Efficiency Measurement of Research Groups using Data Envelopment Analysis and Bayesian Networks

Cristhian F. Ruiz[1], Ricardo Bonilla[2], Diego A. Chavarro[3], Luis A. Orozco[4], Roberto Zarama[5], Xavier Polanco[6]

[1] *cr-ruiz@uniandes.edu.co,* [2] *rbonilla@uniandes.edu.co,* [3] *dchavarr@uniandes.edu.co,*
[4] *lorozco@uniandes.edu.co,* [5] *rzarama@uniandes.edu.co*
Universidad de Los Andes, Carrera 1 Este No. 19 A - 40, Bogotá D.C. (Colombia)

[6] *xavier.polanco@gmail.com*
Universidad del Rosario, Calle 14 No. 6 - 25, Bogotá D.C. (Colombia)

Abstract

Applications of non-parametric frontier production methods such as Data Envelopment Analysis (DEA) have gained popularity and recognition in scientometrics. DEA seems to be a useful method to assess efficiency of research units in different fields of knowledge or disciplines. However, the relations between DEA results and the underlying structure of scientific production of each discipline have not been fully explored. Although there are works that mention the importance to perform studies by scientific disciplines, they do not show how to take into account these differences in the analysis. These studies tend to homogenize the behavior of different communities of practice. In this paper we propose a framework to perform inferences about DEA efficiencies, recognizing the underlying structure of each discipline by means of Bayesian Network (BN) analysis. Two different DEA extensions are applied to calculate the efficiency of research groups, one called CCRO and the other Cross Efficiency (CE). A BN model is proposed as a method to analyze the results obtained from DEA. BNs allow us to recognize peculiarities of each discipline in terms of scientific production and the efficiency frontier. Moreover, BNs bring insight about the relationships between production variables and their impact in each discipline.

Introduction

In this paper we measure the efficiencies of research groups considered as "decision making units" (DMUs). Our proposal combines two methods Data Envelopment Analysis (DEA) and Bayesian Networks (BNs). DEA gives an overview of the relative efficiency of research groups in terms of production. BNs provide a useful representation of the influence between the production variables and DEA efficiency. Our mayor contribution is the use of BN as a graphical probabilistic reasoning tool to assess the influence of production variables in the efficiency. This allows the evaluation of the impact of each related production variable.

In the literature we observe that DEA has become useful to approach science and technology (S&T) efficiency. Bonaccorsi and Daraio (2004) highlight the use of non-parametric efficiency frontier (e.g. DEA) as one of the most adequate methods to evaluate the efficiency of the DMUs in S&T systems. In scientometrics, DEA method has been applied using countries (Bonaccorsi & Daraio, 2004; Rousseau & Rousseau, 1997; 1998), scientific fields (Garg, Gupta, Jamal, Roy, & Kumar, 2005; Meng, Hu, & Liu, 2006), universities (Bonaccorsi, Daraio, & Simar, 2006) or research centres (Korhonen, Tainio, & Wallenius, 2001) as different DMUs. Finally as Guan and Wang (2004) we use research groups as DMUs.

BNs are widely used in data mining, machines learning or decision theory to construct influence diagrams, decision making tools or classifiers. But BNs are less used in scientometric. In scientometric BNs are used to evaluate the collaboration between authors (Lehmann, Jackson, & Lautrup, 2008), to assert the impact of variables in the innovation production process (Kim & Park, 2008), and to approach analysis in Webometrics (Lamirel, Al Shehabi, Francois, & Polanco, 2004). In this paper BN analysis of DEA results is proposed to

Proceedings of ISSI 2009, pages 296–300.
Edited by B. Larsen and J. Leta.

take into account both, the underlying structures of each discipline and to assess the impact of production variables over the efficiency.

Methodology

DEA allows making comparative analysis of the relative efficiency of research groups as DMUs. A DMU is the entity which transforms a certain number of inputs into outputs through a specific process (Cooper, Seiford, & Zhu, 2004). The Measure of efficiency quantifies the distance to an efficiency frontier. Two DEA extensions are applied CCRO proposed by Charnes, Cooper and Rhodes (1978); and cross-efficiency (CE) proposed by Sexton, Silkman and Hogan (1986).

CCRO efficiency calculus is made by means of the construction of an efficiency frontier rather than a central tendency. The efficiency frontier is constructed by DMUs with larger ratio between outputs earned and inputs spent. It means that, the frontier is constructed by the relevant production of the DMUs in the sample. About CCRO, Doyle and Green (1994: 569) argue that high efficiency could be reach not only aiming to all outputs, but having high level of production in a single output. On the other hand, to calculate CE an additional calculation is made over CCRO results. CE compares the value of the DMU efficiency versus the levels of production of the others DMUs (Doyle & Green, 1994: 570). With CE, DMUs are evaluated against relevant production of the sample and versus their peers. CCRO and CE were applied using Restrepo and Villegas (2007)[2] implementation in Matlab®. This implementation was programmed by the optimization *linprog toolbox*, and it solves the linear program by means of *Simplex Method*.

The results obtained with CE extension by discipline are incorporated in a model of BN. A BN is a probabilistic graphical model that represents a set of random variables and their relationship (probabilistic independencies) by a Direct Acyclic Graph (DAG) and a set of probability tables (Heckerman, 1999; Kjaerulff & Madsen, 2008) (figure 4). The DAG structure and the probabilities tables permit the calculus of information transfer to assert the impact of each variable over the efficiency (Pearl, 2000; Spirtes, Glymour & Scheines, 2000). The BN structure and conditional probability tables are constructed using the software Hugin Lite[3]. Two main algorithms are used the NPC constrain based algorithm and the maximization expectation (ME) algorithm. The NPC algorithm is used to construct the BN structure by means of statistical independence test using mutual information calculus. The ME algorithm is used to construct probability tables based on likelihood maximization over the data.

Data

The data for this study came from the ScienTI database[4]. ScienTI is a Colombian repository of research groups and bibliographic references of research products. The sample is composed of 553 research groups organized in 8 sets that correspond to the following disciplines: Electrical Engineering (65), Civil Engineering (43), Physics (94), Chemistry (78), Biology (53), Ecology (45), Law (89), and Economics (86).

To perform DEA we define two input variables and five output variables. The input variables are the number of researchers in the group and the group experience. The group experience is measured by the number of years since foundation. The output variables are papers, books, chapters, proceedings, and thesis. To perform BN analysis, all the values of the output variables plus the efficiency calculated with CE extension are considered as random variables.

[2] Retrieved February 2009, from http://industrial.udea.edu.co/jgvillegas/Pagina%20DEA/index.html
[3] Retrieved February 2009, from http://www.hugin.com/
[4] Retrieved July 2008, from http://thirina.colciencias.gov.co:8081/scienti/

Results

The Two DEA extensions (CCRO and CE) are applied to Colombian research groups. Three main results are obtained; 1. A comparison between CCRO and CE in the whole sample. 2. A comparison of CE between DMUs grouped and not grouped by disciplines. 3. The BNs construction for each discipline using CE results.

When comparing CCRO and CE we observe that CCRO values are always higher than CE values (figure 1). The difference could be explained following Cooper, Seiford and Zhu (2004: 22). They state that CCRO assign "unreasonably low or excessively high values to the multipliers in an attempt to drive the efficiency rating for a particular DMU as high as possible". That means that a DMU can reach high values of efficiency with high production values in only one output. Due to in CE extension a comparison between DMUs is performed: DMUs with low values in some outputs (even if a higher value in few ones) are penalized against DMUs with spread production over all possible outputs.

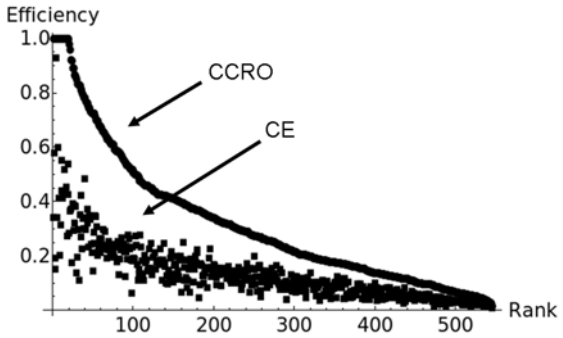

Figure 1. Rank Plot of CCRO and CE values for 553 research groups in Colombia.

The second result is the comparison between CE calculated for each DMU grouped by disciplines and CE calculated against the 553 DMUs of the sample (*Bulk*). Figure 2 is an example of 2 among the 8 disciplines. X-axis is the CE calculated for each DMU taking the whole sample of DMUs and Y-axis the CE only over the DMUs in their discipline. We observe that efficiencies calculated by discipline are usually higher than efficiencies calculated in the whole sample.

Figure 3 shows relevant information from the disciplines not shown. In this figure we extract the amount of spread around the trend (continuous line in figures 2) and the difference between the slopes of the continuous lines and the dashed line. The Continuous line is a linear fit of the data. The dashed line represents the curve $y = x$ (DMU CE by discipline equal to DMU CE in the whole sample). A high value in the X-axis can be interpreted as a large difference when a DMU is evaluated inside its discipline and from outside it. A high value in the Y-axis can be interpreted as the discipline is composed by DMUs that show homogeneous underlying structure of production. This observation supports the decision of making analysis by disciplines, whereby DMUs are evaluated against relevant production of its field and against their peers.

The last result is the BNs construction for each discipline using efficiency calculated with CE extension and the DEA outputs variables as nodes. BNs represent the join probability distribution of selected variables. They represent the dependency relationships between the outputs variables and the efficiency. The construction of the BN was unsupervised. This means the structure and the probability tables reflect the underlying structure of the data. Figure 4 shows 2 BNs examples, Physics and Chemistry, where we observe different structures. This observation suggests that each discipline has its own underlying structure of production. Using BNs we can calculate by means of sensitivity analysis the impact of each variable over the efficiency. Table 1 shows by discipline the mutual information (MI)

between CE and each production variable. The MI quantifies the share information between the two variables and can be interpreted as the amount of information gained over the CE when the second variable is known. A high value in the MI can be interpreted as a high impact from the product variable over the CE. An MI equal to cero means that there is no statistical impact. However, such variable is required to calculate CE.

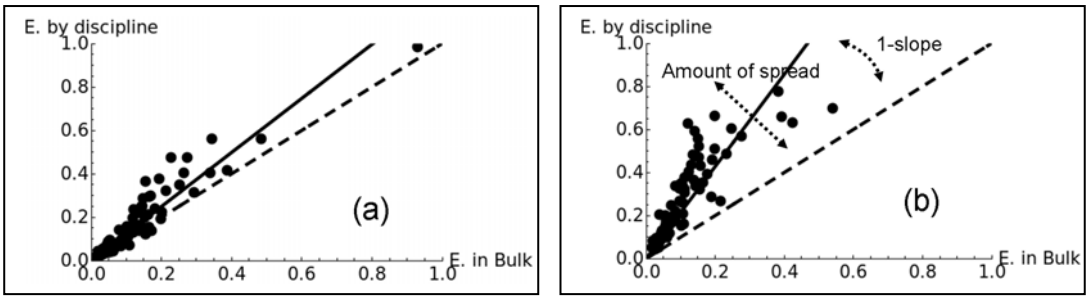

Figure 2. CE values calculated by discipline (a) Physics 94 DMUs (b) Chemistry 78 DMUs versus CE values calculated over the 553 DMUs.

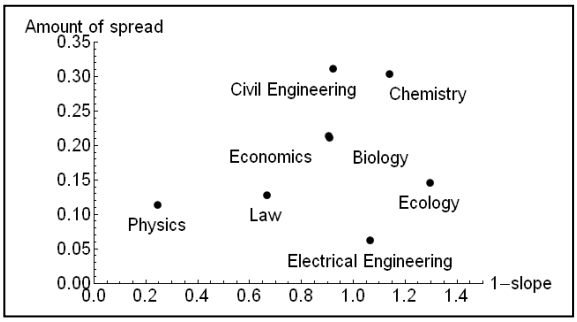

Figure 3. 1-slope and 1-the amount of spread around the linear fit for all disciplines.

Conclusions

We have presented a combination of DEA and BN to calculate and analyze values of efficiency in research groups. As a result we obtain a framework to perform comparisons between disciplines in terms of efficiency. We can infer about the impact of each production variable in the efficiency. These results are examples of an adequate application of efficiency analysis in science and technology. BNs analysis provides a visual representation of the dependency relationship between output variables and efficiency. This representation provides us with an interactive user interface useful to refine intuition about the underlying structure of scientific production. Particularly to the Colombian case of study we contribute to the discussion over the ranking methods of research groups. Actually in Colombia ranking is performed by production, here we present a procedure to evaluate research groups by efficiency and their field of knowledge.

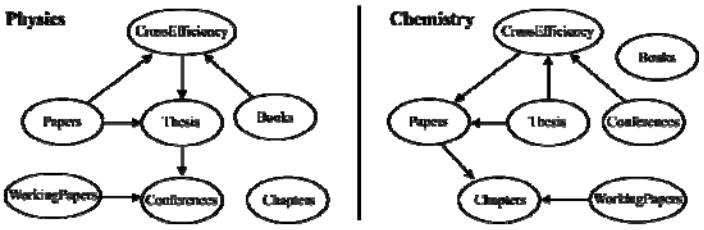

Figure 4. Representation of the relationship between production variables and CE using BNs.

Table 1. Mutual Information between Cross Efficiency and production variables.

Production Variable	Physics	Chemistry	Biology	Ecology	Economics	Law	Civil Engineering	Electrical Engineering
Working Papers	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Conferences	0.11	0.25	0.45	0.37	0.24	0.10	0.34	0.36
Chapters	0.00	0.02	0.01	0.00	0.14	0.06	0.00	0.11
Books	0.01	0.00	0.08	0.00	0.28	0.06	0.14	0.15
Papers	0.63	0.23	0.45	0.15	0.10	0.10	0.41	0.00
Thesis	0.52	0.22	0.38	0.19	0.00	0.08	0.00	0.22

References

Bonaccorsi, A. & Daraio, C. (2004). *Econometric Approaches to the Analysis of Productivity of R\&D Systems. Production Functions and Production Frontiers*. In W. Glänzel, U. Schmoch, M. Zitt, E. Bassecoulard & M. Luwel (Eds.), *Handbook of Quantitative Science and Technology Research* (pp. 51-74). Amsterdam: Springer Netherlands.

Bonaccorsi, A., Daraio, C. & Simar, L. (2006). Advanced indicators of productivity of universities. An application of robust nonparametric methods to Italian data. *Scientometrics, 66*(2), 389-410.

Charnes, A., Cooper, W. W. & Rhodes, E. (1978). Measuring the efficiency of decision-making units. *European Journal of Operational Research, 2*(6), 429-444.

Cooper, W. W., Seiford, L. M., & Zhu, J. (2004). . In J. Zhu, W. W. Cooper, R. D. Banker & L. M. Seiford (Eds.), *Handbook on Data Envelopment Analysis* (pp. 1-39). Kluwer Academic.

Doyle, J. & Green, R. (1994). Efficiency and cross-efficiency in DEA: derivations, meanings and uses. *The Journal of the Operational Research Society, 45*(5), 567-578.

Garg, K. C., Gupta, B. M., Jamal, T., Roy, S. & Kumar, S. (2005). Assessment of impact of AICTE funding on R&D and educational development. *Scientometrics, 65*(2), 151-160.

Guan, J. C. & Wang, J. X. (2004). Evaluation and interpretation of knowledge production efficiency. *Scientometrics, 59*(1), 131-155.

Heckerman, D. (1999). *A Tutorial on Learning with Bayesian Networks*. In Jordan, M. (Ed), *Learning in Graphical Models.* Cambridge: MIT Press.

Kim, H. & Park, Y. (2008). The impact of R&D collaboration on innovative performance in Korea: A Bayesian network approach. *Scientometrics, 75*(3), 535-554.

Kjaerulff, U. B., & Madsen, A. L. (2008). *Bayesian Networks and Influence Diagrams. A Guide to Construction and Analysis*. New York: Springer New York.

Korhonen, P., Tainio, R. & Wallenius, J. (2001). Value efficiency analysis of academic research. *European Journal of Operational Research, 130*(1), 121-132.

Lamirel, J. C., Al Shehabi, S., Francois, C. & Polanco, X. (2004). Using a compound approach based on elaborated neural network for Webometrics: An example issued from the EICSTES project. *Scientometrics, 61*(3), 427-441.

Lehmann, S., Jackson, A. D. & Lautrup, B. E. (2008). A quantitative analysis of indicators of scientific performance. *Scientometrics, 76*(2), 369-390.

Meng, W., Hu, Z. & Liu, W. (2006). Efficiency evaluation of basic research in China. *Scientometrics, 69*(1), 85-101.

Pearl, J. (2000). Causality : models, reasoning, and inference. In J. Pearl (Ed.), (pp. 1-40): Cambridge University Press.

Restrepo, M. & Villegas, J. (2007). Clasificación de grupos de investigación colombianos aplicando análisis envolvente de datos. *Revista Facultad de Ingeniería Universidad de Antioquia*, 42, 105-119.

Rousseau, S. & Rousseau, R. (1997). Data envelopment analysis as a tool for constructing scientometric indicators. *Scientometrics, 40*(1), 45-56.

Rousseau, S., & Rousseau, R. (1998). The scientific wealth of European nations: Taking effectiveness into account. *Scientometrics, 42*(1), 75-87.

Sexton, T. R., Slinkman, R. H. & Hogan, A. (1986). *Data envelopment analysis: Critique and extensions*. In R. H. Silkman (Ed.), *Measuring Efficiency: An assessment of Data Envelopment Analysis* (Vol. 32, pp. 73-105). San Francisco: Jossey-Bass.

Spirtes, P., Glymour, C. N. & Scheines, R. (2000). *Discovery Algorithms for Causally Sufficient Structures*. In P. G. C. N. Spirtes, P., Glymour, C. N. & Scheines, R (Eds.), *Causation, prediction, and search* (pp. 73-122). Cambridge: MIT Press.