HEDG Health, Econometrics and Data Group
THE UNIVERSITY of York

HEDG Working Paper 07/14

The quantile regression approach to efficiency measurement: insights from Monte Carlo Simulations

Chungping. Liu
Audrey Laporte
Brian Ferguson

July 2007

york.ac.uk/res/herc/hedgwp

The Quantile Regression Approach to Efficiency Measurement: Insights from Monte Carlo Simulations

Chunping Liu, MA
Department of Economics
University of Guelph
Guelph, Ontario, Canada N1G 2W1
E-mail: **liuc@uoguelph.ca**
Fax: (519) 763-8497

Audrey Laporte, PhD
Department of Health Policy Management & Evaluation
Faculty of Medicine
University of Toronto
155 College Street, 4th Floor
Toronto, Ontario, Canada M5S 1A8
E-mail **audrey.laporte@utoronto.ca**
Fax: 416-978-7350

Brian Ferguson, PhD
Department of Economics
University of Guelph
Guelph, Ontario, Canada N1G 2W1
Email: **brianfer@uoguelph.ca**
Fax: (519) 763-8497

Abstract:

In the health economics literature there is an ongoing debate over approaches used to estimate the efficiency of health systems at various levels, from the level of the individual hospital- or nursing home –up to that of the health system as a whole. The two most widely used approaches to evaluating the efficiency with which various units deliver care are non-parametric Data Envelopment Analysis (DEA) and parametric Stochastic Frontier Analysis (SFA). Productivity researchers tend to have very strong preferences over which methodology to use for efficiency estimation. In this paper, we use generated experimental datasets and Monte Carlo simulation to compare the performance of DEA and SFA in terms of their ability to accurately estimate efficiency. We also evaluate Quantile regression as a potential alternative approach. A Cobb-Douglas production function, random error terms and a technical inefficiency term with different distributions are used to calculate the observed output. The results, based on these experiments, suggest that neither DEA nor SFA can be regarded as clearly dominant, and that Quantile regression because it yields more reliable estimates, represents a useful alternative approach in efficiency studies.

Keywords: Technical efficiency, data envelopment analysis, stochastic frontier estimation, quantile regression.

I. Introduction:

Efficiency measurement, whether at the level of the individual physician, the hospital or the health care system as a whole, is a topic of continuing interest in the health economics literature with dispute ranging from the appropriate efficiency concept to use to the appropriate measure to use. In fact, the feasibility of efficiency estimation is itself the subject of debate – Newhouse (1994) argues that there are so many problems with any current attempts to accurately measure efficiency that efficiency scores are of virtually no practical policy value. Nevertheless, the ability to measure efficiency continues to be of interest to analysts and to decision-makers at all levels of government who are charged with the responsibility of allocating scare health care resources across competing needs[1].

In this paper we deal with what is termed technical efficiency. A production unit (referred to as a Decision Making Unit or DMU), whether an individual producer or an industry, is said to be technically efficient if its output mix lies on the production possibility frontier defined for its particular input levels. The question of interest is whether, given the set of inputs available and the vector of outputs the DMU has chosen to produce, its output point lies on or below its production possibility frontier. In the case of a single output, of course, technical efficiency refers to whether the producer is operating on or below its production function.

Technical efficiency is not full economic efficiency: there is also the issue of allocative efficiency, which asks whether the producer is not only on the production possibility frontier but at the right point on it given the prices - monetary or shadow - which it faces for its output. In this paper we do not deal with allocative questions, focusing solely on the measurement of technical inefficiency. The aim here is to compare two approaches which have been used fairly widely in the health economics literature, along with a third approach which a few authors have

[1] See Greene (2004) and Jacobs et al. (2006).

experimented with. The two widely used methods, Data Envelopment Analysis (DEA) and Stochastic Frontier Analysis (SFA), have been polarizing elements in the efficiency literature, with each attracting fervent supporters and equally dedicated opponents, and with the advocates of one approach tending to be fierce critics of the other approach[2]. The third approach, Quantile regression analysis, is a technique which has been familiar in the econometrics literature but which has come into wider use in recent years, although not in the context of efficiency measurement[3].

The purpose of this paper is twofold. First, to use a Monte Carlo approach to evaluate the performance of DEA and SFA in the estimation of technical efficiency. Second, to determine whether Quantile regression represents an alternative estimator which avoids a number of the problems associated with these existing measures.

II. Techniques of Efficiency Analysis: An Overview

In our discussion of efficiency measures we follow Farrell(1957), Charnes et al. (1975, 1977 and 1978), and Fare, Grosskopf and Lovell (1985, 1994) in presenting the concept of technical efficiency which deals with whether a Decision Making Unit (DMU) is producing maximal output using a given set of inputs.

In standard microeconomic theory the concept of technical efficiency raises no particular difficulties, especially in the case with which we shall deal here - that of a firm using several inputs to produce a single output. Figure 1 shows the textbook illustration of single input-single output production function. A firm is operating in a technically efficient manner when it is on the frontier (the production function) and it is being technically inefficient when it is operating

[2] See Rowena Jacobs, Peter C. Smith and Andrew Street (2006): Measuring Efficiency in Health Care Cambridge University Press, for a discussion of some issues dividing the SFA and DEA camps.
[3] For one application of Quantile estimation in the efficiency literature, see Bernini et al. (2004).

below the frontier. In terms of Figure 1, firms C, D, E and F are technically efficient since they are on the frontier and firms A and B are technically inefficient since they operate below the frontier. Because the production function of microeconomic theory is a maximum value function, showing the maximum level of output a firm can obtain for any given level of input, it is not possible for the firm to operate above its production function. This point is at the core of the dispute between supporters of DEA and supporters of SFA, so we will return to it below.

Technical inefficiency, then, is defined as the firm lying in the interior of its feasible production set, but while this can obviously be characterized as "lying below its production function", there are a number of ways the degree of inefficiency could, in principle, be measured.

The most obvious direction of measurement is what is referred to as output oriented inefficiency measurement, which measures the vertical distance from the firm's actual production point to the frontier. Basically this approach asks by how much the producer could increase its output with no change in its input use if it were to operate in a fully technical efficient manner - i.e., how far the producer's current, actual production point lies below the production frontier. Again, while the concept is straight forward there are several ways this distance could be measured. The approach which is employed in the Monte Carlo experiments to follow, is to take the point on the production frontier as the basis for comparison and then to assess the firm's actual output as a percentage of its potential output. In this approach, 1 represents full efficiency, and a firm that was operating at 10% below full technical efficiency would have a score of 0.90.[4]

Since the production function is never known in practice, it must be estimated from

[4] In addition to output oriented inefficiency measurement, the literature also defines input-oriented inefficiency measurement. In other words, instead of asking how much more output a producer could get from its current input mix were it to operate in a fully technically efficient manner the question is by how much moving to a technically efficient point on the frontier would allow it to reduce its input use while continuing to produce the same level of output as before. For simplicity we focus on output oriented efficiency.

sample data. Farrell (1957) suggested that it could be estimated using either a non-parametric technology or a parametric form, such as the Cobb-Douglas production function. Charnes, Cooper and others developed the non-parametric Data Envelopment Analysis (DEA) approach while Aigner, Lovell and Schmidt (1977) and Meeusen and van den Broeck (1977) proposed the parametric Stochastic Frontier Analysis (SFA) approach. These are the two most commonly used approaches to estimating efficiency, differing in terms of the econometric approaches and the assumptions used to fit the efficiency frontier. The third approach we consider, Quantile regression, fits into the mix as a semi-parametric method.

DEA is a non-parametric, linear-programming approach because it makes no assumptions about the form of the production frontier, or about the statistical distribution of the inefficiency scores, and does not attempt to estimate the parameters of the production function. Essentially, to use the single input-single output case as an example, it uses linear segments to construct an envelope of all of the observed production points, so that each point in the data set is either on or below the convex hull made up of the linear segments. A point which is on the convex hull is taken to be efficient, a point below it is defined as inefficient with the degree of inefficiency indicated by some measure of the distance to the hull, where the direction of motion towards the hull, as well as the measure to be used, must in general be specified.

Perhaps the most common criticism of the DEA approach is that it takes the definition of the production frontier as a maximum value function too seriously, at least for empirical purposes. This view argues that there will always be some noise in output data, perhaps because of measurement error, perhaps because of random factors which could affect the output of any given production unit at any given time. This would seem particularly likely in health economics applications, where the output measure is often for example, based on mortality. There could

also, presumably, be measurement errors in the input data, especially when, for example, measures of labour quantity or time at work is used to proxy labour effort. Since DEA effectively acts by grouping together observations with the same input levels and selects the observation with the highest output level among them as the most efficient unit for that cluster, the hull which it maps out could be affected to a significant degree by the presence of random disturbances in the data. How serious a problem this might be would, of course, depend on how large the error terms were relative to the output levels.

Critics of the non-parametric DEA methodology generally prefer some version of Stochastic Frontier Analysis (SFA)- an econometric approach which requires one to specify the functional form of the production function. It differs from OLS estimation of a production frontier in that it assumes the presence of two random elements. One is the usual random disturbance term, while the other is an efficiency scaling term, representing the degree of technical inefficiency in the production units in the data set. This methodology assumes that inefficiency, meaning the tendency of some observations to lie below the frontier, can be characterized by a one-sided probability distribution. This does not mean that inefficiency is regarded as a purely random shock - were that the case it could simply be subsumed into the disturbance term, and there would be no reason to assume that any individual unit might be chronically inefficient. The assumption that inefficiency fits a probability distribution is really a way of recognizing the fact that, because economic analysis deals with the behaviour of optimizing agents, we have no good models of pure inefficiency. SFA uses maximum likelihood simultaneously to estimate the production frontier and to allocate deviations between individual observations and the estimated frontier to the two random terms - the one sided density function which represents inefficiency and the standard normal term which represents random

disturbances to output. The main concern with the SFA approach lies in the fact that because it uses maximum likelihood it requires that we make an assumption about functional form for the inefficiency distribution, in practice usually either a half-normal or an exponential distribution, which raises the possibility that misspecification of the inefficiency distribution could bias all of the results of the estimation exercise, including the estimates of the coefficients of the production function. This concern is one on which we shall focus in the Monte Carlo experiments.

The third alternative to be considered in the simulation exercise is Quantile regression, an estimation technique which has come into wider use in empirical work as large micro data sets have become available. Ordinary Least Squares regression yields a conditional expected value function for the dependent variable - a function which allows for the calculation of the expected value of the dependent variable given values of the explanatory variables. In working with large micro data sets it is likely that even well-behaved equations (ones with large t- and F-statistics, for example) have low R^2 values, simply because the data are so widely scattered around the OLS line. Traditionally, in looking at the properties of the scatter of observations around their estimated conditional mean the focus has been simply to check for heteroscedasticity. Quantile regression extends the analysis of the distribution of the observed value of the dependent variable around its expected value by fitting equations characterizing the expected conditional quantiles of that distribution. Thus, just as OLS yields an equation characterizing the way the mean of the observations on the dependent variable is expected to change as the values of the explanatory variables change, so quantile regression produces equations which can be used to observe how the spread of the distribution around the mean changes. Quantile regression is an extension of Least Absolute Deviation (LAD) estimation, which yields an equation for the conditional median of the dependent variable. LAD estimation is sometimes used as a way of reducing the impact of

large outliers on the estimated conditional function for a measure of central tendency, and quantile regression can be used the same way. In our case, quantile regression offers an alternative to OLS as a method of estimating the production frontier. Since inefficient firms will lie below the true frontier, the presence of a handful of highly inefficient producers might bias the OLS-estimated location of the production function downward i.e. may pull the estimated curve below the true frontier. By choosing one of the upper quantiles to estimate - here we have arbitrarily chosen to estimate an equation for the 80^{th} percentile - the effect is to down-weight any unusually low values of the observed dependent variable, on the assumption that they are likely to represent inefficient firms and, presumably, this will yield an estimate of the production frontier which is closer to the true than would obtain using OLS. SFA also attempts to remove the effect of particularly inefficient observations on the estimate of the production frontier, but does it, as noted above, by making assumptions about the parametric form of the distribution of the inefficiency terms.

Quantile regression is a semi-parametric approach, which requires an assumption about the functional form of the frontier (unlike DEA, which does not give estimates of the curvature of the frontier, nor of the marginal productivity of the inputs) but unlike SFA, does not require the imposition of a particular form on the distribution of the inefficiency terms. The true distribution of the inefficiency term is never known in practice, so quantile regression avoids imposing strict assumptions on the inefficiency terms. Quantile regression also avoids the criticism aimed at DEA of not allowing for random error in the observed values of the dependent variable. If we assume that there is a random disturbance term in the observed value of the dependent variable, there is always a concern that the DEA fitted frontier will be dominated by the observations which just happen to have experienced the largest positive shocks. Quantile

regression allows observations to lie above the fitted curve as a result of pure chance. The fitted equation for the chosen quantile can then be used as the estimate of the production frontier where we assume that observations on or above it are efficient and that ones lying below it are likely to be inefficient, and use some measure of the distance from observations below the frontier to the frontier itself as the measure of their inefficiency. Clearly, as in the case of DEA in the presence of a random disturbance term, this process runs the risk of classifying some efficient but slightly unlucky producers (i.e. ones which happen to have had a negative output shock in the period from which the data are drawn) as inefficient, so the best bet is probably (as would also be the case with DEA) to treat small inefficiencies as measurement error and focus on large ones. Our expectation is that the consequences for the estimation of the frontier of mislabelling observations this way will be less for Quantile regression than for DEA.

The literature on efficiency measurement contains a number of papers which compare DEA and SFA results (see Jacobs, Smith and Street (2006) for a discussion and an example) but most of these papers apply the two approaches to real-world data, and compare the efficiency rankings of individual DMUs generated by the two approaches. The concern about this approach is that the true efficiency scores of the individual DMUs are unknown, so we cannot in general say with confidence which approach does better. That issue could be dealt with by applying the two approaches to artificial data sets, in which the efficiency properties of the DMUs are known in advance. This approach is usually done in the context of a Monte Carlo analysis, and while there have been a few Monte Carlo studies (see Gong and Sickels (1992), Bojanic et al.(1998), Yu(1998), Resti(2000) and Banker et al. (2004)), in this paper we add the Quantile regression approach and look at a slightly different set of questions about the results of the three approaches.

III. **Methods**

In this paper we consider the application of all three methods of estimating the degree of output inefficiency in a Monte Carlo framework. We generate values of the dependent variable, output, from a production function whose parameters we have specified, then add a disturbance series generated from a standard normal distribution and add inefficiency terms generated from both half-normal and exponential distributions. (The parameters used are reported in detail below.). These efficiency terms, which indicate how far below the frontier the actual output level for each firm will lie, basically scale down the frontier output value for each firm to yield what we refer to as the actual non-stochastic output level. If a firm has a technical efficiency term of 0.8, its actual non-stochastic output level will be equal to 80% of the frontier output level associated with its input mix. A firm with a technical efficiency score of 1 has its non-stochastic output level on the frontier (its actual, stochastic output level might lie above or below the frontier because of the random disturbance term).

In particular, we use a Cobb-Douglas production function with multiple inputs because it is commonly used, simple and a well-accepted production function form. The Cobb-Douglas production function we use is:

$$y_i = \alpha_0 x_{1i}^{\alpha_1} x_{2i}^{\alpha_2} x_{3i}^{\alpha_3} x_{4i}^{\alpha_4} \qquad (7)$$

Where i=1,…,n, n is the sample size,

α_m is a positive number, m=0,1,…,4 inputs

We also generate random error term, $v_i \sim i.i.d(0, \sigma_v^2)$, and technical efficiency term, u_i. The technical efficiency term u_i generated for the first runs satisfies the half normal distribution, that used for the second set of runs satisfies the exponential distribution. After including the two random terms, v_i and u_i, the Cobb-Douglas production function becomes:

$$y_i = \alpha_0 x_{1i}^{\alpha_1} x_{2i}^{\alpha_2} x_{3i}^{\alpha_3} x_{4i}^{\alpha_4} e^{v_i} e^{-u_i} \quad (8)$$

Where i=1,…,n and n=100.

All input variables are generated uniformly within a certain range. The values of the coefficients on the inputs are chosen so that the production function exhibits decreasing returns to scale, which seems realistic for a production process. Thus in the exercise reported here we set the sum of the coefficients to 0.53. The data description for the four input variables, random error term $v_i \sim \text{i.i.d}(0, 0.01)$, inefficiency terms u_{1i} and u_{2i}, true value of the intercept and true values of the coefficients on the inputs are presented in Table 1 below.

We run 100 replications for each experiment, so for each particular vector of input values there are 100 output values, all distributed around the same efficiency-scaled value of the non-stochastic output level, so that each firm has exactly the same input values and exactly the same degree of inefficiency in all of the 100 replications in a particular experiment. We conduct this exercise three times, once using SFA, once using DEA and once using Quantile regression. Under each method, for each replication we estimate each firm's technical efficiency (TE) score. For each firm, we are then able to compare the mean of the 100 TE estimates generated by each of the three methods with the true TE score which we had built into the data generating process (and which are held constant across replications within an experiment). Since each firm has the same efficiency score across experiments this provides an indication of how well each estimation method does at matching each firm's TE score.

All of the experiments below use the same values of the explanatory variables but we vary the values of the inefficiency term, starting with the half normal distribution and gradually moving away from that particular distribution, to investigate the effect of changes in the

distribution of the actual inefficiency terms on the values yielded by the three different approaches. We do this by increasing the number of efficient units, so that whereas in the first runs virtually all firms have some degree (typically small) of inefficiency, in later runs only a few firms are inefficient, but those relatively highly so. Since the parametric form of the technical inefficiency distribution is specified in the likelihood function for the SFA approach, we are particularly interested in the performance of this approach as the true specification of the inefficiency terms moves further and further away from the parametric form assumed in the estimation.

We expect this gradual shift of firms from less than full efficiency to full efficiency to have some effect on the SFA estimates of the TE scores. The initial experiments, where the true TE scores are drawn from the half normal distribution and the likelihood function for SFA is written assuming a half normal distribution of the TE terms is clearly weighted heavily in favour of SFA. As we increase the proportion of efficient firms, the shape of the probability density function is changing for the empirical TE values, and the more the shape of that distribution changes, the less well specified is the likelihood function of the SFA procedure. We also expect the changes to the distribution of the TE scores to have some effect on the DEA estimates, since increasing the number of efficient firms should improve the fit of the DEA production frontier, although since the firms' output levels still have noise terms attached to them, we have no real sense a priori as to how the DEA results will be affected.

Having undertaken the experiments described above for TE terms generated (initially) with a half normal distribution, we then repeat the exercise for TE terms generated using the other common SFA distribution, the exponential. Apart from the switch in initial distribution, all parameters are as in the first set of exercises, and the likelihood function of the SFA estimator is

written using an exponential distribution for the TE terms, so SFA is initially fully correctly specified. As in the first case we conduct successive runs, gradually moving more and more of the observations up to full efficiency, causing SFA to be increasingly mis-specified.

Next, we investigate the effect on SFA of mis-specification of the likelihood function, by using TE terms drawn from the half normal distribution but running SFA on the assumption that they are drawn from the exponential (we also run DEA and Quantile regression for this case, but since neither of those methods requires that we make assumptions about the shape of the distribution of the TE terms, our interest is in the effect on SFA.) Again, we compound the mis-specification by shifting more and more observations up to full efficiency, and repeating the 100 Monte Carlo replications. Finally, this exercise is repeated for the case where the true TE distribution is exponential but SFA assumes it to be half normal.

VI. Monte Carlo Results

Of particular interest is how misspecification of the true distribution of the TE terms affects the performance of the three estimators. To investigate this, we begin by generating a series of TE scores from a known distribution - in this case we use the two commonly used distributions, the half normal and the exponential. Thus, in the first runs based on the half normal, we begin by assigning each observation a TE value drawn from a half normal distribution, and scaling the efficient output level down by that amount, then factoring in the random disturbance term and using each of the three estimation methods to estimate the true TE term.

In Figure A1 we have performed a Monte Carlo experiment involving 100 replications on 100 data points, for the special case where all of the non-stochastic observations are efficient, so

that the only reason an observation lies off the production function is random noise. In Figure A1, the 'True' line represents the true TE scores. In this figure, the True line is horizontal at 1 because none of the observations have been scaled down.

Of the three estimation techniques, SFA does best when there is no inefficiency - the SFA values are virtually horizontal at 97% efficient. Interestingly, even though there is no actual technical inefficiency, SFA did not identify any of the observations as being fully efficient. The DEA TE scores show the greatest variability, as we would expect, given DEA's sensitivity to the random disturbance term.

Quantile regression did reasonably well, placing most of the observations at between 91-92% efficient. Since we are fitting an equation for the 80^{th} percentile here, in any single run roughly 20% of the observations in the data set should show up as efficient. The reason none of the units are identified as efficient in the quantile regression output in Figure A1 is that each point plotted there is an average efficiency score from 100 replications. Because the 80^{th} percentile curve lies above the true production function, in the absence of any true technical inefficiency the Quantile approach as we use it will tend to underestimate the efficiency of individual units. Our interest in the Quantile approach is in its robustness in the presence of true technical inefficiency.

In our first set of replications, the true TE scores are generated using a half normal distribution, and the likelihood function for the SFA estimation assumes a half normal, so the SFA estimates are based, at least initially, on equations which are correctly specified both in terms of functional form and in terms of the assumed distribution of the inefficiency scores. We say initially, because we will gradually move the actual distribution of the TE scores away from

the half normal distribution from which they were generated, as our test for sensitivity of the estimation techniques to misspecification of the efficiency term.

Figure A2 shows the results from the first, fully correctly specified run when there is one fully efficient unit. The observations are ordered in terms of true technical efficiency, from the most to the least efficient, as shown by the downward slope of the true TE values on the graph. Again, the true TE score is the parameter to be estimated. Interestingly, given that the inefficiency terms are generated using the half normal and that the SFA is run assuming a half normal, SFA cannot be said to clearly dominate the other approaches. SFA underestimates the efficiency of the most efficient units (the ones on the left of the graph, whose TE scores are closest to 1). Furthermore, while SFA does do better than the other methods, in the sense that the SFA series lies closer to the true series than do either the DEA or the Quantile series up to about the 45^{th} observation, from the midpoint on SFA does not do better than Quantile regression, which up until that point had tended to overestimate the TE terms to a greater degree than did SFA. DEA also tracks the trend in the true value well, but the DEA series show some quite noticeable overestimates of the true TE scores, even for some of the least efficient observations.

From this point we begin to introduce the first misspecification, moving the true distribution of TE scores away from the half normal. This is done by moving the units which have the least inefficiency into the fully efficient class, by raising their TE scores to 1. Thus in Figure A3 we have raised the first 15 of 100 TE scores to one, leaving the remainder unchanged. It becomes apparent that SFA underestimates the efficiency of those efficient units to a greater degree than either DEA or Quantile regression do, and while SFA still dominates the other two approaches between about observations 15 and 45, for the second half of our data set, where true technical efficiency is least, the SFA line does quite noticeably worse than the Quantile

regression series in tracking the true TE line. DEA also tends to do better than SFA, although again with a few noticeable overestimates of true efficiency.

In Figure A4, 34 units have been moved up to TE values of 1, and the poor performance of SFA relative, in particular, to Quantile regression is quite marked. SFA tends, systematically, to overestimate the efficiency of the inefficient units and to underestimate the efficiency of the efficient ones. It is worth emphasizing at this point that since these are Monte Carlo results, the SFA line represents the average of 100 runs in which the true TE scores (and the values of the X variables) were held constant and only the values of the random disturbance terms changed across the 100 replications. This suggests that the poor performance of SFA is a general trend.

While the performance of SFA had been tending to worsen as we increased the number of efficient units until we reached the case shown in Figure A4, after this point, interestingly enough, the performance of SFA starts to improve. By the time the first 49 values of TE have been raised up to 1 the SFA line actually lies below not just the Quantile line but below the true line for the entire range of inefficient units (Figure A5). This pattern continues in Figure A6 where there are 60 fully efficient units, and remains consistent through the rest of the runs. By this stage, DEA and Quantile regression are also tending to underestimate the true TE values although not by as much, on average, as does SFA.

Overall, then, it seems that SFA is very sensitive to the particular form of misspecification which we have imposed in the experiments, in which we allow more units to be fully efficient than the half normal assumption which was built into the likelihood function for the SFA estimation was expecting.

In the next set of exercises we generate the true TE scores using the other distribution commonly assumed in the literature, the exponential. We follow the same procedure as above,

starting with true TE scores which are in fact drawn from an exponential distribution and gradually move away from that distribution by moving the least inefficient units up to TE values of 1. In this set of experiments, the likelihood function for the SFA estimator is specified using the exponential distribution for the technical efficiency term, so SFA is, at least initially, correctly specified. As in the previous case, the more relatively efficient observations we move up to full efficiency, the further from the exponential distribution the true TE distribution lies, and the greater the misspecification the SFA procedure must overcome.

Figure A7 shows the result of the Monte Carlo runs when the actual TE scores follow the exponential form which the SFA assumes. As expected, SFA does well, but the series of TE scores generated by the Quantile regression actually lie closer to the true than do the SFA values, and DEA also does quite well, again with the exception of a few cases where it notably overestimates the efficiency of some rather inefficient units. In this case no systematic pattern of bias emerges in the estimated TE scores, even as we move inefficient units onto the efficient frontier. In Figure A8 we show the case where there are 34 fully efficient units - the point at which the problem with SFA had become very clear in the half normal experiments and in Figure A9 the case of 60 efficient units. In general, when the TE scores are generated using an exponential distribution and the SFA likelihood function is written on the assumption that the TE scores come from an exponential distribution, SFA does not display the odd behaviour it did in the half normal case, even when the true distribution of the TE scores moves away from the distribution assumed in the half normal case. In general, though, the graphs for this series of experiments also show that the Quantile regression approach performs at least as well as the SFA approach. DEA demonstrates the same type of behaviour in this case as in the previous one, generally tracking well, but missing quite notably in a few particular cases.

While this latter result sounds encouraging for SFA, it does require that we have correctly specified the likelihood function for the SFA runs. In the next two series of Monte Carlo experiments we consider what happens when that assumption fails.

In the first of these two sets of experiments, we generated the data using a half normal distribution but wrote the likelihood function for SFA assuming an exponential distribution. Figure A10 depicts the case where we have not yet started to deviate from the half normal by moving units up to full efficiency, but it is evident that the effect of the misspecification is very serious for SFA. This pattern remains consistent for the case of 34 fully efficient units (not shown). Notably, the problem with SFA does in a sense correct itself. Figure A11 shows the case of 55 fully units, and SFA is tracking as well as Quantile and DEA estimation. Apparently by this stage so many observations have been moved out of the half normal (remembering that observations which were close to 1 in TE were gradually moved up to 1) that what remains resembles an exponential distribution since the tail of the half normal will be the last part of the distribution to be affected by the migrating values. Even at this point, though, SFA cannot be said to dominate the other two approaches.

For the last set of experiments we turn things around, generating the TE scores using an exponential distribution but assuming a half normal in the likelihood function for the SFA. Figure A12 is the usual starting point, and we see that, while the SFA tracks much better than in the previous case, it consistently underestimates the efficiency of the units, and is dominated throughout by Quantile regression. Because this same pattern continues through the remainder of the experiments, we do not report any more graphs here.

VI. Conclusions:

The results suggest that SFA may be very sensitive to misspecification of the assumed distribution of the technical inefficiency term, especially when the half-normal distribution is involved. SFA works well when the true distribution of the inefficiency scores is exponential, but for that to be useful information, we must know in advance that the inefficiency scores are indeed distributed as exponential. Absent that, the Monte Carlo exercise suggests that SFA can give very misleading results, especially as far as the least efficient firms (the ones in which we are, presumably, most interested) are concerned.

While DEA outperforms SFA in many ways in the experiments, it does have an odd tendency to persistently identify certain very inefficient DMUs as fully efficient. These DMUs can be identified as spikes in the DEA efficiency scores, occurring at the same points along the horizontal axis in each of the graphs of our experiments. Since the effect of the random disturbance term should have been averaged out in the Monte Carlo procedure this is not likely to be a consequence of DEA's sensitivity to disturbances[5]. It can be shown that each of the units whose efficiency was persistently overestimated was a low non-stochastic output unit, and preliminary investigations suggest that these spikes appear to reflect DEA's sensitivity to points at the extremes of the isoquant on which the firm would lie (corner solutions), were it operating in a fully efficient manner. It is important to note that the Monte Carlo approach may be weighted in favour of DEA, since the effects of the random disturbance term should average out in the production of the DEA average TE scores. In a single run, as would be the case when working with real world data drawn from a single year, DEA's fit of the production function is still expected to be sensitive to extreme values of the disturbance term.

[5] The graphs depict the average of 100 estimates of the TE scores, where the only thing which varied across the 100 runs was the random disturbance term, and that was drawn from a mean-zero distribution.

The results seem very strongly to favour the Quantile approach to fitting the frontier and deriving TE scores. Quantile regression performed more reliably than either DEA or SFA in the Monte Carlo runs, and we would expect its favourable performance to apply in a single run on real world data. Quantile regression combines elements of both DEA and SFA: for example, DEA can, in a sense, be regarded as fitting a series of linear splines to the 100^{th} percentile. Quantile regression has the advantage that, as with SFA, we can test for functional form on the production function, and test for the marginal productivity of the different inputs, but it is more robust than SFA to odd distributions of the TE term. While our choice of the 80^{th} percentile as the quantile function to be estimated was arbitrary, it is quite easy to vary the choice of quantile and test for sensitivity of the estimates of the TE scores. Overall, the Quantile regression approach seems to address many of the weaknesses associated with the DEA and SFA approaches and therefore appears to be a more robust approach for the estimation of technical efficiency.

References

Aigner, D.J., C.A.K Lovell and P. Schmidt (1977), "Formulation and Estimation of Stochastic Frontier Production Function Models", Journal of Econometrics, 6, 21-37.

Banker, R.D., Chang, H. and Cooper, W.W. (2004), "A Simulation Study of DEA and Parametric Frontier Models in the Presence of Heteroscedasticity", European Journal of Operational Research 153, 624-640.

Bernini, C., M. Freo and A. Gardini (2004): "Quantile Estimation of Frontier Production Function" Empirical Economics 29, 373-381

Bojanic, A.N., Caudill, S.B. and Ford, J.M., (1998), "Small-Sample Properties of ML, COLS, and DEA Estimators of Frontier Models in the Presence of Heteroscedasticity", European Journal of Operational Research 108,140-148.

Charnes, A., Cooper, W.W. and Rhodes, E. (1975), "Expositions, interpretations, and extensions of Farrell efficiency measures, Management Sciences Research Group Report, Pittsburgh", Camegie-Mellon University School of Urban and Public Affairs.

Charnes, A., Cooper, W.W. and Rhodes, E. (1977), "Measuring the efficiency of decision making units with some new production functions and estimation methods", Center for Cybernetic Studies Research Report CCS 276, Austin, TX, University of Texas Center for Cybernetic Studies.

Charnes, A., Cooper, W.W. and Rhodes, E. (1978), "Measuring the Efficiency of Decision Making Units", European Journal of Operational Research, 2, 429-444.

Fare, R., Grosskopf , S. and Lovell, C.A.K (1985), "The Measurement of Efficiency of Production", Kluwer Academic Publishers, Boston.

Fare, R., Grosskopf, S. and Lovell, C.A.K (1994), "Production Frontiers", Cambridge University Press, Cambridge.

Farrell, M. (1957). "The Measurement of Productive Efficiency," Journal of the Royal Statistical Society, Series A, 120, Part 3, 253-290.

Gong, B.H. and Sickles, R.C., (1992), "Finite Sample Evidence on the Performance of Stochastic Frontiers and Data Envelopment Analysis Using panel Data", Journal of Econometrics 51, 259-84.

Greene, W. (2004): "Distinguishing between heterogeneity and inefficiency: stochastic frontier analysis of the World Health Organization's panel data on national health care systems"_Health Economics 13(10), September, 959-980

Jacobs, R., Smith, P.C. and Street, A., (2006), "A Comparison of SFA and DEA", In "Measuring Efficiency in Health Care", Cambridge University Press, 151-167.

Koenker, R. and Bassett, (1978b), "Regression Quantiles", Econemetrica 46: 33-50.

Koenker, R., and Hallock, K.F. (2001), "Quantile Regression", Journal of Economic Perspectives, 15: 143-156.

Kumbhakar, S. C. and Knox Lovell, C. A. (2000), Stochastic Frontier Analysis, Cambridge: Cambridge University Press.

Meeusen, W. and J. van den Broeck (1977), "Efficiency Estimation from Cobb-Douglas Production Functions With Composed Error", International Economic Review, 18, 435-444.

Newhouse, J. P., (1994), “Frontier Estimation: How Useful a Tool for Health Economics?”, Journal of Health Economics 13 (1994) 317-322.
Resti, A. (2000), “Efficiency Measurement for Multi-product Industries: A Comparison of Classic and Recent Techniques Based on Simulated Data”, European Journal of Operational Research 121, 559-578
Seiford, L.M. (1996), “Data Envelopment Analysis: The Evolution of the State of the Art (1978-1995)”, Journal of Productivity Analysis, 7, 99-138.
Yu, C. (1998), “The Effects of Exogenous Variables in Efficiency Measurement—A Monte Carlo Study”, European Journal of Operational Research 105, 569-580.

Figure 1

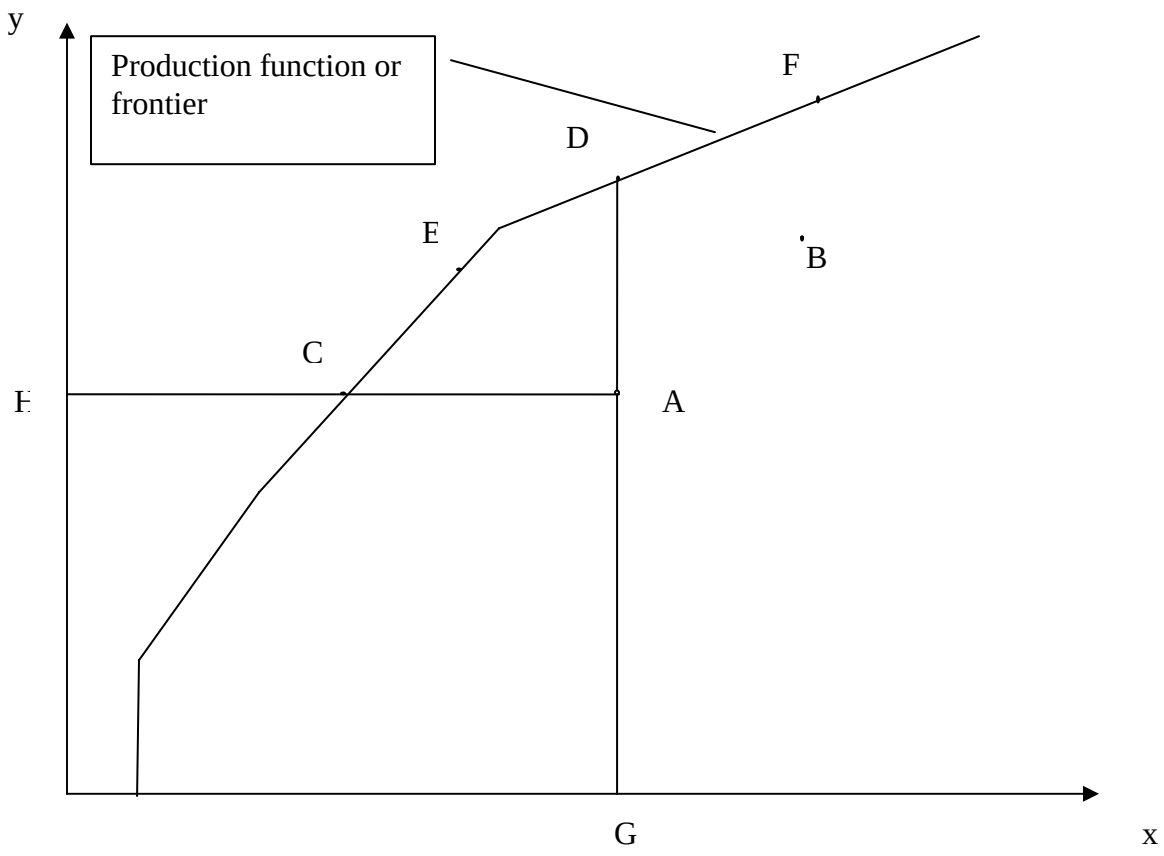

Table 1: DGP Data Description

Variable	Mean	Std. Dev.	Min	Max
x1	2.853314	1.191248	1.00782	4.896069
x2	8.579312	3.847942	2.154894	14.89015
x3	8.36448	4.026787	1.105346	14.99141
x4	442.871	180.6849	144.8173	745.4782
Technical efficiency, u_{1i} (half-normal)	0.3185883	0.2076957	0	0.882275
Technical efficiency, u_{2i} (exponential)	0.18432	0.19769	0	0.98666
True value	$\beta_0 = 50, \ln \beta_0 = 3.912, \beta_1 = 0.2, \beta_2 = 0.1, \beta_3 = 0.08, \beta_4 = 0.15$			

*These are the starting values for the Monte Carlo runs. The actual technical efficiency values are modified across runs as discussed in the text.

Figure A1

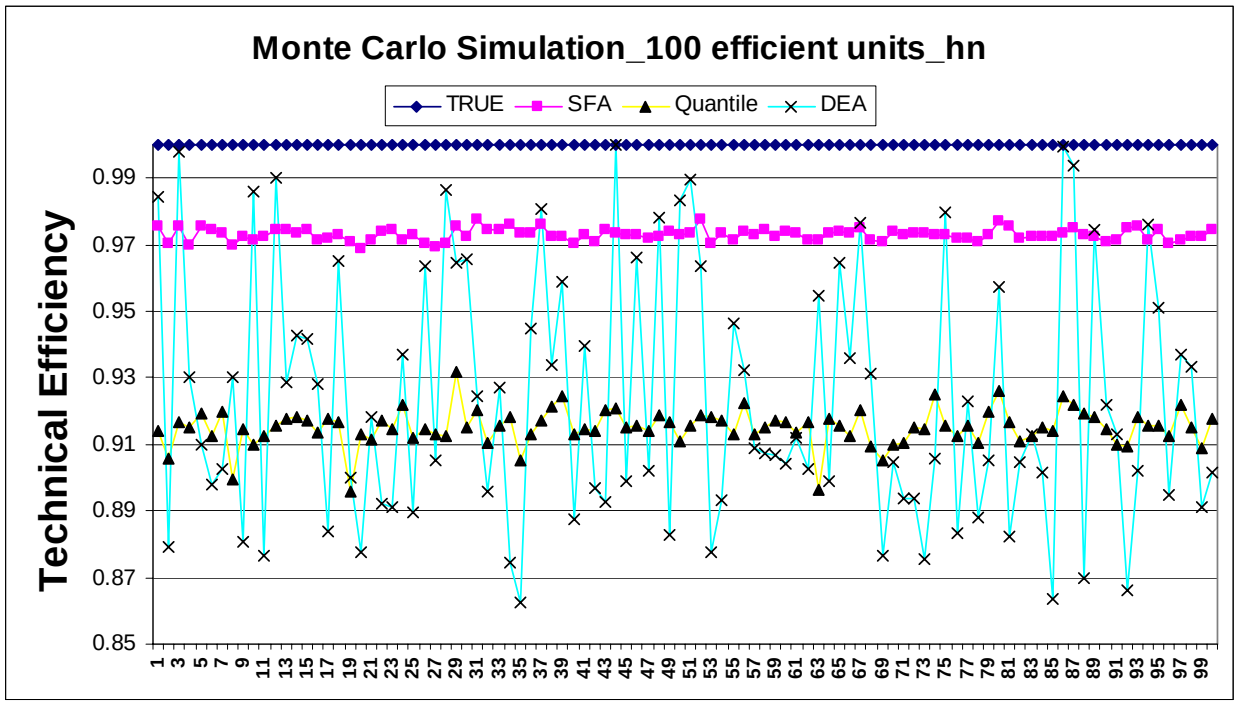

Figure A2

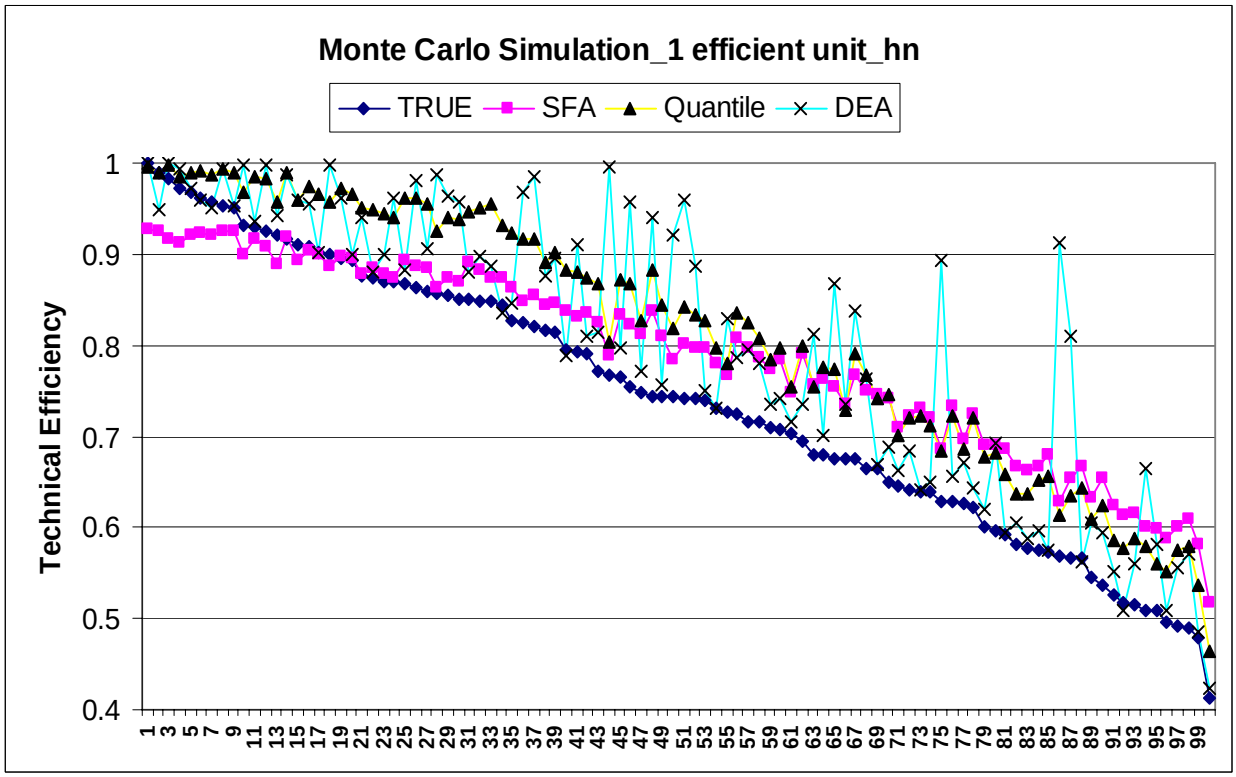

Figure A3

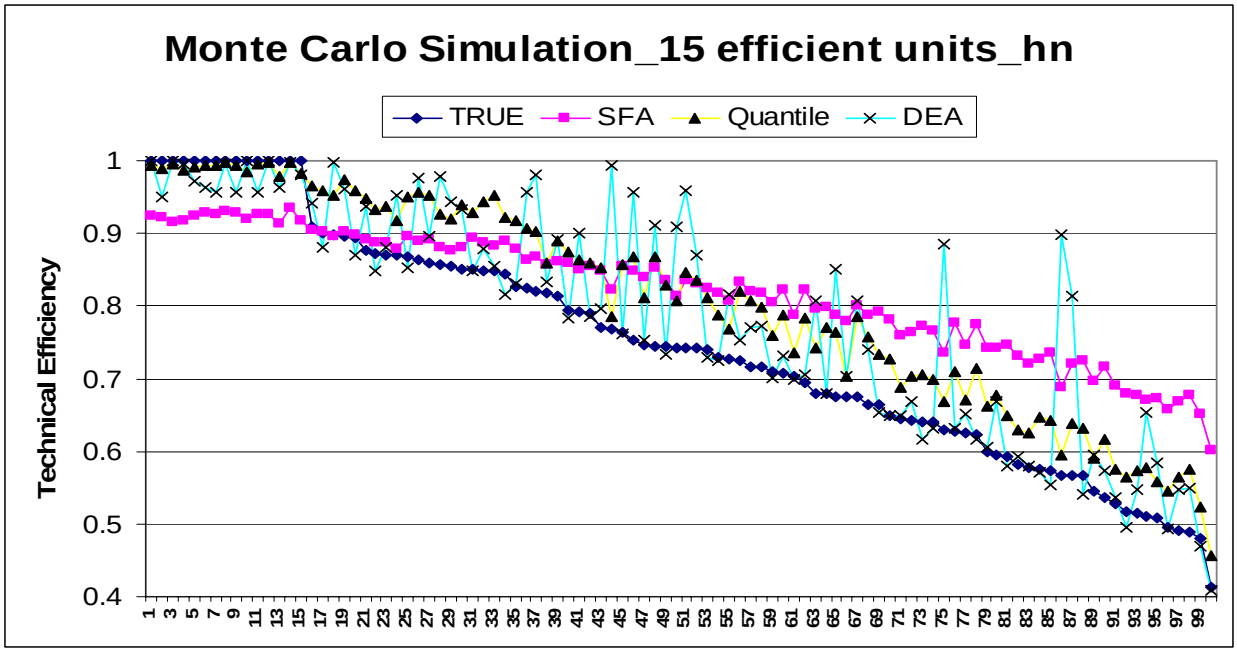

Figure A4

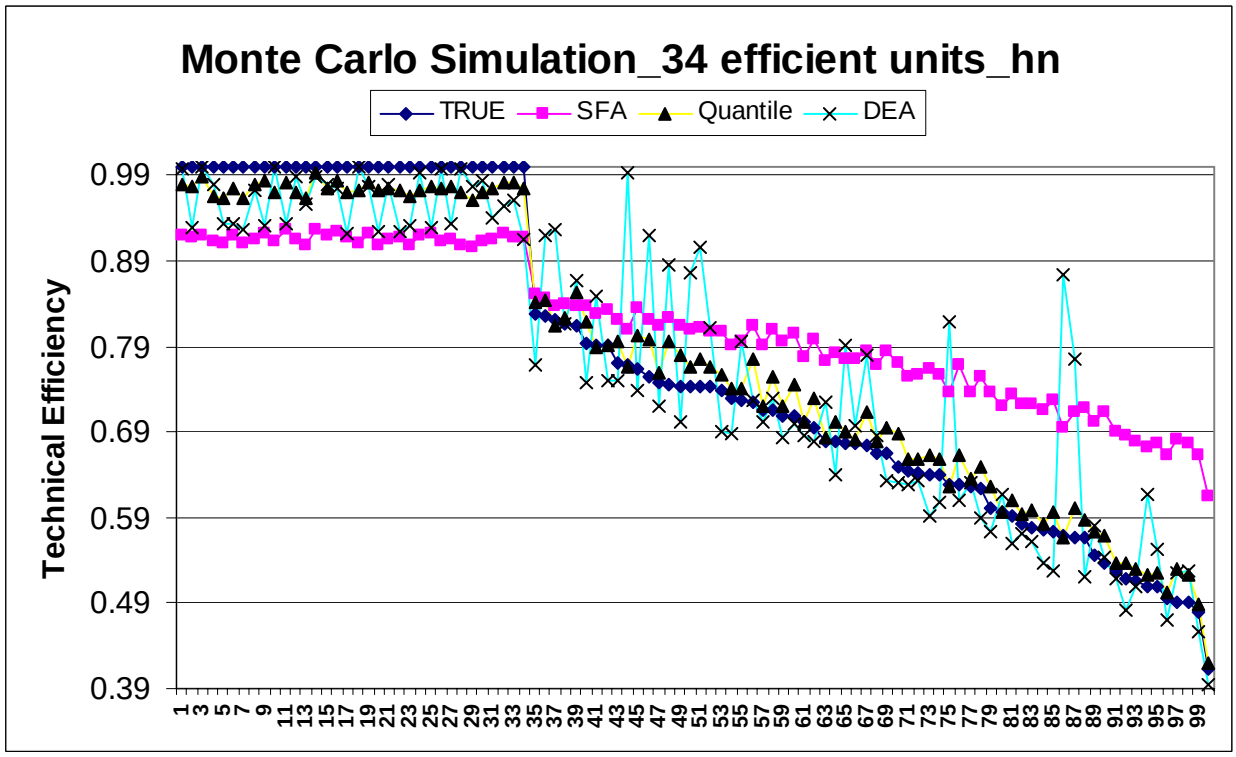

Figure A5

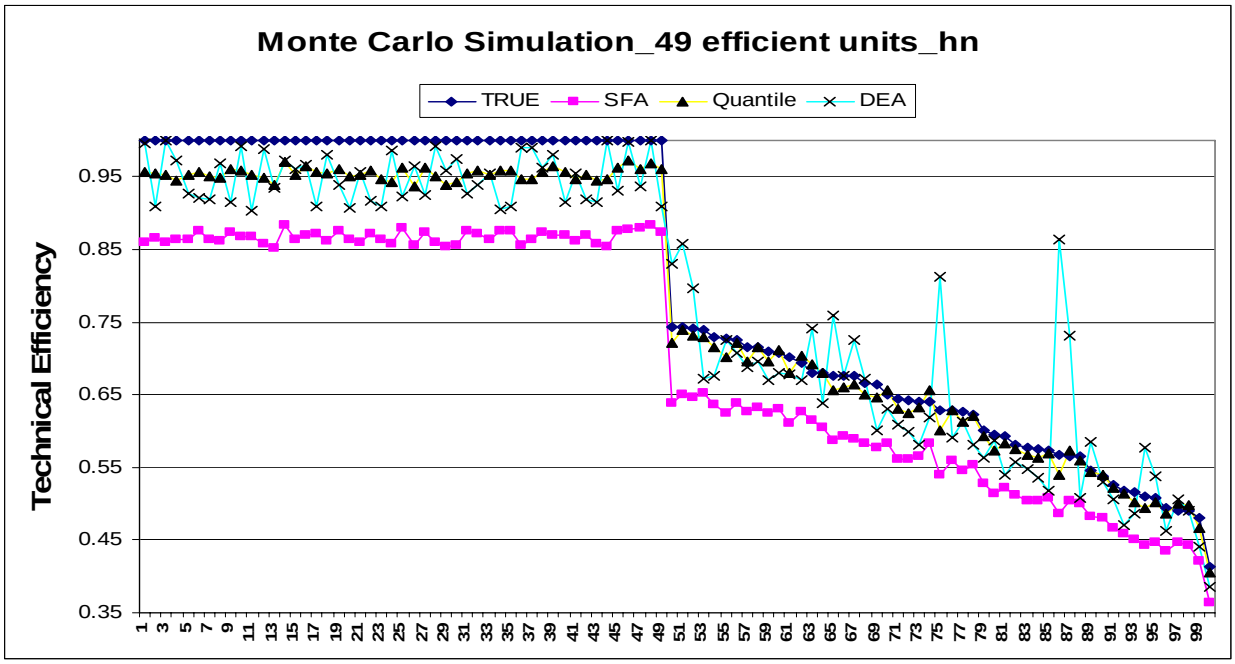

Figure A6

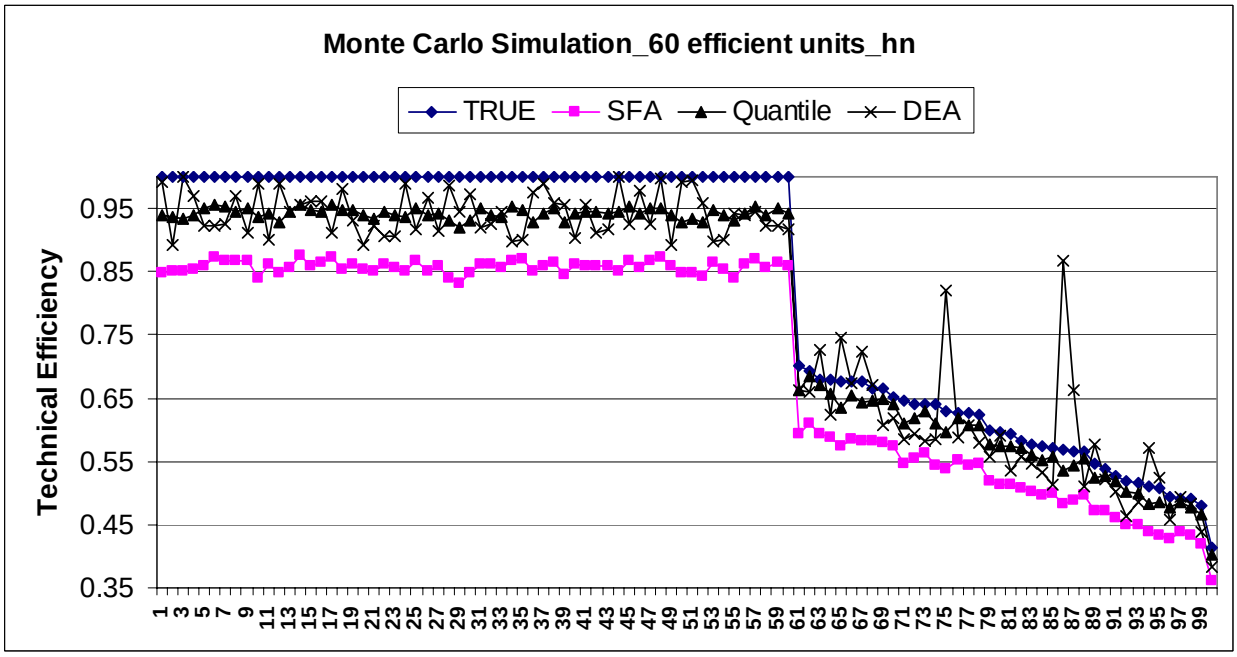

Figure A7

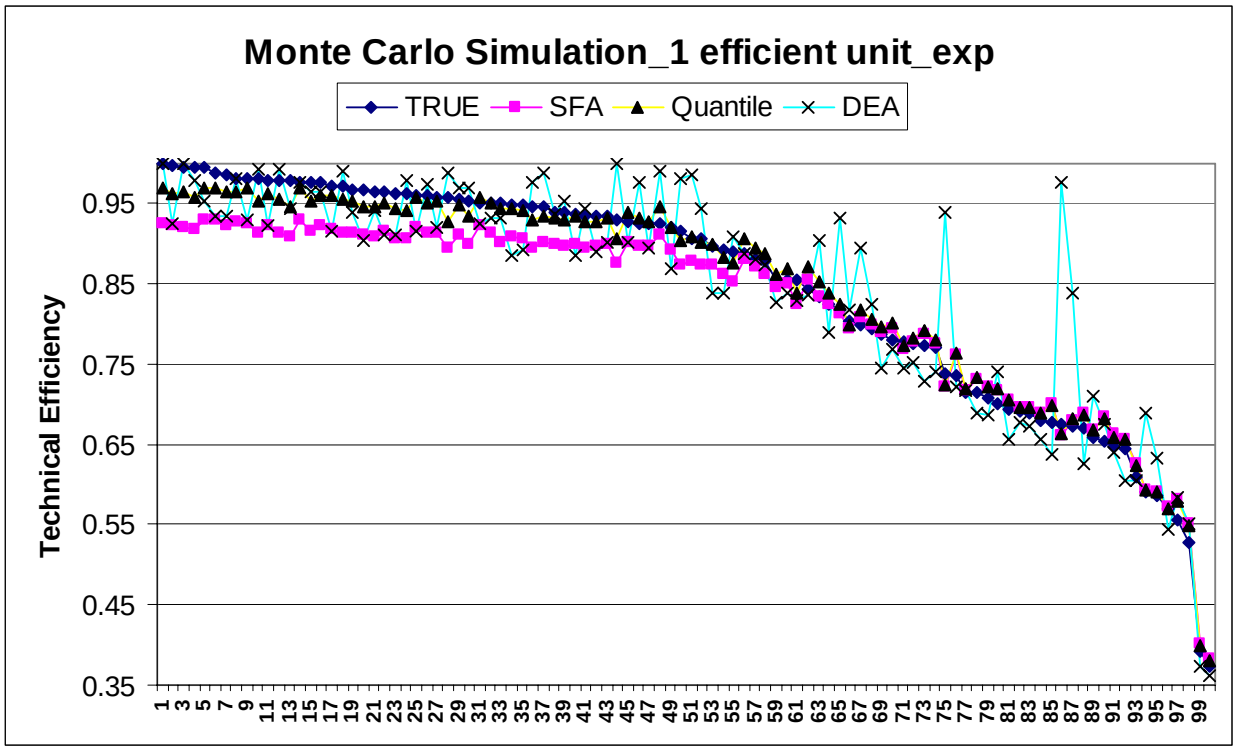

Figure A8

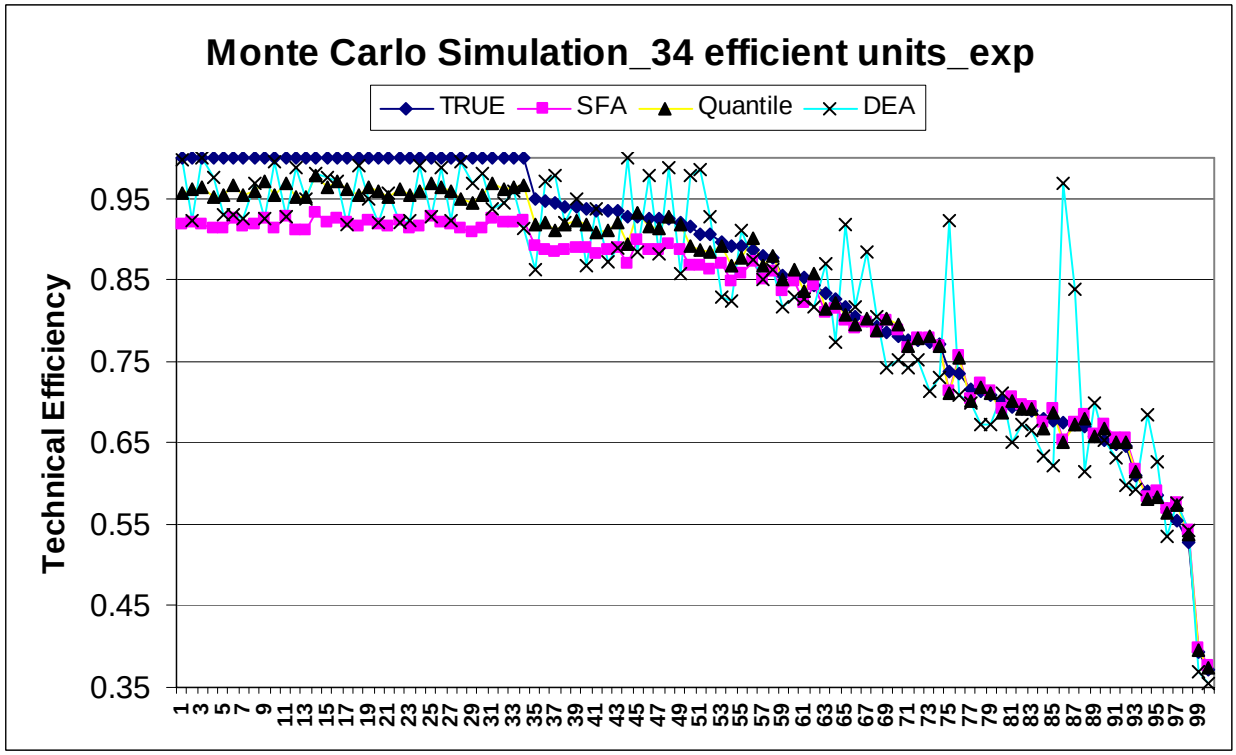

Figure A9

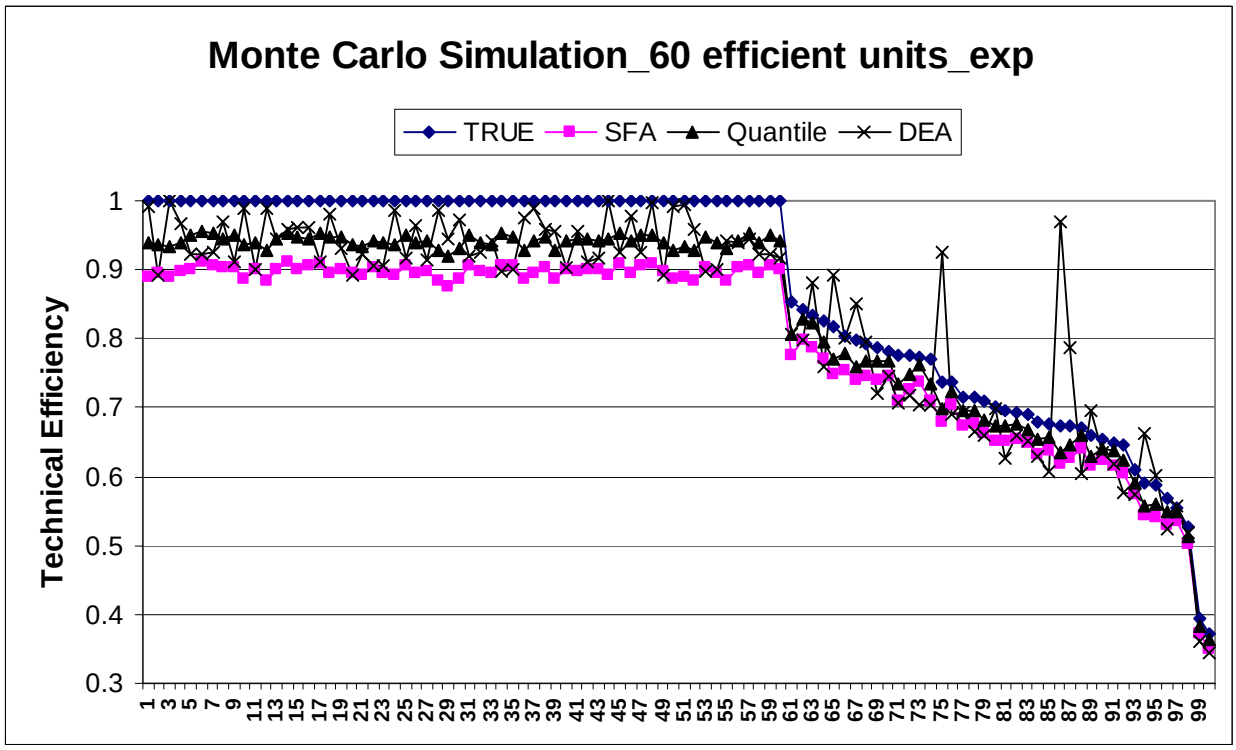

Figure A10

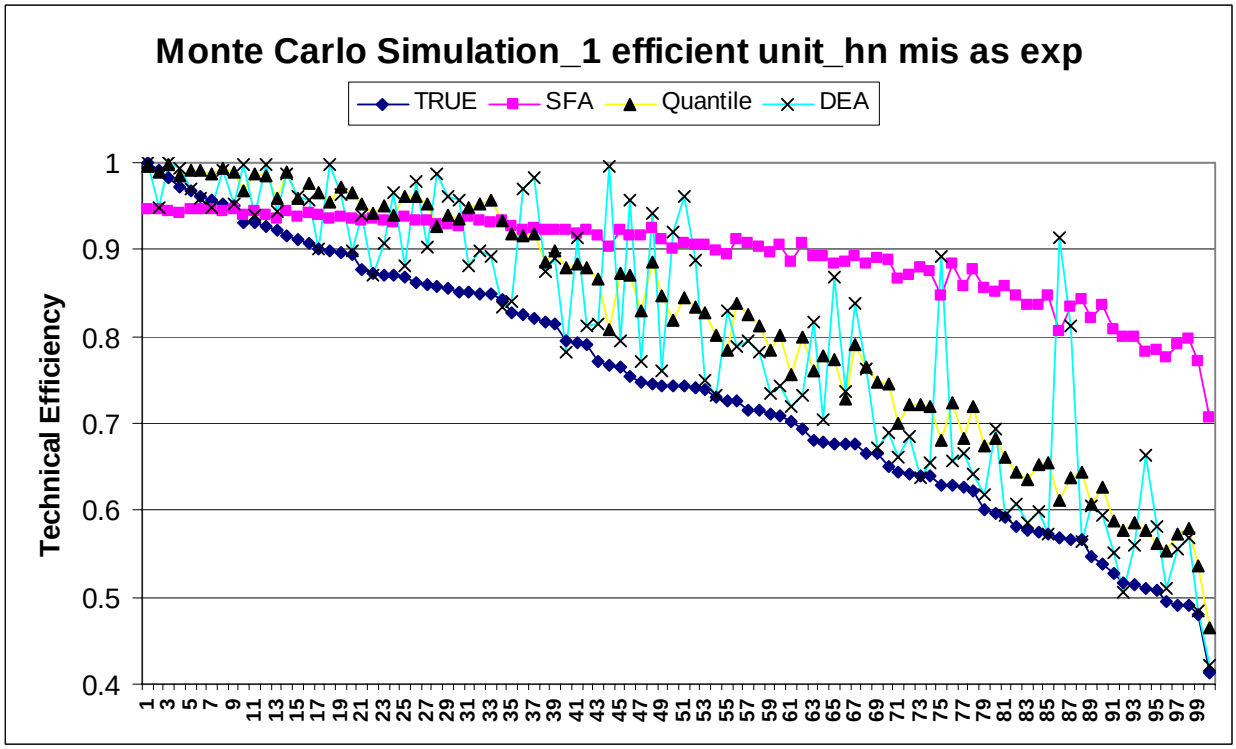

Figure A11

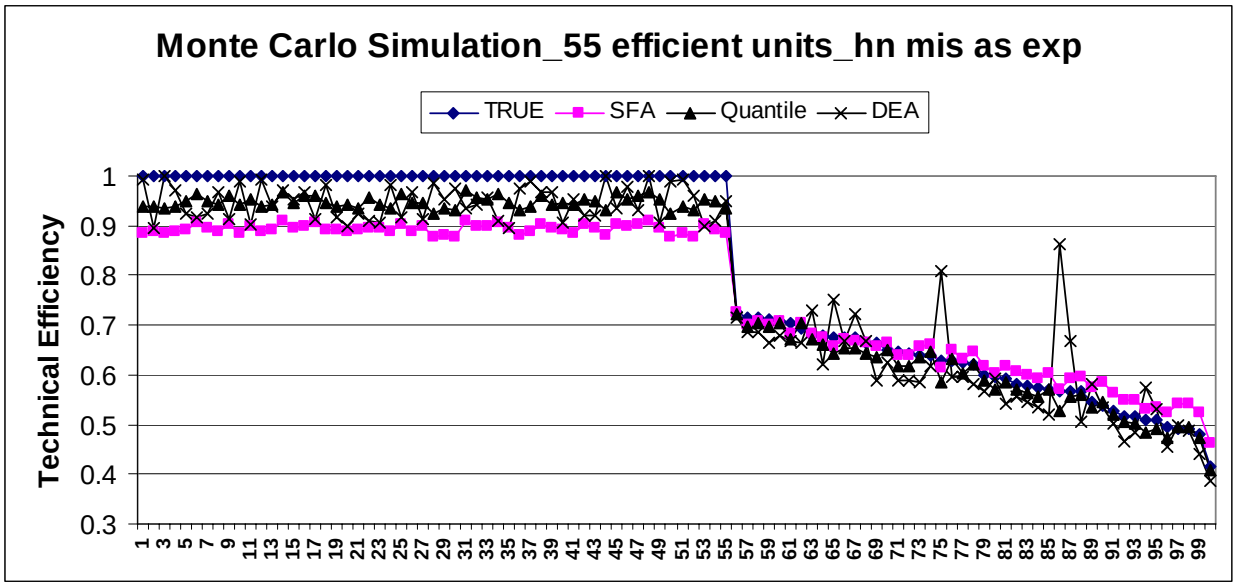

Figure A12