Article

A Hierarchical Fuzzy-Based Correction Algorithm for the Neighboring Network Hit Problem [†]

Andrés Leiva-Araos ^{*,†} and Héctor Allende-Cid [‡]

Escuela de Ingeniería Informática, Pontificia Universidad Católica de Valparaíso, Avenida Brasil, Valparaíso 2241, Chile; hector.allende@pucv.cl

* Correspondence: andres.leiva.a01@mail.pucv.cl

[†] This article presents a work that is the continuation of our article published in the proceedings of the 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 9–12 December.

[‡] These authors contributed equally to this work.

Abstract: Most humans today have mobile phones. These devices are permanently collecting and storing behavior data of human society. Nevertheless, data processing has several challenges to be solved, especially if it is obtained from obsolete technologies. Old technologies like GSM and UMTS still account for almost half of all devices globally. The main problem in the data is known as neighboring network hit (NNH). An NNH occurs when a cellular device connects to a site further away than it corresponds to by network design, introducing an error in the spatio-temporal mobility analysis. The problems presented by the data are mitigated by eliminating erroneous data or diluting them statistically based on increasing the amount of data processed and the size of the study area. None of these solutions are effective if what is sought is to study mobility in small areas (e.g., Covid-19 pandemic). Elimination of complete records or traces in the time series generates deviations in subsequent analyses; this has a special impact on reduced spatial coverage studies. The present work is an evolution of the previous approach to NNH correction (NFA) and travel inference (TCA), based on binary logic. NFA and TCA combined deliver good travel counting results compared to government surveys (2.37 vs. 2.27, respectively). However, its main contribution is given by the increase in the precision of calculating the distances traveled (37% better than previous studies). In this document, we introduce FNFA and FTCA. Both algorithms are based on fuzzy logic and deliver even better results. We observed an improvement in the trip count (2.29, which represents 2.79% better than NFA). With FNFA and FTCA combined, we observe an average distance traveled difference of 9.2 km, which is 9.8% better than the previous NFA-TCA. Compared to the naive methods (without fixing the NNHs), the improvement rises from 28.8 to 19.6 km (46.9%). We use duly anonymized data from mobile devices from three major cities in Chile. We compare our results with previous works and Government's Origin and Destination Surveys to evaluate the performance of our solution. This new approach, while improving our previous results, provides the advantages of a model better adapted to the diffuse condition of the problem variables and shows us a way to develop new models that represent open challenges in studies of urban mobility based on cellular data (e.g., travel mode inference).

Keywords: mobile data; neighboring network hit; fuzzy logic; human mobility; data wrangling

Citation: Leiva-Araos, A.; Allende-Cid, H. A Hierarchical Fuzzy-Based Correction Algorithm for the Neighboring Network Hit Problem. *Mathematics* **2021**, *9*, 315. <https://doi.org/10.3390/math9040315>

Academic Editor: Víctor Yepes

Received: 24 December 2020

Accepted: 20 January 2021

Published: 5 February 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

1.1. Definitions

In order to ensure a better understanding, we will use the following terminology throughout the paper:

Definition 1 (Cellular Site). *A set S of cellular serving nodes consisting in one or several Base Transceiver Station (BTS), that aims to optimize the signaling in a specific coverage area [1,2].*

Normally in a site, there is one BTS per technology (GSM, UMTS, LTE, 5G, etc.) and several antennas per technology. $S = \{s_1, s_2, \dots, s_n\}$ is the set of all n sites (with $n = 201$) in our study, see Figure 1.

Definition 2 (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 user plane events (CDR) and control plane events (XDR) data.

Definition 3 (Cellular Data). Mobile networks generate records produced by a device m when it is turned on and interacts with them. There are two types of cellular data [3]; CDR, which are triggered by user events (e.g., calling, texting, browsing, etc.), or XDR, which are updates of the network at data channel level, capturing background actions of apps and the network. We define H_m as the temporal sequence of recorded events $h_{m,t}$ in the network that belong to the same device m within a day.

Definition 4 (Network Hit (NH)). The recording of an event $h_{m,t}$ generated by the device m at time t in the network. Every NH is associated with a cellular site s_i , where $s_i \in S$.

Definition 5 (Neighboring Network Hit (NNH)). We define a neighboring network hit h' as a network hit where the cellular site does not correspond to the geographic place where the cell phone is located in, giving a wrong position for the device m . In Section 2.1, this problem is explained in detail.

Definition 6 (Trace Data). We define a trace as a correlative sequence two of events, $h_{m,t} \rightarrow h_{m,t+1}$, which are associated with two different cellular sites, $s_t \neq s_{t+1}$.

Definition 7 (Trip or travel). All travel made on public roads with a determined purpose 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 traces.

Figure 1. Visualization of sites in the Valparaíso area in a time period of 15 min. Some sites receive neighboring network hits (NNHs) (red), and others lose network hits (NHs) (green). Most of them are in sectors whose geography allows a clean line-of-sight between sites.

1.2. Context, Purpose, and Significance

The ability to adapt in some cities is rapidly declining. This is particularly true in big conurbations, especially in the developing world [4], where the development plans have huge technical, social, economic, and political complexities. Urban mobility represents an important dimension in which cities are in debt in terms of planning and management; the increase in the population and the explosive growth of the automotive park fleet are two variables that affect this matter. More recently, the necessity to control the pandemic crisis has added new dimensions to the problem [5]. The design of urban mobility plans and policies considers several types of input data. The travel surveys, normally managed by the governments, are among them because they provide data about the urban mobility patterns: where to and from people travel, the trip's purpose (e.g., commuting from home to work), the trip's mode (e.g., bus, car, bicycle, subway), traveling time, as well as other sociodemographic variables. However, surveys are complex to prepare and process, expensive, static, and may have sampling biases [6,7]. Surveys represent only snapshots of a dynamic phenomenon over time, and therefore later versions only allow us to determine large patterns and their changes.

According to the GSMA, 67% of the world's population is connected. Daily there are almost 8.8 billion mobile connections in the world, and 5.3 unique mobile subscribers. About 48% correspond to obsolete technologies such as GSM and UMTS [8]. These devices automatically capture and store people's behavior data. However, the captured data have several challenges to face [3], even if they do not come from outdated sources.

The call detail records (CDR) have been used for more than 20 years to study mobility dynamics, and are, without a doubt, the most investigated type of mobile data in this context. In general, CDR are collected by mobile operators for billing purposes, but they have several scientific uses as shown in surveys [3,9], promising application domains [10], understanding the mobility patterns [11], specific applications such as sensing the human mobility [2,12], or understanding the predictability limits [13]. These data have the potential to help relevant government entities thanks to their volume in spatial and temporal terms (i.e., certain areas of cities, or particular days of the year). With the advent of the Big Data phenomenon, these data begin to be stored, opening the possibility of studying seasonal phenomena. There are a number of well-documented problems [3], however, the quality of this data is also affected due to: post processing, network problems, network configuration, overloaded cell towers, or data collection errors. Most of the previous works make no mention of this phenomenon, or simply suggest discarding these bad records [2], rather than fixing them. As we demonstrated in our previous work [14], this practice in the preparation of the data does not affect their general distribution when we analyze large volumes (e.g., people or geographic areas), however, it introduces significant biases in studies whose scale must be reduced. As far as we know, all of the previous work done on cellular data was done by removing data deemed erroneous or by analyzing large areas, with large volumes of data that reduced the impact of outliers and incorrect data. Although these studies are useful for developing policy and making aggregate decisions, they fail to develop micro-scale conclusions. The opportunity then arises to develop models that support decisions at much smaller geographic scales (i.e., census areas, postal codes, or even smaller ones such as commercial neighborhoods) or in panel studies with small groups of people (i.e., new immigrants), or tracking group of peoples geographically (e.g., COVID-19 vectors), among others. Based on our previous work, we propose a novel way to fix errors and compute travel and distance distributions in the time series represented by the CDRs. Our goal was to improve the state-of-the-art in cellular data-based analytics for small areas or groups of people. Through a series of experiments, we compared our method with previous work and demonstrated that simply discarding (bad) records can significantly affect travel attributes and patterns, which in turn introduces biases in subsequent mobility analyses.

In summary, the main contributions of our paper are:

- A novel and improved fuzzy logic methods for fixing erroneous CDRs and computing trips and distance.
- A detailed evaluation of our methods using real-world based on data from Valparaíso Chile.
- A detailed evaluation of our methods with a real use case.

The rest of the paper is organized as follows. Section 2 provides an overview of understanding the network and its limitations as a mobility data source. Section 3 reviews the state-of-the-art in the research using event-driven cellular data. Section 4 presents our fuzzy reasoning solution approach in solving of the neighboring network hit problem and the trips computing. Section 5 describes the data sets used and the experiments setup. Section 6 describes our results compared with Origin-Destiny Surveys and our previous works. Finally, in Section 7, we make our conclusions and present future next steps and challenges.

2. Understanding the Network and Its Limitations as a Mobility Data Source

2.1. The Problem

Cellular networks, in both physical and logical terms, are extremely complex systems. They are designed around different layers of abstraction that make possible the simultaneous flow of a multiplicity of messages between their different nodes. Taking as reference the ISO/OSI layer model [15], the first dimension of abstraction is the physical layer. In this layer, physical variables are defined as voltage levels, the timing of voltage changes, physical data rates, maximum transmission distances, modulation scheme, channel access method, and physical connectors. When two devices are connected at the ends of the communication channel (e.g., a cellular device with a service antenna), a series of factors appear which influence that connection: the types of antennas, the emission power, the frequency used, the levels of interference that exist in the environment, and the conditions of line of sight. It is precisely in this layer that is required a complex control of these variables to ensure quality and profitable communication. In order to minimize network costs, operators have the necessity to design the network in such a way as to maximize its coverage while minimizing interference factors [16]. This design challenge represents one of the biggest dilemmas in optimizing costs and quality of service. This trade off is especially complex in urban areas where factors such as the height and density of the buildings, the shape of the city, and the height differences in the land, have direct impact in the design of the physical layer. This design condition explains that urban networks have limited coverage at distances between 150 and 500 m [3]. This implies that a device m should be connected to the nearest antenna most of the time. However, environmental conditions are dynamic and changing, which means that devices can connect to distant sites whose distance far exceeds the network's average of nearby sites.

There are important issues in the quality of data. The quality of the CDRs depends on the type of technology (GSM, UMTS, LTE, etc.) and the design of the network. When we analyze and visualize the data it is common to find cases of devices connected to distant antennas (see Figure 1). This is because the main concern of the operators is to maximize the quality of service that is delivered to users (e.g., handover failures, call drop rate, etc.). As we mentioned, the shape of the city and the differences in heights between different areas with clear sight-lines, make it easier for mobile phones to connect to distant antennas. Bay-shaped cities surrounded by hills (e.g., Valparaíso-Chile), sometimes allow a user located at one end of the city to obtain service from an antenna located at the other end, more than 6 km away. This phenomenon occurs when various external factors of the network take on a greater relevance than those considered in its design. Among them, the “mirror” effect of the sea facilitates the propagation of electromagnetic signals, facilitating the connection between the device and the antenna.

Neighboring network hit (NNH) problem. In Figure 2a, we can see the geographical visualization example of a typical trip between points A and B. Sometimes, depending on favorable conditions, the device m (probably the mobile phone of a local bus passenger or a car driver) cannot get connection to the next site in the pathway, s_4 , (e.g., due to network

congestion), and connects with site s_3 distant 5 km from its actual position; a few meters further along its route, it connects with the antenna in site s_4 , to which it should have normally connected given its trajectory. These kinds of issues normally do not affect the travel counting, but they have an important impact in the traveled distance computation, mainly if the effect is more subtle than what is shown in the figure. We call this problem the neighboring network hit or NNH for short.

Figure 2. Two of the most frequent problems that the data wrangling process presents. Both have negative implications in the computation of aggregate distances in urban mobility. (a) Elevation profile between two cell sites in Valparaíso that creates an NNH connection. (b) Example of notable differences between Euclidean, Manhattan, and real distances between two points in the city. This condition is common in hilly cities.

Distance computation problem. Even though we do not know the exactly route that device m uses to move across the city, previous studies make the assumption of displacements based on rectilinear distance. This form of calculation, also known as $L1$ -Norm or Manhattan distance, replaces the Euclidean distance with a new metric in which the distance between two points is the absolute difference of the components of the vectors. In Figure 2b, we can see the actual elevation profile between two nearby cell sites (s_1 and s_2). The green line from points A to B represents the linear (Euclidean) distance between them (1.4 km), which corresponds to 1.7 km considering the $L1$ -Norm (Manhattan distance), but the optimal walking/driving path is 2.9 km, represented by the blue route. Here we can find two problems. First, a device m moving from A to B has an underestimated minimum traveled distance. We call this phenomenon in linguistic terms as *low path linearity*, which is calculated based on the ratio between linear distance and minimum real distance. Second, a device located at point C can perfectly connect to both sites if they have their azimuths facing each other and the antennas down-tilts with convergent angles. If we do not properly compute the minimum travel distance from s_1 to s_2 , by dividing it by the maximum time used to travel them, we will obtain maximum speeds in the city that do not correspond to those actually observed.

Few works mentioned these issues as critical [3,17], but none of them refer to their magnitude or their implications in the studies. Instead, they use well-curated data [11,13,18,19], synthetic data [20,21], or eliminate the outliers [2,22,23]. In all cases, the additional biases introduced into the data are not analyzed, nor their implication in later inferred conclusions. Many of these adverse effects of these procedures are diluted with the use of large volumes of data. However, it is demonstrated its effect when studies have smaller spatial and temporal scales [14]. In Table 1, we include some studies that use cellular and spatio-temporal data, and the management approach of the NNH phenomenon. As you can see, many of them do not consider the problem or at least do not mention it as part of their workflow.

Table 1. Sample of some studies that use cellular data, their case studies, and the NNH’s management approach.

Study Reference	Case Study	NNH Treatment	Data
[2,22]	Human mobility exploration	Record deleting	Pre-processed
[3]	Survey of mobility	Mentioned as critical	NA
[11], [18]	Human mobility patters	Not mentioned	Controlled sample
[12]	Mobility during social events	Spatial isolation	Pre-processed
[20,21]	Synthetic data generation	Not mentioned	Controlled sample
[23]	Impact of location-based game in the pulse of a city	Record deleting	Pre-processed
[24]	Semantic places in people’s lives	Not mentioned	Controlled sample
[25]	City as a holistic, dynamic system	Not mentioned	Pre-processed
[26]	Local and the global structure of a society-wide communication network	Not mentioned	Pre-processed
[27]	Data science for social good	Record deleting	Pre-processed
[28]	Mobility and social inclusion	Record deleting	Pre-processed
[29]	Taxi trips visual exploration	Mentioned, record deleting	Raw data
[30]	Taxi trips data cleaning	Domain Knowledge	Raw data

3. Related Work

Most humans today carry mobile telephones. According to the Global System for Mobile Communications (GSMA), by the end of 2019, there were 5.2 billion unique mobile subscribers, accounting for 67% of the global population [31]. Since all these cell phones make hits on the network with an average frequency of 24.91 (Average obtained from our data, along 2 years.) daily records per device and, in addition, these hits allow to build traces of mobility, is that the use of this information has become an essential source for studies and works that aim to understand, model, and predict urban mobility.

Cellular data have been used for several years in a myriad of applications, particularly in understanding human mobility through daily patterns [11], semantic places in people’s lives [24], holistic view of the city [25], crowd mobility during special events [12], society-wide communication network [26], statistical properties of a communication network [32], social good [27], the impact of location-based game in the pulse of a city [23], and social inclusion [28], gender gaps and inequalities [33], public health [5]. For a more complete review of studies in the area, we refer readers to [3].

Most existing works assume simple models of human mobility to collect or generate data. This is the case of the work in [20] where the authors introduced the Work and Home Extracted REgions (WHERE), and subsequently, the Differential-Privacy WHERE (DP-WHERE) model [21], to produce synthetic CDRs.

In terms of identifying meaningful places, there are different approaches, some of them using parametric models based on passive mobile data in [34], or non-parametric Bayesian approach like in [35]. To deal with the inaccuracy of CRD traces (accuracy in urban areas is approximately 150 to 500 m), several solutions have been proposed, including a study of recurring locations over time [24] and handover during [36] calls. Neither of these approaches take into account the network design variables that affect the quality of the CRDs.

Several works model mobility data using spatial and temporal profiles to: improve the inference of frequent places [2], quantify urban attractiveness [37], determine land use [38], or estimating types of activities [39]. Considering that call activity is different depending on the day and week and the city area, it is possible to derive a classification of the activity profile and define regions as “residential”, “commercial”, or “business”. In [2], the authors propose virtual locations for the antennas, seeking to increase the precision in the geo-referencing of the devices. However, these improvements do not solve the problem of computing trips in small areas. CDRs have relatively high uncertainty in the user’s location and time. This is due to the low rate of *NH* per unit time and the spatial resolution of the network design. Recent works, due to the pandemic outbreak, mobility studies based on cellular data have been reactivated to understand mobility and the result of the health measures adopted [40]. In all these studies, micro-mobility remains a pending challenge to

be solved. Our work focuses on reducing uncertainty by setting outliers to improve travel counting and distance computation in small study units, in terms of areas and numbers of individuals, within the city.

3.1. Research Effort Using Cellular Data

Taking into account the survey in [3], and the comprehensive review of the typology of spatial studies based on mobile phone data in [41], we summarize the research effort in five large areas:

Estimating population distribution. In terms of estimating where the population lives, there are several works: determining the geographical location of home and work-places using parametric models [24,34], and non-parametric Bayesian approaches [35]. In terms of how the estimation of the density of people changes over time, in [17] the authors analyze it by exploring how to use GSM data to recognize high-level properties of user mobility and daily step count. The work in [42] shows how to assist fire and rescue services base on calculating and visualizing mobile phone density. In [43], the authors present the investigation of the calculation and representation of temporally and spatially highly dynamic point data sets based on kernel density estimation (KDE). In terms of the aggregate use of cellular data, in [44], the authors use it to identify the socioeconomic levels.

Estimating types of activities in the city. During the week, call activity by type or region (residential, commercial, or business) is different. It is possible to classify the regions based on the activity profile contained in the CDRs. For example, the work in [37] provides a case study in which different areas of interest in New York are tracked using aggregated cellular data and geo-referenced Flickr photos. Other works attempt to obtain clusters from the data by measuring the activity of cell sites. For example, in [25] the authors obtain clusters of the dynamics of Rome, and in [38] the work aims to automatically identify different land uses (e.g., industrial, commercial, nightlife, recreation, residential, etc.). In [2], the authors propose virtual locations for the antennas, seeking to increase the precision in the geo-referencing of the devices compared with a standard method like Voronoi tessellations.

Estimating mobility patterns. Given its geo-referenced condition, cellular data can be used to estimate commuter's mobility in predefined regions. Several groups of researchers carry out extensive work in this field. Among them, the so-called "Barabasi Lab" (<http://www.barabasilab.com/iritu>) with its open project on "Individual mobility patterns", and the "MIT Senseable City Lab" have a recognized track record. The work in [11] shows us how to track both groups and individuals based on the widespread coverage of mobile sites in urban areas. In [18], the authors demonstrate that human trajectories follow several highly reproducible scaling laws, deprecating the continuous-time random-walk models for human mobility. Subsequent mobility patterns have been used for a wide range of studies, from people migration [45] until road usage [46]. One of the first works of "MIT Senseable City Lab" aims to investigate and how digital technologies, in particular cellphones, are changing the way people live. In [47], they used CDRs to monitor the vehicular traffic status and the movements of pedestrians in Rome, Italy. The work in [36] describes human mobility in several US cities to evaluate the effect of human travel on the environment. In [13,48], the authors work on exploring the limits of predictability in mobility patterns using statistical methods.

Analyzing local events. In more recent years, studies with cellular data have led to issues related to mobility at local events. In particular, several studies attempted to infer human mobility patterns during different kind of emergencies compared with non-emergency events [49], earthquakes [50], and special social events [12,51,52].

Analyzing social networks geography. The impact of geography on interactions through social networks has been approached from statistical perspectives [32], to determine the relative frequency and average duration of communications [53], and to study the social radius of influence [54].

3.2. Open Challenges and Opportunities Related to the Use of CDRs

There are open opportunities and challenges in the use of this type of data.

Limitations of event-driven data. CDRs are event-driven data, which means that they are generated only when the users perform some action, e.g., send an email, search for something on the Internet, make a call, etc. [3]. The geographical position is obtained from the site to which the device connects, and, therefore, the location is updated as these events are recorded. The development of certain urban mobility patterns is affected by the low frequency of events (NHs). To solve this, two approaches are used: sampling high-activity users or sampling internet usage data. In both cases, the sampling process presents complexities and potential representativeness problems. The integrated use of data from different companies to get better samples poses both technical and information privacy challenges.

Limitations in spatial accuracy. Mobile phone network data do not provide accurate localization. The spatial accuracy in urban areas is about 150 to 500 m, so there are limitations in the type of solutions provided with this precision. To this point, some solutions proposed so far are increase the precision based on virtual position of the antennas [2], look at history for recurring locations [24], and look at handover during calls [36].

Managing uncertainties. Due to the limitations of the event-driven data and the spatial accuracy, the uncertainties in the user's status in time and space can be relatively large. As mentioned, two factors that determine this are the low frequency of user localization refresh and the low spatial resolution of the network. In [55], the authors propose a solution to estimate uncertainties in users' position with a trigonometric approach.

Finding comparative data sets. Comparing the results of investigations conducted with cellular data is faced with the scarcity and complexity of ground truth data. Government census and survey data have different spatial and temporal resolutions, and these studies are usually carried out in geographic areas with coverage from several cellular companies (Telcos). Some proposed solutions are self-reported data (e.g., Flickr, Instagram) and social media data (e.g., Facebook, Twitter).

Dealing with privacy and anonymity. The conflicts between data mining and security just begin and promise a tough battle both in the field and in court. There is a growing concern about security and privacy gaps in the use of information collected on a daily basis. The key point here is that we are dealing with data that do not belong to us (the researchers) but the Telco's customers [14].

Real-time data acquisition and processing. Nowadays, real-time analytics gives an extra value in the modeling and predictability of complex processes. Many solutions based on cellular data acting as mobility sensors acquire more value if they operate in real-time (e.g., social event monitoring, traffic optimization, demand prediction, etc.). The development of models based on streaming data generates new challenges in its ingestion and operation. Today it is increasingly common to see data produced every few minutes. However, the high massiveness of such data forces the design of extraction, transformation, loading, and analysis processes in distributed configurations and clusters, and streaming processing.

Social good. There are many pilot projects in which massive cellular data are being used to achieve sustainable development goals. Examples of this are: poverty analysis [56], electrification planning [57], crime prediction [58], CO2 emissions, supporting managing humanitarian disasters [50], among others. An open issue is the lack of standardized cross-countries cross-operators' insights; otherwise, the results cannot be compared.

3.3. Data Quality

As a result of the data privacy concern, the vast majority of studies use data that are properly pre-processed and curated [18], or data that come from the location's services intelligence platforms [59]. Cellular data quality is a real critical issue, but little has been discussed in the studies [29]. Technologies designed for geo-positioning, such as GPS, can also deliver inaccurate data due to the design of the reception algorithms and the

conditions of the physical environment [60]. Cellular networks are subject to the same problem. However, it is not clear its magnitude and/or implications in the analyzes that use it. In many cases, some parameters, at radio access network (RAN) and core network (CN) levels, are configured manually. In [30], the authors argue that data preparation and cleaning is an iterative and essential process for subsequent analysis. In their work, they use GPS data from New York taxis. Preliminary analyses and visualizations show that 4.6% of the data represented ghost rides with taxis geolocated on rivers, oceans, and even outside the United States. Sampling does not solve the problem product of representativeness and coverage factors [3]. The foregoing is aggravated if there are erroneous data that are within ranges considered normal, which go unnoticed without prior in-depth analysis. The problem of NNHs was first proposed and studied in [14]. For further details, in Table 1, we include some studies that use cellular and spatio-temporal data, and the management approach of the NNH phenomenon.

In this work, we apply a hierarchical fuzzy logic model to demonstrate that NNHs have an important impact on the calculation of the mean distances traveled by users. The correction of the NNHs eliminates the biases they introduce, allowing the scale of subsequent studies to be reduced.

4. Fuzzy Reasoning Solution Approach

Fuzzy logic: When the first paper of fuzzy logic [61] was written, there were no technical journals that dared to accept it because at that time it was inconceivable to allow vagueness in the engineering field. The turning point came in 1974 when Ebrahim Mandami applied fuzzy logic to controls for the first time [62]. The fuzzy sets theory was introduced to provide a scheme for handling a variety of problems with an intrinsic characteristic of ambiguity more than a statistical variation [63]. Fuzziness differs from imprecision. In tolerance analysis, imprecision refers to a lack of knowledge about the value of a parameter and is thus expressed as a crisp tolerance interval. This interval is the set of possible values of the parameters. Fuzziness occurs when the interval has no sharp boundaries, i.e., is a fuzzy set $\tilde{A}$. Then, $\mu_{\tilde{A}}(x)$ is interpreted as the degree of possibility that x is the value of parameter fuzzily restricted by $\tilde{A}$ [64]. In this work, we use fuzzy reasoning with Mandami's direct method [62] in order to fix the NNH outliers found in the data.

Fuzzy reasoning: In a typical application, the system is defined using a flat set of rules. That is, all the rules have the same variables in the antecedent part (qualified using different fuzzy terms) and conclude about the same variable. Rules are defined so that antecedents define a fuzzy partition of the application domain. That is, for each possible point in the n -dimensional input space (we assume that there are n variables) there is at least one rule (usually more) that can be applied. Then, when the system is applied to a particular situation (a given input), all rules are fired in parallel (applied all at once to this given input), and for each rule its conclusion is computed. The computation considers the degree to which the antecedent is satisfied; if it is not satisfied at all, the conclusion is an empty set. Subsequently, the final output is computed through the combination of the conclusions of all the rules. This combination usually consists of the union of the conclusions of all rules and a final step of defuzzification. Defuzzification means the process of transforming the union of the conclusions (that is a fuzzy set) into a crisp value (e.g., a numerical value). This process can be seen as either an element selected from a set (in fact, from a fuzzy set) or a fusion process in which the information to be fused the fuzzy set and the outcome is the numerical value [65].

In our work, we define a nine-dimensional domain of variables with a range of 2 to 5 rules each. Given this, we get a complex domain configuration with more than 20,000 rules in the complete set. The model takes advantage of independence in some variables to significantly reduce the required rule-base size when describing the system, without compromising robustness. In order to deal with the curse of dimensionality, we apply the following techniques:

Hierarchical rules architecture: We apply a hierarchical [66] with a prioritized structure [67] approach that decomposes the large-scale system into a finite number of reduced-order subsystems or modules, see Figure 3, thereby eliminating the need for a large-sized inference engine [68]. Each module has the ability to compute a defuzzification process delivering a final output or connecting to the next level.

Parallel-distributed computing: In order to process and fix large amounts of data from NNHs (typically one day of data can have 180 million records), we create a parallel and distributed single program, multiple data architecture.

Figure 3. Hierarchical rules architecture: Rules are grouped into modules according to their roles in the system. Each module has the ability to compute a defuzzification process delivering a final output or connecting to the next level. In the latter case, the input becomes part of the antecedent of a new set of rules, in this way up to the last level if necessary.

In our work, we developed two algorithms base on fuzzy reasoning: *Fuzzy Logic NNH Fix Algorithm (FNFA)* and *Fuzzy Logic Trips Counting Algorithm (FTCA)*. Both are used sequentially to fix NNH data and compute travel counting and distances.

4.1. Fuzzy NNH Fix Algorithm (FNFA)

FNFA aims to fix the NNH records in the dataset, improving the trip count and the computed distance accuracy compared with previous methods. In the latter, each time an anomaly is detected (an NNH), the entire time series is removed. This means eliminating devices that could provide relevant information for further analysis. If the anomaly is not detected, the jump represented by the NNH is considered valid, and the time series is processed, introducing a bias in the computed distances.

In this model, we define a nine-dimensional domain of variables with more than 20,000 rules in the complete domain. In order to create our model, we use the following conceptualization of the traces over antennas sites in a particular time t , see Figure 4.

Figure 4. (a) NH sequence in time series, where s_{t-1} , s_t , and s_{t+1} , represent the previous, current, and future position (antennas) of a device m , centered in time t . (b) The model needs to consider different spatial behaviors, where s_{t+1} could be closer to s_{t-1} than s_t .

The figure shows a rolling window within the NH time series, centered on t (s_t). The sites s , represent the previous, current, and next antennas to which the device m connects. For convenience, we call these sites sA , sB , and sC . The model seeks to detect problems (NNH) in sB and correct them. Then, its possible outputs against a detected NNH can be $s_t = s_{t-1}$ (sB is actually sA) or $s_t = s_{t+1}$ (sB is actually sC). The variables, universe of discourse, and linguistic terms are shown in Table 2.

Table 2. Fuzzy Logic NNH Fix Algorithm (FNFA) model's linguistic variables, universe of discourse, and linguistic terms.

Variable Name	Universe of Discourse	Linguistic Terms
$v = \text{min velocity from A to B (MVAB)}$	$V = \{v \mid 0 \leq v \leq 1000\}, [\text{km/h}]$	$T(v) = \{\text{very slow, normal, high urban, high inter urban, very high}\}$
$w = \text{min velocity from A to C (MVAC)}$	$W = \{w \mid 0 \leq w \leq 400\}, [\text{km/h}]$	$T(w) = T(v)$
$x = \text{distance from A to B (DAB)}$	$X = \{x \mid 0 \leq x \leq 4000\}, [\text{km}]$	$T(x) = \{\text{very short, short, city size, long, very long}\}$
$y = \text{distance from A to C (DAC)}$	$Y = \{y \mid 0 \leq y \leq 4000\}, [\text{km}]$	$T(t) = \{\text{inside macro area, near macro area, far from macro area}\}$
$p = \text{relative position between A and C (RPAC)}$	$P = \{p \mid 0 \leq p \leq 1\}$	$T(p) = \{\text{closer to A, closer to C}\}$
$s = \text{slope between A and B (SAB)}$	$S = \{s \mid 0 \leq s \leq 1\}$	$T(s) = \{\text{low, high}\}$
$l = \text{linearity between A and B (LAB)}$	$L = \{l \mid 0 \leq l \leq 1\}$	$T(l) = \{\text{low, high}\}$
$gw = \text{GoogleMap API walking distance (GWD)}$	$GW = \{gw \mid 0 \leq gw \leq 0.6\}, [\text{km}]$	$T(gw) = \{\text{short, long}\}$
$gd = \text{GoogleMap API driving distance (GDD)}$	$GD = \{gd \mid 0 \leq gd \leq 960\}, [\text{km}]$	$T(gd) = \{\text{short, long}\}$

The model implements the hierarchy shown previously (Figure 3). The architecture is composed of four levels. The first level (*Level 0*), uses an architecture based on the fuzzy reasoning method based on Sugeno and Takagi linear functions [65,69]. The structure of the rules, and its derived outputs are given by:

$$\begin{aligned} \text{Rule 0.1 : } & \text{ IF } gw \text{ is } A_1 \text{ and } gd \text{ is } B_1 \text{ THEN } z_1^* \text{ is } f_1(gw, gd) = (gw_0 + gd_0)/2 \\ \text{Rule 0.2 : } & \text{ IF } gw \text{ is } A_1 \text{ and } gd \text{ is } B_2 \text{ THEN } z_2^* \text{ is } f_2(gw, gd) = gd_0 \end{aligned} \quad (1)$$

where A_i , B_i , are fuzzy sets; gw_0 , and gd_0 are the facts obtained from GoogleMaps walking and driving optimal distance respectively; and z_i^* are the individual rule outputs. The crisp control action is expressed as:

$$z_0 = \frac{\alpha_1 \times (gw_0 + gd_0) + 2\alpha_2 \times gd_0}{2(\alpha_1 + \alpha_2)} \quad (2)$$

where α_i denotes the firing level of the *level 0, i - th* rule, $i = \{1, 2\}$, computed by:

$$\begin{aligned} \alpha_1 &= A_1(gw_0) \wedge B_1(gd_0) \\ \alpha_2 &= A_1(gw_0) \wedge B_2(gd_0) \end{aligned} \quad (3)$$

We can see *Level 0* fuzzification and defuzzification process in Figure 5a. In this level, there is a continuous crisp output, which becomes the input of *Level 1*. The next three levels (1, 2, and 3), have a similar rules structure. Following the structure of *Level 2*:

$$\begin{aligned} \text{Rule 1.1 : } & \text{ IF } x \text{ is } A_1 \text{ and } y \text{ is } B_1 \text{ THEN } z \text{ is } C_1 \\ \text{Rule 1.2 : } & \text{ IF } x \text{ is } A_2 \text{ and } y \text{ is } B_2 \text{ THEN } z \text{ is } C_2 \\ & \dots \\ \text{Rule 1.i : } & \text{ IF } x \text{ is } A_i \text{ and } y \text{ is } B_i \text{ THEN } z \text{ is } FF_2 \\ & \dots \\ \text{Rule 1.n : } & \text{ IF } x \text{ is } A_n \text{ and } y \text{ is } B_n \text{ THEN } z \text{ is } C_n \end{aligned} \quad (4)$$

where A_i , B_i , and C_i are fuzzy sets; and FF_2 is a fuzzy boundary rule. To avoid binary (not fuzzy) thresholds, we created what we call fuzzy boundary rules. With them, we make sure that certain variables (e.g., the minimum speed between sA and sB) have fuzzy limits at both ends (upper and lower).

In Figure 5b–d, we can see an example of this level. Each level of the hierarchy has a defuzzification step. In this step, the consequent of some rules have fuzzy outputs or fuzzy filters. The defuzzification mechanism selected is *Last from Maxima*. Consequently, the defuzzification process gives us both final results ($s_2 = sA$ or $s_2 = sB$), or access to the next level in the hierarchy (FF_2). In the latter case, the flow continues to the next level in the hierarchy (*Level 3*). The next two levels (3 and 4) operate in a similar way using another set of variables and rules, and defusing towards $s_2 = sA$, or $s_2 = sC$.

Fuzzy sets' cores and support points for distance and altitude variables between sites s_i are obtained from original geo-referenced sources of the mobile operators. We use urban and rural density of antennas. Figure 6 shows us the mean and standard deviations of distances for different sets of antennas (top-k).

(a) Level 0: input output (Takagi-Sugeno).

(b) Level 1: Crisp output s_t is A.

Figure 5. Cont.

(c) Level 1: Crisp output s_i is B.

(d) Level 1: Fuzzy output, firing Level 2.

Figure 5. Levels 0 and 1 examples of hierarchy filter defuzzification.

Figure 6. Urban and rural site density.

As an example, the mean distance in sets of 6 sites is 0.86 km with a standard deviation of 0.36 km. The velocity variables are defined based on parameters obtained from experimental sources. Average speeds in the city are obtained from data collected from GoogleMaps API over a year. The optimal minimum distances between sites are obtained from the same source, computed considering walks and trips in vehicles. Some of the fuzzy sets can be seen in Figure 7.

Figure 7. Cont.

Figure 7. Cont.

Relative position between A to C

Slope ratio between A to B

Figure 7. Cont.

Figure 7. Membership functions for antecedent and consequent parts in FNFA. Some of the plots are expressed in logarithmic scale due the extension of Universe of Discourse.

To spark the process, we use the contextual distances between two consecutive sites in the time series. Depending on the distance between sites, it is reasonable to think that people can travel these distances walking or in some type of vehicle. This contextual distance variable uses real values for walking or driving, obtained from the GoogleMaps API. In most of the cases, both distances are different. Later, we add minimum speed variables of a device m , $Vmin_m(s_{t-1}, s_t)$, where $t-1$ and t represent the timestamps of two consecutive events in the time series, moving from two points, s_{t-1} to s_t (or sA and sB respectively). The FNFA carries out its work based on a previous model. Each original data set, containing one day of data, is divided into a series of chunks to which the model is applied using a rolling window over the time series, H_m . Every time the algorithm concludes that there is a wrong record, it modifies the cellular site (geographical position) according to its logic. The minimum travel speed of a device m , $Vmin_m$, will be given by one of the following cases, as shown in Figure 8.

Figure 8. The minimum velocity solution approximation.

Depending on the coverage of the cells of a tower, the sequence of two consecutive events ($h_{m,t-1}$ and $h_{m,t}$) in the time series may have different potential distances in reality.

Definition 8 (Minimum velocity). We define minimum velocity as the velocity of a device m moving from two points, s_{t-1} to s_t (or sA and sB , respectively). Let dl and ds be the largest and smallest Euclidean distance between two coverage points of sites s_{t-1} and s_t , respectively. Similarly, let de be the Euclidean distance between both sites. The distances dl and ds (where $dl > de > ds$) potentially represent the extreme cases for a combination of two consecutive NHs at two different sites s_{t-1} and s_t . The time between two successive NHs has, at any of these potential distances, a single known value $t_{m,s_{t-1} \rightarrow t}$. Since we do not know at what exact moment the movement began and ended, $t_{m,s_{t-1} \rightarrow t}$ represents the longest possible travel time for all cases. That means the object m was in motion for the entire time $t_{m,s_{t-1} \rightarrow t}$. With this in mind, we can only conclude that the speed we get by dividing the distances dl or ds by the time $t_{m,s_{t-1} \rightarrow t}$, will always be the minimum speed possible, $Vmin_m$, for the section distance between s_{t-1} and s_t cellular sites. The minimum speed $Vmin$ is the singleton with which we enter our fuzzy model.

Minimum speeds are contextual to geography. In the city center, the density of cell sites is higher and travel speeds are conditioned by congestion and traffic control mechanisms (e.g., traffic lights, traffic regulations, etc.). Displacements from more distant sites are interpreted as intercity trips, which depending on the distance, allow comparable speeds equivalent to those of commercial flights.

The FNFA implements two groups of parameters as input.

Physical parameters: This group of parameters considers actual physical measurements of the sites obtained from the Operator (Telco) and GoogleMaps API. We use the sites latitude, longitude to calculate optimal walking and driving distances between them. Walking and driving distances are used to discriminate proximity. In order to determine

the linearity between sites (*LAB*), we use the elevation above sea level and the height of the support (e.g., communication tower, building, etc.)

Support parameters: This group includes all the parameters that define the fuzzy sets cores and supports. These parameters basically come from the same sources as above, but in this case, they are determined from an empirical process developed throughout the study. Among these parameters, we find the average speed in urban and interurban areas, the average distance between sites in the city and the elevation ratio between nearby sites.

We got all driving and walking distances from the GoogleMaps API. We built an $n \times n$ 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 [2,70]. 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 consecutive events located in different sites, s_{t-1} and s_t does give an important clue about potential NNHs. This speed is calculated using the maximum displacement time between two sites in the time series, and the minimum possible distance given by the GoogleMaps, d_{dg} . We get this distance as follows:

$$d_{dg} = \frac{\alpha_1 \times (gw_0 + gd_0) + 2\alpha_2 \times gd_0}{2(\alpha_1 + \alpha_2)} \quad (5)$$

where $d_{dg}(s_{t-1}, s_t)$ is the minimum distance of device m moving from site s_{t-1} to site s_t (with $s_{t-1} \neq s_t$); gw and gd are the GoogleMaps optimal distances for walking and driving between both sites, respectively; t_{t-1} and t_t are the observed times between two events in the series, and α_i denotes the firing level of the *level 0*, i -th rule, $i = \{1, 2\}$. As we mention, given the essentially imprecise geo-location nature of the mobile data, we do not know exactly where the device m is located when recording the event h at time t .

As seen in Figure 8, the trajectory of m may be before, after, or between two particular sites. There are two extreme cases: first, the distance d_l with initial and ending points located before and after the two sites (case of maximum distance, red points), and d_s with points in between the sites (case minimum distance, green points). In this analysis there is the fact that:

$$d_{dg} > d_{lr} \quad \forall \quad (s_i, s_j) \in S \quad (6)$$

where d_{lr} is the linear distance from cellular sites s_t and s_{t+1} , obtained using the *L1* method. Using real distances (d_{dg}) instead of estimated (d_{lr}), we are increasing the algorithm precision. As we cannot know the maximum speed since we do not know exactly the distance or the travel time, we assume d_{dg} as an approximation to the actual distance traveled. This assumption is handled differently with the rules of our model depending on whether the time series has consecutive events in three different places ($s_{t-1} \neq s_t \neq s_{t+1}$), or in those cases in which the series has a round trip to the same place ($s_{t-1} = s_{t+1} \neq s_t$); in both cases t is the time subscript. The accuracy of the model increases as the site distances increases as well, that because the *linearity* (*LAB*) tends to one ($d_{dg} = d_{lr}$). Then, the minimum velocity approximation V_{min_m} , will be given by the smallest distance to travel (d_{dg}), divided by the longest possible time ($t_i - t_{i-1}$).

$$V_{min_m}(s_{t-1}, s_t) = \frac{d_{dg}}{t_t - t_{t-1}} \quad (7)$$

Consequently, we can consider that:

$$V_{min_m}(s_{t-1}, s_t) \leq \min_real_vel_m(s_{t-1}, s_t) \quad (8)$$

which is true $\forall$ moving m devices.

We do not know the real speed of the devices, but we can make the assumption that their displacements (when they do), cannot be at speeds higher than those observed in the different urban and interurban contexts. As a reference, we observed average speeds of 20 km/h (downtown), and 42 km/h (inter-areas) in Valparaíso.

4.2. Fuzzy Trips Counting Algorithm (FTCA)

FTCA is designed to count trips and compute distances per device m from traces. The process starts computing traces from the NH s events. We use Mandami's fuzzy reasoning to determine whether or not the events contained in the time series represent a change of location (trace). Not always a single trace represents a single travel. It is common that in a single trip there are several NH with the same site. To control this, we define a set of rules to represent the time spent by a device m that is actually in motion ($s_{t-1} \neq s_t$), in the same place $s_t s_{t+1}$, or transiting to a new site ($s_t \neq s_{t+1}$). The variables, universe of discourse, and linguistic terms are shown in Table 3.

Table 3. Fuzzy Trips Counting Algorithm (FTCA) model's linguistic variables, universe of discourse, and linguistic terms.

Variable Name	Universe of Discourse	Linguistic Terms
c = congestion time (CT)	$C = \{c \mid 0 \leq c \leq 24\}, [\text{h}]$	$T(c) = \{\text{low}, \text{medium}, \text{high}\}$
tos = time on site (ToS)	$ToS = \{tos \mid 0 \leq tos \leq 40\}, [\text{min}]$	$T(tos) = \{\text{low}, \text{medium}, \text{high}\}$
dos = density of sites (DoS)	$DoS = \{dos \mid 0 \leq dos \leq 2\}$	$T(dos) = \{\text{low}, \text{high}\}$

In Figure 9, we see a device taken at random with 12 traces (slanted segments), but only 8 travels (horizontal segments, denoted as T). The horizontal segments represent the time spent in a single site. 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 .

Figure 9. Traces and travels for a random device.

In this model, the rules have the same structure as those mentioned in (4). The fuzzy sets are shown in Figure 10.

Figure 10. Membership functions for antecedent and consequent parts in FTCA. Some of the plots are expressed in logarithmic scale due the extension of Universe of Discourse.

The variable tos represents the time spent by the device m at a specific site. Depending on its length this time may represent a stop at the site rather than a passage through it. The site density dos represents the level of concentration of sites in the rolling window of analysis and is given by:

$$dos = \frac{1}{(1 + d_{AB} + d_{AC})} \quad (9)$$

where d_{AB} and d_{AC} are the distances between sites sA , sB , and sC , respectively. Given the volume of data, this algorithm is also performed based on a parallel computing process. The FTCA output consists of calculating the daily trips of the m devices. The results are grouped by area and compared with similar ones from our previous work, and with the Origin-Destination Survey (see Section 6.1) of the Chilean Transportation Authority. FNFA computes the distances between sites in the trace using optimal walking and/or driving distances obtained from the GoogleMaps API. With this information, FTCA computes the distances for each trip.

5. Experiment Setup

5.1. Input Data Description

In this study, we use a subset of the same raw and duly anonymized data set as in our previous work (see Table 4). The idea was to determine the impact of the FNFA-FTCA fuzzy approach on the results. The CDRs used correspond to a time series characterized by having sparse data. Multiple data sets are analyzed, each with data from a full day. The mean of the events, $h_{m,t}$, per device, m , varies between 28 and 46 per day, depending on the data set, this is one NH every 31–51 min. Even though we do not eliminate any outliers, we delete devices with only one NH (10–60% depending on the dataset). Having just one NH , there are no traces at all (see last column of the table). There are large deviations in NH in all the data sets processed. Some devices with high mobility during the day (e.g., truck, taxi, or bus drivers), or devices for Internet of Things–Machine to Machine (IoT-m2m), can have high daily rates of NH s. After removing the devices with one NH , we obtain means close to 35 NH s, with lower standard deviations 27–28. In Figure 11, the boxplot shows the statistics behavior of the data after filtering.

Table 4. Statistics of filtered data (devices with $NH > 1$), for Valparaíso (Val). A sample of four different days.

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 ($\times 10^2$)
Val: 17/01/01	10.3	41.2	44	62.1	250	19
Val: 18/01/01	11.9	37.8	35	83.7	314	146
Val: 17/06/03	9.7	46.6	45	122.3	209	15
Val: 18/01/03	8.8	29.6	22	65.6	299	143
	40.7

Figure 11. Outcome of filtering the data from devices with just one NH. Dataset: Val: 18/01/01, log scale.

5.2. Approach

To test our algorithms, we used the same two-pronged approach as in our previous study. It consists of running the algorithms on the raw and corrected data and comparing them with the Origin-Destination Survey (OSD for short), provided by the Chilean Transportation Authority. This survey gives us the ground test. The first thing we seek is to achieve the average daily trips indicated in the survey. Having secured these results, we could evaluate the distance variable, which is the main contribution of our work. The experiments carried out in the previous study used real data covering 12 days (275 million records). The previous experiments allowed us to reach an adequate generalization of NFPA-FTCA for different cities. On this opportunity, we execute the sequence NFPA-FTCA from the optimal parameters obtained in previous experiments, but this time fuzzified as indicated in Section 4.1, applied to a Valparaíso data set. Then, in a second step, we use the synthetic data from our previous work to check the efficiency in removing *NNH*. To this data set we add 2% artificial noise (*NNHs*), seeking to eliminate it when applying the algorithm. After testing the effectiveness of NFPA-FTCA on complete data sets, we made a selection of a subset from the original data. To do this, we selected devices (people) that had events in places with a high rate of *NNH*. These sites coincide with our interest in simulating areas of high interest for mobility studies (e.g., shopping centers, business areas, etc.). For each subset of data, we first apply FTCA and then the sequence NFPA-FTCA, to compare the results of trips and distances before and after fixing *NNHs*. To evaluate the performance of NFPA-FTCA, we compared its results with the Origin-Destination Surveys available in the Chilean Transport Authority database and our previous work.

6. Evaluation Results

6.1. Origin-Destination Survey (ODS)

The survey of origin and destination of trips in homes of Greater Valparaíso, consisted of surveying all the residents of 8600 homes of Greater Valparaíso, randomly selected, in the period between August 2014 and June 2015. These surveys are intended to collect information on the trips and the demographics of the people who make them, providing essential information for the development of transport models for the city [71]. The ODSs are expensive, slow to develop, and infrequently studied. Concepción, a major city in Chile, had a 350-day execution plan in 2015, which was unsuccessful. The last time this city carried out an ODS was in 1991 (see Table 5).

Table 5. General status of Origin-Destination Survey (ODS) in the main areas of Chile. The case of Concepción shows us that surveys do not always reach the expected results.

	Inhabitants ($\times 1000$)	Travels/Day ($\times 1000$)	Sampled Homes ($\times 1000$)	Year Study	Average Travels /Day-Person
Santiago	6652	18,461	18,264	2012	2.78
Valparaíso	928	2248	6800	2014	2.27
Concepción	N/A	N/A	8400	2015	N/A

The experiment makes a comparison of the average trips per device computed by FTCA with those in the survey. In Table 6, we see some results for Valparaíso.

Table 6. Gran Valparaíso 2014–2015 O-D survey details.

	Homes ×1000	People ×1000	Travels/Day ×1000	Average Travels /Day-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	2248.1	2.27

In the present work, we apply the new algorithms to Valparaíso macro area, for two different days: 1 January 2018, and one *BigMonday* (First Monday in March after vacations where classes and work activity begin), 6 March 2017. We use different support parameter configurations.

In Figure 12, we show the density of NHs in four specific locations in Greater Valparaíso on New Years, 1 January, 2018. We can see a high activity till dawn in celebration areas near the coast, and a completely different activity in the suburbs of the city.

Figure 12. New Years activities in Great Valparaíso. We see 4 different locations using real (green) and synthetic (red) data. 11NLI is a commercial area, ADUC2 is the epicenter of fireworks, RNCEB is a popular beach, and RODEC is a residential suburb.

In Table 7, we see the experiments of our previous (E_i) and current (FE_i) works differentiated based on the set of algorithms used. Based on our previous experiments (using NFA-TCA), we ran three new experiments in which we applied FNFA-FTCA (FE_2, FE_5, FE_8) adjusting the cores and supports points of the fuzzy sets centered on the binary parameters of the original experiments. The results show us an improvement in the results when compared with the ODS. As an example, considering maximum speeds of displacement in the city between 25 and 30 km/h (the core of “Normal” linguistic term of variable v), and assuming movements smaller than 4 blocks (the core of “very short” linguistic term of variable x) as no-trips (i.e., s_2 is actually sA), we get results between 20–25% better in the

calculation of trips (depending on the experiment). Both FTCA, like its binary counterpart TCA, also proved to be sensitive and accurate. At FE_2 , already with corrected data, we performed the same sensitivity test as in our previous work. In this case, instead of modifying a binary threshold, we increased the value of the right support (in the same magnitude, from 18 to 24 min) of one of the fuzzy sets of the variable *tos* (absorbing congestion scenarios). We obtained travel rates of 2.29, which is 18.9% better than the TCA value and more comparable to the ODS (2.27) in the case of Valparaíso.

Table 7. Some results of experiments from previous and current work. Compared data for Greater Valparaíso.

Experiment	Set of algorithms	1 January 2018				6 March 2017			
		Fixed NNH	% fixed records	Traces	Trips	Fixed NNH	% fixed records	Traces	Trips
Raw data	TCA	0	0.00	7.96	2.53	0	0.00	7.08	2.48
E_1	NFA-TCA	412,235	3.71	6.95	2.42	259,370	2.70	6.27	2.38
E_2	NFA-TCA	432,301	3.89	6.90	2.42	271,720	2.83	6.23	2.37
FE_2	FNFA-FTCA	610,111	5.49	6.06	2.25	383,096	3.99	6.02	2.29
E_3	NFA-TCA	457,942	4.12	6.83	2.39	288,164	3.00	6.18	2.36
E_4	NFA-TCA	494,580	4.45	6.74	2.39	311,901	3.25	6.11	2.35
E_5	NFA-TCA	528,122	4.75	6.77	2.35	322,591	3.36	6.14	2.36
FE_5	FNFA-FTCA	706,015	6.35	6.58	2.30	476,206	4.96	5.95	2.31
E_6	NFA-TCA	547,757	4.92	6.72	2.40	334,780	3.49	6.10	2.35
E_7	NFA-TCA	631,110	5.67	6.59	2.39	375,286	3.91	6.01	2.34
E_8	NFA-TCA	655,735	5.89	6.53	2.38	391,382	4.08	5.96	2.33
FE_8	FNFA-FTCA	760,652	6.83	6.01	2.21	454,003	4.73	5.48	2.16
E_9	NFA-TCA	691,407	6.22	6.44	2.37	414,402	4.32	5.89	2.32

E_8 was an experiment with broad parameters in transit times to a new site (see Section 4.2). This situation probably better reflects a rural setting than an urban one. When fuzzified at FE_8 , the results show a lower result than the ODS (2.16 vs. 2.27). This shows us that the selection of the support points of FE_2 is the best parameterization achieved in the experiments. The latter reinforces our main conclusion and contribution: the impact of NNHs occurs mainly in the field of the distances calculated in the trips rather than in the counting of them. The new approach presented in this study makes its contribution to an essentially diffused problem. With FTCA, we observed an average difference of traveled distances of 9.2 km, which is 9.8% better than the previous TCA (7.8 km). Compared with naive methods (without fixing the NNHs), the improvement rises from 19.6 to 28.8 km (46.9%).

6.2. Synthetic Data

After validating the results with the ODS survey, we used the same technique presented in [14] to generate synthetic data. Essentially, this method consists of replicating the statistical behavior of the real data in 15-minutes intervals. For each interval, we calculate the time distributions of events (NHs) and devices (*m*) that occur at each pair of sites in the antenna data set. This way, we capture the temporal patterns in the data. Then, synthetic data is generated by randomly extracting data from the created distributions. In order to generate a precision pattern for our algorithms, we add 2% noise (NNHs) generated with the same method from the noise identified in pairs of real antennas. In Figure 13, we observe that synthetic data have a good behavior to represent reality, even in particular scenarios such as New Year's Eve (see Figure 12). To test FNFA, we applied the algorithm with the group of Supporting Parameters from the experiment FE_8 and obtained 3.58%

fixed record. Then, we introduced an additional 2% noise, obtaining 5.56% of the corrected records, that is, 99% of the NNHs inserted.

(a) Density of NH per device

(b) Devices with NH=1 analysis

Figure 13. Comparison of statistical behavior between real and synthetic data. Once we confirmed adequate behavior, we introduced controlled noise to evaluate FNFA performance.

With high confidence, we can say that the results obtained in precedent Sections 6.1 and 6.2 are more comparable to the official surveys than our previous work. This allows us to move on to the next experiment, this time evaluating FNFA-FTCA with small groups of data representing small population groups or small geographic areas.

6.3. Applying FNFA and FTCA to Small Groups of Data

Lastly, we ran experiments applying the sequence FNFA-FTCA to small groups of data. With these experiments, we wanted to simulate small groups of people or small areas in the city and demonstrate two things. First, our improved algorithms are able to determine trips with greater precision than our previous version in the data subset (even if they are affected by the noise of type NNH). Second, to confirm that the previous approximations present important deviations in the distances calculated for the trips. All of the above aims to reduce the errors introduced in subsequent analyses and predictions developed with cellular data.

6.3.1. Computing Better Distance Distributions in Small Groups

The experiment consists of applying the sequence to the full data set of both original and fixed data. Although the GoogleMap API increases the distances traveled between sites by using more realistic and precise routes than the previous methods (i.e., L_2), we expected to reduce it by eliminating the ghost jumps that represent the NNHs. However, this apparent adverse effect is offset by the fact that the greater precision in the distances helps to detect more NNHs. In Figure 14, we can see average reductions of 19.8% in Valparaíso. The plots show us results comparing our approach with the naive one (orange and blue lines, respectively). In some cases, the computed distances are greater (Figure 14b). In them, the explanation is based on the type of day. Normally a *BigMonday* involves an abnormal amount of trips due to the end of summer and the beginning of work and school activities. The amount of NNHs in those days becomes more evident (higher jumps in the network), generating a greater elimination of records in the naive approach.

(a) Average velocity in Valparaíso (2018)

(b) Cumulative distance in Valparaíso (2018)

(c) Average velocity in Valparaíso (2017)

(d) Cumulative distance in Valparaíso (2017)

Figure 14. Naive and proposed approaches evaluated on full data sets for 1 January 2018, and 6 March 2017 in Valparaíso.

6.3.2. Analyzing Travels and Distance Distributions in Small Groups

The experiment selects, from the original data, pairs of antennas (s_{t-1} and s_t) with a high rate of NNHs between them. To obtain this subset, two filters were applied: pairs of sites with more than 1000 NNHs reported in one day and more than 2 km faraway each other. For each site combination we select all the m devices that passed through them respecting the direction of movement (i.e., from s_{t-1} to s_t). As expected, the trip count for device m does not alter too much. If the trip has started and an incorrect NH is obtained during it (this is an NNH), the trip counter ($trips+ = 1$) does not increment. On the contrary, a significant impact is observed in the computation of the distance. This is why the travel distributions look similar (see Figure 15), but there are significant differences in the distance distributions (see Figure 16). In all cases, when comparing both approaches, the distance is reduced by at least 40%, as shown in Figure 16b,d.

(a) Trips distributions in Valparaíso (2017)

(b) Trips distributions in Valparaíso (2018)

Figure 15. Trips distributions computed with naive and proposed approach—6 March 2017 and 1 March 2018, respectively. Travel distributions look similar.

(a) Average velocity in Valparaíso (2018)

(b) Cumulative distance in Valparaíso (2018)

Figure 16. Cont.

(c) Average velocity in Valparaíso (2017)

(d) Cumulative distance in Valparaíso (2017)

Figure 16. Naive and proposed approaches evaluated on subsets of data for 1 January 2018, and 6 March 2017 in Valparaíso. When we analyze subsets of sites with a high rate of NNHs, the accumulated distances have important deviations in the naive approximations.

7. Future Work and Conclusions

Mobile data in general, and CDRs in particular, have undoubtedly become an important source for research and development of products and services. They are collected for all active mobile telecommunications devices, which to date total more than 5.3 billion in the world. With the advent of the popularly called BigData, they have begun to be collected to operate networks efficiently, improve user business intelligence, study urban mobility in a wide range of use cases, and, lately, as data to improve society.

Our work leaves us with several conclusions and future experiments. The relevance that this type of data has achieved in various fields of knowledge and innovation motivates us to understand the impacts that erroneous NHs (NNHs) can introduce. To the best of our knowledge, all previous studies using this type of data have worked with well-curated or synthetic data, eliminating the potential impact of incorrect records only as far as they can recognize them. These works only spotted the obvious problems in the data, but due to the inherent complexities of communications networks, there are many more to discover. Our work has shown that removing and maintaining these incorrect records without correcting them negatively influences subsequent analyses. In terms of travel counts, the impact is relatively minor; however, we observe considerable deviations in terms of the distances traveled and their distributions. This deviation is increased if we reduce the space-time scale of the studies.

Regarding Origin and Destination Surveys (ODS), our conclusions coincide with those collected in previous studies. Surveys are expensive, slow, static, very sporadic in time, and not always successful. The samples used in these surveys seek to be representative of cities or macro-areas and not of spatially reduced units of analysis. The main objective of these surveys is to collect information about the trips and the demographics of the people who make them, providing essential information for the development of transport models for the city [72]. The information obtained from them (distances traveled, number of trips, modal split, etc.) is directly related to the quality and general frequency of the process. Our work is inspired by solving these problems by delivering consistent results in terms of trips, significantly improving the distances traveled, and, above all, making the granularity in the analysis more flexible.

CDRs are sparse and geographically inaccurate data. Some years ago, Operators (Telcos) have started to store core network data (XDRs). These data are less sparse and more accurate compared to CDRs (35 versus 600 NH per device-day). However, they are more expensive to capture, store, and process. Like any new data source, it still has the typical problems of raw and uncured data. Our work will undoubtedly help reduce these

complexities. CDRs will continue to be used intensively in the research and development of solutions. Its simplicity, massiveness, and low cost of generation and storage will continue to contribute in many areas. Of course, there are still spaces for improvement in its management and understanding.

Today we see an intensive use of these data in the generation of impact research in the following areas:

Mobility and transportation: Mobility and transportation are vital elements for inhabitants of cities, and a key dimension on authorities agenda.

Smart economy: Knowing where people—residents, and visitors—are concentrated, for what types of activities, and at what times, allows commercial and cultural institutions to improve their offer and segment them based on patterns that include mobility and its semantics.

Public health and safety: The ability to infer population mobility for health and safety reasons is a key aspect of governments' agenda. This ability must include large social events, crisis management, earthquakes, social riots, etc. Today, it is especially relevant to study the impact of urban mobility as a vector of contagion of diseases such as COVID-19.

Land use and sustainability: Urban planning has become a complex challenge; understanding the dynamics of how citizens use urban and suburban spaces is essential for urban planning and the sustainable development of cities.

Social good: correlated with the concept of Smart City, understanding the mobility of specific segments of inhabitants can improve the quality of life in a large list of areas, and help reduce the inequalities that are now visible in our societies.

Our work contributes to improvements in the data source for these studies, making them more reliable and accurate. The current segmentation of users and consumers are not able to adequately model the current complexity. It is necessary to perform analyses at much smaller scales, in greater detail. It is at this point where cellular data cannot hide its problems, and it is there where our work makes its greatest contribution.

We decided to use fuzzy logic as a simple way to understand a complex phenomenon. Fuzzy logic is a methodology that provides a simple way to draw conclusions from ambiguous, imprecise, and incomplete data. In future work, applying more sophisticated fuzzy logic or neural network approaches should give us even better results not only in the distribution of trips and distances but also in dimensions such as modal partition and the purpose of trips. Both issues are not resolved effectively yet. Obtaining these same conclusions in real-time adds new benefits in many fields. We still see a challenge in applying fuzzy logic models in Big Data scenarios.

Synthetic data generation is a reopened area. A few years ago, synthetic data was created to compensate for the lack of it and for an incipient concern about people's privacy. The GDPR (General Data Protection Regulation 2016/679) and other new regulations begin to generate concrete restrictions on the use of this type of data [73]. There is a new opportunity to develop new algorithms to create synthetic data that adequately model the inherent aspects of people's mobility and interactions. The benefits of having synthetic data are given by the possibility of conducting much more research in the areas indicated above, without putting 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 using real data. At this point, the promise of impact value should be greater than the potential vulnerabilities.

Author Contributions: Conceptualization, A.L.-A. and H.A.-C.; data curation, A.L.-A.; investigation, A.L.-A.; methodology, A.L.-A. and H.A.-C.; resources, H.A.-C.; supervision, H.A.-C.; validation, A.L.-A. and H.A.-C.; writing—original draft, A.L.-A.; writing—review and editing, H.A.-C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Pinelli, F.; Di Lorenzo, G.; Calabrese, F. Comparing urban sensing applications using event and network-driven mobile phone location data. In Proceedings of the 2015 16th IEEE International Conference on Mobile Data Management, Pittsburgh, PA, USA, 15–18 June 2015; Volume 1, pp. 219–226.
- Graells-Garrido, E.; Peredo, O.; García, J. Sensing urban patterns with antenna mappings: the case of Santiago, Chile. *Sensors* **2016**, *16*, 1098. [CrossRef] [PubMed]
- Calabrese, F.; Ferrari, L.; Blondel, V.D. Urban sensing using mobile phone network data: A survey of research. *ACM Comput. Surv.* **2015**, *47*, 25. [CrossRef]
- Gakenheimer, R. Urban mobility in the developing world. *Transp. Res. Part Policy Pract.* **1999**, *33*, 671–689. [CrossRef]
- Oliver, N.; Lepri, B.; Sterly, H.; Lambiotte, R.; Deletaille, S.; De Nadai, M.; Letouze, E.; Salah, A.A.; Benjamins, R.; Cattuto, C.; et al. Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle. *Sci. Adv.* **2020**, *6*. [CrossRef]
- Groves, R.M. Nonresponse rates and nonresponse bias in household surveys. *Public Opin. Q.* **2006**, *70*, 646–675. [CrossRef]
- Kuwahara, M.; Sullivan, E.C. Estimating origin-destination matrices from roadside survey data. *Transp. Res. Part Methodol.* **1987**, *21*, 233–248. [CrossRef]
- Intelligence, G. Definitive Data and Analysis for the Mobile Industry. *GSMA-Intelligence*. 2019. Available online: https://www.gsma.com/services/wp-content/uploads/2019/06/GSMAIntelligence_Product_Brochure_2019.pdf (accessed on 5 December 2020).
- Blondel, V.D.; Decuyper, A.; Krings, G. A survey of results on mobile phone datasets analysis. *EPJ Data Sci.* **2015**, *4*, 10. [CrossRef]
- Pan, G.; Qi, G.; Zhang, W.; Li, S.; Wu, Z.; Yang, L.T. Trace analysis and mining for smart cities: Issues, methods, and applications. *IEEE Commun. Mag.* **2013**, *51*, 120–126. [CrossRef]
- Gonzalez, M.C.; Hidalgo, C.A.; Barabasi, A.L. Understanding individual human mobility patterns. *Nature* **2008**, *453*, 779. [CrossRef]
- Calabrese, F.; Pereira, F.C.; Di Lorenzo, G.; Liang, L.; Ratti, C. The geography of taste: Analyzing cell-phone mobility and social events. In *International Conference on Pervasive Computing*; Springer: Berlin/Heidelberg, Germany, 2010; Volume 10, pp. 22–37.
- Song, C.; Qu, Z.; Blumm, N.; Barabási, A.L. Limits of predictability in human mobility. *Science* **2010**, *327*, 1018–1021. [CrossRef]
- Leiva-Araos, A.; Allende-Cid, H.; Khryashchev, D.; Vo, H.T. Tackling the Neighboring Network Hit Problem in Cellular Data. In Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 9–12 December 2019; pp. 2344–2353.
- Zimmermann, H. OSI reference model—The ISO model of architecture for open systems interconnection. *IEEE Trans. Commun.* **1980**, *28*, 425–432. [CrossRef]
- Damnjanovic, A.; Montojo, J.; Wei, Y.; Ji, T.; Luo, T.; Vajapeyam, M.; Yoo, T.; Song, O.; Malladi, D. A survey on 3GPP heterogeneous networks. *IEEE Wirel. Commun.* **2011**, *18*, 10–21. [CrossRef]
- De Jonge, E.; van Pelt, M.; Roos, M. *Time Patterns, Geospatial Clustering and Mobility Statistics Based on Mobile Phone Network Data*; Statistics Netherlands: The Hague, The Netherlands, 2012.
- Song, C.; Koren, T.; Wang, P.; Barabási, A.L. Modelling the scaling properties of human mobility. *Nat. Phys.* **2010**, *6*, 818. [CrossRef]
- Wang, D.; Pedreschi, D.; Song, C.; Giannotti, F.; Barabasi, A.L. Human mobility, social ties, and link prediction. In Proceedings of the 17th ACM SIGKDD, International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 21–24 August 2011; pp. 1100–1108.
- Isaacman, S.; Becker, R.; Cáceres, R.; Martonosi, M.; Rowland, J.; Varshavsky, A.; Willinger, W. Human mobility modeling at metropolitan scales. In Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services, Low Wood Bay, Ambleside, UK, 25–29 June 2012; pp. 239–252.
- Mir, D.J.; Isaacman, S.; Cáceres, R.; Martonosi, M.; Wright, R.N. Dp-where: Differentially private modeling of human mobility. In Proceedings of the 2013 IEEE International Conference on Big Data, Silicon Valley, CA, USA, 6–9 October 2013; pp. 580–588.
- Graells-Garrido, E.; García, J. Visual exploration of urban dynamics using mobile data. In *International Conference on Ubiquitous Computing and Ambient Intelligence*; Springer: Cham, Switzerland, 2015; pp. 480–491.
- Graells-Garrido, E.; Ferrer, L.; Caro, D.; Bravo, L. The effect of Pokémon Go on the pulse of the city: A natural experiment. *EPJ Data Sci.* **2017**, *6*, 23. [CrossRef]
- Isaacman, S.; Becker, R.; Cáceres, R.; Kobourov, S.; Martonosi, M.; Rowland, J.; Varshavsky, A. Identifying important places in people’s lives from cellular network data. In *International Conference on Pervasive Computing*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 133–151.
- Reades, J.; Calabrese, F.; Sevtsuk, A.; Ratti, C. Cellular census: Explorations in urban data collection. *IEEE Pervasive Comput.* **2007**, *6*, 30–38. [CrossRef]
- Onnela, J.P.; Saramäki, J.; Hyvönen, J.; Szabó, G.; Lazer, D.; Kaski, K.; Kertész, J.; Barabási, A.L. Structure and tie strengths in mobile communication networks. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 7332–7336. [CrossRef] [PubMed]

27. Ferres, L. Problems and Opportunities of Working with a Telco's Large Data Sets of Mobile Data. In Proceedings of the Companion Proceedings of The 2019 World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019; pp. 229.
28. Beiró, M.G.; Bravo, L.; Caro, D.; Cattuto, C.; Ferres, L.; Graells-Garrido, E. Shopping mall attraction and social mixing at a city scale. *EPJ Data Sci.* **2018**, *7*, 28. [\[CrossRef\]](#)
29. Ferreira, N.; Poco, J.; Vo, H.T.; Freire, J.; Silva, C.T. Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. *IEEE Trans. Vis. Comput. Graph.* **2013**, *19*, 2149–2158. [\[CrossRef\]](#)
30. Freire, J.; Bessa, A.; Chirigati, F.; Vo, H.; Zhao, K. Exploring What not to Clean in Urban Data: A Study Using New York City Taxi Trips. *Data Eng.* **2016**, *39*, 63–77.
31. Intelligence, G. The Mobile Economy 2020 GSMA-Intelligence. 2021. Available online: <https://data.gsmaintelligence.com/api-web/v2/research-file-download?id=51249388&file=2915-260220-Mobile-Economy.pdf> (accessed on 5 December 2021)
32. Lambiotte, R.; Blondel, V.D.; De Kerchove, C.; Huens, E.; Prieur, C.; Smoreda, Z.; Van Dooren, P. Geographical dispersal of mobile communication networks. *Phys. A Stat. Mech. Its Appl.* **2008**, *387*, 5317–5325. [\[CrossRef\]](#)
33. Laetitia, G.; Michele, T.; Simone, P.; Young, A.; Adler, N.; Stefaan, V.; Ferres, L.; Ciro, C. Gender gaps in urban mobility. *Palgrave Commun.* **2020**, *7*, 1–3.
34. Ahas, R.; Silm, S.; Järv, O.; Saluveer, E.; Tiru, M. Using mobile positioning data to model locations meaningful to users of mobile phones. *J. Urban Technol.* **2010**, *17*, 3–27. [\[CrossRef\]](#)
35. Nurmi, P.; Bhattacharya, S. Identifying meaningful places: The non-parametric way. In *International Conference on Pervasive Computing*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 111–127.
36. Becker, R.A.; Caceres, R.; Hanson, K.; Loh, J.M.; Urbanek, S.; Varshavsky, A.; Volinsky, C. Route classification using cellular handoff patterns. In Proceedings of the 13th International Conference on Ubiquitous Computing, Beijing, China, 17–21 September 2011; pp. 123–132.
37. Girardin, F.; Vaccari, A.; Gerber, A.; Biderman, A.; Ratti, C. Quantifying urban attractiveness from the distribution and density of digital footprints. *Int. J.* **2009**, *4*, 175–200.
38. Soto, V.; Frias-Martinez, E. 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, San Francisco, CA, USA, 12 June 2011.
39. Farrahi, K.; Gatica-Perez, D. What did you do today? Discovering daily routines from large-scale mobile data. In Proceedings of the 16th ACM International Conference on Multimedia, Vancouver, BC, Canada, 26–31 October 2008; pp. 849–852.
40. Buckee, C.O.; Balsari, S.; Chan, J.; Crosas, M.; Dominici, F.; Gasser, U.; Grad, Y.H.; Grenfell, B.; Halloran, M.E.; Kraemer, M.U.; et al. Aggregated mobility data could help fight COVID-19. *Science* **2020**, *368*, 145. [\[CrossRef\]](#)
41. Steenbruggen, J.; Tranos, E.; Nijkamp, P. Data from mobile phone operators: A tool for smarter cities? *Telecommun. Policy* **2015**, *39*, 335–346. [\[CrossRef\]](#)
42. Krisp, J.M. Planning fire and rescue services by visualizing mobile phone density. *J. Urban Technol.* **2010**, *17*, 61–69. [\[CrossRef\]](#)
43. Peters, S.; Krisp, J.M. Density calculation for moving points. In Proceedings of the 13th AGILE International Conference on Geographic Information Science, Guimaraes, Portugal, 11–14 May 2010; Volume 1014.
44. Soto, V.; Frias-Martinez, V.; Virseda, J.; Frias-Martinez, E. Prediction of socioeconomic levels using cell phone records. In *International Conference on User Modeling, Adaptation, and Personalization*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 377–388.
45. Simini, F.; González, M.C.; Maritan, A.; Barabási, A.L. A universal model for mobility and migration patterns. *Nature* **2012**, *484*, 96–100. [\[CrossRef\]](#) [\[PubMed\]](#)
46. Wang, P.; Hunter, T.; Bayen, A.M.; Schechtner, K.; González, M.C. Understanding road usage patterns in urban areas. *Sci. Rep.* **2012**, *2*, 1001. [\[CrossRef\]](#)
47. Calabrese, F.; Colonna, M.; Lovisolo, P.; Parata, D.; Ratti, C. Real-time urban monitoring using cell phones: A case study in Rome. *IEEE Trans. Intell. Transp. Syst.* **2011**, *12*, 141–151. [\[CrossRef\]](#)
48. Lu, X.; Wetter, E.; Bharti, N.; Tatem, A.J.; Bengtsson, L. Approaching the limit of predictability in human mobility. *Sci. Rep.* **2013**, *3*, 1–9. [\[CrossRef\]](#) [\[PubMed\]](#)
49. Bagrow, J.P.; Wang, D.; Barabasi, A.L. Collective response of human populations to large-scale emergencies. *PLoS ONE* **2011**, *6*, e17680. [\[CrossRef\]](#) [\[PubMed\]](#)
50. Lu, X.; Bengtsson, L.; Holme, P. Predictability of population displacement after the 2010 Haiti earthquake. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 11576–11581. [\[CrossRef\]](#) [\[PubMed\]](#)
51. Ferrari, L.; Mamei, M.; Colonna, M. People get together on special events: Discovering happenings in the city via cell network analysis. In Proceedings of the 2012 IEEE International Conference on Pervasive Computing and Communications Workshops, Lugano, Switzerland, 19–23 March 2012; pp. 223–228.
52. Traag, V.A.; Browet, A.; Calabrese, F.; Morlot, F. Social event detection in massive mobile phone data using probabilistic location inference. In Proceedings of the 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, Boston, MA, USA, 9–11 October 2011; pp. 625–628.
53. Blondel, V.; Krings, G.; Thomas, I. Regions and borders of mobile telephony in Belgium and in the Brussels metropolitan zone. *Brussels Studies. La Revue Scientifique électronique pour les Recherches sur Bruxelles/Het Elektronisch Wetenschappelijk Tijdschrift voor Onderzoek over Brussel/J. Acad. Res. Bruss.* **2010**. [\[CrossRef\]](#)

54. Calabrese, F.; Dahlem, D.; Gerber, A.; Paul, D.; Chen, X.; Rowland, J.; Rath, C.; Ratti, C. The connected states of america: Quantifying social radii of influence. In Proceedings of the 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, Boston, MA, USA, 9–11 October 2011; pp. 223–230.
55. Couronne, T.; Olteanu, A.M.; Smoreda, Z. Urban mobility: Velocity and uncertainty in mobile phone data. In Proceedings of the 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, Boston, MA, USA, 9–11 October 2011; pp. 1425–1430.
56. Pokhriyal, N.; Dong, W.; Govindaraju, V. Virtual networks and poverty analysis in senegal. *arXiv* **2015**, arXiv:1506.03401.
57. Martinez-Cesena, E.A.; Mancarella, P.; Ndiaye, M.; Schlöpfer, M. Using mobile phone data for electricity infrastructure planning. *arXiv* **2015**, arXiv:1504.03899.
58. Hossain, S.; Abtahee, A.; Kashem, I.; Hoque, M.M.; Sarker, I.H. Crime Prediction Using Spatio-Temporal Data. In *International Conference on Computing Science, Communication and Security*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 277–289.
59. Klein, B.; LaRocky, T.; McCabey, S.; Torresy, L.; Privitera, F.; Lake, B.; Kraemer, M.U.; Brownstein, J.S.; Lazer, D.; Eliassi-Rad, T.; et al. Assessing Changes in Commuting and Individual Mobility in Major Metropolitan Areas in the United States during the COVID-19 Outbreak. Network Science Institute, Northeastern University. 31 March 2020. Available online: https://uploads-ssl.webflow.com/5c9104426f6f88ac129ef3d2/5e8374ee75221201609ab586_Assessing_mobility_changes_in_the_United_States_during_the_COVID_19_outbreak.pdf (accessed on 5 December 2020)
60. Zhao, K.; Tarkoma, S.; Liu, S.; Vo, H. Urban human mobility data mining: An overview. In Proceedings of the 2016 IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 5–8 December 2016; pp. 1911–1920.
61. Zadeh, L.A. Information and control. *Fuzzy Sets* **1965**, *8*, 338–353.
62. Mamdani, E.H. Application of fuzzy algorithms for control of simple dynamic plant. In *Proceedings of the Institution of Electrical Engineers*; IET: London, UK, 1974; Volume 121, pp. 1585–1588. [\[CrossRef\]](#)
63. De Luca, A.; Termini, S. A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory. In *Readings in Fuzzy Sets for Intelligent Systems*; Elsevier: Amsterdam, The Netherlands, 1993; pp. 197–202.
64. Dubois, D.J. *Fuzzy Sets and Systems: Theory and Applications*; Academic Press: Cambridge, MA, USA, 1980; Volume 144.
65. Torra, V. A review of the construction of hierarchical fuzzy systems. *Int. J. Intell. Syst.* **2002**, *17*, 531–543. [\[CrossRef\]](#)
66. Stufflebeam, J.; Prasad, N.R. Hierarchical fuzzy control. In Proceedings of the FUZZ-IEEE'99. 1999 IEEE International Fuzzy Systems. Conference Proceedings (Cat. No. 99CH36315), Seoul, Korea, 22–25 August 1999; Volume 1, pp. 498–503.
67. Yager, R.R. On a hierarchical structure for fuzzy modeling and control. *IEEE Trans. Syst. Man, Cybern.* **1993**, *23*, 1189–1197. [\[CrossRef\]](#)
68. Jamshidi, M. *Large-Scale Systems: Modeling, Control, and Fuzzy Logic*; Prentice-Hall, Inc.: New Jersey, NJ, USA, 1996.
69. Takagi, T.; Sugeno, M. Fuzzy identification of systems and its applications to modeling and control. *IEEE Trans. Syst. Man Cybern.* **1985**, 116–132. [\[CrossRef\]](#)
70. Li, X.; Pan, G.; Wu, Z.; Qi, G.; Li, S.; Zhang, D.; Zhang, W.; Wang, Z. Prediction of urban human mobility using large-scale taxi traces and its applications. *Front. Comput. Sci.* **2012**, *6*, 111–121.
71. Hurtado, O.S.U.A. *Informe Ejecutivo, EOD de Viajes—Santiago 2012*; Biblioteca Sectra: Santiago, Chile, 2015.
72. Kong, X.; Li, M.; Ma, K.; Tian, K.; Wang, M.; Ning, Z.; Xia, F. Big trajectory data: A survey of applications and services. *IEEE Access* **2018**, *6*, 58295–58306. [\[CrossRef\]](#)
73. Zarsky, T.Z. Incompatible: The GDPR in the age of big data. *Seton Hall L. Rev.* **2016**, *47*, 995.