Pertanika J. Sci. & Technol. 12(1): 1 - 9 (2004)
ISSN: 0128-7680

On Robust Environmental Quality Indices

[1]Azme Khamis & [2]Mokhtar Abdullah

[1]Pusat Pengajian Sains, Kolej Universiti Teknologi Tun Hussein Onn
[2]Pusat Pengajian Sains Matematik, Fakulti Sains & Teknologi
Universiti Kebangsaan Malaysia,
43600 UKM, Bangi, Selangor, Malaysia

Received: 8 December 1998

ABSTRAK

Kajian ini membincangkan rumus baru indeks kualiti alam sekitar yang boleh digunakan untuk memantau parameter udara dan juga parameter kaji cuaca yang lain. Perumusan indeks ini berasaskan kepada analisis komponen utama konvensional dan analisis komponen utama teguh yang dapat memberikan gabungan linear terbaik bagi parameter alam sekitar. Perbandingan telah dilakukan di antara indeks daripada analisis komponen utama konvensional (PCA) dan indeks analisis komponen utama teguh (RPCA). Keputusan menunjukkan bahawa RPCA dapat memberikan satu alternatif gabungan linear yang lebih baik. Contoh berangka mengenai kualiti udara telah dilakukan untuk menunjukkan penggunaan indeks kualiti alam sekitar teguh.

ABSTRACT

This paper discusses a formulation of new environmental quality indices, which can be used for monitoring environmental as well as meteorological parameters. The formulation of the indices is based on conventional and robust principal component analysis, which gives the linear combination of environmental parameters. Comparisons are made between the conventional principal component analysis (PCA) indices and robust principal component analysis (RPCA) indices. The results show that the RPCA gave a better alternative linear combination. A numerical example on air quality was used to illustrate the application of the robust environmental indices.

Keywords: Conventional principal component analysis, robust principal component analysis, quality indices, MLT-estimator, CMB-estimator

INTRODUCTION

Indices or indicators are useful means of observing trends, analysing programs, policy making and informing the public of important concepts in a simple understandable manner. An index is defined as a scheme that transforms the (weighted) values of individual pollutant-related parameters (for example, carbon monoxide concentration or visibility) into a single numbers, or set of number and the result is a set of rules (for example, an equation) that translates parameter values by means of a numerical manipulation into a more parsimonious form (Ott and Thom 1976). Pikul (1974) defined an index, which is a mathematical combination of two or more parameters, which can have utility at least, in an interpretive sense.

In an environmental context, environmental indices are used to give insight into environmental conditions. They should serve as a means to examine the changes in climate, to highlight specific environmental conditions and to help

governmental decision-makers evaluate the effectiveness of regulatory programs. The best measurement of selected parameters, which are reported in a timely and effective manner merely provides the policy maker with large amounts of data. To be useful for evaluation and assessment, these data must be aggregated in a meaningful way to show the right magnitudes and trends.

In discussing environmental aspects, there is more than one parameter that needs to be analyzed. Normally, traditional approaches are used in multivariate data, such as factor analysis, principal component analysis, discriminant analysis, biplot analysis and multidimensional scaling. This paper discusses the implementation of the robust PCA in developing environmental indices.

METHODOLOGY

A principal component analysis (PCA) is concerned with explaining the variance covariance structure through a few linear combinations of original variables. Its general objectives are data reduction and data interpretation. Although p components are required to reproduce the total system variability, often a small number, k, of the principal components, can account for much of this variability. If so, there is as much information in the k components as there is in the original p variables. The original data, consisting of n measurement on p variables, is reduced to one consisting of n measurements on k principal components.

Algebraically, principal components are particular linear combinations of the p random variables $X_1, X_2,..., X_p$. Geometrically, these linear combinations represent the election of a new coordinate system obtained by rotating the original system with $x_1, x_2,..., x_p$ as the coordinate axes. The new axes represent the directions with maximum variability and provide a simpler and parsimonious description of the covariance structure.

The method of principal components is based on a key result from matrix algebra: A $p \times p$ symmetric, nonsingular matrix, such as the covariance matrix Σ, may be reduced to a diagonal matrix L by premultiplying and postmultiplying it by a particular orthonormal Matrix U such that $U'\Sigma U = L$. The diagonal elements of L, $l_1, l_2,..., l_p$ are called the characteristic roots, latent roots or eigenvalues of Σ. The columns of U, $u_1, u_2,..., u_p$ are called the characteristic vectors or eigenvactors of Σ. The characteristic roots may be obtained from the solution of the characteristic equation $|\Sigma - l\, I| = 0$, where I is the identity matrix. This equation produces a p^{th} degree polynomial in l from which the values $l_1, l_2,..., l_p$ are obtained.

If the covariances are not equal to zero, it indicates that a linear relationship exists between these two variables, the strength of that relationship being represented by the correlation coefficient. The principal axis will transform p correlated variables $x_1, x_2,..., x_p$ into p new uncorrelated variables $z_1, z_2,..., z_p$. The coordinate axis of these new variables are described by the characteristic vectors u_i which make up the matrix U of direction cosines used in the transform $z = U'[x - \Sigma]$. The transformed variables are called the principal components of x and the covariance matrix of Z is $\text{cov}(Z) = \text{tr}\ (\lambda_1, \lambda_2, \lambda_3,..., \lambda_p)$.

The majority of techniques have assumed that the data with which we are working are basically 'good'. A number of problems may occur. First, the assumptions regarding the underlying distribution used. Second, there are assumptions of independence of the sample observation. Third, 'outliers' observations. It is possible that outliers can affect the roots and vectors themselves and for that, robust estimation procedures will be required. There are four classes of robust estimators; namely the adaptive estimator, the L-estimator, the M-estimator and the rank test estimator. Let $x_1, x_2, ..., x_n$ be a random sample from a distribution with a probability density function $f(x_i-\theta)$, where θ is the location parameter. Then, the log likelihood function can be written as

$$\ln L(\theta) = \sum_{i=1}^{n} ln\, f(x_i-\theta)$$

$$= -\sum_{i=1}^{n} \eta(x_i-\theta), \quad \text{where } \eta(x) = -\ln f(x)$$

Thus,

$$\frac{d\, ln\, L(\theta)}{d\theta} = -\sum_{i=1}^{n} \frac{f'(x_i-\theta)}{f(x_i-\theta)} = \sum_{i=1}^{n} \eta(x_i-\theta)$$

where $\eta'(x) = \tau(x)$. The solution, $\hat{\theta}$, of $\sum_{i=1}^{n} \tau(x_i-\theta) = 0$ maximise $L(\theta)$ and $\hat{\theta}$ is called the maximum likelihood estimator of θ.

Marrona (1976) introd uced M-estimator for the location vector, μ, and covariance matrix, Σ, for the solution of the system

$$(1/n)\sum_{i=1}^{n} w_1(d_i)(x_i-\mu) = 0$$

and

$$(1/n)\sum_{i=1}^{n} w_2(d_i^2)(x_i-\mu)(x_i-\mu)' = \Sigma$$

where $d_i = [(x_i - \mu)'\Sigma^{-1} (x_i - \mu)]^{1/2}$ is the Mahalanobis distance function and w_1 and w_2 are functions of the technique involved.

The M-estimator

An iteration procedure is needed to calculate the robust value of d_i^2 where $d_i=[(x_i - m^*) S^{*-1} (x_i - m^*)]^{1/2}$, i = 1, 2, ..., n. where m^* and S^* are the new estimators of μ and Σ. Generally, the M-estimators of μ and Σ are given by

$$m^* = \sum_{i=1}^{n} w_1(d_i)x_i \Big/ \sum_{i=1}^{n} w_1(d_i) \text{ and } S^* = \frac{\sum_{i=1}^{n} w_2(d_i^2)(x_i - m^*)(x_i - m^*)'}{f\left[w_2(d_i^2)\right]}$$

where $w_1(d_i)$ and $f[w_2(d_i^2)]$ are some suitable weight functions.

(i) The MLT-estimator

Marrona (1976) suggested the weight functions of w_1 and w_2 as follows. Let

$$w_1(d_i) = \frac{(p+v)}{(v+d_i^2)} = w_2(d_i^2)$$

where v is the degree of freedom associated with the multiple t-distribution. Generally, v is set to 1, the Cauchy distribution. This is the value used by Devlin *et al.* (1981) and $f\{w_2(d_i)\} = 1/n$. So, the likelihood maximum t-estimator (MLT) of mean μ, and covariance matrix Σ, are:

$$m_{MLT} = \frac{\sum_{i=1}^{n} w_1(d_i)x_i}{\sum_{i=1}^{n} w_1(d_i)} \text{ and } S_{MLT} = (1/n)\sum_{i=1}^{n} w_2(d_i^2)(x_i - m^*)(x_i - m^*)'$$

where $w_1(d_i) = (1+p)/(1+d_i^2) = w_2(d_i^2)$, refer to Jackson (1991) for details.

(ii) The CMB-estimator

Campbell (1980) suggested a phi-function ψ, redeclines in $w(d_i) = \psi(d_i)/d_i$ where

$$w(d_i) = \begin{cases} 1 & \text{if } d_i \le \xi \\ \dfrac{\xi}{d_i} exp\left\{\dfrac{-(d_i - \xi)^2}{2c_2^2}\right\} & \text{if } d_i > \xi \end{cases}$$

if $w_2(d_i^2) = [w_1(d_i)]^2$ and $\xi = \sqrt{p} + c_1/\sqrt{2}$. Then $w_1(d_i) = [w_1(d_i)]/d_i$;

$w_2(d_i^2) = [w_1(d_i)]^2$ and $f[w_2(d_i)] = \sum_{i=1}^{n} [w_2(d_i)]^2 - 1$

the c_1 and c_2 are the robust scale estimators to ensure w has robust characteristics. When normality assumption is considered, square root of Fisher transformation

for Chi-square distribution will give d_i approximates to normal distribution with mean $\sqrt{p}$ and variance $1/\sqrt{2}$. Campbell also suggested that $c_1 = 2$ and $c_2 = 1.25$ to produce the robust characteristics which have been suggested by Hampel (1973). This procedure produced weights that decrease at a faster rate than other procedures.

The mean and covariance matrix estimator are

$$m_{MCB} = \frac{\sum_{i=1}^{n} w_1(d_i)x_i}{\sum_{i=1}^{n} w_1(d_i)} \quad and \quad S_{CMB} = \frac{\left\{\sum_{i=1}^{n} w_2(d_i^2)(x_i - m^*)(x_i - m^*)'\right\}}{\left\{\sum_{i=1}^{n} [w_2(d_i)]^2 - 1\right\}}$$

All of the multivariate procedures are sensitive to starting values and usually work best with robust estimates at the beginning. For the starting values in the iteration, the sample mean, $\bar{x}$ and sample covariance matrix, S, are used. If there are extreme values, the robust median estimator, x_m and matrix $S_m = \frac{\sum_{i=1}^{n}(x_i - x_m)(x_i - x_m)'}{n-1}$ are used for $\bar{x}$ and S, respectively. The procedure is repeated until the correlation matrix converges.

The methods discussed earlier produced robust estimates of the mean and variance while the characteristic roots and vectors were obtained from them by conventional PCA. These results are called robust PCA because the starting matrices were robust.

The Development of the Robust Indices

The air quality index introduced by Pikul (1974), is defined as

$$Q_I = 1 - P_I \qquad (1)$$

where P_I is an index of air pollution. The air pollutant index is defined without regard to synergistic effects, which occur as a result of reactions between two or more pollutants. The keys to determining P_I are the index standard for each pollutant, which need not correspond to legal standards. Let

S_{i1} be the standards concentration at 50th percentile (median) for pollutant i
S_{i2} be the standards concentration at 85th percentile (1σ) for pollutant i
S_{i3} be the standards concentration at 95th percentile (2σ) for pollutant i

Then, the standards pollution index from the first principal component, $S_{z1(ik)}$ can be written as $S_{z1(ik)} = \Sigma\omega_{ik} S_{ik}$ where $k = 1, 2, 3$; $i = 1, 2, 3, 4, 5$; and ω_{ik}

the coefficient or weight from the first principal component. The air quality index includes five pollutants, namely ozone (O_3), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO) and suspended particulate matter (SPM).

Let M_{ik}, k =1, 2, 3, correspond to the measured concentration values of pollutant i for the k-th percentile index and let $M_{z1(ik)}$ be the pollutants indices from the first principal components, Z. So we can get the pollutant indices which correspond to the measured concentration values of pollutant i for the k^{th} percentile index; $M_{z1(ik)} = \Sigma\omega_{ik} M_{ik}$ k=1, 2, 3; i = 1, 2, 3, 4, 5; and ω_{ik} are the coefficients and weights from the first principal component.

Then the pollution index is computed as

$$P_I = \frac{1}{M}\left(\sum_{k=1}^{3} \upsilon_{ik} \frac{M_{z1(ik)}}{S_{z1(ik)}}\right) \qquad (2)$$

where the υ_{ik} is the relative weights assigned to each percentile value ($\Sigma\upsilon_{ik} = 1$) and M is a factor that ensures that P_I does not exceed unity.

RESULT AND DISCUSSION

The CPCA Indices

Standard indices have been used for the five primary pollutants shown in Table 1. The standard indices are based on the MAQI (Malaysian Air Quality Index) and RMG (Recommended Malaysian Guideline).

TABLE 1
Standard index based on the RMG and MAQI

Pollutants	Standard index		
	S_{i1}	S_{i2}	S_{i3}
Ozone, O_3 (ppm)	0.1000	0.3270	0.5541
Nitrogen Dioxide, NO_2 (ppm)	0.0600	0.2795	0.4989
Sulphur Dioxide, SO_2 (ppm)	0.0400	0.1265	0.2130
Carbon Monoxide, CO (mg/m^3)	9.0000	24.2634	39.5269
Suspended Particulate Matter, SPM (µg/m3)	150.000	314.9889	479.9777

For this study, the data from 1st to 31st July 1995 from the Kuala Lumpur environmental station were used. To get the standard pollutant index $S_{z1(ik)}$, the coefficient derived from the conventional and robust PCA are multiplied with the pollutant value for each percentile.

Table 2 shows the correlation value (from conventional correlation matrix) between pollutants and it is found that the pollutants have a positive relationship except for O_3 and CO and O_3 and SPM respectively. However, the correlation values are small -0.1205 and -0.2639, respectively.

TABLE 2
The correlation matrix based on conventional PCA

Pollutants	O_3	NO_2	SO_2	CO	SPM
O_3	1.0000				
NO_2	0.4231	1.0000			
SO_2	0.3147	0.4239	1.0000		
CO	-0.1205	0.2135	0.6427	1.0000	
SPM	-0.2639	0.1169	0.9237	0.5428	1.0000

The first eigen value from the correlation matrix is 2.5380 and it explains 50.7% of the variation in the data. As mentioned before, the quality index is calculated based on the first principal component. The pollution index equation is

$$P_1 = 0.2797*NO_2 + 0.0440*O_3 + 0.5848*CO + 0.5245*SO_2 + 0.5501*SPM.$$

Pollutant index, $M_{z1(ik)}$ can be derived by multiplying each M_{ik} with the coefficient (which is derived from the conventional and robust PCA) and then totalled. The air quality index would be computed from equation (1) by substitution from equation (2).

The quality indices throughout July 1995 are displayed in Table 3.

TABLE 3
Indices derived from the first day to the 31st with the conventional PCA

Day	QI	Day	QI
1	0.97120	17	0.95986
2	0.97063	18	0.96771
3	0.97282	19	0.97701
4	0.97855	20	0.96706
5	0.96860	21	0.97477
6	0.96476	22	0.97918
7	0.96188	23	0.97743
8	0.96588	24	0.98261
9	0.97005	25	0.97232
19	0.96570	26	0.97881
11	0.96065	27	0.98127
12	0.95916	28	0.97928
13	0.96007	29	0.96015
14	0.95881	30	0.95901
15	0.95614	31	0.96566
16	0.96288		

This method shows that the air quality in Kuala Lumpur is in good condition.

The RPCA Indices

The calculation of RPCA index is still the same as in CPCA, the only difference is the coefficient former index is derived from a robust estimator. Table 4 shows that *CO* and *SPM* are highly correlated with correlation value of 0.9318 whereas *CO* and SO_2 have a correlation value of 0.6060.

TABLE 4
The correlation matrix based on the robust PCA

Pollutants	O_3	NO_2	SO_2	*CO*	*SPM*
O_3	1.0000				
NO_2	0.5769	1.0000			
SO_2	0.5626	0.5616	1.0000		
CO	0.1678	0.4046	0.6060	1.0000	
SPM	0.0358	0.2960	0.9318	0.4924	1.0000

The first eigen value is 2.8963 and it explains 57.9% of the variation in the data. This means the RPCA can explain more variation than the conventional counterpart. The pollution index equation is

$$P_I = 0.4272*NO_2 + 0.3319*O_3 + 0.5023*CO + 0.5024*SO_2 + 0.4502*SPM$$

Finally, the quality indices throughout July 1995 are displayed in Table 5.

TABLE 5
Indices derived from the first day to the 31st with the robust PCA

Day	QI	Day	QI
1	0.97656	17	0.95472
2	0.97023	18	0.97019
3	0.97038	19	0.97954
4	0.99989	20	0.96483
5	0.96882	21	0.97765
6	0.96748	22	0.97486
7	0.96241	23	0.97612
8	0.96346	24	0.98723
9	0.96918	25	0.97232
19	0.96623	26	0.99940
11	0.95920	27	0.98339
12	0.95271	28	0.99724
13	0.96256	29	0.96609
14	0.95923	30	0.96018
15	0.95928	31	0.96357
16	0.96254		

This method also shows that the air quality in Kuala Lumpur is in good condition.

CONCLUSIONS

The PCA structure can be a better alternative to explain the combination environmental parameters in developing environmental indices. The largest weight in pollution index equation indicates the most influencing factor in air pollution phenomenon. From the pollution index equation, the contribution of each parameter in air pollution can be determined. The results show that the robust estimators are more successful in giving a better alternative result. The CPCA explains only 50.7% of the variation in the data, while RPCA explains 57.97%. However, the CPCA and RPCA give the same caution signals regarding the air condition, but the values of indices are relatively different.

From the structure of the new quality indices, it shows that it gives a better alternative to monitor the level of air quality. Furthermore, the indices are easy to understand, easy to calculate and more comprehensive. This method (RPCA) is very flexible and it can be adapted to any type of environmental parameters, such as water quality, noise pollution, quality of life, etc. If the indices are plotted on the graph, the trends of environmental parameters can be detected and can be used for forecasting purposes.

REFERENCES

AZME KHAMIS. 1996. Pembinaan indeks kualiti udara teguh. MSc. thesis, Universiti Kebangsaan Malaysia.

CAMPBELL, N. A. 1980. Robust procedures in multivariate analysis I: robust covariance estimation. *Applied Statistics* **29:** 231–237.

DEVLIN, S. J. R. GNANADESIKAN and J. R. KETTENRING. 1981. Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association* **76:** 354–362.

HAMPEL, F. R. 1973. Robust estimation a condensed partial survey. *Z. Wahr.verw. Geb.* **27:** 87–104.

JACKSON, J. E. 1991. *A User's Guide to Principal Components.* John Wiley & Sons Inc.

JOHNSON, R. A. and D. W. WICHERN. 1988. *Applied Multivariate Statistical Analysis.* Second edition. Prentice Hall International, Inc.

MARRONA, R. A. 1976. Robust M-estimators of multivariate location and scatter. *Annals of Statistics* **1:** 51–67.

OTT, W. R and G. C. THOM. 1976. *Air Pollution Indices: A Compendium and Assessment of Indices Used in United States and Canada.* Michigan: Ann Arbor Science.

PIKUL, R. 1974. Development of environmental indices. In *Statistical and Mathematical Aspects of Pollution Problems,* ed. J. W. Pratt.