IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Hershey PA 17033-1117, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITP4211

The Contribution of Information Technology To Organizational Productivity: An Exploratory Study Using A Regression Tree

Myung Ko and Kweku-Muata Bryson
Department of Information Systems and The Information Systems Research Institute, Virginia Commonwealth University, Virginia
Tel: (804) 225-3627, Fax: (804) 828-3199, Msko@vcu.edu, Kweku.Muata@isy.vcu.edu

ABSTRACT

As investments in information technology (IT) have continuously increased, identifying the contribution of IT investments has been a major issue of IS research. In this study, we explored the relationship of IT to organizational productivity using a regression tree (RT), a Data Mining technique. Based on a firm-level dataset, our results are consistent with the previous studies in that IT investments make a positive contribution to the firm's productivity. However, our RT based analysis has revealed additional facts. While IS Labor contributes positively to the firm's output, Computer Capital, which represents the market value of IT infrastructure, does not. In addition, contribution of IS Labor to organizational productivity is not uniform. When Labor is within a certain range, IS Labor contributes to the firm's output. Otherwise, the contribution of IS Labor is insignificant. In addition, none of IT contributes in generating the highest value of firm's productivity. This indicates that IT "Productivity Paradox" still exists.

INTRODUCTION

As investments in information technology (IT) have continuously increased, identifying the contribution of IT investments has been a major issue of IS research. The purpose of this paper is to explore whether IT investments make a positive contribution to organizational productivity. While the majority of previous studies have applied Econometrics, we employed a Regression Tree (RT) technique, which provides a different perspective from previous research. In this study, we used a firm-level dataset that has been used in previous studies (Hitt and Brynjolfsson, 1996; Brynjolfsson and Hitt, 1996; Shao and Lin, 2000 & 2001).

Overall, our findings are consistent with previous studies in that the contribution of IT to the firm-level productivity is positive. However, our RT based analysis has revealed additional facts. Computer Capital which represents the market value of IT infrastructure makes insignificant contribution to firm's output while IS Labor contributes significantly. In addition, contribution of the IS Labor to organizational productivity is not uniform. When non-IS Labor is within a certain range, contribution of IS Labor is significant. When non-IS Labor is out of the range, IS Labor does not make any contribution to the firm's productivity. In addition, none of IT contributes in generating the highest value of firm's productivity. This indicates that IT "Productivity Paradox" still exists.

The remainder of the paper is organized as follows. The next section discusses the previous research. Section 3 discusses the overview of a Regression Tree technique in Data Mining. Section 4 briefly describes the experimental data and section 5 discusses the research methodology. Section 6 discusses the empirical results from our study and the paper concludes in the final section.

PREVIOUS RESEARCH ON IT IMPACTS

In this section, we briefly review the previous empirical research on IT investments and their impact on firm's productivity. Table 1 describes the summary of the previous studies. Some earlier firm-level studies have found no relationship between IT investments and productivity. Loveman (1994) examined 60 manufacturing business units based on the general production function and did not find any impact of IT investments on output or labor productivity.

More recent studies provided evidence of a positive relationship between IT investments and organizational productivity. Brynjolfsson and Hitt (1996) used a firm-level data of 367 large firms for the period from 1987 to 1991. Using several econometrics models based on the Cobb-Douglas production function, they related computer capital (market value of central processors and PCs), non-computer capital, IS labor, and non-IS labor to firm's productivity and found that IS spending made a significant contribution to firm-level productivity. In addition, they claimed that the productivity paradox disappeared by 1991.

Dewan and Min (1997) assessed IT substitutability for other inputs using the Constant Elasticity of Substitution (CES) translog and translog production functions. Their study indicated that IT capital was a substitute for capital and labor.

Shao and Lin (2001) examined the impact of IT investments on technical efficiency in the firm's production process using the same dataset of Hitt and Brynjolfsson (1996). The authors used econometrics approach, the Cobb-Douglas and translog stochastic production frontiers and found that IT had a positive effect on technical efficiency in the production process.

Table 1: Research summary of firm-level studies on IT impact

Study	Research Method	Results of IT impact
Loveman (1994)	Econometrics	No evidence of productivity gains from IT investments
Brynjolfsson & Hitt (1996)	Econometrics	Significant contribution of IS spending to firm output
Dewan & Min (1997)	Econometrics	IT capital is a substitute for capital and labor.
Shao & Lin (2001)	Econometrics	IT has a positive effect on technical efficiency and thus, it leads to the productivity growth.

OVERVIEW ON REGRESSION TREES

Data Mining has emerged due to the need for discovering hidden knowledge from the fast-growing huge amounts of organizational databases. An important knowledge structure in data mining activities is the decision trees (DT). A DT is a tree-shaped structure, which consists of nodes, branches, and leaves. For a given decision problem, each non-leaf node is associated with one of the decision variables, each branch from a non-leaf node is associated with a subset of the values of the corresponding decision variable, and each leaf node is associated with a value of the target (or dependent) variable.

Two main types of DTs are classification trees and regression trees. For a classification tree, the target variable takes its value from a discrete domain, and for each leaf the DT associates a probability for each class (i.e. value of the target variable). For a regression tree (RT), the target variable takes its value from a continuous domain, and for each leaf the DT associates the mean value of the target variable.

This paper appears in *Issues and Trends of Information Technology Management in Contemporary Organizations,* the proceedings of the Information Resources Management Association International Conference.

In order to generate a RT, the model dataset is partitioned into at least two parts: the training dataset and the validation (or test) dataset. Once a RT is generated from the training dataset, it is evaluated against the validation (or test) dataset and a subtree that has the lowest error rate is generated.

While the most commonly used performance measure for an RT is *predictive accuracy* (e.g. mean square error, R-squared), *simplicity* and *stability* are also important performance measures. *Simplicity* is referred to as the interpretability of the RT. *Stability* of RT can be checked if the results generated by the RT for the training dataset are similar to ones for the validation dataset. One way to assess the stability of the RT is to compare the predicted mean value of the target variable based on the training dataset and the corresponding value based on the validation dataset.

RTs are similar to regressions since both techniques are used for the prediction. However, the main difference between two techniques is that the RT model uses discontinuous step functions, whereas the regression model uses continuous linear functions (Clark and Pregibon, 1992). Compared to regression models, RTs provide a model with better interpretability because the model represents interpretable English rules or logic statements. One of limitations of traditional regression models is linearity of the functional relationship. To satisfy such a requirement, one needs to use transformed variables and the results are often not easy to interpret. RTs can be also used for an alternative approach for regression problems. There have been instances where a DT has shown clues to datasets while a traditional linear regression analysis could not clearly indicate them (Breiman et al., 1984). However, instability of RTs can be problem with respect to perturbations in data**.** To minimize instability, we can generate multiple trees and choose the best model that fits one's objective.

EXPERIMENTAL DATA

The dataset used in this study is the same dataset of Hitt and Brynjolfsson (1996). Several other studies have also used this dataset in determining IT impact on productivity (Brynjolfsson and Hitt, 1996; Shao and Line, 2000 & 2001). Although the dataset is somewhat old since it covers from 1988 to 1992, adopting the same dataset promotes comparability of research findings that have used different techniques. Thus, the results of our study using regression tree approach can be compared with other studies without any bias.

IT related data were gathered by International Data Group (IDG) from the annual surveys of IT spending by large U.S. firms for the period from 1988 to 1992, and these data are matched to Standard & Poor's Compustat II to obtain financial related data. The dataset includes an unbalanced panel of 370 firms consists of 1252 observations. Table 2 describes the variables used in this study.

Table 2: Variable definitions (Source: Hitt and Brynjolfsson, 1996)

Variable	Description	Source
OUTPUT	Gross Sales deflated by Output Price (see below).	Compustat
VALUE ADDED (V)	Output minus non-labor expense. Non-labor expense is calculated as total firm expenses (excluding interest, taxes, and depreciation) divided by Output Price less Non-IS Labor (see below)	Compustat
COMPUTER CAPITAL (C)	Market value of central processors plus value of PCs and terminals. Deflated by Gordon's deflator for computer systems. Average value of PC determined as weighted average of PC price from Berndt and Griliches (1990) and value of PC from IBM. Resulting estimate is $2,840 in 1990 dollars.	IDG Survey
CAPITAL (K)	Deflated book value of Capital obtained from Compustat less Computer Capital as calculated above.	Compustat
LABOR (L)	Available labor expenses or estimated labor expenses based on sector average labor costs times number of employees minus IS Labor. Obtained from Compustat. Deflated by price index for total compensation.	Compustat
IS LABOR (S)	Labor portion of IS budget. Deflated by price index for total compensation. See above for calculation of Labor expenses.	IDG Survey
OUTPUT PRICE	Output deflator based on 2-digit industry. If not available, sector level deflator for intermediate materials, supplies, and components.	Bureau of Economic Analysis, 1993

RESEARCH METHODOLOGY

We explore the relationship between IT investments and firm's productivity based on a production function, which was used by previous researchers (Loveman, 1994; Hitt & Brynjolfsson, 1996; Dewan & Min, 1997). The production function describes a relationship between specified inputs and outputs in production. In this study, we use four input variables, such as COMPUTER CAPITAL (C), CAPITAL (K), IS LABOR (S), and LABOR (L). As we described in Table 2, COMPUTER CAPITAL (C) represents the total market value of IT Infrastructure (central processors, PCs, and terminals) and IS LABOR (S) represents labor portion of IS budget. While COMPUTER CAPITAL and IS LABOR represent IT investments, CAPITAL and LABOR represent non-IT portion. These four input variables are related to one output (VALUE ADDED (V)) variable using Cobb-Douglas production function as follows:

$$V = C^{\beta_1} K^{\beta_2} S^{\beta_3} L^{\beta_4} \quad (1)$$

where β_1, β_2, β_3 and β_4 are unknown parameters to be estimated. By taking natural logarithms, equation (1) can be expressed in terms of linear regression:

$$\text{Log } V = \beta_0 + \beta_1 \log C + \beta_2 \log K + \beta_3 \log S + \beta_4 \log L + \varepsilon \quad (2)$$

where ε is an error term.

Our approach is to model Log V as a discontinuous function (F) of log C, log K, log S, and log L where F is a regression tree to be estimated. Since the regression tree is invariant under the transformations of input variables, model could be stated as

$$\text{Log } V = F(C, K, S, L) + \varepsilon \quad (3)$$

We used the SAS Enterprise Miner (EM) software, version 8.2 to generate a RT. We partitioned the dataset into Training and Validation (sometimes called Test) using a stratified sampling approach. Approximately 60% of the data was used for Training and 40% for Validation. Two variables, YEARNO (the Year of the Observation) and INUM (the Industry – two digit primary SIC level) are used for stratification variables to ensure that characteristics of both training and validation dataset are close to each other.

Empirical Results from Regression-Tree Based Analysis

The predictive accuracy obtained from our RT in terms of Average Squared Error (ASE) and R squared for the Training dataset are 0.117 and 0.8934 and for the Validation dataset are 0.118 and 0.8968, respectively. Table 3 includes a ruleset obtained from our RT based analysis. Each row in the table represents a rule. The Condition Component columns represent the range of values for the relevant input variables for each rule. The Target columns represent the predicted mean values obtained from the Training and the Validation datasets for the target variable. Standard deviation (SD) for the target variable is enclosed in parentheses in the Training column. The IT Impact column indicates whether IT (Computer Capital or IS Labor) was included in the relevant rule and specifies whether IT makes a contribution to the target value. As shown in Table 3, the ruleset in our study generated the fourteen rules.

The predicted mean values of the target variable from the training dataset and the validation dataset in Table 3 are very close to each other. This indicates the stability of our RT. The ruleset described in Table 3 revealed several facts as following:

1. None of rules selects Computer Capital (C) as an input variable.
2. IS Labor (S) makes a positive contribution to firm's output when non-IS Labor (L) is within a certain range. The bold amounts in the Labor (L) column in Table 3 represent this range. When the non-IS

Table 3: Description of ruleset of RT–Sorted by Labor and Mean Target Value for Training

Condition Component				Target: Mean Value Added log (V)		IT Impact
Non-IT ($ Million)		IT ($ Million)				
Labor (L)	Capital (K)	Computer Capital (C)	IS Labor (S)	Training (SD)	Validation	
[0.000, 118.98]	N/S	N/S	N/S	5.38611 (0.70552)	5.16679	No
[118.98, 403.45]	[0.0, 3448.82]	N/S	N/S	6.19515 (0.42514)	6.20791	No
[118.98, 403.45]	[3448.82, +∞]	N/S	N/S	6.82107 (0.41177)	6.92571	No
[403.45, 694.91]	[0.0, 3840.34]	N/S	N/S	6.72978 (0.28424)	6.75140	No
[403.45, 999.17]	[3840.34,12753.64]	N/S	N/S	7.29434 (0.25999)	7.23583	No
[403.45, 999.17]	[12753.64,+∞]	N/S	N/S	7.67666 (0.20748)	7.73037	No
[694.91, 999.17]	[0.0, 3840.34]	N/S	N/S	7.13068 (0.22992)	7.12547	No
[999.17, 1670.70]	**N/S**	**N/S**	**[0, 51.77]**	**7.66335 (0.26967)**	**7.66475**	**Yes**
[999.17, 2798.02]	**[0.0, 19612.99]**	**N/S**	**[51.77, +∞]**	**8.22403 (0.31983)**	**8.03034**	**Yes**
[999.17, 2798.02]	**[19612.99, +∞]**	**N/S**	**[51.77,+∞]**	**8.71704 (0.33447)**	**8.59984**	**Yes**
[1670.7, 2798.02]	**N/S**	**N/S**	**[0, 51.77]**	**8.03186 (0.24016)**	**8.11727**	**Yes**
[2798.02, 6449.7]	[0.0, 20654.36]	N/S	N/S	8.66923 (0.32431)	8.63342	No
[2798.02, 6449.7]	[20654.36, +∞]	N/S	N/S	9.16245 (0.26933)	9.16241	No
[6449.70, +∞]	N/S	N/S	N/S	9.98549 (0.55757)	9.87332	No

Legend: N/S–Not Selected

Labor (L) amounts are out of the range, IS Labor (S) does not make any difference to the firm's output. Thus, contribution of IS Labor to the firm's output may not be uniform.

3. The mean value for the output is lower for the range where IS Labor (S) contributes than for the range where both non-IS Labor is the highest and there is no IT contribution. This indicates that IS Labor (S) is not a factor generating the highest mean value of the output.
4. Our findings describe an evidence of IT 'Productivity Paradox.'

To ensure the validity of our findings based on our initial RT, we have also generated three additional RTs that varied the Splitting Criterion, the Minimum Number of Observations per Leaf, and the Observations Required for a Split Search. When we review the results from these RTs, we could draw same conclusions as our initial RT. Thus, we are confident that our initial RT is accurate and stable.

CONCLUSION

We explored the relationship of IT to the firm's productivity using a regression tree-based analysis and determined that our results are consistent with the previous studies in that IT makes a positive contribution to the firm's productivity. However, our RT-based analysis discovered additional facts that have not identified in previous studies. While the IS Labor makes a positive contribution to the firm's output, Computer Capital is not a contributing factor to organizational productivity. This indicates that Computer Capital, which represents the market value of IT infrastructure, is less important than IS Labor to organizational productivity. In addition, contribution of IS Labor to the organizational productivity is not uniform. When non-IS Labor is within a certain range, IS Labor contributes to the organizational productivity. Otherwise (either too low or too high), IS Labor does not make any difference to firm's output. Our study also revealed that none of IT contributes to generating the highest value of productivity. Thus, it indicates that IT "Productivity Paradox" still exists.

ACKNOWLEDGMENTS

We are grateful to Erik Brynjolfsson and Lorin Hitt for kindly sharing their dataset with us, thus facilitating this study.

REFERENCES

Brynjolfsson, E. (1993). "The Productivity Paradox of Information Technology." Communications of the ACM **36**(12): 66-76.

Brynjolfsson, E. and L. M. Hitt (1996). "Paradox Lost? Firm-level Evidence on the Returns to Information Systems Spending." Management Science **42**(4): 541-558.

Clark, L. and D. Pregibon (1992). Tree-Based Models. Statistical Models in S. J. M. Chambers and T. J. Hastie. Pacific Grove, California, Wadsworth & Brooks / Cole Advanced Books & Software**:** 377-419.

Dewan, S. and C.-K. Min (1997). "The Substitution of Information Technology for Other Factors of Production: A Firm Level Analysis." Management Science **43**(12): 1660-1675.

Han, J. and M. Kamber (2001). Data Mining Concepts and Techniques, Academic Press.

Hitt, L. M. and E. Brynjolfsson (1996). "Productivity, Business Profitability, and Consumer Surplus: Three Different Measures of Information Technology Value." MIS Quarterly **20**(2): 121-142.

Loveman, G. W. (1994). An Assessment of the Productivity Impact of Information Technologies. Information Technology and the Corporation of the 1990s: Research Studies. T. J. Allen and M. S. Scott Morton, Oxford University Press**:** 84-110.

Shao, B. and W. Lin (2000). "Examining the Determinants of Productive Efficiency with IT as a Production Factor." Journal of Computer Information Systems **41**(1):25-30.

Shao, B. and W. Lin (2001). "Measuring the Value of Information Technology in Technical Efficiency with Stochastic Production Frontiers." Information and Software Technology **43**: 447-456.

Torgo, L. (1997). Functional Models for Regression Tree Leaves. International Conference on Machine Learning.