

Maitotilojen kehittämispotentiaalin arviointi

Timo Sipiläinen¹⁾, Mika Kortelainen²⁾, Sami Ovaska¹⁾ ja Matti Ryhänen³⁾

¹⁾ *MTT Taloustutkimus, Luutnantintie 13, 00410 Helsinki, etunimi.sukunimi@mtt.fi*

²⁾ *Economics, School of Social Sciences, University of Manchester, Oxford Road, UK- Manchester M13 9PL, mika.kortelainen@manchester.ac.uk*

³⁾ *Seinäjoen ammattikorkeakoulu, Maa- ja metsätalouden yksikkö, Ilmajoentie 525, 60800 Ilmajoki, matti.ryhanen@seamk.fi*

Tiivistelmä

Tämän tutkimuksen tarkoituksena oli suomalaisten maitotilojen kehittämispotentiaalin arviointi ja sellaisten tekijöiden etsiminen, jotka selittävät olemassa olevia tuottavuuseroja. Kuten tunnettua, tekniset ja taloudelliset tulokset vaihtelevat merkittävästi tilojen välillä. Vaihtelun olemassa olo viittaa siihen, että tehokkuuden kehittäminen on mahdollista ja että yrittäjien voisi olla mahdollista oppia muiden tilojen parhaista käytännöistä.

Tutkimuksen tutkimusaineistona käytettiin ProAgrian maidontuotannon tulosaineistoa vuodelta 2005. Aineistoa käytettiin, koska siihen sisältyy useita tuotantoa kuvaavia muuttujia, joita ei ole saatavissa muista aineistoista. Ongelmana toisaalta on, että aineisto on ainoastaan yhdeltä vuodelta. Tällöin poikkeavien havaintojen huomioon ottaminen on erityisen tärkeää. Tämä ei koske pelkästään poikkeuksellisen hyviä vaan myös poikkeuksellisen heikkoja tuloksia. Poikkeaviin havaintoihin on kiinnitettävä erityistä huomiota, koska maataloustuotantoon liittyy säänvaihtelujen vuoksi satunnaisuutta. Edelleen monet tuotantoresurssit (mm. maa) ovat laadultaan heterogeenisiä.

Maidontuotannon tehokkuutta ja sen vaihtelua selittäviä tekijöitä tarkastellaan useilla eri menetelmillä. Tutkimuksessa sovelletaan ei-parametrinen DEA (verhokäyrä)menetelmän ohella myös order-m osittaisrintamamenetelmää sekä parametrinen stokastista rintamamenetelmää tehokkuuden mittaamiseen, joita käytettäessä tehdään erilaisia oletuksia tehokkuusrintamaan ja tehottomuuden jakaumaan liittyen.

Edellä mainittuihin tehokkuusmittauksiin yhdistettiin tehokkuuden selitysmallit. Tarkastelussa olivat mukana sekä parametrinen, semiparametrinen että kokonaan ei parametrinen lähestymistapa. Vertailukohtana käytettiin kaksivaiheista DEA - katkaistu regressio –mallia, joka on helppo estimoida mutta jonka käyttöön liittyy eräitä tilastollisia ongelmia. Tämän selitysmalliyhdistelmän lisäksi käytettiin order-m ja tavanomaista regressiota, stokastista tehokkuusvaikutusmallia sekä ehdollisen tehokkuuden selitysmallia. Nämä mallit ovat kukin tilastollisten oletustensa osalta oikeampia kuin ensin mainittu DEA/katkaistu regressio yhdistelmä. Ehdollinen tehokkuuden selitysmalli on kehitetty aivan viime vuosina ja sen etuna on, että panos-tuotostasoihin ja tehokkuuteen vaikuttavien tekijöiden ei tarvitse olettaa olevan toisistaan erillisiä.

Eri mallit tuottavat varsin yhdenmukaisia tuloksia tehokkuusjärjestyksen ja tehokkuuteen vaikuttavien tekijöiden suhteen. Sen sijaan tehokkuuden tasot poikkeavat määritysmenetelmittäin huomattavasti toisistaan.

Tulokset osoittavat myös, että tuottavuuspotentiaalia arvioitaessa tuotanto-olosuhteet on otettava huomioon realististen vertailukohtien saavuttamiseksi. Useimmat tuotantoprosessia kuvaavat indikaattorimuuttujat eivät olleet merkitseviä tuotannon tehokkuuden kannalta. Sen sijaan keskituotos on merkitsevä tehokkuustekijä, mutta korkeimmilla tuotostasoilla se ei enää paranna tehokkuutta. Separoitu- vuusoletuksella on merkitystä maidon rasvapitoisuuden vaikutuksen kannalta. Rasvapitoisuus on merkitsevä tehokkuuteen vaikuttava tekijä kaikissa muissa paitsi ehdollisen tehokkuuden malleissa. Näin ollen maidon korkea rasvapitoisuus ei välttämättä ole yhteydessä tuotannon tehokkuuteen.

Avainsanat: tekninen tehokkuus, DEA, order-m, SFA, parametrinen, ei parametrinen regressio

Introduction

The traditional second stage regression for explaining the differences in DEA efficiency scores has been criticized because of the statistical inconsistencies. Simar and Wilson (2007) have introduced a bootstrapping technique tackling two deficiencies of the traditional two stage regression analysis of DEA efficiency scores: the lack of coherence with the method and data generating process, the ignorance of the serial correlation among estimated efficiencies and of the correlation between the error term of the second stage regression and contextual/non-discretionary variables. On the other hand, Hoff (2007) has compared a set of second stage regression models empirically and concluded that even OLS and Tobit often perform relatively well. A recent comparison also including the Simar and Wilson (SW) approach is provided by Afonso and St. Aubyn (2006). Their estimations suggested very similar results between Tobit and SW approaches. In addition, Banker and Natarajan (2008) have concluded that DEA-based procedures with OLS, truncated regression, or even Tobit estimation in the second stage perform as well as the best of the parametric methods in the estimation of the impact of contextual variables on productivity.

Cazals et al. (2002) have suggested order-m efficiency technique in order to reduce the influence of noise in the nonparametric analysis. Compared to traditional nonparametric DEA and FDH techniques that estimate *full frontier*, order-m procedure estimates *partial frontier*, which allows some observations to be above the frontier thus reducing the effect of outliers. Cazals et al. also showed how one can include contextual/environmental z-variables in partial frontier estimation by conditioning the estimation on the values of z-variables. This so-called conditional efficiency approach was further developed by Daraio and Simar (2005, 2007), who presented a descriptive approach to evaluate the effect of z-variables in the production process. In contrast to two-stage approaches, conditional efficiency avoids the assumption that contextual/environmental variables would only affect on efficiency scores and not on the efficient frontier. In many applications it is therefore a more justified method to examine how various environmental or background variables affect on the production process. Yet, the original conditional efficiency model developed by Cazals et al. (2002) and Daraio and Simar (2005) only allowed continuous environmental variables, even though most applications include also various kinds of categorical exogenous variables. Recently, De Witte and Kortelainen (2009) have developed the method further making it possible to include both continuous and discrete environmental variables and presenting a procedure to examine statistical significance of these variables.

The objective of the study is to compare various methods in the evaluation of efficiency effects. The production takes place in various circumstances. These contextual or environmental variables may constrain or enhance the opportunities for good performance. Therefore, they have to be taken into account when the performance of farms is evaluated. On the other hand, it may be possible to identify management factors or indicators, which may be promoted when we aim at supporting of improvements in the performance of farm managers. For this evaluation it is possible to apply several methods. In this study we apply nonparametric, semiparametric and parametric methods to examine the effect of contextual variables. These methods will be described more in detail later.

Data

The cross-sectional data set from 2005 is used in the study. The data is obtained from the data bank of the Finnish extension service, ProAgria. Originally, close to 2 000 dairy farm observations is covered by the data set. The data are collected during the farm visits mostly for advisory purposes. In order to reduce the effect of good or bad luck on the results we have removed from the sample the farms which have obtained the highest or lowest values with respect to profitability. It has been shown on the basis of bookkeeping data that in a cross-section the differences between well and poorly performing farms are exaggerated when compared to actual long-term differences between farms. Therefore, ten percent of the best and worst performing farms are removed from the data set. After this careful checking of the data 1337 observations are used in the analysis.

Only half of the sample farms are utilised in the final analysis. We sorted the farms according to their identification number and picked up every second farm to the analysis. This procedure was followed because of the high requirements of data processing time in the nonparametric methods.

In the analysis we apply one output, which is energy corrected milk (ECM; standardised milk including 4 % fat and 3.3 % protein). This output measure takes into account not only the quantity but also the quality of milk, i.e., fat and protein content. We chose to use only one output because the

farms are highly specialized in milk. Other products are mainly side products of milk production. In addition, the costs have been allocated to milk production (and to other products) already at the stage of data collection. Thus, only costs related to milk production are recorded.

We considered four input variables: labour (in h), capital (in €), feed (in feed units) and herd variable costs (in €). We have to note that as the analysis constrains to milk production only, it does not directly include the input use of field operations. The only input from crop production is the feed¹. Preliminary regression results showed that the total feed consumption is very closely linked to milk output. This should also be the case since we know that the energy requirement for one kilogram of milk is fairly constant between cows. However, this also raises a question whether the observations of feed consumption have at least partially been calculated on the basis of milk output. Therefore, we dropped this variable from the analysis. Descriptive statistics (mean, standard deviation, minimum and maximum) of the variables are presented in Table 1.

Table 1. Descriptive statistics (n = 668).

	Mean	Std. Dev.	Min	Max
ECM (kg)	204 688	96 877	65 709	874 079
Labour (h)	3 769	1 217	1 373	11 892
Machinery (€)	16 681	14 447	926	133 018
Variable (€)	15 910	9 621	4 455	140 661
Feed (fu)	223 704	110 949	59 004	1 061 646
Cow (n)	25.5	116	7,6	114.6
<u>z-variables</u>				
Fat (%)	4,17	0,31	3,06	5,27
Protein (%)	3,43	0,10	3,14	3,74
Cell count (1000/ltr milk)	167,9	81,0	33,0	877,0
Insemination (n/cow)	2,06	0,51	1	4,3
Milkfu (fu/kg ECM)	0,76	0,06	0,48	1,10
Coefficient of profitability	0,69	0,23	0,26	1,20
Need (% of need)	105	8	43	170
Calvings (n/cow)	2,53	1,51	1,30	4,0

In the data set we have one contextual (z) variable, region, which is expected to capture some of the production environment dependent differences in production possibilities sets. In addition, we have a set of variables that serve as indicators of farm management: the milk yield per cow (classified by quartiles), the fat and protein content of milk (%), the cell count (1 000/ltr milk), the number of insemination (n/cow), the ratio of actual feed consumption and calculated need and the average number of calvings (n/cow).

The farms are grouped to four regions. The location of farms is distributed to different regions as follows: 154 farms are from subsidy areas A and B, 134 from the area C1, 291 from the area C2 and C2P, 89 from the area C3 and C4.

Method

In this study we apply four different methods to explain the factors behind the managerial success. At first a simple two-stage approach is applied to estimate the effects of environmental or managerial variables on efficiency. The two-stage approach calculates efficiency scores with DEA² (using only inputs and outputs) in the first stage and then explains the efficiency scores by z variables using some regression method in the second stage. In the first stage, we estimate standard efficiency scores apply-

¹ As we know, the allocation of costs to various products is challenging. However, we cannot change the data collection process but we have to take the variables as they are measured.

² A detailed description of the DEA model can be found for example in Fried et al. 2008.

ing output oriented DEA both under the assumption of variable (VRS) and constant (CRS) returns to scale, while the second stage applies the truncated regression method by setting a truncation point from below to one. This removes the values of one (efficient farms) from the data set. Thus, the procedure is a combination of nonparametric and parametric approaches.

It has been shown in literature (see Simar and Wilson 2007; Daraio and Simar 2007) that the usual two-stage approach has certain statistical shortcomings. The main problem in the two-stage efficiency explanation models is that DEA scores are serially correlated and the distribution of efficiency scores is limited above or below to one. The serial correlation problem can be alleviated by the bootstrapping method developed by Simar and Wilson (2007). However, as SW approach is fairly complicated and heavy with respect to computational requirements, many applications are still based on simpler methods that do not account for the serial correlation. Interestingly, in practical applications these simple approaches have often produced very similar results for contextual/environmental variables (see Hoff, 2007; Banker and Natarajan, 2008). Because of this, we also apply a (simpler) two-stage approach that does not account serial correlation. However, instead of Tobit regression, which has been shown to be more sensitive to the statistical problems, we use truncated regression.

Another option is to start from the order- m unconditional efficiencies, which are estimated on the basis of partial frontiers. In this partial frontier framework, the efficiency scores are not serially correlated in the way they are in the conventional DEA framework. Another virtue of this approach is that the efficiency scores are not restricted either to be smaller than or equal to (or bigger than or equal to) one because super-efficiency is allowed. This implies that the ordinary least squares method is an asymptotically valid method for estimating the effects of contextual/environmental variables on efficiency and for statistical inference. Like previous two-stage approaches, this method is also a combination of nonparametric and parametric methods. We will use this alternative two-stage approach in our empirical application, and explain it further below in the context of conditional efficiency method.

It is important to note that all two-stage approaches implicitly assume separability between first-stage variables (inputs and outputs) and factors that are included only in the second stage. This separability assumption means that z -variables are allowed only to affect efficiency scores, not to the frontier or efficient levels of inputs and outputs. If z -variables are also important explanatory factor for production technology (or should be included in the first stage), the results given by two-stage approach are biased.

To address the separability issue, one possibility is to use fully parametric stochastic frontier analysis, where it is easy to include environmental variables either as explanatory variables for production technology or for inefficiency effect (see Battese and Coelli 1995). However, in this case we need assumptions both for the functional form and inefficiency distribution. This makes the estimation procedure easier but simultaneously restricts the flexibility of the estimation of relationships between specific factors. In our case, we apply a simple Cobb-Douglas production function and stochastic frontier with the technical inefficiency effect model. Besides inputs, one can include also environmental z -variables into production function estimation if needed. The parameters of the model are estimated by the method of maximum likelihood. We applied the computer program Nlogit3 (Limdep).

While stochastic frontier model allows one to include z variables either as a part of production technology or a part of inefficiency term, the challenge is to decide how to model z variables correctly. Moreover, SFA is based on strong parametric assumptions, which can bias the results even if z variables would have been modelled properly. Because of this, it is preferable to use an approach that avoids strong parametric assumptions, but also allows z variables to affect both on production technology and on efficiency scores. One promising method in this respect is the conditional efficiency method first suggested by Cazals et al. (2002) and Daraio and Simar (2005) and further developed by De Witte and Kortelainen (2009). This fully nonparametric approach provides very useful qualities: 1) no assumptions of specific functional forms for the production function are needed, 2) no separability assumption is required and additionally 3) both continuous and categorical variables can be applied. The last virtue is the extension provided by De Witte and Kortelainen (2009).

In the conditional efficiency analysis (Cazals et al. 2002; Daraio and Simar 2005, 2007) z -variables are directly included in the nonparametric model. As the first step we estimate unconditional efficiency using order- m estimator, which can be seen as the "robust" version of FDH (or DEA). When originally FDH provides non-stochastic estimates for the best practise observations (defines undominated units which produce the highest output vector by given input vector (output orientation)),

the order- m efficiency approach describes the production process using a probabilistic formulation. The probabilistic formulation is based on the examination of the probability that the observation under evaluation is dominated. Basically, we estimate a partial frontier, which allows some observations to lie above the frontier (i.e. "superefficient").

In line with the probabilistic formulation, Cazals et al. (2002) proposed a conditional efficiency approach with respect to some external factors. Daraio and Simar (2005) generalized the idea to a practical methodology of a multivariate case. In the approach the production is conditioned to a given values of z .

The conditional efficiency approach resembles nonparametric (like kernel) methods of regression and density estimation by assuming continuous data. This is extended by DeWitte and Kortelainen (2009). They apply smoothing of discrete variables in a particular manner by using novel kernel methods suggested by Li and Racine (2003, 2007, 2008). Continuous, discrete unordered and ordered variables are treated differently. Standard multivariate (or univariate) product kernel functions are applied for all these data categories (Li and Racine 2007). Multiplying the multivariate kernel functions provides a generalized product kernel function.

The bandwidth choice is the crucial stage of the nonparametric kernel estimation. In our case we apply a data-driven bandwidth choice method adopted from Hall et al. (2004) and Badin et al. (2009), which estimates the bandwidth by the least squares cross-validation method. The only difference from Hall et al. is that in the optimal data-driven bandwidth for conditional efficiency, the reference set is reduced to $x_i \leq x$. Because of this, bandwidths are both observation and z -variable specific.

To examine the effect of z variables in conditional efficiency setting, we follow De Witte and Kortelainen (2009) and explain the ratio of conditional and unconditional efficiency scores by z using local linear regression that smooths both continuous and discrete variables (see Li and Racine, 2004). In addition, we use nonparametric bootstrap procedures to examine whether z variables have a significant effect on production process. Statistical inference is based on the nonparametric significance test by Racine (1997) and Racine et al. (2006), which can be seen as nonparametric counterparts of standard t-tests of OLS regression. The bootstrap tests show whether individual variables have any effect on the fit of the model, i.e. the difference between conditional and unconditional efficiencies.

Results

Efficiency estimates of both data envelopment and stochastic frontier analysis suggest considerable inefficiency of production – even the median values indicate that output could be increased by 40 – 50 %, given inputs (Table 2). In addition, the deviations are large, although we have removed farms from the upper and lower end of the performance (profitability) distribution. The results also show the dependence on the method: in unconditional order- m efficiencies the median inefficiency is less than 10 %. The results also show that the factors like the location, average output class and the fat content of milk considerably affect the efficiency score.

In spite of large differences in efficiency levels, various efficiency scores highly correlate with each other. As expected, the correlation between CRS and VRS scores is close to one but the correlation is very high also between stochastic parametric frontier and unconditional order- m efficiencies (0.740 – 0.814).

Table 2. Efficiency estimates.

	Mean	Median	Minimum	Maximum	St. Dev.
DEA_CRS	1.478	1.430	1.000	5.250	0.238
DEA_VRS	1.416	1.370	1.000	5.080	0.233
Stochastic frontier	1.633	1.485	1.076	3.807	0.362
Unconditional order- m ($m=80$)	1.098	1.029	0.576	3.394	0.230
Conditional order- m ($m=80$)	1.083	1.002	0.878	2.926	0.175

As we described in the method section, we used four approaches in the estimation of efficiency effects of the location and management variables. These were the two stage approach (nonparametric – parametric) in which the output oriented efficiency scores from the DEA estimation were regressed by factors probably affecting the efficiency. The truncated regression without bootstrapping is applied in

the second stage analysis although several authors have criticized this approach because of certain limitations caused by the known deficits. The truncation takes place at one. Thus, the efficient farms taking the value of one are removed from the analysis. At first we test all measured explanatory variables. The clearly insignificant variables are removed and thus we end up to basically three significant variables (the fat content of milk, the region (measured by regional dummies subsidy regions A and B representing the base) and the average milk yield class (classified according to quartiles the lowest one being the base). The result shows that all these variables are significant contributors to technical efficiency: technical inefficiency decreases from south to north; it decreases by fat content of milk and it decreases by increasing average milk yield per cow. The maximum likelihood estimates of regression coefficients are approximately similar independently on the assumption of constant or variable returns to scale

The second stage regression on order-m output efficiency scores is regressed by the same above mentioned variables. The OLS model is significant but its explanatory power is relatively low. Only 10 percent of the variation can be explained by the chosen variables. In addition, the fat content of milk is a less significant contributor to technical efficiency than in DEA/truncated regression models. If we apply by the cow number weighted regression, the fat content of milk is no more significant at 10 percent risk level. The results also show that the marginal effects at the mean are larger in the two stage model of DEA/truncated regression compared to the model of order-m unconditional efficiency/ordinary least squares.

As the third, fully parametric method, we applied a parametric stochastic frontier production function (a Cobb-Douglas function) with technical efficiency effects to our sample of Finnish dairy farms. In the analysis, we used three inputs - labour, machinery and variable inputs - and one output - energy corrected milk. Inputs and the output are measured in logarithmic terms. All three inputs were significant contributors to production with expected signs, and returns to scale were less than one (0.94), indicating decreasing returns to scale in the sample. We should note that in this model the regional dummies are included to the production function part of the equation. The results show that output obtained from the inputs was significantly lower in the Central and especially in Northern Finland than in Southern Finland. In other words, the result shows that production conditions become harder for milk production when we move from south to north.

In the model, technical inefficiency is statistically significant. The technical efficiency effect model shows that only the fat content of milk and the average milk production per cow (classified) are significant contributors to efficiency when regional dummies are estimated as a part of the production function. The increase in both of these variables reduces inefficiency. Instead, the cell count in milk, the need for versus the actual use of feed, the number of inseminations per cow, or the average number of calvings per cow do not have a significant influence on technical efficiency. The inefficiency reducing effect is increasing when the average milk yield increases but at a diminishing rate.

Lastly we report the estimates from z-variables from the conditional efficiency model. According to the bootstrapping tests, the nonparametric conditional efficiency estimation shows that the same factors are approximately equally important in this approach as in the other estimation methods, except the fat content. In this case mean yield has not been measured as a dummy variable but as an ordered variable, which is growing from one to four. The same applies to region but the region is measured as an unordered classified variable with four classes. In both cases kernel smoothing is applied. The fat content is instead treated as an ordinary continuous variable.

The results also show that the bandwidth of the discrete variables is approximately similar and relatively small. The bandwidth of the continuous variable fat content is on average large. This follows from the fact that the bandwidths for some farms of this variable are very large.

Local linear regression can also provide us marginal effects of the contextual/environmental z-variables on performance. However, in this context marginal effects would be more difficult to interpret, as the dependent variable is based on the ratio of conditional and unconditional efficiency score. The general trends are similar to the other methods: the effect of mean yield and fat content of milk are positive with respect to efficiency. Instead, when we move from south to north the efficiency decreases. The evaluation point has some effect on the mean yield per cow and the regional effect.

Discussion

Different approaches suggest very different level of average inefficiency of production. However, the correlation of efficiency scores between methods is high indicating that the order is fairly consistent although the levels of inefficiency may vary a lot. Therefore, inefficiency levels should be interpreted with care when improvement potential is assessed. The improvement potential is clearly affected by the assumptions made about the reference set (potential peers) when the method is chosen.

Similarities in the order are confirmed by the fact that different methods for evaluation of the factors affecting efficiency suggest that similar factors are significant and also the significance levels are similar. In the semiparametric models also the marginal effects closely resemble each other. The only major difference is the insignificance of fat content as a z-variable in the conditional efficiency approach. This approach is the only one that does not assume separability between input levels and z-variables. The result suggests that the separability may not be a valid assumption.

The results show that there are mostly minor differences between methods in this large sample, when inefficiency effects are investigated. This coincides with the observations of earlier studies (Afonso and St. Aubyn 2006; Hoff 2007; Banker and Natarajan 2008). Small differences between methods may also partially be related to the fact that we removed part of the extreme variables. In spite of this, the variation in efficiency is still large.

References

- Afonso, A. & St. Aubyn, M.** (2006) Cross-country efficiency of secondary education provision: A semi-parametric analysis with non-discretionary inputs. *Economic Modelling* 23(3): 476-491.
- Badin, L., Daraio, C. & Simar, L.** (2009) Optimal bandwidth selection for conditional efficiency measures: A data-driven approach. *European Journal of Operational Research*, in press.
- Battese, G.E. & Coelli, T.J.** (1995) A model of technical inefficiency effects in a stochastic frontier production function for panel data. *Empirical Economics* 20: 325-332.
- Banker, R.D. & Natarajan, R.** (2008) Evaluating Contextual Variables Affecting Productivity Using Data Envelopment Analysis. *Operations Research* 56(1): 48-58.
- Cazals, C., Florens, J.P. & Simar, L.** (2002) Nonparametric frontier estimation: A robust approach. *Journal of Econometrics* 106(1): 1-25.
- Daraio, C. & Simar, L.** (2005) Introducing environmental variables in nonparametric frontier models: A probabilistic approach. *Journal of Productivity Analysis* 24(1): 93-121.
- Daraio, C. & Simar, L.** (2007) Advanced robust and nonparametric methods in efficiency analysis. *Methodology and applications. Series: Studies in Productivity and Efficiency.* Springer.
- De Witte, K. & Kortelainen, M.** (2009) Blaming the exogeneous environment? Conditional efficiency estimation with continuous and discrete environmental variables. Paper presented in EWEPA XI, Pisa, 23.-26.6.2009
- Fried, H.O., Lovell, C.A.K. & Schmidt, S.S.** (2008) Efficiency and Productivity, in: H. Fried, C.A.K. Lovell, S. Schmidt (eds) *The Measurement of Productive Efficiency and Productivity Change*, New York, Oxford University Press. p. 3-91.
- Hall, P.J., Racine, J.S. & Li Q.** (2004) Cross-validation and the estimation conditional probability densities. *Journal of the American Statistical Association* 99(468). 1015-1026.
- Hoff, A.** (2007) Second stage DEA: Comparison of approaches for modelling the DEA score. *European Journal of Operational Research* 181: 425-435.
- Kumbhakar, S. & Hjalmarsson, L.** (1993) Technical efficiency and technical progress in Swedish dairy farms. In Fried, H.O., Lovell, C.A.K. & Schmidt, S.S. (eds) *The measurement of productive efficiency – techniques and applications.* Oxford University Press. p. 256-270.
- Li, Q. & Racine, J.S.** (2003) Nonparametric estimation of distributions with categorical and continuous data. *Journal of Multivariate Analysis* 86(2): 266-292.
- Li, Q. & Racine, J.S.** (2004) Cross-validated local linear nonparametric regression. *Statistica Sinica* 14(2): 485-512.
- Li, Q. & Racine, J.S.** (2007) *Nonparametric Econometrics: Theory and Practice.* Princeton University Press.
- Li, Q. & Racine, J.S.** (2008) Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data. *Journal of Business and Economic Statistics* 26(4): 423-434.
- Racine, J.S.** (1997) Consistent significance testing for nonparametric regression. *Journal of Business and Economic Statistics* 15(3): 369-379.
- Racine, J.S., Hart, J. & Li, Q.** (2006) Testing the significance of categorical predictor variables in nonparametric regression models. *Econometric Reviews* 25(4): 523-544.
- Simar, L. & Wilson, P.W.** (2007) Estimation and inference in two-stage, semiparametric models of production processes. *Journal of Econometrics* 136: 31-64.