

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900

Open access books available

185,000

International authors and editors

200M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



Procedure to Prepare and Model Speed Data Considering the Traffic Infrastructure, as Part of a Cyber-Physical System

José Gerardo Carrillo-González,

Jacobo Sandoval-Gutiérrez and Francisco Pérez-Martínez

Abstract

This chapter investigates the relationship between traffic control infrastructure (traffic lights and speed bumps) and the vehicles' travel speeds, for certain hours and days of the week. The authors propose the following procedures: (1) street segmentation, (2) clustering and categorization of speed data, (3) histograms' comparison analysis, (4) outlier detection, (5) modeling, and (6) delivering info to the users. Comparing speed histograms, segments with matching infrastructure presented similarities, regardless of the day of the week. Two techniques to model data were employed: polynomial regression and multinomial logistic regression. The algorithms to predict the travel speed category were also developed. The first technique yields on average 91.3% of data categorized correctly, and the second gets 90.09%. The traffic lights and speed bumps, located on the street segments under consideration, were identified as variables causing different travel speeds. The procedure allows to incorporate more traffic elements and can also be applied to other geographical locations.

Keywords: cyber-physical system, speed bumps, street segments, traffic lights, travel speed

1. Introduction

Traffic conditions have a profound effect on population's quality life. The TomTom traffic index states that in 2017, Mexico City had a travel delay of 66% when compared with normal times of uncongested traffic, placing it as the first in the world rank. The wasted time per day was 59 min, or 227 h per year, with delays in the morning and evening peaks of about 100%. Of the 23 million private cars in Mexico, 72% correspond to metropolitan areas [1]. As a result, those areas are a suitable choice to analyze traffic behavior. In 2010, with a population of 20,116,842 and 0.3 cars per habitant (about 6,035,052 cars), the Mexico City Valley is the most crowded of the country. The number of operating vehicles in a city reduces the average traveling speed and increases pollution [2, 3] and the number of car accidents [4, 5]. The zone under study in this work is located between Mexico City and

Toluca, a region that is part of the Mexico City megalopolis, which makes the area a suitable candidate for analyzing traffic conditions. In this research we developed a procedure to analyze speed tendencies (by comparing histograms) and prepare (set clusters and remove anomalies) and model speed data to be used in an application example: speed prediction. The procedure answers the following question: what is the pathway to generate new information when speed data is available?

Recently, there has been a great effort in studying and analyzing traffic data from different world locations. Travel speed is one way to measure traffic conditions, as is travel time. In [6], the travel time distribution for different kinds of roads is estimated for Beijing. The time intervals to analyze data were set to 15 min, and it was concluded that the best-fitting distribution depends on the congestion level and that the average travel time of all road segments (for all days) can be estimated with acceptable precision using the normal distribution (compared with the log-normal, gamma and Weibull). In [7], travel time prediction is pursued. The variables considered were flow, concentration, and higher order auto-regression, concluding that local linear regression is preferable than global modeling. Characterization of the daily temporal variation of congestion is presented in [8], where a fitted model and live data are combined in a ten-parameter exponential smoothing equation. With the purpose of analyzing historical traffic data, a query processing method with timeline information is proposed in [9], along with an analysis of the congestion dependency along roads. The work presented in [10] estimates the average link speed with vehicles equipped with GPS, and therefore the quantity of equipped vehicles required for estimating the speed was established.

Using traffic data to make predictions is a current challenge, as Google maps traffic and Waze are doing. The purpose in [11] is to use information from Bing Maps to analyze, visualize, and predict traffic jams in Chicago. In addition, a prediction model to correct flow intensities with logistic regression was proposed, where the independent variables were day, hour, street number, and number of pixels (red, yellow, and green). In this work, a tool was developed to extract the roads' traffic intensity from a GIS map service, where colors represent flow intensity: red as congested, green not congested, and yellow in between. In [12], the properties of a community-driven mapping service (Waze) are characterized. Additionally, the authors discuss the use of traffic data to identify traffic accidents and potholes. In [13], a four-phase traffic approach is proposed: (1) data collection and representation, (2) traffic prediction, (3) vehicle selection for re-routing, and (4) alternative route assignment. In our work, we focus our contribution in the first two phases.

The traffic infrastructure elements (such as traffic lights, speed bumps, potholes) involved in driving situations influence driver's behavior, which in turn affects speed and number of accidents. The intention in [14] is the development of statistical models to predict accidents. These models correlate highway characteristics with traffic accidents. The variables considered were classified in groups: section identifiers, cross section related, location, traffic related (e.g., the percentage of trucks on a highway section), alignment, horizontal curvature, and accidents. The regression methods used were Poisson and negative binomial. The statistically significant variables were number of lanes, horizontal curvature, speed limit, tangent length, section length, average annual daily traffic, and peak hour. In addition, accidents are predicted with equations that consider roadway elements such as average daily traffic, commercial and residential units, intersections, speed limits, lane width, and number of lanes.

The work presented in [15] classifies traffic control elements (infrastructure) into three groups according to their effect on accidents. In Group 1 are those elements that reduce the number of accidents, such as speed limit signs,

speed-reducing devices, signalized pedestrian crossings, urban play streets, pedestrian streets, traffic-calming areas, traffic signals at intersections, bus lines and bus stops, parking control, and access control. Group 2 has no statistical effect on accidents: road markings, one-way streets, reversible lanes, traffic control for pedestrians and cyclists, priority control, and yield signs at intersections. Group 3 increases accidents: right turn on red, pedestrian crossing without signs, blinking traffic light, and increasing speed limits. According to [16], the presence of traffic control elements with the purpose to reduce speed or simplify the road users' tasks (e.g., traffic signs) tends to reduce accidents. An obvious consequence of the presence of speed-reducing devices (humps, rumble strips, narrow road width, bollards) is the increase of travel time [17] and the decrease of the average travel speed. One of the conclusions in [15] is that the traffic control elements that reduce accidents also reduce mobility.

Traffic elements such as signals and traffic lights are important in human driving decisions. The work presented in [18] intends to determine the relevance of the static road elements in driving situations using Markov logic networks (MLNs). The information considered to determine the relevance of speed limits and supplementary signs were the position in relation to lanes, vehicle type, date, time, and weather. Then, with first-order logic rules, the relevance of each was inferred. To determine the relevance of traffic lights, the following variables were considered: navigation system, environment perception, spatial relations, and the traffic light state.

The speed changes in the presence of speed bumps were analyzed in [19]. The speed limit on the streets under study is 50 km/h. The speed results measured at the bump location are as follows: about 30% of the cases show an 85th percentile speed higher than the posted limit speed, 26% lie in the range 45–50 m km/h, and the rest is under 45 km/h. The 85th percentile speed (measured after 20–25 m of the bumps' location, at the crosswalk area) tends to increase in 50% of the tested sites, similar result for the 50th percentile case (45%). Nevertheless, for both cases the speed change was not significant, according to the statistical analysis. Another result was obtained comparing the speed at bumps and 100 m away: in most sites, the 85th percentile speed decreases in the range of 1–18% (with respect to the zone without bumps). The statistical analysis concludes for both percentiles that speed values do not change significantly.

The use of cyber-physical system in traffic is a current topic in the literature. In [20], a simulated vehicular cyber-physical system (VCPS) is designed for delivering warnings to the driver and to avoid accidents. With this end, the predicted vehicle motion/location, the driver behavior and the road geometry were considered. Then, the short-term motion of the objective vehicle and the surrounding vehicles are predicted. With the objective vehicle location and the traveled distance among vehicles, the collision risk is estimated, and the driver is notified. In [21], a perceptual Control Architecture of Cyber-Physical Systems (CPSs) is proposed, taking as example a traffic incident management system. The intelligent behavior of this is characterized by the physical-reflex space and cyber-virtual space. In the physical-reflex space, the sensing actuation of the objective scenario is constructed on four levels of traffic infrastructure. In the cyber-virtual space, the decisions (through Bayesian reasoning network) are defined according to three levels: principles, interrelated factors, and situation assessment. In [22] the potential participation of smartphones (equipped with GPS) is discussed to build a traffic information system (to inform the entire transportation network) that is part of the cyber-physical infrastructure system. In [23] a cloud-based cyber-physical system is presented, with the end to find fast routes for the users. The system is presented in four steps: (1) the GPS on taxis are used as mobile sensors to measure the traffic status in the

physical world; (2) the info generated by the taxis is sent to the cloud (cyber world) and mined, and then knowledge is acquired about the taxis' preferred directions and traffic patterns on the roads; (3) the knowledge in the cloud is sent to the users with the Internet; and (4) the recommendations for a specific user are improved using its driving behavior and preferred routes. In [24], a short-term traffic prediction model (combining fuzzy theory with Markov progress) is presented, which is part of a vehicular cyber-physical system; the prediction results are expressed in terms of traffic flow and speed. A proper discussion about the definition of a cyber-physical system, and its relationship with transportation, is in [25].

From a cyber-physical system point of view, in the procedure presented in this work, the cyber part corresponds to the elements in charge to acquire and mine data for generating knowledge and the process to communicate that Intel to the users. The user (a biological entity) and intelligent devices (e.g., the user smartphone, the vehicle computer) reacting in response of the knowledge correspond to the physical part.

The aim of the present work is to introduce a method for analyzing speed data measured on streets where the traffic infrastructure is assumed to be the cause of low speeds. Then, we develop models and algorithms that, working with our data, allow to make predictions. The procedure presented in this work is summarized in the following steps:

- Street segmentation is performed considering traffic control elements (speed bumps and traffic lights).
- Clustering speed data, validated with the silhouette metric.
- With the Chi-Square distance (χ^2), the travel speed histograms of weekdays are compared and also the histograms of segments.
- Mahalanobis distance is used to detect outliers.
- Two techniques (polynomial and logistic regression) were used to develop the models that describe speed data. An algorithm for each modeling technique was developed to predict travel speed.
- Communicate the generated knowledge to the users.

This chapter is organized as follows: Section 1 Introduction; Section 2 Method, which includes theoretical frame (data, clusters, histograms, outliers) and procedure (street segmentation, clustering, comparative analysis of histograms, outlier detection, mathematical models, connecting Intel with users); Section 3 Results (with discussion); and Section 4 Conclusions (with future work).

2. Method

2.1 Theoretical frame

2.1.1 Data

The zone under study is comprised of two streets located in Lerma de Villada, Mexico: Av. Miguel Hidalgo and Av. Reolin Barejon. Data was obtained using the Google Maps Directions API. The time for a vehicle to traverse each segment was

recorded every 15 min, after [6]. We found this time interval to be highly efficient for incorporating relevant data while ignoring redundant information. In this way, the average travel speed on each segment was measured. Three weeks (w_1 , w_2 , and w_3) of data were considered: w_1 from Dec 27, 2016 to Jan 03, 2017; w_2 from Jan 03, 2017 to Jan 10, 2017; and w_3 from Jan 20, 2017 to Jan 27, 2017. The time interval to acquire data was from 6 a.m. to 11:59 p.m. (an interval of 18 h per day) and only in weekdays, i.e., between Monday and Friday.

2.1.2 Clusters

The k -means technique [26] was selected (because it is easy to implement and is commonly used in distinct traffic problems [27–29]) to cluster the speed data of any of the 3 weeks; since these are close in time, it is expected a similar travel speed from 1 week to another, and then we select w_1 . In simple terms, the k -means technique consists in calculating the centroid of each cluster as the mean of the data in the corresponding cluster and is recalculated until convergence.

We apply the k -means technique selecting a number of clusters in the range 3–6; for each case we calculate the silhouette score [30], given in Eq. (1), where $a(i)$ is the average distance from i with the data in the same cluster, $b(i)$ is the minimum average distance from i with the data of each other's cluster, and i is the data index. The silhouette score is in the range -1 to $+1$; a value close to 1 indicates that the speed data is well matched in the selected clusters, while a value close to -1 indicates the opposite situation:

$$ss(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (1)$$

2.1.3 Histograms

Analyzing the speed frequency, by comparing speed histograms of certain locations (special selection) and certain time (temporal selection), we expected to find spatial and temporal relationship about the weekdays when the speed is similar (dissimilar) and the segments where the speed is similar (dissimilar).

The metric employed to compare a pair of histograms is the Chi-Square (χ^2) histogram distance [31], given in Eq. (2), where P and Q are the histograms to be compared and P_i and Q_i contain the speed frequency of the i bin (i is the bin index, the selected bin width is 1):

$$\chi^2(P, Q) = \frac{1}{2} \sum_i \frac{(P_i - Q_i)^2}{(P_i + Q_i)} \quad (2)$$

This metric has the advantage of reducing the importance of the result when bins with large count are compared, as in many natural histograms, the difference of bins with high values is less important [31]. If the metric gets a 0 result, then there is no difference between the compared histograms; as the result value becomes larger, the difference in terms of the speed frequency also becomes higher.

2.1.4 Outliers

We filtered the speed data using the Mahalanobis distance (MD) [32] to detect outliers, i.e., atypical speed not belonging to normal driving behavior, since we are not interested in including this data for modeling. The MD is presented in Eq. (3),

where x_i is a vector containing the time and speed, \bar{x} is a vector with the means, and C_x^{-1} is the covariance matrix:

$$MD_i = \sqrt{(x_i - \bar{x})^T C_x^{-1} (x_i - \bar{x})} \tag{3}$$

2.2 Procedure

2.2.1 Street segmentation

The avenues under study were divided into segments: each segment is denoted s_k , with k as the segment index. On each segment, we have number of speed bumps c_1 , number of traffic lights c_2 , and landmarks c_3 . A segment's length l is set to approximately 500 m, and then on each segment there are specific traffic elements: $s_k = \{c_1, c_2, c_3, l\}$, as shown in **Table 1**.

2.2.2 Clustering

The silhouette score, considering three clusters, is better evaluated, with $ss = 0.7360$. For four, five, and six clusters, we calculated a $ss = 0.7331$, $ss = 0.7194$, and $ss = 0.7105$, respectively. As we were interested in communicating in a simple way the speed category at which is possible to travel, three options (as slow, medium, and normal) seem adequate. A similar approach in Google Maps (traffic option), where the speed is represented considering four options, from fast to slow.

The resultant average speed (in km/h) range of each cluster (or category) is category 1 (5.4112–18.1455), category 2 (18.1455–23.4234), and category 3 (23.4234–36.0750). For w_2 and w_3 , values smaller than 5.4112 fall into category 1, and those larger than 36.0750 fall into category 3.

The percentage of a segment's speed data (from w_1) in a cluster is shown in **Table 2**. It is interesting to note that for all segments, there is a specific cluster that

s_k	c_1	c_2	c_3	l (m)	GPS start coordinate	GPS end coordinate
s_0	2	0	None	501	19.284512, –99.500927	19.285725, –99.505498
s_1	2	0	School, museum, gas station, government offices	500	19.285725, –99.505498	19.286330, –99.510221
s_2	0	1	Banks, center square, school, fast-food restaurants	500	19.286330, –99.510221	19.286711, –99.514964
s_3	3	2	Cultural center, hospital, school offices, kindergarten	501	19.286711, –99.514964	19.286477, –99.519630
s_4	3	0	Telecom company offices, shopping mail	499	19.286477, –99.519630	19.285784, –99.514944
s_5	2	1	Hospital, government offices, cultural forum	500	19.285784, –99.514944	19.284943, –99.510282
s_6	4	1	School, supermarket, hospital	500	19.284943, –99.510282	19.284500, –99.505561
s_7	2	0	None	481	19.284500, –99.505561	19.284403, –99.500993

Table 1.
Segments' characteristics.

Segment	Cluster 1 (%)	Cluster 2 (%)	Cluster 3 (%)
s_0	0.55	9.9	89.53
s_1	11.84	80.71	7.43
s_2	88.98	11.01	0
s_3	93.66	6.33	0
s_4	14.04	85.95	0
s_5	79.61	20.38	0
s_6	83.47	16.52	0
s_7	4.13	3.85	92.01

Table 2.
Percentage of speed data in a cluster.

contains a high percentage of data (at least 79.61%), which validates our clustering results.

2.2.3 Comparative analysis of histograms

First, we consider all segments as a single road, and then the histograms of the speed frequency (from 6 a.m. to 11:59 p.m.) happening on weekdays (in w_1) are compared in pairs, with the Chi-Square metric presented in Eq. (2). The results are shown in **Table 3**, starting with the lowest χ^2 value, i.e., the similar histograms among weekdays, with D_1 = Monday, D_2 = Tuesday, and so on.

Second, the speed data throughout weekdays, but individual segments, was used to conform the histograms of the speed frequency happening on each segment for 5 days (the weekdays of w_1). These histograms were compared in pairs with the χ^2 . **Table 4** shows the results starting with the lowest χ^2 . We found that if the compared segments share similar traffic elements, the speed frequency also is similar, and therefore a low χ^2 is obtained.

D_2-D_3	D_4-D_5	D_3-D_4	D_2-D_4	D_1-D_3	D_1-D_2	D_1-D_5	D_3-D_5	D_2-D_5	D_1-D_4
7.77	10.758	10.936	11.139	14.097	15.168	16.347	16.827	17.609	20.653

Table 3.
Chi-Square distance between histograms with weekdays' data.

s_2-s_5	s_3-s_6	s_5-s_6	s_1-s_4	s_2-s_6	s_0-s_7	s_2-s_3
20.37	34.17	39.01	45.57	46.05	73.75	88.591
s_3-s_5	s_4-s_5	s_4-s_6	s_1-s_5	s_2-s_4	s_3-s_4	s_1-s_6
98.59	167.65	185.37	198.5	212.47	220.04	226.01
s_1-s_2	s_0-s_1	s_1-s_3	s_1-s_7	s_0-s_4	s_4-s_7	s_6-s_7
238.39	269.15	271.36	303.87	321.89	337.16	337.43
s_2-s_7	s_5-s_7	s_0-s_5	s_0-s_6	s_0-s_2	s_3-s_7	s_0-s_3
339.58	340.24	342.78	344.59	346.80	350.76	356.21

Table 4.
Chi-Square distance between histograms with segments data.

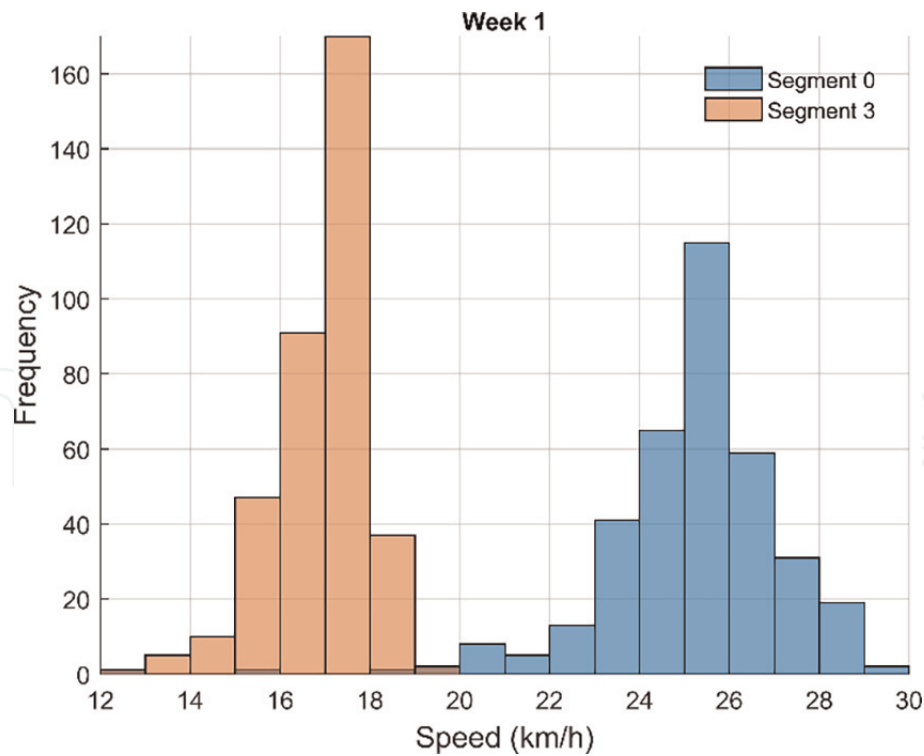


Figure 1.
Dissimilar histograms (s_0 and s_3).

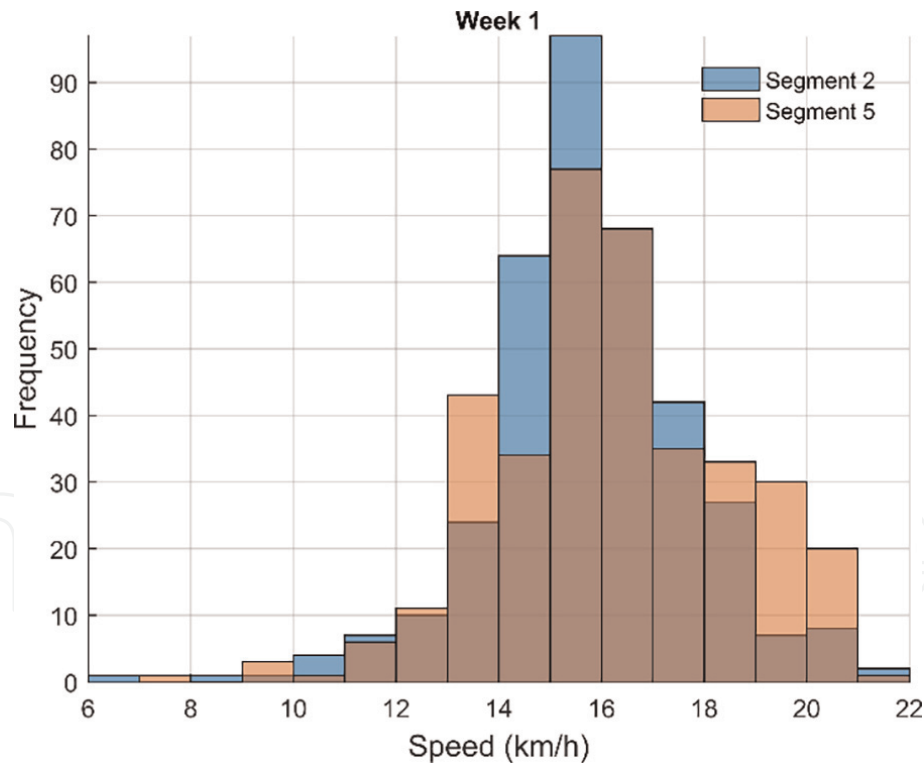


Figure 2.
Similar histograms (s_2 and s_5).

Figure 1 shows the most dissimilar histograms, s_0 and s_3 . Table 1 shows that s_0 has two speed bumps and no traffic lights, while s_3 has three speed bumps and two traffic lights; because the traffic lights on s_3 , we will expect a lower speed in this segment, and this conclusion can be corroborated by looking at Figure 1.

Figure 2 shows the most similar histograms, s_2 and s_5 . Segments s_2 and s_5 share the same number of traffic lights; however, there are two speed bumps in s_5 and 0 in s_2 , then a slight superior speed is expected in s_2 (see Figure 2).

Tables 3 and 4 show that comparing histograms with the speed frequency of individual days (and all segments) are evaluated with a lower χ^2 (the lower value is 7.77, the higher is 20.653) than the observed comparing histograms with the speed frequency of individual segments (and all days), where the lower value is 20.37 and the higher is 356.21. Then, it appears that the travel speed is weakly influenced by the day of the week, since the traffic control elements of the whole road, from day to day, are the same. However, it seems that the segment strongly influences the travel speed, since the traffic control elements, which characterize each segment, modify the speed at which is possible to travel.

To corroborate the abovementioned statement, we use the speed frequency of w_2 . **Figure 3** shows the histograms of the speed frequency of each day (and all segments), where it can be observed the histograms' similarity. **Figure 4** shows the histograms of the speed frequency of each segment (and all days), where it can be observed the histograms' dissimilarities.

2.2.4 Outlier detection

To put an example, the speed data of s_0 and w_1 is presented in **Figure 5**. We calculate the MD of this data (**Figure 5**), and then the probability density of the MD is presented in **Figure 6**, which has mean = 1.2331 and standard deviation SD = 0.6894. From **Figure 6**, a point with value MD > (2*SD + mean) = 2.6119 corresponds to a red point in **Figure 5** and is considered an atypical point. The inequality value, i.e., (2*SD + mean), was established through trial and error.

The speed data from w_1 and w_2 , for all segments, is filtered the same way as the example. The data used in the polynomial regression satisfy MD <= (2*SD + mean) and in the logistic regression MD <= (3*SD + mean).

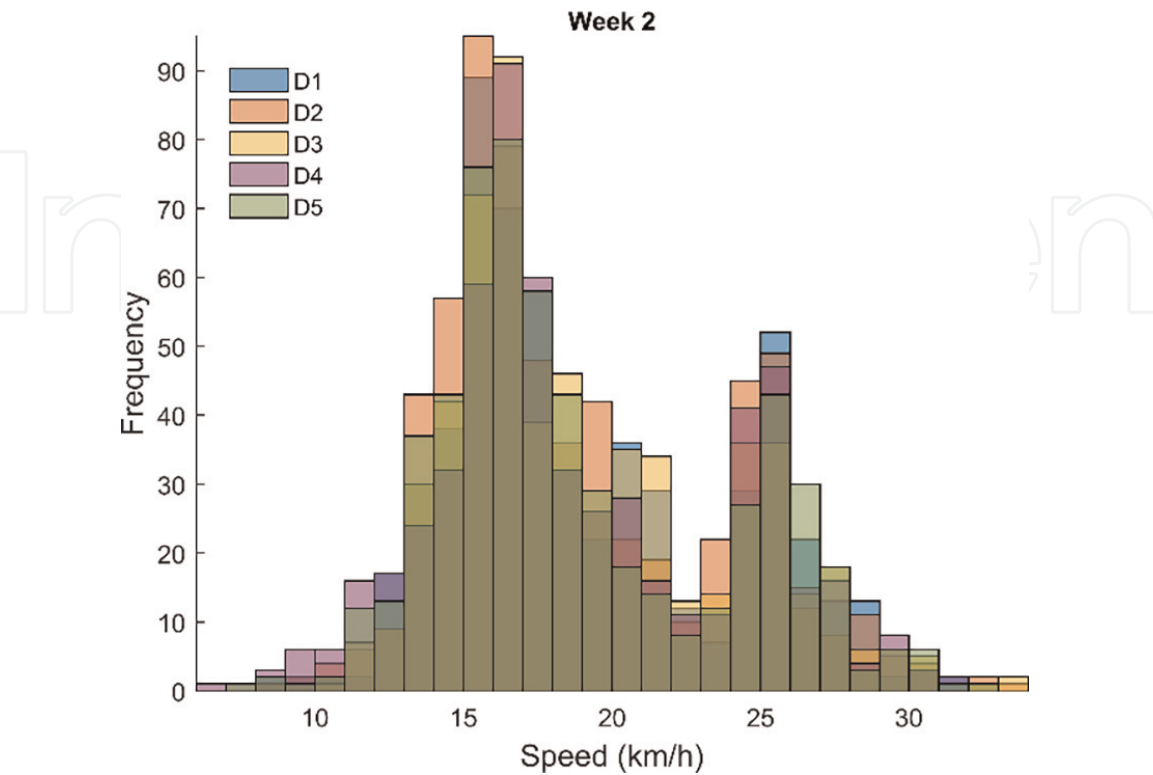


Figure 3.
Seed frequency of days.

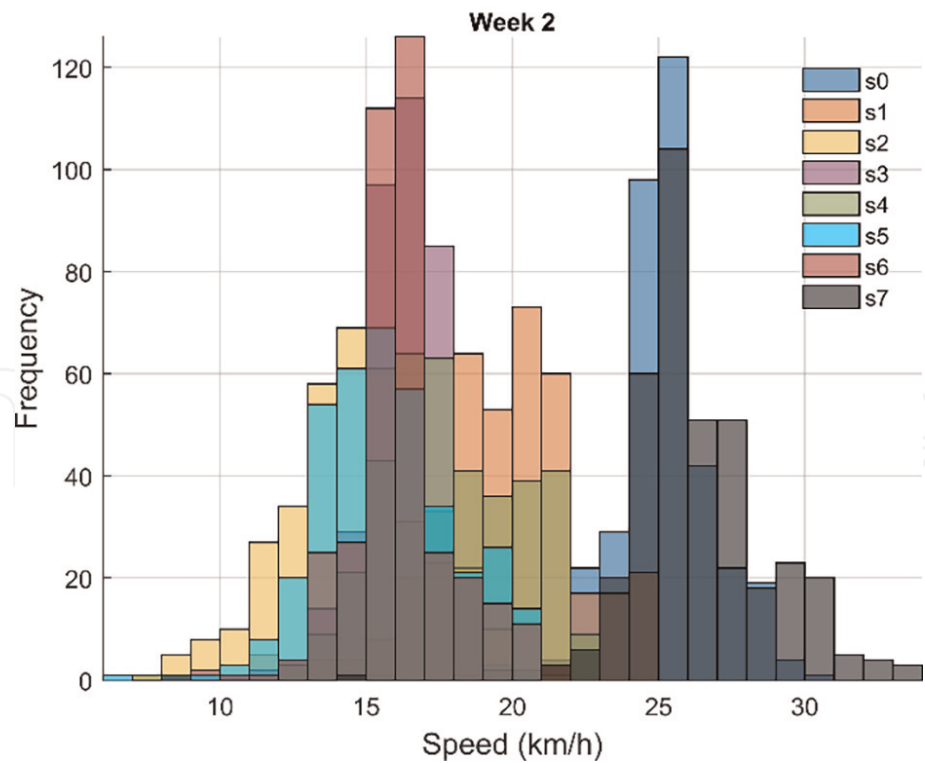


Figure 4.
Speed frequency of segments.

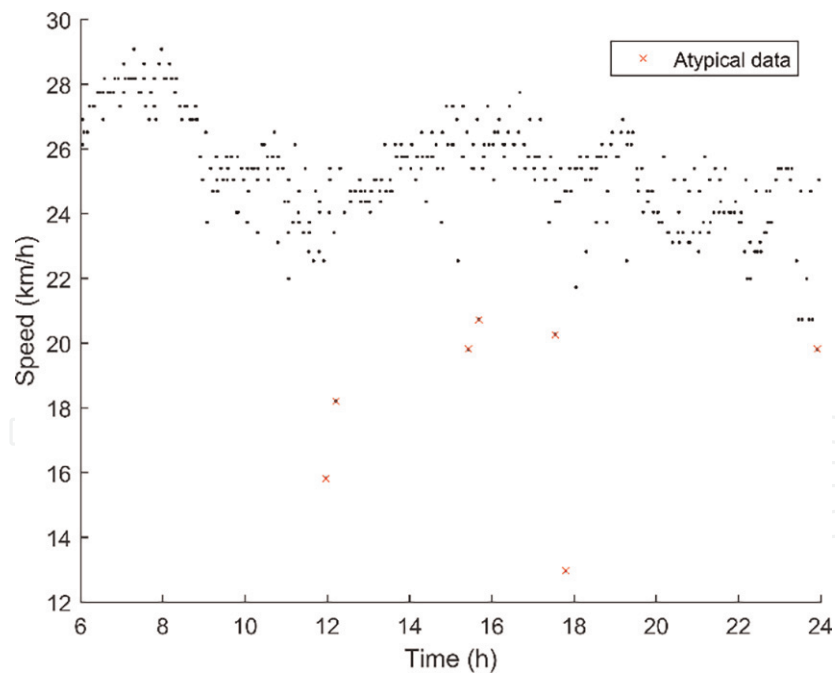


Figure 5.
Time vs. speed: Data of s_0 and w_1 .

2.2.5 Mathematical models

Polynomial: The data of each segment, with time as the independent variable and travel speed as the dependent variable, is modeled with a five-degree polynomial, enabling four-speed trend changes (the common requirement from the observations). The coefficients are calculated with the least-squares regression technique.

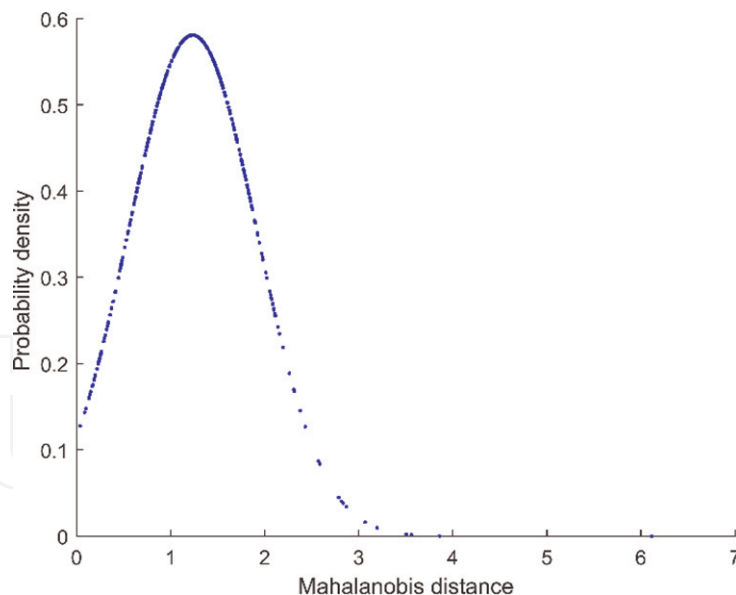


Figure 6.
Mahalanobis distance vs. probability density.

The following terminology is used to describe the model: the data size of all segments is $N = \sum_{k=0}^{k=7} N_k$, with N_k referring the data size of the k segment. The observed i speed is denoted by $y(i)$, while time is $t(i)$. The speed model of segment k and week q is denoted by $M_k^q(i)$, with $k = \{0, 1, 2, 3, 4, 5, 6, 7\}$ and $q = \{1, 2\}$. The model is presented in Eq. (4), where coefficients $\varphi_1 \dots \varphi_6$ were calculated with speed data of the corresponding week (q) and segment (k):

$$M_k^q(i) = \varphi_1 + \varphi_2 t(i) + \varphi_3 t(i)^2 + \varphi_4 t(i)^3 + \varphi_5 t(i)^4 + \varphi_6 t(i)^5 \quad (4)$$

Multinomial logistic: The number of speed bumps and traffic lights (see **Table 1**) are used to explain the speed. With multinomial logistic regression [33], we obtained the logistic model presented in Eq. (5), with $\psi = \{a, b\}$:

$$E_\psi^q(i) = \psi_1 + \psi_2 v_1(i) + \psi_3 v_2(i) + \psi_4 v_3(i) + \psi_5 v_4(i) + \psi_6 v_5(i) \quad (5)$$

The coefficients are denoted by $\psi_1 \dots \psi_6$, and $q = \{1, 2\}$ refers again to the data from w_1 and w_2 , respectively. The explanatory variables are v_1 = day weight, v_2 = number of speed bumps, v_3 = number of traffic lights, v_4 = segment weight, and v_5 = time. The weight of a specific day is calculated as the day average speed (of the speed measured from 6 a.m. to 11:59 p.m.) divided by the sum of the speed average of each weekday. A segment's weight is calculated as the segment's average speed (during weekdays) divided by the sum of the speed average of each segment.

In Eq. (5), E_a^q calculates the relative risk of being in cluster 1 vs. cluster 3 (the reference), and E_b^q calculates the same but for cluster 2 vs. cluster 3. The conversion to probability is given in Eqs. (6)–(8), where R_j^q is the probability belonging to the j category, with $j = \{1, 2, 3\}$:

$$R_1^q(i) = e^{E_a^q(i)} / (1 + e^{E_a^q(i)} + e^{E_b^q(i)}) \quad (6)$$

$$R_2^q(i) = e^{E_b^q(i)} / (1 + e^{E_a^q(i)} + e^{E_b^q(i)}) \quad (7)$$

$$R_3^q(i) = 1 - (R_1^q(i) + R_2^q(i)) \quad (8)$$

2.2.6 Connecting Intel with users

With the developed procedure, knowledge is acquired about the speed at which is expected to travel on the segments under the study. The architecture design (and the implementation) to connect the Intel with the users is out of the scope in this work (planned as future work); nevertheless we present in this section the basic idea.

The algorithms developed (in Appendix A and Appendix B) were programmed in a regular computer; according the procedure presented, the data acquired (from the zone under study) is modeled, and the models are used in the algorithms to generate knowledge. The link between this knowledge and the users could be established through a cell phone app (via the Internet). When a driver is in the proximity of a street segment, the cell phone (with GPS) detects the current location and acquires information for the driver, as the number of bumps and traffic lights, and also the expected travel speed calculated with the proposed algorithms; this info is presented to the driver in a proper way to not distract him, and then the driver can decide the more convenient route. A more challenging design is to communicate the cell phone with the vehicle (assuming that an intelligent system is part of it and can control some functions) and, for example, when the vehicle is approaching a speed bump, it automatically decelerates (if the driver is not reacting adequately).

The program running in a computer, in charged to acquire and mine data for generating knowledge and to establish communication with the responsive elements, conforms the “cyber” part of the system. The elements reacting with intelligence to the Intel delivered, as the driver, the cell phone, and the vehicle, conform the “physical” part of the system. Finally, the cyber and physical parts combined conform a cyber-physical system.

3. Results

3.1 Polynomial regression model and Algorithm 1

The error between the modeled data, with Eq. (4), and the observed data, was calculated with the mean absolute error (MAE) (see Eq. (9)) [34]. Here, $n = N_k$, $y(i)$ and $\hat{y}(i)$ are the observed and modeled data, respectively. **Table 5** shows the

Segment	M_k^1		M_k^2	
	MAE (km/h)	SD (km/h)	MAE (km/h)	SD (km/h)
s_0	0.8269	0.6802	0.7895	0.6321
s_1	1.0101	0.9630	1.1939	0.9916
s_2	0.9523	0.7622	1.1754	1.0670
s_3	0.6198	0.4628	0.6701	0.5917
s_4	0.8435	0.6882	0.9765	0.7416
s_5	0.8971	0.6826	0.9234	0.7408
s_6	0.8438	0.7297	0.7737	0.6680
s_7	0.9376	1.0762	0.7894	0.6653

Table 5.
MAE and SD.

MAE, and its standard deviation (SD), with the data of w_1 and w_2 , and the respective modeled equations:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y(i) - \hat{y}(i)| \tag{9}$$

An algorithm (Appendix A, Algorithm 1) is designed to predict the speed of w_3 using the modeled equations (M_k^1 and M_k^2) and historical data, i.e., the data available from w_3 before the current time. The error between the observed (from w_3) and predicted (with Algorithm 1) travel speeds is calculated with Eq. (9). The MAE, SD, and hits (percentage of data categorized correctly) for w_3 , using Algorithm 1, are shown in **Table 6**.

Figure 7 shows, as example, the observed speed data (in black circles) of w_3 and segment s_0 , the modeled data with w_1 (model M_0^1 , in blue dots) and w_2 (model M_0^2 , in green dots), and the estimated speed with Algorithm 1 (in red plus signs).

3.2 Multinomial logistic regression model and Algorithm 2

Algorithm 2 (see Appendix B) is used to predict the speed category of the observed data from w_3 . $H_1(i)$ and $H_2(i)$ are two data sets obtained from w_1 and w_2 , respectively. These sets save the associated category of the average speed in a time interval from $t(i)-0.5$ to $t(i) + 0.5$ ($0.5 \text{ h} = 30 \text{ min}$) and centered on $t(i)$, of the day and segment under evaluation. $H_3(i)$ is the category speed of w_3 (which is only available for previous data, i.e., prior to (i) , with $i...N$ being the data index. The probability most likely to occur is $P_q(i) = \max\{R_1^q(i), R_2^q(i), R_3^q(i)\} = R_x^q(i)$ and the category is stored in $S_q(i) = x$, where subindex $q = \{1,2\}$ refers to the week. A *threshold* value, selected through trial and error, is used to discard the result in $S_q(i)$ if $P_q(i) < \text{threshold}$. Algorithm 2 predicts the speed category for w_3 , which is stored in $S_3(i)$. Choosing *threshold* = 0.9 gives 90.09% of correct evaluations. This percentage is the summation of cases, where $S_3(i)$ was categorized correctly divided by the total data N.

Afterward, we attempted to predict the speed category of the observed speed in w_3 under the assumption that set $H_2(i)$ is composed only with the average speed of each segment, and not including H_1 . The optimum result was found if *threshold* = ~0.85, with 85.62% of correct predictions. If the *threshold* value is reduced, the positive prediction decreases (because the model fails to predict accurately with that *threshold* value). Similarly, if the *threshold* is increased, it becomes more

Segment	MAE (km/h)	SD (km/h)	Hits (%)
s_0	0.7757	0.7142	92.4
s_1	0.9353	1.0061	85.2
s_2	0.8641	0.7537	89.5
s_3	0.7749	0.7658	94.8
s_4	1.0053	1.0903	85.2
s_5	0.8051	0.7942	93.5
s_6	0.6968	0.6639	96.5
s_7	0.7994	0.9391	93.5

Table 6.
Algorithm 1 prediction results: MAE, SD, and hits.

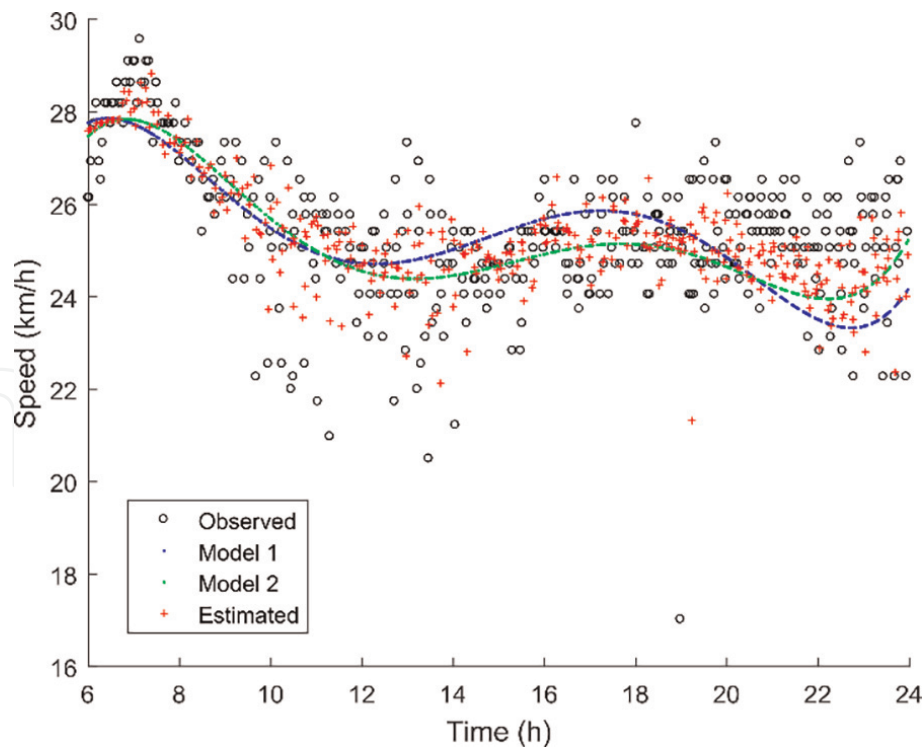


Figure 7.
Time vs. speed: data of s_0 and w_3 .

difficult to satisfy the condition $P_q(i) \geq threshold$, and then the positive prediction also drops because now the set $H_2(i)$ (with the limitation mentioned before) contributes more. **Table 7** shows the percentage of speed data categorized correctly with different threshold values.

3.3 Discussion

A series of steps are employed in a numerical example that, in combination, constitute a new method for speed prediction. The first step, *street segmentation*, divides an avenue in such a way that distributes different traffic elements on different segments. These elements are number of speed bumps, traffic lights, and landmarks, which in turn leads to different speed behavior on each segment. The second step, *clustering*, selects intervals which better fit the travel speed observed, resulting in three categories. Depending on the segment, most of the speed data (approximately at least 80%) is within a specific cluster (category). For example, we infer that the speed behaviors in s_2 and s_5 are similar, since most of the speeds for both fall inside cluster 1. Moreover, speed behaviors of s_0 and s_3 are dissimilar, since most of the speeds belong to different clusters (3 and 1, respectively). In the third step, *comparative analysis of histograms*, we corroborate that for each segment, the

Threshold	Prediction (%)
0.75	81.69
0.80	83.10
0.85	85.62
0.90	84.24
0.95	83.97

Table 7.
Algorithm 2: threshold values and w_3 prediction results.

speed behavior is related to the traffic elements involved. It was observed that the speed histograms of two segments get a low Chi-Square distance if the segments share approximately the same number of speed bumps, traffic lights, and landmarks, independent of the day of the week. A high Chi-Square distance implies the opposite situation, i.e., segments with different number of traffic elements. The fourth step, *outlier detection*, removes atypical speed behavior, e.g., a vehicle circulating slower or faster than the usual. In step five, *mathematical models*, the models explain the speed. From steps 2 and 3, it is already known that on each segment, speed behaves according to the traffic elements involved, and hence the speed data of each segment is modeled independently with a polynomial model, with time as the independent variable. The multinomial logistic model uses as independent variables the number of speed bumps, traffic lights, the time, and two weights. The weights are calculated based on the average of the measured travel speeds considering segments and days. Finally, in step 6, *connecting Intel with users*, the drivers are properly informed about the travel speed expected on the surrounding segments, helping them to continuously adjust their route.

4. Conclusions

The procedure presented in this chapter proposes street segmentation; on each segment, there are traffic elements that we infer may be related with the observed speed frequency. By comparing speed histograms, we found that the speed frequency of all segments is similar among weekdays, and then the speed frequency of a specific segment is similar regarding the day. Considering the speed frequency of all weekdays, and individual segments, the segments with different traffic elements (speed bumps, traffic lights, and landmarks) yield dissimilar traveling speeds. From this observation, two techniques were considered for modeling speed: (1) polynomial regression, where the data of each segment is modeled independently, using time as the independent term, and (2) logistic regression, with several independent variables—number of speed bumps and traffic lights, time, and two weights (from the observed speeds on street segments and weekdays). The models were implemented in algorithms, which use the modeled and historical data. With the polynomial model and Algorithm 1, it was possible to categorize correctly the travel speed in the range from 85.2 to 96.5%, depending on the segment. The multinomial logistic model and Algorithm 2 correctly predict the speed category in 90.09% of the evaluated cases. With these results, we conclude that the proposed procedure is suitable to prepare and model speed data and then to predict the speed category at a low computer processing cost. The procedure is useful to establish the relationship between traffic infrastructure and travel speed.

4.1 Future work

We contemplate as future work the development of the architecture to communicate the expected travel speed (obtained with the proposed procedure) with the users, as well as convert this knowledge in suggestions and decision-making.

Appendix A

In Algorithm 1, if $i \leq \text{deep}$ (line 3), the modeled speed of w_1 and w_2 contributes the same (each multiplied by 0.5). The case $i \geq \text{deep} + 1$ (line 6) enables the

estimation of $\bar{y}(i-1)$ and $\bar{y}(i)$ with known data from w_3 . Variables h_1 and h_2 (see lines from 9 to 14) store the average of the absolute difference between historical and modeled data, from w_1 and w_2 , respectively. h_3 (line 15) stores the absolute difference of the historical and estimated data, from w_3 and index $i-1$. The condition in line 16 verifies that the $y(i-3)$ to $y(i-1)$ speeds are nonempty, i.e., available. $h_1...h_3$ are normalized and converted to weights, named $W_1...W_3$. Because h carries the error, a greater h results in a smaller W , and so forth. In line 18, the predicted speed is calculated using the weights, the modeled speed with w_1 and w_2 , and the estimation with previous data of w_3 . If the condition in line 16 is not true, then in line 21 the speed prediction is calculated with the modeled data and new weights, without the w_3 data.

Algorithm 1

Initial conditions: deep = 3;

1. *for* $k = 0$ *to* $k = 7$
2. *for* $i = 1$ *to* $i = N_k$
3. *if* $i \leq \text{deep}$
4. $\hat{y}(i) = M_k^1(i) * 0.5 + M_k^2(i) * 0.5$
5. *end if*
6. *if* $i \geq \text{deep} + 1$
7. $\bar{y}(i-1) = y(i-2) + (y(i-2) - y(i-3))$
8. $\bar{y}(i) = y(i-1) + (y(i-1) - y(i-2))$
9. $h_1 = 0; h_2 = 0;$
10. *for* $j = 1$ *to* $j = \text{deep}$
11. $h_1 = h_1 + |y(i-j) - M_k^1(i-j)|$
12. $h_2 = h_2 + |y(i-j) - M_k^2(i-j)|$
13. *end for*
14. $h_1 = h_1 / \text{deep}; h_2 = h_2 / \text{deep}$
15. $h_3 = |y(i-1) - \bar{y}(i-1)|$
16. *if* $y(i-1) \notin \emptyset \wedge y(i-2) \notin \emptyset \wedge y(i-3) \notin \emptyset$
17. $h_1 = \frac{h_1}{h_1+h_2+h_3}; h_2 = \frac{h_2}{h_1+h_2+h_3}; h_3 = \frac{h_3}{h_1+h_2+h_3}; W_1 = \frac{1-h_1}{1-h_1+1-h_2+1-h_3};$
 $W_2 = \frac{1-h_2}{1-h_1+1-h_2+1-h_3}; W_3 = \frac{1-h_3}{1-h_1+1-h_2+1-h_3}$
18. $\hat{y}(i) = M_k^1(i) * W_1 + M_k^2(i) * W_2 + \bar{y}(i) * W_3$
19. *else*
20. $h_1 = \frac{h_1}{h_1+h_2}; h_2 = \frac{h_2}{h_1+h_2}; W_1 = 1 - h_1; W_2 = 1 - h_2;$
21. $\hat{y}(i) = M_k^1(i) * W_1 + M_k^2(i) * W_2$
22. *end if*
23. *end if*
24. *end for*
25. *end for*

Appendix B

From Algorithm 2, in lines 3 to 10 it is compared the modeled and historical speed category (from w_1 and w_2), with the historical from w_3 , to determine which is

the accurate. The number of hits of the model and the historical (for weeks 1 and 2) is stored in *score* with sub-index from 1 to 4, for the four cases. In line 12, if the probability $P_2(i)$ is greater or equal than the selected *threshold* and, if $score_2 \geq score_1$, then $S_2(i)$ is the predicted speed category. In line 14, if $P_1(i) \geq threshold$ and, if $score_1 \geq score_2$, then the predicted speed category is $S_1(i)$. If previous conditionals (line 12 and 14) are not evaluated to true, in lines from 16 to 18, the historical with the greater score, H_2 or H_1 , is the selected to predict the speed category.

Algorithm 2

```
Initial conditions:  $score_1 = 0$ ;  $score_2 = 0$ ;  $score_3 = 0$ ;  $score_4 = 0$ ;  
threshold  $\in [0.75, 0.95]$ ;  
1. for  $i = 1$  to  $i = N$   
2. if  $i > 1$   
3. if  $H_3(i - 1) == S_1(i - 1)$   
4.  $score_1 ++$ ; end if  
5. if  $H_3(i - 1) == S_2(i - 1)$   
6.  $score_2 ++$ ; end if  
7. if  $H_3(i - 1) == H_1(i - 1)$   
8.  $score_3 ++$ ; end if  
9. if  $H_3(i - 1) == H_2(i - 1)$   
10.  $score_4 ++$ ; end if  
11. end if  
12. if  $(P_2(i) \geq threshold \wedge (score_2 \geq score_1))$   
13.  $S_3(i) = S_2(i)$ ; else  
14. if  $(P_1(i) \geq threshold \wedge (score_1 \geq score_2))$   
15.  $S_3(i) = S_1(i)$ ; else  
16. if  $score_4 \geq score_3$   
17.  $S_3(i) = H_2(i)$ ; else  
18.  $S_3(i) = H_1(i)$ ; end if  
19. end if  
20. end if  
21. end for
```

IntechOpen

Author details


José Gerardo Carrillo-González^{1,2*}, Jacobo Sandoval-Gutiérrez²
and Francisco Pérez-Martínez²

1 CONACYT, Consejo Nacional de Ciencia y Tecnología, Ciudad de México, México

2 Department of Information and Communications Systems, Universidad
Autónoma Metropolitana, Estado de México, México

*Address all correspondence to: jgcarrilo@conacyt.mx

IntechOpen

© 2019 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

- [1] ONU-Hábitat. Reporte Nacional de Movilidad Urbana en México 2014-2015. 2015
- [2] Marr LC, Grogan LA, Wöhrnschimmel H, Molina LT, Molina MJ, Smith TJ, et al. Vehicle traffic as a source of particulate polycyclic aromatic hydrocarbon exposure in the Mexico city metropolitan area. *Environmental Science & Technology*. 2004;**38**(9): 2584-2592
- [3] Jiang M, Marr LC, Dunlea EJ, Herndon SC, Jayne JT, Kolb CE, et al. Vehicle fleet emissions of black carbon, polycyclic aromatic hydrocarbons, and other pollutants measured by a mobile laboratory in Mexico City. *Atmospheric Chemistry and Physics*. 2005;**5**(12): 3377-3387
- [4] Híjar M, Vazquez-Vela E, Arreola-Risa C. Pedestrian traffic injuries in Mexico: A country update. *Injury Control and Safety Promotion*. 2003;**10**(1-2):37-43
- [5] Ramos A, Silva E, Aguirre A. Fatal car accidents in the metropolitan zone of Mexico City: A geographical and temporal perspective. *Papeles de Poblacion*. 2015;**21**(86):253-282
- [6] Li R, Chai H, Tang J. Empirical study of travel time estimation and reliability. *Mathematical Problems in Engineering*. 2013;**2013**:1-9
- [7] Rupnik J, Davies J, Fortuna B, Duke A, Clarke SS. Travel time prediction on highways. In: *Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (CIT/IUCC/DASC/PICOM)*, 2015 IEEE International Conference on. 2015. pp. 1435-1442
- [8] Chacon S, Kornhauser AL. Analysis, Characterization, and Visualization of Freeway Traffic Data in Los Angeles. 2012
- [9] Imawan A, Indikawati F, Kwon J, Rao P. Querying and extracting timeline information from road traffic sensor data. *Sensors*. 2016;**16**(9):1340
- [10] Long Cheu R, Xie C, Lee D-H. Probe vehicle population and sample size for arterial speed estimation. *Computer-Aided Civil and Infrastructure Engineering*. 2002;**17**(1):53-60
- [11] Tostes AIJ, de LP Duarte-Figueiredo F, Assunção R, Salles J, Loureiro AAF. From data to knowledge: City-wide traffic flows analysis and prediction using Bing maps. In: *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*. 2013. p. 12
- [12] Silva TH, de Melo POSV, Viana AC, Almeida JM, Salles J, Loureiro AAF. Traffic condition is more than colored lines on a map: Characterization of Waze alerts. In: *International Conference on Social Informatics*. 2013. pp. 309-318
- [13] Pan J, Popa IS, Zeitouni K, Borcea C. Proactive vehicular traffic rerouting for lower travel time. *IEEE Transactions on Vehicular Technology*. 2013;**62**(8): 3551-3568
- [14] Milton J, Mannering F. The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies. *Transportation*. 1998;**25**(4):395-413
- [15] Kinderyte-Poškiene J, Sokolovskij E. Traffic control elements influence on accidents, mobility and the environment. *Transport*. 2008;**23**(1): 55-58. Available from: <http://www.tandf>

online.com/doi/abs/10.3846/1648-4142.2008.23.55-58

[16] Peden M, Scurfield R, Sleet D, Mohan D, Hyder AA, Jarawan E, et al. World Report on Road Traffic Injury Prevention. Geneva: World Health Organization; 2004

[17] Baum HM, Wells JK, Lund AK. Motor vehicle crash fatalities in the second year of 65 mph speed limits. *Journal of Safety Research*. 1990; **21**(1):1-8

[18] Nienhüser D, Gump T, Zöllner JM. Relevance estimation of traffic elements using Markov logic networks. In: *Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference on*. 2011. pp. 1659-1664

[19] Pau M, Angius S. Do speed bumps really decrease traffic speed? An Italian experience. *Accident; Analysis and Prevention*. 2001;**33**(5):585-597

[20] Wu C, Peng L, Huang Z, Zhong M, Chu D. A method of vehicle motion prediction and collision risk assessment with a simulated vehicular cyber physical system. *Transportation Research Part C: Emerging Technologies*. 2014;**47**:179-191

[21] Wang Y, Tan G, Wang Y, Yin Y. Perceptual control architecture for cyber-physical systems in traffic incident management. *Journal of Systems Architecture*. 2012;**58**(10): 398-411

[22] Work DB, Bayen AM. Impacts of the mobile internet on transportation cyberphysical systems: Traffic monitoring using smartphones. In: *National Workshop for Research on High-Confidence Transportation Cyber-Physical Systems: Automotive, Aviation, & Rail*. 2008. pp. 18-20

[23] Yuan J, Zheng Y, Xie X, Sun G. Driving with knowledge from the

physical world. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2011. pp. 316-324

[24] Chen C, Liu X, Qiu T, Sangaiah AK. A short-term traffic prediction model in the vehicular cyber-physical systems. *Future Generation Computer Systems*. 2017. In press

[25] Jianjun S, Xu W, Jizhen G, Yangzhou C. The analysis of traffic control cyber-physical systems. *Procedia - Social and Behavioral Sciences*. 2013;**96**:2487-2496

[26] Jain AK. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*. 2010;**31**(8):651-666

[27] Montazeri-Gh M, Fotouhi A. Traffic condition recognition using the k-means clustering method. *Scientia Iranica*. 2011;**18**(4):930-937

[28] Fotouhi A, Montazeri-Gh M. Tehran driving cycle development using the k-means clustering method. *Scientia Iranica*. 2013;**20**(2):286-293

[29] Saunier N, Sayed T. Clustering vehicle trajectories with hidden Markov models application to automated traffic safety analysis. In: *The 2006 IEEE International Joint Conference on Neural Network Proceedings*. 2006. pp. 4132-4138

[30] Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*. 1987;**20**:53-65

[31] Pele O, Werman M. The quadratic-chi histogram distance family. In: *Daniilidis K, Maragos P, Paragios N, editors. Computer Vision – ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part*

II. Berlin, Heidelberg: Springer Berlin
Heidelberg; 2010. pp. 749-762

[32] De Maesschalck R, Jouan-Rimbaud
D, Massart DL. The Mahalanobis
distance. *Chemometrics and Intelligent
Laboratory Systems*. 2000;**50**(1):1-18

[33] Hutcheson G. *The Multinomial
Logistic Regression Model*. Manchester:
Sage Publications; 2009

[34] Willmott CJ, Matsuura K.
Advantages of the mean absolute error
(MAE) over the root mean square error
(RMSE) in assessing average model
performance. *Climate Research*. 2005;
30(1):79-82