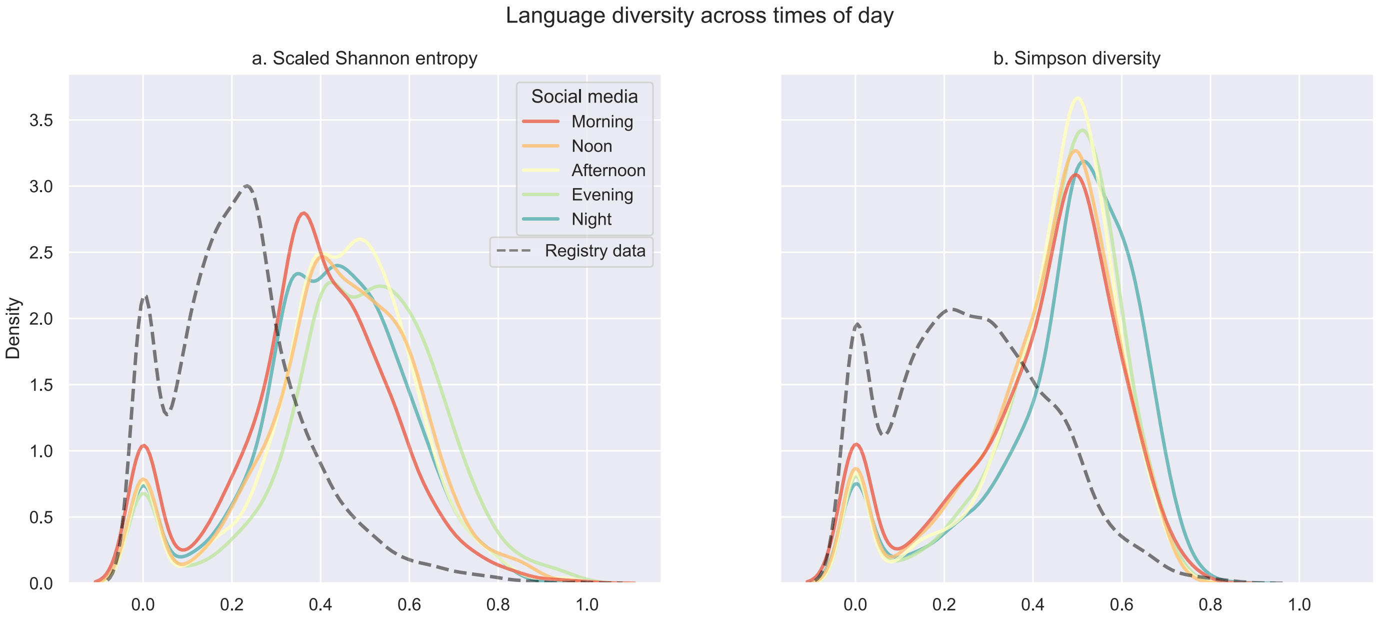


Fig. 2. Kernel density estimations (KDE) for Shannon entropy (a.) and Simpson diversity (b.) over all grid cells in the study area (x-axis). The y-axis gives the kernel density.

Table 3
Mean and standard deviation (in parentheses) for Shannon entropy and Simpson diversity for register and social media data over different times of day.

	Shannon entropy	Simpson diversity
Morning	0.379 (0.184)	0.412 (0.186)
Noon	0.434 (0.186)	0.424 (0.171)
Afternoon	0.437 (0.176)	0.438 (0.164)
Evening	0.480 (0.192)	0.443 (0.165)
Night	0.417 (0.177)	0.469 (0.175)
Register	0.216 (0.154)	0.259 (0.175)

Section 4.3. Fig. 6 shows that most highly diverse clusters are not temporally stable, which reflects everyday mobilities between e.g., home and work. Clusters in residential areas located away from the city centre are less stable temporally, which is likely to be caused by daytime changes in the dynamic population. Notably, linguistically diverse clusters in the register data, which reflect the home locations of potential language users, are positioned close to the cells with low temporal stability. Clusters with high temporal stability, which indicates consistent linguistic diversity throughout the day, can be found in the city centre, transport hubs and touristic landmarks, such as the Suomenlinna World Heritage Site.

5.5. Factors affecting linguistic diversity

In order to understand which factors affect linguistic diversity on social media, we performed ordinary least squares (OLS) and spatial lag model (SLM) regression analyses across different times of day and over 24 h. We used Shannon entropy as a measure of linguistic diversity on social media as the dependent variable, whereas the independent variables were chosen from the socio-economic grid database, as outlined in Section 4.4.

5.5.1. OLS regression

Aspatial OLS regression revealed a weak correlation between the independent variables and linguistic diversity on social media. However, five independent variables are statistically significant and consistently related to linguistic diversity across different times of day: the percentage of (1) dwellings in apartments, (2) the unemployed, (3) students, (4) dynamic population and (5) individuals with a tertiary degree (Table 4). The strongest predictors of linguistic diversity include the percentage of students, dynamic population and individuals with a tertiary degree. Summary tables for OLS regression can be found in Appendix C.

The changes in R^2 values and coefficients over times of day reflect the impact of daily activity patterns on linguistic diversity. The adjusted R^2 values are highest during mornings and nights when most people are home, and lowest with full data, which suggests temporal dynamics affect linguistic diversity. Furthermore, the coefficient for dynamic population grows from noon to evening, which coincides with increasing social media activity (Table 1).

5.5.2. Spatial regression

Spatial regression using a spatial lag model revealed a moderate correlation between the independent variables and linguistic diversity on social media (Table 5). The strongest predictor is spatial lag of Shannon entropy, which indicates that the linguistic diversity of a grid cell is influenced more by the diversity of neighbouring grid cells than other variables that describe the cell itself. Although the coefficients for the percentage of dwellings in apartment buildings and individuals with a tertiary degree are also statistically significant, the coefficients for spatial lag are considerably higher. Furthermore, the values for R^2 increased considerably, indicating the SLM model has a better fit for the data compared to the aspatial OLS model. This suggests that linguistic diversity is a spatial phenomenon. Summary tables for spatial regression can be found in Appendix D.

6. Discussion

6.1. Register and social media data provide different views to linguistic diversity

Our results emphasize the benefits of combining multiple data sources for studying the linguistic diversity of urban environments. Although individual-level register data are rich and spatially accurate, these data do not provide information about daily mobilities and everyday language use (Pennycook & Otsuji, 2015). As such, the register data provide a static view of linguistic diversity. Social media data, in contrast, provides access to everyday language use, which may be linked

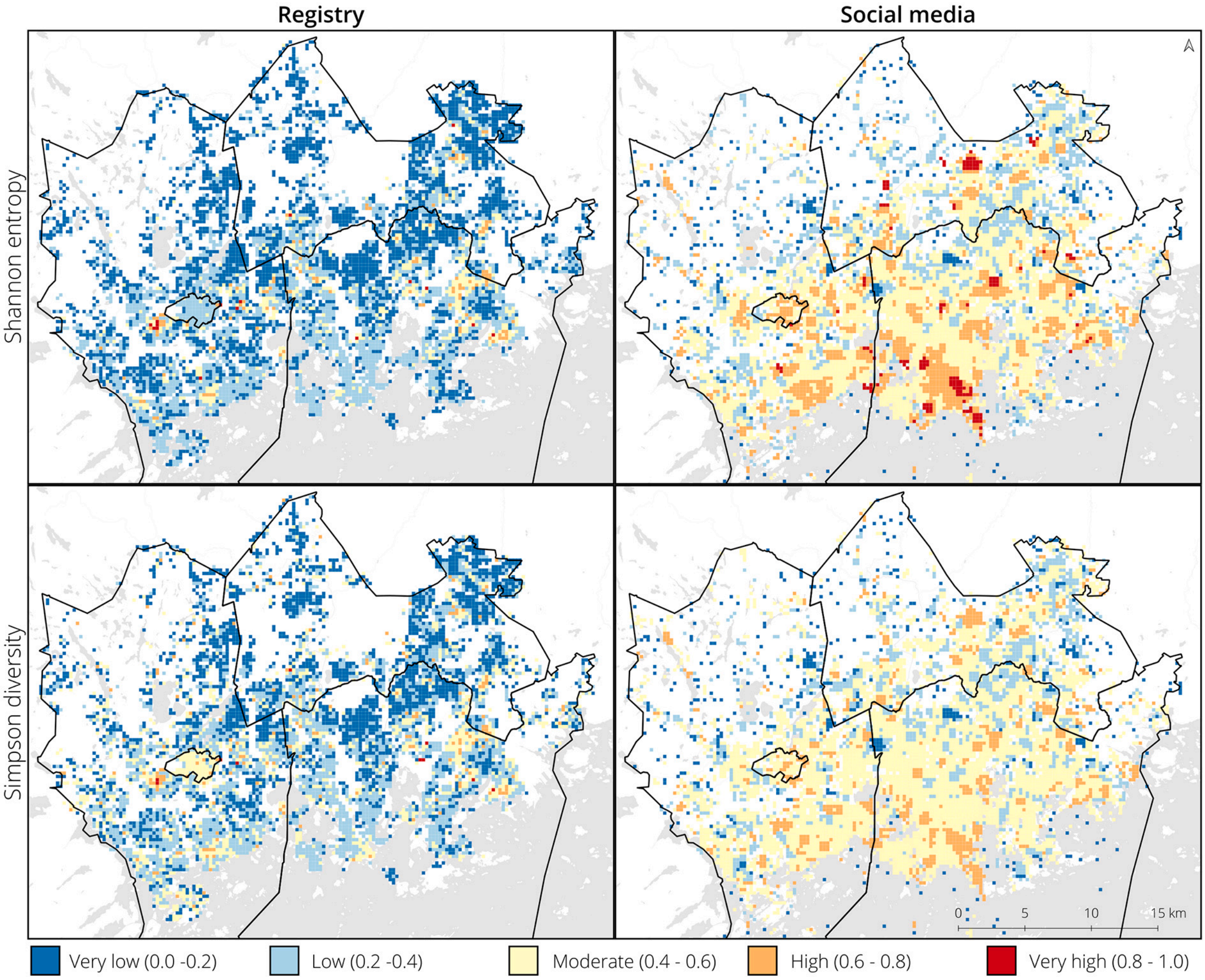


Fig. 3. The spatial distribution of linguistic diversity across the Helsinki Metropolitan Area, as measured using Shannon entropy and Simpson diversity, for register and social media data.

to particular physical locations via geotagging and is not limited to local residents. This information can potentially reveal places where language contact and interactions between different linguistic groups take place. However, language choices are also influenced by the social media platforms and the intended audience (Androutsopoulos, 2015; Hiippala et al., 2019, 2020).

Register and social media data provide fundamentally different views of linguistic diversity and potential for interaction in the Helsinki Metropolitan Area (Järv et al., 2020; Tammaru, Knapp, Silm, van Ham, & Witlox, 2021). When viewed through register data, the Helsinki city centre and the metropolitan area do not appear to be as linguistically diverse, apart from several suburban neighbourhoods with larger populations of immigrants and linguistic minorities. It may be argued that the register-based view of linguistic diversity highlights the presence of *potential* language users, and to some extent, interactions in the private sphere, because the spatial information corresponds to home location. Social media data, in contrast, provide a view of everyday language use as a part of daily mobilities away from home. The Helsinki city centre, main transport hubs, university campuses and tourist areas are all linguistically diverse according to the social media data, which suggests that linguistic diversity emerges as a result of daily mobilities and activities in public and communal spaces.

Contrary to our expectations, some communal spaces located within diverse residential neighbourhoods were not linguistically diverse according to social media data. One such example was a large shopping centre that specializes in ethnic retail, which is located in a linguistically diverse urban subcentre and is known to be socially and culturally important to local immigrant communities (Hewidy & Lilius, 2021). A closer analysis of social media data from this location revealed that the dominant language is Finnish ($n = 171$), which outnumbered observations in six other languages, such as English ($n = 48$), Turkish ($n = 39$), Arabic ($n = 4$), Russian ($n = 2$), Estonian and Thai (both $n = 1$). The lack of observations for the rarer languages may result from the use of social media platforms not covered by this study (see Section 6.5).

The use of social media data follows a diurnal rhythm, remaining low during the night and in the morning and peaking in the evening, which is also reflected in linguistic diversity. However, not all locations follow a similar temporal pattern. We identified several areas in which linguistic diversity remains constantly high, such as the Helsinki city centre and the surrounding neighbourhoods, university campuses, harbours, the airport and tourist destinations. The nature of these locations implies a near-continuous presence of inhabitants, tourists and commuters. However, linguistic diversity, should not be attributed to tourists alone, as previous research has shown that the vast majority of Twitter users from Finland use more than one language on the platform (Hiippala et al., 2020). Notably, temporal variation in linguistic diversity also increases when moving towards suburban centres.

6.2. *Linguistic diversity is a spatio-temporal phenomenon*

Regression analyses using social media data revealed that linguistic diversity is essentially a spatio-temporal phenomenon. The spatial lag

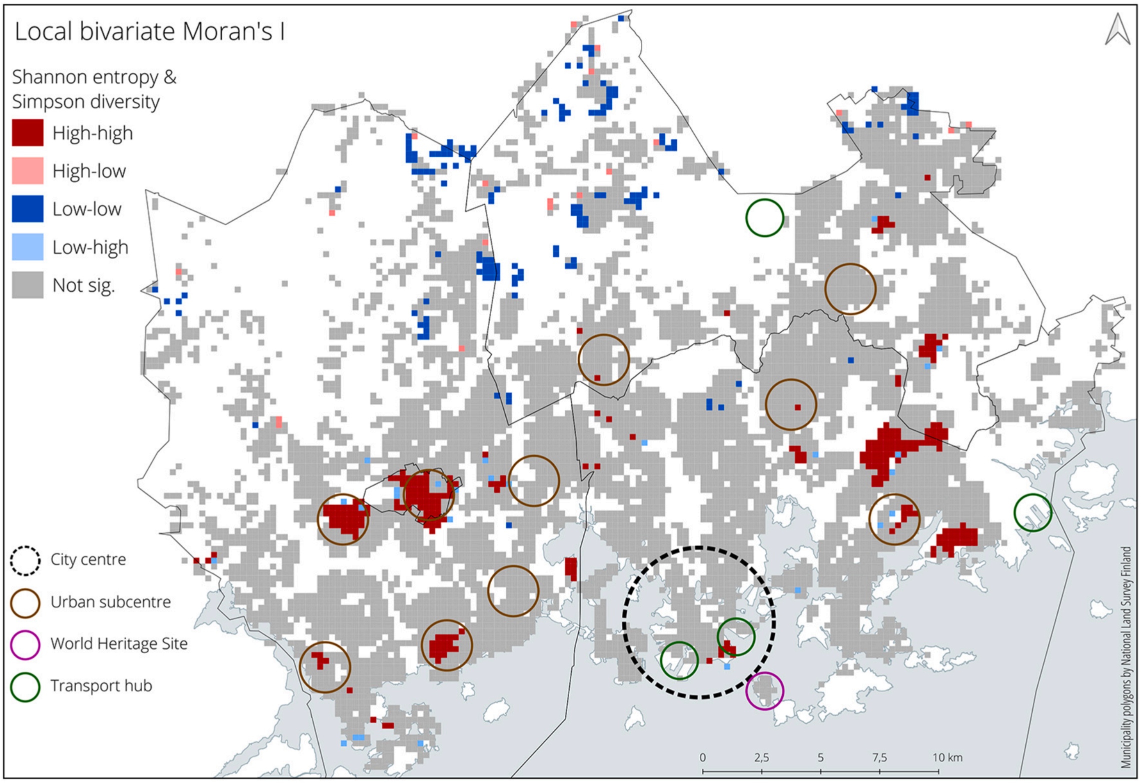

Local bivariate Moran's I
Shannon entropy & Simpson diversity
High-high
High-low
Low-low
Low-high
Not sig.
City centre
Urban subcentre
World Heritage Site
Transport hub
0
2,5
5
7,5
10 km
Municipality polygons by National Land Survey Finland

Fig. 4. Clusters of high and low linguistic diversity in the HMA identified from the register data. The clusters were identified with bivariate local Moran's *I* analysis of Shannon entropy and Simpson diversity. We used 9999 permutations and a pseudo p-value filter of 0.001. High-high clusters represent grid cells with high values for Shannon entropy, whose neighbouring cells have high values for Simpson diversity. Similarly, low-low clusters stand for low values for Shannon entropy and low values for Simpson diversity in the neighbouring cells. High-low and low-high clusters represent outlier cells, where a grid cell with high or low value for Shannon entropy has neighbouring grid cells with opposite values for Simpson diversity.

model revealed that the diversity of neighbouring grid cells is the strongest predictor of linguistic diversity on social media, regardless of the time of day. However, the importance of spatial lag increases during afternoons and evenings together with the overall linguistic diversity (see Table 3), which suggests that linguistic diversity is associated with activities that take place away from home and work. More generally, this finding supports the view that language use is affected by its virtual and physical socio-spatial contexts (Artamonova & Androutsopoulos, 2019; El Ayadi, 2021).

The main socio-economic features that explain linguistic diversity on social media were the proportion of inhabitants with tertiary education and students living in the grid cell. Individuals with tertiary education are likely to speak several languages and have a higher level of income, which enables them to live in central areas with consistently high linguistic diversity. The contribution of students to linguistic diversity, in turn, is not surprising, given that institutions of higher education are both international and multilingual. In addition, students are more likely to use social media, which may suggest that they are overrepresented in the data.

To a lesser extent, linguistic diversity is also explained by the presence of non-residential buildings and dwellings in apartment buildings. Non-residential buildings are related to work, school and leisure, which reinforces the finding that linguistic diversity emerges outside home. Apartment buildings result in higher population density, which also increases the number of potential language users. These findings give a slight indication that the density of the urban forms, which has been recognized as an important factor in fostering socio-economic and cultural activity (Martino, Girling, & Lu, 2021; Oliveira, 2021), might also have a role in the spatio-temporal distribution of linguistic diversity.

Finally, the proportion of jobs in the entertainment, retail and hospitality sectors did not contribute to linguistic diversity, although previous research has connected social media use to leisure (cf. Adelfio et al., 2020). Unlike suggested in previous research, we did not find owner-occupancy and income level to explain linguistic diversity at a statistically significant level (Abascal & Baldassarri, 2015; Bereitschaft & Cammack, 2015). These findings warrant further research, particularly in relation to spatio-temporality of linguistic diversity and its connection to socio-spatial inequalities and segregation.

6.3. Social media and register data can reveal places for potential language contact

Identifying potential sites for language contact is crucial for understanding linguistic diversity in urban environments. Neighbourhoods can contain real or perceived spatio-temporal boundaries, enclosures, or other forms of territorialization that reduce conviviality, for instance, due to segregation (Andersson, Brattbakk, & Vaattovaara, 2017; J. Ye, 2017). Furthermore, intergroup contacts can also be influenced by the degree of integration or segregation of the inhabitants' spatial, social, and cultural networks (Kukk, van Ham, & Tammaru, 2019; Vorobeva, Jauhiainen, & Tammaru, 2021; X. Ye & Andris, 2021), together with their affiliation with a perceived social stratum, which influences their mobility in a city and potential contacts (Järv et al., 2020). Local knowledge is thus needed to contextualize potential areas for language contact identified in the data, to assess whether the interactions are more likely to be positive or negative.

Our analyses allowed identifying linguistically-diverse locations where users of various languages are likely to come in contact with each other. We also found potentially interesting outlier areas in the register data, where cells with low linguistic diversity are surrounded by linguistically diverse cells or vice versa. These areas are potentially relevant places for language contact, as they represent 'estuaries'

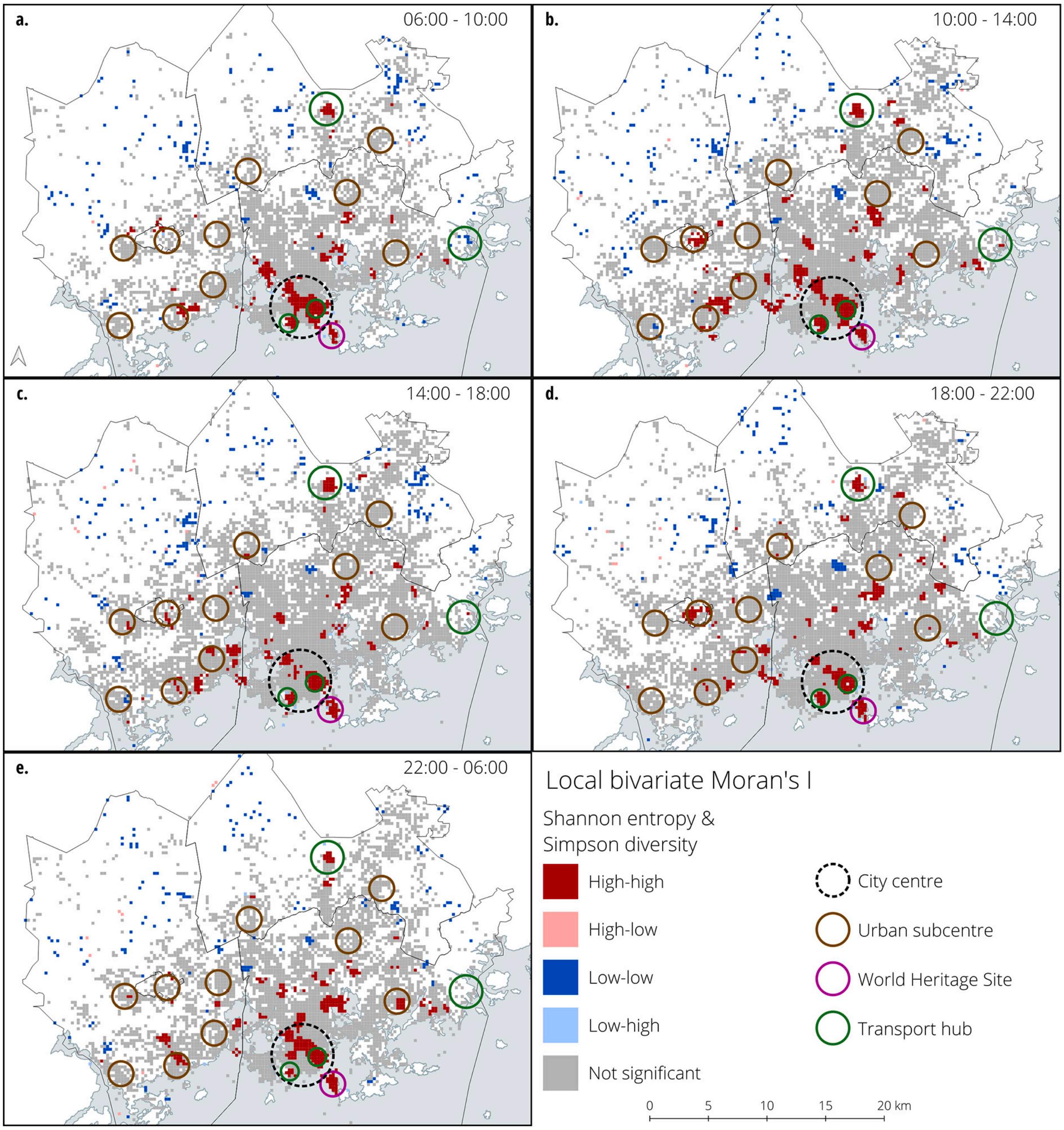


Fig. 5. Temporal variation of the clusters of linguistic diversity identified from social media data. The clusters were identified with bivariate local Moran's *I* analysis of Shannon entropy and Simpson diversity during morning (a.), noon (b.), afternoon (c.), evening (d.), and night (e.). We used 9999 permutations and a pseudo p-value filter of 0.001.

between mono- and multilingual areas, where positive interaction or tension between language groups can take place (Wessendorf, 2014; J. Ye, 2017). Further research is needed to understand the nature of interactions in these areas and their socio-spatial context.

The socio-spatial context in which diversity is encountered determines whether the contact is positive and meaningful (Matejskova & Leitner, 2011; Watson, 2009; Wessendorf, 2014). However, previous research disagrees on whether interactions in public spaces or communal and private spaces are meaningful and positive. Wessendorf (2014), for instance, emphasizes the importance of interactions in private spaces over those in public spaces, as public interactions are often characterized by civility and the desire to avoid conflict. Others consider daily encounters in public and communal spaces to be more important for fostering a sense of community and a sense of belonging (Matejskova & Leitner, 2011; Watson, 2009). Our approach of combining register and social media data can shed light on both types of interaction, as the datasets describe different socio-spatial contexts in which language contact can take place.

6.4. Implications to research on language and geography

Our findings have implications for geographical and linguistic research, particularly in relation to using social media as a source of data. If social media are used as a proxy for estimating the socio-spatial characteristics of the local population (cf. Singleton & Longley, 2009), the effect of daily mobilities must be accounted for, as illustrated by the contrast between register (Fig. 4) and social media data (Fig. 5). This issue is further complicated by language use on social media, as the users

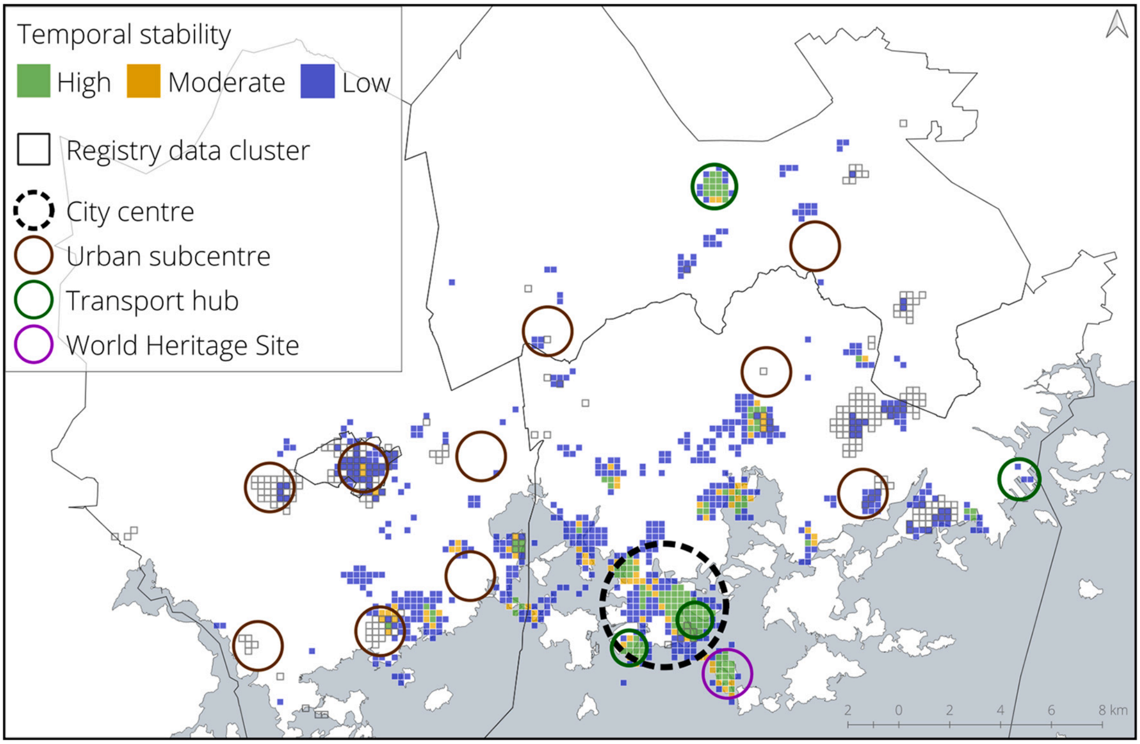


Fig. 6. Spatio-temporal stability of clusters with high linguistic diversity in social media data, based on a bivariate local Moran's *I* analysis of Shannon entropy and Simpson diversity. Clusters with high temporal stability are present four to five times a day, whereas clusters with moderate stability are present three times a day. Low stability clusters are present one to two times a day.

Table 4
Coefficients for OLS regression analysis for Shannon entropy across different times of day and for the full data. A dash indicates that the variable was dropped in the variable selection phase.

	Coefficients					
	Morning	Noon	Afternoon	Evening	Night	Full data
Constant	0.148***	0.224***	0.258***	0.287***	0.220***	0.315***
Non-residential buildings (%)	0.053**	–	0.040**	0.081***	0.077***	0.044**
Dwellings in apartments (%)	0.138***	0.144***	0.122***	0.118***	0.124***	0.121***
Unemployed (%)	0.167**	0.148**	0.216***	0.167**	0.126*	0.158***
Students (%)	0.302***	0.382***	0.178***	0.298***	0.367***	0.275***
Dynamic population (%)	0.222***	0.275***	0.294***	0.302***	0.223***	–
Tertiary degree (%)	0.284***	0.276***	0.239***	0.241***	0.197***	0.231***
Social media posts (%)	0.223**	–	–	–	–	–
Concentration of jobs (%)	–	–	–	–	0.170*	–
Entertainment jobs (%)	–	0.055*	–	–	–	–
Hospitality/retail jobs (%)	–	–	–	0.0260*	–	0.020*
R^2 adjusted	0.226	0.209	0.201	0.191	0.235	0.178

* $p < 0.05$
** $p < 0.01$
*** $p < 0.001$.

can draw on the full repertoire of linguistic resources available to them (Pennycook & Otsuji, 2015), which is reflected in any quantitative measure of linguistic diversity for a given geographical area. A more fine-grained view of social media users may be achieved by using geotags to estimate the home location of individual users (Heikinheimo, Järv, Tenkanen, Hiippala, & Toivonen, 2022) and applying diversity measures to the content they generate (Hiippala et al., 2020). Additionally, identifying potential areas of language contact could potentially provide new perspectives to or augment analyses of socioeconomic geographies of built forms (Martino et al., 2021; Oliveira, 2021). Furthermore, our findings suggest that register data can serve as a point of comparison for social media data by providing information about potential language users and their location in the study area.

6.5. *Methodological considerations and limitations*

6.5.1. *Language choice is platform- and context-dependent*

A comparison of register and social media data revealed considerable differences in the distribution of languages (Table 2). Previous research has found that social media users often draw on multiple languages (Hiippala et al., 2019, 2020), and multilingual individuals, especially immigrants, use their first language mostly at home, and a locally dominant language elsewhere (de Vries, 1990; El Ayadi, 2021; Valentine et al., 2008). This may explain why languages spoken by large linguistic minorities, such as Somali, are not represented in social media data, although the automatic language identification algorithm is capable of detecting these languages. Language choice on social media platforms is also dependent on the context and the intended audience

Table 5
Coefficients for SLM regression analysis for Shannon entropy across different times of day and for the full data. A dash indicates that the variable was dropped in the variable selection phase. Full model summaries for each time of day can be found in the appendices.

	Coefficients					
	Morning	Noon	Afternoon	Evening	Night	Full data
Spatial lag	0.662***	0.675***	0.702***	0.680***	0.664***	0.717***
Constant	0.055***	0.091***	0.085***	0.099***	0.075***	0.101***
Non-residential buildings (%)	0.028*	–	0.025*	0.040***	0.036**	0.021*
Dwellings in apartments (%)	0.040***	0.043***	0.037***	0.033***	0.044***	0.032***
Unemployed (%)	0.061	−0.010	0.034	0.045	0.034	0.023
Students (%)	0.115*	0.130**	0.043	0.056*	0.112**	0.054
Dynamic population (%)	0.073	0.103**	0.078	0.097**	0.069**	–
Tertiary degree (%)	0.089***	0.060***	0.059***	0.069***	0.060***	0.045***
Social media posts (%)	0.079	–	–	–	–	–
Concentration of jobs (%)	–	–	–	–	0.068	–
Entertainment jobs (%)	–	0.029	–	–	–	–
Hospitality/retail jobs (%)	–	–	–	0.018*	–	0.007
R^2	0.648	0.633	0.651	0.618	0.625	0.628

* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

(Androutsopoulos, 2015). Consequently, many speakers of Somali in the Helsinki Metropolitan Area might choose to use Finnish or English publicly on social media, while using Somali in private communications only. The language choice of using Finnish instead of e.g., Somali can be linked to the scarcity of publicly shared social media content about deeply personal linkages to specific places identified by X. Ye and Andris (2021), as Somali might be primarily used in private socio-spatial contexts (de Vries, 1990; El Ayadi, 2021).

The role of English as the lingua franca of social media in Finland and the Nordic countries is very likely to affect our analyses of linguistic diversity (Coats, 2019; Hiippala et al., 2020). Although the dominance of English reduces linguistic diversity, it should be noted that many users are likely to use English as a foreign language. By examining Instagram users' geographical histories and language use at a touristic landmark in Helsinki, Hiippala et al. (2019) showed that the English language gains users from all countries. Thus, the presence of English on social media may in fact signal the presence of multilingual users.

6.5.2. The representativeness of data

The representativeness of social media data warrants attention, as social media users do not represent the entire population. Age, gender, health, socio-economic status, culture, personal choices and geographic location affect the use of digital technologies (Robinson et al., 2020; Robinson et al., 2020; Singleton, Alexiou, & Savani, 2020; X. Ye & Andris, 2021). This 'digital divide' may affect our results, which must be interpreted with these issues in mind, and should not be considered analogous to a systematically collected and representative population sample. Although other data sources, such as mobile phone data, may provide a more representative view, they cannot provide information on everyday language use. The register data, in turn, do not contain information on whether the individual is multilingual or not, nor does it indicate whether the self-reported first language is the main language used by the individual or their level of proficiency (de Vries, 1990; Latomaa, 2012). Furthermore, uneven accessibility to social media data for academic research persists. In early 2021, Twitter opened the full Twitter archive free of charge for academic research (Tornes & Trujillo, 2021), whereas Instagram data remains inaccessible for academic research (for an extended discussion, see Bruns, 2019).

7. Conclusion

Information about languages and language use has much potential in geoinformatics and geography in general. Individual-level register and geotagged social media data provide complementary perspectives into understanding the spatio-temporal patterns of linguistic diversity and where language contact is likely to occur, as register data reflect the presence of potential language users, while social media reflect the daily mobilities of individuals and their everyday language use. Future research efforts should focus on identifying spatial hierarchies among different types of linguistic diversity, discovering links between multilingualism, segregation and gentrification, and validating findings from big data analytics with field work. As urbanization rates increase, and international migration and refugee flows driven by climate change continue, understanding urban diversity, multilingualism and their implications become increasingly important for urban sustainability.

Funding

This work was supported by the Emil Aaltonen Foundation.

CRediT authorship contribution statement

Tuomas Väisänen: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Olle Järv:** Conceptualization, Writing – review & editing, Supervision. **Tuuli Toivonen:** Conceptualization, Resources, Writing – review & editing, Supervision. **Tuomo Hiippala:** Conceptualization, Resources, Writing – original draft, Writing – review & editing, Supervision, Funding acquisition, Project administration.

Declaration of Competing Interest

None.

Acknowledgements

We thank the Emil Aaltonen Foundation for supporting our work through the *Mapping the linguistic landscape of the Helsinki metropolitan area: From linguistic groups to discourse communities (MAPHEL)* project. We are thankful for the Digital Geography Lab data contributors for curating the social media data. We also thank Statistics Finland and their research support services for providing access to individual-level register data.

Appendix A. Software

We 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) Python libraries, and GeoDA (Anselin et al., 2006). We visualized all maps with QGIS (QGIS Development Team, 2021), graphs and plots with Seaborn (Waskom, 2021) and matplotlib (Hunter, 2007) Python libraries.

Appendix B. Variable selection for regression

The full list of independent variables used in OLS and GWR analyses are visible in Tables B1 and B2. Table B1 introduces the variables and the reasons for inclusion in the model. Table B2 describes which variables were dropped from the final model and why. We selected the variables based on what is known in previous literature to influence urban diversity. Furthermore, we added social media post concentration as a variable to control its influence on linguistic diversity.

Table B1
Variables used in the OLS and SLM regression analyses and the conceptual backgrounds by which they were selected.

Variable	Selection background
Unemployed (%)	The unemployed are likely to monolingual, and thus residential areas with high unemployment rates are likely to have low linguistic diversity.
Students (%)	Students from different backgrounds commonly migrate temporarily to study, and stay in student accommodation. Residential areas with student accommodation are thus likely to have higher linguistic diversity.
High education (%)	Highly educated individuals are likely to be fluent in several languages, and thus the areas might be linguistically diverse.
Non-residential buildings (%)	Non-residential buildings contain workplaces, schools, and places of leisure, which attract diverse groups during daily lives.
Dwellings in blocks-of-flat (%)	High-rise residential buildings have a high population density and thus have potentially higher linguistic diversity.
Horeca & Retail jobs (%)	Jobs in the hospitality and retail sectors reflect the locations of restaurants, hotels, and various shops, which are likely to attract individuals with diverse backgrounds.
Dynamic population (%)	Concentration of de facto population in places tend to create potentially higher linguistic diversity levels.
Total job concentration (%)	The concentration of all jobs can potentially capture the influence of the overall job structure.
Entertainment jobs (%)	Concentration of jobs in theatres, cinemas, and other event venues potentially reflects linguistic diversity as social media content is connected to leisure.
Education & Science jobs (%)	Jobs in education and academia are likely to have linguistically diverse individuals from across the globe working there.
Social media post count (%)	Control variable to manage the influence caused by social media post concentration.
International org jobs (%)	International jobs are likely to have multilingual individuals.
Rental dwellings (%)	Rental dwellings are the most affordable type of accommodation, and the majority of immigrants and refugees tend to live in rental dwellings.
Vocational education (%)	Residential areas with vocationally educated individuals are likely to be more monolingual and thus have lower linguistic diversity.
Shannon entropy (register)	Linguistic diversity from the register data to control the influence of linguistically diverse residential areas on linguistically diverse social media content.
Basic education (%)	Residential areas with lowly educated individuals are likely to be more monolingual and thus have lower linguistic diversity.
High income (%)	Individuals with high income levels are likely to be highly educated and to work at an internationally oriented company, thus areas with high income levels have potentially higher linguistic diversity.
Average age	The average age of the population in a grid cell can influence linguistic diversity, as it is more likely that younger inhabitants have diverse or multilingual backgrounds.

Table B2
Variables used in the OLS and SLM regression analyses, including dropped variables. Initial selection was done using expert opinion. A plus symbol (+) indicates variable inclusion, whereas a dash indicates removal from model, either after to Variance Inflation Factor (VIF) filtering or statistical insignificance in stepwise Backwards Regression (BR). Dynamic population was not tested with full social media data.

Variable	Morning	Noon	Afternoon	Evening	Night	Full
Unemployed (%)	+	+	+	+	+	+
Students (%)	+	+	+	+	+	+
High education (%)	+	+	+	+	+	+
Non-residential buildings (%)	+	+	+	+	+	+
Dwellings in blocks-of-flat (%)	+	+	+	+	+	+
Horeca & Retail jobs (%)	- (BR)	- (BR)	- (BR)	+	- (BR)	+
Dynamic population (%)	+	+	+	+	+	(N/A)
Total job count	- (BR)	- (BR)	- (BR)	- (BR)	+	- (BR)
Entertainment jobs (%)	- (BR)	+	- (BR)	- (BR)	- (BR)	- (BR)
Education & Science jobs (%)	- (BR)	- (BR)	- (BR)	- (BR)	- (BR)	- (BR)
Social media post count (%)	+	- (BR)	- (BR)	- (BR)	- (BR)	- (BR)
International org jobs (%)	- (BR)	- (BR)	- (BR)	- (BR)	- (BR)	- (BR)
Rental dwellings (%)	- (VIF)	- (VIF)	- (VIF)	- (VIF)	- (VIF)	- (VIF)
Vocational education (%)	- (VIF)	- (VIF)	- (VIF)	- (VIF)	- (VIF)	- (VIF)
Shannon entropy (registry)	- (VIF)	- (VIF)	- (VIF)	- (VIF)	- (VIF)	- (VIF)
Basic education (%)	- (VIF)	- (VIF)	- (VIF)	- (VIF)	- (VIF)	- (VIF)
High income (%)	- (VIF)	- (VIF)	- (VIF)	- (VIF)	- (VIF)	- (VIF)
Average age	- (VIF)	- (VIF)	- (VIF)	- (VIF)	- (VIF)	- (VIF)

Appendix C. OLS Regression summaries

Here are the OLS regression model summaries for all times of day and full data.

Table C1
OLS Regression Results for morning.

Dep. Variable:	shannon_scaled	R-squared:	0.228
Model:	OLS	Adj. R-squared:	0.226
Method:	Least Squares	F-statistic:	94.93
Date:	Wed, 02 Feb 2022	Prob (F-statistic):	1.23e-121
Time:	15:03:07	Log-Likelihood:	1172.9
No. Observations:	2255	AIC:	−2330.
Df Residuals:	2247	BIC:	−2284.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1481	0.015	9.617	0.000	0.118	0.178
oth_buildings_prop	0.0527	0.018	2.947	0.003	0.018	0.088
students_prop	0.3017	0.068	4.442	0.000	0.169	0.435
block_dwel_prop	0.1378	0.009	15.849	0.000	0.121	0.155
scaled_morning	0.2224	0.064	3.494	0.000	0.098	0.347
unemployed_prop	0.1670	0.061	2.725	0.006	0.047	0.287
postcount_scaled	0.2232	0.080	2.793	0.005	0.067	0.380
high_ed_prop	0.2845	0.026	11.154	0.000	0.234	0.335

Omnibus:	8.139	Durbin-Watson:	1.314
Prob(Omnibus):	0.017	Jarque-Bera (JB):	8.654
Skew:	−0.103	Prob(JB):	0.0132
Kurtosis:	3.222	Cond. No.	36.4

Table C2
OLS regression results for noon.

Dep. Variable:	shannon_scaled	R-squared:	0.210
Model:	OLS	Adj. R-squared:	0.209
Method:	Least Squares	F-statistic:	121.0
Date:	Wed, 02 Feb 2022	Prob (F-statistic):	7.22e-136
Time:	15:06:35	Log-Likelihood:	1435.4
No. Observations:	2732	AIC:	−2857.
Df Residuals:	2725	BIC:	−2815.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2236	0.013	16.915	0.000	0.198	0.250
unemployed_prop	0.1476	0.052	2.842	0.005	0.046	0.249
entertainment_jobs	0.0553	0.024	2.348	0.019	0.009	0.101
block_dwel_prop	0.1438	0.007	19.695	0.000	0.129	0.158
students_prop	0.3819	0.064	5.953	0.000	0.256	0.508
scaled_noon	0.2754	0.056	4.916	0.000	0.166	0.385
high_ed_prop	0.2764	0.022	12.311	0.000	0.232	0.320

Omnibus:	21.905	Durbin-Watson:	1.226
Prob(Omnibus):	0.000	Jarque-Bera (JB):	25.995
Skew:	−0.146	Prob(JB):	2.27e-06
Kurtosis:	3.378	Cond. No.	28.8

Table C3
OLS Regression Results for afternoon.

Dep. Variable:	shannon_scaled	R-squared:	0.202
Model:	OLS	Adj. R-squared:	0.201
Method:	Least Squares	F-statistic:	128.5
Date:	Wed, 02 Feb 2022	Prob (F-statistic):	3.18e-145
Time:	15:07:34	Log-Likelihood:	1789.2
No. Observations:	3047	AIC:	−3564.
Df Residuals:	3040	BIC:	−3522.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2584	0.012	22.408	0.000	0.236	0.281
oth_buildings_prop	0.0405	0.015	2.630	0.009	0.010	0.071
block_dwel_prop	0.1224	0.007	17.801	0.000	0.109	0.136
unemployed_prop	0.2163	0.046	4.687	0.000	0.126	0.307
students_prop	0.1780	0.056	3.185	0.001	0.068	0.288
scaled_afternoon	0.2940	0.063	4.674	0.000	0.171	0.417
high_ed_prop	0.2390	0.020	12.089	0.000	0.200	0.278

Omnibus:	54.420	Durbin-Watson:	1.240
Prob(Omnibus):	0.000	Jarque-Bera (JB):	89.030
Skew:	−0.155	Prob(JB):	4.65e-20
Kurtosis:	3.778	Cond. No.	31.5

Table C4
OLS Regression Results for evening.

Dep. Variable:	shannon_scaled	R-squared:	0.193
Model:	OLS	Adj. R-squared:	0.191
Method:	Least Squares	F-statistic:	102.3
Date:	Wed, 02 Feb 2022	Prob (F-statistic):	1.53e-134
Time:	15:08:16	Log-Likelihood:	1464.7
No. Observations:	3003	AIC:	−2913.
Df Residuals:	2995	BIC:	−2865.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2873	0.013	21.602	0.000	0.261	0.313
block_dwel_prop	0.1184	0.008	14.894	0.000	0.103	0.134
high_ed_prop	0.2409	0.022	10.878	0.000	0.197	0.284
unemployed_prop	0.1669	0.053	3.172	0.002	0.064	0.270
students_prop	0.2977	0.065	4.550	0.000	0.169	0.426
scaled_evening	0.3019	0.049	6.119	0.000	0.205	0.399
horeca_retail	0.0260	0.012	2.150	0.032	0.002	0.050
oth_buildings_prop	0.0805	0.017	4.762	0.000	0.047	0.114

Omnibus:	60.058	Durbin-Watson:	1.168
Prob(Omnibus):	0.000	Jarque-Bera (JB):	79.467
Skew:	−0.249	Prob(JB):	5.55e-18
Kurtosis:	3.622	Cond. No.	29.8

Table C5
OLS Regression Results for night.

Dep. Variable:	shannon_scaled	R-squared:	0.237
Model:	OLS	Adj. R-squared:	0.235
Method:	Least Squares	F-statistic:	120.6
Date:	Wed, 02 Feb 2022	Prob (F-statistic):	1.42e-154
Time:	15:08:45	Log-Likelihood:	1552.4
No. Observations:	2725	AIC:	−3089.

(*continued on next page*)

Table C5 (*continued*)

Dep. Variable:	shannon_scaled	R-squared:	0.237
Df Residuals:	2717	BIC:	−3042.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2205	0.013	17.350	0.000	0.196	0.245
unemployed_prop	0.1260	0.050	2.521	0.012	0.028	0.224
high_ed_prop	0.1972	0.022	9.148	0.000	0.155	0.239
students_prop	0.3675	0.060	6.130	0.000	0.250	0.485
scaled_night	0.2226	0.037	6.064	0.000	0.151	0.295
block_dwel_prop	0.1240	0.008	16.329	0.000	0.109	0.139
oth_buildings_prop	0.0772	0.018	4.363	0.000	0.043	0.112
total_jobs_scaled	0.1701	0.068	2.495	0.013	0.036	0.304

Omnibus:	55.961	Durbin-Watson:	1.321
Prob(Omnibus):	0.000	Jarque-Bera (JB):	68.480
Skew:	−0.279	Prob(JB):	1.35e-15
Kurtosis:	3.541	Cond. No.	32.7

Table C6
OLS Regression Results for full data.

Dep. Variable:	shannon_scaled	R-squared:	0.179
Model:	OLS	Adj. R-squared:	0.178
Method:	Least Squares	F-statistic:	131.7
Date:	Wed, 02 Feb 2022	Prob (F-statistic):	3.74e-151
Time:	15:10:10	Log-Likelihood:	2250.8
No. Observations:	3621	AIC:	−4488.
Df Residuals:	3614	BIC:	−4444.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.3146	0.010	30.905	0.000	0.295	0.335
unemployed_prop	0.1584	0.041	3.820	0.000	0.077	0.240
students_prop	0.2749	0.052	5.260	0.000	0.172	0.377
high_ed_prop	0.2306	0.017	13.310	0.000	0.197	0.265
oth_buildings_prop	0.0441	0.014	3.160	0.002	0.017	0.072
block_dwel_prop	0.1208	0.006	20.330	0.000	0.109	0.132
horeca_retail	0.0195	0.010	2.047	0.041	0.001	0.038

Omnibus:	67.219	Durbin-Watson:	1.070
Prob(Omnibus):	0.000	Jarque-Bera (JB):	111.815
Skew:	−0.159	Prob(JB):	5.24e-25
Kurtosis:	3.800	Cond. No.	29.2

Appendix D. SLM Regression summaries

Here are the spatial lag model regression summaries from GeoDa.

Table D1
SLM Regression results for the full data.

Dep. Variable:	shannon_scaled	No of variables:	8
Model:	SLM	Degrees of Freedom:	3613
Method:	MLE	R-squared:	0.628303
Data:	full_data	Sq. Correlation:	–
Spatial weight:	full_data_queen	Sigma-square:	0.00765018
No. Observations:	3621	S.E. of regression:	0.0874653
Mean dependent var:	0.495671	Log likelihood:	3384.42
S.D. dependent var:	0.143464	AIC:	−6752.84

(*continued on next page*)

Table D1 (*continued*)

Dep. Variable:	shannon_scaled	No of variables:	8
Lag coeff. (Rho):	0.717239	BIC:	−6703.28

	Coef	Std err	z-value	Probability
Spatial lag	0.7172	0.011	63.386	0.000
CONSTANT	0.1009	0.008	12.513	0.000
block_dwel_prop	0.0323	0.004	7.825	0.000
oth_buildings_prop	0.0215	0.009	2.286	0.022
unemployed_prop	0.0234	0.028	0.840	0.401
students_prop	0.0545	0.035	1.549	0.121
horeca_retail	0.0066	0.006	1.031	0.303
high_ed_prop	0.0451	0.012	3.818	0.000

	DF	Value	Probability
Breusch-Pagan test:	6	208.9194	0.00000
Likelihood Ratio test:	1	2267.3070	0.00000

Table D2
SLM Regression results for the morning data.

Dep. Variable:	shannon_scaled	No of variables:	9
Model:	SLM	Degrees of Freedom:	2246
Method:	MLE	R-squared:	0.648409
Data:	morning	Sq. Correlation:	–
Spatial weight:	morning_queen	Sigma-square:	0.00942507
No. Observations:	2255	S.E. of regression:	0.0970828
Mean dependent var:	0.38955	Log likelihood:	1846.13
S.D. dependent var:	0.163728	AIC:	−3674.26
Lag coeff. (Rho):	0.662195	BIC:	−3622.77

	Coef	Std err	z-value	Probability
Spatial lag	0.6622	0.013	50.429	0.000
CONSTANT	0.0551	0.011	5.170	0.000
oth_buildings_prop	0.0279	0.012	2.314	0.021
block_dwel_prop	0.0407	0.006	6.697	0.000
unemployed_prop	0.0608	0.041	1.472	0.141
students_prop	0.1154	0.045	2.518	0.011
scaled_morning	0.0726	0.042	1.690	0.090
high_ed_prop	0.0889	0.017	5.088	0.000
postcount_scaled	0.0791	0.054	1.468	0.142

	DF	Value	Probability
Breusch-Pagan test:	7	104.9281	0.00000
Likelihood Ratio test:	1	1346.3778	0.00000

Table D3
SLM Regression results for the noon data.

Dep. Variable:	shannon_scaled	No of variables:	8
Model:	SLM	Degrees of Freedom:	2724
Method:	MLE	R-squared:	0.632848
Data:	noon	Sq. Correlation:	–
Spatial weight:	noon_queen	Sigma-square:	0.00951876
No. Observations:	2732	S.E. of regression:	0.0975642
Mean dependent var:	0.451703	Log likelihood:	2275.87
S.D. dependent var:	0.161015	AIC:	−4535.73
Lag coeff. (Rho):	0.674939	BIC:	−4488.43

	Coef	Std err	z-value	Probability
Spatial lag	0.6749	0.013	53.882	0.000

(*continued on next page*)

Table D3 (*continued*)

	Coef	Std err	z-value	Probability
CONSTANT	0.0911	0.010	9.379	0.000
block_dwel_prop	0.0433	0.005	8.378	0.000
unemployed_prop	−0.0099	0.035	−0.279	0.780
students_prop	0.1298	0.043	2.966	0.003
scaled_noon	0.1026	0.038	2.681	0.007
high_ed_prop	0.0604	0.015	3.901	0.000
entertainment_jobs	0.0293	0.016	1.830	0.067

	DF	Value	Probability
Breusch-Pagan test:	6	238.9404	0.00000
Likelihood Ratio test:	1	1680.8888	0.00000

Table D4
SLM Regression results for the afternoon data.

Dep. Variable:	shannon_scaled	No of variables:	8
Model:	SLM	Degrees of Freedom:	3039
Method:	MLE	R-squared:	0.651106
Data:	afternoon	Sq. Correlation:	–
Spatial weight:	afternoon_queen	Sigma-square:	0.00791287
No. Observations:	3047	S.E. of regression:	0.0889543
Mean dependent var:	0.45251	Log likelihood:	2810.79
S.D. dependent var:	0.150598	AIC:	−5605.57
Lag coeff. (Rho):	0.701515	BIC:	−5557.4

	Coef	Std err	z-value	Probability
Spatial lag	0.7015	0.011	60.481	0.000
CONSTANT	0.0846	0.008	10.126	0.000
oth_buildings_prop	0.0245	0.010	2.416	0.016
block_dwel_prop	0.0367	0.005	7.815	0.000
unemployed_prop	0.0337	0.031	1.104	0.269
students_prop	0.0433	0.037	1.171	0.242
scaled_afternoon	0.0782	0.042	1.876	0.061
high_ed_prop	0.0590	0.013	4.452	0.000

	DF	Value	Probability
Breusch-Pagan test:	6	180.6144	0.00000
Likelihood Ratio test:	1	2043.2343	0.00000

Table D5
SLM Regression results for the evening data.

Dep. Variable:	shannon_scaled	No of variables:	9
Model:	SLM	Degrees of Freedom:	2994
Method:	MLE	R-squared:	0.617605
Data:	evening	Sq. Correlation:	–
Spatial weight:	evening_queen	Sigma-square:	0.0104593
No. Observations:	3003	S.E. of regression:	0.102271
Mean dependent var:	0.50066	Log likelihood:	2297.61
S.D. dependent var:	0.165385	AIC:	−4577.21
Lag coeff. (Rho):	0.680373	BIC:	−4523.15

	Coef	Std err	z-value	Probability
Spatial lag	0.6804	0.012	55.223	0.000
CONSTANT	0.0992	0.010	9.888	0.000
oth_buildings_prop	0.0403	0.012	3.458	0.001
block_dwel_prop	0.0331	0.006	5.904	0.000
unemployed_prop	0.0452	0.036	1.248	0.212
students_prop	0.0559	0.045	1.241	0.215
scaled_evening	0.0968	0.034	2.844	0.004
high_ed_prop	0.0689	0.015	4.462	0.000

(*continued on next page*)

Table D5 (*continued*)

	Coef	Std err	z-value	Probability
horeca_retail	0.0178	0.008	2.143	0.032

	DF	Value	Probability
Breusch-Pagan test:	7	228.9661	0.00000
Likelihood Ratio test:	1	1665.7697	0.00000

Table D6
SLM Regression results for the night data.

Dep. Variable:	shannon_scaled	No of variables:	9
Model:	SLM	Degrees of Freedom:	2716
Method:	MLE	R-squared:	0.625427
Data:	night	Sq. Correlation:	–
Spatial weight:	night_queen	Sigma-square:	0.00919855
No. Observations:	2725	S.E. of regression:	0.095909
Mean dependent var:	0.424924	Log likelihood:	2325.34
S.D. dependent var:	0.156708	AIC:	−4632.67
Lag coeff. (Rho):	0.663665	BIC:	−4579.48

	Coef	Std err	z-value	Probability
Spatial lag	0.6637	0.013	52.348	0.000
CONSTANT	0.0752	0.009	7.966	0.000
oth_buildings_prop	0.0357	0.012	2.879	0.004
block_dwel_prop	0.0443	0.006	8.047	0.000
unemployed_prop	0.0342	0.035	0.979	0.328
students_prop	0.1189	0.042	2.831	0.005
scaled_evening	0.0694	0.026	2.692	0.007
high_ed_prop	0.0602	0.015	3.946	0.000
total_jobs_scaledl	0.0683	0.048	1.430	0.153

	DF	Value	Probability
Breusch-Pagan test:	7	159.7056	0.00000
Likelihood Ratio test:	1	1545.8685	0.00000

References

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

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

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

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

QGIS Development Team. (2021). QGIS Geographic Information System. https://www.qgis.org

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., & 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

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

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. https://doi.org/10.25080/MAJORA-92BF1922-011

Waskom, M. L. (2021). Seaborn: Statistical data visualization. Journal of Open Source Software, 6 (60), 3021. https://doi.org/10.21105/joss.03021

References

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

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

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

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

Androutsopoulos, J. (2015). Networked multilingualism: Some language practices on Facebook and their implications. *International Journal of Bilingualism, 19*(2), 185–205. https://doi.org/10.1177/1367006913489198

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., Syabri, I., & Smirnov, O. (2002). Visualizing multivariate spatial correlation with dynamically linked windows. In *New Tools for Spatial Data Analysis: Proceedings of the Specialist Meeting* (pp. 1–20).

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

Baldwin, T., Cook, P., Lui, M., MacKinlay, A., & Wang, L. (2013). How Noisy social media text, how diffrnt social media sources?. In *Proceedings of the 6th International Joint Conference on Natural Language Processing* (pp. 356–364). https://aclanthology.org/I13-1041.pdf.

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), 39. https://doi.org/10.1038/s41597-021-01113-4

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

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

Chríost, D. M. G., & Aitchison, J. (1998). Ethnic identities and language in Northern Ireland. *Area, 30*(4), 301–309. Retrieved November 10, 2021, from https://www.jstor.org/stable/20003923.

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

Coats, S. (2019). Language choice and gender in a Nordic social media corpus. *Nordic Journal of Linguistics, 42*(01), 31–55. https://doi.org/10.1017/S0332586519000039

Cvetojevic, S., Juhasz, L., & Hochmair, H. (2016). Positional accuracy of twitter and Instagram images in urban environments. *GI Forum, 4*(1), 191–203. https://doi.org/10.1553/giscience2016 01 s191

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

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., … Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography, 36*(1), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

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.*. 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

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

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

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

Hewidy, H., & Lilius, J. (2021). In the blind spot: Ethnic retailing in Helsinki and the spontaneous placemaking of abandoned spaces. *European Planning Studies.*. 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., 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

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

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.

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

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

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. Springer Science and Business Media.

Järv, O., Masso, A., Silm, S., & Ahas, R. (2020). The link between ethnic segregation and socio-economic status: An activity space approach. *Tijdschrift voor Economische en Sociale Geografie, 112*(3), 319–335. https://doi.org/10.1111/tesg.12465

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

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

Kearns, R. A., & Berg, L. D. (2002). Proclaiming place: Towards a geography of place name pronunciation. *Social & Cultural Geography, 3*(3), 283–302. https://doi.org/10.1080/1464936022000003532

Kiss, T., & Strunk, J. (2006). Unsupervised multilingual sentence boundary detection. *Computational Linguistics, 32*(4), 485–525. https://doi.org/10.1162/coli.2006.32.4.485

Kohvakka, R., Melkas, P., & Tarkoma, J. (2018). Survey on use of information and communications technology by individuals 2018. In *Statistics Finland*. Retrieved April 19, 2022, from https://www.stat.fi/til/sutivi/2018/sutivi_2018_2018-12-04_tie_001_en.html.

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

Kukk, K., van Ham, M., & Tammaru, T. (2019). EthniCity of leisure: A domains approach to ethnic integration during free time activities. *Tijdschrift voor Economische en Sociale Geografie, 110*(3), 289–302. https://doi.org/10.1111/tesg.12307

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

Latomaa, S. (2012). Kielitilasto maahanmuuttajien väestöosuuden mittarina. *Yhteiskuntapolitiikka, 77*(5), 525–534. https://www.julkari.fi/handle/10024/103125.

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

Magdy, A., Ghanem, T. M., Musleh, M., & Mokbel, M. F. (2014). Exploiting geo-tagged tweets to understand localized language diversity. In *Proceedings of Workshop on Managing and Mining Enriched Geo-Spatial Data - GeoRich'14*. https://doi.org/10.1145/2619112.2619114

Magurran, A. E. (2013). *Measuring biological diversity (1st)*. Wiley.

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. In *Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016* (pp. 639–642). https://arxiv.org/abs/1603.02718.

Martino, N., Girling, C., & Lu, Y. (2021). Urban form and livability: Socioeconomic and built environment indicators. *Buildings and Cities, 2*(1), 220–243. https://doi.org/10.5334/bc.82

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

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), 61981. https://doi.org/10.1371/journal.pone.0061981

Morris, E. K., Caruso, T., Buscot, F., Fischer, M., Hancock, C., Maier, T. 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

Oliveira, V. (2021). Urban form and the socioeconomic and environmental dimensions of cities. *Journal of Urbanism: International Research on Placemaking and Urban Sustainability, 0*(0), 1–23. https://doi.org/10.1080/17549175.2021.2011378

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.

Poblete, B., Garcia, R., Mendoza, M., & Jaimes, A. (2011). Do all birds tweet the same?. In *, 1025. Proceedings of the 20th ACM International Conference on Information and Knowledge Management - CIKM 11*. https://doi.org/10.1145/2063576.2063724

Quinn, S. (2016). A Geolinguistic approach for comprehending local Inuence in OpenStreetMap. *Cartographica: The International Journal for Geographic Information and Geovisualization, 51*(2), 67–83. https://doi.org/10.3138/cart.51.2.3301

Robinson, L., Schulz, J., Blank, G., Ragnedda, M., Ono, H., Hogan, B., … Khilnani, A. (2020). Digital inequalities 2.0: Legacy inequalities in the information age. *First Monday*. https://doi.org/10.5210/fm.v25i7.10842

Robinson, L., Schulz, J., Dunn, H. S., Casilli, A. A., Tubaro, P., Carvath, R., … Khilnani, A. (2020). Digital inequalities 3.0: Emergent inequalities in the information age. *First Monday*. https://doi.org/10.5210/fm.v25i7.10844

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

Singleton, A., Alexiou, A., & Savani, R. (2020). Mapping the geodemographics of digital inequality in Great Britain: An integration of machine learning into small area estimation. *Computers, Environment and Urban Systems, 82*, Article 101486. https://doi.org/10.1016/j.compenvurbsys.2020.101486

Singleton, A., & Longley, P. (2009). Geodemographics, visualisation, and social networks in applied geography. *Applied Geography, 29*(3), 289–298. https://doi.org/10.1016/j.apgeog.2008.10.006

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

Tammaru, T., Knapp, D., Silm, S., van Ham, M., & Witlox, F. (2021). Spatial underpinnings of social inequalities: A vicious circles of segregation approach. *Social Inclusion, 9*(2), 65–76. https://doi.org/10.17645/si.v9i2.4345

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* (pp. 49–64). Springer. https://doi.org/10.1007/978-3-319-40902-3_3.

Tornes, A., & Trujillo, L. (2021). Enabling the future of academic research with the Twitter API. Retrieved April 20, 2021, from https://blog.twitter.com/developer/enus/topics/tools/2021/enabling-the-future-of-academic-research-with-the-twitter-api.html.

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

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

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

Vorobeva, E., Jauhiainen, J. S., & Tammaru, T. (2021). Language, networks, and virtual transnationalism: The case of Russian speakers from Estonia living in Finland. *International Migration.*. https://doi.org/10.1111/imig.12969

de Vries, J. (1990). On coming to our census: A layman's guide to demolinguistics. *Journal of Multilingual and Multicultural Development, 11*(1–2), 57–76. https://doi.org/10.1080/01434632.1990.9994401

Wanjiru, M. W., & Matsubara, K. (2017). Street toponymy and the decolonisation of the urban landscape in post-colonial Nairobi. *Journal of Cultural Geography, 34*(1), 1–23. https://doi.org/10.1080/08873631.2016.1203518

Watson, S. (2009). The magic of the marketplace: Sociality in a neglected public space. *Urban Studies, 46*(8), 1577–1591. https://doi.org/10.1177/0042098009105506

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

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

Ye, X., & Andris, C. (2021). Spatial social networks in geographic information science. *International Journal of Geographical Information Science, 35*(12), 2375–2379. https://doi.org/10.1080/13658816.2021.2001722