MultiCell: Urban Population Modeling Based on Multiple Cellphone Networks

ZHIHAN FANG, Rutgers University, USA
FAN ZHANG, SIAT, Chinese Academy of Sciences & Shenzhen Beidou Intelligent Technology Co., Ltd., China, China
LING YIN, SIAT, Chinese Academy of Sciences, China
DESHENG ZHANG, Rutgers University, USA

Exploring cellphone network data has been proved to be a very effective way to understand urban populations because of the high penetration rate of cellphones. However, the state-of-the-art population models driven by cellphone data are typically built upon single cellphone networks, assuming the users in a particular cellphone network used are representative of all residents in the studied city with multiple cellphone networks. This assumption usually does not hold in the real world due to strategic spatial coverages and business concentrations of cellphone companies, which lead to data biases, and thus overfitting of resultant population models. To address this issue, 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; (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. Finally, we cross-validate MultiCell with three transportation systems with more than 8 million passengers to investigate its performances.

CCS Concepts: • **Networks** → **Location based services**; • **Information systems** → *Sensor networks*; Mobile information processing systems;

Additional Key Words and Phrases: cellphone networks, data fusion, data analysis

ACM Reference Format:
Zhihan Fang, Fan Zhang, Ling Yin, and Desheng Zhang. 2018. MultiCell: Urban Population Modeling Based on Multiple Cellphone Networks. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 106 (September 2018), 25 pages. https://doi.org/10.1145/3264916

1 INTRODUCTION

Real-time urban population modeling is essential to many applications, e.g., mobile computing [6], urban planning [20], and location-based services [37]. Traditionally, urban populations have been modeled by surveys, e.g., census data [19], which are comprehensive but typically out-of-date and cannot be used for real-time

Authors' addresses: Zhihan Fang, Rutgers University, Piscataway, NJ, 08854, USA, zhihan.fang@cs.rutgers.edu; Fan Zhang, SIAT, Chinese Academy of Sciences & Shenzhen Beidou Intelligent Technology Co., Ltd., China, Shenzhen, Guangdong, China; Ling Yin, SIAT, Chinese Academy of Sciences, Shenzhen, Guangdong, China; Desheng Zhang, Rutgers University, Piscataway, NJ, USA, desheng.zhang@cs.rutgers.edu.

2474-9567/2018/9-ART106 $15.00
https://doi.org/10.1145/3264916

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 [23], taxi [16], buses [5], subway [25], and smart cards [29]. 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 [16], 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 [12] [23]. In particular, (i) it has been shown that 96% of world population have cellphones and use them regularly [1], which helps us model real-time urban populations [12] 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 [22], 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 [4].

To date, many population models have been proposed based on data from cellphone networks [31] [11] [24]. However, we found almost all these models are implemented by data from single cellphone networks while most cities around the world have multiple networks [23]. 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 [24]. 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 8, 9, 10). 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 paper, 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 [31], 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 [13] 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 2 shows our motivation. Section 3 describes datasets. Section 4 elaborates the model instantiation on multiple cellphone networks. Sections 5 and 6 are the implementation and evaluation, followed by related work in Section 7. Section 8 concludes the paper.

2 MOTIVATIONS

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 [1]. 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 3) 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. 1.

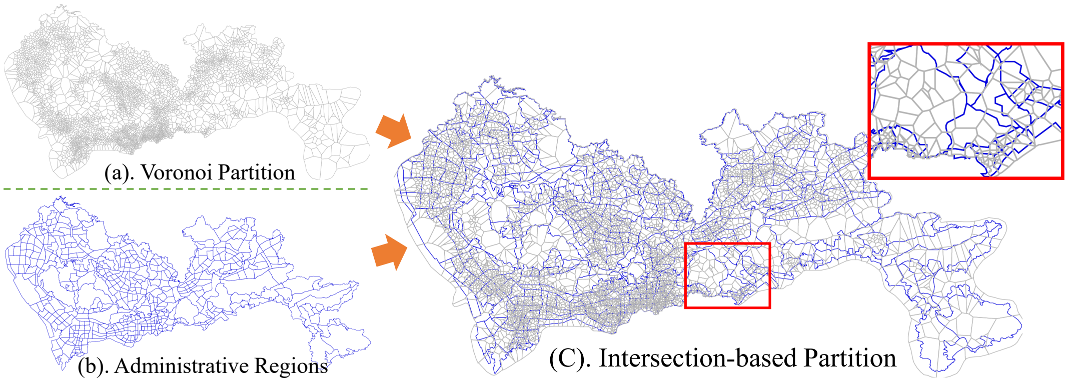

(a). Voronoi Partition
(b). Administrative Regions
(C). Intersection-based Partition

Fig. 1. Map Users to Administrative Regions

Second, we map the number of users in Voronoi partition to administrative regions proportionally based on intersected areas by the function in equation 1.

$$\mathbf{U}(R_x, t, i) = \sum_{l=0}^{n} \frac{|R_x \cap C_l^i|}{|C_l^i|} \times \mathbf{U}(C_l^i, t, i), \tag{1}$$

where $\mathbf{U}(R_x, t, i)$ is the user population in a region R_x based on data from a network i in a time slot t; $\mathbf{U}(C_l^i, t, i)$ is the user population in cell C_l^i; $|R_x \cap C_l^i|$ is the area of $R_x \cap C_l^i$; n is the number of cells intersected with R_x. The average area of the 496 regions is 3.973782 km^2 and the standard deviation is 6.075980 km^2.

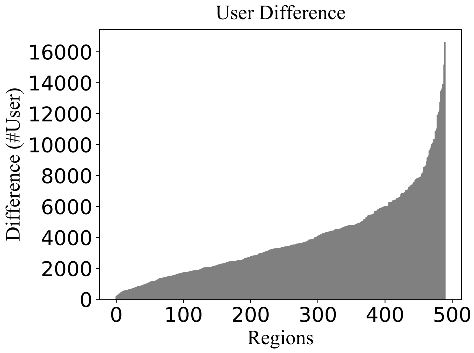


Fig. 2. User Difference

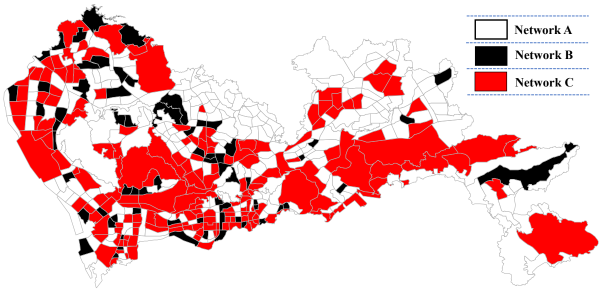


Fig. 3. Users Difference at Urban Region Level

Fig. 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. 3. Fig. 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, 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 3. In this paper, 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.

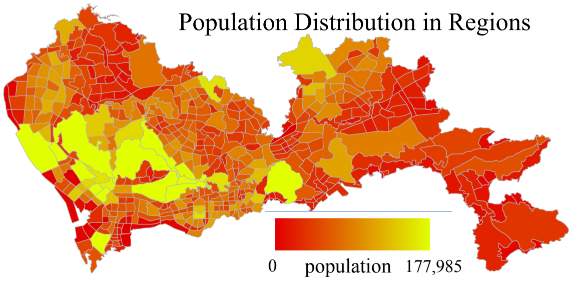


Fig. 4. Population Distribution in Region Level

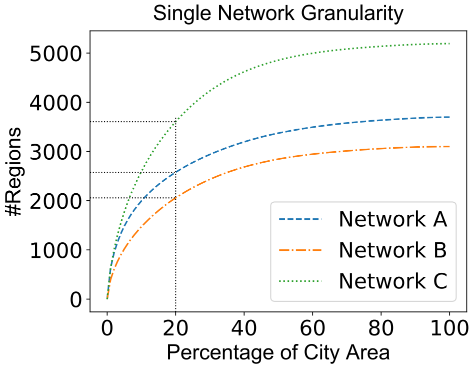


Fig. 5. Cell Coverage

Spatial Granularity: The coarse spatial granularity is the key disadvantage for models driven by cellphone data [23] because we can only infer user locations on the tower level. For example, in Shenzhen with a total area of 1,991 km^2, the three networks we studied have 3,585, 2,977 and 5,174 towers. Each of towers leads to an irregular 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 × 100m = 10 thousand m^2 for dense areas [14]. For example, in Figure 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 [14] 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 paper, 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).

3 DATASETS

3.1 Cellphone Networks

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 in the rest of the paper.

- 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.

We perform a data-driven modeling based on 3 months of data. Based on these networks, we perform various statistical analyses to understand their spatiotemporal features.

Temporal Distribution: We show the number of data records and the number of active users in the three networks of CDR data at the different hour of one day in Figure 6 and Figure 7. We found that these three networks have similar overall patterns. We found a key difference, i.e., Network A has more records from 20:00 to 5:00 after a short decrease from 19:00, whereas other networks do not have such a phenomenon. After we confirm with the operator, this may be because they have a discount plan from 21:00 to 7:00, which leads to a temporal bias for this time period. When comparing the record distribution and active user distribution, we found though the general trends of the three networks are similar, the relative relations are not the same. Network A has the most active users at 10:00 while the least active records during the same time period. This indicates different user behaviors in the three networks, e.g., the frequency of calls per user in Network A is lower than that in Network B and Network C at 10:00. Such a temporal bias leads to imbalanced data when we model populations in a fine-grained temporal partition, e.g., 5 mins.

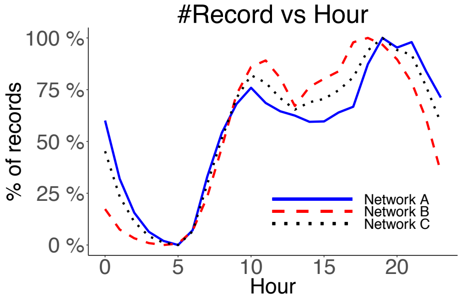


Fig. 6. Record Temporal Distribution

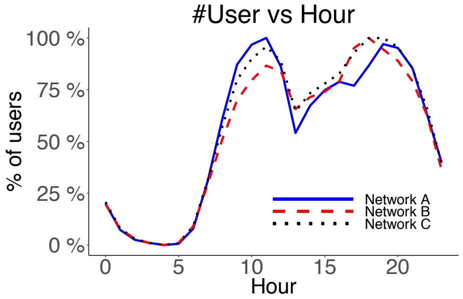


Fig. 7. User Temporal Distribution

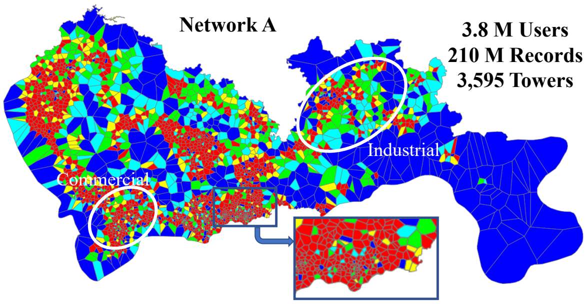


Fig. 8. Spatial-Network A

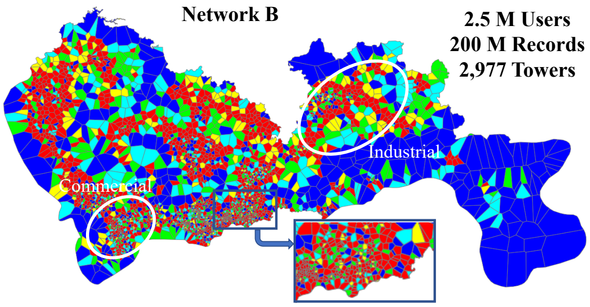


Fig. 9. Spatial-Network B

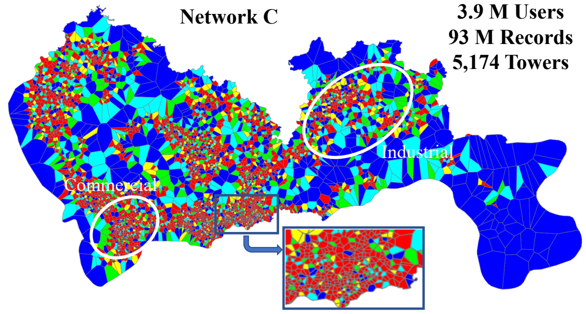


Fig. 10. Spatial-Network C

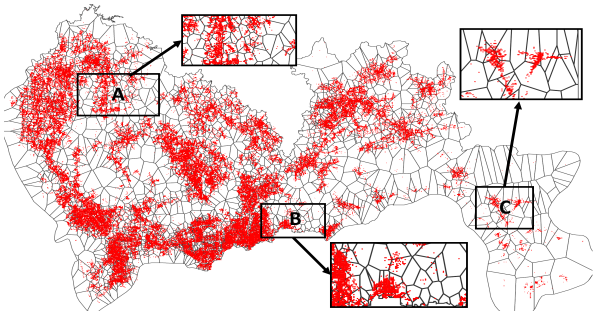


Fig. 11. PoI Distribution

Spatial Distribution: Based on the tower locations of the three networks, we obtain their heatmap-based Voronoi diagrams, which is a partitioning of a plane (e.g., a city) into regions (e.g., cells) based on distances between points (i.e., cell towers) in this plane [28]. In Figure 8, 9 and Figure 10, we found that in general, they have similar patterns. They have more towers in the downtown compared to the suburbs. But we found a few key differences between commercial areas and industrial areas, which reflects business concentrations of different networks. Previous work has shown the commercial and industrial area partition in Shenzhen based on the land use of urban planning [35]. Further, as we mentioned in the motivation section, the current spatial granularity is too coarse for ideal modeling [14]. The differences in the spatial distribution and coarse spatial granularity provide new challenges to fuse multiple network data to model populations.

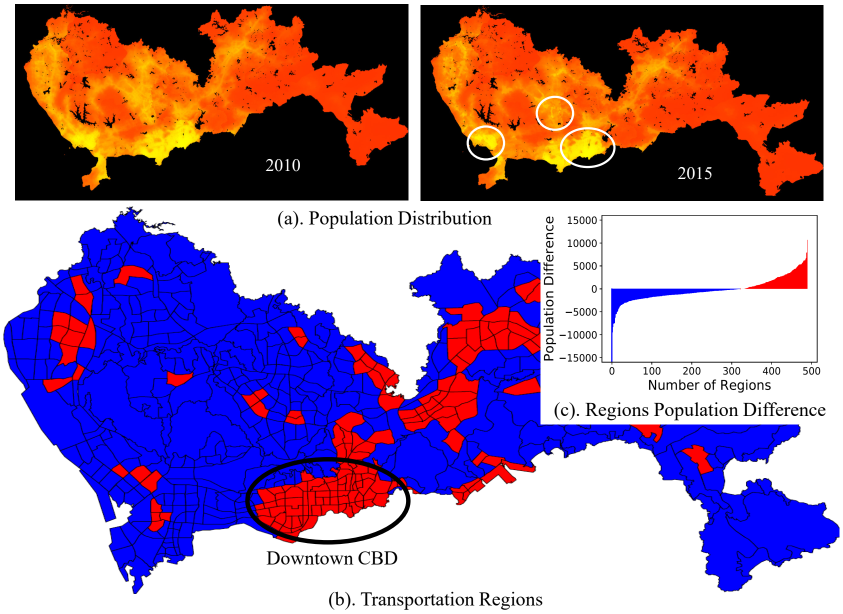
In short, the data from these networks provide valuable insights into our real-time urban population modeling. More importantly, due to their different business models and concentrations, they are spatiotemporally complementary to each other. By taking advantage of data from these networks, MultiCell is significantly different from the state-of-the-art population models based on cellphone data from single networks [31] [11] [24].

3.2 Point of Interests

We utilize point of interests (PoI) in Shenzhen for our modeling. The PoI dataset includes 17 categories and 568,566 locations. The details of PoI categories are given in table 1. Fig 11 visualizes PoIs on the spatial dimension in Voronoi partition of Network A.

Table 1. PoI Distribution in the city

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



(a). Population Distribution

(c). Regions Population Difference

(b). Transportation Regions

Fig. 12. Population Dynamics

3.3 Worldpop as Ground Truth

We utilize Worldpop datasets, which obtain populations by exploring multiple data sources including remote sensing, census, and cellphone data [14]. It is the most accurate population data so far with a 100m × 100m spatial resolution. However, as we mentioned, Worldpop is a static dataset due to the cost of such data collections. Fig 12(a) shows the population distribution in 2010 and 2015. The circled areas are CBD areas. We investigated population distribution in the same administrative region partition. Fig 12(b) presents the population difference between these two years. The red color means a region with a higher population in 2015 while the blue color means a region with a higher population in 2010. Fig 12(c) gives the precise population differences in this regions. We found the population increases in the downtown CBD areas due to the urbanization process in Shenzhen.

4 MODEL: MULTICELL DESIGN

4.1 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 13.

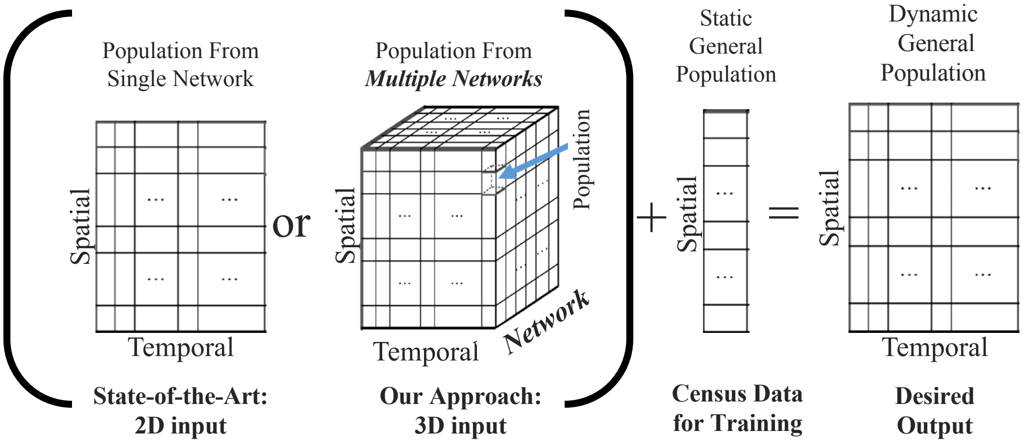


Fig. 13. Core Idea

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., [31]) 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 13) 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.

We show the framework of our MultiCell model according to the data flow in Fig. 14. *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

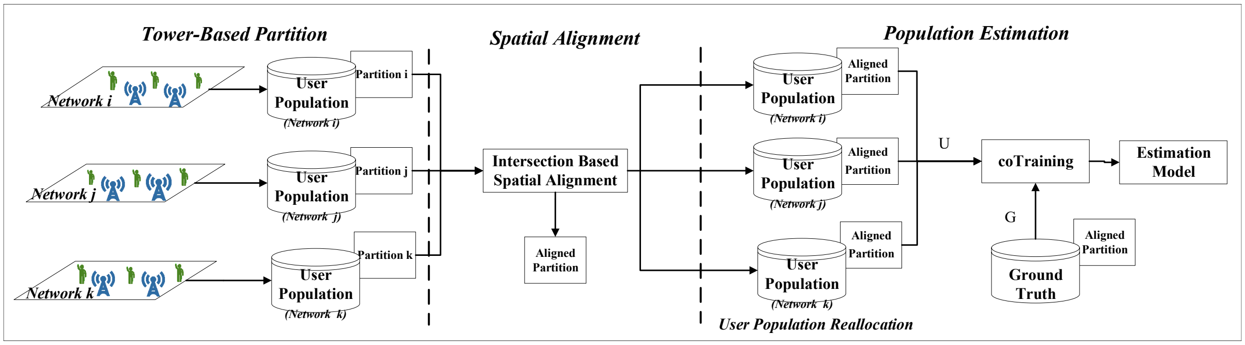


Fig. 14. Model Framework

filter to map cellphone data to the new partition and then design a co-training technique to fuse population estimations obtained by individual networks.

4.2 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 [28]. 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 [34]) 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 [31]. Based on these resultant tower-based partitions, we introduce how to align them as follows.

4.3 Spatial Alignment

For the state-of-the-art models based on single networks [31] [11] [24], 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. 8, 9, 10. 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., [31]. 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 15 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

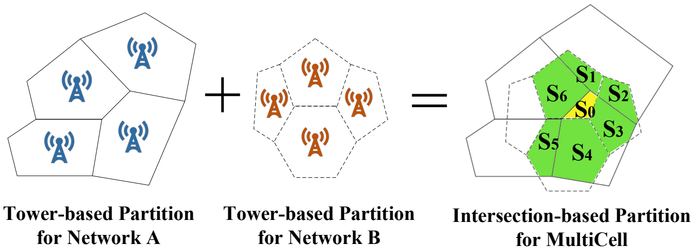

S1
S6
S2
S0
S3
S5
S4
Tower-based Partition for Network A
Tower-based Partition for Network B
Intersection-based Partition for MultiCell

Fig. 15. Cell-Level Alignment

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 [14]. 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 $P^1, P^2, \cdots P^N$ for N networks, we have following constraints for a partition P^i with cells based on a network i: (i) $P^i = \{C^i_1, \cdots, C^i_{|P^i|}\}$ where $|P^i|$ is the total number of cells in P^i; (ii) $\cup_{m=1}^{|P^i|} C^i_m = U$ 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 $S = \{S_1, .., S_x, ..., S_{|S|}\}$ 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.

4.4 Population Estimation

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

4.4.1 Terminologies and Problem Definition: We summarize the notations used in the population estimation in Table 2. Our goal in the population estimation is to estimate the general population $\hat{\mathbf{G}}(S_x, t)$ in subcell S_x at time t given user population C^i_l for $\forall l$ where $i = 1, 2, \cdots, N$, 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 $\mathbf{U}(C^i_l, t, i), \forall l, i = \{1, 2, \cdots, N\}$, and (ii) Phase-2: a general population estimation based on user-population estimation from multiple networks to estimate $\hat{\mathbf{G}}(S_x, t)$ from $\mathbf{U}(S_x, t, i), i = \{1, 2, \cdots, N\}$.

4.4.2 Phase-1: User-Population Estimation. There are many models [31] [11] [24] 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.

$$\mathbf{U}(S_x, t, i) = \sum_{l=1}^{|P^i|} \frac{|C^i_l \cap S_x|}{|C^i_l|} \times \mathbf{U}(C^i_l, t, i), \tag{2}$$

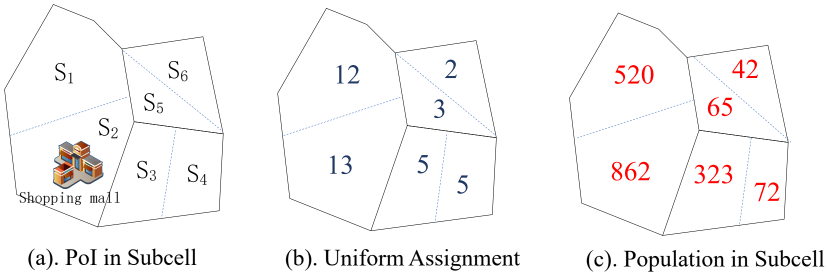
where $\mathbf{U}(S_x, t, i)$ is the user population in a subcell S_x during time t based on data from a network i; $\mathbf{U}(C^i_l, t, i)$ is the user population in a cell C^i_l, which is obtained by the existing work [31]; $|C^i_l \cap S_x|$ is the area of $C^i_l \cap S_x$; $|C^i_l|$ is the area of C^i_l; $|P^i|$ is the total number of cells in the tower-based partition P^i of a network i. The rationale

Table 2. Terminology and Notations

Terminology	Notation	Meaning
Time	t	t_{th} time slot, e.g., 120 means 10AM with 5 minutes time slot
Network	i	i_{th} network
Voronoi Partition	P^i	Voronoi partition based on towers in network i
Cell	C_l^i	i_{th} Voronoi cell with l_{th} tower in network i
Subcell	S_x	x_{th} subcell, subcell is the intersection of Voronoi cells from all networks
User Population	$\mathbf{U}$	number of users (user population)
General Population	$\mathbf{G}, \hat{\mathbf{G}}, \bar{\mathbf{G}}$	population, estimated population from cellphone user, Worldpop population (ground truth)
User Population in Subcell/Cell	$\mathbf{U}(S_x/C_l^i, t, i)$	number of users (user population) in subcell S_x or cell C_l^i
General Population in Subcell/Cell	$\mathbf{G}(S_x/C_l^i, t, i)$	human population (general population) in subcell S_x or cell C_l^i

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 [23].

To address this issue, in this work, we utilize the distribution of PoIs to align cell-level population to subcell-level population. For example, Figure 11 gives the distribution of 586 thousand PoIs among cells based on a tower-based partition. The details of PoI distribution is given in table 1. 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 [23], they are not likely to uniformly distributed across the cell. Fig. 16 shows one example of the influence of PoI on population distribution. There are 6 subcells in 3 Voronoi cells as shown in Fig. 16 (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. 16 (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. 16 (b), which lead to a bias of the user assignment.



(a). PoI in Subcell (b). Uniform Assignment (c). Population in Subcell

Fig. 16. Population in Subcells

As shown in Fig. 11, 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. 15, 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

$$\mathbf{W}(\dot{S_x}, \dot{S_x}(l)) = \frac{1}{\sigma\sqrt{2\mathbf{U}}} e^{-\frac{(\|\dot{S_x}(l) - \dot{S_x}\|_2 - \mu)^2}{\sigma^2}}, \tag{3}$$

where $\dot{S_x}$ is the centroid of the center subcell S_x; $\dot{S_x}(l)$ is a centroid of the lth neighbor subcell $S_x(l)$ of S_x, i.e., from S_1 to S_6 in our example; μ and σ are the mean and standard deviation of distances from all neighbors. By applying this Gaussian filter in Equation 3, we have the formula 4.

$$\mathbf{U}(S_x, t, i) = \sum_{1 \le l \le \mathbf{M}(S_x)} \mathbf{W}(\dot{S_x}, \dot{S_x}(l)) \times \mathbf{U}(S_x(l), t, i) \times \frac{|\dot{S_x}|}{|\dot{S_x}(l)|}, \tag{4}$$

where $\mathbf{M}(S_x)$ is the total number of the neighbor subcells of S_x. We eliminate the influence of the subcell size by $\frac{|\dot{S_x}|}{|\dot{S_x}(l)|}$. With the Equation 4, for a particular subcell S_x and a time slot t, we have N population estimations $(\mathbf{U}(S_x, t, 1), ..., \mathbf{U}(S_x, t, N))$ 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

$$\mathbf{U}(S_x, t, i) = \mathbf{U}(C_l^i, t, i) \times \frac{\mathbf{U}(S_x, t, i)}{\sum_{S_y \in C_l^i} \mathbf{U}(S_y, t, i)} \tag{5}$$

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

4.4.3 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 $\mathbf{U}(S_x, t, i), 1 \le i \le N$ to obtain a general population estimation $\mathbf{G}(S_x, t, \forall)$. 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 [31] that there is an exponential relationship between user population $\mathbf{U}(S_x, t, i)$ estimated by a network i and general population $\mathbf{G}(S_x, t, i)$ inferred by i, i.e.,

$$\mathbf{G}(S_x, t, i) = \alpha^i_{S_x \cdot t} \times (\mathbf{U}(S_x, t, i))^{\beta^i_{S_x \cdot t}}, \tag{6}$$

where $\alpha^i_{S_x \cdot t}$ and $\beta^i_{S_x \cdot t}$ 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 $\mathbf{G}(S_x, t, i)$, given a user population $\mathbf{U}(S_x, t, i)$ 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 $\mathbf{G}(S_x, t, i)$ at different time slots are almost impossible to obtain, as we mentioned in Section 5.1. Some datasets based on census (e.g., Worldpop [14]) 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 13. To address this issue, the state-of-the-art population models [31] 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 $\mathbf{U}(S_x, t, i)$ be the user population estimated by network i; $\mathbf{G}(S_x, t, i)$ be the general population inferred by $\mathbf{U}(S_x, t, i)$. Our objective is to combine $\mathbf{G}(S_x, t, i)$ from different networks $\{1, \cdots, i, \cdots, j, \cdots, n\}$ to estimate $\mathbf{G}(S_x, t, \forall)$, 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 [31] 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 $\mathbf{G}(S_x, t, \forall)$ 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 [14]) and multiple network data as the input to obtain initial parameters. As in Figure 17, 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 $\mathbf{G}(t_1, i)$ and $\mathbf{G}(t_1, j)$. We omit S_x in Figure 17 for concise representation. To estimate the initial parameters α and β in the subcell S_x, we follow the spacial context-aware method proposed in [31]. 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 [27] [33]. 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 $\mathbf{U}(t_1, i)$ and $\mathbf{G}(t_1, i)$ in the same function group. The context-aware model captures spacial dynamics by the PoI distribution of the city. Thus, based on Equation 6 along with user population $\mathbf{U}(t, i)$ and $\mathbf{U}(t, j)$ obtained by data from network i and j, we can obtain two sets of parameters, i.e., $(\alpha^i_{S_x \cdot t_1}, \beta^i_{S_x \cdot t_1})$ and $(\alpha^j_{S_x \cdot t_1}, \beta^j_{S_x \cdot t_1})$ 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 [32] 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.

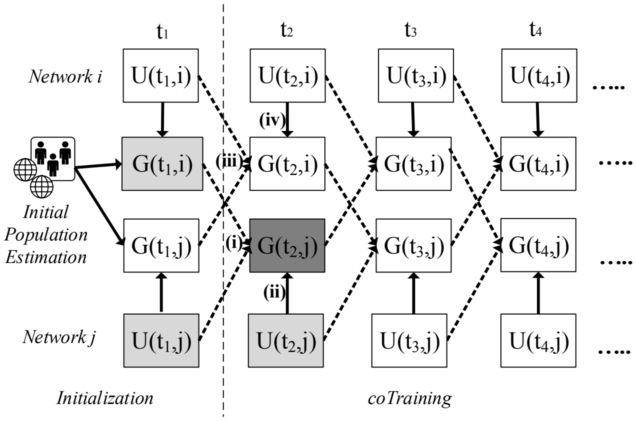


Fig. 17. Cross-Network Data Fusion

(ii) Cross-network Training: With initialization, the key objective of our co-training component is to update two parameters α and β for all networks in the following slots. We show our core idea in Figure 17. 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 17, our co-training starts from time slot t_2: (i) we use $\mathbf{G}(t_1, i)$ and $\mathbf{U}(t_1, j)$ to update two parameters $(\alpha^j_{S_x \cdot t_2}, \beta^j_{S_x \cdot t_2})$; (ii) we use these two updated parameters and new incoming $\mathbf{U}(t_2, j)$ to obtain $\mathbf{G}(t_2, j)$; (iii) we use $\mathbf{G}(t_1, j)$ and $\mathbf{U}(t_1, i)$ to update two parameters $(\alpha^i_{S_x \cdot t_2}, \beta^i_{S_x \cdot t_2})$; (iv) we use these two updated parameters and new incoming $\mathbf{U}(t_2, i)$ to obtain $\mathbf{G}(t_2, i)$.

Note that $\mathbf{G}(t_1, i) = \mathbf{G}(t_1, j)$ since they are equal to the initial estimation based on the census, which leads to $(\alpha^i_{S_x \cdot t_2}, \beta^i_{S_x \cdot t_2}) = (\alpha^i_{S_x \cdot t_1}, \beta^i_{S_x \cdot t_1})$ and $(\alpha^j_{S_x \cdot t_2}, \beta^j_{S_x \cdot t_2}) = (\alpha^j_{S_x \cdot t_1}, \beta^j_{S_x \cdot t_1})$. The reason is that the parameters $(\alpha^i_{S_x \cdot t_1}, \beta^i_{S_x \cdot t_1})$ are inferred from $(\mathbf{U}(t_1, i), \mathbf{G}(t_1, i))$ and the parameters $(\alpha^i_{S_x \cdot t_2}, \beta^i_{S_x \cdot t_2})$ are updated by points $(\mathbf{U}(t_1, i), \mathbf{G}(t_1, j))$ from time slot t_1. We can infer $(\alpha^j_{S_x \cdot t_2}, \beta^j_{S_x \cdot t_2}) = (\alpha^j_{S_x \cdot t_1}, \beta^j_{S_x \cdot t_1})$ in a similar way. However, $\mathbf{G}(t_2, i)$ may not be equal to $\mathbf{G}(t_2, j)$ because based on Equation 6, these two sets of parameters are the same, but $\mathbf{U}(t_2, i)$ and $\mathbf{U}(t_2, j)$ may change compared to $\mathbf{U}(t_1, i)$ and $\mathbf{U}(t_1, j)$ based on real-world data from network i and j. The difference between $\mathbf{G}(t_2, i)$ and $\mathbf{G}(t_2, j)$ makes our cross-network training effective.

To generalize to a multiple network scenario, as in Figure 18, in a slot t, for a network i (e.g., Network 1), we first use the general population estimated by another network $\{1, \cdots, i-1, i+1, \cdots, n\}$ 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 18). 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 17). Finally, the average values of all estimations from all networks are the output of this cross-network training $\mathbf{G}(S_x, t, \forall)$ for a slot t and subcell S_x.

A standard approach to update model parameters is to use the *Least Squares* method [15], 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 7, where μ 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 α and β in our model are obtained from updated Var

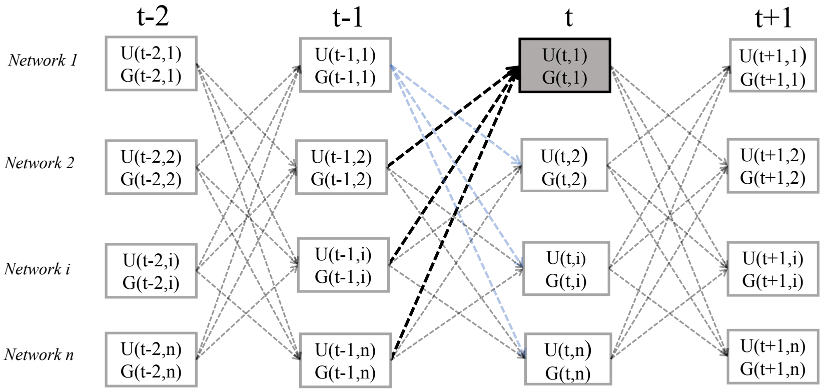


Fig. 18. General Cross-Network Data Fusion

and Cov. This method requires that μ_x, μ_y, Var and Cov are memorized to be utilized at the next time slot.

$$\begin{aligned}
\delta_x^i &= x_i - \mu_x^{i-1};\\
\delta_y^i &= y_i - \mu_y^{i-1};\\
Var(X_i) &= \frac{n-1}{n^2}{\delta_x^i}^2 - \frac{Var(X_{i-1})}{n} + Var(X_{i-1});\\
Cov(X_i, Y_i) &= \frac{n-1}{n^2}\delta_x^i\delta_y^i - \frac{Cov(X_{i-1}, Y_{i-1})}{n} + Cov(X_{i-1}, Y_{i-1});
\end{aligned} \tag{7}$$

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 γ is a ratio depending on the number of networks.

5 MODEL: MULTICELL 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.

(ii) 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.

(iii) Spatial Alignment: Based on the method in Section 4.3, 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 19. We found that even with three networks, we have a much finer granularity compared to three tower-based partitions in Figure 8, 9, 10. As in Fig. 20, with our MultiCell based on subcells, the downtown areas are covered by 36057 subcells, which leads to an average area of 11 thousand m^2. This subcell partition improves our spatial granularity by a factor of 10, compared to the

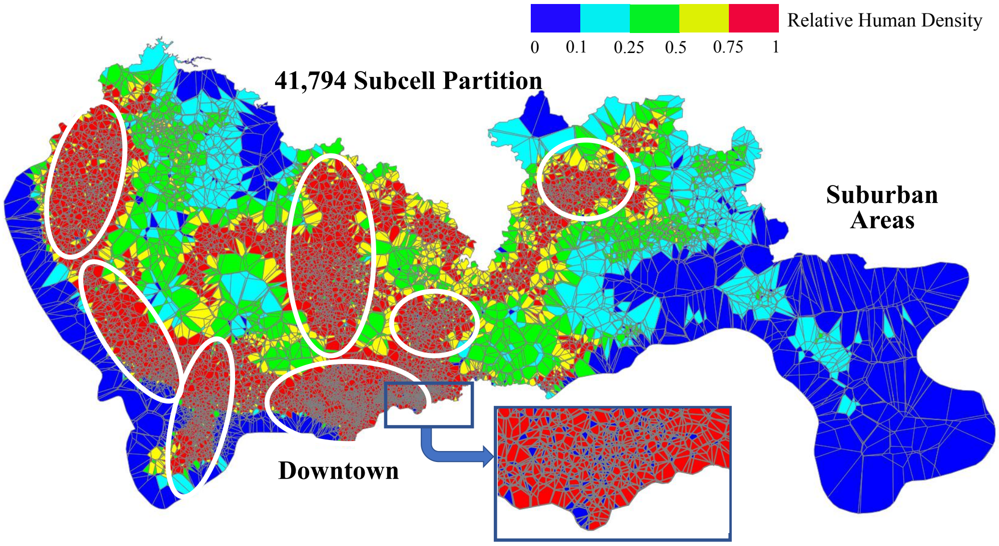


Fig. 19. MultiCell Spatial Partition based on Intersections

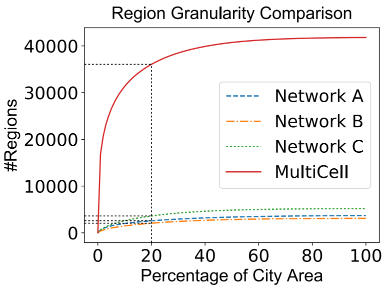


Fig. 20. Cell Coverage

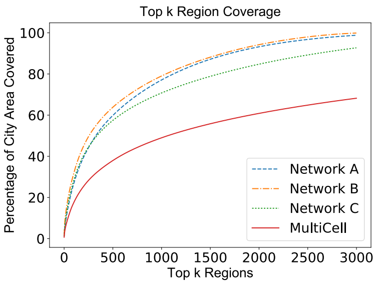


Fig. 21. Top k Region Coverage

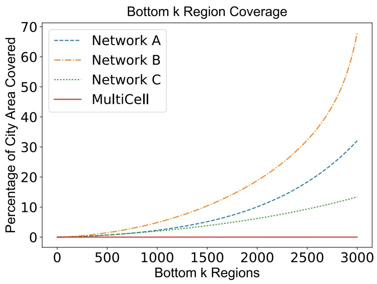
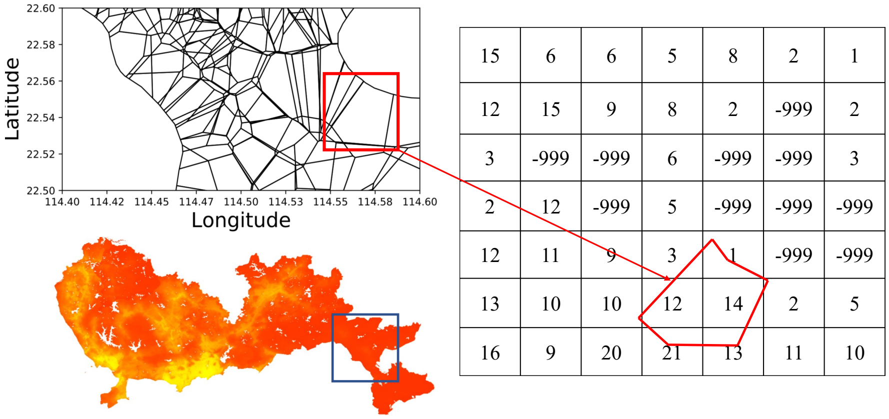


Fig. 22. Bottom k Region Coverage

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 paper 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 21 and Figure 22, 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 6



15	6	6	5	8	2	1
12	15	9	8	2	-999	2
3	-999	-999	6	-999	-999	3
2	12	-999	5	-999	-999	-999
12	11	9	3	1	-999	-999
13	10	10	12	14	2	5
16	9	20	21	13	11	10

Fig. 23. Mapping Worldpop to Population in Subcells

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. 23. First, we calculate the overlapping area of two partitions and then apply function in Equation 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 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.

$$\hat{\mathrm{G}}(R_l, t, i) = \sum_{x=0}^{n} \frac{|R_l \cap S_x|}{|S_x|} \times \hat{\mathrm{G}}(S_x, t, i), \tag{8}$$

We examine this relationship by comparing user populations with the ground truth in Figures 24, 25 and 26 27. 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.

6 EVALUATION

6.1 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 28 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.

(ii) Performance Metrics: Given the extensive usage in population models [31] [11] [24], we utilize the following correlation coefficient and normalized root mean square error (RMSE) as the metrics, respectively.

$$Correlation = \frac{\Sigma_{l=1}^{|S|}[\hat{\mathrm{G}}(S_l,t) - \frac{1}{|S|}\Sigma_{k=1}^{|S|}\hat{\mathrm{G}}(S_k,t)] \cdot [\bar{\mathrm{G}}(S_l,t) - \frac{1}{|S|}\Sigma_{k=1}^{|S|}\bar{\mathrm{G}}(S_k,t)]}{\sqrt{\Sigma_{l=1}^{|S|}[\hat{\mathrm{G}}(S_l,t) - \frac{1}{|S|}\Sigma_{k=1}^{|S|}\hat{\mathrm{G}}(S_k,t)]^2} \cdot \sqrt{\Sigma_{l=1}^{|S|}[\bar{\mathrm{G}}(S_l,t) - \frac{1}{|S|}\Sigma_{k=1}^{|S|}\bar{\mathrm{G}}(S_k,t)]^2}}.$$

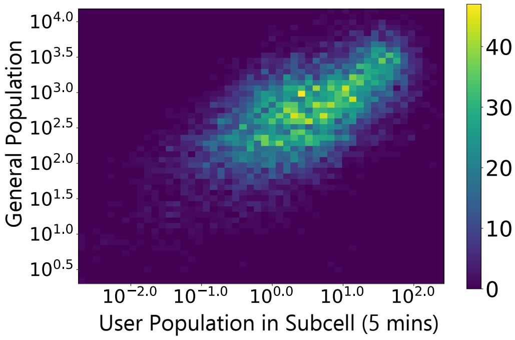


Fig. 24. GT & Network A

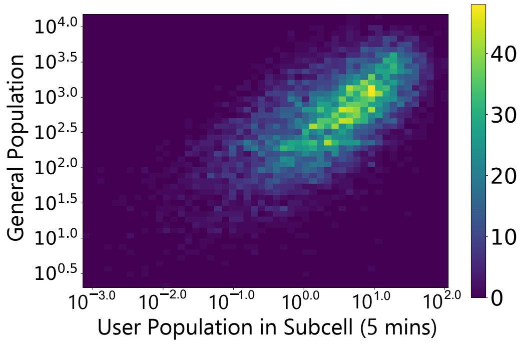


Fig. 25. GT & Network B

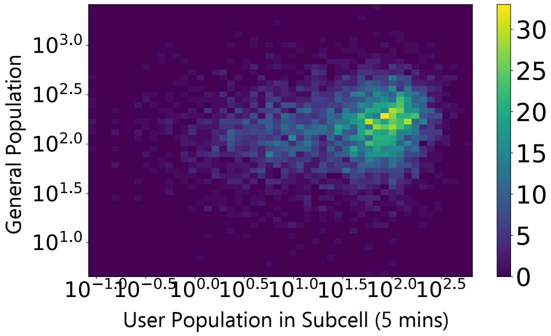


Fig. 26. GT & Network C

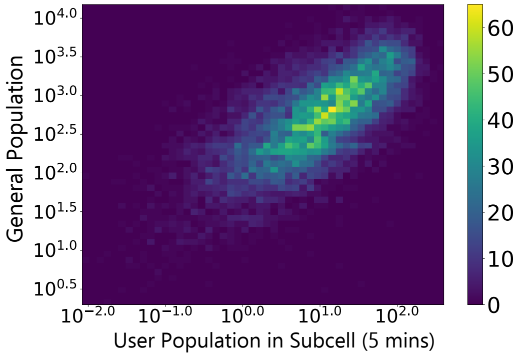
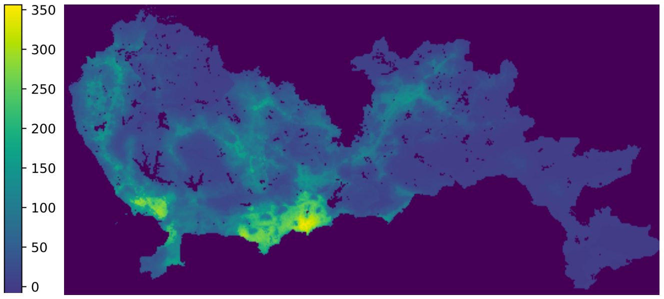


Fig. 27. GT & Shared Net



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

Fig. 28. Ground Truth of Shenzhen Population

$$RMSE = \frac{\sqrt{\frac{1}{|S|}\Sigma_{l=1}^{|S|}[\hat{\mathrm{G}}(S_l,t) - \bar{\mathrm{G}}(S_l,t)]^2}}{\frac{1}{|S|}\Sigma_{l=1}^{|S|}\bar{\mathrm{G}}(S_l,t)}.$$

where $\bar{\mathrm{G}}(S_l, t)$ is the ground truth for the subcell S_l during the time slot t, and $\hat{\mathrm{G}}(S_l, t)$ 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 [31] 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 $U(S_x, t, Shared) = U(S_x, t, A) + U(S_x, t, B) + U(S_x, t, C)$. 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.

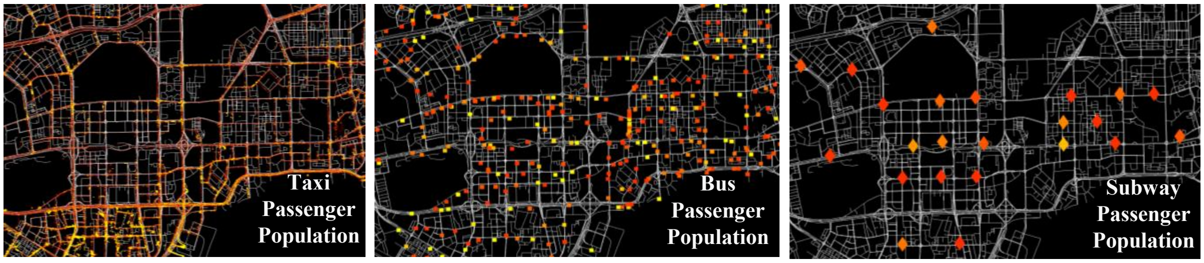


Fig. 29. Transportation Passenger Population

(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 [11] [24]. 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 [31]. It has been showed by the previous research that there are strong correlations between real-time urban population and passenger population from transportation systems [31]. 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 29. For the evaluation, we map these locations into our spatial partition and test the correlations in these locations only.

6.2 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 a 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 30 and 31. 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 32, 33 and 34, 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.

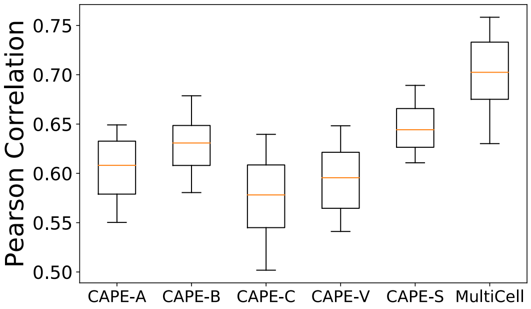


Fig. 30. Correlation

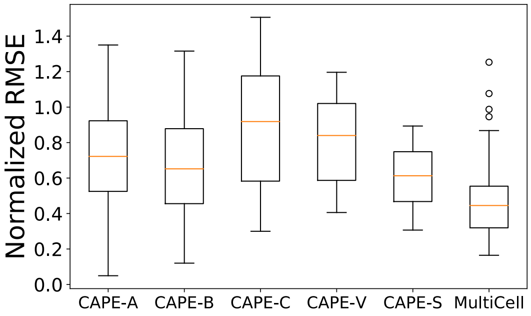


Fig. 31. RMSE

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 [14]. Fig. 12 (c) 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.

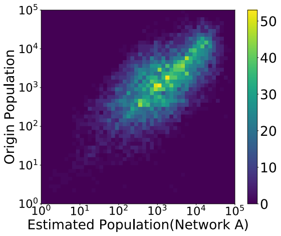


Fig. 32. CAPE-A

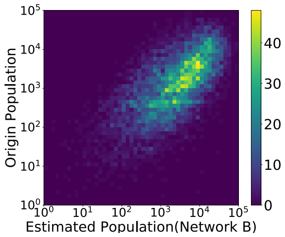


Fig. 33. CAPE-B

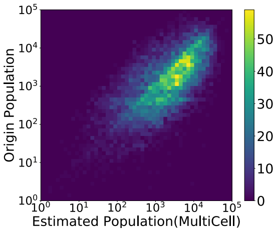


Fig. 34. MultiCell

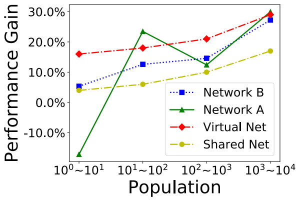


Fig. 35. Population

(ii) Impact of Population: We quantify the impact of populations on our performance gains in Figure 35, 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 8. In Figure 36, 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 37. 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 1. The stations cover 122 out of 496 administrative regions and 286.216km^2 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 38. 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

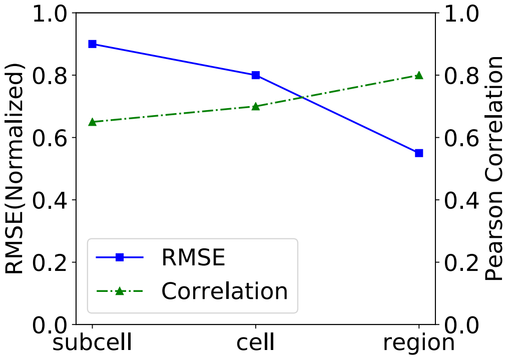


Fig. 36. S-Granularity

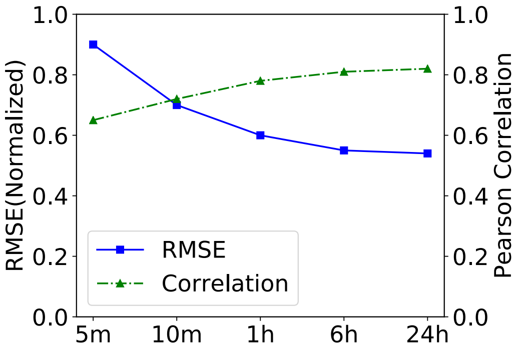


Fig. 37. T-Granularity

regions by Equation 1. The correlations are given in Figure 39. 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 40. 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.

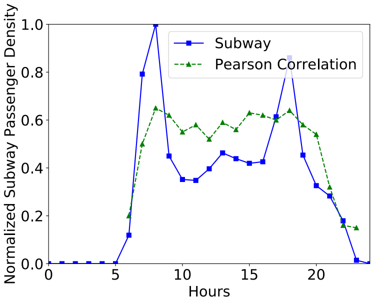


Fig. 38. Subway Correlation

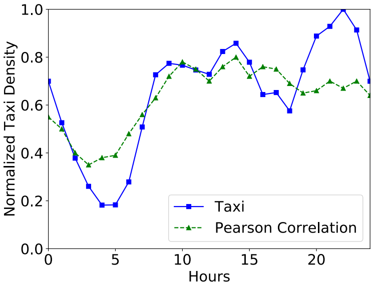


Fig. 39. Taxi Correlation

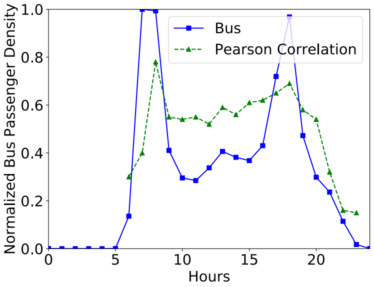


Fig. 40. Bus Correlation

7 RELATED WORK

Analyzing the human population based on multiple networks is crucial for many real-world applications, e.g., urban planning [37] and transportation [2] [21] [16]. In general, our work is directly related to population modeling and system fusion from multiple systems.

7.1 Population Modeling

Due to its various applications and recent advances in data collection techniques, population modeling has been a popular topic since 2000 [9] [10] [3] [8] [24]. These work has 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 [14] 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 [31] and populations in European countries [13]. However, they either only consider single network [31] or theoretically formulate multiple network problems with only synthetic data [13].

7.2 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 [30] [18] [17], e.g., integrating CDR data with census data to model metropolitan-scale human mobility [23]; aligning speeds of buses, trucks and taxis on road segments as spatial granularity to estimate speeds by a statistic model [36]; inferring road maps with OpenStreetMap and GPS trajectories [7]; combining several models to obtain a model with the minimized difference to all source models [26]. 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 [30].

7.3 Summary

Similar to the above 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 MutliCell 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.

8 CONCLUSION

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.

ACKNOWLEDGMENTS

The authors would like to thank anonymous reviewers for their valuable comments. This work is partially supported by Rutgers Global Center Grant, Rutgers Research Council Grant, China 973 Program (2015CB352400), National Natural Science Foundation of China (41401470, 41771441).

REFERENCES

[1] 2016. Cellphone Penetration Rate. (2016). https://en.wikipedia.org/wiki/List_of_countries_by_number_of_mobile_phones_in_use

[2] Javed Aslam, Sejoon Lim, Xinghao Pan, and Daniela Rus. [n. d.]. City-scale traffic estimation from a roving sensor network. In *Proceedings of 10th ACM Conference on Embedded Network Sensor Systems (SenSys '12)*.

[3] Deborah Balk and Gregory Yetman. 2004. The global distribution of population: evaluating the gains in resolution refinement. *Center for International Earth Science Information Network* (2004).

[4] Richard Becker, Ramón Cáceres, Karrie Hanson, Sibren Isaacman, Ji Meng Loh, Margaret Martonosi, James Rowland, Simon Urbanek, Alexander Varshavsky, and Chris Volinsky. 2013. Human Mobility Characterization from Cellular Network Data. *Commun. ACM* 56, 1 (Jan. 2013), 74–82. https://doi.org/10.1145/2398356.2398375

[5] Sourav Bhattacharya, Santi Phithakkitnukoon, Petteri Nurmi, Arto Klami, Marco Veloso, and Carlos Bento. [n. d.]. Gaussian Process-based Predictive Modeling for Bus Ridership *(UbiComp '13)*.

[6] Federica Bogo and Enoch Peserico. [n. d.]. Optimal Throughput and Delay in Delay-tolerant Networks with Ballistic Mobility *(MobiCom '13)*.

[7] Chen Chen, Cewu Lu, Qixing Huang, Qiang Yang, Dimitrios Gunopulos, and Leonidas Guibas. 2016. City-scale map creation and updating using GPS collections. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1465–1474.

[8] John I Clarke. 1978. Population geography. *Progress in human geography* 2, 1 (1978).

[9] Uwe Deichmann. 1996. A review of spatial population database design and modeling. *National Center for Geographic Information and Analysis* (1996).

[10] Uwe Deichmann, Deborah Balk, and Greg Yetman. 2001. Transforming population data for interdisciplinary usages: from census to grid. *Center for International Earth Science Information Network* (2001).

[11] Pierre Deville, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R. Stevens, Andrea E. Gaughan, Vincent D. Blondel, and Andrew J. Tatem. 2014. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences* 111, 45 (2014), 15888–15893. https://doi.org/10.1073/pnas.1408439111 arXiv:http://www.pnas.org/content/111/45/15888.full.pdf

[12] Rex W. Douglass, David A. Meyer, Megha Ram, David Rideout, and Dongjin Song. 2015. High resolution population estimates from telecommunications data. *EPJ Data Science* 4, 1 (16 May 2015), 4. https://doi.org/10.1140/epjds/s13688-015-0040-6

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

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

[15] John Fox. 1997. *Applied regression analysis, linear models, and related methods.* Sage Publications, Inc.

[16] Raghu Ganti, Mudhakar Srivatsa, Anand Ranganathan, and Jiawei Han. [n. d.]. Inferring Human Mobility Patterns from Taxicab Traces *(UbiComp '13)*.

[17] Jing Gao, Wei Fan, Jing Jiang, and Jiawei Han. [n. d.]. Knowledge Transfer via Multiple Model Local Structure Mapping. In *ACM KDD'08*.

[18] Jing Gao, Wei Fan, Yizhou Sun, and Jiawei Han. [n. d.]. Heterogeneous Source Consensus Learning via Decision Propagation and Negotiation. In *ACM KDD'09*.

[19] Philippe Golle. 2006. Revisiting the Uniqueness of Simple Demographics in the US Population. In *Proceedings of the 5th ACM Workshop on Privacy in Electronic Society (WPES '06)*. ACM, New York, NY, USA, 77–80. https://doi.org/10.1145/1179601.1179615

[20] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. 2008. Understanding individual human mobility patterns *(Nature)*.

[21] Shaohan Hu, Lu Su, Hengchang Liu, Hongyan Wang, and Tarek F Abdelzaher. 2015. Smartroad: Smartphone-based crowd sensing for traffic regulator detection and identification. *ACM Transactions on Sensor Networks (TOSN)* 11, 4 (2015), 55.

[22] Sibren Isaacman, Richard Becker, Ramón Cáceres, Stephen Kobourov, James Rowland, and Alexander Varshavsky. [n. d.]. A Tale of Two Cities. In *HotMobile '10*.

[23] Sibren Isaacman, Richard Becker, Ramón Cáceres, Margaret Martonosi, James Rowland, Alexander Varshavsky, and Walter Willinger. [n. d.]. Human Mobility Modeling at Metropolitan Scales *(MobiSys '12)*.

[24] Chaogui Kang, Yu Liu, Xiujun Ma, and Lun Wu. October 2012. Towards Estimating Urban Population Distributions from Mobile Call Data. *Journal of Urban Technology 19(4)* (October 2012).

[25] Neal Lathia and Licia Capra. [n. d.]. How Smart is Your Smartcard?: Measuring Travel Behaviours, Perceptions, and Incentives *(UbiComp '11)*.

[26] Qi Li, Yaliang Li, Jing Gao, Bo Zhao, Wei Fan, and Jiawei Han. [n. d.]. Resolving Conflicts in Heterogeneous Data by Truth Discovery and Source Reliability Estimation. In *ACM SIGMOD'14*.

[27] David Mimno and Andrew McCallum. 2012. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. *arXiv preprint arXiv:1206.3278* (2012).

[28] Atsuyuki Okabe, Barry Boots, Kokichi Sugihara, and Sung Nok Chiu. [n. d.]. *Spatial Tessellations:Concepts and Applications of Voronoi Diagrams.* Wiley.

[29] Lijun Sun, Der-Horng Lee, Alex Erath, and Xianfeng Huang. [n. d.]. Using Smart Card Data to Extract Passenger's Spatio-temporal Density and Train's Trajectory of MRT System *(UrbComp '12)*.

[30] Sihong Xie, Jing Gao, Wei Fan, Deepak Turaga, and Philip S. Yu. [n. d.]. Class-Distribution Regularized Consensus Maximization for Alleviating Overfitting in Model Combination. In *ACM KDD'14*.

[31] Fengli Xu, Pengyu Zhang, and Yong Li. 2016. Context-aware real-time population estimation for metropolis. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 1064–1075.

[32] Shipeng Yu, Balaji Krishnapuram, Rómer Rosales, and R. Bharat Rao. 2011. Bayesian Co-Training. *J. Mach. Learn. Res.* 12 (Nov. 2011), 2649–2680. http://dl.acm.org/citation.cfm?id=1953048.2078190

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

[34] Yasir Zaki, Thomas Pötsch, Jay Chen, Lakshminarayanan Subramanian, and Carmelita Görg. 2015. Adaptive congestion control for unpredictable cellular networks. In *ACM SIGCOMM Computer Communication Review*, Vol. 45. ACM, 509–522.

[35] Desheng Zhang, Jun Huang, Ye Li, Fan Zhang, Chengzhong Xu, and Tian He. 2014. Exploring human mobility with multi-source data at extremely large metropolitan scales. In *Proceedings of the 20th annual international conference on Mobile computing and networking*. ACM, 201–212.

[36] Desheng Zhang, Juanjuan Zhao, Fan Zhang, and Tian He. 2015. UrbanCPS: a cyber-physical system based on multi-source big infrastructure data for heterogeneous model integration. In *Proceedings of the ACM/IEEE Sixth International Conference on Cyber-Physical Systems*. ACM, 238–247.

[37] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. 2014. Urban Computing: Concepts, Methodologies, and Applications. *ACM Trans. Intell. Syst. Technol.* 5, 3, Article 38 (Sept. 2014), 55 pages. https://doi.org/10.1145/2629592

Received February 2018; revised May 2018; accepted September 2018