Discover Artificial Intelligence

Discover

Research

Data collection and analysis applied to intelligent transportation systems: a case study on public transportation

Gabriel Gomes de Oliveira[1] · Yuzo Iano[1] · Gabriel Caumo Vaz[1] · Kannadhasan Suriyan[2]

Received: 23 January 2023 / Accepted: 31 March 2023
Published online: 11 April 2023

Abstract
The big data concept has been gaining strength over the last few years. With the arise and dissemination of social media and high access easiness to information through applications, there is a necessity for all kinds of service providers to collect and analyze data, improving the quality of their services and products. In this regard, the relevance and coverage of this niche of study are notorious. It is not a coincidence that governments, supported by companies and startups, are investing in platforms to collect and analyze data, aiming at the better efficiency of the services provided to the citizens. Considering the aforementioned aspects, this work makes contextualization of the Big Data and ITS (Intelligent Transportation System) concepts by gathering recently published articles, from 2017 to 2021, considering a survey and case studies to demonstrate the importance of those themes in current days. Within the scope of big data applied to ITS, this study proposes a database for public transportation in the city of Campinas (Brazil), enabling its improvement according to the population demands. Finally, this study tries to present clearly and objectively the methodology employed with the maximum number of characteristics, applying statistical analyses (box-and-whisker diagrams and Pearson correlation), highlighting the limitations, and expanding the studied concepts to describe the application of an Advanced Traveler Information System (ATIS), a branch of Intelligent Transportation System (ITS), in a real situation. Therefore, besides the survey of the applied concepts, this work develops a specific case study, highlighting the identified deficiencies and proposing solutions. Future works are also contemplated to expand this study and improve the accuracy of the achieved results.

Keywords Intelligent Transportation System · Big data · Data collection · Data analysis · Public transportation · Urban mobility

1 Introduction

Over the last decades, it has been possible to notice that people's way of living has been changing, aiming for the welfare of themselves and their families. Therefore, there is a worldwide trend of people living near their workplaces, facilitating their daily mobility [1].

A study published by the Journal of Urban Economics [2] shows, based on data from the Danish population, the preferences adopted by scientists and engineers for their workplace choices. The results achieved in this study indicate that the technical workers show a substantial sensitivity to wage differences, but they have stronger preferences for living

✉ Gabriel Gomes de Oliveira, oliveiragomesgabriel@ieee.org; Yuzo Iano, yuzo@unicamp.br; Gabriel Caumo Vaz, gabriel_caumo@yahoo.com.br; Kannadhasan Suriyan, kannadhasan.ece@gmail.com | [1]State University of Campinas – UNICAMP, Campinas 13083-970, Brazil. [2]Study World College of Engineering, Coimbatore, Tamil Nadu 641108, India.

near their families and friends. According to [3], a person spends, on average, 1.1 h every day traveling and dedicating a predictable fraction of the income to move to the workplace.

For the reasons presented above, the ITS (Intelligent Transportation System) integrates information and technologies of data communication and applies them in the transportation field to develop an integrated system of people, roads, and vehicles. It can establish a transportation management system that is fully functional, accurate, and efficient [4, 5].

Intelligent Transportation Systems work with information and control technologies that are the core of their functions [6]. One of its key components is the Advanced Traveler Information Systems [7], which involves data collection, i.e., a strategic action of automatically recording the system's information and parameters in the short and long term. Nowadays, data collection has become increasingly crucial for road transportation operations, especially concerning public transport, creating a big data problem [8].

The expression "big data" is extensive and was first used by NASA scientists to describe a problem of graphical computing. This issue was named "big data" because it occurs when the dataset does not fit into the hardware's main memory or does not fit even into the local disk [9].

The application of the ITS concept is a global trend and, as shown by do Nascimento et al. [10], even countries with a history of technological backwardness are discussing this theme. For that reason, this approach is considered suitable for the proposal of a methodology that could improve the achievement and analysis of data regarding the public transportation service.

This work takes all the aforementioned factors into account and proposes a methodology that quantifies essential parameters for analysis of public transportation service, provided by the government. Therefore, the main contribution of this paper is to provide database parameters of such a service, in order to be possible to connect to an ATIS afterward. The collected and studied variables were temperature, noise, number of people, speed, and delay, and the methodology aims to contextualize the application of ITS, in particular, the Advanced Traveler Information System, whose benefits affect traffic dynamics as a whole [11]. According to the report [12], developed by the United Nations in 2021 and named "Sustainable Transport, Sustainable Development," it is necessary to implement in cities all over the world, independent of their size, projects of intelligent sustainable transport, aimed at the creation of integrated ecosystems that are capable of ensuring the quality of life to the current and upcoming generations. From this perspective, the main objective of this work is to present data and information about urban mobility through innovative technologies and mathematical modeling, making it possible to improve progressively the intelligent transport in the studied city.

Within the presented scope, this work addresses the urban mobility theme and is divided into six sections. The current one has an introductory character for the studied subject and its issues worldwide. The second section presents a survey about the concepts applied to this research: big data and ITS. The third section discusses the materials and methods employed in the development of the proposed methodology, which is described in the fourth section with the metrics adopted to achieve the desired results. The fifth section shows the achieved results and the respective interpretations based on parameters established by health authorities and public agents for the measurement of human comfort in a vehicle. Finally, the last section presents the conclusions of this research, highlighting its relevance in the global scenario, and proposes future works to improve the quality and accuracy of the applied methodology.

2 Literature review

This section is divided into two subsections: Sect. 2.1 demonstrates recently published works with the same scope approached in this research, and Sect. 2.2 shows some case studies applied worldwide.

2.1 Theoretical background

The paper [13] presents a methodology for scientific research concerning the use of big data in transportation systems. The authors gathered and analyzed 115 works of different types (journal papers, conference papers, and dissertations/theses) published from 2003 to 2017. They observed the growing interest of the scientific community (number of publications) for the subject during the aforementioned period.

The work [14] shows an overview of architectures and devices for data collection directed to the application of big data in Intelligent Transportation Systems (ITS). The authors relate three elements of data collection: data sources, tools

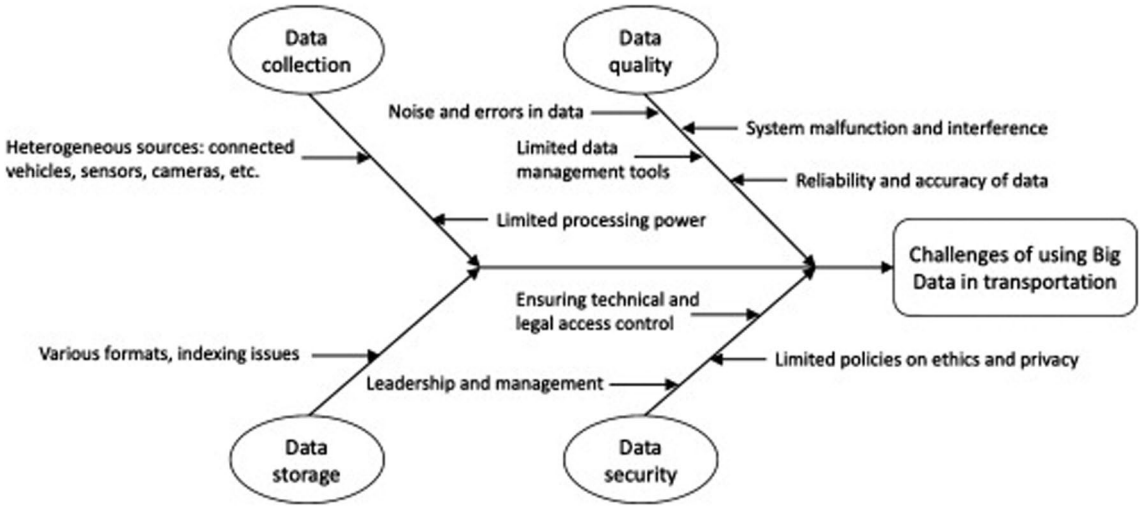

Fig. 1 Challenges of big data in transportation [16]

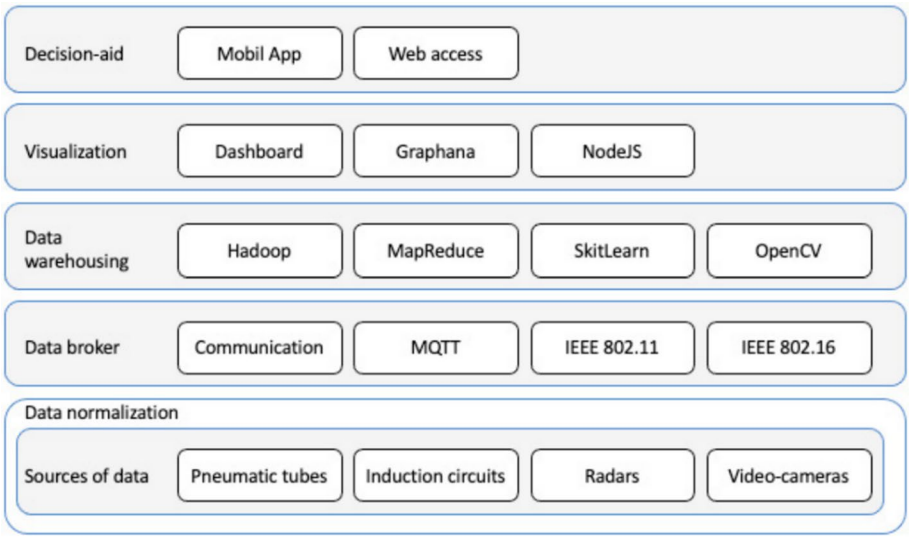

Fig. 2 Architecture from ITS services

through which the data are collected from the sources, and the data that can be achieved from each source. They also propose a machine learning architecture to handle the data within the ITS scope.

The article [15] uses the same methodology of work [14] with a slight difference: it is directed to the application of big data in public transportation, instead of transportation in general. This research focuses on data sources and how much they have been used over the years, e.g., the authors demonstrate that mobile phone data has stood out from 2017 on.

The authors of the paper [16] present concepts and applications of big data in the transportation domain. They also show the challenges of this kind of application, which are depicted in Fig. 1.

In [17], there is a complete architecture for data collection in Intelligent Transportation Systems. The authors structured their architecture from the data source to the visualization and user interface, passing through protocols and frameworks, as shown in Fig. 2.

In work [18], the authors gathered articles concerning state-of-the-art algorithms (machine learning and deep learning) for applications in Intelligent Transportation Systems, showing the evolution of that subject. They presented a variety of models and correspondent applications, as well as a roadmap for future research.

The work [19] presents a computational evolution towards cloud and fog big data analysis, aiming at the intelligent transportation and Internet of Vehicles. The authors are concerned about the necessity of real-time response in latent-sensitive applications, such as unmanned vehicles, and, thinking about it, they propose an architecture to reduce the latency of those applications and provide the basis for further investigation. With a similar concern about the performance of big data analysis

in transportation systems, the authors of work [20] propose an applicable, scalable, efficient algorithm for data storage, which is paramount since, according to the authors, speed measurements of the vehicles of one only carrier in a single day can generate a dataset of multiple Gigabytes.

As shown previously, the ITS concept is currently very well regarded by the global scientific community as a preliminary concept that could solve a problem that affects all countries, especially the large cities: the transportation systems [21, 22].

Within ITS, this work will focus on the Advanced Traveler Information System branch. This theme was first discussed at the end of century XX due to the perception that travelers are an essential part of the transportation systems [23]. For that reason, some assessment metrics started to be studied.

As an example of an application, one can cite the work of Cats et al. [24], which presents a traffic simulation model, called BusMezzo, that aims to evaluate the impacts of real-time traffic information for travelers. In a newer study [25], presents a model, based on Fuzzy logic, to simulate the drivers' choices when an ATIS is applied. Besides the academic works, there are government initiatives that encourage the development of tools for ITS, such as the ITS CodeHub (https://www.its.dot.gov/code) from the U.S. Department of Transportation.

With all those examples in mind, the analyzed data quantify some variables that allow the assessment of comfort and quality of the service provided in the means of public transportation.

This work aims to highlight the importance of information for both the managers and the users of public transportation. A real system of data acquisition and sharing would use the big data concept, but this work makes a small-scale study in one only bus line in the city of Campinas (Brazil) to show how those data would be treated on a municipal scale.

In the following paragraphs, some metrics will be described to justify the collection and analysis of the studied variables and point out the deficiencies in public transportation.

In a document published by the World Health Organization (WHO) [26], establishes that the limits of human comfort for the environmental temperature are 17 °C and 31 °C. Not considering the clothes people wear and the air circulation, which affect the individual thermal feeling, the further the temperature is from this range, the higher the risks for human health, which go from sleepiness and mental slowing to blood pressure problems and heart failure.

In the same document [26], presents a study that relates the noise level with the number of complaining participants. The study showed that the percentage of complaints decreases from 85 to 33% when the noise level decreases from 80 to 55 dBA, and it reaches just 5% when the noise decreases to 50 dBA. The noise, like the temperature, has negative impacts on people's health, which vary from irritability and anxiety to permanent damage to hearing function.

As [27] demonstrated, the roads are dimensioned with specific calculations for civil construction to identify the speed limit on the concerned road, assuring the safety of the users. Those calculations take into account the conditions that increase the accident risk and, therefore, limit the maximum speed. In a case study [28], shows that a hike in speed limit can be associated with the fatality rate in traffic accidents, highlighting the importance of the proper design of roads and respect to the established limits.

In the studies carried out in [29], the basic requirements and parameters for the comfort of public transportation passengers were assessed. One of those studies showed that, beyond a specific level of occupancy, travelers start feeling uncomfortable, and their physiological functions are negatively affected. Thus, the itinerary planning must consider the occupancy level to meet the dynamic needs of the passengers. Therefore, the collection of these data is important as well for the public transport operators to make better planning aiming at the population welfare.

According to [30], public transportation means tend to suffer unpredictable delays, especially in large cities (e.g., New York, Los Angeles, Tokyo, London, and São Paulo), where the users usually travel via a chain of different means (e.g., underground, train, bus, Bus Rapid Transit, Light Rail Transit, and ferry) to move to their destination, a little delay at the beginning of the journey may result in a large delay at its end. For that reason, it is of fundamental importance the synchronization of the transportation systems in these cities and the ITS application because, through this concept, one can integrate the diverse means of a city. It is important to highlight that small and medium-sized cities (e.g., Campinas) do not have a wide variety of transportation means, but they can take advantage of the ITS as well since the public transport operators can improve the quality of the provided services if they have access to the relevant information.

2.2 Case studies

As shown in the previous subsection, in recent years, there is a growing interest in Intelligent Transportation Systems, especially when they are related to big data and data analysis. There are, also, many practical solutions that were developed based on those studies, and some of them are discussed below.

One of the most traditional problems in transportation systems is the management of traffic flow. In [31], the authors apply the concept of Cyber-Physical-Social Systems (CPSS) to achieve signals from both the physical and social spaces. The proposed CPSS-based Transportation System (CTS) is a software-defined transportation system that creates an environment where human factors, transportation systems, and computing technologies are integrated and interact to provide intelligent responses that affect the real world (Fig. 3).

Another conventional approach concerning Intelligent Transportation Systems is traffic light control to optimize the traffic flow. The work [32] proposes a solution in this regard. The authors compared the traditional traffic light control with two algorithms (experiential fuzzy control rule and fuzzy control rule gained by genetic algorithm). Both achieved better results than the traditional model, and, when the second one was used, the vehicles' average waiting time in traffic lights was up to 48.72% lower than the waiting time with the traditional model.

In recent years, new technologies, such as the Internet of Things (IoT) and Machine Learning, have arisen, and many of them can be applied to improve solutions to longstanding problems. It is shown in [33], whose authors gathered real applications around the world and described the problems and how each technology was employed to solve them. As examples, one can cite the use of big data and IoT to solve problems with road irregularities in Boston (USA) and how big data, IoT, and Machine Learning were applied to identify road sections with a high risk of road transportation accidents in Moscow (Russia).

Most works apply the previously mentioned technologies to solve problems, but they can be used to gather knowledge about a transportation system as well. An example is the work [34], which uses carpooling big data to identify multiple centers within a metropolitan region and, based on that information, study strategies to improve the traffic flow between the centers.

Most applications of ITS and big data are performed in developed countries but developing countries have been taking advantage of the improvement of those technologies. As an example, one can cite the article [35], which presents a management system employed in Fortaleza (Brazil) to support smart urban mobility directed to bus transportation. The solution has three layers: data acquisition and processing through big data techniques, investigation of patterns and correlations among the studied variables, and a visualization environment to support urban planners' decisions. This kind of work can also help evaluate the interaction between different smart city urban mobility programs.

3 Methodology

This section is split into two subsections: the first one presents how the studied dataset was built, with details about the analyzed road and data sources (sensors and applications) that were applied; the second subsection highlights the methodology adopted for the developed research.

3.1 Data collection

All the researched data were collected in one only stretch (Fig. 4) to validate and compare the data. It is necessary to observe that the image on the left side shows the whole itinerary from the Barão Geraldo bus station to the Central bus station, and the image on the right side indicates the studied stretch, which goes from point A (430 Albino José Barbosa de Oliveira Ave.) to point B (1968 Carolina Florence St.) by the Professor ZeferinoVaz Highway. The measurements were always made in the same stretch, avoiding interference and ensuring the presented data accuracy.

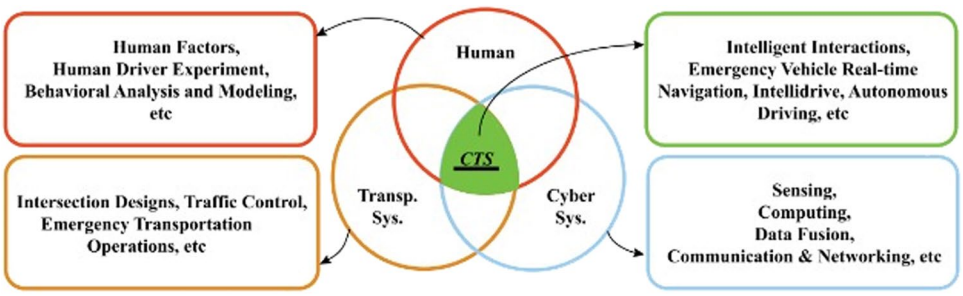

Fig. 3 CPSS-based Transportation System [31]

The data presented in this work were collected with the following tools:

- 1 decibel meter Benetech model GM1351 with measuring range from 30 to 130 dBA, accuracy of ± 1.5 dB, frequency range from 31.5 Hz to 8 kHz, and sample rate of 2 times/second [36];
- 1 thermometer Benetech model GM320 with a digital infrared, temperature range from − 50 °C to 400 °C, accuracy of ± 1.5 °C, resolution of 0.1 °C, repeatability of 1 °C, response time of 500 ms, spectral response from 8 to 14 µm, emissivity of 0.95 preset, distance to spot size of 12:1, operating temperature from 0 °C to 40 °C, operating humidity from 10 to 90 RH, storage temperature from − 20 °C to 60 °C, and power supply of 3 V (2 AAA 1.5 V batteries) [37];
- 1 digital chronometer Tak Shun model TS-1809 with dimensions 95 × 65 × 26 mm and a string of 50 cm [38];
- Application Waze Navigation& Live Traffic (https://www.waze.com), version 4.69.4.900.

All the measurements performed with sensors were collected from a height of 1.5 m.

3.2 Data analysis

After the data collection, the programming language Python 3.8.5 (https://www.python.org) was used in the development environment JupyterLab 2.2.6 (https://jupyter.org) with the libraries Matplotlib 3.3.2 (https://matplotlib.org), Numpy 1.19.2 (https://numpy.org), Pandas 1.1.3 (https://pandas.pydata.org), and Plotly 4.14.3 (https://plotly.com), aiming to organize and analyze the data that will be shown in the results.

The data comprise five variables (temperature, noise, number of people, speed, and delay), meeting the criteria presented on the following topics.

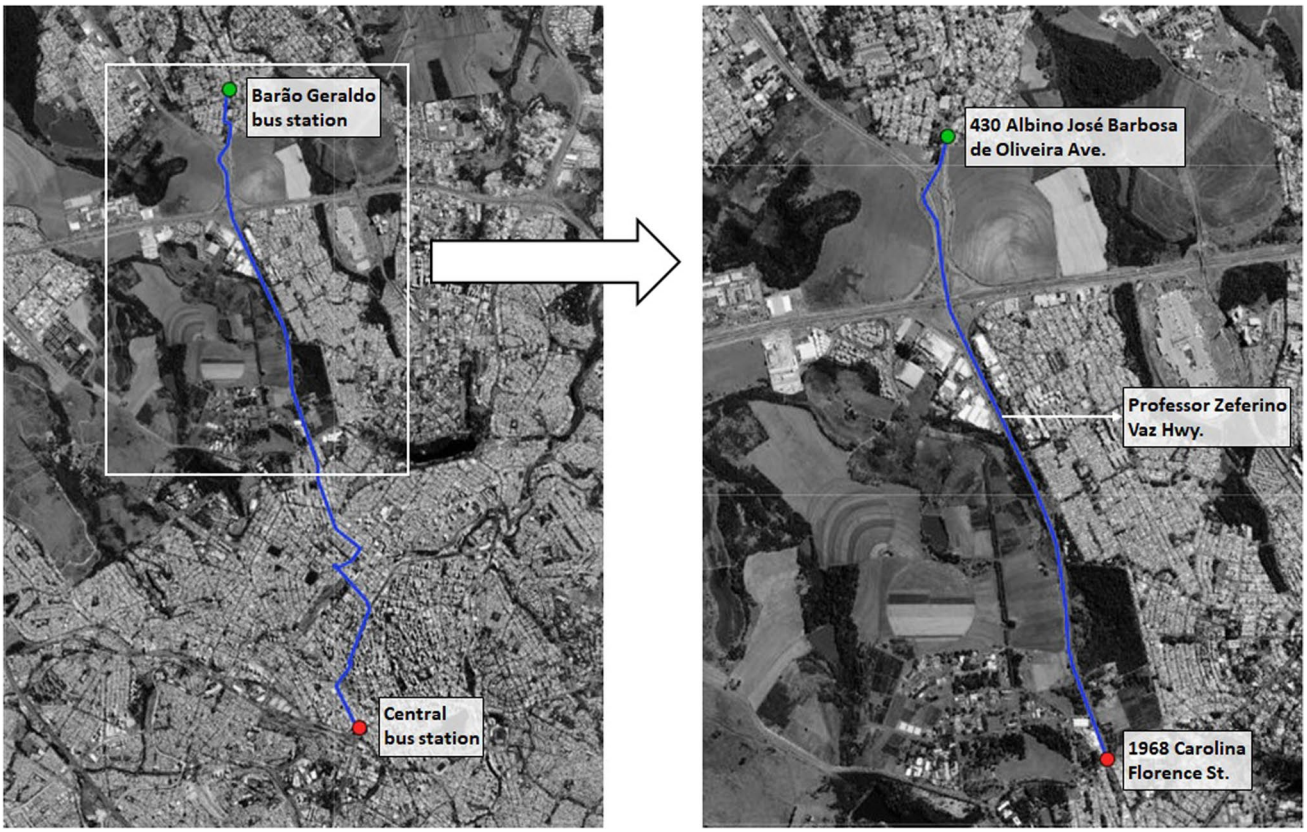

Fig. 4 Studied route (https://www.maps.ie)

3.2.1 Temperature

Since the thermometer described in Sect. 3.1 collects just the temperature in one specific point instead of the ambient temperature, 20 points were selected inside the bus to improve the analyzed data. The metric applied to choose the measurement points was based on the different materials (synthetic fabric, plastics, and metals) and covered the whole extension of the bus from the front end to the back end.

3.2.2 Noise

The decibel meter mentioned in Sect. 3.1 records two values: the minimum noise and the maximum noise. Therefore, in each travel, two values of this variable were collected.

3.2.3 Speed

Two speed values were collected in each travel: the bus speed when it takes the highway, Professor ZeferinoVaz, between points A and B described in Sect. 3.1, and its speed when it enters downtown, after the highway.

3.2.4 Number of people

Two perspectives were observed to collect these data: the number of people who took the bus at the beginning of the itinerary and the number of passengers after the only bus stop in the analyzed stretch, where people could get on and get off the bus.

3.2.5 Delay

This variable refers to the bus departure delay at the start point, i.e., Barão Geraldo's terminal. The time was measured with the digital chronometer, and the reference was the app CittaMobi (https://www.cittamobi.com.br), whose objective is daily to orient the passengers concerning the buses' schedules and itineraries.

4 Results

A database was created for posterior analyses via software through the collection tools presented in Sect. 3.1. The collections were always made in the same itineraries: 08:50 a.m., 01:05 p.m., and 06:51 p.m. The collected data were stored in five CSV (Comma-Separated Values) files, i.e., one file for each variable (temperature, noise, number of people, speed, and delay).On each travel, a total of 27 pieces of data were collected: 20 points of temperature, minimum and maximum noise, two speeds, two numbers of people, and the delay. The measurements were made in three periods each day (morning, afternoon, and evening) and seven days per week, i.e., there were 567 data collected weekly. With the purpose of ensuring more significant coverage of the data, 3 weeks with different characteristics (vacation, carnival, and school) were selected. Therefore, during the research, a total of 1701 raw data were collected, analyzed, and studied.

4.1 Pre-processing

The first step of the data analysis consists of assessing the collected data quality, i.e., the identification of possible measurements and typing errors, focusing on the outliers that must be analyzed more carefully. The atypical data are those whose values are higher than the third quartile plus 1.5IQR (Interquartile Range) or lower than the first quartile minus 1.5IQR [39].

The box-and-whisker diagram is used to make this analysis. As described in the library Plotly (https://plotly.com) documentation, this diagram is built from five points: the minimum value, the three quartiles, and the maximum value, where the second quartile corresponds to the median of the whole data set, the first quartile is the median of the data between the minimum value and the second quartile, and the third quartile is the median of the data between the second quartile and the maximum value.

Once these values are computed, the box-and-whisker diagram is built according to the following steps: a rectangle is defined with the lower end in the first quartile and the upper end in the third quartile, the second quartile is plotted inside the rectangle, and two line segments (whiskers) are created on each side of the rectangle to represent the atypical data limits (1st quartile − 1.5IQR and 3rd quartile + 1.5IQR). All the data outside these limits are considered outliers [39].

The collected data were grouped by days (Monday to Sunday) and by period (morning, afternoon, and evening). The following subsections show each variable distribution in the referred data sets.

4.1.1 Temperature

It is possible to see in the temperature distributions (Fig. 5) that, in the evening period, they tend to be lower due to the nature of the weather, while, in the morning and afternoon periods, the central values are always within the same interval (25 °C to 30 °C), even though the distributions sometimes extrapolate these values. The outliers result from the different materials where the measurements were made, e.g., metal surfaces tend to have a lower specific heat than other materials. That means they heat up or cool down more easily, i.e., the metals have a greater temperature on hot days than the other materials inside the bus, and the opposite occurs on cold days [40].

4.1.2 Noise

The noise values distributions (Fig. 6) have medians concentrated on the interval between 70 and 75 dBA. This trend derives from the constant noise of the combustion engine. The wider distributions occur due to random variables, e.g., people talking, air conditioning, and external noises, resulting from the traffic and occasional civil works. Due to these variables' randomness, it is not possible to observe a pattern in the distributions when the data of a given period of day are compared.

An important detail is that the noise is measured in dBA, i.e., it is a logarithmic variable, which differentiates it from the other collected variables. However, as demonstrated by Golmohammadi et al. [41], concerning the data types, it does not prevent the comparison with continuous or discrete variables.

4.1.3 Number of people

It is not possible to identify a clear pattern in the number of people distributions (Fig. 7). It happens because of the great availability of buses for the studied route, which permits the passengers to take the bus at another time, different from those studied in this research, i.e., the people who take the bus at a given time do not take necessarily the same bus every day. Besides, it is essential to mention the diversity of alternatives to public transportation so that, on some days, the buses are less used.

In the evening period, there is a greater distribution due to the flow of students of the State University of Campinas who use the studied bus line to move to the evening classes.

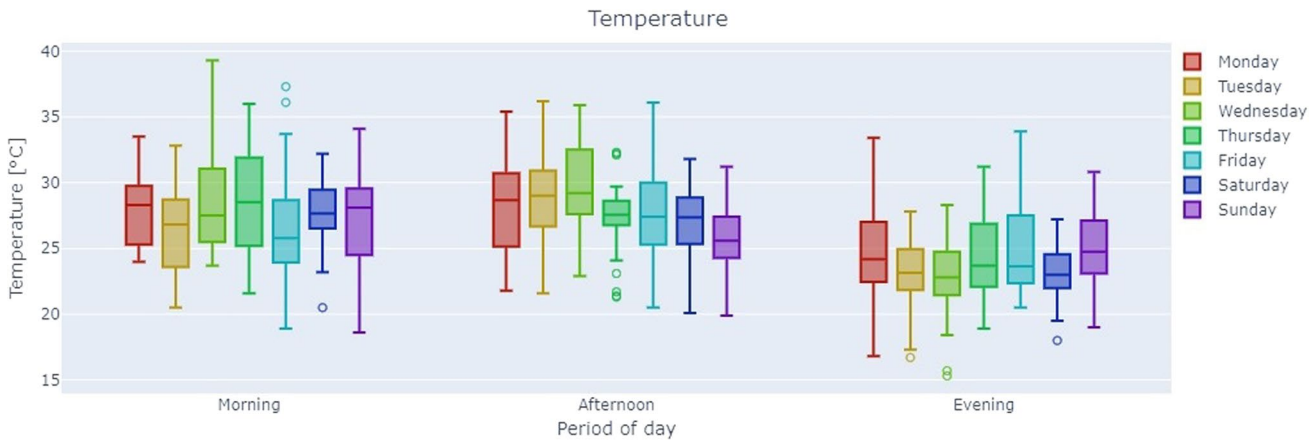

Fig. 5 Temperature measurements distributions

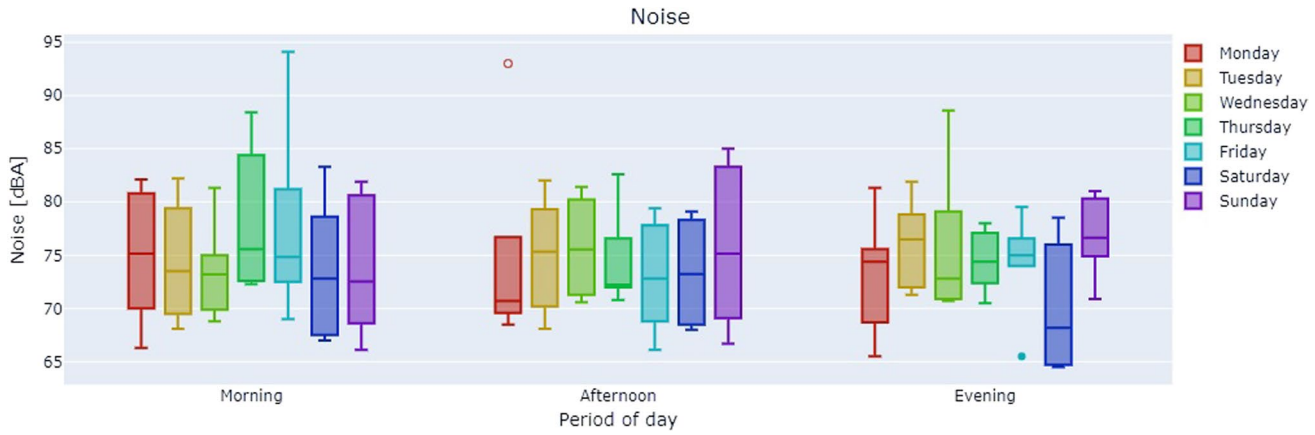

Fig. 6 Noise measurements distributions

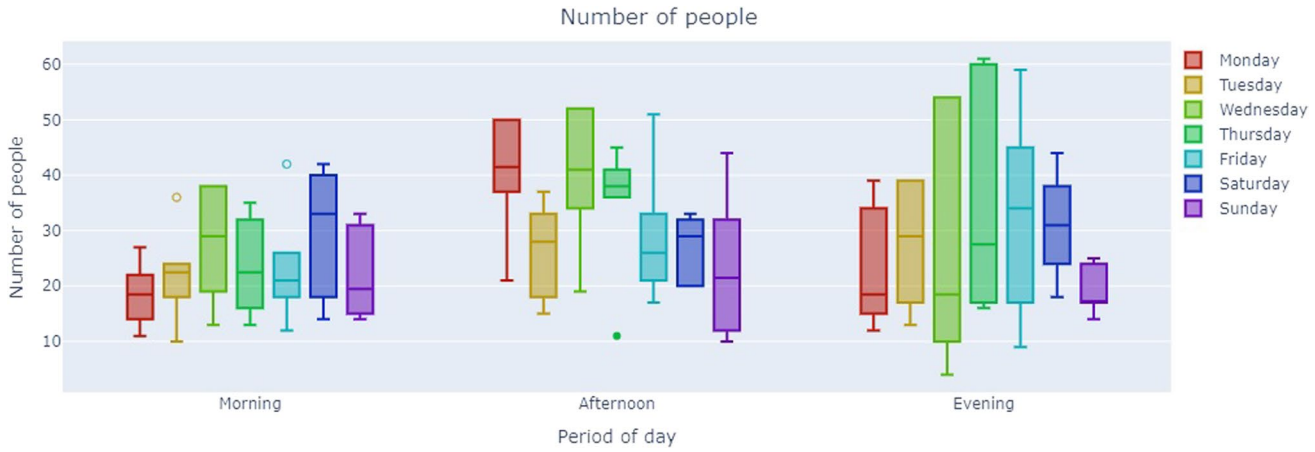

Fig. 7 Number of people distributions

4.1.4 Speed

In the morning period, there are more discrepancies among the daily speed distributions (Fig. 8). It results from the larger number of vehicles on the road at the measurement time, which sometimes causes heavier traffic congestion in this period of the day.

In the other periods, it is possible to note that the central values are, in most cases, close to the maximum road speed (50 km/h + 7 km/h of margin error) defined by the regulatory agency [42]. It suggests that bus drivers mostly drive at the road speed limit with some overrunning of this value. In the evening period, there is a stronger trend of exceeding the speed limit because of many factors, e.g., the lower number of vehicles on the highway and higher flexibility of the speed supervision due to the insecurity at that time.

4.1.5 Delay

The diagrams that represent the delays (Fig. 9) show that, on the weekends, the wait time tends to be lower than on the other days.

Considering the workdays, there is a broader distribution of the collected values in the morning, with a central trend between 5 and 6 min and most of the values below these medians (some of them are zero). It happens due to the characteristic traffic jam in this period of the day.

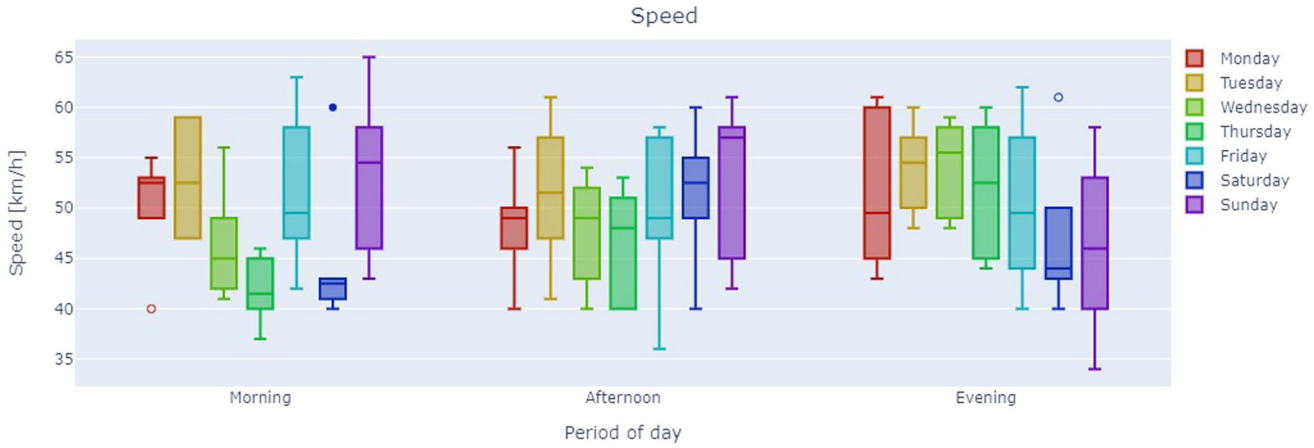

Fig. 8 Speed measurements distributions

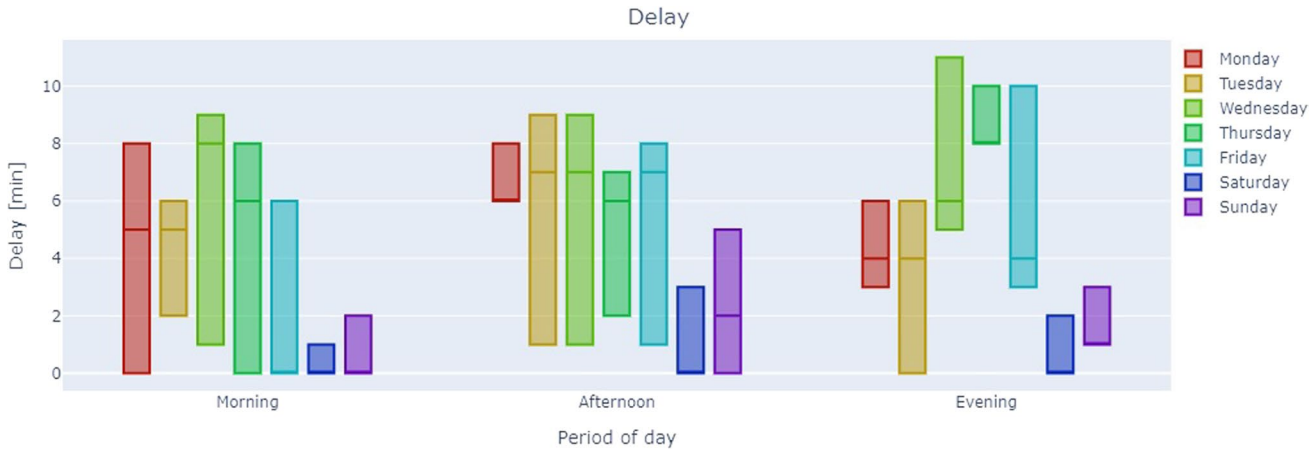

Fig. 9 Delay measurements distributions

This variable behaves in the same way in the afternoon, but the central trend is higher, between 6 and 7 min. However, the minimum values are concentrated between 2 and 3 min. It results from the measurement time, which matches people's lunchtime, so traffic congestion is more intense.

4.2 Results analyses

In the previous subsection, each variable was presented and discussed individually, highlighting the different behaviors in each studied day and period. In this section, the correlation between the variables will be computed and analyzed through the Pearson correlation, which is a linear association between variables, and does not depend on the units. This method results in a coefficient (ρ) for each correlation that ranges from -1 to $+1$, where the absolute value refers to the correlation strength, and the sign shows whether the relationship is positive or negative [43].

The correlation level between two variables obeys the following threshold values [44]:

- $0.0 \leq |\rho| < 0.3$: Negligible correlation;
- $0.3 \leq |\rho| < 0.5$: Weak correlation;
- $0.5 \leq |\rho| < 0.7$: Moderate correlation;
- $0.7 \leq |\rho| < 0.9$: Strong correlation;
- $0.9 \leq |\rho| < 1.0$: Very strong correlation;
- $|\rho| = 1.0$: Perfect correlation.

All the studied variables were correlated between them, resulting in ten correlation values for each analyzed period. In the first analysis, the collected values were grouped by days of the week, whereas, in the second analysis, they were grouped by week (vacation, carnival, and school). The computed coefficients were presented in ten heatmaps, whose graded color scale ranges from −1 to +1, the same range as the Pearson coefficients.

Following the metric adopted in this work, just the moderate and strong correlations were considered. In this research, there were not very strong and perfect correlations.

4.2.1 Daily correlations

In the heatmaps of Fig. 10, there are Pearson coefficients with an absolute value above 0.5 just in the following correlations: temperature and number of people, noise and speed, noise and delay, number of people and speed, and number of people and delay.

- Temperature vs. Number of people: This correlation arises on Mondays ($\rho = 0.51$). As shown in the diagram of Fig. 5, on Mondays, especially in the morning and evening periods, the number of passengers tends to be lower, resulting in better air conditioning system efficiency and, therefore, in a lower temperature, which characterizes a positive correlation on this day.
- Noise vs. Speed: This correlation occurs on Saturdays ($\rho = -0.6$). There are many sources of noise inside and outside the bus, but, in this case, the combustion engine may be mentioned because it tends to be less noisy when a constant speed is kept (50 km/h in the studied stretch). Figure 6 shows that, on Saturdays, specifically in the morning, the speed tends to remain below the road speed limit, a result of more braking and acceleration, which are noise sources related to the combustion engine.
- Noise vs. Delay: This correlation happens on Thursdays ($\rho = -0.64$) and Fridays ($\rho = -0.64$). In this case, it was not possible to identify relevant correlations between the noise and other variables that could directly affect the delay time, i.e., the noise variance is due to external factors. On Thursdays and Fridays, there is a behavioral tendency for people to use their own car to go to work, and that causes more external noise. However, the correlations between the number of people and the delay are just weak and moderate these days. Specifically, in the morning period, the number of passengers is lower, and consequently, the delay is lower as well.
- Number of people vs. Speed: This correlation arises on Thursdays ($\rho = 0.53$). In this case, there is a lack of variables that could support the correlation and also a broad variance among the individual heatmaps (Fig. 10), with values ranging from −0.34 to +0.53. It suggests that the greater dispersion of the data collected on Thursdays concerning the number of people (Fig. 7) and speed (Fig. 8) resulted in a false positive in this correlation.
- Number of people vs. Delay: This correlation occurs on Fridays ($\rho = 0.53$) and Sundays ($\rho = 0.7$). These days are among those when there are fewer passengers on the bus (Fig. 7). It reduces the boarding time at the line beginning and

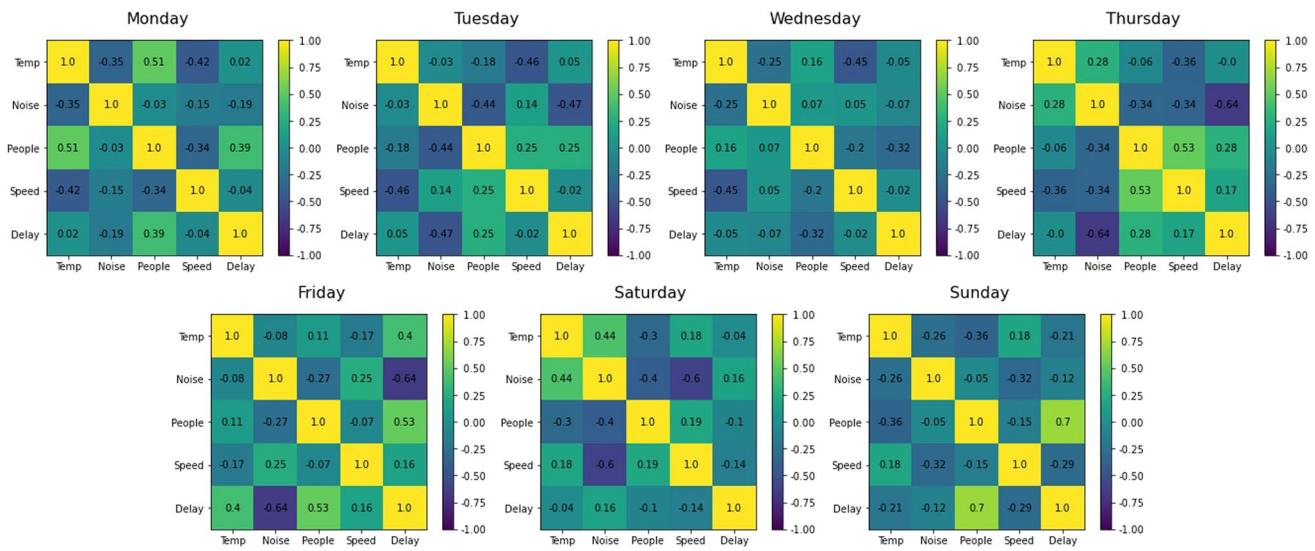

Fig. 10 Daily correlations heatmaps

the stop time during the travel. Therefore, the lower the number of people, the lower the delay time is. Likewise the previous case, there are correlation values on other days that disagree with this analysis, but those days also have a broader dispersion of the collected data.

4.2.2 Weekly correlations

In the heatmaps of Fig. 11, there are Pearson coefficients whose absolute values are greater than 0.5 just in the correlation between the number of people and the delay.

This correlation happens during the vacation week ($\rho = 0.58$). The analyzed bus line is widely used by students from the State University of Campinas. Therefore, during the vacation period, the passenger flow is smaller. From the analysis of this same correlation in the previous subsection, it is possible to conclude that the lower the number of people, the lower the delay time is.

5 Discussion

Based on the achieved results, this work demonstrates the feasibility of the use of data analysis techniques in the public transportation field. The implementation of technological tools and the use of well-established methods for data correlation allowed the identification of problems that would serve as a basis for policies to improve the public transportation service.

The methodology presented was validated in a bus environment but could be extended to any other transportation mean and would be applied in any city, given its applicability. However, some issues that were observed during this study must be taken into account. The first one is the measurements' accuracy because there were cases, like the correlation between the number of people and speed, where the data points were very spread, and they ended up resulting in a false positive correlation. The same occurred in the correlation between the number of people and the delay, though it was less evident.

The second problem was the lack of variables to make a thorough analysis of the correlations. That happened in the correlation between the noise and the delay, in which it was found that there was a neglected external variable affecting the results. Another case that can be mentioned is the temperature which was not strongly correlated with any other variable. Still, it could correlate, for example, with the temperature outside the bus, which was not analyzed. This issue is relevant because it shows the importance of having a complete overview of the investigated conditions.

The third difficulty was the amount of analyzed data. Although there were more than 1700 data collected and studied, there was not enough information to understand the behavior of the variables on holidays, for example, because, in the studied period, there was just one day with this characteristic. Likewise, from the ten possible correlations between the five studied variables, only five could be analyzed with a correlation of at least moderate, and one of them was repeated in both the daily and weekly analyses. With more data, the correlations that were not analyzed could be better understood.

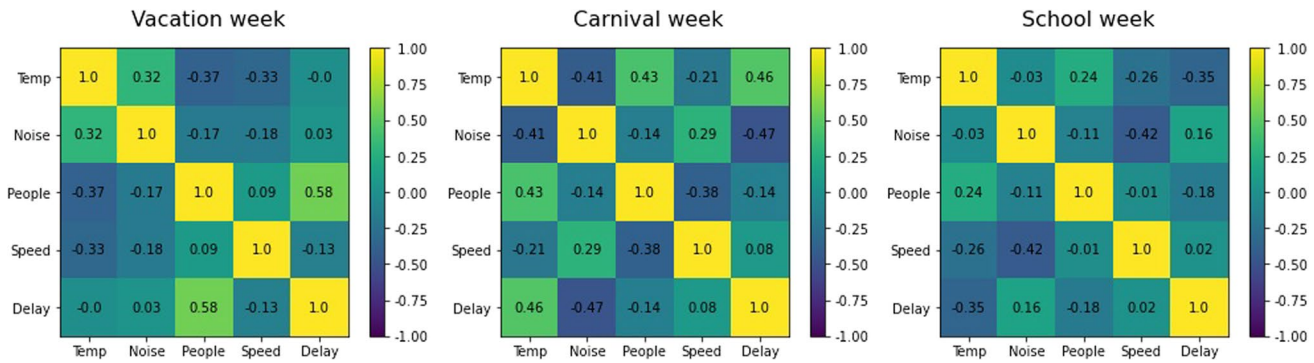

Fig. 11 Weekly correlations heatmap

6 Conclusion

This research proposed a methodology for data acquisition and processing within the scope of urban mobility. With the collected data, it was possible to identify the particular characteristics of each studied variable and analyze the way they interact among themselves. Different approaches were used to cluster and correlate the data, and each one brought new insights to help improve the quality of the public transportation service in the studied city.

In the same way as variables and people have dynamic behavior, the analyses must also be dynamic and ample, grouping the data by periods of the day and days of the week, for example. Thus, it could be possible to draw up improvement strategies for each studied period. Considering the analyses made in this work, even on a small scale, one can observe that, nowadays, the data are important to determine and establish standards and actions in the discussed scenario.

Concerning the transparency for the users and the information for the managers, the implementation of some concepts must be considered. The big data would be applied to create a volume of data that enables relevant analyses and outcomes for planning and managing public transportation. With the same thing in mind, one can mention the Intelligent Transportation System (ITS) and, in particular, the Advanced Traveler Information System (ATIS) may improve the quality of the transportation service regarding the users' communication, comfort, and welfare.

In this regard, this research can serve as a basis for future works by developing automatic systems for data collection, which could generate a more robust dataset and, as consequence, achieve higher accuracy for the decision-making process.

Author contributions GGO—Writing. YI—Editing, analysis. GCV—Data collection. KS—Transportation analysis. All authors read and approved the final manuscript.

Data availability The data that support the findings of this study are available from the corresponding author upon request.

Code availability The code that supports the findings of this study is available from the corresponding author upon request.

Declarations

Competing interests The authors declare no competing interests.

References

1. Brun J, Fagnani J. Lifestyles and locational choices—trade-offs and compromises: a case-study of middle-class couples living in the Ile-de-France region. Urban Stud. 1994;31(6):921–34.
2. Dahl MS, Sorenson O. The migration of technical workers. J Urban Econ. 2010;67(1):33–45.
3. Roth GJ, Zahavi Y. Travel time 'budgets' in developing countries. Transp Res Part A Gen. 1981;15(1):87–95.
4. De Brucker K, Macharis C, Verbeke A. Two-stage multi-criteria analysis and the future of intelligent transport systems-based safety innovation projects. IET Intell Transp Syst. 2015;9(9):842–50.
5. Hou Z, Zhou Y, Du R. Special issue on intelligent transportation systems, big data and intelligent technology. Transp Plan Technol. 2016;39(8):747–50.
6. Jarašūniene A. Research into intelligent transport systems (ITS) technologies and efficiency. Transport. 2007;22(2):61–7.
7. Zhang J, Wang F-Y, Wang K, Lin W-H, Xu X, Chen C. Data-driven intelligent transportation systems: a survey. IEEE Trans Intell Transp Syst. 2011;12(4):1624–39.
8. Wu X, Liu HX. Using high-resolution event-based data for traffic modeling and control: an overview. Transp Res part C Emerg Technol. 2014;42:28–43.
9. Cox M, Ellsworth D, Application-controlled demand paging for out-of-core visualization. In Proceedings. Visualization'97 (Cat. No. 97CB36155), 1997, pp. 235–244.

10. do Nascimento DA, et al. Sustainable adoption of connected vehicles in the Brazilian landscape: policies, technical specifications and challenges. Trans Environ Electr Eng. 2019;3:44.
11. Yang H, Meng Q. Modeling user adoption of advanced traveler information systems: dynamic evolution and stationary equilibrium. Transp Res Part A Policy Pract. 2001;35(10):895–912.
12. Sustainable Transport, Sustainable development:, UN, 2023. https://sdgs.un.org/sites/default/files/2021-10/Transportation%20Report%202021_FullReport_Digital.pdf. Accessed 27 Mar 2022.
13. Ghofrani F, He Q, Goverde RMP, Liu X. Recent applications of big data analytics in railway transportation systems: a survey. Transp Res Part C Emerg Technol. 2018;90:226–46.
14. Zhu L, Yu FR, Wang Y, Ning B, Tang T. Big data analytics in intelligent transportation systems: a survey. IEEE Trans Intell Transp Syst. 2018;20(1):383–98.
15. Welch TF, Widita A. Big data in public transportation: a review of sources and methods. Transp Rev. 2019;39(6):795–818.
16. Neilson A, Daniel B, Tjandra S. Systematic review of the literature on big data in the transportation domain: concepts and applications. Big Data Res. 2019;17:35–44.
17. Montoya-Torres JR, Moreno S, Guerrero WJ, Mejía G. Big data analytics and intelligent transportation systems. IFAC-PapersOnLine. 2021;54(2):216–20.
18. Kaffash S, Nguyen AT, Zhu J. Big data algorithms and applications in intelligent transportation system: a review and bibliometric analysis. Int J Prod Econ. 2021;231: 107868.
19. Darwish TSJ, Bakar KA. Fog based intelligent transportation big data analytics in the internet of vehicles environment: motivations, architecture, challenges, and critical issues. IEEE Access. 2018;6:15679–701.
20. Islam MJ, Sharma A, Rajan H. A cyberinfrastructure for big data transportation engineering. J Big Data Anal Transp. 2019;1(1):83–94.
21. Dimitrakopoulos G, Demestichas P. Intelligent transportation systems. IEEE Veh Technol Mag. 2010;5(1):77–84.
22. Joseph AD, et al. Intelligent transportation systems. IEEE Pervasive Comput. 2006;5(4):63–7.
23. Polydoropoulou A, Ben-Akiva M, Khattak A, Lauprête G. Modeling revealed and stated en-route travel response to advanced traveler information systems. Transp Res Rec. 1996;1537(1):38–45.
24. Cats O, Burghout W, Toledo T, Koutsopoulos HN. Modeling real-time transit information and its impacts on travelers' decisions 2. In Proceedings of the 91st TRB Annual Meeting, 2012, vol. 36, p. 37.
25. Dell'Orco M, Marinelli M. Modeling the dynamic effect of information on drivers' choice behavior in the context of an Advanced Traveler Information System. Transp Res Part C Emerg Technol. 2017;85:168–83.
26. W. H. Organization, "Indoor environment: health aspects of air quality, thermal environment, light and noise," World Health Organization, 1990.
27. de Oliveira CG, Iano Y, "Rodovias e suas características: Estudo de rodovias utilizando e aplicando cálculos para este segmento", 1st ed. NovasEdiçõesAcadêmicas, Latvia, 2020
28. Ossiander EM, Cummings P. Freeway speed limits and traffic fatalities in Washington State. Accid Anal Prev. 2002;34(1):13–8.
29. Kogi K. Passenger requirements and ergonomics in public transport. Ergonomics. 1979;22(6):631–9.
30. Rietveld P, Bruinsma FR, Van Vuuren DJ. Coping with unreliability in public transport chains: a case study for Netherlands. Transp Res Part A Policy Pract. 2001;35(6):539–59.
31. Zheng X, et al. Big data for social transportation. IEEE Trans Intell Transp Syst. 2015;17(3):620–30.
32. Wang C, Li X, Zhou X, Wang A, Nedjah N. Soft computing in big data intelligent transportation systems. Appl Soft Comput. 2016;38:1099–108.
33. Iliashenko O, Iliashenko V, Lukyanchenko E. Big data in transport modelling and planning. Transp Res Procedia. 2021;54:900–8.
34. Liu X, Yan X, Wang W, Titheridge H, Wang R, Liu Y. Characterizing the polycentric spatial structure of Beijing Metropolitan Region using carpooling big data. Cities. 2021;109: 103040.
35. Y. Wang, S. Ram, F. Currim, E. Dantas, and L. A. Sabóia, "A big data approach for smart transportation management on bus network," in 2016 IEEE international smart cities conference (ISC2), 2016, pp. 1–6.
36. "Digital Sound Level Meter GM1351 - Shenzhen Jumaoyuan Science And Technology Co.,Ltd.", Benetechco.net, 2022. http://www.benetechco.net/en/products/gm1351.html. Accessed 15 Feb 2022.
37. "Infrared thermometer GM320 - Shenzhen Jumaoyuan Science And Technology Co.,Ltd.", Benetechco.net, 2022. http://www.benetechco.net/en/products/infrared-thermometer-gm320.html. Accessed 15 Feb 2022.
38. "PutianDexin Electronic CO.,LTD", Taksuncn.com, 2022. http://www.taksuncn.com/Aspx/En/product_detail.aspx?ProductsID=602. Accessed 15 Feb 2022.
39. L. Moreno and A. Morcillo, EstatísticaDescritiva. 2019.
40. Torres CMA, Ferraro NG, Soares PAT, and Penteado PCM, "Física Ciência e Tecnologia", 4th ed. Moderna, São Paulo, 2016
41. Golmohammadi R, Ghorbani F, Mahjub H, Daneshmehr Z. Study of school noise in the capital city of Tehran-Iran. 2010.
42. Conselho Nacional de Trânsito, Resolução n°, 396 de 13 de dezembro de 2011. 2011.
43. de Andrade Martins G. Estatística geral e aplicada. 6th ed., vol. 1 Editora Atlas SA, São Paulo: 2017.
44. Correlação: direto ao ponto. Medium, 2022. https://medium.com/brdata/correla%C3%A7%C3%A3o-direto-ao-ponto-9ec1d48735fb#:~:text=A%20correla%C3%A7%C3%A3o%20de%20Pearson%20mede,ser%20de. Accessed 15 Feb 2022.