Chapter 3
Uncovering the Relationships Between Phone Communication Activities and Spatiotemporal Distribution of Mobile Phone Users

Yang Xu, Shih-Lung Shaw, Feng Lu, Jie Chen and Qingquan Li

3.1 Introduction

The pulses of our cities are largely driven by human activities and their movements. An improved understanding of where people are in space and time would benefit urban and transport planning, and facilitate academic research in a wide range of disciplines (e.g., geography, epidemiology, and economics). Traditionally, our abilities to capture spatial and temporal patterns of population distributions largely rely on census data. Despite of their usefulness in population studies, the collection

Y. Xu
Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Kowloon, Hong Kong
e-mail: yang.ls.xu@polyu.edu.hk

S.-L. Shaw (✉)
Department of Geography, University of Tennessee, Knoxville, TN 37996, USA
e-mail: sshaw@utk.edu

S.-L. Shaw
Guangzhou Institute of Geography, 100 Xianlie Zhong Road, Guangzhou, Guangdong 510070, People's Republic of China

F. Lu · J. Chen
State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, People's Republic of China
e-mail: luf@lreis.ac.cn

J. Chen
e-mail: chenj@lreis.ac.cn

Q. Li
Shenzhen Key Laboratory of Spatial Smart Sensing and Services, Shenzhen University, Shenzhen 518060, People's Republic of China
e-mail: liqq@szu.edu.cn

S.-L. Shaw and D. Sui (eds.), *Human Dynamics Research in Smart and Connected Communities*, Human Dynamics in Smart Cities,
https://doi.org/10.1007/978-3-319-73247-3_3

of census data is costly and time consuming. Moreover, such data provide a static view of population estimates, with update cycles that are relatively long (e.g., 10 years). These issues limit the usability of census data in many application domains, especially the ones (e.g., traffic management, disaster response, and epidemic control) that require timely and spatially detailed population information. Although the ways of estimating population distributions have been enhanced in the past few decades (Dobson et al. 2000; Harvey 2002a, b; Balk 2004; Bhaduri et al. 2007; Stevens et al. 2015), we are still in need of cost-effective ways to capture the whereabouts of people in space and time, which are highly dynamic in its nature.

In recent years, mobile phone data have received much attention in geography and other fields. Several advantages make mobile phone data a valuable resource for studying population dynamics: (1) a high and growing penetration rate of mobile phones around the world[1], (2) various location-aware technologies used in mobile phone positioning (Birenboim and Shoval 2015), and (3) ease of data collection (e.g., little burden on individual participants). Two types of mobile phone data, Erlang and call detail records (CDRs), have been widely used in existing literature to study population distributions (Ahas et al. 2007; Girardin et al. 2009; Reades et al. 2009; Sevtsuk and Ratti 2010). These studies regard phone communication activities as an indicator of the presence of urban population. However, these mobile phone data reveal partial aspects of population dynamics, given the fact that Erlang measures aggregate call volume at cellphone towers, and CDRs are generated during particular types of cellphone activities (i.e., initiating or receiving a phone call/text message). It means many previous studies implicitly assume that phone communication activities could properly reflect the distribution of urban population. Nevertheless, whether this assumption holds has not been investigated. Moreover, few studies have even examined whether phone communication activities could reflect the spatiotemporal distribution of mobile phone users.

To fill the research gap, this study uses a mobile phone data set collected in Shanghai, China to answer an important research question: to what extent could phone communication activities reflect the spatiotemporal distribution of mobile phone users? The mobile phone data set used in this study consists of CDRs plus other cellphone-related logs such as cellular handover and periodic location update. To answer the research question, we extract all CDRs into a separate data set to capture the intensity of mobile phone communications at different places in the city over time. Meanwhile, the complete data set is used to derive the spatiotemporal distribution of mobile phone users. Then, correlation and regression analyses are performed to evaluate the relationships between the two types of distributions. The research findings could reveal the potential bias of using phone communication intensity to reflect the underlying population distribution, and provide useful

[1]According to the International Telecommunication Union (ITU 2015), there are more than 7 billion mobile phone subscriptions by the end of 2015, corresponding to a penetration rate of 97%. The penetration rate in developed countries reaches 121% by the end of 2014 (World Telecommunication Development Conference 2014).

information and guidelines of using large-scale mobile phone data in urban dynamics research.

3.2 Related Work

The advent of mobile phones has changed how people interact with the outside world (Schwanen and Kwan 2008). It also transforms the ways human activities are sensed and understood. Mobile phone location data, which suggest locations visited by people, have been used to better understand different aspects of human dynamics. For example, there have been many studies which use Erlang data to examine the rhythms of urban mobility patterns (Ratti et al. 2006; Reades et al. 2009; Sevtsuk and Ratti 2010). In these studies, the intensity of people's phone communication activities is used as an indicator of the presence of urban population. Similarly, call detail records (CDRs) have been used to uncover collective human activity patterns (Candia et al. 2008) and aggregate population movements (Ahas et al. 2007, 2010a). Although it is reasonable to assume a certain degree of correlation between the cellphone usage and the underlying population, the extent to which they are correlated and how their relationships change over space and time need to be further examined and validated. Some studies based on CDRs have used spatiotemporal patterns of cellphone usage (e.g., call volume) to predict land use types (Soto and Frías-Martínez 2011; Pei et al. 2014) and dense urban areas (Vieira et al. 2010). The reliability of these predictions also depends on the assumption of the relationship between cellphone usage and population distribution.

People organize their daily tasks (e.g., sending emails and browsing websites) "on a timescale that is appropriate to its urgency" (Ball 2010, p. 692). Researchers find that individual cellphone usage possesses a "bursty" nature (Candia 2008; Barabási 2010). People could make several phone calls in a short period of time and then none for hours. That means mobile phone data (e.g., Erlang and CDRs) could lead to a biased view of human activities. In recent years, several studies have investigated the bias of mobile phone data in geographical research (Ranjan et al. 2012; Zhao et al. 2016). However, these studies mainly focus on particular aspects of human mobility patterns (e.g., radius of gyration and movement entropy). The relationships between aggregate cellphone usage and population distribution are not examined.

People often spend a large amount of time at specific locations such as home and workplace. Studies have found that mobile phone data can be used to estimate people's activity "anchor" points (Gonzalez et al. 2008; Cho et al. 2011; Xu et al. 2015, 2016). These activity "anchor" points, especially home locations, are used to estimate urban population distributions (Ahas et al. 2010b; Silm and Ahas 2010). However, the anchor-point based approach associates individuals to one or few fixed locations. It thus provides population estimates that are static or at a coarse temporal resolution. Instead, Kang et al. (2012) compares people's cellphone usage and population distributions derived from LandScan data for Harbin, China at a

finer temporal resolution (1 h). The authors conclude that the proportion between active mobile subscribers and the actual total population varies in different areas, thus cannot reflect the underlying population properly. However, by using two CDR data sets collected in Portugal and France, Deville et al. (2014) find that the density of active mobile phone users can be used to produce spatially and temporally explicit estimations of population densities at national scales. It appears that researchers have not reached a consensus. It is thus important to look deeper into this issue, which has broad implications for human geography and other related fields.

3.3 Study Area and Mobile Phone Data Set

Shanghai is a century old metropolis. The city has a resident population of 24 million as of 2014 and covers an area of 6340 km^2 (Shanghai Bureau of Statistics 2014). It is the largest city in China by population. As a global financial center, its annual gross domestic product (GDP) was ranked the first among all cities in China in the past five years. The city consists of sixteen administrative districts and the Chongming county (Fig. 3.1a). Eight of them on the west bank of Huangpu River (i.e., Putuo, Zhabei, Hongkou, Yangpu, Jingan, Changning, Xuhui and Huangpu), also known as Puxi, are considered as the historic and commercial center of Shanghai (Fig. 3.1b).

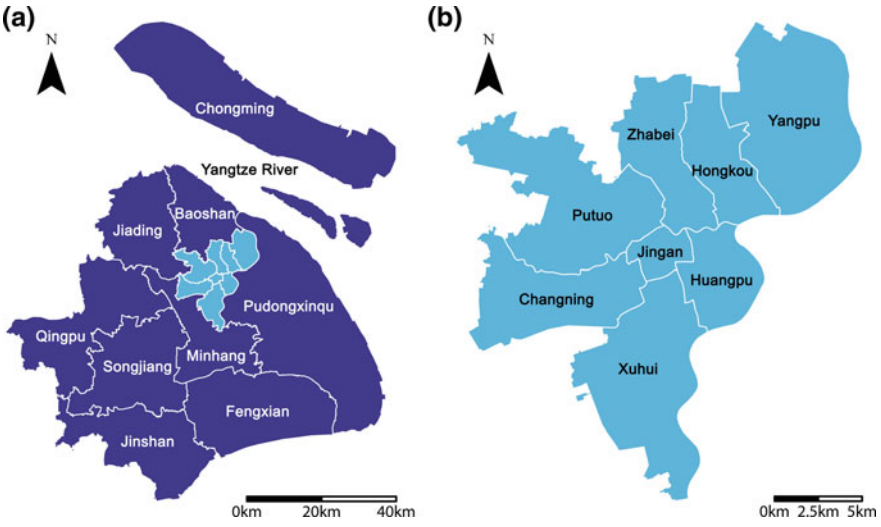

Fig. 3.1 Study area: **a** administrative districts of Shanghai, **b** inset map of the central part (i.e. Puxi) of Shanghai

The mobile phone data set used in this study was collected on a workday in 2012 by a phone service provider in China. As shown in Table 3.1, this data set contains CDRs plus other cellphone-related logs (e.g., regular update, periodic update, cellular handover, power on, and power off). These cellphone-related logs enable us to capture distributions of mobile phone users over space and time better than CDR data. In this data set, each mobile phone record contains information such as the type of event, time (i.e., when the event occurred), and geographic coordinates of the serving cellphone tower. The average nearest distance among cellphone towers operated by this phone service provider in Shanghai is 0.21 km.

Note that we removed mobile subscribers who had *power on* or *power off* event during the study period, since it is difficult to infer their locations when mobile phones are disconnected from the cellular network. The remaining data set after filtering these individuals consists of 698,661 mobile subscribers. As illustrated in Fig. 3.2, we first derive the spatiotemporal distribution of mobile phone users from the complete data set. Meanwhile, we extract all CDRs into a separate data set to capture the intensity of mobile phone communication at different places over time. The relationships between the two types of distributions are then evaluated through correlation and regression analyses.

Table 3.1 Summary of events captured in the mobile phone data set

Type	Event	Description
OT	Phone communication (outbound)	A subscriber makes a phone call or sends a text message
IN	Phone communication (inbound)	A subscriber receives a phone call or text message
RU	Regular update	Triggered by moving from the service area of a cellphone tower to that of another
PU	Periodic update	Triggered by tower pinging if a subscriber has been silent (i.e., no other events detected) for a certain period of time. However, the duration of silence that triggers periodic update is irregular. In addition, mobile phones which are turned off or disconnected from the cellular network do not receive pinging signals from the cellular network
CH	Cellular handover	Transfer of an ongoing phone call from one cellphone tower to another due to a subscriber's movements
ON	Power on	Mobile phone is turned on and connected to the cellular network
OFF	Power off	Mobile phone is turned off and disconnected from the cellular network

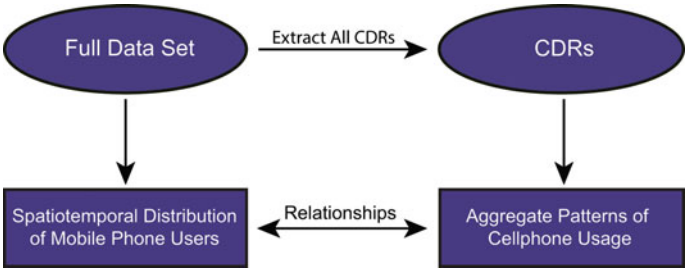

Fig. 3.2 The mobile phone data set consists of call detail records (CDRs) and other cellphone-related logs (e.g., regular update, periodic update, and cellular handover), which make it possible to examine the spatiotemporal relationships between aggregate cellphone usage and phone user distributions

3.4 Research Design

3.4.1 Defining Indicators of Aggregate Cellphone Usage

Mobile phones have become an essential part of people's everyday lives. In recent years, how people use their mobile phones and its societal implications have attracted increasing research interests. When analyzing mobile phone data, previous studies (e.g., Candia et al. 2008; Kang et al. 2012; Yuan et al. 2012) often processed cellphone usage data to reflect either individual phone communication characteristics (e.g., phone call frequency, inter-event time) or collective phone communication activity patterns (e.g., Erlang, call volume). In this study, two indicators of aggregate cellphone usage are selected for the correlation analysis:

- $\boldsymbol{V}$: *volume of calls/text messages*
- $\boldsymbol{N}$: *number of active mobile phone users*.

Note that an individual's cellphone trajectory S can be represented as:

$$S = \{P_1(x_1, y_1, t_1, e_1), P_2(x_2, y_2, t_2, e_2), \ldots, P_i(x_i, y_i, t_i, e_i)\} \tag{3.1}$$

where P_i denotes the ith cellphone record; x_i and y_i denote the longitude and latitude of the serving cellphone tower; t_i and e_i represents the time and type of the corresponding mobile phone event (see Table 3.1), respectively.

Given a geographic area $\boldsymbol{A}$ and a time interval $\boldsymbol{T}$, we define: (1) $\boldsymbol{V}_A^{\boldsymbol{T}}$ as the total number of phone calls/text messages that occurred within the area A during a time interval T, and (2) $\boldsymbol{N}_A^{\boldsymbol{T}}$ as the total number of mobile phone users who have made or received at least one phone call/text message within the area A during a time interval T. The two indicators reflect important characteristics of aggregate cellphone usage, and are generated only using the CDRs extracted from the full data set (i.e., records with event type $\boldsymbol{e}$ being IN or OT in Table 3.1).

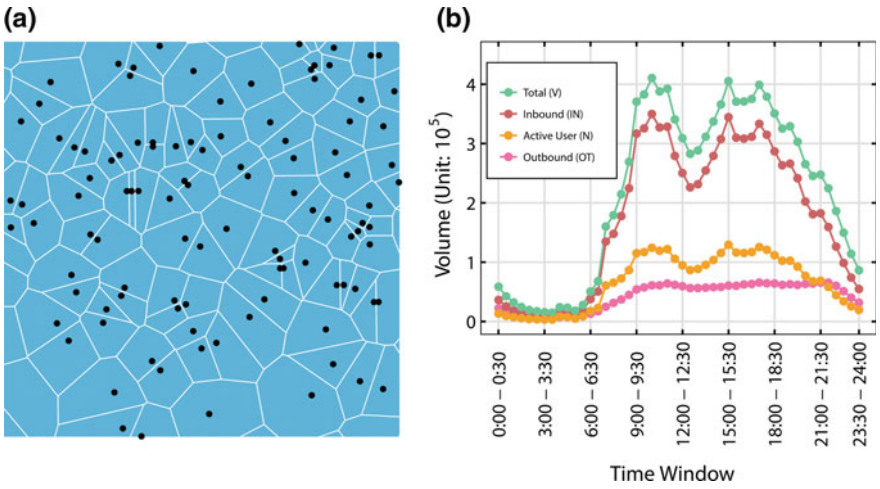

Fig. 3.3 **a** Thiessen polygons are generated based on the spatial distribution of cellphone towers to approximate their service areas. The two indicators (*V* and *N*) are calculated at each Thiessen polygon using the mobile phone records that occurred at the corresponding cellphone tower; **b** Global temporal patterns of aggregate cellphone usage at a 30-min time interval (*V* total volume of phone calls/text messages; *N* total number of *active* mobile phone users; *IN* total inbound phone communication activities; *OT* total outbound phone communication activities)

This study uses Thiessen polygons as the spatial units to derive the cellphone usage indicators. Specifically, Thiessen polygons, which are generated based on the spatial distribution of the cellphone towers, are used to approximate their service areas (Fig. 3.3a). The two indicators can be calculated for each Thiessen polygon using mobile phone records that occurred at the corresponding cellphone tower. Figure 3.3b illustrates the global temporal patterns of the two cellphone usage indicators (as well as the inbound and outbound phone communication activities) at a 30-min time interval. The total volume of phone calls/text messages (*V*) stays relatively low between midnight and 6:00. It starts to increase in the morning, followed by a fluctuation stage (i.e., 10:00–17:00), and then decreases in the evening. The number of *active* mobile phone users (*N*) follows a similar pattern of *V* but has lower intensities. The temporal variations of *V* and *N* indicate that the relationship between aggregate cellphone usage and the *total* number of mobile phone users in the city varies greatly throughout the day. However, how their relationships change over space and time remains unclear and is worth an investigation.

3.4.2 Deriving the Spatiotemporal Distribution of Mobile Phone Users

People don't use their mobile phones regularly over time. As CDRs only record the locations visited by people during their phone communication activities, it is questionable to use such data to infer human dynamics when no phone calls or text messages take place. The mobile phone data set used in this study includes location records generated by other events such as regular update (RU), periodic update (PU), and cellular handover (CH). These mobile phone events enable us to infer individual locations at a finer time interval. For example, the RU and CH events allow an individual's location to be continuously updated when he or she is moving from the service area of one cellphone tower to another. When an individual stays at one particular location or has no phone communication activities, his or her location is still reported by the PU event. Thus, the complete data set (i.e., cellphone-related logs along with the CDRs) enables us to estimate a phone user's location at any given time point no matter he or she is moving.

Hence, given an individual cellphone trajectory S, the phone user's location at a particular time point t can be reasonably estimated using the following criteria: (1) if trajectory S contains at least one mobile phone record after time point t, then the phone user's location is estimated as (x_i, y_i) using the mobile phone record $P_i(x_i, y_i, t_i, e_i)$. Here P_i denotes the first mobile phone record which occurred after time point t; (2) if trajectory S has no mobile phone records after t, the mobile phone's location is estimated using the last mobile phone record which occurred before time point t. By doing so, we can estimate each phone user's location at any given time point t, and aggregate all users at the level of cellphone tower service area. These estimates of the spatiotemporal distributions of phone users can be combined with the two cellphone usage indicators for correlation and regression analysis.

It is necessary to note that these estimates are not without uncertainties. On one hand, it is very difficult to pinpoint a mobile phone user's location when it is travelling among different cellphone tower service areas. On the other hand, given the issues of cellphone load balancing or "ping-pong effect" (Isaacman et al. 2012; Csáji et al. 2013), the x, y coordinates of a cellphone tower associated with a particular mobile phone record might not reflect where a user actually stayed. Hence, it is more appropriate to conduct the correlation analysis at a coarser spatial granularity—for example, using a regular grid with a coarser spatial resolution than the cellphone tower service areas—to mitigate the impact of spatial uncertainty.

3.4.3 Correlation and Regression Analysis

Many existing studies, which use mobile phone data for urban mobility research, have an implicit assumption that phone communication activities are highly

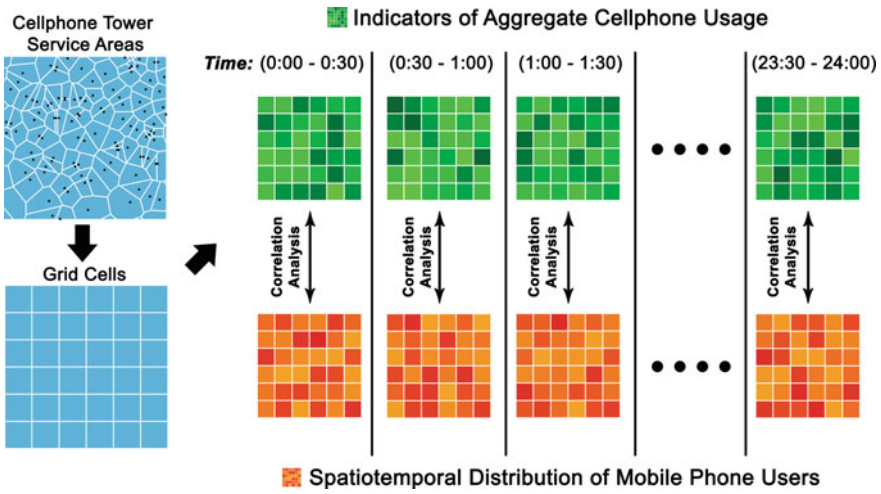

Fig. 3.4 Correlation analysis in a space-time context

correlated to the population size, or at least the number of mobile phone users. While this assumption might hold true, it is important to examine the role of time in such relationships. For example, given a geographic area *A*, although the number of phone calls/text messages may be roughly the same in early morning (e.g., 07:00–07:30), in late afternoon (e.g., 17:00–17:30), and around midnight (e.g., 23:00–23:30), the total number of mobile phone users observed in each time interval could be quite different from each other. This study conducts a correlation analysis using time as a control factor. As illustrated in Fig. 3.4, we first divide the study area into 1km $*$ 1km regular grid cells. By partitioning a day into forty-eight 30-min time windows, we capture the snapshots of aggregate cellphone usage and the number of mobile phone users in each grid cell for each 30-min time window. These snapshots are used to examine their correlations at different times in a day. We choose the 1km $*$ 1km regular grid in order to obtain the estimates of phone user distribution at a relatively fine spatial resolution while minimizing the spatial uncertainty of mobile phone records.

To perform analysis at the selected spatiotemporal resolution, we first calculate the two indicators of aggregate cellphone usage and the total number of mobile phone users at the level of cellphone tower service areas (i.e., Thiessen polygons). If a mobile phone user has more than one mobile phone record during a particular time window, we use the Thiessen polygon that contains the first mobile phone record as his or her representative location. If there is no mobile phone record during a particular time window, it means this user did not move. Thus, the phone user's location can be estimated using the approach described in the previous section. Once this step is completed, we transform the results onto the grid cells. Considering that a Thiessen polygon could overlap with multiple adjacent grid cells, we clip each Thiessen polygon into sub units. For each sub unit, the indicators

of aggregate cellphone usage (V and N) and the total number of mobile phone users (Pop) are prorated based on the proportion of its area to the total area of the corresponding Thiessen polygon. We then calculate V, N and Pop of each grid cell by adding the values of all sub units that fall within the particular grid cell.

For each time interval T, we first analyze the correlation between the number of mobile phone users (Pop) and each of the two cellphone usage indicators using Pearson's correlation coefficients:

$$\rho^{T}_{Pop,V} = \frac{cov(Pop^{T}, V^{T})}{\sigma_{Pop^{T}} * \sigma_{V^{T}}} \tag{3.2}$$

$$\rho^{T}_{Pop,N} = \frac{cov(Pop^{T}, N^{T})}{\sigma_{Pop^{T}} * \sigma_{N^{T}}} \tag{3.3}$$

where: (1) $cov()$ stands for the covariance and σ_X denotes the standard deviation of X; (2) $Pop^{T} = \left[Pop^{T}_{1}, Pop^{T}_{2}, \ldots, Pop^{T}_{m}\right]$, $V^{T} = \left[V^{T}_{1}, V^{T}_{2}, \ldots, V^{T}_{m}\right]$, and $N^{T} = \left[N^{T}_{1}, N^{T}_{2}, \ldots, N^{T}_{m}\right]$; (3) m denotes the total number of grid cells in the study area. The values of $\rho^{T}_{Pop,V}$ and $\rho^{T}_{Pop,N}$ enable us to better assess their correlations during different time periods of a day.

In this study, we introduce two types of regression models that have been suggested in previous studies (Kang et al. 2012; Deville et al. 2014) to further investigate the relationships between the number of mobile phone users and the aggregate cellphone usage:

$$\text{Model 1}: \quad Pop^{T} = a * V^{T} + b \tag{3.4}$$

$$\text{Model 2}: \quad Pop^{T} = a * N^{T} + b \tag{3.5}$$

$$\text{Model 3}: \quad \log_{10}(Pop^{T}) = a * \log_{10}(V^{T}) + b \tag{3.6}$$

$$\text{Model 4}: \quad \log_{10}(Pop^{T}) = a * \log_{10}(N^{T}) + b \tag{3.7}$$

In these regression models, the dependent variable is the total number of mobile phone users in each grid cell during a particular time window (Pop^{T}), and the independent variable is the cellphone usage indicator (V^{T} or N^{T}). Model 1 and Model 2 assume a linear relationship between Pop^{T} and V^{T} (or N^{T}), while Model 3 and Model 4 (i.e., log-transformation models) quantify the power-law relationships between Pop and each of the two cellphone usage indicators. The ordinary least squares (OLS) method is used to derive the parameters of these regression models. As the study day is partitioned into forty-eight 30-min time windows, each model produces 48 sets of parameters. We then use three measures, which are the adjusted R^2, the root mean square error (RMSE), and the mean absolute percentage error

(MAPE), to compare the performance of these regression models at different times in a day[2].

3.5 Results and Discussion

3.5.1 *Correlation Between the Number of Phone Users and the Two Cellphone Usage Indicators*

Figure 3.5a shows the values of $\rho^T_{Pop,V}$ and $\rho^T_{Pop,N}$ and how they change over time. In general, there is a high correlation between the *total* number of mobile phone users (*Pop*) and each of the two cellphone usage indicators during the day time and in the evening. Also, the correlation of *Pop* and the number of *active* mobile phone users (*N*) is always higher than that of *Pop* and the volume of calls/messages (*V*) in the same time window.

According to the temporal variations of $\rho^T_{Pop,V}$ and $\rho^T_{Pop,N}$, the study day can be categorized into several stages. From 07:00 to 21:30, the values of $\rho^T_{Pop,V}$ and $\rho^T_{Pop,N}$ stay above 0.9 and remain relatively stable. A decrease of $\rho^T_{Pop,V}$ and $\rho^T_{Pop,N}$ is observed during 00:00–04:00 and 21:30–24:00, which refer to the time when people have fewer phone communication activities (see Fig. 3.3b). To our surprise, there are some fluctuations of $\rho^T_{Pop,V}$ and $\rho^T_{Pop,N}$ during 03:30–05:00, which is followed by a rising stage (05:00–07:00). These fluctuations, which are somewhat counter-intuitive, encourage us to explore potential explanations. Specifically, we derive several cellphone usage indicators to distinguish inbound and outbound phone communication activities, and further examine their correlations with the number of mobile phones:

- ***N_Inbound*** *Number of active mobile phone users derived from inbound phone communications* (*IN*) *only*
- ***N_Outbound*** *Number of active mobile phone users derived from outbound phone communications* (*OT*) *only*
- ***V_Inbound*** *Volume of inbound calls/messages*
- ***V_Outbound*** *Volume of outbound calls/messages*.

We find that the correlation coefficients of *Pop* versus *N_Outbound* (i.e., $\rho^T_{Pop,N\_Outbound}$ shown in Fig. 3.5b) and *Pop* versus *V_Outbound* (i.e., $\rho^T_{Pop,V\_Outbound}$ shown in Fig. 3.5c) exhibit smooth temporal variations before 07:00. However, the temporal patterns of $\rho^T_{Pop,V\_Inbound}$ and $\rho^T_{Pop,N\_Inbound}$ are very

[2]For Model 3 and Model 4, the three measures (adjusted R^2, RMSE and MAPE) are calculated after converting $\log_{10}(Pop^T)$, $\log_{10}(V^T)$ and $\log_{10}(N^T)$ to the original scale (i.e., Pop^T, V^T, and N^T, respectively).

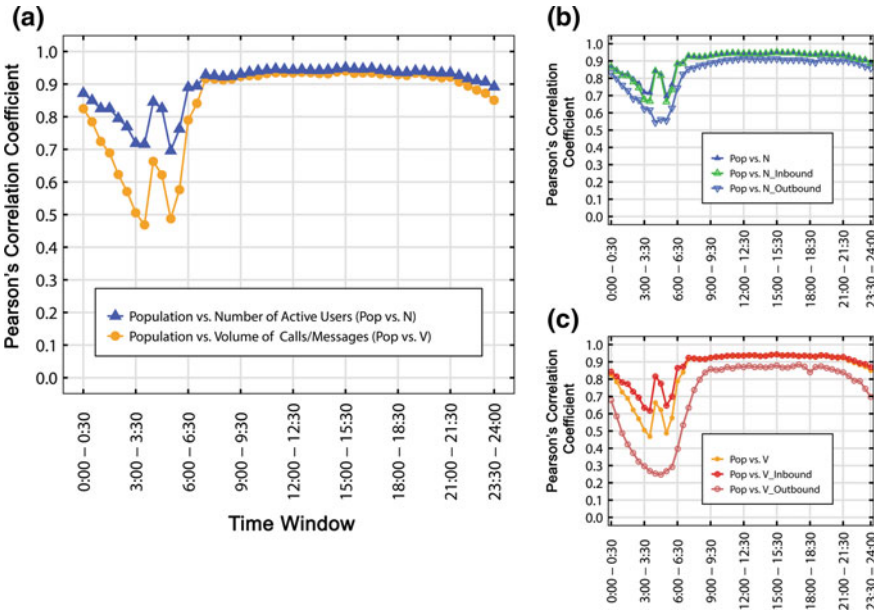

Fig. 3.5 **a** Pearson's correlation coefficients of the *total* number of mobile phone users (*Pop*) and each of the two cellphone usage indicators **b** Pearson's correlation coefficients of *Pop* versus *N_Inbound* (i.e., $\rho^T_{Pop,N\_Inbound}$) and *Pop* versus *N_Outbound* (i.e., $\rho^T_{Pop,N\_Outbound}$). *N_Inbound* denotes the number of *active* mobile phone users derived from the inbound phone communications only. *N_Outbound* denotes the number of *active* mobile phone users derived from the outbound phone communications only; **c** Pearson's correlation coefficients of *Pop* versus *V_Inbound* (i.e., $\rho^T_{Pop,V\_Inbound}$) and *Pop* versus *V_Outbound* (i.e., $\rho^T_{Pop,V\_Outbound}$). *V_Inbound* denotes the volume of inbound calls/text messages, and *V_Outbound* denotes the volume of outbound calls/text messages

similar to that of $\rho^T_{Pop,V}$ and $\rho^T_{Pop,N}$, respectively. This is probably because outbound phone communications are initiated by mobile phone users, while inbound phone communications could include push notifications such as advertisements, weather forecast, news, etc. It is very likely that the fluctuations between 03:30–05:00 are caused by these inbound messages. It also reminds us that the two primary indicators, N and V, are the combined effects of outbound and inbound phone communications, which not only are related to how people use their mobile phones, but also are related to mobile phones' passive interactions with the outside world.

3.5.2 *Comparison of Regression Models*

We include four regression models to further examine the relationships between the total number of mobile phone users and each of the two cellphone usage indicators.

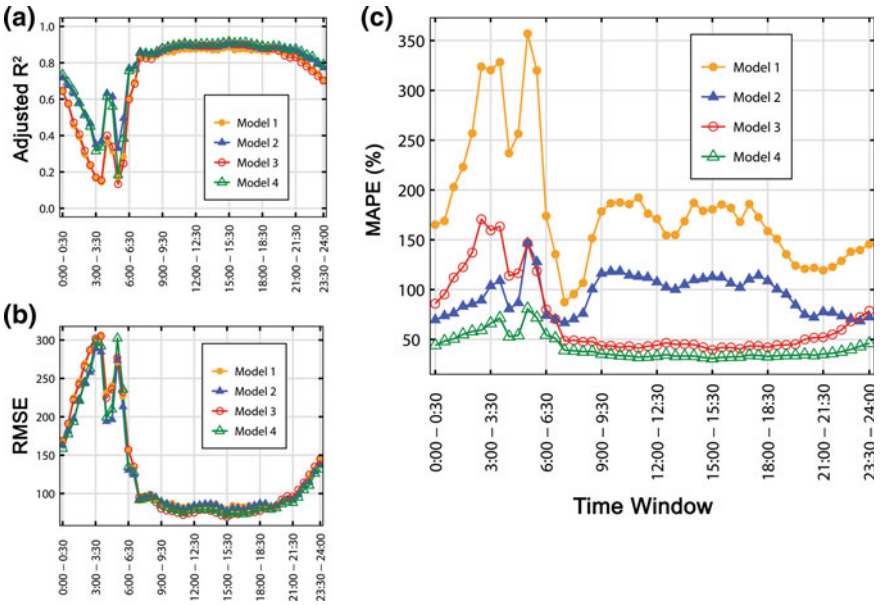

Fig. 3.6 Comparison of the four regression models: **a** Adjusted R^2; **b** Root mean square error (RMSE); **c** Mean absolute percentage error (MAPE)

The adjusted R^2, root mean square error (RMSE), and mean absolute percentage error (MAPE) are used to assess the model performance (Fig. 3.6). As illustrated in Fig. 3.6a, the two models with N^T as the independent variable (i.e., Model 2 and Model 4) have a higher adjusted R^2 than the other two models in each time window. However, when comparing the RMSE of the four models, we find that Model 3 and Model 4 perform better than the other two models during the daytime (07:00–18:00). Notice that the total number of mobile phone users (*Pop*) in the grid cells could vary greatly from each other, it is important to use a normalized measure, which is MAPE in our analysis, to further evaluate the model performance. As illustrated in Fig. 3.6c, Model 3 and Model 4 have a much lower MAPE than the other two models during the daytime. All the three measures suggest that Model 4 performs better than the other three models. The average MAPE of Model 4 between 00:00–07:00 is 58.5%, as compared to 35.1% between 07:00–24:00.

Comparisons of the four regression models indicate that the heterogeneity (or variation) of *Pop* is better explained by the number of *active* mobile phone users (N^T) than by the volume of calls/messages (V^T). As suggested by Barabási (2010), human activities are not random but "bursty"[3]. At a given place during a given time period, V^T is more affected by the individual "burst" of phone communications than

[3]In the context of this book, the author refers the word "burst" to brief periods of intensive human activities (e.g., sending text messages) followed by long periods of no activities.

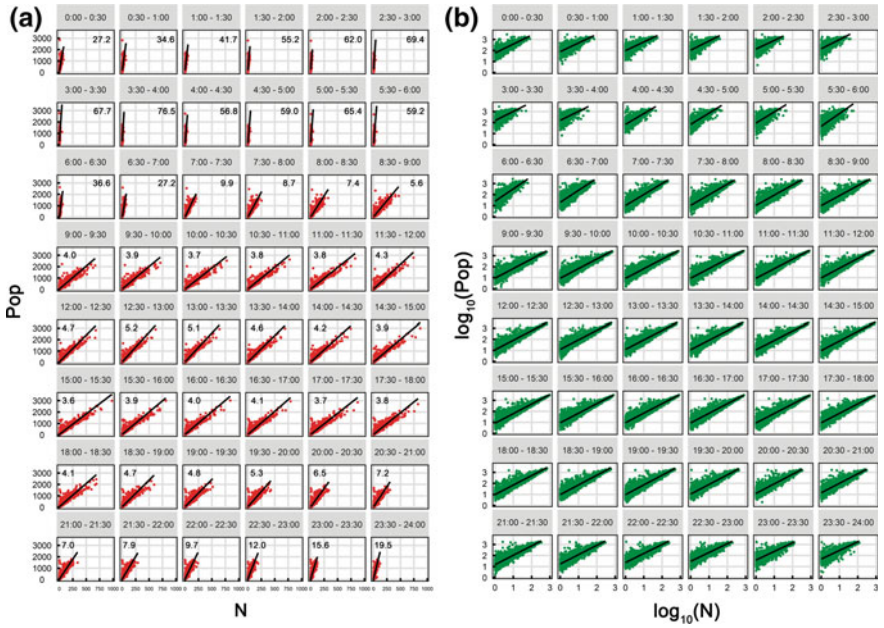

Fig. 3.7 Scatter plots of: **a** *Pop* versus N; **b** $log_{10}(Pop)$ versus $log_{10}(N)$ during different time windows. The black line in each plot denotes the regression line (the numbers in each plot of Fig. 3.7a denotes the coefficient of the regression line)

N^T is, which serves as one potential explanation to our findings. The model comparisons also suggest that the relationships between the total number of mobile phone users and the cellphone usage level are better explained by the log-transformation models (i.e., Model 3 and Model 4) than by the simple linear regression models (i.e., Model 1 and Model 2) when the independent variable (N^T or V^T) is fixed.

Although our findings suggest that the log-transformation models (i.e., Model 4) better describes the relationship between *Pop* and N than the simple linear regression models (i.e. Model 2), it is useful to compare the relationships of *Pop* and N between these two models and examine how the relationships change over time. The scatter plots of *Pop* versus N (Fig. 3.7a) suggest that it is inherently biased to use aggregate cellphone usage to represent the number of mobile phone users. More importantly, the slope of the regression lines indicates that the relationship between *Pop* and N varies greatly throughout a day. That means, even if a place has (or two different places have) the same number of *active* mobile phone users (N) during two different time periods, their values of *Pop* could be quite different from each other.

3.5.3 Cross Validation

We perform a k-fold cross validation to further assess the robustness of Model 4. In particular, the mobile phone data set is partitioned into k subsets with roughly the same size. During the data partition, each individual mobile phone user has an equal probability (i.e., $1/k$) of being assigned to any given subset. Thus, all the subsets after data partition will have approximately the same distribution patterns. During the cross validation process, each time $k-1$ subsets are used as a *training data set*, and the remaining subset is used as a *validation data set*. The training data set is used to produce the parameters of Model 4, which are then used to predict the total number of mobile phone users (*Pop*) of the validation data set. To ensure that each subset covers an adequate number of samples, we choose $k = 3$ for this particular analysis. Specifically, we perform the 3-fold cross validation 10 times—with each time using a new partition of the mobile phone data set—in order to control the impact of data partition on the analysis results. By doing so, we obtain 30 ($C_3^2 * 10$) pairs of training and validation data sets, and several measures (e.g., average MAPE and average RMSE) of the 30 iterations are used to evaluate the model performance.

Note that we also compare the performance of Model 4 based on: (1) the ordinary least squares (OLS) and (2) the population-weighted least squares (PWLS). The model based on PWLS minimizes the sum of squared residuals weighted by the total number of mobile phone users (*Pop*). Thus, the samples (i.e., grid cells) with smaller *Pop* will have less impact on the regression result. This comparison is expected to generate additional insights into the prediction capability of Model 4.

Figure 3.8 illustrates the model performance of OLS and PWLS. We find that the OLS model generates similar MAPE from the 3-fold cross validation (i.e. green line in Fig. 3.8a) and the full data set (i.e., green line in Fig. 3.6c), which indicates the robustness of Model 4. On the other hand, the OLS model generates lower MAPE (Fig. 3.8a) but higher RMSE (Fig. 3.8b) than the PWLS model. The temporal variations of MAPE/RMSE produced by the two models can be better understood by dividing the study day into two stages:

- ***Stage A*** refers to the time periods from 00:00 to 07:00 and from 21:30 to 24:00, when the majority of people rest at home. During this stage, people have fewer phone communication activities (Fig. 3.3b). The correlation between *Pop* and the aggregate cellphone usage varies greatly during this stage (Fig. 3.5), which causes notable fluctuations of prediction accuracy (as shown in Fig. 3.8a, b).
- ***Stage B*** refers to the time period from 07:00 to 21:30, when cellphone communications and other human activities are active. During this stage, the prediction accuracy is much better than that of stage A and remains relatively stable.

We further examine the relationship between the prediction error (i.e. absolute percentage error) and the value of dependent variable (*Pop*) during these two stages

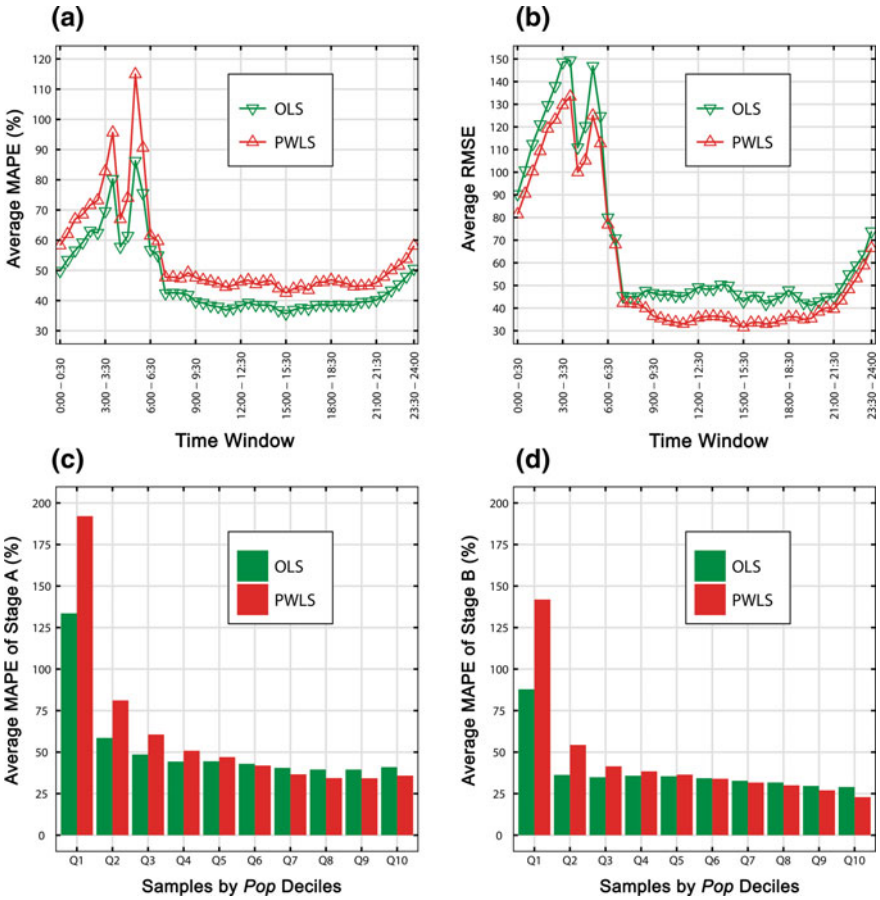

Fig. 3.8 The 3-fold cross validation: **a** Average MAPE of Model 4 based on the ordinary least squares (OLS) method, and the population-weighted least squares (PWLS) method; **b** Average RMSE of OLS and PWLS; **c** Average MAPE of samples by deciles of *Pop* for the stage A (i.e., 00:00–07:00 and 21:30–24:00); **d** Average MAPE of samples by deciles of *Pop* for the stage B (i.e., 07:00–21:30)

produced by OLS and PWLS, respectively. For each model, we aggregate the samples (from validation data sets) during the stage A and the stage B, respectively. For each stage, we organize the samples in an ascending order of *Pop*. We then divide these samples into deciles ($Q1$, $Q2$, … $Q10$) and calculate the average MAPE of each decile. Figure 3.8c illustrates the average MAPE of the samples organized by deciles (generated by the two models) during the stage A. In general, both models yield a better estimation as *Pop* gets larger. Also, PWLS generates better results than the OLS model when the samples have a large value of *Pop* (i.e., $Q6$–$Q10$). An implication is that it is more appropriate to use PWLS than OLS

Table 3.2 The range of value by *Pop* (i.e., the number of mobile phones) decile for stage A (from 00:00 to 07:00 and from 21:30 to 24:00) and stage B (from 07:00–21:30)

Stage A			Stage B		
Pop decile	Minimum value	Maximum value	*Pop* decile	Minimum value	Maximum value
Q1	1.0	20.0	Q1	1.0	8.0
Q2	21.0	32.0	Q2	9.0	13.0
Q3	33.0	45.0	Q3	14.0	18.0
Q4	46.0	61.0	Q4	18.0	25.0
Q5	62.0	80.0	Q5	26.0	34.0
Q6	81.0	109.0	Q6	35.0	46.0
Q7	110.0	150.0	Q7	47.0	63.0
Q8	151.0	214.0	Q8	64.0	92.0
Q9	215.0	304.0	Q9	93.0	168.0
Q10	305.0	984.0	Q10	169.0	1046.0

under certain scenarios (e.g., evacuation) when we want to produce better estimates in populated areas.

Similar patterns are observed for the two models during the stage B (Fig. 3.8d). Both models of this stage, which refers to daytime and early evening, perform relatively well except for $Q1$ and $Q2$. The two models, especially PWLS, do not perform well on $Q1$ and $Q2$ because these two deciles have a small value of *Pop*. As shown in Table 3.2, the maximum value of *Pop* for $Q1$ and $Q2$ during the stage B is 8.0 and 13.0, respectively. The model performance over $Q1$ and $Q2$ are affected more by the unique characteristics of individual phone communication activities due to a smaller number of mobile phone users.

The 3-fold cross validation suggests that it is more reasonable to use Model 4 to approximate the relationship of *Pop* and N, especially during the daytime and in early evening (i.e. Stage B). After removing the samples with very small values of *Pop* (i.e., $Q1$ and $Q2$), we find that the average MAPE of OLS and PWLS models during the stage A changes to 42.6 and 42.7%, respectively. The two models perform better during the stage B, with an average MAPE of 33.0 and 32.8%, respectively. Overall, the PWLS model performs better than the OLS model due to its lower RMSE (Fig. 3.8b).

3.5.4 Spatiotemporal Patterns of Residuals

By performing the 3-fold cross validation, we are able to derive the residuals—measured as the average percentage errors (i.e., $\frac{Y_{predicted}-Y_{observed}}{Y_{observed}} * 100\%$) of 30 iterations using PWLS—at each grid cell for all time windows. A grid cell with a positive or a negative percentage error (during a time window T) suggests that

Model 4 overestimates or underestimates the total number of mobile phone users (*Pop*), respectively. The spatial and temporal patterns of these residuals can help us better understand the relationship between *Pop* and *N*. As Model 4 produces better estimations during the daytime and in early evening (Fig. 3.8), this section only discusses the findings during the stage B, which covers twenty-nine 30-min time windows from 07:00 to 21:30.

The residuals of a grid cell *G* during the stage B can be represented as follows:

$$G = \left\{ erorr^{T_{B1}}, erorr^{T_{B2}}, \ldots, erorr^{T_{Bi}} \ldots, erorr^{T_{B29}} \right\} \tag{3.8}$$

where T_{Bi} denotes the time windows (for example, T_{B1} refers to [07:00–07:30], and TB_{29} refers to [21:00–21:30]). Figure 3.9a shows the residuals of a grid cell that covers part of the *Nanjing Road* pedestrian-only shopping street (Fig. 3.9b), which is one of the world's busiest shopping streets located in Huangpu district in Shanghai. The percentage error of this grid cell remains positive during most of the time windows, which means that the model has mostly overestimated the total number of mobile phone users. The constant overestimation reveals an important fact that a larger percentage of people tend to use their mobile phones in this grid cell (during the daytime and in early evening) as compared to the overall population.

Note that some grid cells may not have mobile phone users (i.e., $Pop = 0$) during particular time windows, which lead to missing values (i.e., $error^{TBi} = NA$) in the residual *G*. As the purpose of this analysis is to find grid cells with similar temporal patterns of residuals, these missing values must be handled appropriately. Figure 3.9c shows the distribution of grid cells with varying number of missing values. It is likely that the grid cells with a large number of missing values reflect less populated areas in Shanghai or areas where mobile phone records are sparse. In this section, we focus on grid cells with no more than 10 missing values (4070 cells in total). As shown in Fig. 3.9d, these grid cells mainly cover the core areas of Shanghai (Fig. 3.1b) and some other administrative districts (e.g., Pudongxinqu, Minhang, Songjiang, Qingpu, Jiading, and Baoshan). As these grid cells have observations during the majority of the time windows, we replace the missing values for each grid cell using a linear interpolation method. For instance, if the percentage error of a grid cell is 10% during time window T_{B1}, and 20% during time window T_{B3}, then the value during T_{B2} (if missing) is estimated as 15%.

We further divide the stage B into three time periods: (1) 07:00–12:00; (2) 12:00–17:00, and (3) 17:00–21:30. For each grid cell, we calculate the percentage of time windows with positive and negative residuals (in each of these three time periods). The temporal patterns of residuals during a particular time period can be characterized as follows:

- If the percentage of time windows with positive residuals is equal to or larger than 75, this time period is labeled as "dominated by overestimations".
- If the percentage of time windows with negative residuals is equal to or larger than 75, this time period is labeled as "dominated by underestimations".
- Otherwise, this time period is labeled as "mixed patterns".

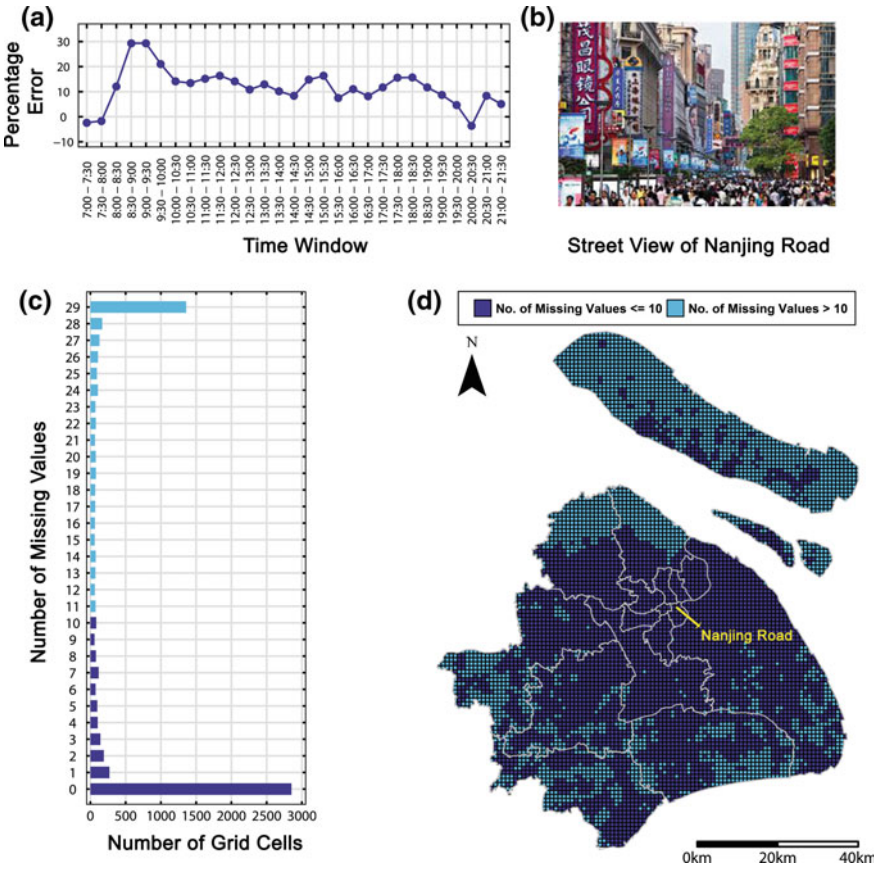

Fig. 3.9 **a** Temporal variations of residuals (i.e., percentage errors) of a grid cell that covers part of the *Nanjing Road* pedestrian-only shopping street; **b** A street view of *Nanjing Road* (picture from Google Image); **c** Distribution of grid cells with varying number of missing values; **d** Geographic distributions of grid cells with no more than 10 missing values

Figure 3.10 illustrates the major types of grid cells ($C1$–$C14$) with distinct temporal patterns. It is interesting to find that none of these grid cells are mixed with time periods dominated by overestimations (i.e., red segments) and underestimations (i.e., green segments). It is likely that there are some inherent characteristics of the built environment which govern the relationships between the aggregate cellphone usage and the total number of mobile phone users. To better understand the geographic context of these grid cells, we map them onto Google Earth and visually examine some of these places through photos, landmarks, and semantic descriptions. As shown in Fig. 3.11, the grid cells with time period(s) dominated by overestimations (i.e., $C1$–$C7$) cover some important commercial and business areas in Shanghai (i.e., grid cells A to F). At these places, more (i.e., a larger percentage of) people tend to use their mobile phones than the average (percentage) of overall

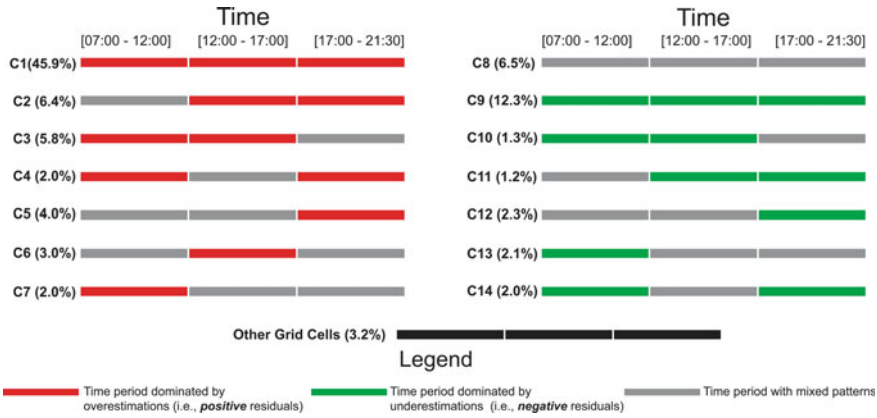

Fig. 3.10 Grid cells with different temporal characteristics (of residuals)

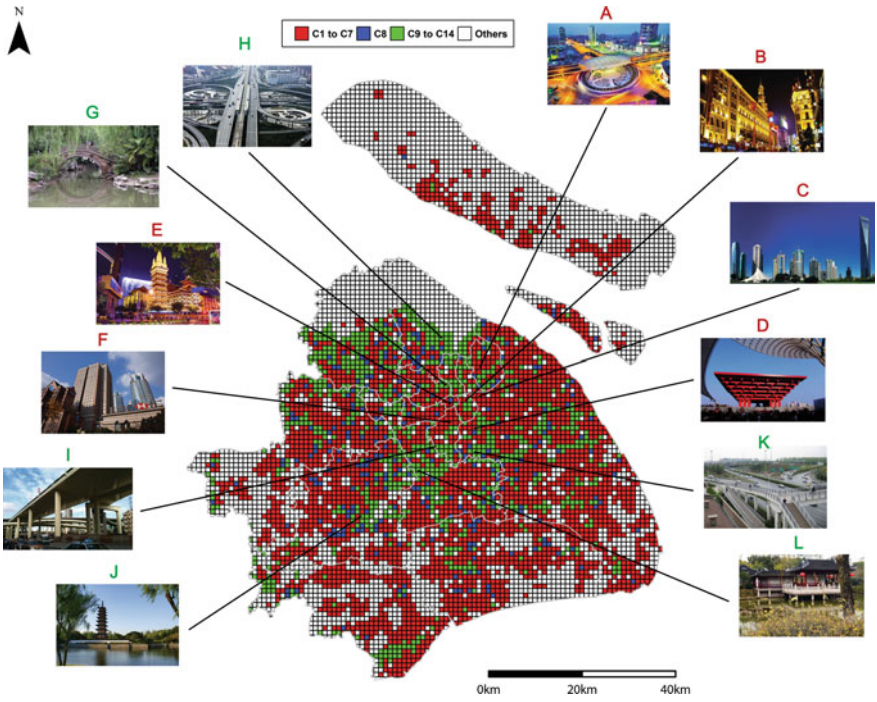

Fig. 3.11 The spatial distribution of grid cells with distinct temporal patterns of residuals. Grid cells A to F, with certain time period(s) dominated by overestimations, refer to some important commercial and business areas in Shanghai. Grid cells G to L, with certain time period(s) dominated by underestimations, represent certain parks (e.g., G, J and L) and places traversed by urban express ways (e.g., H, I, and K), [1]Pictures are captured from Google Image and Panoramio

population. We also find that the grid cells with time period(s) dominated by underestimations (i.e., *C*9–*C*14) include some parks (i.e., grid cells G, J and L) and places traversed by urban express ways (i.e., grid cells H, I and K). At these places, people's cellphone usage is less intense. The temporal patterns of residuals at these selected places suggest that certain characteristics of the built environment—such as land use type, points of interest (POI) and transportation infrastructures—could be considered in the analysis to further understand the behavior of mobile phone usage.

3.6 Conclusion

By using a mobile phone data set that consists of call detail records (CDRs) and other cellphone-related logs (e.g., cellular handover and periodic location update) collected in Shanghai, China, this study evaluates to what extent phone communication activities could reflect the spatiotemporal distribution of mobile phone users. Specifically, we derive two cellphone usage indicators (volume of calls/messages [*V*] and number of *active* mobile phone users [*N*]) as well as the *total* number of mobile phone users observed at different places in the city over time, and examine their relationships through correlation and regression analysis. We find that correlations between the number of mobile phone users and each of the two cellphone usage indicators remain high and stable (with Pearson's correlation coefficient above 0.9) during the daytime and in early evening (i.e., 07:00–21:30). Their correlations are generally lower in other time periods, and exhibit notable fluctuations between 00:00–07:00.

We then introduce four regression models (i.e., two simple linear regression models and two log-transformation models) to further examine relationships between the *total* number of mobile phone users (*Pop*) and the two cellphone usage indicators. Several important findings are discovered. First, comparisons of model performance indicate that the number of *active* mobile phone users (*N*) serves as a better independent variable than the volume of calls/messages (*V*) when explaining spatiotemporal distribution of mobile phone users. The volume of calls/messages—at a given place during a particular time period—is likely affected by individual "burst" of phone communication activities (Barabási 2010), which makes the number of *active* mobile phone users (*N*) a better indicator of the mobile phone user distribution. Second, the log-transformation model performs better than the simple linear regression model (in predicting phone user distribution) when the independent variable is fixed. Although the simple linear regression models do not have the best prediction accuracy, our results illustrate that the relationship between the total number of mobile phone users and the cellphone usage level varies greatly throughout a day. It is likely to generate biased results if we use the intensity of aggregate cellphone usage to directly reflect the mobile phone user distribution or the underlying population distribution, and the degree of bias varies with time. Researchers must be cautious when using phone communication activities to

quantify certain aspects of urban dynamics. Third, the 3-fold cross validation indicates that the log-transformation model (using V as the independent variable) has a prediction error (i.e., mean absolute percentage error) of 32.8% during the daytime and in early evening (i.e., 07:00–21:30), and 42.7% during other time periods (i.e., 00:00–07:00 and 21:30–24:00). The spatiotemporal patterns of residuals suggest that there exist some inherent characteristics of the built environment which govern the relationships between the cellphone usage and the number of mobile phone users. It suggests that CDR data can be used along with other data sources (e.g., land use type, POI, and transportation infrastructures) to deliver robust estimations of phone user distributions.

Mobile phone data can be leveraged to gain better insights into the whereabouts of people in space and time, which suggests that it serves as a promising data source to supplement traditional approaches (e.g., travel surveys) for studying dynamic population distributions. However, challenges still remain. For example, the mobile phone data used in this study are collected from a single phone company. As a city usually includes multiple phone companies, it is necessary to compare whether the relationships between the cellphone usage level and the distribution of mobile phone subscribers are similar across different cellular networks. How to integrate population estimates from multiple cellular networks in order to gain a more compressive view of urban population distribution is of great importance to applications in emergency response, public health, transport planning, among others.

This research examines only the spatiotemporal relationships between the aggregate cellphone usage and the phone user distributions on a weekday. How their relationships vary between weekdays and weekends, and how such relationships are influenced by special events are not examined in this study. Also, how the spatiotemporal resolutions (e.g., size of grid cell, length of time window) would influence the prediction accuracy is worth a further investigation. Future work can focus on these issues and combine other data sources (e.g., land use type and POI) with CDRs to deliver more robust estimations of mobile phone users and dynamic urban population distributions. Findings of this study provide some useful information and guidelines of using large-scale mobile phone data for geographical studies and urban dynamics research.

Acknowledgements This research was jointly supported by the Alvin and Sally Beaman Professorship and Arts and Sciences Excellence Professorship of the University of Tennessee, Natural Science Foundation of China (41231171, 41371377, 41501486, 91546106, 41571431), Key Program of the Chinese Academy of Science (ZDRW-ZS-2016-6-3), and Beijing Key Laboratory of Urban Spatial Information Engineering (2014101).

References

Ahas, R., Aasa, A., Mark, Ü., Pae, T., & Kull, A. (2007). Seasonal tourism spaces in Estonia: Case study with mobile positioning data. *Tourism Management, 28*(3), 898–910.

Ahas, R., Aasa, A., Silm, S., & Tiru, M. (2010a). Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area: case study with mobile positioning data. *Transportation Research Part C: Emerging Technologies, 18*(1), 45–54.

Ahas, R., Silm, S., Järv, O., Saluveer, E., & Tiru, M. (2010b). Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of Urban Technology, 17*(1), 3–27.

Balk, D., & Yetman, G. (2004). *The global distribution of population: evaluating the gains in resolution refinement*. New York: Center for International Earth Science Information Network (CIESIN), Columbia University.

Ball, P. (2010). Predicting human activity. *Nature, 465*(7299), 692.

Barabási, A.-L. 2010. *Bursts: the hidden patterns behind everything we do, from your e-mail to bloody crusades*: Penguin.

Bhaduri, B., Bright, E., Coleman, P., & Urban, M. L. (2007). LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal, 69*(1–2), 103–117.

Birenboim, A., & Shoval, N. (2015). Mobility research in the age of the smartphone. *Annals of the American Association of Geographers, 106*(2), 283–291.

Candia, J., González, M. C., Wang, P., Schoenharl, T., Madey, G., & Barabási, A.-L. (2008). Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical, 41*(22), 224015.

Csáji, B. C., Browet, A., Traag, V. A., Delvenne, J.-C., Huens, E., Van Dooren, P., et al. (2013). Exploring the mobility of mobile phone users. *Physica A: Statistical Mechanics and its Applications, 392*(6), 1459–1473.

Cho, E., Myers, S. A, & Leskovec, J. (2011). Friendship and mobility: user movement in location-based social networks. In *Paper read at Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1082–1090). San Diego, CA: ACM.

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., et al. (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences, 111*(45), 15888–15893.

Dobson, J. E., Bright, E. A., Coleman, P. R., Durfee, R. C., & Worley, B. A. (2000). LandScan: a global population database for estimating populations at risk. *Photogrammetric Engineering and Remote Sensing, 66*(7), 849–857.

Girardin, F., Vaccari, A., Gerber, A., Biderman, A., & Ratti, C. (2009). Towards estimating the presence of visitors from the aggregate mobile phone network activity they generate. In *International Conference on Computers in Urban Planning and Urban Management*.

Gonzalez, M. C., Hidalgo, C. A., & Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature, 453*(7196), 779–782.

Harvey, J. T. (2002a). Estimating census district populations from satellite imagery: some approaches and limitations. *International Journal of Remote Sensing, 23*(10), 2071–2095.

Harvey, J. T. (2002b). Population estimation models based on individual TM pixels. *Photogrammetric Engineering and Remote Sensing, 68*(11), 1181–1192.

International Telecommunication Union. (2014). *World Telecommunication Development Conference* (WTDC-2014): Final Report. (ITU, Dubai, United Arab Emirates).

International Telecommunication Union. (2015). *ICT facts and figures—the world* in 2015. (http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2015.pdf last accessed on February 6, 2016).

Isaacman, S., Becker, R., Cáceres, R., Martonosi, M., Rowland, J., Varshavsky, A., Willinger, W. (2012). Human mobility modeling at metropolitan scales. In *Proceedings of the 10th International Conference on Mobile systems, applications, and services* (pp. 239–252). ACM.

Kang, C., Liu, Y., Ma, X., & Wu, L. (2012). Towards estimating urban population distributions from mobile call data. *Journal of Urban Technology, 19*(4), 3–21.

Pei, T., Sobolevsky, S., Ratti, C., Shaw, S.-L., Li, T., & Zhou, C. (2014). A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science, 28*(9), 1988–2007.

Ranjan, G., Zang, H., Zhang, Z.-L., & Bolot, J. (2012). Are call detail records biased for sampling human mobility? *ACM SIGMOBILE Mobile Computing and Communications Review, 16*(3), 33–44.

Ratti, C., Frenchman, D., Pulselli, R. M., & Williams, S. (2006). Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design, 33*(5), 727–748.

Reades, J., Calabrese, F., & Ratti, C. (2009). Eigenplaces: analysing cities using the space–time structure of the mobile phone network. *Environment and Planning B: Planning and Design, 36* (5), 824–836.

Sevtsuk, A., & Ratti, C. (2010). Does urban mobility have a daily routine? Learning from the aggregate data of mobile networks. *Journal of Urban Technology, 17*(1), 41–60.

Schwanen, T., & Kwan, M.-P. (2008). The internet, mobile phone and space-time constraints. *Geoforum, 39*(3), 1362–1377.

Shanghai Bureau of Statistics. 2014. 2014 年上海市国民经济和社会发展统计公报 [Shanghai Economic and Social Development Statistical Bulletin 2014]. http://www.stats-sh.gov.cn/sjfb/201502/277392.html (last accessed 15 February 2016).

Silm, S., & Ahas, R. (2010). The seasonal variability of population in Estonian municipalities. *Environment and Planning A, 42*(10), 2527–2546.

Soto, V., & Frías-Martínez E. (2011). Automated land use identification using cell-phone records. In *Proceedings of the 3rd ACM International Workshop on MobiArch* (pp. 17–22). ACM.

Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PLoS ONE, 10*(2), e0107042.

Vieira, M. R., Frias-Martinez, V., Oliver, N & Frias-Martinez, E. (2010). Characterizing dense urban areas from mobile phone-call data: Discovery and social dynamics. In *Paper read at Social Computing* (*SocialCom*), *2010 IEEE Second International Conference on.*

Xu, Y., Shaw, S.-L., Zhao, Z., Yin, L., Fang, Z., & Li, Q. (2015). Understanding aggregate human mobility patterns using passive mobile phone location data: a home-based approach. *Transportation, 42*(4), 625–646.

Xu, Y., Shaw, S.-L., Zhao, Z., Yin, L., Lu, F., Chen, J., et al. (2016). Another tale of two cities: understanding human activity space using actively tracked cellphone location data. *Annals of the American Association of Geographers, 106*(2), 489–502.

Yuan, Y., Raubal, M., & Liu, Y. (2012). Correlating mobile phone usage and travel behavior—a case study of Harbin, China. *Computers, Environment and Urban Systems, 36*(2), 118–130.

Zhao, Z., Shaw, S.-L., Xu, Y., Lu, F., Chen, J., & Yin, L. (2016). Understanding the bias of call detail records in human mobility research. *International Journal of Geographical Information Science*, *30*(9), 1738–1762.

Author Biographies

Yang Xu received his Ph.D. from the University of Tennessee, Knoxville. He is currently an Assistant Professor at the Hong Kong Polytechnic University. His research interests include space-time GIS, human mobility mining and modeling, and urban data analytics and visualization.

Shih-Lung Shaw is Alvin and Sally Beaman Professor and Arts and Sciences Excellence Professor of Geography at the University of Tennessee, Knoxville. He also serves as the Interim Associate Provost for international education at the University of Tennessee, Knoxville. He

received his B.S. degree from the National Taiwan University and his M.A. and Ph.D. degrees from the Ohio State University. His research interests cover geographic information science (GIScience), transportation geography, time geography, GIS for transportation (GIS-T), and space-time analytics of human dynamics. His recent research has focused on space-time analytics of human activities and interactions in a hybrid physical-virtual world based on various types of individual tracking data such as cell phone data, online social media data, vehicle tracking data, travel-activity survey data, and population migration data. His research has led to the development of a space-time GIS for representation, analysis, and visualization of individual activities and interactions in a hybrid physical-virtual space. Dr. Shaw is a Fellow of the American Association for Advancement of Science (AAAS). He also received the Edward L. Ullman Award for Outstanding Contributions to Transportation Geography from the Association of American Geographers (AAG) and served as the Head of the Department of Geography at the University of Tennessee, Knoxville.

Feng Lu holds a Ph.D. from the Chinese Academy of Sciences. He is a Professor at the Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China. His research interests include spatial data management and query, text mining and knowledge graph, and complex network analysis.

Jie Chen received his Ph.D. from the Chinese Academy of Sciences. He is an Assistant Professor at the Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China. Her research focuses on human trajectory data analysis and time geography.

Qingquan Li holds a Ph.D. from Wuhan Technical University of Surveying and Mapping. He is a Professor and President of Shenzhen University, China. His research focuses on the integration of GIS, RS and GPS, intelligent transportation, and urban informatics.