Problems of using grouped DMUs for efficiency measurement: Monte Carlo experiments, empirical dimension, and a correction procedure

Bernhard Brümmer[†] and Holger D. Thiele[‡]

Contact address: Bernhard Brümmer
Department of Agricultural Economics
CAU Kiel, 24098 Kiel, Germany

Phone +49 431 880 4449
Email bbruemmer@email.uni-kiel.de

[†] Bernhard Brümmer is research associate at the Institute of Agricultural Economics at the University of Kiel.
[‡] Dr. Holger D. Thiele is assistant professor at the Department of Food Economics and Consumption Studies at the University of Kiel.

Problems of using grouped DMUs for efficiency measurement: Monte Carlo experiments, empirical dimension, and a correction procedure

Abstract

This paper explores the consequences for parametric and non-parametric efficiency levels and rankings when using grouped instead of individual Decision Making Units (DMU). The bias results due to the differences of the grouped DMUs frontier compared to the individual DMUs frontier. Monte Carlo experimentation is used to evaluate the empirical dimension on the estimated efficiency levels and rankings. These results are illustrated with an empirical example using a sample of German farms. The bias in ranking is found to be substantial. Finally, a correction procedure is developed to improve the results when only grouped data are available.

Keywords: Grouped data, aggregation, DEA, SFA.

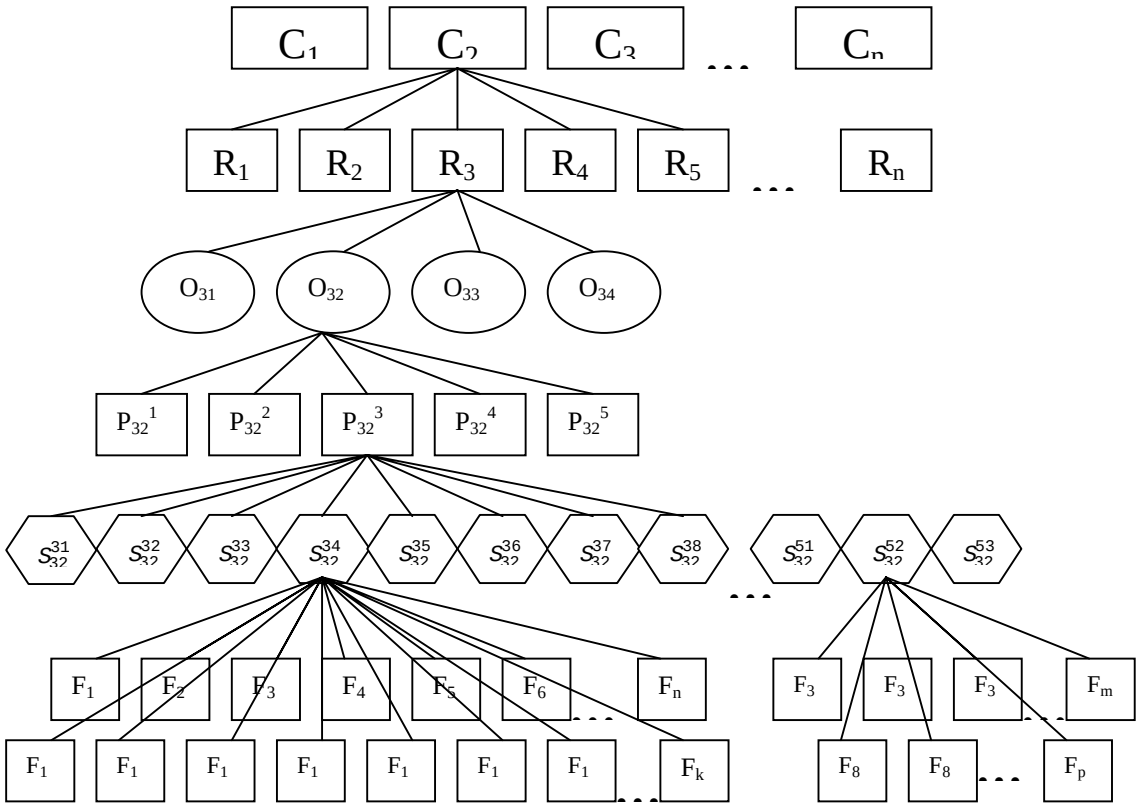
1 Introduction

The measurement of productive efficiency has become increasingly popular over the last decades. Since the introduction of the concept by Farrell (1957), the field of research has been extended from studies which analyze individual Decision Making Units (DMUs) to studies which focus on the relative performance of sectors of the economy, regions within a country, or whole countries. These latter literature often uses no longer observations on individual DMUs but data on groups of individuals (see Färe et al. (1995) for sectors, or Chambers et al. (1996), Färe et al. (1994), Fulginiti and Perrin (1997) for regions and states). Even studies focusing on intra-sectoral efficiency rely sometimes on grouped data because of limitations in the availability of data, for example, in the case of transitional economies (Thiele and Brodersen (1999), Sotnikov (1998)). However, it is clear that the use of data on groups instead of data on individuals will confound the results of a 'bounding' technique such as efficiency measurement much more than the standard averaging techniques (say, least squares or similar). Grouping of Decision Making Units (DMU) results in a bias of efficiency averages and efficiency rankings due to the differences of the frontier based on grouped DMUs compared to the frontier based on the individual DMUs.

A good empirical example in this context is the European Union Farm Accountancy Data Network (FADN), which collects individual data of farms every year from every member state (e.g. Germany: 9119 farms in 1996/97 (BML, 1998)). The official bureaus of statistics group the individual data. Figure 1 shows the aggregation structure of the grouping process. The individual farms ($F_1 \ldots F_n$) are aggregated into regions ($R_1 \ldots R_n$). The farms of each region are aggregated into ownership groups ($O_1 \ldots O_4$), e.g. individual farms, partnerships, etc. The farms of each ownership group are disaggregated into five groups of production types

(P_1…P_5), e.g. crop and livestock production, which are further disaggregated into different size groups (S_1…S_8). The individual DMU's are not available. Due to the fact that in each region one or more groups are empty, the observed number of groups of each region is far below the theoretical maximum of 160 farm groups (R=1, O=4, P=5, S=8). For example, the average number of farm groups for a state in Germany is 25 to 30 each year.

Figure 1. Aggregation structure of firms into grouped DMUs in the European Union, the case of the agriculture sector



Notes: C_1…C_n = Country 1 … n; R_1…R_n= Region 1 … n; O_1…O_4= Ownership type 1 … 4; P_1…P_5= Production type 1 … 5; S_1…S_8= Size 1 … 8; F_1…F_k, F_l, F_m, F_n = Firm 1 … Firm k,l, m, n.

Similar procedures and grouping problems also occur in other sectors of the economy. For conducting efficiency analysis based on grouped data it is necessary to explore the empirical dimension of the bias for the estimated efficiency scores and the relative ordering of the groups.

While the aggregation of inputs and outputs of an individual firm was addressed in many studies (see Diewert (1980), Lovell et al. (1987); Färe and Lovell (1987), Fox (1998)), only some studies focus on aggregation of the DMUs themselves. Cook et al. (1998) introduce the concept of hierarchical DEA in order to view efficiency at various levels of grouping. Banker and Morey (1986) have examined one form of grouping by introduction of categorical variables. Categorical variables allow for a comparison of any DMU with those in its own category and in those categories below it. In contrast to these previous studies we focus on the bias of estimated average efficiency, and on the bias of efficiency ranking based on grouped DMU.

Another issue in this context is that grouping as a specific form of aggregation reduces the number of observations in the sample. Previous studies have shown that the number of firms in an industry plays an important role for the determination of the average efficiency. Zhang and Bartels (1998) provide a detailed analysis for non-parametric efficiency measurement using Data Envelopment Analysis (DEA).The problem of biased average efficiency might be of little importance if the researcher is only interested in relative differences in efficiency, i.e. the ranking of the individuals or groups. However, further problems that might affect not only the average efficiency but also the ranking could arise from different sources: The inefficiency distribution might be heteroscedastic, the number of firms within each group might be different, the aggregation criteria might introduce systematic bias, and issues related to multidimensionality , i.e. the number of in- and outputs in the analysis, might confound the relative ordering of the groups.

To check for the above mentioned factors, we apply Monte Carlo experimentation and rank correlation coefficients to test the reliability with which the individual data as well as the aggregate data will yield similar efficiency rankings by group. In an empirical example, we

illustrate our results from the experiments with a sample of 669 Northern German farms. This dataset has the nice feature that the individual book keeping data provides the basis for the aggregated and grouped figures that are published regularly in official agricultural statistics. Subsequently, we present a simple correction procedure in the spirit of sampling theory, where the original underlying distribution is approximated by a re-sampling of the grouped data. Some conclusions and guidelines for further research follow in the last section of the paper.

2 Theoretical background

To introduce the basic problem in a more formal manner, we will focus on an output oriented efficiency measure under the assumption of a constant returns to scale technology with strong disposability of inputs that fulfills the usual regularity assumptions[1]. Furthermore, we restrict our attention to the case of only one output. This latter assumption seems reasonable since the use of grouped data for efficiency measurement usually requires that different outputs over firms, sectors or nations are transformed to a single monetary valued output. The radial output measure of technical efficiency for producer i (F_i^O) may then be defined as in Färe, Grosskopf, and Lovell (1994, p.63):

$$F_i^O(y_i, x_i) = \max \{\theta : \theta y_i \in P_i(x_i)\} \tag{1}$$

with y_i : output level of producer i,

x_i : input vector of producer i,

θ : arbitrary scalar,

$P(y,x)$: output feasibility set.

[1] For a detailed discussion of these assumptions, see e.g. Färe, Grosskopf and Lovell (1994) or Färe (1988).

The measure F_i^O gives the maximum (proportional) expansion of output given the input vector. Consider an industry J consisting of I producers. The measure $\bar{F}_j^O$ for the industry as a whole should then indicate by how much the total output level $\bar{y}_j = \sum_{i \in J} y_i$ could be raised without changing the actual input endowment $\bar{x}_j = \sum_{i \in J} x_i$. Denoting the frontier output level with y_j^*, this measure can be calculated from the individual efficiencies as the weighted average in Equation 2.

$$\bar{F}_j^O\left(\bar{y}_j, \bar{x}_j\right) = \frac{y_j^*}{\bar{y}_j} = \frac{\sum_{i \in J} y_i^*}{\bar{y}_j} = \sum_{i \in J} F_i^O \frac{y_i}{\bar{y}_j} \tag{2}$$

One possible estimator for the output efficiency measure is the familiar DEA estimator, introduced by Charnes et al. (1978) for the case of constant returns to scale. This estimator for a single DMU is given by the solution to the linear program in Equation 3.

$$\hat{\theta}_i = \max_{\theta, \lambda} \left\{ \theta : \theta y_i \leq Y\lambda, x_i \geq X\lambda, \lambda \geq 0 \right\} \tag{3}$$

For the calculation of an efficiency measure on the grouped level, one would ideally utilize the individual observations to estimate the individual efficiencies. Afterwards, the individual measures may be aggregated, either according to Equation 2 or by using the average estimated score per group. If only information on input-output decisions on the grouped level is available, the usual way to proceed is to estimate the efficiency score by solving the linear program on the grouped level.

$$\hat{\bar{\theta}}_j = \max_{\theta, \lambda} \left\{ \theta : \theta \bar{y}_j \leq \bar{Y}\lambda, \bar{x}_j \geq \bar{X}\lambda, \lambda \geq 0 \right\} \tag{4}$$

By construction, this estimator will be biased towards an efficiency score of one since the grouping procedure will move the estimated boundary of the output set further toward the

origin. Thus, the average level of efficiency within the industry will be overestimated. This effect is reinforced by the decrease in the number of observations as one of the natural consequences of grouping: The boundary hull will always fit better for less observations.

This itself poses only a minor problem if the further analysis would be based on the ranking of the groups, i.e. if we focus on relative efficiency comparisons rather than the absolute differences in the efficiency scores. However, as we shall illustrate in Figure 2, even the ranking might provide a misleading idea of the true distribution of efficiency scores across the groups.

Figure 2. Frontier based on 60 individual DMUs and their 6 groups

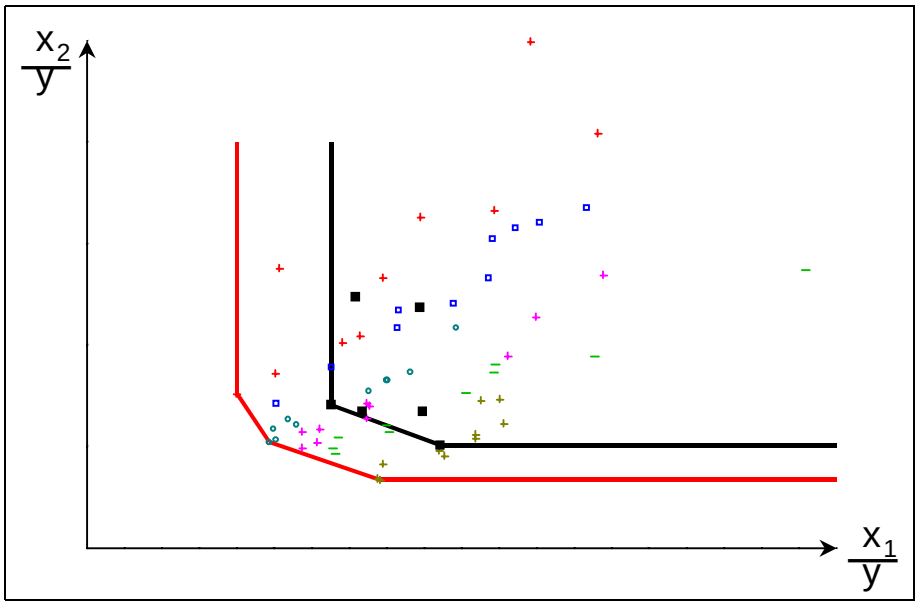


Figure 2 shows a graphical representation of a single-output, two-input CRS technology. The input levels per unit of output of 60 individual decision making units (DMUs) are depicted by the small symbols; the larger 6 symbols give the average input combination of 10 observations which have been grouped by factor intensity. DEA on the individual data would yield the isoquant closer to the origin, while DEA using the grouped data gives an isoquant shifted further away from the origin. We have a different, and probably biased estimate of the

isoquant that serves as a benchmark for the efficiency estimation. Therefore, the level of technical efficiency of the groups when estimated using the individual data is different from that based on the groups.

The estimated results for the six groups are shown in Table 1. The second and third column give the average TE score that has been estimated using individual and grouped data, respectively. An increase in the scores can be observed as expected. The next two columns give the rankings of these scores; even the switch to ordinal measurement does not give consistent results for the two approaches. Ideally, one would like that both approaches give an identical ranking of the groups. Unfortunately, this is not the case.

Table 1. Technical efficiency scores and ranks based on individual and grouped data

Group number	TE individual	TE grouped	Rank individual	Rank grouped	Change in rank
1	0.580	0.911	5	4	+1
2	0.513	0.734	6	6	± 0
3	0.744	1.000	2	1.5	+0.5
4	0.730	0.979	3	3	± 0
5	0.670	0.896	4	5	-1
6	0.752	1.000	1	1.5	-0.5

Another possible estimator for output oriented efficiency measure is the frontier production function approach. In the recent literature, Stochastic Frontier Analysis (SFA) is predominantly used for this branch of efficiency measurement. This requires an explicit parametrization of the production frontier, and distributional assumptions regarding both the systematic as well as the unsystematic error component. Frequently used functional forms include the Cobb-Douglas or the Translog, while the error terms are hypothesized most commonly as normal (unsystematic: random error) and half-normal (systematic: inefficiency). The basic model for the Cobb-Douglas case is then given by Equation (5).

$$\ln y_i = \beta_0 + \sum_j \beta_j \ln x_{ij} + (v_i - u_i) \quad (5)$$

where the $\beta\sigma$ are parameters to be estimated, v_i is a random error term, independently and identically distributed as $N(0,\sigma_v^2)$, and u_i is a non-negative error term, intended to capture output oriented technical inefficiency, which are assumed to be independently distributed as $|N(0,\sigma_u^2)|$. After estimation by some variant of least squares or Maximum Likelihood, a point estimate for technical efficiency is then obtained by evaluating the conditional expectation of $\exp(u_i)$ given the observed residual $(v_i - u_i)$.

The SFA estimator suffers in principle from similar problems as the DEA estimator when it comes to ranking obtained from grouped versus individual data. However, the problem will generally be less severe since SFA may be regarded as a less extreme bounding technique compared to DEA. As a result of the presence of the additional white noise error term, the position of the frontier will be less sensitive when using grouped instead of individual data. Therefore, the subsequent experiments focus on DEA; nevertheless, both DEA and SFA will be used in the empirical example.

One possibility to evaluate the degree of disarray that has been introduced by replacing the individual observations with their grouped counterparts is given by the nonparametric test statistics for association. We consider Spearman's ρ, also known as rank correlation coefficient. For the artificial data set in Figure 3.1-1, the value is 0.928, which still indicates a strong positive association between the two approaches. However, the ranking differs. The subsequent Monte Carlo experiments are used to evaluate the magnitude of this bias in ranking, and to explore its possible sources. We expect that the following factors will be influential for the degree of disarray which is introduced by the grouping procedure:

- Grouping criteria: Generally, we can expect that the grouping procedure will be less

problematic if homogeneous individual observations are aggregated into one group. The probability that this condition holds is likely to decrease with the degree of complexity of the procedure itself. For example, a multi-stage grouping procedure that consists of fairly different criteria (like the one outlined above) introduces more bias than a simple classification of observation according to output level.

- Inefficiency distribution: The bias in ranking is a consequence of the fact that the estimated frontier might be shifted in a non-parallel manner (see Figure 1). The reason for these non-neutral shifts when going from the individual to the grouped data lies in differences across groups regarding the distance of the individual observations to the group mean. The frontier based on individual data will be shifted only little for a very homogeneous group while the difference between the individual-based frontier and the group-based data will be large for a heterogeneous group. This means that a heteroscedastic inefficiency distribution with high cross-group differences in variance will likely generate more bias than a homoscedastic one.

- Variation in group sizes: Non-neutral shifts in the frontier are also more likely to occur when the group sizes vary, with a more heterogeneous number of observations per group leading to a higher bias in ranking. Since the grouping procedure takes away the peaks and troughs of the clusters of individual observations, and this smoothing effect increases with the number of firms within each group, heterogeneity of the group sizes might affect the bias in the estimation procedure. For the extreme case, i.e., at least two groups with only one observation, it could happen that a specific facet of the frontier would not even move at all.

- Loss in discriminatory power: The grouping procedure tends to higher efficiency since the number is lower. This increases the number of fully efficient observations, although

the individual data might give another picture. This effect is more problematic when the number of input variables is high, reflecting another variant of the ‘curse of multi-dimensionality.

Although the direction in which the above factors influence the relative ordering of the observations by their efficiency, the magnitude is difficult to estimate. To shed some more light on the empirical dimension of this problem, we conducted several Monte-Carlo experiments.

3 Monte-Carlo experiments

Data generating process

The experiments in this study assume a single-output, multiple input CRS technology. The data generating process (DGP) is specified according to the assumptions A1 to A4 in Kneip et al. (1996). The inputs are generated as uniformly distributed random numbers over the interval [5,15]. The individual efficiencies *u* are drawn from a normal distribution with mean μ and variance σ_u^2, truncated at below from zero, shifted to the right by one, to ensure that *u* > 1. The potential level of output is given by a Cobb-Douglas technology $y_i = \prod_k x_{ik}^{\beta_k}$ where *k* denotes the number of inputs in the specific experiment. The observed output is generated by multiplying y_i^* with $\frac{1}{u_i}$. Finally, groups are formed according to several different criteria (see below), and the linear program in Equation 4 is solved to estimate the group level output efficiency measure. Our measure of disarray is the bias for Spearman's ρ between the estimated group efficiencies as weighted mean of the individual efficiencies and the group efficiencies as direct estimates from the DEA using grouped data.

In this setting, we consider the following sources of disarray:

Grouping criteria: We have analyzed the impact of the following rules.

- Input-related: Groups are formed according to the level of input 1 or according to the ratio of input 1 to input 2.
- Size-related: Groups are formed according to the level of observed output, according to an input-based measure of firm size, and according to the level of frontier output.

Regardless of which criterion was used, we always introduced some systematic variation in efficiency between the groups by randomly shifting the efficiency level of each group. This was done to ensure that we are not looking for spurious efficiency differences.

Heteroscedastic inefficiency distribution: Variation in the efficiency scores per group is implemented by using a different variance parameter σ_u^2 per group. The expected value of a truncated normally distributed variable is effected by the choice of the variance parameter[2]. In our setting, this would imply that a direct relationship between the variance parameter and say the input endowment would also influence average efficiency. We chose to use a groupwise different variance parameter for the random draws of the true efficiencies; afterwards, a groupwise shift parameter is randomly applied to introduce groupwise different efficiency levels.

Heterogeneity of the group sizes: To evaluate this possible source of bias, we have utilized a random grouping procedure where the realized group sizes are uniformly distributed around the average; for different experiments the range around the average is varied.

Dimensionality: The dimensionality of the DEA problem, i.e. in our setting the number of inputs, might as well have an influence on the reliability of the rankings. The number of

[2] In the truncated model, the expected value is $E(u)=\mu + \sigma_u pdf(\mu/\sigma_u)/cdf(\mu/\sigma_u)$, where pdf and cdf are the standard normal density and distribution, respectively.

inputs is increased from two to four between the different experiments.

The main results of our Monte Carlo experiments[3] are summarized in the following tables. The base scenario is given by the following constellation:

- three inputs are considered,
- the individual observations are subsumed into 50 groups of fixed size,
- grouping criterion is the level of input 1,
- the true inefficiencies a generated with parameter σ_u equal to 0.75, and
- the number of replications of the experiment is set to 1000.

Table 2 shows the results for the variation of the number of groups (columns 2-4) and inputs (columns 5-7). Both variables seem to be unproblematic for our problem, since the rank correlation coefficients are generally well above 0.85. An exception is found for the experiment with a only 25 groups and 500 individuals (group size is 20). For both parts (variation of groups and of inputs), a common pattern can be observed, as the degree of disarray diminishes with the presumed size of the population, i.e., the rank correlation becomes larger in the lower rows of Table 2. For this setting, the number of firms in each group and the number of groups itself seem to increase the reliability of DEA using grouped data. For the each population size in the second part of Table 2, we observe a decrease in the rank correlation ρ when increasing the number of inputs. This is caused by the well known fact that an increasing number of dimensions makes more observations efficient. Since this effect is relatively more important in the case of fewer observations, we observe an decrease of the rank correlation with an increase in the number of inputs.

[3] All numerical work was done using Ox 2.10 (Doornik 1998) and lpsolve 3.0.

Table 2. Mean rank correlation: groups and inputs

	Number of groups			Number of inputs		
Population	25	50	100	2	3	4
500	0.8578	0.8973	0.9204	0.91553	0.89727	0.88632
1000	0.8662	0.9001	0.9181	0.91502	0.90013	0.89184
2000	0.8722	0.9063	0.9207	0.91728	0.90631	0.90233
4000	0.8809	0.9209	0.9318	0.92712	0.92086	0.91682

The influence of different grouping criteria is depicted in Table 3. We analyzed the impact of five different ways in which the grouping is done. Column 2 ('Input 1') shows the results for the base scenario, where the groups are generated by aggregating individual observations that use a similar amount of a specific input (we chose input 1). The rank correlation coefficients follow a similar pattern as outlined above. The next column ('Ratio') gives the results when the ratio of two inputs is used as a criterion. The general picture remains unchanged. However, we note a small decrease in the magnitude of the average rank correlation. Different measures for the firm size have been used as grouping criteria in the following three columns. First, column 4 ('Output')shows the results when the groups are build according to the level of observed output. In this case, we have found remarkably high coefficients, indicating that in our setting the use of the actual output seems to lead to unbiased estimates of the ranking.

Table 3. Mean rank correlation for different grouping criteria

	Criterion used for grouping			
Population	Input 1	Ratio	Output	Size
500	0.8873	0.8607	0.9587	0.8727
1000	0.8974	0.8661	0.9743	0.8580
2000	0.9092	0.8746	0.9848	0.8291
4000	0.9253	0.8867	0.9915	0.7872

This reassuring feature could not be confirmed for the next experiment. The figures in Column 5 indicate a substantial bias in the rank correlation coefficients. Here, we had used an

indicator of total firm size as an grouping criterion. In particular, we also observed that the influence of increasing the number of firms per group seems to be reversed. For the large population of 4000 individuals (which implies a group size of 80), the rank correlation coefficient is only 0.79.

The last single factor that we considered was the influence of an heteroscedastic distribution of the true inefficiencies. We implemented this as a groupwise different parameter σ_u for the truncated normal distribution. The results for the experiment are shown in Table 4, where the single columns give the figures for different average values of this parameter. A remarkable decrease in the rank correlation coefficient can be observed with increasing population size. The lowest mean value is 0.65. On the one hand, this still indicates a strong positive association between the rankings from the individual data and the grouped data, but on the other hand, a substantial amount of disarray, resulting in a different ranking, can be found.

Table 4. Mean rank correlation: Heteroscedastic inefficiency distribution

Population	Average std. dev. of individual efficiencies		
	0.5	0.75	1
500	0.8030	0.8203	0.8369
1000	0.7596	0.7641	0.7765
2000	0.7175	0.7062	0.7133
4000	0.6879	0.6605	0.6586

At last, Table 5 shows the results for a scenario in which we have combined groupwise heteroscedasticity with varying group sizes. The single columns are arranged according to the degree, with which we allowed the realized group size to deviate from the average group size.

For example, in column 2, we have allowed a variation of ± 20 % of the average group size[4]. This implies for the cell in the first row that the realized group sizes are uniformly distributed between 8 and 12. The last column allows for ± 90 %, e.g. the cell in the first row corresponds to group sizes between 1 and 19.

Table 5. Mean rank correlation: Combined scenario

	Variation of group sizes			
Population	± 20 %	± 50 %	± 70 %	± 90 %
500	0.8084	0.8061	0.8199	0.8333
1000	0.7277	0.7291	0.7427	0.7718
2000	0.6336	0.6486	0.6628	0.6826
4000	0.5622	0.5732	0.5884	0.6184

The results for this experiment are interesting in two aspects. First, the rank correlation in this setting is the lowest in magnitude over all the experiments that we carried out. For the large population size of 4000, the mean value of ρ decreases to only 0.56-0.62. Clearly, the bias in ranking is substantial and may lead the researcher to conclusions that may not be supported by the original ranking. Second, we observed that the mean rank correlation increases with increasing variation in the group sizes. This was very surprising since we could imagine that very heterogeneous group sizes would lead to a highly biased estimate of the frontier. However, the differences in the mean values along the rows are small, so this feature might not be so important.

[4] This average size equals the population size divided by the number of groups (the latter is here constantly equal to 50). For example, if the population size equals 500 (4000), we have 10 (80) firms in each group.

4 An empirical example: The case of German farming

Data

We will now illustrate the consequences that arise in practice with an real-world example. We utilize 1998/99 book-keeping data from 709 farms in one German region. The data come from the Farm Accountancy Data Network (FADN). Output and input data of these farms are regularly published in the official statistics in grouped form. Therefore, this data set is adequate for our study as we are able to utilize the 'first best' and 'second best' data set of the region. In the data cleaning process 40 observations were excluded from further analysis (21 part time farms, 7 farms specialized on gardening, 12 farms due to inconsistencies). A cross sectional data set of 669 farms remained in the sample.

We have grouped these 669 farms with the grouping procedure which is used by the official farm statistics in Germany. The first grouping criterion was the type of ownership (Single proprietorships, partnerships). The second criterion was the type of production, based on the revenue share of a production activity to total revenue (>50 % of total revenue from crop, or livestock, or pig/poultry, or mixed production otherwise). The third criterion was the size of the farms, base on the total revenue of each farm. Similar to the official statistics we create 8 size groups for crop and livestock farms (revenue < 50,000, 50-100,000, 100-150,000, 150-200,000, 200-250,000, 250-300,000, 300-350,000, > 350,000 DEM) and 3 size groups for pig/poultry and mixed farms (< 200,000, 200-300,000, > 300,000 DEM), and 1 size group for crop partnership farms, livestock partnership farms and pig/poultry partnership farms (>0 DEM). We end up with a total number of 25 groups. The number of farms in each group varies considerably. We have 2 groups with only 2, and 10 groups with less than or equal to 10 individual observations, while the largest group has 114 members. The average group size is 27 farms with a standard deviation of 28.

We have applied SFA and DEA with one output (total revenue) and four inputs (capital, labor, land, and intermediate inputs) to estimate parametric and non-parametric efficiency measures for individual farms and group of farms. The output variable, the total revenue, covers all production activities in DEM. Assuming a proportionality between service flows from capital and its stock, the variable for capital input is defined as the sum of total assets minus the land assets in DEM. Labor is measured as total farm working units (FWU). Land was measured as the sum of land allocated to agricultural land in hectares (HA). As a last input, the intermediate input was measured as the total intermediate consumption of each farm in DEM (see Table 6).

Table 6. Descriptive statistics of the sample (n=669)

	Total revenue	*Labor*	*Land*	*Capital*	*Intermediates*
Unit	*DEM*	*FWU*	*HA*	*DEM*	*DEM*
Mean	331,183	1.8	70.3	318,910	263,864
Std. Dev.	210,678	0.7	38.7	251,437	175,940
Min.	15,849	0.6	6.0	15,443	29,481
Max.	1,401,965	5.7	273.4	2,050,315	1,269,437

Notes: DEM = Deutsche Mark, FWU = farm working unit, HA = hectares,
Source: Data from FADN Germany, 1998/99, BMELF, Bonn.

Results

For the SFA analysis, three different production functions have been estimated. Two are based on the group data, while the third model utilizes all individual observations. This last model (SFA TL model) uses a translog specification form with four inputs. This functional form was chosen to allow for a high degree of flexibility. However, when we only use the group data as the base for the estimations, the large number of parameters renders the use of the translog nearly impossible since we would end up with only 8 degrees of freedom. Therefore, a Cobb-Douglas (CD) specification was chosen. With this functional form, the

elasticity for land turned out to be insignificant. We have dropped this variable from the CD models. The difference between the two CD models based on group data is that the first (SFA CD^1) is a standard SFA model (normal unsystematic error, half-normal systematic error), while the second (SFA CD^2) allows for a heteroscedastic distribution of the half-normal systematic error term. Heteroscedasticity is implemented as suggested by Caudill and Ford (1994). The distribution of the systematic error term is assumed to be dependent on the input endowment of the firm.

Table 7 shows the results for the comparison between the SFA technical efficiency results between SFA on individual data (n=669) and subsequent grouping into n=25 groups (SFA TL-model) versus SFA directly on grouped data of n=25 groups (SFA CD^1-model or SFA CD^2–model). The estimated average technical efficiency is quite similar across all models. Interestingly, the minimum technical efficiency is lower for the group data based models than for the individual data model.

The impact of grouping on DEA technical efficiency results is also depicted in Table 7. Column two in the lower part of the table shows the DEA technical efficiency based on 25 groups. Column three shows the results for the reference situation, based on the aggregated 669 individuals efficiency scores. While in the grouped data model the mean efficiency is very high (0.96), in the reference scenario based on individual data the technical efficiency is at a much lower level (0.62). According to Zhang and Bartels (1998), one could have expected a lower mean efficiency of the larger sample because of the effect of the increase in sample size. For example, while the DEA group model delivered 13 efficient observations, the DEA individual data model had no efficient observation after grouping. In the latter, the highest efficiency score was 0.87.

Table 7. Average technical efficiency scores of the SFA and DEA models

	Group data SFA CD[1] - Model	Group data SFA CD[2] - Model	Individual data SFA TL-Model
	SFA technical efficiency scores		
Mean	0.82	0.88	0.82
Stddev	0.14	0.15	0.08
Min	0.42	0.42	0.59
Max	0.98	0.98	0.92
	Group data DEA[3] model		Individual data DEA model
	DEA technical efficiency scores		
Mean	0.96		0.62
Stddev	0.06		0.13
Min	0.78		0.45
Max	1.00		0.87

[1] homoscedastic, [2] heteroscedastic, [2] outputoriented variable return to scale

Figure 3 in section 5 shows a graphical representation of the SFA results on the rank correlation analysis (ranks without correction). On the x-axis, we have the group ranks estimated by SFA with individual data and subsequent grouping, while the y-axis contains the ranks for each group as estimated by SFA directly using group means. The rank correlation coefficient is also shown. While the numerically value of 0.78 for Spearman's ρ indicates modest correlation, the graph show several observations – especially in the upper groups – for which the ranking differs not very much or is equal between the two methods (SFA on individual data and subsequent grouping versus SFA directly on grouped data). However, on average the groups are classified quite different.

The graphical representation of the DEA results of the rank correlation analysis is shown in Figure 4 in section 5 (ranks without correction). The mean rank correlation coefficient was 0.41 in this case. The graph shows large substantial deviations from the reference line. We could only find two identical ranks between DEA on individual data and subsequent grouping versus DEA directly on grouped data. The similarity of the ranking seems to be better for the numerically high ranks, i.e. for the inefficient farms, than for the smaller ranks, the more efficient groups. Another issue which was already discussed above can be clearly seen in the figure. When we estimate the group efficiencies directly, we have 13 groups which are

classified as efficient (VRS TE=1.0), and which therefore share the top rank. When the group efficiencies are estimated based on DEA with individual data, the subsequent aggregation of the estimation procedure leads to similar, but numerically different group efficiencies. This also implies different ranks. To summarize, this example indicates a relatively strong bias in ranking for grouped data compared to individual data, which is much stronger if DEA instead of SFA is applied.

5 A correction procedure for group data efficiency

The simulations and the real-world example have shown that efficiency estimates based on grouped data can lead to quite different results compared to an analysis that uses the full information contained in the individual data. Even if the individual data are unavailable for whatsoever reason, the researcher might nevertheless have access to some additional information on the underlying data. If this additional information can be used to approximate the underlying data generating process in a better way than it achieved by just using the grouped data, the situation should improve. For this case, we propose a simple correction procedure in the spirit of repeated sampling that takes into account the heterogeneity regarding group sizes and standard deviations across groups.

Assume that not only the group means (m_j for all J groups) of the variables of interest are known to the researcher but as well the standard deviation (s_j) for each group together with the number of elements per group, N_j. The former piece of information is often readily available, although not regularly published in official statistics. Then the analysis could improved by the repetitive use of pseudo data that is generated with the goal to mimic the underlying individual data. In detail, a naive resampling procedure consists of the following steps.

1. For each of the input and output variables, draw a pseudo sample of size N_j data points for each group assuming a normal distribution $N(m_j, s^2_j)$ for all J groups. This leads to a pseudo data set of the same size as the underlying individual data. If information on the whole variance-covariance matrix for each group is available, this additional information should be considered as well
2. Estimate the technical efficiency scores using the full pseudo data set of $N=\sum_{j\in J} N_j$ observations.
3. Calculate the (output) weighted mean of technical efficiency for each group.
4. Repeat step 1) to 3) B times, where B is a sufficient large number.

We have applied this procedure to the real-world example using SFA and DEA models. Given the advice found in the bootstrapping literature, a number of 1000 iterations should be sufficient. For the examples below, we have used B=1000, although some preliminary experiments with B as low as 200 gave very similar results. For both models, the rank correlation between the average resampled group efficiency and the individual ('true') efficiency is well above 0.8. For SFA, the rank correlation coefficients increases from 0.78 to 0.86 (see Figure 3). In comparison to the previous rank correlation analysis (see Figure 3) the application of the correction procedure increases the number of identical ranks between SFA on individual data and subsequent grouping versus SFA directly on corrected grouped data from 1 to 4.

Figure 4 shows the application of the correction procedure in the case of the DEA (ranks of corrected groups). The DEA approach improves even more: The rank correlation increases from 0.41 to 0.84. Now the graph shows relatively small deviations from the reference line. In contrast to the ranks of uncorrected groups where we could only find two identical ranks between DEA on individual data and subsequent grouping versus DEA directly on grouped data, here we found 7 identical ranks.

Figure 3. Cross plot of group ranks of uncorrected and corrected SFA technical efficiency

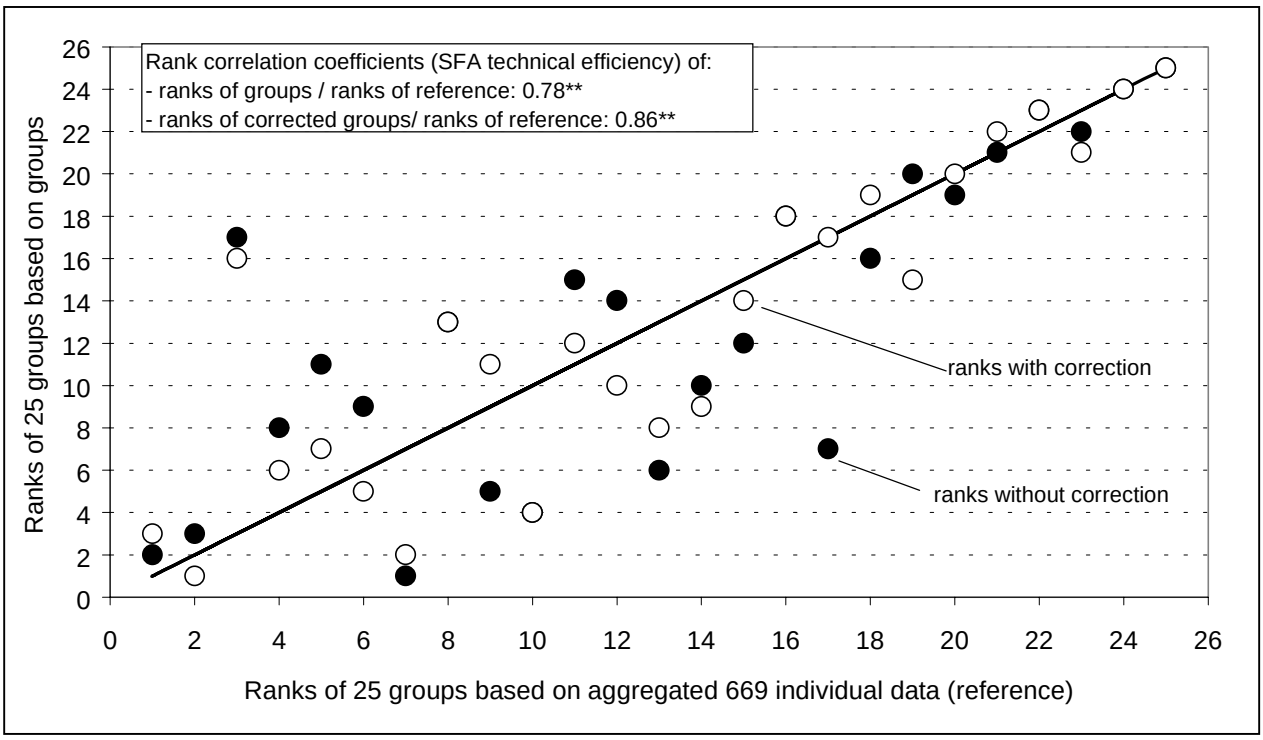
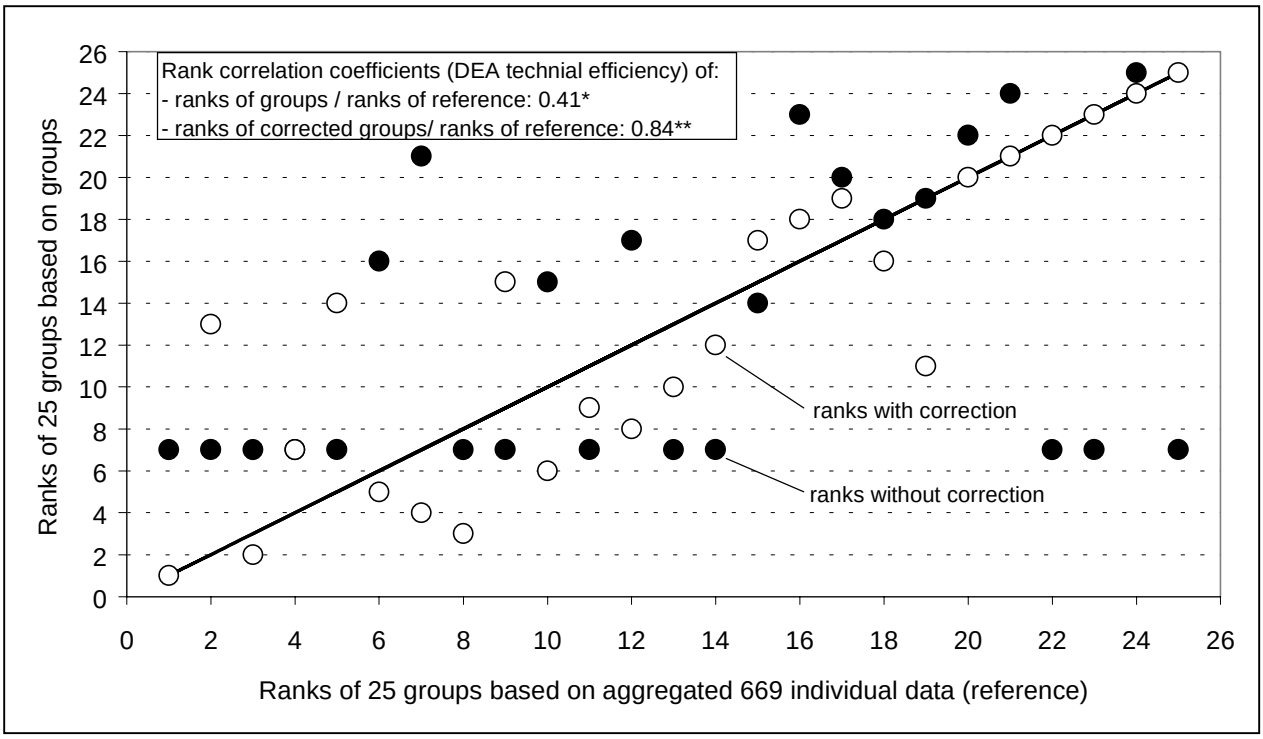


Figure 4. Cross plot of group ranks of uncorrected and corrected DEA technical efficiency



To summarize, the application of this simple correction procedure indicates that efficiency estimates based on grouped data might give relatively similar results to analysis that uses the

full information contained in the individual data.

6 Conclusions

This study has focused on the empirical dimension and on possible corrections for problems caused by using grouped DMUs for efficiency estimation instead of the underlying individual data. First, we have used Monte Carlo experimentation to compare the bias that might arise in the ranking of groups when the efficiency estimation is based on grouped data. The main results are that among the different grouping criteria, 'total firm size' shows the largest bias. Heteroscedasticity in the efficiency distribution alters the ranking considerably.

A real world example – 669 farms of the northern German agriculture sector in 1997/98 – has shown that efficiency estimates based on group data can lead to quite different results compared to an analysis that uses the individual data without any information loss. Mean SFA technical efficiency based on group data was 0.88 and based on individual data 0.82. A significant rank correlation coefficient between the two methods of 0.78 was found. Mean DEA technical efficiency based on group data was 0.96 and based on individual data 0.62. In this case the mean rank correlation coefficient was 0.41. The empirical example indicates that SFA and DEA on the basis of grouped data lead to a substantial bias. Therefore, the interpretation of studies based on grouped data should handled with more care.

In the last section we provide a simple correction procedure if one only has grouped data available but has knowledge on some additional statistics of the underlying distribution of individual observations within in the groups. The procedure essentially augments the grouped data by taking into account the heterogeneity regarding group sizes and standard deviations across groups. The results are promising: the SFA rank corrrelation increases from 0.78 to 0.86. The DEA approach improves even more: The correlation increases from 0.41 to 0.84.

References

Banker, R. and R. Morey (1986). The use of categorical variables in Data Envelopment Analysis. Management Science 32 (12), 1613-1627.

Chambers, R. G., R. Färe, and S. Grosskopf (1996). Productivity growth in APEC countries. *Pacific Economic Review* 1 (3), 181-190.

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

Cook, W. D., D. Chai, J. Doyle, and R. Green (1998). Hierarchies and groups in DEA. The *Journal of Productivity Analysis* 10 (2), 177-198.

Diewert, W. E. (1980). Aggregation problems in the measurement of capital. In D. Usher (ed.), *The Measurement of Capital*. Chicago: The University of Chicago Press for the National Bureau of Economic Research.

Doornik, J. A. (1998). Object-Oriented Matrix Programming using Ox 2.0. Technical report, Oxford: http://www.nuff.ox.ac.uk/Users/Doornik.

Färe, R. (1988). *Fundamentals of Production Theory*. Berlin: Springer Verlag.

Färe, R., S. Grosskopf, and W.-F. Lee (1995). Productivity in Taiwanese manufacturing industries. *Applied Economics* 27 (3), 259-265.

Färe, R., S. Grosskopf, and C. A. K. Lovell (1994). *Production Frontiers*. New York: Cambridge University Press.

Färe, R., S. Grosskopf, M. Norris, and Z. Zhang (1994). Productivity growth, technical progress, and efficiency change in industrialized countries. *American Economic Review* 84 (1), 66-83.

Färe, R. and C. A. K. Lovell (1987). Aggregation and efficiency. In W. Eichhorn (ed.), *Measurement in Economics: Theory and Applications of Economic Indices*. Heidelberg: Physica-Verlag, 639-647.

Farrell, M. J. (1957). The Measurement of Productive Efficiency. *Journal of the Royal Statistic Society A CXX*, 253-290.

Fox, K. J. (1999). Efficiency at different levels of aggregation: Public vs. private sector firms. *Economic Letters* 65, 173-176.

Fulginiti, L. E. and R. K. Perrin (1997). LDC agriculture: Non-parametric Malmquist productivity indexes. *Journal of Development Economics* 53 (2), 373-390.

Kneip, A., B. U. Park, and L. Simar (1996). A note on the convergence of nonparametric DEA efficiency estimates. No. 9603 in CORE Discussion Papers. Louvain: Université Catholique de Louvain.

Lovell, C. A. K., A. Sarkar, and R. Sickles (1987). Testing for aggregation bias in efficiency measurement. In W. Eichhorn (ed.), *Measurement in Economics: Theory and Applications of Economic Indices*. Heidelberg: Physica-Verlag, 187-206.

Thiele, H. and C. M. Brodersen (1999). Differences in farm efficiency in market and transition economies: Empirical evidence from West and East Germany. *European Review of Agricultural Economics* 26 (3), 331-347.

Zhang, Y. and R. Bartels (1998). The effect of sample size on the mean efficiency in DEA with an application to electricity distribution in Australia, Sweden and New Zealand. *The Journal of Productivity Analysis* 9 (3), 187-204.