Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 266 (2025) 317–324

Procedia Computer Science

www.elsevier.com/locate/procedia

12th International Conference on Information Technology and Quantitative Management (ITQM 2025)

Method for choosing models to estimate efficiency in Brazilian sanitation sector

Pedro Henrique de Matos Araujo [a*], Igor Pinheiro de Araújo Costa [a,b,c], Miguel Ângelo Lellis Moreira [c], Luiz Paulo Fávero [a], and Marcos dos Santos [d]

[a]*Escola Superior de Agricultrua de Luiz de Queiroz da Universidade de São Paulo – USP, Piracicaba - SP, 13418-900, Brazil*
[b]*Naval Systems Analysis Center (CASNAV), Rio de Janeiro, RJ 20091-000, Brazil*
[c]*Fluminense Federal University (UFF), Niterói, RJ 24210-240, Brazil*
[d]*Military Institute of Engineering, Urca, RJ 22290-270, Brazil*

Abstract

Estimating the efficiency gains achieved by service providers and sharing these gains with users has become imperative with the advent of the Brazilian Legal Framework of Sanitation. In this context, the objective of this work was to evaluate the methods for estimating efficiency applied to the basic sanitation sector, especially the benchmarking methods, and to propose criteria for choosing the most appropriate model. The operational and financial information used was extracted from the "Sistema Nacional de Informações sobre Saneamento" (SNIS) for the year 2022. Models of Data Envelopment Analysis (DEA), Super-Efficiency DEA (SDEA) and Stochastic Frontier Analysis (SFA) were estimate for the providers available in the database. To ensure comparability among the analyzed providers, they were segregate into large, medium and small providers. Small providers were not discussed in depth, due the SFA model could not be adequately estimate. Finally, the model selection method proposed in this work was apply, verifying the statistical equality between the efficiencies estimated by DEA and SFA models, with preference for the DEA model for large and medium providers, considering the basic premise of minimum extrapolation. Therefore, this work seeks to contribute to the discussion on the suitability of the studied models for estimating the operational efficiency of water supply and sanitation service providers, directly contributing to the implementation of more equitable and sustainable policies in the sector.

Keywords: Data Envelopment Analysis; Stochastic Frontier Analysis; Super-Efficiency Data Envelopment Analysis; Model Selection.

1. Introduction

Natural monopolies arise in markets where the entry costs for supplying a product are excessively high, while marginal costs are relatively low. Commonly, such markets are observed in public service and infrastructure providers, such as gas, electricity, telecommunications and sanitation companies[17].

10.1016/j.procs.2025.08.040

Due to the monopolistic profit-maximizing behavior, sanitation companies have incentives to operate outside the efficient level of production. An alternative to limit this behavior is government regulation for price determination, ensuring that the production level is sufficient to meet the population's need and that company does not incur losses[17].

In the light of the Brazilian New Legal Framework of Sanitation, defined by Federal Law 14.026/2020, the sector aims to universalize water supply and sewerage services by the year 2033, observing quality and efficiency standards in service provision. Furthermore, the aforementioned law defines as a fundamental principle the improvement of quality with efficiency gains and consequent cost reduction for users. Therefore, considering the operational efficiency of sanitation providers at the price determination by regulators has become imperative for the Brazilian context.

To address this issue, some regulatory agencies in the Brazilian sanitation sector has adopted the so-called "Fator X", which aims to reduce prices by a defined percentage[6].

The "Fator X" can be define in four main ways: arbitrary value (*ad hoc*); discounted cash flow; historical index; and benchmarking. As pointed out by Shleifer[15], in natural monopolies characterized by cost-of-service regulation, the best way to ensure operational efficiency of costs is through benchmarking. The benchmarking methods for determining the "Fator X" can also be divided into four groups, which can be used individually or in combination: Corrected Ordinary Least Squares (COLS); Stochastic Frontier Analysis (SFA); Data Envelopment Analysis (DEA) and engineering methods[6].

In this context, the present work focuses on analyzing the application of SFA and DEA methods to the sanitation sector, as these are more robust according to Bragança and Camacho[6]; in addition, a Super-Efficiency Envelopment Analysis (SDEA) was also estimated, whose initial objective has to rank companies classified as efficient through classical DEA modeling[4].

2. Materials and Methods

2.1. Classical Data Envelopment Analysis (DEA)

The DEA model was initially proposed by Charnes[9] as a method for measuring the efficiency of so-called Decision-Making Units (DMU). Production theory is fundamental to DEA literature and assumes that all companies, or DMUs, have access to the same technology. Under the assumption of common technology, the principle of minimum extrapolation is used to estimate an approximation of the real PPF. This principle ensures a conservative estimate of the PPF, as it creates the smallest possible convex set that includes all DMUs used in its construction[4].

Therefore, to estimate the PPF was chosen input-oriented models – minimization of the input with that same level of output – given that in the context of the sanitation sector, outputs are often limited, either by the concession area or by the population's demand. Thus, the DEA model constructed to measure the efficiency of sanitation sector providers, more specifically water supply and sewerage services, was used the assumption of variable returns of scale, given the possibilities of economies and diseconomies of scale[17].

Additionally, the standard minimization[10] was altered to prevent corner solutions, i.e., to prevent the model from considering the non-supply of one of the outputs as efficient, as show in equation (1):

$$\begin{aligned} &\min_{E,\lambda} E \\ \text{subject to} \quad & Ex_0^k \;\geq\; \sum_{j=1}^{J} \lambda_j x_j^k\,, \quad k = 1, \dots, m \\ & y_0^l \leq \sum_{j=1}^{J} \lambda_j y_j^l\,, \qquad l = 1, \dots, n \\ & \prod_{j=1}^{J} \lambda_j \;\geq\; 0 \end{aligned} \tag{1}$$

where, "y": is the output matrix; "x": is the input matrix; "i": is the company subject to maximization; "j": the set of companies considered in the benchmarking; "u": are the output weights; "v": are the input weight. "E": is a scalar representing the efficiency of company "I", according to Farrell's efficiency definition[11]; "λ": is a vector of constants; "m": is the number of inputs; "n": is the number of outputs.

For the estimated DEA model, Exploration Costs were used as inputs, and the quantities of Households Served with Water and Sewerage Services as well as the Volumes of Consumed Water and Treated Sewage were used as outputs.

2.2. Super-Efficiency Data Envelopment Analysis (SDEA)

The SDEA model was proposed by Andersen e Petersen[2] with the objective of differentiating companies that, according to the classical DEA model, are on the PPF. Consequently, was identified the applicability of this approach for regulation and contract elaboration based on DEA model[3].

In essence, the so-called super-efficiency models are based on estimating the efficiency of company "i" given the technological set based on the observations of all agents, except the “i-th”, as show in equation (2):

$$\begin{aligned} &\min_{E,\lambda} E \\ \text{subject to} \quad &Ex_0^k \geq \sum_{j \neq i} \lambda_j x_j^k, \qquad k = 1, \dots, m \\ &y_0^l \leq \sum_{j \neq i} \lambda_j y_j^l, \qquad l = 1, \dots, n \\ &\prod_{j \neq i} \lambda_j \geq 0 \end{aligned} \tag{2}$$

Given the great similarity between eq. (1) and eq. (2), the DEA and SDEA models, the efficiency values below one remains unchanged. Finally, the SDEA model constructed in this work used the minimization of eq. (2).

2.3. Stochastic Frontier Analysis (SFA)

The stochastic approach to production functions was proposed, independently, by Aigner[1] and Meeusen and Van Den Broeck[12]. The central idea of this approach was the addition of an error component to the model to capture statistical noise. Such noise is due to the omission of relevant explanatory variables, as well as errors associated with the chosen distribution.

A simple transformation allows the estimation of the Cobb-Douglas PPF by the stochastic approach. However, the result of this transformation presents a model called multi-input/single-output. So, to ensure the comparability of this model with those discussed earlier, it was necessary to invert the equation so that it assumes the single-input/multi-output form, oriented to production costs, as described by Bogetoft e Otto[4] and explained in eq. (3):

$$\begin{aligned} -x_j &= c(y_j; \alpha) - v_j - u_j \\ v_j &\sim N(0, \sigma_v^2), u_j \sim N_+(0, \sigma_u^2), \qquad j = 1, \dots, J \end{aligned} \tag{3}$$

where, “y”: is, in this equation, the matrix of independent variables, with dimensions “j×m”; “x”: is the dependent variable; “j”: the set of companies considered in the benchmarking; “m”: is the number of inputs; and “u”: is the firm’s inefficiency; “v”: is the stochastic error.

Regarding the specific efficiencies of companies in the sanitation sector in the SFA model, they are given by eq. (4) and, consequently, will be estimated by the maximum likelihood estimator as described by Bogetoft e Otto[4] and explained in eq. (4):

$$TE_i\,(x_i, y_i) = 1 - \frac{\hat{u}_i}{f(x_i, \hat{\beta})} \tag{4}$$

where, “TE”: is the technical efficiency of company “i”; and “$\hat{u}_i$”: is the estimated technical inefficiency for company “i”.

2.4. Proposal for Choosing Between DEA, SDEA and SFA Approaches

For choosing between the aforementioned models, it is proposed to evaluate the paired estimates of efficiency via DEA and SFA through the Student's t-test proposed by Student[16], in case of normality, or the Wilcoxon signed-rank test proposed by Wilcoxon[18] in case of non-normality. For normality verification was used the Shapiro-Wilk test proposed by Shapiro e Wilk[14].

If the tests indicate the central measure of the differences is equal to zero, at the significance level of 5%, it is suggested to choose the DEA approaches as they are more conservative in estimating efficiencies, specifically the SDEA model for formulating incentives for companies already efficient according to the classical DEA model. In case of rejection of the null hypothesis of differences between the estimates, it is suggested to adopt the stochastic approach, due to its greater robustness to extreme values.

2.5. Information Used for Model Estimation

For model estimation, cross-sectional data for the year 2022 with financial and operational information from companies in the sanitation sector will be used. This information was obtained from the "Sistema Nacional de Informações sobre Saneamento" (SNIS), an open data source. In addition to the variables returned by default for any queries made in SNIS, the database includes variables described in Table 1.

Table 1. Selected Variables from SNIS

Variable Code	Description
AG003	Quantity of households actively connected to water services
AG010	Volume of consumed water
ES003	Quantity of households actively connected to sewerage services
ES006	Volume of treated sewage
FN015	Exploration costs
IN009	Hydrometric index

For the purpose of model construction, only observations with hydrometric level above 70% were maintained, so that the volumes obtained are mostly real volumes and not estimated. Additionally, for this analysis, only companies that simultaneously offer water supply and sanitation services were considered to ensure comparability between the evaluated companies.

To maintain comparability between companies analysed for the benchmarking process, they were divided according to the number of households served with water services. Due to the large dispersion of the information, the arbitrary percentages of 5% e 0.5% of the number of connections of SABESP – the largest provider in the database – were used as cut-off points for categorizing companies into large, medium and small providers. The number of companies and the criteria for categorization were summarized in Table 2:

Table 2. Details of Company Size Categorization.

Variable Code	Criteria	Nº of Companies
Large Company	More than 644,898 water connections	15
Medium Company	Between 64,489 and 644,898 water connections	63
Small Company	Less than 64,489 water connections	389

3. Results and Discussion

For the discussion of the results, it was decided to evaluate paired algorithms that have comparable results. That said, the efficiency estimated by DEA was evaluated together with that estimated by SFA, and the ranking of efficiencies from the SDEA model with SFA. This is due to the great similarity between the DEA and SDEA models, given that only companies considered efficient in the DEA model have their estimated efficiency altered when compared to the SDEA model.

Specifically for the small companies, the SFA model found a negative lambda, which resulted in estimated efficiencies above one. This may be an indication of the absence of inefficiency in the model or problems in the base used for estimation, as the second is more likely, given the self-declaratory nature of SNIS, it was decided to remove these observations from the analysis.

3.1. Paired Analysis- DEA and SFA

A trend was identified for the values estimated by SFA to be higher than those estimated by DEA for large companies – average difference of 1.4 p.p. – while for the medium companies, the SFA model presented lower efficiency scores – average difference of 1.1 p.p. – than the DEA model.

Nevertheless, it is noteworthy that the SFA model, being a probabilistic model, has an error component. It was found that for the large companies, more than 99.99% of the variability detected by the model was due to inefficiencies, consequently less than 0.01% was due to the random error component. On the other hand, for the medium-size companies, this value was about 89.22%, implying about 10.72% of variability was due to random error. These values may result from the sample size, as the group of large providers is small compared to medium-size companies.

From Figure 1, a strong correlation between the estimated efficiencies scores is noted. Large providers presented a correlation of 72.28% between efficiencies; while medium providers presented a correlation of 89.55%. For both companies' sizes, the distributions of points fit well to the diagonals of the graphs. Specifically for medium companies, scores above 0.8 estimates by DEA obtained counterparts of lower efficiency in the SFA model.

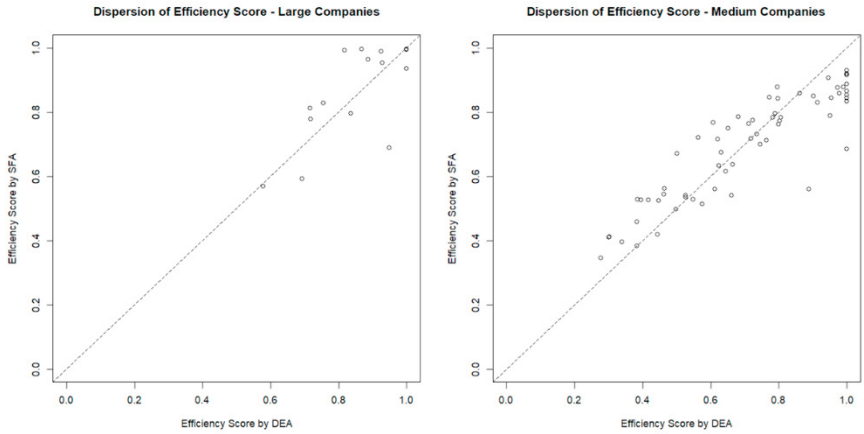

Figure 1. Dispersion of Efficiency Scores – DEA and SFA.

3.2. Paired Analysis- SDEA and SFA

Regarding the ranking of companies by efficiency scores, there were differences for some companies, both for large and medium providers. A possible explanation for this is the SFA model's ability to extrapolate data, allowing the estimation of the PPF in regions where there is few or none data. This is a weakness identify in DEA, as only three large providers were considered efficient, thus the PPF estimated by DEA uses only three companies as a reference.

When evaluating the dispersion of the ranking by SDEA and by SFA, it is noted from Figure 2 that for large companies, there was little adherence to the diagonal of the graph, which is not true for medium providers. This difference in perception may be due to the scale of the graph, and the small amount of data, so when evaluating the correlation of positions, it is noted that the estimated correlation between positions was 61.43% for large companies, and 89.53% for medium companies; correlations very close to those demonstrated by DEA and SFA models. Finally, it is evident that the SDEA and SFA models presented some level of similarity, especially for medium providers. This fact may be due to the sample size, as already discussed.

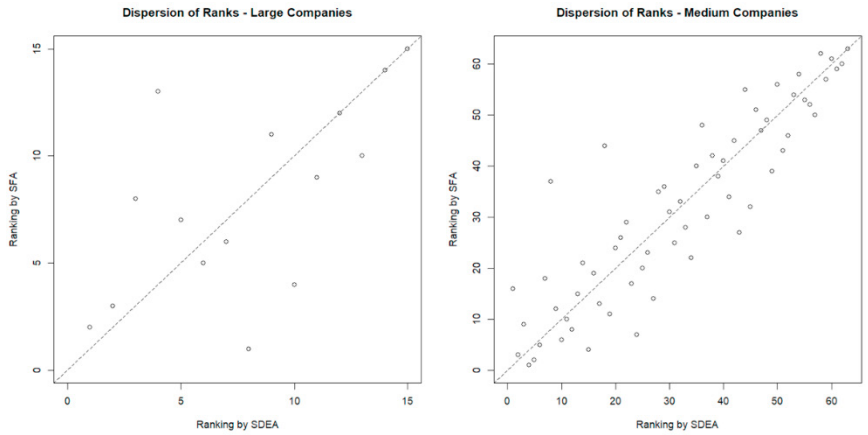

Figure 2. Dispersion of Ranks – DEA and SFA.

3.3. Joint Analysis of DEA, SDEA and SFA

In general, the DEA, SDEA and SFA models presented correlated and satisfactory behavior for large and medium companies in the sanitation sector. As mentioned, the SFA model estimated a negative lambda for small companies, which made their analysis via SFA unfeasible, thus the deterministic approach of DEA is preferable; however, it is imperative to reinforce that there may be problems or inconsistent data in SNIS for small providers due to the self-declaratory nature of the database.

When evaluating efficiencies from the perspective of efficiency incentives, some strengths and weakness of each model stands out. For the DEA model, the assumption of minimum extrapolation is a strength as it ensures that companies are operating at their maximum efficiency; however, there is little room for efficiency incentives in providers already efficient, making the use of other models necessary. Nevertheless, the model is sensitive to extreme values and, consequently, has limitations in considering values with dispersed efficiency scores distributions.

In this sense, the ranking via the SDEA approach has the advantage of excluding the company from its own benchmarking, ensuring the so-called empirical production frontier. Additionally, according to Bogetoft[3], this approach allows the elaboration of rational efficiency incentives through approaches similar to DEA. As negative points, we have, as already discussed for the DEA model, sensitivity to extreme values.

For the SFA models, we see as positive points, due to the insertion of the statistical noise component, the possibility of extrapolation and less sensitivity to extreme values, ensuring a more homogeneous distribution for the estimated efficiencies. Also, when compared to the classical DEA model, it presented a greater field for agency action regarding efficiency incentive mechanisms. However, as stated in this work, some assumptions of the SFA model are more likely to be violated depending on the dataset used, as occurred for small providers. Nevertheless, statistical approaches in small samples, as in the case of large providers, should be evaluated with caution.

3.4. Analysis of Approaches from a Managerial Perspective

Given what has been discussed above, and due to their distinct natures, each method presents its own advantages and limitations, impacting the analysis of inefficiencies, from a managerial perspective, in different ways.

Regarding the identification of inefficiencies, DEA, being a non-parametric technique based on the principle of minimum extrapolation, allows measuring inefficiencies by comparing actual production with the minimum efficient frontier possible. This type of analysis facilitates the identification of possible managerial improvement points for service providers, as it assumes the maximum possible efficiency for each provider. On the other hand, SFA, by incorporating a stochastic error component, allows distinguishing managerial inefficiencies from those resulting from factors external to the provider's control. This provides a broader view of the causes of inefficiencies, also distinguishing which inefficiencies are manageable.

Regarding sensitivity to unobserved variables, the DEA model tends to be more susceptible to the input and output variables used, which can result in significant variations in the results depending on the variables considered. For managers, this offers greater flexibility, allowing performance to be modeled based on different evaluation criteria. In

contrast, the SFA model is less sensitive to omissions of specific variables, as part of the stochastic error absorbs the effects of these variables. From a management perspective, this characteristic makes SFA results more stable over time and less influenced by omitted variables.

In terms of applicability in managerial planning, DEA provides a comparative analysis between providers, allowing the identification of best practices within the sector. On the other hand, SFA enables a more generalist evaluation of efficiency, facilitating the definition of more realistic goals by separating manageable factors from external factors, which are not under the direct control of providers.

In general, DEA and SFA offer complementary approaches for efficiency analysis in the managerial context. While DEA focuses on comparative inefficiencies and provides flexibility in variable selection, SFA presents a more detailed analysis of the causes of inefficiencies, separating managerial influences from external factors. The choice between models should consider the specific needs of the management process, as both provide valuable information for decision-making and efficient resource allocation.

3.5. Model Selection

As already mentioned, it is suggested as a criterion for model adoption the statistical similarity between the estimated score efficiencies. In case there is no statistically significant difference between the estimated values, it is suggested to adopt the DEA approach as it is more conservative, otherwise it is suggested to adopt the stochastic approach due to its robustness to extreme values.

For large providers, the null hypothesis of the normality test was not rejected at a significance level of 5%; the paired test was proceeded under the assumption of normality to verify if there is a statistically significant difference between the estimated efficiencies, in which the null hypothesis was again not rejected at a significance level of 5%. It is suggested for large providers to adopt the classical DEA approach, or SDEA for calculating incentives for providers already efficient in the classical DEA model.

For medium providers, the null hypothesis of the normality test was rejected at a significance level of 5%; the non-parametric paired test was proceeded in which the null hypothesis was not rejected at a significance level of 5%. It is suggested for medium providers, as well as for large providers, to adopt the classical DEA approach, or SDEA for calculating incentives for providers already efficient in the classical DEA model.

4. Final Considerations

This work evaluated the efficiency estimation methods in the context of Brazilian basic sanitation and proposed a systematic for choosing the most appropriate model.

As discussed, a high correlation between the estimated score efficiencies via DEA and SFA was identified. It is noted that the deterministic approach proved to be more robust and less susceptible to having its assumptions violated, while the probabilistic approach proved to be more resilient to extreme values, but with more rigid assumptions.

Regarding the proposed ranking, a high level of similarity between the rankings found was identified, mainly in relation to medium providers. It was inferred that the cause of this dissimilarity for large providers is due to the small sample size.

In general, the analyzed models behaved satisfactorily for large and medium providers in the sanitation sector. As already mentioned, the SFA model estimated a negative lambda for small providers, thus the deterministic approach of DEA is preferable in this case.

The model selection methodology proposed in this work, based on statistical similarity and widely discussed characteristics of the proposed models, indicated the DEA model as the preferred approach, as the estimated efficiencies did not present a statistically significant difference at a significance level of 5%.

Nevertheless, the model chosen by the proposed method may not be ideal depending on the type of analysis desired. This caveat is justified, as the DEA model, being more sensitive to the inclusion and exclusion of variables in the model and having as an assumption the maximization of each provider's efficiency, is more suitable for evaluating manageable inefficiencies and for identifying best practices in the sector. On the other hand, the SFA model, being less sensitive to the exclusion and inclusion of variables, as its error component is capable of absorbing part of the effects of omitted variables, may be more suitable for establishing efficiency goals over time; also, the SFA model is more sensitive to non-manageable inefficiencies, thus enabling the identification of these inefficiencies.

Finally, for future work, it is suggested to treat a joint approach of probabilistic and deterministic models, as their strengths and weaknesses may be complementary.

References

[1] Aigner, D.; Lovell, C. K.; Schmidt, P. 1977. Formulation and estimation of stochastic frontier production function models. Journal of econometrics 6:21-37.

[2] Andersen, P.; Petersen, N.C. 1993. A procedure for ranking efficient units in data envelopment analysis. Management Science 39:1261–1264.

[3] Bogetoft, P. 1994. Incentive Efficient Production Frontiers: An Agency Perspective on DEA. Management Science 40:959–968.

[4] Bogetoft, P.; Otto, L. 2010. Benchmarking with DEA, SFA, and R. Vol. 157. Springer Science & Business Media. New York, NY, USA.

[5] Bogetoft, P.; Otto, L. 2024. Benchmarking with DEA and SFA. R package version 0.32.

[6] Bragança, G. G. F. D.; Camacho, F. T. 2012. Uma nota sobre o repasse de ganhos de produtividade em setores de infraestrutura no Brasil (Fator X). Repositório IPEA. Available at: https://repositorio.ipea.gov.br/handle/11058/5424.

[7] Brasil. 2020. Lei nº 14.026, 15 de julho de 2020. Estabelece as diretrizes nacionais para o saneamento básico. Diário Oficial da União.

[8] Brasil. Sistema Nacional de Informações sobre Saneamento. SNIS – Série Histórica. Site institucional. Available at: http://app4.mdr.gov.br/serieHistorica/.

[9] Charnes, A; Cooper, W.W.; Rhodes, E. 1978. Measuring the efficiency of decision making units. European Journal of Operational Research 2:429–444.

[10] Coelli, T. J.; Rao, D. S. P.; O'Donnell, C. J.; Battese, G. E. 2005. An introduction to efficiency and productivity analysis. Springer Science & Business Media. New York, NY, USA.

[11] Farrell, M. J. 1957. The Measurement of Productive Efficiency. Journal of the Royal Statistical Society 120:253–281.

[12] Meeusen, W.; van Den Broeck, J. 1977. Efficiency estimation from Cobb-Douglas production functions with composed error. International economic review:435-444.

[13] Schwendinger, F. R Interface to 'lp_solve'. R package version 5.5.2.0.

[14] Shapiro, S. S.; Wilk, M. B. 1965. An analysis of variance test for normality. Biometrika 52, 3–4: 591–611.

[15] Shleifer, A. 1985. A theory of yardstick competition. Rand Journal of Economics 16, 3:19-327.

[16] Student. 1908. The probable error of a mean. Biometrika 6, 1:1–25.

[17] Varian, H. R. 2015. Intermediate Microeconomics: a modern approach. 9ed. W.W. Norton & Company. New York, NY, USA.

[18] Wilcoxon, F. 1945."Individual comparisons by ranking methods. Biometrics Bulletin 1, 6:80–83.