SmartGPS: An Android app for collecting urban mobility data

***Milton Giovanny Moncayo*[1,2]**
Ismaïl Saadi[2]
Marc Van Droogenbroeck[2,3]
Mario Cools[2,4,5]

Abstract: *The complexity of urban mobility patterns has been studied many years ago. Thanks to the continuous advance of information and communication technologies (ICTs), several mechanisms have been used to collected mobility data and, consecutively analyze these data using statistical and artificial intelligence techniques with an objective to derive important findings regarding people's mobility behavior. One of these techniques concerns to collecting of data based on GPS. Nowadays, due to the common daily use of smart phones, this technique can be used with considerable ease in research for the identification of urban mobility patterns. This paper presents a new approach to urban mobility data collection using Android GPS devices. It provides an overview of the performance of the SmartGPS application during the data collection process. Furthermore, we present a validation, analysis and visualization of the data using statistical techniques.*

Keywords: "Urban mobility", "Smart Phone", "Android OS", "GPS", "Statistical Analysis".

1. Introduction

Nowadays, several studies provided insight to understand the complexity of urban mobility patterns. In this regard, data collected from sensors of mobile devices combined with GPS information opened new research opportunities. Some approaches have been adopted for identifying the common factors in the daily movement of people, such as trip purpose (Montini, Rieser-Schüssler, Horni, & Axhausen, 2014), mode choice, activity-travel patterns detection, route choice, etc (Wang, He, & Leung, 2017). For example, Byon, Abdulhai, & Shalaby (2009) used GPS traces to find mobile device location with a temporal resolution of 5 minutes and find travelers' transportation mode. Xia, Qiao, Jian, & Chang (2014) incorporated accelerometer data to categorize travelers under their corresponding outdoor transportation mode: walking, bicycling, motorized transport or stationary state. For the same purpose, Fang et al. (2016) collected data from the accelerometer, magnetometer and gyroscope sensors to enhance the accuracy, additionally comparing the different transportation modes with the vehicle classification: motorcycle, car, bus, metro or train.
Another kind of projects deals with the collection and analysis of crowded behavior data, collected through mobile phones applications, and have developed new technologies to mine this collected data (Vlassenroot, Gillis, Bellens, & Gautama, 2015), One such case is the MOVE project, which describes the different steps in the development of tracking applications for smartphones that make use of advanced data mining to build an urban data monitoring system. Another example is SMARTMO (Berger & Platzer, 2015), an innovative approach to the collection of travel behavior data, or in the same way TRAC-IT (Barbeau &

[1] Central University of Ecuador, Engineering Faculty – Physical and Mathematic Sciences, mmoncayo@uce.edu.ec
[2] University of Liège, Local Environment Management & Analysis (LEMA) – Urban and Environmental Engineering (UEE)
[3] University of Liège, Montefiore Institute of Electrical Engineering and Computer Science
[4] KULeuven Campus Brussels, Department of Informatics, Simulation and Modeling
[5] Hasselt University, Faculty of Business Economics

Labrador, 2009) or SWIPE (Faye et al., 2017), who have worked in the construction of GPS based smartphones datasets to study urban mobility patterns.
Besides, data collection through mobile devices is often accompanied by different problems, e.g. battery consumption (Montini, Prost, Schrammel, Rieser-Schüssler, & Axhausen, 2015), ethical issues, privacy or awareness of using the app for research purposes (Odiari, 2018).
In this study, we present a conceptual framework for collecting data through mobile sensors and GPS with Android Operating System. We will show to what extent such disaggregate data can provide more insights into the urban mobility of travelers. In this regard, an advanced statistical analysis of the collected data will be performed.

2. Data and Methods

2.1. SmartGPS App

The first part of the work focused on the development of SmartGPS application under Android OS which is one of the most widely distributed operating systems for mobile devices and always is in constant evolution. In fact, in the last 10 years 17 versions have been released (Android version history, 2019), which implies at least one version each year. (see Table 1)

Table 1 : Android Version History (Android releases, 2018)

Android Version	API Level	Number Version(s)	Date Release	% use
Apple Pie1	1	1,0	2008-09-23	< 0,1
Banana Bread1	2	1.1	2009-02-09	< 0,1
Cupcake	3	1.5	2009-04-25	< 0,1
Donut	4	1.6	2009-09-15	< 0,1
Eclair	5 – 6 – 7	2.0 – 2.1	2009-10-26	< 0,1
Froyo	8	2.2 – 2.2.3	2010-05-20	< 0,1
Gingerbread	9 – 10	2.3 – 2.3.7	2010-12-06	0,2
Honeycomb2	11 – 12 – 13	3.0 – 3.2.6	2011-02-22	< 0,1
Ice Cream Sandwich	14 – 15	4.0 – 4.0.5	2011-10-18	0,3
Jelly Bean	16 – 17 – 18	4.1 – 4.3.1	2012-07-09	3,0
KitKat	19 – 20	4.4 – 4.4.4	2013-10-31	7,6
Lollipop	21 – 22	5.0 – 5.1.1	2014-11-12	17,9
Marshmallow	23	6.0 – 6.0.1	2015-10-05	21,3
Nougat	24 – 25	7.0 – 7.1.2	2016-06-15	28,2
Oreo	26 – 27	8.0 – 8.1	2017-08-21	21,5
Pie	28	9.0	2018-08-06	< 0,1
"Q"	29	10.0	2019-06-01	< 0,1

Android's evolution implies that the applications must adapt quickly to new features and enhancements, which are generally about device design, security and battery saving. One example of it, is the version 8-Oreo, which incorporated a substantial improvement in terms of battery savings, limiting the access to the current location of the device for applications running in background (Android Oreo limitations, 2017).
SmartGPS is a first version of an app created to collect GPS data from Android SmartPhones, specifically designed to run in background. Unlike another apps, it incorporates multiple services to achieve the highest possible accuracy GPS locations, every 15 seconds when the device is in motion. These services operate in a smart way to save battery power. However, to make it possible, it is necessary the devices have Internet connection all the time.

2.2. Data Collection

Data collection was performed with around 400 voluntary people of Quito-Ecuador; most of them were students from Central University of Ecuador, with 58% Men and 42% Women, ages between 20 and 30 years. All participants were informed about the app operation, mainly to the battery power consumption and the GPS track of the device in background.
The application was installed and operated mainly on Samsung, Huawei, Sony and Xiaomi devices, with Android versions 4-KitKat, 5-Lollipop, 6-Marshmallow, 7-Nougat and 8-Oreo, the latter being the predominant version. (see Figure 1).

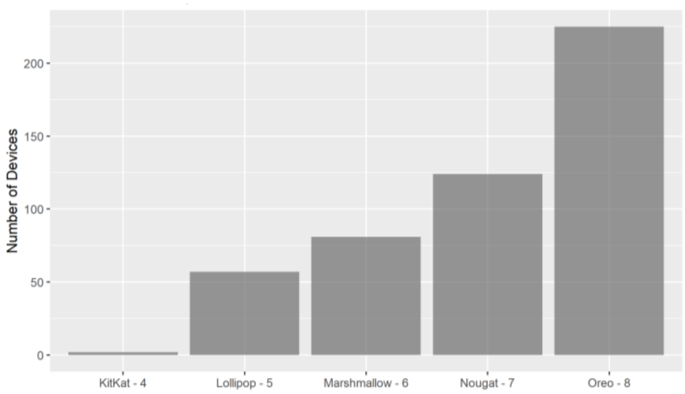

Figure 1 : Android Version Devices

During about three weeks between January and February 2019, data from participants were collected. Daily, the data were stored internally on the mobile device and later sent to a centralized online database. The amount of data collected per Android version is shown in Figure 2

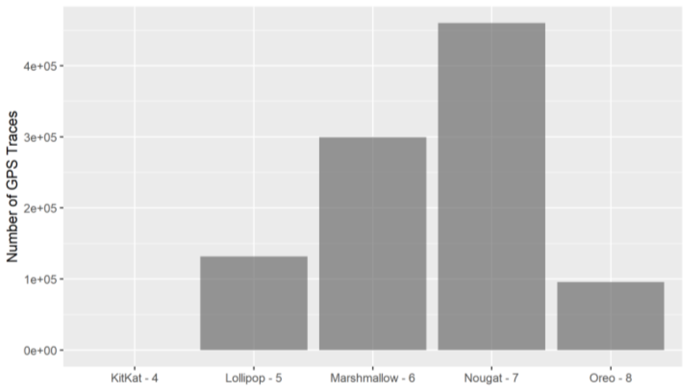

Figure 2 : Android Version Data Collected

Whereas most of devices involved in data collection executed the app over Android 8-Oreo version, the amount of data collected by these devices was not significant compared with the number of the devices that installed the app over the same version. The reason is the improvement incorporated by Google in the API of this version that limits the access to the current location to save battery power.

2.3. Dataset Definition

The dataset structure contains 12 variables; the description of each one is shown in Table 2.

Table 2 : Dataset Definition

Field	Type	Description
dspId	Number	Id to identify each device in the dataset
dspLat	Number	Latitude, in degrees
dspLon	Number	Longitude, in degrees
dspNSat	Number	Number of satellites used by the GPS engine to compute the most recent fix
dspAccu	Number	Average of horizontal accuracy of providers. Approximately 1 = exact location, Approximately 2 = not exact location
dspAlti	Number	The altitude if available, in meters above the WGS 84 reference ellipsoid. If not, 0
dspVelo	Number	The speed if available, in km/h. If not, 0
dspActy	Factor	Activity recognition based on Google API
dspFcIn	Factor	Date in which data was collected
hour	Factor	Hour in which data was collected
weekDay	Factor	Day of week in which data was collected
dspFcUp	Factor	Date in which data was updated after 15 secs. If it is different to dspFcIn, it means there was no movement during this time

3. Results

To show the results of this work, an exploratory analysis and spatial visualization of data was performed using the R software (Team R Core, 2018). For exploratory data analysis, the packages: “dplyr”, “VIM”, “lubridate”, “sqldf” and “vcd” were used, and for the spatial data visualization, we used the “ggmap” package (Kahle & Wickham, 2019).

3.1. Data Collected Quality

Before to start the analysis, it was necessary to check the data quality and sanity. By doing an aggregation for missing values, we found around 28% of missing data, the largest portion of missing data was in the fields: dspAlti, dspActy, dspNSat and dspAccu. On the other hand, 72% of data were complete (see Figure 3).

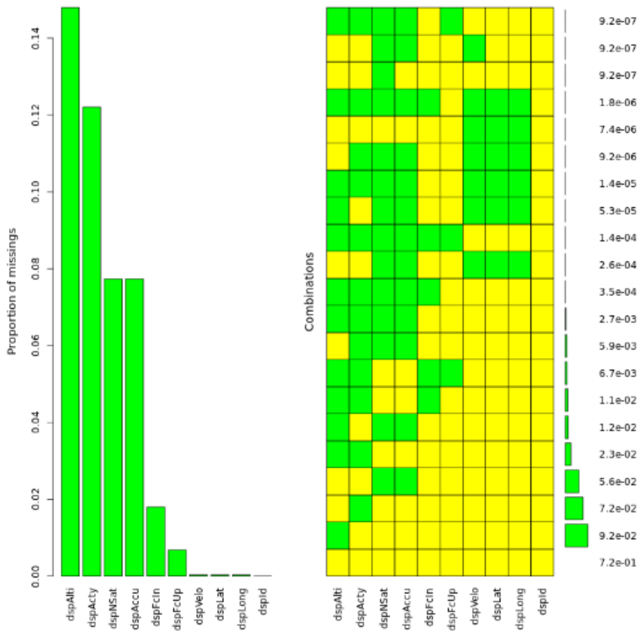

Figure 3 : Aggregation missing/imputed values

In addition, duplicated values and values outside study area (Quito) were found in the dataset. These outliers could cause problems with the data analysis, so that it was necessary to clean the data. Deleting these rows, led to a reduction of 70% of the data

3.2 Exploratory Data Analysis

We examine the correlation between numerical and categorical variable values in Table 3 and Table 4 respectively.

Table 3 : Numerical variables correlation - Pearson Coefficient

	dspLat	dspLon	dspNSat	dspAlti	dspVelo	dspAccu
dspLat	1,0000	0,4749	0,0150	0,0017	0,0085	-0,1161
dspLon	0,4749	1,0000	0,0187	-0,1324	0,0713	-0,0196
dspNSat	0,0150	0,0187	1,0000	0,0690	0,1290	0,0983
dspAlti	0,0017	-0,1324	0,0690	1,0000	0,0397	-0,0575
dspVelo	0,0085	0,0713	0,1290	0,0397	1,0000	-0,0099
dspAccu	-0,1161	-0,0196	0,0983	-0,0575	-0,0099	1,0000

The correlation matrix does not show a direct relationship between the numerical variables of dataset. One of the most influential variables in the table is the number of satellites “dspNSat”. It is possible to observe a possible, but weak, relationship between the number of satellites “dspNSat”, altitude “dspAlti”, velocity “dspVelo” and accuracy “dspAccu”, which tends the prove that when more satellites are visible, the device improves the precision of the location by obtaining better altitude, velocity and accuracy data.

Table 4 : Categorical variables correlation – Cramer’s V Coefficient

	Activity - Week Day	Activity – Hour
Contingency Coefficient	0,108	0,283
Cramer's V	0,048	0,132

Similarly, as shown in Table 4, the categorical correlation matrix shows a possible, but weak relationship, between the activity and the hour of the day in which it is performed.

The plausibility of the missing data values was examined with density diagrams (Figure 4 and Figure 5). These show that data values are within the expected ranges. For example, in Figure 4, the concentration of “Altitude” values refer to Quito's altitude (2.850 m approximately); the "Number of satellites" shows that for most data, locations were calculated with approximately 18 to 23 available satellites.

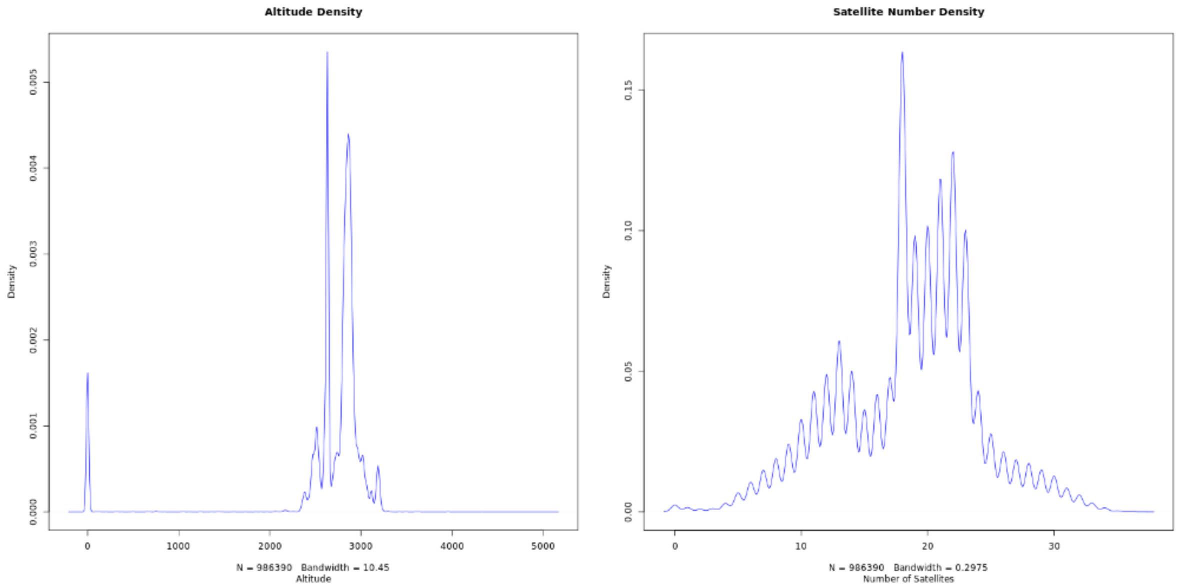

Figure 4 : Altitude and Number of Satellites density diagrams

In Figure 5, the "Accuracy" has values closer to two, because of the well-known margin of error existing in the hardware of mobile devices. Finally, the density diagram of "Activity" indicates that most of the device location readings were taken when it was still.

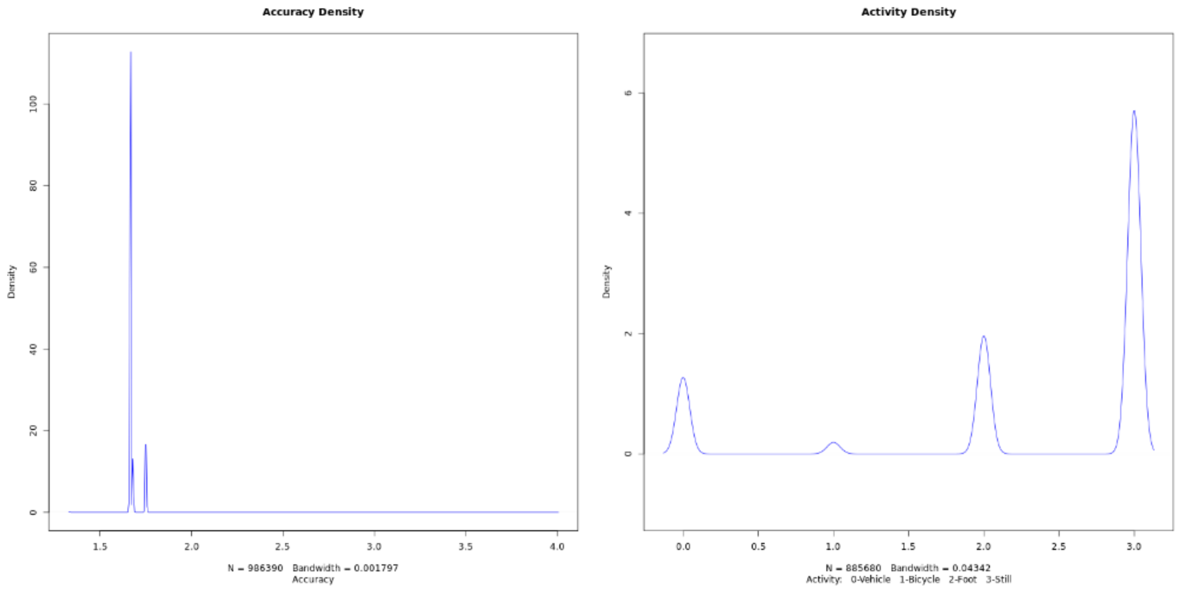

Figure 5 : Accuracy and Activity density diagrams

The last indicator about “Activity” can be ratified with the following diagrams, about the distributions of Activity by week day and hour, respectively. (see Figure 6).

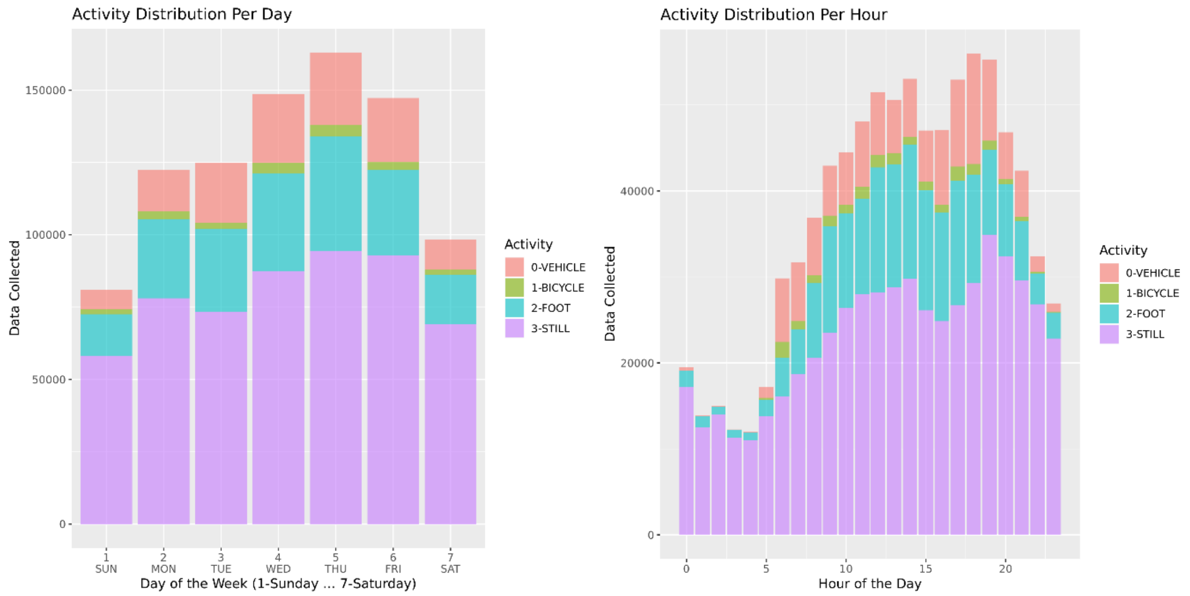

Figure 6 : Activity distribution diagrams. Data collected per week day and hour

The analysis of the distribution of velocity data per day and hour (Figure 7) shows that both “Activity” and “Velocity” have similar distribution functions. It means, their values are high or low in the same week day or in the same hours during the day.

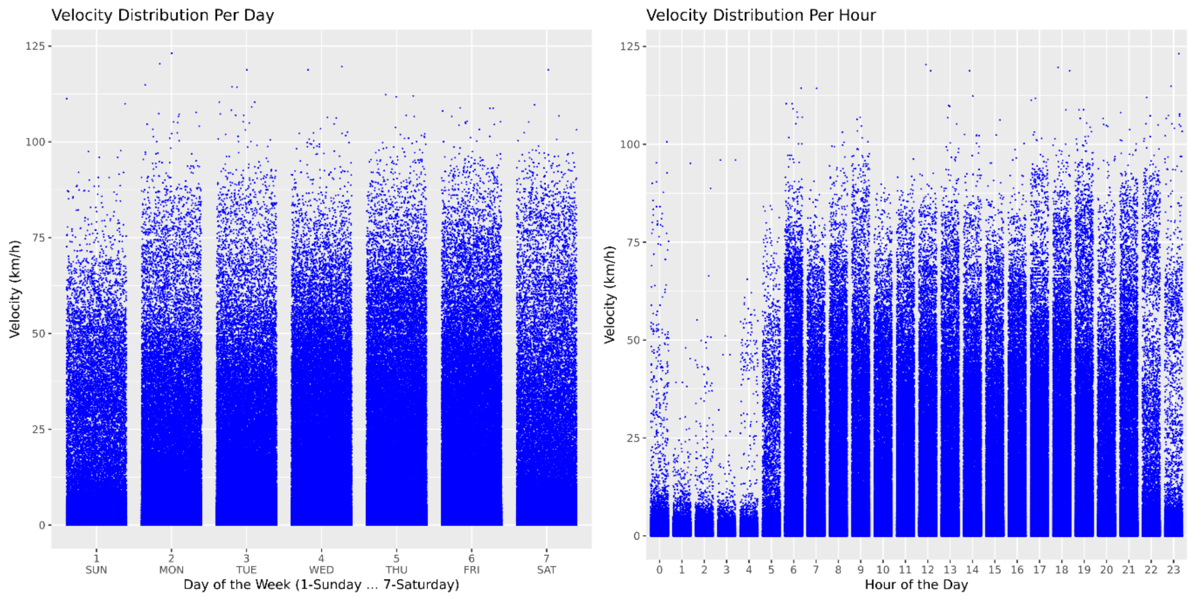

Figure 7 : Velocity distribution diagrams. Data collected per week day and hour

3.3 Spatial Data Visualization

Finally, GPS data collected were plotted on the city map of Quito (see Figure 8).

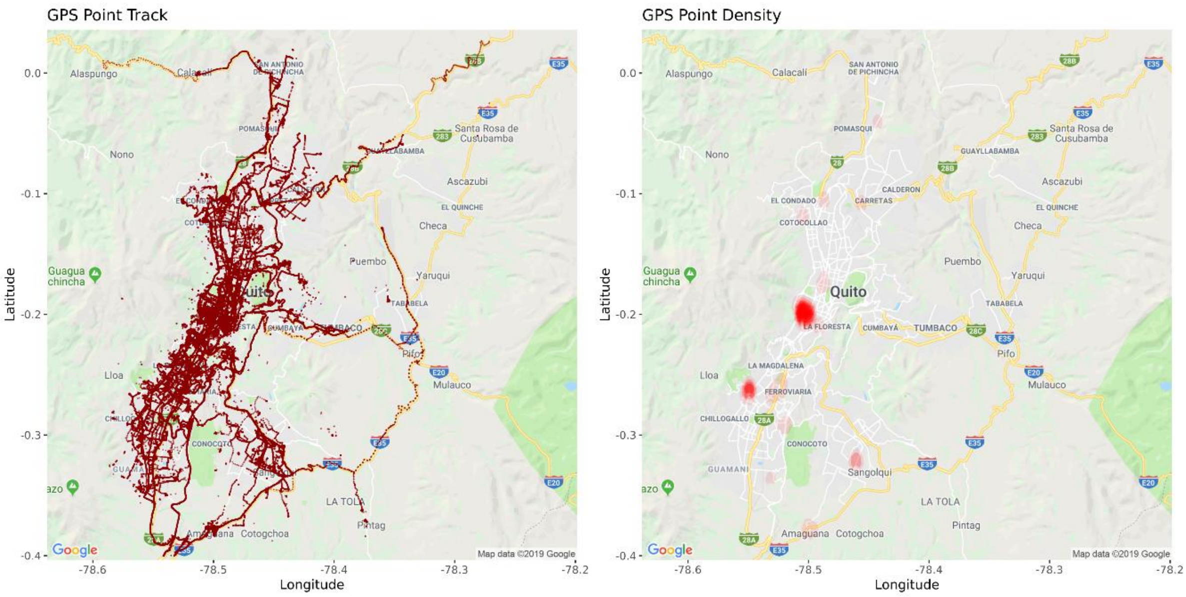

Figure 8 : GPS points plot. Track and density ponts

The maps allow identifying the transit of the devices through the different places of the city. Some places where there exists a high density of points are marked in red, these places correspond to residential or study areas, avenues and highways of Quito.

4. Discussion

Related to SmartGPS app, constant Android version's evolution results in deprecated apps, which on several times forbid its proper functioning in different models of mobile devices. In this work, we used services to save battery power, however, these services behave differently on each device depending on the installed Android version; this problem is highlighted with Android version 8-Oreo.

Related to the data analyzed in this paper, have allowed us to identify some basic urban mobility behaviors, such as hours and days of the week in which people are most active. In the same way, through the location of GPS points on the map, different urban, residential and school areas have been identified, as well as the main highways in Quito. This knowledge can be used to help governments in planning and managing mobility and traffic cities. It is possible to generate a high impact on transport efficiency through the implementation of forecasting models for activity recognition (Work, Studies, Home, Social), movement (Origin-Destination trip, Trip purpose), velocity limits by zones (School, Urban, Residential), etc.

It is also important to mention 3 sensitive factors in the study. The first one is the privacy of people. The second one is related to the power battery consumption. Both generated reluctance of people to participate to the project. The third factor relates to the Internet connection in Ecuador. According to data from INEC (INEC, 2017) and ARCOTEL (ARCOTEL, 2019) approximately 32% of the Economically Active Population (EAP), has access to mobile Internet in the country. It means that an important amount of devices probably did not have Internet connection all the time during data collection and this causes the app not to work properly and generates missing or duplicated data.

Based on this preliminary analysis, several studies can be started, such as individual data analysis of devices, data analysis for a specific day of the week or time of the day. In the same way, it is necessary to continue collecting information and incorporating more fields in the data set, like accelerometer, gyroscope, luminosity, battery level or step count data.
The future work will consist to build a methodology that will allow us to find more specific mobility patterns.

5. Conclusions

It is important to correct errors in data collection, improve the concurrency management of central database, as well as to ensure that devices have connection to the Internet all the time. All of these factors will be essential to ensure best data quality and avoid missing values.
Based on the analyses of the data, a general idea of the mobility behavior in Quito can be formulated, for example, it can be observed that between Wednesday and Friday people show more activity, while on weekends the activity decreases considerably. In the same way, the duration and timing of the peak in movement of people can be deducted.

6. References

Android Oreo limitations. (2017). Android Oreo limitations. Retrieved from https://developer.android.com/about/versions/oreo/background-location-limits

Android releases. (2018). Android Releases. Retrieved from https://developer.android.com/about/dashboards

Android version history. (2019). Android version history. Retrieved from https://en.wikipedia.org/wiki/Android_version_history

ARCOTEL. (2019). Servicio de acceso a internet (SAI) – Agencia de Regulación y Control de las Telecomunicaciones. Retrieved April 20, 2019, from http://www.arcotel.gob.ec/servicio-de-acceso-a-internet-sai2/

Barbeau, S., & Labrador, M. (2009). TRAC-IT: A Software Architecture Supporting Simultaneous Travel Behavior Data Collection and Real-Time Location-Based Services for GPS-Enabled Mobile …. *Proceedings of the …*, 1–21. Retrieved from http://www.locationaware.usf.edu/wp-content/uploads/2010/03/Barbeau-et-al-TRACIT-Software-Architecture-for-GPS-enabled-mobile-phones-FINALPR.pdf

Berger, M., & Platzer, M. (2015). Field evaluation of the smartphone-based travel behaviour data collection app "smartMo." *Transportation Research Procedia*, *11*, 263–279. https://doi.org/10.1016/j.trpro.2015.12.023

Byon, Y. J., Abdulhai, B., & Shalaby, A. (2009). Real-time transportation mode detection via tracking global positioning system mobile devices. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, *13*(4), 161–170. https://doi.org/10.1080/15472450903287781

Fang, S. H., Liao, H. H., Fei, Y. X., Chen, K. H., Huang, J. W., Lu, Y. D., & Tsao, Y. (2016). Transportation modes classification using sensors on smartphones. *Sensors (Switzerland)*, *16*(8), 1–15. https://doi.org/10.3390/s16081324

Faye, S., Louveton, N., Jafarnejad, S., Kryvchenko, R., Faye, S., Louveton, N., … Open, A. (2017). An Open Dataset for Human Activity Analysis using Smart Devices, 0–5.

INEC. (2017). Tecnologías de la Información y Comunicación-TIC | Instituto Ecuatoriano de Estadística y Censos. Retrieved from http://www.ecuadorencifras.gob.ec//tecnologias-de-la-informacion-y-comunicacion-tic/

Kahle, D., & Wickham, H. (2019). ggmap: Spatial Visualization with ggplot2. *The R Journal*, *5*(1), 144. https://doi.org/10.32614/rj-2013-014

Montini, L., Prost, S., Schrammel, J., Rieser-Schüssler, N., & Axhausen, K. W. (2015).

Comparison of travel diaries generated from smartphone data and dedicated GPS devices. *Transportation Research Procedia*, *11*, 227–241. https://doi.org/10.1016/j.trpro.2015.12.020

Montini, L., Rieser-Schüssler, N., Horni, A., & Axhausen, K. (2014). Trip Purpose Identification from GPS Tracks. *Transportation Research Record: Journal of the Transportation Research Board*, *2405*, 16–23. https://doi.org/10.3141/2405-03

Odiari, E. A. (2018). *A Framework for Big Data in Urban Mobility and Movement Patterns Analysis*.

Team R Core. (2018). R: A Language and Environment for Statistical Computing. Retrieved from https://www.r-project.org/

Vlassenroot, S., Gillis, D., Bellens, R., & Gautama, S. (2015). The use of smartphone applications in the collection of travel behaviour data. *International Journal of Intelligent Transportation Systems Research*, *13*(1), 17–27. https://doi.org/10.1007/s13177-013-0076-6

Wang, Z., He, S. Y., & Leung, Y. (2017). Applying mobile phone data to travel behaviour research : A literature review. *Travel Behaviour and Society, 11*(March), 141–155. https://doi.org/10.1016/j.tbs.2017.02.005

Xia, H., Qiao, Y., Jian, J., & Chang, Y. (2014). Using smart phone sensors to detect transportation modes. *Sensors (Switzerland), 14*(11), 20843–20865. https://doi.org/10.3390/s141120843