Estimation of Urban Commuting Patterns Using Cellphone Network Data

Vanessa Frias-Martinez
Telefonica Research, Madrid, Spain
vanessa@tid.es

Cristina Soguero
Telefonica Research, Madrid, Spain
soguero@tid.es

Enrique Frias-Martinez
Telefonica Research, Madrid, Spain
efm@tid.es

ABSTRACT

Commuting matrices are key for a variety of fields, including transportation engineering and urban planning. Up to now, these matrices have been typically generated from data obtained from surveys. Nevertheless, such approaches typically involve high costs which limits the frequency of the studies. Cell phones can be considered one of the main sensors of human behavior due to its ubiquity, and as a such, a pervasive source of mobility information at a large scale. In this paper we propose a new technique for the estimation of commuting matrices using the data collected from the pervasive infrastructure of a cell phone network. Our goal is to show that we can construct cell-phone generated matrices that capture the same patterns as traditional commuting matrices. In order to do so we use optimization techniques in combination with a variation of Temporal Association Rules. Our validation results show that it is possible to construct commuting matrices from call detail records with a high degree of accuracy, and as a result our technique is a cost-effective solution to complement traditional approaches.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms,Experimentation,Measurement.

Keywords

Commuting Patterns, O-D Matrix, Call Detail Records.

1. INTRODUCTION

Commuting patterns are typically represented using commuting matrices, which are a particular case of O-D matrices. O-D matrices characterize the transitions of a population between different geographical regions representing the origin (O) and destination (D) of a route. When building commuting matrices the geographical areas representing origin(O) and destination (D) capture where people live and work. Typically O and D are the same set and represent the towns or neighborhoods of the geographical area under study. Each element of the commuting matrix (i, j) defines the percentage of individuals that live in O_i and work in D_j.

Typically, National Statistical Institutes carry out periodical surveys asking different segments of the population about their commuting patterns [16]. The information obtained is used as input for O-D generation techniques. However, such approach typically involves high costs and the data collected has spatio-temporal limitations, which implies that the matrices generated typically only represents a snapshot of the commuting patterns over time.

In recent years, cell phones have become a pervasive technology with users carrying them at almost all times. The ubiquity of these platforms has transformed cell phones into one of the main sensors of human behavior. In fact, every time a subscriber makes or receives a phone call, or an SMS, or an MMS, information regarding the interaction as well as the geolocation of the user (in the form of the tower used for the communication) is logged for billing purposes. As a result we can find in the literature a variety of studies focussing on using cell phone data for estimating traffic and commuting patterns [8][18] . Following this trend, in this paper we explore the use of the location information contained in Call Detail Records as a means to compute the commuting patterns of a population expressed as an O-D matrix. Such approach overcomes the limitations posed by the use of other proxies (like smart cards, surveys or social security records) and it can be carried out as often as necessary with very limited costs.

Compared to the literature, our approach has the following contributions: (1) We base our study in Call Detail Records, which are already available for billing purposes in a telco operator, and not in specific measurements and/or traces obtained from the cell phone network. As a result our approach is based on a big part of a population and not on a limited number of traced cell phones; (2) We present a new technique for defining and constructing O-D matrices based on a new temporal variation of association rules (TAR, Temporal Association Rules) and (3) Our technique is designed to capture the different cultural commuting schedules of different urban areas.

2. CELLULAR INFRASTRUCTURE

In order to compute the commuting patterns of a population from geolocated cell phone logs, we first give a brief overview about how these pervasive networks work. Cell phone networks are built using a set of base transceiver stations (BTS) that are in charge of communicating cell phone devices with the network. Each BTS tower has a geographical location typically expressed by its latitude and longitude. The area covered by a BTS tower is called a cell. Each

UrbComp'12, August 12, 2012. Beijing, China.

cell is typically divided in three sectors, each one covering 120 degrees. At any given moment, one or more BTSs can give coverage to a cell phone. Whenever an individual makes a phone call, the call is routed through a BTS in the area of coverage. The BTS is assigned depending on the network traffic and on the geographic position of the individual.

CDR (Call Detail Record) databases are generated when a mobile phone connected to the network makes or receives a phone call or uses a service (e.g., SMS, MMS, etc.). In the process, and for invoice purposes, the information regarding the time and the BTS tower where the user was located when the call was initiated is logged, which gives an indication of the geographical position of a user at a given moment in time. Note that no information about the exact position of a user in a cell is known. Also, no information about the location of cell phone is known or stored if no interaction is taking place.

From all the data contained in a CDR, our study uses the encrypted originating number, the encrypted destination number, the time and date of the call, the duration of the call, and the latitude and longitude of the BTS tower used by the originating cell phone number and the destination phone number when the interaction happened. In order to preserve privacy, all the information presented is aggregated and original records are encrypted. No contract or demographic data was considered or available for this study.

3. PROBLEM DEFINITION

A commuting matrix $CM[O, D]$ represents the percentage of population that commutes on an average daily basis from an origin geographical area O to a destination geographical area D. Typically O and D represent the same set of towns, and as a result a commuting matrix is usually a square matrix. Two commuting matrices can be defined: the home-work commuting matrix $CM[H, W]$ and the work-home commuting matrix $CM[W, H]$. In the first case, each row of the commuting home-work matrix $CM[H, W]$, H_i represents the percentage of population that lives in geographical area H_i and commutes to each geographical area W_j. The diagonal of the matrix expresses the percentage of the population that lives and works in the same town. Symmetrically, the work-home commuting matrix $CM[W, H]$ accounts for the population that works in the geographical area W_i and commutes back home to each one of the geographical locations H_j (columns). From this explanation, being N the number of geographical areas considered, it follows that $\sum_{j=1}^{j=N} CM[H_i, W_j] = 1 \forall i \in [1, ..., N]$ and $\sum_{j=1}^{j=N} CM[W_i, H_j] = 1 \forall i \in [1, ..., N]$.

Traditionally, such commuting matrices are computed by National Statistical Institutes (NSIs) that run surveys and questionnaires across the population under study and determine the commutes that citizens carry out on a daily basis. These mobility matrices are typically available at census bureaus. However, as stated earlier, such surveys are expensive and thus carried out every certain number of years.

The goal of this paper is to present a mechanism to estimate the commuting matrix of a geographical area from the information contained in CDR records that can approximate the values provided by traditional questionnaire-based approaches. For that purpose, two mechanisms need to be defined: (1) the construction of commuting matrices from CDR data and (2) an optimization process that identifies which behavioral patterns better define commuting when using CDR data.

4. ESTIMATING COMMUTING MATRICES FROM CDR

In this section we will present the mechanisms needed to characterize the commuting patterns of a population from call detail records (CDR).

4.1 From CDRs to Commuting Matrix

To compute a commuting matrix from CDRs we first need to identify the geographical areas in the region under study that we are going to use as either *home* or *work*. Given that the goal of this paper is to present an alternative method to generate commuting matrices, for each particular case we will select as regions the same ones considered by corresponding NSI. We assign to each region the set of BTSs geographically included in them (i.e. the towers that give coverage to that area). As a result each geographical area considered $g_i, i = 1, ..., N$, with N the total number of geographical areas considered, can be characterized by a set of BTSs $g_i = \{bts_1, bts_2, ..., bts_k\}$.

Once these areas have been characterized, we need to compute – from the CDRs– the individuals that called from an origin area at some point in time and later show calling activity at a destination area. These associations will populate the home-work and work-home commuting matrices.

We can formalize this problem using Association Rules [1]. Association Rules (ARs) were introduced by Agrawal *et al.* as a technique to discover specific item relationships in itemsets [1]. Specifically, given an itemset $X = X_1, X_2, ..., X_n$, an Association Rule of the type $X \rightarrow Y$ implies that whenever X is satisfied, Y is also satisfied, with a given support and confidence. Formally, being P the probability of an itemset:

$$support(X \rightarrow Y) = P(X \bigcup Y) \quad (1)$$

$$confidence(X \rightarrow Y) = P(Y|X) = \frac{P(X \bigcup Y)}{P(X)} \quad (2)$$

Often times, Association Rules(AR) are used to find the tuples that satisfy minimum support and confidence values in a dataset. ARs are calculated using the *Apriori* algorithm presented in [1]. In our context, we seek association rules $H_i \rightarrow W_j$ and $W_i \rightarrow H_j$ that identify tuples characterizing the home to work and work to home commutes. Furthermore, we require these events to happen in a temporal order *i.e.*, the home-work matrix $CM[H, W]$ is populated with pairs of events $H_i \rightarrow W_j$ such that the interaction at a home location H_i always happens earlier in time than an interaction event at work location W_j; analogously, the work-home matrix $CM[W, H]$ is populated with pairs $W_i \rightarrow H_j$ where an interaction event at work location W_i always happens before an interaction at a home location H_j. Because traditional Association Rules do not consider any temporal order, we present a technique designed to capture these elements: *Temporal Association Rules (TARs)*.

4.1.1 Temporal Association Rules

Temporal Association Rules extend association rules by introducing temporal constraints in the relationship between antecedent and consequent [12][6]. For our context, we propose a new Temporal Association Rule (TARs) where items X and Y are required to happen within a specific time interval. Specifically, each association

Algorithm 1: $CM_{CDR} = CMTAR(CDR, (t_{O,start}, t_{O,end}), (t_{D,start}, t_{D,end}))$

```
CM[O,D]
for each Subscriber S do
  for i = 1, ..., |CDR| do
    if time(CDR_i) ∈ [t_{O,start}, t_{O,end}] then
      O = location(CDR_i)
      for j = i, ..., |CDR within 24h| do
        if time(CDR_j) ∈ [t_{D,start}, t_{D,end}] then
          D = location(CDR_j)
          CM(O, D) + +
        end if
      end for
    end if
  end for
end for
for each pair (O_i, D_j) do
  CM[O_i, D_j] = CM(O_i, D_j) / Σ_{j=1}^{j=N} CM(O_i, D_j)
end for
```

Figure 1: ***CMTAR algorithm for the construction of an O-D matrix using Temporal Association Rules (TAR).***

rule $X \rightarrow Y$ is characterized not only by its support and confidence, but also by time intervals at which items X and Y need to happen *i.e.,* $X[T_O] \rightarrow Y[T_D]$, where T_O is the time interval when the antecedent (or origin O) has to happen and T_D the time interval when consequent (or destination D) has to happen. Also while in traditional Association Rules, antecedents and consequents can have more than one element, in our approach X and Y contain just one element, i.e. one geographical area, indicating the Origin(O) and the Destination(D).

In order to reveal commuting patterns from CDRs, we seek to identify the temporal association rules whose confidence represents the percentage of individuals that are at an origin location O_i during a time interval $T_O = [t_{O,start}, t_{O,end}]$ and move to a destination location D_j where they are present during a time interval $T_D = [t_{D,start}, t_{D,end}]$, formally:

$$O_i[t_{O,start}, t_{O,end}] \rightarrow D_j[t_{D,start}, t_{D,end}] \quad (3)$$

Note that $t_{O,end}$ happens before $t_{D,start}$. In our framework, O_i and D_j represent geographical regions and the temporal association rules will either reveal commuting patterns from home to work locations (with O=home location and D=work location) or work to home commutes (with O=work and D=home).

In order to construct a commuting matrix CM, we propose CMTAR, a TAR-based algorithm (see CMTAR Algorithm in Figure 1) that receives as input a set of CDRs and a pair of time intervals T_O and T_D. The algorithm produces as output a Commuting Matrix obtained from CDR records (CM_{CDR}) for the corresponding time intervals. CMTAR identifies for each subscriber S within the CDR dataset, all the pairs $O_i \rightarrow D_j$ such that O_i happens within the interval $[t_{O,start}, t_{O,end}]$ and D_j happens no later than 24 hours within the interval $[t_{D,start}, t_{D,end}]$. Each element of the commuting matrix $CM_{CDR}[O, D]$ is populated with the confidence values associated to each Temporal Association Rule (TAR) $O_i \rightarrow D_j$, with $i, j = 1, ..., N$ (see Equation (2)).

From an implementation perspective, we have implemented CMTAR using a modified *Apriori* algorithm designed to capture the temporal characteristics of TAR. The algorithm assumes that the set of CDRs are grouped for each subscriber S by date and time, being $|CDR|$ the number of CDR entries.

4.2 Optimizing Time Intervals

CMTAR constructs a Commuting Matrix CM_{CDR} using CDR and a set of time intervals that define the Temporal Association Rules. The problem is how to identify which temporal ranges best capture the behavioral fingerprint for the commuting matrix. The objective is to identify the time intervals for the origin and destination of the Temporal Association Rules (T_O and T_D) that produce a Commuting Matrix from CDR (CM_{CDR}) as similar as possible to the original Commuting Matrix provided by the corresponding National Statistics Institute (CM_{NSI}).

A first approach could use brute force to test all possible time intervals, and compute the similarity between CM_{CDR} and CM_{NSI}, being the best solution the one with the highest similarity value. However, due to the large amount of CDR data such approach is not computationally feasible. We propose to use optimization techniques to identify the optimal time intervals that best characterize the commuting patterns. In the following sections, we will present the use of Genetic Algorithms (GA) and Simulated Annealing(SA) to implement the optimization process. Both techniques have been shown to be useful in similar problems [9], and although they are both stochastic, they explore the candidate populations using significantly different approaches.

In our context, for each pair of time intervals T_O and T_D that the optimization technique evaluates, we first need to compute CM_{CDR} using the CMTAR algorithm. In order to evaluate its accuracy, we measure the similarity between CM_{NSI} and CM_{CDR}. As explained, each row in CM_{CDR} represents the set of confidence values for the corresponding TARs for all commutes departing from each geographical area O_i to any destination location ($O_i \rightarrow D_*$). Similarly, each row in CM_{NSI} represents the confidence of the associated TAR from each geographical area O_i to geographical areas D_*. Thus, in order to evaluate the accuracy of CM_{CDR} we need to evaluate the similarity of each row with the corresponding row of CM_{NSI}. For that purpose, we use Pearson's correlation[14] to analyze the similarity between each origin location O_i in CM_{CDR} with CM_{NSI} and the final similarity value is given by the aver-

age Pearson correlation across all origins. Formally the similarity between CM_{NSI} and CM_{CDR} is obtained as:

$$c(O_i) = Pearson(CM_{CDR}[O_i, D*], CM_{NSI}[O_i, D*] \quad (4)$$

$$similarity = \sum_{i=1}^{N} |c(O_i)|/N \quad (5)$$

4.2.1 Optimizing Time Intervals with GA

Genetic Algorithms (GA) are search algorithms based on the mechanics of natural selection tailored for vast and complex search spaces [2]. A GA starts with a population of abstract representations (called chromosomes) of candidate solutions (individuals) that evolves towards an improved sets of solutions. A *chromosome* is composed of several genes that code the value of a specific variable of the solution. Each gene is typically represented as a string of 0s and 1s. During the evolution, individuals from one generation are used to form a new generation, which is (hopefully) closer to the optimal solution. GAs use a fitness function in order to evaluate the quality of the solution represented by a specific individual. In each generation, GA creates a new set of individuals obtained from recombining the fittest solutions of the previous generation (crossover), occasionally adding random new data (mutation) to prevent the population from stagnating. This generational evolution is repeated until some condition (for example number of populations or improvement of the best solution) is satisfied.

In the context of identifying the best time intervals for constructing CM_{CDR}, GA takes as input the set of phone calls (CDRs) from a geographical region and CM_{NSI}, that defines the optimization objective. Each candidate solution produced by GA is designed to capture the time intervals at which commuters call from origin and destination locations. In order to do that, we define a chromosome composed of four different genes. The first two genes represent the starting time and the finishing time at which subscribers make phone calls from the origin locations O. The last two genes represent the starting time and the finishing time at which subscribers make phone calls from destination locations D. Each gene is composed of five bits, which accounts for the 24 hours of the day. Given that we require that $[t_{O,start}, t_{O,end}]$ happens before $[t_{D,start}, t_{D,end}]$, whenever the newly computed chromosomes does not satisfy this restriction, we assume that T_O happens the natural day before T_D.

The fitness of each candidate solution is evaluated using Equation (5), i.e. we define the fitness function as the accuracy of the mobility matrix CM_{CDR} with respect to the NSI mobility matrix, CM_{NSI}. As a first step to evaluate the fitness of a candidate solution, CM_{CDR} has to be generated using CMTAR algorithm with the time slots defined by the genes of the candidate solution.

For example, if a candidate solution proposed by the GA has the values [(06,09),(17,22)], CMTAR computes the temporal association rules $O_i \rightarrow D_j$ that represent calls made or received at location O_i during a morning interval (6am to 9am) and at location D_j during a night period (5pm to 10pm). The confidence values are then used to generate CM_{CDR}, whose fitness is evaluated using CM_{NSI} with Equation (5).

4.2.2 Optimizing Time Intervals with SA

Simulated Annealing (SA) is a probabilistic method designed to find the global minimum of a cost function that may posses several local minima[11]. It works by emulating the physical process whereby a solid is slowly cooled so that its structure is frozen at a minimum energy configuration [3].

The SA metaheuristic starts from a random initial configuration and seeks to find solutions that minimize an energy function $E(x)$ as the temperature T decreases. At each step, the solution explored is accepted as long as the Acceptance Probability Function (APF) that depends both on the energy and on a varying temperature has a higher value than a randomly selected number:

$$P(E(s), E(new), T) > random(0, 1) \quad (6)$$

The APF is selected such that the smaller the value of T the less "uphill" solutions are allowed to be explored, and as T decreases, the more the "downhill" solutions are favored. Such an approach guarantees that the process does not get stuck in local minima reaching a good approximation to a global minimum. This process is repeated multiple times at each temperature value to allow the system to stabilize before decreasing T again.

In our context, SA takes as input CDRs and CM_{NSI}, and outputs CM_{CDR} and the intervals $T_O = [t_{O,start}, t_{O,end}]$ and $T_D = [t_{D,start}, t_{D,end}]$ that best characterize commuting patterns. For that purpose, SA explores randomly selected time intervals seeking the ones that decrease the candidate's energy $E(x)$ until a global minimum is found. Each candidate solution explored by SA is defined as a set of two intervals, one representing the time interval at origin $[t_{O,start}, t_{O,end}]$ and another representing time interval at destination $[t_{D,start}, t_{D,end}]$. Each time in the intervals is represented as a number in $[0, 24]$, checking that T_O happens before T_D. If this condition is not satisfied, the process assumes that T_O happens the natural day before T_D.

Whenever SA explores a new candidate solution, it randomly selects for each time t of each slot a new value from its neighborhood. Given that SA seeks to minimize the energy function, we define it as one minus the correlation coefficient between CM_{CDR} and CM_{NSI} obtained by Equation (5). Finally, the temperature T is decreased following a geometric decrement such that $T_{new} = \alpha * T_{old}$.

5. EXPERIMENTAL EVALUATION

In this section we present an evaluation of the mechanism we have proposed to generate the commuting matrix for the region of Madrid using CDR data. The state has a population of 6.5M and a size of 8,000 Km^2, with the city of Madrid concentrating 3.3M in population, and the rest corresponding to the 48 municipalities in which the region is divided. Figure 2 presents the map of the region and the division in municipalities.

5.1 Datasets

We have used two sources of information from the year 2009: (1) the NSI mobility matrices for the state of Madrid and (2) a CDR dataset of cell phone calls made and received in the state.

NSI matrices represent the home-to-work ($CM_{NSI}[H, W]$) and work-to-home ($CM_{NSI}[W, H]$) commuting patterns during 2009 for the 48 municipalities shown in Figure 2. These municipalities are considered as the Origin O and Destination D sets. Such

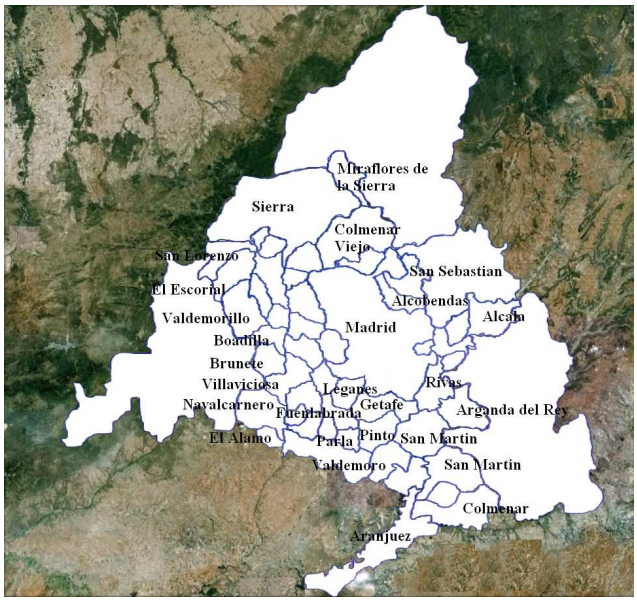

Figure 2: ***Geographical division of the municipalities in which the region of Madrid is divided (including the names of some of them). The association between BTS towers and geographical areas is defined by this borders.***

matrices were built by the local NSI after gathering information regarding the municipality where a person lived and the municipality where a person worked.

The second source of information is a CDR dataset that contains all phone calls, SMS and MMS, that were collected from BTS towers located in the state of Madrid during October and November of 2009, which account roughly for 3.5M unique phones and around 300M interactions. This dataset also includes the geolocation of the BTS towers. In order to filter out mobility patters not related to commuting, we only consider CDR data from Monday through Thursday. Similarly, all bank holidays were filtered. From the two months of traffic available for this study, we will use the data from October for the optimization process, and the data from November will be used to validate the results.

In order to guarantee privacy we implemented a set of elements: (1) All records were anonymized; (2) Data collection and anonymization was done by a third party that was not involved in the analysis; (3) No individual demographic data was available or requested for this study and (4) The information presented is always aggregated in order to further guarantee privacy.

5.2 GA and SA: Configuration

Genetic Algorithms and Simulated Annealing are used to search for the temporal intervals that best represent the times at which people commute using CDRs for the Madrid region. We carry out a total of four experiments: (1) the construction of $CM_{CDR}[H, W]$ using Genetic Algorithms and (2) using Simulated Annealing; and (3) the construction of $CM_{CDR}[W, H]$ using Genetic Algorithms and (4) using Simulated Annealing. The optimization process is the same in all cases, but while the first two use $CM_{NSI}[H, W]$ as the goal of the optimization, the second two use $CM_{NSI}[W, H]$.

For the experimental evaluation, we have used the JGAP implementation of Genetic Algorithms [13] and our own implementation of Simulated Annealing following the description presented in [3]. Both approaches use the CMTAR Algorithm to construct CM_{CDR} for each set of time slots considered, which we have implemented in Java.

Size	Temporal Range	Correlation
10	[20, 21][9, 16]	0.8050
20	[20, 21][9, 10]	0.8219
50	[20, 21][9, 10]	0.8219

Table 1: Optimization results when using Genetic Algorithms for the home-to-work [H,W] commuting matrix.

Size	Temporal Range	Correlation
10	[14, 16][20, 24]	0.9029
20	[15, 16][20, 24]	0.9029
50	[15, 16][20, 23]	0.9059

Table 2: Optimization results when using Genetic Algorithms for the work-to-home [W,H] commuting matrix.

In our experiments, GA uses a distributed architecture where a set of 16 genetic algorithms are run in parallel to explore the quality of different time intervals. Specifically, each process is initialized with a randomly generated population of a set of individuals. At every generation, the reproduction is carried out for a 90% of the total population; the crossover is executed with a 35% of pairs of the selected population by randomly selecting a gene in each individual and exchanging its content with its partner; and the mutation is executed for each gene with a probability of 1/12 and by randomly creating a new gene. The fittest individual is always moved to the next generation, and all the other individuals have a probability of being brought to the next generation proportional to their fitness value. Each process is executed on one core and runs in parallel with the other processes in our architecture of dual-core Intel processors. For our experiments we considered three different population sizes $10, 20, 50$.

On the other hand, the SA implementation starts with an initial temperature of $T_0 = 5$ and decreases its value with the function $T_{new} = 0.65 * T_{old}$ until a threshold value of $T_n = 0.1$ is reached. This cooling criteria allows us to explore a sufficiently large amount of temporal intervals without making the process too long. At each temperature, the SA evaluates three different time intervals and keeps the one that yields the best commuting matrix when compared to CM_{NSI}. Finally, we define as neighborhood solutions the set of temporal intervals that are within a range of four hours before and after the last time explored *i.e.*, $t_{new} \in [t_{old} - 4, t_{old} + 4]$. All the parameters here described were selected because they represented the best performing values across a large evaluation set.

5.3 Optimization Results

In this section, we discuss the results after running GA and SA for constructing $[H, W]$ and $[W, H]$ mobility matrices.

Table 1 and Table 2 show the results after applying GAs for the home to work and work to home commuting matrices, respectively. The tables shows the optimum Temporal Range obtained for each population size considered and the value of the fitness function

Temporal Range	Correlation
[21, 22][11, 16]	0.7863
[21, 22][12, 16]	0.7844
[21, 23][10, 16]	0.7840
[21, 23][14, 18]	0.7808

Table 3: Optimization results when using Simulated Annealing for the home-to-work [H,W] commuting matrix.

Temporal Range	Correlation
[10, 16][20, 23]	0.8949
[14, 17][20, 21]	0.8787
[15, 16][20, 23]	0.8781
[10, 17][21, 22]	0.8724

Table 4: Optimization results when using Simulated Annealing for the work-to-work [W,H] commuting matrix.

(given by Pearson correlation). The Temporal Range is expressed by two intervals, the first one indicates the temporal condition for the origin location and the second one for the destination location.

In Table 1 we observe that using CDRs to compute home-to-work commuting matrices for the region of Madrid we achieve correlation rates of up to 0.82 when compared to the NSI matrices (*ground truth*). This result was obtained with an initial population of 20 candidate solutions and for time slots that define origin as the interactions that took place between 8pm to 9pm of the previous day, and destination as the interactions that took place between 9am to 10am. Smaller populations yielded worse correlation results whereas larger populations did not improve the results. On the other hand, Table 2 shows that the work-to-home mobility matrices computed by GA achieve correlation rates when compared to NSI matrices of up to 0.9059 with an initial population of 20 individuals. In this scenario, the algorithm uses the calls made from 3pm to 4pm to detect the origin location and calls made from 8pm to 11pm (of the same day) to identify the destination location.

Tables 3 and 4 present the results obtained when using SA. For the $[H, W]$ commuting matrices, the best coefficient obtained was of 0.7863 with origin location detected between 9pm and 10pm of the previous day and destination location determined from calls made from 11am to 4pm. In the case of the work-to-home commuting matrices, the highest correlation coefficient is of 0.8949 with a temporal range of 10am to 4pm to detect the origin location and 8pm to 11pm to detect the destination location.

In general the correlation values provided by GA are better than the ones provided by SA. Also, we observe that in both cases, the work-to-home commuting matrices are better modelled from CDRs than the home-to-work (0.90 to 0.82 when using GA, and 0.87 to 0.78 when using SA). This result might be related to the fact that people make more cell phone calls during the day than early in the morning or at night, which provides a larger number of *geographical points* to model commutes from work-to-home than vice versa. Also, it might be an indication that the home-to-work commuting follows a less direct route (*e.g.*, taking kids to school), thus adding noise to the available data.

Finally, the average execution time for GA when considering the

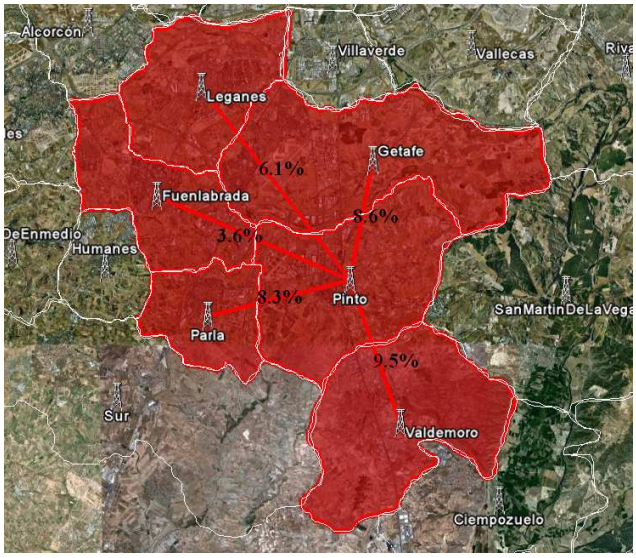

Figure 3: ***Visualization of the Commuting Matrix obtained for the municipality of Pinto in southern Madrid, showing the top five municipalities with the highest confidence value for the work-to-home commuting.***

best solutions obtained for a population of 20 is $2,890$ minutes, while the average processing time for SA for the best solution is $2,699$ minutes.

6. VALIDATION

The experimental results described in the previous section have shown that CDRs can be used to construct commuting matrices that are as good as the one provided by NSI.

In our context, the goal of the validation is to assess whether the time intervals identified for the $[H, W]$ and $[W, H]$ commuting matrices are valid to estimate the commuting matrices of other years, in order to show that CDRs can be used to generate commuting matrices without the need of NSI data. Ideally, the validation process would consider the commuting matrices obtained by the NSI for 2010 and CDR data from 2010, and validate the time intervals using the similarity between CM_{CDR} and CM_{NSI}. Nevertheless, so far, no commuting matrices for 2010 or 2011 have been published by the local NSI.

Considering that limitation, we implement a validation process that uses the 2009 $CM_{NSI}[H, W]$ and $CM_{NSI}[W, H]$ matrices and the November 2009 CDR dataset. The intervals we are going to use are the ones obtained by the GA-based optimization: $[20-21][09-10]$ for the home-to-work commute and $[15-16][20-23]$ for the work-to-home commute. Finally, the validation is done by calculating the similarity between the CDR matrix obtained and the NSI matrix using Equation (5).

Table 5 shows for both home-to-work and work-to-home commutes the Temporal Range used, the correlation values obtained during the Optimization process using the October 2009 CDR dataset, and the Validation correlation between CM_{CDR} and CM_{NSI} using the November 2009 CDR dataset (with its corresponding standard deviation). We observe that the Validation correlation coefficients are within a 10% of the correlation values obtained in the Optimiza-

	Temporal Range	Optimization(Oct09)	Validation(Nov09)
Home-To-Work	$[20, 21][9, 10]$	0.8219	0.765 ($\sigma = 0.46$)
Work-To-Home	$[15, 16][20, 23]$	0.9059	0.9322 ($\sigma = 0.16$)

Table 5: Validation results for the $[H, W]$ and $[W, H]$ commuting matrices obtained with November 2009 CDR data.

Municipality	% of Population	H-W Correlation	W-H Correlation
Madrid	50%	0.9995	0.5818
Alcobendas	2%	0.9885	0.8210
San Fernando	1%	0.9120	0.7411
Moraleja de Enmedio	0.0007%	0.0935	0.9895
Villa Conejos	0.0004%	0.1256	0.9972

Table 6: Individual Correlation values for H-W- and W-H for a set of representative municipalities.

tion process. It is noticeable that in the case of the work-to-home commuting there is a slight increment in the correlation, which, in line with the results discussed in the previous section, being probably caused by an increase of the CDR data available during the time slots considered.

These results show that, although with some differences, the optimization process provides a good approximation of the time intervals needed to compute commuting matrices, and as a result future commuting matrices can be directly estimated from CDR data. This allows for constructing O-D matrices with much more frequency at a fraction of the cost. The reason for the different values between the NSI- and the CDR-generated matrices is mainly caused by the fact that the NSI generates the commuting matrix strictly using individuals that have a declared work location. As a result, CM_{NSI} does not capture any non-work related mobility (which in itself is very difficult to capture using questionnaire-based approaches). Our CDR approach captures all types of mobility (work, leisure, shopping, students, etc.), so the fact that using CDR data we can not completely correlate the results with the NSI is because our matrix contemplates more situations and as such is more realistic.

6.1 Commuting Patterns by Municipality

The correlation coefficient between CM_{CDR} and CM_{NSI} represents an average value between each individual row-to-row correlation. In an attempt to understand the commuting patterns for individual municipalities, we compare the rows of each CDR-based mobility matrix with the rows of its NSI counterpart. Our objective is to do a preliminary study to understand whether there are stronger correlations between both matrices for specific municipalities. Figure 3 presents a visualization of the work-to-home commuting matrix CM_{CDR} for the municipality of Pinto using November 2009 CDR data. It shows the top five TARs with the highest support, *i.e.*, the top municipalities where people that work in Pinto live.

Table 5 shows that the standard deviations for home-to-work and work-to-home correlations are 0.46 and 0.16, respectively. These results reveal that there exist large differences in the correlation values across municipalities, especially for the home-to-work commuting patterns. Table 6 presents the individual correlation coefficients for a set of representative municipalities for the home-to-work and work-to-home commute, including the percentage of the population than they represent. We can observe that the home-to-work correlation coefficients are higher when the municipality has a large number of citizens, *i.e.*, larger cities tend to have more predictable home-to-work commuting patterns than smaller ones. On the other hand, larger municipalities tend to be less predictable in their work-to-home commutes (have smaller correlation values) than smaller towns. This is probably due to the fact that in larger cities citizens tend to do other activities once they get out of work as opposed to smaller towns where people tend to go directly to home. Thus, although on average home-to-work patterns appear to be less predictable than the work-to-home ones (as shown in Tables 1 and 2), that is only the case for small municipalities. In large ones, the opposite holds, whereby the larger the city, the more predictable the home to work mobility matrices are (when compared to the work to home mobility matrix).

These preliminary results seem to indicate that incorporating the size of the municipalities in the optimization process could improve the final correlation values. Also, we consider that having more data to generate the O-D matrix will, to some extent, mitigate the current limitations regarding the predictability of small municipalities (consider that because we only use Monday through Thursday, in the end we have 17 days of traffic for the optimization process).

7. RELATED WORK

The construction of O-D matrices has been typically studied by transportation and urban planning research. Traditional solutions are based on questionnaires and/or in the combination of questionnaires with traffic information. Such solutions typically focus on generalization techniques that construct matrices from partial data. The main approach used to obtain traffic data information is electronic toll collection[10]. This approach is limited because the information provided only reflects a partial view of the route. A possible solution for these limitations is the use of GPS data. In this case, the information contains complete routes but the amount of data available is even more limited [17]. The studies done up to now focus mainly on GPS data available from taxi or bus fleets[17] which highly limits the conclusions.

The use of CDRs to model commuting patterns solves to a large extent the previous limitations. A variety of studies can be found in the literature: Caceres et al.[4] uses GSM simulated traces to construct origin-destination data to measure the flow of vehicles, Zhang et al. [18] presents a model to transform cellular counts into vehicular counts in order to construct commuting matrices, and Sohn et al. [15] introduced cell phone probes in the network to identify trajectories and estimates O-D matrices using handoffs. Our approach has a set of differential factors with these previous studies: (1) we use CDR data that does not contain any handoff information. Handoff information consists on storing the sequence of towers used during a conversation and although they provide more information, cell phone operators do not keep such data due to privacy concerns (also consider that information would actually be useful only if using cell-phones was allowed while driving); and (2) our approach focusses on showing that the traditional questionnaire-based approaches for estimating O-D matrices can be approximated by the technique we present. While the state of the art mainly presents techniques to construct O-D matrices and assumes that the quality of the data will imply good results, in our case the technique we propose uses the information contained in questionnaire-based O-D matrices to tune the parameters. From this perspective, our work has elements in common with Calabrese et al. [5] in the sense that the validation of the technique is done with external O-D matrices. The difference in our case is that we also use that same information to identify the best parameters to construct O-D matrices with CDR in order to approximate the values of traditional approaches. This allows us to present a technique that can be adapted to capture the different cultural schedules of different urban areas.

Some authors identify the construction of O-D matrices using CDR as the identification of home and work for each user, using that information to aggregate origin-destination patterns. The work by Frias-Martinez *et al.* [7], Isaacman *et al.*[8] and Calabrese et al. [5] present algorithms to detect home and work by identifying highly used cell-phone towers. Nevertheless the use of such algorithms has strong limitations that affect the construction of O-D matrices, mainly: (1) the error introduced by the algorithms in the estimation of the locations (which in general is not measurable due to the lack of ground truth data); and (2) the fact that the coverage is limited by the availability of information for each user, *i.e.*, home and work can only be detected for individuals that have a minimum amount of interactions with their cellphone. Depending on the context, this requirement can filter more than 80% of individuals[7], with the corresponding bias in the final matrix.

8. CONCLUSIONS

Traditional methods for the estimation of mobility matrices suffer from a variety of limitations, mainly the bias of the information collected and the cost of gathering such information. To overcome these issues, we have presented a method based on the data collected by cell phone infrastructures to generate commuting matrices. In the literature we can find similar approaches, but in our case we have focussed our study on showing that we can replicate the information contained in questionnaire-based O-D matrices.

Our approach is implemented with CMTAR, a TAR-based algorithm designed to construct commuting matrices from CDR data. The combination of CMTAR with optimization techniques provides an approach that identifies which parameters need to be used to construct commuting matrices that are as similar as possible to the NSI matrices. Our experimental evaluation and validation has showed that we can compute commuting matrices with a high level of accuracy using CDR, and as a result our CDR generated matrices can be used for the same purposes as traditional matrices.

9. REFERENCES

[1] Agrawal, R., Imielinski, T., Swami, A.N. . Mining association rules between sets of items in large databases . In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, 1993.

[2] Goldberg, D. In *Genetics algorithms in search optimization and machine learning, Addison Wesley*, 1989.

[3] Bertsimas, D., Tsitsiklis, J. Simulated Annealing . *Statistical Science*, 8(1):10–15 , 1993.

[4] B. F. Caceres N., Wideberg J.P. Deriving origin destination data from a mobile phone network. *Intelligent Transport Systems, IET*, 1(1):15–26, 2007.

[5] Calabrese F., Di Lorenzo G., Liu L., Ratti C. Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing 10(4)*, pages 36–44, 2011.

[6] Frias-Martinez, E. and Karamchety, V. A Customizable Behavior Model for Temporal Prediction of Web User Access Sequences . In *LNAI 2703*, page , 2003.

[7] Frias-Martinez, V. and Virseda, J. and Rubio, A. and Frias-Martinez, E. Towards large scale technology impact analyses: Automatic residential localization from mobile phone-call data . In *Int. Conf. on Inf. & Comm. Technologies and Development (ICTD)*, 2010.

[8] Isaacman, S. and Becker, R. and Cáceres, R. and Kobourov, S. and Martonosi, M. and Rowland, J. and Varshavsky, A. Identifying important places in peoples lives from cellular network data. *Pervasive Computing*, pages 133–151, 2011.

[9] A. R. Kianmehr K. A fuzzy prediction model for calling communities. *Int. J. Netw. Virtual Organ.*, 8(1/2):75–97, 2011.

[10] Kwon, J., Varaiya, P. Real-Time Estimation of Origin-Destination (O-D) Matrices with Partial Trajectories from Electronic Toll Collection Tag Data. *Transportation Research Record no. 1923*, pages 119–126, 2005.

[11] Laarhoven, P.J.M., Aarts, E.H.L. Kluwer Academic Publisher. In *Simulated Annealing: Theory and Applications*, 1988.

[12] Mannila H., Toivonen H., Inkeri Verkamo A. . Discovery of Frequent Episodes in Event Sequences . *Data Mining and Knowledge Discovery*, 3(1):259–289 , 1997.

[13] Meffert, Klaus et al. . JGAP - Java Genetic Algorithms and Genetic Programming Package . In *http://jgap.sf.net*, 2008.

[14] J. Rodgers and W. Nicewander. Thirteen ways to look at the correlation coefficient. *American Statistician*, pages 59–66, 1988.

[15] K. Sohn and D. Kim. Dynamic origin–destination flow estimation using cellular communication system. *Vehicular Technology, IEEE Transactions on*, 57(5):2703–2713, 2008.

[16] U.S. Census Bureau. *www.census.gov*, 2011.

[17] Veloso M., Phithakkitnukoon S., Bento C., Fonseca N.,Olivier P. Exploratory study of urban flow using taxi traces. In *The First Workshop on Pervasive Urban Applications (PURBA)*, 2011.

[18] Zhang Y., Qin X., Dong S., Ran B. Daily O-D Matrix estimation using cellular probe data. In *89th Annual Meeting Transportation Research Board*, 2010.