Accuracy and Bias in the Identification of Internal Migrants using Cell Phone Data

Lingzi Hong
iSchool
University of Maryland, College Park
lzhong@umd.edu

Jiahui Wu
iSchool
University of Maryland, College Park
jeffwu@umd.edu

Enrique Frias-Martinez
Telefonica Research
Madrid, Spain
e.friasmartinez@telefonica.com

Andrés Villarreal
Department of Sociology
University of Maryland, College Park
avilla@umd.edu

Vanessa Frias-Martinez
iSchool
University of Maryland, College Park
vfrias@umd.edu

***Abstract*—In recent years, there has been an increase in the volume of human internal migration in many countries, mostly due to economic crises, political instability and various types of natural disasters. Internal migrations have been studied lately using a variety of social network data or cell phone traces. All these studies, need to first identify the internal migrants before carrying out any specific analysis. In this paper, we present a comparative analysis of several techniques used in the literature for migrant identification using cell phone traces. Our analyses will focus on the accuracy of the different methods with respect to a ground truth extracted from census data; and on the biases that each method introduces, analyzed for different types of urban and rural internal migrations.**

I. Introduction

Internal migration refers to the migration of individuals from one region to another within the same geopolitical entity, typically within the same country [1], [2]. Considerable attention has been given to the study of migration using ubiquitous spatio-temporal data generated in a passive manner, for example cell phone records or social media data. Such type of rich data enables to carry out large-scale analyses of migration flows as well as the micro-level view to analyze individual behaviors [3]–[5].

At its core, research on internal migration behaviors first requires to identify the internal migrants in the dataset. Methods to identify internal migrants are based on determining home location changes *i.e.,* a person that was living in a location, changes her home permanently or temporarily within the same country. Several methods have been developed to identify home location using spatio-temporal data [6]–[8]; and some of these methods have been applied to identify volumes of internal migrants [3], [9]. However, no work has looked into analyzing the impact that the choice of a home location algorithm might have in the identification of internal migrants, both in terms of accuracy and biases.

In this paper, we focus on the use of cell phone mobility data as a proxy for internal migrations. We present four different state-of-the art methods used to detect the home location of individuals in a cell phone dataset, and we use them to identify potential internal migrants. We evaluate the accuracy of each method by comparing the computed migration matrix with official census data. Additionally, we look into the biases introduced by each method, measured as the differences in accuracy when different types of urban and rural flows are approximated via cell phone data.

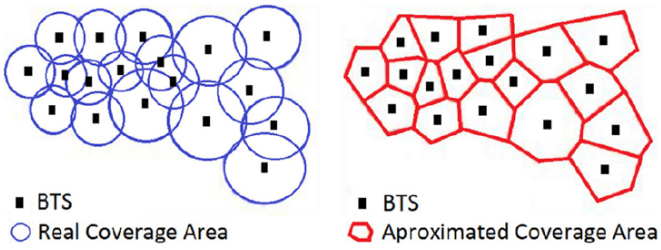

Fig. 1. (left) Original coverage areas per BTS, (right) approximation of coverage areas using Voronoi tesselation.

II. Related Work

Ubiquitous data generated in a passive manner has been used to model internal and international migrations [4], [5], [10]. For example, Zagheni *et al.* used email service logs to identify international migration rates [5]; Weber *et al.* used anonymized log data from Yahoo! services users to generate short-term and medium-term migration flows across countries [10]; and Zagheni *et al.* used Twitter data to model both international and internal migration patterns [4]. On the other hand, researchers have used cell phone data to model mobility patterns and evaluate migrations. For example, Blumenstock *et al.* proposed a macro-level method that used cell phone metadata to identify migrants and quantify volumes and directionality of internal migrations in Rwanda [3], while Isaacman *et al.* did a similar analysis to evaluate internal migrations due to droughts in Colombia [9]. In this paper, we propose a comparison of the state-of-the-art approaches to identify internal migrants using cell phone data, and evaluate the methods in terms of accuracy and bias.

III. Data Description

We use two main data sources, cell phone traces to characterize internal migrants, and census information to evaluate accuracy and biases.

A. Cell Phone Traces

Cell phone networks are built using a set of base transceiver stations (BTS) that are responsible for communicating cell phone devices within the network. Each BTS or cellular tower is identified by the latitude and longitude of its geographical location.The coverage area ranges from less than 1km^2 in dense urban areas to more than 4km^2 in rural areas. For simplicity, it is common in the literature to assume that the cell of each BTS is a 2-dimensional non-overlapping polygon, which is typically approximated using Voronoi diagrams. Figure 1(left) presents a set of BTS with the original coverage for each cell and (right) the simulated coverage obtained using Voronoi diagrams.

Call Detail Records (CDRs) are generated by telecommunication companies for billing purposes. CDRs are created whenever any type of cell phone connected to the network makes or receives a phone call or uses a service (e.g., SMS, MMS). In the process, the BTS details are logged, which gives an indication of the geographical position of the cell phone at the time of the call. Note that no information about the exact position of a user in a cell is known, *i.e.,* we do not have a GPS-type location of the phone within the coverage of a BTS.

In this paper, we use an eight-month aggregated and anonymized CDR dataset for the whole country of Mexico (October 2009 to May 2010). To preserve privacy, original records are encrypted *i.e.,* the researchers did not have access to any actual individual cell phone numbers, but rather to anonymous IDs; and all the information presented in the paper is aggregated *i.e.,* no individual features are reported. From all the information contained in a CDR, our study considers the encrypted originating number, the encrypted destination number, the time and date of the interaction, and the BTS that the cell phone was connected to when the call was placed. No contract or demographic data was considered or available for this study and none of the authors of this paper collaborated in the extraction and the encryption of the original data. The dataset contains 7 billion records from 39K cellular towers that cover the whole country. We eliminate from the dataset all individual IDs (and their corresponding CDR data) whose activity can be assumed to correspond to a machine and not an individual using the approach in [11]. This approach, which uses average measures of reciprocal cell phone contacts and frequency to eliminate anomalous accounts, was applied over the dataset leaving a final number of 48M unique users.

B. Census Data

We have obtained from the Mexican Statistical Institute (INEGI) a migration matrix at the municipality level. The migration matrix is based on the ENADID 2010 survey (National Survey of Demographic Dynamics) [12]. It records the number of people migrating from one municipality to another from 2005 to 2010 across the whole country. To further characterize the municipalities in our analysis, we differentiate between urban and rural using the definition by the OECD, which considers a municipality in Mexico to be rural if the population is below $100,000$ people [13].

IV. Identification of Migrants

A. Definition

We define as internal migrants the individuals in our dataset who have a consistent home location for at least three months and then move to another place, where they also stay for at least three months. With this definition, the internal migrants we identify can be either long-term or short-term (circular) migrants depending on whether they go back or not to their original location after our data collection period finishes [14]. Since the census data we use in ouranalyses measures the internal migration flow at the municipality level, we use one of the four home location methods described in the next section to assign a monthly home to each individual in the dataset. Once all monthly homes have been identified, we select the subset of individuals that can be considered as internal migrants - as explained above- and evaluate accuracy and bias for each home location method.

B. Home Location Methods

In this section we present the four state-of-the-art methods we have considered to identify monthly homes and, consequently, identify internal migrants:

Method 1. Home location algorithms are part of a larger group of algorithms used to identify important places using spatio-temporal mobility information. The main idea behind these algorithms is to define time windows for when people are at home, work or other, and to identify the location of these important places [6], [15].

Method 1 defines as home location the most visited municipality at night, between 6pm to 6am. For each BTS tower visited at night, we extract its municipality and add the number of times each municipality has been visited in each month in the dataset. A municipality is assigned as a home location for a given month if an individual stays at that location for at least 70% of the nights. Otherwise, no home location is assigned. Using this method over the Mexico dataset, we were able to identify $120,627$ internal migrants.

Method 2. This method is similar to Method 1 but changing the time window to a shorter period of night time, from 10pm to 6am. The main motivation to do this is that a tighter temporal range might help to reduce the noise in the set of BTS towers considered as potential home locations, since earlier times might also incorporate locations beyond home, such as work or other. However, since individuals typically have reduced cell phone activity at night (specially in emerging economies like Mexico [16]), it will be harder to assign a home location to many individuals, thus reducing the number of internal migrants identified. In fact, after applying this method to our Mexico dataset, we identify a total of $109,199$ migrants.

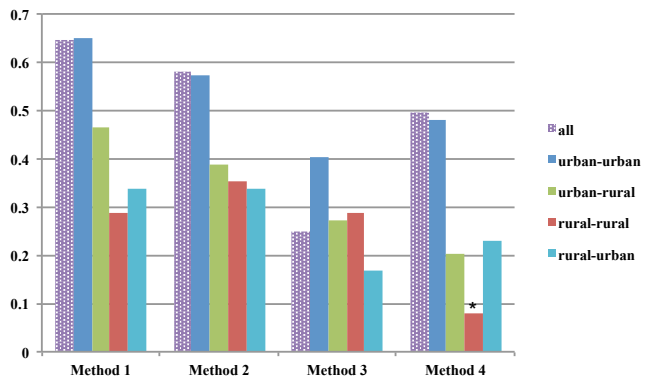

Fig. 2. Correlation coefficients between census- and CDR-based matrices for each method across all pairs of municipalities, and for urban-urban, urban-rural, rural-urban and rural-rural flows. Significance at $p < 0.01$ except for the one with '*'.

Method 3. Time ranges are good when societies tend to have similar day and night time behaviors *i.e.,* people go to similar times to work, back to home, etc. However, that might not be the case in more informal communities. Method 3 explores an approach described by Blumenstock *et al.* where home location is identified as the physical space where an individual spends the majority of her time [3]. Therefore, we identify home location as the municipality where the center of gravity across all visited cellular towers is located, weighted by the cell phone activity in each cell. Since the center of gravity can be heavily affected by long-distance activities, such as traveling, we compute one daily home location; and identify a municipality as monthly home if it has been identified as such at least 70% of the days in a month. After applying this method to our CDR dataset from Mexico, we identified a total of $40,892$ internal migrants.

Method 4. Method 4 represents a combination of Methods 2 and 3 *i.e.,* we combine the temporal window approach (10pm-6am) with the center of gravity approach. The assumption for this approach is that the activities during night hours tend to be closer to one's home location. Specifically, we identify as daily home location the municipality where the night time center of gravity is geographically located. The monthly home location is identified as the location where the individual has been observed at least 70% of the days. This method identifies a total of $1,992$ migrants. It is important to clarify that both Methods 3 and 4 compute a daily home location. If there is no information to identify a home during a week day or weekend day, the last identified position for a week day or weekend is assigned. Individuals must have at least one home location in each week to be considered as potential internal migrants.

V. Evaluation

Each one of the methods used to identify internal migrants inherently introduces a bias caused by the information requirements needed to apply the algorithm. In fact, we have shown that each method identifies a different number of potential internal migrants: while time interval approaches identify larger numbers, ROG-based measures appear to be more conservative and identify smaller numbers of potential internal migrants. This algorithmic bias adds to the already biased CDR dataset available. As a matter of fact, the literature has reported how developing economies tend to have less cell phone-based interactions, both due to economic factors and to the general characteristics of the types of contract available. For example, Rubio *et al.* compared phone usage in a developed and developing economy showing that the developed economy tends to use cell phones much more than the developing economy: almost 75% of individuals in the advanced economy make/receive on average at least 2 calls per day, while only 27% of the population in the developing economy has the same number of calls per day [16].

In order to evaluate the accuracy of the methods proposed and their algorithmic and data biases, we use the internal migrants identified by each approach to compute the CDR-based migration flow matrix, and compare it with the census-based matrix via correlation analysis. To evaluate accuracy, we compare the total internal migration flow between each pair of municipalities *i.e.,* the value for each pair (origin, destination) in the matrix. On the other hand, to evaluate biases we analyze accuracy for four different types of flows: urban to urban (U2U), urban to rural (U2R), rural to urban (R2U) and rural to rural (R2R) migration flows. Figure 2 shows the results.

In a second analysis to further evaluate accuracy, we compare, for each method, the total outbound internal migration flow between the CDR-based and census-based migration matrices *i.e.,* the matrix values of (origin, all destinations) for each municipality, and the total inbound flow from other municipalities *i.e.,* the matrix values of (destination, all origins) for each municipality. Similarly to the previous analysis, biases are identified via a more detailed urban and rural inbound and outbound flow analysis. Figure 3 shows the results.

As can be seen in Figure 2, and focusing on all pairs of municipalities (*all*), methods exclusively based on temporal ranges (Methods 1 and 2) outperform in accuracy methods that use center of gravity measures (Methods 3 and 4) with Pearson's correlation coefficients of up to 0.6457 versus coefficients below 0.4944 (at $p < 0.01$). Although Method 4 is a hybrid method, partially based on a temporal window, the use of the center of gravity appears to negatively impact on its accuracy. However, its correlation coefficient is better than Method 3 which exclusively uses the center of gravity measure. Looking into specific types of urban and rural flows, we observe that methods exclusively based on temporal ranges (Methods 1 and 2) seem to capture better any type of U2U, U2R, R2U or R2R flows with all correlation coefficients being higher than the center of gravity-based methods (Methods 3 and 4) (at $p < 0.01$). It is also important to highlight that internal migration flows that have an urban origin or destination are better detected, across methods, than those with a rural origin or destination municipality *i.e.,* the methods proposed in the literature probably have an inherent bias against rural populations. In fact, the accuracy in the identification of urban flows is probably responsible for the high correlation values of

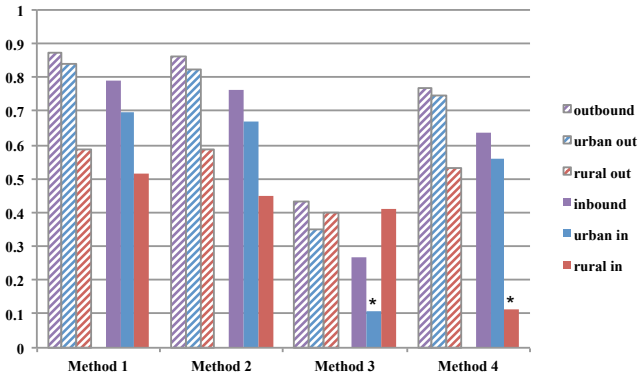

Fig. 3. Correlation coefficients between census- and CDR-based matrices for each method for all inbound and outbound flows; and for urban and rural in- and outbound flows. Significance at $p < 0.01$ except for those with '*'.

the overall internal migration. Methods 3 and 4 have a more reduced bias against rural migrations than Methods 1 and 2. However, this is possibly due to the fact that these methods generally have lower accuracies across the table. The fact that urban migrations are better modeled is probably caused by the phone usage characteristics of developing economies previously introduced. Cell phones tend to be used more in urban areas and as a result more home locations can be identified and migration flows can be better characterized.

Figure 3 shows the correlation coefficients between census-based and CDR-based migration matrices when outbound or inbound flows are considered. As expected, these correlations are better than when all individual pairs are considered since this analysis correlates values that aggregate data from all incoming or all outgoing municipalities. As in the previous analysis, the correlations are better for temporal range methods (Methods 1 and 2) than for center of gravity-based methods (Methods 3 and 4). In terms of biases, we can observe that rural inbound and outbound municipalities appear to have the worse correlation coefficients, indicating again that current approaches to migrant detection appear to be biased against rural population. However, it is important to indicate that the biases are smaller than the ones identified in the previous analysis: while rural versus urban coefficients were up to 50% smaller, here the difference is of at most 30%.

Overall, these results show that: (i) methods that use temporal ranges to identify internal migrants using CDR data, perform better than center of gravity-based methods; (ii) longer temporal ranges show better accuracy than shorter ranges; (iv) current methods show biases against rural population; and that (v) those biases decrease when total outbound or inbound flows are considered, possibly due to decreases in accuracy.

VI. Conclusions

The identification of internal migrants is key for any kind of study on migration behavior. The use of pervasively generated datasets such as cell phone traces, has open the door to automatically identify migrants and migration flows. Nevertheless the characteristics of how these datasets are generated combined with the implementation of algorithms to identify migration flows imply that some bias is introduced in the process. In this paper we have presented four state-of-the-art approaches used for migrant identification and we have measured their accuracy and biases using as ground truth the migration matrix computed from official census data. Our results indicate that the method that uses temporal ranges (from 6pm to 6am) outperforms the other solutions, while all methods show biases against the rural population.

This result could potentially be valid to the great majority of developing economies as they tend to show similar characteristics, namely: (1) high penetration of cell phones; (2) uneven distribution of the population between urban an rural areas and (3) less cell phone activity when compared to developed economies. In the future we plan to run this study with other datasets from other developing economies to measure to which extent these results can be generalized.

References

[1] G. Hugo, "What we know about circular migration and enhanced mobility," *Migration Policy Institute*, vol. 7, 2013.
[2] D. Akeju, "Africa, internal migration," *The Encyclopedia of Human Migration*, 2013.
[3] J. E. Blumenstock, "Inferring patterns of internal migration from mobile phone call records: evidence from rwanda," *Information Technology for Development*, vol. 18, no. 2, pp. 107–125, 2012.
[4] E. Zagheni, V. R. K. Garimella, I. Weber *et al.*, "Inferring international and internal migration patterns from twitter data," in *WWW*. ACM, 2014, pp. 439–444.
[5] E. Zagheni and I. Weber, "You are where you e-mail: using e-mail data to estimate international migration rates," in *ACM WebSci*. ACM, 2012, pp. 348–351.
[6] S. Isaacman, R. Becker, R. Caceres, S. Kobourov, J. Martonosi, M.and Rowland, and A. Varshavsky, "Identifying important places in peoples lives from cellular network data," in *Int. Conf. on Pervasive Computing*, 2011, pp. 133–151.
[7] R. Ahas, S. Silm, O. Järv, E. Saluveer, and M. Tiru, "Using mobile positioning data to model locations meaningful to users of mobile phones," *Journal of urban technology*, vol. 17, no. 1, pp. 3–27, 2010.
[8] R. Montoliu and D. Gatica-Perez, "Discovering human places of interest from multimodal mobile phone data," in *International Conference on Mobile and Ubiquitous Multimedia*. ACM, 2010.
[9] S. Isaacman, V. Frias-Martinez, and E. Frias-Martinez, "Modeling human migration patterns during drought conditions in la guajira, colombia," in *ACM Computing and Sustainable Societies*, 2018.
[10] I. Weber, E. Zagheni *et al.*, "Studying inter-national mobility through ip geolocation," in *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM, 2013, pp. 265–274.
[11] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási, "Structure and tie strengths in mobile communication networks," *Proc. Nat. Academy of Sciences*, vol. 104, no. 18, pp. 7332–7336, 2007.
[12] M. S. I. INEGI, "National survey of demographic dynamics 2014," http://www.beta.inegi.org.mx/proyectos/enchogares/especiales/enadid/2014, [Accessed: 2017-06-15].
[13] O. O. for Economic Co-operation and Development, *Rural Policy Reviews: Mexico*, 2007.
[14] R. D. Bedford *et al.*, *New Hebridean Mobility: a study of circular migration*. Canberra, ACT: Dept. of Human Geography, Research School of Pacific Studies, The Australian National University., 2017.
[15] V. Frias-Martinez, J. Virseda, A. Rubio, and E. Frias-Martinez, "Towards large scale technology impact analyses: Automatic residential localization from mobile phone-call data," in *International Conference on Technologies and Development*. ACM, 2010, p. 11.
[16] A. Rubio, V. Frias-Martinez, E. Frias-Martinez, and N. Oliver, "Human mobility in advanced and developing economies: A comparative analysis." in *AAAI Spring Symposium*, 2010.