DATA SCIENCE
JOURNAL

Horanont, T, et al. 2018. Resembling Population Density Distribution with Massive Mobile Phone Data. *Data Science Journal,* 17: 24, pp. 1–9. DOI: https://doi.org/10.5334/dsj-2018-024

RESEARCH PAPER

Resembling Population Density Distribution with Massive Mobile Phone Data

Teerayut Horanont[1], Thananut Phiboonbanakit[1] and Santi Phithakkitnukoon[2,3]

[1] Sirindhorn International Institute of Technology (SIIT), Thammasat University, Pathum Thani, TH

[2] Department of Computer Engineering, Faculty of Engineering, Chiang Mai University, TH

[3] Excellence Center in Infrastructure Technology and Transportation Engineering (ExCITE), Faculty of Engineering, Chiang Mai University, TH

Corresponding authors: Teerayut Horanont (teerayut@siit.tu.ac.th), Santi Phithakkitnukoon (santi@eng.cmu.ac.th)

As the mobile phone data (CDR data) has gained an increasing interest in research, such as social science, transportation, urban informatics, and big data, this study aims at examining the representativeness of the CDR data in terms of resemblance of the actual population density distribution from three perspectives; operator's market share, urban-rural user population ratio, and user gender ratio. The results reveal that the representativeness of the data does not scale at the same rate with the operator's market share, the urban-rural user population ratio of 80:20 can best represent the population density distribution, and an equal mixture of male and female user population can best resemble the population density distribution. This study is the first investigation into the representativeness of the CDR data. The findings provide useful information, which can serve an insightful guideline when dealing with the CDR data.

Keywords: mobile phone data analysis; call detail records; data representativeness

1. Introduction

Today, a mobile phone is not just a communication device anymore. It has evolved significantly over the past few years with its additional advanced sensing technologies and useful features for handheld use, which makes it an indispensable part of our everyday lives. With its high penetration rate, a mobile phone is being carried by almost everyone these days. When connecting to the cellular network for voice, short message (SMS), or data services, communication logs are collected by the telecom service providers for billing purposes, in forms of the Call Detail Records (CDR) where each record contains a timestamp, corresponding communication activity (e.g., voice, SMS, or data), and location of the connected cellular tower. To use the service, the mobile phone thus needs to connect to the cellular network via a nearest cellular tower. Therefore, each time when the user connects for the cellular service, the user's communication and location information are recorded. Collectively, CDRs constitute a longitudinal behavioral data that can be analyzed methodically to reveal and understand variety aspects of human behavior at different aggregate levels both in time and space.

Mobile phone data has a great advantage over the traditional human behavioral datasets that are mostly collected through surveys and interviews, which could be inaccurate, limited, expensive, and time-consuming. The use of mobile phone data or CDRs therefore has gained an increasing interest in the research community. Mobile phone data has become more available for research studies in recent years (Kiukkonen *et al*., 2010; Blondel *et al*., 2012; Laurila *et al*., 2012; Montjoye and Smoreda, 2014) as its usefulness has been exposed by many studies with benefits in many application domains, such as regional development, urban planning, (Montjoye and Smoreda, 2014) transport engineering, sociology, and so on. Due to the privacy issue, some countries have strict regulations regarding the use of sensitive personal information, such as medical records, travel cards, and CDRs. Even with the data anonymization, the risk of disclosing personal identity still restricts the use of such data in research. Nonetheless, there are still a number of CDR datasets that have been used cautiously in research (Blondel, Decuyper and Krings, 2015).

CDRs contain both communication logs and location traces. The communication logs carry information regarding voice and text communication that includes call duration, timestamp, connected users, and so on, which can be used to characterize and analyze individual social networks. Personal network is a complex system that requires understanding of its properties and mechanism. Social tie strength (Onnela *et al*., 2006), tie persistence (Navarro *et al*., 2017), network structure (Saramäki and Moro, 2015), and information diffusion (Miritello, Moro and Lara, 2011) are among the on-going research affords in social network analysis that are benefited from the use of CDR data. Through analyzing CDR data, researchers have found interesting results. For example, Phithakkitnukoon et al. (Phithakkitnukoon and Dantu, 2011) investigated on mobile social network structure by defining social tie strength based on which ties are classified into three groups, and found that the scaling ratio across these three group sizes is 8, i.e., each group size is scaled by eight from the adjacent group. Aiello et al. (Aiello, Chung and Lu, 2000) observed a power law degree (number of a person's social ties) distribution, which was described by a massive random graph model. The power law degree distribution indicates that the majority of users have a small number of contacts (ties or degrees), while a tiny fraction of users (nodes) are hubs, or super-connectors. Vaz de Melo et al. (Vaz De Melo *et al*., 2010) described the call durations of individual mobile phone users (call duration distribution) with the Truncated Lazy Contractor (TLC) model that has a heavier tail and head than the log-normal distribution, which can be useful for detecting anomalies, generating synthetic dataset, and summarizing a very large number of phone call records.

In addition to the communication logs, the CDRs also provide location traces of individuals that can be used to advance research in human mobility, which is important for understanding transport behavior that requires a massive amount of data to truly explain or model each phenomenon with interdependent properties. A number of studies benefited from the use of CDRs in human mobility research have yielded interesting findings. For instance, Song et al. (Song *et al*., 2010) found that human mobility is highly predictable, showing an upper bound of 93% predictability that significantly reveals regularity in human movement. Phithakkitunukoon et al. (Phithakkitnukoon, Smoreda and Olivier, 2012) further show that human mobility is greatly influenced by social networks, as they found that 80% of the places that we visit are within just 20 km from a person we know, and we are 15% more likely to be traveling near our weak ties than strong ties. Not only the destinations that we travel to, but how we travel there is also influenced by our social networks as Phithakkitnukoon et al. (Phithakkitnukoon *et al*., 2017) show that strong ties are more important to determine if driving is the person's transport mode choice, whereas weak ties are more important to determine if public transit is the person's choice. Understanding human mobility has a useful implication in transport system design and planning. Demissie et al. (Demissie *et al*., 2016) show that CDRs can be used to infer travel demands that facilitate public transport network design, especially for developing countries where traditional travel surveys are costly and infeasible. CDR-derived mobility pattern is shown to be a reasonable alternative – arguably is perhaps a better option because the results are not biased by subjectivity of the surveyed participants' perception.

Previous studies have suitably done analyses with the mobile phone datasets that produced many interesting results with useful implications, as previously mentioned. However, those datasets only provide a partial view of the population as they are constrained by the data obtained from some provider, which is typically a single regional telecom operator that has a certain coverage rate – but not 100%. No study has yet reported on a proper portion of the CDR data that is statistically suitable to represent the whole population. The results gained from the analysis thereby could be biased by the characteristics of the users of a particular cellular network provider. No study has examined the entire CDR data of the users gathered from all regional providers. Therefore, this study aims at filling in this gap by conducting a study of CDR data from all providers in a region to address this issue. Particularly, this study aims at investigating if the CDR data can be used as a proxy to understand population density from three different perspectives; operator's market share, urban-rural user population ratio, and user gender ratio.

2. Dataset

The main dataset used in this study is an anonymized CDR data of all mobile phone service subscribers in a southeast Asian city. The data was provided by all network operators in the city, which accounts for a total of 1,618,265 mobile users, collected over one full month of May 2014. In 2014, the city population was 1,256,654 people, including 614,756 males (49%) and 641,898 of females (51%). This information is based on the most recent population census data collected in 2014, which is the same period with our CDR data. In this country, a population census is taken every 10 years. The number of mobile phone subscribers was higher than the city population was presumably due to the fact that some users carried dual-SIM phones as well as carried more than one phone (e.g., personal and business numbers). Each CDR contains anonymized user IDs (caller's and callee's), connected cellular tower ID, connected cellular tower geo-location, call duration, and timestamp.

There were five cellular network operators in the city, with a total of 4,630 cellular towers providing the service. Each operator provides their service through a cellular network composed of a different number of antennas (cell towers). The number of subscribers and cellular towers of each operator are listed in **Table 1**.

The city area is about 11,600 km^2, divided into 18 districts. The spatial coverage of the cellular towers in each district area differs across the operators depending on their target customers, which generally varies with the area population density. **Figure 1** shows the cellular tower density (in square kilometer) per district. The average cell tower density per district is 0.40 towers/km^2. The population density distribution across all 18 districts is shown in **Figure 2**, in terms of a total population count in each district (**Figure 2a**) and population per area (**Figure 2b**).

The distribution of the usage characteristics of the users of each network operator in terms of connectivity (i.e., frequency of connections), call duration (minutes per call), and mobility (i.e., number of distinct connected cell towers) is shown in **Figure 3**. The corresponding mean and median values are shown in **Table 2**.

3. Methodology and Results

With the CDRs of all subscribers in the city, our goal was to examine the amount and dimension of the data that can resemble the whole population. The population density distribution across all districts was thus used as a measure of representativeness of CDRs to the actual population. The representativeness of the data can vary with the size of the operator's market share, urban-rural population ratio, and gender ratio. These are the dimensions in which our study aims to investigate.

To obtain the population density information from the CDR data, home location of each user must first be inferred, so that the user population of each district can be calculated. Taking the same approach of Phithakkitnukoon et al. (Phithakkitnukoon, Smoreda and Olivier, 2012), the most frequently used cell tower during the nighttime (10PM – 7AM) is identified as the user's home cell tower. Each district population density was then calculated according to the inferred user home locations. **Figure 4** shows the correlation between the CDR-based population density and census information, with the R-squared of 0.89.

Table 1: Number of cellular towers and subscribers of each cellular network operator.

Operator	Number of cellular towers	Number of subscribers
1	1,722	1,056,958
2	1,913	489,835
3	675	68,213
4	85	1,641
5	235	1,618

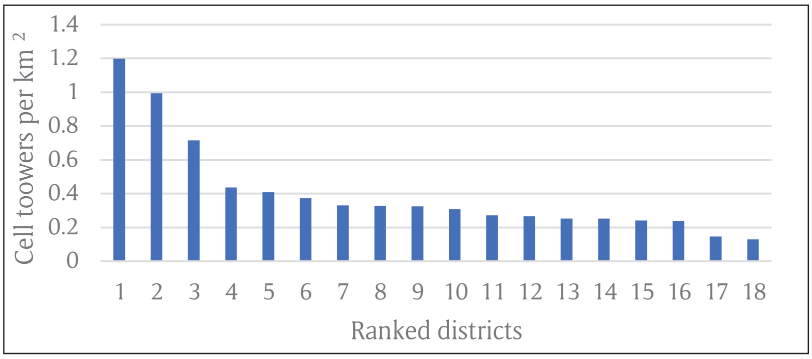
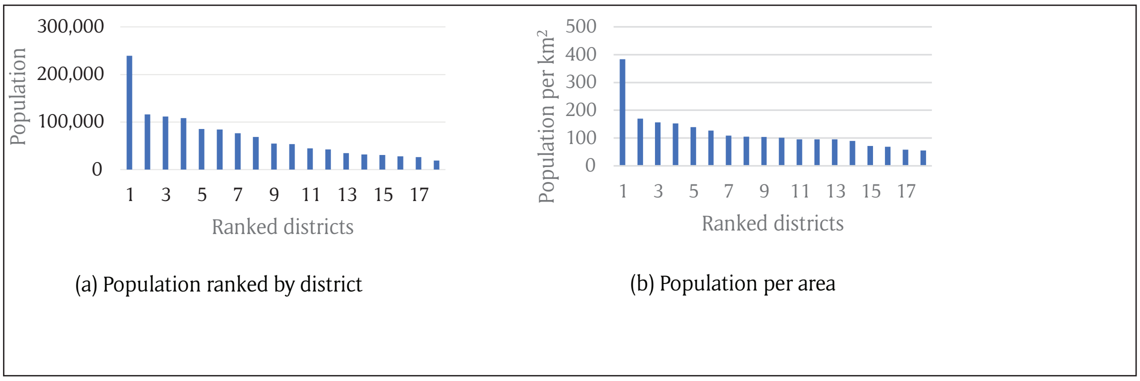


Figure 1: Cellular towers density per district of all operators.



(a) Population ranked by district

(b) Population per area

Figure 2: Census-based population density distribution; total count in each district (a) and population per km^2.

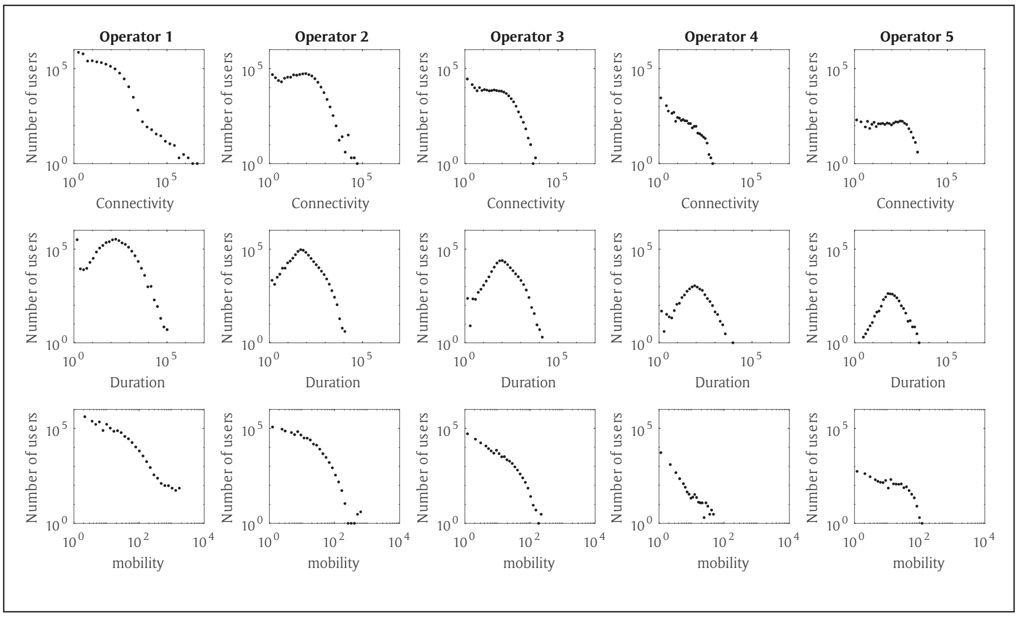


Figure 3: Distributions of user characteristics in each of the five networks.

3.1. Market share

Statistically, CDR data from an operator with a larger market share may better represent the population. Yet data from a smaller market share but with a higher user geographical distribution can conceivably represent the population better than the data from a larger market-share operator. For a given market share, we examined how each operator's data resembles the population.

Similar to **Figure 4**, the correlation between the population density values by census and CDR data from each operator with a different market share was examined. **Figure 5** shows the result of the correlation between each of the five operators against the actual population density by census.

Each operator has a different market share and resembles the population density at a different level as measured by correlation value (R-squared). The observed correlation values (rounded to two decimal places) and the corresponding market share (based on the CDR data) of each operator is listed in **Table 3**. The observed market share in the region of study is similar to other telecom operator market shares in the countries within the south east Asian region, for example, Indonesian (44%, 16%, 15%, 6%, 4%, 3%, 2%, 2%), Thailand (44.3%, 27.4%, 26.2%, 1.78%, 0.18%), Myanmar (66.6%, 20.5%, 13.3%), and Vietnam (75.54%, 22.96%, 1.21%, 0.23%, 0.06%).

The representativeness of the data, which is measured by the correlation value increases with the operator's market share percentage, except for the operators 4 and 5 that have the same market share but different correlation values. Interestingly, the gaps between correlation values are small when compared to the gaps between market share percentages. The correlation values are relatively high across the five operators (0.70–0.88), although the market share percentages vary from as low as 0.10% to as high as 65.31%.

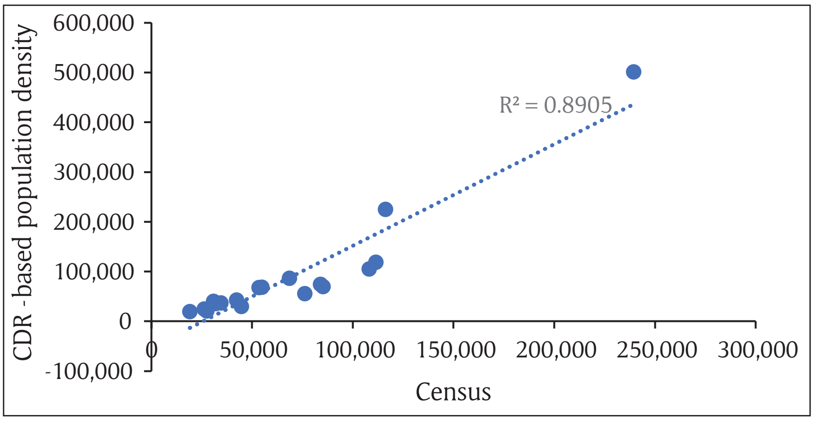


Figure 4: Correlation between the population density by census and the population density derived from the CDR data.

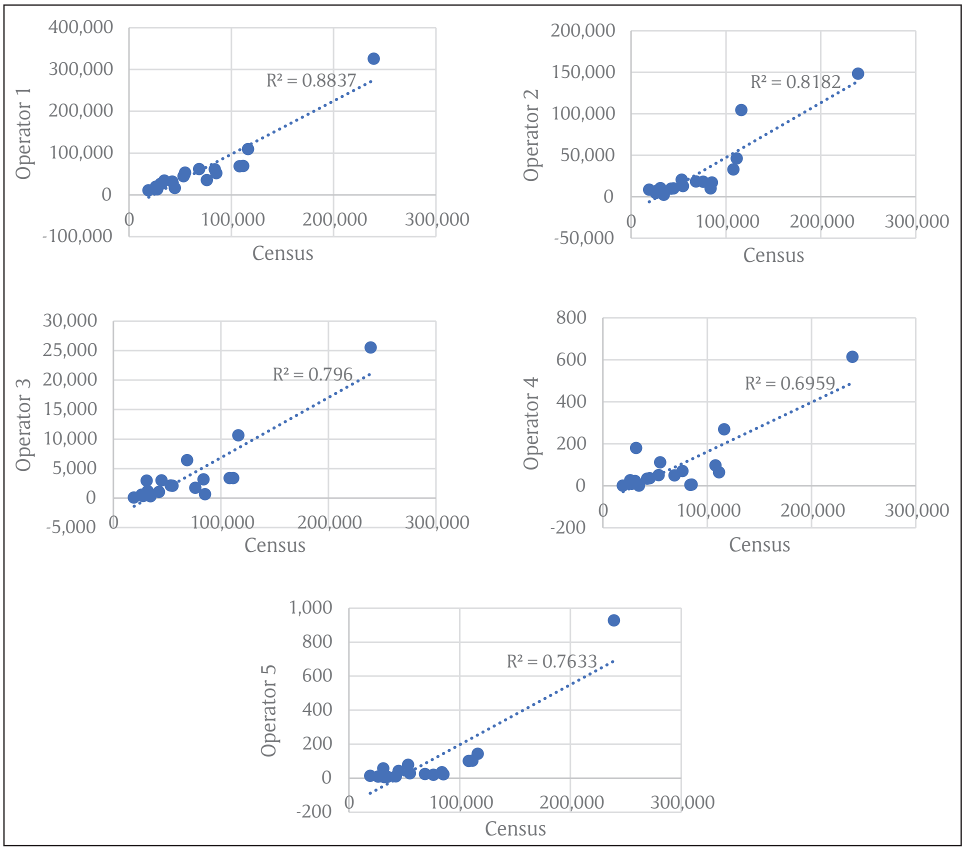


Figure 5: Correlation between the population density by census and the population density derived from the CDR data from five different operators with different market share.

The operator 3 can provide a reasonably high correlation value of 0.80 with a rather low market share of 4.22%. The operator 2 has about seven times larger market share than the operator 3 but the difference in the correlation values is only an increase of 0.02. This suggests that the representativeness of the data does not scale at the same rate with the market share.

3.2. Urban-rural ratio

Portion of urban and rural users can also play an important role in the representativeness of the data to resemble population density distribution. Geographical distribution of users is important. Some operators may dominate in terms of the number of users in some particular areas, while some other operators may attract the users evenly from all areas. The representativeness of their CDR data can therefore be different.

Table 2: Statistical mean and median values of usage characteristics of the users of each network operator.

Operator	Connectivity		Call duration		Mobility	
	Mean	Median	Mean	Median	Mean	Median
1	38.32	4.00	246.32	105.50	6.38	2.00
2	80.58	29.00	83.49	43.56	6.88	4.00
3	44.97	9.00	155.36	84.30	4.80	2.00
4	12.32	2.00	120.46	69.75	1.90	1.00
5	154.86	45.00	113.05	73.33	10.03	5.00

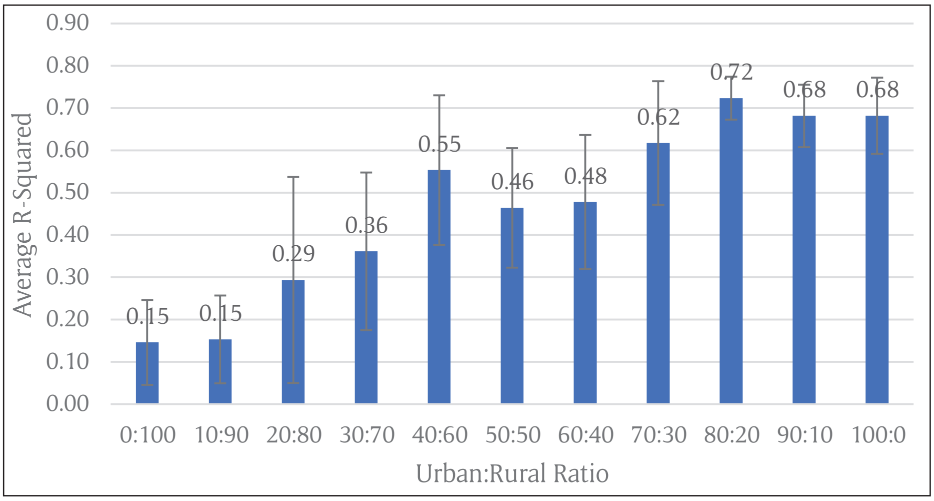


Figure 6: Correlation values between census and CDR-based population densities for different ratios of urban and rural user population.

As the spatial coverage of the cellular towers in each district area generally varies with the area population density, we thus used cellular tower coverage to characterize urban and rural areas. So, we defined the terms "urban area" based on the cellular tower coverage as an area with at least 0.7 cell towers/km^2 and "rural area" as an area with less than 0.7 cell towers/km^2. The threshold of 0.7 cell towers/km^2 was based on the result in **Figure 1** (cell tower density per district) from which we know for a fact that the top three districts are urban areas, and therefore the threshold of 0.7 cell towers per km^2 was chosen to categorize the area. The 18 districts can be divided into 124 sub-districts, which were classified into urban and rural areas according to their district type. This yielded 43 urban and 81 rural sub-districts in total.

Our goal was to examine the representativeness of the data in resembling the population distribution at different ratios of the user population in urban and rural areas. We thus calculated the correlation value (R-squared) between the CDR-based population and the actual population by census for 11 different ratios (percentages of urban and rural population); 0:100, 10:90, 20:80, 30:70, 40:60, 50:50, 60:40, 70:30, 80:20, 90:10, and 0:100. For each ratio, sub-districts were randomly selected from the urban and rural district groups to comprise the ratio, and then the correlation value was calculated. The process was repeated for 10 times for each ratio, so that the result would not be biased towards a single selection. The result is shown in **Figure 6**.

From the result in **Figure 6**, the (average) correlation value appears to rise as the portion of urban population increases. The lowest correlation value is 0.15 for the 0:100 and 10:90 ratios. The highest correlation value is 0.72 (with a low standard deviation of 0.10) at 80:20 ratio. This result suggests that the urban-rural ratio of 80:20 can best represent the population density distribution.

3.3. Gender ratio

Population is a mixture of genders that have different characteristics. Previous studies (Frias-Martinez, Frias-Martinez and Oliver, 2010; Jahani *et al.*, 2017) have shown that male and female mobile phone users statistically have different usage behaviors, such as call initiation, text response time, call duration, number of incoming/outgoing calls, and so on. Gender ratio thus may play an important role in the representativeness of the CDR data.

Table 3: Correlation values (R-squared) against the census population and the market share of each operator.

Operator	Market share (%)	R-squared
1	65.31	0.88
2	30.27	0.82
3	4.22	0.80
4	0.10	0.70
5	0.10	0.76

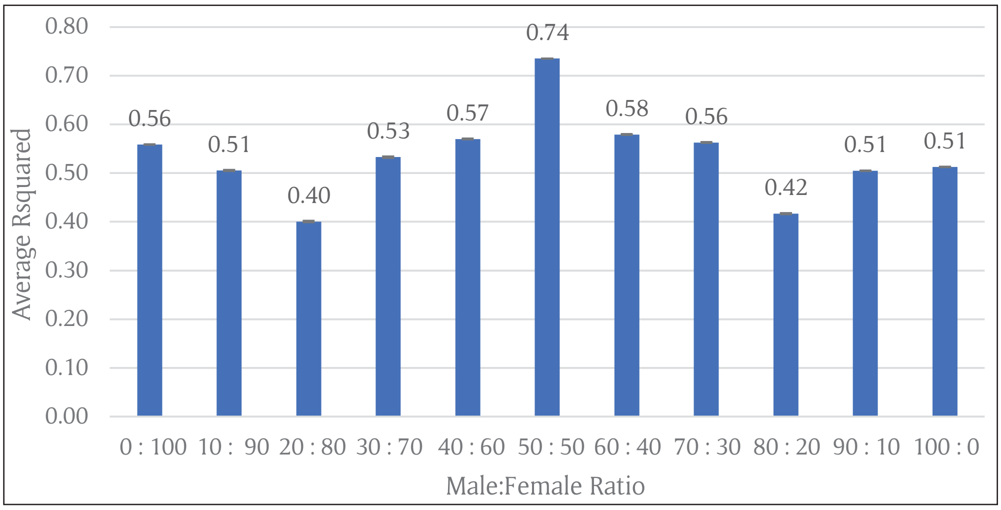


Figure 7: Correlation values between census and CDR-based population densities for different ratios of male and female users.

We first classified each user into male and female groups by using the five criteria from Jahani et al. (Jahani *et al.*, 2017), i.e., night time calls, calls at home, call duration, antenna usage, and call initiation. As a result, there were 780,634 males, 673,447 females, and 164,184 unclassified users (mainly due to inadequate amount of data to satisfy all five criteria). Our goal here was to examine the representativeness of the CDR data given a gender ratio in terms of correlation to the census data, therefore, similar to Section 3.2, we examined the correlation of 11 different gender ratios (percentages of males and females); 0:100, 10:90, 20:80, 30:70, 40:60, 50:50, 60:40, 70:30, 80:20, 90:10, and 0:100. Since the numbers of males and females were different, so for each ratio examination, the total amount of random subjects was 600,000. So that, for example, the 0:100 ratio included a randomly selected 600,000 females and no male, and for the 10:90 ratio, there were 60,000 randomly selected males and 540,000 randomly selected females, and so on. For each ratio examination, this random selection was repeated for 10 times to avoid potential bias towards a single selection. The result is shown in **Figure 7** where the average correlation peaks at the 50:50 ratio (R^2 = 0.74) and hits the lowest value at 20:80 ratio (R^2 = 0.40). The standard deviation value is relatively low across all ratios (the average value was 0.0011). Intuitively, the result suggests that with an equal mixture of male and female users (50:50 ratio), CDR data can best resemble the population density. We would like to note that in the region under study, there is a gender equality i.e., the different behavior, aspirations, and needs of both women and men are considered, valued and favored equally. Their rights, responsibilities, and opportunities do not depend on whether they are male or female. The result observed here somewhat reflects on this fact.

4. Conclusion

Mobile phone data (CDR) has a great advantage over the traditional human behavioral datasets, such as national travel survey, household survey, and traffic count, that are typically collected through questionnaires and interviews, which are expensive and time-consuming, as well as inaccurate as it may be based on recalling past activities, while CDRs provide longitudinal real-life behavioral data. While CDRs have been used to advance human behavior research in various directions with interesting and useful discoveries, those CDR datasets only provide a partial view of the whole population as the data used was limited to the service coverage of some network operators who were the data provider. This study aims at determining the extent to which the CDR data obtained from certain network operators can represent the whole population, by examining the representativeness of the data in forms of the correlation coefficient between the CDR-derived population density and the census data, from three perspectives; operator's market share, urban-rural user population ratio, and user gender ratio. The study reveals that (i) the representativeness of the CDR data does not scale at the same rate with the market share, (ii) the urban-rural user population ratio of 80:20 can best represent the population density distribution, and (iii) an equal mixture of male and female user population can best resemble the actual population density.

The significance of this study is the analytical results obtained from a complete view of all individual mobile phone users (1,618,265 users) in a study area (11,600 km^2). This study is the first investigation on the representativeness of the CDR data in terms of market share, urban-rural ratio, and gender ratio, as no previous study has analyzed a complete view of all mobile phone user data. Nonetheless, there are a number of limitations to this study. The first of these is the spatial coverage of our data that only provides a view of one city. Results from multiple area analysis could potentially be more generative, however a large number of users in this study may have compensated and generalized the results reasonably to a significant extent. The second limitation is the possibility of users carrying multiple phones, which could replicate some individual statistics that affect the overall population density, yet these users are of a considerable portion (1,618,265–1,256,654 = 361,611 users) which accounts for 28.78% for which it's worth a future study exploring whether these users can be identified and filtered, appropriately. A feasible approach can be utilizing a similarity pattern search with some behavioral features across all users or only those who share the same home cell tower. Another potential limitation is the extent to which the findings are applicable beyond a study area (southeast Asian city). As a city of a developing country in southeast region, the city shares a significant similarities with many cities around the globe, we thus believe that the findings are likely to be applicable to other regions as well. If the characteristics of the mobile phone users in other regions are significantly different from this study's, our findings can still be considered valuable at least as an observation of the representativeness of the CDR data in 2014 in a southeast Asian city.

This study reflects on the trend of big data analytics that helps further extend our understanding of the representativeness of a massive CDR data that can be used to mirror distinct perspectives of human behavior. We believe that our findings offer new knowledge and important information for CDR usage in research, principally from a data science perspective.

Data Accessibility Statement

The dataset used in this study was provided to us by the nation telecommunication commission of a southeast Asian country. A sample of data can be made available on request to other researchers for academic, non-commercial purposes by the authors.

Acknowledgements

The authors gratefully acknowledge the financial support provided by Thamasat University under the TU New Research Scholar, No. 16/2559.

Competing Interests

The authors have no competing interests to declare.

Author Contributions

SP and TH conceived of and designed the study. SP and TH analyzed the data and results. TP processed the data. SP, TP, and TH wrote the manuscript. All authors have read and approved the final manuscript.

References

Aiello, W, Chung, F and **Lu, L.** 2000. 'A random graph model for massive graphs'. In: *Proceedings of the thirty-second annual ACM symposium on Theory of computing – STOC'00*. DOI: https://doi.org/10.1145/335305.335326

Blondel, VD, *et al.* 2012. 'Data for Development: the D4D Challenge on Mobile Phone Data'. *arXiv:1210.0137*, 1–10. Available at: http://arxiv.org/abs/1210.0137.

Blondel, VD, Decuyper, A and **Krings, G.** 2015. 'A survey of results on mobile phone datasets analysis'. *EPJ Data Science*. DOI: https://doi.org/10.1140/epjds/s13688-015-0046-0

Demissie, MG, *et al.* 2016. 'Inferring Passenger Travel Demand to Improve Urban Mobility in Developing Countries Using Cell Phone Data: A Case Study of Senegal'. *IEEE Transactions on Intelligent Transportation Systems*, 17(9). DOI: https://doi.org/10.1109/TITS.2016.2521830

Frias-Martinez, V, Frias-Martinez, E and **Oliver, N.** 2010. 'A Gender-Centric Analysis of Calling Behavior in a Developing Economy Using Call Detail Records.'. *... Intelligence for Development*.

Jahani, E, *et al*. 2017. 'Improving official statistics in emerging markets using machine learning and mobile phone data'. *EPJ Data Science*. DOI: https://doi.org/10.1140/epjds/s13688-017-0099-3

Kiukkonen, N, *et al*. 2010. 'Towards rich mobile phone datasets: Lausanne data collection campaign'. *Proceedings ACM International Conference on Pervasive Services (ICPS)*.

Laurila, JK, *et al*. 2012. 'The mobile data challenge: Big data for mobile computing research'. *Proceedings of the Workshop on the Nokia Mobile Data Challenge, in Conjunction with the 10th International Conference on Pervasive Computing*. DOI: https://doi.org/10.1016/j.pmcj.2013.07.014

Miritello, G, Moro, E and **Lara, R.** 2011. 'Dynamical strength of social ties in information spreading'. *Physical Review E – Statistical, Nonlinear, and Soft Matter Physics*. DOI: https://doi.org/10.1103/PhysRevE.83.045102

Montjoye, YDe and **Smoreda, Z.** 2014. 'D4D-Senegal: The Second Mobile Phone Data for Development Challenge'. *arXiv*.

Navarro, H, *et al*. 2017. 'Temporal patterns behind the strength of persistent ties'. *EPJ Data Science*. DOI: https://doi.org/10.1140/epjds/s13688-017-0127-3

Onnela, J-P, *et al*. 2006. 'Structure and tie strengths in mobile communication networks'. *Proceedings of the National Academy of Sciences (PNAS)*, 104(18): 7332–7336. DOI: https://doi.org/10.1073/pnas.0610245104

Phithakkitnukoon, S, *et al*. 2017. 'Inferring social influence in transport mode choice using mobile phone data'. *EPJ Data Science*, 6(1). DOI: https://doi.org/10.1140/epjds/s13688-017-0108-6

Phithakkitnukoon, S and **Dantu, R.** 2011. 'Mobile social group sizes and scaling ratio'. *AI and Society*, 26(1). DOI: https://doi.org/10.1007/s00146-009-0230-5

Phithakkitnukoon, S, Smoreda, Z and **Olivier, P.** 2012. 'Socio-geography of human mobility: A study using longitudinal mobile phone data'. *PLoS ONE*, 7(6). DOI: https://doi.org/10.1371/journal.pone.0039253

Saramäki, J and **Moro, E.** 2015. 'From seconds to months: an overview of multi-scale dynamics of mobile telephone calls'. *European Physical Journal B*. DOI: https://doi.org/10.1140/epjb/e2015-60106-6

Song, C, *et al*. 2010. 'Limits of predictability in human mobility'. *Science*. DOI: https://doi.org/10.1126/science.1177170

Vaz De Melo, POS, *et al*. 2010. 'Surprising patterns for the call duration distribution of mobile phone users'. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. DOI: https://doi.org/10.1007/978-3-642-15939-8_23

How to cite this article: Horanont, T, Phiboonbanakit, T and Phithakkitnukoon, S. 2018. Resembling Population Density Distribution with Massive Mobile Phone Data. *Data Science Journal,* 17: 24, pp. 1–9. DOI: https://doi.org/10.5334/dsj-2018-024

Submitted: 12 June 2018 **Accepted:** 17 September 2018 **Published:** 03 October 2018

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS