Revista Facultad de Ingeniería

Title: **Identification of variables influencing the origin of traffic congestion points**

Authors: Ernesto de la Cruz-Nicolás, Hugo Estrada-Esquivel, Alicia Martínez-Rebollar, Eddie Clemente, Odette Pliego-Martínez

DOI: **10.17533/udea.redin.20250261**

To appear in: *Revista Facultad de Ingeniería Universidad de Antioquia*

Received: October 30, 2024
Accepted: February 28, 2025
Available Online: February 28, 2025

Identification of variables influencing the origin of traffic congestion points

Identificación de las variables que influyen en el origen de los puntos de congestión de tráfico

Ernesto de la Cruz-Nicolás[1*] https://orcid.org/0000-0002-6574-0339, Hugo Estrada-Esquivel[1] https://orcid.org/0000-0002-1466-7581, Alicia Martínez-Rebollar[1] https://orcid.org/0000-0002-1071-8599, Eddie Clemente[1] https://orcid.org/0000-0003-3195-9540, Odette Pliego-Martínez[1] https://orcid.org/0000-0002-1793-4921

[1]Ciencias de la Computación, TecNM-Centro Nacional de Investigación y Desarrollo Tecnológico. Interior Palmira Avenue S/N, Col. Palmira. C.P. 62490. Cuernavaca, Morelos, México.

Corresponding author: Ernesto de la Cruz-Nicolás
E-mail: d21ce090@cenidet.tecnm.mx

KEYWORDS
Urban Congestion; traffic patterns; urban mobility bottlenecks
congestión urbana, patrones de tráfico, cuellos de botella en la movilidad urbana

ABSTRACT: The identification of variables triggering traffic congestion points in cities has been revealed as an extraordinarily complex task, involving characterization, prediction, and forecasting activities of traffic congestion points. The exhaustive literature review indicates that key factors influencing urban traffic congestion include traffic incidents, road infrastructure conditions, the day of the week, time of day, the month, workweeks, and vacation periods. However, there is a lack of quantitative representations of traffic that allows us to accurately assess the degree of influence of relevant traffic variables such as incidents and service locations in the generation of traffic congestion points. In this context, the main objective of this research work focuses on identifying the most significant variables causing traffic congestion points through the use of mathematical techniques. This aims to contribute to the development of models for mitigating traffic congestion. The case study of traffic, incidents, and services in Mexico City is used in this work.

RESUMEN: La identificación de las variables que generan puntos de congestión vehicular en las ciudades se ha revelado como una tarea extraordinariamente compleja, que implica actividades de caracterización, predicción y proyección de los puntos de congestión vehicular. La revisión exhaustiva de la literatura indica que los factores clave que influyen en la congestión vehicular urbana incluyen los incidentes de tráfico, las condiciones de la infraestructura vial, el día de la semana, la hora del día, el mes, las semanas laborales y los períodos vacacionales. Sin embargo, existe una falta de representaciones cuantitativas del tráfico que permitan evaluar con precisión el grado de influencia de variables relevantes, como los incidentes y las ubicaciones de los servicios, en la generación de puntos de congestión vehicular. En este contexto, el objetivo principal de este

trabajo de investigación se centra en identificar las variables más significativas que causan puntos de congestión vehicular mediante el uso de técnicas matemáticas. Esto busca contribuir al desarrollo de modelos para mitigar la congestión vehicular. En este trabajo se utiliza el estudio de caso del tráfico, los incidentes y los servicios en la Ciudad de México.

1. Introduction

One of the main issues of modern cities is the high volume of traffic, which leads to the occurrence of points of congestion or traffic jams, as stated by [13], highlighting its significant impact on urban operations. Traffic congestion can be defined as the situation in which traffic demand exceeds the capacity of the available road infrastructure, as explained by [15]. Traffic congestion can arise in any part of the city due to the presence of various variables such as inadequate and insufficient public transport, design of road infrastructure, traffic policies, road incidents, among others, according to [14]. Aside from these variables, environmental, human and cultural factors contribute to the problem of mobility. Traffic congestion points originate in several zones of the cities and cause inadequate operational performance of the road infrastructure, as described in the works of [16, 18].

The evaluation of the performance of road infrastructure has been addressed through the use of indicators that allow measuring the level of service of streets based on data collected in various locations of the city, especially in those strategic points where traffic congestion is recorded, as described in [17]. The indicators commonly found in the literature to represent traffic are speed, traffic density, flow, congestion index, travel time, and origin-destination analysis. The studies of [23, 24, 25, 26] focus on high-performance measures and indicators to assess traffic congestion behavior.

One of the factors affecting the evaluation of road infrastructure is the multiplicity of measurement indicators, which in turn are often composed of a variety of variables. This complicates the selection of the most relevant and appropriate parameters for conducting an effective assessment of road infrastructure performance. This task becomes even more challenging when considering the need to adapt predictive models of traffic congestion to the specific context of each research, as mentioned in the works of [19, 20, 21, 22].

Multiple models have been developed to identify traffic congestion, taking into account a variety of environmental, health, economic, and other variables. In this context, two research questions arise: (1) What are the environmental variables that trigger traffic congestion? and (2) What is the quantitative value of each variable that contributes to traffic congestion?

In this paper, a multifaceted approach is presented to identify the variables that contribute to the generation of traffic congestion hotspots, which is achieved through the analysis and exploration of a set of variables related to traffic, road incidents, and services. Mexico City has been used as a case study due to its severe traffic congestion, ranking 22nd in the world, according to the ranking conducted by [27].

2. Related Works

Rapid urbanization in major cities around the world has led to an unprecedented increase in traffic congestion, generating health and pollution problems, as noted in [2]. This complexity has driven the

development of numerous studies to address the issue, although the proposed solutions so far are partial due to the large number of variables contributing to vehicle congestion, as highlighted by [12].
Understanding the most influential variables in traffic congestion is crucial for improving the accuracy of predictive models and optimizing mitigation techniques. The research of [1] evaluated essential variables affecting vehicular flow in a specific area, finding only instantaneous relationships between climatic and non-climatic variables. In [4] it is stated that 66% of congestion is due to driver behavior and poor practices, while 34% is attributed to road conditions.
The work of [5] considers that traffic accidents and the attention they receive influence congestion, and the research of [6] used multiple linear regression models to identify variables such as street area per person and vehicle ownership. The works of [7] and [8] focused on factors causing congestion at intersections and stopping points.
It was identified that attraction points, such as shopping centers and schools, generate congestion [9]. A Structural Equation Modeling was employed to analyze predominant factors in geographic areas [10], and it was emphasized that street design and poor driving behaviors are also contributing variables [11].
Recent studies add further complexity to this issue. For example, [37] quantified the impact of incidents on speed reduction. The research of [38] explored factors influencing the operational state of urban traffic, and the study in [39] conducted a thorough review of traffic congestion.
In the case of [40], urban scale factors in the U.S. were analyzed, while in the study of [41], the relationship between urban morphology and congestion was examined. In [42], the impact of land use was investigated, and finally, the work of [43] measured congestion using innovative metrics.
The identification of causal variables has facilitated the development of indicators to assess the quality of road infrastructure. In [3], the authors analyzed relevant indicators and their correlation with population density and public transport availability, highlighting the significant role of the number of taxis in congestion.
Each of these studies identified crucial variables contributing to traffic congestion. In our work, we propose a set of variables related to traffic, road incidents, and services that act as causes of congestion. These variables were analyzed using mathematical techniques to provide data scientists with a more comprehensive understanding of the factors involved. This analysis can inform effective traffic management strategies, road infrastructure planning, and improvements in predictive models.

3. Methodology for identifying variables influencing the origin of traffic congestion points

The identification of causal variables for traffic congestion points was carried out through a methodological approach composed of five main phases. The first phase involved the experimental design, which established a structure that must be followed in each experiment. In the second phase, a case study was established, selecting the city from which traffic, road incidents, and services data were obtained. The third phase focused on data collection, describing the tools that were used to create a dataset of the traffic, road incidents, and services. In the fourth phase, the description of the traffic variables, road incidents, and services was carried out. Finally, in the fifth phase, the steps that needed to be followed for identifying the most influential variables that cause traffic congestion points were detailed.

3.1. Experimental Design Phase

The experiments to identify the variables causing traffic congestion were conducted multiple times until the desired correlation and relationship values were found with each mathematical technique used in this study. For each experiment conducted in this research work, the following proposed structure was used.

a) Experiment description: Analysis of the variables causing traffic congestion.
b) Experiment objective: Identification of the most influential variables in the onset of traffic congestion.
c) Hypothesis definition: In geographical environments, there are certain variables that have more influence than others and are responsible for the origin of traffic congestion.
d) Sample definition: A random sampling method was used to select a representative sample from the large population of data using Equation 1. With this sampling, it was ensured that each member of the population had an equal probability of being selected to be part of the sample.

$$N = \frac{\text{Number of elements selected}}{Number\ of\ elements\ in\ the\ population} \quad (1)$$

e) Definition of the dependent variable and set of independent variables: The dependent variable chosen in this research work is traffic congestion (jamFactor). This variable represents the level of traffic congestion in a specific geographical area or at a particular moment. It refers to the amount of obstruction or blockage of traffic flow in a road network and can be quantified using various independent variables. In this research, a set of independent variables is considered and proposed, which are described in Table 1. Traffic congestion is a common problem in densely populated urban areas and can have significant impacts on transportation efficiency, air quality, public health, and local economies. Therefore, understanding and effectively addressing traffic congestion is crucial for improving mobility and the quality of life in cities.

The experiments were conducted using the R software [34], on a computer with an Intel Core i7 processor running at 2.20GHz, 16 GB of RAM, and a 64-bit Windows operating system.

3.2. Case study definition phase

Mexico City was chosen as the case study due to its significant urban mobility challenges. The study focuses on Mexico City, which is divided into 16 boroughs. According to the TomTom Traffic Index [28] in 2023, travel time increased in all 16 boroughs of Mexico City. The data indicates that the average time to travel 10 km increased by 50 seconds, reflecting a growth in traffic congestion in various boroughs of Mexico City. In this context, the analysis of the traffic data collected daily revealed that traveling 10 km in three time frames (12:00 to 6:00 am, 8:00 am to 2:00 pm, and 4:00 pm to 10:00 pm) requires an average time in minutes that varies depending on the day of the week. In the 12:00 to 6:00 am block, the averages were: 14, 15, 16, 17, 18, 16, and 16 minutes for Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, and Saturday, respectively. For the 8:00 am to 2:00 pm block, the times were 18, 25, 29, 30, 28, 27, and 26 minutes. In the 4:00 pm to 10:00 pm block, the time invested was 18, 25, 36, 33, 29, 25, and 27 minutes.

3.3. Data collection phase

Data collection was carried out continuously in 5-minute intervals throughout the 24 hours of the day, from September 2023 to January 2024, using the methodology proposed by [35]. The data collection focused on traffic and road incident data. The software tools used were Here Maps [29, 30, 31] for collecting traffic and road incident data, and Geoapify [32] for collecting data related to services. The amount of information collected regarding traffic, road incidents, and services is presented in **Table 1.**

Table 1 Dataset of traffic, traffic incidents, and services

Dataset	Records	Number of variables
Traffic	6,857,357	15
Incidents	3,381,244	18
Services	10,000	1

The integration of traffic, road incidents, and services datasets was carried out through a cross-product, resulting in a single integrated dataset named 'Traffic_Incidents_Services'.

3.4. Description of data from the dataset Traffic_Incidents_Services

The Traffic_Incidents_Services dataset includes variables that impact traffic congestion, supported by both literature and empirical evidence. An Analysis of Variance (ANOVA) was conducted to evaluate the p-values of each variable in relation to the dependent variable (jamFactor). A p-value of less than 0.05 indicates significant differences among group means, leading to the rejection of the null hypothesis. **Table 2** presents all variables with significant p-values, such as Bbox_Traffic (0.048), Traffic_Day_Number (0.016), Day_Off_Incident (0.030), Confidence (0.015), Work_Day_Traffic (0.035), Incident_Hour (0.037), Free_Flow (0.044), Traffic_Block_Type (0.021), Incident_Month (0.001), Length (0.030), Bbox_Incident (0.003), Incident_Minute (0.039), Number_Segments (0.023), and Road_Closed (0.026), confirming their statistical relevance in traffic congestion.

Table 2 Description of the variables in the dataset Traffic_Incidents_Services [44]

Number	Dataset	Variable Name	Variable Description
1	Traffic	Bbox_Traffic	Originating traffic bottleneck
2		Confidence	Describe whether the data is real-time or historical
3		Free_Flow	The reference speed (in meters per second) along the street when there is no traffic
4		Length	Street length
5		Number_Segments	Number of segments contained in the monitored street
6		Speed	The expected speed along the street; does not exceed the legal speed limit
7		Speed_Uncapped	The speed along the street; may exceed the legal speed limit
8		Daily_Traffic	Day on which traffic is monitored
9		Day_Off_Traffic	Value of 1 when the day is a holiday and 0 when it is not
10		Traffic_Hour	Traffic monitoring time, formatted as HH: MM
11		Traffic_Month	Month of traffic monitoring
12		Traffic_Minute	Minute of traffic monitoring
13		Traffic_Day_Number	Day number of traffic monitoring (1=Monday, 2=Tuesday...7=Sunday)
14		Work_Day_Traffic	Traffic monitoring on weekdays and weekends (weekday=1, weekend=0)
15		Traffic_Block_Type	Traffic by time blocks (block 1=0-3 hours, block 2=4-7 hours, block 3=8-11 hours... block 6: 20-23 hours)
16	Incidents	Bbox_Incident	Incident origin bottleneck
17		Road_Closed	Indicates if the street is closed due to the incident (0 for open and 1 for closed)
18		Criticality	Represents the severity of the incident. It has the following values: low - less severe, minor, major, and critical
19		Type	Describes the type of incident: accident, construction, congestion among others

20		Description	Provides a description of the incident location
21		Olr	Incident form ID
22		Start_Time	Start time of the incident
23		End_Time	End time of the incident
24		Incident_Day	Day of the incident monitoring (values between 1 and 30, or 1 and 31)
25		Day_Off_Incident	Value of 1 when the day is a holiday and 0 when it is not for the incident monitoring day
26		Incident_Hour	Hour of the incident monitoring
27		Incident_Month	Month of the incident monitoring
28		Incident_Minute	Minute of the incident monitoring
29		Incident_Day_Number	Day number of incident monitoring (1=Monday, 2=Tuesday...7=Sunday)
30		Work_Day_Incident	Incident monitoring on weekdays and weekends (weekday=1, weekend=0)
31		Category_Number	Category of the incident
32		Time	Time in minutes, for example, 0:00 hours is minute 0, 0:05 hours is minute 5, 0:10 hours is minute 10, and so on until 23:55 hours, which is minute 1435
33		Name	Description of the service
34	Services	Bbox_Service	Service delimiter

In this research, the dependent variable is 'jamFactor', which represents traffic congestion using a predetermined scale, as detailed in **Figure 1.** The variables described in **Table 2** will play the role of independent variables. The level of congestion will be determined by the correlation or relationship between these independent variables.

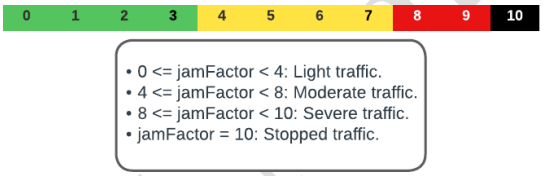

Figure 1. Scale of traffic congestion values (jamFactor) proposed by Here Maps [36]

The Variance Inflation Factor (VIF) was also implemented in the Traffic_Incidents_Services dataset to measure multicollinearity among independent variables in a regression model. A VIF greater than 1 indicates some collinearity, while values above 5 or 10 suggest high multicollinearity, which can affect the stability of the coefficients. In the Traffic_Incidents_Services analysis, variables with VIF > 5 were identified, such as Free_Flow, Speed, Traffic_Hour, Traffic_Block_Type, Road_Closed, Incident_Hour, Bbox_Incident, and Incident_Month, indicating high correlation. Although this suggests redundancy, each variable provides a unique perspective on congestion, reflecting their combined influence on traffic.

3.5. Identification of the variables influencing the onset of traffic congestion points using mathematical techniques.

The identification of variables influencing congestion points was conducted by using several statistical methods. First, the analysis of statistical means reveals that variables with mean values exceeding the overall average indicate a greater impact on congestion. Spearman correlation identifies variables with higher correlation to congestion points, assigning them greater weight based on monotonic relationships. Principal Component Analysis (PCA) uses a rotation matrix to highlight original variables that contribute significantly to principal components, with higher loadings reflecting greater influence on congestion. Structural Equation Modeling (SEM) reveals variable interactions through a factor loading matrix, where

higher loadings indicate stronger impacts on congestion modeling. Subsequently, factor analysis determines relationships among variables, with higher loadings representing greater explained variance related to congestion. Finally, changes in variables such as speed, incident type, or segment length are considered statistically significant (p-value < 0.05) when they exceed a defined threshold, highlighting their relevance in understanding traffic congestion. The identification of relevant variables using each of these mathematical techniques is presented below.

3.5.1 Identification of variables influencing the onset of congestion points through the mean statistic

The statistical analysis of means aims to identify the variables that significantly influence the onset of road congestion. This approach allows for the analysis of the relationship between the independent variables described in **Table 2** and the presence of congestion, assessing their mean values in relation to the dependent variable (congestion). By comparing the means of the independent variables, it is possible to identify those that exhibit significant differences in various scenarios, suggesting that these variables have a considerable impact on the emergence of congestion. The steps are described below:

- Step 1: Obtain the mean values of the independent variables. The arithmetic mean of the independent variables was calculated using equation number 2.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (2)$$

Where:

n is the total number of records in the dataset.

x_i is the individual values in the dataset from i=1 to i=n.

The mean values obtained from the independent variables of the Traffic_Incidents_Services dataset are shown in **Table 3.**

Table 3 Representative mean values of each variable from the dataset Traffic_Incidents_Services

Number	Variable	Mean value	Number	Variable	Mean value
1	Bbox_Traffic	32.319297	17	Criticality	2.802658
2	Free_Flow	10.815257	18	Type	220.314207
3	jamFactor	1.909658	19	Description	6.667524
4	Length	2301.60852	20	Olr	2.711761
5	Number_Segments	0.927815	21	Start_Time	608.070831
6	Speed	9.210867	22	End_Time	191.787018
7	Speed_Uncapped	9.421041	23	Incident_Day	11.173299
8	Daily_Traffic	11.171208	24	Incident_Dayinhabil	0.099745
9	Day_Off_Traffic	0.099745	25	Incident_Hour	14.300548
10	Traffic_Hour	13.45928	26	Incident_Month	10.700803
11	Traffic_Minute	27.226993	27	Incident_Minute	27.300381
12	Traffic_Day_Number	3.909158	28	Incident_Day_Number	3.909862
13	Work_Day_Traffic	0.699212	29	Work_Day_Incident	0.699082
14	Traffic_Block_Type	3.987062	30	block_type_incident	4.180406
15	Bbox_Incidents	32.319297	31	Category_Number	12.255365
16	Road_Closed	1	32	Bbox_Service	32.319297

- Step 2: Graphically represent the relationship between the independent variables and the dependent variable. Box plots are created for each independent variable to visualize the differences in the mean values of the independent variables with respect to the dependent variable.

The mean values of each independent variable are indicated by variations in the horizontal position with respect to the vertically positioned dependent variable. The differences in the means of each independent variable highlight their impact on the dependent variable, as can be observed in **Figures 2 and 3.**

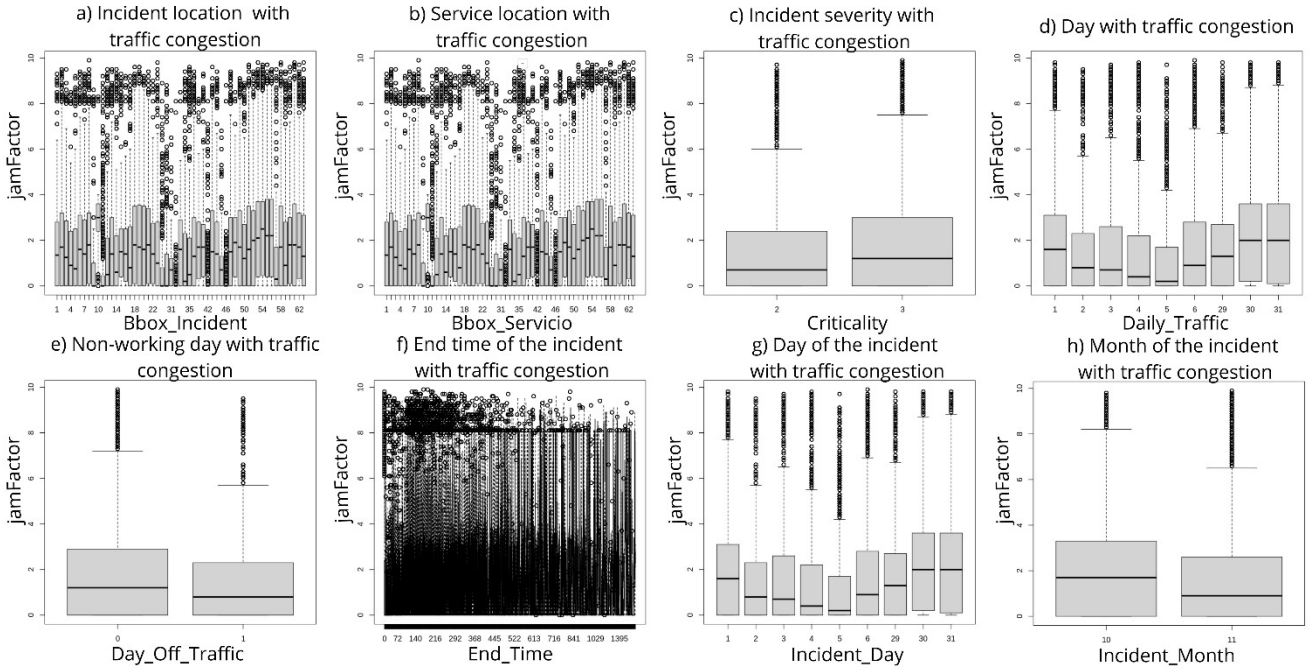

Figure 2. First part of the graphical visualization of the relationship between variables from the dataset Traffic_Incidents_Services and vehicular congestion through the mean of each variable

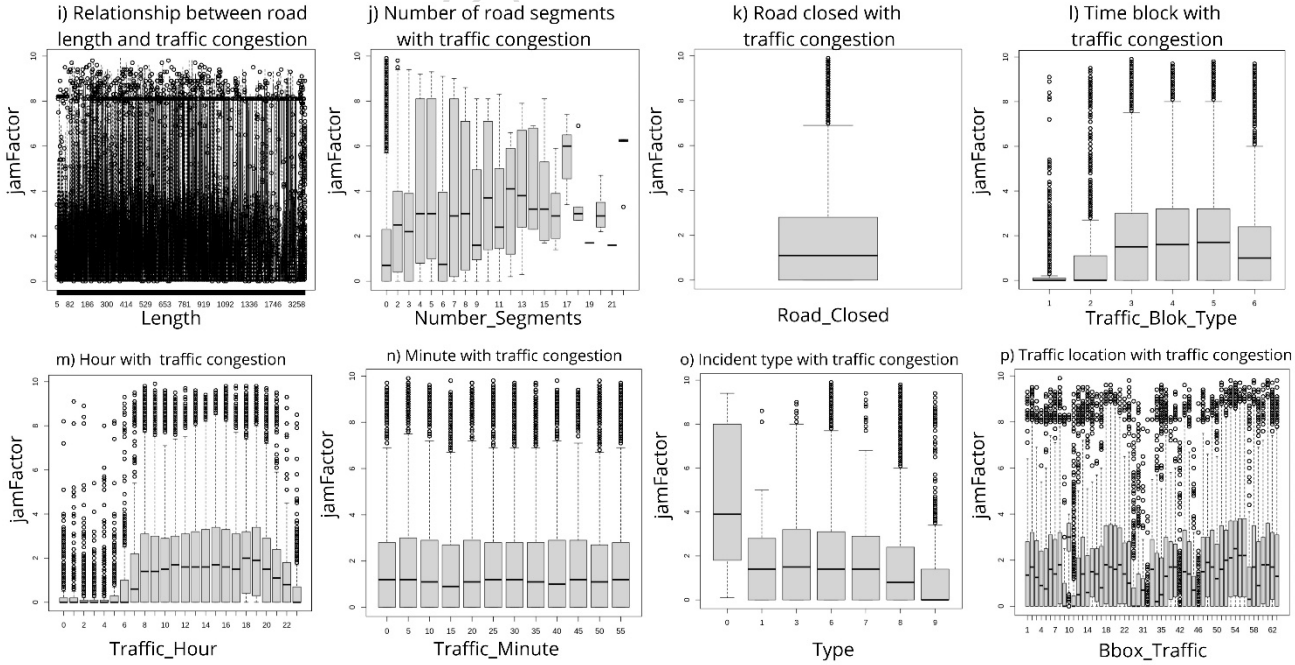

Figure 3. Second part of the graphical visualization of the relationship between variables from the dataset Traffic_Incidents_Services and vehicular congestion through the mean of each variable

The visual analysis of the means of each variable allows for the identification of the impact of independent variables on the dependent variable. A weight of 0 is assigned to those variables whose mean does not show variability, while a weight of 1 is given to those that exhibit significant variability. In **Figure 2**, image a) indicates that Bbox_Incident has a weight of 1, as its mean varies according to the location of the incident, which affects congestion. Image b) shows that Bbox_Service also receives a weight of 1 since the mean number of services influences congestion. In image c), criticality is valued with a weight of 1 due to the variation in its mean impacting congestion. Finally, image d) assigns a weight of 1 to Daily_Traffic, as its mean fluctuates significantly depending on the day of the month when traffic occurs.

The implementation of this mean-based technique enabled the identification of the most influential variables in the congestion process while also facilitating the assignment of relative weights to each variable based on their impact. Accordingly, variables that have no impact on traffic congestion were assigned a value of 0, while those that do have an impact were assigned a value of 1, as shown in **Table 4.** Therefore, the variables that exhibit the more notorious differences in their means across various points or times are considered determining factors in the onset of congestion, which gives them greater weight in predictive models.

Table 4. Variables that most influence according to exploratory statistical analysis

Number	Variable	Value	Number	Variable	Value	Number	Variable	Value
1	Bbox_Traffic	1	7	Traffic_Hour	1	13	Road_Closed	1
2	Length	1	8	Traffic_Minute	1	14	Criticality	1
3	Number_Segments	1	9	End_Time	1	15	Incident_Month	1
4	Type	1	10	Incident_Day	1	16	Bbox_Service	1
5	Daily_Traffic	1	11	Traffic_Block_Type	1			
6	Day_Off_Traffic	1	12	Bbox_Incidents	1			

3.5.2 Identification of variables influencing the onset of congestion points through Spearman correlation.

The Spearman correlation is an effective technique for detecting non-linear relationships, which are common in complex phenomena such as road congestion. In this research, it is used to identify the variables that influence the onset of traffic congestion points using data from the Traffic_Incidents_Services dataset and applying Spearman correlation. This technique is very useful to demonstrate how the values of the independent variables correlate with traffic congestion, providing a more detailed view of how each factor contributes to this phenomenon.

The identification of these relationships facilitates the assignment of relative weights to each variable based on their degree of correlation with the dependent variable (congestion). Variables that exhibit a stronger correlation, whether positive or negative, are considered key factors in the onset of congestion and are therefore assigned greater weight in predictive models.

The following steps outline how to obtain the correlation matrix between the variables, allowing for the visualization of the most significant relationships and the assignment of appropriate weights to each of them:

- Step 1: Correlation analysis between the independent variables. Correlation between the independent variables is conducted using the Spearman technique (see equation 3) due to the nature of the data, which does not follow a normal distribution behavior.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad (3)$$

Where:

n is the number of records being classified, d_i is the difference between the ranks X and Y (x_i - y_i), x_i is the rank of record i with respect to a variable and y_i is the rank of record i with respect to a second variable.

In the correlation between independent variables using Spearman, the goal is to identify significant correlations between independent variables and determine the variables with the greatest influence on the origin of traffic congestion. In **Table 5** and **Figure 5**, the pairs of variables that show a robust correlation are presented. These variables were selected based on a threshold of 0.7 or higher, according to [33].

Table 5. Correlations greater than or equal to 0.7 between the independent variables

Variable 1	Variable 2	Correlation value	Variable 1	Variable 2	Correlation value
Speed	Free_Flow	0.76523058	Incident_Day_Number	Traffic_Day_Number	0.99940959
Speed_Uncapped	Free_Flow	0.75527344	Work_Day_Incident	Work_Day_Traffic	0.99969206
Free_Flow	Speed	0.76523058	Traffic_Hour	Traffic_Block_Type	0.98022436
Speed_Uncapped	Speed	0.99645253	Daily_Traffic	Incident_Day	0.99754481
Free_Flow	Speed_Uncapped	0.75527344	block_type_incident	Incident_Hour	0.97056214
Speed	Speed_Uncapped	0.99645253	Traffic_Day_Number	Incident_Day_Number	0.99940959
Incident_Day	Daily_Traffic	0.99754481	Work_Day_Traffic	Work_Day_Incident	0.99969206
Traffic_Block_Type	Traffic_Hour	0.98022436			

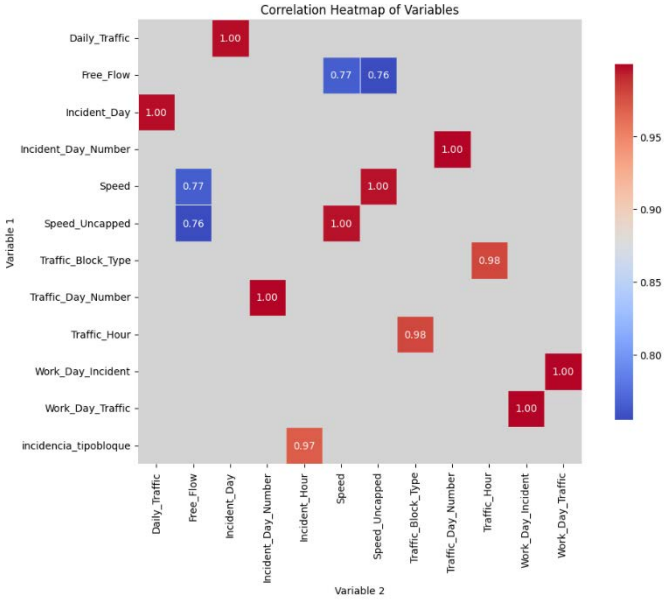

Figure 5. Heat map of correlations greater than or equal to 0.7 between the independent variables.

- Step 2: Correlation between the dependent variable jamFactor and the most significant independent variables. In this research, jamFactor is the dependent variable. A Spearman correlation analysis is conducted between jamFactor and the relevant independent variables described in **Table 5**. The highest correlation values between jamFactor and the significant independent variables are shown in **Table 6**.

Table 6. Correlation of the variable jamFactor with the set of variables in Table 5

Traffic congestion	Variables	Correlation value	Traffic congestion	Variables	Correlation value
jamFactor	Number_Segments	0.2687642	jamFactor	Speed	-0.7261867
jamFactor	Speed_Uncapped	-0.7378836	jamFactor	Traffic_Hour	0.1530089
jamFactor	Work_Day_Traffic	0.1270429	jamFactor	Traffic_Day_Number	-0.1548273
jamFactor	Description	-0.1250564	jamFactor	Traffic_Block_Type	0.1534324
jamFactor	Work_Day_Incident	0.1271877	jamFactor	Type	-0.1271393
jamFactor	Free_Flow	-0.2272753	jamFactor	End_Time	0.1222923
jamFactor	Length	-0.1526781	jamFactor	Incident_Day_Number	-0.1549883

The analysis conducted using the Spearman technique to identify significant causal variables for traffic congestion points is summarized in **Table 7**, which presents the corresponding weight of each variable based on its correlation value. The variables Speed_Uncapped and Speed stand out with correlations of -0.7378836 and -0.7261867, respectively, suggesting that an increase in speed is associated with a decrease in congestion.

Table 6 identifies other variables that also significantly contribute to the onset of congestion. In this analysis, a threshold of 0.7 was established to select the most influential variables. This implies that, although some variables exhibit less pronounced correlations, their impact on the phenomenon of congestion should not be underestimated, as they can provide valuable insights for traffic analysis.

Table 7 Variables of highest significance with the Spearman technique

Number	Variable	Value
1	Speed_Uncapped	0.7378836
2	Speed	0.7261867

3.5.3 Identification of variables influencing the onset of congestion points through the rotation matrix of Principal Component Analysis

The rotation matrix of the Principal Component Analysis (PCA) was used to identify the relationships between the original variables and the principal components. This rotation allows for the identification of relevant variables that contribute to each component, providing a clear understanding of the underlying structure of the data. The values of the rotation matrix are used to assign weights to the variables causing traffic congestion based on their contribution to the components. The variables with higher loadings on the principal components are those that most influence the onset of congestion and are therefore given greater weight in predictive models.

The following steps describe the process to obtain the rotation matrix, which will allow for the visualization of the most significant relationships between the variables and the components, as well as the assignment of appropriate weights to each variable:

- Step 1: Calculation of the covariance matrix. The covariance matrix is computed Σ of the TrafficIncidentsServices dataset (X) using equation 4.

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1} \quad (4)$$

Where:

n is the total number of records, x_i *and* y_i are the individual values of the variables X and Y, respectively. $\bar{x}$ *and* $\bar{y}$ are the means of the variables X and Y, respectively.

- Step 2: Calculation of eigenvectors and eigenvalues. With the covariance matrix Σ in the TrafficIncidentsServices dataset, the calculation of eigenvectors is performed to represent the direction of maximum variance in the data, and the calculation of eigenvalues is performed to represent the magnitude of variances in each direction. This is done by using equation 5.

$$\sum v = \lambda v \quad (5)$$

Where:

v is the eigenvector and λ is the corresponding eigenvalue.

- Step 3: Normalization of the principal components. Once the eigenvectors and eigenvalues are obtained, the principal components matrix is normalized to have a mean of 0 and standard deviation of 1 (σX = 1), using equation 6.

$$X_{norm} = \frac{X\bar{X}}{\sigma X} \quad (6)$$

Where:

σX is the standard deviation of the dataset, X is the dataset and $\bar{X}$ is the mean of the dataset.

- Step 4: Component rotation. After normalizing the principal components, component rotation is performed to simplify the structure of the covariance matrix, using Varimax rotation in this research. The resulting rotation matrix is presented in **Table 8**, and this process is carried out using equation 7.

$$R' = R \cdot Q \quad (7)$$

Where:

R' is the updated rotation matrix after the iteration, R is the current rotation matrix and Q is an incremental rotation matrix calculated during each iteration of the Varimax method using equation 8.

$$Q = \left(\frac{1}{n}\right)\sum_{j=1}^{m}\left(\left(\frac{1}{4}\right)\sum_{i=1}^{p}\left(\frac{\partial}{\partial\psi_{ij}}\left(C_{ij}^2\right)\right)^2 - \frac{1}{2}\sum_{i=1}^{p}\left(\frac{\partial^2}{\partial\psi_{ij}^2}\left(C_{ij}^2\right)\right)\right) \quad (8)$$

Where:

n is the total number of records, m is the number of principal components, p is the number of original variables, ψ_{ij} represents the elements of the rotation matrix R, C_{ij} are the rotated factor loadings, $\frac{\partial}{\partial\psi_{ij}}$ denotes the partial derivative with respect to ψ_{ij}, and $\frac{\partial^2}{\partial\psi_{ij}^2}$ denotes the second partial derivative with respect to ψ_{ij}.

The rotation matrix is presented in **Table 8**, where the factor loadings indicate the relationship between the variables and the principal component. A threshold of 0.7 is established to identify the significant variables, highlighted in orange, which show positive and negative correlations.

Table 8. Analysis of the rotation matrix of the variables with the greatest impact on each component

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17
Length	-0.00002	-0.99992	0.01167	-0.00360	-0.00076	-0.00025	0.00009	0.00003	0.00001	-0.00005	-0.00068	0.00015	0.00001	-0.00012	0.00007	0.00003	-0.00010
Number_Segments	0.00000	-0.00010	-0.00002	-0.00031	0.00023	-0.00152	-0.00253	0.00062	0.00054	0.00438	-0.04180	-0.00203	-0.00577	0.19977	0.00056	-0.09299	0.95679
Traffic_Hour	0.00000	-0.00003	0.00009	-0.00085	0.00047	-0.00638	0.00127	0.00855	-0.00916	0.04251	-0.15049	-0.93133	-0.20778	-0.06298	-0.00900	0.05458	0.00932
Traffic_Minute	0.00000	-0.00002	-0.00004	0.00009	0.00060	0.00124	0.00376	-0.73840	0.67420	0.00610	-0.00093	-0.01193	-0.00422	0.00008	-0.00032	0.00104	0.00005
Type	-0.00005	-0.00083	0.04864	0.56526	-0.82257	-0.03755	-0.00708	-0.00152	-0.00091	-0.00092	-0.00186	-0.00090	0.00191	0.00163	-0.00045	0.00093	-0.00006
Start_Time	0.00006	-0.01208	-0.98888	0.14275	0.03979	-0.00320	-0.00136	-0.00005	-0.00016	-0.00014	-0.00010	-0.00030	0.00057	0.00014	0.00000	0.00020	-0.00001
End_Time	0.00006	0.00172	-0.14005	-0.81199	-0.56653	0.00673	-0.00672	-0.00071	-0.00015	0.00056	0.00083	0.00096	-0.00309	-0.00074	0.00023	-0.00080	-0.00005
Incident_Hour	0.00000	0.00000	0.00006	-0.00388	-0.00011	-0.00105	0.00024	0.00507	0.00754	0.01867	-0.01961	-0.20666	0.94699	0.02835	-0.02181	0.03637	0.00146
Incident_Minute	0.00000	-0.00002	0.00015	-0.00058	0.00179	-0.00132	-0.00264	-0.67427	-0.73841	-0.00192	-0.00006	-0.00098	0.00910	0.00197	-0.00286	0.00250	0.00068
Category_Number	0.00000	0.00006	0.00006	-0.00032	0.00000	-0.00051	0.00291	-0.00078	-0.00099	0.00086	0.00334	0.00463	0.00183	0.14592	0.90076	0.40800	0.01327
Bbox_Traffic	0.00148	0.00034	-0.00037	0.00008	-0.00032	-0.00044	0.00003	-0.00009	0.00001	0.00024	-0.00001	-0.00001	-0.81635	0.01520	0.00150	0.00032	0.00057
jamFactor	-0.09510	-0.73353	0.02580	0.00417	0.09056	0.00890	-0.00563	-0.00532	0.00088	0.01227	0.00053	0.00060	0.00000	0.00000	0.00000	0.00000	0.00000
Speed	-0.00712	-0.14496	0.00707	-0.01994	0.75779	0.03263	-0.00639	-0.00271	0.00442	0.01030	0.00088	-0.00027	0.00000	0.00000	0.00000	0.00000	0.00000
Speed_Uncapped	-0.04193	-0.34282	0.02456	0.02168	-0.63776	-0.02759	0.00685	0.00178	-0.00385	0.00091	-0.00065	-0.00013	0.00000	0.00000	0.00000	0.00000	0.00000
Daily_Traffic	0.00561	-0.00455	0.00658	-0.70641	-0.02046	0.01042	0.00469	0.00031	0.00359	0.00156	0.02450	-0.00307	0.00001	0.00001	0.00345	0.00111	-0.00143
Road_Closed	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00006	-0.00892	0.31672	-0.94796	0.03129
Criticality	-0.05854	0.00273	0.23961	-0.00139	0.01336	-0.07583	0.00947	-0.00347	0.06862	-0.96352	-0.00169	-0.00494	0.00000	0.00000	0.00000	0.00000	0.00000
Description	0.36379	-0.09337	-0.89349	-0.00621	-0.00336	-0.05020	0.00933	0.00037	0.02390	-0.23888	0.00002	-0.00052	0.00000	0.00000	0.00000	0.00000	0.00000
Olr	0.91952	-0.10097	0.37589	0.00849	-0.00435	-0.02191	-0.00010	-0.00062	-0.00116	0.03895	0.00012	0.00216	0.00000	0.00000	0.00000	0.00000	0.00000
Incident_Day	-0.00109	0.00859	-0.00585	0.70565	0.02059	-0.00765	0.00049	0.00229	-0.00254	-0.00178	-0.01678	0.03978	-0.00001	-0.00001	-0.00345	-0.00111	0.00143
Incident_Month	-0.00348	0.00061	0.00054	-0.02456	0.00070	-0.05115	-0.07205	-0.03823	-0.01599	-0.00265	0.20252	0.97373	0.00000	0.00000	0.00000	0.00000	0.00000
block_type_incident	0.00473	-0.00326	-0.00363	-0.00436	0.00200	-0.02731	-0.43159	0.80775	-0.32058	-0.02881	-0.00013	-0.00663	0.00000	0.00000	0.00000	0.00000	0.00000
Bbox_Service	0.00148	0.00034	-0.00037	0.00008	-0.00032	-0.00044	0.00003	-0.00009	0.00001	0.00024	-0.00001	-0.00001	0.39500	-0.71452	-0.00444	0.00501	-0.00773

Using the rotation matrix, critical variables are identified, each with weights indicating their influence on traffic behavior, as shown in **Table 9**. Regarding relevant variables, Length stands out with a weight of 0.99992, suggesting that the length of road segments is nearly determinant in congestion. On the other hand, the Criticality of incidents, with a weight of 0.96352, indicates that they cause greater delays. The variables Traffic_Hour (weight 0.93133) and Incident_Hour (weight 0.94699) suggest that both the time of day and the occurrence of incidents during those hours significantly contribute to congestion.

Furthermore, Incident_Month (weight 0.97373) and Start_Time (weight 0.98888) indicate that certain times of the year may present a higher number of incidents, which in turn leads to more congestion. The variable Traffic_Minute (weight 0.73840) also suggests that the specific minute of traffic can influence vehicle flow. Finally, the Category_Number (weight 0.90076) and Type (weight 0.82257) of incidents reveal that certain types of events generate more disruptions in traffic.

Table 9. Most significant variables using Principal Component Analysis rotation matrix

Number	Variable	Value	Number	Variable	Value	Number	Variable	Value

1	Length	0.99992	9	Olr	0.91952	17	Speed	0.75779
2	Start_Time	0.98888	10	Category_Number	0.90076	18	Incident_Minute	0.73841
3	Incident_Month	0.97373	11	Description	0.89349	19	Traffic_Minute	0.7384
4	Criticality	0.96352	12	Traffic_Block_Type	0.88903	20	jamFactor	0.73353
5	Number_Segments	0.95679	13	Type	0.82257	21	Bbox_Service	0.71452
6	Road_Closed	0.94796	14	Bbox_Traffic	0.81635	22	Daily_Traffic	0.70641
7	Incident_Hour	0.94699	15	End_Time	0.81199	23	Incident_Day	0.70565
8	Traffic_Hour	0.93133	16	block_type_incident	0.80775			

3.5.4 Identifying the variables that influence the origin of congestion points through the factorial load matrix of the Structural Equation Modeling

The factor loading matrix in the Structural Equation Model (SEM) aims to represent the relationship between each observed variable and its corresponding factor. In this matrix, each row is associated with a variable, while each column represents a factor. The values within the matrix, known as factor loadings, reflect the strength and direction of the relationship between the variable and its associated factor. The higher the value of a factor loading, the greater the influence of the variable on the factor.

The analysis of this matrix in the SEM enables the identification of the most significant variables that contribute to the formation of traffic congestion points. This helps to determine which variables have the greatest impact on the latent factors explaining congestion, therefore providing a detailed view of the underlying relationships affecting traffic.

The factor loading matrix also permits to assign relative weights to each variable according to its influence on the latent factors. Variables with higher factor loadings play a key role in explaining congestion and, therefore, receive greater weight in predictive models. The steps to obtain the factor loading matrix in the SEM are outlined below, enabling the visualization of the most significant relationships and assigning appropriate weights to the key variables:

- Step 1: Definition of factors. For the traffic congestion problem, factors are defined as traffic, traffic incidents, and services.
- Step 2: Formulation of equations to measure traffic, traffic incidents, and services factors. Each factor equation is constructed using a set of variables for traffic, incidents, and services. The resulting equations are displayed as 9, 10, and 11, respectively.

$$Traffic = \sim Free\_Flow + jamFactor + Length + Speed + Speed\_Uncapped + Traffic\_Hour + Traffic\_Minute + Traffic\_Day\_Number + Traffic\_Block\_Type \quad (9)$$

$$Incidents = \sim Bbox\_Incident + Type + Start\_Time + End\_Time + Incident\_Day + Incident\_Hour + Incident\_Minute + Incident\_Day\_Number + Traffic\_Block\_Type + Category\_Number \quad (10)$$

$$Services = Bbox\_Servicio \quad (11)$$

- Step 3: Establishing regression among the factors. Regression modeling is conducted among the factors that influence the dependent variable (jamFactor). As a result of this process, equation 12 is obtained, representing the regression among the traffic, incidents, and services variables.

$$jamFactor \sim Traffic + Incidents + Services \quad (12)$$

- Step 4: Establishing correlation equations among the measurable variables. Equations representing the correlations between the variables are formulated, as illustrated in equations 13, 14, 15, and 16.

$$Speed \sim\sim Traffic\_Hour + Traffic\_Minute + Traffic\_Day\_Number \quad (13)$$

$$Free\_Flow \sim\sim Length \quad (14)$$

$$jamFactor \sim\sim Traffic\_Hour + Traffic\_Minute + Traffic\_Day\_Number \quad (15)$$

$$jamFactor \sim\sim Speed \quad (16)$$

- Step 5: Obtaining the factor loadings matrix. To calculate the factor loadings, equations 9 to 16 are used in the Lavaan library of the R software, facilitating the generation of the matrix for the Structural Equation Model. These loadings reflect the relationship between observed variables and latent constructs. The resulting matrix, presented in **Table 10**, provides an overview of these relationships, highlighting in yellow the most significant variables, specifically those with a value equal to or greater than 0.7.

Table 10. Factor loading matrix of the Structural Equation Model

Latent Variable	Operator	Variable	Value	Latent Variable	Operator	Variable	Value
Traffic	=~	Free_Flow	0.5002061	Incidents	=~	Type	-0.23748091
Traffic	=~	jamFactor	0	Incidents	=~	Start_Time	1.00296711
Traffic	=~	Length	1.00004338	Incidents	=~	End_Time	0.56983903
Traffic	=~	Speed	0.34515264	Incidents	=~	Incident_Day	0.01422786
Traffic	=~	Speed_Uncapped	0.33619433	Incidents	=~	Incident_Hour	-0.01648256
Traffic	=~	Traffic_Hour	0.03514468	Incidents	=~	Incident_Minute	-0.01805015
Traffic	=~	Traffic_Minute	0.01307886	Incidents	=~	Incident_Day_Number	-0.00035456
Traffic	=~	Traffic_Day_Number	0.00194125	Incidents	=~	block_type_incident	-0.02422802
Traffic	=~	Traffic_Block_Type	0.03676568	Incidents	=~	Category_Number	-0.03245275
Incidents	=~	Bbox_Incidents	0.06720955	Services	=~	Bbox_Service	1

The results in **Table 11** highlight the key variables in the formation of congestion points, identified through the factor loading matrix of the structural equation model. The variable Start_Time, with the highest factor loading (1.00296711), indicates that trips beginning during peak hours experience higher levels of congestion. The variable Length, with a loading of 1.00004338, shows that longer routes tend to accumulate more vehicles, creating traffic bottlenecks. Finally, Bbox_Service, with a loading of 1, suggests that services related to road segments, such as stops or commercial access points, affect vehicle flow and contribute to congestion.

Table 11. Relevant variables from the factorial load matrix of the Structural Equation Modeling

Number	Variable	Value
1	Start_Time	1.00296711
2	Length	1.00004338
3	Bbox_Service	1

3.5.5 Identification of the variables influencing the origin of congestion points using the factor loadings matrix of Factor Analysis.

The factor loading matrix in Factor Analysis shows the relationships between the observed variables and the extracted factors. Each entry in the matrix represents a variable, while each column corresponds to a latent factor. The values in the matrix, known as factor loadings, indicate both the strength and direction

of the relationship between each variable and the underlying factor. When interpreting this matrix, high values are sought, as they reflect a strong relationship between the variable and the corresponding factor. The purpose of using the factor loading matrix in Factor Analysis is to identify relationships and assign relative weights to the variables based on their importance within the latent factors. Variables with higher factor loadings are those that have a significant influence on the factors explaining the origin of congestion, and thus receive a greater weight in predictive models. The following outlines the steps taken to obtain the factor loading matrix from Factor Analysis, facilitating the visualization of the most important relationships and the assignment of appropriate weights to the key variables.

- Step 1: Obtaining the correlation matrix. Calculating the correlations between all independent variables p of the dataset TrafficIncidentsServices using equation 17.

$$R = \frac{1}{n-1}\left(X - \overline{X}\right)^T\left(X - \overline{X}\right) \quad (17)$$

  Where: n is the number of records, X is the dataset of dimensions and $\overline{X}$ is the mean vector of the variables, calculated for each column of X.

- Step 2: Calculation of eigenvectors and eigenvalues. Using the correlation matrix R of the TrafficIncidentsServices dataset, the calculation of eigenvectors and eigenvalues is performed using equation 18.

$$Rv = \lambda v \quad (18)$$

  Where: R is the correlation matrix, v is the eigenvector associated with the eigenvalue and λ are the eigenvalues of the correlation matrix.

- Step 3: Determination of the appropriate number of factors. To identify the suitable number of factors for factor analysis, the Kaiser criterion is employed, using equation 19.

$$Number\ of\ factors = Number\ of\ eigenvalues > 1 \quad (19)$$

- Step 4: Obtaining the factor loadings. The ***fa*** function from the psyc package in R is used to perform the factor analysis on the TrafficIncidentsServices dataset. The results, displayed in **Table 12**, show columns MR1 to MR6 as the extracted factors, while the rows indicate the factor loadings of each variable. A threshold of 0.7 is set to highlight significant loadings in yellow.

Table 12. Factor loadings of each variable obtained from the factor analysis

Variable	MR1	MR2	MR3	MR4	MR5	MR6	Variable	MR1	MR2	MR3	MR4	MR5	MR6
Bbox_Traffic	-0.44	0.9	-0.03	0.03	0.03	-0.04	Type	0.31	0.11	-0.54	-0.14	0.02	-0.29
Free_Flow	0.57	0.26	0.31	0.08	0.18	-0.06	Olr	0.12	-0.03	-0.29	-0.11	0	-0.2
jamFactor	-0.58	-0.25	-0.14	-0.02	0.04	-0.02	Start_Time	-0.23	-0.03	0.49	0.18	-0.01	0.38
Length	0.38	0.02	0.19	0.06	0.11	-0.04	End_Time	-0.38	-0.1	0.75	0.25	-0.04	0.51
Speed	0.86	0.38	0.35	0.09	0.12	-0.03	Incident_Day	-0.09	-0.04	0.13	-0.08	0.08	-0.05
Speed_Uncapped	0.86	0.38	0.35	0.09	0.12	-0.03	Incident_Hour	-0.08	-0.07	0.16	0.73	-0.47	-0.44
Traffic_Hour	-0.16	-0.13	-0.17	0.56	0.77	-0.06	Incident_Minute	0.01	0	-0.02	-0.02	0.02	0.03
Traffic_Minute	0.01	0.01	0	-0.01	-0.04	0.01	Incident_Day_Number	0.32	0.15	-0.59	0.43	-0.24	0.53
Traffic_Day_Number	0.32	0.15	-0.59	0.43	-0.24	0.53	block_type_incident	-0.08	-0.07	0.15	0.72	-0.46	-0.43
Traffic_Block_Type	-0.16	-0.13	-0.16	0.56	0.77	-0.06	Category_Number	-0.07	0.02	-0.01	-0.01	-0.02	0.02
Bbox_Incidents	-0.44	0.9	-0.03	0.03	0.03	-0.04	Bbox_Service	-0.44	0.9	-0.03	0.03	0.03	-0.04

The results presented in **Table 13** reflect the analysis of the factor loading matrix from the factor analysis model, aimed at identifying significant causal variables in the formation of traffic congestion points. Relevant variables in this stage are Bbox_Traffic, Bbox_Incidents, and

Bbox_Service, each with a loading value of 0.9, suggesting that traffic congestion is associated with geographic areas where incidents occur and where services of interest are located. Similarly, the variables related to speed, Speed and Speed_Uncapped, with values of 0.86, indicate that traffic speed is a relevant factor in congestion. Reduced speed can be influenced by speed limits as well as road conditions. Other variables, such as Traffic_Hour (0.77) and End_Time (0.75), also show a significant correlation with congestion, suggesting that peak traffic periods coincide with the overlap of various activities. Finally, the variables Incident_Hour (0.73) and block_type_incident(0.72) underscore the importance of the timing of traffic incidents and their impact on the generation of congestion.

Table 13. Most significant variables using Factor Analysis

Number	Variable	Value	Number	Variable	Value
1	Bbox_Traffic	0.9	6	Traffic_Hour	0.77
2	Bbox_Incidents	0.9	7	Traffic_Block_Type	0.77
3	Bbox_Service	0.9	8	End_Time	0.75
4	Speed	0.86	9	Incident_Hour	0.73
5	Speed_Uncapped	0.86	10	block_type_incident	0.72

4. Results

The methodology for identifying the variables that influence the origin of traffic congestion points reveals that each mathematical analysis technique offers unique perspectives on the involved variables, although they all converge on key aspects such as speed, street length, incident start time, service location, and occurrence times. A correlation and four significant relationships were identified: the correlation between Speed_Uncapped and Speed suggests that traffic congestion can be calculated using the relationship Speed_Uncapped / Speed, where lower values of Speed_Uncapped indicate greater congestion. The first relationship involves Bbox_Traffic, Speed_Uncapped, Length, and Start_Time, which describe the geographical area of traffic, vehicle speed, street length, and the time of the incident. The second relationship highlights Length, Start_Time, Incident_Month, and Criticality, indicating that street length is associated with the onset of incidents, with shorter streets being more prone to them. The third relationship includes Start_Time, Length, and Bbox_Service, showing that street length and the proximity of services are related to the start time of incidents. Finally, the fourth relationship between Bbox_Traffic, Bbox_Incidents, and Bbox_Service suggests that the combination of these variables indicates a high probability of traffic congestion.

All the analysis techniques used in this research work have enabled the identification of specific variables that are relevant to characterize the generation of traffic congestion points in city streets. This understanding can assist urban planners and traffic managers in making data-based decisions. Several strategies based on the research results could be considered. First, it is essential to prioritize the variables that have shown the greatest impact on congestion, such as traffic speed, day, minute, and hour of traffic. Based on this prioritization, awareness campaigns and educational programs aimed at the population could be developed, encouraging the use of public transport and sustainable mobility alternatives, such as biking and walking.

Furthermore, the implementation of real-time traffic management systems utilizing advanced technologies, such as sensors and data analysis, could be proposed to monitor vehicle flow and dynamically adjust traffic lights. This would allow for an agile response to traffic fluctuations, improving circulation during peak hours.

Additionally, the creation of adequate infrastructure, such as lane expansions at critical points and the implementation of exclusive lanes for public transport, could be an alternative to alleviate congestion.

The planning of this infrastructure should be based on continuous analysis of the areas with the highest congestion, using the data obtained from the research to guide decision-making.

5. Conclusions and discussion

The analysis of the variables causing traffic congestion points in Mexico City has yielded results that highlight both the relevance of these variables and their quantitative weight in generating congestion. Techniques of different types were used to assign these weights, each with advantages and disadvantages that should be considered. Spearman correlation, for example, is a valuable tool for identifying the relationship between individual variables and congestion, assigning weights based on the magnitude of their correlation. However, being a bivariate analysis, it does not capture the interaction among several variables simultaneously, which limits the depth of the analysis. On the other hand, Principal Component Analysis (PCA) reduces data complexity by grouping variables into principal components, allowing them to be assigned weights based on the variability explained by each. Nevertheless, the difficulty in interpreting these linear components can complicate their practical application, as they do not always reflect clear causes.

Factor Analysis, similar to PCA, allows for grouping variables into latent factors and assigning them factor loadings that represent their weight within these groups. This is useful for identifying sets of variables with greater influence and reducing information redundancy. However, like PCA, it lacks a clear causal structure. The most robust technique for this purpose is Structural Equation Modeling (SEM), which not only allows identifying the most influential variables but also assigns them quantitative weights based on causal relationships among them. This is particularly valuable in situations where multiple factors interact in complex ways. Although it is more complex and requires greater technical rigor, SEM offers a more faithful representation of the phenomenon of congestion by modeling these interactions and assigning weights based on the relationships between the variables.

Each of these techniques provides a unique perspective in the analysis of congestion, from the simplicity and speed of Spearman correlation to the sophistication and precision of SEM. Exploratory techniques, such as PCA and Factor Analysis, are useful for identifying general patterns and reducing data complexity, while SEM provides a solid causal structure that assigns weights based on the interrelationship of the variables. The correct combination of these techniques allowed for the assignment of quantitative values that facilitate a more accurate interpretation of each variable's contribution to traffic congestion. This not only enhances our understanding of the factors influencing congestion but also provides a solid foundation for strategic decision-making and the implementation of corrective measures based on solid data.

Declaration of competing interest

We declare that we have no significant competing interests including financial or non-financial, professional, or personal interests interfering with the full and objective presentation of the work described in this manuscript.

Acknowledgment
We want to thank TecNM/CENIDET and CONAHCYT for their invaluable support, which was essential to make this research project possible.

Funding
The authors received no financial support for the research, authorship, and/or publication of this article.

Author contributions
Built the design, worked on the data collection, data preprocessing, data analysis, methodology and wrote the article: E. de la Cruz-Nicolás; contributed with the original idea, worked on the conceptualization, methodology, and carried the supervision of the experimental stage and research guide: H. Estrada-Esquivel; contributed with the original idea, directed and gave the necessary guidelines to efficiently prepare this article: A. Martínez-Rebollar; Contributed with the analysis and interpretation of data, review of the mathematical expressions used in the research, and reviewed of the results obtained in the experiments: E. Clemente; contribute with the digital processing of data and data visualization: O. Pliego-Martínez

Data availability statement
The traffic and road incident data analyzed in this document was obtained from HERE Maps through their API (Application Programming Interface), available at https://developer.here.com/ . Additionally, the service data was obtained through the Geoapify API, which is available at https://www.geoapify.com.

References
[1] Kardani-Yazd, N., Kardani-Yazd, N., and Mansouri Daneshvar, M. R. "A rapid method for evaluating the variables affecting traffic flow in a touristic road, Iran". Environmental Systems Research., vol. 8, no. 34, 2019. [Online]. Available: https://doi.org/10.1186/s40068-019-0162-0
[2] Afrin, T., and Yodo, N. "A survey of road traffic congestion measures towards a sustainable and resilient transportation system". Sustainability, vol. 12, no. 11, 4660, 2020. [Online]. Available: https://doi.org/10.3390/su12114660
[3] Jia, X. "Analysis on influencing factors of traffic congestion in Chongqing and study on countermeasures: Empirical analysis based on principal component analysis". Atlantis Press International BV, pp. 814–822, 2023. [Online]. Available: https://doi.org/10.2991/978-94-6463-200-2_84
[4] Iro, S., and Pat-Mbano, E. C. "Causes of traffic congestion; A study of owerri municipal area of IMO state". American Journal of Environmental Sciences, vol. 18, no. 3, pp. 52–60, 2022. [Online]. Available: https://doi.org/10.3844/ajessp.2022.52.60
[5] Chen, L., Shi, J., Cheng, M., Zhu, H., and Sun, L. "Characteristics of urban road non-recurrent traffic congestion based on floating car data", in *4th International Conference on Electronic Information Technology and Computer Engineering.* 2020, pp. 120-126. [Online]. Available: https://doi.org/10.1145/3443467.3443740
[6] Bian, C., Yuan, C., Kuang, W., and Wu, D. "Evaluation, classification, and influential factors analysis of traffic congestion in Chinese cities using the online map data". Mathematical Problems in Engineering, pp. 1–10, 2016. [Online]. Available: https://doi.org/10.1155/2016/1693729

[7] Mahona, J., Mhilu, C., Kihedu, J., and Bwire, H. "Factors contributing to traffic flow congestion in heterogenous traffic conditions". International Journal for Traffic and Transport Engineering, vol. 9, no. 2, pp. 238–254, 2019. [Online]. Available: https://doi.org/10.7708/ijtte.2019.9(2).09

[8] Yu, J., Wang, L., and Gong, X. "Study on the status evaluation of urban road intersections traffic congestion base on AHP-TOPSIS modal". Procedia, Social and Behavioral Sciences, vol. 96, pp. 609–616, 2013. [Online]. Available: https://doi.org/10.1016/j.sbspro.2013.08.071

[9] Gullotta, G., Loret, E., Stewart, C., and Sarti, F. "Traffic attractors and congestion in the urban context, the case of the city of Rome". Journal of Geographic Information System, vol. 12, no. 06, pp. 545–559, 2020. [Online]. Available: https://doi.org/10.4236/jgis.2020.126032

[10] Rahman, M. M., Najaf, P., Fields, M. G., and Thill, J.-C. "Traffic congestion and its urban scale factors: Empirical evidence from American urban areas". International Journal of Sustainable Transportation, vol. 16, no. 5, pp. 406–421, 2022. [Online]. Available: https://doi.org/10.1080/15568318.2021.1885085

[11] Pi, M., Yeon, H., Son, H., and Jang, Y. "Visual Cause Analytics for Traffic Congestion". IEEE Transactions on Visualization and Computer Graphics, vol. 27, no. 3, pp. 2186–2201, 2021. [Online]. Available: https://doi.org/10.1109/tvcg.2019.2940580

[12] Yue, W., Li, C., Chen, Y., Duan, P., and Mao, G. "What is the root cause of congestion in urban traffic networks: Road infrastructure or signal control?". IEEE Transactions on Intelligent Transportation Systems: A Publication of the IEEE Intelligent Transportation Systems Council, vol. 23, no. 7, pp. 8662–8679, 2022. [Online]. Available: https://doi.org/10.1109/tits.2021.3085021

[13] Hüseyin, G., Youssef, K., Gazi, T. "Evaluation of Traffic Congestion and Level of Service at Major Intersections in Lefkoşa, Northern Cyprus". International Journal of Innovative Technology and Exploring Engineering, vol. 8, no. 8, pp. 2150-2155, 2019.

[14] Mohmad, A., Sukhdeep, S." Congestion modelling and level of service assesment of urban roads in developing countries". Journal of Emerging Technologies and Innovative Research, vol. 7, no. 8, pp. 2230- 2240, 2020.

[15] Ashhad Verdezoto, T. Z., Cabrera, F., and Roa Medina, O. B. "Analysis of vehicular congestion for the improvement of the main road in Guayaquil-Ecuador." Gaceta Técnica, vol. 21, no. 2, pp. 4–23, 2020

[16] Xu, D., Wang, Y., Peng, P., Lin, L., and Liu, Y." The evaluation of the urban road network based on the complex network". IEEE Intelligent Transportation Systems Magazine, vol. 14, no. 3, pp. 200–211, 2022. [Online]. Available: https://doi.org/10.1109/mits.2021.3049351

[17] Badveeti, A., Mir, M. S., and Badweeti, K. "The evaluation of traffic congestion analysis for the Srinagar city under mixed traffic conditions". In: Recent Advances in Traffic Engineering, vol. 69, pp. 85–98, 2020. [Online]. Available: https://doi.org/10.1007/978-981-15-3742-4_6

[18] Lin, P., Weng, J., Yin, B., and Zhou, X. "Urban road network operation quality evaluation method based on high-frequency trajectory data," in *21st International Conference on Intelligent Transportation Systems (ITSC)*, Maui, HI, USA, 2021, pp. 3602-3607.

[19] Aarón, M. A., Gómez, C. A., Fontalvo, J., and Gómez, A. J. "Analysis of Vehicle Mobility in the Department of La Guajira using Simulation: The Case of Riohacha and Maicao." CIT Información Tecnológica, vol. 30, no. 1, pp. 321–332, 2019. [Online]. Available: https://doi.org/10.4067/s0718-07642019000100321

[20] Vázquez, R., Torres, C., and Mariguetti, J. O. "Data Fusion and Analysis for Decision Making in a Complex Vehicle Scenario." Extensionismo, Innovación y Transferencia Tecnológica, vol. 7, no. 1, pp. 196-205, 2021. [Online]. Available: https://doi.org/10.30972/eitt.704777

[21] Zhang, K., Chu, Z., Xing, J., Zhang, H., and Cheng, Q. "Urban traffic flow congestion prediction based on a data-driven model". Mathematics, vol. 11, no. 19, 2023. [Online]. Available: https://doi.org/10.3390/math11194075
[22] Tu, Y., Lin, S., Qiao, J., and Liu, B. "Deep traffic congestion prediction model based on road segment grouping". Applied Intelligence, vol. 51, no. 11, pp. 8519–8541, 2021. [Online]. Available: https://doi.org/10.1007/s10489-020-02152-x
[23] Cvetek, D., Muštra, M., Jelušić, N., and Tišljarić, L. "A survey of methods and technologies for congestion estimation based on multisource data fusion". Applied Sciences (Basel, Switzerland), vol. 11, no. 5, pp. 1-19, 2021. [Online]. Available: https://doi.org/10.3390/app11052306
[24] Diaz, C., Beltran, K., Diaz, C., and Baena, A. "Traffic flow indicators analysis to determine causes of vehicular congestion". ParadigmPlus, vol. 2, no. 2, pp. 1–16, 2021. [Online]. Available: https://doi.org/10.55969/paradigmplus.v2n2a1
[25] Wang, W.-X., Guo, R.-J., and Yu, J. "Research on road traffic congestion index based on comprehensive parameters: Taking Dalian city as an example". Advances in Mechanical Engineering, vol. 10, no. 6, pp. 1-8, 2018. [Online]. Available: https://doi.org/10.1177/1687814018781482
[26] Nguyen, D.-B., Dow, C.-R., and Hwang, S.-F. "An efficient traffic congestion monitoring system on Internet of Vehicles". Wireless Communications and Mobile Computing, pp.1–17, 2018. [Online]. Available: https://doi.org/10.1155/2018/9136813
[27] B. Pishue, "2022 INRIX Global Traffic Scorecard," INRIX., Kirkland, USA, Tech. Rep.-, January. 2023.
[28] TomTom Traffic Index, Mexico City Traffic Report, 2024. [Online]. Available: https://www.tomtom.com/traffic-index/mexico-city-traffic/. Accessed on: Jan. 29, 2024.
[29]Here Maps, Here technologies documentation, 2024. [Online]. Available: https://www.here.com/docs/bundle/traffic-api-developer-guide-v6/page/topics/what-is.html. Accessed on: Jan. 29, 2024.
[30] Here Maps. Developer Guide Routing, 2024. [Online]. Available: https://www.here.com/docs/bundle/sdk-for-ios-navigate-developer-guide/page/topics/routing.html. Accessed on: Jan. 30, 2024.
[31] Here Maps. Incidents 6.2, 2024. [Online]. Available: https://www.here.com/docs/bundle/traffic-api-developer-guide-v6/page/topics/resource-parameters-incidents.html. Accessed on: Jan. 30, 2024.
[32] Geoapify, APIs and components. 2018. [Online]. Available: https://www.geoapify.com/ . Accessed on: Jan. 30, 2024.
[33] Mónica, M. 'Use of Spearman's Correlation in a Physiotherapy Intervention Study.' Movimiento Científico, vol. 8, no. 1, pp. 98-104, 2014. [Online]. Available: https://doi.org/10.33881/2011-7191.mct.08111
[34] R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/
[35] De la Cruz-Nicolás, E., Martínez-Rebollar, A., Estrada-Esquivel, H., Pliego-Martínez, O.A, "Methodology to Obtain Traffic Data and Road Incidents Through Maps Applications", In: Nesmachnow, S., Hernández Callejo, L. (eds) Smart Cities. ICSC-Cities 2023. Communications in Computer and Information Science, 2023, vol 1938, pp. 3-17.
[36] Here Maps, Find Traffic Along a Route, 2024. [Online]. Available: https://developer.here.com/documentation/ios-sdk navigate/4.14.4.0/dev_guide/topics/routing.html#find-traffic-along-a-route . Accessed on: Feb. 26, 2024.

[37] D. Cao, J. Wu, X. Dong, H. Sun, X. Qu, and Z. Yang, "Quantification of the impact of traffic incidents on speed reduction: A causal inference based approach". Accid. Anal. Prev., vol. 157, no. January, pp. 106163, 2021. Available: 10.1016/j.aap.2021.106163.

[38] T. Wang et al., "Research on the Mechanism of Influencing Factors of the Urban Road Traffic Operation State". Wirel. Commun. Mob. Comput., vol. 2022, 2022. Available: 10.1155/2022/7283841.

[39] M. Kumar, K. Kumar, and P. Das, "Study on road traffic congestion: A review". Recent Trends Commun. Electron., no. June, pp. 230–240, 2021. Available: 10.1201/9781003193838-43.

[40] M. M. Rahman, P. Najaf, M. G. Fields, and J. C. Thill, "Traffic congestion and its urban scale factors: Empirical evidence from American urban areas". Int. J. Sustain. Transp., vol. 16, no. 5, pp. 406–421, 2022. Available: 10.1080/15568318.2021.1885085.

[41] M. Wang and N. Debbage, "Urban morphology and traffic congestion: Longitudinal evidence from US cities". Comput. Environ. Urban Syst., vol. 89, no. September 2020, p. 101676, 2021. Available: 10.1016/j.compenvurbsys.2021.101676.

[42] J. Y. Yap, N. Omar, and I. Ismail, "A Study of Traffic Congestion Influenced by the Pattern of Land Use". IOP Conf. Ser. Earth Environ. Sci., vol. 1022, no. 1, 2022. Available: 10.1088/1755-1315/1022/1/012035.

[43] J. Seong, Y. Kim, H. Goh, H. Kim, and A. Stanescu, "Measuring Traffic Congestion with Novel Metrics: A Case Study of Six U.S. Metropolitan Areas". ISPRS Int. J. Geo-Information, vol. 12, no. 3, 2023. Available: 10.3390/ijgi12030130.

[44] De la Cruz-Nicolás, E., Estrada-Esquivel, H., Martínez-Rebollar, A., Pliego-Martínez, O. A., & Clemente, E. Index for assessing the performance level of vehicular traffic on urban streets. Urban Science, vol. 8, no. 4, 2204. Available: doi:10.3390/urbansci8040204