CELLREP: USAGE REPRESENTATIVENESS MODELING AND CORRECTION BASED ON MULTIPLE CITY-SCALE CELLULAR NETWORK

By

CHAOJI ZUO

A thesis submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Master of Science

Graduate Program in Electrical and Computer Engineering

Written under the direction of

Desheng Zhang
Yingying Chen

And approved by

_____

_____

_____

New Brunswick, New Jersey

May 2020

ABSTRACT OF THE THESIS

CellRep: Usage Representativeness Modeling and Correction Based on Multiple City-Scale Cellular Network

by Chaoji Zuo

Thesis Directors:

Prof. Desheng Zhang

Prof. Yingying Chen

Understanding representativeness in cellular web logs at city scale is essential for web applications. Most of the existing work on cellular web analyses or applications is built upon data from a single network in a city, which may not be representative of the overall usage patterns since multiple cellular networks coexist in most cities in the world. In this thesis, we conduct a comprehensive investigation of multiple cellular networks in a city with a 100% user penetration rate. We study web usage pattern (e.g., internet access services) correlation and difference between diverse cellular networks in terms of spatial and temporal dimensions to quantify the representativeness of web usage from a single network in usage patterns of all users in the same city. Moreover, relying on three external datasets, we study the correlation between the representativeness and contextual factors (e.g., Point-of-Interest, population, and mobility) to explain the potential causalities for the representativeness difference. We found that contextual diversity is a key reason for representativeness difference, and representativeness has a significant impact on the performance of real-world applications. Based on the analysis results, we further design a correction model to address the bias of single cellphone networks and improve representativeness by 45.8%.

ACKNOWLEDGEMENTS

I would first like to thank my advisor Professor Desheng Zhang and Professor Yingying Chen for their considerate instructions and generous support for my master thesis in the past year.

I would also show my great thanks to Dr. Zhihan Fang for offering many useful advise and help in the experiments and my thesis. It's proud and happy to be your collaborator. Congratulations on your being graduated as a Ph.D. and wish you well in all your undertakings and hope that you find your career a source of great joy and happiness.

In particular, I want to thank Professor Bo Yuan, Professor Jorge Ortiz for serving in my master thesis committee. Also, I would like to thank Zhongze Tang's template.

Last but not least, I am grateful to my friends, my girl friend Fuling Sun and my family for supporting and inspiring me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Background

Cellular services are essential to our daily life for personal communication and mobile web access. Cellular devices have been increasing from 740 million in 2000 to 7,740 million in 2018 in the world [1] as the increase of cellular web users. Understanding the cellular usage patterns in a city is extremely important for cellular operators to provide reliable services such as mobile web access by improve their infrastructures including tower deployment [2], load balancing [3], and network resilience [4]. To date, many efforts have been focused on cellular usage patterns and applications, e.g., traffic patterns [5] [6], user behaviors [7], tower deployment [6], special events [8], and mobility management [9], based on large-scale data collected by cellular operators or small-scale data collected by individual researchers. These studies have provided valuable insights to understand the performance of cellular networks.

However, most of the above work based on large-scale operator-level data is built upon a single network and assumes users and cellular traffic (e.g., web usage) from single network is a representativeness of all cellular users across different cellular networks in a city [10] [11] [12] [13]. Since different networks have different pricing strategies and user coverage, single-network data is potentially biased to represent all cellular users in applications such as web traffic estimation [6]. Even though some studies are based on the data from multiple networks [14] [15] [16] [17] [18], the data are collected at a small scale, e.g., a dozen devices [17], which are not statistically representative of the generic cellular web usage patterns. To our knowledge, none of the existing work has quantified the bias of single network data (e.g., web access log) and its impact on the real-world applications due to

limited data access.

Recently, thanks to the Smart Cities initiative [19], many cities have been consolidating various data from diverse infrastructures [20]. For example, Shenzhen (i.e., the 4th biggest city in the mainland of China and the twin city of Hong Kong) has been consolidating data from its all three cellular networks for innovative smart city services through different data collection mechanisms, e.g., data trading and purchasing [19], which provide an unprecedented opportunity for the research community to improve our understanding of cellular usage behaviors based on all cellular networks in a city.

1.2 Motivation

The user distribution and tower coverage difference in single networks may cause inaccurate models and bias in real-world applications. However, such bias is often ignored in many existing studies such as population estimation [8], web user estimation [21] due to limited data access. To study the impact of single network biases, we first quantify the difference on coverage in different networks and their user difference. Second, we study the performance of applications based on data from different networks.

Root Cause of Bias of Single Network Data: Many data-driven research studies rely on data from single cellular networks, e.g., modeling human mobility based on CDR (Call Detail Records) data from AT&T [11], inferring internet usage in Shanghai [6]. Those studies assume single network data (e.g., web access record or phone calls) is representative of all cellular activities in the same regions. However, single network data is often biased in data-driven applications due to different tower distributions and target user groups among networks. cellular network operators typically have different business priorities in terms of geographic locations, which leads to a significant difference in cell tower distributions [22]. In fact, tower deployment strategies are dependent on various factors such as communication technologies, usage demand, geographic and demographic information in regions [23]. In particular, we found that the tower coverage differs in the three networks,

as shown in Fig. 1.1 when we model tower coverage by Voronoi partition, which is widely used to estimated cell tower coverage boundary [22]. We found a large difference between the tower coverage, which lead to different quality of services and associated metrics (e.g., advertisement, plan rates, etc) for different networks in same city regions, which lead to different numbers of users for each network in the same region. It is the root cause for bias of single network data when used for real-world applications.

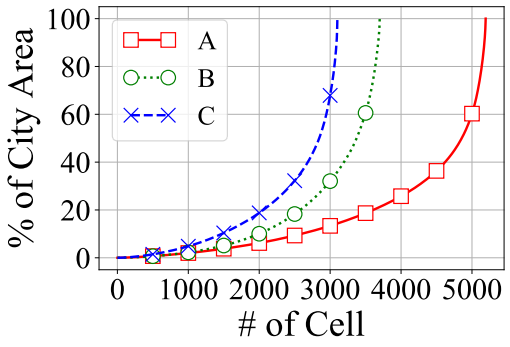


Figure 1.1: Cell Size

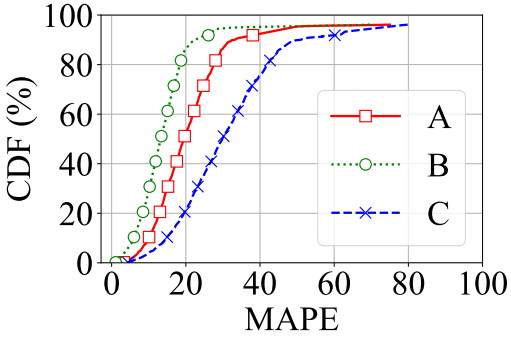


Figure 1.2: Impact of Bias on App.

Impact of Bias on Real-world Applications: Relying on log data records for call, app or Internet service access from three major networks in the Chinese city Shenzhen, we study the impact of data from different networks on a real-world application, which estimates real-time population distribution based on regression models [24] of cellular users. More detailed settings are given in Chapter 6. We use MAPE (Mean Absolute Percent Error) to quantify the performance of the same models with different datasets from three networks. Fig. 1.2 shows CDF distribution of estimation errors on region-level population estimation. We found the performance of the same estimation model differs when using data from different networks. In general, the model based on data from Network B and A show a better performance compared with Network C. The performance difference is caused by different user coverage and usage patterns of networks.

1.3 Overview

In this thesis, we conduct an analysis on cellular network usage **representativeness**, which is defined as *the degree that a single network can be a representative of operational patterns of all cellular users in a region*. The question we want to address in this thesis is ***when***, ***where***, *and to* ***what*** *extent the usage patterns of a given cellular network is biased against the overall patterns of all cellular users across all networks and* ***how*** *we can correct such bias with access only to single-network data.* We infer the overall usage pattern and design quantitative metrics to study cellular network representativeness on multiple diverse networks in the same city. Based on the proposed metrics, we analyze the correlation between representativeness and underlying contextual factors to explore its potential causalities. Our analyses feature large-scale cellular network data for Internet and App access log in Shenzhen, including more than 10 million daily active users from all three cellular networks. The contributions are summarized as follows.

(1). We provide the first investigation on cellular usage representativeness based on multiple diverse cellular networks in the same city. We quantify cellular network representativeness with a distance metric and study the representativeness, its potential causality, and impact on real-world applications. Specifically, we summarize 3 findings and analyze its causality based on real-world contextual data.

- ***finding 1:*** On the spatial dimension, we found that regions with mixed functions such as CBD (Central Business District) area has higher data representativeness compared with regions with single functions such as residential areas.

- ***finding 2:*** On the temporal dimension, we found that the representativeness of a cellular network is highly correlated with user mobility and commuting patterns. We found a 50% lower representativeness during mobility peak hours, e.g., 9am, 5pm, and 8pm, compared with hours with lower mobility demand, e.g., 1pm.

- ***finding 3:*** The performance of population estimation based on single networks is

highly correlated with representativeness. We found a high representativeness leads to a 58.2% lower error of population estimation.

(2). Based on the measurement study and correlation analysis with three contextual datasets (i.e., Point of Interests, Population, and Mobility), we design a learning-based correction model to address data bias in single networks. Further, we evaluate our method based on real-world cellphone web log records from multiple cellular networks covering 100% cellphone users. The results show our method increases the representativeness by 45.8% and then improve the accuracy of population estimation by reducing MAPE from 25.8% to 15.4%.

Moreover, from the correction model, we share our ***finding 4***: Even data from a single network is not a representative of all cellular activities across different networks, with a correction model, 30% of sample data can achieve same representativeness as the data across all networks; 60% of sample data can improve representativeness of a single network by 45.8% on average compared with original single-network data.

1.4 Thesis Structure

The structure of this thesis is organized as follows.

Chapter 1 introduces the background the cellular network, our motivation and findings of human location inference problem.

Chapter 2 discusses related work on cellular web log analyses.

Chapter 3 presents the data we study and the data engineering work to extra the features of our raw data.

Chapter 4 descries our measurement methodology, including the terminologies, metrics and approach in this thesis.

Chapter 5 evaluate our measurement results and findings.

Chapter 6 propose our correlation model to improve the representativeness of a single network.

Chapter 7 draws the discussion.

CHAPTER 2
RELATED WORK

Cellular network is the key infrastructure for Web services. In fact, the trend has been showing that people use their cellular phones for Web services (e.g., Internet Access or App) more often than their phone call [25]. Investigating cellular usage patterns has received considerable attention recently due to data availability. In Table 2.1, we summarize related work based on a two-dimension taxonomy: (i) data collection, i.e., data collected by individual researchers or cellular operators; (ii) investigation scale, i.e., single or multiple networks.

Table 2.1: Cellular Web Log Analyses Survey

Categories		Investigation Scale	
		Single	Multiple
Data Collection Methodology	Individual Researcher	[26] [27] [28] [29] [30]	[14] [15] [16] [18]
	Network Operators	[21] [31] [11] [6] [32] [10] [12] [13] [33]	*CellRep*

2.1 Study on Data from Indivi. Researchers

Many $adhoc$ research projects have various cellular users reporting their data, e.g., locations, web access latency, and signal strength, by installing Apps on cellular devices (e.g., cellphones [30] and connected vehicles [23]). In this approach, researchers obtain detailed data, but the limitation is that the data from a small portion of users and cannot reveal the overall large-scale user patterns.

A Single Network: Given the relatively easier access of single network data collected by individual researchers, lots of work has been proposed to focus on performance and operational patterns of individual networks, such as urban activity inference [30], popular routes construction [26], destination recommendation [27], anomalies spotting [28] and relationships between mobility and PoI [29].

Multiple Networks: Due to the limited accessibility of multiple network data from cellular operators directly, almost all data-driven investigations on multiple networks are limited to small samples of users voluntarily contributing their data from their devices at an application level, e.g., inter-city mobility of Skout users [18], location prediction [15], urban planning based on location-based social network [16], and existing PoI verification [14].

2.2 Study on Data from Network Operators

Cellular network operators passively collected their network data for billing purposes (e.g., CDR data [11]) or web access logs (e.g., internet access data [34]). Compared with detailed data collected by individual researchers, the data collected by cellular operators typically cover all users for a network, yet with coarse granularity on spatial and temporal dimensions.

A Single Network: Extensive studies have been conducted with cellular data for various applications. For example, Call Detail Records (i.e., CDR data) for phone calls or data connection records for data calls are commonly used to model human mobility at a metropolitan scale [11] [35]. Based on cellular data from a single network, researchers (i) conduct spatiotemporal phone call analysis[10], data call analysis [13], mobile traffic analysis and prediction[10] [4] [6], and dynamic urban geo-social connectivity graph construction [21] (ii) trajectories recovery from mobility data [31], (iii) determine the locations of network upgrades [33], and (iv) improve network performance [32]. However, the above work is based on a single network in one city, which may not be representative of usage patterns of all cellular users across different networks.

Multiple Networks: To our knowledge, we conduct the first effort to investigate the usage patterns of all cellular networks in a city. Compared with previous studies in other three categories, we advance the understanding on the usage patterns of multiple diverse cellular networks.

CHAPTER 3
DATASET

3.1 Cellular Networks Datasets

We have been collaborating with the Shenzhen smart city team for data access to all three cellular networks for one month. For privacy and security issues, we use Network A, B, and C in this thesis, instead of using the company name and detailed time. Network A has the largest number of towers, followed by Network B and Network C. Since China only has three cellular service vendors, the dataset achieves 100% penetration rates for cellular devices. Here are the basic information of these three networks.

- Network A deploys 5174 towers serving 3.9 million users;
- Network B deploys 3595 towers serving 3.8 million users;
- Network C deploys 2977 towers serving 2.5 million users.

Even though three networks have different data formats, we reorganize the data to obtain the data log records with five essential attributes including *user ID, timestamp, longitude, latitude, type (e.g., a data call for Internet access)* where the longitude and latitude give the tower location. For example, Network B generates 24.5 million records for a daily data log of voice calls and 185.5 million records for a daily data log of data calls (e.g., app or Web service access). We drop other fields for the minimum data exposure. More details on privacy and ethical issues are given in the Chapter 7 . We show their tower coverage with Voronoi partitions [22] in Fig 3.1.

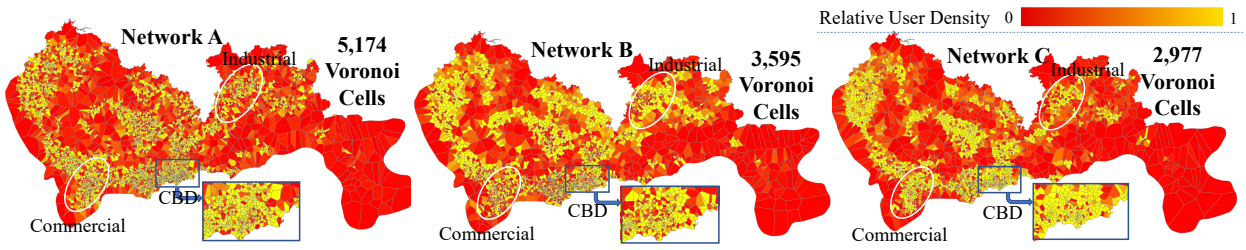


Figure 3.1: User Density in Tower-based Voronoi Diagram

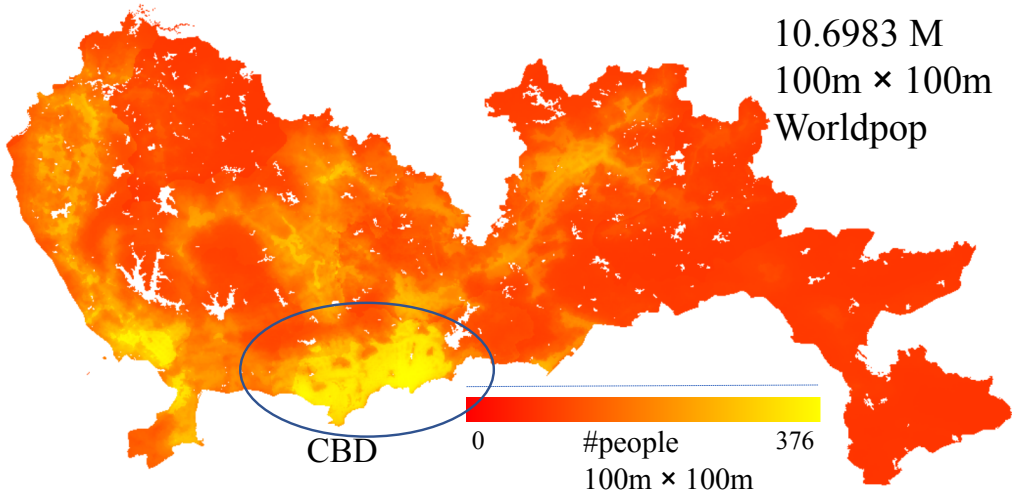


Figure 3.2: Population Distribution

3.2 Contextual Datasets

To study the bias of cellular networks in context, we focus on three most important contextual data during the same period: (1) the total potential users, i.e., population, (2) the reason for a user to use cellular services, i.e., Region Functions with Point of Interests; (3) the physical movement of cellular users, i.e., mobility.

(i) Population: We extract Shenzhen population from Worldpop [36], which gives fine-grained population distribution in $100m \times 100m$ grids. Fig. 3.2 presents the population distribution and statistics in Shenzhen where the CBD (central business district) has a higher population density than other areas. We map population into administrative regions and calculate the population density in regions to study the impact on representativeness in Fig. 5.5.

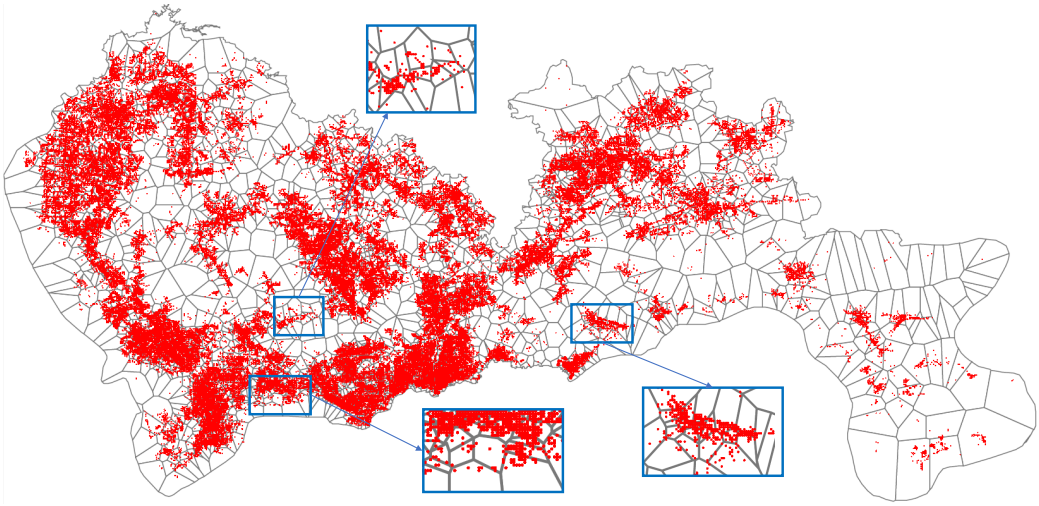


Figure 3.3: Point of Interests

(ii) Point of Interests (PoI): The function of regions is one important factor to determine the spatial cellular patterns [37]. For instance, more web access are made in the downtown Central Business District (CBD) during daytime compared to some residential areas; whereas the nighttime may have a reverse pattern. To quantify region functions, we collect 542,115 PoIs in Shenzhen from an online map service provider. The PoIs are mainly categorized into 5 groups (i.e., residential, office, education, transportation and recreation), and 17 subgroups (i.e., traffic facilities, education, fitness, auto services, culture and media, business, life services, food, tourist attractions, government organizations, shopping, hotels, recreation, medical services, real estates, beauty & spas, finance). Fig. 3.3 visualizes PoIs on a map based on the Voronoi cells of the Network B. We expect that regions with PoIs from different categories may have different web access in cellular networks and lead to a difference of representativeness.

(iii) Mobility: We study cellular operational patterns with three urban-scale mobility datasets in Shenzhen, i.e., (i) a subway system with 8 lines, 194 stations and 4 million users, (ii) a bus system with 1,115 lines, 10,106 stations, 13 thousand buses and 5 million bus passengers, (iii) a taxi system with 15 thousand taxis and 500 thousand passengers, and (iv) a personal car system with 10,043 personal cars which are collected for insurance purposes,

for a correlation analysis on cellular patterns and urban mobility. Fig. 3.4 shows mobility of bus, taxi, subway, and car users, which are also the cellular users given the high penetration rates of cellular services. A lighter color indicates a higher density.

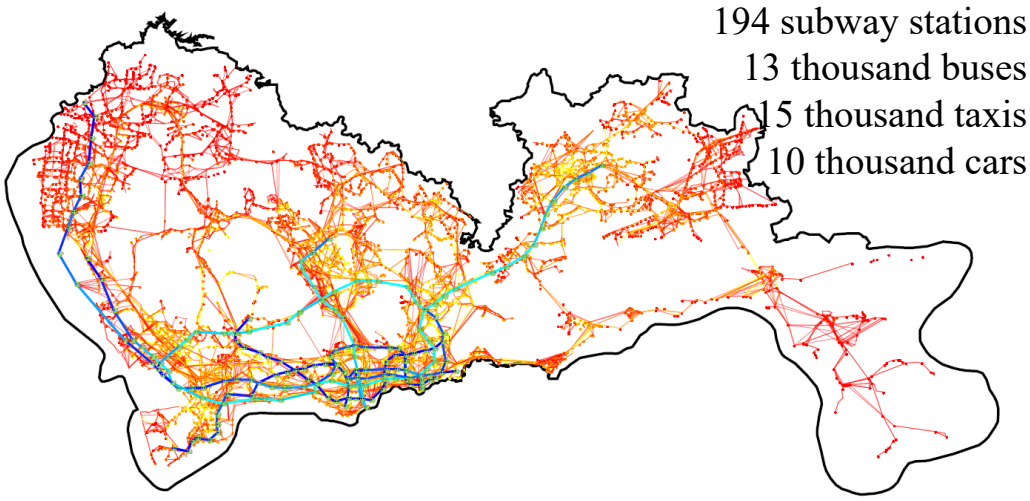


Figure 3.4: Mobility Distribution

CHAPTER 4
MEASUREMENT METHODOLOGY

4.1 Terminologies

We use a lowercase letter for a number, e.g., l presents the number of data records at a specific location during a time period, and a uppercase letter for a collection, e.g., L is a vector of l as a distribution. In general, we have three factors to aggregate loads, i.e., spatial, temporal and networks. We summarized terminologies in Tab. 4.1.

Notation	Meaning
P	a spatial partition
r, R	a region and a region collection
t, T	a time slot and a time slot collection
k, K	a network and a network collection
l, L	a load and a load collection/distribution
$\tilde{L}$	normalized load distribution
$l_k^{r,t}$	a load of network k at region r in time slot t
$L_K^{R,T}$	a load collection given K, T and R
ϵ	a tolerant threshold for representativeness

Table 4.1: Terminologies

Spatial Partition: We introduce two spatial partitions to show the bias of individual networks in a city, i.e., a network-specific tower based partition and a network-agnostic census-based partition. For a single network, a tower partition is generated by a Voronoi graph [22] to estimate the coverage of a tower. The census-based partition is released by city governments according to their road distribution and population distribution. Specifically, Shenzhen has a census-based partition including 491 regions as shown in Fig. 4.1, which shows the dominant cellular network (with most users) for each region, and the average size of regions is 4.06 km^2. Therefore, tower-based partitions are dependent on tower locations in single networks. Instead, since census partition is independent from cellphone

networks, we compare load distribution of different networks under the census-based administrative regions.

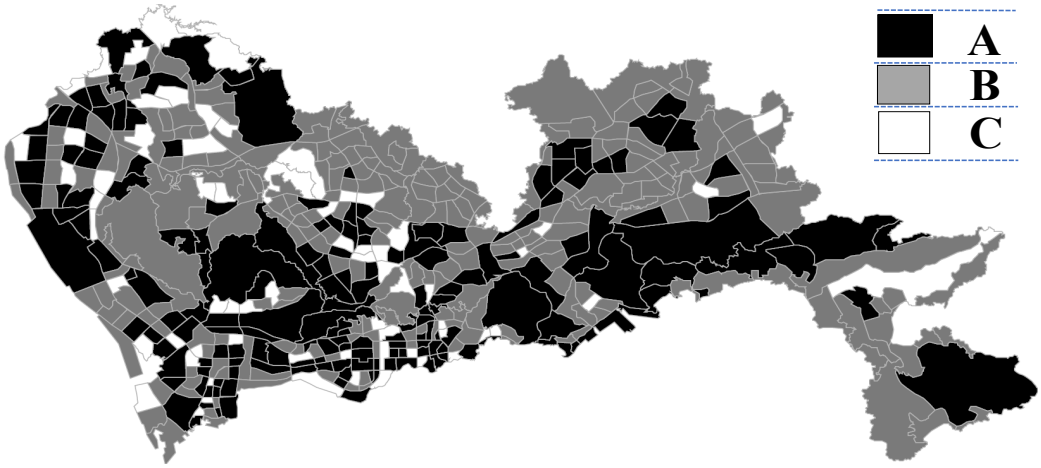


Figure 4.1: Census-based Partition

Temporal Partition: We partition time into 10-minute time slots. In other words, we calculate calculate load for every 10 minutes. As a result, one day is divided into 144 time slots. The 10-minute slot length has been extensively used in various cellular network studies [22] [6] [24].

Load Distribution: The number of phone calls or internet calls is described as *load*. $\mathbf{L} = l_1, l_2, \ldots l_n$ represents the load distribution where l_1 to l_n is the load in a specific region in a specific time slot. As we introduce in table 4.1, we use a subscript k to differentiate loads from different networks, e.g., $\mathbf{L}_k$; we use $\mathbf{L}_\forall$ for loads of total loads in a city by combining all cellphone networks. We use a superscript r and t for load distribution in a region, e.g., $\mathbf{L}^r$, or at specific time, e.g., $\mathbf{L}^t$.

4.2 Measurement Metrics

Representativeness Distance: Intuitively, $\mathbf{L}_i$ is a representative of $\mathbf{L}_\forall$ if $\mathbf{L}_i$ can be scaled to $\mathbf{L}_\forall$ by a scaling factor α. Similarly, to study if a network can be used as a representative of all networks, we use *Representative Distance* θ_k $(0 \leq \theta_k \leq 1)$ which is the maximum norm of the difference between the total load distribution and the scaled load distribution

of network i at region r during the same time slot as in Equation (4.1).

$$\begin{aligned} \theta_k &= \min_{\alpha} \| |\tilde{\mathbf{L}}_{\forall} - \alpha \tilde{\mathbf{L}}_k| \|_{\infty}; \\ \tilde{\mathbf{L}} &= \frac{\mathbf{L} - min(\mathbf{L})}{max(\mathbf{L}) - min(\mathbf{L})};\ \mathbf{L}_{\forall} = \sum_{k=0}^{K} \mathbf{L}_k; \end{aligned} \tag{4.1}$$

We illustrate our idea in the example with a single network $\mathbf{L}_A$ load distribution and the total load among all networks $\mathbf{L}_{\forall}$ in Fig. 4.2. $\mathbf{L}_A$ and $\mathbf{L}_{\forall}$ represent the load distribution during one day for Network A and all three networks, respectively. (1) we normalize both distributions as in the left figure and calculate the maximum norm between the two distributions. (2) we tune a scaling factor α to search for the minimum values of the maximum norm between the two distribution, which is denoted as the representativeness distance between the two distributions as shown in the right figure.

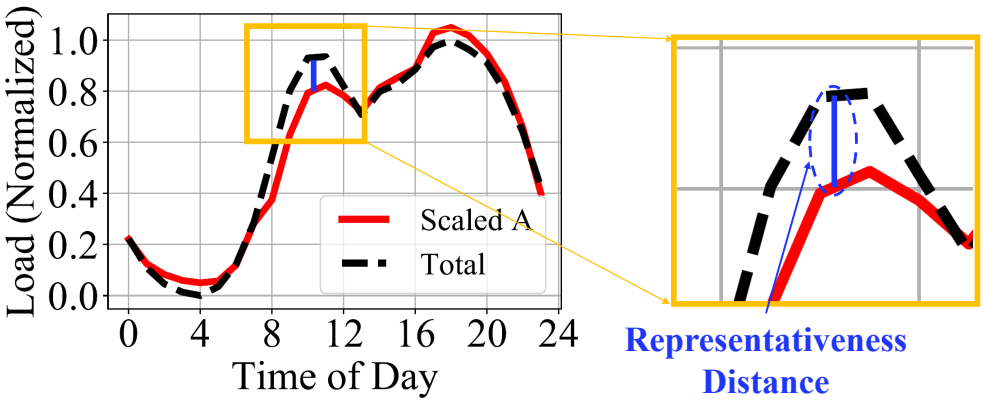


Figure 4.2: Representativeness Distance

We use the maximum norm for two reasons. First, it measures similarity and preserves pair-wise comparison between two load distributions. The pair-wise comparison is important since it measures the representativeness under same spatial-temporal dimension, e.g., l_i in $\mathbf{L}_{\forall}$ and l_i in $\mathbf{L}_A$ describe the load in same region at same time slot. In contrast, other statistical features, e.g., average or similarity, are aggregated results and may ignore the difference between two pairs. Second, it measures the upper bound of the difference between two load distributions and therefore it is a more strict measurement than aggregated

value such as mean and similarity. The upper bound means that difference between loads in the two load distributions is guaranteed to be smaller than the representative distance. In other words, a low value of θ leads to a low value of similarity or mean difference but not vice versa.

Tolerant Parameter ϵ. We define a tolerant parameter ϵ for representativeness. A network k is a representative of all networks if the representativeness distance $\theta_k \leq \epsilon$. Based on load distributions of Network A, B and C, we categorize 491 administrative regions as shown in Fig. 4.1 or time slots into 3 groups: *(i) Total Representative Regions/Time Slots (TR)*, the regions/time slots where every network is representative; *(ii) Partial Representative Regions/Time Slots (PR)*, the regions/time slots where we can find at least one representative network but not all networks; *(iii) No Representative Regions/Time Slots (NR)*, the regions/time slots where no network is representative.

4.3 Measurement Approach

We conduct our study on representativeness from three perspectives:

(1). *Findings and Causalities (Chapter 5.1 and 5.2):* we categorize the measurement results on spatial, temporal dimensions. (i) On the spatial dimension, we study load distribution $\mathbf{L}^r$ for different r. For each region r, $\mathbf{L}^r$ describes loads at different time slots. (ii). On the temporal dimension, we study load distribution $\mathbf{L}^t$ for different slots t. For each time slot t, $\mathbf{L}^t$ describes loads in different regions, i.e., elements l in $\mathbf{L}^t$ are loads from different regions with same time slot t. To better understand the potential reasons for representativeness difference, we study the correlation between representativeness and different factors such as population distribution (i.e., how many potential users); mobility (i.e., will these users change?); point of interest distributions (i.e., why they use cellular there?).

(2). *Case Study (Section 5.3):* we select 4 regions with different contextual information distribution for a detailed study to validate our findings.

(3). *Correction (Chapter 6):* we design a diversity-driven model to alleviate the impact

of representativeness distance with single-network data and public contextual data. We evaluate our correction model with two real-world applications, i.e., population inference and mobility modeling, by studying their performance with corrected representativeness.

CHAPTER 5

MEASUREMENT RESULTS

5.1 Spatial Representativeness

Overall Patterns: Fig. 5.1 shows *Representative Distance* θ distribution of three networks in administrative regions. A lower representative distance indicates high similarity between loads (e.g., cellular traffic on web access) in a single network and loads of all networks. We found the load of Network B is the most similar to the load distribution of all users across all networks. One possible reason is that the load patterns of Network A and C are complementary, while the load pattern of Network B is close to the overall load pattern in the city. Based on the representative distance of three networks, we study the regions in the three groups with different ϵ in Fig. 5.2. When the threshold ϵ increases, the number of total representative (TR) regions increases; the number of no representative (NR) regions decreases; The number of partial representative (PR) regions increases first and then decreases.

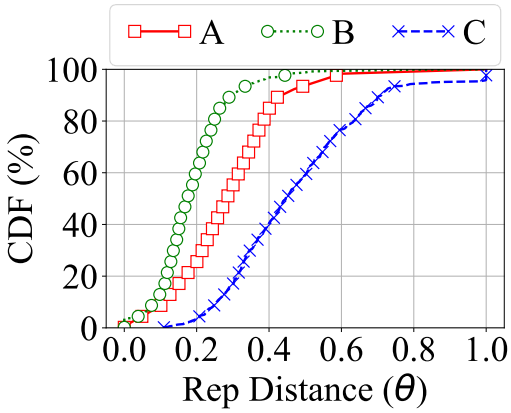


Figure 5.1: Regions

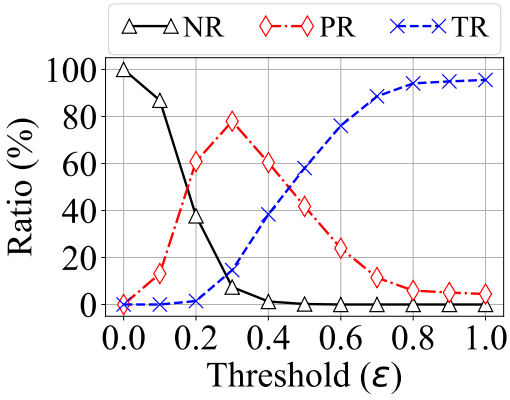


Figure 5.2: Spatial Groups

Impact Factors: To further explain the representativeness of regions in these three groups, i.e., TP, PR, and NR, we study user distribution and their usage patterns in these regions, which are closely related to two types of features, i.e., static features of the regions

(e.g., functions and population) [22] and dynamic features of the users (e.g., mobility) [35]. For example, there are more business activities and users in CBD areas, who prefer the cellular networks with better quality and are more tolerant on costs; whereas college students in educational regions are more sensitive on costs.

Therefore, we take both PoI (points of interests) and static population distribution into consideration for potential reasons for representativeness difference. However, those static features are not sufficient to capture dynamic user distributions since users are moving between different regions during different time of day. Therefore, we introduce a dynamic feature, i.e., user mobility, to analyze its correlation with cellular representativeness. As a result, we study these static and dynamic features as three contextual impact factors, i.e., Point of Interest (PoI), population, and user mobility, which are used to investigate their impact on representativeness in regions to explore the underlying reasons for representativeness differences.

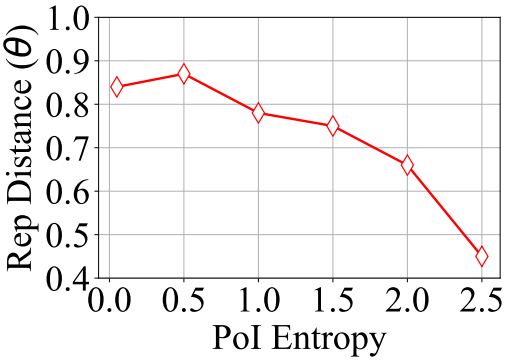


Figure 5.3: PoI v.s. Spatial

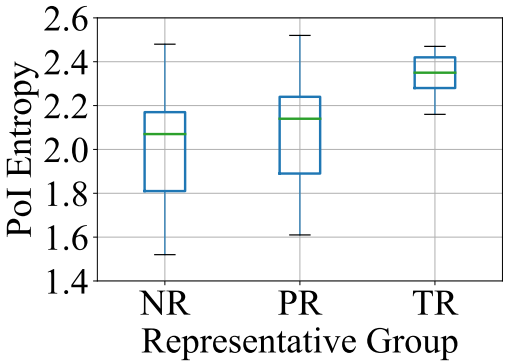


Figure 5.4: Groups v.s. Spatial

Impact Factor 1: Region PoI. For each administrative region in Fig. 4.1, the PoI distribution is described by a 17-dimension vector from 17 subgroups. Since *entropy* is widely used to measure the randomness and diversity of a certain distribution. We study *PoI entropy*, as in Equation (5.1) where x is a 17-dimension vectors and each element x_i is

number of PoIs in one subgroup.

$$H(X) = -\sum_{i=1}^{17} p(x_i) log_2 p(x_i) \tag{5.1}$$

We found that a higher PoI entropy leads to a lower representative distance as in Fig. 5.3. In other words, in the regions with more diverse PoI distributions, the load distribution of a network is more similar to its total load distribution. We further validate this observation in Fig. 5.4 and Tab. 5.1. In Fig. 5.4, we set ϵ as 0.2 to categorize all 491 regions into three groups, i.e., NR (No Representative group), PR (Partial Representative group), and TR (Total Representative group). We found that a high entropy (i.e., more diverse distribution of PoI) in both TR and PR, compared with NR. We give the detailed PoI distribution in

Table 5.1: PoI Distribution in Groups

Group	Cluster				
	Residence	Transport	Office	Recreation	Edu
TR	0.18	0.23	0.21	0.22	0.16
PR	0.14	**0.26**	**0.28**	0.20	0.12
NR	**0.30**	0.18	0.13	0.18	0.21

Tab. 5.1 where we found that (i) the most PoI distributions in TR and PR regions are dominated by the function of Transportation and Office, and (ii) NR regions are dominated by the residence.

Impact Factor 2: Region Population. We extract Shenzhen population from Worldpop [36], which gives fine-grained population distribution in $100m \times 100m$ grids. Fig. 3.2 presents the population distribution and statistics in Shenzhen where the CBD (central business district) has a higher population density than other areas. We map population into administrative regions and calculate the population density in regions to study the impact on representativeness in Fig. 5.5. In regions with high populations, the representative distance is small, which indicates that a single network is more representative in cellular users in these regions.

Impact Factor 3: Region Mobility. We quantify the user mobility of one region by

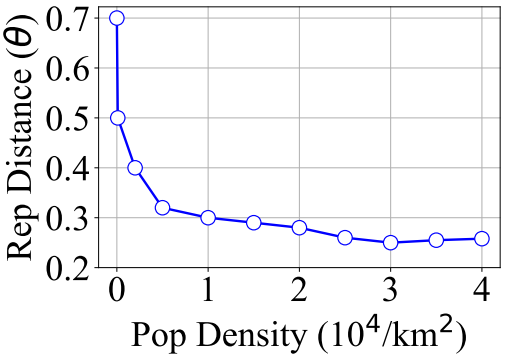


Figure 5.5: Pop. v.s. Spatial

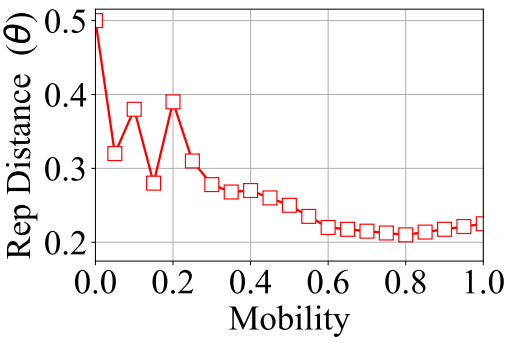


Figure 5.6: Mobility v.s. Spatial

its mobility demand, which is quantified by the number of trips starting from r_i inferred from the four transportation systems as introduced in Section 3.2. To eliminate the impact of region sizes and populations, we use mobility demand index, which is defined as the ratio between mobility demand and population in Fig. 5.6. We found that a high mobility demand index (i.e., a high percentage of moving population) decreases the representative distance. In other words, it increases the representativeness of a single network.

5.2 Temporal Representativeness

Daily Pattern: As shown in Fig. 5.7, we found a lower representativeness distance in Network A and B, but a higher representativeness distance in Network C. All networks show similar patterns including three peaks around 9-10am, 4-5pm, and 8-9pm. Similarly, on the temporal dimension, we study three representativeness groups, i.e., TR (Total Representativeness), PR (Partial Representativeness), and NR (None Representativeness), in Fig. 5.8. Compared with spatial representativeness groups as in Fig. 5.2, temporal representativeness groups present a lower representativeness thresholds. It indicates the spatial dimension has a higher variance of representativess, which motivates us to correct the representativeness mainly from the spatial dimension in Chapter 6.

Impact Factors on Daily Pattern: We analyzed both network and contextual data to study the potential reasons and impact factors on the daily representativeness patterns. We mainly

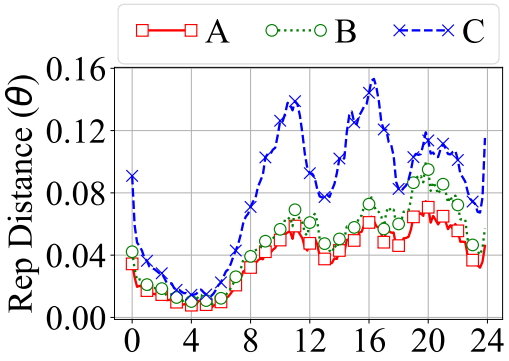


Figure 5.7: Time of Day

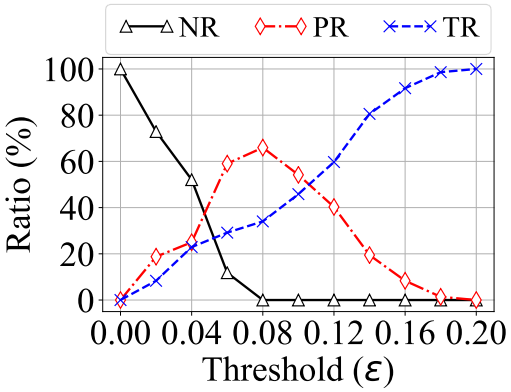


Figure 5.8: Temporal Groups

show the results on user mobility since it is the most important dynamic contextual factors on the temporal dimension compared to population and PoI distributions, which are static features related to spatial distribution of regions. We calculate the entropy of daily origin-destination pair distributions of all taxi and public transportation (i.e., bus and subway) trips based on the mobility data introduced in Section3.2. A lower entropy indicates a less random (i.e., less diverse) distribution of user mobility as shown in Fig. 5.9. In other words,

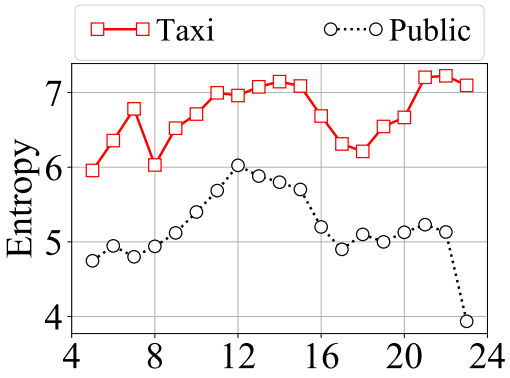


Figure 5.9: Mobility v.s. Temporal

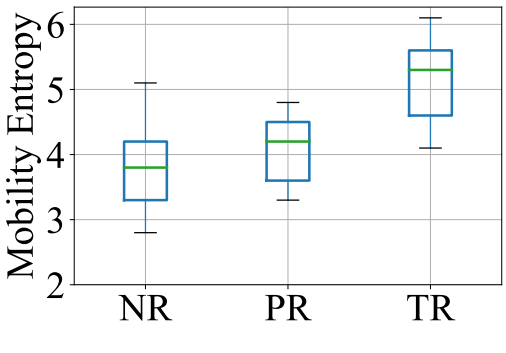


Figure 5.10: Groups v.s. Temporal

most passengers are mainly moving from certain origins to destinations, i.e., from residential regions to office regions or vice versa. We study the impact of mobility entropy on the representativeness by showing the average mobility entropy of three groups in Fig. 5.10. We found the highest mobility entropy in the TR (total representative) group and the lowest in the NR (no representative) group. It suggested that the low diversity of mobility poten-

tially leads to a high representativeness distance, which may be because most passengers are moving between high-demand regions.

(ii) Weekly Pattern: We further study weekly patterns of representativeness as shown in Fig. 5.11. We found a larger representative distance during weekdays than weekends. Besides, the representativeness distance is relatively flat during the day time of weekends. Compared with non-peak time segments, the representative distance is larger in peak segments. Similar to daily patterns, the representativeness difference is potentially caused

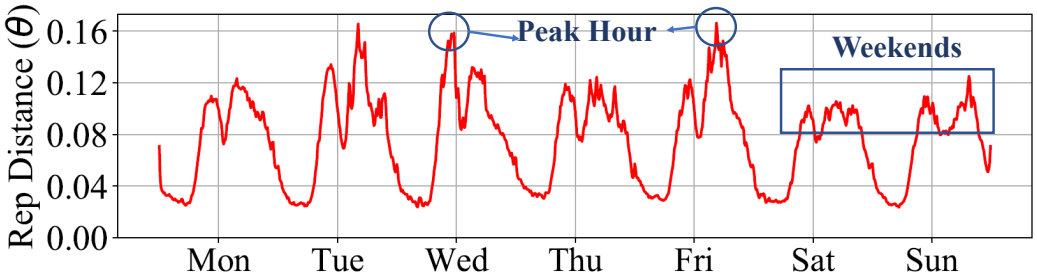
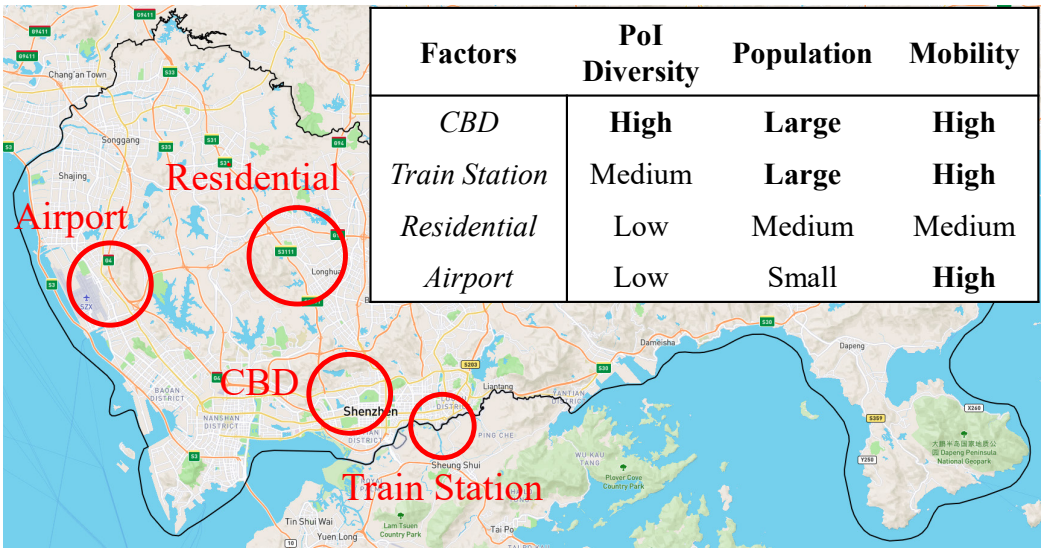


Figure 5.11: Average Representative Distance in One Week

by the user mobility difference. For instance, the mobility traces are more random during weekdays compared with weekends. Due to space limitation, we omit the detailed analysis.

5.3 A Case Study



Factors	PoI Diversity	Population	Mobility
CBD	**High**	**Large**	**High**
Train Station	Medium	**Large**	**High**
Residential	Low	Medium	Medium
Airport	Low	Small	**High**

Figure 5.12: Case Study Areas and Their Contextual Diversity

To dive deeper on the spatial and temporal representativeness, we conduct a case study in four selected regions, i.e, two transportation centers (including the city train station and the airport), the CBD area, and a residential area, which are labeled in Fig. 5.12. We select the four regions for two reasons: (i) they are the most important regions for most cities; (ii) they have very diverse distributions in terms of the contextual factors including PoI, population, and mobility as shown in the Table in Fig. 5.12. The number of towers in the four selected locations with a certain radius is given in Tab. 5.2 from the least number of towers to the most number of towers.

Table 5.2: Tower Distribution on Select Locations

Radius	1 km			2 km		
	A	B	C	A	B	C
airport	2	2	1	17	20	11
residential	25	17	13	84	55	47
train station	38	30	27	134	109	101
CBD	58	56	30	164	178	92

We compare their representative distances in Fig. 5.13, where we found the highest representative distance (i.e., less representative) in the airport area and the lowest representative distance (i.e., more representative) in CBD. It confirmed our previous observations in Section 5.1 that a lower contextual diversity in terms of PoI, population and mobility leads to a larger representative distance, which make a region less representative. For an in-depth study on contextual diversity, we further study the impact of geographical distances from the center of these areas on representativeness in Fig. 5.14. A long distance to the area center (i.e., a larger area with a larger radius) decreases the representativeness distance of the area because it mainly increases its contextual diversity. However, we found that the impacts of distances on four areas are different: the representativeness only decreases slightly around the CBD region; whereas the representativeness decreases significantly around the train station, airport and residential regions. This is because the nearby regions around the CBD area is still downtown so the contextual diversity does not change much with the increasing of geographical distances from the CBD center; whereas the nearby regions

around airport, train station and residential areas have higher contextual diversity with the increasing of geographical distances since they include more diverse regions.

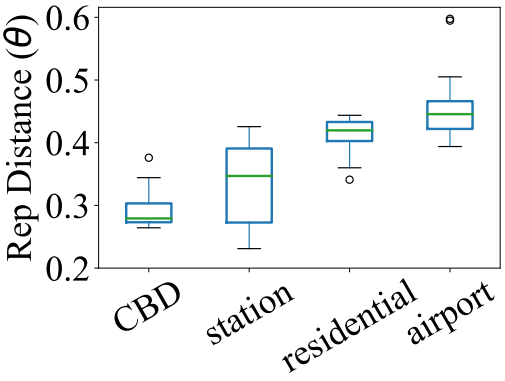


Figure 5.13: Studied Locations

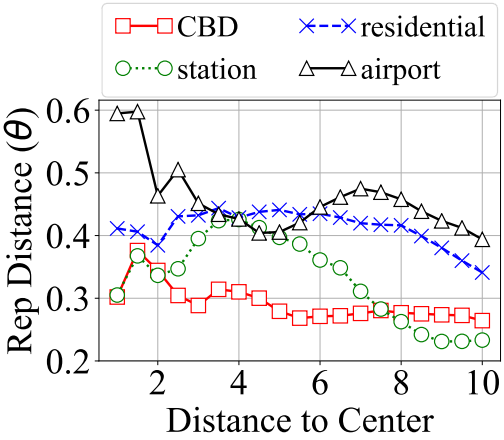


Figure 5.14: Distance to Centers

CHAPTER 6
CORRECTION MODEL

6.1 Motivation

Based on the analyses in the previous section, we found that contextual diversity (i.e., PoI, population, and mobility) is a key reason for cellular network representativeness. In regions with more diverse PoI distribution and mixed functions, higher density of population and more visitors, a single network is more representative for the usage patterns of all networks in a city. Our analysis has the potential to help fellow researchers or network operators with the data from only one network to avoid data bias for their academic research and real-world applications. For instance, they can use a sample of data from a spatial temporal combination with high contextual diversity, instead of all the data from a single network. Therefore, a natural question for us is how to design a correction model to obtain such a data sample, which is resilient to representativeness bias. The key feature of our correlation model is that it is only based on single-network data and public contextual data, and does not require the data from all networks in a city to correct the bias and thus improve representativeness. This is because accessing the data from all networks is very challenging in a real-world setting.

6.2 Problem Definition

We first introduce terminologies for diversity modeling as in Tab. 6.1 and then formalize our target problem.

Notation	Meaning
g, G	a grid and a grid collection
$\mathcal{S}^g, \mathcal{S}^G$	data from a network a grid/grid collection
$\mathcal{S}, \mathcal{S}^U$	both present data from a network for all grids
$\mathcal{S}^r$, $\mathcal{S}^R$	data from a network for a region r or a region subset R
α	a data sampling ratio in terms of $\mathcal{S}$
$\mathcal{M}^g_{from}$	a mobility matrix from grid g
$\mathcal{M}^g_{to}$	a mobility matrix to grid g
$\mathcal{P}^g$	a PoI distribution in grid g
$\mathcal{D}^g$	a Population density in grid g
$\mathcal{V}^g$	a region function distribution for grid g
$\mathcal{E}^g$	a contextual diversity for grid g

Table 6.1: Terminologies

6.3 Terminologies

1. *Spatial Partition:* We use a grid partition in our correction model, which divides a region into grids with equal widths and heights. We use grid partition because it is flexible to change sizes for different spatial granularity, which has been used in many other research [11] [4].

2. *Mobility Matrices:* For each grid g, we construct two matrices to describe its mobility patterns in the grid: a *From* matrix $\mathcal{M}^g_{from}$ to describe the number of passengers moving from grid g to other grids in different time slots. a *To* matrix $\mathcal{M}^g_{to}$ to describe the number of passengers moving to grid g from other grids in different time slots. Therefore, both matrices have $|G|$ rows and $|T|$ columns where $|G|$ is the number of grids; $|T|$ is the number of time slots covering both weekdays and weekends.

3. *PoI Distribution:* For each grid, a PoI vector $\mathcal{P}^g$ is used to describe the PoI distribution; each element in $\mathcal{P}^g$ is number of PoIs in a category, e.g., education, transport. Different from mobility matrices to show dynamic features with time, the PoI vector is a static feature on regions.

4. *Population:* Another static feature is population on a grid, we quantify population on a grid g by population density $\mathcal{D}^g$, i.e., the average number of population per km^2.

5. *Function of Regions:* Since a grid is always mixed with functions (e.g., office area, entertainment, residence, shopping, transportation hub, etc), we model function of regions with a vector $\mathcal{V}^g$ where $|\mathcal{V}^g|$ is the number of prefixed region functions; each element v_i^g in $\mathcal{V}^g$ is a probability that the grid r has a function of region, e.g., eduction. Specifically, we define $\mathcal{V}^g$ as a 5-dimension vector corresponding to five functions of regions, i.e., office, residential, educational, transportation and recreation, which are the main urban region functions used in recent literature [38]. However, different from the traditional definition of region function, which is a static feature for a region, the region function in our study is a dynamic feature on temporal dimensions since we classify region function with temporal mobility data. For example, a grid can be identified as an office area during workdays while as an entertainment area during weekends.

6. *Contextual Diversity:* Intuitively, a grid with a single function, i.e., $\mathcal{V} = \{1, 0, \cdots, 0\}$ represents a low contextual diversity. In contrast, a more uniform distribution of $\mathcal{V}$, e.g., $\mathcal{V} = \{0.1, \cdots, 0.1\}$ represents a high contextual diversity. Therefore, we quantify region diversity $\mathcal{E}^g$ with an entropy of vector $\mathcal{V}^g$, which is one of the most common measurement for randomness of elements in a set [39]. For example, $\mathcal{V}^g = \{0.2, 0.2, 0.2, 0.2, 0.2\}$ has the highest entropy, which indicates a high contextual diversity.

6.4 Target Problem: Diversity-Driven Grid Selection for Data Sampling

Given a data set $\mathcal{S}$ of a network from all grids and a sample ratio α, our target is to select a sub set of grids G from all equally-sized grids to maximize the contextual diversity $\mathcal{E}^G$ under a constraint that the size of $\mathcal{S}^G$ is equal to $\alpha \cdot |\mathcal{S}|$. All the data $\mathcal{S}^G$ from this sub set

of grids G are our data sample. The Equation 6.1 gives the formulation.

$$
\begin{aligned}
& argmax_G \ \mathcal{E}^G \\
s.t. & \sum_{g \in G} |\mathcal{S}^g| = \alpha \cdot |\mathcal{S}| \\
& |G \cap r| \geq 1, \forall r \ \in \ R
\end{aligned}
\tag{6.1}
$$

To avoid missing values on spatial dimension in sampling, we add a constraint $|G \cap r| \geq 1$ to make sure every census-based region r at least gets one of its grids selected. In our setting, a region is always bigger than a grid, and typically has a few grids in it. When we require a smaller region, i.e., finer granularity, we can decrease the grid size to satisfy the constraint.

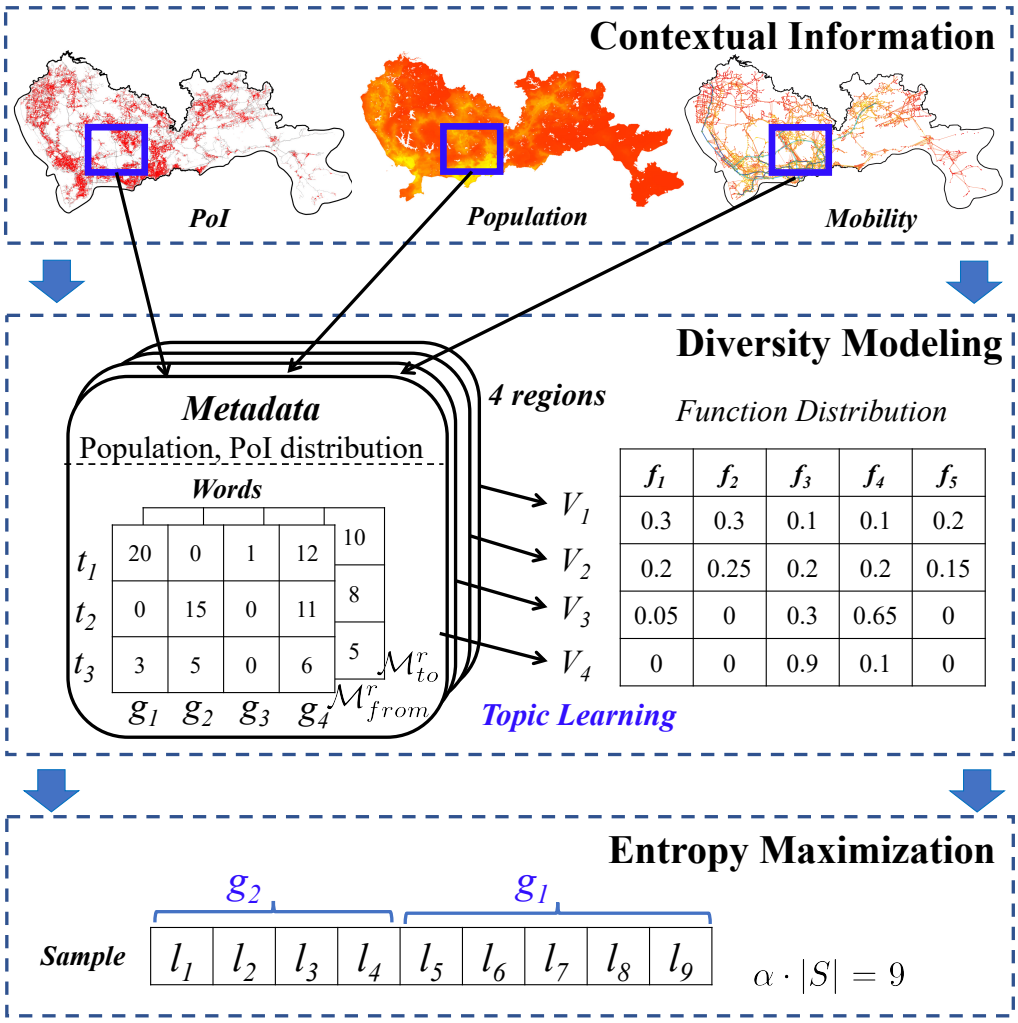


Figure 6.1: Diversity-Driven Sampling

6.5 Diversity-Driven Sampling

Since contextual diversity in terms of PoIs, population and user mobility is a key for representativeness in single networks, we propose a diversity-driven sampling strategy by selecting a few grids to construct a representative dataset (including all the data from the selected grids) from a non-representative single network data to solve the target problem in Equation 6.1. The general idea is to first quantify the contextual diversity in grids (i.e., equally-sized grids) in all regions, and then maximize the contextual diversity in sampling grids. We summarize our model into two steps as in Fig. 6.1 : *(i) diversity modeling*; *(ii) diversity-maximization sampling*. We elaborate on these two steps as follows.

Algorithm 1: Diversity-Driven Sampling

Input: α, $\mathcal{S}^U$, $\mathcal{P}^U$, $\mathcal{D}^U$, $\mathcal{M}^U_{from}$, $\mathcal{M}^U_{to}$
Result: $\mathcal{S}^G$
(1) $metadata \leftarrow (\mathcal{P}^U, \mathcal{D}^U)$;
(2) $words \leftarrow (\mathcal{M}^U_{from}, \mathcal{M}^U_{to})$;
(3) $\mathcal{V}^U \leftarrow$ topicClustering($metadata$, $words$) ;
(4) $G \leftarrow$ initialize() ;
(5) $\mathcal{C} \leftarrow U - G$;
(6) **while** $|S^G| < \alpha \cdot |S^U|$ **do**
(7) $\quad \mathcal{E}^G \leftarrow$ entropy($\mathcal{V}^G$) ;
(8) $\quad \Delta\mathcal{E}^{\mathcal{C}} \leftarrow$ entropyGain($\mathcal{V}^G$,$\mathcal{V}^{\mathcal{C}}$);
(9) $\quad g \leftarrow argmax_{g \in \mathcal{C}}\ \Delta\mathcal{E}^g$;
(10) $\quad G \leftarrow G \cup \{g\}$;
(11) $\quad \mathcal{C} \leftarrow \mathcal{C}/\{g\}$;
(12) $\quad \mathcal{S}^G \leftarrow \mathcal{S}^G \cup \mathcal{S}^g$
(13) **end**

(i) ***Diversity Modeling:*** In diversity modeling, we generate a vector $\mathcal{V} = \{v_1, v_2, \cdots, v_n\}$ for each grid where n is the number of potential region functions (e.g., education, office, etc), and each element v_i in $\mathcal{V}$ is the probability that a grid belongs to a function. In general, a higher entropy on $\mathcal{V}$ indicates a more diverse distribution on region functions, thus a larger contextual diversity in a region. To construct such a $\mathcal{V}$ from contextual information of a grid, e.g., population, mobility and PoI distribution, we apply a topic model [38].

Topic models such as LDA [40] was proposed to model the relation between the word distribution in a document and the topic distribution of the document. Similarly, we infer region functions with topic models along with the input of mobility, PoI and population. Specifically, the detailed mapping from *region function clustering* to *document topic clustering* is as follows: we map *grids* to *documents*; *region functions* to *document topics*; the dynamic feature, i.e., *mobility matrices* to *words*; the static features, i.e., *population and PoI distributions* to *meta data of documents*, e.g., authors, key words of documents. We initialize the topic number as 5 in the clustering and thus the output of a topic model for a document is a vector $\mathcal{V}$ with 5 functions of regions, and each element of the vector indicates the possibility that the document belongs to a topic, i.e., a function of region. Thus, in our region diversity modeling, the topic model is to assign a grid with a distribution of region functions $\mathcal{V}^g$ where $\mathcal{V}^g = \{v_1^g, v_2^g, \cdots, v_n^g\}$ and v_i^g is the possibility a grid g belongs to a region function i. Fig. 6.1 presents a simplified example with 4 grids, 3 time slots, and 5 functions of regions. We map contextual data into grids, and each grid has population and PoI distribution as *metadata*. Besides, both $\mathcal{M}_{from}$ and $\mathcal{M}_{to}$ have 3 rows for 3 time slots and 4 columns for 4 regions. The topic model will generate a 5-dimension vector $\mathcal{V}$ for each region to describe the possibility that the grid has these 5 functions.

(ii) ***Diversity-Maximization Sampling:*** After the first step, each grid has been assigned with a function distribution vector. Based on that, our second step is to create a data sample that meets the sampling requirement and maximizes the contextual diversity of the grids having this data sample. To achieve it, we apply an entropy maximization strategy based on a greedy algorithm. We separate all grids U into two groups, i.e., a selected group G and a unselected group $U - G$. In the initialization, for each region r, we select a grid in r with the highest entropy and put the grid in G to satisfy the second constraint in Equation 6.1. Second, we calculate the entropy gain based on $\mathcal{E}^G$ for every grid in $U - G$. We select the gird g with the highest entropy gain, i.e., the diversity gain, and then update $G = G \cup \{g\}$. Third, we update the $\mathcal{E}^G$ with new G, i.e., including this new grid g. The process will stop

until the number of sample records are satisfied. For the example in Fig. 6.1, the number of sample records are $\alpha \cdot |\mathcal{S}| == 9$ and there are 4 records (i.e., l_1 to l_4) in grid g_2 and 5 records g_1 (i.e., l_5 to l_9). We first select 4 records from g_2 since $\mathcal{V}^{g_2}$ has the largest entropy with one region selected and then select 5 records from g_1 since we have the largest entropy in $\mathcal{V}^{\{g_1,g_2\}}$. The process is described in Algorithm 1.

6.6 Evaluation

6.6.1 Evaluation Settings

we evaluate the sampling strategy with the following settings.

Ground Truth: We use the load of three networks, which covers 100% of cellular users, as the ground truth.

Baselines: we compare the *CellRep* with two baselines: *(1) Single* is based on the raw data from the most representative network for the best performance of a single network, i.e., a network with the lowest representativeness distance from A, B and C without sampling. *(2) CellSam* is a uniform sampling method without considering the contextual diversity.

Metrics: we use representativeness distance θ as the metrics for the evaluation, a lower representativeness distance indicates a higher representativeness.

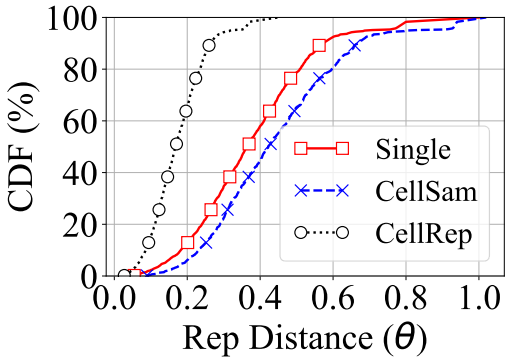


Figure 6.2: Performance

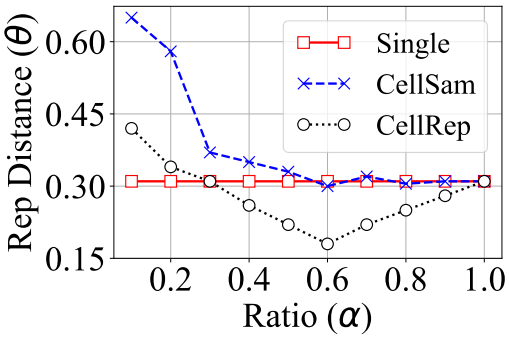


Figure 6.3: Sample Ratio

6.6.2 Overall Results

We compare the performance of *CellRep* with two baselines in Fig. 6.2. Both the baseline model *CellRam* and our *CellRep* increases the representativeness by reducing the representativeness distance due to the higher sample until the ratio is 0.6. In particular, our *CellRep* decreases representativeness distance significantly from 0.31 to 0.16 on average as shown in Fig. 6.3. It shows that with a sophisticated sampling strategy in *CellRep*, even 30% of sample data from a single network can achieve similar representativeness as all single-network data.

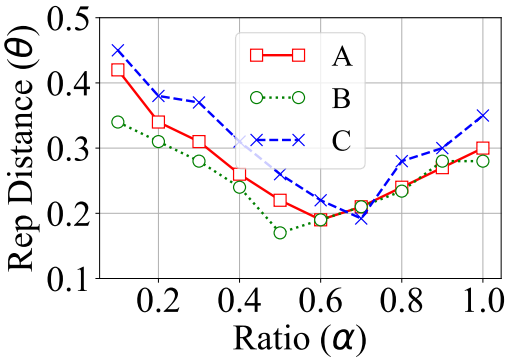


Figure 6.4: Impact of Networks

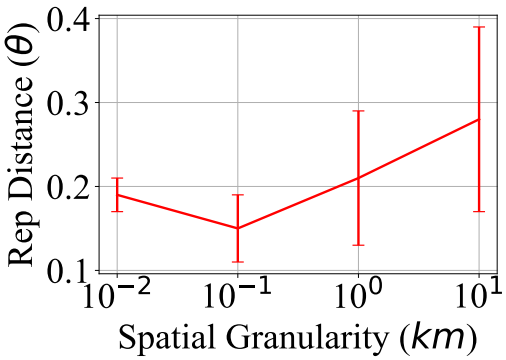


Figure 6.5: Impact of Grid Size

6.6.3 Impact of Factors

We further study the impact of different factors on the performance of *CellRep*. Fig. 6.4 compares the resulted representativeness distance θ with data in the three networks. We found even three networks have different user coverage, they can achieve similar representativeness with *CellRep*. Therefore, *CellRep* shows a robust performance in different networks. Specifically, the representativeness distance can be reduced to smaller than 0.2 in Network A, B, and C with α equal to 0.5, 0.6, and 0.7, respectively. Moreover, we study the impact of spatial granularity in Fig. 6.5, which shows the performance of *CellRep* with different grid sizes. *CellRep* achieves the best performance with a grid size $100m \times 100m$. In general, a finer spatial granularity leads to a better performance.

6.6.4 Impact on Real-World Applications

Relying on the measurement results, we validate the impact of representativeness on population estimation application as introduced in the motivation. Different from the previous work, which improves the inference accuracy, our work focuses on a different angle, which studies the impact of representativeness of cellular data. We implement a contextual-aware population estimation with cellular usage data from single networks [24] and map the estimated population to the administrative regions. We use the Worldpop [36] dataset as the ground truth data for cross-validation and MAPE (Mean Absolute Percent Error) as the evaluation metric. We study the impact of our representative distance on this population estimation in Fig. 6.6, which proves that a higher representativeness distance leads to a worse performance on the application. Fig. 6.7 shows *CellRep* corrects the data bias in this population estimation and improves the performance by reducing the MAPE 40.3% from 25.8% to 15.4% compared with a baseline Single (which use raw data in single networks) and another baseline CellSam, which use a uniform sampling method without considering the contextual diversity.

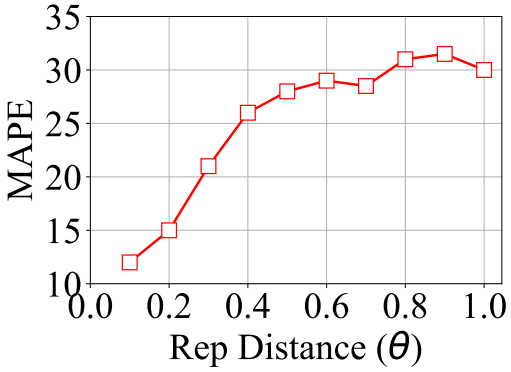


Figure 6.6: Impact on Pop

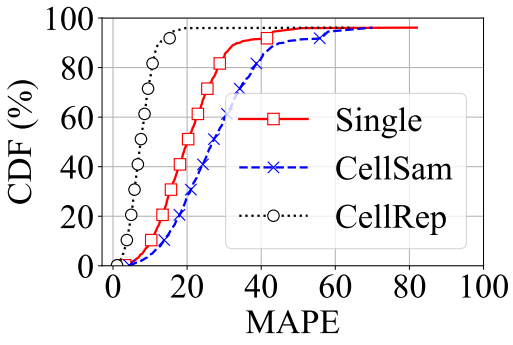


Figure 6.7: Impact of Correction

CHAPTER 7
DISCUSSION

Lesson Learned: We summarize several lessons learned and implications as follows.

1. Contextual diversity is the key factor for network representativeness on both spatial and temporal dimensions. Different contextual information (e.g., PoI distribution, population and mobility) causes different cellular user distribution and leads to representativeness difference of single networks.

2. The representativeness is one of the most important factors for performance for real-world applications. We found a high correlation between representativeness and the performance of a population estimation model. Due to the limited access to cellular activities from multiple networks, most existing applications and research studies are based on single networks. On one hand, a better understanding on representativeness can help understand the performance of existing models. On the other hand, our measurement study paves a way to future cellular web log studies by providing pre-analysis results and insights.

3. A well-designed correction model provides an approach to improving data qualify in single networks by combining open contextual data with single-network data. Our evaluation results show that such a correction model has the potential to improve application performance by intelligently sampling the representative data. The correction model can be applied to many applications related with web services or user distribution such as traffic demand prediction [6], web user estimation [41], hot spot recommendations [42].

Ethical and Privacy Issues: Our study acquired consent to investigate the Cellular web log data for research purposes, which is approved by IRB. The data we investigate (i) Deidenti-

fication: the analyzed data are anonymized by the three cellular operators, and identifiable IDs (e.g., phone numbers or SIM IDs) are replaced by a serial identifier during the analyses. (ii) Coarse-grained Locations: we analyze cellular user behaviors at the level of cell towers, which may cover from a few thousand square meters to a few square kilometers, which cannot reveal detailed locations of users. (iii) Aggregation: Our work was exempted by an IRB process in our affiliation since there is no more than the minimal risk to conduct our research because the tower-level results are based on aggregation, which cannot be traced back to individual cellular users. (iv) Benefits Outweigh Risks: All cellular users consented that their data will be used for cellular network management and improvement. We believe our results have positive impacts on cellular users' by improving their cellular services so the benefit of our data-driven research outweigh the potential risk.

Limitation: A limitation is that our study is based on three networks in one particular city. Due to limited data access, we cannot validate our findings in other cities. However, most cities in the world are covered by multiple cellular networks. We believe that the findings in this work are meaningful to other cities, especially the cities in China since they have the same three cellular operators.

CHAPTER 8
CONCLUSION

As an infrastructure for mobile web service, we conduct a comprehensive study based on multiple diverse cellular networks to understand cellular service representativeness at city scale with more than 10 million cellular users. We quantify the representativeness in single networks and explain the potential reasons for representativeness differences. Based on our analysis, we design a correlation model and then validate its performance based on real world application on population estimation. Our analysis results could be used as a preliminary result to provide insights for research work and applications such as city-scale web service modeling, mobility and population estimation. Last but not least, we design a correction model to improve representativeness in single-network data. Our correction model can improve representativeness of a single network by 45.58% on average.

REFERENCES

[1] S. Statista The Statistics Portal, *Number of mobile (cellular) subscriptions worldwide from 1993 to 2017 (in millions).* `www.statista.com/statistics/262950/global-mobile-subscriptions-since-1993/`, 2018.

[2] M. Tariq, A. Zeitoun, V. Valancius, N. Feamster, and M. Ammar, "Answering what-if deployment and configuration questions with wise," in *ACM SIGCOMM Computer Communication Review*, ACM, vol. 38, 2008, pp. 99–110.

[3] J. Wang, J. Tang, Z. Xu, Y. Wang, G. Xue, X. Zhang, and D. Yang, "Spatiotemporal modeling and prediction in cellular networks: A big data enabled deep learning approach," in *INFOCOM 2017-IEEE Conference on Computer Communications, IEEE*, IEEE, 2017, pp. 1–9.

[4] S. Yang, Y. He, Z. Ge, D. Wang, and J. Xu, "Predictive impact analysis for designing a resilient cellular backhaul network," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 1, no. 2, p. 30, 2017.

[5] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang, "Characterizing and modeling internet traffic dynamics of cellular devices," *ACM SIGMETRICS Performance Evaluation Review*, vol. 39, no. 1, pp. 265–276, 2011.

[6] F. Xu, Y. Li, H. Wang, P. Zhang, and D. Jin, "Understanding mobile traffic patterns of large scale cellular towers in urban environment," *IEEE/ACM Transactions on Networking*, vol. 25, no. 2, pp. 1147–1161, 2017.

[7] H. Falaki, R. Mahajan, S. Kandula, D. Lymberopoulos, R. Govindan, and D. Estrin, "Diversity in smartphone usage," in *Proceedings of the 8th international conference on Mobile systems, applications, and services*, ACM, 2010, pp. 179–194.

[8] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, S. Venkataraman, and J. Wang, "A first look at cellular network performance during crowded events," in *ACM SIGMETRICS Performance Evaluation Review*, ACM, vol. 41, 2013, pp. 17–28.

[9] H. Barbosa, M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini, and M. Tomasini, "Human mobility: Models and applications," *Physics Reports*, 2018.

[10] G. Chen, "Spatiotemporal individual mobile data traffic prediction," PhD thesis, INRIA Saclay-Ile-de-France, 2018.

[11] S. Isaacman, R. A. Becker, R. Cáceres, M. Martonosi, J. Rowland, A. Varshavsky, and W. Willinger, "Human mobility modeling at metropolitan scales," in *MobiSys*, 2012.

[12] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, S. Venkataraman, and J. Wang, "Characterizing and optimizing cellular network performance during crowded events," *IEEE/ACM Transactions on Networking (TON)*, vol. 24, no. 3, pp. 1308–1321, 2016.

[13] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Geospatial and temporal dynamics of application usage in cellular data networks," *IEEE Transactions on Mobile Computing*, vol. 14, no. 7, pp. 1369–1381, 2015.

[14] H.-M. Chuang and C.-H. Chang, "Verification of poi and location pairs via weakly labeled web data," in *Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 743–748.

[15] A. Likhyani, D. Padmanabhan, S. J. Bedathur, and S. Mehta, "Inferring and exploiting categories for next location prediction," in *Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 65–66.

[16] R. Smarzaro, T. F. de Melo Lima, and C. A. D. Jr., "Could data from location-based social networks be used to support urban planning?" In *WWW '17 Companion Proceedings of the 26th International Conference on World Wide Web Companion*, 2017, pp. 1463–1468.

[17] G.-H. Tu, Y. Li, C. Peng, C.-Y. Li, H. Wang, and S. Lu, "Control-plane protocol interactions in cellular networks," *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 4, pp. 223–234, 2015.

[18] R. Xie, Y. Chen, S. Lin, T. Zhang, Y. Xiao, and X. Wang, "Understanding skout users' mobility patterns on a global scale: A data-driven study," *World Wide Web*, 2018.

[19] V. Albino, U. Berardi, and R. M. Dangelico, "Smart cities: Definitions, dimensions, performance, and initiatives," *Journal of Urban Technology*, vol. 22, no. 1, pp. 3–21, 2015.

[20] A. Janecek, D. Valerio, K. A. Hummel, F. Ricciato, and H. Hlavacs, "The cellular network as a sensor: From mobile phone data to real-time road traffic monitoring," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 5, pp. 2551–2572, 2015.

[21] F.-Z. Jiang, K. Thilakarathna, M. Hassan, Y. Ji, and A. Seneviratne, "Efficient content distribution in dooh advertising networks exploiting urban geo-social connec-

tivity," in *WWW '17 Companion Proceedings of the 26th International Conference on World Wide Web Companion*, 2017, pp. 1363–1370.

[22] Z. Fang, F. Zhang, L. Yin, and D. Zhang, "Multicell: Urban population modeling based on multiple cellphone networks," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 3, 106:1–106:25, Sep. 2018.

[23] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "A first look at cellular machine-to-machine traffic: Large scale measurement and characterization," *ACM SIGMETRICS performance evaluation review*, vol. 40, no. 1, pp. 65–76, 2012.

[24] F. Xu, P. Zhang, and Y. Li, "Context-aware real-time population estimation for metropolis," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ACM, 2016, pp. 1064–1075.

[25] C. V. N. Index, "Global mobile data traffic forecast update, 2016–2021 white paper," *Cisco: San Jose, CA, USA*, 2017.

[26] G. Hu, J. Shao, Z. Ni, and D. Zhang, "A graph based method for constructing popular routes with check-ins," *World Wide Web*, vol. 21, no. 6, pp. 1689–1703, 2018.

[27] W. Liu, H. Lai, J. Wang, G. Ke, W. Yang, and J. Yin, "Mix geographical information into local collaborative ranking for poi recommendation," *World Wide Web*, pp. 1–22, 2019.

[28] E. E. Papalexakis, K. Pelechrinis, and C. Faloutsos, "Spotting misbehaviors in location-based social networks using tensors," in *Proceedings of the 23rd International Conference on World Wide Web*, 2014, pp. 551–552.

[29] I. Trestian, K. Huguenin, L. Su, and A. Kuzmanovic, "Understanding human movement semantics: A point of interest based approach," in *Proceedings of the 21st International Conference on World Wide Web*, 2012, pp. 619–620.

[30] A. Noulas, C. Mascolo, and E. Frias-Martinez, "Exploiting foursquare and cellular data to infer user activity in urban environments," in *2013 IEEE 14th International Conference on Mobile Data Management*, IEEE, vol. 1, 2013, pp. 167–176.

[31] F. Xu, Z. Tu, Y. Li, P. Zhang, X. Fu, and D. Jin, "Trajectory recovery from ash: User privacy is not preserved in aggregated mobility data," in *WWW '17 Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 1241–1250.

[32] W. Dong, S. Rallapalli, R. Jana, L. Qiu, K. Ramakrishnan, L. Razoumov, Y. Zhang, and T. W. Cho, "Ideal: Incentivized dynamic cellular offloading via auctions," *IEEE/ACM Transactions on Networking (TON)*, vol. 22, no. 4, pp. 1271–1284, 2014.

[33] M. A. Qureshi, A. Mahimkar, L. Qiu, Z. Ge, S. Puthenpura, N. Mir, and S. Ahuja, "Reflection: Automated test location selection for cellular network upgrades," in *Network Protocols (ICNP), 2017 IEEE 25th International Conference on*, IEEE, 2017, pp. 1–10.

[34] M. Luczak-Roesch, L. Hollink, and B. Berendt, "Current directions for usage analysis and the web of data: The diverse ecosystem of web of data access mechanisms," in *WWW '16 Companion Proceedings of the 25th International Conference Companion on World Wide Web*, 2016, pp. 885–887.

[35] D. Zhang, J. Huang, Y. Li, F. Zhang, C. Xu, and T. He, "Exploring human mobility with multi-source data at extremely large metropolitan scales," in *Proceedings of the 20th annual international conference on Mobile computing and networking*, ACM, 2014, pp. 201–212.

[36] C. L. Forrest R Stevens Andrea E Gaughan and A. J. Tatem., "Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data," *PloS one, 10(2)*, 2015.

[37] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Characterizing geospatial dynamics of application usage in a 3g cellular data network," in *INFOCOM, 2012 Proceedings IEEE*, IEEE, 2012, pp. 1341–1349.

[38] J. Yuan, Y. Zheng, and X. 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*, ACM, 2012, pp. 186–194.

[39] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[40] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[41] A. Balachandran, V. Aggarwal, E. Halepovic, J. Pang, S. Seshan, S. Venkataraman, and H. Yan, "Modeling web quality-of-experience on cellular networks," in *Proceedings of the 20th annual international conference on Mobile computing and networking*, ACM, 2014, pp. 213–224.

[42] S. Hoteit, S. Secci, S. Sobolevsky, C. Ratti, and G. Pujolle, "Estimating human trajectories and hotspots through mobile phone data," *Computer Networks*, vol. 64, pp. 296–307, 2014.