ANALYSIS OF TIME AND COST DEVIATIONS IN SECONDARY ROAD CONSTRUCTION PROJECTS IN COLOMBIA IN THE LAST TEN YEARS

Eduardo Rojas[1], **Jairo Pelaez** [2], **Adriana Gomez**[3]

Department of Civil Engineering, School of Engineering, Pontificia Universidad Javeriana Bogota, Colombia

E-mail Addresses: [1] rojase.d@javeriana.edu.co, [2] pelaez.jairo@javeriana.edu.co, [3] adrianagomez@javeriana.edu.co.

Authors

Abstract: The research project consisted of developing statistical tools that allow the identification of causes and factors that generate deviations in time and cost during the execution of secondary road construction projects in Colombia. The first part of the research project included a review of existing literature; followed by the data collection of projects executed and liquidated in Colombia based on the information available in the open data platform of Colombian government for the last ten years. Subsequently, an exploratory data analysis was performed, mainly of analyzing the behavior of each of the identified variables, both categorical and numerical, through the use and application of statistics. After performing the exploratory data analysis, bivariate and multivariate analysis was carried out, to determine the significant variables (high correlated variables) for time and cost overruns. This allowed the identification of specific parameters for readers to consider before starting the execution of such road infrastructure projects. Results revealed that generally projects frequently present more time deviations than cost deviations; also it was demonstrated that the order of magnitude for cost overruns is significantly lower compared with time overruns. Based on the bivariate analysis it was determined that projects with an estimated budget lower than 1,250 million USD and durations ranging between 100 and 200 days, usually present overruns in cost and time.

Keywords: cost overruns; time overruns; roads; public projects; project management; Public procurement; Statistical tools; Road infrastructure, open data.

1. Introduction

Cost overruns in construction are a latent problem; they occur globally and locally. However, the construction sector at the international level is one of the most important for any developing country, and this is partly due to the size of this industry compared to others, some authors agree that cost overruns are present in 3 out of 5 projects [1]. Infrastructure projects worldwide and nationally frequently present cost deviations due to multiple factors such as changes in scope, lack of planning, design changes, etc.; these causes may be compensable or non-compensable and attributable or not attributable to the project's stakeholders [2].

In the case of Europe, a study in Norway showed that road construction projects have cost overruns of 7.9%, ranging from -59% to +183% [3]. In Germany, it was identified from an analysis of 37 road infrastructure projects that 62.2% of the projects presented cost deviations; in South Korea 137 road projects were studied where 95% presented cost overruns; in the United States a sample of 2668 road projects was analyzed, of which cost overruns were identified in 50% of the sample [4]. On the other hand, in the United States, found cost overruns in the order of 18% in the construction of roads according to Flyvbjerg in 2003 [5]. In accordance with the above is evident that this is a problem that is latent in all regions, now the variability in cost overruns is generally due to factors that affect the proper development of projects and prevent good management of their resources [6].

On the other hand, it has been evidenced that the result of the construction exercise represents an important fraction of the gross domestic product, both for developing and developed economies; these values can range from 7% to 10% for developed countries and from 3% to 6% for underdeveloped countries [7]. As can be seen, deviations are frequent in road infrastructure projects, which can be reflected in different values depending on the type of project, the country of origin, the amount, among others. Based on multiple studies on this problem, it has been possible to determine the level of deviations in infrastructure projects at a global level infrastructure projects at a global level: for rail system projects the underestimation in the cost is 44.5% on average; in transportation projects the variation between the initial and final cost is 28%, followed by road infrastructure projects that present deviations in the order of 20% [8]. In the case of the transportation projects the variation between the initial and final cost is 28%, followed by road infrastructure projects that present deviations in the order of 20% [8].

Cost overruns can be generated during the different stages of the project, from its structuring to its construction and operation, as well as in projects financed by first power countries with highly qualified engineers as in projects led by engineers with little experience [9]. The variability in the initial budget estimate concerning the actual value executed in infrastructure projects is excessively high, so much so that a study showed that 50% of the transportation projects developed in the United States present a deviation in their initial budget estimate [10, 11]. Based on a literature review of 30 academic reports, it was possible to identify multiple factors that affect the execution of construction projects and result in deviations in time and cost. Based on this review,

it was possible to categorize the top 10 most common factors that occur in projects, among which are the following: change in scope, poor supervision, lack of planning, delay in payments, lack of financial flow, delays in the supply of materials, subcontractor default, changes in design and unqualified personnel [10, 11].

According to the above, it is necessary to carry out a research to establish the values that correspond to the deviations in type and cost in the construction of roads in Colombia, which according to the research conducted by Lugcluster 2014, "The road network in Colombia is composed of the Primary Network (Major Highways, in charge of the nation), Secondary Network (in charge of departments) and Tertiary Network (composed of highways or inter-urban roads, in charge of municipalities). Colombia has a road network of 206,700 km, of which 8.54%, i.e. 17,652 km correspond to the primary network, 22.03% corresponding to 45,536 km to the secondary network and 143,573 km 69.46% to the tertiary network." (Assessment Logistics Capacity, 2014), based on the above, for this research the secondary roads are selected, this due to the great importance for Colombia, since these roads connect municipalities and departments, this generates productivity in the regions thanks to the positive impact that originates the uses of the same.

2. Research Method

The research method has been structured in three main stages which were outlined based on the specific objectives previously defined for the development of the research project. A summary of the research method can be observed in figure 1, where the first stage consisted of gathering the empirical data of secondary roads in Colombia; the second stage consisted of the sample characterization and the exploratory data analysis; the third stage, bivariate and multivariate analysis was performed to determine the significant variables that influence the behavior of the dependent variables under study.

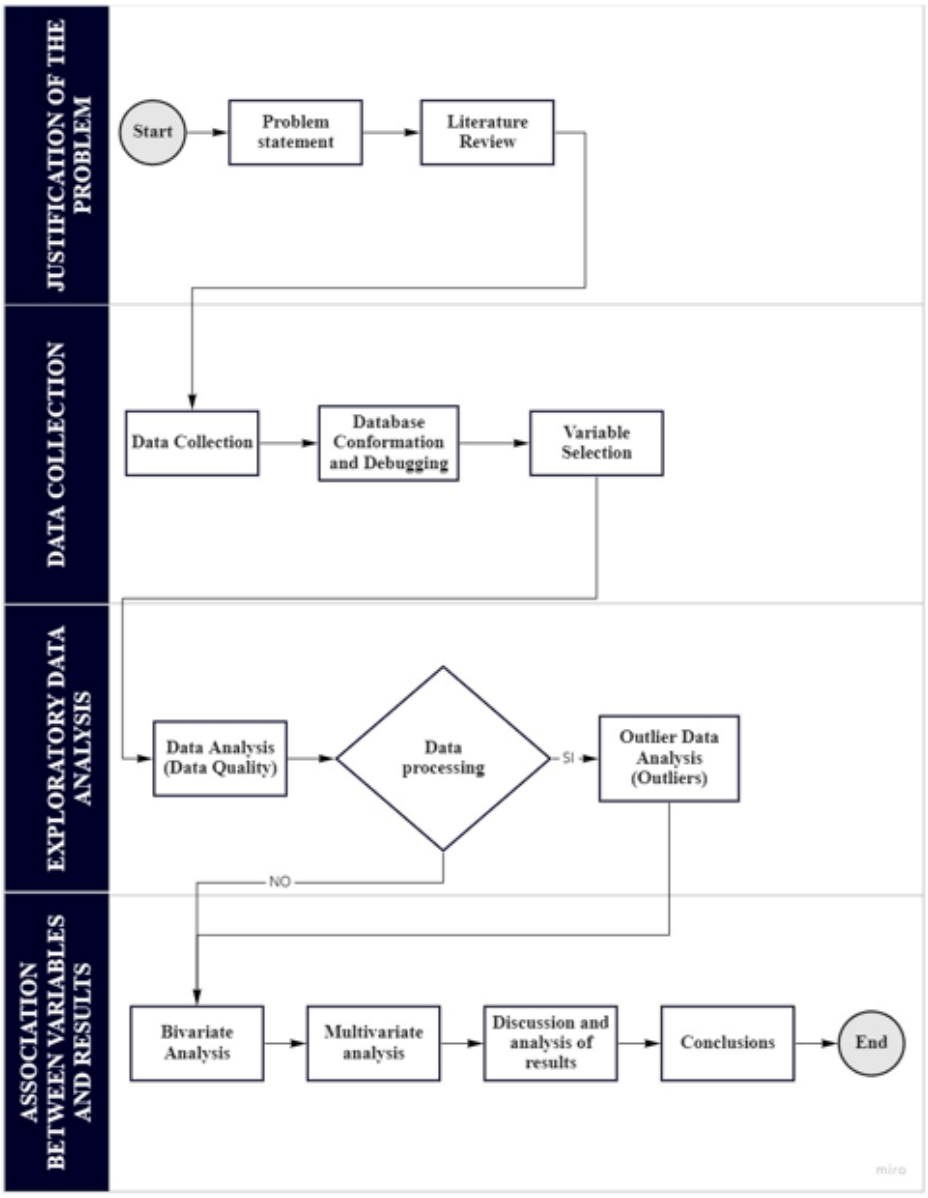

Figure 1. Flow chart for research method.

2.1. Problem statement

The construction sector at national and international level is one of the most important, this is due to the size of this industry in relation to the others, now, the construction industry promotes the economic development of other industries through the demand for goods and services. On the other hand, infrastructure projects worldwide and nationally frequently present cost deviations due to multiple factors such as changes in scope, lack of planning, changes in designs, etc. [12]; globally, it was determined that for rail system projects the underestimation in cost is 44.5% on average; in transportation projects the variation between the initial and final cost is 28%, followed by road

infrastructure projects that present deviations in the order of 20%. [13] Cost overruns can be generated during the different stages of the project, from its structuring to its construction and operation, as well as in projects financed by first power countries with highly qualified engineers as in projects led by engineers with little experience, hence the need to carry out an investigation to identify the deviations in time and cost that originate in the development of road construction projects in Colombia [13].

2.2. Literature Review

It was able to identify multiple factors that affect the execution of construction projects in terms of cost; among these are inflation, lack of cash flow, liquidity, and mainly delays in the execution schedule. The most representative variables associated with time deviations from the initial deadline are changes in scope, design changes, lack of planning, availability of resources, external factors and inadequate supervision [13].

For the review of the existing literature, the specific topic was established, then the keyword is defined, based on this last step the search is structured with the help of a Boolean equation, this search is performed in the major academic databases such as Scopus and Web Science, this search yielded 1. 378 results, which include research articles with topics related to the defined keywords, however, a systematic purification was carried out, which yielded 261 articles that contributed to the object of the research, finally, a manual purification was carried out and 15 articles were identified, which are the bibliographic support of the present research.

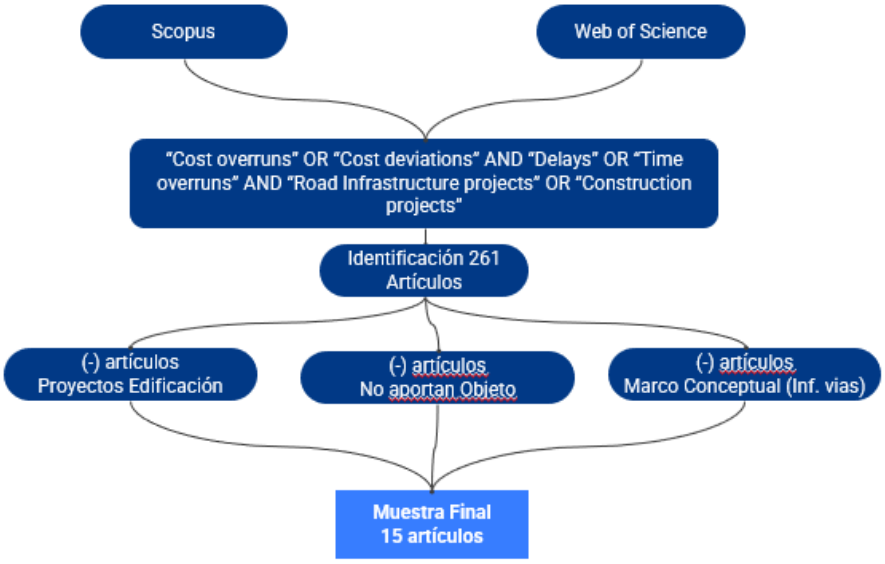

Figure 2. Review Literature

2.3. Data Collection

This stage started with a literature review to identify potential variables that can be selected for this research project, as those were considered in previous studies with similar characteristics. In addition, this literature review identified the research gap. Then, the data gathering was performed using an open data source that stores data from different sectors and industries in Colombia, it should be noted that this information corresponds to public procurement. From this web platform, it is possible to obtain information associated with public award contracts for the construction industry. For this research study, the data collected was extracted from the government contracting platform in Colombia (SECOP I y SECOP II) [1], contemplating executed road construction projects only within the last ten years, as it is subject to the existence of the platform.

2.4. Database Conformation and Debugging

After the export of the database from SECOP I and II, a total of 11 million data was downloaded, which correspond to approximately 65,536 liquidated processes included in the period from 2011 to 2020. From the database, it was carried out a manual debugging process, applying multiple criteria searching using conditionals and rules limiting the research only for processes whose scope includes the maintenance, construction, or rehabilitation of secondary roads. After performing the manual debugging, a sample of 170 processes was obtained. As an additional duty, researchers had to validate the real deadlines and amounts of each process.

2.5. Variable Selection

Once the empirical database was consolidated, the researchers analyzed each of the variables identifying which were numerical or categorical, taking into account that numerical or quantitative variables constitute a data group that is measurable and/or countable, and categorical or qualitative variables are those that cannot be arranged in a specific numerical order [14]. Based on the previous statement, the variables of study were selected and can be identified in Table 1, where they are discriminated by type (categorical or numerical) and nature (independent and dependent). Based on Westland [15] variables were organized according to the project lifecycle phases; for this research variables identified only belonged to two stages. Previous phase, which contains process type, department, region, entity level, project scope, contracting entity, contract award year, estimated duration, advance payment and estimated cost; And execution phase that contains time change order, cost change orders, additional time, additional cost, total cost, and total duration. The dependent variables and which are the main subject of this research, consisting of Time Deviation and Cost Deviation; the deviation for cost and time can be obtained by deduction between the estimated value and the total value, as stated in equation 1 and 2 [16].

$$Cost\ deviation = \frac{Total\ cost - Official\ Budget}{Official\ Budget} \quad \text{[Eq. 1]}$$

$$Time\ deviation = \frac{final\ deadline - Initial\ term}{Initial\ term} \quad \text{[Eq. 2]}$$

Table 1. variables of study.

Phase	Variable	Type	Unit/Values	Description
Previous phase	Process type	Categorical	Public tender, Abbreviate selection, Direct contracting	Modality chosen for the contractor procurement and selection.
	Department	Categorical	Andean region; Pacific; Caribbean; Orinoco; Amazon;	Colombian regions where the project is developed
	Region	Categorical	Geographic Location	Colombian regions where the project is developed
	Entity level	Categorical	Type 1 to 6, 1 being the highest category	Territorial, district, municipality
	Project scope	Categorical	Intervention type	Construction, maintenance, improvement
	Contracting entity	Categorical	Municipalities or departments	Municipality or Other
	Contract award year	Numerical	2010 to 2020	Year of contract subscription
	Estimated duration	Numerical	Time in days	Total duration of the project
	Advance	Numerical	Percentage	Money value corresponding to a percentage of the contract
	Estimated cost	Numerical	Minimum salaries	The contract awarded amount.
Execution phase	Time change order	Numerical	Units	Number of additional days to the initial contract
	Cost change order	Numerical	Units	Value in money additional to the initial
	Additional time	Numerical	Days	Difference between the original deadline and the final contract deadline.
	Additional cost	Numerical	Minimum salaries	The difference between the contract value and the final contract cost.
	Total cost	Numerical	Minimum salaries	Sum of initial value of the contract plus Cost change order
	Total duration	Numerical	Days	Initial term plus additional terms

2.6. Exploratory data analysis

In order to determine the behavior and performance of the variables that were studied, univariate analysis was carried out by determining the descriptive statistics for the numerical variables which allow the identification of patterns and outliers in the data set. For the categorical variables, visual inspection and category distribution (percentage) was performed.

Next, an 'outlier's treatment was elaborated for the all numerical variables, with the purpose to identify atypical data that could affect the behavior of the sample that is being researched and therefore obtained less accurate results. In this sense, the researchers determined to treat only the values from the estimated cost, as it is considered the most important variable, since its nature can determine the size of the projects; higher-cost projects are equivalent to bigger projects in terms

of scope. Extreme outliers were identified and excluded from the dataset, applying the interquartile range (IQR) method that determines the maximum and minimum limit that the data cannot exceed to be considered typical or normal [17].

The procedure for this method consists of organizing data in ascending order; then determine quartile 1, median, quartile 3, and maximum values, followed by IQR through equation 3; finally identify outliers based on the atypical conditions shown in equation 4. Based on this, it was obtained that projects with an estimated value greater than 18,179.78 SMMLV (current legal minimum monthly salary) were eliminated, resulting in a total of 151 projects that were analyzed. The SMMLV is a measure that defines the amount to which every worker is entitled in order to cover his needs and those of his family; this rate is updated annually so that the employer is not affected due to the loss of purchasing power of the currency.

To finalize the exploratory data analysis, an evaluation of the selected variables was performed by determining the descriptive statistics such as measures of central tendency (medium, mean, maximum, minimum, and variance) [18] to identify how is the behavior of the sample and obtain a preliminary data analysis.

$$IQR = Q3 - Q1 \quad \text{[Eq. 3]}$$

$$Q1 - 3IQR < Outlier > Q3 + 3IQR \quad \text{[Eq. 4]}$$

2.7. Bivariate Analysis

To achieve the targets stated for this research study, a bivariate analysis was performed to determine the statistical relationship that represents the independent variables over the behavior of the dependent variables, time and cost deviation; this is called inferential statistics, which its purpose is to determine if there is any relationship between the two variables analyzed [18].

Different statistical methods were implemented depending on the type of variable analyzed. For categorical variable analysis, the Kruskal Wallis test was performed since the dependent variables are numerical and considering that the data set from the different variables did not assemble to a parametric behavior. For this study, the null hypothesis means that the median of the data set for all categorical variables is the same, so if it is rejected, it means that the data set from the variable behaves differently [14]. In those cases where the null hypothesis from the k-w test was rejected, an additional test was performed in order to determine which of the categories from the variable was different from the others; this was achieved through the implementation of the Wilcoxon Mann Whitney test [14, 11]

Each numerical variable was compared with the dependent variables time and cost deviation, through the non-parametric 'Spearman's correlation test because the data set did not assemble a

normal distribution. The 'Spearman's rho range is between -1 and 1; the closer rho is to ±1, the stronger is the relationship between the two variables analyzed. As well the P-value was obtained and supported the previous statement by resulting in statistically significant when is lower than 0.05 [19]. Likewise, a correlation matrix for the numerical variables was elaborated; for this, the Spearman Rho is used to identify the degree of association between the variables; the variables that have a high correlation are those that have greater intensity in its color, and those that the color is very light, are those that have little correlation.

Null hypothesis always denotes that there is no difference between the data set in question, so alternative hypothesis could be understood as statistical relationship between the variables object of study. The P-value represents the probability of occurrence of an event, which is a numerical value between 0 and 1; in this case, the P-Value represents the significance level at which the hypothesis is accepted or rejected. Based on Ali and Bhaskar [14], the null hypothesis is rejected with a P-value lower than 0.05, which means that the data set influences the behavior of the variable under study. In figure 2 it can be observed the significance values considered for this study.

P	Result	Null hypothesis
<0.01	Result is highly significant	Reject (null hypothesis) H0
≥0.01 but <0.05	Result is significant	Reject (null hypothesis) H0
Value ≥ 0.05	Result is not significant	Do Reject (null hypothesis) H0

Figure 2. Flow chart for research method.

2.8. Multivariate Analysis

Multivariate data analysis was carried out, which aimed to simultaneously study the behavior of the data set defined for the present investigation. For this, the Random Forest method is used, which is a learning algorithm, which combines several random decision trees and adds their predictions by averaging [1].

Among the main characteristics of the RF method, it is found that numerical and categorical variables can be included in the same analysis; for the case of the dependent variables that are numerical, deviation in time, and deviation in cost, the literature proposes using regression trees. Random Forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

In the Random Forest algorithm, two control parameters are included, one of them is the number of trees used in the forest and the other is the number of random variables used in each tree; One of the biggest advantages of random forest is its versatility. It can be used for both regression and classification tasks, and 'it's also easy to view the relative importance it assigns to the input features.[2]

3. Results

3.1. Literature Review

As a first step in the research method, a literature review was carried out to understand the context of the subject related to this study. Through this review, it was possible to identify the different causes and reasons that cause time and cost deviation around the world, as well as to determine an order of magnitude for cost and time overruns rates in road construction projects; causes and factors generally are determined through the elaboration of surveys to the project's stakeholders. Each of the analyzed articles focused its research on developing a systematic review which consists of selecting a considerable amount of papers to analyze the frequency and importance with which cost and time overruns are reported in road infrastructure projects.

After performing the revision of eight articles strictly related to the subject from this research, the authors did something similar to a systematic review by classifying the top ten factors that generated the project's overruns from each paper. Once the most frequent factors were classified, through the implementation of a dynamic table in Microsoft Excel it was possible to determine the top-5 causes associated to cost and time deviation of road construction projects. Results indicated that the most common factors for time overruns are delays in progress payment, poor planning, lack of supervision, subcontractor low performance, and lack of resources (machinery). Regarding cost deviation, the top-5 factor resulted in scope change, price material variation, poor planning, design changes, and delays in progress payment [2] [4] [20].

3.2. Univariate Analysis Results

Based on the methodology previously defined, this section contains the results obtained from the sample characterization and the analysis made to the dependent and independent numerical and categorical variables, which let the researchers to understand the behavior of the database for the different variables. In table 2, it is possible to observe the statistical summary of the dependent variables (time deviation and cost deviation) for the 151 projects analyzed. For time deviation, greater values were presented for the different processes by obtaining a mean value of about 45% and projects that last 3 times more than the estimated initial duration; regarding cost deviation, lower values were identified by obtaining a mean value of 8.5% and one project that resulted in a total cost that was almost double the estimated initial cost, even though the Colombian law restricts the amount of additional value for public bidding as maximum as 50% from the contract initial cost.

Table 2. Statistical summary for dependent variables.

Tipo	Max	Min	Mean	Standard Deviation	Coefficient of Variation
Time Deviation (%)	316.7%	0.0%	45.1%	61.8%	137.2%
Cost Deviation (%)	53.17%	0.0%	8.5%	17.1%	201.0%

It can be identified that there are projects that did not have any deviation in time or cost. Also, the coefficient of variation is considerably high for both variables, so it results interesting to visualize how is the 'dataset's behavior for time and cost deviation among the different projects analyzed. As illustrated in figure 3, the distribution value for cost deviation is more uniform precisely because a considerable amount of projects did not present any deviation on its cost, which can be interpreted with Q3 by identifying that the 75% of the projects did not or presented deviation lower than 8.54%, in contrast with time deviation where Q3 reached a value of 66.7%. The boxplot allows the identification of those values that could be considered as outliers but taking into account that the main objective of this study was the analysis of time and cost deviation presented on road infrastructure projects, none of the outliers from the dependent variables were eliminated.

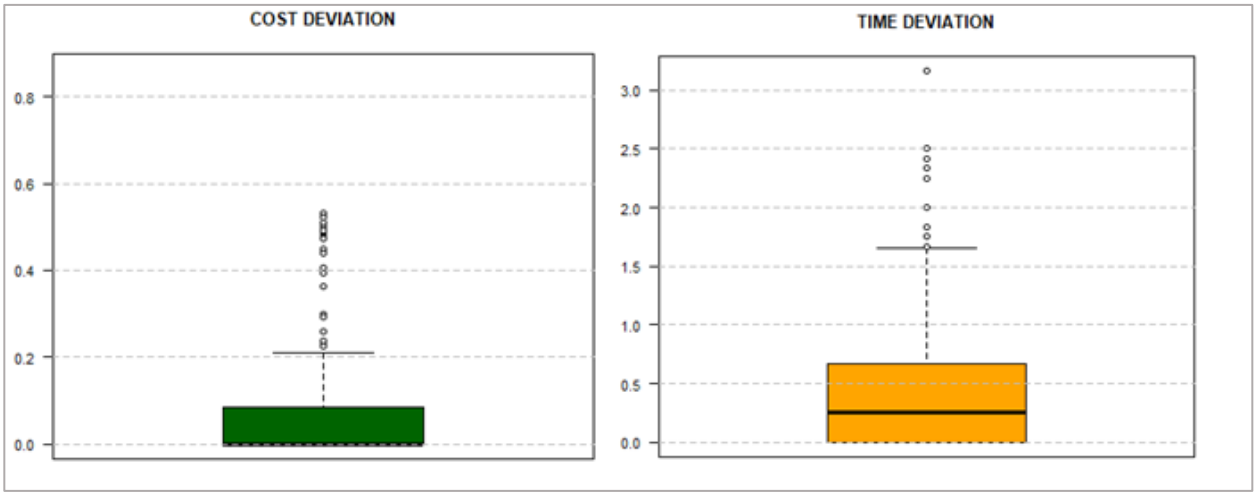

Figure 3. Cost and Time deviation boxplot.

Additionally, it was identified that 61.4% of the projects (94) presented time deviation through its execution, 30.7% (47) presented cost deviation and 28.1% (43) did present delays and cost overruns; it is evident that in the case of this study the deviations in time were higher than the deviations in cost, which according to the literature, at a global level the deviations in time are also higher. Regarding the categorical variables, figure 4 contains the univariate analysis made for each one of them. It was possible to identify that more than 60% of the projects studied were executed in the Andina region from Colombia, which most of them belong to Cundinamarca department. Also, that 76% of the projects contemplated within its scope improvement of existing road infrastructure, 18% construction on new road infrastructure and 6% road rehabilitation. Projects considered in this study are from 2011 to 2019 and were adjudicated through three types of bidding: public bidding (76%), direct contract (20%) and abbreviated selection (4%). The contracting entity was also considered in this study obtaining that 74% of the projects were owned by municipalities and govern ships, which makes sense since secondary roads are those that communicate two municipal cities.

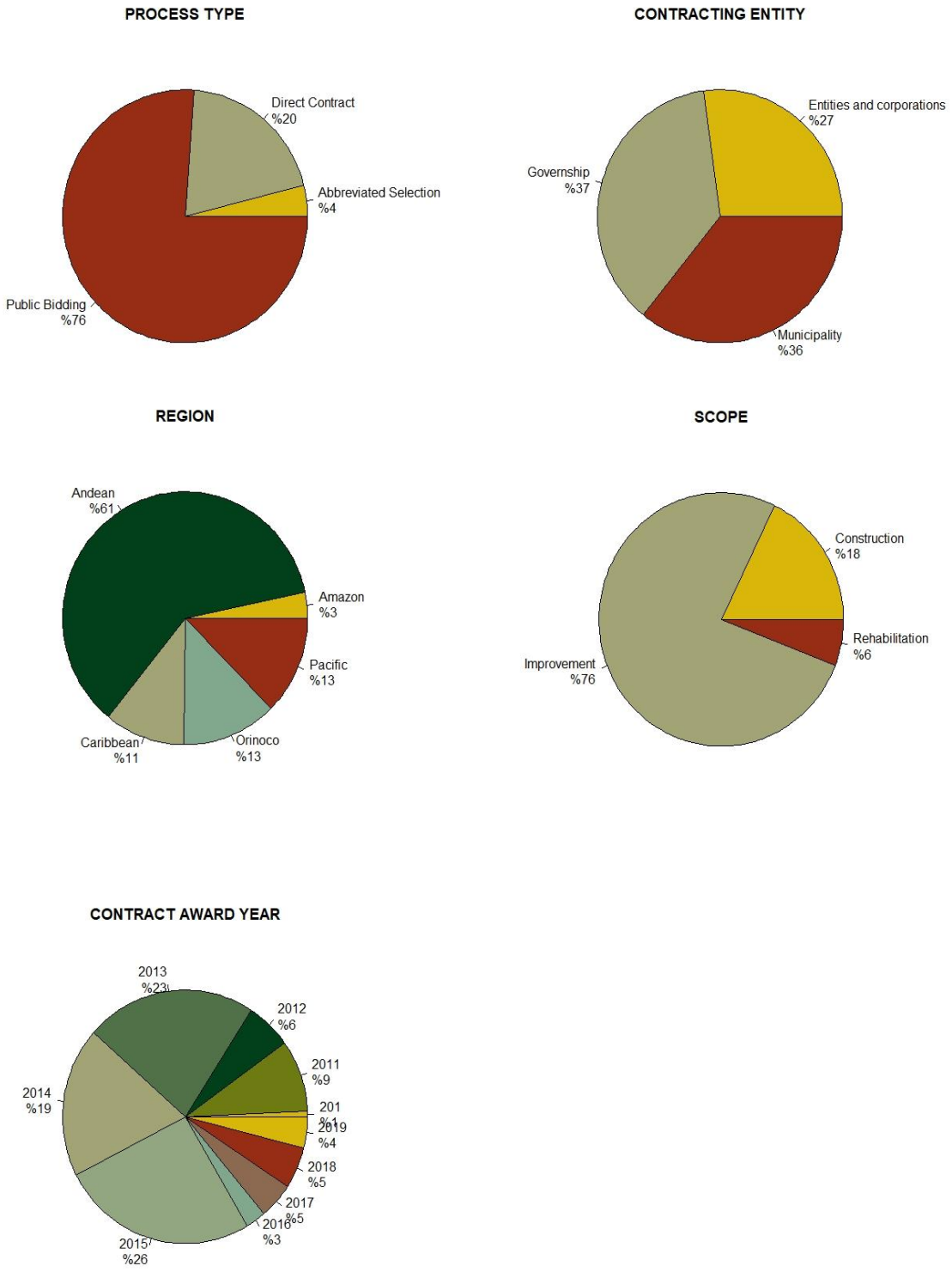

Figure 2. Univariate analysis for categorical variables.

Table 3: Univariate analysis

Variable	Max	Min	Mean	Standard Deviation	Coefficient of Variation
Cost change order	3.00	1.00	0.43	0.75	1,79
Time change order	7.00	1.00	1.14	1.28	1.13
Total duration	780.00	60.00	249.52	139.97	0,56
Additional duration	480.00	60.00	180.58	94.30	0.52
Initial cost	26.607,87	676.55	4.529,0,77	4.948,62	1.09
Total cost	32.886,59	652.45	4.794,32	5.333,18	1.112
Additional cost	8.083,93	1.00	470.58	1.222.01	2.597
Initial duration	480.00	60.00	180.57	94,186	0.52
Advance Payment	0.50	1.00	0.24	1.65	0.71

3.3. Bivariate Analysis Results

This section contains the results obtained from the statistical bivariate analysis elaborated to determine the statistical significance between independent and dependent (cost and time deviation) variables. It includes numerical variable analysis measured through Spearman and 'Pearson's Rho correlation tests and categorical variable analysis developed with non-parametric Kruskal Wallis method.

3.3.1. Significant Numerical Variables

A correlation matrix was elaborated to analyze the statistical relationship between each of the numerical variables. This was carried out by calculating 'Spearman's Rho value for all the numerical variables in order to determine a preliminary association between them. First, an illustrative matrix was elaborated to determine the level of correlation resulted from 'Spearman's Rho test, obtaining a scattering matrix that allow to identify preliminary association; In the way that the scattering points represent a perfect linear form, the relationship between the variables of study is much stronger. However, this does not mean that those variables that do not have a linear correlation between them could be dependent each other. Figure 5 illustrates the results obtained for such numerical analysis, where blue color represents a closer approximation to 'rho's equal to 1, which means that the variables contemplate a positive linear association and red color indicate a negative relationship. This preliminary analysis let the researchers to identify significant association among the numerical variables, with the purpose to avoid future bivariate analysis between dependent and independent variables that could be strongly correlated, by showing Rho values higher than 0.85.

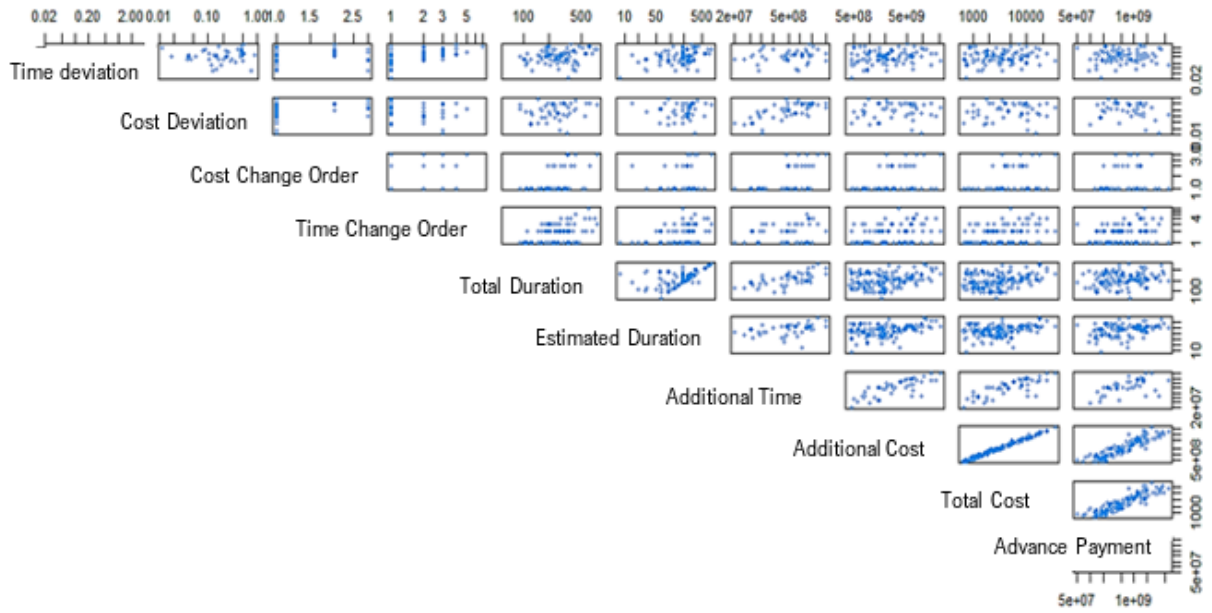

Figure 5. Scattering Relationship Matrix.

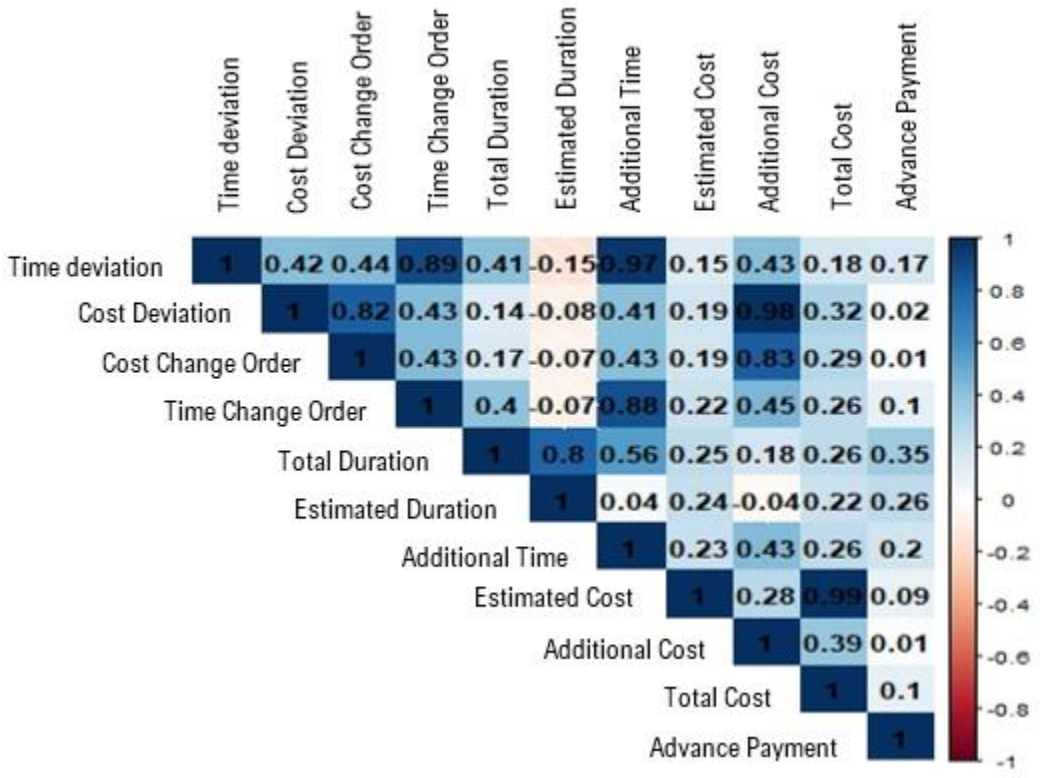

Figure 6. Spearman's Rho Correlation Matrix.

Based on Spearman 'Rho's correlation matrix, researchers determined highly correlated variables such as Time Deviation with number of extension and additional time, cost deviation with number of Cost change order and additional cost. Also, it was possible to identify negative correlated variables such as Time Deviation with initial duration, indicating an inversely proportional relationship. Other strong associations were possible to identify through the results, which allow to discard some of the variables for future analysis like for example Time change order with additional time and Cost change order with additional cost.

Afterwards, each numerical independent variable was analyzed against the time deviation and cost deviation variables with the purpose to determine the level of significance for each of them, assuming a significant correlation for those relationships that contemplate P-Value lower than 0,05. Figure 7 shows the scatterplot graph obtained for the bivariate analysis between initial duration and time deviation, where it is possible to observe that projects with shorter durations (100-200 days) present higher rate time deviations than those with longer durations; as well it was possible identify that additional cost present a dependency with respect to time deviation.

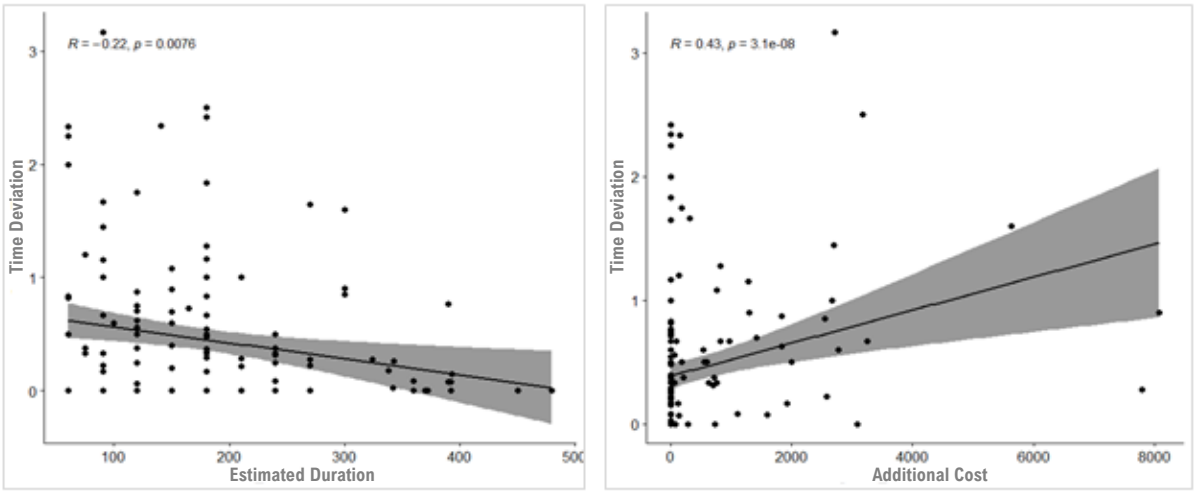

Figure 7. Initial duration and additional cost Vs Time Deviation.

On the other hand, it was observed that regarding cost deviation the variables that resulted to be significant (variables that influence the behavior of the dependent variable) were expected due to its relationship. Figure 8 shows the correlation between cost deviation and time deviation, and it can be concluded that although there is a light directly proportional relationship, many of the projects that present time overruns do not necessarily present cost overruns. Likewise, the scatterplot graph shows that projects with lower budgets present greater cost deviation.

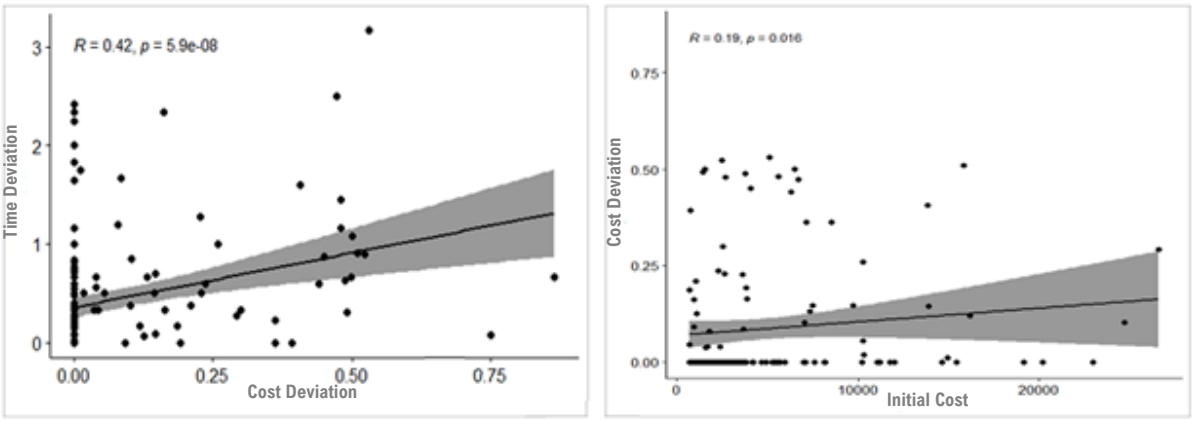

Figure 8. Time Deviation and Initial Cost Vs Cost Deviation.

Table 4 indicates the significant numerical variables with time deviation resulted from the bivariate analysis. Additional cost and cost deviation resulted as the most significant correlation. Initial duration has a negative relationship with the dependent variable. Although advance payment does not have a strong relationship with time deviation, it turned out to be significant; as the value of the advance payment increases, the deviation in time is greater. This can be attributed to the fact that contractors used to be very irresponsible with the advance payment management, and allocated resources from one project to another. This led to the obligation to manage public resources through bank trusts.

Table *4.* *Significant numerical variables vs Time Deviation.*

Variable	Spearman/Pearson 'Rho's	P-Value
Additional Cost	0.43	3.1×10^{-8}
Cost Deviation	0.42	5.9×10^{-8}
Initial Duration	-0.22	0.0076
Advance payment	0.21	0.0084

Regarding bivariate analysis against cost deviation, significant correlations could be observed in table 5. It is possible to identify that most of the significant relationships were expected due to its incidence on the dependent variable. Additional time and initial cost resulted in a significance correlation. The stronger relationship resulted with time deviation, which is the other dependent variable.

Table 5. Significant numerical variables vs Cost Deviation.

Variable	Spearman/Pearson 'Rho's	P-Value
Time Deviation	0.42	5.9×10^{-8}
Additional time	0.41	1.2×10^{-7}
Total Cost	0.32	7.2×10^{-5}
Initial Cost	0.19	0.016

3.3.2. Significant Categorical Variables

Regarding the categorical variables, the bivariate analysis was developed through the Kruskal Wallis test. Regarding time deviation correlations, none of the categorical relationships resulted significant, obtaining p-values greater than 0,05 and therefore no rejecting the null hypothesis. In regards with cost deviation, table 6 shows the significant categorical variables where the project scopes, contracting entity and process type resulted as the most influencing variables over the cost deviation.

Table 6. Significant Categorical Variables vs Cost Deviation.

Variable	P-Value
Project Scope	0.00025
Contracting Entity	0.0092
Process Type	0.0013
Department	0.0123
Entity level	0.0175

To identify which of the categories from the significant variables present the different behavior, the Wilcoxon Mann-Whitney test was performed for project scope, contracting entity, and process type. Results indicated that improvement is the category from scope and institutions & corporations is the category from contracting entity, that presents the most different behavior from the other. Regarding process type, the category that behaves worse corresponds to direct contracting (20% of projects). As illustrated in figure 9, it can be observed that projects adjudicated through this bidding method present lower cost deviations problems in contrast with public bidding (76% of projects) and abbreviate selection (4% of projects). Also, it could be analyzed that although the average cost overruns for improvement scope projects is lower than rehabilitation and construction is the category that presents the highest percentage rates for cost overruns.

Table 7. Wilcoxon Mann-Whitney test OBJECT

	Construction	Improvement
Improvement	0.0021	-
Rehabilitation	1.0000	0.29360

Table 8. Wilcoxon Mann-Whitney test CONTRACTING ENTITY

	Mayoralty	Government
Government	1.0000	-
Institutes and corporations	0.0252	0.0086

Table 9. Wilcoxon Mann-Whitney test PROCESS TYPE

	Direct Contract	Public Bidding
Public Bidding	0.0011	-
Abbreviated Selection	0.0025	1.0000

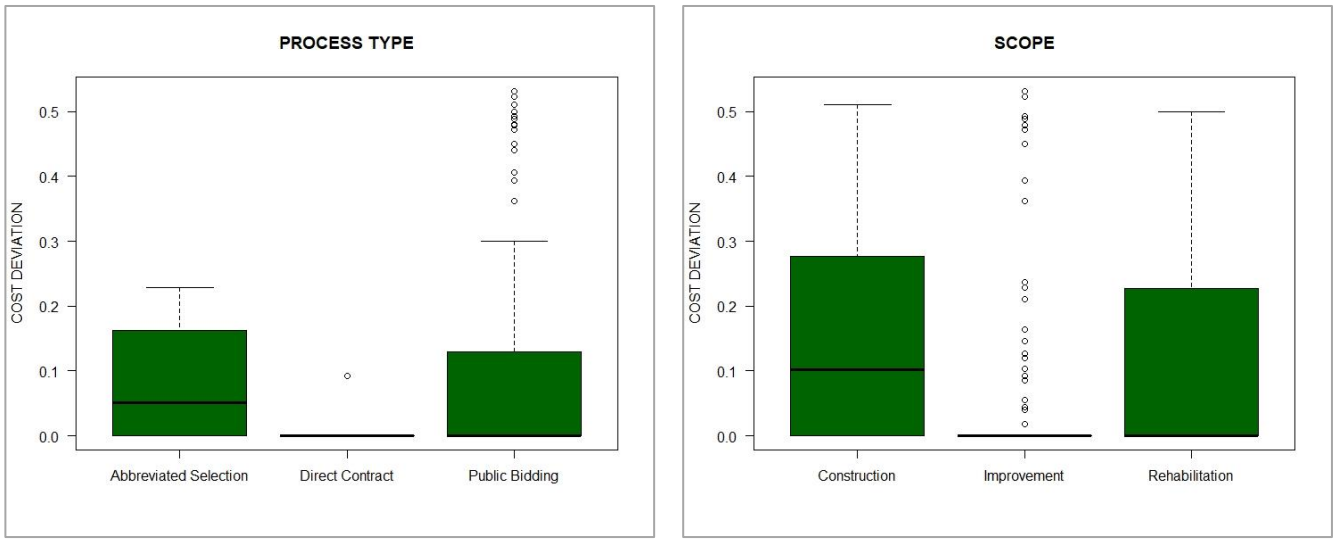

Figure 9. Time Deviation and Initial Cost Vs Cost Deviation.

3.4. Multivariate Analysis Results

The Multivariate analysis brings together statistical methods that focus on simultaneously observing and processing different statistical variables to obtain relevant information. This section contains the results obtained from the multivariate analysis elaborated to determine statistical significance between dependent and independent variables, by analyzing multiple variables at the same time [21]. Decision trees are a prediction model used on a set of data, for which diagrams are made, they are very similar to rule-based prediction systems, which serve to represent and categorize a series of conditions that occur successively, for the resolution of a specific problem. Among the main characteristics we have that it runs efficiently on large databases and gives estimates of which variables are important in the classification; on the other hand, among its main disadvantages we have that it has been observed that random forests are too tight to some data sets [22].

3.4.1. Time deviation

For this research, a series of numerical and categorical variables were analyzed, to which various statistical methods were applied to validate the level of association that each of these variables had, which turns out to be the level of significance of these variables in contrast with others. For this analysis, a random forest test was implemented taking into account the statement that a significant number of trees is necessary to get stable estimates of variable importance and proximity; based on the previous argument, the analysis started with a total of 400 decision trees, which according to the literature is an optimum number for running the random forest analysis [23]. After a comparison was made between the error and the trees, an optimal number of trees was determined, resulting in 127 trees. Numerical and categorical variables were included in this analysis with the purpose to correlate all variables simultaneously. However not all independent variables were tested, only 12 were finally included in the analysis, excluding those high correlated variables such as Cost change order, total cost, Time change order, additional time, and total duration. Also, based on the results obtained from the Wilcoxon Mann-Whitney test, categorical variables (scope, contracting entity, and process type) were re-valuated leaving only 2 categories for each of the mentioned variables, the one that behaves different and the other two as a new group. New categories correspond to Improvement and Non-improvement (object); Institutions & Corporations and Non-Institutions & Corporations (contracting entity); Direct contracting and non-direct contracting (process type). Once the model was executed, the results obtained determined 3 out of 12 predictors based on the reduction of the Out-of-bag error as illustrated in figure 10. The additional cost was determined as the most significant correlation, followed by initial duration and Contract award year.

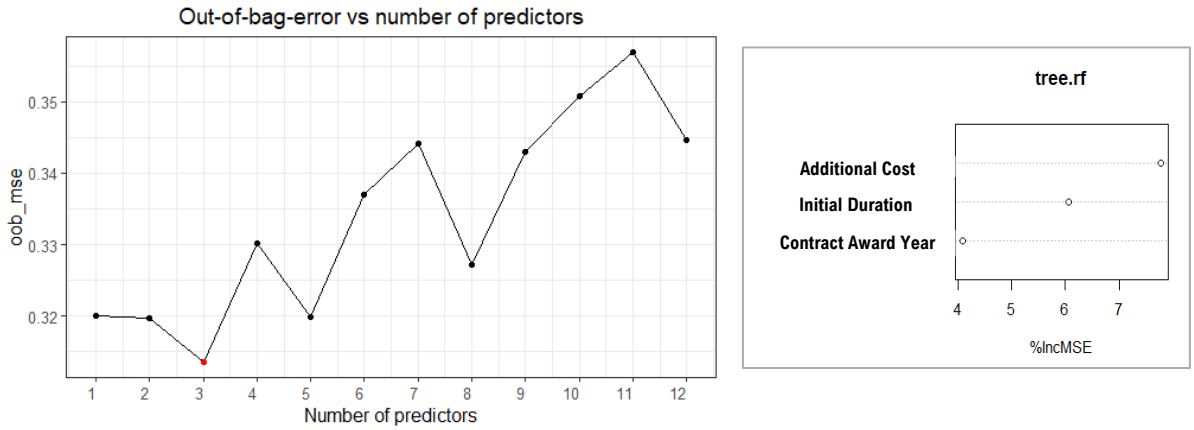

Figure 10. Significant variable for time deviation – Random Forest.

Additional cost and initial duration did result a significant in bivariate analysis too, which means these two are the most influencing variables over the dependent variables, time deviation. Considering that additional cost was n expected variable to be related with deviation, in figure 11, the scatterplot between initial duration and the dependent variables is shown with the purpose to illustrate what is the behavior between them. It can be interpreted that the highest time deviations values are presented in those projects that contemplate a duration lower than 200 days.

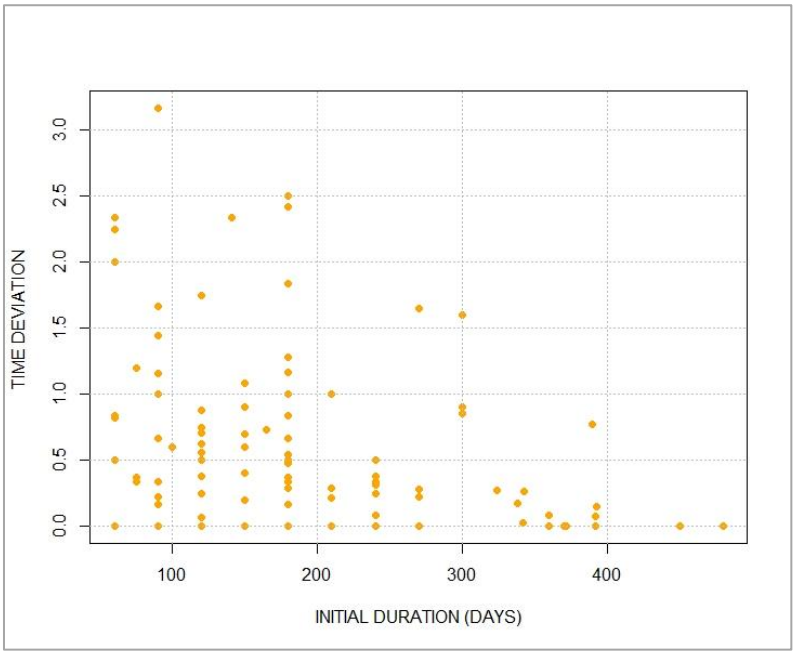

Figure 11. Scatterplot for Initial Duration vs Time Deviation.

3.4.2. Cost deviation

As per the time deviation multivariate analysis, this correlation was performed using the same parameters and characteristics considered for such time analysis. Therefore, a total of 400 decision trees were considered initially, obtaining an optimum number of 119 trees. Based on this, the multivariate analysis was carried correlating all variables at the same time except for Cost change order, total cost, Time change order, additional cost and total duration. As for the time deviation multivariate analysis, the categories from the significant categorical variables (scope, contracting entity and process type) were restructured based on the Wilcoxon Mann Whitney test results. Once the model was executed, the results obtained indicated 3 predictors out of 11 based on the reduction of the Out-of-bag error as illustrated in figure 12. Additional duration was determined as the most significant correlation, followed by process type, and contracting entity.

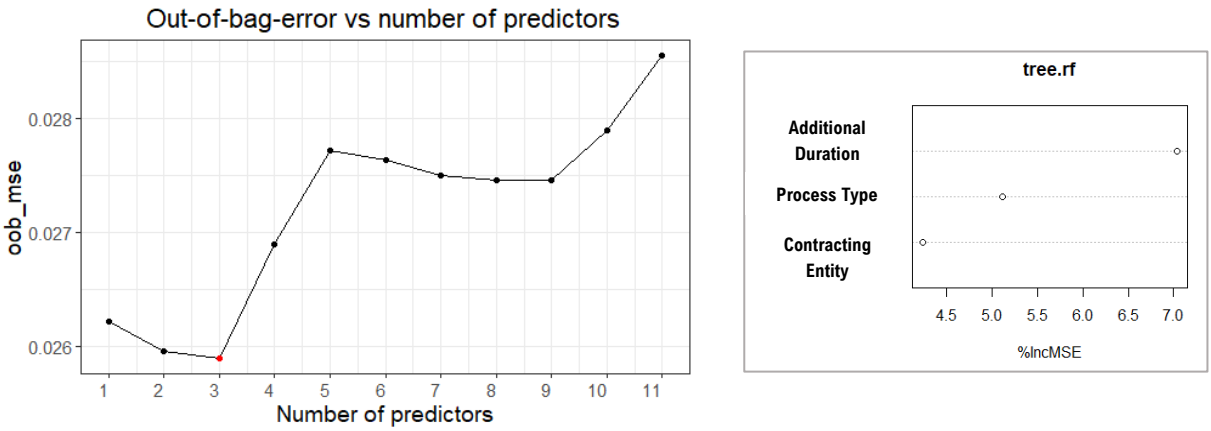

Figure 12. Significant variable for time deviation – Random Forest.

4. Discussion

From the study carried out it was possible to analyze the deviation problems in terms of cost and time that occur in secondary road infrastructure projects in Colombia. The database used for this study contemplated 151 projects, all of them were financed with public resources and were already concluded. Considering that all projects belong to the public sector, all the available information used is of public access. It was possible to identify an average cost deviation of 8.5% considering all the projects studied, which in any case, it should be noted that in some countries the results obtained are greater in comparison with the present research; in comparison with the existing global literature, cost overruns could vary significantly depending on the project's scope. For example, transportation projects present an average cost overrun of 44.7% was identified, 33.8% in tunnel and bridge projects, and 20.4% for road projects, for 258 worldwide (America, Europe, and other) road projects analyzed [5]. Other research studies showed average cost deviations of 14.6% for 169 road projects analyzed in Palestine and 16.3% for 231 highway projects in Australia [24] [25]. The maximum cost overrun obtained for this study is 53.17%; apart from this value all the projects that present cost deviations do not exceed 50-55% which is the maximum additional

amount that could be added to public financed projects, in conformity with Colombia law 80 from 1993. Regarding time deviation it was possible to determine average overruns around 45%, with a maximum value of 316.7%; considering that a big portion of the sample presented time deviation, this result is comparable with previous local findings that identified time overruns values of 52.95% with a maximum of 450% in rural road infrastructure projects [11].

According to the study made by Kamrul Ahsana and Indra Gunawan [26], in Bangladesh there are deviations in time of 34.41%, in the case of China the deviation in time is 13.63%. According to the above, it is observed that in Europe and Asia there is an average deviation of 33.37 [19], which in comparison with the present investigation the values are very close, for the case of the present study there was a deviation in time in the order of 45%.

On the other hand, it was possible to identify that in terms of frequency, projects that present cost deviation correspond to 47 out of 1513 which is equivalent to 310% of the total sample. Most of the projects that did present cost overruns were associated as low magnitude road infrastructure projects, since its initial budget did not exceed 5000 SMMLV; the last statement confirms previous findings by other authors that affirm cost overruns are more likely to be predominant among smaller projects than larger ones [3]. Also, the average amount of projects that present cost overruns approach to existing studies, where it was found that 55% of road projects in United States present cost deviations [27]. In regards with time overruns, 61.4% of the projects presented deviation and it is related with the relationship between budget and cost overruns, where initial duration is inversely proportional to time deviation, which means greater duration projects present lower deviations than smaller ones.

Based on the previous analysis it can be affirmed that frequently projects present greater time deviation values than cost. this is directly related to the statement that not necessarily projects that present time deviation must present cost overruns [24]. According to the literature and research conducted with a methodology like that of this research, the advance payment was not evaluated in the projects, largely due to multiple reasons, such as the fact that some countries do not give advance payments to their contractors. Now, according to the results obtained, it can be seen that most of the processes received advances with percentages that varied from 20% to 50%; however, it could be concluded that the processes in which advances of 50% were generated had the highest cost overruns. In accordance with the above, it is recommended that regularization of the amounts of advances be established since this would generate greater control and impartiality. Additionally, this study identified those categorical and numerical variables that present statiscal significance in respect with time and cost deviation. It could be identified that none of the categorical variables presented any relationship with time overruns. Beside this, it was possible to determine a relationship between projects contemplating road improvement and cost overruns, presenting greater cost overruns than the others. As well projects adjudicated through public bidding and abbreviated selection resulted in higher cost deviations values than those contracted directly; from

which it can be inferred that when entities have the possibility to directly select reliable and trustworthy contractors, projects do not present cost deviation after its execution.

5. Conclusions

The current research consisted of the revision and analysis of empirical data obtained from open-public information of secondary order road infrastructure projects in Colombia for the past 10 years. It allows the identification of those characteristic projects that present time and cost deviations during its execution, identifying a larger number of projects presenting time overruns than cost overruns. The study also provides information regarding contracting entities, regions, and scope of work associated with each of the analyzed projects, determining some parameters that could help readers to take decisions before starting projects with such characteristics.

It can be noticed that 76% of the processes were awarded through public bidding, which can be attributed to the maximum contracting amounts that vary according to the annual budget of each entity; public bidding has no limit on the contracting amount. Additionally, the study showed that 61% of the analyzed projects were executed in the Andean region, this may be due to the high road infrastructure requirements because it is the most productive region in the country.

From the statistical analysis, it can be concluded that the significant variables for time deviation correspond to additional cost, initial duration, and advance payment; and the significant variables for cost deviation are additional time, total cost, initial cost, 'project's scope, contracting entity and process type. Readers should consider that projects with high advance payment amounts present greater time deviations than those contemplating lower amounts; this confirms the need for the entities to regulate the amounts of advance payments, due to the mismanagement by contractors. As well that projects with an estimated budget lower than 5000 SMMLV and durations ranging between 100 and 200 days, usually present overruns in cost and time.

According to the literature review, it was found that in South America there is very little research in this field, which means that it is not possible to make a comparison with similar characteristics countries, such as socio-cultural features and geographic location, which allow identifying a very important gap for future research.

There is a positive correlation between the time deviation of the projects and the amount of the advance payment. The higher the value of the advance payment, the greater the deviation in time. This confirms the need for the entities to regulate the advance payment amounts, due to the mismanagement of the advance payment by the contractors.

Contracts awarded by public bidding present greater deviations in terms of cost. On the other hand, contracts awarded by direct contracting tend not to have cost deviations, which could be attributed to the trust placed by the entities in local and reputable contractors.

It can be seen that there is a lack of research on local matters, which generates few references, so it is recommended to carry out research in different sample areas, including multiple variables

associated with the economic and social conditions of a given region of the country can be analyzed, and the types of contracting entities can also be researched in greater detail, and the processes can be determined with greater reliability.

Colombia, unlike many Latin American countries, the possibility for citizens to know the contracts executed by any public entity, knowing the scope of the projects allows greater transparency and the possibility of being guarantor of the development and fulfillment of the object of the contract, in any case, it is recommended to take advantage of the information found in the Colombia Compra Eficiente platform, for future research, this platform offers the necessary input to develop academic research that can generate benefits to the community in general.

The present study let the readers understand some of the aspects and features from th secondary road construction projects that could present cost and/or time deviation during their execution. This allows taking actions and measures to avoid or at least mitigate such critical problems of cost and time overruns in road construction projects. For example, in further studies, it may be important to deepen in the subject of those contracting entities where the projects present greater deviations and determine the causes that area generating those overruns.

6. References

[1] G. L. E. N. &. H. R. L. A. Mejía, «Caracterización de los sobrecostos en proyectos de construcción de acuerdo con la localizacion geografica: Una revision sistematica entre 1985 y 2016,» *Actas Ingenieria.,* vol. 4, 2017.

[2] J. Shane, S. Anderson y K. Molenaar, «Construction Project Cost Escalation Factors,» *Journal of Management in Engineering,* 2009.

[3] J. Odeck, «Cost overruns in road construction—what are their sizes and determinants?,» *Transport Policy,* vol. 11, nº 43 - 53, 2004.

[4] R. F. O. S. &. C. K. Herrera, «Cost Overrun Causative Factors in Road Infrastructure Projects : A Frequency applied sciences Cost Overrun Causative Factors in Road Infrastructure Projects : A Frequency and Importance Analysis. August.,» 2020.

[5] B. S. M. K. &. B. S. L. Flyvbjerg, «How common and how large are cost overruns in transport infrastructure projects? Transport,» *Transport Reviews,* vol. 23, pp. 71 - 88, 2003.

[6] C. M. M. &. M. K. Kaliba, «Cost escalation and schedule delays in road construction projects in Zambia.,» *International Journal of Project Management,,* vol. 27, pp. 522 - 531, 2009.

[7] R. &. P. M. Stasiak-Betlejewska, «Construction Costs Analysis and its Importance to the Economy,» *Procedia Economics and Finance,*, vol. 34, pp. 35 - 42, 2015.

[8] B. S. M. &. B. S. L. FLYVBJERG, «What Causes Cost Overrun in Transport Infrastructure Projects? Transport Reviews,» vol. 24, pp. 3 - 18, 2002.

[9] A. M. S. &. A. S. Enshassi, «affecting the performance of Construction projects in the Gaza,» *Civil Engineering and Management,* vol. 15, pp. 269 - 280, 2009.

[10] I. P. G. A. G.-C. A. T. Sara Lozano Serna, «Identificación de factores que generan diferencias de tiempo y costos en proyectos de construcción en Colombia,» *Ingenieria y Ciencia* , vol. 14, pp. 117 - 151, 2018.

[11] A. Goméz, A. Sanz, L. Montalban, J. L. Ponz y E. Pellicer, «Identification of Factors Affecting the Performance of Rural Road Projects in Colombia,» MDPI Sustainability, 2020.

[12] H. d. Solminihac, «PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE,» 10 2018. [En línea]. Available: https://www.claseejecutiva.com.co/blog/articulos/sector-de-la-construccion-es-mucho-mas-que-las-empresas-constructoras/. [Último acceso: 10 11 2021].

[13] S. P. G. I. G.-C. A. &. T. A. Lozano Serna, «Identificación de factores que generan diferencias de tiempo y costos en proyectos de construcción en Colombia,» *Ingeniería y Ciencia,* vol. 14, nº 27, pp. 117-151, 2018.

[14] Z. (. Z. B. S. (. S. B. Ali, «Basic statistical tools in research and data analysis,» *INDIAN JOURNAL OF ANAESTHESIA* , vol. 4, 2012.

[15] W. J, «The Project Management Life Cycle: A Complete Step-by-Step Methodology for Initiating, Planning, Executing & Closing a Project Successfully,» London, UK, 2007.

[16] D. Gransberg y M. Villareal, «Construction Project Performance Metrics,» de *AACE International Transactions*, 2002, p. CSC.02.

[17] A. E. y. J. D. G. C. Batanero, «Analisis Exploratorio de Datos: sus posibilidades en la enseñanza secundaria,» 1991, pp. 25-31.

[18] M. S. J. P. Pharmacother., Medidas de tendencia central: La mediana y la moda., 2011.

[19] A. (. A. 1. A. (. A. Heinen, «Spearman rank correlation of the bivariate Student t and scale mixtures of normal distributions,» *Journal of multivariate analysis,* 2020.

[20] A. Gomez, E. Pellicer, J. L. Ponz y A. Sanz, «Factor Generating Schedule Delays and Cost Overruns in Construction Projects,» 2019.

[21] P. S. P. a. J. A. B. Thais Mayumi Oshiro, «How Many Trees in a Random Forest?,» *Springer-Verlag Berlin Heidelberg*, 2012.

[22] A. Chakure, «START IT UP,» 29 Junio 2019. [En línea]. Available: https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f. [Último acceso: 14 12 2021].

[23] A. Liaw y M. Wiener, «Classification and Regression by randomForest,» vol. 2, pp. 18-22, 2001.

[24] I. Mahamid, A. Bruland y a. N. Dmaidi., «Causes of Delay in Road Construction Projects,» *Management in Engineering (ASCE)*, vol. 28, 2012.

[25] G. D. M. &. W. J. K. Creedy, «Evaluation of risk factors leading to cost overrun in delivery of highway construction projects,» *Construction Engineering and Management*, vol. 5, pp. 528-536, 2010.

[26] K. A. a. I. Gunawan, «Analysis of cost and schedule performance of international development projects,» *International Journal of Project Management*, vol. 28, nº 1, pp. 68 - 78, 2010.

[27] C. B. G. M. S. L. a. K. C. S. ordat, «An Analysis of Cost Overruns and Time Delays of INDOT Projects,» *Transportation Research Program, Indiana Department of Transportation and Purdue University*, 2004.

[28] G. Mejía, N. Escandón y L. A. Reyes, «Characterizing construction project overruns by region: A systematic review from 1985 to 2016,» *Actas de Ingeniería*, vol. 3, pp. 24-35, 2017.

[29] G. B. J.-P. V. Erwan Scornet, «Consistency of random forests,» *PROJECT EUCLID*, 2015.

[30] N. S. O. G. M. Jherson Jhadir Bohórquez, «Los Sobrecostos en Proyectos de Infraestructura Vial: Una Revisión Actual / Cost Overruns in Transport Infrastructure Projects: A Current Review,» *Desarrollo e Innovación en Ingeniería*, vol. 3, 2018.

[31] A. S. K. I. a. S. R. H. Doloi, «Analysing factors affecting delays in indian construction projects,» *International Journal of Project Management*, vol. 3, pp. 479 - 489, 2012.

[32] O. A. Olatunji, «A comparative analysis of tender sums and final costs of public construction and supply projects in nigeria,» *Journal of Financial Management of Property and Construction,* vol. 13, pp. 60 - 79, 2008.

[33] E. f. D. Rwakarehe, «Effect of Inadequate Design on Cost and Time Overrun of Road Construction Projects in Tanzania,» *Construction Engineering and Project Management,* vol. 4, 2014.

[34] K. A. a. I. Gunawan, «Analysis of cost and schedule performance of international development projects,» *International Journal of Project Management,* vol. 28, nº 1, pp. 68 - 78, 2010.

[35] N. Sowmya, K. Amol Madhav y P. Sivakumar, «Study on Time and Cost Overruns in Mega Infrastructure Projects in India,» *Journal of The Institution of Engineers (India): Series A,* vol. 100, pp. 139-145, 2019.