Cilt: 10 Sayı: 03 Kategori: 03 Sayfa: 64-75 ISSN No: 1303-6319 Volume: 10 Number: 03 Category: 03 Page: 64-75

A K-NEAREST NEIGHBOR BASED APPROACH FOR DETERMINING THE WEIGHT RESTRICTIONS IN DATA ENVELOPMENT ANALYSIS

Elvan AKTURK HAYAT* **Olcay ALPAY****

ABSTRACT

Data Envelopment Analysis (DEA), a method commonly used to measure the efficiency is becoming an increasingly popular management tool. On the contrary to classical efficiency approaches, the most important advantage of DEA is that researchers can determine the weight restrictions of input and output variables. Variable selection and determination of weight restrictions are important issues in DEA. This work investigates the use of K-nearest neighbor (KNN) algorithm in the definition of weight restrictions for DEA. With this purpose a new approach based on KNN is proposed. Applications are constructed with empirical and real data sets depending on the specific constraints. Performance scores were calculated for both KNN based restricted and unrestricted DEA models and the results are interpreted.

Keywords: Data envelopment analysis, Efficiency, K-nearest neighbor, Weight restrictions.

1. INTRODUCTION

Data Envelopment Analysis (DEA) is a nonparametric technique for measuring the relative efficiency of a set of similar units, usually called the Decision Making Unit (DMU), which use a variety of identical inputs to produce a variety of identical outputs. DEA based on Frontier Analysis was introduced by Farrell in 1957, but the recent series of discussions started with the article by Charnes et al. (Charnes et al., 1978).

DEA provides efficiency score through linear programming when there are multiple inputs and outputs. One of the most important differences of DEA from the other efficiency measurement models is allowance to use input-output weights. In recent years, weight restrictions and value judgments have become one of the major issues in the DEA literature. The traditional DEA formulation allows for unrestricted model weights, which may result in inadequate weight values (zero, for instance, implying that a variable with relevance to the model would not be used for parameter estimation) (Gonçalves et.al., 2013). To deal with this kind of problem, Thompson et al. (1986) were the first to propose the use of weight restrictions in DEA. Many methods for estimating restrictions for the DEA weights have been developed by several researchers in the area, including Charnes et al. (1979); Charnes et al. (1985); Golany (1988); Thompson et al. (1990); Roll et al. (1991); Thanassoulis et al. (1995); Podinovski and Athanassopoulos (1998); Thanassoulis and Allen (1998); Podinovski (1999, 2001,

*Yrd. Doç. Dr. Sinop University, Faculty of Arts and Sciences, Department of Statistics, 57000, Sinop, Turkey, e-mail: elvanhayat@sinop.edu.tr
**Yrd. Doç. Dr. Sinop University, Faculty of Arts and Sciences, Department of Statistics, 57000, Sinop, Turkey, e-mail: olcayb@sinop.edu.tr

2004); Zhu (2003); Allen and Thanassoulis (2004). For a detailed review of such methods, see Allen et al. (1997) and Sarrico and Dyson (2004).

Jahanshahloo et al. used goal programming and Big M method techniques to obtain feasible weights for DMU's in 2005. Dimitrov and Sutton (2010) proposed symmetric weight assignment technique (SWAT) which does not affect feasibility. Mecit and Alp (2012) used correlation coefficients to determine the weights of inputs and outputs and also they compared this new method with cross efficiency evaluation model.

In our study, we proposed the use of a weight restriction technique based on the K-nearest neighbor (KNN) algorithm in order to define variation limits for the DEA model parameters. The KNN based restricted model is applied to three data sets and the results are compared with the unrestricted model.

The rest of paper is organized as follows. In the next section, the basic DEA model and the weight restrictions in DEA, also the proposed approach are briefly explained. In Section 3, the proposed approach and classical DEA model are applied to three data sets and the application results are reported. The first two data sets are from Roll et al.'s (1991) and Beasley's (1990) studies; the last one is a real data related to Turkey health system in 2013. Some concluding remarks are given in the final section.

2. METHODOLOGY

2.1 Data Envelopment Analysis

DEA does not require any assumptions about the functional form of the production function. In the simplest case of a unit having a single input and output, efficiency is defined as output/input. Charnes, Cooper and Rhodes, who developed Farrell's idea, extended the single-output/input ratio measure of efficiency to the multiple output/input measure of efficiency (Cooper et al., 2000).

The efficiency score in the presence of multiple input and output factors is defined as:

$$Efficiency = \frac{weighted\ sum\ of\ outputs}{weighted\ sum\ of\ inputs}$$

The first DEA model was introduced by Charnes et al. in 1978, known as the CCR model. This model measures the total efficiency under the assumption of constant returns to scale (CRS).

CCR is a linear program measuring the efficiencies of DMUs with respect to weighted inputs and outputs (Charnes et al., 1978). The model did not have any restrictions on the weights of inputs and outputs and found the optimal combination of weights that maximizes the efficiency score (Cooper et al., 1996).

Assume that there are n DMUs, each with m inputs and s outputs. The relative efficiency score of a DMU_p is obtained by solving the following proposed model (Charnes et al., 1994).

$$\max \frac{\sum_{k=1}^{s} v_k y_{kp}}{\sum_{j=1}^{m} u_j x_{jp}}$$

$$s.t. \quad \frac{\sum_{k=1}^{s} v_k y_{ki}}{\sum_{j=1}^{m} u_j x_{ji}} \leq 1 \qquad \forall i \tag{1}$$

$$\mathrm{v_k}, u_j \geq 0 \qquad \forall k,j$$

where,

$k = 1,2,..., s$
$j = 1,2,...,m$
$i = 1,2,...,n$

$y_{ki} =$ amount of output k produced by DMU i,
$x_{ji} =$ amount of input j utilized by DMU i,
$v_k =$ weight given to output k,
$u_j =$ weight given to input j.

The fractional program shown as (1) can be converted to a linear program as given in (2).

$$\max \sum_{k=1}^{s} v_k y_{kp}$$

$$s.t. \sum_{j=1}^{m} u_j x_{jp} = 1$$

$$\sum_{k=1}^{s} v_k y_{ki} - \sum_{j=1}^{m} u_j x_{ji} \leq 0 \qquad \forall i \tag{2}$$

$$\mathrm{v_k}, u_j \geq 0 \qquad \forall k,j.$$

The above problem is run n times in identifying the relative efficiency scores of all the DMUs. Each DMU selects input and output weights that maximize its efficiency score. In general, a DMU is considered to be efficient if it obtains a score of 1 and if it has a score of less than 1, it is implied as inefficient.

The weights given by the DEA model may be inconsistent with prior knowledge or accepted views on the relative values of the outputs and the inputs. DEA model can assign lower or higher weights to some inputs and/or outputs than they actually are. The model can give high weights for some inputs and/or outputs which give the impression that these attributes are over represented. As a result, the relative efficiency of a DMU

may not really reflect its performance on the inputs and outputs taken as a whole (Talaue et al., 2011).

In recent years, many new kinds of methods were proposed for weight restrictions such as analytic hierarchy process (AHP) and Delphi. A common characteristic of these approaches is based on specialists' own experiences and subjective judgment, to determine each of the indices that will be used to evaluate. The main disadvantage of this approach is that it is subjective (Allen et al., 1997).

Wong and Beasley (1990) proposed the use of proportions to introduce restrictions in the virtual inputs and outputs, seeking to make the quantification of value judgments easier for decision makers. Thus, they could set weights as varying, for instance, between 10% and 90% of the total contribution of inputs and outputs.

To constitute the weight restrictions in DEA, some methods such as assurance regions type, cone-ratio and absolute weight restriction were developed. In this research, assurance regions method is used to determine the weight restrictions.

2.2 KNN Based Algorithm

Many data mining techniques are based on similarity measures between objects. Measures of similarity may be obtained indirectly from vectors of measurements or characteristics describing each object (Hand et al., 2001).

The KNN prediction model simply stores the entire data set. As the name implies, to predict for a new observation, the predictor finds the *k* observations in the training data with feature vectors close to the one for which we wish to predict the outcome (Ye, 2003). Many applications of nearest neighbor methods adopt a Euclidean metric. Euclidean distance between i^{th} and j^{th} objects is defined as follows (Hand et al., 2001):

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

The nearest neighbor method has several attractive properties. It is easy to program and no optimization or training is required. Its classification accuracy can be very good on some problems, comparing favorably with alternative more unfamiliar methods (Hand et al., 2001). Also, nearest neighbor methods are very simple and therefore suitable for extremely large data sets (Felici and Vercellis, 2008). From a theoretical perspective, the nearest neighbor method is a valuable tool: as the design sample size increases, the bias of the estimated probability will decrease for fixed k (Hand et al., 2001).

In this study, we present a new algorithm based on KNN to determine the weight restriction matrices:

Step-0: Each of input/output numbers must be greater than 2.
Step-1: Determine the input and output variables.
Step-2: Construct the distance matrices using KNN for inputs and outputs.
Step-3: Find min and max values of matrices.
Step-4: Compute the decimal scale bandwidths which include all values in distance matrices.

Step-5: Determine the weight matrices according to bandwidths (in step-4) and relative to each other rates of variables.
Step-6: Construct the constraints with assurance regions method and calculate the efficiencies by DEA.

3. APPLICATION

In this section, three applications are illustrated and DEA is performed using LINDO (Linear, Interactive, and Discrete Optimizer) program. In the first application, we use the data of Roll et al. (1991), in the latter we use the data of 52 universities in Beasley (1990). In the last, we use the selected health statistics data for 12 statistical regions in Turkey. CCR model is used to calculate the efficiency scores and assurance region method is used to compose the restrictions obtained from our KNN based approach. Also, we determine the weight restricted models and compare the efficiencies with unrestricted models.

Roll et al.'s (1991) data consists of 10 DMUs with 3 inputs and 2 outputs. Table 1 summarizes the input and output variables for DMUs. The Euclidean distance matrix and the weight restrictions table for inputs are calculated by proposed KNN based algorithm. Finally, efficiency scores are calculated with DEA and given in Table 2. According to Table 2, 4 DMU reached 100% of efficiency in the unrestricted model, but 3 DMU achieved 100% in the proposed model. In unrestricted model, 2 DMU and in the proposed model 3 DMUs efficiency had less than 70%.

Table 1. Input-output variables (from Roll et. al.'s (1991) study)

DMU	I1	I2	I3	O1	O2
1	1.00	0.80	5.40	0.90	7.00
2	1.50	1.00	4.80	1.00	9.50
3	1.20	2.10	5.10	0.80	7.50
4	1.00	0.60	4.20	0.90	9.00
5	1.80	0.50	6.0	0.70	8.00
6	0.70	0.90	5.20	1.00	5.00
7	1.00	0.30	5.00	0.80	7.00
8	1.20	1.50	5.50	0.75	7.50
9	1.40	1.80	5.70	0.65	5.50
10	0.80	0.90	4.50	0.85	9.00

Table 2. Efficiency scores (%) for unrestricted and proposed model

	Efficiency (%)	
DMU	Unrestricted Model	Proposed Model
1	84.7	80.1
2	97.2	93.2
3	73.4	68.3
4	100.0	98.2
5	82.9	78.6
6	100.0	100.0
7	100.0	100.0
8	66.0	64.1
9	53.2	49.6
10	100.0	100.0

In Beasley's 1990 data set, there were 3 inputs and 8 outputs for 52 DMUs. Kocakoç (2003) used the same data set to determine the constraints for weight restrictions with analytic hierarchy process (AHP). We determine the weight restricted model and compare the efficiencies with AHP and unrestricted models. KNN based algorithm is performed on this data set, and the Euclidean distance matrices and the weight restriction tables are constructed for inputs and outputs. In Table 3 efficiency scores computed for 3 models are given. In the unrestricted model, there is no difference between the 52 DMUs in terms of efficiency. The results obtained from AHP and proposed models are quite similar. In the AHP model 39^{th} and 41^{th} DMU reached 100% of efficiency, whereas in the proposed model only 39^{th} achieved 100%. Average efficiency score is 71.52% in the AHP model, 69.65% in the proposed model.

In the last application, the selected health statistics data related to 12 statistical regions in Turkey consists of 3 inputs and 3 outputs. Physician number (per 100.000), beds number (per 10.000) and inpatient number (%) were taken as the inputs. Also, operation number (per 1000), mortality rate and average hospitalization days were taken as the outputs. Table 4, summarizes the input and output variables for DMUs. Calculated efficiency scores by without restrictions and proposed model are given in Table 5. In considering the efficiency scores obtained by each model, the average efficiency score of unrestricted model and our proposed model is 98.91% and 87.58%, respectively.

Table 3. Efficiency scores (%) for unrestricted, AHP and proposed models

	Efficiency (%)		
DMU	Unrestricted Model	AHP	Proposed Model
University 1	100	65.80	63.19
University 2	100	88.23	89.98
University 3	100	69.17	71.37
University 4	100	65.95	63.51
University 5	100	66.47	63.15
University 6	100	86.13	88.49
University 7	100	63.00	65.04
University 8	100	64.37	58.92
University 9	100	77.53	74.91
University 10	100	62.37	59.26
University 11	100	95.41	91.40
University 12	100	73.29	71.79
University 13	100	70.50	59.22
University 14	100	63.09	58.56
University 15	100	75.90	70.79
University 16	100	59.83	61.02
University 17	100	56.58	55.82
University 18	100	81.42	83.60
University 19	100	68.63	69.25
University 20	100	36.05	31.66
University 21	100	67.92	59.24
University 22	100	68.58	64.76
University 23	100	64.70	61.40
University 24	100	56.70	54.36
University 25	100	58.63	57.38
University 26	100	60.24	59.56
University 27	100	62.38	61.12
University 28	100	78.01	77.58
University 29	100	62.18	58.70
University 30	100	82.02	82.01
University 31	100	89.04	85.02
University 32	100	76.68	75.37
University 33	100	56.18	44.85
University 34	100	74.44	75.10
University 35	100	82.23	80.63
University 36	100	79.01	75.63
University 37	100	66.68	61.87
University 38	100	69.33	65.34
University 39	100	100.00	100.00
University 40	100	68.09	69.48
University 41	100	100.00	99.20
University 42	100	82.55	86.06
University 43	100	74.48	73.68
University 44	100	76.40	72.77
University 45	100	69.35	67.82
University 46	100	52.66	55.15
University 47	100	87.84	84.85
University 48	100	71.07	72.24
University 49	100	78.95	83.64
University 50	100	78.37	73.79
University 51	100	60.71	62.62
University 52	100	74.03	69.72

Table 4. Input-output variables for statistical regions in Turkey

	Inputs			Outputs		
Regions	**# Physician (per 100.000)**	**# Beds (per 10.000)**	**# Inpatient (%)**	**# Operation (per 1000)**	**Mortality rate**	**Average hospitalization days**
Akdeniz	161	23.8	54	66.2	15.5	4.2
Ege	191	27.4	60	61.2	18.7	4.4
Batı Anadolu	274	34.4	53	77.3	16.9	5.0
Güneydoğu Anadolu	124	20.2	63	52.9	11.2	3.4
Batı Karadeniz	156	29.8	67	56.5	18.2	4.9
İstanbul	184	23.4	43	59.6	15.3	5.0
Kuzeydoğu Anadolu	148	29.5	68	56.0	10.7	4.2
Batı Marmara	154	27.2	66	48.1	21.3	4.3
Ortadoğu Anadolu	146	27.7	60	53.6	6.4	4.1
Doğu Karadeniz	160	32.6	64	58.7	19.3	4.7
Doğu Marmara	160	25.8	61	62.5	18.0	4.2
Orta Anadolu	164	27.6	54	65.2	13.2	3.9

Resource: T.C. Minister of Health, Health Statistics Year Book, 2013

Table 5. Efficiency scores (%) for unrestricted and proposed model

	Efficiency (%)	
Regions (DMUs)	**Unrestricted Model**	**Proposed Model**
Akdeniz	100	100
Ege	97	80
Batı Anadolu	100	70
Güneydoğu Anadolu	100	100
Batı Karadeniz	100	88
İstanbul	100	81
Kuzeydoğu Anadolu	96	89
Batı Marmara	100	79
Ortadoğu Anadolu	96	86
Doğu Karadeniz	100	89
Doğu Marmara	100	95
Orta Anadolu	98	94

4. CONCLUSION

In this study, a new approach is proposed for determining the weight restrictions in DEA without any information or expert opinion about constraints. This approach is based on using K-nearest neighbor method establishing of constraint conditions and has several advantages. Firstly, this is a new kind of approach to determine the weight restrictions; it's easy to implement as well. Another advantage of this model is that it does not require expert opinions or value judgments.

Applications are performed to demonstrate the use of the proposed model and calculated efficiency scores for unrestricted model and the proposed model with using different data sets. The first data set, which consists of 10 DMUs with 3 inputs and 2 outputs, is obtained by Roll et al. (1991) study. The second data set from Beasley (1990) consists of 3 inputs and 8 outputs for 52 DMUs. Lastly, in real data application we used the selected health statistics for 12 DMUs in Turkey. As it can be seen from the results of application, the efficiency scores obtained from the proposed and restricted model based on AHP are quite similar. Thus, our proposed model can identify these restrictions objectively if there is no pre-information about weight restrictions. Undoubtedly, it cannot be expected to obtain such results in each time and every application. In a future study the real performance of our model can be evaluated by using simulation study.

5. REFERENCES

Allen R., Athanassopoulos A., Dyson R. G., Thanassoulis, E., 1997. Weights Restrictions and Value Judgements in Data Envelopment Analysis: Evolution, Development and Future Directions, Annals of Operations Research, 73:13 – 34.

Allen R., Thanassoulis E., 2004. Improving Envelopment in Data Envelopment Analysis, European Journal of Operational Research, 154, 363–79.

Beasley J. E., 1990. Comparing University Departments, Omega, International Journal of Management Science 18:171 - 183.

Charnes A., Cooper W. W., Rhodes E., 1978. Measuring the Efficiency of Decision Making Units, European Journal of Operational Research, Vol.2 No: 6, 429-444.

Charnes A., Cooper W. W., Rhodes E., 1979. Short Communication: Measuring the Efficiency of Decision Making Units, European Journal of Operational Research, 3, 339.

Charnes A., Cooper W. W., Golany B., Seiford, L., 1985. Foundations of Data Envelopment Analysis for Pareto-Koopmans Efficient Empirical Production Functions, Journal of Econometrics, 30, 91–108.

Charnes A., Cooper W. W., Lewin A. Y., Seiford L. M. (Eds.)., 1994. Data Envelopment Analysis: Theory, Methodology, and Applications, Boston: Kluwer.

Cooper W. W., Thompson R. G., Thrall R. M., 1996. Introduction: Extensions and New Developments in Data Envelopment Analysis, Annals of Operations Research 66(1):1-45.

Cooper, W. W., Seiford, L. M., Tone, K., 2000. Data Envelopment Analysis, Kluwer Academic Publishers, Boston, USA.

Farrell M. J., 1957. The Measurement of Productive Efficiency, Journal of the Royal Statistical Society, Series A, 120:253–81.

Dimitrov S., Sutton W., 2010. Promoting symmetric weight selection in data envelopment analysis: A penalty function approach, European Journal of Operational Research, 200, 281-288.

Felici G., Vercellis C., 2008. Mathematical Methods for Knowledge Discovery and Data Mining, Information Science Reference, Hershey, New York.

Golany B., 1988. A Note on Including Ordinal Relation among Multipliers in DEA, Management Science, 34, 1029–33.

Gonçalves A. C., Almeida R. M. R. V., Lins M. P. E., Samanez C. P., 2013. Canonical Correlation Analysis in the Definition of Weight Restrictions for Data Envelopment Analysis.

Hand D. J., Mannila H., Smyth P., 2001. Principles of Data Mining, MIT Press.

Jahanshahloo G. R., Memariani A., Hosseinzadeh F., Shoja N., 2005. A feasible interval for weights in data envelopment analysis, Applied Mathematics and Computation, 160, 155–168.

Kocakoç İ. D., 2003. Veri Zarflama Analizi'ndeki Ağırlık Kısıtlamalarının Belirlenmesinde Analitik Hiyerarşi Sürecinin Kullanımı, D.E.Ü.İ.İ.B.F.Dergisi 18(2):1-12.

Mecit E. D., Alp İ., 2012. A New Restricted Model Using Correlation Coefficients as an Alternative to Cross-Efficiency Evaluation in Data Envelopment Analysis, Hacettepe Journal of Mathematics and Statistics, 41(2), 321-335.

Podinovski V. V., Athanassopoulos A. D., 1998. Assessing the Relative Efficiency of Decision Making Units using DEA Models with Weight Restrictions, The Journal of the Operational Research Society, 49, 500–08.

Podinovski V. V., 1999. Side Effects of Absolute Weight Bounds in DEA Models, European Journal of Operational Research, 115, 583–95.

Podinovski V. V., 2001. DEA Models for the Explicit Maximisation of Relative Efficiency, European Journal of Operational Research, 131, 572–86.

Podinovski V. V., 2004. Production Trade-offs and Weight Restrictions in Data Envelopment Analysis, The Journal of the Operational Research Society, 55,1311–22.

Roll Y., Cook W. D., Golany B., 1991. Controlling Factor Weights in Data Envelopment Analysis, IIE Transactions, 23:1, 2-9.

Sarrico C. S., Dyson R. G., 2004. Restricting Virtural Weights in Data Envelopment Analysis, European Journal of Operational Research, 159, 17–34.

Thanassoulis E., Boussofiane A., Dyson R. G., 1995. Exploring Output Quality Targets in the Provision of Perinatal Care in England using Data Envelopment Analysis, European Journal of Operational Research, 80, 588.

Thompson R. G., Singleton, F. D., Thrall R. M., Smith B. A., 1986, Comparative site evaluations for locating a high-energy physics lab in Texas, Interfaces 16:35-49.

Thompson R. G., Langemeier L. N., Lee C. T., Lee E., Thrall R. M., 1990. The Role of Multiplier Bounds in Efficiency Analysis with Application to Kansas Farming, Journal of Econometrics, 46, 93–108.

Talaue C. O., Diesta N. A. N., Tapia C. G., 2011. Weights Restriction by Multiple Decision Makers in Data Envelopment Analysis Using Fuzzy Programming, Proceedings of the 11th Philippine Computing Science Congress, Ateneo de Naga University.

Ye N., 2003. The Handbook of Data Mining, Lawrence Erlbaum Associates, Publishers Mahwah, New Jersey, London.

Zhu J., 2003. Imprecise Data Envelopment Analysis (IDEA): A Review and Improvement with an Application, European Journal of Operational Research, 144, 513–29.

VERİ ZARFLAMA ANALİZİNDE AĞIRLIK KISITLARININ BELİRLENMESİNDE K-EN YAKIN KOMŞULUĞA DAYALI BİR YAKLAŞIM

ÖZET

Genellikle etkinlik ölçümünde kullanılan Veri Zarflama Analizi (VZA), popüler bir yönetim aracı olmaya başlamıştır. Klasik etkinlik yaklaşımlarının tersine, VZA'nın en önemli avantajı, girdi ve çıktı değişkenlerinin ağırlık kısıtlarını araştırmacıların belirleyebilmesidir. Değişken seçimi ve ağırlık kısıtlarının belirlenmesi VZA' da önemli konulardır. Bu çalışma VZA için ağırlık kısıtlarının tanımlanmasında K-en yakın komşuluk algoritmasının kullanımını araştırmaktadır. Bu amaçla K-en yakın komşuluk temeline dayanan yeni bir yaklaşım önerilmiştir. Belirlenen kısıtlara bağlı olarak ampirik ve gerçek veri setleri ile uygulamalar yapılmıştır. K-en yakın komşu temelinde kısıtlı model ve ağılık kısıtlamasız VZA modeli için performans skorları hesaplanmıştır ve sonuçlar yorumlanmıştır.

Anahtar Kelimeler: Ağırlık kısıtları, Etkinlik, Veri zarflama analizi, K-en yakın komşuluk.