A Study of Population Census System Based on Signaling Data: A Case Study with Greater Bay Area Data

Yi Wang[1,2(✉)], Qingxin Zhao[1], Shenhang Xu[3], Qingfeng Zhou[4], Qi Wang[5], Jun Zhang[5], Fan Zhang[5], Ye Li[5], and Chen Tian[3]

[1] Nanjing University of Posts and Telecommunications, Nanjing, China
yiwang@njupt.edu.cn
[2] Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen, China
[3] Nanjing University, Nanjing, China
[4] Shenzhen Urban Planning and Land Resource Research Center, Shenzhen, China
[5] Shenzhen Institute of Beidou Applied Technology, Shenzhen, China

Abstract. This paper introduces a novel system for obtaining regional population data using mobile signaling data to overcome the limitations of traditional methods like censuses and surveys. The system leverages real mobile signaling data from the Guangzhou platform and employs a base station clustering method based on traffic communities to address signaling drift and ensure accurate location data. It enhances population statistic accuracy using eigenvalue curve fitting and optimization methods to determine feature thresholds. Compared to the seventh national census data, the system achieves over 95% precision for the working population and 88% for permanent residents. Additionally, it validates its large-scale application capability through the analysis of temporal and spatial population distribution in the Greater Bay Area. This approach offers a cost-effective and efficient solution for regional population data acquisition, supporting better-informed policy-making and urban planning.

Keywords: Censuses data · Population statistics services · Mobile signaling data · Traffic cell · Greater Bay Area

1 Introduction

Regional population data is vital for government operations and policy making. In 1790, the US had its first modern census. It used six questions and 580 enumerators over 18 months to collect data on occupations etc. [5]. This labor intensive process made scholars look for more efficient methods.

Supported by the Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), under Grant No. GML-KF-22-18.

X. Meng et al. (Eds.): BDSC 2025, CCIS 2622, pp. 72–82, 2026.
https://doi.org/10.1007/978-981-95-0880-8_7

The wide use of mobile phones and urban base stations gives a new solution. Researchers like [1,4] have used mobile phone signaling data to estimate population density from 2015 They refined estimations with the LandScan dataset.

Each mobile device has a unique IMSI. With regular travel records in the database, big data and analytical methods can figure out users' residential status, employment, and mobility patterns. Importantly, the signaling data in this study is encrypted for privacy protection. This article analyzes things like resident population, working population, and dynamic population. However, using this approach needs to deal with three main challenges: (a)Signal Drift: Signal data can drift between overlapping base station coverage areas, causing inaccurate location and population statistics. (b)Accuracy of Population Estimates: It is crucial to accurately estimate the permanent residents and working population by considering the inherent connection between signal data and population data. (c)Large-Scale Validation: Wide-scale validation is essential for municipal or provincial applications.

We propose solutions to these challenges. They can be summarized as the contributions of this paper as shown below:

- It introduces an innovative solution to address signal drift using traffic cell clustering, effectively resolving the issue (Sect. 2).
- It employs a curve fitting method to determine the relationship between feature values and actual population numbers, optimizing eigenvalues and thresholds (Sect. 3).
- The system was validated in a large-scale, densely populated environment using signal data from Shenzhen and the Greater Bay Area from September to November 2020 (Sect. 4).

2 System Architecture and Data Processing

This section outlines the system's structure, including inputs, data processing, and outputs. After that, we will introduce the input data set and its features. Finally, we will detail data processing, especially handling signaling drift. The rest will be covered later.

2.1 System Architecture

The main architecture of this study, as illustrated in Fig. 1, comprises three constituent parts: (a) data input; (b) data processing and metric computation; (c) data output. The data input part needs to gather four different types of data as depicted in Fig. 1. The collected raw data undergoes formatting via automated processes (Sect. 2.2).

The second part has three facets: data sorting, preprocessing, and metric computation. Data sorting arranges signaling data chronologically. Preprocessing removes data with missing key information after formatting and addresses signaling drift (Sect. 2.3). Metric computation uses specific methods to calculate

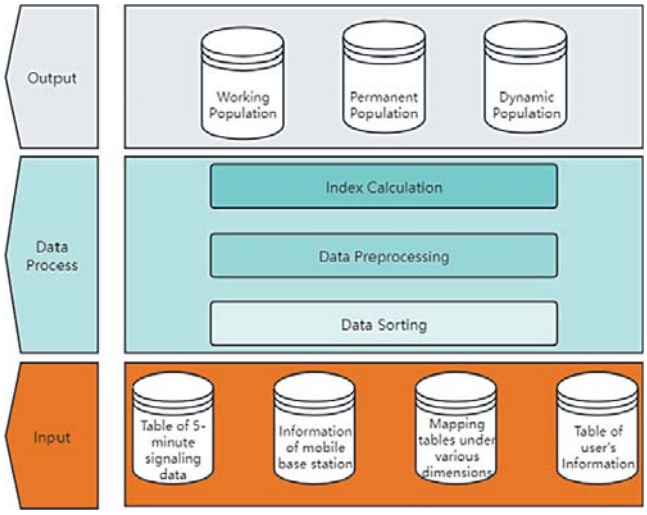

Fig. 1. System structure.

metrics such as permanent residents, working population, dynamic population etc. (Sect. 3).

The output part involves data output and result demonstration. A comprehensive demonstration will be provided in Sect. 4.

2.2 Data Source

As depicted in Fig. 1, this paper utilizes four distinct datasets, namely: (a) a 5-minute base station signaling data table, (b) a mobile base station information table, (c) mapping tables for various dimensions (traffic cells, streets), and (d) a user information table. The study period spans from September 1, 2020, to November 30, 2020, encompassing 90 d, with base station signaling data recorded at 5-minute intervals. The spatial dimensions of the data encompass nine Greater Bay Area cities, namely Shenzhen, Guangzhou, Zhuhai, Foshan, Huizhou, Dongguan, Zhongshan, Jiangmen, and Zhaoqing, with a total of 3,416 traffic cells and 18,000 grids. The 5-minute signaling data table is sourced from China Mobile's Guangzhou branch and encompasses fundamental data, including encrypted IMSI numbers, signaling dates, signaling times, cities, and base station IDs, with a total of 2,148,652 base stations. A detailed description of this 5-minute signaling data is presented in Table 1.

It is essential to clarify that the data we utilized exclusively comprises China Mobile's signaling data, and does not encompass signaling data from China Unicom and China Telecom. It's worth noting that China Mobile's data accounts for approximately 60% of the total mobile user base in the Greater Bay Area.

Additional datasets are the mobile base station information table, mapping tables for different dimensions (traffic sectors, streets), and user information table. Specific details of these data tables are in Table 2.

Table 1. The 5-minute base station signaling data table.

Field	Content	Remarks
msisdn	String	Phone number
imis	String	International mobile subscriber identification number
cgi	String	Cell global identity
h_min	String	Time getting this singling
city	String	The code of city
day	String	Day getting this singling
hour	String	Hour getting this singling

Table 2. The dataset of (b) (c) (d).

Field	Content
Mobile base station information	Cell global identity
	The name of city
	The name of district
	Coverage scene
	Latitude of cell
	Longitude of cell
Mapping for various dimensions	Unique code of street
	The identity of street
	The latitude of street
	The longitude of street
User information table	International mobile subscriber
	-identification number
	The age of user
	The sex of user
	The registration location of imsi
	The month of registration
	Whether the user is in working age

2.3 Data Preprocessing

Raw data from mobile network operators must be cleaned and preprocessed. When a mobile device enters an area with overlapping signals from two base stations, it may frequently switch between them, causing signal oscillation and disrupting the identification of dwell points. We propose a solution for signaling drift based on geographic clustering. We categorize base stations into traffic cells and map their global identities to unique codes. As shown in Fig. 2, the blue area represents a traffic cell's coverage, with base stations indicated by icons. A device connected to base stations within this cell (points A to E) should not switch to a

station outside the cell (e.g., D) unless due to drift. We exclude such interspersed signals to ensure accurate results. We use traffic cells as the smallest unit for population division, treating base stations collectively to minimize signaling drift and enhance location recognition accuracy.

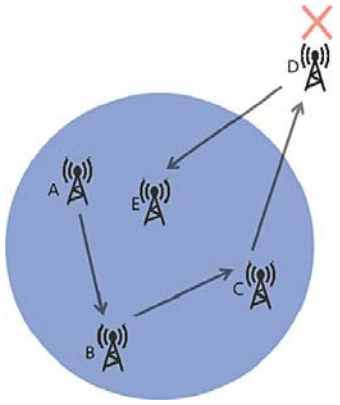

Fig. 2. Delete drift data.

This method involves several steps:

- Excluding data with critical information gaps, such as those lacking user IDs or essential timestamps.
- Establishing a bidirectional mapping between base station information in the signaling data and the corresponding traffic cells, denoted as cell global identity to uniquecode.
- Organizing each user's sampled mobile signaling data in chronological order, as maintaining a time series forms the foundation for subsequent data processing.
- Labeling the current region information as "uniquecode" and comparing it with the adjacent previous region information ("before-uniquecode") and subsequent region information. If "before-uniquecode" equals "after-uniquecode" and "uniquecode" differs from "before-uniquecode", adjusting "uniquecode" to match "before-uniquecode".

3 Statistical Methods for Demographic Data Based on Fitting and Eigenvalues

What information we can extract through the analysis of signaling data include individuals' workplace, mobility patterns, and residential arrangements. Therefore, in the subsequent sections, we will analyze demographic information from three distinct angles: working population, permanent residents, and mobile population.

3.1 Working Population

Our methodology involves selecting optimal features through curve fitting and optimizing their thresholds to enhance accuracy. We identify the working population by considering factors such as an individual's ability to work, age, and duration of stay in a location. We examine the "can-work" field in the mobile user information table, filtering signaling data between 8 AM and 7 PM, and selecting data points where "can-work" is marked as *yes*. We group the data by user and traffic cell, calculating the count of signaling data points (*count*) and the number of days (*count-day*) for each user within each traffic cell. We retain the records corresponding to the maximum *count* value. The results show a linear relationship between the working population and *count-day*, represented by the linear regression function:

$$f(x) = -39267.0667x + 3028546.2444. \tag{1}$$

where x is *count-day*. The correlation coefficient r is calculated by Eq. 2:

$$r = \frac{COV(x, y)}{\sqrt{Var|x| * Var|y|}}. \tag{2}$$

The calculated correlation coefficient of 0.999 indicates a strong fit. However, to match Dongguan's actual population of 5,284,000, *count-day* would unrealistically need to be -56. We find that the "can-work" estimates are lower in most cities, likely due to the limited number of individuals marked as "able to work". So we compute the workforce based on age attributes, defining individuals aged 19–55 as the working-age population. The results for Dongguan show a rapidly decreasing inverse proportion function as the number of days increases, fitted as Eq. 3:

$$f(x) = 1745100 + \frac{6551000}{x}. \tag{3}$$

To enhance the accuracy of working population estimates in Dongguan, we introduce two parameters: *duration-threshold* (minimum continuous stay in a cell) and *daymun-threshold* (minimum days present in a cell per month). Using data from Shenzhen, we analyze the impact of these parameters on the working population. Results show a linear decrease in the working population as *daymun-threshold* increases, with a correlation coefficient of 0.9960 for a duration-threshold of 30. The linear relationship is modeled as:

$$f(x) = -709060x + 17262881. \tag{4}$$

where x represents the *daymun-threshold*. For varying duration-thresholds, the linear coefficient changes. When the duration-threshold increases, multiple inflection points are observed, and a higher-order curve fit is used:

$$f(x) = 9.2726x^3 - 1859.08x^2 + 71672x + 5755658. \tag{5}$$

with a correlation coefficient of 0.9978. Based on these analyses, optimal threshold values for Shenzhen are determined to be duration-threshold = 40 and daymun-threshold = 13, closely approximating the real working population.

3.2 Permanent Population

We use the term "permanent resident population" to refer to individuals who have resided at a location for more than six months. However, due to data constraints, our calculations are based on a single month's data. We assess the count of permanent residents using different threshold values (*count-day*, representing the minimum number of days spent in a cell to be classified as a permanent resident). For example, in Dongguan and Zhaoqing, the results show a linear decline in the permanent resident population as the number of days increases. Comparing these results with actual population data, we find that using *count-day* allows us to approximate the permanent resident population. For Dongguan, the optimal threshold is 28 d, and for Zhaoqing, it is 10 d. These thresholds align the results with the seventh population census data, where Dongguan has approximately 6 million mobile users and Zhaoqing has around 2.4 million people. To accommodate the majority of cities, we select a centrally appropriate threshold, which will be established in Sect. 4 to determine the number of permanent residents.

3.3 Floating Population

Dynamic population, also referred to as instantaneous population, represents the population count at a specific point in time, reflecting the population size at that particular moment. Given the availability of base station signaling data, obtaining dynamic data for a particular traffic area over a defined time period can be accomplished by merely counting the distinct IMSIs within that time interval. This approach enables the derivation of dynamic population data.

4 Value Case Comparison

The methods in Sect. 3 clarify the connection between feature values and statistical outcomes, guiding the selection of appropriate thresholds.

4.1 Working Resident

We chose nine sets of threshold values for the nine cities to align the working population estimates with actual economic census and employment data. Table 3 summarizes these thresholds and the resulting working population figures, showing strong concordance with actual job positions, with discrepancies within a 5% margin and statistical accuracy over 95%.

A heat map (Fig. 3) illustrates the working population distribution across neighborhoods in the Greater Bay Area. Shenzhen and Guangzhou have broad working populations, while Dongguan, Foshan, and Zhongshan have significant working populations due to geographic and industrial factors. In contrast, Jiangmen and Zhaoqing have smaller working populations.

Table 3. Table comparing the working population at different thresholds with the actual working population.

City	Number of positions (ten thousand)	Actual number of positions	Minimum duration
Shenzhen	1243.54	796.21	40
Dongguan	566.71	528.40	50
Huizhou	200.36	145.39	90
Guangzhou	888.50	542.22	50
Foshan	420.10	287.32	60
Zhuhai	134.89	99.76	50
Zhongshan	187.56	147.13	40
Jiangmen	122.24	63.16	40
Zhaoqing	73.26	29.93	70

City	Minimum occurrence days	Working population (ten thousand)	Proportions
Shenzhen	13	794.47	0.9978
Dongguan	13	524.89	0.9934
Huizhou	15	141.09	0.9704
Guangzhou	15	531.95	0.9811
Foshan	15	298.00	1.0372
Zhuhai	13	96.59	0.9683
Zhongshan	15	153.28	1.0418
Jiangmen	20	63.67	1.0081
Zhaoqing	20	29.12	0.9730

4.2 Permanent Resident

Based on our analysis, selecting an optimal threshold for Dongguan and Zhaoqing is impractical due to extreme values. Instead, we use a uniform threshold of 20 d for all nine cities, which provides a reasonable statistical outcome. Given that China Mobile users account for approximately 60 For example, the permanent resident population in Guangzhou is estimated at 18.68 million, with a 61.03% fit to the Seventh National Population Census data. In Dongguan, the estimated population is 10.47 million, with a 93.24% alignment to the census data. The working population in the Greater Bay Area is estimated at 52.672 million using signaling data, compared to the actual 77.95 million, resulting in a ratio of 67.6%. Considering China Mobile's 60% market share, the statistical accuracy for the working population is 88%. A heat map (Fig. 4) shows the distribution of the permanent resident population across districts in the Greater Bay Area.

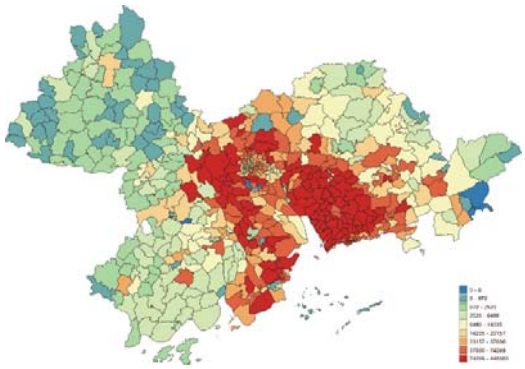

Fig. 3. Street-level working population distribution map in Greater Bay Area.

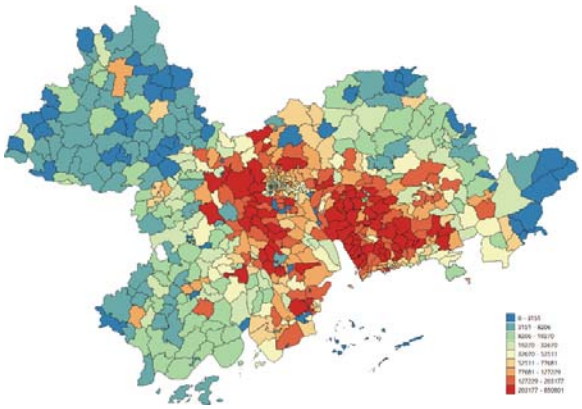

Fig. 4. Street-level permanent resident population distribution map in Greater Bay Area.

4.3 Floating Resident

By analyzing the mobile signaling data for November 2020 in the Greater Bay Area, we conducted a statistical analysis of the dynamic population during both working days and holidays. Spatially, higher population densities are concentrated in Dongguan and Foshan, with moderate levels in Guangzhou, Zhongshan, Zhuhai, and Huizhou, and fewer residents in Zhaoqing and Jiangmen. Some areas maintain a high dynamic population throughout the day. Figure 5 shows that the dynamic population peaks at 18:00 and is lowest at 4:00. The population increases from 4:00 to 8:00, remains stable from 8:00 to 21:00, and decreases from 21:00 to 4:00 the next day. These fluctuations align with daily activities. Major cities like Shenzhen, Guangzhou, and Dongguan show significant daily fluctuations, indicating they attract a large workforce due to employment opportunities.

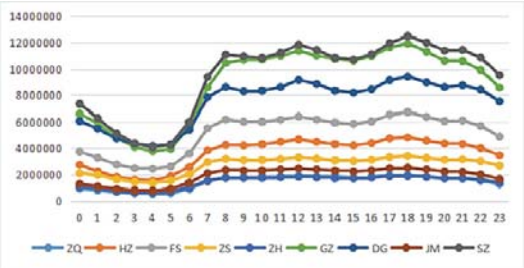

Fig. 5. Distribution of dynamic population in various cities of the Greater Bay Area by time period.

5 Conclusions

This paper presents a method for obtaining regional population data using mobile signaling data from China Mobile's Guangzhou branch. Our method effectively gets population statistics from base station signals, overcoming traditional survey limits. Our feature driven method ensures demographic accuracy by setting feature value thresholds via curve fitting. The model combines base station data and location info to give detailed population statistics. We used this system in the Greater Bay Area, using 2020 signal data to estimate permanent, working, and dynamic populations. Comparisons with the Seventh National Population Census data show good alignment, proving our method is effective.

Acknowledgements. This research was financially supported by the Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), under Grant No. GML-KF-22-18.

References

1. Huang, J., Meng, W., Wu, F.: Research of human flow spatial-temporal distribution and mobility pattern based on signaling data. Comput. Eng. Appl. **55**(23), 53–63 (2016). https://doi.org/10.3778/j.issn.1002-8331.1808-0192
2. O'Hare, W.P.: The importance of census accuracy: uses of census data. In: Differential Undercounts in the US Census: Who is Missed?, pp. 13–24 (2019). https://doi.org/10.1007/978-3-030-10973-8_2
3. Palamà, I., Gringoli, F., Bianchi, G., Melazzi, N.B.: IMSI catchers in the wild: a real world 4G/5G assessment. Comput. Netw. **194**(2), 108137 (2021). https://doi.org/10.1016/j.comnet.2021.108137
4. Ricciato, F., Lanzieri, G., Wirthmann, A., Seynaeve, G.: Towards a methodological framework for estimating present population density from mobile network operator data. Pervasive Mobile Comput. **68**, 101263 (2020). https://doi.org/10.1016/j.pmcj.2020.101263
5. Ruggles, S., Magnuson, D.L.: Census Technology, Politics, and Institutional Change, 1790–2020. J. Am. History **107**(1), 19–51 (2020). https://doi.org/10.1093/jahist/jaaa007

6. Wang, X., et al.: How does socioeconomic status influence social relations? A perspective from mobile phone data. Physica D: Nonlinear Phenomena **615**, 128612 (2023). https://doi.org/10.1016/j.physa.2023.128612
7. Song, S.: Research on regional travel destination choice model based on cellular signaling data. Master's thesis, Southwest Jiaotong University (2022)

An Effective and Efficient Framework for Mining Top-*k* Regional Co-location Patterns

Can Jin[1], Lizhen Wang[2(✉)], Peizhong Yang[1], and Hongmei Chen[1]

[1] School of Information Science and Engineering, Yunnan University, Kunming, China
[2] School of Science and Technology, Dianchi College, Kunming, China
lzhwang@ynu.edu.cn

Abstract. Regional co-location pattern (RCP) mining is a key subfield within spatial co-location pattern mining, focuses on the identification of prevalent co-location patterns within local regions. The RCPs reflect the local relationships between spatial features, which have a wide range of practical application in human daily production and life. Existing RCP mining methods cannot effectively identify the distribution regions of RCPs driven by human activities, and face difficulties in determining a suitable prevalence threshold for mining prevalent RCPs in different instance distribution regions. To address these problems, a novel top-*k* RCP mining framework based on weighted Dirichlet diagram is proposed. The proposed framework first obtains the distribution regions of human activity-driven RCPs through the weighted Dirichlet diagram, and then efficiently detects the top-*k* prevalent RCPs in those areas. In addition, in order to solve the efficiency problem when facing large datasets, a parallel mining scheme is proposed to speed up the RCP mining process. Finally, based on the above methods, a user-friendly demonstration system is developed to promote the application of RCP mining technology in practice.

Keywords: Regional co-location pattern (RCP) mining · Weighted Dirichlet diagram · Top-*k* · Parallelization · Demonstration system

1 Introduction

As data collection technology continues to advance and mobile devices with positioning functions are increasingly utilized, vast volumes of spatial data relevant to human production and life have been generated. Spatial co-location pattern (co-location) mining is a key topic in spatial data mining. It focuses on uncovering subsets of spatial features that prevalently appear nearby, revealing the potential correlation between spatial features, and deeply understanding the laws of human activities. Due to the complication of geospatial and the diversity of human activities, some co-locations only exist in local regions. For example, in bustling commercial districts, upscale restaurants and luxury stores are often located next to each other to meet people's high-end consumption behavior, while this phenomenon is rarely seen in other regions (such as residential districts). Therefore, regional co-location pattern (RCP) mining is proposed to detect co-locations

X. Meng et al. (Eds.): BDSC 2025, CCIS 2622, pp. 83–94, 2026.
https://doi.org/10.1007/978-981-95-0880-8_8