THE ANALYSIS OF DATA PREPARATION TO VALIDATE MODEL VALUES OF INFORMATION TECHNOLOGY

Taufik Hidayat, Rahutomo Mahardiko, and Ali Miftakhu Rosyad

Abstract. Currently, there are some methods of preparing data for validating an IT value model correctly. One challenge in applying data mining to validate model values is to convert data into an appropriate form for this activity. Data mining algorithms can then be applied using the prepared data. The adequacy of data preparation often determines whether this data mining is successful or not. This study aims at creating a method for preparing the data during validation. The basic method used for data preparation is the Returns to Scale (RTS) method because it is easy to use and can be combined with further validation results. This method was applied by employing two models: two-factor and three-factor models. Both models are then compared to see the difference between them. The developed model is then tested on Branchless Banking (BB) and Downstream Petroleum (DP) industries. The results show that the method is applicable to prepare the data for validation. In addition, the results also demonstrate that both industries, DP and BB, have different result on data preparation, meaning that DP and BB have different ITs. This research contributes not only to a technique of preparing data for validating an IT value model by the RTS method but also can be a basis to work for data validation because it can give a result with the behaviour of the industry.

Keywords: IT value; IT value model; Returns to Scale method; data preparation; validation.
Authors:

Taufik Hidayat
Department of Electrical Engineering, Universitas Indonesia, Depok, Indonesia
E-mail: taufik.hidayat23@ui.ac.id
https://orcid.org/0000-0003-0230-9872

Rahutomo Mahardiko
Department of Data Management, PT.BFI Finance Indonesia Tbk, Indonesia
E-mail: rahutomo.mahardiko@bfi.co.id
https://orcid.org/0000-0001-7324-0556

Ali Miftakhu Rosyad
Department of Islamic Education, Universitas Wiralodra, Indramayu, Indonesia
E-mail: miftakhurosyad@gmail.com
https://orcid.org/0000-0002-8845-2364

Corresponding author: Taufik Hidayat, taufik.hidayat23@ui.ac.id

Citation: Hidayat, T., Mahardiko, R., & Rosyad, A. M. (2023). The Analysis of Data Preparation to Validate Model Values of Information Technology. Virtual Economics, 6(2), 23-34. https://doi.org/10.34021/ve.2023.06.02(2)
1. Introduction

In fact, Information Technology (IT) value pattern progress has been extensively conducted in many ways. An IT value model validation verifies the IT value model for the firm used as a research subject. The IT value model validation for the private firm, national, and even governmental institution levels are being developed in the last two decades. The validation compares a two-factor model with a three-factor model. The validation results help the IT value analysis and engineering processes and the validation also improves the degree of accuracy in the IT value analysis and engineering [1–3]. This study will focus on analysing the data preparation before validating model values of IT in the several aspects, especially in the engineering context.

Based on the several studies the validation results were discussed by the previous research. Validation on the model can be performed with many methods, such as: Partial Adjustment Valuation (PAV) [4], Traditional Regression Analysis (TRA) [5], Stochastic Production Frontier (SPF) [6], and Data Envelopment Analysis (DEA) [7,8]. PAV for the validation using constant speed of adjustment was proposed for the first time in 2010, and the validation was also implemented at the country-level. Since then, the validation was studied and implemented in other ways, such as: a firm, a governmental institution and even a country level. All of previous research on PAV was conducted because (1) Partial Adjustment Valuation was more valuable than Data Envelopment Analysis and Traditional Regression Analysis and (2) it used fewer sample measurement than Stochastic Production Frontier. Partial Adjustment Valuation exploits two approaches, such as: static and dynamic Partial Adjustment Valuation at the early time. The recent research on PAV was centred on powerfully Partial Adjustment because of a variable which could be adjusted dynamically based on the obvious situation (macro or micro economic). The research of Partial Adjustment Valuation was also renewed to take innovation by current research on the IT value model in general, and dynamic PAV in particular [9,10]. The research on dynamic PAV could split the data into 2 types. The data that was used for a two-factor model are two independent variables (total capital without IT capital and total labour without IT labour) and a dependent variable (GDP). The data that was used for a three-factor model are three independent variables (total capital without IT capital, total labour without IT labour, and IT capital + IT labour) and a dependent variable (GDP). This result compared a 2-aspect model and a 3-aspect model. The dynamic speed of adjustment due to macroeconomic and microeconomic factors and the validation was implemented at a country-level [11].

The study [12] that used PAV for the validation using the dynamic speed of adjustment was held in branchless banking industry. The newest research [13] also applied PAV. Based on the factual condition, the Bank Indonesia’s interest rate was used as a macroeconomic factor for the dynamic speed of adjustment and the validation also was implemented at a firm level. The needed data were still the same data as the needed data from the last research, except the dependent variable (organizational revenue). This result compared a two-factor model with a three-factor model. On the other hand, the studies [14,15] that used PAV for the validation using adjustment were held in the petroleum industry. Sales growth was used as a microeconomic factor for the dynamic speed of adjustment and the validation was implemented at a company level. The needed data were still the same data as the needed data from the last research, except the dependent variable (organizational revenue). This result compared a two-factor model with a three-factor model [14].

Taufik Hidayat, Rahutomo Mahardiko, and Ali Miftakhu Rosyad
Virtual Economics, Vol. 6, No. 2, 2023
There is a problem while the validation process according to the data because the validation is of a direct effect. The last research on the data preparation was done in 2017 to separate data for a two-factor model. The result showed a positive result of the correction of a 2-factor model with Pearson correlation technique on a hybrid IT value model. However, no research on the data preparation was found on the indirect effect so far. As of today, a country/firm uses IT in their business process. GDP/revenue data in the validation are still uncorrected for a two-factor model. This problem occurs because the GDP/revenue data resulted from the business process that always contains an IT factor. Then, the GDP/revenue data in a two-factor model needs to be corrected. The question is how this research corrects the GDP/revenue data in a two-factor model? This question should be investigated. From this issue, the researchers originally thought that data preparation modelling for validation of an IT value model should be conducted to prevent multiple works during the validation.

This study will focus on the exploration of four aspects of data validity. The first aim is to get a data correction method for a two-factor model. The second aim is the method that can be used to improve the degree of accuracy by correcting the GDP/revenue data for a two-factor model. The third aim is this method can be used to differentiate the GDP/revenue data in 2-factor models and 3-factor models. The final purpose is to differentiate between holistically independent and dependent variables data between 2-factor models and 3-factor models. Besides all primary purposes that have been already explained, this research has the added value to the validation research. Data preparation is useful for the validation.

This research contains the systematized analysis that will describe the following. The second stage points out the literature review concerning the data preparation. The third stage highlights the research design and collection of data. The fourth stage summarizes results and discussion. The fifth part outlines the research conclusion. The last stage determines the research limitations.

2. Literature Review
2.1 Production Function

Quantitative data is a primary requirement in doing IT value model validation by PAV. PAV will explicitly use the production function. Many industries already use the production function briefly, such as material, service, modal, labour, etc. The production theory is a foundation theory for the PAV method. The theory has already been discussed for a long time and recently has been actively developed on the IT value model validation. The theory explains that an organization will use production factors/inputs (material, service, modal, labour and so on) and changes it into outputs. In the production process, the output that has been made is not always the same as the desired output [16]. Technical efficiency may be applied in this condition. The maximum output that can be made may depict organizational performance. The production process may be formulated as follows:

\[ Y_t = f(X_t; \beta) - u_t \]  \hspace{1cm} (1)

where \( f(X_t; \beta) \) is the desired output, \( X_t \) is production inputs (capital, labour, IT and so on), \( u_t \) is technical inefficiency, \( f(X_t; \beta) \) is a function that may be changed by Cobb-Douglas function.
2.2 IT Value Model Validation

The IT value model has been well developed for more than a decade. The model can be found easily in some trusted libraries, such as: IEEE, MIS Quarterly, Tandfonline, Springer, ScienceDirect, etc. The models consist of a direct impact and indirect impact. The direct-impact model highlighted that a digital system could contribute to operational capability, customer service and financial performance in several apparel industries. Meanwhile, the production function was employed to measure IT investment influencing 370 manufacture and services firms over the period of 1988-1992. The IT value was measured by GDP [17].

In some cases, the direct effect cannot explain well the IT effect. However, the IT value model can be explained well through an indirect effect. The use of intermediary (capability) was explained. Since then, the indirect-effect model was developed significantly in some industries and at a countries’ level. However, the validation of the IT value model needs to be addressed better [18].

Lately, the IT value model was validated using PAV. PAV was used based on the production function. It was said to use input and change the input into output. Partial Adjustment Valuation with static speed of adjustment:

\[ Y_t - Y_{t-1} = \delta \times (f(X_t; \beta) - Y_{t-1}) + W_t \]  (2)

where \( Y_t \) is the real result in current \( t \); \( Y_{t-1} \) was the real result in current \( t-1 \), \( f(X_t; \beta) \) is a desired output in time \( t \), \( \delta \) is a speed of adjustment ranging \( 0 - 1 \), \( W_t \) is an error factor that might happen.

The production function of a two-factor model by employing Cobb-Douglas function is:

\[ f(X_t; \beta) = \alpha K_t^{\beta_0} L_t^{\beta_1} e^{-u} \]  (3)

where \( \alpha \) is a vector of unknown coefficients, \( K_t \) is a conventional capital without IT-capital, \( L_t \) is a conventional labour without IT-capital, \( e \) is an error factor that might happen.

The production function of a three-factor model employing Cobb-Douglas function is:

\[ f(X_t; \beta) = \alpha K_t^{\beta_0} L_t^{\beta_1} I_t^{\beta_2} e^{-u} \]  (4)

where \( \alpha \) is a vector of unknown coefficients, \( K_t \) is a traditional capital without IT-capital, \( L_t \) is a traditional labour without IT-capital, \( I_t \) is the sum of IT capital and IT labour, \( e \) is an error factor that might happen.

During the validation, there is no method found on preparing the data except the use of Pearson correlation technique. Then, the data preparation is essential during the validation.
3. The Research Method
3.1 Methodology

This methodology is built to correct the needed data for a two-factor model and a three-factor model. All methodology activities are presented below. First, to look for the sample and necessity data collection must be performed. After finding a sample, the data may be retrieved from the secondary resources. The secondary data should be identified to determine the needed data. The validation result on the downstream petroleum and branchless banking industries gave the holistic view on the broad methodology for this research [15,12].

Second, the data preparation process should be executed. Then, RTS is carried out. The primary reason is to correct the data for two-factor. The second reason, data preparation for measuring performance has been successfully completed. Third, the result of RTS must be interpreted so that data preparation is meaningful [19].

3.2 Data

The data used in this research was borrowed from the validation result on downstream petroleum and branchless banking industries. To continue this research, data preparation activity is performed based on the previous research on the validation of the IT value model. The data for data preparation research are collected and processed, so that the meaning of data collection is useful for the industries [20,21].

There are four variables. Three variables were the investments (total capital without IT capital, total labour without IT labour, and IT capital + IT labour). The last variable is performance (GDP or revenue). IT investment (I) is a mount of IT spending and number of IT labour. The parts of IT spending are the hardware and software purchased by the institution a year. The parts of the IT labour share are all investments in the permanent staff of an IT office, investments in personnel during an IT project and investment in temporary staff of an IT office [22,23].

Capital expenditure (CAPEX) without digital or IT investment (K) could be recovered by decreasing CAPEX with IT expenditure to obtain good percentage. The official investment is a global investment action by the institution regardless the IT expenditure. The good percentage of CAPEX could be obtained simply in the reporting annually. The number of labour regardless IT labour (L) would be received by decreasing the number of labour regardless the IT labour. The variable was globally constant and tentative workers to spend, and others do not include IT labour. The amount of labour itself could be determined in the yearly report and financial reports. GDP or income is obtained easily based on the yearly report. This variable will illustrate the potential of a country or a company to grow every year [24]. The authors will display Table 1 and Table 2 which are the collections of data examples to be used for overcoming the defined problem. Some columns are omitted in Table 2.
Table 1. Sample Time-Series Data Period 2008-2013 in Branchles Banking Industry

<table>
<thead>
<tr>
<th>Year</th>
<th>K</th>
<th>L</th>
<th>I</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td>3.107</td>
<td>3.284</td>
<td>0.225</td>
<td>95.5</td>
</tr>
<tr>
<td>2009</td>
<td>3.745</td>
<td>4.186</td>
<td>0.321</td>
<td>135.3</td>
</tr>
<tr>
<td>2010</td>
<td>4.465</td>
<td>4.545</td>
<td>0.33</td>
<td>187.4</td>
</tr>
<tr>
<td>2011</td>
<td>5.132</td>
<td>5.205</td>
<td>0.336</td>
<td>224.1</td>
</tr>
<tr>
<td>2012</td>
<td>6.079</td>
<td>6.155</td>
<td>0.371</td>
<td>314.3</td>
</tr>
<tr>
<td>2013</td>
<td>6.854</td>
<td>6.866</td>
<td>0.532</td>
<td>389.9</td>
</tr>
</tbody>
</table>

Source: devised by the authors.

This study was conducted in the two ways. The first stage analyses the period of 2008-2013 to obtain the research data in Branchles Banking Industry concerning the data preparation and IT model. So, the second stage covered 2009-2014 to obtain data in Downstream Petroleum Industry concerning PAV. That research was already published in 2015 and 2020.

Table 2. Sample Time-Series Data Period 2009 – 2014 in Downstream Petroleum Industry

<table>
<thead>
<tr>
<th>Year</th>
<th>K</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>3,881.12</td>
<td>437.40</td>
</tr>
<tr>
<td>2010</td>
<td>3,365.95</td>
<td>709.24</td>
</tr>
<tr>
<td>2011</td>
<td>4,174.48</td>
<td>695.90</td>
</tr>
<tr>
<td>2012</td>
<td>5,730.44</td>
<td>1,759.16</td>
</tr>
<tr>
<td>2013</td>
<td>14,202.96</td>
<td>1,869.01</td>
</tr>
<tr>
<td>2014</td>
<td>18,480.61</td>
<td>3,237.35</td>
</tr>
</tbody>
</table>

Source: devised by the authors.

3.3 Returns to Scale (RTS) Method

Estimation result of two models can be compared using RTS method for data preparation purpose. The RTS method was applied because it is easy and can be combined with validation result of Cobb-Douglas from the previous research. This method is the only available on Cobb-Douglas production function. RTS points out the technical property of production which evaluated in result subsequent to a proportional change in all inputs (where all inputs increase by a constant factor). The combination of RTS for two-factor model employs total capital without IT capital, total labour without IT labour, and GDP or revenue. On the other hand, the combination of RTS for a three-factor model employs total capital without IT capital, total labour without IT labour, IT capital + IT labour, and GDP or revenue [25].

There were 2 reasons why RTS was chosen. First, the result of dynamic PAV from Eq. (1) to (4) could be used for data preparation. Second, it is an easy calculation as described in table 1 after the result of dynamic PAV was summarised in Table 4 and Table 5. So, it can be said that the RTS method could be continued without any complicated formula. The combination of RTS is as follows and Table 3:

- Decreasing RTS. It means that new production inputs exceed production result.
- Constant RTS. It means that new production inputs are proportionate to the production result.
Increasing RTS. It means that new production inputs produce bigger production result than production inputs.

**Table 3.** RTS Formulation for two factors and three factor models

<table>
<thead>
<tr>
<th>Combination of RTS</th>
<th>Two-factor model</th>
<th>Three-factor model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decreasing RTS</td>
<td>$\beta_1 + \beta_2 &lt; 1$</td>
<td>$\beta_1 + \beta_2 + \beta_3 &lt; 1$</td>
</tr>
<tr>
<td>Constant RTS</td>
<td>$\beta_1 + \beta_2 = 1$</td>
<td>$\beta_1 + \beta_2 + \beta_3 = 1$</td>
</tr>
<tr>
<td>Increasing RTS</td>
<td>$\beta_1 + \beta_2 &gt; 1$</td>
<td>$\beta_1 + \beta_2 + \beta_3 &gt; 1$</td>
</tr>
</tbody>
</table>

*Source:* devised by the authors.

### 4 Results and Discussions

#### 4.1 Result Estimations

As was argued above, this study contains an analysis of 2 models, and this stage will illustrate the two-factor and three factor models in results validating. Table 4 is estimation results that have been published for branchless banking.

**Table 4.** Unobserved Coefficient to Use Prediction method (non-Linier Least Square) in Branchless Banking

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Two-factor model</th>
<th>Three-factor model</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_0$</td>
<td>1.20245</td>
<td>2.76655</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>-0.94051</td>
<td>-3.09189</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>-160.523</td>
<td>-445.04</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>-31.3483</td>
<td>-1574.69</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>115.594</td>
<td>1632.3</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>-</td>
<td>792.023</td>
</tr>
</tbody>
</table>

*Source:* developed by the authors.

In addition, Table 5 is estimation results in downstream petroleum.

**Table 5.** Unobserved Coefficient to Use Prediction Model (Non-Linier Least Square) in Downstream Petroleum

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Two-factor model</th>
<th>Three-factor model</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_0$</td>
<td>-0.341907</td>
<td>1.37597</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>240.261</td>
<td>-244.409</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>8.31331</td>
<td>19.9813</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.0306634</td>
<td>0.40625</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.576009</td>
<td>0.385247</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>-</td>
<td>-2.8775</td>
</tr>
</tbody>
</table>

*Source:* developed by the authors.

These unobserved coefficients are meaningless until it can be interpreted. $\alpha_0$ and $\alpha_1$ are useful for the validation of the IT value model but it can be ignored for the data preparation research in the IT value model. Next, RTS is useful for data preparation.
4.2 RTS Results

After whole estimation results ($\alpha_0$, $\alpha_1$, $\beta_0$, $\beta_1$, $\beta_2$, $\beta_3$) have been received from Gretl statistical software package with the use of Eq (1) to Eq. (4), RTS is going to be performed. From table 4, a two-factor model ($\beta_1 + \beta_2$) will get 84.246, and a three-factor model ($\beta_1 + \beta_2 + \beta_3$) will have 849.633. Both models have the same category. Two-factor and three-factor models are increasing RTS.

From Table 5, a two-factor model ($\beta_1 + \beta_2$) will get 0.607. Meanwhile, a three-factor model ($\beta_1 + \beta_2 + \beta_3$) will have -2.086. Both models have the same category. Two-factor and three-factor models are decreasing RTS. Table 4 shows increasing RTS which means new production inputs produce bigger production result than production inputs. In other words, increasing RTS will bring more productive value for the industry. Besides that, table 5 is decreasing RTS which means new production inputs produce fewer production result than production inputs. It can be concluded that decreasing RTS will bring less production value for the industry.

5. Discussions

The result of the RTS method gives a better explanation in preparing data to validate it. The 2 (two) factor and 3 (three) factor models may be compared to bring the desired result based on table 1. From this research result in table 4, both models are increasing RTS. It means new production inputs produce bigger production result than production inputs. For comparison purpose, three-factor model value has bigger value than 2 (two) factor model value. Three-factor model value is 849.633. Two-factor model value is 84.246. The RTS measurement provides firms with strategic guidance on how to enhance their efficiency measures and international competitiveness by controlling their operational sizes and/or introducing technology innovation [7]. There is a reason why a three-factor model value is bigger than a two-factor model value though both models are increasing RTS. IT for branchless banking is more meaningful to deliver banking system through retail and postal office than without IT. From this perspective, a customer may feel more convenient with services. Besides that, a customer may be not familiar with branch bank requirement or perceive unsuitable to deal with a teller and branch staff. IT values is a key driver to success in Branchless banking industry. Branchless banking provides the use of IT for lowering channel cost and increasing customer convenience and trust [13]. In contrast, the bank may get good performance without IT, but bank’s performance may be lower than today’s performance with IT. The result of this research also confirms that the validation result in IT values model for branchless banking was positive. Based on the validity results showed that IT enhanced the performance.

Otherwise, the research result in table 5, both models are decreasing RTS. It means new production inputs produce little production result than production inputs. For comparison, a three-factor model value has a smaller value than a two-factor model value. A three-factor model value is -2.086, while a two-factor model value is 0.606724. There is a reason why a three-factor model value is smaller than two-factor model value. IT for downstream petroleum is just an enabler. The study concludes with a model for the downstream petroleum industry's total deregulation to eliminate impediment against the nation's economic growth and development [26]. The main production is still from oil and gas sales. Without IT, the
production is still running by the time [7]. The result of this research shows otherwise than the previous study for the validation of IT value model for downstream petroleum. This study shows that IT is just an enabler. However, the validation of IT value model showed that IT (three-factor model) had positive effect to the performance. It is different from branchless banking. Branchless banking is heavily dependent on IT. Without IT, branchless banking is just a concept.

6. Conclusion

The result of data preparation for the validation has found positive results. These results were achieved by the four targets. The first aim was achieved by applying the RTS method for data correction on a two-factor model (without IT). The second aim was achieved by comparing the previous research and this result on the degree of accuracy. Though, the previous research showed that IT had impact on GDP/revenue, this research explained otherwise. The third aim was achieved by using Eq (1) to Eq. (4) to generate table 4 and table 5. These formulas are common for many models. The final purpose was achieved by the developed methodology in section 3 (research methodology and data collection). This methodology can differentiate many data collections for many models.

Besides these four aims, another result is that the developed methodology for data preparation on validation of IT value model can also be applied on both industries. This result brings greater meaning in resolving data preparation problem during validation. As is known about data preparation for validation, it is hard to separate inputs and outputs for two-factor and three-factor models, but RTS may solve this problem. This research also shows that other industry may apply the methodology to correct the performance of a two-factor model.

The implication for further study is to explore the IT value model for branchless banking must be validated to verify whether it can be applied in the industry or is needed for further research. The validation checks the accuracy of the model’s representation to the branchless banking industry. Moreover, the validation result can ensure whether the IT value model for branchless banking can describe real-world phenomena. Many computers can simultaneously operate in a network for applying the proposed approach to various RAM/RTS measurements. That is another research task. In addition to the computational issue, the proposed approach needs to be applied to many real applications on RAM/RTS. Such real applications will document the practicality of this approach. Finally, it is hoped that this study makes a certain contribution to the RAM/RTS measurement.

7. Limitation of the Study

First, there are lots of methods for data preparation for validation, apart from RTS for validation, such as: SAS Enterprise Miner, Enterprise Miner and pattern-oriented modelling. In the research to follow, a comparison of SAS Enterprise Miner, Enterprise Miner and RTS will be performed.

Second, this research has already compared two models. The further research in data preparation may employ more models to see the better result of RTS. Third, the longer datasets should be studied to comprehend better results on preparation before data validation. Fourth, the forecast analysis of the longer datasets can be done in order to know that IT is capable to
enhance the industry performance.

**Author Contributions:** Conceptualization, T.H. and R.M.; methodology, T.H., R.M.; software, T.H, R.M; validation, T.H., R.M. and A.M.R.; formal analysis, T.H, R.H, A.M.R.; investigation, T.H, R.M.; resources, T.H, R.M.; data curation, T.H, R.M.; writing—original draft preparation, T.H.; writing—review and editing, T.H.; visualization, T.H, R.M.; supervision, T.H.; project administration, T.H.; funding acquisition, T.H, R.M, A.M.R. All authors have read and agreed to the published version of the manuscript.

**Funding:** not applicable.

**Data Availability Statement:** not applicable.

**Acknowledgements** not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.

**References**


*Taufik Hidayat, Rahutomo Mahardiko, and Ali Miftakhu Rosyad*

*Virtual Economics, Vol. 6, No. 2, 2023*


