# Scatter-PCA for Visual Clustering of Spatio-Temporal Data

**Aina Musdholifah and Siti Zaiton Mohd Hashim**,

Soft Computing Research Group, Faculty of Computer Science and Information Systems,
Universiti Teknologi Malaysia, Johor, Malaysia

**Summary**
In order to verify the clustering results, domain experts usually require it to be represented in interpretable and meaningful ways. For that reason, the spatio-temporal clusters obtained are essentially needed to be visualized in an understandable view to support visual exploration of cluster structures. Scatter-Principal Component Analysis (Scatter-PCA) is proposed to visualize the spatio-temporal clustering result. Scatter-PCA combines PCA that projected *m*-dimensional spatio-temporal data into 2 dimensional spatio-temporal data with scatter plot to visualize the structure of clusters. Two spatio-temporal data: crime data and traffic accident data are utilized to validate the visual clustering approach. The experimental results on two clustering result of spatio-temporal data demonstrate the effectiveness of our visual clustering approach to investigate the structure of clusters.

***Key words:***
*Cluster; visualization; interpretation; principal component analysis; scatter plot.*

## 1. Introduction

The extraction of cluster label term is particularly stressed. However, using multiscale representation or visualization can help to accommodate the difficulties of determination of cluster labels [1]. It is easy to determine the cluster labels when the number of clusters is small, e.g. “low crime” and “high crime”. Otherwise, for a large number of clusters, labels will correspond to much more specific areas of investigation in the problem domain.

Visualization involves the use of visual or graphics techniques to represent information, data or knowledge. Visual representation can help the user to get a better understanding of content of dataset, since the human visual system is more inclined to process visual rather than textual information [2]. Visual data mining synergistically combine the computer processing capabilities with the unique and great human capabilities, to gain knowledge on the considered phenomena [3].

At the computational end, methods typically exploit the computational power and the formalisms of statistical inference to search for patterns. Although computational methods can search large volumes of data for a specific type of pattern very quickly, they have very limited pattern interpretation ability. In contrast, visualization methods can help analysts to visually pick out complex patterns, propose explanations and generate hypotheses for further analysis, and present patterns in an easy-to-understand form [4].

The spatio-temporal clusters obtained are also needed to be visualized in an understandable view to support visual exploration of cluster structures. In spatio-temporal data mining the effective analysis of results is crucial [2]. Visualization techniques are fundamental in the support of this task. According to Jain and Dubes (1988), clustering as a tool for exploring data, should be supplemented by visualization techniques to interpret the clustering results. There are many well known visualization techniques in the domain of information visualization, such as scatter plot [5], parallel coordinates plot (Inselberg and Dimsdale, 1990 and Zhou *et al*., 2008), and cartographic map [1].

Three categories of visualization techniques are based on the type of attributes [6]: visualization of data with small number of attribute, visualization of spatio-temporal data, and visualization of data with many attributes. However, this paper proposed scatter–PCA to investigate the structure of clusters obtained by 2-dimensional clustering view. Scatter-PCA is selected since it can visualize the data with small number of attributes, i.e. 2 or 3 attributes.

The rest of this paper; section 2 provides a detailed methodologies used for visualize clustering results including scatter plot, principal component analysis (PCA), and scatter-PCA; section 3 describes the experiments conducted and results obtained; section 4 gives the conclusion and analysis.

## 2. Cluster Visualization Approaches

### 2.1 Scatter plot

Scatter plot (Cleveland and McGill, 1988) is a visualization method that is commonly applied in data visual analysis in two dimensional spaces. It could be used to determine correlation between two variables of multidimensional dataset, initialize discovering clusters

Manuscript received January 5, 2014
Manuscript revised January 20, 2014

and interpret clustering results structure. Scatter plot draws a point that represents two datum of an individual of a dataset.

Cartesian coordinate system is used and defined by two perpendicular axes, resulting in scattering of points. Each object of data is plotted as point in the plane using the value of two attributes, *x* and *y* coordinates. The corresponding dimension or attributes values are represented by the positions of the data points, while correlation between two dimensions is represented in one single scatter plot. If the dataset has more than two dimensions, it requires another scatter plot that is different in color, size or shape of plotting points.

## 2.2 Principal Component Analysis (PCA)

The visualization techniques described in post-processing stage of clustering process is helpful when applied not to the raw data but to the derived scores seeking to summarize the data in some optimal way. There are a number of possible methods of obtaining the respective scores, such as *factor analysis*, *multidimensional scaling*, *canonical analysis*, and *Principal component analysis* (PCA) [7]. However, previous works show the successful application of PCA [8] ,[9] ,[10] and [11].

However, this study used Principal Component Analysis (PCA) as unsupervised data reduction that initially was described by [12] to visualize the clustering results. PCA provide ways to reduce complex data set to a lower dimension which can simplified structure that lie beneath it [13]. Thus, main objective of PCA is to reduce dimensionality of the data with the minimum of loss of information, by retaining as much as possible the variation in the original data set. The dimension reduction is performed by calculating the matrix decomposition of the data and followed by finding principal component of the data that is considered as score seeking of the data summary.

Therefore, this research used matrix decomposition to visualize and interpret clustering result by forcing representations of the data in terms of a small number of substructures. Matrix decompositions use the relationships among large amounts of data and the probable relationships between the components to separate the raw data into the component that underlie them [14]. Furthermore, the definition of matrix decomposition is described below.

Consider a dataset as a matrix, with *n* rows, each of which represents an object, and *m* columns, each of which represents an attribute. The $ij^{th}$ entry of a dataset matrix is the value of attribute *j* for object *i*. Each family of matrix decompositions is a way of expressing a dataset matrix, *A*, as the product of a set of new matrices. More formally, a decomposition matrix can be described by an equation of this form

$$A = C\,W\,F \quad (1)$$

where the size of the matrices are as follows: *A* is *n* x *m* (and we assume for simplicity that $n > m$; in practice *n* much larger than *m*; $n >> m$); *C* is *n* x *r* for some *r* that is usually smaller than *m*; *W* is *r* x *r*, and *F* is *r* x *m*.

There is two matrix decomposition to calculate PCA, the first one is using singular value decomposition, SVD [14]. Another is using eigenvalues and eigenvectors of covariance matrix describing the data set. Both methods yield the same information, but take different routes from a computational standpoint. However, this research used SVD matrix decomposition since it is more numerically robust approach [10].

## 2.3. Scatter-PCA

In this research, combination of PCA with SVD matrix decomposition and scatter plot is performed using these steps below, also shown in Fig.1:

1. Organize a data set as an *m* x *n* matrix, where *m* is the number of attributes (or measurement) and *n* is the number of data.
2. Subtract the mean for each attribute of object $x_i$.
   To avoid the data from domination of certain features, PCA approach use normalization process. This approach starts with Z-score data normalization. The objectives of the normalization process are to reduce the square mean error of approximating the input data by data centering and to get unit variance by standardizing the variables (or data scaling). Using Z-score, an attribute value *V* of an attribute *A* is normalized to *V'* and defined as:
   $$V' = \frac{(V - mean(A))}{std(A)} \quad (1)$$
3. Calculate the SVD of the covariance to get principal component of the data (PCs).
   SVD method of PCA is applied to the normalized dataset to get PC. Applying PCA to the result of step gives the number of PCs obtained is same as the number of original variables.
4. Eliminate the unnecessary PCs.
   To remove the weaker components from this PC set, the variance of PC values and the mean of the variance of PC are calculated. Then, the PCs having variances less than the mean variance is ignoring.
5. Find the reduced projected data
   In the result, the transformation matrix with reduced PCs is formed and this transformation matrix is applied to the normalized data set to produce the new reduced projected dataset.

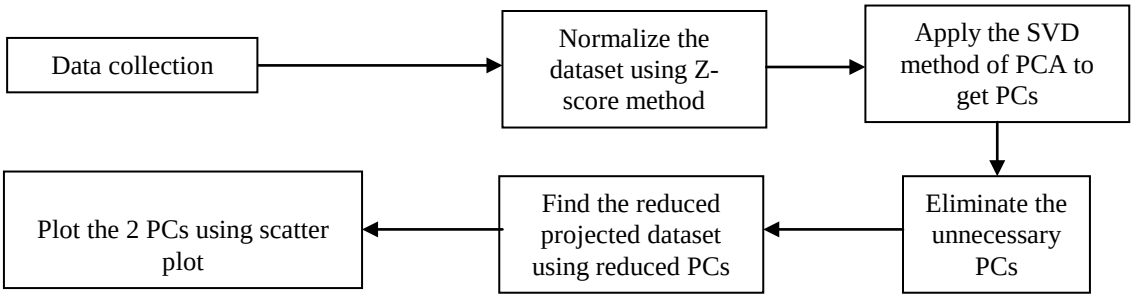


Fig. 1 Proposed Scatter-PCA.

6. Plot the 2 PCs using scatter plot
   From *m*-PCs selects only 2 PCs and plots using scatter plot with different color and symbol to represent different clusters.

## 3. Experimental Results

### 3.1 Visual Clustering on Spatio-Temporal Crime Data

One of the most important responsibilities of government is to protect its citizens from crime and unsafe situations. Spatio-temporal information analysis plays a central role in many security-related applications. The outputs of such analyses can provide useful information to guide the activities aimed at stopping, detecting, and responding to security issues.

Crime is a dynamic event that is stationary neither over space nor over time [15]. Spatio-temporal analysis is an important component of crime analysis since location and time are two critical aspects of most crime related events. The combination of spatial and temporal techniques allows establishing a typology of spatio-temporal characteristic of crime patterns. Finding these patterns helps to optimize effectiveness in the reduction of crime and increase the safety of resident. The clustering process can be used to discover these spatio-temporal patterns of crime offenses by determining clusters within the crime data.

In the first experiment on visualizing the clustering results of spatio-temporal data, crime incident is considered, taken from Pittsburgh for duration January 1990 and December 2001. The crime incidents being examined consist of all crime events recorded at police department and were assembled for a project related to the development of statistical tool, CrimeStat III [16]. While the crime incidents vary in type (24 types of crime recorded in Pittsburgh), the variation in space and time of total crime incidents may give indications as to possible differentials in security service provision or in environmental safety risk. However, this study focuses on burglary crime since most burglary crime is usually crime incident that are reported immediately, for claim purpose.

Table 1 represents the burglary crime data and furthermore the multidimensional spatio-temporal burglary array. The burglary crime data consists of 42 spaces and 144 time periods.

For this experiment TKNN-based clustering algorithm [17] is applied to discover clusters on burglary crime data. However, this paper does not discuss clustering step in details, and focuses only on visualization stage.

The structure of clustering result of burglary crime data is visualized using scatter-PCA. PCA was applied to the correlation matrix of the burglary crime data and followed by scatter method. Fig 2 shows the scatter-PCA of clustering burglary crime data. From 42 beats that represent location of burglary, 39 beats are grouped in cluster 1 and remaining is grouped in cluster 2.

Table 1: The spatio-temporal burglary data

| Beat | Number of burglary | | | |
|---|---|---|---|---|
| | Jan-90 | Feb-90 | … | Dec-01 |
| Beat 11 | 78 | 65 | … | 76 |
| Beat 12 | 2 | 6 | … | 8 |
| . | . | . | … | . |
| . | . | . | … | . |
| . | . | . | … | . |
| Beat 42 | 9 | 6 | … | 20 |

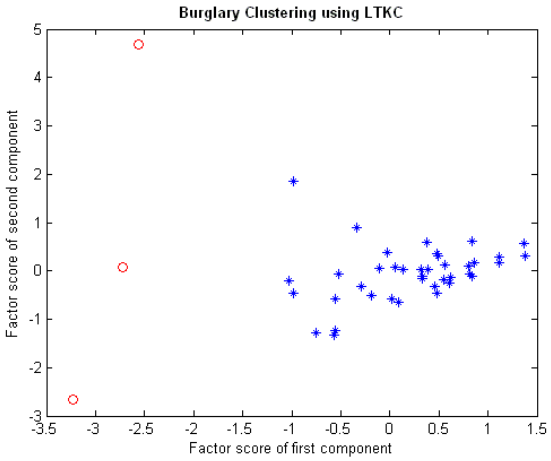


Fig. 2 Scatter-PCA of burglary crime clusters.

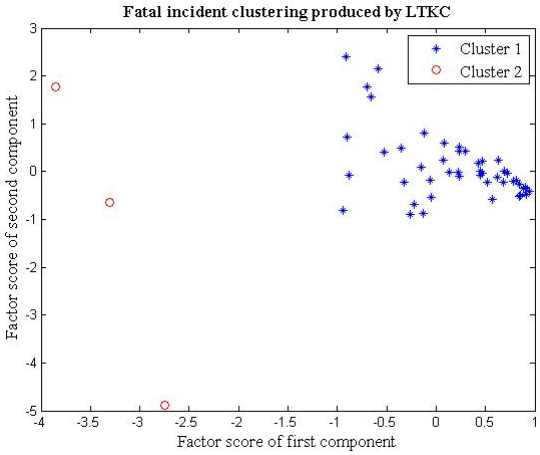


Fig. 3 Scatter-PCA of traffic accident clusters.

### 3.2 Visual Clustering on Spatio-Temporal Traffic Accident Data

Traffic accidents are an important concern of today's governments and societies, due to the high cost human and economical resourced involved. Beshah in [18] states that each year there is more than 1.2 million deaths and 50 injuries occurred in the world. Thus, analyzing accident reports collected from past accidents can be used further in reducing accident severity as well as attracting great interest to traffic agencies and the public at large.

Data mining has been proven giving significant help in improving traffic safety. Among the data mining tasks, the clustering algorithm and visual clustering are mostly applied on spatio-temporal datasets, especially for the traffic dataset [19].

The traffic accident used in this study is fatal crash dataset, which is provided by downloaded Fatality Analysis Reporting Systems, FARS [20]. The dataset recorded vehicle crashes on public roadway of United States and occurred during January 1994 – December 2008. Table 2 represents the fatal accident data. The fatal accident data consists of 51 states and 180 time periods.

Table 2: The spatio-temporal fatal accident data

| States | Number of fatal accidents | | | |
|---|---|---|---|---|
| | Jan-94 | Feb-94 | … | Dec-08 |
| Alabama | 78 | 65 | … | 72 |
| Alaska | 2 | 6 | … | 6 |
| . | . | . | … | . |
| . | . | . | … | . |
| . | . | . | … | . |
| Wyoming | 9 | 6 | … | 13 |

From experimental results on fatal accident data, TKNN-based clustering [17] has produced best clustering with two clusters. To observe the distribution of data in each cluster in the best clustering, scatter-PCA is performed and result is shown in Fig. 3. Three of 51 states are clustered in cluster 2 and remaining is in cluster 1. [17]

## 4. Conclusions and Future Works

In this research, a visualization technique, scatter-PCA suitable for providing informative graphical displays of multivariate spatio-temporal data have been explored and performed. Scatter-PCA with principal component scores provides two-dimensional views of the data. Scatter-PCA has been proven to visualize the clustering result of spatio-temporal data in scatter plot that compare characteristics of all clusters of all attributes within the data. In addition, PCA can be used to examine the compactness and separateness of clusters.

However, it is required to do relative investigation of objects set or clusters, where individual objects are described by value of some attributes, and hence, a set of objects may be characterized by a distribution of attributes values and value combination. In addition, it is necessary to visualize the location of cluster member on the map for investigating the spatial relationship in a cluster. Thus, future works will focus on other visualization techniques that can visualize all data dimensions simultaneously and represent spatial distribution of clustering results in cartographic representation of geographic map.

**Acknowledgments**

This work is supported by a research grant from Universiti Teknologi Malaysia (UTM) VOT number

QJ.130000.7128.01H12. The authors gratefully acknowledge many helpful comments by reviewers and members of Soft Computing Research Group (SCRG) UTM Malaysia in improving the publication.

## References


1. Skupin, A., *The world of geography: Visualizing a knowledge domain with cartographic means.* 2004, PNAS. p. 5274-5278.
2. Kechadi, M.-T., et al., *Mining spatio-temporal datasets: relevance, chalenge and current research directions* in *Data mining and knowledge discovery in real life applications*, J. Ponce and A. Karahoca, Editors. 2009, I-Tech: Vienna, Austria. p. 438.
3. Andrienko, N., G. Andrienko, and P. Gatalsky, *Exploratory spatio-temporal visualization: An analytical review.* Journal of Visual Languages and Computing, 2003. **14**(6): p. 503-541.
4. Guo, D., et al., *Multivariate Analysis and Geovisualization with an Integrated Geographic Knowledge Discovery Approach.* Carthographic and Geographic INformation Science, 2005. **32**(2): p. 113-132.
5. Yuan, X., et al., *Scattering Points in Parallel Coordinates.* IEEE Transactions on Visualization and Computer Graphics, 2009. **15**(6): p. 1001-1008.
6. Tan, P.N., M. Steinbach, and V. Kumar, *Introduction to Data Mining*. 2006: Addison Wesley.
7. Everitt, B.S., *Cluster Analysis*. 3rd ed. 2000.
8. Dash, R., et al., *A hybridized K-means clustering approach for high dimensional dataset.* Interantional Journal of Engineering Science and Technology, 2010. **2**(2): p. 59-66.
9. Gorban, A.N. and A.Y. Zinovyev, *PCA and K-Means Decipher Genome Principal Manifolds for Data Visualization and Dimension Reduction*, A.N. Gorban, et al., Editors. 2008, Springer Berlin Heidelberg. p. 309-323.
10. Hoffman, F.M., et al. *Multivariate Spatio-Temporal Clustering (MSTC) as a Data Mining Tool for Environmental Applications*. in *iEMSs 2008:International Congress on Environmental Modelling and Software Integrating Sciences and Information Technology for Environmental Assessment and Decision Making*. 2008: International Environmental Modelling and Software Society (iEMSs).
11. Xu, H. and A. Ma 'ayan, *Visualization of Patient Samples by Dimensionality Reduction of Genome-Wide Measurements Information Quality in e-Health*, A. Holzinger and K.-M. Simonic, Editors. 2011, Springer Berlin / Heidelberg. p. 15-22.
12. Pearson, K., *LIII. On lines and planes of closest fit to systems of points in space.* Philosophical Magazine Series 6, 1901. **2**(11): p. 559-572.
13. Martinez, W.L., A.R. Martinez, and J.L. Solka, *Exploratory Data Analysis With MATLAB.* 2010: CRC Press.
14. Skillicorn, D., *Understanding complex datasets: data mining with matrix decompositions.* Data Mining and Knowledge Discovery Series, ed. V. Kumar. 2007, Boca Rotan, FL: Chapman & Hall/CRC.
15. Kumar, M.V. and C. Chandrasekar, *Spatial-Temporal Analysis of residential Burglary Repeat Victimization: Case study of Chennai City of Promoters Apartments, India*. International Journal of Research and Reviews in Computing Engineering, 2011. **Vol. 1**(No. 3): p. 101-111.
16. Levine, N., *CrimeStat III: A spatial statistics program for the analysis of crime incident locations (version 3.0)*. 2004., Ned Levine & Associates: Houston, TX/ National Institute of Justice: Washington, DC, .
17. Musdholifah, A. and S.Z. Mohd. Hashim. *Triangular kernel nearest neighbor based clustering for pattern extraction in spatio-temporal database*. 2010.
18. Beshah, T. and S. Hill. *Mining road traffic accident data to improve safety: Role of road-related factors on accident severity in Ethiopia*. 2010.
19. Yin, J., et al. *High-dimensional shared nearest neighbor clustering algorithm*. 2005. Changsha.
20. NCSA, N.C.f.S.a.A. *Fatality analysis reporting system (FARS) web-based encyclopedia*. 2004; Available from: http://www-fars.nhtsa.dot.gov/.


**Aina Musdholifah** received her master degree in computer science from Gadjah Mada University, Indonesia in 2003. Currently, she is a PhD student at Department of Software engineering, faculty of computer science and information system, Universiti Teknologi Malaysia in Malaysia. Her research interest includes data mining in large database and soft computing.

**Siti Zaiton Mohd Hashim** received her master degree in computer science from University of Bradford, UK in 1996. She graduated Ph.D study in soft computing from The University of Sheffield, UK in 2005. Her research interest includes soft computing, data mining and fuzzy system. Currently, she is a deputy dean of academic in faculty of computer science and information system, Universiti Teknologi Malaysia