ENVIRONMETRICS
Environmetrics 2003; **14**: 409–416

On some aspects of data integration techniques with environmental applications

Bikas K. Sinha[1,*,†] and Kirti R. Shah[2,‡]

[1]*Indian Statistical Institute, Calcutta, India*
[2]*University of Waterloo, Canada*

SUMMARY

Multiple Criteria Decision Making (MCDM) is a popular phrase used to describe situations where there is a need for integration of the results of different studies to make an overall judgement. Among the highest priorities towards socioeconiomic development around the world is the Environmental Protection Policy (EPP), and environmental assessment is a key to EPP. In the context of environmental studies, data integration techniques are very appealing and have wider applicability. It is well known that land, air and water are the three sources for determination of the extent of pollution of different regions. The purpose of MCDM is to rank the regions wrt all the sources taken together. For any individual source of pollution, it is trivial to rank the regions from best to worst. However, the problem of integration becomes non-trivial in most cases since the regions do not lend themselves to the same pattern of ranking wrt different sources. In this article we examine critically the performance of two popular composite indices (CI) and suggest some alternatives.

KEY WORDS: environmental indicators; sources; locations; ideal; anti-ideal; weights; composite index

1. INTRODUCTION

Advocated by Hwang and Yoon (1981), Zeleny (1982) and Yoon and Hwang (1995), MCDM is a body of techniques used for meaningful integration of component indices to an overall index in order to decide on the ranking of a number of locations from best to worst. This is based on the premises that in the absence of a natural ideal location, a best alternative would be the one which has the shortest distance from the hypothetical ideal location.

We begin with the mathematical formulation of the problem. We are given a matrix $\boldsymbol{X}$ of order $K \times N$ with rows representing locations (or regions or environmental units) and columns representing sources (or characters), where the element x_{ij} of $\boldsymbol{X}$ denotes an index of the ith location wrt the jth source. Denote by $x_{1:j}$ and $x_{K:j}$, respectively, the smallest and the largest possible values of x_{ij} for fixed $j, 1 \leq j \leq N$. An ideal hypothetical location corresponds to one for which the row vector of x-values is given by $[x_{1:1}, x_{1:2}, \ldots, x_{1:N}]$, comprising the smallest possible values for each source. Likewise, the one based on $[x_{K:1}, x_{K:2}, \ldots, x_{K:N}]$ will be referred to as an anti-ideal location.

The main purpose of such a study is to identify the location which is closest to the ideal in some sense. While doing so, one may also incorporate the fact that such a location is expected to be the farthest from anti-ideal.

To make our search non-trivial, we assume that such ideal or anti-ideal locations do not exist. One way to identify the closest location is to compare the given locations by suitably defining a composite index (CI) based on the distances from the ideal and the anti-ideal. We will set the CI value at 0 for the ideal and at ∞ for the anti-ideal and define the criterion of closeness as one corresponding to the least value of the CI.

There are two popular CIs studied in the literature and applied in practice. One is based on the *Technique for Ordering Preferences by Similarity to Ideal Solution*, abbreviated as TOPSIS; see Zeleny (1982). The other, known as the *Electre* method, is computation-intensive. We will critically examine both the CIs in this article and suggest some needed modifications.

2. NOTATION

Let

$$\begin{aligned} d_{ij} &= \text{distance of } x_{ij} \text{ from } x_{1:j} \\ d_{ij}^{-} &= \text{distance of } x_{ij} \text{ from } x_{K:j}, \qquad 1 \le i \le K, 1 \le j \le N \end{aligned} \tag{1}$$

Thus, d_{ij}'s (respectively d_{ij}^{-}'s) represent distances from the ideal (respectively anti-ideal) location.

Next, let

$$\boldsymbol{d}_{i:N} = [d_{i1}, d_{i2}, \ldots, d_{iN}]' \tag{2}$$

which represents the row-vector of distances d_{ij} for the ith location involving N characters,

$$\boldsymbol{d}_{i:N}^{-} = [d_{i1}^{-}, d_{i2}^{-}, \ldots, d_{iN}^{-}]' \tag{3}$$

which likewise represents a row-vector of distances d_{ij}^{-} for the ith location involving N characters.

In general, a CI is of the form

$$\varphi(\boldsymbol{d}, \boldsymbol{d}^{-}) = \varphi(\boldsymbol{w}, \theta, \boldsymbol{d}, \boldsymbol{d}^{-}) \tag{4}$$

where $\boldsymbol{w} = (w_1, \ldots, w_N)'$ are weights, usually independent of the data, whereas $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)'$ are normalizing constants, usually dependent on the data. In a given context, we will abbreviate CI as $\varphi(\boldsymbol{d}, \boldsymbol{d}^{-})$ or simply as $\varphi(\boldsymbol{d})$.

Below we now proceed to define the TOPSIS-based CI.

Define, for the ith location,

$$L_2(i, IDR) = \left[\sum_j \left\{ w_j d_{i,j}^2 \Big/ \sum_i x_{ij}^2 \right\} \right]^{1/2} \tag{5}$$

based on distances from the ideal and, further,

$$L_2(i, NIDR) = \left[\sum_j \left\{ w_j d_{i,j}^{-2} \Big/ \sum_i x_{ij}^2 \right\}\right]^{1/2} \tag{6}$$

based on those from the anti-ideal.

Then the TOPSIS-based CI for the ith location is given by

$$\varphi_i(\boldsymbol{d}) = \frac{L_2(i, IDR)}{L_2(i, IDR) + L_2(i, NIDR)} \tag{7}$$

It is clear that, for the purpose of comparison of different locations, we can work with

$$\varphi_i^*(\boldsymbol{d}) = \frac{L_2(i, IDR)}{L_2(i, NIDR)} \tag{8}$$

A description of TOPSIS-based CIs can be found in Filar *et al.* (1997).

3. REQUIREMENTS

The general problem of formation of a CI is not easy to resolve. Usually, one cannot come up with a unique choice which should serve for all occasions. However, we may be guided by certain laid-down principles. These principles may not lead to evolution of a CI but we may be in a position to examine the status or validity of existing CIs.

With this objective, we now proceed to study certain other facts regarding formation of CI. To start with, we would expect the following requirements to be satisfied by any CI:

Write $\boldsymbol{d}\{. : N\}$ for an arbitrary N-vector of d-values.

R1: If $\boldsymbol{d}\{. : N\} \leq \boldsymbol{d}^*\{. : N\}$ (co-ordinatewise), then $\varphi(\boldsymbol{d}) \leq \varphi(\boldsymbol{d}^*)$, with strict inequality in the latter in case the same holds in the former (in at least one co-ordinate).

R2: If $\varphi(\boldsymbol{d}\{. : N\}) \leq \varphi(\boldsymbol{d}^*\{. : N\})$, then $\varphi(\boldsymbol{d}\{., \boldsymbol{a} : N + M\}) \leq \varphi(\boldsymbol{d}^*\{., \boldsymbol{a} : N + M\})$ for all additional M-vectors $\boldsymbol{a}, M \geq 1$.

R1 states that $\boldsymbol{d}$ must be better than $\boldsymbol{d}^*$ wrt every meaningful φ in case $\boldsymbol{d}^*$ has a coordinate-wise larger distance (from the ideal) than $\boldsymbol{d}$ wrt at least one co-ordinate.

R2 states that once $\boldsymbol{d}$ performs better than $\boldsymbol{d}^*$ wrt φ, it continues to perform better if both are extended by addition of more features for each of which the distances are the same for both. There is something more to it; we return to the technical details later.

Remark 1: Requirements 1 and 2 seem to be logical from both the theoretical and practical points of view. It would be interesting to investigate how far the CIs suggested in the literature satisfy these requirements. We shall deal with both the TOPSIS-based CI and the one based on the Electre method.

We show below that $\varphi^*(\boldsymbol{d})$ satisfies R1 but does not satisfy R2.

That it satisfies R1 is easy to verify, since

$$\boldsymbol{d}\{. : N\} \leq \boldsymbol{d}^*\{. : N\} \Leftrightarrow \boldsymbol{d}^{*-}\{. : N\} \leq \boldsymbol{d}^-\{. : N\}$$

The following counter-example is to demonstrate that R2 is not generally satisfied by the above choice. In this context, we should note the use of weights.

Example 3. $N = 3$

i/j

	X				D				$\bar{D}$			
	1	2	3	4	1	2	3	4	1	2	3	4
1	7	6	6	3	5	0	2	2	0	2	1	1
2	6	7	5	3	4	1	1	2	1	1	2	1
3	2	8	7	1	0	2	3	0	5	0	0	3
4	5	6	4	4	3	0	0	3	2	2	3	0
min	2	6	4	1								
max	7	8	7	4								

Comparison of $i = 1$ and $i = 2$ excluding the last feature (column 4):

$$\varphi^*(1) = [25w_1/114 + 4w_3/126]/[4w_2/185 + w_3/126]$$
$$\varphi^*(2) = [15w_1/114 + w_2/185 + w_3/126]/[w_1/114 + w_2/185 + 4w_3/126]$$
$$\text{Choose: } w_1 = 0.02, w_2 = 0.70, w_3 = 0.28$$
$$\text{Then } \varphi^*(1) = [1.3275]/[1.7357] = 0.765 > \varphi^*(2) = [0.8813]/[1.2848] = 0.686.$$

Comparison of $i = 1$ and $i = 2$ including the last feature:
The moment we include another feature, we need to modify the weights. The rule is to do so without altering the relative weights of the sources already under the domain of the study. So we use an overhead of P collectively for these sources and attribute the weight $Q = 1 - P$ for the additional source. The new sets of weights are $[Pw_1, Pw_2, Pw_3, Q]$ in this order.

Therefore, the revised values of $\varphi(1)$ and $\varphi(2)$ are given by:

$$\tilde{\varphi}^*(1) = \frac{[P25w_1/114 + P4w_3/126 + Q4/35]}{[P4w_2/185 + Pw_3/126 + Q/35]}$$
$$\tilde{\varphi}^*(2) = \frac{[P15w_1/114 + Pw_2/185 + Pw_3/126 + Q4/35]}{[Pw_1/114 + Pw_2/185 + P4w_3/126 + Q/35]}$$

We now choose $P = 0.90$ to re-evaluate the performances of the first two locations. It turns out that $\tilde{\varphi}(1) = 1.26 < 1.34 = \tilde{\varphi}(2)$, which is a contradiction. Thus even a small amount of perturbation (in terms of an additional location which has only 10 per cent relative weight) may alter the relative performance of the two locations. As a matter of fact, a plot of the ratio of the two expressions as a function of P may clearly provide the set of values of P for reaching such a contradiction.

It would be interesting to note certain elementary facts regarding the formation of φ. We state below some results in this direction and provide examples/counter-examples. Note that the CIs described below are based on $\boldsymbol{d}$'s alone.

Result 1: If φ corresponds to a quadratic form (pd) in the elements of $\boldsymbol{d}$, R1 may be violated.
Example 2: $\boldsymbol{d} = [1,2,3]', \boldsymbol{d}^* = [4,4,4]'$ and $\varphi(.) =$ quadratic form in $\boldsymbol{d}$ involving a pd matrix $\boldsymbol{A}$, where $\boldsymbol{A} = \boldsymbol{P}'\boldsymbol{D}_q\boldsymbol{P}$, with q_1, q_2 and q_3 as the elements in the diagonal matrix $\boldsymbol{D}_q$ and with $\boldsymbol{P}$ as an orthogonal matrix with first row proportional to $\mathbf{1}'$. Let us choose Helmert's matrix for $\boldsymbol{P}$. Then, $\varphi(\boldsymbol{d}) = 12q_1 + 0.5q_2 + 1.5q_3$ while $\varphi(\boldsymbol{d}^*) = 48q_1$. Hence there are choices for the q's for which $\varphi(\boldsymbol{d})$ **exceeds** $\varphi(\boldsymbol{d}^*)$ even though $\boldsymbol{d} \leq \boldsymbol{d}^*$.

Result 2: Example of **best** location for the **whole set** which **ceases** to be the **best** for **every subset**.
Example 3: $N = 3, \boldsymbol{d} = [0.5, 0.5, 0.5]', \boldsymbol{d}_1 = [1.5, 0, 0]', \boldsymbol{d}_2 = [0, 1.5, 0]', \boldsymbol{d}_3 = [0, 0, 1.5]', \varphi(\boldsymbol{d}) = \boldsymbol{d}'\boldsymbol{d}$. **Here, $\boldsymbol{d}$ is best wrt all the three characters taken together but not wrt any subset of two.**

Result 3: Example of **best** location for the **whole set** which **continues** to be the **best** for **every subset**.
Example 4: $N = 3, \boldsymbol{d} = [0.5, 0.5, 0.5]', \boldsymbol{d}_1 = [1.5, 1, 0]', \boldsymbol{d}_2 = [0, 1.5, 1]', \boldsymbol{d}_3 = [1, 0, 1.5]', \varphi_{\boldsymbol{d}} = \boldsymbol{d}'\boldsymbol{d}$. **Here, $\boldsymbol{d}$ is best wrt all the three characters and continues to be the best wrt any subset of two.**

4. MODIFICATIONS

We now suggest two modifications to the TOPSIS-based CI and show that these satisfy the two requirements.

Modification I: We make use of d and d^- together instead of using them separately:

$$\varphi_i(\boldsymbol{d}, \boldsymbol{d}^-) = \left[\sum_j w_j (d_{ij}^2 / d_{ij}^{-2}) \Big/ \sum_i x_{ij}^2\right]^{1/2} + \left[\sum_j{}' w_j R_j^2 \Big/ \sum_i x_{ij}^2\right]^{1/2}$$

where $\sum$ refers to all j for which $d_{ij}^- > 0$ while $\sum'$ refers to all j for which $d_{ij}^- = 0$. Further, R_j is a finite quantity of our choice subject to $R_j \geq \max[d_{ij}/d_{ij}^-]$, the maximum being taken over all i for which $d_{ij}^- > 0$.

Modification II: We make use of d and d^- in additive form:

$$\begin{aligned}\varphi_i(\boldsymbol{d}, \boldsymbol{d}^-) = &\left[\sum_j \left\{ w_j d_{ij}^2 \Big/ \sum_i x_{ij}^2 \right\}\right]^{1/2} \\ &+ \left[\sum_j{}' \left\{ (w_j / d_{ij}^{-2}) \Big/ \sum_i x_{ij}^2 \right\}\right]^{1/2} \\ &+ \left[\sum_j{}'' \left\{ (w_i / r_j^2) \Big/ \sum_i x_{ij}^2 \right\}\right]^{1/2}\end{aligned}$$

where $r_j \leq \min\{d_{ij}^-\}$, the minimum being taken over all $d_{ij}^- > 0$. Again $\sum'$ refers to all j for which $d_{ij}^- > 0$ while $\sum''$ refers to all j for which $d_{ij}^- = 0$.

It follows that both the modified CIs satisfy R1 and R2. Verification of R2 is easy. To see how R1 works, we need to separate out Cases I–III as follows:

Case I: $d_{ij} > 0$; Case II: $0 = d_{ij} < d_{ij}^*$; Case III: $d_{ij} = d_{ij}^* = 0$.

5. NUMERICAL COMPUTATIONS

In this section, we intend to study the performance of the two suggested modifications for the CI. We do so by applying the two formulae on the data set examined by Ross and Sinha (2001) relating to air/water/land for 50 states of the U.S.A. In fact, we start with more general versions of the two formulae, as indicated below, by introducing a parameter k (which assumes the value 2 for the formulae given in the previous section).

Modification I: Generalized version:

$$\varphi_i(\boldsymbol{d}, \boldsymbol{d}^-) = \left[\sum_j w_j \left(d_{ij}^k / d_{ij}^{-k}\right) \Big/ \left[\sum_i x_{ij}^2\right]^{k/2}\right]^{1/2} + \left[\sum{}' w_j R_j^k \Big/ \left[\sum_i x_{ij}^2\right]^{k/2}\right]^{1/2}$$

where $\sum$ refers to all j for which $d_{ij}^- > 0$ while $\sum'$ refers to all j for which $d_{ij}^- = 0$.

Further, R_j is a finite quantity of our choice subject to $R_j \geq \max[d_{ij}/d_{ij}^-]$, the maximum being taken over all i for which $d_{ij}^- > 0$.

We may also work out Modification II along the same lines.

Next we consider the ranks of all the states as assessed by the two formulae using various values of k. Then we group the 50 states ranked according to the first formula into five sets of ten each. Thus, Set 1 will consist of the ten states ranked 1 to 10 according to this formula. Next, we do the same according to the second formula and then for each set compute the intersection numbers. These numbers are shown in the tables below. Table 1 deals with small values of k and Table 2 deals with values of k centered around 2.

It is seen that the values of k have a role to play. It turns out that the TOPSIS-based value $k = 2$ does not seem to be adequate.

Thus, one may use Modifications I and II with a value of k around 0.20.

We have also studied the Electre method. It turns out that this also suffers from certain violations of basic requirements. The details are outlined in the next section.

Further research is, therefore, warranted in this fascinating area.

6. ELECTRE METHOD

The Electre method requires more extensive computations than the TOPSIS-based method. It is used for comparing the status of two locations rather than ranking all of them together like the previous one.

Table 1. Studies of the agreement between two measures of composite indexing (small k)

Serial number	Set 1	Set 2	Set 3	Set 4	Set 5	k
1	10	9	8	9	9	0.01
2	10	10	9	9	9	0.02
3	9	9	9	8	9	0.03
4	9	9	9	8	8	0.04
5	9	9	9	8	8	0.05
6	9	9	9	8	8	0.10
7	9	9	9	8	8	0.15
8	9	9	9	8	8	0.20
9	9	9	9	8	8	0.25

Table 2. Studies of the agreement between two measures of composite indexing (k around 2)

Serial number	Set 1	Set 2	Set 3	Set 4	Set 5	k
1	5	2	4	6	4	1.95
2	5	2	4	6	4	2.00
3	5	3	5	7	4	2.05
4	5	3	4	7	4	2.10
5	5	2	4	6	4	2.15

Starting with the same $\mathbf{X}_{K\times N}$ matrix of observations on pollution indices and the weight vector $\mathbf{w}$ we proceed as follows:

Step 1: Transform $\mathbf{X} = [\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X}_N]$ to $\mathbf{R} = [\mathbf{R_1}, \mathbf{R_2}, \ldots, \mathbf{R}_N]$, where

$$R_i = \frac{X_i}{||X_i||^2}$$

Step 2: Transform $\mathbf{R}$ to $\mathbf{V} = \mathbf{RW}$, where $W = \text{diag}(w_1, w_2, \ldots, w_N)$.
Step 3: Based on $\mathbf{V}$, construct two matrices: $\mathbf{C} = ((c_{ij}))$, where

$$c_{ij} = \sum_{k:v_{ik}\le v_{jk}} w_k$$

and $\mathbf{D} = ((d_{ij}))$, where

$$d_{ij} = \frac{\max_{k:v_{ik}<v_{jk}}|v_{ik} - v_{jk}|}{\max_k |v_{ik} - v_{jk}|}$$

$\mathbf{C}$ and $\mathbf{D}$ are called concordance and discordance matrices, respectively.
Step 4: Compute

$$\overline{c} = \frac{\sum\sum_{i\neq j} c_{ij}}{K(K-1)}$$

and

$$\overline{d} = \frac{\sum\sum_{i\neq j} d_{ij}}{K(K-1)}$$

and construct matrices $\mathbf{F}$ and $\mathbf{G}$ such that

$$f_{ij} = \begin{cases} 1 & \text{when} \quad c_{ij} \le \overline{c} \\ 0 & \text{o.w.} \end{cases}$$

$$g_{ij} = \begin{cases} 1 & \text{when} \quad d_{ij} \le \overline{d} \\ 0 & \text{o.w.} \end{cases}$$

Step 5: Define matrix $\mathbf{E} = ((e_{ij}))$, where $e_{ij} = f_{ij}.g_{ij}$.
$e_{ij} = 1$ implies that location i is better than location j.

Apart from the disadvantage of not giving the relative ranking of the locations simultaneously, there is another serious drawback of the Electre method. It can be checked that $e_{ij} = 1$ does not ensure that $e_{ji} = 0$, and we can get instances where both e_{ij} and e_{ji} are 1, which makes the two locations i and j incomparable.

One may wish to examine whether this method satisfies the two requirements. It is easy to see that R1 is satisfied, but for R2 we have a counter-example.

Example 5:

$$X = \begin{bmatrix} 7 & 6 & 7 & 3 \\ 6 & 7 & 6 & 3 \\ 2 & 8 & 7 & 1 \\ 5 & 6 & 4 & 4 \end{bmatrix}$$

Comparison of $i = 1$ and $j = 2$ excluding the last character, i.e. column 4:
If we take $\mathbf{w} = (\mathbf{0.02}, \mathbf{0.55}, \mathbf{0.43})$, computations yield $\bar{c} = 0.58, \bar{d} = 0.564, c_{12} = 0.55$ and $d_{12} = 1.00$, so $f_{12} = 0$ and $g_{12} = 1$, which leads to $e_{12} = 0$.

Comparison of $i = 1$ and $j = 2$, including the last character, i.e. column 4:
(Note that the 4th character for the 1st and 2nd rows are the same.)

Following the procedure of Example 4 we took the new weight vector as $\mathbf{w} = (0.02P, 0.55P, 0.43P, Q)$, where $Q = (1 - P)$. Now even for Q as low as 0.08, e_{12} changes from 0 to 1. The corresponding changed entries in the concordance and discordance matrices are $\bar{c} = 0.5818, \bar{d} = 0.62$, $c_{12} = 0.586$ and $d_{12} = 1.00$, so $f_{12} = 1$ and $g_{12} = 1$, which yields $e_{12} = 1$.

ACKNOWLEDGEMENTS

We sincerely thank Professor Bimal K. Sinha of the University of Maryland Baltimore County for useful discussions on an earlier draft of this article and for the data set analyzed in Section 5. We are also thankful to Mr. Biresh Giri, M. Stat. Final year Student at the Indian Statistical Institute, Calcutta, for carrying out the computations in this article. At the suggestion of the first author, Mr. Giri worked on a project dealing with these computations.

REFERENCES

Filar JA, Ross NP, Wu M-L. 1997. *US EPA Report* on 'Environmental assessment based on multiple indicators'.
Hwang CL, Yoon K. 1981. *Multiple Attribute Decision-Making: Methods and Applications. A State-of-the-Art Survey.* Springer-Verlag.
Ross NP, Sinha BK. 2001. On some aspects of data integration techniques with applications. Unpublished Manuscript.
Yoon K, Hwang CL. 1995. *Multiple Attribute Decision Making: An Introduction.* Sage Publications.
Zeleny M. 1982. *Multiple Criteria Decision Making*. McGraw-Hill.