NOVA IMS
Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

How Much Data Is Enough To Track Tourists?

The Tradeoff Between Data Granularity And Storage Costs

Inês Correia Pereira

Dissertation presented as partial requirement for obtaining the Master's degree in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

HOW MUCH DATA IS ENOUGH TO TRACK TOURISTS? – THE TRADEOFF BETWEEN DATA GRANULARITY AND STORAGE COSTS

by

Inês Correia Pereira

Dissertation presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Business Intelligence and Knowledge Management

Co Advisor: Leid Zejnilović

Co Advisor*:* Leonardo Vanneschi

November 2019

ACKNOWLEDGEMENTS

First and foremost, I would like to express my most sincere appreciation and gratitude to my supervisors, Professor Leid Zejnilović from Nova School of Business and Economics, and Professor Leonardo Vanneschi from Nova School of Information Management.

I am extremely grateful not only for the Professor Leid Zejnilović's help during the making of this Master's Thesis, but during the duration of my Master's degree in general. The professor has continuously put more trust into my work and ability than I could fandom and has assisted me in achieving the proposed goals every step of the way. I am most thankful for the opportunities that he has provided me in integrating the staff team of the Data Science for Social Good 2018 Summer Fellowship and later on the team at the Data Science Knowledge Center at Nova School of Business and Economics. Both these experiences allowed me to learn more about the practical applications of data science and analytics than any classroom could, and have had an extreme impact on the making of this thesis. His vision and enthusiasm for his work continue to inspire me. It was a great honor to work under his most helpful guidance. In addition, I would like to extend my deepest appreciation to his wife, Sabina Zejnilović, whose help and care is absolutely unrepayable. The discussions I had with Sabina guided me, not only on the right path to take in regards to the thesis, but have most importantly taught me the meaning of selfless care for others. In many stressful points during the making of this thesis, Sabina was there with a guiding hand and a calm spirit. She has inspired me most of all and motivated me when hope seemed lost. I am certain that I would not have been able to accomplish this thesis without her. For that, I will be eternally grateful.

I would also like to show my gratitude for Professor Leonardo Vanneschi's time and attention. Even though the Professor had a handful of Master Thesis students under his advisory, he kindly accepted me under his guidance and helped me in meeting the deadlines. The delivery of this Master's Thesis would not have been possible without his help and collaboration.

Furthermore, I am ever so thankful to Professor Qiwei Han from Nova School of Business and Economics for his patience and help throughout the making of this thesis. He has kindly met with me and answered my many of my questions in multiple occasions and his assistance has had a significant impact on this study. His words of encouragement and passion for his work continue to inspire me and I am thankful for all his help.

I would also like to extend my sincerest appreciation to the members of the Data Science Knowledge Center at Nova School of Business and Economics during the last year. Namely, I would like to show gratitude to Lénia Mestrinho, Andrew Bell, Iñigo de Troya, Susana Lavado, Dierê Fernandez and João Fonseca. Their encouragement has meant the world to me and kept me going in a handful of stressful moments.

In addition, I would like to show my most genuine thankfulness to my friends and family who have supported me throughout this year. Every word of encouragement from them has impacted the outcome of this thesis greatly. I would like to show my gratefulness, in particular, to Leonor Braz Teixeira and João Veiga, who have been my closest accomplices during this demanding degree, and to my close friends Maria Inês Rico, Maria Alves and Elisa Clemente for helping me review my work.

Finally, but definitely most importantly, I would like to show my truest gratitude to my mum and dad. Anything that I have accomplished has been for them and because of them. Nothing can describe how instrumental they were during my Master’s degree and during the making of this thesis in particular. Not only would this adventure have been impossible without their financial support, but it would have also been impossible without their continuous love and care. They have always been my greatest teachers in life and I can honestly say that this thesis was written to make them proud.

ABSTRACT

In the increasingly technology-dependent world, data is one of the key strategic resources for organizations. Often, the challenge that many decision-makers face is to determine which data and how much to collect, and what needs to be kept in their data storage. The challenge is to preserve enough information to inform decisions but doing so without overly high costs of storage and data processing cost. In this thesis, this challenge is studied in the context of a collection of mobile signaling data for studying tourists' behavioral patterns. Given the number of mobile phones in use, and frequency of their interaction with network infrastructure and location reporting, mobile data sets represent a rich source of information for mobility studies. The objective of this research is to analyze to what extent can individual trajectories be reconstructed if only a fraction of the original location data is preserved, providing insights about the tradeoff between the volume of data available and the accuracy of reconstructed paths. To achieve this, a signaling data of 277,093 anonymized foreign travelers is sampled with different sampling rates, and the full trajectories are reconstructed, using the last seen, linear, and cubic interpolations completion methods. The results of the comparison are discussed from the perspective of data management and implications on the research, especially the results of research with lower time-density mobile phone data.

KEYWORDS

Tourism Mobility; Trajectory Reconstruction; Call Detail Records; Signaling Data; Data Sparsity

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

CDR	Call Detail Record
GPS	Global Positioning System
GSM	Global System for Mobile Communication
LAU	Location Area Updates
T-SNE	T-Distributed Stochastic Neighbor Embedding

1. INTRODUCTION

1.1. BACKGROUND AND PROBLEM IDENTIFICATION

Human mobility has long been a subject of interest as understanding it informs diverse disciplines, such as Physics, Sociology, Epidemiology, Transportation, or Complex Systems (Naboulsi et al., 2016). Human mobility's applications range from epidemics (Belik et al., 2011; Colizza et al., 2007; Hufbagel et al., 2004) and the spread of computer viruses (Kleinberg, 2007), to city planning (Horner & O'Kelly, 2001) and traffic forecasting (Kitamura et al., 2000; Calabrese et al. 2011). The understanding of individual mobility patterns can also inform tourism management to design targeted marketing for tourist destinations and attractions, or help in designing and managing attractions the on-site movement (Zheng et al., 2017).

There is an evidence that one can predict individual mobility, since individuals tend to follow regular patterns when moving in a specific location despite uniqueness in individual trajectories and uncertainty in spatio-temporal estimates (Gonzalez et al., 2008; Song et al., 2010; De Montjoye et al., 2013; Iovan et al., 2013). Home-work commutes are a good example of regular patterns in individual mobility (Ahas et al., 2009).

Telecom operator data with mobile phone call detail records is quite popular among researchers, favored for encompassing information about a large percentage of the population over an extended time (Naboulsi et al., 2016). Call Detail Records (CDR) register all the details of an individual's activity over a telecommunication network, except the content of the activity. The CDR's contain information about the registered device, location of a base station where the device is registered, the date- or time-stamp of an event, and the type of the event – a short text message, a call, or a mobile data traffic session (Chen et al., 2018). In 2016, Hoteit et al. referenced as the CDR a dataset that includes data activity over network, significantly increasing the time density of individual location records, as data activity occurs much more frequently than calls and text messages.

Studying individual mobility with the CDR's requires researchers to be aware of its limitations. For instance, Chen et al. (2018) found that: i) trajectories created using CDR's are unevenly distributed in space and time, ii) the spatial data is constrained by the antenna's coverage, iii) the temporal data is sparse and varies from individual to individual, and, iv) due to data sparsity, a big percentage of individual's trajectories are not usable because of the insufficient sampling frequency of the user's movement. For example, Gonzalez et al. (2008), Song, et al. (2010) and Hoteit et al. (2014) kept only 1.67%, 0.45% and 0.07% of their CDR dataset, respectively, after the removal of the individuals whose trajectories did not possess a high enough level of mobility information. Chen et al. (2018) tried to mitigate this issue by using different CDR completion methods to make functional trajectories of the incomplete datasets. This approach was possible thanks to the redundancy, inter-call stability, and nighttime stability that characterize human mobility. Redundancy is very important to ensure that trajectories are predictable and, thus, can be estimated with some accuracy. However, this assumption may not hold as strongly in the context of tourists, as the highly predictable home-work commute does not occur.

Other than CDR's, signaling traffic data collected from cellular devices has also been used in mobility studies (Zhao et al., 2018). Signaling data are the beacons periodically sent by mobile devices to update about their location and the information about the signal strength between the device and its closest base station. Since the signaling event occurs both when a device is idle at a single location, or moving and changing locations, signaling data represents a more accurate depiction of the individual mobility than the CDR's (Fiadino et al., 2012). While there is potential value in data that contain high-time-density of location information for every individual connected to a mobile infrastructure of a telecom provider, storing such amount of data and processing is a challenging and costly task.

1.2. STUDY OBJECTIVES

This thesis seeks to study the tradeoff between the time granularity of phone records and the quality of the estimated tourist trajectories. Hence, the research question: *How much data is needed to learn tourist movements?* To answer this question, a comparison of the trajectories of the same individuals is conducted, constructed from a dataset generated by sampling with variable sampling rate from the signaling data of a European telecom provider. In the case where the data is sparse, the trajectory completion is conducted by applying several completion methods and the outcomes are compared to the trajectories generated from the original dataset.

In order to understand the implications of the results obtained from the application of the different sampling and completion techniques, some basic mobility analysis is performed for comparison. This is achieved by looking into the relationship of commonly used mobility metrics (radius of gyration, total travel distance, and lifetime in days and in hours) and the estimation error to understand for which type of tourists these methods work the best and the worse. Finally, the trajectories of some tourists with different mobility profiles and their stops are plotted, and a k-means is performed to cluster similar types of tourists together. These analyses are made in order to understand under which scenarios changing the sampling rate and/or completion method impacts the results of some algorithms that are typically performed by researchers.

2. LITERATURE REVIEW

2.1. THE STUDY OF HUMAN MOBILITY

Understanding how individuals move by figuring out which locations they visit and the amount of time they stay at each location is extremely important for governments and businesses alike, since its practical implications are manifold. The study of human mobility has applications in different disciplines, such as Physics, Sociology, Epidemiology, Transportation and Networking (Naboulsi et al., 2016) and the level of detail of the study may be done at a macro or micro-level. Whereas at a macro level, one would study the movement of the individuals at a country-level, for example, a country's tourism seasonality (Ahas et al., 2007a), at a micro-level one would look at a smaller maximum span of movement, for example, looking into how individuals move inside a university campus (Bhattacharjee et al., 2004).

Moreover, there are several different data types that are used to study mobility. An extensive literature review conducted by Naboulsi et al. (2016) showed that several human mobility studies have used personal mobile communication data as the main data source. Data stemming from the personal mobile communications is beneficial for these types of studies since it encompasses a larger percentage of the population, depending on the market penetration the mobile operator has, over a period of time that can, theoretically, be as short or as extended as one wishes.

Previous studies have applied personal mobile communications to: the detection of homes (Vanhoof et al., 2018), extraction behavior patterns and identifying user's locations as home and workspace (Leng et al., 2018), exhibition of user's route choice behaviors in long-distance inter-regional journeys (Bwambale et al., 2019), detection of the mode of transportation (Huang et al., 2019), study of patterns in communities (Jundee et al., 2018), clustering of users according to their movement patterns (Tomasini et al., 2018), relationship between mobility and social networks (Puura et al., 2018), and monitoring of food security (Zufiria et al., 2018), to name a few.

Furthermore, it is also relevant to mention that these mobile communications data types can be combined with other sources for a more holistic study. For instance, Oliver & Enrique (2014) combined credit card data collected at BBVA terminals with mobile positioning data to study patterns in tourism flows in Barcelona, while Girardin et al. (2008) compared the use of mobile positioning data with the use of user-generated content of users that use Flickr to store and share their photos, which in turn can be imbedded with geographical attributes to study tourism as well.

There are two types of location data collection in cellular phone networks: active and passive localization (Smoreda et al., 2013). Active localization records positioning data the way a classic survey would, with high granularity over time. In this case, the data gathered from the mobile phones is collected with the purpose of tracing the individual's mobility patterns for a given study. On the other hand, passive localization is used to extract information from data that was collected for purposes other than the conduction of a study, such as data collected for billing, and hence, it is able to provide information on a much broader population.

CDR's (Call Detail Records) are the most popular passive localization records extracted from cellphone data used by researchers (Smoreda et al., 2013) and refer exactly to billing data. However, this is not the only type of passive data. According to Iovan et al. (2013), in addition to CDR's there are two types of mobile passive data: Wi-Fi data and probe data. Wi-Fi data is collected by monitoring the usage of a certain Wi-Fi network (Song et al., 2004) while probe data is issued from mobile network probes. They are anonymized and contain information on communication events such as phone calls and SMS's (incoming or outgoing) as well as handover events (which correspond to antennas' updates during a call) and Location Area Updates (LAU) (Iovan et al., 2013). It is important to notice that datasets composed by either CDR's and probe data records are sparse in space and time, and their spatial features are an approximation of the actual location of the mobile user.

Finally, it is important to state that researchers have also used Global Positioning System (GPS) data to track mobility. Calabrese et al. (2010) found that while GPS provided better data for the implementation of an intelligent transportation system than fixed sensors would, they still posed several issues that limited the study such as: the fact that a significant percentage of the vehicles needed to have a GPS device installed, that mass deployment was costly, that the transmission cost became heavy with high frequency, that GPS's coverage was limited in urban areas (because of canyon obstructions and signal multipath) and due to legislations requiring explicit consent from the vehicle's owner. As such, for the specific purpose of studying intelligent transportation systems, Calabrese et al. (2010) used mobile positioning data even though it is, broadly speaking, not as accurate as GPS data. Nonetheless, the wide adoption of mobile devices and stations both in urban and rural areas make for a good trade-off with accuracy, especially when considering that the performance of the system the authors built after testing against GPS data showed small errors, particularly when temporal data granularity was high. This case study showcases that even though GPS may be used as an alternative in some research papers, the choice between GPS data and mobile positioning data depends on the objectives of the study itself.

2.2. USING PASSIVE MOBILE POSITIONING DATA FOR TOURISM MANAGEMENT

For tourism purposes, the understanding of the individuals' mobility patterns is also very important. It can serve to better market touristic destinations and attractions, help in administrating and designing the touristic attractions and plan the on-site movement (Zheng et al., 2017), grasp how fairs, concerts, sports events, amongst others can impact touristic flows, understanding the relationship between certain Points of Interest and the arriving and departing points of the tourists' journeys (Ahas et al., 2007a), and help city planning (Ahas et al., 2010).

Once more, mobility may be studied at a macro or micro level and may be studied with different data types. For instance, Zheng et al. (2017) were able to study tourist displacement at the level of the touristic attraction by using GPS data and a data mining heuristic algorithm to predict a tourist's next visited attraction based on their own trajectory and the trajectories of previous tourists. In addition, contrary to the complications found by Calabrese et al. (2010) in using GPS data to study intelligent transportation systems, Shoval (2008) managed to use this data type to analyze pedestrian patterns in the historical city of Akko, Israel. Moreover, Scherrer et al. (2018), used data of 71,207 unique users provided by a global navigation assistance app to identify locals and travelers, and, in the city of Ghent, Belgium, Versichele et al. (2014) managed to study visitation patterns in touristic attractions and hotels by employing data mining techniques of association rules in Bluetooth data. Nonetheless, passive mobile positioning data has extensively been used to study tourism.

Ahas et al. (2007a) used data of 21 million roaming call events to inspect tourism flows in the city of Tartu in Estonia. They were able to analyze call activities of tourists from different nationalities in the city as well as explore some simple statistics such as the most popular places for tourists to visit (in number of visits), proving that mobile positioning could in fact be used for tourism management studies. In addition, Ahas et al. (2007b) analyzed the seasonality of Estonian tourism and the characteristics of the tourist's studies by using 9.2 million data records of Network Events (any communication between a user device and a network infrastructure) belonging to foreign mobile phones. The authors found this to allow to predict peaks of tourism influx, which provides the insights needed for tourism management to develop strategies that will allow the redistribution of affluence flows. Finally, in 2008, Ahas et al. collected 12.8 million roaming network event data records of foreign mobiles and analyzed them in order to study the applicability of these datasets in tourist mobility. This study confirmed the findings of the previous ones, especially that mobile positioning data is less similar to accommodation statistics in areas that have a higher number of transit tourists and fewer infrastructures dedicated to them. Moreover, this data is useful to: answer questions related with the nationalities of the tourists in each area, identify flaws in marketing efforts, and to study tourism in isolated regions allowing for the elaboration of development plans for rural areas, which would be very costly to do using traditional methods. These three papers have contributed greatly to the knowledge on the nature of the insights one can get in using this type of spatial-temporal data to study tourism and the impact it can have on tourism management.

Many other researchers followed suit, such as Olteanu et al. (2011) that attempted to understand how visitors of a touristic region use the area and how they visit the touristic attractions by using GSM (Global System for Mobile Communication) network probes from roughly 1.5 million users in Paris. Furthermore, these studies have the capacity to improve decision making based on evidence in tourism management since mobile positioning data can be used to study domestic, inbound and outbound tourism in a region, complementing national statistics (Ahas et al., 2015).

For instance, Zhao et al. (2018), attempted to understand the patterns of tourism flow by taking the size of the travel group into account through the use of Network Event data of 12 million tourists in Xi'an as well as data from Gaode Map, a digital map and navigation service that allowed for the identification of touristic points of interest. This study can lead to possible improvements in resource management directed to touristic activities, such as: travelling packages segmented by travel party size, offering of transportation between attractions, planning and management of the touristic attractions and in the development of personalized recommendations of next attractions.

Additionally, Qin et al. (2019) collected a year's worth of CDR data of tourists in Beijing to analyze their movements at touristic sights; Mamei & Colonna (2018) used CDR data from 600,000 tourists in July 2013 and 600,000 tourists in March 2014 to classify them in clusters of tourists, residents, commuters, people in transit or excursionists; Hu et al. (2018) utilized a CDR dataset to identify the tourists' transportation mode; and Leng et al. (2016) used CDR data of tourists in Andorra from 2014 to 2016 to analyze and evaluate marketing strategies in tourism at a national and local scale.

Finally, a study conducted by Oliver & Enrique (2014) in Barcelona and Madrid that collected roaming data from 680.000 foreign mobile phones and electronic payments made by 169.000 credit cards in BBVA terminals explains the impact Big Data has on city planning since its findings provided many practical recommendations based on several different dimensions related to tourism. In fact, when combined, roaming and electronic payment data can provide insights on: the profile of the most attractive tourist to the region in real time; what type of hotels the city planning team should allow to be built, as well as where these should be built and which nationalities should be targeted in each sub-region); what touristic fairs the tourism management teams should focus on; what advertising methods should be used; what type of services should be provided; and what type of public transportation system should be implemented. All these studies prove the applicability of passive mobile positioning data to tourism management.

Nevertheless, it is important to address the fact that, in an extensive literature review conducted by Li et al. (2018), GPS was the Big Data type that was found to be the most commonly used in tourism research (around 21% of the studies analyzed used this datatype), whereas roaming data was not used as much, which they indicate might be due to privacy concerns from the networks and that GPS data is typically both more continuous than roaming and has a higher precision (Ahas et al., 2008). Nonetheless, Li et al. (2018) found GPS and positioning data to be examples of device data that could be used to study tourism movement. However, user data (e.g. online textual data) and operations data (e.g. web search data) were also found to be used in these studies. Notwithstanding, mobile positioning data presents its advantages and disadvantages that justify its use over other data types in a given study. These will be explored into detail in the following section.

2.3. BENEFITS AND COSTS OF USING MOBILE POSITIONING DATA IN MOBILITY STUDIES

As it has been proven in the previous sections, mobile positioning data can be very useful for the study of tourism management. However, as it is the case with any other data type, mobile positioning records have advantages and disadvantages. According to Ahas et al. (2015), the main strengths of using this type of data are that: (i) the number of trips and of nights spent in a location are consistent with traditional statistics, (ii) it provides coverage to tourists that stay at unpaid/ non-registered accommodations, (iii) it provides the ability to gather more detailed insights in regards to the regions and the country of origin of the tourists, (iv) it offers the ability to identify specific phenomena such as repeated visits, frequency of visits, secondary destinations, among others, (v) deliver practically real time statistics, (vi) enables some automation in the creation of tourism statistics, and (vii) creates the possibility of having cross-border international statistics on tourism. On the other hand, they have also identified some challenges with this data, that are: (i) that data provided by the mobile network operators are difficult to access and the continuity of the access is unclear, (ii) that information collected is on a location basis and it lacks qualitative understandings, (iii) the introduction of some bias due to the data's nature and the user's behaviors, (iv) that it is hard to measure the resulting statistics' quality. For the particular case of tourism, Li et al. (2018) identified the ability to cover larger areas which might also correspond to less visited areas and the ability to extract the tourists' nationality as advantages of this data type, and identified the difficulty in accessing to this data due to privacy issues, and a low accuracy level, especially when GSM communication is limited in the tourism destination.

Nevertheless, although CDR's have been specifically shown to be a widely popular data source amongst researchers in general mobility studies, there are some additional issues associated with their use to study individual positioning. For instance, Fiadino et al. (2012) showed that, since a significant percentage of the individuals generated records associated to a very limited number of different antennas, CDR's produce a biased picture of mobility, generally underestimating how much individuals move. In addition, Jiang et al. (2013), concluded that in urban areas the precision of estimates of the individual's location is, on average, about 300 meters, however, depending on the density of the antennas, this precision can go up to several kilometers (Chen et al., 2016; Järv et al., 2014). Moreover, Wang & Chen (2018) found that without a specific processing of the CDR data, the frequency of human mobility patterns can be overestimated. One such situation occurs, for example, because of ping-pong effect occurs when there is a recurrent handover of the device between 2 or more antennas due to the existence of repeated changes in the strength of the signal (Fiadino et al., 2012).

Chen et al., (2018) found that trajectory data is unevenly distributed in space and time, that spatial data is constrained by the antenna's coverage, that the temporal data is heterogeneous for a given user and, that, due to data sparsity, a big percentage of individual's trajectories are not usable because of the insufficient sampling frequency of the movement of the user. In fact, according to Hoteit et al. (2016) CDR's provide limited accuracy both in the spatial and in the temporal features precisely due to the nature of the data collected. For example, Gonzalez et al. (2008), Song, et al. (2010) and Hoteit et al. (2014) kept 1.67%, 0.45% and 0.07% of their CDR dataset, respectively, after removing individuals with incomplete trajectories that did not meet their required level of mobility information. Chen et al. (2018) tried to mitigate this issue by using different CDR completion methods to make functional trajectories of the incomplete datasets. This approach was possible thanks to the redundancy, which is very important to ensure that trajectories are predictable and, thus, can be estimated with some accuracy; inter-call stability; and nighttime stability that characterize human mobility. However, these assumptions may not hold as strongly in touristic movement as, for instance, the highly predictable home-work commute does not occur in this particular case.

All in all, it is important to notice that the main disadvantage analyzed in the literature on the CDR usage is connected to its sparse nature. According to Chen et al. (2018), obtaining full trajectories is very hard with these datasets. In fact, these authors found that although the completeness of CDR datasets is not dependent on its duration, it depends on its temporal resolution. Their findings showed that, for a 30-day dataset, splitting the data into groups of 15 minutes lead to a mean level of completeness smaller than that of groups of 120 minutes. This demonstrated the tradeoff between obtaining higher completeness levels in the data versus making more detailed analysis with higher resolution rates.

As such, as an alternative to the use of CDR's, researchers may use signaling traffic data collected from cellular devices, which has already been used in mobility studies (Zhao et al., 2018). Signaling data are periodic beacons (for example, every five minutes) between a base station and a user's device, hence providing a much more complete dataset. These new datasets can, in turn, produce more accurate results than CDR's since not only each device's location is being sampled at a higher frequency, but also the data collection happens from all the devices, solving the problem of discarding the majority of the data collected (Fiadino et al., 2012). Nonetheless, high frequency signaling data poses challenges in the storage of greater amounts of data. Therefore, there is a need to identify the tradeoff between CDR and signaling data usage and understand under which scenarios it is better to collected data with a higher granularity and under which scenarios applying some completion techniques to CDR's or less granular signaling data is sufficient to get accurate results.

2.4. CALL DETAIL RECORDS COMPLETION TECHNIQUES

The completion of sparse mobile positioning datasets is of most importance in order to conduct further analyses that require a continuous flow of spatial and temporal records. For instance, the high risk of a place to become overpopulated in peak hours or during a public event requires monitoring in order to ensure safety. Instead of using video analysis to determine this risk, authorities might want to estimate it using Wi-Fi probe data or cellular phone probe data (Wang et al., 2019), or may use another type of mobile positioning data. One way to achieve the needed level of completion for this type of analysis is to use interpolation-based methods to complete the missing data, in which each point in an individual's trajectory is considered to be independent from all others and weights are computed between the missing points and their surrounding points, however, the longer the time span between two points the more ineffective reconstruction becomes (Li et al., 2019).

In order to complete sparse CDR datasets, Hoteit et al. (2014), studied the difference between the ground truth and estimations of localization attained from linear, cubic and last seen interpolations depending on the individual's span of movement calculated by the radius of gyration. These authors found that linear interpolation outperformed the remaining methods for users with a very small span of movement (radius of gyration < 3 km), while cubic interpolation outperformed others for users with a high span of movement (radius of gyration > 32 km).

Moreover, Hoteit et al. (2016) used a GPS dataset to mimic the sparseness of CDR's and complete these new sets. In order to do the completion, the authors used: a last seen approach (which they called "static"), and linear and cubic interpolation approaches ("continuous") given the user's radius of gyration found in Hoteit et al. (2014). Other more complex approaches included: one that assumes that an individual stays at the location of their last call activity 30 minutes prior and posterior to it and that if two activities occur within less than an hour apart the nearest neighbor would be assumed, another approach that completes unknown night locations with the users' identified home location, and, finally, an approach in which the missing locations that occur during the night would have only be completed with the home location if the last seen location of that user was in a radius smaller or equal to 1 km from their home location. The authors found that the first and last of the more complex approaches worked the best, as roughly 95% of the estimations were within 100 meters from the ground truth and in which only 1% of the estimates were at more than 3 kilometers from the ground truth.

Chen et al. (2018) have also implemented a last seen approach that filled in missing locations with the previous nearest neighbor value, and an interpolation approach that infers missing locations through a linear or cubic interpolation given the individual's known locations and radius of gyration. Additionally, they developed a tensor factorization approach, capitalizing on the redundancy assumption, and compared the results against previously used methods. The tensor factorization model proposed by the authors follows three steps: the nighttime enhancement (that fills temporal gaps in the dataset occurring during the night with the individual's identified home location), the temporal improved tensor completion (that uses trajectory's redundancy to infer absent locations with a tensor factorization model), and the cell tower estimation (which takes the inferences from the previous step and allocates these estimations to a cell tower). The authors found promising results for their proposed approach, in which the distance error between the results of the tensor factorization model and the ground truth were smaller than the ones stemming from the remaining techniques. As expected, they found that as the data completeness increased, not only would this distance error decrease, but also the differences between the approaches would become less significant. In this study, unless completeness was below 5%, the authors proposed approach would outperform the others in the minimization of the cell estimation error. However, this only occurred since the more complete the dataset is, the higher the redundancy of human mobility, which improves the results of the inferences. This assumption may not hold for tourism if it is verified that the movement redundancy is significantly lower for tourists than for the general population.

2.5. MOBILE POSITIONING DATASETS SAMPLING TECHNIQUES

In order to study the completion of CDR datasets, researchers have had to mimic CDR sparseness from complete datasets as an initial step through the use of different sampling techniques (Chen et al., 2018; Hoteit et al., 2016; Hoteit et al., 2014). Chen et al. (2018), have solved this issue by determining the level of completeness of the trajectories and keeping only that percentage of the ground truth signaling data dataset (for instance, if the completeness was 0.3, only 30% of the full dataset would be retained). The kept trajectories should simulate the temporal sampling properties of CDR-based trajectories. To ensure that, the authors build an empirical distribution of the probability of the occurrence of an event at each given time slot; set the completeness value; and keep the fraction of the dataset corresponding to the completeness value by randomly selecting time slots that follow the empirical distribution found previously.

Furthermore, Hoteit et al. (2014) implemented a similar approach by computing the empirical probability density function of time between events, taking a random position (coordinates and timestamp) from the ground truth as the starting point (as well as its spatial-temporal attributes) and extracting the next positions using the probability density function.

Notwithstanding, the sampling of mobile positioning data has not been done in the literature for the sole purpose of studying CDR completion methods. In fact, these techniques can be also employed in order to decrease computational load as it is the case in the study conducted by Calabrese et al. (2010) that sampled their dataset with a fixed ratio for a given time resolution. Contrastingly, Zheng et al. (2017) have also sampled a GPS dataset for computational purposes but through a grid-based spatial clustering model in which the specified parameter is a minimum cell density threshold level. This approach may not work as well for CDR or other type of mobile positioning data depending on the level of sparsity in the dataset that may bias the results of the clustering model as the cell density will be highly influenced by the number of cell towers in a specific region.

2.6. STUDY RELEVANCE AND IMPORTANCE

This study aims to extend the current knowledge of the subject by identifying the level of granularity needed when collecting datasets on tourism mobility, an unaddressed research gap. Additionally, this work will add value to the findings in the mobility dataset completion literature by implementing some of the approaches used in the literature (Chen et al., 2018; Hoteit et al., 2016; Hoteit et al., 2014) in the specific context of tourism and identifying the best models to complete these datasets. This application is especially relevant due to the fact that the redundancy characteristic of human mobility that makes it possible to predict human trajectories (Song et al., 2010) may not hold as strongly in tourists due to the loss of routines such as the home-work commute. This lack of redundancy in tourism flows will be explored in further detail in section 4. Moreover, due to their nature, tourists have a shorter lifetime than residents in a specific location and may switch between mobile network operators according to their roaming plans, increasing further the sparseness of the datasets.

It is possible to predict tourism mobility (Olteanu et al., 2011), especially through the use of mobile datasets (Ahas et al., 2009). However, the sparse nature of the CDR data that generally becomes an issue as a grand part of the observations have to be kept out of the analysis (Gonzalez, et al., 2008; Song et al., 2010; Hoteit et al., 2014) is also a concern for this particular application. Moreover, in the study conducted by Olteanu et al. (2011), due to the fact that tourists can connect to different network operators, their arrival and departure points became unclear, which limited the analysis even further to the individuals coming in and out of airports. The authors recognize this limitation and propose that further research should be done to mitigate this sparsity issue in order to generalize the results to the whole tourist population, which strengths the need for this paper.

On the other hand, collecting highly granular data to overcome the CDR's constraints is very costly and can be unfeasible. For instance, the data used in this study consists on IP probe data of 277,093 devices over the course of a week and occupies 3.13 Gigabytes. Assuming that the memory requirements scale linearly, collecting data on 3 million tourists per week over the course of 5 years corresponds to data requirements of approximately 8.83 Terabytes. This translates to thousands of dollars spent in data storage per year to collect information on 3 million users, which corresponds to a small percentage of the tourist population in many countries.

The resolution of this issue will have theoretical applications to the study of tourism mobility. The final results will add to the state of the art by providing a better understanding of the level of granularity of tourist data researchers, authorities, policy makers and marketers have to collect to study their movement as well as the methods that work best at completing the incomplete signaling data or CDR datasets.

3. DATA AND METHODOLOGY

3.1. DATA

The dataset consists of IP probe data of 277,093 devices with foreign SIM cards, connected to a European Provider's network in Tuscany at least once in the week of the 1st of May 2017 to the 7th of May 2017. The dataset contains the following information: i) id that identifies the tourist, ii) the mobile country code that specifies the tourist's country of origin, iii) a timestamp (dd-mm-yyyy hh:mm:ss) and iv) the location of the cell towers in Italy (latitude and longitude). The locations of the cell towers are identifiable via a location id that corresponds to a set of coordinates rounded to 4 decimal points. As the proposed study is computationally intensive, a random subsample of 1,000 tourists was used instead of the full dataset. More information on the differences between the original dataset and this subsample will be provided later on. For the remainder of section 3, let the full dataset be the set of 277,093 tourists and the subsample dataset be the 1,000 tourists subset.

The dataset has a high level of time granularity since the tourists' data was collected every minute in which they perform some type of telecommunication's transaction or every hour if they do not. For instance, let us suppose that tourist X is in location A and makes a call at 13:00, sends a text message at 13:10 and moves to location B at 13:31. Then, this tourist would only have records for the timestamp of 13:00 and 13:10 corresponding to location A and for the timestamp of 13:31 corresponding to location B dues to the Location Area Update (LAU), when he/she should have records for every minute in between 13:00 and 13:31. If, on the other hand, there is a tourist Y that is in location A from 13:00 until 15:00, inclusively, then the tourist only has records for the timestamps of 13:00, 14:00 and 15:00, each at location A. This represented a technical challenge for the study. In fact, if we allow the total ***number of records*** to be the sum of each tourist's number of minutes since their first telecommunication's transaction until their final transaction, then the subsampled dataset should contain a total of 4,488,125 records. Instead, due to the data collection method, it has 295,339 rows, which corresponds to only 6.6% of the total amount of records. Of course, this way is more efficient, as the redundant data is not stored. However, for the purpose of this study, we needed to analyze each tourist on per minute basis. For example, when we wanted to take one sample of each tourist per hour, we needed to ensure that we selected with higher probability a record with the location where a client spent more time. Hence we distinguished ***records*** and ***rows***, and from this point onwards, the records correspond to the number of theoretical records that the dataset should have and rows refer to the number of rows that it actually has, respectively. This issue will be referred to, from this point onwards, as the data collection issue, and we will discuss it in more detail later on.

Another issue we encountered was that the tourist data suffered from some extra data sparsity due to the fact that some foreign visitors have gaps larger than 1 hour between consecutive rows. This may be a result of roaming agreements and not a result of them departing and re-entering Italy multiple times. No action was taken towards this issue, as its prevention is out of the scope of this study. Going forward, this issue will be referred to as the roaming issue.

Some data preparation was carried out during the exploration of the dataset. Firstly, it was defined that all tourists that appear in the full dataset for fewer than a total of 12 hours should be removed, so that the sampling and completion was not hindered by having tourists with too few records.

Secondly, the tourists that appeared in the full dataset only on the first or the last day were removed to minimize the occurrence of incomplete trajectories as much as possible. This action was taken as it was assumed that, for example, it would be more likely that an individual that only appears in the dataset during the last day of the week has just arrived in Italy and, therefore, their full trajectory would never been known.

Thirdly, about 95% of the location ids correspond to non-unique pairs of coordinates, i.e., the same longitude and latitude are recorded for two different cell towers. No action was taken in this case since it was assumed that these towers correspond to different neighboring cell towers in denser areas that only appear as the same in the dataset due to the rounding of latitudes and longitudes to 4 decimal points.

Finally, it was observed that about 3.5% of the total rows in the full dataset were from tourists with more than one location id in a single timestamp. This could be a result of a ping-pong effect, in which there would be a handover between two towers, or could even be a result of an error. The issue was solved by dropping all duplicates after the first record of a timestamp for a given tourist. After these steps of data preparation, 143,625 of the initial 277,093 devices were kept.

3.2. METRICS USED TO DESCRIBE INDIVIDUAL MOBILITY

In order to characterize tourists present in the dataset, some metrics were considered. These metrics were: (i) the total number or cell towers the tourists connected to, (ii) the tourist lifetime as the number of days they appear in the dataset, (iii) the number of tourists in the dataset per day, (iv) the tourists' nationality, (v) the number of tourists in a given hour of the day, (vi) the radius of gyration in km, (vii) the mobility entropy, and (viii) the total travel distance in km. The computation of the radius of gyration, mobility entropy, and total distance travelled are further explained below.

The radius of gyration is a metric of relevance in mobility studies that measures one's span of movement (Schneider et al., 2013; Hoteit et al., 2016). For each tourist t in T, where T represents the set of all tourists, the radius of gyration is computed as follows (Hoteit et al., 2016; Zhao et al., 2019; Zhao et al., 2016):

$$r_t = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(l_t^i - \bar{l}_t)}\ ,$$

where l_t^i represents the ith location visited by tourist t and $\bar{l}_t$ represents the center of mass of all the locations visited by tourist t. The center of mass of the tourist's recorded displacement is the coordinate pair of the weighted mean of the latitude and longitude of that tourist, weighted by a value in a scale of 0 to 1 that represents the difference in time in two consecutive records for that same tourist (further explanation on these weights is provided in the following section). The distance between two pairs of coordinates $(l_t^i - \bar{l}_t)$ was computed with the haversine formula that provides the great-circle distance between two locations (Chopde & Nichat, 2013):

$$d = 2r\sin^{-1}\left(\sqrt{\sin^2\left(\frac{lat_2 - lat_1}{2}\right) + \cos(lat_1) * \cos(lat_2) * \sin^{-1}\left(\frac{lon_2 - lon_1}{2}\right)}\right),$$

where r is the radius of the Earth. For the purpose of this study, let the radius of the Earth be of 6,367 km, which is the average of the radius at the equator, 6,378 km, and the radius at the pole, 6,357 km.

Moreover, the movement entropy is considered another relevant metric of human mobility as it measures the level of randomness in the individual's daily movements (Zhao et al., 2019; Zhao et al., 2016; Qin et al., 2012; Eagle & Pentland, 2006; Chen et al., 2018):

$$E = -\sum_{i=1}^{N} p_i * \log_2 p_i,$$

where N represents the total number of distinct locations for a given tourist and p_i represents the proportion of each location in the total number of distinct locations. A higher entropy level corresponds to a more heterogeneous movement pattern. As such, it is expected that the higher the randomness in the movements, the higher the estimation error will become.

Additionally, the total travel distance has been used in the literature to measure human mobility (Zhao et al., 2016). For the purpose of this study, let the total travel distance be the sum of the haversine distances between the locations of any two consecutive registered data points of a tourist.

3.3. Sampling techniques

3.3.1. Uniform sampling

In order to study the tradeoff between the volume of collected data and accuracy of estimated tourist trajectories, different sampling techniques were applied. The first sampling technique was a uniform sampling approach with different sampling rates. Chen et al. (2018) investigated the completeness of the datasets by using temporal resolutions of 15 minutes, 30 minutes, 1 hour and 2 hours. This study takes these rates as parameters for the uniform sampling process, as well as the additional sampling rates of 45 minutes, 1 hour and 30 minutes, 3 hours and 4 hours. This approach was conducted by randomly selecting 1 record per tourist at the given sampling rate (for example, if the sampling rate is 2 hours each tourist had 1 random timestamp and its associated antenna selected every 2 hours).

Due to the data collection issue, discussed in section 3.1., weights between 0 and 1 were given to each row in order to sample rows with locations where a tourist was for a longer time at a higher frequency. This weight is 1 if the difference between a tourist's data point and its subsequent is higher or equal to 60 minutes and between 0 and 1 if it is between 0 and 60 minutes, exclusively. This method was applied to minimize the over sampling of rows that correspond to locations where a tourist spent less time. For instance, if a tourist is recorded at 13:00 in location A, at 13:05 in location B and at 13:45 in location C, and the sampling method does not take into account the proposed weight, then the locations associated with timestamps 13:00 and 13:05 have the same probability of being sampled. However, less time passes from 13:00 to 13:05 (5 minutes) than from 13:05 to 13:45 (40 minutes). As such, location B should be sampled with a higher frequency than location A. For that reason, the weights of 0.08 and 0.66 are given to locations A and B, respectively, by following the rule where 60 minutes corresponds to a weight of 1. In addition to these weights, the sampling was conducted with replacement to allow for rows that correspond to multiple records to be sampled more than once. Furthermore, the use of the weights and the sampling with replacement is consistent in all approaches proposed in this study and not just the uniform sampling.

However, the rows that were sampled for a given time interval were only the rows whose beginning belongs to that same time interval. This technical approach raised a new issue since it does not consider that the location of a tourist is stored in the previous rows. For example, say there is a timestamp for 13:00 for location A, 13:10 for location B and 13:25 for location C, during the time between 13-13:30, and we wish to sample at a 15 minute rate. In this case, the timestamps of 13:00 and 13:10 are categorized as belonging to the time interval [13:00; 13:15[and the timestamp of 13:25 to the interval [13:15; 13:30[. This means that two locations A and B are available to be sampled in the first interval while only C is available in the second, hence, ignoring the fact that the location B recorded at 13:10 lasts until 13:25, and spills over to the second time interval. This limitation of the methodology will be addressed, for this point forwards, as the spillover issue.

After the sampling is completed, the Latitude and Longitude of all the rows on the subsample dataset that were not selected in the previous step were replaced with NaN's and later estimated.

3.3.2. Uniform sampling with different rates during the day and during the nighttime

Other than redundancy, an assumption made by Hoteit et al. (2016) in order to develop complex completion methods was nighttime stability, in which it was expected that individuals stayed more stable overnight.

If nighttime stability is high, it could be the case that fewer records from each tourist could be collected during the night, which decreases storage requirements without decreasing the estimation accuracy. Therefore, a second sampling approach was implemented. This approach divided the subsampled dataset into two subsets: the daytime sub-dataset (from 9 am to 6 pm, inclusively) and the nighttime sub-dataset (from 6 pm to 9 am, exclusively). The same methodology for the uniform sampling was applied to each of the sub-sets. Then, due to nighttime stability, pairs of sample rates were created, in which the daytime sample rate would always be higher than the nighttime sample rate.

In order to keep the results comparable to the uniform sampling approach the same sampling rates were assumed (15 minutes, 30 minutes, 45 minutes, 1 hour, 1 hour and 30 minutes, 2 hours, 3 hours, 4 hours). A drawback from this method is that the number of daytime or nighttime hours is not divisible by all the sampling rates. This means that for some sampling rates, there were a few classes with fewer records than the others. For example, in the case where we want to take one sample for each 4 hours, we run into an issue as the daytime (from 9 am to 6 pm inclusively) is not divisible by 4, meaning that 2 classes of 4 hours are able to be created, [9:00; 12:00[and [12:00; 16:00[, while the remaining class has only 2 hours, [17:00; 18:00]. This limitation will be referred to, from here onwards, as the day and nighttime sampling issue.

3.3.2.1. Sampling to mimic CDR's

Chen et al. (2018) down sampled their ground-truth data to mimic CDR's by building the empirical distribution of a CDR dataset, setting a desired completeness level to a given rate and retaining the percentage of the ground-truth dataset corresponding to the completeness rate that was set for each trajectory. A different subsampling approach to this one would be to apply a theoretical distribution of CDR arrival to the dataset in order to mimic CDR's. Nonetheless, two drawbacks from using this latter method were identified. Firstly, the CDR arrival is usually studied on a tower level (CDR's registered in a given tower instead of CDR's registered by a given user). Secondly, the existing literature is not consistent regarding the theoretical distribution of the CDR's arrival nor extensive in the identification of the parameters of the corresponding distributions, although it is consensual that the distribution should be skewed to the right. For instance, Ary & Imre (2010) modelled CRD arrival as a combination of the call start distribution (normal) and the call length distribution (lognormal); Iversen (n.d.) modelled call inter-arrival times as a combination of three lognormal distributions; Barcelo & Sánchez (1999) modelled inter-arrival times with an Erlang distribution; and Dang et al. (2004) modelled inter-arrival times with an exponential distribution.

As such, the methodology used for this approach will resemble the method used by Chen et al. (2018). In order to find the empirical distribution of the CDR's, an open source CDR dataset of the city of Milan and Province of Trentino (Barlacchi et al., 2015) was obtained. From this point onwards, this dataset will be called CDR dataset. The CDR dataset that was retrieved was composed by the date-time of the start of the transaction, the country code of the mobile device, and information on the inbound and outbound SMS and calls, and internet traffic. Each record in the CDR dataset represents one CDR. In total, the CDR dataset contains 15,089,165 records and spans for the week of the 1st of November 2013 to the 7th of November 2013. No information on the different users is given.

In order to get the empirical distribution of CDR's over the course of the day, the CDR's were split into bins, each representing a given hour, and the % of CDR's for each hour was computed and is shown in the table below (table 1).

For simplicity purposes, completeness rates were set and assumed for the complete subsampled dataset and not in a per tourist basis, a limitation that will be referred to as the CDR sampling issue. Nonetheless, the completeness rates chosen were the completeness rates found in the uniform sampling approach, in order to compare different approaches. The completeness rate is set as the number of records sampled divided by the number of records in total. As such, 8 different CDR mimicking approaches were applied, corresponding to the same completeness rates of the uniform sampling approach with sample rates of 15 minutes, 30 minutes, 45 minutes, 1 hour, 1 hour and 30 minutes, 2 hours, 3 hours and 4 hours, by computing the number of samples that need to be kept in each hour of the day. This computation is performed as the number of records to be sampled multiplied by the % of CDR's in each hour given the CDR dataset, for each hour of the day. The number of records sampled per hour for each given completion method can be seen in *Appendix A* and are, naturally, consistent with the results of table 1.

Hour	% of CDR’s
00 h	2.4%
01 h	1.9%
02 h	1.6%
03 h	1.6%
04 h	1.7%
05 h	1.9%
06 h	2.6%
07 h	3.6%
08 h	4.7%
09 h	5.5%
10 h	5.7%
11 h	5.7%
12 h	5.9%
13 h	5.5%
14 h	5.6%
15 h	5.6%
16 h	5.8%
17 h	5.8%
18 h	5.7%
19 h	5.4%
20 h	4.8%
21 h	4.2%
22 h	3.6%
23 h	3.1%

Table 1 - Empirical distribution of CDR's during the day.

3.3.3. Completion techniques

As the aim of this study is to reflect on the tradeoff between collecting and storing more data versus estimating complete tourist trajectories from sparse data, three simple completion methods were tested in the datasets originated by the sampling process described in the previous sub-section. These three methods were the last seen, the linear and the cubic approaches, used by Hoteit et al. (2014), Hoteit et al. (2016), and Chen et al. (2018). For simplicity purposes, whenever extrapolation is needed because the first and/or last record(s) were not among the sampled records, then it will be assumed that the tourist remained stationary at the first or last known locations, respectively, even though it is recognized that for many applications the first and last known positions would be assumed as the starting and ending trajectory points. Moreover, for simplicity purposes as well, the latitudes and longitudes were interpolated separately.

The first approach to be applied was the last seen approach, the least expensive method (as the result is already an existing tower and no further assignment between the estimate and its closest tower is needed). It assumes that whenever a tourist's data is missing, that tourist stayed stationary at the previously observed location. Consequently, this method was implemented by filling every non-sampled location with the last known position for that tourist.

The linear interpolation, on the other hand, follows the equation bellow:

$$y = y_0 + (x - x_0) * \frac{y_1 - y_0}{x_1 - x_0},$$

deriving from the fact that the slope of two known points should be the same as the slope of the unknown point and one of the others in a straight line, where x represents the time the record was collected, y=f(x) for either the latitude or the longitude of the tourist's location and (x_i, y_i), i={0,1} ,are known data points. This approach always needs at least two known data points in order to be successfully implemented, however, since the tourists that stayed in the dataset for fewer than 12 hours were removed from the dataset during data preparation, this constraint does not become an issue.

Lastly, the cubic method is a cubic interpolation that is computed by applying a third-degree spline, which interpolates the function by applying polynomial of third degree that uses the function's values and its derivatives at both ends of each of the function's subinterval. This approach requires at least 4 known points in order to be able to interpolate 1 unknown point of a tourist's trajectory. However, the lower the level of granularity is, i.e. lower the sampling rate, the fewer the known observations will be. The tourists that did not possess the 4 needed known points after sampling were excluded from this analysis. *Appendix B* displays the number of tourists for whom it was impossible to apply the cubic method by sampling rate and sampling method (sampling rates/methods omitted from the table correspond to the ones in which this situation is not applicable).

3.3.4. Evaluation methods

After the interpolation was completed and since the ground truth data is taken to be an antenna position, the estimated positions were mapped to the cell towers that were nearest to the estimated position. In the case of the last seen method, the step of mapping to the nearest cell tower was not necessary, since the estimates are the actual tower positions, as it has been stated previously. For the case of linear and cubic methods, the haversine distance between each estimated location and all different antenna locations was calculated and the estimated location replaced by the nearest antenna's coordinates. Subsequently, the haversine distance between the ground truth and the new estimated positions, as well as a Mean Distance Error for each tourist was computed. In order to ensure some statistical significance in the sampling results the sampling and completion processes were performed 20 times, and the final Mean Distance Error per tourist becomes the average of these 20 processes. Due to high computational costs, it was not possible to perform the processes more times.

The Mean Distance Error for the uniform sampling approach was computed as follows:

i) The average error of the estimated positions for a given tourist was calculated as the sum of the distances between the estimated towers' and the actual towers' positions over the number of points interpolated, let's call this the ***average error***. It is important to consider that the weights used during the sampling process were re-applied in the calculation on the average error so that the error of positions in which an individual spends most time in become more important than the errors of positions in which the tourist spends little time;

ii) As the algorithm was run 20 times, the final ***Mean Distance Error*** per tourist becomes the average of the 20 average errors for each tourist.

The same methodology was applied to obtain the Mean Distance Errors for the CDR sampling approach.

Lastly, for the uniform approach with different rates during the day and during the night, the results of the daytime and nighttime subsets for each of the pairs were merged, and the Mean Distance Error for each of these combinations was computed. This process can be summarized as follows:

i) The ***average daytime error for a given tourists*** is computed following the uniform sampling methodology. Recall that the average error for the uniform sampling methodology is the sum of the distances between the estimated towers' and the actual towers' positions over the number of points interpolated;

ii) The ***Mean Distance Error during the day per tourist*** is calculated as the average of each tourist's average daytime error over the 20 runs of the algorithm;

iii) The ***number of interpolated points during the day per tourist*** is calculated. The number of interpolated points are the same in any run of the algorithm since the algorithm takes 1 record per given time period. For example, if we set the sampling rate to 15 minutes, it might take, as a sample, the data point 13:02 from the period of [13:00; 13:15[and 13:28 from the period of [13:15;13;30[, even though these are not truly 15 minutes apart. This methodology was applied due to the data collection and spillover issues;

iv) Steps i), ii), and iii) are repeated for the nighttime;

Steps i), ii), iii) and iv) retrieve 2 tables (1 for the daytime and 1 for the nighttime) with the tourists' id's and their respective Mean Distance Errors during the day and during the night; and 2 tables (1 for the daytime and 1 for the nighttime) with the tourists' ids and their respective number of interpolated points during the day and during the night.

v) The Mean Distance Error during the day for a given tourist is multiplied by that tourist's number of interpolated points during the day. This step is applied to all tourists and repeated for the nighttime to provide the average sum of the distances between the estimated towers' and the actual towers' positions over the 20 runs for both periods. Let's call this the ***average sum of distances***;

vi) The final ***Mean Distance Error*** of this approach, per tourist, is taken as the sum of the ***average sum of distances during the day*** and the ***average sum of distances during the night divided by the total number of points interpolated.***

These extra steps have been taken due to the direct application of the uniform sampling algorithm to the uniform sampling with different rates during the day and during the nighttime method.

The Mean Distance Error per tourist for each sampling and completion rate were analyzed against one another and were plotted against the tourists' lifetime in days and in hours, radius of gyration, and total travel distance. It is to be expected that the higher the radius of gyration and the total travel distance, the higher the Mean Distance Error should be since tourists would have travelled more and, hence, be recorded in a higher number of unique cell towers.

Lastly, in order to understand what the implications of the changing sampling rates and completion methods would be in practical applications, a couple of trajectory and clustering analyses were performed. Firstly, the trajectories of two individuals, one with high radius of gyration and entropy and another with low radius of gyration and entropy, were shown. These trajectories were plotted using the python library scikit-mobility library by Pappalardo et al. (2019). The algorithm used to plot each tourist's trajectory was also set to detect stops of each user if he/she stayed in the same antenna for longer than 1 hour. The algorithm further clustered all the stops by aggregating the geographically closest most visited stops via the application of a DBSCAN. These clusters were then mapped to the trajectories and shown in a user diary of stops, in which a different color corresponds to a different cluster of stops.

Secondly, the k-means algorithm was applied in order to group similar types of tourists together. The features that were analyzed to characterize the tourists are presented in *Appendix C*. These features were then standardized due to the k-means sensitivity to scale. Subsequently, a threshold of 0.7 correlation was set in order to understand which variables to discard. The Pearson correlation matrix can be seen in *Appendix C.* From this correlation matrix, it was found that the Standardized Lifetime in Days and the Standardized Lifetime in Hours had a correlation of 0.83. Thus, the Standardized Lifetime in Days was discarded, as the Standardized Lifetime in Hours provides more information. Additionally, it was also found that the Standardized Average Latitude and the Standardized Average Longitude had a correlation of -0.71. As such, the Standardized Average Longitude was discarded.

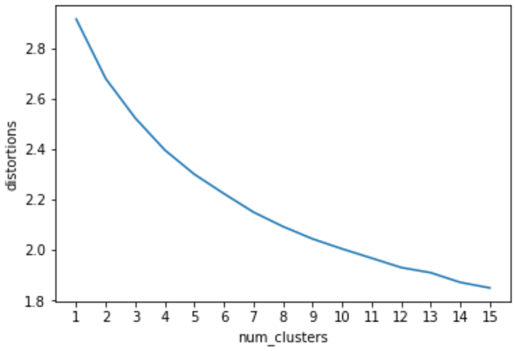


Figure 1 - K-means within cluster sum of squares.

After the choice of features, the within cluster sum of squares was calculated and is presented above in figure 1 (let distortions be the within cluster sum of squares). The ideal number of clusters is not easily identifiable in figure 1 as the elbow is not clear, but after some attempts, the desired number was set to 5. After running the k-means algorithm for the original dataset (let the original dataset be the subsampled dataset for 1,000 tourists without the application of any sampling and completion techniques), the final centroids were taken out and were used as the initialization points for all new k-means clustering's. In other words, these centroids were used as the initialization points for a second run of the k-means in the original dataset and in the datasets retrieved from running the linear interpolation with 1 hour uniform sampling, 15 minutes during the day and 45 minutes during the night uniform sampling, and CDR sampling mimicking 1 hour uniform. These initialization points can be seen below in *Appendix C* (rounded to 2 decimal points).

After the k-means algorithm was performed for all datasets, the t-Distributed Stochastic Neighboring Embedding (t-SNE) was used to perform dimensionality reduction in order to be able to plot the clusters in a 2D graph. Since the t-SNE keeps score of the distances but not of the actual points in space, the visualizations performed are merely for illustration purposes. In order to compare the results for different sampling rates, the % of tourists whose categorization into a cluster changes was calculated.

Typically, the k-means algorithm would be initialized several times with random positions to evaluate which cluster configuration would come up more often. However, as it has been stated above, in order for the clustering results to be comparable, the initialization points were pre-determined and set for all approaches. As the k-means is highly sensitive to the starting points of the algorithm and as these points were chosen somewhat arbitrarily, no conclusions are drawn on the actual meaning of the clusters or on how well they perform. Instead, the k-means is evaluated by comparing the differences in the % of tourists whose categorization into a given cluster changes.

Furthermore, it is important to emphasize that the accuracy of these clustering algorithms (DBSCAN in the trajectory stops clusters and k-means in the tourist clusters) is not being evaluated in this specific study. The purpose of these analyses is solely to compare observable changes in the outputs of these algorithms originated by the changes in the underlying dataset. To originate the datasets of the uniform and CDR sampling approaches, due to the fact that each sampling and completion method was performed 20 times, the antenna considered as the base-station for a user at a given point in time was the mode of the antennas obtained over the 20 runs.

4. RESULTS

4.1. DESCRIPTIVE STATISTICS

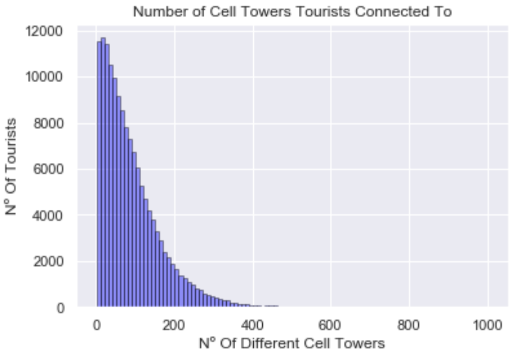


Figure 2 - Number of cell towers tourists connected to on the week of the 1st of May 2017.

Figure 2 shows the distribution of the number of unique cell towers that the tourists connected to, as observed in our one-week full dataset (consider all descriptive analysis in section 4.1 as having been performed on the full dataset). We can conclude that 25% of tourists were registered in at most 33 different cell towers, while 75% were registered in at most 125 unique antennas during the period. The maximum number of antennas that a tourist connected to was 1,004 and the mean was 89.9 unique cell towers, which hints at a large span of movement of the tourist population.

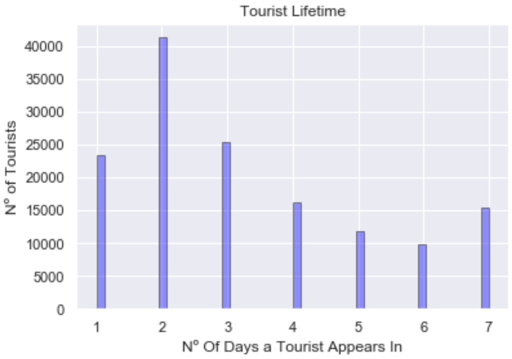


Figure 3 - Number of days tourists appeared in the dataset during the week of the 1st of May 2017.

Figure 3 showcases the tourists' short lifetime despite the fact that the period being considered was of only a week. Moreover, most tourists appear only in a limited number of days during the week, namely 1, 2 and 3 days. In fact, 28.8% of the tourists appeared in the dataset during exactly 2 days, while the mean period of stay was 3.3 days. It is also important to notice that the tourists that appeared for 7 days might, in reality, be in Italy for longer than just a week, which is longer than our observation period. This could explain the increase from the number of tourists that appeared in the dataset for 6 days and the number of tourists that appeared for 7 days.

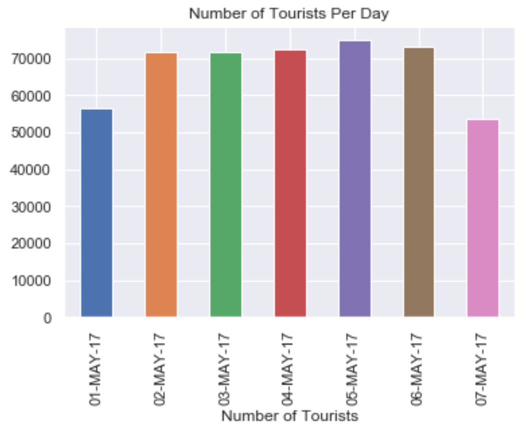


Figure 4 - Number of tourists per day during the week of the 1st of May 2017.

Furthermore, the number of tourists per day is, for the most part, uniformly distributed, as shown in figure 4, only diverging in the first and last day due to the removal of tourists that only appeared in the dataset on the first or last day of the period. In fact, every day from the 2nd until the 6th of May gathers around 15% of the number of tourists per day each, with the 5th of May concentrating almost 16% of the number of tourists per day, whereas the 1st and the 7th of May gather around 12 and 11%, respectively.

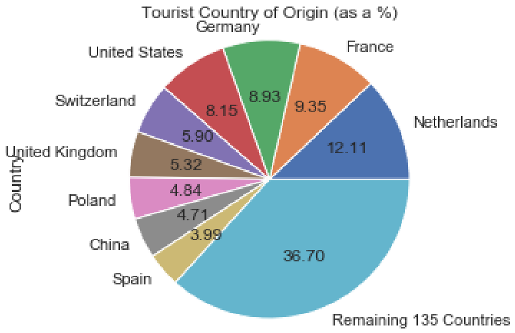


Figure 5 - Tourist country of origin (as a %) during the week of the 1st of May 2017.

As figure 5 shows, the tourists in the dataset have a total of 172 different nationalities. The most popular nationalities during the week of the 1st of May 2017 were Dutch, French, German, American, Swiss, British, Polish, Chinese and Spanish. This seems to be aligned with national statistics as, according to Euromonitor International (2018), the most frequent tourist nationalities in Italy during 2017 were, in descending order, German, French, British, Austrian, American, Swiss, Spanish, Dutch and Polish.

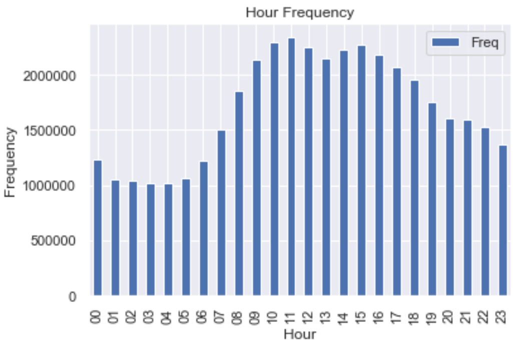


Figure 6 - Hourly distribution of tourists during the week of the 1st of May 2017.

Figure 6 demonstrates that most recorded interactions occur between 9 am and 6 pm, inclusively. Actually, 53.7% of the total number of tourists recorded in any given hour were recorded during this time frame. This could have been caused by an increase in the number of tourists during the daytime, or due to the fact more movements should be recorded during the day under the nighttime stability assumption, both of which would increase the number of signaling data points per tourist in the dataset. The time frame from 9 am to 6 pm, inclusively, will be referred to, from here onwards as daytime, while the time frame from 6 pm to 9 am, exclusively, will be referred to as nighttime.

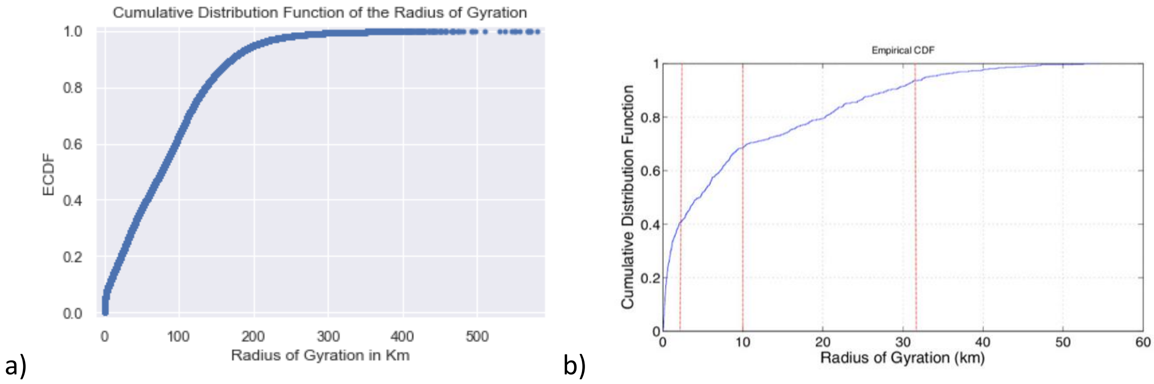


Figure 7 - Cumulative distribution function of the radius of gyration: a) for the tourists in the dataset during the week of the 1st of May 2017; b) for the non-tourist population. Reprinted from 'Estimating human trajectory and hotspots through mobile phone data', by Hoteit, S., Secci, S., Sobolevsky, S., Ratti, C., & Pujolle, G. (2014). Computer Networks, 64, 296-307. Copyright 2014 by Elsevier B.V.

Figure 7a showcases that about 80% of the tourists move in a radius of up to 135.36 km from the center of mass of the tourist's recorded displacement, whereas Hoteit et al. (2014) found that, for the general population, 80% of the individuals would move in a radius of up to about 20 km, which can be observed in figure 7b. This shows that the redundancy assumption previously used in the literature may not hold for tourists, as they tend to move more. In fact, as the average distance between each tower and its closest tower is approximately 1.50 km, on average, 80% of tourists connected to approximately 90 different cell phone towers, whereas, if the individuals from Hoteit et al. (2014) were to have a similar average distance between neighbouring towers, then 80% of these individuals would, on average, have connected to 13 different towers.

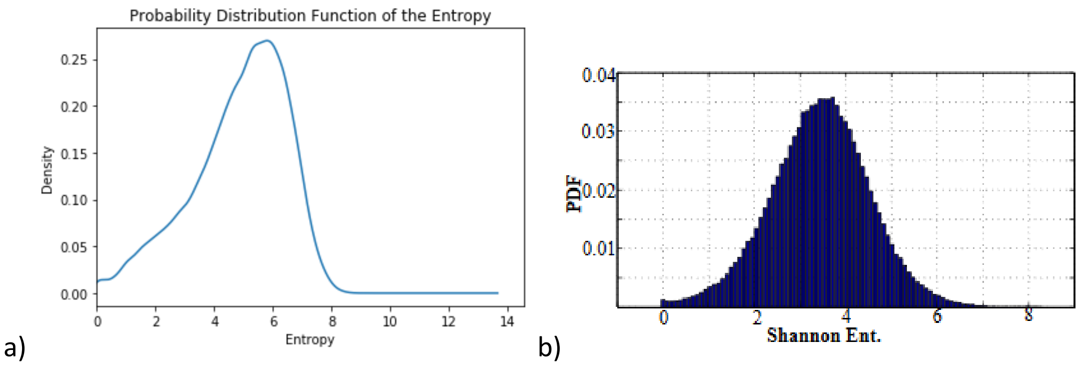


Figure 8 - Probability distribution function of the entropy: a) for the tourist population in the dataset during the week of the 1st of May 2017; b) for the non-tourist population. Reprinted from 'Are call detail records biased for sampling human mobility?', by Ranjan, G., Zang, H., Zhang, Z.L., & Bolot, J. (2012). ACM SIGMOBILE Mobile Computing and Communications Review, 16(3), 33-44. Copyright 2019 by ACM, Inc.

Moreover, by comparing the density function of the entropy of the tourists, figure 8a, with the finding of a non-touristic mobility study in figure 8b (Ranjan et al., 2012), it can be concluded that the tourists' entropy is higher than the full population. This proves that redundancy in movement, an important metric to ensure the predictability of the trajectories, is not as strong in touristic movement. Together with the results observed in the radius of gyration, this demonstrates that tourism data does not have much redundancy, which might influence the quality of the trajectory estimates.

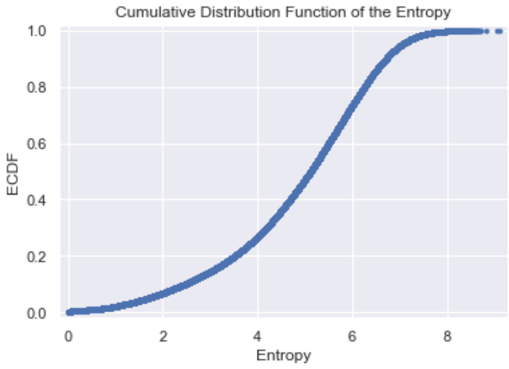


Figure 9 - Cumulative distribution function of the movement entropy during the week of the 1st of May 2017.

Moreover, as it can be observed in figure 9, the cumulative distribution function of the entropy follows an S curve. The maximum entropy value is 9.10 while the minimum is 0.0 (which occurs when a tourist remains stationary at all times in the dataset). The mean entropy level is, approximately, 4.85.

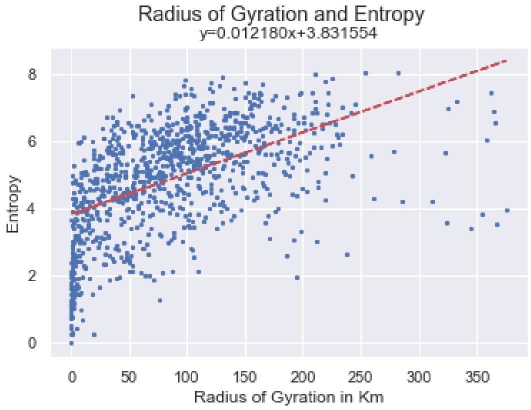


Figure 10 - Radius of gyration and entropy (1,000 tourist subsample).

In addition, as it can be observed in figure 10, there is a positive relationship between these the radius of gyration and the entropy level. This analysis was conducted on the 1,000 tourist subsample for better visualization; nonetheless, the same relationship occurs in the full dataset.

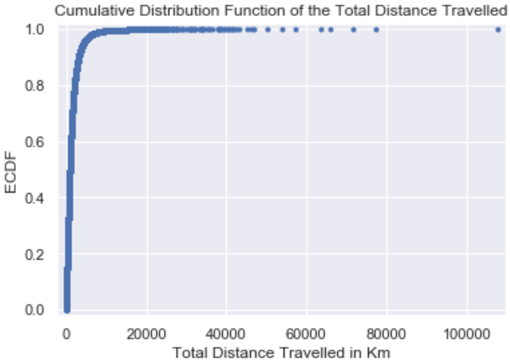


Figure 11 - Cumulative distribution function of the total distance travelled during the week of the 1st of May 2017.

Most of the tourists were concentrated in relatively smaller total distances, as figure 11 displays. In fact, 25% of the tourists traveled at most, approximately, 491.33 km, 50% traveled at most, approximately, 972.24 km, and 75% traveled at most, approximately, 1,825.59 km. The minimum distance travelled is 0 km while the maximum is, approximately, 107,523.06 km. The mean total distance travelled is about 1,472.30 km.

Finally, the distance between each antenna and any other antenna (that corresponds to a different set of coordinate pairs) was computed using the haversine formula previously mentioned. It was found that while the mean distance between towers is 429.47 km, the minimum distance is 0.013 km and the maximum distance is 1,281.66 km. This showcases the extensive coverage of the dataset. Moreover, the average distance between any tower and its closest neighbor is, approximately, of 1.50 km.

4.1.1. Nighttime stability assumption

As stated previously, if nighttime stability can be assumed, then a method of sampling with a higher granularity during the day than during the nighttime can be performed. To understand if this is the case, this sub-section takes some of the previously studied mobility metrics, analyzing and comparing them both during the day and nighttime.

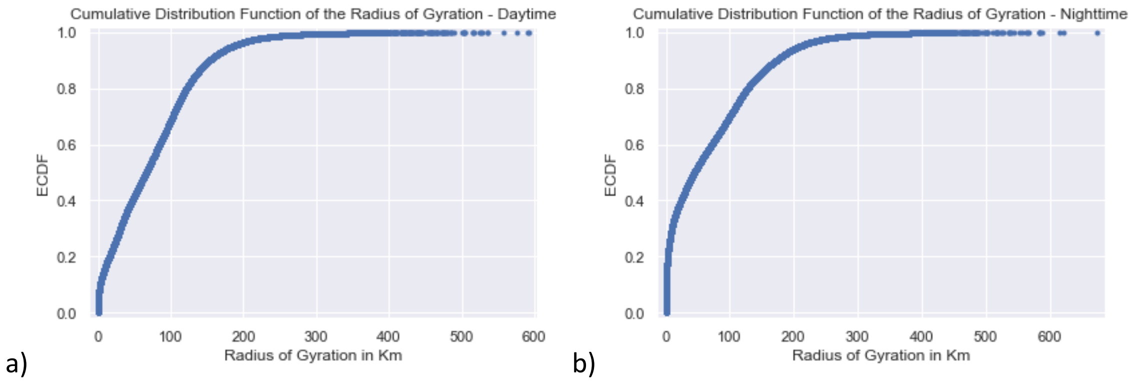


Figure 12 - Cumulative distribution function of radius of gyration during the week of the 1st of May 2017: a) during the daytime; b) during the nighttime.

Figure 12 shows that the radius of gyration during the day grows at a slightly steeper rate than during the night. The average radius of gyration during the day is higher than the average radius of gyration during the night, 74.47 km and 68.67 km, respectively, which is consistent with the nighttime stability assumption, though the difference is not very large. However, the daytime standard deviation is smaller than the standard deviation at nighttime, 66.00 km and 73.48 km, respectively. Roughly, 58 % of tourists have a daytime radius of gyration bigger than the nighttime radius of gyration while 42 % experience the reverse.

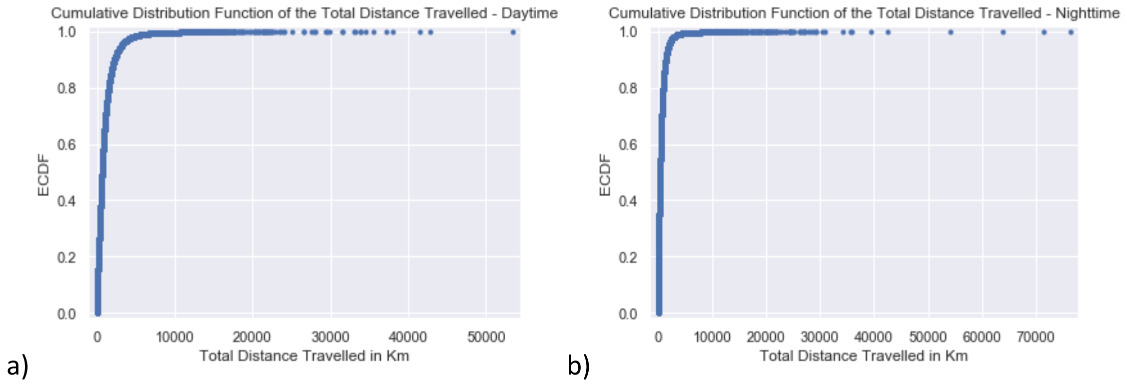


Figure 13 - Cumulative density distribution function of the total travel distance during the week of the 1st of May 2017: a) during the daytime; b) during the nighttime.

Moreover, the total distance travelled (figure 13) shows an average daytime, median and standard deviation (1,076.38 km, 661.26 km and 1,324.28 km, respectively) higher than those of the nighttime (614.52 km, 375.60 km and 1,049.58 km, respectively). It is important to mention, however, that the ping-pong effect is included in the total distance travelled metric and, as such, by itself this finding is not enough to discard the assumption of nighttime stability. Recall that the ping-pong effect happens when repeated handover between antennas due to recurrent changes in signal strength exists.

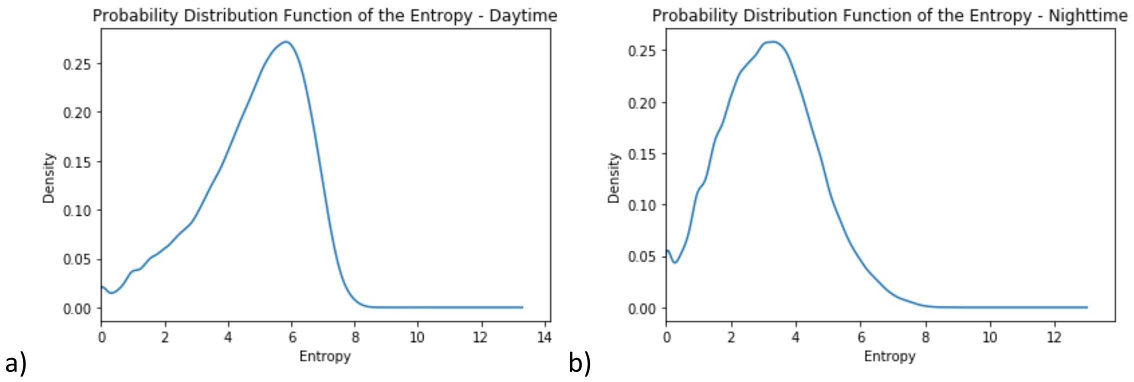


Figure 14 - Probability distribution function of the entropy during the week of the 1st of May 2017: a) during the daytime; b) during the nighttime.

Finally, as it can be observed in figure 14, the entropy level is also, on average, higher in the daytime than in the nighttime (4.82 and 3.16, respectively). The median of this metric is, once again, higher during the day (5.12 in the day and 3.13 in the nighttime), while the standard deviations are more similar (1.64 in the day and 1.48 in the nighttime).

All in all, since the radius of gyration and entropy have proved to be more stable during the nighttime, the assumption of nighttime stability in tourists will not be discarded at this point. Even though it is expected that not all tourists stay in the same location during the night due to, for example, nighttime travelers. On average it will be considered, from this point onwards, that the nighttime stability assumption holds sufficiently and that the uniform sampling with different rates during the day and nighttime will be implemented in a tourist dataset in order to evaluate the merit of this assumption.

As it has been stated before, due to the sampling and completion methods in section 4 being computationally intensive, a random subsample of 1,000 tourists was used to perform all analyses from this point onwards. Nevertheless, efforts were made to keep this subsample similar to the full dataset by comparing the distributions of relevant metrics examined during the data exploration process (number of cell towers tourists connected to, tourist lifetime, number of tourists per day, % of nationalities represented in the dataset, number of records in a given hour of the day, radius of gyration, entropy and total distance travelled) in several subsamples until a subsample with distributions that resembled the originals was discovered. These distributions can be found in *Appendix D*. For simplicity, going forward, we will refer to this dataset subsample simply as the dataset.

4.2. Sampling and completion results

In this section, we will show how the Mean Distance Errors per sampling and completion combination (average of the Mean Distance Errors of all the tourists in each combination) compare against each other and will draw conclusions on the tradeoff between estimate error and granularity.

Furthermore, we will show the relationship between the Mean Distance Error and the mobility metrics presented in the section 3 in order to determine how the estimate error could change given the profile of each tourist. In order to visualize the relationship between the errors and the mobility metrics, the data of the continuous metrics (radius of gyration, total travel distance and lifetime in hours) was binned. The bins correspond to the deciles of each metric, so that the same number of tourists is considered in each bin.

Finally, the trajectories and the antennas that the tourists connect to are visualized for a tourist with high radius of gyration and entropy and for a tourist with low radius of gyration and entropy. A k-means clustering algorithm is also applied to understand how sensitive some practical applications would be to changes in the underlying dataset.

4.2.1. Mean distance error

In order to analyze the tradeoff between the Mean Distance Error and the data granularity, the completeness of each sampling approach was computed as the number of records in the dataset that were sampled divided by the total number of records in dataset. Consider that in this case the number of records sampled is equivalent to the number of rows sampled since sampling was made with replacement.

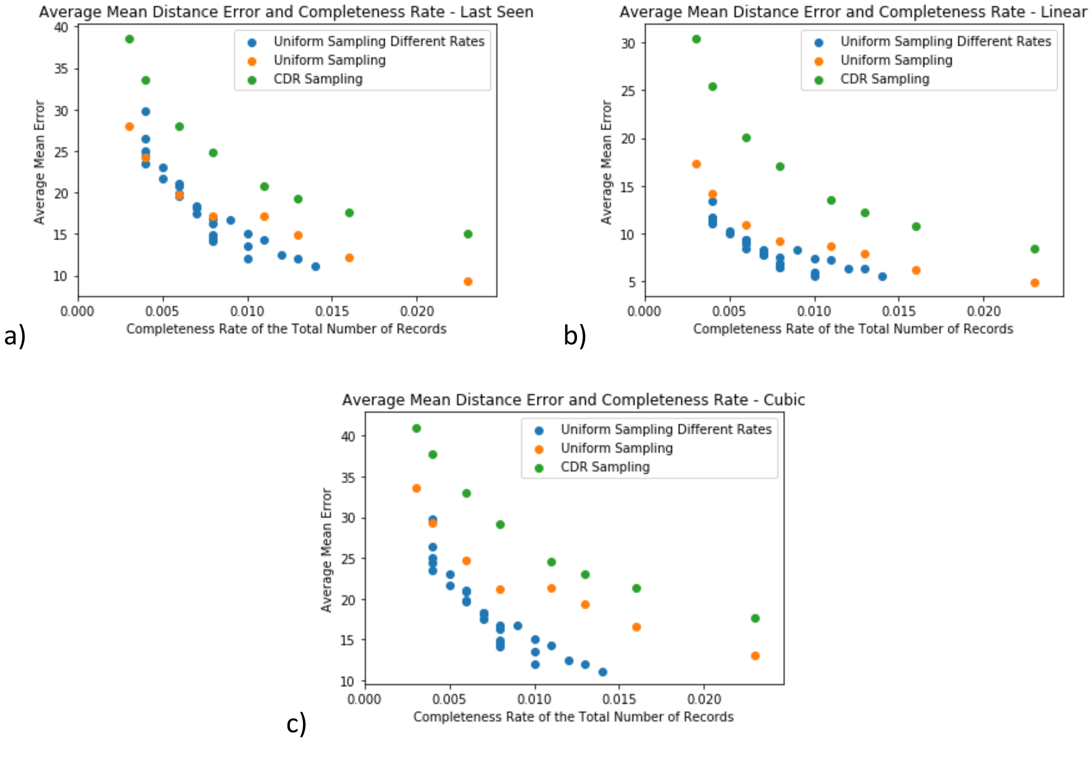


Figure 15 - Average mean distance error and completeness rate: a) last seen interpolation; b) linear interpolation; c) cubic interpolation.

As it can be observed in figure 15 and as it would be expected, typically, the higher the completeness, the better the results of the interpolations (and the worse the cost of storage). This would be expected due to the fact that there are more points that can be used to determine the missing ones, which is consistent with the findings of Hoteit el al. (2014). Although this does not occur for every combination of sampling and completion techniques, the general trend is that sampling more during the day than during the night provides a lower estimate error. The exceptions to this trend were a few combinations of the uniform sampling with different rates during the day and during the night, and the CDR mimicking (the latter will be explained below). For instance, take the case of linear interpolation, for the uniform sampling with a 1 hour and 30 minutes rate (4th orange point from the left with a completeness level of 0.8% and a Mean Error of 9.25 km). There are several combinations of the uniform sampling with different rates during the day and during the nighttime that provide lower Mean Distance Errors for the same level of completeness/ storage costs (rates of 15 minutes during the day and 3 hours during the night, 15 minutes during the day and 4 hours during the night, 30 minutes during the day and 1 hour and 30 minutes during the night, and 30 minutes during the day and 2 hours during the night, corresponding to errors 6.94, 6.47, 6.53 and 6.91 km, respectively).

Appendix E can be referenced to in order to observe the statistics of each of the sample and completion approaches combination. Recall that the completeness rate is very small for all approaches due to the fact that 1 record does not correspond to 1 row in the dataset.

For the specific case of the Uniform Sampling with different rates during the day and during the night, a change in the daytime sampling rate makes the Mean Distance Error grow faster than a change in the nighttime sampling rate (observable in *Appendix E*), which is aligned with the assumption of Chen et al. (2018) of nighttime stability of movement. Despite this, 42% of tourists move more during the night than during the day in the tourist dataset.

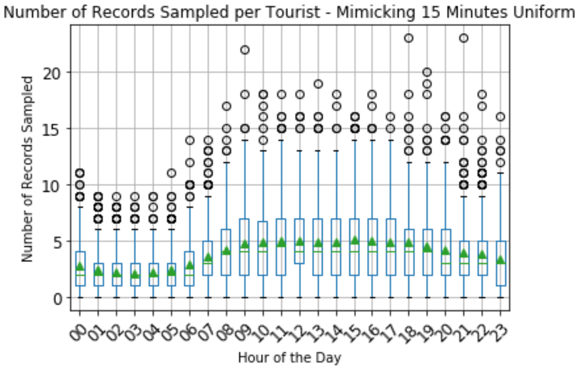


Figure 16 - Number of records sampled per tourist per hour: mimicking 15 minutes uniform sampling.

For the three completion methods (last seen, linear and cubic), it can be observed that CDR sampling has an error that is typically higher than the remaining sampling methods. The reason as to why the CDR sampling performs the worst may be tight with the fact that there is a high dispersion in the number of records sampled per hour per tourist. For instance, as figure 16 displays, for the mimicking of the 15 min uniform sampling, the average number of records sampled per tourist per hour spans from 2.05 to 5.04 (more often during the daytime than during the nighttime). This would be expectable since at a 15 minute rate, we could at most sample uniformly 4 times. However, as the sampling was performed on the whole dataset and not on a per tourist basis, the number of records sampled per tourist is highly variable (with a standard deviation spanning from 1.68 to 3.45, although many outliers can be observed in the boxplots of figure 16), which can explain the higher Mean Distance Error of its final estimates. The number of records sampled per tourist per hour of the remaining CDR sampling rates can be observed in *Appendix F*.

With the exception of the uniform sampling with a sampling rate of 15 minutes, 30 minutes or 4 hours, it can be observed in figure 15 that there is at least one uniform sampling with different rates during the day and during the night that offers a lower Mean Distance Error for a similar level of completeness.

This means that, although some tourists travel during the nighttime, on average, the nighttime stability assumption holds enough so that we can use this new sampling approach.

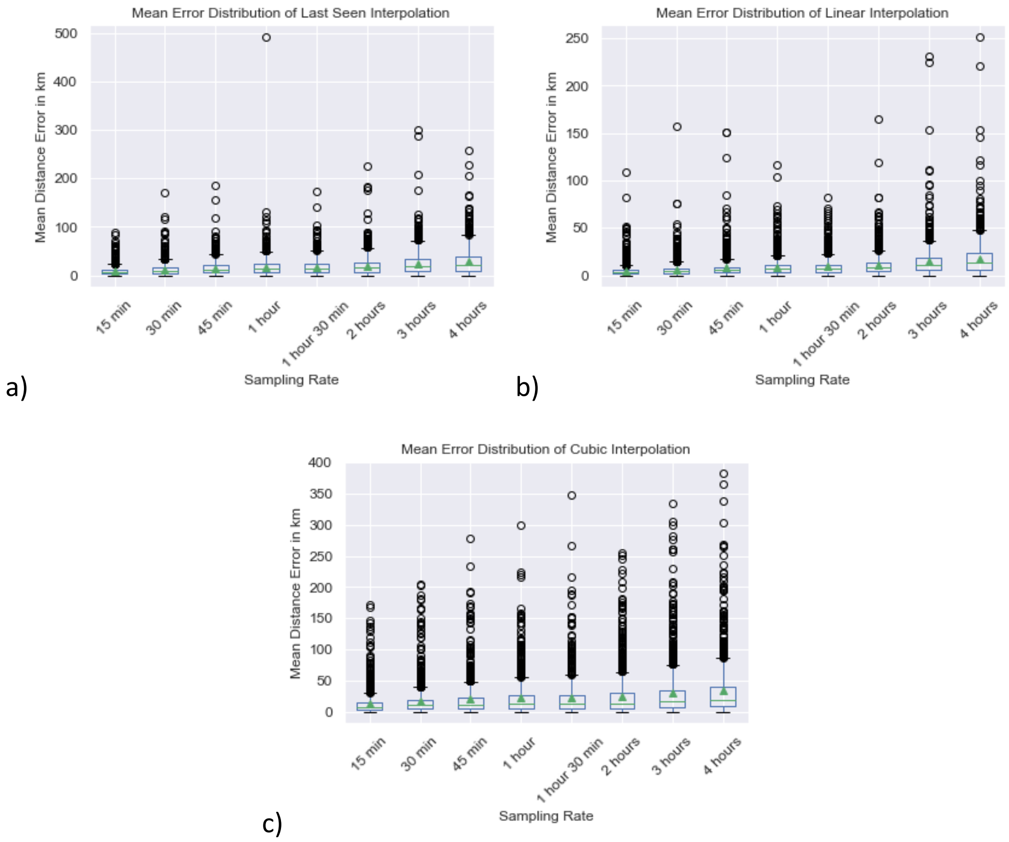


Figure 17 - Mean error per tourist per sampling rate: a) last seen interpolation; b) linear interpolation; c) cubic interpolation.

In order to illustrate how the Mean Distance Error per tourist changes within the same sampling and completion method, a boxplot of the Mean Distance Error per tourist was created for every sampling rate of the uniform sampling method. As it can be seen in figure 17, not only does the average Mean Distance Error increase with the decrease of granularity but the median and the dispersion do as well. Since the data is skewed, the median (indicated by the green line) is smaller than the average (indicated by the green triangle) and the distance between the two increases with the decrease of data granularity due to the increase in dispersion. Furthermore, the dispersion increases with the decrease of data granularity since it is harder to make accurate and consistent estimations the bigger the gap between consecutive records is.

Moreover, it can be observed in figures 15 and 17 that, on average, the linear interpolation method is the best at any sampling rate as both the mean and the median of the Mean Distance Error are smaller than in any other method. Nonetheless, the last seen approach performs similarly to the linear interpolation method when compared to the cubic method. These results will be confirmed and explored in more detail in section 4.4.

4.3. Mean distance error and mobility metrics per sampling rate

In order to illustrate how the Mean Distance Error changed for individuals with different mobility behaviors, the mobility metrics were plotted for a few sampling approaches (uniform sampling with sample rates of 15 minutes and 1 hour; uniform sampling with different rates during the day and nighttime with sample rates of 15 minutes day and 45 minutes night as well as 1 hour day and 2 hours night; CDR sampling mimicking the 15 minutes and 1 hour rates) for the linear completion method. The same plots for the last seen and cubic methods can be observed in *Appendix G*.

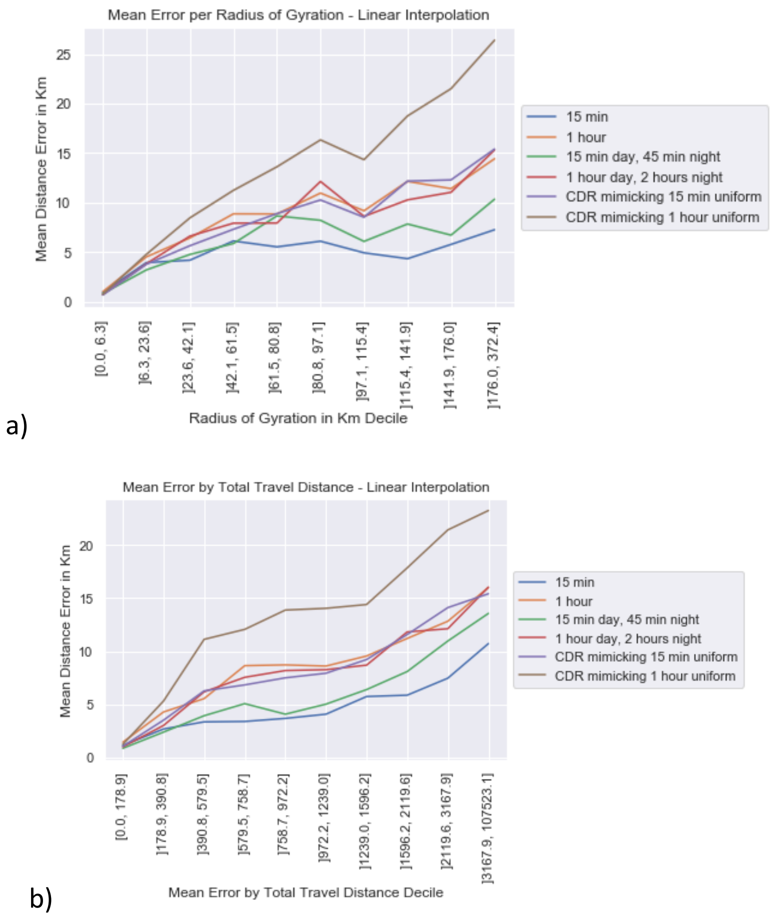


Figure 18 - Mean error versus: a) radius of gyration - linear interpolation; b) total travel distance - linear interpolation.

As it can be observed in figure 18, the two measures of the span of movement of an individual, the radius of gyration and the total distance travelled, tend to have positive relationships with the Mean Distance Error. It is also important to point out that for the last seen and the cubic interpolation methods (*Appendix G*) the Mean Error grows more smoothly when plotted against the radius of gyration than when plotted against the total distance travelled. This might be due to the fact that the total travel distance is affected by the ping-pong effect, while the radius of gyration is less sensitive to it. It also might be due to the fact that an increase of the radius of gyration represents, directly, an increase in the span of movement, meaning that the tourist will be recorded in a higher number of different towers. Additionally, the total distance travelled can increase due to an increase of the number of towers a tourist connects to or to an increase of the frequency of changes between the towers where he/she is recorded.

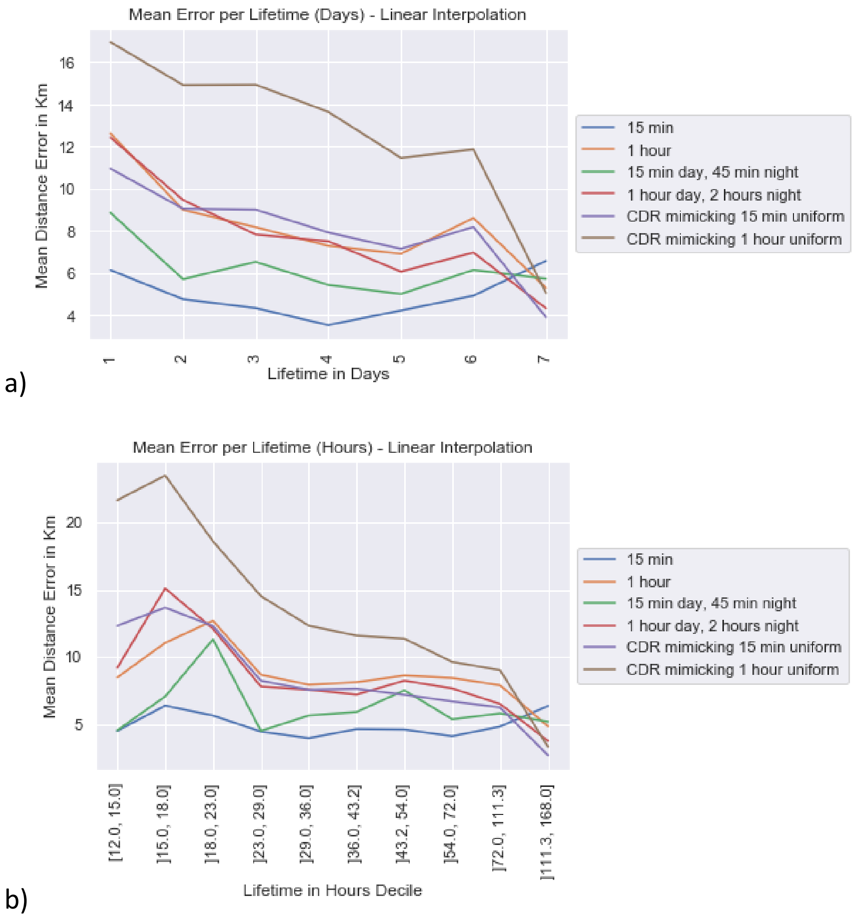


Figure 19 - Mean error versus: a) lifetime in days - linear interpolation; b) lifetime in hours - linear interpolation.

On average, the longer a tourist stays on the dataset, the lower the Mean Distance Error becomes, as shown on figure 19. This could be a result of tourists moving more slowly the more time they stay in their destination (e.g. a tourist that stays for 1 day will need to stay at each location for a smaller amount of time than a tourist that stays for 3 days and wishes to visit the exact same Points of Interest, or stays longer at a hotel). Figure 19 showcases this finding, although the Mean Distance Error increases when the tourist stays for fewer than 18 or fewer than 23 hours, depending on the granularity rate. A possible explanation for this occurrence might be the fact that during the first day tourists move more (e.g. a tourist that arrives in a given city by airplane and needs to travel to their destination city by car/train, or a tourist that only stays in Italy for a day and, hence, connects to more towers due to travelling in and out of the country). Moreover, the spillover issue can explain the variation of the peak in the error for the class]15.0; 18.0] (i.e. uniform 15 minutes, uniform 1 hour during the day and 2 hours during the night, CDR 15 minutes, and CDR 1 hour) and for the class]18.0; 23.0] (i.e. uniform 1 hour and uniform 15 minutes day and 45 minutes night).

Contrary to what is expected, for the uniform sampling with a rate of 15 minutes, in the last decile, the Mean Error increases both for lifetime in days and lifetime in hours. This phenomenon reoccurs in the last seen and cubic interpolation method. As no relevant relationship between the lifetime and other metrics was found to explain this result and as this only happens for the highest granularity level, it was assumed that this was a product of the data collection and the spillover issue. If we consider that individuals with a longer lifetime (7 or more days) move more slowly (less frequently) than those with shorter lifetime (2 or 3 days), then, there would be more instances in which the tourists with longer lifetimes had been recorded only a few times per hour. For a high granularity level such as the 15 minute rate, the spillover effect can skew the results, which, in turn, would make the error go up. The reason as to why this does not occur for the uniform sampling with rates of 15 minute during the day and 45 minute during the night is that the increase in error for the daytime rate was offset by the lower error during the nighttime for this specific decile. This is due to the fact that the 45 minute sampling rate samples from multiple hours (e.g. 00h45, 01h30, 02h15).

Finally, it is important to state that even though the plots in figures 18 and 19 suffer from some discrepancies, the overall trends are that a higher span of movement leads to an increase of the error and a higher lifetime to a decrease of the error. The reason as to why some of these bumps have occurred may be related to the fact that the sampling and completion techniques were performed in a sample of only 1,000 tourists and repeated only 20 times. It is to be expected that these plots smooth out for a higher number of tourists and higher number of runs of the sampling and completion techniques.

4.4. Mean distance error and mobility metrics per completion method

As it has been mentioned in section 2.4, Hoteit et al. (2016) found the cubic interpolation to return lower errors than the last seen interpolation, and Hoteit et al. (2014) found better results for the linear interpolation when the radius of gyration was smaller than 3 km and better results for cubic interpolation when the radius of gyration was higher than 32 km. It is important to notice that a radius of gyration of 32 km or higher for the latter study corresponded to the individuals who were the most mobile. On the other hand, has it has been previously stated, the average alone of radius of gyration of the tourists in the full dataset used for this analysis is of 135.54 km, proving that there are significant differences between the two populations.

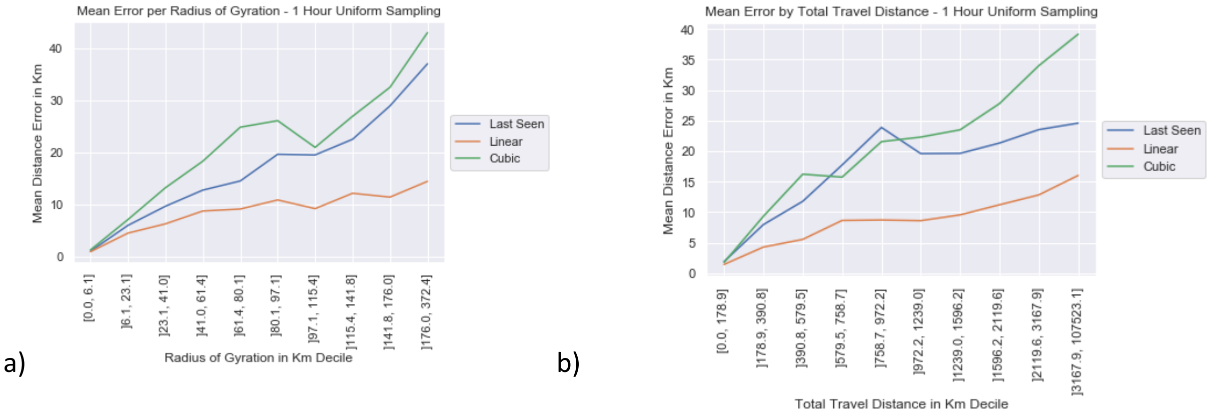


Figure 20 - Mean error for the 1 hour uniform sampling: a) per radius of gyration; b) per total travel distance.

By looking at the results of figure 20a, it can be seen that the linear interpolation outperforms all other approaches and that the cubic interpolation performs the worst regardless of the radius of gyration of the tourists. This is consistent with figure 20b. However, for the total travel distance (figure 20b) it can be seen that for distances between 579.5 and 972.2 km the cubic interpolation outperforms the last seen due to a peak in the error of the last seen approach. This can once more be due to the fact that the sampling and completion processes were done for only 1,000 tourists, only 20 times. Smoother growths can be expected for a bigger population over a larger number of runs.

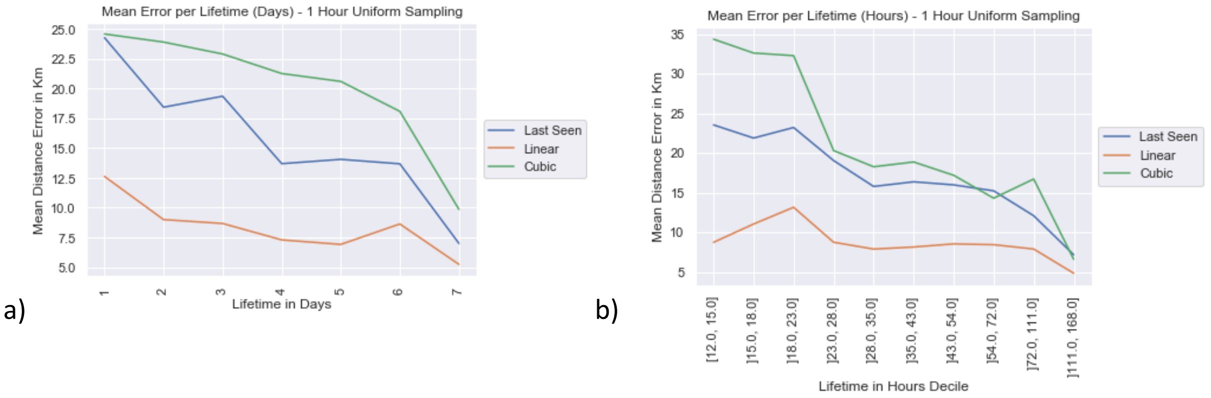


Figure 21 - Mean error for the 1 hour uniform sampling: a) per lifetime in days; b) per lifetime in hours.

Figure 21 reinforces the previous findings, since, as it can be observed, the linear interpolation always outperforms the cubic and last seen. It can also be seen that the cubic approach is in fact the worst of all approaches in minimizing the estimate error. Nonetheless, for the lifetime in hours (figure 21b), the cubic interpolation slightly outperforms the last seen, which can, once more, be tied with the fact that only 1,000 tourists and 20 runs of the algorithm were used.

It is possible that these results contradict those found by Hoteit et al. (2014) due to the data collection issue and the roaming issue that lead to higher gaps in the dataset than expected if each tourist had been recorded exactly once per minute. As the points are not equidistant, the cubic interpolation might be affected and underperform.

Nonetheless, it is also a strong possibility that the results occur due to the fact that tourists move in a much different manner than the non-tourist population. Has it has been confirmed previously in this study, redundancy may not be assumed for tourists while it is generally accepted for the non-tourist population. Furthermore, lifetime is significantly shorter and the radius of gyration and entropy of movement significantly higher. The different underlying behaviour of the tourists may be causing the changes in results when compared to the non-tourist literature.

4.5. Trajectory analysis’ output changes

To illustrate how the reconstructed trajectories change with the different sampling methods, plots showing the trajectories and stops and the daily diaries of two tourists are shown. From these tourists, one has high radius of gyration and entropy (radius of gyration of 211.33 km and entropy level of 7.24 entropy) and another low radius of gyration and entropy (radius of gyration of 9.16 km and entropy level of 1.97). To perform these visualizations, the python library scikit-mobility by Pappalardo et al. (2019) was used.

Furthermore, the k-means clustering algorithm has also been applied to group similar types of tourists together. The results of which have been plotted in a 2D graph by using the t-Distributed Stochastic Neighbor Embedding (t-SNE) as a technique for dimensionality reduction.

In this section, we will perform these analyses for the original dataset and one sample rate per sampling method. The sample and completion methods that were picked for illustration purposes were the linear interpolation of uniform sampling at 1 hour rate, the uniform sampling at 15 minutes rate during the day and 45 minutes rate during the night, and of the CDR mimicking 1 hour uniform completeness.

4.5.1. Trajectories and stops

Although it has been shown that different sample and completion methods have different Mean Distance Errors, in the big scheme of things, these differences may not translate. The clusters of stops (recall that a stop was considered a location in which a tourist stayed for more than 1 hour) provided by the DBSCAN are observed in the trajectory plots below as well as in the tourist diary of stops, in which a different colour corresponds to a different cluster.

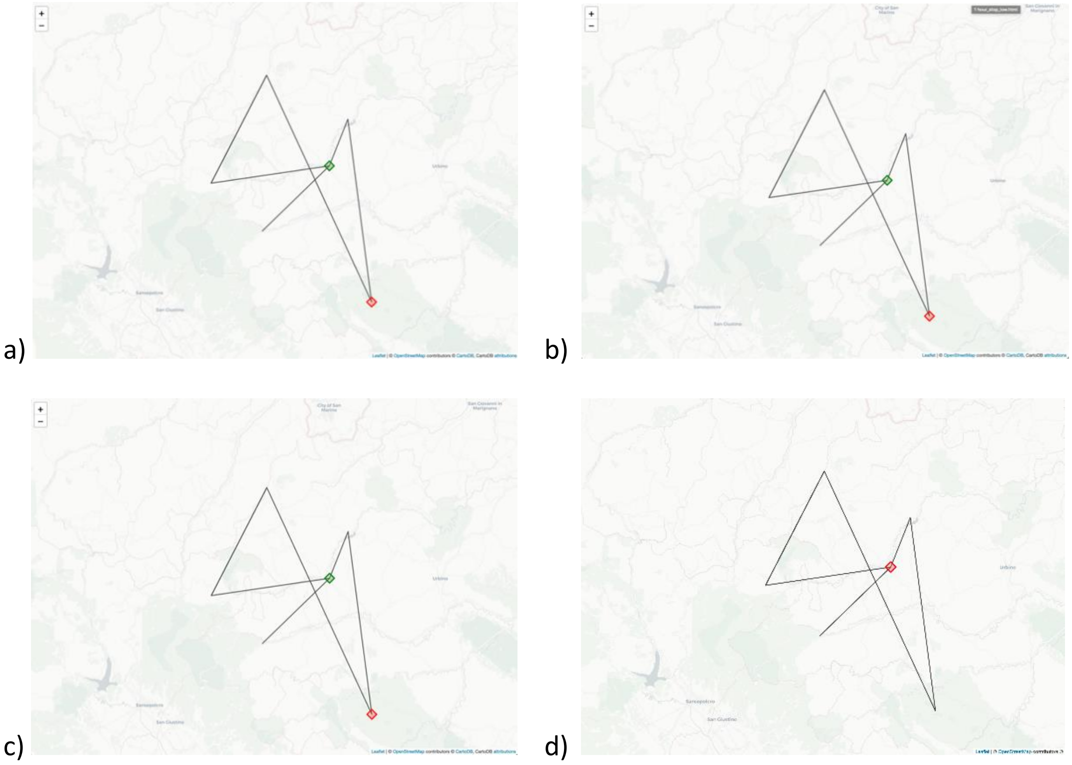


Figure 22 - Trajectory and cluster of stops of a tourist with low radius of gyration (9.16 km) and entropy (1.97): a) original dataset; b) 1 hour sampling; c) 15 minutes day and 45 minutes night sampling; d) CDR mimicking 1 hour uniform completeness.

At first, this study looks at the trajectories and clustered stops of the tourist of low radius of gyration and level of entropy. Even though figure 22 seem to be repeated, in fact, it represents different sampling and completion methods with different Mean Distance Errors that have been previously discussed. Since low radius of gyration returns, on average, better estimation results, it is not surprising that figures 22a, 22b and 22c show a very high level of accuracy for this individual. Figure 22d only picks up 1 cluster of stops and not 2, which may be stemming from the CDR sampling issue discussed previously.

It is important to notice that, even though, the representation of the trajectories looks the same, there may in fact be some underlying errors on the estimations that are not being picked up by the visualization tool. In other words, the estimated trajectories for all the trajectories plotted in this section can still differ from the actual trajectory, even though small variations would not visible to the naked eye.

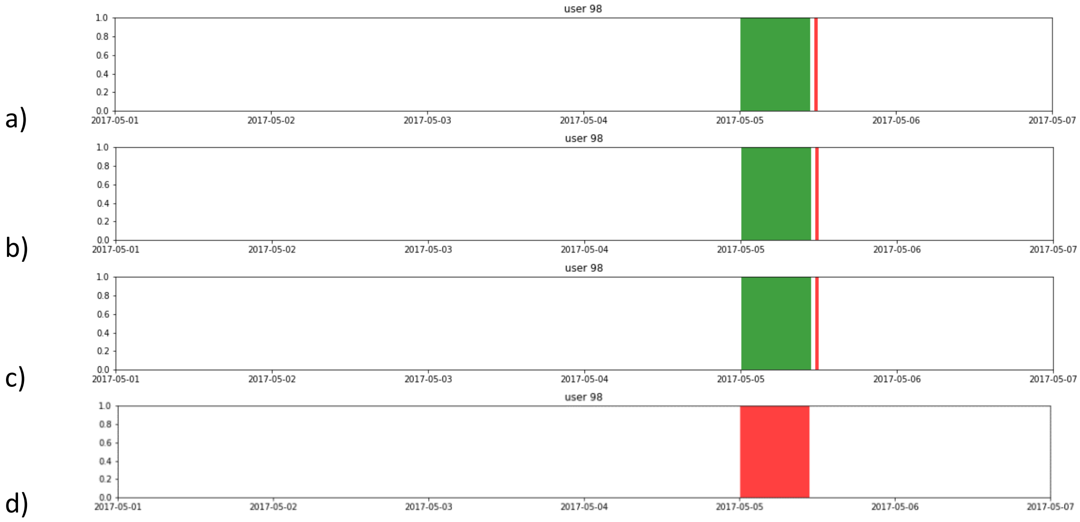


Figure 23 - Tourist diary of a tourist with low radius of gyration (9.16 km) and entropy (1.97): a) original dataset; b) 1 hour sampling; c) 15 minutes day and 45 minutes night sampling; d) CDR mimicking 1 hour uniform completeness.

Regarding the amount of time the tourist stays in each stop cluster, once again, figure 23 showcases very similar results, with only figure 23d deviating slightly, for the same reasons as before. Recall that stop clusters were performed by the DBSCAN algorithm to group together close stopping points. However, the identified stops by the algorithm do not correspond to all the actual stops of a person (towers in which a tourist stayed connected to for more than 1 hour). Instead, they correspond to groups of nearby towers in which the individual stayed for long periods of time. These Tourist Diary plots are relevant to understand whether or not the tourist would stay for longer or shorter in a given cluster for a given method than for the remaining methods, which does not seem to be the case for this particular tourist.

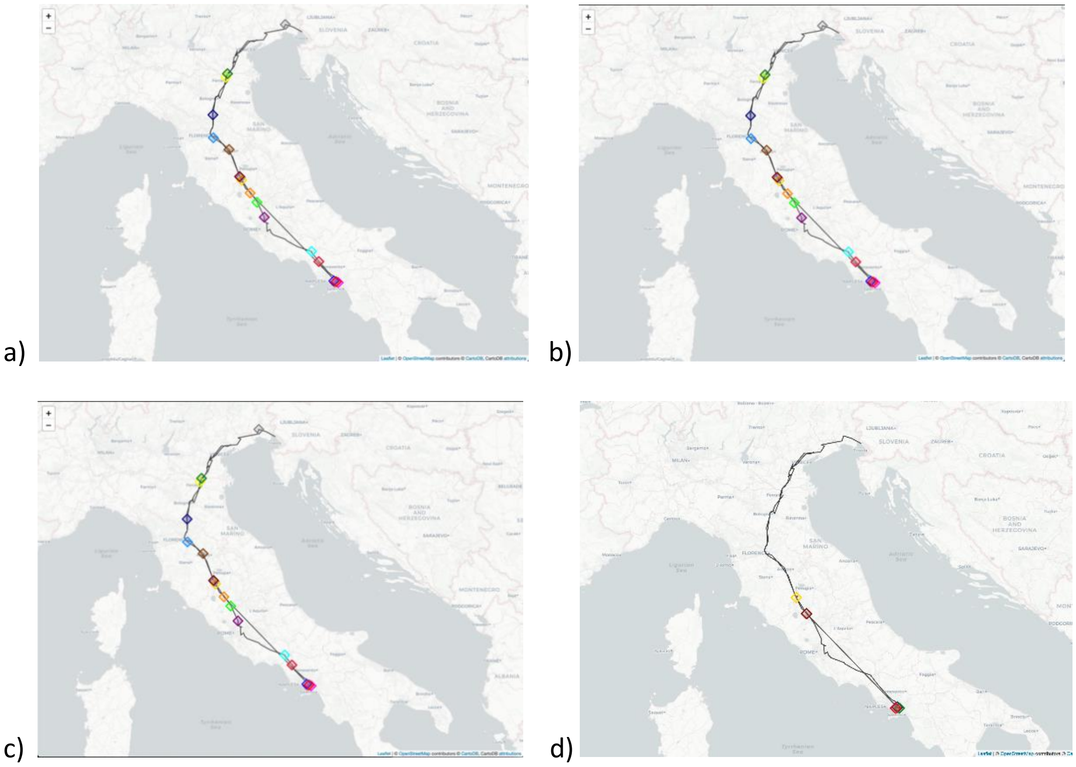


Figure 24 - Trajectory and cluster of stops of a tourist with high radius of gyration (211.33 km) and entropy (7.24): a) original dataset; b) 1 hour sampling; c) 15 minutes day and 45 minutes night sampling; d) CDR mimicking 1 hour uniform completeness.

After looking at the trajectories and clusters of stops of the tourist with a high radius of gyration and entropy level it could be wrongly assumed that all the trajectories and clusters of stops were, once again, the same since figure 24a, 24b, 24c and 24d seem similar at first glance as they are plotted at a macro level. Moreover, the CDR sampling has also picked up fewer clusters of stops than the original dataset or the remaining approaches.

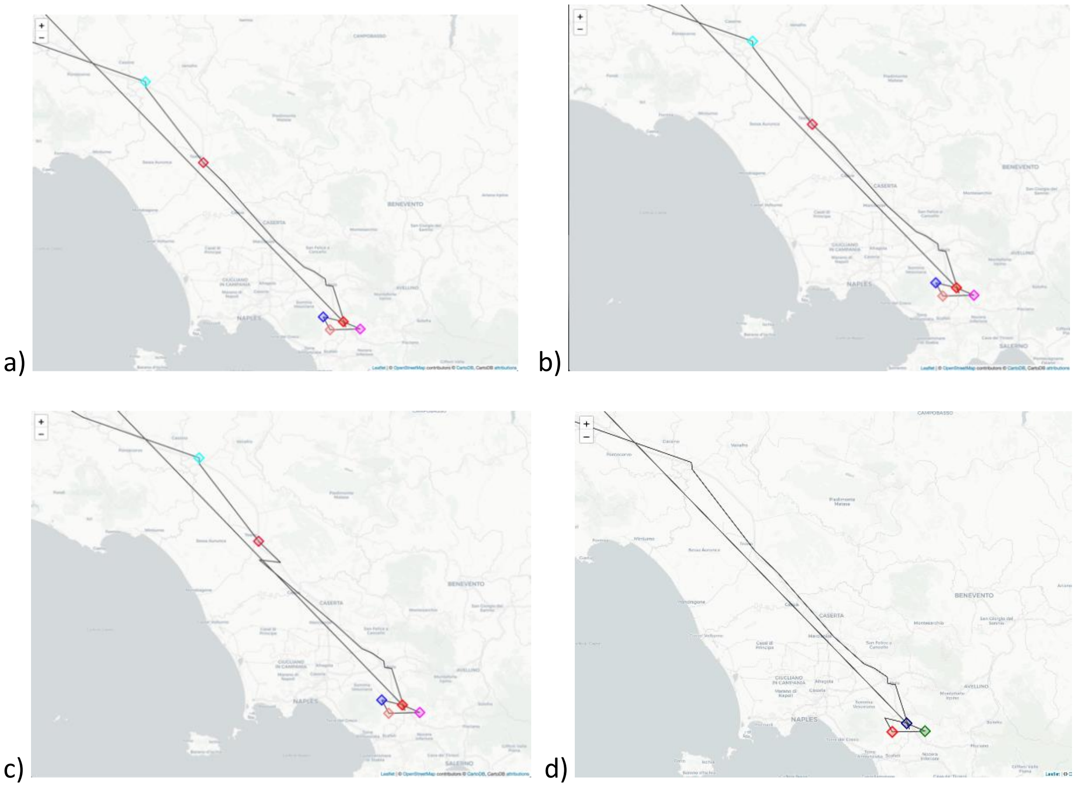


Figure 25 - Zoom in on the trajectory and cluster of stops of a tourist with high radius of gyration (211.33 km) and entropy (7.24): a) original dataset; b) 1 hour sampling; c) 15 minutes day and 45 minutes night sampling; d) CDR mimicking 1 hour uniform completeness.

However, by zooming in on a part of the previous trajectories, it can be observed through figure 25 that the trajectory does in fact change with the change of sampling methods and, for the CDR sampling the position of the clusters differs. This is the most natural conclusion since it had been previously found that the higher the radius of gyration (which is positively correlated to entropy), the higher the error tends to be. Moreover, the change in clusters of stops in the CDR sampling may be due to, once again, the CDR sampling issue. Nonetheless, the sampled and completed sub-trajectory that seems to be the most similar to the original sub-trajectory is the uniform sampling at a 1 hour even though this would have been expected to be the uniform sampling at a 15 minutes rate during the day and 45 minutes during the night. This happens because, even though the overall Mean Distance Error of the former is bigger than the latter (8.19 km Vs 5.89 km) it might still be the case that a given individual contradicts these averages, especially considering that both of these approaches have a standard deviation of 9.09 km and 10.31 km, respectively. Moreover, recall that the radius of gyration had a positive relationship with the error. This explains the fact that differences in the estimated trajectories can be observed in figure 25 for the individual with high radius of gyration and not in figure 22 for the tourist with low radius of gyration.

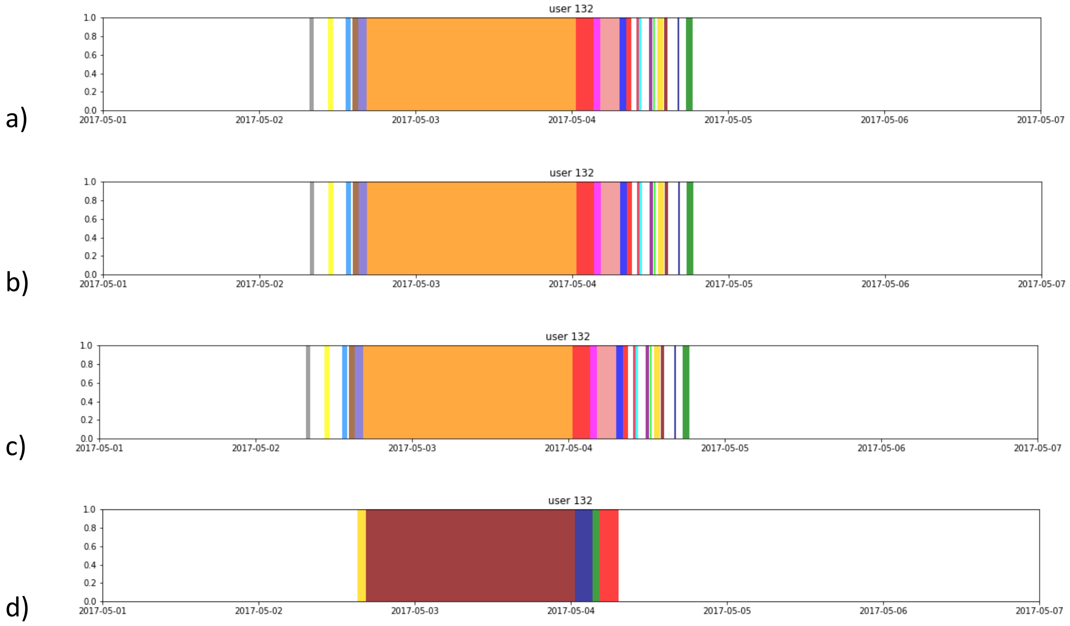


Figure 26 - Tourist diary of a tourist with high radius of gyration and entropy: a) original dataset; b) 1 hour sampling; c) 15 minutes day and 45 minutes night sampling; d) CDR mimicking 1 hour uniform completeness.

Once more, as it can be seen in figure 26, very similar results are obtained for the tourist's diary, with the exception of the CDR sampling method that demonstrates different clusters of stops as it had been previously observed and explained. This only occurred for the CDR sampling method, which can be, once again, due to a limitation of the method itself.

Lastly, it is important to mention the fact that only the best interpolation method was considered in these analyses, which is the method that minimizes the Mean Distance Error (linear interpolation). In fact, the Mean Distance Error of the uniform sampling with 1 hour sample rate, uniform sampling with 15 minutes sample rate during the day and 45 minutes sample rate during the nighttime and the CDR mimicking 1 hour completeness were of, respectively, 8.19, 5.89 and 12.69 km. If it is taken into consideration that the average distance between closest towers is 1.50 km, then these correspond to, on average, making a mistake of 5, 4 and 8 towers respectively.

4.5.2. K-means clustering

As it has been stated beforehand, the k-means algorithm was run in the following standardized features: Entropy, Total Travel Distance, Radius of Gyration, Lifetime in Hours, Number of Unique Towers, First Latitude, First Longitude, Last Latitude, Last Longitude, and Average Latitude. The k-means algorithm was run for 5 clusters with a set of initialization points (*Appendix C*) and the t-SNE was performed for dimensionality reduction in order to be able to plot the results in a 2D graph. The results of these clusters are shown below in figure 27.

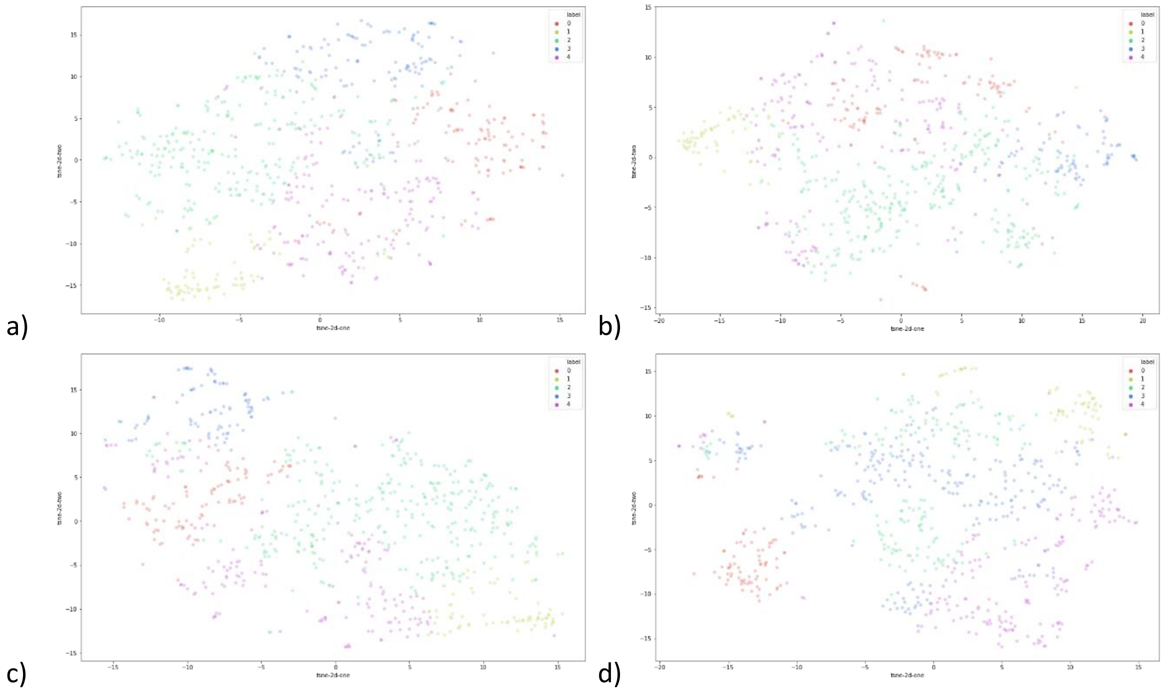


Figure 27 - K-means clusters: a) Full Dataset; b) 1 hour; c) 15 minutes day and 45 minutes night; d) CDR mimicking 1 hour uniform completeness.

Due to the fact that the t-SNE fixes only the distances and not the actual points, it is hard to compare all clusters just by observing them. Nonetheless, it can be seen that the k-means algorithm was not optimized as the clusters are not well defined. This is a reflection of the fact that the initialization points were somewhat arbitrary and not optimized over several different runs. Recall that the k-means is typically initialized several times with different starting seeds to understand which cluster configuration reappears most often. However, for comparability purposes, the same initialization points were set for all approaches.

Despite this, and because this study does not focus on the accuracy of the clustering algorithms but on the changes in data granularity, the % of tourists that were assigned to a different cluster when compared to the 'ground truth' (assumed to be the results of the clusters in the full dataset) was calculated. These percentages were of: 19.31%, 25.60% and 60.55% for the linear interpolation 1 hour uniform, 15 minutes during the day and 45 minutes during the night and CDR mimicking 1 hour uniform approaches, respectively.

The magnitude of these results will be discarded precisely do to the fact that they might be influenced by the performance of the algorithm. Instead, the differences between the underlying datasets are looked into. The average Mean Distance Error of the 15 minutes during the day and 45 minutes during the nighttime uniform and 1 hour uniform are relatively similar (8.76 km and 6.32 km, respectively for the linear interpolation, corresponding to an average error of 6 and 4 towers, respectively, considering average distance between neighboring towers is 1.5 km). As such, it is no surprise that the differences in the percentage of tourists that were assigned to a different cluster between these two approaches are relatively small (6.29%). The fact that the lowest percentage is associated to the 1 hour uniform sampling rate and not to the 15 minutes during the day and 45 minutes during the night is not surprising, since the standard deviation of the tourist's Mean Distance Error is relatively high for both cases (10.16 km and 12.07 km for the 1 hour and the 15 minutes during the day and 45 minutes during the night, respectively). Setting more initialization points and running the algorithm multiple times to check which configurations were the most frequent might be a solution to this occurrence.

The percentage of tourists that were assigned to a different cluster for the CDR mimicking 1 hour uniform is considerably higher than the remaining approaches. This result is expected as the CDR sampling approach has proven to be the one to perform the worse (Mean Distance Error of 13.53 km for the CDR mimicking 1 hour uniform), which may be due to the data sparsity or to the method implemented of sampling on the overall dataset and not per tourist.

Lastly, the sensitivity of the k-means to the changes in the datasets would also be expected since 9 of the 10 features used depends directly on the tower estimation. These features were the: Radius of Gyration, Total Travel Distance, Entropy, Number of Unique Towers, First Latitude, First Longitude, Last Latitude, Last Longitude, Average Latitude (standardized).

5. DISCUSSION

The results presented in the previous section, indicate that, contrary to what was expected for the non-tourist population (Hoteit et al., 2014), the cubic interpolation method was not the method that minimized the Mean Distance Error for the tourist population. In fact, it was the method that performed the worse. This could have been a product of the fact that the points in the dataset that is being used are not equidistant due to the data collection issue, leading the cubic interpolation to be affected as the derivatives are taken at the edges. Nonetheless, it remains a strong possibility that these changes are triggered by significant differences in the movement of the tourist population when compared to the non-tourist population. Overall, the linear interpolation outperformed the remaining approaches in the analyzed sample, although, for many of the higher granularity sample rates, the last seen approach provided estimations with Mean Distance Errors close to the linear approach.

Moreover, it was found that, as it would be expectable, the higher the completion rate of the sampling approach the lower the Mean Distance Error. However, for the uniform sampling with different rates during the day and during the night, changes in the daytime rate caused more impact on the final Mean Distance Error than changes in the nighttime rate. Furthermore, with few exceptions, it was observed that there is at least one uniform sampling with different rates during the day and during the night that offers a lower Mean Distance Error for a similar level of completeness of other sampling approaches, as it can be observed in figure 15. These findings are aligned with the nighttime stability assumption. Furthermore, the CDR sampling proved to be the worst from all sampling methods, which may be linked with the fact that there is a high dispersion of the number of records sampled per hour per tourist, caused by the CDR sampling issue.

It has also been found that, as it would be expectable, the higher the time granularity of the data, the lower the Mean Distance Error tends to become. Additionally, a positive relationship between span of movement (radius of gyration and total distance travelled) metrics with the Mean Distance Error was verified, alongside with a negative relationship between the lifetime of a tourist and the Mean Distance Error have. This proves that the effectiveness of the sampling and completion process depends on the characteristics of the tourists themselves.

Finally, it was shown through the trajectory and stops analysis that changes in the Mean Distance Error has bigger impacts when looking into the trajectories of an individual with a high radius of gyration (that has a positive relationship with error) at a regional level. Therefore, depending on the analysis a researcher wants to make, the optimization of the method that provides lower estimate errors may be more or less critical. Namely, if a researcher's study needs to incorporate a lot of detail (e.g. analyzing how tourists visit different Points of Interest within a given city), than the identification of a sampling method that minimizes errors is important. On the other hand, if the researcher is performing a study at a macro-level (e.g. clustering of different tourists' trajectories at a country level) then this minimization is not as critical. As for k-means clustering of the tourists according to mobility metrics, it was shown that changes in the dataset can impact the results significantly. This reinforces the idea that the choice of granularity level and/or completion method depends on the analysis that the researcher wants to conduct.

5.1. IMPLICATIONS ON PRACTICE

As mentioned above, the minimization of the error when reconstructing a tourist's trajectory is of especial interest for tourism management as it benefits from having as much information as possible by incurring in as few costs as possible. Thus, researchers applying statistical analysis or machine learning algorithms that need to handle mobile positioning data sparsity, as well as tourism management organizations (i.e. city planning offices, governmental agencies for the promotion of national tourism, tourism marketing organizations) are beneficiaries from having the correct trade-off between accuracy and costs.

In order to make this decision, firstly, the organization needs to understand the type of data that it needs for the problem at hand since, as Lin et al. (2018) have pointed out, there are several data types that could be used to explore tourism movement with different benefits and challenges each. This decision will have to factor in, not only the requirements of the analyses that need to be undertaken, but also the feasibility of obtaining a certain data type. In fact, Lin et al. (2018) pointed out that finding the right partnerships to obtain mobile positioning data, for example, can prove to be a difficult task. After having accessed the utility and feasibility of procuring a certain data type, if the organization decides to use CDR's or mobile positioning data, some measure to deal with its sparsity will need to be implemented.

It is suggested that the organizations use the results of this study as a tool when choosing, if possible, the right granularity level of the data and choosing the methods to reconstruct the tourist's trajectories. As such, if the organization requires an analysis with low level of error it can choose, for example, uniform rates of 30 minutes during the day and 1 hour and 30 minutes during the night, whereas, if it allows for larger errors it can choose, for example, uniform rates of 2 hours during the day and 4 hours during the night. When choosing the data granularity, the organization must decide on the level of inaccuracy they will allow in their model, knowing that mobile positioning data is inaccurate in nature as it provides the location of the antenna the tourist connected to and not the actual geographical location. As it has been explored before, this decision will depend on the level at which the analysis the researchers want to conduct operates on (macro or micro-level). The decision on the desired data granularity must take into account the tradeoff between storage costs and the application of one of the proposed completion methods to a less granular dataset, returning less accurate positions, on average.

All in all, using the completion approaches results as a guiding tool to reconstruct full tourist trajectories will allow for a better approximation of the estimates to reality and will enable better decision-making. This could impact touristic destinations and attractions marketing, city planning and administration, attraction's movement design and administrating, among others.

5.2. IMPLICATIONS ON RESEARCH

There are several studies in tourism management that have used CDR's to understand tourist individual mobility. The simulations performed in this study suggest that using positioning data can be an alternative that does not necessarily need to translate into higher storage costs. The type of study the researchers wish to conduct will highly influence the type of data that should be used. Nonetheless, when interpolating missing data in the dataset for both the CDR and the positioning data cases, researchers must not only take into consideration the type of study they are conducting as it has been stated above, but also the characteristics of the population at hand. Evaluating the mobility metrics for a given tourist population is essential to understand whether higher granularity data should be collected.

Furthermore, it is important to keep distinguishing literature of mobility for the general population and literature on tourist mobility. Tourists tend to have a much shorter lifetime, higher entropy in their movements and travel in a higher radius of gyration. The latter two characteristics translate in a lack of redundancy for tourist movement. In addition, despite the fact that this study considered the nighttime stability assumption to hold, this may not be true for all cases. In fact, 42% of the tourists in this study have a higher radius of gyration during the night than during the day. The existence of a high concentration of night travelers will greatly impact the veracity of this assumption, and as such, caution must be taken when using it.

5.3. TRADEOFF BETWEEN ACCURACY AND COST

The collection of highly granular positioning data instead of the CDR's can be very costly and become unfeasible for the researchers. As it has been stated before, this study consists on a dataset of IP probe data of 277,093 devices over the course of a week, which occupies 3.12 Gigabytes of memory. As such, collecting data of 3 million tourists per week over the course of 5 years corresponds to approximately 8.83 Terabytes of memory, if we assume that data requirements scale linearly.

In order to understand how the different granularity levels of data collection affect the data storage costs, the sampled datasets of the uniform sampling approach were stored and their memory requirements analyzed, without the missing positions being completed. In fact, for the uniform sampling, rates of 15 minutes, 30 minutes, 45 minutes, 1 hour, 1 hour 30 minutes, 2 hours, 3 hours and 4 hours corresponded to storage requirements of 7.3 Megabytes, 5.1 Megabytes, 4.1 Megabytes, 3.5 Megabytes, 2.5 Megabytes, 1.9 Megabytes, 1.4 Megabytes and 1.1 Megabytes, respectively. These storage requirements were taken for the set of a 1,000 tourists and not the full dataset.

According to Nasuni. (2016), the average cost of storing 1 Terabyte of data per year is $3,351, with additional $2,000 needed for local backup and another $2,000 for backup at another site, as three copies are usually required. The 8.83 Terabytes of memory needed to store 3 million tourists over 5 years represent a requirement of 1.76 Terabytes per year. If we assume the storage costs of $7,351 per Terabyte, this translates to $12,948 per year, which over the course of 5 years represents a total cost of $64,738. By lowering the storage requirements to, for instance, 3.5 Megabytes per year for 1,000 tourists (corresponding to the 1 hour uniform sampling), we are lowering the requirements to store a year's worth of data on 3 million tourists to half a Terabyte per year. Hence, the cost over the course of 5 years is reduced to $20,123 ($4,024 per year).

This is a significant reduction, corresponding to only 31% of the initial cost of storage and can determine whether or not a research piece has enough funding to continue. Thusly, it is of the most importance to store as little data as possible while still getting the needed level of accuracy that will not hinder the results of the desired analyses. Hence, the need for the results presented on this study on the tradeoff between accuracy and completeness costs, which evidently translate to storage costs. The calculations of the cost and data requirement estimates provided in this section can be found in *Appendix H.*

5.4. LIMITATIONS AND FUTURE WORK

The use of the available dataset and the proposed methodology to compare the performance of multiple sampling and completion methods in the reconstruction of sparse tourist trajectories has some limitations that have been touched on throughout the study. These limitations are summarized and further explored in this section.

Firstly, the dataset was stored in a format that reduces redundant data. The data was collected once per minute if the tourist performs a telecommunication's transaction and once per hour if the tourist does not as it can be assumed he/she remained in the same place. Otherwise, a Location Area Update (LAU) would have been triggered. This means that if a tourist stays in the same location from 13 to 15h, then he/she will have a data point for that location only for the 13:00, 14:00 and 15:00 timestamps when he/she should have data points for every minute during those two hours. It was a technical challenge to perform uniform sampling despite the fact that data is stored such that data points have uneven duration. To overcome this, weighing was used during sampling and error calculation. However, this technical approach also brought upon some limitations, such as the spillover issue.

The spillover issue consists in the fact that, for a given time interval, the data points sampled were only those whose beginning belonged to that same time interval. In other words, suppose that a tourist is recorded at 13:10 and at 13:25 and that we are sampling at a 15 minute rate. In this case, the data point correspondent to the 13:10 timestamp would only be sampled for the 15 minute interval of [13:00; 13:15[even though the location of the tourist from 13:10 to 13:25 is known. This means that, even though the location corresponding to timestamp 13:10 is being sampled at a higher frequency in the first 15 minute period due to the use of weights, it is being ignored in the second. Future work should address this issue by taking into consideration that some positions of an individual for a given time interval can be recorded in data points in the previous time interval.

Another limitation that has emerged from data collection issue has been the roaming issue. This issue occurs due to the fact that tourists may connect to different networks according to the roaming agreements in place. This means that some tourists have gaps in their records larger than 1 hour, which is not a result of them departing and re-entering Italy multiple times. Hence, if a tourist has gaps in their records on the dataset, it is impossible to know whether they stayed stationary (if the gap is smaller than 1 hour), momentarily left the network due to telecommunication networks' roaming agreements, or if it was produced by an error in the system that collects the data. It is important to mention, nonetheless, that this limitation would occur regardless of the data collection issue as tourist's cellphones may leave a certain network at any time due to the roaming agreements.

Moreover, in order to compare results from the uniform sampling and the uniform sampling with different rates during the day and during the night, the same sampling rates were used in both approaches. However, the number of daytime hours or of nighttime hours is not always divisible by the proposed sampling rate. For example, the daytime period (9 am to 6 pm, inclusively) is not divisible by segments of 4 hours each, which provides the emergence of a class of only 2 hours instead of 4, as it has been previously explained. This limitation has been previously referred to as the day and nighttime sampling issue. Future work should address this limitation by building daytime and nighttime sampling rates compatible with the durations of both periods, even if for some cases comparability is lost.

In addition, the CDR sampling issue is a limitation that arose due to the fact that, for simplicity purposes, the sampling method applied to mimic CDR's was performed in the whole dataset and not on a per tourist level. This led to high variations of the number of records sampled per tourist per hour, as it has been previously demonstrated, which hindered the CDR results. Future work should address this issue by performing a CDR mimicking methodology on a per individual basis. Moreover, even though the CDR dataset used to compute the empirical distribution of CDR's was from Italy, it was not specifically for the tourist population. Future work should address this issue by looking at the distribution of CDR's specifically for the tourist population.

Furthermore, for simplicity purposes, the latitude and the longitude were interpolated separately. Some future work is recommended such that the latitudes and the longitudes are interpolated together. Their results should be compared to the ones stemming from a methodology such as the one applied in this study in order to understand if there are significant differences in using a more complex versus a simpler approach when interpolating the positions of tourists.

Additionally, the subsample that was used in order to run the sampling and completion pipeline 20 times was constituted of only 1,000 tourists when the initial set had 277,093 tourists, which limited some of the results. In the future, a higher number of tourists and a higher number of sampling and completion runs should be considered.

Finally, although higher sampling rate levels provided relatively small location estimate errors, future work should develop more complex approaches to be applied to interpolate tourist locations. For instance, Hoteit et al. (2014) and Chen et al. (2018) have been able to successfully implement more complex approaches to estimate the antenna location of an individual in a dataset of sparse mobile positioning records. Nonetheless, caution is advisable in applying their methods directly on a dataset of tourists since it has been proven in this study that tourists not only have a much shorted lifetime than a typical resident, but they also have considerably higher levels of entropy and radius of gyration. For this reason, assumptions made in their approaches (i.e., redundancy in Chen et al., 2018) cannot be expected to hold as strongly for a tourist population.

6. CONCLUSIONS

Understanding patterns in tourist movement is essential for tourism management. This can be achieved using several data types, for instance GPS and mobile positioning data (Li et al., 2018). This thesis studies the tradeoff between time granularity (i.e. having a record every minute VS every hour) and the quality of the reconstructed trajectories. To answer the research question: *How much data is needed to learn tourist movements?,* the dataset of a week's worth of signaling data was sampled uniformly, with different rates during the day and during the night, and to mimic CDR's and reconstructed by using last seen, linear and cubic interpolations.

This study contributes to the existing literature by confirming major differences between the tourist population and the non-tourist population, which annuls the veracity of assumptions, such as redundancy, previously set for the interpolation of non-tourist trajectories. Furthermore, this study concluded that there is a positive relationship between the measures of the span of movement of a tourist (radius of gyration and total travel distance) and the estimate errors, while there is a negative relationship between the tourist's lifetime and error. As such, researchers need to take into account the characteristics of the population at hand when performing interpolation methods. Moreover, the linear interpolation outperformed all other completion methods in this study, while the uniform sampling with different rates during the day than during the night proved to better minimize the estimate error than other sampling methods.

Furthermore, it has been concluded that the level of precision at picking the exact method that minimizes the estimate errors varies with the type of study the researcher wishes to perform. The trajectory and stops analysis led to the conclusion that higher attention is needed when looking at small regions than when making a country level analysis, and the k-means clustering results led to the conclusion that profiling tourists on metrics that relate to the estimated positions is highly dependent on the temporal granularity chosen. Nonetheless, the amount of data stored should be as small as possible to still provide accurate results, as it has also been proven that choosing a dataset with a lower granularity level can provide significant changes in data storage costs, which can in turn determine whether or not a researcher has enough funding to continue their project.

7. BIBLIOGRAPHY

Ahas, R., Aasa, A., Mark, Ü., Pae, T., & Kull, A. (2007b). Seasonal tourism spaces in Estonia: Case study with mobile positioning data. *Tourism management, 28*(3), 898-910. doi.org/10.1016/j.tourman.2006.05.010

Ahas, R., Aasa, A., Roose, A., Mark, Ü., & Silm, S. (2008). Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. *Tourism Management, 3*(29), 469-486. doi.org/10.1016/j.tourman.2007.05.014

Ahas, R., Aasa, A., Silm, S., & Margus, T. (2010). Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area: Case study with mobile positioning data. *Transportation Research Part C: Emerging Technologies, 1*(18), 45-54. doi.org/10.1016/j.trc.2009.04.011

Ahas, R., Aasa, A., Silm, S., & Tiru, M. (2007a). Mobile positioning data in tourism studies and monitoring: case study in Tartu, Estonia. In: M. Sigala, L. Mich, J. Murphy (Eds.), *Information and communication technologies in tourism 2007*, (pp. 119-128). Springer, Vienna.

Ahas, R., Armoogum, J., Esko, S., Ilves, M., Karus, E., Madre, J.L., . . . Tiru, M. (2015). *Eurostat feasibility study on the use of mobile positioning data for tourism statistics, Report.*

Ahas, R., Saluveer, E., Silm, S., & Järv, O. (2009). Modelling Home and Work Locations of Populations Using Passive Mobile Positioning Data. In G. Gartner, W. Cartwright, & M. P. Peterson, *Location based services and TeleCartography II* (pp. 301-315). Berlin, Heidelberg: Springer. doi.org/10.1007/978-3-540-87393-8_18

Arys, B., & Imre, S. (2010). xDR Arrival Distribution. *iPi, 1*, 1.

Barcelo, F., & Sánchez, J. I. (1999). Probability distribution of the inter-arrival time to cellular telephony channels. In *1999 IEEE 49th Vehicular Technology Conference (Cat. No. 99CH36363)*, 1, 762-766. IEEE. doi.org/10.1109/VETEC.1999.778292

Barlacchi, G., De Nadai, M., Larcher, R., Casella, A., Chitic, C., Torrisi, G., ... & Lepri, B. (2015). A multi-source dataset of urban life in the city of Milan and the Province of Trentino. *Scientific data, 2,* 150055. doi.org/10.1038/sdata.2015.55

Belik, V., Geisel, T., & Brockmann, D. (2011). Natural human mobility patterns and spatial spread of infectious diseases. *Physical Review X, 1(1)*, 011001. doi.org/10.1103/PhysRevX.1.011001

Bhattacharjee, D., Rao, A., Shah, C., Shah, M., & Helmy, A. (2004). Empirical modeling on campus-wide pedestrian mobility observations on the USC campus. *IEEE 60th Vehicular Technology Conference*, *4*, pp. 2887-2891. doi.org/10.1109/VETECF.2004.1400588

Bwambale, A., Choudhury, C., & Hess, S. (2019). Modelling long-distance route choice using mobile phone call detail record data: A case study of Senegal. *Transportmetrica A: Transport Science, 15*(2), 1543-1568. Doi.or/10.1080/23249935.2019.1611970

Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., & Ratti, C. (2010). Real-time urban monitoring using cell phones: A case study in Rome. *IEEE Transactions on Intelligent Transportation Systems, 12*(1), 141-151. doi.org/10.1109/TITS.2010.2074196

Chen, C., Ma, J., Susilo, Y., Liu, Y., & Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation research part C: emerging technologies*, *68*, 285-299. doi.org/10.1016/j.trc.2016.04.005

Chen, G., Hoteit, S., Viana, A., Fiore, M., & Sarraute, C. (2018). Individual Trajectory Reconstruction from Mobile Network Data. *[Techinical Report] RT-0495, INRIA Saclay-lle-de-France*, 1-23.

Chopde, N. R., & Nichat, M. (2013). Landmark based shortest path detection by using A* and Haversine formula. *International Journal of Innovative Research in Computer and Communication Engineering*, *1*(2), 298-302.

Colizza, V., Barrat, A., Barthelemy, M., Valleron, A. J., & Vespignani, A. (2007). Modelling the worldwide spread of pandemic influenza: baseline case and containement interventions. *PLoS medicine, 4(1)*, 13. doi.org/10.1371/journal.pmed.0040013

Dang, T., Sonkoly, B., & Molnár, S. (2004). Fractal analysis and modeling of VoIP traffic. In *11th International Telecommunications Network Strategy and Planning Symposium. NNETWORKS 2004,* 123-130. IEEE. doi.org/10.1109/NETWKS.2004.240805

De Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the Crowd: The privacy bounds of human moblity. *Scientific reports, 3*, 1376. doi.org/10.1038/srep01376

Eagle, N., & Pentland, A. S. (2006). Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, *10*(4), 255-268. doi.org/10.1007/s00779-005-0046-3

Euromonitor International. (2018, September). *Tourism Flows in Italy*. Retrieved from http://www.euromonitor.com/

Fiadino, P., Valerio, D., Ricciato, F., & Hummel, K. (2012). Steps towards the extraction of vehicular mobility patterns from 3G signalling data. In *International Workshop on Traffic Monitoring and Analysis* (pp. 66-80). Berlin, Heidelberg: Pringer. doi.org/10.1007/978-3-642-28534-9_7

Girardin, F., Calabrese, F., Dal Fiore, F., Ratti, C., & Blat, J. (2008). Digital footprinting: Uncovering tourists with user-generated content. *IEEE Pervasive computing, 4*(7), 36-43. doi.org/ 10.1109/MPRV.2008.71

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. (2008). Understanding individual human mobility patterns. *nature, 453(7196)*, 779. doi/org/10.1038/nature06958

Horner, M. W., & O'Kelly, M. E. (2001). Embedding economies of scale concepts for hub network design. *Journal of Transport Geography 9(4)*, 255-265. doi.org/10.1016/S0966-6923(01)00019-9

Hoteit, S., Chen, G., Viana, A., & Fiore, M. (2016). Filling the gaps: On the completion of sparse call detail records for mobility analysis. *Proceedings of the Eleventh ACM Workshop on Challenged Networks*, 45-50. doi.org /10.1145/2979683.2979685

Hoteit, S., Secci, S., Sobolevsky, S., Ratti, C., & Pujolle, G. (2014). Estimating human trajectory and hotspots through mobile phone data. *Computer Networks, 64*, 296-307. doi.org/10.1016/j.comnet.2014.02.011

Hu, H., Zhu, X., Hu, Z., Wu, J., & Zhang, X. (2018). Discovering Transportation Mode of Tourists Using Low-Sampling-Rate Trajectory of Cellular Data. *2018 5th International Conference on Systems and Informatics (ICSAI)*, (pp. 1120-1125). doi.org/10.1109/ICSAI.2018.8599469

Huang, H., Cheng, Y., & Weibel, R. (2019). Transport mode detection based on mobile phone network data: A systematic review. *Transportation Research Part C: Emerging Technologies*. doi.org/10.1016/j.trc.2019.02.008

Hufbagel, L., Brockmann, D., & Geisel, T. (2004). Forecast and control of epedemics in a globalized world. *Proceedings of the National Academy of Sciences, 101(42)*, 15124-15129. doi.org/10.1073/pnas.0308344101

Iovan, C., Olteanu-Raimond, A.-M., Couronne, T., & Smoreda, Z. (2013). Moving and calling: Mobile phone data quality measurements and spaciotemporal uncertainty in human mobility studies. In D. Vandenbroucke, B. Bucher, & J. Crompvoets, *Gepgraphic information science at the heart of Europe* (pp. 247-265). Cham: Springer. doi.org/10.1007/978-3-319-00615-4_14

Isaacman, S., Becker, R., Cáceres, R., Martonosi, M., Rowland, J., Varshavsky, A., & Willinger, W. (2012). Human mobility modelling at metropolitan scales. In *Proceedings of the 10th conference on Mobile systems, applications, and* services, 239-252. ACM. doi.org/ 10.1145/2307636.2307659

IVERSEN, V. (n.d.). Modelling voice call interarrival and holding time distributions in mobile networks.

Järv, O., Ahas, R., & Witlox, F. (2014). Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records. *Transportation Research Part C: Emerging Technologies, 38*, 122-135. doi.org/10.1016/j.trc.2013.11.003

Jundee, T., Kunyadoi, C., Apavatjrut, A., Phithakkitnukoon, S., & Smoreda, Z. (2018). Inferring Commuting Flows Using CDR Data: A Case Study of Lisbon, Portugal. *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers* (pp. 1041-1050). ACM. doi.org/10.1145/3267305.3274159

Kitamura, R., Chen, C., Pendyala, R. M., & Narayanan, R. (2000). Micro-simulation of daily activity-travel patterns for travel demand forecasting. *Transportation, 27(1)*, 25-51. doi.org/10.1023/A:1005259324588

Kleinberg, J. (2007). Computing: The wireless epidemic. *Nature, 449(7160)*, 287. doi.org/10.1038/449287a

Leng, Y., Koutsopoulos, H., & Zhao, J. (2018). Profiling presence patterns and segmenting user locations from cell phone data. *Proceedings of ACM Conference (Conference'17).* ACM.

Leng, Y., Noriega, A., Pentland, A. S., Winder, I., Lutz, N., & Alonso, L. (2016). Analysis of tourism dynamics and special events through mobile phone metadata. *arXiv preprint arXiv:1610.08342*.

Li, J., Xu, L., Tang, L., Wang, S., & Li, L. (2018). Big data in tourism research: A literature review. *Tourism Management, 68*, 301-323. doi.org/10.1016/j.tourman.2018.03.009

Li, M., Gao, S., Lu, F., & Zhang, H. (2019). Reconstruction of human movement trajectories from large-scale low-frequency mobile phone data. *Computers, Environment and Urban Systems, 77*, 101346. doi.org/10.1016/j.compenvurbsys.2019.101346

Mamei, M., & Colonna, M. (2018). Analysis of tourist classification from cellular network data. *Journal of Location Based Services, 12*(1), 19-39. doi.org/10.1080/17489725.2018.1463466

Naboulsi, D., Fiore, M., Ribot, S., & Stanica, R. (2016). Large-scale Mobile Traffic Analysis: a Survey. *IEE Communications Survey & Tutorials, 18(1)*, 124-161. doi.org/10.1109/COMST.2015.2491361

Nasuni. (2016). The Hidden Costs of File Storage Revealed in New Infographics. Retrieved November 26, 2019, from https://www.nasuni.com/the-hidden-costs-of-file-storage-revealed-in-new-infographic/

Oliver, V., & Enrique, G. (2014). *Big Data y Turismo: Nuevos Indicadores para la Gestión Turística.* Retrieved from http://www.rocasalvatella.com/sites/default/files/big_data_y_turismo-cast-interactivo.pdf

Olteanu, A. M., Trasarti, R., Couronne, T., Giannotti, F., Nanni, M., Smoreda, Z., & Ziemlicki, C. (2011). GSM data analysis for tourism application. *Proceedings of the 7th International Symposium on Spacial Data Quality (ISSDQ)*.

Pappalardo, L., Barlacchi, G., Simini, F., & Pellungrini, R. (2019). Scikit-mobility: An open-source Python library for human mobility analysis and simultaion. *arXiv preprint arXiv:1907.07062.*

Puura, A., Silm, S., & Ahas, R. (2018). Puura, Anniki, Siiri Silm, and Rein Ahas. The Relationship between Social Networks and Spatial Mobility: A Mobile-Phone-Based Study in Estonia. *Journal of Urban Technology, 25*(2), 7-25. doi.org/10.1080/10630732.2017.1406253

Qin, S., Man, J., Wang, X., Li, C., Dong, H., & Ge, X. (2019). Applying big data analytics to monitor tourist flow for the scenic area operation management. *Discrete Dynamics in Nature and Society 2019*. doi.org/10.1155/2019/8239047

Qin, S. M., Verkasalo, H., Mohtaschemi, M., Hartonen, T., & Alava, M. (2012). Patterns, entropy, and predictability of human mobility and life. *PloS one*, *7*(12), e51353. doi.org/10.1371/journal.pone.0051353

Ranjan, G., Zang, H., Zhang, Z.L., & Bolot, J. (2012). Are call detail records biased for sampling human mobility?. *ACM SIGMOBILE Mobile Computing and Communications Review*, *16*(3), 33-44. doi.org/10.1145/2412096.2412101

Scherrer, L., Tomko, M., Ranacher, P., & Weibel, R. (2018). Travelers or locals? Identifying meaningful sub-populations from human movement data in the absence of ground truth. *EPJ Data Science, 1*(7), 19. doi.org/10.1140/epjds/s13688-018-0147-7

Schneider, C. M., Belik, V., Couronne, T., Smoreda, Z., & Gonzales, M. C. (2013). Unravelling daily human mobility motifs. *Journal of The Royal Society Interface, 10(84)*, 20130246. doi.org/10.1098/rsif.2013.0246

Shoval, N. (2008). Tracking technologies and urban analysis. *Cities, 1*(25), 21-28. doi.org/10.1016/j.cities.2007.07.005

Smoreda, Z., Olteanu-Raimond, A.M., & Couronné, T. (2013). Spatiotemporal data from mobile phones for personal mobility assessment. *Transport survey methods: best practice for decision making*, *41*, 745-767.

Song, C., Qu, Z., & Barabasi, A.L. (2010). Limits of Predictability in Human Mobility. *Science, 327(5968)*, 1018-1021. doi.org/10.1126/science.1177170

Song, L., Kotz, D., Jain, R., & He, X. (2004). Evaluating location predictions wit extensive Wi-Fi mobility data. *IEEE INFOCOM, 2*, 1414-1424. doi.org/ 10.1145/965732.965747

TEDx Talks. (2016, May 12). *How big data will revolutionize tourism management | Marc Cortés | TEDxBarcelonaSalon [Video File].* Retrieved from https://www.youtube.com/watch?v=KENKtzroeK0

Tomasini, M., Bastos-Filho, C., & Menezes, R. (2018). Characterization of Users by Using Hourly and Daily Spatio-Temporal Patterns Extracted from GPS Trajectories. *The Thirty-First International Flairs Conference.*

Vanhoof, M., Reis, F., Ploetz, T., & Smoreda, Z. (2018). Assessing the quality of home detection from mobile phone data for official statistics. *Journal of Official Statistics, 34*(4), 935-960. doi.org/10.2478/jos-2018-0046

Versichele, M., De Groote, L., Bouuaert, M. C., Neutens, T., Moerman, I., & Van de Weghe, N. (2014). Pattern mining in tourist attraction visits through association rule learning on Bluetooth tracking data: A case study of Ghent, Belgium. *Tourism Management*, *44*, 67-81. doi.org/10.1016/j.tourman.2014.02.009

Wang, D. W., Li, L. N., Hu, C., Li, Q., Chen, X., & Huang, P.W. (2019). A Modified Inverse Distance Weighting Method for Interpolation in Open Public Places Based on Wi-Fi Probe Data. *Journal of Advanced Transportation, 2019*. doi.org/10.1155/2019/7602792

Wang, F., & Chen, C. (2018). On data processing required to derive mobility patterns from passively-generated mobile phone data. *Transportation Research Part C: Emerging Technologies*, *87*, 58-74. doi.org/10.1016/j.trc.2017.12.003

Zhao, X., Lu, X., Liu, Y., Lin, J., & An, J. (2018). Tourist movement patterns understanding from the perspective of travel party size using mobile tracking data: A case study of Xi'an, China. *Tourism Management, 69*, 368-383. doi.org/10.1016/j.tourman.2018.06.026

Zhao, Z., Shaw, S. L., Xu, Y., Lu, F., Chen, J., & Yin, L. (2016). Understanding the bias of call detail records in human mobility research. *International Journal of Geographical Information Science*, *30*(9), 1738-1762. doi.org/10.1080/13658816.2015.1137298

Zhao, Z., Shaw, S. L., Yin, L., Fang, Z., Yang, X., Zhang, F., & Wu, S. (2019). The effect of temporal sampling intervals on typical human mobility indicators obtained from mobile phone location data. *International Journal of Geographical Information Science*, *33*(7), 1471-1495. doi.org/10.1080/13658816.2019.1584805

Zheng, W., Huang, X., & Li, Y. (2017). Understanding the touristic mobility using GPS: Where is the next place? *Tourism Management, 59*, 267-280. doi.org/10.1016/j.tourman.2016.08.009

Zufiria, P. J., Pastor-Escuredo, D., Úbeda-Medina, L., Hernandez-Medina, M. A., Barriales-Valbuena, I., Morales, A. J., . . . Hidalgo-Sanchís. (2018). Identifying seasonal mobility profiles from anonymized and aggregated mobile phone data. Application in food security. *PloS one, 13*(4). doi.org/10.1371/journal.pone.0195714

8. APPENDIX

8.1. APPENDIX A – NUMBER OF RECORDS SAMPLED IN THE CDR SAMPLING APPROACH

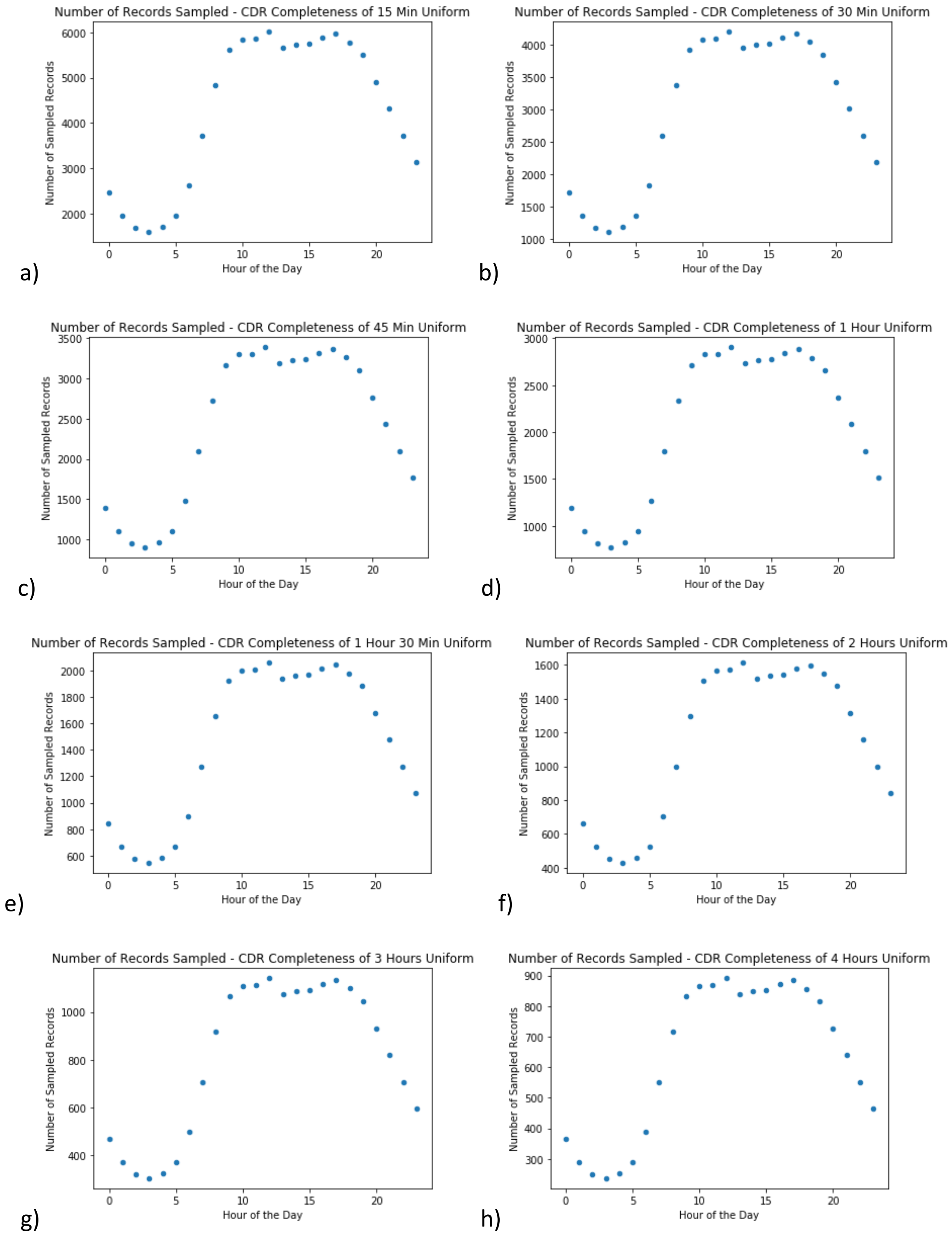


Figure 28 - Number of records sampled per hour: a) CDR with completeness level of 15 min uniform; b) CDR with completeness level of 30 min uniform; c) CDR with completeness level of 45 min uniform; d) CDR with completeness level of 1 hour uniform; e) CDR with completeness level of 1 hour 30 min uniform; f) CDR with completeness level of 2 hours uniform; g) CDR with completeness level of 3 hours uniform; h) CDR with completeness level of 4 hours uniform.

Recall that, in the dataset, 1 record is not equivalent to 1 row since the data was not collected every minute. As the data was collected once per minute if the user makes a call, sends a SMS or starts a mobile data session, and once per hour if the user does not, the number of sampled records is lower than it would be if the dataset was actually per minute. In fact, only 6.6% of the total number of minutes that should have been collected were. Nonetheless, the dataset is treated as if data was collected every minute through the use of weights when sampling that allow for positions that were recorded with lower granularity to be sampled with a higher frequency since they represent positions when users are stable. Furthermore, sampling was made with replacement in order for rows to be able to be sampled more than once as they might correspond to multiple records.

8.2. Appendix B – Number of tourists that can and cannot be used in the cubic completion method

Sampling Method	Sampling Rate	Number of tourists that can be used for cubic interpolation	Number of tourists that cannot be used for cubic interpolation
Uniform	4 h	998	2
Uniform - Daytime	15 min	995	5
Uniform - Daytime	30 min	995	5
Uniform - Daytime	45 min	994	6
Uniform - Daytime	1 h	989	11
Uniform - Daytime	1 h 30 min	976	24
Uniform - Daytime	2 h	975	25
Uniform - Daytime	3 h	931	69
Uniform - Daytime	4 h	789	211
Uniform - Nighttime	15 min	981	17
Uniform - Nighttime	30 min	974	24
Uniform - Nighttime	45 min	972	26
Uniform - Nighttime	1 h	967	31
Uniform - Nighttime	1 h 30 min	954	44
Uniform - Nighttime	2 h	948	50
Uniform - Nighttime	3 h	879	119
Uniform - Nighttime	4 h	900	98

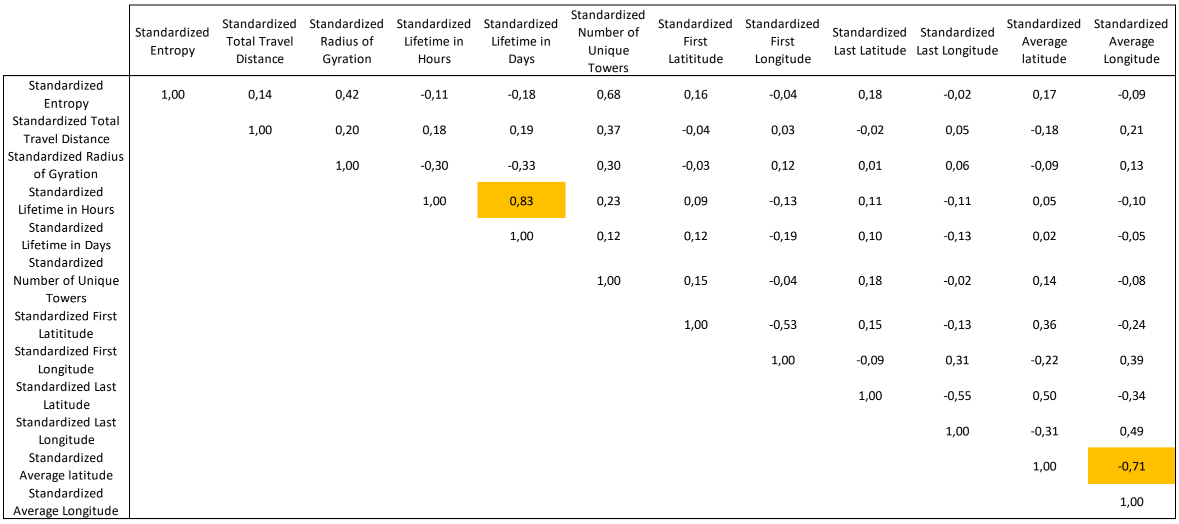
Table 2 - Example of the number of tourists that can and cannot be used in the cubic completion method during an iteration of the cubic interpolation.

While the number of daytime tourists adds up to 1,000, the number of nighttime tourists adds up to 998, which is due to the fact that 2 tourists only appear during the day.

8.3. Appendix C – K-means methodology

Variable	Description
Radius of Gyration (km)	*See section 3.2*
Total Travel Distance (km)	*See section 3.2*
Entropy	*See section 3.2*
Lifetime in Days	Number of days a user is recorded in the dataset
Lifetime in Hours	Number of hours a user is recorded in the dataset
Number of Unique Towers	Number of different antennas each user connected to
First Latitude	First latitude recorded for each user
First Longitude	First longitude recorded for each user
Last Latitude	Last latitude recorded for each user
Last Longitude	Last longitude recorded for each user
Average Latitude	Average of the latitudes of all the antennas each user connected to weighted by the duration of their stay at each antenna
Average Longitude	Average of the longitudes of all the antennas each user connected to weighted by the duration of their stay at each antenna

Table 3 - Features considered for k-means.



	Standardized Entropy	Standardized Total Travel Distance	Standardized Radius of Gyration	Standardized Lifetime in Hours	Standardized Lifetime in Days	Standardized Number of Unique Towers	Standardized First Latititude	Standardized First Longitude	Standardized Last Latitude	Standardized Last Longitude	Standardized Average latitude	Standardized Average Longitude
Standardized Entropy	1,00	0,14	0,42	-0,11	-0,18	0,68	0,16	-0,04	0,18	-0,02	0,17	-0,09
Standardized Total Travel Distance		1,00	0,20	0,18	0,19	0,37	-0,04	0,03	-0,02	0,05	-0,18	0,21
Standardized Radius of Gyration			1,00	-0,30	-0,33	0,30	-0,03	0,12	0,01	0,06	-0,09	0,13
Standardized Lifetime in Hours				1,00	0,83	0,23	0,09	-0,13	0,11	-0,11	0,05	-0,10
Standardized Lifetime in Days					1,00	0,12	0,12	-0,19	0,10	-0,13	0,02	-0,05
Standardized Number of Unique Towers						1,00	0,15	-0,04	0,18	-0,02	0,14	-0,08
Standardized First Latititude							1,00	-0,53	0,15	-0,13	0,36	-0,24
Standardized First Longitude								1,00	-0,09	0,31	-0,22	0,39
Standardized Last Latitude									1,00	-0,55	0,50	-0,34
Standardized Last Longitude										1,00	-0,31	0,49
Standardized Average latitude											1,00	-0,71
Standardized Average Longitude												1,00

Figure 29 - Pearson correlation for the k-means features.

Highlighted in orange are the correlations that exceed the threshold of 0.7 that was set.

Features	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Standardized Radius of Gyration	0.36	-1.10	-0.28	-0.25	0.95
Standardized Total Travel Distance	-0.03	0.06	-0.33	-0.07	0.62
Standardized Entropy	0.79	-0.97	-0.38	0.22	0.33
Standardized Lifetime in Hours	-0.41	2.17	-0.48	-0.46	0.45
Standardized Number of Unique Towers	0.03	-0.43	-0.49	-0.37	1.35
Standardized First Latitude	-1.05	0.02	0.43	-0.59	0.57
Standardized First Longitude	0.79	-0.22	-0.39	0.58	-0.32
Standardized Last Latitude	0.65	0.02	0.08	-1.41	0.50
Standardized Last Longitude	-0.31	-0.27	-0.22	1.18	-0.23
Standardized Average Latitude	0.11	0.05	0.39	-1.21	0.27

Table 4- Features considered for k-means.

8.4. Appendix D – Data exploration of the subsample of 1,000 tourists

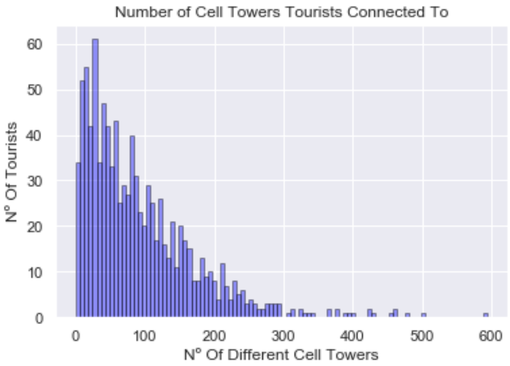


Figure 30 - Number of cell towers tourist connected to during the week of the 1st of May 2017 - 1,000 tourists subsample.

The mean number of cell towers the 1,000 tourists connected to was 93.38 towers with a standard deviation of 81.94. The minimum was 1 tower while the maximum was 593 and the median 73.5.

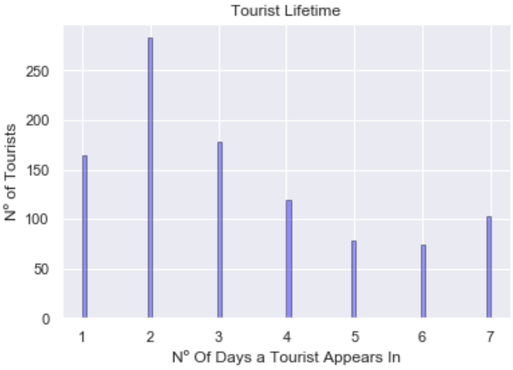


Figure 31 - Tourist lifetime during the week of the 1st of May 2017 - 1,000 tourists subsample.

The mean lifetime of the 1,000 tourists was 3.30 days with a standard deviation of 1.90. The median was 3 days and 25% of the 1,000 tourists stayed up to 2 days.

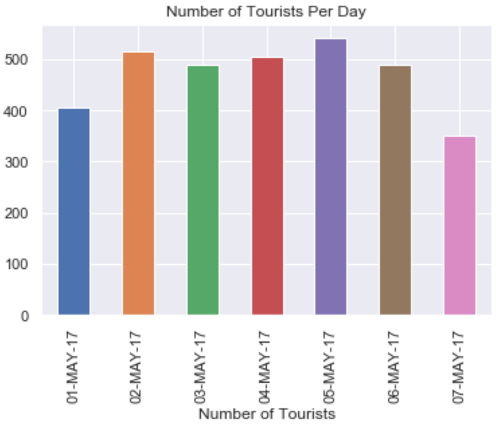


Figure 32 - Number of tourists per day during the week of the 1st of May 2017 - 1,000 tourists subsample.

The mean number of tourists per day in the subsample was 470.86 tourists with a standard deviation of 67.86. The minimum was 350 tourists while the maximum was 541 and the median 489 tourists.

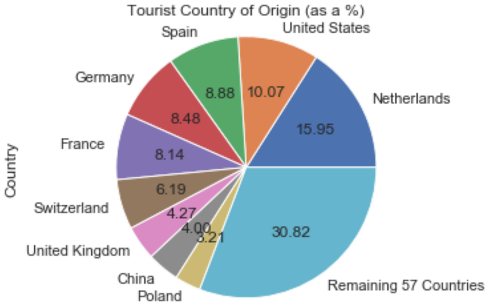


Figure 33 - Tourist country of origin day during the week of the 1st of May 2017 - 1,000 tourists subsample.

The Netherlands, US, Spain, Germany and France were the top nationalities of 1,000 tourists in the subsample.

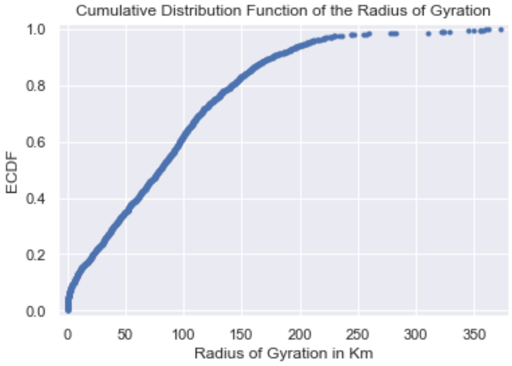


Figure 34 - Cumulative distribution function of the radius of gyration during the week of the 1st of May 2017 - 1,000 tourists subsample.

The mean radius of gyration in the subsample was 87.05 km with a standard deviation of 68.06. 25% of the tourists had a radius of gyration of 31.95 km, while 50% and 75% had a radius of gyration of 79.46 km and 126.40 km, respectively.

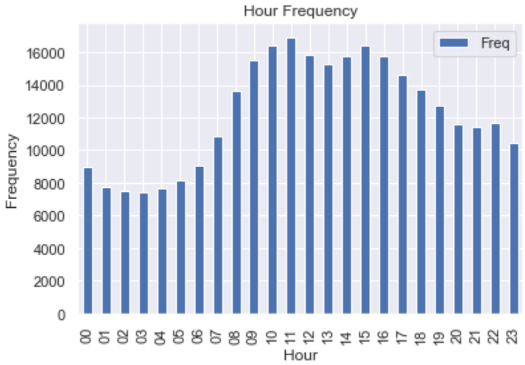


Figure 35 - Hour frequency during the week of the 1st of May 2017 - 1,000 tourists subsample.

Most recorded interactions occur between 9 am and 6 pm, inclusively. In fact, 52.93% of the total number of tourists recorded in any given hour were recorded during this time frame.

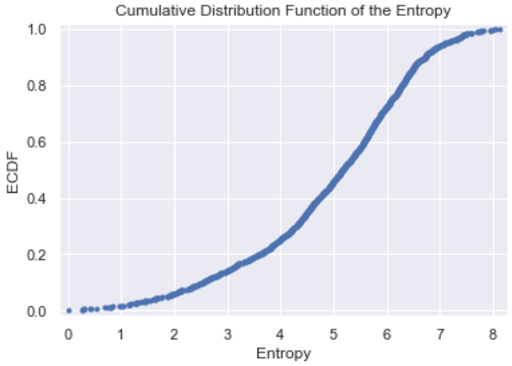


Figure 36 - Cumulative distribution function of the entropy during the week of the 1st of May 2017 - 1,000 tourists subsample.

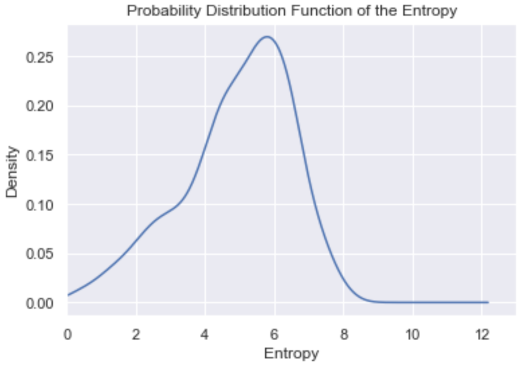


Figure 37 - Probability distribution function of the entropy during the week of the 1st of May 2017 - 1,000 tourists subsample.

The entropy in the subsample was 4.92 with a standard deviation of 1.59. The minimum was 0 (stationary tourists) while the maximum was 8.12 and the median 5.18, respectively.

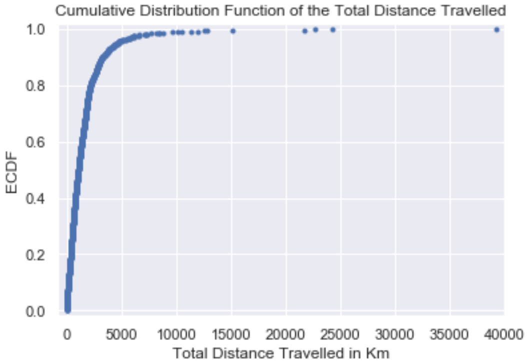


Figure 38 - Cumulative function of the total distance travelled during the week of the 1st of May 2017 - 1,000 tourists subsample.

The total distance travelled of the 1,000 tourists was 1,613.80 km with a standard deviation of 2,355.10. The minimum total distance travelled was 470.16 km while the maximum was 39,206.78 km and the median 1,081.60 km, respectively.

8.5. Appendix E – Mean distance error statistics (in km)

Uniform Sampling

Sampling Rate	Interpolation Method	Mean Error - Mean	Mean Error - Standard Deviation	Mean Error - Median
15 min	Last Seen	9.33	9.94	7.33
30 min	Last Seen	12.27	13.04	9.66
45 min	Last Seen	14.90	14.99	11.61
1 h	Last Seen	17.14	21.65	13.22
1 h 30 min	Last Seen	17.16	16.31	13.82
2 h	Last Seen	19.77	20.46	15.07
3 h	Last Seen	24.18	25.13	18.84
4 h	Last Seen	28.01	27.70	21.22
15 min	Linear	4.92	7.38	3.33
30 min	Linear	6.17	8.52	4.45
45 min	Linear	7.86	11.21	5.61
1 h	Linear	8.76	10.16	6.43
1 h 30 min	Linear	9.25	9.45	7.23
2 h	Linear	10.87	12.00	8.51
3 h	Linear	14.18	16.63	10.95
4 h	Linear	17.28	18.78	13.77
15 min	Cubic	12.99	18.76	7.62
30 min	Cubic	16.62	24.30	9.70
45 min	Cubic	19.41	27.03	11.25
1 h	Cubic	21.41	28.81	12.46
1 h 30 min	Cubic	21.27	29.27	26.97
2 h	Cubic	24.70	33.70	12.94
3 h	Cubic	29.34	41.16	15.97
4 h	Cubic	33.64	45.02	18.86

Table 5 - Uniform sampling mean distance error statistics (in km).

Uniform Sampling with Different Rates During the Day and Nighttime

Mean: Last Seen Interpolation

Day/Night	30 min	45 min	1 h	1 h 30 min	2h	3h	4h
15 min	11.09	12.02	12.47	12.02	13.50	14.16	14.49
30 min		14.35	15.06	14.89	16.33	17.53	18.21
45 min			16.72	16.80	18.17	19.82	20.81
1 h				18.33	19.60	21.68	23.03
1 h 30 min					21.10	23.48	25.01
2 h						24.40	26.44
3 h							29.86

Table 6 - Mean of last seen interpolation's mean distance error for the uniform sampling with different rates during the day and nighttime.

Mean: Linear Interpolation

Day/Night	30 min	45 min	1 h	1 h 30 min	2h	3h	4h
15 min	5.53	6.32	6.35	5.56	5.97	6.94	6.47
30 min		7.22	7.34	6.53	6.91	8.18	7.74
45 min			8.28	7.51	7.79	9.35	8.96
1 h				8.33	8.46	10.31	10.00
1 h 30 min					9.34	11.32	11.12
2 h						11.69	11.52
3 h							13.46

Table 7 - Mean of linear interpolation's mean distance error for the uniform sampling with different rates during the day and nighttime.

Mean: Cubic Interpolation

Day/Night	30 min	45 min	1 h	1 h 30 min	2h	3h	4h
15 min	17.64	19.60	19.43	17.19	19.65	20.50	21.31
30 min		23.34	23.41	21.13	23.76	25.01	26.58
45 min			26.07	23.85	26.47	28.15	30.29
1 h				25.73	28.11	30.00	32.71
1 h 30 min					29.51	31.56	34.29
2 h						34.15	38.14
3 h							42.57

Table 8 - Mean of cubic interpolation's mean distance error for the uniform sampling with different rates during the day and nighttime.

Standard Deviation: Last Seen Interpolation

Day/Night	30 min	45 min	1 h	1 h 30 min	2h	3h	4h
15 min	11.11	11.47	12.01	11.66	13.17	14.08	14.08
30 min		13.40	13.95	13.58	14.87	16.07	16.29
45 min			14.43	14.43	15.99	17.53	17.88
1 h				16.49	17.89	19.78	20.52
1 h 30 min					20.10	22.65	23.46
2 h						23.23	24.51
3 h							28.27

Table 9 - Standard deviation of last seen interpolation's mean distance error for the uniform sampling with different rates during the day and nighttime.

Standard Deviation: Linear Interpolation

Day/Night	30 min	45 min	1 h	1 h 30 min	2h	3h	4h
15 min	9.46	12.07	9.26	6.78	7.65	12.08	8.20
30 min		13.31	10.14	7.13	8.70	12.86	9.25
45 min			11.00	7.95	10.05	14.49	10.65
1 h				9.04	11.05	15.19	11.85
1 h 30 min					11.35	15.98	12.34
2 h						17.67	14.36
3 h							16.76

Table 10 - Standard deviation of linear interpolation's mean distance error for the uniform sampling with different rates during the day and nighttime.

Standard Deviation: Cubic Interpolation

Day/Night	30 min	45 min	1 h	1 h 30 min	2h	3h	4h
15 min	21.41	23.68	22.06	18.72	23.04	26.47	23.48
30 min		28.79	27.29	23.81	28.81	31.95	30.39
45 min			29.88	26.32	26.47	35.18	33.79
1 h				29.08	34.87	37.14	36.87
1 h 30 min					36.52	39.22	38.94
2 h						43.33	44.17
3 h							48.08

Table 11 - Standard deviation of cubic interpolation's mean distance error for the uniform sampling with different rates during the day and nighttime.

Median: Last Seen Interpolation

Day/Night	30 min	45 min	1 h	1 h 30 min	2h	3h	4h
15 min	8.93	9.86	10.22	9.86	11.04	11.44	11.73
30 min		11.75	12.52	12.56	13.59	14.56	15.17
45 min			14.14	14.23	15.17	16.33	17.02
1 h				15.25	16.19	18.10	19.11
1 h 30 min					17.30	19.22	20.41
2 h						19.50	21.21
3 h							23.65

Table 12 - Median of last seen interpolation's mean distance error for the uniform sampling with different rates during the day and nighttime.

Median: Linear Interpolation

Day/Night	30 min	45 min	1 h	1 h 30 min	2h	3h	4h
15 min	3.51	3.81	4.09	3.75	3.74	4.29	6.47
30 min		4.71	5.03	4.88	4.79	5.62	5.44
45 min			5.73	5.55	5.27	6.31	6.18
1 h				6.27	5.78	7.17	7.15
1 h 30 min					6.76	8.17	8.26
2 h						8.26	8.25
3 h							9.86

Table 13 - Median of linear interpolation's mean distance error for the uniform sampling with different rates during the day and nighttime.

Median: Cubic Interpolation

Day/Night	30 min	45 min	1 h	1 h 30 min	2h	3h	4h
15 min	11.41	11.77	12.23	11.69	12.37	13.00	13.97
30 min		14.03	14.56	14.33	14.77	16.14	17.85
45 min			16.45	15.59	16.72	17.97	20.89
1 h				17.12	17.76	19.77	22.29
1 h 30 min					18.55	20.47	23.28
2 h						21.89	25.63
3 h							28.84

Table 14 - Median of cubic interpolation's mean distance error for the uniform sampling with different rates during the day and nighttime.

Uniform Sampling To Mimic CDR’s

Completeness Rate Mimicking Uniform	Interpolation Method	Mean Error - Mean	Mean Error - Standard Deviation	Mean Error - Median
15 min	Last Seen	15.05	16.50	10.60
30 min	Last Seen	17.67	19.26	12.65
45 min	Last Seen	19.30	20.29	13.80
1 h	Last Seen	20.71	21.38	15.24
1 h 30 min	Last Seen	24.82	25.38	18.10
2 h	Last Seen	27.98	27.67	20.70
3 h	Last Seen	33.58	32.34	24.35
4 h	Last Seen	38.50	36.26	28.23
15 min	Linear	8.44	9.76	5.85
30 min	Linear	10.85	12.56	7.39
45 min	Linear	12.18	13.88	8.62
1 h	Linear	13.53	15.39	9.55
1 h 30 min	Linear	17.07	19.07	12.37
2 h	Linear	20.07	21.46	14.89
3 h	Linear	25.37	25.60	19.12
4 h	Linear	30.43	29.88	22.72
15 min	Cubic	17.60	22.72	10.49
30 min	Cubic	21.28	27.63	12.86
45 min	Cubic	23.08	29.09	14.01
1 h	Cubic	24.58	30.41	15.17
1 h 30 min	Cubic	29.20	34.94	18.89
2 h	Cubic	32.97	37.68	22.38
3 h	Cubic	37.70	39.78	26.39
4 h	Cubic	41.07	40.58	30.82

Table 15 - Sampling to mimic CDR’s mean distance error statistics (in km).

The *completeness rate mimicking uniform* column in table 15 represents the completeness rates obtained during the uniform sampling method, which were used for this approach as well. This is in reference to the fact that the number of records sampled during the uniform approach were fixed and used as the number of records to sample to mimic CDR’s, in order for both approaches to be comparable.

Completeness Rate per Sampling Rate

Sampling Rate	Completeness Rate	Sampling Rate	Completeness Rate
15 min	2.3%	30 min day 2 hours night	0.8%
30 min	1.6%	30 min day 3 hours night	0.7%
45 min	1.3%	30 min day 4 hours night	0.7%
1 h	1.1%	45 min day 1 hour night	0.9%
1 h 30 min	0.8%	45 min day 1 hour 30 night	0.8%
2 h	0.6%	45 min day 2 hours night	0.7%
3 h	0.4%	45 min day 3 hours night	0.6%
4 h	0.3%	45 min day 4 hours night	0.6%
15 min day 30 min night	1.4%	1 hour day 1 hour 30 night	0.7%
15 min day 45 min night	1.3%	1 hour day 2 hours night	0.6%
15 min day 1 hour night	1.2%	1 hour day 3 hours night	0.5%
15 min day 1 hour 30 night	1%	1 hour day 4 hours night	0.5%
15 min day 2 hours night	1%	1 hour 30 day 2 hours night	0.6%
15 min day 3 hours night	0.8%	1 hour 30 day 3 hours night	0.4%
15 min day 4 hours night	0.8%	1 hour 30 day 4 hours night	0.4%
30 min day 45 min night	1.1%	2 hours day 3 hours night	0.4%
30 min day 1 hour night	1%	2 hours day 4 hours night	0.4%
30 min day 1 hour 30 night	0.8%	3 hours day 4 hours night	0.4%

Table 16 - Completeness rate per sampling rate.

Recall that the CDR sampling mimics the completion *rate of the uniform* sampling. The completeness rates presented above were rounded to 3 decimal points.

8.6. Appendix F – Number of records sampled per tourist per hour – CDR sampling

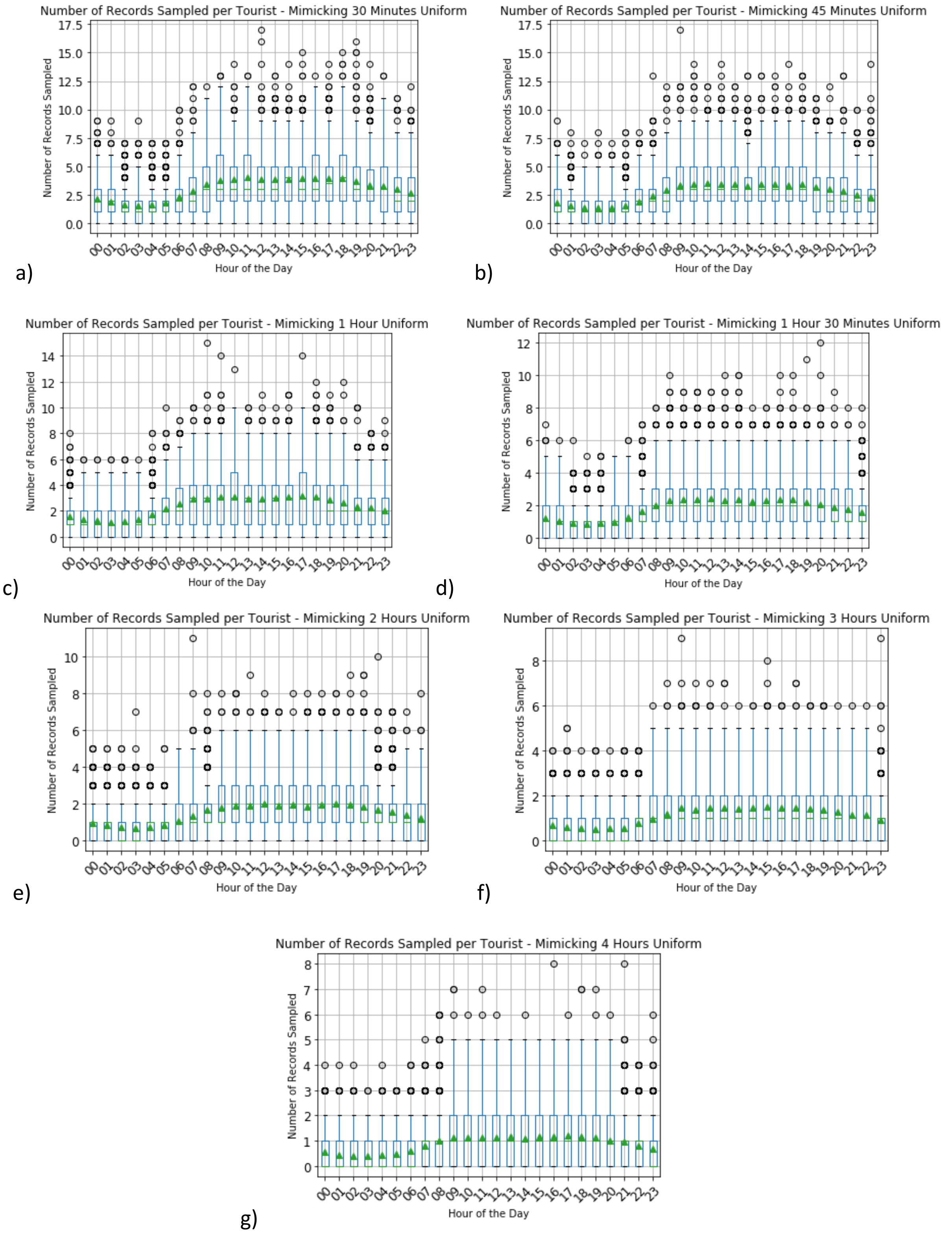


Figure 39 - Number of records sampled per tourist per hour: a) mimicking 30 minutes uniform sampling; b) mimicking 45 minutes uniform sampling; c) mimicking 1 hour uniform sampling; d) mimicking 1 hour 30 minutes uniform sampling; e) mimicking 2 hours uniform sampling; f) mimicking 3 hours uniform sampling; g) mimicking 4 hours uniform sampling.

8.7. APPENDIX G – MEAN DISTANCE ERROR (KM) AND MOBILITY METRICS – LAST SEEN AND CUBIC INTERPOLATION METHODS

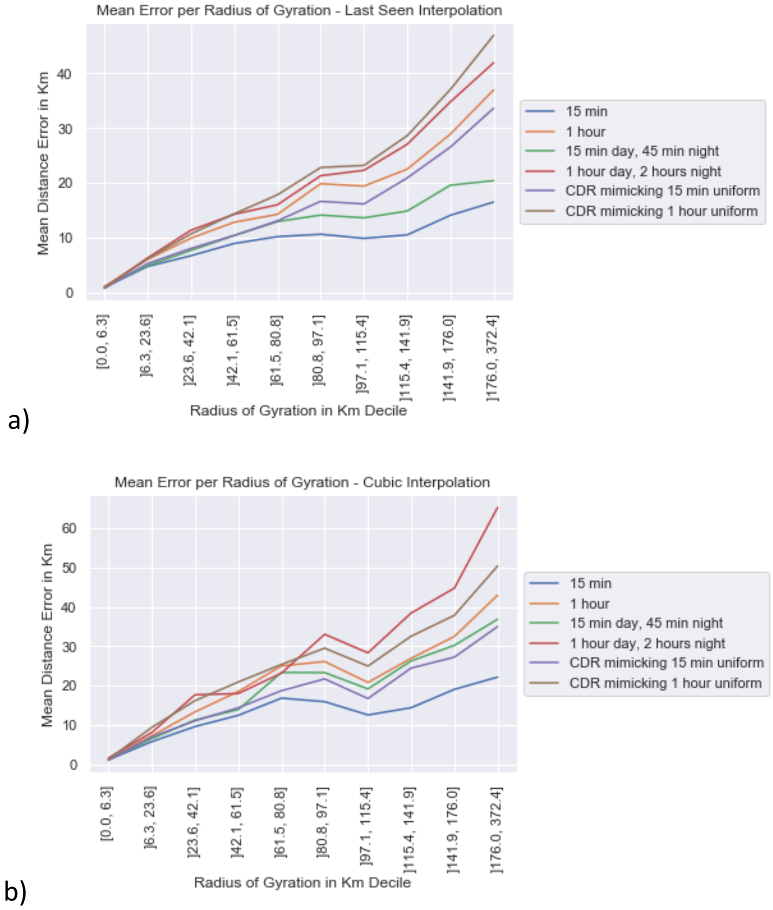


Figure 40 - Mean error versus radius of gyration: a) last seen interpolation; b) cubic interpolation.

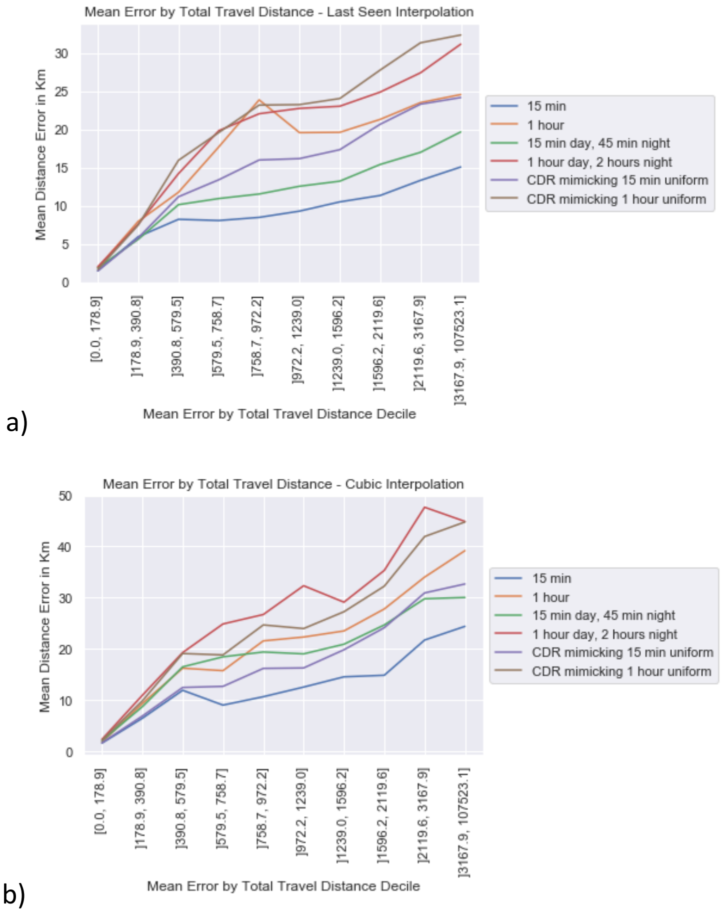


Figure 41 - Mean error versus total travel distance: a) last seen interpolation; b) cubic interpolation.

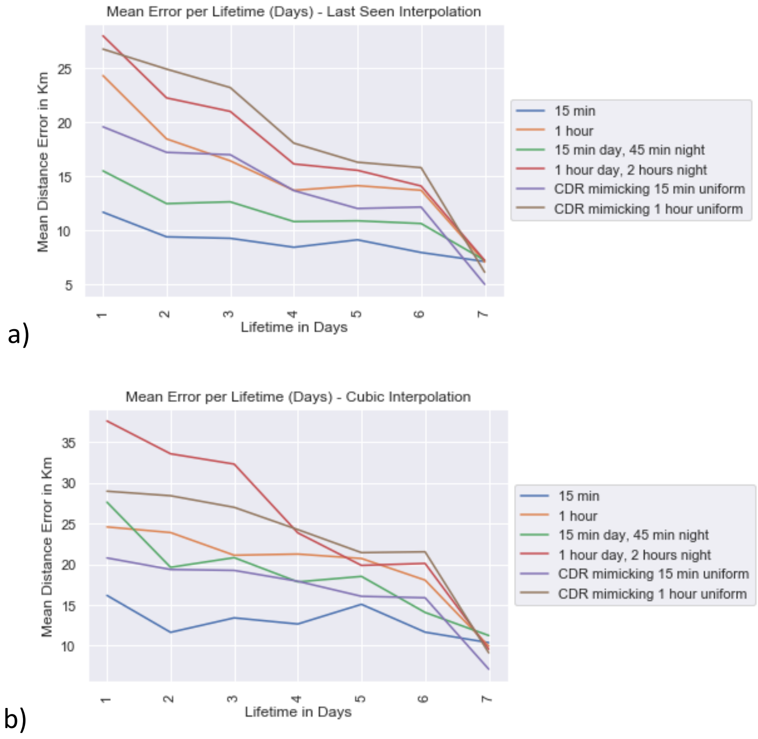


Figure 42 - Mean error versus lifetime in days: a) last seen interpolation; b) cubic interpolation.

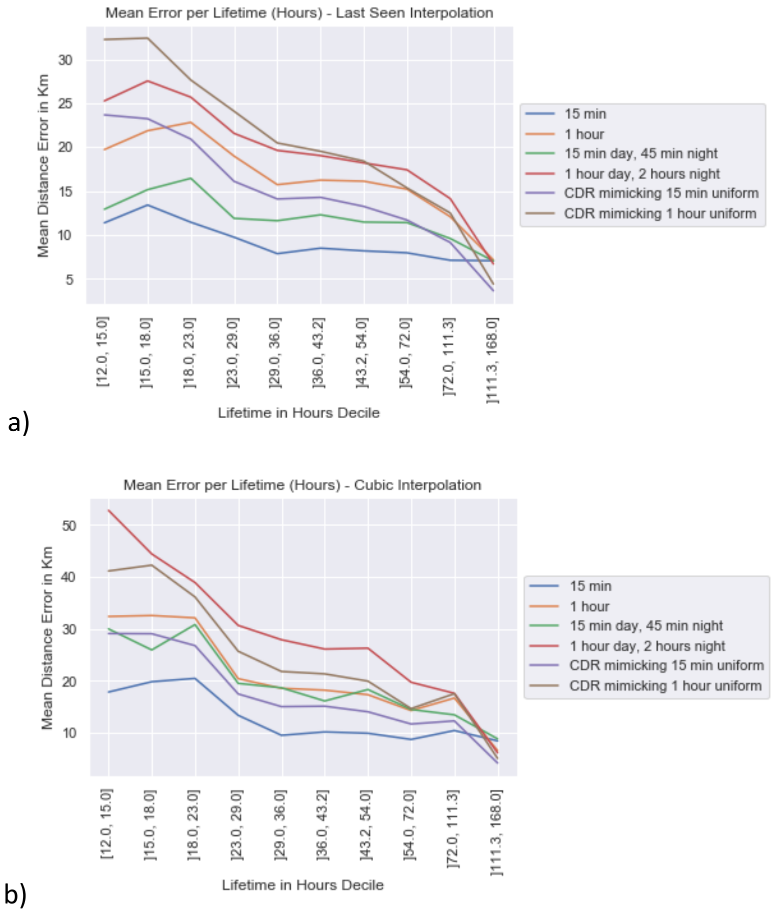
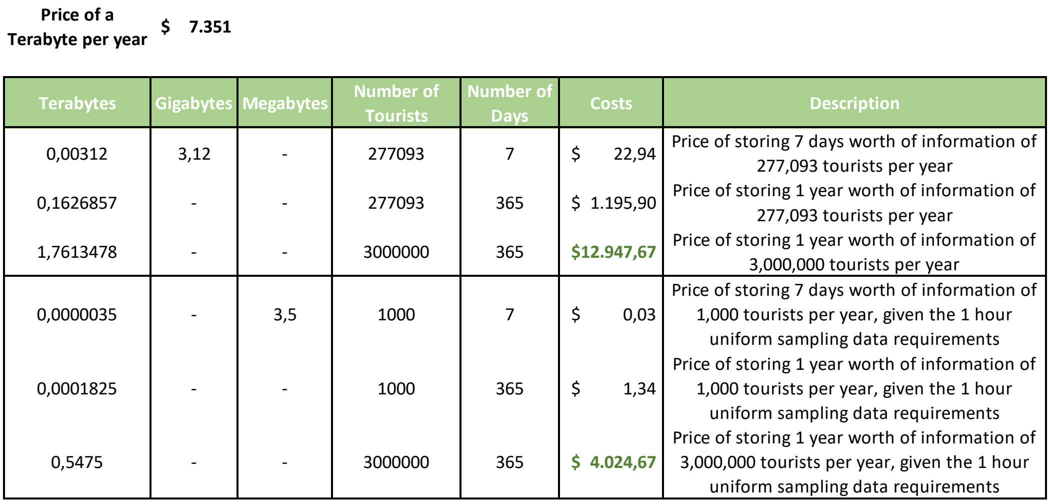


Figure 43 - Mean error versus lifetime in hours: a) last seen interpolation; b) cubic interpolation.

8.8. Appendix H – Data requirements and storage costs estimates



Price of a Terabyte per year **$ 7.351**

Terabytes	Gigabytes	Megabytes	Number of Tourists	Number of Days	Costs	Description
0,00312	3,12	-	277093	7	$ 22,94	Price of storing 7 days worth of information of 277,093 tourists per year
0,1626857	-	-	277093	365	$ 1.195,90	Price of storing 1 year worth of information of 277,093 tourists per year
1,7613478	-	-	3000000	365	**$12.947,67**	Price of storing 1 year worth of information of 3,000,000 tourists per year
0,0000035	-	3,5	1000	7	$ 0,03	Price of storing 7 days worth of information of 1,000 tourists per year, given the 1 hour uniform sampling data requirements
0,0001825	-	-	1000	365	$ 1,34	Price of storing 1 year worth of information of 1,000 tourists per year, given the 1 hour uniform sampling data requirements
0,5475	-	-	3000000	365	**$ 4.024,67**	Price of storing 1 year worth of information of 3,000,000 tourists per year, given the 1 hour uniform sampling data requirements

Figure 44 - Data requirements and storage costs estimates

These estimates assume that the price of a Terabyte provided by Nasuni. (2016) is still accurate at the time of the writing of this thesis, and that the data requirements scale linearly.