Agricultural Information Research 16(2), 2007. 44–51
Available online at www.jstage.jst.go.jp/

Original Paper

Principal Component Analysis and Self-Organizing Map for Visualizing and Classifying Fire Risks in Forest Regions

Suwardi Annas*, Takenori Kanai and Shuhei Koyama

Osaka Prefecture University, Gakuen cho 1-1, Nakaku Sakai, Osaka 599-8531, Japan

Abstract

Dataset compiled from spreading hot spots, responsible for fire risk in many regions of Indonesian forests, are complex, primarily induced by the large size of the observed regions and high variation of hot spot distribution. The challenge in analyzing this type of dataset is to develop statistical techniques that facilitate the analysis, visualization, and interpretation of the results. Techniques, such as multivariate analysis and artificial neural networks, have been applied to resolve the high-dimensional space in such large datasets. Each method uses a different rationale for how the relationship between the input parameters will be preserved during analysis. This study presents the use of a principal component analysis (PCA) and a self-organizing map (SOM) to reduce the high dimensionality of the input variables and, subsequently to visualize the dataset into a two-dimensional (2-D) space. The results indicate that the first two principal components of the PCA provide a large percentage of cumulative variance to explain the data patterns. However, a comparison of the data projection, SOM is better suited than PCA in visualizing the fire-risk distribution in forests. The SOM color-coding and labeling also effectively visualized a classification system of fire risk via node clusters, in such a way that the fire risks level according to their hot spot locations in forest is easily interpreted.

Key words

dimensionality reduction, fire risk distribution, hot spot, noise pattern

Introduction

Forest fires in Indonesia continue to increase in both frequency and size, damaging forest resources and adversely affecting living conditions across Southeast Asia. The extended fire risk in forests is primarily caused by hot spots that occur in many regions with high temperatures. As reported by AFP (2002) that large number and cluster of hot spots that persist over time are good indicators of fire problems. Hence, the collections of the hot spots data that are potential in emerging fire risk in Indonesian forests are central to the present analysis. The sample data was observed from the hot spot occurrences in many forest regions all year round in which their frequencies, intensities, and widths vary. The structure of data that compiled based on models of hot spot is available with large space features and noise distribution patterns. The challenge in analyzing these datasets is to develop analysis techniques to facilitate a solution to the problem.

Visualisation and classification methods have become standard tools to handle the complexity of the data because they enable a ready representation and interpretation. When the number of dimensions is large, a multivariate analysis technique such as PCA can be used to reduce the dimensions before subjecting the output factors to a clustering routine (Kiang *et al*., 2004). The present study utilizes PCA to extract the size-dimension information and to construct a linear projection of the dataset in the 2-D plane of the map. PCA is a multivariate, statistical data analysis that can be used for processing and visualizing data (Tipping and Bishop, 1999). Although the PCA is a powerful technique for extracting data (Kwan *et al*., 2001), sometime its visualization was not suitable to represent the complex structure of datasets (Laitinen *et al*., 2002). In this condition, a SOM algorithm is an alternative method for the optimal visualization and clustering of datasets. It is often promised as an effective tool for exploratory analysis of data (Koua, 2003).

The SOM is a nonlinear statistical technique for transforming

* Corresponding Author
E-mail: suwardi@envi.osakafu-u.ac.jp

and visualizing multi-dimensional data in a lower-dimensional space (Kohonen, 1998; Himberg, 2000; Mancuso, 2001). There are many applications that have implemented the use of SOM technique for exploratory analysis of data. The SOM was designed for solving problems that involved clustering and visualization (Flexer, 2001; Kiang, 2001). SOM clustering with color-coding is a way to group data, according to its properties (Kaski and Kohonen, 1998; Kaski, 2001). The SOM method also has advantages in the classification of satellite images data. For example, two-stage SOM was successfully applied to clustering of weather satellite cloud images (Honda and Konishi, 2001) and a multiple SOM was quite suitable for remote sensing classification under various data and simple-design conditions (Wan and Fraser, 2000). The robustness of those SOM applications to handle the data visualization and classification that motivates to utilize the SOM approach for exploring this analysis data problem.

The objectives of study are focused to implement both PCA and SOM methods to provide insights on the data extraction and visualization. First, PCA is utilized for extracting the high dimensionality of the input variables and project the dataset onto a 2-D space. Second, SOM algorithm is used for data visualization and, subsequently, to create a classification system of fire risks via the node clusters on the SOM map.

Data description

This study analyzed dataset for hot spots that are responsible for fire risk in the forest regions of Sumatera and Kalimantan, where forest fires occurred frequently throughout 2000–2003 in Indonesia. The sources of the hot spots data were originally detected by Advanced Very High Resolution Radiometer (AVHRR) satellite imagery from the ASEAN Specialized Meteorological Centre (ASMC) in Singapore. Here, hot spots are defined as image pixels whose brightness temperature exceeds a pre-defined threshold value. A temperature threshold was adopted by ASMC for hot spot detection, related to a brightness temperature of 321.3 K (ASMC, 2005).

The Indonesian State Ministry of Environment (ISME) Bureau as part of forest fire monitoring in Indonesia has developed a hot spot database by calculating the frequency of spreading hot spot occurrences in forest regions, according to their coordinates location from ASMC satellite imagery, with GIS software (Arc View). As reported by Abberger *et al*. (2002) that hot spot data might be useful as early fire detection information if they are translated into descriptions of estimated fire locations. This indicates that the existence of hot spots in a forest region is potential in emerging fire risk according to their location in the forest regions.

In this study, the dataset that acquired from the ISME Bureau are compiled based on the frequency of hot spot occurrences in a region of forest. The observed regions of sample data consist of 72 regions in Sumateran forest and 38 regions in Kalimantan forest. Details of the input variables consist of the 11-dimensional spaces based on the average of hot spots data during four years from 2000 to 2003 on a monthly basis for the period January-November. Note that the December hot spot data is not used for data analysis because it is not available of the year 2001. The resulting large dataset with heterogeneity of the distribution patterns make the analysis and interpretation a difficult task. This study, therefore, discusses the methods used to analyze the dataset in the following sections.

Analysis methods

Feature extraction using PCA

PCA is a way for extracting the data features by reducing the number of dimensions, without much loss of information. In this case study we applied the PCA to identify the fire risk patterns in hot spot data for many regions of forest.

In the PCA process we compiled the hot spot data into a matrix $\mathbf{A}={}_n\mathbf{X}_m$, where the n-rows are associated with the number of observed regions in relevant hot spot events and the m-columns with the number of input variables (months). And let $\bar{\mathbf{x}}_k$ is the mean of m variables in the matrix **A**, the covariance matrix is given by

$$\Phi_{jk}=\frac{1}{n}\sum_{i=1}^{n}(x_{ji}-\bar{x}_j)(x_{ik}-\bar{x}_k), j\neq k \qquad (1)$$

where j=1, 2, ..., n, k=1, 2, ..., m. There are two main steps in order to find a few orthogonal features, called principal components (PCs) of the matrix **A**, as follows:

Step 1. Calculate the eigenvectors and eigenvalues of the covariance matrix. Since the covariance matrix from equation (1) is square, a set eigenvalue (λ_1, λ_2, ..., λ_m) can be found by solving the determinant equation $|\Phi-\lambda I|=0$. A set nonzero eigenvector $\mathbf{E}=[\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_m]$, corresponding to the relevant eigenvalue is found by using $(\Phi-\lambda_i I)\mathbf{e}_i=0$, i=1, 2, ..., m. Then, a diagonal nonzero eigenvalues matrix (Λ) can be constructed from the sorted eigenvalues $\lambda_1\geq\lambda_2\geq...\geq\lambda_m$.

Step 2. Find the PCs of the covariance matrix Φ that can be generated using a process of singular value decomposition, which is given by

$$\Phi=\mathbf{E}\Lambda\,\mathbf{E}^T \qquad (2)$$

The set of PCs is then represented as a linear combination of the original variables of $PC_m=\mathbf{e}_m^T\mathbf{x}$.

In order to interpret the level of fire risk for different months,

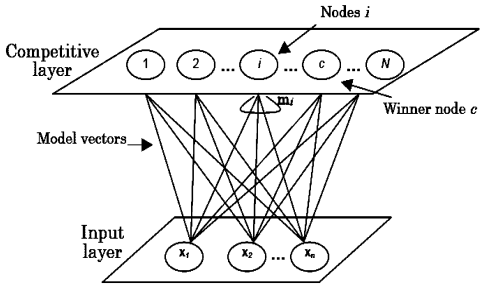

Fig. 1 The network structure of the input and competitive layers of the SOM.

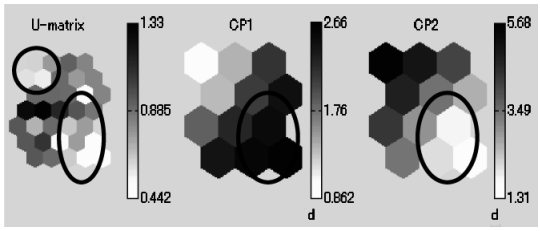

Fig. 2 U-matrix and CP generation example for a 4×3 hexagonal SOM. Two circles in the U-matrix display data clusters via nodes. Circled nodes in CP1 represent high data values, but low in CP2.

the coefficient values of the PCs could be used. Further, the PCs projection would be developed for expressing the fire risk patterns according to regions of hot spot events. The first two PCs are usually used to project a 2-D plane of data, if they provide at least 80% of cumulative variance (Johnson and Wichren, 1998). The percentage of variance (PV) for the PCs can be calculated by

$$\mathrm{PV}=\frac{\lambda_k}{\lambda_1+\lambda_2+\ldots+\lambda_m}\times 100\% \quad (3)$$

Self-organizing map

The SOM also used the same data matrix corresponds to hot spot data that was analyzed in the PCA as input data. The SOM network is divided into an input layer, containing a set of observation vectors $\mathbf{x}_i=[x_1, x_2, \ldots, x_n]^T\in\Re^n$, and a competitive array layer of nodes, as illustrated in Fig. 1. Each node i in the competitive layer is then associated with all data vectors in the input layer by the model's vector $\mathbf{m}_i=[m_{i1}, m_{i2}, \ldots, m_{in}]\in\Re^n$, ($i$=1, 2, ..., N)

The connection between the two layers represents a map of real high-dimensional data onto a low-dimensional (usually 2-D) display of nodes. In the training process, the best-matching node (winning node) is found using the criterion of greater similarity,

$$\|\mathbf{x}-\mathbf{m}_c\|=\min_i\{\|\mathbf{x}-\mathbf{m}_i\|\} \quad (4)$$

The models of the winning node are then updated in accordance with the rule,

$$\mathbf{m}_i(t+1)=\mathbf{m}_i(t)+h_{ci}(t)[\mathbf{x}(t)-\mathbf{m}_i(t)] \quad (5)$$

where t denotes the index of the iteration step, $\mathbf{x}(t)$ is the vector-valued input sample of $\mathbf{x}$ in the iteration t. Here, the $h_{ci}(t)$ is called the neighborhood function around the winning node c. During training, $h_{ci}(t)$ is a decreasing function of the distance between the i-th and c-th model on the map node. For convergence it is necessary that $h_{ci}(t)\to 0$ when $t\to\infty$. More detail of the SOM algorithm can be found in the Kohonen (2001).

In this study, for training the hot spot sample to this SOM algorithm, the MATLAB software with SOM toolbox was utilized. The most appealing features of the SOM toolbox are that the source code can be modified during analysis.

Visualization of data by SOM

Two types of the SOM visualization were implemented such as a Unified distance matrix (U-matrix) and Component Planes (CP), as shown in Fig. 2. The hexagonal grid was used to display the network nodes on the SOM, because the inter-neural distance of the hexagonal more coincides with the Euclidean metric distance than rectangular (Starikov, 2000).

First, U-matrix represents the distance between neighboring nodes on the SOM map. With the SOM algorithm the distance measure is calculated between the vector of the node weight and its neighbors. U-matrix has more hexagons than CP because it shows not only the distance values at the map nodes but also the distance between map nodes. The distance values were then used to define the nodes with different grayscale of color shades. The high values on the U-matrix mean large distance (darker shades) between neighboring nodes, and thus indicate a frontier region between clusters. The lighter shade of nodes represents the similarity values among nodes (cluster) on the SOM map.

Second, CP was used to visualize the distribution of data values for different variables via the nodes map. Color-bar (d) of CP corresponds to the reference vector pattern on the nodes. The reference vector of the node such as that shades of darker mean high value, gray moderate value, and shades of light low values. These color shades information were then used to clarify the level of fire risk for different variables on monthly basis. Further, a combination visualization of the U-matrix and CP enables to obtain both cluster structures and the correlations between the variables from the same picture (Fig. 2). It can be simultaneously visualized using the SOM visual inspection. This visualization can also be used to create an understanding of which variables are respect to the clusters.

Results and discussion

PCA extraction and projection

PCA has extracted the 11-dimensional space (i.e., months) as the input data variables of dataset using a covariance matrix. Table 1 gives a summary of the eigenvalues and the variances of data from the first five PCs. The eigenvalues illustrated that the

Table 1 Cumulative variance explained of the first five components

Component	Eigenvalues	Variance (%)	Cumulative (%)
1	46.02	63.71	63.71
2	15.18	21.01	84.72
3	5.388	7.458	92.18
4	2.627	3.636	95.81
5	0.903	1.249	97.06

Extraction method by PCA

first two PCs explain 84.72% cumulative variance of the dataset. The higher percentage of the variance indicated that these PCs are suitable used to explain the distribution pattern of the fire risk, according to hot spot details in forest regions for different months. Therefore, the original variables (months) can be weighted as a linear combination form into PC1 and PC2 as follow:

$$\text{PC1}=0.13\ (\text{Jan})+0.35\ (\text{Feb})+1.01\ (\text{Mar})+0.18\ (\text{Apr})+0.54\ (\text{May})+1.0\ (\text{Jun})+1.59\ (\text{Jul})+3.89\ (\text{Aug})+4.13\ (\text{Sep})+2.95\ (\text{Oct})+0.4\ (\text{Nov}) \quad (6)$$

$$\text{PC2}=0.23\ (\text{Jan})+0.82\ (\text{Feb})+2.32\ (\text{Mar})+0.59\ (\text{Apr})+1.3\ (\text{May})+0.99\ (\text{Jun})+2.06\ (\text{Jul})-0.32\ (\text{Aug})-1.17\ (\text{Sep})-0.57\ (\text{Oct})+0.1\ (\text{Nov}) \quad (7)$$

The absolute value of the coefficient could help to interpret the PCs. For example, based on the equation (6), PC1 indicated that the month of September, with the largest coefficient (4.13), receives the greatest of fire risk and, conversely, the month of January, with the smallest coefficient (0.13), receives the lowest of fire risk. Further, according to equation (7), PC2 indicated that the month of March, with the coefficient (2.32), receives larger of fire risk and lower of fire risk occurred for November, with coefficient (0.1).

The overlapping plot (Fig. 3) from both PCs between the forest regions and the months of hot spot occurrences represent that the PC1 axis (1st component) contained the maximum amount of the variance. The PC2 axis (2nd component) contained the maximum amount of the variance orthogonal to the first. As it can be observed that the regions of forest with the smallest variance of fire risk occurrences are closest to the center of the orthogonal axes. Conversely, the forest regions with the largest variance of fire risk are far away from the center.

PCA plot also shows the correlation pattern between samples (regions) relative to the certain input variable (months). For example, the regions with larger variance of fire risk occurrences are grouped at around the September plot. In contrary, the regions with smaller variance of fire risk are grouped at around of the January plot (near the center).

However, the present result illustrated that although PCA explained most the cumulative variance of data, its scatter plot is sometime difficult to interpret. In particular, many regions in the PCs scores from –1 to 1 coupled in the around center of the orthogonal axes, as shown inside of the circle (Fig. 3). This result supports the finding study by Brosse *et al*. (2001) that the drawback was afforded to data complex, in which being poorly represented on the PCA plane. Therefore, we proposed to use the SOM algorithm as an alternative method for exploring the dataset.

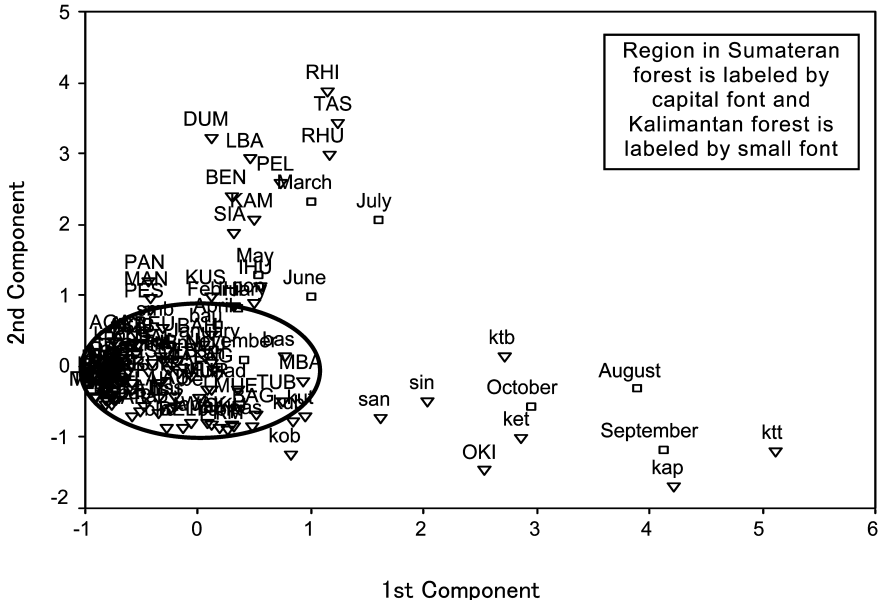

Fig. 3 The 2-D scatter plot of the first two PCs of data represents the distribution pattern of fire risk via regions relative to the month of hot spots event.

Table 2 The correlation coefficient values between the variables on monthly basis

Months	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov
Jan	1.00										
Feb	0.43	1.00									
Mar	0.44	0.87	1.00								
Apr	0.47	0.51	0.57	1.00							
May	0.50	0.72	0.74	0.76	1.00						
Jun	0.45	0.62	0.69	0.62	0.77	1.00					
Jul	0.47	0.58	0.67	0.68	0.84	0.77	1.00				
Aug	0.17	0.17	0.23	0.14	0.22	0.46	0.47	1.00			
Sep	0.10	0.15	0.16	0.00	0.11	0.43	0.32	0.81	1.00		
Oct	0.22	0.20	0.20	0.09	0.20	0.47	0.36	0.66	0.90	1.00	
Nov	0.40	0.31	0.26	0.49	0.40	0.47	0.38	0.43	0.48	0.53	1.00

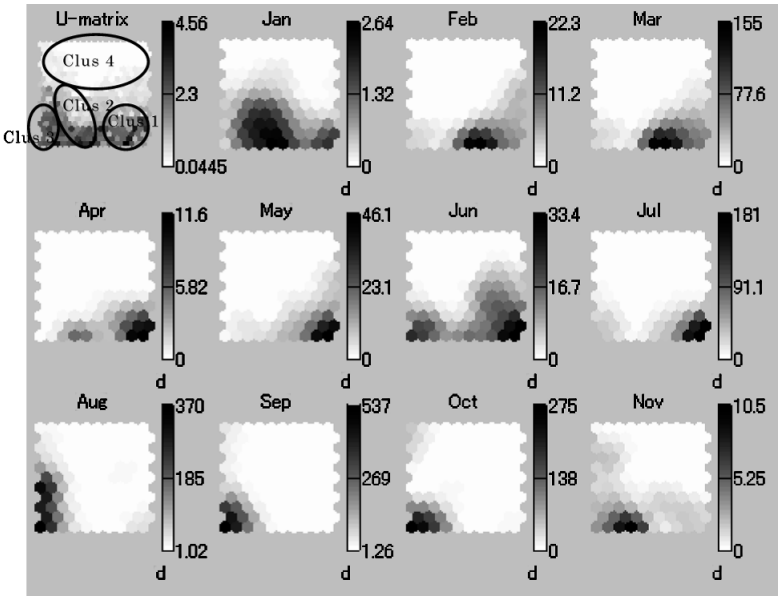

Fig. 4 Visualization of U-matrix and CP (i.e., months). U-matrix represents the data clusters, and CP represents the fire risk distribution patterns for each month.

SOM visual inspection

The SOM visual inspection combines the CP and U-matrix in the 2-D nodes map, as shown in Fig. 4. The maps are connected to adjacent hexagonal nodes with sizes 11×10, by adapting the 110 observed regions in the Sumateran and Kalimantan forests. There are no explicit rules for choosing the number of nodes (Hautaniemi *et al*., 2003), but one principle is that the size should allow easy detection of the structure of SOM (Wilppu, 1997). The CP is used for visualizing the different input variables. Here, each CP represents the fire risk that measured based on the average of hot spot values during four years (2000 to 2003) on a monthly basis. The level of fire risks of differing months can be studied from the density of color shades in the nodes network for each CP map. A darker shade corresponds to high fire risk, a gray shade that represents a moderate fire risk, and a lighter shade that corresponds to a low fire risk.

The visualization of CP also allows in inspecting the possible correlations between the variables of input data. By inspecting grayscale representation from the grid nodes, an even partial correlation of CP may be identified. For example, As we can be observed that the color shades of SOM nodes in Fig. 4 shows that fire risk, emerging from hot spots of the February have similar distribution to March, May similar to July, August similar to September, and September very similar to October. This fact is reflected by the analysis of the correlation of the data (Table 2). The four highest correlation coefficients are the September–October (0.90), followed February–March (0.87), May–July (0.84), and August–September (0.81). Whereas the month of January, April, June, and November give low correlation each other.

The top left of Fig. 4 provides a visualization of the U-matrix that represents the relative distances measure between the network nodes, marked by color shades. A large distance between

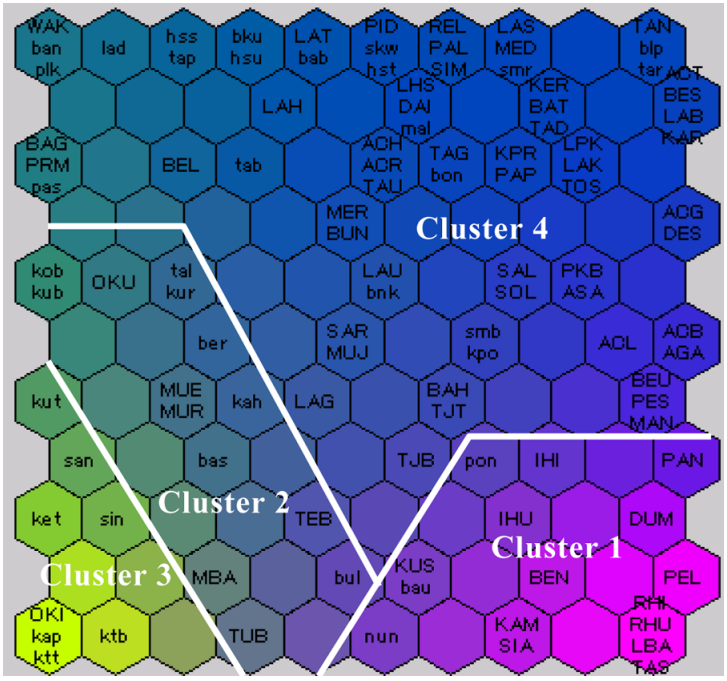

Fig. 5 SOM color-coding and distance measure represent the clustering of fire risk in relevant regions of hot spot events. Note that the regions of Sumateran forest are labeled by capital font and Kalimantan forest are labeled by small font.

adjacent nodes is shown as a darker shade and thus represents a border region between clusters. Areas similar to each other correspond to a lighter shade, which indicates that the values of the input data vectors are similar, and thus show a cluster. It can be seen, for example, in Fig. 4 that U-matrix provides cluster information that was represented with some circle on the map. However, the map of U-matrix indicates the situation where the distance measure is not reliable to show a representative data cluster. As reported by Kiang *et al*. (2004) that sometime it is difficult to visually group the output from SOM when the map is highly populated. In this case study, a difficult decision when only used the Euclidian distance to select the clusters in which may be incapable.

Selecting clusters in the SOM

To overcome the SOM deficiency in clustering data, as currently described via the U-matrix, a combination of the distance measure and the SOM color-coding are further used to visualize cluster of fire risk by region. The SOM color-coding is a way to group data, according to its properties (Kaski, 2001). For this purpose, the individual map of CP is reconstructed to visualize the clusters, as shown in Fig. 5. On the SOM map, the labels of the forest regions with similar hot spot values automatically have similar colors on the grid nodes and have a small distance measure to each other. Large distance measures on the map are also automatically assigned different colors and clusters.

To select a cluster, we first identify the group regions based on the discoloration of the nodes. When there are situations that the colors of nodes are unclear to indicate the cluster differences, the distance measures are then used to verify the clusters. Based on these criterions, we develop cluster membership for fire risk regions and found four clusters, as shown in Fig. 5. In this particularly data clustering, although both SOM methods have been used to select the clusters, some regions were difficult to assign to a certain cluster. In the clustering context, as reported by Hautaniemi *et al*. (2003) that all clustering algorithms including the SOM share a problem of deciding boundaries of the clusters, it is required further specification of the SOM clusters. However, the clearly labeling of regions on the SOM nodes would make easily to interpret of the results if compared than PCA plot.

The benefit of the SOM visualization that it can be used to detect the level of fire risks in the forest regions by inspecting the relation position between the node clusters on the map from both Figs. 4 and 5. As it can be observed that all CPs in Fig. 4 presents a lighter shade of nodes (low value) on top of the map. Those nodes correspond to the nodes position that suited on top map in Fig. 5 with also low fire risk. In contrary, the CP presents of nodes with high values on bottom, but different patterns. Therefore, the relation of both maps easily clarified that the most regions in *Cluster 1* (Fig. 5) receive high of fire risk from February to July and moderate for January (Fig. 4). Regions in *Cluster 2* only receive high of fire risk for January. Further, the most regions in *Cluster 3* receive high fire risks from August to November, and regions in *Cluster 4* are low fire risks for all months.

Comparison of methods

The use of PCA and SOM were first compared in terms of data extraction. The applied dataset illustrated that both methods are a suitable method for extraction the high dimensional data onto a low dimensional representation. PCA was very well to explain the variance of dataset because the extraction process provided a high percentage of the first two PCs. In contrast, SOM presented the variance of the input data based on a visual identification of the relative distances between the nodes network, according to the color shades on the U-matrix and CP. Second, both methods also developed data projection involving visual classification patterns. In this complex data application, scatter plot PCA was not suitable method to visualize the data classification into a 2-D space since the more forest regions hold similar hot spot occurrences. Conversely, the SOM map gave an excellent classification and visualization of fire risk in forest regions via the node clusters. These results support the finding study by Brosse *et al*. (2001) that SOM constituted a more reliable data representation method when complex ecological datasets were used. In a different study by Laitinen *et al*. (2002) also reported that SOM proved a useful interpretative method for analysis of large size datasets.

Conclusion

SOM and PCA have been applied to visualize and classify fire risk distribution in forest regions based on hot spot dataset. The results indicated that PCA has explained most the cumulative variance of data; unfortunately, the PCA projection was difficult to reveal a representative data pattern when the applied data available with large-scales. However, the SOM appears as a flexible method to represent the complexity of the data patterns. Especially, the CP of SOM most effective to visualize the fire risk distribution and the possible correlation between input variables on monthly basis.

In the context data classification, although the U-matrix of SOM was difficult to provide a representative data cluster, SOM color-coding and labeling was helpful in visualizing the clusters data structure via the map nodes. Therefore, the fire risks level according to their region clusters in forest can be easily interpreted. A problem of the SOM clustering, such specification of the clusters is sometimes still needed when the coloring is not clearly to indicate at the cluster borders. This is a challenge of the next study, to develop an analysis technique in order to yield a high classification rate of the SOM output.

Acknowledgement

The authors thank the Indonesian State Ministry of Environment Bureau for assistance on the data for hot spots of forest fires.

References

AFP (2002) Burning Question about Fire. Burning Issues: Thinking for more effective fire management, Asia Forest Partnership. Available online at http://www.asiaforests.org/scripts_afp/.

ASMC (2005) Fire monitoring and detection by remote sensing. ASEAN Specialized Meteorological Centre, Singapore. Available online at Available online at http://intranet.mssinet.gov.sg/asmc/asmc.html.

Abberger, H. M., B. M. Sanders and H. Dotzauer (2002) The development of community-based approach for an integrated forest fire management system in East Kalimantan, Indonesia. *In Communities in Flames: Proceedings of an International Conference on Community Involvement in Fire Management*, P. Moore, D. Ganz, L. C. Tan, T. Enters and P. B. Durst, Food and Agriculture Organization of the United Nations, Regional Office for Asia and the Pacific, Bangkok, Thailand. pp. 53–65.

Brosse, S., J. L. Giraudel and S. Lek (2001) Utilisation of non-supervised neural networks and principal component analysis to study fish assemblages. *Ecological Modeling*, 146, 159–166.

Flexer, A. (2001) On the use of self-organizing maps for clustering and visualization. *Intelligent-Data-Analysis*, 5, 373–384.

Hautaniemi, S., O. Yli-Harja, J. Astola, P. Kauraniemi, A. Kallioniemi, M. Wolf, J. Ruiz, S. Mousses and O. Kallioniemi (2003) Analysis and visualization of gene expression microarray data in human cancer using self-organizing maps. *Machine Learning*, 52, 45–66.

Himberg, J. (2000) SOM based cluster visualization and its application for false coloring. In *Proceedings of the International Joint Conference on Neural Networks*, *IEEE*, 3, 587–592.

Honda, R. and O. Konishi (2001) Temporal rule discovery for time-series satellite images and integration with RDB. In *Principles of Data Mining and Knowledge Discovery*, L. De Raedt, A. Siebes (eds), Springer-Verlag, Berlin Heidelberg. pp. 204–215.

Johnson, R. A. and D. W. Wichern (1998) *Applied multivariate statistical analysis*, International Edition, 4, United States of America: Prentice-Hall.

Kaski, S. (2001) SOM-based exploratory analysis of gene expression data. In *Advances in Self-Organising Maps*, N. Allinson, H. Yin, L. Allinson and J. Slack (eds), Springer-Verlag, London. pp. 124–131.

Kaski, S. and T. Kohonen (1998) Tips for processing and color-coding of self-organizing maps. In *Visual Explorations in Finance with Self-Organizing Maps*, G. Deboeck and T. Kohonen (eds), Springer-Verlag, London. pp. 195–202.

Kiang, M. Y. (2001) Extending the Kohonen self-organizing map networks for clustering analysis. *Computational Statistics and Data Analysis*, 38, 161–180.

Kiang, M. Y., M. Y. Hu and D. M. Fisher (2004) An extended self-organizing map network for market segmentation a telecommunication example. *Decision Support Systems*, DECSUP-11061; No of Pages 12. Available online at http://www.sciencedirect.com.

Kohonen, T. (1998) Self-organizing map. *Neurocomputing*, 21, 1–6.

Kohonen, T. (2001) Self-organizing maps. *3rd Ed, Springer-Verlag*, Berlin Heidelberg.

Koua, E. L. (2003) Using self-organizing maps for information visualization and knowledge discovery in complex geospatial datasets. In *Proceedings of the 21st International Cartographic Conference (ICC)*, Cartographic Renaissance, Durban, South Africa, pp. 1694–1702.

Kwan, C., R. Xu and L. Haynes (2001) A new data clustering and its applications. In *Proceeding of SPIE-The International Society for Op-*

tical Engineering, 4384, 1–5.

Laitinen, N., J. Rantanen, S. Laine, O. Antikainen, E. Rasanen, S. Airaksinen and J. Yliruusi (2002) Visualization of particle size and shape distributions using self-organizing maps. *Chemometrics and Intelligent Laboratory Systems*, 62, 47–60.

Mancuso, S. (2001) Clustering of grapevine (Vitis vinivera L.) genotypes with Kohonen neural networks, *VITIS*, 40, 59–63.

Starikov, A. (2000) Self-Organizing Maps—Mathematical Apparatus, BaseGroup Lab of data analysis technology. Available online at http://www.basegroup.ru/neural/som.en.htm.

Tipping, M. E. and C. M. Bishop (1999) Mixtures of probabilistic principal component analysers. *Neural Computation*, 11, 443–482.

Wan, W. J. and D. Fraser (2000) A multiple self-organizing map scheme for remote sensing classification, In *Proceedings of the Multiple Classifier System*, 300–309.

Wilppu, E. (1997) The visualization capability of self-organizing maps to detect deviations in distribution control, TUCS Technical Report No 153, Turku School of Economics and Business Administration, Finland.

Received October 23, 2006
Accepted February 13, 2007
Environmental information