School of Economic Sciences

Working Paper Series
WP 2017-6

BAYESIAN DEA, THE CURSE OF DIMINESIONALITY, AND MISCLASSIFIED DMUs

Mehmet Guray Unsal, Daniel Friesner
and Robert Rosenman

April 2017

WASHINGTON STATE UNIVERSITY
World Class. Face to Face.

BAYESIAN DEA, THE CURSE OF DIMINESIONALITY, AND MISCLASSIFIED DMUs

Mehmet Guray Unsal, Daniel Friesner and Robert Rosenman*

Abstract

In this paper we use Bayesian methods to characterize and adjust for two potential biases in data envelopment analysis, the first from decision-making units (DMUs) that may be misclassified as efficient, and the second related to the curse of dimensionality (COD). We propose a theoretically appropriate triangular prior distribution for the proportion of misclassified DMUs and use it to derive the concordant posterior distribution. Simulation analysis suggests that a triangular prior may outperform an ignorance prior in small samples where COD-biases are most likely to exist. An application applies the model to data from the Turkish electric industry.

Key Words: Data Envelopment Analysis, Incidence of Inefficiency, Simulation, Bayesian Priors

JEL: C11, C14

Highlights

- We account for the curse of dimensionality in the misclassification of DEA estimates.
- A left skewed (discrete triangular) distribution is used to model prior misclassification beliefs.
- Simulations suggest that a left skewed prior outperforms an ignorance prior in small samples.

* Mehmet Guray Unsal (*Corresponding Author*), Research Assistant, Statistics Department, Uşak University, (email: mgunsal@gazi.edu.tr, mehmet.unsal@usak.edu.tr), Daniel Friesner, Professor and Associate Dean, College of Health Professions, Dept. 2650, P.O. Box 6050, North Dakota State University, Fargo, ND 58108-6050, (email: Daniel.Friesner@ndsu.edu, voice phone 701-231-9509, fax 701-231-7606), Robert Rosenman, Professor, School of Economic Sciences, Room 101E, Hulbert Hall, Washington State University, Pullman, WA, 99164-6210, (email: yamaka@wsu.edu, voice phone 509-335-1193, fax 509-335-1173).

BAYESIAN DEA, THE CURSE OF DIMINESIONALITY, AND MISCLASSIFIED DMUs

1. Introduction

Data envelopment analysis (DEA) has become one of the most commonly used empirical methods to evaluate the relative efficiency of production processes. Within the economics and operations research literatures, several hundred studies using DEA have been published (Gattoufi et al., 2003; Hollingsworth, 2003; Hatami-Marbini et al., 2011). When applying data envelopment analysis (DEA) to real world data, it is not uncommon for a researcher to have many inputs and outputs relative to the number of observations, attenuating the ability to discriminate between observations. This situation is often referred to as the "curse of dimensionality" (COD). Without a sufficiently large and varied selection of DMUs (or a sample that approaches the size of the population) it is impossible to ensure that the efficient frontier is appropriately mapped, thus distorting the efficiency analysis. Indeed, the fact that DEA uses an extreme value approach to calculate the frontier makes empirical applications of DEA highly sensitive to the COD.

Generally researchers have two options to address the COD. The first is to postulate a theoretical frontier against which all firms (both those empirically realized and potentially realized) can be assessed (Badin et al., 2014). Alternatively, if the frontier is to be defined empirically, the bias can be reduced by reducing the ratio of the number of inputs and outputs to the sample size. One way to accomplish the second approach is to expand the number of observations. Since this is rarely possible, ratio reduction is often accomplished by applying principal components-based DEA (PCA-DEA), in which principal components analysis is used to reduce the dimensionality of the production process prior to, or concurrently with, the application of DEA (Ueda and Hoshiai, 1997; Adler and Golany, 2002; Yap et al., 2011; Augustyniak, 2014; Araujo, et al., 2014). One drawback to PCA-DEA is that it is difficult to meaningfully interpret the efficient frontier and resulting efficiency estimates.

Given the limitations just described, some researchers simply acknowledge that the results **(**Gijbels et al., 1999; Kneip et al., 2003; Badin and Simar, 2009; Badin et al., 2014**)** may be biased. However, without any means to characterize the extent of the COD, it is unclear exactly *how biased* the results are, and by extension, whether researchers and policy makers should consider that study's results as valid and reliable.

In a recent paper, Friesner, et al. (2013) used Bayesian methods to estimate the general prevalence of bias that exists in DEA estimates. Although their focus was on misclassified decision making units (DMUs), the approach can correct for bias regardless of cause. Assuming specifically that researchers draw data without replacement from a finite population they show that if such biases[1] occur, DEA scores will be mis-measured because some DMUs classified as "fully efficient" will not *truly* be so, that is, they are misclassified. In their Bayesian approach, the researcher imposes prior beliefs about the extent of the bias (from whatever cause) and combines this with the data yielding posterior distributions, which can be analysed using various metrics (credible regions, etc.). Prior beliefs may come from earlier studies, knowledge of the industry being studied, or statistical properties associated with a sampling method. Absent other information, Friesner, et al. (2013) suggest using an "ignorance" prior (also known as an "uninformative" prior).

In this paper we improve on Friesner et al. (2013) by characterizing and adjusting specifically for the possibility of COD-related bias. Based on the nature of COD bias, we suggest a more appropriate prior distribution for the proportion of misclassified DMUs. While doing so we more completely clarify the full set of assumptions underlying their methodology. Simulation analyses shows our results generally superior to those obtained with the Friesner et al. (2013) ignorance prior. We finish with an application to the Turkish electric industry.

2. Adjusting for The Curse of Dimensionality

2.1 Output oriented production in theory and in DEA

Our main contribution is to improve on the Friesner et al. (2013) in the choice of the prior, focusing on the COD. This is crucial in applied DEA studies, where researchers are often faced with data limitations (including small sample sizes relative to the number of inputs and outputs, drawn from finite populations) that may create or exacerbate the COD. While it is known that COD-related biases disappear as the sample size increases (holding the number of inputs and outputs constant), what is unknown is the rate at which the bias disappears as the ratio

[1] In what follows, we employ Friesner et al.'s (2013) definition of "bias", which denotes a difference between the empirically realized DEA scores and their scores in the absence of any confounding effects which distort the efficient frontier. It should not be confused with its more common use in statistics. In our application of Friesner et al.'s methodology, we implicitly hold constant any confounding factors besides the COD.

of the sample size to the number of inputs and outputs increases. Nonetheless, general information about the rate at which the COD bias diminishes can provide a simple rule of thumb for the selection of a reasonable prior distribution.

To investigate this issue, we construct a simple simulation exercise with an output-oriented production function. In an output-oriented production function, a DMU uses a px1 vector of inputs to produce a qx1 vector of outputs. Efficiency is defined based on producing maximum output given these inputs. If we denote $\mathbf{w} \in \mathrm{R}^{p}_{\geq 0}$ as the vector of inputs and $\mathbf{y} \in \mathrm{R}^{q}_{\geq 0}$ as the vector of outputs, the output-oriented production technology is theoretically defined by the following (Debreu, 1951; Farrell, 1957; Fare and Primont, 1995):

$$\mathrm{T} = \left\{ (\mathbf{w},\ \mathbf{y}) : \mathbf{w} \in \mathrm{R}^{p}_{\geq 0},\ \mathbf{y} \in \mathrm{R}^{q}_{\geq 0}\ \textit{s.t.}\ \mathbf{w} \text{ can produce } \mathbf{y} \right\} \tag{1}$$

Assuming i) that all DMUs in the population use one or more inputs to produce one or more outputs; ii) the existence of a convex efficient frontier; and iii) free disposability of outputs, the the production technology defined in (1) leads to the standard output-oriented distance function measure:

$$d = d(\mathbf{w},\mathbf{y}) \equiv \inf_{\theta} \{\theta \mid (\mathbf{w}, \frac{\mathbf{y}}{\theta}) \in T\} \tag{2}$$

The goal of (output-oriented) DEA is to empirically characterize (1) and (2). In its standard formulation, DEA empirically generates the efficient frontier, and evaluates DMUs relative to the frontier, using the following linear program:

$$\max\ \rho\ \ \textit{s.t.}\ \begin{Bmatrix} -\rho\mathbf{y} + \mathbf{Y}\boldsymbol{\lambda} \geq \mathbf{0} \\ \mathbf{w} - \mathbf{W}\boldsymbol{\lambda} \geq \mathbf{0} \\ \mathbf{1}_n{}'\boldsymbol{\lambda} = 1 \\ \boldsymbol{\lambda} \geq \mathbf{0} \end{Bmatrix} \tag{3}$$

where $\mathbf{W}$ is a $p \times n$ matrix of inputs, $\mathbf{Y}$ is a $q \times n$ output matrix , $\mathbf{1}_n$ is an $n \times 1$ vector of 1's and $\boldsymbol{\lambda}$ represents an $n \times 1$ vector of convexity weights. DEA efficiency scores (which attempt to characterize *d* in equation 2) are calculated by computing the inverse of ρ.

2.2 Developing the distribution of inefficient DMUs through simulation

The model formulated in (3) implicitly assumes that the observed DMUs being evaluated are sufficient to accurately and precisely characterize the true frontier. When not the case, and

absent additional information, researchers have little to go on when characterizing potential biases in their estimates, including those associated with the COD.

A common formulation of a DEA study occurs when the population (N) is finite and the frontier is defined relative to only those input-output combinations that are actually observed. Bias may be further exacerbated if a sample is drawn using sampling without replacement. Such situations give little theoretical guidance in how to account for any forces that may bias the construction of the efficient frontier, including COD. While unstated in their manuscript, Friesner et al. (2013) implicitly assume no knowledge about the source of the bias.

We more narrowly assume a bias caused only by the COD which allows us to theoretically justify a specific prior distribution. Consider a finite population and a sample of DMUs chosen without replacement. Each DMU uses a fixed input and m variable inputs to produce s outputs. Given a fixed set of DMUs, the COD is examined by adjusting the number of outputs and variable inputs (all DMUs always use a single fixed input). Tables 1and 2 show the variable input-output (m, s) combinations we use and some basic statistics for a sample size of 20 DMUs. The fixed and variable inputs are aggregated to produce the outputs through the following production technology, which applies to each of the DMUs: $\ln Y_i = \ln \beta + \sum_{j=1}^{m} \alpha_j \ln X_j - u_i + v_i$ where Y_i, i=1,…,s denotes each output; X_j, j=1,…,m denotes each variable input; β is the fixed input, which we assume is the same for all 20 DMUs; u_i represents a normally distributed random disturbance for each output $u_i \sim N(\mu, \sigma)$; and v_i represents the truncated normal disturbance denoting inefficiency: $v_i \sim N^+(\mu, \sigma)$ (Giraleas et al., 2012).

DEA is applied for each given input-output combination, and the incidence of inefficiency among the sample is calculated. Consistent with Friesner et al. (2013) we conduct the simulation twice, once assuming a single ray technology and once with a two-ray technology, where the latter is enforced by the constraint: $\sum_{j=1}^{m} \alpha_j = 1$. Each simulation was repeated 100,000 times, and the results were summarized using expected values, variances, skewness and kurtosis. While 100,000 replications may be more than the traditional requirements for simulation analyses, the authors determined that the costs of the additional computation time were more than offset by a greater certainty that the resulting distributions accurately and precisely characterize the shape of the resulting distributions (Manly, 2006). This additional certainty is crucial considering that the simulation exercises largely determine the choice for the prior distribution of the incidence of inefficiency.

The column labeled "Expected Percent of Inefficient DMUs" is the expected value of the percentage of inefficient DMUs. The variance measures the dispersion of this statistic. Skewness and kurtosis also characterize the empirical distribution of this metric. As illustrated in Tables 1 and 2 and Figures 1 and 2, the expected values for the empirical distribution of the incidence of inefficiency have a tendency to decrease as the number of inputs and outputs increases. This shows that the discriminatory power of the model weakens as the ratio of the sample size to the number of inputs and outputs decreases. Skewness coefficients indicate a tendency towards a positively skewed situation from a negatively skewed one[2] (Unsal et al., 2014). More specifically, when the sum of inputs and outputs is low compared to the size of the sample, the empirical distribution of incidence of inefficiency is negatively skewed. However, as the sum of input and output variables approaches the sample size, the empirical distributions become positively skewed (Unsal et al., 2014). These findings suggest that a reasonable prior distribution for the incidence of inefficiency should be negatively skewed when there is little chance of COD bias and positively skewed when the COD is expected to be more severe.

One obvious prior distribution that is flexible, parsimonious and can characterize both negatively and positively skewed distributions is the discrete version of the triangular distribution (Evans and Olson, 2003; McLaughlin and Hays, 2008). This distribution is commonly used in simulations and numerical modelling because it can take any range of shapes (and can thus approximate many different distributions) and need only be described by its mode, minimum and maximum support values. A negatively skewed version of this distribution can be expressed as:

$$P(\pi) = \begin{cases} \dfrac{2n\pi}{n(n-1)} & \dfrac{1}{n} \le \pi < 1 \\ 0 & elsewhere \end{cases}, \min(\pi) = \frac{1}{\mathrm{n}}, max(\pi) = \frac{\mathrm{n}-1}{\mathrm{n}}, \quad \mathrm{Mean}_\pi = \mathrm{E}(\pi) = \frac{2n-1}{3n}$$

$$\mathrm{Mode}_\pi = \tfrac{n-1}{n}, \quad \mathrm{Mean}_\pi < \mathrm{Mode}_\pi \tag{4}$$

where $\pi \in \Omega = \left\{ \frac{\gamma}{N}, \, for \; \gamma = 0,1,...,N \right\}$.

2.3 Improving DEA estimates with Bayesian methods

[2] The right skewed structures in empirical studies are related to "non rule of thumb" cases in the literature. The "rule of thumb" says that classical DEA models may lose the power of discrimination among of DMU efficiency when there are a large number of variables in the analysis compared with number of observations.

Friesner et al. (2013) define π as the proportion of inefficient firms in the population, which is contained in the closed unit interval $\pi \in [0,1]$. In most empirical applications of DEA, DMUs are drawn from a finite population, making π discrete: $\pi \in \Omega = \left\{ \frac{\gamma}{N}, \; for \; \gamma = 0,1,...,N \right\}$ where γ is the number of inefficient firms in the population. If sampling is conducted without replacement[3], and if there are K inefficient DMUs in the population, the likelihood of drawing a sample of n DMUs, x of which are inefficient, is given by the hypergeometric density:

$$f(x \mid N,\pi,n) = \begin{Bmatrix} \dfrac{\binom{N(1-\pi)}{n-x}\binom{N\pi}{x}}{\binom{N}{n}} & \max[0, n-(N-K)] \le x \le \min[n,K] \\ 0 & otherwise \end{Bmatrix} \tag{5}$$

where $\binom{m}{z} \equiv \frac{m!}{z!(m-z)!}$ is the standard combination calculation.

The fundamental contribution of Friesner et al. (2013) is to apply a Bayesian prior distribution based on information from the sample or from *a priori* knowledge about the distribution of the parameters of interest. The resulting posterior distribution allows the computation of a Bayesian estimator, the minimum posterior quadratic risk (MPQR). Friesner et al. (2013) assume an ignorance prior for π: $p(\pi) = \frac{1}{N+1} \; for \; \pi \in \Omega$. It follows that

$$f(x,\pi \mid N,n) = f(x \mid N,\pi,n)\,p(\pi) = \left[(N+1)\binom{N}{n} \right]^{-1} \binom{N(1-\pi)}{n-x}\binom{N\pi}{x} \tag{6}$$

where supports for (6) are suppressed for simplicity. By summing the joint density over π (to create a marginal distribution), and subsequently taking the ratio of the joint and marginal distributions, we arrive at a posterior distribution for the distribution of inefficiency with the following form (where, as before, supports are suppressed for simplicity):

$$f(\pi \mid x,N,n) = \frac{f(x,\pi \mid N,n)}{f(x \mid N,n)} = \frac{\binom{N(1-\pi)}{n-x}\binom{N\pi}{x}}{\sum_{\pi \in \Omega} \binom{N(1-\pi)}{n-x}\binom{N\pi}{x}} \tag{7}$$

[3] Some researchers assume random sampling with replacement, and approximate the hypergeometric probabilities with binomial probabilities. While this simplifies the characterization of the posterior distribution, it is typically indefensible since N is usually not close to infinity. See Friesner et al. (2013), footnote 20 for more details.

The posterior distribution in (7) fails to account for the fact that DEA constructs efficiency scores based on the data available in the sample of n DMUs. Hence, x (the number of inefficient DMUs in the sample) is never known with certainty, except that it is almost certainly under-estimated. To account for this, let x_* be the *true* number of inefficient DMUs in the sample, and let x be the number identified as inefficient by applying DEA to the sample. In the simplest case where the DEA technology is defined by a single ray, it follows that $x_* = x + e$, where $e = \{ {}^{n_E}_{0}$ is the latent measurement error that occurs when counting the number of inefficient firms, and n_E is a count variable representing the number of misclassified DMUs in the sample. The joint distribution for $\{x_*, \pi\}$, conditional on the empirically observed values $\{x, n_E\}$, over the support $\{x_*, \pi\} \in \Psi \times \Omega$ is given by

$$f(x_*, \pi | N, n, x, n_E) = \frac{f(x_*, \pi | N, n)}{\sum_{x_* \in \Psi} \sum_{\pi \in \Omega} f(x_*, \pi | N, n)} = \frac{\binom{N(1-\pi)}{n-x_*}\binom{N\pi}{x_*}}{\sum_{x_* \in \Psi} \sum_{\pi \in \Omega} \binom{N(1-\pi)}{n-x_*}\binom{N\pi}{x_*}} \tag{8}$$

where $\Psi = \{x, x + n_E\}$ and Ω is as previously defined.

By marginalizing out the latent variable x_*, the posterior distribution of π is conditioned solely on observed information, and is given by:

$$f(\pi | N, n, x, n_E) = \sum_{x_* \in \Psi} f(x_*, \pi | N, n, x, n_E) = \frac{\sum_{x_* \in \Psi} \binom{N(1-\pi)}{n-x_*}\binom{N\pi}{x_*}}{\sum_{x_* \in \Psi} \sum_{\pi \in \Omega} \binom{N(1-\pi)}{n-x_*}\binom{N\pi}{x_*}} \tag{9}$$

and $\Psi = \{x, x + n_E\}$ is the set of feasible misclassified DMUs. This posterior distribution can be characterized by its expected value (i.e., the minimum posterior quadratic risk, or MPQR, value) and/or credible regions. For example, the MPQR estimate is:

$$E(\pi) = \frac{\Sigma_{x_* \in \Psi} \Sigma_{\pi \in \Omega} \pi \binom{N(1-\pi)}{n-x_*}\binom{N\pi}{x_*}}{\Sigma_{x_* \in \Psi} \Sigma_{\pi \in \Omega} \binom{N(1-\pi)}{n-x_*}\binom{N\pi}{x_*}} \tag{10}$$

Extensions of (9) and (10) to multi-ray DEA technologies can be incorporated by allowing e to take on multiple values. For example, in the two ray (denoted by rays A and B) case, $x_* = x + e,\ e \in Unique\{0, n_E^A, n_E^B, n_E\}$ and $\Psi = Unique\{x, x + n_E^A, x + n_E^B, x + n_E\}$.

The framework just described can be recast using the discrete triangular prior defined in (4) instead of the ignorance prior used by Friesner et al. (2013). As they did, assume the probability of being efficient for none and all of the DMUs are both assumed to be zero. Incorporating the discrete triangular prior with a hypergeometric sampling process yields a posterior distribution for π analogous to equation (7), but which more thoroughly accounts for the researcher's belief about the extent of COD-related bias, givng the MPQR estimate

$$E(\pi)=\frac{\sum_{\pi=1/n}^{(n-1)/n}\pi^{2}\binom{N(1-\pi)}{n-x}\binom{N\pi}{x}}{\sum_{\pi=1/n}^{(n-1)/n}\pi\binom{N(1-\pi)}{n-x}\binom{N\pi}{x}} \qquad \pi=\frac{1}{n},\frac{2}{n},...,\frac{n-1}{n}. \tag{11}$$

3. Implementing the Discrete Triangular Prior to Account for The Curse of Dimensionality

Of particular interest is how the use of discrete triangular priors leads to improved performance over the ignorance-based priors of Friesner et al. (2013). To check, simulations are conducted using the approach outlined in the previous section. In each case, the production function is limited to m=1 variable input and s=2 outputs to facilitate the use of the negatively (left) skewed discrete triangular prior described in the previous section. An ignorance prior is used as the baseline against which the negatively skewed prior can be compared. The use of a low number of inputs and outputs allows multiple samples sizes to be used, which are consistent with many empirical applications of DEA, yet are also consistent with the choice of the triangular prior. For each replication, the, *n* DMUs are randomly selected without replacement from a simulated population. Information gleaned from the sample (via hypergeometric probabilities) is combined with prior information (i.e., a left-skewed triangular distribution) to generate posterior distributions adjusted for the possibility of misclassification.

Traditional DEA theory the minimum acceptable sample size is based on the formula Q = max {3* (m+s), m*s}, where m and s are the number of input and output variables respectively. This would argue for a sample of at least 9 in our simulations. To check the value of the Bayesian approach in general and our proposed prior in particular, we conduct simulations eight times, once for each sample size: n=5, n=10, n=15, n=20, n=40, n=60, n=80 and n=100, so we

have one sample size below what would be expected in traditional DEA, one just above that value, and six simulation samples that allow us to observe the value of the different approaches as sample size grows.

All simulations are based on 10,000 replications, which (having already established the prior distribution) should be sufficient to accurately characterize each underlying distribution (Manly, 2006). In each replication and for a given sample size, a random data set is constructed, and the output-oriented DEA algorithm described earlier is applied to the data. The empirical incidence of inefficiency is calculated, and the posterior distributions are generated (once using the ignorance prior and once using the triangular prior). To facilitate additional comparisons, the posterior distributions for the incidence of inefficiency with and without, the correction for misclassification are calculated. After the 10,000 replications are completed, the results are summarized using mean values.

3.1 The Single Ray Technology

Assuming a single ray technology, all firms are absolutely and directly comparable, as they lie on a single ray emanating from the origin to the frontier of the production technology. The results of the single ray simulation are shown in Tables 3. As the sample size increases, all DEA estimates of the incidence of ineficiencyusing either prior distribution and correcting for latent misclassification or not, generate estimates that are more consistent with the true population incidence of inefficiency. The proposed triangular prior with the correction for misclassification dominates all estimates in terms of having expected values that most closely match the true incidence of inefficieny.[4] The gains that accrue for correcting for misclassification are substantial. Comparing the two priors, the triangular (corrected) prior dominates at lower sample sizes, especially for n=5, the sample size below what is normally acceptable in traditional DEA. However, the difference between the triangular prior and the ignorance prior diminishes as the sample size grows, especially once the sample size exceeds 40. While the advantage of the triangular prior over the ingnorance prior is strongest when not adjusting for misclassifcation

[4] This information is not reported in the manuscript, but is available from the authors upon request.

errors, the overall approach is most valuable with latent variables, as can be seen by the closeness of estimates to the true value.

3.2 The Two or Multiple Ray Technology

The results of the simulation in which DMUs may lie along one of two or multiple distinct rays are shown in Table 4 and are consistent with what we found assuming a single ray technology. As in the single ray case, increased sample size generates DEA estimates, whether uncorrected or corrected, and under both corrected prior beliefs, that are more consistent with the true population incidence of inefficiency. Again, the proposed triangular prior with the correction for misclassification dominates all estimates in terms of having expected values that most closely match the true incidence of inefficieny.[5] . Once again the gains that accrue for correcting for misclassification are substantial, the triangular prior is especially dominant at lower sample sizes, and the gains that accrue from using the triangular (corrected) prior belief over and above a standard ignorance prior appear diminish after the sample size exceeds 40 observations.

4. An Empirical Application using Turkish Electric Companies

To we see how the negatively (left) skewed prior performs when applied to a real world data set we present a case study using DMUs collected from the Turkey Electric Companies Corporation *(TEİAŞ)* during the year 2013. An advantage of this application is that we have the population so can assess the incidence of inefficiency when the frontier is described empirically by a population of DMUs. By using a subsample of the population we can evaluate how our triangular prior compares against traditional DEA and the Friesner, et al. (2013) naive prior. There are 21 DMUs in the population, each of which reports 3 variable inputs available for analysis: the number of employees (a measure of labor), the length of power cable (in kilometers) under the purview of a company (a measure of capital), and the number of transformers managed by a company (a second measure of capital). The single output consists of the number of customers for each firm. Application data for each of the inputs and output are provided in *Appendix* .

DEA was first applied to this population assuming an output-oriented, single ray technology. Under this construct, there are 18 inefficient companies; thus the naive incidence of

[5] As with the single ray technology it also generates lower mean absolute errors.

inefficiency is 0.8571. Next, 15 companies are selected using random sampling without replacement, and the posterior distribution for the incidence of inefficieny is calculated using both Friesner et al.'s (2013) ignorance prior and the new discrete triangular prior, but without correcting for misclassification and the end of the single ray (i.e., the frontier).[6] The expected value (MPQR estimate) for the incidence of inefficency under the ignorance prior is 0.7214, while the analogous estimate using the discrete triangular prior is 0.7262. After adjusting for the biases that exist due to the COD, the MPQR estimate of ineeficiency under an ignorance prior is 0.8359, and for the discrete triangular prior is 0.8386. As in the simulation analysis, the results adjusting for misclassification, and which adjust for COD bias, are closer to the population parameter value of incidence of inefficiency.

If one assumes a two ray technology, there are 14 inefficient DMUs, thus the incidence of inefficiency is 0.6667. When 15 companies are selected wtih random sampling without replacement, the uncorrected, expected value for the posterior distribution (MPQR) estimate of inefficiency using an ignorance prior is 0.5821, while it is 0.5864 under the triangular prior. After adjusting for misclassification bias, these estimates are 0.6502 and 0.6557, respectively. Once again our proposed prior distribution gives estimates which are are closer to the population incidence of inefficiency.

These results, along with those from Sections 3, suggest that in situations where the COD is most likely to distort DEA estimates, a discrete triangular prior distribution out-performs an ignorance prior. The intuition behind the result is straightforward. While DEA scores from any sample may be subject to bias and misclassification, the bias is more pronounced in smaller samples due to the COD. The use of a skewed prior distribution (rather than a uniform prior, which assigns equal probabilities to all feasible support values) more appropriately assigns prior misclassification probabilities in anticipation of the COD, and leads to posterior distributions which provider superior model fit when the COD actually exists. As the sample size increases and the risk of COD-related biases decrease, the dominance of the triangular prior decreases, and the ability of the posterior distribution to

[6] The number 15 was chosen so that we did not have the population and according to the rule of thumb that the number of DMUs in a DEA analysis should be greater than one plus 3 times the number of inputs plus outputs. We have three inputs and one output, so this threshold value is 13. At that point, the choice of 15 from the range of 14 to 20 was arbitrary.

accurately and precisely predict the magnitude of DEA misclassification converges with that of the ignorance prior.

An additional implication of our analysis is that it provides a simple rule of thumb for applied researchers who are concerned about the possible biases that result from the COD. More specifically, we find that, with simple production processes, the effects of the COD disappear after a sample size of 40 - 60 observations. While our results do not provide a comprehensive analysis of where the effects of the COD become insubstantial, they do provide a simple means for future researchers to identify a similar rule of thumb for production processes with different input-output combinations. In other words, future replications of our simulation analysis can identify benchmarks for specific production processes, which can then guide decisions concerning how much data to collect from DMUs (sampled without replacement from finite populations) to avoid the curse of dimensionality. In such cases, one may follow our methodology and adopt a triangular prior, or (if a sufficient sample size is collected) use Friesner et al.'s more straightforward (and computationally simpler) ignorance prior to adjust for misclassification of the efficient frontier.

5. Conclusions, Limitations, and Suggestions for Future Research

In this paper, we posit an improved Bayesian method of estimating the incidence of inefficiency which accounts for both misclassification that occurs in DEA and COD which may exacerbate misclassification. We use a triangular prior based on its flexibility to model a variety of different distributional shapes, as well as its theoretical and computational simplicity. Results from an empirical application and a simulation analysis suggest that in small samples (where the COD is most likely to distort DEA estimates of inefficiency) the proposed prior distribution out-performs an ignorance prior.

While our analysis provides some interesting insights, it is intended as a first step, and our findings should be interpreted with a degree of caution. We assume, for example, a specific type of skewness in our discrete triangular distribution, which may not be applicable to all production processes. We also assume that DMUs are sampled without replacement from a finite population, as this is the most common practice in empirical research. However, researchers who have the luxury of assuming sampling with replacement from a

finite population, or sampling (with or without replacement) from an infinite population may be able to derive and employ computationally simpler prior (and posterior) distributions than those contained in this manuscript.

Acknowledgement

This study was supported by Scientific and Technological Council of Turkey (TÜBİTAK) (project no. 2214-A).

References

Adler, N., & Golany, B. (2002). Including principal component weights to improve discrimination in data envelopment analysis. *Journal of the Operational Research Society*, 53(9): 985-991.

Araujo, C., Barros, C., & Wanke, P. (2014). Efficiency determinants and capacity issues in Brazilian for-profit hospitals. *Health Care Management Science*, 17: 126-138.

Augustyniak, W. (2014). Efficiency change in regional airports during market liberalization. *Economics & Sociology*, 7(1): 85-93.

Badin, L., & Simar, L. (2009). A bias-corrected nonparametric envelopment estimator of frontiers. *Econometric Theory*, 25: 1289-1318.

Badin, L., Daraio, C., & Simar, L. (2014). Explaining inefficiency in nonparametric production models: The state of the art. *Annals of Operations Research*, 214: 5-30.

Charnes, A., Cooper, W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2(2): 429–444.

Debreu, G. (1951). The coefficient of resource utilization. *Econometrica*, 19: 273-292.

Evans, J., & Olson, D. (2003). *Statitsics, Data Analysis, and Decision Modeling, 2nd Edition.* Prentice Hall, Upper Saddle River, NJ.

Fare, R., & Primont, D. (1995). *Multi-Output Production and Duality: Theory and Applications*. Kluwer Academic Publishers, Boston, MA.

Farrell, M. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society, Series A*, 120: 253-281.

Friesner, D., Mittelhammer, R., & Rosenman, R. (2013). Inferring the incidence of industry inefficiency from DEA estimates. *European Journal of Operational Research*, 224(2): 414–424.

Gajewski, B., Lee, R., Bott, M., Piamjariyakul, U., & Taunton, R. (2009). On estimating the distribution of data envelopment analysis efficiency scores: an application of nursing homes' care planning process. *Journal of Applied Statistics*, 36 (9): 933–944.

Gattoufi, S., Oral, M., & Reisman, A. (2004). Data envelopment analysis literature: a bibliography update (1951-2001). *Journal of Socio-Economic Planning Sciences*, 38(2-3), 159-229.

Gijbels, I., Mammen, E., Park, B., & Simar, L. (1999). On estimation of monotone and concave frontier functions. *Journal of the American Statistical Association*, 94 (445), 220–228.

Giraleas, D.,Emrouznejad, A., & Thanassoulis, E. (2012). Productivity change using growth accounting and frontier-based approaches – Evidence from a Monte Carlo analysis. *European Journal of Operations Research,* 222(3), 673–683.

Hatami-Marbini, A., Emrouznejad, A., & Tavana, M. (2011). A taxonomy and review of the fuzzy data envelopment analysis literature: two decades in the making. *European Journal of Operational Research,* 214(3): 457-472.

Hollingsworth, B. (2003). Non-parametric and parametric applications measuring efficiency in health care. *Health Care Management Science*, 6(4): 203-218.

Kneip, A., Simar, L., &Wilson, P. (2003). Asymptotics for DEA estimators in non-parametric frontier models,Discussion paper #0317, Institut de Statistique, Universite Catholique de Louvain, Louvain-la-Neuve, Belgium.

Manly, B.F.J. (2006). *Randomization, Bootstrap and Monte Carlo Methods in Biology, 3th Edition.* Taylor & Francis Group, Chapman&Hall/CRC, Boca Raton, FL.

McLaughlin, D., & Hays, J. (2008). *Healthcare Operations Management.* Health Administration Press, Chicago, IL.

Simar, L., & Wilson, P. (2007). Estimation and inference in two-stage, semiparametric models of production processes. *Journal of Econometrics,* 136: 31–64.

Ueda, T., & Hoshiai, Y. (1997). Application of principal component analysis for parsimonious summarization of DEA inputs and/or outputs. *Journal of the Operations Research Society of Japan*, 40(4): 466-478.

Unsal, M.G., Alp, I., & Bal, H. (2014). Empirical Distribution of Incidence of Inefficiency. *Journal of Selcuk University Natural and Applied Science*, 3(2): 43-50.

Yap, G., Ismail, W., & Isa, R. (2013). An alternative approach to reduce dimensionality in data envelopment analysis. *Journal of Modern Applied Statistical Methods,* 2012(1), 128-147.

Table 1. Empirical Distribution Results for a Single Ray Technology

Number of DMUs	**Number of Inputs**	**Number of Outputs**	**Rate***	**Expected Percent Inefficient DMUs**	**Variance**	**Skewness**	**Kurtosis**
20	2	1	5.00	0.8427	0.0020	-0.4505	3.1375
20	2	3	3.33	0.6107	0.0088	-0.1540	2.9242
20	2	5	2.50	0.5070	0.0110	-0.0697	2.9030
20	2	8	1.82	0.4168	0.0119	0.0097	2.8560
20	2	10	1.54	0.2675	0.0101	0.2962	2.9299
20	2	12	1.33	0.2351	0.0094	0.3971	2.9364
20	2	14	1.18	0.1743	0.0072	0.7903	3.2996
20	2	16	1.05	0.1531	0.0063	1.0418	3.6850

Table 2. Empirical Distribution Results for a Two Ray Technology

Number of DMUs	Number of Inputs	Number of Outputs	Rate*	Expected Percent Inefficient DMUS	Variance	Skewness	Kurtosis
20	1	1	6.67	0.8003	0.0027	-0.3417	3.0446
20	1	2	5.00	0.6421	0.0072	-0.1807	2.9407
20	1	3	4.00	0.4965	0.0099	-0.0301	2.9262
20	1	4	3.33	0.3717	0.0106	0.0963	2.9101
20	1	5	2.86	0.2714	0.0095	0.2659	2.9252
20	1	6	2.50	0.1962	0.0075	0.5502	3.0740
20	1	7	2.22	0.1427	0.0053	1.0743	3.7350
20	1	8	2.00	0.1085	0.0037	1.6961	4.9301

*Note: The rate column indicates the ratio of number of DMUs to total number of variables plus one.

Table 3. Estimated Incidence Values for a Finite Population, Sampling without Replacement and a Single Ray Technology

Sample Size	True Value	Non-Latent Variable		Latent Variable	
		Proposed	Friesner et. al.	Proposed	Friesner et. al.
n=5	0.9992	0.2164	0.2091	0.7503	0.7338
n=10		0.4010	0.3588	0.8224	0.8137
n=15		0.4467	0.4231	0.8583	0.8532
n=20		0.5312	0.5100	0.8766	0.8746
n=40		0.7438	0.7377	0.9282	0.9265
n=60		0.8242	0.8213	0.9622	0.9617
n=80		0.8649	0.8633	0.9745	0.9741
n=100		0.8883	0.8872	0.9804	0.9803
n=5	0.9980	0.2337	0.2038	0.7632	0.7502
n=10		0.5035	0.4632	0.8654	0.8599
n=15		0.6277	0.6066	0.9040	0.9010
n=20		0.7074	0.6941	0.9252	0.9233
n=40		0.8091	0.8045	0.9604	0.9599
n=60		0.9017	0.9002	0.9724	0.9721
n=80		0.9358	0.9350	0.9787	0.9785
n=100		0.9486	0.9481	0.9824	0.9823
n=5	0.9900	0.3711	0.3144	0.7625	0.7494
n=10		0.5794	0.5461	0.8643	0.8587
n=15		0.6854	0.6650	0.9048	0.9019
n=20		0.8296	0.8043	0.9263	0.9244
n=40		0.8534	0.8505	0.9598	0.9593
n=60		0.8726	0.8715	0.9679	0.9676
n=80		0.9051	0.9044	0.9771	0.9769
n=100		0.9588	0.9586	0.9789	0.9788
n=5	0.9267	0.2095	0.2053	0.7500	0.7333
n=10		0.4535	0.4137	0.8420	0.8338
n=15		0.5861	0.5669	0.8690	0.8641
n=20		0.6499	0.6359	0.8798	0.8754
n=40		0.6638	0.6580	0.8993	0.8976
n=60		0.7105	0.7077	0.9040	0.9031
n=80		0.8109	0.8098	0.9050	0.9045
n=100		0.8542	0.8537	0.9097	0.9094

Table 4. Estimated Incidence Values for a Finite Population, Sampling without Replacement and Two or Multiple Rays Technology

Sample Size	True Value	Non-Latent Variable		Latent Variable	
		Proposed	Friesner et. al.	Proposed	Friesner et. al.
n=5	0.9976	0.4162	0.3567	0.7539	0.7380
n=10		0.5018	0.4614	0.8492	0.8414
n=15		0.5851	0.5616	0.8660	0.8613
n=20		0.6422	0.6259	0.8748	0.8730
n=40		0.7663	0.7608	0.9277	0.9260
n=60		0.8294	0.8267	0.9618	0.9613
n=80		0.8588	0.8571	0.9736	0.9733
n=100		0.8784	0.8772	0.9797	0.9796
n=5	0.9760	0.4061	0.3468	0.7479	0.7302
n=10		0.5071	0.4679	0.8442	0.8358
n=15		0.5888	0.5662	0.8780	0.8727
n=20		0.6318	0.6158	0.9072	0.9046
n=40		0.7628	0.7575	0.9396	0.9387
n=60		0.8100	0.8073	0.9509	0.9504
n=80		0. 8414	0. 8397	0.9542	0.9538
n=100		0.8532	0.8521	0.9566	0.9563
n=5	0.9450	0.4020	0.3438	0.7474	0.7295
n=10		0.5012	0.4615	0.8399	0.8310
n=15		0.5785	0.5553	0.8749	0.8693
n=20		0.6444	0.6295	0.8943	0.8905
n=40		0.7643	0.7598	0.9152	0.9136
n=60		0.8342	0.8324	0.9242	0.9233
n=80		0.8672	0.8663	0.9255	0.9250
n=100		0.8809	0.8803	0.9279	0.9275
n=5	0.7933	0.3167	0.2743	0.6791	0.6462
n=10		0.3417	0.2989	0.7207	0.7036
n=15		0.3816	0.3579	0.7403	0.7297
n=20		0.4027	0.3791	0.7767	0.7679
n=40		0.5829	0.5756	0.7849	0.7810
n=60		0.6536	0.6504	0.7868	0.7856
n=80		0.7007	0.6991	0.7891	0.7869
n=100		0.7332	0.7324	0.7902	0.7880

Fig 1. Histograms for Empirical Distributions of Incidence of Inefficiency in Single Ray Technology

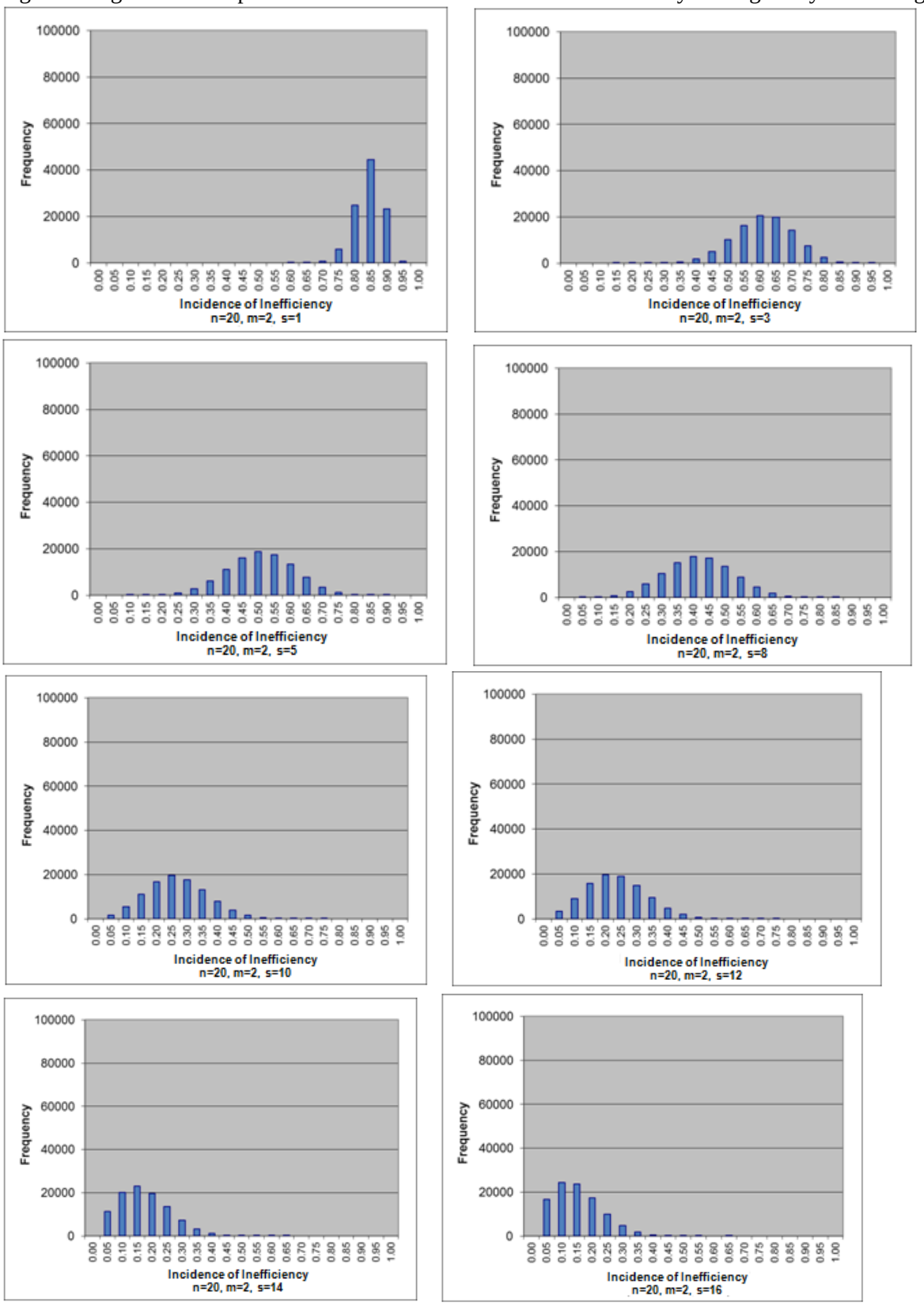
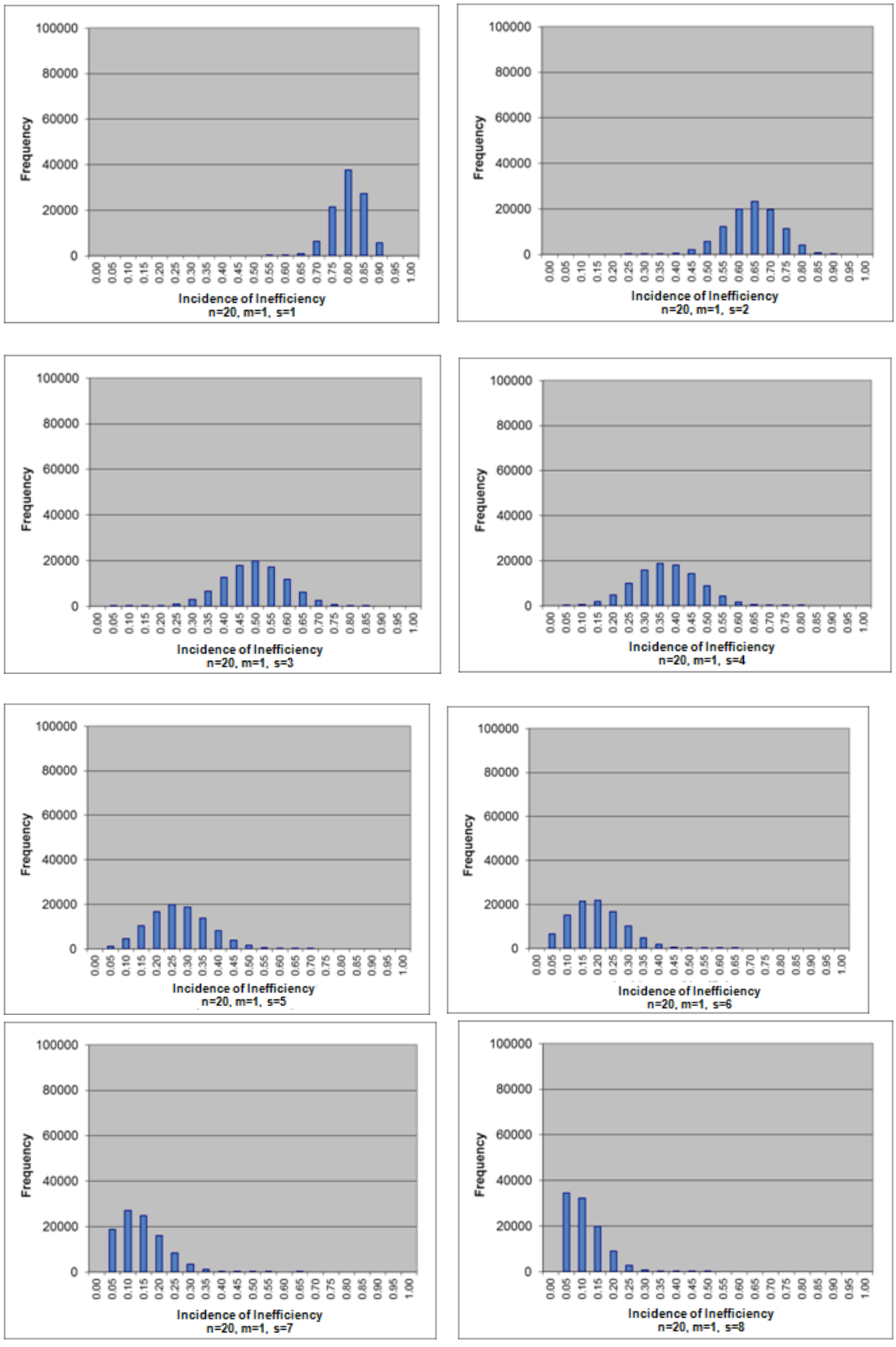


Fig 2. Histograms for Empirical Distributions of Incidence of Inefficiency in Multiple Ray Technology



Appendix. Application Data for Turkish Electric Companies

	Number of Transformator (x_1)	Number of Staff (x_2)	Lenght of Cables, km (x_3)	Number of Customer (y)
E1	36617,00	1306,00	50143,40	1220735,00
E2	8386,00	523,00	30408,30	458746,00
E3	10481,00	1072,00	47015,90	794177,00
E4	10632,00	360,00	49598,60	1103980,00
E5	9349,00	319,00	39438,00	734391,00
E6	9945,00	450,00	38215,20	709790,00
E7	35606,00	1635,00	76602,50	2878630,00
E8	41284,00	626,00	75274,10	1667752,00
E9	27088,00	1284,00	101616,90	3448457,00
E10	15614,00	781,00	55754,40	1676546,00
E11	24825,00	1001,00	47410,30	2488437,00
E12	21086,00	1079,00	55491,10	2445117,00
E13	9505,00	360,00	17949,40	849714,00
E14	5617,00	1681,00	18429,80	2388702,00
E15	13248,00	760,00	4542206,00	1435516,00
E16	19230,00	650,00	39803,50	1362922,00
E17	12229,00	1522,00	32468,40	4202132,00
E18	5462,00	634,00	17917,90	578438,00
E19	19116,00	464,00	66472,50	1555424,00
E20	9314,00	455,00	23375,40	524972,00
E21	16340,00	605,00	80527,90	1610685,00