Traffic flow estimation using higher-order speed statistics

Submission date: 07/31/2012

Edouard Bulteau[1]
Master's student, Systems Engineering
Department of Civil and Environmental Engineering
University of California, Berkeley
Berkeley, CA 94720, USA
edouard.bulteau@polytechnique.org

Romain Leblanc
Master's student, Systems Engineering
Department of Civil and Environmental Engineering
University of California, Berkeley
Berkeley, CA 94720, USA
romain.leblanc@polytechnique.org

Sebastien Blandin
Research Scientist
IBM Research Collaboratory, Singapore
sblandin@sg.ibm.com

Alex Bayen
Associate Professor, Systems Engineering
Department of Electrical Engineering and Computer Sciences
Department of Civil and Environmental Engineering
University of California, Berkeley
642 Sutardja Dai Hall, Berkeley, CA 94720-1764, USA
bayen@berkeley.edu

Word count:
Number of words (Body): 4,125
Number of words (Abstract): 190
Number of figures: 7 (250 each)
Number of tables: 3 (250 each)
Total: 6,815

[1]Corresponding author

Abstract

In this article, we consider the problem of estimating traffic flow on a multi-lane road using a set of point speeds, either crowd-sourced or collected from the fixed infrastructure. We specifically investigate the relation between higher-order speed moments and the expected value of traffic flow. The algorithm proposed is based on the selection of optimal covariates constructed as speed moments, for a class of conditional mean predictors. The second contribution of this article consists in the analysis of specific components of the speed moments with significant correlation with flow values. In particular, we show that for more than 75% of the fixed sensing devices considered, the correlation coefficient between the inter-lanes speed variance and the aggregate flow is more than 0.75. Additionally, for more than 70% of these fixed sensing devices the lane speed variance increases with flow. The third contribution of this article consists of identifying the explanatory features for the high correlation between speed moments and flow values. The algorithms presented in this article are trained and tested on a large dataset from the Mobile Millennium system, collected in the Bay Area from August 2009 to October 2009.

1 Introduction

1.1 Motivation

Macroscopic traffic flow modeling is at the core of real-time traffic monitoring. The main assumption of macroscopic traffic modeling is that vehicles traveling on a road behave similarly to a fluid in a pipe, hence the state of traffic can be completely captured by three macroscopic variables: the *density* ρ of vehicles, the *flow* q of traffic and the *space-mean speed* v of vehicles (Cassidy and Coifman (*1*), Knoop et al. (*2*)).

The *Lighthill-Whitham-Richards* (LWR) equation, expressing the conservation of mass, is classically used to model traffic flow. By definition, ρ, q and v also satisfy the relation $q = \rho v$. However, a third relation between these three quantities is needed in order to have a well-defined system.

The so-called *fundamental diagram* (Newell (*3*), Daganzo (*4*)) of traffic flow expresses an empirical relation between density, flow and speed defined for any two couples of these three quantities. Lighthill and Whitham (*5*) describe in their classical work the fundamental diagram as the fact that "at any point of the road the flow (vehicles per hour) is a function of the concentration (vehicles per mile)". Historically, research work has been focused on modeling the relation for the couple density-flow.

The recent explosion of mobile devices (Mobile Millennium (*6*)) measuring speeds requires innovative approaches allowing the estimation of traffic flow from reported speeds only. This problem is a major open problem in traffic flow modeling. For the triangular model, it it not possible to infer flow from speed as flow is a multi-valued function of speed, in particular for high speeds.

Recent work from Blandin et al. (*7*) has shown a promising venue for estimating traffic flow from speed via the computation of individual speed variances. The goal of this work is to investigate innovative approaches to further push the research frontier in this topic. Using speed and flow measurements recorded by 116 fixed radar sensors in the Bay Area, in this article we investigate the design of a speed-flow mapping based on statistical analysis.

1.2 General background

Research on fundamental relations between macroscopic traffic quantities date back to the work of Greenshields (8), who proposed a parabolic flow-density relation corresponding to a linear speed-density relation. Over the past decades, traffic scientists developed a wide range of mathematical models for uninterrupted traffic flow. In particular, the *triangular fundamental diagram* (Daganzo (*4*)) models the relation between traffic flow and density by a triangular relation.

In the triangular fundamental diagram, the uncongested phase is characterized by an increasing linear relation between flow and density. The congested phase is characterized by a decreasing affine relationship between flow and density. In the uncongested phase, for this model, the speed is constant equal to the free-flow speed, whereas in the congested phase, the speed is decreasing as the density increases.

In the uncongested phase, the flow is a multi-valued function of speed. This models the fact that up to a certain density, the spacing between vehicles traveling on a roadway segment is sufficiently large for drivers not to be affected by other vehicles. Thus, while the speed-to-flow conversion is straightforward in the congested phase, this very conversion is impossible in the uncongested phase as speed is theoretically constant at the free-flow speed for the whole phase.

1.3 Related work

While a lot of efforts focus on estimating vehicle density (Alvarez-Icaza et al. (*9*), Atrimy (*10*), Coifman (*11*), Gazis and Knapp (*12*), Gazis and Szeto (*13*), Jerbi et al. (*14*), Panichpapiboon and Pattara-atikom (*15*)), analyzing the relation between traffic flow and individual traffic speed is much less documented. The analysis of the relation between traffic flow and individual traffic speeds and the ability to estimate flow using using a set of point speeds, either crowd-sourced or collected from the fixed infrastructure, currently constitutes a major research question for traffic modeling theory.

Jørgensen (*16*) analyzes a large amount of combined speed and flow observations collected for the Ring 3 Motorway of the greater Copenhagen area over a 2 month period. The measurements are analyzed with the aim of extracting information which can be used to realistically and reasonably model speed and flow in a road traffic assignment model. It is observed that the US Bureau of Public Road (BPR) speed-flow curve bpr (*17*) is an adequate average description of the non queuing conditions in static assignment models.

Blandin et al. (*7*) show that the conventional speed-flow mapping produces significantly more accurate results than a speed variance-flow approach in the case of an increasing uncongested branch in (q, v) coordinates. However, a classical assumption of the free-flow phase is a flat uncongested branch in (q, v) coordinates. In that case, (7) show that the speed variance regression is more accurate than the classical approach.

Wang et al. (*18*) focus essentially on the speed variance as a function of the density. They show that the empirical variance takes a parabolic shape which first increases to a local maximum and then decreases as traffic density increases. Although some flow estimation schemes are proposed, most of them are not designed to be used in the uncongested phase.

The remainder of this article is organized as follows. Section 2 describes the dataset used in the subsequent analysis. Section 3 presents the mathematical formulation of the estimation problem and introduces the covariates considered. Numerical and experimental results for the flow estimation algorithm are provided in Section 4. The features of the covariate that gives the best mapping with the flow are detailed in Section 5. Finally, Section 6 concludes the paper.

2 Large scale Bay Area dataset

2.1 Sensing devices

The dataset used for the analysis presented in this paper consists of speeds and flows of vehicles traveling on highway and freeway segments, as recorded by 116 fixed sensing devices r_k ($k = \{1, .., 116\}$) in the Bay Area, California between August and October 2009.

2.2 Measurements

The sensing devices output speed (miles per hours) and flow (vehicles per minute) measurements per lane of roadway at a sampling period $\Delta t = 1$ min. These measurements are made available in the *Mobile Millennium* system hosted at the Center of California for Innovative Transportation. For every sampling time $m\Delta t$ (with $m \in \mathbb{N}$ the time index), we have a direct access to $v_i(m\Delta t)$ the average speed on lane i during the time interval $[(m-1)\Delta t, m\Delta t]$, with $n_i(m\Delta t) = q_i(m\Delta t)\Delta t$ the number of cars measured on lane i during the previous period Δt. Additional static information at the location of the sensing device is also available, for instance the number of lanes L and their characteristics (HOV or not), the direction the sensing device faces (N, NE, E, SE, S, SW, W or NW) and the speed limit.

3 Mathematical formulation

3.1 Problem Statement

In this section we formulate the problem of finding a mapping between flow values and traffic variables available in an infrastructure-free environment. For a given dataset $\mathcal{D}$ of joint speed flow measurements, we build a set $X(\mathcal{D})$ of traffic quantities derived from point speeds. The estimation problem, posed as an optimization problem, reads as:

$$x_{opt} = \underset{x \in X(\mathcal{D})}{\operatorname{argmin}} \|\hat{q}(x) - q\| \tag{1}$$

where:

- q denotes the measured flow,
- $\hat{q}(\cdot)$ denotes a flow estimator,
- $\|.\|$ denotes a norm chosen to measure the error.

3.2 Covariates considered

Finding a mapping between the flow and speed data for the uncongested state requires to make sure that the considered covariates are built using speed data from the uncongested state. The critical speed $\overline{v}_c$ is defined as the speed at which the congested and uncongested branches of the speed/flow diagram intersect and partitions the set of speed data V into two subsets $C = [0, \overline{v}_c[$ and $U = [\overline{v}_c, +\infty[$ is the set of speed data in the free-flow state. The critical velocity $\overline{v}_c$ is computed from observed data for each sensing device and the covariates considered in the remainder of this article are built at each time of interest for which the average speed belongs to U.

In this section we present the most relevant covariate of each category, whose performance in terms of flow estimation are compared in Section 4.2. The two covariates we present are the flow-weighted speed $\overline{v}_q$ and the speed variance over time of the individual speeds Var_T^{is}.

The flow-weighted speed at time index m reads:

$$\overline{v}_q(m\Delta t) = \frac{\sum_{i=1}^{L} n_i v_i(m\Delta t)}{\sum_{i=1}^{L} n_i(m\Delta t)} \tag{2}$$

where:

- L is the number of lanes of the roadway segment at the sensing location,
- $v_i(m\Delta t)$ is the average speed on lane at time index m,
- $n_i(m\Delta t)$ is the number of cars on lane i at time index m.

The second covariate considered, suggested by Blandin et al. (*7*) is the variance over time of the aggregate speeds. It is the variance of the speeds measured on time interval. In this work, the size T of the time intervals is taken to be 11 min to be consistent with Blandin et al. (*7*). This choice is a trade-off between a short interval size required for consistency of the assumption of stationary traffic state, and a large number of speed observations required for accurate variance computation. In this article we focus our work on finding a mapping between the flow and variables built from individual speeds.

The variance over time of the individual speeds at time index m over a period T reads:

$$Var_T^{is}(m\Delta t) = \frac{1}{N-1} \sum_{i=1}^{L} \sum_{j=m-k}^{m+k} n_{ij}(v_{ij} - \overline{v}(m\Delta t))^2 \tag{3}$$

where:

- $[(m-k)\Delta t, (m+k)\Delta t]$ is the time interval of interest around the time index m. Note that $(2k+1)\Delta t = T$ and for our analysis we take $k = 5$ to have $T = 11$ min.
- L is the number of lanes of the roadway segment at the sensing location,
- v_{ij} is the average speed of cars on lane i at time index j,
- n_{ij} is the number of cars with speed v_{ij},
- $N = \sum_{i=1}^{L} \sum_{j=m-k}^{m+k} n_{ij}$ is the total number of cars observed in the time interval of interest,

- $\overline{v}(m\Delta t)$ is the flow weighted speed in the time interval of interest: $\frac{1}{N}\sum_{i=1}^{L}\sum_{j=m-k}^{m+k} n_{ij}v_{ij}$

The superscript is and subscript T are used to remind the reader that this variance aims to evaluate the variance over the period T of the individual speed (is). Note that this expression of the variance assumes that all the cars on lane i at time index j drive at the same speed v_{ij}, the average speed over this set of car. This assumption is discussed in 5.2.

This variance can be decomposed into components involving explicitly the variance over time of the individual speeds on each unique lane. Indeed for each lane i, we can introduce $\overline{v}_i(m\Delta t)$ the average speed on lane i during the time interval of interest and n_i the number of cars passing on lane i during the same interval:

$$\overline{v}_i(m\Delta t) = \frac{1}{n_i}\sum_{j=m-k}^{m+k} n_{ij}v_{ij}$$

$$n_i = \sum_{j=m-k}^{m+k} n_{ij}$$

Using (3) we can write $Var_T^{is}(m\Delta t)$ as:

$$\begin{aligned} Var_T^{is}(m\Delta t) &= \frac{1}{N-1}\sum_{i=1}^{L}\sum_{j=m-k}^{m+k} n_{ij}[(v_{ij}-\overline{v}_i(m\Delta t)) + (\overline{v}_i(m\Delta t)+\overline{v}(m\Delta t))]^2 \\ &= \frac{1}{N-1}\sum_{i=1}^{L}[\sum_{j=m-k}^{m+k} n_{ij}(v_{ij}-\overline{v}_i(m\Delta t))^2 + \sum_{j=m-k}^{m+k} n_{ij}(\overline{v}_i(m\Delta t)-\overline{v}(m\Delta t))^2] \\ &\quad + \frac{2}{N-1}\sum_{i=1}^{L}[(\overline{v}_i(m\Delta t)+\overline{v}(m\Delta t))\sum_{j=m-k}^{m+k} n_{ij}(v_{ij}-\overline{v}_i(m\Delta t))] \end{aligned} \quad (4)$$

We can introduce the individual speed variances over time for a single lane i:

$$Var_{T,i}^{is}(m\Delta t) = \frac{1}{n_i-1}\sum_{j=m-k}^{m+k} n_{ij}(v_{ij}-\overline{v}_i(m\Delta t))^2$$

In (4) the first term can be expressed with the lane variances, and the last term can be simplified. The total variance can thus be written as:

$$Var_T^{is}(m\Delta t) = \frac{1}{N-1}\sum_{i=1}^{L}(n_i-1)Var_{T,i}^{is}(m\Delta t) + \frac{1}{N-1}\sum_{i=1}^{L} n_i(\overline{v}_i(m\Delta t)-\overline{v}(m\Delta t))^2 \quad (5)$$

Equation (5) shows that the total individual speed variance over time is composed of the weighted sum of the variances on each lane, and of the weighted sum of the lanes average speed deviation from the overall speed average. This last component can be seen as a variance of the average speed over the lanes. The respective weights of each of these two components in the total individual speed variance is investigated in 5.2.

3.3 Flow estimation method

Our aim is to find a mapping between the variables described above and the flow, to then be able to produce flow estimations from raw speed data. The accuracy of the method used is determined using the root mean square error. For a given predictor variable x (in our case either flow-weighted speed or speed variance), we construct the flow predictor as:

$$\hat{q}(x_i) = E(q|\lfloor x \rfloor = x_i) \quad (6)$$

where $\lfloor x \rfloor$ is the floor of the variable x. The conditional mean estimator is justified by Sherman (*19*) in the following lemma:

Lemma 1 *If the probability density function associated with a variable q is symmetric around the mean and unimodal then $E(q)$ is the optimal estimator of q in the sense of the quadratic error.*

In our case for each category x_i we assume that the probability density of q verifies these properties, hence the conditional mean $E(q|x = x_i)$ is the optimal estimator of q when $x = x_i$. Using equation (6), equation (1) becomes:

$$x_{opt} = \underset{x \in X(\mathcal{D})}{\operatorname{argmin}} \| E(q|\lfloor X \rfloor = x) - q \| \tag{7}$$

where:

- q denotes the measured flow,
- $\hat{q}(\cdot)$ denotes the estimator described in equation (6),
- $\|.\|$ denotes a norm chosen to measure the error.

3.4 Algorithm

For each of the two covariates proposed we use the method illustrated in Figure 1 for assessing the estimator performance for flow estimation. We use a cross-validation technique, where the cross-validation subsets consist of the months of August 2009 and September 2009 - October 2009.

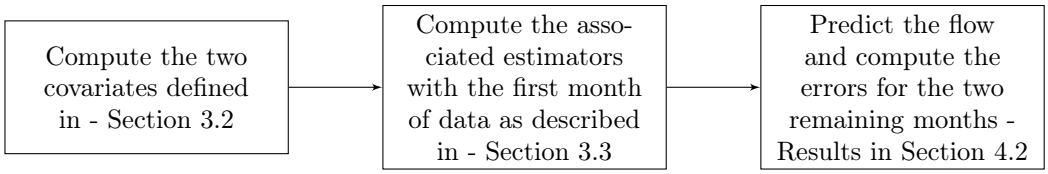

FIGURE 1 Methodology for best mapping selection

4 Flow estimation results

4.1 Properties of conditional flow distributions

In this section, we empirically validate the assumption made in the previous section on the symmetry around the mean, and unimodality, of the conditional probability of the flow. Figure 2 represents the quartiles and the mean of the conditional flow distribution as a function of the speed variance (Figure 2a) and as a function of the flow-weighted speed (Figure 2b) for different sensing devices. For the two covariates the symmetry of the quartiles around the median (which stays close to the mean) partially validates the use of Lemma 1. In order to assess the unimodality of the conditional distributions, we propose to represent in Figure 3 the conditional flow distributions and their mean for different values of speed variance (Figure 3a) and flow-weighted speed (Figure 3b).

The majority of the 116 sensing devices exhibit unimodal conditional distributions with a good symmetry around the mean in the case of the speed variance. The conditional distributions for the flow-weighted speed are much less well-behaved (Figure 3b).

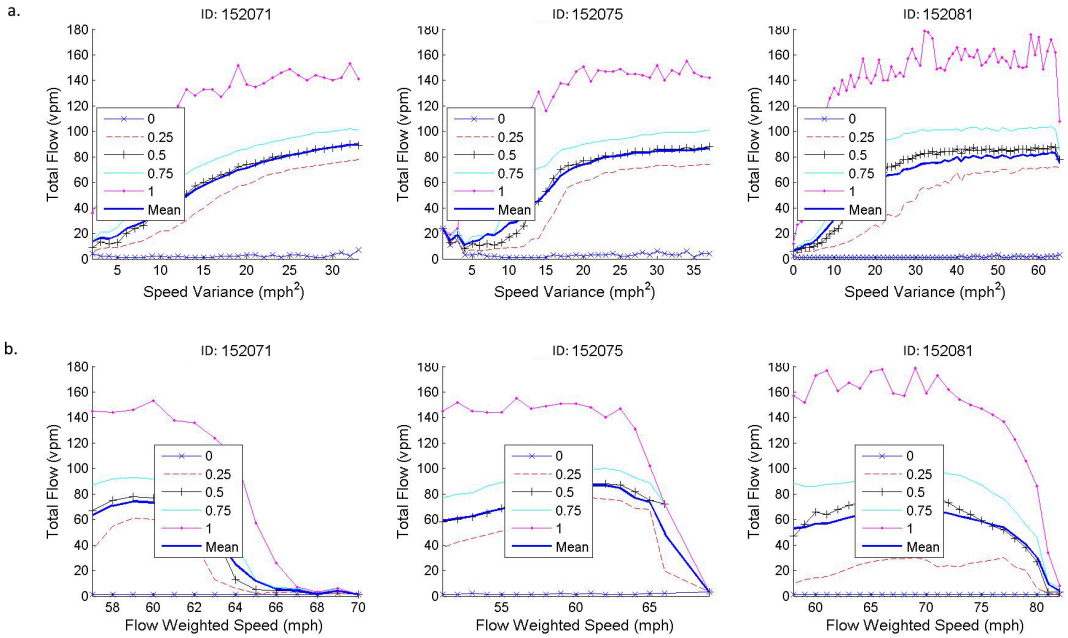

FIGURE 2 Quartiles and mean of the flow distribution conditioned by: a. The speed variance b. The flow-weighted speed. Results are given for three sensing devices (identifiers are displayed at the top of each graph)

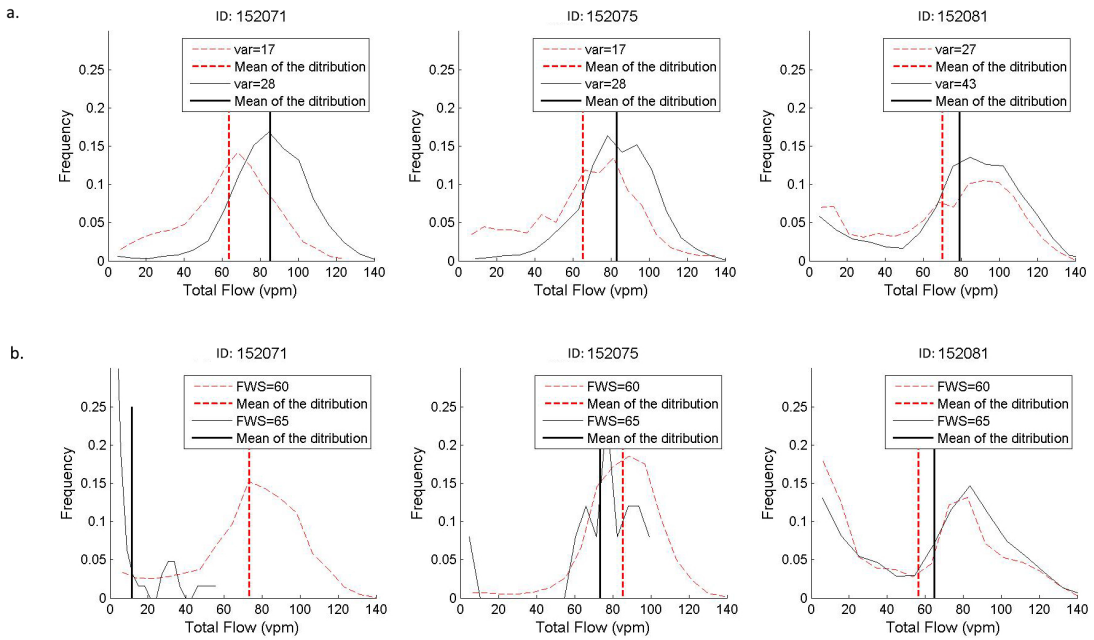

FIGURE 3 Flow distribution for given: a. Speed variances b. Flow-weighted speeds

These results show that:

- the first and third quartiles of the conditional flow distribution are closer to the mean and median in the case of the speed variance,
- the distribution of the flow conditioned by the speed variance satisfy the assumption of unimodality and symmetry around the mean which does not seem to be always the case for the distribution conditioned by the flow-weighted speed.

In the following section we present the performance results for these two estimators.

4.2 Flow estimation errors

For each sensing device r the root mean squared error (20) of the estimator $\hat{q}(\cdot)$ is given by:

$$\varepsilon_r = \sqrt{\frac{\sum_{i=1}^{N}(\hat{q}(x_i) - q_i)^2}{N-1}}$$

where:

- $\hat{q}(x_i)$ is the estimator of the flow at time of measurement i
- q_i is the measured flow at time i
- N is the total number of flow measurements in uncongested state

We represent in Figure 4 the cumulative distribution of the error over all sensing devices for the two methods. The blue plain line corresponds to the speed variance method, whereas the red dashed line corresponds to the flow weighted speed method. We notice that the plain line is always above the dashed line. It indicates that for a given error ε the proportion of sensing devices which have an error inferior to ε is always higher for the speed variance method than for the flow weighted speed method.

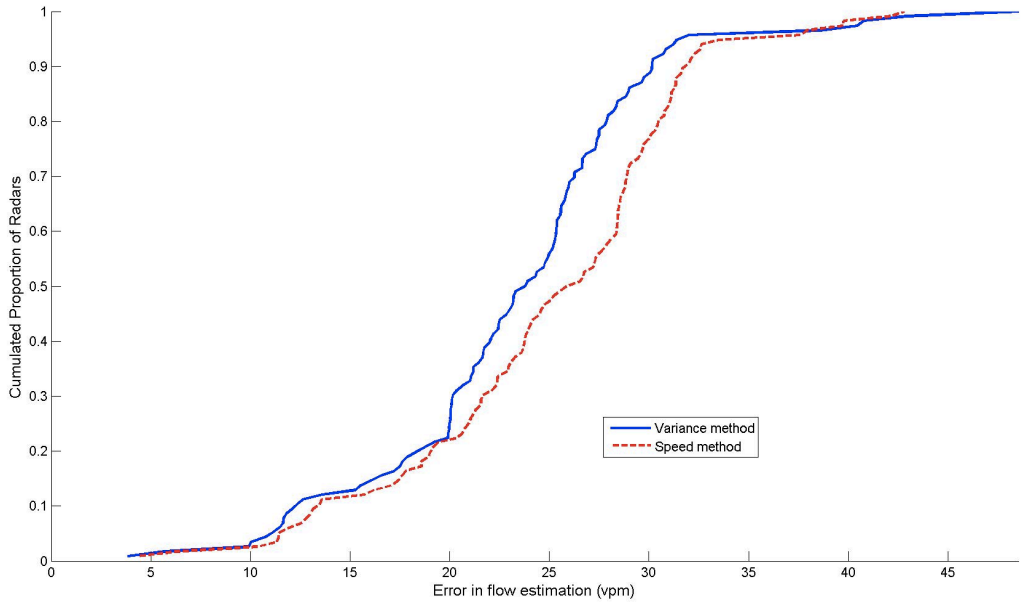

FIGURE 4 Cumulative distribution of the error

In the rest of the article, we investigate more thoroughly the relation between flow and speed variance.

5 Features of the speed variance over time

5.1 Correlation with the flow

In this section we analyze the relation between flow and speed variance by computing the distribution of the correlations between these two quantities for each sensing device. Figure 5 represents the flow estimator as a function of the speed variance. For all the presented sensing devices, a clear positive relation between the two variables is displayed.

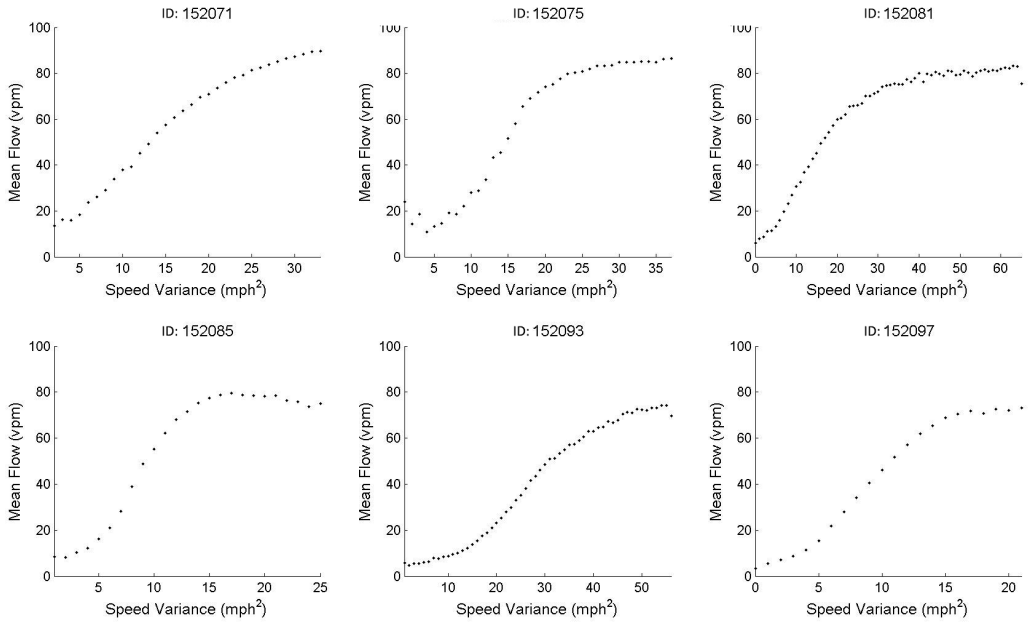

FIGURE 5 Estimator of the flow vs. speed variance over time

In order to get a better insight into the relation between speed variance and flow, we compute the Pearson's correlation coefficients between the speed variance and the mean flow conditioned by the speed variance. We represent the distribution of these coefficients over the 116 sensing devices in Figure 6.

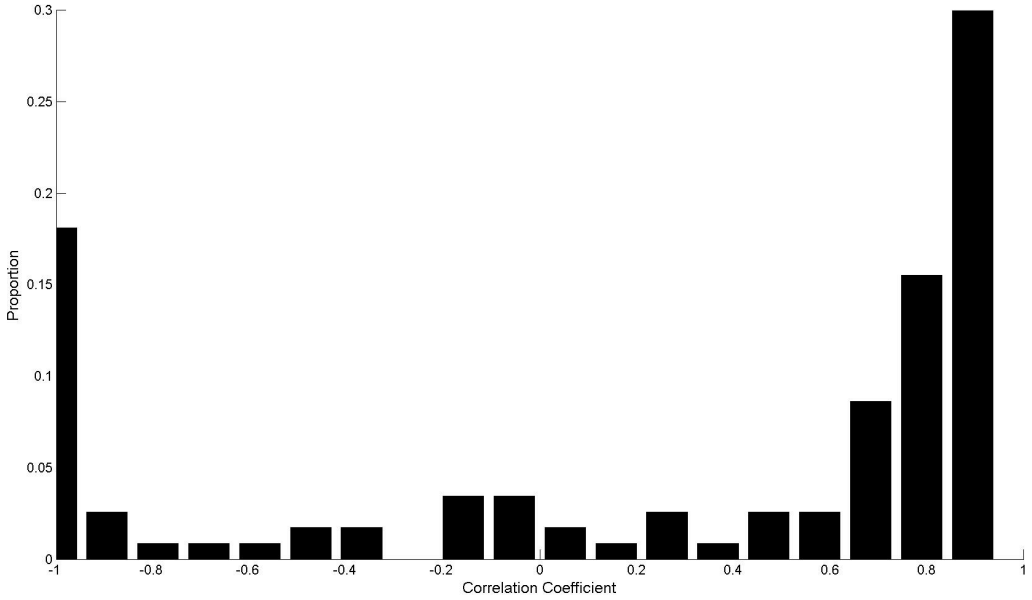

FIGURE 6 Distribution of the correlation coefficients

Figure 6 illustrates that for more than 75% of the sensing devices, the absolute correlation coefficient between the two described variables is more than 0.75. The histogram in Figure 6 illustrates both the strong relation between speed variance and the mean of the conditional distribution of the flow and the relevance of the mapping introduced in equation (1).

It is also important to notice that for most of these sensing devices (70%) individual speed variance increases with the flow. To understand this phenomenon and what could be the physical reasons of this relation, it is necessary to consider specific components of the speed variance.

5.2 Physical interpretation

We showed in (5) that the total variance results from the addition of two components: the first one is a weighted sum of the variances over time of the speed on each lane (we call it component 1) and the second one is a weighted variance over the lane of the lane average speed (called component 2). The first component quantifies the variability of the cars speed on the lanes over a given time interval. The second component represents how different lanes differ in terms of the average lane speed of the cars.

In order to explore the impact of each of these components in the total variance, we compute separately each of them and look at the ratios $r_1 = component1/totalvariance$ and $r_2 = component2/totalvariance$. We compute for each sensing device the mean of these ratios for all the uncongested states and analyze the distribution of these means over all sensing devices. We find that the median of the ratio r_2 is 60%. It shows that the second component explains slightly more of the total variance than the first component.

A more significant result describes the dependence of r_2 on the value of the total variance. Indeed Figure 7 represents the mean of r_2 conditioned by the value of the total variance. We find that for all the sensing device, an increase of the total variance is associated with an increase of r_2. This result seems to indicate that the increase of the total variance is mainly driven by an increase of the speed variance between the lanes. Thus the physical process which leads to an increase of the variance is less due to the increase of the speed variance within one same lane than to the difference of average speeds between the lanes.

The fact that the speed variance over time is not the main driver of the total variance justifies the assumption made in the definition of the total variance in equation (3) that along each time period Δt all the cars on a same lane have identical speed equal to their average speed.

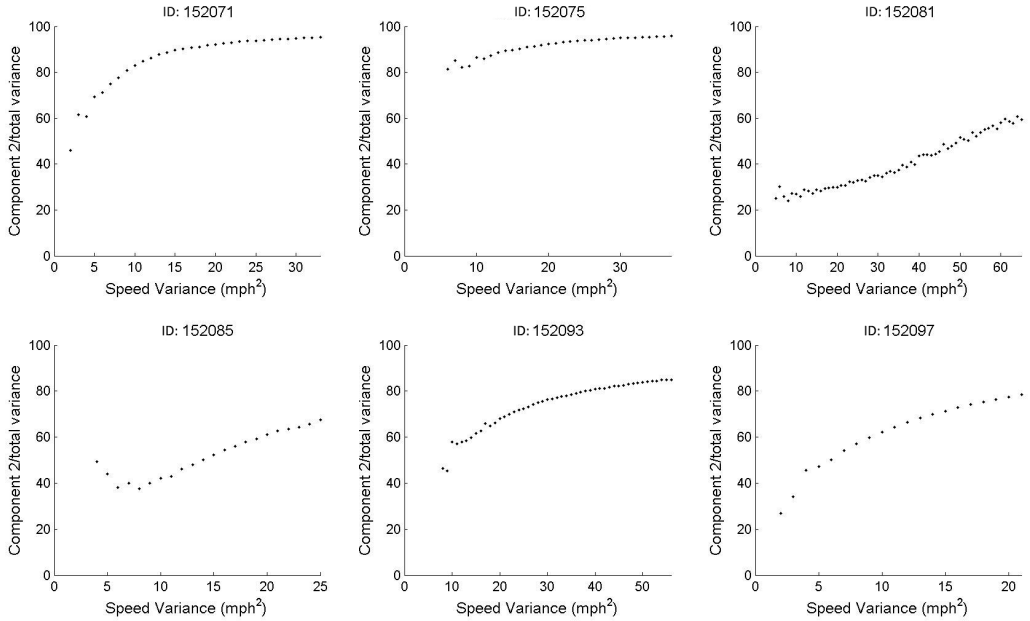

FIGURE 7 Weight of the inter-lane variance component in the total variance depending on the total variance

These results can be used to physically explain the phenomenon observed in the previous part (5.1) displaying for the majority of the sensing devices an increase of the total speed variance associated with an increase of the flow. Indeed an increase of the flow (note that we are still in free flow situation) could lead to a strong lane heterogeneity of the cars and thus to an increase of the speed variance due to an increase of the speed difference between the lanes. On the contrary, at low flow values, the distribution is fairly homogeneous between lanes because the drivers do not feel the need to stay into a specific lane and the overall variance is low.

Nevertheless it is important to note that for 30% of the sensing devices the relation between flow and speed variance is negative. In the following part we study how specific parameters impact the sign of the relation between flow and speed variance.

5.3 Analysis of factors influencing the correlation

Our goal is to determine if some features of the sensing device and its location could explain why some sensing devices show an increasing relation between flow and speed variance whereas other sensing devices exhibit a decreasing relation. Our analysis is based on two techniques: the Chi-test for independence is used to consider categorical parameters and a logistic regression model is employed to consider more type of parameters.

5.3.1 Chi-test of independence and Cramer's V

Cramer's V, denoted ϕ_c, represents the intercorrelation of two discrete variables (Cramér (*21*)). It varies from 0 to 1 and can reach 1 only when the two variables are equal. Our objective is to understand the relation between the fact that a sensing device displays a strong positive correlation (i.e. $r \geq 0.75$) between the speed variance and the flow, and some other characteristics of the sensing device and its location -such as direction, speed limit, number of lanes or type of sensing device (two different types of devices, based on two different measurement process, are used). The formula for the ϕ_c coefficient is:

$$\phi_c = \sqrt{\frac{\chi^2}{N(k-1)}} \tag{8}$$

where:

- χ^2 is derived from Pearson's chi-squared test,
- N is the total number of observations,
- k is the number of rows or columns, whichever is less.

Below are the results of the Cramer's coefficients and Chi-squared test of independence between the listed possible explanatory variables and the property of the sensing device of exhibiting an increasing or decreasing relation between flow and speed variance:

Explanatory variable	χ^2	ϕ_c	p-value
Direction	3.013	0.188	0.390 (*)
Speed limit	1.672	0.140	0.643 (*)
Sensing device type	48.749	0.757	<0.001
Number of lanes	5.496	0.254	0.358 (*)

TABLE 1 Cramer's coefficients

Standards for interpreting Cramer's V were proposed in (22):

$0.10 \leq \phi_c \leq 0.30$	Small effect
$0.30 \leq \phi_c \leq 0.50$	Medium effect
$\phi_c \geq 0.50$	Strong effect

TABLE 2 Standards for interpreting Cramer's V as proposed by (22)

Using the above standard and the p-values of the chi-squared test for independence we notice that in our case only one variable seems to have an important impact on the sign of the relation between speed variance and flow. This variable is the sensing device type ($p-value \leq 10^{-4}$ and $\phi_c = 0.757$). The fact that a sensing device is of type 1 is a strong predictor of a negative relation between individual speed variance and flow.

5.3.2 Logistic regression model

The statistical analysis performed in 5.3.1 enables us to test independence between categorical variables. To complete our analysis and to take into account continuous and ordinal predictors, we perform a binary logistic regression (Hosmer and Lemeshow (*23*)). The output of this model is for a sensing device the sign of the relation between speed variance and flow, and the explanatory variables are the speed limit, the type of sensing device, the number of lanes and the distance from the closest on-ramp/off-ramp. Based on the maximum likelihood procedure, this model returns for one category of the output ('positive relation

between speed variance and flow') an intercept and regression coefficients for each predictor. For the second output category ('negative relation between flow and speed variance') the coefficients are taken to be zero. These coefficients can then be used to determine for a given set of explanatory variables observations, the probabilities of being in a positive or negative case. We present in Table 3 the results of this analysis conducted using all the sensing devices.

Variable	Coefficient	t-value	p-value
Intercept	4.702	0.533	0.594
Speed limit	0.020	0.142	0.887
Type	-1.163	-3.623	<0.001
Number of lanes	1.530	2.037	0.042
Distance from an exit	-0.833	-0.969	0.333

TABLE 3 Coefficients of the logistic regression model

These results confirm the significant impact of the sensing device type on the sign of the correlation between speed variance and flow. It also reveals the impact of the number of lanes on this outcome. Indeed the coefficient associated with the number of lanes has a low p-value ($p-value \leq 0.042$). Because this coefficient is positive, it implies that the higher the number of lanes is, the higher the probability for the sensing device of displaying a positive relation, is.

This result supports the interpretation proposed in 5.1: a higher number of lanes could lead to the categorization of the lanes into heterogeneous flow regimes. If we are in the case of a low number of lanes, a driver has to follow the flow and has limited freedom in the choice of his lane and of his speed: this increase of the flow leads to a decrease of the speed variance. Conversely, with a high number of lanes, a driver is able to choose a lane to keep a given speed. The difference between the lanes is more likely to increase compared to the case where we have a low number of lanes: an increase of the flow should lead to an increase of the average speed between the lane and thus drives the increase of the total speed variance.

6 Conclusion

In this article, we designed a novel algorithm for traffic flow estimation from fixed radars speed measurements. We showed that a fundamental relation exists between traffic flow and speed variance, which provides significant flow estimation improvements compared to standard regression techniques based on mean speed.

We showed that for most of the 116 sensing devices considered, the main contribution to the positive correlation between speed variance and flow is due to a difference of speed between lanes.

The analysis presented in this article illustrates that speed measurements from fixed radars data provide fundamentally new insights into the field of macroscopic traffic monitoring, in particular regarding novel properties of the relation between flow and speed on multi-lane highways.

Extensions to this work include systematic calibration of the parameters used for data processing in this work (such as the time interval used to compute the speed variance over time), and a more thorough analysis of the stationarity property of this higher-order fundamental diagram. Also more work should be done to assess the relevance of the described method when using probe data from mobile devices. In particular the impact of speed population variance on the flow estimation algorithm could be explored.

Acknowledgements

We want to thank the California Center for Innovative Transportation (CCIT) team for their help with this project, in particular Scott Myers for helping with data extraction, but also Joe Butler, Ali Mortazavi, Saneesh Apte and Jonathan Felder. Finally, we want to thank the California DOT for their ongoing support to UC Berkeley and their interest in data fusion and hybrid data. Particular thanks go to John Wolf, Joan Sollenberger and Nicholas Compin for their guidance.

References

[1] Cassidy, M. and B. Coifman, Relation among average speed, flow, and density and analogous relation between density and occupancy. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1591, 1997, pp. 1–6.

[2] Knoop, V., S. Hoogendoorn, and H. Zuylen, Empirical differences between time mean speed and space mean speed. *Traffic and Granular FlowÕ07*, 2009, pp. 351–356.

[3] Newell, G., A simplified theory of kinematic waves in highway traffic, part I: General theory. *Transportation Research Part B: Methodological*, Vol. Volume 27, No. 4, 1993, pp. 281–287.

[4] Daganzo, C., The cell-transmission model, part II: network traffic. *Transportation 21 Research Part B: Methodological*, Vol. 29, No. 2, 1995, pp. 79–83.

[5] Lighthill, M. and G. Whitham, On Kinematic Waves. II. A Theory of Traffic Flow on Long Crowded Roads. *Proceedings of the Royal Society A Mathematical Physical and Engineering Sciences*, Vol. 229, No. 1178, 1955, pp. 317–345.

[6] Mobile Millennium, *Mobile Millennium*. http://traffic.berkeley.edu, 2012.

[7] Blandin, S., A. Salam, and A. Bayen, Individual speed variance in traffic flow: analysis of Bay Area radar measurements. *91th Annual Meeting of the Transportation Research Board*, 2011.

[8] Greenshields, B. D., A study of traffic capacity. *Highway Research Board Proceedings*, Vol. 14, 1935, pp. 448–477.

[9] Alvarez-Icaza, L., L. Munoz, X. Sun, and R. Horowitz, Adaptive observer for traffic density estimation. Boston, MA, 2004, pp. 2705–2710.

[10] Atrimy, M., Local density estimation and dynamic transmission-range assignment in vehicular ad hoc networks. *IEEE Trans. Intell Transp. Syst.*, Vol. 8, No. 3, 2007, pp. 400–412.

[11] Coifman, B., Estimating density and lane inflow on a freeway segment. *Transp. Res. Part A*, Vol. 37, No. 8, 2003, pp. 689–701.

[12] Gazis, D. and C. Knapp, On-line estimation of traffic densities from timeseries of flow and speed data. *Transp. Sci.*, Vol. 5, No. 3, 1971, pp. 283–301.

[13] Gazis, D. and M. Szeto, Design of density-measuring systems for roadways. *Transp. Res. Rec.*, Vol. 495, 1974, pp. 44–52.

[14] Jerbi, M., S. Senouci, T. Rasheed, and Y. Ghamri-Doudane, An infrastructure-free traffic information system for vehicular networks. Baltimore, MD, 2007, pp. 2086–2090.

[15] Panichpapiboon, S. and W. Pattara-atikom, Exploiting Wireless Communication in Vehicle Density Estimation. *IEEE Transaction on vehicular technology*, Vol. 60, No. 6, 2011, pp. 2742–2751.

[16] Jørgensen, O. N. R., Estimation of speed–flow and flow–density relations on the motorway network in the greater Copenhagen region. *IET Intelligent Transport Systems*, 2008.

[17] *Traffic assignment manual*. US Dept. of Commerce, Urban Planning Division, Washington, DC, 1964.

[18] Wang, H., D. Ni, Q.-Y. Chen, and J. Li, Stochastic Modeling of the Equilibrium Speed-Density Relationship. *J. Adv. Transp.*, 2011.

[19] Sherman, S., Non-mean-square error criteria. *IRE Transactions on Information Theory*, Vol. 4, No. 3, 1958, pp. 125–126.

[20] Wasserman, L., *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics, Springer, 2004.

[21] Cramér, H., *Mathematical methods of statistics*. Princeton mathematical series, Princeton University Press, 1946.

[22] Cohen, J., *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates, 2nd ed., 1988.

[23] Hosmer, D. and S. Lemeshow, *Applied Logistic Regression*. Wiley Series in Probability and Statistics: Texts and References Section, Wiley, 2000.