HUMAN MOBILITY MODELING BASED ON HETEROGENEOUS URBAN SENSING SYSTEMS

by

ZHIHAN FANG

A dissertation submitted to the
School of Graduate Studies
Rutgers, The State University of New Jersey
In partial fulfillment of the requirements
For the degree of
Doctor of Philosophy
Graduate Program in Computer Science

Written under the direction of
Desheng Zhang
and approved by

__________________________________________

__________________________________________

__________________________________________

__________________________________________

New Brunswick, New Jersey
May, 2020
Recently, with an increasing number of people living in cities, it introduces new challenges in human mobility such as traffic congestion and energy consumption, which are caused by dense human population distribution, unbalanced infrastructure deployment, or insufficient understanding of travel demand. Thus, it is essential to improve the mobility of urban residents on a daily basis, which can be achieved by accurately modeling human mobility with ubiquitous urban sensing data from heterogeneous urban sensing systems, e.g., on-board GPS systems including taxis, buses, personal vehicles, and portable device systems such as cellphones. Existing studies modeling human mobility are mostly built upon single systems. However, people in cities take multiple transportation modalities on a daily basis, where a single sensing system limits a comprehensive understanding and modeling of human mobility.

In the dissertation, we aim to model human mobility at the metropolitan scale, by utilizing spatio-temporal data of heterogeneous sensing systems already collected for billing or management purposes. We design, implement and evaluate a data-driven framework named urbanSense with three modules for human mobility modeling (e.g.,
travel distance, travel time, travel speed): (i) a sensing module to collect and preprocess human mobility sensing data from 8 urban sensing systems crossing 3 domains (i.e., transportation, communication, and payment); (ii) a measurement module where we present a measurement work named SysRep to measure the data bias of urban sensing systems for human mobility modeling. In SysRep, we quantify the data bias of urban sensing systems as representativeness of sensing systems. We analyzed potential reasons for representativeness and found the representativeness is highly correlated with contextual factors such as population, mobility, and Points of Interest. We further design a correction model to improve the representativeness of sensing systems. The evaluation results show the proposed correction model can improve the representativeness of single systems by 45%. (iii) a prediction module to model human mobility from heterogeneous urban sensing systems. In particular, we present two works: one work named MultiCell for real-time population modeling and the other work named MAC for travel time prediction. In MultiCell, we design two techniques to model real-time population from multiple cellular networks: a spatial alignment technique to align different spatial partitions into a uniform spatial partition; a co-training technique to learn the relation between active cellphone users of different networks and population distribution simultaneously. MultiCell is implemented with Call Detail Records (CDR) of three major networks in China in the same city covering 100% cellphone users. The evaluation results prove the effectiveness of MultiCell by reducing the modeling error by 27% compared with the start-of-the-art models. In MAC, we decompose travel time of multiple transportation systems (i.e., subway, taxi, bus, and personal vehicle) into fine-grained travel time based on different travel stages (e.g., walking, riding, waiting time). Moreover, we design a time-series model based on Long Short-Term Memory (LSTM) architecture to predict the travel delay under the impact of different anomalies. We implement and evaluate MAC with data collected from 37 thousand vehicles and 5 million smart cards. The results show MAC reduces the prediction error by 31% compared with state-of-the-art methods. Finally, we discuss some lessons learned and potential applications of our framework.
Acknowledgements

First of all, I would like to give straightforward but truthful, deepest gratitude and appreciation to my advisor, Prof. Desheng Zhang, for his guidance, inspiration, care and tremendous patience on me during my Ph.D study. Desheng is the person who brought me into research, and spent days and nights on my papers and dissertations. Without his guidance, it would be impossible for me to complete Ph.D and this dissertation. During the years in working with him, I am impressed, influenced and shaped by his wisdom, kindness and diligence. I am sincerely grateful for all mental, academic, financial supports from him. The discussions with him have always been constructive and inspiring.

I would like to thank Prof. Yongfeng Zhang, Prof. Jie Gao, Dr. Ruilin Liu, for serving for my defense committee and providing me with insightful feedback on my thesis work and presentation. I would like to acknowledge Prof. Casimir Kulikowski, Prof. Hui Xiong, Prof. Yongfeng Zhang for serving for my qualifying committee and providing insightful feedback on my presentation and research directions.

I would also like to thank Prof. Shuai Wang (Southeast University, China) for his valuable comments, selfless guidance and encouragement on my research papers and academic careers. It is a great experience and honor to work with him, from whom I gained rich domain knowledge.

I was fortunate to join ARSENAL group and would like to express my sincere appreciation to all excellent group members including Xiaoyang Xie, Guang Wang, Yu Yang, Zhou Qin, Shuxin Zhong and Fan Zhang, and other undergraduate or master students contributing to my research projects including but not limited Boyang Fu, Chaoji Zuo, Bliss Hu, Jiaxu Su, Xiaofa Lin. I also thank many of my friends sharing my 5 years of life in Rutgers including but not limited Liyang Wang, Yikun Xian,
Qiaoying Huang, Shuhan Wang, Jian Song, Yujun Pan, Qi Chang, Yikai Zhang, Yizhe Zhu, Kun Wang, Ji Zhang, Cong Wang, Yang Yu.

I would also like to express my deepest thanks to my family. I am deeply indebted to my parents for their love, encouragement and selfless support. I sincerely thank my wife, Anran Chen for her accompany, support, dedication and sharing every important moment of life since 12 years ago.
Dedication

To my parents and my wife.

To my advisor Prof. Desheng Zhang for his guidance and care.
# Table of Contents

Abstract ................................................................. ii  
Acknowledgements .............................................. iv  
Dedication .............................................................. vi  
List of Tables .......................................................... x  
List of Figures ........................................................ xi  

1. Introduction ....................................................... 1  

2. Framework and Datasets .......................................... 6  
   2.1. Heterogeneous Urban Sensing Data .......................... 8  
       2.1.1. Cellular Sensing System Data ....................... 8  
       2.1.2. Vehicular Sensing System Data .................... 8  
       2.1.3. Payment Sensing System Data ..................... 10  
   2.2. Contextual Data ................................................ 11  
       2.2.1. Population ........................................... 11  
       2.2.2. Point of Interests (PoI) ............................ 11  
       2.2.3. Road Networks ........................................ 12  
       2.2.4. Anomalies ............................................. 12  

3. Measurement for Sensing Representativeness .................. 15  
   3.1. Introduction .................................................. 15  
   3.2. Motivation ................................................... 17  
   3.3. Measurement .................................................. 19  
       3.3.1. Terminologies ....................................... 19
3.3.2. Methodology ................................................. 22
3.3.3. Evaluation .................................................. 23
3.3.4. A Case Study ............................................... 28
3.4. Correction ..................................................... 29
  3.4.1. Motivation ................................................ 29
  3.4.2. Terminologies .......................................... 30
  3.4.3. Methodology ............................................ 32
  3.4.4. Evaluation ............................................... 35
3.5. Discussion .................................................. 36
3.6. Related Work ............................................... 38
3.7. Summary ..................................................... 40

4. Prediction Model for Population .................................. 41
  4.1. Introduction ............................................... 41
  4.2. Motivation ................................................ 44
  4.3. Methodology ............................................... 47
    4.3.1. Core Idea ........................................... 47
    4.3.2. Tower-based Partition ................................ 48
    4.3.3. Spatial Alignment ................................... 49
    4.3.4. Population Estimation .............................. 50
    4.3.5. Implementation ...................................... 59
  4.4. Evaluation ................................................ 63
    4.4.1. Evaluation Methodology ............................ 63
    4.4.2. Evaluation Results ................................ 66
  4.5. Related Work ............................................... 71
  4.6. Summary .................................................. 72

5. Prediction Model for Travel Time .................................. 73
  5.1. Introduction ............................................... 73
  5.2. Motivation ................................................ 75
5.2.1. Fined-grained Travel Time .............................. 75
5.2.2. Multiple Transportation Systems .......................... 76
5.2.3. Anomaly Events ........................................ 76
5.2.4. Summary ............................................... 77

5.3. Methodology .............................................. 77
5.3.1. Background: Understanding Passenger Behaviors .......... 78
5.3.2. Terminologies ......................................... 78
5.3.3. Travel Time Decomposition: Aboveground .................. 79
5.3.4. Travel Time Decomposition: Underground ................... 81
5.3.5. Measurement Method .................................... 84

5.4. Evaluation ............................................... 85
5.4.1. Waiting Time Patterns .................................. 87
5.4.2. Riding Time Patterns ................................... 89
5.4.3. Impact of Anomalies .................................... 90

5.5. Application: Delay Time Prediction .......................... 92
5.5.1. Methodology ........................................... 92
5.5.2. Evaluation ............................................. 95

5.6. Discussion ................................................. 99

5.7. Related Work .............................................. 100
5.8. Summary ................................................ 102

6. Future Work and Conclusion .................................. 103

6.1. Future Directions .......................................... 103
6.1.1. Sensing: Infrastructure Improvement ...................... 103
6.1.2. Measurement: Human Mobility Evolving .................... 104
6.1.3. Prediction: Individual Human Mobility ...................... 104
6.1.4. Novel Applications and Services .......................... 105

6.2. Conclusion ................................................ 106

References .................................................. 107
List of Tables

2.1. Anomaly Categorization .............................................. 13
2.2. Description of Accident Data Set ............................... 13
2.3. Sample of Unexpected Anomaly Data Set ...................... 14
2.4. Sample of Expected Anomaly Data Set ......................... 14
3.1. Terminologies ........................................................... 20
3.2. PoI Distribution in Groups .......................................... 25
3.3. Tower Distribution on Select Locations ....................... 28
3.4. Terminologies ........................................................... 30
3.5. Cellular Web Log Analyses Survey .............................. 38
4.1. Terminology and Notations ......................................... 51
4.2. PoI Distribution in the city ........................................ 52
5.1. Travel Time Measurement Survey ............................... 100
# List of Figures

2.1. System Framework ................................................. 6  
2.2. Cellular Sensing Systems ........................................ 9  
2.3. Vehicular Sensing Systems ....................................... 9  
2.4. Payment Sensing Systems ......................................... 10  
2.5. Contextual Data Distribution ..................................... 12  
3.1. Cell Size ......................................................... 18  
3.2. Impact of Bias on App. ........................................... 18  
3.3. Census-based Partition ............................................ 20  
3.4. Representativeness Distance ...................................... 22  
3.5. Regions ......................................................... 23  
3.6. Spatial Groups ................................................. 23  
3.7. PoI v.s. Spatial ................................................. 24  
3.8. Groups v.s. Spatial .............................................. 24  
3.9. Pop. v.s. Spatial ................................................ 25  
3.10. Mobility v.s. Spatial ............................................ 25  
3.11. Time of Day .................................................... 26  
3.12. Temporal Groups ............................................... 26  
3.13. Mobility v.s. Temporal ......................................... 27  
3.15. Average Representative Distance in One Week .............. 27  
3.16. Case Study Areas and Their Contextual Diversity .......... 28  
3.17. Studied Locations .............................................. 29  
3.18. Distance to Centers ............................................ 29  
3.19. Diversity-Driven Sampling ...................................... 32
4.24. Correlation .............................................. 66
4.25. RMSE ....................................................... 66
4.26. CAPE-A .................................................... 68
4.27. CAPE-B .................................................... 68
4.28. MultiCell .................................................. 68
4.29. Population ................................................ 68
4.30. Population Dynamics ................................. 69
4.31. S-Granularity ............................................ 70
4.32. T-Granularity ............................................ 70
4.33. Sub. Correlation ....................................... 71
4.34. Taxi Correlation ........................................ 71
4.35. Bus Correlation .......................................... 71
5.1. Fine-grained Travel Time ............................. 75
5.2. Waiting Time ............................................. 75
5.3. Waiting Time Ratio ..................................... 75
5.4. Travel Demand .......................................... 77
5.5. Demand Trend .......................................... 77
5.6. Storm ....................................................... 77
5.7. Waiting Time Inference ............................... 80
5.8. Riding Time Inference ................................. 80
5.9. Subway Passenger Behavior ......................... 82
5.10. Subway Passenger Behavior ......................... 85
5.11. Bus Passenger Behavior ............................. 85
5.12. Sub.-Morning Peak ................................. 86
5.13. Sub.-Evening Peak ................................. 86
5.14. Sub.-Regular ................................. 86
5.15. Subway Inference ................................. 86
5.16. Subway Inference Regression ...................... 86
5.17. Bus-Morning Peak ................................. 87
5.47. Systems ............................................. 98
5.48. Anomalies ............................................. 98
Chapter 1
Introduction

According to the United Nations, we are undergoing a rapid process of urbanization where more than 50% of the world’s population has already been moved into urban areas in 2014, and this number is projected to rise to 70% by 2050 [124]. With the urbanization, an increasing number of people are living in cities with dense population density and high travel demand, which bring many challenges for urban human mobility [123]. According to survey, the U.S. people spend 26 minutes on traveling over 18 miles on average for daily commute [80], and the average daily travel distance increases to 40 miles with all purposes [81]. The booming demand of human mobility introduces enormous pressures to city infrastructures such roads and subway lines. The conflicts between the increasing human travel demand and the limited infrastructure supply degrade human travel experience, causing economic, social, and environmental costs. Previous studies have revealed it costs people in Washington D.C. 82 hours on travel delay on average for one year [82]. In 2014, the peak hour commuters in New York City suffer from an annual delay of 74 hours with a drop of 10% in the average speed and an annual economic loss of $15 billion for all transportation modals, and even for the less-congested NYC subway system, the cost for its delays is around $389 million [93]. One of the key solutions to improve travel experience in cities is to understand and model human mobility with data collected in urban-scale sensing systems [123].

There are many works focusing on human mobility and travel behavior modeling on large-scale mobility sensing data. Most of existing studies are based on single urban sensing systems, e.g., human mobility modeling on cellular networks [50]; travel time inference on taxis [105], buses [132], and subways [57]; location estimation on personal vehicles [109]. However, human mobility are not limited to a single system, e.g., subway,
bus, taxis, or personal vehicles (including for-hire vehicles) [26]. As a result, those approaches are potentially biased by data sources, lacking comprehensive investigations of human mobility in cities.

The key reason for the above limitation is the logistics boundary among different urban sensing systems, which makes it difficult for researchers to have data access for heterogeneous sensing systems for a particular city. Recently, due to upgrades of city infrastructures, many urban transportation or mobile systems have been equipped with sensing devices such as on-board GPS. Besides, with the decreasing cost of data transmission, a large amount of mobility data have been collected with those sensing devices and transmitted to central servers. Based on ubiquitous data collected for social good, in the vision of smart cities, numbers of city governments have been consolidating various data from their urban sensing systems as whole, e.g., transportation and communication. As part of the smart city initiative of Shenzhen city, we are fortunate to have data access to heterogeneous urban sensing systems. However, it is challenging to model human mobility from heterogeneous systems for two reasons: (i) Diverse spatio-temporal granularities of heterogeneous sensing systems. For instance, an on-board GPS device records vehicle locations with fine temporal granularity, i.e., every 5 to 15 seconds. In comparison, cellular networks have a coarse spatial-temporal granularity. A connection record is generated when a user is connected to a nearby tower, e.g., for a phone call or data connection. The connection record is always with coarse spatial and temporal granularity, e.g., a cellular tower (spatial granularity) and every 1 to 10 hours (temporal granularity). (ii) Diverse sensing partitions of heterogeneous sensing systems. Since various sensors are deployed in heterogeneous sensing systems, it introduces diverse spatial coverage and spatial partitions. For instance, the coverage of cellular networks is the radius of cellular towers, which is always modeled by Voronoi partitions [28]. In comparison, in subway systems, the coverage of a subway station always depends on the walking distance of passengers [15]. In taxi and personal vehicle systems, grid partitions are always applied [109]. Therefore, it is challenging to unify the diverse spatial partitions and model human mobility in a uniform spatial partition.

In this dissertation, our goal is to measure and model human mobility at urban
scale based on data collected by heterogeneous urban sensing systems. We design, implement, and evaluate a human mobility modeling framework named urbanSense. With heterogeneous urban sensing systems including transportation, communication, and payment systems, we conduct a comprehensive study to measure and model human mobility with heterogeneous sensing systems. In general, urbanSense has three modules: (i) a sensing module to collect travel behaviors data from heterogeneous sensing systems; (ii) a measurement module to measure the weakness and bias in single sensing systems; (iii) a prediction module to model travel behaviors with heterogeneous sensing data. In general, urbanSense advances current studies from two perspectives: (i) A comprehensive analysis and comparison to understand unique features of multiple sensing systems; (ii) An explosive framework to utilize strength and complimentary features of heterogeneous sensing systems for human mobility modeling. We summarize our contribution as follows.

- Conceptually, we design the first framework named urbanSense to comprehensively understand and model human mobility from heterogeneous urban sensing systems. We conduct a detailed study to explore the sensing bias and unique features in individual urban sensing systems, which are used to drive complimentary fusion of heterogeneous systems for human travel behavior modeling. Our work is based on real-world urban sensing systems in the same city covering around 10 million users. Specifically, we implement and evaluate urbanSense with 8 urban sensing systems crossing three domains (i.e., transportation, communication, and payment). To the best of our knowledge, urbanSense is one of the largest systems to understand and model urban travel behaviors with real-world datasets.

- In the sensing module of urbanSense, we investigate 8 urban sensing systems to explore their sensing granularity and data quality. Specifically, urbanSense is built upon 3 vehicular fleets with on-board GPS devices (i.e., taxi, bus, and personal vehicle systems); 3 major cellular networks covering 100% cellphone users in the same city; and 2 automated fare collection systems covering 3 million bus
passengers and 4 subway passengers. We present sensing data formats, data maintenance, and data cleaning in the sensing module. Besides, we collect external data sources as contextual information of those sensing systems including Point of Interests (PoI) distribution, population distribution, and road networks.

- In the measurement module of urbanSense, to explore the bias in single sensing systems, we propose a metric, i.e., representativeness distance, to quantify the representativeness of single systems. We analyze the representativeness of single sensing systems under various spatial and temporal dimensions. Moreover, we investigate the correlation between representativeness and contextual factors such as population, PoI, and mobility. We summarize several findings based on our analysis. Specifically, 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. On the temporal dimension, we found that the representativeness of sensing systems 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. We study the performance of a real-world application on population estimation and its correlation with representativeness. The performance of population estimation based on single systems is highly correlated with representativeness. We found a high representativeness leads to a 58.2% lower error of population estimation. More importantly, we design a correction model based on single sensing systems and public datasets to improve the representativeness of single sensing systems.

- In the prediction module of urbanSense, after measuring weakness of single sensing systems, we utilize the strength of heterogeneous systems and design models for travel behavior modeling. In dissertation, we use two travel behavior metrics as examples for our travel behavior modeling, i.e., population and travel time. In the first work, we model real-time human population based on heterogeneous cellular networks. We design a model called MultiCell to model real-time urban
populations from multiple cellphone networks with two novel techniques: (i) a network realignment technique to integrate individual cell-tower spatial distributions from multiple cellphone networks for finer granular population modeling; and (ii) a data fusion technique based on cross-network training to design a population model based on multiple network data. We implement MultiCell in the Chinese city Shenzhen based on three cellphone networks with 10 million active users and their daily data records at 11 thousand cell towers. We evaluate MultiCell by comparing it to the state-of-the-art models driven by single cellphone networks, and the evaluation results show that MultiCell outperforms them by 27% in terms of accuracy. We cross-validate MultiCell with three transportation systems with more than 8 million passengers to investigate its performances. In the second work, we study the fine-grained travel time patterns in multiple transportation systems, i.e., bus, taxi, personal vehicle, and subway system, under the impact of urban anomalies. Specifically, (i) we investigate implicit components, including waiting and riding time, in multiple transportation systems; (ii) we measure the impact of real-world anomalies on travel time components; (iii) we design a learning-based model for travel time component prediction with anomalies. Different from existing studies, we implement and evaluate our measurement framework on multiple data sources including four city-scale transportation systems, which are (i) a 14-thousand taxicab network, (ii) a 13-thousand bus network, (iii) a 10-thousand private vehicle network, and (iv) an automatic fare collection system for a public transit network (i.e., subway and bus) with 5 million smart cards.

The remaining part of the dissertation is organized as follows. Chapter 2 introduces the framework of urbanSense and its data sources. We present a data bias measurement study in Chapter 3 to understand representativeness of single systems, followed by two works for real-time population modeling in Chapter 4 and travel time estimation in Chapter 5. We discuss our future work and conclude the dissertation in Chapter 6.
Chapter 2
Framework and Datasets

In this chapter, we introduce the framework of \textit{urbanSense}, in which we study the urban travel behaviors with three modules. Moreover, we present urban sensing systems and external data sources as contextual information in our analyses.

We illustrate the framework of \textit{urbanSense} system in Fig. 2.1 with three major modules: (i) a \textit{sensing} module; (ii) a \textit{measurement} module; and (iii) a \textit{prediction} module. We briefly introduce the three modules as follows. (i) The \textit{sensing} module is to collect and process spatio-temporal records from heterogeneous sensing systems (e.g., buses, taxis, personal vehicles, subways and cellular networks) for the data format with unified spatial and temporal granularity. A record collected from a sensing system consists of both spatial (e.g., GPS location, station name) and temporal (e.g., time, day, month,
year) information of a vehicle or a cellphone. However, there are erroneous and noise records in sensing data caused by GPS shifting and abnormal users. Moreover, because of the diverse sensing devices and communication frequencies of the sensing systems, the collected sensing records are with different spatial and temporal granularity. Therefore, the data sensing module has two functions, i.e., data clean and processing data into the certain spatial and temporal dimension. Moreover, we introduce external contextual datasets for our correlation analysis. We elaborate on the sensing systems and data preprocessing in the later sections of this chapter.

(ii) The measurement module is to analyze the unique features and weakness of single sensing systems for capturing human travel behaviors. In the dissertation, we take representativeness as an example for a deep analysis. In particular, we propose a measurement metric, representativeness distance to quantify the data bias in single sensing systems. We found representativeness plays an important role in real-world applications with urban sensing data. We measure representativeness of sensing systems under various spatial, temporal, and contextual factors, e.g., PoIs. We found the key reason for representativeness is the contextual diversity. We model the contextual diversity as a combination of population, mobility and PoI distribution. In general, a higher contextual diversity leads to a higher representativeness, i.e., a denser population, a higher entropy of mobility and PoI distribution. Based on the findings, we design a correction model to improve the representativeness of sensing systems. With the proposed correction model, we can improve the representativeness by 30% for a better performance of real-world applications. The details of the measurement study to understand sensing systems are given in Chapter 3.

(iii) The prediction module is to overcome the weakness of single sensing systems and predict human travel behaviors from heterogeneous systems. In the dissertation, we demonstrate two works to model travel behaviors, i.e., real-time population modeling and travel time estimation. In the first work, real-time locations of crowds are of great importance for many real-world applications such as mobile computing and social networks. We model real-time human population distribution at specific locations in cities. In the second work, we decompose direct travel time from data into logistic
stages (e.g., walking, riding, waiting) in four sensing systems (i.e., subway, taxi, bus, and personal vehicle systems). We predict travel time and travel delay under the impact of anomalies, e.g., car accidents. The details of the two works are given in Chapter 4 and Chapter 5.

2.1 Heterogeneous Urban Sensing Data

In the section, we first introduce urban sensing systems and their data formats, and then the contextual information for our analysis and modeling. Further, we describe the preliminary processing for the spatio-temporal data.

2.1.1 Cellular Sensing System Data

We have been collaborating with three major cellphone networks in Shenzhen for data access to model urban population. In this version of MultiCell implementation, we consider three cellphone networks from complementary perspectives. For privacy issues, we use Network A, B and C.

- Network A includes 3.8 million active users and different types of cellphone usages, e.g., phone call, message, data connection, around the whole Shenzhen city. On average, the daily data in Network A contain 210 million records across 3595 towers.

- Network B includes 2.5 million active users from 2977 towers in Shenzhen city. On average, the daily data in Network B contain 200 million records across 2977 towers in the whole Shenzhen city.

- Network C includes 3.9 million active users and the only type of usage of the record is call. It contains 93 million daily records in 5174 towers.

2.1.2 Vehicular Sensing System Data

We introduce three vehicular fleets and visualize their spatial distribution in Figure 2.3 where a lighter color (yellow) indicates a denser distribution of vehicles.
Fig 2.2: Cellular Sensing Systems

Fig 2.3: Vehicular Sensing Systems

- **Bus**: There are two data sources in the bus system: The first data source is an automated fare collection (AFC) system, which records the trip records of bus passengers with smartcard swiping. The bus AFC data record includes passenger id, tap-on station, tap-on time, tap-out station, tap-out time, bus id, bus route. The second data source is on-board bus GPS trackers, which has two components: a GPS recorder and a communication component. The GPS recorder logs the current bus GPS and time information, and the communication component sends the data to a cloud server. The record collected by the GPS tracker includes id, longitude, latitude, time. We access the Shenzhen bus fleet including 976 bus lines, 13 thousand buses and their data. The bus fleet has a regular pattern due to their operating routes.

- **Taxi**: For the taxi system, in addition to the on-board GPS tracker, one more bit is recorded to indicate if the taxi is occupied by any passenger. Therefore, one taxi record includes id, longitude, latitude, time, occupied or not. The taxi fleet in Shenzhen has 14 thousand taxis generating one status record every 30 seconds.

- **Personal Vehicle**: We access Private Vehicle (PV) GPS trace data by an on-board GPS recorder, which is installed by an insurance company for pay-as-you-go insurance programs. The GPS location is sent to the insurance company. The PV
GPS records are similar to bus GPS records, containing id, longitude, latitude, time. We have access to this private vehicle network with more than 293 thousand vehicles, among which 10 thousand vehicles are in Shenzhen.

### 2.1.3 Payment Sensing System Data

- **Subway Payment System**: An automated fare collection (AFC) system is utilized to collect passenger’s trip origins and destinations in the subway system when passengers tap in and tap out of subway stations. The Shenzhen subway AFC system includes 8 subway lines, 215 subway stations, 4 million daily users. A data record includes passenger id, timestamp, AFC machine id, station name, IN/OUT where IN/OUT is a status for either tap in or tap out a station, timestamp records the time when passenger tap in/out the subway system.

- **Bus Payment System**: Similar to subway AFC systems, automated fare collection machines are deployed on buses. Therefore, a bus payment system records passenger id, timestamp, AFC machine id, bus id, bus route. The Shenzhen bus AFC system includes 13 thousand buses. There are three major differences between subway AFC systems and bus AFC systems: (i) A bus AFC record is generated when a passenger get on a bus when a subway record is generated when a passenger tap in and tap out the subway system; (ii) A bus AFC machine moves with a bus. A bus AFC record does not include a station name. Therefore we cannot infer the location where a passenger get on a bus based on the bus AFC records. Instead, the station can be inferred by integrating bus AFC records with
on-board GPS data from vehicular sensing records; (iii) A subway AFC system can always capture activities of all passengers in the subway system while a bus AFC system can only capture passenger activities with metro cards, e.g., it cannot capture passengers who pay travel fee with cash.

2.2 Contextual Data

We collect various types of contextual datasets, e.g., population distribution, Point of Interests (PoI) distribution, road networks, and anomalies.

2.2.1 Population

We extract Shenzhen population from Worldpop [31], which gives fine-grained population distribution in $100m \times 100m$ grids. Fig. 2.5a 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 regions with different spatial partition due to the coverage of sensing devices. We study the impact of population on human mobility in measurement study and integrate population as a contextual factor in the human mobility prediction.

2.2.2 Point of Interests (PoI)

The function of regions is one important factor to determine human mobility patterns on the spatial dimension [85]. For instance, more human mobility activities 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. 2.5b visualizes PoIs on a city map. We expect that regions with PoIs
from different categories may have different human mobility patterns.

### 2.2.3 Road Networks

Human mobility always follows the topological distribution of road networks. We introduce Shenzhen road network for human mobility modeling. OpenStreetMap [41] is one of the most active crowd-sourcing map services. To date, there are 72,676 road segments in Shenzhen from OpenStreetMap, of which the total length is 10,711 km. Since OpenStreetMap includes detailed branches and paths and differentiates lanes from opposite directions, we found the road length calculated from the OpenStreetMap database is larger than the road reported by governments. The geographic distribution of roads is presented in Fig. 2.5c. We take topological and statistical distribution of road networks in human mobility modeling.

### 2.2.4 Anomalies

Transportation accidents are one of the major unexpected factors that affect human mobility, e.g., travel time. However, accident datasets cannot be collected through city infrastructures directly with highly accurate spatial and temporal information. We list the investigated anomalies in two categories, i.e., expected and unexpected anomalies, in
Table 2.1, and show anomalies on city map in Fig. 2.5d where a larger circle indicates a higher degree of anomaly. We describe the procedures to collect anomaly data as follows.

<table>
<thead>
<tr>
<th>Anomaly Category</th>
<th>Anomaly Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected</td>
<td>concerts, musical festivals, college entrance exams, subway fire drill, sports</td>
</tr>
<tr>
<td>Unexpected</td>
<td>station entrance of stagnant water, subway delay, facility malfunction, accidents</td>
</tr>
</tbody>
</table>

**Unexpected Anomalies**

(i) *Retrieval*. We firstly build a web spider to extract and obtain all the posts from official accounts of the Shenzhen Transportation Police from *Sina Weibo*, i.e., the Chinese version of *Twitter*. We filter the raw data with the keywords *road condition* or *accident*. By this approach, we focus on the relevant information and use the texts posted on the website for anomaly analysis.

<table>
<thead>
<tr>
<th>Description keywords</th>
<th>Anomaly Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>scrape, small cars, slow</td>
<td>1</td>
</tr>
<tr>
<td>bus, van, injured, several cars</td>
<td>2</td>
</tr>
<tr>
<td>fire, dead, construction</td>
<td>3</td>
</tr>
</tbody>
</table>

(ii) *Structuralizing*. We extract the useful information from the texts we have crawled in the previous step. We apply regular expressions to obtain the time and location of the accidents. To further quantify the extent of anomaly, we classify them into different *anomaly levels* according to the keywords in the description provided. For example, if the keywords such as “small cars” and scrape” are mentioned, we consider it as a low level of severity. The details of anomaly level and its corresponding keywords are provided in Table 2.2.

(iii) *Matching locations*. *Geocoding* is the process of converting addresses (e.g., “1600 Amphitheatre Parkway, Mountain View, CA” or “Intersection of 5th street and 62nd street in NYC”) into geographic coordinates (e.g., latitude 37.423021 and longitude
-122.083739). *Geocoding* is conducted with two steps, i.e., *Entity Extraction* and *Location Matching* with Google Map API [69].

(iv) *Data cleaning*. We clean out some unreasonable and unrealistic data such as the ones with coordinates out of the region of Shenzhen. Finally, we obtain Table 2.3 to describe every item in the data set of unexpected anomaly.

<table>
<thead>
<tr>
<th>Events</th>
<th>Description</th>
<th>Time</th>
<th>Location</th>
<th>Latitude</th>
<th>Longitude</th>
<th>Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>many cars collided</td>
<td>2016-06-01</td>
<td>Beihuan Ave, Shahe West Overpass</td>
<td>22.5682</td>
<td>113.9543</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>two cars scraped</td>
<td>2016-06-01</td>
<td>NanPing Exp, TangLangShan Road</td>
<td>22.5674</td>
<td>113.99177</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>a van turned over</td>
<td>2016-06-04</td>
<td>ShenShan Highway</td>
<td>22.5398</td>
<td>114.0195</td>
<td>3</td>
</tr>
</tbody>
</table>

**Expected Anomalies**

Expected social events such as Marathon, concerts and music festivals have a large impact on passengers’ travel behavior. We collect these social events by crawling and analyzing online news. The crawling procedures are similar to the irregular anomaly extraction, which includes *event retrieval*, *event structuralizing*, *location matching* and *data cleaning*. We show the samples of the data set of regular anomaly in Table 2.4.

<table>
<thead>
<tr>
<th>Events</th>
<th>Description</th>
<th>Time</th>
<th>Location</th>
<th>Latitude</th>
<th>Longitude</th>
<th>Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Marathon</td>
<td>2016-12-04</td>
<td>Baoan People's Government</td>
<td>22.5536</td>
<td>113.8843</td>
<td>3</td>
</tr>
<tr>
<td>B</td>
<td>Midi Music Festival</td>
<td>2016-01-01</td>
<td>Universiade Shenzhen Gymnasium</td>
<td>22.6933</td>
<td>114.2190</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>Emergency Drill</td>
<td>2016-06-26</td>
<td>Futian Subway Station</td>
<td>22.5416</td>
<td>114.0526</td>
<td>1</td>
</tr>
</tbody>
</table>
Chapter 3
Measurement for Sensing Representativeness

3.1 Introduction

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 [92] 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 [96], load balancing [104], and network resilience [115]. To date, many efforts have been focused on cellular usage patterns and applications, e.g., traffic patterns [88] [111], user behaviors [22], tower deployment [111], special events [83], and mobility management [5], 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 [12] [50] [84] [87]. However, 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 [111]. Even though some studies are based on the data from multiple networks [13] [63] [90] [100] [107], the data are collected at a small scale, e.g., a dozen devices [100], 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 [1], many cities have been consolidating various data from diverse infrastructures [51] [25] [24]. 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 [23] [102] [78] [27], 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.

We conduct the first 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 is when, where, to what extent, and why 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.

- 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. We study the performance of a real-world application on population estimation and its correlation with network representativeness and summarize our 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.

- 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.

- Last but not least, based on our analysis, we share several implications and discuss the potential impact of our study. Our analyses are involved with large-scale cellular web log data covering 10 million users in the same city with unique data access and structured contextual datasets.

3.2 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 [83], web user estimation [54]
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 [50], inferring internet usage in Shanghai [111]. 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 [28]. In fact, tower deployment strategies are dependent on various factors such as communication technologies, usage demand, geographic and demographic information in regions [86]. In particular, we found that the tower coverage differs in the three networks, as shown in Fig. 3.1 when we model tower coverage by Voronoi partition, which is widely used to estimated cell tower coverage boundary [28]. 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.

![Fig 3.1: Cell Size](image1)

![Fig 3.2: Impact of Bias on App.](image2)

**Impact of Bias on Real-world Applications:** Relying on log data records for

1. ...
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 [113] of cellular users. More detailed settings are given in Section 3.4. We use MAPE (Mean Absolute Percent Error) to quantify the performance of the same models with different datasets from three networks. Fig. 3.2 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.

**Summary:** To quantitatively understand where, when, why and to what extent cellular data bias happens, and more importantly how to correct and alleviate such data bias from only one network, we conduct a comprehensive empirical study based on real-world cellular network data from three cellular providers covering 100% cellular users in Shenzhen. We model the bias in cellular data with the term representativeness, which measures the degree of single network data or sample data of a single network to be representative enough for all users across all networks in a city. Besides, we study the potential contextual reasons for representativeness differences and implement a real-world application to study the correlation between their performance and representativeness.

### 3.3 Measurement

#### 3.3.1 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.

**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 [28] 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. 3.3, 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.

**Temporal Partition:** We partition time into 10-minute time slots. In other words, we 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 [28] [111] [113].

<table>
<thead>
<tr>
<th>Notation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P$</td>
<td>a spatial partition</td>
</tr>
<tr>
<td>$r, R$</td>
<td>a region and a region collection</td>
</tr>
<tr>
<td>$t, T$</td>
<td>a time slot and a time slot collection</td>
</tr>
<tr>
<td>$k, K$</td>
<td>a network and a network collection</td>
</tr>
<tr>
<td>$l, L$</td>
<td>a load and a load collection/distribution</td>
</tr>
<tr>
<td>$\tilde{L}_r$</td>
<td>normalized load distribution</td>
</tr>
<tr>
<td>$l_{k}^{r,t}$</td>
<td>a load of network $k$ at region $r$ in time slot $t$</td>
</tr>
<tr>
<td>$L_{K}^{R,T}$</td>
<td>a load collection given $K$, $T$ and $R$</td>
</tr>
<tr>
<td>$\epsilon$</td>
<td>a tolerant threshold for representativeness</td>
</tr>
</tbody>
</table>

Table 3.1: Terminologies
Load Distribution: The number of phone calls or internet calls is described as load. \( L = l_1, l_2, \ldots, l_n \) represents the load distribution where \( l_1 \) to \( l_n \) is the load in a specific region in a specific time slot. We use a subscript \( k \) to differentiate loads from different networks, e.g., \( L_k \); we use \( L_\forall \) 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., \( L^r \), or at specific time, e.g., \( L^t \).

Representativeness Distance: Intuitively, \( L_i \) is a representative of \( L_\forall \) if \( L_i \) can be scaled to \( L_\forall \) by a scaling factor \( \alpha \). Similarly, to study if a network can be used as a representative of all networks, we use Representative Distance \( \theta_k \) \((0 \leq \theta_k \leq 1)\) 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 (3.1).

\[
\theta_k = \min_{\alpha} \| \tilde{L}_\forall - \alpha \tilde{L}_k \|_\infty; \\
\tilde{L} = \frac{L - \min(L)}{\max(L) - \min(L)}; \quad L_\forall = \sum_{k=0}^{K} L_k; \quad (3.1)
\]

We illustrate our idea in the example with a single network \( L_A \) load distribution and the total load among all networks \( L_\forall \) in Fig. 3.4. \( L_A \) and \( L_\forall \) 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 \( \alpha \) 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. 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 \( L_\forall \) and \( l_i \) in \( L_A \) 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 $\theta$ leads to a low value of similarity or mean difference but not vice versa.

![Fig 3.4: Representativeness Distance](image)

**Tolerant Parameter $\epsilon$.** We define a tolerant parameter $\epsilon$ 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. 3.3 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.

### 3.3.2 Methodology

We conduct our study on representativeness from three perspectives:

(1). **Findings and Causalities (Section 3.3.3 and 3.3.3):** we categorize the measurement results on spatial, temporal dimensions. (i) On the spatial dimension, we study load distribution $L_r^r$ for different $r$. For each region $r$, $L_r^r$ describes loads at different time slots. (ii). On the temporal dimension, we study load distribution $L_t^r$ for different slots $t$. For each time slot $t$, $L_t^r$ describes loads in different regions, i.e., elements $l$ in $L_t^r$ 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 3.3.4): we select 4 regions with different contextual information distribution for a detailed study to validate our findings.

(3). Correction (Section 3.4): 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.

3.3.3 Evaluation

(I) Spatial Representativeness. Fig. 3.5 shows Representative Distance $\theta$ 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 $\epsilon$ in Fig. 3.6. When the threshold $\epsilon$ 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.

![Fig 3.5: Regions](image1)

![Fig 3.6: Spatial Groups](image2)

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) [28] and dynamic features of the users (e.g., mobility) [123]. 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 distribu-

![Fig 3.7: PoI v.s. Spatial](image1)

![Fig 3.8: Groups v.s. Spatial](image2)

tion 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.

**Impact Factor 1: Region PoI.** For each administrative region in Fig. 3.3, 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 in 17 subgroups for each region $- \sum_{i=1}^{17} p(x_i) \log_2 p(x_i)$ where $x_i$ is the number of PoIs in a subgroup $i$. We found that a higher PoI entropy leads to a lower representative distance as in Fig. 3.7. 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. 3.8 and Tab. 3.2. In Fig. 3.8, we set $\epsilon$ 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 Tab. 3.2 where we found

<table>
<thead>
<tr>
<th>Group</th>
<th>Cluster</th>
<th>Residence</th>
<th>Transport</th>
<th>Office</th>
<th>Recreation</th>
<th>Edu</th>
</tr>
</thead>
<tbody>
<tr>
<td>TR</td>
<td></td>
<td>0.18</td>
<td>0.23</td>
<td>0.21</td>
<td>0.22</td>
<td>0.16</td>
</tr>
<tr>
<td>PR</td>
<td></td>
<td>0.14</td>
<td>0.26</td>
<td>0.28</td>
<td>0.20</td>
<td>0.12</td>
</tr>
<tr>
<td>NR</td>
<td></td>
<td>0.30</td>
<td>0.18</td>
<td>0.13</td>
<td>0.18</td>
<td>0.21</td>
</tr>
</tbody>
</table>

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 [31], which gives fine-grained population distribution in 100m × 100m grids. Fig. 2.5 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. 3.9. 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 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 2.1. 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. 3.10. 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.

(II) Temporal Representativeness. As shown in Fig. 3.11, 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. 3.12. Compared with spatial representativeness groups as in Fig. 3.6, temporal representativeness groups present a lower representativeness thresholds. It indicates the spatial dimension has a higher variance of representativeness, which motivates us to correct the representativeness mainly from the spatial dimension in Section 3.4.

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 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 Section 2.1. A lower entropy indicates a less random (i.e., less diverse) distribution of user mobility as shown in Fig. 3.13. In other words, most passengers are mainly mov-
ing 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. 3.14. 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 potentially leads to a high representativeness distance, which may be because most passengers are moving between high-demand regions.

**Weekly Pattern:** We further study weekly patterns of representativeness as shown in Fig. 3.15. 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 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.
### 3.3.4 A Case Study

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. 3.16. 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. 3.16. The number of towers in the four selected locations with a certain radius is given in Tab. 3.3 from the least number of towers to the most number of towers.

<table>
<thead>
<tr>
<th>Factors</th>
<th>PoI Diversity</th>
<th>Population</th>
<th>Mobility</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBD</td>
<td>High</td>
<td>Large</td>
<td>High</td>
</tr>
<tr>
<td>Train Station</td>
<td>Medium</td>
<td>Large</td>
<td>High</td>
</tr>
<tr>
<td>Residential</td>
<td>Low</td>
<td>Medium</td>
<td>Medium</td>
</tr>
<tr>
<td>Airport</td>
<td>Low</td>
<td>Small</td>
<td>High</td>
</tr>
</tbody>
</table>

![Fig 3.16: Case Study Areas and Their Contextual Diversity](image)

**Table 3.3: Tower Distribution on Select Locations**

<table>
<thead>
<tr>
<th>Radius</th>
<th>1 km</th>
<th>2 km</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>airport</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>residential</td>
<td>25</td>
<td>17</td>
</tr>
<tr>
<td>train station</td>
<td>38</td>
<td>30</td>
</tr>
<tr>
<td>CBD</td>
<td>58</td>
<td>56</td>
</tr>
</tbody>
</table>

We compare their representative distances in Fig. 3.17, 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 3.3.3 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. 3.18. 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.

Fig 3.17: Studied Locations  
Fig 3.18: Distance to Centers

3.4 Correction

3.4.1 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.

3.4.2 Terminologies

We first introduce terminologies for diversity modeling as in Tab. 3.4 and then formalize our target problem.

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

Table 3.4: Terminologies

**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 [50] [115]. (2) *Mobility Matrices*: For each grid $g$, we construct two matrices to describe its mobility patterns in the grid: a *From* matrix $M^g_{from}$ to describe the number of passengers moving from grid $g$ to other grids in different time slots. a *To* matrix $M^g_{to}$ 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 \(D_g\), 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 \(v^j_r\) in \(\mathcal{V}_g\) is a probability that the grid \(r\) has a function of region, e.g., education. 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 [120]. 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 \(E_g\) with an entropy of vector \(\mathcal{V}_g\), which is one of the most common measurement for randomness of elements in a set [89]. 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.

**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 \(\alpha\), our target is to select a sub set of grids \(G\) from all equally-sized grids to maximize the contextual diversity \(E_G\) under a constraint that the size of \(S^G\) is equal to \(\alpha \cdot |S|\). All the data \(S^G\) from this sub set of grids \(G\) are our data sample. The Equation 3.2 gives the
formulation.

$$\arg\max_{G} \mathcal{E}^G$$

s.t. $\sum_{g \in G} |S^g| = \alpha \cdot |S|$ \hspace{1cm} (3.2)

$|G \cap r| \geq 1, \forall r \in R$

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.

3.4.3 Methodology

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 3.2. 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. 3.19: (i) diversity modeling; (ii) diversity-maximization sampling. We elaborate on these two steps as follows.

ALGORITHM 1: Diversity-Driven Sampling

\begin{algorithm}
\begin{algorithmic}
\State \textbf{Input:} $\alpha, S^U, P^U, D^U, M^U_{\text{from}}, M^U_{\text{to}}$
\State \textbf{Result:} $S^G$
\State metadata $\leftarrow (P^U, D^U)$;
\State words $\leftarrow (M^U_{\text{from}}, M^U_{\text{to}})$;
\State $V^U$ $\leftarrow$ topicClustering(metadata, words);
\State $G$ $\leftarrow$ initialize() ;
\State $C$ $\leftarrow$ $U - G$;
\While{$|S^G| < \alpha \cdot |S^U|$}
\State $E^G$ $\leftarrow$ entropy($V^G$) ;
\State $\Delta E^C$ $\leftarrow$ entropyGain($V^G, V^C$);
\State $g$ $\leftarrow$ argmax$_{g \in C} \Delta E^g$;
\State $G$ $\leftarrow$ $G \cup \{g\}$;
\State $C$ $\leftarrow$ $C / \{g\}$ ;
\State $S^G$ $\leftarrow$ $S^G \cup S^g$
\EndWhile
\end{algorithmic}
\end{algorithm}

(i) \textbf{Diversity Modeling:} In diversity modeling, we generate a vector $V = \{v_1, v_2, \cdots, v_n\}$ for each grid where $n$ is the number of potential region functions (e.g., education, office, etc), and each element $v_i$ in $V$ is the probability that a grid belongs to a function. In general, a higher entropy on $V$ indicates a more diverse distribution on region functions, thus a larger contextual diversity in a region. To construct such a $V$ from contextual information of a grid, e.g., population, mobility and PoI distribution, we apply a topic model [42] [43]. Topic models such as LDA [8] 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 \textit{region function clustering to document topic clustering} is as follows: we map \textit{grids to documents}; \textit{region functions to document topics}; the dynamic feature, i.e., \textit{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 $V$ 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 $V^g$ where $V^g = \{v^g_1, v^g_2, \ldots, v^g_n\}$ and $v^g_i$ is the possibility a grid $g$ belongs to a region function $i$. Fig. 3.19 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 $M_{from}$ and $M_{to}$ have 3 rows for 3 time slots and 4 columns for 4 regions. The topic model will generate a 5-dimension vector $V$ 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 3.2. Second, we calculate the entropy gain based on $E^G$ for every grid in $U - G$. We select the grid $g$ with the highest entropy gain, i.e., the diversity gain, and then update $G = G \cup \{g\}$. Third, we update the $E^G$ 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. 3.19, the number of sample records are $\alpha \cdot |S| = 9$ 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 $V^{g_2}$ has the largest entropy with one region selected and then select 5 records from $g_1$ since we have the largest entropy in $V^{\{g_1,g_2\}}$. The process is described in Algorithm 1.
3.4.4 Evaluation

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 SysRep 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 \( \theta \) as the metrics for the evaluation, a lower representativeness distance indicates a higher representativeness.

Fig 3.20: Performance  
Fig 3.21: Sample Ratio

Overall Results: We compare the performance of SysRep with two baselines in Fig. 3.20. Both the baseline model SysSam and our SysRep increases the representativeness by reducing the representativeness distance due to the higher sample until the ratio is 0.6. In particular, our SysRep decreases representativeness distance significantly from 0.31 to 0.16 on average as shown in Fig. 3.21. It shows that with a sophisticated sampling strategy in SysRep, even 30% of sample data from a single network can achieve similar representativeness as all single-network data.

Impact of Factors: We further study the impact of different factors on the performance of SysRep. Fig. 3.22 compares the resulted representativeness distance \( \theta \) with data in the three networks. We found even three networks have different user coverage, they can achieve similar representativeness with SysRep. Therefore, SysRep 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 \( \alpha \) equal to 0.5, 0.6,
and 0.7, respectively. Moreover, we study the impact of spatial granularity in Fig. 3.23, which shows the performance of SysRep with different grid sizes. SysRep achieves the best performance with a grid size $100m \times 100m$. In general, a finer spatial granularity leads to a better performance.

**Impact on Real-World Applications:** Relying on the measurement results, we validate the impact of the 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 [113] and map the estimated population to the administrative regions. We use the Worldpop [31] 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. 3.24, which proves that a higher representativeness distance leads to a worse performance on the application. Fig. 3.25 shows SysRep 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 SysSam, which use a uniform sampling method without considering the contextual diversity.

### 3.5 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 quality 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 [111], web user estimation [3], hot spot recommendations [44].

**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) Deidentification: 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.

3.6 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 [47]. Investigating cellular usage patterns has received considerable attention recently due to data availability. In Table 5.1, 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>
<thead>
<tr>
<th>Table 3.5: Cellular Web Log Analyses Survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Categories</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Data Collection Methodology</td>
</tr>
<tr>
<td>Individual Researcher</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Network Operators</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

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 [72] and connected vehicles [86]). 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.

(i) **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 [72], popular routes construction [45], destination recommendation [65], anomalies spotting [75] and relationships between mobility and PoI [99].

(ii) **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 [107], location prediction [63], urban planning based on location-based social network [90], and existing PoI verification [13].

**Study on Data from Network Operators** Cellular network operators passively collected their network data for billing purposes (e.g., CDR data [50]) or web access logs (e.g., internet access data [68]). 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 [50] [123]. Based on cellular data from a single network, researchers (i) conduct spatiotemporal phone call analysis[12], data call analysis [87], mobile traffic analysis and prediction[12] [115] [111], and dynamic urban geo-social connectivity graph construction [54] (ii) trajectories recovery from mobility data [112], (iii) determine the locations of network upgrades [79], and (iv) improve network performance [19]. 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.

### 3.7 Summary

As an infrastructure for mobile web service, we conduct the first 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. Finally, we design a correction model to improve representativeness in single-network data.
Chapter 4
Prediction Model for Population

4.1 Introduction

Real-time urban population modeling is essential to many applications, e.g., mobile computing [9], urban planning [39], and location-based services [129]. Traditionally, urban populations have been modeled by surveys, e.g., census data [38], which are comprehensive but typically out-of-date and cannot be used for real-time population modeling. Recently, the real-time population study gains great attention because of high penetration rates of location tracking devices and advanced urban infrastructure systems, e.g., cellphone [49], taxi [34], buses [7], subway [56], and smart cards [94]. Based on these systems, we can infer real-time locations of system users, and then model aggregated urban populations.

Among all these systems, although transportation systems offer record data at unprecedented temporal scale [34], cellphone network system has been considered as an effective way to model urban population because of its high penetration rates, low-cost data collection and alleviated privacy concerns [20] [49]. In particular, (i) it has been shown that 96% of world population have cellphones and use them regularly [29], which helps us model real-time urban populations [20] that are challenging to be modeled by other data sources; (ii) the cellphone data are already automatically collected by the cellphone companies for billing purposes [48], which leads to low marginal costs; (iii) the cellphone data are collected at the cell tower level and do not need GPS, which alleviates the energy and privacy issues [6].

To date, many population models have been proposed based on data from cellphone networks [113] [18] [55]. However, we found almost all these models are implemented by data from single cellphone networks while most cities around the world have multiple
networks [49]. The assumption behind these models is that the users in single cellphone networks are representative of all residents using multiple networks in the same city [55]. However, as we validated by our data, different networks have different spatial concentrations due to their strategic plans and market shares (e.g., in Figure 2.2). As a result, the data from one network are typically biased against the users of other cellphone networks in the same city, which leads to overfitting of the models driven by single networks.

In this work, to address this issue, we design a population model called MultiCell based on data from multiple cellphone networks. Inspired by the previous work based on single cellphone networks [113], MultiCell models urban populations but provides new insights from multiple network perspectives. It seems straightforward to simply merge data from different networks together and then feed them to existing population models by considering multiple networks as a large virtual network, which is suggested by [21] with synthetic data. However, we argue that naive data merging leads to biased population models because in practice, different cellphone companies have different spatial distributions of cell towers, and resultant data are biased towards their underlying spatial distributions as discussed in our motivation. To address this challenge, we utilize two novel techniques, i.e., network realignment and cross-network data fusion, to estimate urban population in finer spatial-temporal granularities with multiple networks.

Specifically, our contributions are as follows:

- To our knowledge, we conduct the first study on urban populations based on multiple real-world cellphone networks. Conceptually, we advance existing models based on cellphone networks from two dimensions (i.e., spatiotemporal) to three dimensions (i.e., spatiotemporal and network). Our study is based on real-world data capturing more than 10-million users. We provide empirical evidence for two facts: (i) cellphone data from individual networks have a spatial bias against users in other networks; (ii) integrating multiple networks enables finer-grained population modeling while keeping original spatial structures.
• With these data-driven insights, we design a population model MultiCell based on multiple cellphone networks. We address a core challenge for multi-granularity data fusion from different networks with two techniques: (i) we design a network realignment technique to integrate individual spatial partitions of multiple cellphone networks for fine-granular population modeling; (ii) we design a data fusion technique based on cross-network training for a population model based on multiple networks.

• We implement MultiCell with three cellphone networks in the Chinese city Shenzhen based on three months of cellphone data. These networks have 3.8 million, 2.5 million, and 3.9 million daily active users with 3595, 2977, and 5174 cell towers, respectively. The total daily data records for these three networks are more than 500 million. It covers all cellphone users and achieves 96% population penetration rate. To our knowledge, MultiCell is one of the largest urban phenomenon models in terms of user numbers, and more importantly, the first population model driven by multiple real-world cellphone networks.

• We evaluate MultiCell by comparing it to state-of-the-art models driven by single cellphone networks, and the results show that MultiCell outperforms them by 27% in terms of accuracy. We further evaluate MultiCell with various transportation data to investigate the correlation between our population model and transportation ridership. We found that our population model has a high correlation with taxi, bus and subway systems with more than 6 million daily passengers. This is the first work investigates urban population from such a comprehensive multi-system perspective.

As follows, Section 4.2 shows our motivation. Section 4.3 elaborates the model instantiation on multiple cellphone networks. Sections 4.3.5 and 4.4 are the implementation and evaluation, followed by related work in Section 4.5. Section 4.6 concludes the work.
4.2 Motivation

Spatial Biases: The previous research on cellphone data-driven population modeling relies on data generated from a single network. However, cellphone network companies typically have different business priorities in terms of geographic locations, which leads to significantly different cell tower distributions and thus different user numbers. In fact, the cell tower spatial distribution of a cellphone network is dependent on various factors including the technologies they are using, the region-specific geographic and demographic information [29]. As a result, different networks in the same city may have very different tower distributions, which lead to a bias for population modeling if only data from a single network are utilized. To provide empirical evidence, we utilize data from three networks (details are in Section 2.1) to calculate cellphone user population for 496 administrative regions in Chinese city Shenzhen. To calculate the population in regions, we first calculate the intersected areas of Voronoi partition and administrative regions as shown in Fig. 4.1.

\[ U(R_x, t, i) = \sum_{l=0}^{n} \frac{|R_x \cap C^i_l|}{|C^i_l|} \times U(C^i_l, t, i), \] (4.1)

where \( U(R_x, t, i) \) is the user population in a region \( R_x \) based on data from a network \( i \) in a time slot \( t \); \( U(C^i_l, t, i) \) is the user population in cell \( C^i_l \); \( |R_x \cap C^i_l| \) is the area of \( R_x \cap C^i_l \); \( n \) is the number of cells intersected with \( R_x \). The average area of the 496
regions is 3.97382 km² and the standard deviation is 6.075980 km².

Fig 4.2: User Difference  

Fig 4.3: Users Difference at Region Level

Fig. 4.2 gives the difference on cellphone users in these 496 regions between cellphone network with most users and cellphone network with least users in a region. We ranked these regions based on the difference in user numbers. We found that network A has more users in 249 regions; network B has more users in 67 regions; network C has more users in 180 regions. To explore if there are any spatial patterns by any network, we further visualize these regions to show user populations of three networks in Fig. 4.3. Fig. 4.3 gives all 496 regions in Shenzhen. The blank regions have more users in the network A; whereas the dark regions have more users in the network B and the red regions have more users in the network C. Compared with the population distribution in Fig. 4.4, we found that there are no clear spatial patterns about the regions dominated by any of networks. It indicates that if only data from one network are used for modeling, the resultant models may experience overfitting in the regions dominated by this network, and vice versa. Further, a straightforward method to combine data from multiple networks for modeling cannot work because different networks have different concentrations in different regions as in Figure 4.3. In this work, we aim to explore the possibility of combining multiple networks to model real-time population with a new technique based on co-training to iteratively utilize multiple networks to optimize population models.

Spatial Granularity: The coarse spatial granularity is the key disadvantage for models driven by cellphone data [49] because we can only infer user locations on the tower level. For example, in Shenzhen with a total area of 1,991 km², the three networks we studied have 3,585, 2,977 and 5,174 towers. Each of towers leads to an irregular
Population Distribution in Regions

Fig 4.4: Population Distribution in Region Level

Fig 4.5: Cell Coverage

cell with an average area of 530 thousand m$^2$, 640 thousand m$^2$ and 385 thousand m$^2$ respectively in three networks. However, the desired spatial granularity for population study is 100m $\times$ 100m = 10 thousand m$^2$ for dense areas [31]. For example, in Figure 4.5, we show the number of cells in the Shenzhen downtown from three networks. We found that in the Shenzhen downtown (i.e., roughly 20% of Shenzhen), there are 2578, 2057 and 3605 cells for three networks respectively, which leads to the average cell areas of 140 thousand m$^2$, 180 thousand m$^2$ and 100 thousand m$^2$. But they are still one order of magnitude larger than the desired granularity. Moreover, the penetration rate of single networks is low. The user distribution on the spatial dimension is biased in single networks. In our implementation, our model cover 100% cellphone users and 96% of the total population in Shenzhen. It eliminates the spatial bias caused by the user distribution and reduces the spatial granularity to finer-grained regions, which is described in later sections.

**Temporal Dynamics:** Due to the limitation of access to real-time dynamic population data, traditional regression techniques estimate human population in a city by static models. For instance, Worldpop [31] dataset provides population distribution in 2010 and 2015. Single network estimation models can be only updated in 2010 and 2015 and remain static in any time between these two years due to the lack of ground truth for model training. Since the temporal dynamics exists in human mobility, e.g., inter-cities or intra-city, the static model introduces the bias in the temporal dimension. In this work, relying on the strengths of multiple cellphone networks, we designed a highly dynamic model for the population estimation.

**Summary:** By comparing single and multiple network scenarios, we found (i) the
data from single networks have spatial biases at different urban regions, which motivates us to fuse data from multiple networks to address biases; (ii) the single network has a coarse spatial granularity for modeling, which motivates us to study intersections of cells from multiple networks to explore a finer spatial granularity; Almost all existing work aims to train a model based on cellphone data from two dimensions (e.g., temporal or spatial), e.g., finding data for the similar time slots, or finding data for the similar locations. Conceptually, MultiCell advances existing models based on cellphone networks from two dimensions (i.e., spatial and temporal) to three dimensions (i.e., spatial, temporal and networks).

4.3 Methodology

4.3.1 Core Idea

We introduce how our core philosophy of multiple networks advances the state-of-the-art population modeling based on single networks in Figure 4.6.

![Core Idea Diagram](image)

We have our desired output: a dynamic general population model on the right to show urban-scale real-time populations where we have a temporal dimension (e.g., a slot), a spatial dimension (e.g., a cell) and an entry which is a general population in this cell during this slot. To obtain this output, the state-of-the-art models (e.g., [113]) utilize (i) user population from single networks (i.e., the first 2D matrix) and (ii) static general
population data without temporal dynamics (i.e., a ground truth vector such as census data collected every 10 years). Since static general population data used as ground truths only have spatial dynamics but no temporal dynamics, existing models typically use spatial training by selecting spatial data (e.g., different rows in the first matrix), which leads to limitations. In contrast, our work utilizes multiple networks to form a tensor (i.e., a 3D input as in Figure 4.6) and then combines static general population data to obtain the desired output. As a result, our work provides a new dimension (e.g., different layers of the tensor), which provides valuable diversity.

![Model Framework](image)

**Fig 4.7: Model Framework**

We show the framework of our MultiCell model according to the data flow in Fig. 4.7. **MultiCell** has three key components: (i) **Tower-Based Partition** where we generate tower-based cell regions for individual cellphone network; (ii) **Spatial Alignment** where we utilize heterogeneous tower distributions from multiple networks to obtain a fine partition for population modeling; (iii) **Population Estimation** where we first apply a Gaussian filter to map cellphone data to the new partition and then design a co-training technique to fuse population estimations obtained by individual networks.

### 4.3.2 Tower-based Partition

In the tower-based partition, we divide a city into different cells based on cell towers belonging to the same network. Given a particular network with a fixed number of towers, we apply the Voronoi diagram to generate a partition based on locations of these towers, similar to the previous work [73]. This partition divides a city into different cells where every point in a cell is closest to its massive centroid, i.e., a tower in our case. Note that this kind of partitions is based on the case that cellphones are...
connected to the geographical closest tower. Even though there are cases where cellphones are connected to a farther tower because of specific communication technologies (e.g., congestion control [121]) used by different networks, we cannot obtain such detailed information based on cellphone data, and so almost all existing models driven by cellphone data are under this assumption [113]. Based on these resultant tower-based partitions, we introduce how to align them as follows.

### 4.3.3 Spatial Alignment

For the state-of-the-art models based on single networks [113] [18] [55], their spatial partition is straightforward because all cell towers belong to a single network, which leads to a non-overlapping Voronoi partition, e.g., as shown in Fig. 2.2. But as shown in our motivation, such a partition typically has large cells due to limited cell towers. In contrast, MultiCell has cell towers from different networks. A straightforward yet trivial solution is to combine all cell towers and data from different networks to form a large virtual network and then apply an existing population modeling technique, e.g., [113]. But such a solution leads to inaccurate modeling where all users belonging to a large cell have to be assigned to much smaller subcells. In this work, to address this issue, we first perform tower-based partitions for each network separately, and then spatially align all these partitions together at the cell level. This cell-level spatial alignment ensures that users in different networks are still distributed within original cells. As follows, we introduce these two components, respectively.

Based on multiple tower-based partitions, we integrate them for a cell-level alignment where the cells from one network intersect the cells from other networks. Thus, we utilize these intersections to form a new partition, i.e., an intersection-based partition. Such an intersection-based partition has a finer granularity than all tower-based partitions because a cell in an original tower-based partition can be intersected by many cells from other networks. We define these intersections as subcells, which are our spatial unit for modeling. The user population in subcells is dependent on all original cells from all networks. 

Figure 4.8 gives an example of our cell-level alignment with two networks. Based on two
tower-level partitions for two networks, we have the intersection-based partition with 15 subcells in MultiCell, among which 7 subcells are shown. These subcells typically (i) have much smaller areas than the original cells from single networks, and (ii) still have original spatial cell structures (e.g., coverage boundaries) compared to a uniform grid partition [31]. Based on these subcells, we model in a finer granularity compared to the existing work focusing on cell-level modeling. As follows, we formalize this cell-level alignment based on tower-based partitions. Given \( N \) tower-based partitions \( P^1, P^2, \ldots, P^N \) for \( N \) networks, we have following constraints for a partition \( P^i \) with cells based on a network \( i \): (i) \( P^i = \{ C^i_1, \ldots, C^i_{|P^i|} \} \) where \( |P^i| \) is the total number of cells in \( P^i \); (ii) \( \bigcup_{m=1}^{|P^i|} C^i_m = U \) where \( U \) is the whole city area, i.e., any partition covers the whole city; (iii) \( C^i_m \cap C^i_n = \emptyset \) where \( m \neq n \), i.e., there is no overlap between cells from the same partition; (iv) \( S_x = C^i_m \cap C^j_n \) where \( i \neq j \), i.e., a set of subcells \( S = \{ S_1, \ldots, S_x, \ldots, S_{|S|} \} \) based on the intersection-based partition where any cell \( C^i_m \) intersects any other cell \( C^j_n \) not from the same tower-based partition.

Based on this intersection-based partition \( S \) with subcells, we estimate population in these subcells as follows.

### 4.3.4 Population Estimation

In this subsection, we formalize our population estimation problem and presents our two-phase model for population estimation.

**Terminologies and Problem Definition:** We summarize the notations used in the
population estimation in Table 4.1. Our goal in the population estimation is to estimate the general population \( \hat{G}(S_x, t) \) in subcell \( S_x \) at time \( t \) given user population \( C^i_l \) for \( \forall l \) where \( i = 1, 2, \ldots, N \), \( N \) is the number of cellphone networks.

We design a two-phase fusion model to estimate the general population: (i) Phase-1: a user population estimation for single networks to estimate \( U(S_x, t, i) \) from \( U(C^i_l, t, i) \), \( \forall l, i = \{1, 2, \ldots, N\} \), and (ii) Phase-2: a general population estimation based on user-population estimation from multiple networks to estimate \( \hat{G}(S_x, t) \) from \( U(S_x, t, i), i = \{1, 2, \ldots, N\} \).

### Table 4.1: Terminology and Notations

<table>
<thead>
<tr>
<th>Terminology</th>
<th>Notation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>( t )</td>
<td>( t_{th} ) time slot, e.g., 120 means 10AM with 5 minutes time slot</td>
</tr>
<tr>
<td>Network</td>
<td>( i )</td>
<td>( i_{th} ) network</td>
</tr>
<tr>
<td>Voronoi Partition</td>
<td>( P^i )</td>
<td>Voronoi partition based on towers in network ( i )</td>
</tr>
<tr>
<td>Cell</td>
<td>( C^i_l )</td>
<td>( i_{th} ) Voronoi cell with ( l_{th} ) tower in network ( i )</td>
</tr>
<tr>
<td>Subcell</td>
<td>( S_x )</td>
<td>( x_{th} ) subcell, subcell is the intersection ( x_{th} ) subcell</td>
</tr>
<tr>
<td>User Population</td>
<td>( U )</td>
<td>number of users (user population)</td>
</tr>
<tr>
<td>General Population</td>
<td>( G, G, G )</td>
<td>population, estimated population from cellphone user, Worldpop population (ground truth)</td>
</tr>
<tr>
<td>User Population in Subcell/Cell</td>
<td>( U(S_x/C^i_l, t, i) )</td>
<td>number of users (user population) in subcell ( S_x ) or cell ( C^i_l )</td>
</tr>
<tr>
<td>General Population in Subcell/Cell</td>
<td>( G(S_x/C^i_l, t, i) )</td>
<td>human population (general population) in subcell ( S_x ) or cell ( C^i_l )</td>
</tr>
</tbody>
</table>

**Phase-1: User-Population Estimation.** There are many models [113] [18] [55] working on cell-level estimation for single networks. Existing models obtain population for each cell individually and do not consider spatial correlation. Therefore, they cannot be applied to our model directly. Our key objective is to align estimated cell-level population to subcell-level population, i.e., a mapping from a population distribution in a tower-based partition to population distribution in an intersection-based partition.

A straightforward method to directly assign an estimated cell-level user population to subcell-level can be based on the overlapping areas as follows.

\[
U(S_x, t, i) = \sum_{l=1}^{\left|P^i\right|} \frac{|C^i_l \cap S_x|}{|C^i_l|} \times U(C^i_l, t, i), \tag{4.2}
\]

where \( U(S_x, t, i) \) is the user population in a subcell \( S_x \) during time \( t \) based on data from a network \( i \); \( U(C^i_l, t, i) \) is the user population in a cell \( C^i_l \), which is obtained by
the existing work \[113\]; \(|C_i \cap S_x|\) is the area of \(C_i \cap S_x\); \(|C_i|\) is the area of \(C_i\); \(|P^i|\) is the total number of cells in the tower-based partition \(P^i\) of a network \(i\). The rationale behind this straightforward population alignment is based on the assumption that users are uniformly distributed inside a cell. However, this assumption is not practical in the real world because the detailed infrastructures (e.g., roads and buildings) inside a cell mainly decide the population distribution. It has been shown that residents are more likely staying nearby points of interests (PoI), instead of uniformly distributed across a region \[49\].

![Fig 4.9: PoI Distribution](image)

To address this issue, in this work, we utilize the distribution of PoIs to align cell-level population to subcell-level population. For example, Fig. 4.9 gives the distribution of 586 thousand PoIs among cells based on a tower-based partition. The details of PoI distribution is given in table 4.2.

![Table 4.2: PoI Distribution in the city](image)

<table>
<thead>
<tr>
<th>Category</th>
<th>Traffic Facilities</th>
<th>Education</th>
<th>Fitness</th>
<th>Auto Services</th>
<th>Culture and Media</th>
<th>Finance</th>
</tr>
</thead>
<tbody>
<tr>
<td># of PoIs</td>
<td>19260</td>
<td>4018</td>
<td>3275</td>
<td>11254</td>
<td>2357</td>
<td>12053</td>
</tr>
<tr>
<td>Category</td>
<td>Business</td>
<td>Life Services</td>
<td>Food</td>
<td>Tourist Attractions</td>
<td>Government Organizations</td>
<td>Beauty &amp; Spas</td>
</tr>
<tr>
<td># of PoIs</td>
<td>127722</td>
<td>11254</td>
<td>68084</td>
<td>3167</td>
<td>9823</td>
<td>18663</td>
</tr>
<tr>
<td>Category</td>
<td>Shopping</td>
<td>Hotels</td>
<td>Recreation</td>
<td>Medical Services</td>
<td>Real Estates</td>
<td></td>
</tr>
<tr>
<td># of PoIs</td>
<td>153657</td>
<td>12860</td>
<td>14007</td>
<td>13060</td>
<td>57601</td>
<td></td>
</tr>
</tbody>
</table>

As shown by three zoom-in areas, i.e., A, B, and C, we found that most of PoIs are not uniformly distributed inside a cell. In the suburb cell B, their PoIs are mostly distributed along the roads, instead of uniformly distributed across the cell. Thus, since cellphone users are likely to stay nearby PoIs \[49\], they are not likely to uniformly distributed across the cell. Fig. 4.10 shows one example of the influence of PoI on
population distribution. There are 6 subcells in 3 Voronoi cells as shown in Fig. 4.10 (a). One PoI (i.e., a shopping mall) is located at the bottom right corner of the left cell (i.e., Cell 1). Cell 1 has 25 records from 25 users in 10 minutes. Cell 2 has 10 records from 10 users, and Cell 3 has 5 records from 5 users. The ground truth of the population distribution is as shown in Fig. 4.10 (c), in which the subcells closer to the shopping mall as a PoI have much more people than other subcells. However, if we apply a uniform assignment that assigns all users to subcells based on the subcell area size (i.e., does not consider the shopping mall as a PoI at all), we will have a user distribution in the subcells as shown in Fig. 4.10 (b), which lead to a bias of the user assignment.

As shown in Fig. 4.9, PoIs are not uniformly distributed in cellphone cells. It indicates a nonuniform distribution of users in a cell. To overcome this issue, we apply a customized Gaussian filter to the straightforward uniform alignment. For example, in Fig. 4.8, given the intersection-based partition, we assign the population for subcell $S_0$ based on data from Network B alone. We take the neighbor subcells of $S_0$ into considerations, i.e., the green subcells from $S_1$ to $S_6$. The weight of each subcell from $S_1$ to $S_6$ decreases as the distance from its center to the center of $S_0$ increases. Formally, it follows the Gaussian distribution as

$$W(S_x, \hat{S}_x(l)) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(\|S_x - \hat{S}_x(l)\| - \mu)^2}{2\sigma^2}}, \quad (4.3)$$

where $\hat{S}_x$ is the centroid of the center subcell $S_x$; $\hat{S}_x(l)$ is a centroid of the $l$th neighbor subcell $S_x(l)$ of $S_x$, i.e., from $S_1$ to $S_6$ in our example; $\mu$ and $\sigma$ are the mean and
standard deviation of distances from all neighbors. By applying this Gaussian filter in Equation 4.3, we have the formula 4.4.

\[
U(S_x, t, i) = \sum_{1 \leq l \leq M(S_x)} W(\dot{S}_x, \dot{S}_x(l)) \times U(S_x(l), t, i) \times \frac{|\dot{S}_x|}{|S_x(l)|}, \tag{4.4}
\]

where \(M(S_x)\) is the total number of the neighbor subcells of \(S_x\). We eliminate the influence of the subcell size by \(|\dot{S}_x|/|S_x(l)|\). With the Equation 4.4, for a particular subcell \(S_x\) and a time slot \(t\), we have \(N\) population estimations \((U(S_x, t, 1), ..., U(S_x, t, N))\) based on \(N\) networks. To keep the total number of the user distribution in cells, we apply a normalization function to make the total number of users in related subcells equal to that in the original cell. Our final user population estimation model for a network \(i\) is given by

\[
U(S_x, t, i) = U(C_i^t, t, i) \times \frac{U(S_x, t, i)}{\sum_{S_y \in C_i^t} U(S_y, t, i)} \tag{4.5}
\]

\(C_i^t\) is the cell to which the subcell \(S_x\) belongs. \(U(C_i^t, t, i)\) is the number of users in cell \(C_i^t\) at time \(t\). As follows, we introduce how to fuse \(N\) user population estimations to obtain a general population.

**Phase-2: General-Population Estimation.** In this phase, we use a particular subcell \(S_x\) and a time slot \(t\) as an example to show how to fuse \(N\) user population estimations \(U(S_x, t, i), 1 \leq i \leq N\) to obtain a general population estimation \(G(S_x, t, \forall)\). Similarly, we have general population estimations for all subcells, and thus an urban-scale real-time general population model.

Based on the existing models driven by single networks, it has been shown [113] that there is an exponential relationship between user population \(U(S_x, t, i)\) estimated by a network \(i\) and general population \(G(S_x, t, i)\) inferred by \(i\), i.e.,

\[
G(S_x, t, i) = \alpha_{S,x,t}^{i} \times (U(S_x, t, i))^{\beta_{S,x,t}^{i}}, \tag{4.6}
\]

where \(\alpha_{S,x,t}^{i}\) and \(\beta_{S,x,t}^{i}\) are the parameters we want to estimate in three dimensions, e.g., spatial \(S_x\), temporal \(t\), and network \(i\). After we have these parameters, we directly obtain \(G(S_x, t, i)\), given a user population \(U(S_x, t, i)\) obtained by data from Network \(i\).
However, in population modeling, these two parameters are extremely challenging to obtain. A standard approach to obtaining them is based on training with data obtained by different time slots (i.e., temporal cross-validation). However, we lack enough training data because the ground truths of urban population $G(S_x, t, i)$ at different time slots are almost impossible to obtain, as we mentioned in Section 5.1. Some datasets based on census (e.g., Worldpop [31]) can infer urban population in general, but they do not have detailed population at different time of day, i.e., a slot, as motivated in Figure 4.6. To address this issue, the state-of-the-art population models [113] are using spatial dynamics to obtain more training data from regions with similar functions, i.e., spatial cross-validation. Built upon this technique, we show how to utilize multiple networks as a third dimension (i.e., network cross validation) to provide more training data.

**Problem Definition:** Given a network $i$ at a subcell $S_x$ at a specific time slot $t$, let $U(S_x, t, i)$ be the user population estimated by network $i$; $G(S_x, t, i)$ be the general population inferred by $U(S_x, t, i)$. Our objective is to combine $G(S_x, t, i)$ from different networks $\{1, \cdots, i, \cdots, j, \cdots, n\}$ to estimate $G(S_x, t, \forall)$, which is the output of our data fusion model, i.e., the general population inferred by all networks together.

**Key Challenge:** Since single networks introduce bias in both spatial dynamics and temporal dynamics, The paper [113] reduced the spatial bias in the estimation based on single networks by grouping PoIs to functions of regions. However, the bias in temporal dynamics increases with time in a static model. For example, if the function of one region is changed from a residential region to a commercial region, the relation between phone call activities and populations will change correspondingly, which is modeled by regression parameters. Therefore, the bias in existing single networks and static estimation models increases as time evolves. To solve the challenge, our data fusion model is seeking a way to control bias increase on the temporal dimension when the fine-grained spatial partition reduces bias on spatial dimension. The assumption is that the general human population is identical although the user population differs in networks. Thus, we utilize the estimation results of different networks to control the bias in temporal dynamics. We reduce the bias introduced by both spatial and
temporal dynamics by co-training the user population and the general population of multiple networks in the same time slot.

**Single Networks:** For cell $C_l$, if the relation between the number of the user population and general population is $G = \alpha \times U^\beta$, which is described by the model $M = (\alpha, \beta)$. Given a user population for next $n$ time slot is $U(t+1), U(t+2), \cdots, U(t+N)$ and the general population is $G(t+1), G(t+2), \cdots, G(t+n)$. Therefore, for each time slot, we update $M$ by the real-time input $(U(t+i), G(t+i))$. $U(t+i)$ is obtained in real time by the number of active cellphone users in cell $C_l$. However, the general population in $C_l$, which is $G(t+i)$, is almost impossible to obtain in each time slot. Therefore, the sparsity of the general population $G(t+i)$ on the temporal dimension limits a single network model $M$ to dynamically evolve with time.

**Multiple Networks:** Compared with single networks, in MultiCell, we solved two problems with a two-component model. The first component builds an initial model $M$ for each subcell $S_x$. The second component provides an estimated $G(t+i)$ at time slot $t+i$ for the model updates. Therefore, our data fusion model is a dynamic model based on two components. (i) an initialization component where we initially estimate regression parameters based on an estimated general population $G(S_x,t,\forall)$ and multiple network data; (ii) a cross-network component where we only utilize real-time multiple network data (i.e., no estimated general population) to update the initially-estimated parameters in the initialization component as the time evolves.

(i) **Initialization:** We use an estimated urban population (obtained by census [31]) and multiple network data as the input to obtain initial parameters. As in Figure 4.11, for a subcell $S_x$, given two networks $i$ and $j$, we first use a general population estimation based on census data as initial estimations for both $G(t_1,i)$ and $G(t_1,j)$. We omit $S_x$ in Figure 4.11 for concise representation. To estimate the initial parameters $\alpha$ and $\beta$ in the subcell $S_x$, we follow the spatial context-aware method proposed in [113]. The context-aware model first groups subcells to different functional groups based on the PoI distribution. We categorize PoIs to seven categories, i.e., business, residence, education, entertainment, industry, scenery spot, suburb, according to previous works [71] [120]. We apply a k-mean clustering algorithm to cluster regions to
7 functional groups based on PoIs in the region. The function of one region depends on the main category of PoIs. The model estimates regression parameters based on $U(t_1, i)$ and $G(t_1, i)$ in the same function group. The context-aware model captures spacial dynamics by the PoI distribution of the city. Thus, based on Equation 4.6 along with user population $U(t, i)$ and $U(t, j)$ obtained by data from network $i$ and $j$, we can obtain two sets of parameters, i.e., $(\alpha^i_{S_x-t_1}, \beta^i_{S_x-t_1})$ and $(\alpha^j_{S_x-t_1}, \beta^j_{S_x-t_1})$ for $i$ and $j$, respectively. Since census data are static in the temporal dimension, after this initial slot $t_1$, we do not have new census data to update these parameters. The key technique we design based on co-training [119] is to utilize data from multiple networks to provide new data for updating these parameters as the time evolves, which is our key contribution to advance state-of-the-art models based on single networks.

(ii) **Cross-network Training**: With initialization, the key objective of our co-training component is to update two parameters $\alpha$ and $\beta$ for all networks in the following slots. We show our core idea in Figure 4.11. The dashed arrows indicate the process of cross-network updating parameters, and the solid arrows indicate the process of obtaining the general population by single network data.

As shown in Figure 4.11, our co-training starts from time slot $t_2$: (i) we use $G(t_1, i)$
and $U(t_1, j)$ to update two parameters $(\alpha_{S_x-t_2}^j, \beta_{S_x-t_2}^j)$; (iii) we use these two updated parameters and new incoming $U(t_2, j)$ to obtain $G(t_2, j)$; (iii) we use $G(t_1, j)$ and $U(t_1, i)$ to update two parameters $(\alpha_{S_x-t_2}^i, \beta_{S_x-t_2}^i)$; (iv) we use these two updated parameters and new incoming $U(t_2, i)$ to obtain $G(t_2, i)$.

Note that $G(t_1, i) = G(t_1, j)$ since they are equal to the initial estimation based on the census, which leads to $(\alpha_{S_x-t_2}^i, \beta_{S_x-t_2}^i) = (\alpha_{S_x-t_1}^i, \beta_{S_x-t_1}^i)$ and $(\alpha_{S_x-t_2}^j, \beta_{S_x-t_2}^j) = (\alpha_{S_x-t_1}^j, \beta_{S_x-t_1}^j)$. The reason is that the parameters $(\alpha_{S_x-t_1}^i, \beta_{S_x-t_1}^i)$ are inferred from $(U(t_1, i), G(t_1, i))$ and the parameters $(\alpha_{S_x-t_2}^j, \beta_{S_x-t_2}^j)$ are updated by points $(U(t_1, i), G(t_1, j))$ from time slot $t_1$. We can infer $(\alpha_{S_x-t_2}^j, \beta_{S_x-t_2}^j) = (\alpha_{S_x-t_1}^j, \beta_{S_x-t_1}^j)$ in a similar way. However, $G(t_2, i)$ may not be equal to $G(t_2, j)$ because based on Equation 4.6, these two sets of parameters are the same, but $U(t_2, i)$ and $U(t_2, j)$ may change compared to $U(t_1, i)$ and $U(t_1, j)$ based on real-world data from network $i$ and $j$. The difference between $G(t_2, i)$ and $G(t_2, j)$ makes our cross-network training effective.

To generalize to a multiple network scenario, as in Figure 4.12, in a slot $t$, for a network $i$ (e.g., Network 1), we first use the general population estimated by another network $\{1, \ldots, i-1, i+1, \ldots, n\}$ and the user population estimated by $i$ during the previous slot $t-1$ to cross-update parameters for Network $i$ for the current slot $t$ (i.e., dashed lines in Figure 4.12). Then we use the updated parameters along with the user population estimated by $i$ during the current slot $t$ to obtain the general population estimated by $i$ for this slot $t$ (i.e., solid lines in Figure 4.11). Finally, the average values of all estimations from all networks are the output of this cross-network training $G(S_x, t, \forall)$ for a slot $t$ and subcell $S_x$.

A standard approach to update model parameters is to use the Least Squares method [32], but it leads to a high computational cost in our dynamic population estimation model since the model changes as the time evolves. To reduce the computational cost in parameter updates, we utilize a dynamic computing method combined with a memorization technique. In particular, with the formulas in Equation 4.7, where $\mu$ is the mean, $Var$ is the variance, and $Cov$ is the covariance, the computational cost is reduced to a constant time when the data point $(x_i, y_i)$ is added to the existing regression model. The two parameters $\alpha$ and $\beta$ in our model are obtained from updated...
Var and Cov. This method requires that $\mu_x$, $\mu_y$, $\text{Var}$ and $\text{Cov}$ are memorized to be utilized at the next time slot.

$$\delta_x^i = x_i - \mu_{x}^{i-1};$$
$$\delta_y^i = y_i - \mu_{y}^{i-1};$$
$$\text{Var}(X_i) = \frac{n-1}{n^2}\delta_x^i \delta_x^i - \frac{\text{Var}(X_{i-1})}{n} + \text{Var}(X_{i-1});$$
$$\text{Cov}(X_i, Y_i) = \frac{n-1}{n^2}\delta_x^i \delta_y^i - \frac{\text{Cov}(X_{i-1}, Y_{i-1})}{n} + \text{Cov}(X_{i-1}, Y_{i-1});$$

As a result, the computational cost to update models at the time slot $t$ is $O(\gamma |S|)$ where $|S|$ is the spatial complexity and $\gamma$ is a ratio depending on the number of networks.

### 4.3.5 Implementation

To illustrate the feasibility of MultiCell, We implement MultiCell based on three major cellphone carriers in Shenzhen with a near-100% penetration rate.

(i) **Data Management:** Due to the data-driven nature of MultiCell, we introduce how to obtain and manage our cellphone data as follows. For security reasons, we are not allowed to directly access the carrier servers. Instead, we obtain these data off line. Such a large amount of data requires significant efforts for efficient management, querying, and processing. We employ a high-performance cluster with Spark for data processing. The details are given as follows: (i) 12 HP machines with 2 Tesla K80c
each; (ii) 10 Dell machines with 4 Tesla K80c each; (iii) 4 Xeon E5-2650 with a half TB memory each; (iv) A series of 800GB SSD and 15TB of spinning-disk spaces; (v) 2 PB additional disk space.

**Data Preprocessing:** Due to the large size of our cellphone data, we performed a detailed cleaning process to filter out duplicate, error, and incomplete data.

**Spatial Alignment:** Based on the method in Section 4.3.3, we implement spatial alignment with three networks in Shenzhen, which generates 3 partitions $P^1$, $P^2$ and $P^3$. We first integrate $P^1$ and $P^2$ to obtain an intersection partition and then integrate $P^3$ with it. To visualize the result, we show our intersection-based partition and subcells based on real-world cell tower data in a heat map in Figure 4.13. We found that even with three networks, we have a much finer granularity compared to three tower-based partitions in Figure 2.2. As in Fig. 4.14, with our MultiCell based on subcells, the downtown areas are covered by 36057 subcells, which leads to an average area of 11

---

**Graphs and Diagrams:**

**Figure 4.13:** MultiCell Spatial Partition based on Intersections

**Figure 4.14:** Cell Coverage

**Figure 4.15:** Top k Region Coverage

**Figure 4.16:** Bottom k Region Coverage
This subcell partition improves our spatial granularity by a factor of 10, compared to the single network data-driven modeling. Even with three networks, we achieve a granularity much closer to the desired granularity of 10 thousand m$^2$ in Worldpop. Note that in this work we use three networks as a concrete implementation of MultiCell based on multiple networks, we believe a model based on four or more networks can have subcells smaller than the desired spatial granularity. In particular, as shown by the zoom-in area, we have much more subcells in the downtown, compared to the suburban areas. For several business areas in different districts shown by the circles, we also have a much finer granularity. Quantitatively, in Figure 4.15 and Figure 4.16, we show the city area percentage covered by Top and Bottom K subcells. We found that MultiCell improves the spatial granularity of areas in the whole city. MultiCell is based on an extremely fine-grained partition, especially in the Bottom-K subcells.

(iv) Population Estimation We implement MultiCell on three dimensions $(41794, k, 3)$ where $k$ is the number of time slots of one day. Several temporal granularities from 5 minutes to 24 hours are investigated in the implementation. When the time granularity is small (e.g., 5 minutes), there are sparse regions with no user activity. This issue is alleviated dramatically by applying the Gaussian filter in the user population estimation procedure. To further address the issue, we utilize the user population from previous nonempty time slots in the same region in the implementation. For three networks,
the log-scale user populations are correlated with the log-scale ground truth linearly. It suggested a power-law distribution can model this relationship by Equation 4.6 with two parameters to learn. We map the general population in Worldpop to the general population in subcells or cells by a method given in Fig. 4.17. First, we calculate the overlapping area of two partitions and then apply function in Equation 4.1 to calculate the population in subcells. For human-unreachable areas (e.g., lakes), the Worldpop marks the grid as special values $-999$. As a result, we removed subcells with only human-unreachable areas to reduce the computational cost. For other subcells, we ignore the human-unreachable grids to calculate the general population in the subcell.

We apply the following Formula 4.8 to map estimated population in subcells to cells and regions based on the size of intersected areas where $S_x$ is the subcell and $R_l$ is a mapped region or cell, $|R_l|$ is the size of the region $R_l$, $n$ is the number of subcells intersected with $R_l$.

$$
\hat{G}(R_l, t, i) = \sum_{x=0}^{n} \frac{|R_l \cap S_x|}{|S_x|} \times \hat{G}(S_x, t, i),
$$  

(4.8)

Fig 4.18: GT & Network A

Fig 4.19: GT & Network B

Fig 4.20: GT & Network C

Fig 4.21: GT & Shared Net

We examine this relationship by comparing user populations with the ground truth in Figures 4.18, 4.19 and 4.20, 4.21. We introduce a baseline called Shared Net to naively
sum up user densities in the subcell from three networks in the same spatiotemporal dimension. Network A, B and Shared Net show strong linear relation, while Network C is partially skewed because the data we access are preprocessed by operators for privacy issues.

4.4 Evaluation

4.4.1 Evaluation Methodology

We introduce five evaluation components as follows.

(i) **Ground Truths:** In this project, we use 2010 Worldpop data for training, and we use 2015 Worldpop data for evaluation. A heat map of 2015 Worldpop data is shown in Figure 4.22 where the spatial resolution is very high, and we can identify a few urban clusters. We map the population of 100m × 100m grips of Worldpop to our subcells for the ground truth in our partition.

![Fig 4.22: Ground Truth of Shenzhen Population](image)

**Total Population:** 10.6983 M  
**Granularity:** 100m × 100m  
**Data Source:** Worldpop

(ii) **Performance Metrics:** Given the extensive usage in population models [113] [18] [55], we utilize the following correlation coefficient and normalized root mean square error (RMSE) as the metrics, respectively.

\[
\text{Cor} = \frac{\sum_{l=1}^{|S|} [G(S_l, t) - \frac{1}{|S|} \sum_{k=1}^{|S|} G(S_k, t)] \cdot [\hat{G}(S_l, t) - \frac{1}{|S|} \sum_{k=1}^{|S|} \hat{G}(S_k, t)]}{\sqrt{\sum_{l=1}^{|S|} [G(S_l, t) - \frac{1}{|S|} \sum_{k=1}^{|S|} G(S_k, t)]^2} \cdot \sqrt{\sum_{l=1}^{|S|} [\hat{G}(S_l, t) - \frac{1}{|S|} \sum_{k=1}^{|S|} \hat{G}(S_k, t)]^2}}.
\]
\[ \text{RMSE} = \sqrt{\frac{1}{|S|} \sum_{l=1}^{|S|} [ \hat{G}(S_l, t) - G(S_l, t)]^2} \]

where \( G(S_l, t) \) is the ground truth for the subcell \( S_l \) during the time slot \( t \), and \( \hat{G}(S_l, t) \) is our result. The higher metrics indicate a better accuracy of our model.

(iii) Baseline Approaches: We use five baseline approaches CAPE-A, CAPE-B, CAPE-C, CAPE-V and CAPE-S, which are based on a state-of-the-art model called Context-Aware Population Estimation [113] driven by data from five different networks, i.e., single networks A, B, C, Virtual Net and Shared Net. Context-Aware Population Estimation model clusters regions to 7 function groups based PoI distributions, i.e., residence, entertainment, business, industry, education, scenery spot, suburb. Then it builds a regression model for each group. Virtual Net considers all towers in different networks as one virtual network. We generate Voronoi partition based on towers from three cellphone networks. Therefore, Virtual Net has 11,746 towers or cell partitions. The average area of cell partition is 0.166 km\(^2\). It generates fine-grained tower-based Voronoi partition but changes the coverage range of existing towers. Shared Net calculates the user population in subcells as the total number of users in three cellphone networks, where \( U(S_x, t, \text{Shared}) = U(S_x, t, A) + U(S_x, t, B) + U(S_x, t, C) \). We apply our spatial alignment in CAPE-S since it is based on subcells. CAPE-V and CAPE-S are baselines to combine three networks together. Similar to Virtual Net and Shared Net, MultiCell utilizes data from three networks A, B and C, but the key differences are our subcell-based participation and resultant cross-network data fusion. We use the cellphone data from 8pm to 12pm to estimate the population in the city and compare the result with the ground truth.

(iv) Impacts of Factors: We evaluate three real-world factors and their impacts. (a) Subcell Population: To investigate the impact of different population on the accuracy of models, we group subcells together by four different scales based on the population, and test the performance gains of our model with increasing urban populations. (b) Temporal Granularity: We evaluate the impact of the temporal granularity by grouping all cellphone data together by a time interval of 5 mins, 10 mins, 1 hour, 6 hours, and 24 hours. (c) Spatial Granularity: We evaluate the impact of spatial
granularity by selecting three partitions, i.e., subcells, 491 regions, 11 districts. The default setting is 5 mins at subcells.

![Fig 4.23: Transportation Passenger Population](image)

(v) **Cross-Validation with Transportation Systems:** A key challenge for all urban population modeling is the lack of direct ground truths of the real-time large-scale population [18] [55]. Therefore, as a state-of-the-practice method, many existing works utilize data from urban transportation systems to indirectly evaluate their real-time population modeling results [113]. It has been showed by the previous research that there are strong correlations between real-time urban population and passenger population from transportation systems [113]. Thus, if the correlation between the results of a population model and passenger population is strong, it suggests that the performance of this model is high. In our evaluation, we consider (i) a 14 thousand taxicab network with a 460 thousand daily ridership, (ii) a 13 thousand bus network with 976 lines and a 4.3 million daily ridership, and (iii) a 127-station subway network with a 1.4 million daily ridership. These three systems captured 10.5 million rides and 6.5 million passengers per day. Different from cellphone networks capturing users locations when using phones, transportation systems can only capture passengers when they enter or exit the transportation systems. We can map these three kinds of passengers to taxi GPS locations, bus stops, and subway stations, respectively, which are visualized in Figure 4.23. For the evaluation, we map these locations into our spatial partition and test the correlations in these locations only.
4.4.2 Evaluation Results

In our evaluation, we first process cellphone records on a Spark cluster. The configuration of the cluster is described in the previous data management. Second, we build and run our MultiCell model in a local machine with an Inter(R) Xeon(R) E5-1660 v3 CPU, a NVIDIA Tesla K40c graphics card, 32.0 GB Ram and 3TB available storage. For each batch (i.e., time slot) of training, the data size is 571 KB. The training data includes subcell ID and 3 separate numbers of users for the subcell in a specific time slot. Therefore, for one-day data with 5-minute time slot, the data size is around 164 MB. For each batch (time slot), the training takes 0.123 seconds for model updates in 41,794 subcells.

(i) Model Accuracy: In this subsection, we evaluate the performance of our model by RMSE and correlation. We show the accuracy comparison with five context-aware baselines in Figures 4.24 and 4.25. From the results, we found that our MultiCell model significantly reduces RMSE by 28%, 23%, 44%, 33% and 17% and then enhances correlation by 14%, 11%, 25%, 18% and 9% on average compared with five baselines, respectively. It indicates that our model produces much more accurate estimation.

In multiple network models, CAPE-V changes cellphone tower coverage on spatial dimension. It decreases average tower coverage. Therefore, CAPE-V performs worse than single network model CAPE-A, CAPE-B. While CAPE-S reduces the single network bias by incorporating user population from multiple cellphone networks, it fails to capture time dynamics compared with our model. By comparing CAPE-A, CAPE-B, CAPE-C, CAPE-V and CAPE-S, we found that in general CAPE-S has a better
performance than single network models in both RMSE and correlation since Shared Network captures more user activities than single networks and it keeps the original tower coverage by applying our spatial alignment technique. Among single network models, CAPE-C performs worst due to the quality of data we access. For the aforementioned reasons and space limitation, we ignore CAPE-C for detailed comparisons in further evaluations.

To study the relationship between errors and populations, we plot the distribution of estimated populations and the ground truth of populations as a heat map for three models CAPE-A, CAPE-B and MultiCell in Figures 4.26, 4.27 and 4.28, respectively. The hot colors, e.g., from yellow to red, indicate more subcells; whereas the cool colors, e.g., from yellow to blue, indicate fewer. We found that (i) CAPE-A often overestimates populations compared to the ground truth if the original population is high; (ii) CAPE-B slightly underestimates populations compared to the ground truth if the original population is high. In contrast, we found that MultiCell is distributed more evenly around the ground truth with a slight trend to overestimate when the population is high.

Further, we compare the difference between the training data Worldpop 2010 and test data Wordlpop 2015 since both datasets present population distribution in the night [31]. Fig. 4.30 shows the population difference in administrative regions. We further calculate the RMSE and correlation between Worldpop 2010 and Worldpop 2015. The RMSE is 0.189599 and correlation is 0.99293. However, Worldpop 2010 is a static dataset and CDR provides the model ability to capture population change in a short time slot, e.g., 10 minutes.

(ii) Impact of Population: We quantify the impact of populations on our performance gains in Figure 4.29, by grouping all subcells into four groups based on their population and then show the performance gain of our models. The performance gain is calculated as the relative difference of RMSE between the baseline model and MultiCell. We found that MultiCell performs similarly or worse when the population is low (e.g., lower than 10) due to randomness in these lowly-populated regions, but MultiCell outperforms CAPE-A, CAPE-B, CAPE-V and CAPE-S significantly for regions with
the populations from 10 to 10,000 by 27.3%, 29.1%, 28.6% and 16.9%, respectively. Because the subcells with high populations are more important for real-world services, MultiCell is more practical than these three baselines.

(iii) Impacts of Spatial Granularity: We merge our subcells to different administrative regions to test the performances of all models by formula 4.8. In Figure 4.31, we found that the average performance of all models improve significantly as the spatial granularity of models decreases for bigger areas. In particular, at the district level, we have the best performance, which indicates the estimated population of MultiCell is almost identical to the population given by the ground truth. The reason for this phenomenon is that the randomness of human mobility is less significant if we estimate the population of large areas. It suggests our model can scale to large areas. More aggregation is better on performance but has coarser spatial granularity.

(iv) Impacts of Temporal Granularity: We merge cellphone data into five kinds
of slots, i.e., 5 mins, 10 mins, 1 hour, 6 hours, and 24 hours, respectively. Since Worldpop is static data, we use the same ground truth. We tune the training data with the user population in different time slots. By cross-validating with the ground truth, we evaluate all models and show the average performance change in Figure 4.32. We found that the performances of all models improve when the length of time slots increases. It suggests that a lower temporal granularity leads to better performances, but it is less useful in real-time applications, e.g., taxi dispatching. But as the length continues to increase, the performances of all models do not become higher significantly. Modeling population in 10-minute or 1-hour time slot is a reasonable balance between performance and temporal effectiveness.

(v) Cross-Validation with Subway: The subway passenger population is calculated at station level where both entering and exiting passengers paying with smart cards are captured. Given a time slot. We calculate the estimated population in the 496 administrative regions where subway stations locate based on the formula in Equation 4.1. The stations cover 122 out of 496 administrative regions and 286.216km² area. Then we
vectorize both estimated population and subway passengers in regions in the same time slot and calculate its correlation. We compared correlation coefficients between these subway populations and the estimated population from MultiCell with 1-hour time slot in Figure 4.33. We found that the correlation is fluctuated based on the commuting patterns. When the subway systems are operating, the correlations are high during the evening and morning rush hours, but they are low during the non-rush hours.

(vi) Cross-Validation with Taxi: The taxi passenger populations are obtained by pickup and drop-off events inferred by taxi GPS data in Shenzhen. We aggregate the pickup and drop-off location to 496 administrative regions by Equation 4.1. The correlations are given in Figure 4.34. We found that the correlation is low in the early morning and high during the daytime or early night. This is because both taxi numbers and passengers are fewer in the early morning, which leads to low taxi passenger population, while the general population obtained by our model is still high.

(vii) Cross-Validation with Bus: We calculate the real-time bus passenger population by using data from smartcards. We use the similar method to an aggregate number of passengers to 496 administrative regions and calculate the correlation coefficients in Figure 4.35. We found that from 5 AM in the morning where most bus lines start to operate, the correlation becomes higher until the morning rush hour is over in Shenzhen around 9AM. Then the correlation decreases in general during the daytime but increases again around the evening rush hour, and decreases until all bus lines stop to operate around 11 PM. Such a correlation change is based on the daily
commutes. In general, the correlation with estimated population fluctuates with the change of passenger density.

Fig 4.33: Sub. Correlation  Fig 4.34: Taxi Correlation  Fig 4.35: Bus Correlation

4.5 Related Work

Analyzing population based on multiple networks is crucial for many real-world applications, e.g., urban planning [129] and transportation [2] [46] [34]. In general, our work is directly related to population modeling and system fusion from multiple systems.

Population Modeling: Due to its various applications and recent advances in data collection techniques, population modeling has been a popular topic since 2000 [16] [17] [4] [14] [55]. These works have been focusing on simple area weighting methods or dynamic modeling to redistribute population obtained from census within finer-grained urban regions. Along with this direction, WorldPop [31] is the state-of-the-art method, which leverages the remote sensing to estimate the world population based on static data but cannot obtain the real-time population. With the increasing popularity of cellphones, many models driven by the cellphone data are proposed, e.g., cellphone data-driven models are proposed for urban populations in Shanghai [113] and populations in European countries [21]. However, they either only consider single network [113] or theoretically formulate multiple network problems with only synthetic data [21].

Multiple System Fusion: Our work is also related to data fusion based on multiple systems. Several studies have been proposed to theoretically fuse data from different systems to improve modeling performances [108] [36] [35], e.g., integrating CDR data with census data to model metropolitan-scale human mobility [49]; aligning speeds of buses, trucks and taxis on road segments as spatial granularity to estimate speeds by
a statistic model [125]; inferring road maps with OpenStreetMap and GPS trajectories [10]; combining several models to obtain a model with the minimized difference to all source models [62]. However, the above models either have dynamic ground truth for constant training or have been projected to a coarser spatial granularity, e.g., blocks, districts, cities [108].

**Our Work:** MultiCell is targeted at real-time urban sensing for fine-grained human populations based on data fusion at urban scale. However, based on the above analysis, almost all urban population models based on cellphone networks have been focusing on single cellphone networks; whereas our MultiCell system is based on real-world data from multiple networks with a novel technique for cross-network data fusion, which is our key contribution to advance the state-of-the-art population models driven by cellphone data.

### 4.6 Summary

In this work, we motivate, design, and implement an urban-scale population model called MultiCell based on data from three cellphone networks with 10 million users in the Chinese city Shenzhen. With MultiCell, we addressed a key challenge for cellphone data based population modeling, i.e., individual cellphone networks are biased for population modeling, by a network alignment technique and a cross-network data fusion technique. We evaluated MultiCell by comparing it to state-of-the-art models driven by single cellphone networks, and the results show that MultiCell outperforms them by 27% in terms of accuracy. We hope the results we demonstrated in our MultiCell model could be used for other multi-network data-driven modelings at large scale.
Chapter 5
Prediction Model for Travel Time

5.1 Introduction

According to the United Nations, we are undergoing a rapid process of urbanization where 54% of the world’s population has already been moved into urban areas in 2014, and this number is projected to rise to 70% by 2050 [124]. Thus, it is essential to improve the mobility of urban residents on a daily basis, which can be achieved by accurately providing travel time estimations for improving passenger confidence when selecting different transportation systems, e.g., subway, taxi, bus, and private vehicles. However, urban anomalies, e.g., transportation accidents [131] and social events [127], have major impact on travel time across all transportation systems in cities. In this work, our goal is to understand, quantify, and predict the impact of urban anomalies on the travel time of different transportation systems, which is essential to many real-world applications, e.g., emergency response [66], trip planning [67], and location-based services [124].

In recent years, as a result of urbanization, urban transportation systems have been equipped with advanced sensing and communication devices to generate massive amount of data [64], which provide an excellent opportunity for travel time estimation. Researchers have accumulated abundant knowledge for travel time modeling through various state-of-the-art models [95, 127, 130, 131, 122, 110, 101]. In this work, we focus on a combination of three key factors as follows to advance the state-of-the-art models [26] [30]. (i) **Fine-Grained Travel Time**: since most transportation data are mainly collected for management purposes and do not provide direct measurement of different stages of a trip, most existing studies have been focused on end-to-end travel time [95, 130], instead of fine-grained travel time. (ii) **Urban Anomalies**: the existing
studies on urban anomalies were mainly focused on the anomaly detection from traffic flows [127, 131] instead of measuring the impact of anomalies on travel time. (iii) **Multiple Transportation Systems**: Due to limited data access, almost all existing work has been focused on an individual transportation modality, e.g., taxi [98, 105] or subway [52], instead of multiple transportation systems.

To the best of our knowledge, little work, if any, has been conducted under the above three factors. This is because it is challenging to access large-scale data on urban mobility and anomalies with fine-grained spatiotemporal coverage across different transportation systems. In the vision of smart cities, many of them, e.g., New York City, Beijing and Shenzhen, have been collecting urban-scale data across different systems to improve urban efficiency. Some of these data have been made available for researchers to understand urban mobility, e.g., travel time[130]. However, since most of these data are collected for billings across different systems, they lack direct measurement of detailed components of the travel time, e.g., walking time and waiting time.

In this work, to address the above challenges, we design a measurement framework called MAC to Measure the impact of Anomalies on different travel time Components of heterogeneous transportation systems. In particular, we utilize various existing data sources (e.g., vehicle GPS, fare transactions, etc) to infer walking time, waiting time, and riding time for taxi, bus, subway and private vehicles. As a result, the key novelty of MAC is that it measures the impact of urban anomalies on fine-grained travel time components of multiple transportation systems, by utilizing data already collected for billing and management purposes. The key contributions of the work are as follows:

- We utilize various transportation infrastructures and their data for travel time measurement under urban anomalies. To our knowledge, the utilized data is fairly complete, i.e., including taxi cab, bus, subway and private vehicle data for the same city Shenzhen. The data covers more than 78% of 11 million permanent residents. We further collect a city-scale dataset of urban anomalies including expected and unexpected anomalies as ground truth for analyses.

- We design a framework called MAC to investigate the fine-grained travel time
components and the impact of anomaly events. In particular, we design a model to infer the waiting time and riding time and then validate the inferred travel time components through case studies. Furthermore, we analyze the travel time patterns on the inferred travel time components under normal and anomaly events. Finally, we design a learning model to integrate contextual information for the prediction of delay time in anomalies across different transportation systems.

- We implement and evaluate MAC by integrating data from four transportation systems including (i) a 15-thousand taxicab network, (ii) a 13-thousand bus network, (iii) an automatic fare collection system for a public transit network (i.e., subway and bus) with 5 million smartcards, and (iv) a 10-thousand private vehicle network. Our prediction model achieves 86.5% prediction accuracy on delay time prediction. Our research efforts lead to several insights and lessons learned, which are helpful for understanding city-scale fine-grained travel time under extreme anomaly condition across different transportation systems.

5.2 Motivation

5.2.1 Fined-grained Travel Time

We compare the fined-grained travel time, i.e., riding, waiting, and walking, with coarse-grained travel time, i.e., total travel time or riding time. First, we study the average waiting time, walking time and riding time that subway passengers spend between two major subway stations with high traffic, i.e., central station. Fig. 5.1 plots the fine-grained travel time distribution between these two stations in the subway system. We find that even though passengers spend a large portion of travel time on riding, the
waiting time and walking time are not negligible and have higher fluctuation. Fig. 5.2 compares the waiting time with the total travel time in all subway trips. On average, the waiting time accounts for 16% of the total travel time in the subway system. Moreover, we compare the waiting time with the total travel time in the bus, taxi and subway systems. Fig. 5.3 shows the CDF of the ratio of waiting time in the total travel time in all trips. We find that the waiting time accounts for more than 20% of the total travel time in 20% of the trips in the taxi system. In the bus system, more than 20% of the travel time is spent on waiting in 57% of all trips. In the subway system, passengers spend more than 20% of the travel time on waiting for trains in 40% of all trips. Compared with the coarse-grained travel time, the fined-grained travel time accurately describes passengers’ travel time in different stages. The details of travel time decomposition will be provided in Section 5.3.4.

5.2.2 Multiple Transportation Systems

Passengers with travel demand in a city dynamically choose one or multiple of the transportation modalities, e.g., bus, subway, taxi, and private vehicle, based on travel purposes and traffic conditions. For instance, in our analysis, we find that commuters traveling between home and work locations prefer to take public transportation systems or drive their own cars instead of riding taxis. In contrast, visitors mostly take taxis at airports or train stations. We show the travel demand (i.e., the number of passengers) and the demand trend (i.e., the normalized travel demand) in the four transportation systems in Fig. 5.4 and Fig. 5.5. We find some obvious differences among these four systems in terms of travel demands. The two public transportation systems have two peaks of travel demand during rush hours. The taxi demand increases significantly in the evening due to different travel purposes. Therefore, a single transportation system is not the representative for travel behaviors in a city due to its biased mobility patterns.

5.2.3 Anomaly Events

Anomaly events, such as social events and transportation accidents, affect the traffic flow, which leads to travel delay. For example, the severe tropical storm Nida in 2016
has caused severe travel delay. We quantify the impact of the storm on the travel time by the increased travel delay, which is calculated as the relative difference between travel time. We study its impact of travel time on both above-ground and underground transportation systems. Fig. 5.6 shows the increased travel delay during the storm. The storm significantly increases the travel time on above-ground systems, i.e., buses, taxis, and private vehicles by around 40% to 70%. It also affects the service of several subway lines and stations and leads to travel delay in the underground subway system as well.

5.2.4 Summary

In summary, we investigate the features of fine-grained travel time and anomalies. Even if passengers spend a large portion of time on riding during a trip, the waiting time is not negligible as it makes up around 20% of the total travel time. Due to mobility difference in different transportation systems, a single transportation system is not representative of all passengers’ travel behavior in the city. This motivates us to measure the impact of anomalies on fine-grained travel time, i.e., waiting time and riding time on four different transportation systems to better understand their impacts on urban mobility.

5.3 Methodology

This section describes how we decompose trips into travel components to better understand urban mobility. We first describe passengers’ behaviors with different transportation modalities. Second, we elaborate on travel time decomposition to infer fine-grained travel time components.
5.3.1 Background: Understanding Passenger Behaviors

In most cities, the major travel modalities include subways, buses, taxis and private vehicles. In order to decompose the total travel time into fine-grained time components, we first aim to understand the travel behaviors in passengers using different transportation systems.

(i) **Taxi.** In the taxi system, passengers walk from an origin to the pickup location and wait for the next available taxi. Next, passengers take a taxi to a drop-off location, after which they walk to their final destination. Most of the travel time is spent on *waiting* and *riding*.

(ii) **Bus.** If passengers choose to take a bus, similar to the taxi, passengers first walk to a bus station. Then passengers wait in the bus station for the next available bus to their destination station. Similar to taxi passengers, bus passengers spend most of the travel time on *waiting* and *riding* as well.

(iii) **Private Vehicle.** For private vehicle passengers, since there is generally no waiting time for pickup, most of the travel time is spent on *riding*.

(iv) **Subway.** For the subway system, passengers’ behavior is decomposed into *walking*, *waiting* and *riding*. Starting from an automated fare collection machine in a station, passengers walk to a train waiting platform. Next, passengers wait in the platform for the next available train. In the third phase, passengers take the train to the destination station and then walk to the tap-out machine to exit the subway system. Subway passengers spend most of the travel time in *walking*, *waiting* and *riding*.

**Summary.** By analyzing passengers’ behavior, we specifically investigate (i) *riding time* for personal vehicles, (ii) *waiting time* and *riding time* for taxis and buses, and (iii) *waiting time* and *riding time* for subway systems. The analysis does not include in-station *walking time* since the dynamics of in-station walking time is negligible.

5.3.2 Terminologies

In this subsection, we describe the terminologies and their abbreviations in the travel time decomposition.
Transportation System. We use four initial characters to represent the four investigated transportation systems. Specifically, we use $B$ to represent the city bus system, $S$ for the city subway system, $T$ for the taxi system and $P$ for personal vehicles. We categorize the four systems into above-ground group $A$ and underground group $U$, since above-ground systems share the same mobility pattern and road traffic in travel time components.

Travel Time Components. The total travel time consists of three travel time components, i.e., walking, waiting and riding, based on semantics of passengers’ behavior. A travel time component is described as $\tau^M_{status}(s, t)$ where $M$ and $status$ are two global parameters to describe the transportation modality/system and the status of passengers (e.g. walking or waiting). The two spatiotemporal parameters $s$ and $t$ are location and time respectively. For travel time components involving two locations, e.g., riding time with an origin and a destination, we use two parameters, i.e., $\tau^S_{status}(s_1, s_2, t)$, to present the locations. In addition, an individual-level travel time component is denoted as $\tau^S_{status}(s, t, i)$, where $i$ represents the id of a particular passenger.

Anomaly Event. Similar to a travel time component, an anomaly event $E_{category}^{level}(s, t)$ is associated with two global parameters and two spatiotemporal parameters where the two global parameters $level$ and $category$ are the extent of the anomaly event defined in the dataset section, and the category of the event (i.e., expected or unexpected event). The two spatiotemporal parameters $s$ and $t$ describe the location and time of the anomaly event.

5.3.3 Travel Time Decomposition: Aboveground

Given city-scale mobility data in four transportation systems, we describe how to infer the fine-grained travel time components in a specific transportation system. Specifically, we elaborate the inference of (i) the waiting time in taxi, bus, and subway systems, (ii) the riding time in the four systems, and (iii) the in-station walking time in the subway system.

(i) Waiting Time $\tau^A_{waiting}$: We use the taxi system as an example to illustrate the idea
of waiting time inference. As shown in Fig. 5.7, given taxi traces passing through location $s$, the upper bound of the waiting time is inferred by the time difference of two successive available taxis which accommodate at least one passenger. For buses, the waiting time is inferred by the time difference of two successive buses from the same route. Therefore, we divide the city into $50m \times 50m$ grids, which leads to $1800 \times 920$ grids in the city. In addition, the waiting time is affected by the number of passengers in the waiting queue. For example, if there are two passengers waiting in the same location for taxis at the same time without ride sharing, the statistical waiting time is the average waiting time of the two passengers. For the first passenger, the waiting time is obtained through the waiting time estimation of the next available taxi. For the second passenger, in addition to the waiting time of the next available taxi, the extra time is the queuing time that equals the waiting time of the first passenger. This total queuing time is determined by the travel demand which is estimated through the historical records of the taxi data.

$$\tau_{waiting}^A(s, t) = \frac{\sum_{n=1}^{D(s, t)} (r_{i+n}.time - r_i.time)}{D(s, t)};$$

$$s = r_i \text{location}; \quad r_i \text{.time in t time slot};$$

Fig 5.7: Waiting Time Inference

Fig 5.8: Riding Time Inference

For location $s$, e.g., a grid for taxis and a station for buses, we estimate the waiting time $\tau_{waiting}^A(s, t)$ at time $t$ in Equation (5.1) where $r_{i+n}$ is next $n$ available vehicles after vehicle $i$ and $D(s, t)$ is the travel demand at location $s$ and time $t$. We model the demand by the number of passengers in the historical data, which is captured by the taxi status changing from 0 (empty) to 1 (occupied). Since it is almost impossible to
infer the large-scale exact waiting time of individuals, our inference model focuses on
the upper bound of individual roadside waiting time based on taxi GPS locations.

(ii) Riding Time $\tau_{riding}^A$: The riding time inference from historical data has been
extensively investigated in the previous works [105] [132]. Instead of estimating travel
time on road segments, we infer the riding time in grids before and after an anomaly
happens. We first investigate the riding time from the traffic flows that have to pass
the grid in the normal cases. Then, to compare the travel time difference before and
after an anomaly event in the central red area, we compare the average riding time from
eight directions, i.e., from the top to the bottom, from the left to the right, from the
upper left to the bottom right, from the upper right to the bottom left and the reverse
directions, as shown in Fig. 5.8.

5.3.4 Travel Time Decomposition: Underground

Different from above-ground systems, which mainly use GPS devices to track vehicle
locations, underground systems depend on stationary devices, e.g., automated fare col-
collection systems, to track passenger flows. For each trip, the total travel time is captured
by autonomous fare collection machines, which record time and location when pas-
engers tap in or tap out the subway system, i.e., the origin and destination stations.
Therefore, the total travel time of taking the subway is the time difference between
the tap-in time and tap-out time of a passenger. With travel flows from an origin
station $s_o$ to a destination station $s_d$, the total travel time between the two locations
is $\tau_{total}^U(s_o, s_d, t, i)$. As described in passenger behavior analyses, the total travel time
is decomposed into walking, waiting and riding time components, as shown in Equa-
tion (5.2).

\[
\tau_{total}^S(s_o, s_d, t) = \tau_{walking}^S(s_o, t) + \tau_{waiting}^S(s_o, t_1) + \tau_{riding}^S(s_o, s_d, t_2) + \tau_{walking}^S(s_d, t_3);
\]

\[
t_1 = t + \tau_{walking}^S(s_o, t); \quad t_2 = t_1 + \tau_{waiting}^S(s_o, t_1); \quad t_3 = t_2 + \tau_{riding}^S(s_o, s_d, t_2);
\]

(5.2)
(i) Waiting Time $\tau_{\text{waiting}}$: To illustrate the inference of the waiting time for a passenger $i$ in a station $s_{o}$, we first define the fluent travel time. We infer the fluent travel time with the following observation in subway systems. Most of the trips include four stages: i) the walking time in the original station, ii) the waiting time in the original station, iii) the riding time from original station to the destination station, and iv) the waiting time in the destination station. Since the riding time following the subway schedule is constant and the difference of walking time among passengers is negligible, the minimal travel time occurs when “lucky” passengers catch the subway without waiting. We define the fluent travel time as the travel time without waiting which is inferred based on those “lucky” passengers. Therefore, the fluent travel time is the travel time between two stations without the waiting time, which is estimated by the minimum travel time between two stations. The waiting time is the time difference between the total travel time and the fluent travel time. The inference is described in Equation (5.3).

$$
\tau_{\text{total}}(s_{o}, s_{d}, t, i) = \min_{i} \tau_{\text{total}}(s_{o}, s_{d}, t, i);
$$

$$
\tau_{\text{waiting}}(s_{o}, s_{d}, t, i) = \tau_{\text{total}}(s_{o}, s_{d}, t, i) - \tau_{\text{fluent}}(s_{o}, s_{d}, t);
$$

(ii) Walking Time $\tau_{\text{walking}}$: We use an example to show how to extract the walking time in a station, i.e., station $B$, in Fig. 5.9 where the dash arrow represents the walking time in the station and the solid line represents the riding time between two stations.

Given three fluent trips (i.e., trips without waiting time), trip $a$ from station A to station C, trip $b$ from station A to station B, trip $c$ from station B to station C, B is the intermediate station of A and C. The fluent travel time in the three trips is
decomposed into walking and riding time, as shown in Equation (5.4).

\[
\tau_{\text{fluent}}^S(A, C, t') = \tau_{\text{walking}}^S(A, t') + \tau_{\text{riding}}^S(A, C) + \tau_{\text{walking}}^S(C, t'') \\
\tau_{\text{fluent}}^S(A, B, t') = \tau_{\text{walking}}^S(A, t') + \tau_{\text{riding}}^S(A, B) + \tau_{\text{walking}}^S(B, t) \tag{5.4} \\
\tau_{\text{fluent}}^S(B, C, t) = \tau_{\text{walking}}^S(B, t) + \tau_{\text{riding}}^S(B, C) + \tau_{\text{walking}}^S(C, t'')
\]

The riding time of trip \( a \) is the summation of the riding time of trip \( b \) and the riding time of trip \( c \).

\[
\tau_{\text{riding}}^S(A, C) = \tau_{\text{riding}}^S(A, B) + \tau_{\text{riding}}^S(B, C) \tag{5.5}
\]

In Equation (5.4), we find the second equation plus the third equation minus the first equation, we can remove the \( \tau_{\text{walking}}^S(C, t'') \). After applying Equation (5.5), the riding time will be removed. Therefore, the walking time of station B is inferred by the fluent travel time difference of the three trips, which is described in Equation (5.6).

\[
\tau_{\text{walking}}^S(B, t) = \frac{1}{2}[\tau_{\text{fluent}}^S(A, B, t') + \tau_{\text{fluent}}^S(B, C, t) - \tau_{\text{fluent}}^S(A, C, t')]; \\
\quad t = t' + \tau_{\text{fluent}}^S(A, B, t'); \tag{5.6}
\]

(iii) Riding Time \( \tau_{\text{riding}}^S \): Given the waiting time and walking time, according to Equation (5.2), the riding time inference is straightforward, which is the difference of the total travel time and other time components, as shown in Equation (5.7).

\[
\tau_{\text{riding}}^S(s_o, s_d, t_2) = \tau_{\text{total}}^S(s_o, s_d, t) - \tau_{\text{walking}}^S(s_o, t) - \tau_{\text{waiting}}^S(s_o, t_1) - \tau_{\text{walking}}^S(s_d, t_3); \tag{5.7}
\]

Summary. With the decomposition method, we infer the walking time and waiting time in subway stations, as well as the riding time between subway stations. In the implementation, we find there are multiple entrances in large stations, e.g., 12 entrances in Futian station which is adjacent to the city train station, and the walking time difference among different entrances is not negligible. Fortunately, since there are always high traffic demands in large stations, it provides enough observations in each entrance. On the other hand, different entrances can be identified by the id of fare collection machine in the dataset. Therefore, we identify the large stations by a threshold on the number of entrances, i.e., fare collection machines. Instead of modeling travel time in
the station level, we infer the walking time, riding time and waiting time on entrances in the large stations. When calculating the average waiting time in the station or riding time between stations, we aggregate inferred time from multiple entrances.

5.3.5 Measurement Method

Anomaly Measurement: We investigate the impact of an anomaly at location \( s \) and time \( t \) on a travel time component \( \tau^M_{\text{status}}(s,t) \) in multiple transportation systems in a city. With the anomaly measurement task, we are able to study the impact of an anomaly event \( E \) on the travel time components in multiple transportation systems at the city level. As a result, it enables us to compare the robustness of different transportation systems under abnormal conditions at extreme fine-grained travel patterns, which hopefully will provide insights for urban planning [33], navigation [70], travel planning [114] and anomaly detection systems [11].

Metrics: In addition to the direct comparison of travel time components, we use delay time in Equation (5.8) since travel time components differ at different spatiotemporal dimensions, and the delay time is defined as the time difference between the time that passengers spent under the impact of anomalies and the time that passengers spent without anomalies.

\[
d^M_{\text{status}}(s,t|E) = \tau^M_{\text{status}}(s,t|E) - \tau^M_{\text{status}}(s,t) \tag{5.8}
\]

System-level measurement. To understand the impact of anomalies, for every system, we study \( d^M_{\text{status}} \) under different types of anomalies with certain degrees. For an event \( E_{\text{category}}(s,t) \), we compare the anomaly events with different category and study their impact on travel time in individual systems.

Inter-system comparison. To compare the robustness of different transportation systems under the impact of anomalies, we study the impact of anomalies of the same category on different transportation systems.
5.4 Evaluation

In this section, we first briefly evaluate our travel time decomposition methods with case studies since it is the foundation of the anomaly measurement. Second, we study the fine-grained travel time patterns in the four systems. Third, we investigate the impact of anomalies on the travel time by the delay time.

Since the riding time estimation is a statistical aggregation of observations and the waiting time is inferred from the data, we validate the waiting time inference by case studies. We conduct case studies by videotaping the passenger flow in both subway stations and bus stations to validate the $\tau_{\text{waiting}}^M$ for both underground and above-ground systems as shown in Fig. 5.10 and Fig. 5.11. We collect videos recording passengers’ waiting behaviors in the subway stations and the bus stations in both peak hours (i.e., morning rush hours and evening rush hours) and non-peak hours (i.e., regular time) in one week, which covers 1181 subway passengers and 677 bus passengers. For the subway system, we count passengers’ waiting time by recording the time they arrive at the station and the time they get on the train. Due to the space limitation, we show the waiting time distribution during two peak hours (08:30-09:30 and 17:30-18:30) and one regular time (13:30-14:30) in Futian subway station in Fig. 5.12, Fig. 5.13, and Fig. 5.14, which cover the waiting behavior of 241 out of 1181 subway passengers in our dataset.

Fig. 5.15 shows the inferred waiting time from subway transaction records. We first use a noise reduction by a distance constraint between points to remove noises.
Second, we divide the points into clusters by a clustering algorithm, e.g., DBSCAN. In the third step, we apply an in-cluster regression algorithm on the inferred waiting time. We show the process in Fig. 5.16. We then compare the observations with the regression center at the same time. The average root mean square error is 0.829 minutes in the case study, which shows the inference can capture passengers’ waiting behaviors in the subway station. We use the waiting time in bus stations to validate the aboveground waiting time inference. Fig. 5.17 to Fig. 5.19 shows the waiting time of 138 passengers in Futian bus station in both peak hours and regular time. The waiting time distribution in the subway system is different from that in the bus system, i.e., the waiting time is more regular in the subway system compared with the bus system. The difference is caused by the waiting patterns and schedules of the two systems. In the subway system, subway passengers wait for the same train while bus passengers wait for different buses. Therefore, the waiting time in the subway system is linearly related with the time when the passengers start the waiting while the waiting time in the bus system is additionally determined by a specific bus arrival time. Since the above-ground waiting time inference
estimates the upper bound of the waiting time in the bus system and the taxi system, we use the precision score as the accuracy metrics. If the passengers’ waiting time is in the estimated bounds, the true positive will increase by 1. In this way, the precision score is 96.8% for the bus system in the case study.

5.4.1 Waiting Time Patterns

**Above-ground System.** Based on our waiting time inference of above-ground transportation systems, we study the waiting time of the three transportation systems, i.e., taxi, bus and subway. We compare the aboveground waiting time distribution of weekdays and weekends in Fig. 5.20. Comparing the weekday pattern with weekend pattern, we find that the main difference of waiting time is located at around 5pm, which is the peak hour in the afternoon.

To study the difference in the spatial dimension, we visualize the waiting time distribution of the peak on the spatial dimension in Fig. 5.22 and Fig. 5.23 for taxi systems. We find that the waiting time is longer in the downtown area and the transportation junctions such as the airport and the train station. Similar pattern has been found in
the bus system as shown in Fig. 5.24 and Fig. 5.25. The waiting time is higher on peak hours in the weekdays in the bus system and around 18.9% of the buses cannot follow the regular time table schedules.

![Fig 5.22: Taxi Waiting - Weekdays](image1)
![Fig 5.23: Taxi Waiting - Weekends](image2)

**Underground System.** Fig. 5.21 compares waiting times on weekdays and weekends. On weekdays, the waiting time is longer in non-peak hours during daytime and nights. However, in the two peak hours, the waiting time is lower compared with non-peak time. The reason is that subway operators send more trains in the peak hours, which decreases the time interval between two consecutive trains in a subway station. Besides, the waiting time difference is small between weekdays and weekends during peak hours. The reason is that passengers in peak hours are most commuters, who are sensitive with time. Those passengers take the train although the trains are highly loaded. However, during the non-peak hours, since there are more passengers on weekdays, unlike commuters, those passengers prefer to wait for next available trains if the current one is highly loaded. Therefore, it increases the waiting time compared with weekends.

Fig. 5.26 and Fig. 5.27 present the waiting time distribution on the spatial dimension. Although the spatial distribution is similar between weekdays and weekends, we
find higher waiting time in the downtown areas during weekdays. Fig. 5.28 shows the

weekly pattern of the waiting time distribution. Motivated by the cellphone tower clu-
stering by traffic patterns [103], we apply k-mean clustering on the normalized weekly
patterns. According to the tuning of parameters, the best \( k \) is 4. The result is shown in
Fig. 5.29, in which each line indicates different functions of regions where the stations
locate.

5.4.2 Riding Time Patterns

Since the three above-ground systems present very similar riding time patterns due to
the shared road networks, we present their average riding time distribution in Fig. 5.30.
The riding time in the aboveground systems is inferred by the average time that above-
ground vehicles spend in passing a grid \((50m \times 50m)\). In weekdays, especially during
the peak hours, the riding time increases dramatically due to poor traffic conditions.
Moreover, the peak hours shift from 8am and 6pm on weekdays to 11am and 5pm in
weekends, which are the time for lunch or dinner. For the subway system, we find that
the riding time on weekdays and weekends between the same origin and destination
stay almost unchanged. Therefore, we compare the riding time of individual passengers to study their travel behaviors. Fig. 5.31 shows the cumulative distribution where passengers take much more riding time during weekends. The reason is that a large portion of trips on weekdays are commuting between home, work and central business areas. Since passengers live around their work locations, the home-work distance is not far. In contrast, the trips during weekends are random trips, e.g., from home location to parks with families. Therefore, the distances of trips during weekends are longer compared with those on weekdays.

5.4.3 Impact of Anomalies

We measure the overall impact of anomalies on the travel time components in the four transportation systems. We show the cumulative distribution of delay waiting time under the impact of two categories of anomalies in Fig. 5.32, Fig. 5.33 and Fig. 5.34. In general, the unexpected anomalies have a larger impact on the travel time components compared with the expected anomalies. Compared with the bus and taxi system, the subway system has lower delay time in expected anomalies but higher delay time in unexpected anomalies. The reason is that unexpected anomalies cause a large impact on the subway operation (e.g., closing subway lines or changing the schedule of subway trains).

The taxi system shows higher delay time in expected events compared with the bus system. In the analysis, we find that in expected anomalies, such as concerts, the increase of travel demand increases the waiting time. In the bus waiting time inference,
we assume that buses accommodate all passengers. Therefore, the taxi delay time is caused by traffic delay and the availability of taxis while the bus delay time is caused by the traffic conditions. In terms of unexpected events, i.e., accidents, the taxi systems show a lower waiting time increase. One reason is that taxis are more agile with poor road conditions.

We show the cumulative distribution of delay riding time under the impact of two categories of anomalies in Fig. 5.35, Fig. 5.36 and Fig. 5.37. We ignore the riding time analysis for the subway system since we find that the riding time is rarely affected by anomalies inside or close to the subway stations. As shown in Fig. 5.8, to capture the travel delay caused by the detours in the anomalies, we compare the travel time of eight directions of traffic flows between the normal cases and anomalies. Since the three above-ground systems share the same road infrastructures, the riding time changes are similar in the three transportation systems. Among the three transportation systems, the taxi system is most stable and the bus system is the least stable in terms of anomalies. The possible reason is that taxis choose detours while buses have a constant route, which may cover the regions where accidents happen. Compared with personal car drivers, the taxi drivers are more experienced. This leads to a lower delay in riding time in the taxi system.
5.5 Application: Delay Time Prediction

In this section, we use the measurement results to predict delay time in anomalies. As shown in Fig. 5.38, given an observation that an anomaly $E_{\text{level}}^{\text{category}}(s, t_0)$ happens at location $s$ and time $t_0$, our target is to predict the delay time in the following $n$ time slots $\{d(s, t_0), d(s, t_1), \cdots, d(s, t_{n-1})\}$ for a travel time component $\tau$.

5.5.1 Methodology

As shown in Fig. 5.39, the prediction framework consists of three components: (i) an anomaly information feeder to feed anomaly features, (ii) a spatial information feeder to feed spatial features, and (iii) a long short-term memory (LSTM) learning model.

Fig 5.38: Delay Time Prediction

Fig 5.39: Prediction Framework
that takes a sequence of delay time after the anomaly and predicts the future delay time. We describe the details of each component as follows.

(i) **Anomaly Information.** In the measurement study, we find that the impact of anomaly events differs with categories and levels at different time. Therefore, we use three features of the anomalies in our prediction including *anomaly category, level* and *time*. Since the three features are all categorical features, which cannot be directly used as the input for the learning model, we add an embedding layer before feeding them into the learning model.

(ii) **Spatial Information.** Figs. 5.22-5.25 have shown that the spatial dimension influences the travel time components significantly. We use different features related to the locations where the anomaly happens. Besides the geographic value of the location, the spatial features include *population, PoIs* (i.e., point of interests), and *road types*. The population information is extracted from the Worldpop data set [76], which provides population distribution with high resolution, i.e., 100m × 100m. We map the population under the preset partition to our partition based on the intersection area of the two partitions [28]. We access the road networks and PoIs from an online service provider OpenStreetMap [41]. We categorize roads into four classes, i.e., main roads,
secondary roads, path and highway, and the PoIs into five groups, i.e., educational, residential, office, recreation and transport, based on the labels in OpenStreetMap. We use the distribution of PoIs and road segments in regions as spatial features. We visualize the population distribution in Fig. 5.40, the road network in Fig. 5.41, the PoI distribution in Fig. 5.42, and the anomaly distribution in Fig. 5.43. We find higher population on road segments and more PoIs in the central business district, i.e., the middle bottom area on the map, compared with other areas. The distribution of spatial contextual information, i.e., population, roads and PoIs, shows a positive correlation with the occurrence of anomalies on the spatial dimension. To reduce the impact of data scales, we apply minmax scaler on the numerical values of all regions for normalization.

(iii) Delay Time Learning. Due to the spatial-temporal nature of human mobility, the delay time distribution presents a high correlation with the spatial and temporal information. Recurrent Neural Network (RNN) is especially suitable to capture the temporal and spatial evolution of human moving and delay evolution after anomalies. Compared with a regression model, which restricts a constant relation between input and output, e.g., a polynomial relation, RNN presents higher flexibility on hidden relations. Besides, the configuration flexibility makes it suitable to integrate spatial and temporal dependency. However, previous studies [91] have shown that traditional RNNs fail to capture the long temporal dependency for the input sequence due to the vanishing gradient and exploding gradient problems. To address these drawbacks, Long Short-Term Memory (LSTM) is a special RNN architecture for sequence labeling and prediction tasks [91]. Therefore, we apply a time series learning LSTM model combined with spatial information and anomaly information to capture the delay time dynamics.

Fig. 5.44 illustrates the internal structure of LSTM cell, which consists of three gates, $g_t$ is the input node at time $t$, which takes activation in the standard way from the input layer $x_t$ and previous hidden layer $h_{t-1}$; $i_t$ is the input gate, similar to $g_t$, which takes the input $x_t$ and $h_{t-1}$ with a sigmoid activation; $s_t$ is a self-connected internal state with a fix unit weight, which is designed to solve the vanishing or exploding problem; $f_t$ is the forget gate and used to flush the internal state; $o_t$ is the output gate.

In our learning model, we construct a two-layer LSTM model including an encoding
layer and decoding layer. The encoding layer takes two inputs (taxi system, bus system and private car system, which cover passengers’ major modalities in a city). We store the data in a relational database and build index on location, time slot, waiting time, riding time, transportation system, and anomaly. We store the data in a relational database and build index on location, time slot, waiting time, riding time, transportation system, and anomaly. We store the data in a relational database and build index on location, time slot, transportation system to construct travel time component tensors for efficient queries.

5.5.2 Evaluation

Learning Data: We build the learning model upon the dataset in our measurement study. The learning dataset includes four transportation systems, i.e., subway system, taxi system, bus system and private car system, which cover passengers’ major modalities in a city. Each record in the dataset includes the following attributes: location, time slot, waiting time, riding time, transportation system, and anomaly. We store the data in a relational database and build index on location, time slot, transportation system to construct travel time component tensors for efficient queries.
**Cross Validation:** We apply a 5-fold cross-validation, in each round, we use 1-fold as testing data. We train the models with different travel time components in multiple transportation systems. Specifically, when estimating delay time $d_{\text{taxi\_waiting}}$, which is the delay waiting time in the taxi system, we train the model with historical $d_{\text{taxi\_waiting}}$.

**Metrics:** We compare the predicted result $\hat{d}$ with the testing data $\bar{d}$ in terms of the delay time by Mean Absolute Percent Error (MAPE) defined in Equation 5.9.

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \frac{|\hat{d}_i - \bar{d}_i|}{\bar{d}_i}$$  (5.9)

**Baseline Approaches:** We compare our model with three baseline methods.

- **ARIMA:** ARIMA is autoregressive integrate moving average model, which is a time series model to predict future points in the series. ARIMA model takes the delay time from previous $n$ time slots as input and the delay time in the following time slot as the prediction target. To improve the prediction accuracy and reduce the noises introduced by anomaly types, we separate the data into groups by anomaly levels and categories and train a regression model on each data group. The external information, such as population and road networks, is not available in this prediction model [127].

- **MLE:** We implement the state-of-the-art model of travel time estimation [117], in which we fit a multinormal distribution in the anomaly area and surrounding regions. We set the delay time with different traveling paths as observations and maximize the likelihood of the estimated delay time based on the observations. We apply the Expectation-Maximization (EM) algorithm for the model convergence. The $MLE$ model is implemented in different anomaly categories and levels.

- **MAC-:** We train three LSTM models: (i) a LSTM model without spatial information and anomaly information, (ii) a LSTM model with spatial information but without anomaly information, and (iii) a LSTM model with anomaly information but without spatial information. We integrate the three models and adopt the best performance as the output of $MAC$-.
**Implementation:** We implement our prediction model on travel time components inferred from the four transportation systems. Our model and baseline models are implemented with Keras and Tensorflow libraries. We train and evaluate our design on a server with 8 Nvidia K40C GPUs. We set the learning rate as 0.01. For each LSTM layer, we set the number of cells as 60 and initialize the LSTM parameter with random values between -0.01 to 0.01.

**Evaluation Results:** We predict the waiting delay and riding delay after an anomaly in the four transportation systems. Fig. 5.45 presents the performance of riding delay prediction after the anomaly for one-hour period. The MLE is a fitting model with historical delay time. However, different from travel time estimation, which shows regularity in the historical data, even though we categorize anomalies into different categories and levels and implement the model in a specific category and level, the delay time shows a high variance in anomalies. As a result, the delay time estimation depends on the time series features and time series models achieve better performance compared with the state-of-the-art travel time estimation model. The three time series models show similar performance trends in terms of time, where all models achieve the worse performance in the first 5 minutes. This is caused by two factors: (i) the uncertainty of the location, the level of the anomaly and the traffic condition, and (ii) the sparse after-anomaly observations as prediction features since ARIMA and MAC take the previous delay time as input features for the prediction. After a half hour of an anomaly happened, the performances of the learning-based time series model become stable. Compared with the two baselines, our model achieves better performance in terms of the prediction accuracy and stability. The performance of the learning model on waiting delay prediction is given in Fig. 5.46, which shows similar results as the riding time estimation. However, the waiting time estimation error is larger compared with riding time.

**Impact of Factors:** We further investigate the impact of two factors on the prediction performance, i.e., the transportation systems and anomaly categories. Fig. 5.47 shows the performance distribution on the testing data. The box plot shows the minimum, first quarter, medium, third quarter and maximum of the distribution, which are used
to present the average and the variance of the distribution. Even the subway system is the most stable in terms of travel time compared with above-ground systems in normal cases, among the four transportation systems, our algorithm has the lowest prediction error in the taxi system and the highest prediction error in the subway system. This is due to the anomalies in the subway system cause large-scale changes of traffic flow, especially the unexpected anomalies, which is difficult to be captured by limited observations. Comparing the three above-ground systems, we find that our algorithm achieves the best performance in the taxi system and it has the worst performance in the bus system. The reason is that the delay time of taxis is aggregated from vehicles, where much more taxis serve as travel time sensors, providing accurate observations as training features. Instead, since only riding time is predicted in private vehicles, the prediction error is the smallest in private vehicles among all above-ground systems.
5.6 Discussion

**Lessons learned:** Based on the measurement results, we summarize a few lessons learned as follows. (i) Dividing travel time into fine-grained components helps us understand the impact of various factors on different parts of travel time, specifically urban anomalies. (ii) Compared with above-ground transportation systems, the anomalies have a larger impact on underground systems, i.e., subway systems, although underground systems are not affected by above-ground traffic condition. (iii) Unexpected anomalies have a larger impact on riding time and waiting time in all transportation systems compared with expected anomalies. (vi) While it is challenging to model and predict human mobility in terms of fine-grained travel time, spatial related contexts show strong correlations with human mobility and can improve the predictability significantly.

**Generalization:** We design, evaluate and implement the system in the Chinese city Shenzhen. Without access to the data in other cities, we cannot verify the effectiveness of our system in other cities. In particular, cities outside China have different policies for data releasing, which creates barriers to generalize our system in those cities with 4 transportation systems. However, since existing works have shown the accessibility to city-scale dataset in a single system such as London subway system [57], Beijing [105] and NYC [128] taxi system, Singapore bus system [132], we believe the analysis method and prediction model can be generalized to other cities with similar spatio-temporal features. Moreover, our analysis about the impact of anomalies on different transportation modalities can be referred by city administrators and urban planners to better manage the city transportation. For the benefits of peer researchers, we negotiate with the data provider and agree to release our sample dataset including one day of taxi GPS data, one day of subway smart card transaction records and one day of bus GPS records. Since private vehicle dataset contains personal information and is not a major design in our analysis, we will not release the private vehicle dataset.

**Privacy Protections:** While modeling the travel delay is important for individuals, we have to protect the privacy of participants involved. In this project, all data analyzed
is anonymized by the collaborators, so data cannot be used to trace back to individual users explicitly. We only store and process data that is useful for the travel time modeling project, and exclude other information for the minimal exposure. All data are collected legally under the consent of the users.

**Potential Societal Impacts:** In the case study, we have shown that our measurement result could be utilized to predict the delay time when anomalies occur in a city. Our study can be applied to more potential applications with better societal impacts. (i) One application could be heterogeneous travel modes. Generally, people only take one mode of transportation towards their destinations if they do not have to transfer. Through understanding the travel delay in each area, people may be able to choose multiple transportation modes to their destinations depending on what kind of mode combinations achieve the minimum travel time. (ii) Another application could be the arrangement of public transportation. The travel delay of each transportation mode implies the balance of demand and supply in the area. Administrators can rearrange the existing transportation supply (e.g., increasing more buses in some lines) or guide the transportation setting (e.g., guiding the bus or subway line construction in the future).

### 5.7 Related Work

Travel time analysis has been investigated by considerable studies because of its importance to people’s daily life. We summarize existing works in Table 5.1 with a two-dimension taxonomy: (i) transportation modality, i.e., single modality or multiple modalities; (ii) granularity, i.e., with or without travel time decomposition.

<table>
<thead>
<tr>
<th>Categories</th>
<th>Grularity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coarse-Grained</td>
</tr>
<tr>
<td>Transportation Modality</td>
<td>Single</td>
</tr>
<tr>
<td></td>
<td>Coarse-Grained</td>
</tr>
<tr>
<td></td>
<td>[37] [57] [58] [61]</td>
</tr>
<tr>
<td></td>
<td>[97] [105] [109] [133]</td>
</tr>
<tr>
<td></td>
<td>Multiple</td>
</tr>
</tbody>
</table>

**Studies on Single Modality:** Most existing studies on travel time estimation are based on single modality due to the isolation of transportation systems, such as subways,
buses, and taxis.

(i) **Coarse-grained travel time**: Given the task of estimating travel time, existing studies focus on the total travel time between the two locations in a city by bus [97], subway [57, 58], taxi [37, 61, 105, 133], or private vehicles [109]. The travel time estimation models are carefully designed for a specific transportation system or modality. The generalization is not investigated due to system isolation as well as data accessibility.

(ii) **Fine-grained travel time**: The actual travel time is not only the travel time in the transportation system such as between getting into the bus and getting off the bus. Instead, it is decomposed into different stages according to travel patterns. For example, the travel time is decomposed into waiting time in a bus station and riding time in a bus [53, 132] in bus systems. Similarly, the travel time is mostly divided into waiting time in a pickup location and the riding time in a taxi [40, 60]. For the subway system, the travel time is mainly divided into waiting time, riding time and transfer time [59].

**Studies on Multiple Modalities**: Studies involved with multiple modalities either integrate multiple mobile systems or investigate the differences and similarities between different mobile systems.

(i) **Coarse-grained travel time**: Due to the limitation of multi-modal data access, little work has been focused on coarse-grained travel time for multiple modalities, i.e., riding time from one location to another location based on different systems. Some of the representative studies in this category include a multi-view learning model to explore human mobility using transportation and cellphone data [123], and an integration model to infer real-time traffic speeds with multi-source large-scale infrastructure data [126].

(ii) **Fine-grained travel time**: To the best of our knowledge, we are the first to investigate the fine-grained travel time pattern with multiple modalities. Building upon four transportation systems, our study aims to provide a comprehensive analysis on the fine-grained travel time pattern.
5.8 Summary

In this work, we study the fine-grained travel time distribution in four transportation systems, which covers 10 million passengers in the Chinese city Shenzhen. In particular, we investigate the impact of anomalies on the travel time components in terms of expected anomalies and unexpected anomalies. Finally, we design a context-aware learning model to predict the fine-grained travel time under the impact of anomalies. Based on the above efforts, we provide a few valuable insights for fellow researchers to understand urban-scale human mobility with fine-grained travel behaviors. More importantly, our results have the potential to help the city government to manage urban traffic given expected anomalies and unexpected anomalies, which significantly improves urban efficiency and resilience.
Chapter 6
Future Work and Conclusion

In this chapter, we discuss several future directions based on our current work then conclude our dissertation.

6.1 Future Directions

In future work, we will continue measuring and predicting human mobility with urban sensing systems. In general, we will focus on problems in urban-scale human mobility. On one hand, we will conduct measurement studies to understand the strength and weakness of different urban sensing systems, and propose techniques to solve the weakness of single systems. On the other hand, we will model human mobility based on the data from heterogeneous urban sensing systems at both the collective level and individual level, e.g., modeling and predicting travel time, distance, locations, hot spots, etc. Besides, the measurement and prediction studies provide feedback to urban sensing systems, and help improve the systems. We elaborate on our future work with concrete directions.

6.1.1 Sensing: Infrastructure Improvement

urbanSense relies on large-scale and reliable sensing data. With analysis and modeling on those sensing data, we have a better understanding on current infrastructures and urban sensing systems. On one hand, the conflicts between human travel demand and infrastructure supply can potentially provide guidance for city planners to improve those infrastructures. With real-time human population distribution, we can identify hot spots in the city and improve infrastructure accordingly, e.g., cellular tower deployment and upgrades for 5G. With a detailed travel time distribution in the city,
we can automatically detect road segments with congestion and further improve road network deployment. On the other hand, the findings in the dissertation can be applied to improve the urban sensing systems. For instance, we can deploy vehicle-to-vehicle communication module on on-board sensing devices of transportation systems based on their mobility patterns to improve spatio-temporal coverage of current sensing systems.

6.1.2 Measurement: Human Mobility Evolving

With long-term data collected in urban sensing systems, we can measure and understand evolving patterns in human mobility and travel behaviors. The human mobility measurement can be conducted in terms of certain metrics such as travel time, distance or locations. We will study the evolving patterns on heterogeneous sensing systems from two perspectives, i.e., comparison of different sensing systems, and their interactions. Heterogeneous sensing systems capture human mobility and travel behaviors from different perspectives. For example, people take public transportation to commute between home and work locations on weekdays while drive personal vehicles on weekends. We can compare human mobility patterns in public transportation systems and personal vehicles based on heterogeneous sensing data. Moreover, a sensing system is impacted by another system, and their human mobility can be complimentary. Therefore, we can measure their long-term interactions among different sensing systems. For example, the upgrade of a subway system, e.g., new subway lines and stations, will not only impact human mobility patterns in the subway system but also impact human mobility patterns in other systems such as taxis and buses, e.g., increasing or decreasing travel demand in those systems. A better understanding of such interactions will benefit applications such as urban planning and mobile computing.

6.1.3 Prediction: Individual Human Mobility

In the dissertation, we only model collective travel behaviors, i.e., population and travel time. Nowadays, there is an increasing demand for personalized services for human mobility and travel demand [106]. We will expand our work from collective human mobility to individual human mobility, which is important for many real-world applications such
as personalized navigation and usage-based insurance for individual drivers. The historical records in sensing systems include an anonymous ID, which can be used to model individual human mobility. We mainly focus on two categories of individual mobility modeling, i.e., human mobility inference and human mobility prediction. Human mobility inference is to infer the missing observations caused by sensing gaps in sensing systems. For example, subway fare collection systems record the station name and time when a user tap in or tap out subway systems. The travel details between two origin and destination remain unknown due to the limitation of the sensing systems. We can infer travel routes, locations, time of individual passengers based on their travel patterns. Human mobility prediction is to predict future travel behaviors based on individual historical travel records. For example, usage-based insurance companies are interested to predict future risk of drivers, which can be quantified by certain metrics such as travel time, travel distance, and speed variance [106] [74].

6.1.4 Novel Applications and Services

Apart from sensing, understanding and modeling human mobility, we can design novel applications based on heterogeneous sensing data. We use one novel application, i.e., automatic road map inference, as an example. In general, our target is to automatically construct a digital road map from GPS traces. However, it is challenging to infer road map from GPS traces for two reasons. First, a practical map requires not only road structures but also road properties, e.g., road categories, to provide practical services (e.g., navigation), which is challenging to infer based on just locations of vehicles. Second, road structures are complicated in cities such as interchanges and parallel roads. Due to the GPS noises, it is challenging to separate those interchanges and parallel roads from GPS observations. We can utilize the diversity of heterogeneous sensing systems to address those challenges. We found the distribution of heterogeneous vehicular fleets differs on different types of roads, e.g., more buses on residential roads but less on highways. Based on the observations, we can design models to improve the accuracy of road map inference, and infer the road types.
6.2 Conclusion

In this dissertation, we design a urban sensing and modeling framework named *urbanSense* to understand and model human mobility at urban scale. We implemented and evaluated *urbanSense* on heterogeneous data passively collected with 8 urban-scale sensing systems crossing 3 domains, i.e., transportation, communications, and payment, covering 10 million residents in Chinese city Shenzhen. Based on the heterogeneous sensing data, we first studied *representativeness* of urban sensing systems and identified the bias of single sensing system for urban sensing and travel behavior modeling. To address the drawback of single sensing systems, we combine multiple systems to model urban human mobility. Specifically, we model two of the most important metrics on urban travel behaviors, i.e., real-time human population and travel time on fine-grained spatio-temporal granularity.

Conceptually, the main contribution of the dissertation is exploring the bias in single urban sensing systems and strength of heterogeneous sensing systems in human mobility modeling at urban scale. Technically, we designed a metric named *representativeness* to quantify the bias of different urban sensing systems, i.e., heterogeneous cellular networks. To utilize the strength of heterogeneous systems, we build learning-based models for human mobility modeling. For human population modeling, we integrate heterogeneous data with two key techniques to infer real-time human population, i.e., a *spatial alignment* technique to convert diverse spatial partitions in heterogeneous sensing system to a uniform partition, a *coTraining* technique to fuse multiple sensing systems simultaneously on the uniform spatial partition. For travel time prediction, we predict fine-grained travel time of heterogeneous sensing systems with a time-series learning model, which combines contextual information with historical travel behaviors. We evaluate *urbanSense* with large-scale real-world sensing data. The evaluation results show that *urbanSense* achieves a better performance compared with state-of-the-art models. Finally, our *urbanSense* framework has a few practical implications, which has the potential to provide guidance for future large-scale sensing system designs and resultant data-driven predictions.
References


[35] Gao, J., Fan, W., Jiang, J., and Han, J. Knowledge transfer via multiple model local structure mapping. In ACM KDD’08.


[40] Guan-De, Q., Yao, P., Shi-Jian, L., and Gang, P. Predicting passengers’ waiting time by mining taxi traces.


[62] Li, Q., Li, Y., Gao, J., Zhao, B., Fan, W., and Han, J. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *ACM SIGMOD ’14*.


