RASHI 1 (1) :5-10 (2016)

DATA INTEGRATION TECHNIQUES IN AGRICULTURAL SCIENCES

Mamunur Rashid and [1]Bikas K. Sinha

Department of Mathematics, DePauw University, Indiana, USA

[1]Retired Professor, Indian Statistical Institute, Kolkata, India

ABSTRACT

We intend to discuss some standard and non-standard Data Integration Techniques with an illustration from a recently published paper [3] on ranking of several DNA extraction methods for extraction of DNA from soil samples. The study was undertaken to carry out a relative comparison of several different methods of soil extraction under several distinct DNA analysis criteria. From a practical point of view, it is highly unlikely that a single DNA extraction strategy can be optimum for all the selected criteria. Hence there is a need for data integration to arrive at an overall ranking of the methods, keeping all the different judgment criteria in mind.

***Keywords*:** Agricultural Soil; Biochar; DNA Extraction; Multiple Criteria Decision Making; Technique for Order Preference by Similarity to Ideal Solution [TOPSIS]: Poultry Manure

1. Introduction

The key reference to this paper is [3]wherein the authors undertake a statistical study of ranking a given set of competing and alternative DNA extraction methods for agricultural soil, using TOPSIS Method [TM] - a specific Multiple Criteria Decision Making [MCDM] Algorithmic Tool/Technique. This method along with another less popular ELECTRE METHOD are thoroughly discussed in [2, 4].It is indeed commendable that the authors in [3]ventured in this relatively unexplored area of what is known as 'Data Integration Techniques'. There are certain typos in the published manuscript. However, that in no way takes away the credit to be attributed to the authors for so nicely discussing various features of the technique and its computational details.

We propose to undertake various theoretical / computational issues in the implementation of TM. The paper is organized as follows. In Section 2, we start with a brief description of the computational algorithm underlying the TM in a theoretical framework and this we do by borrowing the notations as in [3].This is geared towards providing maximum comfort to the readers - at least to those who are familiar with [3].Next we discuss some related practical issues in the same theoretical framework. In Section 3, we rework on the data set already described and analyzed in [3].In passing, we point out the computational mistakes in[3]. But, as we said before, the authors already earned a lot of credit by simply being familiar with TM. It is only natural that we, as statisticians, provide further insights into the intricacies of application of TM in real data, as was the purpose in [3].We close the paper with some remarks in Section 4.

2. TM : Computational Algorithm in a theoretical framework and related issues

Assume there are 'm' DNA extraction methods and there are 'n' criteria for evaluation of these methods. Each method is judged by the 'score' it receives after its application and subsequent evaluation with respect to each criterion.

We denote by $X = ((x_{ij}))$ the positive-valued score matrix of order representing the extraction methods as rows of the matrix X and the evaluation criteria as the columns of the matrix X. In order that an extraction method is adjudged the best with respect to a specific evaluation criterion, it is tacitly assumed that the score for this method has to exceed those of all others in the list. The objective of the study is to arrive at an 'over-all' ranking of the extraction methods, by taking into account their performance across all the evaluation criteria. It may so happen that the natural choice of one or more evaluation criterion lend themselves to 'minimum-the-best' criterion. In such a case, one suggestion is to change the scores for all extraction methods [across that column of the X-matrix] by taking their reciprocals. In fine, one has to ensure that all the scores for each evaluation criterion have the same interpretation in terms of 'max-to-min' going hand-in-hand with 'best-to-worst'. At times, the X-matrix is also termed as 'Decision Matrix'.

It is clear that for one single evaluation criterion, the ranking of extraction methods is trivial. Also as and when all the criteria values exhibit same relative positions of different extraction methods, the solution

Email: mrashid@depauw.edu

is easy to arrive at. Non-trivial situations arise when there are 'waive-like' patterns in the data and this is most expected scenario in practice with real data.

One natural and simple-minded approach has been to work out the average score for each method of extraction – by averaging the scores across all the evaluation criteria. That means, we simply compute the row averages in the *X*-matrix of scores and use them for ranking of the methods. There are obvious limitations to this approach since it does not take into account the variations among the scores [of different extraction methods] under each evaluation criterion. It deals with one method at a time. Apart from this, the point to be noted is that while we are working out the average score, we are assuming that all the evaluation criteria are equally important and hence they possess the same weights. This has been a point of concern to the scientists and the data analysts have worked out a solution to this problem. Naturally, we should call upon 'subject experts' and utilize their knowledge in ascertaining relative weights of the different evaluation criteria. Failing to have access to such experts' inputs, data-driven techniques have been suggested in theliterature. One such technique is based on Shannon Entropy Measure, nicely explained in [3].There are two other data-driven techniques applied in such cases.

We will discuss and apply all three techniques for evaluation of weights of different evaluation criteria. Once the weights are determined, the formulae for applying the weights and computing 'composite indices' for different extraction techniques are the same to arrive at their individual rankings.

We describe the necessary steps as are explained in the paper [3] with reference to Entropy Weight Measure.

Step 1. Transferring the decision matrix to the normalized mode

In order to compute the entropy measure for the j^{th} criterion, the related values in the decision matrix are first normalized as P_{ij} **:**

$$P_{ij} = \frac{x_{ij}}{\sum_{i=1}^{m} x_{ij}} \tag{1}$$

Step 2. Calculating the entropy of dataset for each criterion

In this step, the entropy of the j^{th} criterion, j^{th} criterion, E_j is calculated as follows:

$$E_j = -\alpha \sum_{i=1}^{m} p_{ij} \ln \quad (i = 1,2,3,...,m;\ j = 1,2,3,...,n) \tag{2}$$

Alternatives	***Criteria***	C_1	C_2	C_3	...	C_n
	(Weights	w_1	w_2	w_3	...	w_n)
A_1		x_{11}	x_{12}	x_{13}	...	x_{1n}
A_2		x_{21}	x_{22}	x_{23}	...	x_{2n}
A_3		x_{31}	x_{32}	x_{33}	...	x_{3n}
⋮		⋮				
A_m		x_{m1}	x_{m2}	x_{m3}	...	x_{mn}

Fig 1: A typical decision matrix in MCDM

where, $\alpha = 1/\ln(m)$; *"m"* is the total number of alternatives (in this study, the DNA extraction methods over different samples).

Next, the operation of subtraction is used to measure the degree of diversity relative to the corresponding anchor value (unity), D_j using the following formula:

$$D_j = 1 - E_j \tag{3}$$

Step 3. Defining criteria weights

The entropy weight *'W'* of each criterion is calculated using

$$W_j = \frac{D_j}{\sum_{j=1}^{n} D_j} \tag{4}$$

So far, we explained the steps in ascertaining the weights as per entropy measure. Another method is based on the notion of 'Coefficient of Variation' [CV] defined as CV=sd/mean. Weights are taken to be directly proportional to the respective CV's. We will also describe another method, known as 'method of reversal'.

Once the weights are chosen [by any convenient method], these weights are then incorporated into the so-called TM to calculate an overall score for each DNA extraction method. The TM was chosen because of its high speed, accuracy, and compatibility [5].The algorithm of this technique is summarized as follows:

1) Transfer the decision matrix to the normalized mode :

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} x_{ij}^2}} \qquad (i = 1,2,3,...,m;\ j = 1,2,3,...,n) \tag{5}$$

2) Weigh the normalized decision matrix :

$$v_{ij} = W_j \times r_{ij} \qquad (i = 1,2,3,...,m;\ j = 1,2,3,...,n) \tag{6}$$

3) Define the "ideal positive" V_j^+ and "ideal negative (nadir)" V_j^- solutions :

$$\begin{cases} \{V_1^+, V_2^+, \cdots, V_n^+\} = \{(\max_i V_{ij} \mid j \in J), (\min_i V_{ij} \mid j \in J') \mid i = 1,2,\cdots,m\} \\ \{V_1^-, V_2^-, \cdots, V_n^-\} = \{(\max_i V_{ij} \mid j \in J), (\min_i V_{ij} \mid j \in J') \mid i = 1,2,\cdots,m\} \end{cases} \quad (7)$$

Instead of making adjustments in the scores, in the definition of 'ideal' and 'anti-ideal', one can use maximum and minimum in the reverse direction. That is why the notations $'J'$ and J' have been introduced in the above. It is tacitly assumed that the 'J' – collection' corresponds to the right order and the J' - collection' corresponds to the reverse order.

4) Measure the distances, d_i^+ and $d_i^-, i = 1,2,3,...,m$ and from the ideal and negative ideal solutions :

$$\begin{cases} d_i^+ = \left[\sum_{j=1}^{n} (V_{ij} - V_j^+)^2\right]^{\frac{1}{2}}, i = 1,2,3...,m \\ d_j^+ = \left[\sum_{j=1}^{n} (V_{ij} - V_j^-)^2\right]^{\frac{1}{2}}, i = 1,2,3...,m \end{cases} \quad (8)$$

In (8), the 'distance measure' used is referred to as 'Euclidian distance' or 'Euclidian Norm', denoted by L_2

5) Determine the relative closeness of alternatives to ideal solution by computing what is known as Composite Index [CI]' :

$$CI_i = \frac{d_i^-}{d_i^- + d_i^+}, i = 1,2,3,...,m \quad (9)$$

These composite indices are used for final ranking of the methods, the rule being: max – to – min for ranks 1 – to –*m*.

3. Data and Results Based on the Analyses

Performances of eight different DNA extraction methods were studied under seven decision criteria C_1-C_7 for soil, SM [soil:manure, 99:1(w/w)] and SMB [soil:manure:biochar, 98:1:1(w/w)] and for this, CIs were computed in the original article. However, there were some computational mistakes in [3]. The original data set are shown in table 1.It was found that except for C_1, C_3 and C_6, all others had insignificant weight / effect in the overall rankingof the methods. Henceforth, we work with only these three features.

In what follows, we will deal with three different approaches for ascertaining the weights of the three features: C_1, C_3 and C_6. Entropy measure is explained in [3] and also in the above. Use of CV is a routine task. The third is 'Reversal Method' [1]as explained below.

1. Start with equal weights for all the *n*columns and rank the m rows following either L_1 or L_2 distance measure.

2. Reverse the role of rows and columns, and rank the columns using the overall indices of the rows derived in Step 1 as their weights.

3.Now rank the rows afresh using the overall indices of the columns derived in Step 2 as their weights.

Before proceeding further, we display the weights as determined by all the three methods for each of the data sets *viz.,* Soil, SM and SMB.

Below we also show the results under a different

Table 1: Table of wights using different methods

Weight	*Soil*			*SM*			*SMB*		
	C_1	C_3	C_6	C_1	C_3	C_6	C_1	C_3	C_6
Entropy	0.4206	0.4276	0.1517	0.2766	0.4251	0.2983	0.2882	0.3892	0.3226
CV	0.3950	0.3888	0.2162	0.3155	0.3843	0.3002	0.3225	0.3669	0.3106
Reversal	0.1667	0.5000	0.3333	0.1667	0.5000	0.3333	0.1667	0.5000	0.3333
Geometric mean	0.3149	0.4542	0.2310	0.2470	0.4391	0.3139	0.2528	0.4207	0.3265

Table 2: Original date on DNA extraction methods under seven decision criteria soil, SM, and SMB

Methods	Sample	C1	C2	C3	C4	C5	C6	C7
Ultra Clean	Soil	8.21	1.75	1.30	2.	2.	0.83	4.46
Conventional	SM	14.21	1.71	1.43	2	2	0.83	4.46
	SMB	13.03	1.77	1.21	2	2	0.83	4.46
Ultra Clean	Soil	5.59	1.63	1.22	1	2	0.92	4.46
AlternativE	SM	10.77	1.50	0.81	1	2	0.92	4.46
	SMB	9.40	1.53	0.88	1	2	0.92	4.46
Power Soil	Soil	6.76	1.67	1.28	1	2	1	5.54
Conventional	SM	10.52	1.96	1.67	1	2	1	5.54
	SMB	8.75	1.90	1.57	1	2	1	5.54
Power Soil	Soil	3.78	1.74	2.07	1	2	1.08	5.54
Alternative	SM	10.48	1.55	0.86	1	2	1.08	5.54
	SMB	11.97	1.55	0.96	1	2	1.08	5.54
Fast Spin	Soil	17.70	1.73	0.25	3	1	1.16	6.58
Conventional	SM	20.00	1.77	0.39	3	1	1.16	6.58
	SMB	23.87	1.84	0.60	3	1	1.16	6.58
Fast Spin	Soil	23.29	1.70	0.34	3	1	1.25	6.58
Alternative	SM	20.36	1.77	0.50	3	1	1.25	6.58
	SMB	21.45	1.79	0.47	3	1	1.25	6.58
	Soil							
E.Z.N.A	SM	4.33	1.67	0.43	3	1	3.17	4.38
Conventional	SMB	7.26	1.85	1.57	3	1	3.17	4.38
		7.69	1.87	1.70	3	1	3.17	4,38
E.Z.N.A	Soil	25.98	1.56	0,30	3	1	3.17	4.38
Alternative	SM	25,90	1.57	0.55	3	1	3.25	4.38
	SMB	15.73	1.65	0.47	3	1	3.25	4.38

Table 3: Original (as in [3])and revised rankings of DNA extraction methods using (L2, Entropy)

Methods	Soil		SM		SMB	
	Original	***Revised***	***Original***	***Revised***	***Original***	***Revised***
Ultra Clean Conventional	4	4	2	1	2	1
Ultra Clean Alternative	6	6	3	4	4	6
Power Soil Conventional	5	5	1	2	1	2
Power Soil Alternative	1	1	4	7	3	5
Fast Spin Contentional	7	7	6	8	5	3
Fast Spin Alternative	2	3	5	6	6	7
E.Z.N.A Conventional	8	8	7	3	7	4
E.Z.N.A Alternative	3	2	8	5	8	8

Note: [+]The calculation is based on C_1, C_3, and C_6, where C_6 is the reciprocal of the original data to make it in increasing order (larger is better)

form of the 'distance measure', called L_1 norm which corresponds to 'mean deviation'. Weight measures used correspond to Entropy, CV and Reversal Method [explained below].

Table 4: Rankings of DNA extraction methods using (L_1, Entropy)

Methods	Soil	SM	SMB
Ultra Clean Conventional	2	1	1
Ultra CleanAlternative	6	3	5
Power Soil Conventional	5	2	2
Power Soil Alternative	1	5	4
Fast Spin Contentional	7	7	3
Fast Spin Alternative	3	6	7
E.Z.N.A Conventional	8	4	6
E.Z.N.A Alternative	4	8	8

Table 5: Rankings of DNA extraction methods using weight obtained by CV method for both distance measures

Methods	Soil		SM		SMB	
	(L_2, CV)	(L_1, CV)	(L_2, CV)	(L_1, CV)	(L_2, CV)	(L_1, CV)
Ultra Clean Conventional	2	2	1	1	1	1
Ultra Clean Alternative	6	6	4	3	7	6
Power Soil Conventional	5	5	2	2	2	2
Power Soil Alternative	1	1	8	5	6	4
Fast Spin Contentional	7	7	7	6	3	3
Fast Spin Alternative	3	3	6	4	5	5
E.Z.N.A Conventional	8	8	3	7	4	7
E.Z.N.A Alternative	4	4	5	8	8	8

Table 6: Rankings of DNA extraction methods using weight obtained by Reversal method[++] for both distance measures

Methods	Soil		SM		SMB	
	(L_2, CV)	(L_1, CV)	(L_2, CV)	(L_1, CV)	(L_2, CV)	(L_1, CV)
Ultraclean Conventional	2	2	1	1	1	1
UltraCleanAlternative	4	4	4	4	5	5
PowerSoilConventional	3	3	2	2	1	1
PowerSoilAlternative	1	1	5	5	4	4
FastSpinContentional	7	6	7	7	6	6
FastSpinAlternative	5	5	6	6	7	7
E.Z.N.AConventional	8	8	3	3	3	3
E.Z.N.AAlternative	6	7	8	8	8	8

[++] *L_2 distance measure is initially used to calculate the weights; however, L_1 distance measure may also be used.*

Additionally, we calculate another weight obtained by the geometric mean based on the weights of entropy, CV, and reversal methods, and then the rankings are given for both distance measures.

Table 7: Rankings of DNA extraction methods using weight obtained by the geometric mean for both distances

Methods	Soil		SM		SMB	
	(L_2, CV)	(L_1, CV)	(L_2, CV)	(L_1, CV)	(L_2, CV)	(L_1, CV)
Ultra Clean Conventional	2	2	1	1	1	1
Ultra Clean Alternative	4	4	4	3	4	3
Power Soil Conventional	3	3	2	2	2	2
Power Soil Alternative	1	1	5	5	5	5
Fast Spin Contentional	7	7	8	7	8	7
Fast Spin Alternative	5	5	6	6	6	6
E.Z.N.A Conventional	8	8	3	4	3	4
E.Z.N.A Alternative	6	6	7	8	7	8

4. Conclusion

It this paper we discuss several techniques for ascertaining the over-all ranks of competing methods when judged against several alternative decision criteria. The weights [i.e., the relative importance] of the criteria are to be derived based on the 'data matrix'. Several methods for determination of the weights are discussed. Also two distinct distance measures are presented.

An illustrative example from a recent paper [3] is taken up for explanation of the computational details.

REFERENCES

Cascales, G. and Lamata, T. (2012) [1]: On Rank Reversal & TOPSIS Method. *Mathematical Computer Modelling.* **56:** 123-32.

Hwang, L. C., and Yoon, K. (1981) [2]: Multiple Attribute Decision Making Methods and Applications. Springer-Verlag, New York.

Pakpour, S., Snizhana, O., Prasher, Shiv.,Milani, A., and Chénier,M. (2013)[3]: DNA extraction method selection for agricultural soil using TOPSIS multiple criteria decision-making model.*Amer. J. Molecular Biology.* **3:** 215-28.

Yoon, K. P., and Hwang, C. (1995)[4]: Multiple attribute decision making: An introduction. *Sage University Paper series on Quantitative Applications in the Social Sciences,* 07-104. Thousands Oaks, CA: Sage.

Zou, Z. H., Yan, Y., and Sun, J. N. (2006)[5]: Entropy method for determination of weight of evaluating indicators in fuzzy synthetic evaluation for water quality assessment. *Journal of Environmental Sciences-China,* 18, 1020-1023.