Geography Compass 10/6 (2016): 239–252, 10.1111/gec3.12269

Exploring Georeferenced Mobile Phone Datasets – A Survey and Reference Framework

Yihong Yuan[1]* and Martin Raubal[2]

[1]*Department of Geography, Texas State University*

[2]*Institute of Cartography and Geoinformation, ETH Zurich*

Abstract

Nowadays, mobile phones and other information and communication technology (ICT) devices collect large numbers of measurements about their users. This review paper provides an overview of georeferenced mobile phone datasets by exploring and summarizing the metadata of multiple datasets based on a literature review. It also presents an abstract model to depict the connections of these datasets, serving as the basis for potential spatio-temporal data mining and geographic knowledge discovery tasks. The summarized metadata table and proposed research topics can be applied as a reference framework to future studies in the area of mobile computing, geospatial data mining, and mobile data collection in the age of instant access.

1. *Introduction*

Nowadays, mobile phones and other ICT devices collect large numbers of measurements about their users, including intra-personal behaviors (e.g., individual call frequencies) and inter-personal level behaviors (e.g., social networks) (Gonzalez, et al. 2008, Ahas, et al. 2010, Phithakkitnukoon, et al. 2010, Farrahi, et al. 2012, Sagl, et al. 2014), which provide a wide range of spatio-temporal data sources to be used for geographic knowledge discovery and spatial data mining, such as human travel behavior and mobility patterns (Batsanov, et al. 2001, Miller 2009, Kang, et al. 2010, Song, et al. 2010, Becker, et al. 2011, Silm and Ahas 2014).

Mobile phones are capable of recording location information by several means (Adusei, et al. 2002, Wang, et al. 2002, Brimicombe and Li 2009), such as using the global positioning system (GPS), using the service-provider assisted GPS (A-GPS) or simply logging the connected cellular tower. Although the completeness varies for different datasets, a typical mobile phone dataset contains three categories of user information (Yuan and Raubal 2010): (i) spatio-temporal tracking information (i.e., event-based information recorded when a certain type of event occurs or real-time locations of stand-by phones collected by service providers); (ii) service usage information (i.e., the frequency and duration of voice, text, and other types of service usage); and (iii) demographic profiles, if available (including individual-level profiles, such as age or gender; and super-individual profiles, such as social aspects or cultural backgrounds (Reades, et al. 2007, Dashdorj, et al. 2013, Liu et al., 2013a, 2013b).

However, because the data precision, accuracy, and richness vary for different datasets, there have not been sufficient studies on how much information gets recorded in georeferenced mobile phone data, and to what extent researchers have utilized these data in the fields of Geographic Information Science (GIScience) and various application scenarios. Oxendine and Waters (2014) reviewed the usage of crowdsourced mobile data in minimizing risk, but this literature survey was primarily conducted in the field of urban evacuation and risk analysis.

Calabrese *et al.* (2015) provided a systematic review of the studies on urban sensing using mobile phone data. However, our paper is conducted from a different perspective: (i) The focus is put on the datasets instead of the technologies: We investigate the categories of information available in georeferenced mobile phone datasets based on a literature review on the Web of Knowledge,[1] the Association for Computing Machinery (ACM) digital library,[2] and top GIScience journals (*Transactions in GIS*, *Geography Compass*, and the *International Journal of Geographic Information Science*) (c.f. journal rankings from Caron et al., 2008 and Google Scholar search[3]). In practice, researchers rarely have access to the raw data; hence, questions such as "which data fields are more relevant and why?", or "which groups should I contact if I am interested in a certain type of mobile phone data?", are crucial for new researchers in this area. (ii) We focus on both "individual-oriented" and "aggregate-oriented" research in modeling human mobility from mobile phone usage (c.f. Section 4).

Although ICTs have impacted modern society in many other perspectives besides geography, the scope of this review focuses on the GIScience field, and only a selected set of data attributes directly related to location-based analytics will be examined. Section 2 provides a brief background of mobile phone tracking, and Section 3 summarizes existing datasets as well as their shared characteristics. We also discuss the potential research topics in Section 4. This paper will be of use to new GIScience researchers to conduct an initial review of georeferenced mobile phone datasets. The summarized metadata table and proposed topics can also be applied as a reference to future studies in the area of mobile computing, spatio-temporal data mining, and mobile data collection.

2. Mobile Phone Tracking

Mobile phone tracking refers to the process of acquiring the location of target phones via various techniques (Kasemsan and Ratsameethammawong 2010), which can be divided into Lagrangian-based and Eulerian-based measurements (Hong and Lee 2011). The former refers to the process of following a device as it moves (e.g., use in-device GPS trackers). The latter is to have an observation site geographically fixed, such as monitoring traffic load in cell towers to estimate population distribution. The generation of datasets can be further stratified by utilized positioning method, including but not limited to the following.

1 GSM-based (originally Groupe Spécial Mobile)

This method is widely available due to the wide spread and low cost of GSM facilities (Ahas 2005, Ratti, et al. 2005, Bayir, et al. 2010, Calabrese, et al. 2013). The location of the handset is determined by the identity (ID) of the connected base tower; therefore, both accuracy and precision of the data are highly related to the density of cellular infrastructures (usually 300–500 m in urban areas) (Yuan, et al. 2012). The information collected includes call logs, cell tower IDs, application usage, and phone status (charging, idle, etc.). In practice, more advanced technologies have been applied to improve the location accuracy based on triangulation, such as "timing advance" (TA), "received signal strength" (RSS) (Calabrese, et al. 2015), and "cell tower triangulation", which measures the angles between three or more nearby sensors to pinpoint the location more precisely.

The raw usage data of mobile tracking technologies include various kinds of usage detail records (UDRs). For GSM-based services, the details of a phone call are usually documented as a CDR (Table 1) produced by a telephone exchange. A typical CDR includes the following items (Horak 2007):

1. the phone number of the subscriber originating the call (calling party),
2. the phone number receiving the call (called party),

Table 1. Example of simplified CDR.

Source	Destination	Cell ID	Time answered	Duration
1350*******	1360*******	111	14:36:24	12 min

3. the starting time of the call (date and time),
4. the call duration,
5. cell identification,
6. a sequence number identifying the record,
7. the disposition or the results of the call (indicating, for example, whether the call failed),
8. call type (voice, SMS, etc.),
9. any fault condition encountered, and
10. other related fields (e.g., the route by which the call entered/left the exchange).

However, the manufacturers and service providers can decide which information is emitted and how it is formatted; therefore, researchers may not get complete information from CDRs. Note that the usage record of SMS and cellular data plans can be viewed as specific types of CDRs.

2 Wi-Fi-based

Many mobile phones today are equipped with a Wi-Fi or Internet option. Similar to GSM-based tracking, this technique tracks the network cell that a mobile device registers to. The availability of an Internet Protocol (IP) network is highly based on Wi-Fi signal types and signal coverage, and the range is around 75–150 ft (around 20–45 m) (Kasemsan and Ratsameethammawong 2010).

Generally, the usage of all Wi-Fi-based services can be retrieved from an Internet protocol detail record (IPDR). Similar to CDRs, each IPDR contains the "5 Ws": who, when, what, where, and why:

1. who: the identification of network user(s), service provider, etc.,
2. when: the time and duration of connection,
3. what: the content of the transferred data packet,
4. where: IP address, location of connected users,
5. why: the purpose of connection (e.g., web surfing and playing online games).

Additional service-specific attributes can be added to the usage records (e.g., quality of service, bandwidth, and latency). IPDRs can be encoded using two available formats: extensible markup language (XML) and external data representation (XDR). The IP address can be used to locate specific users when available. Navizon, a Miami-based mobile technology firm, provides a new cell phone tracking service allowing its users to track their phones based on the transmitted and received IP signals.[4] However, the majority of IPDR from mobile phone users are not collected by cellular phone companies; instead, they are collected by Internet service providers (via IPDR collectors and transmitters).

3 Bluetooth-based

Bluetooth is a technology designed to connect devices and exchange data over short distances based on short-wavelength radio waves. This technology only functions within a limited range (e.g., 10–100 m); hence, it can be particularly useful for studying the social and spatial interactions between individuals (Jenkins 2013). Bluetooth logs typically do not store longitude/latitude directly; however, the location information can be further extracted based on the relative distances between a certain device and pre-located Bluetooth receivers (Murakami, et al. 2007, Versichele, et al. 2012, Li, et al. 2014). The log files normally include the identification of connected devices, media access control (MAC) addresses, connection time, device type, event type (i.e., device arrival or device left), etc.

4 GPS-based

Nowadays, assisted GPS (A-GPS) or differential GPS (D-GPS) modules are usually incorporated into smartphones to improve the tracking accuracy. The difference between A-GPS and a regular GPS handset is that A-GPS communicates with servers in base towers, and the locations are calculated by servers instead of handsets; therefore, the GPS-based tracking of mobile phones can be viewed as a hybrid method that incorporates both GSM and GPS techniques (Djuknic and Richton 2001, Alanen, et al. 2006).

The official format for standard GPS data is NMEA-0183, which is a publicly used standard defined by the National Marine Electronics Association (NMEA) to transmit GPS positions. Based on this standard, all data are transmitted in three basic forms of sentences: talker sentences, proprietary sentences, and query sentences. Each sentence starts with a "$" character and ends with <CR><LF>.[5]

In the mobile phone service market, A-GPS uses an assistance server (mobile location server) via cellular networks. The server either knows or learns the position of the closest cell phone tower and uses A-GPS to calculate the user's exact position. Therefore, the usage logs of these services are usually included as part of the CDRs. In addition, the development of GPS modules and smartphones also enables more flexibility for the researchers to collect customized data, such as developing data collection applications (apps) to acquire data directly from users.

Other tracking methods may be available, such as the inertial navigation system in smartphones (Liu et al., 2013a, 2013b, Quarmyne and Pachter 2014), which uses a motion sensor to continuously monitor the position of a device without external references.

3. A Summary of Existing Mobile Phone Datasets

Although the accuracy, precision, resolution, and completeness vary for different datasets, the raw data of mobile phone usage generally provide us with three types of information to profile the user behavior.

1 Spatio-temporal tracking information

The accuracy and precision of recorded trajectories depend on the location technique, and the collected data are typically generated as sample points, which require further processing to interpret user activity patterns. For instance, a mobile switching center (MSC) can detect that a mobile device is approaching the edge of its cell regardless of whether the user initiated a voice call (Alguliyev, et al. 2010).

2 Service usage information

One of mobile phones' main functions is to enable two-way radio telecommunication over a cellular network of base stations (Agar 2004). Hence, service usage is another aspect of information recorded in mobile phone datasets, which also includes pieces of social interaction information of mobile phone users.

4 User profiles

The demographic profiles of phone users (i.e., age, gender, occupation, and address) are also documented by the service provider when the users register for services. Due to privacy and policy issues, this information is stored in separate databases, and the levels of detail vary for different regions. However, the aggregated-level social-economic profiles (e.g., social conditions) of the service area may also be obtained.

The remainder of this section reviews the content and structure of existing mobile phone datasets, and Section 4 constructs a framework to categorize the information stored in mobile phone usage data. The datasets are selected based on a literature review from the web of knowledge and the ACM digital library. The former is a comprehensive search engine which covers many high-ranking Geography journals, including *Annals of the Association of American Geographers*, *Computers Environment and Urban Systems*, etc., whereas the latter (ACM digital library) focuses on the computer science community. When searching for topic "mobile phone data" in these two databases in the research areas of geography, sociology, transportation, and urban studies, over 100 records were returned. To better incorporate the geography community, we also conducted a literature search in three top GIScience and geography journals (*Transactions in GIS*, *Geography Compass*, and the *International Journal of Geographic Information Science*), using keywords "mobile phone data" and publication date in 2015. We further eliminated the papers that are not related to georeferenced mobile phone data or did not provide the metadata (~50 papers). Several papers utilized the subset of the same datasets; therefore, we also eliminated duplicates (~20 papers). After this filtering process, we obtained the metadata of 47 georeferenced mobile phone datasets and summarized the information focusing on the following six properties (corresponding to the three types of information discussed earlier): spatio-temporal tracking information (*temporal duration*, *area covered*, *location technique*, *and location accuracy*), service usage information (*service usage type*), and user profiles (*sample size*, *participant type*, *and included user attributes*).

1 Sample size: the number of users/records covered by a particular dataset.
2 Temporal duration: the time span during which the data were collected.
3 Area covered: the geographic area in which the data were created. In several studies, the geographic area is not specified, because data are acquired for specified users regardless of the actual locations to which they have traveled (user-oriented).
4 Location technique and location accuracy: As stated in Section 2, there are four major types of location techniques: GSM-based, Wi-Fi-based, GPS-based, and Bluetooth-based.
5 Service usage type: voice call, SMS, Internet surfing, etc.
6 User (participant type) and additional user attributes: The background information of participants is not always available in the datasets (Wesolowski, et al. 2013). Researchers are able to initially collect personal information through human participant experiments, but the availability of personal information in UDRs is often limited.

Table A-1 (see Appendix) lists the basic information of these datasets. Based on this table, we can observe the following:

1 Generally, there are two distinct sources of data.

a. Data collected by human participant experiments [actively collected, e.g., by distributing experimental devices (i.e., smartphones)]. The sample size is relatively low due to limited devices available and the cost of recruiting human participants. A typical example for this category is the reality mining dataset,[6] which has 100 participants.

b. Existing usage record data (passively collected from mobile service providers): Here, the sample size is only restricted by the access privilege of corresponding researchers and the scope of research topics; therefore, the sample size can easily reach a magnitude of "millions".

2 The locations of the datasets are mapped in Figure 1. The majority of studies focuses on Asia, Europe (Estonia in particular), and North America (United States in particular), including both developing and developed countries. Among all countries, China, Estonia, and the US are the most active in terms of the number of publications and research groups. It is also worth noting the difference between obtaining and utilizing georeferenced mobile phone data in China and the US. Due to limited data infrastructure in both private and public sectors, most Chinese datasets were obtained due to a collaboration between researchers and the direct service providers (China mobile and China telecom in particular, which occupy over 90% of the telecommunication market in China). However, in the US, there are more opportunities for collaborative research with third-party companies such as Airsage, Inc,[7] which specializes in providing movement data to the transportation industry. An interesting topic for future study is to investigate how the social, economic, and policy status affects the development of ICT studies.

3 Due to the high cost of human participant experiments, most datasets are acquired directly from CDRs with relatively low spatial accuracy and precision (Ahas, et al. 2008, Ahas, et al. 2010). Actually, all positioning techniques have data-quality issues: GPS signals can be impacted by the number of satellites connected, the weather condition, the landscape, or spatial setting of a certain location. For Wi-Fi-based and Bluetooth-based tracking, different devices may have varying coverage range and locational accuracy, or they are even not traceable at all if a virtual private network (VPN) is utilized. Geocoding an IP address or Bluetooth record also brings in data uncertainty. For instance, the company MaxMind provides various levels of accuracy for locating IP addresses at different prices.[8]

4 Access to personal information of phone users in CDRs is usually restricted to researchers due to privacy issues; however, for human participant experiments, user profiles are commonly available per the agreement between researchers and participants (Farrahi and Gatica-Perez 2010). A primary concern with big data research is whether people have given consent for the use of their data (e.g., their phone usage patterns). For instance, GPS signals collected from

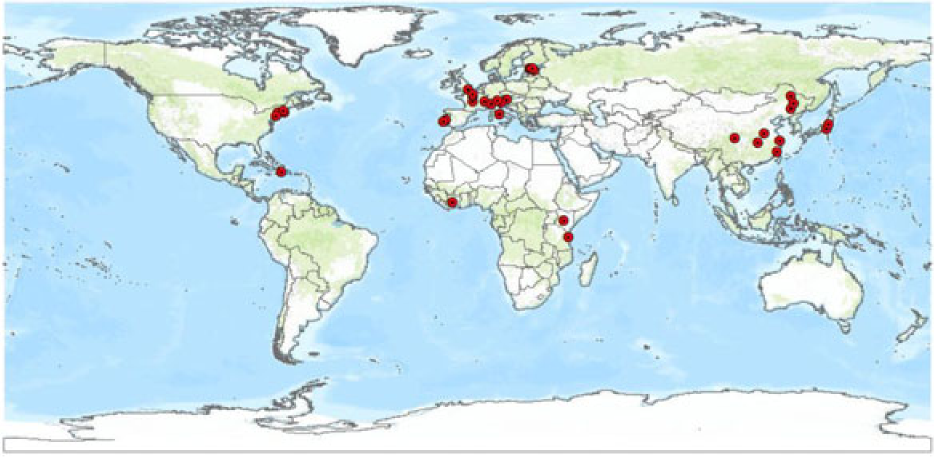

Fig. 1. Study areas of selected sample datasets.

smartphone applications are subject to the terms and conditions in the user agreement. The storage of such data is also subject to the data retention policy in the user agreement (e.g., data should be removed after 30 days). For CDR data, different countries may have various restrictions regarding how to utilize these signals legally. Multiple companies have released aggregated CDR data in their publications, such as AirSage Inc. and AT&T labs.[9] These datasets are useful in the fields of city planning and transportation. However, individual tracking data are generally confidential and can cause more legal problems than aggregated data. Although the usage of georeferenced mobile phone data is limited due to privacy issues, there have been several efforts to provide open-source or publicly accessible data, such as the Lausanne Data Collection Campaign.[10]

5 Researchers have also raised ethical concerns regarding utilizing such personally identifiable information (PII) in georeferenced mobile phone datasets, including corporate or state surveillance, marketing campaigns, and improper usage even for research purposes only (Boyles, et al. 2012, Calabrese, et al. 2015). It is worth noting that the general public often expresses concerns regarding the usage of georeferenced mobile phone data in various scenarios and the research topics in Tables 2 and 3 (c.f. Section 4) may impose privacy issues at a different level. Although PII is generally removed from such datasets (e.g., by assigning randomly generated unique false IDs), individual-oriented research such as trajectory visualization can risk exposing user identifications even with fully anonymized data. For instance, family members can easily identify a certain individual based on estimated work/home locations. Such problem is minimized in aggregate-level research. Researchers from the AT&T Lab have also utilized representative "synthetic call records" to mathematically obscure any data that could tend to identify people (Talbot 2013).

6 Service usage information has been widely adopted to analyze the social interaction between phone users. However, few studies have linked this social interaction with spatio-temporal tracking information and investigated these two types of information in a synergistic way. An interesting research question is to identify the interaction between "connected presence" and "physical presence" by calculating the co-location time points of these users with people in their mobile social networks.

To further structure the findings and the details of the 47 mobile phone datasets in the Appendix, we provide an abstraction of research questions in Section 4 to help with structuring the types of information stored in mobile phone datasets. It can also be applied as a reference framework to other future studies in the area of mobile computing and mobile data collection.

4. Research Questions Based on Georeferenced Mobile Phone Data

Based on the categorized information in Section 3, it is feasible to classify the research questions that can be explored based on georeferenced mobile phone datasets. These questions fit into two categories: individual-oriented and aggregate-oriented [Tables 2 (individual-oriented research) and 3 (aggregate-oriented research)]. The former focuses on analyzing mobility patterns from users' perspective, while the latter aims to study the movement patterns from a city (or even a country's perspective). These tables synergize and extend the categories of spatio-temporal aggregations in Andrienko *et al.* (2011) (see table 1 in Andrienko et al., 2011) from several perspectives: (i) The original framework only focused on user movement patterns from an aggregated perspective, but we further extend this framework to represent individual behavior patterns. (ii) We synergize and simplified the original framework to better serve new researchers in the field. For instance, one of the categories in Andrienko *et al.* (2011) Table 1 [$T \times T \rightarrow (S \times S \rightarrow A)$ – matrix arrangement of flow maps] is not represented in our model as

Table 2. Individual-oriented research topics.

Scenarios	Examples	Potential applications
$U(A) \rightarrow (S)$ Spatial distribution of user activities	Analyze points of interest (POIs) associated with users, i.e., regularly visited places.	Characterize users based on what types of places they visit and suggest new places.
$U(A) \rightarrow (T)$ Temporal distribution of user activities	Analyze special time units associated with users, i.e., regular time to make phone calls.	Characterize users based on the temporal pattern of their activities and provide a prediction to optimize call traffic loads.
$U(A) \rightarrow (S \times T)$ Spatio-temporal distribution of user activities	Analyze user trajectory patterns.	Classify user evacuation routes after an earthquake and identify outlier users that may need special attention.
$U(A) \times U(A) \rightarrow (S)$ Spatial distribution of user interaction/correlation	Analyze the location distribution of both ends of phone calls.	Analyze the spatial decay effect in mobile communication, e.g., do people who live close to each other make more or less phone calls? Provide references for mobile advertising companies.
$U(A) \times U(A) \rightarrow (T)$ Temporal distribution of user interaction/correlation	Analyze the time unit in which the particular user pairs make phone calls.	Classify connections between user pairs to enrich user profiles (e.g., two people who only call each other during work hours are most likely professionally connected).
$U(A) \times U(A) \rightarrow (\mathrm{S} \times \mathrm{T})$ Spatio-temporal distribution of user interaction/correlation	Analyze the trajectory patterns of a particular user pair *A* and *B*.	Extract users with similar trajectory patterns and improve friend recommendation in social network apps.

Table 3. **Urban-oriented research topics.**

Scenarios	Example	Potential applications
$S \rightarrow$ A Summary of attribute values at each location	Analyze the mobility count in each location, i.e., identify urban hotspots of mobility.	Detect the cluster of population in different areas and examine whether a city has sufficient public facilities (e.g., schools and hospitals).
$T \rightarrow$ A Summary of attribute values for each time unit	Analyze the mobility count for each time unit.	Examine the activeness of a population during different time units (e.g., hourly, daily, and monthly).
$S \times S \rightarrow$ A Correlation/interaction between the locations	Analyze the correlation between two locations regarding their mobility count.	Identify similarity districts with similar population clusters and provide input for urban planning.
$T \times T \rightarrow$ A Correlation/interaction between time points	Analyze the correlation between morning and afternoon patterns regarding their mobility count.	Identify time spans with similar activity patterns and help service providers to optimize traffic loads.
$S \rightarrow (T \rightarrow A)$ Time series of summarized attribute values in each location	Analyze the hourly time series of mobility count in each location.	Identify the time-dependent cluster pattern of a certain area (e.g., district with lots of bars may need a larger dispatch of police officers).
$T \rightarrow (S \rightarrow A)$ Summary of attribute values associated with each time unit	Analyze the spatial distribution (i.e., as an array) of mobility count for each time unit.	Analyze the spatial distribution of a population at a certain time (e.g., how much more public transportation do we need during holiday seasons?).
$S \times S \rightarrow (T \rightarrow A)$ For each pair of locations, time series of correlation/interaction between the locations by time intervals	Analyze the hourly communication flow between location cells.	Characterize the dynamics of mobile communication between two urban areas, re-define urban divisions based on the level of mobile connections, and analyze social network distribution on a city scale.
$T \times T \rightarrow (S \rightarrow A)$ For each pair of time units, aggregate attributes representing changes in different locations.	Spatial matrix of aggregated mobility count change	Identify which area has the fastest mobility increase during rush hours; helps to identify the changes of activities in different urban divisions

it is more meaningful for maintaining the completeness of a mathematical model rather than providing a clear and informative outline to new researchers. (iii) We demonstrate how this framework can be customized and transformed to represent more specific research questions, such as spatio-temporal clustering in Table 4.

The abbreviations are defined as follows:

1 *U*: users; *S*: space; *T*: time; *A*: attribute.
2 *U*(*A*): users with attribute *A*. Note that *A* can be a null value, in which case *U*(*A*) stands for all the users in a certain data sample.

In real-world applications, it is also feasible to modify or extend Tables 2 and 3 . For instance, clustering techniques can facilitate the extraction of the spatio-temporal characteristics of user mobility patterns (Yuan and Raubal, 2012a, Yuan and Raubal, 2012b). Table 3 can be easily modified to investigate spatio-temporal clustering methods that can be applied to mobile phone datasets at various spatio-temporal scales (Table 4).

Table 4. Different scenarios and corresponding clustering features.

Scenarios	Clustering feature
$S\rightarrow(T\rightarrow A)$ Time series of summary attribute values in each location	Time series
$T\rightarrow(S\rightarrow A)$ Summary attribute values associated with each time unit	Spatial series
$S\times S\rightarrow(T\rightarrow A)$ For each pair of locations, time series of flows between locations	Vector time series
$T\times T\rightarrow(S\rightarrow A)$ For each pair of time units, aggregate attributes representing changes between the spatial configurations.	Spatial series of the change of aggregated attributes
$T\times T\rightarrow(S\times S\rightarrow A)$ For each pair of time units, aggregate moves (flows) of objects between locations.	Matrix in spatial dimension

For example, the hourly phone call frequencies can be viewed as regular time series. In this case, many well-developed clustering techniques for time series data can be applied, such as the longest common subsequence (LCSS) described in Hirschberg (1977).

The constructed model provides a guideline for analyzing mobile phone data. Table 5 shows the correspondence between popular exemplary studies and the topics listed in Tables 2 and 3.

For instance, previous research by the authors improved the traditional edit distance algorithm by incorporating both spatial and temporal information into the cost functions (Yuan and Raubal 2014). The idea of extending the traditional algorithm was inspired by combining three topics in Table 2 [$U(A)\rightarrow(S\times T)$, $U(A)\times U(A)\rightarrow(S)$, and $U(A)\times U(A)\rightarrow(S\times T)$], so that researchers can preserve both space and time information from string-formatted CDR data. Another previous research went one step further from identifying aggregated mobility patterns by combining two topics [$S\rightarrow(T\rightarrow A)$ and $S\times S\rightarrow(T\rightarrow A)$]. Using hourly time series, we extract and represent the *dynamic mobility patterns* in different urban areas applying a dynamic time warping (DTW) algorithm. Follow-up research by the authors also adopted a similar framework [$S\rightarrow(T\rightarrow A)$ and $S\times S\rightarrow(T\rightarrow A)$]

Table 5. The correspondence between exemplary topics and the constructed model.

Research question	Formalized topic
Exploring individual activity space	$U(A)\rightarrow(S)$
	$U(A)\rightarrow(T)$
Trajectory similarity measure of mobile phone users	$U(A)\rightarrow(S\times T)$
	$U(A)\times U(A)\rightarrow(S)$
	$U(A)\times U(A)\rightarrow(S\times T)$
Modeling urban clusters	$S\rightarrow A$
	$T\rightarrow A$
	$T\rightarrow(S\rightarrow A)$
Modeling urban rhythms	$S\rightarrow(T\rightarrow A)$
	$S\times S\rightarrow(T\rightarrow A)$

to investigate the hourly mobility patterns in Harbin (a Chinese city), Paris (a French city), and Tallinn (an Estonian city) (Ahas, et al. 2015). The results show important differences in time use in Chinese, Estonian, and French cities. These studies can be utilized by researchers and policy makers to understand the dynamic nature of urban areas, as well as updating environmental and transportation policies.

5. Conclusions and Future Work

Mobile phone data as an input to the analysis of human mobility have the potential to transform research in diverse fields, such as geography, transportation, and economics. This review investigated the richness and availability of three types of information individually by reviewing existing mobile phone datasets, as well as constructing a detailed framework to discern data types stored in mobile phone datasets. Forty-seven datasets were analyzed focusing on the following three perspectives: *spatio-temporal tracking information*, *service usage information*, and *demographic profiles*.

Based on the summarized information, Tables 2 and 3 formalized the potential topics from two perspectives: individual-oriented and aggregate-oriented. Note that the different categories in these two tables are often combined in real-world studies. These research questions are generated as a framework for our future research agenda on reducing the complexity of dynamic data and analyzing human mobility patterns in mobile phone datasets. In practice, these questions can also be extended to accord with specific topics. For instance, Table 4 listed an extension of the topics focusing on clustering analysis. The objective of this review is to provide a generalizable framework and a summary table for exploring human mobility patterns from mobile phone data. We did not aim at covering every single dataset. Instead, we conducted a literature review as a demonstration of how the itemized topics in Tables 2 and 3 can be connected to real-world examples. An interesting extension can be conducted based on the movement ecology approach proposed by Nathan *et al.* (2008) to answer questions, such as "which external factors affect a movement and how?" The detailed information listed in Table A-1 also offers a valuable resource to facilitate institutional collaboration. For future research, we will further improve the model constructed in this work as well as extend the review of metadata when new datasets become available. This research can also be extended as a benchmark to assess the completeness of mobile phone data.

Acknowledgements

This research was supported by the Swiss National Science Foundation [project # 205121_141284]. The reviewers provided excellent feedback, which helped us improve the content and clarity of this paper.

Notes

* Correspondence address: Yihong Yuan, Department of Geography, Texas State University, San Marcos, TX 78666, USA. E-mail: yuan@txstate.edu

[1] http://apps.webofknowledge.com/

[2] dl.acm.org

[3] https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=soc_geographycartography

[4] http://www.bizjournals.com/boston/blog/bottom_line/2012/04/cell-phone-tracking.html

[5] The NMEA 0183 Protocol, http://www.tronico.fi/OH6NT/docs/NMEA0183.pdf

[6] http://reality.media.mit.edu/dataset.php

[7] http://www.airsage.com/

[8] https://www.maxmind.com/en/home

[9] http://www.research.att.com/

[10] https://research.nokia.com/page/11367

References

Adusei, I. K., Kyamakya, K. and Jobmann, K. (2002). Mobile positioning technologies in cellular networks: an evaluation of their performance metrics. *2002 Milcom Proceedings, Vols 1 and 2* 1239–1244.

Agar, J. (2004). *Constant Touch: A Global History of the Mobile Phone*. Cambridge, UK: Icon Books.

Ahas, R. (2005). Mobile phones and geography: social positioning method. In: *Power over Time-space: Inaugural Nordic Geographers Meeting*. Lund, Sweden, pp.1–8.

Ahas, R., Aasa, A., Roose, A., Mark, U. and Silm, S. (2008). Evaluating passive mobile positioning data for tourism surveys: an Estonian case study. *Tourism Management* 29, pp. 469–486.

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, pp. 2017–2039.

Ahas, R., Silm, S., Jarv, 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, pp. 3–27.

Alanen, K., Wirola, L., Kappi, J. and Syrjarinne, J. (2006). Inertial sensor enhanced mobile RTK solution using low-cost assisted GPS receivers and internet-enabled cellular phones. 2016 *IEEE/ION Position, Location and Navigation Symposium* 1-3, pp. 920–926.

Alguliyev, R., Imamverdiyev, Y., Zargari, H. and Bairami, S. (2010). Relationship between mobile switching center information and social behavior in week days and holidays in a telecommunication area. *Fourth International Conference on Digital Society: Icds 2010, Proceedings* 136–138.

Andrienko, G., Andrienko, N.,., Bak, P., Keim, D., Kisilevich, S. and Wrobel, S. (2011). A conceptual framework and taxonomy of techniques for analyzing movement. *Journal of Visual Languages and Computing* 22, pp. 213–232.

Batsanov, A. S., Bryce, M. R., Chesney, A., Howard, J. A. K., John, D. E., Moore, A. J., Wood, C. L., Gershtenman, H., Becker, J. Y., Khodorkovsky, V. Y., Ellern, A., Bernstein, J., Perepichka, I. F., Rotello, V., Gray, M. and Cuello, A. O. (2001). Synthesis and crystal engineering of new halogenated tetrathiafulvalene (TTF) derivatives and their charge transfer complexes and radical ion salts. *Journal of Materials Chemistry* 11, pp. 2181–2191.

Bayir, M. A., Demirbas, M. and Eagle, N. (2010). Mobility profiler: a framework for discovering mobility profiles of cell phone users. *Pervasive and Mobile Computing* 6, pp. 435–454.

Becker, R. A., Caceres, R., Hanson, K., Loh, J. M., Urbanek, S., Varshavsky, A. and Volinsky, C. (2011). A tale of one city: using cellular network data for urban planning. *IEEE Pervasive Computing* 10, pp. 18–26.

Boyles JL, Smith A, and Madden M (2012). Privacy and data management on mobile devices. Pew Research Center.

Brimicombe, A. and Li, C. (2009). *Location-based Services and Geo-information Engineering*. Wiley-Blackwell: Chichester, UK; Hoboken, NJ.

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J. Jr. and Ratti, C. (2013). Understanding individual mobility patterns from urban sensing data: a mobile phone trace example. *Transportation Research Part C: Emerging* 26, pp. 301–313.

Calabrese, F., Ferrari, L. and Blondel, V. D. (2015). Urban sensing using mobile phone network data: a survey of research. *ACM Computing Surveys* 47(2), pp. 1–20.

Caron, C., Roche, S., Goyer, D. and Jaton, A. (2008). GIScience journals ranking and evaluation: an international Delphi study. *Transactions in GIS* 12, pp. 293–321.

Dashdorj, Z., Serafini, L., Antonelli, F. and Larcher, R. (2013). Semantic enrichment of mobile phone data records. In: *Proceedings of the 12th International Conference on Mobile and Ubiquitous Multimedia*. Luleå, Sweden: ACM New York, NY.

Djuknic, G. M. and Richton, R. E. (2001). Geolocation and assisted GPS. *Computer* 34, pp. 123–125.

Farrahi, K., Emonet, R., Ferscha, A. and Ieee (2012). Socio-technical network analysis from wearable interactions. *2012 16th International Symposium on Wearable Computers (Iswc)* 9–16.

Farrahi, K. and Gatica-Perez, D. (2010). Probabilistic mining of socio-geographic routines from mobile phone data. *IEEE Journal of Selected Topics in Signal Processing* 4, pp. 746–755.

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

Hirschberg, D. S. (1977). Algorithms for longest common subsequence problem. *Journal of the ACM* 24, pp. 664–675.

Hong, J. M. and Lee, J. H. (2011). Optimal mobile switching center positioning and cells assignment using Lagrangian heuristic. *IEICE Transactions on Fundamentals of Electronics* E94a, pp. 2425–2433.

Horak, R. (2007). *Telecommunications and Data Communications Handbook*, Hoboken N.J.: Wiley-Interscience.

Jenkins, R. I. (2013). *Bluetooth and Wireless Local Area Networks: Security Guides*. New York: Novinka.

Kang, C. G., Gao, S., Lin, X., Xiao, Y., Yuan, Y. H., Liu, Y. and Ma, X. J. (2010). Analyzing and geo-visualizing individual human mobility patterns using mobile call records. *2010 18th International Conference on Geoinformatics*.

Kasemsan, M. L. K. and Ratsameethammawong, P. (2010). Moving mobile phone location tracking by the combination of GPS, Wi-Fi and cell location technology. In: *Business Transformation through Innovation and Knowledge Management: An Academic Perspective*, Vol. 1-4, pp.979–985.

Li, S., Lou, Y. S. and Liu, B. (2014). Bluetooth aided mobile phone localization: a nonlinear neural circuit approach. *ACM Transactions on Embedded Computing Systems* 13(4), pp. 1–15.

Liu, F., Janssens, D., Wets, G. and Cools, M. (2013a). Annotating mobile phone location data with activity purposes using machine learning algorithms. *Expert Systems with Applications* 40, pp. 3299–3311.

Liu, Y., Dashti, M. and Zhang, J. (2013b). Indoor localization on mobile phone platforms using embedded inertial sensors. *2013 10th Workshop on Positioning, Navigation and Communication (Wpnc)*.

Miller, H. J. (2009). Geographic data mining and knowledge discovery: an overview. In: Miller, H. J. and Han, J. (eds) *Geographic Data Mining and Knowledge Discovery*. 2nd ed. CRC Press: London, pp.3–32.

Murakami, H., Ito, A., Watanabe, Y. and Yabe, T. (2007). Mobile phone based ad hoc network using built in bluetooth for ubiquitous life. *Eighth International Symposium on Autonomous Decentralized Systems, Proceedings* 137–143.

Nathan, R., Getz, W. M., Revilla, E., Holyoak, M., Kadmon, R., Saltz, D. and Smouse, P. E. (2008). A movement ecology paradigm for unifying organismal movement research. *Proceedings of the National Academy of Sciences of the United States of America* 105, pp. 19052–19059.

Oxendine, C. E. and Waters, N. (2014). No-notice urban evacuations: using crowdsourced mobile data to minimize risk. *Geography Compass* 8, pp. 49–62.

Phithakkitnukoon, S., Horanont, T., Di Lorenzo, G., Shibasaki, R. and Ratti, C. (2010). Activity-aware map: identifying human daily activity pattern using mobile phone data. In: Salah, A. A., Gevers, T., Sebe, N. and Vinciarelli, A. (eds) *HBU 2010* LNCS. Heidelberg: Springer, pp.14–25.

Quarmyne, J. and Pachter, M. (2014). Inertial navigation system aiding using vision. *2014 American Control Conference (Acc), pp.* 85–90

Ratti, C., Sevtsuk, A., Huang, S. and Pailer, R. (2005). Mobile landscapes: Graz in real time. In: *The 3rd Symposium on LBS & TeleCartography*. Vienna, Austria: Lecture Notes in Geoinformation and Cartography, Springer pp. 433–444.

Reades, J., Calabrese, F., Sevtsuk, A. and Ratti, C. (2007). Cellular census: explorations in urban data collection. *IEEE Pervasive Computing* 6, pp. 30–38.

Sagl, G., Delmelle, E. and Delmelle, E. (2014). Mapping collective human activity in an urban environment based on mobile phone data. *Cartography and Geographic Information Science* 41, pp. 272–285.

Silm, S. and Ahas, R. (2014). Ethnic differences in activity spaces: a study of out-of-home nonemployment activities with mobile phone data. *Annals of the Association of American Geographers* 104, pp. 542–559.

Song, C. M., Qu, Z. H., Blumm, N. and Barabasi, A. L. (2010). Limits of predictability in human mobility. *Science* 327, pp. 1018–1021.

Talbot, D. (2013). How to mine cell-phone data without invading your privacy. http://www.technologyreview.com/news/514676/how-to-mine-cell-phone-data-without-invading-your-privacy/.

Versichele, M., Neutens, T., Delafontaine, M. and Van de Weghe, N. (2012). The use of Bluetooth for analysing spatiotemporal dynamics of human movement at mass events: a case study of the Ghent Festivities. *Applied Geography* 32, pp. 208–220.

Wang, S. S. P., Green, M. and Malkawi, M. (2002). Mobile positioning technologies and location services. *Rawcon 2002: IEEE Radio and Wireless Conference, Proceedings, pp.* 9–12,

Wesolowski, A., Eagle, N., Noor, A. M., Snow, R. W. and Buckee, C. O. (2013). The impact of biases in mobile phone ownership on estimates of human mobility. *Journal of the Royal Society Interface* 10, DOI: 10.1098/rsif.2012.0986.

Yuan, Y. and Raubal, M. (2010). Spatio-temporal knowledge discovery from georeferenced mobile phone data. In: Gottfried, B., Laube, P., Klippe, A., Weghe, N. V. D. and Billen, R. (eds) *MPA'10 – 1st Workshop on Movement Pattern Analysis*. Zurich, Switzerland: pp.121–126.

Yuan, Y. and Raubal, M. (2012a). Extracting dynamic urban mobility patterns from mobile phone data. In: *Geographic Information Science – 7th International Conference*, Lecture Notes in Computer Science. Columbus, USA: Springer, pp.354–367.

Yuan, Y. and Raubal, M. (2012b). Similarity measurement of mobile phone user trajectories – a modified edit distance method. In: *Workshop on "Progress in Movement Analysis – Experiences with Real Data"*. Zurich, Switzerland.

Yuan, Y. and Raubal, M. (2014). Measuring similarity of mobile phone user trajectories – a spatio-temporal edit distance method. *International Journal of Geographical Information Science* 28, pp. 496–520.

Yuan, Y., Raubal, M. and Liu, Y. (2012). Correlating mobile phone usage and travel behavior – a case study of Harbin, China. *Computers, Environment and Urban Systems* 36, pp. 118–130.

Appendix

Due to page limits and to facilitate public sharing, Table A-1 (a summary of metadata) and a complete list of reviewed datasets are hosted on Slideshare and can be downloaded from the following: http://www.slideshare.net/mobile_data/table-a1-a-summary-of-georeferenced-mobile-phone-datasets-52652066.