IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 6, 2019, accepted November 3, 2019, date of publication November 11, 2019,
date of current version November 21, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2952911

Inferring and Modeling Migration Flows Using Mobile Phone Network Data

SORANAN HANKAEW[1], SANTI PHITHAKKITNUKOON[1], MERKEBE GETACHEW DEMISSIE[2], LINA KATTAN[2], ZBIGNIEW SMOREDA[3], AND CARLO RATTI[4]

[1]Department of Computer Engineering, Chiang Mai University, Chiang Mai 50200, Thailand
[2]Department of Civil Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada
[3]Sociology and Economics of Networks and Services Department, Orange Labs, 92320 Châtillon, France
[4]SENSEable City Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Corresponding author: Santi Phithakkitnukoon (santi@eng.cmu.ac.th)

This work was supported in part by the Thailand Research Fund under Grant RSA6180014, and in part by the Eyes High Postdoctoral Fellowship at the University of Calgary.

ABSTRACT Estimating migration flows and forecasting future trends is important, both to understand the causes and effects of migration and to implement policies directed at supplying particular services. Over the years, less research has been done on modeling migration flows than the efforts allocated to modeling other flow types, for instance, commute. Limited data availability has been one of the major impediments for empirical analyses and for theoretical advances in the modeling of migration flows. As a migration trip takes place much less frequent compared to the commute, it requires a longitudinal set of data for the analysis. This study makes use a massive mobile phone network data to infer migration trips and their distribution. Insightful characteristics of the inferred migration trips are revealed, such as intra/inter-district migration flows, migration distance distribution, and origin-destination (O-D) movements. For migration trip distribution modelling, log-linear model, traditional gravity model, and recently introduced radiation model were examined with different approaches taken in defining parameters for each model. As the result, the gravity and log-linear models with a direct distance (displacement) used as its travel cost and district centroids used as the reference points perform best among the other alternative models. A radiation model that considers district population performs best among the radiation models, but worse than that of the gravity and log-linear models.

INDEX TERMS Migration flows, trip distribution modeling, mobile phone network data, gravity model, log-linear model, radiation model.

I. INTRODUCTION

Movement of human beings both individuals as well as groups over short and long distances has long been studied. It is essential to understand and to be able to model the movement to effectively predict the human mobility. The way of living of human beings has always been inextricably linked with their movement. Earlier movements were primary influenced by factors such as climate change and inhospitable landscapes, while socio-economic factors such as employment, living condition, and food are mainly driving the modern movements.

Human mobility over short distances are mostly regular trips [1] such as daily commute to work and shopping that can be home-based or non-home-based, while long distance trips are largely associated with temporary or permanent change of the place of residence. The permanent change of residence or *migration* is a type of long distance trips that are influenced by socio-political, economic, and ecological factors [2] such as occupation opportunities and family. In addition to migration, long distance trips are made for other purposes, such as business, pleasure; and particularly include trips made on motorways, train and airplane. Long distance trips account for a much smaller portion, less regular, and are not well represented in most regional as well as national travel demand models. However, it is important that these trips are included in the travel demand models. For instance, long distance trips

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Awais Javed.

over 80km only account 2.3% of all trips, but about a third of total vehicle kilometers travelled within Great Britain [3]. Because long-distance trips account for a substantial portion of total kilometers travelled, especially on high capacity routes, the infrastructure needs for long-distance trips are expensive and can take years to build. There is also a substantial and increasing environmental impact associated to long-distance trips [4]. Thus, appropriate planning and operation of these trips is required to reduce their impacts.

Trip distribution modeling is the second step of the four-step transportation planning process widely used for forecasting travel demand, which consists of trip generation, trip distribution, mode choice, and route assignment, designed to respond to questions such as how many trips will be generated?, where will these trips take place?, what is the travel mode used in each trip?, and what is the route choice of each trip, respectively [5]. In this study, we investigate long distance trips, particularly interested in migration trip distribution modeling, which aims at examining and essentially describing the distribution of this kind of long distance trips by statistical models. Modeling long distance trips is more challenging than the commuting trips because of its irregularity nature – as such, the collected information regarding this kind of trips is limited.

Travel demand and other human behavioral modeling are generally derived from the observatory data of real behavior using revealed preference survey. In some cases, stated preference surveys are conducted to analyze passengers' evaluation of transportation services, especially when there are hypothetical transportation service alternatives and new attributes. Most surveys collect information about an individual, their household, and a diary of their journeys on a given day. Traffic count, roadside interview, and questionnaire are usually used in conducting a travel survey. Major travel surveys are conducted typically once a decade due to the high costs and laborious efforts [6]. Although a travel survey can provide detailed mobility information, it can be outdated due to a large time gap between the surveys, and it can also be erroneous due to the inaccurate responses to the travel survey questionnaires which are normally based on recalling some information regarding past journeys.

The vast majority of previous studies on migration flows rely on census and population registers data to infer migration flows. These aggregate data may be appropriate to describe rough patterns of international migration, urban-rural migration and some general directions of migration patterns [7]. However, censuses and population register often fall to capture patterns of temporary and seasonal migration flows, especially for specific years between censuses and for recent trends [8], [9]. Massey and Capoferro [10] highlighted that censuses are biased toward documented citizens and may not track migration flows associated to undocumented or international migrants. Furthermore, migration-specific surveys are time consuming (with the prospect of longitudinal interviews with migrants), and are prohibitively expensive for most planners and government agencies [11]. On the other hand, big data sources can track the movements of a large portion of the population, and provide unprecedented spatial and temporal accuracy [12].

With the recent advances in ICT, sensors such as GPS tracking units have been used increasingly in travel surveys [13]. However, due to the privacy issues and regulations, e.g., the EU general data protection regulation, collecting such data at large scale is difficult and challenging. Recent attempts have produced data that are limited to specific type of tracked individuals, such as university students [14] and customers of a particular service provider where the data was obtained in exchange of some incentives [15]. Privacy concerns largely prevent this type of detailed mobility data to be available and utilized extensively, and hence not easily exploitable for an extensive trip distribution modeling.

Recently, the focus has been shifted towards using opportunistic sensing data produced from various sources that can provide insights regarding the spatial distribution and temporal evolution of movements of people. Opportunistic sensing data is a data that is collected for one purpose but also creates an opportunity for another purpose. Mobile phone's call detail records (CDR) is a type of the opportunistic sensing data where the data is originally collected for billing purposes but it can also be useful for human mobility study. CDRs are the communication and corresponding location records of the mobile users. When a mobile phone user connects to the cellular network by making or receiving a phone call or using internet, the communication (e.g., call duration, timestamp, callers and callee's identifications) and location of connected cellular tower are recorded for billing. With the location records of individual users, CDR has been used in human mobility studies. The use of cellular network data has been explored for the development of large scale mobility sensing since the early 2000s [16].

While CDR has been used to investigate various aspects of transportation issues including large-scale urban sensing [17], [18], traffic parameter estimation [19]–[21], O-D flow estimation [22]–[24], and land use inference [25], [26], efforts to apply CDR data to infer migration trips and modeling migration flows is still overlooked. This study improves on previous studies of the same topic by using CDR data to advance our understanding and modeling of migration flow. Thus, this study makes the following distinct contributions: (i) development of an heuristic based approach to infer migration flows using CDR data; and (ii) development of trip distribution models for distributing migration flows in a country scale. Our objective is to draw actionable insights from country-wide migration flows. Transport planners can use these insights to establish better travel demand planning strategies. For instance, the inferred migration flows can be integrated in the regional and provincial models as long-distance trips; and results from our analysis can be used as indicator of demographic changes, and general directions of internal migration patterns.

This paper is structured as follows. The next section (Sect. II) is dedicated to the discussion of the methodological

FIGURE 1. Analysis process.

framework, which includes data description, subject selection, migration flow inference, and migration flow modeling in which migration trip expansion, trip distribution models, and generalized cost are discussed. In the following section (Sect. III), estimation results of three trip distribution models with several considered approaches for the models' parameters are presented. The paper is then concluded in the last section (Sect. IV) with a summary, limitations, and future directions.

II. METHODOLOGY

Our analysis process started with a mobile phone network dataset (CDR) from which a set of subjects was extracted. For each subject, a residence location was identified based on the most frequently used cell tower locations during the nighttime. Migration was then inferred based on the change of the residential locations (homes). A set of statistical models was examined for describing the migration trip distribution. Our analysis process is summarized in Figure 1.

A. DATA DESCRIPTION

In this study, we used anonymized CDR data collected from 1,891,928 mobile phone users (which accounts for approximately 18% of the population) in Portugal over the course of 14 months (April 2, 2006 – June 30, 2007 where half of September and entire October 2006 records are missing). The data was provided to us from one of the largest telecom operators in Portugal. Each record includes the caller ID, callee ID, caller's connected cell tower ID, callee's connected cell tower ID, duration of the call, and timestamp. Each time the mobile phone user makes or receives a call, i.e., connecting to a cell tower, the nearest cell tower location is recorded.

To safeguard personal privacy, individual phone numbers were anonymized by the operator before leaving their storage facilities, and were identified with a security ID (hash code). The dataset does not contain information relating to text messages (SMS) or data usage (Internet). There is a total of 6,358 cell towers, each on average serves an area of 14 km2, which reduces to 0.13 km2 in urban areas such as Lisbon and Porto. Only cell towers located within the Continental Portugal were considered in this study, not including the autonomous regions of Portugal i.e., Azores and Madeira, which are islands.

The data includes over 500 million records (cellular network connections) over the 14 months. To have a sense of how connectivity is distributed over time, Figure 2 shows the average number of records monthly, daily, weekly, and hourly. Low connectivity in September is mainly due to the missing records. Overall, the connectivity is intuitive. High connectivity is observed in the summer period (May – Aug) and holiday season (Nov – Jan). Connectivity rises from

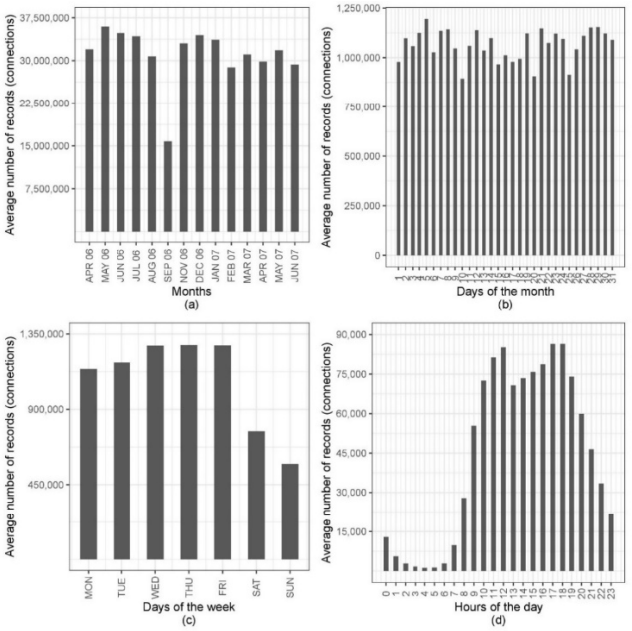


FIGURE 2. Average number of records (cellular network connections); (a) over 14 months (half of September data are missing), (b) days of the month, (c) days of the week, (d) hours of the day.

Monday to Friday and drops during the weekend. Hourly, the connectivity peaks about noon and 5PM – 6PM in the evening.

B. SUBJECTS

To ensure fine-grained mobility information for our analysis, we selected the mobile phone users who had at least five connections in each of the 14 months, which yielded 538,394 users. As we were interested in migration trip which is the change of residence, so we initially needed to identify the location of residence for each subject and determine if there was a change.

For each of the 14 months, a residence was identified using the same approach as Phithakkitnukoon *et al.* [23] i.e., assigning the most frequently used cell tower location during the night hours (10PM – 7AM) as the approximate location of residence. Consequently, only subjects with night hour connectivity in every month were considered further in our analysis (i.e., 148,215 subjects). Each of these subjects had a different number of residential locations detected across the 14 months, varying from 1 to 14 different residential locations. Only one (same) residence detected throughout 14 months implies that there was no change of residence, on the other hand, a detection of different locations of residence can imply that there were changes of residence. Figure 3 shows a histogram of the number of residences detected.

There were 27,004 subjects whose residence did not change throughout 14 months. Following [23], these subjects can be used to derive the population density across 18 districts of Portugal. As such, Table 1 and Figure 4 shows a comparison between the CDR-based and census population density

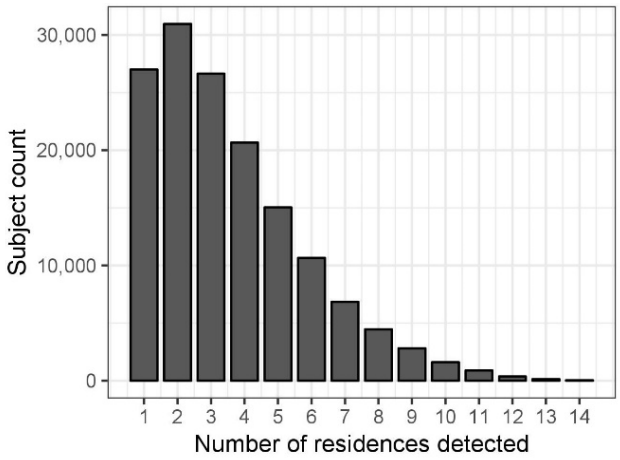


FIGURE 3. Histogram of the number of residences detected.

TABLE 1. Comparison of population density distributions between CDR-based and census data.

District	Census population	CDR-based population
Lisbon	2,250,533	5,185
Porto	1,817,117	6,274
Setubal	851,258	2,640
Braga	848,185	2,377
Aveiro	714,200	1,301
Leiria	470,930	567
Santarem	453,638	1,451
Faro	451,006	393
Coimbra	430,104	781
Viseu	377,653	1,005
Viana do Castelo	244,836	562
Vila Real	206,661	1,122
Castelo Branco	196,264	391
Evora	166,706	727
Guarda	160,939	170

FIGURE 4. Map of Portugal.

distributions, as well as the corresponding map of Portugal with the locations of 18 districts. Statistically, the result is comparable to the census information (in line with [23])

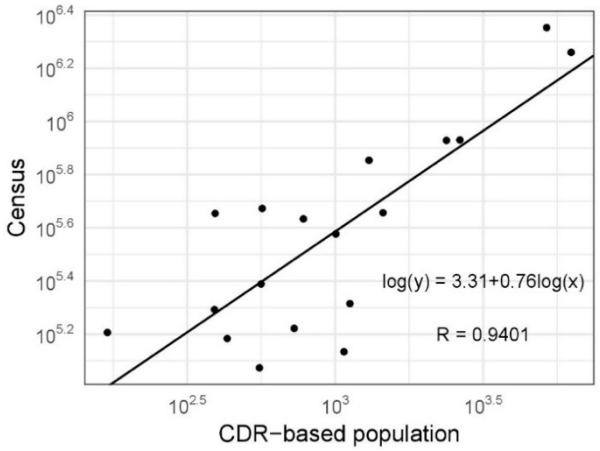


FIGURE 5. Comparison between CDR-based and census population density information.

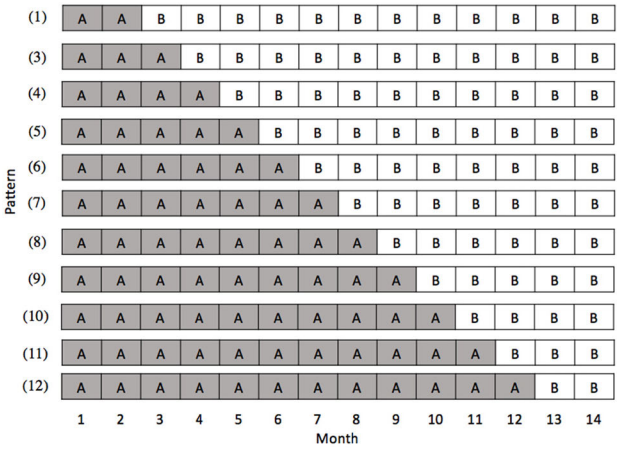


FIGURE 6. Considered patterns of migration in our analysis, where A and B are the previous and new residence.

with the correlation coefficient (R-value) of 0.9401, as shown in Figure 5. Hence, the detected change of residence can be reasonably used for our further analysis of migration.

C. MIGRATION FLOW INFERENCE

We defined migration as a scenario where the subject's residential location changes from one place to another, given that the time at the previous location must be at least two months prior to the change and the time at the new location must be at least two months after the change. There was a possibility that some subjects may migrate more than once. However, in this study, we focused only on the subjects who migrated once (over the period of 14 months). Figure 6 shows all possible patterns of migration where A and B are the previous and new residence. As the result, there were 2,107 subjects who migrated once based on our definition of migration.

Among these 2,107 migrations detected, there were 1,681 intra-district and 426 inter-district migrations. Geographically, Figure 7 shows the amount of intra-district migrations and the flow of inter-district migrations across the country. Temporally, Figure 8 shows the monthly distribution of these

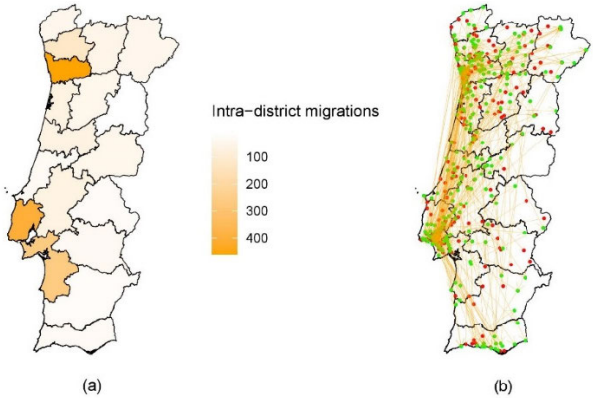


FIGURE 7. Detected (a) intra-district migrations and (b) inter-district migrations, where green and red dots indicate the previous and new residential locations.

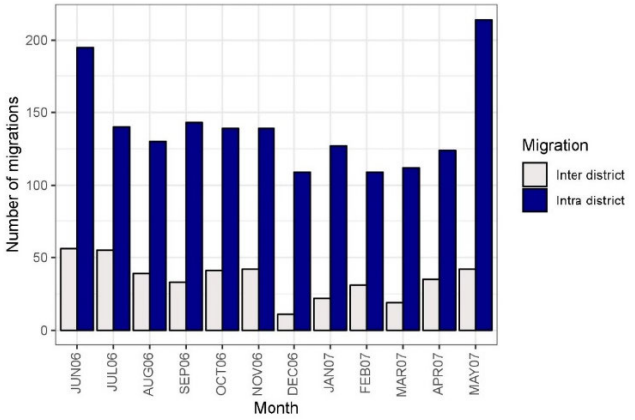


FIGURE 8. Monthly distribution of intra-district and inter-district migrations.

intra-district and inter-district migrations, from which it can be observed that people migrated within district mostly during summertime especially May and June. Likewise, the inter-district migrations pose a high flow in the summer (May to August), but intuitively a low flow during the winter.

Spatially, Figure 9 shows the distribution of inter-district migration flow distance. The distance was measured as a travel distance on the road network using the Google Distance Matrix API.[1] The nearest distance is 55.99 km, migrating from Porto to Braga. The farthest distance is 633.83 km, migrating Faro to Vila Real. The average distance is 184.05 km. The migration distances in Figure 9 can be observed into two groups around the average distance (indicated by a dash line). The below average-distance group is composed of migrations to nearby districts, while the above average-distance group consists of long-distance migrations that mostly are between the big cities, such as Lisbon and Porto. In general, Fig. 9 shows the migration propensity decreases as distance increases.

In the form of Origin-Destination (O-D) matrix that describes the people movement from the 'origin' which is the previous residence to the 'destination' which is the new residence, Figure 10 shows a checkerboard plot ranked by the origins, i.e., amount of originated inter-district migrations (or outflows). An O-D matrix is normally used for transportation system planning. It informs about the volumes of traffic, which represents the transport demand. In this case, it is a transport demand for migration.

Lisbon has the highest outflow followed by Porto, Setubal, and Braga. Top destinations of Lisbon's outflows are Setubal, Santarem, and Porto. Setubal and Santarem are nearby districts to Lisbon, while Porto is another highly populated district located about 320 km north of Lisbon. Top destinations of Porto's outflows are Lisbon, Braga, and Aveiro. Braga and Aveiro are nearby districts to Porto, while Lisbon is relatively much farther district to the south of Porto and Lisbon is the capital and largest city of Portugal.

[1]https://developers.google.com/maps/documentation/distance-matrix/start

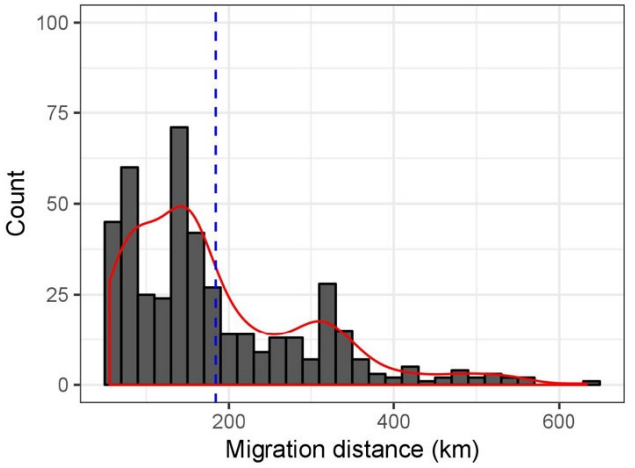


FIGURE 9. Distribution of the migration distances, where a solid red line and dash line indicate the trend and average, respectively.

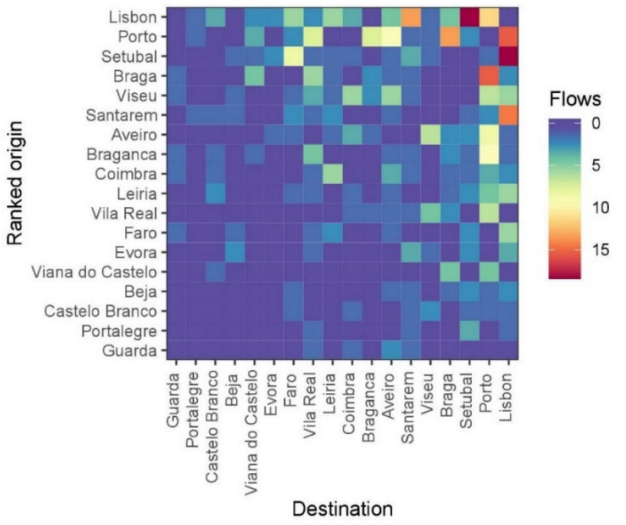


FIGURE 10. O-D flows of inter-district migrations ranked by the origins (previous residence).

In terms of the inflows, Lisbon is the top destination for migration followed by Porto. Setubal, Porto, and Santarem are the top inflows for Lisbon, while Braga, Lisbon, and

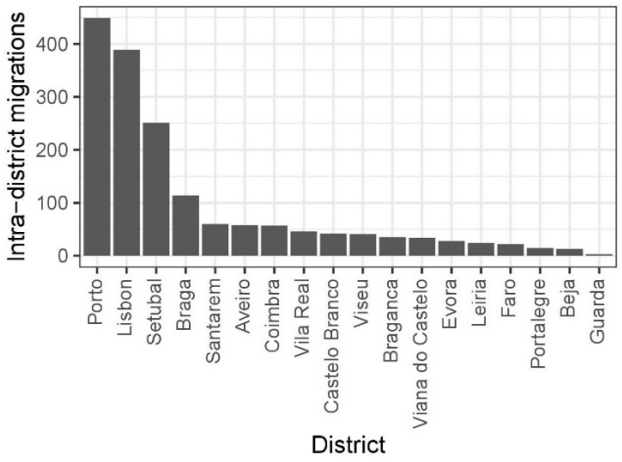


FIGURE 11. Number of intra-district migrations ranked by district.

Braganca are the top inflows for Porto. Braganca is around 200 km northeast from Porto. Locations of the aforementioned municipalities can be seen in Figure 4.

In addition to the migration flows between districts, there were also people who moved within the same district. There were a total 1,681 intra-district migration flows. Figure 11 shows the number of intra-district migration flows ranked by district. Interestingly, Porto has larger flows than Lisbon, which is the capital and largest city. Setubal and Braga are ranked third and fourth, respectively.

D. MIGRATION FLOW MODELING

In the case of travel demand forecasting model, a trip generation model can be used to estimate the total trip productions and attractions of each district. Then, a trip distribution model can be used to distribute trips (from a trip generation model) among destinations. There is no comprehensive travel survey data regarding migration flows in Portugal that can be used to develop a trip generation model. In this study, we've developed a simple procedure to expand the sample migration O-D flows to describe the migration flow behavior of the total population. Then, the total permanent migration trips originating from each district, and the total permanent migration trips destined to each district are calculated to represent migration specific trip generation and attraction roles of each district. We've also examined different trip distribution models to distribute migration flows in Portugal.

1) MIGRATION TRIP EXPANSION

The migration flows collected from the CDR data in Section II (C) are partial. The next step is to expand the sample migration O-D matrix in order to represent the migration flow behavior of the total population. Sample migration O-D flow can be expanded to a population migration O-D flow through an expansion factor which is based on sampling ratio. A simple procedure is developed to calculate: (i) the total permanent migration trips originating from each district (MO_i); and (ii) the total permanent migration trips destined to each district (MD_j).

In order to obtain MO_i, first, the number of sampled residents of each district is calculated. A total of 148,215 users, whom we have enough information to infer whether they migrated permanently or not are used. The number of sampled residents is based on the number of residence detected in each district. This means, distributing the 148,215 users to all the districts based on their first time residence location, which was inferred from CDR data. The second step is calculating the total number of migrants originated from each district from the sampled resident data (MO_{si}). Then, an expansion factor (f_i) is derived for each district as the ratio between the total number of population of each district (population who are between the age of 20 to 59) and the number of sampled users identified as residents of that district based on the CDR data. Finally, the total permanent migration flow originated from each district (MO_i) is obtained by multiplying MO_{si} and f_i. To calculate MO_i, a census near the timespan of the CDR data is used as a secondary source of data describing the population which can be related to the sample. We could not find appropriate secondary source of data describing the total migration flow attractions in each district. We assumed that the model for total permanent migration trips originating from each district was reasonable. Thus, we control the number of total permanent migration trips, so that the number of total permanent migration trip origins equals the number of permanent migration trip destinations.

2) MIGRATION TRIP DISTRIBUTION MODELING

Trip distribution modelling is the second component of the classical four-step transportation travel demand forecasting model. We examined three well known models to distribute migration flows in Portugal. The first model is the log-linear model, which is based on statistical estimation of flows as a function of several exploratory variables that includes the characteristics of origin and destination zones and travel cost parameter. The second one is the gravity model that produces the most likely set of flows given various constraints reflecting the total trip productions and generations. The third one is the radiation model, which is originated from diffusion dynamics and inspired by Stouffer's framework of intervening opportunities [27]. It is based exclusively on the spatial distribution of population and it is parameter free.

a: GRAVITY MODEL

Inspired by an analogy with the Newton's gravitational law, the gravity model describes the volume of trips made between two areas that can be considered as being proportional to their population but inversely proportional to the travel cost between them. Mathematically, the relationship is defined by (1), as following.

$$T_{ij} = A_i O_i B_j D_j f(C_{ij}), \tag{1}$$

where T_{ij} is the number of trips between district i and district j, O_i and D_j are the total trip ends of districts i and j respectively, and $f(c_{ij})$ is a generalized travel cost function between districts i and j. The two sets of balancing factors

$A_i = 1/\sum_j B_j D_j f(C_{ij})$ and $B_j = 1/\sum_i A_i O_i f(C_{ij})$, which ensure that the estimates of T_{ij}, when summed across both rows and columns of the matrix equal the known O_i and D_j totals [28]. The popular versions for cost function, $f(C_{ij})$, are negative exponential function ($e^{-\beta C_{ij}}$), inverse power function (C_{ij}^{-n}), and combined function ($C_{ij}^{-n} e^{-\beta C_{ij}}$), where β and n are exponential and power parameters for the cost function, respectively.

The gravity model has long been applied in economics (e.g., trade) and transportation studies (e.g., mobility). The classic early applications of the model were by studies of international trade flows by Fisk and Linnemann [29], and urban trip distribution patterns by Bouchard and Pyers [30]. Later studies (e.g., [31], [32]) still show that the relationship holds well. There are however some limitations of the gravity model such as being deterministic (i.e., it cannot account for fluctuations in the number of travelers), its systematic predictive inconsistency and discrepancies [33], its parameters are not always easy to calibrate [34], and the lack of emphasis on social aspects that may be significant [35].

b: LOG-LINEAR MODEL

Log-linear model has been applied to the analysis of values contained within contingency tables. Our analysis of the O-D flow between 18 districts can be considered as a two-way contingency table. A study by [7] has showed how the doubly constrained, multiplicative model in (2), can be equivalent with statistical (additive) log-linear model in (3). Using a similar notation to [7], the multiplicative version of log-linear model describing the full system can be written as shown in (2).

$$T_{ij} = \tau \tau_i^O \tau_j^D \tau_{ij}^{OD}, \tag{2}$$

where, τ is the overall main effect representing the level of migration, τ_i^O and τ_j^D are the origin and destination 'main effects' represented by categorical variables, respectively. Each of them has 18 levels (the total number of districts in Portugal), τ_{ij}^{OD} is an origin-destination interaction component representing the physical or social distance between districts not explained by the other three components with parameters, where $n = 18$ (i.e., 306 in total for our case, where intra-district migrations are not included).

By taking a natural logarithm, the multiplicative log-linear model in (2), can be expressed as a log-linear (additive) model as follows in (3), where $ln(\tau) = \lambda$.

$$ln\left(T_{ij}\right) = \lambda + \lambda_i^O + \lambda_j^D + \lambda_{ij}^{OD} \tag{3}$$

c: RADIATION MODEL

The radiation model is originated from diffusion dynamics and inspired by Stouffer's framework of intervening opportunities [27]. It is based exclusively on the spatial distribution of population and it is parameter free. The radiation model [33] describes the volume of trips made between two areas as a process of job selection that consists of job seeking and job selecting. Job selection considers the number of employment opportunities in each area to be proportional to the resident population, while job selecting whose criteria is to choose the nearest job with a benefit higher than the best offer available in the resident area.

Mathematically, as defined in (4), the model's equation relates the origin population, the destination population, and the total population within the circle centered from the origin area with the distance radius between the origin and the destination areas. Finally, the number of trips from origin to destination is calculated as a part of the total out flows of the origin.

$$T_{ij} = T_i \frac{m_i n_j}{(m_i + s_{ij})(m_i + n_j + s_{ij})}, \tag{4}$$

where T_{ij} denotes the number of trips made from origin i to destination j, T_i is total number of trips departing from location i, m_i is the population of area i, n_j is the population of area j, and s_{ij} is the enclosed population in the circle with radius the distance between areas i and j excluding the populations of i and j.

3) GENERALIZED COST

a: DIFFERENT MEASURES OF TRAVEL COST

Travel cost is the cost incurred to complete a trip between two traffic analysis zones (i.e., districts in our case). Measuring the travel cost is thus important for evaluating migration flows as it represents the disutility of travel as perceived by the trip maker. The actual cost of a trip can be influenced by many factors such as travel time, distance, fuel consumption, and individual efforts (e.g., physical work involved, willingness, travel period, etc.), so it is not quite straightforward to quantify the actual travel cost. Usually, a measure combining all the monetary and non-monetary costs of a journey can be used and this value is referred to as the generalized cost of travel. The generalized cost can be formulated as a linear function of the all the monitory and non-monetary costs of the journey weighted by coefficients which attempt to represent the relative importance as perceived by the trip maker [36]. Estimating these coefficients is beyond the scope of this study and three measures of travel cost based on distance and monetary values are examined separately, as follows:

i) DISPLACEMENT: Displacement is a linear distance between two geolocations, computed using Haversine formula, given by (5).

$$d = 2R \cdot arcsin \times \left(\sqrt{\sin^2\left(\frac{\Delta lat}{2}\right) + \cos(\text{lat1}) \cdot \cos(lat2) \cdot \sin^2\left(\frac{\Delta lon}{2}\right)}\right), \tag{5}$$

where Δ lat = lat2 - lat1, Δlon = lon2 - lon1, and R is the Earth's radius (for which we used the average radius of 6,371km).

ii) ROAD NETWORK DISTANCE: Road network distance is a shortest path from one point to another in a road network,

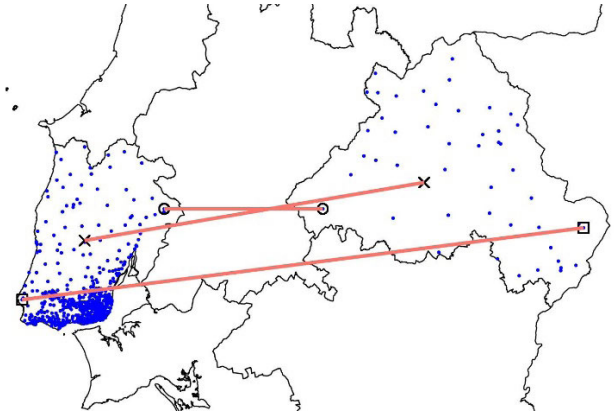


FIGURE 12. An example of different approaches in determining the reference points for travel cost measure.

which is obtained from using the Google Maps Distance Matrix API.

iii) MONETARY COST: Monetary cost is a monetary value for traveling from one location to another in a road network estimated by a taxi fare, which can be obtained from using the Taxi Fare Finder API.[2]

b: REFERENCE POINTS

Measuring a travel cost needs points of reference of the two zones to determine the origin and destination geolocations from which a relative numerical value can be calculated. Here we considered three approaches for determining the reference points, as follows.

i) DISTRICT CENTROIDS: District Centroids using the centroid of the district as the reference geolocation point [37]. Figure 12 shows an example of using district centroids as the reference points of two districts, marked with a cross ('x').

ii) FARTHEST CELL TOWER LOCATIONS: Farthest Cell Tower Locations using the locations of the cell towers that are located farthest apart from each other between the two districts as the reference points. One is located in one district and the other one is located in the other district, as shown in an example in Figure 12, marked with a square ('□').

iii) NEAREST CELL TOWER LOCATIONS: Nearest Cell Tower Locations using the locations of the cell towers that are nearest to each other as the reference points where each one belongs to each of the two districts, as shown with a circle ('O') in an example in Figure 12.

III. RESULTS

A trip distribution model is developed to predict or estimate the number of trips that will be made between a pair of zones, e.g., migration flows in our case. There are various models each differs in their characterization of the incorporated factors that are assumed to affect the trip distribution. Here we consider three models; the traditional gravity model, log-linear model, and a radiation model. The three modeling approaches are explored to estimate a total of 27 trip distribution models.

A. LOG-LINEAR MODEL

The first group of models, Model 1.1.1 to Model 1.3.3, are estimated using log-linear model. The use of 18 district level zoning system resulted a set of 306 inter-district permanent migration O-D flows. We started with the estimation of Model 1.1.1. Fitting a doubly constrained log-linear model in (3), involves categorical variables associated with an origin component representing the relative 'pushes' from each district (18 categorical variables), a destination component representing the relative 'pulls' to each district (18 categorical variables). A travel cost variable, which is continuous is used to capture an origin-destination interaction.

We applied the Generalized Linear Model (GLM) function implemented in R-Programming to estimate the log-linear model in (3), (R Core Team, 2016). The observed migration O-D flows are counts and assumed to follow a Poisson distribution. Thus, the Poisson model in the GLM framework assumed the estimated permanent migration O-D trips is assumed to follow a Poisson distribution with a mean that is logarithmically linked to a linear combination of the origin and destination specific categorical variables and the travel cost variable. For illustration purposes, the estimation result of Model 1.1.1 is shown in (6), where, $\widehat{M1.1.1_{ij}}$ is the estimated permanent migration O-D flow between districts i and j for Model 1.1.1. The $orig_1$ and $dest_1$ variables were used as the reference categories for the origin-specific and the destination-specific categorical variables, respectively. Therefore, the model results obtained for the other categorical variables should be interpreted relative to the reference categories, which are associated to Lisbon district. The coefficients of all the categorical variables are negative and significant at the 95% confidence level, which show migration inflow and outflow are low for other districts than the reference category (Lisbon). The coefficient of the travel cost ($\ln C_{ij}$) variable is statistically significant at the 95% confidence level and its sign is negative.

$$\begin{aligned}\widehat{M1.1.1_{ij}} &= \exp(12.346 - 0.997\text{orig}_2 - 1.701\text{orig}_3 \\ &\quad - 3.595\text{orig}_4 + \ldots - 2.555\text{orig}_{18} - 1.786\text{dest}_2 \\ &\quad - 1.652\text{dest}_3 - 1.051\text{dest}_4 + \ldots - 2.388\text{dest}_{18} \\ &\quad - 1.055\ln C_{ij}), \end{aligned} \tag{6}$$

Figure 13 shows a comparison of the observed (CDR-based) and estimated inter-district migration trips for the nine log-linear models. Figure 13 a shows result of the log-linear model presented in (6), which is compared against the CDR-based migration flows. We also computed a more conventional correlation value to compare linear relationship between the observed trips against the model outputs. A complete list of R values is provided in Table 2.

[2]https://www.taxifarefinder.com

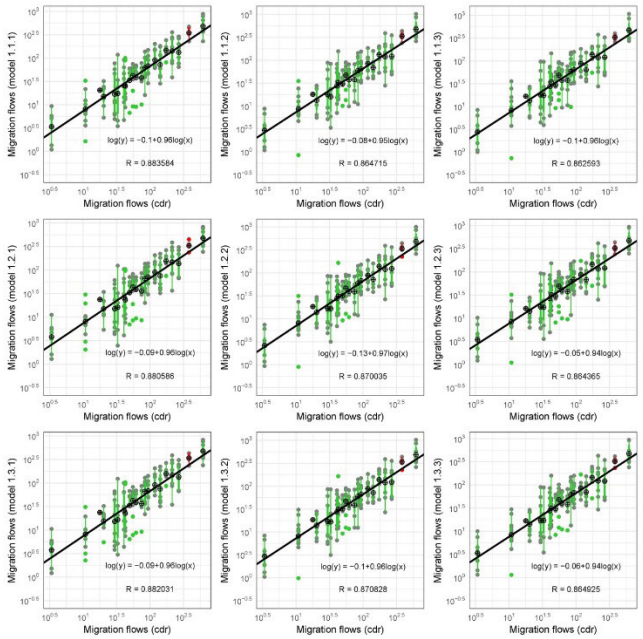


FIGURE 13. Correlation results of different approaches for log-linear models. Grey points are scatter plot for each pair of districts. A box is colored green if the fitted regression line lies between the 9th and the 91st percentiles in that bin and is red otherwise. The black circles with cross lines correspond to the mean number of estimated migration flows in that bin.

Table 2 shows estimation results of nine log-linear models. As can be seen from Tables 2, all nine estimated models have fairly high R-values, ranging from 0.8625 to 0.8835, which indicates a good fit between the observed and the estimated migration O-D flows. The travel cost (lnC_{ij}) parameters of the nine doubly constrained log-linear models are presented in Table 2. In section II (D.3), description of the different approaches to estimate the inter-district travel cost measures are discussed. The coefficient of the travel cost parameter for all nine models is negative, which makes sense since the travel cost variable reflects the travel cost associated to migration between the origin and destination districts and the attractiveness of a destination district decreases with a higher travel cost. A difference in the value of the travel cost parameter is seen between the models that were estimated using a similar travel cost specification. For instance, in the case of Model 1.1.1 and Model 1.2.1, district centroid is used as a reference point to measure the travel cost (travel distance) between the origin and destination districts. However, the travel cost parameter for Model 1.1.1 (-1.055) is larger than the travel cost for Model 1.2.1 (-1.227). One of the main reasons is that the actual travel distance on a road network is used for Model 1.2.1, which results in a higher average travel cost when compared to Model 1.2.1. Consequently, the higher average travel cost in Model 1.2.1 results in a model that is more sensitive to the distance-decay effect, so the migration interaction between two districts declines as the travel cost (measured in travel distance) between them increases.

TABLE 2. Result summary of the log-linear models.

Model	Travel cost	Reference point	Travel cost parameter (lnC_{ij})	R-value	p-value	F-test*
1.1.1	Displacement	Centroid	-1.055	0.883584	$<10^{-16}$	1082
1.1.2	Displacement	Farthest	-0.856	0.864715	$<10^{-16}$	901.1
1.1.3	Displacement	Nearest	-0.694	0.862593	$<10^{-16}$	883.8
1.2.1	Road distance	Centroid	-1.227	0.880586	$<10^{-16}$	1050
1.2.2	Road distance	Farthest	-0.900	0.870035	$<10^{-16}$	946.8
1.2.3	Road distance	Nearest	-0.811	0.864365	$<10^{-16}$	898.2
1.3.1	Taxi fare	Centroid	-1.285	0.882031	$<10^{-16}$	1065
1.3.2	Taxi fare	Farthest	-1.042	0.870828	$<10^{-16}$	954
1.3.3	Taxi fare	Nearest	-0.927	0.864925	$<10^{-16}$	902.8

*F-statistic = 3.91

B. GRAVITY MODEL

Although with its drawbacks, the gravity model does not bear any significant computation burden and is easily scalable with real-world populations while using only a few data variables. Table 3 shows estimation results of nine doubly constrained gravity models for permanent migration flows. Fitting the gravity model in (1), requires estimation of its parameters to reasonably reproduce the observed (benchmark) migration flow pattern. The list of parameters includes A_i, B_j, and a parameter for the travel cost. In our case, the parameters A_i and B_j are calibrated during the estimation of the gravity models, while the parameter for the travel cost is obtained from Table 2 (the corresponding log-linear model estimates). For instance, to estimate Model 2.1.1, 18 A_i parameters, 18 B_j parameters, and 1 travel cost parameter, such as -1.055 from (6) is used.

For each model, a regression line was fitted and a correlation (R-value) was measured between the actual or observed migration flows (i.e., CDR based) and the gravity model's projection or estimated migration flows. The result is shown in Figure 14 and summarized in Table 3.

The model that shows the highest correlation ($R = 0.8835$, $p < 10^{-16}$) is the Model 2.1.1, a gravity with displacement distance as the travel cost and using district centroids as the reference points, followed by the Model 2.3.1, a gravity with taxi fare considered as the travel cost and using district centroids as the reference points ($R = 0.8820$, $p < 10^{-16}$). Among the models that shows lowest correlations are Model 2.1.3, gravity with displacement and using nearest cell tower locations as the reference points ($R = 0.8625$, $p < 10^{-16}$) and Model 2.2.3, gravity with road distance as its travel cost and nearest cell tower locations being reference points ($R = 0.8643$, $p < 10^{-16}$). The correlation values are comparably high across all nine models, ranging from 0.8835 to 0.8625. The complete ranking based on the correlation values is Models 2.1.1, 2.3.1, 2.2.1, 2.3.2, 2.2.2, 2.3.3, 2.1.2, 2.2.3, and 2.1.3, from highest to lowest respectively. By considering the average of each travel cost measurement, the displacement (0.8820) is ranked first, followed road distance (0.8685), then taxi fare (0.8638). If grouped by the reference point approach, the ranking is district centroid (0.873042), farthest cell towers (0.8721), and nearest cell towers (0.8693). In summary, the combination of displacement distance as the travel cost and district centroid has the best fitting for the gravity

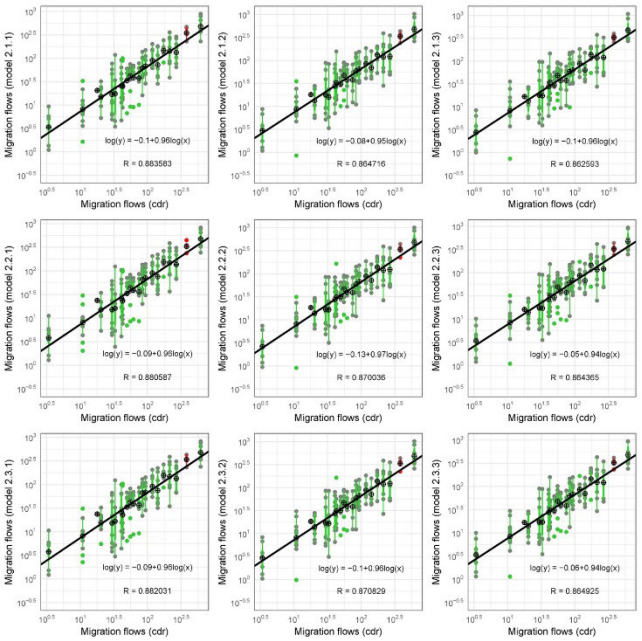


FIGURE 14. Correlation results of different approaches for gravity models. Grey points are scatter plot for each pair of districts. A box is colored green if the fitted regression line lies between the 9th and the 91st percentiles in that bin and is red otherwise. The black circles with cross lines correspond to the mean number of estimated migration flows in that bin.

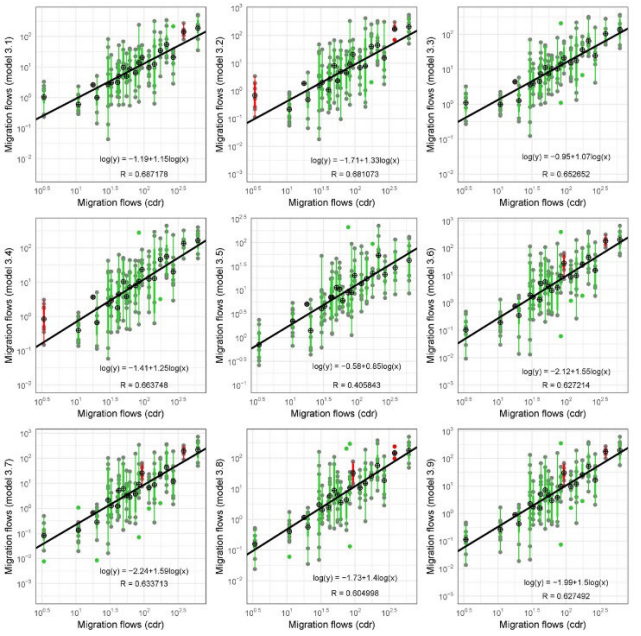


FIGURE 15. Correlation results of radiation models. Grey points are scatter plot for each pair of districts. A box is colored green if the fitted regression line lies between the 9th and the 91st percentiles in that bin and is red otherwise. The black circles with cross lines correspond to the mean number of estimated migration flows in that bin.

TABLE 3. Result summary of the gravity models.

Model	Travel cost	Reference point	R-value	p-value	F-test*
2.1.1	Displacement	Centroid	0.883583	$<10^{-16}$	1082
2.1.2	Displacement	Farthest	0.864716	$<10^{-16}$	901.1
2.1.3	Displacement	Nearest	0.862593	$<10^{-16}$	883.8
2.2.1	Road distance	Centroid	0.880587	$<10^{-16}$	1050
2.2.2	Road distance	Farthest	0.870036	$<10^{-16}$	946.8
2.2.3	Road distance	Nearest	0.864365	$<10^{-16}$	898.2
2.3.1	Taxi fare	Centroid	0.882031	$<10^{-16}$	1065
2.3.2	Taxi fare	Farthest	0.870829	$<10^{-16}$	954
2.3.3	Taxi fare	Nearest	0.864925	$<10^{-16}$	902.8

*F-statistic = 3.91

model. When considering the travel cost and reference point approaches individually, the district centroid and taxi fare are the best approaches.

C. RADIATION MODEL

The log-linear and gravity models were estimated to distribute migration specific trips in Portugal. We also compared the CDR-based migration flow against the radiation model estimates, where estimation is made based on key determinants of migration, such as business establishment data, employment data, etc. The radiation model is mainly based on the population or mass-based parameters (m_i, n_j, s_{ij}), which signify spatial attractiveness that presumably affects the trip distribution. Here, we estimated the radiation model by varying the source of data for these parameters. Different data sources are used to characterize the attractiveness of the area based on which nine radiation models are assembled. Models 3.1 and 3.2 are based on the census data, where one uses total population of the district and the latter uses the district population per km^2. Models 3.3 and 3.4 are based on the data obtained from the Google Places API[3], which provides lists of local businesses where one uses total number of places (local businesses) and the latter uses the number of places per km^2. Model 3.5 is based on the Global Competitiveness Index (GCI), which measures the economic performance on 12 pillars of competitiveness [39]. We used the GCI by city data of Portuguese district capitals published by the Público newspaper on 30 September 2006. Model 3.6 uses the census data on employment [40], which reports the number of people aged 15 or over who performed some work for a wage or salary during the reference year. Model 3.7 uses the census data on unemployment [40], which indicates the number of people aged between 15 and 74 who neither had a job nor were at work during the reference year. Model 3.8 uses the census data on death of residents, which reports the number of residents who passed away or permanently disappeared during the reference year. Lastly, Model 3.9 uses the census data on private households, indicating the number of groups of people residing in the same unit who are related to each other (by law or 'de facto') during the reference year. The reference year of all census data used here was 2011, as it was the closest available census information to our CDR data period.

For each examined radiation model, a regression line was fitted and a correlation (R-value) was measured between the actual or observed migration flows (i.e., CDR based)

[3]https://developers.google.com/places/web-service/intro

TABLE 4. Result summary of the radiation models.

Model	Population or mass terms (m_i, n_j, s_{ij})	R-value	p-value	F-test*
3.1	District population	0.687178	$<10^{-16}$	272
3.2	District population per km^2	0.681073	$<10^{-16}$	263
3.3	Number of local businesses	0.652652	$<10^{-16}$	225.6
3.4	Number of local businesses per km^2	0.663748	$<10^{-16}$	239.4
3.5	Global Competitiveness Index	0.405843	$<10^{-13}$	59.94
3.6	Employment	0.627214	$<10^{-16}$	197.2
3.7	Unemployment	0.633713	$<10^{-16}$	204
3.8	Death of residents	0.604998	$<10^{-16}$	175.5
3.9	Private households	0.627492	$<10^{-16}$	197.4

*F-statistic = 3.91

and the model's estimate. The resulting regression is shown in Figure 15 and statistical values are summarized in Table 4.

The result shows that Model 3.1, the district population-based model has the highest correlation value ($R = 0.6871$, $p < 10^{-16}$) followed by Model 3.2, the distance population per km^2-based model ($R = 0.6810$, $p < 10^{-16}$), while the GCI-based model surprisingly is the least accurate model with a correlation value of as low as 0.4058 ($p < 10^{-16}$). Based on the correlation values, the ranking is Models 3.1, 3.2, 3.4, 3.3, 3.7, 3.9, 3.6, 3.8, and 3.5, from the highest to lowest values respectively. The models that are based on the population and local businesses appear to hold a better relation with a relatively high correlation value (R is above 0.65).

With different approaches in defining the mass terms for the radiation model, all results do not exceed the performance of the gravity and log-linear models. In fact, the resulting correlation values that reflect on the fitting of the model to the CDR-based migration trips are relatively much lower than that of the other two models. This could be due to the fact that the radiation model is based exclusively on the population size and other mass terms, hence the travel cost is not considered, which is also one of key influential factors to capture an origin-destination interaction.

This migration trip distribution modeling is performed here to demonstrate that our intuitive approach for migration trip inference is effective, which can be reasonably described by well-known models. As observed, the resulting estimations vary with the models that are different in their characterization of the incorporated factors assumed to affect the migration.

IV. CONCLUSION

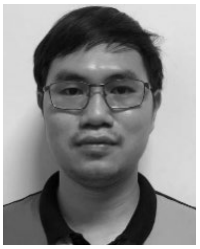
Mobile phone has become an indispensable part of our lives. With its equipped sensory components and communication technologies, it has also become our personal behavioral sensor. Collectively, communication records can be used to better understand behaviors and provide insightful information about people and city characteristics. By taking the approach of utilizing a massive mobile phone network data (CDR: call detail records) in this study, we were able to reasonably infer about the migration trips (or change of residence) for which different trip distribution models were examined to mathematically describe such human mobility. We described our methodology that includes data preprocessing, subject selection, and migration trip inference from which some exploratory results were revealed, such as intra/inter-district migration flow characteristics, migration distance distribution, and migration origin-destination (O-D) movements. For migration trip modeling, the log-linear model, traditional gravity model and recently introduced radiation model were examined with different approaches taken in defining parameters for each model. For the log-linear and gravity models, displacement, road network distance, and monetary cost were each considered as an alternative travel cost, while the reference points were experimentally defined as the district centroids, farthest cell tower locations, and nearest cell tower locations. For the radiation model, district population, number of local businesses, Global Competitiveness Index, employment, unemployment, death of residents, and private households were experimentally defined as the population or mass parameters. We believe that our approach and findings contribute to the body of knowledge in the human mobility research community.

There are nonetheless some limitations of our study. Firstly, there were only subjects who migrated once were inferred and considered in this study. The results obtained here could potentially vary with consideration of multiple migration scenarios. Secondly, due to the intra-zonal trip distance, only the inter-district migration was considered in trip distribution modeling in this study. It is thus worth an investigation of the intra-district migration trip distribution modeling in a future study by exploring different possible approaches for estimating the trip distance. Thirdly, there was a lack of ground truth for the inferred migration trips, which was mainly due to the data unavailability. Although the ground truth was not available, the migration inference was still validated based on a reasonable interpretation from the census-based population density evaluation. Moreover, the inferred migration flow in this paper is based on approximately 18% of the population, which do not represent the entire population. In this research, sampled residents of each district are used to expand sample migration O-D flow to a population migration O-D flow. Future studies should attempt the inclusion of socioeconomic and demographic information of sample users to properly explain the composition of the sampled data. In our analysis, an inverse power function is applied for distance decay parameter estimation. The rate at which migration flow propensities drop off with increased cost varies from place to place and future studies should explore this issue in detail. Area of potential future improvement includes developing separate models for short distance and long distance migration flows to understand the effects of economic, social capital, and amenity variables on the mobility behavior of migrants.

REFERENCES

[1] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[2] M. Piesse, "Factors influencing migration and population movements—Part 1," Future Directions Int., Nedlands, WA, Australia, Tech. Rep., 2014.
[3] F. T. Charlene Rohr, J. Fox, A. Daly, B. Patruni, and S. Patil, "Modelling long-distance travel in the UK," in *Proc. Eur. Transport Conf.*, 2010, pp. 1–22.
[4] *Roadmap to a Single European Transport Area-Towards a Competitive and Resource Efficient Transport System*, Eur. Commission, Brussels, Belgium, 2011.
[5] M. G. Mcnally, "The four step model," in *Handbook of Transport Modelling*. Bingley, U.K.: Emerald Group Publishing Ltd, 2007.
[6] P. R. Stopher and S. P. Greaves, "Household travel surveys: Where are we going?" *Transp. Res. A, Policy Pract.*, vol. 41, no. 5, pp. 367–381, 2007.
[7] J. Raymer, "The estimation of international migration flows: A general technique focused on the origin-destination association structure," *Environ. Planning A, Econ. Space*, vol. 39, no. 4, pp. 985–995, 2007.
[8] J. E. Blumenstock, "Inferring patterns of internal migration from mobile phone call records: Evidence from Rwanda," *Inf. Technol. Develop.*, vol. 18, no. 2, pp. 107–125, 2012.
[9] E. Zagheni, V. R. K. Garimella, I. Weber, and B. State, "Inferring international and internal migration patterns from Twitter data," in *Proc. 23rd Int. Conf. World Wide Web*, 2016, pp. 439–444.
[10] D. S. Massey and C. Capoferro, "Measuring undocumented migration," *Int. Migration Rev.*, vol. 38, no. 3, pp. 1075–1102, 2006.
[11] K. Beegle, J. de Weerdt, and S. Dercon, "Migration and economic mobility in Tanzania: Evidence from a tracking survey," *Rev. Econ. Stat.*, vol. 93, no. 3, pp. 1010–1033, 2011.
[12] M. G. Demissie, "Combining datasets from multiple sources for urban and transportation planning: Emphasis on cellular network data," Univ. Coimbra, Coimbra, Portugal, 2014.
[13] L. Shen and P. R. Stopher, "Review of GPS travel survey and GPS data-processing methods," *Transp. Rev.*, vol. 34, no. 3, pp. 316–334, 2014.
[14] A. Cuttone, S. Lehmann, and M. C. González, "Understanding predictability and exploration in human mobility," *EPJ Data Sci.*, vol. 7, no. 1, pp. 1–2, 2018.
[15] S. Phithakkitnukoon, T. Horanont, A. Witayangkurn, R. Siri, Y. Sekimoto, and R. Shibasaki, "Understanding tourist behavior using large-scale mobile sensing approach: A case study of mobile phone users in Japan," *Pervasive Mobile Comput.*, vol. 18, pp. 18–39, Apr. 2015.
[16] N. Caceres, J. P. Wideberg, and F. G. Benitez, "Review of traffic data estimations extracted from cellular networks," *IET Intell. Transp. Syst.*, vol. 2, no. 3, pp. 179–182, 2008.
[17] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti, "Real-time urban monitoring using cell phones: A case study in Rome," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 1, pp. 141–151, Oct. 2011.
[18] C. Ratti, A. Sevtsuk, S. Huang, and R. Pailer, "Mobile landscapes: Graz in real time," in *Location Based Services and TeleCartography*. New York, NY, USA: Springer, 2005, pp. 433–444.
[19] H. Bar-Gera, "Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from Israel," *Transp. Res. C, Emerg. Technol.*, vol. 15, no. 6, pp. 380–391, 2007.
[20] M. G. Demissie, G. H. de Almeida Correia, and C. Bento, "Intelligent road traffic status detection system through cellular networks handover information: An exploratory study," *Transp. Res. C, Emerg. Technol.*, vol. 32, pp. 76–88, Jul. 2013.
[21] H. X. Liu, A. Danczyk, R. Brewer, and R. Starr, "Evaluation of cell phone traffic data in Minnesota," *Transp. Res. Rec.*, vol. 2086, no. 1, pp. 1–7, 2008.
[22] M. G. Demissie, S. Phithakkitnukoon, and L. Kattan, "Trip distribution modeling using mobile phone data: Emphasis on intra-zonal trips," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 7, pp. 2605–2617, Jul. 2019.
[23] S. Phithakkitnukoon, Z. Smoreda, and P. Olivier, "Socio-geography of human mobility: A study using longitudinal mobile phone data," *PLoS ONE*, vol. 7, no. 6, 2012, Art. no. e39253.
[24] F. Calabrese, G. Di Lorenzo, L. Liu, and C. Ratti, "Estimating origin-destination flows using mobile phone location data," *IEEE Pervasive Comput.*, vol. 10, no. 4, pp. 36–44, Apr. 2011.
[25] J. L. Toole, M. Ulm, M. C. González, and D. Bauer, "Inferring land use from mobile phone activity," in *Proc. ACM SIGKDD Int. Workshop Urban Comput.*, 2012, pp. 1–8.
[26] M. G. Demissie, G. Correia, and C. Bento, "Analysis of the pattern and intensity of urban activities through aggregate cellphone usage," *Transp. A, Transp. Sci.*, vol. 11, no. 6, pp. 502–524, 2015.
[27] S. A. Stouffer, "Intervening opportunities: A theory relating mobility and distance," *Amer. Sociol. Rev.*, vol. 5, no. 6, pp. 845–867, Dec. 1940.
[28] A. G. Wilson, "The use of entropy maximising models, in the theory of trip distribution, mode split and route split," *J. Transp. Econ. Policy*, vol. 3, no. 1, pp. 108–126, 1969.
[29] P. R. Fisk and H. Linnemann, "An econometric study of international trade flows," *Econ. J.*, vol. 77, no. 306, pp. 366–368, 1967.
[30] R. J. Bouchard and C. E. Pyers, "Use of gravity model for describing urban travel: An analysis and critique," *Highway Res. Rec.*, vol. 88, no. 88, pp. 1–43, 1965.
[31] M. M. M. Abdel-Aal, "Calibrating a trip distribution gravity model stratified by the trip purposes for the city of Alexandria," *Alexandria Eng. J.*, vol. 53, no. 3, pp. 677–689, 2014.
[32] M. Lenormand, A. Bassolas, and J. J. Ramasco, "Systematic comparison of trip distribution laws and models," *J. Transp. Geogr.*, vol. 51, pp. 158–169, Feb. 2016.
[33] F. Simini, M. C. González, A. Maritan, and A.-L. Barabási, "A universal model for mobility and migration patterns," *Nature*, vol. 484, no. 7392, pp. 96–100, 2012.
[34] F. Gargiulo, M. Lenormand, S. Huet, and O. B. Espinosa, "Commuting network models: Getting the essentials," *J. Artif. Soc. Soc. Simul.*, vol. 15, no. 2, p. 6, 2012, doi: 10.18564/jasss.1964.
[35] S. Kraft and J. Blazek, "Spatial interactions and regionalisation of the Vysočina Region using the gravity models," *Acta Univ. Palackianae Olomucensis-Geograph.*, vol. 43, no. 2, pp. 65–82, 2012.
[36] J. de D. Ortúzar and L. G. Willumsen, *Modelling Transport*. Hoboken, NJ, USA: Wiley, 2011.
[37] M. G. Demissie, S. Phithakkitnukoon, L. Kattan, and A. Farhan, "Understanding human mobility patterns in a developing country using mobile phone data," *Data Sci. J.*, vol. 18, no. 1, pp. 1–13, 2019.
[38] *R: A Language and Environment for Statistical Computing*, R Found. Stat. Comput., Vienna, Austria, 2016.
[39] X. Sala-i-Martin and E. V. Artadi, "The global competitiveness index," Global Econ. Forum, Cologny, Switzerland, Global Competitiveness Rep., 2004.
[40] Francisco Manuel dos Santos Foundation. *PORDATA Database of Contemporary Portugal*. Accessed: Jun. 8, 2019. [Online]. Available: https://www.pordata.pt/en/Municipalities/Search/5/



SORANAN HANKAEW received the B.Sc. degree in statistics from Chiang Mai University, Thailand, where he is currently pursuing the degree with the Department of Computer Engineering. His research interests include urban data analytics and intelligent transportation systems.

SANTI PHITHAKKITNUKOON received the B.S. and M.S. degrees in electrical engineering from Southern Methodist University, USA, in 2003 and 2005, respectively, and the Ph.D. degree in computer science and engineering from the University of North Texas, USA. He is currently an Associate Professor with the Department of Computer Engineering, Chiang Mai University, Thailand. Before joining Chiang Mai University, he was a Lecturer in computing with The Open University, U.K., a Research Associate with Newcastle University, U.K., and a Postdoctoral Fellow with the SENSEable City Laboratory, Massachusetts Institute of Technology, USA. His research is in the area of urban informatics.

MERKEBE GETACHEW DEMISSIE received the M.Sc. degree in transport systems from the Royal Institute of Technology (KTH), Sweden, in 2009, and the Ph.D. degree in transportation systems from the MIT-Portugal Program, in 2014. He is currently a Research Associate with the Department of Civil Engineering, University of Calgary, Canada. Before joining the University of Calgary, he was a Postdoctoral Fellow with the University of Coimbra and the Instituto Pedro Nunes, Portugal. His main research interests include transport demand modeling, intelligent transportation systems, data mining, and machine learning.

LINA KATTAN is currently a Professor of civil engineering with the University of Calgary, Canada. She also holds an Urban Alliance Professorship in Transportation Systems Optimization. Her research program focuses on advanced traffic management and information systems, including intelligent transportation systems (ITS), traffic control, the application of artificial intelligence to ITS, connected and autonomous vehicles, network microsimulation modeling and analysis, dynamic traffic assignment, dynamic demand modeling, and traveler behavioral modeling in response to traffic and transit information.

ZBIGNIEW SMOREDA received the Ph.D. degree in sociology from Paris-Est University. He is currently a Senior Researcher with the Orange Labs. Before integrating (future) Orange in 1995, he was an Assistant Professor with Warsaw University, a Researcher and a Lecturer with GRIFS (Université de Paris 8), a Researcher with GAST (France Télécom), and with Observatoire Mondial des Systèmes de Communication. His work in SENSE/Orange Labs is related to sociology of communication and in particular to social uses of ICT and social network forms and transformations associated with technologies.

CARLO RATTI received the Ph.D. degree in architecture from the University of Cambridge. He is currently an Architect and Engineer, who practices architecture in Turin, and teaches at MIT, where he also directs the SENSEable City Laboratory. His research interests include urban design, human–computer interfaces, electronic media, and the design of public spaces. He is a member of the Ordine degli Ingegneri di Torino, the Architects Registration Board (U.K.), and the Association des Anciens Eléves de l'École Nationale des Ponts et Chaussées.