TRANSPORTES
www.revistatransportes.org.br
anpet
associação nacional de pesquisa e ensino em transportes
www.anpet.org.br

The use of traffic data from automatic monitoring systems to obtain day-to-day time series of vehicle traffic volumes and origin-destination flows in urban networks

O uso de dados de sistemas de monitoramento automático de tráfego para obter séries temporais dia-a-dia de volumes de tráfego e fluxos origem-destino em redes urbanas

Joana Maia Fernandes Barroso[1], **João Lucas Albuquerque Oliveira**[2], **Francisco Moraes de Oliveira Neto**[3]

[1]Federal University of Ceará, Ceará – Brazil, joana@det.ufc.br
[2]Federal University of Ceará, Ceará – Brazil, joao@det.ufc.br
[3]Federal University of Ceará, Ceará – Brazil, moraes@det.ufc.br

Recebido:
28 de maio de 2020
Aceito para publicação:
30 de novembro de 2020
Publicado:
18 de agosto de 2021
Editor de área:
Sara Ferreira

Keywords:
Day-to-day traffic volumes.
Day-to-day origin-destination flows.
Day-to-day traffic dynamics.
Transport data analysis.

Palavras-chave:
Volumes de tráfego dia a dia.
Fluxos origem-destino dia a dia.
Dinâmica do tráfego dia a dia.
Análise de dados em transportes.

DOI:10.14295/transportes.v29i2.2385

ABSTRACT
The understanding of travel pattern dynamics in the urban environment is essential for the transportation systems planning and operation. Recently, the increasing availability of massive traffic data from traffic monitoring systems, including automatic number plate recognition systems (TMS-ANPR), can allow an understanding of the day-to-day variability of traffic flows in large urban network systems. However, to enhance the data quality for analysis, it is essential to carry out a previous data treatment. This work presents a method for treatment of TMS-ANPR data. The main product of this data treatment are the day-to-day time series of traffic volumes and OD flows for different periods of a typical day, allowing the analysis of the multiday dynamic of travel behavior and of the model assumptions stated in the literature about such dynamic behavior. The proposed method, which can be applied to any type of TMS-ANPR, was applied to generate time series data from the TMS-ANPR of Fortaleza city, contributing to identify suspicious and atypical data, to define representative patterns of vehicular traffic and to estimate series of OD flows.

RESUMO
A compreensão do padrão de deslocamentos no meio urbano é essencial para o planejamento e operação dos sistemas de transporte. Recentemente, a crescente disponibilidade de dados massivos de tráfego a partir de sistemas de monitoramento de veículos, equipados com sistemas de reconhecimento de placas (SMV-RP), pode permitir uma compreensão da variabilidade dia a dia dos fluxos de tráfego na malha viária de grandes centros urbanos. No entanto, antes de qualquer análise é essencial realizar um tratamento dos dados. Este trabalho apresenta um método para tratamento de dados de SMV-RP. O principal produto deste tratamento de dados são as séries temporais volumes de tráfego e fluxos origem-destino (OD) para diferentes períodos de um dia típico, permitindo a análise da dinâmica dos padrões de viagem e das premissas dos modelos propostos na literatura para representar tal dinâmica. O método proposto, que pode ser aplicado para qualquer tipo de SMV-RP, foi aplicado para gerar séries temporais de tráfego do Sistema de SMV-RP da cidade de Fortaleza, contribuindo para identificar dados suspeitos e atípicos, definir padrões representativos de tráfego veicular e estimar séries de fluxos OD.

1. INTRODUCTION

The road travel pattern in a city can be represented mainly by two variables: the origin-destination (OD) flows, which indicates the number of trips made between zones in a study area over a given period of day; and the traffic volume, which indicates the demand in nodes and arcs in the transportation network over a specific time interval or daily period. Both the OD flows distribution and the traffic volume magnitude represent basic information for transportation planning and design, as well as traffic management and control (Cremer and Keller, 1987).

The OD flows are the result of the trip decisions of a population aiming to carry out activities located in the urban environment. On the other hand, the traffic volumes are the result of the distribution of OD flows on the network and may be defined as the number of vehicles that travel on a road section in a certain direction over a specific time interval (Roess and McShane, 2004). OD flows require a lot of effort to be directly measured, requiring individual interviews or plate surveys. In contrast, the development of traffic monitoring systems opened up the possibility of acquiring data on traffic volumes in an automatic way at low cost (Pitombeira *et al.*, 2017). This sparked an increasing interest in indirectly estimating the OD matrix through mathematical models by using data on traffic counts on the network (Pitombeira-Neto and Loureiro, 2016; Pitombeira-Neto *et al.*, 2018).

According to Cascetta (2009), knowledge of users' travel pattern is essential for the formulation and implementation of travel demand models that help in decision-making about the system supply. One important aspect to understand in modelling the travel behavior is the day-to-day as well as daily temporal variation of the traffic volumes and OD flows (Pitombeira-Neto *et al.*, 2018). Currently, this travel dynamic can be obtained from different data sources, such as data from mobile phones (Järv *et al.*, 2014), which can generate information about the movement of mobile users, Global Positioning System (GPS) (Li *et al.*, 2004; Anda *et al.*, 2017), which provide information on vehicle movement on the road network, and smart card data (Anda *et al.*, 2017; Milne and Watling, 2019), which can help understanding the public transport demand. As stated by Pitombeira-Neto *et al.* (2018), recent approaches consider the development of day-to-day dynamic models to represent the traffic states on a network. However, there is a lack of research dealing with the issue of assessing the modelling assumptions stated about multiday dynamic of travel behavior.

In the city of Fortaleza, Brazil, there are approximately three hundred sites monitored by a Traffic Monitoring System, including an Automatic Number Plate Recognition component (TMS-ANPR). The system includes a set of sensors installed on the road network, located mainly in arterial roads, that record the passage and speed of each vehicle. The system is also equipped with video cameras and a license plate recognition system that can capture and read, through an Optical Character Recognition (OCR) algorithm, the number plates of the detected vehicles. It is worth noting the cameras and the OCR algorithm are not capable of capturing and reading the license plate of all vehicles. From this dataset, it is possible to extract the vehicle volume recorded by each equipment, as well as to associate license plate readings between equipment from different regions of the city and obtain an estimate of the OD flows between regions.

Despite its great availability, data from TMS-ANPR of Fortaleza can have some limitations for use in demand studies, depending on the number and distribution of equipment. Since the main purpose of the system is traffic enforcement, the equipment are mainly concentrated in arterial roads, which makes it impossible to observe many of the routes used by users in the central area. In addition, travel information obtained from the association of license plate readings from

two equipment does not guarantee that it is in fact an OD flow, given the small amount of equipment in certain areas. Possible equipment failures during the operation of the data collection system, both regarding vehicle detection and plate reading, should also be considered. Besides, in both day-to-day and within-day traffic variability applications it is necessary to define off-peak and peak periods, i.e., the separation of the day into periods where vehicle flow can be considered stable. In transportation planning, this definition of different periods within a day with constant traffic flow is important for understanding the day-to-day variability of traffic volumes and OD flows (Cheng *et al.*, 2012; Stathopoulos and Karlaftis, 2001). Specifically, in the analysis of day-to-day traffic dynamics, this definition is essential for the analysis of correlation between consecutive days of traffic volumes and OD flows.

In this work, we give a step forward for the understanding of multiday dynamic of travel behavior by acquiring adequate data on day-to-day traffic volumes and OD flows. Such knowledge is essential for real time traffic operations and to identify the mechanism behind travel behavior. Thus, the main goal of this paper is to propose a methodology for treatment of the TMS-ANPR data that can be used for empirical analysis of the day-to-day travel dynamic. To this end, the specific objectives are: i) to treat original TMS-ANPR data eliminating the suspicious data caused by failures in the traffic sensors, by the limitations of the license plate recognition system, and by atypical traffic variations; ii) to propose a method based on clustering analysis for defining typical periods of day with constant traffic, allowing to analyze day-to-day variation of traffic volume and OD flows; iii) to evaluate the data obtained by inferring if the probability distributions of the obtained variables fit what is stated in the literature. The main product of this data treatment are the day-to-day time series of traffic volumes and OD flows. Therefore, this treated data can then be applied to assess the multiday dynamic of travel behavior and verify the model assumptions stated in the literature about such dynamic behavior. To our knowledge, such analysis, which is out of scope of this work, has not been done yet using empirical data. The treatment method can be applied to any type of TMS system equipped with ANPR system and can be used for generating data for other purposes such as operational analysis of traffic control systems.

2. BACKGROUND

As stated by Loureiro *et al.* (2009), an efficient alternative for urban traffic management consists of implementing traffic management centers (TMCs), which collect, model, and store data relating to traffic conditions. Traffic monitoring system (TMS) is an essential component of the TMC that automatically collects and stores traffic data by loop detectors (traffic sensors) placed on the road or at intersection approaches and sends, at each second, to the management center through private phone lines.

Besides the function of collecting traffic data, the TMS can also be equipped with a ANPR system, allowing to track vehicle between different locations. According to Oliveira-Neto *et al.* (2013), these systems were developed with the main objective of interpreting the alphanumeric characters on vehicle plates without human intervention. They typically rely on four main components: an imaging acquisition processor, a license plate detection system, a character segmentation and recognition engine and a computer to store the data. The ANPR technology is a mature but imperfect technology. As stated by Oliveira-Neto *et al.* (2012, 2013), the ANPR accuracy is around or less than 60%, depending on the model, installation, variation of the license plates in the traffic stream, lighting conditions, and other factors. Exploring the fact that

the most errors made by ANPR hardware are only one or two misread characters of the vehicles plates, Oliveira-Neto *et al.* (2012, 2013) proposed a method for matching imperfect readings between two locations, even when the ANPR accuracies are unknown, increasing the number of matches or observed trips between two locations.

The data generated by TMS and APNR, or TMS-APNR, support real-time information systems, as well as assist researchers with a better understanding of travel behavior in urban network systems. The extraction of vehicular traffic information such as average speed, travel time, volume and OD flows from TMS-ANPR has been explored in the literature, as in Castillo *et al.* (2008) and Rao *et al.* (2018) who used license plate data, along with traffic volume counts on network arcs, to estimate an OD matrix, as well as in Bertini *et al.* (2005) and Liu *et al.* (2011), who used license plate data to predict vehicle travel time on the network. However, as can be seen in these studies, the focus is on using the TMS data for reconstructing OD flows and/or predict the traffic network performance. Not much detail of the previous stage of data treatment is presented and there is no concern of treating the data for the purpose of analyzing the multi-day travel behavior, which is the main purpose of this work.

The data treatment of TMS-ANPR data is an essential step before any traffic analysis and demand modelling. At this step, we seek to eliminate any bias, whether due to failures in automatic collection or when it is only possible to collect a sample for specific categories of the population. The importance of data treatment is highlighted by Oliveira and Loureiro (2006) who presented a method for data treatment of traffic data obtained from several loop detectors of the Real-Time Traffic Control System of Fortaleza, with the main objective of identifying outliers or atypical data. As pointed out by Cheng *et al.* (2012), if a prior analysis of the data is not performed any conclusion based on the analysis of the variables of interest may be misleading.

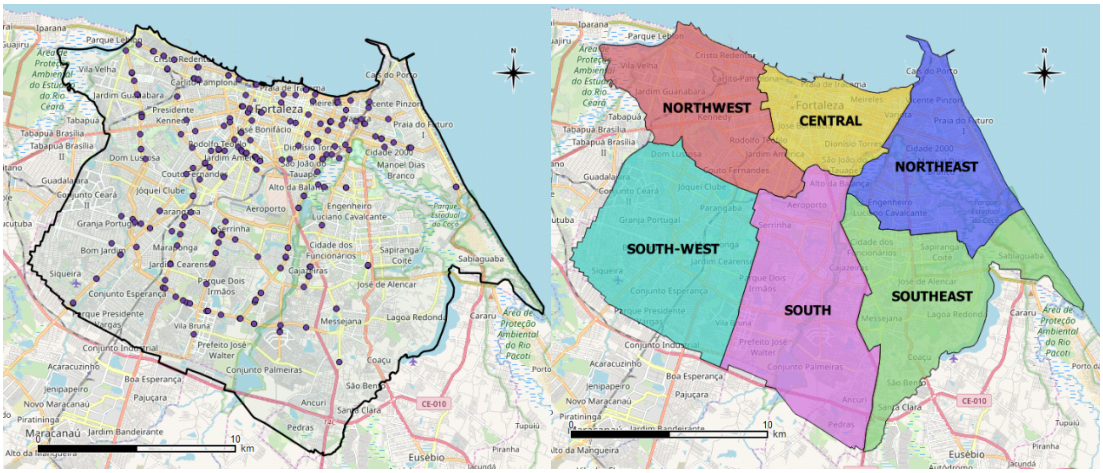
Furthermore, before any analysis after the treatment phase, it is also important to define periods in which the traffic is stable during a typical day (i.e., periods of peak and off-peak with constant traffic flow). As stated by Cheng *et al.* (2012) and Stathopoulos and Karlaftis (2001), these periods are usually defined to represent variation of traffic in urban environment, with the purpose of designing different strategies of traffic control according to the demand variation. According to Cheng *et al.* (2012) an arbitrary categorization of traffic data into different time periods is not adequate to isolate specific states of network traffic. Stathopoulos and Karlaftis (2001) highlight the importance of defining different periods with similar traffic characteristics in terms not only of traffic volumes but also related to OD flows, following the idea that commuting patterns are caused by the activities that are carried out in certain periods of day. Another important issue related to traffic patterns and the definition of the traffic states is that different traffic profiles can be observed in different locations on an urban network, resulting in different peak and off-peak periods. Some authors approached this problem by using clustering techniques. As suggested by Weijermars (2007), clustering techniques (e.g., k-means algorithm) can be used to classify traffic volume profiles and can be useful for identifying peak times according to each profile, considering that traffic may vary according to the region and trip direction.

Finally, the assumptions about probability models to represent the day-to-day traffic volumes and OD flows in urban networks is also an important issue discussed in the literature. Although this is not really a part of data treatment, a first analysis of the variables could be done to verify the hypothesis about the probability distributions stated in the literature. As stated by Pitombeira-Neto *et. al* (2017) and Pitombeira-Neto *et. al* (2018), several models

have been proposed to estimate OD flows from traffic volumes. To account for the variability on OD flows, the early models assumed that OD flows are the result of a Poisson process (Vardi, 1996; Tebaldi and West, 1998). Hazelton (2000, 2001, and 2003) proposed to approximate the distribution of OD flows with a multivariate normal density, which has more tractable computational properties. In this latter case, the traffic volumes can be also modeled by a multivariate normal distribution, and each traffic volume for a given arc modeled by a normal distribution. All these earlier works have in common the main assumption that the OD flows are the result of a stationary process in which the mean OD and variances are constant day after day at the same period of day. Pitombeira-Neto and Loureiro (2016) and Pitombeira-Neto *et al.* (2018 and 2020) also have suggested the normal distribution as approximation for the Poisson process, but they relaxed the assumption that the OD flows, and consequently the traffic volumes, are independently distributed random variables, by assuming some dynamic structure for day-to-day variation. We state in this paper that for the case of weak dependence (i.e., the state of the system – represented in terms of route flows or route costs – does not depend strongly on the previous history of the system's states), it is possible to verify the hypothesis of normality.

3. METHODS

The TMS-ANPR data used in this study was collected in the city of Fortaleza, Brazil, in 2017. Three companies are responsible for the TMS-ANPR equipment. The system was designed to monitor drivers who exceed the speed limit or cross the red light. A total of 358 sites (including intersections and middle block) on the road network were monitored in 2017 (Figure 1a). Comma-Separated Values (CSV) files of data collected by each company are provided by the Municipal Traffic and Citizenship Authority (AMC). The files contain for each vehicle detected a record with the following information: equipment identification code, date and time of detection, lane in which the vehicle was detected, the speed limit of the road site, the measured speed, an estimated vehicle size in meters, an estimated vehicle classification and the reading of the detected vehicle license plate. The vehicle license plate is encrypted for privacy purposes.



(a) Equipment sites (b) Study regions

Figure 1. Equipment sites and study regions

The study area, Fortaleza urban network, was divided in six regions, as shown in Figure 1b. The regions were constructed originally from the census tracts, by joining tracts with similar employment and demographic characteristics. For more details, see Lima (2017).

The result was a central region, with a mixed type of land use (i.e., residential, and commercial uses) and five others peripherical regions, with mostly residential characteristic, but with distinct levels of income. The southeast region was excluded from the analysis due the small number of devices in this region. Therefore, this definition of distinct regions, with respect to socioeconomic characteristics and with a coarse level of spatial aggregation, allows not only to obtain the OD flows between regions from the TMS-ANPR data, but mainly to analyze the day-to-day variability of OD flows relating this dynamic with the context of the urban environment (i.e., the land use and socioeconomic characteristics of the different regions). Regarding the OD flows from TMS-ANPR data, it is believed that at this aggregation level most of traffic flows between two devices located at two different regions are origin-destination trips between those two regions.

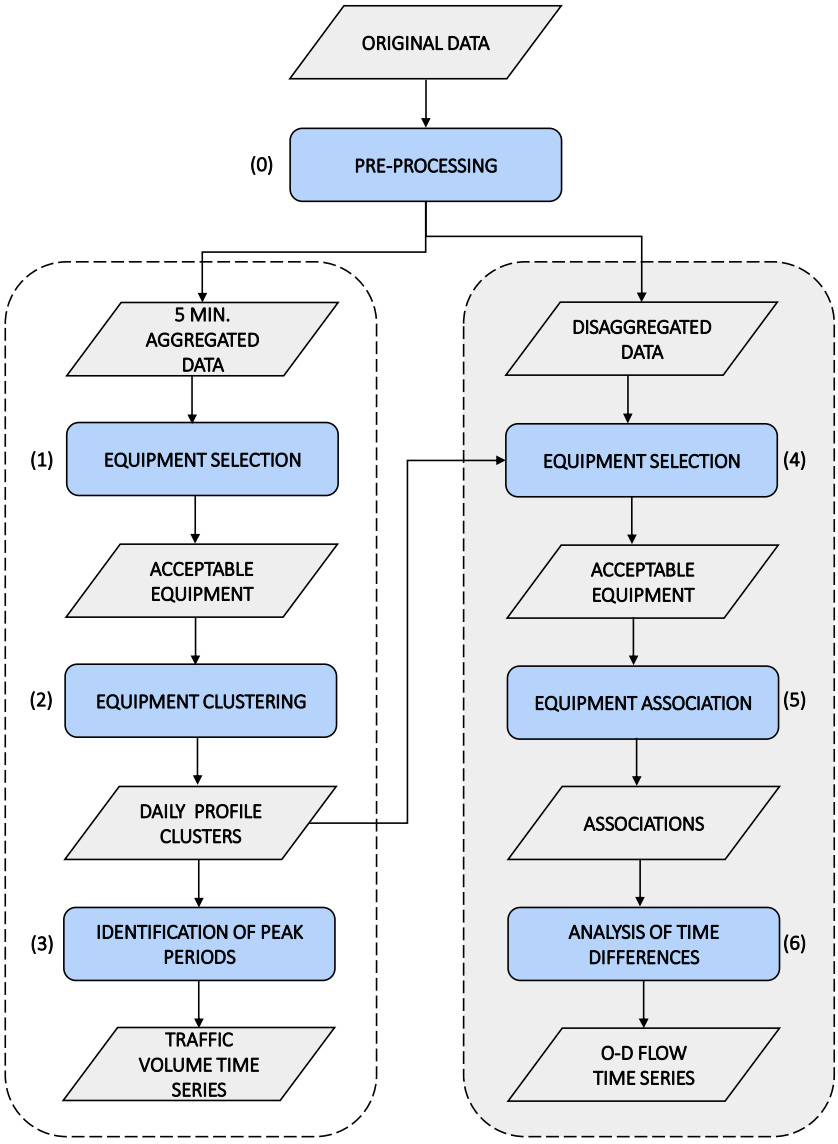


Figure 2. Proposed method for data treatment

Figure 2 shows the steps of the proposed method of the TMS-ANPR data treatment. The 2017 data was treated to extract reliable time series of values for the two variables of interest for this study: day-to-day traffic volumes (vehicles/hour) and day-to-day OD flows between regions for

the main periods, peak and off-peak, of typical days (excluding holidays and weekends). The data was treated according to the following steps: equipment selection for the analysis of traffic volumes, definition of peak and off-peak periods of analysis, selection of devices for the analysis of OD flows, detection of outliers or anomalies on the time series of traffic volumes and OD flows. The first step was to organize the data by each company into a single file by day. Next, the data was aggregated at 5-minute intervals. This time interval was set to be short enough to represent the daily variation of traffic flows, allowing to identify the different traffic states during a typical day, and large enough to make it possible to detect any missing or faulty data in the dataset.

3.1. Day-to-day traffic volume extraction

3.1.1. Equipment selection

The equipment selection step (1) aims to ensure the selection of equipment that has an operation considered acceptable to obtain the data of interest during the analysis period. Initially, it was selected only the equipment that worked over all months along a given year. Besides this criterion, it was also considered as a selection criterion the proportion of 5-minute intervals between 5:00 a.m. and 10:00 p.m. with non-zero traffic volume. Since most of the surveillance equipment are in arterial roads of the city, it is expected, within the defined period for a given day, that all intervals have traffic volume greater than zero. Any interval with zero traffic volume would be classified as a faulty data or missing data.

Based on this proportion, a day of traffic observations was classified as either acceptable or not acceptable. The threshold defined to classify any day of observations into these two groups was defined based on a trade-off between the number of days of traffic volume necessary to describe the day-to-day variation of traffic and the number of 5-minute intervals necessary to describe the within-day traffic variation. To this end, a sensitivity analysis was performed to assess the effect of the proportion of non-zero intervals on the number of acceptable days for each equipment. This analysis was also performed by different time intervals of aggregation (10 min, 15 min and 30 min). Finally, the data for a given equipment was considered acceptable for analysis when the proportion of acceptable days along a given year was greater than 90%, assuming that with this proportion of valid days would be possible to analyze the day-to-day variation of traffic for a given year of data.

3.1.2. Traffic volume profile identification

The next step refers to equipment clustering (2) according to their relative average traffic volume profile. The traffic profile for each equipment is defined by calculating the average volume at each 5-minute intervals over a given number of typical days, corresponding to a week, a month, or a year. The clustering analysis was performed by applying the k-means algorithm as suggested by Song *et al.* (2019). The k-means technique defines groups based on a predefined number of clusters and a measure of similarity between each element to be grouped. There are different techniques of clustering in the literature, so the k-means was chosen assuming that some types of traffic profiles in an urban network are usually expected given that the activities are usually concentrated in certain areas of the city and take place at specific hours along the day. Therefore, the number of clusters could be previously defined based on the knowledge of the temporal and spatial distribution of activities.

The similarity measure adopted to compare the traffic volume profiles between every two sites was the Euclidean distance between the profiles. Since the main goal was to identify groups with similar shape, instead of using the 288 absolute values (corresponding to the number of 5-minute intervals in a day) of the traffic volume profiles, the relative values (fraction of the total daily traffic volume) at each 5-minute interval throughout the day was considered. To define the number of clusters, the average silhouette width criterion was adopted. This criterion seeks to maximize the separability (variation between groups) and compactness (variation within each group) of the clusters. Finally, to validate this number of clusters, a sensitivity analysis was performed by varying the number of clusters around this initial value, and visually analyzing the average profile for each group, as well as the spatial distribution of the equipment groups in the network. The final number of clusters was chosen to represent the different variations of traffic volume according to the road site of the equipment in the network and the direction of the monitored traffic.

The k-means method was compared to the Functional Clustering method (Jacques and Preda, 2014), which is a technique designed for data generated by a process that occur on continuous space (e.g., continuous time space). A probabilistic approach was adopted in this case that consists in assuming a density probability on a finite number of parameters describing the profiles for each cluster. However, the type of profiles obtained by this technique were not much different than the profiles using the k-means. Therefore, the k-means was chosen due to its simplicity.

3.1.3. Peak and off-Peak periods identification

The identification of peak periods (3) is performed for each profile identified in the previous step, aiming to define periods of the day when the vehicle volume can be considered stable, also including the identification of an off-peak period. To this end, it was applied the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering algorithm proposed by Ester *et al.* (1996), which is suitable for identifying arbitrary clusters of different sizes and identifying elements that do not belong to any cluster, called noise, without the need to provide preliminary information about the groups.

The DBSCAN technique is based on spatial density of elements, so points that are tightly packed together are grouped, while elements that lie on low density regions or regions with high variation are classified as noise elements or outliers. The use of this algorithm instead of k-means was due to its capability to identifying the different states of traffic for a given traffic profile, corresponding to different densities of traffic volumes by time of day, and also to identifying the transitions between every two states of traffic, that we defined as periods of traffic with high variation of intensity. In sum, DBSCAN is best suited for generating groups whose elements have similar intensity, while k-means generates groups whose elements have similar variation trends. After this clustering analysis, the day-to-day series of traffic volume for each daily period can be generated.

Basically two global parameters are required for the execution of the DBSCAN algorithm: the radius of the search circle, ε, so every two points located within this radius are said to belong to the same group; and the minimum number of points, MinPts, that are considered to form a cluster within the radius ε. Points belong to same group if they are either inside the search circle formed by at least MinPts points or they can be reached by any point of a set of points inside another search circle with at least MinPts points. Points that are not packed together (those that do not fit to any of these criteria) are classified as noise points.

3.2. Day-to-day OD flow extraction

3.2.1. Equipment selection and daily periods of analysis

The next step (4) concerns to the selection of suitable equipment for extracting the OD flow from the sample of equipment previously selected for volume extraction. As pointed out in Sections 1 and 2, the TMS-ANPR has the limitation of not registering all license plates of the vehicles detected and not covering the entire road network. However, since most of arterial roads are monitored (Figure 1b), it is believed that a set of the equipment can be selected to provide data of the day-to-day flow variation between urban areas, for specific time periods of a day. The equipment suitable for this analysis were selected based on the proportion of plates read by each equipment (reading rates).

The reading rates for each equipment are calculated for different daily periods of analysis (i.e., morning, midday, and afternoon), which are determined based on the traffic states defined in Subsection 3.1.3. To determine the daily periods for OD flow analysis, the average traffic profiles of every two regions are associated to define time periods that include the same state of traffic observed at the two different regions (e.g., the morning period for OD flow observation between a peripherical and a central region includes both morning peaks for those two regions).

The equipment selection criterion for each daily period defined was based on the standard deviation of reading rates for a given year of observation. Therefore, for each equipment a maximum threshold of 0.12 for the standard deviation of the reading rates was set by seeking to select equipment with small variation of the reading rates (reducing the effect of the reading proportion on the observed day-to-day traffic pattern between regions) but keeping a minimum number of equipment for each region that allows to analyze day-to-day variation of OD flow.

3.2.2. License plate association between regions

The equipment association step (5) is simply the identification of trips made by vehicles between regions during the daily periods of analysis. Since the APNR accuracy is unknown, trips between two regions are identified by counting the number of exact matches between plate readings within the adopted analysis period. It is assumed that two exact license plate readings are unlikely to come from different vehicles. To avoid the possibility of replications (when vehicles are registered on more than two equipment at a sequence), the first reading record is associated with the last exact reading record, eliminating intermediate records of the same vehicle.

The great concern here at this point of treatment is how to represent the pattern OD flow between city regions through the flow of passage observed by associating plate readings of equipment located at different regions. Firstly, we already argued that each exact match of reading plates between regions is likely to represent a trip between the same regions, since the adopted zones are large areas representing different land use and socioeconomic characteristics in the city. Secondly, although the day-to-day series of OD matrices obtained from the proposed process may not represent the whole OD flow patterns among all regions in absolute terms, it allows to analyze the day-to-day OD flow variation for each pair of regions, which is the main goal of this data treatment. Finally, as we will see at the Subsection 3.2.3, the difference in time between reading associations was also used as criterion to define a trip based on expected travel times between regions.

3.2.3. Travel time filter

The last step (6) of the method concerns the analysis of the time differences between reading associations. The analysis of the distribution of the resulting time differences allows to eliminate associations with very short time differences, possibly corresponding to associations between very closely equipment located on the same arterial street that tends to generate many observations (which are likely to represent a traffic volume associated with any pair of regions), leading to a misrepresentation of the OD flows. Further, the observed time differences are compared to the travel times obtained by the Open Street Map (OSM) platform through the 'osrm' package developed by Giraud (2018) which is part of the R software.

Thus, the time differences were filtered according to the following criterion: $0.85\times t_{osm} < \Delta t < 3\times t_{osm}$, where Δt is the time difference between reading associations, and t_{osm} is the corresponding time obtained by OSM. Any time difference outside this interval is assumed to be unreliable, since it is likely to be originated from an association of misreading plates (i.e., association between very distant equipment but with very short time difference, or association between very closely equipment, located few blocks away at the same road, but with long time difference). The upper bound of $3\times t_{osm}$ of the time provided by the OSM was set to consider the possibility of intermediate stops. The lower limit of 85% was mainly defined to incorporate the possibility of synchronization failures between equipment, which may result in shorter time differences due to small differences in time clock.

3.3. Outlier detection in the time series and probability distribution of the traffic variables

After generating the initial time series of day-to-day traffic volumes and day-to-day OD flows, an analysis of outliers is performed to identify atypical values that were not detected by the previous analyzes. Atypical variations in the data can occur due to accidents, weather conditions, among other reasons. Since the main interest lies on generating traffic flow series that represent typical day-to-day variations of traffic, such atypical events should be left out. The criterion adopted was to consider an extreme value, or outlier, an observation outside the following range (see distance-based methods in Aggarwal, 2015): $[\tilde{x} - 3s, \tilde{x} + 3s]$, where $\tilde{x}$ is the median and s is the sample standard deviation for a sample obtained using a time window of 20 days. This 20-day time window was defined to estimate the data dispersion and to represent the local pattern of the data, avoiding eliminating any observation due to variation between months. A shorter time window (e.g., 7 days) could be considered, but in this case the sample size would be smaller.

To analyze which probability distribution may represent the variation of the time series obtained, the histograms of the treated data considering only the typical months (without January, July and December) of the year were compared with the Normal distribution, by applying the Shapiro-Wilk test. In fact, it is believed that the values of traffic variables fluctuate around a central measure, since the series are generated from stable periods of traffic throughout the day, or periods of commuting, where the traffic flows tend to repeat day by day.

4. APPLICATION TO THE URBAN NETWORK OF FORTALEZA, BRAZIL

This section presents the results obtained from the application of the proposed method for treatment of the 2017 TMS-ANPR data in Fortaleza, Brazil.

4.1. Generating the day-to-day traffic volumes

From the total of 358 equipment, 271 devices with detection data for all months were initially selected. The number of equipment with acceptable days was defined by varying the threshold for the proportion of non-zero 5-minute intervals from 60% to 100%, as shown in Figure 3. It was found that for a proportion up to 90%, around 200 days were classified as acceptable and that above this proportion the number of acceptable days decreases considerably, as compared with other time intervals of 10 minutes, 15 minutes and 30 minutes. Hence, a threshold of 90% for the data aggregated at 5-munute intervals was adopted, which corresponds to a balance between desirable sample size (number of days) and data quality (data representing the daily variation of traffic flow). With the indication of acceptable days for each equipment, the second criterion was applied regarding the minimum proportion of acceptable days (at least 90% of the total workdays present in the sample), resulting in a sample of 179 equipment, representing about 50% of the initial number.

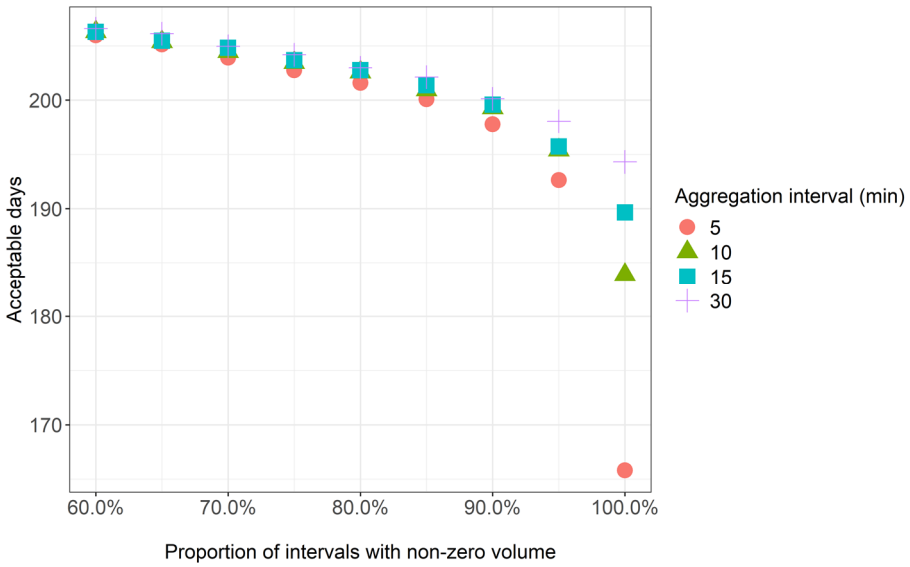


Figure 3. Effect of the proportion of time intervals with non-zero volume on the number of acceptable days

4.1.1. Traffic volume variation patterns

As for the definition of daily traffic profiles, the k-means method was applied to determine the daily traffic profiles for typical and atypical months (January, July, and December), and for different weekdays, to account for possible seasonal variation at the daily traffic profiles. The method suggested three traffic patterns to represent the different traffic dynamics existing in the city, as shown in Figure 4, which shows the average traffic profiles for typical and atypical months. As can be seen in Figure 4, these 3 profiles represent three main traffic patterns corresponding to the mainly flow directions of traffic in the city of Fortaleza: with links of intense traffic going towards the central region in the morning (profile 3), links of intense traffic going outwards the central region in the late afternoon (profile 1), and some locations with two less intense peaks and an off-peak period of less intensity between them (profile 2). The results did not show evidence of differences between the traffic profiles of the weekdays. According to Figure 4, it was observed a change of traffic pattern between atypical and typical moths. Several traffic segments of profile 2 and 3 switched to profile 1 during atypical months.

This pattern changing probably reflects the change in daily activities during vacation months, which usually have less intense traffic during the morning period and higher intense traffic during the afternoon period.

Regarding the typical months, the spatial locations, together with the traffic directions, of the equipment for each group reveal these three traffic patterns detected by the clustering analysis. Equipment of type 3 profile are located mostly in arterial roads within residential regions and whose direction is predominantly toward the central region (suburb to central direction), while profile 1 equipment are also mostly located closed to residential areas of the city usually in the opposite direction, revealing the predominantly commuting characteristic of travel, moving towards the center early in the day and returning to residence at the end of the day. Profile 2 equipment indicates locations where traffic generally does not change significantly as a function of the time of day and are usually on major roads in the central area, with no dominant direction of traffic, linking either commercial or residential neighborhoods, which may account for two well-defined peaks beyond the plateau between them. Therefore, the results revealed traffic patterns that are consequence of an urban environment in which most of the activities are concentrated on a single region or central region of the city, generating intense traffic in certain locations and periods of the day.

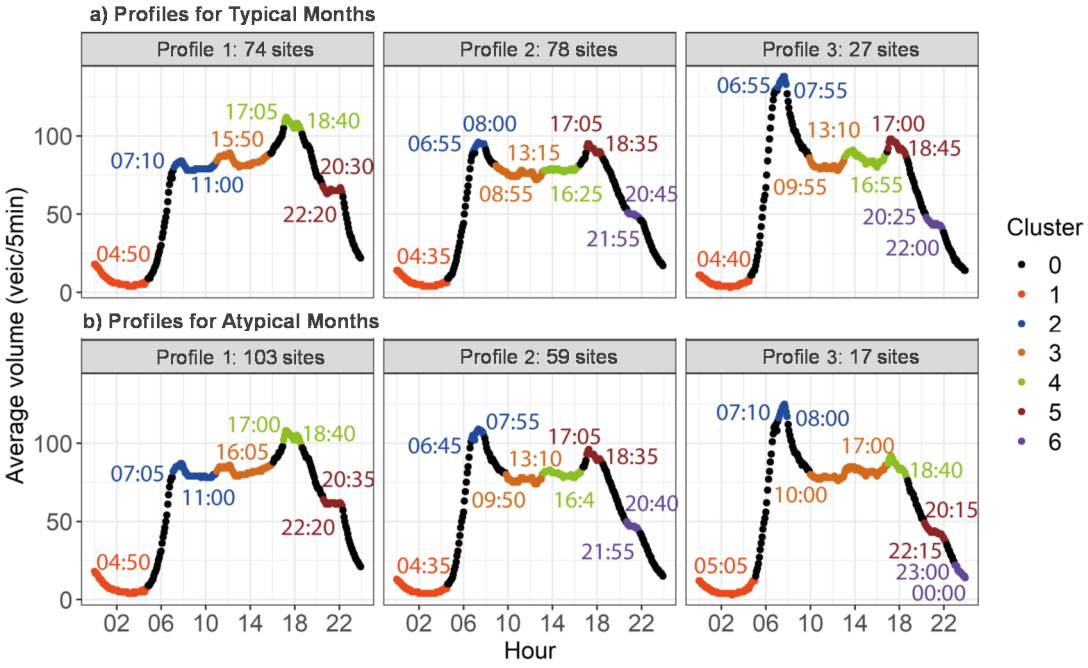


Figure 4. Average daily traffic profiles and the time periods for typical and atypical months

From the identified profiles, the DBSCAN parameters were defined by a sensitivity analysis based on two criteria: the silhouette width and the number of traffic periods. The quality of the clusters is better for higher values of the silhouette width. Figure 5 shows this analysis for the profile 2 of typical months. The silhouette indicator was calculated only for observations that were classified at any group (excluding the noises). Since the daily traffic profile is quite variable, short values for both MinPts and *e* result in too many groups or even no group (only noise). As the parameters increase the DBCAN algorithm tends to identify only one single group. Hence, the criterion of the number of groups was used to define a set of parameters that better represent the daily variation of traffic at each profile.

For the case shown in Figure 5, the adopted parameters were: MinPts = 8 intervals of 5 minutes and ε = 14. The MinPts of 8 corresponds to a period of at least 40 minutes for each cluster. As for ε = 14, it is the maximum search radius obtained by observing a cumulative distance graph of the 8 (defined MinPts) nearest neighbors. Recall that the value of *e* has no meaning since the distances between observations are calculated from both attributes of volume (vehicles/5 minutes) and time (minutes). The same analysis was done for other traffic profiles. The clusters identified for each profile, are presented in Figure 4. The observations attributed to cluster 0 are the noises and, as expected, correspond to those transition intervals between stable periods of traffic.

The traffic peaks identified revealed that the size and intensity of the morning peak periods are quite different for the three profiles, with the profile 3 presenting the most intense peak, between 6:55 a.m. and 8:00 a.m. The time intervals of the late afternoon peaks, mostly between 5:00 p.m. and 7:00 p.m., are similar for the three profiles, with the profile 1 presenting the sharpest peak. It is worth noting that only profile 1 has a well-defined midday peak, between 11:00 a.m. and 4:00 p.m., probably representing trips mainly for lunch purposes. Profile 3, on the other hand, presents a well-defined peak between 1:00 p.m. to 5:00 p.m., perhaps representing trips mainly for non-work purposes (e.g., health and shopping), since the equipment of this group are in the suburb areas and monitor the traffic going toward the central region. Finally, it is worth noting the profile 2 do not present any sharp peak, as observed in the other two profiles, and present a long period of low traffic variation between 8:00 a.m. and 4:40 p.m., probably representing trips of different purposes having the central area as either origin or destination. The major difference between atypical and typical months is that the profile 3 for atypical months present not much variation for the period between 10 a.m. to 5 p.m.

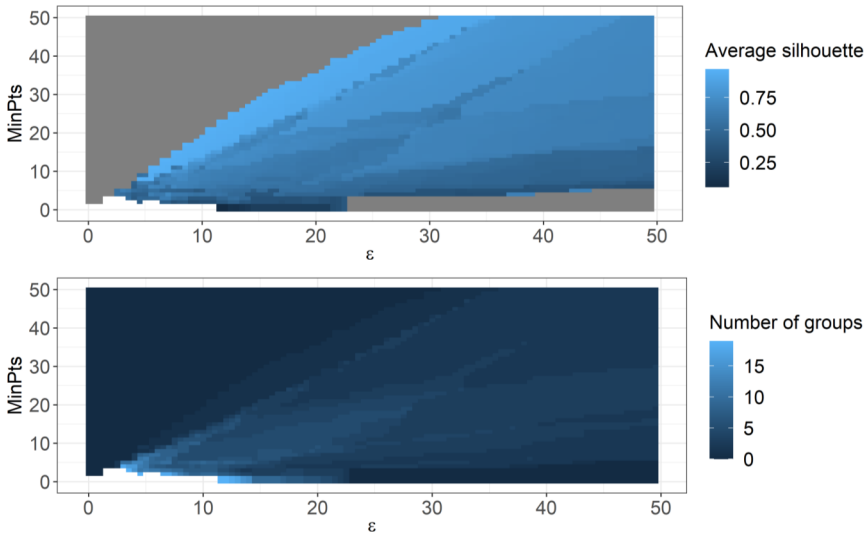


Figure 5. Sensitivity analysis of DBSCAN parameters for the traffic profile 2 of the typical months

4.1.2. Day-to-day time series of traffic volumes

The 179 time series of traffic volumes were generated for each daily period. Three time periods were assumed for the analysis for each profile: morning peak, afternoon peak and a midday period. The midday period was defined to be between 1:00 p.m. to 5:00 pm for all profiles. Figure 6 shows the result of the time series for the morning peak of a given

equipment, classified as profile 3, along with the timeline trend, the identified outliers, and the histogram of the variable. It was observed for the most series of traffic volumes that the atypical months (January, July, and December) presented less intense traffic.

For the histogram in Figure 6, the Shapiro-Wilk normality test, with p-value = 0.52, showed evidence that the distribution of the traffic volume variable does not differ significantly from the normal distribution. The test was performed for all series of traffic volumes, considering only the typical months, yielding for most samples to no rejection of the null hypothesis of normality, at a significance level of 5%. It is worth to say that the traffic volume presented over-dispersion when compared to that expected by a Poisson variable. This indicates that the traffic volumes are not completely random, since they are the result of user decisions (trip or route decisions), that in turn are influenced by the temporal and spatial distribution of activities on the city and probably by the previous history of travel decisions.

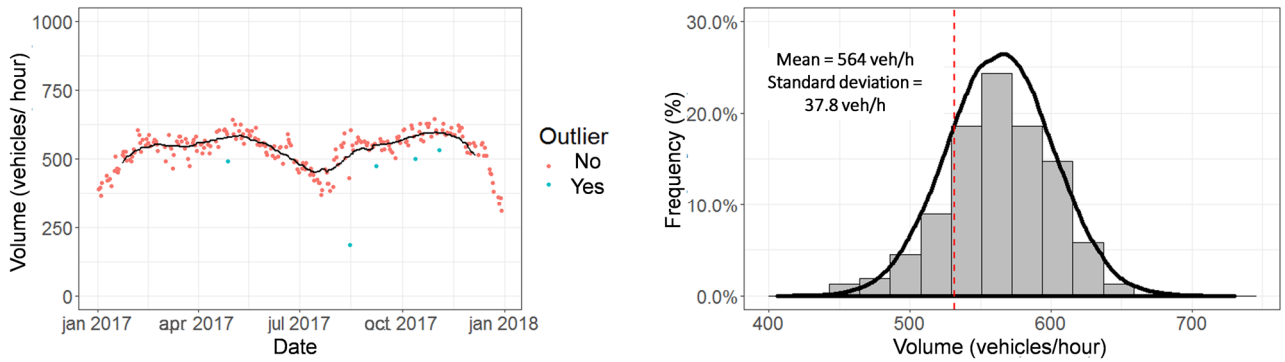


Figure 6. Time series and histogram of the traffic volume for the morning peak

4.2. Generating the day-to-day OD flows

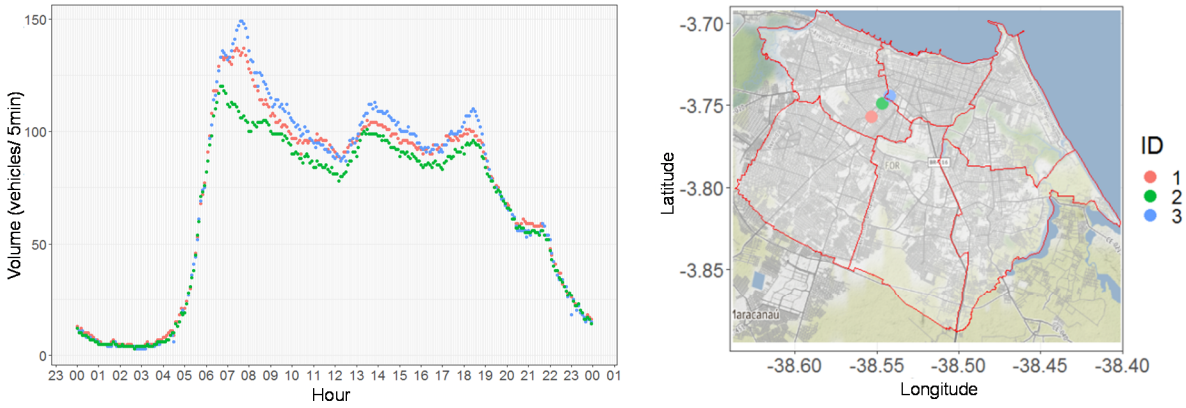
Regarding the daily periods for OD flow analysis, Figure 7a illustrates how the daily time period for an OD pair was defined. The figure shows a sequence of hourly traffic volume profiles of three devices located on a roadway connecting the northwest region to central region. As can be seen the most predominant profile between these two regions is profile 3, which was used to define the time periods for the corresponding OD flow analysis.

For each daily period (morning, afternoon, and midday periods), the analysis of the reading rates variation along the 2017 year resulted in 106 devices suitable for license plate association. Figures 7b and 7c show the variation in the proportion of readings throughout 2017 for one equipment that was selected and another that was rejected, respectively.

After performing the license plate association between regions and the travel time treatment based on the OSM tool, the day-to-day OD flows for each daily period were generated. Figure 8 shows the time series for the morning peak of the OD flow between the northwest and central regions, along with the identified outliers, the timeline trend, and the histogram of the variable. Like Figure 6a, it was observed for the most series of OD flows that the atypical months (January, July, and December) presented less trips.

The Shapiro-Wilk's normality test, with p-value = 0.19, showed evidence, at 5% significance level, that the OD flow distribution does not differ significantly from the normal distribution. The test was performed for all OD pairs, considering only typical months, and the null

hypothesis of normality could not be rejected at most tests, at a significance level of 5%. As with the traffic volumes, it was observed an overdispersion for the OD flows, even higher than the variation of the traffic volumes. This is not only a result of the variation of the OD flow along the year, but also of the variation in reading rates and of the spatial distribution of the selected equipment. Considering that the detected trips were random sampled, it is possible to analyze the dynamic of OD flow between regions looking at the generated time series of day-to-day OD flow. This analysis is out of the scope of this paper.

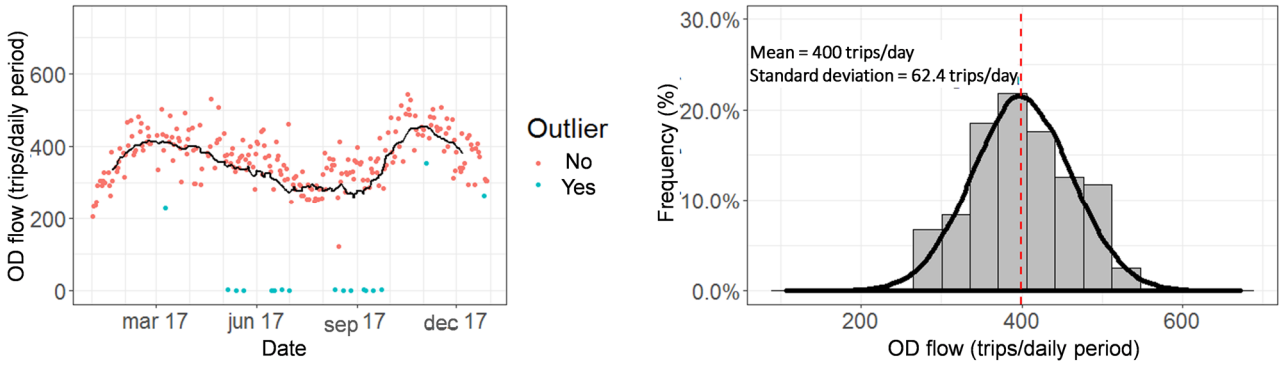


(a) Sequence of hourly traffic volume profiles of type 3

(b) Reading rates for a selected equipment

(c) Reading rates for a non-selected equipment

Figure 7. Sequence of traffic volume profiles and reading rates



(a) Series of day-to-day OD flow

(b) Histogram of OD flow for typical days

Figure 8. Time series and histogram of the traffic volume for the morning peak

5. CONCLUSIONS

This work presented a methodology for TMS-ANPR data treatment for generating day-to-day time series of traffic volumes and OD flows in urban networks. The study contributes to the use of data from TMS systems for the generation of time series that allow evaluating the variability of urban traffic, thus supporting studies that aim to empirically verify theoretical assumptions about day-to-day assignment methods and OD matrix estimation models, such as the probability distributions adopted and the temporal correlation of the variables.

As discussed so far, the process of handling large volumes of data from automatic collection systems is not only essential for reliable analysis, but also it is the first step in understanding the dynamics of traffic in an urban environment. In other words, in addition to the cleaning (eliminating failures and anomalies in the data) and organization procedures, the treatment stage can also include the definition of stable traffic periods and a preliminary descriptive analysis of the resulting data allowing to identify traffic patterns that will assist in raising hypotheses to be tested about the phenomenon of interest. Such hypotheses may include probability distribution of the variables, seasonal effects (monthly, weekly, and daily difference) and difference of traffic patterns between regions of the urban network. Specifically, this work contributed to the definition and analysis of traffic variation patterns, applying clustering techniques, which in the case of Fortaleza-CE revealed the tendency of commuting towards the central region of the city, where most commercial and service activities are concentrated.

The study also contributed to the use of TMS-ANPR data to generate OD flows in urban areas, which usually requires a great effort to collect. Such data allows to analyze the day-to-day variability of OD flows that is essential for the urban planning of major cities. Therefore, we stress the effort in this study to adopt a division of study area in regions representing different land use and socioeconomic characteristics and that allow to obtain an adequate sample of OD trips, to select a set of equipment for analysis that worked properly most of the year (i.e., considered reliable according to the proportion of non-zero traffic observed by 5-minute intervals) and with low variation of reading rates, and to treat the obtained OD flows based on travel time expected from the OSM tool avoiding some bias in the data.

One limitation of this study was that several equipment was not suitable for generating traffic volume and OD flow series, probably due to failures (e.g., lack of synchronization between equipment, plate reading failure and failure of vehicle registration). Another limitation of the study concerns to the distribution of equipment in the analyzed area, with some regions having few equipment. The effect of the spatial distribution of the TMS-ANPR equipment (installed for enforcement purposes) on the network is an important issue for further studies. Furthermore, the accuracy of the ANPR system of Fortaleza is unknown. This issue has been addressed in previous studies that proposed methods to improve the matching of imperfect license plate readings, even when the ANPR accuracy is unknown (Oliveira *et al.*, 2012 and 2013). Such techniques can be incorporated into the proposed method and should be a subject for future work. Finally, the method of data treatment can support future studies about the influence of the network performance on the multiday dynamic of traffic volumes and OD flows.

ACKNOWLEDGEMENTS

This Study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – and the Conselho Nacional de Desenvolvimento Científico e Tecnológico – Brasil (CNPq).

REFERENCES

Aggarwal, C. C. (2015) *Data mining: the texbook*. Springer International Publishing, Switzerland. DOI: 10.1007/978-3-319-14142-8.

Anda, C.; A. Erath and P. J. Fourie (2017) Transport modelling in the age of big data. *International Journal of Urban Sciences*, v. 21, p. 19–42. DOI: 10.1080/12265934.2017.1281150.

Bertini, R. L.; M. Lasky and C. M. Monsere (2005) Validating predicted rural corridor travel times from an automated license plate recognition system: Oregon's frontier project. *IEEE Conference on Intelligent Transportation Systems, Proceedings*, ITSC, v. 2005, p. 706–711. DOI: 10.1109/ITSC.2005.1520134.

Cascetta, E. (2009) *Transportation System Analysis: Models and applications*. Ed. Springer (2ª ed.), New York, USA.

Castillo, E.; J. M. Mennéndez and P. Jimenez (2008) Trip matrix and path flow reconstruction and estimation based on plate scanning and link observations. *Transportation Research Part B: Methodological*, v. 42, n. 5, p. 455–481. DOI: 10.1016/j.trb.2007.09.004.

Cheng, T.; J. Haworth and J. Wang (2012) Spatio-temporal autocorrelation of road network data. *Journal of Geographical Systems*, v. 14, n. 4, p. 389-413. DOI: 10.1007/s10109-011-0149-5.

Cremer, M. and H. Keller (1987) A New Class of Dynamic Methods for the Identification of Origin-Destination Flows. *Transportation Research Part B: Methodological* , v. 21, n. 2, p. 117–132. DOI: 10.1016/0191-2615(87)90011-7.

Ester, M.; H. P. Kriegel; J. Sander and X. Xu (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*. AAAI Press, Portland, Oregon, USA, p. 226–231.

Giraud, T. (2018) OSRM: Interface Between R and the OpenStreetMap-Based Routing Service OSRM. R package version 3.1.1. Available in: <https://cran.r-project.org/web/packages/osrm/index.html> (consulted on 08/18/2021).

Hazelton, M. L. (2000) Estimation of Origin-Destination Matrices from Link Flows on Uncongested Networks. *Transportation Research Part B: Methodological*, v. 34, n. 7, p. 549–566. DOI: 10.1016/S0191-2615(99)00037-5.

Hazelton, M. L. (2001) Inference for Origin-Destination Matrices: Estimation, Prediction and Reconstruction. *Transportation Research Part B: Methodological*, v. 35, n. 7, p. 667–676. DOI: 10.1016/S0191-2615(00)00009-6.

Hazelton, M. L. (2003) Some comments on origin–destination matrix estimation. *Transportation Research Part A: Policy and Practice*, v. 37, p. 811–822. DOI: 10.1016/S0965-8564(03)00044-2.

Jacques, J. and C. Preda (2014) Functional data clustering: A survey. *Advances in Data Analysis and Classification*, v. 8, p. 231–255. DOI: 10.1007/s11634-013-0158-y.

Järv, O.; R. Ahas and F. Witlox (2014) Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records. *Transportation Research Part C: Emerging* Technologies, v. 38, p. 122–135. DOI: 10.1016/j.trc.2013.11.003.

Li, H.; R. Guensler; J. Ogle and J. Wang (2004) Using Global Positioning System Data to Understand Day-to-Day Dynamics of Morning Commute Behavior. *Transportation Research Record: Journal of the Transportation Research Board*, v. 1895, p. 78–84. DOI: 10.3141/1895-11.

Lima, L. S. (2017) *Espraiamento Urbano por Autossegregação e seus Impactos na Acessibilidade Urbana de Fortaleza.* Dissertação de Mestrado – Programa de Pós-Graduação em Engenharia de Transportes – PETRAN, Departamento de Engenharia de Transportes, Universidade Federal do Ceará, Fortaleza, Brasil, 2017. Disponível em: <http://www.repositorio.ufc.br/handle/riufc/30015> (acesso em 18/08/2021).

Liu, G.; Z. Ma; Z. Du and C. Wen (2011) The Calculation Method of Road Travel Time Based on License Plate Recognition Technology. *Communications in Computer and Information Science*. p. 385–389. DOI: 10.1007/978-3-642-22418-8_54.

Loureiro, C. F. G.; H. B. Meneses; F. M. Oliveira-Neto and M. M. Castro-Neto (2009) Managing Congestion in Large Brazilian Urban Area through Logical Interface between SCOOT and GIS Platform. *Transportation Research Record: Journal of the Transportation Research Board*, v. 2099, p. 76–84. DOI: 10.3141/2099-09.

Milne, D. and D. Watling (2019) Big data and understanding change in the context of planning transport systems. *Journal of Transport Geography*, v. 76, p. 235–244. DOI: 10.1016/j.jtrangeo.2017.11.004.

Oliveira, M. V. T. e C. F. G. Loureiro (2006) Análise dos Padrões de Variação Espaço-Temporal do Volume Veicular no Ambiente Urbano de Fortaleza. *Anais do XX Congresso de Pesquisa e Ensino em Transportes, ANPET*, Brasília, v. 1, p. 149–161.

Oliveira-Neto, F. M.; L. D. Han and M. K. Jeong (2012) Online license plate matching procedures using license-plate recognition machines and new weighted edit distance. *Transportation Research Part C: Emerging Technologies*, v. 21, p. 306–320. DOI: 10.1016/j.trc.2011.11.003.

Oliveira-Neto, F. M., L. D. Han and M. K. Jeong (2013) An online self-learning algorithm for license plate matching. *IEEE Transactions on Intelligent Transportation* Systems, v. 14, p. 1806–1816. DOI: 10.1109/TITS.2013.2270107.

Pitombeira-Neto, A. R. and C. F. G. Loureiro (2016) A Dynamic Linear Model for the Estimation of Time-Varying Origin–Destination Matrices from Link Counts. *Journal of Advanced Transportation*, v. 50, n. 8, p. 2116-2129. DOI: 10.1002/atr.1449.

Pitombeira-Neto, A. R.; C. F. G. Loureiro and L. E. Carvalho (2018) Bayesian Inference on Dynamic Linear Models of Day-to-Day Origin-Destination Flows in Transportation Networks. *Urban Science*. v. 2, n. 4, p. 117. DOI: 10.3390/urbansci2040117.

Pitombeira-Neto, A. R.; F. M. Oliveira-Neto and C. F. G. Loureiro (2017) Statistical models for the estimation of the origin-destination matrix from traffic counts. *Transportes (Rio de Janeiro)*, v. 25, p. 1-13. DOI: 10.14295/transportes.v25i4.1344.

Pitombeira-Neto, A. R.; C. F. G. Loureiro and L. E. Carvalho (2020) A Dynamic Hierarchical Bayesian Model for the Estimation of Day-to-Day Origin-Destination Flows in Transportation Networks. *Networks and Spatial Economics*. v. 20, n. 2, p. 499–527. DOI: 10.1007/s11067-019-09490-5.

Rao, W.; Y.-J. Wu; J. Xia; J. Ou and R. Kluger (2018) Origin-destination pattern estimation based on trajectory reconstruction using automatic license plate recognition data. *Transportation Research Part C: Emerging Technologies*, v. 95, p. 29–46. DOI: 10.1016/j.trc.2018.07.002.

Roess, R. P. and W. R. McShane (2004) *Traffic Engegineering*. Ed. Pearson/Prentice-Hall, New Jersey, USA.

Song, J.; C. Zhao; S. Zhong; T. A. S. Nielsen and A. V. Prishchepov (2019) Mapping Spatio-Temporal Patterns and Detecting the Factors of Traffic Congestion with Multi-Source Data Fusion and Mining Techniques. *Computers, Environment and Urban Systems*, v. 77, p. 101364. DOI: 10.1016/j.compenvurbsys.2019.101364.

Stathopoulos, A. and M. G. Karlaftis (2001) Temporal and Spatial Variations of Real-Time Traffic Data in Urban Areas. *Transportation Research Record: Journal of the Transportation Research Board*, Washington, D.C., USA, v. 1768, n. 1, p. 135-140. DOI: 10.3141/1768-16.

Tebaldi, C. and M. West (1998) Bayesian Inference on Network Traffic Using Link Count Data. *Journal of the American Statistical Association*. v. 93, n. 442, p. 557–573. DOI: 10.1080/01621459.1998.10473707.

Vardi, Y. (1996) Network Tomography: Estimating Source-Destination Traffic Intensities from Link Data. *Journal of the American Statistical Association*, v. 91, n. 433, p. 365–377. DOI: 10.1080/01621459.1996.10476697.

Weijermars, W. A. M. (2007) *Analysis of urban traffic patterns using clustering*. University of Twente, Enschede / Delft. Available in: <https://research.utwente.nl/en/publications/analysis-of-urban-traffic-patterns-using-clustering> (consulted on 08/18/2021).