

# A Clustering Based Methodology for Determining the Optimal Roadway Configuration of Detectors for Travel Time Estimation

Bekir Bartin and Kaan Ozbay

**Abstract**— This paper deals with the problem of finding the optimal roadway segment configuration for road-based surveillance technologies to estimate route travel times accurately. This problem is inherently a space discretization problem regardless of which travel time estimation function is used. The ad-hoc solution to this problem is the equidistant segment configuration, such as every half-mile, every one-mile. It is shown in this paper that the space discretization problem can be expressed as the common clustering problem. The novelty of the proposed approach is the use of preliminary vehicle trajectory data to obtain statistically significant traffic regime at the study route. Clustering of sample space-time trajectory data is proposed as a viable methodology for solving the optimal roadway segment configuration problem.

## I. INTRODUCTION

ACCURATE route travel time estimation is essential for two different aspects of traffic management and planning. First, travel time estimates are often used in travel time prediction algorithms as a benchmark value, as discussed in [1][2][3]. Second, these estimates are invaluable for determining off-line performance measures for various policy applications. For example, travel time variability is an emerging performance measure increasingly used by decision makers and transportation planners in many transportation investment decisions. Accurately estimated travel times are especially useful for quantifying such performance measures [4][5][6].

Traffic surveillance is currently conducted by using two distinct approaches. These are (a) *road-based surveillance technologies*, such as in-road detectors, road-side detectors, and (b) *vehicle-based surveillance technologies*, such as probe vehicles, automatic vehicle location (AVL) systems, automatic vehicle identification (AVI) systems.

Both surveillance technologies have several advantages and disadvantages as described in detail by [7]. Vehicle-based technologies are useful for measuring travel times accurately; however, they are not widely used because of high implementation costs. The use of road-based technologies, on the other hand, is still prevalent. Their implementation costs are also high, mostly owing to the need for multiple units within the area of interest.

This paper is motivated by the importance of determining

how many traffic road-based surveillance units (RBSU) should be deployed and where to be deployed along a given route to obtain travel time estimations with least errors. The ad-hoc approach to this problem has been the equidistant segment configuration (e.g. every quarter-mile, half-mile or one-mile intervals).

*RBSU deployment problem:* Given a route with known traffic characteristics find the optimal number and location of roadway segments for RBSU deployment that will minimize the estimation error of a selected travel time function.

*Proposition:* There may be various other reasons for adopting the ad-hoc deployment for many possible applications. Within the travel time estimation context, it is argued in this paper that given the traffic and geometric conditions there exists an optimal segment configuration of a highway route that gives better travel time estimates than those of the ad-hoc deployment.

*Novelty of Proposed Methodology:* The use of preliminary route data for determining the number and locations of roadway segments for RBSU deployment. With the use of a simplistic travel time estimation function, the space discretization problem can be solved using known clustering techniques.

## II. RELATED WORK

The focus of the road-based surveillance technology literature has been to improve (a) the accuracy of detector readings [8][9], (b) the accuracy of travel time estimation between two traffic detectors [10][11][12][13] (c) the accuracy of route travel time prediction algorithms that rely on travel time estimations between predefined segments along a route [1][2][3].

Relatively little research has been conducted to investigate the RBSU deployment problem. Thomas [14] investigated the relationship between several selected travel detector locations and the link travel characteristics on a 3-mile arterial road in Arizona using CORSIM simulation software. This study reported network specific correlations between the four loop detector outputs and the link travel characteristics.

Oh et al. [15] used a hypothetical CORSIM simulation network to investigate the optimal loop detector locations based on various numbers of roadway lanes, link lengths, speed limits, green signal times, and traffic volumes. They searched for the optimal detector locations based on the proximity of the speed values read by loop detectors and the average link speed. They concluded that the optimal detector location is related to link length and green time.

B. Bartin, Research Associate, Civil & Environmental Engineering Department, Rutgers University. 623 Bowser Rd. Piscataway, NJ 08854 (phone: 732-445-3162; fax: 732-445-0577; e-mail: bbartin@rci.rutgers.edu).

K. Ozbay, Associate Professor, Civil & Environmental Engineering Department, Rutgers University.

Ozbay et al. [16] investigated the effect of sensor location on the travel time estimation during recurrent and non-recurrent congestion on I-76 in southern New Jersey highway network. Ozbay et al. [16] reported that increasing the number of RBSU does not always improve the accuracy of travel-time estimation during non-recurrent congestion. They also reported that the accuracy of travel time estimation during an incident vary considerably with (a) Incident Characteristics: These include location, duration, or severity degree; (b) Locations of surveillance units: Depending on the vehicle-speed profile along the route, a different selection of sensor, detector locations yield different results; and (c) Travel-Time Function: Different travel-time estimation functions yield different results.

Optimal segment configuration problem is certainly not restricted to road-based traffic surveillance technologies. Since the cost of deploying these technologies is almost prohibitive for large scale networks, the similar problem exists in the deployment of vehicle-based technologies, where one often needs to decide, for example, how many AVI stations to deploy and where to deploy. Sherali et al. [17] studied the problem of determining the optimum number of AVI readers to obtain maximum travel time variability. Yang and Miller-Hooks [18] approached the problem on a network level without specifying any category of surveillance technology.

There is no commonly accepted methodology for the RBSU deployment problem. The simulation approach as employed in Oh et al. [15], Ozbay et al. [16] and Thomas [14] can only supply network specific results. Moreover, their results depend highly on the validity of the developed simulation model. The methodologies presented in Sherali et al. [17] and Yang and Miller-Hooks [18], on the other hand, are specific to determining travel time variability, where travel times can be determined by VBSU accurately.

### III. FORMULATION OF RBSU LOCATION PROBLEM

It is often desired to obtain space-time trajectories of vehicles using discrete approximation in cases where this information is not fully available. The idea is to represent space-time trajectory in steps of a small interval,  $\Delta$ . Clearly, once a function is discretized, a certain loss of information is introduced. Furthermore, the higher the magnitude of  $\Delta$ , the more the loss of information is. The lack of continuous space-time trajectories restricts us to observe the function values only at discrete points.

Suppose that the study route of length  $L$  is divided into  $n$  number of infinitesimal intervals of size  $\Delta$ , where  $\Delta = L/n$ . Each location  $j = 1..n$  along the route is a candidate location for RBSU deployment. Each RBSU shall be associated with one **disjoint** roadway segment  $s$ , defined by segment boundary locations  $b_{s-1}$  and  $b_s$ , where  $s = 1..k$  and  $1 \leq b_s \leq n$ , and  $k$  is the number of available RBSU for deployment

Fig 1 demonstrates the definitions of the variables on a sample roadway scheme. The figure shows two RBSU with their corresponding two disjoint roadway segments. The

first segment  $s = 1$  and the second segment  $s = 2$  are defined by the boundaries  $[b_0, b_1]$  and  $[b_1, b_2]$ , respectively. The lengths of segment one and two can be calculated by  $(b_1 - b_0)\Delta$  and  $(b_2 - b_1)\Delta$ , respectively.

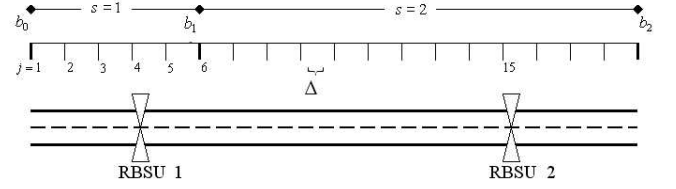


Fig 1. A sample roadway scheme with corresponding variables

Suppose that space-time trajectory information of  $m$  randomly selected vehicles is available at each location  $j$  along the route. Let  $y_{ij}$  denote the inverse instantaneous speed of vehicle  $i$  at location  $j$ . Then  $t_i$ , the actual travel time of vehicle  $i$ , can be approximated as follows:

$$t_i \cong \sum_{j=1}^n y_{ij} \Delta \quad (1)$$

Suppose that there exists only one RBSU available for deployment. There are  $n$  possible deployment locations along the route. Let  $\hat{t}_{ij}$  denote the estimated travel time of vehicle  $i$ .  $\hat{t}_{ij}$  is based on the information collected at location  $j$ , and can be calculated by a simplistic travel time estimation function, such as  $\hat{t}_{ij} = y_{ij} \cdot L$ . Let  $\varepsilon_{ij}$  denote the squared travel time estimation error of vehicle  $i$  at location  $j$ . Then,  $\varepsilon_{ij}$  can be written as,

$$\varepsilon_{ij} = (t_i - \hat{t}_{ij})^2 = \left( \sum_{j=1}^n y_{ij} \Delta - y_{ij} n \Delta \right)^2 \quad (2)$$

Note that  $L = n \Delta$ . Let us define the variable  $\bar{y}_i$ , which is the inverse of the harmonic mean of observed speeds of vehicle  $i$  along the route, given by  $\bar{y}_i = \sum_{j=1}^n y_{ij} / n$ . Then (2)

can be rewritten as in the following form.

$$\varepsilon_{ij} = L^2 (\bar{y}_i - y_{ij})^2 \quad (3)$$

It is assumed that travel time estimation errors  $\varepsilon_{ij}$  of  $m$  randomly selected vehicles observed at location  $j$  are independent. Thus, the average of  $\varepsilon_{ij}$ , denoted by  $\bar{\varepsilon}_j$  is an unbiased estimate of the mean error at location  $j$ .  $\bar{\varepsilon}_j$  is given by,

$$\bar{\varepsilon}_j = \sum_{i=1}^m L^2 (\bar{y}_i - y_{ij})^2 / m \quad (4)$$

Let  $p_j$  represents the probability of location  $j$  being selected out of  $n$  possible locations for deployment. In other words, each location  $j$  is assigned with a probability value

that is drawn from a discrete probability function selected for the route. Then, the average travel time estimation error,  $\bar{\varepsilon}$ , along the route can be presented as follows.

$$\bar{\varepsilon} = L^2 \sum_{j=1}^n \sum_{i=1}^m p_j (\bar{y}_i - y_{ij})^2 / m \quad (5)$$

Let us now suppose that there are  $k$  number of RBSU available. Because each RBSU shall be associated with a disjoint roadway segment, there are  $k$  segments that need to be monitored along the route. Let  $p_{sj}$  denote the probability of location  $j$  being selected for deployment within segment  $s$ , where  $s = 1 \dots k$ . Therefore, (5) can be rewritten as follows.

$$\bar{\varepsilon} = \sum_{s=1}^k (b_s - b_{s-1})^2 \Delta^2 \sum_{j=b_{s-1}}^{b_s} \sum_{i=1}^m p_{sj} (\bar{y}_i - y_{ij})^2 / m \quad (6)$$

Suppose that  $p_{sj}$  is drawn from a uniform probability distribution function. Namely, each location  $j$  within a given segment  $s$  has equal chances of being selected for deployment. Then,  $p_{sj}$  can be expressed by the term  $1/(b_s - b_{s-1})$ . Consequently, in (6) the term

$$\sum_{j=b_{s-1}}^{b_s} \sum_{i=1}^m p_{sj} (\bar{y}_i - y_{ij})^2 / m$$

can be changed to  $\sum_{j=b_{s-1}}^{b_s} \sum_{i=1}^m (\bar{y}_i - y_{ij})^2 / [m(b_s - b_{s-1})]$ .

Rearranging the summation signs in this term, (6) becomes,

$$\bar{\varepsilon} = \frac{\Delta^2}{m} \sum_{s=1}^k \sum_{i=1}^m (b_s - b_{s-1}) \sum_{j=b_{s-1}}^{b_s} (\bar{y}_i - y_{ij})^2 \quad (7)$$

Note that the term

$$(b_s - b_{s-1}) \sum_{j=b_{s-1}}^{b_s} (\bar{y}_i - y_{ij})^2$$

is the Euclidean sum of squared distances of  $y_{ij}$  of vehicle  $i$  along the segment  $[b_{s-1}, b_s]$ <sup>1</sup>. Since the variables  $\Delta$  and  $m$  are constants, (7) represents the total Euclidean sum of squared distances for  $m$  vehicles.

It can be observed in (7) that  $\bar{\varepsilon}$  can be minimized by dividing the route into  $k$  segments within which individual vehicles' inverse speeds dispersed closely in space with respect to each other. Namely, the solution can be found where the within-segments Euclidean sum of squares of  $y_{ij}$  is minimized. The objective function in (7) is similar to that of common clustering algorithm, which is based on minimum within-group distance criterion.

*Remark:* It is clear that minimization of  $\bar{\varepsilon}$  in (7) is not defined when the vehicles' speeds are zero. Such cases are

<sup>1</sup> Sum of squared errors (SSE) of  $n$  independent samples is  $SSE = (n-1)s^2$ , where  $s^2$  is the variance of the  $n$  samples. Euclidean sum of squared distances between  $n$  points is equal to  $n \cdot SSE$ .

often expected during heavy congestion. Simple adjustments to data points can be done for avoiding such problems. It is obvious that these coarse shifts in data points would result in loss of data. Therefore, they should be sufficient enough to make (7) valid, and small enough prevent high loss of information.

#### A. Discussion of Modeling Assumptions

The two major assumptions are used in the formulation of the RBSU deployment problem. These assumptions are discussed as follows.

1)  $\hat{t}_{ij}$  is based on the information collected at location  $j$ , and can be calculated by a simplistic travel time estimation function,  $\hat{t}_{ij} = y_{ij} \cdot L$ . This estimation function assumes a uniform speed trajectory for each vehicle over space.

Such simplistic speed trajectory functions have been extensively used in the literature [9][13][3]. As mentioned previously, finding the optimal segment configuration that minimizes travel time estimation error is a space discretization problem. Using this assumption the problem can be converted to the common clustering problem. In the lack of the first assumption, other available mathematical techniques need to be sought for solving this problem.

2)  $p_{sj}$  denote the probability of location  $j$  being selected for deployment within segment  $s$ , where  $s = 1 \dots k$ . It is assumed that  $p_{sj}$  is drawn from a uniform probability distribution function. In practical applications, RBSU are usually deployed approximately in the middle of predefined segments. See [11], for example. In this case, the selection of Gaussian probability distribution along each segment is well suited for  $p_{sj}$ . The problem can be further performed based on this distribution. In this case, clustering can still be performed where each data point  $y_{ij}$  has varying weights. This task is left as a future work.

#### IV. NUMERICAL ANALYSIS

This section is intended to demonstrate the potential benefits of identifying roadway segments using the proposed approach. These benefits are quantified based on the minimization of travel time estimation errors as shown in (7).

Due to the lack of space-time trajectory ground truth data, the proposed approach is tested using simulated traffic data obtained from a hypothetical freeway route modeled in PARAMICS micro-simulation software. It should be noted that the proposed methodology does not rely on simulation results. The simulation model here is merely used to obtain travel times and space-time trajectories of vehicles under given traffic conditions.

#### A. Study Network

The study route is shown in . It consists of nine origin-destination (OD) zones and three interchanges. The route of interest is between OD pair (1,2). It is approximately 12.5 miles (20 km). The distances between each monument as

labeled in Fig 2 are shown in Table 1. The network is loaded with hypothetical OD demand matrices that lead to congested traffic along the route.

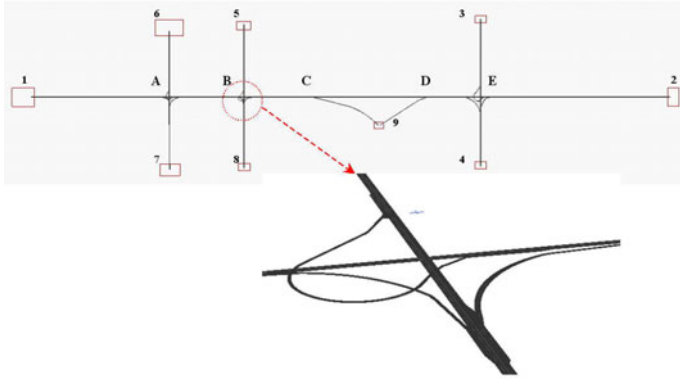


Fig 2. Study route modeled in PARAMICS

Only the traffic flow in the westbound direction of the study route is considered. The OD pairs with high traffic volumes are (1,4), (1,9), (4,2) and (9,2). There are two distinctive congestion patterns on this study route. (a) **Delay before exit E**: This is due to the high demand destined to zone 4 and the high through traffic demand destined to zone 2. (b) **Delay before exit C**: This is because of vehicles (mostly vehicles between (1,9) weaving to change lanes to take the exit to zone 9. Considering the high through traffic volume at this location i.e. (1,4), (1,2), (8,2) and (8,4), long queues are observed before exit C.

TABLE I  
DISTANCES BETWEEN MONUMENTS

Location	Distance
I-A	2.52 miles
A-B	1.49
B-C	1.45
C-D	2.22
D-E	0.78
E-2	3.97

### B. RBSU Location Problem

Suppose that there are 5 RBSU available for traffic monitoring along this given route. The question is then how to deploy these available units so that travel time estimation error is minimized. The analyses in this section include not only the entire route i.e. OD pair (1,2), but also the intermediate OD pairs as well, i.e. OD pairs (1,9), (1,4), (9,2) and (4,2) as shown in Fig 2.

The following questions are addressed in this section. (a) What is the sufficient number of probe vehicle data for statistically significant estimation of segment boundaries? (b) What is the optimum number of segments for RBSU deployment?

#### 1) Configuration of Segments

The main route between OD pair (1,2) is divided to  $n=1000$  discrete locations, where  $\Delta=20$  meters. Vehicles' speed data are collected at intervals of  $\Delta=20$  meters using the PARAMICS application programming interface (API).

A random set of vehicle trajectory data collected from various simulation runs is selected to better represent most of the possible traffic states. Different random numbers in simulation runs result in different number of vehicles released from the zones, different vehicle release times, vehicle characteristics, etc. Different demand profiles change the demand within user-defined intervals and result in fluctuation in demand over time.

The study route is simulated with different random seed values and demand profiles until the average network travel time is within 99% confidence interval with 1% relative error. The confidence interval is obtained as  $[11.9 - 12.1 \text{ minutes}]^2$ .

From the collected vehicle data of multiple simulation runs, 50 independent vehicle trajectories traveling between OD pair (1,2) are selected. This data set is used for the clustering analysis. The global  $k$ -means clustering analysis proposed by [19] is performed based on the minimization of the objective function given in (7).

Table 2 summarizes the clustering results.

Note that some segment boundaries are different for the

TABLE II  
SEGMENT BOUNDARIES OF SELECTED OD PAIRS  
BASED ON CLUSTERING APPROACH

OD Pair	Segment Boundaries
(1,2)	[1,340], (340,429], (429,599], (599,679] and (679,1000]
(1,9)	[1,340] and (340,440]
(1,4)	[1,340], (340,429], (429,599] and (599,680]
(9,2)	[622,679], (679,1000]

intermediate OD pairs because they are adjusted based on the location of exit points C and D in Fig 2.

The solution considers the roadway connecting the OD pair (4,2) as one segment although it is approximately 4-miles. Further clustering of the dataset indicates that this roadway should be considered as a complete segment even after the tenth clustering solution. This result can be attributed to the fact that the clustering analysis based on the minimization of the objective function in (7) focuses on the segments where there are higher variances in traffic characteristics.

The expected boundaries based on equidistant approach are shown in Table 3.

These values are determined simply by equally

TABLE III  
SEGMENT BOUNDARIES OF SELECTED OD PAIRS BASED ON EQUIDISTANT  
APPROACH

OD Pair	Segment Boundaries
(1,2)	[1,200], (200,400], (400,600], (600,800] and (800,1000]
(1,9)	[1,220] and (220,440]
(1,4)	[1,170], (170,340], (340,510], and (510,680]
(9,2)	[622,811] and (811,1000]

partitioning the selected OD pairs. The purpose of analysis is to show the differences in travel time estimation error

<sup>2</sup> Average network travel time is the average of travel times of all vehicles simulated in the network.

with the same number of segments between each OD pair, but with different segment configuration approaches. For example, clustering analysis results in two segments between the OD pair (1,9). Therefore, for the equidistant approach the route between this OD pair is divided into 2 equally distant segments.

### 2) Comparison of Travel Time Estimation Error

Travel time estimation errors of the segment configurations shown in Table 2 and Table 3, are compared using a total of approximately 7500 vehicle records collected in various simulation runs.

Table 4 shows the percentage reduction in the travel time estimation error owing to the proposed approach as compared with the equidistant deployment approach.

Table 4 indicates that there is substantial reduction in travel time estimation error using the segments defined by the clustering approach.

TABLE IV  
REDUCTION IN TRAVEL TIME ESTIMATION ERROR USING THE PROPOSED APPROACH

Simulation Run	Percent Reduction in Total Travel Time Estimation Error			
OD pair	(1,2)	(1,9)	(1,4)	(9,2)
1	46.2 %	38.6 %	48.8 %	14.7 %
2	56.2	57.9	51.4	17.8
3	56.1	60.7	48.1	13.2
4	45.0	55.6	58.3	27.9
5	59.6	58.6	48.2	23.7
6	53.7	43.6	45.9	16.3
7	56.7	47.5	60.5	16.9
8	49.9	41.3	51.8	21.0
9	56.0	67.9	49.5	20.6
10	58.8	50.1	60.5	12.2

The average of the absolute differences between the actual travel times and the estimated travel times are shown in Table 5. The values given in Table 5 are simply the square root of  $\bar{\epsilon}$  in (7) between the selected OD pairs. It can be seen that the clustering approach to RBSU deployment problem results in better travel time estimations than the ad-hoc approach.

### 3) Determining the Vehicle Trajectory Sample Size

One should test the sensitivity of the clustering analysis results with respect to sample vehicle trajectory data. The

TABLE V  
COMPARISON OF ABSOLUTE DIFFERENCES BETWEEN THE ACTUAL AND ESTIMATED TRAVEL TIMES

	OD pair			
	(1,2)	(1,9)	(1,4)	(9,2)
Equidistant Approach	11.5 min	7.8	9.9	13.9
Clustering Approach	8.2	5.5	6.9	12.5

question is how much the segment boundaries would change if different sets of vehicle data were chosen. Table 6 shows the results of the clustering analysis based on 10 different sets of 50-vehicle trajectory data between the OD pair (1,2).

Table 6 shows that the results of segment boundaries with respect to different sample sets do not considerably vary much. It is assumed the boundary points obtained based on different datasets follow a Gaussian distribution.

Then, the results show that 99% of all possible 4-boundary points will be within less than 2-3% of the real

boundary points. If Student- $t$  distribution were used, then the relative errors would vary only by 0.026.

The assumption of a probability distribution of boundary points for different datasets is an essential leeway in estimating the sufficient number of space-time trajectories. Utilizing basic statistical methods, a sequential approach can be applied to define a statistically sufficient number of sample set based on a postulated relative error ( $\beta$ ) and a confidence level ( $\alpha$ ).

One could easily determine the sufficient number of datasets using the data given in Table 6. For instance, following the results of the first two datasets, i.e. 100 probe vehicles in total, one could deduce a 97.5% confidence

TABLE VI  
SEGMENT BOUNDARIES OF CLUSTERING ANALYSIS FOR VARIOUS DATASETS

Dataset	$b_1$	$b_2$	$b_3$	$b_4$
1	338	431	600	680
2	342	430	576	679
3	339	429	564	678
4	343	429	597	679
5	342	432	580	679
6	343	429	601	680
7	320	451	616	679
8	336	431	598	680
9	353	430	592	679
10	346	430	601	680
Average	340.2	432.2	592.5	679.3
97.5 %	[333.3-	[426.8-	[580.2-	[678.8-
C.I.	347.1]	437.6]	604.8]	679.8]
Relative Error	0.0204	0.0126	0.0207	0.0008

interval of the boundaries with a 5% relative error as [334.8,345.2], [429.2,431.8], [557.1,678.2], [678.2,680.8]. The average boundaries based on the first two datasets are [1,340], (340,430], (430,588], (588,680], (680,1000]. In other words, based on the first two data sets, it is expected that with  $p=0.975$  probability, the actual boundaries would fall within the confidence intervals. In fact, it can be observed that the averages of 10 datasets given in Table 6 are very close to the average of first two datasets.

### 4) Determining the Optimum Number of Segments

It is often required to determine the optimum number of segments, namely the number of RBSU that will yield the minimum travel time estimation error. The clustering approach is also well suited for this purpose. Clustering procedure can be terminated when the cumulative gain from clustering becomes minimal. Fig 3 demonstrates the cumulative gain versus the number of segments between OD pairs (1,2), (1,4), (1,9) and (9,2).

A flattening at the OD pair (1,2) curve begins at the 4-cluster solution (88.9%), and the line is essentially flat after the seven-cluster solution (97.8%). Therefore, the graph implies 4 but at most 7 segments within the dataset for the OD pair (1,2).

It can also be noticed that the flattening of the cumulative gain curves for the OD pairs (1,4), (1,9) and (9,2) are faster than the flattening of the curve for the OD pair (1,2). For the

OD pair (1,4) the curve becomes almost flat at the 4-cluster solution. For the OD pairs (1,9) and (9,2), it starts at the 2-cluster solution.

For the OD pair (1,2), if the ad-hoc solution were chosen where RBSU were deployed for example at every half-a-mile, the solution would yield approximately 25 surveillance units. It can be observed in Fig 3 that the relative gain between the seven-cluster solution and the 15-cluster solution is only 1.5%, if the surveillance units were deployed based on clustering results. One could easily argue that the marginal return from the extra units is not scalable. For the study route, ad-hoc approach results in a costly solution to the RBSU deployment problem.

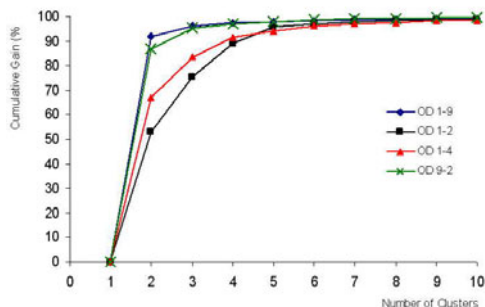


Fig 3. Cumulative Gain in Clustering

## V. CONCLUSIONS

It is implied in this paper that the direction in investigating the travel time estimation problem should not be only towards the development of complex estimation algorithms and vehicle trajectory functions, but should be towards finding the segments where travel time estimation errors are minimized for a given estimation function. It is shown in this paper that with the use of a simplistic travel time function, the RBSU deployment problem can be expressed as a clustering problem. The novelty of the proposed approach is the use of preliminary traffic data to observe the possible traffic regimes of the study route. Clustering analysis of this dataset gives the optimal RBSU deployment for the route of interest.

The proposed clustering approach not only finds the segments for the RBSU deployment, but also determines the optimum number of segments using the percent-gain of clustering approach. The ability of estimating the sufficient number of segments to monitor for better travel time estimation is of great importance in freeway traffic management.

An example problem of the RBSU deployment problem is investigated. Clustering of vehicle trajectory data is used to determine the number of segments to monitor. Necessary vehicle trajectory dataset is obtained from hypothetical highway network developed PARAMICS simulation software. The travel time estimation errors of the clustering approach and the equidistant approach are also compared.

It is also shown that the number of vehicle trajectory data can be approximated assuming a Gaussian distribution of the cluster boundaries of each dataset. Non-parametric statistics

can also be applied to determine the sufficient sample data size for this application See [20] for details.

## REFERENCES

- [1] Kwon, J., Coifman, B., and Bickel, P.J., Day to day travel time trends and travel time prediction from loop detector data, Presented at the 78<sup>th</sup> Annual Meeting of the Transportation Research Board, 1999.
- [2] Oda, T., An algorithm for prediction of travel time using vehicle sensor data, In *Third International Conference on Road Traffic Control*, pp.40-44, Institute of Electrical Engineers, 1990.
- [3] Rice, J.A. and van Zwet, E. (2004) A simple and effective method for predicting travel times on freeways, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 5, No. 3, pp. 200-207.
- [4] Chen C., Skabardonis, A. and Varaiya, P. (2003) "Travel time variability as a measure of service." Presented at the 82<sup>nd</sup> Annual Meeting of the Transportation Research Board.
- [5] Lomax, T. Schrank, D., Turner S., Margiotta, R. "Selecting Travel Reliability Measures", NCHRP Draft Report, May 2003.
- [6] Van Lint, J.W.C., Tu, H., and Van Zuylen, H.J. (2004) "Travel time reliability on freeways," Proceedings of 10th world conference on transport research (WTRC), Istanbul, Turkey.
- [7] Turner, S.M. Advanced techniques for travel time data collection. Presented at the 75<sup>th</sup> Annual Meeting of the Transportation Research Board, 2000.
- [8] Coifman, B. (1999), Using dual loop speed traps to identify detector errors, Presented at the 78<sup>th</sup> Annual Meeting of the Transportation Research Board.
- [9] Dailey, D.J. travel time estimates using a series of loop volume and occupancy measurements. Presented at the 76<sup>th</sup> Annual Meeting of the Transportation Research Board, 1997.
- [10] Coifman, B. (2002), Estimating travel times and vehicle trajectories on freeways using dual loops detectors, *Transportation Research Record Part A*, Vol.36, pp.351-364.
- [11] Cortes, C, Lavanya, R., Oh, J. and Jayakrishnan, R. (2002). General-purpose methodology for estimating link travel time with multiple point detection of traffic, *Transportation Research Record 1802*, pp. 181-189.
- [12] Dailey, D.J. (1993), Travel time estimation using cross-correlation techniques, *Transportation Research Part B*, Vol. 27B, No.2, pp. 97-107.
- [13] Gazis, D.C. and Knapp, C.H. (1971). Online estimation of traffic densities from time-series of flow and speed data, *Transportation Science*, Vol. 5, No. 3, pp. 283-301.
- [14] Thomas, G. (1999). "The relationship between detector location and travel characteristics on arterial streets." *Institute of Transportation Engineers Journal*. Vol. 69. Issue 10. pp. 36-42.
- [15] Oh, S., Ran, B. and Choi, K. (2003). "Optimal detector location for estimating link travel speed in urban arterial roads." Presented at the 82<sup>nd</sup> Annual Meeting of the Transportation Research Board.
- [16] Ozbay, K., Bartin, B. and Chien, S. "South Jersey Real-Time Motorist Information Systems: Technology and Practice". *Transportation Research Record 1886*, 2004.
- [17] Sherali, H.D., Desai, J., Rakha, H., and El-Shawarby, I. A discrete optimization approach for locating automatic vehicle identification readers for the provision of roadway travel times, Presented at the 82<sup>nd</sup> Annual Meeting of the Transportation Research Board, 2002.
- [18] Yang, B. and Miller-Hooks, E. Determining critical arcs for collecting real-time travel information, Presented at the 81<sup>st</sup> Annual Meeting of the Transportation Research Board, 2001.
- [19] Likas, A., Vlassis, N. and Verbeek, J. J. (2003). "The global k-means clustering algorithm." *Pattern Recognition*. Vol. 36. pp. 451-461.
- [20] Noether, G. E. (1991) *Introduction to Statistics: The nonparametric way*. Springer-Verlag.