Electronic Preprint for Journal of Information Processing Vol.20 No.3

Regular Paper

A Quantitative Analysis-based Algorithm for Optimal Data Signature Construction of Traffic Data Sets

Jasmine A. Malinao[1,a)] Richelle Ann B. Juayong[1] Rona May U. Tadlas[1]
Jhoirene B. Clemente[1] Erlo Robert F. Oquendo[1] John Boaz Lee[1]
Ma. Sheilah Gaabucayan-Napalang[2,†1] Jose Regin F. Regidor[2,3]
Henry N. Adorna[1]

Received: September 13, 2011, Accepted: February 3, 2012

Abstract: In this paper, a new set of *m*-dimensional Power Spectrum-based data signatures is derived to obtain better Vector Fusion 2-dimensional visualizations of a time series and periodic *n*-dimensional traffic data set as compared with visualizations produced from using the entire set of *n*-dimensional Power Spectrum representations in literature, where $m \ll n$. We were able to ascertain that 4-dimensional data signatures provide empirically optimal representations with respect to the data set used. We have achieved ≈97.6% reduction in terms of data representation of the original *n*D data set with the signatures. We propose an algorithm that determines how good the selected set of *m*-dimensional signatures represents the *n*-dimensional data set in 2 dimensions in quantitative terms. We use the Vector Fusion visualization algorithm in transforming each signature from *m* dimensions into 2 dimensions. An improved set of qualitative criterion is drawn to measure the goodness of the 2-dimensional data signature-based visual representation of the original *n*-dimensional data set. Finally, we provide empirical testing, discuss the results, and conclude the contributions of the proposed methods.

Keywords: data signatures, discrete fourier transform, vector fusion, power spectrum, X-means

1. Introduction

A data signature, as defined in Ref. [1], is a mathematical data vector that captures the essence of a large data set in a small fraction of its original size. It had been shown in previous studies [2], [3] that Fourier-based data signatures employed on time series traffic data sets provide better characterization on sets of traffic flow behavior and unravel previously unknown information from the data set. In particular, these studies show the effectiveness of using such type of signatures to produce an optimal cluster model from the 2006 North Luzon Expressway (NLEX) Balintawak-Northbound (BLK-NB) traffic volume data set. The data set had to be preprocessed, partitioned, and projected as discrete input time domain signals. Each signal, representing a week with 168 hourly traffic volume entries, is then decomposed through the Fast Fourier Transform and its corresponding set of Power Spectrum components is computed. A data signature is obtained by selecting the first 85 components to represent each week in its frequency domain. The X-Means [4] clustering algorithm was used to group all similar weeks and extract a number of outliers by using these data signatures.

Shown in **Fig. 1** is the time domain visualization of the data set with rows (representing weeks of 2006) structured contiguously conforming to the cluster model produced through X-Means on the data signatures, denoted as XMeans(F,85). The horizontal axis reflects the 168 hourly total traffic volume of each week from Sunday to Saturday. Each pixel is colored based on the current traffic volume of a time step in a week. This image is produced using the Iterative Data Image Rotated Bar Graph (iDIRBrG)-based approach in Ref. [5].

Using Fig. 1 of the time domain data set, it is used to determine inter-cluster and intra-cluster similarities and differences. Outliers are also easily identified with the significant changes highlighted in various sections in their rows. In addition to these results, analysts are also capable of mining out weeks which belong to a cluster that possesses peculiar behavior among its cluster co-members. These weeks are referred to as *potential outliers* in Ref. [3]. Cluster and outlier analysis using this cluster model is also detailed in the same paper. However, using a 2-dimensional Vector Fusion (VF) visualization technique [7], [8], a similar cluster, outlier, and potential outlier analysis can also be accomplished in a more simplified and straightforward manner than the iDIRBrG-based visualizations. We initially used the 168-dimensional Power Spectrum components of each week and use VF to obtain a scattergram of the data set. The scattergram points are then colored using the cluster model information ob-

[1] Department of Computer Science (Algorithms and Complexity), College of Engineering, University of the Philippines, Diliman, Quezon City 1101, Metro Manila, Philippines
[2] National Center for Transportation Studies, University of the Philippines, Diliman, Quezon City 1101, Metro Manila, Philippines
[3] Institute of Civil Engineering, College of Engineering, University of the Philippines, Diliman, Quezon City 1101, Metro Manila, Philippines
[†1] Presently with School of Urban and Regional Planning, University of the Philippines, Diliman, Quezon City 1101, Metro Manila, Philippines
[a)] jamalinao@dcs.upd.edu.ph

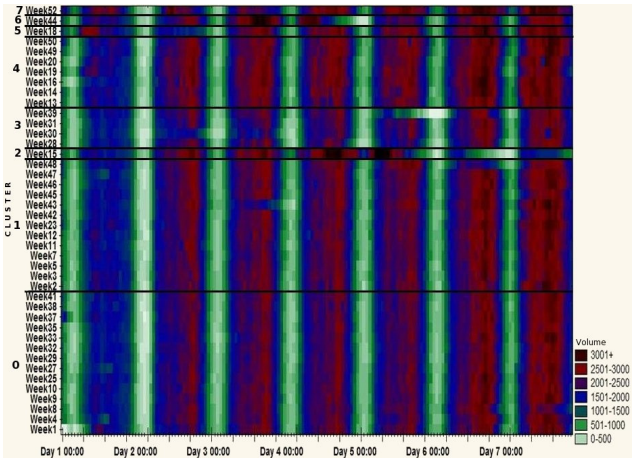

Fig. 1 iDIRBrG-based [5] visualization of the time-domain data set. The y-axis is composed of weeks rearranged according to the results of frequency-domain clustering (using the 85-dimensional data signatures) through X-means. The lines separate the clusters from one another.

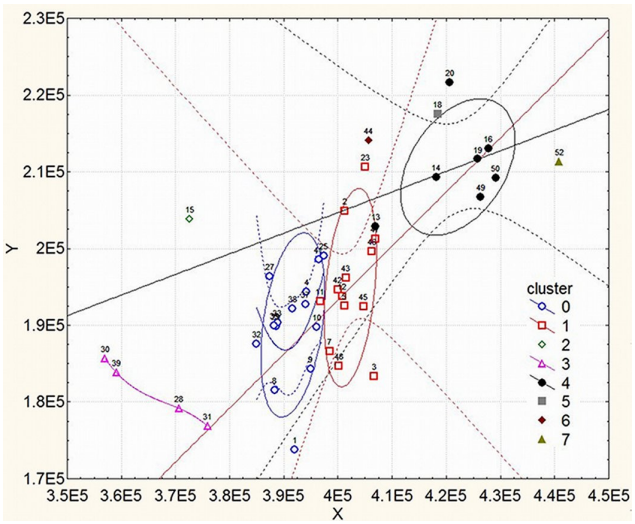

Fig. 2 VF(85,168) for the 2006 NLEX BLK-NB traffic volume data set.

tained through X-Means(F,85).

Shown in **Fig. 2** is the VF Visualization [7] VF(85,168) where each point (representing a week) is colored using the results of XMeans(F,85). The notation VF(X,Y) means that we have used the first X and Y components of the Power Spectrum as inputs to the X-Means clustering algorithm and the VF visualization algorithm, respectively. In the latter part of the paper, the second component Y may also be a set of elements associated to a Power Spectrum component A_i for some $i \in \{1, 2, \ldots, 168\}$. Co-members in a cluster are identifiable using a unique color assigned to them. Additionally, a cluster's behavior may further be described by generating its confidence ellipses, bands and best-fit curves [6].

With the use of the entire set of Power Spectrum components of each week, a "good enough" VF visualization is obtained from the data set. However, a previous work in Ref. [3] has shown the possibility of obtaining a signature from a small subset of the n-dimensional Power Spectrum components while still achieving a good cluster model of the data set. This model, in fact, was concluded as a better one when compared to the model obtained by using the original n-dimensional Power Spectrums of the data set. Thus, this study aims to establish the following results,

- determine an optimal set of data signatures with smaller dimensionality m, where $m \ll n$, for optimal VF visualization purposes;
- improve the qualitative goodness criteria [3] by measuring the improvements of the 2-dimensional VF representation of the data over its m-dimensional data signature;
- provide novel information from the data regarding actual traffic incidents by using the newly obtained optimal m-dimensional data signatures in the VF visualization to project data points with interesting or peculiar behavior far enough from their co-members in the cluster. These points may be considered as potential outliers [6].
- effectiveness of representation of the m-dimensional data signatures using the algorithm on time series periodic traffic data sets. Through this process, we show the robustness and reliability of the signatures for 2-dimensional VF data representation. We also show that this characteristic is maintained while achieving a simplistic, intuitive, abstract, yet readily-interpretable representations of large n-dimensional traffic data sets. These interpretations are achieved without the need to refer to the time-domain n-dimensional data image visualizations.

In Section 2, we give definitions and notations to the concepts building the theoretical backbone of this study. Provided in Section 3 are the details on how "good" visualizations are obtained through qualitative and quantitative approaches. We further strengthen our qualitative results by introducing an algorithm to obtain a quantitative value measuring the consistency of the visualization of the vector-fused data signatures with respect to their actual Euclidean distances. In Section 3.3, we give characterizations of Power Spectrum values so as to select reasonable components for data signature construction. Finally, we show empirical tests on various data signatures on various data sets and detailed discussions on their results in Section 4 and conclusions in Section 5.

2. Basic Definitions and Notations

2.1 The Data Sets

The data sets used in this work have the following characteristics: time-series, periodic and multidimensional. In particular, we used the 2006–2009 NLEX Balintawak Northbound (BLK-NB) data set provided by the Manila North Tollways Corporation (MNTC) through the National Center of Transportation Studies (NCTS). A record contains hourly entries accumulated via an automatic detector embedded in every lane of NLEX's segments. The detector inserts the mean spot volume in its record for each lane per hour, thus, 168 data points are collected in each week. We totaled these values in all four lanes to obtain a 52×168 data matrix, i.e., 52 weeks with 168 data points each.

2.2 Power Spectrums

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-dimensional

weekly partition in the NLEX BLK-NB time-series traffic data set as discrete signals, we obtain their Power Spectrums through the Discrete Fourier Transform (DFT) decomposition as shown below,

$$\theta(t) = \mu_0 + \sum_{k=1}^{n-1}\left(a_k \cos\frac{2\pi k}{n} - ib_k \sin\frac{2\pi k}{n}\right),$$

where μ_0 is the component referred to as the *offset* of the signal translated from the horizontal axis. Using DFT, a vector of real numbers can produce a vector F of frequency components of the same length, where $F = (a_0 \pm b_0 i, a_1 \pm b_1 i, a_2 \pm b_2 i, \ldots, a_n - 1 \pm b_{n-1} i)$, $\mu_0 = a_0 \pm b_0 i$ and $i^2 = -1$. For the resulting n-dimensional vector, we produce distinct values for $a_0 \pm b_0 i, a_1 \pm b_1 i, a_2 \pm b_2 i, \ldots, a_{n/2} \pm b_{n/2} i$ and the succeeding values are their complex conjugates.

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 of the corresponding sine-cosine pair and then getting its square root. The Power Spectrum A_k of the signal, $k = 0, 1, \ldots, n-1$ is given by,

$$A_k = \sqrt{{a_k}^2 + {b_k}^2}.$$

2.3 Vector Fusion Visualization

In literature [7], a method is introduced to provide a 3-dimensional perspective of any given n-dimensional data vector by using the Single-point Broken-line Parallel coordinates (SBP) algorithm. Each instance in a given n-dimensional data set is projected in 3-dimensional as a vector resultant of its components. The paper [8] simplifies this visual representation such that an n-dimensional data point $\mathbf{w} = [w_1, w_2, \ldots, w_n]$ is represented as a 2-dimensional resultant point in a scattergram by summing all of data point's component w_j using a precomputed angle θ_j with w_{j-1}, $j = 1, 2, \ldots, n$. Shown below is the formula to compute the 2-dimensional coordinates $(SPBx, SBPy)$ for an n-dimensional data point $\mathbf{w}$.

$$\begin{aligned}\mathbf{w} &= w_1 e^{i\theta_1} + w_2 e^{i\theta_2} + \ldots + w_n e^{i\theta_n} \\ &= \sum_{j=1}^{n} w_j \cos(\theta_j) + i \sum_{j=1}^{n} w_j \sin(\theta_j) \\ &= (w_{sumX}, w_{sumY}) = (SPBx, SBPy)\end{aligned}$$

where $\theta_j = (j-1)180°/n$, n is the dimension of the vector, and w_j is the value of the j^{th} dimension.

3. Methodology

3.1 A Qualitative Goodness Measure of the VF Visualization

After obtaining the set of Power Spectrum components for each n-dimensional week in the data set, m of these components shall be selected and used as its data signature for VF visualization purposes, where $m \leq n$. To obtain relationships of the weeks using the scattergram visualization, we shall use the cluster model XMeans(F,85) to color each scatter point. Then, we determine how well the 2-dimensional scattergram represents the pre-computed point-to-point, intercluster, and intracluster relationships obtained from XMeans(F,85). Shown below is the criteria [3] that we have improved in this work.

(1) **Closeness of co-members.** A good visualization should show reasonable visual proximity of points belonging to a common cluster. Regions occupied by clusters should have minimal overlaps in the visualization.

(2) **Visibility of all points.** No total occlusion should exist among the points.

(3) **Outlier detection.** Outliers seen after using X-Means clustering algorithm should have a significant distance from all other weeks such that they can easily be pinpointed in the visualization.

(4) **Detection of potential outliers.** Potential outliers, i.e., weeks within a cluster that show "interesting" behavior as seen in the iDIRBrG-based visualizations, should be found near or at the periphery of the region occupied by a cluster. These should be projected far from their co-members in the VF visualization.

The detection of potential outliers by use of projection of convex hulls along the periphery of a cluster region and determining whether they are "far enough" from their co-members are both highly subjective processes. Candidate potential outliers may exist along the convex hull but may be significantly spatially near cluster centroids compared to other points that may have a larger spatial distance but are not along the hull. Thus, a previous work Ref. [6] formalized the definition of potential outliers using regression curves, confidence bands, and confidence ellipses. We shall use this method to provide us a list of these points in our empirical tests.

(5) **Characterizing relationship across clusters.** A good visualization should aid users in efficiently determining what characteristics differentiate one cluster from another in the data set.

(6) **Consistency of the 2-dimensional representation of the data points** Suppose a point R has a smaller data signature Euclidean distance from a point S compared to a point Q. Then, the (SBPx, SBPy) Euclidean distance of R to S must also be smaller compared with R's distance to Q.

Exploring the first five criterions is easily done. Figure 2 in Section 1 was shown in the previous work [3] to be the best VF visualization of the 2006 NLEX BLK-NB traffic volume data set based on these first 5 criterions. In this paper, we propose an algorithm that checks the consistency stated in the last criterion.

3.2 Exploring Criterion 6 via Benchmark Model M_1 and VF Test Model M_2

Criterion 6 requires us to initially model how points, represented by their data signatures, relate to one another in terms of their Euclidean distances. This is the benchmark model M_1.

To build the benchmark model M_1, let $R = [r_0, r_1, \ldots, r_m]$ and $S = [s_0, s_1, \ldots, s_m]$ be data signatures from two arbitrary data points from the data set D. Let $\delta(R, S)$ be the Euclidean Distance between R and S, $\delta(R, S) = \sqrt{\sum_{i=1}^{m} (r_i - s_i)^2}$. By computing all the Euclidean distances of all data signatures in D, M_1 shall have a distance matrix for all the points of D.

We build the test model M_2 by initially using the VF visualization algorithm to generate the (SBPx, SBPy) 2-dimensional representation of the m-dimensional data signature in D. Let $R' = (SBPx_1, SBPy_1)$ and $S' = (SBPx_2, SBPy_2)$ be the vector-fused data signature R and S, respectively.

Let $\delta(R', S')$ be the Euclidean Distance from R' to S'. Then, compute all the Euclidean distances of all vector-fused data signatures in D to obtain a distance matrix of D for M_2. Finally, using the distance matrices of M_1 and M_2, we compare how consistent is M_2's 2-dimensional representation of the original m-dimensional signatures of M_1 by using the algorithm below,

Algorithm for Quantitative Analysis

(1) Let N be the number of weeks in the data set D. For each week R in D, create a list L_R containing every other week $S \in D$ arranged from the smallest to the largest Euclidean Distance $\delta(R, S)$ from R of M_1.

(2) Get the maximum distance $MaxD$ which is equal to the distance from R to the last week in the list L_R. Let $d = MaxD/N$. Let $P(i)$ be the set of weeks in D in the i^{th} partition in the list L_R, i.e., the set of weeks found at the distance $q, q \in (i * d, (i+1) * d]$, where $i = 0, 1, \ldots, N-1$. Let $Count(i)$ be equal to the number of weeks in $P(i)$, where $i = 0, 1, \ldots, N-1$.

(3) Create a list L'_R containing every other week $S \in D$ ranked from the one with the closest Vector-Fusion Visualization Euclidean distance $\delta(R', S')$ in M_2 to the farthest (with respect to R).

(4) Let $P'(i)$ of M_2 be the i^{th} partition in the list L'_R containing the set of weeks ranked from r, where $\forall r$,

$$r \in \left[\left(\sum_{j=0}^{i-1} |P(j)|\right) + 1, \sum_{j=0}^{i} |P(j)|\right].$$

(5) For each partition i, $i = 0, 1, \ldots, N-1$, compute for the number of matches at the i^{th} partition of M_1 using R, denoted as $M(R, i)$, where $M(R, i) = P(i) \cap P'(i)$. Then, compute for the errors in M_2 with respect to M_1 at the i^{th} partition using R, denoted as $A(R, i)$,

$$A(R, i) = \frac{\sum_{\forall S : S \in P(i)} |i - i'|}{N},$$

where $S \in P(i')$ of M_2. Finally, compute for the consistency of M_2 with M_1 using R on the i^{th} partition, denoted as $Cons(R, i)$, where

$$Cons(R, i) = M(R, i) - A(R, i).$$

(6) Compute for the overall consistency of M_2 with M_1 using R as $OC(R)$,

$$OC(R) = \sum_{i=0}^{N-1} Cons(R, i).$$

Finally, compute for the model consistency of M_2 with the benchmark model M_1 using all data points in the data set D, denoted as MC, were

$$MC = \sum_{\forall R : R \in D} OC(R).$$

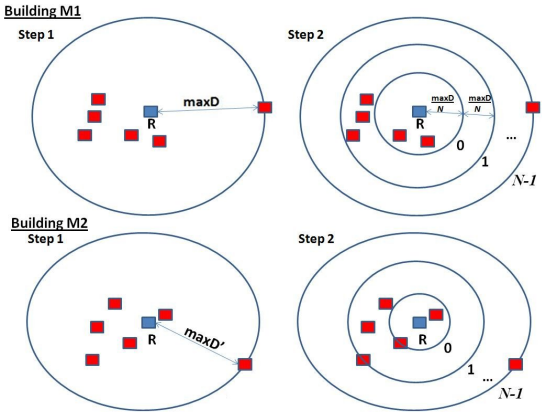

Fig. 3 Benchmark and Test Models M_1 and M_2, respectively.

The algorithm above is illustrated in **Fig. 3**. Step 1 builds M_1 and M_2 by computing $maxD$ and $maxD'$ as the distance of a week to the farthest point in the m and 2 dimensions, respectively. Note that model M_1 is constructed using the original signatures while model M_2 is achieved from vector-fused signatures. In Step 2 of building M_1, the relationships of week R with all other weeks in the data are established and ranking of weeks based on proximity are determined by projecting concentric circles from the farthest point to R. These circles determine partition numbers associated to the other weeks in the data set. This information is accounted for in performing Step 2 of building M_2. The relations of R with all other points in M_1 are then checked and the consistency of these relations with respect to its counterpart in M_2 is analyzed. Since there exist distortions in applying the VF algorithm for converting from an m-dimensional space to a 2-dimensional space, we can then identify a data signature construction that yields optimal visualizations for data sets. The algorithm performs all of the aforementioned processes for each week in the data, thus, it determines the consistency of the projections in $O(N^2)$ time and space complexity, where N is the number of weeks in the data set analyzed. A perfect representation of M_1 in M_2 would have every R and all other weeks placed in their correct partitions for both models. In such case, the algorithm outputs the maximum quantitative analysis value of N^2.

3.3 On Obtaining Candidate Optimal Data Signatures for Vector Fusion Visualization

In recent studies [2], [3], Power Spectrum-based data signatures of each week in the 2006 NLEX BLK-NB traffic volume data set were used in obtaining optimal cluster models through the X-Means clustering algorithm. An optimal cluster model can be obtained by using the first 85 components of the Power Spectrums of each input rather than using all of them. **Figure 4** shows the values of the first 85 components of the Spectrum. When projecting the last few components, we just obtain a mirror image of the visualization below.

As seen in Fig. 4, almost all weeks had the 7th dominant Power Spectrum component (also known as *harmonic*). The first 7 harmonics show significant variations compared to the succeeding values. Note the 42nd harmonic shows a significant increase of the value from the normally-decreasing values of previous val-

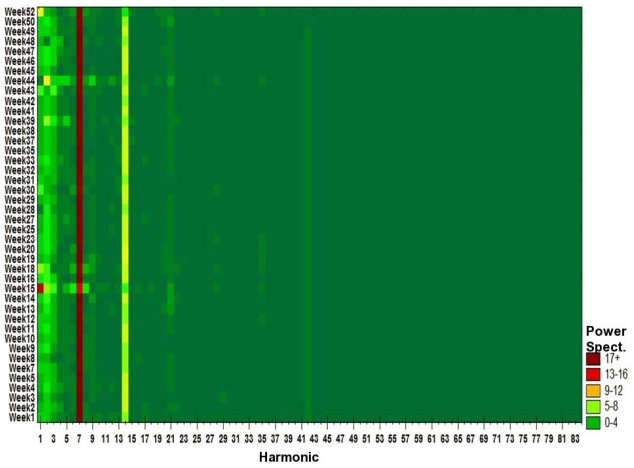

Fig. 4 Power Spectrums $A_1, A_2, \ldots, A_{85}$ for 2006 Traffic Data Set.

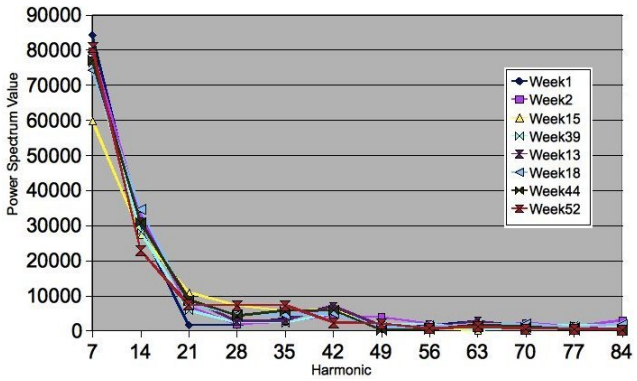

Fig. 5 Dominant Power Spectrums Values $A_7, A_{14}, \ldots, A_{84}$ of the Data's sampled weeks.

ues of Harmonics 7, 14, 21, 28, and 35. Most weeks already had low values in harmonics 21, 28, 35 and increase at the 42nd harmonic. The succeeding harmonics already show values converging to zero. Thus, it is reasonable to use harmonics $7, 14, \ldots, 42$ and the offset as candidate components of the data signatures of the weeks in the NLEX data set for visualization purposes.

With a closer look at the candidate harmonics in **Fig. 5** using a set of sampled weeks from different clusters of the data (inclusive of outliers), it can be seen that the most variation of the Power Spectrum values are in A_7, A_{14}, and A_{21}. Thus it is also interesting to obtain a visualization of the data set using the data signature composed of the Power Spectrum components A_0, A_7, A_{14}, and A_{21}. Finally, note that Week 15 has its first harmonic to be dominant. A relatively large and peculiar set values for the first few harmonics of Week 44 can also be seen in Fig. 4 and Fig. 5. By taking advantage of Weeks 15 and 44 nonconforming dominant harmonics, we can choose to use components of factor 7 as data signature, thus highlighting them as apparent outliers in the data set.

4. Results and Discussion

Implementing the algorithm on the benchmark and test models M_1 and M_2 using varied data signatures of each data point in D, the following values for Model Consistency (MC) are obtained as shown in **Table 1**. The MC of VF(85,168), which is the previously-known optimal model Ref. [3], is clearly defeated by

Table 1 Model Consistency (MC) of M_2 with M_1 for the data sets.

Model	2006	2007	2008	2009
VF(85,{0,7,14,21})	293.00	109.00	303.00	136.00
VF(85,{0,7,14,21,28,35,42})	207.00	77.00	177.00	100.00
VF(85,43)	93.00	37.00	73.00	65.00
VF(85,168)	100.00	60.00	87.00	84.00
VF(85,85)	102.00	46.00	69.00	55.00

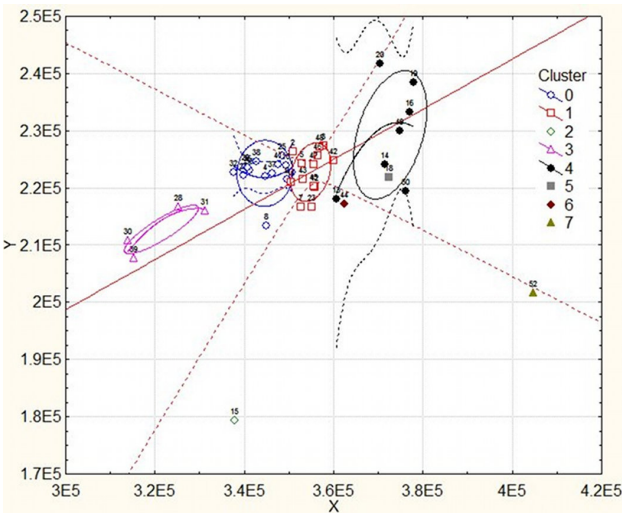

Fig. 6 VF(85,{0,7,14,21}) for 2006 NLEX BLK-NB data set.

the MC of proposed 4-dimensional data signatures constructed using the Power Spectrum components {0, 7, 14, 21} for all the data sets tested.

For illustrative purposes, we show results and perform analysis on the 2006 NLEX BLK-NB data set using the proposed algorithm and the candidate data signatures for VF visualizations. This can be replicated on other data sets.

Figure 6 shows the VF visualization VF(85,{0,7,14,21}) of the 2006 data set. To support the quantitative results, we check whether VF(85,{0,7,14,21}) is better compared with VF(85,168) in the qualitative criteria [3] using this data.

In terms of Criterion 1, VF(85,168) slightly surpasses the quality of VF(85,{0,7,14,21}) by an outlier Week 18 (i.e., Cluster 5). Nevertheless, both visualizations had minimal overlaps in some of their clusters such as Clusters 1 and 0. In general, closeness of co-members is maintained for both VF models. Both models also satisfy Criterion 2. In terms of Criterion 3, known outliers, i.e., Weeks 15, 51, and 44 are easily seen in both visualizations. It is notable that both Weeks 44 and 18 seem to have similarities with Cluster 4 in both VF visualizations. However, VF(85,168) had clearly projected Week 18 far enough from this cluster as compared to VF(85,{0,7,14,21}). The latter had actually placed this outlier within Cluster 4's region while the former projected it along the edge of the confidence ellipse of the cluster. In the detection of potential outliers, we had to refer back to the time domain iDIRBrG visualization in Fig. 1 to obtain a set of this points by intra-cluster analysis which we identified as Weeks 1, 30, 31, 43, and 48. It should be noted that a previous result in Ref. [2] has shown that Week 30 has the set of the smallest traffic volume values for year 2006. By using definitions of potential outliers and categories thereof as defined in Ref. [6], it can be observed that VF(85,{0,7,14,21}) slightly outperforms VF(85,168) by the

former's capability of detecting Week 31 as a potential outlier. However, for both models, Week 43 has not been detected as this type. As for the fifth criterion, it can be seen that both models clearly project clusters from the leftmost to the rightmost parts of the visualization in terms of ascending magnitudes of traffic volume values. In summary, for the first five qualitative goodness-of-representation, the two VF visualization models clearly have a competitive quality in terms of representing the relationships of the data signatures in their original m-dimensional space in the transformed 2-dimensional ($SBPx$, $SBPy$) space.

5. Conclusions

In this paper, we were able to obtain an optimal data signature that is more effective in representing points in the data set for VF visualization compared with using the entire set of Power Spectrum components. We added another qualitative criterion, i.e., Criterion 6, to further check the goodness of the data set visualization. An algorithm was formulated to check how each proposed vector-fused data signature visualization performs with respect to this criterion. Different data signature constructions were formulated and checked using the algorithm with results showing that the Power Spectrum components A_0, A_7, A_{14}, and A_{21} provide the best VF visualization quantitative value among the evaluated models in this paper for the 2006–2009 NLEX BLK-NB traffic volume data sets. By performing this analysis on multiple data, it was shown that the 4-dimensional data signatures provide robust representations for data visualization. Further validation of this model has also shown its competitiveness when compared with the previously-known optimal model Ref. [3] VF(85,168) in terms of the first five criterions. Thus, the quantitative analysis-based algorithm had shown that 4-dimensional models can be used to represent high dimensional data sets without sacrificing the amount of information that can derived from the visualizations. This representation accounts for a total of ≈97.6% dimensionality reduction with even better results in visualizing the data set based on the proposed algorithm. This reduction is crucial when analysts apply any additional exploratory data mining techniques such as clustering. Finally, with reliable, yet simple, 2-dimensional visuals produced by use of VF and data signatures, analysts are now capable of pinpointing weeks that may exhibit "interestingness" due to their spatial distance from all other points (or co-members) in the visualization.

Acknowledgments The authors would like to thank Roberto V. Bontia, VP-Toll Operations of MNTC, and Nicolas Manalo, VP Traffic Operations, of Tollways Management Corporation for providing the research team the NLEX data set. E.R. Oquendo would like to thank the University of the Philippines (UP) Visayas for his Fellowship Grant. J. Clemente, R.A. Juayong, and R.M. Tadlas are assisted by ERDT in their graduate studies. J.B. Lee would like to thank the UP Information Technology Training Center for his graduate studies scholarship. J. Malinao and H. Adorna are partially supported by DOST-PCIEERD through the ERDT project entitled *Information Visualization via Data Signatures*.

References

[1] Wong, P.C. et al.: Data Signatures and Visualization of Scientific Data Sets, *IEEE Computer Graphics and Applications*, Vol.20, No.2, pp.12–15 (2000).
[2] Malinao, J.A. et al.: Data Signatures for Traffic Data Analysis, *National Conference on Information Technology Education*, Capitol University (2009).
[3] Malinao, J.A. et al.: Patterns and Outlier Analysis of Traffic Flow using Data Signatures via BC Method and Vector Fusion Visualization, *Proc. 3rd International Conference on Human-centric Computing* (2010).
[4] Pelleg, D. and Moore, A.: X-means: Extending K-means with efficient Estimation of the Number of Clusters, *Proc. 17th International Conference on Machine Learning* (2000).
[5] Becerral, J.G. et al.: Traffic Data Analysis and Visualization via BC Method, *Philippine Computing Journal* (2010).
[6] Oquendo, E.R.F. et al.: Characterizing Classes of Potential Outliers through Traffic Data Set Data Signature 2D nMDS Projection, *Philippine Information Technology Journal*, Vol.4, No.1, pp.37–42 (2011).
[7] Johnson, R.: Visualization of Multidimensional Data with Vector-fusion, *IEEE Trans.*, pp.298–302 (2000).
[8] Xue, X. and Henderson, T.: Feature fusion for basic behavior unit segmentation from video sequences, *Robotics and Autonomous Systems*, Vol.57, Issue 3, pp.239–248 (2009).

Jasmine A. Malinao is an Assistant Professor in the Department of Computer Science, College of Engineering, UP Diliman. Her interests include Data Signatures, Visualization, Algorithmics, Design and Implementations.

Richelle Ann B. Juayong

Rona May U. Tadlas

Jhoirene B. Clemente

Erlo Robert F. Oquendo

John Boaz Lee

Ma. Sheilah Gaabucayan-Napalang

Jose Regin F. Regidor

Henry N. Adorna is an Associate Professor in the Department of Computer Science, College of Engineering, UP Diliman. He heads the Algorithms and Complexity Laboratory of the Department. He is the Project leader of the DOST-PCIEERD-funded ERDT research project "Information Visualization via Data Signatures," 2009–2011. His interests are in the Mathematical Foundations of Computer Science, in particular, Automata and Formal Language Theory, Discrete Mathematics and Algorithms for Hard Problems.