Proving and Implementing the Effectiveness of iDnc for Optimal 4D Data Signature-based Cluster Models of NLEX Traffic Density Data Sets

Reynaldo G. Maravilla, Jr., Elise Raina A. Tabanda, Rona May U. Tadlas
Jasmine A. Malinao, Henry N. Adorna

Department of Computer Science, Algorithms and Complexity Laboratory,
University of the Philippines, Diliman, Quezon City 1101 Philippines
Email: {rgmaravilla, eatabanda, rutadlas}@up.edu.ph,
{jamalinao, hnadorna} @dcs.upd.edu.ph

ABSTRACT

This paper used the improved cluster validity index, known as iD_{nc}, of an existing cluster validity index D_{nc}. It is implemented to determine which data signature construction generates the optimal cluster models for periodic time series density data set. Furthermore, we proved the effectiveness of iD_{nc} to show that the said cluster validity index is indeed improved from D_{nc}. Among the candidate data signatures identified in previous literature, the 4D data signature was found to be the most effective in generating the optimal cluster models for traffic density data set based on iD_{nc} computations. We showed the robustness of the 4D data signature construction by implementing these computations and the Kruskal-Wallis Test on multiple segments of the North Luzon Expressway (NLEX). With the resulting optimal cluster model for density, traffic analysis using the data signature visualization technique data image was done on 4 different NLEX segments.

Key words: Data Signatures, Traffic Density Analysis, North Luzon Expressway, Data Image, Cluster Validity Index

1. INTRODUCTION

A recent study dealt with density analysis for studying traffic behavior[1]. If we are to consider traffic congestion, density is an accurate indicator. Density considers the occupied space of the road and the speed of the vehicles and it can give a better estimate of the real behavior of the traffic flow through time. However, the optimal data signature-based cluster model for traffic behavior studies is only defined in volume[2]. With the recent traffic behavioral studies focusing on density analysis in NLEX, its optimal cluster model must also be established.

In order to generate optimal cluster models, we will use the improved cluster validity index iD_{nc} introduced in [2]. This is the improved index of the Dunn-like index D_{nc} in literature. In this paper, we will present the formal proof of the effectiveness of iD_{nc} over D_{nc} in generating optimal cluster models.

Traffic data set can be efficiently and precisely analyzed using Fourier-based data signatures. Papers on volume identified the candidate optimal data signatures for volume in both clustering[2] and visualization[3] by virtue of the data set characteristics, i.e. periodic. The same holds for density, therefore, the data signature-based method will be used for density data sets. Then we will select the data signature that will produce the optimal cluster model for density among the candidates. These data signatures will be clustered and visualized using data image. Analysis will be done on the generated data images of the NLEX segments.

The data set recorded and provided by National Center for Transportation Studies (NCTS) is in time mean speed. The data set, therefore, will be preprocessed to produce and represent realistic characterizations of traffic flow in NLEX.

Section 1.1 discusses the definitions, concepts, and notations used in this paper. Section 2 discusses the formal proof of the effectiveness of iD_{nc} as a cluster validity index. Section 3 shows the steps conducted to build the density models of the NLEX segments. It also includes the steps in representing the density models as data signatures which are to be clustered and visualized using data image. The resulting iD_{nc} computation and data signature-based visualization models are explained in Section 4. Finally, the conclusions and recommendations for this study are discussed in Section 5.

1.1 Definitions

1.1.1 The Data Sets

The data sets provided by NCTS in this study on the NLEX Balintawak (Blk), Bocaue (Boc), Meycauayan (Mcy), and Marilao (Mrl) segments in the year 2006 are periodic. These data sets contain hourly time mean speed and mean volume of each of the four segments. The hourly data set is sufficient for analysis as it has been validated to be accurate in a previous study[1].

The data sets are preprocessed in a previous study in which average time mean speeds must meet the minimum speed requirement of 40 kph. Eleven weeks, five weeks, twenty-six weeks, and seven weeks are eliminated for the Blk, Boc, Mcy, and Mrl data sets, leaving us with 41 weeks, 47 weeks, 26 weeks, and 45 weeks, respectively.

From these data sets, we produce traffic density data sets that are also per segment, with each density data set covering all four lanes of each segment.

1.1.2 Major Traffic Variables

1. **Volume q.** Volume is the hourly mean of the number of vehicles per lane.

2. **Time Mean Speed u_t.** Time mean speed is the mean of the speeds u_i of the n vehicles passing through a specific point within a given interval of time.

$$u_t = \frac{\sum_{i=1}^{n} u_i}{n}$$

3. **Space Mean Speed u_s.** Space mean speed is the speed based on the average travel time of n vehicles in the stream within a given section of road.

$$u_s = \frac{n}{\sum_{i=1}^{n} \frac{1}{u_i}}$$

4. **Density k.** Density k is the number of vehicles over a certain length of a road.

$$k = \frac{q}{u_s}$$

Space mean speed is used in estimating the density because it considers the space between the vehicles.

1.1.3 Estimation of Space Mean Speed from Time Mean Speed

Since the data set provided contains only the time mean speed and space mean speed is required in determining density, we estimate the space mean speed from the time mean speed using Rakha-Wang equation[4] to get u_s, where $\overline{u}_s \approx \overline{u}_t - \frac{\sigma_t^2}{\overline{u}_t}$ There will be a 0 to 1 percent margin of error in the estimation.

1.1.4 Data Signature

A data signature, as defined in [5] is a mathematical data vector that captures the essence of a large data set in a small fraction of its original size. These signatures allow us to conduct analysis in a higher level of abstraction and yet still reflect the intended results as if we are using the original data.

Various Power Spectrum-based data signatures[3, 6] had been employed to generate cluster and visualization models to represent periodic time series data. Fourier descriptors such as Power Spectrums rely on the fact that any signal can be decomposed into a series of frequency components via Fourier Transforms. By treating each nD weekly partitions in the NLEX BLK-NB time-series traffic volume data set[6] as discrete signals, we can obtain their Power Spectrums through the Discrete Fourier Transform(DFT) decomposition.

Power Spectrum is the distribution of power values as a function of frequency. For every frequency component, power can be measured by summing the squares of the coefficients a_k and b_k of the corresponding sine-cosine pair of the decomposition and then getting its square root, where the variable $k = 0, 1, \ldots, n-1$. The Power Spectrum A_k of the signal is given by, $A_k = \sqrt{{a_k}^2 + {b_k}^2}$. Studies on NLEX traffic volume have shown that the set $\{A_0, A_7, A_{14}, A_{21}\}$ is an optimal data signature for both visualization[3] and clustering [2]. Methods in [2] validate the optimality of the 4D data signature by showing an improved Dunn-like index. The 4D data signature used for clustering achieved statistical competence among all other data signatures. The study achieved $\approx 97.6\%$ original data reduction for production of an optimal cluster model for Dunn-like variables.

1.1.5 Data Visualization via Data Image

In this study, we make use of data images to visualize the data set. A data image is a graphical representation that transforms the given multidimensional data set into a color range image. Observations are made through the colors' given characteristics and respective magnitudes. In our given data set, weeks are represented by the y-axis arranged by their cluster membership and days by the x-axis (with 1 as Sunday, 2 as Monday, and so on). The weeks are arranged according to their clusters. Clusters are determined by using the X-Means Clustering algorithm[7] that takes the 4D data signatures of the weeks in the data set as its input.

1.1.6 Cluster Validity Indices

Cluster validity indices are used to measure cluster compactness and separation for cluster models (or *schemes*) obtained from either fuzzy or crisp clustering algorithms without the presence of user-defined criterion[2].

1. **Dunn-like index D_{nc}**
A Minimum Spanning Tree(MST)-based Dunn-like index is defined as follows:

   Let a cluster c_i and the complete graph G_i whose vertices correspond to the vectors of c_i. The weight w_e of an edge e of this graph equals the distance between its two end points $\mathbf{q}, \mathbf{r}$. Let E_i^{MST} be the set of edges of the MST of the graph G_i and e_i^{MST} the edge in E_i^{MST} with the maximum weight. Then the diameter of G_i, denoted as $diam_i^{MST}$, is defined as the weight of e_i^{MST}. Then the Dunn-like index is given by the equation,

$$D_{nc} = \min_{i=1,\ldots,nc} \left\{ \min_{j=i+1,\ldots,nc} \left(\frac{d(c_i, c_j)}{\max_{k=1,\ldots,nc} diam_k^{MST}} \right) \right\},$$

   where $d(c_i, c_j) = \min_{\mathbf{q} \in c_j, \mathbf{r} \in c_j} d(\mathbf{q}, \mathbf{r})$, is the dissimilarity function between two clusters c_i and c_j. For brevity, we denote $d(c_i, c_j)$ with z.

2. **Improved Dunn-like index iD_{nc}**
In the computation of Dunn-like index D_{nc}, cluster separation is only measured by using $\mathbf{x} \in c_i$ and $\mathbf{y} \in c_j$ that derives $d(c_i, c_j)$. By considering that $\mathbf{x}$ and $\mathbf{y}$ may be potential outliers[8] of their clusters, we strengthen this measure by incorporating the information of the relationships of both of these points to their co-members, and more particularly, to their respective centroids $\overline{x}_i$ and $\overline{x}_j$, through A and B, where A is the

distance from $\mathbf{x}$ to its cluster's centroid and B is the distance from $\mathbf{y}$ to its cluster's centroid. Additionally, the inter-cluster information attributed to Y as the distance of two centroids for c_i and c_j shall also be considered in building the improved index iD_{nc}, as follows:

$$iD_{nc} = \min_{i=1,\ldots,nc} \left\{ \min_{j=i+1,\ldots,nc} \left(\frac{d(c_i, c_j) + Y + 1}{A + B + 1} \right) \right\}.$$

Illustrated in Figure 1 are the variables used to measure D_{nc} and iD_{nc}.

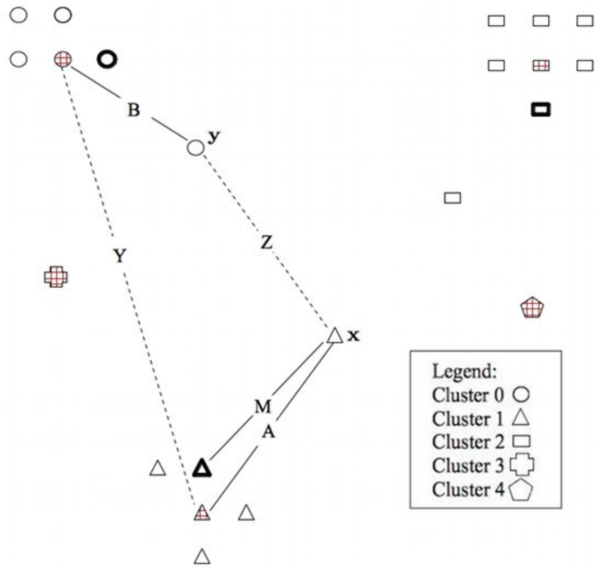

Figure 1: Sample Scheme with 5 Clusters, with Cluster Centroids with Horizontal and Vertical Hatchings

2. PROVING THE EFFECTIVENESS OF ID_{NC} AS CLUSTER VALIDITY INDEX

In Figure 1, the points with darkened edges, along with their clusters' potential outliers give $diam_k^{MST}$ for any cluster c_k in the scheme. In this scheme, Cluster 1 gives the value $M = \max_{k=1,\ldots,nc} diam_k^{MST}$.

Suppose we build another model S_2 basing from the existing model (referred hereon as S_1) in Figure 1. In this new model, we modify the inter-cluster information of clusters 0 and 1. We can either add or subtract $\epsilon_1 > 0$ and $\epsilon_2 > 0$ from z and Y, respectively. By adding ϵ_1 to z in S_2, S_2 would give a larger D_{nc} than S_1 regardless of the fact that cluster separation may have deteriorated by subtracting ϵ_2 to Y. On the hand, when subtracting ϵ_1 from z in S_2, S_1 yields a better D_{nc} than S_2 even when Y had been added with ϵ_2. It is therefore essential that we consider the general effect of all types of modifications on cluster separation and compactness from S_1 to S_2. We provide generalizations in succeeding discussions and prove that our proposed improved index, iD_{nc}, gives better characterization of the inherent relationships in the scheme.

In order to ensure that the suggested improvements and the revisions incorporated to the original cluster validity index D_{nc} are observable in iD_{nc}, the following concerns are investigated, as follows,

- the consistency of D_{nc} and iD_{nc} in determining a better cluster model between the schemes S_1 and S_2
- the capability of iD_{nc} to effectively measure cluster validity in certain configurations of S_1 and S_2 that D_{nc} failed to do so

In the discussion that follows, cluster validity index computation shall be focused for two clusters. We fix the global parameter that holds the maximum diameter of all the minimum spanning trees built for each cluster in a scheme. We shall focus on three particular information: centroid-to-centroid, centroid-to-potential outlier, and single-linkage distances to describe inter-cluster and intra-cluster relationships in the scheme. However, note that we maintain the original global minimum cluster validity index iD_{nc} to describe the entire scheme containing at least two clusters.

Let $\bullet, \overline{\bullet} \in \{>, <\}$ and $\alpha, \overline{\alpha} = \{0, 1\}$.

Let A and B be the intra-cluster distances measured from centroids $\overline{x}_i$ and $\overline{x}_j$ of the clusters c_i and c_j, respectively, where $i \neq j, i, j \in \{1, 2, \ldots, nc\}$, to their farthest co-members $\mathbf{a} \in c_i$ and $\mathbf{b} \in c_j$ for both schemes S_1 and S_2.

Let $\mathbf{x} \in c_i, \mathbf{y} \in c_j$, $||\overline{\mathbf{xy}}|| = d(c_i, c_j)$, Y be the distance from $\overline{x}_i$ to $\overline{x}_j$, and $M = \max_{k=1,2,\ldots,nc} diam_k^{MST}$ of S_1.

Let $Y + (-1)^{\alpha}(\epsilon_2)$ be the distance from $\overline{x}_i$ to $\overline{x}_j$ in scheme S_2, where $\epsilon_2 > 0$, $\alpha \in \{0, 1\}$. Furthermore, let $||\overline{\mathbf{xy}}|| + (-1)^{\alpha}(\epsilon_1)$ $= d(c_i, c_j)$ in S_2, where $\epsilon_1 > 0$, $\alpha \in \{0, 1\}$.

Thus,

$$D_{nc}{}^{(S_1,0)} = \frac{||\overline{\mathbf{xy}}||}{M} \text{and } D_{nc}{}^{(S_2,\alpha)} = \frac{||\overline{\mathbf{xy}}|| + (-1)^{\alpha}(\epsilon_1)}{M},$$

$$iD_{nc}{}^{(S_1,0)} = \frac{||\overline{\mathbf{xy}}|| + Y + 1}{A + B + 1} \; \textit{and}$$

$$iD_{nc}{}^{(S_2,(\alpha,\overline{\alpha}))} = \frac{(||\overline{\mathbf{xy}}|| + (-1)^{\alpha}(\epsilon_1)) + (Y + (-1)^{\overline{\alpha}}(\epsilon_2)) + 1}{A + B + 1}.$$

Theorem 1. *Given any two schemes S_1 and S_2, the following cases hold,*

1. *for $\alpha = \overline{\alpha}$, $iD_{nc}^{(S_1,0)} \bullet iD_{nc}^{(S_2,(\alpha,\overline{\alpha}))}$ iff $D_{nc}^{(S_1,0)} \bullet D_{nc}^{(S_2,\alpha)}$*
2. *for $\alpha \neq \overline{\alpha}$,*
   (a) *if $D_{nc}^{(S_1,0)} \bullet D_{nc}^{(S_2,\alpha)}$ and $(-1)^{\overline{\alpha}}\epsilon_2 + (-1)^{\alpha}\epsilon_1 \bullet 0$ then $iD_{nc}^{(S_1,0)} \overline{\bullet} iD_{nc}^{(S_2,(\alpha,\overline{\alpha}))}$, where $\bullet \neq \overline{\bullet}$.*
   (b) *if $D_{nc}^{(S_1,0)} \bullet D_{nc}^{(S_2,\alpha)}$ and $(-1)^{\overline{\alpha}}\epsilon_2 + (-1)^{\alpha}\epsilon_1 \overline{\bullet} 0$ then $iD_{nc}^{(S_1,0)} \bullet iD_{nc}^{(S_2,(\alpha,\overline{\alpha}))}$, where $\bullet \neq \overline{\bullet}$.*
   (c) *$iD_{nc}^{(S_1,0)} = iD_{nc}^{(S_2,(\alpha,\overline{\alpha})}$ iff $|(-1)^{\alpha}(\epsilon_2)| = |(-1)^{\overline{\alpha}}(\epsilon_1)|$.*

PROOF. Let $z = ||\overline{\mathbf{xy}}||$, $z' = z + (-1)^{\alpha}\epsilon_1$, $Y' = Y + (-1)^{\overline{\alpha}}$.

For $\alpha = \overline{\alpha}$, if $z' - z < 0$, we know that $Y' - Y < 0$ or if $z' - z > 0$, then $Y' - Y > 0$. Thus $(z + (-1)^{\alpha}(\epsilon_1))$ and $(Y + (-1)^{\overline{\alpha}}(\epsilon_2))$ are either both larger or smaller than z and Y, respectively, thus when $\frac{z}{M} \bullet \frac{z+(-1)^{\alpha}(\epsilon_1)}{M}$, we have

$$\frac{z+Y+1}{A+B+1} \bullet \frac{z+(-1)^{\alpha}\epsilon_1 + Y + (-1)^{\overline{\alpha}} + 1}{A+B+1}, \bullet \in \{>,<\}.$$

Conversely, when $(Y + (-1)^{\overline{\alpha}}(\epsilon_2)) > Y$, we know that $(z + (-1)^{\alpha}(\epsilon_1)) > z$, or when $(Y + (-1)^{\overline{\alpha}}(\epsilon_2)) < Y$, we have $(z + (-1)^{\alpha}(\epsilon_1)) < z$. Thus, when $\frac{z+Y+1}{A+B+1} \bullet \frac{z+(-1)^{\alpha}(\epsilon_1)+Y+(-1)^{\overline{\alpha}}+1}{A+B+1}$, we obtain

$$\frac{z}{M} \bullet \frac{z+(-1)^{\alpha}(\epsilon_1)}{M}, \bullet \in \{>,<\}.$$

We formulate the following proofs to the indicated items in item 2.

(a) For $\alpha \neq \overline{\alpha}$, if $z' - z < 0$, we know that $Y' - Y > 0$ since $\overline{\alpha} = 0$. Additionally, with $\alpha = 1$, we obtain $\frac{z}{M} > \frac{z+(-1)^{\alpha}(\epsilon_1)}{M}$. Given that $(-1)^{\alpha}(\epsilon_1) + (-1)^{\overline{\alpha}}(\epsilon_2) > 0$, it is therefore true that $\epsilon_2 > \epsilon_1$. Let $\varepsilon = (-1)^{\alpha}(\epsilon_1) + (-1)^{\overline{\alpha}}(\epsilon_2)$. Since $\varepsilon > 0$, it implies that

$$\frac{z+Y+1}{A+B+1} < \frac{z+Y+1+\varepsilon}{A+B+1}.$$

Additionally, $\alpha \neq \overline{\alpha}$, if $z' - z > 0$, then $Y' - Y < 0$. With $\alpha = 0$, we know that $\frac{z}{M} < \frac{z+(-1)^{\alpha}(\epsilon_1)}{M}$. Given that $(-1)^{\alpha}(\epsilon_1) + (-1)^{\overline{\alpha}}(\epsilon_2) < 0$, then $\epsilon_1 > \epsilon_2$. Using the same computation of ε, where $\varepsilon < 0$, it is therefore true that

$$\frac{z+Y+1}{A+B+1} > \frac{z+Y+1+\varepsilon}{A+B+1}.$$

(b) For $\alpha \neq \overline{\alpha}$, if $z' - z > 0$, we know that $Y' - Y < 0$ since $\alpha = 0$. Additionally, with $\alpha = 0$, we obtain $\frac{z}{M} < \frac{z+(-1)^{\alpha}(\epsilon_1)}{M}$. Given that $(-1)^{\alpha}(\epsilon_1) + (-1)^{\overline{\alpha}}(\epsilon_2) > 0$, it is therefore true that $\epsilon_1 > \epsilon_2$. Let $\varepsilon = (-1)^{\alpha}(\epsilon_1) + (-1)^{\overline{\alpha}}(\epsilon_2)$. Since $\varepsilon > 0$, it implies that

$$\frac{z+Y+1}{A+B+1} < \frac{z+Y+1+\varepsilon}{A+B+1}.$$

Additionally, $\alpha \neq \overline{\alpha}$, if $z' - z < 0$, then $Y' - Y > 0$ with $\alpha = 1$. With $\alpha = 1$, we know that $\frac{z}{M} > \frac{z+(-1)^{\alpha}(\epsilon_1)}{M}$. Given that $(-1)^{\alpha}(\epsilon_1) + (-1)^{\overline{\alpha}}(\epsilon_2) < 0$, then $\epsilon_1 > \epsilon_2$. Using the same computation of ε, where $\varepsilon < 0$, it is therefore true that

$$\frac{z+Y+1}{A+B+1} > \frac{z+Y+1+\varepsilon}{A+B+1}.$$

(c) Given $|(-1)^{\alpha}(\epsilon_2)| = |(-1)^{\overline{\alpha}}(\epsilon_1)|$, where $\alpha \neq \overline{\alpha}$, we know that $\epsilon_1 = \epsilon_2$ and $(-1)^{\alpha}(\epsilon_2)$ is just an additive inverse of $(-1)^{\overline{\alpha}}(\epsilon_1)$. Let $\varepsilon = (-1)^{\alpha}(\epsilon_1) + (-1)^{\overline{\alpha}}(\epsilon_2)$. Thus, $\varepsilon = 0$, and

$$\frac{z+Y+1}{A+B+1} = \frac{z+Y+1+\varepsilon}{A+B+1}.$$

Conversely, $\frac{z+Y+1}{A+B+1} > \frac{z+Y+1+\varepsilon}{A+B+1}$ holds when $\varepsilon = (-1)^{\alpha}(\epsilon_1) + (-1)^{\overline{\alpha}}(\epsilon_2) = 0$, where $\alpha \neq \overline{\alpha}$, thus $|(-1)^{\alpha}(\epsilon_2)| = |(-1)^{\overline{\alpha}}(\epsilon_1)|$.

However, both of the aforementioned relations also show how the original index D_{nc} measure cluster validity differently with the improved index iD_{nc}, i.e. $D_{nc}^{(S_1,0)} = \frac{z}{M}$ and $D_{nc}^{(S_2,\alpha)} = \frac{z+(-1)^{\alpha}(\epsilon_1)}{M}$, $\alpha \in \{0,1\}$, therefore

$$D_{nc}^{(S_1,0)} \neq D_{nc}^{(S_2,\alpha)}.$$

□

3. METHODOLOGY

3.1 Building Effective Density Models from Sparse Data Points

1. From the preprocessed data set, we extract the 4 segments' mean volume and time mean speed per hour.
2. We estimate the space mean speed from the time mean speed by first getting the variances among the time mean speeds of the four lanes of each segment. We apply the Rakha-Wang equation to get the space mean speed per hour of the segments. To maintain consistency, the computed space mean speeds undergo preprocessing to eliminate values that are below 40 kph.
3. We estimate each segment's density k values using their corresponding given mean volume and space mean speed per hour.

3.2 Data Signature-based Clustering and Visualization of the Density Models

1. From the generated density data sets from the previous section, we generate the the data signature of each segment's weeks.
2. Clustering is then done using the X-means clustering algorithm [7]. The data sets' candidate optimal index data signatures have the Power Spectrum values of $\{A_0, A_7, A_{14}, A_{21}\}$, $\{A_0, A_7, A_{14}, A_{21}, A_{28}, A_{35}, A_{42}\}$, $\{A_0, A_1, ..., A_{\frac{n}{2}}\}$, and $\{A_0, A_1, ..., A_n\}$ which are in 4D, 7D, $\frac{n}{2}$D, and nD, respectively, as discussed in [2].
3. The iD_{nc} values of the data sets are computed on different dimensions. Comparison of the iD_{nc} values of the different dimensions is then done to determine the optimal cluster model for each density data set. The non-parametric method Kruskal-Wallis Test is used to identify which of the data signatures help generate the optimal cluster model of density on the 4 segments.
4. We visualize the traffic density values of the time domain data set using data images where rows represent the values of each week, structured contiguously based on the clustering result determined by iD_{nc}, and each pixel is colored based on the actual values of the density in a time slot. Analysis is done on these data images to pinpoint which among the segments more frequently exhibit free flow, traffic disruptions, traffic incidents, and road congestions.

4. RESULTS AND DISCUSSIONS

4.1 Graphs of the Segments' Density Data Sets

From the preprocessed data, we computed the variances of the hourly time mean speed per segment. High values of variances are evident from the graphs because of lane congestions during certain hours. With the computed variances, hourly space mean speeds of the segments were produced. Densities of the segments were then computed.

The calculated hourly densities of the Blk, Boc, Mcy, and Mrl segments are shown in Figures 2, 3, 4, and 5, respectively. As seen from the values of the graphs, the density values of the Blk segment is relatively higher than the density values of the 3 other segments. The graphs' behavior are similar with the exception of Mcy segment. Mcy segment's graph is more varied than the other graphs because it has only half of its original weeks (26 out of 52), whereas the others only lost at most 11 weeks.

All graphs show relatively higher values on the 2000^{th}-3000^{th} hour (1000^{th}-1500^{th} hour in Mcy), 6000^{th}-7000^{th} hour (5000^{th}-6000^{th} in Blk, 3500^{th}-4000^{th} in Mcy), and final hours of the graphs. These hours represent the Holy Week in mid-April, All Saint's Day/semestral break at the end of October until the start of November, and the Christmas holidays, respectively.

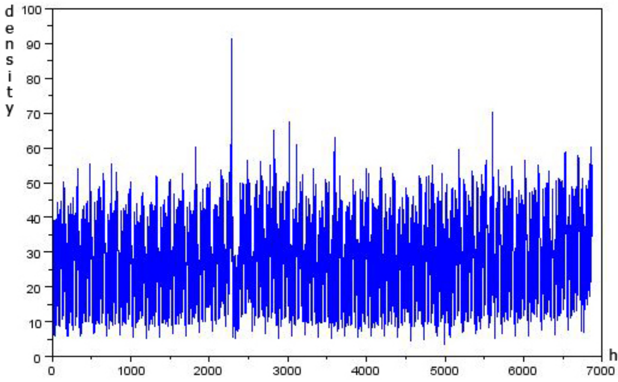

Figure 2: Hourly Densities of the Blk Segment

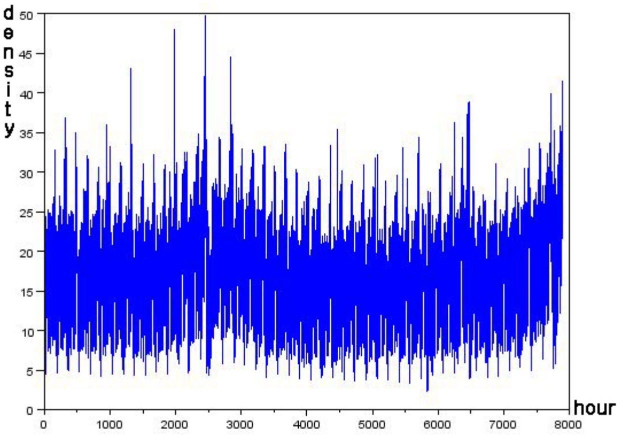

Figure 3: Hourly Densities of the Boc Segment

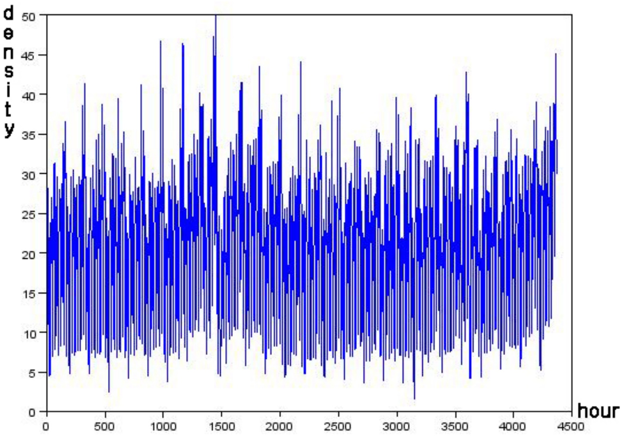

Figure 4: Hourly Densities of the Mcy Segment

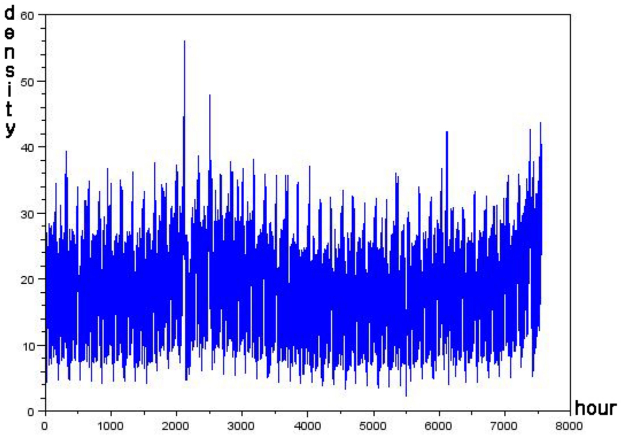

Figure 5: Hourly Densities of the Mrl Segment

4.2 Data Signature-Based Cluster and Visualization Models

The following improved Dunn-like index iD_{nc} values were obtained for the different segments' data sets as shown in Table 1.

Representation	Blk	Boc	Mcy	Mrl
4D	**1.238**	1.088	**1.226**	0.808
7D	1.180	**1.150**	1.196	**0.881**
$\frac{n}{2}D$	0.585	0.703	0.737	0.669
nD	0.682	0.647	0.739	0.554

Table 1: iD_{nc} of Cluster Models using Data Signatures and entire Power Spectrums of the Density Data Sets

The non-parametric method Kruskal-Wallis test is then done to determine the data signature with the optimal dimension for the density data sets of Blk, Boc, Mcy, and Mrl.

Shown in Tables 2 and 3 are the values showing the Mean Ranks and Test Statistics of the data signatures used in obtaining the cluster validity index iD_{nc} for the segments Blk, Boc, Mcy, and Mrl.

Representation	N	Mean Rank
4D	4	**12.75**
7D	4	12.25
$\frac{n}{2}D$	4	4.75
nD	4	4.25

Table 2: Ranks of Used Data Dignatures for the Density Data Sets using iD_{nc}.

	iD_{nc} value
Chi-square	11.34
df	3
Asymp. Sig.	0.01

Table 3: Test Statistics^{a,b} of Used Data Signatures for the Density Data Sets using iD_{nc}. a. Kruskal Wallis Test. b. Grouping variable:Representation

As we can see from the mean ranks in Table 2, 4D data signature has the highest rank among the given density data sets on the said segments. Also, it is suggested that the computed index iD_{nc} using the different dimensioned data signatures of the density data sets for the said segments are significantly different because the P-value (Asymp. Sig. value) of the Kruskal-Wallis Test on the data signature is ¡ 0.05 (at 0.01). This means that the quality of the cluster models derived from using data signature with higher dimensions (i.e. 7D, $\frac{n}{2}$D, nD) is statistically inequivalent to that of the 4D data signatures to describe the data set.

4.3 Notes on Traffic Behavior for the NLEX Segments

As seen from the data images of the segments' densities in Figures 6, 7, 8, and 9, the following weeks have relatively high densities in the 4 segments: week 15 (day 4) and week 44 (days 3 and 4). The sudden increase in density is due to the departure for the holidays on Holy Week and All Saint's Day. The following weeks have relatively low densities in the 4 segments: week 1 (day 1), week 39 (day 5 to 6), week 15(days 5, 6, and 7), and week 44 (day 4). Instances of the sudden decrease in density occurred on some days of a holiday vacation. Majority of the people planning on a vacation have already left, leaving a few to depart on the following days (week 15's days 5, 6, and 7). Other instances of a relatively low density turnout are also attributed to travel advisories due to typhoons (week 39 and 44 - typhoons Milenyo and Paeng).

The Blk segment is found to exhibit higher densities more frequently. Among the 4 segments, it is the only one that has a record of about 90 vehicles per kilometer. Mrl and Boc segments, on the other hand, are the segments that exhibit free flow more frequently. Mcy segment is too compromised (only half of the weeks are valid) so analysis might not be reliable and accurate.

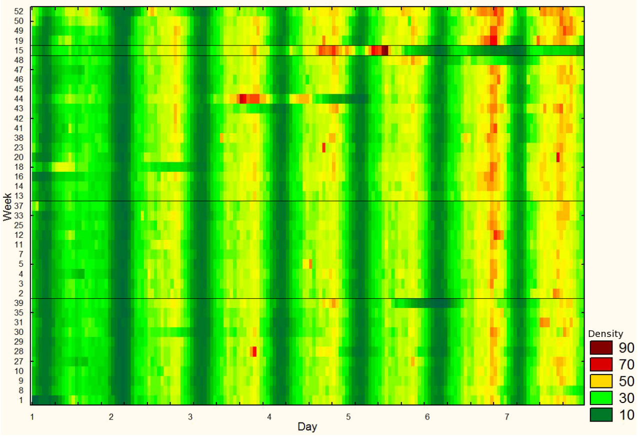

Figure 6: Data Image of the Blk Segment's Hourly Traffic Density Data

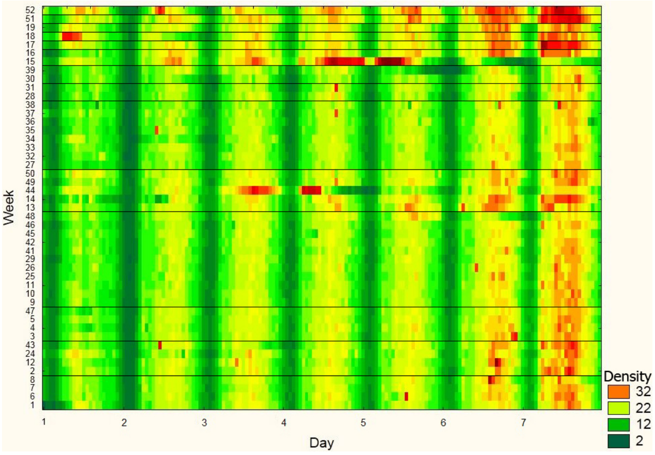

Figure 7: Data Image of the Boc Segment's Hourly Traffic Density Data

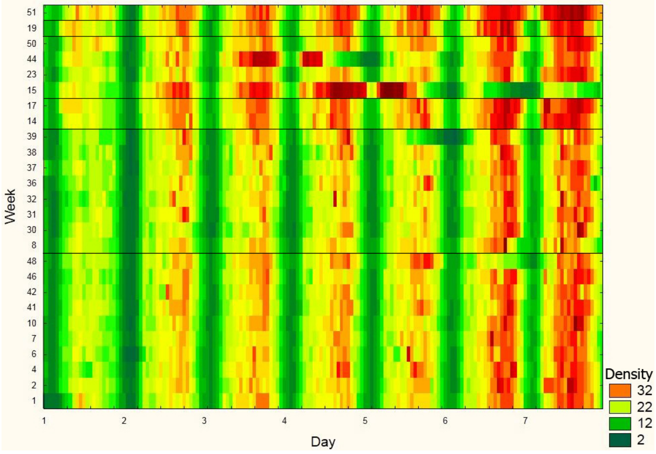

Figure 8: Data Image of the Mcy Segment's Hourly Traffic Density Data

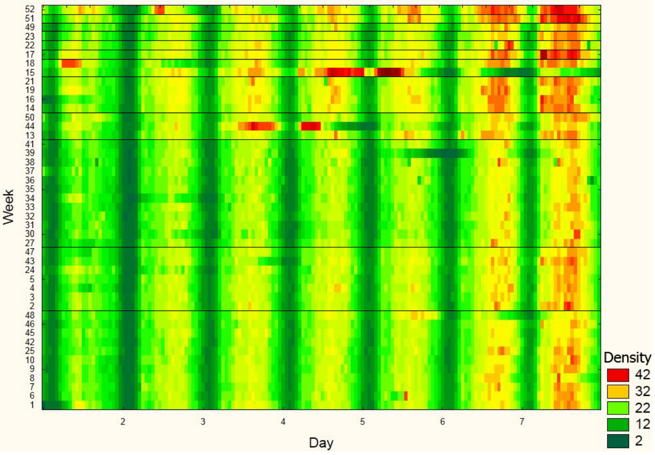

Figure 9: Data Image of the Mrl Segment's Hourly Traffic Density Data

5. CONCLUSIONS AND RECOMMENDATIONS

In this paper, we used density for traffic analysis as a congestion indicator in traffic. 4D data signature-based cluster models were obtained to characterize traffic behavior through time in NLEX. Traffic density analysis on different segments produced similar behaviors on periods where significant spikes are observed (e.g. high density in April because of Holy Week).

This paper also proved that the cluster validity index iD_{nc} is indeed an improved variant of D_{nc} in both theory and application for periodic traffic density data set. Using iD_{nc}, the produced density data signatures' optimal cluster model of the given traffic data set is found to be the 4D cluster model, the same as volume's optimal cluster model. The 4D data signature cluster model's iD_{nc} value was also the numerically highest among the studied dimensions (i.e., 4D, 7D, $\frac{n}{2}$D, nD) and it also had the highest mean rank. The higher dimensions were not statistically equivalent with the 4D cluster model. This resulted in a more efficient analysis without compromising its accuracy.

Traffic density analysis on NLEX's 4 segments using data image showed the traffic behavior of the expressway accurately because of the color coded scheme implemented in the visualization technique. Because of this, congestions can now be observed more easily and accurately.

Multi-year traffic density analysis on a segment is recommended for further validation of the 4D data signature cluster model optimality.

6. ACKNOWLEDGEMENTS

The researchers like to thank Dr. Ma. Sheilah Gaabucayan-Napalang and Dr. Jose Regin Regidor for validating our results and providing the data sets. Mr. Maravilla thanks DOST-SEI for his undergraduate scholarship.

This work is partially supported by a grant from DOST-PCIEERD through an ERDT project entitled *Information Visualization via Data Signatures.*

7. ADDITIONAL AUTHORS

8. REFERENCES

[1] Maravilla, R., Tabanda, E., Malinao, J., Adorna, H.: Data Signature-based Time Series Traffic Analysis on Coarse-grained NLEX Density Data Set. Communication and Networking 266, Part 2, Springer 2012, pp. 208-219. ISBN 978-3-642-27200-4. (2012)

[2] Malinao, J., Tadlas, R.M., Juayong, R.A., Oquendo, E.R., Adorna, H.: An Index for Optimal Data Signature-based Cluster Models of Coarse- and Fine-grained Time Series Traffic Data Sets. Proceedings of the National Conference for Information Technology Education. (2011)

[3] Malinao, J., Juayong, R.A., Oquendo, E., Tadlas, R., Lee, J., Clemente, J., Gabucayan-Napalang, Ma.S., Regidor, J.R., Adorna, J.: A Quantitative Analysis-based Algorithm for Optimal Data Signature Construction of Traffic Data Sets. In Proceedings of the 1st AICS/GNU International Conference on Computers, Networks, Systems, and Industrial Engineering (CNSI 2011). (2011)

[4] Rakha, H., Wang, Z.: Estimating Traffic Stream Space-Mean Speed and Reliability from Dual and Single Loop Detectors. (2005)

[5] Wong, P., Foote, H., Leung, R., Adams, D., Thomas, J.: Data Signatures and Visualization of Scientific Data Sets. Pacific Northwest National Laboratory, USA, IEEE. (2000)

[6] Malinao, J., Juayong, R.A., Becerral, J., Cabreros, K.R., Remaneses, K.M., Khaw, J., Wuysang, D., Corpuz, F.J., Hernandez, N.H., Yap, J.M., Adorna, A.: Patterns and Outlier Analysis of Traffic Flow using Data Signatures via BC Method and Vector Fusion Visualization, In Proc. of the 3rd International Conference on Human-centric Computing (HumanCom-10). (2010)

[7] Pelleg, D., Moore, A.: X-means: Extending K-means with efficient Estimation of the Number of Clusters. Proceedings of the 17th International Conf. on Machine Learning. (2000)

[8] Oquendo, E.R., Clemente, J., Malinao, J., Adorna, H.: Characterizing Classes of Potential Outliers through Traffic Data Set Data Signature 2D nMDS Projection, In Philippine Information Technology Journal, Vol.4, No.1. (2011)