Build Simul (2024) 17: 3–21
https://doi.org/10.1007/s12273-023-1043-z

Modeling urban scale human mobility through big data analysis and machine learning

Yapan Liu, Bing Dong (✉)

Department of Mechanical & Aerospace Engineering, Syracuse University, 263 Link Hall, Syracuse, NY 13244, USA

Abstract

In the United States, the buildings sector consumes about 76% of electricity use and 40% of all primary energy use and associated greenhouse gas emissions. Occupant behavior has drawn increasing research interests due to its impacts on the building energy consumption. However, occupant behavior study at urban scale remains a challenge, and very limited studies have been conducted. As an effort to couple big data analysis with human mobility modeling, this study has explored urban scale human mobility utilizing three months Global Positioning System (GPS) data of 93,000 users at Phoenix Metropolitan Area. This research extracted stay points from raw data, and identified users' home, work, and other locations by Density-Based Spatial Clustering algorithm. Then, daily mobility patterns were constructed using different types of locations. We propose a novel approach to predict urban scale daily human mobility patterns with 12-hour prediction horizon, using Long Short-Term Memory (LSTM) neural network model. Results shows the developed models achieved around 85% average accuracy and about 86% mean precision. The developed models can be further applied to analyze urban scale occupant behavior, building energy demand and flexibility, and contributed to urban planning.

Keywords

urban human mobility
big data analysis
urban scale occupant behavior
recurrent neural networks

Article History

Received: 10 February 2023
Revised: 28 April 2023
Accepted: 12 May 2023

1 Introduction

1.1 Background

In the United States (U.S.), the buildings sector consumes about 76% of electricity use and 40% of all primary energy use and associated greenhouse gas emissions (DOE 2015). Occupant behavior, which includes the presence of people in the space, interactions between the occupant and building systems, and occupant adaptations to the built environment, has drawn increasing attention because of its impact on the energy consumption of buildings. Many researchers have investigated occupant behaviors at a single building scale, and very limited studies have been conducted at community scale or urban-scale (Barbour et al. 2019; Wu et al. 2020; Lu et al. 2021). With the development of information technologies, such as mobile technology, urban sensing, and IoT, big data that was generated by those technologies provides opportunities to better understand occupant behavior at urban scale (Salim et al. 2020).

Meanwhile, the International Energy Agency (IEA) Energy in Buildings and Community (EBC) Annex 67 (Jensen et al. 2017) investigated the energy flexibility in buildings, and developed the definition of it as the ability to manage building energy demand and generation considering local climate conditions, occupant needs, and energy network requirements. This definition clearly shows that building occupant behavior is a significant factor that impacts building energy demand and its flexibility. Recently, the Building Technologies Office from the United States Department of Energy (DOE) has initiated research on Grid-interactive Efficient Buildings (GEB) (DOE 2022), to make building operations coordinate with the grid regarding the amount and timing of energy use and reduce greenhouse gas emissions from buildings. Both GEB and Annex 67 require advanced understandings of building energy demand and energy flexibility.

Researchers identified occupant behavior as the main cause of the uncertainty of building energy performance (Yan et al. 2015). On a city level, the urban scale building

E-mail: bidong@syr.edu

energy models (UBEMs) aspires to become key planning tools for the holistic optimization of buildings, urban design, and energy systems in neighborhoods and districts (Happle et al. 2018). However, existing UBEMs mainly use physical simulation models of energy flows in and around groups of buildings, to represent the impact of the urban context on building energy demand (Fonseca and Schlueter 2015; Cerezo Davila et al. 2016; Reinhart and Cerezo Davila 2016). And inappropriate choice of occupant behavior model could result in oversized district energy systems, leading to over-investment and low operational efficiency (Happle et al. 2018). Therefore, it is critical to investigate urban scale occupant behavior and integrate it to the urban scale energy modeling process. Studies (Barbour et al. 2019; Dong et al. 2019; Wu et al. 2020) have investigated occupant behaviors at urban scale, the authors developed building occupancy models and integrated with UBEMs to understand their impact on energy consumption. Results (Wu et al. 2020) show that the reduction of predicted cooling and heating energy were up to 40% and 60% respectively.

Furthermore, increasingly research efforts have been focused on human mobility study to understand the occupant behavior at urban scale. As part of human mobility study, recent studies extracted occupant profiles by processing GPS data from various data sources (Wu et al. 2020), including social media data (Lu et al. 2021), and Call Detail Records (CDR) data (Barbour et al. 2019). Other studies focused on modeling and predicting urban human mobility, such as future movement patterns and next locations (Feng et al. 2018; Wang et al. 2019b; Yang et al. 2020). Because of the temporal and spatial nature of human mobility data, studies have investigated the mobility patterns using a clustering-based algorithm naming the DBSCAN (Cesario et al. 2013; Huang and Wong 2015; Tang et al. 2015). And Recurrent Neural Network (RNN) models have been developed to learn human mobility patterns and predict next locations (Liu et al. 2016; Huang 2017; Khoroshevsky and Lerner 2017; Feng et al. 2018; Wang et al. 2019a; Guo et al. 2020; Yang et al. 2020).

In this study, we focus on developing the urban scale human mobility models utilizing GPS data collected from smart mobile phones. As forementioned, considering the spatial and temporal characteristics of human mobility data, this study adopted DBSCAN and RNN to identify, model and predict patterns of human mobility for 93,000 users. The developed models can be further applied to analyze urban scale occupant behavior, building predictive control, building energy demand and energy flexibility.

1.2 Structure of this paper

This study modeled daily human mobility patterns from raw GPS data, and trained RNN models for 12-hour ahead mobility predictions. The paper is organized as follows: Section 2 covers literature review of the current building occupant behavior study and human mobility study. We detailed out the method of big data processing and analysis, as well as human mobility modeling and prediction in Section 3. Section 4 presents the results and discussions of this study, including stay point extraction, DBSCAN clustering and location labeling, LSTM model training and testing, and model performance evaluation. Section 5 concludes this paper with conclusions, scientific contributions, and limitations of this study.

2 Literature review

2.1 Motivation

The IEA EBC has developed the Annex 66 (Yan et al. 2017) and Annex 79 (O'Brien et al. 2020) programs to advance the research of occupant behaviors in buildings. Among the scientific findings, data-driven modeling of occupant behavior is considered as a promising approach due to the fact of increasing data sources and rapid development of various sensing technologies. Additionally, research concludes that the building energy demand is largely impacted by occupant behaviors, and it could cause performance gap with insufficient consideration of occupant behaviors (Happle et al. 2018). Therefore, it is critical to investigate urban scale occupant behavior and include it into the urban scale energy modeling process. Occupant behavior in this study refers to the presence of people in the space, interactions between the occupant and building systems, and occupant adaptations to the built environment. Dong et al. (2022) developed a global building occupant behavior database including nine different categories of occupant behavior as well as indoor and outdoor environmental measurements. Those categories are door status, fan status, HVAC measurement, lighting status, occupancy measurement, occupant number measurement, plug load, shade status and window status.

Previous study (Wu et al. 2020) has developed a novel mobility-based approach to study urban scale occupant behavior and showed promising results of its impacts on building energy consumption. Similar studies have been conducted to explore occupant behavior through urban human mobility (Jiang et al. 2016; Barbour et al. 2019; Kang et al. 2021; Lu et al. 2021). But the model used in Barbour et al. (2019) and Jiang et al. (2016) is a statistical model which adopts Markov chain for temporal choices and a rank-based exploration and preferential return model for spatial choices. This study aims to develop a purely data driven occupant behavior model from raw GPS data. Human

mobility studies include indoor mobility (Biczók et al. 2014; Tang et al. 2016; Trivedi et al. 2021) and outdoor mobility (Pan et al. 2013; Chen et al. 2014; Wang and Taylor 2016; Lin et al. 2018; Chang et al. 2021a) research areas. The indoor mobility refers to movements between locations or zones within an indoor environment at single building scale. In contrast, the outdoor mobility relates to human movements among various locations (e.g., buildings, roads, etc.) at city level. This study focuses on understanding outdoor human mobility among different buildings at urban scale. With the knowledge of how people travel at urban scale, occupant behavior models (e.g., occupant's presence in a building) can be inferred (Jiang et al. 2016; Wu et al. 2020). This section provides a comprehensive literature review of the occupant behavior and human mobility studies.

2.2 Modeling occupant behavior at an urban scale

Urban building energy modeling is fundamental for optimization of building operations and urban planning. And occupant behavior is one of the main factors that largely impact building energy demand. Happle et al. (2018) reviewed different occupant behavior modeling approaches used by various urban building energy models. It concluded that, occupant behavior models at urban scale are still limited. And it also projected that advanced urban scale occupant behavior models are essential to holistic urban planning and energy infrastructure optimization. To develop occupant behavior models at urban scale, big data and advanced modeling methods are needed. Salim et al. (2020) has conducted a holistic review of available urban scale occupant-centric data cross different disciplines, those data can be applied to model occupant behavior and energy usage patterns at urban scale. It summarized available occupant-centric data source into six different categories, such as survey data, building data, Internet of Things (IoT) sensor data, crowdsourcing data, city spatial data, and mobility data. Among those data sources, the mobility data covers GPS data, CDR dataset, social media check-ins, data from location-based services (LBS), and transportation data. To address the challenges to model occupant behavior at urban scale, Dong et al. (2021) studied existing modeling methods in building science domain and beyond. Since the urban scale building applications still heavily rely on occupant behavior models at single building level, the paper has identified and discussed potential modeling approaches from other domains such as transportation, epidemiology, disaster management, and smart retail domain. The study finds out both recurrent neural networks and graphical networks have drawn much attention, and has shown promising results as well.

Recent studies have investigated occupant behavior and its impact on energy consumption at urban scale. Happle et al. (2020) developed data-driven and context-specific urban occupancy modeling methods based on LBS (location-based service) data collected from web mapping services in the downtown neighborhoods of 13 different U.S. cities. The study concludes that current standard occupancy schedules significantly overpredicted weekly building occupancy, and have significant impacts on district scale energy demand simulations. Built on the TimeGeo framework (Jiang et al. 2016), Barbour et al. (2019) used the CDRs (Call Detail Records) of 1.92 million anonymous mobile users to develop urban scale occupancy model and simulated 3.54 million people in the building energy modeling study. Compared with standard building occupancy rates defined by the Department of Energy (DOE), the study observed energy consumption reduction could reach 15% for residential buildings and 21% for commercial buildings. The TimeGeo framework was established on a time-inhomogeneous Markov chain model for modeling temporal choices, and a rank-based exploration and preferential return (r-EPR) model for generating spatial choices. Another study (Kang et al. 2019) implemented max normalization and K-means clustering methods, to develop occupancy models for different types of buildings from GPS data collected by social network software. Dong et al. (2019) derived urban scale occupancy patterns at the individual building level, and compared them with synthetic schedules from DoE reference models. Energy study of example buildings showed significant differences up to 50% for large office buildings and 30% for strip malls.

By analyzing raw GPS data collected by smart mobile phone users, Wu et al. (2020) presented a new approach to derive empirical occupancy profiles for various building types in San Antonio, Texas. The study combined mobility data with building data (998 buildings) to capture more realistic occupant dynamics within different building types. Based on the derived occupancy rates, simulation results show that the cooling and heating energy demand could be reduced by up to 40% and 60% respectively. Lu et al. (2021) investigated typical building occupancy schedules using data from social networks. Results showed that building occupancy profiles derived from various sources show similar trends. Similarly, Kang et al. (2021) also utilized GPS data collected from social networks to study typical weekly occupancy profiles for non-residential buildings. It adopted cluster analysis to extract the typical patterns of weekly occupancy profiles, 16 buildings were selected and tested to demonstrate the proposed approach. Compared with ASHRAE Standard 90.1, results showed that the standard could underestimate or overestimate building occupancy

conditions for different types of buildings at different times of the day.

2.3 Human mobility

As an interdisciplinary field, human mobility study has drawn an increasing research attentions in recent years (Wang et al. 2019b; Yang et al. 2022). It refers to the movement of people among different locations at different times of the day. As the evolving of modern technologies, multi-source of big data provides new opportunities to explore urban human mobility across different disciplines, such as traffic analysis, disaster management, and epidemiology. For instance, social network services and mobile computing enable urban planners to investigate city dynamics (Pan et al. 2013). Taking advantage of the human mobility data and social media data, Pan et al. (2013) developed a system to detect and describe traffic anomalies in big cities. By studying the driver's routine behavior on an urban road, the proposed system can identify anomalous traffic patterns or events. The system was tested with social media datasets and GPS trajectory datasets of taxi in Beijing. In another study, a social media based traffic congestion monitoring system was evaluated by the Twitter data and INRIX probe speed datasets in two U.S. major cities (Chen et al. 2014). With smart card data collected by Tokyo Metro and social media data from Twitter, researchers from Japan visualized the integrated traffic and social media analysis (Itoh et al. 2014). Another study shows that better road traffic speed prediction can be achieved by fusing traditional speed sensing data with social media data, trajectory sensors from map and traffic service platforms (Lin et al. 2018). Jiang et al. (2019) proposed a model to predict cyclists' destinations based on data from Mobike, which contains a GPS/3G module, and used Shanghai Mobike trajectory data to explore bike-sharing systems traffic flow and demand prediction, and address bike lane planning issues (Bao et al. 2017; Jiang et al. 2019).

By analyzing individuals' mobility data collected from Twitter, Wang and Taylor (2016) examined how natural disasters influence human mobility patterns in urban populations. Results showed that human mobility patterns are unlikely to deviate from the fundamental power-law during a natural disaster, but natural disasters can significantly change human mobility patterns even where the fundamental power-law still holds. Mohammadi and Taylor (2017) introduced a multivariate autoregressive model to predict buildings' energy demand using mobility data collected from Twitter. Another research presented a social media-based approach to assess the severity and location of disaster impacts on highways (Chen et al. 2020). The author investigated disaster impacts on highways brought by Hurricane Harvey in Houston, the results showed the presented approach is feasible and applicable. On the other hand, a study was conducted to estimate massive population displacement during or after natural disasters. Wilson et al. (2016) analyzed human mobility patterns based on call detail records. They investigated the population left soon after the 2015 Nepal earthquake, population flow destinations, and return rates, which indicate where humanitarian aid should be directed and help to identify recovery and reconstruction progress.

More recently, human mobility study has been conducted to understand impacts of COVID-19 pandemic (Buckee et al. 2020; Chang et al. 2021a, 2021b; Gozzi et al. 2021; Schulte-Fischedick et al. 2021; Liu et al. 2022). Buckee et al. (2020) suggested that the aggregated mobility data can provide approximately real time information about changes in human mobility patterns at urban scale. This will lead to efficient interventions on preventing the spreads of COVID-19. Chang et al. (2021a) introduced a decision-support tool to quantify the impact of human mobility dynamics on COVID-19 infection rates. With the knowledge of locations that infected individuals visited, the model can provide detailed analysis and inform more effective and equitable policy responses for COVID-19 (Chang et al. 2021b). Another study (Gozzi et al. 2021) used anonymized mobile phone data to estimate the effects of social inequalities cross communities on the mitigation of COVID-19. Other studies (Schulte-Fischedick et al. 2021; Liu et al. 2022) estimated the carbon emission changes upon the mobility during COVID-19.

2.3.1 Machine learning in modeling human mobility

Recent studies of human mobility focused on using machine learning approaches to derive mobility patterns, model and predicting the mobility trajectories. Considering the spatial characteristic of the mobility data, clustering algorithm like DBSCAN has been adopted by different studies to identify locations of interest and explore human mobility at urban scale. Liu et al. (2021) proposed a novel space-time analytical framework to study the AOIs (areas of interest) using the taxi GPS data in Manhattan, NYC. The study applied the ST-DBSCAN (Spatial-temporal Density-Based Spatial Clustering of Applications with Noise) algorithm and successfully identified 31 unique AOIs that highly correlated to famous places, landmarks, transit stations, etc. Currently, taxi trip plays an important role in the daily movements of urban residents, Tang et al. (2015) uncovered urban human mobility by analyzing city scale of GPS data from taxis in Harbin, China. The study adopted the DBSCAN algorithm to identify the clusters of pick-up and drop-off locations. The proposed approach was tested in a city area by splitting

the area into unique transportation districts, mobility patterns were modeled from distribution of taxi trips. Similar studies (Bonnetain et al. 2021; Smolak et al. 2021) have also taken advantage of DBSCAN to analyze and model human mobility at urban scale. Smolak et al. (2021) used DBSCAN to cluster stay point and concluded that the data processed with DBSCAN are more predictable. Bonnetain et al. (2021) utilized network signaling data to capture stationary activities from individual mobile devices, and rebuilt a fine-grained human mobility trajectories at urban scale. However, this study focused on estimating the number of trips over time inferred via the proposed framework. The analysis of individual mobility, stay duration, and mobility prediction are limited. Jurdak et al. (2015) analyzed geotagged tweets of more than six million users in Australia. DBSCAN algorithm was used in this study to identify locations from raw data and reduce the vagueness.

Other studies (Liu et al. 2016; Huang 2017; Khoroshevsky and Lerner 2017; Feng et al. 2018; Wang et al. 2019a; Guo et al. 2020; Yang et al. 2020) applied neural networks to model human mobility at urban scale by predicting next locations. Given the temporal nature of the mobility data, RNN and its variant were mainly adopted in the literature. Feng et al. (2018) combined a multi-modal embedding RNN with a historical attention model to capture both sequential transitions and periodicity in the mobility data. Yang et al. (2020) developed the Flashback framework for modeling sparse user mobility traces, it explicitly uses spatio-temporal contexts to search past states with high predictive power for next location prediction. Liu et al. (2016) extended RNN to the Spatial Temporal Recurrent Neural Networks (ST-RNN) by incorporating both spatial distance information and time interval information. Guo et al. (2020) integrated attention mechanism to RNN model, and introduced Attentional Recurrent Neural Network framework (ARNN) to uncover both sequential regularity and transition regularity in the mobility data. Another study (Wang et al. 2019a) analyzed the characteristics of human mobility from real world GPS dataset, and extracted information of spatio-temporal regularities and user mobility preferences. Based on extracted feature, the study adopted LSTM network for multi-user destination prediction.

2.4 Summary of research gaps and research contributions

As shown in Figure 1, this section reviewed state of arts in both Occupant Behavior and Human Mobility domains. The discussions spread from single or multiple buildings level to urban scale. It covers modeling techniques, data sources and various applications. Through literature review, we have summarized current key knowledge gaps and our contributions in the occupant behavior and human mobility studies, as the following:

(1) *Gap*: Lack of a prediction model of urban scale occupant behavior. Studies (Happle et al. 2018; Kang et al. 2021; Lu et al. 2021) have developed building occupancy profiles at urban scale based on various data sources. However, a generalized urban scale occupant behavior modeling framework is still needed. With the information of predicted occupant behavior at an urban scale, one can assess and better predict building energy demands, which contributes to better optimal building controls and increased energy flexibility. In this study, *our contribution* is to develop novel urban scale occupant behavior models using machine learning approaches based on raw GPS data that models individual occupant movement.

(2) *Gap*: Insufficient understandings of the spatial-temporal patterns of the human mobility modeling at urban scale with fine granularity and a constant time (e.g., a whole day). Prior studies (Liu et al. 2016; Huang 2017; Khoroshevsky and Lerner 2017; Feng et al. 2018; Wang et al. 2019a; Guo et al. 2020; Yang et al. 2020) focus on predicting only single next location or a point of interest (e.g. next location or destination), while our study focuses on exploring and predicting the complete trajectory of human mobility for a whole day. In this study, *our contribution* is to derive daily mobility patterns in terms of various city-wide locations and predict future 12-hour ahead mobility patterns using LSTM model. The proposed mobility models can be utilized to understand and

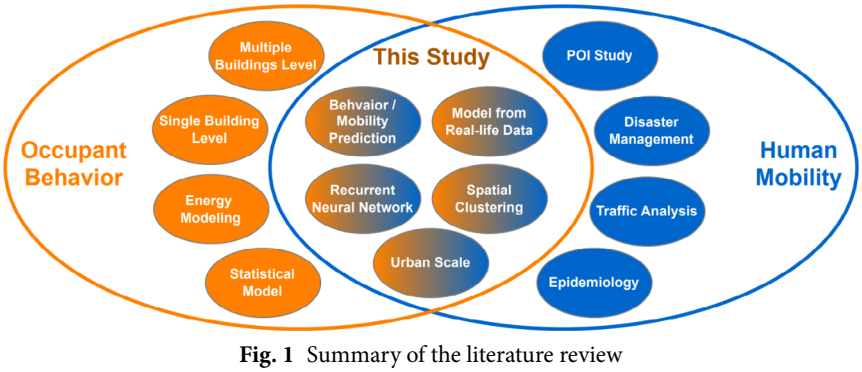

Fig. 1 Summary of the literature review

predict urban scale occupant behavior, assist to develop more realistic building predictive control algorithms and study building energy demand and its flexibility.

3 Method

Urban scale human mobility data has temporal and spatial characteristics, data collected from different users show various mobility patterns as shown in Figure 4 Those characteristics make human mobility data high dimensional and bring challenges to capture and model human mobility behavior at urban scale. The dataset in this study includes GPS records with high accuracy and granularity. However, noises could be introduced under some special circumstances, such as high-profile buildings around the GPS terminal, weak GPS signal. Meanwhile, human mobility data usually have a big size because of the large amounts of users and the fine granularity of data collection. It requires high computational power resources to process human mobility data. This section covers the methods that were adopted to process raw mobility datasets, as well as to model and predict daily human mobility patterns.

3.1 Overview of the method

Figure 2 provides an overview of the approach in this study. It includes (1) data processing, (2) data modeling and (3) performance evaluation. In data processing, we have implemented the "stay point detection" algorithm to extract stay points from the raw datasets with a specified time and distance threshold. DBSCAN algorithm was used to cluster the extracted stay points and identify important places such as home, work, and other locations. Then mobility patterns were constructed for each user represented by important places. Those patterns were later fed into the LSTM model for training and testing. Based on the trained LSTM models, sequence-to-sequence prediction was conducted to predict future human mobility patterns. At the end, we evaluated the performance of the developed LSTM models. In the case study for a typical user, confusion matrix was presented to evaluate prediction results, parameters such as precision and accuracy were calculated to quantify the model performance. The entire process was implemented on a Linux machine which has an AMD Ryzen 9 3950X 16-Core Processor, two NVIDIA Quadro RTX 5000 GPUs with 16Gb memory each, and 115 Gb system memory. The model training and testing process took about four days to complete.

3.2 Data preprocessing

3.2.1 *Data set description*

Dataset used in this study were collected from anonymized mobile phone users in the state of Arizona ranges from October 1st to December 31st in 2016. Raw data sets are in compressed CSV format with a total size of 307 GB. Table 1 listed all the variables from this dataset. The raw data includes a Unix format timestamp in Coordinated Universal Time (UTC) format, latitude, longitude, altitude, and horizontal accuracy of GPS records. Latitudes and longitudes are stored in decimal values as degrees, altitudes and horizontal accuracy are stored as decimal values in meters. Each row of records is associated with an anonymized unique ID which represents the device where data collected from. The temporal resolution of raw data is in seconds, and the spatial resolution represented by latitudes and longitudes in degrees with seven decimal places. The level

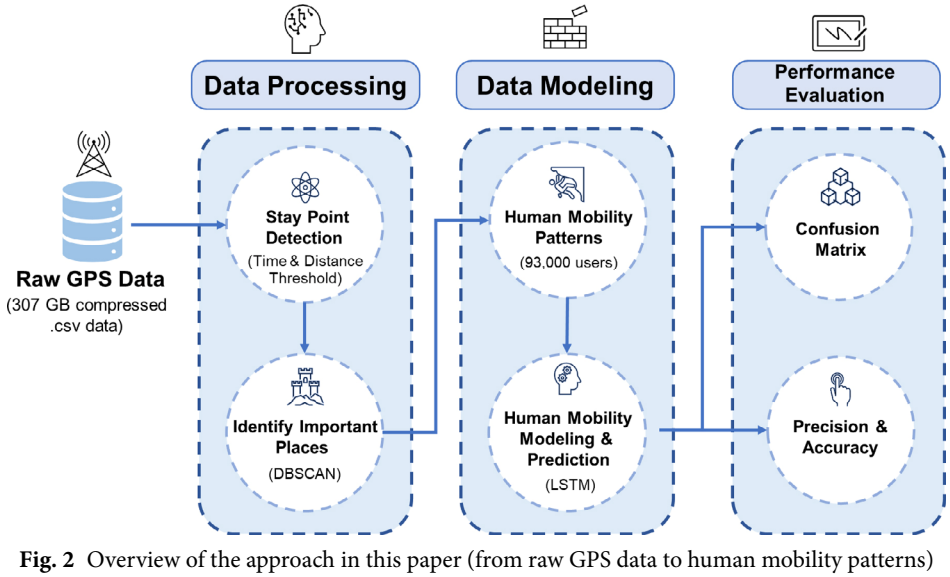

Fig. 2 Overview of the approach in this paper (from raw GPS data to human mobility patterns)

Table 1 List of measuremetns in the raw data

Name	Description
Timestamp	UTC timestamp of the GPS record in Unix format
Unique ID	Anonymized unique ID to represent the device
Latitude	Decimal value of the latitude
Longitude	Decimal value of the longitude
Altitude	Decimal value of the altitude
Horizontal Accuracy	Horizontal accuracy of the GPS record in meters

of horizontal accuracy can be as precise as 10 meters, but it differs between various device types (Pepe et al. 2020). Our stay point detection algorithm, discussed in the Section 3.2.3, can eliminate noise data points using a distance threshold (250 meters) and time threshold (10 minutes). This will also minimize the impact on the DBSCAN algorithm (ε = 250 meters) caused by location accuracy. Similar raw data from the same data source have been used to extract building occupancy profiles (Dong et al. 2019; Wu et al. 2020), investigate urban mobility and accessibility (Akhavan et al. 2019), and study commuting and travel patterns (Sadeghinasr et al. 2019).

3.2.2 *Select user of interest from raw data*

Figure 3 provides an overview of the raw data from 12 PM to 1 PM on Monday of October 31, 2016. The figure shows most data points are gathered around the Phoenix area. We have selected the raw data from Arizona State focusing on Phoenix Metropolitan Area. We have pre-selected a zip code area with the largest percentage of utility customers which adopted PV (Photovoltaics) and battery storage system. This area will be the focus of future building-to-grid research, which is currently not in the scope of this work. Then, the authors implemented spatial join between the raw GPS data and selected zip code area, as a result, 93,000 users were obtained from the raw data set for a case study, size of the selected data is around 19.2 GB in parquet format. Parquet is an open source, column-oriented data storage format which is designed for efficient data storage and retrieval. Because of the nature of big data, this raw data selection process has many challenges such as insufficient system memory, low processing speed, long file loading time. We have taken advantage of the Python Dask Library to load the raw data in chunks and improve the processing speed.

Figure 4 shows 3D plots of the raw data from four different users. In the figure, x-axis and y-axis represent latitude and longitude respectively, z-axis is the time of the day from 0 to 24 which also represented by a color scheme from blue to red. The morning and afternoon commute routes can be clearly observed. As shown in the figure, users stayed around the same location (e.g., home location) in the early morning and commuted to a different location (e.g. work location), then stayed at that location for several hours. In the evening, users started to commute back to the same location as the early morning, and stayed there for the rest of the night. Those trajectories clearly showed human mobility patterns at different times of a day. This study focuses on recognizing where users stayed in the day, and extracting those stay locations.

3.2.3 *Stay point detection*

This study focuses on understanding how users move from one location to another at urban scale. Therefore, we have

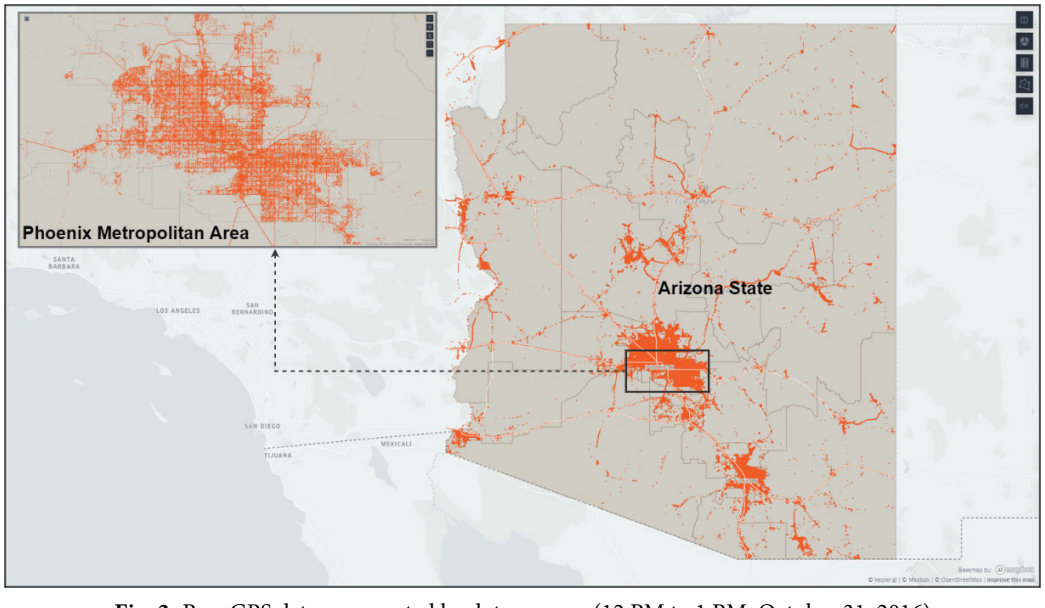

Fig. 3 Raw GPS data represented by dots on map (12 PM to 1 PM, October 31, 2016)

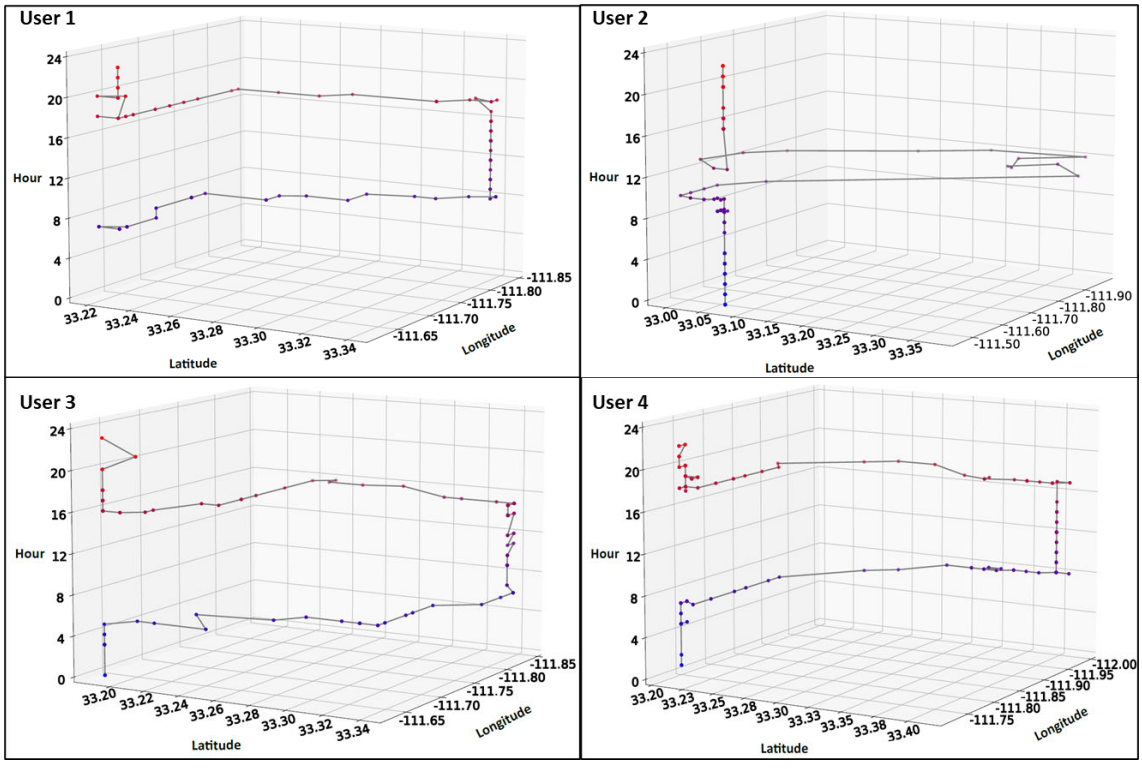

Fig. 4 Trajectory of sample users represented by GPS raw data at different times of the day

constructed trajectories for each user based on the raw data. Raw data points (P_1, P_2, P_3, …, P_n) include timestamp, longitude, and latitude for a sample user. As discussed in the beginning of Section 3, noised can be introduced under various unusual circumstances. Meanwhile, since this study focused on identifying locations where users stayed, data points collected while the user was traveling or commuting can be treated as noises. Hence, we implemented the "Stay Point Detection" algorithm, which has been broadly adopted in literature (Damiani et al. 2014; Jiang et al. 2016; Khoroshevsky and Lerner 2017; Suzuki et al. 2019). Figure 5 shows a complete sample daily trajectory of a user, the green circles are raw data points represented by datetime, latitude and longitude. The gray dashed circle shows a group

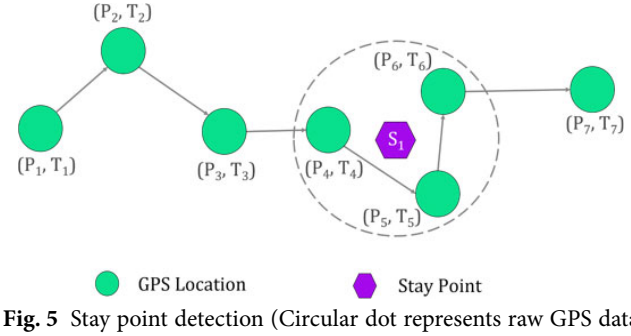

Fig. 5 Stay point detection (Circular dot represents raw GPS data point, hexagon represents stay point. Dashed circles represent time threshold 10 minutes, and distance threshold 250 meters)

of data points within the distance threshold (250 meters) and time threshold (10 minutes), those thresholds were adopted from literature (Jiang et al. 2016; Barbour et al. 2019). Within the dashed circle, a stay point represented by a purple polygon was extracted to represent one of the user's stay locations.

Each stay point represents where the user stayed within a distance threshold and time threshold, it has a longitude, latitude, arrival time and departure time. As shown in the Eq. (1), points P_i, P_{i+1}, …, P_{i+n} represent a group of data points within the specified thresholds. The datetime of the first data point (P_i) and the last data point (P_{i+n}) in this group represents the arrival and departure time for this stay point (S_j). Stay duration (T_{stay}) is the time difference between the datetime of P_{i+n} and P_i. Longitude and latitude of the stay point are represented by lon_{S_j} and lat_{S_j} that calculated by the average latitude and longitude of points P_i, P_{i+1}, …, P_{i+n}. Further analysis will be conducted to identify important locations where the user frequently stayed among those stay points.

$$
\begin{aligned}
T_{\text{stay}} &= T_{P_{i+n}} - T_{P_i} \\
\text{lat}_{S_j} &= \frac{\sum\left(\text{lat}_{P_i} + \text{lat}_{P_{i+1}} + \ldots + \text{lat}_{P_{i+n}}\right)}{n} \\
\text{lon}_{S_j} &= \frac{\sum\left(\text{lon}_{P_i} + \text{lon}_{P_{i+1}} + \ldots + \text{lon}_{P_{i+n}}\right)}{n}
\end{aligned}
\tag{1}
$$

3.2.4 *Identify stay locations*

Stay point reveals the location where the user stayed at urban scale, then a clustering algorithm was used to cluster the stay points into different location groups. DBSCAN algorithm is a well-known clustering algorithm which is commonly used in machine learning and data mining studies (Schubert et al. 2017; Chen et al. 2018; Liu et al. 2019). Recent studies (Jurdak et al. 2015; Tang et al. 2015; Bonnetain et al. 2021; Liu et al. 2021; Smolak et al. 2021) implemented DBSCAN to identify locations of interest and investigate human mobility at urban scale. As Figure 6 shown, given a set of stay points from a user, DBSCAN can group the stay points that are close to each other within specified distance and include a minimum number of points.

$$N_\varepsilon(x) = \{ y \in D : d(x, y) \leq \varepsilon \} \quad (2)$$

In Eq. (2), $N_\varepsilon(x)$ denotes the neighborhood of x, and $|N_\varepsilon(x)|$ represents the total number of points in the neighborhood of point x. D is a set of stay points extracted from the user's raw GPS data. ε is the radius of the circle around each data point to check the density. $d(x, y)$ is the distance between points x and y from D. $N_{\min}$ specifies the minimum number of points within radius of ε to be considered as a cluster.

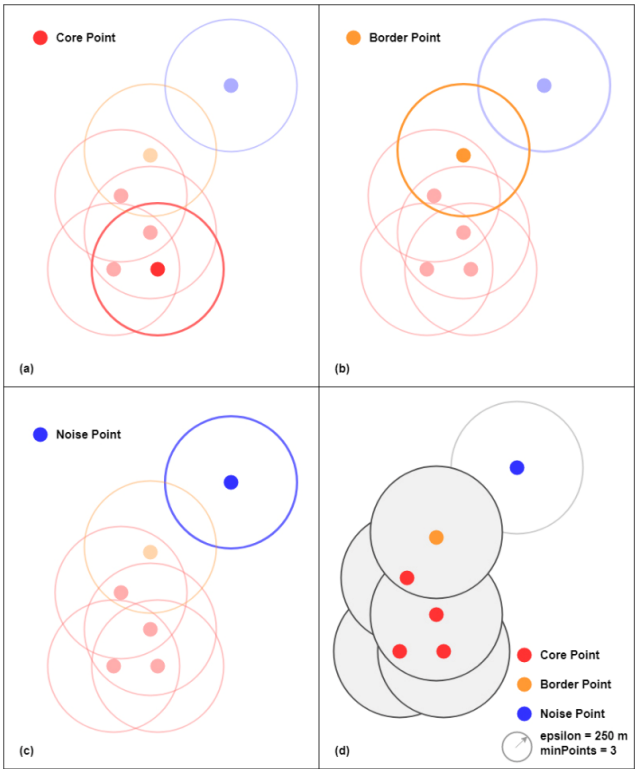

Fig. 6 Illustration of DBSCAN algorithm (ε is 250 meters, minimum three points to be considered as a cluster): (a) core point; (b) border point; (c) noise point; (d) final cluster results

Based on a set of stay points as shown in Figure 6, DBSCAN algorithm works as followings,

- if $x, y \in D$, $|N_\varepsilon(x)| \geq N_{\min}$, $y \in N_\varepsilon(x)$ and $|N_\varepsilon(y)| < N_{\min}$, y is the border point of the neighborhood of point x (Figure 6(b)), and x is the core point (Figure 6(a)).
- However, if $x, y \in D$, $|N_\varepsilon(x)| \geq N_{\min}$, $y \notin N_\varepsilon(x)$ and $|N_\varepsilon(y)| < N_{\min}$, y is a noise point (Figure 6(c)), and x is the core point (Figure 6(a)).
- Core points and its border points will form a cluster as the shaded area shown in Figure 6(d), and noise points will be put into the noise cluster.

Figure 7 shows the sample results after implementing the DBSCAN algorithm (ε = 250 meters, $N_{\min}$ = 3), cluster number −1 is the noise cluster that contains all the noise stay points. Cluster 0 to 8 are individual clusters find by the algorithm. As highlighted in the figure, each cluster may include various number of stay points. Considering the distance threshold from the process of stay point detection, ε was also set as 250 meters. $N_{\min}$ was set as 3 to get most clusters which cover different location types, meanwhile given that each stay point was already represent numerous raw data points. Next, in the mobility modeling process, those clusters will be evaluated by arrival time, departure time, and the duration of stays.

3.3 Human mobility modeling

In this study, the human mobility patterns are represented by different types of stay locations such as home, work, and other locations. Since our data were collected from smart mobile phone users, we have assumed that the phone is around the user all the time and each user has only one home location. By implementing DBSCAN algorithm in previous step, user's stay points have been labeled as different cluster numbers. Each cluster may have various density, stay duration, arrival and departure time. To model human mobility, it is necessary to assess those clusters and understand the connections among those stay points. Next, the time windows of home stay, work stay, and other stay were pre-defined as followings:

- $|T_{\text{home}}|$ denotes the time window of home stays from 8 PM to 6 AM the second day;
- $|T_{\text{work}}|$ denotes the time window of work stays from 9 AM to 6 PM the same day;
- Other times will be considered as other stays.

For each user, all the stay points will be visited and labeled as one of those three types of stays based on following judgements:

- If the date time of a stay point (T_{S_j}) is within the time window of home stays ($|T_{\text{home}}|$) and the duration of this stay (T_{stay}) is equal or greater than 1.5 hours, then the stay point (S_j) will be labeled as home location;

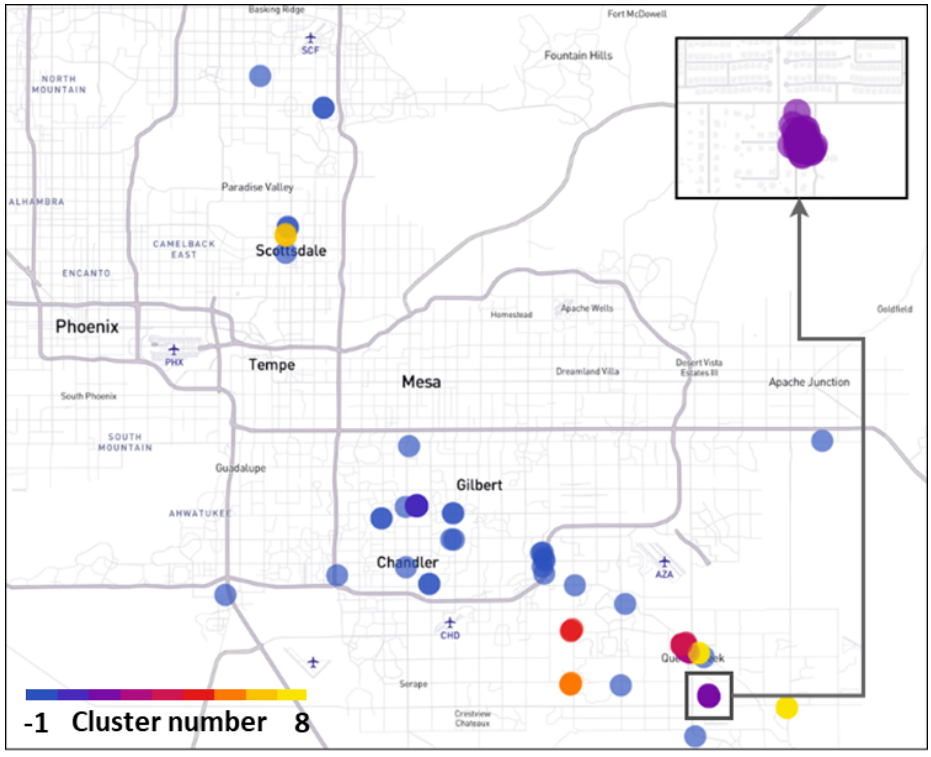

Fig. 7 Example results of the DBSCAN algorithm (Nine clusters were identified including the noise cluster)

- For some rare conditions if the stay duration is equal or greater than 24 hours, the stay point (S_j) will also be labeled as home location;
- However, if the date time of a stay point (T_{S_j}) is within the time window of work stays ($|T_{\text{work}}|$) and the duration of this stay (T_{stay}) is equal or greater than 1.5 hours, then the stay point (S_j) will be labeled as work location;
- Otherwise, the stay point (S_j) will be labeled as other locations.

Afterwards, the cluster with dominant home labels will be considered as home type of cluster, and all the stay points within the cluster will be labeled as home locations. Similarly, the cluster with dominant work labels will be labeled as work type of cluster, and all the stay points within the cluster will be labeled as work locations. Otherwise, the stay points will be considered as other locations within the other type of cluster. The assumption was made that each user has only one home location, but could have multiple work and other locations. For the home cluster, a home location can be extracted as the centroid of this cluster. For the work clusters, location of work stays can be calculated as the centroid of the work clusters. Other locations can be identified as the centroid of the clusters that included other locations. In the end, for each user, a home location was extracted, a ranking list of work locations, and a ranking list of other locations were constructed respectively together with the probabilities of those locations appeared in the data. This could be further used to interpret human mobility patterns to exact locations.

Figure 8 visualized the results of the above human mobility modeling process. All circles stand for stay points, orange colored stay points belong to home type of cluster, blue colored stay points represent work type of cluster, green colored stay points were labeled as other type of cluster. The radius of each circle stands for length of stay duration, the larger the circle and the longer the stay duration. As a result of this modeling process, all the stay points have been labeled either home, work, or other type of locations.

Next, daily human mobility patterns can be constructed based on the various location types of stay points. The pattern includes one data point for every hour of the day. As shown in Figure 9, every daily mobility pattern has 24 stay points represented by H (home), W (work), or O (other) locations. Those mobility patterns are then grouped by day of week for further analysis.

3.4 Model training and prediction

Recurrent Neural Network (RNN) has been broadly used for time series prediction, and sequence-to-sequence prediction as well (Marino et al. 2016; Rahman et al. 2018; Fan et al. 2019; Kim et al. 2021; Mughees et al. 2021). Studies have used RNN for next location prediction or travel destination prediction to understand human mobility at

urban scale (Liu et al. 2016; Huang 2017; Khoroshevsky and Lerner 2017; Feng et al. 2018; Wang et al. 2019a; Guo et al. 2020; Yang et al. 2020). However, studies for modeling and prediction of daily human mobility patterns for a whole day with constant prediction horizon are still very limited. This study adopted a variant of RNN modules named LSTM to learn and predict daily human mobility patterns at urban scale. Figure 10 shows the overall structure of this LSTM neural network model. It works as following: given stay points S_1, S_2, ..., S_j as inputs (represented by location types) to the model, LSTM cell generates the hidden state and provides an output based on the input; both the hidden state and output will be passed to the next LSTM cell as inputs to generate new hidden state and next output; this process continues until a desired length of pattern is generated. In summary, the LSTM model utilized for the prediction task receives sequences of stay points represented by location types as inputs and outputs, as demonstrated in Figure 10. For the prediction task, the model takes in a sequence of 12-hour data points and generates one data point at each step until the output length reaches 12, effectively completing a pattern of a full day spanning 24 hours.

Based on the processed data from 93,000 users in Phoenix

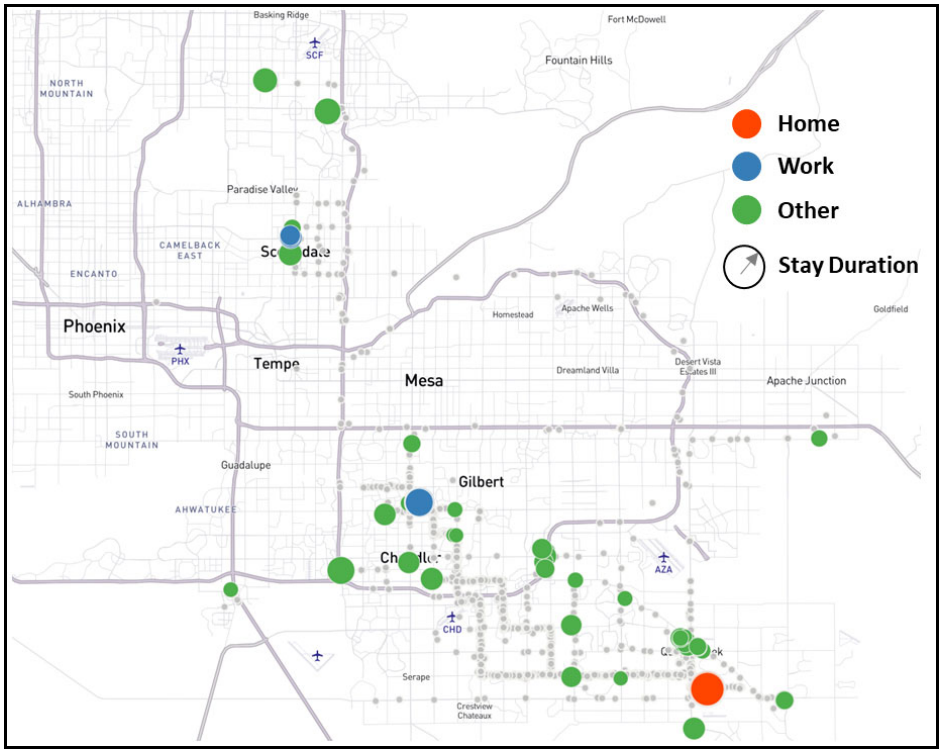

Fig. 8 Location types of different clusters of stay points (home, work, and other clusters)

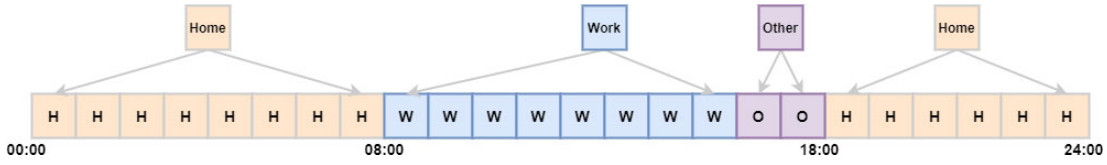

Fig. 9 Sample of daily human mobility patterns (represented by hourly location type)

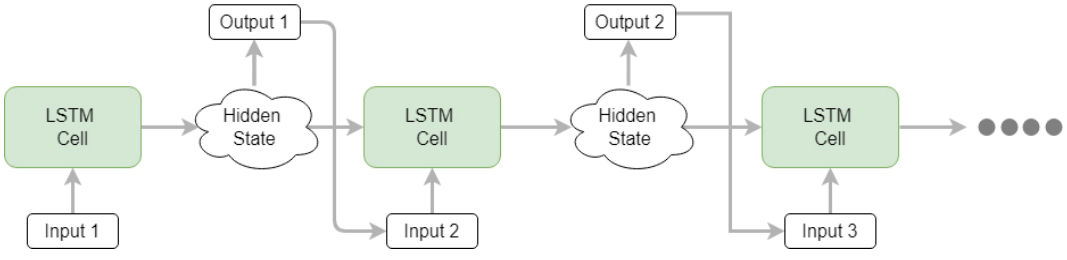

Fig. 10 LSTM sequence to sequence model

Metropolitan Area, daily human mobility patterns were constructed for each user. However, those categorical data represented by location patterns cannot directly be fed into the LSTM model. We have adopted the one-hot encoding to convert the input categorical data (home, work, and other) to a binary vector with values of 1 and 0. Since the categorical value in this study is relatively simple (only three different location types), one-hot encoding fits well with the experiment settings. Shadow neural networks were used to construct the LSTM models. The hidden dimension size was set as five, training epochs were 1000 with learning rate scheduler started at 0.1. Finally, an LSTM model was developed for each user, with 80% of the user's data for model training and 20% of the user's data for model testing. This study focuses on understanding the dynamics of human mobility patterns over the course of a day. In this study, the input sequences and target sequences were designed to cover a duration of 23 hours of a day while training the model. This duration was selected to capture a comprehensive view of the movements and transitions of individuals throughout most of their day. During training, the input sequence represents a 23-hour window of human mobility patterns, which is used to predict the target sequence, representing the next step in the mobility pattern. This approach allows the model to learn and understand the underlying patterns and dependencies within human mobility patterns over a day, which can then be used to predict future mobility patterns. It's important to note that the target sequence is one step ahead of the input sequence, which means that the model is predicting the next step in the mobility pattern based on the input sequence. This ensures that the model is trained to predict future mobility patterns, rather than simply memorizing, and reproducing the input sequence. In the model validation process, the user's daily mobility pattern was predicted based on the first 12 hours' sequence data. This study has developed an LSTM model for each user to capture and model the complex dynamics of urban human mobility. While the models share the same structure, they were trained and tested using different datasets. The experiment was conducted using the hardware described in Section 3.1, and the results showed that each user takes approximately eight seconds to complete both the training and testing process. The following section presents detailed results of this LSTM model training and testing process.

4 Results and discussions

4.1 Results

This section presents the results of data pre-processing, analyzing, and modeling of the selected 93,000 users in Phoenix Metropolitan Area, Arizona. The data covers three months from October 1, 2016, to December 31, 2016. Figure 11 shows the raw data of a typical user plotted on the map, the color scheme represents different hours of the day. Overlap between data points and the street network on the map can be observed, it clearly shows the user's commute routes along the roads. Noted that some spots on map have a higher density of raw data points which indicate potential stay locations (home, work or other). Next, in data processing, stay points were successfully detected from users' raw data with a distance threshold of 250 meters, and a time threshold of 10 minutes. As illustrated in Figure 12, the blue circles represent stay points, and the gray dots refers to the raw data points of the typical user. The following analysis was based on those stay points extracted from raw data. In the above Figure 8, DBSCAN algorithm and the mobility modeling process successfully identified the home, work, and other locations as well as stay durations from the stay points.

Based on the labeled stay points, user's daily mobility patterns were constructed, represented by different location types. Figure 13 shows the daily human mobility patterns in three months represented by different location types. It can be observed that the user commuted between home and work locations most of the time and visited other locations occasionally. The figure also shows that the user stayed at home most of the time. Those patterns indicate that the user's daily human mobility pattern is highly predictable. Afterwards, those daily patterns were grouped together by weekdays and weekends. Then, 80% of the data was used to train the LSTM model, and 20% of the data were used for model testing. Figure 14 visualizes the patterns of training data on weekdays and weekends in 24 hours from the typical user. Most of the patterns started and ended at home location, very few patterns stared or ended at work location on weekdays. This aligns with the high dimensional nature of mobility data, as the user's location varies at different times of the day and the mobility pattern also varies on different days of the week. Testing data were used to validate the trained LSTM model with a 12-hour prediction horizon. Figure 15 shows the results of model testing both on weekdays and weekends. In 24 hours of a day, the data from 0AM to 12PM were treated as inputs into the model to predict the user's mobility of next 12 hours (indicated by green shaded area in the figure). The blue patterns are ground truth, and the orange patterns are prediction results by the trained model. To better illustrate the differences between ground truth and prediction results, the orange patterns were shifted to above of the ground truth. As results show, the trained LSTM model can learn and predict the dynamics of daily human mobility patterns. The following analysis quantified the model performance with evaluation matrix.

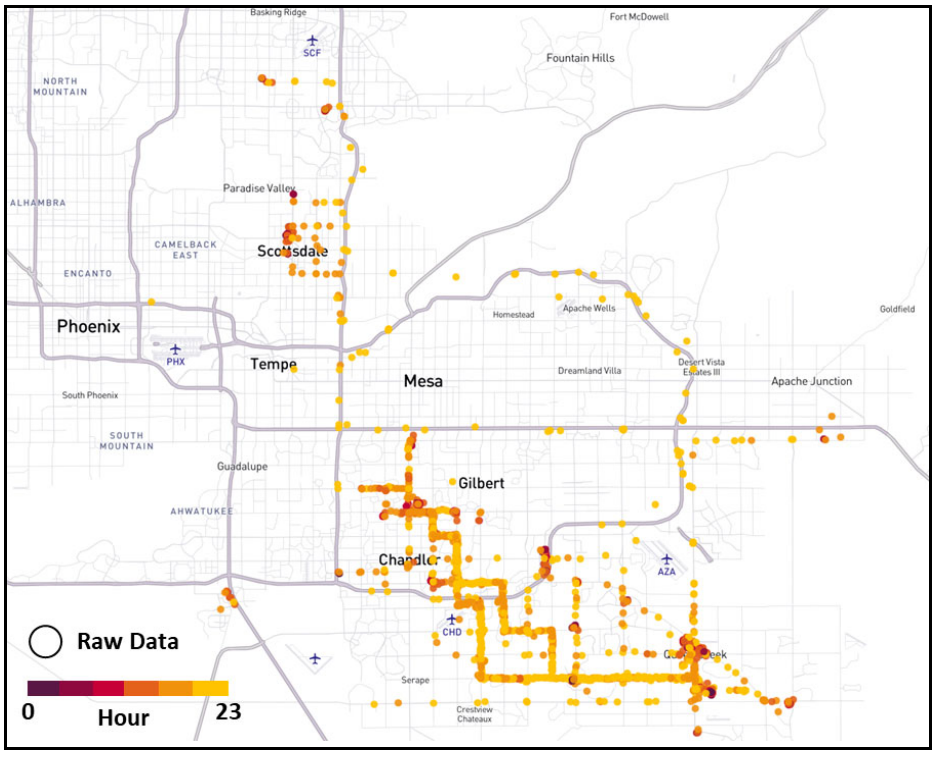

Fig. 11 Raw GPS data on map of a typical user

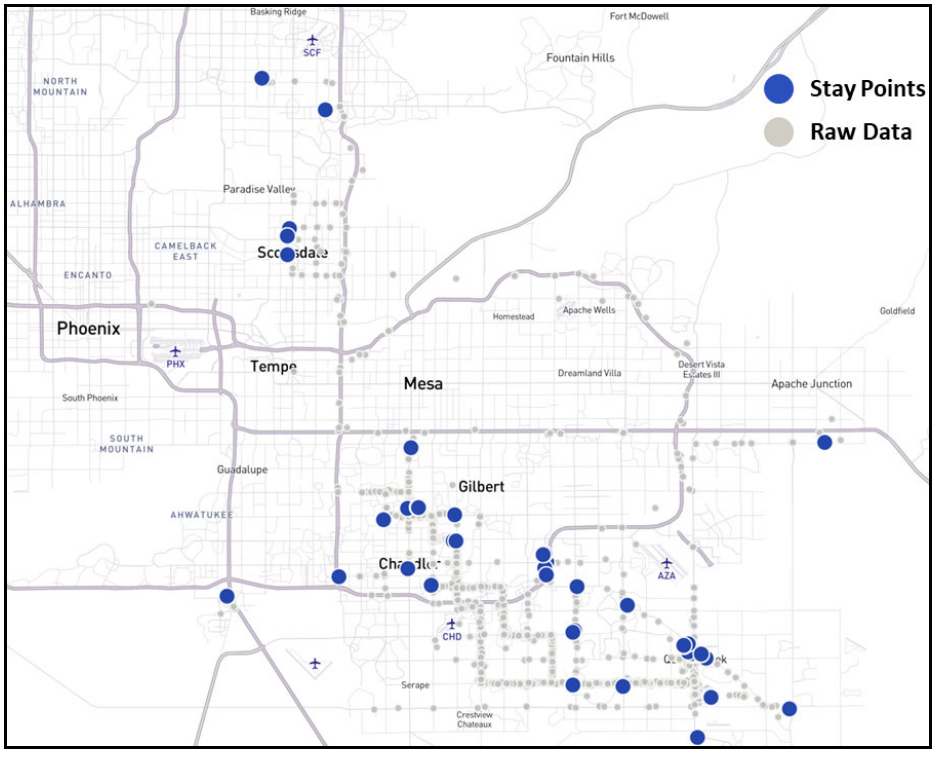

Fig. 12 Stay points on map detected from raw GPS data of a typical user

Figure 16 shows the confusion matrix of the LSTM model for a typical user, it compares the actual values with predicted results for different location types. True positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are commonly used to evaluate the performance of machine learning models on categorical data. In this case study, the location types are categorical data. Take the evaluation of "home" location as an example,

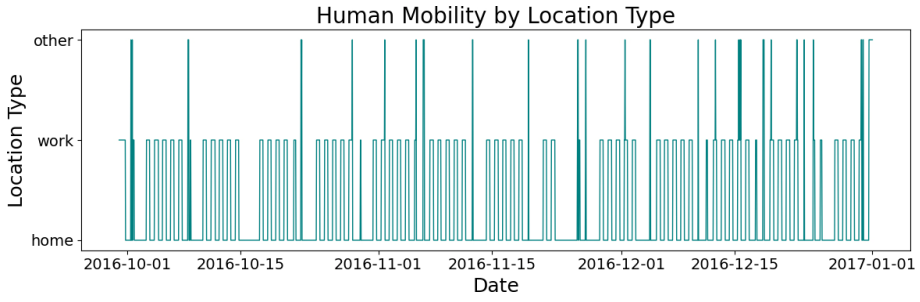

Fig. 13 Human mobility patterns represent by location type (one month data at hourly level)

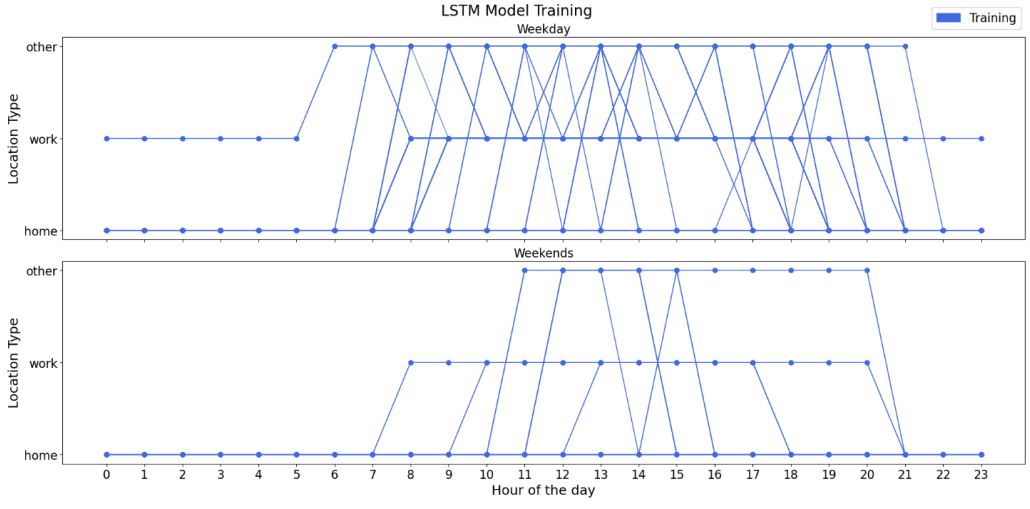

Fig. 14 Sample of LSTM model training data (80% of the data were used for model training)

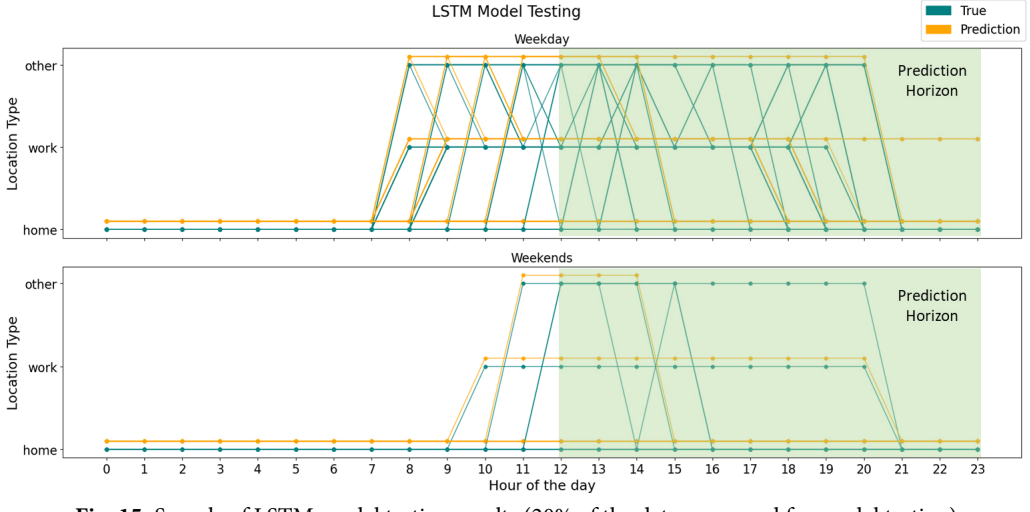

Fig. 15 Sample of LSTM model testing results (20% of the data were used for model testing)

TP means when the true value is home, and the predicted value is also home; FP means when the true value is not home, but the predicted value is home; FN means when the true value is home, but the predicted value is not home; and the rest cases belong to TN. In this figure, the darker cell means more data points fall in that category. The matrix indicates that among the data records of this typical user, home locations accounted for the majority, followed by work and other locations. To evaluate the performance of this LSTM model, both precision and accuracy have been selected. Accuracy relates to how close the prediction value is to the actual value. And precision refers to how close are the predictions to each other. Those metrics have been used in the literature (Li and Huang 2013; Chatterjee et al. 2017; Ghosh and Ghosh 2018) to evaluate the performance of machine learning models. Eq. (3) shows the process

of calculating precision and accuracy.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$
$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (3)$$

Based on the confusion matrix in Figure 16, calculation results show that the LSTM model of this typical user has an overall 91% accuracy, and around 91% weighted average of precision. We then analyzed the accuracy and precision of the model for all tested users. Figure 17 shows the overall accuracy distribution of the LSTM model. The average prediction accuracy of weekdays' data is 85.05%, and 85.26% for data on weekends. Figure 18 shows the overall precision distribution of the LSTM model. The average prediction precisions are 86.03% on weekdays and 86.09% on weekends. Meanwhile, both figures also show the LSTM model predicted 100% accurately and precisely for large amount of mobility patterns. This will be discussed in the following subsection.

4.2 Discussions

Model testing results showed that this study predicted daily human mobility patterns both accurately and precisely, with overall 85% accuracy and 86% precision. As shown in Figure 11 and Figure 12, the total number of stay points is much smaller than the total number of raw data points. Since the focus of this study is understanding the locations where users stayed and its datetime, the raw data points of commuting, or with very short stay duration were excluded when processing. Stay points were extracted from a group of raw data points that were within the specified time threshold and distance threshold. The values of those thresholds were adopted based on published research work (Jiang et al. 2016; Khoroshevsky and Lerner 2017; Barbour et al. 2019; Suzuki et al. 2019) and tested in our dataset.

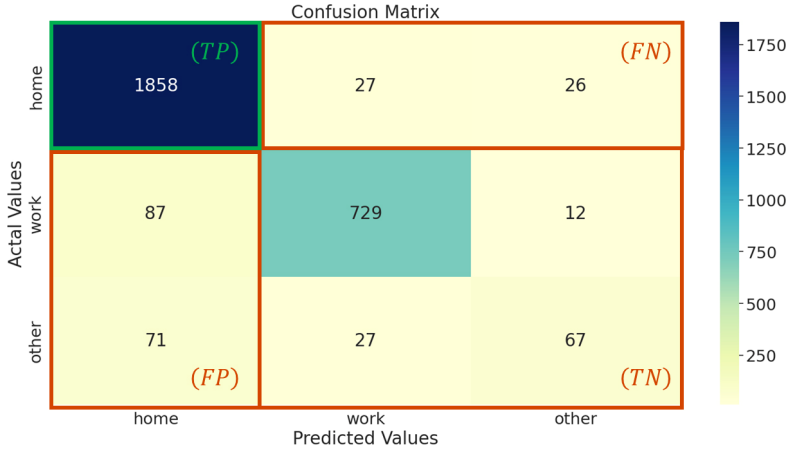

Fig. 16 Confusion matrix of the LSTM model for a typical user

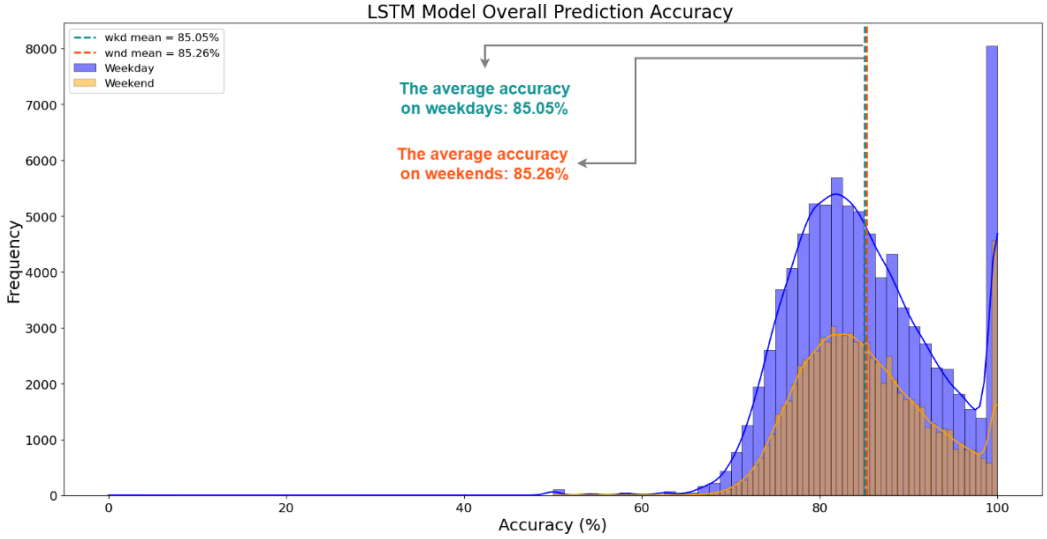

Fig. 17 Overall accuracy distribution of the LSTM model

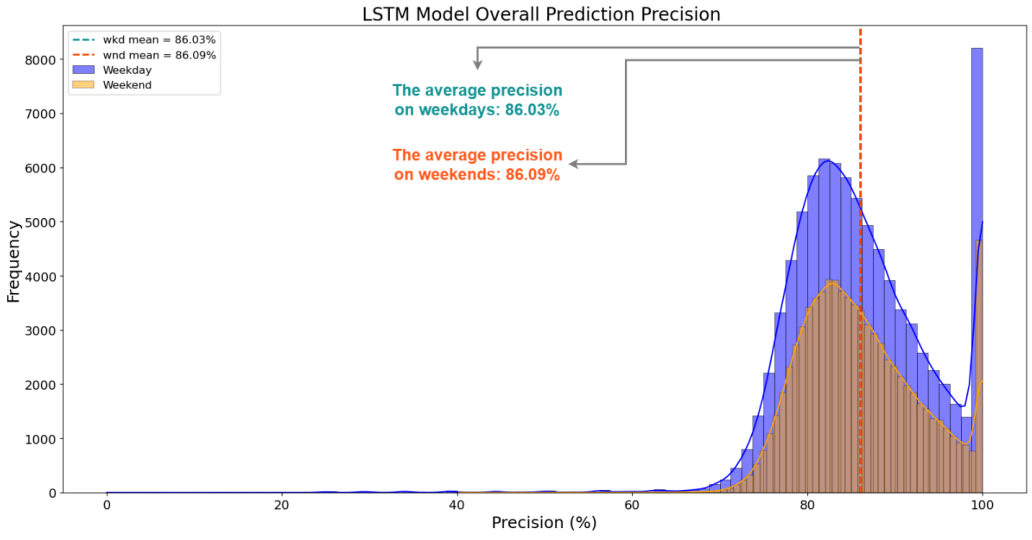

Fig. 18 Overall precision distribution of the LSTM model

Figure 8 clearly showed that the home, work, and other locations were successfully identified from stay points. In this process, the assumption is each user only has one home location, if multiple home locations are detected, the one with the greatest visit frequency will be considered as the default home location. For users with multiple work or other locations, a ranking list of locations with corresponding probabilities was constructed based on users' mobility history. Those two lists can be further used to decode mobility patterns to exact locations.

While the models showed relatively high accuracy and precision as anticipated, this study simplified users' locations into three categories and constructed mobility patterns using the sequence of location types. In Figure 16, we can observe large numbers of correct predictions for both home and work locations from this typical user. However, quite a few false positives and false negatives were observed for other types of locations. This is expected due to the low frequency of other locations that appeared in daily mobility patterns. As shown in Figure 13, the user commuted between home location and work location most of the time. Considering the low frequency together with the stochastic nature of human mobility, the LSTM model cannot capture and predict the patterns of other locations as good as other types of locations. However, some prediction results showed equal or close to 100% accuracy or precision, this is attributed by high regularity mobility patterns of some users. Although the patterns of users may vary, the current structure of the LSTM model only takes into account the output and hidden state from the previous step to generate a single output for the next step. As a result, we believe that the difficulty of the task at each step remains constant using the current model regardless of input data.

As previous studies (Barbour et al. 2019; Wu et al. 2020) showed, more representative occupancy profiles can be derived from human mobility based approach. Compared to the standard reference provided by the U.S. Department of Energy, simulation results of different building types revealed potential heating and cooling energy savings up to 60%. Mobility models in this study contribute to the understandings of human movement patterns at urban scale, which leads to better understanding of daily travel distance, arrival and departure time to home, office, and other locations. Furthermore, it will expand the knowledge of, deriving urban scale building occupancy profiles, predicting city energy demand by communities or districts at different times of the day, planning for building to grid integration and energy flexibility.

5 Conclusions

In this study, we have investigated urban scale human mobility utilizing GPS data collected from smart mobile phones. Three months' raw data of 93,000 users in Phoenix Metropolitan Area were processed to detect users' stay points based on specified time threshold and distance threshold. Built on user's stay points, the DBSCAN clustering algorithm was used to identify different clusters of stay locations. Those clusters were further examined and labeled as home, work, or other locations, by analyzing the arrival time, departure time, and stay duration. On top of that, we have built daily mobility patterns for those users represented by different types of locations. In addition, this study proposed a novel approach to predict urban scale human mobility patterns using a type of recurrent neural network models named LSTM. Shadow neural networks were applied to constructed LSTM models for each user. The models were successfully trained and tested based on daily human mobility patterns.

Testing results of 12-hour ahead prediction show that the model can predict daily human mobility patterns accurately and precisely. The overall accuracy is around 85% and average precision is close to 86%.

Limited studies existed to model and predict human mobility patterns at urban scale. Compared with current studies, the scientific contributions are summarized as follows: (1) It proposed a big data approach based on smart phone GPS data to study urban scale human mobility. This study also conquered the computational challenges caused by large size of raw datasets. (2) Developed human mobility modeling approach and built mobility models for 93,000 users at urban scale in Phoenix Metropolitan Area of Arizona State. (3) Trained and tested LSTM models to predict daily human mobility patterns with 12-hour prediction horizon, resulted in high accuracy and precision. (4) The models that have been developed can be utilized to conduct a more in-depth analysis of occupant behavior at an urban scale, develop more realistic building predictive control algorithms, as well as examine the energy demand of buildings and building energy flexibility.

As an effort to integrate big data analysis with human mobility modeling and prediction, this study has the following limitations: (1) Since this study focused on locations where users stayed at urban scale, raw data points that were collected while commuting or traveling were excluded by the stay point detection algorithm. (2) This study assumed that each user have only one home location, but could have multiple work and other locations. (3) A universal evaluation life schedule was used to examine each stay points and identify home, work, and other locations. This study did not consider the special life schedules like working at night and staying at home during daytime.

Acknowledgements

This work was supported by the U.S. National Science Foundation (Award No. 1949372 and No. 2125775); and in part supported through computational resources provided by Syracuse University.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article. Bing Dong is an Associate Editor of *Building Simulation*.

Ethical approval

This study does not contain any studies with human or animal subjects performed by any of the authors.

Author contribution statement

Bing Dong: conceptualization, methodology, investigation, resources, writing—review, funding acquisition, supervision. Yapan Liu: conceptualization, methodology, investigation, formal analysis, writing—original draft, writing—review & editing, visualization.

References

Akhavan A, Phillips NE, Du J, et al. (2019). Accessibility inequality in Houston. *IEEE Sensors Letters*, 3: 1–4.

Bao J, He T, Ruan S, et al. (2017). Planning bike lanes based on sharing-bikes' trajectories. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, Canada.

Barbour E, Davila CC, Gupta S, et al. (2019). Planning for sustainable cities by estimating building occupancy with mobile phones. *Nature Communications*, 10: 3736.

Biczók G, Díez Martínez S, Jelle T, et al. (2014). Navigating MazeMap: Indoor human mobility, spatio-logical ties and future potential. In: Proceedings of 2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS), Budapest, Hungary.

Bonnetain L, Furno A, El Faouzi NE, et al. (2021). TRANSIT: Fine-grained human mobility trajectory inference at scale with mobile network signaling data. *Transportation Research Part C: Emerging Technologies*, 130: 103257.

Buckee CO, Balsari S, Chan J, et al. (2020). Aggregated mobility data could help fight COVID-19. *Science*, 368: 145–146.

Cerezo Davila C, Reinhart CF, Bemis JL (2016). Modeling Boston: A workflow for the efficient generation and maintenance of urban building energy models from existing geospatial datasets. *Energy*, 117: 237–250.

Cesario E, Comito C, Talia D (2013). Towards a cloud-based framework for urban computing, the trajectory analysis case. In: Proceedings of 2013 International Conference on Cloud and Green Computing, Karlsruhe, Germany.

Chang S, Pierson E, Koh PW, et al. (2021a). Mobility network models of COVID-19 explain inequities and inform reopening. *Nature*, 589: 82–87.

Chang S, Wilson ML, Lewis B, et al. (2021b). Supporting COVID-19 policy response with large-scale mobility-based modeling. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining.

Chatterjee S, Sarkar S, Hore S, et al. (2017). Particle swarm optimization trained neural network for structural failure prediction of multistoried RC buildings. *Neural Computing and Applications*, 28: 2005–2016.

Chen P-T, Chen F, Qian Z (2014). Road traffic congestion monitoring in social media with hinge-loss Markov random fields. In: Proceedings of 2014 IEEE International Conference on Data Mining, Shenzhen, China.

Chen B, Liu Y, Shi W (2018). Vehicle personnel identification model based on optimized ST-DBSCAN algorithm. In: Proceedings of 2018 IEEE 4th Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, China.

Chen Y, Wang Q, Ji W (2020). Rapid assessment of disaster impacts on highways using social media. *Journal of Management in Engineering*, 36(5): 04020068.

Damiani ML, Issa H, Cagnacci F (2014). Extracting stay regions with uncertain boundaries from GPS trajectories: A case study in animal ecology. In: Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Dallas, TX, USA.

DOE (2015). Chapter 5: Increasing Efficiency of Building Systems and Technologies. In: An Assessment of Energy Technologies and Research Opportunities. U.S. Department of Energy.

DOE (2022). Grid-Interactive Efficient Buildings. Available at https://www.energy.gov/eere/buildings/grid-interactive-efficient-buildings.

Dong B, Wu W, Wang Q, et al. (2019). Derive urban scale occupant behavior profiles from mobile position data: A Pilot Study. In: Proceedings of the 16th International IBPSA Building Simulation Conference, Rome, Italy.

Dong B, Liu Y, Fontenot H, et al. (2021). Occupant behavior modeling methods for resilient building design, operation and policy at urban scale: A review. *Applied Energy*, 293: 116856.

Dong B, Liu Y, Mu W, et al. (2022). A global building occupant behavior database. *Scientific Data*, 9: 369

Fan C, Wang J, Gang W, et al. (2019). Assessment of deep recurrent neural network-based strategies for short-term building energy predictions. *Applied Energy*, 236: 700–710.

Feng J, Li Y, Zhang C, et al. (2018). DeepMove: Predicting human mobility with attentional recurrent networks. In: Proceedings of the 2018 World Wide Web Conference.

Fonseca JA, Schlueter A (2015). Integrated model for characterization of spatiotemporal building energy consumption patterns in neighborhoods and city districts. *Applied Energy*, 142: 247–265.

Ghosh SK, Ghosh S (2018). Modeling individual's movement patterns to infer next location from sparse trajectory traces. In: Proceedings of 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Miyazaki, Japan.

Gozzi N, Tizzoni M, Chinazzi M, et al. (2021). Estimating the effect of social inequalities on the mitigation of COVID-19 across communities in Santiago de Chile. *Nature Communications*, 12: 2429.

Guo Q, Sun Z, Zhang J, et al. (2020). An attentional recurrent neural network for personalized next location recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34: 83–90.

Happle G, Fonseca JA, Schlueter A (2018). A review on occupant behavior in urban building energy models. *Energy and Buildings*, 174: 276–292.

Happle G, Fonseca JA, Schlueter A (2020). Context-specific urban occupancy modeling using location-based services data. *Building and Environment*, 175: 106803.

Huang Q, Wong DWS (2015). Modeling and visualizing regular human mobility patterns with uncertainty: An example using twitter data. *Annals of the Association of American Geographers*, 105: 1179–1197.

Huang Q (2017). Mining online footprints to predict user's next location. *International Journal of Geographical Information Science*, 31: 523–541.

Itoh M, Yokoyama D, Toyoda M, et al. (2014). Visual fusion of mega-city big data: An application to traffic and tweets data analysis of Metro passengers. In: Proceedings of 2014 IEEE International Conference on Big Data.

Jensen SØ, Marszal-Pomianowska A, Lollini R, et al. (2017). IEA EBC annex 67 energy flexible buildings. *Energy and Buildings*, 155: 25–34.

Jiang S, Yang Y, Gupta S, et al. (2016). The TimeGeo modeling framework for urban mobility without travel surveys. *Proceedings of the National Academy of Sciences of the United States of America*, 113: E5370–E5378.

Jiang J, Lin F, Fan J, et al. (2019). A destination prediction network based on spatiotemporal data for bike-sharing. *Complexity*, 2019: e7643905.

Jurdak R, Zhao K, Liu J, et al. (2015). Understanding human mobility from twitter. *PLoS One*, 10: e0131469.

Kang X, Yan D, Sun H, et al. (2019). An approach for obtaining and extracting occupancy patterns in buildings based on mobile positioning data. In: Proceedings of the 16th International IBPSA Building Simulation Conference, Rome, Italy.

Kang X, Yan D, An J, et al. (2021). Typical weekly occupancy profiles in non-residential buildings based on mobile positioning data. *Energy and Buildings*, 250: 111264.

Khoroshevsky F, Lerner B (2017). Human mobility-pattern discovery and next-place prediction from GPS data. In: Schwenker F, Scherer S (Eds), Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction. Cham, Switzerland: Springer International Publishing. pp. 24–35.

Kim CH, Kim M, Song Y (2021). Sequence-to-sequence deep learning model for building energy consumption prediction with dynamic simulation modeling. *Journal of Building Engineering*, 43: 102577.

Li Z, Huang G (2013). Re-evaluation of building cooling load prediction models for use in humid subtropical area. *Energy and Buildings*, 62: 442–449.

Liu Q, Wu S, Wang L, et al. (2016). Predicting the next location: a recurrent model with spatial and temporal contexts. In: Proceedings of the AAAI Conference on Artificial Intelligence.

Lin L, Li J, Chen F, et al. (2018). Road traffic speed prediction: A probabilistic model fusing multi-source data. *IEEE Transactions on Knowledge and Data Engineering*, 30: 1310–1323.

Liu X, Huang Q, Gao S (2019). Exploring the uncertainty of activity zone detection using digital footprints with multi-scaled DBSCAN. *International Journal of Geographical Information Science*, 33: 1196–1223.

Liu Y, Singleton A, Arribas-bel D, et al. (2021). Identifying and understanding road-constrained areas of interest (AOIs) through spatiotemporal taxi GPS data: A case study in New York City. *Computers, Environment and Urban Systems*, 86: 101592.

Liu J, Tian J, Lyu W, et al. (2022). The impact of COVID-19 on reducing carbon emissions: From the angle of international student mobility. *Applied Energy*, 317: 119136.

Lu X, Feng F, Pang Z, et al. (2021). Extracting typical occupancy schedules from social media (TOSSM) and its integration with building energy modeling. *Building Simulation*, 14: 25–41.

Marino DL, Amarasinghe K, Manic M (2016). Building energy load forecasting using Deep Neural Networks. In: Proceedings of IECON 2016—42nd Annual Conference of the IEEE Industrial Electronics Society, Florence, Italy.

Mohammadi N, Taylor JE (2017). Urban energy flux: Spatiotemporal fluctuations of building energy consumption and human mobility-driven prediction. *Applied Energy*, 195: 810–818.

Mughees N, Mohsin SA, Mughees A, et al. (2021). Deep sequence to sequence Bi-LSTM neural networks for day-ahead peak load forecasting. *Expert Systems with Applications*, 175: 114844.

O'Brien W, Wagner A, Schweiker M, et al. (2020). Introducing IEA EBC annex 79: Key challenges and opportunities in the field of occupant-centric building design and operation. *Building and Environment*, 178: 106738.

Pan B, Zheng Y, Wilkie D, et al. (2013). Crowd sensing of traffic anomalies based on human mobility and social media. In: Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Orlando, FL, USA.

Pepe E, Bajardi P, Gauvin L, et al. (2020). COVID-19 outbreak response, a dataset to assess mobility changes in Italy following national lockdown. *Scientific Data*, 7: 230.

Rahman A, Srikumar V, Smith AD (2018). Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks. *Applied Energy*, 212: 372–385.

Reinhart CF, Cerezo Davila C (2016). Urban building energy modeling—A review of a nascent field. *Building and Environment*, 97: 196–202.

Sadeghinasr B, Akhavan A, Wang Q (2019). Estimating commuting patterns from high resolution phone GPS Data. arXiv:1907.03744

Salim FD, Dong B, Ouf M, et al. (2020). Modelling urban-scale occupant behaviour, mobility, and energy in buildings: A survey. *Building and Environment*, 183: 106964.

Schubert E, Sander J, Ester M, et al. (2017). DBSCAN revisited, revisited: Why and how You should (still) use DBSCAN. *ACM Transactions on Database Systems*, 42: 19.

Schulte-Fischedick M, Shan Y, Hubacek K (2021). Implications of COVID-19 lockdowns on surface passenger mobility and related CO_2 emission changes in Europe. *Applied Energy*, 300: 117396.

Smolak K, Siła-Nowicka K, Delvenne JC, et al. (2021). The impact of human mobility data scales and processing on movement predictability. *Scientific Reports*, 11: 15177.

Suzuki J, Suhara Y, Toda H, et al. (2019). Personalized visited-POI assignment to individual raw GPS trajectories. *ACM Transactions on Spatial Algorithms and Systems*, 5: 16.

Tang J, Liu F, Wang Y, et al. (2015). Uncovering urban human mobility from large scale taxi GPS data. *Physica A: Statistical Mechanics and Its Applications*, 438: 140–153.

Tang B, Jiang C, He H, et al. (2016). Probabilistic human mobility model in indoor environment. In: Proceedings of 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada.

Trivedi A, Silverstein K, Strubell E, et al. (2021). WiFiMod: Transformer-based indoor human mobility modeling using passive sensing. In: Proceedings of ACM SIGCAS Conference on Computing and Sustainable Societies.

Wang Q, Taylor JE (2016). Patterns and limitations of urban human mobility resilience under the influence of multiple types of natural disaster. *PLoS One*, 11: e0147299.

Wang C, Li R, Zhao Z, et al. (2019a). Statistics-enhanced destination prediction model for multi-users based on deep learning. In: Proceedings of 2019 IEEE 19th International Conference on Communication Technology (ICCT), Xi'an, China.

Wang J, Kong X, Xia F, et al. (2019b). Urban human mobility: Data-driven modeling and prediction. *ACM SIGKDD Explorations Newsletter*, 21(1): 1–19.

Wilson R, Erbach-Schoenberg EZ, Albert M, et al. (2016). Rapid and near real-time assessments of population displacement using mobile phone data following disasters: the 2015 Nepal earthquake. *PLoS Currents*, 8. https://doi.org/10.1371/currents.dis.d073fbece328e4c39087bc086d694b5c.

Wu W, Dong B, Wang Q, et al. (2020). A novel mobility-based approach to derive urban-scale building occupant profiles and analyze impacts on building energy consumption. *Applied Energy*, 278: 115656.

Yan D, O'Brien W, Hong T, et al. (2015). Occupant behavior modeling for building performance simulation: current state and future challenges. *Energy and Buildings*, 107: 264–278.

Yan D, Hong T, Dong B, et al. (2017). IEA EBC Annex 66: Definition and simulation of occupant behavior in buildings. *Energy and Buildings*, 156: 258–270.

Yang D, Fankhauser B, Rosso P, et al. (2020). Location prediction over sparse user mobility traces using RNNs: Flashback in hidden states! In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, Yokohama, Japan.

Yang X, Zhuge C, Shao C, et al. (2022). Characterizing mobility patterns of private electric vehicle users with trajectory data. *Applied Energy*, 321: 119417.