

Choice of Cumulative Percentage in Principal Component Analysis for Regionalization of Peninsular Malaysia Based on the Rainfall Amount

Shazlyn Milleana Shaharudin¹(✉) and Norhaiza Ahmad²

¹ Department of Mathematics, Faculty of Science and Mathematics,
Universiti Pendidikan Sultan Idris, 35900 Tanjong Malim, Perak, Malaysia
shazlyn@fsmt.upsii.edu.my

² Department of Mathematics, Faculty of Science,
Universiti Teknologi Malaysia, UTM, 81310 Johor Bahru, Johor, Malaysia
norhaiza@utm.my

Abstract. Principal Component Analysis (PCA) is a popular method used for reduction of large scale data sets in hydrological applications. Typically, PCA scores are applied to hierarchical cluster analysis to redefine region. However, the choice of cumulative percentage of variance for PCA scores and identifying the best number of clusters can be difficult. In this paper, we investigate the effect of determining the number of clusters by comparing (i) standardized and unstandardized PCA scores on different cumulative percentages of variance (ii) to determine number of clusters using Calinski and Harabasz Index. We have found that Calinski and Harabasz Index is most appropriate to determine the best number of clusters and that standardized PCA scores within the range of 65% to 70% cumulative percentage of variance give the most reasonable number of clusters.

Keywords: Principal component analysis · Hierarchical analysis · Calinski and Harabasz Index

1 Introduction

Several researchers have used a variety of methods to define climatic types and delineate zones of similar climate. One of the methods that is popular among researchers to define region is through the combined use of principal component analysis (PCA) and cluster analysis. PCA is a data reduction technique where it allows the researcher to reorient the data, thus the first few dimensions account as much of the available information as possible. When working with less dimensions of the data set, this makes it easier to visualize the data and identify interesting patterns [1]. In defining region, the fundamental modes of PCA are considered for the clustering process [2]. Typically, in extracting components, one of the three rules, scree plot, Kaiser's rule and proportion of explain variance are taken as guideline [3, 4]. Scree plot proposed by [5] is a graphical approach which involved plotting the variance accounted for by each

principal component in the order of descending eigenvalues from the largest to the smallest. For a large dimension data set especially rainfall data set, this approach is unsuitable to be used. This is due to the steep curve followed by a bend which are not clearly visible to get the cutoffs of the number of principal components. When the scree plot is not diagnostic, Kaiser's rule may come in handy. This method retains regarding the amount of variance accounted for those components. In this rule, eigenvalues greater than average eigenvalue (i.e. $\lambda > 1$) are retained because these axes summarize more information than any single original variable [17]. Hence, those components with $\lambda > 1$ is obtained to determine the number of principal components. Occasionally in the data set, some eigenvalues are close to 1, thus it also might be consider as a significance of principal components to obtain. As a result, this method has been criticized by [18, 19]. For high dimensional data set, [5] recommended using 70% cumulative percentage of variance as a rough guide to cutoff the number of principal components. To test whether 70% cumulative percentage of variance is appropriate in defining region, the range of the cumulative percentage (65% until 90%) is tested to choose a suitable range to cutoff the principal components to define climate region. However in defining region, extracting the correct number of component is crucial because it dictates the true regional boundaries. As far as we know, there is no literature showing how to choose the appropriate number of components based on the breakdown point of the number of clusters.

The cutoffs of the number of components depends on the structure of the data set. In climate data, especially rainfall data set in Peninsular Malaysia, it involves many zero bound data which signifies that the observation is less than 1.0 mm [7]. These zero bound data might influence the choice of cumulative percentage of variance. It can be seen clearly when cluster analysis is employed to standardized and unstandardized principal component score, it demonstrates that the number of clusters are sensitive to the standardization data.

In this study, we establish a procedure to choose the best cumulative percentage of variance to obtain in defining region. We also need to investigate the effect of standardized and unstandardized principal component score to mitigate the effect of zero bound data.

2 Data

Daily rainfall totals for 33 years period 1975–2007 were obtained from 75 stations across Peninsular Malaysia. The rainfall data set considered for the purpose of this study is a matrix, comprising data from 75 stations and 365 days which constitute enough data to allow for defining region. In this present study, a wet day is defined as a day with at least 1 mm of rainfall [7]. Figure 1 shows the geographical coordinates of the stations in this study.

Before clustering process were employed, the needed standardization of the daily rainfall data was examined. Standardization is an important part in this analysis since mean and variance are likely to be small in consequence of zero bound data. The standardization will affect the result of clustering analysis where the rainfall stations are likely to be clustered together even if the stations are poorly correlated. Some

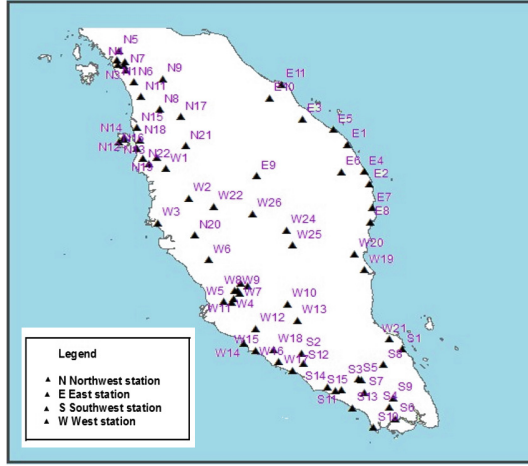


Fig. 1. The location of 75 rainfall stations in Peninsular Malaysia

adjustments are required in the usual standardization method due to a problem in zero bound data where the data were standardized by dividing the daily mean for that station, as given by

$$x_{ij}^* = \frac{x_{ij}}{\frac{365}{p} \sum_{j=1}^p x_{ij}} \quad (1)$$

where the denominator represents the daily mean rainfall at station i , calculated from the $p = 365 \times 33$ that is the daily observations. Daily rainfall is then expressed as a proportion of the mean of the daily total [6].

To validate the results in this study, we analyze another rainfall data set from other country which has same characteristic of rainfall data in Peninsular Malaysia. Daily rainfall data from 11 rainfall stations were obtained from Indonesia. The data were recorded from 2003 until 2005. The data set was assembled as data matrix of Peninsular Malaysia where the rows in the matrix represent the rainfall observation i.e. 365 rainfall days and 11 stations were represented in the columns of the data matrix. Standardization is also necessary to overcome the problem in dealing with zero bound data which is similar to the rainfall data in Peninsular Malaysia.

3 Methods

3.1 Principal Component Analysis

Principal components of the scaled rainfall data were computed based on the correlation matrix in order to extract the main modes of variation of the data and to reduce the from large dimension to low dimension. This procedure requires that several decisions be made in obtaining the best cumulative percentage of variance or in other

words, the best number of extracted components to retain. As mentioned previously in the introduction, several methods have been obtained in extracting the number of components to retain. In this study, we used explained variance to determine the best number of components to obtain. When using this method, the challenge lies in selecting an appropriate threshold percentage. If we choose the higher percentage such as 90% and above, we may encounter difficulties such as inflating the importance of noise and results in poorly defined regions. On the contrary, if we choose low variation of cumulative percentage, the observations that are not well represented will be clustered together due to low scores for all of the components. Therefore, we construct this study in order to determine the best range of cumulative percentage of variance in the defined region.

3.2 Calinski and Harabasz Index

Cluster analysis using k means method was then performed on principal component score matrix. The drawback of the k means method is the requirement for the number of clusters must be specified before the algorithm is applied. To counter this issue, we apply Calinski and Harabasz Index as a guide for us in quantifying the best number of clusters for our data set. Calinski and Harabasz Index is computed as

$$[\text{trace } B / (k - 1)] / [\text{trace } W / (n - k)] \quad (2)$$

where:

- n = total number of items k = number of clusters
- B = between pooled within cluster sum of square
- W = cross product matrix

The maximum value of the index was used to indicate the correct number of partitions in the data set.

4 Results and Discussion

In this section, we will discuss on the choice of cumulative percentage to cut off the number of principal components and the sensitivity of the number of cluster to the choice of cumulative percentage. We also show the effect of clustering result when using standardized and unstandardized principal component score. To validate the results of defining region in Peninsular Malaysia, we had compared the results with the rainfall data of Indonesia that has similar characteristic with Peninsular Malaysia.

The choice of cumulative percentage of variance will reflect the number of components to retain. As an example, we can see clearly in Table 1 where when 65% cumulative percentage was chosen, the number of components to retain is nine while when we chose 70% cumulative percentage, the number of components to retain is 13. The most significant effects shown on the choice of cumulative percentage of variance where it is sensitive to the number of cluster obtained. For instance, in Fig. 2, when we had selected effect 65% cumulative percentage, the number of cluster to retain was

Table 1. Results of standardized principal component score and number of clusters obtained using Calinski and Harabasz Index for Peninsular Malaysia

Eigenvalues	Cumulative percentage (%)	No. of component	No. of cluster	Calinski & Harabasz Index (max.value of index)
1.09	65	9	3	250.58
0.89	70	13	5	224.32
0.70	75	18	4	201.32
0.60	80	23	2	185.55
0.48	85	30	2	171.84
0.38	90	39	2	160.63

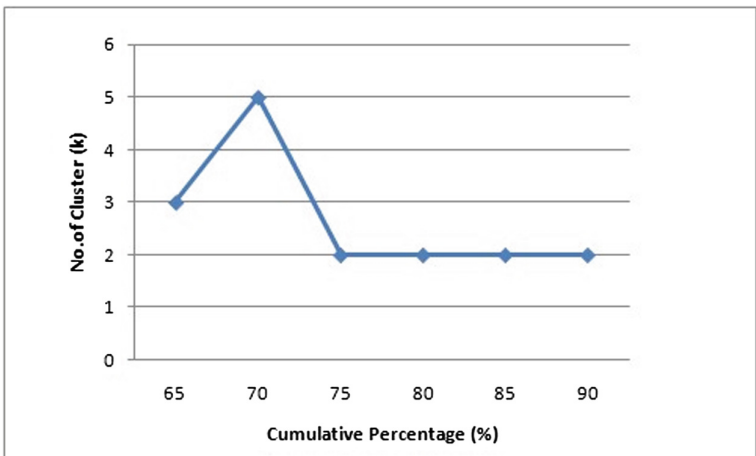


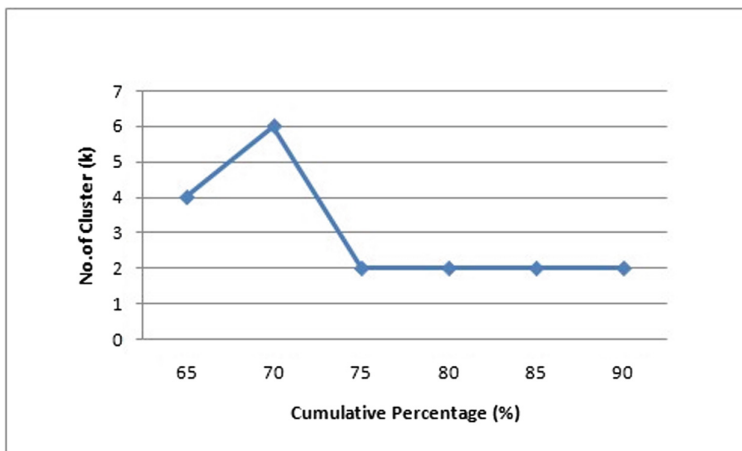
Fig. 2. Determined number of cluster for standardized principal component score for Peninsular Malaysia

three. When 5% additional cumulative percentage of variance is retained, the number of cluster changed from three to five. If we look through the Fig. 3, it gave the same result where 65% cumulative percentage of variance obtained two number of cluster. Meanwhile when 70% cumulative percentage of variance was retained, the number of cluster became six. However, the defined regions the selection of cumulative percentage above than 70% was not a good decision as a cut off for the number of principal components. As clearly presented in Figs. 1 and 2, the resulting number of cluster remained the same even with the additional 5% of the variance in every phase. Moreover, the number of cluster obtained from Figs. 1 and 2 is too small because in the defined region, we need more clusters to allow regions to benchmark their cluster against other regions [20]. This result is supported by [14] which stated that a few number of clusters i.e. two clusters would be insufficient to define region when dealing with analysing considerable extent of regions. This statement is proved by [11] where a sensitivity of the clustering results to the number of principal components retained has been noted elsewhere [21, 22].

Table 2. Results of standardized principal component score and number of clusters obtained using Calinski and Harabasz Index for Indonesia

Eigenvalues	Cumulative percentage (%)	No. of component	No. of cluster	Calinski & Harabasz Index (max.value of index)
0.98	65	5	4	108.68
0.86	70	6	6	93.13
0.74	75	7	2	83.90
0.72	80	8	2	69.75
0.68	85	9	2	69.60
0.58	90	10	2	60.20

Because of the sensitivity of the clustering results to the number of retained principal components, the correct number of components to retain needs to be identified. It is important that the variation between the clusters is represented in the direction of at least one of the principal components [12]. Accordingly, it is best to err towards retaining significantly more principal components rather than too few [13]. If there are too few components to retain, the observations that are not well represented will cluster together because they have low scores for all the components meanwhile inclusion of too many principal components inflates the importance of noise and results in poorly defined regions [6]. Clustering results are not as sensitive to the choice of cumulative percentage of variance when the component scores are left unstandardized compared to when they are standardized. If we look through Tables 3 and 4, the number of cluster remains the same even though we have increased 5% cumulative percentage of variance in every phase. This situation happened due to the lowest-order modes, which define the noise element of the data, are given minimal weighting. Therefore, we need standardized principal component score to ensure all the temporal modes are given equal weight and rainfall distribution patterns that occur frequently are treated as equal to unusual patterns and to noise components.

**Fig. 3.** Determined number of cluster for standardized principal component score for Indonesia

In order to obtain the best number of cluster, Calinski and Harabasz index was employed in principal component score matrix. According to Table 1 for Peninsular Malaysia, the values for cluster numbers run from two to five while for Indonesia in Table 2, the cluster was obtained from two to six. The optimum number of cluster was established as three for Peninsular Malaysian and four for Indonesia where each recorded the maximum value of index among the others.

Table 3. Results of unstandardized principal component score and number of clusters obtained using Calinski and Harabasz Index for Peninsular Malaysia

Eigenvalues	Cumulative percentage (%)	No. of component	No. of cluster	Calinski & Harabasz Index (max.value of index)
1.09	65	9	2	265.78
0.89	70	13	2	229.45
0.70	75	18	2	208.46
0.60	80	23	2	167.49
0.48	85	30	2	154.28
0.38	90	39	2	150.20

Table 4. Results of unstandardized principal component score and number of clusters obtained using Calinski and Harabasz Index for Indonesia

Eigenvalues	Cummulative percentage (%)	No. of component	No. of cluster	Calinski & Harabasz Index (max.value of index)
0.98	65	5	2	265.78
0.86	70	6	2	229.45
0.74	75	7	2	208.46
0.72	80	8	2	167.49
0.68	85	9	2	154.28
0.58	90	10	2	150.20

5 Conclusion and Recommendations

This study has shown that PCA method is particularly well adapted to the regionalization of rainfall region. It allows the grouping of stations with similar characteristics and recognition of climatic regions in the alpine domain [21]. Typically, in defining climate region, it will need the largest cluster to retain. If we only have fewer groups, we have to face the problem in differentiating the new region defined and it will give us difficulty to analyze it. Hence, the following recommendations are made for cluster analysis cum PCA to define new region:

- (1) If there are too few components, observations that are not well represented will be clustered together due to low scores on all of the components but if more number of components to retain or more cumulative percentage are taken, the result of

defining region become poor as it will inflate the importance of noise. Therefore, the most suitable cumulative percentage to define region is between 65% until 70%.

- (2) The principal component scores should be standardized as it will make the clustering result become sensitive to the number of component to retain.
- (3) Validity index is recommended to be used when determining the best number of cluster to define region.

Generally, we have a lot of methods in defining region such as modeling method and regression method. Our proposed method may also be used by researchers to define climate region in their countries. All of the recommendations above can be used as guideline for other researchers with similar topics related to this paper. Having mentioned this, it is not a claim that all of the result is entirely accurate for all cases as it is based on rainfall data in Peninsular Malaysia and Indonesia. Both of these countries are part of the Asia, hence the weather and seasons are different compared to the other zones.

Acknowledgments. The authors would like to thank Universiti Pendidikan Sultan Idris for their financial funding.

References

1. Lattin, J.: *Analyzing Multivariate Data*. Curt Hinrichs, Canada (2003)
2. Romero, R., Ramis, C., Guijarro, J.A.: Daily rainfall patterns in the spanish mediterranean area: an objective classification. *Int. J. Climatol.* **19**, 95–112 (1999)
3. Alvin, C.R.: *Methods of Multivariate Analysis*. Wiley, Hoboken (2002)
4. Cliff, N.: *Analyzing Multivariate Data*. Harcourt Brace, San Diego (1987)
5. Cattell, R.B.: The scree test for the number of factors. *Multivar. Behav. Res.* **1**, 245–276 (1966)
6. Jolliffe, I.T.: Discarding variables in principal component analysis. I: artifical data. *Appl. Stat.* **21**, 160–173 (1972)
7. Mimmack, G.M., Mason, S.J., Galpin, J.S.: Choice of distance matrices in cluster analysis: defining region. *J. Clim.* **14**, 2790–2797 (2000)
8. Suhaila, J., Deni, S.M., Wan Zin, W.Z., Jemain, A.A.: *Sains Malays.* **39**(4), 533–542 (2010)
9. Wilks, D.S.: *Statistical Methods in the Atmospheric Sciences*, p. 467. Academic Press, Cambridge (1995)
10. Aldenderfer, M.S., Blashfield, R.K.: *Cluster Analysis*. Sage Publications, Inc., Beverly Hills (1984)
11. Mielke, P.W.: *J. Atmos. Sci.* **42**, 1209–1212 (1985)
12. Bunkers, M.J., Miller, J.R., DeGaetano, A.T.: *J. Clim.* **9**, 130–146 (1996)
13. Jolliffe, I.T.: *Principal Component Analysis*. Springer Series in Statistics, p. 271. Springer, Heidelberg (1986)
14. Chang, W.C.: On using principal components before separating a mixture of two multivariate normal populations. *J. Appl. Stat.* **32**, 267–275 (1983)
15. Pelczar, I.J., Cisnerous-Iturbe, H.L.: Identification of rainfall patterns over the valley Mexico. In: 11th International Conference on Urban Drainage, pp. 1–9 (2008)
16. Fovell, R.G., Fovell, M.Y.C.: *J. Clim.* **6**, 2103–2135 (1993)

17. Donald, A.J.: Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology* **74**(8), 2204–2214 (1993)
18. Grossman, G.D., Nickerson, D.M., Freeman, M.C.: Principal component analyses of assemblage structure data: utility of tests based on eigenvalues. *Ecology* **72**, 341–347 (1991)
19. Rexstad, E.A., Miller, D.D., Flather, C.H., Anderson, E.M., Hupp, J.W., Anderson, D.R.: Questionable multivariate statistical inference in wildlife habitat and community studies. *J. Wildl. Manag.* **52**, 794–798 (1988)
20. Mercedesm, D., Michael, E.P, Scott, S.: Defining Clusters of Related Industries (2013)
21. Baeriswyl, P.A., Rebetez, M.: Regionalization of precipitation in switzerland by means of principal component analysis. *Theor. Appl. Climatol.* **58**, 31–41 (1997)
22. Bunkers, M.J., Miller, J.R., DeGaetano, A.T.: Definition of climate regions in the northern plains using an objective cluster modification technique. *J. Clim.* **9**, 130–146 (1996)
23. DeGaetano, A.T.: Delineation of mesoscale climate zones in the Northeastern United States using a novel approach to cluster analysis. *J. Clim.* **9**, 1765–1782 (1996)