January 2010

WORKING PAPER SERIES

2010-ECO-01

Comparing Efficiency Across Markets: An Extension and Critique of the Zhang and Bartels (1998) Methodology

Ruben Chumpitaz
IESEG School of Management, LEM-CNRS (UMR 8179)

Kristiaan Kerstens
CNRS-LEM (UMR 8179), IESEG School of Management

Nicholas Paparoidamis
IESEG School of Management, LEM-CNRS (UMR 8179)

Matthias Staat
University of Mannheim, Germany

IESEG School of Management
Catholic University of Lille
3, rue de la Digue
F-59000 Lille
www.ieseg.fr
Tel: 33(0)3 20 54 58 92
Fax: 33(0)3 20 57 48 55

Comparing Efficiency Across Markets: An Extension and Critique of the Zhang and Bartels (1998) Methodology†

Ruben Chumpitaz*

Kristiaan Kerstens*

Nicholas Paparoidamis*

Matthias Staat**

Abstract:

The use of non-parametric frontier methods for the evaluation of product market efficiency in heterogeneous markets seems to have gained some popularity recently. However, the statistical properties of these frontier estimators have been largely ignored. The main point is that non-parametric frontier estimators are biased and that the degree of bias depends on specific sample properties, most importantly sample size and number of dimensions of the model. To investigate the effect of this bias on comparing market efficiency, this contribution estimates the efficiency for several datasets for two main product categories. Following Zhang and Bartels (1998), these results comprise re-estimates for the larger samples limiting their size to that of the smaller samples when the model dimensions for different samples are identical. Furthermore, sample sizes are adjusted to mitigate the eventual differences in dimensions in specification. This allows comparing market efficiency for different markets on a more equal footing, since it reduces the bias effect to a minimum making the comparison of market efficiency possible. However, the article also points out the critical limitations of this Zhang and Bartels (1998) approach in certain respects. Apart from reporting these negative results, we also offer some suggestions for future work.

Keywords: Market Efficiency, Heterogeneous Product Markets, Bias, Monte-Carlo Simulation

* CNRS-LEM (UMR 8179), IESEG School of Management, 3 rue de la Digue, F-59000 Lille, France, Tel: +33 320545892, Fax: +33 320574855. Correspondence to K. Kerstens: k.kerstens@ieseg.fr.

** University of Mannheim, Department of Economics, D-68131 Mannheim, Germany, Tel. +49 6211811894, Fax. +49 6211811893, staat@uni-mannheim.de.

July 2007 / Third Revision: November 2009

† We thank three referees for their suggestions that greatly improved the quality of the paper.

1. Introduction

Recently, based upon the theory of hedonic price functions, a number of studies assessing the efficiency of heterogeneous product markets using non-parametric frontier estimators (Data Envelopment Analysis (DEA)) have appeared (see, e.g., Staat and Hammerschmidt (2005) for a review).[1] Indeed, the advantage of being able to evaluate differentiated products and their prices has made DEA a standard tool for the evaluation of market efficiency in the marketing, management and economics literatures alike. As the exchange between Hjorth-Andersen (1992), Maynes (1992) and Ratchford and Gupta (1992) reveals, alternative approaches like measures of price dispersions or price quality relations are not informative as to the degree of market efficiency. This strand of the literature is somewhat akin to the use of frontier-based inefficiency estimators in labour economics to estimate "hedonic" wage frontiers to establish deviations resulting from imperfect information (one example is the matching efficiency of regional labour markets: see, e.g., Ibourk et al. (2004)). Polachek and Robst (1998) is –to our knowledge- the sole study corroborating the incomplete information interpretation of these wage inefficiency estimates by comparing these to independent direct measures of workers' knowledge of the world of work. Thus, inefficiency estimates of heterogeneous product and labour markets can be attributed at least in part to imperfect information among consumers and employees.

However, the advantages of this methodology come at a cost that has hitherto been largely ignored by current practice price frontier applications. The understanding of this specific problem has been facilitated by recent insights into the statistical properties of these frontier estimators (see Daraio and Simar (2007, chap. 3), Simar and Wilson (2000) for a survey and especially Gijbels et al. (1999)). Namely, (price) frontier estimators are inherently biased and this bias depends on specific properties of the underlying data. The bias is not only related to the number of observations in the sample and to the number of inputs and outputs in the model, but also to the density of observations around the relevant segment of the frontier. The reason why efficiency scores obtained from samples with different properties cannot be directly compared is that non-parametric frontier estimators provide a local and inner approximation of the true, but unknown frontier (technology). Roughly speaking, the more observations there are in a sample, the better the approximation of the true frontier. The better this approximation is, the closer the efficiency estimates resemble the true efficiency. Put differently, with a poor approximation of the frontier there is possibly a substantial bias for the efficiency estimates. Obviously, different samples with specific properties in terms of sample size lead to different qualities of

[1] Rosen (1974) offers a theoretical framework to study market equilibria for differentiated commodities differing along multiple characteristics.

approximations and hence different degrees of bias. Similar to the sample size bias, the more input and output dimensions are included in a given technology, the more serious the bias problem becomes.

This makes the comparison of average product efficiency interpreted as "market efficiency" across markets difficult when the samples for the markets studied differ in size and when products are evaluated on the basis of different numbers of characteristics.[2] If so, one cannot infer from the average efficiency scores that one market is more efficient than another (or alternatively phrased in terms of inefficiency, that higher mark-ups on differentiated products and hence lower consumer surplus are realized in one market vis-à-vis another), let alone employ statistical methods to analyse the determinants of market efficiency. This, however, is precisely what has been attempted in some of the existing studies on market efficiency (e.g., Kamakura et al. (1988)).

However, this problem need not distract from the attractiveness of measuring and comparing market efficiency with frontier based approaches provided one can properly account for this above bias. It had been noted by Gstach (1995) as well as by Zhang and Bartels (1998) some time ago, that comparing results across samples in a naïve way is clearly problematic. Zhang and Bartels (1998) demonstrated their case using three different samples of electricity utilities and showed a pragmatic way to arrive at results that can be readily compared.[3]

Our focus on the Zhang and Bartels (1998) study is justified by two main arguments. First, this article is often credited as being among the first demonstrating the impact of sample size on average efficiency (see, e.g., Balcombe, Fraser and Kim (2006) for a study comparing average efficiencies across frontier methodologies). Second, and more importantly, this article has been a source of inspiration to some articles trying to circumvent this problem of comparing average efficiency across samples (see Dexter et al. (2008) and De Witte and Marques (2010), among others).

Some often cited rules of thumb in the frontier literature maintain that certain relations between the number of observations and the number of variables should be observed. For instance, Vassiloglu and Giokas (1990: p. 593) suggest that the sample should have at least twice as many observations as there are variables in the model. In the same vein, Dyson et al.

[2] In the production frontier literature, the term "structural efficiency" has alternatively been employed to denote efficiency measures for the performance of a group of production units (i.e., an industry). Furthermore, though not explicitly phrased in terms of consumer welfare, it is clear that the measure of market efficiency developed in this contribution is somehow related to applied welfare notions (e.g., consumer surplus) employed in the economics and marketing literature on consumption behaviour.

[3] As a matter of fact, Zhang and Bartels (1998) point out that a similar problem exists when comparing average technical efficiency scores across samples obtained from stochastic parametric frontiers. An early attempt to construct confidence intervals for firm and time means of non-parametrically and econometrically estimated efficiency scores in small samples is found in Atkinson and Wilson (1995).

(2001) maintain that the number of observations should be at least twice the product of the number of inputs and the number of outputs. These authors maintain that observing these rules when specifying a model should lead to well-differentiated results. These rules point to the fact that for low numbers of observations in relation to the number of inputs and outputs the approximation of technology may become too poor to reveal anything about the efficiency of the observations. Even when researchers follow these rules and thus obtain well-differentiated results for a single market, this would not resolve the problem of comparisons across markets. Therefore, our paper insists on the necessity to compare product efficiency across different markets on an equal footing.

Notice that the problem addressed here is far more general to the use of non-parametric frontier estimators than it might appear at first, since there are a number of other instances where results obtained from samples of different sizes are compared. Two obvious cases that come to mind are (i) surveys of published studies pertaining to the same industry[4] (ii) studies based on comparing efficiency estimates between unbalanced panels where the sample size changes over time.[5]

In the present study, based on some efficiency estimates for markets for computer hardware, we illustrate how a naïve application of DEA to the problem of comparing market efficiency across markets fails to generate sound conclusions. Following Zhang and Bartels (1998) we re-estimate the results for our larger samples limiting their size to the number of observations found in the smaller of the available samples. In the spirit of the Zhang and Bartels (1998) method, we suggest adjusting sample sizes to mitigate the eventual differences in dimensions included in the specification. These strategies should ideally reduce the bias effect to a minimum and allow for a comparison of market efficiency across markets without confounding effects.

However, both a priori arguments and the empirical analysis show that the Zhang and Bartels (1998) method and its extension do not offer a general solution for the problem at hand. The systematic application of this methodological correction in this contribution and especially

[4] To mention but a few studies related to the first case, neither Hollingsworth et al. (1999) and Hollingsworth (2003) on health care service providers, nor Athanassopoulos (2004), Berger et al. (1999), Berger and Humphrey (1997), or Paradi et al. (2004) on efficiency studies related to bank and bank branches even mention this bias issue. For example, Hollingsworth et al. (1999: p. 165) compare average efficiencies of hospitals with different ownership type stating that: "… public sector hospitals have the highest mean efficiency (0.96) and the highest median (0.96), compared with not–for–profit hospitals which have a lower mean efficiency (0.80) and a lower median (0.84)." without mentioning any sample properties.

[5] An example of the second case is the use of a sequential technology to compute efficiency using all cumulative data observed in the periods up to the period being considered, which allows measuring technical progress but precludes observing any technical regress. While most applications (e.g., Shestalova (2003)) ignore this problem altogether, already Färe et al. (1989: page 665) noted that "… one may wish to ensure that the reference sets … contain the same number of observations."

the indication of its critical limits should therefore pave the way to a more systematic discussion on how to compare market efficiencies across different product categories, and in general on how to compare efficiency scores across different samples and/or different specifications. These limitations of the Zhang and Bartels (1998) article were not noticed before in the literature.

This study is organised as follows. The next section gives a brief survey of the literature and discusses in some detail the problems that may arise due to the bias of the estimators used. Next, we provide a description of the non-parametric frontier estimation methodology. This section also elaborates on the need for the Zhang and Bartels (1998) approach or an alternative methodology in general and in market efficiency studies in particular. The following section contains a description of the data used. Thereafter, we present the results obtained. A final section concludes.

2. Product Market Efficiency

2.1 A Succinct Review of the Literature on Market Efficiency

Efficiency of choice in the marketing literature has been measured in a variety of ways. Past studies exploring efficiency of consumer choice tend to define consumer inefficiency based on price-quality correlations (e.g., Morris and Bronson (1969)), price dispersions (e.g., Maynes and Assum (1982)) and a concept similar to Lancaster's (1966) efficiency frontier (e.g., Kamakura et al. (1988)). In addition, analyzing price dispersion has become increasingly popular in economics (see the Blinder et al. (1998) survey). While the early literature was mainly interested in macroeconomic implications in terms of business cycles and unemployment (e.g., Carlton (1989)), recent contributions also focus on consequences related to firm strategies, industrial organization, etc. (e.g., Warner and Barsky (1995)).

Briefly assessing the main methodologies employed in marketing, measuring price dispersion is useful for (fairly) homogeneous goods and services only. Otherwise, the eventual differences in quality characteristics must be accounted for. Furthermore, these studies cannot provide any indication as to the degree of informational imperfection in the market. Research on price-quality correlations has often used quality rankings of Consumer Reports to investigate the relationship between market prices and objective quality (see, e.g. Bodell et al. (1986), Faulds et al. (1995)). Most studies on price-quality correlations found a positive but weak correlation, and at times even a significantly negative correlation leading researchers to conclude that substantial inefficiencies prevail in many markets (see Ratchford and Gupta (1990), as well as Hjorth-Andersen (1992)). However, there is no reason to believe that these price-quality correlations

provide any indication about the degree of market efficiency (see, e.g., the argument between Hjorth-Andersen (1992), Maynes (1992) and Ratchford and Gupta (1992)). Ratchford and Gupta (1992) argue in favour of the use of price-characteristics frontiers to delineate the subset of efficient products (in line with, e.g., Kamakura et al. (1988)), i.e., products worthwhile buying by fully informed consumers with according preferences.

While price-quality correlations make it necessary to aggregate the quality dimension of a product into a single index, non-parametric frontier estimators determine the relative efficiency of products taking into account price and all multi-dimensional quality aspects simultaneously. Heterogeneous consumers may prefer different product attributes and a one-dimensional quality index, which ideally reflects the preferences of a "representative" consumer, may produce misleading results. Even in the absence of information on consumer preferences, these efficiency measures at least provide an easily computable index of efficiency in markets with differentiated products. This explains why there are also a number of price-characteristics frontier studies where only a single market is scrutinised in detail. A full fledged analysis of market efficiency would ideally have to comprise the market shares of individual products, analyse transaction rather than list prices, consider dynamic aspects of market efficiency, etc. Because of lacking data, this is mostly neglected and for the same reason our analysis is unable to consider these aspects.

While the bias problem discussed in the introduction is already relevant for single market studies, it is certainly highly problematic to compare market efficiencies *across* markets when data properties and model specifications differ. The bias problem certainly pertains to the standard frontier methodology applied by, e.g., Kamakura et al. (1988).[6] These authors studied 20 markets in an effort to quantify potential welfare gains from eliminating inefficient buys. In each market, between 18 and 47 products were observed and each product was characterised by 2 to 10 characteristics. The authors found 52% of all products to be inefficient, average inefficiency being at 10% and conclude that inefficiency varies substantially over markets. Much of the variation can be explained by differential consumer search strategies related to the product price but is also driven by factors such as purchasing frequency, budget share and involvement. A later study by Ratchford et al. (1996) based on the same methodology

[6] Note that the bias problem is not limited to standard frontier approaches. For instance, in his pioneering study, Hjorth-Andersen (1984) analyzed the efficiency of 127 markets to assess whether prices are valid quality indicators. In the markets analyzed, 5 to 34 different products were observed on each market and products were characterised by 3 to 16 characteristics. Efficiency was assessed by a simple vector dominance comparison, which is similar to another non-parametric frontier estimation method known as the Free Disposal Hull (Deprins et al. (1984)). The analysis revealed that 54% of all markets were inefficient and that the average inefficiency across all markets was at 13%. Hjorth-Andersen (1984) concludes that prices are not a perfect signal for quality, but that welfare losses due to inefficient buys are much lower than previously thought.

comprised 60 markets with an average of 17 products and compared frontier measures with price-quality correlations. The results based on frontier estimators implied an average inefficiency of 18%. All frontier measures employed are highly correlated, but at the same time the correlation with the price-quality measures is low.

While non-parametric frontier estimators seem by now a standard tool for product benchmarking (see, e.g., Fernandez-Castro and Smith (2002) or Lee et al. (2004)), the statistical properties of these estimators and the implications for the interpretation of results have been largely ignored. Therefore, it is interesting to review the results derived in the market efficiency literature in view of this issue. For instance, the fact that, e.g., Kamakura et al. (1988) in their study comprising 20 markets find above average "market efficiency" for datasets with a below average number of observations and an above average number of parameters (and vice versa) raises the question whether this may – at least in part – be due to different degrees of bias affecting the results for different markets. Hence, their conclusions on the relation between price/budget share, purchasing frequency and involvement must be viewed with some caution. Equally so, the high correlation between all frontier measures found in Ratchford et al. (1996) cannot be interpreted as evidence that these results are robust. Instead, different frontier estimators may suffer from the same type of bias which may contribute to the high correlation.

2.2. Hedonic Price-Quality Relations: Non-Parametric Frontier Estimation and the Nature of the Problem at Hand

The characteristics approach to consumer theory developed by Lancaster (1966) writes utility not as a function of a vector of goods but of their characteristics. Characteristics are normally assumed to be objective, in contrast to the concept of attributes widely used in psychology and marketing. In economics, building upon the characteristics approach to consumer theory, Rosen (1974) developed a substantive theoretical framework to study market equilibria for heterogeneous commodities differing along multiple characteristics (see Mendelsohn (1987) for an early review). Basically, one seeks to obtain an implicit price for the vector of observed characteristics to aggregate these into a measure of value. Recently, there emerged a series of applications of non-parametric frontier specifications imposing minimal assumptions (mainly monotonicity and convexity) to characterise the price quality correspondence and to explicitly measure the eventual presence of price inefficiencies.

The remainder of this section on the estimation of non-parametric frontier efficiency of production starts with some basic definitions. Since we only intend to briefly summarise the main arguments of an existing literature (see Simar and Wilson (2000)), we keep this

presentation in line with earlier contributions and formulate it in terms of the production approach. A production possibility set describes which amounts of some *p* inputs *x* can produce some *q* outputs *y*:

(1) $$\Psi = \left\{(x, y) \in \mathbb{R}_+^{p+q} \,\middle|\, x \text{ can produce } y\right\}.$$

In our case, outputs are product characteristics whereas the input is the price of the product. As developed below, an efficiency measure is a price-performance ratio based on the simultaneous assessment of multiple outputs and can be interpreted as a measure of customer value (see Staat et al. (2002)). An input requirement set $X(y)$ is defined as:

(2) $$X(y) = \left\{x \in \mathbb{R}_+^{p} \,\middle|\, (x, y) \in \Psi\right\}.$$

The assumptions maintained with respect to these sets are that a) Ψ is closed and convex and that $X(y)$ is closed and convex for all *y*; b) nonzero production of *y* requires nonzero inputs *x*; and c) *x* and *y* are strongly disposable. The efficient boundary of the input requirement set $X(y)$ is defined as:

(3) $$\partial X(y) = \left\{x \,\middle|\, x \in X(y), \theta x \notin X(y)\ \forall\ 0 < \theta < 1\right\},$$

and $\theta_k = \min\left\{\theta \,\middle|\, \theta x_k \in X(y_k)\right\}$ is the input-oriented efficiency measure for a given combination of inputs and outputs (x_k, y_k). It indicates the proportional reduction of observed inputs (prices) that would make the evaluated observation efficient.

The sets Ψ and $X(y)$ as well as the efficient boundary $\partial X(y)$ are not directly observed, but for any given sample of observations $\mathcal{S} = \left\{(x_i, y_i) \,\middle|\, i = 1, ..., n\right\}$, the sample equivalents of (2), $\hat{X}(y)$, and (3), $\partial \hat{X}(y)$, as well as of θ can be derived. Specifically, $\hat{\theta}_k$ is the estimate of θ_k obtained by solving:

(4) $$\hat{\theta}_k = \min\left\{\theta \,\middle|\, y_k \le \sum_{i=1}^{n} \lambda_i y_i;\ \theta x_k \ge \sum_{i=1}^{n} \lambda_i x_i;\ \theta > 0;\ \sum_{i=1}^{n} \lambda_i = 1;\ \gamma_i \ge 0,\ i = 1, ..., n\right\}.$$

The efficiency measure is calculated as the optimal proportional reduction of inputs for observation *k*, given that the benchmark units (the terms containing the λ_i) produce at least as much output with no more inputs than $\hat{\theta}_k x_k$. Efficient products in terms of qualities and price jointly constitute the piece-wise linear reference technology. The condition $\Sigma_{i=1}^{n} \lambda_i = 1$ maintained in (4) leads to an evaluation based on a variable returns to scale technology.[7]

[7] Without the latter condition, one allows for a free scaling up or down of price and characteristics, which is not warranted given the nature of our data (see also below).

Efficient products obtain an efficiency score of unity, while inefficient products obtain a score below unity.

These input-oriented efficiency estimates based on non-parametric frontier methods are positively (upwards) biased. Since the observed frontier $\partial\hat{X}(y)$ can only be as good as the theoretical frontier $\partial X(y)$, but never better, the benchmark based on sample observations is in all likelihood weaker than $\partial X(y)$. Hence, the upward bias of the efficiency scores $\hat{\theta}$.

Theoretical results on the bias, which would allow correcting for it, are only available for the one-input and one-output case. Assuming a monotone, concave production function with a frontier function $g(\cdot)$ that is twice continuously differentiable at x_0, Simar and Wilson (2000) state the following expression for the asymptotic bias:[8]

$$\text{(5)} \qquad \text{asymp. bias of } \hat{g}(x_0) = -n^{-2/3}\left(-g''(x_0)/2\Big/ f\left(x_0, g(x_0)\right)^2\right)^{1/3} c_1,$$

where c_1 is a constant and $f(\cdot)$ is the density. This bias depends on sample size *n* as well as on "the curvature of the frontier and the magnitude of the density at the frontier" (Simar and Wilson (2000: p. 59)). It should be intuitively clear that that this bias decreases in density and increases in curvature. Thus, in (i) large samples with a (ii) high density of observations around a frontier and with a (iii) mild curvature, one should expect a relatively small bias. By contrast, when (i) the sample is small, (ii) the density of observations around the frontier is low, and (iii) the frontier exhibits kinks (changes in curvature), then a relatively large bias is to be expected. It should be evident that this bias is exacerbated with a rising number of characteristics used for the evaluation of observations.

For the case with more than one input and/or more than one output, the bootstrap seems to be the only way to correct for the bias in DEA-type estimators. First, a naïve bootstrap approach would be to resample with replacement samples of size *n* from the original data, but it is well-known that this method is inconsistent. Second, a simple and appealing idea is the sub-sampling bootstrap whereby sub-samples of smaller size are drawn. While Kneip et al. (2008) have shown that this is consistent, the exact size of the sub-samples is critical for smaller data sets, but the determination of this size remains an open issue. Finally, there are bootstrap methods that employ smoothing techniques to approximate a distribution of the efficiency scores from which pseudo scores are re-sampled. However, these techniques are somewhat involved: for instance, it may be required to smooth the distribution of the efficiency estimates, to reflect efficiency scores at the limit of their distribution, to transform the data from Cartesian

[8] See their section 3 and the results obtained by Gijbels et al. (1999). As a matter of fact, this expression given in Simar and Wilson (2000) pertains to the output oriented case.

to spherical coordinates, to calculate pseudo data from estimates of pseudo scores, these pseudo data can in turn be used to estimate bootstrap efficiency scores, (Simar and Wilson (2000)).[9]

Some statistical procedures for testing various restrictions in the context of nonparametric models of technical efficiency do exist in the literature. For instance, tests for whether inputs or outputs are irrelevant and aggregation tests have been formulated and bootstrap estimation procedures yielding appropriate critical values for the test statistics have been provided in combination with evidence on the true sizes and power of these tests statistics obtained from Monte Carlo experiments (see, e.g., Simar and Wilson (2001)).[10]

However, these test statistics are designed to compare nested model specifications only. To the best of our knowledge no general statistical procedures have been proposed in the literature allowing comparing potentially non-nested model specifications and models based on potentially different samples (differing, among others, in sample size or nature of the data (time series, cross-section, panel, …)). Furthermore, the above mentioned statistical test procedures are demanding in terms of informational requirements. In particular, these tests assume perfect data availability in that the underlying samples of inputs and outputs must be readily available. However, when comparing results across different studies (e.g., in view of formulating policy advise), often more limited information is available: for instance, sample size, number and nature of inputs and outputs, average efficiency levels (or some other summary statistic), etc.

2.3. Zhang and Bartels (1998) on Comparing Frontier Estimates across Samples and Specifications

Zhang and Bartels (1998) using data on electric utilities demonstrate that average efficiency is lower when there are more observations in a model for a given number of variables used. They argue in favour of using a Monte Carlo-type of approach limiting the size of larger samples to the size of the smallest sample in order to derive average sample efficiencies to be compared across samples in a pragmatic way. We follow Zhang and Bartels (1998) in drawing (without replacement) random sub-samples from larger samples such that they match the size of the smaller samples obtained for a different product of the same category. By repeating this process a large number of times and averaging over the results we obtain the expected market efficiency for larger samples if only a smaller sample had been available. In this way, we make some

[9] Gstach (1995) already proposed a smoothed bootstrap technique in a more ad hoc fashion.
[10] Another example of such a nested test is related to testing hypotheses regarding returns to scale (e.g., Simar and Wilson (2002)).

progress towards disentangling the sample size effect as described by Zhang and Bartels (1998) from (expected) differences in market efficiency of products from the same category.

However, one should notice that the Zhang and Bartels (1998) method provides no correction for bias in a technical sense, but simply ensures that results share a similar degree of bias. Note also that the application of this approach artificially limits the precision of the estimates. Indeed, reducing the number of observations decreases the level of precision to the one for the market with the smallest sample size. Thus, the gain in one desirable property – increased comparability– comes at the loss of another desirable property –the overall precision of the estimates.

Furthermore, the original Zhang and Bartels (1998) article only focuses on remedying differences in sample size for models with the same number of parameters. Since the non-parametric estimators have a rate of convergence that is inversely related to the number of parameters in the model (e.g., Kneip et al. (1998)), the bias increases with the number of parameters. To maintain the precision of the estimates when parameters are added to a model, the number of observations must increase considerably. The simulation results by, e.g., Pedraja-Chaparro et al. (1999) are compatible with the theoretical results obtained by Kneip et al. (1998) in that the number of observations must ideally double for each parameter added to a specific model to retain the same level of precision for the estimates. Thus, one way to deal with the fact that different models are estimated using different numbers of parameters is to adjust the number of observations in the samples accordingly.

An alternative for adjusting the number of observations is to simply drop some parameters from the models containing relatively more parameters, or to aggregate some parameters into a single parameter. However, Orme and Smith (1996) demonstrate that dropping a parameter that is highly correlated with another parameter from the model or dropping a parameter that is basically uncorrelated with the rest of the parameters may have very different effects on the results. Therefore, it is not obvious how dropping or aggregating parameters contributes to the solution of the underlying problem. Consequently, we explore each of these strategies in turn.

In brief, when different numbers of parameters are available for different markets, this can be considered by either adjusting the sample size accordingly or by dropping parameters from the model to arrive at comparisons based upon an equal number of parameters and observations. However, the latter strategy is only possible when there are relatively more observations for the products with more parameters and thus cannot be generalized to all situations in which a comparison of markets is needed.

Consider, for instance, the markets for hair conditioners and dishwashers evaluated by Kamakura et al. (1988). There are 47 observations for hair conditioners which are evaluated by two attributes, but only 25 observations for dishwashers which are evaluated along ten attributes (see their Table 4, p. 299). The fact that the average efficiency of hair conditioners is estimated at only 71.3%, whereas that of dishwashers is estimated at 91.1% is no surprise. One should expect that 25 observations evaluated on 10 characteristics turn out to appear more efficient than 47 observations benchmarked on just 2 characteristics if the true underlying efficiencies of the markets are not too different. While one ideally would like to adjust the data such that comparisons across markets become sensible, one should realise that this may become practically impossible for certain combinations of data and model characteristics. For instance, to keep the entire information available for conditioners, we would need 6400 ($=25*2^{(10-2)}$) observations on dishwashers. Alternatively, we could use both attributes for conditioners, but only one (out of ten) for dishwashers, resulting in two roughly comparable settings: 47 observations and two characteristics vs. 25 observations and one characteristic. A final possibility would be to keep two characteristics for dishwashers and limit the number of observations for conditioners to just 25. Without commenting on the open question of which markets can be meaningfully compared to one another under ideal circumstances, we simply point out that none of these technically feasible comparisons seem to make much sense. Hence, practical issues may limit the scope for making meaningful comparisons between markets and their informational efficiencies.

Therefore, a priori the Zhang and Bartels (1998) method cannot be universally applied. Furthermore, it is informationally demanding in that the underlying samples of inputs and outputs must be accessible. This observation calls for the search for more general methods capable to handle any configuration of sample sizes and specifications to which we return in the conclusions. Now, we first turn to the presentation of the data which we utilise to illustrate the possibilities and limitations of the Zhang and Bartels (1998) approach.

One admittedly limited way out is to concentrate on making international comparisons for a single market and then adjusting for differences in the sample size for a given number of dimensions. Another way out is to focus on comparing efficiency within the same market over time (e.g., using discrete time indexes (see Chumpitaz et al. (2009) for an example), provided we can define coherent product life cycles for some of the markets involved. Maybe one can even think about combining both of these strategies. Obviously, these strategies would severely limit the use of hedonic frontiers in a marketing context.

3. Data: Sample Description

To investigate whether the empirical results and hence the conclusions derived in earlier market efficiency studies may in fact have been influenced by the properties of the estimators applied, this contribution assesses the market efficiency for two product categories using several datasets for computer parts. The data used in the current analysis are taken from hardware tests published in the German computer magazine "CHIP" in 2005. These hardware test results were then available at the website of this magazine (www.chip.de) and have subsequently been updated. The information provided is similar to that contained in the Consumer Reports data used in previous studies, but "CHIP" specializes in computers and computer related products. We utilize data on two product categories: (i) hard disk drives (HDD), and (ii) CD/DVD-writers. Since "CHIP" updates prices from internet vendors and we can safely assume that haggling is impossible for internet transactions, we are among the first studies analysing transaction rather than list prices, as done by most studies in the literature.

These products have been selected because one does not expect the consumer's attitude to vary between them. The buyers of these products can be considered expert buyers. Since they themselves normally fit these computer parts into the computers and since this requires substantial technical expertise, it is likely that we deal with "prosumers". At the same time, the price ranges in which these products sell are rather similar, so are the purchasing frequency, the involvement, and most likely any other aspect of shopping behaviour. Hence, we would expect similar market efficiency levels for the markets analyzed as far as the shopping behaviour of customers is concerned.[11] Of course, there may be other reasons for differences in market efficiency, like brand and retailer attributes, the phase of the product life cycle, the market structure, etc.[12]

These specific aspects of the data allow isolating the effect of sample size and model dimensions on average efficiency from other factors on the consumer side that may potentially lead to differences in average market efficiency. Efficiency differences may, however, exist because some products are clearly standard products whereas others pertain to more specialized needs. For the hard disks, the standard is IDE drives, while SCSI drives continue to exist along with these more common types of drives. Similarly, different form factors for external drives continue to play a role in the market for hard disks. Likewise, standard CD drives/writers are

[11] The market efficiency reported in the studies by Hjorth-Andersen (1984), Kamakura et al. (1988) or Ratchford et al. (1996) was most likely affected by differences in the above mentioned attitudes, since the market efficiency of very heterogeneous product categories was investigated.

[12] Of course, the market efficiency results are conditional upon the correct specification of the price characteristics hedonic relationship. It is well-known that when this relationship is misspecified due to unknown characteristics, then the interpretation of these efficiency estimates is problematic (see Varian (1988)).

now fitted into nearly every computer sold and DVD drives/writers are about as common as CD drives. We may surmise that the markets for the most common type of product are the ones with the fiercest competition and therefore the highest average efficiency, whereas the less common or newly marketed products are in an earlier stage of their product life cycle such that the maturity of the market and hence market efficiency is lower. Also some products are at the end of their life cycle and may be about to be phased out.

All products are evaluated in the test laboratory of CHIP with the same test set-up. HDDs as wells as CD/DVD writers were fitted into identical computers running the same software. For all HDDs (internal and external; based on IDE, SATA or SCSI technology; and the special case of notebooks (NB)) the attributes access time, transfer rate, data base performance, noise and power consumption function as outputs (characteristics).[13] Since all HDDs are evaluated with the same number of characteristics, it is sufficient to generate samples of equal size to compare the average efficiency of these markets on an equal footing. Here, the number of observations ranges between 6 (1" HDDs) and 33 (NB HDDs).

The situation is different for the CD/DVD-drives, since the number of characteristics varies slightly between 6 and 7 per product. Therefore, we provide a table listing these characteristics. Table 1 lists the products (rows) and the characteristics by which these products are evaluated (columns). Since all products are also evaluated by their price, this column is not represented in the table. The number of observations ranges between 5 and 31 for CD/DVD writers (see parentheses in the first column in Table 1).

Table 1: Products Categories and Evaluated Parameters

CD/DVD-Writers						
Type (# Obs.)	Specific	Write	Read	Features	Noise	Performance UDF
CD (5)	Manual	R/RW	CD	x		x
DVD (17)		DVD/CD	DVD/CD	x	x	
DVD slim (31)		DVD/CD	DVD/CD	x	x	

The fact that the number of observations *as well as* the number of characteristics differs across products complicates the comparison across these markets significantly. For CD writers the read/write and UDF performance are core features and also the documentation is relevant (listed in column 1 of Table 1). For DVD-writers, documentation is not considered, but noise level is now a relevant characteristic that was not considered for CD writers. Also, since DVD-

[13] In fact, these five criteria contain aspects of mobility for the external HDDs that are not contained in the evaluation of the other drives, while the SATA drives are also evaluated with respect to the performance on specific applications which are not relevant for the rest of the drives. CHIP provides no further details on how these slight variations across HDDs are integrated into the five criteria reported.

writers are really CD and DVD drives combined, the read and the write performance for both DVDs and CDs are considered separately. In the end, this results in 6 characteristics for CDs and 7 characteristics for DVDs.

CHIP simply aggregates the values for the single characteristics with a fixed set of weights and then arrives at a ranking for the products based on this weighted aggregate. Ideally, these weights should reflect the preferences of some "representative" consumer. But, one should realize that if it made sense to evaluate these products in such a way, then there would be no need for differentiated product variants in the first place since they could never coexist in the market if all consumers behaved like a "representative" consumer. CHIP seems aware of this: its website now allows readers to change the standard weights used by the magazine online according to their own preferences and then provides the corresponding ranking.

Since CD writers are the products with the lowest number of observations (5 as opposed to the 17 and 31 observations for the two types of DVD writers) and evaluated on the basis of a smaller number of characteristics (6 compared to 7), as explained at the end of section 2, there are different strategies to end up with a comparison of average market efficiency on an equal footing. While comparing all products on the basis of 5 observations regardless of the number of characteristics in the model does not lead to a comparison of equally precise/biased estimates (since with an equal number of observations the model with more characteristics tends to be more biased), we have two options to achieve this.

First, we can adjust the number of observations according to the different number of characteristics. In our case, this implies comparing the results for the CD market obtained with 5 observations on the basis of a model with 6 characteristics to results for DVD-drives obtained with a model with 7 characteristics and datasets for which the number of observations has been artificially limited to 10 (since one more characteristic necessitates doubling the number of observations).[14]

Second, we can also drop one characteristic from the DVD models and evaluate both the CD- and the DVD-writer markets based on 5 observations and 6 parameters. However, as Orme and Smith (1996) demonstrate, the results may change drastically depending on whether the dropped parameter (in our case, a characteristic) is correlated or not with other parameters.[15] Yet another variation on this second strategy is to aggregate parameters.

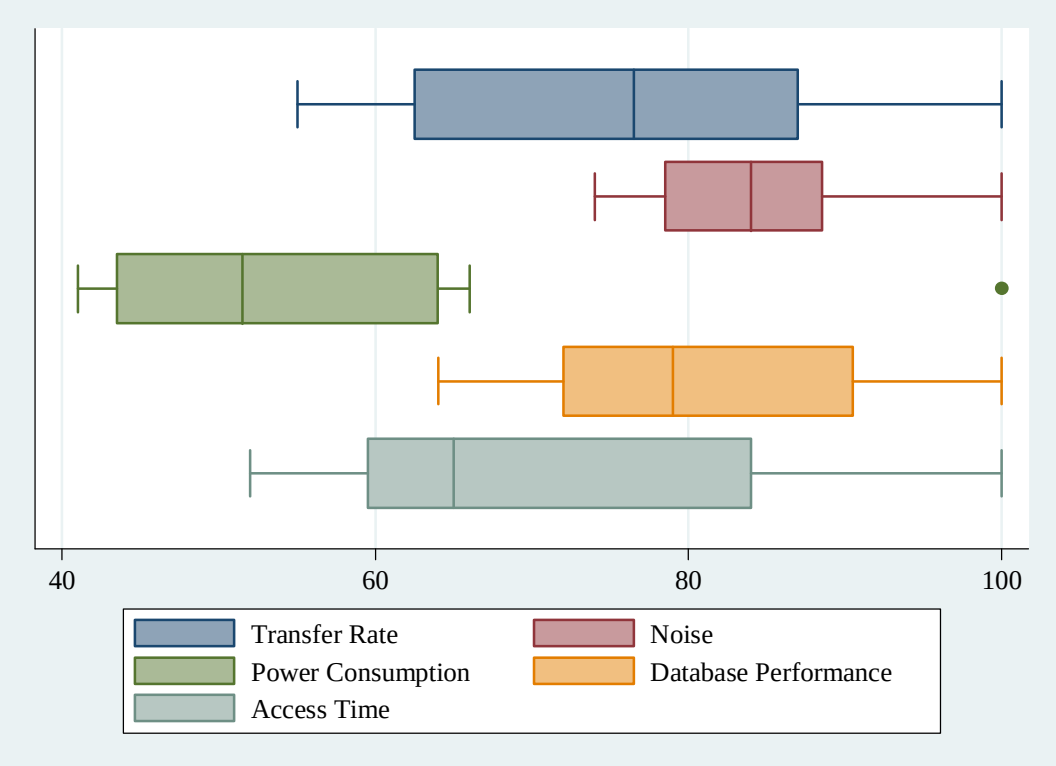
[14] Notice that this strategy is only possible when there are relatively more observations for the products evaluated with more characteristics. If this condition is not met, then this option is unavailable and one can only adopt the second strategy described below.

[15] Recall that dropping a parameter that is perfectly correlated with another parameter does not change the results at all, whereas dropping a parameter that is not correlated with any of the others often results in a decrease in average efficiency.

Before proceeding to the results of the simulation exercises, it is useful to stress several noteworthy aspects of the data. First, the number of products per market and the number of attributes observed here are in about the same range as in the early study by Hjorth-Andersen (1984), while slightly larger data sets were used by Kamakura et al. (1988). Hence, as far as the sample properties are concerned we would expect the same type of variation of efficiency between product categories as in these studies and even though the samples analysed are relatively small, they offer a typical and realistic case study. By contrast, one may suspect that technical products like the ones analyzed here are much more homogeneous than the products analyzed in other studies.

Second, to give a visual impression of the relative heterogeneity among even these technical products, Figure 1 provides box plots for the five characteristics of the 12 SCSI drives in the sample. The box plots have the usual interpretation: the box reflects the interquartile range, the whiskers include 75% of observations, and the dots are outliers outside the range of three standard deviations. Note that the data are scaled such that a larger value implies a better performance (e. g., a larger value for “noise level” implies “lower” noise) and that the optimal performance in each category is normalized to 100.

Figure 1: Box Plots for SCSI Drives



It is obvious that the values for all five characteristics span over a considerable range and that these drives are not homogeneous, but seem to be relatively differentiated. For instance, only for the noise level there is no observation below the value of 50. Furthermore, in the power consumption dimension, there is clearly a potential outlier situated outside the outer fences.[16] Obviously, this sample information may hide considerable individual variations. For example, in the case of power consumption one drive outperforms all others by far using less than half the power of the second best drive in the sample. Furthermore, some drives have an almost identical product concept that differs markedly from other drives in the same sample.

Therefore, while the attitude with which consumers shop for these products is likely to be identical across markets and while the estimation method ensures that bias effects are minimized, there is sufficient differentiation among the products such that inefficiency could be identified if present. Remember that the bias depends critically on the density of the observations around the relevant segment of the frontier: this implies that the distribution of product characteristics within the same market plays an important role for the resulting bias and that no a priori assessment of bias is possible.

4. Empirical Results

Table 2 presents the results for the HDDs, the first product category. The table is organised as follows: on the main diagonal, the average efficiencies for all products from a standard, input-oriented variable returns to scale specification are displayed. Remember that it is this average efficiency displayed on the diagonal which is interpreted as a measure of market efficiency in the studies discussed above, resting on the evaluation of the entire samples for which the size is listed in the second column. The column headers give the number of observations used for estimation and the first column lists the product type. The off-diagonal cells list results that were obtained by drawing, as described above, smaller sub-samples.

For instance, the 1" HDDs, for which there are only six observations, seem to constitute a perfectly efficient market (see bottom row). However, this may be the consequence of the very small number of observations. Other average efficiencies on the main diagonal, where standard DEA results for original sample sizes are reported, range between below 80% to 89%. Furthermore, note that while the relationship between sample size and average efficiency is by no means a linear one – other effects such as the density of observations around specific

[16] It is important to underscore the fact that especially deterministic frontier methods like DEA may suffer from the presence of outliers. A specialised literature has developped to identify outliers in this context: see, e.g., Fox et al. (2004), Seaver and Triantis (1995), Simar (2003), and Wilson (1993).

segments of the frontier play a role – a clear tendency for smaller samples to be attributed higher average efficiency can be observed. Standard IDE drives seem to be the most inefficient product, but this may again be due to the fact that the sample for IDE drives is the second largest sample in this product category. When the expected average efficiency is calculated for IDE drives for smaller sample sizes these drives appear to be relatively more efficient the smaller the sample becomes (compare the results given in the row "IDE" to the results for the other product types in the respective columns). This is in line with the intuition that a market with a huge trade volume – the market for the "standard" product – should in fact be among the more efficient markets. This would have been contradicted by the results generated on the basis of a naïve application of the frontier model to the original samples.

Table 2: Results for Hard Disk Drives

Type	Obs.	33	22	21	19	13	12	6
HDD NB	33	82.30%	86.21%	86.52%	87.63%	91.13%	90.63%	95.87%
HDD IDE	22		79.82%	80.78%	82.83%	88.34%	90.03%	95.67%
HDD SATA	21			81.02%	81.42%	85.22%	84.71%	89.65%
HDD 3.5"	19				85.41%	88.23%	89.67%	95.57%
HDD 2.5"	13					88.52%	88.94%	94.06%
HDD SCSI	12						82.92%	91.83%
HDD 1"	6							100%

As another example, the second most efficient product according to the standard results (main diagonal) are 2.5" HDDs with an average efficiency of 88.5%. Since these drives are needed for specific purposes only, one could conjecture that it is unlikely that this market would be among the more efficient ones. When comparing them to other product types on the basis of like sample size, these drives are in the midrange in terms of efficiency and not in any particularly efficient position.

Similar effects can be observed for the product category of CD/DVD writers in Table 3. This table is structured very much like Table 2 above. Looking at the left part, comparisons are made correcting for sample size, while maintaining the original model specification irrespective of the number of dimensions (i.e., 6 parameters for CDs and 7 for DVDs). For instance, the column headed "5" lists the average market efficiency for the category of CD/DVD writers based on 5 observations only, i.e., for DVD writers with a slim size factor and for standard DVD writers the results are based on five observations to make them comparable to the CD writer group of products where only 5 models are left, even though respectively 17 and 31 observations were available for these two product types. We have listed the standard results for the DVD-markets under the respective heading and as above simulated results for the market

with the most observations (31 for slim sized DVD writers) for all smaller sample sizes (here only one additional simulation for 17 observations).

Table 3: Results for CD/DVD-Writers

	Adjusting for sample size only			Adjusting for sample size & dimensions		
	31	17	5	*10 for DVD/ 5 for CD*	*5 (less Noise)*	*5 (less DVD read/CD write)*
DVD writer	88.60%	92.11%	97.87%	*95.15%*	*97.28%*	*96.78%*
DVD writer Slim		95.20%	98.65%	*96.70%*	*93.81%*	*98.44%*
CD writer			90.67%	*90.67%*	*90.67%*	*90.67%*

From the standard results listed on the main diagonal in the left part of the table – where all observations in the respective samples were used for estimation – one may infer that the market for standard DVD writers is the most inefficient of the three, since the average of 88.60% is the lowest on the main diagonal. An interpretation of the same standard results that ignores the effects just discussed would consider the market for DVD writers with a slim size factor the most efficient market, because its average efficiency of 95.20% is the highest listed on the main diagonal. The market for CD writers is positioned in between both extremes (90.67%).

Notice that the second column with heading “17” can be interpreted along similar lines if one were only interested in comparing the two markets for DVD writers, but disregard the CD writer market. In the latter case one takes the sample size of slim DVD writers as a starting point for the comparison. This picture changes markedly when looking at the column headed “5”. This column allows comparing all 3 product types based on the same sample size, namely the size of the smallest market. The markets for both types of DVD writers seem to be about equally efficient (98.65% and 97.87%, respectively), while the market for CD writers seems substantially less efficient. The latter results compare market efficiency across product types on a more equal footing than the results on the main diagonal, but the difference between the DVD and CD results may be exaggerated because there is one more parameter in the DVD model.

Turning attention to the right part of Table 3, we also provide results for the DVD markets for 10 observations, i.e., twice the number of observations available for the CD market (where products are evaluated with a model that has one characteristic less and the number of observations stays put to the original 5 observations). As explained above, one more characteristic and twice the number of observations should make these results comparable to the ones for the CD market. Another way to generate comparable results is to drop characteristics from the DVD models. As mentioned before, this may lead to different changes in results

depending on how the characteristic dropped from the model correlates with the rest of the characteristics. With real data, there are neither perfectly correlated nor completely uncorrelated characteristics that could be dropped. We have chosen to drop one characteristic that was not correlated with any of the other characteristics in the model, namely noise level, and one characteristic that was strongly and significantly correlated with another one: for slim size DVD writers and for normal DVD writers we dropped DVD reading performance respectively CD writing performance, both having a correlation coefficient above 0.5 with another characteristic and significant at the 5% level. Notice that all results added to adjust for the different number of characteristics of the models have been put in italics.

Looking at the first column of the right part, the average efficiency for the DVD markets drops slightly compared to the column headed “5”. The inefficiency of the CD market is confirmed and it seems that CD writers are an outdated product that has been superseded by DVD drives that write CDs as well. The remaining CD writers are rather few, seem to be phasing out, and the market does not appear to be very efficient anymore. The newer DVD drives have currently a larger sales volume and are traded much more efficiently.

Finally, we discuss the results for the models for DVD writers where one parameter was dropped. These results are listed in the two last columns of Table 3. The changes of the results should be interpreted with care. On the one hand, when dropping a variable that is uncorrelated with any other, one destroys a maximum of information and therefore one would expect a palpable change in results. This happens for slim size DVD writers where efficiency drops by nearly 5% due to ignoring the noise level characteristic, but dropping the same characteristic for standard DVD writers changes almost nothing. On the other hand, dropping a positively correlated characteristic leads to results that are somewhat more comparable to the results obtained by doubling the number of observations for the model with an extra parameter.

These results for a variety of hard disk drives and CD/DVD writers both provide evidence about the potential bias in average market efficiency due to differences in sample size and dimensionality. For instance, for both product categories, the ranking of markets based on efficiency varies considerably when the size of the samples is adjusted to allow for comparison. These results should provide a fair warning against the current practice of taking average market efficiency results at face value and comparing them across markets. Especially the regressions of average efficiencies on price level (see Kamakura et al. (1988)) and potentially other

variables seem problematic, considering that the dependent variable is composed of biased efficiency scores.[17]

5. Conclusions

An empirical application on a few varieties of computer hardware components sold in the German market has served to illustrate the possibilities and limitations of the Zhang and Bartels (1998) procedure in comparing inefficiency levels across markets. The main message of this empirical illustration is that all currently reported results on product market efficiency in the literature should be interpreted (and compared) with great care. Irrespective of informational requirements, the key methodological message from this contribution is that there currently seems to be no simple means available for comparing average efficiencies of datasets of different size and involving different specifications. This is a serious lacuna requiring more attention in the future.

The Zhang and Bartels (1998) method lacking general applicability, it is important that other avenues are explored to compare efficiency scores across different samples and/or different specifications. Apart from the remarks on recent bootstrapping proposals in subsection 2.2., in the recent literature one can find several potentially promising alternatives. One recent potential solution is the use of the order-*m* estimators proposed in Cazals, Florens and Simar (2002) that do not suffer from the curse of dimensionality at all and that furthermore tend to be robust for any eventual outliers. Martins-Filho and Yao (2008) propose another nonparametric order α frontier model with very similar properties. An earlier proposal for constructing confidence intervals for average efficiency scores in both non-parametric and parametric frontiers is found in Atkinson and Wilson (1995). Yet another recent estimator proposed by Allon et al. (2007) employs convex entropic nonparametric estimators to estimate concave production frontiers. Finally, without the ambition of completeness, further new frontier estimators have been proposed in Bouchard et al. (2005) and Post et al. (2002), among others. However, the small sample properties of most of these alternative frontier estimators are unknown. Therefore, a comparative analysis is being called for to test the relative strengths and weakness of these estimators.

This search for a proper remedy is important when drawing conclusions from this type of research for both public (e.g., industrial sector analysis) and private (e.g., decisions about entering or leaving a market in terms of potential surpluses) policies. One major practical

[17] Inferential problems related to the explanation of efficiency patterns in a second stage analysis have recently been investigated in Simar and Wilson (2007).

implication to facilitate this search for a proper remedy is that the data used in these market efficiency studies should ideally be available for future studies.[18]

References

Allon, G., M. Beenstock, S. Hackman, U. Passy, A. Shapiro (2007) Nonparametric Estimation of Concave Production Technologies by Entropic Methods, *Journal of Applied Econometrics*, 22(4), 95-816.

Athanassopoulos, A. D. (2004) Assessing the Selling Function in Retailing: Insights from Banking, Sales Forces, Restaurants and Betting Shops, in: W.W. Cooper, L.M. Seiford, J. Zhu. (eds.) *Handbook on Data Envelopment Analysis*, Kluwer, Boston, 455-480.

Atkinson, S.E., P.W. Wilson (1995) Comparing Mean Efficiency and Productivity Scores from Small Samples: A Bootstrap Methodology, *Journal of Productivity Analysis*, 6(2), 137-152.

Balcombe, K., I. Fraser, J.H Kim (2006) Estimating Technical Efficiency of Australian Dairy Farms Using Alternative Frontier Methodologies, *Applied Economics***,** 38(19), 2221-2236.

Berger, A.N., R.S. Demsetz, P.E. Strahan (1999) The Consolidation of the Financial Services Industry: Causes, Consequences, and Implications for the Future, *Journal of Banking and Finance*, 23(2-4), 135-194.

Berger, A., D. Humphrey (1997) The Efficiency of Financial Institutions: International Survey and Directions for Future Research, *European Journal of Operational Research*, 98(2), 175-212.

Blinder, A., E.R.D. Canetti, D.E. Lebow, J.B. Rudd (1998) *Asking about Prices: A New Approach to Understanding Price Stickiness*, New York, Russel Sage Foundation.

Bodell, R., R. Kerton, R. Schuster (1986) Price as a Signal of Quality: Canada in the International Context, *Journal of Consumer Policy*, 9(4), 431-444.

Bouchard, G., S. Girard, A. Iouditski, A. Nazin (2005) Some Linear Programming Methods for Frontier Estimation, *Applied Stochastic Models in Business and Industry*, 21(2), 175-185.

Carlton, D. W. (1989) The Theory and Facts About How Market Clears: Is Industrial Organization Valuable for Understanding Macroeconomics?, in: R. Schmalensee, R. D. Willig (eds.) *Handbook of Industrial Organization, Volume 1*, North Holland, Amsterdam, 909-946.

Cazals, C., J.-P. Florens, L. Simar (2002) Nonparametric Frontier Estimation: A Robust Approach, *Journal of Econometrics*, 106(1), 1-25.

Chumpitaz, R., K. Kerstens, N. Paparoidamis, M. Staat (2009) Hedonic Price Function Estimation in Economics and Marketing: Revisiting Lancaster's Issue of "Noncombinable" Goods, *Annals of Operations Research*, online first.

Daraio, C., L. Simar (2007) *Advanced Robust and Nonparametric Methods in Efficiency Analysis: Methodology and Applications*, Springer, Heidelberg

Deprins, D., L. Simar, H. Tulkens (1984) Measuring Labor Inefficiency in Post Offices, in: M. Marchand, P. Pestieau, H. Tulkens (eds.) *The Performance of Public Enterprises: Concepts and Measurements*, North Holland, Amsterdam, 243-267.

Dexter, F., L. O'Neill, L. Xin, J. Ledolter (2008) Sensitivity of Super-Efficient Data Envelopment Analysis Results to Individual Decision-Making Units: An Example of Surgical Workload by Specialty, *Health Care Management Science*, 11(4), 307-318.

[18] In an effort to take the lead, we make the data used in this study available upon simple request.

De Witte, K., R.C. Marques (2010) Designing Performance Incentives: An International Benchmark Study in the Water Sector, *Central European Journal of Operations Research*, forthcoming.

Dyson, R.G., R. Allen, A.S. Camanho, V.V. Podinovski, C.S. Sarrico, E.A. Shale (2001) Pitfalls and Protocols in DEA, *European Journal of Operational Research*, 132(2), 245-259.

Färe, R., S. Grosskopf, E. Kokkelenberg (1989) Measuring Plant Capacity, Utilization and Technical Change: A Nonparametric Approach, *International Economic Review*, 30(3), 655-666.

Faulds, D.J., O. Grunewald, D. Johnson (1995) A Cross-National Investigation of the Relationship between the Price and Quality of Consumer Products: 1970-1990, *Journal of Global Marketing*, 8(1), 7-25.

Fernandez-Castro, A.S., P.C. Smith (2002) Lancaster's Characteristics Approach Revisited: Product Selection Using Non-Parametric Methods, *Managerial and Decision Economics*, 23(2), 83-91.

Fox, K.J., R.J. Hill, W.E. Diewert (2004) Identifying Outliers in Multi-Output Models, *Journal of Productivity Analysis*, 22(1), 73-94.

Gijbels, I., E. Mammen, B.U. Park, L. Simar (1999) On Estimation of Monotone and Concave Frontier Functions, *Journal of the American Statistical Association*, 94(445), 220-228.

Gstach, D. (1995) Comparing Structural Efficiency of Unbalanced Subsamples: A Resampling Adaptation of Data Envelopment Analysis, *Empirical Economics*, 20(3), 531-542.

Hjorth-Andersen, C. (1984) The Concept of Quality and the Efficiency of Markets for Consumer Products, *Journal of Consumer Research*, 11(2), 708–718.

Hjorth-Andersen, C. (1992) Alternative Interpretations of Price-Quality Relations, *Journal of Consumer Policy*, 15(1), 71-82.

Hollingsworth, B. (2003) Non-Parametric and Parametric Applications Measuring Efficiency in Health Care, *Health Care Management Science*, 6(4), 203-218.

Hollingsworth, B., P.J. Dawson, N. Maniadakis (1999) Efficiency Measurement of Health Care: A Review of Non-Parametric Methods and Applications, *Health Care Management Science*, 2(3), 161-172.

Ibourk, A., B. Maillard, S. Perelman, H.R. Sneessens (2004) Aggregate Matching Efficiency: A Stochastic Production Frontier Approach, France 1990–1994, *Empirica*, 31(1), 1-25.

Kamakura, W.A., T.B. Ratchford, J. Agrawal (1988) Measuring Market Efficiency and Welfare Loss, *Journal of Consumer Research*, 15(3), 289-302.

Kneip, A., B.U. Park, L. Simar (1998) A Note on the Convergence of Nonparametric DEA Estimators for Production Efficiency Scores, *Econometric Theory*, 14(6), 783-793.

Kneip, A., L. Simar, P.W. Wilson (2008) Asymptotics and Consistent Bootstraps for DEA Estimators in Nonparametric Frontier Models, *Econometric Theory*, 24(6), 1663-1697.

Lancaster, K. (1966) A New Approach to Consumer Theory, *Journal of Political Economy*, 74(1), 132-157.

Lee, J.-D., A. Repkine, S.-W. Hwang, T.-Y. Kim (2004) Estimating Consumers' Willingness to Pay for Individual Quality Attributes with DEA, *Journal of the Operational Research Society*, 55(10), 1064-1070.

Martins-Filho, C., F. Yao (2008) A Smooth Nonparametric Conditional Quantile Frontier Estimator, *Journal of Econometrics*, 143(2), 317-333.

Maynes, E. S. (1992) Salute and Critique: Remarks on Ratchford and Gupta's Analysis of Price-Quality Relations, *Journal of Consumer Policy*, 15(1), 83-96.

Maynes, E.S., T. Assum (1982) Informationally Imperfect Consumer Markets: Empirical Findings and policy. Implications, *Journal of Consumer Affairs*, 16(1), 62-87.

Mendelsohn, R. (1987) A Review of Identification of Hedonic Supply and Demand Functions, *Growth and Change*, 18(1), 82-92.

Morris, R.T., C.S. Bronson (1969) The Chaos of Competition Indicated by Consumer Reports, *Journal of Marketing*, 33(3), 26-34.

Orme, C., P. Smith (1996) The Potential for Endogeneity Bias in Data Envelopment Analysis, *Journal of the Operational Research Society*, 47(1), 73-83.

Paradi, J.C., S. Vela, Z. Yang (2004) Assessing Bank and Bank Branch Performance: Modelling Considerations and Approaches, in: W.W. Cooper, L.M. Seiford, J. Zhu. (eds.) *Handbook on Data Envelopment Analysis*, Kluwer, Boston, 349-400.

Pedraja-Chaparro, F., J. Salinas-Jiménez, P. Smith (1999) On the Quality of the Data Envelopment Analysis Model, *Journal of the Operational Research Society*, 50(6), 636-644.

Polachek, S., J. Robst (1998) Employee Labor Market Information: Comparing Direct World of Work Measures of Workers' Knowledge to Stochastic Frontier Estimates, *Labour Economics*, 5(2), 231-242.

Post, T., L. Cherchye, T. Kuosmanen (2002) Non-Parametric Efficiency Estimation in Stochastic Environments, *Operations Research*, 50(4), 645–655.

Ratchford, B.T., J. Agrawal, P.E. Grimm, N. Srinivasan (1996) Toward Understanding the Measurement of Market Efficiency, *Journal of Public Policy and Marketing*, 15(2), 167-184.

Ratchford, B.T., P. Gupta (1990) On the Interpretation of Price-Quality Relations, *Journal of Consumer Policy*, 13(4), 389–411.

Ratchford, B.T., P. Gupta (1992) On Estimating Market Efficiency, *Journal of Consumer Policy*, 15(3), 275-293.

Rosen, S. (1974) Hedonic Prices and Implicit Markets: Production Differentiation in Pure Competition, *Journal of Political Economy*, 82(1), 34-55.

Seaver, B., K. Triantis (1995) The Impact of Outliers and Leverage Points for Efficiency Measurement and Evaluation Using High Breakdown Procedures, *Management Science*, 41(6), 937-956.

Shestalova, V. (2003) Sequential Malmquist Indices of Productivity Growth: An Application to OECD Industrial Activities, *Journal of Productivity Analysis*, 19(2-3), 211-226.

Simar, L. (2003) Detecting Outliers in Frontier Models: A Simple Approach, *Journal of Productivity Analysis*, 20(3), 391-424.

Simar, L., P.W. Wilson (2000) Statistical Inference in Nonparametric Frontier Models: The State of the Art, *Journal of Productivity Analysis*, 13(1), 49–78.

Simar, L., P. Wilson (2001) Testing Restrictions in Nonparametric Efficiency Models, *Communications in Statistics: Simulation & Computation*, 30(1), 159-184.

Simar, L., P.W. Wilson (2002) Non-parametric Tests of Returns to Scale, *European Journal of Operational Research*, 139(1), 115-132.

Simar, L., P. Wilson (2007) Estimation and Inference in Two-stage, Semi-Parametric Models of Production Processes, *Journal of Econometrics*, 136(1), 31-64.

Staat, M., H.H. Bauer, M. Hammerschmidt (2002) Structuring Product-Markets: An Approach Based on Customer Value, *Marketing Theory and Applications*, 13(1), 205-212.

Staat, M., M. Hammerschmidt (2005) Product Performance Evaluation: A Super-Efficiency Model, *International Journal of Business Performance Management*, 7(3), 304-319.

Varian, H. (1988) Revealed Preference with a Subset of Goods, *Journal of Economic Theory*, 46(1), 179-185.

Vassiloglou, M., D. Giokas (1990) A Study of the Relative Efficiency of Bank Branches: An Application of Data Envelopment Analysis, *Journal of the Operational Research Society*, 41(7), 591-597.

Warner, E.J., R.B. Barsky (1995) The Timing and Magnitude of Retail Store Markdowns: Evidence from Weekends and Holidays, *Quarterly Journal of Economics*, 110(2), 321-352.

Wilson, P.W. (1993) Detecting Outliers in Deterministic Nonparametric Frontier Models with Multiple Outputs, *Journal of Business & Economic Statistics*, 11(3), 319-323.

Zhang, Y., 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, *Journal of Productivity Analysis*, 9(3), 187-204.