The Impact of Environmental Characteristics on Mental Health - A Sentiment Analysis in Switzerland based on Twitter Data

GEO 511 Master's Thesis

Author

Nicolas Schmidheiny

15-709-934

Supervised by

Dr. Oliver Grübner

Dr. phil. Markus Wolf

Faculty representative

Prof. Dr. Sara Irina Fabrikant

30.04.2022

Department of Geography, University of Zurich

Abstract

Mental health research became increasingly relevant in recent years. It is estimated that one in twenty adults worldwide suffers from depression. Social stigma often prevents people from seeking help, leading to high dark figures in psychological statistics. A growing number of studies show that the physical and societal environment play an important role for mental well-being. More specifically, the neighbourhood in which people live can significantly affect their mental health outcomes, by exposing them to certain risk factors, or by featuring tranquillizing characteristics. Big data approaches using social media like Twitter have emerged to a highly promising new research method to analyse psychological characteristics and mental health in society. Geolocated social media data enable the analysis of spatial patterns of sentiments, which can be used as proxies for mental well-being.

This work set out to investigate if Twitter data can be used to detect associations between neighbourhood environmental characteristics and sentiments expressed in tweets on a user-level, by performing a regression analysis. Initially, roughly 12.5 million tweets from Swiss users were considered. The Twitter data used in this thesis were beforehand analysed with Botometer, the M3-method and a DBSCAN-approach. This enabled to control for bots, organisations, gender, and age, as well as using the presumed homeplace locations of the analysed users to define their neighbourhood area. Using the NLP-systems EMOTIVE, Stresscapes and LIWC, rates in positive and negative emotions, and stress were assigned to the users, representing the response variables. The physical and societal neighbourhood characteristics were represented by available greenspace, exposure to traffic noise, and socio-economic position, which posed the explanatory variables of the regression model. High resolution spatial data for the extent of Switzerland were used to calculate the neighbourhood variables within a 500 m buffer around the homeplace for each user. Using ArcGIS Pro Modelbuilder and Python, a workflow was implemented to account for distance decay. Additionally, data from the Swiss federal office of statistics were used to control for urban-rural differences.

Out of the 70’333 available users, 733 were used in the final regression analysis, after proceeding through a careful selection process. On average, 236 tweets per user defined their sentiment rates. Significant negative associations were found between traffic noise and positive emotions. Although accounting for numerous influencing factors and controlling for bias, the pursued approach shows several limitations, where insufficient sample size may be one of the most prominent. Nevertheless, promoting further research in assessing the impact of environmental characteristics on mental well-being using geolocated tweets could provide valuable new insights.

Zusammenfassung

Die Forschung zur psychischen Gesundheit hat in den letzten Jahren zunehmend an Bedeutung gewonnen. Schätzungen zufolge leidet einer von zwanzig Erwachsenen weltweit an einer Depression. Die soziale Stigmatisierung hält die Menschen oft davon ab, Hilfe zu suchen, was zu einer hohen Dunkelziffer in psychologischen Statistiken führt. Eine wachsende Zahl von Studien zeigt, dass das physische und gesellschaftliche Umfeld eine wichtige Rolle für das psychische Wohlbefinden spielt. Insbesondere die Nachbarschaft, in der die Menschen leben, kann sich erheblich auf ihre psychische Gesundheit auswirken, indem sie sie bestimmten Risikofaktoren aussetzt oder beruhigende Eigenschaften aufweist. Big-Data-Ansätze, die soziale Medien wie Twitter nutzen, haben sich zu einer vielversprechenden neuen Forschungsmethode entwickelt, um psychologische Merkmale und die psychische Gesundheit in der Gesellschaft zu analysieren. Geolokalisierte Social-Media-Daten ermöglichen die Analyse räumlicher Stimmungsmuster, die als Indikatoren für das psychische Wohlbefinden verwendet werden können.

In dieser Arbeit wurde untersucht, ob Twitter-Daten verwendet werden können, um Assoziationen zwischen Umgebungsmerkmalen und den in Tweets ausgedrückten Stimmungen auf Nutzerebene zu erkennen, indem eine Regressionsanalyse durchgeführt wurde. Zunächst wurden etwa 12,5 Millionen Tweets von Schweizer Nutzern berücksichtigt. Die in dieser Arbeit verwendeten Twitter-Daten wurden zuvor mit Botometer, der M3-Methode und einem DBSCAN-Ansatz analysiert. Dies ermöglichte die Kontrolle für Bots, Organisationen, Geschlecht und Alter, sowie die Verwendung der vermuteten Wohnorte der analysierten Nutzer, um ihre Nachbarschaft zu definieren. Mit Hilfe der NLP-Systeme EMOTIVE, Stresscapes und LIWC wurden den Nutzern Werte für positive und negative Emotionen und Stress zugeordnet, die die Antwortvariablen darstellen. Die physischen und gesellschaftlichen Merkmale der Nachbarschaft wurden durch verfügbare Grünflächen, die Belastung durch Verkehrslärm und die sozioökonomische Position dargestellt, die die erklärenden Variablen des Regressionsmodells darstellten. Schweizweite, hochaufgelöste räumliche Daten wurden verwendet, um die Nachbarschaftsvariablen innerhalb eines Puffers von 500 m um den Wohnort für jeden Nutzer zu berechnen. Mit Hilfe von ArcGIS Pro Modelbuilder und Python wurde ein Workflow implementiert, um den Entfernungsabfall zu berücksichtigen. Zusätzlich wurden Daten des Schweizerischen Bundesamtes für Statistik verwendet, um die Unterschiede zwischen Stadt und Land zu berücksichtigen.

Von den 70'333 verfügbaren Nutzern wurden nach einem sorgfältigen Auswahlverfahren 733 in die endgültige Regressionsanalyse aufgenommen. Im Durchschnitt definierten 236 Tweets pro Nutzer ihre Stimmungswerte. Es wurde ein signifikanter negativer Zusammenhang zwischen Verkehrslärm und positiven Emotionen festgestellt. Trotz der Berücksichtigung zahlreicher Einflussfaktoren und der Kontrolle von Verzerrungen weist der verfolgte Ansatz mehrere Einschränkungen auf, von denen die unzureichende Stichprobengröße eine der wichtigsten sein dürfte. Nichtsdestotrotz könnte die Förderung weiterer Forschung zur Bewertung der Auswirkungen von Umweltmerkmalen auf das psychische Wohlbefinden unter Verwendung von geografisch verorteten Tweets wertvolle neue Erkenntnisse liefern.

Acknowledgements

The submission of this thesis concludes more than six years of my studies at the University of Zürich. I would like to thank all those who have accompanied and supported me during this time. I am particularly grateful for the support, advice, and motivation given by:

Dr. Oliver Grübner; Thank you so much for the close supervision from the beginning onwards, the countless hours spent on Zoom meetings, and for getting me back on track every now and then.

Dr. phil. Markus Wolf; Thank you for co-supervising me in this thesis and for your interesting inputs from a psychological perspective.

Dr. Martin Sykora & Dr. Suzanne Elayan; Thank you for helping me setting up this thesis and for answering all my technical questions so extensively and patiently.

Philipp Schmidheiny & Jeannine Suremann; Thank you for the careful proofreading and giving me valuable feedback.

Dr. Magnus Heitzler; Thank you for providing the awesome shaded relief for my maps.

Larissa Schädler, Lino Asper, Silvan Caduff, Michael Niederberger, Nicolas Steinmann: Thank you for giving me permission to use our greenspace dataset.

My Friends & Family; Thank you for reminding me that other things than this thesis exist.

Adrian von Ziegler, Ludovico Einaudi, and Grandbrothers; Thank you for making such inspiring and motivating music and for putting it on Spotify.

Table of Contents

List of Figures

List of Tables

Acronyms

API	Application Programming Interface
AOI	Area Of Interest
BAFU	Bundesamt für Umwelt
BLN	Bundesinventar der Landschaften und Naturdenkmäler
CI	Confidence Interval
EDA	Exploratory Data Analysis
FID	Feature Identifier
IDW	Inverse Distance Weighting
MCA	Multi-Criteria Analysis
NLP	Natural Language Processing
PCC	Pearson Correlation Coefficient
SEM	Socio-Ecological Model
SEP	Socio-Economic Position
WHO	World Health Organisation

Chapter 1 | Introduction

The World Health Organization (WHO) estimates that globally, 5% of adults suffer from depression. It is a major contributor of the overall global burden of disease and is a leading cause of disability worldwide. Although effective treatments are known, far too many people receive no treatment, because there is a lack of resources or a lack of trained health-care providers. However, one of the most important barriers to effective care is social stigma associated with mental disorders, which leads to a high number of unreported cases (Evans-Lacko et al., 2018). Also, the assessment of community-level psychological characteristics remains a great challenge to this day. Traditional approaches using phone surveys and household visits are costly and are limited by their spatial and temporal precision (Auchincloss et al., 2012).

However, in the last decade, there has been a renewed focus on the links between the psychological characteristics of people and the characteristics of the places in which they live. This field, known as geographical psychology, aims to understand psychological phenomena on the basis of their spatial distribution and their interactions with features of the environment (Chen et al., 2020). Health geography, a closely related field and subdiscipline of human geography, studies the role of place, location and geography in health, well-being and disease (Dummer, 2008). Research has shown that the place where people live significantly affects their health outcomes (Tunstall et al., 2004). Within health geography, spatial epidemiology describes and analyses geographical differences in diseases in terms of different risk factors. Those include demographic, socio-economic and environmental characteristics, as well as genetic predispositions and behavioural risk factors. The research field traditionally focusing on disease outbreaks like yellow fever and cholera in the 1800s (Elliott & Wartenberg, 2004), continued to evolve and brought new specialised subfields such as digital spatial epidemiology, using big data approaches to study (mental) health outcomes in space.

Big Data methodology using social media has great potential to contribute to mental health research and should therefore be strongly promoted to further investigate the emerging and evolving mechanisms of geographic differences in psychological phenomena and health outcomes (Chen et al. 2020). The rapid growth in the number of social media users worldwide generates an immense quantity of data. This data can be studied and analysed, offering new opportunities to detect and track patterns of social phenomena, such as public health concerns. The information contained in these social media data may include spatial information in the form of coordinates, which allows to apply advanced spatial analysis to detect geographical trends in human behaviour and public health concerns and is now increasingly used in mental health research (Naslund et al., 2019).

1.1 Motivation and Goal

Over the last decade, a growing number of research studies have shown that social media holds great potential for mental health research (Dredze, 2012; Naslund et al., 2019; Gruebner et al., 2017). In health geography, digital spatial epidemiology has emerged to a highly promising research field, successfully linking emotions expressed in social media data with health outcomes (Eichstaedt et al., 2015; Jashinsky et al., 2014). However, so far, the spatial granularity of these studies has mostly been very coarse, meaning that trends in health outcomes and mental well-being were only found on high spatial aggregation levels. This raises the question, if patterns and associations between space and indicators of mental health found in social media, can also be recognised in lower-scale geographies such as urban districts or even in neighbourhoods.

Therefore, the broader aim of the study is the assessment whether social media data like Twitter tweets can be used to detect associations between spatial phenomena at a neighbourhood scale and mental well-being on a user-level. The study area is restricted to the political boundary of Switzerland. In *Subsection 2.2*, similar studies using Twitter data are introduced which were conducted in the form of macro- to meso-level analyses, operating on the geographic scale of, for example, states or counties. However, in this thesis the goal is to conduct an ecological analysis on a smaller spatial unit, namely the neighbourhood, making it a meso- to micro-level analysis. A further distinct difference between the approach in this thesis and studies conducted so far, is the aggregation of tweets onto the user, instead of aggregating single tweets onto administrative boundaries. In this way, it is tried to associate the mental well-being of a user with the characteristics of the neighbourhood in which the user lives. As a measure for mental well-being, different variables are used, which were added to the Twitter datasets using natural language processing algorithms (NLP) categorizing single tweets into emotions and stress-levels. The applied NLP-algorithms are described in *Subsection 2.3*.

Research in geographical psychology and digital spatial epidemiology using social media could reveal spatial patterns of society in mental health and mental well-being. It would provide new insights in the general mental state of the population related to the environment and bypass the distortion caused by unrecorded cases of mental illness due to social stigma. Consequently, preventive measures could be taken to combat the environmental problems having negative influences on mental health outcomes.

1.2 Structure

Chapter 2 introduces the theoretical background around the state of research, and crucial concepts. In Chapter 3 the research gaps to be filled are addressed and the research objectives of the thesis are introduced. The data used in this study is introduced in Chapter 4, along with its spatial and non-spatial structure, as well as its source. Chapter 5 describes the methodological approach for the accomplishment of the research objectives in detail. In Chapter 6 the results of the research objectives are presented and illustrated using tables and visualisations. The discussion around the results and their underlying methodology is opened in Chapter 7. The outcome of the thesis is critically analysed and interpreted, pointing out the strengths and weaknesses of the applied methods. And finally, in Chapter 8, the findings of the thesis are concluded and potential improvements for future work is proposed.

Chapter 2 | Theoretical Background

The theoretical background about the research context of this study, as well as overviews on crucial concepts and applied algorithms are introduced in this chapter. First, the socio-ecological model, the concept and definitions of neighbourhood, and crucial environmental factors on mental health are introduced. Then, the role of social media in mental health research at the example of related studies is described, and an overview of Twitter is given. And finally, different important methods and algorithms applied on the Twitter data used in this thesis are briefly explained.

2.1 Socio-Ecological Model

In order to embed the influence of both societal and physical neighbourhood characteristics on (mental) health into a theoretical perspective, in the following, the socio-ecological model (SEM) is introduced. The American psychologist Urie Bronfenbrenner introduced the socio-ecological model for the first time in the 1970s as a conceptual model to understand human development. In the 1980s it formalized as a theory and was illustrated by nesting circles which place the individual human being in the centre surrounded by multiple influential systems (Kilanowski 2017). Today, countless variations and re-interpretations of the model exist, however the basic idea has remained the same: The microsystem which is closest to the individual, includes the interactions and relationships of the immediate surroundings, and has therefore the strongest influence. The microsystem is embedded in multiple other systems which are all enclosed by the most outer system, usually referred to as macrosystem. The macrosystem has the least influence on the individual, and includes attitudes and ideologies of a culture, or in adapted models also physical factors such as weather and topography (Bornstein and Davis 2014).

Caesar et al. (2020) introduce an adapted SEM for health behaviour, which is illustrated in *Figure 1*.

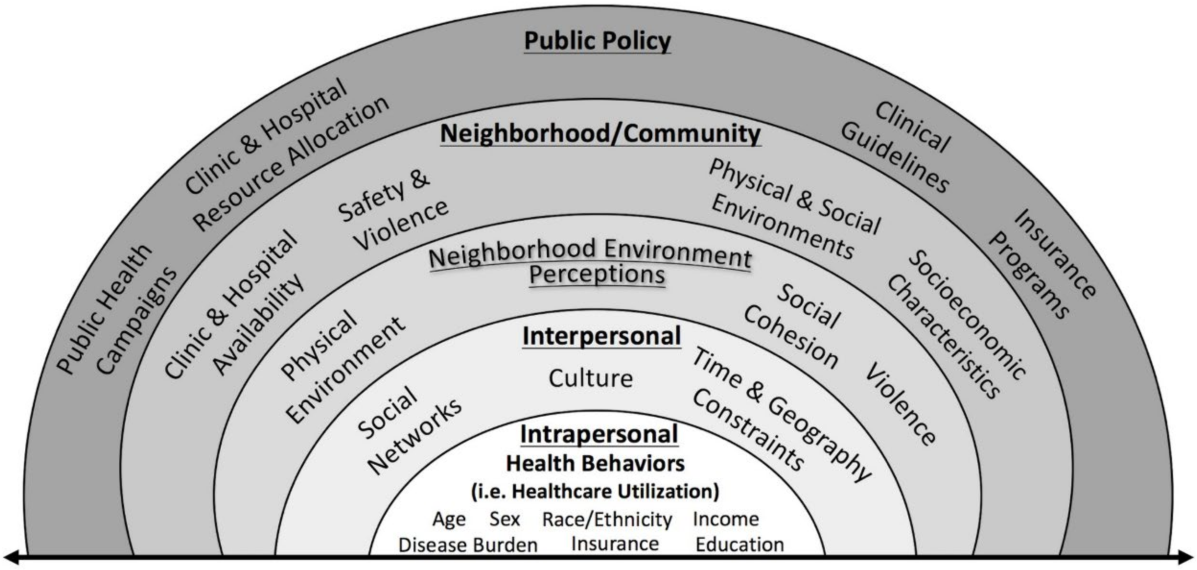

Public Policy
Clinic & Hospital Resource Allocation
Clinical Guidelines
Public Health Campaigns
Insurance Programs
Neighborhood/Community
Safety & Violence
Physical & Social Environments
Clinic & Hospital Availability
Socioeconomic Characteristics
Neighborhood Environment Perceptions
Physical Environment
Social Cohesion
Violence
Interpersonal
Culture
Social Networks
Time & Geography Constraints
Intrapersonal
Health Behaviors
(i.e. Healthcare Utilization)
Age Sex Race/Ethnicity Income
Disease Burden Insurance Education

Figure 1: Socio-ecological model for health behaviour (Caesar et al, 2020).

In this model, the influencing factors on the neighbourhood scale, such as socio-economic characteristics, and physical and social environments are placed in the third and fourth system between the interpersonal system and the public policy system. This SEM enables a better understanding of how strong the influence of the neighbourhood may be on the individual, when put in relation to the intrapersonal or interpersonal factors.

2.2 Neighbourhood

As already emphasized, literature shows that neighbourhood characteristics are crucial for (mental) health outcomes (Cutrona et al., 2006; Goldsmith et al., 1998). Also, time spent at home represents in average 70% of the time budget, which is why public health research lays focus on environmental exposure factors at home and in the immediately surrounding neighbourhood (Tenailleau et al., 2014). However, this raises the question, how the term neighbourhood is usually defined in terms of spatial extent.

In context of cities in the United States, the spatial extent of a neighbourhood is often defined as the area of census tracts (O'Campo et al., 2015). However, this simple approach is limited by the fact that administrative partitions do not capture the actual activity spaces of citizens very accurately. Martí et al. (2021) introduce more "meaningful neighbourhood boundaries" by using *Google Places* data as a source of information on urban activity patterns. Panczak et al. (2012) used buildings and road network data to define overlapping neighbourhood areas in Switzerland.

These more sophisticated approaches are however quite complex, require advanced know-how, and their implementation is very time consuming. Simpler commonly used approaches include buffer techniques to represent the immediate living neighbourhoods of the studied subjects and are used as approximations for the "walking neighbourhood" where most of the daily needs are met (Tenailleau et al., 2015). For example, Tenailleau et al. (2014) assessed residential exposure to urban noise by defining the local living neighbourhoods using 50 to 400 meters buffers around homeplaces. Standard "walking neighbourhood" definitions are also found to be as large as 1-kilometre buffers or 1-mile buffers, which is around 1.6 kilometres (Smith et al., 2010).

2.3 Environmental Factors on Mental Health

It is increasingly recognized that mental health is not only affected by personal characteristics, but also by environmental exposures, which can either contribute to a worsening or to the improvement of a mental condition (Helbich, 2018). O'Campo et al. (2015) stated that hundreds of cross-sectional and more recently longitudinal studies have linked neighbourhood area characteristics, both physical and social, to a range of health behaviours and outcomes such as distress, anxiety, and depression. With growing urbanization, the number of people exposed to risk factors originating from the urban social (e.g., poverty) or physical environment (e.g., traffic noise) is ever increasing. The exposure to these risk factors leads to increased stress, which is negatively associated with mental health outcomes (Gruebner et al., 2017).

In the following subsections, the importance of greenspace, traffic noise and socio-economic position on mental health is described. These three factors build the main components in the analysis of this thesis.

2.3.1 Greenspace

The benefits of greenspace for humans living in cities or settlements is widely discussed in literature. There seems to be broad agreement on the positive impacts of accessible greenspace for human health and well-being especially in or near cities and settlements (Taylor & Hochuli, 2017; Beyer et al., 2014). The trend in increasing numbers of people suffering from depressive disorders, may be related to increased urbanisation, with more than 77% of people in the world's more developed regions now residing in urban areas, and to reduced access to "natural" spaces which aid stress reduction. Epidemiological studies find that individuals living in the greenest urban areas tend to have better mental health than those in the least green areas (Alcock et al., 2014).

James et al. (2015) further express the importance of not only the existence of greenspace, but also the closeness to inhabitants, framing it as "neighbourhood greenness", really emphasizing the notion of easy access and proximity. The value of proximity to greenspace becomes evident when considering the fact that real estate prices are highest near greenspaces (Camargo, 2016).

2.3.2 Traffic Noise

A great and growing environmental problem in residential areas is noise from transport. Findings from a large body of studies show that traffic noise causes non-auditory stress effects such as changes in the physiological systems, various cognitive deficits (e.g., poor sustained attention, memory/concentration problems), sleep disturbances, and emotional/motivational effects (Gidlöf-Gunnarsson and Öhrström, 2007). In a more recent study, exposure to road traffic noise above 65 decibel was associated with changes in blood pressure and cardiovascular biochemistry, which is an indicator for increased stress (Kupcikova et al., 2021).

2.3.3 Socio-Economic Position

It is widely recognised that poorer individual socio-economic circumstances are generally associated with less favourable (mental-) health outcomes (Panczak et al., 2012). At a community level, low socio-economic position (SEP) may lead to greater concerns about neighbourhood safety and decrease the amount of physical activity in the community, which has consequent impacts on mental health (Macintyre et al., 2018). Mann et al. (2022) used multilevel modelling to analyse data collected from 7866 participants aged 40 to 65 years in Brisbane, who participated in a study called HABITAT (How Areas in Brisbane Influence health and AcTivity). They concluded that both individual-level SEP and neighbourhood disadvantage are associated with mental well-being. Moreover, in a study based on a sample of 1010 adults with diabetes, knowing that persons with diabetes have higher rates of depression, the impact of neighbourhood SEP on the severity of the depression was investigated. The researchers found that lower neighbourhood SEP was significantly associated with poorer (mental-) health outcomes (Gary-Webb et al., 2011).

2.4 Social Media in Psychology and Mental Health Research

Social media data have been used extensively in marketing for quantifying specific personality traits and dimensions as for example the ‘Big 5’ (openness to experience, conscientiousness, extraversion, agreeableness, neuroticism) from Facebook data (Schwartz et al., 2014). Social media has established itself as a very powerful data source in politics and business and is now also increasingly used for large scale health monitoring, as well as for mental health research. The analysis of social media is particularly promising in the mental health domain, since users of Twitter, Facebook, etc., provide raw and unfiltered insights to their thoughts, behaviours, and feelings by posting often very intimate messages, which may be indicative of emotional well-being (Conway & O’Connor, 2016).

There have been multiple studies in the US performing spatial analyses using Twitter data on county or on state level. Eichstaedt et al. (2015) for example used 148 million geolocated tweets across 1’347 counties to predict heart disease mortality, based on emotions derived with natural language processing and machine learning. It was found that negative emotions in tweets were highly correlated with heart disease mortality figures, even more highly correlated than official socio-economic, demographic, and health statistics (Eichstaedt et al., 2015). Jashinsky et al. (2014) used Twitter data, some of them geolocated, to track suicide risk factors in the US at state level. The suicidal tweets per state were compared against national data of actual suicide rates from the Centres for Disease Control and Prevention, observing a strong correlation.

Furthermore, sentiment analysis using geolocated Twitter data has been conducted in the context of disasters. One example is the study carried out by Gruebner et al. (2018), where roughly 1 million geolocated tweets that were analysed with an advanced sentiment analysis algorithm called EMOTIVE (see *Subsection 2.7.1*), have been used to calculate discomfort rates for 2137 New York City census tracts, before, during, and after Superstorm Sandy. Discomfort was defined as the combination of different negative emotions. They found increased discomfort after the storm as compared to during the storm, with prominent spatial clusters in Staten Island. In another similar study, also in the context of a traumatic event, a spatial analysis of emotions was performed by Gruebner et al. (2016) using geolocated tweets in Paris during the terrorist attacks in November 2015. They found spatial clusters of tweets expressing sadness in the area of the attacks. Moreover, Edry et al. (2021) developed a web-based geovisualization tool as a proof-of-concept analysis, for the spatio-temporal surveillance of negative emotions and stress found in Tweets in New York City before and during the COVID-19 pandemic. They identified hotspots of stress in the census tracts of Manhattan and Brooklyn, appearing after the lockdown in April 2020.

The mentioned studies demonstrate various examples of how social media, especially Twitter, has been used in mental health research so far. In Chapter 3, research gaps to be filled are pointed out, which lead to the research objectives this thesis attempts to cover.

2.5 Twitter

Twitter is an American social media service where users can share information in a real-time news feed by posting brief messages known as “tweets” about their experiences and thoughts (Maclean et al., 2013). The widely used free social networking tool, founded in 2006, allows people to write tweets of a maximum length of 280 characters, after doubling the former limit of 140 characters in November 2017 (Gligori et al., 2020). The tweets may include images, videos, links to web pages and similar online material. In health research using big data approaches, Twitter is one of the most frequently used social media services, because it is a publicly available resource that can be widely used for research purposes (Jordan et al., 2018; Sinnenberg et al., 2017).

2.5.1 Twitter Popularity in Switzerland Compared to the World

The top 10 leading countries based on number of Twitter users are listed in *Table 1* (Statista, 2022). In January 2021, in Switzerland around 746’000 (world: 353.1 million) Twitter accounts existed, which was around 9.9% (world: 5.8%) of the total population aged 13 and older, of which 22.4% (world: 31.5%) were reported to be female users and 77.6% (world: 68.5%) to be male (Datareportal, 2021a; 2021b).

Table 1: Top 10 leading countries based on number of Twitter users.

Country	Number of Users [Million]	Percentage of Population
United States	76.9	~ 23%
Japan	58.95	~ 47%
India	23.6	~ 1.7%
Brazil	19.05	~ 9%
Indonesia	18.45	~ 6.7%
United Kingdom	18.4	~ 27%
Turkey	16.1	~ 19%
Saudi Arabia	14.1	~ 40%
Mexico	13.9	~ 10.7%
Thailand	11.45	~ 16%

2.5.2 Botometer

Botometer (formerly BotOrNot) is a popular bot detection tool for Twitter tweets. Bots or social bots are inauthentic social media accounts which are partially controlled by algorithms. A social bot, also known as sybil account, automatically produces content and interacts with humans on social media (Davis et al., 2016). Bots are used for orchestrated spreading of misinformation, large-scale opinion manipulation, as well as adding confusion to online debates (Ferrara et al., 2016; Subrahmanian et al., 2016; Broniatowski et al., 2018). Efficient and reliable bot detection methods are crucial to estimate the proportion of automated inauthentic accounts and their influence on social media (Yang et al., 2020). Botometer calculates a score for Twitter accounts, where low scores indicate a low likelihood of being a social bot, and high scores indicate a high likelihood of being a social bot. The score is calculated using over a thousand features to characterize the account's profile, friends, social network structure, temporal activity patterns, language, and sentiment. The features are used by multiple machine learning models to finally compute the bot scores, which range from 0 to 100 percent (Yang et al., 2020).

2.6 M3-Method

The M3-method is a multimodal deep neural system operating on social media accounts to infer probabilities, whether the user is male or female, whether the user belongs to defined age categories, and whether the user is an organisational account. M3 stands for multimodal, multilingual, and multi-attribute abilities. It operates on 32 different languages, and uses social media profile image, username, "screen name" and biography to infer the described variables. The M3-method was applied onto the Twitter data used in this thesis (see *Subsection 4.1.2*) (Wang et al., 2019).

2.7 Natural Language Processing

Natural Language Processing (NLP) dates back to the 1950s and can be viewed as the intersection of artificial intelligence and linguistics, with the aim to make computers understand the statements or words written in human languages (Nadkarni et al., 2011; Khatter et al., 2017). In the following subsections, three NLP-systems are introduced which were applied on the Twitter data used in this thesis.

2.7.1 EMOTIVE

EMOTIVE (Extracting the Meaning Of Terse Information in a Visualization of Emotion) is an ontology-based NLP-system, which can detect 8 different emotions in social media messages. The emotions include Anger, Confusion, Disgust, Fear, Happiness, Sadness, Shame, and Surprise. EMOTIVE was developed as a reaction on existing systems only distinguishing between positive and negative emotions. The ontology on which the system is built, was constructed by performing an in-depth study of language containing emotional expressions. This was done by an English language and literature PhD level research associate during a three-month time-period, where around 600 MB of cleaned tweets were analysed. The ontology does not only contain single words, but also multi-word phrases (Sykora et al., 2013).

2.7.2 Stresscapes

Stresscapes is an ontology-based system that automatically captures, measures and monitors expressions of stress on various text-based social media massages. It was developed to combat the rise of chronic stress-related diseases, by enabling a better understanding of the causes behind the stress social media users experience, as well as the spatial patterns where stress occurs. The underlying ontology model consists of over 1530 terms, of which 1100 are multi word phrases, such as "stressed out", or "under pressure". It was built with the help of a researcher with experience in discourse analysis and linguistics, which combed through thousands of social media messages of ‘stressful’ events (Elayan et al., 2020).

2.7.3 LIWC

LIWC (Linguistic Inquiry and Word Count) is an NLP software similar to EMOTIVE and Stresscapes, which was however not designed specifically for the analysis of social media messages, but rather for the analysis of different text genres, such as emails, speeches, or transcribed everyday language for psychological research (Golder & Macy, 2011). The output is in the form of dozens of different linguistic categories, all of them reporting the percentage of words which fell into that specific category (Cohn et al., 2004). As an example, the two categories or variables used in this thesis are *posemo* and *negemo*, indicating the percentage of words in the text associated with positive or negative emotions respectively.

2.8 Cluster Analysis with DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a spatial data clustering algorithm, which searches for data points having dense neighbourhood, grouping them together as clusters. It is one of the most frequently used and most cited clustering algorithms in scholarly literature (Shinde & Sankhe 2017). In studies with Twitter data, DBSCAN has been widely used for cluster detection, due to its high suitability for user-level analysis with large datasets (Phillips et al., 2019). Three inputs are required when applying the DBSCAN algorithm on geolocated tweets: The minimum number of points required to form a cluster, the search radius (maximum distance from one point to the closest neighbour), and the geolocated tweets. DBSCAN was applied onto the Twitter data used in this thesis, to identify the presumed homeplace location of certain users (see *Subsection 4.1.2*).

Chapter 3 | Research Gaps and Objectives

This chapter points out the research gaps concerning mental health research based on Twitter data using sentiment analysis. Moreover, the general research question and the more specific research objectives pursued in this thesis are presented.

3.1 Research Gaps

As described in *Section 2.1*, Twitter data has been widely used on macro-level analysis aggregating single geolocated tweets onto large administrative units as states or counties in the United States (Eichstaedt et al., 2015; Jashinsky et al., 2014). Twitter has also been used on smaller spatial aggregation units as census tracts in New York City, also using single geolocated tweets (Edry et al., 2021). Furthermore, spatial analysis using geolocated tweets has been applied in the context of traumatic events, detecting spatial clusters of discomfort or negative emotions at specific locations (Gruebner et al., 2016; Gruebner et al., 2018).

However, what seems to remain mostly unstudied, is the potential impact of everyday physical and societal environmental characteristics on sentiments expressed in tweets at a neighbourhood scale. Also, most studies using Twitter data in a spatial context so far have been operating with single geolocated tweets, whilst user-level analysis has been less explored. Thus, ecological studies using Twitter data on user-level need more attention in mental health research and would potentially provide useful new insights concerning associations between environmental factors and tweeting behaviour.

3.2 Research Question and Objectives

Based on the research gaps, the aim was to investigate possible associations between physical and societal environmental factors and the sentiments of Twitter users, which they express in their tweets. Research has provided strong evidence of neighbourhood characteristics directly and indirectly affecting health behaviours and outcomes (see *Section 2.2*). Thus, it was decided to link neighbourhood characteristics of the presumed Twitter user's homeplace to the user's emotions- and stress-related tweets. For this purpose, the chosen approach was to conduct a regression analysis between the Twitter data as the outcome, and the environmental factors as the predictors.

Due to their high relevance found in the literature (see *Section 2.3*), the following three neighbourhood characteristics have been chosen, which form the main explanatory variables of the general regression model:

- Available Greenspace
- Exposure to Traffic Noise
- Socio-economic Circumstances

To account for bias, the following control variables are added:

- Gender
- Age
- Urban/Rural Differences

The answering of the following research question builds the primary aim of the thesis:

> *To which degree are physical and socio-economic phenomena in the neighbourhood of a Twitter user's homeplace, associated with the emotions found in the user's Tweets?*

The posed research question implies three research objectives (RO):

RO_1: Select Twitter users for the analysis, based on the following criteria:

- The homeplace location has been identified
- A substantial number of Tweets must be available
- The user is not a bot, or an account managed by an organisation
- The identified homeplace location is realistic (e.g., not on a glacier)

RO_2: Estimate the neighbourhood variables greenspace, traffic noise, and SEP for each Twitter user's homeplace.

- What is the quality and suitability of the generated variables?
- What are the limitations, and which potential improvements could be made?

RO_3: Perform a regression analysis between neighbourhood variables as predictors and emotions as outcome variables.

- Which significant associations can be found and what is their effect size?
- How does the replacement of traffic noise during daytime with traffic noise during night change the models?

In the next chapter, the required data for the accomplishment of the three research objectives is introduced.

Chapter 4 | Data

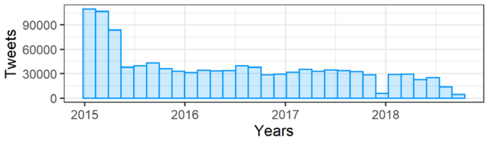
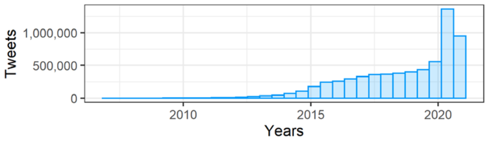
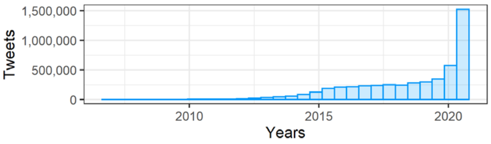
This chapter introduces the data used in this thesis. The chapter is divided into subsections, each of which describes one of the applied datasets in detail. The main datasets used, were on the one hand Twitter tweets posted from Swiss users, and on the other hand three high resolution spatial datasets to describe the physical and social characteristics of the neighbourhoods. The first one being available greenspace, the second being traffic noise, and the third being socio-economic position (SEP). The spatial extent of all data is defined by the area of Switzerland.

4.1 Twitter Data

At this point it is important to mention, that the Twitter data used in this thesis was provided by the health geography research group of the University of Zürich. The data was used before in other studies collaborating with the Loughborough University, and multiple processes enriched the data with further variables, which are described in *subsection 4.1.2*. All variables presented in that subsection were added to the data beforehand and were not part of the author's methodological processes.

The thesis is based on Twitter data which has its origin in the following three large Twitter datasets: The geolocated Twitter dataset, containing 1'115'893 tweets with a geolocation ($\text{tweets}_{\text{geoloc}}$), the rehydrated English Twitter dataset, containing 6'427'025 English tweets ($\text{tweets}_{\text{english}}$), and the rehydrated non-English Twitter dataset, containing 4'961'291 non-English tweets, which have been translated into English ($\text{tweets}_{\text{non-english}}$). In *Table 2*, the temporal extent of each of the three datasets is given. Certain tweets in the two rehydrated dataset $\text{tweets}_{\text{english}}$ and $\text{tweets}_{\text{non-english}}$ may date back until 2006, however no more than 7% are older than from 2015 in both datasets. The concept of rehydration is briefly explained in the next section.

Table 2: Temporal extent of the three Twitter datasets

Dataset	Oldest Tweet	Most Recent Tweet	Histogram of Tweets
$\text{tweets}_{\text{geoloc}}$	2015.01.01	2018.09.06	
$\text{tweets}_{\text{english}}$	2006.12.12	2020.09.22	
$\text{tweets}_{\text{non-english}}$	2007.01.09	2020.09.22	

4.1.1 Data Access

The Twitter data was accessed through the Twitter API on the Developer Platform, which provides tweets on request for research purposes. There are different access levels, where the level "Academic Research" grants access to retrieving up to 10 million Tweets per month, advanced search operators, and access to full-archive search. API rehydration is a standard process where specific requests are programmatically sent to the Twitter API, at which Twitter automatically responds by providing the specific requested tweets in the form of JSON code (Twitter Developer Platform, 2022). The two datasets $\text{tweets}_{\text{english}}$ and $\text{tweets}_{\text{non-english}}$ were rehydrated by using the user IDs of the dataset $\text{tweets}_{\text{geoloc}}$. In this way, up to 200 of the most recent tweets for each user were retrieved (date of rehydration: 2020.09.22), also including non-geolocated tweets.

4.1.2 Data Structure

The Twitter data was provided in a comma-separated, tabular format (.csv file), in which a single row represents one single tweet. When retrieved from the Twitter API, each tweet features more than 60 variables, including a unique identifier of the tweet, a unique identifier of this tweet's user, a username, the coordinates from where the tweet was posted (if geolocated), the date and time when the tweet was posted and finally, of course, the text of the tweet itself.

The most relevant variables are listed in *Table 3* and are based on a fictitious tweet to preserve user privacy. An example value for each variable is also shown to enable a better understanding of the different variables.

Table 3: Most relevant tweet attributes with example values of a fictitious tweet

Attribute	Example Value
id	935919123491283597123592349
user_id	32152139490172
user_name	Nicolas Schmidheiny
latitude	47.314178
longitude	8.466400
raw_geo	{u'type': u'Point', u'coordinates': [8.4664, 47.314178]}
created_at	2022-03-01 15:26:38
text	I am so happy!
lang	en
user_total_number_of_tweets_ever	27
profile_image_url_https	https://pbs.twimg.com/profile_images/<unique_code>.jpg

As mentioned earlier, many new variables have already been added to the datasets within previous studies, which are described in the following. Using the natural language processing algorithm EMOTIVE, introduced in *Subsection 2.7.1*, 8 new variables have been added to categorize the tweets into basic emotions. These include the following emotions: Anger, Confusion, Disgust, Fear, Happiness, Sadness, Shame and Surprise. The values are of the data type integer and range from 0 to 9, depending on the emotion, indicating the magnitude of the emotion found in the tweet. Additionally, Stresscapes, introduced in *Subsection 2.7.2* was applied onto the data to represent the stress-level found in the tweets by an integer value ranging from 0 to 20 (*stress.overall.score*). An overview of the added variables including example values for the text "I am so happy!" is given in *Table 4*. From now on, for convenience, these 9 variables are referred to as sentiments.

Table 4: The 9 sentiment-variables added through EMOTIVE and Stresscapes

Attribute	Example Value
anger	0
confusion	0
disgust	0
fear	0
happiness	6
sadness	0
shame	0
surprise	0
stress.overall.score	0

Moreover, the coordinates of the geolocated tweets have been used to conduct an activity spaces analysis using the clustering algorithm DBSCAN, introduced in *Section 2.8*. The tweets forming the detected spatio-temporal clusters have been categorized as either workplace (8 am to 5 pm) or homeplace (5 pm to 8 am) locations, depending on the time of the day. The two added variables *workplace* and *homeplace* (see *Table 5*) take either the value 1 (tweet was posted at workplace / homeplace) or the value 0 (tweet is not part of a cluster). The minimum number of geolocated tweets to build a cluster was set at 5. The search radius was set at 50 meters. However, in this study only the variable *homeplace* is of interest.

Table 5: Activity space variables added through the DBSCAN approach

Attribute	Example Value
homeplace	1
workplace	0

Furthermore, Botometer, introduced in *Subsection 2.5.2* was applied onto the data to control for bot-like activity of Twitter accounts. The two variables *cap_english* and *cap_universal* (see *Table 6*) give a score, taking continuous values from 0 to 1 to represent the chance from 0 to 100% if the tweet originates from an automated account. CAP stands for Complete Automation Probability. The variable *cap_english* is specific for English tweets and *cap_universal* is for all languages. In this thesis, only *cap_universal* was considered in the analysis and was later renamed to *is_bot* for convenience.

Table 6: Variables to represent the probability of the user being a bot

Attribute	Example Value
cap_english	0.124
cap_universal	0.113

Finally, the M3-Method introduced in *Section 2.6* was applied on a user level, to estimate further variables about user demographics and whether the account is managed by an organization or not. User profile information including the profile image are used to estimate the approximate age, represented by the four variables *age_less_than_inclusive_18*, *age_19_to_29*, *age_30_to_39* and *age_over_than_inclusive_40*, all of them taking values between 0 and 1 to represent the chance from 0 to 100% of the user falling into the corresponding age category. The sum of these four age category variables is always 1 = 100%. The two variables *gender_male* and *gender_female* also take values between 0 and 1 to represent the chance from 0 to 100% of the user having the respective gender and for every user the sum of these two variables is obviously 1 = 100%. The last M3-variable to be mentioned is named *is_org* and again takes values between 0 and 1 to represent the chance from 0 to 100% of the user actually being an account managed by an organization. For convenience, in the further course of the work the two variables *age_less_than_inclusive_18* and *age_over_than_inclusive_40* have been renamed to the shorter names *age_under_19* and *age_40_plus*. The M3-variables are shown in *Table 7*.

Table 7: M3-variables inferring user demographics and probability of being an organization

Attribute	Example Value
age_less_than_inclusive_18	0.0879
age_19_to_29	0.4690
age_30_to_39	0.2436
age_over_than_inclusive_40	0.1995
gender_male	0.9983
gender_female	0.0017
is_org	0.1027

4.2 Greenspace Data

The dataset used to estimate the quality and quantity of available greenspace within a neighbourhood, has its origin in a project that was conducted by the author and colleagues within the frame of the master's course "GEO888 – GIS for Environmental Modelling" at the University of Zürich. In the following, the undertaken approaches to generate the dataset are briefly explained.

As described in *Subsection 2.3.1*, the availability of greenspace is crucial for mental well-being. Hence, it must be defined, what availability actually means. Availability of greenspace should include the quantity, the quality, and the accessibility of green areas, where people can recreate in their leisure time. The author and colleagues combined the latter two factors into one single variable by creating an index for the extent of Switzerland, which takes multiple different spatial criteria into account. The calculation of the index is based on a multi-criteria analysis (MCA) approach, where various spatial input datasets in the form of vector or raster layers have been processed and multiplied with one another, resulting in a single greenspace raster dataset, covering the entire area of Switzerland with a spatial resolution of 50 meters.

The data to compute the index were provided by the Federal Office of Topography (Swisstopo), the Federal Office for the Environment (BAFU), and Landsat satellite imagery from NASA. The following open-source datasets were mainly used: The digital elevation model *SwissALTI3D* (Swisstopo, 2018), Landsat data (Google Earth Engine, 2021) to calculate the normalized difference vegetation index (NDVI), a buildings dataset *swissTLMRegio_Building* and a landcover dataset *swissTLMRegio_LandCover* (Swisstopo 2021), road and railway noise data (Bundesamt für Umwelt, 2018). All input datasets were converted to raster data and rescaled to a cell size of 50 meters.

To determine the spatial coverage and the quality of greenspace, the NDVI, landcover types, road and railway noise, as well as the distance to water bodies were used. Areas with a NDVI lower than a certain threshold were considered as non-green areas and were therefore excluded. Landcover types such as forests and dry grasslands were given additional weight due to their recreational value and their ecological importance. The same applies to areas belonging to the Federal inventory of landscape and natural monument (BLN). Areas affected by road and railway noise above 65 dB were excluded, and areas affected by only low or even no noise pollution were given additional weight. Green areas next to lakes and rivers until a maximum distance of 200 meters were also given additional weight, since water bodies, also referred to as *blue spaces*, are seen as having a positive impact on the recreational value of greenspace (Wheeler et al., 2015).

For the examination of the accessibility two factors were considered: Firstly, the slope of the terrain was used to analyse whether certain green areas are too steep for people to comfortably walk on them. A steepness of 30 degrees was used as a threshold to exclude too steep and therefore inaccessible areas. Additionally, to account for elder people and the fact that less steep terrain is considered more convenient for recreational activities, more weight was given to less steep or even flat terrain. Secondly, areas further away than 1000 meters from the next building were excluded, as they were considered as outside of walking distance from the closest residents. Furthermore, areas closer to buildings were given more weight. A visualization of the dataset can be seen in *Figure 2*.

A more detailed description of the methodological approaches and used data to develop the greenspace index can be found in the *Appendix A.1*.

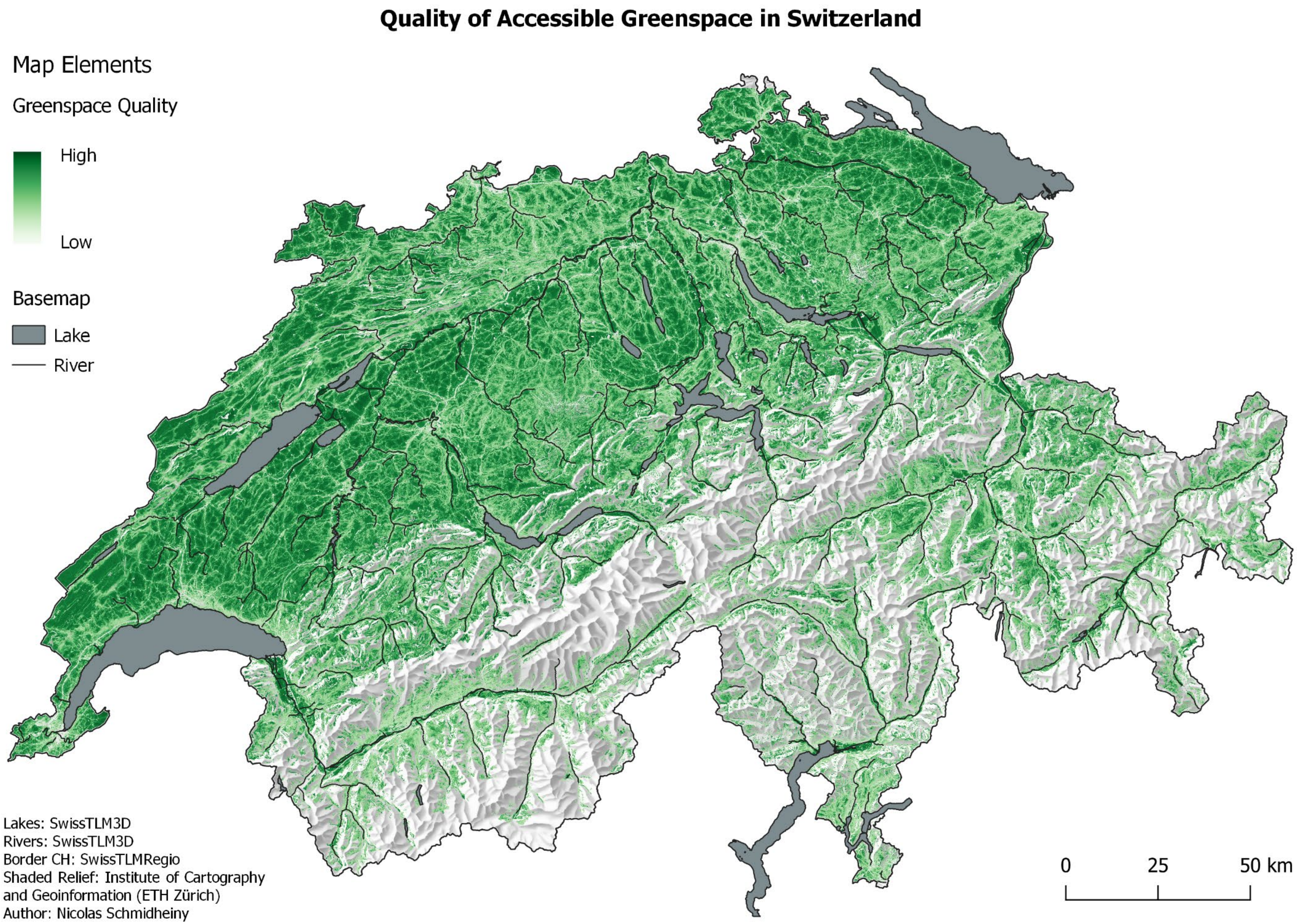

Quality of Accessible Greenspace in Switzerland
Map Elements
Greenspace Quality
High
Low
Basemap
Lake
River
Lakes: SwissTLM3D
Rivers: SwissTLM3D
Border CH: SwissTLMRegio
Shaded Relief: Institute of Cartography and Geoinformation (ETH Zürich)
Author: Nicolas Schmidheiny
0
25
50 km

Figure 2: Visualization of the greenspace dataset showing greener and less green areas in Switzerland, indicating the quality of greenspace.

4.3 Traffic Noise Data

As emphasised in *Subsection 2.3.2*, exposure to excessive and chronic noise from traffic is one of the most important environmental problems for physical and mental health. To estimate the exposure to traffic noise within a neighbourhood, four high resolution datasets provided by the BAFU were used. Road noise during daytime, which is defined from 6 a.m. to 10 p.m., and road noise during night, which is defined from 10 p.m. to 6 p.m., as well as railway noise during daytime and during night. The datasets are freely available as raster data in the GeoTIFF format and are the product of a noise monitoring system and database called SonBase which was developed by BAFU in 2009 and was updated in 2015 (BAFU, 2018). The cell size of these nationwide raster datasets is 10 meters. For the calculations of the road noise, 68'000 kilometres of the entire Swiss road network were used, and for the railway noise, roughly 4'000 kilometres of the railway network were considered. For the purpose of the analysis of this thesis, the two daytime noise datasets were combined to one single dataset, representing traffic noise during the day, and the same also applies for the two night-noise datasets. This pre-processing step is explained in more detail in *Section 5.2.2* of the next chapter. The daytime traffic noise dataset is illustrated in *Figure 3*.

4.4 Socio-economic Position Data

The data for socio-economic position (SEP) used in this work was kindly provided by the Institute of Social- and Prevention-Medicine (Institut für Sozial- und Präventivmedizin) of the University of Bern.

Panczak et al. (2012) developed an index for socio-economic position in Switzerland, at a neighbourhood scale called Swiss-SEP. They used Census 2000 data of all 2.95 million residential households in Switzerland, which were spatially referenced by using the geographic coordinates of 1.27 million residential buildings. The 1.27 million overlapping neighbourhood areas were defined based on a road network connectivity approach. Each Neighbourhood is centred on a residential building and consists of about 50 of the nearest households. (Panczak et al. 2012) The Swiss-SEP was conceptualised as a combination of the following four domains: income, education, occupation, and housing conditions. Each of them was represented by one variable, which was calculated with data aggregated on a neighbourhood level. Median rent in Swiss Francs per square meter of the 50 nearest rented flats was used to approximate household income. For the education variable, the ratio of households inhabited by a person with primary education or less was used. Occupation was similarly represented by the ratio of households headed by a person in a manual or unskilled profession. Finally, for the housing conditions variable, the mean number of inhabitants per room was used.

The data was provided as CSV file with coordinates of the neighbourhood centroids. It is updated regularly (last update: June 2021) and now contains 1'527'173 observations. *Figure 4* shows a visualization of the dataset, where the single neighbourhoods are shown as spatial point data.

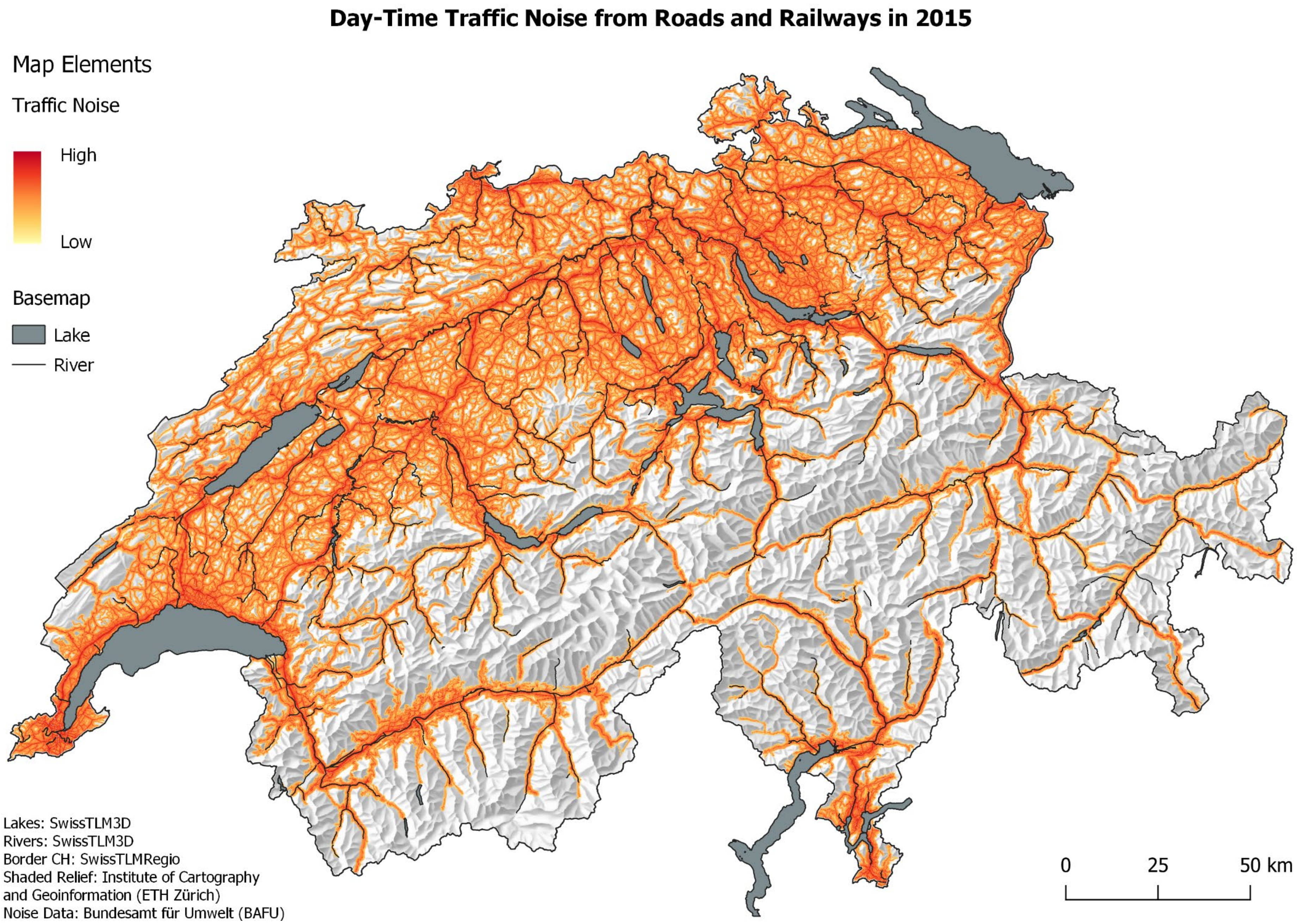

Day-Time Traffic Noise from Roads and Railways in 2015
Map Elements
Traffic Noise
High
Low
Basemap
Lake
River
Lakes: SwissTLM3D
Rivers: SwissTLM3D
Border CH: SwissTLMRegio
Shaded Relief: Institute of Cartography and Geoinformation (ETH Zürich)
Noise Data: Bundesamt für Umwelt (BAFU)
0
25
50 km

Figure 3: Visualization of daytime traffic noise in Switzerland.

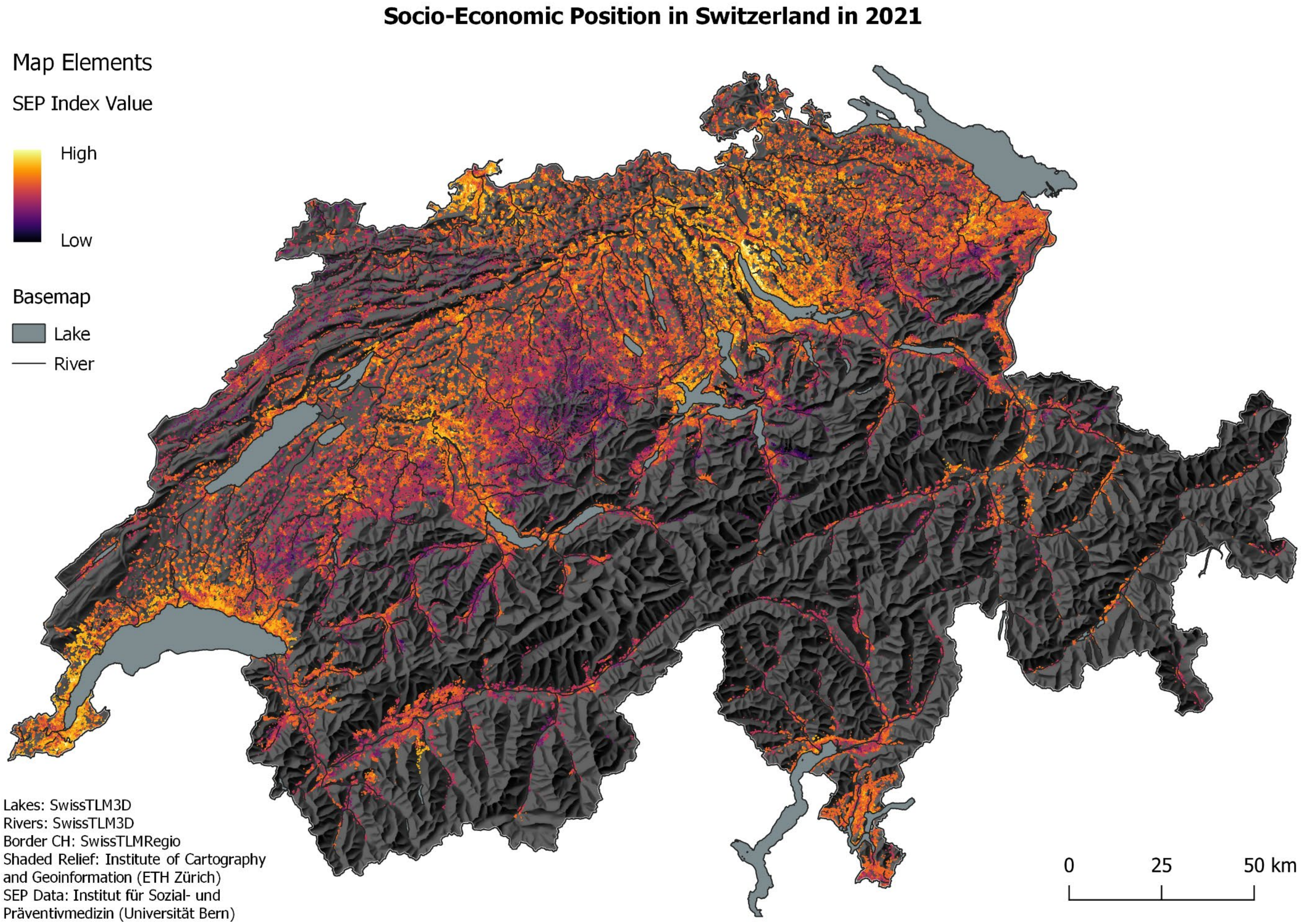

Socio-Economic Position in Switzerland in 2021
Map Elements
SEP Index Value
High
Low
Basemap
Lake
River
0
25
50 km
Lakes: SwissTLM3D
Rivers: SwissTLM3D
Border CH: SwissTLMRegio
Shaded Relief: Institute of Cartography and Geoinformation (ETH Zürich)
SEP Data: Institut für Sozial- und Präventivmedizin (Universität Bern)

Figure 4: Visualization of the Swiss-SEP in Switzerland.

4.5 Urban/Rural Typology Data

To account for urban-rural differences in the definition of greenspace, and potentially also for the tweeting behaviour of users, urban/rural typology data on a communal aggregation level from BFS was used (Bundesamt für Statistik, 2022). The freely available dataset published in 2012, divides the Swiss communes into the three categories urban *(Städtisch)*, intermediate *(Intermediär)*, and rural *(Ländlich)*. The categorisation is derived from a further subdivided classification, where the communes are classified depending on population density, total population, and accessibility criteria. However, for the scope of this thesis, a division of the data into the three mentioned classes is sufficient. a more fine-grained classification is not of interest because it would result in fewer data points in each class and hence decrease statistical significance. The data is available as CSV file. *Figure 5* shows a choropleth map with the Swiss communes coloured depending on their topology category.

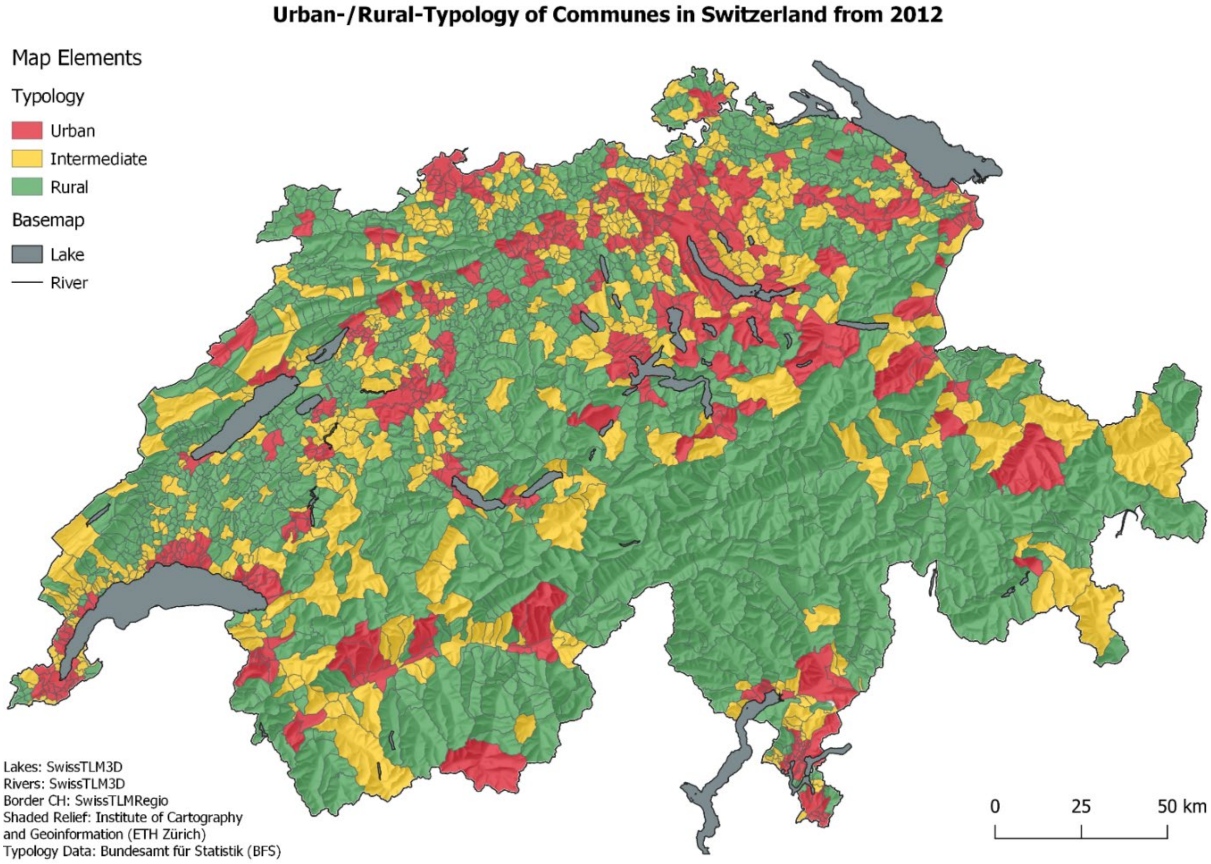

Urban-/Rural-Typology of Communes in Switzerland from 2012
Map Elements
Typology
Urban
Intermediate
Rural
Basemap
Lake
River
Lakes: SwissTLM3D
Rivers: SwissTLM3D
Border CH: SwissTLMRegio
Shaded Relief: Institute of Cartography and Geoinformation (ETH Zürich)
Typology Data: Bundesamt für Statistik (BFS)
0
25
50 km

Figure 5: Choropleth map showing urban-/rural-typology of the Swiss communes.

Chapter 5 | Methodology

In this chapter, the methodological processes applied in the thesis are elaborated. The overall procedure is basically structured in four main steps: Pre-processing of the input data (*Section 5.2*), the selection of users based on specific criteria (*Section 5.3*), calculating the neighbourhood variables (*Section 5.4*), and finally the regression analysis (*Section 5.5*). These four main processes each contain further sub-processes which are described in detail in the corresponding subsections.

5.1 Software and Scripting

The exploration, processing, analysis, and visualisation of the data was done using different software and scripting languages including R, Python, ArcGIS Pro and QGIS. R was the mainly used software. It was used for Twitter data exploration and processing, merging different datasets, as well as statistical analysis of the final dataset and visualization of the results. ArcGIS Pro was used to process the data of the neighbourhood variables and the modelling of the distance decay workflow. Python was used for time efficient parallel execution of ArcGIS processing tools. Finally, QGIS was used for the visual exploration of geolocated tweets and Twitter user homeplace locations, due to its high rendering and drawing performance of geodata, and for the exporting of the maps in high resolution.

5.2 Data Processing

The following subsections describe the processing steps carried out to prepare the data used in the regression analysis. First, the processing of the Twitter datasets is explained in detail. Then, the processing of the traffic noise datasets, and the SEP dataset is described, and finally, the processing steps for the urban-/rural-typology dataset are explained. The greenspace dataset introduced in *Section 4.2* did not need further changes, which is why it is not mentioned in this Section.

5.2.1 Twitter Datasets

As described in *Section 4.1*, the Twitter data analysed in this thesis was provided in three large datasets. Each of them needed separate pre-processing before merging them to one single Twitter dataset. As there were multiple processing steps, the different steps are

Clipping to AOI

Starting with tweets$_{geoloc}$, featuring 106 variables, of which the greater part is not of interest for the purpose of the analysis, only the relevant ones were selected in order to significantly reduce file size. This step decreased memory usage and enabled shorter computing time for all following processing steps. The retained variables are exactly those described in

Subsection 4.1.2, but without the M3-variables shown in *Table 7*, because they were not relevant for the processing steps executed on this dataset. After exporting the dataset as shapefile (.shp) using the R function *st_write()*, a first explorative visualization of the spatial distribution of the geolocated tweets was carried out in QGIS. It instantly revealed that some data points were lying outside the border of Switzerland. Therefore, the dataset had to be clipped onto the area of Switzerland. As the coordinates of Twitter data are provided in the standard *WGS84* projection, they first had to be converted to the Swiss *LV95* projection, to match the projection of all the other spatial datasets using the R function *spTransform()*. Of the previous 1’115’893 data points, only 2207 were excluded after clipping to the area of interest using the R function *st_intersection()*, which is only about 0.2%. The spatial distribution of the tweets in tweets$_{\text{geoloc}}$ within Switzerland are visualized in a map in *Figure 7*.

Grouping by User

Because the analysis is carried out on a user level and not on the single tweet level, the data had to be grouped by user. The emotion variables and the stress variable can take integer values from 0 to around 20, depending on the variable, to reflect the magnitude of the tweet’s sentiment. This certainly offers an information gain for the single tweet, but when grouping the tweets to the user level, complications arise. Summing up 10 tweets, each having the value 1 for the emotion *Sadness* is not the same as having only 1 tweet with the value 10 for the same emotion. Because this problem is not trivial to solve, and there is a lot of room for subjective decisions on how to capture sentiment magnitude, it was decided to create 9 new variables, 1 for each emotion and 1 for stress, which simply take the values 0 and 1. 0 if the respective sentiment is not present in the tweet, and 1 for all the other possible values, that is, if the sentiment is greater than 0. When looking at the number of occurring sentiments in the entire tweets$_{\text{geoloc}}$ dataset, shown in *Figure 6*, it is evident that *Happiness* is by far the most frequent sentiment, followed by *Stress* and *Sadness*. All the other sentiments tend to be expressed less often by Twitter users, when posting a tweet. Because the ratio of tweets expressing any sentiment in general is quite low, that is around 6 to 7%, it was decided to combine the 6 negatively afflicted emotions (*Sadness, Anger, Fear, Confusion, Disgust*, and *Shame*) into a new variable *emo_neg_01*, to enhance statistical power. This variable takes the value 1 if at least 1 of the 6 emotions is present, and 0 otherwise. The same was done including all 8 emotions into a variable *emo_01*, to assess if the tweet shows any emotional content.

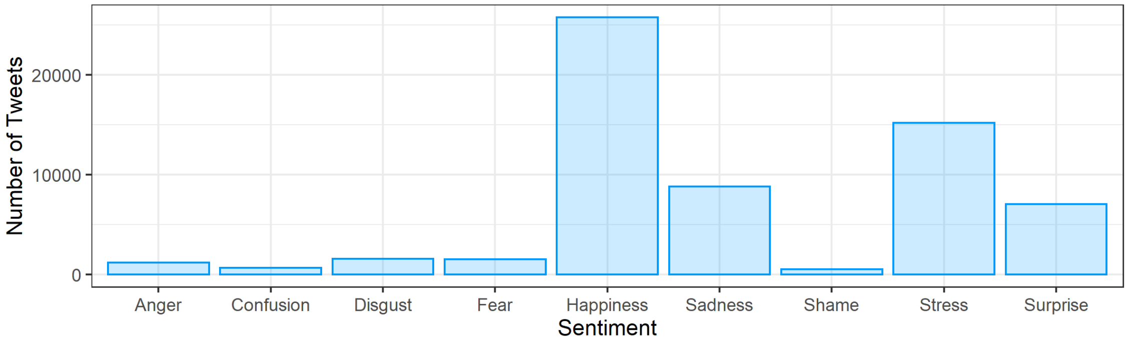


Figure 6: Barplot showing number of tweets expressing a sentiment in the dataset tweets$_{\text{geoloc}}$

Next, the data was prepared to be grouped by user, again creating a new variable for every sentiment variable counting the number of tweets showing the respective sentiment. Furthermore, since the coordinates of the single tweets lose their meaning when grouping the data, the latitude and longitude variables of certain users' detected homeplace locations had to be extracted and added to the corresponding users.

Merging the Three Datasets

After grouping the tweets of the tweets$_{english}$ and tweets$_{non-english}$ datasets in the exact same manner as for the tweets$_{geoloc}$ dataset, the datasets were merged by *user_id* using the R function *merge()*. The merged dataset (users$_{all}$) features 70'333 users. After merging, the counts of every variable for each user had to be summed up. By dividing the summed-up variables through the number of total tweets for each user, the rate of the different sentiment occurrences was calculated and saved as new variables (see *Table 8*). The Example values in the table are real values of an anonymous user. For example, a value of 0.00885 for the variable *rat_anger* means that in 0.885 % of this user's tweets the emotion Anger was found.

Table 8: New variables showing the rate of tweets of a user expressing this sentiment

Attribute	Example Value
rate_anger	0.00885
rate_confu	0.00000
rate_disgu	0.00442
rate_fear	0.00442
rate_happy	0.00885
rate_sad	0.02212
rate_shame	0.00000
rate_surpr	0.01327
rate_emo	0.05310
rate_neg	0.03540
rate_stress	0.07522

Having aggregated the data onto the user level, user specific variables from Botometer and the M3-method were extracted from the tweets$_{geoloc}$ dataset and added to the users$_{all}$ dataset, again merging by the *user_id*.

Filtering Users with Detected Homeplace

Out of the 70'333 users in the dataset users$_{all}$, for 1213 users a homeplace was detected by the DBSCAN clustering algorithm applied in a previous study. These users were selected and saved as a new dataset users$_{home}$. The variables *longitude* and *latitude* were used to create point geometries with the R function *st_as_sf()* and the dataset was exported as shapefile using the R function *st_write()*.

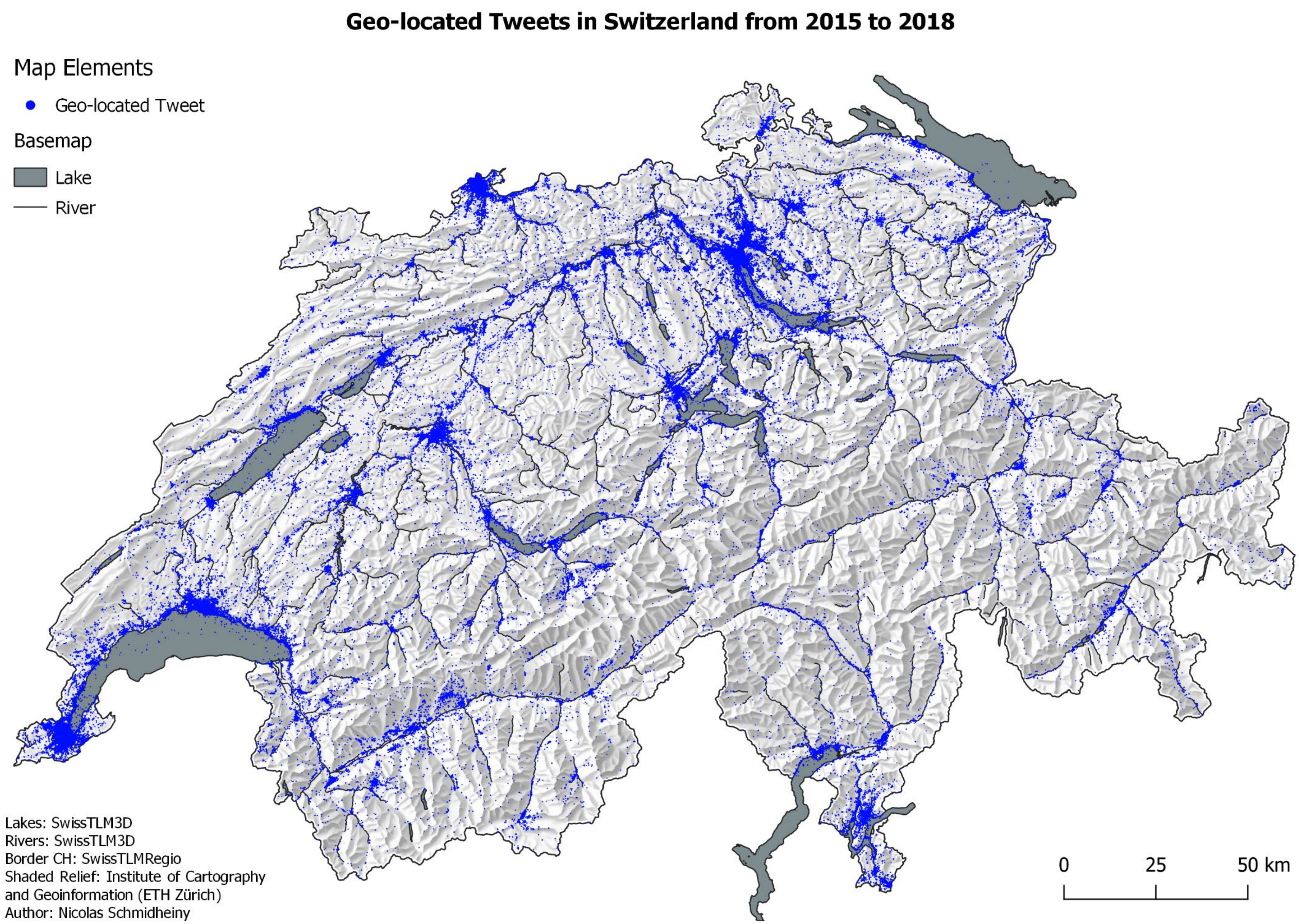

Geo-located Tweets in Switzerland from 2015 to 2018
Map Elements
● Geo-located Tweet
Basemap
▨ Lake
— River
Lakes: SwissTLM3D
Rivers: SwissTLM3D
Border CH: SwissTLMRegio
Shaded Relief: Institute of Cartography
and Geoinformation (ETH Zürich)
Author: Nicolas Schmidheiny
0
25
50 km

Figure 7: Visualization of the spatial distribution of geolocated tweets in Switzerland between 2015 and 2018.

Preparing Data for LIWC

The idea to apply LIWC (see *Subsection 2.7.3*) onto the Twitter data only came to mind during an already advanced stage of the analysis. Therefore, in contrast to the other algorithms applied on the Twitter data used in this thesis, LIWC was not yet applied. For this reason, the datasets first had to be fed to the LIWC tool. Before feeding the Twitter data into LIWC, it had to be rearranged in a specific way. From the three datasets $\text{tweets}_{\text{geoloc}}$, $\text{tweets}_{\text{english}}$, and $\text{tweets}_{\text{non-english}}$, only the variables *user_id*, *id*, and *text* were selected. Afterwards, the three datasets were combined to one large dataset, containing all 12’504’209 single tweet texts with their corresponding *id* and *user_id*, using the R function *rbind()*. Given that tweet $\text{tweets}_{\text{english}}$ and $\text{tweets}_{\text{non-english}}$ were rehydrated from $\text{tweets}_{\text{geoloc}}$, they may contain some number of tweets from the latter one. These potential duplicates were removed by using the R function *duplicated()*. After that, the variable *id* was also removed. As LIWC is designed for large text sizes and thus only provides usable results for texts consisting of at least a few hundred words, all tweet texts have been concatenated to one single text for each user using the R function *paste0()*. In this way, instead of having certain texts containing only a single word (e.g., “bored”), on average the texts to be analysed by LIWC now featured about 500 words. After analysing the data with LIWC, the output featured dozens of statistical variables like word count, a performance value etc, of which only the variables *posemo* and *negemo* were of interest for the scope of this thesis. These two variables were then added to the $\text{users}_{\text{all}}$ dataset *(*see *Table 9).* A value of 2.60 for *posemo* means that 2.6 % of the words in all the tweets of a user were considered as being emotionally positive, while a value of 1.09 for *negemo* means that 1.09 % of the words were considered as being emotionally negative.

Table 9: LIWC variables posemo (positive emotions) and negemo (negative emotions)

Attribute	Example Value
posemo	2.60
negemo	1.09

5.2.2 Traffic Noise Datasets

As described in *Section 4.3*, four high resolution raster datasets were used to estimate traffic noise in the neighbourhoods. Because the difference between the perception of road noise and railway noise is irrelevant for the purpose of the analysis, the datasets were combined to a general traffic noise dataset. However, the difference in noise between daytime and night was of interest, hence it was decided not to combine the daytime with the night datasets. First, the two datasets for daytime noise were combined. The same procedure was then applied for the night noise datasets.

Using the geoprocessing tool *Cell Statistics* in ArcGIS, the layers were combined by using *Maximum* as the overlay statistic. The tool compares the values of both layers for each cell and picks the higher noise value. This simple approach was chosen since noise does not stack up linearly and because the highest noise value was deemed to be relevant. In a next step, using the *Resample* tool in ArcGIS, the combined raster dataset was resampled from a 10 meters resolution to a 50 meters resolution for three reasons: A cell size of 10 meters is a unnecessarily high resolution and would most probably not improve the quality of the analysis.

The resampling from 10 m to 50 m drastically reduces file size which consequentially reduces computing time for subsequent processing. Since the greenspace dataset introduced in *Section 4.2* has a 50 m resolution, the resolutions of both datasets were matched, and the raster grids were made congruent by snapping them together. Because the emitted traffic noise follows the road and railway networks and only expands a few hundred meters from the network axis, large areas remain unaffected from traffic noise, especially in the alpine regions. Cells covering these areas have the value *NoData* and had to be converted to 0, using the tool *Raster Calculator* in ArcGIS to be included in the following calculations. Consequentially, the value of 0 was also assigned to the cells outside the border of Switzerland, which had to be undone again by clipping the dataset onto the area of Switzerland using the ArcGIS tool *Clip*.

5.2.3 Socio-economic Position Dataset

As already mentioned in *Section 4.4*, the dataset for socio-economic position was provided as a large CSV file featuring 1’527’173 observations and 17 variables. The 17 variables include X- and Y-coordinates in the LV03 projection, which had to be transformed into LV95 using the R function *spTransform()*. Then, the point geometries were created using the R function *st_as_sf()* and the dataset was exported as shapefile using the R function *st_write().* In a next step, the dataset was loaded into ArcGIS Pro in order to convert the shapefile into a raster file with a cell size of 50 meters using the tool *Point to Raster*. As value field, the SEP index value (variable *ssep3*) was chosen and as cell assignment type *Mean* was chosen. The grid was snapped to the greenspace raster, making it congruent to the other two neighbourhood variable datasets.

5.2.4 Urban-/Rural-Typology Dataset

The urban-/rural-typology dataset from the year 2012 provided by BFS is available as CSV file, as already described in *Section 4.5*. It features 2255 rows, one for each commune, and only the three variables *ID* (official BFS number of the commune), *Name* (name of the commune), and *Kategorien* (typology category). In Switzerland, certain communes tend to merge from year to year, typically for administrative or political reasons. Usually, rather remote communes with a low population unite to one single commune. According to BFS on January 1st 2022, the number of existing communes was 2148.

Since the dataset does not include any variable containing the geometry of the commune boundary, it had to be added in a further step. Because the author could not find a dataset containing the geometries of the 2012 communes, it was decided to take a more up-to-date shapefile from the year 2021. After joining the two datasets in ArcGIS Pro by *ID* and BFS number respectively, some polygon features did not take any typology category, because certain BFS numbers do not exist any longer. The typology category values were then manually added by visually comparing the dataset with the interactive web map displayed on the Statatlas website of BFS (Bundesamt für Statistik, 2022).

5.3 Selection Criteria for Users

Having already selected the users with a presumed homeplace location and created a dataset $users_{home}$ containing these, the remaining steps regarding *RO_1* (see *Section 3.2*) were carried out next and are described in the following.

Assuming that the sentiments of Twitter users expressed in their tweets depend on the neighbourhood characteristics to a certain degree, three criteria must be fulfilled to obtain a meaningful result in the regression analysis: First, The Twitter user must be a real person, rather than an automated account (bot) or an account managed by an organisation. Second, the presumed homeplace location of the user must be realistic, meaning that a location on a glacier, on a lake or in a sports stadium, for obvious reasons, most likely is not the user's real homeplace. Third, the number of tweets of a user must have a substantial size to ensure statistical power (Biau et al., 2008).

The first criterion was met by excluding users which exceed a defined threshold for the Botometer variable *cap_universal* and the M3-variable *is_org* (introduced in *Subsection 4.1.2*). As a recap: Both variables hold continuous values between 0 and 1, reflecting the probability of the user being an automated account or an organization respectively. When choosing the thresholds, two things had to be considered: On the one hand, it was of interest to minimize false positives (Banerjee et al., 2009) by setting the threshold value as low as possible. On the other hand, keeping the sample size as large as possible was crucial to ensure statistical power, and setting the threshold too low would result in excluding too many users. Having this in mind and carefully inspecting the distributions, a Threshold value of 0.5 was defined for the bot variable, excluding 17 users with a value above. As for the organisations variable *is_org*, a threshold of 0.9 was chosen, excluding 224 users with a value above.

For the second criterion, it was first necessary to approximate the spatial coverage of residential areas in Switzerland. To simply use a dataset containing all building footprints would be very unprecise, as uninhabited buildings and facilities such as industrial areas, recreational areas, sports facilities, and event halls should not be included. Instead, the SEP data (see *Section 4.4*) was used as a point feature dataset, containing very precise locations of roughly 1.5 million Swiss households. Using the ArcGIS tool *Buffer*, a buffer of 200 meters around the point locations was created, to represent the residential areas in Switzerland. Although 200 meters may be considered as quite generous, it seemed reasonable to choose a radius of this size for the following reasons: The point locations of the households tend to be centred in the middle of the building footprints, and certain buildings may have a relatively large extent. Additionally, the uncertainty of the Twitter user's homeplace location is not exactly known, and by choosing a too small buffer radius, there would be a risk to randomly exclude certain users.

To verify the plausibility of the above-described approach, a second approach was carried out. A dataset of the statistics of population and households (STATPOP) of the year 2020, provided by the BFS was used. It is freely available as a CSV file and was first converted into a shapefile. The dataset divides the area of Switzerland into hectare-sized cells, each cell containing statistics about demographics like total population (which is equal to population density, given that every cell has the same area of one hectare), and population per age category. These precise statistics are irrelevant in this matter, however, all cells having a population greater than 0 show that the respective area is inhabited. Because the spatial division of the country through the grid lines is entirely arbitrary, it may often occur that a building is split into two cells, or that a building is located directly at the border of a cell. For this reason, a buffer of 100 meters was computed around the cells, extending the spatial coverage of residential area in such a way, that slightly shifted homeplace locations of certain Twitter users are still included. This approach resulted in a very similar number of realistic homeplace locations, namely 1020 of the 1213 users in the users$_{home}$ dataset, versus 1023 in the above approach using the SEP data. Hence, 193 users showed unrealistic homeplace locations and were therefore excluded (see *Figure 8*).

As for the third criterion, being the minimum number of tweets of a user, the threshold was set at 100 tweets after inspecting the histogram. This seemed to be a substantial number of tweets to be analysed, whilst not excluding too many users. Applying this threshold, 85 users were excluded.

As these criteria overlap, an excluded user may belong to more than one exclusion criterion. In addition to these three criteria, some more users had to be removed from the dataset, due to missing values for the M3-variables. In summary, of the 1213 users 733 fulfil all criteria and build the final dataset users$_{regression}$ to be analysed in the regression models.

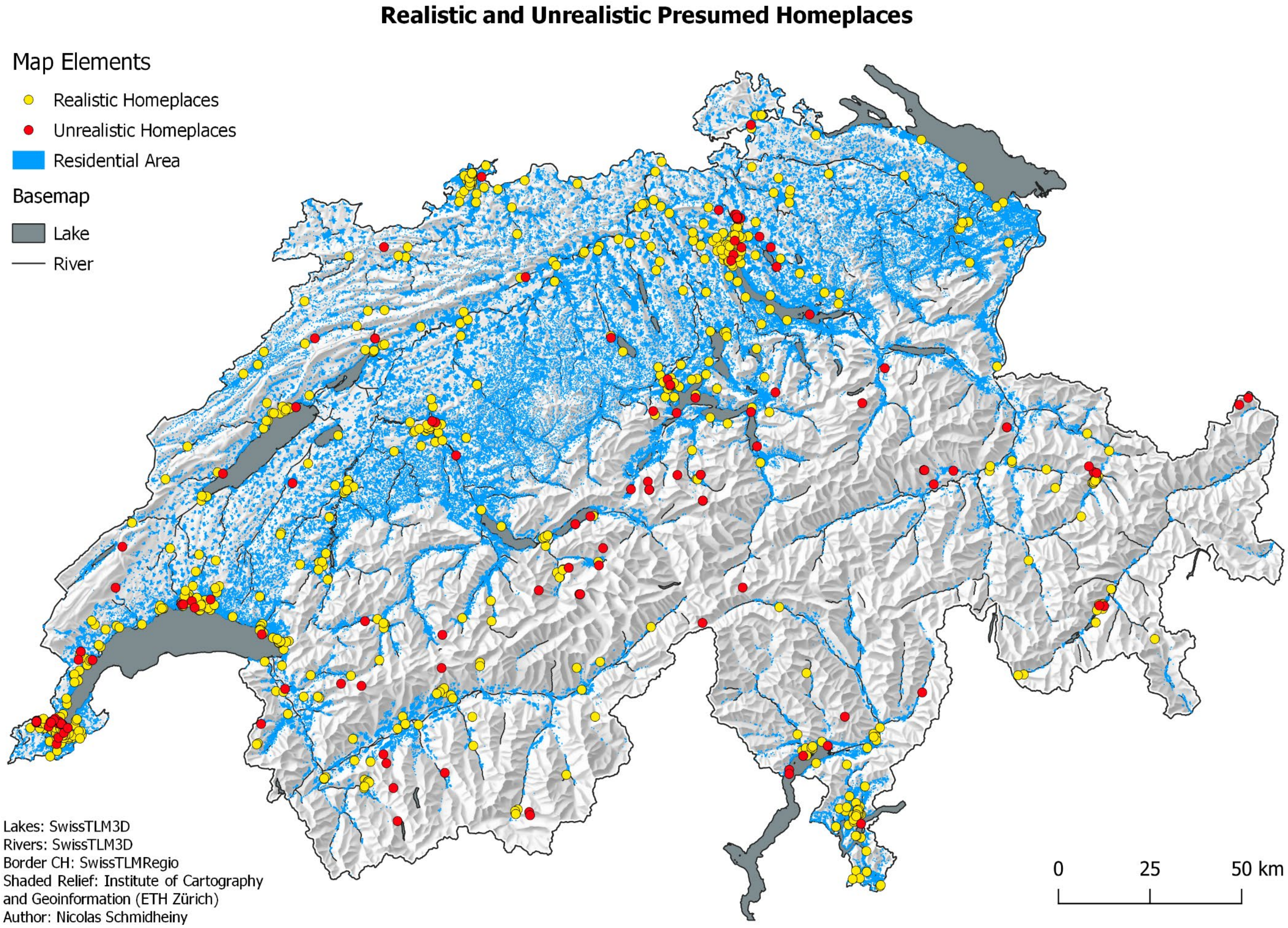


Figure 8: Visualization of realistic and unrealistic presumed homeplace locations of users in the users$_{home}$ dataset.

5.4 Estimating Neighbourhood Variables

Having extracted the homeplace locations of users with a presumed homeplace, as a next step, the neighbourhood areas had to be defined. Because no assumptions about the movement pattern of the users can be made, the area around their homeplace where they potentially spend their time must be equal among all users analysed. Perhaps the most simplistic and trivial way to represent the neighbourhood around a point location, is to draw a circle with a specific radius, as seen in other studies (see *Section 2.2*). Choosing the same radius for every point location, instead of defining neighbourhoods individually based on e.g., road networks, ensures that the different neighbourhoods have a constant area.

As mentioned in *Section 2.2*, in ecological studies it is common practice to define the “walking neighbourhood” as the area within a 1 kilometre or 1 mile distance around the homeplace location (Smith et al., 2010). Assuming an average walking speed of 5 kilometres per hour (Murtagh et al., 2021), a walking distance of 5 minutes would be equal to a metric distance of about 417 meters, and 834 meters for a 10-minute walking distance respectively. To choose a bit more convenient values, a radius of 500 meters and 1000 meters seemed adequate. It was decided to estimate the neighbourhood variables with both radius sizes, since both seemed equally plausible.

After having defined the spatial extent of a neighbourhood, a further assumption had to be made concerning the time spent in different locations within the neighbourhood. Since this analysis tries to quantify the impact of the physical and social environment on a user’s tweeting behaviour, the closeness and the time exposed to a certain spatial phenomenon is crucial. For example, being exposed to the noise of a construction site for half an hour a day, will not affect the stress level of a person as much, as being exposed to the same noise throughout the entire day. Likewise, being exposed to that noise from a distance of, for example, 200 meters will not affect the stress level as much, as being located directly next to the construction site.

Obviously, the time spent at different locations is not equally distributed within the neighbourhood area. It is trivial to assume that on average, a person spends more time at the homeplace location than in any other random location within the neighbourhood. The resident will most probably also spend more time directly around the homeplace location, say within a range of 50 meters, for example to dispose of the green waste and the garbage bags, than at any location 500 meters away. Therefore, it is evident, that some form of weighting is needed to more accurately represent the impact of social and physical characteristics in a neighbourhood on the sentiments of a Twitter user, expressed in the user’s tweets.Since a weighting tailored to each individual neighbourhood would by far be too complex to implement, some form of Inverse Distance Weighting (IDW) seemed reasonable. This would give the neighbourhood variables the most weight exactly at the homeplace location, and the least weight at the perimeter of the neighbourhood area. This effect is also known as Distance Decay (Yin et al., 2019). How this IDW was implemented, is elaborated in *Subsection 5.4.1*.

The neighbourhood variable *traffic noise* was estimated once for daytime noise *(noise_day)*, using the aforementioned buffer radii, and once for noise during the night *(noise_night)*, using a buffer of 100 meters. The choice for the much smaller radius for night noise was made with the consideration, that the person is sleeping during the night and therefore not mobile. Nevertheless, instead of simply taking the traffic noise at the exact homeplace location as a point sample seemed too prone to error, because of the uncertainty of the estimated homeplace location’s accuracy. Moreover, no distance decay was applied to calculate *noise_night* due to the assumption of the person having a fixed location during the time period assigned to this variable, being from 10 pm to 6 am.

5.4.1 Modelling Distance Decay

There are different functions to model distance decay, including exponential, logistic, gaussian, and linear, of which all would arguably be suitable in this case (Bauer & Groneberg, 2016; de Vries et al., 2009). Because the implementation of the exponential, logistic, and the gaussian distance decay is considerably more complex, whilst presumably not gaining much better fitting results, it was decided to choose a linear approach. In the following, the implementation of the distance decay approach using ArcGIS Pro and Python is described in detail. Note that it was carried out for all 1213 users of the $\text{users}_{\text{all}}$ dataset, and not only for the 733 users of the $\text{users}_{\text{regression}}$ dataset. This enabled the possibility to later include more than 733 users from $\text{users}_{\text{home}}$, in case the author should have retrospectively adjusted the exclusion criteria and thresholds.

Workflow ArcGIS Pro

First, buffers of 500 meters and 1000 meters were created around the 1213 homeplace locations, using the tool *Buffer* in ArcGIS, generating the dataset *Neighbourhoods*. Next, the tool *Split Raster* was used once for each of the three neighbourhood variable raster datasets as input raster, once with the 500 meters buffer layer and once with the 1000 meters buffer layer as the split method. The resulting 1213 rasters (*Neighbourhood Variable Raster*) were stored in the same folder, that is, for each combination (e.g. greenspace – 500 m) one separate folder. These rasters would now each contain all the cells within the defined neighbourhoods, each cell holding a value for one of the three examined neighbourhood characteristics greenspace, traffic noise, and socio-economic position. In a next step, these cell values needed to be weighted according to the chosen linear IDW. For the weighting process, a model was created in the ArcGIS Modelbuilder, which sequentially performs a number of calculations for each of the 1213 neighbourhood rasters as input data. The processing steps designed in the model, shown in *Figure 9*, were applied on all three neighbourhood variables for both 500 meters and 1000 meters radius.

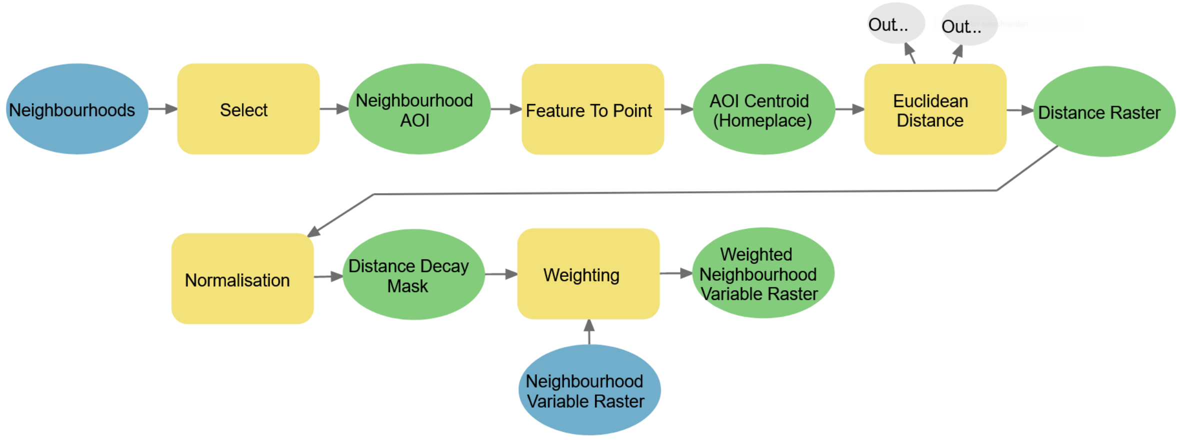


Figure 9: *Workflow in the ArcGIS Pro Modelbuilder to implement distance decay*. The blue elements are the inputs, the yellow elements are the processing tools, the green elements are the interim processing outputs. The final output is the Weighted Neighbourhood Variable Raster.

In the following, each processing step is described in detail and the interim processing outputs are illustrated with figures.

Step 1 - Select: The dataset *Neighbourhoods* was used as input data to select the neighbourhood area of a specific user via unique FID as the Area of Interest (AOI) for all the following steps and was saved as output *Neighbourhood AOI* (see *Figure 10*).

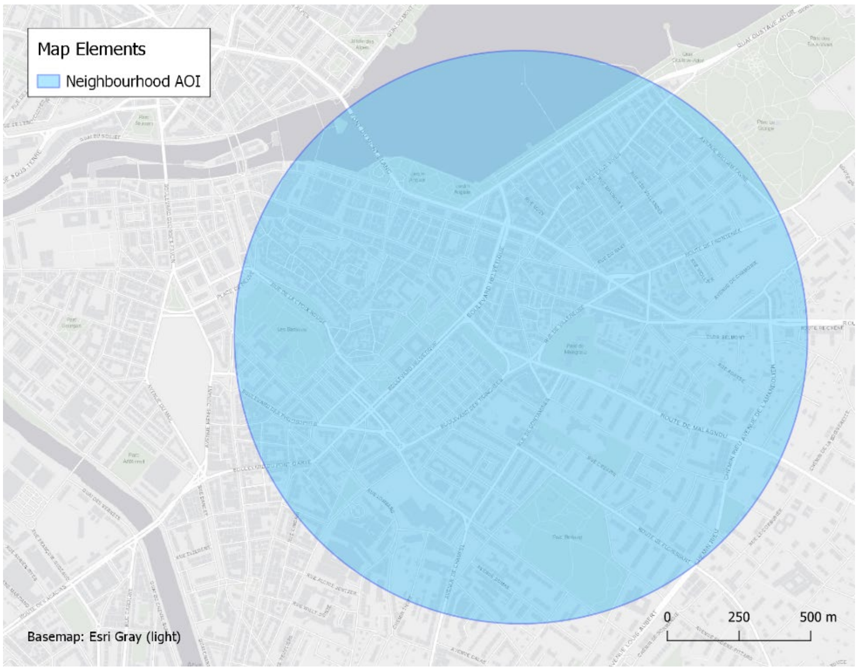


Figure 10: Neighbourhood Area of Interest (AOI) with a radius of 1000 meters in Geneva.

Step 2 - Feature to Point: The polygon feature *Neighbourhood AOI* was used as input data to create the centroid point, which de facto is the homeplace location on which the neighbourhood area is based on and was saved as output *AOI Centroid (Homeplace)* (see *Figure 11*). This step may seem unnecessarily circumstantial, as it would also be possible to select the homeplace location via unique FID from the users$_{home}$ dataset. However, this approach added some convenience to the workflow, which is why the author decided to do it in this manner.

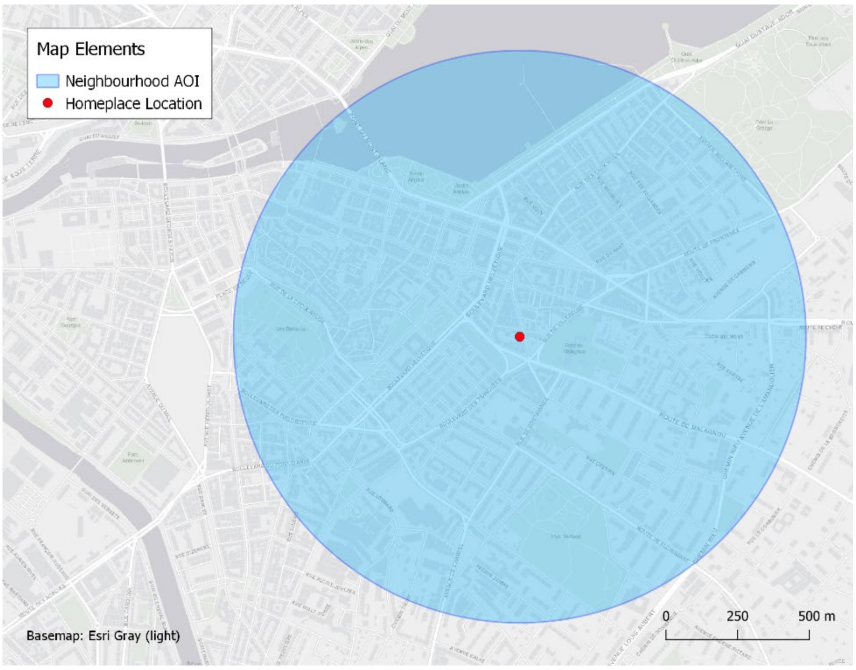


Figure 11: *AOI Centroid (Homeplace)* derived from the *Neighbourhood AOI.*

Step 3 - Euclidean Distance: The point feature *AOI Centroid (Homeplace)* is used as input to calculate the Euclidean distance from the point outwards, with a maximum distance equal to the radius of the AOI, i.e., 500 meters or 1000 meters respectively, and a cell size of 50 meters. The output was saved as *Distance Raster* (see *Figure 12*) and covers the AOI with cell values ranging from 0 in the centre to 1000 (or 500 if radius is 500 m) at the perimeter, representing the distance in meters from each cell centre to the homeplace location. Note that the cells are not perfectly symmetrical around the homeplace location, because the underlying algorithm aligns the raster grid with the point of origin, choosing only one of the four cells around the origin as the closest cell with the value 0. However, this small shift of the grid is neglectable and has hardly any influence on the result.

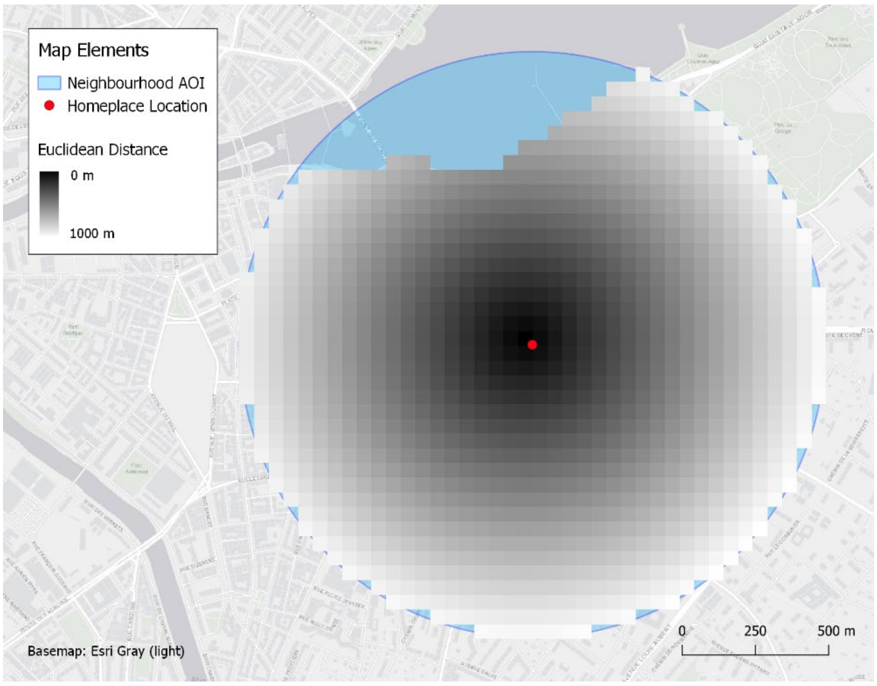


Figure 12: *Distance Raster* indicating close cells in black and distant cells in white.

Step 4 - Normalisation: The raster *Distance Raster* is used as input to calculate a normalised weight mask to approximate linear distance decay. The normalization of the values between 0 and the maximum distance was calculated using the following formula:

$$Weight = 1 - \frac{Euclidean\ Distance + 1}{Radius_{neighbourhood} + 40}$$

Three numeric examples with Euclidean Distance of 0, 500, and 1000, and Radius of 1000:

$$Weight_0 = 1 - \frac{0+1}{1000+40} = 1 - \frac{1}{1040} = 1 - 0.001 = 0.999 \ \approx \mathbf{1}$$

$$Weight_{500} = 1 - \frac{500+1}{1000+40} = 1 - \frac{501}{1040} = 1 - 0.482 = 0.518 \ \approx \mathbf{0.5}$$

$$Weight_{1000} = 1 - \frac{1000+1}{1000+40} = 1 - \frac{1001}{1040} = 1 - 0.963 = 0.037 \ \approx \mathbf{0}$$

As already mentioned above, the raster grid is shifted by half of a cell size, i.e., 25 meters in both X and Y axis, leading to maximum distance values of slightly above 1000 meters, potentially up to 1037.5 meters. Adding 40 to the neighbourhood radius prevents the resulting weight to take negative values. The output raster was saved as *Distance Decay Mask* (see *Figure 13*), and includes continuous values between 0 and 1, which serve as factors to weight the cell values of the three neighbourhood variables greenspace, traffic noise, and socio-economic position.

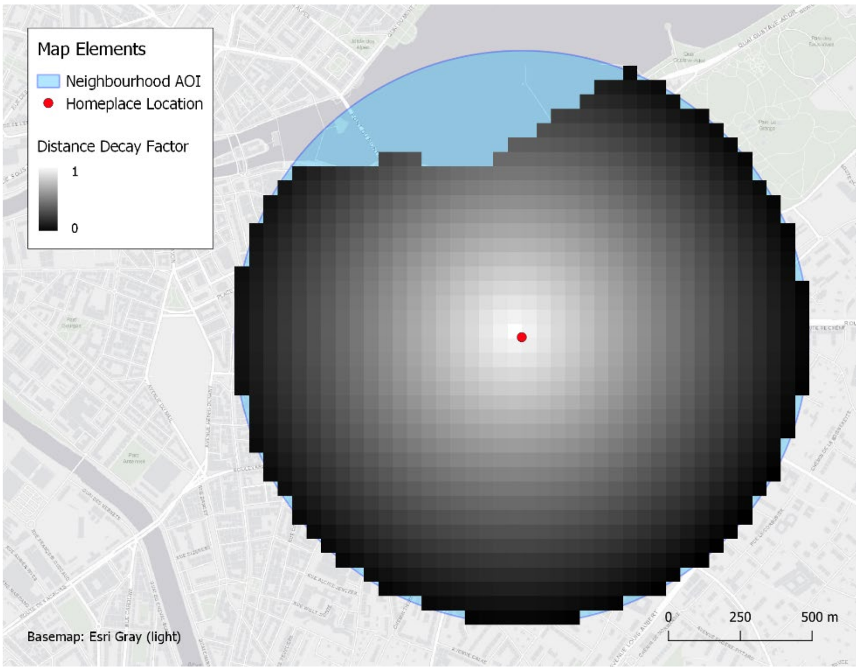


Figure 13: Normalised *Distance Raster* taking continuous values between 0 and 1 (distance decay factor).

Step 5 - Weighting: The raster *Distance Decay Mask* is used as input to be multiplied with the *Neighbourhood Variable Raster*, here at the example of the greenspace raster (see *Figure 14*), which was previously clipped to the AOI. The final output of the modelled process chain is the *Weighted Neighbourhood Variable Raster*, shown in *Figure 15* at the example of the weighted greenspace raster.

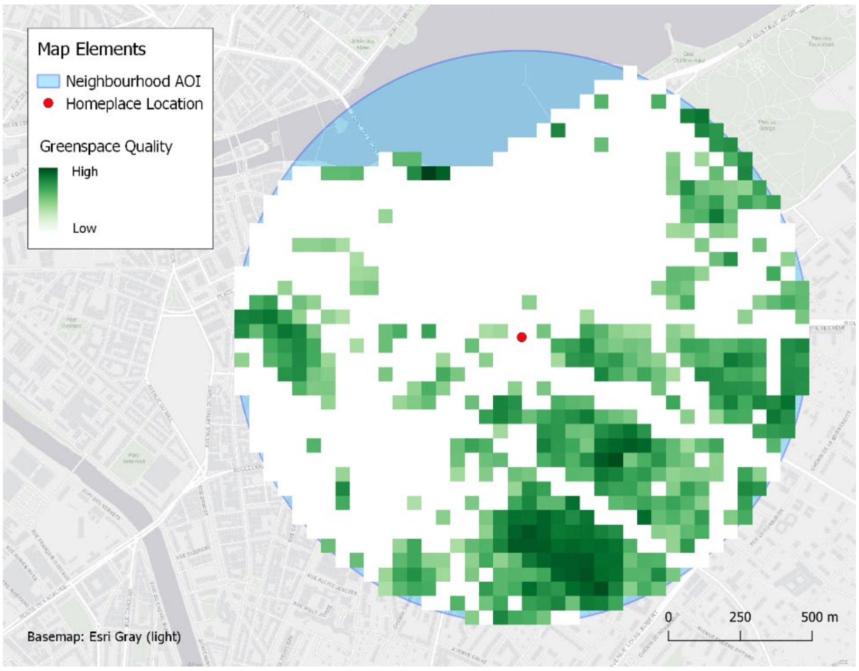


Figure 14: Greenspace raster clipped to AOI. Dark green cells indicate areas with high greenspace quality, light green cells indicate areas with low greenspace quality. White cells mean there is no greenspace.

When comparing *Figure 14* with *Figure 15*, it can be clearly seen how the colour saturation of the green cells tends to decrease towards the perimeter after the inverse distance weighting. The lower the colour saturation of a cell, the lower the underlying value. Hence, when calculating the mean value within the AOI, the outer cells have lesser influence on the result than the ones close to the homeplace.

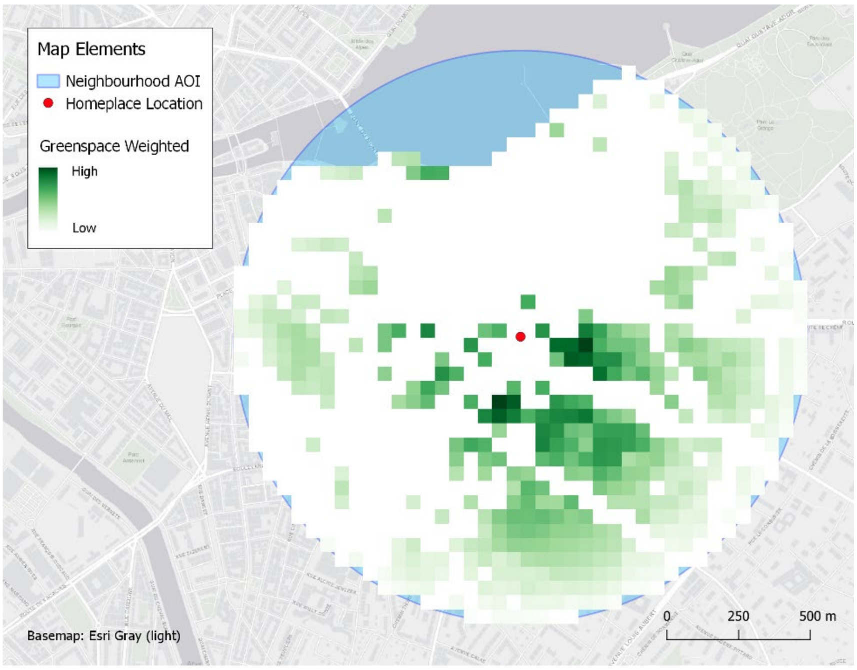


Figure 15: Weighted greenspace raster. Here, the green tones do not directly represent greenspace quality, but rather reflect the impact of the greenspace at that particular location on the resident living at the red dot. High quality greenspace further away may have less positive influence on the resident than close greenspace with moderate quality.

These 5 processing steps had to be applied on 1213 neighbourhoods, which is why the procedure had to be automated. Although the ArcGIS Pro Modelbuilder features an iteration functionality, it was decided to export the model as a Python script. Experience shows that running the process chain in ArcGIS Pro has a longer computing time and tends to be more error prone, than executing it as a Python script. This process is explained in the following subsection.

5.4.2 Parallel Processing

After exporting the process chain to Python, some adjustments and additions in the script had to be made. First, all processing steps were embedded in a *for loop*, iterating through all the different neighbourhoods. Secondly, the mean cell value of the resulting weighted raster at the end of each iteration was calculated and stored in an array and finally in a CSV file. This value was later assigned to the corresponding user living in that neighbourhood via FID. Finally, the script was copied, once for every neighbourhood variable (greenspace, traffic noise, and SEP), and once for each neighbourhood radius (500 m, and 1000 m), making 6 scripts in total. After adjusting some parameters and the input and output paths according to variable and radius, the first script was tested by running it via Command Prompt. When reading the printed computing time of about 1 minute per iteration, it became immediately clear that scaling it up to 1213 iterations would result in a computing time of almost a day. Multiplying it by the number of scripts would take up nearly an entire week. Since certain errors and misconceptions may only be detected after having executed the whole procedure, making it necessary to be repeated one or multiple times, doing it in this manner was not feasible. The 1213 iterations had to be divided into multiple equal parts and needed to be run simultaneously to reduce computing time. It was decided to split the input data into 8 parts, each part being processed by a separate Python script. This led to a computing time of 2 to 3 hours per neighbourhood variable and radius. After adjusting the input and output directory paths according to the script numbers 1 to 8, all scripts were successfully run and the resulting CSV files containing the mean values of the neighbourhood variables for each homeplace were joined with the users$_{home}$ dataset by FID. In total, 7 neighbourhood variables were generated in this process. The variable names are *green_500, green_1000, day_500, day_1000, ssep_500, ssep_1000*, and *night_100*. *Green_500* and *green_1000* representing greenspace for a 500 m radius, and a 1000 m radius, *day_500* and *day_1000* for daytime traffic noise, *ssep_500* and *ssep_1000* for SEP, and finally, *night_100* for traffic noise during night. The general Python script can be found in the *Appendix A.3*.

5.5 Regression Analysis

Having estimated and added the neighbourhood variables greenspace, traffic noise, and SEP to the dataset users$_{home}$, the main preparatory steps for the regression analysis were completed. In this section, the last remaining steps within the data preparation phase are explained, an exploratory data analysis (EDA) is performed, and finally, the approach for the regression analysis is described. The 5 variables listed in *Table 10* were introduced in *Subsection 5.2.1* and represent the outcome variables for the regression models. For each outcome variable, a separate regression model was built, which is explained in detail in *Subsection 5.5.2*.

Table 10: Variables used as the outcome variables for the regression models

Variable	Description
rate_neg	Ratio of tweets with negative emotions / total tweets
rate_happy	Ratio of tweets with the emotion happiness / total tweets
rate_stress	Ratio of tweets with stress / total tweets
negemo	Percentage of words considered as emotionally negative in all tweets
posemo	Percentage of words considered as emotionally positive in all tweets

5.5.1 Choice of Regression Model

After having prepared the dataset for the regression analysis, the next step was to choose the right type of regression model. To start, the assumptions for a multiple linear regression were checked. The first assumption to be met is an existing linear relationship between each explanatory variable and the outcome variable (Schreiber-Gregory & Bader, 2018).

Figure 16 shows that no linear relationship was found when plotting each of the 5 outcome variables with each of the 7 neighbourhood variables.

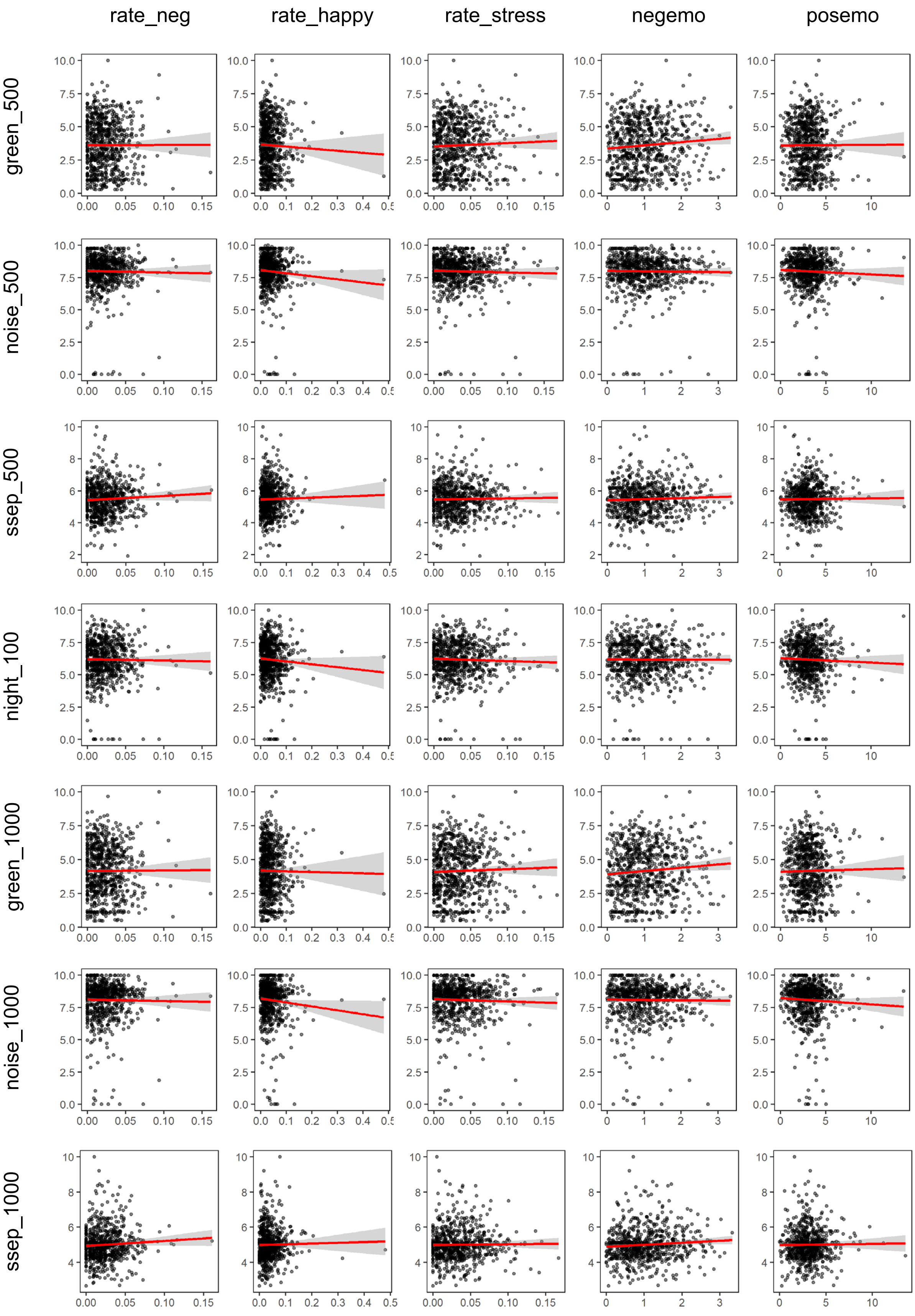


Figure 16: Scatterplots between all 7 neighbourhood variables and the 5 outcome variables for the regression model. No clear linear relationship is visible.

Therefore, the assumptions for a multiple linear regression were not met. Instead, it was decided to perform a multiple logistic regression. To do so, new variables were generated, dividing the continuous values of the outcome variables into a binary classification with the values 0 and 1. These new variables were named *rate_neg_01, rate_happy_01, rate_stress_01, negemo_01, posemo_01* and replace the continuous outcome variables for the logistic regression models.

5.5.2 Multiple Logistic Regression

The aim was to build a multiple logistic regression model for each of the 5 outcome variables, using the three neighbourhood variables greenspace, traffic noise, and SEP as the main explanatory variables. One variable representing gender, one variable for age, and one variable representing urban-/rural-typology were added to the regression models as control variables.

Before defining the exact regression models, the different potential explanatory variables were tested for multicollinearity. The resulting correlation matrix is illustrated in *Figure 6.10* and discussed in *Subsection 6.13* of the next Chapter.

The basic regression model is defined as follows:

$$
\begin{aligned}
Y = \beta_0 + \beta_1 * Gender + \beta_2 * Age + \beta_3 * Typology + \beta_4 * Greenspace + \beta_5 * Noise \\
+ \beta_6 * SEP
\end{aligned}
$$

In addition, interaction terms between all the explanatory variables were added to the model:

$$
\begin{aligned}
Y = \beta_0 + \beta_1 * Gender + \beta_2 * Age + \beta_3 * Typology + \beta_4 * Greenspace + \beta_5 * Noise \\
+ \beta_6 * SEP + \beta_7 * Gender \times Age + \beta_8 * Gender \times Typology + \beta_9 \\
* Gender \times Greenspace + \beta_{10} * Gender \times Noise + \beta_{11} * Gender \times SEP \\
+ \beta_{12} * Age \times Typology + \beta_{13} * Age \times Greenspace + \beta_{14} * Age \times SEP \\
+ \beta_{15} * Age \times Noise + \beta_{16} * Typology \times Greenspace + \beta_{17} \\
* Typology \times SEP + \beta_{18} * Typology \times Noise + \beta_{19} * Greenspace \times SEP \\
+ \beta_{20} * Greenspace \times Noise + \beta_{21} * Noise \times SEP
\end{aligned}
$$

In R, the function *stepAIC()* was used on the model for each outcome variable separately. In this manner, the optimal set of variables for each outcome variable is produced. The resulting regression equations are shown in the *Subsections 6.3.1* and *6.3.2*.

Chapter 6 | Results

The results of the thesis are described and visualised in this chapter. First, the descriptive statistics about the analysed dataset $\text{users}_{\text{regression}}$ are shown. Then, the representativeness of $\text{users}_{\text{regression}}$ is reviewed, comparing it to $\text{users}_{\text{all}}$. And finally, the regression models for each outcome variable are presented and their outputs are described in detail. The results of the regression analysis elaborated in this chapter are all based on the neighbourhood variables for the 500 meters radius. From now on, the 4 neighbourhood variables *green_500*, *day_500*, *ssep_500*, and *night_100* are referred to as *greenspace*, *day_noise*, *sep*, and *night_noise*.

6.1 Characteristics of Analysed Users

This section shows the spatial distribution of the homeplace locations of the analysed users, the distributions of the most relevant variables, as well as correlations between the variables.

6.1.1 Spatial Distribution

Figure 17 shows the spatial distribution of the homeplace locations of the 733 users analysed in the regression analysis ($\text{users}_{\text{regression}}$). Most of the points are concentrated in the area in and around the city of Geneva (~ 190 homeplaces), the city of Zürich (~ 75 homeplaces), the city of Basel (~ 58 homeplaces), Lausanne (~ 49 homeplaces), and Lugano (~ 30 homeplaces). More than a quarter of the 733 total users are therefore located in or around the city of Geneva, which is about two and a half times more than in and around the city of Zürich. Although there are single locations in rural areas and in alpine regions, the vast majority of the points is concentrated in the cities. An alternative visualization showing clusters with the number of homeplace locations within the clusters can be found in *Appendix A.2*. It may give a better overview of the actual number of points at specific locations and reveals that there is a disproportionally high number of homeplace locations in Davos.

6.1.2 Distributions of Variables

In this subsection, the distributions of the most relevant variables of $\text{users}_{\text{regression}}$ are illustrated with histograms, which are briefly described. The following descriptive statistics for each variable are pointed out: *mean* (average value), *min* (lowest value), *max* (highest value), *median* (median value), *sd* (standard deviation). The red vertical line marks the mean of the variable.

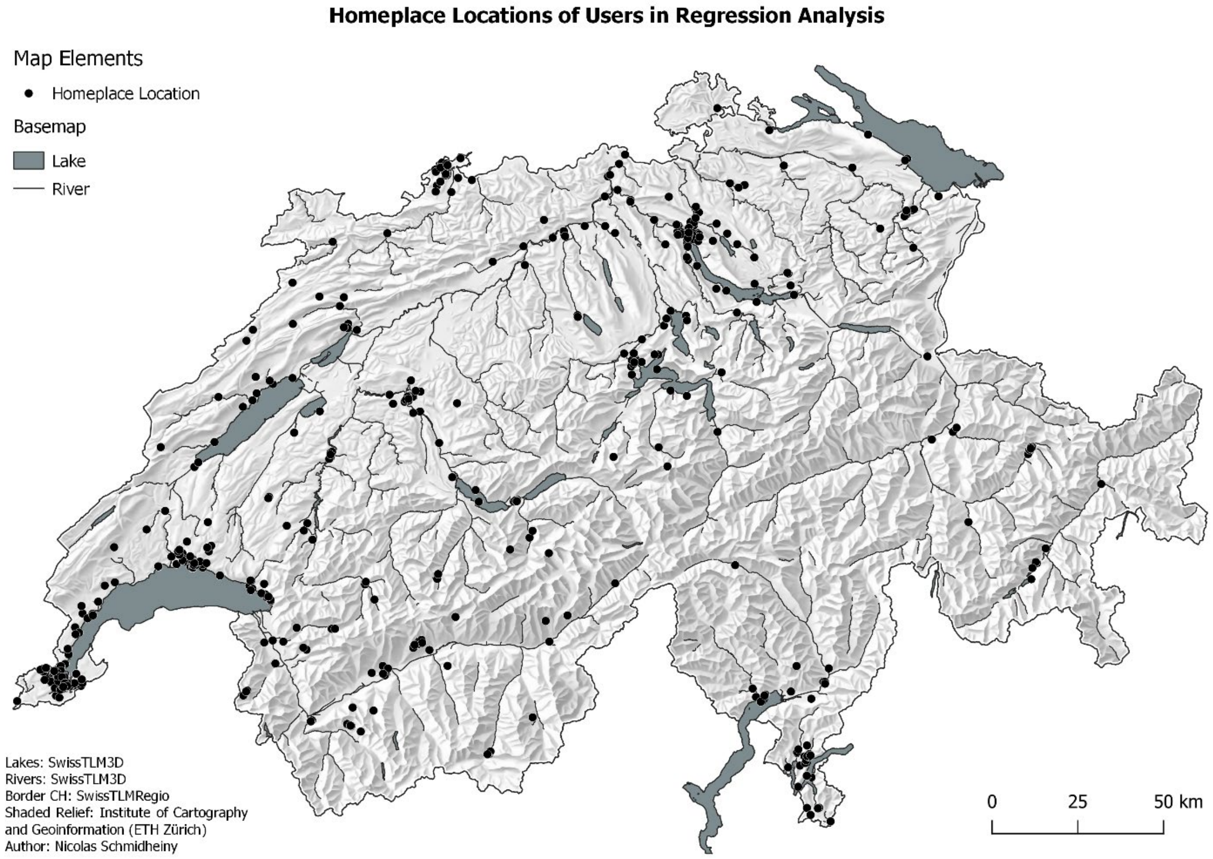

Homeplace Locations of Users in Regression Analysis
Map Elements
● Homeplace Location
Basemap
Lake
River
Lakes: SwissTLM3D
Rivers: SwissTLM3D
Border CH: SwissTLMRegio
Shaded Relief: Institute of Cartography and Geoinformation (ETH Zürich)
Author: Nicolas Schmidheiny
0
25
50 km

Figure 17: Visualization of the spatial distribution of the analyzed users' presumed homeplace locations

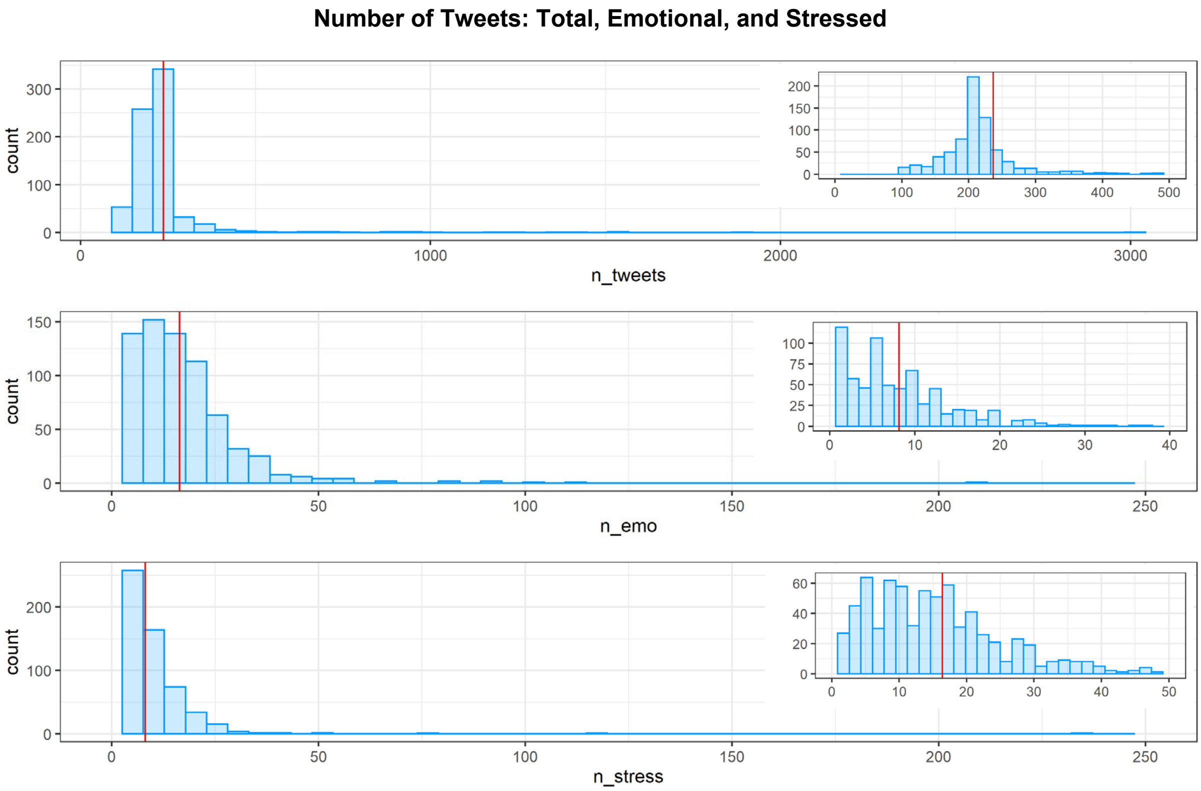


Figure 18: Histograms showing the number of tweets (*n_tweets*), the number of emotional tweets (*n_emo*), and the number of "stressed" tweets (*n_stress*) of each user in the dataset $\text{users}_{\text{regression}}$.

The variable *n_tweets* stands for the total number of tweets per user. For most of the users around 200 tweets were available. This is due to the rehydration process described in *Subsection 4.1.1*, where the last 200 tweets of each user have been collected. For *n_tweets,* these are the descriptive statistics: *mean = 236.5, min = 100, max = 2995, median = 210, sd = 181.5*. The min is exactly 100 because of the cut-off mentioned in *Section 5.3*. Because not only the tweets from the rehydrated datasets $\text{tweets}_{\text{english}}$ and $\text{tweets}_{\text{non-english}}$ have been used, but also the ones from $\text{tweets}_{\text{geoloc}}$, for some users there are more than 200 tweets available. Only for 21 users there are more than 500 tweets. The distribution is roughly bell shaped between 100 and 400 and shows a few far outliers on the right side.

The variable *n_emo* stands for the number of tweets per user which EMOTIVE successfully classified into an emotion category. For *n_emo*, these are the descriptive statistics: *mean = 16.4, min = 0, max = 209, median = 14, sd = 14.6*. There are 12 users with 0 emotional tweets. 16 users have more than 50 emotional tweets. The distribution is strongly skewed and shows a few outliers on the right side.

The variable *n_stress* stands for the number of tweets per user where Stresscapes assigned a stress value greater than 0. In other words, *n_stress* counts the number of "stressed" tweets per user. For *n_stress*, these are the descriptive statistics: *mean = 8.1, min = 0, max = 236, median = 6, sd = 11.9.* The number of users showing 0 "stressed" tweets is 56. The number of users having more than 25 "stressed" tweets is 16. The distribution is right skewed and shows very few far outliers on the right side.

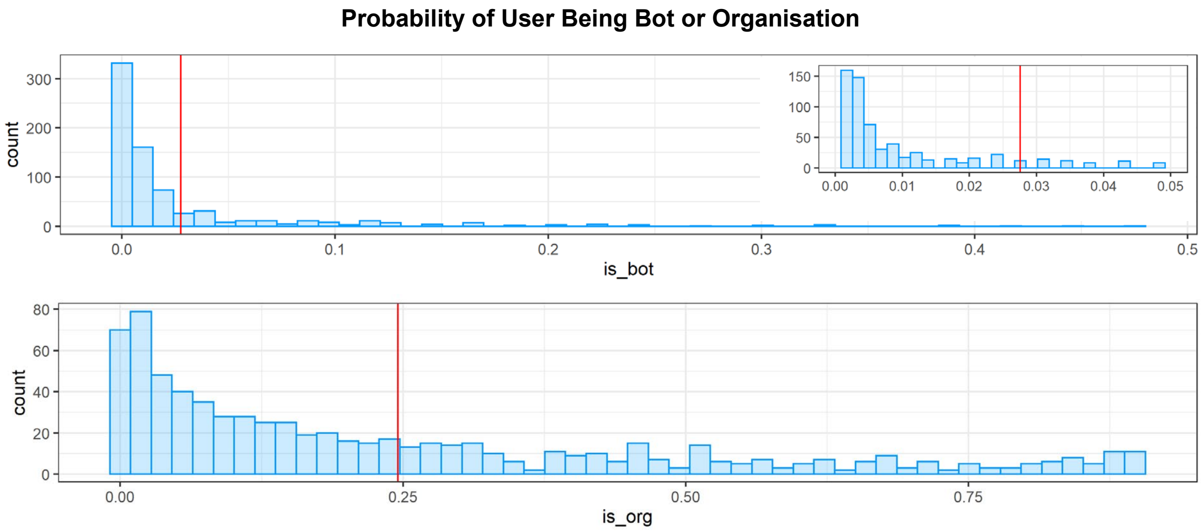


Figure 19: Histograms of probabilities of user being an automated account (*is_bot*) or an account managed by an organisation (*is_org*).

The variable *is_bot* stands for the probability of the user being an automated account (bot). For *is_bot*, these are the descriptive statistics: *mean = 0.028, min = 0, max = 0.475, median = 0.006, sd = 0.058.* Only 3 users have a 0% probability of being a bot. The maximum is 0.475, because of the exclusion of users having a probability higher than 50% (see *Section 5.3*). The number of users having a probability higher than 10% of being an automated account is 55. The distribution is strongly right skewed and shows a few far outliers on the right side.

The variable *is_org* stands for the probability of the user being an account managed by an organisation. For *is_org*, these are the descriptive statistics: *mean = 0.245, min = 0, max = 0.897, median = 0.148, sd = 0.255*. The maximum probability is 0.897, because of the cut-off at 0.9 (see *Section 5.3*). A value above 50% is found in 134 users, and 53 users have a probability higher than 75% of being an account managed by an organisation. The distribution is right skewed with two small additional modi at around 0.5 and at the right end.

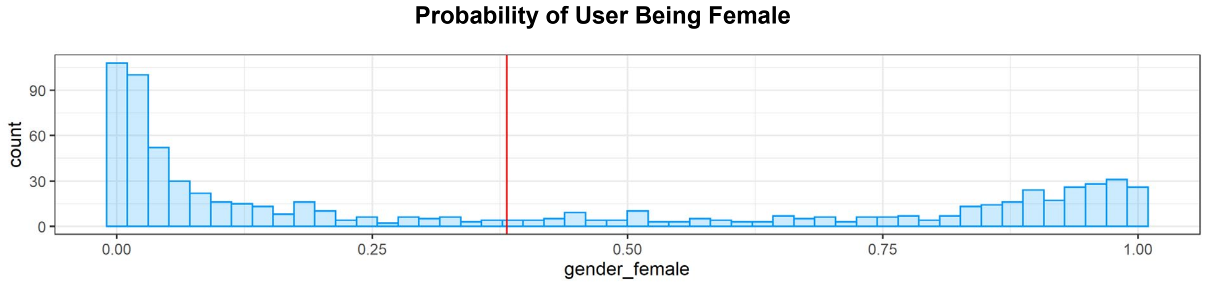


Figure 20: Histogram of probability of user being female.

The variable *gender_female* stands for the probability of the user being of the female gender. The variable *gender_male* has the exact same distribution, but mirror inverted, because the sum of the two gender variables always equals 1. Since one of the two variables is redundant, the variable *gender_female* was chosen for the analysis. For *gender_female*, these are the descriptive statistics: *mean = 0.382, min = 0, max = 0.9996, median = 0.177, sd = 0.391*. Only 1 user has a value of 0. For 277 users, the probability of being of the female gender, is higher than 50%. The distribution is clearly bimodal, with the main mode close to 0, and the second mode close to 1. Between 0 and 0.25 the distribution is right skewed, between 0.25 and 0.75 it tends to be uniformly distributed, and between 0.75 and 1 it is left skewed.

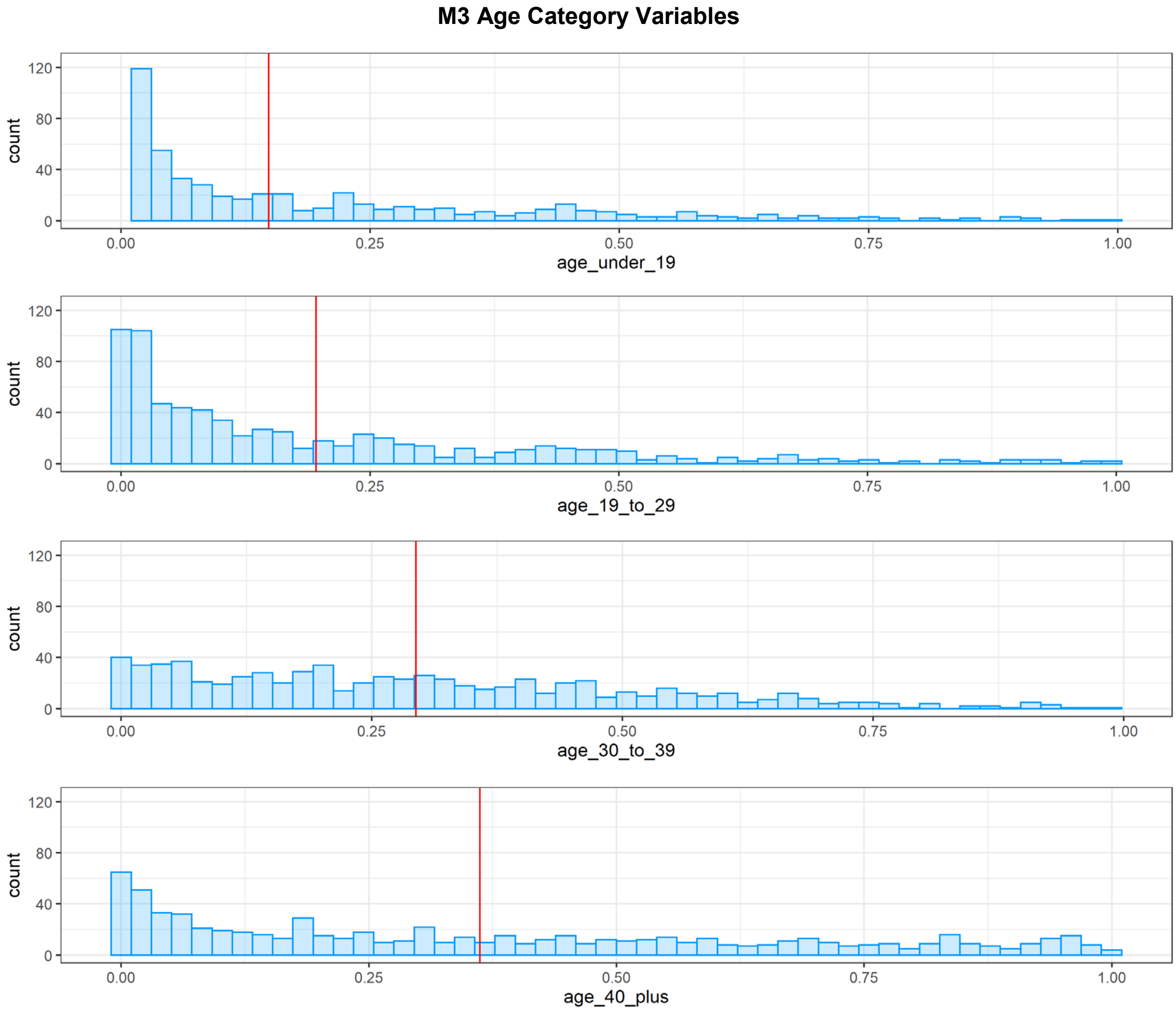


Figure 21: Histograms of probabilities of user belonging to the respective age categories.

The four age variables derived with the M3-method divide the estimated age of a user into 4 categories: below 19 *(age_under_19)*, 19 to 29 *(age_19_to_29)*, 30 to 39 *(age_30_to_39)*, and above 39 *(age_40_plus)*.

The variable *age_under_19* stands for the probability of the user being Age under 19. For *age_under_19*, these are the descriptive statistics: *mean = 0.148, min = 0, max = 0.994, median = 0.044, sd = 0.206*. 7 users have the minimum probability of 0. The number of users with a value greater than 50% is 59, and with a value greater than 25% there are 160 users. The distribution is strongly right skewed.

The variable *age_19_to_29* stands for the probability of the user being of age between and including 19 to 29. For *age_19_to_29*, these are the descriptive statistics: *mean = 0.196, min = 0, max = 0.995, median = 0.106, sd = 0.222*. Only 1 user has the probability of 0. A probability above 50% of being in this age category is found in 77 users, and above 25% there are 219 users. The distribution is right skewed.

The variable *age_30_to_39* stands for the probability of the user being of age between and including 30 to 39. For *age_30_to_39*, these are the descriptive statistics: *mean = 0.294, min*

= 0, max = 0.988, median = 0.26, sd = 0.224. 3 users have the value 0. For 141 users the probability of being in this age category is higher than 50% and for 377 users the probability is above 25%, which is roughly half of the sample size. The distribution is slightly right skewed.

The variable *age_40_plus* stands for the probability of the user being of age 40 or older. For *age_40_plus*, these are the descriptive statistics: *mean = 0.362, min = 0, max = 0.9998, median = 0.298, sd = 0.306*. 6 users have the value 0. The number of users having a probability higher than 50% of belonging to this age category is 241, which is roughly a third of the sample. 395 users have a probability higher than 25% of belonging to this age category. The distribution tends to be right skewed between 0 and 0.25 and tends to be uniformly distributed between 0.25 and 1.

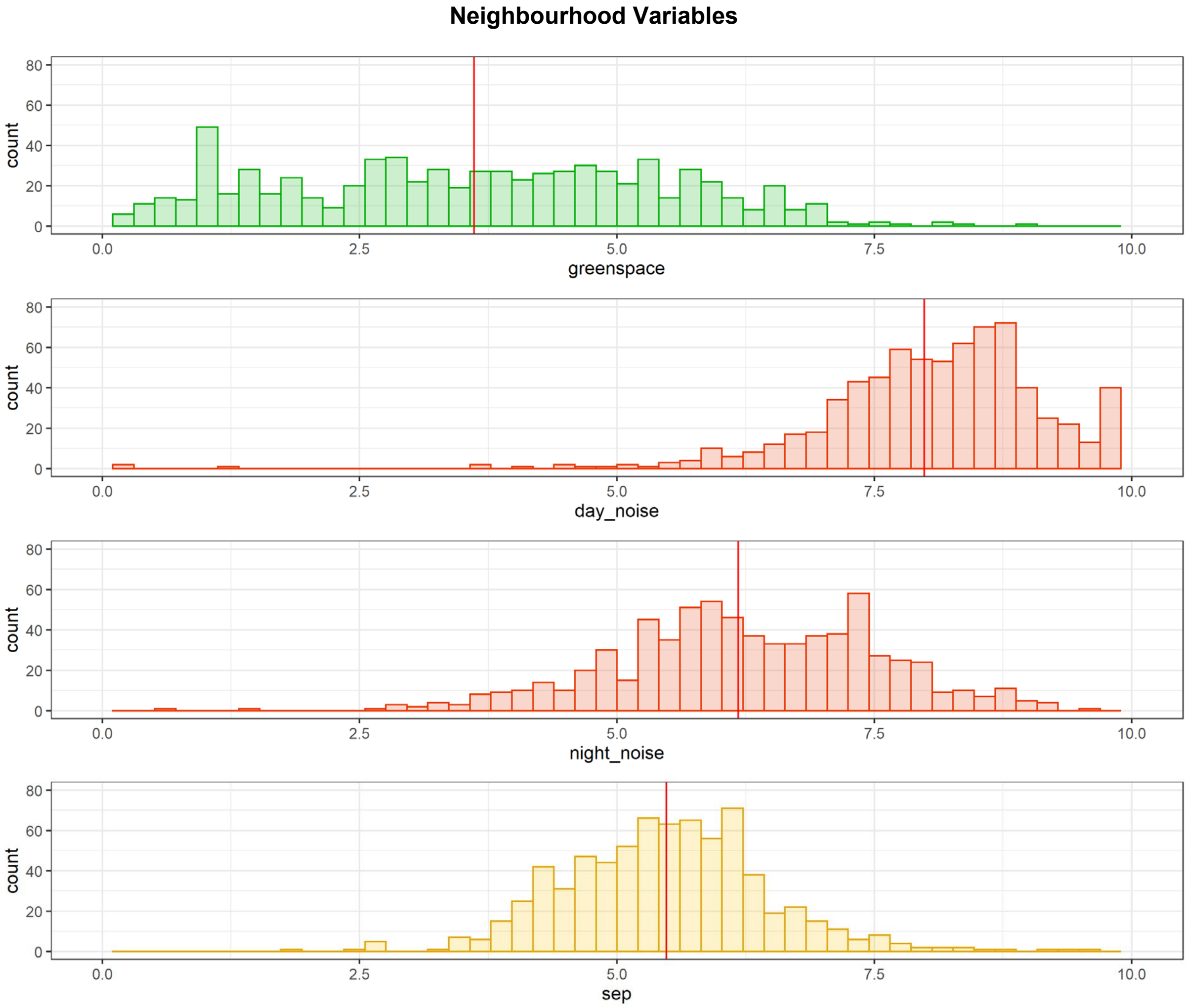


Figure 22: Histograms of the neighbourhood variables based on a 500 meters radius.

The values of the four neighbourhood variables were rescaled between 0 and 10 to make them more convenient, since the values have no directly interpretable meaning. Therefore, the maximum value of the four variables is always 10.

The variable *greenspace* represents the "greenness" of the neighbourhood in which the homeplace location of a user is centred in. The variable combines the amount, the closeness, and the quality of the available greenspace within 500 meters of the homeplace location. Higher values mean that the neighbourhood is "greener", lower values mean that the neighbourhood is less green. For *greenspace*, these are the descriptive statistcs: *mean = 3.61,* *min = 0.27, max = 10, median = 3.69, sd = 1.86*. The distribution is roughly symmetric around the mean, with a slight tendency to a bell shape. There are a few outliers on the right.

The variable *day_noise* represents the exposure of the user to traffic noise during day time, that is, between 6 am and 10 pm, within 500 meters of the homeplace location. Higher values mean higher exposure to traffic noise, lower values mean less exposure. For *day_noise*, these are the descriptive statistcs: *mean = 7.98, min = 0, max = 10, median = 8.19, sd = 1.39*. There are 8 users with the value 0. The distribution tends to be bell shaped with some far outliers on the left.

The variable *night_noise* represents the exposure of the user to traffic noise during night, that is, between 10 pm and 6 am. For this variable, it is assumed that the user is sleeping and has therefore a fixed position. For *night_noise*, these are the descriptive statistcs: *mean = 6.18, min = 0, max = 10, median = 6.2, sd = 1.49*. There are 11 users with the value 0. The distribution has the tendency of a flattened bell shape with a few outliers on the left.

The variable *sep* represents the socio-economic position of the neighbourhood in which the homeplace location of a user is centred in. Higher values mean that the socio-economic position is higher, lower values the opposite. For *sep*, these are the descriptive statistcs: *mean = 5.48, min = 1.91, max = 10, median = 5.49, sd = 1.01*. The distribution is bell shaped with no significant outliers.

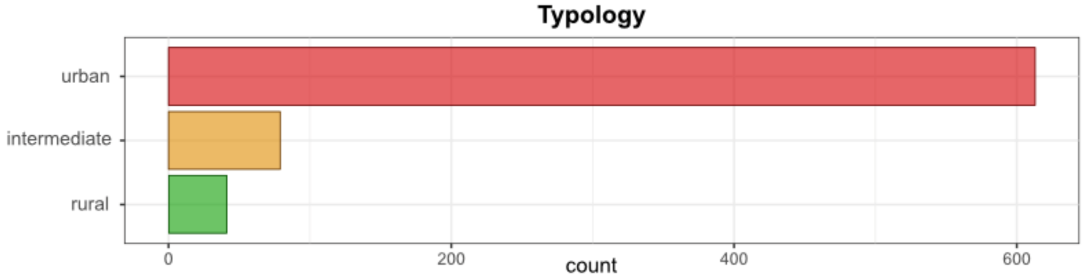


Figure 23: Barplot of number of users in each of the three urban-/rural-typology.

The variable *typology* classifies the homeplace locations of the users into the three categories rural, intermediate, and urban based on the dataset introduced in *Section 4.5*. Out of the 733 total users, 613 fall into the category urban, 79 into the category intermediate, and 41 into the category rural.

Since the users belonging to the category urban heavily outnumber the users of the other two categories, it was decided to produce a new binary variable named *urban_01*. This new variable assigns the value 0 to users belonging to the typology rural or intermediate, and the value 1 to the ones belonging to the typology urban. The variable *urban_01* replaced the variable *typology* in the regression models.

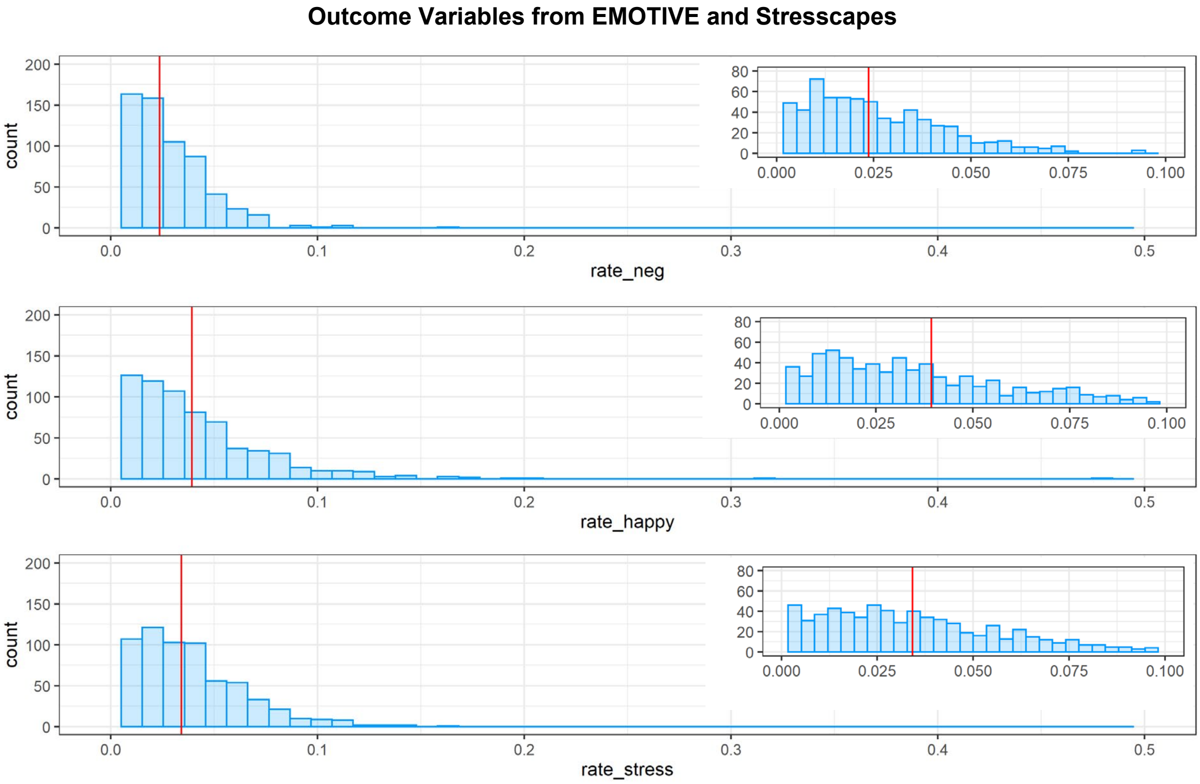


Figure 24: Histograms of rates of emotionally negative tweets (*rate_neg*), tweets expressing the emotion happiness (*rate_happy*), and tweets indicating stress (*rate_stress*).

The variable *rate_neg* stands for the ratio between the number of tweets where EMOTIVE found at least 1 negative emotion, and the total number of tweets. For *rate_neg*, these are the descriptive statistics: *mean = 0.024, min = 0, max = 0.161, median = 0.02, sd = 0.02*. The value 0 is found in 83 users. Only 5 users expressed negative emotions in more than 10% of their tweets. The distribution is right skewed with no significant outliers.

The variable *rate_happy* stands for the ratio between the number of tweets where EMOTIVE found the emotion “happiness”, and the total number of tweets. For *rate_happy*, these are the descriptive statistics: *mean = 0.039, min = 0, max = 0.481, median = 0.031, sd = 0.038*. The value 0 is found in 34 users. 40 users expressed the emotion “happiness” in more than 10% of their tweets. The distribution is right skewed with only 2 significant outliers on the right.

The variable *rate_stress* stands for the ratio between the number of tweets where the NLP-algorithm Stresscapes recognized stress, and the total number of tweets. For *rate_stress*, these are the descriptive statistics: *mean = 0.034, min = 0, max = 0.168, median = 0.029, sd = 0.027*. The value 0 is found in 56 users. There are 19 users, for which in more than 10% of the tweets stress was recognized. The distribution is right skewed with no significant outliers.

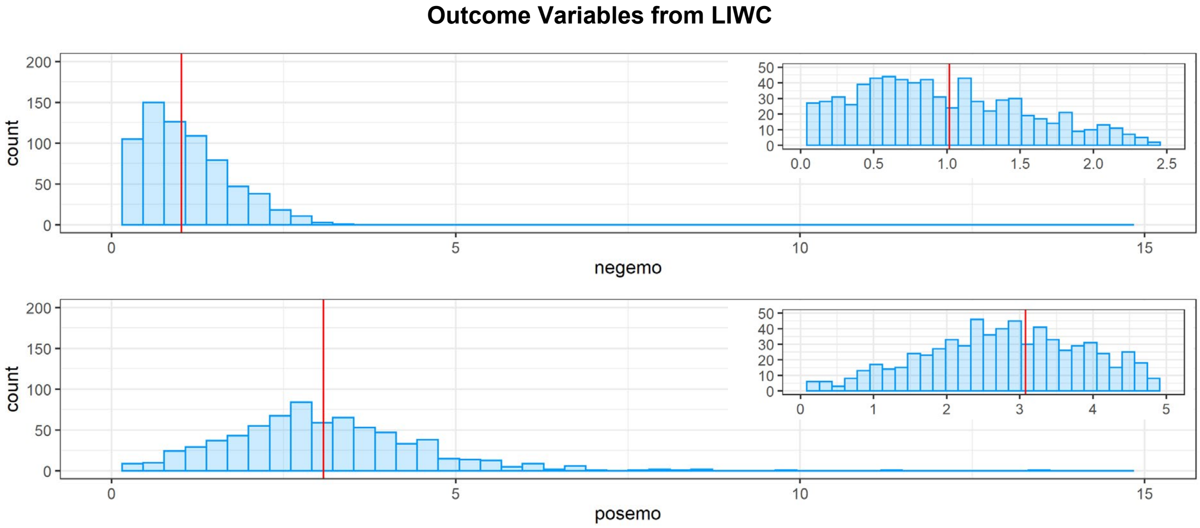


Figure 25: Histograms of percentage of words in all tweets for each user expressing negative emotions (*negemo*), and positive emotions (*posemo*).

The variable *negemo* stands for the percentage of words in all tweets which were associated with negative emotions, using the NLP-algorithm LIWC. For *negemo*, these are the descriptive statistics: *mean = 1.02, min = 0, max = 3.33, median = 0.88, sd = 0.652*. There are 7 users with the value 0. The distribution is right skewed with no outliers.

The variable *posemo* stands for the percentage of words in all tweets which were associated with positive emotions, using the NLP-algorithm LIWC. For *posemo*, these are the descriptive statistics: *mean = 3.08, min = 0, max = 13.58, median = 2.93, sd = 1.49*. Only 1 user has the value 0. There are 9 users, for which more than 7.5% of their tweeted words were associated with positive emotions. The distribution has a bell shape with a few significant outliers on the right.

6.1.3 Correlation Between Variables

In this section, the correlation between the different explanatory variables, as well as between the different outcome variables are examined. For this purpose, two correlation matrices were created. One for the explanatory variables, and one for the outcome variables.

Figure 26 shows the correlation matrix of the explanatory variables as a heatmap. The explanatory variables include: *gender_female* to represent gender, the four age categories *age_under_19*, *age_19_to_29*, *age_30_to_39*, and *age_40_plus* to represent age, and finally the four neighbourhood variables *greenspace, day_noise, night_noise,* and *sep* to represent the physical and social characteristics of the neighbourhood. The values in the cells are the Pearson correlation coefficients (PCC) between the two corresponding variables on the X- and on the Y-axis. The correlations can take values between -1 and 1, where -1 indicates a perfect negative correlation and 1 a perfect positive correlation.

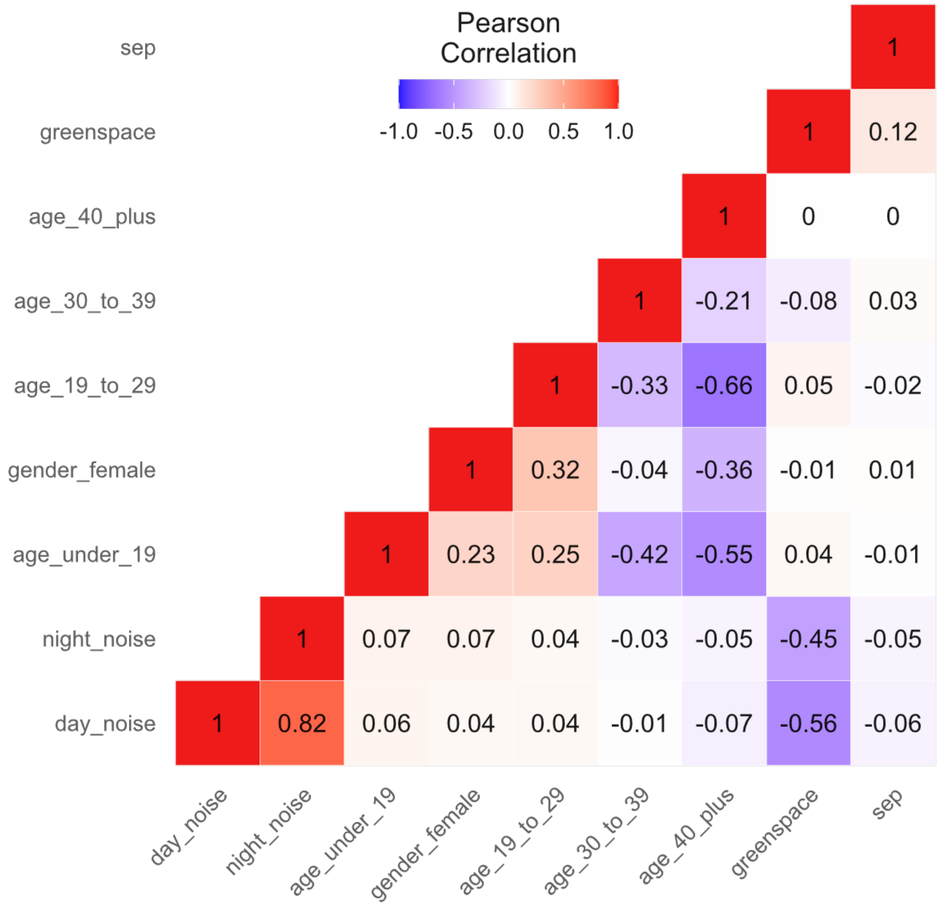


Figure 26: Correlation matrix as a heatmap showing correlations between explanatory variables

In *Table 11*, all correlations above an absolute value of 0.2 are listed from highest to lowest absolute PCC.

Table 11: Pearson correlation coefficient between explanatory variables

Variable 1	Variable 2	Pearson Correlation Coefficient
day_noise	night_noise	0.82
age_19_to_29	age_40_plus	-0.66
greenspace	day_noise	-0.56
age_under_19	age_40_plus	-0.55
greenspace	night_noise	-0.45
age_under_19	age_30_to_39	-0.42
gender_female	age_40_plus	-0.36
age_19_to_29	age_30_to_39	-0.33
gender_female	age_19_to_29	0.32
age_under_19	age_19_to_29	0.25
age_under_19	gender_female	0.23
age_30_to_39	age_40_plus	-0.21

In *Figure 27*, scatterplots between the three neighbourhood variables *greenspace*, *day_noise*, and *night_noise* are illustrated to visually underline the Pearson correlation coefficients between the pairs.

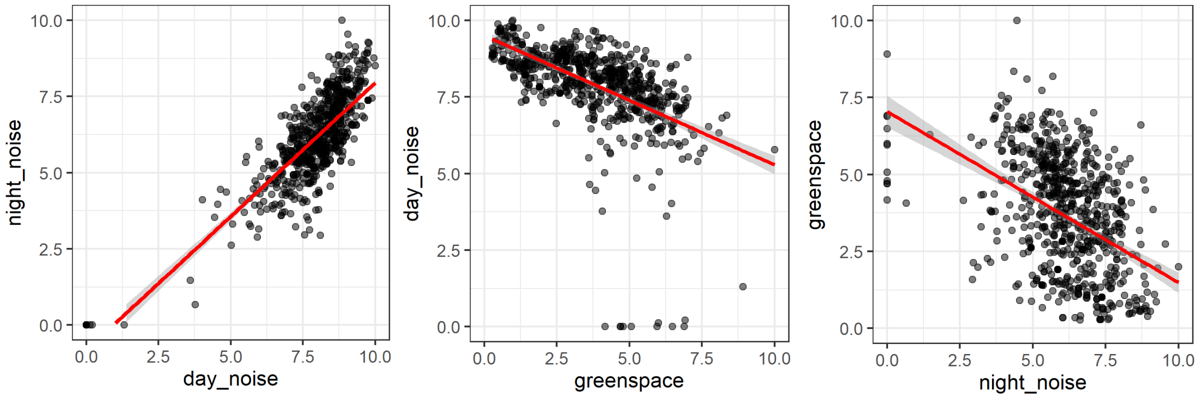


Figure 27: Scatterplots of the neighbourhood variable pairs with the highest PCCs.

When inspecting the scatterplots, it can be confirmed that there is a strong positive linear association between *day_noise* and *night_noise*. The variables greenspace and *day_noise* tend to have a moderately strong negative linear association. Lastly, *night_noise* and *greenspace* show a rather weak negative linear association.

Figure 28 shows the correlation matrix of the outcome variables as a heatmap. *rate_neg* shows the same PCC with *rate_stress* as with *negemo*, which is 0.65. The PCC between *negemo* and *rate_stress* is very similar with a value of 0.62. The variables associated with positive emotions, being *posemo* and *rate_happy,* show a PCC of 0.53.

The Scatterplots of these 4 mentioned correlation pairs are shown in *Figure 29*. All 4 variable pairs show a similar correlation pattern, that is, a moderately strong positive linear association.

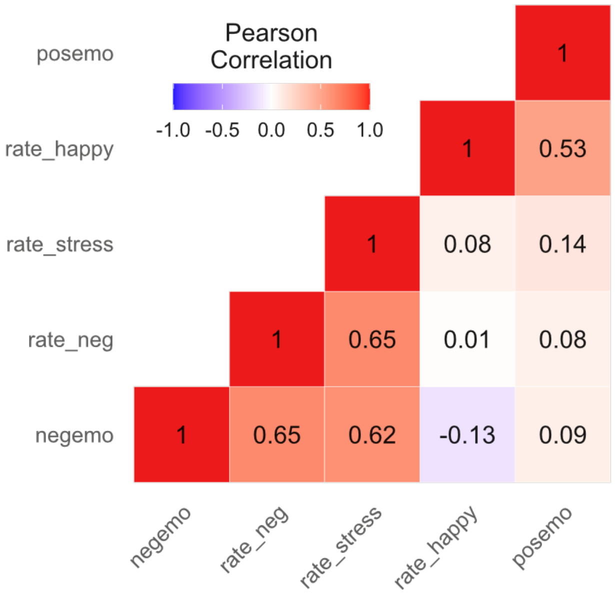


Figure 28: Correlation matrix as a heatmap showing correlations between outcome variables

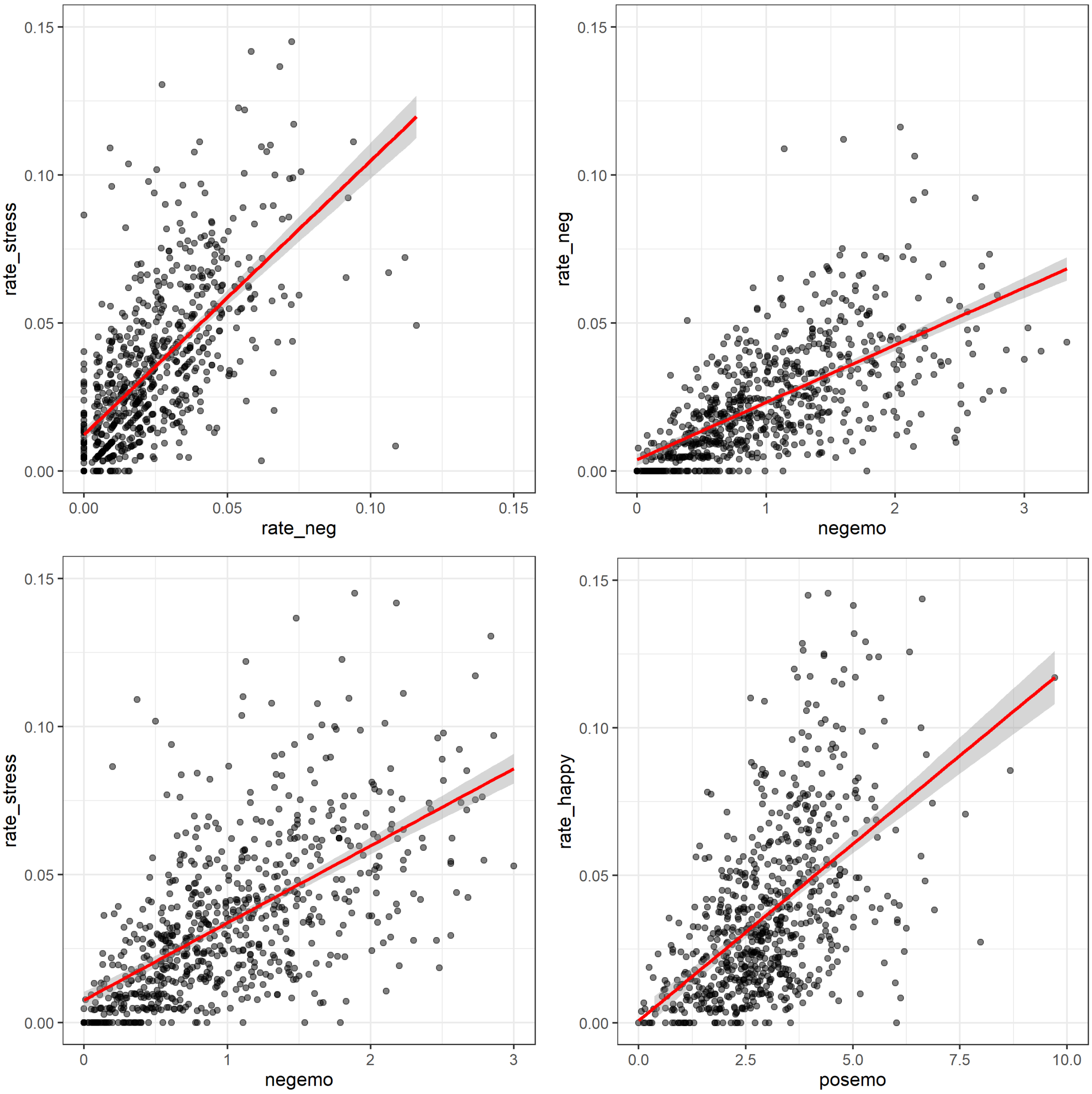


Figure 29: Scatterplots of pairs of outcome variables with the highest PCCs.

6.2 Representativeness of Subset

In this section, the representativeness of the subset users$_{regression}$ is assessed, by comparing the variable distributions to the distributions of the dataset users$_{all}$. The distributions are compared using violin plots. The red point in the violin plots marks the mean of the variable, and the blue point marks the median. Some plots were cut at the y-axis, to focus on the range where most data points lie within.

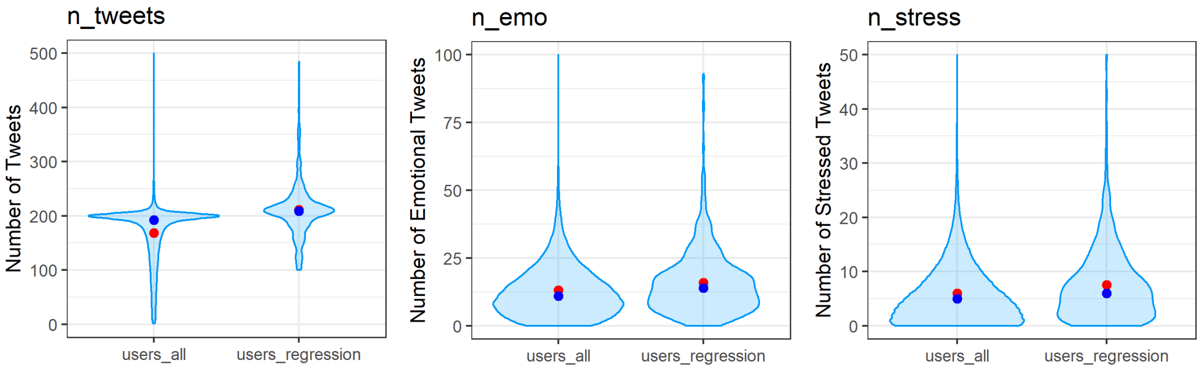


Figure 30: Violin plots comparing the distributions of the number of tweets (*n_tweets*), number of emotional tweets (*n_emo*), and number of stressed tweets (*n_stress*) per user.

The variable *n_tweets* of users$_{all}$ has a lower mean than in users$_{regression}$, due to the exclusion of users with less than 100 tweets in users$_{regression}$. Also, the distribution of users$_{all}$ is steeper around the mode, and the median is slightly lower. The variable *n_emo* has slightly higher mean and median values in users$_{regression}$, and roughly between the values 5 and 20, the distribution is more uniform. For the variable *n_stress*, the distributions as well as the differences in the distributions are very similar to the variable *n_emo*.

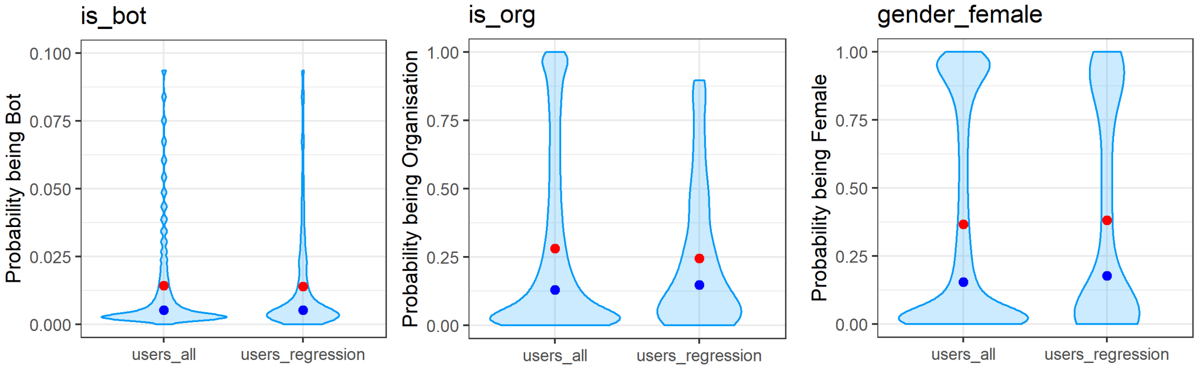


Figure 31: Violin plots comparing the distributions of the probabilities of the user being an automated account (is_bot), an account managed by an organisation (is_org), or being female (gender_female).

For *is_bot*, the main difference in the distributions is towards 0, where users$_{all}$ has a higher mode with a steeper slope. The variable *is_org* has a lower mean and a higher median in users$_{regression}$ due to the cut off at 0.9. Further, users$_{regression}$ shows a less steep slope towards 0. When comparing the variable *gender_female* between both datasets, it can be observed that the bimodal distribution is less extreme for users$_{regression}$. The mass in the two modes is partially shifted towards the center.

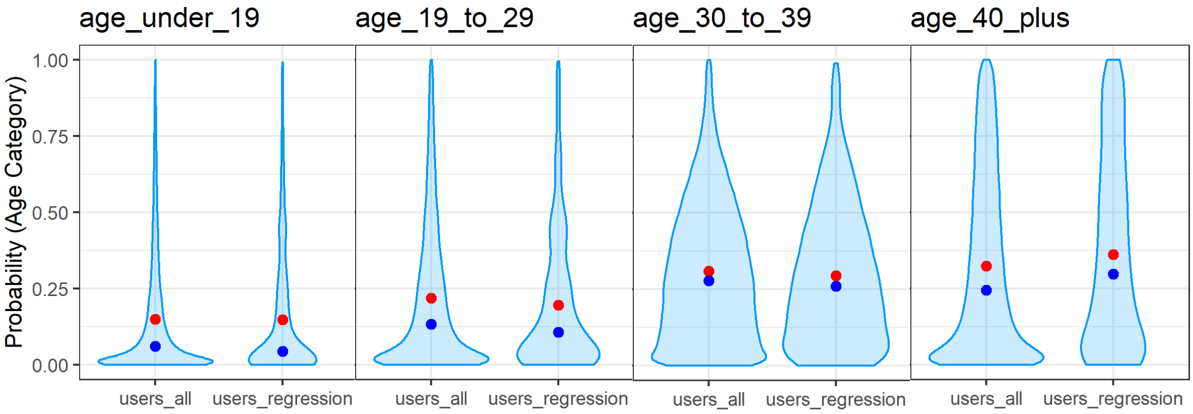


Figure 32: Violin plots comparing the distributions of the probabilities of the user belonging to the respective age categories.

The variable *age_under_19* shows a slightly lower median in users$_{\text{regression}}$. In age_19_to_29, both mean and median have a slightly lower value in users$_{\text{regression}}$. The same applies for the variable age_30_to_39. Finally, in age_40_plus a both higher mean and median can be observed in users$_{\text{regression}}$. Allthough the distributions in the four age category variables look relatively similar in both datasets, generally speaking, the slope towards 0 is less steep in all four variables.

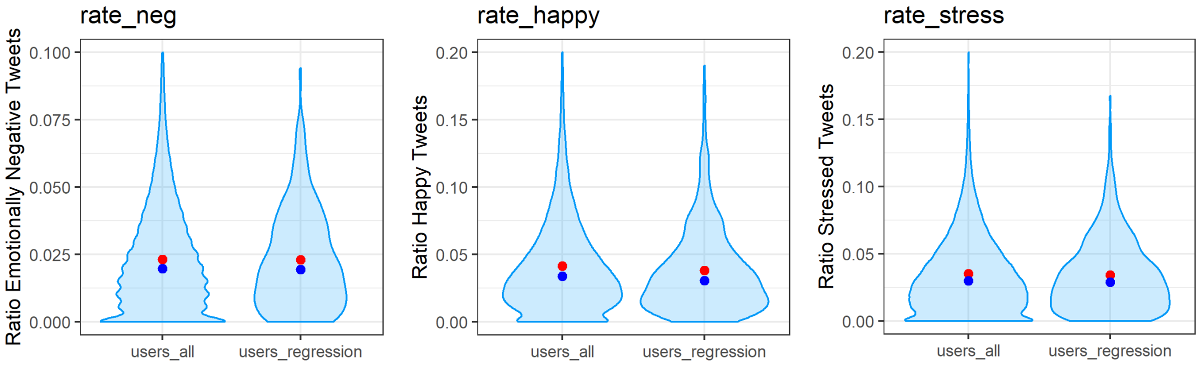


Figure 33: Violin plots comparing the rates of emotionally negative tweets (*rate_neg*), tweets expressing the emotion happiness (*rate_happy*), and tweets indicating stress (*rate_stress*).

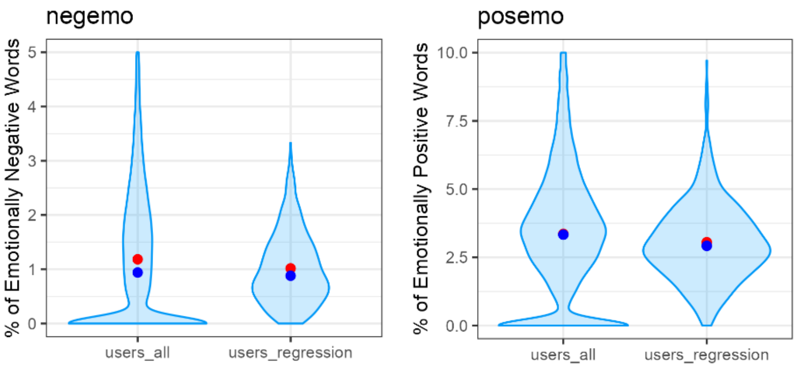


Figure 34: Violin plots comparing the percentages of words in all tweets for each user expressing negative emotions (*negemo*), and positive emotions (*posemo*).

For all three outcome variables *rate_neg*, *rate_happy*, and *rate_stress*, the distributions between the two datasets look very similar. However, one difference which can be observed in all three variables, is the monotonous decrease between the mode and 0 in $\text{users}_{\text{regression}}$, while in $\text{users}_{\text{all}}$, the distribution has a turning point and increases again towards 0. The differences in mean and median are barely observable in all three plots.

For the two outcome variables *negemo* and *posemo*, the differences are quite large, especially for *negemo*. The violin plots of $\text{users}_{\text{all}}$ show an accumulation of values at or near 0. The differences in the means are also well observable.

Additionally, a two-sample t-test to assess the difference in means of the two datasets for each of the compared variables was carried out. It was once tested between $\text{users}_{\text{regression}}$ and $\text{users}_{\text{all}}$, and once between $\text{users}_{\text{regression}}$ and $\text{users}_{\text{filtered}}$. The latter one is the same dataset as $\text{users}_{\text{all}}$, but without users having less than 100 tweets, or a probability higher than 0.5 of being a bot, or a probability higher than 0.9 of being an organisational account. In short, the same exclusion criteria were applied as for the dataset $\text{users}_{\text{regression}}$. *Table 12* shows the mean values of each variable for all three datasets, and the p-values for each variable, once for each of the two compared dataset pairs.

The p-values for the differences in means between $\text{users}_{\text{regression}}$ and $\text{users}_{\text{all}}$ are all under a significance level of 5%, except for the variables *gender_female*, *age_under_19*, *age_30_to_39*, *rate_neg*, and *rate_stress*. When calculating the p-values for the differences in means between $\text{users}_{\text{regression}}$ and $\text{users}_{\text{filtered}}$, all values show a significance level below 5%, expect for *is_bot*, *gender_female*, *age_under_19*, and *rate_neg*. Additionally, the variable *rate_stress* has a p-value of 0.011 which is just over the significance level of 1%.

Table 12: Mean values of the three compared datasets and p-values for difference in means.

variable	mean			p-value	
	$\text{users}_{\text{all}}$	$\text{users}_{\text{regression}}$	$\text{users}_{\text{filtered}}$	$\text{users}_{\text{all}}$	$\text{users}_{\text{filtered}}$
n_tweets	169.69	236.48	185.85	$2.2 * 10^{-16}$	$1.3 * 10^{-13}$
n_emo	13.16	16.40	14.60	$3.1 * 10^{-9}$	0.00088
n_stress	6.07	8.14	6.82	$3.0 * 10^{-6}$	0.0028
is_bot	0.043	0.027	0.026	$3.8 * 10^{-13}$	0.556
is_org	0.281	0.245	0.214	0.00016	0.0011
gender_female	0.366	0.381	0.373	0.294	0.535
age_under_19	0.149	0.148	0.145	0.816	0.715
age_19_to_29	0.217	0.195	0.229	0.0083	$6.3 * 10^{-5}$
age_30_to_39	0.308	0.294	0.325	0.088	0.00018
age_40_plus	0.324	0.362	0.300	0.00094	$7.8 * 10^{-8}$
rate_neg	0.024	0.023	0.025	0.634	0.068
rate_happy	0.043	0.039	0.044	0.0088	0.0021
rate_stress	0.035	0.034	0.037	0.137	0.011

6.3 Regression Models

In this section, the regression models are introduced, and the resulting outputs are described. For each of the five dependent variables, two regression models were produced. One with traffic noise during day time as an explanatory variable, and one with traffic noise during night.

Since the R function *stepAIC()* was used to fit the best regression model for each dependent variable, the number of explanatory variables may differ among the dependent variables. Some explanatory variables were automatically excluded from the initial model as they did not contribute to an improvement of the model. Furthermore, since the four age category variables have the tendency to correlate with each other, as seen in *Subsection 6.1.3*, it was decided to only include the variable *age_under_19* into the models. This variable, representing the youngest age category, tended to contribute the most to the models.

In the *Subsections 6.3.1* and *6.3.2*, the equations of the five different logistic regression models are noted down. To improve readability, the variable names have been replaced with more convenient names. Interaction terms indicating the interaction between two variables are noted down with a multiplication cross (×). In the *Tables 13* and *14*, the regression model outputs are listed, indicating the coefficient values with their standard errors in brackets, and whether the independent variables are significant. The significance of an independent variable is marked with one star (*) if its p-value is under 5%, two stars (**) if its p-value is under 1%, and three stars (***) if its p-value is under 0.1%.

6.3.1 Regression Models with Day Noise

$$
\begin{aligned}
&\mathit{EMOTIVE\ Negative\ Emotions} \\
&\quad = \beta_0 + \beta_1 * \mathit{Gender\ Female} + \beta_2 * \mathit{Age\ under}\ 19 + \beta_3 \\
&\quad * \mathit{Typology\ Urban} + \beta_4 * \mathit{SEP} + \beta_5 \\
&\quad * \mathit{Gender\ Female} \times \mathit{Typology\ Urban} + \beta_6 * \mathit{Age\ under}\ 19 \times \mathit{SEP}
\end{aligned}
$$

$$
\begin{aligned}
&\mathit{EMOTIVE\ Happiness} \\
&\quad = \beta_0 + \beta_1 * \mathit{Gender\ Female} + \beta_2 * \mathit{Age\ under}\ 19 + \beta_3 \\
&\quad * \mathit{Typology\ Urban} + \beta_4 * \mathit{Greenspace} + \beta_5 * \mathit{Day\ Noise} + \beta_6 * \mathit{SEP} \\
&\quad + \beta_7 * \mathit{Gender\ Female} \times \mathit{Age\ under}\ 19 + \beta_8 \\
&\quad * \mathit{Gender\ Female} \times \mathit{Typology\ Urban} + \beta_9 * \mathit{Gender\ Female} \times \mathit{Day\ Noise} \\
&\quad + \beta_{10} * \mathit{Typology\ Urban} \times \mathit{Day\ Noise} + \beta_{11} * \mathit{Day\ Noise} \times \mathit{SEP}
\end{aligned}
$$

$$
\begin{aligned}
&\mathit{Stresscapes\ Stress} \\
&\quad = \beta_0 + \beta_1 * \mathit{Gender\ Female} + \beta_2 * \mathit{Age\ under}\ 19 + \beta_3 \\
&\quad * \mathit{Typology\ Urban} + \beta_4 * \mathit{Gender\ Female} \times \mathit{Age\ under}\ 19
\end{aligned}
$$

$$\begin{aligned} LIWC\ Negative\ Emotions \\ &= \beta_0 + \beta_1 * Gender\ Female + \beta_2 * Age\ under\ 19 + \beta_3 \\ &* Typology\ Urban + \beta_4 * Greenspace + \beta_5 * Day\ Noise + \beta_6 * SEP \\ &+ \beta_7 * Age\ under\ 19 \times Typology\ Urban + \beta_8 * Greenspace \times Day\ Noise \end{aligned}$$

$$\begin{aligned} LIWC\ Positive\ Emotions \\ &= \beta_0 + \beta_1 * Gender\ Female + \beta_2 * Age\ under\ 19 + \beta_3 \\ &* Typology\ Urban + \beta_4 * Day\ Noise + \beta_5 \\ &* Gender\ Female \times Age\ under\ 19 + \beta_6 * Gender\ Female \times Day\ Noise \end{aligned}$$

For the model with *EMOTIVE Negative Emotions* as the dependent variable, the following independent variables are significant: *Gender Female* with a coefficient of 1.022, and a standard error of 0.499. *Age under 19* with a coefficient of 4.322, and a standard error of 2.014. And finally, *Typology Urban* with a coefficient of 0.706, and a standard error of 0.293. In other words, all three control variables are significant, while all three neighbourhood variables are not.

For the model with *EMOTIVE Happiness* as the dependent variable, the following independent variables are significant: *Typology Urban* with a coefficient of 2.524, and a standard error of 1.261. *Greenspace* with a coefficient of -0.142, and a standard error of 0.058. *Day Noise* with a coefficient of -0.819, and a standard error of 0.409. The interaction term *Gender Female* × *Age under 19* is very significant (**) with a coefficient of -3.154, and a standard error of 1.146. And finally, the interaction term *Gender Female* × *Typology Urban* with a coefficient of -1.352, and a standard error of 0.630.

For the model with *Stresscapes Stress* as the dependent variable, none of the independent variables are significant.

For the model with *LIWC Negative Emotions* as the dependent variable, only the independent variable *Age under 19* is significant. However, the variable is very significant (**) with a coefficient of 2.623, and a standard error of 0.998.

For the model with *LIWC Positive Emotions* as the dependent variable, the following independent variables are significant: *Gender Female* is very significant (**) with a coefficient of 4.750, and a standard error of 1.459. The interaction term *Gender Female* × *Age under 19* with a coefficient of -2.651, and a standard error of 1.121. And finally, the interaction term *Gender Female* × *Day Noise* with a coefficient of -0.445, and a standard error of 0.176.

Table 13: Regression results for models of all 5 outcome variables including daytime traffic noise

Regression Results - Day noise

	Dependent variable:				
	EMOTIVE Negative Emotions	EMOTIVE Happy	Stresscapes Stress	LIWC Negative Emotions	LIWC Positive Emotions
	(1)	(2)	(3)	(4)	(5)
Gender Female	1.022* (0.499)	-0.612 (1.475)	0.375 (0.235)	-0.087 (0.198)	4.750** (1.459)
Age under 19	4.322* (2.014)	0.225 (0.680)	-0.365 (0.668)	2.623** (0.998)	0.320 (0.668)
Typology Urban	0.706* (0.293)	2.524* (1.261)	-0.054 (0.205)	0.435 (0.277)	0.134 (0.224)
Greenspace		-0.142* (0.058)		-0.229 (0.271)	
Day Noise		-0.819* (0.409)		-0.154 (0.168)	-0.036 (0.074)
SEP	0.161 (0.092)	-0.807 (0.494)		0.115 (0.078)	
Gender Female × Age under 19		-3.154** (1.146)	2.180 (1.158)		-2.651* (1.121)
Gender Female × Typology Urban	-0.928 (0.538)	-1.352* (0.630)			
Age under 19 × Typology Urban				-1.478 (1.081)	
Gender Female × Day Noise		0.349 (0.198)			-0.445* (0.176)
Age under 19 × SEP	-0.700 (0.359)				
Typology Urban × Day Noise		-0.277 (0.166)			
Greenspace × Day Noise				0.044 (0.031)	
Day Noise × SEP		0.116 (0.063)			
Constant	-1.408* (0.592)	6.137 (3.185)	-0.007 (0.211)	-0.135 (1.513)	-0.076 (0.573)
Observations	733	733	733	733	733
Log Likelihood	-496.007	-481.694	-494.703	-490.821	-489.045
Akaike Inf. Crit.	1,006.014	987.388	999.406	999.643	992.090

6.3.2 Regression Models with Night Noise

$$EMOTIVE\ Negative\ Emotions = \beta_0 + \beta_1 * Gender\ Female + \beta_2 * Age\ under\ 19 + \beta_3 * Typology\ Urban + \beta_4 * Greenspace + \beta_5 * Night\ Noise + \beta_6 * SEP + \beta_7 * Gender\ Female \times Typology\ Urban + \beta_8 * Age\ under\ 19 \times SEP + \beta_9 * Greenspace \times Night\ Noise$$

$$EMOTIVE\ Happiness = \beta_0 + \beta_1 * Gender\ Female + \beta_2 * Age\ under\ 19 + \beta_3 * Typology\ Urban + \beta_4 * Greenspace + \beta_5 * Night\ Noise + \beta_6 * SEP + \beta_7 * Gender\ Female \times Age\ under\ 19 + \beta_8 * Gender\ Female \times Typology\ Urban + \beta_9 * Age\ under\ 19 \times SEP$$

$$Stresscapes\ Stress = \beta_0 + \beta_1 * Gender\ Female + \beta_2 * Age\ under\ 19 + \beta_3 * Typology\ Urban + \beta_4 * Gender\ Female \times Age\ under\ 19$$

$$LIWC\ Negative\ Emotions = \beta_0 + \beta_1 * Gender\ Female + \beta_2 * Age\ under\ 19 + \beta_3 * Typology\ Urban + \beta_4 * Greenspace + \beta_5 * Night\ Noise + \beta_6 * Age\ under\ 19 \times Night\ Noise$$

$$LIWC\ Positive\ Emotions = \beta_0 + \beta_1 * Gender\ Female + \beta_2 * Age\ under\ 19 + \beta_3 * Typology\ Urban + \beta_4 * Night\ Noise + \beta_5 * Gender\ Female \times Age\ under\ 19 + \beta_6 * Gender\ Female \times Night\ Noise$$

For the model with *EMOTIVE Negative Emotions* as the dependent variable, the following independent variables are significant: *Gender Female* with a coefficient of 1.060, and a standard error of 0.502. *Age under 19* with a coefficient of 4.365, and a standard error of 2.040. And finally, *Typology Urban* is very significant (**) with a coefficient of 0.860, and a standard error of 0.314. Again, same as in the model using day time noise, all three control variables are significant, while all three neighbourhood variables are not.

For the model with *EMOTIVE Happiness* as the dependent variable, the following independent variables are significant: *Gender Female* is highly significant (***) with a coefficient of 1.941, and a standard error of 0.562. *Greenspace* with a coefficient of -0.101, and a standard error of 0.050. *Night Noise* is very significant (**) with a coefficient of -0.183, and a standard error of 0.060. And finally, the interaction term *Gender Female* × *Age under 19* is very significant (**) with a coefficient of -2.999, and a standard error of 1.145.

For the model with *Stresscapes Stress* as the dependent variable, none of the independent variables are significant. The model is identical with the one with *Stresscapes Stress* as dependent variable in *Subsection 6.3.1*.

For the model with *LIWC Negative Emotions* as the dependent variable, the following two independent variables are significant. Firstly, *Age under 19* with a coefficient of 5.250, and a standard error of 2.121. Secondly, *Greenspace* is very significant (**) with a coefficient of 0.136, and a standard error of 0.049.

For the model with *LIWC Positive Emotions* as the dependent variable, the following independent variables are significant: *Gender Female* is highly significant (***) with a coefficient of 3.250, and a standard error of 0.959. The interaction term *Gender Female* $\times$ *Age under 19* with a coefficient of -2.675, and a standard error of 1.118. And finally, the interaction term *Gender Female* $\times$ *Night Noise* with a coefficient of -0.333, and a standard error of 0.145.

Table 14: Regression results for models of all 5 outcome variables including traffic noise during night

Regression Results - Night noise

	Dependent variable:				
	EMOTIVE Negative Emotions	EMOTIVE Happy	Stresscapes Stress	LIWC Negative Emotions	LIWC Positive Emotions
	(1)	(2)	(3)	(4)	(5)
Gender Female	1.060* (0.502)	1.941*** (0.562)	0.375 (0.235)	-0.100 (0.198)	3.250*** (0.959)
Age under 19	4.365* (2.040)	2.935 (2.151)	-0.365 (0.668)	5.250* (2.121)	0.320 (0.666)
Typology Urban	0.860** (0.314)	0.334 (0.313)	-0.054 (0.205)	0.252 (0.229)	0.026 (0.213)
Greenspace	0.306 (0.172)	-0.101* (0.050)		0.136** (0.049)	
Night Noise	0.193 (0.121)	-0.183** (0.060)		0.103 (0.065)	-0.020 (0.069)
SEP	0.146 (0.093)	0.147 (0.095)			
Gender Female × Age under 19		-2.999** (1.145)	2.180 (1.158)		-2.675* (1.118)
Gender Female × Typology Urban	-0.985 (0.541)	-1.002 (0.571)			
Gender Female × Night Noise					-0.333* (0.145)
Age under 19 × Night Noise				-0.598 (0.322)	
Age under 19 × SEP	-0.711 (0.364)	-0.516 (0.378)			
Greenspace × Night Noise	-0.038 (0.026)				
Constant	-2.939** (1.018)	0.304 (0.747)	-0.007 (0.211)	-1.338* (0.539)	-0.155 (0.433)
Observations	733	733	733	733	733
Log Likelihood	-493.967	-485.637	-494.703	-492.198	-490.173
Akaike Inf. Crit.	1,007.934	991.273	999.406	998.395	994.347

6.3.3 Effect Size of Significant Neighbourhood Variables

In the following, the effect size of the significant explanatory neighbourhood variables are described by reporting the odds ratios with their corresponding 95% Confidence Intervals (CI).

The odds ratios were calculated by e^{β}, where β is the coefficient value (log-odds) of the predictor variable. The 95% confidence intervals were calculated by using the formula

$e^{(\beta +/- 1.96 * SE)}$, where SE is the standard error of the coefficient β.

Models including Day Noise

It was found that, holding all other predictor variables constant, the odds of a user having an above-median rate of Happiness-tweets decreases by 13.2% (95% CI [0.028, 0.226]) for a one -unit increase in the predictor variable *Greenspace*.

It was also found that, holding all other predictor variables constant, the odds of a user having an above-median rate of Happiness-tweets decreases by 55.9% (95% CI [0.017, 0.802]) for a one -unit increase in the predictor variable *Day Noise*.

Models including Night Noise

It was found that, holding all other predictor variables constant, the odds of a user having an above-median rate of “Happiness”-tweets decreases by 9.6% (95% CI [0.003, 0.18]) for a one-unit increase in the predictor variable *Greenspace*.

Further, it was found that, holding all other predictor variables constant, the odds of a user having an above-median rate of “Happiness”-tweets decreases by 16.7% (95% CI [0.063, 0.26]) for a one-unit increase in the predictor variable *Night Noise*.

Finally, it was found that, holding all other predictor variables constant, the odds of a user tweeting an above-median percentage of emotionally negative words increases by 14.6% (95% CI [0.041, 0.261]) for a one-unit increase in the predictor variable *Greenspace*.

Chapter 7 | Discussion

The obtained results and the underlying data and methodological approaches are discussed in depth in this chapter. First, the Twitter variables are examined. Then, the generated neighbourhood variables are discussed. The representativeness of the subset used for the regression analysis is elaborated in third section. Finally, the approach and the results of the multiple logistic regressions is elaborated.

7.1 Twitter Variables

The Twitter variables of the subset $\text{users}_{\text{regression}}$ are discussed in this section. The first subsection interprets the found results, the second subsection outlines the uncertainties and limitations of the variables, and the third subsection reflects the approaches and suggestions for improvements are made.

7.1.1 Interpretation of Results

In this subsection, the characteristics of the variables in the dataset $\text{users}_{\text{regression}}$ are interpreted and discussed, which were visualised and described in *Subsection 6.1.2*.

for the variable *n_tweets*, the mean is 236.5 and the median is 210, which means that on average, loosely speaking just over 200 tweets per user were available for the analysis. This may initially seem to be a reasonable number of tweets. However, when considering that on average, only around 16 tweets are found to be emotional according to EMOTIVE, and only around 8 show indications of stress according to Stresscapes, the situation changes. Obviously, this fact directly transfers to the low average values found in the outcome variables *rate_neg*, *rate_happy*, and *rate_stress*. The average ratio between tweets expressing negative emotions and the total number of tweets is around 2%. The emotion happy, on average, is found in roughly 4% of all tweets, and indications of stress, on average, are found in around 3% of all tweets. The outcome variables *negemo* and *posemo* generated with LIWC, show similarly low rates in emotional content. The variable *negemo* shows that on average, about 1% of all words in the tweets are associated with negative emotions. The rate of words associated with positive emotions is around 3% according to *posemo*. These low rates in emotionality expressed in tweets demand a high number of total tweets per user, in order to be sufficiently statistically meaningful. For example, if the average number of total tweets per user is 50, the average number of tweets expressing negative emotions would be 1. Moreover, the number of users having randomly 0 tweets expressing negative emotions, would most probably be quite high. Vice versa, users having randomly 2 tweets expressing negative emotions, would already have a rate twice as high as the average. This example shows, how the standard error would strongly influence the statistical significance.

The distribution of the variable *is_bot* shows that the probability of unfavourably having included automated inauthentic accounts in the analysis is very small, which is also reflected in the low average probability of around 3%.

For the variable *is_org*, representing the probability of the account being managed by an organisation, the average probability is around 25%. In other words, one out of four users could potentially be an account managed by an organisation. This fact most likely causes a not negligible bias on the analysis.

The variable *gender_female*, which represents the probability of the user having the female gender, shows a bimodal distribution, with the two modes towards 0 and 1 respectively. However, the mode towards 0 (most likely male) is significantly larger than the one towards 1 (most likely female). This is also reflected in the mean value being 0.382. The mean value could be interpreted as the proportion of female users, which would be 38.2%.

The four age category variables *age_under_19*, *age_19_to_29*, *age_30_to_39*, and *age_40_plus*, derived with the M3-method, each represent the probability of the user belonging to the respective age category. On average, the probability of a user being under 19 years old is around 15%, being 19 to 29 years old is around 20%, being 30 to 39 years old is around 29%, and being 40 or older is around 36%. Expressed in this way, it does not seem very intuitive. The probabilities could however be interpreted as the proportions of users belonging to the respective age category. In this manner, it can be implied that the users are not evenly distributed among the four age categories, but the majority of the users probably belong to the older two categories, that is, around 2 out of 3 users.

7.1.2 Uncertainties and Limitations

The accuracy of the variables firstly depends on the accuracy of the analysis performed by the underlying Algorithms. This adds the first factor of uncertainty, which is however very difficult to quantify. Secondly, the variables derived with the M3-method (age categories, *gender_female*, *is_org*), as well as *is_bot*, derived with Botometer, hold probabilities rather than categorical values. It certainly brings some advantages; however, a probability is an uncertainty by its definition.

Moreover, as introduced in *Subsection 4.1.2*, the NLP-algorithms EMOTIVE and Stresscapes not only detect tweets containing emotions or indicating stress, but they also assign an integer value ranging up to 20 depending on the variable. This value is a measure for the amplitude of the found sentiment, which could be used as a weight. A weighting of the sentiment using this value was however not performed, due to its non-trivial nature, when it comes to the aggregation of all tweets onto the user level. By choosing to do so, potentially important information was not included into the resulting variables, which may distort the analysis by an additional factor.

A further limiting factor leading to uncertainty is the combination of the 6 emotions considered as negative (sadness, anger, disgust, confusion, fear, shame) into one variable to represent all negative emotions. The reason for this step was the fact that the frequency of the single negative emotions tends to be very small. In this manner, the statistical power could be increased compared to taking only the emotion sadness, for example. However, this aggregation of emotions eliminates the meaning of the single emotion. This procedure assumes that confusion is equally negative than sadness or any other of the 6 emotions. Although the resulting bias is very abstract and basically impossible to quantify, it adds a further factor of uncertainty to the analysis.

Additionally, as described in the previous subsection, the average probability of a user being an account managed by an organisation is about 25%. This imposes further uncertainty in the analysis, as there may be a reasonably high number of accounts which are falsely handled as if they were real persons.

7.1.3 Reflections

The threshold for excluding users possibly being bots above a certain probability, was set at 50%. Choosing this threshold, the average probability of a user being an automated account is about 3% which seems very solid. However, the threshold could arguably be chosen even lower, at 25% or even 20%, to further lower the chance of falsely including bots in the analysis. The justification for this optional adjustment is that only few users would additionally be excluded.

As for the threshold excluding users possibly being organisational accounts, it should definitely be set significantly lower than at 90%. However, because of the relatively small sample size of initially roughly 1200 users, a compromise hat to be found. By setting the threshold to 50%, for example, a large part of the users would be excluded, which drastically reduces the sample size. As a consequence, in order to set the threshold lower, a much larger initial sample size is required, which enables to finally keep a reasonable number of users for the analysis.

A further improvement could possibly be done for the three outcome variables *rate_neg*, *rate_happy*, and *rate_stress*. As mentioned in the previous subsection, the magnitude of a sentiment was not considered when creating the outcome variables. The resulting information loss should probably be avoided as far as possible, by introducing a weighting approach.

7.2 Neighbourhood Variables

In this section, the generated neighbourhood variables are discussed. The results are interpreted in the first subsection, then, the numerous uncertainties and limitations are discussed in the second subsection. In the third subsection, ides for possible improvements are pointed out.

7.2.1 Interpretation of Results

The interpretation of the generated neighbourhood variables is generally very difficult, because of their abstract nature. Nevertheless, in the following the most important characteristics of the generated variables and their distributions are discussed and interpreted as far as possible.

The variable *greenspace* represents greenspace within a neighbourhood and is certainly the most abstract of the neighbourhood variables. It stands for the amount, the quality, and the accessibility of greenspace within a radius of 500 meters around the homeplace location and is inversely weighted with increasing distance. It is an approach to measure whether a user lives in a rather “green” environment or in a less “green” environment, expressed with a continuous index ranging from 0 to 10. The distribution has the tendency of a uniform distribution roughly between 0 and 7.5. Only a few users seem to live in “very green” neighbourhoods, having a value over 7.5. The fact that the mass of the distribution stretches over the largest part of its extent (standard deviation of 1.86), tells that there seem to be large differences in the availability, quality, and accessibility of greenspace between the users. If the differences were to be very small, the potential effect of the variable on the outcome variable would presumably also be smaller and less significant. The shape of the distribution means that no further transformation is required, which is a positive aspect of the variable.

The variable *day_noise* represents the perceived traffic noise during daytime within a neighbourhood. More accurately, it is the average traffic noise in dB perceived in the neighbourhood, inversely weighted with increasing distance to the homeplace location, between 6 am and 10 pm. Furthermore, the values were rescaled to lie between 0 and 10. Obviously, the values cannot be interpreted as dB anymore, due to the underlying averaging, weighting and rescaling processes. Instead, values towards 0 mean that the respective user lives in a neighbourhood with low traffic noise exposure during daytime, whereas values towards 10 mean high traffic noise exposure. The histogram shows an approximately bell-shaped distribution between 5 and 10, while there are only few values below 5. This would mean that the vast majority of the analysed users live in neighbourhoods with rather high traffic noise exposure during daytime. This may be plausible, given that the largest part of the users apparently live in urban communes. However, the distribution may not be optimal due to its high concentration of values in the upper end, leaving almost none in the lower half.

The variable *night_noise* represents the perceived traffic noise during night at a homeplace location. The values are calculated differently as for *day_noise*, by simply taking the average traffic noise between 10 pm and 6 am, within a 100 meters radius around the homeplace location. The resulting values were again rescaled to lie between 0 and 10. The distribution has a bell-shaped tendency between 2.5 and 10. Therefore, the variable leads to the interpretation that the majority of the users has a moderate to rather high exposure to traffic noise during night.

Finally, the variable *sep* represents the socio-economic position within a neighbourhood. The values were calculated by inversely weighting the rasterised SEP values (see *Subsection 5.2.3*) with increasing distance to the homeplace location, and then taking the average. A rescaling of the values restricted the range between 0 and 10. As the original SEP values are already index values, the interpretation of the generated variable *sep* does not change. Values towards 10 mean that the neighbourhood is resided by persons with high SEP, values towards 0 mean the opposite. The distribution has a strong tendency to a normal distribution, which was to be expected, since the underlying SEP data is also normally distributed.

7.2.2 Uncertainties and Limitations

The generated neighbourhood variables largely contribute to the whole sum of uncertainties in the analysis. Not only do the original datasets, on which the variables are based, contain uncertainties and inaccuracies, but so do the approaches to estimate the value representing a neighbourhood characteristic.

The two main additional factors contributing to uncertainty, are on the one hand the definition of the neighbourhood extent, and on the other hand the choice of the distance decay function. The definition of the neighbourhood extent is equal for every homeplace location. Individually defining the extents depending on certain factors would by far be too elaborate for the scope of this thesis, although it would enable to better represent the reality. By choosing the same definition for every homeplace location, it is assumed that every user has the exact same activity space, which is obviously not true. Moreover, the linear distance decay to account for the fact that nearer phenomena have a stronger influence on a subject than more distant ones, may be the less suitable approach than, for example, an exponential distance decay.

The variable *greenspace* is based on the greenspace dataset introduced in *Section 4.2*. The dataset was produced by including a number of different factors, all of them contributing to a final index value, which should represent the quality as well as the accessibility of greenspace for the entire extent of Switzerland. The dataset has two main limitations. The first one is the fact that it tries to combine both quality and accessibility into one variable. This makes the resulting variable much harder to interpret as if the two factors were held separately. And the second one is the fact that it was made for both urban and rural areas, instead of just focusing on urban areas. The effect of available greenspace on mental health is generally studied in the context of urban areas, as rural areas are basically defined by the high presence of greenspace. Because the perception of the availability and quality of greenspace may strongly differ between citizens living in rural villages and those living in cities, two different in indices would probably be more appropriate.

The two variables *day_noise* and *night_noise* have their origin in the datasets introduced in *Section 4.3*. These variables also have certain limitations. An important one to mention concerns the suitability of the data for an analysis of such a spatial scale as in this thesis. The BAFU, which provides the datasets, states that the data is intended for the use on a national or regional scale. They point out, that the data may not be very suitable for local analysis. As the noise values are only estimates, and not actual measurements, in some cases the local error may be quite large. It is discouraged to use the data to define noise exposure for single buildings. However, averaging noise values within a 100-meters or 500-meters radius respectively, is most likely less problematic. This directly leads to a further limitation, being the idea of averaging noise values in decibel (dB). As a linear increase in the unit decibel is not equal to a linear increase in noise, but a 10 dB increase means a doubling of sound volume, simply averaging noise levels within an area may not be a very appropriate approach. For example, in a raster of the dimensions 3x3 with only values of 30 dB, the average value is 30. Now replacing one cell value with 120 dB, which is equal to the noise emitted from a chain saw (Iac acoustics, 2022), the average noise value within that 3x3 raster is 40 dB, which is still perceived as very quiet. Of course, this is a very theoretical example and may never appear as such in the real world, also because the noise in one areal unit expands into other adjacent areal units. Still, it seems to be a factor contributing to uncertainty, which should be kept in mind.

7.2.3 Reflections

The uncertainties within the neighbourhood variables could be reduced by introducing the approaches discussed in this subsection.

The neighbourhood areas could be defined in a more sophisticated way, for example by using street networks, rivers and buildings to approximate the natural activity space of a user living in that neighbourhood. A similar approach, using street networks, was used to define neighbourhoods for the SEP dataset, as described in *Section 4.4*. In doing so, the area of the environment having influence on a user would potentially better correspond to reality. Obviously, this approach requires an automation of the process, which would pose a time-consuming challenge to implement.

As for the neighbourhood variables, improvements especially for the estimation of the variables representing greenspace and traffic noise could be made. Greenspace could be defined differently depending on the neighbourhood being in an urban or in a rural area. Focusing on urban areas, layers of higher resolution could be used to better capture the microstructure of greenspace related features, such as single trees. Additionally, quality and accessibility of greenspace could be stored separately in two different variables, which then would be added individually to the regression model. Regarding the representation of traffic noise, loud areas should probably have more weight in the calculation. This is due to the fact that the unit decibel seems not suitable for being averaged over larger areas with highly heterogenous noise levels, as explained in the previous subsection.

7.3 Representativeness of Subset

The representativeness of the subset users$_{\text{regression}}$ is interpreted in the following two subsections. First the spatial distribution is briefly discussed, and then the comparison of the distributions of the variables between the subset and users$_{\text{all}}$ is interpreted.

7.3.1 Spatial Distribution

Considering the fact, that Zürich is the largest city of Switzerland with a population of approximately 436’000 (Stadt Zürich, 2021), and Geneva is the second largest city with approximately 203’000 inhabitants (Bundesamt für Statistik, 2020), the spatial distribution of the analysed sample is quite strongly disproportionate to the population density when focusing on these two cities.

However, when comparing the spatial distribution of users$_{\text{regression}}$ with the spatial distribution of all geolocated tweets in Switzerland (tweets$_{\text{geoloc}}$), it can be observed, that persons in Geneva in fact tend to tweet a lot. Overall, besides a few exceptions, the spatial distributions of users$_{\text{regression}}$ and tweets$_{\text{geoloc}}$ are quite similar, making users$_{\text{regression}}$ a satisfying sample concerning the spatial representativeness of Twitter users.

7.3.2 Distributions of Variables

In *Section 6.2* the representativeness of the subset users$_{\text{regression}}$ was assessed, by visually comparing the distributions of the variables with those of the dataset users$_{\text{all}}$. Furthermore, a t-test was performed to assess the significance of the difference in means between the variables of both datasets.

The general tendency which can be observed in the distributions of all variables, is the steeper slope around the mode in the dataset users$_{\text{all}}$. This shape which reminds of the stem of a wine glass when using a violin plot, is mainly due to the much larger sample size of users$_{\text{all}}$, which leads to a smaller variance. A further noticeable difference in the distributions is the wider “footprint” in the dataset users$_{\text{all}}$ for the following variables: *n_emo*, *n_stress*, *rate_neg*, *rate_happy*, *rate_stress*. This is simply a consequence of the inclusion of users having less than 100 tweets, which increases the number of users having 0 tweets expressing emotions or stress. Obviously, also the two variables *n_tweets* and *is_org* show quite strong differences because of the thresholds at 100 and 0.9, excluding users below or above those values respectively in users$_{\text{regression}}$. However, other than that, the distributions seem to have fairly similar shapes.

The results of the t-tests, assessing the difference in means for each variable, revealed that the means of most variables are significantly different between the two datasets. However, for some variables it was to be expected, on the one hand, and on the other hand it is not very relevant at all. Because of the three thresholds set to exclude users depending on the number of tweets, probability of being a bot or an organisational account, the significantly different mean of variables affected by those thresholds can be explained. The most important variables, which are used in the regression analysis, being *gender_female* and *age_under_19* as control variables, do not show a significantly different mean. The outcome variables *rate_neg* and *rate_stress* also show a very similar mean between both datasets. The outcome variable *rate_happy* has a rather small p-value of 0.0088, which may be considered as being significant, depending on the choice of the significance level.

Overall, it can be concluded that the subset users$_{\text{regression}}$ is not an optimal, but arguably still a reasonable representation of the larger dataset users$_{\text{all}}$.

7.4 Regression Analysis

This final section discusses the results of the regression analysis. In the first subsection, associations found in the model outputs are interpreted, and the related uncertainties and limitations are discussed in the second subsection. In the last subsection, reflections about improvements are described.

7.4.1 Interpretation of Results

In this subsection, the significant variables in the models are interpreted and discussed. First, the regression results including daytime noise are elaborated, and then the regression results including night noise.

Regression Results – Day Noise

Gender Female shows significant positive associations with the outcome variables *EMOTIVE Negative Emotions* and *LIWC Positive Emotions*. This could be interpreted as a tendency of female Twitter users generally tweeting more emotional content than men. This tendency can be explained by the findings of Kring & Gordon (1998), who state that women are generally more emotionally expressive than men. Park et al. (2012) also find more positive and more negative emotions expressed in tweets by female Twitter users.

Age under 19 shows significant positive associations with the outcome variables *EMOTIVE Negative Emotions* and *LIWC Negative Emotions*. This leads to the assumption of adolescents expressing generally more negative emotions in tweets than older users. Bailen et al. (2018) reviewed relevant literature on adolescents' experience of four specific dimensions of emotion, being emotional frequency, intensity, instability, and clarity. Compared to adults, they find that adolescents experience greater emotional intensity, greater emotional instability, and more frequent high-intensity positive and negative emotions. Furthermore, it was shown that older adolescents, compared to younger adolescents, experience more frequent negative and less frequent positive emotions (Frost et al. 2015). The research department of the statistics platform Statista showed that in the United States in April 2018, 3% of teenagers aged 15 to 17 used Twitter, compared to only 1% of teenagers aged 13 to 14 (Statista, 2018). Although no literature was found on the Twitter user demographics of adolescents in Switzerland, it can be assumed that the proportions are similar. This would potentially explain the positive associations of *Age under 19* with the two outcome variables representing negative emotions.

Typology Urban shows significant positive associations with the outcome variables *EMOTIVE Negative Emotions* and *EMOTIVE Happiness*. Similarly, as in the case of the variable Gender Female, it suggests that Twitter users living in urban areas tend to more often express emotions in their tweets, than rural users. No literature was found to support this suggestion, however, due to a higher anonymity in cities, users may have less inhibition to publicly share their emotions via tweets. The lower anonymity in rural areas may lead to social stigma, when sharing too personal information online.

Greenspace shows significant negative associations with the outcome variable *EMOTIVE Happiness*. This finding suggests that the higher the availability and quality of greenspace in a neighbourhood, the less happiness is found in the users' tweets. Obviously, this is the opposite of what was to be expected, since greenspace has a great positive influence on mental well-being, as described in *Subsection 2.3.1*. A possible explanation for this could be, that the users living in rural areas, having higher Greenspace values, generally tend to express less emotions in tweets. And since *Greenspace* has a larger effect size on *EMOTIVE Happiness* than on *EMOTIVE Negative Emotions*, the unexpected negative association is only found in the former variable. Furthermore, the effect size of the negative association is quite small.

Same as *Greenspace*, *Day Noise* shows significant negative associations with the outcome variable *EMOTIVE Happiness*, but with a stronger effect size. It suggests that the higher the exposure to traffic noise during daytime in a neighbourhood, the less happiness is found in the users' tweets. As described in *Subsection 2.3.2*, traffic noise has a negative impact on mental well-being, which could explain the negative association between the two variables.

The interaction of *Gender Female* and *Age under 19* shows significant negative associations with the outcome variables *EMOTIVE Happiness* and *LIWC Positive Emotions*. It suggests that adolescent female users tend to post less tweets with content considered as emotionally positive. This could partially be explained by the findings of Frost et al. (2015) who state that older adolescents experience less frequent positive emotions.

The interaction of *Gender Female* and *Typology Urban* shows a significant negative association with the outcome variable *EMOTIVE Happiness*. It suggests that female users living in urban neighbourhoods tend to tweet less Tweets containing the emotion happiness. This result is rather counter intuitive and difficult to interpret, as both female users and users living in urban neighbourhoods tend to express more emotions in their tweets, as described earlier. The association is therefore failed to be explained.

Finally, the interaction of *Gender Female* and *Day Noise* shows a significant negative association with the outcome variable *LIWC Positive Emotions*. It suggests that female users, living in a neighbourhood with high exposure to traffic noise during daytime, tend to express less positive emotions in their tweets. As mentioned earlier, traffic noise has a negative impact on emotional well-being. Furthermore, female users tend to more frequently express emotions, and consequently also changes in the frequency of emotional expression become more strongly visible for female users. Therefore, a possible explanation could be, that additional impact on the reduced expression of positive emotions due to exposure to traffic noise can be found in female users.

Regression Results – Night Noise

When replacing Day Noise with Night Noise in the 5 different models, most models remain very similar, and most explanatory variables show very similar significant associations with the outcome variables. For this reason, only the most striking difference regarding significant associations between independent and dependent variables is described in the following:

Greenspace shows in addition to the significant negative association to *EMOTIVE Happiness*, a significant positive association to *LIWC Negative Emotions*. This additionally suggests that users living in “greener” neighbourhoods tend to express more negative emotions in their tweets. Again, this association is the opposite of what has been expected and cannot be explained with the same arguments as for the negative association with *EMOTIVE Happiness*. This leads to the question of what effects greenspace could have on users to make them express more negative emotions. One example that comes to mind, where greenspace potentially has a negative impact on health, is pollinosis, also known as “hay fever”. In Switzerland, it is estimated that one in four persons suffers from a pollen allergy, with an ever-increasing tendency (Universitätsspital Zürich, 2022). This example shows that greenspace may not only have a positive influence on health and consequently on mental health. The combination with other unknown factors of greenspace having a negative influence on (mental) well-being could ultimately lead to this unexpected result.

7.4.2 Uncertainties and Limitations

This subsection addresses the uncertainties and limitations of the regression analysis approach.

First, it must be recalled that the performance of the regression models can only be as good as the quality and the accuracy of the involved variables. As discussed in the *Sections 7.1* and *7.2*, the variables show a considerable number of uncertainties and intransparencies which directly lead to uncertainties in the models. Consequently, the interpretation of certain significant associations between independent and dependent variables become quite challenging. Hence, although certain explanatory variables seem to have significant associations with certain outcome variables, the true reason behind the association may be a whole different one then assumed.

The sentiments found in the users’ tweets, which are represented by the 5 outcome variables, are tried to be explained by environmental variables which are restricted to the users’ neighbourhood. This means that indirectly, all tweets of a user are somehow treated as if they were all posted within the boundaries of the defined neighbourhood, which does not correspond to reality. This is a clear limitation of the study, which was deliberately accepted, however, for two reasons. One, there were not sufficient geolocated tweets to conduct a classic hotspot analysis of emotions in space using spatial binning techniques or cluster analysis. Two, assessing the environmental variables greenspace, traffic noise and SEP for every single geolocated tweet location would have been much too costly, and presumably also rather pointless.

7.4.3 Reflections

After interpreting the regression analysis results and discussing the uncertainties and limitations of the chosen approaches, possible improvements are briefly introduced.

The choice of a binary logistic regression model, using a median split to generate the outcome variables, may not be the most suitable approach. When splitting the initially continuous outcome variables into the binary categorization of “above median” and “below median”, a large part of information gets lost. The extreme values at the upper and lower range suddenly get the same value as the values next to the median. An alternative approach to tackle this problem, would be the use of an ordinal logistic regression. For this purpose, the outcome variables would be split up into three to five categories, using the quantiles as thresholds. For example, using 5 categories, the rate in negative emotions would be classified into “very low”, “low”, “moderate”, “high”, and “very high”.

A further alternative approach worth considering, which is not specifically addressing the regression analysis, but rather the whole study design, is to focus only on cities. This has already been mentioned earlier in *Subsection 7.2.3*. When focusing solely on urban areas, the typology variable would no longer be needed, reducing the total number of variables in the regression models. Urban and rural differences in tweeting behaviour of users would no longer affect the regression results. And most importantly, the neighbourhood variables, especially greenspace could be modelled much more accurately to better fit the definition of urban greenspace.

Chapter 8 | Conclusion

This work is an attempt to link sentiments found in Twitter data with neighbourhood scale environmental characteristics. Literature suggests that sentiments expressed in social media may be indicative for mental well-being, which in turn is influenced by environmental factors. In this study, the major physical and societal influencing factors are represented by the availability of greenspace, exposure to traffic noise, and the socio-economic circumstances within a neighbourhood. Since generally only a small proportion of tweets are geolocated, the analysis was performed on a user-level, which allowed to also include tweets without coordinates. In this way, on average around 236 of the most recent tweets were used to characterize the user's rate of positive and negative emotions, and stress, by applying the three NLP-tools EMOTIVE, Stresscapes, and LIWC. With DBSCAN, the user's presumed homeplace location was identified, representing the overall spatial context to be linked to the sentiments expressed in the user's tweets.

In this thesis, the defined objectives were achieved along with secondary results providing useful additional insights. With regard to RO_1, being the selection of users based on specific criteria, the following was found: Out of 70'333 users, for 1213 users a homeplace location was identified using DBSCAN, which is a proportion of around 1.7%. The plausibility of these presumed homeplace locations was then assessed using two different datasets, which both approximate the spatial coverage of residential area in Switzerland. It showed that out of the 1213 users with a detected homeplace, 193 locations lied significantly outside the residential area. This means that at least 15.9% of the presumed homeplaces derived by the DBSCAN-approach are unplausible. Furthermore, the exclusion of users with high probabilities of being a bot (threshold set at 50%) or being an organisational account (threshold set at 90%), revealed that in all 70'333 users, the average probability was 4.3% for being a bot and 28.1% for being an organisation.

For RO_2, a workflow for the automated estimation of variables representing the neighbourhood characteristics was modelled in ArcGIS Pro and exported to Python scripts for optimization and parallel processing. A linear distance weighting approach was implemented in the workflow, to account for the effect of distance decay, considering that closer environmental characteristics have more influence than distant ones.

Moreover, concerning RO_3, multiple logistic regression analysis was applied, to assess possible associations between sentiments on user-level as outcomes and the neighbourhood variables generated for RO_2 as predictors, while including different control variables. The outcome variables indicating above- or below-median rates of positive and negative emotions, and stress, were created using the NLP-systems EMOTIVE, Stresscapes, and LIWC. The neighbourhood variables or environmental variables, taking continuous index values between 0 and 10, representing *greenspace*, *traffic noise*, once during daytime and once during night, and socio-economic position (*SEP*). As control variables, probabilities of the user being female and being of age under 19, as well as urban-rural differences were added to the models. Significant negative associations were found between *traffic noise* (both daytime and night) and *happiness* from EMOTIVE. Additionally, if the user tended to be female, traffic noise (both daytime and night) also had significant negative associations with *positive emotions* from LIWC. *Greenspace* was found to be significantly negatively associated with *happiness* from EMOTIVE and significantly positively associated with *negative emotions* from LIWC. Although having a rather small effect size, the found associations of *greenspace* with the outcome variables are the opposite of what was expected. These findings may partially be explained by factors like pollinosis, although the high uncertainties within the variable *greenspace* cannot be excluded as the root cause for the unexpected results. For *SEP*, no significant associations were found.

Still, probably the most important findings of the thesis are the numerous limitations and pitfalls of the pursued approaches and the possible alternatives and improvements for future work, which can be derived from it. Thus, to conclude the thesis, suggestions for future research are proposed in the following. The results showed that the proportion of emotional tweets is low, as well as the proportion of users for which a presumed homeplace location is detected. This implies the necessity of a large sample size for both number of users and number of tweets per user. Hence, Switzerland as the study area may be less suitable, as the United States or Japan, for instance, where Twitter is very popular. Furthermore, focusing solely on urban areas would bring the advantage of an urban-specific definition of greenspace, enabling a more accurate modelling of the variable and its associations with emotions and stress. Still, the general approach of considering any of the most recent tweets of a user to calculate its rate of sentimental tweets, no matter the location at the time of tweeting, may be the strongest limiting factor. It infers the assumption, that the environmental characteristic of the neighbourhood the user lives in, are associated also with sentiments found in tweets which were posted whilst not being anywhere near home. In other words, all tweets of a user are treated as if they were posted within the neighbourhood. Analysing environmental factors on a neighbourhood scale and at a user-level may therefore not be feasible, and alternative approaches on the spatial unit of urban districts or census tracts and on single tweet-level should be considered.

Overall, it can however be concluded that with further research, the spatial analysis of Twitter data could become a valuable method to assess the impact of environmental characteristics on mental well-being. It could potentially replace costly traditional surveys and would pose a very efficient alternative with large-scale applicability.

Literature

Auchincloss, AH, Gebreab SY, Mair C, and Diez Roux AV. 2012. "A Review of Spatial Methods in Epidemiology, 2000-2010." *Annual Review of Public Health* 33 (April): 107–22. https://doi.org/10.1146/ANNUREV-PUBLHEALTH-031811-124655.

Bafu, Umwelt. 2018. "Lärmbelastung in Der Schweiz Lärmbelastung in Der Schweiz." www.bafu.admin.ch/uv-1820-d.

Banerjee, Amitav, UB Chitnis, SL Jadhav, JS Bhawalkar, and S Chaudhury. 2009. "Hypothesis Testing, Type I and Type II Errors." *Industrial Psychiatry Journal* 18 (2): 127. https://doi.org/10.4103/0972-6748.62274.

Bauer, Jan, and David A Groneberg. 2016. "Measuring Spatial Accessibility of Health Care Providers-Introduction of a Variable Distance Decay Function within the Floating Catchment Area (FCA) Method." https://doi.org/10.1371/journal.pone.0159148.

Beyer, Kirsten M M, Andrea Kaltenbach, Aniko Szabo, Sandra Bogar, F Javier Nieto, and Kristen M Malecki. 2014. "Exposure to Neighborhood Green Space and Mental Health: Evidence from the Survey of the Health of Wisconsin." *Int. J. Environ. Res. Public Health* 11: 11. https://doi.org/10.3390/ijerph110303453.

Biau, David Jean, Solen Kernéis, and Raphaël Porcher. 2008. "Statistics in Brief: The Importance of Sample Size in the Planning and Interpretation of Medical Research." *Clinical Orthopaedics and Related Research*. https://doi.org/10.1007/s11999-008-0346-9.

Bornstein, Daniel B., and William J. Davis. 2014. "The Transportation Profession's Role in Improving Public Health." *ITE Journal (Institute of Transportation Engineers)* 84 (7): 18–24. www.ite.org.

Broniatowski, David A., Amelia M. Jamison, Si Hua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra C. Quinn, and Mark Dredze. 2018. "Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate." *American Journal of Public Health* 108 (10): 1378–84. https://doi.org/10.2105/AJPH.2018.304567.

Bundesamt für Statistik (2022). Räumliche Gliederungen der Schweiz, Räumliche Typologien, Stadt/Land-Typologie 2012. URL: https://www.atlas.bfs.admin.ch/maps/13/de/12362_12361_3191_227/20389.html. Accessed on 12.12.2021.

Bundesamt für Umwelt (2021). swissTLMRegio, Das kleinmassstäbliche digitale Landschaftsmodell der Schweiz, Produktinformation. URL: https://www.swisstopo.admin.ch/content/swisstopo-internet/de/geodata/landscape/tlmregio/_jcr_content/contentPar/tabs_copy/items/dokumente/tabPar/downloadlist/downloadItems/184_1634221561733.download/2021_10_TLMRegio_InfoD_bf.pdf. Accessed on 23.03.2021.

Camargo, Rafael. 2016. "The Effects of Urban Green Spaces on House Prices." https://doi.org/10.13140/RG.2.2.32791.19365.

Chen, Hao, Kaisheng Lai, Lingnan He, and Rongjun Yu. 2020. "Where You Are Is Who You Are? The Geographical Account of Psychological Phenomena." *Frontiers in Psychology* 11 (March): 536. https://doi.org/10.3389/FPSYG.2020.00536/BIBTEX.

Cohn, Michael A., Matthias R. Mehl, and James W. Pennebaker. 2004. “Linguistic Markers of Psychological Change Surrounding September 11, 2001.” *Psychological Science* 15 (10): 687–93. https://doi.org/10.1111/j.0956-7976.2004.00741.x.

Conway, Mike, and Daniel O’Connor. 2016. “Social Media, Big Data, and Mental Health: Current Advances and Ethical Implications.” *Current Opinion in Psychology* 9 (June): 77–82. https://doi.org/10.1016/J.COPSYC.2016.01.004.

Cutrona, Carolyn E., Gail Wallace, and Kristin A. Wesner. 2006. “Neighborhood Characteristics and Depression: An Examination of Stress Processes.” *Current Directions in Psychological Science* 15 (4): 188. https://doi.org/10.1111/J.1467-8721.2006.00433.X.

Datareportal (2021a): The full Digital 2021 global report. Twitter: Advertising Audience Overview. URL: https://datareportal.com/reports/digital-2021-global-overview-report. Accessed on 09.04.2022.

Datareportal (2021b): Digital 2021: Switzerland. Twitter: Advertising Audience Overview. URL: https://datareportal.com/reports/digital-2021-switzerland?rq=switzerland. Accessed on 09.04.2022.

Davis, Clayton Allen, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. “BotOrNot.” In , 273–74. https://doi.org/10.1145/2872518.2889302.

Dredze, Mark. 2012. “How Social Media Will Change Public Health.” *IEEE Intelligent Systems* 27 (4): 81–84. https://doi.org/10.1109/MIS.2012.76.

Dummer, Trevor J.B. 2008. “Health Geography: Supporting Public Health Policy and Planning.” *CMAJ* 178 (9): 1177–80. https://doi.org/10.1503/CMAJ.071783.

Edry, Tamar, Nason Maani, Martin Sykora, Suzanne Elayan, Yulin Hswen, Markus Wolf, Fabio Rinaldi, Sandro Galea, and Oliver Gruebner. 2021. “Real-Time Geospatial Surveillance of Localized Emotional Stress Responses to COVID-19: A Proof of Concept Analysis.” *Health & Place* 70 (July): 102598. https://doi.org/10.1016/J.HEALTHPLACE.2021.102598.

Eichstaedt, Johannes C., Hansen Andrew Schwartz, Margaret L. Kern, Gregory Park, Darwin R. Labarthe, Raina M. Merchant, Sneha Jha, et al. 2015. “Psychological Language on Twitter Predicts County-Level Heart Disease Mortality:” 26 (2): 159–69. https://doi.org/10.1177/0956797614557867.

Elliott, Paul, and Daniel Wartenberg. 2004. “Spatial Epidemiology: Current Approaches and Future Challenges.” *Environmental Health Perspectives* 112 (9): 998–1006. https://doi.org/10.1289/EHP.6735.

Evans-Lacko, S., S. Aguilar-Gaxiola, A. Al-Hamzawi, J. Alonso, C. Benjet, R. Bruffaerts, W. T. Chiu, et al. 2018. “Socio-Economic Variations in the Mental Health Treatment Gap for People with Anxiety, Mood, and Substance Use Disorders: Results from the WHO World Mental Health (WMH) Surveys.” *Psychological Medicine* 48 (9): 1560–71. https://doi.org/10.1017/S0033291717003336.

Ferrara, Emilio, Clayton Davis, Filippo Menczer, Alessandro Flammini, and Onur Varol. 2016. “The Rise of Social Bots.” *Commun. ACM* 59: 96–104. https://doi.org/10.1145/2818717.

Frost, Allison, Lindsay T. Hoyt, Alissa Levy Chung, and Emma K. Adam. 2015. “Daily Life with Depressive Symptoms: Gender Differences in Adolescents’ Everyday Emotional Experiences.” *Journal of Adolescence* 43 (August): 132–41. https://doi.org/10.1016/J.ADOLESCENCE.2015.06.001.

Gary-Webb, Tiffany L, Kesha Baptiste-Roberts, Luu Pham, Jacqueline Wesche-Thobaben, Jennifer Patricio, Xavier Pi-Sunyer, Arleen F Brown, Lashanda Jones-Corneille, and Frederick L Brancati. 2011. "Neighborhood Socio-economic Status, Depression, and Health Status in the Look AHEAD (Action for Health in Diabetes) Study." https://doi.org/10.1186/1471-2458-11-349.

Gidlöf-Gunnarsson, Anita, and Evy Öhrström. 2007. "Noise and Well-Being in Urban Residential Environments: The Potential Role of Perceived Availability to Nearby Green Areas." *Landscape and Urban Planning* 83 (2–3): 115–26. https://doi.org/10.1016/J.LANDURBPLAN.2007.03.003.

Gligorić, Kristina, Ashton Anderson, and Robert West. 2020. "Adoption of Twitter's New Length Limit: Is 280 the New 140?" https://archive.org/details/twitterstream.

Golder, Scott A., and Michael W. Macy. 2011. "Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength across Diverse Cultures." *Science* 333 (6051): 1878–81. https://doi.org/10.1126/science.1202775.

Goldsmith, Harold F., Charles E. Holzer, and Ronald W. Manderscheid. 1998. "Neighborhood Characteristics and Mental Illness." *Evaluation and Program Planning* 21 (2): 211–25. https://doi.org/10.1016/S0149-7189(98)00012-3.

Google Earth Engine (2021): USGS Landsat 8 Level 2, Collection 2, Tier 1. URL: https://developers.google.com/earthengine/datasets/catalog/LANDSAT_LC08_C02_T1_L2. Accessed on 20.04.2022

Gruebner, Oliver, Martin Sykora, Sarah R. Lowe, Ketan Shankardass, Sandro Galea, and S. V. Subramanian. 2017. "Big Data Opportunities for Social Behavioral and Mental Health Research." *Social Science & Medicine* 189 (September): 167–69. https://doi.org/10.1016/J.SOCSCIMED.2017.07.018.

Gruebner, Oliver, Martin Sykora, Sarah R Lowe, Ketan Shankardass, Ludovic Trinquart, Tom Jackson, S V Subramanian, and Sandro Galea. 2016. "Mental Health Surveillance after the Terrorist Attacks in Paris." *The Lancet* 387 (10034): 2195–96. https://doi.org/10.1016/S0140-6736(16)30602-X.

Helbich, Marco. 2018. "Toward Dynamic Urban Environmental Exposure Assessments in Mental Health Research." *Environmental Research* 161 (February): 129–35. https://doi.org/10.1016/J.ENVRES.2017.11.006.

Iac acoustics (2022). Comparative Examples of Noise Levels. URL: https://www.iacacoustics.com/blog-full/comparative-examples-of-noise-levels.html. Accessed on 25.04.2022.

Alcock, Jan, White Mathew P, Wheeler Benedict W, Fleming Lora E, and Depledge Michael H. 2014. "Longitudinal Effects on Mental Health of Moving to Greener and Less Green Urban Areas." *Environmental Science & Technology* 48 (2): 1247–55. https://doi.org/10.1021/ES403688W.

Naslund, John A, Gonsalves Pattie P, Gruebner Oliver, Pendse Sachin R, Smith Stephanie L, Sharma Amit, and Raviola Giuseppe. 2019. "Digital Innovations for Global Mental Health: Opportunities for Data Science, Task Sharing, and Early Intervention." *Current Treatment Options in Psychiatry* 6 (4): 337–51. https://doi.org/10.1007/S40501-019-00186-8.

James, Peter, Rachel F Banay, Jaime E Hart, and Francine Laden. 2015. "A Review of the Health Benefits of Greenness." *Current Epidemiology Reports* 2 (2): 131–42. https://doi.org/10.1007/s40471-015-0043-7.

Jashinsky, Jared, Scott H. Burton, Carl L. Hanson, Josh West, Christophe Giraud-Carrier, Michael D. Barnes, and Trenton Argyle. 2014. "Tracking Suicide Risk Factors through Twitter in the US." *Crisis* 35 (1): 51–59. https://doi.org/10.1027/0227-5910/a000234.

Jordan, Sophie E, Sierra E Hovet, Isaac Chun-Hai Fung, Hai Liang, King-Wa Fu, and Zion Tsz Ho Tse. 2018. “Using Twitter for Public Health Surveillance from Monitoring and Prediction to Public Response.” https://doi.org/10.3390/data4010006.

Khatter, Kiran, Sukhdev Singh, Diksha Khurana, and Aditya Koli. 2017. “Natural Language Processing: State of The Art, Current Trends and Challenges Lognormal Distribution from Bayesian View Point View Project Natural Language Processing: State of The Art, Current Trends and Challenges.” https://www.researchgate.net/publication/319164243.

Kilanowski, Jill F. 2017. “Breadth of the Socio-Ecological Model.” *Journal of Agromedicine* 22 (4): 295–97. https://doi.org/10.1080/1059924X.2017.1358971.

Kupcikova, Zuzana, Daniela Fecht, Rema Ramakrishnan, Charlotte Clark, and Yutong Samuel Cai. 2021. “Road Traffic Noise and Cardiovascular Disease Risk Factors in UK Biobank.” *European Heart Journal* 42 (21): 2072–84. https://doi.org/10.1093/EURHEARTJ/EHAB121.

Maclean, Fiona, Derek Jones, Gail Carin-Levy, and Heather M Hunter. 2013. “Understanding Twitter.” *Understanding Twitter Article in British Journal of Occupational Therapy*. https://doi.org/10.4276/030802213X13706169933021.

Mann, Emily M., Kristiann C. Heesch, Jerome N. Rachele, Nicola W. Burton, and Gavin Turrell. 2022. “Individual Socio-economic Position, Neighbourhood Disadvantage and Mental Well-Being: A Cross-Sectional Multilevel Analysis of Mid-Age Adults.” *BMC Public Health* 22 (1). https://doi.org/10.1186/S12889-022-12905-7.

Martí, Pablo, Leticia Serrano-Estrada, Almudena Nolasco-Cirugeda, and Jesús López Baeza. 2021. “Revisiting the Spatial Definition of Neighborhood Boundaries: Functional Clusters versus Administrative Neighborhoods.” *Journal of Urban Technology*. https://doi.org/10.1080/10630732.2021.1930837.

Murtagh, Elaine M., Jacqueline L. Mair, Elroy Aguiar, Catrine Tudor-Locke, and Marie H. Murphy. 2021. “Outdoor Walking Speeds of Apparently Healthy Adults: A Systematic Review and Meta-Analysis.” *Sports Medicine*. Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s40279-020-01351-3.

Nadkarni, Prakash M., Lucila Ohno-Machado, and Wendy W. Chapman. 2011. “Natural Language Processing: An Introduction.” *Journal of the American Medical Informatics Association* 18 (5): 544–51. https://doi.org/10.1136/AMIAJNL-2011-000464.

O’Campo, Patricia, Blair Wheaton, Rosane Nisenbaum, Richard H. Glazier, James R. Dunn, and Catharine Chambers. 2015. “The Neighbourhood Effects on Health and Well-Being (NEHW) Study.” *Health & Place* 31 (January): 65–74. https://doi.org/10.1016/J.HEALTHPLACE.2014.11.001.

Panczak, Radoslaw, Bruna Galobardes, Marieke Voorpostel, Adrian Spoerri, Marcel Zwahlen, and Matthias Egger. 2012. “A Swiss Neighbourhood Index of Socio-economic Position: Development and Association with Mortality.” *Journal of Epidemiology and Community Health* 66 (12): 1129–36. https://doi.org/10.1136/jech-2011-200699.

Schreiber-Gregory, Deanna, and Karlen Bader. 2018. “Logistic and Linear Regression Assumptions: Violation Recognition and Control.” *Midwest SAS User Group*, no. May: 1–21. https://www.researchgate.net/publication/341354759.

Schwartz, H Andrew, Johannes C Eichstaedt, Margaret L Kern, and David Stillwell. 2014. “Automatic Personality Assessment Through Social Media Language.” *Article in Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspp0000020.

Sinnenberg, Lauren, Alison M. Buttenheim, Kevin Padrez, Christina Mancheno, Lyle Ungar, and Raina M. Merchant. 2017. “Twitter as a Tool for Health Research: A Systematic Review.” *American Journal of Public Health* 107 (1): e1–8. https://doi.org/10.2105/AJPH.2016.303512.

Smith, Graham, Christopher Gidlow, Rachel Davey, and Charles Foster. 2010. “What Is My Walking Neighbourhood? A Pilot Study of English Adults’ Definitions of Their Local Walking Neighbourhoods.” *International Journal of Behavioral Nutrition and Physical Activity* 7: 34. https://doi.org/10.1186/1479-5868-7-34.

Stadt Zürich (2021). Präsidialdepartement, Statistik, Themen, Bevölkerung. URL: https://www.stadt-zuerich.ch/prd/de/index/statistik/themen/bevoelkerung.html. Accessed on 17.04.2022.

Statista (2018). Most Used Social Media Platforms of Teenagers in the United States as of April 2018, by Age Group. URL: https://www.statista.com/statistics/945390/teenagers-social-media-platforms-the-most-usa-age/. Accessed on 16.04.2022.

Statista (2022): Leading countries based on number of Twitter users as of January 2022. URL: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/. Accessed on 09.04.2022.

Swisstopo (2018). SwissATLI3D, Das hoch aufgelöste Terrainmodell der Schweiz. URL: https://www.swisstopo.admin.ch/content/swisstopo-internet/de/geodata/height/alti3d/_jcr_content/contentPar/tabs_copy/items/dokumente/tabPar/downloadlist/downloadItems/846_1464690554132.download/swissALTI3D_detaillierte%20Produktinfo_201802_DE.pdf. Accessed on 23.03.2021.

Subrahmanian, VS, Amos Azaria, Skylar Durst, Vadim Kagan, Aram Galstyan, Kristina Lerman, Linhong Zhu, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. “Shuyang Gao (USC), Tad Hogg (Institute for Molecular Manufacturing) Farshad Kooti (USC).” Pacific Social.

Taylor, Lucy, and Dieter F. Hochuli. 2017. “Defining Greenspace: Multiple Uses across Multiple Disciplines.” *Landscape and Urban Planning* 158 (February): 25–38. https://doi.org/10.1016/J.LANDURBPLAN.2016.09.024.

Tenailleau, Quentin M., Frédéric Mauny, Daniel Joly, Stéphane François, and Nadine Bernard. 2015. “Air Pollution in Moderately Polluted Urban Areas: How Does the Definition of ‘Neighborhood’ Impact Exposure Assessment?” *Environmental Pollution* 206 (August): 437–48. https://doi.org/10.1016/J.ENVPOL.2015.07.021.

Tenailleau, Quentin, Sophie Pujol, Hélène Houot, and Daniel Joly. 2014. “Assessing Residential Exposure to Urban Noise Using Environmental Models: Does the Size of the Local Living Neighborhood Matter? MIXTURE : Multi-Exposure in Urban Environment : A Spatial Multi-Scale Approach of Human Exposure to Noise and Atmospheric Pollution View Project Malaria in Pregnancy View Project.” *Article in Journal of Exposure Science & Environmental Epidemiology*. https://doi.org/10.1038/jes.2014.33.

Tunstall, H V Z, Shaw M, and Dorling D. 2004. “Places and Health.” *J Epidemiol Community Health* 58: 6–10. https://doi.org/10.1136/jech.58.1.6.

Twitter Developer Platform (2022). Twitter API. URL: https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api. Accessed on 02.04.2022.

Universitätsspital Zürich (2022). Krankheiten & Therapien, Heuschnupfen. URL: https://www.usz.ch/krankheit/heuschnupfen/. Accessed on 08.04.2022.

Vries, Jacob J. de, Peter Nijkamp, and Piet Rietveld. 2009. “Exponential or Power Distance-Decay for Commuting? An Alternative Specification.” *Environment and Planning A* 41 (2): 461–80. https://doi.org/10.1068/a39369.

Wang, Zijian, Scott A. Hale, David Adelani, Przemyslaw A. Grabowicz, Timo Hartmann, Fabian Flöck, and David Jurgens. 2019. “Demographic Inference and Representative Population Estimates from Multilingual Social Media Data.” *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*, 2056–67. https://doi.org/10.1145/3308558.3313684.

Wheeler, Benedict W, Rebecca Lovell, Sahran L Higgins, Mathew P White, Ian Alcock, Nicholas J Osborne, Kerryn Husk, Clive E Sabel, and Michael H Depledge. 2015. “Beyond Greenspace: An Ecological Study of Population General Health and Indicators of Natural Environment Type and Quality.” *International Journal of Health Geographics* 14 (1). https://doi.org/10.1186/s12942-015-0009-5.

Yang, Kai-Cheng, Onur Varol, Pik-Mai Hui, and Filippo Menczer. 2020. “Scalable and Generalizable Social Bot Detection through Data Selection.” *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (01): 1096–1103. https://doi.org/10.1609/AAAI.V34I01.5460.

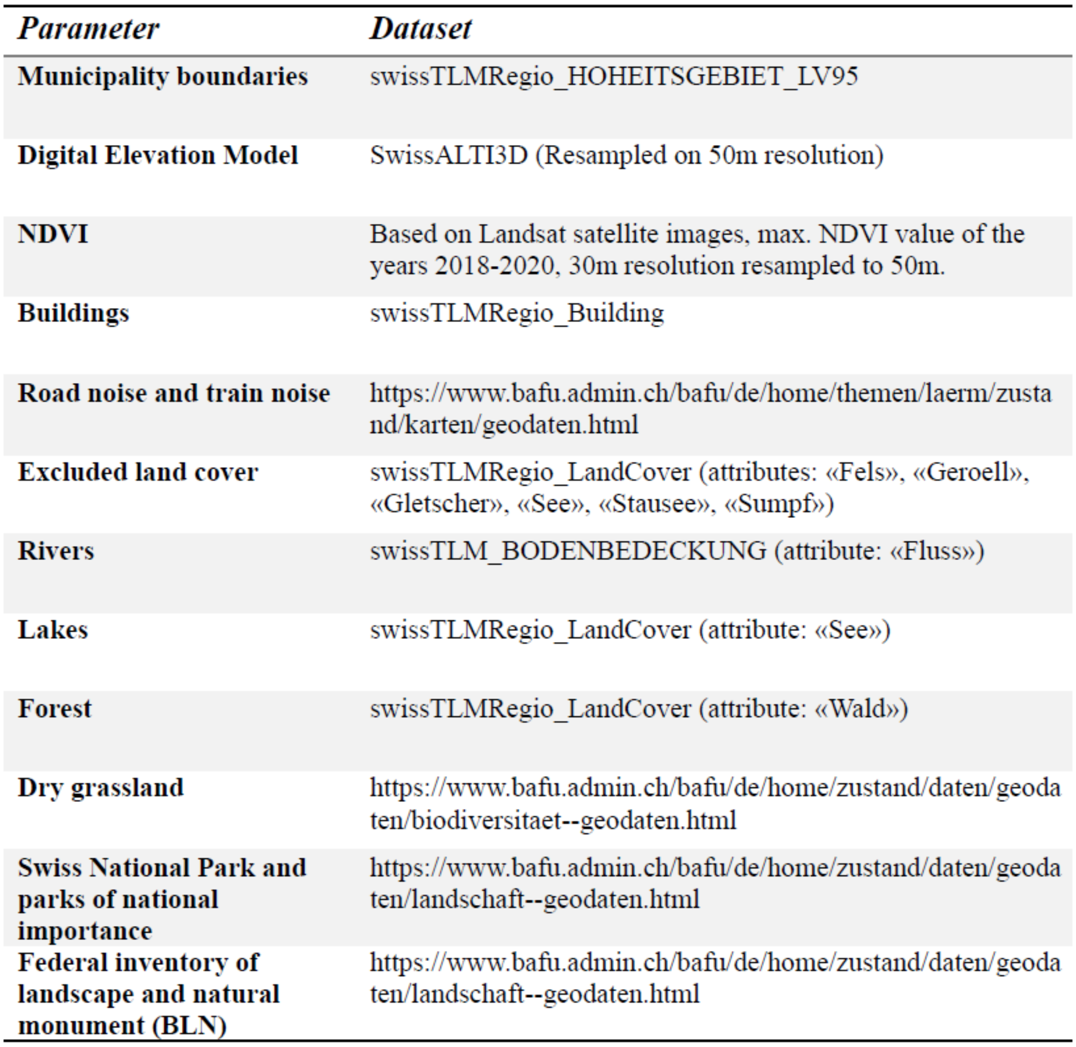
Yin, Zhang Cai, Zhang Hao Nan Jin, Shen Ying, Hui Liu, San Juan Li, and Jia Qiang Xiao. 2019. “Distance-Decay Effect in Probabilistic Time Geography for Random Encounter.” *ISPRS International Journal of Geo-Information* 8 (4). https://doi.org/10.3390/ijgi8040177.

Appendix

A.1 Greenspace Dataset

As mentioned in Section 4.2, the greenspace dataset was made by the author and 5 fellow students (L. Asper, S. Caduff, M. Niederberger, L. Schädler, and N. Steinmann) within a Geography Master's course at the University of Zürich. In this section, the used data and the implementation for the greenspace dataset used in this thesis are described. All processing was done in ArcGIS Pro.

A.1.1 Data



Parameter	*Dataset*
Municipality boundaries	swissTLMRegio_HOHEITSGEBIET_LV95
Digital Elevation Model	SwissALTI3D (Resampled on 50m resolution)
NDVI	Based on Landsat satellite images, max. NDVI value of the years 2018-2020, 30m resolution resampled to 50m.
Buildings	swissTLMRegio_Building
Road noise and train noise	https://www.bafu.admin.ch/bafu/de/home/themen/laerm/zustand/karten/geodaten.html
Excluded land cover	swissTLMRegio_LandCover (attributes: «Fels», «Geroell», «Gletscher», «See», «Stausee», «Sumpf»)
Rivers	swissTLM_BODENBEDECKUNG (attribute: «Fluss»)
Lakes	swissTLMRegio_LandCover (attribute: «See»)
Forest	swissTLMRegio_LandCover (attribute: «Wald»)
Dry grassland	https://www.bafu.admin.ch/bafu/de/home/zustand/daten/geodaten/biodiversitaet--geodaten.html
Swiss National Park and parks of national importance	https://www.bafu.admin.ch/bafu/de/home/zustand/daten/geodaten/landschaft--geodaten.html
Federal inventory of landscape and natural monument (BLN)	https://www.bafu.admin.ch/bafu/de/home/zustand/daten/geodaten/landschaft--geodaten.html

Figure 35: Data used to generate the greenspace index dataset

A.1.2 Implementation

Excluded areas

Excluded areas comprise rocks, glaciers, debris, lakes, water reservoirs as well as bogs. These were extracted from land cover data of swissTLMRegio *(Select by attributes)*. The vector data of land cover and buildings were combined *(Union)* and then converted to a raster *(Polygon to Raster)*. By using *Raster Calculator*, the raster values were rescaled to values of 0 (excluded areas) or 1 (potential greenspace).

Greenness

To illustrate greenness, an NDVI image for Switzerland from the Landsat satellite was produced with Google Earth Engine. To ensure that all pixels have a usable pixel value and that the satellite image is not disturbed by clouds and other effects, a composite of several satellite images was created. Satellite images captured between April and October in the years 2018, 2019, and 2020 were included. Since we are interested in the greenness of a pixel, winter months are not needed because most vegetation areas lose their leaves. All pixels identified as clouds by the Landsat cloud detection algorithm have been excluded. The NDVI was calculated for all available satellite images based on the formula below.

$$NDVI = \frac{NIR - RED}{NIR + RED}$$

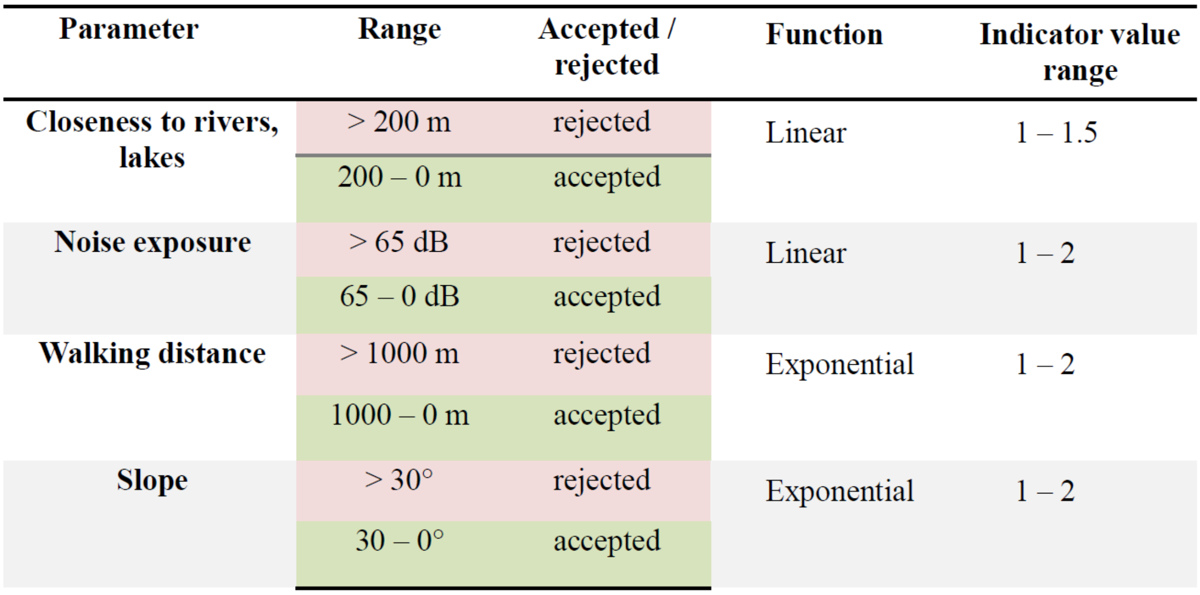
A single image was created from the NDVI composites, by assigning each pixel the highest NDVI value available from all pixels over the time period of 3 years. By considering multiple images and including the highest value, disturbances such as cloud cover or cloud shadow which were not detected with the default Landsat algorithm could be fixed. Landsat satellite captures images with a spatial resolution of 30 m, which was rescaled to 50 m so that less storage space is needed and the spatial resolution corresponds to the end product. NDVI values lower than 0.40 were excluded and the final raster was then rescaled to values from 4 to 10.

Landscape quality

The landscape quality is described by various factors. At first, the closeness of greenspace to rivers and lakes was evaluated by merging both vector data (Union), by rasterizing the data (Polygon to raster), and by calculating the distance to rivers and lakes (Euclidean Distance, max. distance of 200 m). The resulting absolute distance (m) was then rescaled linearly to values between 1 (distance > 200 m) and 1.5 (at a river/lake) (Tab. 2). Noise data was first summarised by extracting the maximum value in dB from both noise data sets (roads and railways, Cell Statistics). Values higher than 65 dB were completely excluded since the recreational value of greenspace at high noise levels is not given. The values were then rescaled linearly from 1 (highest acceptable noise exposure) to 2 (no noise) using Raster Calculator. Lastly, landscape properties that indicate high quality were given additional weight, namely: forests due to their recreational value, dry grassland with its ecological importance as well as areas belonging to the Federal inventory of landscape and natural monument (BLN) and the Swiss National Park and parks of national importance. In the following, these data are referred to as ‘value-added objects’. All data of the value-added objects were merged (Union) and converted into a raster (Polygon to raster). If an area did belong to one of these ‘value added objects’ a value of 1.1 was given (bonus of 0.1). If an area did not belong to any of these objects, a value of 1 was assigned (no effect in final raster multiplication).

Accessibility

To calculate the walking distance, the Euclidean distance to buildings was computed (max. distance of 1’000 m). These distances were then rescaled to values ranging from 1 (1’000 m away from buildings) to 2 (close to buildings) using a function which exponentially decreases with increasing distance (Raster Calculator). As steep areas are not accessible to all people, the slope of the surface was included into the accessibility and calculated based on the DEM (Slope). Slopes larger than 30° were entirely excluded. For slopes between 0 and 30°, values between 1 (steep) and 2 (flat) were assigned using an exponential function (Raster calculator).



Parameter	Range	Accepted / rejected	Function	Indicator value range
Closeness to rivers, lakes	> 200 m	rejected	Linear	1 – 1.5
	200 – 0 m	accepted		
Noise exposure	> 65 dB	rejected	Linear	1 – 2
	65 – 0 dB	accepted		
Walking distance	> 1000 m	rejected	Exponential	1 – 2
	1000 – 0 m	accepted		
Slope	> 30°	rejected	Exponential	1 – 2
	30 – 0°	accepted		

Figure 34: Range of values used for rescaling final raster data for the parameters closeness to rivers and lakes, noise exposure, walking distance and slope.

Indicator implementation

After the data was generated following the pre-processing steps, the actual indicator implementation was done. The layers were multiplied without any further weighting (Raster Calculator). The indicator was then rescaled to values between 0 and 10 and exported to a 50 m resolution raster resulting in a general greenspace indicator.

A.2 Cluster Map of Users$_{\text{regression}}$

On the next page, a cluster map visualizes the spatial distributions of users$_{\text{regression}}$ in such a manner, that the number of points in a dense spot are directly readable from the map (see *Figure A.1*).

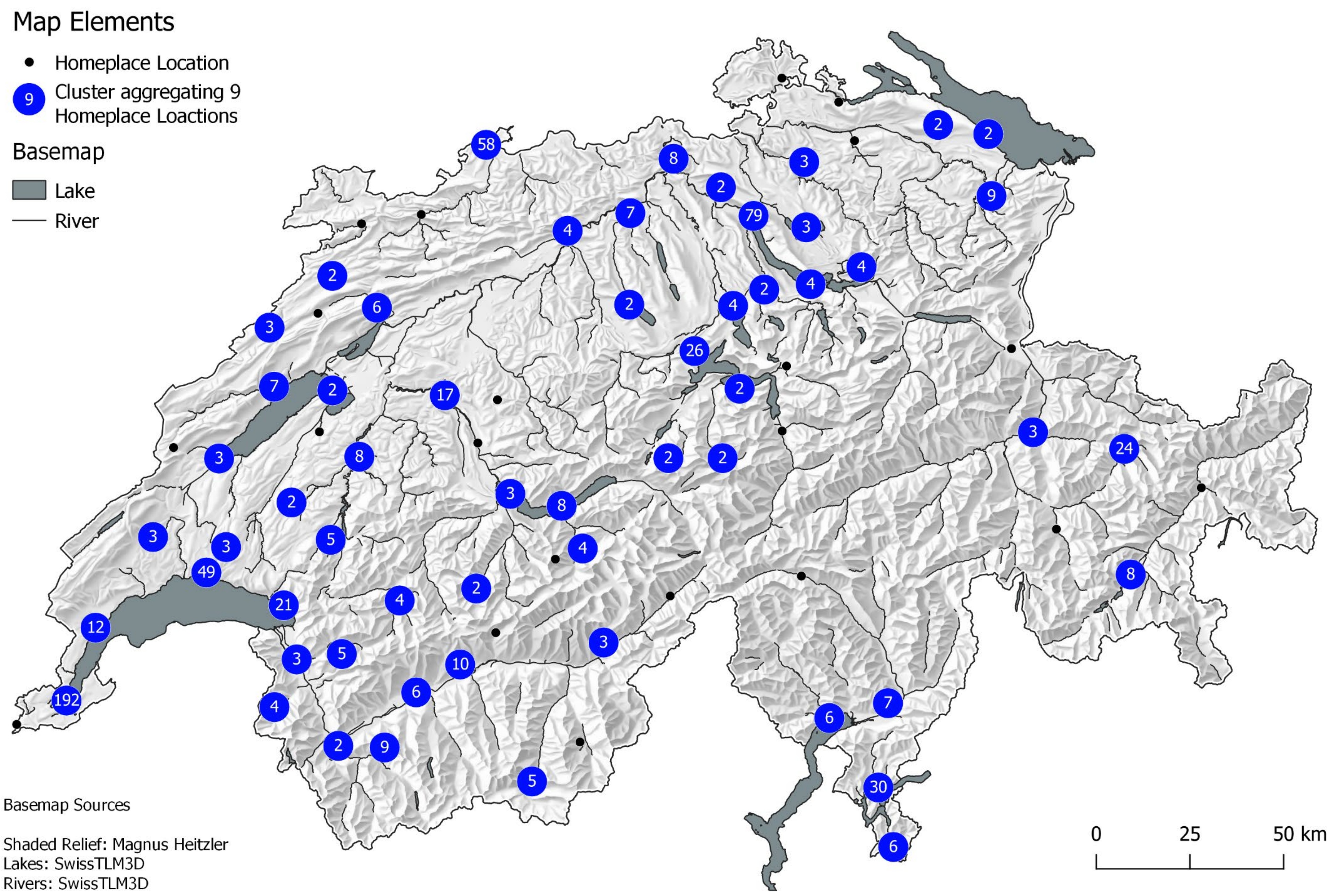


Figure 37: Visualisation of spatial distribution of homeplace locations as clusters.

A.3 Python Code for Parallel Processing

```
# -*- coding: utf-8 -*-
# -------------------------------------------------------------------------
--
# Created on: 2021-12-15
# -------------------------------------------------------------------------
--

# Set the necessary product code
# import arcinfo

# Import arcpy module
import arcpy, os, time, traceback, csv, sys
from simpledbf import Dbf5
import numpy as np

# Check out any necessary licenses.
if arcpy.CheckExtension("3D") == "Available":
    arcpy.CheckOutExtension("3D")

if arcpy.CheckExtension("spatial") == "Available":
    arcpy.CheckOutExtension("spatial")

if arcpy.CheckExtension("ImageAnalyst") == "Available":
    arcpy.CheckOutExtension("ImageAnalyst")

if arcpy.CheckExtension("ImageExt") == "Available":
    arcpy.CheckOutExtension("ImageExt")


def delete_shapefile(shp_path):
    path = shp_path[:-4]
    endings = [".cpg", ".dbf", ".prj", ".sbn", ".sbx", ".shp", ".shp.xml",
".shx"]

    for ending in endings:
        file_path = path + ending

        if os.path.isfile(file_path):
            os.remove(file_path)

def delete_raster(ras_path):
    ras_files = os.listdir(ras_path)
    endings = [".adf", "log", ".xml"]
    for ras_file in ras_files:
        for ending in endings:
            if ras_file.endswith(ending):
                ras_file_path = ras_path + "/" + ras_file
                os.remove(ras_file_path)

    os.rmdir(ras_path)


arcpy.env.overwriteOutput = True
```

```
sha_directory = "."

# ADJUST Greenspace / Traffic_Noise / SSEP
neighborhood_variable = "Greenspace"

# ADJUST 500m or 1000m
neighborhood_radius = "500m"

# ADJUST to 540.0 for 500m neighborhood_radius and 1040.0 for 1000m
neighborhood_radius
raster_calc_radius = 540.0

# ADJUST 1-8
script_number = "1"

batch_input = neighborhood_radius + "_" + script_number
batch_process = "Processing_" + script_number

base_tile_directory = "S:\\group\\m-health\\03 projects\\MSc
Schmidheiny\\ArcGIS_Projects\\ArcGIS_Project_01\\Rasters\\" +
neighborhood_variable + "\\" + neighborhood_radius + "\\" + batch_input +
"\\"
base_temp_directory = "S:\\group\\m-health\\03 projects\\MSc
Schmidheiny\\ArcGIS_Projects\\ArcGIS_Project_01\\Rasters\\" +
neighborhood_variable + "\\" + neighborhood_radius + "\\" + batch_process +
"\\"
base_tables_directory = "S:\\group\\m-health\\03 projects\\MSc
Schmidheiny\\ArcGIS_Projects\\ArcGIS_Project_01\\Tables\\" +
neighborhood_variable + "\\" + neighborhood_radius + "\\"

homeplace_users_buffer = "S:\\group\\m-health\\03 projects\\MSc
Schmidheiny\\ArcGIS_Projects\\ArcGIS_Project_01\\Rasters\\" +
neighborhood_variable + "\\" + neighborhood_radius + "\\Input_" +
script_number + "\\homeplace_users_buffer_" + neighborhood_radius + ".shp"


files = os.listdir(base_tile_directory)
tif_files = [file for file in files if file.endswith(".TIF")]


# Create empty matrix to store neighborhood variable mean
neighborhood_variable_mean = np.zeros((1213,2))


print(len(tif_files))

for tif_file in tif_files:
    raw_tile_name = tif_file[:-4]
    input_tile = base_tile_directory + "\\" + tif_file

    # ADJUST DEPENDING ON NEIGHBORHOODVARIABLE AND RADIUS
    tile_number = raw_tile_name[11:]

    # Saving Processing Results in separate folders
    temp_folder = base_temp_directory + "\\" + raw_tile_name + "\\"
    if not os.path.exists(temp_folder):
        os.makedirs(temp_folder)

    print("Processing tile: " + raw_tile_name)
    start_time = time.time()
    while True:
        try:

            # Process: Select AOI
            neighborhood_AOI_shp = temp_folder + "neighborhood_AOI.shp"
            arcpy.Select_analysis(homeplace_users_buffer,
neighborhood_AOI_shp, '"FID" =' + tile_number)


            # Process: Feature To Point (Feature To Point) (management)
            AOI_centroid_shp = temp_folder + "AOI_centroid.shp"
            with arcpy.EnvManager(scratchWorkspace=r"S:\group\m-health\03
projects\MSc
Schmidheiny\ArcGIS_Projects\ArcGIS_Project_01\ArcGIS_Project_01.gdb",
workspace=r"S:\group\m-health\03 projects\MSc
Schmidheiny\ArcGIS_Projects\ArcGIS_Project_01\ArcGIS_Project_01.gdb"):
                 arcpy.management.FeatureToPoint(in_features =
neighborhood_AOI_shp, out_feature_class = AOI_centroid_shp,
point_location="INSIDE")


            # Process: Euclidean Distance (Euclidean Distance) (sa)
            euclidean_distance_tif = temp_folder + "euclidean_distance.tif"
            Euclidean_Distance = euclidean_distance_tif
            with arcpy.EnvManager(extent = input_tile,
                                  snapRaster = input_tile):
                outEucDistance =
arcpy.sa.EucDistance(in_source_data=AOI_centroid_shp, maximum_distance  =
raster_calc_radius, cell_size = "50",
out_direction_raster=euclidean_distance_tif, distance_method="PLANAR")
                outEucDistance.save(Euclidean_Distance)


            # Process: Extract by Mask
            euclidean_distance_masked_tif = temp_folder +
"euclidean_distance_masked.tif"
            outExtractByMask =
arcpy.sa.ExtractByMask(euclidean_distance_tif, input_tile)
            outExtractByMask.save(euclidean_distance_masked_tif)


            # Process: Raster Calculator (Raster Calculator) (ia)
            euclidean_distance_norm_tif = temp_folder +
"euclidean_distance_norm.tif"
            Raster_Calculator = euclidean_distance_norm_tif
            EucDistRaster = arcpy.sa.Raster(euclidean_distance_masked_tif)
            outRasterCalc = 1.0 - ((EucDistRaster + 1.0) /
raster_calc_radius)
            outRasterCalc.save(Raster_Calculator)


            # Process: Raster Calculator (2) (Raster Calculator) (ia)
            green_500m_0_weight_tif = temp_folder + "green_500m_weight.tif"
            Raster_Calculator_2 = green_500m_0_weight_tif
            EucDistNormRaster =
arcpy.sa.Raster(euclidean_distance_norm_tif)
            InputRaster = arcpy.sa.Raster(input_tile)
            outRasterCalc = EucDistNormRaster * InputRaster
            outRasterCalc.save(Raster_Calculator_2)
            # Get input Raster properties
            inRas = arcpy.Raster(green_500m_0_weight_tif)

            # Convert Raster to numpy array
            myarr = arcpy.RasterToNumPyArray(inRas, nodata_to_value = -999)

            def nan_if(arr, value):
                return np.where(arr == value, np.nan, arr)

            arrMean = np.nanmean([nan_if(myarr, -999)])


            # Writing greenspace mean into matrix
            neighborhood_variable_mean[tile_number,0] = tile_number
            neighborhood_variable_mean[tile_number,1] = arrMean


            np.savetxt(base_tables_directory + neighborhood_variable + "_"
+ neighborhood_radius + "_" + script_number + "_mean_temp.csv",
neighborhood_variable_mean, delimiter = ",")


            break

        except:
            print("Warning: Something went wrong for tile: " +
raw_tile_name)
            arcpy.gp.AddError(traceback.format_exc())

    end_time = time.time()
    print("elapsed time: ", str(end_time - start_time))

np.savetxt(base_tables_directory + neighborhood_variable + "_" +
neighborhood_radius + "_" + script_number + "_mean_final.csv",
neighborhood_variable_mean, delimiter = ",")

print("FINISHED")
```

Personal Declaration

I hereby declare that the submitted thesis is result of my own, independent work.

All external sources are explicitly acknowledged in the thesis.

Nicolas Schmidheiny, 30.04.2022