# ESTIMATION OF A LOADING/UNLOADING PARKING DURATION MODEL FOR COMMERCIAL VEHICLES IN URBAN AREAS

**Lina María Vasco Díaz**

Universidad Nacional de Colombia
Facultad de Minas, Departamento de Ingeniería Civil
Medellín, Colombia
2024

# ESTIMATION OF A LOADING/UNLOADING PARKING DURATION MODEL FOR COMMERCIAL VEHICLES IN URBAN AREAS

**Lina María Vasco Díaz**

**Tesis presentada como requisito parcial para optar al título de:**

Magister en Ingeniería - Infraestructura y Sistemas de Transporte

**Director:**
John Jairo Posada Henao, PhD.

**Co-director:**
Carlos Alberto González-Calderón, PhD.

**Línea de investigación:**
Freight transportation

**Grupo de investigación:**
Grupo de Investigación en Vías y Transporte (VITRA)
Semillero de transporte de carga y logística (FTL)

Universidad Nacional de Colombia
Facultad de Minas, Departamento de Ingeniería Civil
Medellín, Colombia
2024

*Dedicatoria*

*A Dios por todas las bendiciones y oportunidades recibidas.*
*A mis padres y hermanos, por ser siempre pilares, ejemplo y motivación en mi vida.*
*A mi novio y mis amigos, por ser consejeros, cómplices y haber hecho de este proceso algo inolvidable.*
*A mis profesores y compañeros, que con su guía, apoyo y cariño hicieron esto posible.*

*"Una Inversión en conocimiento, siempre paga el mejor interés".*
*-Benjamin Franklin*

# Agradecimientos

Quiero expresar mis agradecimientos a los profesores John Jairo Posada Henao y Carlos González-Calderón por ser mis guías en este proceso y brindarme su apoyo tanto académica como personalmente de manera incondicional a lo largo de toda mi formación.

A los profesores Carlos Iván Rivera y Diana Ramírez por ser una guía y un apoyo durante todo el desarrollo de esta tesis. Gracias por su tiempo, apoyo y dedicación en mi formación académica y profesional.

A las profesoras Claudia Muñoz y Marcela Aldana por ser consejeras, compañeras, amigas y una gran motivación en los momentos que más lo necesitaba, gracias por creer en mí y ayudarme a explotar mi potencial.

A todos los compañeros de especialización y maestría que han sido un gran apoyo en cada paso de este viaje.

A la Universidad Nacional de Colombia por abrirme sus puertas y permitirme seguir mi formación como una ingeniera integral. Pero, sobre todo ha sido un espacio de crecimiento personal, aprendizaje en diversos aspectos y una segunda alma máter que me ha dejado recuerdos y personas inolvidables. Agradezco profundamente todas las oportunidades y apoyo bridado.

# Declaración de obra original

Yo declaro lo siguiente:

He leído el Acuerdo 035 de 2003 del Consejo Académico de la Universidad Nacional. «Reglamento sobre propiedad intelectual» y la Normatividad Nacional relacionada al respeto de los derechos de autor. Esta disertación representa mi trabajo original, excepto donde he reconocido las ideas, las palabras, o materiales de otros autores.

Cuando se han presentado ideas o palabras de otros autores en esta disertación, he realizado su respectivo reconocimiento aplicando correctamente los esquemas de citas y referencias bibliográficas en el estilo requerido.

He obtenido el permiso del autor o editor para incluir cualquier material con derechos de autor (por ejemplo, tablas, figuras, instrumentos de encuesta o grandes porciones de texto).

Por último, he sometido esta disertación a la herramienta de integridad académica, definida por la universidad.

**Lina Maria Vasco Diaz**

Fecha

22/10/2024

# Resumen


**ESTIMACIÓN DE UN MODELO DE DURACIÓN DE ESTACIONAMIENTO DE VEHÍCULOS COMERCIALES PARA CARGUE Y DESCARGUE EN ZONAS URBANAS**

A lo largo de los años, la entrega de bienes y servicios se ha enfrentado cada vez más a retos como la falta de zonas de parqueo para las entregas, el aumento de la demanda por parte de los usuarios y externalidades negativas como mayores niveles de accidentes y contaminación. En esta investigación se utiliza un modelo de regresión Lasso para estimar un modelo de duración del parqueo para actividades de carga y descarga. El modelo, desarrollado a partir de 247 observaciones de operaciones de camiones en Medellín (Colombia), se utiliza para identificar las principales variables que influyen en la duración del estacionamiento en el Área Metropolitana de Medellín. Los resultados del modelo indican cómo las principales variables que intervienen en el tiempo de estacionamiento son el peso de la carga (Ton), las unidades por Pallet, el número de empleados del establecimiento, el número de equipos manuales, el número de equipos mecánicos, y si la operación fue asistida con equipos de carga/descarga. También se descubrió que el modelo de regresión Lasso tiene un gran potencial para comprender y predecir la duración del estacionamiento de los viajes de camiones comerciales. La tesis concluye con recomendaciones para futuras investigaciones basadas en los resultados obtenidos al aplicar la metodología propuesta en este estudio.


**Palabras clave:** *Parqueo en zonas urbanas, Vehículos comerciales, Regresión Lasso, Duración del estacionamiento.*

# Abstract


### ESTIMATION OF A LOADING/UNLOADING PARKING DURATION MODEL FOR COMMERCIAL VEHICLES IN URBAN AREAS

Over the years, the delivery of goods and services has increasingly faced challenges such as a lack of parking spaces for deliveries, increased demand from users, and negative externalities such as higher levels of accidents and pollution. This research uses a Lasso regression model to estimate a parking duration model for loading and unloading activities. The model, developed from 247 observations of truck operations in Medellin (Colombia), is used to identify the main variables influencing parking duration in Medellin's metropolitan area. The model results indicate how the main variables that intervene in the parking time are cargo weight (Ton), units per Pallet, number of employees, number of manual equipment, number of mechanical equipment, and whether the operation was assisted with loading/unloading equipment. It was also found that the Lasso regression model has great potential to understand and predict the parking duration of commercial truck trips. The thesis concludes with recommendations for future research based on the results obtained by applying the methodology proposed in this study.


**Keywords:** *Urban Parking, Commercial Vehicles, Lasso regression, Parking duration*

# Content

**Resumen** .......... **7**

**Figure List** .......... **10**

**Table List** .......... **10**

**1. Introduction** .......... **11**
Objectives .......... 14

**2. Background** .......... **15**
2.1 Freight transportation challenges .......... 16
2.1.1 Negative externalities .......... 16
2.1.2 Demand for parking and parking duration .......... 16
2.1.3 Freight transportation in the global south .......... 17
2.2 Parking duration models .......... 17
2.3 Lasso regression and machine learning models .......... 18

**3. Methodology** .......... **20**
3.1 Sample size and data preparation .......... 20
3.2 Outliers' identification .......... 21
3.2.1 Multivariate analysis .......... 21
3.2.2 Univariate analysis .......... 21
3.3 Outlier management .......... 22
3.4 Correlation Analysis .......... 23
3.5 Model estimation .......... 24
3.5.1 Lasso Regression .......... 24
3.5.2 Standardization .......... 25
3.5.3 K-fold cross-validation .......... 25
3.6 Classification scheme .......... 25

**4. Case Study** .......... **27**
4.1 Data Description .......... 28
4.2 Sample size and data preparation .......... 28
4.3 Outlier identification .......... 30
4.4 Variables distribution .......... 33
4.5 Correlation Analysis .......... 34
4.6 Model estimation .......... 36
4.7 Segmented Model Estimation .......... 41
4.8 Duration classification scheme .......... 44
4.9 Proposed strategies .......... 47

**5. Conclusions and recommendations** .......... **49**

**6. References** .......... **53**

# Figure List

Figure 1. Winsorization method representation ........ 22
Figure 2 Medellin Metropolitan area location. ........ 27
Figure 3. Inter curtail range ........ 31
Figure 4. Variables distribution ........ 33
Figure 5. Spearman's correlation matrix ........ 34
Figure 6 Lasso Path plot Final model ........ 38
Figure 7 Statistics related to the type of vehicle ........ 40
Figure 8 Lasso Path plot Final segmented model ........ 43
Figure 9 Parking duration classification scheme MMA ........ 45

# Table List

Table 1. Analyzed Variables description ........ 13
Table 2. Grouping Variables ........ 28
Table 3. Multivariate analysis variables ........ 30
Table 4. Data frames used for model estimation ........ 35
Table 5. Main models result ........ 37
Table 6 Characteristics of the final model ........ 38
Table 7 Segmented database ........ 41
Table 8 Main models result with segmented database ........ 42
Table 9 Characteristics of the final segmented model ........ 43

# 1 Introduction

Freight transport plays a crucial role in the economy and the development of communities (Holguin-Veras et al., 2011). During the last two decades, there has been worldwide economic growth, which has contributed to the globalization of supply chains and, in turn, has been reflected in a substantial increase in the movement of goods in all segments and means of transportation (Barbero & Guerrero, 2017). The increase in cargo movement represents a challenge for freight carriers as they must develop efficient strategies to meet this need to transport goods and services in the shortest possible time (Schmid et al., 2018).

In Latin America, land freight transportation is responsible for mobilizing more than 70 percent of the region's national cargo, representing an essential statistic for the region's logistics (Barbero & Guerrero, 2017). Factors such as e-commerce are part of globalization that has driven the increase in freight movement. However, parking spaces in cities are limited, so this means that more commercial vehicles are competing for the same number of spaces (Reed et al., 2024). This situation represents a challenge for planners and government authorities, who must find ways to maintain the positive aspects of transportation and control the associated negative externalities such as pollution, congestion, and fossil fuel consumption (Kalahasthi et al., 2022). Within urban areas, many commercial transactions are carried out, which implies high traffic of commercial vehicles, which increases negative externalities such as congestion and pollution, among others (Ramirez-Rios et al., 2023). Therefore, these areas must be considered to understand and improve the conditions of the logistics chain.

Negative externalities also arise because of urban planning deficiencies in freight transportation. The collective parking search behavior reflects the failure of curb management and the imbalance between the supply and demand of curb space (Xiao & Jaller, 2023). Due to limited parking spaces, commercial vehicles increase their cruising time when searching for a place for loading and unloading activities, increasing congestion, pollution, and accident risks. Therefore, a correct understanding of the demand for parking in commercial areas is the first step to focusing public policies efficiently (Dalla Chiara & Goodchild, 2020).

A gap in the literature is understanding the key factors influencing parking duration. Amaya et al. (2023) identified the main variables influencing drivers' parking decisions for loading and unloading activities. These variables included parking search time, walking time, and cost per hour of parking. It was shown that truck drivers prefer regulated parking lots followed by double-line parking. Although they avoid illegal parking places, they prioritize making deliveries on time.

Prior research has illustrated how the duration of vehicle parking is directly related to parking demand and how policies can influence parking behavior. Campbell et al. (2018) analyzed parking duration from Freight Trip Generation (FTG) estimates in New York City, and their findings suggest the longer a vehicle remains parked, the greater the need for parking in that area will be (Campbell et al., 2018). Sharman et al. (2012) investigated the adoption of policies, and their implementation in pilot tests showed potential to address the problem of commercial vehicle parking in urban areas. They demonstrated the usefulness of studying this problem and the need for public policy guidance from a technical perspective.

The literature review of 34 research studies conducted by Ghizzawi et al. (2024), highlights the significant challenges commercial vehicles face when parking in densely populated areas due to insufficient infrastructure and policies that do not meet their needs. It also points to the growing interest in understanding parking dynamics in emerging economies. Therefore, this research helps fill the knowledge gap by implementing a machine learning model to estimate a model of commercial vehicle parking duration in urban areas within this context in the global south.

Silva et al. (2024) points out the importance of proposing public initiatives to improve logistics conditions in the city and conducting future research to gather the results so that decisions can be made based on the results. The lack of empirical information regarding the dynamics of freight transportation and the demand for parking spaces hinders effective city planning.

With primary information on the duration of commercial vehicle parking in urban areas, the estimated models are expected to not only provide a clearer picture of parking dynamics in these contexts but also offer valuable recommendations for formulating public policies that can improve logistics efficiency and mitigate associated adverse effects such as congestion and pollution in urban areas.

To facilitate the interpretation of the results of the estimated model,

**Table** 1 shows the description of the variables analyzed.

In some cases a second type of equipment was necessary to complete the activity (for example, the cargo was unloaded with a forklift and moved with other equipment), which corresponds to equipment 2; however, as these cases had very little information with respect to the total of the database, the column was eliminated because the blanks could not be filled in and only the equipment used in all cases corresponding to equipment 1 was considered.

Table 1. Analyzed Variables description

| | | | |
|---|---|---|---|
| **Code** | Code of each survey observation | **In-vehicle storage** | Type of product storage in the vehicle |
| **ID** | Numbering of each observation in alphabetical order | **In-house storage** | Type of product storage In-house |
| **City** | City where the survey was conducted | **Products** | Specification of type of product transported |
| **Direction** | Address where the survey was conducted | **SCTG codes (Standard Classification of Transported Goods)** | Classification of the type of product according to SCTG code from U.S. department of commerce |
| **Classification of observed vehicle** | classification of vehicle observed in loading and/or unloading activity | **Total weight of cargo (Ton)** | Weight of transported cargo in tons |
| **Operation** | Classification of the operation between loading and unloading | **Packaging** | Product packaging type |
| **Journey** | Daytime or extended working hours | **Number of pallets** | Number of pallets used in the loading or unloading activity |
| **Position 1 (administrative employees)** | Job title held by the employee in position 1 who perform administrative activities | **Units per Pallet** | Packaging units placed on pallets |
| **Number Employees in position 1 (administrative employees)** | Number of employees in the mentioned position | **Loading and unloading infrastructure** | Whether the facilities have an internal or external unloading infrastructure |
| **Salary Position 1 (administrative employees)** | Salary of employees in the mentioned position | **Total Operation Time** | Total time of loading or unloading activity duration |
| **Position 2 (operational employees participating in physical activity)** | Job title held by the employee in position 2 who carry out operational activities related to loading and unloading | **Number of employees** | sum of employees in position 1 and 2 |
| **Number Employees in position 2 (operational employees participating in physical activity)** | Number of employees in the mentioned position | **Number of manual equipment** | Number of manual equipment used in the activities |
| **Salary Position 2 (operational employees participating in physical activity)** | Salary of employees in the mentioned position | **Number of mechanical equipment** | number of mechanical equipment used in the activities |
| **Equipment 1 Description** | Description of the type of equipment used for loading or unloading the goods from the truck | **Type of equipment** | Classification of whether the equipment is manual or mechanical |
| **Number of Equipment 1** | Number of equipment used for loading or unloading the goods from the truck | **MACRO descriptions products** | Products classification according to the MACRO system in Colombia |
| **Assisted Delivery** | If the emplpyees use Equipment during the operation or not | | |

## Objectives

- **General**

To optimize freight operation time related to loading and unloading activities by modeling the parking duration time of commercial vehicles in urban areas.

- **Specific**
  - To identify the main variables affecting commercial vehicle loading and unloading times.
  - To obtain a zonal classification scheme for the parking duration of commercial vehicles in an urban area.
  - To propose strategies to optimize loading and unloading times for commercial vehicle

# 2 Background

Cargo transportation is a necessity for humanity. Everything must be moved from one point to another, from transporting raw materials to medications. However, although communities and technology have continued to grow and improve over time, we still face problems of the past, such as accidents and pollution, among others (Ortúzar & Willumsen, 2011).

The study of the factors related to the delivery of goods in urban areas has been approached from various perspectives. Sanchez-Diaz et al. (2020) studied the variables related to inefficiencies in the delivery of goods in urban areas of Sweden. His research revealed that approximately 55% of the time spent on delivery routes, drivers are engaged in loading or unloading activities and customer service. Additionally, they identified possible initiatives to improve delivery time efficiency, such as route optimization and enhancing equipment for handling goods during deliveries.

Amaya et al. (2023) examined drivers' preferences regarding parking locations for goods delivery in urban areas. The three main factors influencing the choice of a parking space are search time, walking time to the destination, and the cost of nearby parking. However, they also point out that although drivers avoid parking in prohibited areas, they are willing to do so if necessary to meet delivery times, as this is their top priority.

Burns et al. (2024) developed a curbside parking optimization model to efficiently assign parking spaces for commercial vehicles for goods deliveries using real data from Aspen and Pittsburgh. discovering that parking reservations in high-demand parking scenarios with short reservation spaces can reduce congestion, translating into annual savings in congestion-related costs and pollution emissions between US -$100,000 and US $300,000. Demonstrating that proper parking space management contributes to improving urban mobility.

Analyzing curbside parking, Castrellon & Sanchez-Diaz (2024) considered the effects of 4 different initiatives related to the management of these spaces, such as the optimization of public space, data sharing that makes it easier for users to check space availability, parking limits, and enforcing regulations. After conducting a meta-analysis that examined the effects of these alternatives, the authors concluded that these initiatives can, in fact, positively influence the reduction of emissions, improve traffic, and accelerate delivery times in urban areas.

On the other hand, Pinto et al. (2019) studied the delivery patterns of businesses in the urban area of Bergamo, Italy. By estimating a model using a two-stage approach, they determined the optimal location and size of loading bays in that area to reduce street congestion caused by improper parking and improve delivery efficiency.

To better understand the challenges and the negative externalities associated with transportation, as well as the technologies implemented to increase delivery efficiency in urban areas, the following will be described

## 2.1 Freight transportation challenges

### 2.1.1 Negative externalities

The negative externalities of transportation are related to water pollution (fuel spills, oils, among others), air pollution (gas emissions), congestion, traffic accidents, and noise, all of which negatively impact the environment and the quality of life of individuals (Chatziioannou et al., 2020).

In urban areas, there is a large volume of trips related to the delivery of goods due to the presence of commercial establishments which generate freight traffic. This leads to an increase in the negative externalities associated with transportation (Ramirez-Rios et al., 2023).

An important concept to define when thinking about freight travel in urban areas is "cruising for parking" as explained by Dalla Chiara & Goodchild (2020) as the action of drivers circling their destination while searching for a parking space. Their study discovered that the average cruising for parking time in Seattle (USA) is 2.3 minutes per trip, representing approximately 28% of the total travel time.

Cruising for parking contributes to the increase in traffic in urban areas, which is directly related to the rise in congestion and the increase in vehicle pollutant emissions. This phenomenon shows the impact of the transportation sector, which is responsible for 90-95% of carbon monoxide emissions and 30% of carbon dioxide emissions worldwide (Profillidis et al. 2014).

### 2.1.2 Demand for parking and parking duration

Previous studies have shown that the duration of parking for commercial vehicles in urban areas is related to the demand for parking spaces. Campbell et al. (2018) discovered that the longer a vehicle was parked during delivery, the greater the demand for parking spaces and cruising for parking time for other vehicles. Additionally, they proposed strategies to control this parking demand, such as off-peak delivery programs and staggered deliveries, which showed promising results in reducing demand.

When several drivers collectively search for a parking space in a specific area, it reflects a failure in managing parking demand in that area, increasing negative externalities and adversely affecting the quality of life of the residents in that community (Xiao & Jaller, 2023). Urban spaces are limited; on the other hand, the growth rate of commercial vehicles is increasing. Therefore, understanding the demand for space becomes necessary as a first step in focusing public policies on this issue (Dalla Chiara & Goodchild, 2020).

Alho et al. (2022) implemented various strategies for managing the demand for parking spaces for commercial vehicles in urban areas through simulation, such as increasing the capacity of loading bays, adjusting parking prices, and providing guidance to drivers on where to park, among others. The results demonstrated that the implementation of well-designed strategies improves delivery efficiency, reducing costs for drivers and minimizing negative externalities.

### 2.1.3 Freight transportation in the global south

The globalization of supply chains has positively influenced countries' economies and freight transportation. Due to the economy's growth in the last two decades, phenomena such as the relocation of production and outsourcing have emerged, allowing the inputs, parts, and components of final products to have different origins. However, this has also caused freight transportation to evolve with new technologies and face new challenges in optimizing the transport of goods, as it accounts for approximately 5% of the total GDP of countries. Therefore, greater efficiency in the logistics chain represents significant savings for carriers and consumers, positively contributing to the economies of those countries (Barbero & Guerrero, 2017).

## 2.2 Parking duration models

There are studies where various models have been implemented to estimate the parking durations of commercial vehicles in urban areas. Kalahasthi et al. (2022) analyzed models such as Log-normal (assume that the logarithm of the time follows a normal distribution, it is useful when the duration time has a skewed distribution), exponential, and Weibull, with the latter being the model with the best statistical characteristics for determining parking durations. Similarly, Schmid et al. (2018) analyzed log-logistic (similar to the log-normal but has a longer tail, which allows for analysis when longer duration times are more frequent than in a normal distribution), exponential, and Weibull models, among others, to estimate the parking duration of commercial vehicles in New York, with Weibull providing the best estimation results.

However, Weibull models, which work with historical data, have a disadvantage because they require a robust database to achieve acceptable reliability (Murthy et al., 2004).

Zou et al. (2016) applied disaggregated data from downtown Manhattan to estimate a Cox proportional hazards model to determine the parking duration of commercial vehicles on public streets. This semiparametric model considers that load type, location, and vehicle size influence the probability of a vehicle vacating a parking space within a given time. The authors discovered that smaller food delivery vehicles have shorter delivery durations. Using this as a reference to propose space management strategies to reduce negative externalities.

On the other hand, studies such as that of Castrellon et al. (2023) implement algorithms and machine learning methods to analyze factors that influence the dynamic demand for parking on urban roads. The results showed that these tools facilitate understanding demand's dynamic behavior. This facilitates decision-making in public space management in these areas, as public policies can be adapted more flexibly and efficiently to the changing parking needs.

In the literature review conducted by Ghizzawi et al. (2024) it is assessed that there is a significant shortage of studies on commercial vehicles; out of the 34 studies reviewed, only 3 use machine learning to describe, model, or estimate vehicle behavior. Therefore, this thesis contributes to the literature regarding the application of this model with commercial vehicles in urban areas.

## 2.3 Lasso regression and machine learning models

Various studies have used Lasso regression to estimate parking times and behaviors. Using machine learning methods, Gao & Ozbay (2017) identified the main variables influencing double parking in densely populated New York areas. The authors implemented Lasso regression as the primary technique to select the most relevant variables. When faced with multiple correlated variables in the dataset, they noticed that Lasso only selects the most relevant one and eliminates the others. Therefore, within this research, the correlated variables were separated into different data frames to avoid this phenomenon and better analyze the results of the estimated models. The study identified that factors such as traffic volume, proximity to commercial areas, block length, and number of parking spaces are relevant and contribute to the double parking of commercial vehicles.

Considering factors that influence parking demand, Feng et al. (2019) studied how behavior and parking demand in China were influenced by weather factors, day of the week, holidays, and the year's season. The authors implemented several machine learning techniques, including Lasso regression, to predict user parking behavior and regulate model variance by eliminating the least significant variables. The results showed that people tend to increase their demand for parking when weather conditions are more adverse, such as high temperatures, wind speeds, and heavy rain.

The authors also used different machine-learning methods to calculate parking duration. Parmar et al. (2021) estimated a vehicle parking duration model in urban areas using artificial neural networks. The study analyzed two types of land use: offices and markets/commercial areas. The authors discovered that economic factors such as hourly parking rates, personal income, and trip purposes are the main determinants of parking duration. This research demonstrated that using machine learning techniques allows for highly accurate predictions of parking patterns and provides valuable information for creating effective and sustainable parking management policies.

Yang et al. (2019) used machine learning models to predict parking duration and analyze the demand for parking spaces. The authors considered variables such as parking meter transactions, traffic speed, and weather conditions in downtown Pittsburgh to reduce congestion caused by the search for parking. The authors found that Lasso regression yielded satisfactory prediction results with a mean absolute percentage error (MAPE) of 14.3%. However, it was surpassed by the neural network method with a MAPE of 10.6%.

Although various machine learning techniques, such as artificial neural networks (ANN) and random forests, have proven to be highly effective in predicting parking duration, their complexity level exceeds this thesis's scope. Therefore, the Lasso regression model has been chosen for this research, showing consistent and reliable results. Additionally, it facilitates the selection of relevant variables and allows for an efficient model with a simple and manageable approach, which aligns with this work's objectives. As far as the author knows, this is the first study in the Global South.

# 3 Methodology

The methodological process to estimate the parking duration model started with an exhaustive literature review to gain a broad perspective on parking duration models and the guidelines for analyzing the information and the modeling process. This step identified and analyzed the different models to estimate parking duration. We selected the most appropriate one according to the research objective, which is to obtain a practical model with the main variables that influence the parking time of commercial vehicles in urban areas.

The methodological approach to estimate the parking duration model considered the following tasks: (i) data collection and preparation, (ii) identification and management of outliers using a multivariate and univariate analysis, (iii) correlation analysis of data, and (iv) model estimation using Lasso regression.

## 3.1 Sample size and data preparation

Different methods are analyzed to determine the sample size (as shown in **Appendix 1**). However, since the population is finite and its size is known, the method proposed by López-Roldán & Fachelli (2015), is implemented for cases with these characteristics and not to overestimate the sample size with more general formulas that imply a more significant investment of resources. As is shown in equation 1:

$$n = \frac{Z^2 * p * q * N}{(N-1) * e^2 + Z^2 * p * q} \qquad (1)$$

Where:

**n=** Sample size

**N=** Total population.

**Z=** Confidence level of probability that the actual value of what is being studied in the population is found in the calculated sample (determined by the investigator's judgment).

**p=** Expected rate of affirmative response to the survey (determined by the investigator's judgment or information from previous surveys).

**q=** Expected negative response rate to the survey (determined by the investigator's judgment or information from previous surveys).

**e=** Maximum estimation error (determined by the investigator's judgment).

## 3.2 Outliers' identification

### 3.2.1 Multivariate analysis

For a multivariate analysis, the Mahalanobis distance is used to measure the separation between observations in multidimensional space relative to the centroid of the dataset. This method has been selected because, unlike other multivariate methods such as principal component analysis or factor analysis, its results are more accessible to interpret, and it considers the correlation of variables in its analysis, which is beneficial for the Lasso regression method. In the Mahalanobis method, a more considerable distance indicates a higher likelihood that the observation deviates from the expected behavior of the data. To confirm the deviation of the data from the mean, the probability that the observations follow the chi-square distribution is calculated; those with a probability of less than 0.001 are considered outlier observations (Hair et al., 2014).

However, this method does not provide insights into the individual contribution of each variable. Therefore, it must be completed with univariate analysis to examine each variable independently.

### 3.2.2 Univariate analysis

To develop a univariate analysis, the interquartile range (IQR) is a robust measure of dispersion used to describe how data is distributed around the mean (Greenberg, 2009). This method was selected because of its ease of application and because it is a preliminary step in applying the Winsorization method. Unlike the standard deviation, the IQR focuses on the central dispersion of the data, excluding the most extreme values (Kwak & Kim, 2017). The process to calculate it is shown below:

- **Q1 (First quartile):** Represents the limit where the lowest 25% of the data set is below. To calculate Q1, it is necessary to sort the data in ascending order and then to identify the values at the 25th percentile.
- **Q3 (Third quartile):** The third quartile indicates the value where 75% of the data set lies below. To determine Q3, one must sort the data in ascending order and identify the 75th percentile. Q3 delimits the upper limit of the third quarter of the data.
- **IQR (interquartile range):** The interquartile range is a measure that indicates the dispersion within the central 50% of the data. It is calculated as the difference

between the third quartile (Q3) and the first quartile (Q1). This measure provides information about the variability concentrated in the central part of the data set, excluding any outliers. The interquartile range is increased 1.5 times upward and downward to identify the outer limits, as shown in equations 2 and 3.

$$Outlier\ below\ limit = Q1 - 1{,}5 * IQR \quad (2)$$

$$Outlier\ above\ limit = Q3 + 1{,}5 * IQR \quad (3)$$

Where:

IQR= Interquartile Range

$Q3$= Upper quartile

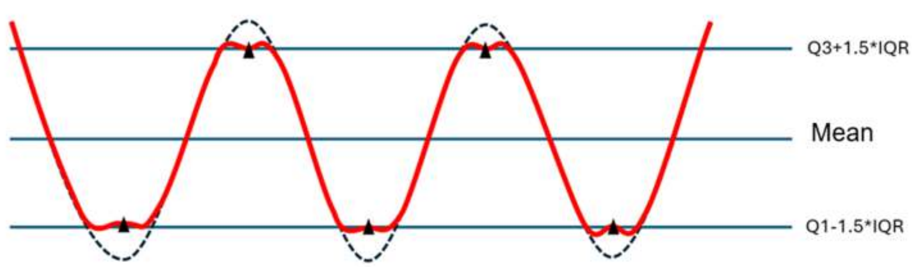
$Q1$= Lower quartile

## 3.3 Outlier management

The data considered outliers are commonly removed from the database to clean the information. However, some alternatives do not sacrifice observations.

Winsorization is a statistical method that prioritizes the integrity of the database. It involves replacing the values corresponding to a boundary percentile instead of removing them from the database. Although there are other methods to deal with outliers, Winsorization was chosen because it offers advantages such as not eliminating observations from the database by sacrificing their size and not replacing them with values such as the mean of the data. Therefore, it allows a more complex analysis of the database, and the integrity of the database is maintained for the most part. To define the analysis range, the extreme limits of the IQR are used (Kwak & Kim, 2017). Where the data beyond the boundaries is replaced by the upper boundary or the second-highest lower value where no outliers are observed, as shown in **Figure 1**.

Figure 1. Winsorization method representation



The red line represents the behavior of the database; when the observations exceed the limits set in the IQR, the Winsorization method returns the outliers to the normal range.

## 3.4 Correlation Analysis

Once the database is clean, a correlation analysis of the variables needs to be done to identify the level of dependence between them to avoid variables that have a high correlation level and could lead to an erroneous interaction when describing the behavior of the data in a model.

For this purpose, the Pearson correlation method can be applied for variables with a normal distribution as Washington et al. (2003) explained. Equation 4 is used to calculate the Pearson correlation for a sample:

$$r = \frac{COV(X,Y)}{s_x s_y} \qquad (4)$$

Where:

r= Sample correlation parameter
X= Independent variables
Y= Dependent Variable
COV= covariances between X and Y
Sx,y= Sample standard deviation

On the other hand, Spearman's correlation coefficient should be applied for variables that do not comply with the normality assumption, as explained by Wackerly et al. (2008). The equation for calculating this coefficient is presented in Equation 5.

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \qquad (5)$$

Where:

$r_s$= Sample correlation parameter
$d_i$= R(xi) - R(yi); with R(xi) denote the range of xi between x1, x2,..., xn and with R(yi) denote the range of yi between y1, y2,...yn.
n= Number of observations

The correlation analysis is performed with all the numerical variables, and the correlation parameter varies between (-1,1). The variables with a correlation coefficient more significant than |0.60| would not be included in the same model since they are highly correlated.

## 3.5 Model estimation

Several models, such as Weibull and Hazard, are commonly used to estimate durations by analyzing historical data, although they have slightly different approaches. Weibull models focus mainly on the distribution of times until an event occurs; this is usually used to estimate the possible failure of materials for scheduled maintenance (Murthy et al., 2004) and to estimate vehicle parking duration times (Kalahasthi et al., 2022).

Hazard models are more focused on the probability of occurrence of a specific event at a particular time by analyzing the characteristics of the study subjects (Sharman et al., 2012). The disadvantage of these models is that when working with historical data, a robust database is needed to increase the reliability of the models, which translates into more significant resources to obtain and analyze the information.

Machine learning models can also be implemented for regressions, such as Ridge and Lasso; unlike the previous models, they do not work with historical data, representing a significant advantage. Although both models can be implemented, given that both are responsible for adding penalties to non-significant variables (G. James et al., 2013). For this case, Lasso regression is chosen for the regression method.

This is because when estimating duration models, a wide variety of variables are analyzed, as shown in

**Table** 1. Specifically, the Lasso regression eliminates non-relevant variables by forcing them to be zero, which adds a selection factor that the Ridge model does not have and is very significant for the modeling process.

### 3.5.1 Lasso Regression

The Lasso regression is shown in Equation 6 (G. James et al., 2013).

$$RSS + \lambda \sum_{j=1}^{p} |\hat{\beta}_j| \quad (6)$$

Where:

RSS: Residual sum of squares

$\lambda$: Turning parameter Lambda

$p$: Is the total number of parameters in the model (excluding the intercept term)

$j$: It is the index that goes through all the coefficients from 1 to $p$

$\hat{\beta}_j$: Shrinking coefficients

$\sum_{j=1}^{p} |\hat{\beta}_j|$: The sum of the absolute value of estimated parameters except for the intercept

To estimate the model and satisfy Equation 6, it is necessary to standardize the data and calculate an optimal Lambda value that allows the variables to be chosen correctly. Hence, standardization and cross-validation methods must be used.

### 3.5.2 Standardization

The data is standardized to fit the properties of a standard normal distribution, which helps the gradient descent learning algorithm converge quickly. To do so, it is necessary to make the mean zero and the standard deviation one by applying equation 7 (Raschka & Mirjalili, 2017).

$$x'_j = \frac{x_j - \mu_j}{\sigma_j} \quad (7)$$

Where:

$x'_j$: Standardized value

$x_j$: Each feature j in our dataset.

$\mu_j$: Sample Mean

$\sigma_j$: Sample standard deviation

### 3.5.3 K-fold cross-validation

In k-fold cross-validation, the training dataset is split into folds without replacement, where k-1 folds are used for model training and one-fold for performance evaluation. This procedure is repeated several times to obtain models and performance estimates (Raschka & Mirjalili, 2017). With the previous results, the values of Lambda obtained are plotted, and the one that minimizes the error is chosen. Thus, Lasso's regularization coefficient is adjusted to select the main variables of the regression model.

## 3.6 Classification scheme

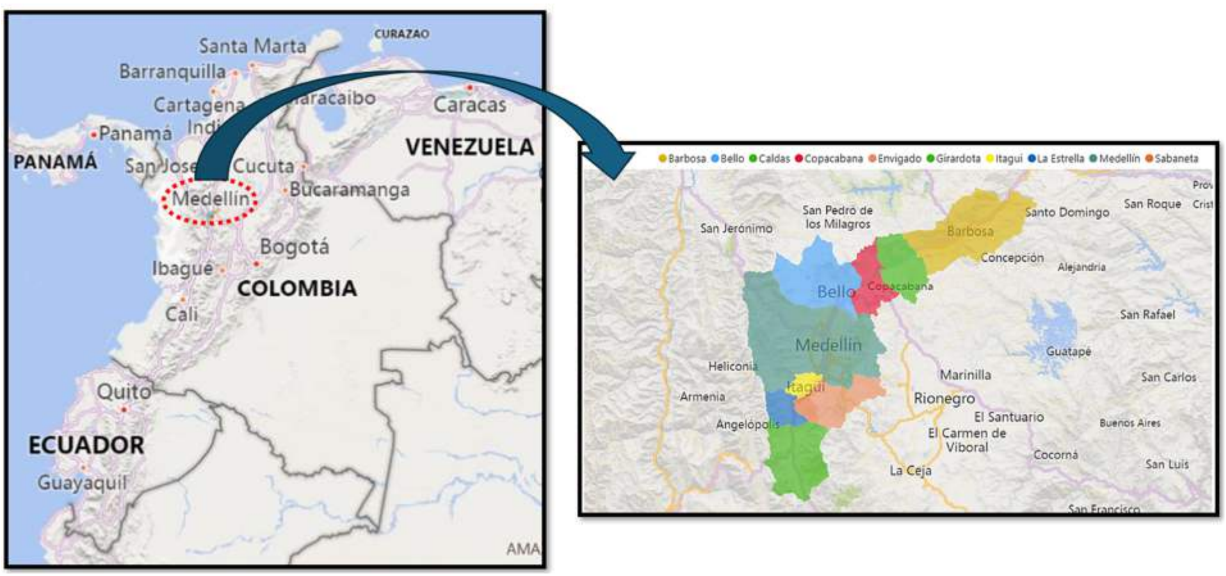
Geographic information systems (GIS) are fundamental tools that help humans manipulate, organize, interpret, and visualize spatial data to identify patterns or trends in the data. They are helpful for urban transportation planning, among others (Scott & Janikas, 2010).

# 4 Case Study

This section presents the characteristics of the study area where the proposed methodology was implemented, and the data used for the analysis. The research focuses on the Medellín Metropolitan Area (MMA) in Colombia, classified as an upper-middle income country with GNI per capita between US $4,466 and US $13,845 (World Bank Group, 2024). This focus on upper-middle-income countries offers a relevant and replicable study, given that those in the same classification share similar socioeconomic characteristics. In addition, MMA stands out in a dynamic and promising development context (Comunas de Medellín, 2024), which provides an opportunity to understand an emerging urban context's economic and social dynamics.

The Medellin Metropolitan Area is composed of 10 municipalities whose location is shown in **Figure 2**, with an area of approximately eight thousand nine hundred square kilometers and four million three hundred thousand inhabitants. (Área Metropolitana del Valle de Aburrá, 2019)

**Figure 2** Medellin Metropolitan area location.




Source: Modified based on Área Metropolitana del Valle de Aburrá, (2019)

The data used in the analysis are obtained from the database of the loading study conducted in MMA in 2018 Área Metropolitana del Valle de Aburrá & Universidad Nacional de Colombia, (2018). The database has 250 pieces of data, which were analyzed, identified outliers, and debugged before estimating the parking duration models for loading and unloading activities.

## 4.1 Data Description

The variables used in the analysis of the estimated model can be broadly divided into two large groups:

- **Socioeconomic variables:** they comprise aspects such as employee positions, salaries, and type of equipment available, among others. These contribute to understanding how these factors influence activity durations.
- **Spatial variables:** include information on the geographic location of the establishments, as well as the characteristics of the infrastructure for loading and unloading them, to consider the physical environment in the analysis of activities

## 4.2 Sample size and data preparation

This research's database was obtained from the Freight Transportation Survey conducted by Área Metropolitana del Valle de Aburrá & Universidad Nacional de Colombia (2018). The sample size was calculated considering a finite population of 437,000 daily loading and unloading activities within the Medellin Metropolitan Area and applying the formula posted by López-Roldán & Fachelli (2015) as it is shown in equation 1.

To ensure a confidence level of 95%, replace the value of Z=1.96. Additionally, the values of p and q equal to 0.5 are replaced due to the lack of information about the characteristics of the population. Likewise, a maximum error rate of 10% was established, resulting in an initial sample size of 96 surveys. However, 250 surveys were conducted during data collection, which reduced the error rate to approximately 6.2%, intending to have greater precision when estimating future models and representing the population under study.

From the initial surveys, three observations were identified with missing information. Therefore, those three observations were eliminated to avoid making additional imputations, leaving 247 surveys for the analysis. With the 247 observations, a descriptive statistics analysis was conducted on the database, identifying the characteristics of the categories within and grouping the variables with similar characteristics according to the guidelines proposed by entities such as the Federal Highway Administration, (2017) and U.S. Department of transportation, (2021). This results in **Table 2**, where the summary of the grouping given in the previous criteria is shown. After this, we proceed to identify the outlier data.Table 2. Grouping Variables

**Vehicles Type**

| Vehicle Classification Observed | | Federal Highway Administration vehicle classification |
|---|---|---|
| **Group 1** | Motorcycle | Class 1: Motorcycles |
| **Group 2** | Pickup-Van | Class 3: Four tire, single unit |
| **Group 3** | C2P-C2G-Little truck-Light car | Class 5: Two axle, six tire, single unit |
| **Group 4** | C3 | Class 6: Three axle, single unit |
| **Group 5** | C2S2 | Class 8: Four or less axle, single trailer |
| **Group 6** | C3S2 | Class 9: Five axle tractor, semi-trailer |
| **Group 7** | C3S3 | Class 10: Six or more axle, single trailer |
| | **Packaging Type** | |
| **Group 1** | Lumps-Bales | |
| **Group 2** | Bulk- Hopper | |
| **Group 3** | Boxes-Sacks-Jars | |
| **Group 4** | Bags-Gages-baskets | |
| **Group 5** | Not Apply- Empty- Barrels- Rolls- Others | |
| | **In-vehicle and In-House storage** | |
| **Group 1** | On the floor- Bulk- bales- Barrels | |
| **Group 2** | Stowaged-baskets-Bags-Boxes | |
| **Group 3** | Shelfs | |
| **Group 4** | Hoisting- Others | |
| | **MACRO Classification** | |
| **Group 1** | Agriculture and Livestock-Prepared foods | |
| **Group 2** | Construction materials and equipment-Rubble | |
| **Group 3** | Industrial, mineral and chemical products | |
| **Group 4** | Manufactured products-Others | |

The equivalence of vehicle classification according to the FHA is added to facilitate the reader's interpretation of the classification of trucks in Colombia.

The grouping of variables was carried out to facilitate the training process of the regression model. Since the variables are categorical, they must be treated as binary variables. (dummies). This implies that it is necessary to create a column in the database for each subcategory within a category; therefore, by generalizing the number of categories, data processing is facilitated, computational resources are optimized, and a more efficient analysis is allowed without compromising the model's accuracy.

The survey information was reviewed, and typing errors and misinterpretations were corrected to calculate the outliers with the cleaned database.

## 4.3 Outlier identification

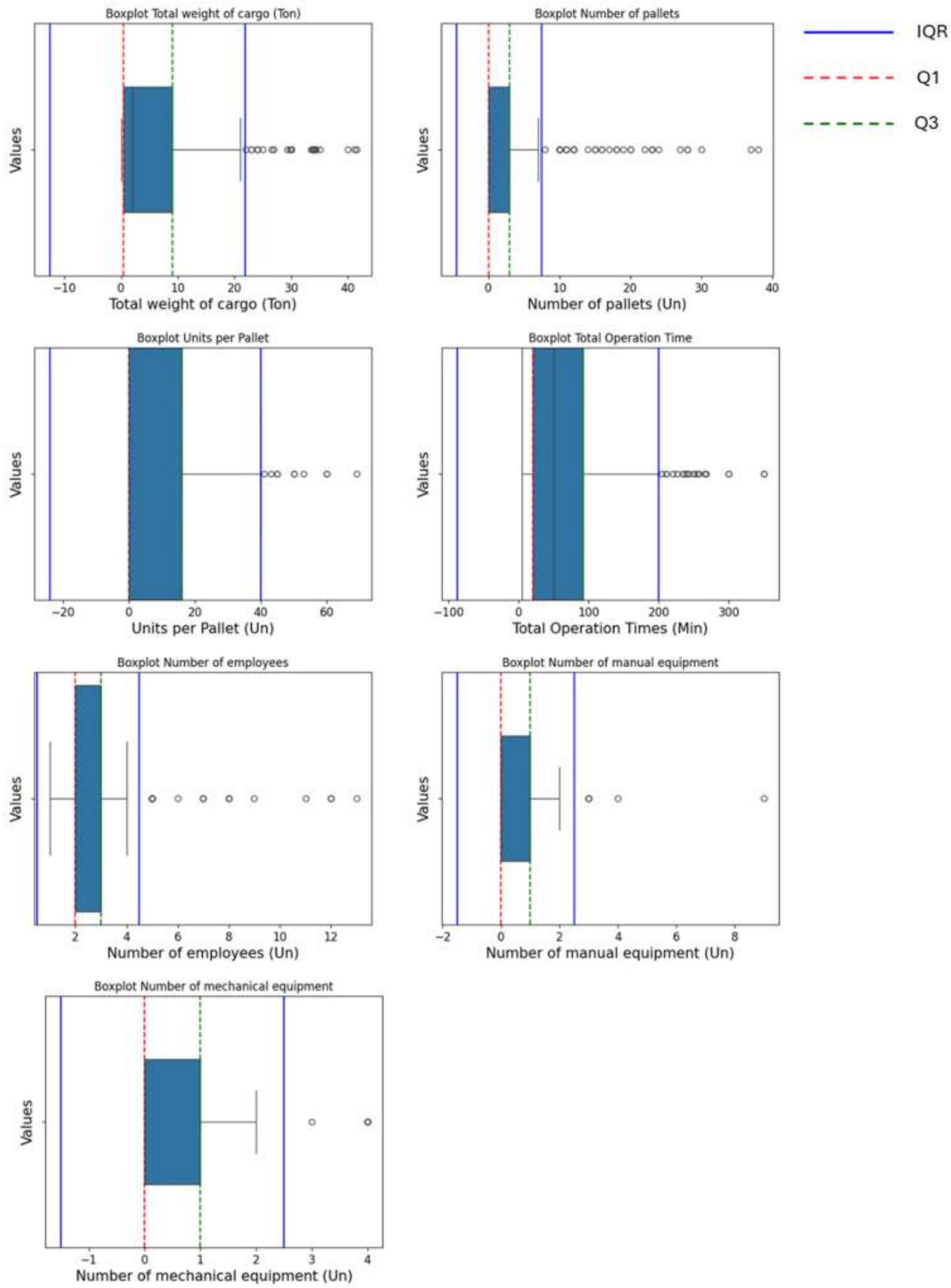
As presented in the methodology, the database is first examined through a multivariate analysis by calculating the Mahalanobis distance (Hair et al., 2014). For this purpose, the variables are shown in **Table 3**. This is because numerical and continuous variables are needed to calculate the Mahalanobis distance. Although salaries were numerical and continuous variables, they were not considered for the outlier analysis because they are on a very different scale from the other variables and were deemed irrelevant.

Table 3. Multivariate analysis variables

| Variables | Units |
|---|---|
| Total Weight of Cargo | Ton |
| Number of pallets | Un |
| Units per Pallet | Un |
| Total Operation Time | Min |
| Number of employees | Un |
| Number of manual equipment | Un |
| Number of Mechanical Equipment | Un |

The Mahalanobis distance was chosen as a reliable method for conducting multivariate analysis; additionally, the statistical tool SPSS is available, which allows it to be calculated reliably and efficiently. With this program, the Mahalanobis distances of each observation in the database are calculated, resulting in 17 observations with distances more significant than the rest. Then, the probability for all observations is estimated to see if they follow the Chi-square distribution as a criterion for identifying outlier observation values. With this, it is confirmed that 17 observations exhibit atypical behavior compared to the rest of the database; however, this is not a sufficient criterion to define them as erroneous data or to apply any treatment. Therefore, an univariate analysis is conducted on the variables, examining them individually with the interquartile range (IQR) to determine the extreme values, as shown in **Figure 3**.

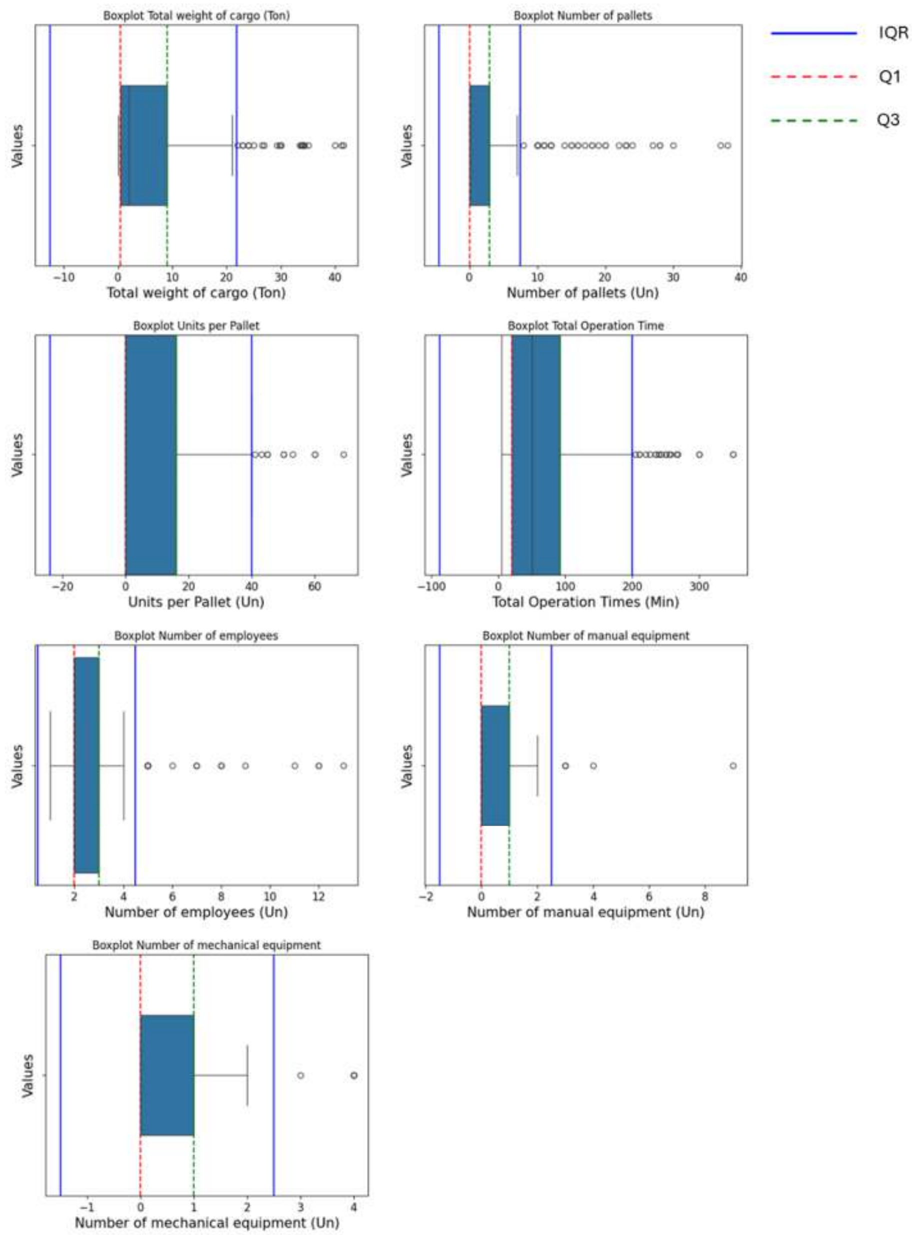
Figure 3. Interquartile range without Winsorization




When conducting the univariate analysis, observations that could statistically be considered out of the ordinary were found. Still, when analyzing all the variables together, they were identified as logical variables, for example, observations of large vehicles, where the load weight and total operating time were much higher than average but made sense when related to the type of vehicle.

When applying the IQR method, two observations were identified outside the typical ranges of data distribution concerning the number of units of goods per pallet and the total operation time. The Winzorization method was applied, and the maximum value of the IQR replaced the two outliers. The Mahalanobis distance was recalculated in the database with this modification, as shown in **Figure 4**.

Figure 4 Interquartile range with Winsorization

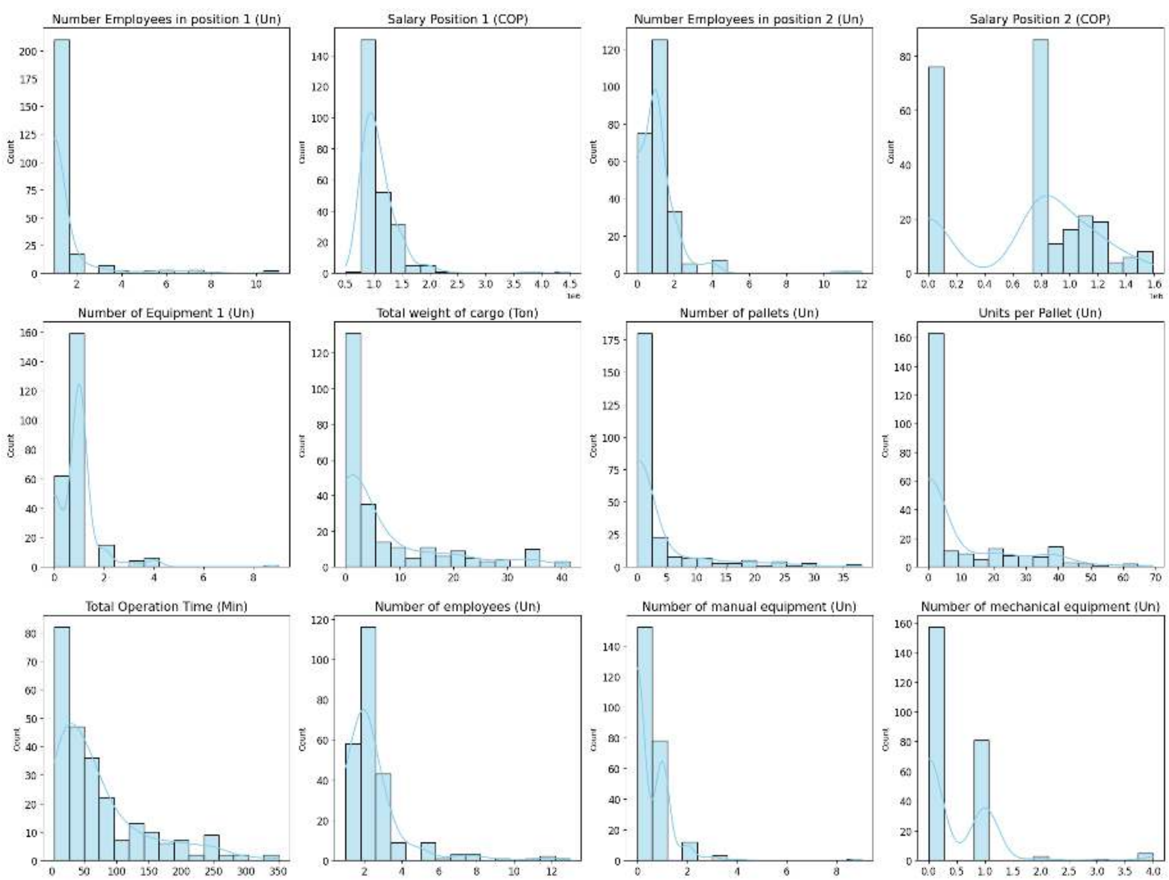



Although there is not a significant graphical difference, this is because the amount of data changed relative to the total data is approximately 0.8%, which does not represent a significant change but increases the model's reliability by ensuring that after applying the Winsorization method, the regression model is not estimated with outliers within the database. Additionally, given the high consistency in the data and the need to make very few changes, it can be concluded that the database contains reliable primary information.

## 4.4 Variables distribution

Unlike the multivariate analysis, where seven variables were chosen to calculate the Mahalanobis distance of the observations in the database shown in **Table 3**, it is necessary to analyze the distribution of all the numerical variables in the database (twelve variables). For this purpose, histograms are created for each of them, as shown in **Figure 5**.

Figure 5. Variables distribution

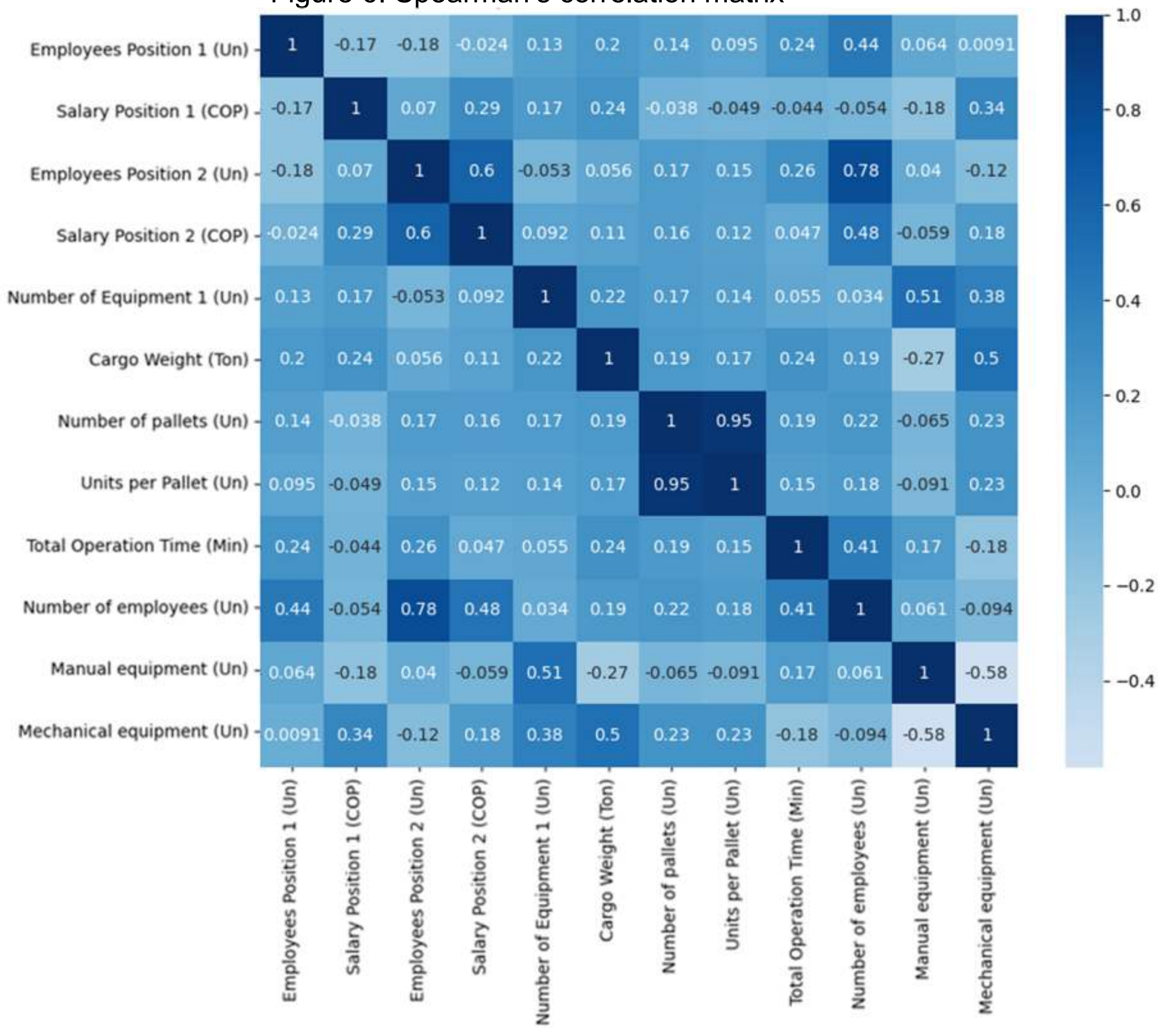



When analyzing the histograms of the numerical variables in the database, it is identified that they do not follow a normal or Gaussian distribution; therefore, the correlation coefficients between the variables are calculated using Spearman's method for non-parametric statistics, as mentioned in the methodology (Wackerly et al., 2008). The numerical variables in the database are used for calculating the correlation matrix, corresponding to those shown in **Figure 5**.

## 4.5 Correlation Analysis

Applying the correlation analysis for non-parametric variables, the Spearman correlation matrix is shown below in **Figure 6**. For this research, highly correlated variables are defined as those with a correlation coefficient greater than 0.6 according to the author's criteria after reviewing the literature and based on personal experience.

Figure 6. Spearman’s correlation matrix



The initial database, where all the variables mentioned in Table 1 are found, is defined as data frame 1 or Df 1. When separating all the numerical variables shown in **Figure 5** the so-called data frame 2 or Df 2 is created to analyze their distribution. Additionally, the variables defined as highly correlated are separated into different data frames to prevent any regression model malfunction, as shown in **Table 4**.

When analyzing **Figure 6**, variables with significantly lower correlation coefficients than the rest can be identified, such as the salaries of employees in positions 1 and 2 and the number of equipment 1, which have coefficients of 0.044, 0.047, and 0.055, respectively. When conducting a conceptual analysis of the relationship of these variables with the total operation time, no possible relationship is found that would justify considering them when estimating the models. For the above reasons and to improve the model fit by eliminating unnecessary variables, it is decided to remove these three variables from Data Frame 3 (Df 3) onwards.

As was mentioned above, the Lasso regression is not conditional to not using highly correlated variables in the same model. Therefore, the highly correlated variables were divided into seven different data frames, as shown in **Table 4**, to avoid having two highly correlated variables in the same data frame and to determine which combination of variables resulted in a model with better predictive characteristics.

Table 4. Data frames used for model estimation

| | Data frames | | | | | | |
|---|---|---|---|---|---|---|---|
| **Variables** | **Df2** | **Df3** | **Df4** | **Df5** | **Df6** | **Df7** | **Df8** |
| Employees position 1 (Un) | x | x | x | | | | |
| Salary Position 1 (COP) | x | | | | | | |
| Employees position 2 (Un) | x | x | x | | | | |
| Salary Position 2 (COP) | x | | | | | | |
| Number of Equipment 1 (Un) | x | | | | | | |
| Cargo Weight (Ton) | x | x | x | x | x | x | x |
| Number of pallets (Un) | x | | x | | | x | x |
| Units per Pallet (Un) | x | x | | x | x | | |
| Total Operation Time (Min) | x | x | x | x | x | x | x |
| Number of employees (Un) | x | | | x | x | x | x |
| Manual equipment (Un) | x | x | x | x | | | x |
| Mechanical equipment (Un) | x | x | x | x | | | x |
| Assisted Delivery (Dummy) | | x | x | x | x | x | x |

**Df: Data frame*

## 4.6 Model estimation

The R studio's glmnet library was used to estimate models. After defining the data frames, all with 247 observations, each was divided into 80% of the data to train the Lasso model and 20% to test the final model using the Caret library, which provides a robust and balanced stratified separation compatible with cross-validation processes. For the separation of the data sets, the ***"CreateDataPartition"*** function from the Caret library is used, with which the R program separates the data frame into 198 observations for the training dataset used to estimate Lasso regression models and 49 observations for the testing dataset, which is subsequently used for the validation of the estimated models with each data frame.

In this way, the values of $R^2$ and RMSE were calculated by comparing the data from the computed models and the test database. The standardization process was carried out after separating the databases to avoid data leakage, which occurs when information from the test set leaks into the training set and can lead to an overestimation of the model (Müller & Guido, 2016).

When standardizing separate databases, the best value for the Lambda parameter is calculated to determine the model with fewer prediction errors. The results of the models were estimated with each data frame, as shown in **Table 5**, with the $R^2$ and de root mean square error (RMSE) of each of them.

In Lasso regression, the focus is on regularization rather than traditional statistical significance tests such as the T-test or P-value, this means that even if individual variables are not statistically significant evaluated by conventional methods, they may still contribute to the overall predictive accuracy of the model. Lasso selects and retains variables based on their ability to improve the model’s performance as a whole and not based on their significance levels (Ranstam & Cook, 2018).

The results are shown in **Table 5** are from Data Frame 3, the data frame from which calculations were made to estimate regression models. Data frame 1 would be the complete database, and data frame 2 was constructed with all the numerical variables to identify the statistical distribution of each one, for which there are no model results.

Table 5. Main models result

| Variable description |
|---|
| **[1] Cargo Weight (Ton)** |
| **[2] Number of Employees (Un)** |
| **[3] Manual equipment (Un)** |
| **[4] Mechanical equipment (Un)** |
| **[5] Employees position 1 (Un)** |
| **[6] Employees position 2 (Un)** |
| **[7] Units per pallet (Un)** |
| **[8] Number of pallets (Un)** |
| **[9] Assisted Delivery (Dummy)** |
| **[10] Intercept (Min)** |

| Data Frames | Variable / Coefficients | $R^2$ | RMSE |
|---|---|---|---|
| 3 | [10] 65.814<br>[5] 9.910<br>[6] (1.609)<br>[1] 26.292<br>[7] 7.651<br>[3] 27.377<br>[4] 9.824<br>[9] (23.665) | 0.026 | 58.024 |
| 4 | [10] 65.814<br>[5] 9.854<br>[6] (1.771)<br>[1] 25.720<br>[8] 3.532<br>[3] 26.354<br>[4] 10.214<br>[9] (23.231) | 0.0229 | 58.413 |
| 5 | [10] 65.814<br>[1] 23.214<br>[7] 7.337<br>[2] 0.903<br>[3] 28.231<br>[4] 12.133<br>[9] (23.423) | -0.005 | 58.788 |
| 6 | [10] 65.814<br>[1] 17.072<br>[7] 4.237<br>[2] 7.993<br>[9] (6.707) | 0.103 | 62.06 |
| 7 | [10] 65.814<br>[1] 17.180<br>[8] 1.157<br>[2] 8.064<br>[9] (6.596) | 0.109 | 62.195 |
| 8 | [10] 65.814<br>[1] 22.758<br>[8] 2.836<br>[2] 1.031<br>[3] 27.139<br>[4] 12.444<br>[9] (22.904) | 0.001 | 59.165 |

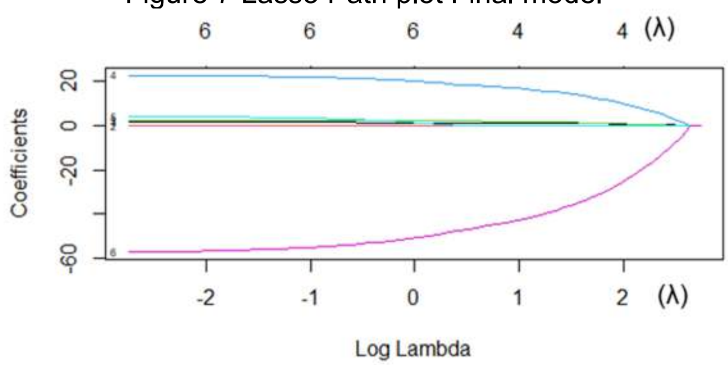
Categorical variables shown in **Table 2** were omitted in the model estimation since the Lasso regression did not interpret them correctly during preliminary tests, which generated errors in the R study. It is presumed that this could be due to the number of columns with binary values (one and zero), which could negatively affect the model's fit.

After estimating the models using all the data frames, the final model was selected based on the $R^2$ and RMSE values. After analyzing the results, it was identified that the model corresponding to data frame 3 presented the lowest RMSE value. However, the model derived from data frame 7 showed a considerably higher $R^2$ and a different minimum RMSE of approximately 4 minutes compared to data frame 3. Analyzing the balance between both indicators, model 7 is considered more robust and is therefore chosen as the final model, as is shown in **Table 6.**

Table 6 Characteristics of the final model

| Variables | Coefficients | $R^2$ | RMSE |
|---|---|---|---|
| Intercept (Min) | 65.814 | | |
| Cargo Weight (Ton) | 17.180 | | |
| Number of Pallets (Un) | 1.157 | 0.109 | 62.195 |
| Number of employees (Un) | 8.064 | | |
| Assisted Delivery (Dummy) | (6.596) | | |

Figure 7 Lasso Path plot Final model




When generating the Lasso path plot of the final model shown in **Figure 7,** it can be corroborated that as the Lambda penalty value increases, the model reaches an equilibrium point (Convergence) where the irrelevant variables are precipitated to zero, and the most important ones are maintained, fulfilling the purpose of the model. The values shown at the top and bottom on the x-axis correspond to the lambda values taken by the iterations until a convergence point is reached, and the lines shown in the graph are the behaviors of the coefficients of the variables when replacing that lambda value within the chosen Lasso model.

- **Model results analysis**

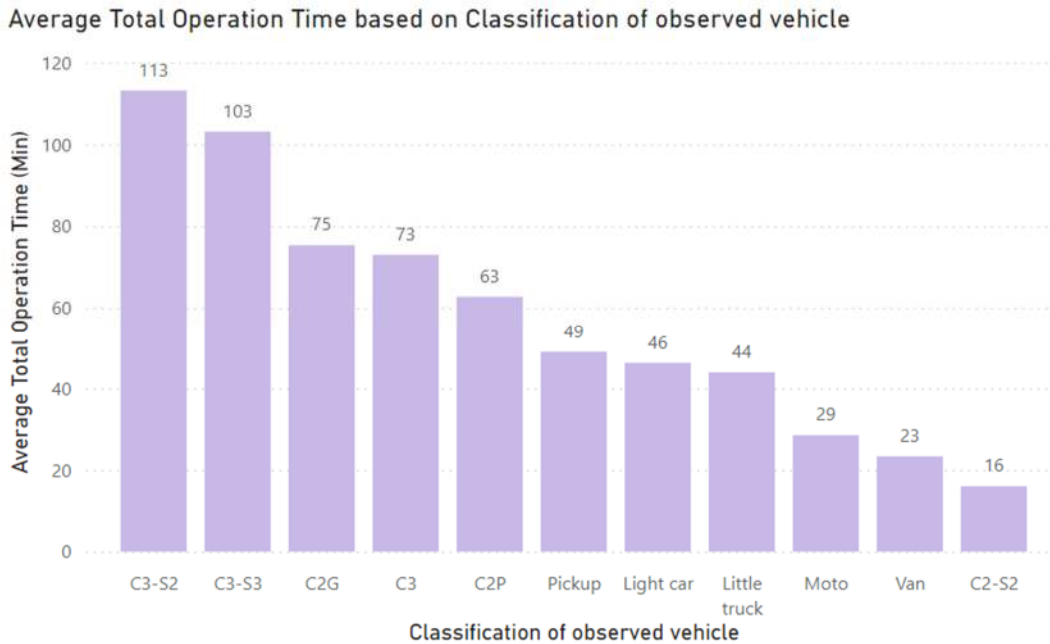
The Lasso model results in **Table 6** show that the significant variables that influence the total operating time for the loading and unloading of goods are statistically significant, and most are conceptually valid, except for one coefficient that has signs contrary to intuition, which is the number of employees.

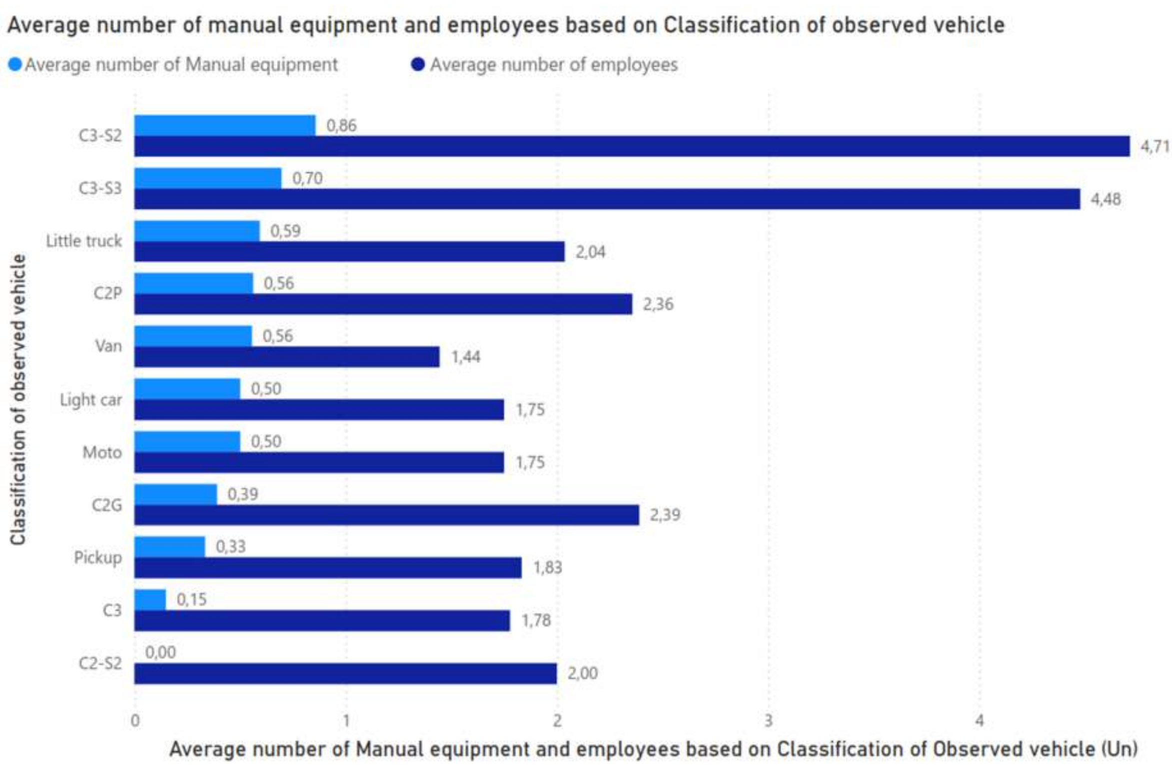
This coefficient can be explained by analyzing the behavior of the observations in the database. **Figure 8** shows the average operating times according to the truck type and the average number of employees involved.

The average total operating time graph based on vehicle type shows that the amount of time for the loading and unloading activity increases as the size of the vehicle increases. This finding can be explained as larger vehicles carry larger cargo volumes, implying more excellent resources for their load and unload. Thus, the model associates an increase in workers with more time spent on the loading and unloading operation. Although the data did not differentiate by truck type, these results highlight it as an essential factor to consider.

Examining the other variables that are part of the final model, it is considered that they have conceptual logic. The weight of the load in tons and the number of pallets, as their presence increases in a loading and unloading activity, implies a greater utilization of resources such as labor, equipment, etc., increasing the total duration of the activity. Conversely, suppose the equipment is used to assist in loading or unloading. In that case, the total time of the activity is reduced compared to if the activity were carried out without any equipment, which is why it is considered that these coefficients make sense and present conceptually valid signs.

Figure 8 Statistics related to the type of vehicle



## 4.7 Segmented Model Estimation

After analyzing the results of the general model and the behavior of two of the main variables shown in **Figure 8**, although the results are considered logical and meet the objective, it was decided to segment the database with 247 observations into two groups, as shown in **Table 7** due to their similar total operation time. Separating large and medium vehicles from small vehicles to analyze in more detail, reduce the data variance, and see if the level of fit improves.

Table 7 Segmented database by vehicle type

| Group 1: Large and Medium vehicles | Group 2: Small vehicles |
|---|---|
| C3-S3 | Moto |
| C3-S2 | Van |
| C2-S2 | Pickup |
| C3 | Little truck |
| C2G | Light Car |
| C2P | |
| **197 observations** | **50 observations** |

When dividing the database with 247 observations according to the type of vehicles, there is a first group with 197 observations and a second group with 50, as shown in **Table 7,** to be further divided into training and test datasets to calculate new segmented models. However, upon reviewing the group belonging to small trucks and highlighting that 20% of the database should be reserved for the test set, it is considered that the number of observations is too low to define a robust model for small vehicles. Therefore, only the segmented model for large and medium trucks is calculated.

The partitioning of the segmented database with 197 observations is carried out, of which 158 were assigned to the training set (80% of the database) and 39 observations were assigned to the test set to be used in the validation process (20% of the database).

Following the established guidelines for the complete database, Lasso regression models are estimated to use the data frames outlined in **Table 4**, employing the 197 observations of large and medium vehicles. Obtaining in this way the results are shown in **Table 8.**

Table 8 Main models result in a segmented database

| Variable description |
|---|
| **[1] Cargo Weight (Ton)** |
| **[2] Number of Employees (Un)** |
| **[3] Manual equipment (Un)** |
| **[4] Mechanical equipment (Un)** |
| **[5] Employees position 1 (Un)** |
| **[6] Employees position 2 (Un)** |
| **[7] Units per pallet (Un)** |
| **[8] Number of pallets (Un)** |
| **[9] Assisted Delivery (Dummy)** |
| **[10] Intercept (Number)** |

| Data Frames | Variable / Coefficients | $R^2$ | RMSE |
|---|---|---|---|
| 3 | [10] 72.478<br>[5] 14.014<br>[6] (1.085)<br>[1] 19.064<br>[7] 0.608<br>[3] 18.227<br>[4] 3.379<br>[9] (21.498) | 0.319 | 64.974 |
| 4 | [10] 72.478<br>[5] 13.750<br>[6] 0.376<br>[1] 21.728<br>[8] (8.292)<br>[3] 19.433<br>[4] 4.423<br>[9] (21.457) | 0.253 | 64.559 |
| 5 | [10] 72.478<br>[1] 14.510<br>[2] 3.037<br>[3] 18.725<br>[4] 6.323<br>[9] (20.828) | 0.332 | 66.262 |
| 6 | [10] 72.478<br>[1] 13.974<br>[7] (0.942)<br>[2] 7.144<br>[9] (14.407) | 0.198 | 68.174 |
| 7 | [10] 72.478<br>[1] 16.257<br>[8] (7.374)<br>[2] 8.970<br>[9] (13.878) | 0.143 | 67.865 |
| 8 | [10] 72.478<br>[1] 17.582<br>[8] (9.770)<br>[2] 5.187<br>[3] 20.074<br>[4] 7.477<br>[9] (20.724) | 0.256 | 65.705 |

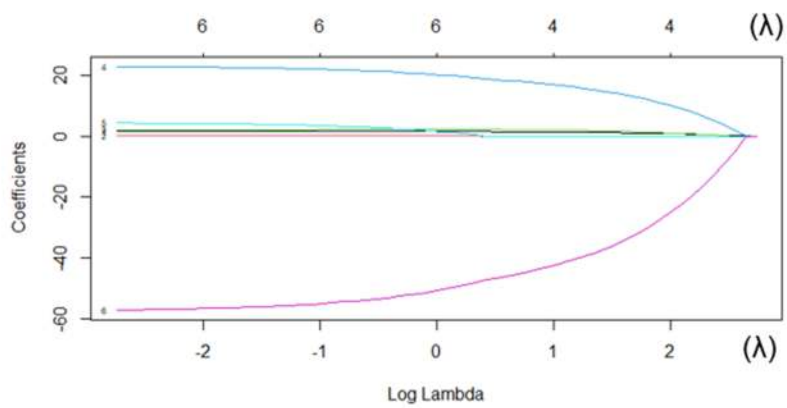
Analyzing the results of the segmented models, the $R^2$ values of the new models improved significantly compared to the initial models, which allows us to deduce that separating the database based on vehicle types contributed to enhancing the overall fit of the models.

When comparing the models with a higher $R^2$, the models from data frames 3 and 5 are compared as they show a better fit than the others. When comparing the two models, it is identified that the one with the lowest RMSE corresponds to data frame 3. However, upon reviewing the $R^2$ values, it is observed that the model resulting from dataset 5 has a higher $R^2$ value and a 2-minute RMSE difference compared to dataset 3. Therefore, it is considered more robust in comparison and is defined as the definitive segmented model, as shown in **Table 9.**

Table 9 Characteristics of the final segmented model

| Variables | Coefficients | $R^2$ | RMSE |
|---|---|---|---|
| Intercept (Min) | 72.478 | 0.332 | 66.262 |
| Cargo Weight (Ton) | 14.510 | | |
| Numer of employees (Un) | 3.037 | | |
| Manual Equipment (Un) | 18.725 | | |
| Mechanical Equipment (Un) | 6.323 | | |
| Assisted Delivery (Dummy) | (20.828) | | |

Figure 9 Lasso Path plot Final segmented model




Analyzing the Lasso path plot of the final segmented model shown in **Figure 9**, it can be corroborated that as the Lambda penalty value increases, the model reaches an equilibrium point (convergence) where the irrelevant variables are precipitated to zero, and the most important ones are maintained, fulfilling the purpose of the model.

- **Segmented model results analysis**

Finally, the model chosen as definitive for the segmented database is the one obtained with dataset 5, whose characteristics are shown in **Table 9**. Like the general model, it is considered that the variables have the correct signs, except for the number of workers, which is thought to have a counterintuitive sign. It could be inferred that the behavior of the workers is shown in **Figure 8.** This type of vehicle may influence the model to associate the increase in workers with a longer total operating time.

It is deemed an acceptable value to analyze the $R^2$. However, it is not very valuable considering the variability factors associated with transportation issues such as vehicle size, type of packaging, cargo weight, etc.

By performing the segmented model and comparing it with the general model, improvements in model fit can be seen. This demonstrates that the Lasso regression model can be used to estimate parking durations for loading and unloading activities satisfactorily.

## 4.8 Duration classification scheme

Creating a zoning scheme in Qgis begins by locating the points with the addresses of the 247 observations taken for the database. For the definition of class intervals in a frequency distribution, we implement the Sturges rule, which is commonly used in engineering to calculate the number of intervals and their width (Walpole et al., 2012). The Sturges rule and the Width (w) are shown in Equations 8 and 9, respectively:

$$k = 1 + 3.322 log_{10}(n) \qquad (8)$$

Where:

*k*: Number of classes
*n*: Amount of data in the database

$$w = \frac{range}{k}$$

Where:

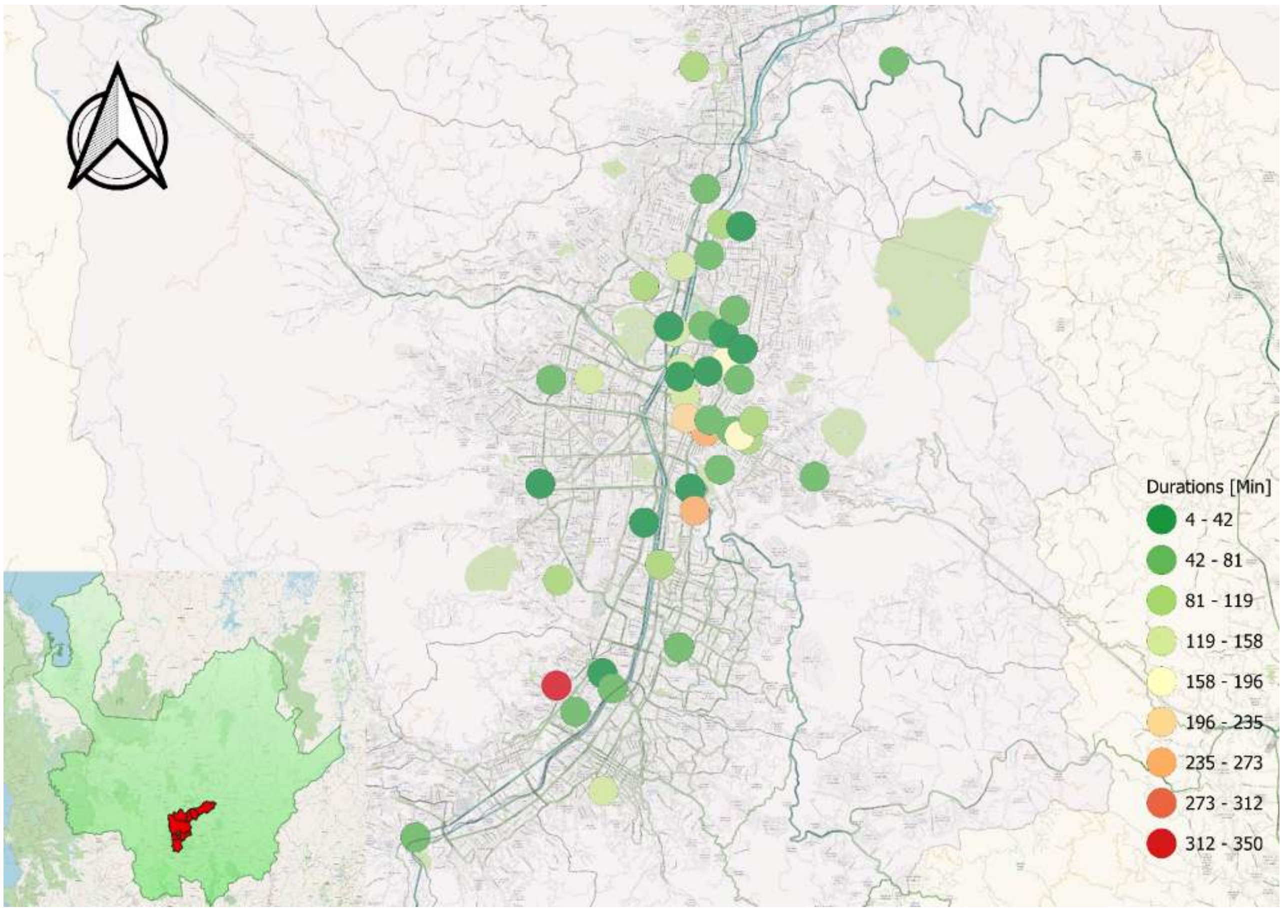
*w:* width of the class interval

*range*: difference between the maximum value and the minimum value of the database

*k*: Number of classes

When calculating the class intervals for the 247 observations, there is a total of 9 classes, each with an interval width of 38.44 minutes. Considering the calculated intervals, the representation is made as shown in **Figure 10.**

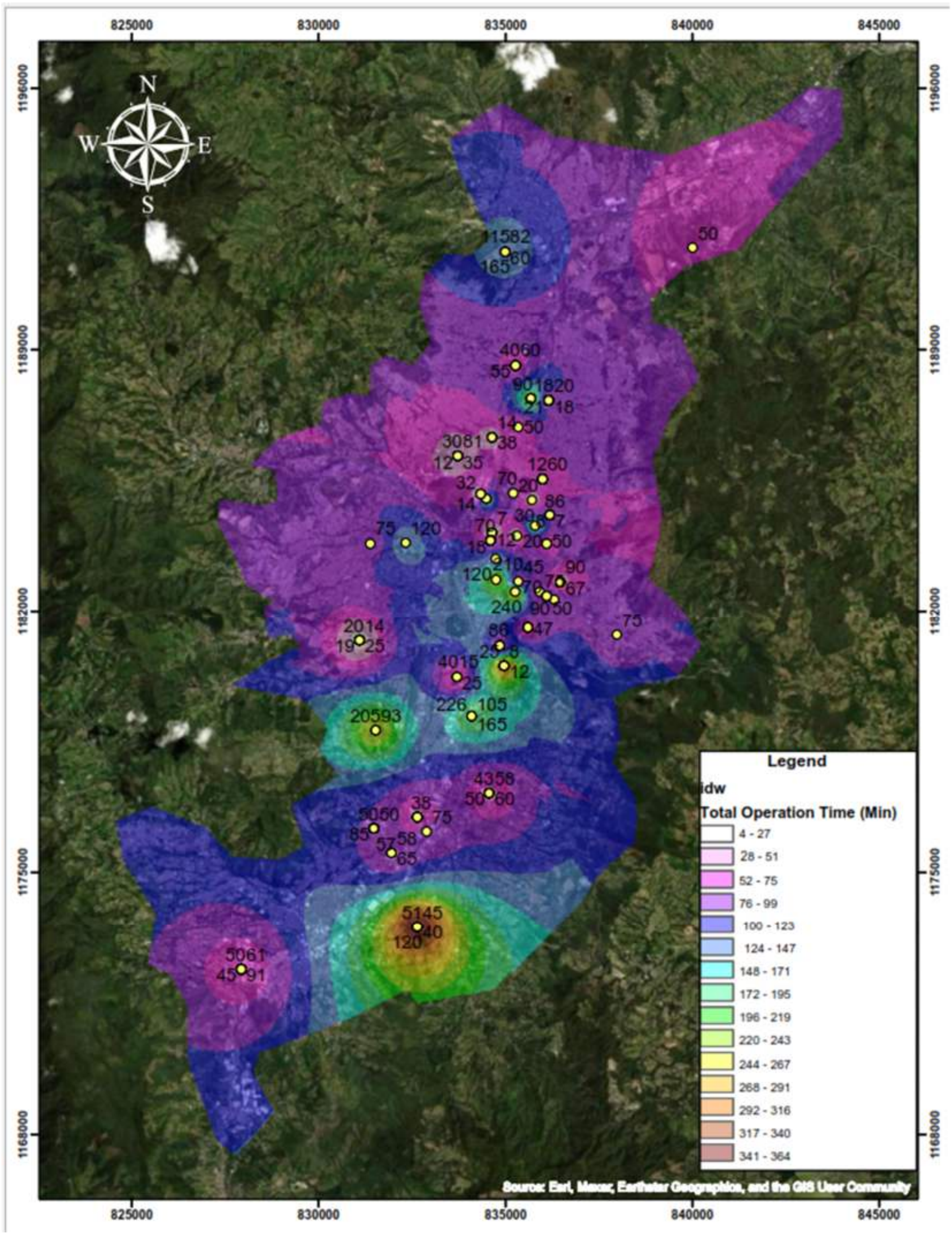
Figure 10 Observations of spatial localization




Then, an initial location of the points is made, and interpolation is carried out using a tool present in the QGIS program called Inverse Distance Weighting or IDW interpolation, which is a technique used in QGIS to create a raster surface from points with associated characteristics, in this case, the duration of parking.

For the interpolation, the total operation time associated with each point located on the map was taken as a reference, and the IDW interpolation method was applied to show the different variations in parking duration in the Metropolitan Area of Medellín, also considering the density of the points in the study areas, resulting in the map shown in **Figure 11.**

Figure 11 Parking duration classification scheme MMA




Upon analyzing the parking classification scheme, it is identified that the area with a concentration of points is in the center of Medellín. However, in this area, the spots with the most extended parking duration are not found, which leads to the understanding that the statement made by Schmid et al. (2018) could apply, where it mentions that vehicles parked in improper areas tend to take less time during deliveries. While in the Itagüí Zone, which is industrial, the maximum parking duration times are observed despite having a low concentration of observations. It is presumed that it is related to land use and the flow of vehicles that come to load or unload activities and demand parking spaces.

For the scheme's development, the equal intervals method was used for the class intervals, as it is widely used and allows for a uniform distribution of data across the class intervals. Fifteen intervals were chosen due to the variability of the data after conducting several tests with different intervals, it was considered that this amount allowed the zoning map to be displayed to make the information more accessible and enabled the reader to identify the zones more clearly, along with the labels of the observation IDs.

## 4.9 Proposed strategies

The study of possible public policies and strategies is a tool that contributes to well-planned decision-making aimed at the efficient management of public space in urban areas.

For example, Sanchez-Diaz et al. (2020) conducted a study focused on determining the variables that influence the efficiency of loading and unloading activities of commercial vehicles in urban areas, identifying that the equipment was a significant factor, and confirming the results obtained in the models of this research.

Therefore, one of the proposed strategies is to improve the type of equipment used in loading and unloading activities, as performing these activities with assistance significantly helps to reduce their duration and, in turn, decreases the time vehicles are parked, enhancing mobility.

Sharman et al. (2012) conducted pilot tests of public strategies that demonstrated the potential to address mobility issues within urban areas, highlighting the importance of studying parking demand in urban zones to guide public policies. As polite tests to implement within MMA, it is proposed to carry out strategies such as:

- Off-hour deliveries, as demonstrated, help reduce congestion during peak hours of the day and decrease emissions of polluting gases.
- Establishment of urban consolidation centers (UCC) that allow for reduced circulation of commercial vehicles in urban areas and implement more sustainable alternatives such as cargo bicycles, which do not require a large parking space.

On the other hand, there are the results presented by Castrellon & Sanchez-Diaz (2024), where they applied strategies such as:

- Assignment of loading and unloading spaces at the edges of platforms can significantly reduce delivery times by eliminating factors such as long walks for drivers to reach their destination with the goods.
- Sharing information by public space managers, such as drivers and customers, through technology to help drivers find available parking spaces and reduce cruising times in search of parking, thereby decreasing negative externalities such as congestion and emissions.
- Limiting the parking duration, this alternative can help drivers focus on completing their delivery within a specified time, as distractions often arise or time optimization during delivery is not seen as necessary.
- Implement and monitor the established regulations to control loading and unloading activities to help prevent illegal parking and streamline vehicle flow.

Finally, we can find strategies like those proposed by Zou et al. (2016), which are more related to the lack of loading and unloading spaces in urban areas and the difference in characteristics such as:

- Making deliveries during nighttime hours to reduce congestion during the day and, in turn, other related externalities such as accidents and pollutant emissions.
- Improve the management of sidewalks and loading zones through programs such as regulated parking times and space distribution schemes, which mainly benefit vehicles with longer delivery times.
- Implementation of smaller vehicles for deliveries, favoring mobility in areas with higher traffic flows.
- Designing specific policies for sectors; this particular strategy is considered very relevant because it can be seen on the zoning map that there are sectors with a higher number of observations and more critical durations. This can be related to factors such as the type of merchandise and the type of vehicle, which highlights the importance of generating strategies according to the needs and conditions of areas.

# 5 Conclusions and Recommendations

This study has addressed parking duration for commercial vehicles engaged in loading and unloading activities, particularly within Medellin's Metropolitan Area. By employing a Lasso regression model, we have gained valuable insights into the factors that significantly influence parking duration. Based on 247 observations of truck operations, 80% were for training Lasso models and 20% for testing estimated models. Our analysis highlights critical variables such as cargo weight and the types of equipment used, which are crucial in determining how long commercial vehicles occupy parking spaces.

The approach of not removing outliers but instead applying the Winsorization method effectively preserved the integrity of the dataset while mitigating the influence of extreme values. The model's accuracy was improved by addressing outliers in this way and conducting a thorough correlation analysis to eliminate highly correlated variables. This methodological choice ensured that the final Lasso regression model reflected the relationships between parking duration variables more accurately without being skewed by atypical data or multicollinearity issues.

Utilizing Lasso regression for model estimation proved to be a robust choice for this study. The Lasso regression technique, which performs both variable selection and regularization, enabled the identification of critical variables that influence parking duration while reducing overfitting. This approach aligns well with the research objective of developing a practical model for estimating parking duration for commercial vehicles in urban areas of the Global South, highlighting the method's effectiveness in managing complex, real-world data scenarios.

The findings indicate that the Lasso regression model is an effective tool for understanding and predicting parking duration. It offers a robust framework for analyzing and managing urban freight operations. The model’s ability to capture influential variables provides a solid foundation for developing targeted policies and strategies to address parking challenges faced by commercial vehicles.

This work provided some methodological insights from applying the Lasso regression model. We found that identifying variables with a low correlation with the objective function before developing the regression is critical. These can negatively affect the model results due to the added noise during data analysis that affects the prediction adjustment of the final model.

Although categorical variables can be expressed as dummies, when integrated into data frames to estimate models, they typically increase the prediction error, possibly due to the presence of ones and zeros that impact the prediction adjustment concerning other data. As a result, we suggest implementing continuous variables to estimate the final model after exploring their correlation with the target variable using a statistical coefficient.

This paper's findings indicate that the Lasso regression model effectively identified and prioritized variables with a strong correlation to total operating time, demonstrating its utility in refining the model by excluding less relevant variables. By focusing on variables with meaningful relationships to total operating time and using the root mean square error (RMSE) and $R^2$ for model comparison, the study determined that Data Frame 7 provided the most accurate model with an RMSE of 62.195 minutes and an $R^2$ value of 0.109. This model offered the best performance in predicting loading and unloading times, indicating that Lasso regression can successfully enhance the precision of parking duration estimates in urban areas.

Upon reviewing the results of the estimated lasso regression models, both general and segmented, we suggest using the segmented model to estimate the loading and unloading durations of commercial trucks in urban areas for large and medium-sized vehicles due to its goodness of fit and the lower prediction error obtained concerning the general model. For the case of small trucks, it is necessary to estimate a model with more information that better represents the study population.

The segmented model obtained with large and medium commercial vehicles, which is recommended for estimating loading and unloading times, has the following structure:

$$TOT = 72.478 + 14.510[1] + \ 3.037\ [2] + 18.725\ [3] + 6.323\ [4] - 20.828\ [5]$$

Where:

TOT= Total Operation Time (Min)

[1] = Cargo Weight (Ton)

[2] = Number of employees (Un)

[3] = Manual Equipment (Un)

[4] = Mechanical Equipment (Un)

[5] = Assisted Delivery (Dummy)

After implementing the segmentation of the database into large and medium vehicles on the one hand and small vehicles on the other, a model derived from data frame 5 was obtained, with an $R^2$ value of 0.332 and an RMSE of 66.262 min, demonstrating that a more detailed analysis of the type of vehicle yields better results. For small vehicles, gathering more information is recommended to determine a more robust model with a more significant number of observations.

Analyzing the developed zonal classification scheme, it is identified that loading and unloading observations conducted near the city's central area tend to take longer than those performed in areas closer to the edges of the Medellin Metropolitan Area. Except for the observations in the municipality of Itagui, where durations of 350 minutes are noted, this can be explained because this area is industrial and has multiple factories and industrial warehouses, for which most of the observations are of large trucks.

The longer durations near the city center can be explained by the fact that drivers have less space to carry out loading and unloading activities, in addition to a high flow of both commercial and private vehicles and pedestrians, which can hinder the development of the activity.

To achieve the objective of proposing strategies that improve the durations of loading and unloading commercial vehicles in urban areas, different methods that have been proposed in pilot tests in various parts of the world were studied. However, it is considered that by studying the conditions present in the Metropolitan Area of Medellin, it is concluded that those that could be replicated in the study area are the ones proposed by Zou et al. (2016).

Due to space limitations, nighttime delivery strategies are recommended, as shown in the literature review, to efficiently reduce pollutant emissions and accident levels for users related to logistics activities.

Additionally, given the limited space, it is crucial to implement the available area in the most efficient way possible, so it is considered that proposing schemes and strategies to manage the curb parking can help improve mobility in the city center and reduce delivery times for goods.

Finally, due to the variability in parking times in the metropolitan area zones, it is recommended that strategies be devised through a strategic division depending on the needs of the metro area zones. This is to establish more efficient policies and strategies that improve mobility in the study area and delivery times.

Future research should build upon these findings by exploring additional variables and testing the model in different urban contexts. This study not only provides a detailed analysis of factors influencing parking duration but also demonstrates the potential of Lasso regression in advancing the field of urban freight management. By leveraging these insights, policymakers and urban planners can develop more effective solutions to alleviate the challenges associated with commercial vehicle parking, ultimately contributing to a more efficient and sustainable urban transportation system.

# 6 References


Comunas de Medellín. (2024, April 13). *Medellín firma acuerdo para incentivar el desarrollo económico del país desde las regiones*. Comunasme Comunicación Digital de Medellín. https://comunasdemedellin.com/medellin-firma-acuerdo-para-incentivar-el-desarrollo-economico-del-pais-desde-las-regiones/

Alho, A., Oh, S., Seshadri, R., Dalla Chiara, G., Chong, W. H., Sakai, T., Cheah, L., & Ben-Akiva, M. (2022). An agent-based simulation assessment of freight parking demand management strategies for large urban freight generators. *Research in Transportation Business and Management*, *43*. https://doi.org/10.1016/j.rtbm.2022.100804

Amaya, J., Encarnación, T., & Delgado-Lindeman, M. (2023). Understanding Delivery Drivers' Parking Preferences in Urban Freight Operations. *Transportation Research Part A: Policy and Practice*, *176*. https://doi.org/10.1016/j.tra.2023.103823

Área metropolitana del valle de Aburrá. (2019). *Municipios, información básica del territorio a nivel Municipio*. Aburrá Datos. https://www.metropol.gov.co/observatorio/Paginas/areadatos-mapa.aspx#HeaderMap

Área Metropolitana del Valle de Aburrá, & Universidad Nacional de Colombia. (2018). *Estudio del transporte de carga en el Área Metropolitána del Valle de Aburrá*.

Barbero, J. A., & Guerrero, P. (2017). *El transporte automotor de carga en américa latina, soporte logístico de la producción y el comercio*. www.iadb.org

Bartlett, J. E., Kotrlik, J. W., & Higgins, C. C. (2001). Organizational Research: Determining Organizational Research: Determining Appropriate Sample Size in Survey Research Appropriate Sample Size in Survey Research. In *Information Technology, Learning, and Performance Journal* (Vol. 19, Issue 1).

Burns, A. J., Michalek, J. J., & Samaras, C. (2024). Estimating the potential for optimized curb management to reduce delivery vehicle double parking, traffic congestion and energy consumption. *Transportation Research Part E: Logistics and Transportation Review*, *187*. https://doi.org/10.1016/j.tre.2024.103574

Campbell, S., Holguín-Veras, J., Ramirez-Rios, D. G., González-Calderón, C., Kalahasthi, L., & Wojtowicz, J. (2018). Freight and service parking needs and the role of demand management. *European Transport Research Review*, *10*(2). https://doi.org/10.1186/s12544-018-0309-5

Castrellon, J. P., & Sanchez-Diaz, I. (2024). Effects of freight curbside management on sustainable cities: Evidence and paths forward. *Transportation Research Part D: Transport and Environment*, *130*. https://doi.org/10.1016/j.trd.2024.104165

Castrellon, J. P., Sanchez-Diaz, I., & Kalahasthi, L. K. (2023). Enabling Factors and Durations Data Analytics for Dynamic Freight Parking Limits. In *Transportation Research Record* (Vol. 2677, Issue 2, pp. 219–234). SAGE Publications Ltd. https://doi.org/10.1177/03611981221115086

Chatziioannou, I., Alvarez-Icaza, L., Bakogiannis, E., Kyriakidis, C., & Chias-Becerril, L. (2020). A structural analysis for the categorization of the negative externalities of transport and the hierarchical organization of sustainable mobility's strategies. *Sustainability (Switzerland)*, *12*(15). https://doi.org/10.3390/su12156011

Dalla Chiara, G., & Goodchild, A. (2020). Do commercial vehicles cruise for parking? Empirical evidence from Seattle. *Transport Policy*, *97*, 26–36. https://doi.org/10.1016/j.tranpol.2020.06.013

Federal Highway Administration. (2017, January 31). *Introduction To Vehicle Classification*. Office of Research, Development, and Technology, Office of Infrastructure, RDT. https://www.fhwa.dot.gov/publications/research/infrastructure/pavements/ltpp/13091/002.cfm

Feng, N., Zhang, F., Lin, J., Zhai, J., & Du, X. (2019). Statistical Analysis and Prediction of Parking Behavior. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11783 LNCS*, 93–104. https://doi.org/10.1007/978-3-030-30709-7_8

Gao, J., & Ozbay, K. (2017). A data-driven approach to estimate double parking events using. *Transportation Research Board 96th Annual Meeting*, 1–15.

Ghizzawi, F., Galal, A., & Roorda, M. J. (2024). Modelling parking behaviour of commercial vehicles: a scoping review. *Transport Reviews*. https://doi.org/10.1080/01441647.2024.2305202

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate data analysis* (7th ed.). Pearson Education Limited.

Holguin-Veras, J., Jaller, M., Destro, L., Ban, X., Lawson, C., & Levinson, H. (2011). Freight generation, freight trip generation, and the perils of using constant trip rates. *Sage Journal*, *Volume 2224*(Issue 1).

James, G., Gareth, M., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning : with applications in R*. Springer. https://doi.org/10.1007/978-1-4614-7138-7

Kalahasthi, L. K., Sánchez-Díaz, I., Pablo Castrellon, J., Gil, J., Browne, M., Hayes, S., & Sentís Ros, C. (2022). Joint modeling of arrivals and parking durations for freight loading zones: Potential applications to improving urban logistics. *Transportation*

*Research Part A: Policy and Practice*, *166*, 307–329. https://doi.org/10.1016/j.tra.2022.11.003

Kish, L. (1995). *KISH Survey Sampling*. Wiley-Interscience.

Kwak, S. K., & Kim, J. H. (2017). Statistical data preparation: Management of missing values and outliers. *Korean Journal of Anesthesiology*, *70*(4), 407–411. https://doi.org/10.4097/kjae.2017.70.4.407

López-Roldán, P., & Fachelli, S. (2015). *Metodología de la investigación social cuantitativa: Vol. 1st edition*. Universitat Autónoma de Barcelona.

Greenberg, B. (2009). *Introduction to the Practice of Statistics Sixth Edition* (6th ed.). W.H. Freeman and Company.

Morin, J. F., Olsson, C., & Atikcan, E. O. (2021). *Research methods in the social sciences an A-Z of key concepts*. Oxford University Press.

Müller, A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python a guide for data scientists* (1st ed.). O'Reilly Media.

Murthy, P., Xie, M., & Jiang, R. (2004). *Weibull Models* (D. Blading & P. Bloomfield, Eds.). John Wiley & Sons, Inc.

Ortúzar, J. de D., & Willumsen, L. G. (2011). *Modelling Transport* (J. de D. Ortuzar & L. G. Willumsen, Eds.; 4th ed.). WILEY John Wiley & Sons, Ltd.

Otzen, T., & Manterola, C. (2017). Técnicas de Muestreo sobre una Población a Estudio Sampling Techniques on a Population Study. In *Int. J. Morphol* (Vol. 35, Issue 1).

Parmar, J., Das, P., & Dave, S. M. (2021). A machine learning approach for modelling parking duration in urban land-use. *Physica A: Statistical Mechanics and Its Applications*, *572*. https://doi.org/10.1016/j.physa.2021.125873

Pinto, R., Lagorio, A., & Golini, R. (2019). The location and sizing of urban freight loading/unloading lay-by areas. *International Journal of Production Research*, *57*(1), 83–99. https://doi.org/10.1080/00207543.2018.1461269

Profillidis, V. A., Botzoris, G. N., & Galanis, A. T. (2014). Environmental Effects and Externalities from the Transport Sector and Sustainable Transportation Planning-A Review. *International Journal of Energy Economics and Policy*, *4*(4), 647–661. www.econjournals.com

Ramirez-Rios, D. G., Kalahasthi, L. K., & Holguín-Veras, J. (2023). On-street parking for freight, services, and e-commerce traffic in US cities: A simulation model incorporating demand and duration. *Transportation Research Part A: Policy and Practice*, *169*. https://doi.org/10.1016/j.tra.2023.103590

Ranstam, J., & Cook, J. A. (2018). Lasso regression. *British Journal of Surgery*, *105*(10), 1348. https://doi.org/10.1002/bjs.10895

Raschka, Sebastian., & Mirjalili, Vahid. (2017). *Python machine learning : machine learning and deep learning with Python, scikit-learn, and TensorFlow*. Packt Publishing.

Reed, S., Campbell, A. M., & Thomas, B. W. (2024). Does parking matter? The impact of parking time on last-mile delivery optimization. *Transportation Research Part E: Logistics and Transportation Review*, *181*. https://doi.org/10.1016/j.tre.2023.103391

Sanchez-Diaz, I., Palacios-Argüello, L., Levandi, A., Mardberg, J., & Basso, R. (2020). A time-efficiency study of medium-duty trucks delivering in urban environments. *Sustainability (Switzerland)*, *12*(1). https://doi.org/10.3390/SU12010425

Schmid, J., Wang, X. (Cara), & Conway, A. (2018). Commercial vehicle parking duration in New York City and its implications for planning. *Transportation Research Part A: Policy and Practice*, *116*, 580–590. https://doi.org/10.1016/j.tra.2018.06.018

Scott, L. M., & Janikas, M. V. (2010). Spatial Statistics in ArcGIS. In *Handbook of Applied Spatial Analysis* (pp. 27–41). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-03647-7_2

Sharman, B. W., Roorda, M. J., & Habib, K. M. N. (2012). Comparison of parametric and nonparametric hazard models for stop durations on urban tours with commercial vehicles. *Transportation Research Record*, *2269*, 117–126. https://doi.org/10.3141/2269-14

Silva, K., Labegalini, A., Lima, R. da S., & Yushimito, W. (2024). Improving freight parking needs in Brazilian cities with city logistics initiatives. *Case Studies on Transport Policy*, *16*. https://doi.org/10.1016/j.cstp.2024.101184

U.S. Department of transportation. (2022). *Commodity Flow Survey Standard Classification of Transported Goods (SCTG) SCTG COMMODITY CODES*.

Wackerly, D. D., Mendenhall III, W., & Scheaffer, R. L. (2008). *Mathematical statistics with applications* (7th ed.). Engage Learning.

Walpole, R. E., Myers, R., Myers, S., & Ye, K. (2012). *Probability & Statistics for Engineers & Scientists* (Vol. 9th). Pearson Education, Inc.

Washington, Simon., Karlaftis, M. G., & Mannering, F. L. (2003). *Statistical and econometric methods for transportation data analysis*. Chapman & Hall/CRC.

World Bank Group. (2024). *World Bank Country and Lending Groups*. https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups

Xiao, R. I., & Jaller, M. (2023). Prediction framework for parking search cruising time and emissions in dense urban areas. *Transportation*. https://doi.org/10.1007/s11116-023-10455-4

Yang, S., Ma, W., Pi, X., & Qian, S. (2019). A deep learning approach to real-time parking occupancy prediction in transportation networks incorporating multiple spatio-temporal data sources. *Transportation Research Part C: Emerging Technologies*, *107*, 248–265. https://doi.org/10.1016/j.trc.2019.08.010

Zou, W., Wang, X. (Cara), Conway, A., & Chen, Q. (2016). Empirical Analysis of Delivery Vehicle On-Street Parking Pattern in Manhattan Area. *Journal of Urban Planning and Development*, *142*(2). https://doi.org/10.1061/(ASCE)UP.1943-5444.0000300

## APPENDIX 1. SAMPLE SIZE ESTIMATION METHODS

With the aim of better understanding the formulas for sample size, and the concepts of finite and infinite population. When defining it from the sampling perspective, the difference between a finite or infinite population lies in the relative importance of the sample size n concerning the population size N. If the sample represents less than 5% of the population, or it has more than 100,000 units it is considered an infinite population; otherwise, it is considered finite (López-Roldán & Fachelli, 2015).

Within this appendix, the methods for calculating sample size for surveys are specified along with their respective advantages and disadvantages, as well as the justification for the selection of the chosen method.

### 1. SIMPLE RANDOM SAMPLING

This is the most basic formula for calculating the sample size for a finite or infinite population. It is used when one seeks to represent a population with a specific level of confidence and margin of error (Otzen & Manterola, 2017).

**Formula**:

$$n = \frac{Z^2 * p * q}{E^2}$$

Where:
n: Sample size
Z: Value of the standard normal distribution
p: Expected rate of affirmative response to the survey conducted (determined by the investigator's judgment or information from previous surveys).
q: Expected negative response rate to the survey conducted (determined by the investigator's judgment or information from previous surveys).
E: Maximum estimation error

➢ **Advantages**

- Ease of understanding and application.
- Provide a simple calculation based on the confidence level and margin of error, which are easily accessible data because they are defined by the author.
- It is applicable when there is not much information regarding the study population.

➢ **Disadvantages**

- It does not consider the size of the population, which can lead to overestimating the sample size and wasting resources.
- A uniform distribution of the data is assumed, which may not be realistic if the population variability is high.

### 2. FINITE POPULATIONS METHOD

When working with finite populations, the classical formula is adjusted to consider the total population size. (López-Roldán & Fachelli, 2015).

**Formula:**

$$n = \frac{N * Z^2 * p * q}{E^2 * (N - 1) + Z^2 * p * q}$$

Where:

n: Sample size
N: Population size
Z: Value of the standard normal distribution
p: Expected rate of affirmative response to the survey conducted (determined by the investigator's judgment or information from previous surveys).
q: Expected negative response rate to the survey conducted (determined by the investigator's judgment or information from previous surveys).
E: Maximum estimation error

- **Advantages**

- It offers a more accurate estimate compared to the previous equation by knowing the total population size, reducing the total number of surveys needed.
- It allows for a more efficient implementation of resources by not overestimating the sample size.

- **Disadvantages**

- It requires knowing the total size of the population, which is not always easy to determine.
- It still depends on the level of confidence and the margin of error, which can lead to suboptimal decisions if these values are not selected properly.
- Assume that the population is homogeneous, which means there is no variability among the observations; if this is the case, the required sample size to obtain a representative sample of the types of variables of interest would be underestimated.

It is used when the population is divided into subgroups (strata) and it is necessary to ensure that each stratum is represented in the sample. The sample size in each stratum can be calculated proportionally to the size of the stratum (Bartlett et al., 2001).

**Formula (for an i stratum):**

$$n_i = \frac{N_i}{N} * n$$

Where:

$n_i$: Stratum i sample size
$N_i$: Size of stratum i
N: Population size
n: sample size (Calculated with the simple random sample method)

- **Advantages**

- Ensures that all subgroups of the population are represented in the sample.
- It reduces the estimation error if the strata are correctly defined and vary significantly from each other.

- **Disadvantages**

- It is more complex to implement and requires detailed knowledge of the population strata.
- If the strata are not adequately defined or there are no significant differences between them, the benefit of the method is reduced.
- He needs more time and resources to design the sample and collect the data.

### 3. KISH'S METHOD

This method for determining the sample size is used when planning to design a complex survey, that is, one that includes additional sources of variability, such as stratification, clusters, or other subdivisions that do not present in simple random sampling. This generates a factor defined as the "Design Factor" that must be considered in the sample size calculation to avoid underestimating or overestimating the necessary number of surveys (Kish, 1995).

**Formula:**

$$n_{adjusted} = n * DEFF$$

Donde:

$n_{adjusted}$: Adjusted sample size
$n$: Sample size if it were a simple random sample
$DEFF$: design effect

$$DEFF = 1 + \rho(b - 1)$$

Where:

$\rho$: Intraclass correlation coefficient, which measures the homogeneity within clusters.

$b$: It is the average cluster size or the average number of observations per stratum or cluster.

- DEFF = 1 implies that there is no increase in variability, that is, the complex survey behaves the same as a simple random survey.
- DEFF > 1 indicates that the survey design has increased variability.
- DEFF < 1 is rare, but it can occur if the design is very efficient, which reduces variability.

➢ **Advantages**

- It applies to more complex sample models as it considers the "design effect," improving accuracy.
- It is useful when different types of sampling, such as stratified and cluster sampling, are combined.
- It allows obtaining valid estimates when the sample is not completely random.

➢ **Disadvantages**

- Requires advanced knowledge of survey design and how to calculate the design effect (DEFF)
- More difficult to implement and justify if there is no experience with complex surveys
- Increases the sample size compared to simple random sampling.

### 4. CLUSTER SAMPLING METHOD

It is used when the population is naturally clustered in groups (neighborhoods, schools, etc.). Instead of sampling individuals directly, clusters are selected and then everyone within them is surveyed or a subsample is taken. The sample size calculations depend on the variability between clusters and how many clusters are selected (Morin et al., 2021).

**Formula:**

$$\boldsymbol{ncl = \frac{n * DEFF}{m}}$$

Where:

ncl: Adjusted sample size for each cluster
n: sample size calculated through simple random sampling
*DEFF:* Desing effect defined in Kish´s Method
*m*: Number of individuals within each cluster

➢ **Advantages**

- It can be more economical and practical, especially if the population is geographically dispersed.
- It is useful when it is not feasible to survey individuals from the entire population.
- It allows grouping sampling units to reduce field costs.

➢ **Disadvantages**

- It increases the variability within the clusters (bias), which may require a larger sample size to achieve the same precision.
- It is less precise than simple random sampling if the clusters are not homogeneous.

## 5. VARIANCE METHOD FOR INFINITE POPULATIONS

The variance method equation or formula for estimating the sample size in infinite populations is commonly used in sampling studies when the total population size is unknown (Ortúzar & Willumsen, 2011).

**Formula:**

$$n = \frac{S^2}{Se(\tilde{x})^2}$$

Where:

n: Sample size
$S^2$: Population variance, which measures the dispersion of the data concerning its mean.
$Se(\tilde{x})$: The variance is the square of the standard deviation.
Standard error of the mean, which is a measure of how accurately a sample represents the entire population. This value is related to the level of confidence and the margin of error allowed in the study.

additionally, the correction for a finite population is given in the following expression:

$$n = \frac{n'}{1 + \frac{n'}{N}}$$

Where:
n: Adjusted sample size for a finite population
n': Sample size for an infinite population estimated with the previous equation
N: Population size

## 6. CARGO STUDY DEVELOPED IN THE MEDELLIN METROPOLITAN AREA

The information presented in the freight transportation study developed in 2018 by Área Metropolitana del Valle de Aburrá & Universidad Nacional de Colombia (2018), where equations for the sample size of finite and infinite populations are proposed, is reviewed below:

- **INFINITE POPULATION (N>2000)**

$$n_0 = \frac{Z^2_{\propto/2^{s^2}}}{e^2}$$

Where:

$n_0$: Sample Size
$Z$: Value of the normal distribution at the significance of a
$a$: significance level
$s$: Standard deviation
$e$: Admissible error

- **FINITE POPULATION (N<2000)**

The correction for a finite population is expressed as follows:

$$n = \frac{n_0 N}{n_0 + (N - 1)}$$

Where:

n: Sample Size
$n_0$: Number of establishments in each area
$N$: Population considered for a specific zone

When reviewing the methods to determine the sample size, the advantages and disadvantages of each were considered; however, the method for finite populations was selected. This is because, by knowing the size of the study population, a more precise estimation of the sample size can be made, allowing for more efficient use of technical resources while maintaining the precision and quality of the research.

A summary table is presented with the methods studied in this appendix along with their respective formulas and applicability for finite and/or infinite populations.

Table 10 Summary sample size methods

| | | POPULATION APPLICABILITY | |
|---|---|---|---|
| **METHOD** | **EQUATION** | **FINITE** | **INFINITE** |
| **SIMPLE RANDOM SAMPLING** | $n = \frac{Z^2 * p * q}{E^2}$ | YES | YES |
| **FINITE POPULATION METHOD** | $n = \frac{N * Z^2 * p * q}{E^2 * (N-1) + Z^2 * p * q}$ | YES | NO |
| **STRATIFIED SAMPLING** | $n_i = \frac{N_i}{N} * n$ | YES | YES |
| **KISH'S METHOD** | $n_{adjusted} = n * DEFF$ <br> $DEFF = 1 + \rho(b-1)$ | YES | YES |
| **CLUSTER SAMPLING METHOD** | $ncl = \frac{n * DEFF}{m}$ | YES | YES |
| **VARIANCE METHOD FOR INFINITE POPULATION** | $n = \frac{S^2}{Se(\tilde{x})^2}$ | NO | YES |
| **VARIANCE METHOD FOR FINITE POPULATION** | $n = \frac{n'}{1 + \frac{n'}{N}}$ | YES | NO |
| **INFINITE POPULATION (N>2000)** | $n_0 = \frac{Z^2_{\propto/2} s^2}{e^2}$ | NO | YES |
| **FINITE POPULATION (N<2000)** | $n = \frac{n_0 N}{n_0 + (N-1)}$ | YES | NO |

# References

Área Metropolitana del Valle de Aburrá, & Universidad Nacional de Colombia. (2018). *Estudio del transporte de carga en el Área Metropolitána del Valle de Aburrá*.

Bartlett, J. E., Kotrlik, J. W., & Higgins, C. C. (2001). Organizational Research: Determining Organizational Research: Determining Appropriate Sample Size in Survey Research Appropriate Sample Size in Survey Research. In *Information Technology, Learning, and Performance Journal* (Vol. 19, Issue 1).

Kish, L. (1995). *KISH Survey Sampling*. Wiley-Interscience.

López-Roldán, P., & Fachelli, S. (2015). *Metodología de la investigación social cuantitativa: Vol. 1st edition*. Universitat Autónoma de Barcelona.

Morin, J. F., Olsson, C., & Atikcan, E. O. (2021). *Research methods in the social sciences an A-Z of key concepts*. Oxford University Press.

Otzen, T., & Manterola, C. (2017). Técnicas de Muestreo sobre una Población a Estudio Sampling Techniques on a Population Study. In *Int. J. Morphol* (Vol. 35, Issue 1).

Ortúzar, J. de D., & Willumsen, L. G. (2011). *Modelling Transport* (J. de D. Ortuzar & L. G. Willumsen, Eds.; 4th ed.). WILEY John Wiley & Sons, Ltd.