Context-aware Real-time Population Estimation for Metropolis

Fengli Xu, Jie Feng
Tsinghua University
Beijing, China
xfl15@mails.tsinghua.edu.cn

Pengyu Zhang
Stanford University
California, USA
pyzhang@stanford.edu

Yong Li
Tsinghua University
Beijing, China
liyong07@tsinghua.edu.cn

ABSTRACT

Achieving accurate, real-time, and spatially fine-grained population estimation for a metropolitan city is extremely valuable for a variety of applications. Previous solutions look at data generated by human activities, such as night time lights and phone calls, for population estimation. However, these mechanisms cannot achieve both real-time and fine-grained population estimation because the data sampling rate is low and spatial granularity chosen is improper. We address these two problems by leveraging a key insight — people frequently use data plan on cellphones and leave mobility signatures on cellular networks. Therefore, we are able to exploit these cellular signatures for real-time population estimation.

Extracting population information from cellular data records is not easy because the number of users recorded by a cellular tower is not equal to the population covered by the tower, and mobile users' behavior is spatially and temporally different, where static estimating model does not work. We exploit context-aware city segmentation and dynamic population estimation model to address these challenges. We show that the population estimation error is reduced by 22.5% on a cellular dataset that includes 1 million users.

ACM Classification Keywords

H.4.m Information Systems Application: Miscellaneous; H.2.8 Database Management: Database Applications - data mining

Author Keywords

Population estimation; urban computing; context aware computing; mobile sensing; big data.

INTRODUCTION

Achieving real-time and spatially fine-grained population estimation in a metropolitan city is extremely valuable for a variety of applications, including city planning, transportation scheduling and disease control [1]. However, previous solutions [2, 3, 4] cannot support accurate, real-time, and spatially fine-grained population estimation at low cost. Achieving this goal is hard because the first three requirements suggest that we need to sample the population frequently. For example, we can directly count people one by one. However, such solution costs lots of money and time, and cannot be used in practical systems. As an alternative, some previous works look at data collected from people's activities and try to build a linkage between the data and the population [4, 5]. Table 1 summarizes the advantages and disadvantages of these works. None of them can achieve accurate, real-time, and spatially fine-grained population estimation at low cost. Let us discuss each mechanism, which inspires our system and design.

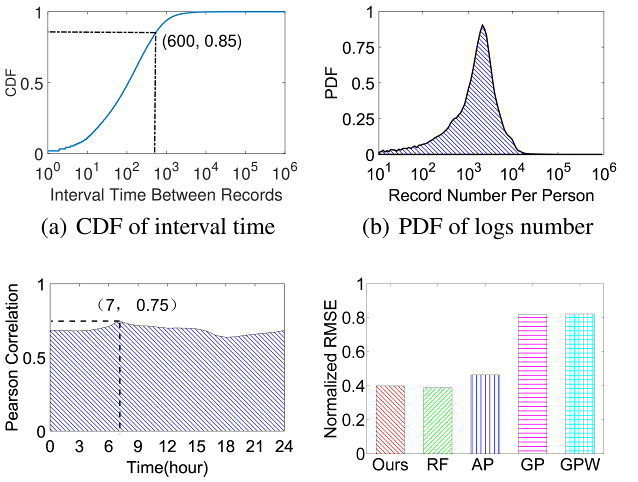
	Accuracy	Real-time	Spatial Granularity	Cost
Population census [2]	High	×	Low	High
Remote sensing [3]	High	×	High	Low
Phone call records [4]	Low	✓	Low	Low
Ours	High	✓	High	Low

Table 1. Comparing the advantages and disadvantages of several mechanisms used in population estimation.

Most of governments carry out regular census because it provides high accuracy even though at a cost of money and high latency [2]. Bhaduri et al.[3] look at images collected by satellites and use remote sensing to estimate population distribution. This method does lower the cost and is able to provide 100m×100m spatial resolution. However, it cannot track the dynamic variation of population distribution during the day because it relies on the mapping between night time lights and population. Thus, it still suffers high latency, since the estimation results can only be produced during night time. Pierre et al.[4] estimate the population by looking at phone call records, and achieve real-time but inaccurate estimation. Its inaccuracy comes from the sparsity between two neighboring phone calls where people may move to another location. As a result, we cannot estimate the population in a particular area accurately at the scale of minutes. Thus, we ask what kind of data source should we use to build the linkage between the data and population?

We look at cellular data accessing records because of three reasons. First, cellular data access happens more frequently, usually at the scale of minutes. It naturally provides finer time-domain granularity in sampling population, and as a result, has the potential to improve the estimation accuracy. Second, cellular infrastructure is ubiquitously deployed, and the distance between two neighboring cellular towers is only 200∼300m in urban area[6]. Therefore, we can leverage this infrastructure to achieve spatially fine-grained population es-

UbiComp '16, September 12-16, 2016, Heidelberg, Germany
© 2016 ACM. ISBN 978-1-4503-4461-6/16/09...$15.00
DOI: http://dx.doi.org/10.1145/2971648.2971673



(a) CDF of interval time (b) PDF of logs number

(c) Correlation between population and mobile user numbers (d) Performance comparing with previous works

Figure 1. Illustration of the quality of our dataset.

timation. Third, cellular data accessing records are logged passively. We do not need to deploy additional software or hardware on user side. Therefore, we focus on cellular data accessing records and investigate how we can estimate population based on them.

Estimating real-time population distribution for metropolitan via cellular data accessing records is not easy because of three reasons. The first challenge is that the number of people accessing the cellular network is not simply equal to the population covered by cellular towers. Many factors contribute to this gap, including disconnected smartphones, people who do not have smartphones, etc. Therefore, we need to model the correlation between the cellular data accessing records and population, and design an algorithm to bridge this gap. Second, it is difficult to find a proper spatial resolution to enable the analysis of physical context behind the estimated population. For example, how many people stay in a residential zone at a given time. Such information is extremely valuable to understand urban mobility patterns, which have the potential to enable various applications, such as transportation scheduling. However, since cellular towers simply segment the city based on their coverage, it is hard to acquire the physical context of population distribution. Third, the data plan usage of mobile users is heterogeneous in spatial and temporal domain. For example, people in residential area use their data plan more frequently compared to people in industrial area, and people use their data plan more frequently during daytime compared to midnight. These complicated user behaviors make it difficult to estimate the population with a simple and static model. Therefore, in order to achieve accurate and real-time population estimation, we need to design a context-aware and dynamic model.

We address above three challenges by designing a novel system. First, our system segments the city based on urban functions regions, which are basic units with one specific function, such as residential area and business district. The division is done by leveraging the roads structure and marking the urban function of each segmented area with points of interest. Instead of using fixed grids that will lose physical context, we choose a segmentation that is able to provide the sweet point of the physical context and spatial granularity tradeoff. Second, our system combines a long-tailed power-law model and an regression model to accurately estimate the population. Third, we enable dynamic population estimation by leveraging a mechanism that is able to model mobile users' heterogeneous behavior in spatial and temporal domain.

Our empirical evaluation on a cellular dataset with 1 million users shows that we are able to achieve real-time and accurate population estimation. Our system reduces estimation error by 22.5% compared with a baseline method that does not consider physical context. We show that such accurate population estimation enables several new applications, including understanding dynamic population migration and scheduling subway transportation.

DATASETS

Now, we introduce the datasets we utilize to enable accurate and real-time population estimation. In addition, we also provide basic visualization to demonstrate their merit.

Mobile Cellular Data Accessing Trace

The utilized dataset is an anonymized cellular data accessing traces collected from the mobile network of Shanghai, one of the major metropolitan in China. It passively records detailed information of cellular data traffic consumed by mobile users, including the ID of mobile devices (anonymized), ID of the base stations, location of base stations, start-end time of data connection and the volume of traffic consumed in each connection. This dataset is large-scale in terms of tracking more than 1,000,000 mobile users on over 9,000 base stations for one month of August, 2014. It provides real-time location information of mobile users with high spatial and temporal resolution, which is ideal for context-aware and real-time population estimation.

To explicitly demonstrate the benefits of utilizing this cellular data, we present several basic visualizations about its characteristics in Figure 1. Subplots (a) and (b) show the empirical Cumulative Distribution Function (CDF) of interval time between two consecutive records and the empirical Probability Density Function (PDF) of the number of records per users, respectively. From the results, we can observe that more than 85% consecutive records happen in less than 600 seconds and most of mobile users have more than 1,000 records in total. In contrast, the average inter-event time of consecutive mobile phone calls is 8.2 hours [1], which indicates that the cellular data accessing records are much more fine grained than the call detail records. These observations demonstrate that our dataset has extensive records of mobile users, and have fine temporal granularity to guarantee the credibility of real-time population estimation. Figure 1(c) presents the correlation coefficient between the number of mobile users and night time population obtained from census during a day. From the results, we can observe that they have a strong correlation during the whole day, especially at around 7AM. This is because most of people stay at home during the night and

Dataset	Resolution	Real-time	Usage	Public Available
Cellular	Base station	√	Estimation	×
Worldpop[5]	$100m \times 100m$	×	Ground truth	√
Census[11]	Admin. area	×	Evaluation	√
Transport[12]	GPS location	√	Evaluation	×

Table 2. The summary of datasets.

begin to become active in the morning. In addition, the maximum correlation reaches up to 0.75, which is much higher than 0.45 reported in call detail records dataset[7]. It shows that the mobile data accessing records we utilize capture population distribution much better than the call records dataset. To further illustrate the benefit of our data, we apply an existing simple estimation model from [4] and cross validate the results with the ground truth data of census. We utilize the normalized RMSE, a widely adopted evaluation metrics in population estimation[4, 5, 8], to evaluate the accuracy of estimation, where smaller normalized RMSE means more accurate estimation. In Figure 1(d), we compare our results with the current state of the art projects that utilizing other datasets, including Asiapop Project (AP)[8], Global Rural Urban Mapping Project (GP)[9], Gridded Population of the World (GPW)[10] and random forest approach (RF)[5]. We can observe that, even though we utilize an existing intuitive model, the estimation obtained from our data is significantly more accurate than those reported in existing works of AP, GP and GPW. These observations further demonstrate that our dataset is ideal for accurate and real-time population distribution estimation.

Ground Truth and Evaluation Datasets

In order to build the population estimation model and evaluate the system performance, we need accurate population data to serve as ground truth and evaluation datasets. To achieve this goal, we utilize data collected from Worldpop[5], census and transport datasets, which are summarized in Table 2 and introduced as follows.

Worldpop[5] provides night time population by exploiting the data collected from multiple sources, including remote sensing images, census and call detail records. With 100m×100m spatial resolution, it achieves most accurate night time population distribution. Thus, we utilize this dataset as the ground truth to calibrate parameters of our system.

The census data[11] is open by Shanghai government, which provides accurate night time population of 188 administration areas and 16 districts of Shanghai in 2014. District is a more fine-grained administration division than the whole city, while the administration area is the most fine-grained administration division. This dataset is used to evaluate the accuracy of our estimation.

The transportation dataset[12] contains detailed records of taxi as well as subway for one month in Shanghai, which records the time stamps and GPS coordinates of citizens transporting via taxi and subway. This dataset is large-scale in terms of tracking over 10 million taxi trips and 100 million subway trips. Since human mobility is closely correlated with the dynamic population distribution, the transportation dataset is utilized to validate the approach of real-time estimation.

Spatial Granularity	Resolution	Accuracy	Physical Context
Adminstration district	Low	High	×
100m×100m grid	High	Low	×
Ours	High	High	√

Table 3. Features of different spatial granularity.

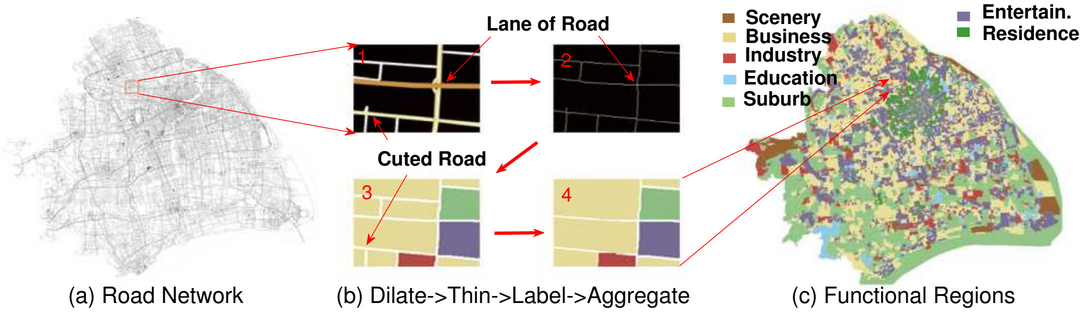
HOW CAN WE ESTIMATE POPULATION?

In order to estimate population from cellular records, we need to answer the following three questions. First, which spatial granularity should we use for population estimation? Second, how to fuse the data collected from multiple sources? Third, how to build a dynamic model to estimate population from the cellular data accessing records?

Spatial Granularity for Population Estimation

When we estimate the population of a specific area, we need to decide how large the targeted area is, which is referred as spatial granularity. However, the spatial granularity of the existing population estimation approaches has several shortcomings, which are summarized in Table 3. Adminstration district is utilized by census[2] and call records approach[4]. Such resolution is too low for urban environment, because each administration district covers a large area, which leads to loosing physical context. On the contrary, a small grid of 100m×100m is adopted by remote sensing approach to segment the city, which is a super-high spatial resolution[5]. However, it also cannot preserve physical context, because urban functional regions are often segmented into different grids. In addition, due to the randomness of human behaviour, such small regions cause high variance in population estimation, which further reduces the estimation accuracy. Therefore, we are motivated to design a spatial granularity that is able to preserve physical context, and yield high resolution and accurate population estimation.

To achieve this goal, we need to first develop a method to segment the city into regions with proper size. Space Syntax[13] is a powerful tool to extract the structure of urban area. However, this method mainly focuses on streets, while in our application the intersection regions are of greater interest, because most of population distribute in the regions. On the other hand, road network, including street, highway, etc., is a natural partition of a city[14, 15]. In addition, segmentation with different level of roads is a flexible approach with fine-grained regions in downtown and coarse-grained regions around suburb, which is closely related to the population distribution. Thus, we utilize the road network to segment the city into different regions. In Shanghai, freeways and city expressways plus the urban arterial roads form a natural segmentation of the urban area as showing in Figure 2(a). Intuitively, we consider each segmented region as a basic smallest unit that carries the context of urban functions by Points of Interest (PoI), which are locations associated with specific functions like residents or shopping malls, and often fall inside these regions where people perform socioeconomic activities, i.e., staying home and going to work.



(a) Road Network (b) Dilate->Thin->Label->Aggregate (c) Functional Regions

Figure 2. The illustration of city segmentation and functional region identification.

ID	Function	Utilized PoI
#1	Residence	residence, life services.
#2	Entertainment	food, hotel, gym, shopping, leisure.
#3	Business	finance, office building, company, trading area.
#4	Industry	factory, industrial estate, economic development zone.
#5	Education	school, campus.
#6	Scenery spot	scenery spot.
#7	Suburb	villages, towns.

Table 4. The utilized PoI categories and taxonomies.

Region\PoI	Resid.	Enter.	Busi.	Indus.	Edu.	Scen.	Sub.
Residence	**0.76**	0.29	0.46	0.01	0.05	0.02	0.01
Entertaiment	0.29	**0.66**	0.38	0.07	0.05	0.04	0.12
Business	0.21	0.24	**0.73**	0.14	0.04	0.02	0.17
Industry	0.09	0.14	0.40	**0.66**	0.03	0.02	0.29
Education	0.14	0.22	0.22	0.08	**0.72**	0.03	0.17
Scenery spot	0.13	0.22	0.19	0.03	0.02	**0.77**	0.11
Suburb	0.06	0.08	0.17	0.10	0.02	0.02	**0.86**

Table 5. The mean value of TF-IDF vectors for every functional type.

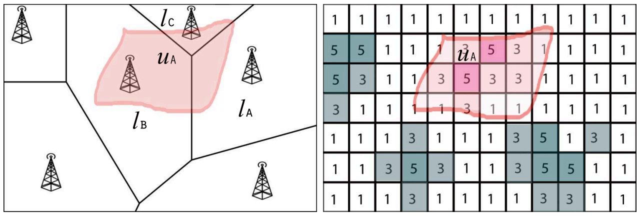
In our system, we utilize an raster-based model to represent the road network and utilize morphological image processing techniques to deal with the problem of city segmentation[15]. In contrast to the vector-based model using geometric primitives such as points, lines and polygons to represent spatial objects, which requires intensive computation when performing topological analysis, the raster-based model based on a binary image, e.g., '0' stands for road segments and '1' stands for blank space, is more computational efficient and brief for road topology analysis. In order to remove the unnecessary details and noise information in the road network for segmentation, such as cuted roads, lanes of a road and overpasses as showing in Figure 2(b-1), we first perform a dilation operation to eliminate all the cuted roads, and then thicken the roads to fill the small holes and smooth out unnecessary details. Finally, we obtain the skeleton of the road networks by performing a thinning operation[16] to recover the size of a region while keeping the connectivity between regions as showing in Figure 2(b-2). The last step is to perform a connected component identification that finds the smallest unit of regions by clustering '1'-labeled consecutive grids.

After segmenting the city into proper regions, we need to identify the physical context, i.e., urban function, of each regions. Cranshaw[17] proposed an approach to discover urban functional regions based on the check-in data in FourSquare. However, the penetration rate of FourSquare is very low in Shanghai, which indicates that the check-in data is insufficient to support accurate identification. On the contrary, the PoIs data can reflect function of a region and usually can be open accessed through the APIs of map service providers, which makes it easier to be generalized to other cities. Thus, we crawl 0.75 million PoIs of Shanghai city from BaiduMap, and calculate the distribution of PoIs in each region unit. To measure the importance of a PoI in one region properly, we utilize a numerical statistic named term frequency inverse document frequency (TF-IDF)[18], which is designed to reflect how important a word is to a document. In our system, it is used to measure the importance of a specific type PoI to its located region. Specifically, for a given region unit $r \in R$, where R is the set of all region unit, the number of PoIs in each PoI category can be counted, and we further calculate a PoI vector, $[\text{TF-IDF}_1^r, \text{TF-IDF}_2^r, ..., \text{TF-IDF}_F^r]$, where F is the number of PoI categories, and TF-IDF_i^r is the TF-IDF value of i-th PoI category in region r, which can be calculated as the following:

$$\begin{cases} \mathrm{IDF}_i^r = \log(R/\ ||\{r|\text{the } i_{th} \text{ PoI category } \in r\}||), \\ \text{TF-IDF}_i^r = n_i^r \cdot \mathrm{IDF}_i^r /\ N^r, \end{cases} \qquad (1)$$

where n_i^r is the number of PoIs belongs to the i-th category in region r and N^r is the total number of PoIs in region r.

Then, we divide them into seven categories according to [19, 15], i.e., residence, entertainment, business, industry, education, scenery spot and suburb, shown in Table 4. These PoIs cover all the major region types that people commute between every day to engage in different socioeconomic activities, e.g., working, shopping, and entertaining. After calculating the TF-IDF_i^r of all regions, we apply a K-means clustering algorithm on the TF-IDF vectors to cluster all the region units into seven urban functions, which are shown in Figure 2(b-3). To achieve the proper spatial resolution, we aggregate the neighboring region units with the same type urban function into a larger region, which is shown in Figure 2(b-4). To avoid too many regions are aggregated and loss spatial granularity, we use the highways to limit the aggregation, i.e., only aggregate the regions within the zone separated by the highways. Finally, we mark the regions of different urban function with different colors in Figure 2(c). To evaluate the effectiveness of our system, we calculate the mean value of TF-IDF across identified urban functional regions and present in Table 5, which shows that all seven clusters are well distinguished with all the diagonal elements as the highest value in the corresponding row and column. We denote the identified



(a) Mapping of mobile users (b) Mapping of Worldpop

Figure 3. Schematic illustration of multi-source data fusion.

urban functional regions as $\{u_1, u_2, ..., u_U\}$, where U is the number of all regions.

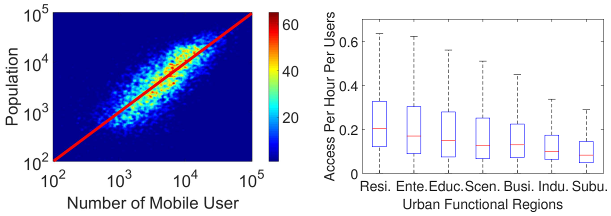
Multi-source Data Fusion

Since our datasets are collected from multi-sources, it is non-trivial to conduct data fusion. To achieve real-time population estimation, we first extract the number of mobile users in each base station at each time slot. Given M base stations $(l_1, l_2, ..., l_M)$ and T time slots $(1, 2, ..., T)$, we define the number of mobile users at base station l_i at time slot t as $\rho_{l_i}^t$. Mobile cellular networks are composed of cells, i.e., geographic zones around a base station. Each cellular connection and data communication can be located by identifying the geographic coordinates of the associated cell. In order to obtain real-time population in the scale of urban functional regions, we need to map the number of cellular users to the scale of regions. According to the locations of all base stations, we first utilize Voronoi diagram to partition the city and obtain their coverage area. Specifically, the Voronoi diagram partitions the areas for each BS as $\{b(l_1), b(l_2), ..., b(l_M)\}$, where any location $p_i \in b(l_i)$ satisfies that for any point $l_j \neq l_i$, the Euclidean distance between p_i and l_i is smaller than that between p_i and l_j. Through this way, we build the Voronoi polygons based on the spatial location of base station. In Figure 3(a), we use the towers to mark the locations of base stations and the black lines to represent the borders of each base station's coverage. Then, we map the number of cellular users, $\rho_{l_i}^t$ for base station l_i at time slot t, into the identified urban functional region, which is the region marked with red color in Figure 3(a). Due to these two different city segment methods, their boundaries of the same area should have intersections and overlaps. Thus, we derive the recorded mobile users of each functional region based on the proportion of Voronoi polygons intersecting that region. In Figure 3(a), the mobile users of functional region u_A is derived by aggregating the mobile users of base stations l_A, l_B and l_C based on the proportion of intersecting area. Specifically, mobile users for urban functional regions u_i at time slot t, denoted by $\rho_{u_i}^t$, is obtained by the following expressions,

$$\rho_{u_i}^t = \sum_{l_j} \rho_{l_j}^t \frac{A(b(l_j) \cap u_i)}{A(b(l_j))}, \tag{2}$$

where $A(b(l_j))$ is the area size of base station l_j's coverage, and $A(b(l_j) \cap u_i)$ is the intersection area of $b(l_j)$ and functional region u_i. Then, we denote the number of users recorded in U urban functional regions at time slot t as the vector of $\rho_u^t = [\rho_{u_1}^t, \rho_{u_2}^t, ..., \rho_{u_U}^t]$.

To fusion the ground truth of Worldpop dataset, we utilize a similar approach to deal with the mobile network data. Specifically, we estimate the night time population of each urban functional regions u_i, denoted as θ_{u_i}, by aggregating the population of the 100m grid squares in the WordProp dataset based on the intersecting area. In Figure 3(b), the population of functional region u_A is derived by aggregating the population of the intersecting squares based on the proportion of intersecting area. More specifically, the expression is presented as follows,



(a) Correlation between number of mobile users and population (b) Mobile network access rate in different functional regions

Figure 4. Characteristics of user behaviour in general and in different functional regions, respectively.

$$\theta_{u_i} = \sum_j \rho_{g_j} \frac{A_{(g_j \cap u_i)}}{100 \times 100}, \tag{3}$$

where $A_{(g_j \cap u_i)}$ is the intersection area of 100m grid square g_j with functional region u_i. Through this way, we obtain the training dataset of each functional regions, i.e., $[\theta_{u_1}, \theta_{u_2}, ..., \theta_{u_U}]$, and are ready to carry out real-time population estimation.

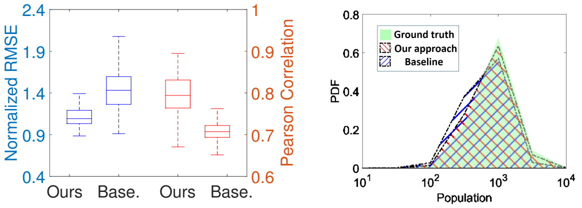
Real-time Population Estimation

In order to estimate real-time population, we first utilize the obtained mobile data accessing recorded users and ground truth for training a static model to estimate the night time population, and then adjust the trained model to estimate real-time population. Since the recorded number of mobile users has highest correlation with night time population at 7AM as shown before, we compare the recorded number of users at the time slot of 7AM $\rho_{u_i}^t (t = 7)$ with the ground truth population θ_{u_i} in Figure 4(a). We can observe that the log-scale mobile users are linearly correlated with the log-scale population, which indicates that a power-law distribution is suitable to model the relationship between the number of mobile users and population. We denote the estimated population of each functional region as $\hat{\theta}_u = [\hat{\theta}_{u_1}, \hat{\theta}_{u_2}, ..., \hat{\theta}_{u_U}]$ and propose an estimation model as follows:

$$\hat{\theta}_u = \alpha(\rho_u^t)^\beta, \tag{4}$$

In this model, parameter α represents the scale ratio and β denotes the superlinear effect of population θ_u on the number of mobile users ρ_u^t. In order to estimate these two parameters, we transform (4) to $\log \hat{\theta}_u = \log \alpha + \beta \log \rho_u^t$. Since Figure 4(a) shows that the linear correlation between log-scale population and log-scale number of mobile users is strong, a standard linear regression model is sufficient to accurately learn parameters α and β. Therefore, we fit the parameters by applying a linear regression on the training data.

Since mobile users' behaviour is correlated with urban functional regions, the static parameters α and β cannot estimate



(a) Evaluation via correlation and normalized RMSE

(b) Evaluation via PDF

Figure 5. Performance evaluation via correlation, RMSE and PDF.

the population across different functional regions accurately. To demonstrate such spatial heterogeneity of behaviour, Figure 4(b) shows that mobile network accessing rate per user differs significantly across different urban functional regions. Specifically, the users in resident area access mobile data network twice more frequently than those in industry area. This motivates us to design a context-aware and time variant model to characterize the spatial and temporal heterogeneity of mobile users' behaviour. To achieve this, we train the data according to different types of regions, and obtain the parameters as $(\alpha_1, \beta_1), (\alpha_2, \beta_2), ..., (\alpha_7, \beta_7)$ for the seven identified types of functional regions.

Utilizing the above parameters with region types, we can estimate accurate population of any functional regions by the number of mobile users, i.e., denoted by $\rho^t_{u_i^j}$ with region type j at time $t = 7$, by the following static model,

$$\hat{\theta}^{t=7}_{u_i^j} = \alpha_j \left(\rho^{t=7}_{u_i^j}\right)^{\beta_j}. \tag{5}$$

Now, we expand our estimation model into a dynamic model to estimate real-time population during one day. To achieve this goal, we need to handle the heterogeneity of mobile users' behaviour in temporal domain. We address this problem by utilizing the factor that β captures the superlinear effect between recorded mobile user number and population, which is a characteristic of different urban functional regions, while α captures the intensity of human activity that varies along with the time. Furthermore, we exploit the fact that the population of Shanghai city does not vary significantly during the day. Therefore, we expand the model into a dynamic one by scaling parameter α while with β fixed. We denote the scaling factor of α as R_t, then the dynamic α^t_j can be computed by,

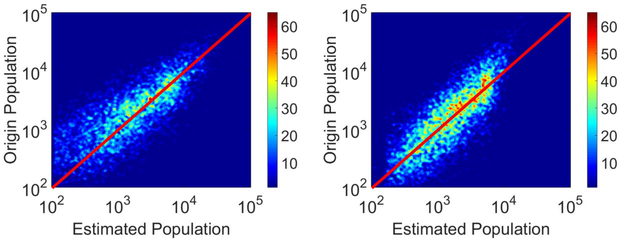
$$R_t = \sum_i \theta_{u_i} \Bigg/ \sum \alpha_j \left(\rho^t_{u_i^j}\right)^{\beta_j}, \qquad \alpha^t_j = \alpha_j \times R_t. \tag{6}$$

Finally, we derive the following dynamic population estimating model as follows,

$$\hat{\theta}^t_{u_i^j} = \alpha^t_j \left(\rho^t_{u_i^j}\right)^{\beta_j}, \qquad t = 1, 2, ..., T. \tag{7}$$

EVALUATION

Now, we evaluate the performance of our population estimation system from two perspectives: quantifying the accuracy of our population estimation and evaluating real-time estimation with transport dataset.



(a) Baseline

(b) Our approach

Figure 6. Visualization of the estimated population.

Cross Validation

We first evaluate our population estimation by adopting correlation coefficient and normalized root mean square error (RMSE) as the metrics, which are widely used in measuring the accuracy of population estimation[4, 5, 8]. We denote the pearson correlation coefficient and normalized RMSE as C and ε, respectively, which are defined as,

$$\begin{cases} \varepsilon = \dfrac{\sqrt{\frac{1}{U}\sum_{i=1}^{U}\left(\hat{\theta}^t_{u_i} - \theta_{u_i}\right)^2}}{\frac{1}{U}\sum_{i=1}^{U}\theta_{u_i}}, \\ C = \dfrac{\sum_{i=1}^{U}\left(\hat{\theta}^t_{u_i} - \frac{1}{U}\sum_{i=1}^{U}\hat{\theta}^t_{u_i}\right)\left(\theta_{u_i} - \frac{1}{U}\sum_{i=1}^{U}\theta_{u_i}\right)}{\sqrt{\sum_{i=1}^{U}\left(\hat{\theta}^t_{u_i} - \frac{1}{U}\sum_{i=1}^{U}\hat{\theta}^t_{u_i}\right)^2}\sqrt{\sum_{i=1}^{U}\left(\theta_{u_i} - \frac{1}{U}\sum_{i=1}^{U}\theta_{u_i}\right)^2}}. \end{cases} \tag{8}$$

The higher C or lower ε indicates more accurate estimation. Since one of our main contributions is taking the urban functional regions into consideration, we define the basic power-law model that does not exploit physical context as the baseline. To evaluate the performance gain of our system, we show the accuracy comparison with the baseline in Figure 5. From the results, we find that our estimation model significantly reduce ε by 22.5% and enhances C by 12.5% on average comparing with the baseline, which indicates that our system produces much more accurate estimation. On the other hand, in Figure 5(b) we present the empirical PDF of the estimated population distribution with the ground truth. Obviously, the PDF of our approach deviates less from the originate data than the baseline, which reveals the underlying reasons why our system achieves more accurate estimation results. However, we still have little knowledge about how do the population and physical context impact the accuracy of population estimation. Therefore, further evaluations should be provided with the population and physical context considered.

To investigate the relationship between estimation error and population, we show the distribution of estimated population and original population as a heatmap in Figure 6, where the red color means more functional regions and blue color means less. From the results, we can observe that the baseline approach tends to estimate less people than the ground truth when the original population is high, and estimate more when the original population is low. On the contrary, the estimated population in our approach always distributed evenly around the ground truth, which explains why our approach produces more accurate estimation in another view. To quantify how the estimation error varies with the population, we classify all the regions into four groups based on their population and

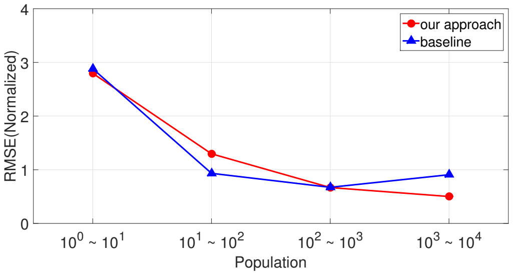
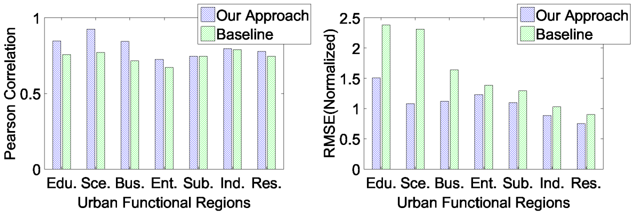


Figure 7. Estimation error varies with population.

show ε of different groups in Figure 7. From the results, we can observe that our system performs a bit worse than the baseline when the population is between 10 and 100, while the performance is similar when the population is between 1 and 10. These differences are mainly influenced by the randomness of human behaviour, which plays an important role when the population is low. On the contrary, the performance of our system is much better than baseline when population is high. Especially when the population is between 1,000 and 10,000, our system reduces ε by 44.7%. Since the regions with high population are more important in estimation, our system performs better in general.

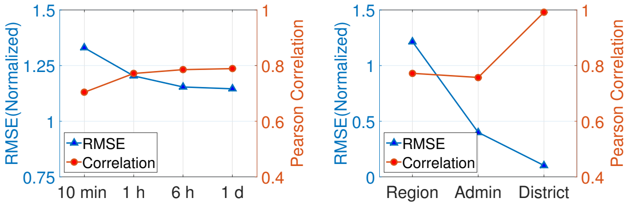
To further evaluate the performance of our system across different urban functional regions, we present normalized RMSE(ε) and correlation(C) of seven different urban functional regions in Figure 8. From the results, we can observe that our system estimates population more accurately than the baseline in every urban functional region. More specifically, in school, park and business areas our system obtains largest performance boost, where the reduce of ε ranges from 31.7% to 53.2% and the improvement of correlation is more than 11.8%. It suggests that in these three regions mobile users' behaviour in cellular network differs significantly than in other regions. Therefore, without considering the urban functional regions, the baseline performs much worse than our system in these areas.

Finally, we evaluate the system's performance trade-off with spatial and temporal granularity. We aggregate the cellular accessing data into different size of time slot, which are 10 minutes, 1 hour, 6 hour and 1 day respectively. Then, we evaluate the night time estimations by cross validating with the ground truth provided by WorldPop and show the results in Figure 9(a). From the result, we find out that the system accuracy improves as the time slot increases, which however decreases the temporal granularity. When time slot increases from 10 minutes to one hour, the correlation is increased by 9.6% and normalized RMSE is decreased by 10.0%. However, when the time slot increases to 6 hours, the accuracy does not improve much and the performance of our system become stable. By utilizing the census data, we can obtain accurate population in administration area and district scales. Therefore, we can utilize the census data to evaluate the accuracy of our system at different spatial granularity, which is presented in Figure 9(b). From these results, we find out that as the spatial granularity decreases, system's performance significantly improves. At the district level, ε is close to 0 and C is close to 1, which means that the estimated population is almost identical with the original population. The underlying reason is that when the predicted regions cover more mobile users, the randomness of human behaviour will have less impacts, which on the other hand shows the scalability of our approach.



(a) Correlation (b) NRMSE

Figure 8. Accuracy evaluation at different urban functional regions.



(a) Temporal dimension (b) Spatial dimension

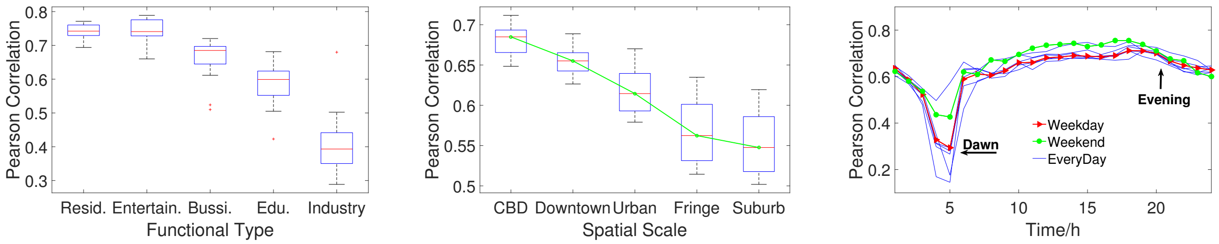
Figure 9. Accuracy evaluation at different granularity.

We evaluated the performance of our system with the baseline, and the results reveal that our system significantly improves the accuracy of real-time population estimation by reducing ε by 22.5% and improves correlation by 12.5% in general. More specifically, the accuracy of our system is related with the population and urban functional regions, and the most significant performance gain is achieved in large population area and education regions. Furthermore, the accuracy is highly correlated with the spatial and temporal granularity.

Cross Validation with Transport Data

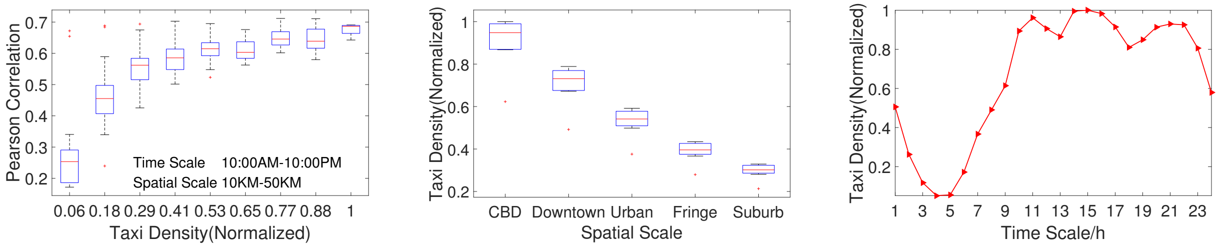
Because of lacking direct ground-truth of real-time population distribution, we utilize the transport data to evaluate the real-time estimation of our system. Transport data records human daily urban transfer and mobility, which directly accounts for the variation of population distribution. Thus, there is a strong relevance between the transport data and real-time population distribution. In other words, if we can obtain an reliable and strong correlation between the transport data and estimated population, the accuracy of our system can be ensured.

Particularly, we use the taxi data that covers most of regions to serve as evaluation dataset. For region u_i in time slot t, we have two vectors: $[\rho^t_{u_1}, \rho^t_{u_2}, ..., \rho^t_{u_U}]$ (the population of each functional region) and $[\gamma^t_{u_1}, \gamma^t_{u_2}, ..., \gamma^t_{u_U}]$ (the transport active level of each functional region, which is derived as the sum of the number of the arriving taxies and the departing taxies). To evaluate the accuracy of our system, we calculate the correlation coefficients between these two vectors, and show them in Figure 10. From the results, we can observe correlations under different conditions of functional regions, spatial scale and time scale. The spatial scale is defined as the radius



(a) Correlation changes with function (b) Correlation changes with space (c) Correlation changes over time

Figure 10. The results of cross validation with taxi data for real-time estimation evaluation.



(a) Correlation changes with taxi density (b) Taxi density changes with space (c) Taxi density changes over time

Figure 11. Observations of taxi density distribution that influences the correlation evaluation.

of the investigated area, which is centered by central business district (CBD). In terms of region function, as showing in Figure 10(a) we find that in function types of residence and entertainment, the correlation reaches up to 0.75, while other function types show low correlation. For example, the correlation in industry area can be as low as 0.3. Inspired by such significant correlation difference, we examine the correlation variation from the downtown to suburb, which is decreasing from 0.7 to 0.55 as Figure 10(b) shows. In addition, time is also a critical factor that influences the correlation. As Figure 10(c) shows, in the evening the correlation maintains at high level, while it reaches to the lowest point near 0.3 before the dawn. Besides, the validation results on the weekend(the green line) and in the weekday(the red line) are also different. In conclusion, high correlation between the taxi user and population estimation is observed for most of time. However, under some conditions such as before the dawn and in the suburbs, we obtain relatively lower correlation. A natural question to ask is what is the underlying reasons for this phenomena?

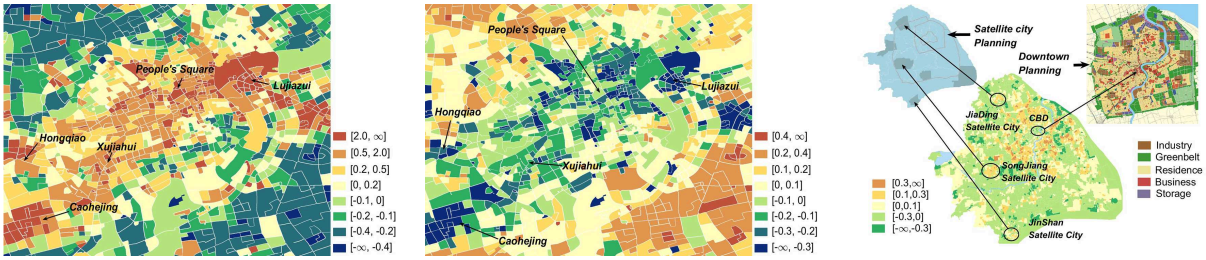
We investigate the underlying reasons in Figure 11. As Figure 11(a) shows, the correlation increases as the taxi density increases. Further, Figure 11(b) and (c) reveal how the spatial scale and time influence the taxi density, which surprisingly show similar changing trend displayed in Figure 10(b) and (c). Observing these results, we find that the underlying reason of obtaining low correlation is the low taxi density, because the taxi active level cannot reflect the population accurately when its density is low. In conclusion, we cannot obtain precise cross validation when taxi density is too low. However, if we cross validate the estimation results with transport data in the proper time (i.e., evening) and spatial scale (i.e., downtown) that guarantee enough taxi density, the correlation between transport data and our population estimation reaches over 0.7, which strongly demonstrates the accuracy of our real-time estimation.

OBSERVATIONS AND APPLICATIONS

In this section, we turn to discuss interesting observations and applications enabled by our real-time and fine-grained population estimation. We demonstrate that the dynamic population distribution with spatial granularity of functional regions and temporal granularity of one hour is sufficient to support the observation of morning-evening rush, monitoring population migration, and recommending locations of new subway stations.

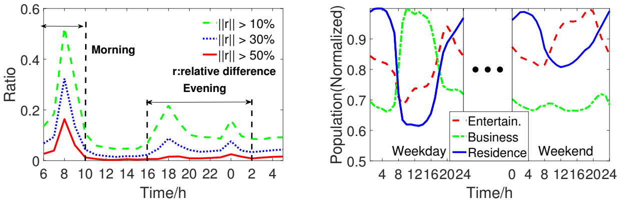
Morning-Evening Rush Visualization

With the fine-grained population estimation, we can obtain population of any functional regions at the temporal granularity of one hour, which enables the observation of urban dynamics, i.e., population migration during the morning and evening rush. We use relative difference as the indicator calculated as $r_{u_i} = (\rho_{u_i}^{t+1} - \rho_{u_i}^{t})/\rho_{u_i}^{t}$. We calculate the relative difference of every region in the period of morning rush (7:00-9:00) and evening rush (17:00-19:00), and quantize it to eight levels coded with different colors, which is shown in Figure 12 with (a-b) visualizing downtown perspective and (c) visualizing overall perspective. Hourly changes are evident in Figure 12(a-b): downtown shows significant increase in the morning while the residence around downtown shows large decrease. Typically, we observe landmarks in Shanghai like Lujiazui and People's Square appearing as the center of population aggregation. Figure 12(c) shows the temporal patterns during the evening rush where residence in the downtown and satellite cities are characterized by a large population increase, while the entertainment and office area undergo significant decrease. From Figure 12(c), we can observe that most of population live in the residence around the downtown, comparing with less population living in the satellite



(a) Morning rush of downtown (b) Evening rush of downtown (c) Evening rush of the city

Figure 12. Population variation in the downtown during the morning and evening rush time.



(a) Migration peak analysis (b) Migration of regions

Figure 13. Population migration patterns.

cities, which shows that the downtown residence has a great capacity of hosting population. Obviously, the population migration is directly corresponding to the location and function, which gives us insights to estimate the population from the perspective of regional context. With the context-aware estimation, we are able to map the population with clear physical meaning, which is one of the major advantages of our work.

Dynamic Population Migration

Context-aware real-time population estimation enables us to investigate dynamic population migration in the metropolis. Figure 13 shows the migration patterns of one day and one week. Specifically, we set three threshold as 10%, 30% and 50% to calculate the proportion of the regions with absolute value of r, denoted by $||r||$, exceeding the threshold per hour, and present the results in Figure 13(a). The curves with different threshold show similar population migration pattern that has three peaks corresponding to the morning and evening commute hours. The first sharper peak at 8:00 belongs to the morning commuting hours that lasts 3 hours. The remaining two lower peaks at 18:00 and 24:00 belong to evening commuting hours that lasts longer to about 6 hours. Thus, we find the morning commuting has shorter duration, which is only half of the commuting hours in the evening. Besides, the volume of morning commuting peak is much higher than the two peaks of evening, which tells us that in metropolis population migration during morning is much more intensity than in the evening. These investigations demonstrate that real-time population estimation enables the potential of obtaining city microcosmic structure and instantaneous dynamics.

Figure 13(b) shows us the daily and weekly population variation from the perspective of functional regions, where we compare the normalized population variations of entertainment, business and residence regions to show the migration routine of working population. As the weekday shows, the population of residence decreases quickly from 7:00 while the population of business increases fall behind it with 0.5 hour with similar duration. In addition, when the population of business decreases from 16:00, the population of residence and entertainment increases. The population of entertainment decreases quickly after the peak at 20:00 while the population of residence increases continually. In the weekend, the population of business remains stable at low level while more people stay at home or in leisure regions. These depict us the typical routine of working population at both day and week level. Enabled by real-time estimation, we are able to monitor the population variation in typical regions and their interactions, which is helpful for a variety of applications like public safety and emergency treatment.

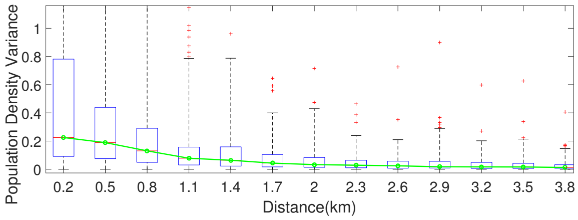
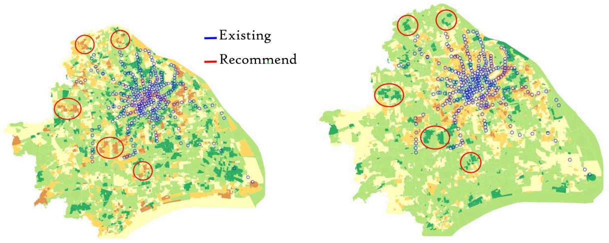


Figure 14. Population variance at different distance from subway.

Subway System Evaluation

We now demonstrate one application enabled by context-aware population estimation — subway system evaluation. In the past, government plans subway system by conducting market survey to obtain population migration patterns. It is usually based on static migration models. With the help of real-time population estimation, we are able to investigate dynamic model to evaluate the location of subway station. Figure 14 shows our evaluation results for the city subway system with the real-time estimated population, i.e., we investigate the relationship between the population density variance with the distance to the nearest subway station. The decreasing tendency of population density variance is obviously observed when the distance increases, which indicates that most existing stations are built reasonably to play an important role in servicing the population migration for nearby regions. Further, we are able to find better subway station locations based on the dynamic population distribution. In terms of morning rush and evening rush, Figure 15 shows several examples of subway station locations, where the blue circles represent the existing subway stations and red circles represent the recommending new subway stations. Most of regions with rapid population changes have existing subway stations. However, regions located by the red circle show rapid population change but without any subway station, which are potential locations of new subway stations. Comparing with the



(a) Morning view (b) Evening view

Figure 15. New subway station location recommendation.

statistic data, real-time population gives us a comprehensive and timely understanding of commuting demand. Places with higher population migration rate, which are more suitable for building station than those just with higher static population density, can be located accurately by our system. All of these helps in the planning and scheduling of transportation system.

Other Potential Applications

In this subsection, we discuss two more applications enabled by our system — anomaly detection and tourist's movement analysis.

Anomaly detection: Rapidly detecting the unnormal gatherings, such as parades and sports events, is of great importance for public safety in the management of modern metropolis. With the help of accurate and real-time population estimation provided by our system, we are able to develop a method to rapidly detect such anomaly events. The basic idea is that we first extract the patterns of population variation in each region, and then compare real-time estimated population with the expectation value of the patterns. When the deviation is larger than a predefined threshold, an anomaly event may happen. In addition, matching the detected anomaly events with the real-world traces can also help us evaluate the accuracy of dynamic population estimation.

Tourist's movement analysis: Tourism is an important business in modern metropolis. Understanding how the tourists move and distribute across the city can significantly benefit the tourist business locating and promoting. Our system is ideal for investigating the movement of tourists, because it can not only passively monitor the location of mobile users but also easily identify the tourists by looking at the duration of their records. Therefore, based on our proposed system, we are able to estimate real-time tourists distribution that can be exploited to investigate the tourist's movement.

RELATED WORK

In this section, we summarize relevant works from two perspectives — population mapping and mobile network data.

Population Mapping: Knowing where people are is a critical social and technical problem. A growing interest in the global mapping of human populations emerged from 1990s [20, 21]. Many researchers use simple area weighting methods [22, 23] or dasymetric modeling approaches [8, 24, 25] to redistribute census population within administrative units. The state of the art of traditional method is the project WordProp, which leverages the remote sensing techniques for inferring the population, but it cannot achieve real-time estimation [5]. With the rapid development of global mobile communication system, many advanced work in this field by using other sources data appear. For example, Deville et al. [4] develop a mechanism that is able to estimate the population density at national scale with mobile phone call records. Ricciato et al. [26] design a method to fuse the mobile phone call records from different operators and estimate the population density at pan-European level. However, these approaches suffers from low sampling rate and consequently loss accuracy. In this paper, we estimate real-time population in urban environment via cellular data accessing logs with high sampling with the aim of capturing the urban dynamics.

Mobile Network Data: The proliferation of mobile phones and networks offers an unprecedented observations and solutions for studying sociospatial behaviors. Mobile network data enables many emerging research fields like human mobility[1, 27, 28, 29], social network[30, 31] and urban planning[32, 33]. In terms of human mobility, Gonzalez et al. [1] and Song et al. [27, 34] find that human trajectories show a high degree of temporal and spatial regularity. From the individual aspect, Isaacman et al. [28] and Bayir et al. [29] develop methods to find important places from the individual trajectory. Combing the cell phone data and online location-based social network data, Cho et al. [31] find that social relationships can accounts for about 20% of human movement. Besides, many researchers investigate the collective response of population to emergency [35, 36]. In addition, Isaacman et al. [37] model how large populations move within different metropolitan areas, Fan et al. [38] predict the collective movement in rare crowd events, and Shimosaka et al. [39] utilize a bilinear poisson regression model to predict mobility flow in the city.

Summary: Different from previous approaches, we study fine-grained real-time population estimation by cellular data accessing logs with short interval and more records in the urban scale, by mapping population with the help of region's context. To the best of our knowledge, these data have not been assessed in their capability of mapping human population at fine spatial and temporal scale, which enables the possibility of context-aware real-time population estimation and guarantees the confidence of our investigation.

CONCLUSION

In this paper, we design, to the best of our knowledge, the first system to estimate context-aware and real-time population distribution via a large-scale mobile data accessing records. Extensive evaluations and analysis reveal that our system reduces the estimation error by 22.5% and show several important observations of urban mobility as well as one application enabled by our system. We believe that our study provides a new angle to achieve real-time and accurate population estimation for metropolis, and paves the way for extensive urban computing applications.

ACKNOWLEDGMENTS

This work is supported by National Basic Research Program of China (973 Program) (No. 2013CB329105), National Nature Science Foundation of China (No. 61301080), and Research Fund of Tsinghua University (No 20161080099)

REFERENCES

1. Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

2. Malaysia. Jabatan Perangkaan. *1980 population and housing census of Malaysia. General report of the population census, Vol. 1*. Jabatan Perangkaan Malaysia, 2005.

3. Budhendra L Bhaduri, Eddie A Bright, and Jerome E Dobson. Landscan: Locating people is what matters. *Geoinfomatics*, 5, 2002.

4. Deville Pierre, Linard Catherine, Martin Samuel, Gilbert Marius, Forrest R Stevens, Andrea E Gaughan, Vincent D Blondel, and Andrew J Tatem. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences of the United States of America*, 111(45):15888–15893, 2014.

5. Forrest R Stevens, Andrea E Gaughan, Catherine Linard, and Andrew J Tatem. Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PloS one*, 10(2):e0107042, 2015.

6. J. G. Andrews, F. Baccelli, and R. K. Ganti. A new tractable model for cellular coverage. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pages 1204 – 1211, 2010.

7. Chaogui Kang, Yu Liu, Xiujun Ma, and Lun Wu. Towards estimating urban population distributions from mobile call data. *Journal of Urban Technology*, 19(4):3–21, 2012.

8. Andrea E Gaughan, Forrest R Stevens, Catherine Linard, Peng Jia, and Andrew J Tatem. High resolution population distribution maps for southeast asia in 2010 and 2015. *PloS one*, 8(2):e55882, 2013.

9. W. R. Tobler, U. Deichmann, J. Gottsegen, and K. Malloy. The global demography project. *National Center for Geographic Information and Analysis (NCGIA)*, 1995.

10. Yetman G. Balk DL. The global distribution of population: Evaluating the gains in resolution refinement. *Center for International Earth Science Information Network (CIESIN)*, 2004.

11. National bureau of statistics of china `http://data.stats.gov.cn/`.

12. Shanghai open data apps `http://www.datashanghai.gov.cn/`.

13. Vassilis Kostakos. *Space Syntax and Pervasive Systems*. 2010.

14. Hector Gonzalez, Jiawei Han, Xiaolei Li, Margaret Myslinska, and John Paul Sondag. Adaptive fastest path computation on a road network: a traffic mining approach. In *Proceedings of the 33rd international conference on Very large data bases*, pages 794–805. VLDB Endowment, 2007.

15. Jing Yuan, Yu Zheng, and Xing Xie. Discovering regions of different functions in a city using human mobility and pois. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 186–194. ACM, 2012.

16. Louisa Lam, Seong-Whan Lee, and Ching Y Suen. Thinning methodologies-a comprehensive survey. *IEEE Transactions on pattern analysis and machine intelligence*, 14(9):869–885, 1992.

17. Justin Cranshaw, Raz Schwartz, Jason I. Hong, and Norman Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. *Social Science Electronic Publishing*, 2012.

18. Mark Steyvers and Tom Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440, 2007.

19. David Mimno and Andrew McCallum. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. *arXiv preprint arXiv:1206.3278*, 2012.

20. John I Clarke. Population geography. *Progress in human geography*, 2(1):163–169, 1978.

21. Uwe Deichmann. *A review of spatial population database design and modeling*. National Center for Geographic Information and Analysis, 1996.

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

23. Uwe Deichmann, Deborah Balk, and Greg Yetman. Transforming population data for interdisciplinary usages: from census to grid. *Washington (DC): Center for International Earth Science Information Network*, 200(1), 2001.

24. DL Balk, U Deichmann, G Yetman, F Pozzi, SI Hay, and A Nelson. Determining global population distribution: methods, applications and data. *Advances in parasitology*, 62:119–156, 2006.

25. Catherine Linard, Marius Gilbert, Robert W Snow, Abdisalan M Noor, and Andrew J Tatem. Population distribution, settlement patterns and accessibility across africa in 2010. *PloS one*, 7(2):e31743, 2012.

26. Fabio Ricciato, Peter Widhalm, Massimo Craglia, and Francesco Pantisano. *Estimating Population Density Distribution from Network-based Mobile Phone Data*. Publications Office of the European Union, 2015.

27. Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

28. Sibren Isaacman, Richard Becker, Ramón Cáceres, Stephen Kobourov, Margaret Martonosi, James Rowland, and Alexander Varshavsky. Identifying important places in people's lives from cellular network data. In *Pervasive computing*, pages 133–151. Springer, 2011.

29. Murat Ali Bayir, Murat Demirbas, and Nathan Eagle. Discovering spatiotemporal mobility profiles of cellphone users. In *World of Wireless, Mobile and Multimedia Networks & Workshops, 2009. WoWMoM 2009. IEEE International Symposium on a*, pages 1–9. IEEE, 2009.

30. Renaud Lambiotte, Vincent D Blondel, Cristobald De Kerchove, Etienne Huens, Christophe Prieur, Zbigniew Smoreda, and Paul Van Dooren. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317–5325, 2008.

31. Eunjoon Cho, Seth A Myers, and Jure Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM, 2011.

32. Jonathan Reades, Francesco Calabrese, Andres Sevtsuk, and Carlo Ratti. Cellular census: Explorations in urban data collection. *IEEE Pervasive Computing*, (3):30–38, 2007.

33. Víctor Soto and Enrique Frías-Martínez. Automated land use identification using cell-phone records. In *Proceedings of the 3rd ACM international workshop on MobiArch*, pages 17–22. ACM, 2011.

34. Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, 2010.

35. James P Bagrow, Dashun Wang, and Albert-Laszlo Barabasi. Collective response of human populations to large-scale emergencies. *PloS one*, 6(3):e17680, 2011.

36. Linus Bengtsson, Xin Lu, Anna Thorson, Richard Garfield, and Johan Von Schreeb. Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in haiti. *PLoS Med*, 8(8):e1001083, 2011.

37. Sibren Isaacman, Richard Becker, Ramón Cáceres, Margaret Martonosi, James Rowland, Alexander Varshavsky, and Walter Willinger. Human mobility modeling at metropolitan scales. In *Proceedings of the 10th international conference on Mobile systems, applications, and services*, pages 239–252. ACM, 2012.

38. Zipei Fan, Xuan Song, Ryosuke Shibasaki, and Ryutaro Adachi. Citymomentum: an online approach for crowd behavior prediction at a citywide level. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 559–569. ACM, 2015.

39. Masamichi Shimosaka, Keisuke Maeda, Takeshi Tsukiji, and Kota Tsubouchi. Forecasting urban dynamics with mobility logs by bilinear poisson regression. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 535–546. ACM, 2015.