Tackling the Neighboring Network Hit Problem in Cellular Data

Andrés Leiva-Araos
Pontificia Universidad Católica de Valparaíso
Valparaíso, Chile
andres.leivaa@gmail.com

Denis Khryashchev
CUNY Graduate Center
New York, USA

Héctor Allende-Cid
Pontificia Universidad Católica de Valparaíso
Valparaíso, Chile

Huy T. Vo
Department of Computer Science
CUNY City College
New York, USA

***Abstract*—Most humans today carry mobile telephones. According to the GSMA, daily there are almost 10 billion mobile connections in the world. About two-thirds correspond to obsolete technologies such as 2G and 3G. These devices automatically capture behavioral data from human society and store it in databases around the world. Nevertheless, the data capture has several challenges to deal with, even more if we consider the one that comes from old sources. To the best of our knowledge, all previous works eliminated erroneous data. Deleting traces in a time series can lead to deviations and biases in its analysis (e.g. computing the distributions of the distances when people's commute). In this work, we present two algorithms to solve the problem of the network of neighbors (NNH) and calculate the distributions of trips and distances with higher precision. The NNH problem arises when a mobile device connects to cellular sites other than those defined in the network design, complicating the analysis of space-time mobility. We use cellular device data from three cities in Chile, obtained from a mobile phone operator and duly anonymized. To test our results, we compare it with the Government's Origin-Destination Survey and use a novel method to generate synthetic data to which errors are added in a controlled manner to evaluate the performance of our solution. We conclude that our algorithms improve results compared to naive methods. with improvements in the count of trips and, mainly, in the distance distributions.**

***Index Terms*—mobile data, data cleaning, binary logic, network errors**

I. Introduction

Some cities are growing faster than their ability to adapt to change. This is particularly true in complex conurbations where the developing plan has complex technical, economic and political variables. One of the main aspects in which cities are in debt is in the planning and management of urban mobility; this is due to the increase of the population and the explosive growth of the automotive park. The design of urban policies for mobility has several inputs into consideration. The travel surveys are among them because they provide a lot of information about the city and its urban patterns: where people travel (and from), the purpose of the trip (e.g., commuting to work), means of transportation (e.g., car, bike, subway), time spent traveling, as well as other sociodemographic variables. However, surveys are expensive, static, require enormous effort to prepare them, and may have sampling biases or reporting errors [1], [2]. They represent just images of a dynamic phenomenon, and therefore, subsequent surveys only capture the big patterns and their changes.

Most humans today carry mobile telephones. According to the GSMA, daily there are almost 10 billion mobile connections in the world. About two-thirds correspond to obsolete technologies such as 2G and 3G [3]. These devices automatically capture behavioral data from human society and store it in databases around the world. Nevertheless, captured data has several challenges to deal with [4], even more if we consider the one that comes from old sources.

For more than a decade, mobile data has been used to analyze and understand the dynamics of mobility. Within these data, the most common and abundant are called *Call Detail Records* (CDR), which are used by telecommunications companies to bill their consumers. In general, CDR data have several uses as shown in surveys [4]–[6], and specific applications such as sensing the human mobility [7]–[10]. These data have the potential to help urban planners and policy makers thanks to their volume, spatial and time granularity of analysis (e.g., areas in the cities, or particular days of the year) However, in addition to well-documented problems [4], the quality of these data also suffers due to bad data as a result of, e.g. weak GPS signals, overloaded cell towers, or uncertainty in data collection. Most previous works simply suggest to discard these bad records [10], instead of fixing them. While this practice does not affect the overall data distribution, given the high number of total records, it can introduce significant biases in studies at a smaller scale.

To the best of our knowledge, all previous work was carried out in large areas, with vast volumes of information that reduced the impact of outlier data. Those studies are useful for developing policies and making aggregated decisions. However, the opportunity arises to develop studies and make decisions at much smaller geographical scales (e.g., commercial neighborhoods), or in small groups of people (e.g., immigrants). We propose a novel way of correcting errors and compute travel distributions in the time series represented by

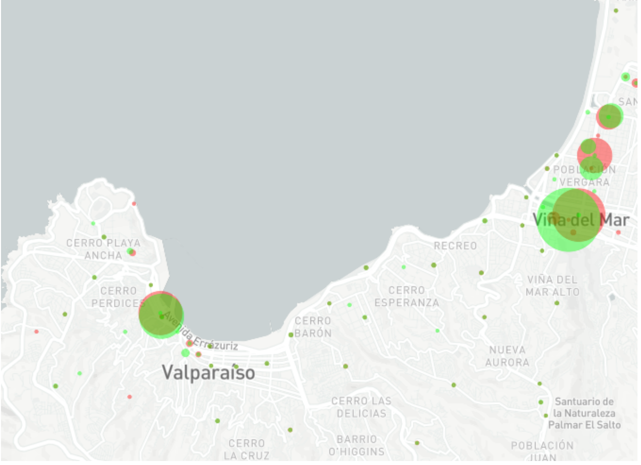


Fig. 1: Visualization of sites that acquire NNHs (red) and sites that lose NHs (green). Most of them (Top-K) are in sectors with a line-of-sight in the bay.

the CDRs in order to improve the state of the art in the analysis of data for small areas or groups of people. Through a number of case studies, we demonstrate that, due to the sparseness of data in these areas, simply discarding (bad) records can significantly affect travel patterns, which in turns introduce biases into subsequent mobility analysis.

In summary, the main contributions of our paper are:

- A novel methods for fixing erroneous call detail records and computing trips and distance using binary logic.
- A benchmark for our approach using synthetic data.
- A detailed evaluation using real-world scenarios based on data from three major cities in Chile.

The rest of the paper is organized as follows. Section II provides an overview of the problem and terminology in dealing with cellular data. Section III reviews the state of the art in the research using event-driven cellular data. Section IV presents our approach in solving of the Neighboring Network Hit problem including our fixing algorithm and its complexity. In Section V and VI, we describe the setup of our experiments and the evaluation results. Finally, in Chapter VII, we made our conclusion, and present the next steps and challenges.

II. Background

A. Terminology

We will use the following terminology throughout the paper:

Cellular Site. A set S of Base Transceiver Station (BTS) and antennas (or serving node), that are grouped to optimize the signaling [10], [11] in a specific coverage area. Normally in a site, there are one BTS per technology (2G, 3G, LTE, etc.) and several antennas per technology.

Mobile Connected Device (MCD). A physical object m that has an IP stack, enabling two-way communication over a network interface (e.g. mobile phones, tablets) that generates CDR (Call Detail Records) and/or IPDR (Internet Protocol Detailed Record) data.

Cellular Data. Most telephone networks generate records produced by a MCD m when it is turned on and interacts with the network. There are two type of cellular data; triggered by events or updates of the network [4].

Network Hit (NH). The recording of an event h_i generated by a Mobile Connected Device m in the network. It is associated with a cellular site $\{s_i, (s_i \in S)\}$, where S is the set of antennas in the study area. In our study we work with three different areas: Concepción, Santiago, and Valparaíso. The three of them represent a large variety of population and geography complexity.

Neighboring Network Hit (NNH). We define a neighboring network hit h' as a Network Hit where the site antenna does not correspond to the geographic place where the cell phone is located in. This is due to the network configuration, the layout and size of the surrounding buildings, the weather conditions and/or when the cells reach their maximum concurrent connection capacity.

Trace Data. A temporal sequence t_m of network hits h that belong to the same mobile connected device m within a day: $t_m : h_1 \rightarrow h_2 \rightarrow ... \rightarrow h_n$. And $T_i = \{t_1, \ldots, t_m\}$ is the dataset of the trace data of all mobile connected devices m within a day.

Input Dataset. The set D_i of CDR and/or IPDR records for the period i (usually a single day).

Trip or travel. All travel made on public roads with a purpose determined between two places (origin and destination) at a certain time of day; which can be performed in various ways of transport and consist of one or more stages.

B. Neighboring Network Hit (NNH) problem

Cellular networks are extremely complex communication systems. They are composed of different layers of abstraction, physical and logical, to make possible the flow of a multiplicity of messages between different points in the network. The first dimension of abstraction is the so-called physical layer (referred to the OSI layer model [12]), in which, as the name indicates, the physical variables of the network are relevant. Among the most important physical variables to take into consideration are: the type of antenna of the devices at the ends of the communication channel (e.g.: user terminal and the serving node or antenna), the power emitted by both devices, the frequency band used, the level of interference in the communication channel, and the physical conditions of the environment in which the devices are located (altitude relative to sea level, line-of-sight conditions between the antennas and the devices, amount and location of dimming points between both devices, etc.). It is in this physical layer or physical channel of communication where one of the more complex control phenomena of the cellular networks occurs. This phenomenon is the trade off and balance between maximization of network coverage and minimization of interference in the communication channel. This issue is more common in urban areas than in rural ones. Maximizing the level of coverage is important because it has direct impact on network costs [13], however in urban areas the complexity of the physical channel grows exponentially with the shape of the city, and the density and height of buildings. Due to this balance between coverage and

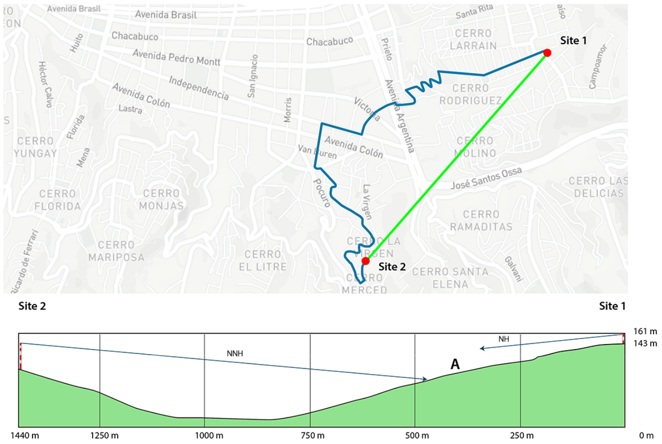


Fig. 2: Elevation profile between two cell sites in Valparaíso with actual heights, distances and downtilts.

interference, what usually happens in consolidated networks is that in urban areas the antenna coverage is bounded only a few blocks, which implies that most of the time a user (device m) should have service from an antenna near him/her. However, due to the difficulty in controlling the physical conditions, devices may connect with antennas that far exceed the average network distance.

The data quality is indexed to the type of technology (2G, 3G, LTE, etc.). When the main objective is the quality of service, e.g.: call drop rate close to zero, it is common to find cases of user's devices connected to distant antennas. A factor that influences the way that cell phones get connected to far away antennas is the shape of the city. Cities with bay shapes surrounded by hills allows that sometimes, when several external network factors take relevance, a user located at one end of the bay obtains service from an antenna located at the other end, more than 5 kilometers away. This phenomenon happens despite the network optimization for bounded coverage. Two of these external factors are the "mirror" effect of the sea, which facilitates the propagation of electromagnetic signals, and the clear line-of-sight between the device and the antenna. In Fig. 2, we can see the actual elevation profile between two nearby cell sites. The green line represents the linear distance between them (1.4 km), and the blue route corresponds to the optimal walking/driving path (2.9 km). A device located in A can perfectly connect to both sites if they have their azimuths facing each other and the down-tilts of the antennas with convergent angles. Sometimes, the minimum distance to travel from Site 1 to Site 2, divided by the maximum time spent, does not correspond with the maximum speed in the city; unequivocal symptom of an *NNHs*.

There is an important issue in the quality of data. Several works [4], [14] mentioned this issue as critical but none of them refers to its magnitude or its implications in the studies. Instead, they use well-curated data [7], [8], [15], [16], synthetic data [17], [18] or eliminate the outliers [10], [19], [20]. In all cases there is no analysis of the additional biases introduced to the data, nor the impact in important subsequent inferred ratios like average trips per user, distances, transportation mode, etc. Many of these adverse effects are diluted in large volumes of data when the analyzes are large-scale, but we do not know the effect when the granularity of the studies focuses at neighborhoods or small groups of people.

III. Related Work

Massive mobile phone data have been explored in various urban applications, particularly in understanding human mobility through daily patterns (e.g. [7], [21], [22]) as well as in special events (e.g. [9], [23], [24]). We refer the readers to [4] for a comprehensive review of the area. Nevertheless, most existing works assume that data have been either collected or generated based on a simple model of human mobility. For example, Isaacman et al. [17] introduced the WHERE (Work and Home Extracted REgions) modeling approach to produce privacy-preserving CDRs to simulate synthetic population. A subsequent DP-WHERE (Differential-Privacy WHERE) [18] model, yield good results but over the basis of using empirical probability distributions to predict behaviors from synthetic data. In terms of estimating the spatial distribution of the data, two common approaches are determining the geographical location of home and work places [25], and describing and evaluating a non-parametric Bayesian approach for identifying places from sparse GPS traces [26]. Since mobile phone network data does not provide accurate localization (the precision in urban areas is about 150 to 500m), several workaround have been proposed including look at history for recurring locations [21], and look at handover during calls [27]. However, neither of these approaches take into account the travel and elevation profile being targeted in our work.

Recent works [10], [22], [28], [29] also model mobility data by estimating types of activities in using both spatial and temporal profiles. During the week, the call activity of a residential region, a commercial or a business is different. It may be possible to derive a classification from the call activity profile of a region, thus allowing to classify regions as “residential”, “commercial” or “business”. In [10], the authors propose virtual placements for the antennas and improves city coverage of CDR data by geo-locating devices to more city areas than using standard methods city areas than using standard methods. Nevertheless, the uncertainties in the user’s location and time, if solely derived from CDR data, can be relatively large. This is due to the low frequency of user’s localization update and the spatial resolution of mobile phone network data. Our work focuses on reducing the uncertainty by fixing outlier records.

Limitations of event-driven data. Event-driven data are generated only when the user takes some action, i.e., sends an SMS, makes a call, sends an email, etc. Thus, the location of the user is updated only when these events occur. The problem is that certain types of urban patterns need very frequent location data. Some approaches proposed to solve this problem are: sampling only highly active users and sampling

Internet usage data. In the first case, the main problem is how to choose users that represents a good sample of citizens' behavior, and in the second case, the sources of data not always have the expected sampling frequency. Although there is a high penetration rate of smart phones, it is complex to integrate data from multiple companies to achieve a better coverage.

Data Quality. The clear majority of studies use data sets that are properly cured and preprocessed [15]. Data quality has been brought up as a critical issue many times but little about its magnitude and/or implications has been discussed in the studies. [30] refers to the quality of the GPS receiver algorithm might also lead to inaccurate GPS positions. The same problem occurs in a cellular network which is subject to a complex configuration process, in many cases manual, at the level of its radio access network (RAN) and core network (CN) components.

In [31], the authors argue that data cleaning must be an integral part of the inherently iterative data analysis cycle: data cleaning must be applied on the fly. In their study they use GPS data from New York taxis. The study showed that 4.6% of the data represented ghost trips, and an unidentified but significant number of obviously erroneous data such as taxis geo-localized in rivers, in the ocean, and even outside North America.

In our case, given the large volume represented by the cellular data, sampling appears as an option of doubtful quality [4]; this is aggravated if within the data there are outliers that are not evident without a prior analysis. Outlier detection is an iterative process that needs to combine domain knowledge, thus, human in the loop, with a set of automatic metrics and tools, which is the proposed framework in this paper.

IV. Our Approach

At the core of our approach, we developed two algorithms using Crisp Logic [32]: *Network Hit Fix Algorithm (NFA)* and *Trips Counting Algorithm (TCA)* that are used for fixing incorrect data, and computing travel patterns in small, sparse-sampled areas, respectively.

A. Network Hit Fix Algorithm (NFA)

NFA aims to fix the *NNH* records in the data set and reduce the biases in trips and distance computed later due to deleting records. We define a set of 29 rules containing thresholds used for defining splitting predicates (constraints). For example, we use contextually defined distance (walking or driving) variable between two sites ($distance(s_i, s_j)$) and minimum speed of a device m ($min\_vel_m(s_i, s_j)$, where i and j are two consecutive events in the time series) moving from two points, s_i to s_j. We define several thresholds in each variable in order to classify a specific *NH* as correct or not (*NNH*). Due to the nature of this approach, any variation in the variables can lead to a completely different categorization. The crisp set is defined in such a way as to dichotomize the individuals in some given universe of discourse into two groups: members and nonmembers. The classification of individuals can be done using a indicator or characteristic function:

$$\mu_A =: E \to \{0, 1\}$$

$$\mu_A = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases} \tag{1}$$

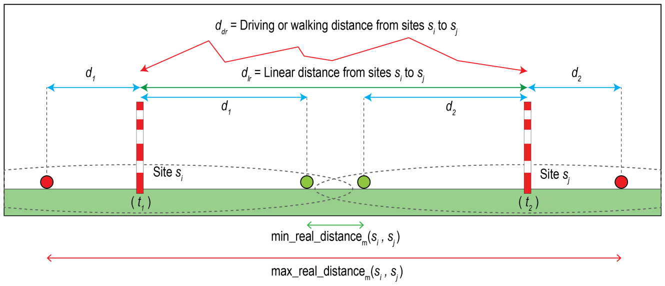

d_{dr} = Driving or walking distance from sites s_i to s_j
d_1
d_{lr} = Linear distance from sites s_i to s_j
d_2
d_1
d_2
Site s_i
Site s_j
(t_1)
(t_2)
$min\_real\_distance_m(s_i, s_j)$
$max\_real\_distance_m(s_i, s_j)$

Fig. 3: The minimum velocity solution approximation

Overview. The *NFA* carries out its work based on a previous parametrization and the application of chained binary filters. Every record is subjected to a series of logical tests to determine if it needs to be corrected or not using a rolling window over the series. The algorithm does not delete records but instead modifies the cellular site (geographical position) of those it considers erroneous. All of the previous works remove such kind of records, which introduce biases when analyzing small areas or small panels of devices (m).

The minimum travel speed (V_m) of a device m will be given by one of the following cases, as shown in Fig. 3. Depending on the coverage of the cells of a tower, the sequence of two consecutive *NHs* (h_i and h_{i+1}) in the time series may have different potential distances in reality. The distances d_1, d_2 and d_3 (where $d_2 > d_1 > d_3$) potentially represent the extreme cases for a combination of two consecutive *NHs* at two different sites s_1 and s_2. The sequence of two successive *NHs* has, at any of the potential distances, a single known time $t_{h_{1\to2}}$. Since we do not know at what exact moment the movement began and ended, $t_{h_{1\to2}}$ represents the longest travel time for all cases. That means the object m was in motion for the entire time $t_{h_{1\to2}}$. In this case, we will conclude that the speed we get from dividing the distances d_1, d_2 or d_3 by the time $t_{h_{1\to2}}$ will always be minimum V_m for the section distance between s_1 and s_2 cellular sites. The minimum speed V_m is compared parametrically with the context of the location of the sites. Nearby sites in the city center require to travel at maximum urban speeds. On the contrary, distant sites allow comparable speeds equivalent to that of a commercial flights.

Binary Filters. There are five groups of binary filters.

1) *Marginal displacements* In Group 1, we seek to correct errors in the data (consecutive records with the same timestamp but in different cell sites) or eliminate marginal displacements (which do not represent a trip).
2) *Network line-of-sight* In Group 2, we fix problems associated with the physical configuration of the network.

In this case, we seek to correct *NH* associated with physically close sites. Both sites have line-of-sight, but their linear distances are less than minimum walking distance displacements (see Figure 1).

3) *Urban min velocity* In Group 3, there are eight filters to discriminate V_m problems higher than those observed in urban areas.
4) *Suburban min velocity* In Group 4, there are nine filters to search for speed deviations in major urban areas (e.g., two adjacent or nearby cities). In this case, we compare the speeds with those that, on average, can be observed on interurban roads or highways.
5) *Inter region min velocity* Finally, in Group 5, with eight filters, records whose minimum speeds exceed those of commercial flights are detected and corrected. The latter is a typical case that occurs when the core network team insert repeated cell site identifiers in the management systems.

Algorithm. The *NFA* considers two groups of parameters as input. The first group of parameters considers the physical distances between the sites to discriminate proximity. Optimal distances are used to drive or walk between the different sites, the linear distances between the sites, and their heights above sea level and the height of the support (for example tower, building, etc.). The second group corresponds to setting parameters. These include the average speed in urban areas, the average speed between interurban areas, the distance between sites and the elevation ratio between nearby sites. The algorithm uses a heuristically defined variable window to analyze the time series. We got all driving and walking distances from the GoogleMaps API. We built a $n \times n$ distances matrix (where n is the number of sites cell phones included in the set S), so here we have a significant improvement over *rectilinear distance* ($L1$ *distance* or ℓ_1) approaches used in previous studies [10] [33]. We rely on the assumption that two consecutive *NHs* in different sites are performed at comparable distances with the linear distances between sites. We do not know the exact moment when a device starts or finishes moving, but we do know that it does so by changing the cell site over time. The data analyzed do not include the hand over records while the device is moving and connecting from one site to another, adding complexity to the analysis. The minimum speed V_min of device m between two sites, s_i and s_{i+1} does give an important clue about potential *NNHs*. This speed is calculated using the maximum displacement time between two sites in the time series. We use a rolling window analysis on the time series; then we compute V_min to sites that are not necessarily consecutive. A more formal definition is provided below.

$$V_{min}(s_i, s_j) = \begin{cases} \frac{d_w(s_i,s_j)}{t_2-t_1} & d_{dr}(a_i,a_j) < K \\ \\ \frac{d_{dr}(s_i,s_j)}{t_2-t_1} & d_{dr}(a_i,a_j) \geqslant K \end{cases} \tag{2}$$

where, d_w and d_{dr} are the GoogleMaps optimal distances for walking and driving respectively, between sites s_i and s_j; t_i, t_j are the observed times between two events in the series, and K is a tuning parameter that represents the distance threshold between walking and driving. We do not know the real speed of the devices but we can make the assumption that their displacements (when they do), can not be at speeds higher than those observed in the different urban and interurban contexts (as a reference the average speed in the Great Valparaíso is 32 Km / h). The *NFA* algorithm is shown in Algorithm 1.

Algorithm 1: NFA Algorithm

```
Input  : dataset ← $D$
Input  : walking/driving threshold ← $K$
Input  : set of urban thresholds ← $ut$
Input  : set of suburban thresholds ← $st$
Input  : set of regional thresholds ← $rt$
Input  : GoogleMaps distances ← $GMD$
Output: fixed records dataset $F$
def dist($s_1$, $s_2$, $GMD$):
    compute optimal walking/driving distance
    return(dist)
def v_min($s_1$, $s_2$, $dist$):
    compute min velocity
    return(vm)
def fix_nh(D, K, ut, st, rt):
    while records in D:
        if $record_i \neq record_{i-1}$:
            if $v\_min(A,B) == 0 \vee dist(A,B) < K$:
                $new\_site \leftarrow apply\,(Group\ 1\ rules)$
            elif $v\_min(A,B) < ut \wedge dist(A,B) < ut$:
                $new\_site \leftarrow apply\,(Group\ 2\ rules)$
                $new\_site \leftarrow apply\,(Group\ 3\ rules)$
            elif $v\_min(A,B) < st \wedge dist(AB) < st$:
                $new\_site \leftarrow apply\,(Group\ 4\ rules)$
            elif $v\_min(A,B) < rt \wedge dist(AB) < rt$:
                $new\_site \leftarrow apply\,(Group\ 5\ rules)$
            $fix\_record(record_i \leftarrow new\_site)$
    return (F)
```

B. Trips Counting Algorithm (TCA)

TCA is designed to count trips per device m from traces. The same Binary Logic technique is used to determine whether or not the traces contained in the time series represent a change of location that is considered (or not) as a trip. The first process consists in compute traces from the NHs events (original and fixed). Then, we used Binary Logic to determine whether or not the traces contained in the time series represent a change of location that is considered (or not) as a trip. From this computation, we get the average trips per device m per day. The total trips average per day for a specific area is compared it with the Origin-Destination Survey (see section VI-A) from the Chilean Transportation Authority. This process also compute the travel distance based on the optimal walking and/or driving distances from GoogleMaps. This is a significant improvement from previous work while all of them used *rectilinear distance* ($L1$ *distance* or ℓ_1) methods. The TCA algorithm is shown in Algorithm 2.

```
Algorithm 2: Trips Counting Algorithm
  Input  : dataset $F$
  Input  : threshold parameters $\leftarrow \ p$
  Input  : GoogleMaps distances $\leftarrow \ GMD$
  def count_trips($F, \ p$):
      while records in $F$:
          $traces \leftarrow compute\ traces$
      for $t$ in traces do
          if $site \neq old\_site$:
              $t\_trace += 1$
              $t\_trips, t\_distance \leftarrow compute\ trip\,(p)$
      return $(t\_trace, t\_trip, t\_distance)$
```

V. Experiment Setup

To test our algorithms, we use a twofold approach. In both cases, we look to correlate the results in terms of average trips per person (device m) per day with the Origin-Destination Survey (OSD for short), provided by the Chilean Transportation Authority. First, we test them using real data sets that cover 12 days (212 million records, from four different days and three different cities). We execute the *NFA* and *TCA* algorithms with different parameters on the real data looking for a correlation with the ODS outputs. Then, in a second step, we created synthetic data to which we introduced 2% of artificial noise (*NNHs*) to fix it and ensure proper operation. Having tested the algorithms, we selected from the real data a small portion of most affected records by the *NNHs* problem, and apply the *TCA* algorithm in order to compute the trips in small samples of m devices correctly.

To evaluate the performance of *NFA* and *TCA*, we compare the results they produce against the Origin-Destination Surveys available in the Chilean Transportation Authority database. These studies also provide us with parametric data such as the average speeds in the city suggesting the different contexts (time and type of day). To specifically test *NFA*, we used synthetic data to which we introduced controlled noise, seeking to eliminate it when applying the algorithm.

A. Input Data Description

The data used represents a time series with sparse data. There are few records per unit time (typically one day). The mean NHs observed varies between 20 to 45 per day, one peer every 30 or 70 minutes (see Table I).

We can see an important amount of devices with only one *NH* (no traces at all). There is a high variance in the *NHs* in all cases, even though we filtered the devices with just one record (10-60%). This is because there are devices with high mobility during the day like machine taxi or bus drivers, or machine (m2m[1] devices. When applying a filter to the deviated records, we obtain means of *NHs* close to 35 with lower standard deviations 27-28. In Figure 4 the boxplot shows this behavior applied before and after filtering the outliers is observed. Given the imprecise geolocation nature of the Cellular Data, we do not know exactly where the device m is located when recording the event h at time t. As seen in Fig. 3, the trajectory of m may be before, after or between two particular sites. There are two extreme cases: first, the distances d_1 and d_2 are before and after the sites (case of maximum distance, red points), and that d_1 and d_1 are between the sites (case minimum distance, green points).

	Records statistics after filter devices with NH = 1					
City:Date	Count $(\times 10^6)$	μ NH / device	M_e NH / device	σ	Dev. NH>1 $(\times 10^2)$	Dev. NH=1 $(\times 10^2)$
S:17/01/01	42.8	30.0	20	64.4	1,409	967
S:18/01/01	52.2	39.2	41	59.8	1,330	541
S:17/06/03	60.4	32.0	22	66.1	1,887	1,145
S:18/01/03	53.9	30.6	28	45.2	1,763	715
V:17/01/01	10.3	41.2	44	62.1	250	19
V:18/01/01	11.9	37.8	35	83.7	314	146
V:17/06/03	9.7	46.6	45	122.3	209	15
V:18/01/03	8.8	29.6	22	65.6	299	143
C:17/01/01	5.1	39.0	38	79.2	131	14
C:18/01/01	6.8	37.3	34	36.0	183	92
C:17/03/06	6.5	43.4	45	71.3	150	14
C:18/01/03	6.6	28.8	20	45.1	230	118
	275,0					

TABLE I: Statistics of raw and filtered data.

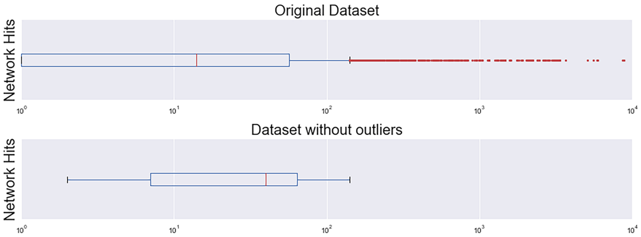


Fig. 4: Outcome of filtering the data from outliers

In this analysis there is a fact that:

$$d_{dr} > d_{lr} \quad \forall \quad (s_i, s_j) \in S \tag{3}$$

Where d_{dr} is the contextually driving or walking optimal distance from GoogleMap API, and d_{lr} is the linear distance from cellular sites s_1 and s_2, obtained using the equirectangular approximation method. Using d_{dr} instead of d_{lr}, we are increasing the algorithm precision. Then, the maximum and minimum real velocities of m between two sites s_i and s_j are given by:

$$max\_real\_vel_m(s_i, s_j) = \frac{d_{dr} + d_1 + d_2}{t_2 - t_1} \tag{4}$$

$$min\_real\_vel_m(s_i, s_j) = \frac{d_{dr} - (d_1 + d_2)}{t_2 - t_1} \tag{5}$$

[1] Machine to Machine device

	Inhabit-	Travels	Sampled	Year	Travels
	ants	per day	homes	Study	/ day
	(×1000)	(×1000)	(×1000)		/ person
Gran Santiago	6,652	18,461	18,264	2012	2.78
Gran Valparaíso	928,6	2,248.1	6,800	2014	2.27
Gran Concepción	N/A	N/A	8,400	2015	N/A

TABLE II: General status of ODS in the great areas

Consequently, the *minimum velocity approximation* (V_m), is given by the fact that if we consider $V_m(s_i, s_j) \Leftrightarrow min\_real\_vel_m(s_i, s_j)$ as a worst but perfectly possible case, we can find that if:

$$V_m(s_i, s_j) > P \quad \Rightarrow NH_i \quad is \quad NNH_i \tag{6}$$

where P is a set of contextual velocity defined in the geographical parameters.

VI. Evaluation Results

A. Origin-Destination Survey

The central objectives of the *ODSs* are, first, collect detailed information about travel that are carried out in a specific area and of the people who do them. And in second term, meet the requirements of information for the strategic transport model for the city [34]. The *ODSs* are costly, slow and infrequent studies (Gran Concepción has a 350-day execution plan and the last time it was done was in 1991 (see Table II). To compare with the *ODSs*, we need to compute travels from traces. Not always a single trace represents a single travel. It is frequent that in a single travel there are several NHs with the same site. To control this, we define a heuristic threshold parameter p, which represents the time in the same site (see Algorithm 2). This parameter considers as a single travel those traces with an stop of less than p minutes in the same site. In Figure 5, we see a device taken at random with 19 traces but only 8 travels denoted as T; The shorter the horizontal lines, the greater the probability that the device m is passing through the cell site s_i on a single trip T.

	Total	People	Total	Average travels
	homes	×1000	travels/day	/ day
	×1000		×1000	/ person
Con-Con	13.8	47.4	118.6	2.50
Viña del Mar	109.4	321.8	852.2	2.67
Valparaíso	98.4	259.1	697.3	2.36
Quilpué	49.1	165.3	348.4	2.11
Villa Alemana	36.5	135.0	231.6	1.72
Total	307.2	928.6	2,248.1	2.27

TABLE III: Gran Valparaíso 2014-15 O-D Survey detail.

We made several experiment over twelve real dataset with 275 million record in total, consisting in three macro areas (Gran Santiago, Gran Valparaíso, and Gran Concepción), and for each of them four different days: two January 1^{st}, 2017 and 2018, and two $BigMonday^2$, March 6^{th} 2017, and March 1^{st} 2018 (see Table I), with different parameter configurations. In Figure 8 we show the *NHs* density in four specific areas in Valparaíso in January 1^{st} 2018. We can see high activity at the beginning of the day (3:00 am) in New Year's celebration areas near the coast, and a completely different activity in residential areas outside the city.

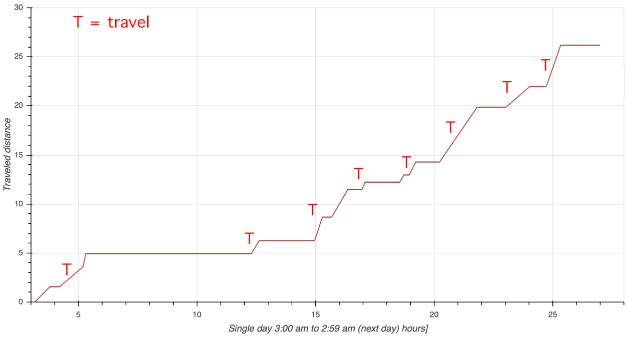


Fig. 5: Traces and travels for a random device

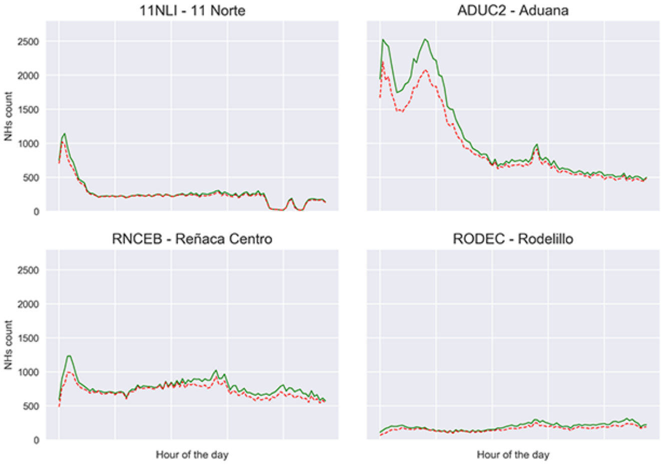


Fig. 6: New Year activities at 4 sites using real (green) and synthetic (red) data.

The idea was to approximate to the average trips values of the ODS. We can see some of the experiment results in Table IV, in this case for Gran Valparaíso data.

We modeled some of the experiments with real parameters of the city ($E2, E5, E8$). In some cases we force some parameters to see their correlation with the results ($E4, E9$). The results show us an approximation with the expected results. As an example, considering maximum speeds of displacement in the city between 25 and 30 km/h, and assuming movements smaller than 4 blocks ($\approx$ $500m$) as no−trips, it gives us results between 20 and 25% better in the calculation of trips. The travel algorithm parameter (p), also turned out to be sensitivity. For instance, increasing from 18 to 24 average displacement minutes per trip (quite conservative figures for congested cities) we obtained travel rates of 2.37, comparable with the ODS (2.27) in the case of Gran Valparaíso. We got similar and comparable results for Gran Santiago data. The next validation we made was to process the data using a basic

[2] First Monday of each year in which classes and work activity begin after holidays

	Parameters					January, 1st 2018				March, 6st 2017			
	ut1	st1	ut2	st2	ut	Fixed NNH	% fixed records	Traces	Trips	Fixed NNH	% fixed records	Traces	Trips
Original						0	0.00	7,96	2.53	0	0.00	7.08	2.48
E1	35	120	0.3	5	400	412,235	3.71	6.95	2.42	259,370	2.70	6.27	2.38
E2	30	120	0.3	5	400	432,301	3.89	6.90	2.42	271,720	2.83	6.23	2.37
E3	25	120	0.3	5	400	457,942	4.12	6.83	2.39	288,164	3.00	6.18	2.36
E4	20	120	0.3	5	400	494,580	4.45	6.74	2.39	311,901	3.25	6.11	2.35
E5	35	120	0.4	5	400	528,122	4.75	6.77	2.35	322,591	3.36	6.14	2.36
E6	30	120	0.4	5	400	547,757	4.92	6.72	2.40	334,780	3.49	6.10	2.35
E7	30	120	0.5	5	400	631,110	5.67	6.59	2.39	375,286	3.91	6.01	2.34
E8	25	120	0.5	5	400	655,735	5.89	6.53	2.38	391,382	4.08	5.96	2.33
E9	20	120	0.5	5	400	691,407	,6.22	6.44	2.37	414,402	4.32	5.89	2.32

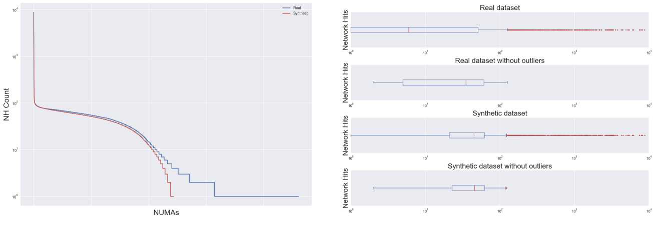
TABLE IV: Some experiments results for Gran Valparaíso data.

fixing and deleting the wrong record. We tried to emulate previous works. We got a trip rate of 2.49; just 4.6% over the fixed data. But the great difference was in the distances traveled rising from 21.0 to 28.8 km (37% plus in the case of deleted records dataset).

B. Synthetic Data

After the validation of the results against the survey, we developed a method to create synthetic data based on the statistical behavior of the real data. First, we grouped the records of the real data set based on date, time, duration, and hashed phone number collecting the pairs of cellular sites visited by the device during one event (e.g. phone call). We filtered out events that had more than 2 cellular sites as outliers (0.01% of the data set). Next, for every pair of cellular sites, we computed the distributions of event time, duration, and hashed phone numbers per every 15-minute interval within a day to capture temporal patterns efficiently. Then, we generated the synthetic data randomly pulling values from the distributions matching the time, duration, and hashed phone numbers with the cellular sites. Finally, we added 2% of noise generated in a similar fashion based on the noise detected in the real data.

Statistical analysis of the synthetic data shows that they exhibit a suitable behavior to represent reality (see Fig. 7). The same promising results are observed when comparing cellular sites with typical activities during the New Year's night as shown in Fig. 8.



(a) Density of NH per device

(b) Outliers analysis

Fig. 7: Comparison between real and synthetic data for validation. After ensuring an adequate data behavior, we introduced controlled noise to evaluate the performance of the NNHF.

We applied the *NFA* algorithm to the synthetic data using the $E8$ group of parameters, and we obtained 3.38% of fixed record. Then we introduced 2% of additional noise (based on the real error data profile). After a new excecution, our algorithm fixed 5.56% of the records, with almost 100% of the introduced *NNHs* fixed. With high confidence, we can say that the results are comparable to the published surveys, thus, allowing us to move forward with the next challenges of our research.

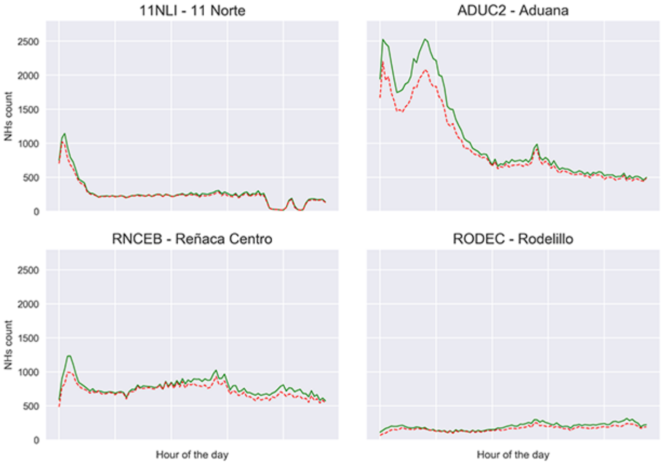


Fig. 8: New Year activities at 4 sites using real (green) and synthetic (red) data.

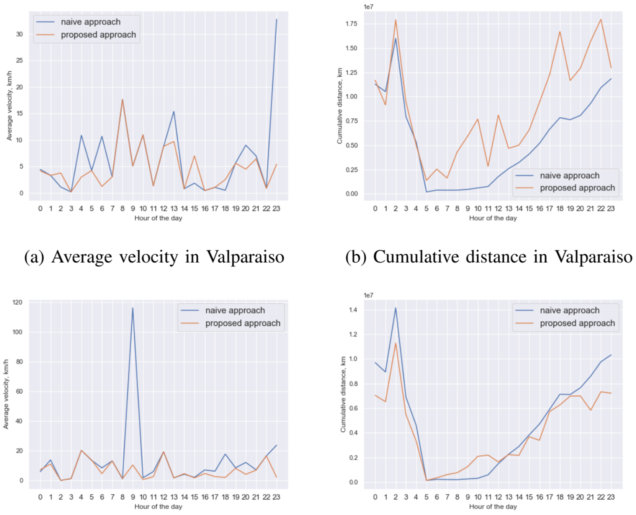
C. Applying NFA and TCA to small groups of data

The last experiments we ran were to apply the algorithms to small groups of data; i.e.: small groups of people, or small areas in the city. With this we wanted to demonstrate two things not addressed in previous work. First, our algorithms are capable of computing trips with comparable precision in data subset (even if they are affected by *NNH* type noise). And second, demonstrate that previous approximations present important deviations in the distances computed for the trips, introducing biases in further steps.

Computing better distance distributions The experiment consisted of applying the algorithms to the full dataset of original and corrected data. We expected to see a reduction of the distances because the elimination of the *NNH*. Even though the GoogleMap API increases the distances between sites by making the paths more real and accurate than previous methods ($L2$), this allows, on the other hand, to detect more

NNHs. After fixing the time series, the accumulated distance is reduced. In Fig.11, we can see some of the results comparing the naive approach (blue lines) and ours (orange lines).In terms of magnitudes, the reductions were 17.4%, 19,8%, and 27,2% in Santiago, Valparaíso, and Concepción respectively.

Analysing travels and distance distributions in small groups The last experiment consisted in selecting a group of pair of antennas with height rate of *NNHs* between them, and applying two filters: more than 1,000 *NNHs* and more than 2 Km distance between the two sites. For these combination of sites (s_1 and s_2) we selected all the devices m that passed through one of these combinations, respecting the direction of movement (from s_1 to s_2). When m starts a trip we compute $trips+ = 1$ independently if in that travel m gets a wrong *NH* (that means a *NNH*). So the impact should be more significant in distance computation than travel counting. That is why the travels distributions look similar (see Fig.9), but there are significant differences in distances distributions (see Fig.10) between the naive approach (blue) and ours (orange). In all cases, the distance are reduced at least in 20%.



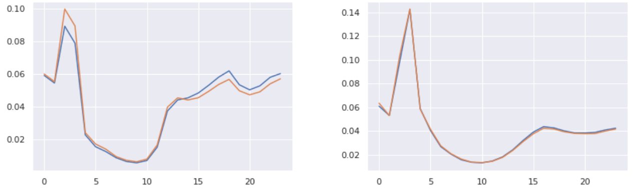
(a) Average velocity in Valparaiso

(b) Cumulative distance in Valparaiso

(c) Average velocity in Concepcion

(d) Cumulative distance in Concepcion

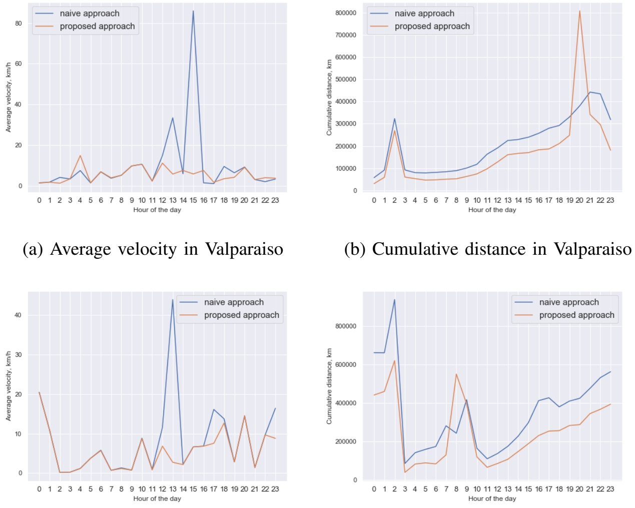
Fig. 11: Naive and proposed approaches evaluated on data for March 6, 2017



(a) Trips distributions in Valparaíso

(b) Trips distributions in Concepción

Fig. 9: Trips distributions computed with naive and proposed approach - March 6, 2017 and January 1,2017 respectively



(a) Average velocity in Valparaiso

(b) Cumulative distance in Valparaiso

(c) Average velocity in Concepcion

(d) Cumulative distance in Concepcion

Fig. 10: Naive and proposed approaches evaluated on data for January 1, 2017

VII. Future Work and Conclusion

The work done leaves us with several conclusions and future experiments. Although many works have been proposed using data generated by mobile traces, according to our best knowledge, all of them have worked with well curated or synthetic data eliminating the potential impact of wrong records. In our study, we looked for understand the erroneous *NHs* (*NNHs*) and how they affect the outcomes of the analysis. Previous work just detected the obvious problems in the data, but due to the nature of the network there are a lot more in it. Both eliminating or keeping these wrong records without fixing them, impact further analysis. We demonstrated that the incorrect management of these kind of problems introduces biases in the analysis; few in the trips counting but significant in the traveled distances distributions.

Regarding *EOD* we conclude two things. First, they are expensive, slow, static and very sporadic over time. In addition, the survey sample of respondents seeks to be statistically representative of the city and not of subgroups or sub-areas of interest. And second, one of the main uses of them is given by the need for data to carry out urban planning. In line with this, the distances traveled and the number of travels (and their modal partition) are directly related to transport planning, pollutant emissions, etc. Our work is aligned in solving these problems by being able to deliver results in smaller areas or groups and by substantially improving the computation of distance distribution.

Several years ago, Telcos begun to store data from the core of the network. This data has been becoming more complex with time, and today is much more dense (600+ *NHs* per device per day) and accurate (in the geo-location of events). However, it is more challenging and expensive capture, store, and process it. Also, like any new source, it is not exempt from the typical problems of uncured raw data. Our work will undoubtedly help reduce these times. For some more time, the old CDRs from billing and mediation process will continue to contribute to the study of mobility and urban semantics, which will require better management and understanding of them.

We decided to use binary logic as a simple way to understand a complex phenomenon. In future work the application of fuzzy logic or neuronal networks approaches should give us even better results not only in travels and distances distributions, but also in dimensions such as modal partition, and purpose of the trips. The compared benefits are enormous, even more so if we consider the possibility of carrying out the analyzes and obtaining conclusions in near-real times.

Another re-opened area of investigation is the generation of synthetic data. With the new regulations that are emerging around the privacy of the data (e.g., GDPR), there is a real probability that in the future there will be greater restrictions on the use of data [35]. This is an opportunity to deepen in the development of new methods to create synthetic data (based on the understanding of the real one). If we have abundant synthetic data, it is possible to carry out many research projects, which do not put people's privacy at risk. Once these studies reach adequate levels of maturity in relation to obtaining answers to their research questions, progress can be made in the use of real data. At this point, the promise of value will be greater than the possible vulnerabilities.

VIII. Acknowledgements

This work was supported in part by Alfred P. Sloan Foundation G-2018-11069, National Science Foundation Award 1827505, and Vingroup Innovation Award VINIF.2019.20.

References

[1] R. M. Groves, "Nonresponse rates and nonresponse bias in household surveys," *Public opinion quarterly*, vol. 70, no. 5, pp. 646–675, 2006.

[2] M. Kuwahara and E. C. Sullivan, "Estimating origin-destination matrices from roadside survey data," *Transportation research part b: methodological*, vol. 21, no. 3, pp. 233–248, 1987.

[3] G. Intelligence, "Definitive data and analysis for the mobile industry," February 2019. [Online]. Available: https://www.gsmaintelligence.com/

[4] F. Calabrese, L. Ferrari, and V. D. Blondel, "Urban sensing using mobile phone network data: a survey of research," *Acm computing surveys (csur)*, vol. 47, no. 2, p. 25, 2015.

[5] V. D. Blondel, A. Decuyper, and G. Krings, "A survey of results on mobile phone datasets analysis," *EPJ data science*, vol. 4, no. 1, p. 10, 2015.

[6] G. Pan, G. Qi, W. Zhang, S. Li, Z. Wu, and L. T. Yang, "Trace analysis and mining for smart cities: issues, methods, and applications," *IEEE Communications Magazine*, vol. 51, no. 6, pp. 120–126, 2013.

[7] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *nature*, vol. 453, no. 7196, p. 779, 2008.

[8] 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.

[9] F. Calabrese, F. C. Pereira, G. Di Lorenzo, L. Liang, and C. Ratti, "The geography of taste: Analyzing cell-phone mobility and social events." in *Pervasive*, vol. 10. Springer, 2010, pp. 22–37.

[10] E. Graells-Garrido, O. Peredo, and J. García, "Sensing urban patterns with antenna mappings: the case of santiago, chile," *Sensors*, vol. 16, no. 7, p. 1098, 2016.

[11] F. Pinelli, G. Di Lorenzo, and F. Calabrese, "Comparing urban sensing applications using event and network-driven mobile phone location data," in *Mobile Data Management (MDM), 2015 16th IEEE International Conference on*, vol. 1. IEEE, 2015, pp. 219–226.

[12] H. Zimmermann, "Osi reference model–the iso model of architecture for open systems interconnection," *IEEE Transactions on communications*, vol. 28, no. 4, pp. 425–432, 1980.

[13] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3gpp heterogeneous networks," *IEEE Wireless communications*, vol. 18, no. 3, 2011.

[14] E. De Jonge, M. van Pelt, and M. Roos, "Time patterns, geospatial clustering and mobility statistics based on mobile phone network data," in *Paper for the Federal Committee on Statistical Methodology research conference, Washington, USA*, 2012.

[15] C. Song, T. Koren, P. Wang, and A.-L. Barabási, "Modelling the scaling properties of human mobility," *Nature Physics*, vol. 6, no. 10, p. 818, 2010.

[16] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi, "Human mobility, social ties, and link prediction," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. Acm, 2011, pp. 1100–1108.

[17] S. Isaacman, R. Becker, R. Cáceres, M. Martonosi, J. Rowland, A. Varshavsky, and W. Willinger, "Human mobility modeling at metropolitan scales," in *Proceedings of the 10th international conference on Mobile systems, applications, and services*. Acm, 2012, pp. 239–252.

[18] D. J. Mir, S. Isaacman, R. Cáceres, M. Martonosi, and R. N. Wright, "Dp-where: Differentially private modeling of human mobility," in *Big Data, 2013 IEEE International Conference on*. IEEE, 2013, pp. 580–588.

[19] E. Graells-Garrido and J. García, "Visual exploration of urban dynamics using mobile data," in *International Conference on Ubiquitous Computing and Ambient Intelligence*. Springer, 2015, pp. 480–491.

[20] E. Graells-Garrido, L. Ferres, D. Caro, and L. Bravo, "The effect of pokémon go on the pulse of the city: a natural experiment," *EPJ Data Science*, vol. 6, no. 1, p. 23, 2017.

[21] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, and A. Varshavsky, "Identifying important places in people's lives from cellular network data," in *International Conference on Pervasive Computing*. Springer, 2011, pp. 133–151.

[22] J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti, "Cellular census: Explorations in urban data collection," *IEEE Pervasive computing*, vol. 6, no. 3, 2007.

[23] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási, "Structure and tie strengths in mobile communication networks," *Proceedings of the national academy of sciences*, vol. 104, no. 18, pp. 7332–7336, 2007.

[24] R. Lambiotte, V. D. Blondel, C. De Kerchove, E. Huens, C. Prieur, Z. Smoreda, and P. Van Dooren, "Geographical dispersal of mobile communication networks," *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 21, pp. 5317–5325, 2008.

[25] R. Ahas, S. Silm, O. Järv, E. Saluveer, and M. Tiru, "Using mobile positioning data to model locations meaningful to users of mobile phones," *Journal of urban technology*, vol. 17, no. 1, pp. 3–27, 2010.

[26] P. Nurmi and S. Bhattacharya, "Identifying meaningful places: The non-parametric way," in *International Conference on Pervasive Computing*. Springer, 2008, pp. 111–127.

[27] R. A. Becker, R. Caceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky, "Route classification using cellular handoff patterns," in *Proceedings of the 13th international conference on Ubiquitous computing*. ACM, 2011, pp. 123–132.

[28] F. Girardin, A. Vaccari, A. Gerber, A. Biderman, and C. Ratti, "Quantifying urban attractiveness from the distribution and density of digital footprints," *International Journal*, vol. 4, pp. 175–200, 2009.

[29] V. Soto and E. Frias-Martinez, "Robust land use characterization of urban landscapes using cell phone data," in *Proceedings of the 1st Workshop on Pervasive Urban Applications, in conjunction with 9th Int. Conf. Pervasive Computing*, 2011.

[30] K. Zhao, S. Tarkoma, S. Liu, and H. Vo, "Urban human mobility data mining: An overview," in *Big Data (Big Data), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1911–1920.

[31] J. F. A. B. F. Chirigati and H. V. K. Zhao, "Exploring what not to clean in urban data: A study using new york city taxi trips," *Data Engineering*, p. 63, 2016.

[32] V. Novák and I. Perfilieva, "The principles of fuzzy logic: Its mathematical and computational aspects," in *Lectures on Soft Computing and Fuzzy Logic*, A. Di Nola and G. Gerla, Eds. Heidelberg: Physica-Verlag HD, 2001, pp. 189–237.

[33] X. LI, G. PAN, Z. WU, G. QI, S. LI, D. ZHANG, W. ZHANG, and Z. WANG, "Prediction of urban human mobility using large-scale taxi traces and its applications."

[34] O. S. U. A. Hurtado, "Informe ejecutivo, eod de viajes - santiago 2012," March 2015.

[35] T. Z. Zarsky, "Incompatible: The gdpr in the age of big data," *Seton Hall L. Rev.*, vol. 47, p. 995, 2016.