# Data Signature-based Time Series Traffic Analysis on Coarse-grained NLEX Density Data Sets

Reynaldo G. Maravilla, Jr., Elise Raina A. Tabanda,
Jasmine A. Malinao and Henry N. Adorna

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

***Abstract***

*In this study, we characterized traffic density modeled from coarse data by using data signatures to effectively and efficiently represent traffic flow behavior. Using the 2006 North Luzon Expressway North Bound (NLEX NB) Balintawak (Blk), Bocaue (Boc), Meycauayan (Mcy), and Marilao (Mrl) segments' hourly traffic volume and time mean speed data sets provided by the National Center for Transportation Studies (NCTS), we generated hourly traffic density data set. Each point in the data was represented by a 4D data signature where cluster models and 2D visualizations were formulated and varying traffic density behaviors were identified, i.e. high and low traffic congestions, outliers, etc. Best-fit curves, confidence bands and ellipses were generated in the visualizations for additional cluster information. From a finer-grained 6-minute interval NLEX Blk-NB density data set, the coarser-grained hourly density data set of Blk was validated for consistency and correctness of results. Finally, we ascertained probable causes of the behaviors to provide insights for better traffic management in the expressway.*

***Keywords:*** *Data Signatures, Traffic Density Analysis, North Luzon Expressway, Non-Metric Multidimensional Scaling, Data Image*

This paper is a revised and expanded version of a paper entitled "Data Signature-based Time Series Traffic Analysis on Coarse-grained NLEX Density Data Set" presented at the Future Generation Communication and Networking (FGCN) international conference held in Jeju Island, Korea on December 9, 2011.

## 1. Introduction

Previous traffic behavior studies dealt only with volume analysis [1]. If we are to consider 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. Expressways, most of the time, should exhibit very low densities and high speeds, but spikes in the density graphs of 2006 NLEX Blk-NB segment data show otherwise. Domain experts identified inconsistencies and pointed out that the outliers determined are unrealistic for expressways.

The study aims to show that the proposed density model is effective in estimating the traffic behavior of NLEX, with emphasis on *u* being the space mean speed and not time mean speed. The data set recorded and provided by NCTS is in time mean speed. The data set,

therefore, will be preprocessed to produce and represent realistic characterizations of traffic flow in NLEX.

With the NLEX segments Blk's, Boc's, Mcy's, and Mrl's densities produced by the model, we will analyze the traffic flow by building a model for hourly traffic space mean speed and volume in NLEX. Data signatures will then be produced to represent the hourly traffic density data points. These data signatures and the time-domain data set cluster model will be visualized using the Non-Metric Multidimensional Scaling [2] and Data Images, respectively. Then, the intercluster and intracluster relationships of these data points will be examined. Data set outliers and potential outliers will be identified and analyzed using the methods in [1, 3]. We validate our results in the hourly density data with the 6-minute data set of Blk.

Section 1 discusses the definitions, concepts, and notations used in this paper. Section 2 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. The resulting 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. Data sets:** The data sets provided by NCTS in this study on the NLEX north bound segments in the year 2006 are periodic. One of the five data sets used in the study consists of the hourly time mean speed and mean volume of the Blk segment. Another data set, consisting of the Blk segment's 6-minute time mean speed and mean volume, is used to validate the segment's hourly data set. The other three data sets consist of the hourly time mean speed and mean volume of Boc, Mcy, and Mrl segments.

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, eleven weeks, five weeks, twenty-six weeks, and seven weeks are eliminated for the Blk (hourly), Blk (6-minute), Boc, Mcy, and Mrl data sets, leaving us with 41 weeks (168 hours), 41 weeks (1680 hours), 47 weeks, 26 weeks, and 45 weeks, respectively.

**1.1.2. Traffic flow:**

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 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.

5. **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]:

$$\bar{u}_s \approx \bar{u}_t - \frac{\sigma_t^2}{\bar{u}_t}.$$

There will be a 0 to 1 percent margin of error in the estimation.

**1.1.3. 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 *n*D weekly partitions in the NLEX BLK-NB time-series traffic volume data set[3] 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 $k = 0, 1,\ldots, n\text{-}1$. 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 [6] and clustering [7]. Methods in [7] 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.

**1.1.4. Data visualization:** In this study, we incorporate two methods, namely data images and Non-metric multidimensional scaling to analyze the data set. The first method is used to present the time-domain traffic density data. The second one projects the 4D signatures into a simpler 2D visualization for traffic analysis.

1. **Data image.** 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[8] that takes the 4D data signatures of the weeks in the data set as its input.

2. **Non-metric multidimensional scaling.** Non-metric multidimensional scaling (nMDS) is another visualization technique that maps multidimensional data set onto a 2D space. It computes the dissimilarity of the data points using Euclidean distance, Correlation, Mahalanobis, and other distance measures discussed in

the literature[2]. nMDS includes a minimization of the stress or loss function to determine an optimal projection of the points in the Euclidean space given the known relationships in the higher dimension.

3. **Confidence measures.**
   - *Best fit curve.* Linear, quadratic, cubic, quartic, and quintic curves are generated and fitted to their respective clusters. Root Mean Squared Deviation (RMSD) formula is applied to the curves to determine the best fit curve. RMSD gets the difference of the actual nMDS $y$ values observed and the predicted $y$ values ( $\hat{y}$ ) of the curve model. The one with the lowest value will then be the best fit curve.
   - *Confidence bands.* From the constructed best fit curve, the confidence band is extended above and below the curve by

$$\sqrt{c}\sqrt{\frac{SS}{DF}}t_{\alpha}(DF)$$

     where $c = G|x \times \Sigma \times G'|x$, $G|x$ is the gradient vector of the parameters at a particular value of $x$, $G'|x$ is the transposed gradient vector, $\Sigma$ is the variance-covariance matrix, $SS$ is the sum of squares for the fit, $DF$ is the degrees of freedom, and $t_{\alpha}DF$ is the student's $t$ critical value based on the confidence level $\alpha$ and the degrees of freedom $DF$.
   - *Confidence ellipse.* A confidence ellipse, as defined in [9], uses intervals for both $X$ and $Y$ values of the nMDS scatterplot. The interval is projected horizontally and vertically respectively. The confidence ellipse is formed using the equation $\overline{Z} \pm R \times I$, where $\overline{Z}$ is the mean of either $X$ or $Y$, $R$ is the range of either $X$ or $Y$, and $I$ is the confidence level $1 - \alpha$.

4. **Potential outliers.** Potential outliers, as previously defined in [3,9] are points projected "near" or at the periphery of a region occupied by its cluster in the 2D visualization.
   - *Absolute potential outliers*. An absolute potential outlier is a data point that lies outside the confidence band and ellipse of its respective cluster. This point is not represented by its cluster's best fit curve.
   - *Valid potential outliers*. A valid potential outlier is a data point that lies outside the confidence ellipse, but lies within the confidence band of its cluster and is still represented by the best fit curve.
   - *Ambiguous potential outliers*. An ambiguous potential outlier is a data point that is bounded by either two confidence bands or two confidence ellipses of different clusters or is inside a confidence ellipse, but outside of its cluster's confidence band.

## 2. Methodology

### 2.1. 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.
4. To validate the produced hourly density data of the segments, steps 1-3 for modeling hourly traffic density are also conducted for modeling 6-minute traffic density of the Blk segment.

### 2.2. Data signature-based clustering and visualization of the density models

1. Given the hourly (and 6-minute) density data sets generated from the previous section, the values of each week is transformed from its time domain to its frequency domain representation through the Discrete Fourier Transform and generate its 168D (and 1680D) Power Spectrum values for the hourly (and 6-minute interval) data sets. Then, a 4D data signature is constructed from the Power Spectrum values of each week consisting of the components $A_0$, $A_7$, $A_{14}$, and $A_{21}$.
2. Using all the data signatures of the weeks of each density data set as input to the X-Means clustering algorithm[8], we build each data set's cluster model to identify groups of weeks that may have high, regular, and low traffic density (i.e. congestion) and pinpoint outliers and potential outliers in the model. We also pinpoint the time frames where these various traffic behaviors are identified. The traffic density analysis will be presented to domain experts for their assessment. With the resulting assessment, we will provide suggestions to traffic control management for business-related decisions.
3. The nMDS 2D visualizations are produced using the data signature representations of each week in the five data sets, incorporating the clustering results of X-Means clustering algorithms by coloring the 2D projections of the data signatures with respect to the assigned color information of their cluster.
4. The confidence bands, confidence ellipse, and best fit curve at 90% confidence interval per cluster of Blk's hourly and 6-minute density data sets are generated to determine their set of potential outliers.
5. The traffic density values of the five time domain data sets are visualized as data images where rows represent the values of each week, structured contiguously based on the clustering result, and each pixel is colored based on the actual values of the density in a time slot. Darkened lines separate the clusters and outliers from one another.

## 3. Results and Discussions

### 3.1. Graphs of the segments' density data sets

From the preprocessed data, we computed the variances of the hourly time mean speed. The variances are consistent, but there are relatively high values of variance are found. It is because other lanes are congested during a specific hour, therefore, time mean speed variation is evident. From the computed variances, the hourly space mean speeds of the segment are produced. Spikes from processed space mean speeds are still observed, but they are relatively shorter than the spikes produced in the raw space mean speed graph. With the new set of space mean speeds, consistency in the density graph is expected.

The calculated hourly and 6-minute densities from the mean volume and processed space mean speeds of Blk are shown in Figure 1(a) and Figure 1(b), respectively. The graphs show consistent values of density except on spikes where traffic incidents could have happened. The graphs show similar behavior of traffic density. Consistency in this matter shows us that the produced hourly density model is precise. To further validate the hourly density model, we then perform data visualization techniques in comparing the two data sets.

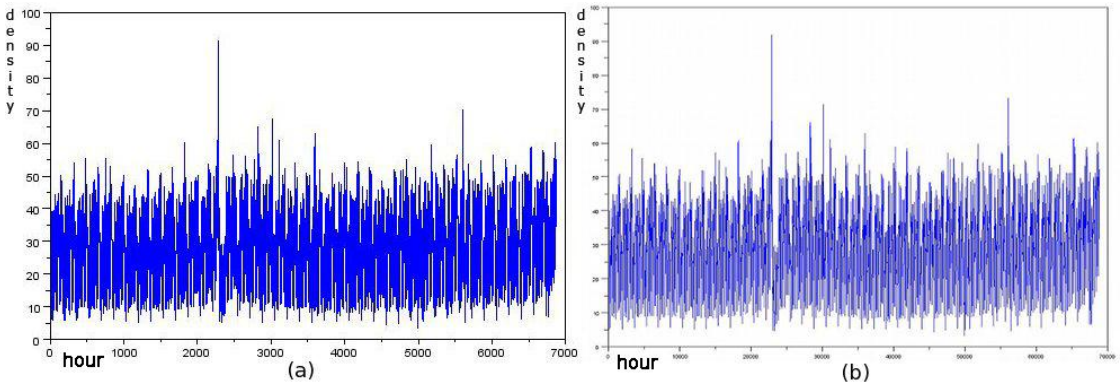



**Figure 1. (a) Hourly Densities of the Blk Segment**
**(b) 6-minute Densities of the Blk Segment**

The calculated hourly densities of the Boc, Mcy, and Mrl segments are shown in Figures 2, 3, and 4, 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' behaviors 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 Saints' Day or semestral break at the end of October until the start of November, and the Christmas holidays, respectively.

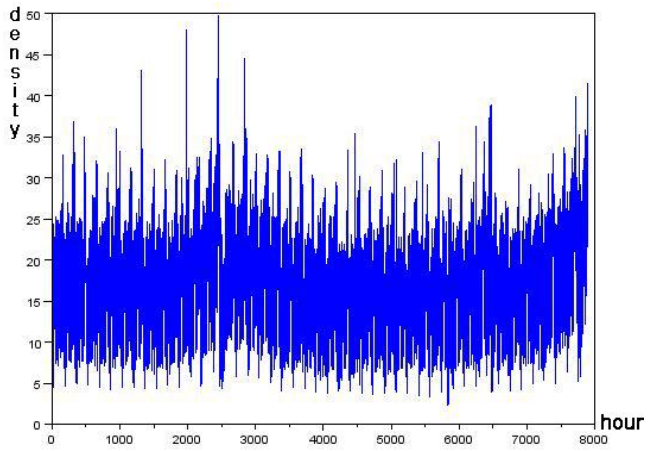



**Figure 2. Hourly Densities of the Boc Segment**

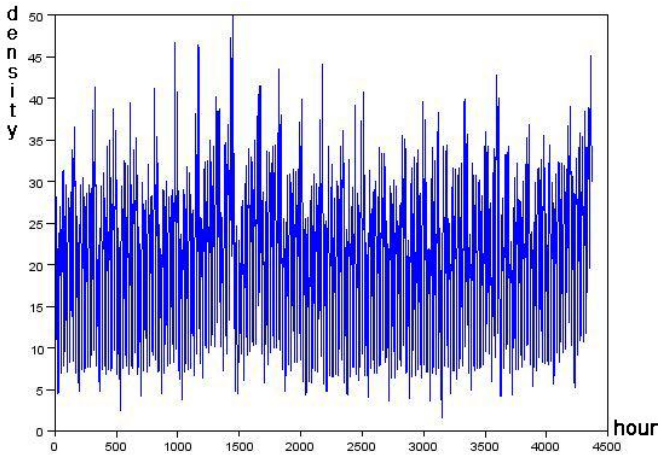



**Figure 3. Hourly Densities of the Mcy Segment**

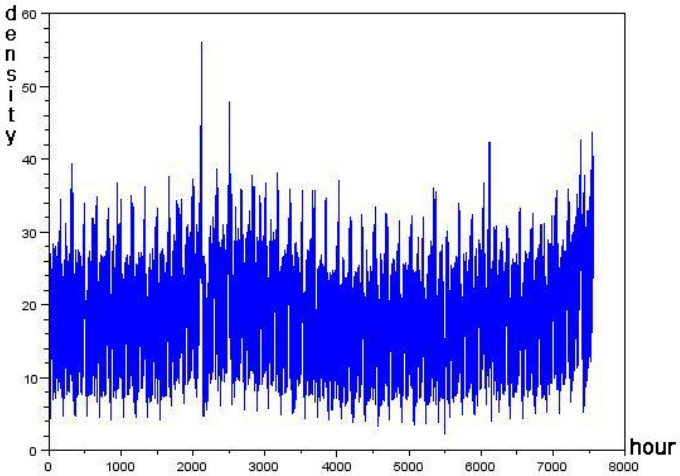



**Figure 4. Hourly Densities of the Mrl Segment**

### 3.2. Data Signature-based Visualization Models

**3.2.1. Cluster models:** The cluster models of the NLEX segments Blk, Boc, Mcy, and Mrl, as shown in Tables 1, 2, 3, and 4 were generated using the data signatures of the hourly traffic density data sets.

**Table 1. Clustering of Blk's Hourly Density Data Set's Weeks**

| Cluster | Week/s |
|---|---|
| 0 | 1, 8, 9, 10, 27, 28, 29, 30, 31, 35, 39 |
| 1 | 2, 3, 4, 5, 7, 11, 12, 25, 33, 37 |
| 2 | 13, 14, 16, 18, 20, 23, 38, 41, 42, 43, 44, 45, 46, 47, 48 |
| 3 | 15 |
| 4 | 19, 49, 50, 52 |

**Table 2. Clustering of Boc's Hourly Density Data Set's Weeks**

| Cluster | Week/s |
|---|---|
| 0 | 1, 6, 7, 8 |
| 1 | 2, 12, 24, 43 |
| 2 | 3, 4, 5, 47 |
| 3 | 9, 10, 11, 25, 26, 29, 41, 42, 45, 46, 48 |
| 4 | 13, 44, 49, 50 |
| 5 | 27, 32, 33, 34, 35, 36, 37, 38 |
| 6 | 28, 31 |
| 7 | 30, 39 |
| 8 | 15, 16 |
| 9 | 17 |
| 10 | 18 |
| 11 | 19 |
| 12 | 51 |
| 13 | 52 |

**Table 3. Clustering of Mcy's Hourly Density Data Set's Weeks**

| Cluster | Week/s |
|---|---|
| 0 | 1, 2, 4, 6, 7, 10, 41, 42, 46, 48 |
| 1 | 8, 30, 31, 32, 36, 37, 38, 39 |
| 2 | 14, 17 |
| 3 | 15, 23, 44, 50 |
| 4 | 19 |
| 5 | 51 |

**Table 4. Clustering of Mrl's Hourly Density Data Set's Weeks**

| Cluster | Week/s |
|---|---|
| 0 | 1, 6, 7, 8, 9, 10, 25, 42, 45, 46, 48 |
| 1 | 2, 3, 4, 5, 24, 43, 47 |
| 2 | 27, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 41 |
| 3 | 13, 44, 50 |
| 4 | 14, 16, 19, 21 |
| 5 | 15, 18 |
| 6 | 17 |
| 7 | 22 |
| 8 | 23 |
| 9 | 49 |
| 10 | 51 |
| 11 | 52 |

Among the 4 hourly density data sets, it was Mrl that was found to have the most outliers, with 6 weeks as outliers (Weeks 17, 22, 23, 49, 51, and 52). The second density data set with the most outliers is Boc with 5 weeks as outliers (Weeks 17, 18, 19, 51, and 52). The third density data set with the most outliers is Mcy with 2 weeks outliers (Weeks 19 and 51). Blk had the least outliers, only having 1 week as an outlier (Week 15).

Using the DSIV tool, the data signature for the 6-minute density data set of the Blk segment is also produced and the clusters of the weeks are generated. The clusters of this data set are shown in Table 5.

**Table 5. Clustering of Blk's 6-minute Density Data Set's Weeks**

| Cluster | Week/s |
|---|---|
| 0 | 1, 8, 9, 10, 27, 28, 29, 30, 31, 35, 39 |
| 1 | 2, 3, 4, 5, 7, 11, 12, 25, 33, 37 |
| 2 | 13, 14, 16, 18, 20, 43, 44, 47 |
| 3 | 15 |
| 4 | 23, 38, 41, 42, 45, 46, 48 |
| 5 | 19 |
| 6 | 49 |
| 7 | 50 |
| 8 | 52 |

The weeks of the hourly and 6-minute density data sets have the same cluster model except for two variations. Cluster 2 in the hourly model is split into two different clusters (Clusters 2 and 4). The weeks of Cluster 4 of the first model were separated as outliers in the second one. Thus, Clusters 3, 5, 6, 7, and 8 in the 6-minute density data set are all outliers. These divisions occurred because of the weeks' high intracluster distance. 6-minute density data set has more points to consider than the hourly density data set.

**3.2.2. NMDS visualizations:** The nMDS visualizations of the data signatures from the hourly traffic density data set of the NLEX segments are shown in Figures 5, 6, 7, and 8, respectively. In an nMDS plot, the points which are reflected with the same symbols and coloring belong to the same cluster.

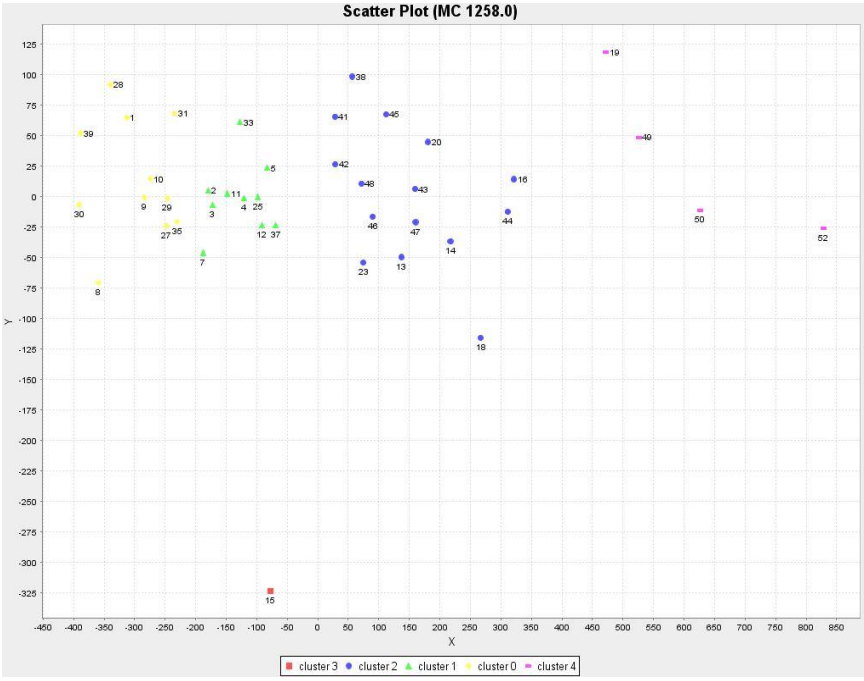



**Figure 5. NMDS plot of Blk's Hourly Density Data Set**

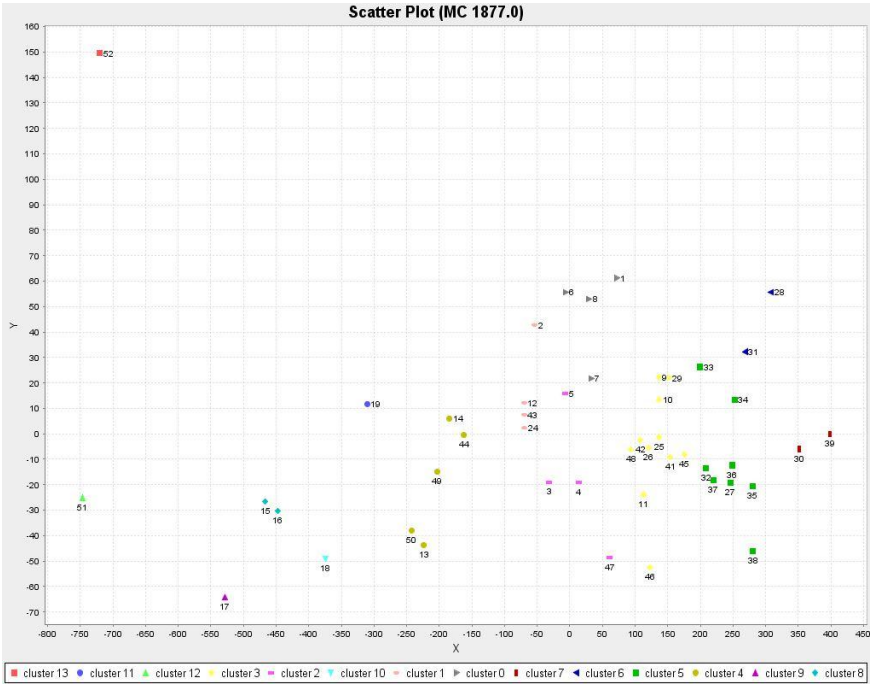



**Figure 6. NMDS Plot of Boc's Density Data Set**

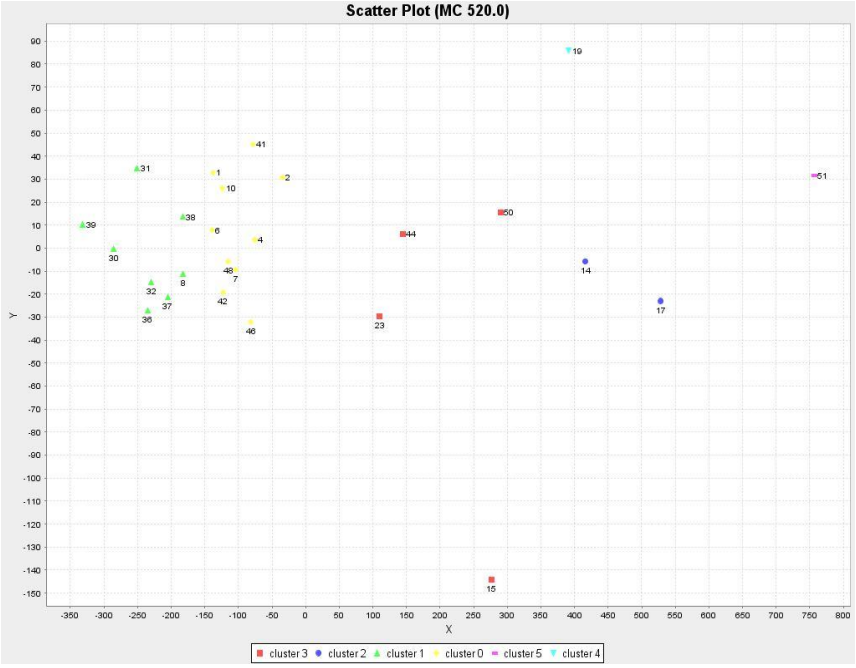



**Figure 7. NMDS Plot of Mcy's Density Data Set**

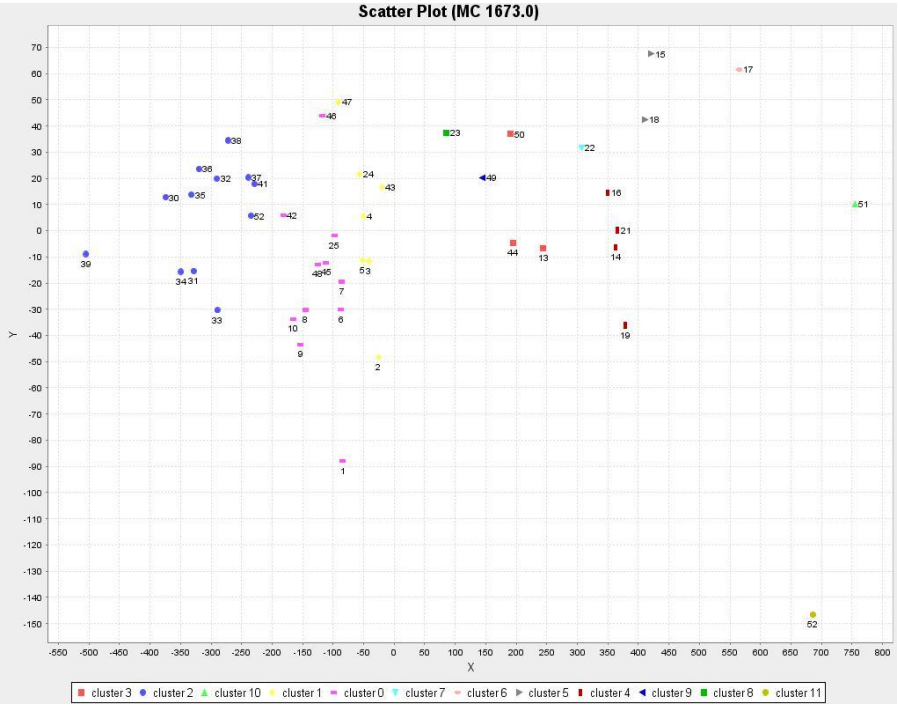



**Figure 8. NMDS Plot of Mrl's Density Data Set**

Using the Blk density nMDS visualization, curves are fitted to each cluster's data points. The best fit curves determined by RMSD are found to be linear as shown in Table 6. No curve was fitted for Cluster 3 due to the lack of data necessary in constructing the curve.

**Table 6. Best Fit Curves of Blk's Hourly Density Data Set's Clusters**

| Cluster | Curve |
|---|---|
| 0 | $1529 - 0.0789x$ |
| 1 | $686.5867 + 0.0934x$ |
| 2 | $2165.6912 - 0.1948x$ |
| 4 | $2664.2925 - 0.259x$ |

The best fit curves are used to determine the confidence bands and confidence ellipse of each cluster. With the resulting best fit curves, confidence ellipses, and confidence bands, outliers and potential outliers are examined. The visualization with each cluster's best fit curves, confidence bands and ellipse is shown in Figure 9.

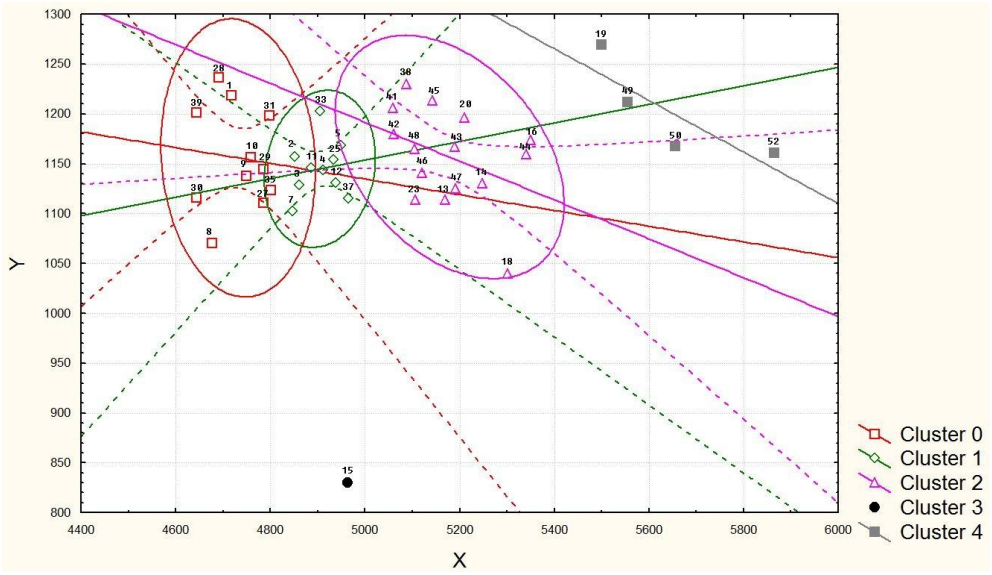



**Figure 9. NMDS Plot with Confidence Measures of Blk's Hourly Density Data Set**

Cluster 4's ellipse is not considered because it only has a few points and it covers all points of Cluster 2, making all these points ambiguous. Cluster 4 of the hourly density data set contains no potential outliers because it is relatively far from the points of other cluster, preventing them from being covered by other clusters' ellipses. Cluster 3 is an outlier of the hourly density data set because it is the only has one data point. All potential outliers are found to be ambiguous since all points are covered by their own confidence ellipse. Table 7 shows each cluster and its potential outliers.

**Table 7. Potential Outliers of Blk's Hourly Density Data Set's Clusters**

| Cluster | Ambiguous Potential Outliers |
|---|---|
| 0 | Wk1, Wk8, Wk28, Wk29, Wk31, Wk35 |
| 1 | Wk2, Wk3, Wk5, Wk7, Wk11, Wk33, Wk37 |
| 2 | Wk13, Wk16, Wk18, Wk20, Wk23, Wk38, Wk45, Wk46, Wk47 |

The nMDS plot of Blk's 6-minute density data set is also produced and shown in Figure 10.

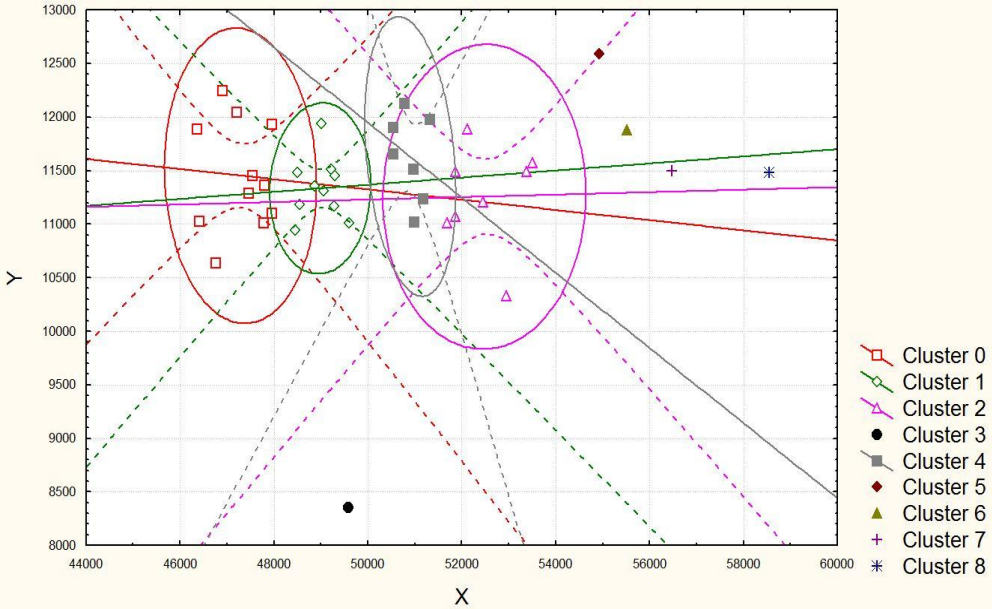



**Figure 10. NMDS Plot with Confidence Measures of Blk's 6-minute Density Data Set**

The potential outliers of Blk's 6-minute density data set are found to be ambiguous. As seen in Tables 7 and 8, Blk's hourly and 6-minute density data sets have produced similar ambiguous potential outliers. As mentioned earlier, Cluster 2 of the hourly data set had split into two clusters in the 6-minute data set which is why the potential outliers of the first data set's Cluster 2 were also split into two clusters in the second data set (Cluster 2 and 4).

**Table 8. Potential Outliers of Blk's 6-minute Density Data Set's Clusters**

| Cluster | Ambiguous Potential Outliers |
|---|---|
| 0 | Wk1, Wk8, Wk27, Wk28, Wk31, Wk35 |
| 1 | Wk2, Wk3, Wk7, Wk11, Wk33, Wk37 |
| 2 | Wk13, Wk18, Wk20, Wk43, Wk47 |
| 4 | Wk23, Wk38, Wk42, Wk45, Wk46, Wk48 |

The details and evidences of the points being potential outliers in their clusters in all segments are found in Section 3.2.3.

**3.2.3. Data image visualizations:** The data images of the hourly density data sets are also analyzed to determine the time frames of regular and irregular densities. The weeks of the hourly data sets' data images are arranged according to their clusters. Figure 11 shows the data image of Blk segment's hourly traffic density data set.

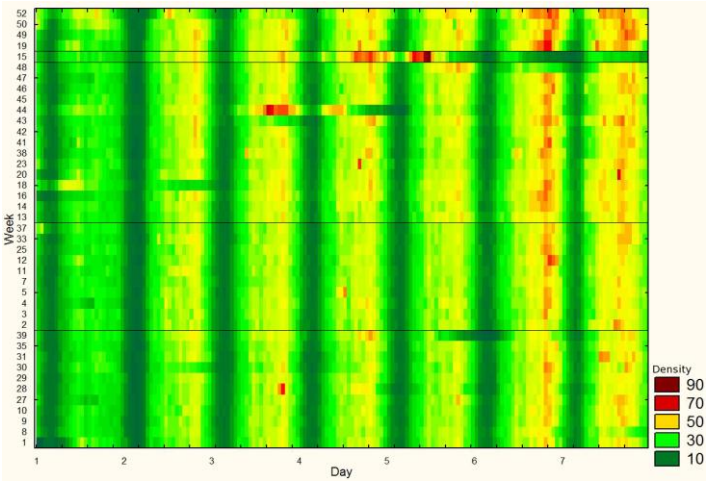



**Figure 11. Data Image of Blk's Hourly Density Data Set**

The densities that are consistent in their value with respect to the same day throughout the year are the regular densities. Irregular density values are inconsistent with respect to the regular values of their day. As seen in Figure 11, week 18's Day 2 (Monday) has irregular density because it has relatively lower density than the other Monday's of the year. Week 15's Day 4 (Wednesday) has higher density than the other Wednesday's of the year, making it irregular. The weeks of Cluster 2 should be the time frames of regular density since Cluster 2 has the highest number of weeks among the clusters. But since incidents are inevitable, irregular densities can be observed in Cluster 2.

The data image of the 6-minute density data set of the Blk segment is also generated. The weeks of the 6-minute data set's data image are arranged in such a way that they follow the order of the weeks of the Blk segment's hourly data set's data image. This is done for a more convenient comparison between the two data images.

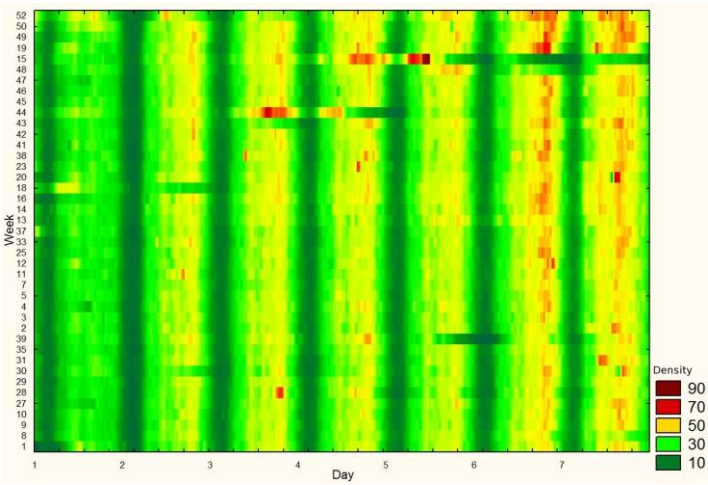



**Figure 12. Data Image of Blk's 6-minute Density Data Set**

Figures 11 and 12 exhibit similar behaviors. This further validates the accuracy of the hourly density model. With the validation, the hourly data set is sufficient in representing the whole traffic data set.

As we can see from the produced cluster model, nMDS visualizations, and the best fit curves, confidence bands and ellipses, the 6-minute density data set has more potential outliers than the hourly density data set. It can be ascertained that most of the potential outliers in the hourly model are also discovered as such in the 6-minute data, validating our initial result.

Figures 13, 14, and 15 also show the data images for Boc, Mcy, and Mrl segments' hourly density data sets. They are clustered differently but overall, the behavior is the same with Blk's when it comes to events that is not exclusive to the segments (e.g. holidays). Irregularities are also present in all data images. The details describing these discovered irregularities are discussed in Section 3.2.3.

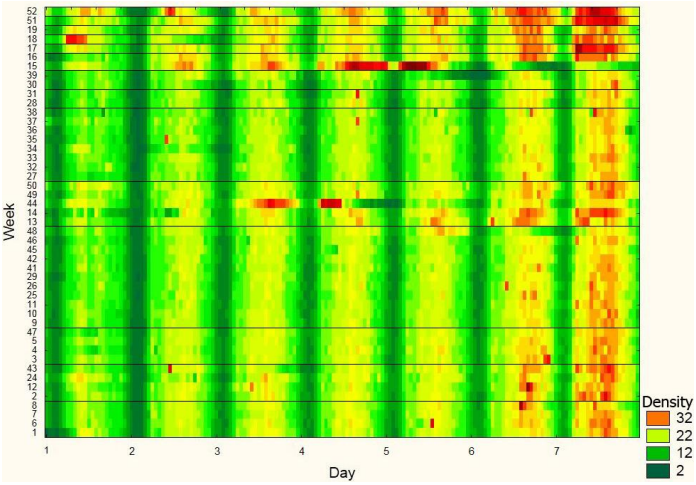



**Figure 13. Data Image of Boc Segment**

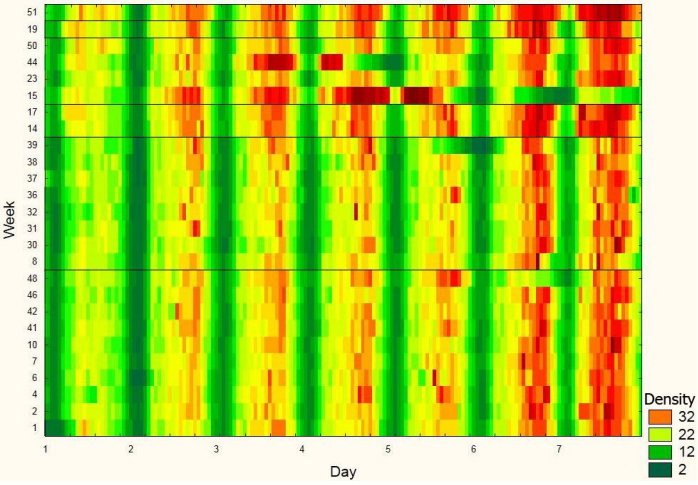



**Figure 14. Data Image of Mcy Segment**

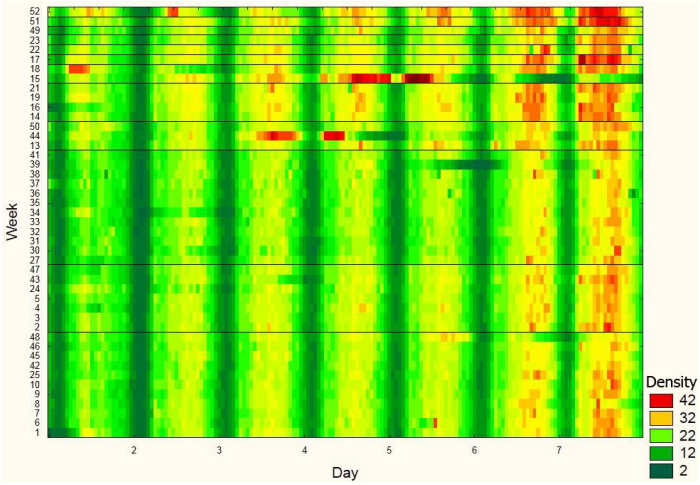



**Figure 15. Data Image of Mrl Segment**

**3.2.4. Visualization analysis:** For Blk's hourly traffic density data set, we were able to find evidences of some of the points in the cluster model that we have produced as being potential outliers in Table 7 as shown below. We also state here the identified in traffic flow using Figure 11. We look for the events that triggered the irregular densities in the data image by focusing on the time frames of relatively high, relatively low, and extremely high density values.

- Certain days of Cluster 0's weeks 1, 8, 29, and 35 have relatively lower density than the common density for that particular day in the cluster. For example, week 1's Sunday has lower density than other Sundays of the same cluster. On the other hand, certain days of this cluster's weeks 28, and 31 have relatively higher density than the common density for that particular day in the cluster.
- Some of Cluster 1's weeks have similar behavior with Cluster 0's weeks. This is because the confidence ellipses of Clusters 0 and 2 cover these weeks, making them ambiguous. However, these weeks belong to Cluster 1 because of distinct behaviors exhibited only by the weeks in Cluster 1. We can see this in week 2, where its density is relatively lower on Friday, but relatively higher on Saturday which is a characteristic of Cluster 1.
- Cluster 2's potential outliers have days that have a different density value than the common one. Week 13, for instance, has relatively lower densities on Friday and Saturday, as compared to its co-weeks in the cluster.
- Extremely low densities are not classified specifically because this is a regular occurrence in expressways during midnight until early morning.
- Extremely high density values are observed during important holidays in the country. This is reflected by the time frames with a dark red range. The weeks are weeks 15 and 44. The start of Holy Week happens on week 15 (Day 4) and the All Saints' Day holiday happens on week 44 (Day 3). These are the days when people visit their provinces located at the north of Luzon.

- Sudden increase of density values during days with consistent density values usually observed are irregular densities that are classified as relatively high density values. Reasons behind this sudden increase include accidents on weeks 12 (Day 6), 20 (Day 7), and 28 (Day 3) and departure for holidays on weeks 50 and 52.
- Relatively low densities are observed on sudden decrease in densities the same way relatively high densities are observed on sudden increase in densities. We observe that the days after the highest density value occurred have relatively low density values. This is due to the large number of people departing at the same time at the official start of the vacation for the Holy Week. 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 and week 16's day 1).
- During weeks 39 and 44, typhoons *Milenyo* and *Paeng*, respectively, struck the country. Travel advisory from weather domain experts prevented the people from traveling which is why there is a low density turnout on the said time frames. There is also a low density turnout during some holidays. Christmas Eve (week 32's day 1) and New Year (week 1's day 1) are observed in the country with people staying inside their houses to celebrate. Day 2 of week 18 also has relatively low density. This is due to the Labor Day holiday. Most of the professional drivers who pass by NLEX are on holiday.
- The Data Image also reflects relatively low densities during Day 1 of weeks 4, 27, and 47. This is due to the many people watching the fights of the boxer Manny Pacquiao. Pacquiao-Morales 2 happened during week 4 (January 22, Philippine time). Pacquiao-Larios happened during week 27 (July 2, Philippine time). Pacquiao-Morales 3 happened during week 47 (November 19, Philippine time).

For some of the potential outliers in the 6-minute density data set's model which are not found in the hourly density data set's, the following observations were derived. Cluster 0's week 27 has relatively lower density on Sunday than the usual density value of Sundays in the same cluster. Week 43 of Cluster 2 and week 42 of Cluster 4 are both similar to each other's cluster. The same goes for the majority of the potential outliers in both clusters. This is due to their clusters' confidence ellipses enclosing each other's points. This is further supported by the hourly data set's clustering wherein both clusters' points belong to one cluster (Cluster 2 of hourly data set).

As seen from the data images of the 3 segments' densities in Figures 13, 14, and 15, the following weeks have relatively high densities: 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 Saints Day. The following weeks have relatively low densities among the 3 segments: week1 (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.

## 4. Conclusions and Recommendations

We have shown in this paper that data signature-based density analysis can provide an efficient and effective representation of traffic behavior. Using the space mean speed instead of time mean speed produces realistic results because it considers the rate of movement of vehicles within a given section. Density analysis, together with thorough preprocessing of the data set, produces an effective congestion indicator.

With the data signature representation of the hourly density output data, analysis on traffic outliers can be conducted efficiently. With the same preprocessing and procedures done on the 6-minute density data set, comparison with the hourly density data set yielded similar results. Thus, the hourly density data set is validated to be accurate enough to be used in traffic congestion analysis. With the validation of a larger scaled data set, there are less data points to process, providing efficiency without compromising its accuracy.

With the outliers and potential outliers determined by our study, expressway management can have an efficient analysis of traffic behavior that can be used in anticipating traffic flow patterns. Traffic incidents can be addressed more efficiently to reduce accidents and other traffic obstructions. Additionally, to come up with a more generalized behavior of the whole expressway, it is recommended that multi-year traffic analysis on NLEX segments be conducted.

## Acknowledgements

The researchers like to thank Dr. Ma. Sheilah Gaabucayan-Napalang and Dr. Jose Regin Regidor for validating the results and providing the data sets. Mr. Maravilla likes to thank 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*.

## References

[1] Malinao J, Juayong R, Corpuz F, Yap J, Adorna H, "Data Signatures for Traffic Data Analysis", Proceedings of the 7th National Conference for Information Technology Education, **(2009)**.

[2] Cox T, Cox M, "Multidimensional Scaling", **(1994),** pp. 42-69.

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

[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)**.

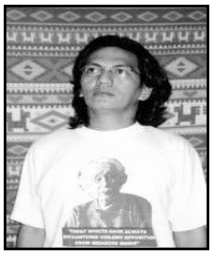
[6] Malinao J, Juayong R, Oquendo E, Tadlas R, Lee J, Clemente J, Gabucayan-Napalang M, Regidor J, Adorna H, "A Quantitative Analysis-based Algorithm for Optimal Data Signature Construction of Traffic Data Sets", Proceedings of the 1st AICS/GNU International Conference on Computers, Networks, Systems,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

[7] Malinao J, Tadlas R, Juayong R, Oquendo E, 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)**.

[8] 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)**.

[9] Oquendo E, Clemente J, Malinao J, Adorna H, "Characterizing Classes of Potential Outliers through Traffic Data Set Data Signature 2D nMDS Projection", Philippine Information Technology Journal, vol. 4, no. 1, **(2011)**.

## Authors



**Henry N. Adorna**

He is an Associate Professor and member of the Faculty of the Department of Computer Science, College of Engineering, University of the Philippines Diliman. He heads the Algorithms and Complexity Laboratory of the Department. He is currently the Project leader of the research project "Information Visualization via Data Signatures", 2009-2011, funded by the Engineering Research and Development for Technology (ERDT). His interests are in the Mathematical Foundations of Computer Science, in particular, Automata and Formal Language Theory, Discrete Mathematics and Algorithms for Hard Problems.

**Jasmine A. Malinao**

She is an Assistant Professor of the Department of Computer Science in the University of the Philippines Diliman and is a member of the Algorithms and Complexity Laboratory. Her interests include Data Mining and Representation, Visualization, Algorithmics, Design and Implementations.

**Reynaldo G. Maravilla, Jr.**

He is an undergraduate student in the Department of Computer Science and is a member of the Algorithms and Complexity Laboratory.

**Elise Raina A. Tabanda**

She is an undergraduate student in the Department of Computer Science and is a member of the Algorithms and Complexity Laboratory.