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

Calibration of the empirical fundamental relationship using very large databases

Calibração da relação fundamental do tráfego a partir de bases de dados muito grandes

Juliana Mitsuyama Cardoso[1], **Lucas Assirati**[2], **José Reynaldo Setti**[3]

[1]University of São Paulo, São Paulo – Brazil, julianam.cardoso@usp.br
[2]São Carlos School of Engineering, São Paulo – Brazil, assirati@usp.br
[3]São Carlos School of Engineering, São Paulo – Brazil, jrasetti@usp.br

Recebido:
29 de fevereiro de 2020
Aceito para publicação:
2 de dezembro de 2020
Publicado:
30 de abril de 2021
Editor de área:
Sara Ferreira

Keywords:
Traffic stream models.
very large databases.
Model fitting.
Genetic algorithm.

Palavras-chaves:
Modelos de correntes de tráfego.
Bases de dados muito grandes.
Ajuste de modelos.
Algoritmo genético.

DOI:10.14295/transportes.v29i1.2317

ABSTRACT
This paper describes a procedure for fitting traffic stream models using very large traffic databases. The proposed approach consists of four steps: (1) an initial treatment to eliminate noisy, inaccurate data and to homogenize the information over the density range; (2) a first fitting of the model, based on the sum of squared orthogonal errors; (3) a second filter, to eliminate outliers that survived the initial data treatment; and (4) a second fitting of the model. The proposed approach was tested by fitting the Van Aerde traffic stream model to 104 thousand observations collected by a permanent traffic monitoring station on a freeway in the metropolitan region of São Paulo, Brazil. The model fitting used a genetic algorithm to search for the best values of the model parameters. The results demonstrate the effectiveness of the proposed approach.

RESUMO
Neste artigo, descreve-se um procedimento para ajustar modelos de correntes de tráfego a partir de bases de dados muito grandes. O procedimento proposto consiste em quatro etapas: (1) um tratamento inicial nos dados para eliminar observações espúrias (ruído) e homogeneizar a informação ao longo de toda a gama de densidades observada; (2) um ajuste inicial do modelo, baseado na soma dos erros quadráticos ortogonais; (3) uma segunda filtragem de dados, visando eliminar os *outliers* que sobreviveram ao tratamento inicial para eliminação do ruído; e (4) um segundo ajuste final do modelo. O método proposto foi testado ajustando-se o modelo de correntes de tráfego de Van Aerde a um conjunto de 104 mil observações coletadas por uma estação permanente de monitoramento de tráfego instalada numa autoestrada na região metropolitana de São Paulo. A calibração do modelo usou um algoritmo genético para procurar os melhores valores dos parâmetros do modelo. Os resultados obtidos demonstram a eficiência do método proposto.

1. INTRODUCTION

Traffic stream analysis usually uses data collected by traffic sensors at permanent traffic monitoring stations (PTMS), which, working continuously over months and years, can accumulate a very large amount of observations. However, this data includes noise that can skew the models. Road construction, accidents, bad weather, sensor malfunction, and incidents that affect the traffic stream behavior are unusual conditions that do not represent regular road operation.

It can be expected that in a very large database (VLDB) covering several years, a considerable portion of the speed-flow observations includes noise.

This study proposes a process for fitting traffic stream models using very large data sets. The proposed method includes a first treatment to standardize the volume of information across the range of traffic stream densities and reduce noise. The model is then calibrated by a two-step optimization process to eliminate the noise that survived the first filter. The proposed procedure used a genetic algorithm (GA) to search for the best values for model parameters, but any other optimization technique could be used. The traffic stream model chosen to test the process was the one proposed by Van Aerde (Rakha, 2009), as it is a well-known model; any other model, however, could have been used.

2. LITERATURE REVIEW

2.1. Traffic stream models and their calibration

A traffic stream model describes the macroscopic relationships between flow (q), speed (u) and density (k) (May, 1990) in a system of equations consisting of:

$$0 = \frac{\partial q}{\partial x} + \frac{\partial k}{\partial t} \quad (1)$$

$$q = u \cdot k \quad (2)$$

$$u = f(k) \quad (3)$$

where Eq.1 is a continuity equation concerning the space (x) and time (t) domain; Eq. 2 is a state equation, and Eq.3 establishes a relation (function f) between speed and density (Lu *et al*., 2010). The fundamental hypothesis for defining a model is that, for each specific location, there is a relationship among u, k, and q, called the fundamental equation (Eq. 2), which contains the solutions for a steady-state model for this traffic stream (Kerner, 2004). Traffic stream models are the subject of constant studies given their informational capacity regarding the characteristics of the roads and drivers (Coifman, 2014).

The empirical fundamental relationship (the speed-density or speed-flow models) is fitted using data collected by traffic monitoring stations (Hall *et al.*, 1992; Coifman, 2014), which provide, for a given observation time interval, measurements of flow rate, speed, and occupancy (of which density can be estimated), to which one an error is associated. It is well known that the regression of a dependent variable y as a function of an independent variable x such that $y = f(x)$ does not produce the same relation as the regression of x as a function of y (Draper & Smith, 1980). To avoid this problem, in a regression, one must clearly define which independent variables and which dependent variables, since it assumes that only the dependent variable contains a measurement error. When fitting a traffic model, this means that the regression $u = f(k)$, which presupposes that the density k is the explanatory (independent) variable and that the speed u is the dependent variable, minimizes only the error associated with the estimated speed. Rakha & Arafeh (2010) demonstrate that this is not the case with traffic stream models since it is not easy to define which the independent variable is and that, depending on the situation, any of the three variables may be the determining factor for traffic behavior, and all three variables inherently carry a measurement error.

To overcome this problem, Rakha & Arafeh (2010) suggested that the calibration should be based on a neutral regression, which does not require the determination of the dependent variable, and the adjustment seeks to minimize the normalized orthogonal quadratic error of the fundamental diagram of the chosen model. This optimization model can be described as:

$$\text{minimize } E = \sum_i \left\{ \left(\frac{u_i - \hat{u}_i}{\tilde{u}} \right)^2 + \left(\frac{q_i - \hat{q}_i}{\tilde{q}} \right)^2 + \left(\frac{k_i - \hat{k}_i}{\tilde{k}} \right)^2 \right\} \quad (4)$$

subject to:

$$\hat{u}_i = f(\hat{k}_i) \quad \forall i, \quad (5)$$

$$\hat{q}_i = \hat{k}_i \times \hat{u}_i \quad \forall i, \quad (6)$$

$$\hat{q}_i, \hat{k}_i, \hat{u}_i \geq 0 \quad \forall i, \quad (7)$$

where E is the estimated orthogonal quadratic error; u_i, q_i and k_i observed speed, flow, and density values for the i-th observation; $\hat{u}_i$, $\hat{q}_i$ and $\hat{k}_i$ are the estimated values for speed, flow, and density for a i-th observation; and $\tilde{u}$, $\tilde{q}$ and $\tilde{k}$ are the maximum observed values for speed, flow, and density. The neutral regression method can be applied to any traffic stream model (Rakha & Arafeh, 2010) and can be solved using any optimization technique.

This formalism supports the development of empirical flow-density and speed-density relationships, based on empirical observations of the values of u and q (Kerner, 2004). Permanent traffic monitoring stations, where sensors count and classify vehicles and measure their speed, are used to collect traffic stream data. Because each road segment has its own peculiarities, these empirical observations of traffic variables lead to a unique fundamental diagram (Knoop & Daamen, 2017). The collection of empirical data, however, is associated with some problems that need attention before the data is ready for model fitting.

2.2. Empirical data for fitting traffic stream models

The use of raw traffic data to calibrate empirical fundamental relationship is linked to many problems (Knoop & Daamen, 2017): (*i*) the traffic stream may not be in equilibrium during the observation period; (*ii*) the traffic stream is heterogeneous; (*iii*) the detectors have limitations (such as not being able to detect stationary vehicles) and are subject both to failure and measurement errors; (*iv*) the number of vehicles measured during an interval is always integer; and (*v*) the average speed recorded by the sensor is the time-mean speed. Regarding this last aspect, Knoop et al. (2009) compared the time-mean speed and the space-mean speed using individual vehicle data for a motorway segment, showing that "the space-mean speed gives a better fit for the fundamental diagram". The authors also point out that the use of time mean speeds affect mostly the congested flow region, underestimating the jam density and the shockwave speed.

Fitting traffic models to empirical data bring up another problem, which is the noise inherent to the traffic sensor data. There may be incidents (e.g., roadwork, traffic bans, accidents, bad weather, sensor malfunctions, etc.) during the period over which the data is collected that are not representative of the normal operation. This noise (i.e., inaccurate information) can negatively affect the quality of the fitted model. In the absence of reliable information about such incidents, it is necessary to create a way to filter the raw data to minimize the noise. Models used to detect freeway incidents, which incorporate techniques such as fuzzy logic, wavelets, and neural networks to reduce noise and increase their reliability (Karim & Adeli, 2002) demonstrate the importance of raw data filtering.

2.3. The use of large databases for the calibration of traffic models

With the increasing availability of data collected by PTMS, the use of large databases for the calibration of traffic models became more common. The literature shows traffic models calibrated using data collected during peak hours on specific days (Ma & Abdulhai, 2002; Hourdakis

et al., 2003; Yang & Ozbay, 2011; Balakrishna et al., 2007; Henclewood et al., 2013) and covering several days or weeks (Jha et al., 2004; Toledo et al., 2004; Qin et al., 2004; Lee & Ozbay, 2009; Zhang et al., 2008; Knoop et al., 2009). When traffic data covers several months and even years, the database comprises tens of thousands of observations and is considered a very large database (VLDB). Dealing with VLDB, Dervisoglu et al. (2009) reported using 27,000 observations obtained good results in fundamental diagram calibration observing the breakdown point. Qu et al. (2015) used 48,000 observations and a weighted least square method to calibrate both light-traffic/free-flow conditions and congested/jam conditions separately; and Zhong et al. (2016) have used 10,000 observations for a cell transmission model fitting which implied the division of the data in analysis regions (cells) and training sets.

Using a VLDB to fit a traffic model requires automating the calibration procedure, due to the sheer amount of data processed. The following sections in this paper describe the traffic data and the proposed approach.

3. THE PROPOSED APPROACH

To calibrate the fundamental diagram using a very large database (VLDB), the proposed approach consists of the following steps: (1) data aggregation; (2) noise reduction; (3) first-stage model calibration; and (4) second-stage model calibration. All steps are based on average speed (u) and density (k) data, because there is a monotonous relationship between these two variables (Wu, 2002): u never increases with an increase in k – i.e., low densities imply in high average speeds and vice-versa. The next sections explain the proposed approach.

3.1. Data aggregation

A very large traffic database includes observations on congested and non-congested flow regimes, with the former being much rarer than the latter, even for locations where traffic jams are very common. Thus, a scatterplot of (u, k) data will show many more data points representing uncongested flows than congested flows. This unbalance will bias the model, resulting in a poorly fit model, whichever calibration procedure is used. The best way to eliminate this bias is to aggregate the data into density classes, in such way that all classes have the same weight in the calculations to fit the model (Wu, 2002; Rakha & Arafeh, 2010).

Instead of providing individual vehicle data (speed, class and timestamp), from which space-mean speed and density could be easily derived, typical PTMS data consist of vehicle counts and average speeds for predetermined time intervals (5 or 15 minutes). From such data, density for a given observation interval may be estimated using $k = q/v$, assuming that the time-mean speed v is an adequate estimate of the space-mean speed. With the (u, k) data, the next step is choosing the range of the density classes. The selected range must provide a sufficiently large number of data points for fitting the model and the choice will depend on the available data and other intervening factors.

Within each class, several individual observations of speed and density will be available. However, a single pair of values (u, k) should be obtained for each density class. Coifman (2014) chose the average speed median and the density median. In this study, the selection of the value for speed and density employed the cumulative distribution, using a predefined percentile, similar to the approach used by Punzo & Montanino (2016) The next step in the method is noise attenuation.

3.2. Noise reduction

In a traffic stream, high speeds are associated with low density values. Thus, observations made at very low densities should theoretically result in average speeds close to the free flow speed. In a VLDB, it is possible to find very low densities linked to very low speeds, representing anomalous operating conditions (road maintenance, lane closures, bad weather, etc.). Ideally, information on periods of anomalous operating conditions would be available to purge this noise from the database.

In many cases, however, such information is not readily available or is not dependable. For these situations, a noise attenuation step improves the quality of the fitted model. A preliminary dataset analysis indicated that the initial data aggregation procedure eliminates some, but not all, of this noise. Evidence of anomalous data is the presence of data points with low densities and low speeds after the first step. In this study, investigations on possible ways to reduce the noise in low-density data points indicated that eliminating very low-density data points from the dataset would affect the fitted model minimally. The proposed approach employs a filter that requires the selection of k_{low}, a lower threshold value for density – that is, any data point with $k_i \leq k_{low}$ is purged from the dataset. The value of this density threshold is highly dependent on the dataset. A good indication of this threshold is the density beyond which speed never increases with an increase in density.

The model calibration in two stages also helps to reduce noisy data associated with high-density observations, as explained in the next sections.

3.3. First-stage model calibration

Once most of the noise is purged from the data, the first stage of the model calibration ensues. The user must choose the traffic stream model and the optimization method that will be used in this step. For instance, Wang et al. (2011) and Ni (2016) review several traffic models that could be used for this purpose.

Any optimization method may be used to calibrate the model with a carefully chosen objective function. At the end of this stage, a fitted model $u = f_1(k)$ is available to support the second-stage calibration.

Some researchers (Rakha & Arafeh, 2010; Wang et al., 2011) used a single calibration stage and obtained very good models; none of them, however, employed a VLDB to fit the model. The absence of reliable information on the occurrence of anomalous operating conditions to purge the VLDB, however, might imply in some loss of accuracy due to noise escaping the filters of the first two steps in the method. Hence, a second-stage calibration is included to further refine the fitted model.

3.4. Second-stage model calibration

The second-stage calibration uses the first-stage model $u = f_1(k)$ to remove outlying data points missed by the initial filter. To do so, for each density class i, an estimate of the average speed $\hat{u}_i = f_1(k_i)$ is calculated and compared to the average speed u_i for that class. If $|\hat{u}_i - u_i| \geq T$, where T is a predetermined acceptable tolerance, that data point is considered as an outlier and discarded. The value of T should be selected carefully, after inspection of the data and the fitted model.

The dataset obtained after this last filter is then used to fit the model through the same optimization method used in the first-stage calibration. The proposed approach can be easily automated using any programming language. The next sections explain how this was done in this study.

4. CASE STUDY

To demonstrate the effectiveness of the proposed approach, it was applied to the calibration of the fundamental diagram of a segment of freeway in the metropolitan region of São Paulo, Brazil, using a database comprising more than 104,000 observations.

4.1. Description of traffic data

The permanent traffic monitoring station (PTMS) selected to test the proposed approach is installed on a freeway section without significant longitudinal grades, where access is controlled and is outside the area of influence of on- or off-ramps. In addition, because capacity is routinely reached at this location, the data contains observations in the uncongested and the congested flow regions. This PTMS is located on a major freeway in the metropolitan region of São Paulo (SP070, km 39.5 East), on a three-lane segment that can be considered a basic freeway section. The closest off-ramp is located at approximately 4 km downstream from the PTMS; the nearest on-ramp is located around 3 km upstream from the sensor. Traffic data were provided by ARTESP (São Paulo State Transportation Agency) and cover the period from September 1, 2011 to December 31, 2017.

The PTMS chosen collects traffic data using inductive loop sensors. Data records consist of date, time, heavy vehicle count, passenger car count, and average speed for 15-minutes intervals. The average speed is the time-mean speed for the 15-minute interval and not the space-mean speed, as would be desirable. For this study, only the observations recorded between 5 AM and 10 PM were used, as it was considered that the traffic at late hours and dawn is not representative, due to the low volume of passenger-cars and the large percentage of heavy vehicles. In this section of the freeway, the speed limit for passenger cars (120 km/h) is higher than the speed limit of heavy vehicles (90 km/h). The average speed recorded by the PTMS is the average speed of all vehicles (cars and heavy vehicles) traveling over the segment.

As described in another paper (Cardoso et al. 2019), the data were treated to eliminate observations made in rainy weather, based on information from weather radar from IPMet/UNESP. In addition to this treatment, in processing data for VLDB composition, observations that showed apparent errors of sensor malfunctioning were excluded (such as the presence of repeated values several times, hugely discrepant values regarding the time series, absence of information and so on). After this step, the database contained 103,606 observations. However, these data contain noise, because there was no information on roadwork, accidents, sensor malfunction, and other incidents that may interfere with the regular operation of the traffic.

4.2. Data aggregation

To be successful, the calibration of a traffic model requires information on uncongested and congested conditions. Even in a freeway that regularly experiences traffic jams (such as the one chosen for this study), it is far more common to find uncongested 15-minute periods than it

is to find congested 15-minute periods in one year. Therefore, there will be an imbalance in information on congested and uncongested conditions, biasing the fitted model.

Figure 1 illustrates the problem of using raw data for model fitting. In the plot, the darker the color of a data point, the greater the frequency of the values represented by that data point, as the gray scale on the graph legends shows. The number of observations made under congested conditions (high density, low speed and low flow rates) is much smaller than the number of observations made under uncongested conditions. Furthermore, there is a great concentration of observations between 5 and 10 pce/km/lane, with speeds around 110 km/h and flow rates between 400 and 1200 pce/h/lane. The blue lines represent a model fitted to the 103,606 data points. The resulting 900 pce/h/lane capacity is much lower than the observed maximum flow rates, indicating the bias caused by the large number of observations in the darker gray area in Figure 1, as they will have a higher weight in the estimation of the fitted model error. The best way to reduce this undesirable effect consists of aggregating the raw data so that all density ranges have the same weight in model calibration (Rakha & Arafeh, 2010).

For the data aggregation, flow rates were converted from veh/h/lane into pce/h/lane using the PCE value adopted by ARTESP ($E_T = 2.5$) and densities were estimated using Equation 2, where q is the flow rate, in pce/h/lane, and u is the average speed of all vehicles (in km/h) and the density k is given in pce/km/lane.

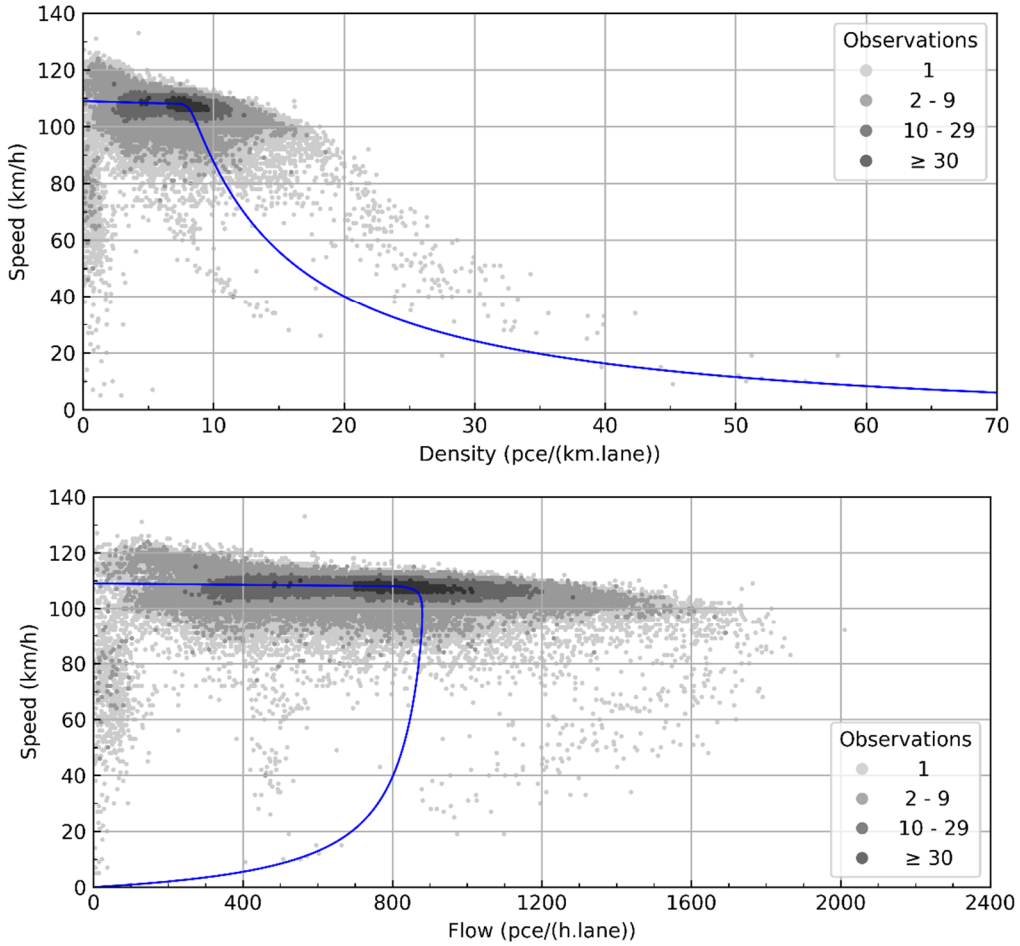

Figure 1. The need for raw data aggregation into classes is demonstrated by fitting a traffic model to the raw data. The large number of observations with density between 5 and 10 pce/km/lane biases the fitted model (blue lines), resulting in a much smaller capacity than the observed maximum flow rates.

The class range selected for this study is 0.25 pce/km/lane, because it provided a sufficiently large number of data points for fitting the model. For each class, the mean, median and 85th-percentile of speed and density were calculated, as shown in the histogram in Figure 2. The values for speed and density for each class were the 85th-percentiles of the cumulative speed and density distributions for the class (as shown in Figure 2). The 85th percentile was chosen on the assumption that it better represents the average speed of cars, given that the raw data is the average speed of all vehicles (cars and heavy vehicles) and that, in this segment of the free-way, the posted speed limit for cars is 30 km/h higher than the speed limit for heavy vehicles.

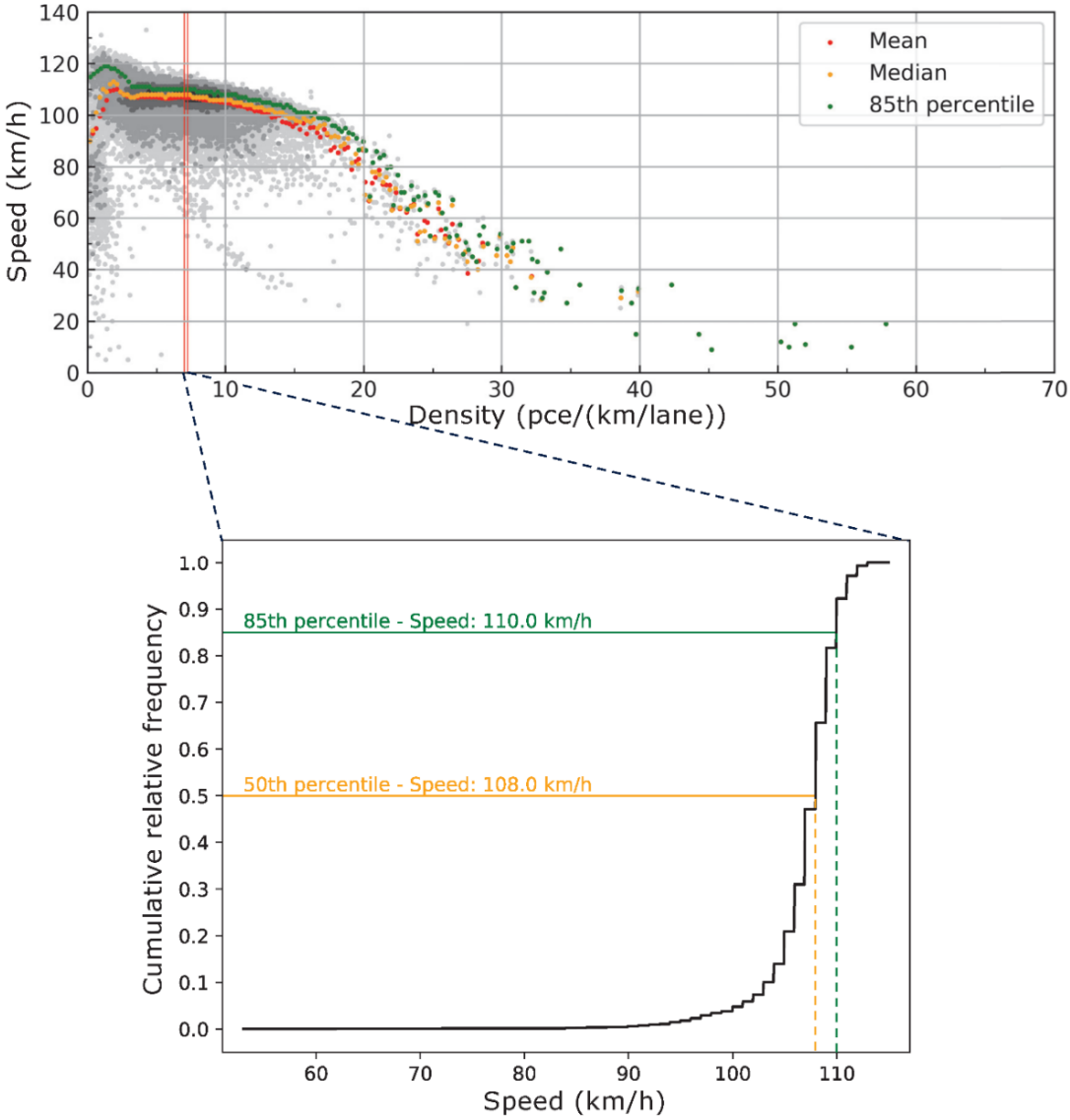

Figure 2. Aggregation of raw field data to homogenize the information over the observed range of density (SP070, km 39.5 East). The cumulative histogram illustrates the observed average speed distribution for densities in the range of 7.00 to 7.25 cpe/km/lane.

The aggregation of data transformed the 103,606 observations into 146 pairs of speed and density data, ensuring that the information about the behavior of the traffic stream is homogeneously distributed over the range of observed densities. Figure 2 shows the raw data, the mean and the median for each density class. It is possible to notice that, despite covering more than

6 years, the field data contains little information about the region close to the jam density ($k > 40$ veic/km/lane). This is a limitation of the data set; if the data had been collected at shorter intervals (for instance, at 5-minute intervals instead of 15-minute intervals), perhaps higher densities would have appeared more frequently in the traffic data.

4.3. Noise reduction

Figure 1 also shows the effect of noise in the raw field data, which is more apparent on the speed-density graph. The region where density is low ($k < 5$ pce/km/lane) and speed is also low ($u < 80$ km/h) represent highly anomalous operating conditions, in which vehicles are traveling at speeds far below the posted speed limit (120 km/h) with inter-vehicular spacing greater than 200 m, when one would expect that such spacing should correspond to speeds close to the free speed.

Figure 2 shows that the data aggregation eliminates some of the noise, but some noise still remains in the low-density region. For a traffic stream behaving normally, the speed-density relationship is a monotonically decreasing function – that is, the speed decreases or stays constant with the increase of density. When the raw data contains anomalous operating conditions (such as, when $k < 5$ veic/km/lane, in Figure 2), this condition does not exist. To reduce the noise associated with the identified low-density anomalous observations, only data points with density greater than k_{low} should be used in the calibration. For this site, k_{low} was selected as 5 pce/km/lane and the application of this filter reduced to number of data points to 126 from 146.

4.43. Selection of the traffic model

The proposed approach can use any traffic model. For this case study, the model selected was the Van Aerde model (Van Aerde, 1995), because it is a versatile model that can model both uncongested and congested flows with the same equation. The Van Aerde model combines the Pipes and Greenberg models (Lu et al., 2010) and can represent free and congested flows through a single mathematical function, without the need to establish breakpoints that separate these two regimes (Van Aerde, 1995; Rakha, 2009). Due to its mathematical structure, this model can adequately represent the behavior of traffic flow on freeways, two-lane roads, or even urban arterial roads (Rakha & Crowther, 2003) and, because of this versatility, the German capacity manual HBS (*Handbuch für die Bemessung von Straßenverkehrsanlagen*) has adopted the Van Aerde model (FGSV, 2015).

The Van Aerde traffic stream model is based on four parameters: the free flow speed u_f, the speed at capacity u_c, the flow rate at capacity q_c, and the jam density k_j (Rakha, 2009). Mathematically, the model is expressed by:

$$q = u \cdot k \tag{8}$$

$$k = \frac{1}{c_1 + \frac{c_2}{u_f - u} + c_3 u} \tag{9}$$

where c_1, c_2, and c_3 are constants that can be calculated using the following equations (Demarchi, 2003):

$$c_1 = \frac{u_f}{k_j u_c^2}\left(2u_c - u_f\right) \tag{10}$$

$$c_2 = \frac{u_f}{k_j u_c^2}\left(u_f - u_c\right)^2 \tag{11}$$

$$c_3 = \frac{1}{q_j} - \frac{u_f}{k_j u_c^2}. \quad (12)$$

4.4. Selection of the optimization technique for model calibration

Van Aerde & Rakha (1995) used a hill-climbing search to fit the traffic model to the data; Rakha & Arafeh (2010) adopted a multistage search to find the set of parameters that best fit the data. In this study, a genetic algorithm (GA) was used to fit the Van Aerde traffic model to the data, because genetic algorithms are able to better explore the solution space from a multitude of points and, therefore, are less susceptible to entrapment by local minima.

Genetic algorithms are a stochastic search method that mimics the theory of evolution and natural selection, in the sense that individuals best adapted to the environment (the better solutions to the problem) are more likely to survive. In GAs, a fitness function measures the degree of adaptation to the environment of an individual – i.e., the quality of a given solution (Goldberg 1989, p. 9).

The fitness function adopted for the GA is given by Eq. 13. By normalizing the flow, density, and speed values, the optimization problem can be expressed by:

$$\text{minimize } E = \sum_i \left\{ \left(\frac{u_i - \hat{u}_i}{\tilde{u}} \right)^2 + \left(\frac{q_i - \hat{q}_i}{\tilde{q}} \right)^2 + \left(\frac{k_i - \hat{k}_i}{\tilde{k}} \right)^2 \right\} \quad (13)$$

$$\text{with} \quad \hat{k}_i = \frac{1}{c_1 + \frac{c_2}{u_f - u_i} + c_3 u_i} \quad \forall i, \quad (14)$$

$$\hat{u}_i = u_f \left(1 - \frac{\hat{k}_i}{k_j} \right) \quad \forall i, \text{ and} \quad (15)$$

$$\hat{q}_i = \hat{k}_i \times \hat{u}_i \quad \forall i, \quad (16)$$

subject to specific restrictions for this application, which are:

$$\hat{q}_i \geq 0 \quad \forall i, \quad (17)$$

$$\hat{k}_i \geq 0 \quad \forall i, \quad (18)$$

$$\hat{u}_i \geq 0 \quad \forall i, \quad (19)$$

$$u_c \leq 0.9\, u_f, \quad (20)$$

$$u_f \in [(0.9 \times u_{\text{lim}}), (1.1 \times u_{\text{lim}})] \quad (21)$$

$$u_c \in [50, 105] \text{ km/h}, \quad (22)$$

$$q_c \in [1000, 3000] \text{ pce/h/lane, and} \quad (23)$$

$$k_j \in [65, 125] \text{ pce/h/lane} \quad (24)$$

where u_{lim} is the posted speed limit, and all the other variables have already been defined.

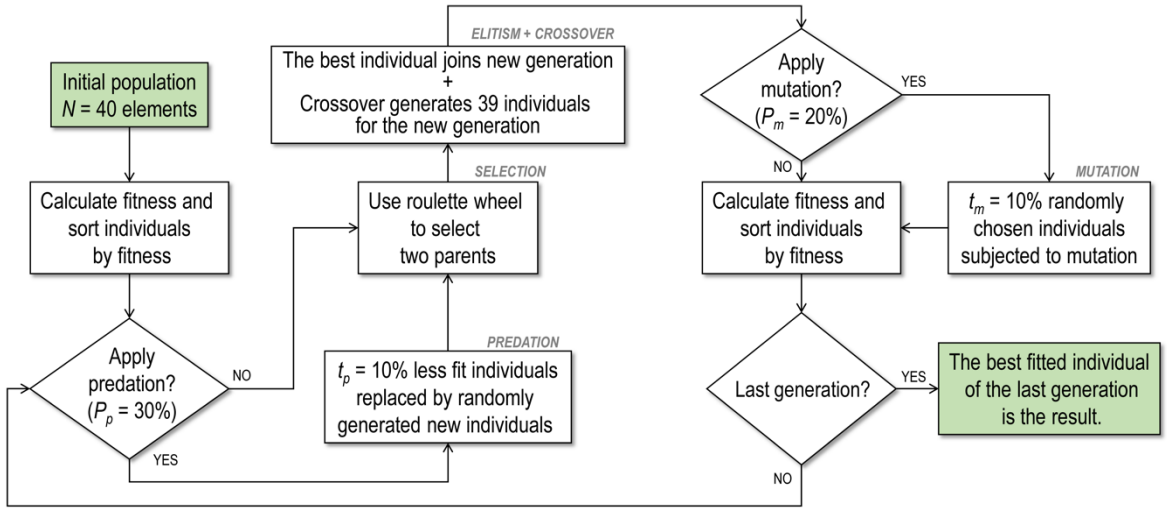

Figure 3. Overview of the genetic algorithm used to fit the Van Aerde model to the traffic data

The flowchart of Figure 3 shows the structure of the genetic algorithm, which was coded in Python v.3.7.0. The process starts with the creation of a population of N individuals. Due to the particularities of each problem, there is no general rule for defining the population size N (Diaz-Gomez & Hougen, 2007). The size N of the initial population is a critical factor as it has a major influence on the computational cost of the process. If N is too small, the algorithm might converge early, while a too large N might waste computational resources due to the high number of iterations required to improve the result (Arabas et al., 1994). The classic approach that defines N by the rule $N = 10 \times D$, with D being the number of genes (Storn, 1996), was chosen. As, for this study, $D = 4$, then $N = 40$.

Each individual or solution consists of a randomly created set of four model parameters (genes): free-flow speed (u_f), speed at capacity (u_c), flow rate at capacity (q_c), and jam density (k_j). In the generation of individuals, a series of checks ensures that the solution is feasible. Eqs.17 to 24 express these restrictions.

Once the initial population is established, the fittest individual must be found. To do this, the following steps are taken for all solutions in the population:

1) c_1, c_2 and c_3 are calculated from u_f, u_c, q_c, and k_j using Eqs. 10 to 12;
2) Eq.14 and c_1, c_2 and c_3 are used to calculate $\hat{k}_i$;
3) $\hat{u}_i$ is computed using Eq.15;
4) $\hat{q}_i$ is found using Eq.16; and
5) The individual's fitness, the orthogonal quadratic error E_i is calculated using Eq. 13.

Once the process has been completed for the population, the individuals are ranked according to their fitness E_i, with the lowest value of E ranking first in the list.

The evolution of this population (i.e., the search for the best solution) happens through the application of genetic operators (elitism, selection, crossover, mutation and predation) in combination with each individual's fitness.

Predation culls less adapted individuals (the worst solutions), replaced by randomly created new individuals (Srinivas & Patnaik, 1994). With each generation, there is a fixed chance for predation to occur. In this study, predation eliminates the worst individuals at a rate of $t_p = 10\%$, with a probability of $P_p = 30\%$ occurring with each iteration (Sivanandam & Deepa, 2007).

Elitism, selection and crossover are the operators used to create the new generation. Elitism places the best-fitted individual of one generation into the next, ensuring that a good solution will not be lost by chance during the selection of individuals to generate offspring. The $(N - 1)$ other individuals of the new generation are created from two parents chosen by the roulette wheel method (Chambers, 2000), based on the fitness of the parents: better-adapted individuals have a greater number of offspring in the next generation. To do so, the sum S of all errors E_i is calculated and the probability p_i of choosing an individual is inversely proportional to its contribution to S. After both parents are selected by this method, a random draw, with equal probability of occurrence, is conducted to determine if one, two, three, or four genes will come from one parent, with the complementary genes coming from the other parent. This combination of genes (crossover) creates a new individual for the next generation. The process is repeated until the individuals necessary to complete the future generation population have been created.

From time to time, the mutation operator is used to increase population variability. Mutation is applied at a constant rate $t_m = 10\%$ of the population with a random probability of occurrence $P_m = 20\%$ per generation. Mutation makes it possible to escape scenarios of little variability in the population, where descendants tend to be exact replicas of parents. In such cases, mutation is an opportunity to generate a new and better individual from a stagnating population (Coley, 1999).

The process of evolution continues through the generations until the maximum number of generations is reached, or the fitness value stabilizes. The individual best adapted to the environment of this generation represents the best solution.

For this study, the number of generations used is 1000, to explore as many solutions as possible. This number of iterations, perhaps exaggerated, was selected due to the low computational cost to reach this level, about 10 minutes, and also because more generations did not result in better solutions in the tests performed.

The parameters of a GA are usually chosen on a pragmatic way, seeking to maintain diversity at a gene-level, a population-level, or even a combination of both, to obtain good-quality solutions avoiding premature convergence, as well as considering the computational costs of each alternative (Diaz-Gomez & Hougen, 2007). The values chosen for this GA are very common in optimization problems and were tried in test runs. Anyhow, Reeves (1993) points out that a properly selected fitness function is far more important to ensure that an optimization state is reachable from any starting point within the search space than the chosen values for the GA parameters.

4.5. First-stage model calibration

The GA was then used to search for the best values for the Van Aerde model parameters, for the traffic data. Figure 4 shows the calibrated model (black line) over the raw data. The colored points on the graph show the data used for model calibration. The values found for the parameters of the Van Aerde model at the end of this initial stage were: (*i*) free-flow speed $u_f = 110$ km/h; (*ii*) speed at capacity $u_c = 89$ km/h; (*iii*) flow rate at capacity $q_c = 1761$ pce/h/lane; and (*iv*) jam density $k_j = 65$ pce/km/lane.

4.6. Second-stage model calibration

To refine the model obtained after the previous step, a second stage, consisting of a new filter followed by a further model fitting, was employed. This filter, mathematically expressed as $|\hat{u}_i(k_i) - u_i(k_i)| \geq T$ km/h, was applied to the 126 data points shown in Figure 4, eliminating those for which the absolute difference between the estimated speed and the observed speed for the correspondent density was higher than tolerance T. For this application, a tolerance $T = \pm 10$ km/h was used. This second filter eliminated 17 observations from the set of 126 observations initially used (marked in red in Figure 4).

The value of tolerance T was chosen in a pragmatic way, similarly to that used for the selection of the GA parameters. Several values were tested, trying to balance the need for the elimination of residual noise and the need for preserving the greatest amount of information for model calibration. The value adopted for T, 10 km/h, was assumed to be a good compromise between these two conflicting objectives.

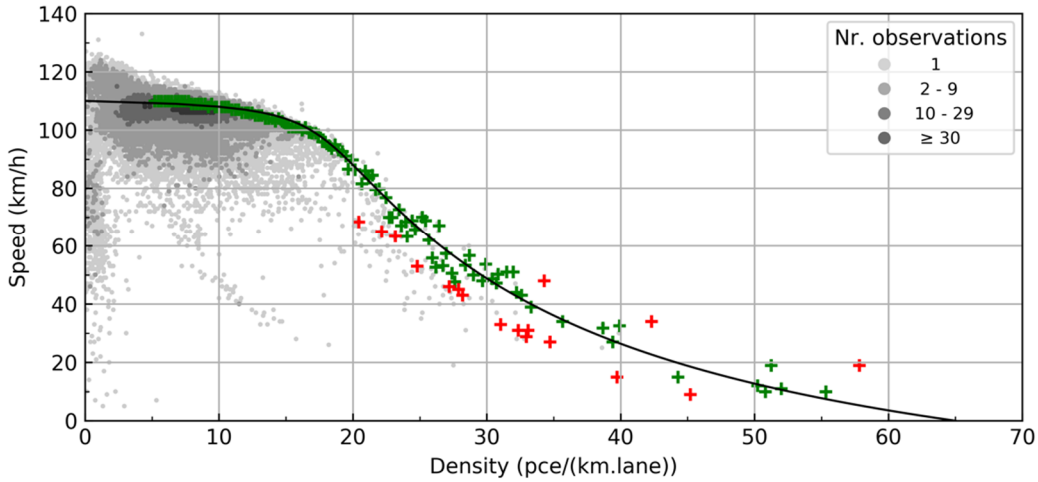

Figure 4. The Van Aerde model fitted at the end of the first stage using the genetic algorithm: the gray points are the raw data and the crosses represent the 126 data points used for the model calibration

Figure 5 shows the result of the second-stage model calibration. The values for the parameters of the Van Aerde model fitted to the 117 data points used in the second-stage calibration were: (*i*) free-flow speed $u_f = 110$ km/h; (*ii*) speed at capacity $u_c = 89$ km/h; (*iii*) flow rate at capacity $q_c = 1773$ pce/h/lane; and (*iv*) jam density $k_j = 65$ pce/km/lane. In Figure 5, the blue lines describe the model fitted to the raw PTMS data, whereas the red lines are the results of the proposed approach.

Little difference can be observed between the models fitted in the first stage and second stage calibration. This is due to two aspects, the first of which is the lack of information on traffic flows with speed lower than 25 km/h and density greater than 40 pce/km/lane, which influences the estimation of jam density. The GA search is more effective when more information (observed values) is available because this has a greater effect on the fitness value of a solution. The second aspect is that this result shows that steps 1 (data aggregation) and 2 (noise filter) in the proposed approach are quite efficient in removing the data noise, at least for the particular data set.

5. DISCUSSION OF RESULTS

To evaluate the proposed approach, the GA stopping criterion was a very large number of generations (1000) and the calibrated model goodness of fit of was evaluated by a Q metric defined as:

$$Q_i = \alpha \cdot \exp(-\beta \cdot E_i/\gamma), \tag{25}$$

where E_i is the orthogonal square error of the best solution of the i-th generation, defined by Equation 13; and α, β and γ are scale factors whose values were arbitrarily chosen to be $\alpha = 100$, $\beta = 5$ and $\gamma = 1$. Equation 25 shows that $0 < Q_i \leq 100$; that is, the larger the orthogonal quadratic error E_i, the lower the value of Q_i and if $E_i = 0 \Rightarrow Q_i \to 100$. Q is a measure of the goodness of fit of the model as the genetic algorithm evolves.

Figure 6 shows the evolution of Q as the number of solutions tested by the GA increases with each new generation. Note that each generation involves testing at least 39 solutions – more if predation and mutation operators are applied in that generation.

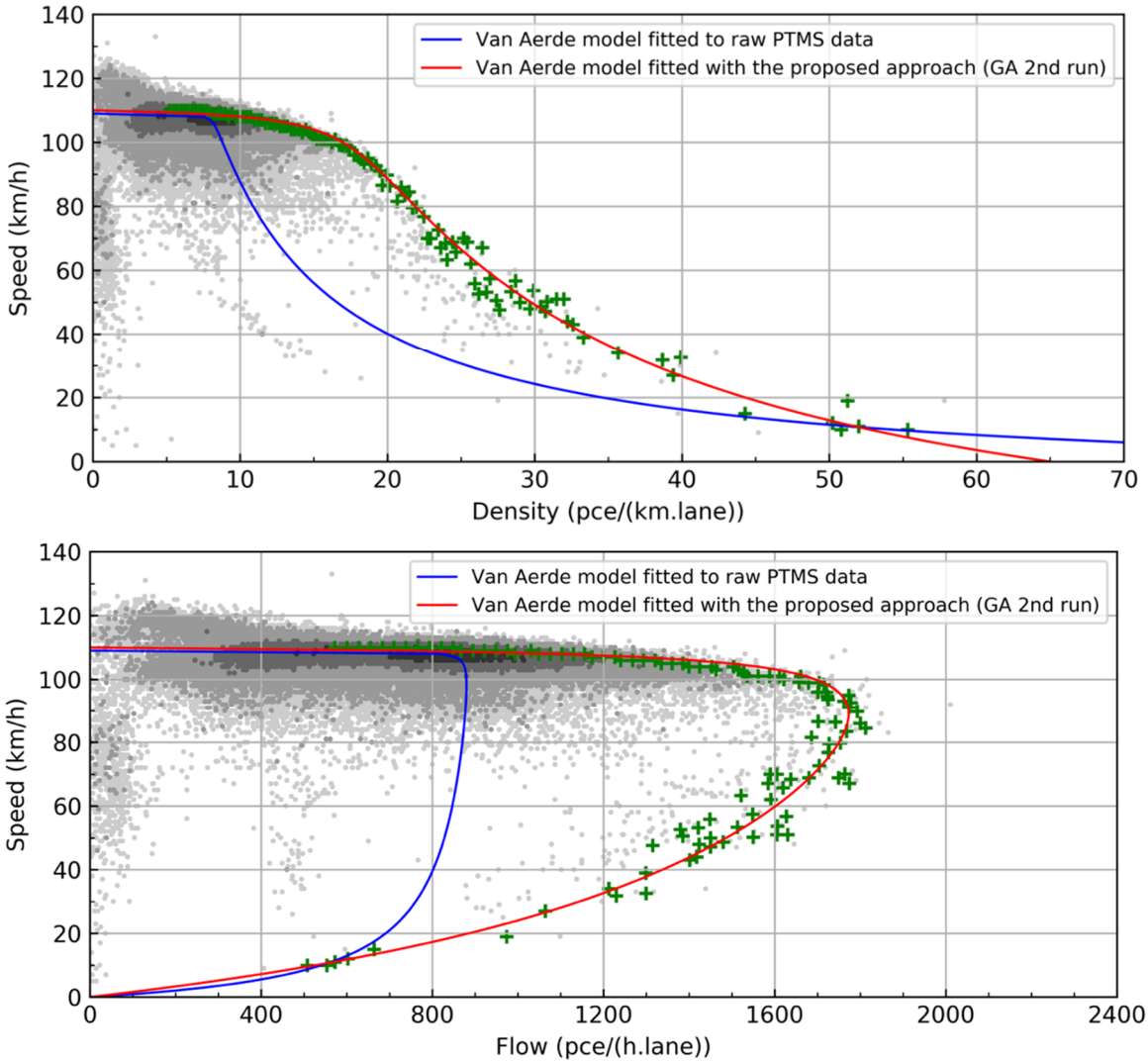

Figure 5. Comparison of the Van Aerde model fitted to the raw PTMS data (104,000 observations, in blue) and fitted with the proposed approach (in red): speed vs. density and speed vs. flow rate graphs

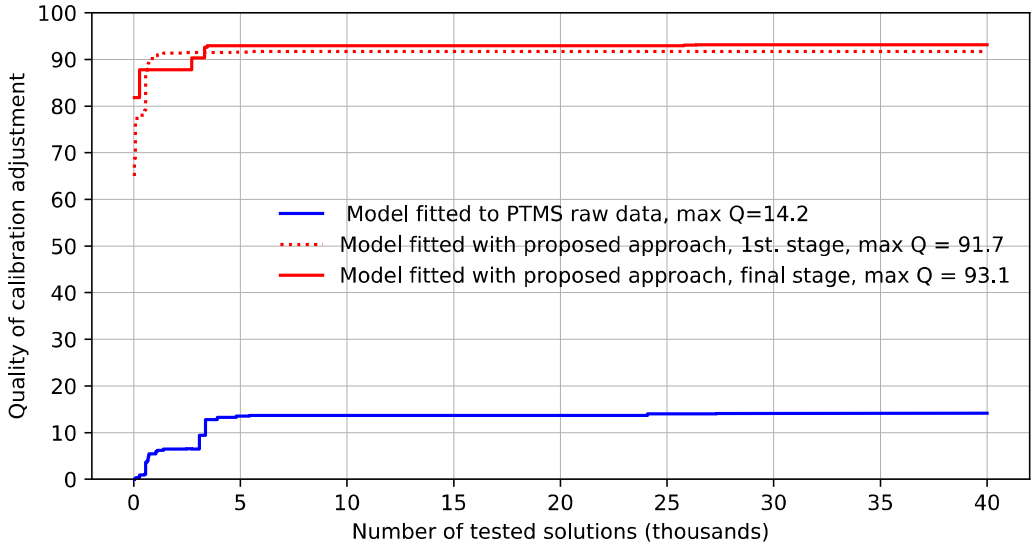

Figure 6. Evolution of the calibrated model's goodness-of-fit as a function of the number of solutions tested by the genetic algorithm.

Fitting the Van Aerde model to the raw PTMS data results in a maximum fitting quality of 14.2 after testing approximately 27360 solutions. The low quality of fit is explained by the great number of observations under low flow rates, which justifies the need to apply the proposed approach to homogenize the information over the full range of observed densities.

Using the proposed approach, which fits the model to filtered data, results in $Q = 91.7$ after the first-stage calibration, with just 5480 solutions tested. The second-stage calibration results in a small increase in the model's goodness-of-fit: $Q = 93.1$ after 26400 solutions tested. The small increment in the goodness of fit obtained in the final stage indicates that the proposed approach eliminates most of the noise, at least for this particular data set. In that way, one can consider the second stage representing a fine tune of the calibration process. If there are available computational resources and the desire to obtain a detailed solution, it is recommended to carry out the second stage. In the case of a lack of computational resources, the method can be interrupted after the execution of the first stage, without significant losses.

Because this GA starts with an initial population of 40 randomly created individuals, it is much more likely that fairly good solutions would appear within the initial generations, as was the case in this study, as the values of Q for the first generation were 65.01, in the first stage, and 81.85, in the second stage. However, this fact does not guarantee that same final results will be verified at the same number of tested solutions, but only suggests that the proposed approach can converge quickly to a good result thanks to the stochastic nature of the population created at the start of the algorithm.

6. FINAL CONSIDERATIONS

This paper has demonstrated that direct use of traffic data from a VLDB does not result into a properly calibrated traffic stream model, if the VLDB, like the one used in this study, includes a large number of repeated observations and noise. To solve these problems, the proposed approach reduces raw data into narrow density bins so that information is equally distributed over the range of observed densities in such way that all bins had equal weight on the fitted model and noise was minimized.

The implementation of a genetic algorithm allowed for an efficient way of searching for the search of the best solution because a GA is more likely to initiate the iterative search process from a superior solution and can better exploit the feasible set that satisfies the problem constraints. The proposed approach can be easily adapted to other traffic stream models and/or search mechanisms.

ACKNOWLEDGEMENTS

The authors would like to thank two anonymous referees whose comments and suggestions helped to clarify and improve the manuscript greatly. ARTESP and IPMet kindly provide d the data used in this research. This research was financed by grants from the Coordination for the Improvement of Higher Education Personnel (CAPES) - Finance Code 001, and CNPq (Bolsa de Produtividade em Pesquisa).

REFERENCES

Arabas, J., Z. Michalewicz, & J. Mulawka (1994). GAVaPS-a genetic algorithm with varying population size. In Proc. of the 1st IEEE Conference on Evolutionary Computation. IEEE World Congress on Computational Intelligence, p. 73–78. IEEE. DOI: 10.1109/icec.1994.350039

Balakrishna, R., C. Antoniou, M. Ben-Akiva, H. N. Koutsopoulos, & Y. Wen (2007) Calibration of Microscopic Traffic Simulation Models: Methods and Application. *Transportation Research Record: Journal of Transportation Research Board*, v. 1999, p. 198–207, DOI: 10.3141/1999-21

Cardoso, J. M., L. Assirati, & J. R. Setti (2019). Influência das condições meteorológicas na operação de rodovias de pista dupla paulistas. In Anais do XXXIII Congresso Nacional de Pesquisa em Transportes, Balneário Camboriú. ANPET.

Chambers, L. D. (2000). *The Practical Handbook of Genetic Algorithms: Applications.* Chapman and Hall/CRC. DOI: 10.1201/9781420035568

Coifman, B. (2014) Revisiting the empirical fundamental relationship, *Transportation Research Part B: Methodological*, v. 68, p. 173-184. DOI: 10.1016/j.trb.2014.06.005

Coley, D. A. (1999). *An introduction to genetic algorithms for scientists and engineers.* World Scientific Publishing Company. DOI: 10.1142/3904

Demarchi, S. H. (2003). Uma nova formulação para o modelo fluxo-velocidade-densidade de Van Aerde. In CNT/ANPET (Ed.), *Transporte em Transformação* – 7, p. 77–94. Brasilia, DF: LGE.

Dervisoglu, G., G. Gomes, J. Kwon, R. Horowitz, & P. Varaiya (2009). Automatic calibration of the fundamental diagram and empirical observations on capacity. Paper presented at the 88th Annual Meeting of the Transportation Research Board.

Diaz-Gomez, P. & D. Hougen (2007). Initial population for genetic algorithms: A metric approach. In *Proc. of the 2007 Intl. Conference on Genetic and Evolutionary Methods, GEM 2007*, Las Vegas, pp. 43–49.

Draper, N. & H. Smith (1980). *Applied Regression Analysis* (2nd. ed. ed.). New York: John Wiley & Sons. DOI: 10.2307/1267833

FGSV (2015). *Handbuch für die Bemessung von Straßenverkehrsanlagen: HBS 2015 Forschungsgesellschaft für Straßen und verkehrswesen*. Cologne: FGSV.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Boston: Addison-Wesley Longman.

Hall, F., V. F. Hurdle, & J. H. Banks (1992). Synthesis of recent work on the nature of speed-flow and flow-occupancy (or density) relationships on freeways. *Transportation Research Record: Journal of Transportation Research Board* v. 1365, p. 12–18.

Henclewood, D., W. Suh, M. O. Rodgers, & M. Hunter (2013) Statistical calibration for data-driven microscopic simulation model. Presented at 92nd Annual Meeting of the Transportation Research Board, Washington, D.C., 2013.

Hourdakis, J., P. G. Michalopoulos, & J. Kottommannil (2003) A Practical Procedure for Calibrating Microscopic Traffic Simulation Models. *Transportation Research Record: Journal of Transportation Research Board*, v. 1852, p. 130–139, DOI: 10.3141/1852-17.

Jha, M., G. Gopalan, A. Garms, B. P. Mahanti, T. Toledo & M. E. Ben-Akiva (2004) Development and calibration of a large-scale microscopic traffic simulation model. *Transportation Research Record: Journal of Transportation Research Board*, v. 1876, p. 121–131, DOI: 10.3141/1876-13

Karim, A. & H. Adeli (2002). Comparison of fuzzy-wavelet radial basis function neural network freeway incident detection model with California algorithm. *Journal of Transportation Engineering*, v. 128, n. 1, p. 21–30. DOI: 10.1061/(asce)0733-947x(2002)128:1(21)

Kerner, B. S. (2004). *The Physics of Traffic – Empirical Freeway Pattern Features, Engineering Applications, and Theory*. Berlin Heidelberg: Springer.

Knoop, V. L. & W. Daamen (2017). Automatic fitting procedure for the fundamental diagram. *Transportmetrica B: Transport Dynamics*, v. 5, n. 2, p. 129–144. DOI: 10.1080/21680566.2016. 1256239.

Knoop, V; S. P. Hoogendoorn & H. Van Zuylen (2009) Empirical differences between time mean speed and space mean speed. In: Traffic and Granular Flow '07. Springer: Berlin, Heidelberg, p. 351–356, DOI: 10.1007/978-3-540-77074-9_36

Lee, J.-B. & K. Ozbay (2009) New calibration methodology for microscopic traffic simulation using enhanced simultaneous perturbation stochastic approximation approach. *Transportation Research Record: Journal of the Transportation Research Board*, v. 2124, p. 233–240, DOI: 10.3141/2124-23

Lu, S., Y. Jun, H. Mahmassani, G. Wenjun, & K. Bum-Jin (2010). Data mining-based adaptive regression for developing equilibrium speed-density relationships. *Canadian Journal of Civil Engineering*, v. 37, n. 3, p. 389–400. DOI: 10.1139/L09-158

Ma, T., and B. Abdulhai (2002) Genetic Algorithm-Based Optimization Approach and Generic Tool for Calibrating Traffic Microscopic Simulation Parameters. *Transportation Research Record: Journal of Transportation Research Board*, v. 1800, p. 6–15.

May, A. D. (1990). *Traffic Flow Fundamentals*. Upper Saddle River, NJ, USA: Prentice Hall.

Ni, D. (2016) *Traffic Flow Theory: Characteristics, Experimental Methods, and Numerical Techniques*. Oxford: Butterworth-Heinemann, p. 51–71. DOI: 10.1016/B978-0-12-804134-5.00004-0

Punzo, V., & M. Montanino (2016) Speed or spacing? Cumulative variables, and convolution of model errors and time in traffic flow models validation and calibration. *Transportation Research Part B: Methodological*, v.91, p. 21–33, DOI: 10.1016/j.trb.2016.04.012

Qin, X., & H. S. Mahmassani (2004) Adaptive calibration of dynamic speed-density relations for online network traffic estimation and prediction applications. *Transportation Research Record: Journal of Transportation Research Board*, v.1876, p. 82–89, DOI: 10.3141/1876-09

Qu, X., S. Wang, & J. Zhang (2015). On the fundamental diagram for freeway traffic: a novel calibration approach for single-regime models. *Transportation Research Part B: Methodological*, v. 73, p. 91–102. DOI: 10.1016/j.trb.2015.01.001

Rakha, H. (2009). Validation of Van Aerde's simplified steadystate car-following and traffic stream model. *Transportation Letters*, v. 1, n. 3, p. 227–244. DOI:10.3328/TL.2009.01.03.227-244.

Rakha, H. & M. Arafeh (2010). Calibrating steady-state traffic stream and car-following models using loop detector data. *Transportation Science,* v. 44, n. 2, p. 151–168. DOI: 10.1287/trsc.1090.0297

Rakha, H. & B. Crowther (2003). Comparison and calibration of FRESIM and INTEGRATION steady-state car-following behavior. *Transportation Research Part A: Policy and Practice,* v. 37, n. 1, p. 1–27. DOI: 10.1016/s0965-8564(02)00003-4

Reeves, C. (1993) Using Genetic Algorithms with Small Populations. In: Proceedings of the Fifth International Conference on Genetic Algorithms (ICGA93), p. 92–99.

Sivanandam, S. N. & S. N. Deepa (2007). *Introduction to Genetic Algorithms*. Berlin: Springer. DOI: 10.1007/ 978-3-540-73190-0

Srinivas, M. & L. M. Patnaik (1994). Adaptive probabilities of crossover and mutation in genetic algorithms. *IEEE Transactions on Systems, Man, and Cybernetics,* v. 24, n. 4, p. 656–667. DOI: 10.1109/21.286385

Storn, R. (1996). On the usage of differential evolution for function optimization. In *Proceedings of North American Fuzzy Information Processing*, p. 519–523. IEEE. DOI: 10.1109/nafips.1996.534789

Toledo, T., M. E. Ben-Akiva, D. Darda, M. Jha, & H. N. Koutsopoulos (2004) Calibration of Microscopic Traffic Simulation Models with Aggregate Data. *Transportation Research Record: Journal of Transportation Research Board*, v.1876, p. 10–19, DOI: 10.3141/1876-02.

Van Aerde, M. (1995). Single regime speed-flow-density relationship for congested and uncongested highways. Paper presented at the 74th Annual Meeting of the Transportation Research Board, Washington, D.C. Paper No. 950802.

Van Aerde, M. & H. Rakha (1995). Multivariate calibration of single regime speed-flow-density relationships. In Pacific Rim TransTech Conference. 1995 Vehicle Navigation and Information Systems Conference Proceedings. 6th International VNIS, p. 334–341. IEEE. DOI: 10.1109/vnis.1995.518858

Wang, H., Jia Li, Qian-Yong Chen & Daiheng Ni (2011). Logistic modeling of the equilibrium speed–density relationship, *Transportation Research Part A: Policy and Practice*, v. 45, n. 6, p. 554–566. DOI: 10.1016 /j.tra.2011.03.010

Wu, N. (2002). A new approach for modeling of Fundamental Diagrams, *Transportation Research Part A: Policy and Practice*, v. 36, n. 10, p. 867–884. DOI: 10.1016/S0965-8564(01)00043-X.

Yang, H. & K. Ozbay (2011) Calibration of microsimulation models to account for safety and operation factors for traffic conflict risk analysis. Presented at 3rd International Conference on Road Safety and Simulation, September 14–16, 2011, Indianapolis, Ind.

Zhang, M., J. Ma, S. P. Singh & L. Chu (2008) Developing calibration tools for microscopic traffic simulation Final Report Part III: Global calibration—O-D estimation, traffic signal enhancements, and a case study. UCB-ITS- PRR-2008-8. California PATH Research Report, June 2008.

Zhong, R., C. Chen, A. H. Chow, T. Pan, F. Yuan, & Z. He (2016). Automatic calibration of fundamental diagram for first-order macroscopic freeway traffic models. *Journal of Advanced Transportation*, v. 50, n. 3, p. 363–385. DOI: 10.1002/atr.1334