Analysing the Tradeoff between Resource Consumption and Information Gain in the Gathering of Geolocation Data Using Smarphones

Thierry Silva Barros[1], **Claudio E. C. Campelo**[1]

[1]Systems and Computing Department
Federal University of Campina Grande – UFCG
Campina Grande – PB – Brazil

thierry.barros@ccc.ufcg.edu.br, campelo@dsc.ufcg.edu.br

Abstract. *Geolocation data have been widely used for the comprehension of various social phenomena. Nowadays, such data are produced on a large scale by people using their smartphones. However, the capture of this kind of data, in a mobile device, can be expensive, consuming device resources. On the other hand, the capture with low frequency may impair the quality and consistency of the information collected. In this context, we conducted a comparative study on the performance across different data collection frequencies to analyse the impact on resource consumption and data quality. Afterwards, an evaluation was performed in order to show the pros and cons of the different capture frequencies and estimate a frequency best suited for different usage scenarios.*

1. Introduction

Geolocation is the identification of a object's geographical location in the real world. This information commonly used to identify an electronic device's physical location. Geolocation data have been used to help comprehend various social phenomena. Nowadays, this kind of data is produced in large scale by people using their smartphones equipped with different types of sensors. Some examples of data generated by smartphones that may contain references to locations include photos, videos and posts in social networks. In addition, several applications retrieve user's location data with a certain frequency, in order to offer targeted products and services based on their locations.

Geolocation data are also used to conduct research with relevant social impact, such as those related to transport policies, public safety, traffic engineering and other topics related to urban planning. Moreover, this kind of data has been exploited to investigate people's trajectories, which express characteristics of human behaviour, enabling different kinds of studies, particularly in large urban centres [Kong et al. 2018]. Furthermore, location-based services has been using this type of data to predict the trajectory of users to recommend products and services based on their destinations or along their routes [Herder et al. 2014].

Although advances in mobile technology have increased the devices' capacity, both in terms of processing capabilities and battery life, these resources are still considered limited. In spite of this, application developers have been largely ignoring the cost for capturing geolocation data in terms of resource consumption. Very frequent geolocation data collection may cause a significant impact on resource consumption, reducing the

devices' performance and providing bad experience for applications' users. On the other hand, low frequency captures may cause loss of crucial data. Consequently, information derived from these data may become inaccurate or uncertain. A deeper understanding of this tradeoff may help developers to minimise resource consumption while maintaining levels information granularity that do not impair their analysis.

For instance, a recommender system that captures geolocated data with 60-second frequency may lose crucial information about the user's locomotion occurred within this time interval, consequently offering services and information which may not be useful, frustrating the customers. On the other hand, an application with high collection frequency can considerably drain the battery of the user's device, or slow it down. Deciding the appropriate collection frequency is not a trivial task, since it is necessary to analyse the impact for the application's user and also the impact on the quality of information that may be derived from the data. Generally, arbitrary values are assigned by developers, without theoretical or experimental foundations.

Thus, in this work, we conducted a comparative study to observe the pros and cons of choosing different frequencies of data collection, in terms of effectiveness and efficiency. For this study, we developed an android application capable of collecting geolocation data along with other sensor data from smartphones. Ten volunteers were recruited to participate in the experiment, over a period of 4 weeks, using the application for 5 consecutive days (weekdays) for each collection frequency. At the end of each day, data about the consumption of smartphones' resources were collected from the volunteers. We then analysed different collection frequencies and compared them based on the consumption of smartphones' resources and the loss of information derived for different application scenarios. In this paper, we also discuss whether the choice of frequency has a significant impact on these two variables and whether specific frequencies appear to be more suitable for certain contexts.

The analysis of the tradeoff between information gain and collection efficiency was guided based on the use of data by geolocation analysis algorithms. The main algorithm used was the Dynamic Time Warping [Vaughan and Gabrys 2016], an algorithm used for comparison and alignment of two time series, which is commonly used in geolocation surveys to compare the similarity between trajectories. We also implemented a variation of the DBSCAN algorithm (Density-based Spatial Clustering of Applications with Noise) [Luo et al. 2017], targeted to identify stop regions in trajectory analysis problems.

The rest of this paper is structured as follows. In Section 2, we discuss the research methodology proposed in this work. Then, in Section 3, we present the obtained results. Finally, Section 4 concludes the paper and points to future work.

2. Methodology

This section describes the methodology adopted to analyse the tradeoff between different geolocalised data collection frequencies and information quality derived from these data.

2.1. Defined indicators

To enable the analysis of the proposed tradeoff, we first defined a set of indicators related to the resources consumption and the quality of produced information. The three

indicators related to resource consumption are:

Battery Consumption: This indicator refers to the power consumption of the geolocation gathering application on the participant's device. Modern smartphones provide information on battery consumption by different applications, that is, the amount of battery spent by a certain application (not skewed by the use of other applications on the smartphone). The high battery consumption is one of the main resistance factors for using applications that capture geolocation data, thus we consider this a crucial indicator to be analysed.

RAM Memory Consumption: This indicator refers to the total memory allocated by location gathering application on the participant's device. High RAM consumption may impact the smartphone performance, decreasing the device's responsiveness and generating user discomfort. For this reason, we also consider this a very relevant aspect to be observed.

Amount of data transmitted: This indicator refers to the amount of data sent by the application to a cloud server over the network. High data transmission rates can directly impact the monetary amount paid by the users with internet packages (specially in non-developed countries), reducing their interest for the app. Hence, this is also a decisive indicator to be investigated.

Apart from the privacy concerns, we believe those indicators represent the main factors that make people avoid using apps that activate the GPS sensor very frequently. Moreover, relevant researches with the aim of tracking people's location have been conducted using considerably reduced databases [El Faouzi et al. 2011, Zheng 2015, Parent et al. 2013], and related issues have been reported regarding the volunteers' engagement.

We also defined two indicators to assess the quality of the information produced from the geolocation data: trajectories inferred from raw data (i.e., routes taken by the users); and stop regions, which are places within the users' trajectories where they stayed for a certain time (these places also include places of origin and destination). The latter is more related to the semantic aspects of the trajectory [Xiang et al. 2016] and is of high interest by both industry and academy, since it can be used to infer other relevant information, such as: types of places; human activities; points of interest; among others. This information, in turn, can be used to perform different kinds of analyses, such as those related to urban mobility patterns and trajectory prediction [Feng and Zhu 2016, Mazimpaka and Timpf 2016, Kong et al. 2018].

Additionally, we defined indicators based on the data collected from other smartphone's sensors: ambient lighting; proximity between the user and the smartphone; screen locked / unlocked; and audio status (i.e., normal, muted or in silence mode). From that data, we could estimate, for example, how accurate would be the information of whether the place the user is located is well lit. This kind of information can be considerably impacted by the granularity of geolocation data, since the light level normally changes as the users change their location.

2.2. Data capture strategy

To carry out this study, we used a mobile application developed in the research laboratory [Barros T. 2019], which can capture both geolocation data and smartphone sensors data. The research was divided into three stages. The first stage consisted in capturing the devices' resource consumption and geolocation data. To carry out these tasks, we count with the participation of 10 volunteers recruited by the researchers. The participants were subjected to a observational study where they installed the application on their device for data collection over a period of 20 days. The types of transportation used by the volunteers to move around the city were vehicles or buses.

Apart from the geolocation data, we collected data about the consumption of smartphones' resources at 4 different data collection frequencies: 15, 30, 60 and 120 seconds. Each frequency was observed for a period of 5 consecutive days, so that we estimate with greater precision the confidence interval of resource consumption of each collection frequency. The data about resource consumption was collected manually, since the application does not have functionalities for collecting consumption data in a automatic way. At the end of each day, the volunteers informed, through a messaging application, the data consumption of each resource.

Before starting the data acquisition phase, all volunteers were trained on how to obtain data from smartphone resources consumption and how to format the daily report. At the end of the first stage, four datasets were produced, one for each frequency collection, containing geolocation data at each different frequency, and 3 dataframes, one for each resource, containing data about the consumption of each resource by frequency.

2.3. Information generation

The second stage consisted in generating information from captured data. To generate information on resource consumption, confidence intervals for each resource are calculated by collection frequency. This estimate was used for verifying whether there is a significant difference in relation to resource consumption among the different collection frequencies. In order to calculate the confidence interval of each frequency, we first calculated the average values of the 5 days for each user; then, using this, the sample standard deviation could be calculated, and thus obtaining the confidence intervals.

In regard to the analysis of geolocation information, it should be noted that the contexts are different in each of the 4 datasets, that is, the participants visited different places and performed different trajectories in each of the 5-day periods. Thus, it is not possible to carry out an adequate comparative analysis between them. For example, we intend to carry out analysis in terms of the loss of geolocation information, such as the accuracy of the trajectory produced from a set of geographic coordinates. Thus, this requires to compare the same trajectory produced from coordinates collected at different frequencies, and consequently it becomes necessary to keep the same spatial context (i.e., to use spatial data produced in the same period of time).

To address this issue, we produced datasets of 30-, 45- and 60-second frequencies by temporally aggregating the data from the 15-second frequency dataset, which was built in the first 5 days of the data collection phase. By applying this methodology, we ensured that all 4 datasets have the same spatial context. The procedure consisted in

removing values from the original dataset. For example, to generate a dataset of 30-second frequency from 15-second frequency dataset, we just removed half of the data points (alternately). After deriving such 3 "simulated" datasets, algorithms were applied on the data to generate geolocation information. The algorithms produced information about: trajectory identification performed by users, stop regions, information from the embedded sensors and the use of smartphone.

2.4. Metrics and information analysis

Finally, in the third stage, an analysis of the information generated in the previous stages was performed, in order to compare the tradeoff between the consumption of smartphone resources and the geolocation data quality. To perform the resource consumption analysis, we carried out a comparative study between the values of the confidence intervals for the consumption value of each resource by frequency collection, aiming at evaluating whether there is a significant difference between the values of each frequency. After analysing the consumption of each smartphone's resource by frequency, a comparative analysis was performed to quantify the impact that each frequency had in the loss of geolocation information, using the 3 derived datasets and the original one.

The metric used to compare geolocation data quality, in relation to the trajectories produced, was the level of dissimilarity, calculated though the Dynamic Time Warping algorithm [Lerato and Niesler 2019]. This algorithm provides a numerical value related to the measure of distance between trajectories, which can be interpreted as the level of dissimilarity, or loss of information, between the trajectories.

In order to identify stop regions, we made an ad-hoc implementation based on DBSCAN. In addition, to assess the quality of information on stop regions, the metric adopted was the similarity in the number of stop regions identified in each derived dataset and the original ones. A similar metric was adopted to calculate the similarity in relation to smartphone usage. For example, taking into account the original dataset compared with the 30-second derived dataset, to calculate the similarity between information on whether the smartphone screen is on or off, it is only necessary to check, in each pair, in sequence, of the original dataset (which was mapped to a single value of the derived set), the number of occasions when the screen was off and compare with the number of occasions when the screen was off in the derived dataset, since each pair in the original dataset value is mapped to a single value in the derived dataset.

The same strategy was also applied to calculate the loss of information in relation to the smartphone audio. In order to calculate similarity for the others information (i.e., ambient lighting, smartphone usage and identification of the audio mode), the metric used was the average of each sequence of values in the original dataset, which was mapped to the derived dataset, and calculate the difference, in percentage, in relation to the value of the derived dataset. Thus, the sum of all differences is the value of dissimilarity. The greater the value, the greater the loss of information. The next section presents the results obtained.

2.5. Devices (smartphones)

Strategy to select and prepare the devices: To carry out the experiments and measurements, the first step was to select volunteers who have Android devices, so that the data

capture application could be installed. Most smarphones were Samsung Galaxy Note 3, Motorola Moto G5, Xiaomi Redmi Note 8, among other similar models. The Android versions were between 5.0 and 9.0. We required at least version 5.0 since it allows monitoring the resources consumed by applications individually (some older versions of Android only offered general device consumption). The next step was to install the *My Data Manager*[1] tool on the devices to monitor data consumption by application. Since Android supports the monitoring of CPU usage and energy consumption, no third party software was needed for obtaining this information.

Discussion on the use of different devices: The devices differ both in hardware configurations and in terms of Android versions. We decided to use a diversity of devices with the aim of obtaining a more accurate and generalised simulation of reality, that is, to obtain a greater representation of how the consumption of resources occurs in devices with different characteristics. The use of devices with different characteristics does not pose a risk to the validity of the obtained results, since the consumption information was captured individually by application, that is, the values analysed corresponds only to the resources consumed by the application used to conduct the research. In addition, we have ensured the same set of devices were used in different stages of this research (when different frequencies of collection were adopted). Thus, since the same diversity of devices are encountered in each group (each frequency), it possible to compare the average values obtained in each group, without posing a risk to the validity of the experiment.

3. Results and Discussions

By following the methodology discussed above, it was performed a comparative analysis of each resource consumption by collection frequency, which were calculated in terms of confidence intervals. In addition, it was also obtained the comparative analysis of the geolocalised data of the dataset with greater granularity (dataset with frequency time of 15 seconds), with the derived datasets (datasets of 30, 60 and 120 seconds). Finally, it was performed a comparative analysis for proximity sensors and ambient lighting, and for the indicators smartphone usage: whether the screen was locked or unlocked and if the audio was in normal, mute or silent mode.

In Figure 1, it is possible to notice that the data collection with a frequency of 15 seconds had the highest resources consumption. In contrast, the statistical test of the confidence intervals indicated that there were no significant difference in consumption of the resources, among the collection frequencies 30 and 60 seconds, because in all cases their confidence intervals had intersection of values. In other words, this result indicates that using the 60-second collection frequency does not present significant gains in the resources consumption in comparison to collection frequency of 30 seconds. Regarding the collection frequency of 120 seconds, it presented the best performance, that is, lower consumption, for data and battery resources, compared to all other frequencies analysed. However, there was no significant difference for the consumption of CPU compared to the frequencies of 30 and 60 seconds. Thus, it was possible to perceive that, for the consumption of resources, in certain cases, there is no advantages of using collection frequencies with lower collection granularity.

[1]My Data Manager - https://www.mydatamanagerapp.com

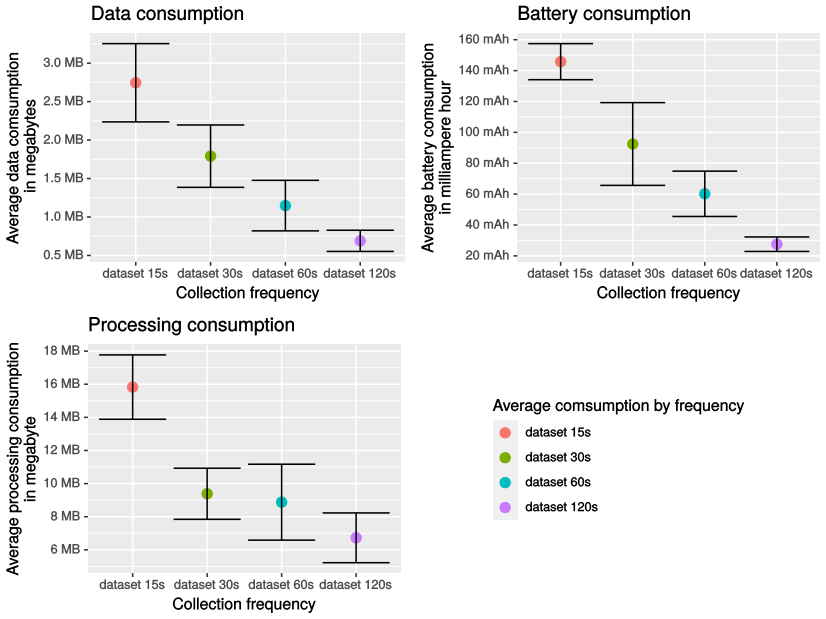

Figure 1. Charts of confidence intervals for each resource consumption by collection frequency

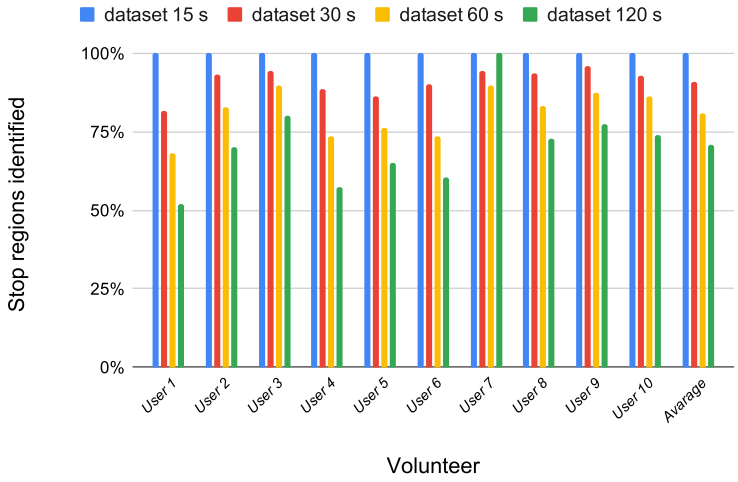

Figure 2. Chart of the number of stop regions identified by the original compared to derived datasets

In Figure 2, it can be seen the number of stop regions that were identified by each collection frequency. These stop regions took into account a radius of maximum distance of 50 meters, and a minimum time internal of 5 minutes. That is, for a stopping region to be identified, the user could not distance himself more than 50 meters from the region and stay at least 5 minutes in that region. The dataset with collection frequency of 30 seconds managed to capture an average of 90% of all stop regions, while the dataset with

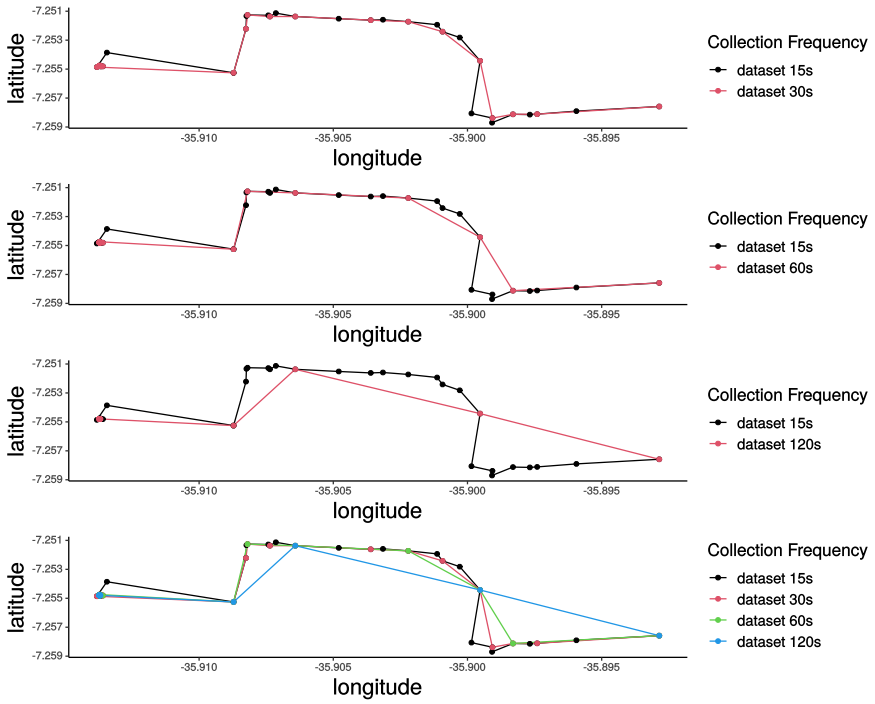

Figure 3. Graphs of a trajectory produced by dataset with a frequency of 15 seconds, compared to the same trajectory produced by derived datasets

a frequency of 60 seconds captured on average 80% and the dataset with a frequency of 120 captured only 70% of the stop regions. Hence, we can perceive a gradual increase in the number of stop regions that were not identified by the derived datasets. For the 30-second frequency, a loss of only 10% in the number of stop regions, for some usage scenarios, may not be so problematic. On the other hand, the adoption of a collection frequency of 120 seconds may significantly impact the application or research that relies on that information, given the average loss of 30% in the number of stop regions identified.

The distributions of the number of stop regions that were identified for each participating user are shown in the boxplots of Figure **??**. Depending on the movement pattern of each user, a different number of stop regions are identified. The variation was between 100 and 250, users who move more tend to have more stop regions in their trajectories, than users who stay stationary for a long period of time in the same region.

In Figure 3, it is possible to notice the loss of information of the trajectories produced by the derived datasets with collection frequencies of 30, 60 and 120 seconds, compared with the actual trajectory produced by the original dataset with collection frequency of 15 seconds. The loss of information on trajectories for the dataset with a frequency of 60 seconds, compared to the dataset with frequently of 30 seconds, grew by an average of 30%. For the dataset with frequency of 120 seconds, compared to the 30-second dataset, the loss of information grew by an average of 46%, indicating a significant increase in the loss of trajectory information produced by the derived datasets. Besides that, it is worth highlighting that this loss of information was obtained by taking into account only the first 5 days of study, where no volunteers have moved for a long period of time (such as a long

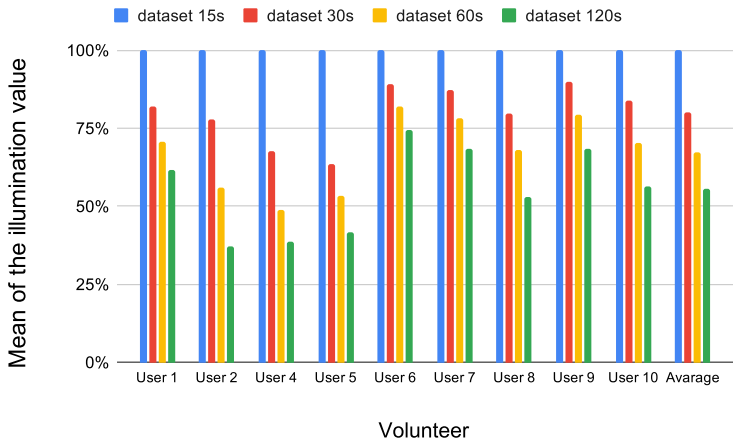

Figure 4. Average lighting values graph estimated by derived datasets compared to the actual values from the original dataset

road trip). If longer trajectories had been performed in the period studied, the observed loss of information could be even greater.

The averages values for information produced from the ambient lighting sensor are shown in Figure 4. As it is an indicator with a wide variation in its value over time, the loss of information was significantly large for the derived datasets, compared to the original dataset. This average value information was calculated using the average difference, in percentage, of the value estimated by the derived dataset compared to the actual value of the original dataset. The values obtained from the datasets with a frequency of 30, 60 and 120 seconds were, on average, 80%, 66%, 55% of the original value, respectively, indicating significant loss of information. Furthermore, if we remove the data in moments where the user was sleeping, on which the lighting was practically constant for a long period, the loss of information is even greater. For searches using this type of information, this difference may represent an significant estimate error in the survey. For example, a researcher/developer can try to determine whether the user was in a working environment or not, through the value of lighting, because the Brazilian Standard[2] determines that in offices and other working environments the ideal illuminance values should be from 500 to 1000 lux. Therefore, if data is captured from lighting with a hit rate of just 55% of the real value, the estimated value can easily be outside this range values indicating possibly misleading information, according to that the user would not be in a work environment.

The other sensor captured was the proximity sensor, which captures the distance that the front of the smartphone is from a particular object. In this case, the average inaccuracy of the values produced from the derived datasets was lower, when compared to the inaccuracy produced from the sensor lighting data. This is mainly due to the fact that it has less variation over time. The average of correct answers for collections with a frequency of 30, 60 and 120 seconds were 89%, 83% and 76%, respectively. This data can be important, for example, to determine situations on which the smartphone screen

[2]Standard NBR 5413 - Interior Illumination `http://ftp.demec.ufpr.br/disciplinas/TM802/NBR5413.pdf`

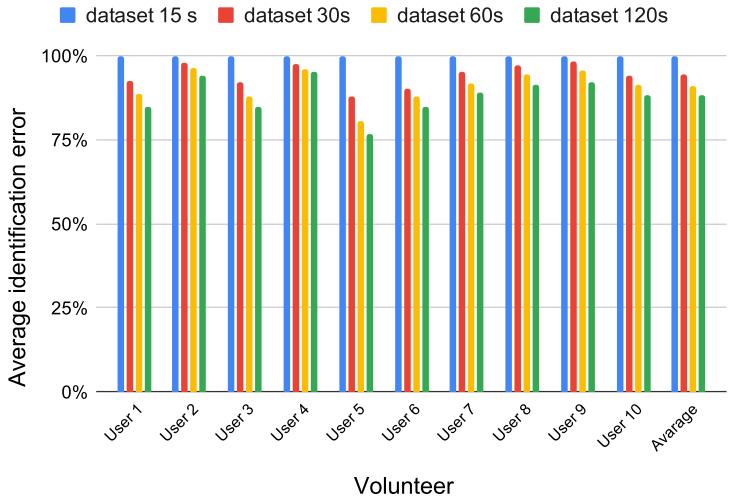

Figure 5. Graph of the average precision of the identification of the screen on or off

was close to the user's face, indicating that he could be in a call.

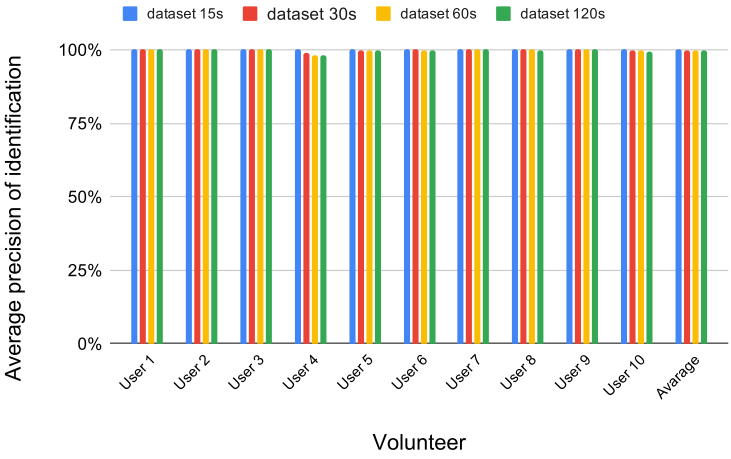

Figure 6. Graph of the average precision of the identification of the silent audio mode

Finally, in Figure 5, it is possible to observe the precision in relation to smartphone usage information, that is, if the screen was locked or unlocked. On average, the precision, was very high, for all the derived datasets, usually close to 90%. This is due to the fact that users, in general, leave the screen locked for a period of time longer than 1 or 2 minutes, or use the smartphone, with the screen unlocked, for a long period of time. The other data captured about the use of the smartphone was the audio mode (whether it was in normal, silent or silent mode). Similarly, as shown in Figure 6, the precision was also considerably high, averaging around 95% for all derived datasets. These results demonstrate that the

frequency of collection has no significant impact on loss of information of smartphone usage.

4. Conclusion

The frequency collection of geolocation data can directly impact on the performance of smartphones and also on quality of applications that use these types of data to different needs. The result obtained by the comparative analisys presented in this article is an important artifact to support development teams and researchers in their decisions regarding the frequency collection of geolocation data, so that they maintain user satisfaction without compromise the quality of the captured data.

In this study, it was possible to observe the advantages and disadvantages that each collection frequency presented in comparison to the others. Regarding the consumption of resources, as expected, the frequency with greater granularity (collection frequency of 15 seconds) had the worst performance, that is, the highest cost in the consumption of resources. The 30-second and 60-second frequencies, in turn, showed no significant difference in resource consumption between them. These results indicate that there are no advantage in choosing any one of the two frequencies (in relation to the resource consumption).

The frequency with less granularity of collection (120 seconds) obtained the best performance in relation to battery and data consumption. Nonetheless, it did not show any performance gains in relation to CPU usage, compared to the frequencies of 30 and 60 seconds. From this results, it is possible to see that, regarding the smartphone's resources consumption, in certain situations, there are no advantages in changing between these frequency of collection.

In addition, in relation to the quality of the data produced, it was possible to perceive that, regarding spatial data, the loss of information was considerably large for the frequencies 60 and 120 seconds. For example, the collection frequency of 120 seconds only managed to capture 70% of all stop regions. Moreover, it was possible to notice that, in relation to the information obtained from smartphone's sensors data, such as the lighting sensor, the loss of information was even greater, with hits of just 55% of the reference value. Finally, in relation to smartphone usage data, the loss of information was considerably smaller, showing that there is no great difference in information loss depending on the collection frequency.

In this perspective, the comparative analysis showed that, depending on the situation, a collection frequency may be more indicated than another. For example, in scenarios where one needs spatial data or sensor data, but does not need a smartphone's low resource consumption, the use a frequency with a 15-second collection would be the most adequate to obtain the data with higher quality and better precision. On the other hand, in scenarios where just smartphone usage data are needed, the adoption of frequencies with less granularity would not significantly impact the quality of data, therefore being a good alternative for decrease the cost of resource consumption. In general, the 30-second collection frequency was the one that obtained the best tradeoff between the frequencies analysed, as it presented a reasonable performance in resource consumption (similar to consumption with a frequency of 60 seconds), while obtaining a considerably better performance in relation to the quality of the information produced from data (when compared

to frequencies of 60 and above).

As future work, it is intended to replicate the experiment with a larger number of volunteers (at least 20), for a longer period of time (40 days). We also intend to analyse other factors, such as the activities performed by the users at specific times and other collection frequencies. Finally, other types of information derived from data should be considered, such as those related to quality of life metrics.

References

Barros T., Campelo, C. (2019). *Plataforma para fomentar a produção e a disponibilização de dados geolocalizados*. Federal University of Campina Grande, Campina Grande, PB, Brazil. Technical Report - Technological Initiation Program (CNPq-UFCG).

El Faouzi, N.-E., Leung, H., and Kurian, A. (2011). Data fusion in intelligent transportation systems: Progress and challenges – a survey. *Information Fusion*, 12:4–10.

Feng, Z. and Zhu, Y. (2016). A survey on trajectory data mining: Techniques and applications. *IEEE Access*, 4:1–1.

Herder, E., Siehndel, P., and Kawase, R. (2014). Predicting user locations and trajectories. volume 8538, pages 86–97.

Kong, X., Li, M., Ma, K., Tian, K., Wang, M., Ning, Z., and Xia, F. (2018). Big trajectory data: A survey of applications and services. *IEEE Access*, 6:58295–58306.

Lerato, L. and Niesler, T. (2019). Feature trajectory dynamic time warping for clustering of speech segments. *EURASIP Journal on Audio, Speech, and Music Processing*, 2019.

Luo, T., Zheng, X., Xu, G., Fu, K., and Ren, W. (2017). An improved dbscan algorithm to detect stops in individual trajectories. *ISPRS International Journal of Geo-Information*, 6:63.

Mazimpaka, J. D. and Timpf, S. (2016). Trajectory data mining: A review of methods and applications. *Journal of Spatial Information Science*, (13):61 – 99.

Parent, C., Spaccapietra, S., Renso, C., Andrienko, G., Andrienko, N., Bogorny, V., Damiani, M. L., Gkoulalas-Divanis, A., Macedo, J., Pelekis, N., Theodoridis, Y., and Yan, Z. (2013). Semantic trajectories modeling and analysis. *ACM Comput. Surv.*, 45(4).

Vaughan, N. and Gabrys, B. (2016). Comparing and combining time series trajectories using dynamic time warping. *Procedia Computer Science*, 96:465–474.

Xiang, L., Gao, M., and Wu, T. (2016). Extracting stops from noisy trajectories: A sequence oriented clustering approach. *ISPRS International Journal of Geo-Information*, 5:29.

Zheng, Y. (2015). Trajectory data mining: An overview. *ACM Trans. Intell. Syst. Technol.*, 6(3).