DOI 10.20544/HORIZONS.B.03.1.16.P16
UDC 656.13.05:004.652

ROAD TRAFFIC FLOW CHARACTERIZATION USING RELATIONAL DATABASES[1]

Ilija Hristoski
"St. Kliment Ohridski" University, Bitola
Faculty of Economics, Marksova St, 133
Prilep, Republic of Macedonia
ilija.hristoski@uklo.edu.mk

Marija Malenkovska Todorova
"St. Kliment Ohridski" University, Bitola
Faculty of Technical Sciences, Makedonska Falanga St, 33
Bitola, Republic of Macedonia
marija.malenkovska@tfb.uklo.edu.mk

Abstract

Characterizing road traffic flows at particular road sections is not only an integral but also a fundamental part of the preparation process of any traffic flow study. In this particular case, this means an exact specification of a set of traffic flow variables, including average flow rate, average headway, time mean speed, space mean speed, as well as traffic density, all being computed by travelling directions and by specific time intervals, whose optimal number and duration are *a priori* not known. In addition, this activity also includes obtaining the distribution of vehicles' relative frequencies vis-à-vis their corresponding categories. Within this paper, we propose a methodology for a complete characterization of road traffic flows at a particular road section of a rural two-lane highway during the observation time, based strictly on the usage of relational databases and SQL scripts. Specifically, we propose a framework for post-processing of raw traffic data, built on the development of a corresponding Entity-Relational (E-R) diagram, its translation into a relational schema, as well as a physical implementation of a relational database in Microsoft® SQL Server®. We also develop specific SQL scripts necessary to obtain automatically the entire gamut of traffic flow variables from database table records.

[1] Original scientific paper

Keywords – road traffic flow; workload characterization; relational databases; SQL scripts

INTRODUCTION

Traffic flow characterization is a crucial activity for both transportation systems and infrastructure planning. Such characterization relies on specific methods for obtaining traffic data, including measurements at a road point, over a short road section, along a length of the road, or by the employment of the so-called “moving observer method”. No matter which measurement method is used, traffic flow characterization is, generally, a process of quantitative describing the complex interactions between vehicles and road infrastructure, with an aim of understanding and developing an optimal road network with efficient flow of traffic and minimal traffic congestion problems. This task is accomplished by an exact specification of a set of traffic flow variables of interest, including traffic density, space and time headways, average flow rate, time mean speed, space mean speed, and a travel time over a known length of a road, and occupancy [1].

As being a part of transportation/traffic engineering, the problem of characterizing traffic flows has been studied by many authors. Most of the efforts are mainly focused on the automated acquisition of raw data (pre-processing). For instance, a usage of computer vision and artificial neural networks for automating the process of counting and classifying the vehicles has been proposed in [2], whilst a system based on multiple cameras has been proposed in [3]. In addition, a characterization of road traffic flow from measured data of speed and time-headway - the relationship between density, flow rate, and speed (post-processing) has been described in [4].

Automating the process of traffic flow characterization by any means and/or in any phase (pre- or post-processing) should result in a number of benefits, including higher consistency and reliability of the output statistics, as well as more efficient, and less time-consuming analyses. Within this paper, we focus on the post-processing of already collected traffic data, which are going to be organized into relational database tables, as well as on the methodological framework needed to carry out the road flow characterization, using SQL scripts.

PROBLEM STATEMENT

Having minded the method of measuring along a specified length of a road section for obtaining traffic data, the problem of traffic characterization can be stated as follows:

A specific road section with a known length L[2] on a rural two-lane highway consists of two traffic lanes, where vehicles drive in two, mutually opposite directions (Fig. 1). Vehicles generally belong to different categories C_i (i = 1, 2, …, M)[3]. In reality, vehicle arrivals follow the Poisson distribution (for light traffic conditions), i.e. time intervals between two consecutive vehicles are exponentially distributed. For each particular vehicle passing through the road section, one should register its category (C_i), and drive direction (1 = 'A-to-B' or 2 = 'B-to-A'). The driving speed (v_i) of the i-th vehicle [km/h] (i = 1, 2, 3, …, N) passing through the road section can be determined as a ratio between the section length (L) [m] and the section travelling time t_i [s], as in (1)[4]. The latter one is a difference of the time points when the vehicle has exited (t_2) [hh:mm:ss] and has entered (t_1) [hh:mm:ss] the road section.

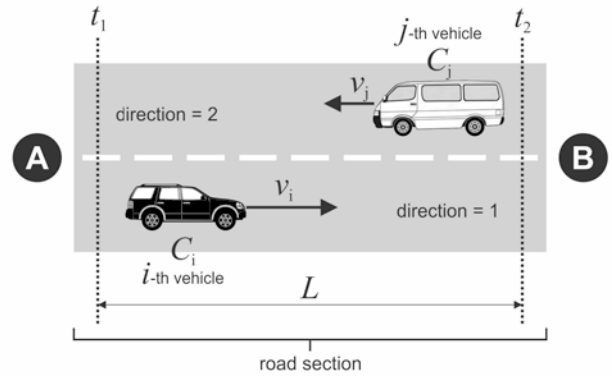

Fig.1. Graphic representation of a road section with two traffic lanes

$$v_i = \frac{L\ [\mathrm{m}]}{t_2(i)\ [\mathrm{s}] - t_1(i)\ [\mathrm{s}]} \cdot \frac{36}{10} \left[\frac{\mathrm{km}}{\mathrm{h}}\right];\ \ i = 1,\ 2,\ ...,\ N \qquad (1)$$

A traffic flow or traffic volume of a traffic stream (q) is defined as the total number of vehicles in the stream, passing a fixed reference point (i.e. a control point, a road section) over a specific unit of time, e.g. vehicles per

[2] The supposed length of the road section is L = 1,000 [m]

[3] We have taken into account three (M = 3) different categories of vehicles, including passenger cars (CatID = 1), busses (CatID = 2), and trucks (CatID = 3), whose frequencies are uniformly distributed. The driving speed of vehicles have been drawn from specific Normal distributions, i.e. passenger cars: ~N(μ = 90; σ = 6.66), busses: ~N(μ = 70; σ = 6.66), and trucks: ~N(μ = 50; σ = 6.66) [km/h]

[4] For the purposes of this study, we have artificially generated 30,000 records of traffic data, simulating real measurements on a road section during a 24-hours observation time

hour [5]. The average traffic flow rate $\overline{q}$ can be computed as in (2), where N is the total number of vehicles counted, T is the total elapsed time for all N vehicles, and $\overline{h}$ is the average headway, giving that the total elapsed time is the sum of the particular headways h_i of the i-th vehicle (i = 1, 2, …, N) [1].

$$\overline{q} = \frac{N}{T} = \frac{N}{\sum_{i=1}^{N} h_i} = \frac{1}{\frac{1}{N} \cdot \sum_{i=1}^{N} h_i} = \frac{1}{\overline{h}} \quad (2)$$

In addition, the time mean speed $\overline{v}_t$ (3) is an arithmetic mean of the speeds v_i of all i (i = 1, 2, …, N) vehicles, passing through the road segment during the observing time, whilst the space mean speed $\overline{v}_s$ (4) is the average speed of vehicles measured at an instant of time over a specified stretch of the road L [1]. Given the average traffic flow rate $\overline{q}$ and space mean speed $\overline{v}_s$, one can easily obtain the traffic density g, as in (5).

$$\overline{v}_t = \frac{1}{N} \cdot \sum_{i=1}^{N} v_i \quad (3)$$

$$\overline{v}_s = \frac{L}{\sum_{i=1}^{N} \frac{t_i}{N}} = \frac{N \cdot L}{\sum_{i=1}^{N} t_i} = \frac{N \cdot L}{\sum_{i=1}^{N} \left[t_2(i) - t_1(i) \right]} \quad (4)$$

$$g = \frac{\overline{q}}{\overline{v}_s} \quad (5)$$

Knowing this, the ultimate goal of the traffic flow characterization, regarding a particular road section, is to specify the various-length time intervals T_1, T_2, …, T_P, exhibiting mutually different, yet internally similar traffic flows $\overline{q}_j$ (j = 1, 2, …, P), based on vehicle frequencies being obtained regarding a pre-defined time unit Δt. The optimal number of time intervals, P, and the duration of each of them are *a priori* not known. Such characterization has to be carried out for each driving direction at a particular road section during the observation time (e.g. 24 hours). Besides the inclusion of average traffic flows, the characterization has to include also an evaluation of all previously mentioned traffic variables, as well as the distribution of vehicles' relative frequencies vis-à-vis their corresponding categories (Fig. 2).

In this particular case, the goal is to fully automatize and carry out the process of traffic flow characterization using a relational databases approach.

Our proposed methodology involves the following steps:

- Physical implementation of a relational database, which consists of the following sub-phases:
  - Creation of an entity-relationship (E-R) diagram;
  - Extracting the relational schema of the database out of the E-R diagram;
  - Specification of the specific data types and constraints for the fields within tables;
  - Physical implementation of the relational database in Microsoft® SQL Server®;

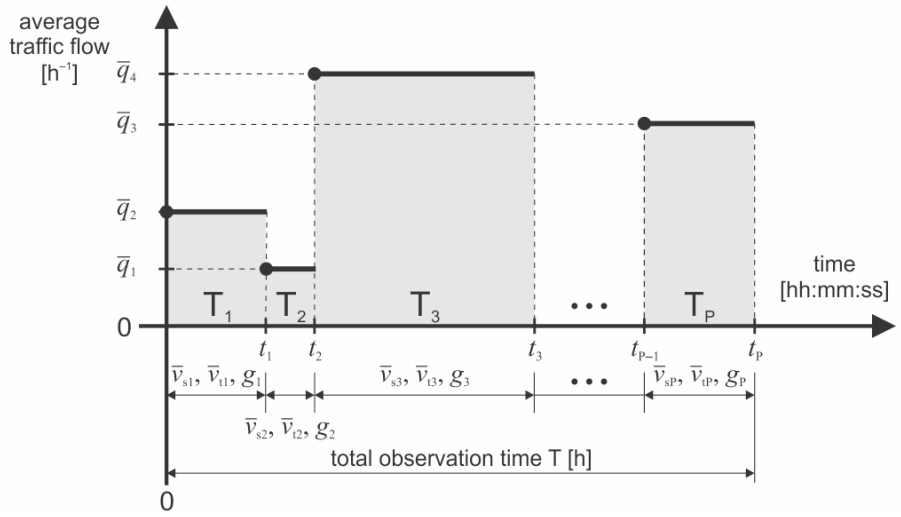

Fig.2. A schematic representation of a traffic flow characterization for a given road section and a specific driving direction

- Filling up the relational database tables with relevant data, being a result of specific control point's traffic observation;
- Generation of vehicles' frequencies, given a particular time unit, i.e. time interval (e.g. 1 minute, 5 minutes, 10 minutes, 15 minutes, 20 minutes, 30 minutes, 1 hour, …);
- Finding out the optimal number of bins, and the optimal bin width applying the Freedman-Diaconis rule or another relevant criterion;
- Mapping a corresponding bin number to each particular record about each particular vehicle, passing through the control point;
- Specification of contiguous time intervals containing the same bin number;
- Calculation of the set of variables characterizing the traffic flows in both directions, for each time interval being identified previously.

The basic table within the database, which contains raw data ready for further analysis, is the table PASSES_THROUGH (Fig. 3).

ILIJA\SQLEXPRE....PASSES_THROUGH

catID	vehicleSerNo	dateAndTime1	dateAndTime2	travelingTime	segmentSpeed	direction	cpointID
3	3985	2015-12-09 03:16:00.000	2015-12-09 03:17:03.000	63,00	57,14	2	1
1	3986	2015-12-09 03:16:02.000	2015-12-09 03:16:45.000	43,00	83,72	1	1
1	3987	2015-12-09 03:16:05.000	2015-12-09 03:16:43.000	38,00	94,74	2	1
3	3988	2015-12-09 03:16:06.000	2015-12-09 03:17:05.000	59,00	61,02	1	1
2	3989	2015-12-09 03:16:07.000	2015-12-09 03:16:56.000	49,00	73,47	1	1
2	3990	2015-12-09 03:16:13.000	2015-12-09 03:17:06.000	53,00	67,92	2	1

Fig.3. An excerpt from the table PASSES_THROUGH, showing raw data

The above table's data resembles a real situation, schematically depicted in Fig. 5[5]. These data also corresponds to the situation presented in Fig. 1.

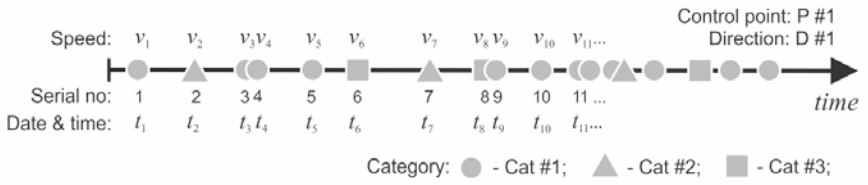

Fig.5. Schematic representation of raw data in the table PASSES_THROUGH

From here, the next step starts with a generation of vehicles' frequencies in table ANALYSIS, by driving direction, from the records already present in table PASSES_THROUGH, given a particular time unit, i.e. a time interval that will specify the granularity/resolution (e.g. 1 minute, 5 minutes, 10 minutes, 15 minutes, 20 minutes, 30 minutes, 1 hour, …). We have used a time unit of Δt = 15 minutes, as the most appropriate period for aggregating available traffic data, which is, also, a recommended time frame vis-à-vis the optimal representation of traffic flows. The schematic representation of this step is depicted in Fig. 6, whilst the first 10 records of the table ANALYSIS are presented in Fig. 7[6].

The next step is the specification of the optimal number of bins and the optimal bin width, for each driving direction, within the table BINS_DEFINITION, based on the records in the table ANALYSIS. The aim is to define the boundaries of a frequency distribution with equally spaced

[5] There is no strict 1:1 mapping of the table values presented in Fig. 4 vis-à-vis the linear distribution of vehicles in Fig. 5.

[6] There is no strict 1:1 mapping of the table values presented in Fig. 7 vis-à-vis the linear distribution of vehicles in Fig. 6.

bins. One can use his/her rule of thumb, however, there are several statistical methods for obtaining the optimal values of these parameters, including the Sturges' rule, the Scott's rule, as well as the Freedman-Diaconis' rule. In our calculations, we have used the last one [6]. Speaking about 'optimality', there is always a trade-off between the number of bins (or bin width), vis-à-vis the granularity, i.e. the resolution of the frequency distribution. In fact, if there are too many bins, one can get a 'broken comb' look, which does not give a sense of the distribution. On the contrary, if there are too few bins, the frequency distribution would not also portray the underlying data very well.

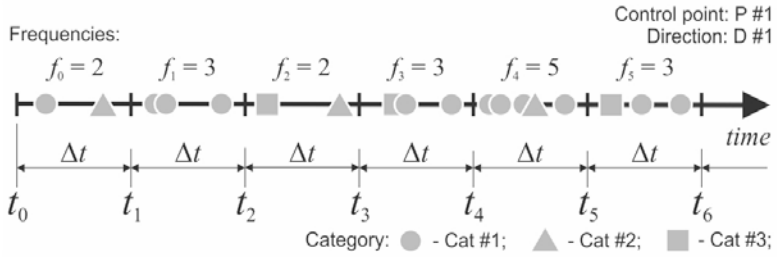

Fig.6. Obtaining vehicles' frequencies in time intervals of Δt = 15 minutes, by a specific driving direction

ILIJA\SQLEXPRESS...IC - dbo.ANALYSIS

	cpointID	direction	absolute_time	frequency	bin
▶	1	1	1	146	NULL
	1	1	2	158	NULL
	1	1	3	154	NULL
	1	1	4	158	NULL

Fig.7. The frequency of vehicles that belong to all categories, driving in direction 1, in the first 60 minutes ($4 \times \Delta t$) of the observation period (table ANALYSIS). The column *absolute_time* contains the values of k (k = 1, 2, 3, …, 96), where 0 [min] < $k \times \Delta t \leq$ 1,440 [min]

The Freedman-Diaconis' rule is proposed to be used to select the optimal size of the bins (bin width) when the number of observations in the sample is large. The general equation for the rule is given by (6):

$$w = \frac{2 \cdot IQR(x)}{\sqrt[3]{n}} \quad (6)$$

In (6), w is the optimal bin width, $IQR(x)$ is the inter-quartile range of the data, i.e. $Q3 - Q1$, and n is the number of observations in the sample x, which consists of the vehicles' frequencies driving in both directions.

The rule is based on the goal of minimizing the sum of squared errors between the histogram bar height and the probability density of the underlying distribution. As such, it generally produces more bars. Therefore, we have deliberately 'widen' the originally computed optimal bin width for 5 times, which still yields pretty good granularity. An excerpt from the table BINS_DEFINITION is given in Fig. 8.

ILIJA\SQLEXPRESS....BINS_DEFINITION

cPointID	direction	binNo	low	high
1	1	1	-12,00	20,00
1	1	2	20,01	52,01
1	1	3	52,02	84,02
1	1	4	84,03	116,03
1	1	5	116,04	148,04
1	1	6	148,05	180,05
1	1	7	180,06	212,06

Fig.8. Obtaining the parameters of the frequency distribution (bin width boundaries: fields *low* and *high*, and the bin number: field *binNo*), for all categories of vehicles driving in the same direction (direction = 1)

Once bin definitions for each driving direction have been written in the table BINS_DEFINITION, the next step is to assign the corresponding bin numbers to the frequencies within the table ANALYSIS (Fig. 9).

ILIJA\SQLEXPRESS...IC - dbo.ANALYSIS

cpointID	direction	absoluteTime	frequency	binNo
1	1	1	146	5
1	1	2	158	6
1	1	3	154	6
1	1	4	158	6
1	1	5	132	5
1	1	6	158	6
1	1	7	145	5
1	1	8	146	5
1	1	9	138	5
1	1	10	169	6

Fig.9. The assigned bin numbers (field *binNo*) for all vehicles, driving in direction 1, in the first 150 minutes ($10 \times \Delta t$) of the observation period (table ANALYSIS)

The schematic representation of this step is depicted in Fig. 10[7].

The next step is to specify the contiguous time intervals containing the same bin number within the table CHARACTERIZATION, based on the

[7] There is no strict 1:1 mapping of the table values presented in Fig. 9 vis-à-vis the linear distribution of vehicles in Fig. 10.

records of the table ANALYSIS, for each driving direction. As a result, the fields *sTime* and *eTime* have been defined, which denote the boundaries of each adjacent period belonging to a different bin number (Fig. 11).

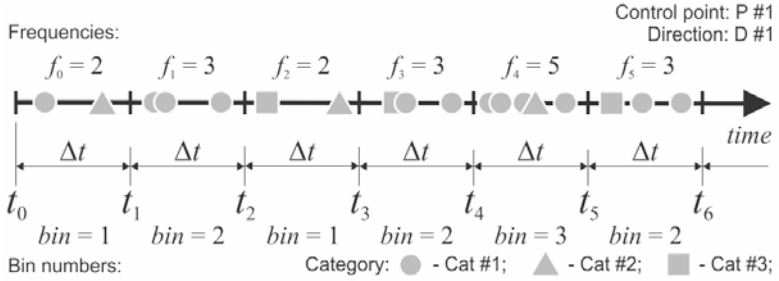

Fig.10. Assigning bin numbers to specific time intervals, based on the frequencies of vehicles, for each period of Δt = 15 minutes, and specific driving direction

ILIJA\SQLEXPRE...HARACTERIZATION

	cPointID	direction	periodSerNo	startTime	endTime	sTime	eTime
▶	1	1	1	1	1	00:00:00	00:14:59
	1	1	2	2	4	00:15:00	00:59:59
	1	1	3	5	5	01:00:00	01:14:59
	1	1	4	6	6	01:15:00	01:29:59
	1	1	5	7	9	01:30:00	02:14:59
	1	1	6	10	10	02:15:00	02:29:59

Fig.11. Specification of the time boundaries (fields *sTime* and *eTime*) with adjacent periods, for a specific driving direction, for the first six intervals/periods [00:00:00 – 02:29:59] (table CHARACTERIZATION)

In the final step, all variables of interest are calculated by specific periods, being identified in the previous step, and all corresponding fields in the table CHARACTERIZATION are updated, including *averageFlowRate*, *averageHeadway*, *timeMeanSpeed*, *spaceMeanSpeed*, and *trafficDensity*. The values in the fields *passengerCars*, *busses*, and *trucks* comprise the relative frequencies' distribution of vehicles by their category, such that their sum is always 1.00 for each record, i.e. each period (Fig. 12). All these calculations have been done taking into account data records from the table PASSES_THROUGH, and the fields *sTime* and *eTime* in the table CHARACTERIZATION.

sTime	eTime	averageFlowRate	averageHeadway	timeMeanSpeed	spaceMeanSpeed	trafficDensity	passengerCars	busses	trucks
00:00:00	00:14:59	584,00000	0,00171	71,59103	66,73438	8,75111	0,33	0,34	0,33
00:15:00	00:59:59	1880,00000	0,00053	71,09579	66,16612	28,41333	0,33	0,33	0,33
01:00:00	01:14:59	528,00000	0,00189	70,77114	66,11938	7,98556	0,30	0,45	0,25
01:15:00	01:29:59	632,00000	0,00158	70,39589	65,27427	9,68222	0,39	0,26	0,35
01:30:00	02:14:59	1716,00000	0,00058	71,97937	67,22967	25,52444	0,37	0,34	0,29
02:15:00	02:29:59	676,00000	0,00148	71,69586	67,00441	10,08889	0,34	0,34	0,32

Fig.12. Road traffic characterization for vehicles driving in direction 1, in the first six intervals [00:00:00 – 02:29:59] of the observation period
(table CHARACTERIZATION)

CONCLUDING REMARKS

Traffic flow characterization can be accomplished in a myriad of ways, depending on the acquisition methodology (both software and hardware usage), as well as the format and volume of raw data. In addition, post-processing of acquired data can include usage of dedicated software, spreadsheets, statistical software, but in most cases, it is based on the usage of relational databases, which offer unprecedented storage capacity, processing efficiency, increased reliability, minimized redundancy and maximized performances.

The proposed methodology has been tested on a set of artificially generated 30,000 records; it has proven to be both consistent and reliable. Given the proposed conceptual database model, the set of developed SQL scripts provides a dependable and unified framework that can be successfully applied on any underlying real dataset, offering a great flexibility vis-à-vis the granularity of the traffic characterization.

The proposed framework remains a solid basis for further improvements, as well as for adding programming features in order to build up a stand-alone software solution.

REFERENCES

[1] G-L. Chang. (2016, February 11th). Traffic Flow Theory (Traffic Stream Characteristics) (–) [online]. Available: http://ocw.nctu.edu.tw/course/ ftf011/Lec1-01.pdf

[2] M. M. Chan and M. C. Ríos. (2016, January 5th). "Technical Report on Automated Traffic Characterization" (2007) [online]. Available: https://772f25bf9a0f2baf81f05981b62d265385bb9a4e.googledrive.com /host/0B92aE0wdpf7TdnVHSTJaMWgxYTg/2007/pisis-2007-04.pdf

[3] R. Khoshabeh, T. Gandhi and M. M. Trivedi. (2016, January 5th). Multi-camera Based Traffic Flow Characterization & Classification (2007) [online], Available: http://cvrr.ucsd.edu/publications/2007/ RKhoshabeh_ITSC07_multicam.pdf

[4] S. Suzuki *et al.*, "Characterization of road traffic flow from measured data of speed and time-headway -relationship between density (k), flow rate (q) and speed (V)," presented at the SICE Annual Conference 2007 in Takamatsu (SICE2007), Takamatsu, Japan, 2007, pp. 1657–1661.

[5] P. Chakroborty and A. Das, "Traffic Flow," in *Principles of Transportation Engineering*, PHI Learning Private Ltd, New Delhi, India, 2012, Ch. 4, pp. 55–122.

[6] D. Freedman and P. Diaconis, "On the histogram as a density estimator: L_2 theory," in *Probability Theory and Related Fields*, Heidelberg: Springer Berlin, Germany, 1981, pp. 453–476.