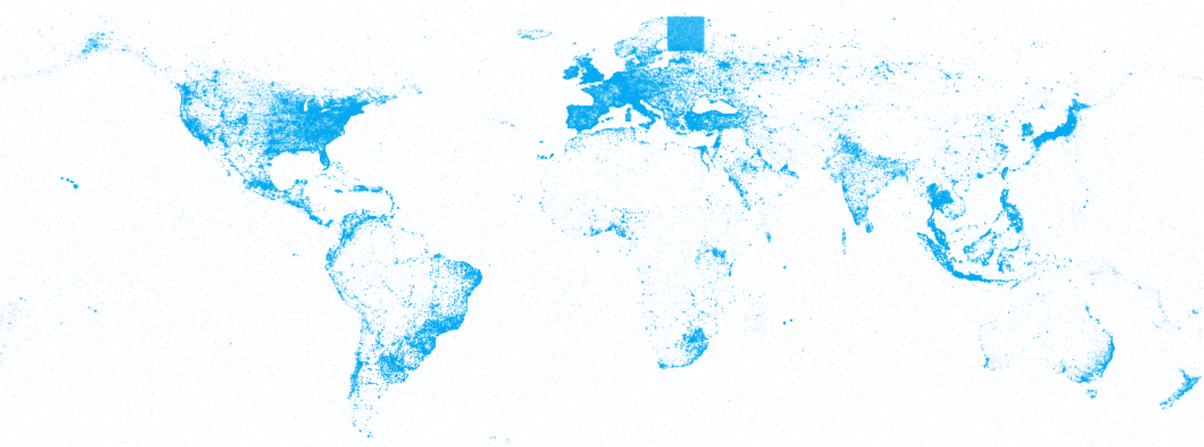
University of Zurich^UZH
Department of Geography

Master's Thesis

GEO 511 Masterarbeit

Geotagged Tweets as a Proxy for Human Settlements

An exploratory search for spatial relationship in noisy Big Data



David Hanimann

11-709-672

Submitted on the 26th of January, 2018

Faculty Member

Prof Dr Ross Purves

Geocomputation Unit
Department of Geography
University of Zurich

Supervisor

Dr Emanuele Strano

Dept. of Civil and Environmental Engineering
Massachusetts Institute of Technology (MIT)
Cambridge, MA, USA
emanuele.strano@gmail.com

Abstract

Twitter data reflects what people do and think. Many tweets contain geotags, that allow to spatially locate the tweet content. The utility of geolocated Twitter data for land use classification has been explored by researchers. Most studies focus on relatively small study areas due to uneven global adoption of Twitter. It is assumed that geotweets are mainly generated within human settlements and predominantly in cities. However, this has never been assessed systematically.

In this study the potential of geotagged Twitter data to locate human settlements is estimated. Therefore the spatial overlap of geotagged Twitter data with human settlements is assessed. As ground truth serves the Global Urban Footprint (GUF) mask, a high resolution global model depicting human settlements. Furthermore, the influence of geotemporal patterns, weather, settlement size and user characteristics are observed and used to filter out tweets outside human settlement. Finally the amount of Twitter data to get full global coverage of human settlements is estimated using Monte Carlo simulation. Contrary to most studies deducted so far, this thesis has a global perspective. Special attention is given to the reliability of Twitter data, therefore two independently sampled Twitter datasets are applied.

The observed overall spatial overlap of Twitter and GUF is roughly 80%. The overlap is dependent on geographic region, daytime, weekday, season and user characteristics. No influence of weather was observed. The potential of Twitter data as classifier is limited. In a simulation it is assessed that data volumes of billions of geotagged tweets would be required use Twitter data as primary source for classifying human settlement. Also, the reliability of Twitter data must be questioned as the two independently sampled Twitter data sets often yield different results.

The high spatial overlap of Twitter data and human settlements can be seen as a good precondition for Twitter data and social media data in general to be used as proxy for human settlement. Yet way larger data volumes are required to do so. Nevertheless, the results can be used to put studies of small areas of interest into a global context. Significant variations in temporal patterns over space can be attributed to human behaviour. The reliability of global Twitter data from Twitter's Streaming API is disputable due to unrepresentative returned volumes.

Acknowledgements

This thesis is a blend of ideas coming from many people. I would like to sincerely thank my supervisors Professor Ross Purves from the University of Zurich and Dr. Emanuele Strano from the Massachusetts Institute of Technology. They provided me with valuable inputs and motivation. A special thank goes to Dr. Thomas Esch and his team at the German Aerospace Center. They supported me with necessary data and fruitful ideas. I would like to mention Dr. Curdin Derungs and Manuel Bär from the Department of Geography at University of Zurich, who supported me with their expertise for the very tricky parts. Finally the critical revision of my thesis by Florian Landerer, Nils Styger and Raphael Hanimann deserves many thanks.

Contents

List of Figures

List of Tables

Abbreviations

API	Application Programming Interface
CSR	Complete Spatial Randomness
ECDF	Empirical Cumulative Distribution Function
GMT	Greenwich Mean Time
GPS	Global Positioning System
GUF	Global Urban Settlement data set (Esch et al., 2017)
NND	Nearest Neighbour Distance
NNI	Nearest Neighbour Index
OSM	Open Street Maps
PCA	Principal Component Analysis
RF	Random Forest
RFID	Radio Frequency Identification
$\boldsymbol{T_{eu}}$	Twitter data set queried over Europe
$\boldsymbol{T_{world}(EU)}$	Subset of tweets from T_{world} within the spatial extent of T_{eu}
$\boldsymbol{T_{world}}$	Twitter data set queried worldwide
VGI	Volunteered Geographic Information
WLAN	Wireless Local Area Network
WSS	Within Sum of Squares

Chapter 1

Introduction

1.1 Motivation

Since the launch of the first iPhone in 2007 the world has changed dramatically. Smart mobile devices connected to the internet surround us without temporal nor spatial limits. This has changed the way we 'communicate, navigate, work and entertain ourselves' (Duke and Montag, 2017). Psychologists and sociologists eagerly attempt to understand the transformations in humans and societies driven by smartphones and their possibilities. Notwithstanding, smartphones are not only information-feeding and communication-enabling gadgets. As a by-product of our daily use of smartphones and internet we generate massive amounts of data. Today, an almost inconceivable data volume of 5 exabytes ($= 10^6$ terabytes) is generated on planet earth every other day. This corresponds to the accumulated data volume that had been created by humans by 2003 (Akoka et al., 2017). The abundance of data has given rise to the importance of data science and data analysts that try to extract valuable information from that data. This relatively recent phenomenon has been described with the still inchoately defined term *Big Data*.

The descriptive has arisen because big data is in many regards different from traditional data. A common characterization of big data are the 5 V: huge Volume, high Velocity, low Veracity, high Variety and high Value (Jin et al., 2015). In the first place, big data is about immense data volumes. This poses a challenge to computing power and capacity, as big data analytics may exceed current processing performance (Volume). Though, big data is not only 'big', but it is consequently generated at high speed. This challenges data analysts to store and exploit near real-time data (Velocity). On Twitter roughly 6'000 tweets are sent every second[1]. Big data is highly variable in structure and content (Variability). It comprises structured, semi-structured and completely unstructured data such as text or images. Semi-structured data formats like JavaScript Object Notation (JSON) have become standard in data exchange over the web. Algorithmic concepts to find structures in this data, such as statistical learning, are

[1] http://www.internetlivestats.com, accessed 30-11-2017

required. Further, Big Data is less reliable than traditional data (*Veracity*). One major challenge is to clean the data and distinguish useful from noise data. In social media automated bots contribute large amounts to the data. Last but not least, the *value* of Big Data is undoubtedly outstanding. Leading companies like Google or Facebook are at the core dealing with big data. It is estimated that from 2013 to 2020 90% of the growth in IT industry is driven by big data (Jin et al., 2015). Not only economy, but also research institutes have entered this door. By November 2017 Google Scholar lists 251'000 papers with the keyword *Big Data*.

While this widely accepted delineation of big data sees high temporal density of data as an inherent characteristic, it does not mention space. Mobile devices build the core technology for the permanent access to the world wide web. These devices are normally GPS-enabled or make use of other technologies to be location-aware, such as RFID, WLAN or cell phone tower triangulation. The data produced with mobile devices is spread across space and often explicitly geolocated. Thus, big data is also about geography and its analysis an issue for GIScience, the science of geographic information (Graham and Shelton, 2013).

Geography elucidates big data

Geographers have investigated the potentials and limitations of geographic big data. One of the most intriguing types seems to be social media data. Numerous social media platforms have established geolocation support (e. g. Twitter, Flickr, Facebook) or are inherently spatial in nature (e. g. Foursquare). Some allow their data to be harvested via API's, which makes it very attractive for researchers to make use of. As data from social media reflects what people are 'looking, hearing, feeling' it can be seen as a sensor of real-world phenomena (Takahashi et al., 2011). Geographic social media data unravels where people go, what they think about places and how they use them. Related studies have been published on diverse topics ranging from demography, human mobility or land use classification. A plethora of studies has proven the high potential of georeferenced social media data in geography and also pinpointed out limitations.

The potential benefits of social media data in a geographic context can be summarized in two branches. (1) They may serve as an alternative source of data to accomplish tasks that have been done using traditional data. Social media data thus complement or replace other means of analysis at relatively low cost. For example Patel et al. (2016) found that population density as obtained from costly census data can to some degree be estimated using geolocated tweet density. Social media data has also been used as control data for existing population density model validation (e. g. Lin and Cromley (2015)). Hawelka et al. (2014) were able to successfully correlate global human mobility patterns with tourism data provided by the World Economic Forum. (2) Social media data can semantically enrich existing data and extend their resolution. Wang et al. (2016) and Frias-Martinez and Frias-Martinez (2014) *enrich* urban land use classes with temporal human activity derived from social media. Adnan et al. (2013) extract ethnicity from Twitter data and come up with a high-*resolution* ethnic map of London. High resolution can also be enhanced temporally, for example in population distribution (Steiger, Westerholt, et al., 2015). A more exploratory study by Lloyd and Cheshire (2017) derives retail center locations and their customer catchment areas. Another question that has been subject to crowdsourced data analytics is human perception of space, e. g. Chesnokova et al. (2017), Jenkins et al. (2016). Arribas-bel et al. (2015) proclaim valuable applications of such data in the context of *smart cities*.

Insightful study results have, however, shed light on the manifold challenges. Contrary to traditional data, social media data is not produced for the purpose of its analysis and information retrieval. The data is therefore blurred and imprecise. E. g. in terms of geolocation the coordinates can be derived by different means (GPS, WLAN) that have different levels of precision (Blanford et al., 2015). Precision is in this regard traded off against larger data volumes. Likewise, accuracy of results, i. e. relation of data to the phenomenon at inspection, is often hard to assess (Goodchild, 2013). Fore example, population distribution derived from social media may be biased by the prevalent demographic groups on the respective platform. Assessment of accurate information is hampered by the unrepresentative nature of the data. Data gathered from social media —Quelle— platforms are not representative for human population. Additionally, social media users are contributing to a varying degree. In a geographic context, social media penetration is also unevenly adopted across the globe. A last pitfall is the absence of accurate validation data. E. g. high resolution mapping of ethnicity (Adnan et al., 2013) or space perception (Chesnokova et al., 2017) can hardly be validated by other data, as traditionally such phenomena can't be assessed at high spatial resolution. It is common practice to use control data as plausibility check rather than as actual validation. Lack of validation data is the curse of studies that aim to extract novel kinds of information from geographic big data.

Twitter data and Human Settlements

This thesis puts a much simpler, easier-to-validate and more fundamental question at its origin. Where are social media data located in space? More precisely, *are Twitter data located where human settlements are?* It is sensible to assume that people are predominantly staying inside areas where man-made infrastructure is present. Infrastructure is the precondition for reachability. In very general terms it can therefore be assumed that where there's built-up area, there are probably people and where there's no built-up area, there are probably no people. In fact this relationship has been used to estimate population density from remotely sensed built-up area models (Lin and Cromley, 2015).

If the geolocation of social media data is to reflect where people are, as pointed out by Takahashi et al. (2011), it could be inferred that social media data are mainly situated on built-up areas and within settlements. Studies have already glimpsed at this relationship, e. g. Leetaru et al. (2013). A set of Twitter data collected from 2009 to 2013 by (Rios, 2013) illustrates this (Figure 1.1). The analysis of the spatial relationship of built-up areas and Twitter data distribution may hence validate or question the assumption that population distribution can be derived from built-up areas.

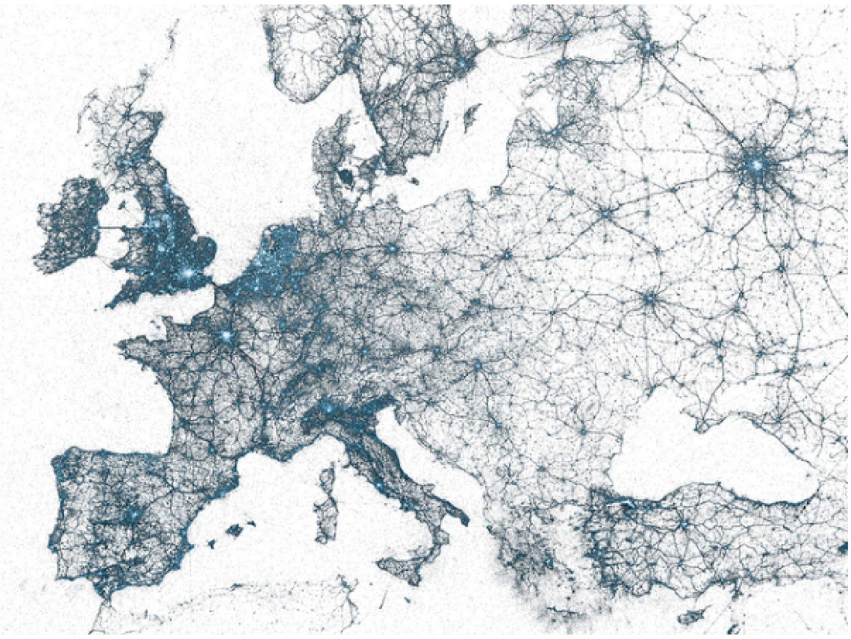


Figure 1.1: Georeferenced tweets sampled from 2009 to 2013 (Rios, 2013).

Land cover classification

If georeferenced Twitter data really *is* located where human settlements are, then it should be viable to infer the location of human settlements through Twitter data. The classification of built-up land cover on large areas has traditionally been accomplished using remote sensing products from air- and spaceborne sensors. The general procedure is to assign signatures to spectral or radiometric measurements that relate to a specific type of land cover. Often ancillary data like DEM's complement these (Esch et al., 2017). There are numerous models depicting human settlements globally, such as Globcover[2] or NASA Night Lights[3]. The main challenge in remote sensing and in the extraction of built-up areas in particular is the trade-off between spatial, spectral and temporal resolution. Global products predominantly rely on coarse-resolution ($<$ 100m) imagery (Potere et al., 2009). However, the typical physical scale of built-up areas is at 10 - 20 m (Small, 2003). In 2017 Esch et al. (2017) released a high-resolution product (12m) derived from radiometric measurements. The model depicts human settlements at an unprecedented accuracy. Radiometric measurements can alleviate problems inherent to spectral methods such as cloud cover.

Despite these recent advances, several challenges remain. Data quality and atmospheric effects are of particular exigency for the detection of small-area class of human settlements. The definition of land cover classes (e. g. urban land, human settlements, built-up area) is varying and dependent on the raw data (Potere et al., 2009). The layer produced by Esch et al. (2017) for example depicts only built-up

[2] http://due.esrin.esa.int/page$_{g}lobcover.php, accessed 2017-12-03$

[3] https://www.nasa.gov/, accessed 2017-12-03

objects with a vertical component, hence, e. g. roads are not mapped. Finally, high spatial and temporal resolution is still involving high costs (Potere et al., 2009).

In recent years a new potential source of land cover classifiers has risen: humans. Geo-Wiki is a platform that asks volunteers to validate land cover models (Fritz et al., 2009). A similar project is OpenStreetMap (OSM), a platform that lets citizens create polygons for objects on earth surface. Such approaches are a low cost alternative with potentially high spatial and temporal resolution. However, the acquisition relies on people that actively participate in mapping earth. Therefore, the data is commonly summarized as Volunteered Geographic Information (VGI). This data is introducing a new kind of bias. Densely populated areas in developed countries are much more completely reported than remote places in developing countries (Barrington-Leigh and Millard-Ball, 2017). This is a limitation that has been attempted to tackle through gamification (Baer, 2017).

Contrary to VGI, social media data are generated in huge volumes without active human 'work' involved. Opposite to remote sensing, the temporal and spatial resolution of the data does not inherently conflict due to physical limitations. The costs for data acquisition are very low. To take advantage of this resource, it has to be assessed in how far social media data indicate built-up land cover. Studies on spatial social media data mentioned above aim at describing *land use*. Opposite to land cover, the description of physical properties of the earth surface, land use is the description of earth according to how humans use it. Researchers see high potential for social media to fill this gap. − Quelle– For this purpose, such studies extract textual or geotemporal patterns from the data. However, the applicability in land cover classification has to the knowledge of the author not been studied. Despite this, the data has sparked interest in the domain of remote sensing and already been embodied in several studies involving remote sensing (e. g. Chakraborty et al., 2015; Lin and Cromley, 2015).

The global study

Most studies on geographic social media focus on study areas of limited size (rare exceptions are Graham, Stephens, et al. (2013), Takhteyev et al. (2012) and Leetaru et al. (2013)). When comparing the results of studies, one needs to take into account that observations may vary globally. While demographic variables about Twitter usage have been studied by means of personal interrogation (Blank, 2016), observations drawn directly from Twitter data have not. E. g. diurnal Twitter activity has been described in a number of papers –Quelle–. But It is unknown, whether this feature is globally constant. This is a major research gap that hampers the validation and comparability of existing research.

This thesis incorporates a global data set in spite of the manifold entailing obstacles. These range from uneven geographic adoption of social media, excessive data volumes, time alignment problems or projection problems. However, a global approach does not only face difficulties. It potentially reveals spatially dependent characteristics of Twitter data. Such patterns may be helpful to put local study results into a wider geographic context.

1.1.1 Research Questions

The previous section has disentangled a number of research gaps in the context of geographic social media data. (1) A comprehensive description of the relation of Twitter data and human settlements puts to test the assumption that people are staying within built-up areas. (2) Social media data, such as Twitter, may complement and enhance the description of space as done with remote sensing or VGI. On the one hand, Twitter data may enhance remote sensing products, since it does not mutually exclude high spatial and temporal resolution. On the other hand, contrary to VGI, it does not require motivated people to get involved. Twitter data is created independently in large volumes. This potential has not been explored so far. (3) The vast amount of research on land use using Twitter data demands for a global framework to put results into context. This ensures that the rich semantics extracted from Twitter data are valid globally.

This study pursues a number of novel endeavours. For the first time the spatial overlap of geotagged Twitter data and human settlements is assessed quantitatively and at high resolution. Contrary to the great majority of studies, this feature is traced down on a global level. Factors that are expected to influence the spatial overlap of human settlement and Twitter data are proposed and tested. Finally, the potential of Twitter data as a proxy for human settlements is estimated through filtering by the proposed factors and assessment of the spatial pervasiveness of tweets. Accordingly, the following research questions are put forward:

RQ 1 *Does the presence of geolocated tweets indicate built-up land cover?*

RQ 2 *What are factors that influence whether the geotag of Twitter data is on settlement area or outside?*

RQ 3 *Is a global classification of built-up land cover possible with Twitter data?*

1.2 Background

In this section, a short introduction on Twitter and its applicability in geographic research will be given. First, the functioning of the Twitter platform is described. Second, biases and major caveats that arise are delineated. Then a review of existing literature on geographical Twitter data research is portrayed. In the context thereof a set of methods and guidelines for this study are presented.

1.2.1 What is Twitter?

Twitter is probably the most famous and far reaching microblogging service worldwide. Its concept is to enable anyone to share ideas and information. The mechanics are straightforward: Users can post short text messages of maximal 140[4] characters or images (*tweet*) that are visible for a selected audience

[4] As of November 2017 messages are allowed for up to 280 characters (https://www.nzz.ch/wirtschaft/twitter-verdoppelt-laengen-limit-auf-280-zeichen-ld.1326983, accessed 2017-12-03). The sampling period of the data used for this thesis is not affected.

(normally everyone). A user can follow other users to track his/her activity (e. g. Katy Perry is the most followed Twitter user with over 100M followers by August 2017[5]). Tweets can be denoted with *hashtags*. Hashtags allow to relate a tweet to a topic, e. g. #ZurichOpenair. The platform is operated by Twitter Inc., a company that was founded in 2006 and has since become one of the most visited websites on the internet[6]. In 2010 Twitter reached a daily tweet volume of 50 M, in 2013 this number increased to 500 M tweets per day [7], a level where Twitter more or less remained constant until now. Twitter claims to have 328 M active users monthly, whereof 82% gain access to the platform through mobile devices[8].

Twitter data features some advantageous characteristics in the context of this study. (1) Data volume of the service: Twitter has a huge user community of over 300 million users and with it a large data volume. (2) Accessibility: There are other services like Facebook or Whatsapp that have a way larger user base and even higher data volumes. Yet these don't give easy access to their data, partly because of the higher sensitivity of their data. Tweets are open for public and freely accessible through Twitter's Streaming API. (3) Nature of the data: E. g. Flickr gives easy access to a large data base, too. The purpose of Flickr is to share images. It can be assumed that georeferenced images tend to be sent from remote / scenic places such as National Parks (Li et al., 2013). Hence the spatial overlap of Flickr points and built-up area may be less sticking. Twitter on the other hand gives no obvious reasons for this bias to be assumed.

Twitter and Geography

As of 2009, Twitter implements geolocation sharing per tweet (Leetaru et al., 2013). There are two types of geo-localization to be differentiated. (1) The precise geolocation with coordinates supplied by the GPS-receiver or other means of positioning (e. g. cell phone tower triangulation, WLAN) internal to the users (mobile) device.[9] The positioning system have varying precision, i. e. a random offset of the generated position from the user's true position exists. (2) A set of coordinates that is place-related to a point of interest, neighbourhood, city, country or continent level. Rather than from positioning systems these coordinates most likely stem from a database where place names are holding a spatial reference. E. g. for the city of Zurich the corresponding coordinates returned by Twitter are always *Lat/Long 8.55/47.3667*. Place-relating coordinates is not only a way to semantically enrich the geodata but also a mean of obfuscation.

Obfuscation is a method to attenuate privacy concerns that come along with geolocation sharing. For the same reason, geolocation is turned off by default on Twitter. The user has to actively opt for enabling location addition to tweets. When users do so, locations are added on the obfuscated level only. To enable precise geolocations, the user faces further obstacles. Having enabled obfuscated location sharing, the user needs to explicitly look for the option to enable GPS-positioning. This can only be

[5] https://twittercounter.com/pages/100, accessed 2017-08-29
[6] https://www.alexa.com/topsites, accessed 2017-08-29
[7] http://www.internetlivestats.com/twitter-statistics, accessed 2017-08-29
[8] https://about.twitter.com/de/company, accessed 2017-08-18
[9] For a comprehensive discussion on positioning methods refer to Roxin et al. (2007)

done on Twitter's mobile application for Android and for iOS[10]. From the web page on internet browsers, regardless whether on mobile devices or desktop machines, it is not possible to share precise geolocation. Twitter's API returns geolocation in one or several of the following fields: a point feature that is either precise or artificially matched to a place name, a rectangular bounding box that is place-related or a text string.

Privacy considerations and the intricacy of the process of enabling geolocation lead to a generally low adoption of this functionality. Leetaru et al. (2013) find that only 1.8% of the tweets have a place-related and 1.6% of the tweets exact geotag. Other studies found similarly low values, e. g. 0.6 % Takahashi et al. (2011), 1.45 - 3.17% Morstatter et al. (2013), < 1% Li et al. (2013). The studies are not clear about which acquisition methods were used to come up with these numbers. This can probably explain the differing numbers. Also differing study areas may lead to differing share of precise geolocations. In any case the rate of adoption of geoservices is very low and does surely not exceed few percent.

Besides on-site geolocation-sharing by the user (as intended by Twitter's interface), geolocation can be administered by other means. There are third-party applications that require the user to grant access to their Twitter profile. One example is *swarmapp*[11]. This is a mobile device app that automatically tracks a person in space for the purpose of a diary generation, and posts certain locations on Twitter. The validity of geolocations is further complicated by the fact that geolocation on computers and mobile devices can be manipulated. Freely accessible software allows to overwrite the location derived from physical positioning systems (e. g. *Manual Geolocation*[12]). Whatever reasons one might have to fake his geolocation, the ease of doing so exposes Twitter users to do so.

Accessing Twitter data

The live Twitter data can be queried via Twitter's Streaming API. Access is granted upon creation of a registered *Twitter Application*. In this way the volume returned is limited to 1% of the whole Twitter traffic. The process of selecting the 1% can partly be curtailed by query parameters such as location or keywords. It is not transparent how Twitter subsets the Twitter data stream. Studies suggest that the sample is slightly biased in terms of e. g. top hashtags (Kumar et al., 2013). A costly way to overcome this limitation is the Twitter Firehose API that returns 100% of the Twitter data stream. Due to the high costs and extensive data volumes, most researches stick with the Streaming API sample.

1.2.2 Who tweets?

Twitter users do not represent the real world's population in three regards. (1) Demographic variables like age, gender and education correlate with Twitter usage. (2) The geographic location, i. e. culture and economic circumstances influence Twitter penetration. (3) A significant amount of the registered accounts are not belonging to single humans but rather to institutions (e. g. corporations) or fully

[10] https://support.twitter.com/articles/484789, accessed 2017-11-30

[11] https://de.swarmapp.com, accessed 2017-11-30

[12] https://chrome.google.com/webstore/, accessed 2017-12-03

automated (spam) bots. Moreover, geographic Twitter data isn't representative of Twitter's user base. (1) Some users are contributing more to the platform than others. (2) Sharing of geolocation is biased by mainly cultural factors.

Twitter and Demography

Blank (2016) investigated in demographic variables that correlate with Twitter usage. He found that 'Twitter users fit the profile of young, well-educated, and wealthy elites'. It is therefore hard to infer characteristics of any real-world population from Twitter data. In a purely geographic context this bias may not play an overwhelming role, unless demographic variables correlate in space. This seems to be the case indeed. In the same study by Blank it is estimated that rural people are less likely to be Twitter users. Another study by Klotz et al. (2017) shows that in places where poor people live (slums) there is less Twitter data. Whatever influence demographic variables might have on the presence and absence of Twitter data in space, this thesis focuses on the spatial distribution of tweets. The observed patterns may or may not be ascribed to demographic characteristics.

Geographical bias (Twitter penetration)

The spatial distribution (penetration) of Twitter data is highly biased in two regards. (1) At small scale the density of tweets varies between continents, regions and countries. E. g. African countries have much less georeferenced Twitter data than North America. The Netherlands have a much higher density than Belgium. On a larger scale the adoption of Twitter in cities is higher than in the countryside. In general it is assumed for Twitter data to be generated in Western European, Panamerican and East Asian cities (Leetaru et al., 2013).

It is probably for this reason that most studies on Twitter focus on confined areas with a high Twitter data density. The application of precise (GPS-) geotags for a specific research on a global level has to this date not been found. In fact most studies don't even regard places outside cities.

Bot or not?

The term *social* media suggests that Twitter users are people. However, not only accounts from single human users prowl about Twitter. Fully and partly automated programs, commonly summarized as bots, deliver large amounts of tweets (Chu et al., 2010). Fully automated accounts tweet updates from e. g. weather stations, seismic stations, Twitter trends, or just random content. Other users post messages on e. g. job opportunities, traffic jams and upcoming events or the like. Chu et al. (2010) distinguish *legitimate* and *malicious* bots, where legitimate correspond to the classes mentioned above. Malicious bots are fully automated users that tweet spam from hacked accounts. Legitimate bots account for large amounts of the data produced on Twitter. Malicious tweets are only a small fraction of Twitter's traffic (Grier et al., 2010).

Automated users often contribute meaningless data. In a geographic context, a bot may post geolocations that are not relevant. In Figure 1.2 the geolocated tweets of an automated bot and a real human are plotted. The bot posts random (meaningless) locations, where the real user's location point to relevant locations. Hence, an analysis of human Twitter user's behaviour has to exclude such accounts. Several attempts have been made to classify users as bots or normal human users. Grier et al. (2010) looked for bots by evaluating the URLs posted by users. Chu et al. (2010) found evidence for user characteristics in the frequency of tweets. They distinguished humans, cyborgs and bots. There are also ready-made bot classifiers available online (e. g. Botometer[13]). However, a standardized manner of cleaning Twitter data does not exist. Bot detection is still an ongoing research, no definite solutions of high enough robustness have been found by the author. This is also due to the fact that the most extensive exploration in the field can not resolve a further issue: no hard line can be drawn to separate bot from human. There are for example accounts kept by fire fighters or the fire brigade that post tweets from incidental sites. Depending on the study, these are desirable or not. Sensitive exclusion criteria lead to a loss of much data, insensitive criteria include more unwanted data. Therefore, studies often apply individual and rudimentary tactics that meet their requirements. Criteria used to subset the data range from number of tweets per user (Longley and Adnan, 2016), sorting by user name (Blanford et al., 2015), tweet keywords for specific topics (Allen et al., 2016) or similar tweet content (Lloyd and Cheshire, 2017). More common in literature is, however, that no mentions about data cleansing appear.

Instead of classifying (and eventually filtering) certain groups of users in advance, the behaviour of different groups of users is monitored. This requires methods that regard single user characteristics. For the sake of consistency, a simplistic terminology that is expected to summarize Twitter users appropriately is used henceforth. Here, bots are defined as accounts that without any human assistance post tweets. These may be live weather data, randomized processes, trends on Twitter or earthquake warnings measured from seismographic stations. Services on the other hand are defined as Twitter accounts that are created for the purpose of merchandising or information broadcasting. In comparison to bots, services tweet human-generated contents, but probably multiple users are involved. Examples are employment ads, firefighter alarms, traffic jams or local news. The categories *bot* and *service* are summarized as *nonpersonal* users. In contrast *personal* users are the model Twitter user accounts controlled by single humans.

Contribution bias

In scientific literature there is a consent that Twitter data, but also social media data in general is biased by a small number of very active and a large number of very inactive users (Longley and Adnan, 2016), (Longley, Adnan, and Lansley, 2015). Leetaru et al. (2013) even conclude that 'a very small number of core users thus drive the majority of Twitter's traffic'. In their data they find that the top 15% of the most active users account for 85% of the tweet volume and the top 5% for 48% of the tweet volume.

[13] https://botometer.iuni.iu.edu/

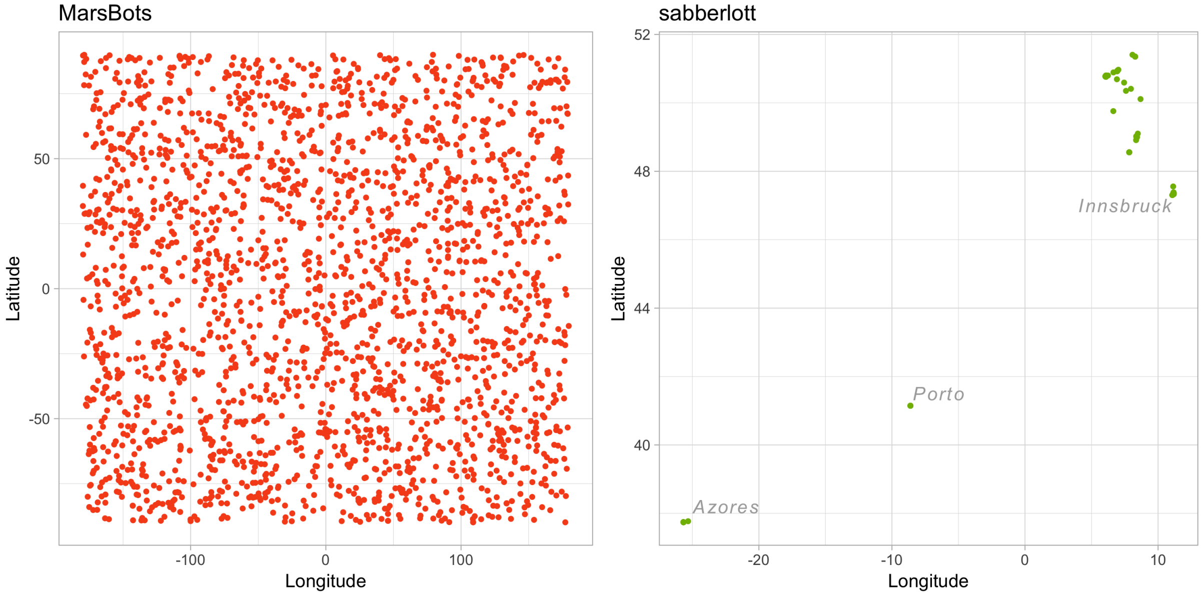


Figure 1.2: Point pattern generated from a user without (left) and with (right) spatial relevance.

This phenomenon is in Li et al. (2013) termed *contribution bias*. It is partly driven by nonpersonal users but also present within real human users. Hence Twitter data is not only unrepresentative of the world population, but also in terms of Twitter's users.

Who tweets with geolocation?

There is another kind of bias introduced through the specific subset of precisely geolocated tweets used. The Twitter users sharing their geolocation are only a small fraction of all Twitter. Leetaru et al. (2013) assess a share of marginal 1.6% of the users sharing their GPS-location. A more recent study by Sloan and Morgan (2015) discusses this issue. It is shown that there are small albeit significant differences in the geotagging behaviour of demographic groups. The disparities of geotag adoption between gender, age and profession are smaller than 1%. More peculiar are adoption rates among tweet languages. Some languages like Turkish (8.3%), Indonesian (7.0%) or Portuguese (5.9%) show a relatively high rate of geotag adoption. Very small rates are found in Korean (0.4%), Japanese (0.8%), Arabic (0.9%) German and Russian (2.0% each) tweets. Cultural aspects hence seem to influence the adoption rate of geolocated tweets. We can only assume that the language depends on countries. But it is impossible to get adequate numbers for geotag sharing in different geographic areas, because tweets without geotag are not spatially referenced. While this has substantial impact on social science study design, it may not be of concern for the pure geolocation. However, it can be expected that the presence of geotagged tweets in geographic regions is not only a function of Twitter penetration but so of the geotagging adoption.

1.2.3 Methods of Spatial Twitter analysis

Spatial Twitter data research can be grouped with respect to the attribute fields employed. Steiger, Albuquerque, et al. (2015) discerns (1) research using all information (including the tweet content) of tweets from (2) puerly spatio-temporal research. Studies using the tweet content normally extract semantics from tweets and locate it in space. Spatio-temporal reserach can according to a non-systematic literature review by the author be split up into three main subfields: (a) Geotemporal: The timestamp of tweets and temporal volumes of Twitter data per time are analysed for the extraction of human activity over time and space (e. g. Longley and Adnan, 2016). (b) Interactions: The interaction between tweets is assessed using geolocation, user id and re-tweeting behaviour. In doing so patterns of human mobility and connectedness are assessed (e. g. Blanford et al., 2015). The methods used in such studies are not suitable for the scope of this thesis. (c) Spatial: The application of the geolocation only is used to estimate population densities (e. g. Lin and Cromley, 2015). Methodologically, this study is interested in spatial and spatio-temporal patterns. A set of core literature on this topic is listed in Table 1.1.

Spatial patterns

The presence of social media data is seen as an indicator for human activity. Accordingly studies focusing on spatial distribution of tweets are normally dealing with estimation of population densities. For this purpose the density of tweets is assessed what involves some sort of aggregation. Spatial aggregation is normally done on grid cells or administrative units. Measures for Twitter density estimation are kernel density estimation (KDE) (Li et al., 2013)or quadrat counting (Jendryke et al., 2017; Patel et al., 2016).

The list of research using the geolocation of tweets only is short, as the scope of applications is limited to characteristics relating to human activity in space. In 1.1 all the research found by the author on purely spatial Twitter analysis is listed (gray).

Geotemporal patterns

Temporal variability in absolute observed tweet volumes are here referred to as *Twitter activity signatures* (Yang and Leskovec, 2011). These can be associated with human activity. The simplest pattern described in literature is the lower Twitter activity during night (e. g. Li et al., 2013). Most commonly research focuses on diurnal patterns of Twitter activity (e. g. Soliman et al., 2017). Tweet volumes also vary among days of the week. Weekend days show a different pattern than weekdays (Longley, Adnan, and Lansley, 2015). The variation of such patterns in space is in numerous studies used as an indicator for different types of land use (e. g. Li et al., 2013; Longley, Adnan, and Lansley, 2015; Yang and Leskovec, 2011). Frias-Martinez, Soto, et al., 2012 could assign differing temporal patterns to land use classes such as business districts, leisure and residential areas. Since these patterns are spatially variable, they may also influence the spatial overlap of Twitter with GUF.

The above mentioned studies usually focus on study areas of the extent of a single city. Different studies are hardly comparable, as the observed geotemporal patterns often vary. For example the daily

Study	Goal	Applied Methods	Study area
Lin and Cromley (2015)	Control data for population density model	Grid / administrative units and density function	Connecticut USA
Patel et al. (2016)	Population density modelling	Grid and administrative level volumes	Indonesia
Lloyd and Cheshire (2017)	Derive retail centre locations	KDE of selected tweets	United Kingdom
Jendryke et al. (2017)	Urban land use classification	Comparison of built-up land cover and Twitter volumes on grid level (quadrat count)	Shanghai
Longley and Adnan (2016)	Characterize Land Use by geotemporal activity	Administrative level and night-day-time patterns in Twitter content diversity	London
Li et al. (2013)	Analyse Twitter distribution	KDE and daytime / weekday patterns	USA
Leetaru et al. (2013)	Description of geotemporal patterns	Weekdays vs weekends, grid	Global
Helwig et al. (2015)	Model varations in spatiotemporal patterns	Daytime and weekdays on grid, 2h intervals	USA
Frias-Martinez and Frias-Martinez (2014)	Derive urban land use classes from geotemporal patterns	Use of diurnal (20 min intervals) and weekday patterns for clustering	New York, London, Madrid
Adnan et al. (2013)	Mapping spatio-temporal patterns with respect to user ethnicity	Administrative units and daytime/nighttime	United Kingdom
García-Palomares et al. (2018)	Characterize Land Use by geotemporal activity	Administrative levels and diurnal patterns	Madrid
Arribas-bel et al. (2015)	Characterize Land Use by geotemporal activity	Clustering of similar diurnal and weekday patterns at administrative unit level	Amsterdam

Table 1.1: Reviewed studies on spatial (gray) and spatiotemporal Twitter analysis.

and weekly patterns as found in Twitter data in London by Longley, Adnan, and Lansley (2015) and in Amsterdam by Arribas-bel et al. (2015) are not fully conform. While Longley finds maximum tweet volumes in the evening, Arribas-bel's data show peaks at noon. Two explanations are possible: Either the different study areas show different tweet behaviours. Or else temporal offset between the studies (changes in behaviour) or the data querying process leads to unequal results. In either case the obtained results are not generalizable.

Tweet Content

The actual tweet message in combination with geolocation reveals what people feel, think and do in geographic areas (Hahmann et al., 2014). However, not only the place influences the content of tweets, but also variables like daytime, season or weather. By means of sentiment analysis Modoni and Tosi (2016) showed that there is a relation between weather conditions and the reported mood on Twitter. If the tweet content is influenced by weather, it is not far to suppose that also the tweet location is affected by weather conditions. To the author's knowledge no research has been conducted on this topic. However, common sense would tempt us to expect that rainy weather makes people stay inside buildings. If this is to be true, then geolocation of tweets during rain are more likely to indicate human settlements.

1.3 Dealing with Twitter data intricacies

To answer the research questions as presented above in the context of the research presented, a number of issues have to be considered: (1) The Unrepresentative user base and bots and services present in the data, (2) the geographic bias in Twitter data distribution, (3) the geolocation acquisition method (GPS, WLAN, resmapling) of tweets is not explicitly given, (4) results in geographic Twitter data research vary, and (5) temporal Twitter activity patterns vary in space. In this section, methods that allow to approach the research questions taking into account these caveats will be disclosed.

Users

Twitter data is unrepresentative of the world population. Few users drive the majority of Twitter's traffic, a phenomenon that has led to the coercion of the term *contribution bias*. Bots and services exacerbate this circumstance by contributing vast amounts of data with (semi-) automatically generated content. Twitter analysis is compelled to investigate its content by user to assess whether different user groups behave differently in the context of the research questions. Well-established methods of removing undesired users from the data do not exist.

In this study users are not filtered in a preprocessing step. Rather user's behaviour is observed by means of common strategics and one novel approach. (1) According to the definition of personal and nonpersonal users, a classification scheme is set up to manually classify users accordingly. (2) In the context of the contribution bias the user activity is observed. (3) Bot and prevailing service accounts are classified with the help of tokens in their user name into semantic classes such as 'traffic' or 'weather'. (4) A new approach is tested that analyses the spatial distribution of geotweets per user. The spatial distribution of a user's geolocations is tested for Complete Spatial Randomness (CSR). Randomly distributed points do not bear any relevant geographic information. The spatial overlap of human settlements and Twitter data is analysed regarding these characterisations.

Geographic bias

The geographic distribution of tweets is highly biased. It is assumed that tweets are overrepresented in cities. Twitter penetration is uneven amongst countries. A global study needs to carefully regard that overall data is only representing parts of the world. The patterns under study have to be analysed independently in different geographic regions.

In this study observed phenomena such as spatial overlap of GUF and Twitter data are analysed separately in spatial regions in form of regular grids. Furthermore, for the first time the often claimed discrepancy between Twitter usage in cities and rural regions with geolocated data is investigated. This is done by comparing the Twitter density with human settlement size. Last, the geographic bias is incorporated in the estimation of the potential of Twitter data as classifier for human settlements. The calculations consider, that tweets are often close to other tweets.

Geotag precision and accuracy

Geotags can be created with physical localization sensors or by means of artificial manipulation. The linked chain from geotag generation by a user to geotag retrieval via Twitter's API is traversing Twitter's obfuscation process. A study that assumes GPS-generated geotags must be aware that imprecise (error from positioning system), inaccurate (resampled geolocation) or plainly meaningless geolocations can be present and even account for the bulk of the data.

In this study the identical geolocations of tweets are summarized. Tweet locations are also put into the context of surrounding tweets. The nearest neighbour distance for every tweet is calculated and its relevance for the spatial overlap with human settlements is assessed.

Replicability

Results from large-scale studies are not comparable due to the lack of global studies. It is unknown to what degree characteristics of Twitter data vary in space and to what extent the process of querying Twitter data is leading to different results. A replicable methodological framework allows its result to be approximated with independently sampled data. Results from spatial Twitter data research can be compared, if methods and data are replicable.

In this study the patterns found in the data are analysed globally. In doing so the degree of spatial variation can be assessed. Further, two independently sampled data sets are applied and the results are compared. One data set is of global coverage, the other only over Europe. If the results turn out to be unequal, replicability of the results from Twitter data studies in general have to be questioned. Either the data shows varying patterns in space, or the data querying process and preprocessing might largely impact the data.

Temporal patterns

Researchers have made use of the fact that land use and land cover classes show differing patterns of human activity (Longley, Adnan, and Lansley, 2015), what is reflected in Twitter activity signatures. Therefore tweets have to be analysed relative to their temporal context.

The spatial overlap of Twitter data with GUF is evaluated temporally. Tweets are therefore aggregated to hourly intervals and to days of week. The observations are compared spatially. It is expected that these temporal patterns of spatial overlap with GUF are similar in close areas. The patterns of overlap of Twitter data and human settlement are also compared with the spatial distribution of Twitter activity signatures.

Chapter 2

Methods

2.1 Data

2.1.1 Global Urban Footprint (GUF)

GUF is a binary raster data set that depicts human settlements globally at a resolution of 0.4 arcsec (12 m). It is derived from high-resolution TanDEM-X and TerraSAR-X radar images captured in 2011-2012. The authors (Esch et al., 2017) assess an overall accuracy for GUF of 85% which is unprecedented for this kind of product. For practical reasons a resampled version of 2.8 arcsec (84 m) resolution is applied.

Additionally, a vectorized version of the settlements is derived. The settlement pixels are dissolved so that a set of neighbouring positive pixels constitute a single polygon. A polygon represents an area of connected human settlement (patch) but does not hold any underlying semantic value (e. g. city). There are 11'289'316 settlement patches. These vary in size from $1'267\,m^2$ (one single pixel) to $4.17\,B\,m^2$.

The shape of the derived polygons is arbitrary, especially in terms of its size. The area covered by a patch is not necessarily representative of the degree of urbanization in a certain location. For example a river may divide a single city in two halves. Cities with no river on the other hand are not separated. Hence the size of a patch is only a very limited dimension. In order to add a measure of urbanization to the arbitrarily defined patch extent, a new attribute is added to this data. This attribute is the percentage of settlement area within a radius of 10 km from the centroid of a patch (see Figure 2.1). It embeds the patch into its environment. A patch of certain size in an urban area gets a much higher value than a patch of the same size in a rural context.

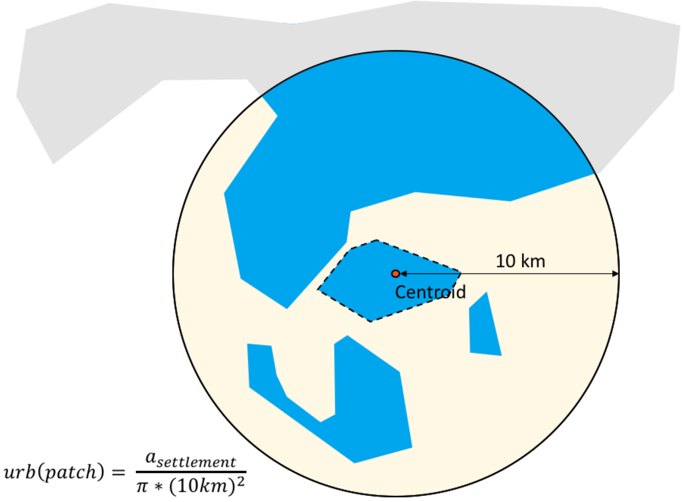


Figure 2.1: Sketch for the calculation of the degree of urbanization: The degree of urbanization of the dashed patch ($urb(patch)$) is given by the share of settled area (blue) within a 10km buffer around its centroid.

2.1.2 Twitter data

Twitter data is downloaded via the freely accessible Twitter Streaming API[1]. It is possible to download a limited amount of maximally 1% (Morstatter et al., 2013) of the live tweet volume that meet requested properties (e. g. keywords). For the purpose of this thesis, data is queried by geolocation, which returns georeferenced tweets only. These can be precisely located or obfuscated tweets. How Twitter selects the subset of returned tweets is not clear. Since the share of georeferenced tweets is around 1%, it can be assumed that most tweets meeting the requested properties are returned.

There are two distinct Twitter datasets employed. (1) Data from a query covering the greater area of Europe (T_{eu}) and (2) a dataset at full global coverage (T_{world}). T_{eu} was collected from 2017-05-18 12:23 to 2017-10-07 14:42 querying the parameters *message, user name, user id, user location, place type, place name, place country, timestamp, precise location, bounding box.* The geographical extent of the query is defined by a bounding box from -29.0 Lng / 33.0 Lat to 51.0 Lng / 72.0 Lat. 137'082'879 tweets were downloaded in this time period. Most of the geolocated tweets are obfuscated to city, country or continent level. The number of remaining precisely located tweets is 18'781'571. T_{world} was collected from 2016-04-21 10:25 to 2017-01-20 10:08. The queried attributes are limited to *userid, time in GMT and Long/Lat coordinates.* The subset of precisely georeferenced data contains 17'852'042 distinct tweets. Figure 2.2 summarizes the sampling periods of T_{world} and T_{eu} along with the daily amount of tweets collected. Significant drops in the volume are due to outages of the client computer during the sampling process.

Generally the whole data set is considered for analysis and hardly any cleaning is performed. However

[1] https://dev.twitter.com/streaming/overview, accessed 2017-11-11

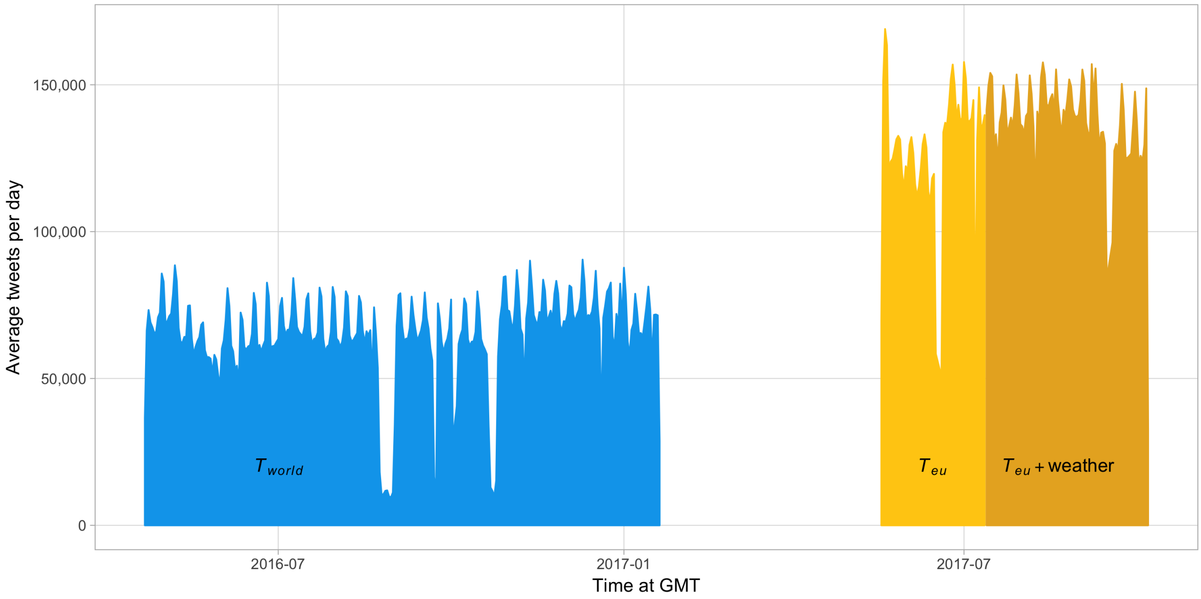


Figure 2.2: Daily tweet volumes over the sampling period of the two datasets T_{world} (blue) and T_{eu}. Weather data sampling period started later than T_{eu}.

two types of cleansing by geolocation are executed. First, in T_{world} there is a pile of tweets that are located exactly on the north or south pole. These tweets are removed from the data, since they are arguably artefact geolocations. Second, in T_{eu} there are tweets outside the queried area. These are tweets with both precise and obfuscated geolocations. The bounding box of the obfuscated location intersects the queried area and is therefore returned by Twitter's API. Since the precise location is outside the queried area, they are removed (175'618 tweets).

There are 17'852'042 distinct tweets in raw T_{world}, but only 7'148'044 distinct geolocations. In T_{eu} the ratio is even more striking with 3'303'659 out of 18'605'953 tweets being distinct geolocations. The likelihood for two GPS-located points to be in the exact same location twice is very low. These points are very likely to be automatically generated or derived from inaccurate measurement or resampling methods. From the response of the Twitter Streaming API obfuscated and GPS-located geolocation can't be discerned. For purely spatial analyses, duplicate geolocations are therefore regarded as one single location.

2.1.3 Other data

Weather data

Weather condition was downloaded to accompany T_{eu} from OpenWeatherMap[2]. This is a platform that allows to query live weather data for requested cities by bounding box. The following parameters were

[2] https://openweathermap.org, accessed 2017-08-01

queried: *City Name, Snow, Rain, Wind, Clouds, Temperature, Timestamp, Geometry* for the same area like T_{eu} from 2017-07-13 07:04 to 2017-10-07 14:42. In Figure 2.2 the period of weather data sampling is indicated. The weather conditions are given per city (7'372 cities in total).

Time zones

Twitter and weather data timestamps are given in Greenwich Mean Time (GMT). In order to obtain the local time the data has to be spatially joined with world time zones. The necessary geodata was acquired from NaturalEarthData[3].

2.2 Spatiotemporal patterns

Twittter activity signatures

Geotemporal patterns in terms of tweet volume are assessed with the given Twitter data. On the temporal axis, the data is analysed at resolutions of (a) hour of day and (b) weekday. Successively, the patterns are analysed spatially. In order to analyse temporal patterns in space, the study area is split into a regular grid of 300 x 300 km (at equator) for T_{world} and a grid of 200 x 200 km for T_{eu}.[4] Twitter data outside land masses is ignored for two reasons. On one hand over the sea there are no human settlements. On the other hand there are generally not enough tweets for a representative result. The grid is hence clipped to the land masses, resulting in grid cells of varying shape and size. All calculations regarding area have to be normalized to the area of the grid cell. Grid cells containing less than a minimum of 200 tweets are also considered unrepresentative and therefore neglected.

The Twitter volumes over the course of a day form a spatially characteristic pattern (García-Palomares et al., 2018). To verify this with the data at hand, two methods are applied to characterise diurnal patterns. The tweets are therefore aggregated by hour. First, the hour of maximum Twitter traffic (peak hour) is extracted for every grid cell. This is a very simple description for the diurnal Twitter pattern. Second, the whole diurnal pattern is clustered. Frias-Martinez, Soto, et al. (2012) propose to use a k-means algorithm to find similar diurnal Twitter activity patterns. The k-means input is a vector of 24 variables per sample: $0, 1, \ldots, 23$, where 0 is the amount of tweets per grid cell between 00.00 and 01.00. The diurnal pattern is normalized in order to account for varying absolute tweet volumes in different grid cells. The normalized values are the percentage of tweets relative to the amount of tweets at peak hour. E. g. a pattern with peak hour at 12:00 with 1'000 tweets and 900 tweets at 13:00 gets the value 1 at peak and 0.9 at 13:00. These normalized Twitter activity values correlate. E. g. during the night Twitter volume is low everywhere. During the day, there are higher variabilities. Figure 2.3 a) illustrates this by means of hourly boxplots over all grid cells. Therefore the variables are decorrelated using Principal Component Analysis (PCA). The five first principal components explain 86.5% of the variance (Figure

[3] http://www.naturalearthdata.com/downloads/10m-cultural-vectors/timezones, accessed: 2017-06-22

[4] The grids were created in Universal Transverse Mercator projection (EPSG:3857).

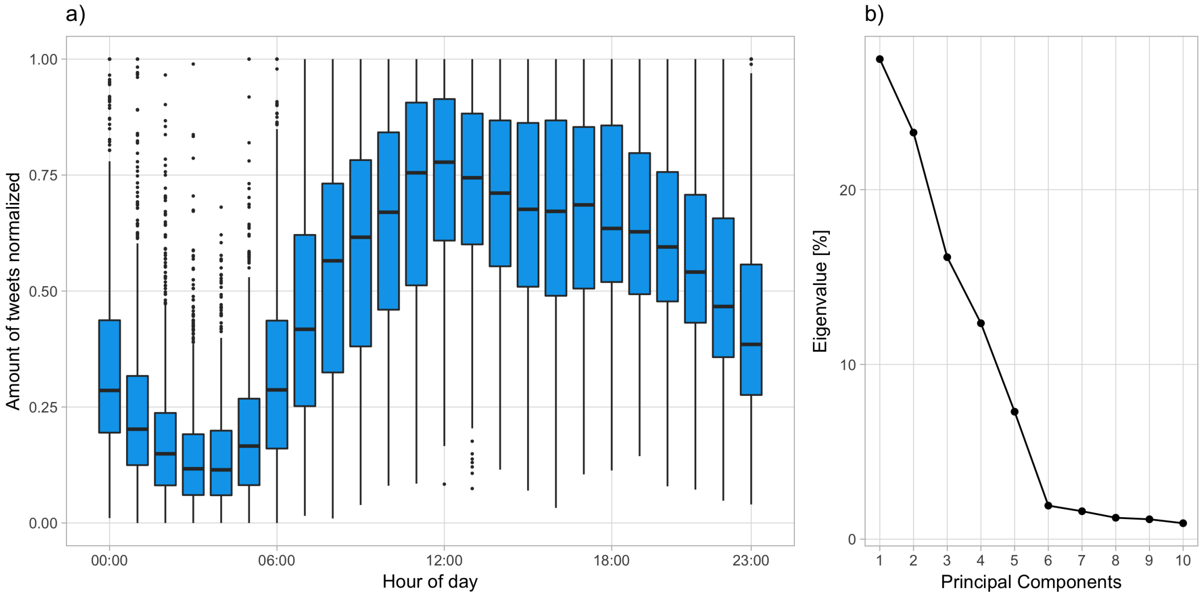


Figure 2.3: a) Distribution of the hourly volumes by grid cell. The volume of Twitter data per grid cell is normalized from 0 to 1. The hour of day where grid cells have the highest amount of Twitter data (relative to the absolute amount per grid cell) is 12:00. **b)** Result from PCA: Eigenvalues of the dimensions with highest load.

2.3 b)).

The five selected principal components are then clustered with a k-means algorithm. The number of clusters is defined by looking at the behaviour of within cluster sum of squares (WSS). This measure is an indicator for the homogeneity of the clusters obtained. Putting all values into one cluster leads to a low homogeneity, many clusters lead to more homogeneous clusters, but on the downside also to more clusters. The optimal value is defined where WSS is not further lowered significantly at increasing clusters. In this case six clusters seem to reflect the structure of the data. The k-means algorithm is hence run for six clusters. As a result, a cluster is assigned to every grid cell.

Spatiotemporal patterns of overlap with GUF

The spatial overlap of GUF and tweets is given as the share of tweets on settlements by the total amount of tweets in a specified area. To assess how far the geotemporal patterns are related to the correspondence with human settlements, tweets are spatially joined to GUF. Every tweet henceforth belongs to one of the two classes 0 (outside settlement) and 1 (on settlement) depending on its relative location to GUF. Tweets are then aggregated to hours and weekdays. The spatial overlap of Twitter and settlement by hour of day and weekday is then examined spatially by grid cell.

2.3 Classification of Users

Users are differentiated in several ways: (1) by manually checking accounts, (2) by user name, (3) by number of tweets and (4) by point pattern distribution.

Classification by hand

Users are classified as personal and nonpersonal as defined in Chapter 1. This is done for T_{world} only. This method leaves full control over the classification process to the author and gives deep insight to the data set. Selected Twitter accounts are inspected by the author on Twitter's website. Indicators for nonpersonal accounts considered are (1) (very) high tweet frequency, (2) similar tweet structure, (3) content, e. g. advertisement, (4) regular intervals of tweets and (5) number of followers.

(1) Automated accounts often post tweets in intervals of seconds to several minutes or hours. Personal users normally do not exhibit such high Twitter activity. One exception worth mentioning are users that deploy third-party geo-applications that have access to his/her Twitter account. The most prevalent application found is swarmapp.com[5]. This app automatically tweets the user's location. Because real people's activity is posted, these users are classified as personal, despite the high tweet frequency. (2) Both automated accounts as well as services tend to post tweets in an predefined format. For example accounts that push meteorological measurements automatically show a hard-coded tweet structure. Likewise services that are not fully automated like job advertisers follow a predefined format. (3) The content of the tweets is probably the most meaningful but also the most arbitrary indicator. Many accounts can clearly be assigned to a specific purpose they serve. Examples are meteo stations, traffic news feeder or job advertisers. Others definitely show human thought behind every tweet. But many accounts lie between the two and leave room for speculation. For example some users have a meteo measurement feeder installed on their personal account. (4) Automated accounts often exhibit not only high posting frequencies, but also very regular intervals. (5) Normally, automated users and services have fewer followers in relation to the amount of tweets they post.

These five factors are considered separately and rated subjectively. In a first step a subset of 200 randomly selected users is classified. However, the chances to hit a nonpersonal user are very low, indicating that most users are actually personal users. In a second attempt the users are listed alongside their corresponding number of tweets in the dataset and ordered from the most to least active user. It is sensible that the most active tweeters are usually nonpersonal users. For the sake of illustration, the most active users are examined. The top ten of this list are given in Table 2.1. These users look very much like nonpersonal users by just looking at the user name. Manual inspection verifies this presumption. *511NY* is a traffic information website from New York, *SONICjobs* an employment adder, *EveryFinnishNo* claims to post every number in Finnish (be this possible or not). The tweets of the secluded leader of this ranking *googuns_ lulz* are randomly generated content[6]. In total the 1'500 most

[5] https://www.swarmapp.com

[6] http://victorz.ca/bots/googuns, accessed 2017-08-14, 16:22

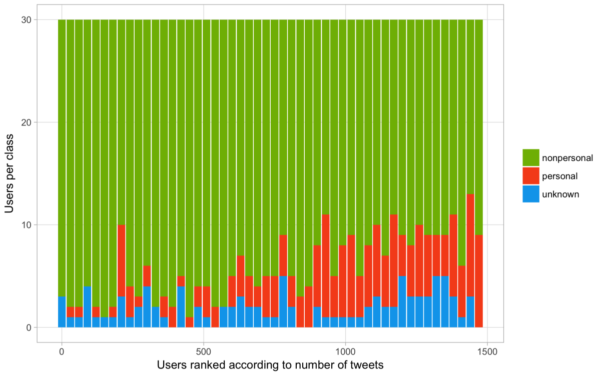


Figure 2.4: Share of personal, nonpersonal and unknown user accounts of top 1'500 ordered by user activity (bin width = 30 users).

active users are classified as *personal* and *nonpersonal*.

The manual classification of users bears three disadvantages. First, the distinction is hard to operationalise and hence results in subjective judgements. Second, the process of classification is time consuming. The results can hardly be transferred to other Twitter data sets, as the users may change over time and with the study area. And third, the semantics of *personal* and *nonpersonal* might not be an adequate schema. There are nonpersonal Twitter accounts that post meaningful locations, e. g. firefighter operation locations or real estate locations. For these reasons the users are additionally handled by automatable and reproducible means.

	T_{world}	n Tweets	T_{eu}	n Tweets
1	googuns lulz	41'222	Every Finnish Number	153'379
2	511NY	30'674	OV Radar	116'159
3	infosrv	28'954	Solar Realtime Edent	109'945
4	Chatter ng	22'466	BrugOpen	106'459
5	EveryFinnishNo	20'271	Pen-Y-Renglyn	105'301
6	PenYRenglyn	19'714	Trendinalia España	65'641
7	VirtualJukebox	19'391	Trendinalia UK	59'019
8	SONICjobs	18'142	L'hora catalana	46'640
9	propertiesindia	18'034	infosrv	46'209
10	slappervader	16'889	Trendinalia France	43'760

Table 2.1: Absolute number of tweets of the top 10 users in both data sets.

Classification by number of tweets

The classification by hand has revealed that nonpersonal users are found mainly among the frequent tweeters. As a mean of overcoming tedious manual work and in the context of contribution bias, users can simply be analysed by their activity. In Figure 2.4 the share of nonpersonal users is shown along the user activity. In the top 1'500 most active users it can be seen that the probability for a user to be personal increases with a decrease of activity, a trend that is expected to proceed further into less active users.

Ranking users according to their activity is hence expected to embed the user among similar users. The correspondence of Twitter data with GUF might be affected by the characteristics ordered like this.

Classification by user name

Besides the distinction between personal and nonpersonal, Twitter accounts can be grouped into more detailed classes. Many of the nonpersonal accounts follow a clear purpose. E. g. some are posting job opportunities, others weather information or current Twitter trends. The prevailing groups of accounts are assessed by manual inspection of the tweets and given in Table 2.2. It is assessed whether different groups exhibit different spatial overlap with settlements.

In order to make the process of group extraction fast and as objective as possible, the user names of the respective accounts are scanned for keywords. Regular expressions (regex) allow to extract substrings out of strings such as user names. The tokens corresponding to each category are given in Table 2.2. The tokens are also collected during the course of manual Twitter account inspection.

Category	Token
trendinalia	trendinalia, trendsmap
job	job, career, tmj, work
weather	weather, meteo
news	feed, news, reports, info, live
traffic	traffic, road, travel
bot	bot

Table 2.2: Categories and corresponding keywords as derived from manual classification.

Classification by geographic point pattern distribution

T_{world} and T_{eu} clearly are no randomly distributed point processes. Notwithstanding, single users often show a random or dispersed behaviour. These are most likely to be automated bots or services. In any case random point processes are never to reveal any underlying spatial structure. In order to detect these accounts, users with more than 50 distinct tweet locations are classified according to spatial distribution.

The null hypothesis in the test is *'The point pattern is random or regular'*. Hence a one-side test for CSR *or* regularity against clustering is applied. The patterns of single user's geolocations are very diverse. Typical users have many clustered tweets in the surroundings of their home town and some outliers in far-away destinations. Common tests for CSR such as the Nearest Neighbour Index (NNI) are relying on *mean* NND. The NNI requires a certain degree of evenness in the distribution of the points, which is not guaranteed in Twitter user's geolocations. Therefore the following median-based approach is executed:

The observed NND (P_{obs}) is compared to the NND of an expected random point pattern (P_{exp}) in the area of the minimum enclosing rectangle of the observed point pattern. As of observation, P_{exp} follows a Poisson distribution with the following form $P_{exp}(x) = \lambda^x * e^{-\lambda}/x!$, where λ is the median of $P_{exp}(x)$. An estimate of P_{exp} at the extent of P_{obs} is obtained through a simulation of 30 randomly generated point processes. From these the average median NND λ is derived. We seek for the probability that the median of P_{obs} is lower or equal to λ ($P(\lambda \leq \lambda_{obs})$). This is then given as $F_x(\lambda_{obs})$, where $F_x = \int P_{exp}(x)dt$. Hence the resulting p-value is interpreted as an indicator for the probability at which P_{obs} follows P_{exp} and can be considered random.

2.4 Point distribution

Tweets tend to cluster in urban areas. Spatial outliers may indicate places where people rarely go and that are probably not settled. An easily implementable and interpretable measure for spatial outliers is the nearest neighbour distance (NND). High relative NND values indicate outliers. For both datasets the NND is calculated independently for all distinct points. Hence NND = 0 is not possible. The NND of the tweets is compared to their position relative to GUF.

2.5 Influence of Weather

Weather parameters on OpenWeatherMap can only be requested in limited frequency. The weather however doesn't change rapidly and the weather parameters on OpenWeatherMap are only updated in intervals of approximately three hours. Therefore data is queried every three hours. The data can only be requested for selected places (7'372 distinct places in the queried area). It should be noted that these locations are not necessarily set according to the location of meteo stations, but rather according to cities. The values are interpolated from physical measurements.

The weather condition in the moment of the tweet t and a location l hence needs to be estimated. This is accomplished in two steps: (1) Linear interpolation along the time axis and (2) subsequent Inverse Distance Weighting (IDW) interpolation in terms of space. Let the collective of weather value locations be S where $s_i \in S$ represents a distinct location and $i = \{1, \ldots, 7'372\}$. Each s_i is defined by a unique location l_i and holds a number of weather measurements $m_i = (precipitation, wind, temperature, clouds, snow)$ at a given time. A single tweet is also given with coordinates l and timestamp t. For this purpose the 5 nearest neighbouring weather locations S_{nn} are selected. From these the measurement shortest before $m_i(0)$ and after $m_i(1)$ the timestamp of the tweet are selected. Linear interpolation of the measured values $m_i(0)$ and $m_i(1)$ over time is conducted to get an estimate of the weather conditions at the time of tweeting t. The value of an linearly interpolated value is given as:

$$m^*(t) = m_i(0) + (t - t_i(0)) * \frac{m_i(1) - m_i(0)}{t_i(1) - t_i(0)}$$

, where $t_i(0)$ is the time of the measurement right before the time of the tweet and $t_i(1)$ right after. Then the estimated values m^* at t of the k nearest weather data points are used to do Inverse Distance Weighting interpolation (IDW). This method allows take into account the distance of s_i from the tweet location in weighing the value of each station with the inverse of its squared distance. The formula applied looks like following:

$$z^*(x) = \begin{cases} \dfrac{\sum_{i=1}^{N} m^* * d(x,x_i)^{-2}}{\sum_{i=1}^{N} d(x,x_i)^{-2}} & \text{if } d(x,x_i) > 0 \\ z_i & \text{if } d(x,x_i) = 0 \end{cases}$$

, where $z^*(x)$ is the interpolated value for a tweet, m^* is the estimated weather value at t_0 at weather station i and $d(x,x_i)$ is the distance from the tweet to weather point i. The obtained values are then analysed relative to the position of tweets.

2.6 Classification of Tweets

With the above characteristics, it is attempted to identify tweets that are likely to be outside human settlement. Yet, the numerous characteristics do not provide any outstandingly clear exclusion criteria. Some collinearity exists. It is therefore unfeasible to use hard classification criteria per class. Instead a random forest (RF) model is trained (Breiman, 2001). RF is a supervised classification algorithm that performs well with noisy input data. The procedure is based on random decision trees. Every tree consists of a random sample of test data that predicts the output class based on given input variables. In this case the inputs are derived from the above discussed characteristics. Following inputs are generated per tweet: *continent, user type by regex, p-value of user point distribution, bounding box area of user point distribution, number of tweets per user, number of identical points to this tweet, hour of day, month, weekday, NND.* The model is trained for the corresponding GUF class where a tweet is placed on. A random sample of 5% of the data sets is used to train the model. There are 50 trees generated per model. The model is validated using 10 independently sampled subsets of the data.

2.7 The role of GUF patch size

The above implemented measures are all aiming to describe a single tweet's probability of being on built-up area. However, tweets are geographically biased. It has been described that Twitter data is restricted to certain geographic regions of the world (Leetaru et al., 2013). Likewise, it is assumed that Twitter data is mainly found in urban areas rather than in rural towns. To test this assumption the patch size and the degree of urbanization per patch is compared with the amount of tweets contained in the respective patch. Due to unevenly distributed Twitter data, a set of models is calculated for different geographic locations. If the tweets are evenly distributed, the number of tweets per patch is dependent on the patch size. Hence the density would be the same everywhere. It is expected that the tweet density of large patches and patches with a high degree of urbanization is higher than the density in small and less urbanized patches.

2.8 Detection of Settlement

Explain assumptions The given data does show where settlements are. A problematic issue is that it doesn't show all settlements. Therefore it is attempted to estimate how much data is needed to cover all human settlements. Settlement pixels that are covered with at least one tweet are henceforth referred to as *detected pixels*. It is assumed that the overlap of tweets and settlement area is constant. Under this assumption, the amount of tweets that have to be collected to detect all pixels of a given raster with settlement is calculated. If tweets were evenly distributed, a linear relationship between the amount of tweets and the amount of detected pixels can be expected. However, tweets are geographically biased, therefore duplicate and close-by tweets are present in the data. These don't contribute to the detection of more settlement areas. Therefore the amount of duplicate geolocations in different data set sizes needs to be estimated. For this purpose a Monte Carlo simulation is run on subsets of the given data. For every size the average number of detected and duplicate /close-by geolocations is computed.

Close-by is here defined as half the resolution of GUF ($= 84\ m/2 = 42\ m$). Two tweets that are closer than 42m are more likely to be in the same pixel than in different pixels. Out of some close-by tweets all but one tweet are regarded as close-by. With this value the detection rate of the Twitter data sets is modelled. For a given quantity q of Twitter data t_q the number of detected raster cells t_{det} is $t_q - t_{guf=0} - t_{close} - t_{dupl}$, where t_{close} is the number of close-by geolocations that are on settlement, t_{dupl} is the number of identical geolocations that are on settlement and $t_{guf=0}$ is the number of tweets outside settled area. t_{det} is estimated through observation of randomly selected subsets of the data at hand of varying size. The share of t_{det} is then defined as $t_r = t_{det}/t_q$. t_r is assumed to decrease with higher volumes due to an increase in t_{dupl} and t_{close}.

We then fit the detection rate $t_{det} \sim t_q$ to a non-linear model. t_{det} follows an exponential function of the form $t_{det} = a * t_q^b$, where a and b are the parameters to be estimated. Non-linear least square analysis is performed to obtain these. The resulting values for a and b define a function $t_{det}(t_q)$. The amount of Twitter data required to detect all pixels of human settlement can be estimated analytically at $t_{det}(n_{pixels})$.

Chapter 3

Results

The spatial overlap of Twitter data and GUF is henceforth given in percentage of tweets overlapping GUF settlement from all selected tweets. The overall spatial overlap of GUF and Twitter data is 84.1% for T_{world} and 79.9% for T_{eu}. The significant, albeit small difference of the two data sets can partly be explained by the extent of the sampling areas. A subset of T_{world} within the spatial extent of T_{eu} (henceforth abbreviated $T_{world}(EU)$) has an overall overlap of 83.1%. The correlation in European countries is hence lower than elsewhere. However there remains a considerable offset between the two datasets. The aforementioned numbers relate to distinct tweets. When summarizing identical points, the spatial overlap of T_{world} is 83.7% and 76.5% for T_{eu}. Also $T_{world}(EU)$ has a lower overlap of 81.4% when unique geolocations are considered. These numbers resemble more impressive when relating it to the chance of random points to overlap with settlement. For example the area of the queried area of T_{eu} is $1.9 * 10^7 km^2$ while only $2.6 * 10^5 km^2$ are covered with built-up area. The spatial overlap of a random point pattern would be 1.3%. Tweets outside built-up area tend to be in the vicinity of settlements. In T_{world} 38.3% of all tweets outside settlements are within 84 meters from settlement, which is the GUF raster resolution. In T_{eu} there are 34.4% of the tweets outside settlement within 84 meters. In sum there are 90.7% (T_{world}) and 84.1% (T_{eu}) of all distinct geolocation on settlement or as far away as the resolution of GUF. Table 3.1 summarizes the key data of the two data sets.

	Spatial extent	Sampled days	Total tweets	Distinct tweets	Overlap	Overlap dist.
T_{world}	World	275	17.9 M	7.1 M	84.1 %	83.7 %
T_{eu}	Europe	143	18.6 M	3.3 M	79.9 %	76.5 %
$T_{world}(EU)$	Europe	275	3.7 M	1.5 M	83.1 %	81.4 %

Table 3.1: Overview of the two independently sampled Twitter data sets T_{eu} and T_{world} and the tweets from T_{world} within the spatial extent of T_{eu} ($T_{world}(EU)$). Overlap is the spatial overlap of GUF and the Twitter data set, overlap dist. is the overlap of distinct geolocations only.

3.1 Spatiotemporal patterns

3.1.1 Geographical distribution

Figure 3.1 gives an overview of the spatial distribution of the two Twitter data sets. The spatial distribution of the tweets T_{world} across the globe is uneven in several regards. On a global level the majority of the tweets is concentrated in Europe, North America, Indonesia, Japan and parts of South America.

The tweets T_{eu} are largely congruent to T_{world}, with the exception that the density is 2.3 times higher. It contains 3'303'659 distinct tweet locations as compared to 1'455'984 in the same area of T_{world}. Western European countries show higher tweet densities, particularly the Netherlands and UK. The tweets tend to be in densely populated areas, especially cities. There are differences on a small scale between national boundaries. E. g. the Netherlands show a clearly higher tweet density than its neighbouring countries. In the area of Finland there is an area of dense tweets of rectangular shape in both data sets that are generated by a bot.

3.1.2 Temporal Twitter activity signatures

Temporal patterns in the volume of Twitter data can be observed at weekly and daily time scales. Figure 2.2 depicts the number of tweets received per day during the sampling period. Aside from the heavy drops due to client server outages, two distinctive phenomena can be discerned: (1) There are variations in the overall amount of tweets gathered. These can't be related to any obvious cause. T_{eu} has on average a higher sampling rate than T_{world} by the factor 2. On an average fully sampled day (without technical problems) T_{world} receives 68'728 tweets and T_{eu} 137'523. (2) There are regular oscillations present almost throughout the whole sampling period. The frequency of these are the duration of one week. On weekends the amount of tweets is on average 26.4% (T_{world}) and 15.8% (T_{eu}) higher than during weekdays. The weekday of minimal Twitter traffic is in both datasets Monday with 62'287 and 129'201 tweets for T_{world} and T_{eu}, respectively. The weekday with most Twitter traffic is Saturday (78'702 and 149'559 tweets).

On a daily time scale the tweet volume varies, too. In both datasets there is an overall minimum Twitter activity between 3 am and 4 am with an average of 646 and 1'200 tweets every day. The maximum in T_{world} has highest Twitter volumes at 5 pm to 6 pm with 4'467 tweets, where T_{eu} has its maximum from 20 pm to 21 pm with 9'463 tweets on average.

The Twitter activity by hour of day and day of week is given in Figure 3.2. From Monday to Friday there is a slight increase in Twitter volume. It can also be observed that on weekends the rise of Twitter activity is slightly delayed compared to weekdays. The two datasets show some differences. In T_{world}, Saturday and Sunday show a different pattern, but not so in T_{eu}. Furthermore, the amount of tweets in the evening in T_{eu} show a more accentuated peak than in T_{world}. The difference of minimum Twitter activity in the night and maximum Twitter activity during the day is in T_{eu} much higher than in T_{world}. $T_{world}(EU)$ exhibits yet another pattern of Twitter activity with weekdays peaking in the morning and

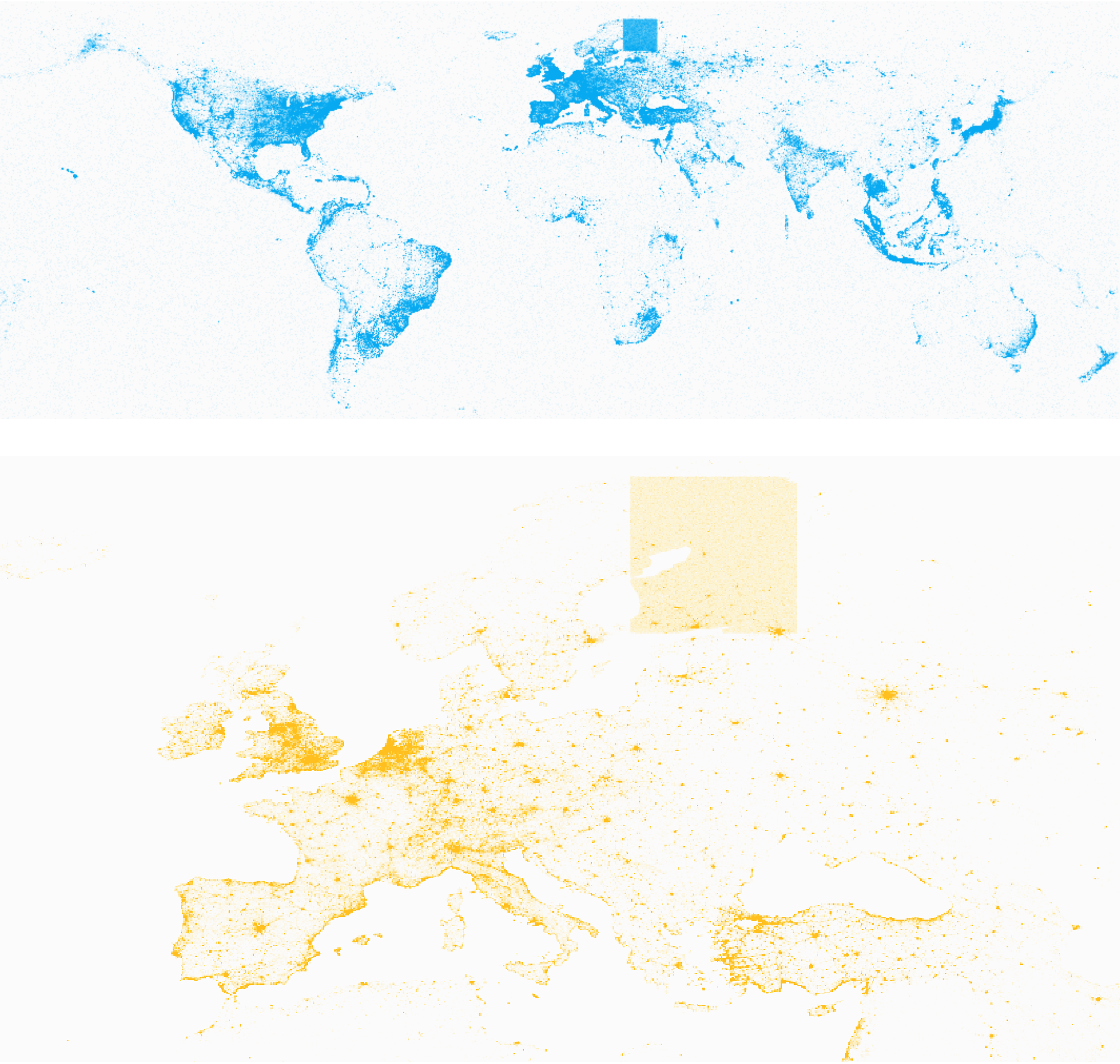


Figure 3.1: Spatial distribution of the tweets of T_{world} (blue) and T_{eu} (yellow).

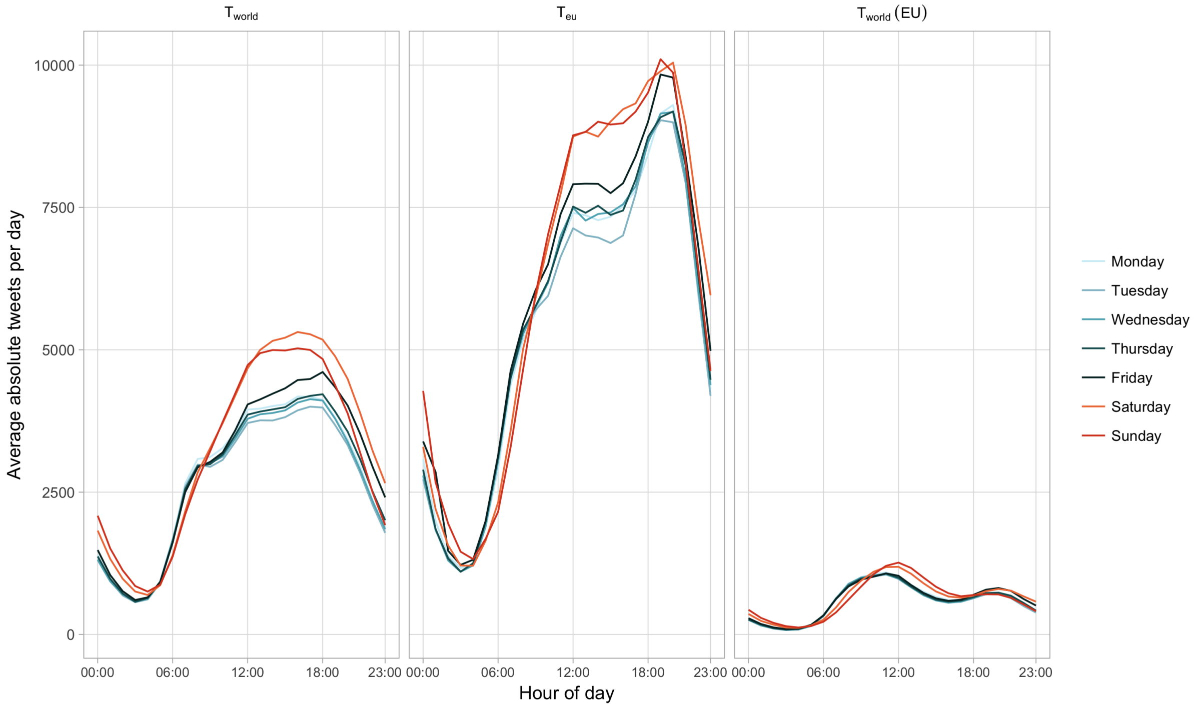


Figure 3.2: Daily tweets volume globally grouped by day of week for T_{world} (left) and T_{eu} (middle) and $T_{world}(EU)$. Only fully sampled days are used and average amounts are displayed.

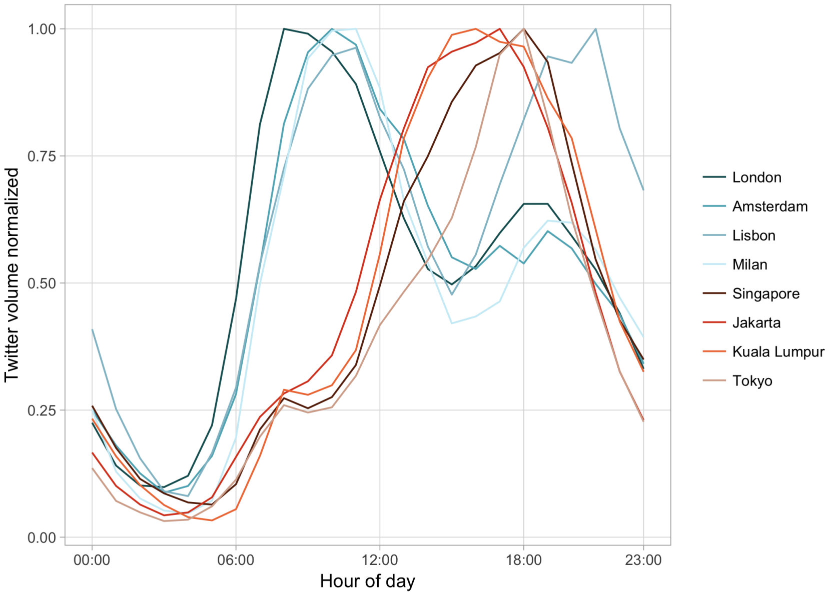


Figure 3.3: Daily tweets volume of four major cities in Western Europe and four major cities in Eastern Asia.

weekends peaking at noon.

3.1.3 Spatiotemporal variation

The temporal patterns vary in space. Figure 3.3 illustrates this phenomenon on hourly aggregated daily Twitter volume from T_{world}. The tweets of eight cities with a high Twitter activity are selected. Four cities lie in Eastern Asia (Singapore, Jakarta, Kuala Lumpur and Tokyo), four in Western Europe (London, Amsterdam, Milan and Lisbon). In order to make the differing total amounts of tweets comparable, the y-axis is normalized to 0 - 1. The two geographic regions group into two separate patterns. European cities peak around noon, then drop down in the afternoon before in the evening the activity rises again. Contrary the Asian cities peak in the late afternoon.

This behaviour is mapped in space. Figure 3.4 depicts the hour of day with maximum amount of tweets on a regular grid over the land surface. Grid cells with less than 200 tweets are considered unrepresentative and coloured gray. All other grid cells are coloured according to the hour of maximum Twitter volume starting from black at midnight to blue at 6 am to yellow at noon and to red at 6pm. Many observations can be done here: Globally there are four regions with considerable tweet densities. North America has maxima in the afternoon although being rather inconsistent. South America on the other hand peaks in the late evening and at midnight. Greater Europe has maxima before and around noon and (South-) East Asia in the evening. When zooming in there are more interesting patterns. Southern European countries tend to peak later in the afternoon than northern European countries. The Iberian west coast shows a very distinct pattern of late-evening peaks. Eastern Asia shows a later peak

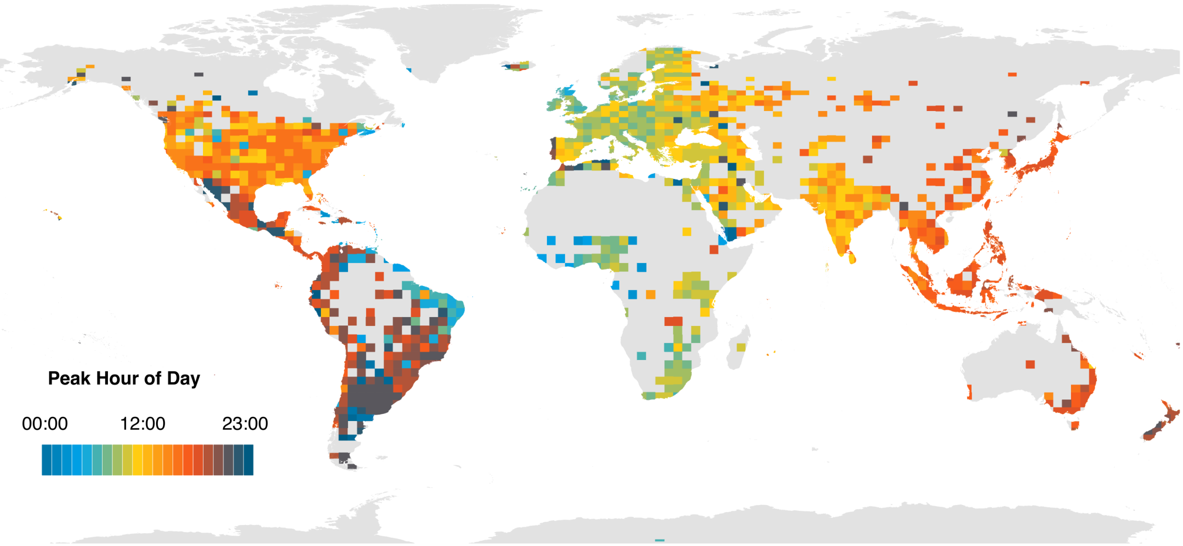


Figure 3.4: Twitter peak hours on a $300km^2$ grid over land.

in the eastern countries (Japan and Philippines). In South America there are three homogeneous regions. In eastern South America Twitter traffic peaks in the morning. South of it the peak is at 6 pm and in Southern Argentina it's at 21.00. Summed up, the peak hour of a grid cell tends to be similar to the peak hour of its neighbouring grid cells.

Clustering of the daily Twitter routine

The spatial correlation of the diurnal tweet pattern is even more pronounced when clustering the whole temporal routine. A k-means with five clusters and the first five PC's results in a spatially highly correlated set of grid cells (Figure 3.5). The United States form a homogeneous area (yellow). Latin America is divided into two clusters, one comprising the rough area of Argentina (orange) and one with the rest of Latin America including Central America (red). Europe and Africa including Turkey form a single cluster (green). India is in a single cluster, where parts of Russia are included (violet). The last cluster comprises Eastern Asia and Australia (blue). The cluster centres are plotted alongside the world map. The differences are, as expected, found during the day, in the night all clusters show a low activity.

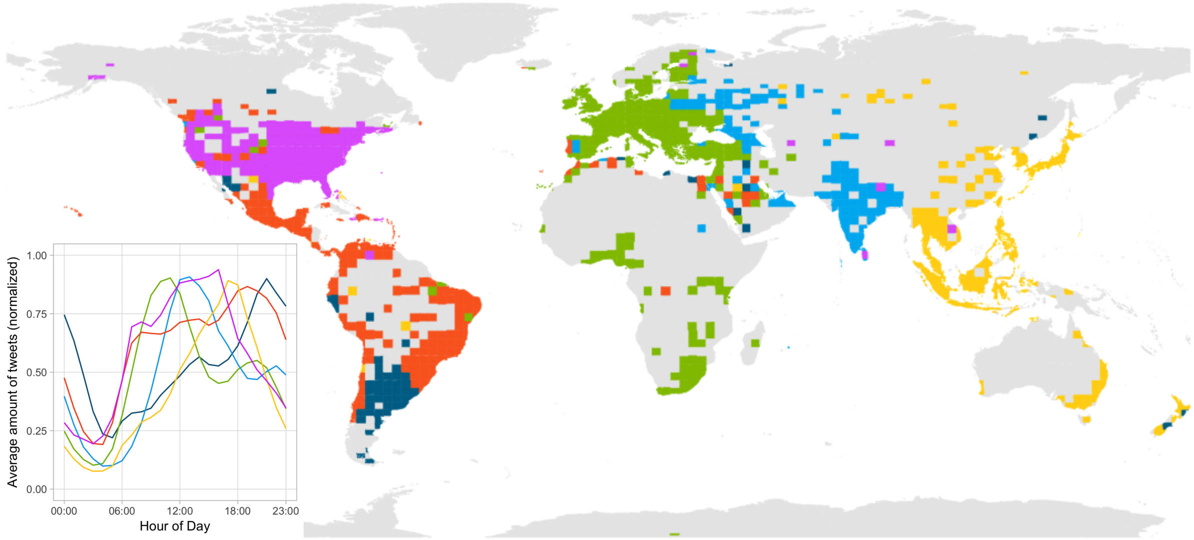


Figure 3.5: Cluster membership of grid cells as obtained from PCA 1 - 5. The according cluster centres are given.

3.1.4 Spatiotemporal overlap with GUF

The spatial overlap of Twitter data with GUF shows spatial as well as temporal trends. (1) Temporally, variations at diurnal, weekly and seasonal time scales are observed. (2) Spatially differing patterns on continent level can be seen.

Temporal patterns of spatial overlap with GUF

An aggregation of tweets by day and its correspondence with GUF is shown in Figure 3.6. Similar to the absolute volumes of tweets per day, there are regular oscillations present throughout the whole sampling period of both data sets. However, the peaks of this pattern correspond to the weekdays (contrary to the absolute volumes). Hence, on weekends the spatial overlap of GUF and tweets is lower than during the week. Clearly visible is also the difference in the magnitude of the oscillation: T_{world} has a smaller difference of correspondence with GUF between weekdays and weekends than T_{eu}. This phenomenon can largely be explained by differing geographic extents of the data sets. $T_{world}(EU)$ is mapped in light colour in Figure 3.6. The amplitude of this data is similar to the amplitude of T_{eu} indicating a stronger dependency of the spatial overlap of tweets and GUF on the day of week in Europe.

One more thing that can be observed in Figure 3.6 are periods of drop in spatial overlap in mid-year months. Again this is a special characteristic of European tweets, as $T_{world}(EU)$ and T_{eu} share a more heavy drop than T_{world}. It is plausible to relate these drops to the weather conditions that are dependent on seasons in Europe. But still the drop can be observed in T_{world} as well, which might be explained by the dominance of tweets from geographic areas with seasons (Europe, North America, Japan).

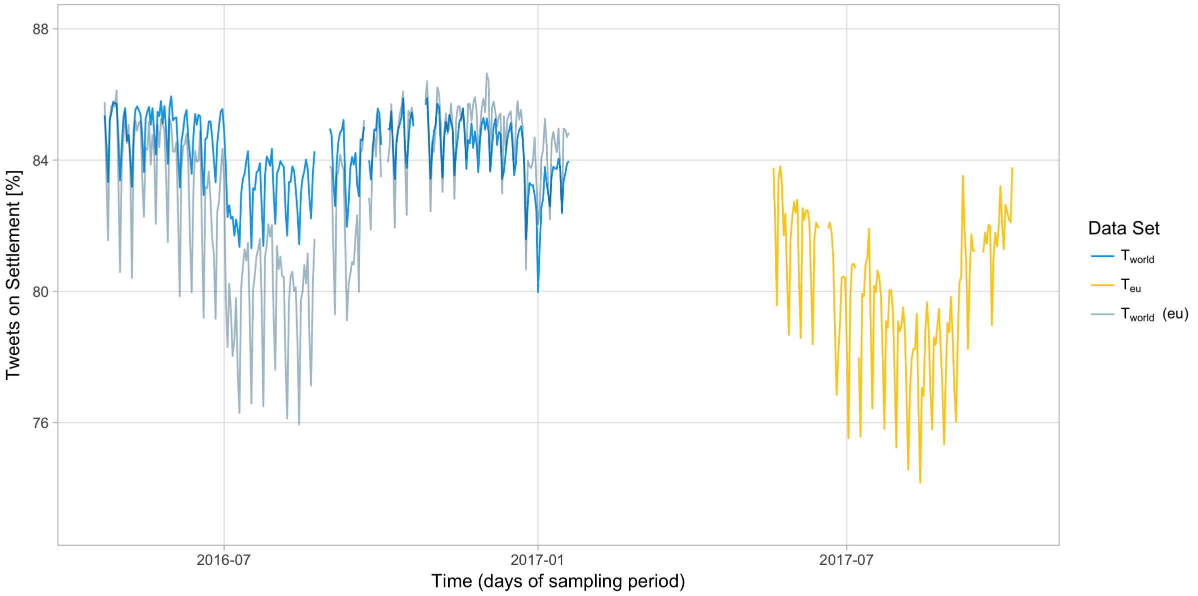


Figure 3.6: Spatial overlap of GUF and T_{eu} (yellow) and T_{world} over the sampling periods. The subset of T_{world} at the spatial extent of T_{eu} is drawn in light blue.

T_{world} experiences a drop in spatial overlap in the last few weeks. This drop can either be associated with a season dependent change of tweeting behaviour. This might also be due to changes in Twitter's privacy rules or default geotag sharing options. This would also explain the lower spatial overlap of tweets and GUF in T_{eu}.

The spatial overlap of GUF and tweet locations aggregated by hour of day and weekday are shown in Figure 3.7. The spatial overlap is lowest in the night (3am - 4am) in both datasets. Saturdays (orange) and Sundays (red) show a different pattern than weekdays: The minimum spatial overlap in the night is higher than the minimum spatial overlap on weekdays. During the day, spatial overlap stays low as compared to weekdays. On Sundays the correspondence is generally lower than on Saturdays.

There are considerable differences between the datasets. In T_{world}, weekends roughly follow the same pattern as weekdays: During the night the spatial overlap drops and rises constantly towards the late evening. In T_{eu}, Saturday and more thoroughly Sunday do show a pattern independent from weekdays' pattern. The minimum spatial overlap is on Sunday during the day. While on weekdays the drop of spatial overlap is by roughly 5% between 3 am and 7am, on weekends there is no change in spatial overlap of the same magnitude. The discrepancy can again be explained by the differing geographical area. $T_{world}(EU)$ shares the characteristic described for T_{eu}, with the exception that the overall spatial overlap is higher in $T_{world}(EU)$.

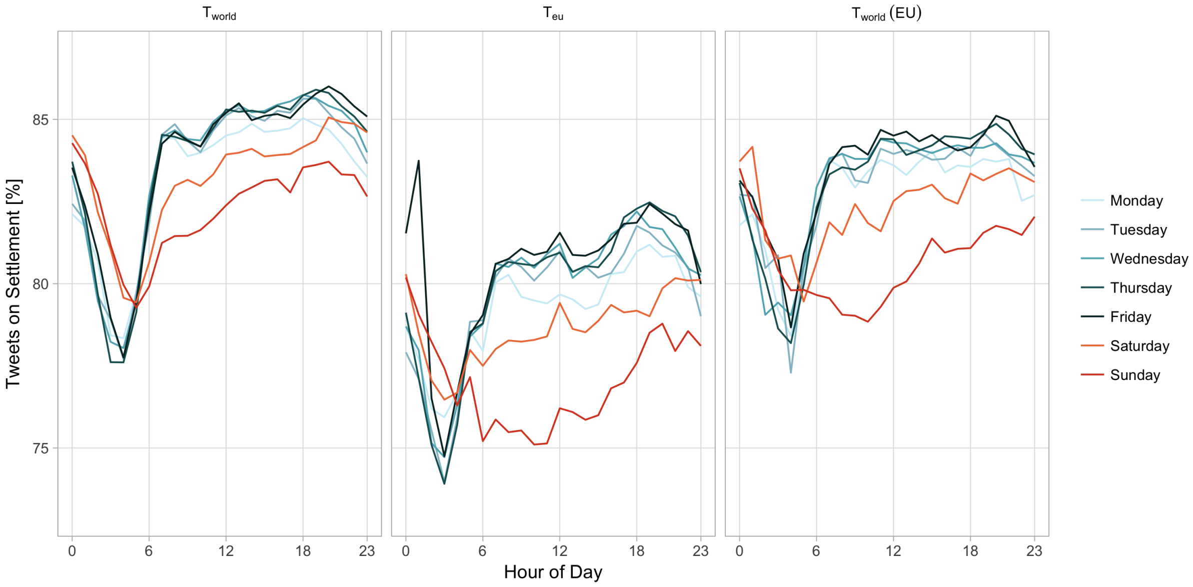


Figure 3.7: Percentage of tweets on settlement area for the two data sets on daily course, split by day of week. Although T_{eu} has a lower spatial overlap with GUF and different patterns than T_{world}. $T_{world}(EU)$ shows patterns similar to T_{eu}.

Geographic variation in spatial overlap

The different patterns in T_{eu} and T_{world} described above indicate that geography reveals structures in the data. In fact, the spatial overlap with GUF is not a spatially randomly distributed variable. In Figure 3.8 the percentage of distinct tweet locations on settlement per total distinct tweet locations is mapped on a grid. The grid cells are coloured orange when their value is below the median spatial overlap of all cells and green otherwise. The median is 77.5% in T_{world} and $<$ 69.0 % in T_{eu}. Coloured gray are cells with less than 200 distinct tweet locations.

T_{world}: It can be observed that mainly Latin American countries lie above the median. North America, India and parts of Europe are generally below the median. Northern Europe is dominated by the automated bot *Every Finnish Number* that posts random geolocations with a low spatial overlap with GUF. The spatial pattern of correspondence is not as distinctive as is the case for the temporal variation in tweet volume (Figure 3.4).

T_{eu}: In Scandinavia the spatial overlap is consistently below the median, not only in the area influenced by *Every Finnish Number*. For the rest, large areas of France, Spain, UK and Turkey show spatial overlap above median. Eastern Europe has both, areas of higher and areas of lower spatial overlap.

A comparison of the two datasets reveals consistent patterns in the area of Europe. The patterns described above are present in both data sets. But also patches of lower overlap in the area of the Alps, Slovenia and Western France are consistent. Eastern Europe has in both data sets grid cells above and below the median. The two data sets are almost congruent. The patterns present show non-random

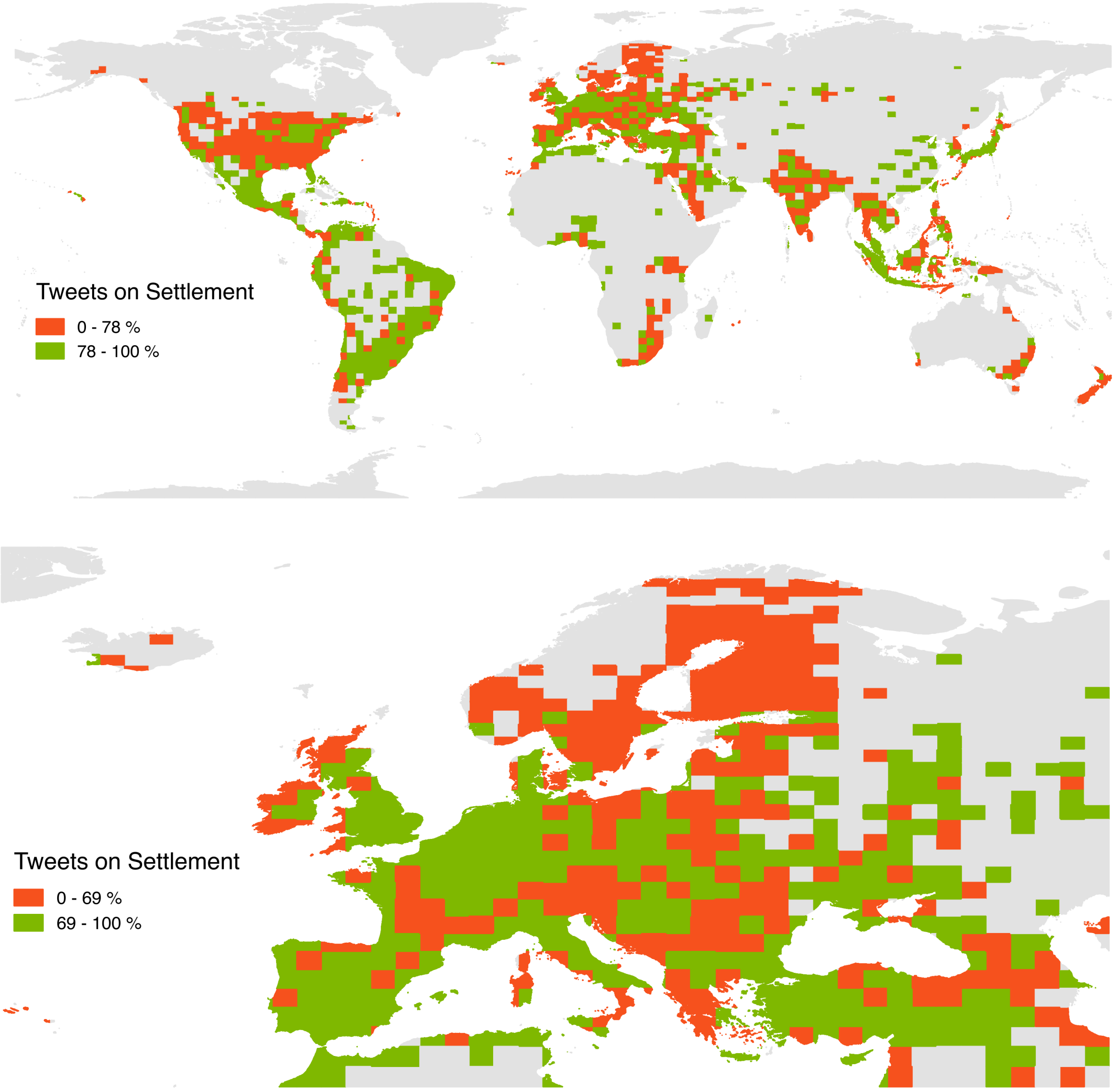


Figure 3.8: Spatial overlap of GUF and T_{world} (upper) and T_{eu}. Green grid cells indicate a spatial overlap above the median of all cells, orange below.

spatial arrangement. The grid cell values in Europe are congruent in both data sets, a strong indication for the robustness of the obtained result. Reasons for this consistent collocation in the two data sets can be manifold. Cultural differences of tweeting behaviour is only one possibility. It may just as well be the constitution of the user base in these locations, such as the dominant bot in Finland.

Spatiotemporal patterns by weekday and hour of day are spatially variable. However, consistent patterns can not be observed at large scales of $300km^2$. At small scale aggregation to continents distinctive patterns are visible (Figure 3.9). Africa and Australia both exhibit a very noisy pattern without obvious trends. This is probably due to the low number of tweets on these continents. Asia, Europe, South and North America have a more pronounced daily routine. On all continents there is a tendency for spatial overlap to be lower on weekends with Sunday generally below Saturday. Europe and South America both show little variation in spatial overlap with GUF throughout a day. In both subsets the spatial overlap is on weekdays lowest during the night.[1] North America and Asia clearly have lower spatial overlap during the night. In Asia, the spatial overlap is close to 90% in the evening. Overall maximum spatial overlap is found in South America, what is in line with results in Figure 3.8.

3.1.5 Point distribution

Tweets tend to be spatially clustered. The nearest neighbour index for both data sets is close to 0. Geographical outliers are more likely to be outside human settlements. The average NND ($\overline{D}$) for T_{world} is 566.7 m. Separate values for tweets by their location respective to GUF reveal that $\overline{D}_{guf1} = 415.9m$ and $\overline{D}_{guf0} = 1'379m^2$. T_{eu} has a lower $\overline{D}$ of 302.5 with $\overline{D}_{guf1} = 116.4m$ and $\overline{D}_{guf0} = 909.0m$. This difference can to some degree be explained by the differing extent. T_{world} at the extent of Europe has the following values: $\overline{D} = 363.4m$, $\overline{D}_{guf1} = 145.7m$ and $\overline{D}_{guf0} = 1'316.8m$. These values are close to T_{eu}. The remainder of the discrepancy can be attributed to the higher point density of T_{eu}, which may also increase the probability of small nearest neighbours. This shows that tweets with far away nearest neighbours are generally less likely to be on human settlement than tweets with close nearest neighbours.

Figure 3.10 shows the cumulative nearest neighbour distribution for the two datasets split by their relative position to GUF. It can be observed that tweets outside GUF tend to have nearest neighbours that are further away. It is worth noting that nearest neighbours are calculated using all distinct points inside or outside settlement.

3.2 Influence of User Characteristics

An aggregation of tweets by user reveals that there are 3’172’437 distinct users in T_{world} and 1’043’168 in T_{eu}. This adds up to an average user activity of 5.6 and 18.0 tweets per user. The user activity is a

[1] Note that the spatial subset of Europe in this graphic is based on political continent boundaries and therefore smaller than $T_{world}(EU)$.

[2] $\overline{D}_{guf1}$ is the average NND of the subset of tweets on settlement, $\overline{D}_{guf0}$ the NND of the subset of tweets outside settlement.

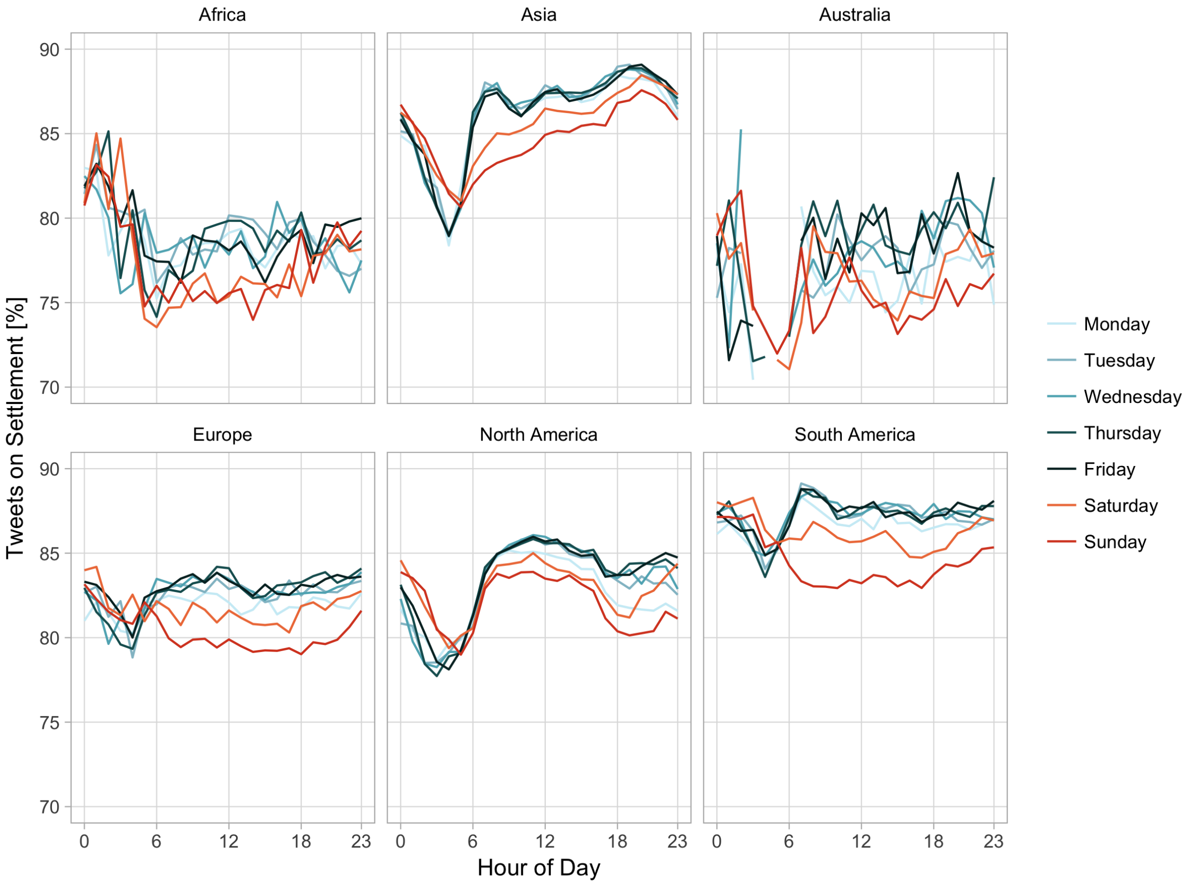


Figure 3.9: Temporal patterns of spatial overlap of GUF and subsets of T_{world} by continent.

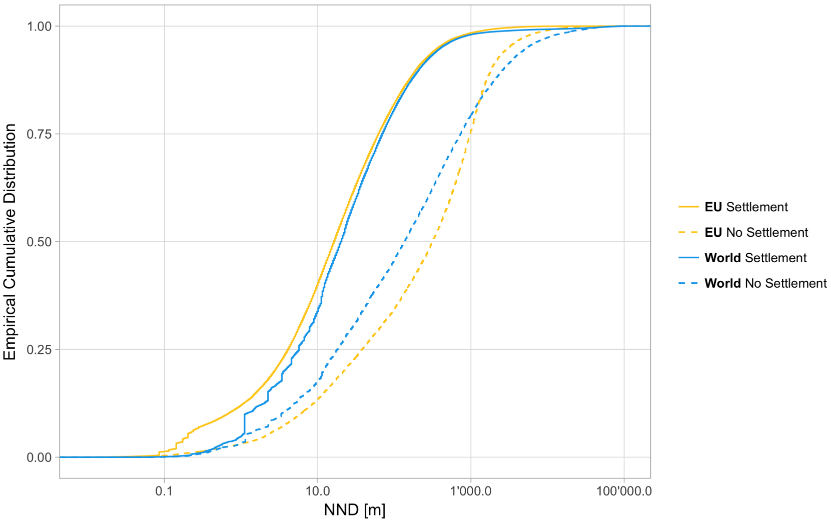


Figure 3.10: Cumulative NND for both data sets split by the tweet's relative position to GUF.

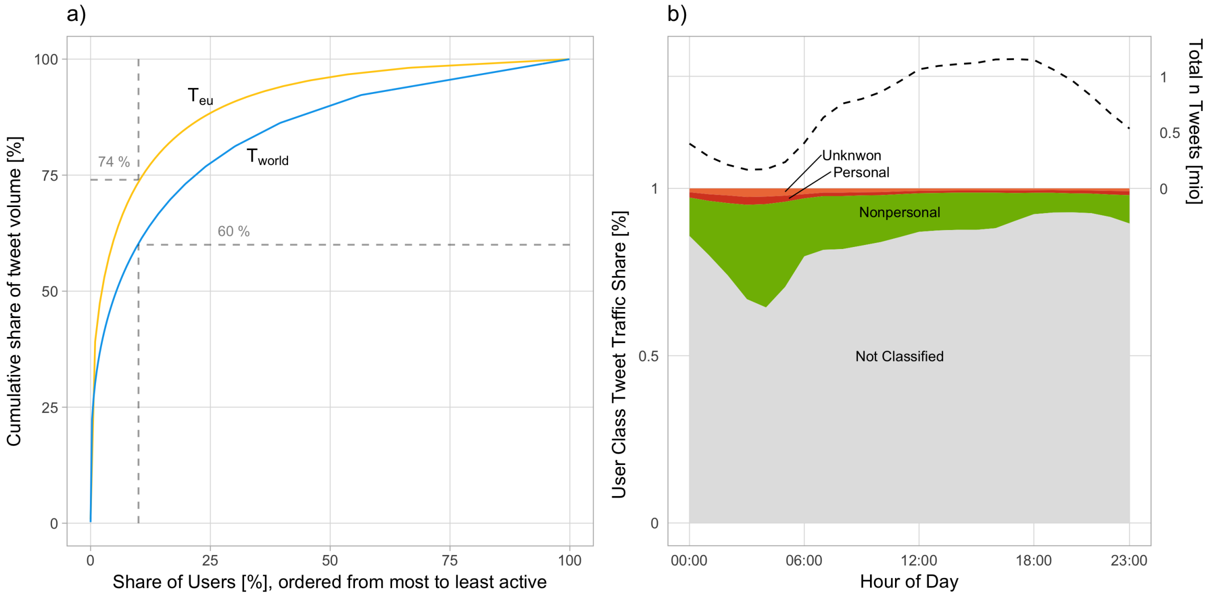


Figure 3.11: **a)** Cumulative tweet volume by user activity. **b)** Daily routine of tweet volume of manually classified users (nonpersonal and personal).

very biased variable. Median number of tweets are 2 and 3 for the two data sets and the mode is 1 in both. Figure 3.11 a) depicts this bias in a cumulative distribution. On the x-axis the users are ordered by the amount of tweets. The y-axis is the cumulative amount of tweets. For example the 10% of users that are most active account for 60% of the tweet volume in T_{world}. T_{eu} on the other hand has 74% of the tweet volume from its top 10% users. A data set with an equal amount of tweets per user would give a diagonal from 0/0 to 100/100. Hence Figure 3.11 a) illustrates that both data sets exhibit a strong contribution bias and the contribution bias is higher for T_{eu} than T_{world}.

Classification of users by hand (T_{world} only)

Out of the 1'500 top users 191 are personal and 1'210 are nonpersonal. So there are only 1'401 classified. The remaining 99 users can't be classified because their accounts no longer existed at the time of classification. Either they are not found (no further information given) or they are blocked by Twitter. These accounts are labelled separately. Moreover there are accounts that are private, which makes it hard to classify. If there is no obvious indication for nonpersonal user (e.g. high number of tweets while low number of followers), these accounts are labelled personal.

The overlap of all tweets with GUF is 84.1%. While 75.6% of the content generated by nonpersonal users is on GUF, the spatial overlap of all classified personal users is 80.8%. Personal users arguably have a higher spatial overlap than nonpersonal users. However, the spatial overlap of users classified as personal is still lower than overall. The spatial overlap for all except the classified users is 85.3%. Figure 3.11 illustrates the hourly share of personal, nonpersonal and unknown users on the total tweet volume. Generally the behaviour of these classes is similar. At times of low Twitter traffic the share of the classified users increases. Nonpersonal users are overrepresented at night. The increased share of users classified as personal may also be interpreted as a characteristic of these very active users. The lowest share is in the evening. Under the assumption that nonpersonal users are less likely to be on settlement, this corresponds to the findings of Figure 3.7 where the highest spatial overlap with GUF is found in the evening, too.

As nonpersonal users dominate during the night, the difference in tweet volume between day and night is more pronounced when looking at personal users only. The fact that the classified personal users have a lower spatial overlap than the overall spatial overlap leaves two interpretations. Either personal users are actually more likely to post from areas outside settlement. In this case the distinction of personal and nonpersonal users is not relevant for this study. The alternative case is that frequent users behave differently from less frequent users. Hand-classification would then be superfluous, as a simple ordering according to user activity would be more adequate.

User activity

The influence of user activity on the spatial overlap of Twitter data and GUF is less pronounced than expected. The most active users ($>$ 10'000 tweets) show a significantly lower spatial overlap with GUF than other users. Figure 3.12 b) shows the local trends of both datasets on a semi-logarithmic scale. The curves are local means derived from a generalized additive model of the user activity and user's spatial overlap with GUF. Both datasets have a spatial overlap of ±80% for users with one single tweet. From there the spatial overlap rises up to around 90% at users with 100 tweets. T_{world} shows a decreasing spatial overlap from there on below 20%. T_{eu} has another peak of spatial overlap at users with 1'000 tweets and decreases then. The higher the tweets per user, the higher is the standard deviation. This is due to the fact that only few observations (users) post large amounts of data. Summing up, the highest spatial overlap is found in mid-range users (100 and 100-1'000 tweets for T_{world} and T_{eu} respectively).

User name keywords

The investigations of individual user groups has led to groups of varying size. Most users are found in the categories *job* (10'394 and 1'205 users), *news* (7'844 and 2'274 users) and *traffic* (2'722 and 1'433 users for T_{world} and T_{eu}). The large discrepancy between the two data sets can be explained by the total number of users in each of them. In total the amount of users in all classes is only 0.8% and 0.6% of the whole user base in the data. However, the tweet volume captured from the entire tweet volume is 10.8% of T_{world} and 7.6% of T_{eu}. The share of *distinct* geolocations on the other hand is again a small fraction of 3.1% and 1.5% within the classes out of all distinct geolocations. This indicates that the classes are including users that tweet the same geolocation over and over again.

In Figure 3.12 a) the spatial overlap with GUF is listed for all classes and both data sets. The dashes indicate the overall spatial overlap. The spatial overlap of the classes with GUF is, in line with

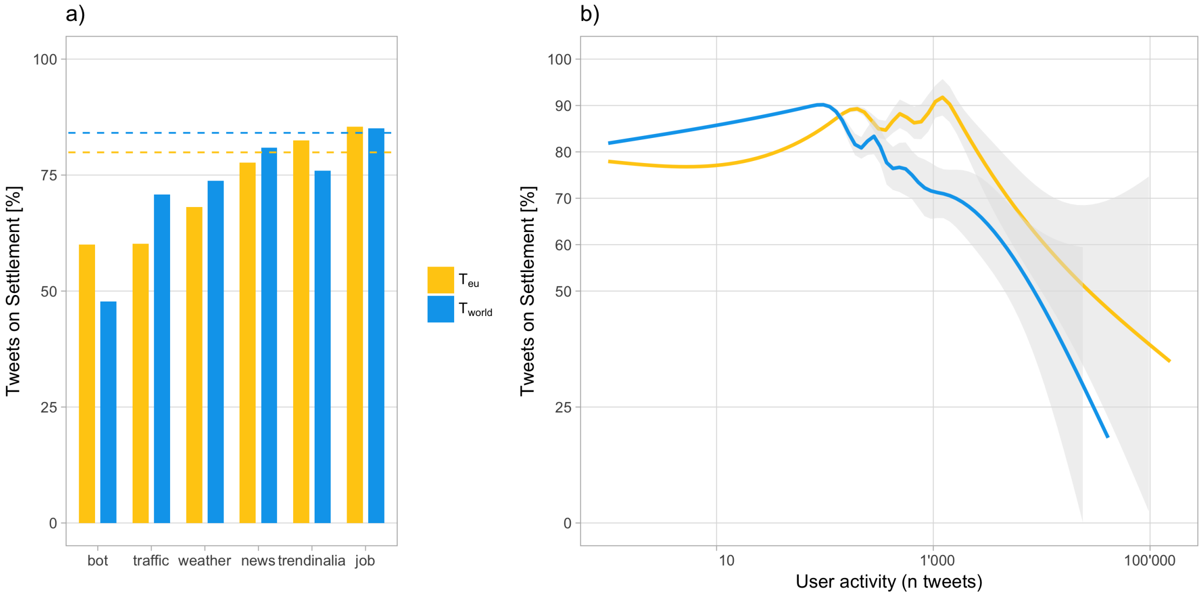


Figure 3.12: a) Spatial overlap of GUF per user class as obtained from regex. **b)** Correspondence with GUF of users according to their activity. Note the logarithmic x-axis.

expectations, generally lower than the overall spatial overlap. Exceptions are the category *job* where both datasets show a higher spatial overlap than the average and *trendinalia* where T_{eu} is slightly higher than average. The category *job* can be considered reliable due to the large amount of distinct geolocations. *Trendinalia* yet has only 122 and 53 distinct geolocations and is therefore less reliable. The lowest overlap with GUF show the categories *bot* and *traffic*.

Accounts in the category *traffic* are typically placing geolocations in tweets where there's traffic rush. The low spatial overlap with GUF can be explained by the fact that GUF does not comprise flat built-up areas like roads. It is likely that this category highly correlates with road locations. It is noticeable that none of the categories shows a spatial overlap nearly as low as random geolocations. The lowest category *bots* with roughly 50% spatial overlap shows still high clustering on human settlements (random would be 1%).

User characterization by point pattern process

In T_{world} the number of users with more than 50 distinct geolocations is 12'720. From these 9'816 are significantly clustered at a confidence level of 95%, 2'904 are random or dispersed (22.8%). In T_{eu} there are 28'912 users with more than 50 distinct geolocations. 3'272 users show a significantly random or dispersed pattern (= 11.3%). The randomness of a user's geolocations do however not directly indicate low correspondence with GUF. Many users with random points have geolocations within a very small area. These tend to be all on settlement. If one takes into account the area of the minimal enclosing rectangle, users with high spatial overlap can to some degree be separated from others. In Figure 3.13

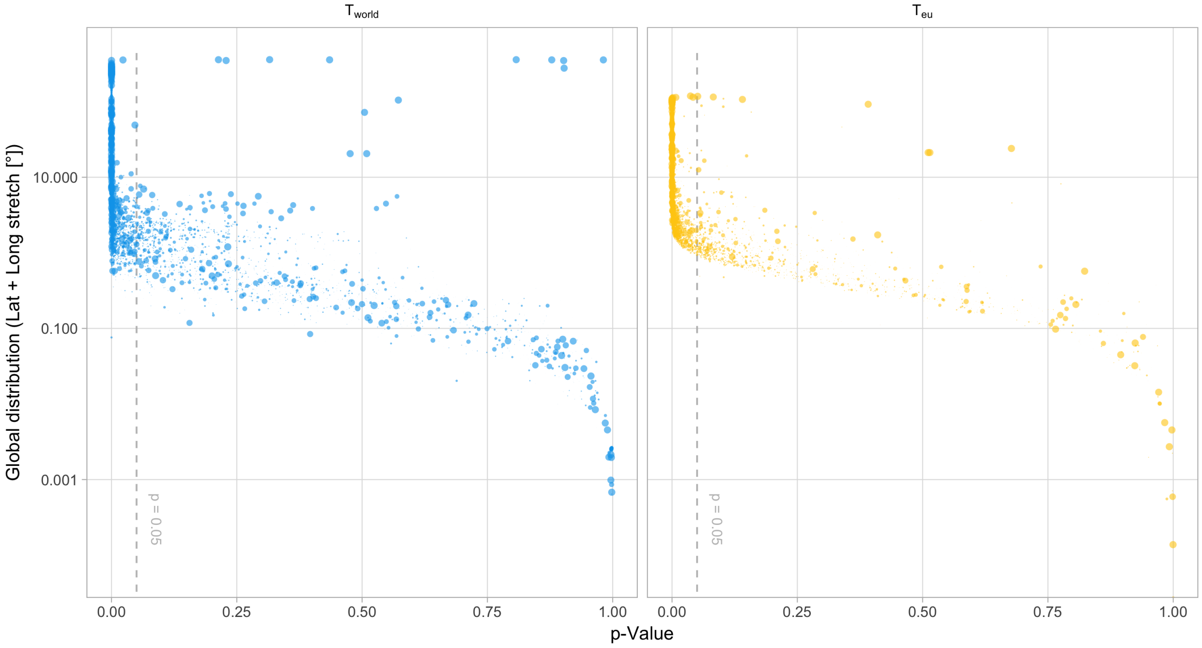


Figure 3.13: Users according to their p-value and spatial outstretch. Large points stand for users with many tweets outside settlement area. The spatial stretch is given as the sum of maximum latitudinal and longitudinal distance.

the users are given according to their p-value and spatial stretch. The spatial stretch is given as the sum of latitudinal and longitudinal maximum distance.[3] Small points have a high spatial overlap with GUF and vice versa. Note that the y-axis is logarithmically distorted. We can observe that the significantly random or regular users ($p(x) > 0.05$) tend to have either very large or very small areas. Generally users with a spatially restricted scope are more likely to be non-clustered. The users covering a large area that are random or dispersed exhibit a low spatial overlap with GUF. Similarly, the spatially very narrowly confined users are more likely to have a low spatial overlap with GUF. However, many users with high correlation are there, too. This is, because depending on where the user's small tweet area is, they are all on settlement or all outside. In either way, users that are not significantly clustered should be regarded critically, as the point pattern is not bearing spatial information.

3.3 Influence of Weather (T_{eu} only)

The dependency of tweet's spatial overlap with GUF on environmental conditions can be exemplified seasonally. The correlation decreases in summer in higher latitudes. In latitudes without seasons, the correspondence remains constant throughout the whole sampling period (Figure 3.14).

[3] This is due to technical problems in Postgis in calculating very large areas on spheres.

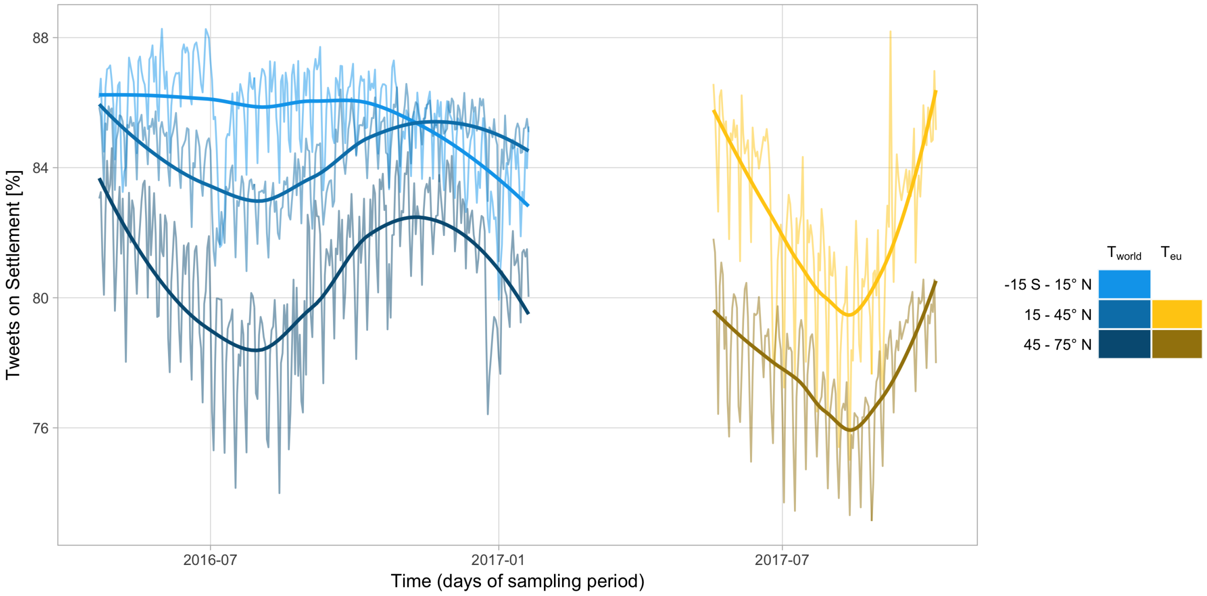


Figure 3.14: Spatial overlap of Twitter and GUF by Latitude. The bold lines represent the local mean of the daily correspondence. Higher latitudes exhibit lower correlation in summer.

The weather data interpolation has led to 548’020 tweets with rain out of 11’020’568 classified tweets. This low ratio is probably due to the summer season during the sampling period. Surprisingly the correlation with weather data is exactly opposite to what was expected. When there is rain, the correspondence with GUF is lower (63.5%) than without rain. A geographically split analysis on a 200km * 200km grid does not alter this result in any region. The spatial overlap of GUF and tweets is in all grid cells below average of all tweets. Also for the other weather variables cloud cover, temperature, wind and snow, the spatial overlap of tweets and GUF is not higher in uncomfortable weather conditions. The values are furthermore analysed on a grid and normalized to average weather conditions per grid cell. Also this method did not detect a trend for higher spatial overlap with GUF in adverse weather conditions.

3.4 Influence of GUF patch size and degree of urbanization

The average GUF patch size is 0.09 km^2 at a total area of 1.05 billion km^2. The median size is 0.01 km^2. Hence the vast majority of patches is very small. In the extent of T_{eu} the average size is 0.08 km^2, the median 0.01 km^2 and the total area is 0.26 billion km^2. In an evenly distributed dataset the number of tweets linearly increases with the patch area, so that point density is constant. However the number of distinct geolocations only slightly correlates with the area of GUF patches. For T_{world} the correlation is $R^2 = 0.2845$, T_{eu} has a slightly higher value of $R^2 = 0.3015$. This can be explained with the fact that only a small fraction of the patches, namely 2.1% have at least one tweet on top. Although it is more likely for large patches to be covered with a tweet, other factors seem more important. The most obvious

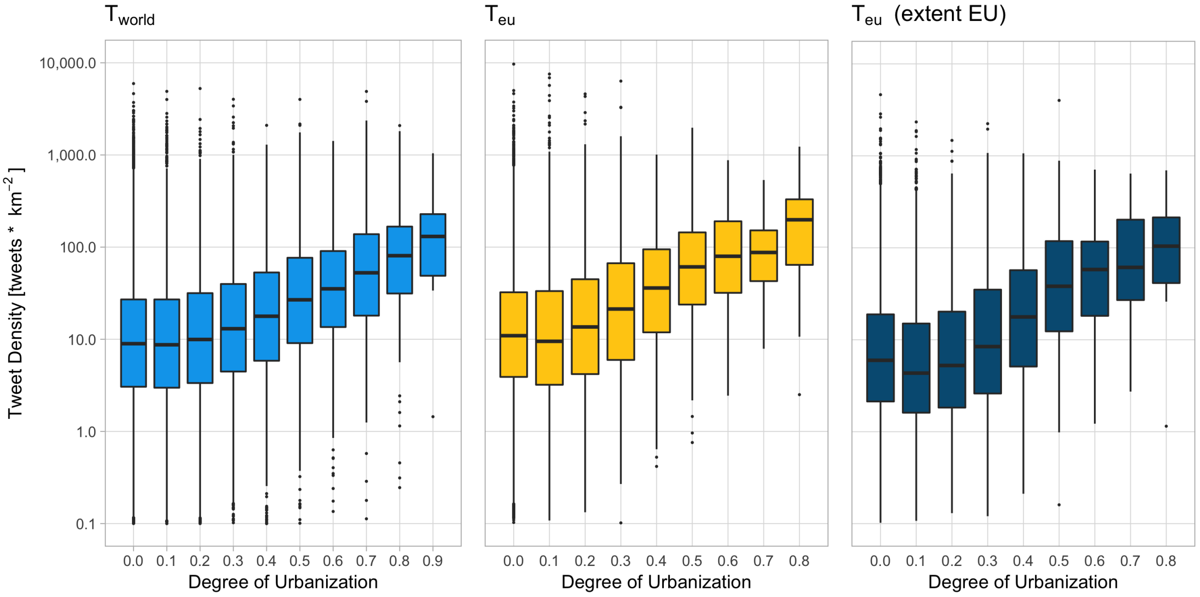


Figure 3.15: Tweets density over degree of urbanization. Only patches with at least one tweet are included. Note the logarithmic y-axis.

of them is geography. Large parts of the world are not covered with tweets, what biases the correlation. This might also be the reason why the correlation in T_{eu} is slightly higher, as Europe has a higher average Twitter penetration than the world.

The Twitter density per patch is overall correspondingly low. For T_{world} the average number of tweets per km^2 is 0.64 with a minimum of 0 and a maximum of 54'436. T_{eu} has an average density of 2.0 $tweets * km^{-2}$ with a minimum of 0 and a maximum of 43'647. An overall correlation between the density of tweets and the patch area can't be expected due to the lack of data in many places. In areas with high Twitter density, the correlation is expected to be higher. A segmentation of the world into a 300 km^2 grid however shows that there is no significant correlation between tweet density and patch size in any region. With the exception of some isolated grid cells (most of them situated in South America), there is no cell with an R^2 of more than 0.1. A similar pattern is present in T_{eu}. Rather, patches of small size show a number of outliers that have very high tweet densities. This can partly be explained with the overproportionally high number of small patches, why outliers are more probable. Also small patches tend to produce outliers as the small area leads to very high densities with just a few tweets on it.

GUF patch degree of urbanization

The derived degree of urbanization of GUF patches are given in percentage of settlement area in the radius of 10 km from the centroid of a patch. On average a GUF patch has 6.9% settlement area in its surrounding area. The distribution is skewed towards smaller amounts with a median of 3.7% and a maximum of 94.8%.

The influence of the degree of urbanization on tweet density is more pronounced. Higher degree of urbanization leads to a higher tweet density in both data sets and $T_{world}(EU)$, as can be seen in Figure 3.15. However this only holds true because patches with no tweets are excluded. The variability increases with lower degree of urbanization. In conclusion it can be stated that on average tweets are more dense in urban areas, but they may be dense in less urban areas too. Probably the patches with low degree of urbanization and high tweet density stem from regions with generally high Twitter penetration.

3.5 Classification of tweets

The random forest classifier has an overall accuracy of 0.88 for T_{eu} and 0.89 T_{world}. Precision is 0.88 for T_{eu} and 0.9 for T_{world}. This is the percentage of true positives by all positives. Hence, this number illustrates that in both datasets the tweets classified as settlement are to 90% actually on settlement on T_{world} and 88% in T_{eu}. As compared to the raw tweets where 84% and 79% are on settlement, this is an increase. This comes at the the cost of some false negatives. Recall is the share of true positives in the true positives and false negatives. In this case the false negatives are tweets that are on settlement but not classified as such. Both data sets have a recall of 98%. Hence 2% of the tweets on settlement are wrongly classified as being outside settlement.

	Precision	Recall	Accuracy
T_{eu}	0.88 ±0	0.98 ±0	0.88 ±0
T_{world}	0.9 ±0.01	0.98 ±0	0.89 ±0.01

Table 3.2: Classification results of RF model. Values are means of 10 predictions of 0.1% randomly selected tweets of the respective data sets. The standard deviation is rounded to two digits after comma.

A precision of 90% shows that the classifier can distinguish tweets on settlement from tweets outside settlement. When the result is seen in relation with the distance to the nearest settlement, the power of the method becomes even more striking. In Figure 3.16, random samples classified with the trained RF are shown. The x-axis represents the distance to the nearest settlement and y stands for the ECDF. The two curves list all tweets classified as 0 (no settlement) or 1 (settlement). It can be observed that the tweets classified as settlement tend to be closer to human settlements. The classifier can not only select tweets from the raw data sets that are more likely to be on human settlement, but also the false positives tend to be close to human settlement.

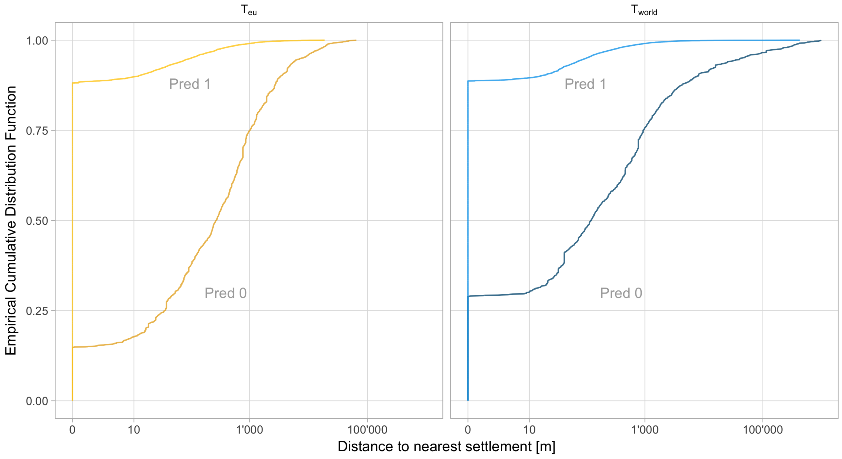


Figure 3.16: Cumulative Distance of 1% randomly sampled tweets by the respective predicted class settlement (1) and outside settlement (0). The x-axis is logarithmic.

3.6 Detection of Settlements

In total GUF has 11'289'316 patches of connected touching settlement pixels. 11'054'157 (= 97.9%) of them are not covered with any tweet from T_{world}. From the patches with area of more than $1M\ m^2$ (97'749) there are still 44'504 = 46% not covered with tweets. In the extent of Europe there are 3'099'100 patches from where 3'028'116 are not covered with tweets (= 97.7%). The estimation of detected pixels obtained through a Monte Carlo simulation are shown in Figure 3.17 a). $t_{guf=0}$, the tweets outside settlement area, stays constant in both data sets, which is sensible due to the random selection process. In T_{eu}, t_{close} is constant with slight decrease at increasing sample size. T_{world} shows a slight trend towards increase in the share of t_{close}. The main process that can be observed in both data sets is that t_{dupl} increases at the expense of t_{det}. However, T_{eu} has a lower share of detected pixels at any data set size. This can be explained by the differing geographic extent of the two data sets. Worldwide the chances for two tweets to be close or identical is smaller than in the restricted area of Europe.

In Figure 3.17 b) the values of t_{det} as simulated for varying sample size are given as points. The points exhibit an exponential behaviour. The fitted curves are overlaid. For T_{eu} the function takes the form $y = 117.532 * x^{0.561}$, where the independent variable x is the sample size and y the number of detected pixels. E. g. at a sample size of 10 M tweets the function estimates that 993'473 pixels are detected. T_{world} with its higher detection rate has the following fitted model: $y = 77.18 * x^{0.64}$. At a sample size of 10 M it would accordingly detect 2'352'314 pixels. With the help of this function and statistics from GUF the amount of Twitter data needed to detect all settlement pixels is estimated. In GUF, there are in total 175.2 M pixels with human settlement. This is a share of 0.7% of all classified pixels in GUF (24 B). The amount of tweets needed to detect all human settlements is then given as $17.52 * 10^6 = 77.18 * x^{0.64}$ what results in $x = 8'380'807'008$. Hence according to this extrapolation an amount of > 8 B precisely

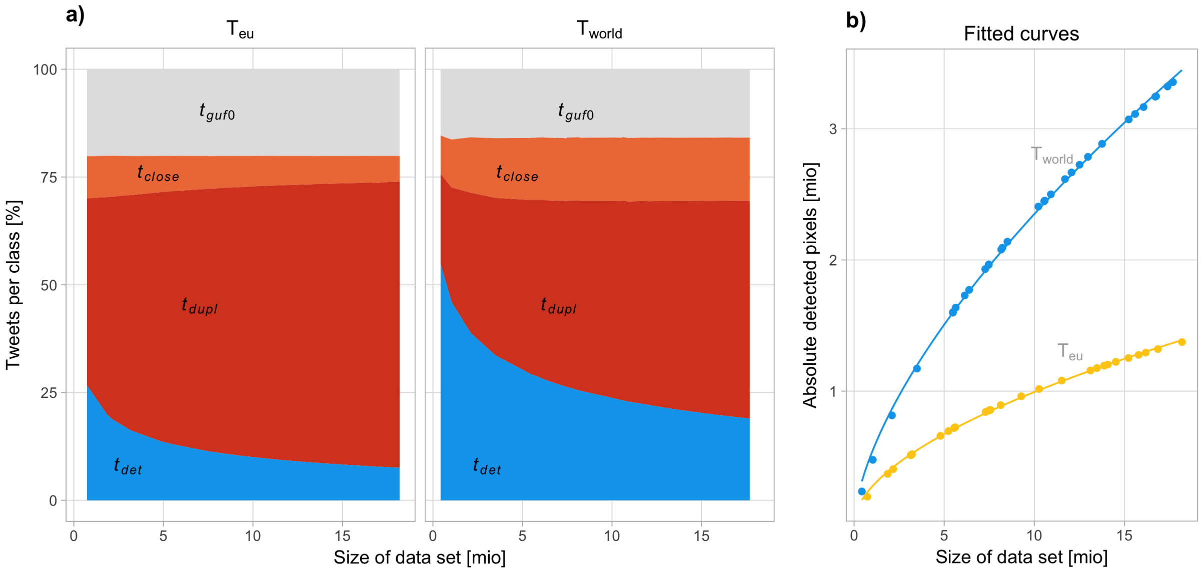


Figure 3.17: Results obtained from Monte Carlo simulation. a) show the share of the parameters t_{guf0}, t_{close}, t_{dupl} and t_{det} for the two data sets. b) displays the absolute values of t_{det} and the fitted curve.

geotagged tweets are needed to detect all human settlements. The same procedure applied to T_{eu} and the GUF data in this area yields an even higher number of 10.6 B required tweets. The total number of pixels in this area is 2.4 B whereof 49.5 M are labelled settlement (2.1%) These values are summarized in Table 3.3.

The data sets were sampled in a period of 275 (T_{world}) and 143 (T_{eu}) days returning roughly 18 M tweets each. The sampling period to obtain a tweet volume that is sufficient to detect all pixels would then be 81'527 days for T_{eu} and 129'101 days for T_{world}.

	n Settlement Pixels	n tweets	$t_{det}(T)$	$T(t_{det} = all)$	Required sampling days
T_{eu}	49.5 M	18.6 M	1.4 M	10.6 B	81'527
T_{world}	175.2 M	17.9 M	3.4 M	8.4 B	129'101

Table 3.3: Summary of detection rates of the two data sets. $t_{det}(T)$ stands for the detected pixels with the given data, $T(t_{det} = all)$ is the required tweet volume to detect all pixels.

Chapter 4

Discussion

This study has revealed interesting patterns in the spatial and temporal relationship of tweets and human settlements. An overall high spatial overlap is significantly influenced by spatial, temporal, user-specific and partly environmental variables. In the following it is examined if and to what degree the results can answer the research questions of this thesis.

4.1 Research Question 1

Does the presence of geolocated tweets indicate built-up land cover?

With a spatial overlap on GUF of roughly 80%, Twitter data is an indicator for human settlements. The spatial overlap is by a factor of ±100 higher than would be expected from random geolocations. The 20% of the tweets outside human settlements tend to be in the vicinity of human settlements. 92.7% of the distinct tweet locations in T_{eu} and 97.1% of the distinct tweet locations in T_{world} are within $1km$ from a GUF settlement patch.

This can be interpreted in three ways: (1) Either near-by tweets may actually be on human settlement and the applied ground truth (GUF) is wrong. However, according to Esch et al. (2017) GUF is one of the most accurate products depicting human settlements. Twitter geolocations on the other hand have, depending on the positioning system employed, varying accuracies of meters to kilometers (Roxin et al., 2007). Due to this, Twitter data probably isn't more accurate than GUF in predicting human settlement. (2) Yet GUF does not depict flat built-up areas (Esch et al., 2017). Supposing that humans tweet mainly from any kind of built-up area, the tweets outside human settlement could indicate built-up areas that are not depicted by GUF, such as roads or town squares. This conjecture is supported by the fact that flat infrastructure like road is predominantly in the vicinity of settled areas. Also Rios (2013) reports that transportation routes are covered with Twitter data (see Figure 1.1). Further studies may incorporate other kinds of infrastructure in order to test this. (3) A third explanation is that the tweet locations reflect places of human activity. They tend to be close to built-up area in the same way like humans tend to be close, but not necessarily right upon built-up area. People probably also visit city parks or the nearby forests on a regular basis. This is also supported by the fact that tweets are more likely to be close to human settlement than far away. This explanation is in line with the assumption that Twitter data reflects human population distribution at a fine resolution (Patel et al., 2016).

It is probable that a combination of the latter two processes described above contribute to the distribution of Twitter data. Thus, the idea that Twitter data might be more accurate in detecting human settlement than GUF is rejected while the explanations claiming that Twitter data reflects any kind of infrastructure and/or human activity are supported. This suggests that Twitter data is only an indirect indicator for human settlement and a direct indicator for human activity. As stated in the introduction of this thesis, models of built-up land cover only depict features that the data allows to detect. Twitter data seems to detect human activity rather than settlements. Yet the two phenomena spatially overlap, what is used to estimate population density from remotely sensed built-up areas (Bhaduri et al., 2002). However, Twitter data may provide higher resolution information on population distribution than physical measurements do. Future research should therefore focus on quantitatively assessing the relationship of human activity and Twitter data. Furthermore, if tweets from within human settlements can be discerned from tweets outside settlement, Twitter can also serve as an accurate proxy for human settlements. This is the focus of the second research question.

4.2 Research Question 2

What are factors that influence whether the geotag of Twitter data is on settlement area or outside?

Properties of tweets that were assumed to affect the spatial overlap of Twitter and human settlement are time, geography, user characteristics, weather conditions and the relative position of a tweet to other tweets. With the help of a random forest these properties were used to discern tweets outside settlement from tweets on settlement.

4.2.1 Spatiotemporal setting

The spatially differing temporal Twitter activity signatures suggest that the use of Twitter space-dependent. Also the overlap of tweets and GUF varies spatially. Latin America has a relatively high spatial overlap while North America generally shows a lower degree of overlap. Northern Europe has a relatively low spatial overlap. Whether these differences can be explained with different human behaviour can not be definitely concluded. However, compared to the Twitter activity signatures, which are not consistent in the T_{eu} and $Tworld$, the spatial overlap is spatially consistent.

Also temporal patterns of overlap are varying geographically. The maximum overlap of tweets and GUF for example is in the evening in Asia, at noon in North America and in the morning in South America (see Figure 3.9). The temporal pattern of the whole dataset T_{world} is similar to the patterns of Asia and North America. This is due to the fact that Twitter data set is dominated by Eastern Asian and North American tweets. Hence, any temporal patterns found in the full data set are in fact reflecting the regions with highest Twitter penetration. Discussing explanations for temporal variation in overlap of GUF and tweets has to be based on spatial subsets of the data with more or less constant Twitter penetration. How these spatial subsets are to be selected poses another challenge: If space is subdivided into very small areas, the number of tweets within that area may be too small for a representative result. On the other hand, large subdivisions lead to a mixture of patterns. This can be illustrated by the different temporal overlap patterns of Europe in Figure 3.9 and $T_{world}(EU)$ in Figure 3.7. The spatial subset of the former is based on political boundaries, where in the latter by a bounding box of the queried area and therefore slightly larger. This issue is referred to as Modifiable Areal Unit Problem (Wong, 2004).

Some features of the observed spatiotemporal patterns can be attributed to human behaviour. The lower overlap during weekends is very probable ascribable to the fact that on off-days people are able to head outside instead of staying inside working environments. The observed decrease of overlap of Twitter and GUF in Northern hemispherical summer months is likely to be due to the fact that people go out in warm weather. On the other hand, with the same logic it is assumed that during night people reside in their homes and thus the spatial overlap of tweets and GUF is high. Twitter data shows on all continents (except for Africa) the opposite: the spatial overlap of tweets and GUF during the night is lower compared to the rest of the day. A possible explanation for this unexpected result is that the relative contribution of bots posting random geolocations to the Twitter traffic increases during the night. According to the results of manual classification, the share of bots during the night actually increases (see Figure 2.4). Bots and services are not particularly active at night, but people are particularly inactive what leads to a high share of nonpersonal users. In this light, studies using temporal characteristics of Twitter data have to be judged critically. E. g. Lin and Cromley, 2015 used geolocated tweets sent between 6 pm and 8 am to identify residential areas. This is only possible when nonpersonal users are excluded from the data (what hasn't been done in Lin and Cromley's study).

Nonpersonal users are not only represented to a varying degree temporally, but also spatially. Over Finland the user *Every Finnish Number* posts most of the geolocated tweets in this region. The geolocations are most likely randomly generated. In the region of Finland, the overlap of GUF and tweets is therefore lower than average (see Figure 3.8). The same bias can be suspected in other regions. The interpretation of geographic variations in the overlap of GUF and tweets has to account for this. The high spatial overlap in South America can either be due to the fact that people in South America tend to stay within human settlements or due to a lower share of nonpersonal users. Thus, the distinction of desired from undesired users is crucial for a reliable interpretation of these results.

4.2.2 Classification of Users

The pronounced contribution bias demands for an investigation of the data at user level. The most active users are often not of the kind that is desired, but rather likely to be bots with no, or services with little meaning in spatial regards. Several approaches have been tested, among which some have turned out to be more feasible than others.

Manual classification

Classification by hand bears the disadvantage of high workload and more importantly subjective decisions. Millions of users pose a major hindrance for a nearly representative sampling. The classification process can only partly be standardized (here through given criteria). It might especially be that the perception of the content of a Twitter user page is biased by the author's cultural background. Accounts foreign to the author are unlike harder to classify. Those foreign accounts are checked more extensively and with use of translation tools. Despite this, comparing users in different cultural regions can not be done. Another drawback of this approach is that it's hardly generalizable. Different Twitter data sets are made up of different users, as these are changing over time and by request parameters.

However, the manual assessment of driving actors can also be seen as a window into the blackbox of big data. In this case, the manual classification not only led to a deepened understanding of the data, but was crucial for the development of further analysis methods. It is argued that getting to know to the data is a prerequisite for analysing Twitter data. Bots as well as the contribution bias are not specific to Twitter data. They are inherent characteristics of social media data (Li et al., 2013). The need for manual data sighting is not limited to Twitter data, but a necessity for any social media data analysis.

Automated classification

The user classification according to their tweet activity, selected keywords and geolocation distribution is relevant for the overlap with GUF. The most active tweeters are clearly bots or services that post tweets overlapping GUF to a lower degree than average. Likewise, selected keywords could extract groups of users that show varying overlap with GUF. This approach was completed with English keywords only, certainly the predominant language on Twitter. But the approach could be enhanced by incorporating more languages.

The assessment of the point pattern distribution per user is relevant for the overlap of tweets and GUF, too. Users that do not tweet meaningful geolocations can be distinguished from users with meaningful geolocation. Most importantly, users that (by visual inspection) are unambiguously personal users with meaningful geolocations get a low p-value. This is a major advantage over methods relying on *mean* nearest neighbour distances (Lagache et al., 2013). Though, users that by means of this method get a high p-value do not necessarily have many points outside human settlement. Most of the users with a high p-value have many points in either a very small or a very large area (see Figure 3.13). In the case of small area, the tweets are often all on GUF. Although the p-value does not predict the overlap of a

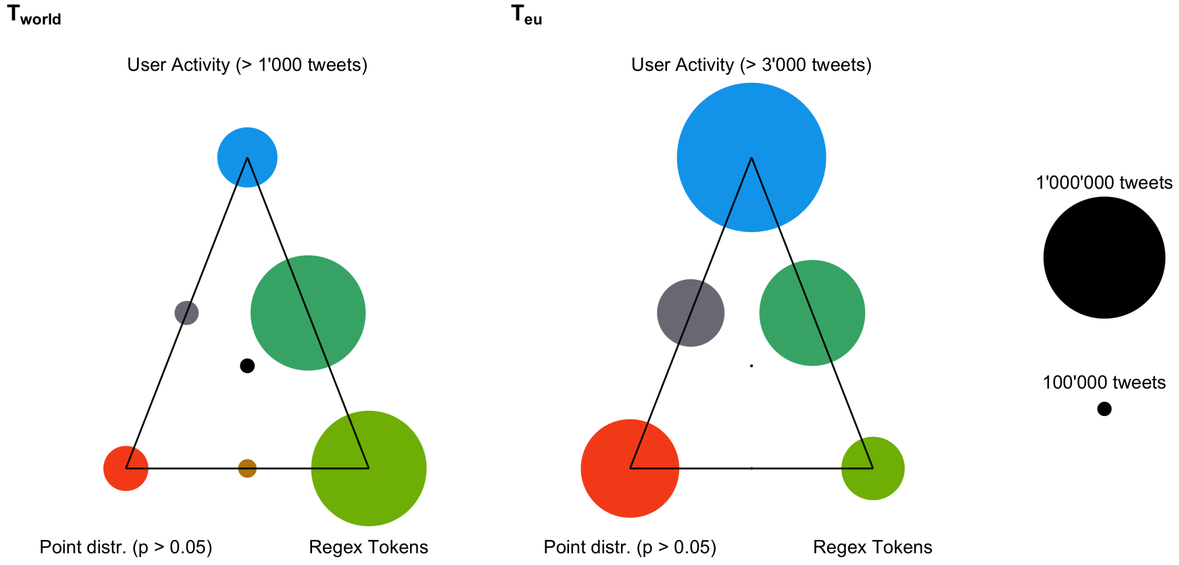


Figure 4.1: Amount of tweets from undesired users by *user activity, point distribution* and *regex keywords*. Circle sizes indicate the number of tweets per user class. On the nodes of the triangle are summed tweets from users that are only in one of the user classes, on the edges in both neighbouring classes, in the center in all three user classes.

users tweets with GUF, it identifies users with random or regular point distributions and can therefore be recommended to be applied in further studies on geotagged social media data. The method may be enhanced in two regards: First, the sea area could be excluded for the simulation of random point patterns, as this may be done so by nonpersonal accounts, too. Second, nonpersonal users may use different coordinate systems for geolocation generation. In this thesis the simulated random points were generated in WGS84 coordinates. This could additionally be done in other (projected) coordinate systems.

The three automated methods combined identify 3'150'164 tweets in T_{world} and 3'963'382 in T_{eu} as tweets from undesired users when thresholds given in Figure 4.1 are applied. This is a share of 17.7% for T_{world} and 21% for T_{eu} of the whole data volume. Although these methods found such a large part of the data to be sent from nonpersonal users, many studies do exclude only a small fraction from their raw data. E. g. Jenkins et al. (2016) remove no data, Lin and Cromley (2015) remove nearby geolocations only, Soliman et al. (2017) exclude 'redundant tweets without true geographic coordinates'. The results of such studies have to be regarded critically. When Twitter data is queried by geolocation only, monitoring or cleansing of undesired users during the study is of particular exigency. While filtering by keywords in the Streaming API might omit many users posting random or similar content, a geolocation query returns all tweets that are geolocated not regarding tweet contents. With the here applied methods to classify users, the same users are sometimes detected by two or all three methods. Figure 4.1 illustrates this with circles indicating the number of tweets detected by each method. In the corners of the triangle the number of tweets detected by one single method are given. On the edges, tweets detected by the two neighbouring methods are summed and in the center tweets detected by all three methods. Particularly the methods looking at regex keywords and user activity have often detected the same users. However, there are hardly any users classified as nonpersonal by all three methods. Many users are only detected by one of the methods. On one hand this is a legitimization for the incorporation of several user classification methods in this study. On the other hand, this exemplifies that Twitter data cleansing is important and depending on the applied method leading to different results. The methods used in this study are by no means covering all types of undesired users. E. g. one aspect that is not discussed here are what Chu et al. (2010) name *cyborgs* and *malicious bots*. Cyborgs are bots that imitate human behaviour and are therefore hard to detect with the methods applied here. Malicious bots post content on hacked accounts and are equally hard to detect.

In 2014, Driscoll and Walker (2014) called for a 'common language for Twitter research'. In this spirit, research has to be conducted on how to standardize Twitter data cleansing integrating many approaches that describe Twitter users. A standardized framework for the detection of undesired users would make Twitter studies more comparable. This might also involve the development of a programming library for this purpose that can be used by researchers. An easy-to-use and freely available programming interface would leverage the use of standardized methods.

4.2.3 Weather data

The seasonal variability in ovrerlap of GUF and tweets (Figure 3.14) clearly suggests that weather-related behaviour of staying inside or outside of buildings is represented in Twitter data. However, there was no correlation found between the weather data and the overlap of GUF and tweets. The interpolated values were analysed in spatially and temporally restricted subsets. In doing so, a bias induced by different weather regimes per region was circumvented. None of the weather parameters *rain, cloud coverage, temperature, snow, wind* showed to have an influence on the overlap of GUF and tweets. Therefore it is tantalising to assume, that either the weather data or the applied methods are not suitable.

A first uncertainty is introduced by the fact that the weather data is not based on direct measurements but rather on resampled values to town locations. Therefore the spatial distribution is uneven. Furthermore, these values are then interpolated temporally and spatially to the timestamp and the location of tweets. Due to these two intermediated manipulations of the data, the weather values calculated per tweets might be to far offset from the actual weather conditions.

A second issue is that other changes in tweeting behaviour than expected might be induced by weather condition. E. g. bad weather might lead to a decrease in Twitter activity amongst personal users. If so, nonpersonal users would have a larger share and hence the spatial overlap of tweets and GUF would decrease.

At this point, a definite conclusion on the influence of weather data on overlap of tweets and GUF can't be drawn. It is probable that the sum of uncertainties introduced by the weather data, the Twitter data and the applied methods lead to an insignificant result when using ancillary weather data. The observed seasonal variation of Twitter data overlap with GUF suggests, that weather influences the tweeting location of people.

4.2.4 Point distribution

NND is a simple but in this case powerful measure. Tweets with close nearest neighbours are more likely to be on settlement. Put differently, the NND is a measure for local point density. High tweet density enhances the likelihood for human settlement. The caveat of using NND as indicator of a tweet's position relative to GUF is, that spatial outliers are always regarded as outside human settlement. The application of this method has to be regarded carefully, otherwise tweets in big clusters dominate and fine granular information is lost.

Another weakness of the NND is that it is dependent on the data density. Data sets with high tweet density have generally lower NNDs. This especially poses a problem when the density of tweets varies not only due to ground truth, but due to a geographic bias. This is the case in Twitter data. In this case tweets on settlement in countries with a low Twitter penetration are likely to have a high NND. Then the NND is not a good predictor.

Last, the NND could be substituted with an average k-NND. This method takes, instead of the one nearest neighbour, the k nearest neighbour distances and is therefore more robust.

4.2.5 Patch Size

It was observed that the degree of urbanization of a GUF patch influences the tweet density. This confirms the presumption in many research papers, that tweets are overrepresented in cities (e. g. Soliman et al., 2017). The correlation of degree of urbanization of a patch and tweets density is overshadowed by uneven geographical distribution. Only the exclusion of patches with no tweets leads to a visible trend 3.15. The applied measure of urbanization (percentage of settlement area in the surrounding) has shown to be a better indicator than the GUF patch size.

However, if population density was incorporated per patch, which is normally higher in cities, the observed effect might even out. It remains unclear whether urban people are using Twitter more than people from the countryside. This would require a comparison with population data.

4.2.6 Classification with Random Forest

The application of the RF has shown that tweets on settlement can with the help of the above discussed factors to some degree be distinguished from tweets outside settlement. Especially when taking into account the distance to the nearest settlement of the classified tweets, the classifier shows its full strength. The classifier predicts many false positives. But these tend to be near settlement. This result is argues for the validity of the used factors discussed in this thesis.

	Gini Index T_{eu}	Gini Index T_{world}
NND	103792.62	72469.50
Duplicated geolocation	38951.61	24950.67
User activity	36289.41	24332.89
Hour of day	19251.72	20169.08
User point distr.	11810.08	4691.21
BBox area	11103.54	6866.43
Day of week	8983.18	9368.65
Month	6312.18	11118.51
User keyword	3406.77	3143.74
Continent	0.22	7182.94

Table 4.1: Factor importance (Gini Index) for the RF model. High value indicate high relative importance. For details on how the values are generated refer to the R package *randomForest*.

One of the strengths of the random forest algorithm is its ability to report the relative importance of the input factors for the classification (Breiman, 2001). In Table 4.1 the Gini Index is given for the applied factors for both data sets. In both data sets, the NND is the most important factor. In T_{eu} the importance of the factor *continent* is very low due to the spatial extent of the data set. Amongst the factors defining the spatiotemporal setting (gray), the hour of day is the most important factor.

The dominance of the NND may introduce a geographical bias in the prediction accuracy. Places with low Twitter penetration have fewer tweets and thus contribute less to the model generation. The NND is therefore representative for places with high tweets densities. The classifier is then more likely to classify tweets with relatively high NNDs as *outside settlement* in countries with a low Twitter penetration, because the density is lower and hence the average NND is higher.

It would therefore be advisable in further studies to watch the accuracy of the classification in spatial subsets. Further improvements could be achieved using regression random forests on the distance of tweets from GUF instead of classification on *inside* or *outside settlement* (see ibid.).

4.3 Research Question 3

Is a global classification of built-up land cover possible with Twitter data?

Due to the insufficient coverage, a global classification of human settlements is not possible with the given Twitter data only. More data does however not lead to a significantly higher coverage of tweets to an extent that human settlements are globally covered. Twitter data is in fact limited to certain areas of the world (Leetaru et al., 2013). A larger Twitter data set generated by a longer sampling period produces more data mainly in these regions with high Twitter penetration. To estimate how much Twitter data would be required to get global coverage, a curve was fitted for different sample sizes of the given data. According to this estimate 10.6 B (T_{eu}) and 8.4 B (T_{world}) precisely geolocated tweets would be required, what corresponds to a sampling period of 223.4 and 353.7 years, respectively.

The differences in the estimates for T_{eu} and T_{world} exemplify uncertainties in this estimation. t_{det} is with 1.4 M in T_{eu} significantly smaller than in T_{world} with $t_{det} = 3.4M$, although the data set are almost equal in size. The higher detection rate in T_{world} (3.4 M out of 17.9 M = 19.9%) compared to T_{eu} (7.5%) results in a higher estimate for the required data volume to detect all pixels in T_{eu}. Considering the smaller spatial extent of T_{eu}, a lower number of required tweets would be expected for T_{eu}. Other uncertainties may arise from the following assumptions made. It is assumed that the tweeting behaviour of people stays constant. Changes in tweeting frequency, geolocation sharing or geographic adoption may increase or decrease the detection rate. The pixels size of the classification raster is fixed at 84 meters. Setting the output raster resolution differently may lead to a different result. E. g. Schneider et al. (2009) presented a global map of urban extent at 500 m resolution.

Despite these uncertainties, the results nevertheless exemplifies that a global or even Europe-wide classification with Twitter data only is far beyond producibility. However, due to the high spatial overlap of Twitter data and human settlements, the use of Twitter data as complementary or control data for land cover classifiers bears high potential. This has already been proposed by Miao (2017).

4.4 Comparison of T_{eu} and T_{world}

The two Twitter data sets used do often show different results. Differences are found in the downloaded volume per day, the number of distinct geolocations, the overall spatial overlap with GUF (Table 3.1), the user diversity (Figure 3.11 a)), the diurnal volumes (Figure 3.2), and the user classification results (Figure 3.12). In this section the differing results for the two data sets are discussed in the context of the data acquisition process and their different spatial and temporal extent.

4.4.1 Amount of geolocated tweets

In this study the data is queried by geolocation. The process of downloading Twitter data is not expected to introduce any uncertainties. However, the different volumes per day call for a critical view on the downloading process. A first source of uncertainty is probably the implementation of the data crawler infrastructure. Twitter's Streaming API relies on REST requests. Hence, any software can be used to stream the data. In this case the software was written in Python using tweepy (T_{eu}) and in R (T_{world}). Variations in Twitter data volume raise the question whether the crawling infrastructure can influence the amount of data retrieved.

Another issue is the changing Twitter specifications. Just recently, Twitter has changed the maximum message length of a tweet to 280 instead of formerly 140 characters [1]. Such changes go without notice when it comes to more trivial functionalities like geolocation sharing. Changes in the user interface of the Twitter mobile app (e. g. changing the location to the geolocation sharing activation button in the

[1] https://help.twitter.com/de/using-twitter, accessed 2017-12-03

app menu) may lead to a rise or decline of geotagged tweets. T_{eu} and T_{world} are downloaded in different time frames. It is therefore possible that the amount of geotagged tweets may have changed during this time period due to changes in Twitter's specifications. This has wide implications for the comparability of Twitter data sets. Only two years ago Blanford et al. (2015) report that 91% of the geolocated tweets are located by GPS. In 2017, as indicated by the many duplicate points in the given data, the amount of GPS-geolocated tweets is substantially lower.

A possible scenario for the observed differences in Twitter volume and user activity distribution could be delineated as follows: Twitter chnages either the default settings for geolocation sharing or the location of the geotag sharing button in the app menu. More people don't share geolocation or only obfuscated geolocations. However, the nonpersonal users do not access Twitter through the mobile application and pursue to post precise geolocations. The retrieved data consists of fewer users and contains more duplicate (obfuscated) geolocations. The spatial overlap with GUF may decrease, because nonpersonal users contribute more to the retrieved data.

4.4.2 Diurnal and weekly patterns

A comparison of two datasets has shown that diurnal and temporal patterns are inconsistent. For example T_{world} has peak hours in the morning and T_{eu} in the afternoon at the same spatial extent.

One possible explanation for differing results is the 1% download limit of Twitter's Streaming API. In scientific literature it is widely accepted that geolocated tweets do not significantly exceed the 1% limit at any time (Kumar et al., 2013). As of testing, tweets that were sent by the author were not always returned by the Streaming API when querying T_{eu}. This is evidence that not all tweets are returned. It is not clear how many and which tweets are omitted. In any case geotemporal analyses are therefore unreliable. If geotagged tweets are to exceed the 1% limit in certain time slots, it has to be presumed that geotagged tweet *share* is variable. This is what Leetaru et al. (2013) found. As long as the geotagged tweets do not exceed the 1% limit, all geolocated tweets are returned. Once this limit is surpassed, Twitter returns only a subset of all tweets. Then comparing absolute tweet volumes becomes invalid.

Let's illustrate this process with an example (see Figure 4.2). For the sake of convenience, the absolute tweet volume is assumed to be constant over time in this example. When total tweet volume is 10'000 tweets per time, then the returned number of tweets is 100. We assume two longitudinal differing places (orange and green). The total amount of geolocated tweets is given in 4.2 a). The geolocated tweets exceed the 1% limit from 4:00 am to 8:00 pm due to a diurnal pattern in the green location. The corresponding response to a request from Twitter is given in b). Due to random sampling, the location orange exhibits a diurnal pattern that is not present in the full geolocated data set. The circumstance is further complicated by the fact that the total tweet traffic is likely to vary with time. In Figure 4.2 c) the tweets for data from T_{world} in GMT are grouped by continent. Here the absolute amount of data varies. But it is unknown whether this amount is always the full amount of returned tweets. The activity signatures per continent influence the total amount of tweets and the geolocation sharing adoption per

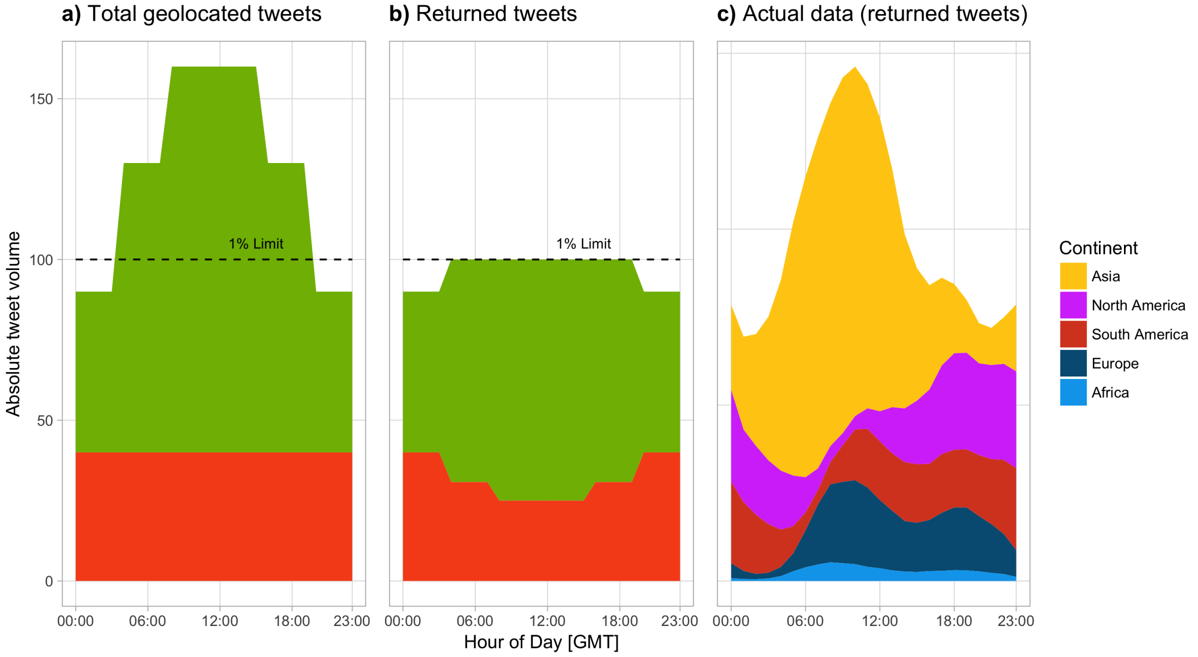


Figure 4.2: A query bias introduced when geolocated tweets exceed the 1% limit of Twitter traffic. a) Shows a simplified absolute geolocated tweet volumes for two region in different time zones in stacked position at GMT. b) Shows how many tweets are absolutely returned by a query per area. In c) the returned tweets per continent from T_{world} at GMT are given.

continent determine the share of geolocated tweets.

The phenomenon described biases a geotemporal analysis only when geolocation share varies and partly or permanently lies above the 1% threshold. If this is the case in a systematic manner (e. g. due to varying geotag adoption rates in space as described by Sloan and Morgan (2015)), any geotemporal analyses have to be questioned. In this thesis a global inspection of Twitter peak hours and clustering of the daily routine was done. The results, interpreted in this light, lead to a different conclusion. The *observed* spatially varying Twitter activity signatures are the product of spatially varying Twitter penetration and geolocation sharing adoption in combination with *real* spatially variable Twitter activity signatures. The spatial clustering of similar Twitter activity signatures as given in Figure 3.5 could then be explained with their geographical position relative to other areas, whose Twitter activity influence the 1% value.

This theory is corroborated by the fact that the spatial overlap with GUF is more stable than the purely temporal patterns. Figure 3.7 shows that the GUF correlation in T_{eu} and T_{world} show a similar pattern (especially the behaviour of weekends) despite that the absolute volumes as given in Figure 3.2 look very different. T_{eu} is likely to contain a more complete set of this data, since the query area is smaller than in T_{world}. A smaller query extent reduces the chances for the relevant tweets to exceed the 1% limit. Hence, Twitter data from different geographic query extents can not be compared.

The same problem has been described by Driscoll and Walker (2014) for specific keywords that at some point exceed the 1% limit. In a geographic context, this problem is particularly interesting, because it biases the amount of data geographically. Research relying on spatiotemporal Twitter activity signatures has to take into account, that the query extent used in Twitter's Streaming API determines the amount of data returned.

Chapter 5

Conclusion

This thesis is an attempt to correlate the distribution of geolocated tweets with human settlements. The influence of spatiotemporal patterns, user characteristics, weather conditions and settlement size was investigated. With the help of these elements tweets on human settlement could be distinguished from tweets outside human settlement using a random forest predictor. Finally, the potential coverage of Twitter data on human settlements was estimated using Monte Carlo simulation. The work can be described as explorative in the sense that there is hardly any research using densities of precise geolocation to classify land cover.

Main Findings

It can be concluded that geolocated Twitter data is a strong indicator for human settlements. Further findings are:

- The fact that people are outside settlements on weekends and in summer seems to be reflected in Twitter data
- Spatial overlap of Twitter data and human settlements is spatially variable
- While temporal Twitter activity signatures are differing dependent on the Twitter data set, the temporal patterns of spatial overlap are robust
- Twitter data cluster in centres with a high degree of urbanization
- Twitter data tend to be on or close to human settlement
- Nonpersonal users contribute a considerable amount to the Twitter data
- Twitter's Streaming API can generate artificial patterns due to its 1% limit

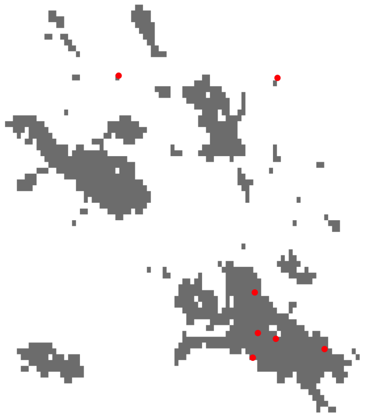


Figure 5.1: An illustrative excerpt from T_{world} (red) overlapping GUF.

However, at the current state of Twitter it is not possible to accumulate a data sample that would merely allow a global classification of human settlements. Twitter data is highly clustered in space and does never detect all settlements. Even in regions with a high Twitter penetration, many settlements are not covered with tweets. An excerpt of GUF and T_{world} Figure 5.1 summarizes this conclusion. This thesis started with the assumption that where there's built-up area, there are probably people, and where there's no built-up area, there are probably no people. This assumption is corroborated by Twitter data. As of the results presented, let the statement be completed: *Where there are tweets, there are probably people. But where there are no tweets, there are not necessarily no people!*

Future Work

Future work on Twitter data and social media has to focus on five main issues:

First, the user community of Twitter has to be analysed more thoroughly. This thesis suggests that the amount of data created by bots is tremendous. There are no standard cleansing methods. Means of cleansing have to be developed that can be easily implemented by researchers. Prerequisite for data cleansing by user is a proper definition of user categories that goes beyond *bot or not*. There are far more kinds of users to be distinguished. This would lead to more valid results and ensure that studies can be compared more effectively.

Second, further studies that focus on globally existing patterns in Twitter data are required to pull local studies out of their narrow scope. Differences in the results of different studies with geographic context have not yet been discussed in the light of cultural impact on e. g. Twitter activity signatures. Little is known about geographical differences in the usage of social media. Today studies of different geographic locations can only be compared with great caution.

Third, the potential of social media data as proxy for human settlements is not limited to Twitter data.

It would be interesting to do the same analysis with data from different social media platforms. This would lead to a deepened understanding of the spatial overlap of human settlements and social media data. Moreover, other social media platforms show complementing penetration levels across countries.

Fourth, the overlap of human settlement and tweets can be used in other research contexts. For example tweet clusters of tweets outside human settlement are likely to indicate scenic or tourist hot spots.

And fifth, the scope of the work can be widened in several regards. The *where* of Twitter data can possibly more exhaustively be explained using transportation networks. Tweet content may also be a source of information in this regard that has not been exploited in this thesis.

Closing Words

In literature the legitimation for studying geotemporal social media data like Twitter are manifold. One very popular motivation is to help urban planners in sensing land use and land cover. (Frias-Martinez and Frias-Martinez, 2014), (Longley, Adnan, and Lansley, 2015)) If, by any means, something can be concluded with certainty, then it is that Twitter data is not suitable for assisting urban planning authorities. Any results from analyses of this data are highly biased. (1) Tweeters are only a small fraction of the population that is by no means representative of the whole population. (2) The share of tweets from bots and services as well as very frequent tweeters is overwhelming (Blank, 2016). This circumstance is furthermore complicated by the fact that disambiguation of individual Twitter users from bots and services is not easy. Any decision making based on this information will be guided by a small fraction of the population and non-humanoid accounts. A democratic legitimization to use this data in public authorities is not given[1].

Nevertheless, the study of social media data should be an important research goal for many reasons: (1) If the number of social media users increase in the the future, more data will be generated. More data leads to more accurate insights into real-world processes. (2) Social media data seems to be highly connected to real-world processes at a first glance. e. g. the spatial distribution widely correlates with human settlement. The diurnal patterns match with human habit of sleeping and activity. It is tempting to relate any pattern found in this data to the real-world. A critical view at social media data is crucial to detect any snap judgements done in research before it becomes state of the art and eventually influences individual's lives. (3) The potential of data generated by people is undoubtedly given and with it dangers. Other sources of social media platforms like Facebook, Whatsapp but also location data collected by Google are probably more representative of a population. On one hand in this data lies much more information that could be exploited and yield better results. On the other hand the fact that this data is more personal and bears more valuable information makes the data unsuitable for public access. But the data is also undisclosed because it is the capital of large companies. The question is raised about who gets access to social media data. Profit-driven IT companies like Google hold tremendous amounts

[1] For a discussion on this topic see Salkin (2011)

of highly personal data. Researchers on the other hand are stuck with few sources of very biased less accurate data. This hampers knowledge generation from valuable information that exists.

Widening the view is also essential when assessing human settlements with big data. Twitter data is not a globally applicable data stream for this purpose. Whatsoever, the results have to be regarded in a larger context of a rapidly changing world with social media platforms emerging and more and more people using them. As Twitter indicates, people are accessing social media mainly from within human settlements. Automatic means allow to remove noise that is generated by undesired users and outliers that are often outside human settlements. The preconditions for human settlement classification are given and maybe existing inherently in social media data. The availability of denser data with higher coverage will raise the question of classification of human settlements with social media data again in future.

Bibliography

Adnan, Muhammad, Guy Lansley, and Paul A Longley (2013). “Twitter geodemographic analysis of ethnicity and identity in Greater London”. In: *16th AGILE international conference on geographic information science*. Vol. 44, pp. 1–7.

Akoka, Jacky, Isabelle Comyn-Wattiau, and Nabil Laoufi (2017). “Research on Big Data – A systematic mapping study”. In: *Computer Standards & Interfaces* 54.January, pp. 105–115. DOI: `10.1016/j.csi.2017.01.004`. URL: `http://linkinghub.elsevier.com/retrieve/pii/S0920548917300211`.

Allen, Chris et al. (2016). “Applying GIS and Machine Learning Methods to Twitter Data for Multiscale Surveillance of Influenza”. In: *PLOS ONE* 11.7. Ed. by Mansour Ebrahimi, pp. 1–10. DOI: `10.1371/journal.pone.0157734`. URL: `http://dx.plos.org/10.1371/journal.pone.0157734`.

Arribas-bel, Daniel et al. (2015). “Cyber Cities : Social Media as a Tool for Understanding Cities”. In: *Applied Spatial Analysis* 8, pp. 231–247. DOI: `10.1007/s12061-015-9154-2`.

Baer, Manuel (2017). “Development, Implementation, Assessment and Analysis of a Real-Time Tile-Based Location-Based Game for Geographic Information Mining Regarding Land Cover Data”. Master Thesis. University of Zurich.

Barrington-Leigh, Christopher and Adam Millard-Ball (2017). “The world’s open-source street map is more than 80% complete”. In: *Plos One* 12.8, pp. 1–21. DOI: `10.1371/journal.pone.0180698`. URL: `http://people.ucsc.edu/%7B~%7Dadammb/OSM/`.

Bhaduri, Budhendra, Edward Bright nad Phillip Coleman, and Jerome Dobson (2002). “LandScan: Locating People is What Matters”. In: *Geoinformatics* 5.2, pp. 34–37.

Blanford, Justine I et al. (2015). “Geo-Located Tweets. Enhancing Mobility Maps and Capturing Cross-Border Movement”. In: *PLOS ONE* 10.6. Ed. by Renaud Lambiotte, pp. 1–16. DOI: `10.1371/journal.pone.0129202`. URL: `http://dx.plos.org/10.1371/journal.pone.0129202`.

Blank, Grant (2016). “The Digital Divide Among Twitter Users and Its Implications for Social Research”. In: *Social Science Computer Review*, pp. 1–19. DOI: `10.1177/0894439316671698`. URL: `http://ssc.sagepub.com/cgi/doi/10.1177/0894439316671698`.

Breiman, Leo (2001). “Random forests”. In: *Machine Learning* 45.1, pp. 5–32. DOI: `10.1023/A:1010933404324`. arXiv: `/dx.doi.org/10.1023{\%}2FA{\%}3A1010933404324 [http:]`.

Chakraborty, Arnab et al. (2015). “Open data for informal settlements: Toward a users guide for urban managers and planners”. In: *Journal of Urban Management* 4.2, pp. 74–91. DOI: 10.1016/j.jum.2015.12.001. URL: http://linkinghub.elsevier.com/retrieve/pii/S2226585615000229.

Chesnokova, Olga, Mario Nowak, and Ross S Purves (2017). “A crowdsourced model of landscape preference”. In: *13th International Conference on Spatial Information Theory (COSIT 2017)*. 19. Dagstuhl: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, pp. 1–13.

Chu, Zi et al. (2010). “Who is Tweeting on Twitter: Human, Bot, or Cyborg?” In: *Acsac 2010Proceedings of the 26th Annual Computer Security Applications Conference*, pp. 21–30. DOI: 10.1145/1920261.1920265. URL: http://portal.acm.org/citation.cfm?doid=1920261.1920265.

Driscoll, Kevin and Shawn Walker (2014). “Working within a black box: Transparency in the collection and production of big twitter data”. In: *International Journal of Communication* 8, pp. 1745–1764.

Duke, Eilish and Christian Montag (2017). “Smartphone addiction, daily interruptions and self-reported productivity”. In: *Addictive Behaviors Reports* 6.October 2016, pp. 90–95. DOI: 10.1016/j.abrep.2017.07.002. URL: http://linkinghub.elsevier.com/retrieve/pii/S2352853217300159.

Esch, Thomas et al. (2017). “Breaking new ground in mapping human settlements from space - The Global Urban Footprint”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 134.December 2017, pp. 30–42. URL: https://arxiv.org/abs/1706.04862.

Frias-Martinez, Vanessa and Enrique Frias-Martinez (2014). “Spectral clustering for sensing urban land use using Twitter activity”. In: *Engineering Applications of Artificial Intelligence* 35, pp. 237–245. DOI: 10.1016/j.engappai.2014.06.019.

Frias-Martinez, Vanessa, Victor Soto, et al. (2012). “Characterizing urban landscapes using geolocated tweets”. In: *Proceedings - 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust and 2012 ASE/IEEE International Conference on Social Computing, SocialCom/PASSAT 2012*, pp. 239–248. DOI: 10.1109/SocialCom-PASSAT.2012.19.

Fritz, Steffen et al. (2009). “Geo-Wiki.Org: The Use of Crowdsourcing to Improve Global Land Cover”. In: *Remote Sensing* 1.3, pp. 345–354. DOI: 10.3390/rs1030345. URL: http://www.mdpi.com/2072-4292/1/3/345/.

García-Palomares, Juan Carlos et al. (2018). “City dynamics through Twitter: Relationships between land use and spatiotemporal demographics”. In: *Cities* 72.February, pp. 310–319. DOI: 10.1016/j.cities.2017.09.007. arXiv: 1705.07956. URL: http://dx.doi.org/10.1016/j.cities.2017.09.007.

Goodchild, Michael F (2013). “The quality of big (geo)data”. In: *Dialogues in Human Geography* 3.3, pp. 280–284. DOI: 10.1177/2043820613513392. URL: http://journals.sagepub.com/doi/10.1177/2043820613513392.

Graham, Mark and Taylor Shelton (2013). “Geography and the future of big data, big data and the future of geography”. In: *Dialogues in Human Geography* 3.3, pp. 255–261. DOI: 10.1177/2043820613513121. URL: http://journals.sagepub.com/doi/10.1177/2043820613513121.

Graham, Mark, Monica Stephens, and Scott Hale (2013). “Featured Graphic. Mapping the Geoweb: A Geography of Twitter”. In: *Environment and Planning A* 44, pp. 100–102. DOI: 10.1068/a45349. URL: http://journals.sagepub.com/doi/10.1068/a45349.

Grier, Chris et al. (2010). “@ spam : The Underground on 140 Characters or Less”. In: *Proceedings of the 17th ACM conference on Computer and communications security*, pp. 27–37. DOI: 10.1145/1866307.1866311. URL: http://portal.acm.org/citation.cfm?id=1866307.1866311.

Hahmann, Stefan, Ross Purves, and Dirk Burghardt (2014). “Twitter location (sometimes) matters: Exploring the relationship between georeferenced tweet content and nearby feature classes”. In: *Journal of Spatial Information Science* 9. DOI: 10.5311/JOSIS.2014.9.185. URL: http://josis.org/index.php/josis/article/view/185.

Hawelka, Bartosz et al. (2014). “Geo-located Twitter as proxy for global mobility patterns”. In: *Cartography and Geographic Information Science* 41.3, pp. 260–271. DOI: 10.1080/15230406.2014.890072. URL: http://www.tandfonline.com/doi/full/10.1080/15230406.2014.890072.

Helwig, Nathaniel E et al. (2015). “Analyzing spatiotemporal trends in social media data via smoothing spline analysis of variance”. In: *Spatial Statistics* 14, pp. 491–504. DOI: 10.1016/j.spasta.2015.09.002. URL: http://linkinghub.elsevier.com/retrieve/pii/S2211675315000767.

Jendryke, Michael et al. (2017). “Putting people in the picture: Combining big location-based social media data and remote sensing imagery for enhanced contextual urban information in Shanghai”. In: *Computers, Environment and Urban Systems* 62, pp. 99–112. DOI: 10.1016/j.compenvurbsys.2016.10.004. URL: http://dx.doi.org/10.1016/j.compenvurbsys.2016.10.004.

Jenkins, Andrew et al. (2016). “Crowdsourcing a Collective Sense of Place”. In: *PLOS ONE* 11.4. Ed. by Tobias Preis, e0152932. DOI: 10.1371/journal.pone.0152932. URL: http://dx.plos.org/10.1371/journal.pone.0152932.

Jin, Xiaolong et al. (2015). “Significance and Challenges of Big Data Research”. In: *Big Data Research* 2.2, pp. 59–64. DOI: 10.1016/j.bdr.2015.01.006. URL: http://dx.doi.org/10.1016/j.bdr.2015.01.006.

Klotz, Martin et al. (2017). “Digital deserts on the ground and from space”. In: *Urban {Remote} {Sensing} {Event} ({JURSE}), 2017 {Joint}*. IEEE, pp. 1–4. URL: http://ieeexplore.ieee.org/abstract/document/7924562/.

Kumar, Shamanth, Fred Morstatter, and Huan Liu (2013). “Analyzing Twitter Data”. In: *Twitter Data Analytics*. May. Springer, pp. 35–48.

Lagache, Thibault et al. (2013). “Analysis of the Spatial Organization of Molecules with Robust Statistics”. In: *PLoS ONE* 8.12, e80914. DOI: 10.1371/journal.pone.0080914. URL: http://dx.plos.org/10.1371/journal.pone.0080914.

Leetaru, Kalev et al. (2013). “Mapping the global Twitter heartbeat: The geography of Twitter”. In: *First Monday* 18.5. URL: http://firstmonday.org/article/view/4366/3654?%7B%5C_%7D%7B%5C_%7Dhstc=225085317..1436140800094.1436140800095.1436140800096.1%7B%5C&%7D%7B%5C_

%7D%7B%5C_%7Dhssc=225085317.1.1436140800097%7B%5C&%7D%7B%5C_%7D%7B%5C_%7Dhsfp=1314462730.

Li, Linna, Michael F Goodchild, and Bo Xu (2013). "Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr". In: *Cartography and Geographic Information Science* 40.2, pp. 61–77. DOI: 10.1080/15230406.2013.777139. URL: http://www.tandfonline.com/doi/abs/10.1080/15230406.2013.777139.

Lin, Jie and Robert G Cromley (2015). "Evaluating geo-located Twitter data as a control layer for areal interpolation of population". In: *Applied Geography* 58, pp. 41–47. DOI: 10.1016/j.apgeog.2015.01.006. URL: http://linkinghub.elsevier.com/retrieve/pii/S0143622815000107.

Lloyd, Alyson and James Cheshire (2017). "Deriving retail centre locations and catchments from geo-tagged Twitter data". In: *Computers, Environment and Urban Systems* 61, pp. 108–118. DOI: 10.1016/j.compenvurbsys.2016.09.006. URL: http://linkinghub.elsevier.com/retrieve/pii/S0198971516302666.

Longley, Paul A and Muhammad Adnan (2016). "Geo-temporal Twitter demographics". In: *International Journal of Geographical Information Science* 30.2, pp. 369–389. DOI: 10.1080/13658816.2015.1089441. URL: http://www.tandfonline.com/doi/full/10.1080/13658816.2015.1089441.

Longley, Paul A., Muhammad Adnan, and Guy Lansley (2015). "The geotemporal demographics of twitter usage". In: *Environment and Planning A* 47.2, pp. 465–484. DOI: 10.1068/a130122p.

Miao, Zelang (2017). "Towards an Automatic Framework for Urban Settlement Mapping from Satellite Images : Applications of Geo-referenced Social Media and One Class Classification". In: *Geophysical Research Abstracts* 19.EGU2017-4767.

Modoni, Gianfranco E. and Davide Tosi (2016). "Correlation of weather and moods of the Italy residents through an analysis of their tweets". In: *Proceedings - 2016 4th International Conference on Future Internet of Things and Cloud Workshops, W-FiCloud 2016*, pp. 216–219. DOI: 10.1109/W-FiCloud.2016.53.

Morstatter, Fred et al. (2013). "Is the sample good enough? Comparing data from twitter's streaming api with twitter's firehose". In: *Proceedings of ICWSM*. Cambridge. URL: https://arxiv.org/abs/1306.5204.

Patel, Nirav N et al. (2016). "Improving Large Area Population Mapping Using Geotweet Densities: Improving Large Area Population Mapping Using Geotweet Densities". In: *Transactions in GIS*. DOI: 10.1111/tgis.12214. URL: http://doi.wiley.com/10.1111/tgis.12214.

Potere, David et al. (2009). "Mapping urban areas on a global scale: Which of the eight maps now available is more accurate?" In: *International Journal of Remote Sensing* 30.24, pp. 6531–6558. DOI: 10.1080/01431160903121134.

Rios, Miguel (2013). *The Geography of Tweets*. URL: https://blog.twitter.com/official/en%7B%5C_%7Dus/a/2013/the-geography-of-tweets.html (visited on 07/10/2017).

Roxin, A. et al. (2007). "Survey of Wireless Geolocation Techniques". In: *IEEE Globecom Work- shops*, pp. 1–9. URL: http://www.gsem.fr/old/download/ROX%7B%5C_%7D07b.pdf.

Salkin, Patricia E (2011). “Social networking and land use planning and regulation: Practical benefits, pitfalls, and ethical considerations”. In: *Pace L. Rev.* 31, p. 54. URL: `http://heinonline.org/hol-cgi-bin/get%7B%5C_%7Dpdf.cgi?handle=hein.journals/pace31%7B%5C&%7Dsection=5`.

Schneider, A., M. A. Friedl, and D. Potere (2009). “A new map of global urban extent from MODIS satellite data”. In: *Environmental Research Letters* 4.4. DOI: `10.1088/1748-9326/4/4/044003`.

Sloan, Luke and Jeffrey Morgan (2015). “Who Tweets with Their Location? Understanding the Relationship between Demographic Characteristics and the Use of Geoservices and Geotagging on Twitter”. In: *PLOS ONE* 10.11. Ed. by Tobias Preis, e0142209. DOI: `10.1371/journal.pone.0142209`. URL: `http://dx.plos.org/10.1371/journal.pone.0142209`.

Small, Christopher (2003). “High spatial resolution spectral mixture analysis of urban reflectance”. In: *Remote Sensing of Environment* 88.1-2, pp. 170–186. DOI: `10.1016/j.rse.2003.04.008`.

Soliman, Aiman et al. (2017). “Social sensing of urban land use based on analysis of Twitter users’ mobility patterns”. In: *PloS ONE* 12.7, e0181657. DOI: `10.1371/journal.pone.0181657`.

Steiger, Enrico, João Porto de Albuquerque, and Alexander Zipf (2015). “An Advanced Systematic Literature Review on Spatiotemporal Analyses of Twitter Data”. In: *Transactions in GIS* 19.6, pp. 809–834. DOI: `10.1111/tgis.12132`.

Steiger, Enrico, René Westerholt, et al. (2015). “Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data”. In: *Computers, Environment and Urban Systems* 54, pp. 255–265. DOI: `10.1016/j.compenvurbsys.2015.09.007`. URL: `http://dx.doi.org/10.1016/j.compenvurbsys.2015.09.007`.

Takahashi, Tetsuro, Shuya Abe, and Nobuyuki Igata (2011). “Can Twitter be an alternative of real-world sensors?” In: *International Conference on Human-Computer Interaction*. Springer, pp. 240–249. URL: `http://link.springer.com/chapter/10.1007/978-3-642-21616-9%7B%5C_%7D27`.

Takhteyev, Yuri, Anatoliy Gruzd, and Barry Wellman (2012). “Geography of Twitter networks”. In: *Social Networks* 34.1, pp. 73–81. DOI: `10.1016/j.socnet.2011.05.006`. URL: `http://linkinghub.elsevier.com/retrieve/pii/S0378873311000359`.

Wang, Yandong et al. (2016). “Mapping Dynamic Urban Land Use Patterns with Crowdsourced Geo-Tagged Social Media (Sina-Weibo) and Commercial Points of Interest Collections in Beijing, China”. In: *Sustainability* 8.11, p. 1202. DOI: `10.3390/su8111202`. URL: `http://www.mdpi.com/2071-1050/8/11/1202`.

Wong, David W. S. (2004). “The Modifiable Areal Unit Problem (MAUP)”. In: *WorldMinds: Geographical Perspectives on 100 Problems*. Ed. by D. G. Janelle, B. Warf, and K. Hansen. 1934. Dordrecht: Springer, pp. 571–575. ISBN: 1402016123. DOI: `10.1007/978-1-4020-2352-1_93`. URL: `http://link.springer.com/10.1007/978-1-4020-2352-1%7B%5C_%7D93`.

Yang, Jaewon and Jure Leskovec (2011). “Patterns of Temporal Variation in Online Media”. In: *ACM International Conference on Web Search and Data Minig (WSDM)*. Hong Kong: Stanford InfoLab, pp. 177–186.

Personal Declaration

I hereby declare that the submitted thesis is the result of my own, independent work. All external sources are explicitly acknowledged in the thesis.

....