MSc Thesis (GEO 511)

Studying Human Mobility Through Geotagged Social Media Content

Timo Grossenbacher

07-707-821

timo.grossenbacher@uzh.ch

Handed in on the 31st of January, 2014

Supervisors: Prof. Dr. Ross S. Purves (UZH), Florian Straub (ETHZ)

Faculty Member: Prof. Dr. Ross S. Purves (UZH)

Geocomputation Unit (GIScience)
Department of Geography
University of Zürich

Chair of Geoinformation Engineering
Institute of Cartography and
Geoinformation
ETH Zürich

Acknowledgements

When I wrote this thesis, everybody was talking about "Big Data". I always wondered — what is it, and what does "big" actually mean? Is it big in terms of numbers, or is it big in terms of complexity? And what are people actually doing with it?

I still do not know the answers to these questions, but now, I can at least say that I am able to run my own, little "Big Data" analysis. Without the help of the following people, this would not have been easy, or mabye not even possible. I therefore want to genuinely thank:

- Prof. Dr. Ross Purves, my principal supervisor, for the many meetings, the critical comments, the helpful hints, and for bringing me back down to earth.
- Florian Straub, my co-supervisor, and Grant McKenzie, for encouraging me to do what I really wanted to do and for the important advice about the mechanics of scientific work.
- Kathrin Freire and Christoph Freymond of the Swiss Federal Statistical Office, for providing me with the invaluable authoritative data needed in this thesis and for patiently answering all my questions.
- Dr. Ralph Straumann, for proofreading my thesis, giving me true expert advice, and, most importantly, for keeping me from writing too much unscientific nonsense.
- My friends and family, who at least *tried* to understand what I do.
- Last but not least, Nicole, for her continuous support and for luring me away from the computer when I needed it.

"For every two degrees the temperature goes up,
check-ins at ice cream shops go up by 2%."

Andrew Hogue, Head of Data, Foursquare

Abstract

In recent years, the Web has seen a steadily growing pile of voluntarily geotagged content on social media, such as georeferenced Tweets or photos. Such content is often made publicly available and has thus sparked interest in the marketing and research community alike. From a GIScience perspective, geotagged content can be looked at as a form of Volunteered Geographic Information (VGI) which may be used to answer all sorts of (geo-)analytical questions. The research community has thus been very enthusiastic about the unprecedented insights into the spatiotemporal behavior of people, as well as the opportunities offered by such data to study human mobility, e.g., the dynamic flows of people between neighborhoods of a city over the course of a day.

Unfortunately, as diverse and promising the new possibilities are, so are the dangers of misinterpreting these data. For instance, socio-demographic representativeness can not be assumed; contrarily, social media are predominantly used by a certain cohort in the Western world. Secondly, not only is geotagged content produced by a still very small percentage of all social media users, it also suffers from the same "participation insequality" phenomenon detected in other types of social media content; meaning that the most prolific users produce an overproportionally large share of content. As this thesis shows, most research efforts in the field of human mobility ignore these issues altogether. Particularly, geotagged content is often used "as is", i.e., conclusions are drawn based on mere spatial aggregations of individually unreferenced, decoupled content. That is, the individual properties of the users behind that content are seldom considered explicitly. Additionally, even though the techniques employed to derive information from this kind of VGI are often innovative, the results of such analyses are rarely compared and validated with authoritative data.

In this work, millions of geotagged Twitter messages are collected and analyzed in order to derive insights about the spatiotemporal behavior of Twitter users. Through a sophisticated exploratory analysis, it is first verified that the collected users actually live within a specified study area and that their data satisfy certain criteria. After that, a rule-based heuristic is used to extract so-called semantic places, such as “home” and “work”, of each user. Based on these and with the help of high-quality authoritative, (geo-)demographic data, it is assessed whether the Twitter data are spatially and socio-demographically representative. As a use case for the study of human mobility, commuter balances are extracted from the semantic places of individual users, and compared with authoritative data.

The results show that a very large majority of the collected data does not satisfy the requirements of being used in a representative survey based on individual users. Moreover, it is found that most geotagged Twitter messages are indeed produced by a rather small minority, which is likely to bias any inferences made from the messages themselves. The results further show that there are stark inequalities not only in terms of demographic factors such as age, but also in terms of the spatial representation of the population. In other words, some linguistic and socio-economic regions seem to be significantly stronger represented on Twitter than others. Together, these findings hint at a socially, culturally, and economically unequally represented population, which raises concerns about the validity of inferences made with such data.

As this thesis shows, though, geotagged Twitter messages can still be looked at as a potential data source for studying human mobility. In fact, the evaluation with official data shows that the inferred commuter balances are a reasonably good indicator for actual flows of people; however, it appears that such inferences are only valid in areas where enough geotagged content is available. This leads to the following conclusions: Firstly, while inferences made from individual spatiotemporal behavior are likely to be more accurate than such made from the decoupled content itself, the data needed for the former are very sparse. Secondly, even though the data are clearly not representative of the general population, they still appear to contain valid indicators of human mobility.

Geotagged content from sources such as Twitter might thus serve as a potential data source for studying mobility, but the data need careful preprocessing and validation in order to be suitable for such an endeavor. The work at hand is among the first to demonstrate this by using an innovative, user-centric approach to spatial knowledge discovery from social media data.

Keywords: VGI, Geotagging, Twitter, Human Mobility, Spatial Data Analysis, Geographical Knowledge Discovery in Databases.

Contents

List of Figures

List of Tables

List of Listings

Acronyms

API Application Programming Interface 1, 5, 10, 12, 21, 37, 39, 40, 47, 131

DBMS Database Management System 43, 135

FSO Swiss Federal Statistical Office 32–36, 49, 72, 74, 86, 92, 94, 144

GIS Geographic Information System 3, 135

GIScience Geographic Information Science 2–4, 16

GPS Global Positioning System 2, 12, 14, 15, 17, 29

GWR Geographically Weighted Regression 23

HTTP Hypertext Transfer Protocol 39

IP Internet Protocol 12

JSON Javascript Object Notation 39, 40, 43, 47, 49, 131

KDE Kernel Density Estimation 55

LBS Location-based Service 9, 10, 12, 15

LBSN Location-based Social Network 1, 3, 9, 10, 14, 15, 18, 21, 24

MAUP Modifiable Areal Unit Problem 23, 99, 105

MBR Minimum Bounding Rectangle 40, 45, 53

MTUP Modifiable Temporal Unit Problem 20, 99

OGC Open Geospatial Consortium 135

PCA Principal Component Analysis 23

REST Representational State Transfer 39, 131

SQL Structured Query Language 43

UGC User-Generated Content 1, 23, 29, 30

URL Uniform Resource Locator 40

UTC Coordinated Universal Time 49, 131

VGI Volunteered Geographic Information 2–7, 9, 18, 21–24, 27–29, 64, 71–73, 91, 107, 108, 111–115

WGS 84 World Geodetic System 84 . . . . 40, 49

1. Introduction

Since Tim O'Reilly (2005) coined the term *Web 2.0* eight years ago, the Internet has seen a tremendous rise of social platforms and networks (generally referred to as *social media*). With Facebook[1] recently crossing the one-billion-user-line (Schroeder, 2012) and Twitter[2] growing faster than ever (Fiegerman, 2012), a constant stream of User-Generated Content (UGC) is produced. So-called Application Programming Interfaces (APIs) allow third parties to programmatically access and download partial or full datasets. This has sparked interest in many researchers to analyze the data; for instance, to discover social structures (Ugander et al., 2011; Kumar et al., 2010; Mislove et al., 2007), reveal privacy leaks (Mislove et al., 2010; Lindamood et al., 2009), and infer moods and sentiments from textual data (Bollen et al., 2011; Diakopoulos & Shamma, 2010).

1.1. Context and Review

A special group of social networks, Location-based Social Networks (LBSNs), such as Foursquare[3] or the now offline Gowalla[4], have introduced the possibility to share a user's geographical position with his or her social network, usually through *checking into* a particular venue or by *geotagging* a status message or some media. The growing popularity of some of these LBSNs has forced well-established social networks such as Facebook to add similar features, which allow people to annotate their updates with explicit geographical references. Therefore, researchers can nowadays access and analyze a steadily growing pile of geographical data from these networks.

[1] http://www.facebook.com

[2] http://www.twitter.com

[3] http://www.foursquare.com

[4] https://en.wikipedia.org/wiki/Gowalla

1.1.1. Location Data and Movement Analysis

In literature, there exists a wide variety of studies with location data, but most of them use data collected in experimental setups, for instance, through Global Positioning System (GPS) loggings (Hofmann-Wellenhof et al., 1993). Such *conventional* sources of data often yield sufficiently accurate, regular, and frequent measurements of position. They are therefore suited for the study of *movement* and its physical properties such as speed and direction, an area where the last decade has seen a wealth of methodological research (Laube et al., 2005; Andersson et al., 2008; Dodge et al., 2008; Z. Li et al., 2010). Using these as well as other approaches, e.g., techniques from machine learning, researchers have tried to make inferences about the spatiotemporal behavior of humans. For instance, Phithakkitnukoon et al. (2010) infer individual activity patterns from mobile telephony cell information. Monreale et al. (2009) cluster many individual trajectories and predict the next location of an individual. Another application is the learning and inference of transportation modes based on raw data from a wearable GPS logger (Liao et al., 2007b).

If a user repeatedly and reasonably frequently updates his or her location on a social media site, one can also derive a *trajectory* trough space and time. However, in contrast to conventional data sources, the regularity and frequency of such updates may greatly vary from one user to another, trajectories are difficult to model, and in many cases one cannot derive movement parameters such as speed and direction. The latter two points are mainly due to the fact that location updates from social media are very sparse and temporally fragmented (Ferrari et al., 2011; N. Andrienko et al., 2012). Thus, only very few studies exist which explicitly make use of individual trajectories derived from location updates, for example the prediction of someone's location based on the last location of his or her Twitter friends (Sadilek et al., 2012).

Therefore, while the widespread dissemination of location sensing technology has made possible to thoroughly study movement, the same kind of research with location updates from social media is still in its infancy. This is not only because such data are a rather new phenomenon but also because they exhibit the above mentioned deficiencies, which will be discussed in more detail in Section 1.1.3.

1.1.2. Studying Human Mobility Through VGI

From a Geographic Information Science (GIScience) perspective, location updates can be looked at as some form of Volunteered Geographic Information

(VGI) and placed in the *egocentric*, *geosocial* domain, as outlined by Elwood et al. (2012). In fact, some studies in GIScience have already made use of such locational information (both textual and explicitly referenced), for example, to quickly detect and locate disasters (De Longueville et al., 2009; Sakaki et al., 2010; Schade et al., 2013).

A promising potential for egocentric, geosocial VGI is the study of human mobility. Given the justified assumption that individual spatiotemporal activity follows a certain routine (Song et al., 2010), trajectories built from a reasonable number of location updates could be aggregated and used to model intra- and interregional mobility, i.e., the spatial and temporal patterns of people moving within and between regions. Location updates from social media are a "natural" form of location disclosure in the sense that users voluntarily share their positions and do not participate in some sort of experiment — a sort of "activity diary" as envisioned in *time geography* (Hägerstraand, 1970). Although recognizing the sparse and noisy nature of such data, Ferrari et al. (2011, p. 9) for instance, consider them as "live traces describing, (...), the way in which people live and interact ...". These data could therefore complement or even replace authoritative data on human mobility and fill gaps in spatial data infrastructure, as proposed by Sui and Goodchild (2011, p. 1742). These scholars see the fusion of Geographic Information Systems (GISs) with social media as an

> "... unprecedented opportunity to have a better understanding of the spatial dynamics of human behavior and societal transformation, (...)."

Research with VGI in the area of human mobility has so far almost exclusively focused on coarse measurements of travel and simple, distance-based models of mobility (Noulas et al., 2012; Z. Cheng et al., 2011), as well as socio-spatial relationships (Leetaru et al., 2013; Allamanis et al., 2012; Takhteyev et al., 2012; Scellato et al., 2011; Cho et al., 2011). The former try to find new, generalizable rules about human mobility or aim to validate existing mobility theories. For example, Noulas et al. (2012) analyze a very large Foursquare dataset and find that the average distance traveled by people is a function of *place* density[5] and varies only slightly across different cities. The latter kind of studies look for relationships between someone's social network and the geographical distances in this network. Cho et al. (2011) use data from two LBSNs to study the influence of social connections on mobility, and come to the conclusion that short-range travel is less impacted by social network structure than long-range travel. Even

[5] A "place" is a Foursquare venue one can check into.

though these studies are often able to fit plausible models, their approach of finding universalities in data from social media has several shortcomings which are illustrated in the next section.

1.1.3. Challenges of VGI

The assessment of VGI as a data source for spatial analysis tasks is an ongoing research challenge in GIScience (Purves, 2011; Sui & Goodchild, 2011; Flanagin & Metzger, 2008). Especially for applications in the social sciences, accuracy, validity, and representativeness of the data need to be considered and assessed when statements about society as a whole are to be made.

The reasons why VGI from social media can lead to false conclusions are manifold. First of all, socio-demographic representativeness of the data is hardly ever given. Contrarily, social media are predominantly used by a certain cohort of the overall population in the Western world, which is, again, only a fraction of the global population (L. Li et al., 2013; Graham, 2012). Secondly, data from social media are often biased towards *outliers*, a small fraction of prolific users who contribute an overproportional amount of data (Haklay, 2012). This holds true especially for geotagged data, because only a small percentage of users actually employs these features, as will be seen later. Lastly, geographic information on social media is not only produced by humans but a variety of *non-humanoid actors* such as automatic broadcasting bots and spammers. In a very recent paper, Crampton et al. (2013, p. 132) address these problems and argue that

> "... there is little that can be said definitively about society-at-large using only these kinds of user-generated data, as such data generally skews toward a more wealthy, more educated, more Western, more white and more male demographic."

The above mentioned studies, while often claiming to infer universal rules, almost never take into account these issues. Additionally, their results cannot be used to quantify mobility patterns between regions inhabited by a few thousand people, but are rather coarse generalizations over very large areas.

1.2. Motivation and Goal

> "... The quality and credibility of [spatiotemporal data from social media] for scientific research and decision-making still need further investigation. We need to explore new ways in which the fusion of

> GIS with social media can be deployed to promote the human-as-sensor paradigm ... in spatial-data generation."

Given this call for a better understanding of spatiotemporal data from social media by Sui and Goodchild (2011, p. 1742), the current state of research, and the intricacies of egocentric, geosocial VGI outlined above, the motivation for this work is twofold:

1. To the author's knowledge, no research exists which assesses the suitability of egocentric, geosocial VGI to complement or replace data from authoritative sources in the domain of human mobility. In particular, no case study exists which actually *validates* findings gained from egocentric, geosocial VGI by comparing them with data from authoritative sources.

2. Approaches which use egocentric, geosocial VGI as data source for the study of human mobility often neglect the accompanying issues of i) socio-demographic bias, as well as of ii) outliers and non-humanoid actors.

Therefore, this thesis has the goal *to assess the representativeness and validity of egocentric, geosocial VGI for quantifying small-scale, regional human mobility in a predefined study area*. This is done by applying a *user-centric approach*, i.e., by collecting location updates and inferring travel routines on a per-user basis. As data source, the public Twitter API is used. With the help of statistical indicators, outliers and non-humanoid actors are identified and removed from the dataset. So-called *individual spatiotemporal routines* are then computed, and *semantic places* such as "home" and "work" are extracted. These routines are assigned to suitably aggregated, regional units, so that both socio-demographic representativeness and intra- and interregional flows of people can be evaluated.

1.3. Research Questions

To extract meaningful patterns of human mobility from VGI, a set of coherent and robust methods is required. These methods should both be suited to data from social media D_{vgi} and produce results which match the structure and format of the reference, i.e., data from authoritative sources D_{as}. Furthermore, a confined region defines the study area A within which data are collected and analyzed. Together, these three components — methods, data, and study area — form a methodological framework which is used to answer the research questions outlined below.

It is important to consider that, given the uncertain nature of the data at hand, a different framework would possibly yield different answers, but, because of practical reasons, only one possible framework is assumed and used throughout this thesis. Thus, given a particular methodological framework, this thesis aims to answer the following research questions:

RQ1 *How representative is the spatial and socio-demographic distribution of users found in D_{vgi} of the overall population as measured in D_{as}?*

First and foremost, the term representativeness needs to be clearly defined. It is used here in the sense of how well the socio-demographic properties of a society as described through D_{as} are reflected in a sample of D_{vgi} that stems from the same geographical study area. Socio-demographic properties may be operationalized through many indicators such as language, culture, age, gender, economic status, and education. While some of these are hard to measure with the data at hand, others, such as age, gender and language, might be obtained more easily. As to a certain degree, socio-demographic properties are *spatially auto-correlated* (Cliff & Ord, 1970; O'Sullivan & Unwin, 2003), it makes sense to consider the (residential) location of users as an important component in the evaluation of representativeness. Namely, the spatial distribution of users may act as a *proxy* for certain socio-demographic factors, although these factors are hard to operationalize. For instance, it is known that people living in cities have a different lifestyle, and possibly, a different socio-economic background, from people living in rural areas, which is often manifested in political orientation and voting preferences (Lipset & Rokkan, 1967).

RQ2 *How do patterns of intra- and interregional mobility as inferred from D_{vgi} compare to mobility quantifications found in D_{as}?*

In other words, how are intra- and interregional flows of people reflected in D_{vgi} and are they similar to the ones in D_{as}? The answer to this question should not only help to assess the quality and validity of egocentric, geosocial VGI in this domain. It should also provide first insights about the feasibility of complementing or replacing authoritative sources with VGI.

RQ3 *How does spatial and socio-demographic representativeness as measured in* RQ1 *influence the results of* RQ2*?*

In other words, can potential deviations in human mobility in D_{vgi} from those in D_{as} be explained with unequally represented regions in D_{vgi}? If this is the case, it does not only mean that spatial and, possibly, socio-demographic representativeness is not given in D_{vgi}, but also that this significantly affects higher-level analyses of the data. As part of this question, one can ask whether it is possible to incorporate knowledge about spatial representativeness of D_{vgi} to calibrate such analyses.

Given these research questions, the overall aim of this thesis is thus i) a feasibility study which assesses the comparability of VGI to existing, authoritative data, and ii) a thorough analysis of the representativeness of those data.

1.4. Thesis Outline

- In the following Chapter 2, the necessary theoretical background in the field of geodemography and spatiotemporal analysis is provided, and the concept and process of *geotagging* is explained. At the same time, related work is highlighted.

- Chapter 3 introduces the study area and the various kinds of data sources used throughout this thesis.

- Chapter 4 explains the methodological process from the data collection to the eventual comparison of egocentric, geosocial VGI to authoritative data.

- Chapter 5 presents the results with regard to the representativeness and validity of egocentric, geosocial VGI for quantifying human mobility.

- In Chapter 6, the results are questioned with regard to the robustness of the applied methodological framework, and the research questions are addressed.

- Chapter 7 summarizes the findings of this work and makes recommendations for current and future research with egocentric, geosocial VGI.

2. Background

This chapter provides the necessary theoretical background, from the mechanisms of *geotagging* to the concept of *spatiotemporal routine*, which is heavily made use of in the methodology of this thesis. At the same time, it presents academic work which is related to this thesis in terms of the type of data used or the type of analytical questions asked. Specifically, applications of egocentric, geosocial VGI in the fields of *geodemography* and *human mobility* are highlighted since these constitute the validation background and the concrete use case of this thesis, respectively.

2.1. Geotagging

Geotagging, in this context, refers to the addition of explicit *geographical information* to digital content such as photos, videos, weblogs, as well as social media status updates, by the user who produces that content (Goodchild, 2007). In many of such systems, geographical information can take a multitude of forms such as place names or geographical coordinates. Moreover, geographical information can be described using different levels of spatial granularity, from the country level down to particular street addresses, which again implies different levels of precision (Worboys, 1998).

For clarity, one must distinguish between geotagging from explicitly disclosing a geographical position (for the sake of it). In the latter case, many applications, so-called Location-based Services (LBSs), exist (Schiller & Voisard, 2004). These provide users with location-based information or reward them with coupons or discounts for checking into a particular shop or restaurant (see for example Rao, 2012). As a special case of LBSs, LBSNs, such as Foursquare, combine aspects of social networks with explicit disclosure of geographical information (see Cramer et al., 2011 for a good introduction to the mechanics of Foursquare). This was described by some as "social-driven location sharing" (Lindqvist et al., 2011), and the underlying motivations for participating in such a network might be different from those involved in geotagging (Section 2.1.3).

In this thesis, data from LBSs and LBSNs play a minor, but still not negligible role in that some of their location data are automatically pushed to Twitter accounts (Section 3.3.5).

The following section gives an overview of geotagging functionality and a brief look at the underlying technology. It also examines what motivates people to geotag their content. In the end, the implications and dangers for personal privacy are discussed.

2.1.1. Examples of Geotagging Functionality

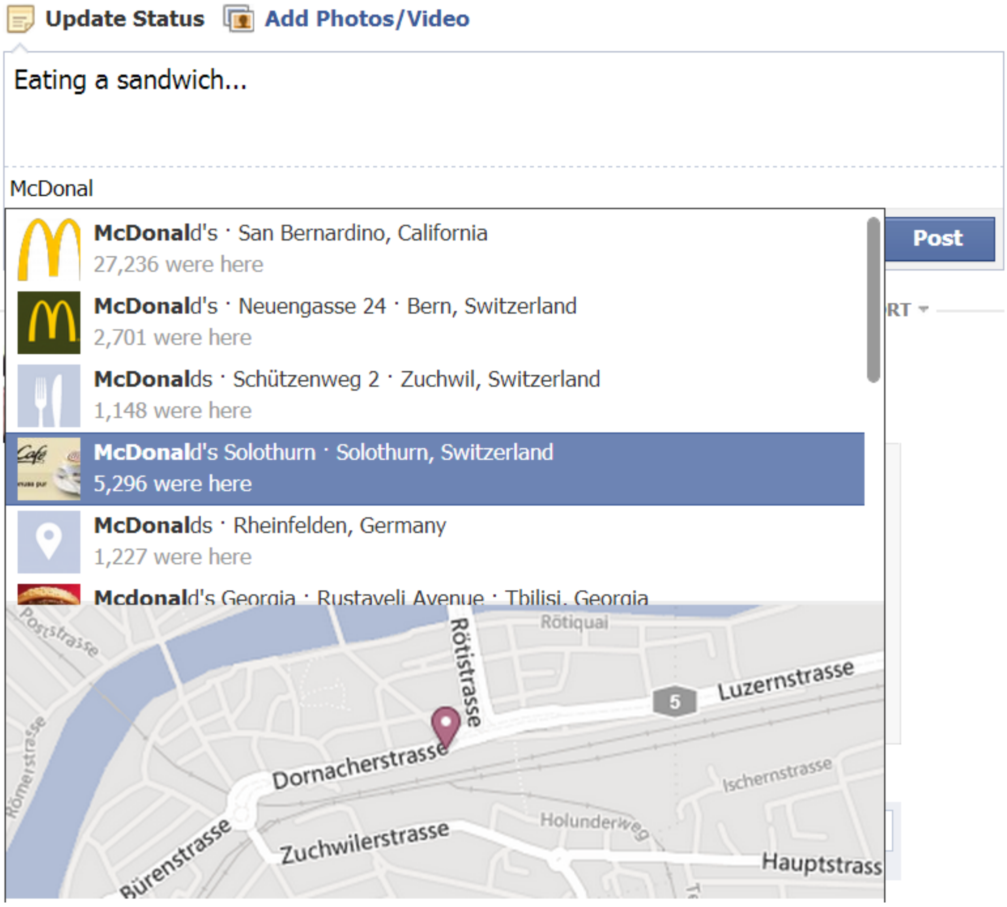
Many social platforms and networks have incorporated geotagging features into their user interfaces. In general, users can either directly specify a place name or choose from a list of suggestions nearby the current location. Based on the privacy settings of a particular user, his or her location updates are visible to different audiences, from only close friends to the whole world (Tang et al., 2010).

Facebook allows its users to annotate any status message or uploaded media with geographical metadata (Figure 2.1). Apart from cities, users can also choose other granularities such as countries and neighborhoods. If a particular venue, such as a restaurant, is registered on Facebook, users can also directly check into this venue without having to write a status message (Facebook Help Center, 2014).

On Flickr[1], a popular image and video hosting website, users can manually annotate uploaded content with geographical locations — if the image or video has not already been georeferenced by the GPS sensor of the camera. A map interface allows the user to choose an appropriate level of detail for the place description (Figure 2.2). For example, a set of images could either be *bulk-tagged* with "Zürich, Switzerland", or specific images could receive fine-grained neighborhood information such as "Oberstrass, Zürich, Switzerland" (Hollenstein & Purves, 2010). In either case, location is not only represented by place names but also by unambiguous geographical coordinates. Apart from this feature, users can also annotate their pictures with so-called *tags*, i.e., short textual descriptions which usually consist of only one or two keywords (Flickr Help FAQ, 2014). Since many of these tags contain implicit geographical information (e.g. "beach", "mountain"), they can be used in conjunction with the corresponding geographical coordinates to infer popular descriptions of places (Purves et al., 2011).

Twitter began incorporating geolocation features into their API in late 2009

[1] http://www.flickr.com



Eating a sandwich...

McDonal

McDonald's · San Bernardino, California
27,236 were here

McDonald's · Neuengasse 24 · Bern, Switzerland
2,701 were here

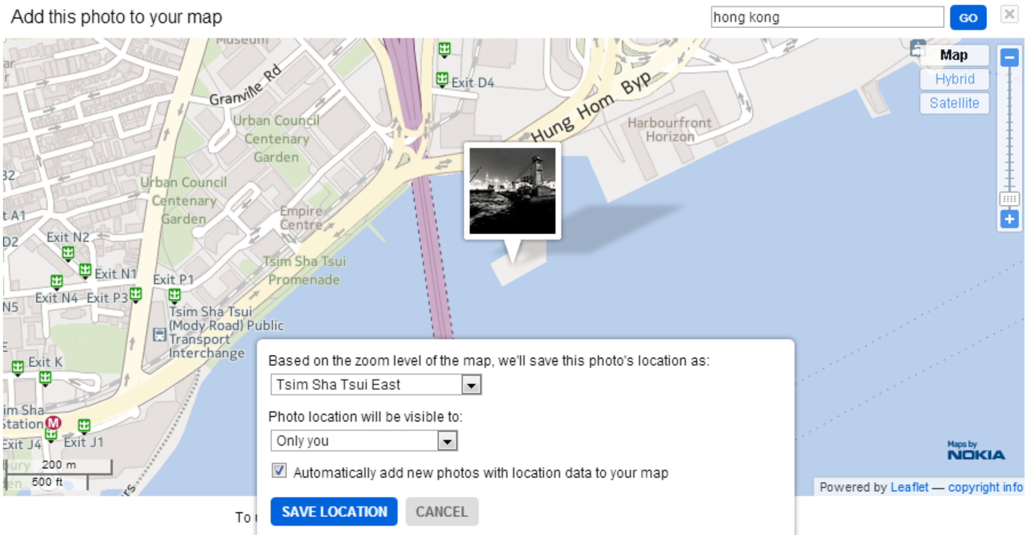
McDonalds · Schützenweg 2 · Zuchwil, Switzerland
1,148 were here

McDonald's Solothurn · Solothurn, Switzerland
5,296 were here

McDonalds · Rheinfelden, Germany
1,227 were here

Mcdonald's Georgia · Rustaveli Avenue · Tbilisi, Georgia

Figure 2.1: *Facebook geotagging functionality. When writing a status message or posting a photo or video, the user can type in his or her location or choose from a list of suggestions.*



Add this photo to your map

hong kong

Based on the zoom level of the map, we'll save this photo's location as:
Tsim Sha Tsui East

Photo location will be visible to:
Only you

☑ Automatically add new photos with location data to your map

Figure 2.2: *Flickr geotagging functionality. By navigating and panning on a map, the user can place an image or video and specify the desired level of spatial granularity.*

and made geotagging functionality available from early 2010 (Sarver, 2009). Similar to Facebook, users can annotate their Tweets[2] with place names from a list or let their mobile device sense their current GPS position. The mechanisms of the Twitter geotagging functionality are covered in more detail in Section 3.3.2.

In conclusion, systems and features for sharing and recording users' geographical positions are becoming increasingly popular, not only in the form of altogether new applications but also on established social media. Some actually argue that location data are the new "currency of the web" and that an actual "location revolution" is happening right now, with the market for LBSs becoming a multi-billion-dollar one (McLaren & Kennedy, 2013).

2.1.2. Technology

For geotagging to be possible, the client, that is, the device of the user, needs to somehow communicate its current position to the server. Even though it is often unclear how this is specifically implemented in different applications, the underlying technologies are likely one or a combination of the following.

Since most personal computers do not have an integrated GPS module or cellular antenna, location is often inferred from their current Internet Protocol (IP) address. This can be done with different lookup techniques (Padmanabhan & Subramanian, 2001), but the precision of the resulting location may be very limited (Gueye et al., 2007).

Most social platforms offer corresponding *apps*, i.e., simple applications which were developed for a particular mobile operating system and have a limited set of functionality when compared to desktop applications. These apps make use of the various location APIs which are exposed by the mobile device. This ranges from access to the GPS module over mobile telephony cell location to WiFi access point IDs (Adams et al., 2003). The accuracy and precision of the resulting geographical information can range from a few meters up to many kilometers; it depends on multiple factors, among which are the type of the sensor, completeness of reference databases (e.g. for WiFi access points), and quality of *reverse-geocoding* of coordinates into places (Spirito, 2001; Y.-C. Cheng et al., 2005).

To summarize, many technologies for sensing someone's location exist, and it is often not clear how exactly they are made use of in different applications. Positional accuracy, for instance, may range from a few meters to a few

[2] On Twitter, a status message is typically called "Tweet", see Section 3.3.

kilometers, and assumptions need to be made because content providers do not reveal the details of their systems. This poses a challenge for researchers dealing with geotagged content; they need to rely on potentially unreliable meta data.

2.1.3. Motivations for Geotagging

In order to enable geotagging functionality, users normally have to *opt-in*, i.e., they have to explicitly agree to disclose their location and share it with a particular audience. This and other reasons might be responsible for the still relatively small amount of geotagged content on the Web. In fact, its share of all content often lies below 10%. On Flickr, for instance, only 3–5% of all photos are geotagged (Friedland & Sommer, 2010; Catt, 2009). Several studies found figures of the same magnitude for Twitter, from 0.4% over 0.6% to a very recent 2% (Z. Cheng et al. (2010), Takahashi et al. (2011), Leetaru et al. (2013), respectively). Of course, these figures depend on the kind of data analyzed and the mechanisms which were used to collect the Tweets. The last of the mentioned surveys actually comes to the conclusion that geotagging behavior varies greatly among different regions and cities, e.g. 4.6% for New York and only 1% for Istanbul (Leetaru et al., 2013). From the temporal order of these sources, one can observe a slight increase in geotagged content over the years, but overall adoption stays very low.

For people who choose to use geotagging, however, motivations to do so are manifold. Geotagging photos, for instance, might be useful for the later compilation of a personal travel diary or just for quickly locating and filtering them, as Friedland and Sommer (2010) argue. In a lab experiment conducted by Wagner et al. (2010), a number of students were presented with fictional scenarios where they had to decide whether or not to disclose their location on a social networking application, and in what detail. The results show that willingness and motivation to do so are highly dependent on the kind of location to be broadcast and on the recipients. For instance, participants were rather reluctant to let their bosses know they were at home, whereas this posed no problem when they were at work. Participants also stated that they would only share their location when there is a specific need and purpose, but certainly not constantly. In another study, Barkhuus et al. (2008) claim that location sharing is not only done for communication and coordination purposes, but also for portraying a certain lifestyle and for self-representation. Using similar arguments, Tang et al. (2010) distinguish between *purpose-driven* and *social-driven* location sharing, where the former happens for mainly pragmatic reasons and the latter for promoting and sustaining social capital within a network. While purpose-driven sharing mainly applies to *one-to-one* communication, social-driven makes most sense in the domain of social networks with *one-to-many* communication. After conducting interviews and a small survey, the authors conclude that social-driven sharing is often used to attract attention and boost self-representation. The authors also point out that *one-to-many* sharing, as it is the case on Twitter, involves more complex reasoning than *one-to-one* or *one-to-few* sharing, where location is broadcast to a small audience only. Lastly, Cramer et al. (2011) conducted several interviews with and a survey on users of Foursquare. Beyond the categories of purpose- and social-driven sharing, they discover motivations endemic to LBSNs (see Section 2.1.1), such as check-ins for discounts, discovering new venues, and getting to know new people, as well as pure curiosity about who else is checked into a particular place at the moment. Concerning the ability to push check-ins to Twitter and Facebook, and thus to a larger audience, the results show that participants normally do this with only a small fraction of their check-ins. According to the participants, the main reason for this behavior are not privacy considerations but fear of "annoying" the larger Twitter or Facebook audience with irrelevant "spam".

From these findings one can deduct that those locations which are actually shared are often highly selective. They are the result of a complex decision making process, involving factors such as the type (and size) of audience, the type of place to be shared, external or monetary benefits, gain of social capital, and privacy considerations.

2.1.4. Privacy

After a series of incidents, such as the proliferation of an application which allowed to locate and identify women nearby one's current position by accessing publicly accessible Facebook profiles and location updates (Brownlee, 2012), as well as other demonstrations of possible abuses (Perez, 2012; Pot, 2011), users have become aware of the possible threats associated with repeatedly broadcasting their whereabouts.

Apart from possible abuses, e.g., *stalking*, the automatic inference of movement profiles or the re-identification of supposedly anonymous users poses another, possibly even more severe threat to privacy. In recent years, academic research has started to focus on *geoprivacy* and a wealth of studies has been published. Krumm (2007), for example, evaluates various techniques for inferring somebody's residential address, based on a set of GPS positions. In an early attempt to show the vulnerability of location samples to re-identification, Gruteser and Hoh (2005) achieve to extract sequences of fully anonymized GPS positions[3] which belong to the same user. These tracks could be used to re-identify the user based on his or her frequently visited places. Using a semi-automatic approach including interactive visualizations, Jedrzejczyk et al. (2009) analyze the paths of several individuals with the help of geographical and textual information gained from an LBSN. Afterwards, they confront the respective individuals with these findings and report on their reactions, which range from indifference to outrage. In a follow-up article, the authors question focus groups and conduct interviews in order to evaluate the impact of real-time feedback on location sharing practices (Jedrzejczyk et al., 2010). They find that people tend to restrict access to their location traces or stop posting locations altogether as soon as they become aware that somebody else is observing them.

One can conclude that, even though LBSs and LBSNs are becoming increasingly popular, the fear of privacy invasions is certainly among the main reasons for people not to share their geographical positions. This was also indicated by the research insights presented in Section 2.1.3. It remains to be seen whether and how this problem will be dealt with, and what role, in general, privacy will play in the future.

2.2. Spatiotemporal Routine

In the scientific community, individual, spatiotemporal movement of people and other objects has been researched for many decades. From Hägerstraand's prominent *time geography* (1970), where space and time are viewed as inseparable concepts, to Hornsby and Egenhofer's *geospatial lifelines* which are modeled over multiple granularities (2002), many concepts have been defined, discussed, and applied to real-world data. In this section, fundamental concepts of movement data, and the process of inferring individual travel routines from them, will be highlighted.

2.2.1. Concepts of Movement Data

As the three components which form a spatiotemporal *trajectory*, i.e., a series of visited locations ordered by time, N. Andrienko et al. (2008) introduce *space*, *time*, and *population*. The former two can only be described through a certain *reference system*, for example, space can be referenced by coordinates but also by

[3] Each position was anonymized so that one could not tell which track it belongs to.

dividing it into certain areas. Just as space, time can be structured into areas or temporal periods, for instance, all events happening between 15:00 and 16:00. Moreover, it can be referenced by the standard Gregorian calendar or in nested, cyclical systems, e.g., day of the week or week of the year. Which reference system is chosen depends on the particular problem, the data at hand, and the analytical questions to be answered. The third component, population, consists of the set of *moving objects* or *entities* which occupy different spaces at different times.

The main contribution of the authors, though, is the establishment of a taxonomy of analytical questions relevant to movement data. In their framework, both time and population form the reference system, that is, the *independent* variables which determine the current position of a moving object. On top of that, they distinguish between so-called *elementary* and *synoptic* questions. The former deal with moments in time and single entities — a possible question could be "Where was entity i at time t?". The latter deal with groups of entities and behavior over temporal intervals, so a possible question could be "Did group i of moving objects spatially converge with group j during interval δt?". Synoptic questions often seek to detect *behavior* of collective or single moving entities over time, e.g., "What are the typical trajectories of bees of a certain hive?". Another important concept mentioned by the authors is therefore *pattern*, which is the *representation* of behavior in some language, e.g., a graphical, mathematical, or textual description.

Findings patterns in movement data of humans, animals, or other moving objects has been one of the most vividly discussed research topics in the last few years. Gudmundsson et al. (2012) give a compact overview over common tasks and problems encountered when analyzing movement traces and list possible fields of applications. These range from behavioral ecology over mobility and transportation to movement in abstract spaces[4]. Methodological research in the field of GIScience can take many forms. For instance, Dodge et al. (2008) call for a general agreement on the relevant types of movement patterns. Having such a taxonomy would greatly facilitate the development of pattern recognition algorithms that are efficient, effective, and — most importantly — generically applicable. The authors therefore establish a classification which distinguishes, for instance, generic patterns, such as "concentration" and "divergence", from behavioral patterns, such as "fighting" or "flocking". Special attention is also directed to the interaction of both the spatial and temporal dimension of move-

[4] See Cöltekin et al. (2009) for a usability study which analyzes eye movements.

ment. Laube and Purves (2011) focus on the scale-dependency of analyses with movement data. Using GPS positions of grazing cows, they show that the choice of a particular temporal scale (e.g., an aggregation of all GPS readings during one minute) may influence the outcome just as strongly as the choice of the spatial aggregation. The authors thus stress the importance of cross-scale sensitivity analyses if movement parameters such as speed, turning angle, or sinuosity are to be estimated. Often, traditional (spatial) data mining techniques are also combined with *visual analytics* to leverage the human brain's capacity for visual thinking (G. Andrienko et al., 2010).

In summary, finding and formulating patterns is a means of reducing complexity and achieving a high-level description and interpretation of the spatiotemporal behavior found in data, and this particular notion of pattern will be used throughout the thesis. Although the above mentioned studies also strive to find and describe patterns, they usually rely on near-continuous, frequently sampled location data — a kind of data that is thoroughly different from geotagged social media content, which will be illustrated further below.

2.2.2. Spatiotemporal Routine

While a particular location can be looked at from a physical perspective (e.g. using coordinates or colloquial descriptions), it can also be annotated with domain- or subject-specific *semantics*. This leads to the notion of *place*, which "relates geography to human existence, experiences and interaction" (Tuan, 1977, as cited in Edwardes and Purves, 2007, p. 109). For an individual, a place might have a significant meaning, for instance, it might be someone's home. Indeed, places frequently visited by individuals usually fulfill specific functions or activities, e.g., living, working, leisure, social interaction, et cetera — in other words, they *afford* a certain action or meaning (Jordan et al., 1998). This leads to the simple definition of place used throughout this thesis: A confined region which affords a particular function for someone, such as housing or working.

At the same time, human life and thus human movement typically follows a certain *routine*, which means that the same, few places are repeatedly and regularly visited over time. Specifically, these places are usually visited or occupied during certain, cyclically ordered times, i.e., only during certain times of a day or days of a week, or even only during certain seasons. Although rather intuitive, this was also detected by a range of seminal papers using different types of individual movement data (Song et al., 2010; Gonzalez et al., 2008; Brockmann et al., 2006). Here it is assumed that such routines can be detected in egocentric, geosocial VGI as well, and described as patterns. The notion of spatiotemporal routine with regard to an individual object

$$SR_o = \{l(t) | freq(l(t)) \geq \alpha\} \tag{2.1}$$

is thus defined as a set of repeatedly visited places $l_1(t), l_2(t), ..., l_n(t)$ at specific times t, where $freq(l(t))$ denotes an appropriate function to measure the frequency of each $l_i(t)$ in relation to the remaining $l(t)$ and α is an appropriate threshold for this frequency. Note that the temporal ordering of the visited places is irrelevant. On the other hand, each $l_i(t)$ is associated with a particular function, as described above. This *semantic place annotation* is thus directly dependent on the time t at which $l_i(t)$ is frequently visited. For instance, if a place is frequently occupied during evening or night time, its possible semantic annotation could be "home" or "leisure" — given that knowledge about typical human behavior is provided. Using this and similar concepts, several research papers have striven to extract and label semantic places from a variety of data sources such as activity diaries (Krumm & Rouhana, 2013; Partridge & Golle, 2008), mobile phone logs (Wang et al., 2012), LBSN check-ins (Ye et al., 2011) and GPS data (Liao et al., 2007a; Zhou et al., 2007). They usually applied rule-based methods or machine learning techniques and verified their results with dedicated test datasets.

A spatiotemporal routine can thus be looked at as a simple pattern for the description of an object's routine. Even though it stays very general, the definition is sufficient for the purposes of this thesis. It will be applied and further explained in Section 4.4.3.

2.2.3. Regularity in Spatiotemporal Routines

In order to justify the search for spatiotemporal routines and — in the best case — be able to detect them, the movements of an object must be regular to a certain degree. A simple but effective approach for quantifying such regularity is proposed by Cranshaw et al. (2010). In their work, each *location observation* $e \in E_o$, i.e., each spatiotemporal event of a particular object o, is represented by a vector containing the components visited location L, day of the week D, and hour of the day H. To study regularity, they restrict this vector to different combinations of each component $R \subset \{L, H, D\}$, e.g., $\{L\}, \{L, D\}$ or $\{L, H\}$. Given $R = \{L, D\}$, for example, the observations $e(R)$ would be all actually observed *configurations* of locations and days of week for a particular object. The probability for a particular configuration r is defined as

$$p(r) = \frac{|\{e \in E_o : e(R) = r\}|}{|E_o|} \tag{2.2}$$

All observed configurations of R thus give a probability distribution $P(R)$, which can be used to compute the so-called *entropy*, a well-known concept from information theory (Shannon, 1948). Entropy can be colloquially described as a measure for disorder (or randomness) in a system. The more unpredictable a system behaves, the higher is its entropy and the more information is needed to encode such behavior. Cranshaw et al. (2010) use this concept as a description of regularity in someone's spatiotemporal routine — a routine with high regularity is easily predictable and thus associated with low entropy. Entropy is formally defined as:

$$Entropy(E_o, R) = -\sum_{r \in E_o(R)} P(r) \log P(r) \tag{2.3}$$

Using the simple case of $R = \{L\}$ and assuming that a user visits three different locations of which one is visited very frequently and the other two only marginally, this results in a low entropy. From another perspective, this time using $R = \{L, H\}$, a low entropy means that an object only visits one or very few locations for a given hour of the day or that a particular location is only visited during a very restricted temporal interval. Therefore, a low entropy corresponds with a high regularity for a given restriction R. Even though it is rather simple, this measure is well suited to get a preliminary impression of a possible manifestation of a spatiotemporal routine. In other words, for a spatiotemporal routine as defined above to be recognizable, at least some restrictions of R need to have a sufficiently low entropy. The concept will thus be used as a precursor for the computation of spatiotemporal routines in Section 4.4.2.

2.2.4. From Continuous Trajectories to Episodic Movement Data

As could be seen from the examples in the introduction, the growing dissemination of location sensing technology allows researchers to study individual and collective movement behavior based on real-world, near-continuous location data. These data are often so densely sampled that actual movement parameters such as speed, direction, and other second- and third-order characteristics can be inferred and fed to sophisticated pattern detection algorithms. Such movement data is *time-based*, i.e., sampled at a specified, relatively short rate, whereas the type of data used in this thesis is *event-based*, which means that position and time are only recorded whenever a certain event occurs — in this case, when a user decides to broadcast a geotagged Tweet.

N. Andrienko et al. (2012, p. 241) call this type of movement data *episodic* — "data about spatial positions of moving objects where the time intervals between the measurements may be quite large and therefore the intermediate positions cannot be reliably reconstructed ..." — and acknowledge the lack of methods for and experience with this kind of data, although some studies combining data from fixed, *location-based* sensors with visual analytics are mentioned (Stange et al., 2011; Wood et al., 2011). Whereas continuous movement data may fully reflect the spatiotemporal routine of an object, episodic movement data can be looked at as an irregular, infrequent sample thereof, which brings with itself a range of uncertainties (see next section).

In literature, there exist only a few studies so far which use episodic movement data for inference of individual movement regularities or spatiotemporal routines. For example, Williams et al. (2012) use three different data sources, namely a metropolitan transport system, a university campus WiFi system, and Foursquare, to study patterns in visiting regularity. All of the three datasets have in common that users have to voluntarily check into a particular location (electronic payment gates, WiFi hotspots, and Foursquare venues, respectively). As a regularity measure the authors employ an approach from neurophysiology (see Kreuz et al., 2009) which compares *sequences* of visits to a particular location between multiple weeks. According to that measure, the more similar such sequences are, the more regularly visited is a location. The study finds that, while most users visit not more than one or two locations *at all*, actually only a small fraction of users visits *one or more* locations *regularly*, with the highest proportion found in the transport system data. This is somewhat astounding when compared to the findings of the studies presented in Section 2.2.2, but could have several reasons. Either the threshold for a location to be deemed as regularly visited was set too high, or the data were indeed too sparse. Another reason for the apparently irregular behavior of users could be the focus on weekly visiting patterns and not on intraday regularities, which might have yielded significantly different results[5].

Indeed, and as will be seen in Section 2.3, most research using episodic data focuses on the collective behavior of multiple users, bypassing the computation of individual routines. Even though such approaches are often pragmatic and easy to apply, they have several limitations, some of which were already mentioned in Section 1.1.3.

[5] Also see the Modifiable Temporal Unit Problem (MTUP) as defined by Cöltekin et al. (2011).

2.2.5. Uncertainties of Episodic Movement Data

The irregularity and sparsity of episodic movement data result in a range of uncertainties mentioned by N. Andrienko et al. (2012):

1. A lack of information about the spatial positions of objects between recorded positions (*continuity*).
2. A lack of precision concerning the recorded positions due to imprecise sensors (*accuracy*).
3. A lack of precision concerning the number of recorded objects at a particular sensor (*coverage*).

Note that the second uncertainty is actually not necessarily inherent and restricted to episodic movement data as one could also regularly and frequently sense an imprecise location, e.g. through mobile telephony cell information. At the same time, episodic data may also be recorded with reasonably precise sensors, as it is the case with the data used in this thesis. The third uncertainty does not apply to the kind of data used in this thesis, as it assumes fixed sensors such as Bluetooth signal detectors.

In the case of egocentric, geosocial VGI, the sample is, on the other hand, directly determined by the user himself. This means that the researcher cannot control the mechanisms of data production and publication, which results in additional, specific uncertainties:

4. A lack of understanding of the *exact* mechanisms of the data provider, in this case, the Twitter API. Although the features that are made use of in this thesis are reasonably well documented, it is not guaranteed that the system behaves as advertised (also see Section 3.3.3).
5. A lack of knowledge about the *sampling motivation* of a user. This means, for instance, that it cannot be known whether a user prefers to only sample his or her location whenever he or she stays at a particular place, e.g., at home, or whether this user *actually* spends most of his or her time there (also see Section 2.1.3).
6. A lack of assurance that a user tells the truth. LBSN do generally not implement mechanisms which verify the authenticity of geotagged content.
7. A lack of assurance that a user fulfills certain criteria in order to be suited for the task at hand. It is, for instance, not fully known whether a user is

human, whether he or she actually resides in the study area, et cetera (also see Section 4.3).

While some of these points can somehow be handled with appropriate data processing steps, others must be taken as given and assumptions about the data must be made. In the remainder of this thesis, especially during the computation of individual spatiotemporal routines, these uncertainties are therefore explicitly taken into consideration.

2.3. Geodemography

Besides the concepts related to episodic movement data and spatiotemporal routine, two others are of vital importance for this thesis. The first, *geodemography*, is often used as a framework for putting geotagged social media into socio-demographic context. It is the "science of profiling people based on where they live", according to one definition (rda research, 2013), and involves gathering and inferring demographic statistics about a particular region's population (Harris et al., 2005). The underlying assumption is that different socio-demographic groups are not equally distributed in space and that characteristics of residents are (positively) spatially auto-correlated (Cliff & Ord, 1970), with *segregation* being the extreme manifestation of this phenomenon (Gregory et al., 2009).

Although the purpose of collecting and analyzing such data is often targeted advertising, governmental statistical offices make use of the concept for the purpose of regional, socio-demographic and socio-economic comparison. They also use indicators to derive typologies of space such as the ones presented in Section 3.2, and this often involves statistical classification of political entities, such as municipalities.

In the context of this thesis, the geodemographic typologies introduced later on will mainly be used to assess the representativeness of the collected Twitter data. As the next sections will show, there also exist other usage contexts in literature.

2.3.1. Geodemographics as a Framework

Geodemographics of certain regions, usually defined by statistical offices, are sometimes used in literature as a framework or *benchmark* for evaluating the socio-demographic properties of egocentric, geosocial VGI. This section critically examines some of the research conducted in this field.

L. Li et al. (2013) analyze the background of Twitter and Flickr users by comparing the number of geotagged Tweets and photos, respectively, to socio-economic properties of Californian counties. In this "ecological correlation" approach, the counties thus act as reference units for comparing social media data with census data. The authors first employ a simple heuristic to determine local residents and exclude the rest, which are deemed to be visitors. Interestingly, not this number of local residents per county, but the Tweet density, normalized by county population, is used as a dependent variable in their model. Through Principal Component Analysis (PCA), the socio-economic factors which explain most of the variance in the dependent variable are found. This research has several shortcomings, of which the fact that the spatial distribution of Tweets and not users is taken as dependent variable, is the most severe. In doing so, outliers, i.e., heavily contributing users, are likely to bias the result. Therefore, while being an innovative approach for studying socio-economic properties of users, it certainly needs more careful consideration of the manifold noise in the data.

Kent and Capello (2013) use a similar approach for identifying demographics of VGI-contributing users, but with a different motivation. They argue that emergency response centers who use VGI as additional data source for gaining *situational awareness* of disastrous events need to know how consistent and reliable those data are. The authors see their study as a preliminary exploration of how this could be evaluated. In a first step, they collect geotagged messages and images from Twitter, Flickr and Instagram[6], which were broadcast during a wildfire event and classify them into event-related and event-unrelated groups. By using different spatial analytical techniques, they find that event-related UGC follows a different spatial pattern than other content, thus it is indeed determined not only by non-spatial but also spatial factors. In order to detect these factors and compare the spatial distribution of UGC with demographic properties, the authors aggregate the event-related content on the level of census blocks. Through applying Geographically Weighted Regression (GWR), they find demographic and spatial variables which explain a great deal of the frequency of wildfire-related content. While this research succeeds in linking VGI with authoritative, real-world data, the authors themselves acknowledge its limitations. For example, the configuration of census blocks is, in this case, highly heterogeneous and large, less densely populated areas might be less representative than smaller ones. This is related to the so-called Modifiable Areal Unit Problem (MAUP) (Openshaw, 1983), that is, the outcome of such

[6] A social photo upload service, http://instagram.com

analyses depends on the particular choice of spatial boundaries, a fact which is not accounted for in this study.

These examples have in common that already existing, census-block-level statistics are used to characterize users who produce geotagged social media content. While quantitative benchmarking as shown above is a certainly desirable development, comparing regional statistics with mere quantities of event data has its pitfalls, among whose is the already mentioned user contribution bias or "participation inequality" phenomenon (J. Nielsen, 2006; Ochoa & Duval, 2008; Haklay, 2012), which will be further explored in Section 4.3.2.

2.3.2. Geodemographics as a Result

Another type of research that can be found in the literature is concerned with inferring socio-demographic and cultural properties of regions *through* egocentric, geosocial VGI. Studies in this field often have the goal to characterize or distinguish areas based on how and, especially, *when* people in these areas use social media. The underlying assumption is that certain activities, or functions, are only conducted at certain times. Looking at the volume of the constant stream of social media activity at various times could thus give each region an inherent signature, which could in turn be used to derive clusters of regions with similar signatures and activities. Thus, in this type of studies, geodemographics do not act as a benchmark or reference but are rather the product of the data analysis.

For instance, Cranshaw et al. (2012) try to characterize urban neighborhoods according to information found in an LBSN. In particular, they collect Foursquare check-ins at venues spread over Pittsburgh, PA, and cluster these venues based on geographical and social distance. The latter is calculated by computing a vector of all users for each venue, where each component of the vector is the number of times the respective user has checked into this venue. Thus, venues with a lot of users in common are considered more similar. The resulting clusters are called *livehoods*, "a dynamic, almost live view of the social flows of people throughout the different parts of a city" (p. 64). This can indeed be looked at as a valid conceptualization of a neighborhood as opposed to the strict definition set by planning officials. The resulting livehoods are then evaluated qualitatively by interviewing both local residents and domain experts, and the evaluation shows that most of the resulting neighborhoods make sense in terms of culture and demography. While in this case, heavily contributing users and outliers might not be biasing the result because every user is equally weighted, demographic bias definitely plays a role, since some groups simply lack any representation in the data. However, the authors correctly interpret the very *absence* of Foursquare data in particular neighborhoods as livehoods which are being inhabited by low income, predominantly African-American residents.

Also using Foursquare data, Rösler and Liebig (2013) compute vectors of check-in frequencies for each of about 1,000 Foursquare venues in the Cologne, Germany, area. Each 24-dimensional vector contains the *overall* number of check-ins per hour at this venue, again not accounting for *distinct* users.Using these vectors and their spatial position as distance metrics, temporally and spatially similar venues are then clustered. Since each Foursquare venue is already categorized according to one of about 10 different activity types (for instance "shop & service"), clusters can be characterized by the activity types of the venues found in them, resulting in specific *activity profiles* for each cluster. Using these profiles, the authors manage to separate clusters related to nightlife from clusters related to workplaces. In the end, the authors derive spatially contiguous clusters, which serves the goal of having a complete, activity-based classification of the urban region of Cologne. This, in turn, could be used as additional input in town development planning processes, so the authors argue.

Wakamiya et al. (2011) try to extract *crowd behavior patterns*, i.e., spatiotemporal differences in population activity, from Twitter. They first partition the study area (Japan) into irregularly sized regions, based on a spatial clustering of geotagged Tweets. For each of these extracted regions, they then count i) the number of Tweets, ii) the number of tweeting users, and iii) the number of tweeting users who post from different locations, during specified time intervals (morning, afternoon, evening, night). Crowd behavior patterns are defined as typical, frequently occurring changes in these numbers between different time intervals, and the authors are able to extract four of them, which leads them to four types of distinct regions ("bedroom town [sic]", "office town", "nightlife town", "multifunctional town"). For example, a region where all three above mentioned indicators steadily rise from morning to evening and reach a maximum in the evening is designated a "nightlife town". Evaluation is done qualitatively based on regional knowledge and satellite images. Unfortunately, it is not very clear whether the approach actually produces interpretable results, especially because the semantic extraction of four significant patterns based on three indicators seems to be somewhat arbitrary. As it is often the case with the other research presented here, this work suffers from the fact that it tries to infer knowledge about the mobility of a population without looking at the representativeness of the sample dataset. In this case, the authors aim to enhance the study of urban characteristics with a novel method, yet they forget that an urban area also houses socio-demographic cohorts which are unlikely to be represented on Twitter.

In an earlier study, Fujisaka et al. (2010) collect geotagged Tweets over a one-week period in Japan and spatially cluster them to delineate discrete regions. They then look at individual users and check whether they enter (i.e. the user has never sent a geotagged Tweet from this region before) or leave a region (i.e. the user did not send a geotagged Tweet from this region afterwards). Doing so, they are able to calculate so-called measures of aggregation and dispersion, respectively. Unfortunately, the presented approach makes many assumptions which are hard to validate. For instance, people could still stay in a certain region but stop broadcasting location updates, yet they are considered to be leaving the region, a problem that is also discussed by G. Andrienko et al. (2013b).

Lastly, in a very recent article, Fuchs et al. (2013) use textual analysis of Tweets to infer *semantics* of frequently visited places. Using a dataset of several thousand geotagged Tweets, they are able to categorize their textual contents into a set of predefined topics such as "family", "work" or "health". By aggregating Tweets of the same topic and visualizing them in two-dimensional, temporal histograms (days of week versus times of days), it becomes apparent that some of these topics have distinct temporal patterns, e.g., increased activity on evenings and weekends for the topic "sports". In order to relate topics to geographical places, the authors then apply a clustering algorithm to each user's trajectory and extract significant places. These are in turn aggregated over the whole user base and for each cluster of significant places, frequently occurring topics are determined. By doing this, the authors are able to relate geographical regions with certain types of activities. Although this approach is innovative in that it tries to infer semantics of places by looking not only at the coordinates of a Tweet but also at its textual content, it is hard to tell whether it produces meaningful results. First of all, the authors define an almost too large of a number of topics, thus it is possible that Tweets are arbitrarily categorized into a non-expressive topic. Secondly, the results of the method are not evaluated, neither with ground truth about, for example, socio-economic urban characteristics, nor by manually looking at individual Tweet content. Nonetheless, one has to bear in mind that the focus of this work lies more on exploratory, geovisual analysis rather than inference, and that it is able to provide interesting findings about distinct temporal patterns of activities.

To conclude, although the approaches presented above are often innovative in terms of methodology and data sources, they suffer from a series of shortcomings, of which the lack of assessment of demographic representativeness is the most striking. It is particularly problematic that the demographics of the user base are seldom explicitly addressed, and possible mitigations are lacking altogether in literature. Additionally, prolific users are seldom addressed and may often distort the results when content *per se* is aggregated. Another problem is the frequent absence of quantitative comparison to existing geodemographic data, although this might be due to the different spatial configurations resulting from aggregation of the data. Since this thesis uses authoritative data as a quantitative reference, new spatial configurations, even if they inherently existed in the data, cannot be depicted. This might not fully reflect the structure of VGI, but is an inevitable compromise when validating such data.

2.4. Human Mobility

Another promising field of application for egocentric, geosocial VGI is the computation and modeling of *human mobility*. In this context, human mobility is defined simply as the varying spatial distribution of a population at different times, a phenomenon that is often researched under the term *population dynamics* (Bhaduri et al., 2007). This also comprises the quantification of flows of people between spatial entities. Studies concerning human mobility can thus involve anything from visualizing the usual paths of people moving through an airport (Jochem et al., 2012), over estimating daytime and nighttime population distributions in cities (McPherson & Brown, 2003; Freire et al., 2011), to studying daily commuter flows between parts of a city or between whole regions (T. A. S. Nielsen & Hovgesen, 2008).

Traditionally, gaining fine-grained population distribution data involved building models based on data from various static sources such as census data, transport hub locations, distribution and size of companies, et cetera, and applying sophisticated techniques such as *dasymetric mapping* (Bhaduri et al., 2007; Mennis & Hultgren, 2006). In the last few years, technical innovation and the wide dissemination and ubiquitous use of mobile devices have attracted a lot of interest from the research community, authorities, and mobile telephony operators, because detailed and continuous datasets are increasingly being made available by the latter. Locations of network usage allow an unprecedented insight into the distribution of network participants at different times, and therefore quite a few studies use mobile phone data to quantify the distribution and flows of people in urban areas (Loibl & Peters-Anders, 2012; Sevtsuk & Ratti, 2010; Mohan et al., 2008; Ratti et al., 2006).

Recently, the increasing amount of VGI has encouraged researchers to experiment with this new kind of data, and thus some studies are presented in the next section. In this thesis, the computation of intra- and interregional commuter flows will serve as a case study in the field of human mobility for comparing egocentric, geosocial VGI to authoritative data.

2.4.1. Human Mobility Through Egocentric, Geosocial VGI

With the goal of computing flows of people for disaster management, Aubrecht et al. (2011) propose an approach to compare the number of people checked into particular Foursquare venues at particular times to *working population density* census data. They collect check-ins which were posted during normal workdays and tessellate them into a regular grid in order to fit the format of the census data. Unfortunately, no details about the statistical comparison or results are presented. Also, and as the authors themselves reflect, neither socio-demographic properties of Foursquare users are assessed nor outliers are explicitly considered.

In order to extract temporal patterns of people's whereabouts in New York, Ferrari et al. (2011) analyze a dataset of geotagged Tweets. They collect all Tweets posted in Manhattan during a certain interval and categorize them into equidistant time slots. For each time slot, they then compute spatial clusters of high Tweet abundance and assign each cluster to a zip-code defined area. Using probabilistic topic models, they are able to infer patterns of daily routine, i.e., which areas are frequently visited in which temporal order. Socio-demographic properties of the contributing users are not considered, and the distribution of Tweets is not compared to population density. As in many other examples, not individual behavior but rather the temporal variation in spatial hotspots of Tweet activity is modeled.

In an attempt to study tourist behavior in Rome, Girardin et al. (2008) utilize a large dataset of geotagged photos from Flickr and foreign mobile phone call records to find areas of high tourist abundance. In particular, they divide the data into a regular grid of 250x250m cells and compute the temporal variations in the presence of visiting photographers and foreign mobile phone users. In addition to that, they chronologically order the photos of each photographer to reconstruct his or her movement through the city. In an aggregated form, these movements and visited places reveal information about tourists' most favorite places and popular travel routes. The authors recognize the importance of proper validation for these sources of data, and attempt to compare their findings with ticket sales at popular sightseeing spots. While they find a correlation between the sales and mobile phone usage, they are not reflected in UGC from Flickr. The authors conclude that such analyses are only meant to complement conventional data sources, but alternate strategies for validation with real-world datasets are to be found.

2.5. Summary and Research Gaps

This chapter started with introducing the properties and intricacies of egocentric, geosocial VGI as opposed to conventional data sources such as GPS. It then presented the scientific field of movement pattern analysis and showed that this kind of research normally relies on near-continuous, frequently sampled location data. It further formulated the concept of spatiotemporal routine, which is of vital importance for the case study in this thesis, and made the shift from conventional movement data to episodic movement data, towards which egocentric, geosocial VGI may be counted.

In the second part of this chapter, various research, which explicitly uses episodic movement data from social media to make inferences about spatiotemporal behavior of people, was presented. One can generally conclude that studies using egocentric, geosocial VGI are still very sparse. This might be due to the novelty of this data source, but also because working with episodic data has its difficulties. Indeed, several reoccurring problems can be identified; they confirm the need for more, fundamental research in this area. In particular, the following research gaps need to be addressed:

1. Properties and, especially, the representativeness of the analyzed users are seldom inquired. This has implications for spatial analyses, especially when the population is not equally represented in space. As universal statements about users on social media can likely not be made, a basic assessment of representativeness should precede any analysis working with such data.

2. The user base is often looked at as "given", meaning that outliers and non-humanoid actors are rarely taken into consideration. Methods for defining, identifying, and quantifying these less desired groups of users are thus needed. Furthermore, their impact on the outcome of analyses needs to be assessed.

3. Human mobility and other geographical phenomena are often modeled in

a crowd-centric way — conclusions are usually drawn based on absolute amounts of UGC and not on individual behavior. This may be problematic, because the "participation inequality" bias detected in other kinds of UGC (Ochoa & Duval, 2008) has to be assumed for geotagged content, too (see Section 4.3.2 for references). This has two implications for the research community: Firstly, the impact of such a bias on the outcome of analyses needs to be studied, and secondly, new methods for working with individual, episodic data are needed. As these data are profoundly different from conventional movement data, well-established methods of movement pattern analysis are not suited to fill this gap, but may possibly be adapted.

4. Lastly, while some authors call for evaluation and validation with authoritative data sources, this is almost never explicitly done. New kinds of authoritative data and new fields of applications, where both sources of data can be compared, need to be found.

3. Study Area and Data Sources

3.1. Study Area

As the confined study area for which the data are collected and compared, the country of Switzerland was chosen. This choice has several reasons, among which the relatively small area and population, the local knowledge of the author, and the quality and abundance of authoritative data are just a few. Being small and very densely populated, Switzerland is particularly interesting in terms of its cultural-linguistic and socio-economic heterogeneity. This setting provides a good basis for comparing the properties of the collected Twitter data according to their geographical and thus socio-demographic origin. Table 3.1 shows key figures for Switzerland (BFS, 2013b, 2013c, 2012c, 2010). The definition of *permanent residency* encompasses Swiss citizens with residency in Switzerland and foreign nationals with a residency or settlement permit for more than 12 months.

3.1.1. Political Organization

Switzerland is hierarchically organized into 26 federal cantons, 147 districts, and approximately 2,500 municipalities (Figure 3.1). While the cantons and municipalities enjoy a high degree of political autonomy, the districts are not actual political institutions but mere organizational units. They are often utilized to coordinate collaborations between neighboring municipalities and do normally not act as legal entities. In fact, every canton is free to decide how to organize itself in terms of districts, and some cantons do not even know this level — from a statistical viewpoint, these cantons themselves are looked at as districts. Districts are often used for statistical analyses on the federal level, because they can act as suitable aggregations of the underlying municipal samples, which are in many cases not representative. One needs to consider, though, that districts themselves vary greatly in terms of population size, from a mere 2,108 inhabitants in the smallest district to over 450,000 in the largest, in 2011 (BFS, 2012c).

Table 3.1: *Key figures for Switzerland.*

	Switzerland
Area	41,280 km^2
Permanent residents (2011)	7.95 m
Increase in permanent residents (2000–2011)	10%
Population density (2011)	192.7 per km^2
Real GDP (2011)	587 bn Swiss Francs
Increase in real GDP (2000–2011)	21%
Largest cities (2011)	Zürich (379,915 inhabitants), Genf (194,458), Basel (172,091), Lausanne (130,421), BERN (127,515)
Official languages (2010)	German (65.6% total population share), French (22.8%), Italian (8.4%), Romansh (0.6%)
Percentage of people above age 14 regularly using the Internet (2011)	~79%

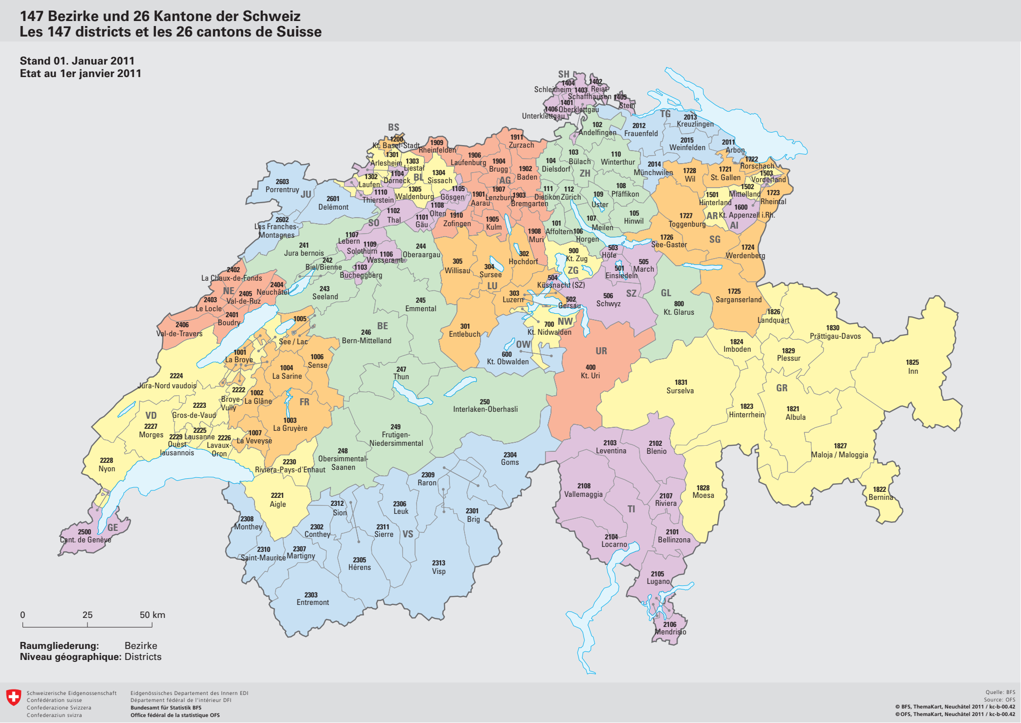


Figure 3.1: *The 147 districts and 26 cantons of Switzerland, 2011. Source: Swiss Federal Statistical Office (FSO).*

3.1.2. Employment and Commuter Mobility in Switzerland

In 2011, about 78% of the permanent residents of Switzerland, aged 15–64, belonged to the *working population* (BFS, 2013e). This definition encompasses all persons aged above 15 who work at least an hour per week, i.e., whose work can be counted towards the GDP. There is no upper age bound in this definition, but the retirement age in Switzerland currently lies at 65 years for men and 64 years for women. About 3% were officially registered as unemployed and looking for a job (SECO, 2012).

Approximately 13% of the working population worked part-time (50% and less), about 19% worked 50–90% of the time, and the majority, 66%, worked full-time (90–100%). The average work week for people working full-time encompassed 41.4 hours (BFS, 2013a, 2012d).

Approximately 9 in 10 people belonging to the working population used to leave their home to go to work[1] (defined here as *mobile working population*), and about 7 in 10 commuted to a place outside their home municipality, including places abroad. On average, the mobile working population's daily commute required 14 kilometers, and 30 minutes were needed for this. Roughly 0.8 million people were categorized as being in education (school kids, students, apprentices) and leaving their home to go to school (defined here as *mobile population in education*) (BFS, 2013e).

There does not exist any data on the "typical" day of such people, but one can guess from Figure 3.2 that most people leave home around 07:00–08:00 and return around 17:00–18:00, with some going home for lunch.

3.2. Authoritative Data

On the federal level, statistical surveys on a wide array of topics are usually carried out or commissioned by the Swiss Federal Statistical Office (FSO) ("Bundesamt für Statistik (BFS)"). While the last full census dates back to 2000, smaller, so-called *structural surveys*, which cover many themes such as households, religion and employment, are carried out since 2010. Usually, around 200,000 persons[2] are questioned by means of written questionnaires or Internet forms. From their answers, regionally comparable statistics are then extrapolated (BFS, 2008).

[1] People without a usual workplace, e.g. travelling salesmen, are exempt from this definition.

[2] Aged 15 or more, only permanent residents, approximately 3% of the population.

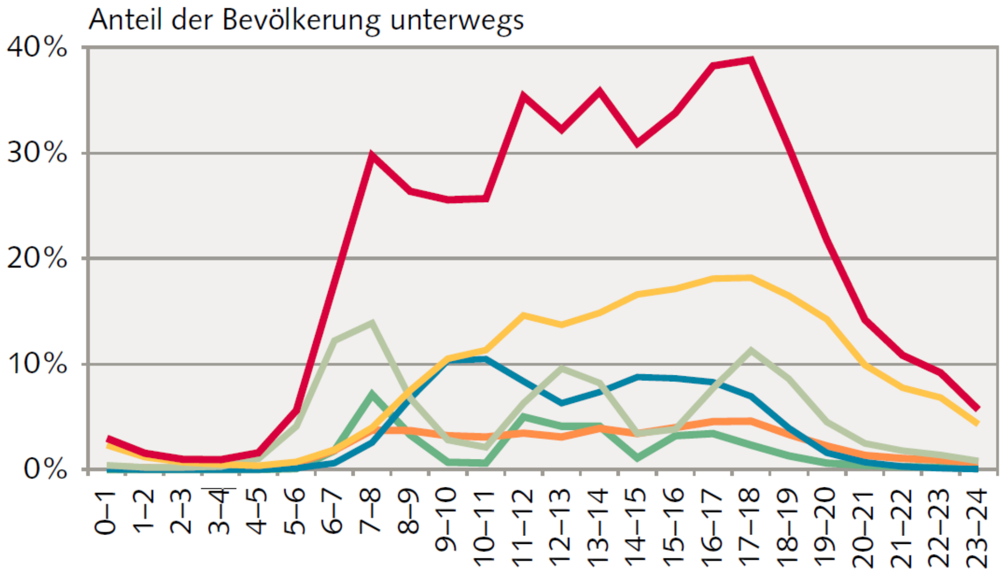


Figure 3.2: *Peak hours for commuters, 2010. The light green line shows the relative volume of work-related traffic by time of day, while the dark green line shows the volume of education-related traffic. Source: BFS (2013e).*

3.2.1. Structural Survey on Mobility and Transport 2011

The last structural survey concerning mobility and transport was conducted in 2011[3] and published in the summer of 2013 (BFS, 2013e). Approximately 330,000 permanent residents of Switzerland above age 15 were randomly sampled[4] and questioned about the origin and destination of their work place or school, means of getting there, and time and distance needed. Out of this sample, about 280'000 responded with valid data, although some of them did not provide enough data to be used in the procedure explained below. The data of people belonging to the mobile working population and the mobile population in education, as defined above, were used to compute so-called *origin-destination-matrices*, which tell how many people commute from each district to any other. These, in turn, were transformed to tables giving the number of staying[5], outgoing, and incoming commuters for each district (the so-called *commuter balance*), as data on the level of municipalities are often not significant.

Although the FSO had not yet officially published these figures at the time of writing, the author was provided with two datasets, one for work and one for education. Both of these do not only include the extrapolated numbers, but also the count of actual observations, as well as corresponding confidence intervals. As some districts have quite a small population, many of the extrapolations on the district level are still quite uncertain, with confidence intervals sometimes

[3] This is the reason why most of the above key figures are presented for 2011.

[4] Including additional samples financed by cantonal authorities.

[5] In this case, the place of work or education and the starting point of the commute lie in the same district.

amounting to over 50% of the extrapolated value, especially for data on people in education, where the population is considerably smaller (K. Freire of the FSO, personal email communication, June 18th, 2013).

Extrapolation from counted observations followed a three-step procedure. First, a sample, stratified by so-called *drawing zones*[6], was drawn, with every person in the sample having a certain drawing probability, whose inverse gave the basic weight for each person. Secondly, these weights were adjusted in order to correct for people who had not answered the survey, using data from similar, previous studies. Lastly, the weights were calibrated so that certain, known totals for each drawing zone were reached. The weights were then multiplied with the count of actual observations, which gave the final, extrapolated values (C. Freymond of the FSO, personal email communication, October 17th, 2013).

It should be noted that the commuter balances do not have an explicit temporal dimension, as it does not matter how many days per week a person goes to work or school in order for him or her to be included. Still, although one cannot say that they represent "daily" amounts, they probably are a good estimation thereof, since, as was shown above, a large majority of people works full-time.

3.2.2. Geodata, Spatial Divisions and Demographic Data

In terms of geodata, municipal boundaries of late 2011 were used to be in accordance with the thematic data available (BFS, 2012b). Those are available as geometric *shapefiles* in three different generalization levels, of which the most detailed was taken. For each municipality, a unique identifier and the identifier of the enclosing district are also provided.

In order to statistically compare different regions and geographical areas and, for instance, to make recommendations as to which municipalities might receive federal subsidies, the FSO defines and maintains a series of *spatial divisions*, which can be categorized into *regional-political*, *analytical*, and *typological* divisions (BFS, 2011). Table 3.2 gives an overview of the divisions which are of relevance for this thesis. Spatial divisions are based on the political entities of municipalities, i.e., once one possesses data on the level of municipalities, one can aggregate those according to the municipalities' membership to certain spatial divisions. In the context of this thesis, the spatial divisions and especially the municipal typologies can be used to infer, at least coarsely, the socio-

[6] Usually, cantons are used as drawing zones, although for a few cantons, the major cities are considered separately.

Table 3.2: *Spatial divisions of Switzerland as defined by the FSO.*

Category	Division	Description
Analytical	Greater regions	The seven greater regions are the only division based on cantons instead of municipalities, and were created in accordance with the statistical system of the European Union ("NUTS 2"). They are defined as relatively coherent political, economic and societal units.
	Linguistic regions	The four linguistic regions categorize municipalities according to the dominant language spoken and act as one of the most important divisions, since they reflect the publicly perceived, cultural "parts" of the country.
	Urban / rural	This division distinguishes *core cities* of agglomerations from suburban (agglomerations) and rural municipalities, as well as so-called *isolated cities*. Agglomerations are contiguous regions of multiple municipalities, which together amount to at least 20,000 inhabitants. Core cities are defined as municipalities with more than 10,000 inhabitants and which are part of an agglomeration, while isolated cities are not part of an agglomeration.
	Metropolitan regions	A division built on top of agglomerations, which distinguishes the 5 biggest metropolitan regions from the remaining urban and rural municipalities (Figure E.1).
Typological	Municipality types	In this system, each municipality is classified into 22 types according to various indicators, ranging from employment factors (including commuter balances) and tax volume to density of buildings and tourism activity.

demographic properties of Twitter users classified as residing in these municipalities.

Besides geographical boundaries, basic demographic facts of municipalities are also available from the FSO, namely from "Statatlas", the interactive statistical atlas. There, 2011 register data about population count, population density and further attributes can be downloaded (*demographic data*).

3.3. Twitter

Twitter was founded in 2006 by Twitter Inc., a San Francisco, CA, based company, and became publicly available in July of the same year. Twitter is a social network with so-called *microblogging* at its core, as it enables its users to write and broadcast up to 140 character long messages ("Tweets"). Twitter experienced rapid growth over the years, recently arriving at 200 million active users and over 400 million Tweets sent out each day (Wickre, 2013), compared to 5,000

Tweets per day in 2007 (Beaumont, 2010).

Social interaction on Twitter is relatively straightforward: users can have *followers*, i.e., other, *following* users who subscribe to their Tweets. Tweets can be *retweeted*, meaning that the same content is broadcast again by the retweeting user to his or her followers. This enables multiplication of reach and thus possibly fast spread of information. Tweets can also be *favorited* and *replied to*, with a reply being just a Tweet directed at a specific user.

Tweets can contain so-called *entities*, among which are *mentions* (denoted by an @, e.g., "Hey @jack!"), and *hashtags* (denoted by an #, e.g., "I like #geography"). Mentions are useful for social interaction because the mentioned user is immediately notified. Hashtags help to associate Tweets with particular subjects or events (e.g., "On my way to #gisconference2013").

Even though users can restrict the visibility of their Tweets to their followers, most accounts are public and thus visible to everyone, including non-registered users.

3.3.1. Demographics

Aside from the academic efforts to gain more understanding about Twitter user demographics (see Section 2.3.1), there also have been non-academic attempts which harvested large numbers of profiles and Tweets via the Twitter API (see Section 3.3.3). Such surveys are usually conducted by marketing and analytics companies who have access to the full Twitter stream and are seldom published. While the few ones that *are* published often present striking figures, the results have to be interpreted with care because, often, their methodology is not made transparent.

A study conducted by Sysomos Inc., based on 11.5 million profiles, comes to the conclusion that 53% of users are female, 66% of users are aged 15–24 and 21% are aged 25–34, based on self-disclosed age information (A. Cheng et al., 2009). Since the study was made in Twitter's early days, the results are biased towards users from the United States (65%) and other Western countries. According to a more recent survey which looked at 36 million profiles, still 51% of users are from the United States and 17% from the United Kingdom (Beevolve, 2012). Average gender distribution coincides with the results of the above study, but varies by geographical region. The study also looked at self-disclosed age information and found that a large majority (74%) is between 15 and 25 years old, followed by 15% of between 26 and 35 year old people. Another study by Smith and Brenner (2012) interviewed 1,729 American adults, aged 18 or older, who use the Internet, and found that 15% of them were actively using Twitter, of which 40% are aged 18–29, 35% are aged 30–49 and the rest is aged above 49. The survey also found that 40% of users are living in urban, 50% in suburban and 10% in rural areas. Other sources claim that, when Twitter started in the United States, the age distribution was skewed towards people in their late twenties and thirties, but teenagers caught up as Twitter became more and more mainstream (Lipsman, 2009). One could expect that this process also occurs in other countries and regions, as soon as Twitter is becoming more popular.

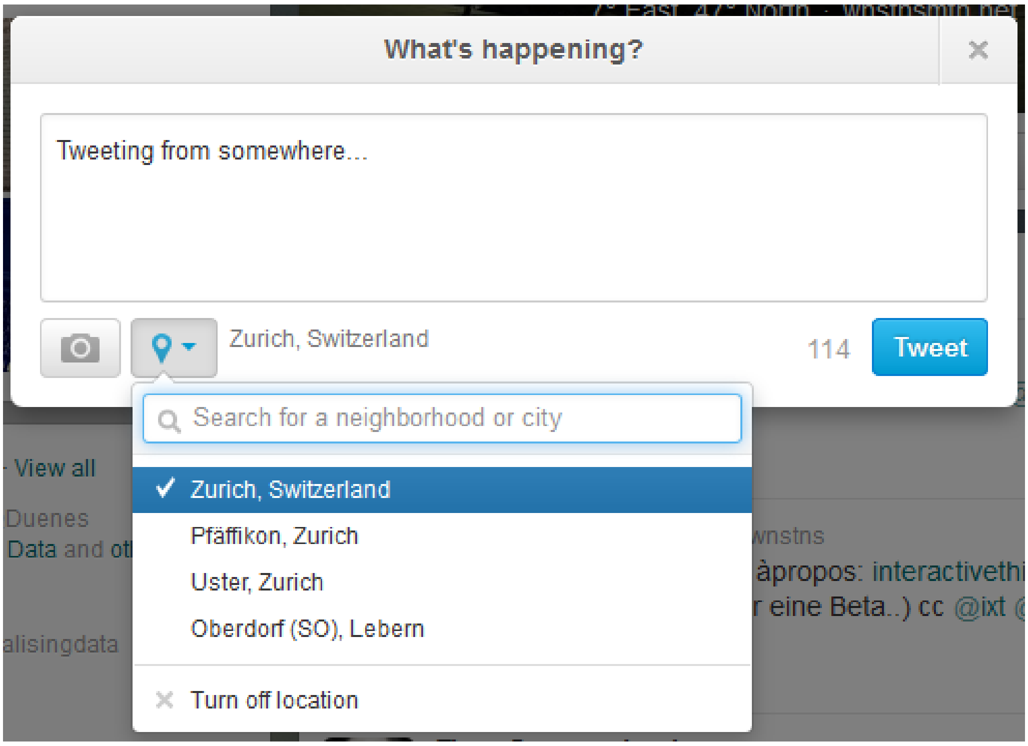
From these few examples it becomes clear that (socio-)demographic analyses of the Twitter user base are hard to obtain as well as difficult to interpret and compare. One fact which one can extract from these studies, tough, is that demographics may vary significantly between geographical regions, especially between the Global North and South.

3.3.2. Geotagging

On Twitter, there basically exist two explicit geotagging use cases. Once a user has opted-in, i.e., explicitly requested the activation of the geotagging feature, he or she can annotate Tweets with either his or her *precise* location or with a *place* (Twitter Help Center, 2013). The former uses device location (see Section 2.1.2) in order to attach precise — but not necessarily accurate — geographical coordinates to the Tweet. The latter must be chosen from a list of suggestions, which means that one cannot specify a non-existing or not yet cataloged place name (Figure 3.3). While the former use case is common for third-party apps on mobile devices, the latter is primarily used on personal computers without location sensors. According to Twitter Help Center (2013), precise coordinates are also stored in the latter case, in order to "improve the accuracy of [Twitter's] geolocation systems (for example, the way [Twitter] defines neighborhoods and places)".

Places can be specified using different levels of granularity, which are internally denoted as "admin", "country", "city", "neighborhood" and "poi" (see Section 3.3.5 and 4.3.1). When looked at from the user's perspective, this means that he or she can choose something like "Switzerland" ("country") but also "Twitter 3rd Floor Lunch Room" ("poi"). For some geographical regions, however, not all types are available. For instance, "neighborhood" is generally only available in large cities.

Even when opted-in, a user can choose not to geotag a particular Tweet but, by default, this feature is enabled, therefore making it likely that people



What's happening?

Tweeting from somewhere...

Zurich, Switzerland

Figure 3.3: *Annotating a Tweet with a place. The user is presented with a list of suggestions but can also search for a particular place in the catalog.*

continuously share their location. After having opted-out, users have the possibility to delete *all* their past geographical metadata at once (Twitter Help Center, 2013).

3.3.3. Twitter API

Twitter offers developers a comprehensive interface to most of its public data, the so-called Twitter API. It consists of the Streaming API, which allows access to a constant flow of Tweets filtered according to some criterion, and of the Representational State Transfer (REST) API[7] (Twitter Developers, 2013d). Only the latter is used for the purposes of this thesis and is therefore examined in more detail below.

The Twitter REST API is an implementation of the RESTful API as defined by Fielding (2000). This means that it accepts Hypertext Transfer Protocol (HTTP) requests (*GET*, *POST*, et cetera.) and returns responses in various formats, among which is Javascript Object Notation (JSON) (Crockford, 2006). JSON is a human-readable format for data storage that can be parsed by a multitude of programming languages. It defines a simple key-value storage format, as shown in the example in listing 3.1.

The Twitter API exposes a set of endpoints, called *resources*, in the form

[7] By the time of writing, the current version was 1.1.

of Uniform Resource Locators (URLs), which take mandatory or optional arguments and return specific kinds of data. For example, http://api.twitter.com/1.1/statuses/user_timeline.json?screen_name=xyz?count=5 returns the five most recent Tweets of user "xyz". In order to be able to receive data, the request needs a valid authentification token, i.e., it needs to be made on behalf of a registered *application* which is, in turn, owned by an authenticated user. Among other reasons, this is used to enforce limits on the number of requests that can be issued in a particular time window (see next section). Listing A.1 and A.2 Appendix A show an example GET request and the corresponding response from the server.

3.3.4. Limits

Twitter imposes several limits on the amount of data that can be fetched via its API. Each endpoint is subject to a limit regarding the number of results it returns per 15 minutes (Twitter Developers, 2013c). On top of that, particular resources such as *GET search/tweets* (see Section 4.2.1) only return a small fraction of the overall data volume. It is not fully clear to what extent content is filtered, but for the Streaming API, numbers such as only 1% of the full volume exist (Twitter Developers, 2012b).

3.3.5. Geographical Data Format

Whenever a Tweet is geotagged in some form, a combination of the JSON keys "geo", "coordinates" and "place" is populated with values (see lines 62–64 in Listing A.2).

The "geo" field contains World Geodetic System 84 (WGS 84) latitude and longitude coordinates of a single point (Listing 3.1).

Listing 3.1: *Contents of the "geo" field*

```
"geo":{
  "type":"Point",
  "coordinates":[
     47.20609337,
     8.57881543
  ]
}
```

The "coordinates" field is identical, except that coordinates are displayed in reversed order. Finally, the "place" field contains geographical metadata for a particular place (as introduced in Section 3.3.2), including the name, a Minimum Bounding Rectangle (MBR), the country it belongs to, the particular place type, and a unique place ID (Listing 3.2).

Listing 3.2: *Contents of the "place" field*

```
"place":{
  "full_name":"Neuheim, Zug",
  "url":"https://api.twitter.com/1.1/geo/id/acc076bb4ce682f7.json",
  "country":"Switzerland",
  "place_type":"city",
  "bounding_box":{
     "type":"Polygon",
     "coordinates":[
        [
           ...
        ]
     ]
  },
  "country_code":"CH",
  "attributes":{

  },
  "id":"acc076bb4ce682f7",
  "name":"Neuheim"
}
```

Based on the particular existence and combination of these three values, the type of geotagging can be inferred, and thus it is decided whether the Tweet is suited for futher processing (see Section 4.3.1).

As previously mentioned, some third-party apps, such as Foursquare, allow users to push their status updates to other social networks, such as Twitter and Facebook. At least in the case of Foursquare — which is arguably the largest contributing third party app[8] —, the format of the geographical data does not remarkably change. Since third-party status updates only constitute a small minority of all Tweets[9], and expose the same geographical data format as native Tweets, they are not treated differently in this thesis.

[8] In the study of Z. Cheng et al. (2011), for instance, more than 50% of all geotagged Tweets coming from third-party apps are Foursquare check-ins.

[9] Mahmud et al. (2012) found the proportion to be 6.6%.

4. Methods

As can be seen from Figure 4.1, the methodological procedure is mainly divided into three different subprocesses: data collection (Section 4.2), preprocessing and data cleansing (Section 4.3), and comparison with authoritative data (Section 4.4). These subprocesses, in turn, consist of several small, successively executed steps, which are explained in detail in the following sections.

4.1. Software and Scripting

With a few exceptions, the processing steps in the remainder of this chapter were scripted with the Python[1] language. The rich abundance of third party packages for Python allowed to incorporate additional software, such as Database Management Systems (DBMSs) and the statistical environment R[2]. Table B.1 in Appendix B gives an overview of the applied software.

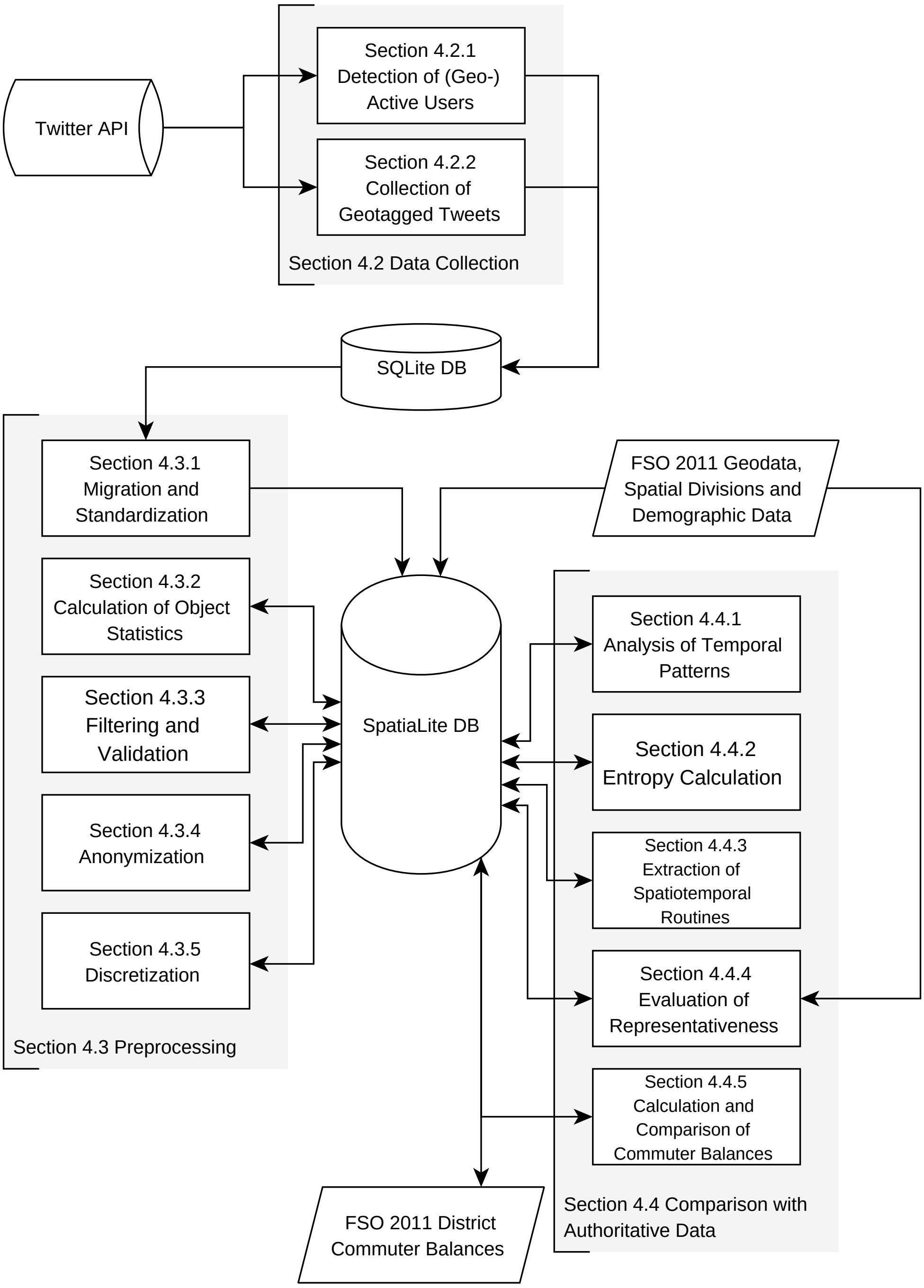
Due to Python's possibilities for modular programming, all the subprocesses involved, even the production of plots, could be chained together and data could be seamlessly transferred from one step to the next. This means that, during the whole data collection process, all the processes implemented so far could be executed fully automatically and the interim results could be used to implement the next steps or alter the already existing ones.

Figure 4.1 presents two different types of databases. The first, SQLite, was used for storing the "raw" geographical information, i.e., the geographical descriptions in JSON, as described in Section 3.3.5. Later on, the data was transformed to a "standardized" format, meaning that the punctual coordinates of geotagged Tweets were explicitly stored as spatial points in a SpatiaLite geodatabase (see Section 4.3.1). This had the advantage that spatial operations such as the *point-in-polygon* analysis used to discretize geotagged Tweets (Section 4.3.5) could be conducted directly using already implemented DBMS functionality and the intuitive Structured Query Language (SQL).

[1] http://python.org

[2] http://r-project.org

Figure 4.1: *Methodological framework. Data processing steps are denoted as rectangles, authoritative data sources as trapezoids. Read from top to bottom.*


Twitter API
Section 4.2.1 Detection of (Geo-) Active Users
Section 4.2.2 Collection of Geotagged Tweets
Section 4.2 Data Collection
SQLite DB
Section 4.3.1 Migration and Standardization
Section 4.3.2 Calculation of Object Statistics
Section 4.3.3 Filtering and Validation
Section 4.3.4 Anonymization
Section 4.3.5 Discretization
Section 4.3 Preprocessing
SpatiaLite DB
FSO 2011 Geodata, Spatial Divisions and Demographic Data
Section 4.4.1 Analysis of Temporal Patterns
Section 4.4.2 Entropy Calculation
Section 4.4.3 Extraction of Spatiotemporal Routines
Section 4.4.4 Evaluation of Representativeness
Section 4.4.5 Calculation and Comparison of Commuter Balances
Section 4.4 Comparison with Authoritative Data
FSO 2011 District Commuter Balances

4.2. Data Collection

The data collection process consisted of two concurrently running scripts, where one searched for *(geo-)active* users (Section 4.2.1) and the other repeatedly harvested new geotagged Tweets from these users (Section 4.2.2). Except for minor interruptions due to technical problems, these scripts were running continuously on a netbook with Linux Ubuntu 12.04LTS[3] installed. The collection process lasted from **January 2nd, 2013**[4] until **November 12th, 2013**, and gathered **12,069,967** geotagged Tweets from **24,733** distinct users. The data collection process will be described in detail in the following sections.

4.2.1. Detection of (Geo-)Active Users in the Study Area

Since the work of this thesis focuses on a particular study area, only Twitter users who dwell in this area are of interest. While users can specify a home location in their profile, this can be anything from precise coordinates to non-existent fantasy places, and many users ignore it altogether (Hecht et al., 2011; Z. Cheng et al., 2010). Therefore, it was neither realistically possible nor sensible to search for users based on this piece of information. Instead, a heuristic consisting of several steps was employed, of which some are explained here and others in Section 4.3.

One particular Twitter resource allows to search for geotagged Tweets by passing it a place ID, as mentioned in Section 3.3.5, as search term (Twitter Developers, 2013a). It returns the 100 most recent Tweets located within the polygon that delineates the place[5]. Every 15 minutes, a query was made to this endpoint, passing it the place ID of Switzerland, which returned up to 100 geotagged Tweets. This time window is due to rate limits and it must be assumed that not all geotagged Tweets which were written during such an interval could be captured[6]. However, it can be assumed that reasonably active users were discovered sooner or later, anyway (see further below in this section).

[3] http://www.ubuntu.com/desktop

[4] Detection of (geo-)active users started a bit earlier, namely on December 27th, 2012.

[5] It is not exhaustively clear whether the corresponding polygon or the MBR is used in this case, because using this particular endpoint with a place ID is a rudimentarily documented feature (Twitter Developers, 2012a). It is assumed that the former is the case, but the MBR denoted by 5.95 ° E / 45.82 ° N (lower left) and 10.49 ° E / 47.81 ° N (upper right), approximates Switzerland pretty well, too. Whatever is the case, Tweets outside the study area were filtered out in a later phase, anyway (see Section 4.3).

[6] Each 15 minutes, about 84 Tweets were returned on average (standard deviation: 29), and about half of all the queries returned 100 Tweets, meaning that, possibly, more were actually written.

The distinct authors of these geotagged Tweets were then looked up using another resource (Twitter Developers, 2013b), which returned the 200 most recent Tweets per user. Those were checked against a set of criteria to decide whether to start tracking the user:

1. Overall number of Tweets. If a user had posted less than 50 Tweets, including those without geographical metadata, he or she was discarded. This criterion was applied in order to preventively filter out users who were very new to Twitter or hardly ever active.

2. Age of the 5th oldest Tweet. If a user's 5th oldest Tweet, including those without geographical metadata, was older than a week, he or she was discarded. This criterion was applied as an additional measure to avoid passive or barely active users.

3. Percentage of geotagged Tweets. If a user had posted less than 25% of geotagged Tweets, he or she was discarded[7]. This criterion was applied in order to gain users who were reasonably frequently geotagging their Tweets.

As will be seen in Section 4.3, another set of filters had to be applied in order to remove the manifold noise in the set of collected users. Some of these filters could have also been applied in this stage, but because of the continuous inflow of users and Tweets, the data collection had to be kept as minimal and performant as possible.

Since (geo-)active users, by definition, repeatedly posted within the study area, their probability of eventually being discovered by the search routine was very high. In fact, as Figure 4.2 shows, the rate of collected users per day converged to a relatively stable level after just a few days. This phenomenon could have several reasons: it either means that the search routine was able to quickly identify the majority of (geo-)active Twitter users in Switzerland, and later additions to the database were either new adopters of Twitter, users who had moved to Switzerland, or users who were temporarily visiting Switzerland. On the other hand, it could be that in the beginning, most of the prolific users (see Section 4.3.2) were collected at once. Later on, the routine successively identified users who rarely geotag their Tweets, because their chance of appearing in search results is lower. Whatever is the case, the intricacies of this process did likely not influence the later analysis.

[7] In the first few weeks, the threshold was set to 50%, but later relaxed (also see Figure 4.2).

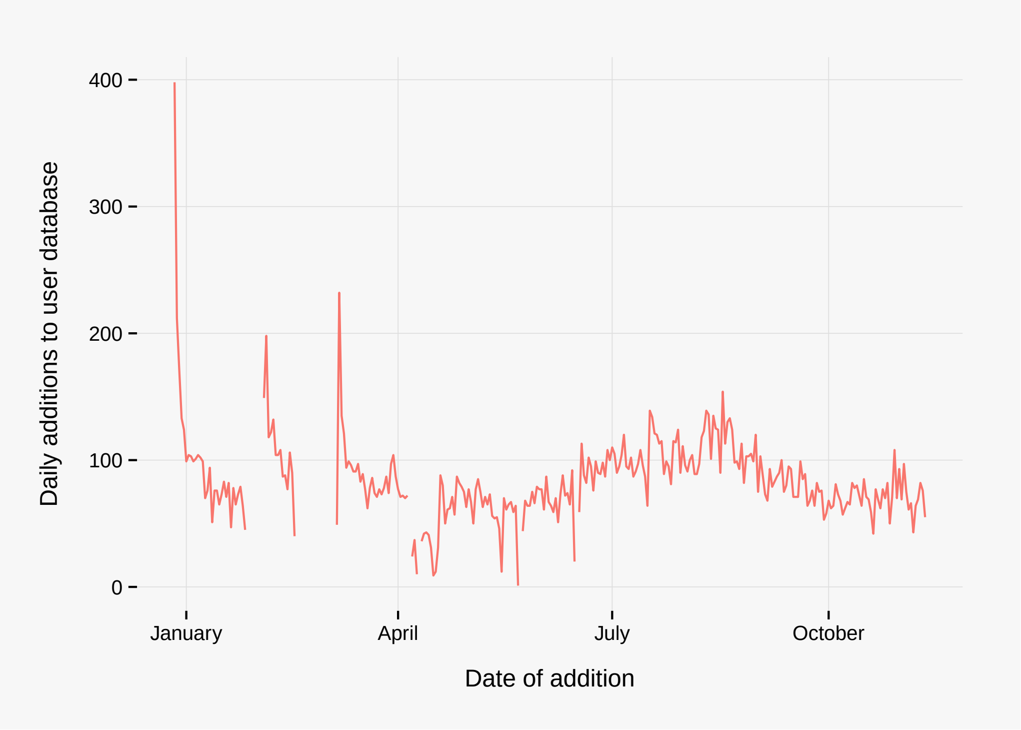


Figure 4.2: *Daily additions to the user database. Gaps are due to technical problems. The second spike occurred after relaxing one of the tracking criteria, while the third was probably due to the precedent interruption of roughly two weeks.*

4.2.2. Collection of Geotagged Tweets

Concurrently to the process described above, users in the database were queried for geotagged Tweets. Each 15 minutes, the 180[8] users who had been checked least recently (or never before) were fetched from the database and subsequently, their 200 most recent Tweets[9] since the time of the last check, including those without geographical metadata, were traversed using another API endpoint (Twitter Developers, 2013b). In order for a Tweet to be deemed as geotagged and thus be stored, at least one of the fields "geo" and "coordinates" had to be non-empty (see Section 3.3.5). It was only in the first few weeks when also Tweets were collected which *only* had contents in the "place" field, namely for computing statistics on the nature of geotagged Tweets (Section 4.3.1). The JSON content of these fields was then stored directly as text in the SQLite database. Even though the mentioned endpoint is only capable of returning 200 Tweets per request, this turned out to be sufficient to recover most if not all Tweets, since users were checked as frequently as every two or three days, on average. In the rare cases of very prolific users (Section 4.3.2), it is thinkable that some Tweets could not be captured; however, they can certainly be neglected, as the spatiotemporal routines computed from these Tweets usually become stable after a certain support of data has been reached (see Section 4.4.3).

[8] Again, this number is due to rate limits.

[9] Retweets were excluded because they contain the location of the user of the original Tweet.

It has to be noted that, at this stage, no geographical filter was applied, therefore Tweets outside of Switzerland were stored in the database, too. This was particularly the case for people who only posted a few geotagged Tweets within Switzerland when they were detected, and then left the country, while still being tracked. Additionally, during the course of the collection process, some users either deleted their account or restricted public access to it — these were not queried anymore, but remained in the database.

One might ask why it was not tried to retrospectively fetch all Tweets dating back to a certain date, e.g., December 2012, so that each sample would cover the same timespan. By using the above mentioned endpoint (Twitter Developers, 2013b), it would theoretically be possible to fetch up to the 3,200 most recent Tweets for a given user; however, this is complicated and resource-intensive, as only 200 Tweets can be fetched at a time (Twitter Developers, 2014). Another problem that could have occurred with such an approach are the large volumes of Tweets posted by prolific users. For instance, if a user had been discovered in July 2013, his or her 3,200 most recent Tweets would possibly have dated back only to February 2013, and not to December 2012. Apart from these technical problems, the methodological process simply did not require the samples to be of the same length or covering the same period — rather, the preprocessing stage ensured that the samples were reasonably dense and Tweets were coming from at least 30 different days (Section 4.3.3).

4.3. Preprocessing

The goal of the preprocessing phase was to extract the users who are ultimately desirable for the analysis: users who live in Switzerland and whose geotagged Tweets appropriately reflect their true spatiotemporal routines (Section 2.2). In fact, the properties of the collected data were highly irregular and a lot of undesired side effects, resulting from the applied collection methodology, prevailed. Firstly, the collection routine did not check for the actual residence of the users it decided to start tracking. Secondly, once a user was being tracked, the routine also collected events occurring outside of the study area. Thirdly, no measure of automatically telling whether a user appeared to be a so-called *non-humanoid actor*[10] had been implemented. Lastly, the data was in a raw format which required appropriate standardization and pruning before being usable in a spatial

[10] A non-humanoid actor, or bot, is defined here as a service or institution which automatically sends out geotagged Tweets based on occurrences of real-life events or other criteria, for example, a service which notifies followers about free parking spots in a city.

analysis (Section 4.3.1).

In order to detect and discard undesirable users, basic statistics about each user were computed (Section 4.3.2). These were heuristically used as decision criteria to preliminarily discard non-humanoid actors, *stationary* users[11], and users who are likely to reside outside of Switzerland (Section 4.3.3). To validate this removal procedure, a subset of the remaining users were randomly sampled and their profiles manually inspected. After that, profile information was no longer needed and users could be anonymized (Section 4.3.4).

To summarize, the primary objective of the preprocessing phase was the removal of undesirable users. At the same time, as much data as possible had to be retained in order for the analysis to produce reasonably significant findings. The challenge of preprocessing was thus to strike a balance between assuring data quality and retaining data quantity.

4.3.1. Migration and Standardization

After the data collection process had been finished, the SQLite database file was *migrated* to a SpatiaLite database. Before migration, the SpatiaLite database was manually created and populated with two tables containing the FSO data. The first table consisted of multi-polygon geometries, the district membership, the categorization according to any relevant spatial division system, as well as demographic data of 2,515 Swiss municipalities ("FSO 2011 Geodata, Spatial Divisions and Demographic Data" in Figure 4.1). The second contained estimated numbers of staying, outgoing and incoming commuters for each of the 147 districts, including confidence intervals and the counts of actual observations ("FSO 2011 District Commuter Balances").

The migration procedure basically converted the "raw" geographical information contained in geotagged Tweets into point geometries with WGS 84 coordinates, and also made sure that the dates and times of posting were correctly stored[12]. Apart from that, the *type* of each Tweet was inferred based on how geographical information was stored in the JSON content (see Section 3.3.5). Table 4.1 shows how the three, mutually exclusive types were defined and to what percentage they amounted[13].

Since Tweets of type C do not contain geographical coordinates and geocod-

[11] Users always tweeting from the exact same place.

[12] All dates were originally collected in Coordinated Universal Time (UTC) format but now converted to the local timezone (UTC+1 or UTC+2, depending on Swiss daylight savings time).

[13] Based on a preliminary analysis conducted after roughly four weeks of collection, involving 643,323 Tweets.

Table 4.1: *Different types of geotagged Tweets and their relative occurrence (N=643,323).*

Type	Description	Percentage
A	Tweets which contain data in all location-specific fields "geo", "coordinates", and "place". These are most likely Tweets where the user specified his or her precise location which in turn was reverse geocoded to a particular place name. The vast majority (75.1% of all Tweets) had the "city" place type.	79.7%
B	Tweets for which only the fields "geo" and "coordinates" are specified, i.e., which do not contain a particular place name. These are most likely Tweets where the user specified his or her precise location, which could not be reverse geocoded.	4.7%
C	Tweets for which only the field "place" is specified, i.e., which do not contain a precise geographical location. These are most likely Tweets where the user chose a predefined place name from a list, without exposing his or her precise location.	15.6%

ing them would have been a tedious and error-prone task, they were discarded (except for the sake of calculating the just presented statistical figures). Moreover, they only amounted to a minor percentage of all geotagged Tweets. It was verified that Tweets of type A and B contained precise — but not necessarily accurate — geographical coordinates, no matter the type of spatial granularity, and were thus retained for the following analysis.

In order to drastically reduce the computation cost of the following steps, users were ordered according to the number of Tweets they had posted, and the 20 most prolific users were removed from the dataset. This led to a significant speed up of computation time, because their share of data to be processed was overproportionally high (also see Section 4.3.2). Lastly, users without any Tweets of type A or B were discarded. After migration, the database eventually contained **10,599,669** Tweets (corresponding to a retention of **87.5**% of the "raw" data) of **24,349** distinct users (**98.5**% retention).

In the process of migration, users were renamed to *objects* and Tweets to *events*, a terminology which will be used from now on.

	Indicator	Explanatory remarks
n_o	Number of events	
d_o	Distance travelled [km]	The sum of distances between each subsequent event.
sd_o	Standard distance [km]	The standard deviation from the mean location (the "center"). It is a measure for spatial dispersion of a point cloud. A high standard distance corresponds to a large spatial dispersion (e.g. travel to many different, distant places), whereas a low standard distance corresponds to a small spatial dispersion.
tp_o	Temporal period [min]	The mean of the timespans between subsequent events. A low period corresponds to more frequent events.
s_o	Average speed [km/h]	
pe_o	Percentage of events inside study area	
pa_o	Percentage of *active periods*	See further below.
sa_o	Spread of *active periods*	See further below.

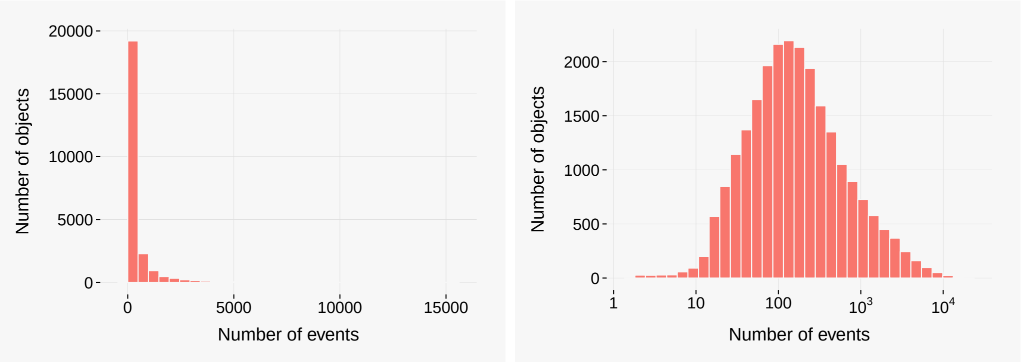
Table 4.2: *Object statistics. For each object, a range of statistical indicators were computed.*

4.3.2. Calculation and Sighting of Object Statistics

For each object, a range of different statistics was calculated based on the spatial and temporal properties of its corresponding events (Table 4.2). These *indicators* allowed first, exploratory insights about the spatiotemporal behavior of objects and served as decision criteria for discarding or keeping them. Because some of the indicators had to be log-transformed to be properly displayed in the following figures, only objects where n_o, d_o, sd_o, s_o, and $tp_o > 0$ are displayed below (N=24,019, N=10,556,128 events).

As can be inferred from Figure 4.3, the distribution of number of events per object is heavily long-tailed. Figure 4.4 illuminates this from another perspective. It can be seen that a small minority of very prolific users contributes a large amount of the (geotagged) data on Twitter. The distribution conforms to *Pareto's Principle* very well, since about 80% of the data are produced by 20% of all users (Hardy, 2010). This has been described by J. Nielsen (2006) as the "participation inequality" phenomenon and has also been observed in Flickr data (Purves et al., 2011) and in other Twitter usage contexts (e.g., Blau and Neuthal, 2012; Ross et al., 2011).

Interestingly, the same long-tailed distribution can also be found for distance traveled, average speed, standard distance, and temporal period. One could therefore assume that this is because these variables are correlated with the number of events per object, but that is not the case (Figure 4.5). The relatively high correlation with temporal period (Figure 4.5d) can be explained by



(a)

(b)

Figure 4.3: *Number of events per object before filtering (a) and its logarithm to the base of 10 (b) (N=24,019 objects).*

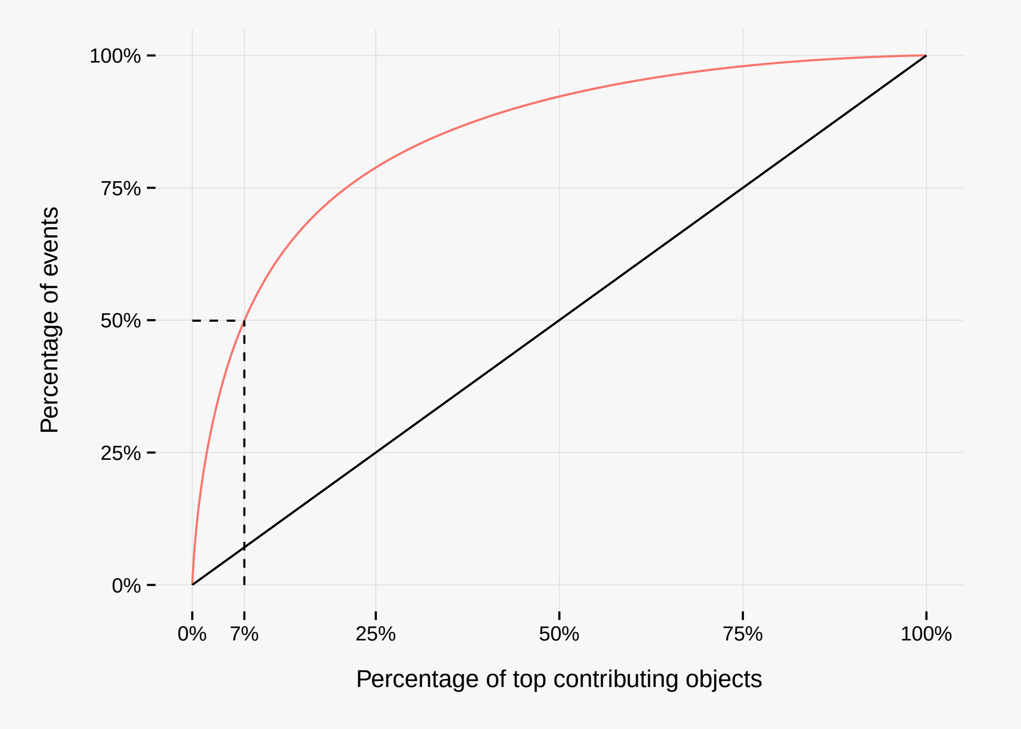
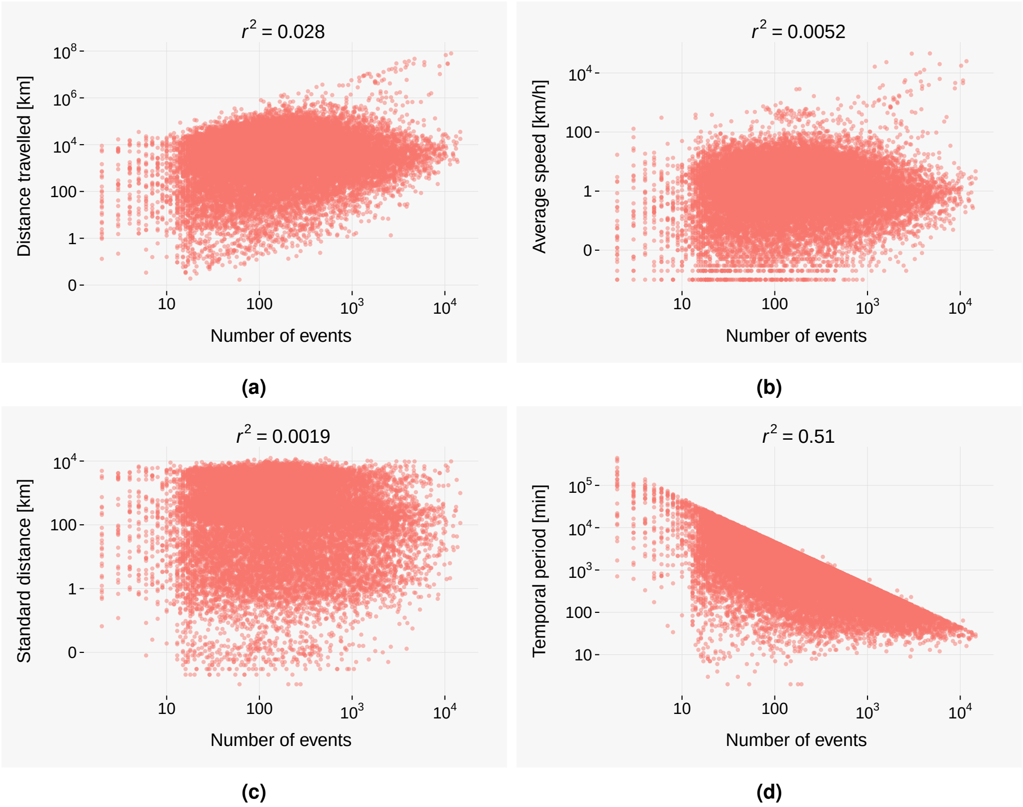


Figure 4.4: *Percentage of events created by objects before filtering. The red curve shows the percentage of all events created by the* n*th percentile of the top contributing objects. For example, 50% of all events were created by just 7% of all objects. The red curve would coincide with the black diagonal line if every object contributed the same number of events (N=24,019 objects, N=10,556,128 events).*



(a) **(b)** **(c)** **(d)**

Figure 4.5: *Relationship between the number of events and distance traveled (a), average speed (b), standard distance (c) and temporal period (d) per object before filtering (N=24,019 objects). Axes show the logarithm to the base of 10.* r^2 *shows the strength of a linear correlation between two variables under the assumption that they are approximately normally distributed.*

the physical boundary that is caused by the limited observation duration.

One can conclude that the *perceived* travel behavior of an object, expressed in terms of travel distance, speed, and spatial dispersion, as well as the posting behavior, expressed in terms of posting frequency (temporal period), is not correlated with the number of events. In other words, it does not depend on the amount of information collected or known about this object, which is intuitive and desirable.

As previously mentioned, objects which do not reside in the study area were possibly tracked, too. A simple measure to detect these is to compute the relative share of events in Switzerland pe_o. As Figure 4.6 illuminates, there seems to be a divide between objects: most of them either posted nearly all of their events inside Switzerland; or they posted only a few or none at all. Clearly, the latter is the large majority, which shows that the data collection process tracked a lot of foreign users ($pe_o = 0$) and gathered a lot of temporary visitors or, possibly, non-humanoid actors (Section 4.2.1). It has to be assumed that the prevalence of many objects with $pe_o = 0$ are due to the fact that the initial spatial query for geotagged Tweets within Switzerland indeed uses a MBR, and not the exact boundaries of the country (Section 4.2.1). This may have resulted in the discovery of many objects who dwell in the close vicinity of Switzerland,

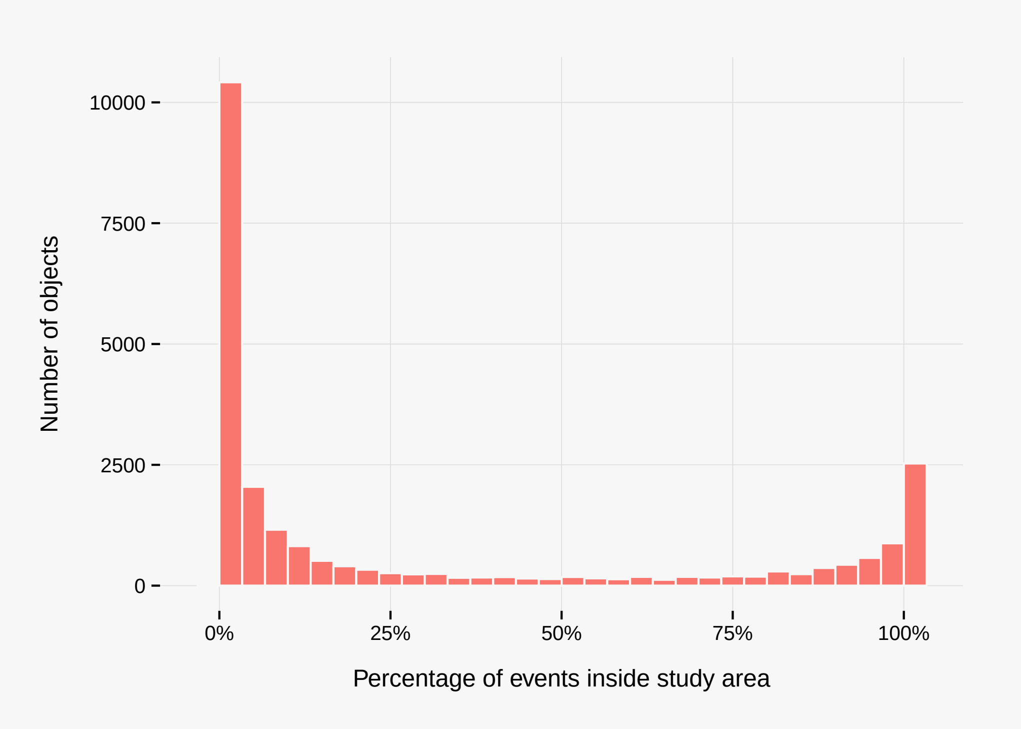


Figure 4.6: *Percentage of events inside study area before filtering (N=24,019 objects).*

e.g., in Southern Germany or in Eastern France.

This prevalence of a significant number of objects which are not desirable must be dealt with during preprocessing. However, simply setting a threshold on the percentage of events inside the study area is not sufficient. Namely, it is not an actual indicator of time spent in Switzerland, because single events are just points in time and one does not know what happens in between them. Given an object has a value of 90% percent for this measure, it does not mean that the object spent 90% of its time in Switzerland. Contrarily, it could be that the corresponding user preferred to broadcast geotagged Tweets from within Switzerland, but refused or was not able to do so abroad, even though he or she spent considerable time outside of Switzerland. It could also be that people are *only* geotagging their Tweets when they are visiting a certain part of the world, e.g., for reasons of self representation (see Section 2.1.3).

Residing in a certain country implies continuity. For instance, it is unlikely that somebody switches his or her residential country on a daily basis. This principle was made use of to compute another indicator variable, the percentage of so-called *active periods* pa_o. An active period is defined as a time frame in which the object created sufficiently enough or *dense* events within the study area, i.e., a timespan without longer interruptions. Per definition, an object is considered to be active as long as there are not k consecutive inactive *days*[14]. The

[14] Actually, in order for a day to be deemed as active, *all* the events posted during this day need to

day is taken as temporal aggregation level because it is neither too coarse nor too fine. The *percentage* of active periods is simply the overall length of active periods in relation to the timespan from the object's first event until the end of data collection.

The computation of active periods can be regarded as a density computation, i.e., it is calculated how densely active days occur. Therefore, the well-known Kernel Density Estimation (KDE) method (Wand & Jones, 1995) is suited for this kind of problem. Colloquially speaking, the method works by applying a kernel distribution with a certain bandwidth b to each point in a dataset (active days in this case) and summing up the values of the distribution. This results in a continuous, dimensionless density value defined on the whole input domain. For the purpose of the task at hand, sections, where the density reaches 0, are considered to be interruptions of active periods, and a uniform kernel function is used. Instead of k, b must be specified, but k can be deducted from b. Figure 4.7 gives an example of an object where $pe_o = 52\%$, whereas the percentage of active periods pa_o, computed with the density calculation, is just 13% (Figure 4.7a) and 10% (Figure 4.7b), depending on the chosen bandwith. The choice of b obviously affects the outcome of pa_o and should thus be made carefully. To illustrate this, a sensitivity analysis was conducted (Figure 4.8)[15]. pa_o was computed for different values of b, with a reference of $b = 14$. The resulting set of objects satisfying the threshold $pa_o \geq 25\%$ was then compared to the reference set, both in terms of set size and set overlap. The latter is defined by the *Jaccard index*

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{4.1}$$

where A and B are two different sets. An index of 1 signifies total similarity, whereas one of 0.3, for instance, signifies 30% overlap. From Figure 4.8, it can be seen that the choice of $b = 14$ makes sense since the deviations in set length and Jaccard index when using slightly different values of b are marginal. Contrarily, setting b too low results in a significant reduction of objects passing the pa_o threshold. Interestingly, at $b = 15.75$, the number of objects which still fulfill the criteria drops slightly, but continues to rise afterwards.

While pa_o reflects the amount of time for which it can be assumed that

be within the study area. This is because certain bots create so many events that, even though most events are outside of Switzerland, there would still be an event within Switzerland per day — without applying this criterion, these bots would qualify as normal users with very long active periods. The measure is thus also a technique to detect and remove bots.

[15] Based on a dataset of roughly 8 months of collection, involving 6,226,295 events from 17,224 distinct objects.

Figure 4.7: *Active periods of an object with different bandwidths* b = 14 *(a) and* b = 7 *(b). The light gray curve is the density surface calculated using a uniform kernel. The red lines mark active periods.*

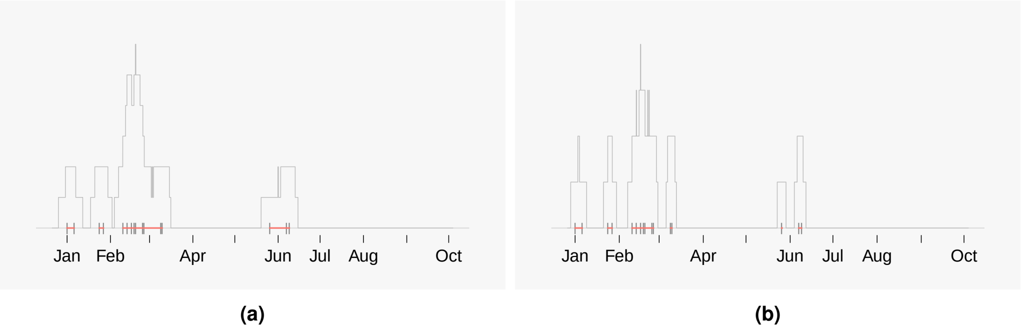
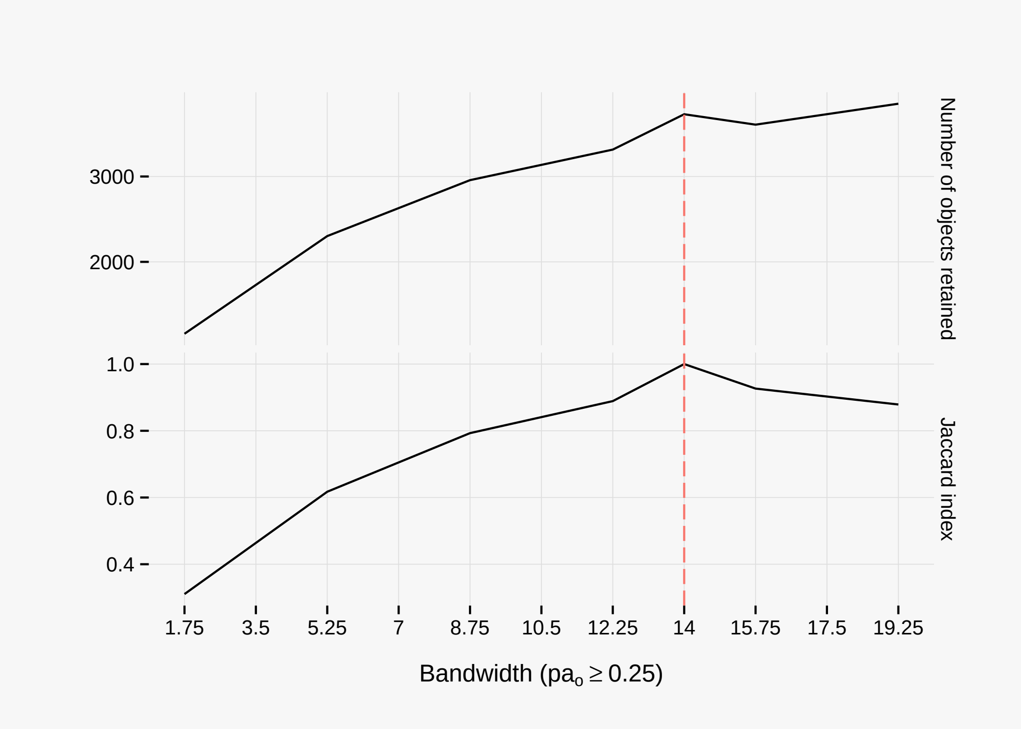


Figure 4.8: *Sensitivity analysis for the choice of the bandwidth. The upper chart shows the number of objects which would be retained if a certain* b *would be chosen, given $pa_o \geq 25\%$. The lower chart shows the relative set overlap between the resulting and the reference set of* b *= 14 (red dashed line) (N = 17,224 objects).*



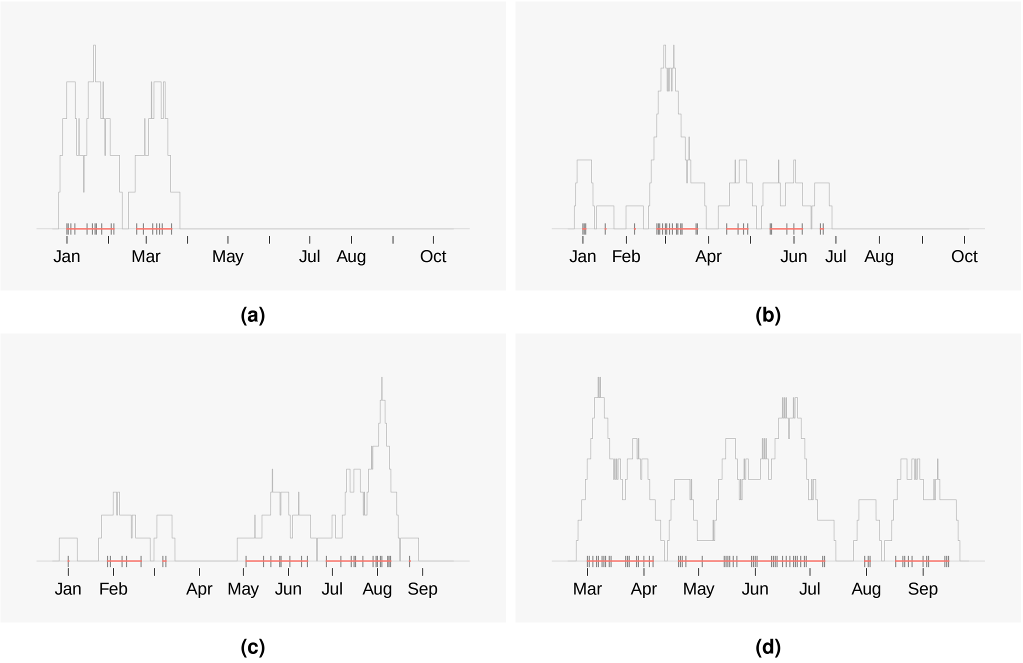


Figure 4.9: *Spread of active periods for different objects.* $sa_o = 8.8$, $pa_o = 0.26$ *(a);* $sa_o = 15.6$, $pa_o = 0.26$ *(b);* $sa_o = 31.5$, $pa_o = 0.43$ *(c);* $sa_o = 39.8$, $pa_o = 0.64$ *(d).*

object spent in Switzerland pretty well, it cannot tell *when* active periods happened and how they were spread over the time of data collection. For instance, the object in Figure 4.9a has an overall pa_o of 26%, spanning two full months. Thus, this object is likely to qualify for further analysis. Still, it might be that it left Switzerland after March or even stopped using Twitter altogether. For this reason, an additional statistic, spread of active periods sa_o, was computed. It is defined as

$$sa_o = \sigma_{ado} * pa_o \tag{4.2}$$

where σ_{ado} is the standard deviation of active days — in other words, of the time that has passed from the first active day to every other. This value is multiplied with pa_o in order to also value objects which have a generally frequent posting behavior (compare Figure 4.9d with Figure 4.9c, whose sa_o would be similar if not corrected for pa_o). Figure 4.9 shows that, even though the resulting value cannot be as easily interpreted as pa_o, it is a good measure of temporal spread of active periods. For instance, Figure 4.9a has the same pa_o as 4.9b, but a considerably lower sa_o.

In conclusion, this section showed that the properties of the collected objects and their events vary heavily. Extreme values in the data, such as very long travel distance, high standard distance, and unrealistically high average speed, indicate the prevalence of non-humanoid actors, errors, or purposefully faked data. On the other hand, low values of percentage of active periods and of standard distance might hint to non-active and stationary users, respectively. While no indicator can exclusively deal with a certain type of unwanted objects, together they are likely able to clean the data from the majority of those. This will be the subject of the next section.

4.3.3. Filtering and Validation

Before the applied filtering criteria are examined in detail, the properties of objects worth to retain are summarized once more. Where possible, corresponding indicator variables are highlighted.

1. A desirable object should have a reasonable number of events, so that spatiotemporal routines can be discovered.
2. It should not be stationary, meaning that it should not send Tweets from always exactly the same position, which also hints at a non-humanoid or immobile object. Such objects can be excluded by setting a minimal standard distance or by setting a minimal distance traveled.
3. It should permanently reside in Switzerland or spend at least a long enough time in Switzerland. A very high standard distance, for example, could mean that the object travels frequently abroad, effectively resides at another place, and visited Switzerland only once. Another indicator that can be used is percentage of events in Switzerland, although its flaws have been shown above. Thus, a minimal required percentage of active periods and a minimal spread of active periods might be most suited. As could be seen in Section 3.2, only permanent residents are included in the commuter statistics, and this requires at least twelve months of residence. This criterion is relaxed here because of the temporally constrained data collection process.
4. There should exist a continuous, ongoing, and temporally spread sample of the object's whereabouts. For this, the spread of active periods and percentage of active periods indicators can be used.
5. It should be humanoid. For example, unusually high average speeds and standard distances hint to frequently changing locations, which could, for instance, be produced by a service which notifies people of events happening at these locations. At the same time, stationary objects, as defined above, are often non-humanoid, too.

To filter for objects who fulfill these criteria, a set of logically conjunct indicator thresholds were applied. In order to study the effects of possible combi-

	$\mathbf{n_o} \geq$	$\mathbf{sd_o} \geq$	$\mathbf{sd_o} <$	$\mathbf{s_o} <$	$\mathbf{pe_o} \geq$	$\mathbf{da_o} \geq$	$\mathbf{sa_o} \geq$	**Objects**	**Remarks**
	0	0	∞	∞	0	0	0	100%	Reference set
A	100	0.1	10000	10	0.5	60	30	3.8%	Most desirable
B	...	...	...	...	...	30	20	6.1%	Lower minimal da_o, lower minimal sa_o
C	50	...	...	...	...	...	...	7.9%	Lower minimal n_o
D	100	...	...	...	...	...	10	9.3%	Higher minimal n_o, lower minimal sa_o
E	75	...	...	...	...	...	...	10.6%	Lower minimal n_o
F	50	...	...	...	...	...	...	12.5%	Lower minimal n_o
G	...	...	...	...	0.25	...	...	13.1%	Lower minimal pe_o

Table 4.3: *Effects of filtering criteria on the number of retained objects. Read from top to bottom, ... signifies repetition of the previous value (N=17,224 objects).*

nations of threshold values on the number of retained objects, a small sensitivity analysis[16], which is presented in Table 4.3, was conducted. For the purpose of better interpretability, not pa_o, but its multiplication with days since first event (dsa_o), $da_o = pa_o * dsa_o$, is displayed. This makes it possible to specify a threshold which is independent of the actual duration of the data collection process.

One particularly striking result of this analysis is that only very few objects fulfill the desirable criteria. Apparently, the most desirable scenarios A and B cannot be used because too many objects would be rejected. When successively adjusting different indicator values, the percentage of retained objects slowly rises, but stays generally low. The strong influence of n_o on the percentage of retained objects is interesting, too, as can be seen in the change from scenario B to C and from scenario D to F. One would expect that the number of events an object posted is already incorporated into pa_o and sa_o, but the data shows that there must be quite a lot of objects with sparse but well-spread events. To achieve a certain compromise between a loss of too many objects and a reasonable quality of the data, scenario E was chosen. This eventually resulted in the retention of **2,380** objects (corresponding to a retention of **9.8%** of the previously retained objects) having **2,052,913** events (**19.4%** retention).

After that, a subset of $N = 298$ objects[17] was randomly sampled from the resulting set of scenario E. The Twitter profiles of those were then manually looked at, and the following criteria were checked:

- Is the account humanoid? If not, what kind of non-humanoid actor is it?
- Does the account holder reside in Switzerland? This can be guessed based

[16] Based on a dataset of roughly 8 months of collection, involving 6,226,295 events from 17,224 distinct objects.

[17] Originally, 300 were sampled, but 2 had deleted their account in the meantime.

Table 4.4: *Prevalence of non-humanoid actors in the filtered sample (N=298).*

Humanoid	Non-humanoid, company account	Non-humanoid, non-stationary
99.0% ± 1.1%	0.3% ± 0.7%	0.7% ± 0.9%

Table 4.5: *Residence distribution in the filtered sample (N=298).*

Inside study area	Outside study area	Unknown
97.0% ± 1.9%	1.0% ± 1.1%	2.0% ± 1.6%

on information found in the profile, particularly self-declared location information and language as well as topics used in Tweets.

- What is the approximate age of the account holder? This can be guessed based on information found in the profile, e.g., description, profile picture, and uploaded images. Although such an estimation is subject to high uncertainty, an overall picture can still be established. In order to compare the obtained results with previous findings as presented in Section 3.3.1, approximately the same age ranges were used.
- What is the gender of the account holder? The same procedure and concerns as for age estimation apply.

Tables 4.4,4.5,4.7 and 4.6 show the results for the respective estimates. Intervals are based on 95% confidence and assumption of normal distribution for the errors. Clearly, the goal of having only humanoid actors residing in Switzerland is almost fully met and the few exceptions are negligible. There was one profile in the sample which acts as an account for a company and two profiles which automatically broadcast certain events. One regularly publishes temperatures for various positions of the Bodensee lake and the other notifies people about nightlife events happening in the city of Zürich. These were most likely not discarded because they both operate in a rather small region. There were also two humanoid profiles of people who clearly reside in another country than Switzerland, and several where this could not be determined with high certainty.

The gender distribution shows an equal share of males and females when taking into account the sample error (Table 4.6). Regarding age distribution, one can say that teenagers amount to a very large share of overall users, while people in their twenties and thirties are common too (Table 4.7). Apparently, older people only amount to a very marginal share. Overall, the figures are in good accordance with demographic statistics presented in Section 3.3.1.

In terms of qualitative findings, an interesting difference regarding age and social background of users across the two major language regions was ob-

Female	**Male**	**Unknown**
50.3% ± 5.7%	45.0% ± 5.6%	4.7% ± 2.4%

Table 4.6: *Gender distribution in the filtered sample (N=298).*

< **20**	**20–40**	> **40**	**Unknown**
53.4% ± 5.7%	36.6% ± 5.5%	3.0% ± 1.9%	7.0% ± 2.9%

Table 4.7: *Age distribution in the filtered sample (N=298).*

served. In the French-speaking part, Twitter seems to be considerably more popular among young users (aged below 20), and the service is often used like a chat or messaging application rather than a pure *one-to-many* broadcasting tool, which often results in a high number of events per object. It also appears that quite a large part of the examined profiles of French-speaking people belong to young immigrants in the Geneva (Genf) region. On the other hand, in the German-speaking part, Twitter seems to be most popular among professionals in the field of media, public relations, design and computer science — a mostly well-educated, high-income and male cohort. Even though these observations cannot be backed by statistical evidence, they should be kept in mind when interpreting the results of this thesis — they certainly hint at a very heterogeneous Twitter landscape in Switzerland.

In addition to this validation, indicator variables were compared to those before filtering. A quick visual overview shows that most distributions of and relationships between variables stayed similar. As Figure 4.10a shows, the number of events per object did not follow a log-normal distribution anymore due to the truncation at $n_o = 75$, but was still skewed. In fact, still 9% of all objects contributed 50% of all events, as compared to the 7% before the filtering. The distribution of pe_o shows that about 75% of all remaining objects had 85% or more of their events inside the study area, which is also desirable (Figure 4.10b).

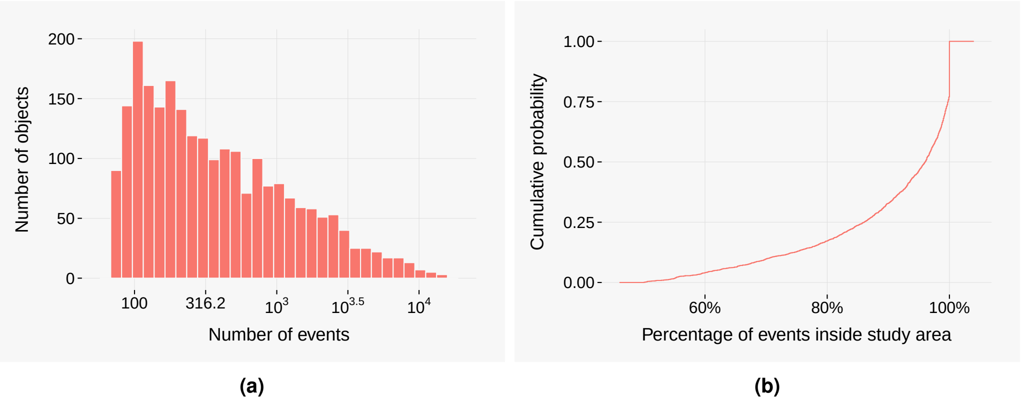


Figure 4.10: *Number of events (a) and percentage of events inside study area (b) per object after filtering (N=2,380 objects).*

To conclude, even though only about 10% of users could be retained, the validation justified these measures. It could now be assumed that the data was in a state of quality sufficient for further analysis.

4.3.4. Anonymization

In order to preserve the privacy of the collected and tracked users and to prevent possible inference attacks on their exact whereabouts, the only two links to their Twitter profiles — unique Twitter user ID and screen name — were fully removed. Thus, from this moment on, all objects in the SpatiaLite database were just (pseudonymized) sequences of locations without any connection to the actual Twitter profiles.

It has to be noted that previous versions of the SpatiaLite database or the preceding SQLite database still contained the mentioned links, and could have been used to reestablish a connection to Twitter profiles. The author hereby declares that this was not the case and reasonable measures (i.e., encryption) to protect the data were taken.

4.3.5. Discretization

In order to prepare the data for further analysis, the remaining events had to be enhanced with additional information. First, all remaining events *outside* of the study area were removed. Then, the table holding the municipalities data (see Section 4.3.1) was geometrically intersected with each event, resulting in a direct mapping of that event to a municipality and thus to its spatial division and socio-demographic properties (Section 3.2.2).

In addition to these spatial semantics, the time and the day of each Tweet was discretized into hourly intervals from 0 to 23 and weekdays from 0 to 6, respectively, with 0 signifying 00:00–01:00 and Sunday, respectively. An event e_i belonging to object o is thus defined as a tuple

$$e_i = \langle o_i, m_i, h_i, d_i \rangle \tag{4.3}$$

where $o_i \in O$ is the object identifier, m_i one of a limited set of municipalities M, $h_i \in H$ the hour of the day and $d_i \in D$ the day of the week.

At the end of the discretization process, the database still consisted of **2,380** objects having **1,913,512** events (**93.2%** retention). Due to the removal of events outside of Switzerland and the threshold of $pe_o \geq 50\%$, some objects lost up to 50% of their events.

4.4. Comparison with Authoritative Data

4.4.1. Analysis of Temporal Patterns

One would assume that the volume of events follows a certain daily or weekly pattern, and such patterns may help to find universalities that may be used in the extraction of spatiotemporal routines (Section 4.4.3).

Summarizing events was straightforward — for each day which was ever collected (and week, respectively), all the hourly (and daily, respectively) events were counted, e.g., the number of events during the time slot of 05:00–06:00 for each day. These amounts were then averaged over all days (and weeks, respectively). Each time slot can thus be looked at as a sample that grew bigger with every day data was being collected. One has to keep in mind, though, that the number of events was considerably smaller at the beginning of the collection process because fewer objects were in the database.

However, the mere number of events per time slot is not representative of the actual number of people broadcasting their whereabouts because of participation inequality (Section 4.3.2). Therefore, in order to know how many *different* people were "active" during a certain slot, the number of *distinct* objects per slot was also summarized. This assures that an object is only counted once per hour or per day, no matter how many events it posted during that slot. This approach has disadvantages in itself, for instance, regional figures as presented below do not precisely sum up to the figures for the whole study area, but it is closer to the notion of "present" or "active" population.

4.4.2. Entropy Calculation

After discretization, the data was ready to be used for the calculation of $Entropy(E_o, R)$ values for each object o as introduced in Section 2.2.3. Particularly, entropies for $R = \{M\}$, $R = \{M, H\}$, $R = \{M, D\}$ and $R = \{H, D\}$ were computed (see Section 4.3.5 for an explanation of the symbols). Since there might be many more distinct observations in $E_o(R = \{M, H\})$ than in $E_o(R = \{M, D\})$, for example, the respective entropies are hardly comparable. Thus, the *normalized* entropy

$$Entropy_{norm}(E_o, R) = \frac{Entropy(E_o, R)}{Entropy_{max}(E_o, R)} = \frac{Entropy(E_o, R)}{\log |E_o(R)|} \tag{4.4}$$

was calculated. $|E_o(R)|$ signifies the number of distinct observations in $E_o(R)$ — the so-called *schedule size*[18].

[18] Note that $|E_o| = n_o$.

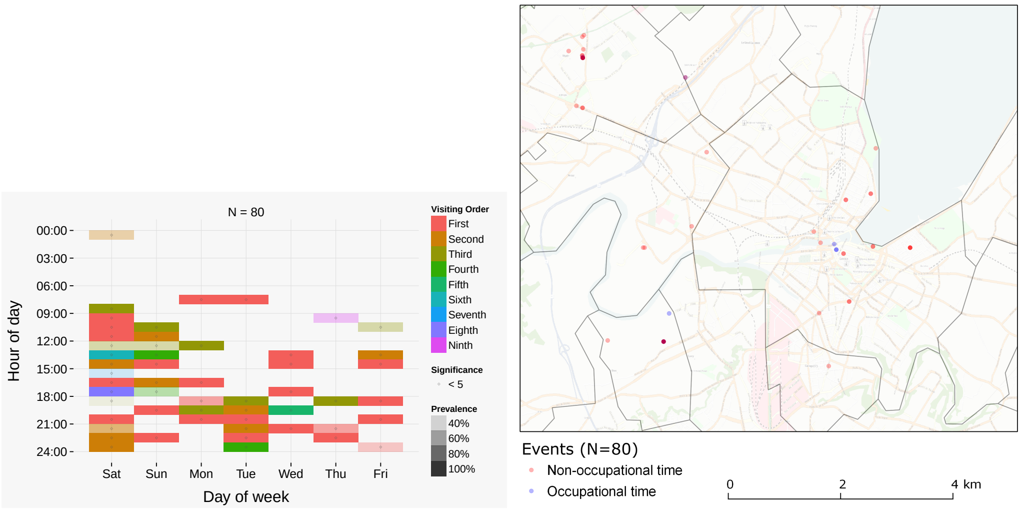
In addition to these entropy values, the percentages of the n-most visited locations were computed, for instance, the share of the two most visited locations of all visited locations. These figures are yet another indicator for regularity in visiting patterns because they tell whether an object mostly returns to the same few locations or not. Together, these statistics help — at least to a certain degree — to judge whether the data about an object is sufficient to derive a spatiotemporal routine from it, and whether the dataset as a whole is suited for the extraction of spatiotemporal routines. The findings, which are presented in Section 5.2, generally reaffirm that this is the case.

4.4.3. Extraction of Spatiotemporal Routines

In order to assess the representativeness of the collected VGI and to compare commuter balances as found in Twitter with official ones, the municipality of both residence and work or school (from now on referred to as *occupation*) had to be determined for every object. Even though authoritative commuter balances are only available on the district level, semantic places had to be extracted on the level of municipalities, because the spatial divisions as presented in Section 3.2.2 are based on municipalities and not on districts.

Using the above definition of SR_o, the goal of this processing step was to find municipalities which are regularly and reasonably frequently visited *at a particular time*, and thus afford a certain function or bear a certain meaning. Or, in the sense of Section 2.2.2, municipalities had to be labeled with the semantic annotations "home" and "occupation". Taking into account the fact that most people in Switzerland are likely to be occupied from Monday to Friday and during 08:00 and 17:00 (Section 3.1.2), it seemed legitimate to partition t, expressed through H and D, into *non-occupational* and *occupational* time, with the delimitation being the same for each object. Occupational time thus encompasses each weekday from 08:00 to 17:00, while the rest is considered to be non-occupational time. This allows to classify municipalities based on their occurrence, or prevalence, in each of those two types and incorporates SR_o's notion of temporal dependence.

The first step in finding a decent methodology for labeling municipalities was to explore common types of behavior found in the event data. Thus, a so-called *carpet plot*, which shows the spatiotemporal distribution of events at one glance, was constructed for each object. The axes show both temporal dimensions, and the spatial dimension M, i.e., the visited municipalities, is depicted as colored tiles. In order to enhance readability, always the same color



(a)

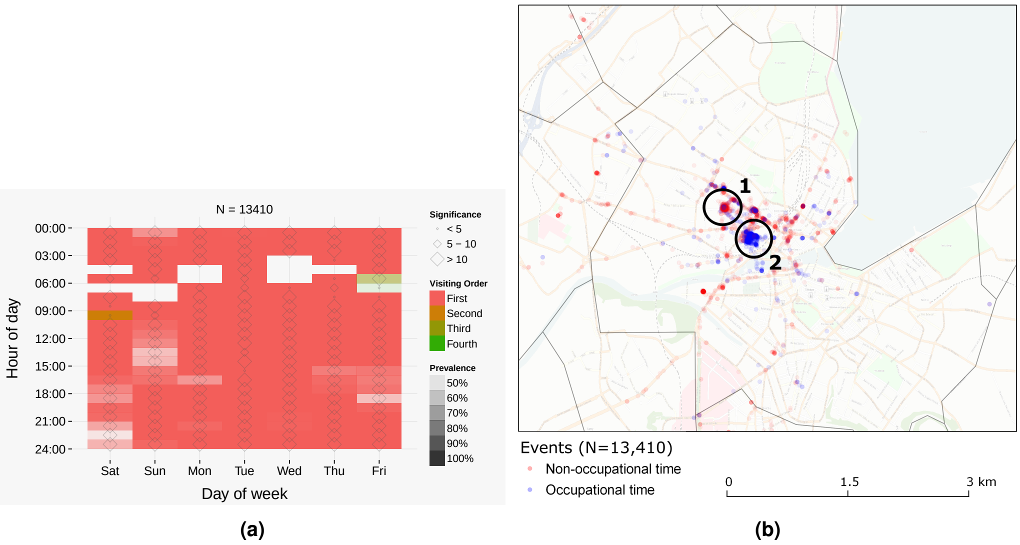
(b)

Figure 4.11: *Type A spatiotemporal pattern (a) and spatial distribution of corresponding events (b). Municipality boundaries are depicted as gray lines. Base map source: http://openstreetmap.org.*

scale is used — the overall most visited municipality is depicted in red, the second most visited in brown, et cetera. Since several *different* municipalities might have been visited during a certain slot, only the most visited is depicted, but its share among all municipalities visited during this slot is conveyed with opacity, thus allowing to assess its relative *prevalence*. On the other hand, prevalent municipalities become relatively *significant* only when they have been visited more than just a few times, which is depicted with a rhombus symbol. Together these visual clues help to assess the existence and form of patterns, and thus allow to judge whether an object's data as a whole are suited for the detection of function-bearing municipalities. In order to get an overview of the different types of patterns found in objects' data, a few dozens of objects were randomly sampled and visually examined. From this preliminary examination it appeared that objects can generally be divided into four types.

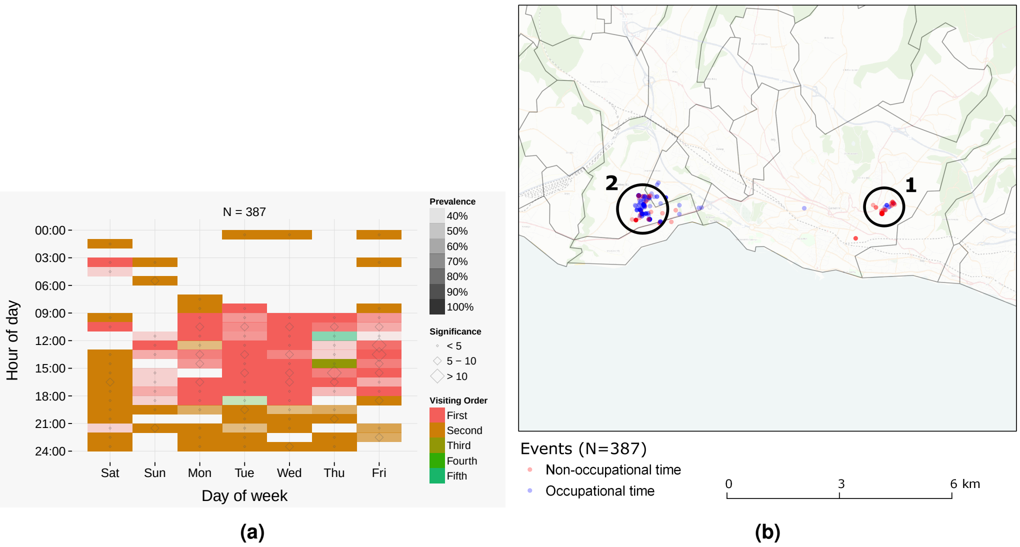
Type *A* concerns objects whose events seem not to follow a pattern at all and exhibit generally high entropy values (Figure 4.11a), which is often due to a lack of a sufficient number of events. In other words, there might actually exist a pattern but, with the data at hand, it does not become apparent. In such cases, there is usually not even a clearly predominant municipality, which could be heuristically classified as residential municipality.

Type *B* concerns objects where the most visited municipality is likely to be the place of both residence and occupation. This is reflected in the significant spread of the same municipality over both occupational and non-occupational time (Figure 4.12a). Since it is assumed that the spatiotemporal behavior on Twitter reflects someone's true routine and that all objects have an occupation, one must conclude that, in this case, the place of occupation lies in the same



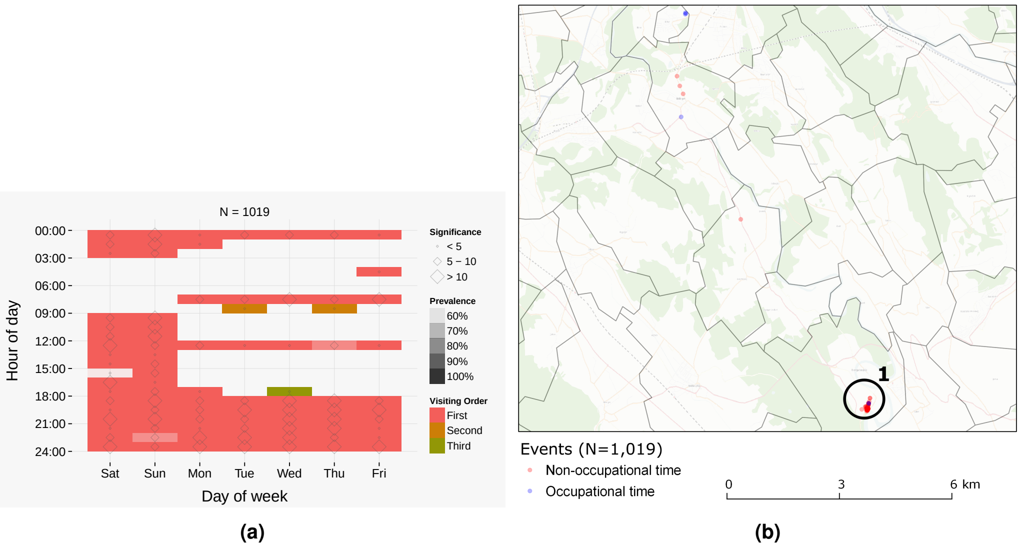
(a) (b)

Figure 4.12: *Type B spatiotemporal pattern (a) and spatial distribution of corresponding events (b). Possible places of residence are denoted with "1", places of occupation with "2". Municipality boundaries are depicted as gray lines. Base map source: http://openstreetmap.org.*



(a) (b)

Figure 4.13: *Type C spatiotemporal pattern (a) and spatial distribution of corresponding events (b). Possible places of residence are denoted with "1", places of occupation with "2". Municipality boundaries are depicted as gray lines. Base map source: http://openstreetmap.org.*



(a) (b)

Figure 4.14: *Type D spatiotemporal pattern (a) and spatial distribution of corresponding events (b). Possible places of residence are denoted with "1". Municipality boundaries are depicted as gray lines. Base map source: http://openstreetmap.org.*

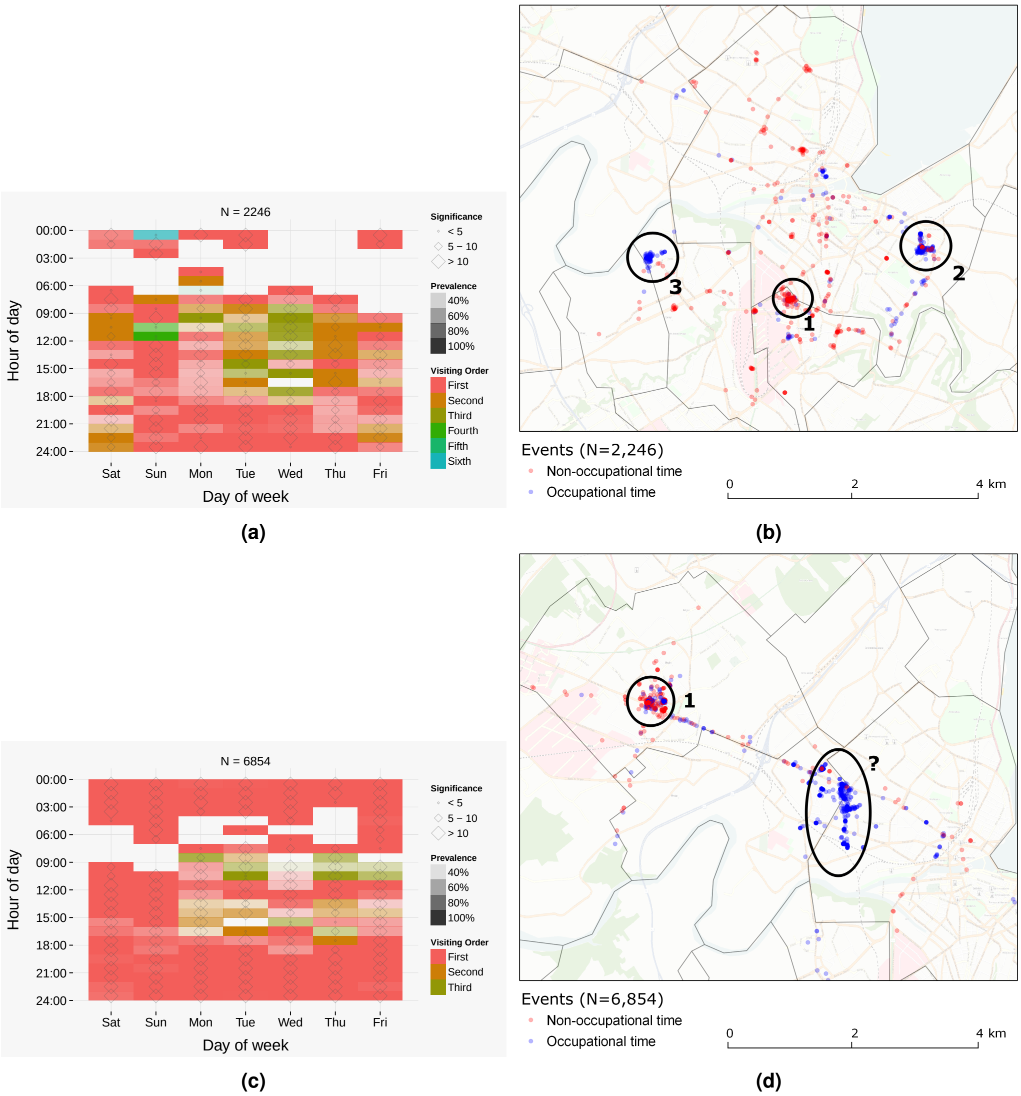
municipality.

Type *C* objects show a clear spatiotemporal pattern in terms of separability of non-occupational and occupational time, i.e., each subset of events is dominated by a *single, distinct* municipality (Figure 4.13a). This means that both the place of residence and occupation could be extracted just by looking for the most dominant municipality in each subset.

Type *D* concerns objects which are, to some degree, a combination of type *A* and *C*. The municipalities visited during non-occupational time usually exhibit a clear pattern, i.e., non-occupational time is dominated by a single municipality, which is most likely the place of residence. On the other hand, occupational time does not show an explicit predominance of a particular municipality, or there is just not enough data, as it is the case in Figure 4.14a. Sometimes, occupational time is dominated by two or more different municipalities, which might be due to the fact that a person is occupied at different places, for instance, a person who usually goes to work in the beginning and end of the week and to school on Wednesday (Figure 4.15a), or in the morning and in the afternoon, respectively (Figure 4.15c). The opposite case, where the municipality of occupation is clearly recognizable and the municipality of residence is somewhat unclear, was only rarely detected, neither in the exploratory sighting nor during the manual classification described below.

Although the goal of this step was to extract locations on the rather coarse-granular level of municipalities, it is helpful to take a look at the actual spatial distribution of events in order to further understand spatiotemporal routines. Therefore, the events of the object were plotted on top of a map of municipal boundaries and colored according to whether they were posted during occupational or non-occupational time. To make frequently visited places easily detectable, the events were plotted as transparent dots. Intuitively, function-bearing places should appear as dense clusters of chromatic homogeneity. For type *A*, events are spread over multiple municipalities and no such clusters can be visually detected (Figure 4.11b), whereas in Figure 4.12b, the same municipality is visited in a lot of different places, but one can still recognize at least two such clusters. Even though the events, in this case, are spread over many different places, one can conclude that the municipality of residence and of occupation are very likely the same. For the object belonging to type *C*, it is clear that one municipality is usually visited during occupational time and the other during non-occupational time, although some events were also posted in other municipalities (Figure 4.13b). Lastly, while one can recognize a manifestation of a residential cluster in Figure 4.14b, it is impossible to tell where the object

Figure 4.15: *Type D spatiotemporal patterns (a,c) and spatial distribution of corresponding events (b,d). Possible places of residence are denoted with “1”', “2” and “3”. Municipality boundaries are depicted as gray lines. Base map source: http://openstreetmap.org.*



(a)

(b)

(c)

(d)

Type A	Type B	Type C	Type D
15% ± 3.5%	28.8% ± 4.4%	22.8% ± 4.1%	33.5% ± 4.6%

Table 4.8: *Distribution of spatiotemporal routine types among objects (N=400). Intervals are based on 95% confidence and assumption of normal distribution of the errors.*

stays during occupational time. When looking at the special cases of type *D*, the two places of occupation of the object depicted in Figure 4.15a are recognizable (Figure 4.15b). Although one can find two temporally separated occupational municipalities for the object depicted in Figure 4.15d, the spatial distribution of events does not clearly convey this. Possible places of occupation are part of one large cluster that spans two different municipalities — judging from this, it is hard to tell whether the object is actually occupied at two different places. This is still likely, though, because, as can be seen from the carpet plot, the object seems to go to one municipality in the morning, returns home for lunch — which explains the blue dots in the residential municipality — and then goes to another municipality in the afternoon.

To conclude, the actual spatial distribution of events generally reaffirms the insights gained from the carpet plots. It becomes particularly clear that, given enough events are available, distinct clusters of residence or occupation emerge and that these could probably be used to extract function-bearing municipalities, too, e.g., by conducting a cluster analysis. In some cases, the spatial distribution also reveals uncertainties and peculiarities that are hardly detectable by just looking at the carpet plots. Nonetheless, the approach taken here is more simple and thus preferred, as it essentially reduces a spatial problem to a non-spatial pattern detection task, which will be described in the following paragraphs.

In order to quantify the relative share of each type, N=400 objects were randomly sampled and manually labeled by looking at their carpet plots. As it was the case with the validation of the filtered results in Section 4.3.3, this classification suffers from some degree of uncertainty, since types could not always be clearly separated from each other. Nonetheless, it is assumed that misclassifications average out and that it is sufficient to have a rough estimate of the true distribution, which is shown in Table 4.8.

According to these figures, over 80% of the individual routines seem to be suited for the extraction of function-bearing municipalities (type *B*–*D*). Be-

Table 4.9: *Classification confusion matrix of decision tree for identifying type-A-objects (N=400).*

classified as →	**A**	**B**	**Total**
A	41	19	60
B–D	11	329	340
Total	52	348	

fore that, the objects exhibiting type-A-behavior needed to be detected and discarded. This likely could not achieved by only filtering for a simple threshold, for instance a minimum number of events n_o. Instead, a binary classifier which takes into account not only one but several attributes of an object seemed more appropriate. The 400 manually classified objects, together with their respective entropy values and other descriptive figures, were thus used to train a C4.5 decision tree (Quinlan, 1993), an algorithm commonly employed in machine learning. Beside its simplicity in implementation and relatively good performance, it also has the advantage of outputting a comprehensible classification model. Using the Weka data mining toolkit (Hall et al., 2009), such a tree was trained and its performance was assessed by means of 10-fold cross validation[19].

The learner eventually identified both attributes $Entropy_{norm}(E_o, R = \{M\})$ and percentage of the most visited municipality as decision criteria, achieving an overall accuracy of 92.5% correctly classified instances. Namely, objects are classified as belonging to type A whenever their $Entropy_{norm}(E_o, R = \{M\})$ exceeds 0.68 *and* the overall most visited municipality amounts to less than 50.7%. While this learner performs very well in terms of *true positive rate* for objects of type B–D (objects of type B–D correctly classified as such) and *false positive rate* for objects of type A (objects of type B–D wrongly classified as type A), the false positive rate of objects of type B–D is relatively high[20] (Table 4.9). When applied to the full dataset, the decision tree classified **185** of **2,380** objects as showing type-A-behavior (**7.8**%), leaving **2,195** objects for further analysis.

With only — or mostly — objects of type B–D remaining in the dataset, an extraction algorithm which acknowledges the various types of patterns needed to be found. Instead of examining the particular properties of such types in further detail, a *rule-based* heuristic, based on already computed indicators, was chosen.

In order to extract the residential municipality, denoted as m_o^{res}, it was

[19] The original dataset was partitioned into 10 parts of equal size and each part served as test set in each iteration, while the rest was used to train the decision tree.

[20] This means that objects of type A may sometimes be wrongly classified as B–D, however, most of those wrongly classified objects were likely to be discarded anyway because of the rules described in the next paragraph.

checked whether a certain minimal number of events happened during non-occupational time ($|E_{o,t=non-occupational}| \geq 30$) *and* whether these events did happen across several different time slots ($|E_{o,t=non-occupational}(R = \{H, D\})| \geq 15$). If this was the case, the most visited municipality during non-occupational time was classified as residential, if not, m_o^{res} was labeled as "unknown".

The procedure for the extraction of occupational municipalities was slightly more complicated, mainly due to two reasons: occupational time generally exhibits higher entropy in terms of $R = \{M\}$ than non-occupational time, as will be seen below (Section 5.2). Secondly, in the survey upon which authoritative commuter balances are based, it is possible for participants to specify not only one but *two* different addresses of occupation — one for work and one for education, if this applies. This can possibly result in two different occupational municipalities, denoted as $m_{o,1}^{occ}$ and $m_{o,2}^{occ}$, a fact for which the algorithm should account. The first part of the algorithm is the same as for m_o^{res}: if not enough events happened during occupational time *or* if the schedule size for $R = \{H, D\}$ is to small, both $m_{o,i}^{occ}$ are deemed "unknown", using the same thresholds as above. If this if *not* the case *and* if the second most visited location during occupational time differed by a share of more than 30% from the most visited location, *both* were extracted as $m_{o,i}^{occ}$, if not, only the most visited location was. For instance, if the most visited location amounted to 65% of all locations during occupational time, and the first- and second-most visited *together* to 96%, 31% are due to the second-most visited location and both places are thus extracted as $m_{o,i}^{occ}$.

In conclusion, applying this algorithm thus resulted in a SR_o described through possibly none, one or two $m_{o,i}^{occ}$, and none or one m_o^{res} for each object. Objects where m_o^{res} is unknown could not be used in the evaluation of spatial representativeness, presented in the following section, whereas objects where both $m_{o,i}^{occ}$ are unknown could be used for the evaluation of representativeness but not for the calculation of commuter balances as presented in Section 4.4.5.

4.4.4. Evaluation of Representativeness

Having a reasonably certain indication of the place of residence for most of the objects allows to summarize the number of Twitter users residing in each municipality. This figure, in turn, can be evaluated with regards to official population count data, thus indicating whether a municipality is likely over- or underrepresented in geosocial, egocentric VGI. Since findings in Section 3.3.1 and 4.3.3 suggest that most of the users on Twitter are either young adults or teenagers, and that people aged above 60 are barely present, it was decided to use the population of people aged between 15 and 64 as reference. Not only does this definition increase comparability, it is also in accordance with the one of working population (see Section 3.1.2).

As the actually usable data from VGI is very sparse, directly comparing data on the level of municipalities does not lead to very expressive results. Therefore, population counts of municipalities were also aggregated according to their membership to types of certain spatial divisions (Section 3.2.2). By doing so, not only can more significant findings be derived. One can also try to make statements about certain socio-demographic properties of Twitter users, since those are encoded in some of the spatial divisions (see Section 2.3.1 for similar approaches). In order to visualize the disparities between the data from Twitter and actual population counts, the remaining **2,178** objects with known residential municipality were distributed over the respective groups of municipalities according to the actual population in these groups, which yielded the *expected* population. This figure could then be directly compared to the *observed* number of Twitter users per type.

4.4.5. Calculation and Comparison of Commuter Balances

In accordance with the system defined by the structural survey on mobility and transport by the FSO (Section 3.2.1), commuter balances of districts are described as tuples

$$d_i = \langle n_i, stay_i, out_i, in_i, cb_i \rangle \tag{4.5}$$

where $d_i \in D$ signifies one of the 147 districts of Switzerland, n_i the number of objects where m_o^{res} is defined and part of d_i, $stay_i$ the number of objects where m_o^{res} is defined and part of d_i and where at least one of $m_{o,i}^{occ}$ is defined and also part of d_i, out_i the number of objects where m_o^{res} is defined and part of d_i and where at least one of $m_{o,i}^{occ}$ is defined and part of $d_{j,j \neq i}$ and, lastly, in_i the number of objects where m_o^{res} is defined and part of $d_{j,j \neq i}$ and where at least one of $m_{o,i}^{occ}$ is defined and part of d_i. The actual *commuter balance*, expressed through a single number, is defined as $cb_i = (in_i - out_i)/n_i$. A positive cb_i signifies a surplus of commuters, i.e., the district "attracts" commuters, while a negative cb_i signifies a deficit, i.e., the district "provides" commuters. Note that this statistic does not account for the number of commuters that are occupied in the same district ($stay_i$).

As was mentioned in Section 3.2.1, official commuter balances are available separately for the mobile working population and the mobile population in education (see Section 3.1.2 for definitions). On average, people in education

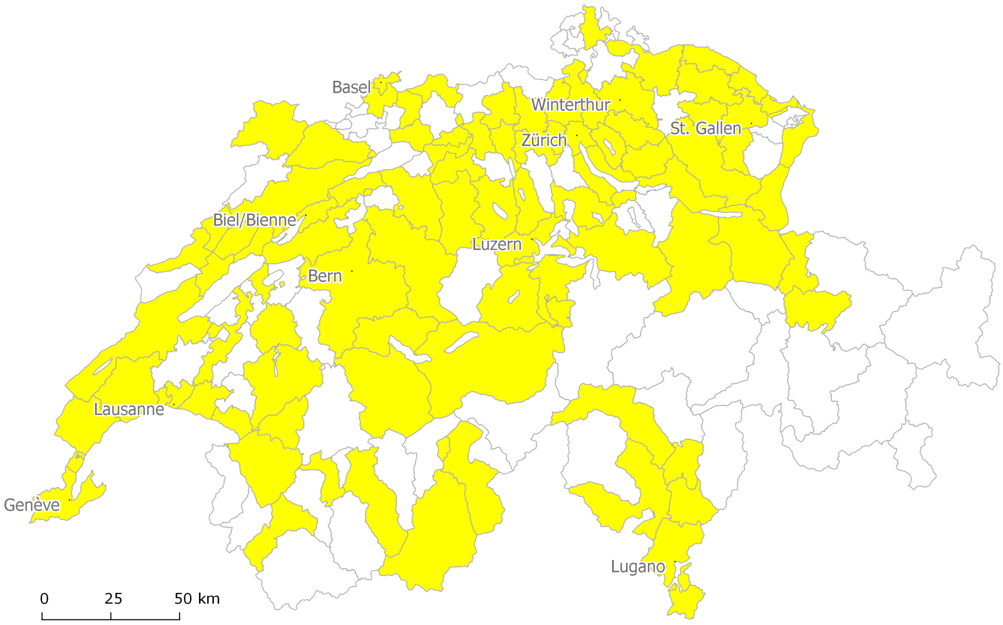


Figure 4.16: *Districts used in the comparative analysis (shown in yellow). Also shown are the ten biggest cities of Switzerland.*

amounted to only 16.4% (standard deviation: 2.6%) of all residential commuters per district. Since this distinction can not be made for the Twitter data, both official datasets were added together and used as a common reference. It should be noted that, theoretically, a person could work *and* go to school and thus appear twice in the resulting sum, but this could also be the case with Twitter users, as, possibly, two different occupational municipalities could have been detected.

Another aspect worth to consider are the partially very large confidence intervals for the estimations in the official data. As was mentioned previously, this primarily concerns the data about people in education (Figure D.1 in Appendix D). In order to have trustworthy reference figures, districts with two-sided 95%-confidence intervals amounting to over 25% and 50% of the estimated values were excluded from the analysis (mobile working population and mobile population in education, respectively), which resulted in the retention of 90 districts. As can be seen from Figure 4.16, the retained districts are pretty well spread over the study area. While the districts of all major cities are represented, rural districts in the mountainous south-east and south-west as well as some relatively sparsely populated areas in the rest of Switzerland were excluded. One can assume that such districts also have a very small number of residential Twitter users (compare with the findings in Section 5.3.1).

In addition to the quantitative comparison between individual districts, commuter balances in VGI were re-calculated on the level of cantons. Here, a quantitative comparison could not be conducted *per se*, because official figures were not available on this level and summing up values on the level of districts does not lead to the same results. For instance, somebody could be an outgoing commuter on the level of districts but still be a staying commuter on the level of cantons, and a mere aggregation would, in this case, lead to an overestimation of outgoing commuters. Nonetheless, a qualitative, side-by-side comparison with official data was possible because the FSO had published a graph showing the balances for all cantons (BFS, 2013e). The same publication also contains an origin-destination matrix which shows the relative share of people commuting from a municipality in a specific spatial type to a municipality in another type. For instance, the matrix shows which share of people commuted from a core city to a rural municipality or from a suburban municipality to another suburban municipality, et cetera. Since such type-specific balances could be easily computed for the Twitter data as well, they were also compared with each other. Lastly, a map showing the numbers of staying, outgoing and incoming commuters in each of the agglomerations in 2010 was published by the FSO (BFS, 2013d). Even though the data stems from 2010, it is likely that the map accurately reflects commuter balances of 2011, at least for the biggest agglomerations. Therefore, commuter balances as found on Twitter were re-calculated for the ten biggest agglomerations and once more compared.

5. Results

5.1. Temporal Patterns

5.1.1. Daily Patterns

Concerning daily patterns, the overall distinct object volume shows significant variations over time (Figure 5.1), and the volume of events is similarly distributed (Figure C.1 in Appendix C). During an average day, object volume has its first maximum around 07:00–08:00, when most people go to their occupation (compare with Section 3.1.2). It then continually rises until noon, when most people take a break from their occupation. Interestingly, the large majority of objects are active during the evening hours, with a global maximum at around 22:00, when most people are at home.

If the volume of objects is categorized according to certain spatial divisions (Section 3.2.2), several things become apparent (Figure 5.2 and 5.3). First of all, there exists a striking imbalance between volumes of the French- and the German-speaking region. Municipalities in the French-speaking region seem to have a considerably higher share of active objects at any time of the day, with amounts sometimes more than twice as large. Nonetheless, all linguistic parts of Switzerland seem to show mostly similar daily patterns, although regions with more absolute volume appear to have more pronounced, extreme patterns. This is most likely due to a smoothing effect, i.e., less active objects in absolute terms make it harder to perceive actual patterns. The patterns for the volume of events are again very similar, although it seems that, in regions with many active objects, overproportionally more events are broadcast, which may hint at a more prolific behavior in these regions (Figure C.2 and C.3 in Appendix C, respectively). When looking at the volumes according to the division between urban and rural municipalities, the rather minuscule share of rural municipalities becomes apparent. However, also here, all types show a very similar daily pattern, even though the evening peak is less pronounced in core cities than in agglomeration municipalities. This might be a sign of people subsequently leaving work

Figure 5.1: *Hourly volume of distinct objects (whole study area) (N=2,380 objects). Error bars signify 95%-confidence intervals.*

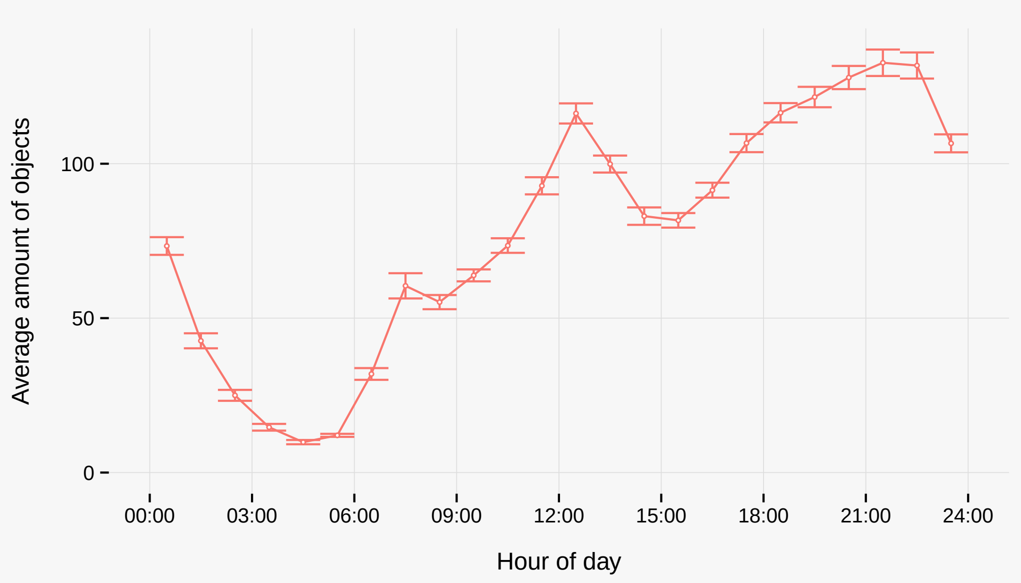
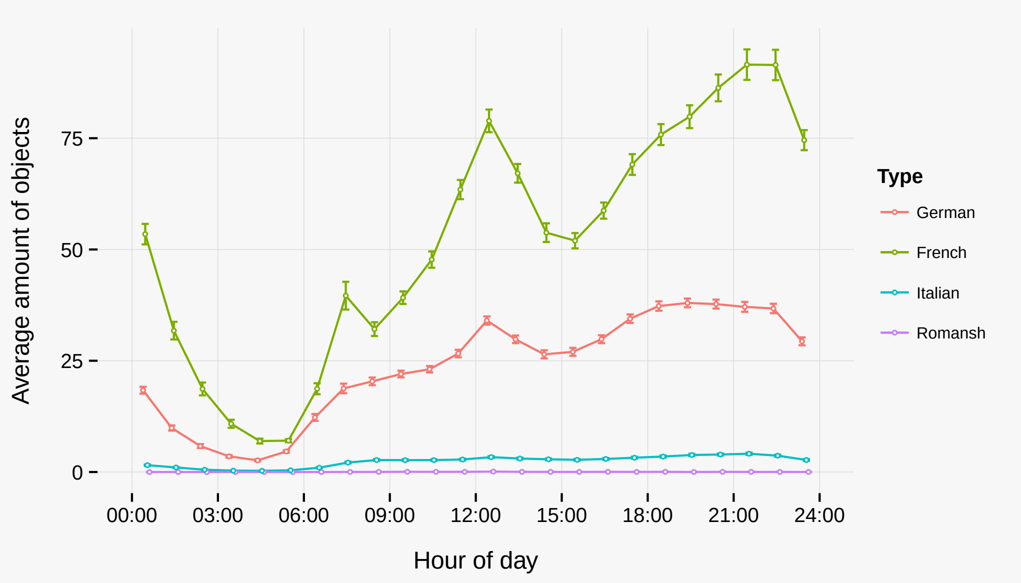


Figure 5.2: *Hourly volume of distinct objects grouped by linguistic region (N=2,380 objects). Error bars signify 95%-confidence intervals.*



or school in the core city and going home to the suburbs.

5.1.2. Weekly Patterns

While daily patterns are quite characteristic, there are no significant weekly patterns (Figure C.4 and C.5 in Appendix C), with a few exceptions. For instance, there seems to be a barely noticeable but still significant rise in distinct objects during the weekend in rural municipalities (Figure 5.4), which might be due to touristic activity or because of people who live in the city during the week and outside of the city during the weekends (e.g., students). At the same time, activity in core cities seems to reach a global minimum during Sundays, but this might just be a random effect. As mentioned above, object activity in individual

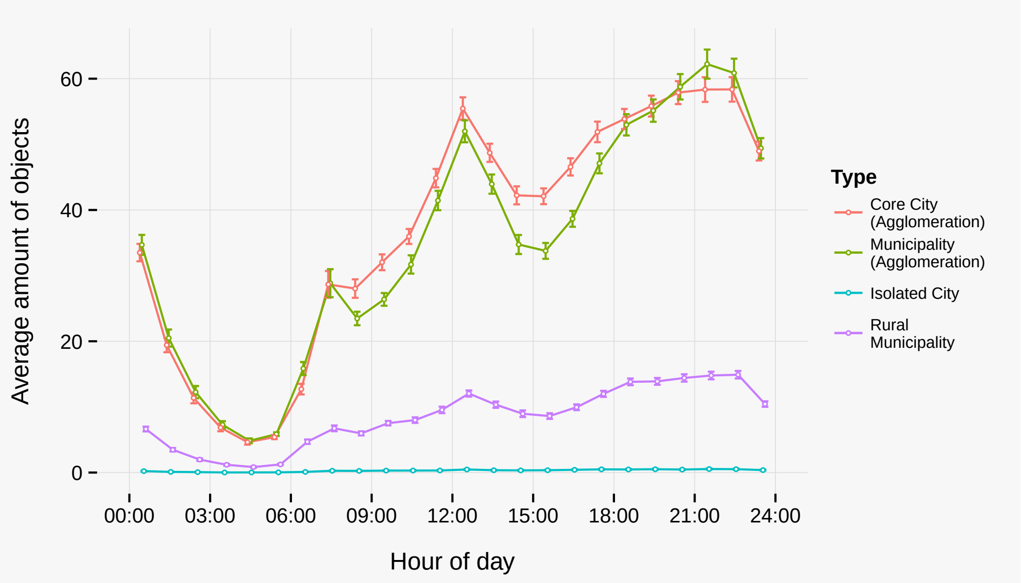


Figure 5.3: *Hourly volume of distinct objects grouped by urban and rural municipalities (N=2,380 objects). Error bars signify 95%-confidence intervals.*

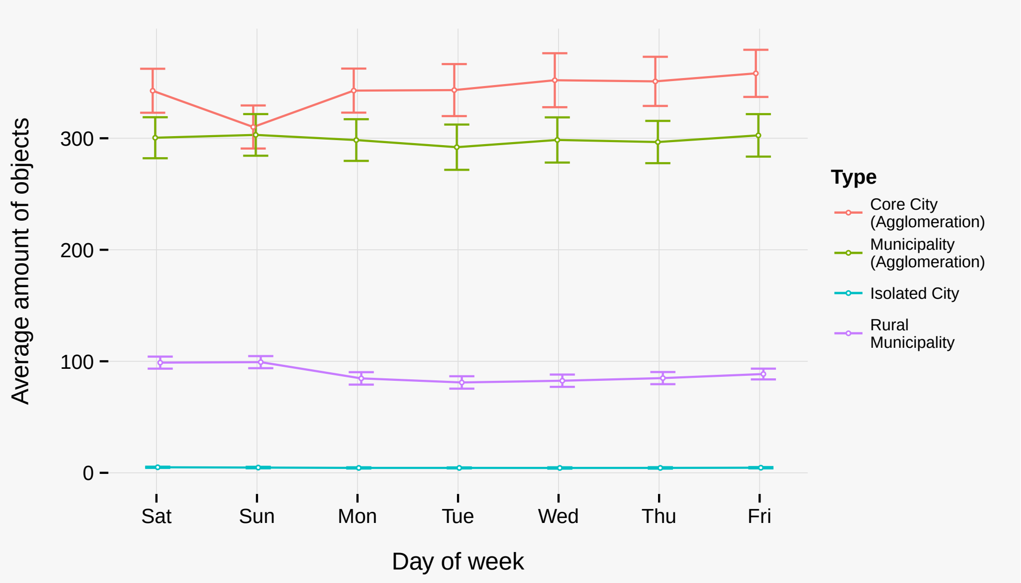
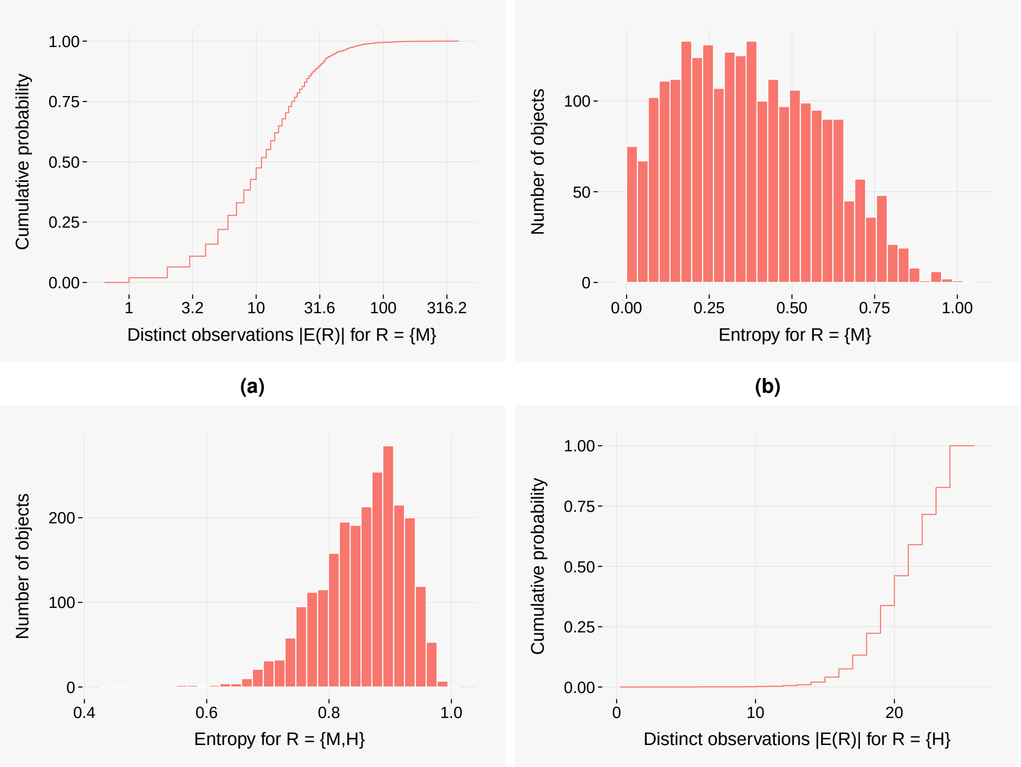


Figure 5.4: *Daily volume of distinct objects grouped by urban and rural municipalities (N=2,380 objects). Error bars signify 95%-confidence intervals.*

regions does not sum up because of the way it is computed.

To conclude, even though such daily and weekly patterns cannot be directly used to evaluate representativeness because they are not based on where objects actually live, they still hint at a regionally non-representative distribution of Twitter users. Furthermore, they show that event volume is more or less steady during the week, which allows to gain information from weekends, too. As could be seen, most objects are active during noon and in the evening, which certainly helps to extract spatiotemporal routines, as both of these daytimes lie in occupational and non-occupational time, respectively.



(a)

(b)

(c)

(d)

Figure 5.5: *Entropies and schedule sizes for different R. Schedule sizes of $R = \{M\}$ (logarithm to the base of 10) (a), entropies of $R = \{M\}$ (b), entropies of $R = \{M, H\}$ (c) and schedule sizes of $R = \{H\}$ (d) (N=2,380 objects).*

5.2. Regularity in Spatiotemporal Routines

5.2.1. Entropies and Schedule Sizes

Only about 2–3% of all objects visited only one municipality throughout the collection process, and about 5% visited only two or less municipalities (Figure 5.5a). Since the entropy for visited municipalities is often below 0.5, one can guess that a few municipalities are visited very often while the majority is visited only sparsely (Figure 5.5b). Contrarily, the entropy for municipality-hour-pairs is generally high[1] (Figure 5.5c), which could mean that different locations are visited multiple times a day or *vice versa* and that there are no strict visiting patterns, although this might be due to the high temporal resolution. In fact, more than 50% percent of all objects were active during at least 20 different hours of the day, and no object was active during less than 6 different hours (Figure 5.5d). Another interesting aspect that can be found is the non-existing relationship between entropy and schedule size for municipalities (Figure 5.6). In other words, entropy does not increase for objects who visited more municipalities, which is ultimately desirable.

[1] The distribution of entropies for $R = \{M, D\}$ is similar to the one for $R = \{M, H\}$, and is thus not shown here.

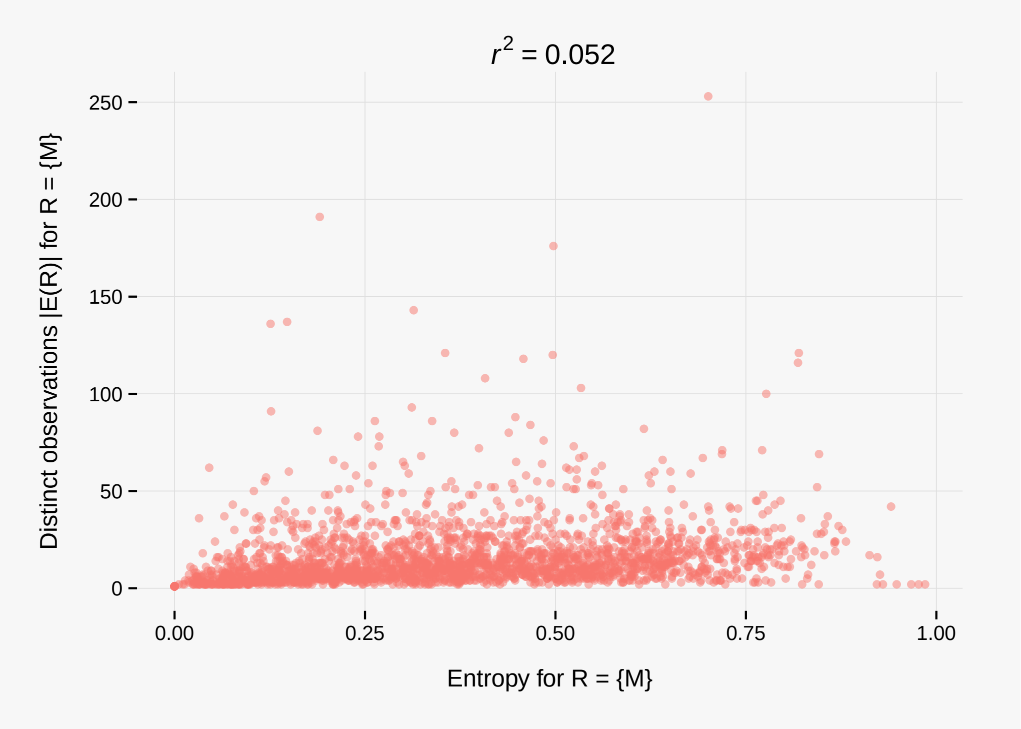


Figure 5.6: *Relationship between schedule sizes and entropies for $R = \{M\}$ (N=2,380 objects).*

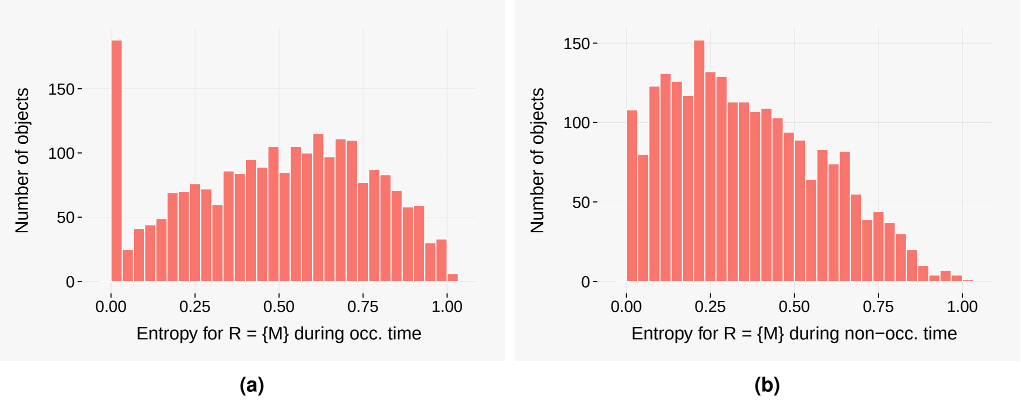


Figure 5.7: *Entropies for $R = \{M\}$ in occupational (a) and non-occupational time (b) (N=2,380 objects).*

Moreover, when $Entropy_{norm}(E_o, R = \{M\})$ is calculated separately for events happening in occupational and non-occupational time, it appears that occupational time generally exhibits a higher entropy, with a median of 0.51 (Figure 5.7a). The median entropy value for non-occupational time is with 0.33 considerably lower (Figure 5.7b). However, a significant number of objects only visited one municipality during occupational time, resulting in an entropy value of 0. A visual inspection showed that these mostly belong to pattern type *B* (Section 4.4.3). If one excludes these exceptional cases, the median of entropy values during occupational time amounts to a even higher value (0.55), reaffirming that the distributions are indeed different.

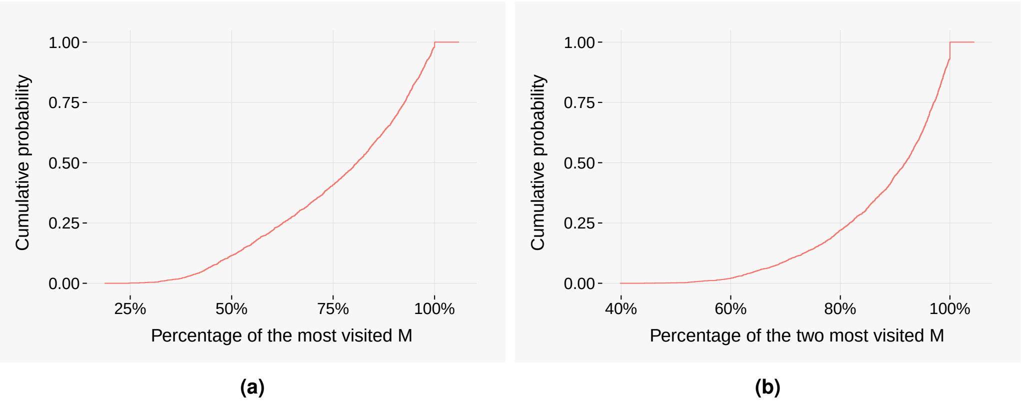


Figure 5.8: *Frequency of most visited locations per object. Most visited location (a) and two most visited locations (b) (N=2,380 objects).*

5.2.2. Most Visited Locations

Concerning the frequency of the n-most visited location, the data show that 75% of all objects visited their most frequent location in about 60% (or more) of their events (Figure 5.8a). The second- and first-most visited locations together even amount to approximately 80% (or more) for 75% of all objects (Figure 5.8b).

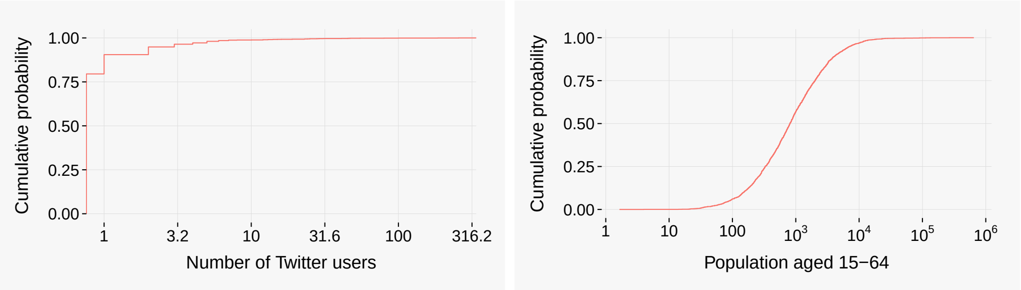
Together, these findings reveal that there is a certain regularity or continuity concerning visited locations — most objects tend to visit the same municipalities over and over again. The properties of the data therefore generally justify the steps taken to extract spatiotemporal routines, as detailed in Section 4.4.3.

5.3. Evaluation of Representativeness

5.3.1. Representation of Individual Municipalities

For about 80% of the 2,515 Swiss municipalities, no residents were found in the Twitter data (referred to as *residential Twitter users* from now on), and for over 95%, less than 10 residents were found (Figure 5.9a). Contrarily, the majority of municipalities of Switzerland are home to around 1,000 people, and the distribution appears to be log-normal, or rank-sized (Figure 5.9b), as it is often the case for cities of a country (Berry, 1961).

However, if the municipalities *without* any residential Twitter users are excluded from the dataset, one can observe a moderate correlation with the actual population aged 15–64 (Figure 5.10). This is expressed through a *coefficient of determination*, r^2, of 0.38, which signifies that about 40% of the variation in one variable can be explained by the other variable (Nagelkerke, 1991). While this correlation is rather weak, it still helps to assure that the method of extracting residential municipalities did not produce completely random results and that municipalities are indeed, at least roughly, represented according to their pop-



(a)

(b)

Figure 5.9: *Distribution of population count across municipalities (logarithm to the base of 10). As inferred from Twitter (a) and actual population aged 15–64 (b) (N=2,515 municipalities, N=2,178 objects).*

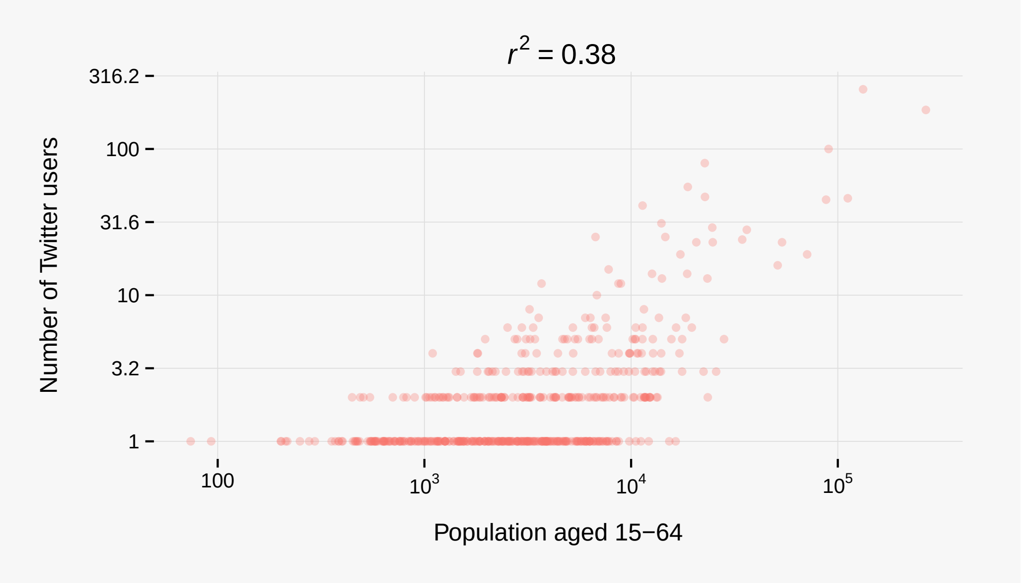


Figure 5.10: *Relationship between actual population aged 15–64 and the number of residential Twitter users per municipality. Only municipalities with at least one residential Twitter user are included (N=515 municipalities, N=2,178 objects). Axes show the logarithm to the base of 10.*

ulation count. Moreover, the rather low r^2 is mostly due to the high variance of municipalities for which only very few Twitter users (< 3) were classified as residents. Such municipalities, associated with a high degree of uncertainty, constitute the majority, which poses a challenge for directly comparing the Twitter data to fine-grained authoritative data. In fact, there even exist 5 municipalities with more than 10,000 inhabitants but only one residential Twitter user, and all of them are core cities or municipalities of agglomerations.

A broad look at the geographical density of residential Twitter users (Figure 5.11) reveals a basic similarity with the actual population density (Figure 5.12). As one would expect, all major cities seem to be relatively densely inhabited by Twitter users and the rural and mountainous regions are only sparsely populated, although the countryside appears to be spotted by patches of seemingly random occurrences of Twitter users. If one computes the *expected* number of residential Twitter users per municipality, as it was done for spatial divisions (Section 4.4.4), and subtracts this figure from the count of actually *observed*

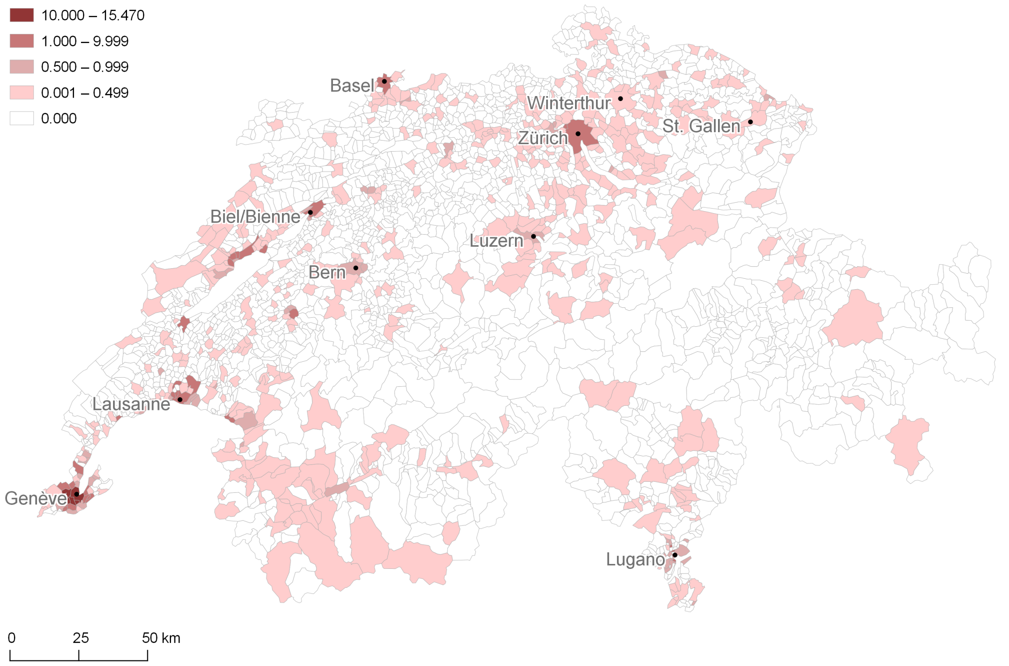


Figure 5.11: *Residential Twitter users per km^2. Also shown are the ten biggest cities of Switzerland. "No data" corresponds to municipalities having no residential Twitter users.*

users, one can visualize in which regions municipalities are over- or underrepresented. On the map in Figure 5.13, municipalities which are underrepresented on Twitter therefore have negative density values; those which are overrepresented have positive values. As can be seen, most of the ten largest cities are slightly or highly overrepresented, with the exception of Winterthur and St. Gallen, which are slightly underrepresented, as well as Basel and Luzern, which are correctly represented. It also appears that the French-speaking region in the western part of Switzerland contains a higher share of overrepresented municipalities than the rest of the country. Without conducting an extensive analysis, one might also suspect that patterns of under- and overrepresentation are spatially auto-correlated (O'Sullivan & Unwin, 2003). This is probably related to the fact that over- or underrepresented regions occur where the population density is reasonably high, which, itself, is auto-correlated.

5.3.2. Representation of Municipalities Aggregated According to Spatial Divisions

While individual municipalities are barely comparable due to sparse data, the comparison of municipalities aggregated according to various spatial divisions gives more insights about the socio-demographic representativeness of the Twitter data (see Section 3.2.2 for an overview of the spatial divisions used in this thesis). One of the first, striking examples of unequal representation of the ac-

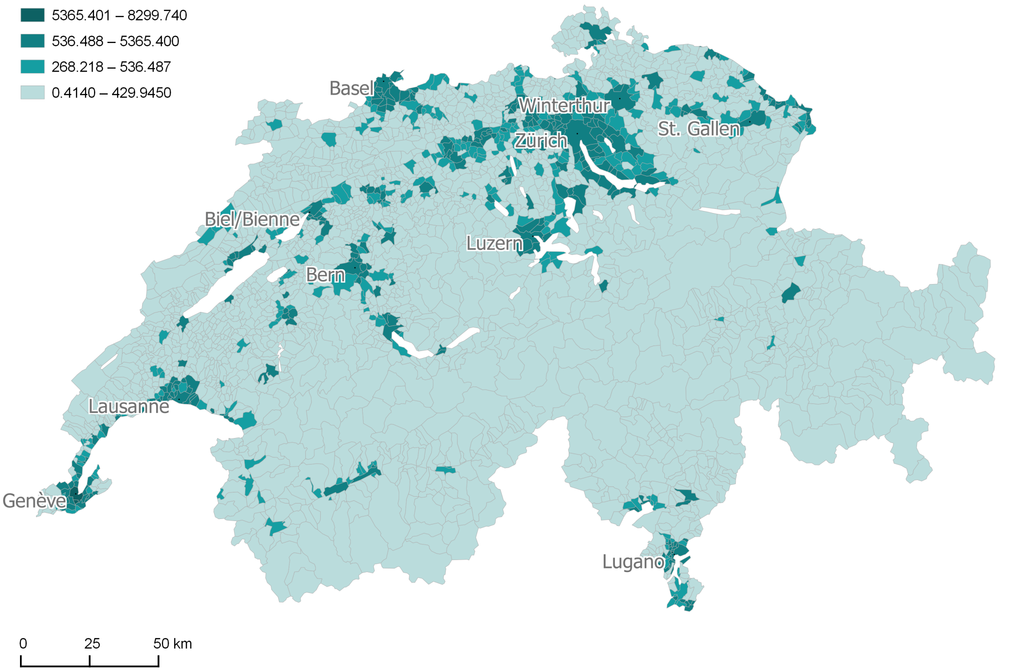


Figure 5.12: *Population aged 15–64 per km^2, 2011. Also shown are the ten biggest cities of Switzerland. Class boundaries are scaled according to those in Figure 5.11.*

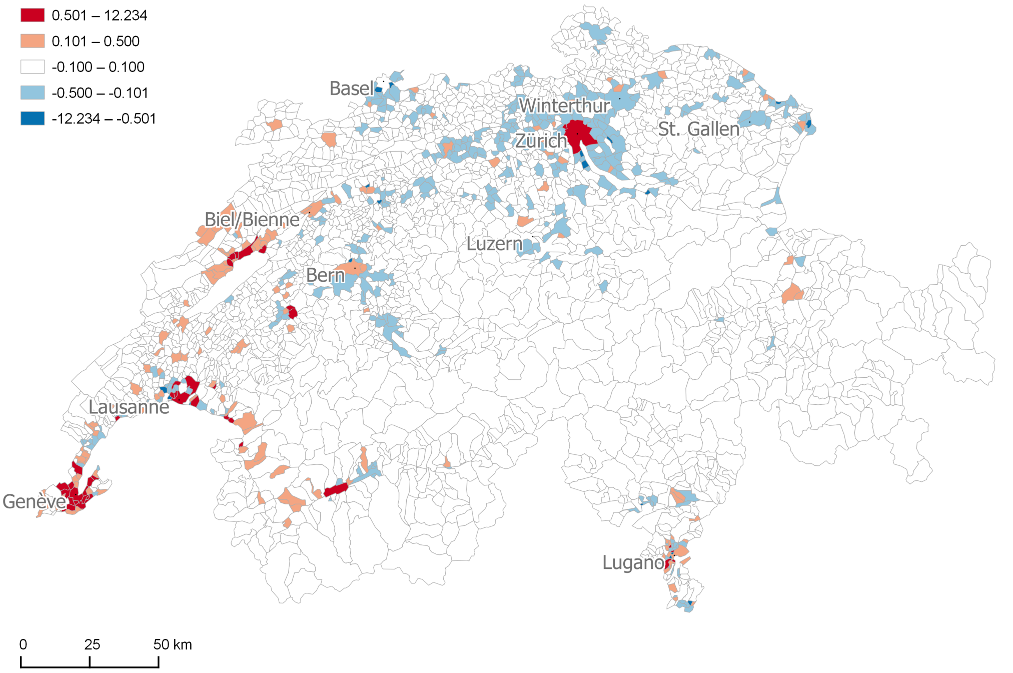
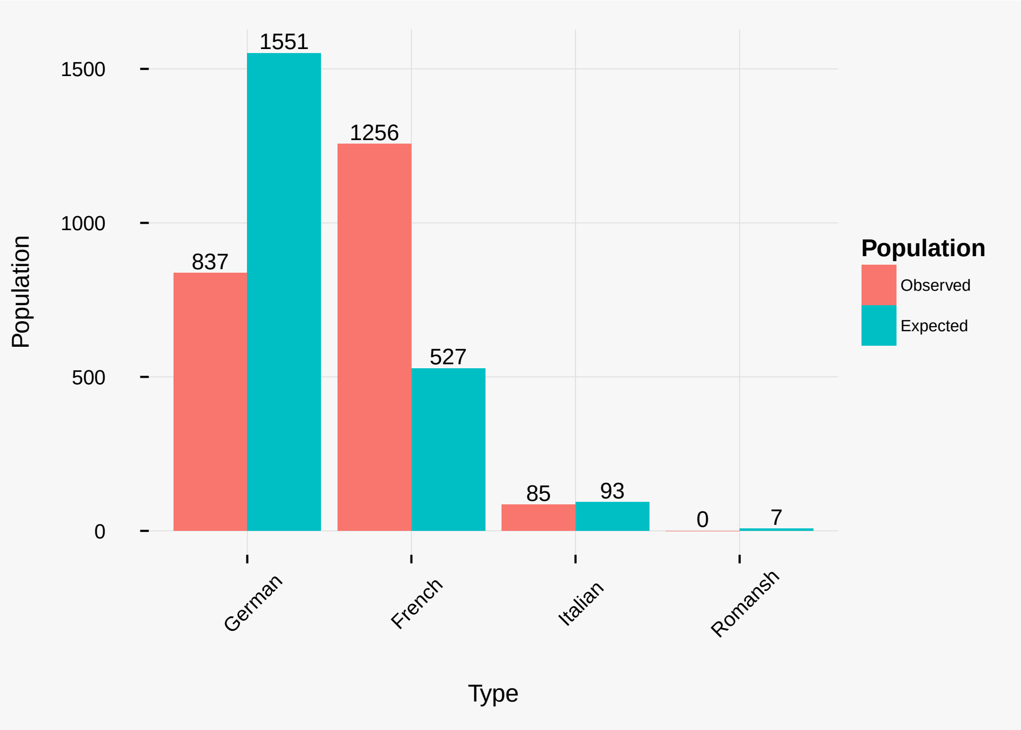


Figure 5.13: *Difference between actually observed and expected residential Twitter users per km^2. Also shown are the ten biggest cities of Switzerland.*

Figure 5.14: *Expected and observed number of residential Twitter users grouped by linguistic regions (N=2,178 objects).*



tual population on Twitter is the strongly overrepresented French-speaking part (Figure 5.14). French-speaking Twitter users are considerably more numerous than German-speaking users, even though the actual population living in the French-speaking part only amounts to approximately a third of that living in the German-speaking part. The seven greater regions can be used to further examine this inequality (Figure 5.15). Clearly, all greater regions except the "Region Lémanique" (Lake Geneva region), which encompasses the French-speaking cantons of Genf and Waadt as well as the bilingual canton of Wallis, seem to be more or less underrepresented. It is likely that the majority of Twitter users of the Lake Geneva region resides in the city of Genf and to some degree in the canton of Waadt. The canton of Wallis, situated in the very south-western part of the country, is probably underrepresented, too, as it mainly consists of rural and touristic municipalities, which are generally underrepresented (see below). An even more detailed picture of the situation is given by the distribution according to the five main metropolitan regions (Figure E.1). It becomes apparent that the Genf-Lausanne ("Genève-Lausanne") region is likely solely responsible for the overall strong overrepresentation of the French-speaking part of Switzerland, as its population is almost three times overrepresented on Twitter (Figure 5.16).

A second dimension along which the unequal representation of different types of municipalities becomes visible, is the division between urban and rural areas (Figure 5.17). In general, core cities of agglomerations, which together

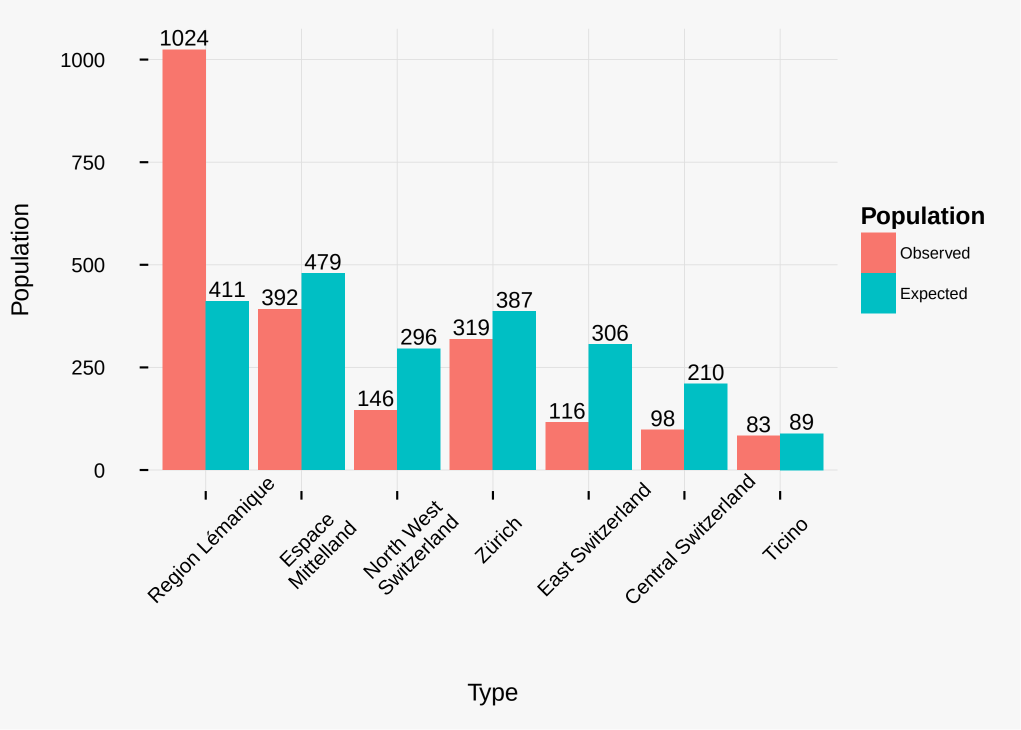


Figure 5.15: *Expected and observed number of residential Twitter users grouped by greater regions (N=2,178 objects).*

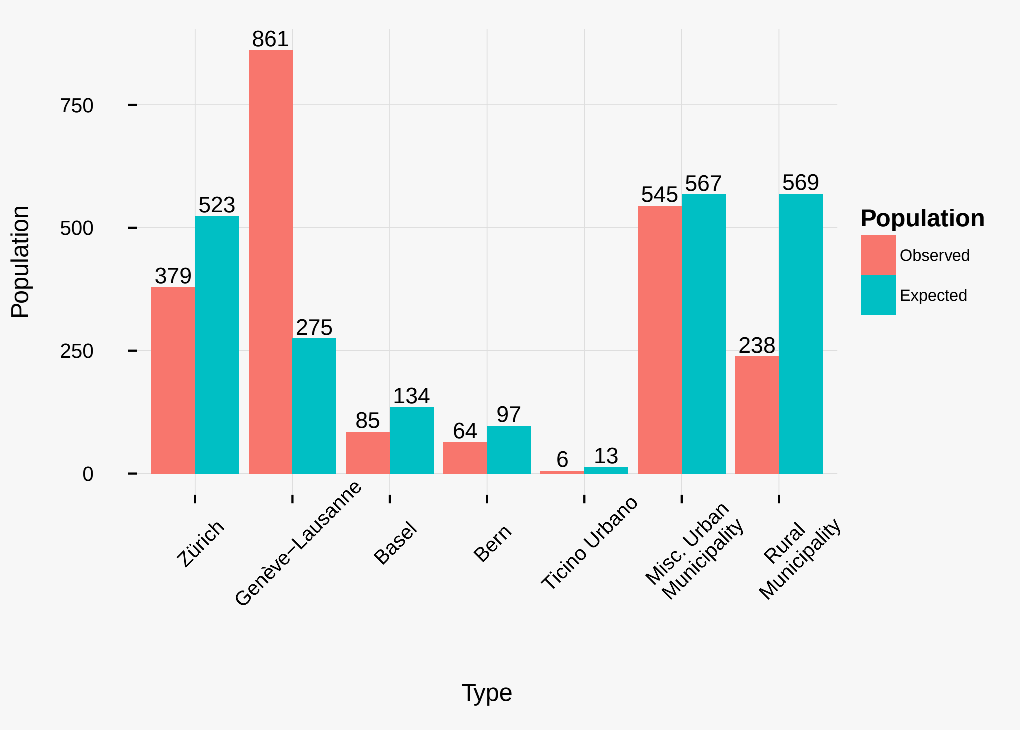
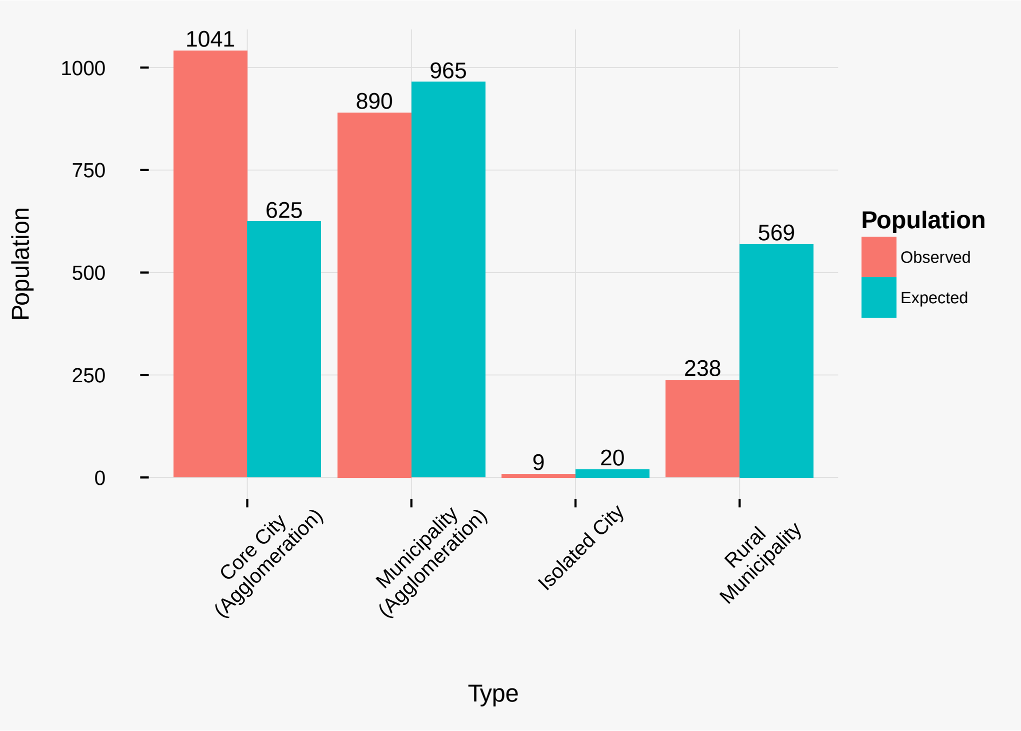


Figure 5.16: *Expected and observed number of residential Twitter users grouped by metropolitan regions (N=2,178 objects).*

Figure 5.17: *Expected and observed number of residential Twitter users grouped by urban and rural municipalities before correcting for linguistic regions (N=2,178 objects).*



housed about 2.2 million inhabitants in 2011, are strongly overrepresented on Twitter. This surplus in observed residential Twitter users is mostly at the expense of rural municipalities, which together housed about 2.1 million people. Nonetheless, also suburban municipalities and isolated cities seem to be slightly underrepresented on Twitter, even though this cannot be stated with a lot of confidence.

In order to further examine the stark divide between core cities and rural municipalities, the imbalance caused by linguistic differences (Figure 5.14) was corrected for. Namely, the *observed* number of residential Twitter users of each municipality was adjusted to the *expected* number of users if linguistic regions were correctly represented. Interestingly, after this calibration, the inequality between core cities and the rest is even more pronounced (Figure 5.18).

Therefore, one can assume that the unequal representation of urban and rural areas is largely independent from the unequal representation of French- and German-speaking regions. For instance, it can be ruled out that the strong overrepresentation of urban areas is *only* due to the overrepresentation of the Genf-Lausanne region (Figure 5.16).

Lastly, one can compare municipalities based on their type as defined by Joye et al. (1988) on behalf of the FSO. This typology is based on a hierarchical center-periphery model, which takes into account several dozens of indicators, such as population count or the percentage of outgoing commuters, as

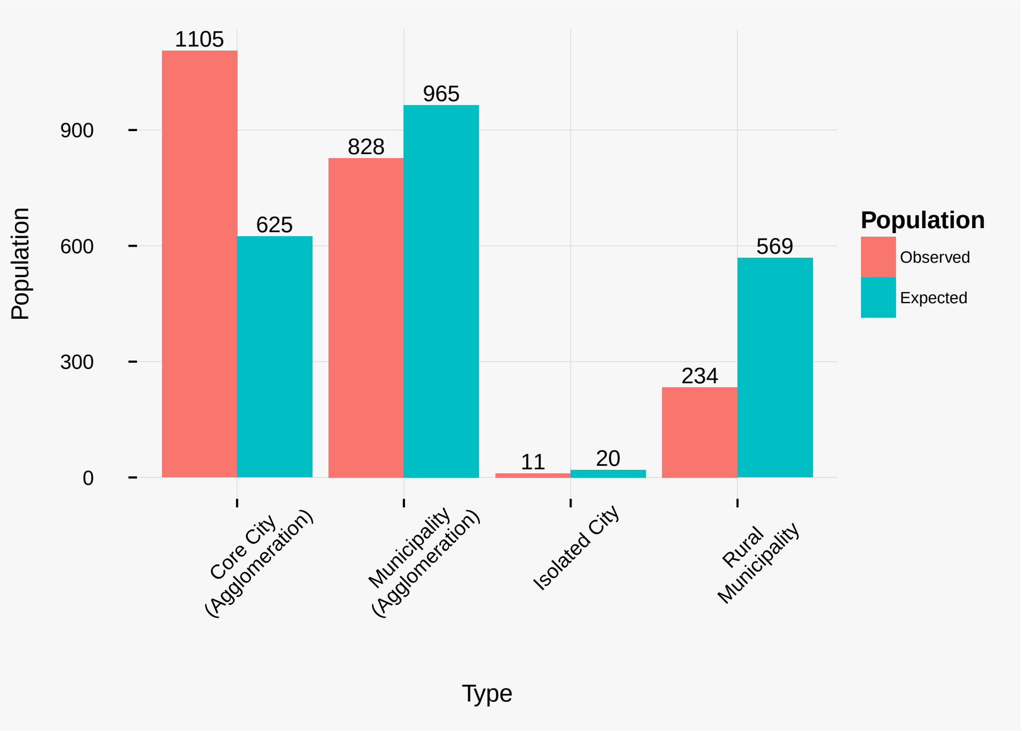
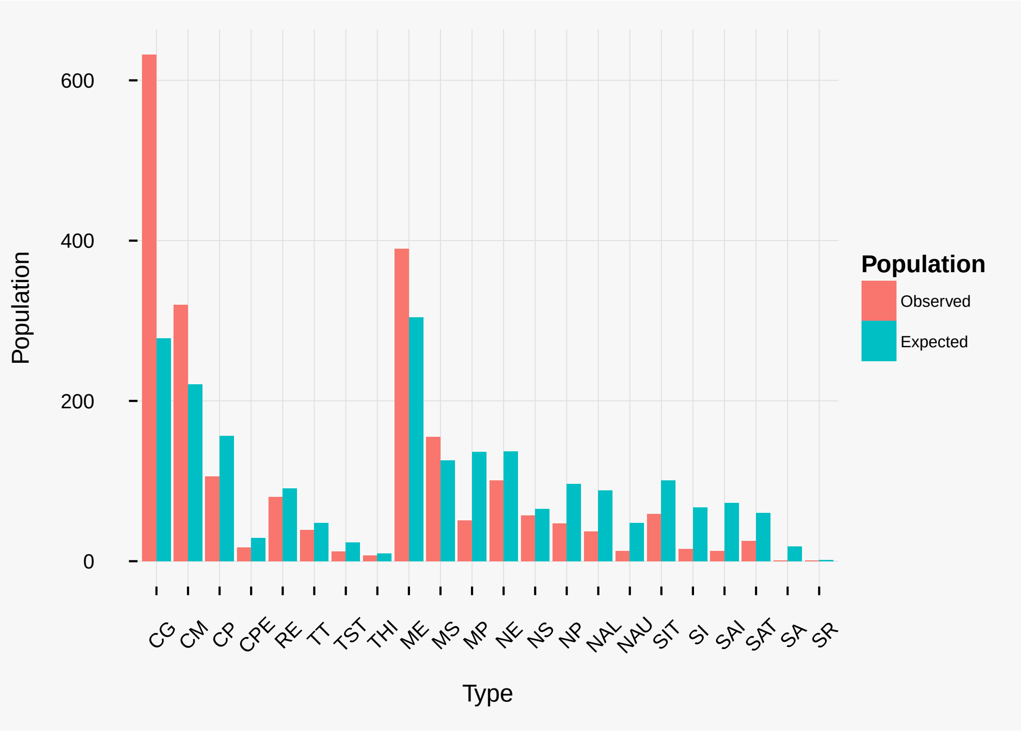


Figure 5.18: *Expected and observed number of residential Twitter users grouped by urban and rural municipalities after correcting for linguistic regions (N=2,178 objects).*

well as existing classifications. For instance, a municipality belongs to the type *CG* (large center) if it is a core city (as defined elsewhere) and has more than 30,000 inhabitants. Over the years, these thresholds were only slightly adjusted (Schuler & Joye, 2005) and classifications are still made on a yearly basis. While Figure 5.19 shows expected and observed values for all 22 types, only a few will be explicitly considered and explained here. The reasons for this are twofold. First, the typology itself is quite complex and often too detailed for the purposes of this thesis, and second, many types probably lack statistical significance because of the low volume of residential Twitter users found in those.

First and foremost, the highly overrepresented large centers (*CG*) attract attention, especially because middle centers (*CM*, core city and more than 14,000 inhabitants) are a lot less overrepresented, and small and peripheral centers (*CP* and *CPE*, respectively) are even underrepresented. So-called *workplace municipalities*, which afford a high share of work opportunities and where a lot of people commute to, seem to be overrepresented slightly, but only if they are part of a metropolitan area (*ME*). If they are not part of a metropolitan area (*NE*), the opposite is the case. Other significantly underrepresented municipalities are those belonging to the type of *MP* and *NP*, which are *periurban* municipalities of metropolitan and non-metropolitan areas, respectively. The definition of periurban encompasses municipalities which i) are part of an agglomeration, ii) do not belong to the definition of *suburban* municipalities, which, in turn, encom-

Figure 5.19: *Expected and observed number of residential Twitter users grouped by 22 municipality types (N=2,178 objects).*



passes municipalities with a relatively high share of apartment buildings, iii) do not belong to the above definition of workplance municipalities, and iv) are not classified as *high income* municipalities (*RE*). Lastly, municipalities outside of agglomerations with a high share of outgoing commuters (*NAL*, *NAU*) as well as industrial (*SI*) and agrarian (*SAT*, *SAI*) municipalities seem to be significantly underrepresented.

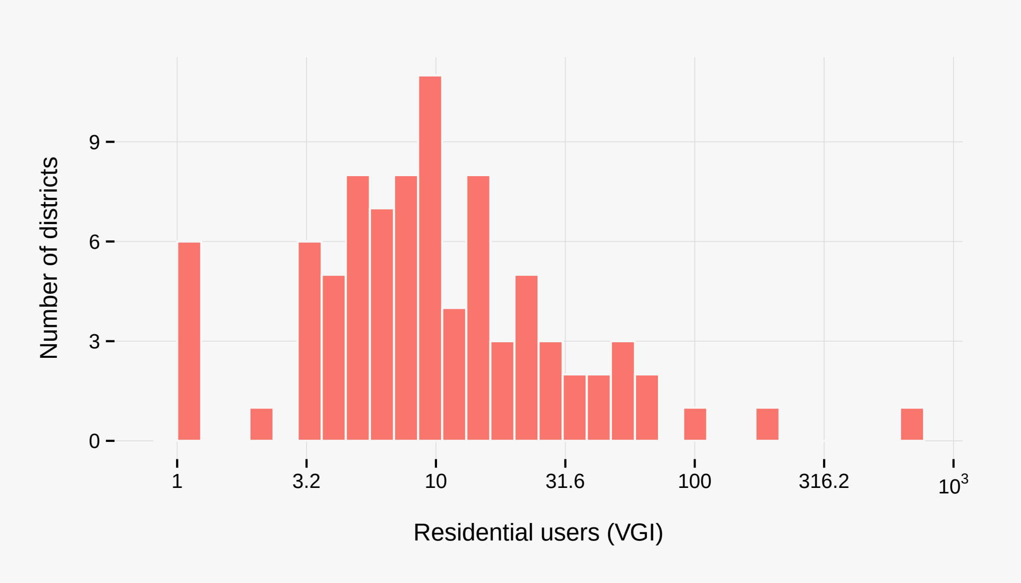
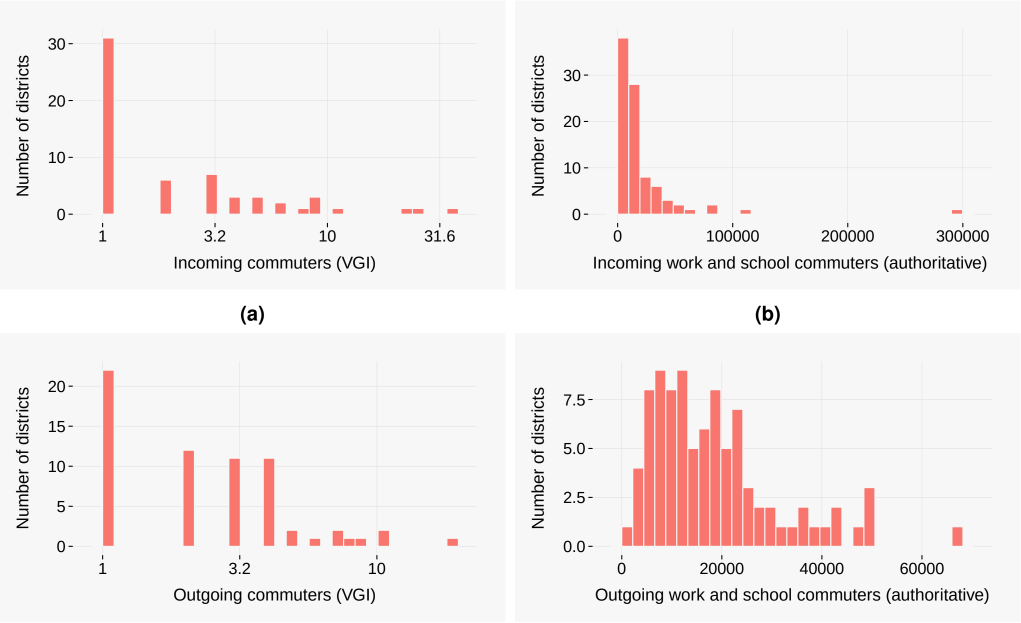


Figure 5.20: *Distribution of n_i per district for Twitter data. Axis shows the logarithm to the base of 10.*

5.4. Comparison of Commuter Balances

5.4.1. Comparison of Individual Districts

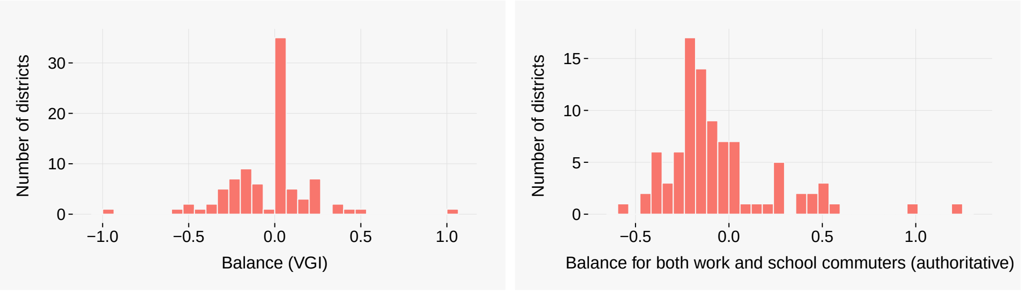
Although residential Twitter users of districts (denoted with n_i in Section 4.4.5) are almost log-normally distributed (Figure 5.20), this is not the case for incoming and outgoing Twitter commuters (in_i and out_i, respectively) (Figure 5.21a and 5.21c, respectively). While only 3 of the analyzed 90 districts have no residential Twitter users, about 30 districts have either no incoming or no outgoing commuters, and 13 have neither. In the authoritative data, both in_i and out_i seem to come from a different distribution. in_i appears to be heavily skewed with many districts having only a small number of incoming commuters and a few having a lot of them (Figure 5.21b), while out_i is more evenly distributed (Figure 5.21d). In other words, the majority of districts is the origin for several 1,000 to 10,000 commuters, while the destination of these commuters lies in only a few districts, which provide work and education opportunities for hundred thousands of people. This is also reflected in the distribution of cb_i in the authoritative data, as the large majority of districts has a negative commuter balance (Figure 5.22b). This can not be said for the distribution of cb_i for Twitter users, where most districts are centered around 0, i.e., have neither a clearly positive nor a negative balance (Figure 5.22a). Concerning staying commuters ($stay_i$ in Section 4.4.5), the distribution for Twitter data (Figure 5.23a) is similar to the one of out_i (Figure 5.21c), and only 6 districts do not have any staying commuters. In contrast, the authoritative data (Figure 5.23b) show a rather log-normal distribution similar to the one of in_i (Figure 5.21b).



(a) **(b)**

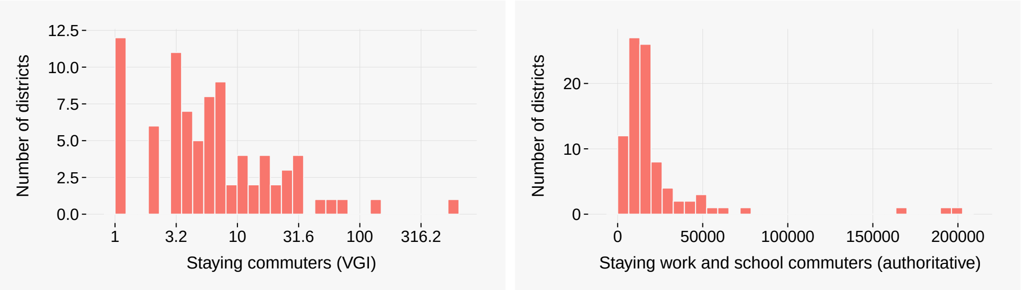
(c) **(d)**

Figure 5.21: *Distribution of in_i and out_i per district for Twitter (a,c) and authoritative data (b,d). Axes for Twitter data show the logarithm to the base of 10.*



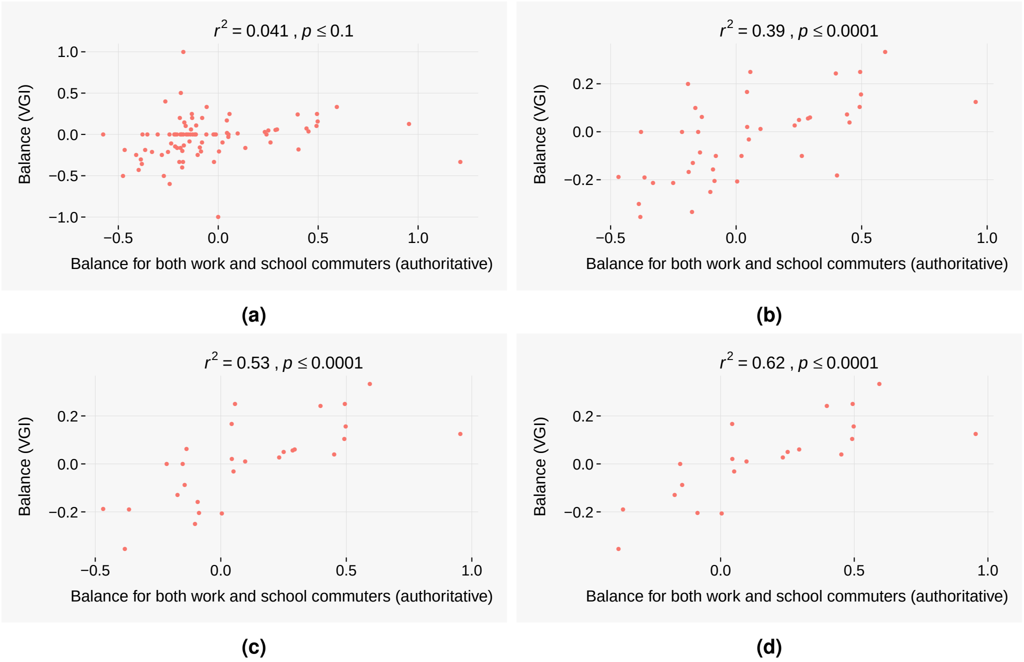
(a) **(b)**

Figure 5.22: *Distribution of cb_i per district for both Twitter (a) and authoritative data (b) (N=90 districts).*



(a) **(b)**

Figure 5.23: *Distribution of $stay_i$ per district for both Twitter (a) and authoritative data (b) (N=90 districts). Axis for Twitter data shows the logarithm to the base of 10.*



(a)

(b)

(c)

(d)

Figure 5.24: *Comparison of cb_i per district between authoritative and Twitter data. For all districts (a) (N=90 districts), for districts with $n_{i,vgi} \geq 10$ (b) (N=41), for districts with $n_{i,vgi} \geq 15$ (c) (N=28) and for districts with $n_{i,vgi} \geq 20$ (d) (N=21).*

If the commuter balances of both authoritative data and VGI are actually compared for each of the 90 districts, no similarity or correlation can be detected at first (Figure 5.24a). It seems that, especially for districts which neither have a positive nor a negative cb_i on Twitter, the variance of the authoritative cb_i is quite high. Moreover, many districts exist which have a positive balance on Twitter, but a negative one in the authoritative data. These two phenomena are mainly responsible for the low correlation.

In Section 5.3.1, one could observe that municipalities with a low number of residential Twitter users could not be used to infer actual population values with high certainty, either. Therefore, it makes sense to look at the correlation between districts which have a certain minimal *support* of n_i on Twitter, as these are likely to be more significant in terms of cb_i. In fact, as one raises the threshold for inclusion of districts from at least 10 residential Twitter users (Figure 5.24b) over at least 15 users (Figure 5.24c) to at least 20 users (Figure 5.24d), the apparently linear correlation successively becomes stronger, while still being highly significant even when only 21 districts remain. Districts where both authoritative and Twitter commuter balances do not have the same algebraic sign are becoming more and more sparse, too. It also appears that balances are generally less pronounced on Twitter, and extreme cases ($|cb_i| \geq 0.5$) begin to disappear as soon as one raises the minimal support.

For the sake of completeness, and because commuter balances are not normally distributed, as one could see in the plots above, Spearman's ρ, which mea-

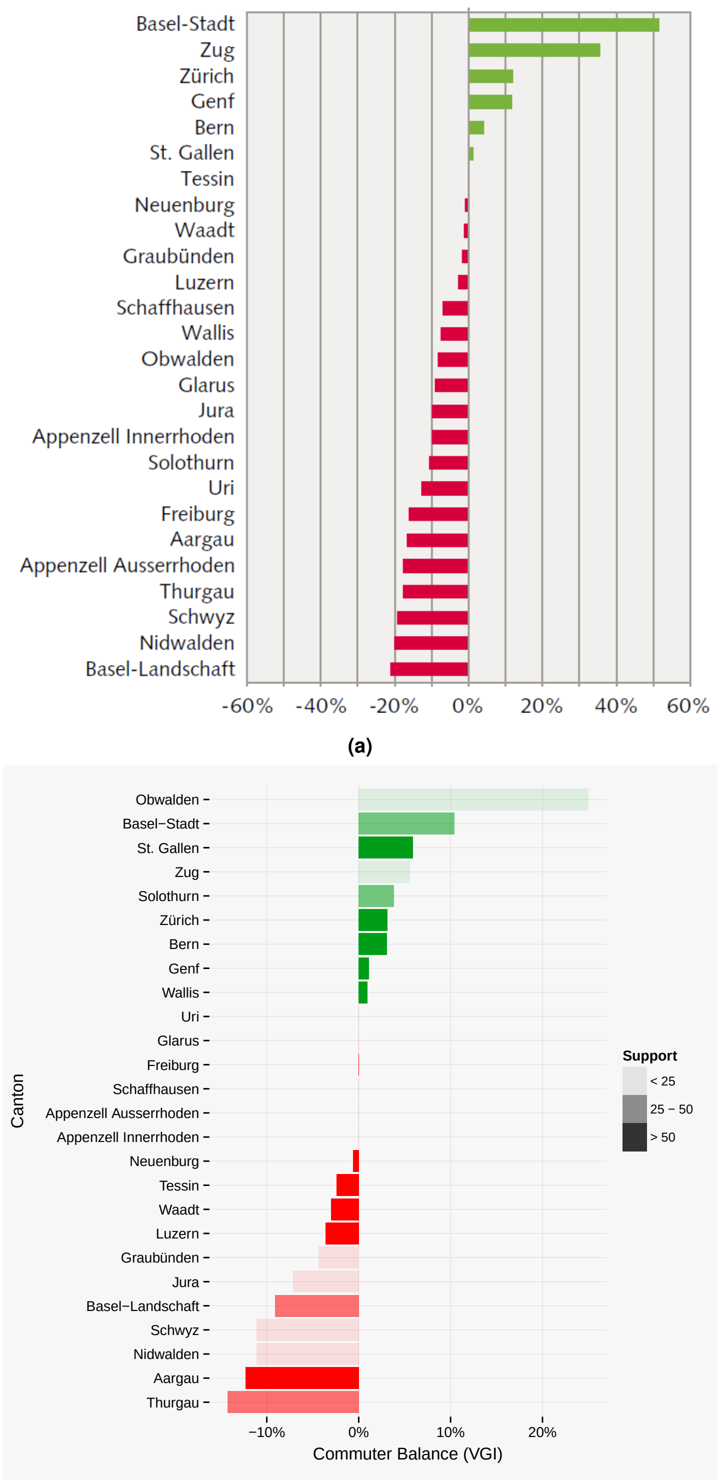
Table 5.1: *Association of cb_i between authoritative and Twitter data as measured through Spearman's ρ and Pearson's correlation coefficient.*

$\mathbf{n_i} \geq$	N	ρ	r
$-\infty$	90	0.41	0.2
10	41	0.65	0.62
15	28	0.76	0.73
20	21	0.84	0.79

sures the *association* between the *ranks* of two variables, was computed. A ρ of ± 1 signifies a perfect monotonical relation between the two variables, i.e., as one increases or decreases one variable, the other is always increased or decreased too (or *vice versa* in the case of $\rho = -1$). The resulting values for both ρ and the previously plotted Pearson's correlation coefficient $r = \sqrt{r^2}$ are displayed in Table 5.1. Clearly, the association between the ranked variables, as measured through ρ, is always higher than the correlation, as measured through r, although only slightly. As it is the case with the correlation coefficient, the higher the support, the closer the association between variables. Furthermore, the difference between ρ and r is most visible for the case where all 90 districts are included — even though no or only a very weak linear correlation can be detected, the variables are at least moderately associated in terms of their ranks.

5.4.2. Comparison of Districts Aggregated According to Spatial Divisions

As was mentioned in Section 4.4.5, the FSO also published cantonal commuter balances for 2011 (Figure 5.25a), but the underlying data are not available, and it is not ultimately clear whether the data only concerns the mobile working population or also the mobile population in education (see Section 3.1.2 for definitions). As it is the case with individual districts, most cantons have a negative balance, while only six have a positive one. On Twitter, this ratio is more even, as almost the same number of cantons have positive and negative balances (Figure 5.25b). Again, the balances are generally less pronounced on Twitter, and most fall within ±10%. To more accurately reflect the real balances, the values should thus probably be stretched by a factor of 2. If one only looks at cantons with a support higher than 50 residential Twitter users, the distribution looks quite similar to the one in the authoritative data. For instance, all cantons which have a positive balance in the authoritative data also have more incoming commuters on Twitter. Still, there are some cantons which have a rather large support but are not correctly represented, such as the canton of Wallis, which has a negative balance in the authoritative data, or the canton of Tessin, which does not have a

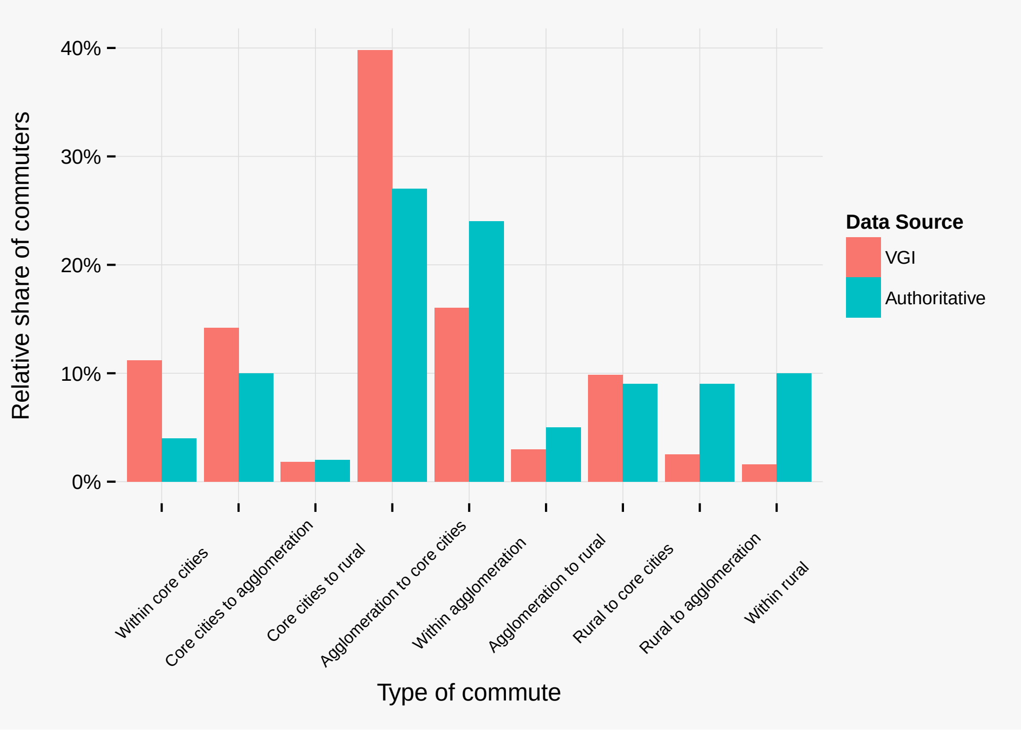


(a)

(b)

Figure 5.25: *Commuter balances per canton for authoritative (BFS, 2013e) (a) and Twitter data (b). In Figure (b), the support signifies the number of residential Twitter users per canton (n_i).*

Figure 5.26: *Commuter flows between and within urban and rural areas. The y-axis shows the percentage of people commuting from a specific spatial type to another. Flows extracted from Twitter are shown in red, those from authoritative sources (BFS, 2013e) in blue. Authoritative data only consider the mobile working population. Core cities include isolated cities.*



negative balance in the authoritative data. Freiburg, even though having quite a large support (not visible in the plot), is not correctly represented as well. If one considers cantons with a support lower than 50 residential Twitter users, some cantons strongly deviate from their authoritative counterparts, such as the cantons of Obwalden, Solothurn and Graubünden. As for individual districts, Spearman's ρ was calculated for the association between the order of the cantons. A high positive value means that cantons are in more or less the same order, while a high negative value signifies that cantons are in opposite order. If all cantons are included, ρ amounts to 0.62. If only those having a support of $n_i \geq 25$ are considered, $\rho = 0.72$. Interestingly, ρ decreases slightly to 0.71 if only those having a support of $n_i \geq 50$ are included. However, one can certainly conclude that, in terms of commuter balance, the order of cantons as seen on Twitter is relatively congruent with the order seen in the authoritative data, and that this congruence is more strong for cantons with a minimal support of residential Twitter users.

Another comparison can be made for commuter movements between and within urban and rural municipalities (Section 3.2.2). The figures published by the FSO (BFS, 2013e) suggest that most people commute from agglomeration municipalities to core cities (including isolated cities) or to other agglomeration municipalities, and comparatively few people travel within core cities and from core cities to rural municipalities. People who have the origin of their commute in rural municipalities together amount to almost 30% of all commuters (Figure 5.26). On Twitter, some clear deviations from these figures can be observed. Movements within core cities and from agglomerations to core cities are strongly overestimated, while those originating from rural municipalities are mostly underestimated. Apart from these disparities, the magnitudes of flows are more or less correctly represented on Twitter. It should be noted that the authoritative data do, in this case, only consider the mobile working population, while the data from Twitter is assumed to also contain people in education (see Section 4.4.5).

Lastly, commuter figures for the biggest agglomerations as inferred from Twitter (Figure 5.28) can be qualitatively compared to official data from 2010 (Figure 5.27), which, again, only consider the mobile working population. Because of the problem of low support on Twitter, only the figures for the ten biggest agglomerations were computed. If one compares the mere balance, i.e., whether an agglomeration has a surplus or a deficit of commuters, most agglomerations coincide with the official data. The only exception is Luzern, where the official data show a slightly positive balance compared to the negative balance on Twitter. Except for Luzern and Winterthur, the ratio between incoming and outgoing commuters seems to be more or less correct, too. A clear deviation from the official data can only be detected for staying commuters, who seem to be tendentially — and, in some cases, heavily — overestimated. This is especially the case for agglomerations where n_i is very large: Genf (Genève), Zürich, Basel and Bern. For this reason, it is difficult to compare the ratio of incoming/outgoing and staying commuters between both Twitter and authoritative data.

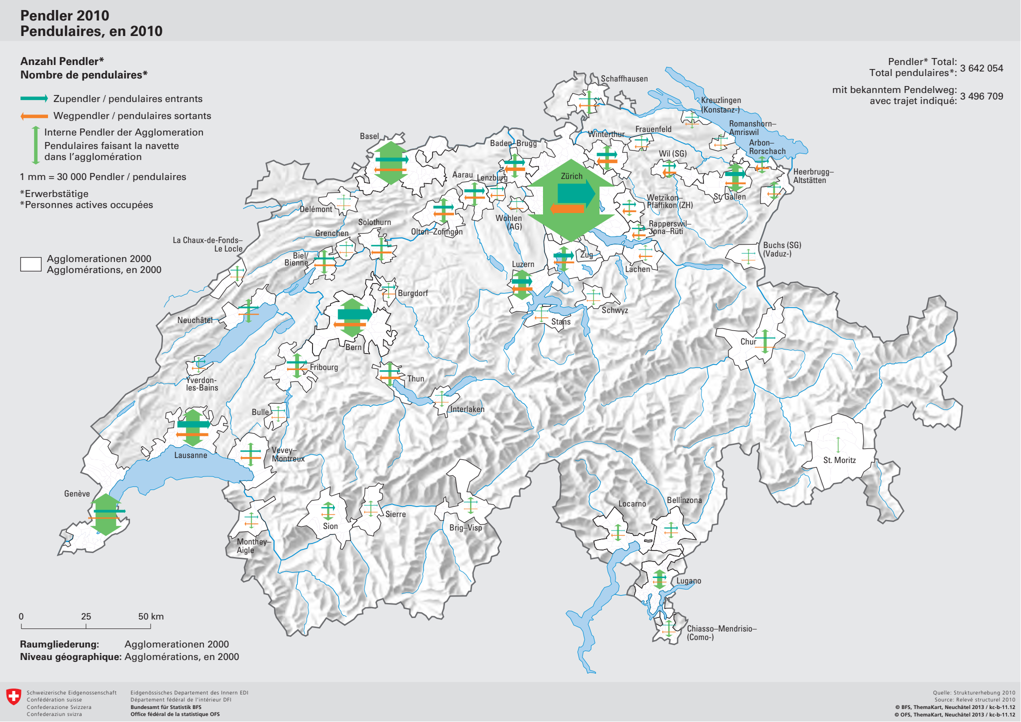

Pendler 2010
Pendulaires, en 2010
Anzahl Pendler*
Nombre de pendulaires*
Zupendler / pendulaires entrants
Wegpendler / pendulaires sortants
Interne Pendler der Agglomeration
Pendulaires faisant la navette dans l'agglomération
1 mm = 30 000 Pendler / pendulaires
*Erwerbstätige
*Personnes actives occupées
Agglomerationen 2000
Agglomérations, en 2000
Pendler* Total:
Total pendulaires*: 3 642 054
mit bekanntem Pendelweg:
avec trajet indiqué: 3 496 709
Schaffhausen
Kreuzlingen (Konstanz-)
Romanshorn–Amriswil
Arbon–Rorschach
Frauenfeld
Winterthur
Baden–Brugg
Basel
Wil (SG)
Heerbrugg–Altstätten
St. Gallen
Wetzikon–Pfäffikon (ZH)
Zürich
Aarau
Lenzburg
Wohlen (AG)
Delémont
Solothurn
Olten–Zofingen
Grenchen
La Chaux-de-Fonds–Le Locle
Biel/Bienne
Rapperswil–Jona–Rüti
Buchs (SG) (Vaduz-)
Luzern
Zug
Lachen
Burgdorf
Schwyz
Stans
Neuchâtel
Bern
Chur
Fribourg
Thun
Yverdon-les-Bains
Interlaken
Bulle
St. Moritz
Lausanne
Vevey–Montreux
Genève
Locarno
Bellinzona
Sion
Sierre
Brig–Visp
Monthey–Aigle
Lugano
Chiasso–Mendrisio–(Como-)
0
25
50 km
Raumgliederung: Agglomerationen 2000
Niveau géographique: Agglomérations, en 2000
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Eidgenössisches Departement des Innern EDI
Département fédéral de l'intérieur DFI
Bundesamt für Statistik BFS
Office fédéral de la statistique OFS
Quelle: Strukturerhebung 2010
Source: Relevé structurel 2010
© BFS, ThemaKart, Neuchâtel 2013 / kc-b-11.12
© OFS, ThemaKart, Neuchâtel 2013 / kc-b-11.12

Figure 5.27: *Work commuters of agglomerations, 2010. Staying commuters are denoted as vertical, green arrows. Incoming and outgoing commuters are denoted as horizontal blue and orange arrows, respectively. 1mm of width of an arrow stands for 30,000 commuters. Source: BFS (2013d).*

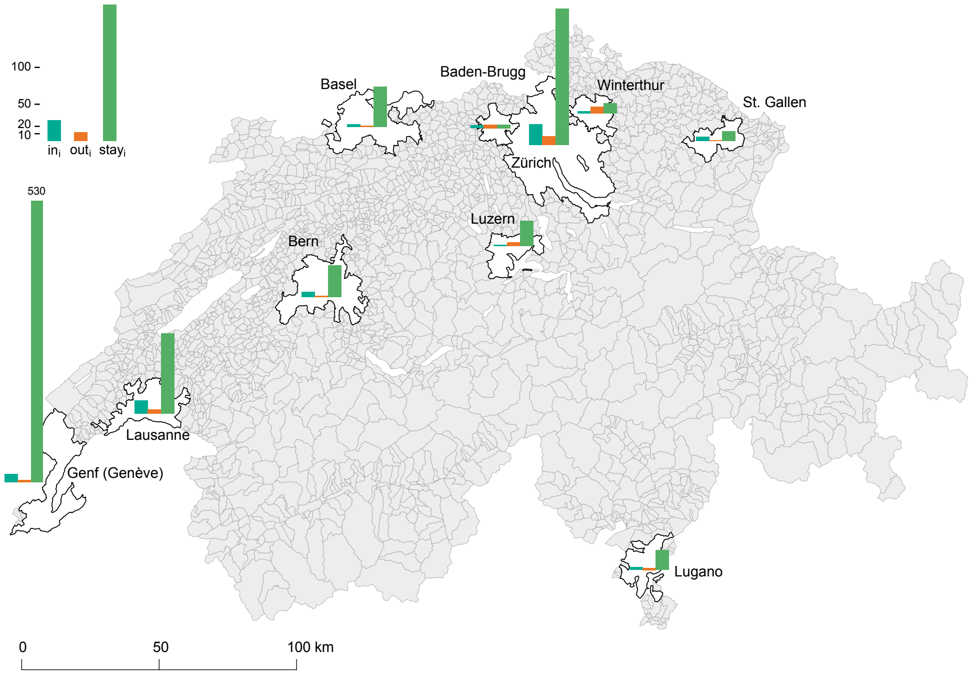


Figure 5.28: *Commuters of the ten biggest agglomerations as inferred from Twitter. Colors correspond to those used in Figure 5.27. Note that the value of $stay_i$ for Genf (Genève) is truncated.*

6. Discussion

In the last chapter it was shown that it is possible to detect spatial and temporal patterns which are comparable to official data and which are plausible in the sense that they do not deviate too strongly from what is considered to be "reality". In other words, it can be said with high certainty that the figures computed and the patterns detected are not random.

On the other hand, every sufficiently specialized methodological process produces results which are not random, and it must be asked whether such a process accurately reflects the structure within the data at hand, in this case, the "real" spatiotemporal behavior of people. Although the apparent similarity with "reality" is an indicator for this, it can not definitely be ruled out that another methodological framework would lead to a different outcome. It is not the goal of this work to optimize such a framework so that it reflects "reality" as close as possible. However, the sensitivity of the data mining process as a whole can and must still be assessed, because it could be that seemingly minor parameter adjustments lead to significantly different results. Therefore, before putting the results into the context of the research questions, their robustness is critically examined.

6.1. Uncertainties of the Methodological Framework

While some aspects of the methodological framework, such as the uncertainties inherent to type of data used, as illustrated in Section 2.2.5, are given and can not be circumvented, others can be more deeply investigated. In particular, the following stages and decisions might have influenced the outcome of the methodological process:

1. The criteria for searching and tracking Twitter users during data collection.
2. The preprocessing stage, particularly the choice of variables and thresholds for detecting and removing non-desirable users.

3. The spatial and temporal discretization of the collected Tweets.

4. The extraction of semantic places as spatiotemporal routines, namely, the parameters of the rule-based heuristic and the particular delimitation of occupational and non-occupational time.

5. The choice of aggregation levels on which the results were compared with the authoritative data, and the uncertainties associated with the authoritative data themselves. These have already been discussed in Section 4.4.5.

6.1.1. Data Collection and Preprocessing

Even though the data collection process could not have been implemented in a very different fashion, the filtering criteria for tracking users could have been stricter or more relaxed (Section 4.2.1). However, even if more (or less) users had been tracked and collected in the first place, they would eventually have been removed (or kept) during the preprocessing stage. That stage was indeed subject to some ambiguity — as Table 4.3 showed, adjusting the indicator thresholds resulted in a successively shrinking (or growing) user base, and it was tried to retain a high quality while keeping as much of the data as possible (Section 4.3.3). As the subsequent, manual validation showed, this was achieved, even though the large majority of the data needed to be discarded. It can be assumed that slightly different parameter thresholds would not have significantly affected the final analysis, as marginal adjustments only led to minor changes in the number of retained objects.

The parameters used in the preprocessing stage are thus assumed to have only marginally affected the qualitative aspects of the comparison with authoritative data. Nonetheless, it could be that Twitter users in different parts of the country show different usage patterns and were therefore *selectively* filtered during the preprocessing stage. However, this biased selection is likely to stay similar if the filtering criteria are only slightly adjusted. For instance, during the visual inspection of profiles in Section 4.3.3 it was detected that users in the French-speaking part of Switzerland, especially in the metropolitan region of Genf-Lausanne, appear to be more prolific, and thus have a higher chance of being included in the analysis than users who only seldom broadcast Tweets. However, slightly adjusting the threshold for the minimal number of events per object would not really have changed this imbalance. Either way, any analysis working with such data and trying to infer patterns only from users fulfilling certain criteria is prone to that sort of uncertainty and, possibly, bias.

6.1.2. Spatial and Temporal Discretization of Tweets

Another type of uncertainty arises from the discretization of the continuous, physical properties associated with each Tweet into semantic, not explicitly spatial information (Section 4.3.5). Firstly, a different aggregation, both spatially and temporally, would have led to different entropy values and thus to potentially different — and possibly valid — classifications of types of spatiotemporal patterns. This phenomenon can again be summarized under the name of the previously introduced MAUP and MTUP, respectively (Openshaw, 1983; Çöltekin et al., 2011). Namely, a more coarse discretization, e.g., into districts instead of municipalities or into 4-hour-blocks, respectively, would have smoothened the patterns and might have led to the discovery of regularities where there are actually none. Contrarily, a more fine-grained discretization would have made it more difficult to distinguish signal from noise.

Secondly, imposing a predefined *spatial* configuration onto a continuous phenomenon might not accurately reflect the structure of the underlying data. A semantic place in the sense of Section 2.2.2 might span multiple municipalities (see for example Figure 4.15d), or might not be clearly attributable to one municipality due to uncertainties in the data, e.g., stemming from inaccurate GPS readings. However, as the data needed to be spatially and temporally discretized in order to be compared with official sources, and as those uncertainties are assumed to play a minor role only, they were not explicitly addressed.

Another approach would be to use a regular grid to spatially discretize Tweets, as proposed by Morzy (2007) for conventional location data. This would have the advantage that all spatial entities would have exactly the same size; however, the Tweets could not be analyzed with regards to socio-demographic data, as such data is normally only available on the level of political entities.

6.1.3. Sensitivity of Extracting Spatiotemporal Routines

The procedure of finding semantic places of users through a rule-based heuristic (Section 4.4.3) is the main data processing step of this thesis, and the uncertainties associated with it probably affected the outcome of the analysis more profoundly than any other, previously mentioned points. The results of the entropy analysis generally showed that an underlying regularity can be found in the data of most Twitter users (Section 5.2). This is in accordance with findings gained with other kinds of (episodic) movement data (Gonzalez et al., 2008; Song et al., 2010; Cranshaw et al., 2010). Moreover, the visual inspection of spatiotemporal bevahior reaffirmed that the routines of the large majority of users have more or less the same structure. Therefore, it made sense to use a simple, rule-based heuristic to extract frequently visited, function-bearing municipalities, and to use regularity measures in the form of entropy values and schedule sizes as attributes for these rules. In fact, a study concerned with mobile phone logs used a very similar rule-based method to classify "home" and "work" places, and achieved a high classification accuracy[1] (Wang et al., 2012). The authors also found that sophisticated machine learning techniques, which took into account individual properties of users, were not able to surpass the accuracy of the rule-based method. Other plausible approaches to extract semantic places from georeferenced Tweets have been recently discussed in literature, but are not fully automatic like the approach proposed here as they incorporate visual analytics (G. Andrienko et al., 2013a, 2013b).

On the other hand, there exists no way of actually verifying the results of this process without taking a lot of effort, for instance, through textual analysis of Tweets, or by conducting interviews with users. Without such measures, it is almost impossible to judge which configurations of rules or thresholds most accurately reflect the "real" spatiotemporal behavior of users. Thus, the thresholds for the rules used in the heuristic are uncertain in the sense that they were manually and rather arbitrarily chosen. Table 6.1 redefines the thresholds used in the heuristic and also explains other symbols for clarity. Table 6.2 shows the different scenarios that will be compared below (A is the actually applied scenario). As was mentioned in Section 4.4.3, no residential or occupational municipalities were defined for objects who did not fulfill the thresholds in non-occupational or occupational time, respectively.

Setting these values too low or too high could significantly affect the extraction of municipalities. For example, if the schedule sizes $|E_{o,t=non-occ.,occ.}(R = \{...\})|$ were set too low, municipalities that are not significant or meaningful would be qualified as residential or occupational (*false positives*). On the other hand, if they were set too high, a large number of residential and occupational municipalities would not be recognized as such (*false negatives*). Furthermore, the choice of the threshold for detecting a second occupational municipality, $m^{occ}_{o,2}$, might severely affect the results — if set too low, randomly visited locations (noise) might be classified as $m^{occ}_{o,2}$, if set too high, actual places of occupation might be wrongly discarded.

A second dimension of uncertainty is associated with the delimitation of both non-occupational and occupational time. As not all people have the same

[1] The authors were in possession of a test set which allowed them to validate their results.

Table 6.1: *Redefinition of thresholds and symbols used in the extraction of spatiotemporal routines (compare with Section 4.4.3).*

Formal definition or symbol	Abbreviation	Definition
$\lvert \mathbf{E}_{\mathbf{o,t=non-occ.}} \rvert$	α	Total number of events posted during non-occupational time.
$\lvert \mathbf{E}_{\mathbf{o,t=non-occ.}}(\mathbf{R} = \{\mathbf{H}, \mathbf{D}\}) \rvert$	β	Number of events posted during different hour-day-combinations (non-occupational time).
$\lvert \mathbf{E}_{\mathbf{o,t=occ.}} \rvert$	δ	Total number of events posted during occupational time.
$\lvert \mathbf{E}_{\mathbf{o,t=occ.}}(\mathbf{R} = \{\mathbf{H}, \mathbf{D}\}) \rvert$	ϵ	Number of events posted during different hour-day-combinations (occupational time).
	γ	If the second most visited location during occupational time differed by a share of more than γ from the most visited location (i.e., the most visited location amounted to 70% of all locations, and the second most visited location to less than $70 - \gamma$%), it was also extracted as occupational municipality ($m^{occ}_{o,2}$).
$\mathbf{m}^{\mathbf{occ}}_{\mathbf{o,1}}$		First occupational municipality, as detected by the heuristic.
$\mathbf{m}^{\mathbf{occ}}_{\mathbf{o,2}}$		Second occupational municipality, as detected by the heuristic.
$\mathbf{m}^{\mathbf{res}}_{\mathbf{o}}$		Residential (non-occupational) municipality, as detected by the heuristic.

Table 6.2: *Different scenarios for extracting spatiotemporal routines. In order for a residential municipality to be extracted, thresholds α and β need to be reached. In order for at least one occupational municipality to be extracted, thresholds δ and ϵ need to be reached.*

	A	B	C	D	E
$\boldsymbol{\alpha} \geq$	30	1	80	30	30
$\boldsymbol{\beta} \geq$	15	1	40	15	15
$\boldsymbol{\delta} \geq$	30	1	60	30	30
$\boldsymbol{\epsilon} \geq$	15	1	30	15	15
$\boldsymbol{\gamma}$	30%	10%	50%	30%	30%
Definition of occupational time	Weekdays from 08:00 to 17:00	Weekdays from 08:00 to 17:00	Weekdays from 08:00 to 17:00	Weekdays from 09:00 to 16:00	Weekdays from 07:00 to 18:00
Remarks	Actually applied scenario	Very relaxed	Very strict	Smaller frame for occupational time	Larger frame for occupational time

Table 6.3: *Effects of different scenarios for extracting spatiotemporal routines on the number of detected municipalities. The figures show the number of objects (in percent of all objects) for which certain types of municipalities were detected (N=2,195 objects).*

Scenario	$\mathbf{m_o^{res}}$ detected	$\mathbf{m_{o,1}^{occ}}$ detected	$\mathbf{m_{o,2}^{occ}}$ detected
A	2,178 (99.2%)	1,665 (75.9%)	242 (11.0%)
B	2,195 (100.0%)	2,193 (99.9%)	1235 (56.3%)
C	1,645 (74.9%)	927 (42.2%)	0 (0.0%)
D	2,178 (99.2%)	1,665 (75.9%)	242 (11.0%)
E	2,178 (99.2%)	1,665 (75.9%)	242 (11.0%)

working hours, it must be asked how different temporal delimitations would affect the results. There is actual evidence about when most people usually work, given through Figure 3.2 and other information in Section 3.1.2. Therefore, a slightly shorter (scenario *D*) and a slightly longer (scenario *E*) work day were considered, too (Table 6.2). People working during the weekend or in night shifts are not accounted for by the heuristic, as they constitute a small minority[2].

In summary, based on how these thresholds are chosen, different numbers of non-occupational and occupational municipalities are detected (Table 6.3). In the applied scenario *A*, an m_o^{res} is detected for almost all objects, and an $m_{o,1}^{occ}$ for most, while an $m_{o,2}^{occ}$ is only detected for very few objects. This is more or less in accordance with the fact that only a small minority of people work *and* go to school or work at *two* different places. Objects without an m_o^{res} could not be used for the evaluation of representativeness and those without a known $m_{o,i}^{occ}$ could not be used for the calculation and comparison of commuter balances. If the criteria were heavily relaxed (scenario *B*), all three kinds of municipalities would be detected for the majority of the objects, whereas if the criteria were considerably stricter (scenario *C*), an occupational municipality would be found for less than half of the objects. Thus, if this scenario were applied, both the evaluation of representativeness and the comparison of commuter balances would be more difficult because of less support. Different frames for occupational time would not influence the detection of semantic places, as it seems, at least not quantitatively, which confirms the robustness of the chosen delimitation.

If different scenarios were applied to the evaluation of representativeness as presented in Section 5.3, only minor differences could be observed. For example, the correlation between population aged 15–64 and residential Twitter users of municipalities would barely be different (Table 6.4). The only notable differ-

[2] 5.5% of the permanent residents normally work in night shifts, and 8.3% do this only from time to time (BFS, 2012a).

	A	B	C	D	E
N municipalities	515	515	430	515	515
r^2	0.38	0.38	0.33	0.38	0.38

Table 6.4: *Effects of different scenarios for extracting spatiotemporal routines on the relationship between actual population aged 15–64 and the number of residential Twitter users per municipality. Compare with Figure 5.10.*

ence lies in scenario C, where the correlation would be slightly weaker (Figure F.1 in Appendix F). This can be attributed to the fact that an m_o^{res} would be detected for fewer objects, which, in turn, would result in more municipalities having fewer or no residential Twitter users and thus more variance in terms of actual population, as was discussed with the help of Figure 5.10. Regarding municipalities aggregated according to spatial divisions (Section 5.3.2), the qualitative comparison between scenarios A and B as well as between scenarios A and D/E does not reveal significant or even perceivable differences. Again, only between scenarios A and C some uncommonalities could be detected for linguistic regions and metropolitan areas, although these would likely not be statistically significant (Figure F.2 in Appendix F).

Additionally, the five different scenarios were applied to the actual use case of the thesis, the comparison of commuter balances, cb_i, on the individual district level (Section 5.4.1, for a definition of cb_i see Section 4.4.5). Table 6.5 shows the values of indicators for the association between authoritative and Twitter data for all scenarios (compare with Figure 5.24 and Table 5.1). First of all, the robustness of the time frame for occupational time can definitely be confirmed, as slightly larger or smaller work days would not affect the qualitative outcome of the comparison *at all*. Secondly, the same relationship between n_i and r as well as ρ could be observed in all scenarios, i.e., as one raises n_i, r and ρ are increased monotonically. Thirdly, both scenarios B and C would result in slightly lower values for r and ρ, but the differences would likely lie within statistical error. The difference would be stronger in scenario B, and a glance at the distribution of cb_i per district as seen on Twitter reveals that there would be quite a few districts with extremely positive values (Figure F.3 in Appendix F). The reason for this might be the unrealistically frequent extraction of $m_{o,2}^{occ}$. These additional occupational municipalities would likely distort the computation of commuter balances and allow for such outliers. One can thus assume that the

Table 6.5: *Effects of different scenarios for extracting spatiotemporal routines on the association of commuter balances, cb_i, between authoritative and Twitter data, as measured through Spearman's ρ and Pearson's correlation coefficient. n_i denotes the number of residential Twitter users per district. Compare with Figure 5.24 and Table 5.1.*

$\mathbf{n_i} \geq$	Scenario	N districts	ρ	r
$-\infty$	A	90	0.41	0.20
	B	90	0.35	0.13
	C	90	0.44	0.36
	D / E	90	0.41	0.20
10	A	41	0.65	0.62
	B	41	0.50	0.39
	C	29	0.64	0.60
	D / E	41	0.65	0.62
15	A	28	0.76	0.73
	B	28	0.68	0.52
	C	19	0.78	0.69
	D / E	28	0.76	0.73
20	A	21	0.84	0.79
	B	21	0.87	0.77
	C	16	0.93	0.77
	D / E	90	0.84	0.79

more relaxed rule for detecting the second occupational municipality (Table 6.2) would be responsible for this deviation from A. Namely, the second-most visited municipality would often be automatically declared as $m^{occ}_{o,2}$ for objects who have visited more than one municipality during occupational time.

In conclusion, it appears that the extraction of spatiotemporal routines produces mostly robust results under the variation of rule thresholds. The particular choice of thresholds for the actually applied scenario A proved to be appropriate in that not too many second occupational municipalities, which could have distorted the computation of commuter balances, were extracted. As already mentioned, the goal of this process was not to find an optimized configuration of parameters which approximates the authoritative data as close as possible. Rather, it was shown that, even though extracted semantic places could not be directly validated, a simple heuristic achieved to find an underlying, spatiotemporal structure in geotagged Tweets — a structure which persists even when the parameters are brought to their extremes.

6.2. Reflections on the Research Questions

In this section, the research questions asked in Section 1.3 are put into the context of the results and the state of the art. It becomes apparent that, while **RQ1** and **RQ2** can be fully answered and discussed with the obtained results, **RQ3** would require a more sophisticated analysis and, particularly, a greater data support in order to be exhaustively answered.

6.2.1. Spatial and Socio-Demographic Representativeness

The inference of socio-demographic properties based on the membership to a certain type of municipality naturally suffers from a range of statistical biases and uncertainties. First of all, inferences about the nature of *individuals* based on inferences about the group these individuals belong to are subject to *ecological fallacy* (Piantadosi et al., 1988). In other words, one cannot strictly assume that the properties of the group (of municipalities) account for each and every individual, not even for the majority of individuals. Secondly, spatially aggregated statistical properties are always subject to the MAUP as introduced in Section 2.3. As an example in the context of this thesis, one can think of a municipality belonging to the type "high income", while its neighboring municipality was classified as "agrarian". Nonetheless it is possible that, for instance, a rich neighborhood stretches across both municipalities. It is therefore difficult to directly derive socio-demographic properties from the spatial distribution of Twitter users, nonetheless, the results allow to gain a coarse impression thereof and help to answer the first research question:

RQ1 *How representative is the spatial and socio-demographic distribution of users found in D_{vgi} of the overall population as measured in D_{as}?*

From the results presented in Section 5.3, it can clearly be stated that users of Twitter who frequently geotag their Tweets are *not* regionally representative of the actual population. Contrarily, one could describe a "typical", geoactive Twitter user as having an overproportionally high probability of living in the French-speaking part of Switzerland, living in an urban area, namely in a large or medium-sized center, and living in a municipality of a metropolitan area. On the other hand, it is less likely to detect a user who dwells in the German-speaking region or lives in the rural, namely in the industrial and agrarian, part of Switzerland. It was also found that, while the population is equally represented in terms of gender, young people (aged 15–30) strongly prevail, although there are regional differences. As was mentioned in the introduction, other socio-demographic dimensions such as education, economic status, and culture cannot be directly inferred from the Twitter data. On the other hand, differences in terms of those between urban and rural areas as well as between linguistic regions of Switzerland are a reality (Büchi, 2003; Leuthold et al., 2007; Brügger et al., 2009). Therefore, the spatially unequal distribution of Twitter users certainly indicates that the population is also unequally represented in terms of those dimensions.

Moreover, it has to be assumed that, in Switzerland, Twitter is still only at the brink of leaving its early adopter phase. As a very recently published study with N=1,114 subjects shows, only 18% of Swiss Internet users have an account, of whom roughly a third actively uses it, as compared to the 58% who participate in other social networks (Latzer et al., 2013). The early adopter phase of a new technology is usually dominated by a mostly urban, well-educated, tech-savvy and male cohort (Lipsman, 2009), and the visual inspection of profiles in Section 4.3.3 partially confirmed this. On the other hand, it was also found that, in the French-speaking part of Switzerland, users are often in their teens and appear to be very prolific, which could be a reason for the Genève-Lausanne metropolitan region being so highly overrepresented.

In summary, the willingness to geotag content or, possibly, to participate on Twitter at all seems to vary from one type of region to another. This can be explained through socio-demographic and thus spatial disparities, since that motivation is certainly determined by someone's social, cultural and economic background (Section 2.1.3). With regards to the state of the art, it was confirmed that authoritative, demographic data may be used as a framework for assessing the properties of content producers, as it was done by L. Li et al. (2013) and Kent and Capello (2013). However, as "participation inequality" (J. Nielsen, 2006; Ochoa & Duval, 2008; Haklay, 2012) is now a proven fact for geotagged content, too, the above presented approaches of aggregating mere content are subject to severe bias caused by prolific users.

Although the claims of Crampton et al. (2013), saying that user-generated data is produced by a wealthy, more educated, and more male demographic, could not be directly replicated, it can certainly be stated that there are certain tendencies in terms of the socio-economic and cultural background of users. However, making statements about society may still be possible, as the validation of the inferred commuter balances showed (see next section).

6.2.2. The Inference of Commuter Balances as a Use Case for VGI

RQ2 *How do patterns of intra- and interregional mobility as inferred from D_{vgi} compare to mobility quantifications found in D_{as}?*

The results of the comparison of individual districts show a moderately strong and significant linear correlation with official data, but only if the support of VGI is high enough ($r^2 = 0.39$ for districts with more than 10 residential Twitter users, $r^2 = 0.62$ for districts with more than 20 residential Twitter users). In particular, the more Twitter users are found to reside in a district, the more the commuter balance of this district is likely to be correlated with its authoritative counterpart. This phenomenon can possibly be explained from two perspectives. It may mean that the correlation is biased towards districts with a high population count (as these intuitively have more users in D_{vgi}) and that it is not existent in other districts, independent of the number of residential Twitter users found in them. Or, it may mean that the correlation between D_{vgi} and D_{as} is valid for all districts, i.e., for the whole study area, but can only be detected for districts with enough support.

Even though a definite answer can not be given to this question, the strength of the correlation is still remarkable if one considers that the commuter balances were extracted through a simple, unvalidated heuristic. This certainly shows that working with episodic movement is indeed possible (N. Andrienko et al., 2012), but more research is needed in terms of how the ultimate results of such methods can be validated. Moreover, it has to be recounted that the demographics of Twitter are shifted towards a younger cohort than that of the majority of the working population in Switzerland (Section 3.3.1 and 4.3.3). If authoritative commuter balances were available exclusively for people aged 20–40, one could investigate whether the correlation would be even stronger. However, one can not directly derive actual commuter balances from those found in D_{vgi}, because those are generally less strongly pronounced than the authoritative balances and sometimes have the wrong algebraic sign. Thus, in order to estimate actual balances from those seen in D_{vgi}, a linear model could be learned from the data. By doing so, one could — with the given data — vaguely estimate the actual commuter balances of about 40, densely populated districts. It was also shown that, when districts are aggregated to cantons, the approximate balance and the rank of most cantons can be estimated, although here, too, cantons with a low support in D_{vgi} should be excluded from the estimation.

A closer look at the biggest agglomerations showed that, while balances are generally in accordance with those found in D_{as}, staying commuters are strongly overrepresented. It is assumed that this is due to the way the heuristic for the extraction of semantic places was implemented. Namely, the actual residential municipality was often also declared as first or second occupational municipality, even though the user does not have an occupation there. This may happen because people stay at home during occupational time, for instance because they are on sick leave or have holidays, or, for instance, because people return early from work and post Tweets which are then still classified as having been sent during occupational time. Either way, it would require a more sophisticated extraction algorithm to account for such cases. Not only are staying commuters overrepresented, incoming and outgoing commuters are underrepresented when compared to the number of residential Twitter users, n_i. Since n_i is used as the denominator for computing balances, its relative overestimation explains the generally smaller range of commuter balances as compared to official figures. To compensate for this, one could either equally increase the number of both incoming and outgoing commuters or decrease the number of residential Twitter users per district. In doing so, the resulting commuter balances would be stretched and thus be closer to authoritative figures, although the strength of the correlation would not change.

In literature, attempts to use egocentric, geosocial VGI to quantify human mobility are still very sparse, and their results are rarely evaluated with an authoritative benchmark (Girardin et al., 2008; Ferrari et al., 2011; Aubrecht et al., 2011). The results obtained here clearly show that an evaluation with the help of official sources is indeed possible. In this vein, the findings gained here are a step towards a better understanding of the merits and pitfalls of VGI from social media, as requested by Purves (2011) as well as Sui and Goodchild (2011).

RQ3 *How does spatial and socio-demographic representativeness as measured in* RQ1 *influence the results of* RQ2*?*

While the first two research questions can mostly be answered with the obtained results, the third would require another analysis if taken literally. In order to measure whether equally represented spatial regions actually led to an even stronger correlation between D_{vgi} and D_{as} in terms of commuter balances, one would have to construct a model that corrects for unequal representation. Such a model would be quite complex, since various dimensions of representativeness would have to be incorporated. Another approach would be to compare commuter balances separately for different spatial divisions, e.g., it could be analyzed whether urban districts are more strongly correlated with official data than rural districts. Since the data are already too sparse to compare districts globally, however, such an endeavor would need considerably more data.

The direct influence of the unequal regional representativeness on the extracted commuter balances is thus difficult to grasp, however, evidence thereof could still be detected. For example, in Figure 5.28 it could be seen that the agglomeration of Genève has an unnaturally high number of staying commuters, a fact which can definitely be attributed to the general strong overrepresentation of this region. Secondly, the comparison of commuter flows between and within urban and rural areas showed that flows originating or ending in core cities are overrepresented, while such originating and ending in rural areas are underrepresented. The reason for this might quite possibly be the generally unequal representation of these types of settlements.

In conclusion, it can probably be assumed that the unequal representation of different types of regions influences the correct inference of commuter balances in some way. However, restrictions of the available data, both qualitatively and quantitatively, render such an analysis very difficult. A higher priority, anyway, should be to have a more definite answer on **RQ2** before starting to inquire **RQ3**.

7. Conclusion

This work set out to explore how georeferenced data in the form of egocentric, geosocial VGI can be used to complement or even replace authoritative data. The study of human mobility, in this case expressed through commuter balances, is an interesting and promising application for such analyses. While the research community has been very enthusiastic about exploring use cases for VGI, studies often take such data "as is" and do not question it with regards to representativeness. Even more importantly, most, if not all, studies that were reviewed in this thesis focus on the geotagged content itself but ignore the producers of the content, whose power to bias the results of any analysis is often underestimated. The motivation behind this thesis was thus not to promote yet another way of merely aggregating chunks of geotagged content, but to thoroughly investigate the properties, limitations, and pitfalls of such data as a potential source of input to geographical analyses.

7.1. Achievements

The following list gives a detailed overview of what was achieved in this work:

- An extensive literature review was conducted in order to identify the shortcomings of studies working with egocentric, geosocial VGI, and the main research gaps were outlined.
- Through a simple, yet effective data collection procedure, almost 25,000 (geo-)active Twitter users and over 12 million geotagged Tweets were gathered over the course of a year. To the author's knowledge, only very few studies in this field of research exist which continuously tracked producers of VGI over such an extended timespan and gathered so much individually referenceable data.
- A detailed exploratory analysis was conducted to investigate the properties of the collected data, and particularly, of the tracked users. In or-

der to assure that these users were indeed humans and originated from the study area, desirability criteria were formulated and operationalized through spatial and temporal indicator variables.

- The desirability criteria were together used as rules to detect and discard undesirable users. The visual inspection of a representative sample showed that the remaining users conformed with the formulated criteria.

- The participation inequality phenomenon commonly found in social media could be confirmed for geotagged content on Twitter.

- Based on a visual inspection, it was confirmed that a specific spatiotemporal pattern can be found in the Tweet history of most users. With the help of the concept of information entropy, it was also shown that the large majority of users visits the same few places with high regularity.

- This thesis is among the first studies to make inferences from individual producers of content and not from the apparently biased content itself. Namely, through a rule-based heuristic, residential and occupational municipalities of each user were detected and extracted.

- Having obtained residential municipalities of Twitter users allowed to compare the data with official population data, both on the individual municipal level and aggregated according to geodemographic divisions. The latter was used to assess which regions are over- and underrepresented among geoactive Twitter users. To the author's knowledge, this is the first study to look at small-scale regional representativeness of geosocial, egocentric VGI.

- Knowledge about residential and occupational municipalities was used to compute commuter balances, which were compared with authoritative data, both on the individual district level and in aggregated form. The results confirmed that commuter balances, as a use case for the study of human mobility, can be extracted from VGI in a format so that they can be compared with authoritative data. To the author's knowledge, such a thorough, quantitative evaluation of an analysis done with VGI has not been conducted before.

7.2. Insights

Twitter as a provider of egocentric, geosocial VGI is a complex source of data associated with many uncertainties. For instance, the data is likely to be biased towards very prolific users, and the envisaged origin and authenticity of content producers cannot be taken for granted, as it can be in traditional surveys. In this vein, it was shown that the localization of users through geotagged Tweets is only possible if a lot of information has been gathered — merely searching for Tweets posted within a specified area does not automatically return users who also dwell in this area. In order to derive meaningful insights from egocentric, geosocial VGI, the data must thus be carefully preprocessed, which is likely to result in the retention of only a small fraction of what was initially gathered. As geotagging is still not very popular amongst users of social media, most, if not all, analyses seeking to make inferences from it are thus prone to data sparsity, even though the enthusiastic research community often claims the opposite.

Another problem associated with egocentric, geosocial VGI is representativeness. While it is quite intuitive that demographic, e.g., age-related, representativeness is not given, this thesis has shown that, also in terms of geographical regions, the population is highly unequally represented. Even for a relatively small, densely populated country like Switzerland, it must be assumed that people living in different geodemographic regions are not equally willing to geotag content, or to produce content at all. Certainly, it has to be admitted that the overrepresentation of urban areas, especially of center cities, is not so much of a surprise. The fact that there are stark differences in terms of linguistic and thus, to a certain degree, cultural representation, is quite striking, however. Moreover, the analyzed data indicates that in different linguistic regions, VGI is produced by different socio-demographic groups. This implies that, depending of how the data is preprocessed and filtered, different socio-demographic groups may be more or less prominently represented. This may also have biased the analysis conducted here, but must be accepted as an inevitable characteristic of such data. Hereby it should also be noted that the findings in terms of representativeness may or may not be generalized to Twitter users in general, as only *geoactive* users were the subject of this study. It is quite possible that geotagging is an even more unequally represented phenomenon than general Twitter usage, especially in terms of cultural differences.

While the finding that different regions and socio-demographic cohorts are unequally represented most certainly holds true in the rest of the world, it should be noted that these imbalances probably differ from place to place.

For instance, while in Switzerland, large centers are grossly overrepresented, this may not be the case in the United States, where Twitter is generally more popular. A major insight of this work is therefore that Twitter and likely any other provider of egocentric, geosocial VGI, is used in a different fashion and by different socio-demographic groups in different geographical spaces. Every geographical analysis should thus start with a thorough assessment of small-scale representativeness and incorporate these findings into the interpretation of the obtained results.

The second major insight of this thesis is that, if high-quality authoritative data are available, they can and should be used to validate inferences made from VGI. In this case, they were used to show that, in theory, reasonable indicators for human mobility can be inferred from Twitter data, even though spatial and socio-demographic representativeness can not be assumed. For some use cases, such as the inference of commuter balances, one may not even need socio-demographically representative data to gain a coarse impression of reality, and certain dimensions of representativeness might be more important than others. However, as the inferred results cannot be used to directly replace authoritative data — at least not at this stage— the value of VGI for such analyses should not be overestimated. The main problem appears to be data sparsity, and it might be possible that, with more data, significant correlations could have been found for the whole study area. More importantly, more data would allow to better investigate the influence of unequally represented regions on the outcome of the comparison.

In conclusion, while egocentric, geosocial VGI seems to be an interesting data source for all kinds of geographical analyses, it needs to be treated with a lot of care when representative conclusions are to be drawn. Even when authoritative data are available as reference, analyses are likely to suffer from data sparsity and uncertainties which can not ultimately be avoided but must be handled with appropriate measures.

7.3. Future Work

Possible future research can be divided into three main categories. First of all, more research is needed to address the questions that could not be answered in this thesis. For instance, in order to better understand regionally different usage contexts and motivations for geotagging, one could take regional samples — made possible through the inference of residential municipalities — and try to infer more information about socio-demographic properties, e.g., through inspection of Twitter profiles or through (semi-)automatic analysis of textual Tweet content. In general, it should be more deeply investigated how exactly spatial and socio-demographic representativeness influences results of (geographical) analyses. At the same time, other applications than the study of human mobility should be considered, too.

Secondly, it should be assessed whether the results obtained here can be reproduced in other usage contexts, i.e., in other regions and with other sources of VGI. Most importantly, it should be tried to run the same or a slightly different analysis with a higher support of data, e.g., through more sophisticated data collection methods or by looking at a region or country where Twitter is more popular. It could also be tried to enhance the Twitter data with such from other social networks such as Facebook or Foursquare, although this would likely complicate the analysis.

Lastly, episodic movement data, towards which egocentric, geosocial VGI can be counted, are associated with a range of additional uncertainties that greatly limit the possibilities of certain types of analysis. More research is thus also needed in terms of new, privacy-preserving techniques for analyzing such data, so that future studies can resort to established methods. On the other hand, it should have become clear by now that analyzing VGI from sources such as Twitter always requires highly specialized procedures, and that the results of such analyses are thus difficult to compare with each other. The establishment of generalizable research frameworks and "best practices" for working with such data would therefore certainly benefit further endeavors.

Bibliography

Adams, P. M., Ashwell, G. W. B., & Baxter, R. (2003). Location-based services – An overview of the standards. *BT Technology Journal*, *21*, 34–43.

Allamanis, M., Scellato, S., & Mascolo, C. (2012). Evolution of a location-based online social network: Analysis and models. In *Proceedings of the 2012 ACM conference on internet measurement* (145–158).

Andersson, M., Gudmundsson, J., Laube, P., & Wolle, T. (2008). Reporting leaders and followers among trajectories of moving point objects. *GeoInformatica*, *12*, 497–528.

Andrienko, G., Andrienko, N., Bosch, H., Ertl, T., Fuchs, G., Jankowski, P., & Thom, D. (2013a). Thematic patterns in georeferenced tweets through space-time visual analytics. *Computing in Science & Engineering*, *15*, 72–82.

Andrienko, G., Andrienko, N., Demsar, U., Dransch, D., Dykes, J., Fabrikant, S. I., … Tominski, C. (2010). Space, time and visual analytics. *International Journal of Geographical Information Science*, *24*, 1577–1600.

Andrienko, G., Andrienko, N., Fuchs, G., Raimond, A.-M. O., Symanzik, J., & Ziemlicki, C. (2013b). Extracting semantics of individual places from movement data by analyzing temporal patterns of visits. In *Proceedings of the first ACM SIGSPATIAL international workshop on computational models of place (COMP'13)*.

Andrienko, N., Andrienko, G., Pelekis, N., & Spaccapietra, S. (2008). Basic concepts of movement data. In *Mobility, data mining and privacy* (15–38). Springer.

Andrienko, N., Andrienko, G., Stange, H., Liebig, T., & Hecker, D. (2012). Visual analytics for understanding spatial situations from episodic movement data. *KI-Künstliche Intelligenz*, *26*, 241–251.

Aubrecht, C., Ungar, J., & Freire, S. (2011). Exploring the potential of volunteered geographic information for modeling spatio-temporal characteristics of urban population. In *Proceedings of 7VCT* (13–16).

Barkhuus, L., Brown, B., Bell, M., Sherwood, S., Hall, M., & Chalmers, M. (2008). From awareness to repartee: Sharing location within social groups. In *Proceedings of the SIGCHI conference on human factors in computing systems* (497–506).

Beaumont, C. (2010, February 23). Twitter users send 50 million Tweets per day [The Telegraph]. Retrieved July 19, 2013, from http://www.telegraph.co.uk/technology/twitter/7297541/Twitter-users-send-50-million-tweets-per-day.html

Beevolve. (2012, October 10). An exhaustive study of twitter users across the world. Retrieved July 19, 2013, from http://www.beevolve.com/twitter-statistics/

Berry, B. J. (1961). City size distributions and economic development. *Economic Development and Cultural Change*, *4*, 573–588.

BFS. (2008). *Data collection programme of the federal census*. Bundesamt für Statistik BFS. Neuchâtel.

BFS. (2010). *Ständige Wohnbevölkerung ab 15 Jahren nach Hauptsprache 2010*. Bundesamt für Statistik BFS. Neuchâtel.

BFS. (2011). *Raumgliederungen der Schweiz 2011*. Bundesamt für Statistik BFS. Neuchâtel.

BFS. (2012a). *Abend-, Nachtarbeit nach Geschlecht, Nationalität, Altersgruppen, Familientyp, 1991–2012*. Bundesamt für Statistik BFS. Neuchâtel.

BFS. (2012b). *Generalisierte Gemeindegrenzen der Schweiz, GEOSTAT-Datenbeschreibung*. Bundesamt für Statistik BFS. Neuchâtel.

BFS. (2012c). *Ständige Wohnbevölkerung nach Alter, Kanton, Bezirk und Gemeinde 2010–2012*. Bundesamt für Statistik BFS. Neuchâtel.

BFS. (2012d). *Wöchentliche Normalarbeitszeit der Vollzeitarbeitnehmenden nach Geschlecht, Nationalität und Wirtschaftsabschnitten, 1991–2012*. Bundesamt für Statistik BFS. Neuchâtel.

BFS. (2013a). *Beschäftigungsgrad nach Geschlecht, Nationalität, Altersgruppen, Familientyp 1991–2013*. Bundesamt für Statistik BFS. Neuchâtel.

BFS. (2013b). *Internetnutzung in der Schweiz 1997–2013*. Bundesamt für Statistik BFS. Neuchâtel.

BFS. (2013c). *Mobility and transport - pocket statistics 2013*. Bundesamt für Statistik BFS. Neuchâtel.

BFS. (2013d). *Pendler 2010*. Bundesamt für Statistik BFS. Neuchâtel.

BFS. (2013e). *Pendlermobilität in der Schweiz 2011*. Bundesamt für Statistik BFS. Neuchâtel.

Bhaduri, B., Bright, E., Coleman, P., & Urban, M. L. (2007). LandScan USA: A high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal*, *69*, 103–117.

Blau, I. & Neuthal, T. (2012). Twitter as a platform for an Israeli community of information science professionals. *Issues in Informing Science & Information Technology*, *9*, 177–186.

Bollen, J., Pepe, A., & Mao, H. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of the 5th international AAAI conference on weblogs and social media* (450–453).

Brockmann, D., Hufnagel, L., & Geisel, T. (2006). The scaling laws of human travel. *Nature*, *439*, 462–465.

Brownlee, J. (2012, March 30). This creepy app isn't just stalking women without their knowledge, it's a wake-up call about Facebook privacy [update] [Cult of Mac]. Retrieved July 18, 2013, from http://www.cultofmac.com/157641/this-creepy-app-isnt-just-stalking-women-without-their-knowledge-its-a-wake-up-call-about-facebook-privacy/

Brügger, B., Lalive, R., & Zweimüller, J. (2009). Does culture affect unemployment? Evidence from the Röstigraben. *CESifo Working Paper Series*. 2714th ser.

Büchi, C. (2003). *Röstigraben: Das Verhältnis zwischen deutscher und französischer Schweiz – Geschichte und Perspektiven*. Verlag Neue Zürcher Zeitung.

Catt, R. D. (2009, February 4). 100,000,000 geotagged photos (plus) [Flickr Code Blog]. Retrieved July 18, 2013, from http://code.flickr.net/2009/02/04/100000000-geotagged-photos-plus/

Cheng, A., Evans, M., & Singh, H. (2009). *Inside Twitter: An in-depth look inside the Twitter world*. Sysomos Inc.

Cheng, Y.-C., Chawathe, Y., LaMarca, A., & Krumm, J. (2005). Accuracy characterization for metropolitan-scale wi-fi localization. In *Proceedings of the 3rd international conference on mobile systems, applications, and services* (233–245).

Cheng, Z., Caverlee, J., & Lee, K. (2010). You are where you tweet: A content-based approach to geo-locating Twitter users. In *Proceedings of the 19th ACM international conference on information and knowledge management* (759–768).

Cheng, Z., Caverlee, J., Lee, K., & Sui, D. Z. (2011). Exploring millions of footprints in location sharing services. In *Proceedings of the 5th international AAAI conference on weblogs and social media* (81–88).

Cho, E., Myers, S. A., & Leskovec, J. (2011). Friendship and mobility: User movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining* (1082–1090).

Cliff, A. D. & Ord, K. (1970). Spatial autocorrelation: A review of existing and new measures with applications. *Economic Geography*, *46*, 269–292.

Cöltekin, A., De Sabbata, S., Willi, C., Vontobel, I., Pfister, S., Kuhn, M., & Lacayo, M. (2011). Modifiable temporal unit problem. In *Proceedings of the ISPRS/ICA workshop "Persistent problems in geographic visualization"*.

Cöltekin, A., Heil, B., Garlandini, S., & Fabrikant, S. I. (2009). Evaluating the effectiveness of interactive map interface designs: a case study integrating usability metrics with eye-movement analysis. *Cartography and Geographic Information Science*, *36*, 5–17.

Cramer, H., Rost, M., & Holmquist, L. E. (2011). Performing a check-in: Emerging practices, norms and 'conflicts' in location-sharing using Foursquare. In *Proceedings of the 13th international conference on human computer interaction with mobile devices and services* (57–66).

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

Cranshaw, J., Schwartz, R., Hong, J., & Sadeh, N. (2012). The Livehoods project: Utilizing social media to understand the dynamics of a city. In *Proceedings of the 6th international AAAI conference on weblogs and social media* (58–65).

Cranshaw, J., Toch, E., Hong, J., Kittur, A., & Sadeh, N. (2010). Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM international conference on ubiquitous computing* (119–128).

Crockford, D. (2006). RFC 4627: The application/json media type for JavaScript Object Notation (JSON). Retrieved July 24, 2013, from https://tools.ietf.org/html/rfc4627

De Longueville, B., Smith, R. S., & Luraschi, G. (2009). 'OMG, from here, I can see the flames!': A use case of mining location based social networks to acquire spatio-temporal data on forest fires. In *Proceedings of the 2009 international workshop on location based social networks* (73–80).

Diakopoulos, N. A. & Shamma, D. A. (2010). Characterizing debate performance via aggregated Twitter sentiment. In *Proceedings of the 28th international conference on human factors in computing systems* (1195–1198).

Dodge, S., Weibel, R., & Lautenschütz, A.-K. (2008). Towards a taxonomy of movement patterns. *Information Visualization*, *7*, 240–252.

Edwardes, A. J. & Purves, R. S. (2007). A theoretical grounding for semantic descriptions of place. In *Web and wireless geographical information systems* (106–120). Springer.

Elwood, S., Goodchild, M. F., & Sui, D. Z. (2012). Researching volunteered geographic information: Spatial data, geographic research, and new social practice. *Annals of the Association of American Geographers*, *102*, 571–590.

Facebook Help Center. (2014). Location. Retrieved January 16, 2014, from https://www.facebook.com/help/115298751894487

Ferrari, L., Rosi, A., Mamei, M., & Zambonelli, F. (2011). Extracting urban patterns from location-based social networks. In *Proceedings of the 3rd ACM SIGSPATIAL international workshop on location-based social networks* (9–16).

Fiegerman, S. (2012, December 18). Twitter now has more than 200 million monthly active users [Mashable]. Retrieved July 17, 2013, from http://mashable.com/2012/12/18/twitter-200-million-active-users/

Fielding, R. T. (2000). *Architectural styles and the design of network-based software architectures* (Ph.D. Thesis, University of California).

Flanagin, A. J. & Metzger, M. J. (2008). The credibility of volunteered geographic information. *GeoJournal*, *72*, 137–148.

Flickr Help FAQ. (2014). Tags. Retrieved January 16, 2014, from https://secure.flickr.com/help/tags/

Freire, S., Aubrecht, C., & Wegscheider, S. (2011). Spatio-temporal population distribution and evacuation modeling for improving Tsunami risk assessment in the Lisbon metropolitan area. *Proceedings of the 2011 conference on geoinformation for disaster management*.

Friedland, G. & Sommer, R. (2010). Cybercasing the joint: On the privacy implications of geo-tagging. In *Proceedings of the fifth USENIX workshop on hot topics in security (HotSec 10)* (1–8).

Fuchs, G., Andrienko, G., Andrienko, N., & Jankowski, P. (2013). Extracting personal behavioral patterns from geo-referenced Tweets. In *Proceedings of AGILE 2013*.

Fujisaka, T., Lee, R., & Sumiya, K. (2010). Discovery of user behavior patterns from geo-tagged micro-blogs. In *Proceedings of the 4th international conference on ubiquitous information management and communication* (p. 36).

Girardin, F., Calabrese, F., Fiore, F., Ratti, C., & Blat, J. (2008). Digital footprinting: Uncovering tourists with user-generated content. *IEEE Pervasive Computing*, *7*, 36–43.

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. (2008). Understanding individual human mobility patterns. *Nature*, *453*, 779–782.

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, *69*, 211–221.

Graham, M. (2012, September 3). Big data and the end of theory? [The Guardian Datablog]. Retrieved May 7, 2013, from http://www.guardian.co.uk/news/datablog/2012/mar/09/big-data-theory

Gregory, D., Johnston, R., Pratt, G., Watts, M., & Whatmore, S. (2009). *The dictionary of human geography* (5th). Wiley-Blackwell.

Gruteser, M. & Hoh, B. (2005). On the anonymity of periodic location samples. In *Security in pervasive computing* (179–192). Springer.

Gudmundsson, J., Laube, P., & Wolle, T. (2012). Computational movement analysis. In *Springer handbook of geographic information* (pp. 423–438). Springer.

Gueye, B., Uhlig, S., & Fdida, S. (2007). Investigating the imprecision of IP block-based geolocation. In *Passive and active network measurement* (237–240). Springer.

Hägerstraand, T. (1970). What about people in regional science? *Papers in Regional Science*, *24*, 7–24.

Haklay, M. (2012, May 3). 'NOBODY wants to do council estates': Digital divide, spatial justice and outliers. Retrieved May 2, 2013, from https://povesham.wordpress.com/2012/03/05/nobody-wants-to-do-council-estates-digital-divide-spatial-justice-and-outliers-aag-2012/

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, *11*, 10–18.

Hardy, M. (2010). Pareto's law. *The Mathematical Intelligencer*, *32*, 38–43.

Harris, R., Sleight, P., & Webber, R. (2005). *Geodemographics, GIS and neighbourhood targeting*. Wiley.

Hecht, B., Hong, L., Suh, B., & Chi, E. H. (2011). Tweets from Justin Bieber's heart: The dynamics of the location field in user profiles. In *Proceedings of the 2011 annual conference on human factors in computing systems* (237–246).

Hofmann-Wellenhof, B., Lichtenegger, H., & Collins, J. (1993). *Global positioning system. Theory and practice.* Springer.

Hollenstein, L. & Purves, R. S. (2010). Exploring place through user-generated content: Using flickr tags to describe city cores. *Journal of Spatial Information Science*, 21–48.

Hornsby, K. & Egenhofer, M. J. (2002). Modeling moving objects over multiple granularities. *Annals of Mathematics and Artificial Intelligence*, *36*, 177–194.

Jedrzejczyk, L., Price, B. A., Bandara, A. K., & Nuseibeh, B. (2009). *I know what you did last summer: Risks of location data leakage in mobile and social computing*. Department of Computing Faculty of Mathematics, Computing and Technology, The Open University.

Jedrzejczyk, L., Price, B. A., Bandara, A. K., & Nuseibeh, B. (2010). On the impact of real-time feedback on users' behaviour in mobile location-sharing applications. In *Proceedings of the 6th symposium on usable privacy and security*.

Jochem, W. C., Sims, K., Bright, E. A., Urban, M. L., Rose, A. N., Coleman, P. R., & Bhaduri, B. L. (2012). Estimating traveler populations at airport and cruise terminals for population distribution and dynamics. *Natural Hazards*, 1–18.

Jordan, T., Raubal, M., Gartrell, B., & Egenhofer, M. (1998). An affordance-based model of place in GIS. In *Proceedings of the 8th int. symposium on spatial data handling* (98–109).

Joye, D., Schuler, M., Nef, R., & Bassand, M. (1988). *Typologie der Gemeinden der Schweiz: Ein systematischer Ansatz nach dem Zentren–Peripherie–Modell*. ARE. Bern.

Kent, J. D. & Capello, H. T. (2013). Spatial patterns and demographic indicators of effective social media content during the Horse Thief canyon fire of 2012. *Cartography and Geographic Information Science*, *40*, 78–89.

Kreuz, T., Chicharro, D., Andrzejak, R. G., Haas, J. S., & Abarbanel, H. D. (2009). Measuring multiple spike train synchrony. *Journal of Neuroscience Methods*, *183*, 287–299.

Krumm, J. (2007). Inference attacks on location tracks. In *Proceedings of the 5th international conference on pervasive computing* (127–143).

Krumm, J. & Rouhana, D. (2013). Placer: Semantic place labels from diary data. In *Proceedings of the 2013 ACM international joint conference on pervasive and ubiquitous computing* (163–172).

Kumar, R., Novak, J., & Tomkins, A. (2010). Structure and evolution of online social networks. In *Link mining: Models, algorithms, and applications* (337–357). Springer.

Latzer, M., Just, N., Metreveli, S., & Saurwein, F. (2013). *Internet-Anwendungen und deren Nutzung in der Schweiz 2013. Themenbericht aus dem World Internet Project – Switzerland 2013.* Universität Zürich. Zürich.

Laube, P., Imfeld, S., & Weibel, R. (2005). Discovering relative motion patterns in groups of moving point objects. *International Journal of Geographical Information Science*, *19*, 639–668.

Laube, P. & Purves, R. S. (2011). How fast is a cow? Cross-Scale analysis of movement data. *Transactions in GIS*, *15*, 401–418.

Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., & Shook, E. (2013). Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday*, *18*.

Leuthold, H., Hermann, M., & Fabrikant, S. I. (2007). Making the political landscape visible: Mapping and analyzing voting patterns in an ideological space. *ENVIRONMENT AND PLANNING B PLANNING AND DESIGN*, *34*, 785.

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

Li, Z., Ding, B., Han, J., Kays, R., & Nye, P. (2010). Mining periodic behaviors for moving objects. In *Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining* (1099–1108).

Liao, L., Fox, D., & Kautz, H. (2007a). Extracting places and activities from gps traces using hierarchical conditional random fields. *The International Journal of Robotics Research*, *26*, 119–134.

Liao, L., Patterson, D., Fox, D., & Kautz, H. (2007b). Learning and inferring transportation routines. *Artificial Intelligence*, *171*, 311–331.

Lindamood, J., Heatherly, R., Kantarcioglu, M., & Thuraisingham, B. (2009). Inferring private information using social network data. In *Proceedings of the 18th international conference on world wide web* (1145–1146).

Lindqvist, J., Cranshaw, J., Wiese, J., Hong, J., & Zimmerman, J. (2011). I'm the mayor of my house: Examining why people use Foursquare - a social-driven location sharing application. In *Proceedings of the SIGCHI conference on human factors in computing systems* (2409–2418).

Lipset, S. M. & Rokkan, S. (1967). *Party systems and voter alignments: Cross-national perspectives*. Free Press.

Lipsman, A. (2009, September 2). What Ashton vs. CNN foretold about the changing demographics of Twitter [comScore, Inc]. Retrieved July 19, 2013, from http://www.comscore.com/Insights/Blog/What_Ashton_vs._CNN_Foretold_About_the_Changing_Demographics_of_Twitter

Loibl, W. & Peters-Anders, J. (2012). Mobile phone data as source to discover spatial activity and motion patterns. In *Proceedings of the G1_Forum 2012: Geovisualization, society and learning* (524–533).

Mahmud, J., Nichols, J., & Drews, C. (2012). Where is this Tweet from? Inferring home locations of Twitter users. In *Proceedings of the 6th international AAAI conference on weblogs and social media* (511–514).

McLaren, R. & Kennedy, E. (2013). Data is the new currency in the location revolution – who will supply the data? In *Proceedings of the annual World Bank conference on land and poverty*.

McPherson, T. N. & Brown, M. J. (2003). Estimating daytime and nighttime population distributions in US cities for emergency response activities. In *Proceedings of the 84th AMS annual meeting*.

Mennis, J. & Hultgren, T. (2006). Intelligent dasymetric mapping and its application to areal interpolation. *Cartography and Geographic Information Science*, *33*, 179–194.

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. (2007). Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on internet measurement* (29–42).

Mislove, A., Viswanath, B., Gummadi, K. P., & Druschel, P. (2010). You are who you know: Inferring user profiles in online social networks. In *Proceedings of the 3rd ACM international conference on web search and data mining* (251–260).

Mohan, P., Padmanabhan, V. N., & Ramjee, R. (2008). Nericell: Rich monitoring of road and traffic conditions using mobile smartphones. In *Proceedings of the 6th ACM conference on embedded network sensor systems* (323–336).

Monreale, A., Pinelli, F., Trasarti, R., & Giannotti, F. (2009). WhereNext: A location predictor on trajectory pattern mining. In *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining* (637–646).

Morzy, M. (2007). Mining frequent trajectories of moving objects for location prediction. *Machine Learning and Data Mining in Pattern Recognition*, 667–680.

Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*, 691–692.

Nielsen, J. (2006, October 9). Participation inequality: Encouraging more users to contribute [Nielsen Norman Group]. Retrieved August 5, 2013, from http://www.nngroup.com/articles/participation-inequality/

Nielsen, T. A. S. & Hovgesen, H. H. (2008). Exploratory mapping of commuter flows in England and Wales. *Journal of Transport Geography*, *16*, 90–99.

Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., & Mascolo, C. (2012). A tale of many cities: Universal patterns in human urban mobility. *PLoS ONE*, *7*, e37027.

Ochoa, X. & Duval, E. (2008). Quantitative analysis of user-generated content on the web. In *Proceedings of webevolve2008: web science workshop at WWW2008* (1–8).

Openshaw, S. (1983). *The modifiable areal unit problem*. Geo Books Norwich.

O'Reilly, T. (2005, September 30). What is web 2.0 [O'Reilly Media]. Retrieved May 3, 2013, from http://oreilly.com/web2/archive/what-is-web-20.html

O'Sullivan, D. & Unwin, D. J. (2003). *Geographic information analysis*. John Wiley & Sons.

Padmanabhan, V. N. & Subramanian, L. (2001). An investigation of geographic mapping techniques for internet hosts. *ACM SIGCOMM Computer Communication Review*, *31*, 173–185.

Partridge, K. & Golle, P. (2008). On using existing time-use study data for ubiquitous computing applications. In *Proceedings of the 10th international conference on ubiquitous computing* (144–153).

Perez, S. (2012, August 16). Following Twitter suspension, WeKnowYourHouse returns, continues to post twitter users' addresses, home photos [TechCrunch]. Retrieved July 18, 2013, from http://techcrunch.com/2012/08/16/following-twitter-suspension-weknowyourhouse-returns-continues-to-post-twitter-users-addresses-home-photos/

Phithakkitnukoon, S., Horanont, T., Di Lorenzo, G., Shibasaki, R., & Ratti, C. (2010). Activity-aware map: Identifying human daily activity pattern using mobile phone data. *Human Behavior Understanding. Lecture Notes in Computer Science*, *6219*, 14–25.

Piantadosi, S., Byar, D. P., & Green, S. B. (1988). The ecological fallacy. *American Journal of Epidemiology*, *127*, 893–904.

Pot, J. (2011, April 14). Creepy shows just how much geolocation data you broadcast online [MakeUseOf]. Retrieved July 18, 2013, from http://www.makeuseof.com/tag/creepy-shows-geolocation-data-broadcast-online/

Purves, R. S. (2011). Methods, examples and pitfalls in the exploitation of the geospatial web. In *The handbook of emergent technologies in social research*. Oxford University Press.

Purves, R. S., Edwardes, A., & Wood, J. (2011, August 13). Describing place through user generated content. *First Monday*, *16*.

Quinlan, J. R. (1993). *C 4.5: Programs for machine learning*. Morgan Kaufmann.

Rao, L. (2012, January 24). Location-based shopping app Shopkick now 3 million users strong; 1B deals viewed [TechCrunch]. Retrieved July 22, 2013, from http://techcrunch.com/2012/01/24/location-based-shopping-app-shopkick-now-3-million-users-strong/

Ratti, C., Williams, S., Frenchman, D., & Pulselli, R. M. (2006). Mobile landscapes: Using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design*, *33*, 727–748.

rda research. (2013). About geodemographics. Retrieved August 26, 2013, from http://www.rdaresearch.com.au/geodemographics

Rösler, R. & Liebig, T. (2013). Using data from location based social networks for urban activity clustering. In *Geographic information science at the heart of Europe* (55–72). Springer.

Ross, C., Terras, M., Warwick, C., & Welsh, A. (2011). Enabled backchannel: Conference Twitter use by digital humanists. *Journal of Documentation*, *67*, 214–237.

Sadilek, A., Kautz, H., & Bigham, J. P. (2012). Finding your friends and following them to where you are. In *Proceedings of the fifth ACM international conference on web search and data mining* (723–732). WSDM '12. New York, NY, USA: ACM.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on world wide web* (851–860).

Sarver, R. (2009, November 19). Think globally, tweet locally [Twitter Blog]. Retrieved October 7, 2013, from https://blog.twitter.com/2009/think-globally-tweet-locally

Scellato, S., Noulas, A., Lambiotte, R., & Mascolo, C. (2011). Socio-spatial properties of online location-based social networks. In *Proceedings of the 5th international AAAI conference on weblogs and social media* (329–336).

Schade, S., Díaz, L., Ostermann, F., Spinsanti, L., Luraschi, G., Cox, S., . . . Longueville, B. D. (2013). Citizen-based sensing of crisis events: Sensor web enablement for volunteered geographic information. *Applied Geomatics*, *5*, 3–18.

Schiller, J. & Voisard, A. (2004). *Location-based services*. Elsevier.

Schroeder, S. (2012, October 4). Facebook hits one billion active users [Mashable]. Retrieved July 17, 2013, from http://mashable.com/2012/10/04/facebook-one-billion/

Schuler, M. & Joye, D. (2005). *Typologie der Gemeinden der Schweiz: 1980–2000*.

SECO. (2012). *Die Lage auf dem Arbeitsmarkt, Dezember 2011*. Staatssekretariat für Wirtschaft SECO. Bern.

Sevtsuk, A. & Ratti, C. (2010). Does urban mobility have a daily routine? Learning from the aggregate data of mobile networks. *Journal of Urban Technology*, *17*, 41–60.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423.

Smith, A. & Brenner, J. (2012). *Twitter use 2012*. Pew Research Center.

Song, C., Qu, Z., Blumm, N., & Barabási, A. L. (2010). Limits of predictability in human mobility. *Science*, *327*, 1018–1021.

Spirito, M. A. (2001). On the accuracy of cellular mobile station location estimation. *IEEE Transactions on Vehicular Technology*, *50*, 674–685.

Stange, H., Liebig, T., Hecker, D., Andrienko, G., & Andrienko, N. (2011). Analytical workflow of monitoring human mobility in big event settings using bluetooth. In *Proceedings of the 3rd ACM SIGSPATIAL international workshop on indoor spatial awareness* (51–58).

Sui, D. Z. & Goodchild, M. F. (2011). The convergence of GIS and social media: Challenges for GIScience. *International Journal of Geographical Information Science*, *25*, 1737–1748.

Takahashi, T., Abe, S., & Igata, N. (2011). Can Twitter be an alternative of real-world sensors? *Human-Computer Interaction: Towards Mobile and Intelligent Interaction Environments. Lecture Notes in Computer Science*, *6763*, 240–249.

Takhteyev, Y., Gruzd, A., & Wellman, B. (2012). Geography of Twitter networks. *Social Networks*, *34*, 73–81.

Tang, K. P., Lin, J., Hong, J. I., Siewiorek, D. P., & Sadeh, N. (2010). Rethinking location sharing: Exploring the implications of social-driven vs. purpose-driven location sharing. In *Proceedings of the 12th ACM international conference on ubiquitous computing* (85–94).

Tuan, Y.-F. (1977). *Space and place: The perspective of experience*. University of Minnesota Press.

Twitter Developers. (2012a). Finding Tweets about places. Retrieved July 26, 2013, from https://dev.twitter.com/docs/finding-tweets-about-places

Twitter Developers. (2012b). Firehose and filter streaming API. Retrieved July 26, 2013, from https://dev.twitter.com/discussions/9716

Twitter Developers. (2013a). GET search/tweets. Retrieved July 25, 2013, from https://dev.twitter.com/docs/api/1.1/get/search/tweets

Twitter Developers. (2013b). GET statuses/user_timeline. Retrieved July 26, 2013, from https://dev.twitter.com/docs/api/1.1/get/statuses/user_timeline

Twitter Developers. (2013c). REST API v1.1 limits per window by resource. Retrieved July 26, 2013, from https://dev.twitter.com/docs/rate-limiting/1.1/limits

Twitter Developers. (2013d). REST API v1.1 resources. Retrieved July 22, 2013, from https://dev.twitter.com/docs/api/1.1

Twitter Developers. (2014). Working with timelines. Retrieved January 18, 2014, from https://dev.twitter.com/docs/working-with-timelines

Twitter Help Center. (2013). FAQs about the tweet location feature. Retrieved July 22, 2013, from https://support.twitter.com/articles/78525-faqs-about-tweet-location

Ugander, J., Karrer, B., Backstrom, L., & Marlow, C. (2011). The anatomy of the Facebook Social Graph. *arXiv:1111.4503*.

Wagner, D., Lopez, M., Doria, A., Pavlyshak, I., Kostakos, V., Oakley, I., & Spiliotopoulos, T. (2010). Hide and seek: Location sharing practices with social media. In *Proceedings of the 12th international conference on human computer interaction with mobile devices and services* (55–58).

Wakamiya, S., Lee, R., & Sumiya, K. (2011). Crowd-based urban characterization: Extracting crowd behavioral patterns in urban areas from Twitter. In *Proceedings of the 3rd ACM SIGSPATIAL international workshop on location-based social networks* (77–84).

Wand, M. M. P. & Jones, M. C. (1995). *Kernel smoothing*. Crc Press.

Wang, Y., Zhu, Y., & Sun, Y. (2012). Nokia mobile data challenge: predicting semantic place and next place via mobile data. *Nokia Mobile Data Challenge 2012*.

Wickre, K. (2013, March 21). Celebrating #twitter7 [Twitter Blog]. Retrieved July 19, 2013, from https://blog.twitter.com/2013/celebrating-twitter7

Williams, M., Whitaker, R., & Allen, S. (2012). Measuring individual regularity in human visiting patterns. In *Proceedings of the 2012 international conference on social computing* (117–122).

Wood, J., Slingsby, A., & Dykes, J. (2011). Visualizing the dynamics of London's bicycle-hire scheme. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *46*, 239–251.

Worboys, M. (1998). Imprecision in finite resolution spatial data. *GeoInformatica*, *2*, 257–279.

Ye, M., Shou, D., Lee, W.-C., Yin, P., & Janowicz, K. (2011). On the semantic annotation of places in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining* (520–528).

Zhou, C., Frankowski, D., Ludford, P., Shekhar, S., & Terveen, L. (2007). Discovering personally meaningful places: An interactive clustering approach. *ACM Transactions on Information Systems (TOIS)*, *25*, 12.

A. Code

As can be seen, in this case, the JSON response consists of structured and hierarchical information about each Tweet (from line 43), including the time when the Tweet was sent in UTC (field "created_at" on line 45), and information about the user (line 56–61).

Listing A.1: *An example GET request to the Twitter REST API, requesting the most recent Tweet of Justin Bieber (Retweets excluded).*

```
GET /1.1/statuses/user_timeline.json?count=1&include_rts=false&screen_name=justinbieber HTTP/1.1
X-HostCommonName:
    api.twitter.com
Authorization:
    OAuth oauth_consumer_key="...",oauth_signature_method="HMAC-SHA1",oauth_timestamp="1374664745",oauth_nonce="1635874679",oauth_version="1.0",oauth_token="...",oauth_signature="q8kravuAKfKEZXMKtY2Q\%2Bsxdgew\%3D\"
Host:
    api.twitter.com
X-Target-URI:
    https://api.twitter.com
Connection:
    Keep-Alive
```

Listing A.2: *The REST API response to the above GET request.*

```
1  HTTP/1.1 200 OK
2  
3  content-type:
4      application/json;charset=utf-8
5  x-frame-options:
6      SAMEORIGIN
7  x-rate-limit-remaining:
8      177
9  last-modified:
10     Wed, 24 Jul 2013 11:25:53 GMT
11 status:
12     200 OK
```

```
date:
    Wed, 24 Jul 2013 11:25:53 GMT
x-transaction:
    d363cce1f87c5964
pragma:
    no-cache
cache-control:
    no-cache, no-store, must-revalidate, pre-check=0, post-check=0
x-xss-protection:
    1; mode=block
x-rate-limit-limit:
    180
expires:
    Tue, 31 Mar 1981 05:00:00 GMT
set-cookie:
    lang=en
set-cookie:
    guest_id=v1%3A137466515325033936; Domain=.twitter.com; Path=/; Expires=Fri, 24-Jul-2015 11:25:53 UTC
content-length:
    2635
x-rate-limit-reset:
    1374665645
server:
    tfe
strict-transport-security:
    max-age=631138519
x-access-level:
    read-write-directmessages


[
  {
    "created_at": "Tue Jul 23 21:16:05 +0000 2013",
    "id": 359784004611870700,
    "id_str": "359784004611870723",
    "text": "@YeshuaTheGudwin we just work hard. making music. being creative. #art",
    "source": "web",
    "truncated": false,
    "in_reply_to_status_id": 359768073793830900,
    "in_reply_to_status_id_str": "359768073793830912",
    "in_reply_to_user_id": 241760469,
    "in_reply_to_user_id_str": "241760469",
    "in_reply_to_screen_name": "YeshuaTheGudwin",
    "user": {
      "id": 27260086,
      "id_str": "27260086",
      "name": "Justin Bieber",
      ...
    },
    "geo": null,
    "coordinates": null,
    "place": null,
    "contributors": null,
    "retweet_count": 21629,
    "favorite_count": 15789,
    "entities": {
      "hashtags": [
        ...
      ],
      "symbols": [],
      "urls": [],
      "user_mentions": [
        ...
      ]
    },
    "favorited": false,
    "retweeted": false,
    "lang": "en"
  }
]
```

B. Software

Table B.1: *Used software and database systems.*

Software	Description, area of application
R 2.15.3	An open source programming language and software environment for statistical analysis and visualization (http://r-project.org); used here primarily for exploratory data analysis
Quantum GIS 1.8.0	An open source GIS with a wide array of functionality for spatial analysis and geovisualization; used here primarily for validating results of the various data processing steps and for geovisualization (http://qgis.org)
SQLite 3.7.10	A DBMS for managing simple, lightweight, and file-based databases (http://sqlite.org); used here for the storage of "raw" geographical information
SpatiaLite 3.0.1	A spatial extension to SQLite, implements standards of the OGC and provides advanced spatial operations, data structures, and projections; used here for the storage and spatial analysis of "standardized" geographical information (http://www.gaia-gis.it/gaia-sins)
Weka 3.7.10	A data mining and machine learning toolkit with a graphical user interface
Eclipse "Juno" with PyDev 2.7.0	An open source Integrated Development Environment for executing and debugging Python scripts
Python 2.7.4	The actual Python interpreter (http://python.org)
rpy2 2.3.6	Python wrapper for the R package, allows inclusion of R code in Python scripts
numpy 1.5.1	Python package for mathematics and statistics
sqlite3 2.6.0	Python interface to the SQLite 3 DBMS
pyspatialite 2.3.1	Python interface to the SQLite 3 and SpatiaLite DBMS
fiona 0.10	Python package for reading from and writing to shapefiles; used here for scripted imports of shapefiles into the database

C. Additional Temporal Patterns

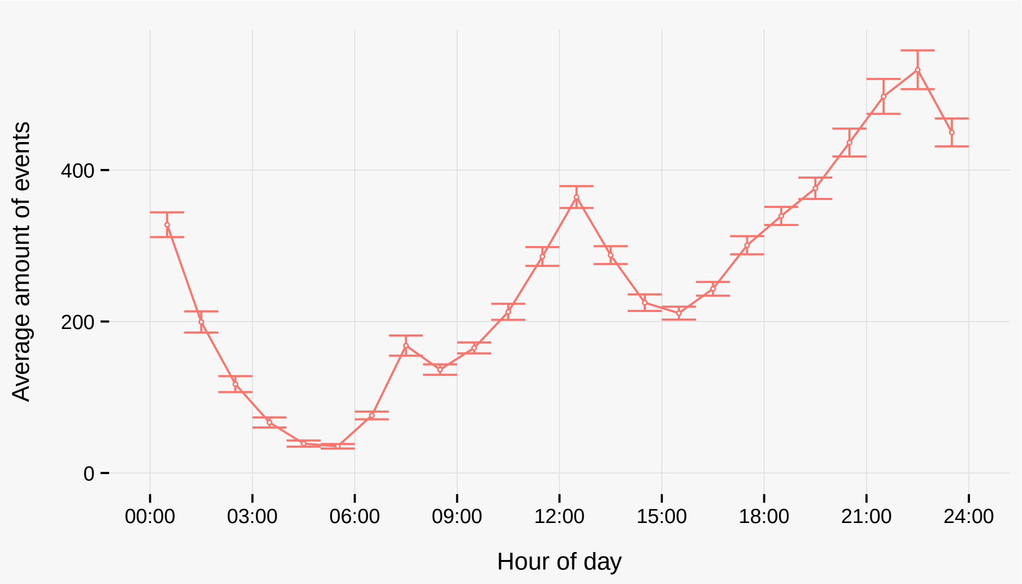


Figure C.1: *Hourly volume of events (whole study area) (N=1,913,512 events). Error bars signify 95%-confidence intervals.*

Figure C.2: *Hourly volume of events grouped by linguistic region (N=1,913,512 events). Error bars signify 95%-confidence intervals.*

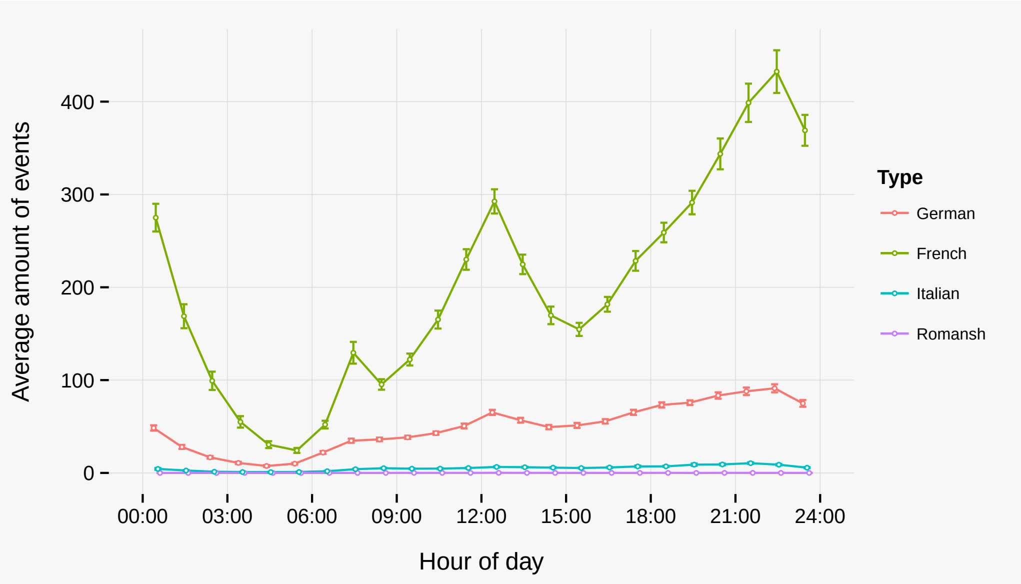
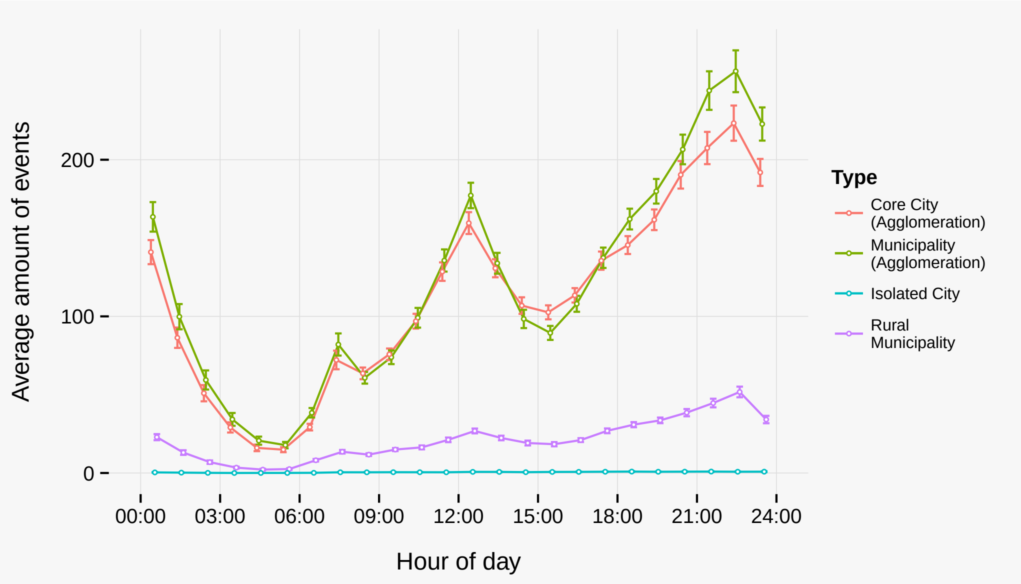


Figure C.3: *Hourly volume of events grouped by urban and rural municipalities (N=1,913,512 events). Error bars signify 95%-confidence intervals.*



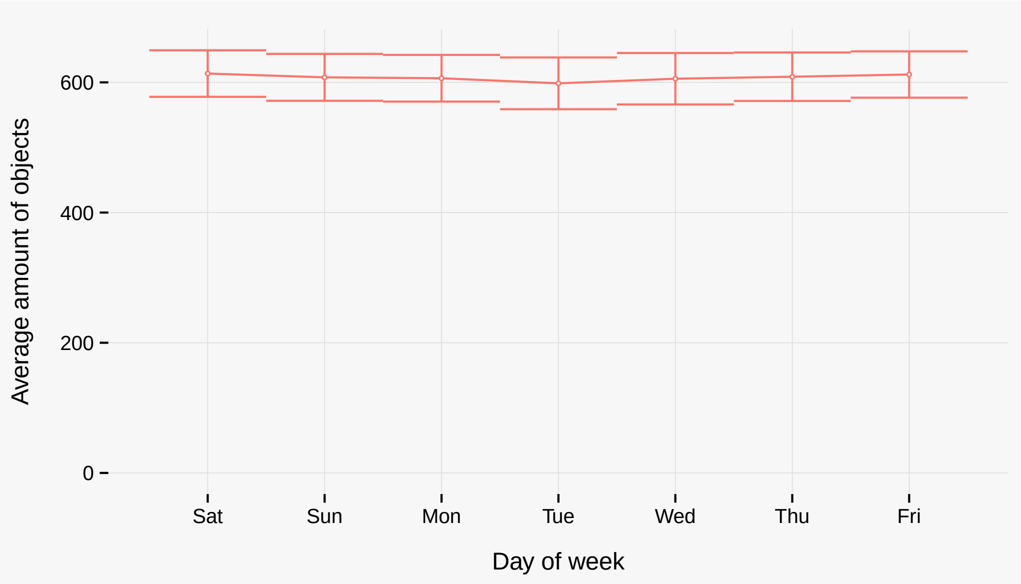


Figure C.4: *Daily volume of distinct objects (whole study area) (N=1,913,512 events). Error bars signify 95%-confidence intervals.*

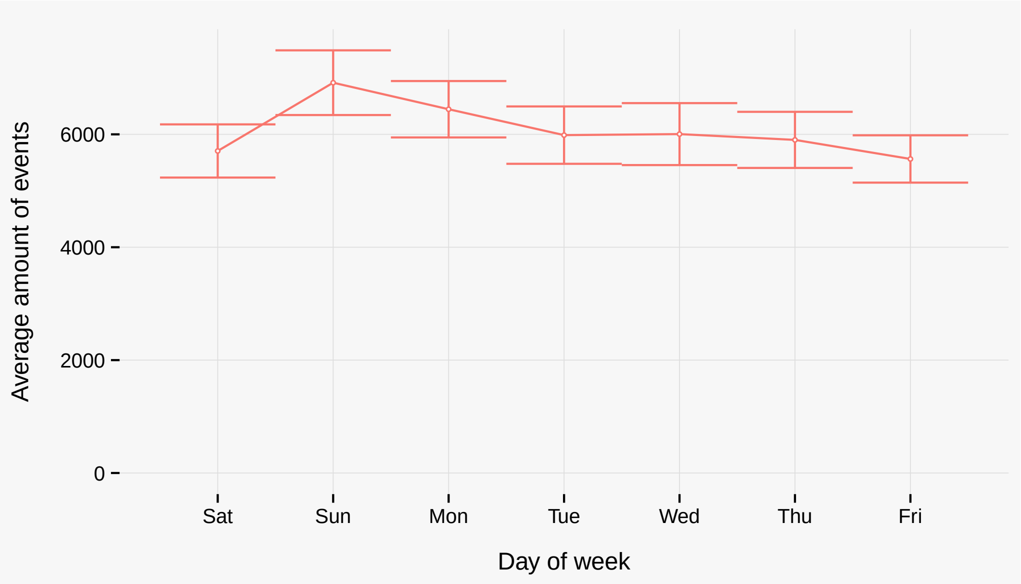
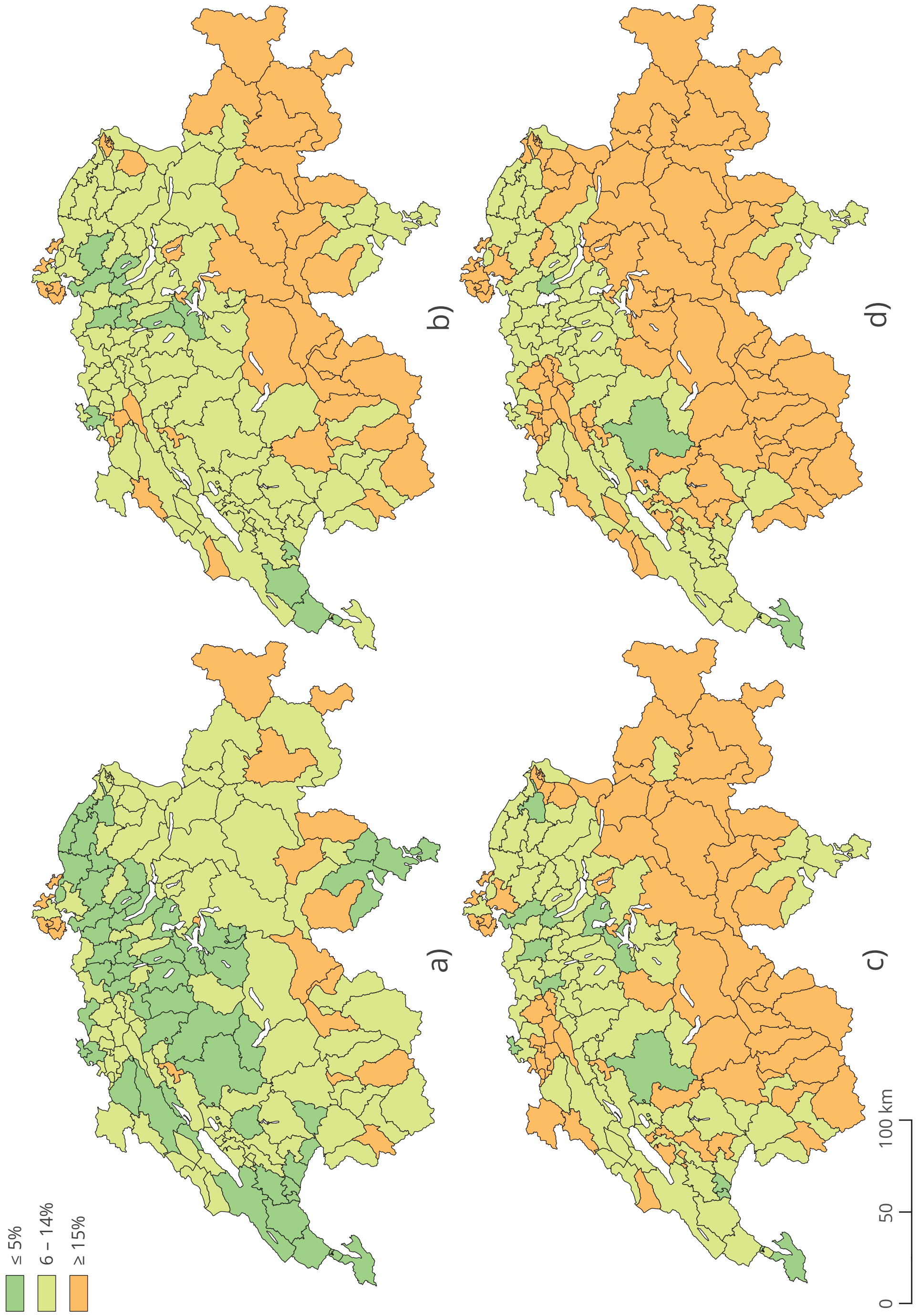


Figure C.5: *Daily volume of events (whole study area) (N=1,913,512 events). Error bars signify 95%-confidence intervals.*

D. Confidence Intervals of Authoritative Commuter Balance Estimations

Figure D.1: *Relative share of two-sided 95%-confidence intervals of estimated values for commuter balances per district (authoritative data). n_i for the mobile working population (a), out_i for the mobile working population (b), in_i for the mobile working population (c) and n_i for the mobile population in education (d).*



E. Metropolitan Regions of Switzerland

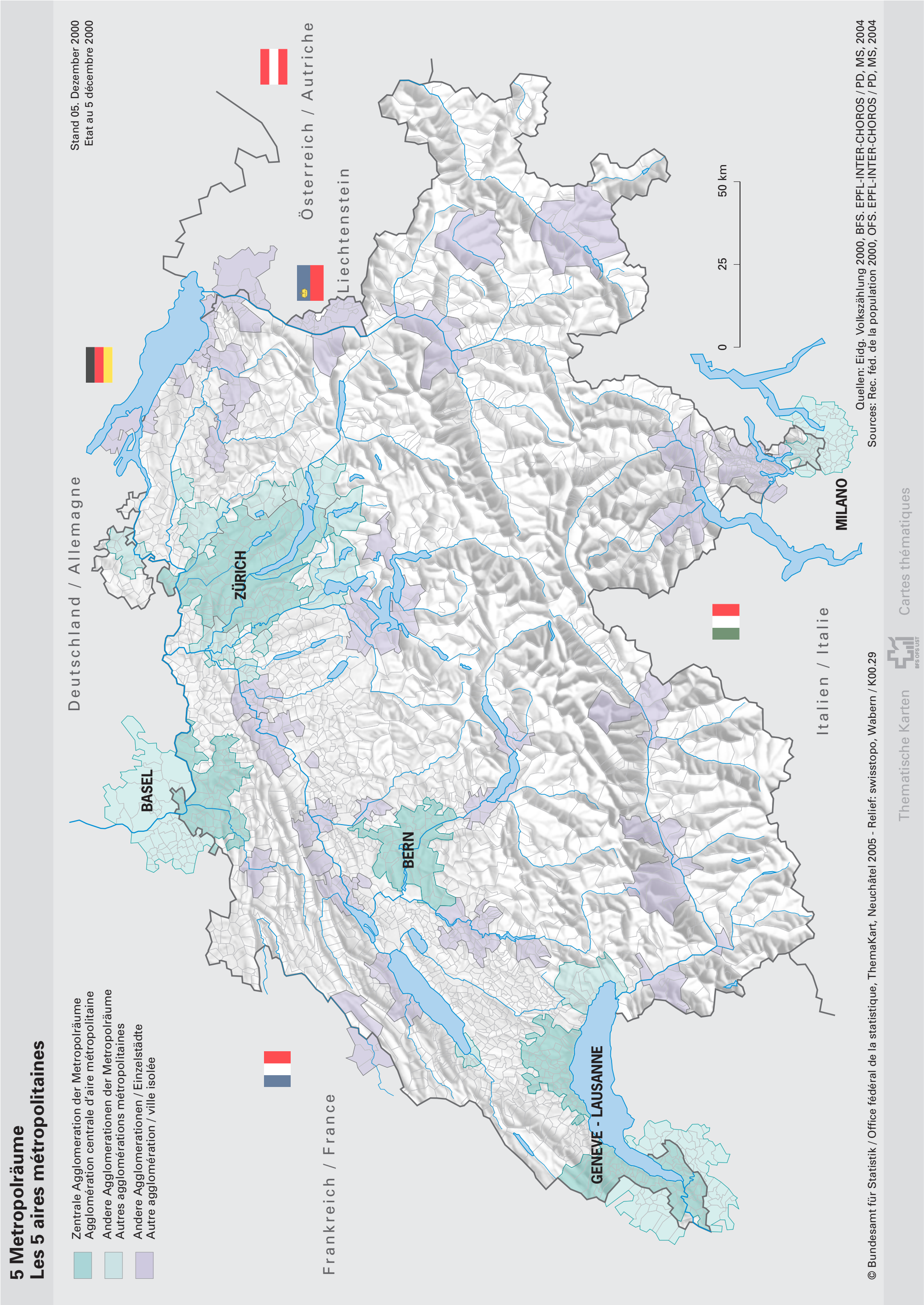

5 Metropolräume
Les 5 aires métropolitaines
Zentrale Agglomeration der Metropolräume
Agglomération centrale d'aire métropolitaine
Andere Agglomerationen der Metropolräume
Autres agglomérations métropolitaines
Andere Agglomerationen / Einzelstädte
Autre agglomération / ville isolée
Stand 05. Dezember 2000
Etat au 5 décembre 2000
Deutschland / Allemagne
Österreich / Autriche
Liechtenstein
Frankreich / France
Italien / Italie
BASEL
ZÜRICH
BERN
GENEVE - LAUSANNE
MILANO
0
25
50 km
© Bundesamt für Statistik / Office fédéral de la statistique, ThemaKart, Neuchâtel 2005 - Relief: swisstopo, Wabern / K00.29
Quellen: Eidg. Volkszählung 2000, BFS. EPFL-INTER-CHOROS / PD, MS, 2004
Sources: Rec. féd. de la population 2000, OFS. EPFL-INTER-CHOROS / PD, MS, 2004
Thematische Karten
Cartes thématiques

Figure E.1: *The five metropolitan regions of Switzerland, 2000, denoted in turquoise. Source: FSO.*

F. Results Obtained Through Other Extraction Scenarios

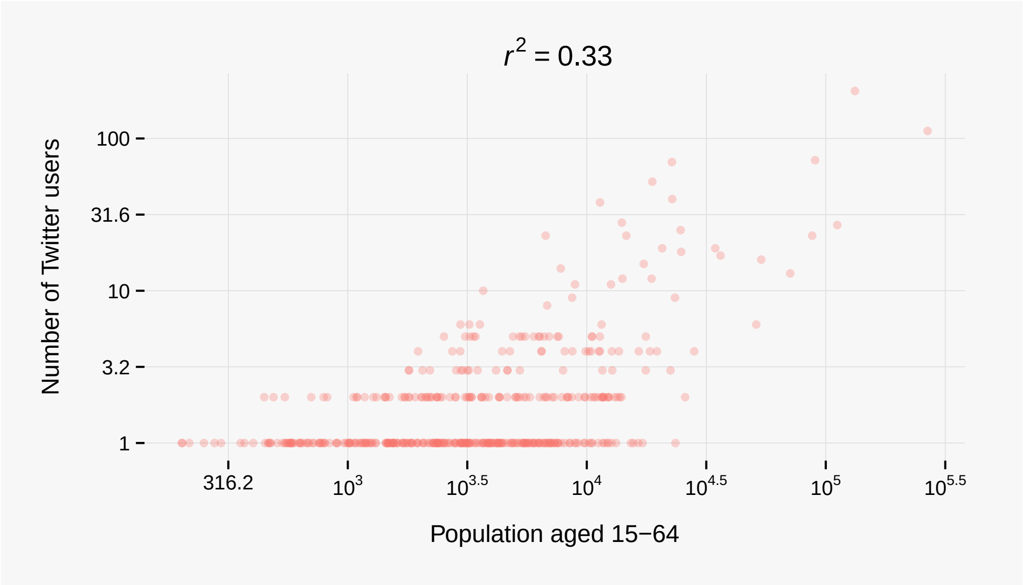
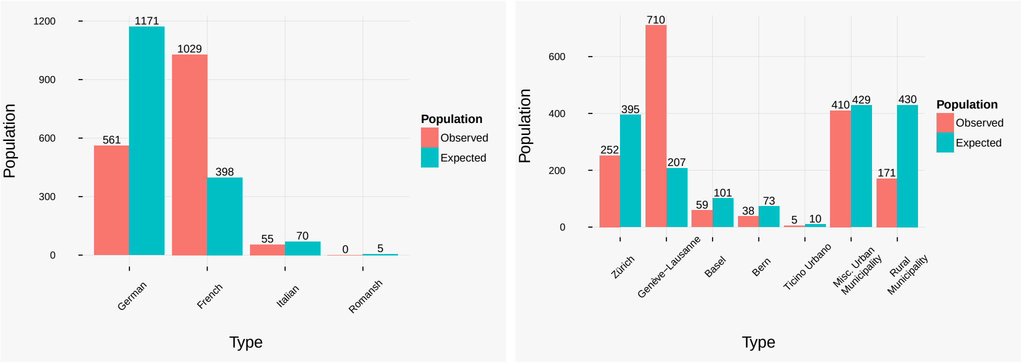


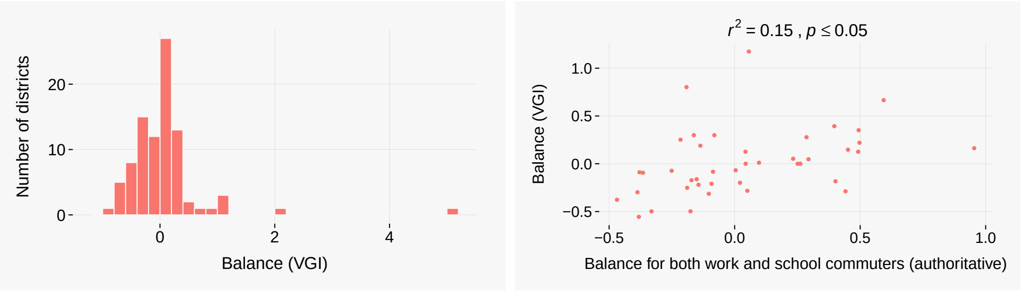
Figure F.1: *Relationship between actual population aged 15–64 and residential Twitter users per municipality as inferred through scenario C. Only municipalities with at least one residential Twitter user are included (N=430 municipalities, N=1,645 objects). Axes show the logarithm to the base of 10. Compare with Figure 5.10.*



(a)

(b)

Figure F.2: *Expected and observed number of residential Twitter users grouped by linguistic regions (a) and metropolitan regions (b) as inferred trough scenario C (N=1,645 objects). Compare with Figure 5.14 and Figure 5.16, respectively.*



(a)

(b)

Figure F.3: *Distribution of cb_i for Twitter data (a) and comparison of cb_i between authoritative and Twitter data (b) as inferred through scenario B. Compare with Figure 5.22a and Figure 5.24b, respectively.*

G. Personal Statement

I hereby declare that I have created this work completely on my own and used no other sources or tools than the ones listed, and that I have marked any citations accordingly.

Hiermit versichere ich, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie Zitate kenntlich gemacht habe.

Zürich, January 2014
Timo Grossenbacher