*13th International Defense and Homeland Security Simulation Workshop*
20th International Multidisciplinary Modeling & Simulation Multiconference

2724-0029 © 2023 The Authors.
doi: 10.46354/i3m.2023.dhss.007

# Demand analysis for failure prediction for decision making in the STC Metro

Adrián Vilchis Serrano[1], Alma Elia Vera Morales[1], and Aida Huerta Barrientos[2*]

[1] Faculty of Engineering, UNAM, Ciudad Universitaria, México City, 04510, México.
[2] Faculty of Engineering & Centro de Ciencias de la Complejidad, UNAM, Ciudad Universitaria, México City, 04510, México.

*Corresponding author. Email address: aida.huerta@comunidad.unam.mx

## Abstract

Failure prevention planning as well as user demand analysis and forecasting are crucial for providing efficient services, particularly for massive transportation systems. The Mexico City subway is one of the massive transportation system operational with 12 lines spanning 226 km and 195 stations. It transports from origin to destination more than 1.057 billion passengers annually, including 4.1 million on weekdays. As user demand continues to rise, from the operational point of view it is essential to examine the relation between demand and failures. This study aims to explore whether higher user demand corresponds to biggest failure rates and to identify subway station-related factors influencing failure occurrences. Employing statistical techniques, specifically multiple linear regression and using R™ software, this study focused on variables such as demand, subway station, subway facilities, time between failures, and subway line. The analysis revealed a strong interaction between user demand and failures, with external agents accounting for 31% of total failures. By focusing on the most significant failures and addressing external factors, the subway of Mexico City can prioritize efforts to reduce service delays. We consider that the findings contribute to understand the phenomenon in order to implement more effective strategies to improve subway service reliability.

**Keywords:** failures, demand, subway, statistics, multiple linear regression, prediction.

## 1. Introduction

The Mexico City subway (STC Metro) is internationally recognized as one of the most relevant massive transportation systems. It is the main means of transportation in Mexico City, iconic for the country's capital, and plays a fundamental role in the mobility of the Metropolitan Zone of the Valley of Mexico (ZMVM) (Flores de la Mota and Huerta-Barrientos, 2016).

The current network has 12 lines made up of 226 km of tracks and 195 stations, of which 44 are correspondence stations and 127 are transit stations, there are also 12 terminal stations with correspondence and 12 terminals without correspondence. Due to their type, there are 115 underground stations, 55 superficial and 25 elevated (STC Metro, 2017).

The STC Metro, is a decentralize public organization those days started working on September 4th1969, offering service the 365 of the year with different schedules between workdays and weekends. The network is made up of 226 kilometers of railways. The infrastructure of the System is mainly made up of three elements: Rolling Stock, Fixed Installations and Civil Works.

The STC Metro network is basically radial. The STC Metro currently transports 1,057 million people a year, equivalent to 4.1 million people on a weekday. The stations with the highest influx of users, with more than 100 thousand people on average per day, are Indios Verdes (subway Line 3), Pantitlán (subway Line “A”), and Constitución de 1917 (subway Line 8). Figure 1 shows a simplified map to see in an easier way its distribution in Mexico City, at the same time, the lines are shown, with their respective color and number, these were built in increasing order. The ZMVM is delimited by the integration of 16 CDMX mayors, 59 municipalities of the State of Mexico and one municipality of the State of Hidalgo. It covers an area of 7,866.1 km2 and houses a population of 20.8 million inhabitants.

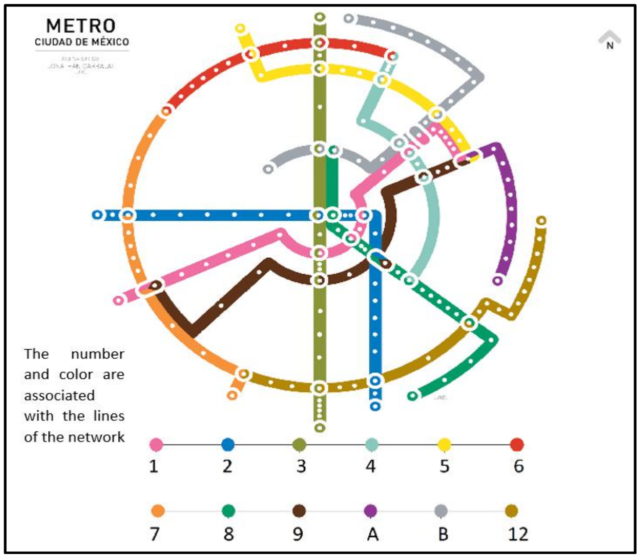

METRO
CIUDAD DE MÉXICO
N
The number and color are associated with the lines of the network
1
2
3
4
5
6
7
8
9
A
B
12

**Figure 1.** Simplified visual representation of the network distribution in Mexico City by JIC website.

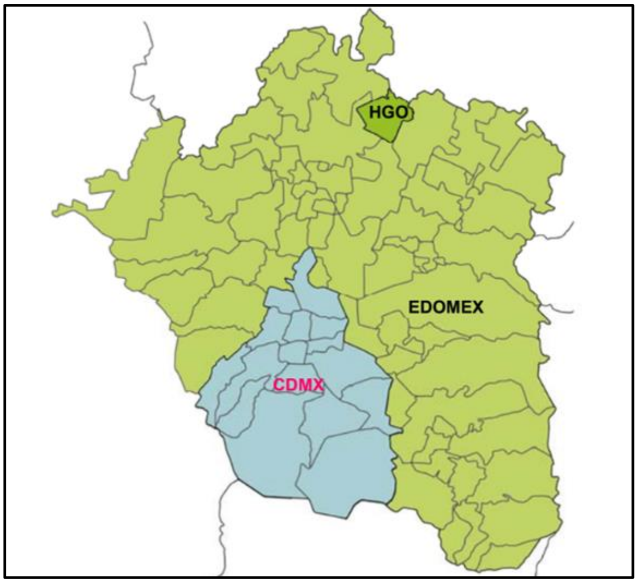


**Figure 2.** Map of the ZMVM, to which the STC Metro provides service. SCT Plan Maestro Del Metro 2018 – 2030.

The main elements of the System infrastructure are as follows:

- The STC Metro Network has a total of 388 trains (321 tires and 67 ferrous), it is made up of 17 models, 4 of them ferrous and the rest pneumatic. 3 types of manufacturing technology are distinguished: 98 trains are of the electromechanical type, JH (camshaft); 193 have the direct current electronic traction control system (Chopper) and the remaining 97 have an alternating current electronic traction control system (Asynchronous).
- The fixed installations play a fundamental role in the operation of the STC Metro, whose operation ensures the circulation of rolling stock throughout the Network; The purpose of the toll system is to control access and exit of users to the stations (Bautista-Martínez *et al.*, 2016).

Regarding the infrastructure of the STC Metro, the workshops that supports the preventive and corrective maintenance are physically located in the following subway stations: Zaragoza, Taxqueña, Ticomán, Ciudad Azteca, La Paz, El Rosario, Constitución de 1917 and Tláhuac.

The natural wear and tear over time, the different types of trains, the difficult modification of the fixed installations, and the technological lag, modify the useful life of the main components, increasing the amount of maintenance, costs, having recurring failures, having collapses when the system is in high demand. Some of the most representative failures in the operation are described below: subway doors, object on track, emergency lever, fight, and evicted train to meet demand.

### *1.1. Causes of the operating failures.*

From the operational perspective, the main problem that causes dissatisfaction in the quality of the service is related to delays in the circulation of trains, which are caused by the lack of rolling stock and breakdowns in trains and fixed installations. During the period 2013 to 2015, the highest percentage of registered events corresponding to 93.13% of all of these, were events with average affectation times of less than 5 minutes (STC, 2017). Every day, at the busy hours the subway capacity is exceeded, causing incidents and failures that affect the correct provision of the service. Due to this situation and to avoid incidents and/or accidents, People Control and Dosage Maneuvers (MCDU) have been implemented in 20 Stations of the network and allocation of cars to women, girls, and boys, thereby increasing the safety of these, however, this measure causes a slight delay in the march of the trains. There is a significant deterioration in the stations that affects the quality of the service provided, a situation that occurs in almost all lines that make up the subway network. (STC, 2017).

### 1.2. *Demand and its relationship with failures in the transportation system*

Based on statistics about the operation of the STC Metro published on its website, it is important to note that in recent days the user demand has increased causing a progressive deterioration with a greater impact on fixed installations and rolling stock. In Figure 3, it is presented the evolution of user demand in the STC Metro for the period from 2012 to 2019. It is important to note that the data covering the period from 2020 to 2022 were eliminated due to atypical behavior due to SARS COV-2.

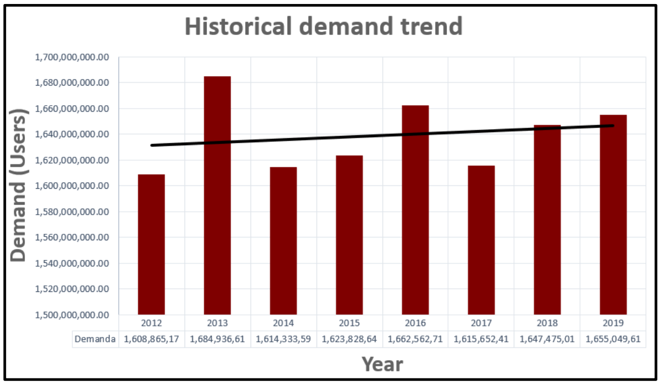


**Figure 3.** History of the user demand in the STC Metro (2012-2019).

According to the diagnosis made by STC Metro in 2017, the accessibility to the network and its proper functioning increases in the busy hours, particularly affecting correspondence stations and terminals. Likewise, the rain and other events unrelated to the internal operation of the system causes delays and some failures, mainly in wagons, trains and sometimes in fixed installations, due to the large number of users that require the service.

The justification for carrying out this study is to show how the user demand impact the increase in failures in the STC Metro network. Additionally, it is important to evaluate the impacts that other variables have on the everyday operation of the STC Metro network through data analysis in order to identify those that are provided by external agents due to demand and give them attention and follow-up in a timely manner.

This paper is organized into four sections. In the following section the systematic literature review about data analysis using statistical methods is described. The methodological approach used for data analysis was multiple linear regression and the main results of the analysis carried out using R™ software are shown in Section 3. Results obtained from R™ software are discussed in Section 4. Finally in section 5 concluding remarks and the relationship between demand and failures in the STC Metro networks are presented.

## 2. Systematic literature review

We followed the literature review process proposed by Machi and McEvoy (2009). The Figure 4 describes the steps for conducting a literature review.

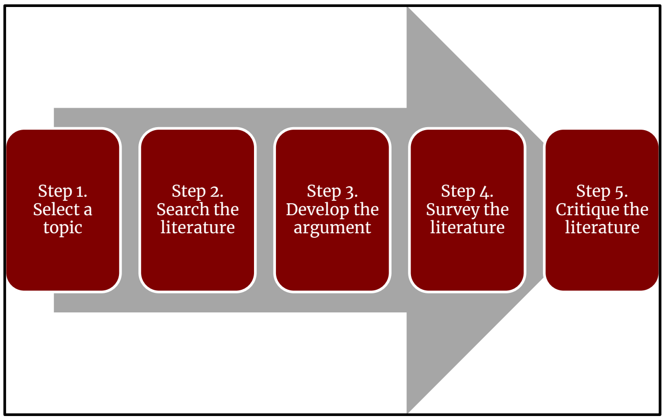

Step 1. Select a topic
Step 2. Search the literature
Step 3. Develop the argument
Step 4. Survey the literature
Step 5. Critique the literature

**Figure 4.** The literature review model, Machi and McEvoy (2009).

### 2.1. *Select a topic*

In the scientific literature we found some simulation studies about the STC Metro related to users, accidents and operational failures (Dillarza-Andrade *et al.*, 2017; Vera-Morales and Huerta-Barrientos, 2016; Portillo-Villasana and Huerta Barrientos, 2016; Bautista-Martínez *et al.*, 2016). However, for the purpose of this study we consider data management and its adequacy, as well as mathematical modeling prior to implement a simulation model.

### 2.2. *Search the literature*

For the bibliographical research, the Scopus™ database was used. We started the bibliographical research using *demand* AND *analysis* AND *subway* as keywords. For our first iteration we got 459 results, then we continued the search using the keywords: *demand* AND *statistics* AND *linear* AND *regression*, getting 638 results, after that, the next investigation was using the keywords *linear* AND *regression* AND *failure* AND *prediction* AND *simulation*, getting 152 results.

### 2.3. *Develop the argument*

The main objective of the literature search was to find the literature related to the failure and demand of a system and the related statistical mathematical methods that are being used to study this, and to see how this topic relates to transportation systems. In the second part of the literature search, the objective was to find the relationship between mathematical modeling and simulation, failures, and demand, identifying that machine learning can be used in combination with simulation models to improve their performance and predictive capacity.

### 2.4. Survey the literature

We downloaded the biographical information from the Scopus™ database and then used the VOSviewer™ software to group keywords based on match. VOSviewer™ software is a tool to visualize bibliometric networks and the relationship of terminology and specific keywords based on their distance in the network (Van Eck & Waltman, 2014).

Figures 5 and 6 presents the results of the search using the key words: demand, statistics, and linear regression used and the relationship with subway systems.

Figures 7 shows the relationship of the computer simulation with the main applications of mathematical models. Figure 8 presents the interrelations that were found in the literature between the *demand* and the *prediction of failures*.

### 2.5. Critique the literature

Based on the research carried out in the Scopus™ database and related studies, we can identify the importance of correctly managing data, and how mathematical models and machine learning can help simulation, to understand the relationship of failures in transport systems for correct decision making and risk prevention. Regarding modeling, linear regression was the most common method used for data prediction, however, we were able to identify that it has strong support in machine learning and data mining.

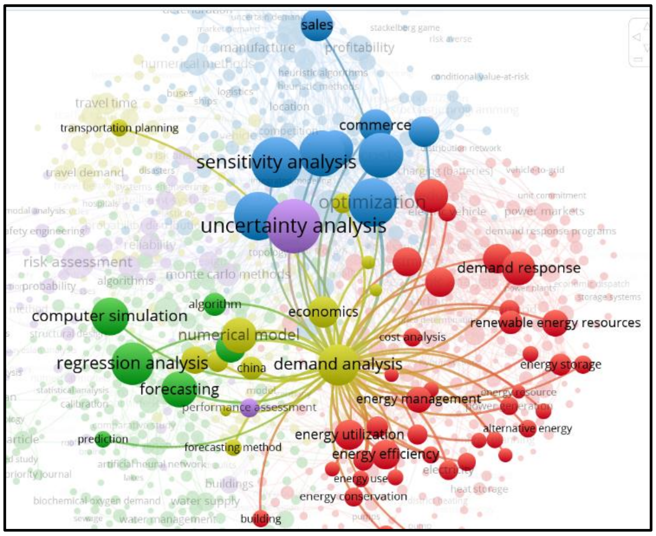


**Figure 5.** Search for key terms in Scopus™: *demand* AND *analysis* AND *subway*, based on the co-occurrences using VOSviewerTM software for search based on ScopusTM database.

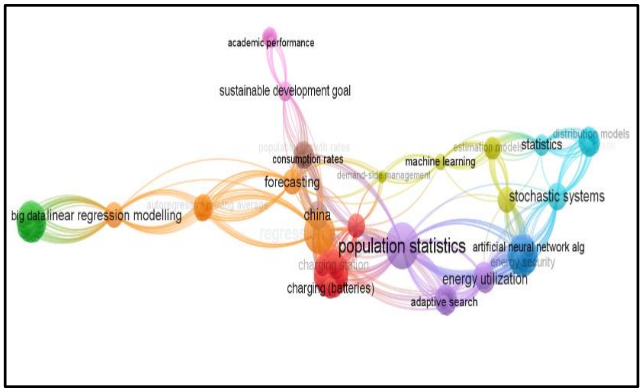


**Figure 6.** Clustering key words in Scopus™: *demand*, AND *statistics*, AND *linear* AND *regression.* Related Results (683 218), based on the co-occurrences using VOSviewer™ software for search based on Scopus™ database.

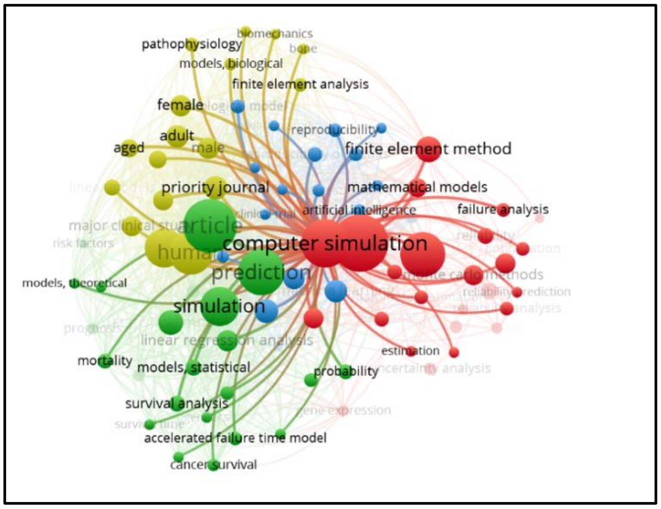


**Figure 7.** Clustering key words in Scopus™: *demand* AND *linear* AND *regression* AND *failure* AND *prediction*, based on the co-occurrences using VOSviewer™ software for search based on Scopus™ database.

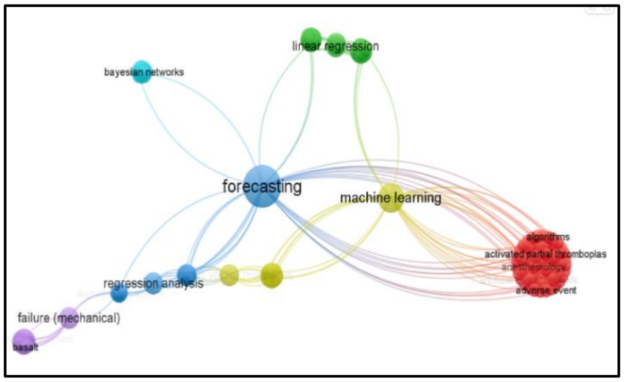


**Figure 8.** Clustering key words in Scopus™: *demand AND linear AND regression AND failure AND prediction*, based on the co-occurrences using VOSviewer™ software for search based on Scopus™ database.

## 3. Materials and methods

### 3.1. Data

For this research work, two databases were used (BD Failure registered in the period January - September 2022 and BD Demand registered in the period January - September 2022) provided by the Operations Management of the STC Metro. The database that contains the faults that occur in the subway, has 16,000 records throughout the network. For this study, 3 subway lines with high demand were selected, using 800 records for the analysis. The data was cleaned and organized, through the use of tables in Excel, renaming and classifying the data for its correct integration into R™ software. Since the databases were independent, the time periods had a gap, therefore, all the data were organized so that they were presented quarterly.

For this study, only the data of subway lines: 2, 3 and 4 of the STC Metro were considered, since they are lines that have outdoor and underground stations, the set of these lines represents 34.48% of the total demand of the STS Metro network. When integrating the databases, individual aspects per station were considered, such as: subway station (what station it is), subway line (what line it belongs to), type of station (way station, correspondence, terminal or terminal-correspondence), facility type (Underground, surface or elevated), user demand (Number of users that entered that station), failures (Failures that occurred at that station), time (Time delay due to failure) and frequency (Number of times of a failure).

### *3.2. The statistics software*

For the implementation and analysis of the model, a software was required that would allow a mathematical analysis of the data and that in turn would present graphs for an easy interpretation, for the present investigation the open-source R™ software was chosen.

### *3.3. Methods*

This study presents a methodological approach that integrates, manipulation and treatment of massive data, statistical inference, multiple linear regression, programming in R™ software. To understand the behavior of the failures with respect to the demand and to certain physical characteristics of the stations or rolling stock of the STC Metro. It was decided to use a multiple linear regression model to evaluate the relationship between the response variable and multiple predictor variables and determine the relative contribution of each one of them to the model. In this case, the aim is to know the degree of influence that the demand, time, facility, station has with respect to the failures that occur on lines 2, 3 and 4 of the SCT Metro network.

Let be the general mathematical structure of a multiple linear regression model:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta nXn + \varepsilon \quad (1)$$

Where:

- Y = Dependent variable
- X1, X2, ..., Xn = Are the n independent variables
- β0 = It is the intercept or the constant
- β1, β2, ..., βn = Are the regression coefficients that indicate the change in Y for each unit change in the corresponding independent variable.
- ε = The random error that cannot be explained by the variables included in the model.

### *3.4. Model Implementation*

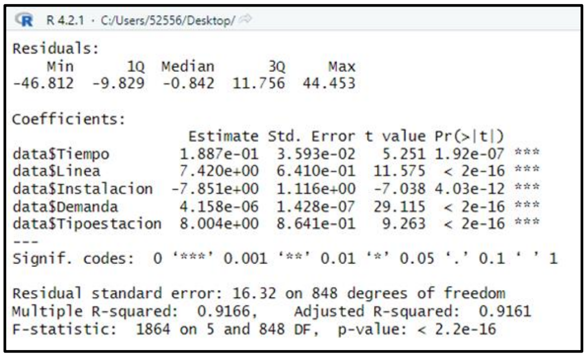
The proposed model for total failures of subway lines 2, 3 and 4 of the STC Metro is:

$$\text{Totalfailures} = \beta 0 + \beta 1\text{Time} + \beta 2\text{Line} + \beta 3\text{Facility} + \beta 4\text{Demand} + \beta 5\text{StationType} + \varepsilon$$

Where:

- Total failures: is the dependent variable or response variable.
- β0: is the intercept, the value of Totalfailures when all the independent variables are zero.
- β1, β2, β3, β4, β5: are the regression coefficients that indicate the relationship between each independent variable and the dependent variable.
- Time, line, installation, demand, station type: they are the independent variables or predictors.
- ε: It is the error term or residual, which represents the difference between the observed values and the values predicted by the model.

This model was entered into the R™ software and the data analysis began. When running the "Summary" function of the software, it provided us with a summary of the most basic statistical information, as shown in Figure 9.



```
Residuals:
    Min      1Q  Median      3Q     Max 
-46.812  -9.829  -0.842  11.756  44.453 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
data$Tiempo      1.887e-01  3.593e-02   5.251 1.92e-07 ***
data$Linea       7.420e+00  6.410e-01  11.575  < 2e-16 ***
data$Instalacion -7.851e+00  1.116e+00  -7.038 4.03e-12 ***
data$Demanda     4.158e-06  1.428e-07  29.115  < 2e-16 ***
data$Tipoestacion 8.004e+00  8.641e-01   9.263  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.32 on 848 degrees of freedom
Multiple R-squared:  0.9166,	Adjusted R-squared:  0.9161 
F-statistic:  1864 on 5 and 848 DF,  p-value: < 2.2e-16
```

**Figure 9**. Elementary statistical data of the proposed model.

As it is noted, Figure 9 shows that the model explains approximately 92% of the variability, there is a small *p* value, which indicates that the model is significant, that is, all the variables are significantly associated with the dependent variable. Once the basic statistics have been reviewed and inferences or advanced conclusions made and as the methodology says, the assumptions of: Linearity, Normality, Homoscedasticity, Multicollinearity, and identification of influential variables must be verified. These assumptions were reviewed using R™ software, and it gave us the following results.

## 4. Results and Discussion

Below the results of the assumptions to be checked to validate the reliability and significance of the variables in the model are described.

### 4.1. Linearity

For this assumption, all the variables were analyzed using a Pearson correlation to analyze the linearity of the data. We have 3 statistically significant variables out of 6. Table 1 shows the results of the Pearson tests, explaining the type of correlation and whether it is significant or not. The importance of verifying that all the variables are associated with each other is because it can be considered that they do not contribute anything to the model, generating deviations and limiting its quality.

### 4.2. Normality

Figure 11 shows the results obtained by the R™ software for the normality assumption. There is significant evidence to reject the null hypothesis of normality of the residuals, which suggests that the residuals do not follow a normal distribution. The residuals do not follow a normal distribution, it is important to take into account that in linear models with a large sample size, normality tests may be sensitive to small deviations from normality.

**Table 1.** Linearity test results.

| | Pearson correlatio n | p-value | There is correlation | It is statistically significant? |
|---|---|---|---|---|
| Demand -total | 0.7707 | <2.2e-16 | Strong | Yes |
| Demand-Time | 0.1623 | 1.83e-06 | Weak positive | Yes |
| Demand -Line | -0.0739 | 0.0308 | Weak positive | No |
| Demand-Facility | -0-0187 | 0.5851 | Very weak positive | No |
| Demand-S type | -0.5642 | <2.2e-16 | Moderate positive | Yes |
| Demand-Station | 0.0219 | 0.5216 | Very weal positive | No |

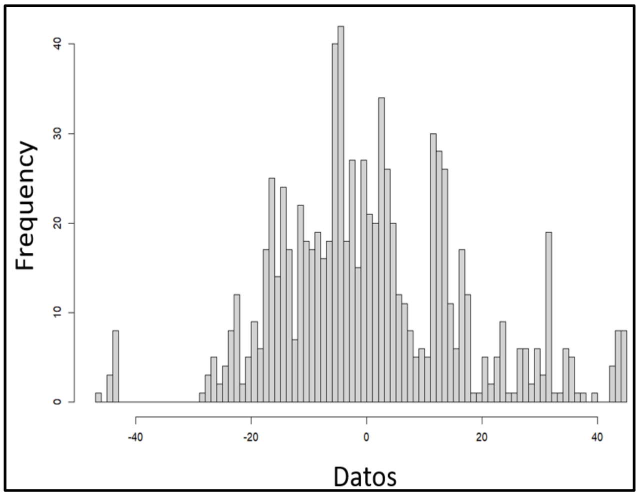


**Figure 10.** Representation of normal distribution of data.

It is possible to observe a normal distribution with variations at the extremes, so the lag could be attributed to sensitivity due to the large sample size. The importance of verifying normality is due to the fact that the global validation tests of the model could not be applied, since they require that there be normality in the residuals.

### 4.3.Homoscedasticity

Figure 11 shows the results produced by the R™ software for this assumption. Figure 11 shows that the variance of the errors is not constant throughout the range of the predictors. Therefore, the errors may be correlated with some other variable that is not included in the model and this may affect the precision of the estimates.

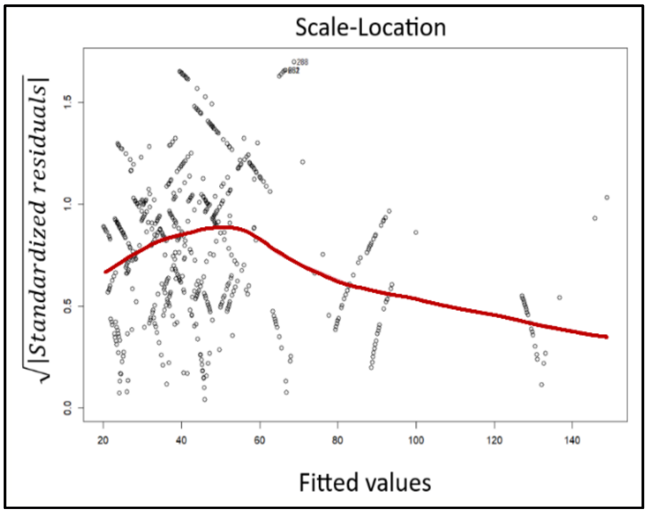
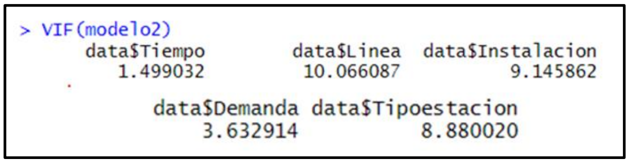


**Figure 11.** Representation of homoscedasticity provided by R™ software.

### 4.4. Multicollinearity

To corroborate the multicollinearity, we have the VIF test, as can be seen in Figure 12, the VIF results indicate that there is no severe multicollinearity in the model. The VIF values are all less than 10, which suggests that there are no significant multicollinearity problems. In this case, the highest value of the VIF is 10.07, which indicates that the "Line" variable is highly correlated with the other explanatory variables. This may indicate the presence of multicollinearity, which must be considered when interpreting the model results.



```
> VIF(modelo2)
      data$Tiempo          data$Linea  data$Instalacion
         1.499032           10.066087          9.145862

           data$Demanda data$Tipoestacion
               3.632914          8.880020
```

**Figure 12.** VIF Test Results by R™ software.

### 4.5. Influential Values Identification

To corroborate the identification of influential values we have the Cook's distance test or statistic, the R software directly provides us with a graph for reviewing the cook's distance, the red line should be horizontal and not vertically exceed the number 1 (see Figure 13). There are no points or outliers in the model, they are all within the established range, the outliers are outside the limit marked with the number 1, We can see that the trend line marks a horizontal line with a slight deviation.

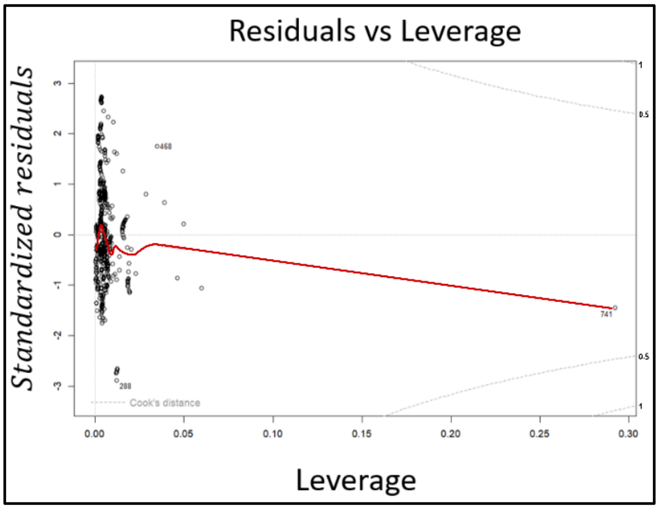


**Figure 13.** Results of Cook's test on R™ software.

## 5. Justification

Data handling and multiple linear regression play fundamental roles in computer simulation and underscore their importance in obtaining accurate and meaningful results. First, efficient data handling is essential for accurate simulation. The quality of the data used in the simulation directly impacts the reliability and validity of the results obtained. The data handling process involves tasks such as cleaning, transformation, and selection of relevant features. By ensuring that the data is error-free, complete, and representative, the potential for bias is reduced and the quality of the resulting simulation model is improved. Second, multiple linear regression plays a crucial role in simulation. Unlike simple linear regression, multiple linear regression allows you to model complex relationships between multiple independent variables and one dependent variable. This is especially relevant in simulations involving systems with multiple input factors that affect the behavior or outcome of the system. By using multiple linear regression in simulation, interactions between variables can be showed and quantified more accurately, resulting in a more realistic and representative model.

## 6. Conclusions

The regression coefficients are significant (with a confidence level of 95%) and have an interpretation consistent with the logic of the problem, that is, the variables demand, and type of station are closely related to the failures that occur in subway lines 2, 3 and 4 of the STC Metro Network.

There is a strong relationship between breakdowns and demand, so the greater the number of users the metro has, the greater the number of breakdowns, either in fixed installations or in trains. It is possible to simulate several scenarios with the model in the R™ software to visualize the critical states of the system. It can be inferred that failures are mostly caused by an external factor, such as the number of users. When the demand exceeds the supply of transport, it causes incidents during the operation that affect the correct provision of the service.

The way in which the data is taken or administered is wrong, since by presenting it only on a quarterly basis, it limits a more in-depth and specific analysis.
After verifying that the demand generates failures, the database was manipulated and when classifying the failures caused by the demand, these represent 31% of the total failures in the network. In general, the model is adequate to explain the variability in the fault data. It meets 3 of 5 assumptions, although these can be adjusted for greater reliability.

### 6.1. Limitations and future research

A limitation was identified to the presentation and administration of the databases by the STC Metro. The way the data was handled by the STC Metro limited our options for handling it, as the databases were separated into different time periods. This circumstance complicated the analysis, since our shortest period of time is quarterly, eliminating the option of being able to do it per day, week, or even month.

The task of combining both databases was performed manually, data by data. This approach represented a restriction for the investigation, since we had more than 16,000 data. As a result, we were forced to limit the scope of the investigation and focus on the analysis of a particular couple of lines. In future investigations, it is expected to be able to use the data from all the lines, in a period of time that can help us understand the behaviors on dates of interest, such as business days, weekends, vacations, holidays, massive events, etc.

Likewise, apply the use of computational technology, such as artificial intelligence to generate a model through neural networks that can generate predictions and analysis more quickly and accurately. Regarding the data, it is possible to apply mathematical transformation tools for the assumptions that are not met, such as:

- Sensitivity analysis: It could be consider adjusting the model using different combinations of variables or their transformations to assess whether the model remains robust.
- Cross Validation: Cross validation can be performed to assess the performance of the model on new data sets and to assess the generalizability of the model.
- Influence analysis: In addition to the Cook's distance test, other influence tests can be considered to identify if there are data points that have a disproportionate impact on the model.
- Measurement proposal: Measure different variables that are not yet averages or carry out the measurement in a different way from the current one. This to adjust the model as much as possible and find the one that is optimal, ensuring its quality and confidence for decision-making in the STC Metro.

## Funding

The sources of funding for this study are: CONAHCYT and the Faculty of Engineering of the National Autonomous University of Mexico (UNAM) and PAPIME UNAM.

## Acknowledgements

The authors greatly appreciate the partial financial support by the National Autonomous University of Mexico (UNAM). Also, the authors thanks to PAPIME PE104022.

## References

Bautista-Martiínez, H. O., Huerta-Barrientos, A., Portillo-Villasana, G. J. (2016). Minimizing the impact of escalator failures in metro Tacubaya subway station on user´s mobility. Proceedings of the 2016 European Modeling and Simulation Symposium, EMSS 2016, pp. 224-230.

Cheng, C.W., & Chen, D.W. (2018). The researches on subway demand forecast at station level: Smart card data, Space Syntax and Points of Interests. China: IOP Conf. Series: Materials Science and Engineering.

Devore, J. L. (2015). Probabilidad y estadística para ingeniería y ciencias. México: Cenage Learning.

Dillarza-Andrade, J., Huerta-Barrientos, A., Salazar Dñiaz, G., Pérez-Bonilla, J. (2016). Simulation of boarding pedestrian in Mexican subway: the case of Pantitlan terminal station. Proceedings of the 2017 European Modeling and Simulation Symposium, EMSS 2017, pp. 468-473.

Flores de la Mota, I. and Huerta-Barrientos, A. (2016). A proposal for optimizing urban mobility in Mexico City based on the public transport network. Proceedings of the 2016 European Modeling and Simulation Symposium, EMSS 2016, pp. 282-286.

Galal, A.A., & Charles, S.B. (2011). Comparative analysis and prediction of traffic accidents in Sudan using artificial neural networks and statistical methods. Sudan: Proceedings of the 30th Southern African Transport Conference (SATC 2011).

Hines, W. W., Montgomery, D. C., Goldsman, D. M., & Borror, C. M. (2014). Estadística y probabilidad para ingeniería. México: Grupo Editorial Patria.

Mendenhall, W. M., & Sincich, T. L. (2016). Statistics for engineering and the sciences. Florida: CRC Press.

Sistema de Transporte Colectivo Metro (2017). Plan Maestro 2018-2030. Mexico City, Mexico.

Portillo-Villasana, G. J. and huerta-Barrientos, A. (2016). A simulation model for assigning secure waiting areas on subway platforms to minimize accidents. Proceedings of the 2016 European Modeling and Simulation Symposium, EMSS 2016, pp. 242-248.

Vera-Morales, A.E. and Huerta-Barrientos, A. (2016). Simulation optimization of pedestrian evacuation in Mexican subway: the case of Pino Suarez station. Proceedings of the 2016 European Modeling and Simulation Symposium, EMSS 2016, pp. 175-182.