Proc. of the Third Intl. Conf. on Advances in Economics, Management and Social Study - EMS 2015

ISBN: 978-1-63248-058-3 doi: 10.15224/ 978-1-63248-058-3-95

NMM-StoNED: a normal mixture model based stochastic semi-parametric benchmarking method

Xiaofeng Dai

***Abstract*—This paper presents a novel benchmarking tool, NMM-StoNED, which identifies the best practices closely located with each decision making unit (DMU) in the input-output space. Unlike the conventional techniques such as DEA where the success recepies of the benchmarks may not be transferable to all DMUs given their differences in, e.g., the operational scales, best practices identified by this method do not suffer from these problems and offer more practical values. NMM-StoNED is a specific configuration of the clustering and efficiency estimation algorithms in the benchmarking framework previously presented. This combination is able to cluster DMUs into less ambiguous groups and model the inefficiencies in a stochastic semi-nonparametric framework, which produces more accurate results than conventional benchmarking techniques such as DEA or other combinations such as the integration of K-means and StoNED. The performance comparison between NMM-StoNED and DEA has previously been reported, and the superiorities of StoNED over other productive efficiency analysis methods have been thoroughly investigated. Here we focus on showing the advantages of NMM in the clustering based benchmarking framework, for which, an empirical study using the Finland energy regulation data was conducted. This study contributes in its systematic evaluations on the performance of NMM-StoNED under various conditions which provide solid specifications on this algorithm, availing its practical use.**

***Keywords*—benchmarking, normal mixture model (NMM), data envelopment analysis (DEA), stochastic semi-nonparametric envelopment of data (StoNED)**

I. Introduction

Benchmarking, the process of comparing the performance of one decision making unit (DMU) against that of the DMUs with the 'best practice', has multiple applications, including offering the general insight of a given business sector, facilitating the manager on decision making, and providing the backbone of incentive provision for the regulators in the context of multiple agents [1]. DEA (data envelopment analysis) is conventionally applied in benchmarking, where the intensity weights strictly positive from the frontier estimation are considered as the best practices in benchmarking [2]. However, the success formula of the benchmarks identified may not be transferrable to a given DMU if they differ greatly on their, e.g., input and output structure. Also, as a deterministic method geared towards efficiency estimation, DEA doesn't take consider the stochasticity in its modelling framework. Thus, DEA is sensitive to both the heterogeneity and random noise of the DMUs in benchmarking.

Xiaofeng Dai
JiangNan University
China P. R.
Email: xiaofeng.dai@me.com

We have proposed a clustering based benchmarking framework in [4], where it segments the DMUs into groups based on user-specified metrics (e.g., the input-output vectors or their projections on the estimated frontier) using a clustering technique, and the benchmark(s) are identified according to the efficiency scores estimated from productive efficiency analysis within each cluster. We have shown that such a framework is flexible in choosing the clustering and efficiency analysis algorithms and these problems could be efficiently solved if the method at each step is appropriately selected. However, for what combination this method achieves the best performance is still left for discussion.

Typical clustering approaches can be classified into three categories, i.e., the hierarchical methods, the partitioning methods, and the model-based methods [5]. Hierarchical algorithms recursively combines or splits a set of objects into bigger or smaller groups based on a certain distance measurement and stops when meeting a certain criterion [6]. Methods of this class are conceptually intuitive and computationally simple which, however, could not determine the number of groups automatically, needs expert domain knowledge to define the distance measurement and is problem-specific. Partitioning methods iteratively reallocate data points across groups until no further improvement is obtainable [5], [11], with K-means being the most representative algorithm of this class [11]. Partitioning methods are widely used due to their computational simplicity and nonparametric structure which, however, needs pre-specification of the number of clusters. Model-based techniques optimise the fitness between the data and the model where the data is assumed to be generated [14]. Model based methods are superior over other methods in their automatic determination of the number of clusters, robustness to outliers, and probabilistic nature [5]. Among others, NMM (normal mixture model) is the most widely applied method of this class since normal distribution is the most commonly encountered distribution in practice.

Traditional productive efficiency analysis methods can be grouped based on two properties, i.e., parametric or nonparametric, and deterministic or stochastic. Many statistical methods can be used for productive efficiency analysis, with the most widely applied being DEA and SFA (stochastic frontier analysis), where DEA is non-parametric but deterministic and SFA is stochastic but parametric [8]. StoNED (stochastic semi-nonparametric envelopment of data) is a recently developed technique that melds the merits of DEA and SFA where the inefficiencies are estimated in a stochastic semi-nonparametric fashion. Unlike the semiparametric variate of

Proc. of the Third Intl. Conf. on Advances in Economics, Management and Social Study - EMS 2015

ISBN: 978-1-63248-058-3 doi: 10.15224/ 978-1-63248-058-3-95

SFA, StoNED builds directly on the axioms of the production theory such as free disposability and convexity instead of making any assumptions on the functional form or smoothness [9]. On the other hand, StoNED uses information of all observations in the data set to estimate the frontier rather than a few influential ones as adopted by DEA, making it less sensitive to outliers than DEA besides its insensitivity to the random noise.

Given the advantages of NMM and StoNED in clustering and efficiency estimation, respectively, we are motivated to fit these two algorithms in the clustering based benchmarking framework presented in [4]. This method, named NMM-StoNED here, detects the heterogeneous structure of the data, groups similar DMUs into unambiguous clusters, and ranks them within each cluster by the estimated efficiencies according to which the best practice is identified for each group. The superiorities of NMM-StoNED over DEA have been demonstrated in [4] using Finland energy regulation data from EMA (Energy Market Authority), and the advantages of StoNED over other efficiency analysis methods such as DEA and SFA have been studied in [10]. Here we focus on evaluating the performance of combining NMM with StoNED as compared with integrating other clustering techniques with StoNED in benchmarking. For this, we compared NMM with K-means, the most widely applied clustering technique due to its simple yet powerful features, in this clustering based benchmarking framework with an empirical study.

The rest of paper is organized as 'Method', 'Empirical study' and 'Conclusion'. The technical details of NMM and StoNED are described in the 'Method' section. In the 'Empirical study', the 'Data and methods' and 'Results and discussion' are described by sub-sections. The 'Conclusion' section finalizes this paper by summarizing the work and main contributions, and pointing out the future direction.

II. Method

The proposed method, NMM-StoNED, combines the NMM and StoNED into a unified framework. One can either measure the efficiencies of all DMUs using the whole data set before clustering, or compute the efficiencies using segment frontier after clustering if the number of DMUs in each cluster is sufficiently large [4]. The first alternative was used here given the limited size of our empirical data. The estimation process comprises of 1) estimating the efficiencies of all DMUs from the whole data set using StoNED; and 2) clustering DMUs using NMM and identifing the best practices in each group.

A. Efficiency estimation using StoNED

Given the standard multiple-input $\mathbf{r}_i$, single-output y_i, cross-sectional productive efficiency analysis model $y_i = f(\mathbf{r}_i) - u_i + v_i, \forall i = 1, \ldots, N$. where f satisfies monotonicity and concavity, $u_i > 0$ is an asymmetric inefficiency term and v_i is a stochastic noise term, StoNED uses a two-stage strategy in efficiency estimation [9]. In Stage 1, the shape of the function f is estimated by convex nonparametric least squares (CNLS) regression. In Stage 2, the inefficiency u is computed from the variances (σ_u^2, σ_v^2), which are estimated based on the skewness of the CNLS residuals (obtained from Stage 1) using, e.g., the method of moments. In the second stage, additional distributional assumptions are typically assumed, including, e.g., the asymmetric distribution for u_i with positive mean μ and finite variance σ_u^2, and a symmetric distribution for v_i with zero mean and constant finite variance σ_v^2.

Mathematically, the first stage is equivalent to (1) to (4) [9],

$$\min_{\boldsymbol{v},\boldsymbol{\alpha},\boldsymbol{\beta}} \sum_{i=1}^{n} \varepsilon_i^2 \quad \text{such that} \tag{1}$$

$$y_i = \alpha_i + \boldsymbol{\beta}_i^{'}\mathbf{r}_i + \varepsilon_i \tag{2}$$

$$\alpha_i + \boldsymbol{\beta}_i^{'}\mathbf{r}_i \leq \alpha_h + \boldsymbol{\beta}_h^{'}\mathbf{r}_i, \forall h, i = 1 \ldots n \tag{3}$$

$$\boldsymbol{\beta}_i \geq \mathbf{0}. \forall i = 1 \ldots n \tag{4}$$

where α_i and $\boldsymbol{\beta}_i$ are coefficients specific to observation i and v_i captures its random noise. In Stage 2, the inefficiency is computed using the distribution of the CNLS residuals $\hat{\varepsilon}_i$ (note that $\varepsilon = v_i + u_i$). Assuming that the inefficiency and noise terms follow the half-normal and normal distributions, respectively, the 2nd and 3rd central moments of the composite error distribution are

$$M_2 = [\frac{\pi - 2}{\pi}]\sigma_u^2 + \sigma_v^2, \quad M_3 = -(\sqrt{\frac{2}{\pi}})[\frac{4}{\pi} - 1]\sigma_u^3, \tag{5}$$

which can be estimated using the CNLS residuals

$$\hat{M}_2 = \sum_{i=1}^{n} (\hat{\varepsilon}_i - \bar{\varepsilon})^2/n, \quad \hat{M}_3 = \sum_{i=1}^{n} (\hat{\varepsilon}_i - \bar{\varepsilon})^3/n. \tag{6}$$

The standard deviations of the inefficiency and error term are then computed from

$$\hat{\sigma}_u = \sqrt[3]{\frac{\hat{M}_3}{(\sqrt{\frac{2}{\pi}})[\frac{4}{\pi} - 1]}}, \quad \hat{\sigma}_v = \sqrt{\hat{M}_2 - [\frac{\pi - 2}{\pi}]\hat{\sigma}_u^2}. \tag{7}$$

The conditional distribution of the inefficiency u_i given ε_i is a zero-truncated normal distribution with mean $\mu_\star = -\varepsilon_i\sigma_u^2/(\sigma_u^2 + \sigma_v^2)$ and variance $\sigma_\star^2 = \sigma_u^2\sigma_v^2/(\sigma_u^2 + \sigma_v^2)$. Let ϕ and Φ represent the standard normal density function and the standard normal cumulative distribution function, respectively, the inefficiencies are computed by $E(u_i|\varepsilon_i) = \mu_\star + \sigma_\star[\frac{\phi(-\mu_\star/\sigma_\star))}{1-\Phi(-\mu_\star/\sigma_\star)}]$.

B. Cluster-specific benchmark identification using NMM

In this step, the metrics dominating the heterogeneity of the data and following (or convertible to) normal distribution were specified and used as the input of NMM. If the input does not follow normal distribution or is a composite of multiple distributions, the mixture model of the corresponding distribution or a joint mixture model [3] would be required.

Assume that each observation $\mathbf{r}$ is drawn from g mixed normal distributions where, for each normal distribution f_i, it has the prior probability π_i and parameters θ_i, NMM optimises the fitness between the data and model $f(\mathbf{r}; \Theta) = \sum_{i=1}^{g} \pi_i f_i(\mathbf{r}; \boldsymbol{\theta}_i)$. Note that $\Theta = \{(\pi_i, \boldsymbol{\theta}_i) : i = 1, \ldots, g\}$ denotes all unknown parameters, $0 \leq \pi_i \leq 1$ for any i and $\sum_{i=1}^{g} \pi_i = 1$. Expectation Maximization (EM) algorithm is used to iteratively estimate the parameters by maximising the data log-likelihood $\log L(\Theta) = \sum_{j=1}^{N} \log([\sum_{i=1}^{g} \pi_i f_i(\mathbf{r}_j; \boldsymbol{\theta}_i)])$, where $R = \{\mathbf{r}_j : j = 1, \ldots, N\}$ and N is the total number of observations. The problem is casted in the framework of incomplete data using a

Proc. of the Third Intl. Conf. on Advances in Economics, Management and Social Study - EMS 2015

ISBN: 978-1-63248-058-3 doi: 10.15224/ 978-1-63248-058-3-95

dummy variable I_{ji} to indicate whether $\mathbf{r}_j$ comes from component i. Thus, $\log L_c(\Theta) = \sum_{j=1}^{N}\sum_{i=1}^{g} I_{ji}\log\left(\pi_i f_i(\mathbf{r}_j;\boldsymbol{\theta}_i)\right)$. At the m^{th} iteration of the EM algorithm, the E (expectation) step computes the expectation of the complete data log-likelihood Q

$$\begin{aligned} Q(\Theta;\Theta^{(m)}) &= E_{\Theta^{(m)}}(\log L_c|R) \\ &= \sum_{j=1}^{N}\sum_{i=1}^{g}\tau_{ji}^{(m)}\log\left(\pi_i f_i(I_j;\boldsymbol{\theta}_i)\right), \end{aligned} \tag{8}$$

and the M (maximisation) step updates the parameter estimates to maximize Q. The algorithm is iterated until convergence. Note that I's are replaced with τ's in (8), where $\tau_{ji} = E[I_{ji}|\mathbf{r}_j, \hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_g; \hat{\pi}_1, ..., \hat{\pi}_g]$. The set of parameter estimates $\left\{\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_g; \hat{\pi}_1, ..., \hat{\pi}_g\right\}$ is a maximizer of the expected log-likelihood for given τ_{ji}'s, and each $\mathbf{r}_j$ is assigned to its component by $\{i_0|\tau_{ji_0} = \max_i \tau_{ji}\}$. In NMM, the probability density function of f_i is defined as $f_i(\mathbf{r}_j;\theta_i) = \frac{1}{(2\pi)^{\frac{p}{2}}|\boldsymbol{V}|^{\frac{1}{2}}}\exp\big(-\frac{1}{2}(\mathbf{r}_j-\mu_i)^T\boldsymbol{V}^{-1}(\mathbf{r}_j-\mu_i)\big)$. Note that $\boldsymbol{V} = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2)$, $|V| = \prod_{v=1}^{p}\sigma_v^2$ and p is the dimension of the observations, whose parameters are estimated iteratively over the following equations [12].

$$\hat{\mu}_i^{(m+1)} = \sum_{j=1}^{N}\tau_{ji}^{(m)}\mathbf{r}_j \Big/ \sum_{j=1}^{N}\tau_{ji}^{(m)}$$

$$\hat{V}_i^{(m+1)} = \sum_{j=1}^{N}\tau_{ji}^{(m)}(\mathbf{r}_j-\mu_i^{(m+1)})(\mathbf{r}_j-\mu_i^{(m+1)})^T \Big/ \sum_{j=1}^{N}\tau_{ji}^{(m)}$$

$$\hat{\pi}_i^{(m+1)} = \sum_{j=1}^{N}\tau_{ji}^{(m)}/N \qquad \tau_{ji}^{(m)} = \frac{\pi_i^{(m)} f_i(\mathbf{r}_j;\theta_i^{(m)})}{\sum_{i=1}^{g}\pi_i^{(m)} f_i(\mathbf{r}_j;\theta_i^{(m)})}$$

Bayesian information criterion (BIC) [13], the most widely used model selection method, was used here to determine the best fitting model as well as the optimal number of clusters if not particularly specified BIC $= -2\log L(\hat{\theta}) + d\log(pN)$, where d represents the number of free parameters.

Once the DMUs are properly segregated, we rank the DMUs within each cluster by their efficiencies, and the best practice(s) within each cluster are considered the benchmarks of other units belonging to this group. As pointed out in [4], the 'best practice' may not achieve 100% efficiency, and is called the 'relative benchmark' to differentiate it from the 'absolute benchmark' which achieves, and more than one 'absolute benchmark' may exist for one group if multiple DMUs achieve 100% efficiency. Relative benchmark is defined as $h = \{i|\max_{i=1}^{N_{G_j}}\zeta_i\}, \max_{i=1}^{N_{G_j}}\zeta_i < 1$, and absolute benchmark is $h = \{i|\zeta_i \geq 1\}, \max_{i=1}^{N_{G_j}}\zeta_i \geq 1$, where h denotes the frontier, ζ_i represents the efficiency of DMU i ($i \in 1 \ldots N_{G_j}$ in group G_j), N is the number of DMUs, g is the number of groups identified, and G_j has N_{G_j} DMUs.

III. Empirical study

A. Data and methods

Our empirical data comes from the Energy Market Authority (EMA) website (www.emvi.fi), which consists of 85 electricity suppliers and are the six-year average over the period 2005-2010 [4], [7]. Recently, EMA has replaced the conventional DEA and SFA by StoNED after a rigorous evaluation process [7]. Also, provided with the advantages of StoNED in overcoming the pitfalls of DEA and SFA [10], we fitted StoNED in this framework, and focused on evaluating the performance of NMM in improving the accuracy of efficiency estimation when combined with StoNED. For the purpose of comparison, K-means, a simple yet powerful and most widely applied clustering technique, was chosen.

We used the cost frontier model, $x_i = C(\mathbf{y}_i)\cdot\exp(\delta z_i + u_i + v_i)$, as adopted by EMA [7], in this empirical study, where C denotes the frontier cost function. This model adds a contextual variable z and its weight δ to the conventional cost frontier model. The variable z is the proportion of the underground cables in the total network length which captures the heterogeneity of the electricity suppliers in Finland, since the underground cables are widely used in urban and suburban regions but not in rural areas. In this model, the total cost (x) is used as the single input, and three variables, i.e., 'Energy transmission' (GWh of 0.4 kV equivalents, y_1), 'Network length' (km, y_2), and 'Customer number' (y_3) are specified as the outputs ($\mathbf{y}$). We used the three output-input ratios from productive efficiency analysis as the input variables for clustering, i.e., 'Energy transmission/Efficient cost' (r_1), 'Network length/Efficient cost' (r_2), and 'Customer number/Efficient cost' (r_3), where the efficient cost is computed as the estimated cost frontier '$C(\mathbf{y}_i)$' to take into account the efficiencies in segmentation. In addition, the actual cost was used in the inputs, i.e., 'Energy transmission/Actual cost' (r_1), 'Network length/Actual cost' (r_2), and 'Customer number/Actual cost' (r_3), to exclude the influence of the efficiencies in the analysis as a comparison. Note that the efficient cost is computed as the actual cost multiplied by the firm efficiency. We used the descriptive statistics of the clustered groups to evaluate the clustering accuracy, assuming that better clustering results in more distant inter-group means, less cross-group overlaps and lower within-group standard deviations.

B. Results and discussion

The 85 firms were grouped into four clusters, which consist of 26, 33, 24 and 2 DMUs, respectively, for clusters 1 to 4. The descriptive statistics, including mean, standard deviation and parameter ranges of $r_1 \ldots r_3$ and 'Energy transmission/Network length', are summarized for groups clustered by NMM and K-means in Table 1. Efficient and actual costs are used as the denominator of the inputs in the upper and lower panel of Table 1, respectively.

Let's first analyze the scenarios where efficient cost is used for computing the clustering inputs. It is seen that the groups clustered using NMM are characteristic of the four types of electricity networks in Finland, but with K-means the statistics are not as representative as such especially for the 4^{th} cluster (the industrial network). Specifically, the rural area consumes less energy than the other regions given its sparse population in Finland and there is no significant difference among suburban,

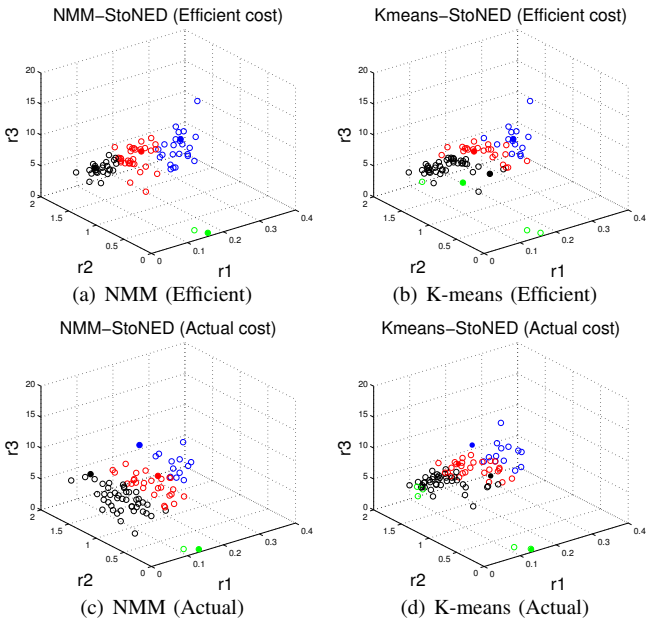
Proc. of the Third Intl. Conf. on Advances in Economics, Management and Social Study - EMS 2015

ISBN: 978-1-63248-058-3 doi: 10.15224/ 978-1-63248-058-3-95

urban and industrial customers. This property is represented by r_1, and better captured in NMM-clustered groups than those clustered by K-means, since the distance between cluster 1 and the average of the other clusters is (0.124+0.158+0.156)/3-0.075=0.071 in NMM-clustered groups which is larger than that of K-means, i.e., (0.137+0.162+0.124)/3-0.095=0.046 (Table 1). The distance between the customer and electricity producer decreases from the rural to the industrial group, leading to a declining trend in the 'Network length' from clusters 1 to 4. This is well-captured by r_2 in NMM-clustered groups but is violated by the industrial cluster when the groups are clustered by K-means (i.e., the distance is 0.735 in the industrial group which is bigger than 0.529, the distance in the urban cluster). The number of customers increases from the rural to urban regions, and only a few industrial customers exist in Finland. This property is captured by r_3 in both NMM and K-means clustered groups. However, as the standard deviation of the group means is slightly larger in NMM-clustered groups than that in the K-means case, we'd say that groups are more clearly separated by NMM than K-means regarding this parameter. Here, we also examined the 'Energy transmission/Network length', since it merges r_1 and r_2 (the parameters that capture the principle differences between NMM and K-means in separating these groups given their statistics) and should represent the major distinction between the four groups as well as different clustering techniques. As seen from Table 1, the standard deviation of the group means is much larger in NMM-clustered groups (0.623) than that in the case of K-means (0.288), the average standard deviation of the groups is lower in case of NMM (0.209) compared with K-means (0.285), and there is no adjacent group overlap in NMM separated clusters but is 0.323 on average in the case of K-means. Thus, it is concluded that NMM could separate the four types of electricity suppliers into more appropriate groups compared with K-means in this empirical study.

The same conclusions can be drawn when the actual cost is used in the inputs as seen from Table 1. Thus, NMM performs better than K-means in this real case application regardless of whether the efficiencies are taken into account in computing the clustering inputs. However, using efficient cost in the inputs indeed groups the DMUs into more distant clusters than using the actual cost no matter whether NMM or K-means is used. For example, the averages of the standard deviation and overlapping range are lower in most cases when the efficient cost is used than those computed using the actual cost, indicating a higher within-group homogeneity and a larger inter-group distance when efficiencies are included in grouping. Also, the standard deviations of the group means are mostly larger when the efficient cost is used in the inputs than those computed using the actual cost, which again shows a larger inter-group distance among the four clusters.

The superiority of NMM over K-means in separating the rural, urban, suburban and industrial electricity networks in Finland is also illustrated in Figure 1. In this figure, each color represents one type of electricity supplier. There is a clear trend from the rural to urban areas (colored in black, red, and blue, respectively) along the three axes and the industrial cluster (shown in green) is distinctively separated from the other groups in NMM clustering, regardless of whether the efficiency is taken into account; yet when K-means is used, the boundaries become ambiguous especially for the industrial group where a few units are scattered into the rural cluster. More importantly, notice that the filled dots (representing the best performing DMU in a given cluster) may differ when different clustering techniques are used, resulting in different benchmarks for a given DMU. Take the industrial group as an example, its best performing unit is within the rural area in K-means clustering when efficient cost is used in the inputs which, once chosen as the benchmark for the industrial group, will become an unrealistic goal for this cluster given their large differences in, e.g., the input-output space.



(a) NMM (Efficient) (b) K-means (Efficient)
(c) NMM (Actual) (d) K-means (Actual)

Figure 1. Comparison of NMM and K-means using EMA data. The efficient cost (a and b) and actual cost (c and d) are used in the inputs. The filled dots are the best performing unit for each cluster. 'Black', 'red', 'blue' and 'green' represent 'rural', 'suburban', 'urban' and 'industrial' networks, respectively.

IV. Conclusions

We present a combination of the NMM based clustering and the StoNED efficiency estimation technique in the benchmarking framework previously presented in [4]. It inherits the advantages of NMM such as automatic determination of the number of clusters and insensitivity to random noise, and the benefits of StoNED in its stochastic and semi-parametric modelling. With one empirical application we show that the DMUs could be clustered into groups having less ambiguous boundaries than other clustering techniques such as K-means. The superiorities of StoNED over other productive efficiency analysis methods such as DEA and SFA have been previously studied in [10]. Further, the benefits of combining

Proc. of the Third Intl. Conf. on Advances in Economics, Management and Social Study - EMS 2015

ISBN: 978-1-63248-058-3 doi: 10.15224/ 978-1-63248-058-3-95

Efficient cost	NMM					K-means				
Mean	Cluster 1	Cluster 2	Cluster 3	Cluster 4	STD(Mean)	Cluster 1	Cluster 2	Cluster 3	Cluster 4	STD(Mean)
r_1	0.075	0.124	0.158	0.156	0.034	0.095	0.137	0.162	0.124	0.024
r_2	1.403	1.110	0.550	0.117	0.497	1.273	0.911	0.529	0.735	0.273
r_3	5.961	8.219	11.846	0.265	4.204	6.504	9.785	13.227	1.986	4.149
ET/NL	0.054	0.114	0.314	1.583	0.623	0.080	0.174	0.341	0.827	0.288
STD					Mean(STD)					Mean(STD)
r_1	0.016	0.026	0.020	0.012	0.019	0.027	0.022	0.020	0.041	0.028
r_2	0.083	0.131	0.153	0.043	0.103	0.196	0.268	0.177	0.628	0.317
r_3	0.977	1.776	2.281	0.228	1.316	1.015	0.784	1.707	1.729	1.309
ET/NL	0.014	0.026	0.108	0.686	0.209	0.039	0.090	0.113	0.898	0.285
[min, max]					Mean(OL)					Mean(OL)
r_1	[0.038,0.102]	[0.096,0.168]	[0.120,0.210]	[0.144,0.168]	0.026	[0.038,0.168]	[0.084,0.174]	[0.135,0.210]	[0.059,0.168]	0.052
r_2	[1.226,1.611]	[0.785,1.374]	[0.221,0.801]	[0.074,0.161]	0.055	[0.642,1.611]	[0.357,1.374]	[0.221,0.923]	[0.074,1.506]	0.667
r_3	[3.651,8.410]	[3.763,11.954]	[8.009,18.491]	[0.038,0.493]	2.864	[4.552,8.085]	[8.297,11.302]	[11.612,18.491]	[0.038,3.763]	0
ET/NL	[0.023,0.078]	[0.079,0.169]	[0.177,0.612]	[0.897,2.269]	0	[0.023,0.237]	[0.064,0.489]	[0.153,0.612]	[0.039,2.269]	0.323

Actual cost	NMM					K-means				
Mean	Cluster 1	Cluster 2	Cluster 3	Cluster 4	STD(Mean)	Cluster 1	Cluster 2	Cluster 3	Cluster 4	STD(Mean)
r_1	0.093	0.137	0.163	0.156	0.027	0.097	0.133	0.163	0.105	0.026
r_2	1.269	0.887	0.600	0.117	0.421	1.246	0.903	0.600	0.907	0.229
r_3	6.658	9.494	12.915	0.265	4.643	6.572	9.806	12.915	3.105	3.653
ET/NL	0.081	0.188	0.318	1.583	0.607	0.087	0.183	0.318	0.663	0.218
STD					Mean(STD)					Mean(STD)
r_1	0.027	0.023	0.021	0.012	0.021	0.030	0.023	0.021	0.043	0.029
r_2	0.228	0.289	0.238	0.043	0.200	0.233	0.306	0.238	0.647	0.356
r_3	1.549	1.885	2.278	0.228	1.485	1.307	1.600	2.278	2.437	1.906
ET/NL	0.045	0.109	0.128	0.686	0.242	0.050	0.111	0.128	0.867	0.289
[min, max]					Mean(OL)					Mean(OL)
r_1	[0.038,0.150]	[0.095,0.174]	[0.135,0.210]	[0.144,0.168]	0.053	[0.038,0.168]	[0.095,0.174]	[0.135,0.210]	[0.059,0.168]	0.048
r_2	[0.542,1.611]	[0.331,1.296]	[0.221,1.102]	[0.074,0.161]	0.508	[0.542,1.611]	[0.331,1.374]	[0.221,1.102]	[0.074,1.506]	0.828
r_3	[3.651,10.579]	[5.393,12.803]	[8.981,18.491]	[0.038,0.493]	3.003	[3.763,9.478]	[6.913,12.803]	[8.981,18.491]	[0.038,5.700]	2.129
ET/NL	[0.023,0.266]	[0.073,0.489]	[0.128,0.612]	[0.897,2.269]	0.185	[0.023,0.266]	[0.073,0.489]	[0.128,0.612]	[0.039,2.269]	0.346

Table 1. Descriptive statistics of groups clustered using efficient (upper panel) and actual (lower panel) costs in the inputs. ET/NL is Energy transmission/Network length. 'STD(Mean)' represents the standard deviation of the mean. 'Mean(STD)' is the average of the standard deviation. Overlap is computed between every adjacent 2 ranges, 'Mean(OL)' is the average length of 3 overlaps among 4 clusters.

NMM and StoNED as compared with the traditional DEA in benchmarking has been previously demonstrated by an empirical application in [4]. Thus, the performance of the proposed configuration in the clustering based benchmarking framework [4], i.e., NMM-StoNED, has been well-surrounded and is suggested to use if no specific needs to meet.

With the metrics selected as the input of clustering, we obtained four mutually exclusive clusters, each corresponds to a well-defined type of energy supplier. It is worth mentioning that with different metrics as the inputs, the clustering results may differ. Thus, one need to identify the principle statistics dominating the heterogeneity of the DMUs if not otherwise specified before clustering. If the input metrics do not follow or are not convertible to the normal distribution, a mixture model of the corresponding distribution or a joint mixture model [3] need to be used. Also, the computational complexity increases with the number of inputs. Therefore, techniques such as principle component analysis are needed to capture the main properties needed for clustering.

This paper successfully applies NMM-StoNED to energy regulation data which, however, is not restricted to such an area. It is applicable to any problems where the distribution of the evaluating metric is or convertible to normal distribution. Here we focus on applying NMM-StoNED in the cross-section setting, which could be used for panel data as well. To solve more practical benchmarking problems especially those that are problematic using conventional methods, more applications are worthwhile to explore.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (31471251) and the Fundamental Research Funds for the Central Universities (JUSRP11507). I thank Prof. Timo Kuosmanen for his insightful advice.

REFERENCES

[1] P. Bogetoft and K. Nielsen, "Internet based benchmarking", Group Decision and Negotiation, vol. 14, 2005, pp. 195-215.

[2] L. Botti, W. Briec and G. Cliquet, "Plural forms versus franchise and company-owned systems: a DEA approach of hotel chain performance", Omega, vol. 37, 2009, pp. 566-578.

[3] X. F. Dai, T. Erkkila, O. Yli-Harja and H. Lahdesmaki, "A joint finite mixture model for clustering genes from independent Gaussian and beta distributed data", BMC Bioinformatics, vol. 10, 2009, doi:10.1186/1471-2105-10-165.

[4] X. F. Dai and T. Kuosmanen, "Best-practice benchmarking using clustering methods: Application to energy regulation", Omega, vol. 42, 2013, pp. 179-188.

[5] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation", Journal of the American Statistical Association, vol. 97, 2002, pp. 611-631.

[6] S. C. Johnson, "Hierarchical clustering schemes", Psychometrika, vol. 32, 1967, pp. 241-254.

[7] T. Kuosmanen, "Stochastic semi-nonparametric frontier estimation of electricity distribution networks: Application of the StoNED method in the finnish regulatory model", Energy Economics, vol. 34, 2012, pp. 2189-2199.

[8] T. Kuosmanen and A. L. Johnson, "Data envelopment analysis as nonparametric least-squares regression", Operations Research, vol. 58 , 2010, pp. 149-160.

[9] T. Kuosmanen and M. Kortelainen, "Stochastic non-smooth envelopment of data: Semi-parametric frontier estimation subject to shape constraints", Journal of Productivity Analysis, vol. 38, 2012, pp. 11-28.

[10] T. Kuosmanen, A. Saastamoinen and T. Sipilainen, "What is the best practice for benchmark regulation of electricity distribution? comparison of DEA, SFA and StoNED methods", Energy Policy, vol. 61, 2013, pp. 740-750.

[11] J. MacQueen, "Some methods for classification and analysis of multivariate observations", In Proceedings of the 5th berkeley symposium on mathematical statistics and probability, 1967, pp. 281-297.

[12] G. J. McLachlan and D. Peel, "Finite mixture models". New York, USA: John Wiley & Sons, 2000.

[13] W. Pan, "Incorporating gene functions as priors in model-based clustering of microarray gene expression data", Bioinformatics, vol. 22, 2006, pp. 795-801.

[14] S. Zhong and J. Ghosh, "A unified framework for model-based clustering", Journal of Machine Learning Research, vol. 4, 2003, pp. 1001-1037.