UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

Machine Learning applied to Last Mile Operations: Development of an Artificial Intelligence Model for Detention Analysis in Urban Freight

David Alejandro Calahorrano Salazar
Daniel Rolando Masaquiza Chango
Henry Martin Gavilanes Rangles

Ingeniería Industrial

Trabajo de fin de carrera presentado como requisito
para la obtención del título de
INGENIERO INDUSTRIAL

Quito, 22 diciembre de 2020

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

Machine Learning applied to Last Mile Operations: Development of an Artificial Intelligence Model for Detention Analysis in Urban Freight

David Alejandro Calahorrano Salazar
Daniel Rolando Masaquiza Chango
Henry Martin Gavilanes Rangles

Nombre del profesor, Título académico **Carlos Suárez, PhD**

Quito, 22 diciembre de 2020

© DERECHOS DE AUTOR

Nombres y apellidos:	David Alejandro Calahorrano Salazar
Código:	00130589
Cédula de identidad:	1724062029
Nombres y apellidos:	Daniel Rolando Masaquiza Chango
Código:	00130735
Cédula de identidad:	1804258703
Nombres y apellidos:	Henry Martin Gavilanes Rangles
Código:	00132626
Cédula de identidad:	1716625544
Lugar y fecha:	Quito, 22 de diciembre de 2020

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

Hoy en día la optimización de los recursos es un aspecto vital en el desarrollo de una empresa. Es necesario aprovechar todos los recursos disponibles, especialmente cuando el deseo de los clientes es obtener productos más rápidos y más baratos. Además, la mayoría de las empresas dedicadas a la distribución de productos no explotan todos los recursos adquiridos. Este es el caso del sistema de rastreo GPS, en el que sus datos GPS no se utilizan para análisis posteriores. La falta de estudios y de un mayor análisis de los datos del GPS inspiró el siguiente estudio. Con el fin de lograr políticas públicas más eficientes y mejores rutas de operación de última milla, decidimos analizar los datos del GPS de carga urbana, proporcionados por la OTUC, con el objetivo principal de obtener el tiempo de detención de cada vehículo de carga urbana y clasificar la detención como detención de tráfico, detención de entrega o detención de descanso. Después de clasificar cada detención utilizando metodologías de minería de datos y aplicando modelos de inteligencia artificial como K-Means y HDBSCAN, el estudio demuestra que es posible desarrollar una inteligencia artificial capaz de agrupar los registros de detención del GPS por características similares y dar diferentes agrupaciones como resultado principal. Los clústeres obtenidos son aportaciones importantes para futuras investigaciones y posibles aplicaciones, como en la toma de decisiones de políticas públicas o en la optimización de rutas logísticas, lo que permite a las entidades públicas y privadas optimizar la logística urbana.

Palabras clave: Aprendizaje automático, logística urbana, operaciones de última milla, análisis de datos geoespaciales, HDBSCAN, Cluster.

ABSTRACT

Nowadays the optimization of resources is a vital aspect in a company's development. It is necessary to take advantage of all resources available specially when customers' desire is to get faster and cheaper products. In addition, most of the companies dedicated to product distribution does not exploit all the resources acquired. This is the case of GPS tracking system where its GPS data is not used for further analysis. The lack of studies and further GPS data analysis inspired the following study. In order to achieve more efficient public policies and better last mile operation's routes, we decided to analyze urban freight GPS data, provided by OTUC, with the main objective of getting the time in detention of each urban freight vehicle and classify the detention as a traffic detention, delivery detention or rest detention. After classifying each detention by using data mining methodologies and applying artificial intelligence models such as K-Means and HDBSCAN, the study demonstrate that it is possible to develop an artificial intelligence capable of grouping GPS detention records by similar features and giving different clusters as the main output. The clusters obtained are important inputs to future research and potential applications such as in public policy decision making or logistic routes optimizations allowing both, private and public entities to optimize urban logistics.

Key Words: Machine Learning, Urban Logistics, Last Mile Operations, Geospatial Data Analysis, HDBSCAN, Cluster.

TABLE OF CONTENTS

INTRODUCTION 10

LITERATURE REVIEW 14

Global Positioning system 14

Related Topics 15

K-Means 17

Hierarchical Density Based Spatial Clustering Application with Noise (HDBSCAN) 18

METHODOLOGY 19

EXPLORATORY DATA ANALYSIS. 21

DATA PRE-PROCESSING AND CLEANING 24

DETENTIONS IDENTIFICATION ALGORITHM 26

DETENTIONS CLASSIFICATION MODEL 28

K-Means Model 28

HDBSCAN Model 31

STUDY RESULTS AND POTENTIAL APPLICATIONS 33

Study Results and Potential Applications 33

CONCLUSIONS 40

REFERENCES 42

TABLE`S INDEX

Table 1. *Identification Algorithm Results* 33
Table **2**. *K-means Cluster 1 Results.* 34
Table **3**. *K-means Cluster 2 Results.* 35
Table 4. *K-means Cluster 3 Results.* 36
Table 5. *K-means Cluster 4 Results.* 37
Table 6. *Clustering Results Summary*. 38

FIGURE`S INDEX

Figure 2. Quantity of GPS Records per Week. 21
Figure 3. Printing Frequency of GPS Records per Hour of the Day... 22
Figure 4. The analysis of the key variable [Speed]. 23
Figure 5. Data pre-processing and Cleaning Flow.. 24
Figure 6. Total Quantity of GPS Records per Variable.. 25
Figure 7. Data Algorithm Flow 26
Figure 8. Distribution of the Variable Time in Detention. 29
Figure 9. Distribution of the Variable Time in Detention Logarithmic Transformation. 30
*Figure 11***.** HDBSCAN Cluster Results. 34
*Figure 12***.** HDBSCAN Cluster Results. 35
*Figure 13***.** *HDBSCAN Cluster Results.* 36
*Figure 14***.** *HDBSCAN Cluster Results.* 37
Figure 15. HDBSCAN Cluster Results. 39

INTRODUCTION

According to the United Nations, nowadays more than half of the world 's population lives or works in the urban areas of cities around the world and predictions show that it will reach a top of 70% in the following thirty years (Laranjeiro, 2019). The efficient movement of daily consumption products such as groceries, industrial supplies and other staples of modern life is critical to reach competitiveness and develop urban economies (Hess, 2015). However, developing countries have been struggling with satisfying the demand for additional road capacity, a demand created by the exponential growth of population and economic activity. Urban freight transport and logistics activities have experienced an exponential increment all around the world (Sharman, 2011). Indeed, according to Kin, 80 percent of the domestic's products are produced in cities and most consumption happens inside the urban area. As a result, a reliable Urban Freight Transport system is needed (2017).

For a long time, different geolocation data has been used for different applications. One of them is to observe patterns, identify possible solutions and aim for different and better road systems in traffic and transportation management. (Spaccapietra et al., n.d.). By the same token, Global Positioning System (GPS) data has gained popularity when measuring freight performance is needed. GPS systems presents advantages over traditional measuring tools such as traffic cameras, loop detectors and others. By using GPS, the system can capture continuous vehicle traces data (Du and Aultman-Hall, 2007). However, GPS data has its own limitations like spatial inaccuracy provoked by urban canyon or misleading information caused by signal loss (Du and Aultman-Hall, 2007). Xia Yang states that access to large GPS databases could be beneficial in emerging countries where data collection is scarce since other sources of data unavailable or inactive (2014).

The increment of logistics activities and the lack of geospatial data has created the necessity of finding optimal solutions in order to supply market's needs (Khan, 2001). Not to mention the constant pressure city authorities suffer in order to mitigate the negative impact of freight transportation and the environmental damage (Khan, 2001). The enormous contrast between needing information and having actual data, has stablished all the parameters needed for a forward data analysis and the creation of an establishment responsible for satisfying these needs (BID, 2013). Under those circumstances, many different countries have decided to stablish a logistics observatory to attend these necessities (BID, 2013). Indeed, several countries in the Latin American region are currently in the process of creating their logistics observatories or have already stated the willing to do so (BID, 2013).

Ecuador is one of the countries in the Latin American region developing a logistics observatory (BID, 2013) since there is an exponential growth of population which is directly proportional to mobility needs and urban logistics development (MDMQ, 2014). By the same path, Ecuador has found the need to create different entities in charge to deal with all logistics aspects caused by urban population increase. Specific, in Quito, according to the mobility diagnosis done by the Municipio Metropolitano de Quito for the Metropolitan Territory Development Plan (PMOT), if socio economic conditions are maintained in the city of Quito, Ecuador, when 2030 year is reached the city vehicle fleet would reach 1 150 000 vehicles approximately creating an unsustainable situation for city vehicular mobility (MDMQ, 2014). By the same hand, a various quantity of typical set of policies were applied hopping to mitigate and control mobility congestion (MDMQ, 2014). However, these policies, including vehicle access restriction, vehicle mobility restrictions and other regulations, are not often completely evaluated before their implementation, in consequence, an inefficient mobility inside urban areas are created and commercial strategies are affected (Greaves and Figliozzi, 2008). By these consequences, the need of creating sustainable public mobility plans and other different commercial strategies for the private sector for these reasons data collection is vital in order to have a more in-depth analysis (Laranjeiro, 2019).

Under those circumstances, the project of creating the Urban Freight Transport Observatory (OTUC) is stablished in Quito, Ecuador. Moreover, OTUC's main functions are orientated to the evaluation of existing available data such as geospatial urban freight data, collect and disseminate existing information, among other duties (MDMQ, 2014). At the moment two companies' fleet geospatial data are available, with a total of approximately 4.5 million records per month, for forward analysis. Filling the short-term need of analyzing these datasets to get the number of detentions and assign a detention classification, either delivery detention, traffic detention, etcetera, in order to get specific information about gas consumption, greenhouse gases emission and possible potential applications such as public mobility plans. (BID, 2013).

In addition, all datasets provided have 12 common features, the most important used in our study are: date and hour, longitude, latitude, velocity and car label. For Big Data tools to be applied, datasets must meet the four main characteristics of Big Data, which are volume, velocity, variety and veracity (Hiraharan, 2004). This characterization has opened the possibility for applying data science tools in order to determine a certain predictive model that recognize and classify records into specific clusters with same detention characteristics creating a potential output for urban logistic applications. These artificial intelligence models are created through the application of machine learning (Hess, 2015).

In doing so, the present research proposes the development of an artificial intelligence model able to identify and classify each detention from geospatial urban freight datasets. Second, the presentation of results that showcase the different detention's classification as an important input for future potential applications like public mobility policy and to give valuable insights for private companies and help them to modify commercial strategies that will make urban logistics more efficient (Romano, 2019). Finally, the analysis of the detention's obtained with detailed insights such as variation through week, schedule congestion along the historical time, and others. This analysis demonstrates the potential value of a geospatial data analytics for different diagnosis of urban logistics and planning, standing from the perspective of a city lever, providing a macroscopic view of urban logistics activities of the datasets studied (Greaves and Figliozzi, 2008). It's worth saying that if more geospatial datasets would've been available from different commercial sectors, a better understanding of how urban logistics for commercial corporations could've been reached since by the day we only have datasets from just two commercial sectors. However, the standardization created by the model for the datasets gives the possibility of using an artificial intelligence model in future investigations with different commercial sector's datasets. In addition, the main purpose of the automation of the artificial intelligence model is to help future analysis of huge GPS datasets.

LITERATURE REVIEW

Global Positioning system

Global Positioning System, known as well as GPS, is a space-based system for positioning, navigation, and timing created by the Department of Defense of the United States of America. Developed with the main objective of merging synergistic Navy and Air Force programs for timing and space-based navigation. Taking the advantages of the system, it has been used for many applications in different areas such as telecommunication, emergency services, electronic commerce, electrical power distribution, transportation among others (McNeff, 2002).

Conveniently, this data was never left unrecorded, indeed data about transport movement has been collected using a bunch of different static devices like, road detectors or sensors measuring traffic flow measures. The problem of collecting data with any GPS, is the uncertainty and bias generated by unprecise and inaccurate instruments. However, data acquisition changed in an exponential way with the availability of the GPS devices allowing to measure traffic data as records in sequences of discrete positioning signals transmitted by the cars' GPS (Spaccapietra et al., n.d.).

GPS data devices record navigation information such as latitude, longitude, speed, heading, and altitude, some of them could give more specifics alerts like engine status of a vehicle, in addition the personalization that a GPS has, allows to go through different settings, for example, data can be recorded every minute, every second, etc as long as the GPS receives signal from the corresponding satellites. (Du and Aultman-Hall, 2007). Above all, the erroneous calibration or high-rise buildings or structures, especially in big cities, are obstacles for GPS signals that could give less accurate information. (Schuessler and Axhausen, 2009).

Fulfilling the characteristics like large volume and by-product nature, GPS information can be classified as big data. However, methodology applied to the processing and analysis geospatial of data for their eventual application into the freight transportation modelling does not have a deeper exploration. (Romano et al., 2019). Despite the sheer volume and granularity of data, this opens a new opportunity for extensive analysis and data mining of a completely different nature between geospatial data and other kind of data (Hariharan and Toyama, 2004).

Related Topics

Many companies around the world use Global Positioning System to self-vehicle monitoring (Spaccapietra et al., n.d.). Even though a huge volume of data collected is available, just a few companies use geospatial data in order to obtain useful metrics and generate different efficient applications. And fewer agree to share these data to neither organizations nor observatories for further analysis (Romano, 2019). As a matter of fact, there is a lack of available literature about identifying and analyzing freight vehicles' detentions from geospatial data. However, researches have made a good use of information available and have used it to create new models of the freight system. (Sharman and Roorda, 2011).

From GPS data available, Greaves and Figliozzi developed an algorithm able to identify when the urban freight fleet from commercial vehicles in Melbourne stops (2008). The algorithm works by analyzing the difference between consecutive GPS location points to determine if a vehicle stopped ever (Greaves and Figliozzi, 2008). The study proposes a parameter of 240 seconds as a correct threshold to flag a group of records as a stop. In the same way, the investigation states that the geographic distance between locations of a vehicle at consecutive geospatial points was equally important (Greaves and Figliozzi, 2008). Finally, it is important to mention that the accuracy rating of the GPS in this study is given as 6 meters. Therefore, if a vehicle had moved more than the accuracy rating of the device it was flagged as invalid and dropped from calculations (Greaves and Figliozzi, 2008).

In the other hand, Xia Yang uses a Support Vector Machine (SVM) to identify delivery stops applied to Global Positioning System datasets in the city of New York (2014). Developers created an algorithm which segregates freight fleet stops into two groups, delivery stops and non-delivery stops. In addition, Yang use three classification features: stop duration, distance from a stop to the center of the city, and a binary distance to the closest major bottleneck like a toll booth or a tunnel (2014). It was noticed that the current minimum delivery time has a variation between ten minutes and twenty minutes. In addition, the study presents a specific threshold speed of fourteen kilometers per hour, using this parameter, researchers were able to determine when the vehicles stop, giving a major probability to detect all stops. In other words, the algorithm identifies as a stop the 0 km/h and 14 km/h speeds since, in time difference terms, this difference is approximately 30 seconds and is not statistically representative between the 10- and 20-minutes threshold

Whereas Sharman and Roorda propose an algorithm able to record delivery stops if and only if trucks' dwell time exceeded 180 seconds (3 minutes). Equally important, insignificant truck movement was not reported because it was considered as a fluctuation of the GPS when a truck idles. In consequence, it was decided that the record was removed only if distance between two consecutive data points was less than 20 meters (Sharman and Roorda, 2011).

Detentions by itself could give a few or no information about urban freight behavior and urban mobility. In consequence, the process of clustering is needed in order to get significant insights about urban logistics. Clustering is the process of grouping large data sets by similar characteristics or features in order to differentiate records from one another. (Briant 2006). In addition, clustering methodology uses partition of objects' sets into groups such that objects in the same group have more similar features to each other than objects in different groups.

Cluster analysis is a major tool in many areas dedicated to the research based in high volume databases, it is commonly used in the field of engineering and scientific applications. Not only clustering is used for grouping data records according to similar characteristics, but it is useful when data segmentation, discretization of continuous attributes, data reduction, outlier detection, noise filtering, pattern recognition and image processing is wanted and needed (Birant and Ktu, 2006).

K-Means

Clustering of unlabeled data can be performed with multiple clustering algorithms. Moreover, considering the scenario where the classification of detentions is based on the time in detention of a vehicle; K-Means is the perfect use case. The model works efficiently with flat geometry variables and it allows for a very large scalability (Clustering, n.d.).

K-means algorithm has the main objective to find optimal solutions between the record and the clustering by associating a number of records to a specific cluster and grouping them between others that are placed within the stablished Euclidian distance (Morissette and Chartier, 2013). In other words, it is a point-based clustering method that begins when the initial cluster center is placed at an arbitrary position and then proceed to move at each step the cluster center in order to eradicate the clustering error (Aristidis et al., 2003).

However, the most important disadvantage of the method is its sensitivity to cluster center's initial position (Zheng and Wang, 2013). In addition, Bryce Sharman states that data processing in this specific model supports to logistic decision components because, unlike most other models, k-means is intended to represent vehicle routing and building up schedules over long periods of time to reflect daily, weekly, etc (2011).

Hierarchical Density Based Spatial Clustering Application with Noise (HDBSCAN)

GPS positional data (latitude and longitude) behaves as a non-flat geometry variable, hence a model based on density is the best approach in an effort to cluster GPS data based on the position of the vehicle (Clustering, n.d.). HDBSCAN model presents itself as a great option in a scenario where uneven clusters of different shapes are needed. More importantly, the model is able to discriminate noise from signal. This means that not necessarily all the GPS data entries will be assigned to a cluster.

HDBSCAN model was developed as a major update to DBSCAN model. HDBSCAN's main objective is to create a flat cluster having the assumptions that there is noise in the data, in other words, the data obtained as an input is not one hundred precise (McInnes et al., 2017). What HDBSCAN does is, group by cluster all the points that fall in between the hyperparameter called ε (Epsilon) is the distance from the center to the border of the cluster (Malzer, 2019). HDBSCAN uses a manually minimum value of Epsilon given by the programmer and learns from specific features to reach a maximum distance and group the points found in the area. These points will be grouped in the cluster if and only if they share common features, if not, those will automatically fall in other cluster where its features are shared in common (Rahman et al., 2016).

METHODOLOGY

After doing a meticulous research about different coding and machine learning project development methodologies, a hybrid combination between Comendador's Methodology and CRISP-DM methodology is stablished for the algorithm and artificial model development. According to Julio Comendador's Methodology in the paper "A GPS analysis for urban freight distribution" (2012) the first step is to define every general tool that could be applied for the algorithm development. For this study, the tools found and classified were the geospatial datasets, coding literature and coding software for data cleaning and code construction.

In the other hand, CRISP-DM methodology analyzes the use of data mining techniques in order to build up an artificial intelligence model for a specific purpose (Wowczko, 2015). This methodology groups every task to be done into six concatenated phases:

1- Business Understanding

- To understand why the project or study is important for the business and how it will be useful for the different stakeholders.

2- Data Understanding

- To understand the different dataset's features, what information they provide and the usefulness of the records.

3- Data Preparation

- To use different tools in order to drop useless data, outliers and standardize the different datasets in order to use them as input for the artificial intelligence model.

4- Modelling

- To determine the data mining task, choose data mining technique and use algorithms to perform the tasks.

5- Model Evaluation

- To analyze the accuracy and precision of the results given by the model deployment and to fix any error of the model, and to interpret the results of the model.

6- Deployment

- To integrate the model into the different operational systems and running it on real record to produce useful and valuable information for decision making.

As a result, the following methodology is proposed in order to achieve the objective of the study.

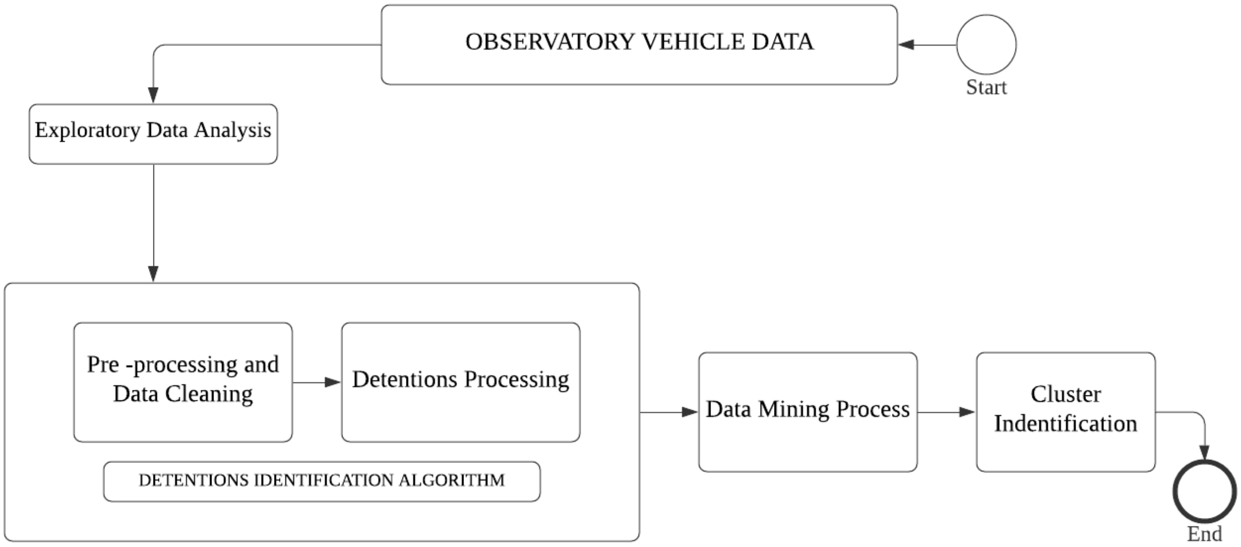


Figure 1. Research Methodology

EXPLORATORY DATA ANALYSIS.

The first step before processing and modelling data is to understand and analyze the different characteristics and information of the different geospatial datasets obtained through different data analytics tools and software such as Tableau, QGIS, Kepler Gl, between others (Yu, 2020). Then, the Exploratory Data Analysis is started, for the present study five geospatial datasets are studied. Each dataset belongs to a specific month from September 2019 to January 2020 giving us a total of 25,475,082 records. Also, it is found out that neither the number of records per car or the number of records per month aren't the same through time, as shown in Figure 2.

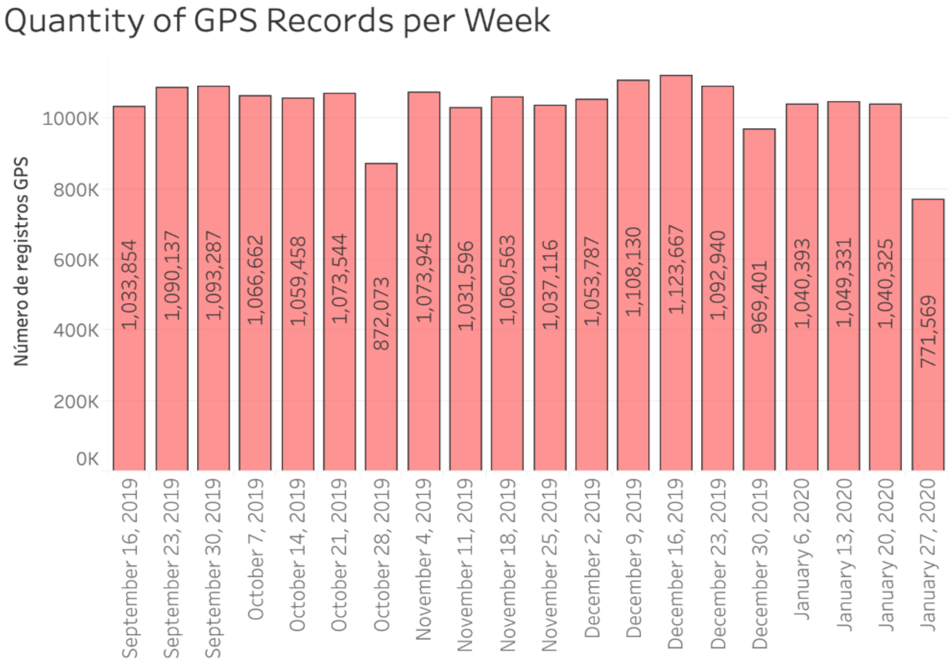


Figure 2. Quantity of GPS Records per Week.

The 25 million records show an average standard deviation of 3,462 records per week. The EDA gives that in the different GPS datasets a record is registered every 12 seconds in average, where it reaches a maximum of 14.7 seconds between 6 pm and 6 am, and a minimum of 8.3 seconds between 6 am and 10 am, as shown in Figure 3.

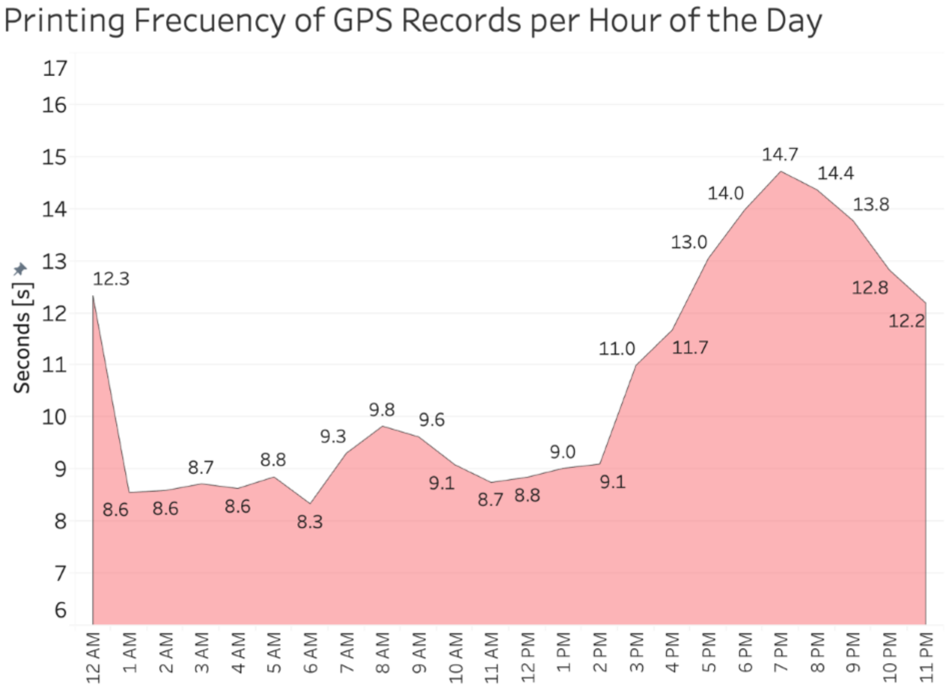


Figure 3. Printing Frequency of GPS Records per Hour of the Day

The next step is the analysis of the key variable and other target variables for the detention algorithm development (Laranjeiro, 2019). The following variables are presented as target variables and Car Velocity [km/h] as the key variable:

- Vehicle: Car (Total of 83 vehicles in the fleet)
- Date-Hour: Record printed into the GPS dataset in date and hour format.
- Latitude: Is expressed in angular measures that vary from 0º of the equator to 90ºN (+90º) of the North Pole or 90ºS (-90º) of the South Pole
- Longitude: It shows the East or West direction from the reference meridian 0º, expressed in angular measurements from 0º to 180ºE (+180º) and 180ºW (-180º).
- Vehicle Velocity: Velocity of the vehicle [km/h]

In order to get clean data, first, the aggregation of average velocities by Hour of the Day, Day of the Week and Day of the Month. Then, the analysis of the key variable gives as a result that between 7 pm and 5 am most of the variable result is zero. This data behavior is considered as normal since there are not any delivery out of work schedule. Also, the average velocity per vehicle decreases in the most concurred hours of the day, such as 8 am and 5 to 6 pm and on Sundays the velocity tends to zero throughout the day, also considered as a normal behavior since almost every vehicle does not have any activity on Sundays. It is important to bring out that this particular behavior in the days of the week is similar through all the studied months as shown in Figure 4.

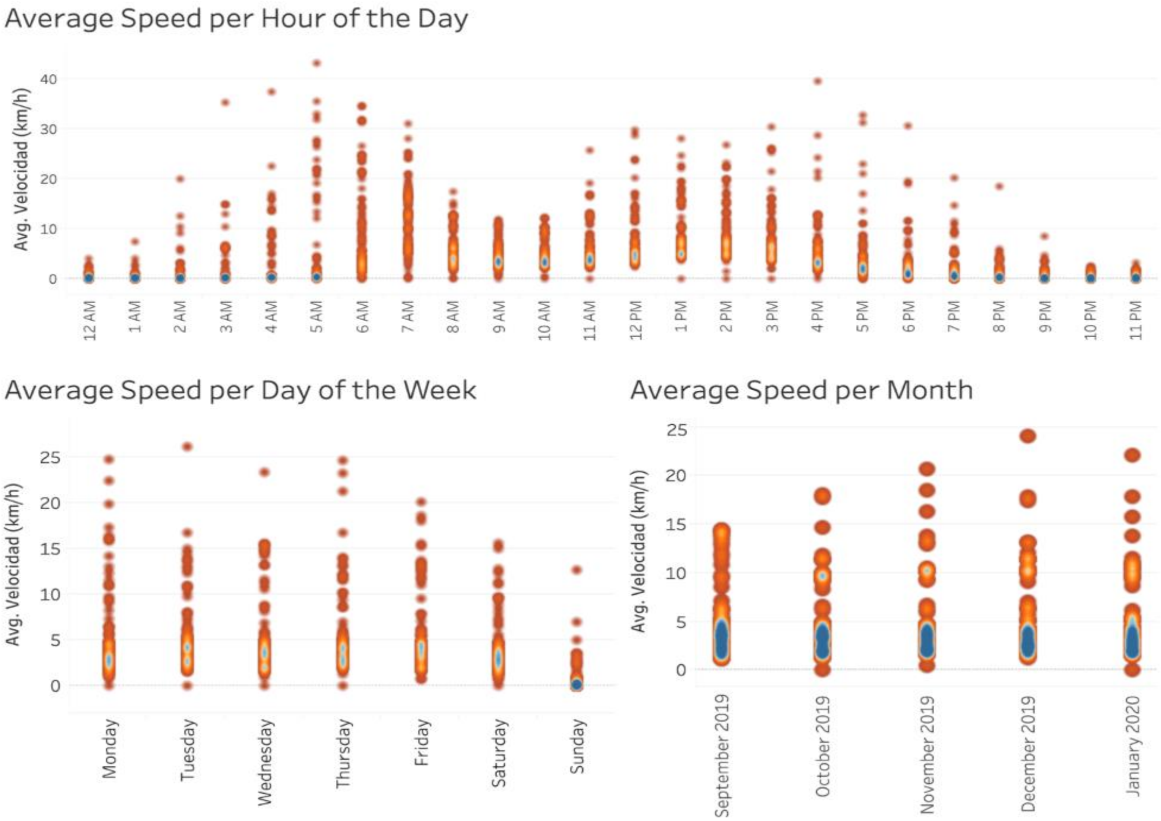


Figure 4. The analysis of the key variable [Speed].

DATA PRE-PROCESSING AND CLEANING

Before we get to the algorithm and model development, data pre-processing and cleaning steps are needed in order to guarantee that only useful information gets to the next stage (Laranjeiro, 2019). Therefore, the following group of concatenated steps has been applied:

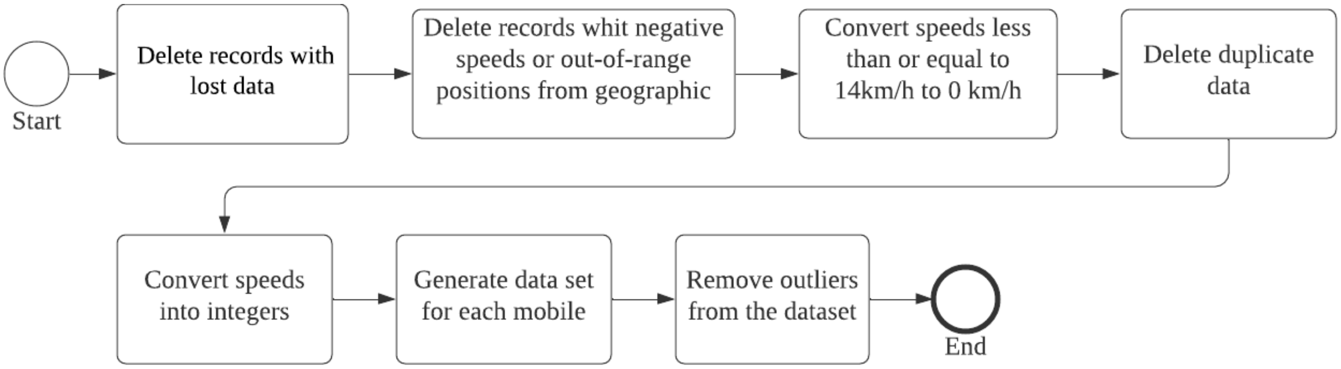

Start
Delete records with lost data
Delete records whit negative speeds or out-of-range positions from geographic
Convert speeds less than or equal to 14km/h to 0 km/h
Delete duplicate data
Convert speeds into integers
Generate data set for each mobile
Remove outliers from the dataset
End

Figure 5. Data pre-processing and Cleaning Flow.

1. Delete records with missing data since it is labeled as a GPS' signal lost or variability.
2. There is no negative velocity value used in a vehicle day to day so those records are dropped. Also, records with latitude and longitude values out of range are eliminated.
3. Next, records that have a velocity value less than or equal to 14 kilometers per hour are assumed as 0 kilometers per hour and considered as a short detention.
4. Then, duplicate data records are dropped so a unique record is kept.
5. After that, velocity values are rounded to the nearest integer.
6. Sixth, a dataset for each vehicle is created in order to analyze every vehicle separately.
7. Finally, data out of the third quartile is removed since those are considered as outliers.

After finishing with the pre-processing and cleaning data procedure it is found out that there are either single records with missing data as shown in Figure 6 or negative velocities or out of range values from latitude or longitude features nor duplicate records. In the other hand, the procedure allowed to identify multiple GPS' signal lost, this signal lost was identified specially at midnight when a record was printed with a velocity greater than zero whereas previous and following records showed velocity of zero. These records were labeled as zero kilometers per hour. Finally, the Tukey interquartile range method let detect and eliminate outliers for the upper boundary (Conagin et al., 2008).

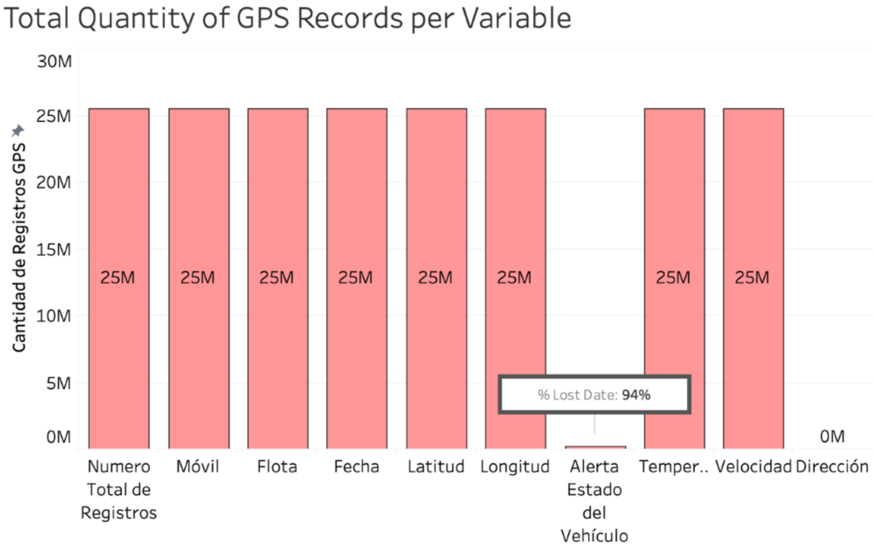


Figure 6. Total Quantity of GPS Records per Variable.

Once assured the data was completely pre-processed, the following step was to transform the velocity variable data type from a float variable to an integer variable to make the algorithm processing more efficient (Beauchamp and Olson, 1973). In order to have a smoother method for this variable, each vehicle was processed independently by creating a single dataset for every vehicle.

DETENTIONS IDENTIFICATION ALGORITHM

In the previous step to classify detentions, it is needed to get the actual value of the detention measured in a time variable, in this case, minutes. Moreover, the huge amount of data available in the study forces to look forward automatization so a detention identification algorithm is developed. The algorithm works automatically once the dataset input is give, it follows a group of concatenated steps in order to get the detention length. The algorithm works in five divided steps that follows as it is shown in figure 7.

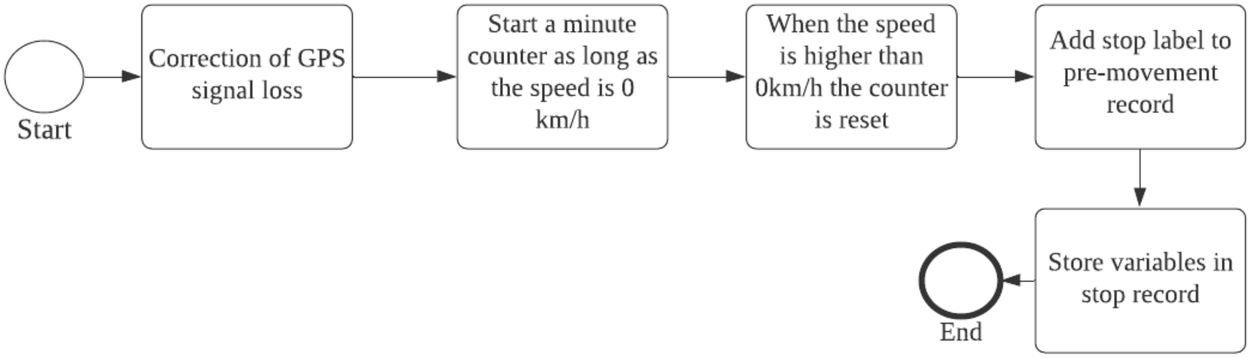

Start
Correction of GPS signal loss
Start a minute counter as long as the speed is 0 km/h
When the speed is higher than 0km/h the counter is reset
Add stop label to pre-movement record
Store variables in stop record
End

Figure 7. Data Algorithm Flow

1. There is a natural uncertainty of a 20 meters radius from de GPS signal that gives a point with a velocity equal to zero different values of latitude and longitude as if it was moving (Zhao et al., 2013). The algorithm corrects this position and change de position of the record to the previous record's position in order to eliminate this uncertainty.
2. All dataset records are first sorted from oldest to newest. Then, the algorithm identifies the vehicle's velocity feature into the datasets so in case the velocity is greater than zero, the algorithm starts a counter responsible of measuring the difference

between the actual time record (n) and the time record (n-1), it is important to clarify that this time difference is accumulated as long as the velocity is 0 kilometers per hour.

3. Then, if and only if the velocity is greater than zero, the counter resets and the difference of time obtained is recorded as a detention.
4. By the same hand, if the velocity record (n) is equal to 0 kilometers per hour and the next velocity record (n+1) is greater than 0 kilometers per hour, a detention label is assigned on the register (n).
5. Finally, over the (n) register labelled as detention, the algorithm executes a “fill forward” that allows to save the position (longitude and latitude) where the detention started, and the total time accumulated during the detention.

After the algorithm is run, the file is filtered so we can work only over the detentions obtained and a dataset file is saved automatically. This dataset contains the full records of the detentions the algorithm gives, and it is characterized to have the following features: Vehicle, Date, Detention Initial Hour, Start of the Detention Position (Latitude and Longitude), Detention Final Hour, End of Detention Position (Latitude and Longitude) and, the most important, Time in Detention. The combination of these features is needed as an input for the Detention Classification Model.

DETENTIONS CLASSIFICATION MODEL

The adaptation of the CRISP-DM methodology that is stablished for the present research development into data mining is the determination of the classification of the detentions obtained from de Detention Identification Model based on the Time in Detention feature. In this step is required to analyze and select what kind of artificial intelligence model is going to be used based on the features and characteristics the input dataset has.

After analyzing all features obtained from the input dataset, it is obtained that there is an absence of a label that allows us to build any supervised machine learning model (Gira et al., 2005). Therefore, based in the features of longitude and latitude and time in detention the selection of K-Means and HDBSSCAN non-supervised models are applied.

K-Means Model

According to Bryce Sherman (2011), one of the most efficient models for cluster classification of vehicle detentions based on the time in detention is the non-supervised clustering model K-Means. Before the model is built, the analysis of the variable time in detention is needed (Greaves and Figliozzi, 2008). This analysis shows that the variable results have a bias that tend to the right as shown in Figure 8.

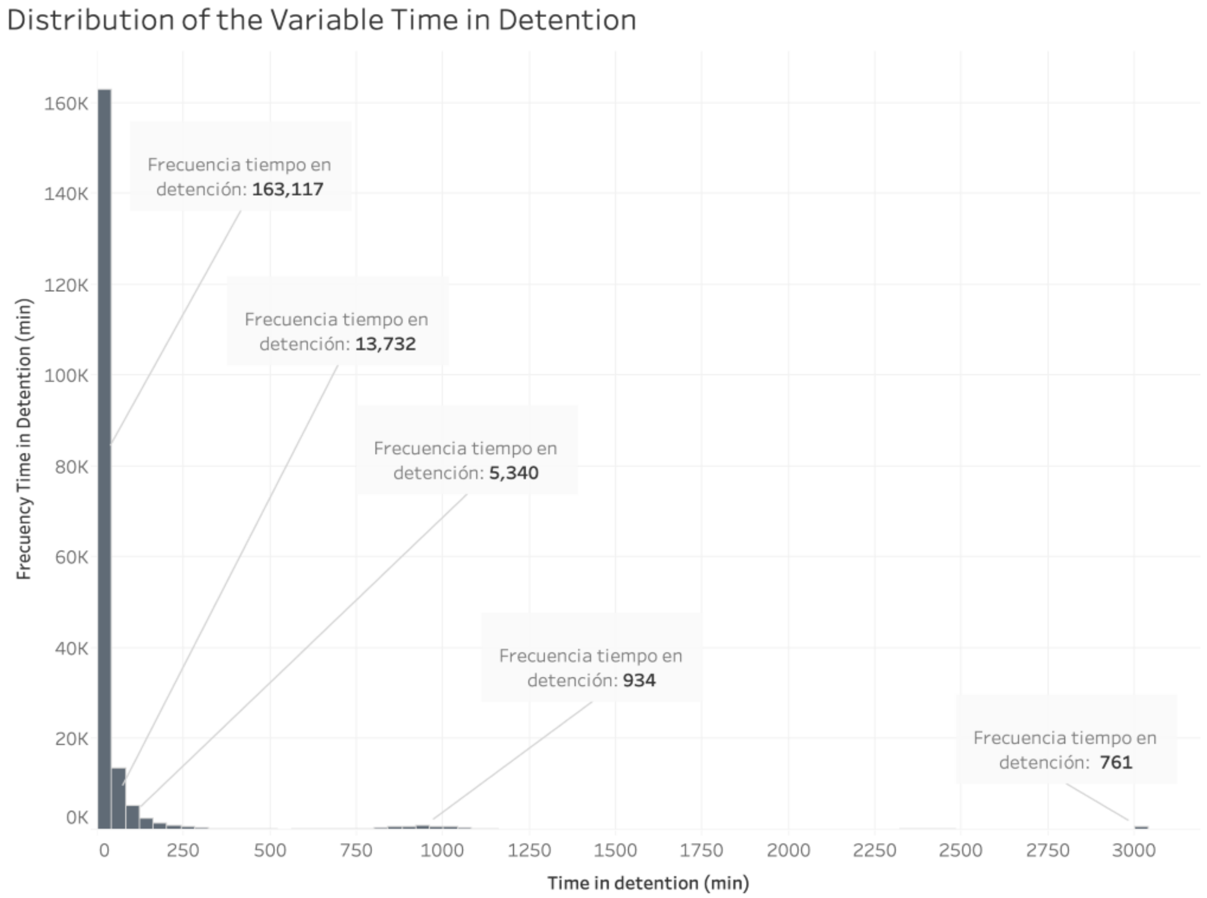


Figure 8. Distribution of the Variable Time in Detention.

However, it is common since a great number of detentions are short and recurrent as if it was urban traffic. According to Laurence Morissete and Sylvain Charteir (2013) if K-Means model is used, the variables used must be continuous and have a normal distribution. To deal whit this problem John Beauchamp and Jerry Olson (1973) recommend a normal distribution a base 10 logarithm transformation over the data of the variable of study, this new variable is called Time in Detention and as a result a new feature based on the vehicle's time in detention logarithm measured in minutes is created as shown in Figure 9.

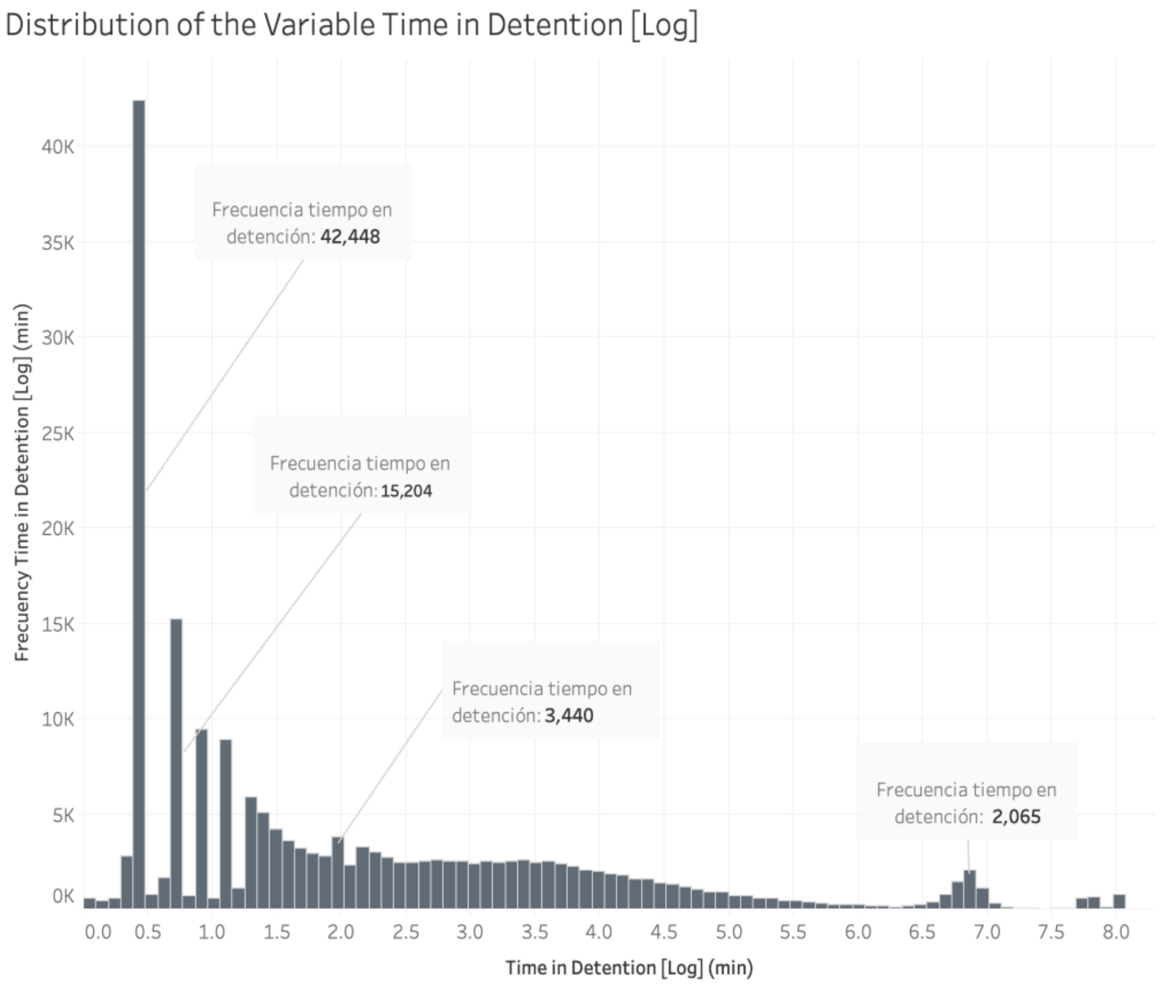


Figure 9. Distribution of the Variable Time in Detention Logarithmic Transformation.

In continuance, the iterative process of executing the model using multiple hyperparameter combinations of the model is started until the model by itself finds an optimal solution (Morissete and Charteir, 2013). For the K-Means model the K hyperparameter was changed with different values, the K value is the number of clusters desired. Looking forward to obtaining an automatically data flow for model execution, the Silhouette Score tool was used to obtain the optimal value of K (Wang et al., 2017). Despite software and time limitations a range between 3 and 6 clusters was defined for more processing efficiency.

Once the K range is established the K-Means model evaluates 4 iterations for each vehicle and identifies the highest score of the different iterations and this one is then selected and defined as the optimal solution for this particular vehicle. The algorithm works automatically and repeats this process for each and every vehicle dataset that is given as input to de model.

HDBSCAN Model

After running the K-Means model a variable labeled with the number of clusters it was associated was created for each dataset evaluated in the previous model. Once the variable is created, the aggregation inside each cluster for every vehicle dataset is stablished with the main objective of finding effective aggregated metrics based on the detention's geo-localization and the average time in detention measured in minutes.

In order to find the effective aggregated metrics, the HDBSCAN model is applied as a data mining technique (Birant and Kut, 2007). The HDBSCAN model uses only the detention's geo-localization as a variable. In other words, the model uses de longitude and latitude to find optimal clusters grouped by position. Considering Jianhe Du and Lisa Aultman-Hall (2007) recommendation this model was independently applied for each cluster and vehicle's dataset so the model by itself could group the detentions that share common characteristics based on geo-localization and uses a set of hyperparameters that find the most and best quantity of possible clusters as shown in Figure 10.

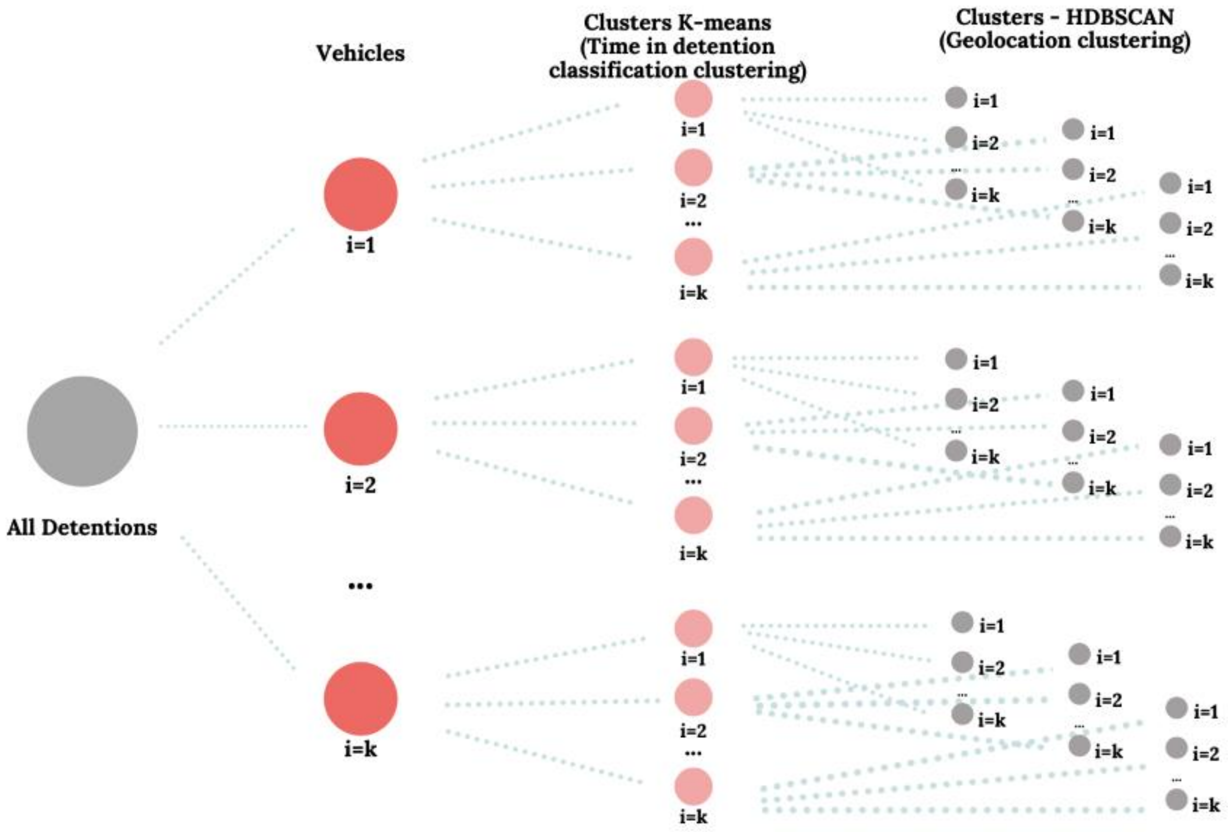

Vehicles
Clusters K-means (Time in detention classification clustering)
Clusters - HDBSCAN (Geolocation clustering)
All Detentions
i=1
i=2
...
i=k

*Figure 10***.** Identification Model Flow.

The fixed hyperparameters used in the model development were:

- Min_Cluster_Sizw = 2
- Cluster_Selection_Epsilon = 0.0001
- Min_Samples = None

The HDBSCAN model gives as a result the segregation for each of the vehicle's detention cluster into different geo-localization clusters according to specific characteristics and features. Once the model gets the final number of segregated clusters for each vehicle dataset, the type of detention label is assigned to each cluster obtained.

STUDY RESULTS AND POTENTIAL APPLICATIONS

Study Results and Potential Applications

Once the Detentions Identification Model run, the following results were given:

Total Number of Detentions	Total Number of Vehicles	Average Quantity of Detentions per Vehicle
199,111	82	2,429

Table 1. *Identification Algorithm Results*

The identification model consolidates all vehicles detentions in a single dataset and sort the Date-Hour variable from oldest to the newest. This dataset will be the input for the classification models.

The following results show an example of the classification models deployment for the vehicle Q02. The first model implemented is the K-Means model. It allows to have the optimal number of clusters based in the variable Time in Detention [min]. For each cluster then the HDBSCAN model was deployed. This model allows to cluster detentions depending on how close they are from each other considering their geolocation (Latitude, Longitude) in the map. This allows to aggregate the K-Means metrics to reduce the number of errors and get a better sense of how the detentions are occurring in the map.

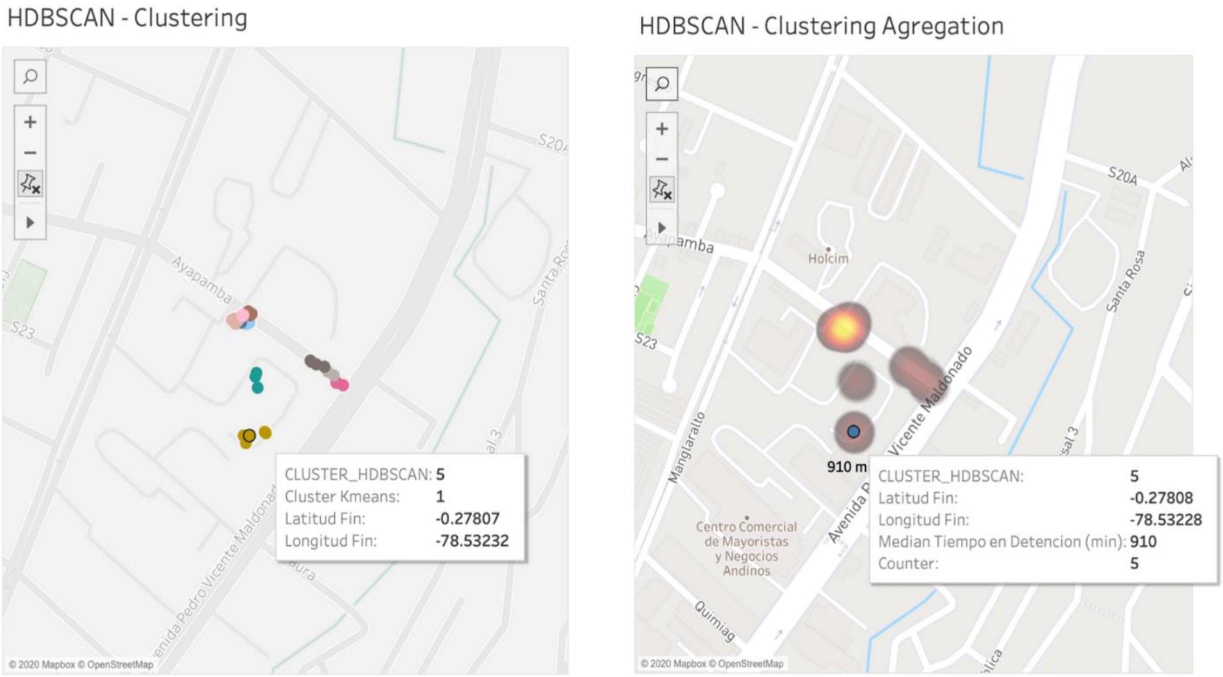
The optimal number of K-Means clusters for this vehicle detentions is four. In continuance, each cluster results are analyzed and exposed in detail.

In the Table 2, cluster 1 is analyzed and it is given that is possible to see that this cluster of detentions reports an average of 1,330 minutes, 268 total detentions and 22 HDBSCAN clusters.

These results indicate that this detention cluster resembles to a long duration detention. In the Figure 11 we can see how the detentions where the vehicle stay over the night are being captured under this “long duration detention” K-Means cluster. Then the HDBSCAN aggregates this detention depending on how close they are form each other. It is possible to see in the picture for the Cluster HDBSCAN 5 a median of 910 minutes in detention.

Cluster K-Means - 1	
Avg. Time in Detention [min]	1,330
Median Time in Detention [min]	1,006
Count of Detentions	268
Number of HDBSCAN Clusters	22

Table **2**. *K-means Cluster 1 Results.*



HDBSCAN - Clustering

CLUSTER_HDBSCAN: 5
Cluster Kmeans: 1
Latitud Fin: -0.27807
Longitud Fin: -78.53232

HDBSCAN - Clustering Agregation

CLUSTER_HDBSCAN: 5
Latitud Fin: -0.27808
Longitud Fin: -78.53228
Median Tiempo en Detencion (min): 910
Counter: 5

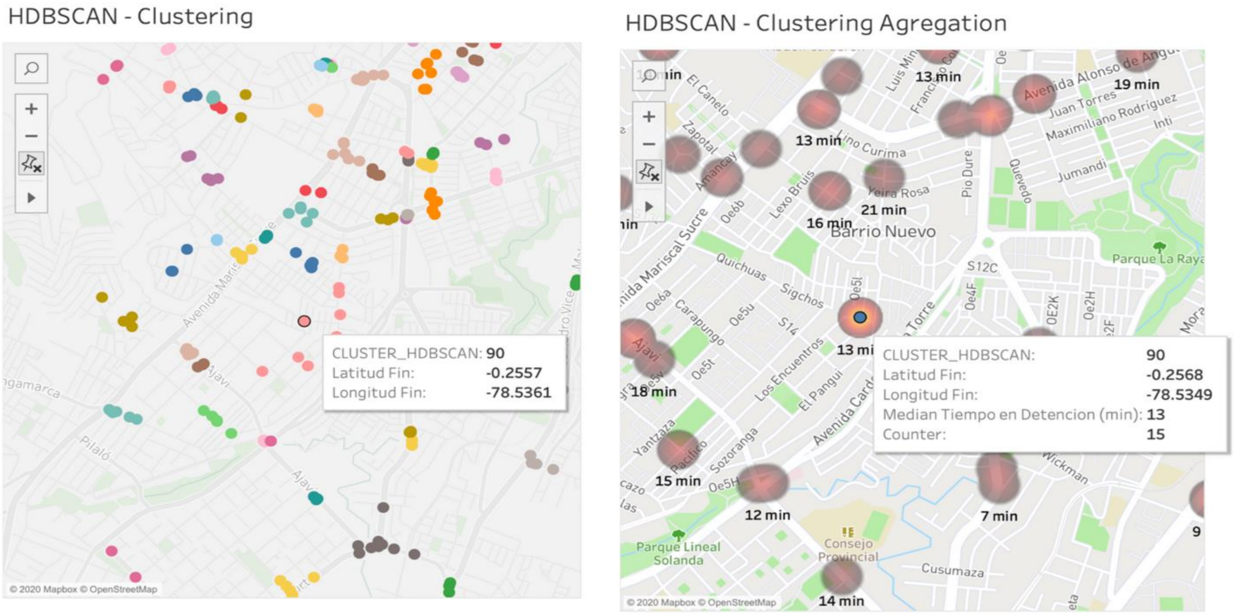
Ayapamba
S23
S20A
Santa Rosa
Avenida Pedro Vicente Maldonado
© 2020 Mapbox © OpenStreetMap
Holcim
Manglaralto
Vicente Maldonado
Avenida
910 m
Centro Comercial de Mayoristas y Negocios Andinos
Quimiag
© 2020 Mapbox © OpenStreetMap

*Figure 11***.** HDBSCAN Cluster Results.

In the Table 3, cluster 2 is analyzed and showed that it is possible to see that the cluster of detentions reports an average of 13 minutes, 519 total detentions and 119 HDBSCAN clusters. These results indicate that this detention cluster resembles to a short delivery detention. In the Figure 12 it is possible to see how the detentions the detentions that are close to a commerce are being captured under this "short delivery detention" K-Means cluster. Then the HDBSCAN aggregates this detention depending on how close they are form each other. It is possible to see in the picture for the Cluster HDBSCAN 90 a median of 13 minutes in detention.

Cluster K-Means - 2	
Avg. Time in Detention [min]	13
Median Time in Detention [min]	12
Count of Detentions	519
Number of HDBSCAN Clusters	119

Table **3**. *K-means Cluster 2 Results.*



CLUSTER_HDBSCAN: 90
Latitud Fin: -0.2557
Longitud Fin: -78.5361

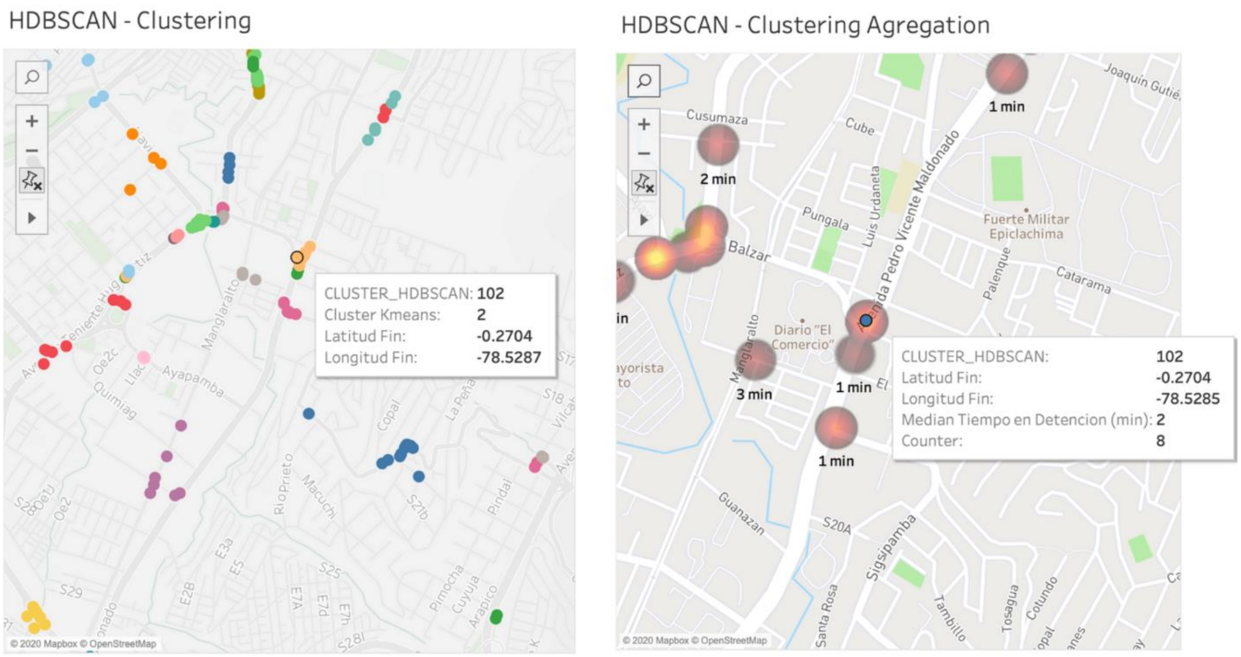
CLUSTER_HDBSCAN: 90
Latitud Fin: -0.2568
Longitud Fin: -78.5349
Median Tiempo en Detencion (min): 13
Counter: 15

*Figure 12***.** HDBSCAN Cluster Results.

In the Table 4, cluster 3 is presented and its results shows that is possible to see that the cluster of detentions reports an average of 2 minutes, 626 total detentions and 141 HDBSCAN clusters. These results indicate that this detention cluster resembles to a detention in traffic. In the Figure 13 it is possible to see how the detentions in the middle of the transit, avenues and streets are being captured under this "detention in traffic detention" K-Means cluster. Then the HDBSCAN aggregates this detention depending on how close they are form each other. It is possible to see in the picture for the Cluster HDBSCAN 102 a median of 2 minutes in detention.

Cluster K-Means - 3	
Avg. Time in Detention [min]	2
Median Time in Detention [min]	1
Count of Detentions	626
Number of HDBSCAN Clusters	141

Table 4. *K-means Cluster 3 Results.*


HDBSCAN - Clustering

CLUSTER_HDBSCAN: 102
Cluster Kmeans: 2
Latitud Fin: -0.2704
Longitud Fin: -78.5287

HDBSCAN - Clustering Agregation

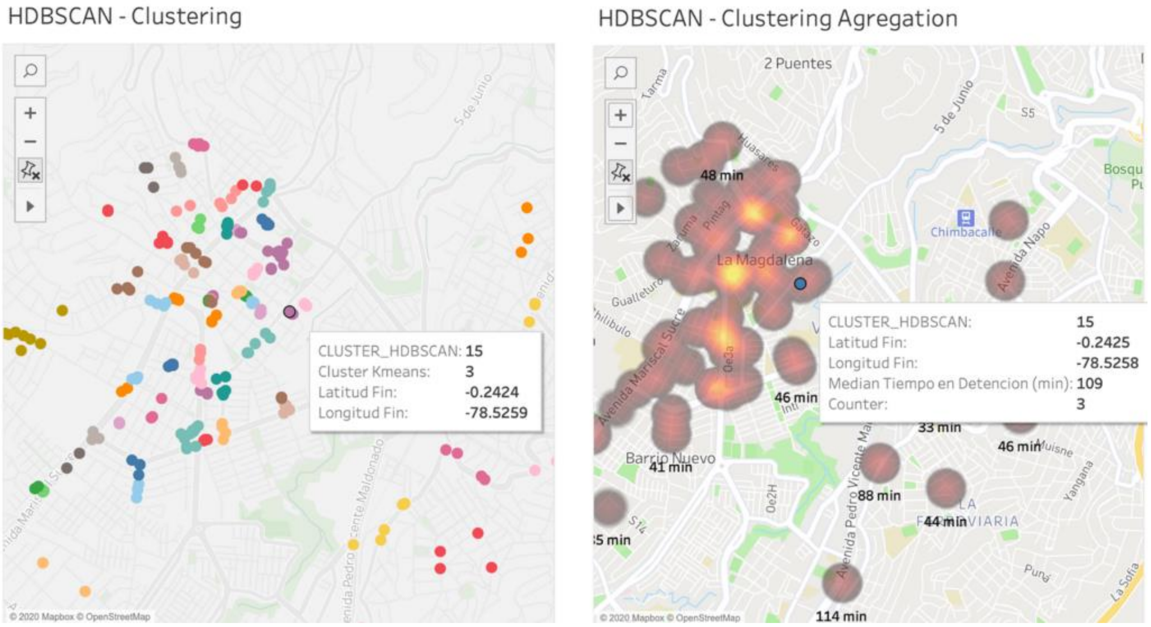
CLUSTER_HDBSCAN: 102
Latitud Fin: -0.2704
Longitud Fin: -78.5285
Median Tiempo en Detencion (min): 2
Counter: 8

Figure 13. HDBSCAN Cluster Results.

Finally, in the Table 5 it is possible to see that the cluster 4 of detentions reports an average of 75 minutes, 430 total detentions and 102 HDBSCAN clusters. These results indicate that this detention cluster resembles to a long delivery detention. In the Figure 14 it is possible to see how the detentions close to bigger commerce are being captured under this "long delivery detention" K-Means cluster. Then the HDBSCAN aggregates this detention depending on how close they are form each other. It is possible to see in the picture for the Cluster HDBSCAN 15 a median of 109 minutes in detention.

Cluster K-Means - 4	
Avg. Time in Detention [min]	75
Median Time in Detention [min]	59
Count of Detentions	430
Number of HDBSCAN Clusters	102

Table 5. K-means Cluster 4 Results.



HDBSCAN - Clustering

CLUSTER_HDBSCAN: 15
Cluster Kmeans: 3
Latitud Fin: -0.2424
Longitud Fin: -78.5259

HDBSCAN - Clustering Agregation

CLUSTER_HDBSCAN: 15
Latitud Fin: -0.2425
Longitud Fin: -78.5258
Median Tiempo en Detencion (min): 109
Counter: 3

Figure 14. HDBSCAN Cluster Results.

Table 6 shows the comparison between the four clusters found. It is important to emphasize the significant difference between them in terms of the metrics average time in detention, number of detentions and number of HDBSCAN clusters.

Figure 15 shows the relationship between these variables. For the detentions in transit which have a very small average time in detention, the number of detentions and the number of HDBSCAN clusters is higher. On the other hand, for long duration detentions the number of detentions and the number of HDBSCAN clusters is much lower. This makes sense as long duration detentions are less recurrent and usually occur at specific locations. While in transit detentions are much more recurrent and occur at different locations.

General Metrics	**Cluster K-Means - 1**	**Cluster K- Means - 2**	**Cluster K-Means - 3**	**Cluster K-Means - 4**
Avg. Time in Detention [min]	1,330	13	2	75
Median Time in Detention [min]	1,006	12	1	59
Count of Detentions	268	519	626	430
Number of HDBSCAN Clusters	22	119	141	102

Table 6. *Clustering Results Summary.*

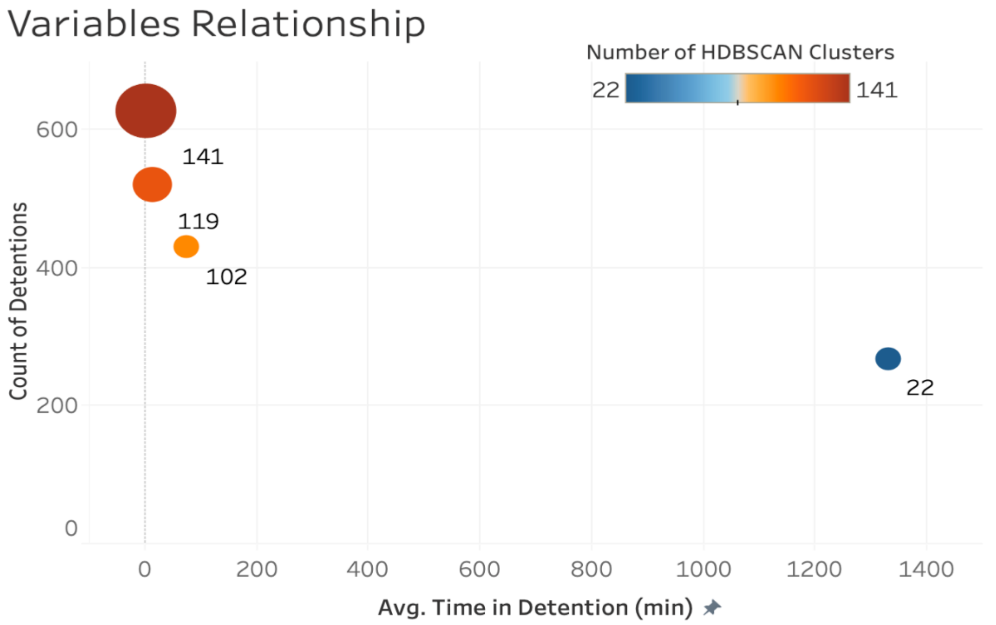


Figure 15. HDBSCAN Cluster Results.

The clusters’ result presented for both of the non-supervised models developed all along the study have a precision over the sixty percent when applying the Silhouete Score tool. For the K-Means model it was obtained a 62 percent of precision whereas for the HDBSCAN model it was obtained a 64 percent of precision. It is necessary to emphasize that the Silhoute Score tool doesn’t evaluate the precision between the clusters and the real detention but it evaluates the precision between how the record’s characteristics resembles to the cluster’s characteristics ((Morissete and Charteir, 2013).

After analyzing all the results obtained from the different stages of the study, a bunch of potential applications are possible if companies or researches decide to do a deeper analysis of the detentions and its classification (Poister and Van Slyke, 2002). One of many potential applications of this study’s results is the production plan in a transportation system, specifically in operations schedule, the same that contains the departure and arrival information of different company’s fleet (Etschmaier, 1987).

CONCLUSIONS

The detention identification algorithm gives reliable data when talking about the time in detention variable since it does not require further logistic parameters that could make the results vary between the naturality of one city to another. The algorithm works on the data given and uses fixed parameters based on statistic and previous GPS's data studies.

The artificial intelligence model, HDBSCAN, generated effective clusters when talking about the identification of geospatial points of traffic or vehicular congregation and grouping them by common or similar characteristics giving us the detention classification for each of the vehicle datasets studied. These allowed us to divide the detentions into different groups like traffic detentions, delivery detentions and resting detentions. However, because of the naturality of the model, some vehicles had more than three clusters. Therefore, it complicated a deeper analysis since there are no public logistic studies from Quito that gives us more parameters, we could use to give a specific label to these clusters.

Furthermore, the artificial intelligence model is as robust as it can be to receive, process and analyze whatever set of geospatial datasets if and only if this dataset has the standard characteristics specified in the model. Also, the detention classification clusters obtained from the different artificial intelligence models could be used as inputs for potential applications for the optimization of logistic routes by analyzing the points where traffic detentions are congregated. Improvement in the overall last mile operations of a company convey into higher margins and profits. An optimized distribution logistics system yields to a relevant competitive advantage; either better capacity of responsiveness or more efficiency could be attained.

In the other hand, public entities could use the model's results for the design of urban freight policies and infrastructure by the public policy application. For example, the creation of specific discharge bays for urban freight inside the urban area where traffic gets affected by unforeseen actions like delivery detentions. Different applications must make emphasis on the goods' movement inside the urban area of the city.

REFERENCES

Aristidis, L., Nikos, V., & Verbeek, J. (2003). The global k-means clustering algorithm. *Pattern Recognition, 36*(2), 451-461. doi:ISSN 0031-3203

Beauchamp, J. J., & Olson, J. S. (1973). Corrections for Bias in Regression Estimates After Logarithmic Transformation. *Ecology, 54*(6), 1403-1407. doi:10.2307/1934208

Banco Interamericano de Desarrollo. (2013). Observatorio Regional de Transporte de Carga y Logística. *Departamento de Transporte y Medio Ambiente,* 8-17. Nota Técnica IDB-NT-508.

Birant, D., & Kut, A. (2007). ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data & Knowledge Engineering, 60*(1), 208-221. doi:10.1016/j.datak.2006.01.013

Comendador, J., López-Lambas, M. E., & Monzón, A. (2012). A GPS Analysis for Urban Freight Distribution. *Procedia - Social and Behavioral Sciences, 39,* 521-533. doi:10.1016/j.sbspro.2012.03.127

Conagin, A., Barbin, D., & Demétrio, C. G. (2008). Modifications for the tukey test procedure and evaluation of the power and efficiency of multiple comparison procedures. *Scientia Agricola, 65*(4), 428-432. doi:10.1590/s0103-90162008000400016

Du, J., & Aultman-Hall, L. (2007). Increasing the accuracy of trip rate information from passive multi-day GPS travel datasets: Automatic trip end identification issues. *Transportation Research Part A: Policy and Practice, 41*(3), 220-232. doi:10.1016/j.tra.2006.05.001

Greaves, S., Figliozzi, M. (2008). Collecting Commercial Vehicle Tour Data with Passive Global Positioning System Technology Issues and Potential Applications. *Transportation Research Record: Journal of the Transportation Research Board*, 158-165.

Hariharan, R., & Toyama, K. (2004, October 20). Project Lachesis: Parsing and Modeling Location Histories. Retrieved October 18, 2020, from https://link.springer.com/chapter/10.1007/978-3-540-30231-5_8

Hess, S., Quddus, M., Rieser-Schüssler, N., & Daly, A. (2015). Developing advanced route choice models for heavy goods vehicles using GPS data. *Transportation Research Part E: Logistics and Transportation Review, 77*, 29-44. doi:10.1016/j.tre.2015.01.010

Kin, B., Verlinde, S., & Macharis, C. (2017, March 24). Sustainable urban freight transport in megacities in emerging markets. Retrieved December 19, 2020, from https://www.sciencedirect.com/science/article/pii/S2210670716305868

Malzer, C., Baum, M. (2019). A Hybrid Approach To HierarchicalDensity-based Cluster Selection. *arXiv:1911.02282*, 2-7.

Morissette, L., & Chartier, S. (2013). The k-means clustering technique: General considerations and implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology, 9*(1), 15-24. doi:10.20982/tqmp.09.1.p015

M.M. Etschmaier, Operational Planning and Control in Transportation Systems, IFAC Proceedings Volumes, Volume 20, Issue 3, 1987, Pages 147-153, ISSN 1474-6670, https://doi.org/10.1016/S1474-6670(17)55889-9.

McNeff, J. (2002). The Global Positioning System. *IEEE TRANSACTIONS ON MICROWAVE THEORY AND TECHNIQUES, VOL. 50, NO. 3*, 645-652.

Mcinnes, L., Healy, J., & Astels, S. (2017). Hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software, 2*(11), 205. doi:10.21105/joss.00205

Municipio del Distrito Metropolitano de Quito. (2014). Diagnóstico De La Movilidad En El Distrito Metropolitano De Quito Para El Plan Metropolitano De Desarrollo Territorial (Pmot), 5-10. Retrieved From http://gobiernoabierto.quito.gob.ec/wp-content/uploads/documentos/pdf/diagnosticomovilidad.pdf

Sharman, B., Roorda, M. (2011). Analysis of Freight Global Positioning System Data Clustering Approach for Identifying Trip Destinations. *Transportation Research Record Journal of the Transportation Research Board*, 83-91

Spaccapietra, S., Parent, C., & Spinsanti, L. (n.d.). Trajectories and Their Representations. *Mobility Data,* 3-22. doi:10.1017/cbo9781139128926.002

Poister, T. H., & Slyke, D. M. (2002). Strategic Management Innovations in State Transportation Departments. *Public Performance & Management Review, 26*(1), 58. doi:10.2307/3381298

Rahman, M. F., Liu, W., Suhaim, S. B., Thirumuruganathan, S., Zhang, N., & Das, G. (2017). Density Based Clustering over Location Based Services. *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*. doi:10.1109/icde.2017.103

Romano, A., Tkanori, S,. Hong, M., Jeong, K., Jing, P., Ben-Akiva, M,. (2019). Exploring Algorithms for Revealing Freight Vehicle Tours, Tour-Types, and Tour-Chain-Types from GPS Vehicle Traces and Stop Activity Data. *Journal of Big Data Analytics in Transportation.* 175-177.

Laranjeiro, P. F., Merchán, D., Godoy, L. A., Giannotti, M., Yoshizaki, H. T., Winkenbach, M., & Cunha, C. B. (2019). Using GPS data to explore speed patterns and temporal fluctuations in urban logistics: The case of São Paulo, Brazil. *Journal of Transport Geography, 76*, 114-129. doi:10.1016/j.jtrangeo.2019.03.003

Lee, J. (2007). Comparison of GPS-Equipped Vehicles and its Archived Data for this Estimation of Freeways Speeds. (Master thesis, Georgia Institute of Technology, Atlanta, United State).

Khan, A. M. (2001). Reducing Traffic Density: The Experience of Hong Kong and Singapore. *Journal of Urban Technology, 8*(1), 69-87. doi:10.1080/10630730120052181

Wang, F., Franco-Penya, H., Kelleher, J. D., Pugh, J., & Ross, R. (2017). An Analysis of the Application of Simplified Silhouette to the Evaluation of k-means Clustering Validity. *Machine Learning and Data Mining in Pattern Recognition Lecture Notes in Computer Science*, 291-305. doi:10.1007/978-3-319-62416-7_21

Wang, J., Chai, R., & Xue, X. (2016). The Effects of Stop-and-go Wave on the Immediate Follower and Change in Driver Characteristics. *Procedia Engineering, 137,* 289-298. doi:10.1016/j.proeng.2016.01.261

Wowczko, I. (2015). A Case Study of Evaluating Job Readiness with Data Mining Tools and CRISP-DM Methodology. *International Journal for Infonomics, 8*(3), 1066-1070. doi:10.20533/iji.1742.4712.2015.0126

Yang, X., Sun, Z., Ban, X. J., & Holguín-Veras, J. (2014). Urban Freight Delivery Stop Identification with GPS Data. *Transportation Research Record: Journal of the Transportation Research Board, 2411*(1), 55-61. doi:10.3141/2411-07

Yu, C. H. (2010). Exploratory data analysis in the context of data mining and resampling. *International Journal of Psychological Research, 3*(1), 9-22. doi:10.21500/20112084.819

Zheng, D., & Wang, Q. (2013). Selection algorithm for K-means initial clustering center. *Journal of Computer Applications, 32*(8), 2186-2188. doi:10.3724/sp.j.1087.2012.02186

Zhao, W., Mccormack, E., Dailey, D. J., & Scharnhorst, E. (2013). Using Truck Probe GPS Data to Identify and Rank Roadway Bottlenecks. *Journal of Transportation Engineering*, *139*(1), 1-7. doi:10.1061/(asce)te.1943-5436.0000444