Identifying home locations in human mobility data: an open-source R package for comparison and reproducibility

Qingqing Chen¹, Ate Poorthuis²

¹ Department of Geography, University at Buffalo - SUNY, 105 Wilkeson Quad, Ellicott Complex, North Campus, Buffalo, NY 14261, USA

² Department of Earth and Environmental Sciences, KU Leuven, Celestijnenlaan 200E, 3001, Leuven, Belgium

Published as: Qingqing Chen, Ate Poorthuis (2021). Identifying home locations in human mobility data: an open-source R package for comparison and reproducibility. *International Journal of Geographical Information Science* [Online First]

<https://doi.org/10.1080/13658816.2021.1887489>

Abstract

Identifying meaningful locations, such as home or work, from human mobility data has become an increasingly common prerequisite for geographic research. Although location-based services (LBS) and other mobile technology have rapidly grown in recent years, it can be challenging to infer meaningful places from such data, which – compared to conventional datasets – can be devoid of context. Existing approaches are often developed ad-hoc and can lack transparency and reproducibility. To address this, we introduce an R software package for inferring home locations from LBS data. The package implements pre-existing algorithms and provides building blocks to make writing algorithmic ‘recipes’ more convenient. We evaluate this approach by analyzing a de-identified LBS dataset from Singapore that aims to balance ethics and privacy with the research goal of identifying meaningful locations. We show that ensemble approaches, combining multiple algorithms, can be especially valuable in this regard as the resulting patterns of inferred home locations closely correlate with the distribution of residential population. We hope this package, and others like it, will contribute to an increase in use and sharing of comparable algorithms, research code and data. This will increase transparency

¹ Email address: qchen47@buffalo.edu

² Email address: ate.poorthuis@kuleuven.be

and reproducibility in mobility analyses and further the ongoing discourse around ethical big data research.

Keywords: identifying home locations; location-based services (LBS); Twitter; R package; comparison and reproducibility

Introduction

The analysis of human mobility has become an integral part of a variety of academic (sub-) disciplines. Such analyses can range from understanding commuting flows in cities (e.g. Hincks *et al.* 2018) to migration flows within and between countries (e.g. Belyi *et al.* 2017) and have applications in fields beyond Geography that range from urban planning to epidemiology (e.g. Balcan *et al.* 2009, Maat *et al.* 2016). Although analyses of human mobility have been conducted for well over a century (cf. Minard's 1858 map of global migration), they conventionally rely on resource-intensive data sources such as administrative data, census, or manually collected travel surveys³. The arrival of the Global Position System (GPS) has opened up a host of new possibilities for mobility analysis by making it more affordable and convenient to collect spatio-temporally precise and accurate mobility data (Kwan 2004), which was soon followed by data derived from mobile phone cell towers (e.g. Ahas *et al.* 2015). In recent years, the onset of social media platforms and location-based services (LBS), often including a geolocation feature, has created another set of potential data sources for mobility analysis (e.g. Tu *et al.* 2017).

With the proliferation of mobility data sources also come new challenges. One of the foremost differences between many new data sources and more conventional fieldwork-based data is that there is no explicit or direct interaction between researchers and research subjects.

³ The Chicago Area Transportation Study, which is one of the first examples of a large-scale travel survey, cost \$650,000 in 1956 for just the survey fieldwork (Black 1990).

This has several implications, including for privacy and ethics, but also means that it is challenging or even impossible to ask additional questions to a research ‘participant’. Such sources may indeed yield mobility data in the form of precise locations and timestamps but without specific meaning attached to those locations. Social media users rarely publish additional location information (for both technical and privacy reasons), which makes inferences about places relatively challenging. We know when and where a person went. But *what* were they doing there? *Why* did they go there? *How* did they get there? Depending on the research question at hand, the answers to some of these questions can be quite essential. However, new sources of mobility data might not be able to provide these answers because they are, in many ways, devoid of social context.

To address this, a new line of research has developed that is focused on inferring, what Ahas et al. (2010) call, ‘meaningful’ locations from mobility data. The most obvious example of such a location is the place of a person’s home or residence, but this notion can be extended to work, recreation, and other types of locations. In the last 10 years, a myriad of different approaches to inferring meaningful locations has been developed, sometimes as the main impetus for methodological articles (e.g. Ahas *et al.* 2010, Cheng *et al.* 2010, McGee *et al.* 2013) as well as a means-to-an-end in empirical analyses of human mobility (e.g. Huang and Wong 2016, Siła-Nowicka *et al.* 2016, Xin and MacEachren 2020).

Although similarities between approaches do exist, a consensus or single best approach has not emerged in the current state-of-the-art. What’s more, the actual algorithms used are not always discussed in detail within publications and the source code is rarely released publicly, which makes comparing algorithms as well as reproducing results all the more challenging. This is further complicated by the fact that a ground-truth for the evaluation of the inference process often simply does not exist and mobility data itself is rarely made available to the larger research community because of both technical, ethical and social reasons. Together, this has

resulted in a situation where an increasing number of papers use a process of location inference that is implemented with unique, custom algorithms that are difficult to subsequently evaluate or re-use by the larger research community.

To alleviate this issue, we developed an R package named ‘*homelocator*’ aimed at bringing consistency and replicability to the domain of inferring meaningful locations. The package allows users to write structured, algorithmic ‘recipes’ or use existing, built-in recipes to identify meaningful locations in LBS data and enables the comparison across different algorithms on the same dataset. To contextualize this work, we first conduct a broad review of the different existing approaches for identifying locations and show how the current state-of-the-art relies on different, often bespoke solutions. We then discuss the approach and design of the R package and its implementation of several pre-existing algorithms. To provide an empirical evaluation, we describe an aggregated and de-identified LBS dataset, consisting of data from Singapore sent between 2012-2016. The dataset is made available in conjunction with the R package to provide a baseline for a fair and quantitative comparison across algorithms and approaches. We illustrate our approach by applying four existing algorithms on this dataset and highlight the affordances provided by a subsequent ensemble approach, which enhances the reliability of the inferred meaningful locations.

Existing approaches to detecting meaningful locations

Human mobility is of interest to many researchers because it captures part of the interactions between people and their spatial context. A movement by a person is not simply or randomly embedded into an abstract geographical space but has a meaning attached to both the movement itself and the places it connects to (Sila-Nowicka *et al.* 2016). People with different cultural, social, and economic characteristics will ultimately create different underlying patterns of human spatial activity (Huang and Wong 2016). Understanding these activity patterns may help to better understand the urban spatial structure, the usage and functionality of urban areas, and

the needs of public transportation services for different population groups (Siła-Nowicka *et al.* 2016). For example, Yuan and Raubal (2016) explore activity patterns to understand the socio-spatial relationships and correlation between demographic factors and the usage of urban space for different communities. Xin and MacEachren (2020) recently proposed a workflow to understand event-related travel behaviour. Many of these applications rely on a common step in their analysis: the identification of meaningful locations in mobility data. Accordingly, identifying meaningful places in travel trajectories has become relatively commonplace in mobility analyses. These meaningful places can be home, work offices, schools, health centers, or other ‘third’ places like coffee shops, bars, museums and so on where people spend their time for leisure.

Although there are many sources for mobility data, we focus on data derived from social media platforms and location-based services (LBS) in this paper. Smartphones have become ubiquitous devices that go wherever their users go, and the usage of social media such as Twitter, Facebook, Wechat and Weibo has grown dramatically. Thus, mobile phone location, as well as a notable amount of information from social media (e.g. spatio-temporal timestamps, user profiles, micro-blogs, etc) has provided a prolific source of data for analyzing meaningful locations for large groups of the population. It should be noted that such data cannot be treated the same as a well-designed random survey and biases are likely to exist (Longley and Adnan 2016, Lovelace *et al.* 2016). Nonetheless, when cognizant of this bias, useful insights about social processes can still be derived, especially when other datasets might not be readily available (cf. Crampton *et al.* 2013). In this section, we discuss prior work on identifying meaningful locations for users based on three specific categories: (1) approaches based on the content of messages or user profiles; (2) approaches based on underlying social networks between users; and (3) approaches based on the spatio-temporal features of messages. We

summarize a selection of approaches across the three categories in *Table 1* and discuss these in more detail in the next section.

Publication	Main Approach			Localization Focus		Spatial Granularity	Data Source	Ground Truth	Data/Code Publicly Available	Implemented in homelocator
	Message / Profile content	Social Network	Spatio-temporal Attributes	Home Location	Tweet Location					
Eisenstein et al. (2010)						S/R	Twitter	GPS-tag in data		
Backstrom et al. (2010)						C	Facebook	Self-reported		
Cheng et al. (2010)						C	Twitter	GPS-tag in data		
Ahas et al. (2010)						AP	Mobile phone	Population Register		
Chandra et al. (2011)						C	Twitter	GPS-tag in data		
Kinsella et al. (2011)						S/C/Z	Twitter	GPS-tag in data		
Hecht et al. (2011)						S	Twitter	Self-reported		
Chang et al. (2012)						C	Twitter	Manually labelled		
Sadilek et al. (2012)						C	Twitter	Cross-validation		
Li et al. (2012)						C	Twitter	Self-reported		
Rout et al. (2013)						C	Twitter	Self-reported		
Schulz et al. (2013)						C	Twitter	GPS-tag in data		
Jurgens (2013)						P	Twitter	Self-reported		
McGee et al. (2013)						C	Twitter	GPS-tag in data		
Compton et al. (2014)						C	Twitter	GPS-tag in data/Self-reported		
Li et al. (2014)						S/C/ST	Twitter	Survey		
Katragadda et al. (2014)						C	Twitter	Manually labelled		
Mahmud et al. (2014)						S/R/C	Twitter	GPS-tag in data		
Efstathiades et al. (2015)						Z	Twitter	Self-reported		
Lin et al. (2018)						P	Twitter	Residential Parcels		

☒ Yes ☐ No Note: R (Region), S (State), C (City), AP (Anchor-Point), Z (Zip code), ST (Street), P (Geo-point)

Table1. Overview of existing approaches to detecting meaningful locations

Based on the content of messages or user profiles

Content-based methods identify meaningful location(s) based on certain characteristics of the content of social media messages. These can range from specific (dialect) terms linked to a local geographic context, place names and toponyms, or words, phrases or topics with certain geographical associations (Eisenstein *et al.* 2010). In this way, this group of methods does not require more conventional personal identifying information, IP addresses, or location data which may alleviate privacy concerns in some situations but also raises ethical questions (cf. our discussion on privacy and ethics in the subsequent section). Nonetheless, this approach faces many challenges in terms of the semantic complexities in natural language. In the case of social media, the issue of word sparsity – with Twitter’s 140/280-character limit being a prime example – creates an additional complication with non-standard/shorthand words and acronyms.

Additionally, different languages may be used (even in the same message) and the self-reported location may be incorrect, vague, incomplete or even nonsensical. All of these issues are especially pronounced in places with smaller populations that may ‘suffer’ from a data sparsity problem (Cheng *et al.* 2010). To overcome these difficulties when conducting content analysis, researchers have proposed three approaches to infer location information of users at different levels of granularity from country to city or even zip code level. Roughly speaking, these are based on (1) a gazetteer lookup; (2) probabilistic models; and (3) multi-method and multi-data source approaches.

The first, and perhaps most straightforward approach is to analyse words or terms in the content of the social media message and compare them with a gazetteer like *Geonames* (Schulz *et al.* 2013) or use online services like *Yahoo! Placemaker* (Kinsella *et al.* 2011) to do the same. In effect, this identifies a spatial entity based on one or more place names in the message. Often this ‘naive’ technique is used to determine a baseline or reference to which other algorithms are compared. For example, Schulz *et al.* (2013) use multiple gazetteers to match references in social media messages with known geographic entities. Multiple references are combined through a ‘polygon stacking’ technique – effectively a spatial voting mechanism – to estimate the location of the message. Kinsella *et al.* (2011) utilized *Yahoo! Placemaker* to identify specific geographic information in both tweets and the self-reported location in a user's profile. Although the gazetteer approach is relatively simple, it has some obvious limitations – one of which is that it cannot identify locations that are simply not part of the gazetteer.

Second, building forth on this ‘naive’ gazetteer matching, other approaches employ statistical models and methods on the message content before ultimately matching a location to a gazetteer entry. For example, Hecht *et al.* (2011) use a multinomial naive Bayes model to classify words with a regional focus. They first develop a ground truth dataset, as we have seen in the previous paragraph, by comparing the self-reported locations in user profiles with both

Wikipedia and Yahoo geocoders (Hecht *et al.* 2011). They then train a machine learning classifier on the text of the tweets for those users to estimate both country and state location. Such a classifier can potentially be used to infer locations for users who do not self-report their location.

As mentioned before, the self-reported locations in user profiles can be broad or unreliable, which makes it difficult to infer locations at granularities finer than the country or state level. A promising method in this respect adopts a probabilistic framework that ties location probabilities to specific words, phrases, topics and local terms. For example, if a user tweets words like ‘*Merlion*’ and ‘*red dot*’ – words that reference cultural markers in Singapore, Singapore is more likely to be the user's home location. More specifically, the probabilistic framework refers to a statistical predictive model that infers users' home locations by classifying words based on their spatial usage pattern or by capturing topics with high regional affinities. For example, Eisenstein *et al.* (2010) take a topic modelling approach by assuming individual word tokens can reflect both the topic and the geographical aspects of that topic. Other researchers focus on individual words specific to a particular location instead of broader topics. The probability distribution of each word across locations is then estimated to produce a per-user location probability across all possible locations. The location with the maximum probability can then be labelled as the user's home location (Cheng *et al.* 2010, Chang *et al.* 2012).

Third, and linking with the other two main approaches, we find work that combines the content of the message with additional information such as time or social network connections (Chandra *et al.* 2011, Li *et al.* 2012). For example, Chandra *et al.* (2011) build on Cheng *et al.*'s (2010) work and add a consideration of the underlying social interactions to estimate the home location of a user. Specifically, they focus on both tweets and associated reply-tweets to evaluate the spatial distribution of terms throughout a complete conversation

between users. Other features like the coordinates of tweets, time zone, tweeting behaviour can also be used to create new ensemble methods that further improve the accuracy of content analysis for identifying a user's home location (Kinsella *et al.* 2011, Schulz *et al.* 2013, Mahmud *et al.* 2014).

Based on social connections and networks

Besides the content-based approach, an alternative method related to location inference leverages the underlying social network that represents following-follower or friendship relationships between users. The intuition behind the proposed method is that a user is more likely to connect to people who are, often literally, close to them in actual life. A modest but practical approach is to count the frequency of known locations among the friends of a user and assign the most popular location among friends as the user's location (Jr *et al.* 2011). However, the effectiveness of this approach has clear limitations when a user has too few or too many followers as the data is either too scarce or too abundant to make reliable inferences.

A similar but more complicated method is proposed by Jurgens (2013), in which they use the idea of label propagation for inferring locations. First, a single location is assigned to each user based on the geometric median of their locations. Second, an iterative procedure is applied in which a random location of a user is replaced with the geometric median of their friends' locations. This is repeated until convergence is reached. For users with large friend networks convergence might be difficult so Compton *et al.* (2014) applying a weighting to the social ties according to the volume of interactions between users and their friends. Then at each iteration in the label propagation, they restrict the total variation over geographic distances between connected users to a certain constraint. In this way, more reliable locations can be propagated into subsequent iterations to improve location inference accuracy.

Here too, we find approaches that employ probabilistic models or machine learning techniques to infer users' locations based on social networks. For example, Backstrom *et al.*

(2010) estimate residential locations by employing Facebook data, where self-reported addresses are used as the ground truth. They examine the relationship between geographic distance and friendship and, unsurprisingly, find an inverse correlation between distance and the possibility of being friends. Based on this relation, they built a probabilistic model to estimate a user's location by evaluating the likelihood of all locations of his or her friends and the location with the maximum likelihood is selected as the user's home location. Other researchers have tried to integrate the strength of social ties into such models through factors like the population density, the number of interactions between users and their friends, the number of followers, the distribution of different types of connections (Sadilek *et al.* 2012, McGee *et al.* 2013, Rout *et al.* 2013).

Based on spatio-temporal metadata

The third way for inferring meaningful locations leverages the spatio-temporal features of geo-tagged messages. Concretely, the method infers key locations (e.g. home, workplace, etc) of individuals based on their everyday online social activities or behaviour at different timeframes. For example, Efstathiades *et al.* (2015) divide a day into three time frames, which they call 'rest', 'leisure' and 'active' time, and assume a person is most likely to stay at home during 'rest' and 'leisure' time. As such, they assign the postal code that a user most frequently tweets from during these two time frames as their home location. Lin and Cromley (2018) use the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) approach to attempt to identify precise point-level home locations. They use the most frequently visited night-time cluster as the basis for the most likely home location. However, the algorithm is quite sensitive to its input parameters (e.g. cluster radius and minimum number of observations), particularly when data is sparse and unevenly distributed in space (Liu *et al.* 2019). Most importantly, attempting to pin-point home locations with such a fine spatial granularity raises significant

ethical issues (cf. our discussion on privacy and ethics in the subsequent section), especially when the resulting data and the methods become publicly available.

Finally, beyond LBS data, mobile phone data is another fertile source of information for meaningful location inference (Ahas *et al.* 2010), but it also faces the challenge of the uneven spatial resolution of call detail records (Järv *et al.* 2017). Perhaps more so than LBS data, the use of mobile phone data brings up serious privacy and ethics concerns (discussed in the next section).

What stands out here, is that there is a wide variety of different approaches without a clear ‘best-practice’ and this is further complicated by the fact that a ground-truth most often doesn’t exist, and it is thus difficult to evaluate the effectiveness of each approach. This has yielded a research landscape in which most work that relies on inferring meaningful locations does so in a bespoke way – and because terms and conditions often prevent the sharing of LBS data – it is challenging to evaluate the robustness of subsequent findings.

It is pertinent to mention that the adjacent field of movement ecology has made significant strides in this regard, with a recent review identifying 58 packages for the R programming language alone that are created for the analysis of animal movement (Joo *et al.* 2020). Within this field, home range estimation – the area that an animal moves across for its normal activities (e.g. eating, caring for its young, etc., see Burt 1943) is conceptually close to what is often referred to as activity space in human mobility analysis, and by extent home location detection. There are several software packages for home range analysis (e.g. Lyons *et al.* 2013, Tracey *et al.* 2014, Signer and Balkenhol 2015, Calabrese *et al.* 2016, Signer *et al.* 2018, Calenge and Fortmann-Roe 2020, etc) and this approach might also be of use in certain kinds of human mobility analysis. However, animal movement data is generally collected through GPS devices with consistent and granular temporal and spatial resolution. In human mobility analysis, this consistency and resolution is not always available and, as we discuss in

the next section, might also not be desirable to collect from an ethical standpoint. As such, taking inspiration from the software landscape in movement ecology, we develop a similar package specifically tailored for human mobility data.

Privacy and ethics

As we discussed, social media and LBS have become a mainstay in many millions of people's lives, creating detailed datasets of their spatio-temporal footprints in the process. While this has enabled a plethora of social and spatial research, this technology (and its use by researchers) also has specific privacy and ethical consequences.

Considerations for privacy have a longer history in GIScience, for example considering the trade-offs between the accuracy of an analysis and maintaining a certain degree of privacy for research participants (Kwan *et al.* 2006, Olson *et al.* 2006). Point data, even when only disclosed on published maps, can often be reverse engineered (Curtis *et al.* 2006, Kounadi and Leitner 2014). To alleviate this, geomasking (or perturbing) techniques can be applied to such data (Swanlund *et al.* 2020) but LBS provide additional challenges to privacy as they can contain more precise, expansive, and linkable data than conventional point data sets (Gao *et al.* 2019). De-anonymization and re-identification are very serious concerns. As such, with the proliferation of these platforms, the salience and specific consideration of *geoprivacy* have also risen (Keßler and McKenzie 2018).

As Elwood and Leszczynski (2011) already pointed out a decade ago, privacy on such platforms is never binary but always negotiated – with distinct potential differences between a person's expected and actual privacy. This makes charting an appropriate, ethical course for researchers in this domain certainly not trivial. In this work, we take inspiration from Zook *et al.*'s (2017) 'ten simple rules for responsible big data research'. These rules (e.g. 'guard against re-identification', 'debate the tough, ethical choices') are by no means simple - especially not

when considering the trade-offs between maintaining the privacy and the accuracy and usefulness of the subsequent analysis - but provide a useful, albeit general foundation.

Beyond privacy, there is also a broader ethical question to consider. First, for conventional social science research, clear ethical standards exist, including the concept of informed consent. Although many big data sets are technically ‘public’, this does not mean users have given their explicit and informed consent to be included in such data, let alone for its use in a myriad of research projects (cf. Zimmer’s (2018) discussion of the Kirkegaard OKCupid data release). As a post-hoc consent process might not (always) be trivial (Norval and Henderson 2017), this is an additional reason to pay careful attention to the ethics, potential consequences and harm that a dataset or analysis can result in. Second, there are significant issues around access to LBS datasets. While some data is available publicly, companies increasingly put walls around this data (Longley and Adnan 2016, Lovelace *et al.* 2016) and limit access to company employees or simply to those who pay for data products – often unbeknownst to the people contained within such data (Thompson and Warzel 2019). As Taylor (2016) aptly illustrates in the case of West African mobile phone data, who has access, and how, are important questions with complicated answers, especially as increasing datafication has intensified the connection between data and capitalism (Sadowski 2019).

It is clear that any researcher working with LBS data will have to consciously navigate these issues. An important example of what can go wrong otherwise is provided by Prabhu and Birhane (2020) in their thorough analysis of the large ImageNet dataset, which is relatively foundational in the field of computer vision. They lay bare how significant issues surrounding the sourcing of the data, its labelling and its subsequent use in AI models bring harm to several social groups. Such harm could also happen in the creation and use of LBS datasets, especially when we consider the process of inferring meaningful locations, which can be sensitive.

Although no one-size-fits-all solution exists, we proceed based on the following principles:

- (1) Spatial aggregation: for many geographic research questions and analyses, it is not necessary to know precise locations. Aggregation to the largest spatial unit that still allows for answering the question at hand should be conducted. For example, in this work we aggregate to 750-meter hexagons.
- (2) Limiting data use: many research ultimately does not need exact data on individual users but is focused on larger patterns and processes instead. Following practices common in statistics agencies, any data on locations with only a few linked users is removed. Along the same lines, inferring meaningful locations based on a person’s friends or the implicit content of their messages is far removed from the original intent with which data is shared. For this reason, we focus only on timestamp and location information.
- (3) Perturbation and random noise: to further prevent re-identification or the subsequent linking with other datasets, the only two data points used (location and timestamp) are randomly perturbed and random noise is added through swapping data points between users to provide an additional layer of de-identification and plausible deniability.

Unfortunately, protecting privacy and performing failsafe de-identification is still an open problem without a clear silver-bullet solution (Fung *et al.* 2010, Fiore *et al.* 2020). This is especially poignant for what is often referred to as ‘micro’ trajectory data, which consists of precise spatio-temporal data points collected at consistent intervals, collected through GPS or mobile phone data (cf. Thompson and Warzel 2019). Pseudonymizing and geomasking alone (Fiore *et al.* 2020), or even aggregation (de Montjoye *et al.* 2013) might not be enough to prevent a re-identification ‘attack’ on such granular data.

The LBS dataset used in this work does not have the same level of granularity as micro trajectory data. In contrast it generally covers only several dozen to several hundred data points

per user over a period of 5 years. Nonetheless, to minimize any potential harm, we have combined all of the aforementioned steps (aggregation, limiting variables, spatio-temporal perturbation, random noise) to create a de-identified dataset. This yields a scenario in which one could still re-identify a LBS user if they had prior knowledge about several significant 750m grid cell locations that user sent data from and this also uniquely separates the user from all other users in the dataset. However, beyond what you already know about that user, this would only tell you which other 750m cells they *might* have visited in the same city. After all, the perturbation and random noise provide plausible deniability and the spatial aggregation in the urban context of Singapore would prevent disclosure of any sensitive locations (each grid cell contains many hundreds of addresses).

These steps are discussed in more detail in the next section. To our knowledge, open datasets on human mobility are relatively rare in a landscape where access to data is limited or needs to be purchased. With the approach outlined here, we hope to find the right balance between the actual use of LBS data and methods for a better understanding of urban processes and the ethical use of that data and retention of a user’s privacy. As such, we release both our package and the dataset discussed here publicly in order to provide a foundation for further reproduction, comparison, and evaluation of different algorithms.

Methodology

To address the aforementioned issue, we present a software package written within the R ecosystem called ‘*homelocator*’, which provides a consistent framework and interface for the adoption of different approaches for identifying meaningful locations. Although the package can be used to infer different types of locations, we only focus on residential locations in this paper to give a particular illustration of the value of this approach. We have chosen to implement the package in the R language as it provides an open-source, comprehensive and flexible environment for geospatial data analysis, visualization and modelling with both

foundational spatial functionality (e.g. Lovelace *et al.* 2019, Bivand *et al.* 2020, Pebesma *et al.* 2020, etc) and domain-specific packages (cf. Joo *et al.*'s (2020) overview of 58 R packages for the analysis of animal movement). Although a similar landscape exists for Python (e.g. Rey and Anselin 2010), we have chosen R primarily for its tidyverse approach (Wickham *et al.* 2019), a consistent and newcomer-friendly approach to data analysis, which our package is designed to be compatible with.

Each approach to inferring locations can be written as a 'recipe' that opens the door for researchers to easily achieve an apples-to-apples comparison across algorithms and approaches. The functions of the package are designed to be flexible enough to allow researchers to create new variables or tune the thresholds used in existing recipes or algorithms. In doing so, they can infer meaningful locations within the specific context and requirements of their research (question).

The package and its source code can be accessed as a GitHub repository: <https://github.com/spatialnetworkslab/homelocator>. The package can be installed in R using the following command:

```
install_github("spatialnetworkslab/homelocator")
```

More detailed installation instructions as well a quick-start worked example are included as supplementary materials. In addition, documentation for the functions of the package and a tutorial on its use can be found on its documentation website: <https://homelocator-website.netlify.app/>.

The package has four built-in 'recipes'. The default 'recipe' is called '*HMLC*', which weighs data points across multiple time frames to 'score' potentially meaningful locations (discussed in detail in the section below). Two additional built-in recipes have been translated from approaches in the existing literature. The first, named '*OSNA*' (Efsthathiades *et al.* 2015), only considers data points sent on weekdays and divides a day into three time frames - 'rest

time’, ‘leisure time’ and ‘active time’. The algorithm finds the most ‘popular’ location during ‘rest’ and ‘leisure’ time as the home locations for users. The second, which we call ‘*APDM*’ (Ahas *et al.* 2010), calculates both the average and standard deviation timestamps in each location for each user. Finally, we include a simple or ‘naive’ algorithm, ‘*FREQ*’, which simply selects the most frequently ‘visited’ location.

Although existing algorithms for inferring meaningful locations based on spatio-temporal metadata differ in their specific approach, they all need three specific attributes for each data point: a unique identifier for the person or user, a unique identifier for the spatial location for the data point, and finally a timestamp that reflects the time the data point was created. In the ‘*HMLC*’ approach, we set a series of relatively stringent pre-conditions (cf. *Table 2*) to prevent inferring false positives or potentially ‘meaningless’ locations. It should be noted that these pre-conditions can be defined by researchers based on their specific research topic. We provide default values here as an illustration of the approach. Sporadic users do not post enough messages to make any claims about their meaningful locations, so we remove any user with less than 11 total tweets. Conversely, another type of these users can be spam or robot accounts run by algorithms. These generally have a large number of data points, much higher than human Twitter users. To avoid including this kind of user, we set the precondition to remove the top 1% most active users. Similarly, tourists or other transient visitors need to be excluded for many urban research questions. Data points sent by these users may only appear in a few days across several locations in the whole study period, so any users with data points sent during fewer than 11 distinct days, at fewer than 11 different places will be removed as well. For the remaining set of users, we also filter out locations that a user visited less than 11 times or were visited during fewer than 11 distinct hours of the day.

Preconditions	Min. Requirement
Potential fake or spam accounts	Top 1% of the most active users based on the frequency of data points
Data points sent by a user in total	> 10
Distinct locations a user visited in total	> 10
Data points sent by a user at a location	> 10
Distinct hours of the day a user was active at a location	> 10
Distinct days a user was active at a location	> 10
Number of days between the first and last activity at a location	> 10

Table 2. Default pre-conditions used for selecting ‘meaningful’ locations in the ‘*HMLC*’ recipe (Note: thresholds for each condition are tuneable based on specific research questions)

After this initial filtering phase, the algorithm uses a ‘scoring’ phase to determine the most likely home location. The timestamp of each data point can be used to determine different, meaningful time frames, such as ‘weekend’, ‘weekday’, ‘work time’ (i.e. between 9 AM and 6 PM on weekdays), ‘rest time’ (i.e. anytime that is not ‘work time’), and ‘weekend morning’ that can be used to further analyse the behaviour of each user and weight data points that are more likely to correspond to a residential location. *Table 3* summarizes these different variables and their weightage. Based on this, a final score for each location is created with the highest score being the most likely home location.

In this paper, we focus only on the most likely home location, so each user is only assigned a single home. However, it is important to note that some users may have multiple homes, for example students sleep at school dormitories on weekdays and back to their own/parents’ home on weekends – not an uncommon scenario in the Singapore context. If desired, a researcher can also extract these additional potential home locations with the package.

Activity	Expression	Weight	Score
Normalized tweets sent at a location	$\frac{T_p}{\max(T_p), p \in P_u}$	0.1	$0.1 \times \frac{T_p}{\max(T_p), p \in P_u}$
Normalized distinct active hours at a location	$\frac{N_{h,p}, \forall p \in P_u}{24}$	0.1	$0.1 \times \frac{N_{h,p}, \forall p \in P_u}{24}$
Normalized distinct days user visited at a location	$\frac{N_{d,p}}{\max(N_{d,p}), \forall p \in P_u}$	0.1	$0.1 \times \frac{N_{d,p}}{\max(N_{d,p}), \forall p \in P_u}$
Normalized distinct day of week user visited at a location	$\frac{N_{dow,p}}{7}, \forall p \in P_u$	0.1	$0.1 \times \frac{N_{dow,p}}{7}, \forall p \in P_u$
Normalized distinct months user visited at a location	$\frac{N_{m,p}}{12}, \forall p \in P_u$	0.1	$0.1 \times \frac{N_{m,p}}{12}, \forall p \in P_u$
Normalized No. of days between the 1 st & last activity at a location	$\frac{N_{activity,p}}{\max(N_{activity,p}), \forall p \in P_u}$	0.1	$0.1 \times \frac{N_{activity,p}}{\max(N_{activity,p}), \forall p \in P_u}$
Percentage of tweets sent on weekends	$\frac{T_{wk}}{T_{wk} + T_{wd}}$	0.1	$0.1 \times \frac{T_{wk}}{T_{wk} + T_{wd}}$
Percentage of tweets sent in weekends morning	$\frac{T_{wk,am}}{T_{wk,am} + T_{wk,pm}}$	0.1	$0.1 \times \frac{T_{wk,am}}{T_{wk,am} + T_{wk,pm}}$
Percentage of tweets sent during rest time	$\frac{T_{rt}}{T_{rt} + T_{wt}}$	0.2	$0.2 \times \frac{T_{rt}}{T_{rt} + T_{wt}}$
Symbol with explanation	$T_{wk} \& T_{wd}$	Tweets sent on weekends & tweets sent on weekdays	
	$T_{wk,am} \& T_{wk,pm}$	Tweets sent in weekends morning & tweets sent in weekends afternoon	
	$T_{rt} \& T_{wt}$	Tweets sent during rest time & tweets sent during work time	
	T_p	Number of tweets sent at a location	
	P_u	User visited locations	
	$N_{h,p}$	Distinct active hours at a location	
	$N_{d,p}$	Distinct days at a location	
	$N_{dow,p}$	Distinct day of week at a location	
	$N_{m,p}$	Distinct months at a location	
	$N_{activity,p}$	Number of days between the 1 st and last activity at a location	

Table 3. Default scoring standards used in the ‘HMLC’ recipe (Note: the weight for each activity is tuneable according to a researcher’s own importance)

Built-in ‘recipes’ can be used easily within the package by calling the one-stop function `identify_location` with a selected ‘recipe’ name (e.g. ‘APDM’, ‘FREQ’, ‘HMLC’, ‘OSNA’). The input data frame with one row per data point should include variables for ‘user’, ‘location’ and ‘timestamp’.

The package allows for the customization of ‘recipes’, or writing them completely from scratch, through the provision of a set of basic building blocks (see Table 4). Many of the existing algorithms in the literature can be expressed through these building blocks, implemented as various key functions.

(1) `validate_dataset`: validates the required variables are present in the input dataset. This

is a useful first step, especially in the context of reproducible workflows with different input datasets.

- (2) `enrich_timestamp`: derives additional variables from timestamp column (e.g. ‘weekday’, ‘hour of the day’) that are often used as intermediate variables in home location algorithms
- (3) A set of functions that make it easier to operate on ‘nested’ data. Nesting data (i.e. the storage of an entire table inside of a column) is a common procedure in the tidyverse approach to data manipulation. For location inference, it is often necessary to operate on doubly nested data as analysis is needed *per* person, *per* location. To aid the mental gymnastics required for this, we provide nested versions of common tidyverse functions such as ‘mutate’ and ‘summarise’, as well as functions specific to location inference such as `score_nested`, which allows the user to perform a scoring phase, after the initial filtering. Nesting has an additional advantage as it allows for the subsequent parallel processing of data, which can dramatically speed up computing time.
- (4) Finally, a function named `extract_location` extracts the most likely home location for each user based on the filtering and scoring in the previous steps.

```
# Use of algorithm with built-in recipes (e.g. HMLC, APDM, OSNA, FREQ)
home <- identify_location(df, ..., recipe = "HMLS") # df is an input table with one row per data point and variables
for 'user', 'location' and 'timestamp'

# Validate input dataset, ensuring necessary columns are present
validate_dataset(df, user, timestamp, location)

# Derive additional temporal variables (e.g. day of the week)
enrich_timestamp(df, timestamp) # derive temporal features from timestamp

# Nest columns (creates a nested tibble, allowing subsequent parallel processing)
nest_verbose(df, c(timestamp, location)) # nest timestamp and location columns

# Add additional column for each data point within nested tibble
```

```

mutate_nested(df_nest, wd_or_wk = if_else(wday %in% c(1, 7), "weekend", "weekday")) # create
weekday/weekend column

# Convert value to percentage format
prop_factor_nested(df_nest, var = wd_or_wk) # calculate percentage of weekday and weekend within nest

# Summarise based on column within nest
summarise_nested(df_nest, n_tweets = n())# calculate the number of tweets within nest
summarise_double_nested(df_nest, nest_cols = nest_cols_nm, n_tweets = n()) # calculate the number of tweets
within double nest

# Filter based on column within nest
filter_nested(n_points_per_location > 10) # only keep locations with more than 10 data points

# Score variable within nest
score_nested(df_nest, user = user, location = location,
             s_n_tweets_loc = 0.1*(n_tweets_loc/max(n_tweets_loc)) # assign weight and score variable of number
of tweets sent at a location

# Extract most likely meaningful location after previous filter and weighting steps
extract_location(df_scored, show_n_loc = 1, keep_score = FALSE, score) # extract only 1 meaningful location
based on score value for each user

```

Table 4. Basic building blocks of ‘homelocator’ package

Data and implementation

To illustrate and evaluate the package and proposed approach, we use a data repository of all geotagged tweets sent worldwide, derived from DOLLY project (Poorthuis and Zook 2017). We extract all geotagged tweets sent in Singapore from 2012 to 2016, which consists of around 22.4 million geotagged messages sent by approximately 405.6k users.

Following our earlier discussion of privacy and ethics, we de-identify and aggregate this dataset, while balancing with the ultimate aim of inferring meaningful locations. The resulting dataset can be downloaded from: <https://doi.org/10.6084/m9.figshare.13394102>. This procedure is summarized in *Figure 1*. Specifically, we only retain a (randomly generated) user ID, a location attribute, and a date/time attribute while any other variables such as content of

tweet, user description, etc are removed from the source data. We also remove users with too few data points, as well as those with so many data points that they are likely to be bots. Subsequently, for each user, a 5% portion of their tweets is randomly swapped with other users, effectively introducing a small amount of noise.

The next step consists of both temporal masking (timestamps are shifted by a random number of minutes - and the day of the week is swapped with a similar day (weekday or weekend)) and geomasking (locations are offset by a random distance < 100 meters). We also aggregate locations to 750m hexagonal grid cells. These aggregated 750m hexagonal grid cells are used as ‘locations’ in the remainder of the paper.

This aggregation resolution is used as a sensible resolution for the Singapore case, but it is important to emphasize that the size of the grid cell is an open choice for researchers depending on their specific research questions. For example, larger grid cells may be needed if the research is conducted in a less densely populated region. To analyze the potential effect of a change in resolution, we have performed a sensitivity analysis on different grid cell sizes from 250m to 1,500m with a 250m interval. A complete analysis is available in the supplementary materials together with the code and data (cf. Data Availability statement). Overall, the effect of resolution on the number of identified home location is minimal (< 2 percentage point difference between the smallest and largest resolution), and the consistency of the identified home locations is high as well. A switch to a different resolution changes the identified home location only in 3-5% of users.

Finally, locations visited by fewer than 5 people, or users with inferred homes at a location that have fewer than 5 inferred residential locations are also removed to prevent re-identification. This final aggregated and de-identified dataset consists of around 20.1 (89.5% of original dataset) million data points created by approximately 130.3k (32.13% of original dataset) users. To aid in subsequent apples-to-apples comparison across algorithms and

approaches, this aggregated and de-identified dataset is made publicly available in conjunction with the package.

Figure 1. Workflow of de-identification approach

It is important to note that the de-identification approach outlined here does result in a loss of both accuracy and precision for the purpose of inferring meaningful locations. However, we argue that this loss is relatively inconsequential for many human mobility analyses and is a worthwhile trade-off vis-a-vis the ethical use of this dataset and any potential harm that can result from its use. To evaluate this trade-off, we have applied each of the four algorithms on both original and de-identified datasets (cf. supplementary materials, see Data Availability statement). More than 80% of the users are labelled with the same home location in both datasets, and the median ‘error’ for mismatches is less than 2 kilometres for most algorithms (i.e. a neighbouring hexagon is selected instead).

We use the de-identified dataset to evaluate the three existing algorithms discussed above, in addition to our own algorithm. Computation is carried out on a MacBook Pro with

2.3 GHz Dual-Core Intel Core i5 and 16 GB Memory. To speed up computing and make processing large datasets possible, each algorithm is processed in parallel (2 cores). Of course computing time will differ based on the complexity of the algorithm and the size of the dataset. For reference, the ‘*HMLC*’ algorithm takes ~2.8 hour to process the 20.1 million records in the dataset, while it only takes ~1.9 hour for the ‘*FREQ*’ approach.

Results

We present the results of the application of the four algorithms on the same dataset (cf. *Table 5*). The code and data needed to reproduce this analysis and the figures in this section can be accessed at: <https://github.com/spatialnetworkslab/identifying-meaningful-locations>. To repeat, all four algorithms have the same purpose: to infer a user’s most likely home location. To allow for a spatial comparison, we calculate the odds ratio for the number of inferred homes in Singapore (cf. *Figure 2*). The odds ratio (OR) is helpful to visualize how significant the feature in a specific location is compared to the ‘expected’ pattern based on general use and it is often used in this context in social media analysis (Shelton *et al.* 2015). In this paper, we calculate the odds ratio as follows:

$$OR = \left(\frac{N_{users \text{ with inferred homes at a place}}}{N_{users \text{ with inferred home at all places}}} \right) / \left(\frac{N_{users \text{ without inferred homes at a place}}}{N_{users \text{ without inferred homes at all places}}} \right) \quad (1)$$

where N refers to the total numbers.

As *Figure 2* clearly shows, different algorithms may yield very different results. We also see clear differences in how many users an algorithm is able to infer a location for. For example, ‘*OSNA*’ identifies a home location for 89.1% of all users. However, the spatial distribution of those users deviates quite a bit from the other three approaches. One possible reason for this deviation is that the ‘*OSNA*’ algorithm is less ‘strict’ and includes users with relatively few data points that are filtered out in the other approaches. Our own approach - ‘*HMLC*’ is relatively close to the ‘*APMD*’ approach and identifies only ~25.7% of all users. These two algorithms show relatively similar spatial patterns. There are 28,007 (21.5%) users

that have been assigned a location by all four algorithms, which we refer to as ‘shared’ users. Even though the approaches are quite distinct, 21,863 (78.1% of 28,007) of these shared users get assigned the same home location in all four different approaches.

In this way, the use of the package also opens the door to the adoption of ensemble methods, especially in contexts where prevention of false positives is important. An ensemble method can apply different algorithms on the same dataset and only extract the inferred location for a user when the results of all algorithms match each other.

Approach	Results
APDM (Ahas <i>et al.</i> 2010)	Identified 40,374 (31.0%) users’ homes of 130,311 users
FREQ	Identified 47,263 (36.3%) users’ homes of 130,311 users
HMLC	Identified 33,488 (25.7%) users’ homes of 130,311 users
OSNA (Efsthathiades <i>et al.</i> 2015)	Identified 116,104 (89.1%) users’ homes of 130,311 users
Ensemble	Identified 21,863 (78.1%) users’ homes of 28,007 shared users

Table 5. Results of four individual algorithms and one ensemble approach for identifying home locations for users

Figure 2. Geographical distribution of inferred home locations for users by four different approaches

Since a ground-truth is not available due to the nature of the dataset, we also conduct a manual labelling procedure in order to further evaluate the performance of all methods discussed above (i.e. four implemented recipes and the ensembled method). In particular, we randomly select 1,000 users with identified home locations (i.e. 200 users for each method) and extract all tweets that were sent from their identified home location grid cell. This yields a total dataset of about 146,000 tweets sent by 1,000 users. We then manually read the content of all tweets for each user sent from their home location. If there are clear ‘home’ patterns in the content of the tweets (e.g. ‘reached homed’, ‘home sweet home’, ‘back home from ...’, ‘stay alone at home’, ‘just woke up’, ‘get out of the bed’, etc), we label that user/location as a true positive. However, even when reading the tweet content it can be difficult or even impossible to ascertain whether the location is really the home location. Thus, for all users where either the location is clearly not their home, or where there is simply no definitive positive indication, we labelled that user/location as a false positive. While this will likely also include a number of locations that are the user’s home after all, the true positive rate will give a clear insight in the performance of the algorithm nonetheless.

We evaluate the performance of each method by calculating its precision based on the hand-labelled ground truth – the true positive rate over 200 users. As *Table 6* clearly shows, the precision of a method is inversely proportional to the number of home locations the method is able to identify. More ‘stringent’ algorithms have a higher precision but this comes at a cost of being able to identify fewer locations. For example, the ensemble method – by definition – infers the fewest number of home locations of all the methods, but it has the highest precision (81.5%). On the other hand, the ‘OSNA’ recipe is able to infer 89.1% of users’ home locations, but its precision is only 38.5%.

Method	True Positive	Total	Precision (%)
Ensemble	163	200	81.50
HMLC	154	200	77.00
APDM	121	200	60.50
FREQ	112	200	56.00
OSNA	77	200	38.50

Table 6. Precision of identified home locations in different methods based on the hand-labelled ground truth dataset

To show why inferring meaningful locations can be an important step, especially when compared with using raw tweet density, we compare the geographic distributions of both tweets, users and inferred home locations in *Figure 3*. The density of unique ‘shared’ users visiting the central region is much higher than the actual number of inferred home locations. This is quite logical as the central region functions both as a CBD – the place of employment for most Singaporeans – and as a leisure and entertainment centre (including attractions such as Orchard Road and Marina Bay). As such, the spatio-temporal behaviour of people in these central areas is likely to be much less ‘home-like’. On the other hand, the ensemble approach seems to pick up quite well on the city’s densest residential areas (cf. *Figure 3d*). Singapore’s ‘heartlands’ are relatively dense neighbourhoods, centred around public housing estates developed by the country’s Housing Development Board. Unfortunately, population counts in Singapore are not available at a fine spatial scale⁴, but we can verify the output of the ensemble approach with the 2015 population per planning zone⁵. *Figure 4* shows a comparison between the normalized number of actual residents and the normalized number of inferred residents. Although there are some discrepancies (e.g. some central areas still have more inferred than actual residents)

⁴ Open datasets, such as [WorldPop](#), do exist on a finer spatial scale than planning zones, but the estimation process underlying these datasets is not reliable in the case of Singapore, especially in industrial areas where such datasets (erroneously) show a high number of estimated residents.

⁵ [Singapore Residents by Planning Area and Type of Dwelling, June 2015](#), a dataset published publicly on [Data.gov.sg](#)

overall the correlation between the two is strongly positive ($R = 0.96$, $p < 0.001$). This illustrates that a combination of algorithms can be quite a fruitful way to infer the underlying residential geography of a population and could thus provide a foundation for further urban analyses.

Nonetheless, it is important to repeat that despite the strong correlation between inferred home locations and the residential population, Twitter data does not constitute a representative sample of the population and should not be treated as such.

Figure 3. Geographic distributions of (top left to bottom right): (a) normalized number of tweets; (b) normalized number of unique users; (c) normalized number of inferred home locations; (d) normalized number of residents (2015). Normalization is calculated based on min-max normalization.

Figure 4. The correlation between the normalized number of inferred residents and the normalized number of actual residents in Singapore in 2015. Normalization is calculated based on total-sum normalization.

Conclusion

In this article we have implemented four different approaches for determining home locations from the literature and applied them to a de-identified and aggregated LBS dataset in Singapore. We do so through an R package – named ‘*homelocator*’ – that has been developed for this purpose. As current approaches in the literature often rely on custom algorithms – making evaluation and comparison more challenging – we hope this package makes comparison across different approaches and algorithms easier, more accessible, and more customizable in the context of inferring meaningful location from LBS and similar datasets. In doing so, the package also allows for the combination of multiple algorithms in an ensemble approach and

makes implementation of new algorithms more convenient through the provision of a set of common functions.

Although different algorithms may yield significantly different results, if we focus only on users for which all four algorithms in our case study agree on the most likely home location, the resulting spatial patterns closely correlate with the actual population distribution. In addition, based on a manually-constructed ground-truth, the precision of this approach is conservatively estimated at 81.5%. Accordingly, the outlined ensemble approach for identifying home locations is promising (i.e. use of multiple algorithms on the same dataset and only assign home location for users if a quorum of algorithms agrees), especially in cases where the prevention of false positives is important. Although we only focus on residential locations in this paper, other meaningful locations, such as places of work and leisure, could also be inferred with a similar approach. Algorithms for identifying these types of meaningful locations could be developed for the package in the future.

We are cognizant that the use of LBS data, and inferring meaningful locations in particular, brings forward specific privacy considerations and obliges us to seriously consider research ethics and potential harm. For this reason, we de-identify the LBS data used through spatial and temporal perturbation, random noise as well as spatial aggregation. Although this undoubtedly affects the precision and accuracy of any location inference, we argue that this is a fine trade-off for most analytical applications at the urban scale. Considering that some researchers may not have access to such a dataset, we have made the de-identified dataset available alongside the package to aid in further methodological development.

The work presented here can certainly be expanded. Although other datasets, such as mobile phone data, should in theory be compatible with the approach here, we have only tested the package with LBS data. Furthermore, additional algorithm development could focus on using a set of data points that is as small as possible (Stanley *et al.* 2018). This could potentially

alleviate the ‘need’ for massive data trawling operations, although this too does come with potential ethical implications. Foremost, we hope that through packages and libraries like the one presented here, future work that relies on the inference of meaningful locations becomes less ‘custom’ (with each researcher writing their own algorithm implementation) and more transparent. Using common and comparable algorithms, sharing code and, where possible, data will increase transparency and reproducibility in the field and will aid in furthering the ongoing discourse around ethical big data research.

Data and code availability statement

Along with this paper, we publish three specific data & code artifacts:

1. The homelocator R package itself, with code hosted on: <https://github.com/spatialnetworkslab/homelocator>. The package contains a small test sample dataset so readers can quickly try out the package. A documentation site, including a tutorial and installation instructions, is available at: <https://homelocator-website.netlify.app/>.
2. The deidentified LBS dataset, which can be downloaded from Figshare: <https://doi.org/10.6084/m9.figshare.13394102>.
3. All code & data to reproduce the analysis and figures within this paper. This is available as a series of RMarkdown files hosted on: <https://github.com/spatialnetworkslab/identifying-meaningful-locations>. It includes a Binder link so readers can explore the analysis directly from their browser.

Acknowledgment

This research, led together with the Housing and Development Board, is supported by the Singapore Ministry of National Development and the National Research Foundation, Prime Minister’s Office under the Land and Liveability National Innovation Challenge (L2 NIC) Research Programme (L2 NIC Award No. L2NICTDF1-2017-4). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Housing and

Development Board, Singapore Ministry of National Development and National Research Foundation, Prime Minister's Office, Singapore.

References

- Ahas, R., Aasa, A., Yuan, Y., Raubal, M., Smoreda, Z., Liu, Y., Ziemlicki, C., Tiru, M., and Zook, M., 2015. Everyday space–time geographies: using mobile phone-based sensor data to monitor urban activity in Harbin, Paris, and Tallinn. *International Journal of Geographical Information Science*, 29 (11), 2017–2039.
- Ahas, R., Silm, S., Järv, O., Saluveer, E., and Tiru, M., 2010. Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. *Journal of Urban Technology*, 17 (1), 3–27.
- Backstrom, L., Sun, E., and Marlow, C., 2010. Find me if you can: improving geographical prediction with social and spatial proximity. In: *Proceedings of the 19th international conference on World wide web - WWW '10*. Presented at the the 19th international conference, Raleigh, North Carolina, USA: ACM Press, 61.
- Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J.J., and Vespignani, A., 2009. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106 (51).
- Belyi, A., Bojic, I., Sobolevsky, S., Sitko, I., Hawelka, B., Rudikova, L., Kurbatski, A., and Ratti, C., 2017. Global multi-layer network of human mobility. *International Journal of Geographical Information Science*, 31 (7), 1381–1402.
- Bivand, R., Altman, M., Anselin, L., Assunção, R., Berke, O., Bernat, A., Blanchet, G., Blankmeyer, E., Carvalho, M., Christensen, B., Chun, Y., Dormann, C., Dray, S., Gómez-Rubio, V., Gubri, M., Halbersma, R., Krainski, E., Legendre, P., Lewin-Koh, N., Li, A., Li, H., Ma, J., Mallik, A., Millo, G., Mueller, W., Ono, H., Peres-Neto, P., Piras, G., Reder, M., Tiefelsdorf, M., Westerholt, R., and Yu, D., 2020. *spdep: Spatial Dependence: Weighting Schemes, Statistics*.
- Burt, W.H., 1943. Territoriality and Home Range Concepts as Applied to Mammals. *Journal of Mammalogy*, 24 (3), 346.
- Calabrese, J.M., Fleming, C.H., and Gurarie, E., 2016. ctmm: an r package for analyzing animal relocation data as a continuous-time stochastic process. *Methods in Ecology and Evolution*, 7 (9), 1124–1132.
- Calenge, C. and Fortmann-Roe, contributions from S., 2020. *adehabitatHR: Home Range Estimation*.

- Chandra, S., Khan, L., and Muhaya, F.B., 2011. Estimating Twitter User Location Using Social Interactions--A Content Based Approach. *In: 2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing*. Presented at the 2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust (PASSAT) / 2011 IEEE Third Int'l Conference on Social Computing (SocialCom), Boston, MA, USA: IEEE, 838–843.
- Chang, H., Lee, D., Eltaher, M., and Lee, J., 2012. @Phillies Tweeting from Philly? Predicting Twitter User Locations with Spatial Word Usage. *In: 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. Presented at the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 111–118.
- Cheng, Z., Caverlee, J., and Lee, K., 2010. You are where you tweet: a content-based approach to geo-locating twitter users. *In: Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10*. Presented at the the 19th ACM international conference, Toronto, ON, Canada: ACM Press, 759.
- Compton, R., Jurgens, D., and Allen, D., 2014. Geotagging One Hundred Million Twitter Accounts with Total Variation Minimization. *2014 IEEE International Conference on Big Data (Big Data)*, 393–401.
- Crampton, J., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M., and Zook, M., 2013. *Beyond the Geotag: Situating 'Big Data' and Leveraging the Potential of the Geoweb*. Rochester, NY: Social Science Research Network, SSRN Scholarly Paper No. ID 2253918.
- Curtis, A.J., Mills, J.W., and Leitner, M., 2006. Spatial confidentiality and GIS: re-engineering mortality locations from published maps about Hurricane Katrina. *International Journal of Health Geographics*, 5 (1), 44.
- Efstathiades, H., Antoniadis, D., Pallis, G., and Dikaiakos, M.D., 2015. Identification of key locations based on online social network activity. *In: 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. Presented at the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 218–225.
- Eisenstein, J., O'Connor, B., Smith, N.A., and Xing, E.P., 2010. A latent variable model for geographic lexical variation. *In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. USA: Association for Computational Linguistics, 1277–1287.

- Elwood, S. and Leszczynski, A., 2011. Privacy, reconsidered: New representations, data practices, and the geoweb. *Geoforum*, 42 (1), 6–15.
- Fiore, M., Katsikouli, P., Zavou, E., Cunche, M., Fessant, F., Hello, D.L., Aivodji, U.M., Olivier, B., Quertier, T., and Stanica, R., 2020. Privacy in trajectory micro-data publishing : a survey. *arXiv:1903.12211 [cs]*.
- Fung, B.C.M., Wang, K., Chen, R., and Yu, P.S., 2010. Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*, 42 (4), 14:1-14:53.
- Gao, S., Rao, J., Liu, X., Kang, Y., Huang, Q., and App, J., 2019. Exploring the effectiveness of geomasking techniques for protecting the geoprivacy of Twitter users. *Journal of Spatial Information Science*, 2019 (19), 105–129.
- Hecht, B., Hong, L., Suh, B., and Chi, E.H., 2011. Tweets from Justin Bieber’s heart: the dynamics of the location field in user profiles. *In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 237–246.
- Hincks, S., Kingston, R., Webb, B., and Wong, C., 2018. A new geodemographic classification of commuting flows for England and Wales. *International Journal of Geographical Information Science*, 32 (4), 663–684.
- Huang, Q. and Wong, D.W.S., 2016. Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us? *International Journal of Geographical Information Science*, 30 (9), 1873–1898.
- Järv, O., Tenkanen, H., and Toivonen, T., 2017. Enhancing spatial accuracy of mobile phone data using multi-temporal dasymetric interpolation. *International Journal of Geographical Information Science*, 31 (8), 1630–1651.
- Joo, R., Boone, M.E., Clay, T.A., Patrick, S.C., Clusella-Trullas, S., and Basille, M., 2020. Navigating through the r packages for movement. *Journal of Animal Ecology*, 89 (1), 248–267.
- Jr, C.A.D., Pappa, G.L., Oliveira, D.R.R. de, and Arcanjo, F. de L., 2011. Inferring the Location of Twitter Messages Based on User Relationships. *Transactions in GIS*, 15 (6), 735–751.
- Jurgens, D., 2013. That’s What Friends Are For: Inferring Location in Online Social Media Platforms Based on Social Relationships. *In: ICWSM*.
- Keßler, C. and McKenzie, G., 2018. A geoprivacy manifesto. *Transactions in GIS*, 22 (1), 3–19.
- Kinsella, S., Murdock, V., and O’Hare, N., 2011. ‘I’m eating a sandwich in Glasgow’: modeling

- locations with tweets. In: *Proceedings of the 3rd international workshop on Search and mining user-generated contents*. New York, NY, USA: Association for Computing Machinery, 61–68.
- Kounadi, O. and Leitner, M., 2014. Why Does Geoprivacy Matter? The Scientific Publication of Confidential Data Presented on Maps. *Journal of Empirical Research on Human Research Ethics*, 9 (4), 34–45.
- Kwan, M., 2004. Gis methods in time-geographic research: geocomputation and geovisualization of human activity patterns. *Geografiska Annaler: Series B, Human Geography*, 86 (4), 267–280.
- Kwan, M.-P., Casas, I., and Schmitz, B., 2006. Protection of Geoprivacy and Accuracy of Spatial Information: How Effective Are Geographical Masks? *Cartographica: The International Journal for Geographic Information and Geovisualization*.
- Li, R., Wang, S., and Chang, K.C.-C., 2012. Multiple location profiling for users and relationships from social network and content. *Proceedings of the VLDB Endowment*, 5 (11), 1603–1614.
- Lin, J. and Cromley, R.G., 2018. Inferring the home locations of Twitter users based on the spatiotemporal clustering of Twitter data. *Transactions in GIS*, 22 (1), 82–97.
- Liu, X., Huang, Q., and Gao, S., 2019. Exploring the uncertainty of activity zone detection using digital footprints with multi-scaled DBSCAN. *International Journal of Geographical Information Science*, 33 (6), 1196–1223.
- Longley, P.A. and Adnan, M., 2016. Geo-temporal Twitter demographics. *International Journal of Geographical Information Science*, 30 (2), 369–389.
- Lovelace, R., Birkin, M., Cross, P., and Clarke, M., 2016. From Big Noise to Big Data: Toward the Verification of Large Data sets for Understanding Regional Retail Flows. *Geographical Analysis*, 48 (1), 59–81.
- Lovelace, R., Nowosad, J., and Muenchow, J., 2019. *Geocomputation with R*. CRC Press.
- Lyons, A.J., Turner, W.C., and Getz, W.M., 2013. Home range plus: a space-time characterization of movement over real landscapes. *Movement Ecology*, 1 (1), 2.
- Maat, K., Wee, B. van, and Stead, D., 2016. Land Use and Travel Behaviour: Expected Effects from the Perspective of Utility Theory and Activity-Based Theories: *Environment and Planning B: Planning and Design*.
- Mahmud, J., Nichols, J., and Drews, C., 2014. Home Location Identification of Twitter Users. *arXiv:1403.2345 [cs]*.
- McGee, J., Caverlee, J., and Cheng, Z., 2013. Location prediction in social media based on tie

- strength. In: *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management - CIKM '13*. Presented at the the 22nd ACM international conference, San Francisco, California, USA: ACM Press, 459–468.
- de Montjoye, Y.-A., Hidalgo, C.A., Verleysen, M., and Blondel, V.D., 2013. Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports*, 3 (1), 1376.
- Norval, C. and Henderson, T., 2017. Contextual Consent: Ethical Mining of Social Media for Health Research. *arXiv:1701.07765 [cs]*.
- Olson, K.L., Grannis, S.J., and Mandl, K.D., 2006. Privacy Protection Versus Cluster Detection in Spatial Epidemiology. *American Journal of Public Health*, 96 (11), 2002–2008.
- Pebesma, E., Bivand, R., Racine, E., Sumner, M., Cook, I., Keitt, T., Lovelace, R., Wickham, H., Ooms, J., Müller, K., Pedersen, T.L., and Baston, D., 2020. *sf: Simple Features for R*.
- Poorthuis, A. and Zook, M., 2017. Making Big Data Small: Strategies to Expand Urban and Geographical Research Using Social Media. *Journal of Urban Technology*, 24 (4), 115–135.
- Prabhu, V.U. and Birhane, A., 2020. Large image datasets: A pyrrhic win for computer vision? *arXiv:2006.16923 [cs, stat]*.
- Rey, S.J. and Anselin, L., 2010. PySAL: A Python Library of Spatial Analytical Methods. In: M.M. Fischer and A. Getis, eds. *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*. Berlin, Heidelberg: Springer, 175–193.
- Rout, D., Bontcheva, K., Preotăciuc-Pietro, D., and Cohn, T., 2013. Where's @wally?: a classification approach to geolocating users based on their social ties. In: *Proceedings of the 24th ACM Conference on Hypertext and Social Media - HT '13*. Presented at the the 24th ACM Conference, Paris, France: ACM Press, 11–20.
- Sadilek, A., Kautz, H., and Bigham, J.P., 2012. Finding your friends and following them to where you are. In: *Proceedings of the fifth ACM international conference on Web search and data mining - WSDM '12*. Presented at the the fifth ACM international conference, Seattle, Washington, USA: ACM Press, 723.
- Sadowski, J., 2019. When data is capital: Datafication, accumulation, and extraction. *Big Data & Society*, 6 (1), 2053951718820549.
- Schulz, A., Hadjakos, A., Paulheim, H., Nachtwey, J., and Mühlhäuser, M., 2013. A Multi-Indicator Approach for Geolocalization of Tweets. In: *ICWSM*.
- Signer, J. and Balkenhol, N., 2015. Reproducible home ranges (rhr): A new, user-friendly R package for analyses of wildlife telemetry data. *Wildlife Society Bulletin*, 39 (2), 358–

- Signer, J., Fieberg, J., and Avgar, T., 2018. Animal Movement Tools (amt): R-Package for Managing Tracking Data and Conducting Habitat Selection Analyses. *arXiv:1805.03227 [q-bio]*.
- Sila-Nowicka, K., Vandrol, J., Oshan, T., Long, J.A., Demšar, U., and Fotheringham, A.S., 2016. Analysis of human mobility patterns from GPS trajectories and contextual information. *International Journal of Geographical Information Science*, 30 (5), 881–906.
- Stanley, K., Yoo, E.-H., Paul, T., and Bell, S., 2018. How many days are enough?: capturing routine human mobility. *International Journal of Geographical Information Science*, 32 (7), 1485–1504.
- Swanlund, D., Schuurman, N., and Brussoni, M., 2020. MaskMy.XYZ: An easy-to-use tool for protecting geoprivacy using geographic masks. *Transactions in GIS*, 24 (2), 390–401.
- Taylor, L., 2016. No place to hide? The ethics and analytics of tracking mobility using mobile phone data. *Environment and Planning D: Society and Space*, 34 (2), 319–336.
- Thompson and Warzel, 2019. Opinion | Twelve Million Phones, One Dataset, Zero Privacy - The New York Times [online]. Available from: <https://www.nytimes.com/interactive/2019/12/19/opinion/location-tracking-cell-phone.html> [Accessed 23 Oct 2020].
- Tracey, J., Sheppard, J., Zhu, J., Sinkovits, R., Chourasia, A., Lock-Wood, G., and Fisher, R., 2014. *R package 'mkde' manual (2D and 3D movement-based kernel density estimates)*.
- Tu, W., Cao, J., Yue, Y., Shaw, S.-L., Zhou, M., Wang, Z., Chang, X., Xu, Y., and Li, Q., 2017. Coupling mobile phone and social media data: a new approach to understanding urban functions and diurnal patterns. *International Journal of Geographical Information Science*, 31 (12), 2331–2358.
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemond, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H., 2019. Welcome to the Tidyverse. *Journal of Open Source Software*, 4 (43), 1686.
- Xin, Y. and MacEachren, A.M., 2020. Characterizing traveling fans: a workflow for event-oriented travel pattern analysis using Twitter data. *International Journal of Geographical Information Science*, 0 (0), 1–20.
- Yuan, Y. and Raubal, M., 2016. Analyzing the distribution of human activity space from mobile

- phone usage: an individual and urban-oriented study. *International Journal of Geographical Information Science*, 30 (8), 1594–1621.
- Zimmer, M., 2018. Addressing Conceptual Gaps in Big Data Research Ethics: An Application of Contextual Integrity. *Social Media + Society*, 4 (2), 2056305118768300.
- Zook, M., Barocas, S., Boyd, D., Crawford, K., Keller, E., Gangadharan, S.P., Goodman, A., Hollander, R., Koenig, B.A., Metcalf, J., Narayanan, A., Nelson, A., and Pasquale, F., 2017. Ten simple rules for responsible big data research. *PLOS Computational Biology*, 13 (3), e1005399.