# Pitfalls of Analyzing a Cross-company Dataset of Software Maintenance and Support

Masateru Tsunoda
Kinki University
Osaka, Japan
tsunoda@info.kindai.ac.jp

Kenichi Ono
Kinki University
Osaka, Japan
1110370048f@kindai.ac.jp

***Abstract*— It is important to establish a benchmark of work efficiency for software maintenance and support. For maintenance and support service providers, the benchmarking is the basis for improvement of their work. For the service purchasers, it is useful to check work efficiency of contracted service provider. To establish a benchmark of work efficiency for software development activities, a cross-company dataset is often used. Data points included in it are collected from various organizations. ISBSG (International Software Benchmarking Standards Group) builds the cross-company dataset of software maintenance and support. In the analysis, we found some pitfalls when one analyzes ISBSG software maintenance and support dataset. If the pitfalls are ignored, spurious relationships would be found. In this paper, we showed some pitfalls, and how to avoid it. It would be very useful for researchers, because ISBSG software maintenance and support dataset may be widely used in the near future.**

***Keywords*— spurious relationships; productivity; cross-company dataset; benchmarking; adjusted variance explained**

## I. INTRODUCTION

Recently, a number of software users contract with software developers for maintenance and support of enterprise software. After software is released, software maintenance and support is necessary to sustain availability of the software. Software maintenance does not mean only removing faults found after software release. Software needs extensions or modifications of its functions due to changes in a business environment, and software maintenance also indicates them. ISO/IEC 14764 [8] classifies software maintenance into followings:

- Corrective maintenance: modifications of faults found after software release.
- Preventive maintenance: corrective modifications before potential faults become actual faults, after software release.
- Adaptive maintenance: modifications to keep software availability against environmental changing after software release.
- Perfective maintenance: modifications for conservation or improvement of software performance or maintainability after software release.

Also, software support becomes more important for both software users and developers than ever. Activities of software support consist of problem investigation, queries and quick service, and user help and advice [7]. The activities are the part of IT operations explained in ITIL (Information Technology Infrastructure Library) [3].

It is important to establish a benchmark (reference values to compare an organization's work efficiency with others [11]) of work efficiency for software maintenance and support. For maintenance and support service providers, the benchmarking is the basis for improvement of their work. The improvement will enhance price competitiveness of the providers. For the service purchasers, it is useful to check work efficiency of contracted service provider. If the efficiency is low, the purchaser should consider to contract other providers to suppress maintenance and support cost.

To establish a benchmark of work efficiency for software development activities, a cross-company dataset is often used. Data points included in it are collected from various organizations. In regard to software development activities, ISBSG (International Software Benchmarking Standards Group) [5] builds the dataset, and it is widely used for analysis [11][16] and effort estimation [12][14], in addition to benchmarking.

ISBSG also builds a dataset of software maintenance and support. Data points included in it were collected from various companies. A cross-company dataset of software maintenance and support is rare. So, like software development dataset, the dataset of software maintenance and support may be widely used for analyze, estimation, and benchmarking in the near future. However, as long as we know, the dataset is not analyzed in detail.

First, we used the dataset, to analyze relationships between work efficiency and attributes such as hardware. This is because if there are some strong relationships between them, stratification is needed before benchmarking. There are some researches which analyzed relationships between work efficiency and attributes [9] on software maintenance or support. However, there are few researches which used cross-company dataset. Replicated study is important on software engineering research area. So, we tried to analyze the dataset.

However, in the analysis, we found some pitfalls when one analyzes ISBSG software maintenance and support dataset. If the pitfalls are ignored, spurious relationships would be found. In this paper, we showed some pitfalls, and how to avoid it. It would be very useful for researchers, because ISBSG software maintenance and support dataset may be widely used in the

TABLE I. DESCRIPTION OF ATTRIBUTES

| Attribute | Scale | Description |
|---|---|---|
| Data quality rating | Nominal | Data quality of a data point reviewed by ISBSG. The value is A, B, C, or D. |
| Size Approach | Nominal | Function point method used to measure application size of a data point |
| Application size | Numerical | Application size of a data point measured by a function point method |
| Calculated maintenance hours | Numerical | Total maintenance hours for one year. "Calculated" means normalizing the value when the measurement duration is not one year. |
| Calculated support hours | Numerical | Total support hours for one year. "Calculated" means normalizing the value. |
| Maintenance FP rate | Numerical | Signifies work efficiency of maintenance activities. It is calculated by dividing calculated maintenance hours by application size |
| Support FP rate | Numerical | Signifies work efficiency of support activities. It is calculated by dividing calculated support hours by application size |
| Application set | Nominal | Identifier signifying where a data point was collected from. When data points were submitted from an organization, same number is given to application set of them. |
| Industry sector | Nominal | Industry sector of the organization where software is maintained and supported. |
| Hardware | Nominal | Hardware on which software is maintained and supported. |
| Programming language | Nominal | Programming language mainly used in software which is maintained and supported. |

TABLE II. CORRELATION COEFFICIENTS BETWEEN SIZE AND HOURS

| | | Calculated maintenance hours | Calculated support hours | Application size |
|---|---|---|---|---|
| **Calculated maintenance hours** | $\rho$ | 1.00 | 0.58 | 0.61 |
| | **p-value** | . | *0.00* | *0.00* |
| | **Number of data points** | 105 | 63 | 105 |
| **Calculated support hours** | $\rho$ | 0.58 | 1.00 | 0.64 |
| | **p-value** | *0.00* | . | *0.00* |
| | **Number of data points** | 63 | 68 | 68 |
| **Application size** | $\rho$ | 0.61 | 0.64 | 1.00 |
| | **p-value** | *0.00* | *0.00* | . |
| | **Number of data points** | 105 | 68 | 113 |

near future. So, main contribution of our research is to show the pitfalls.

In Section 2, we explain ISBSG software maintenance and support dataset, and research method. Then, in Section 3, we analyze the relationships between attributes. Section 4 introduces related works. In the end, Section 5 concludes the paper with a summary.

## II. ANALYSIS METHOD

### A. Dataset

In the analysis, we used cross-company dataset of software maintenance and support, collected by ISBSG (International Software Benchmarking Standards Group) [5]. The dataset includes 478 data points and 103 attributes, and was collected from 19 organizations. There are some versions of the dataset, and we used Release 4, published on March 2010 [7]. Attributes used in the analysis are shown in Table I. Note that ISBSG does not explain pitfalls shown in this paper.

In the analysis, we excluded data points in which data quality rating was C or D. Data quality rating is evaluated by ISBSG and when the value is C or D, data quality of the data point is low. So, we excluded them to enhance reliability of the analysis. When the software development dataset collected by ISBSG [6] is analyzed, such data is often eliminated [10].

Also, we selected data points whose sizing methods of FP are same. The sizing method is used to measure FP. When the method is different, FP becomes different even if same software is measured. So, we selected data points, considering the method. In the preliminary analysis described in section III, we selected data points whose sizing method is IFPUG FP method, because they were selected in the research [10] explained above. In other analyses, we used data points whose sizing method is other than IFPUG FP method.

### B. Relationship Analysis

To see strength of a relationship of ratio scale attributes, we used Spearman's rank correlation coefficient to avoid influence of outliers. In what follows, "correlation" and $\rho$ indicate the Spearman correlation. In the analysis, we set the significance level at 0.05. In what follows, letters in italics indicate p-values < 0.05.

To analyze a relationship of a ratio scale attribute and a nominal scale attribute, we used ANOVA. Before applying ANOVA, ratio scale attributes were log transformed, to avoid influence of outliers. When p-value is significant, it means there is a relationship between an attribute and FP rate. Adjusted variance explained ($\omega^2$) indicates the strength of the relationship. It is calculated using the following equation [19].

$$\omega^2 = \frac{SSB - (k-1)MSE}{SST + MSE} \quad (1)$$

In the equation, $SSB$ is the sum of squares between groups, $SST$ is the sum of squares total, $MSE$ is the mean square error, and $k$ is the number of groups. A larger value indicates there is a stronger relationship.

## III. ANALYSIS RESULTS

### A. Preliminary Analysis

In the analysis, we focused on maintenance FP rate and support FP rate. The former denotes the efficiency of the maintenance of software applications, and the latter does the efficiency of the support. The efficiency directly relates to the maintenance cost and the support cost. So, they are very important for vendors and users. The calculated maintenance hours and calculated support hours are numerator of the FP rates, and application size is denominator of the rates. When analyzing software development data, productivity is often analyzed [2][13]. Productivity denotes efficiency of software development. The numerator is FP, and the denominator is effort. That is, FP rate is similar to the reciprocal of productivity.

As preliminary analysis, to understand the characteristics of maintenance FP rate and support FP rate, we analyzed relationships between calculated maintenance hours, calculated support hours, and application size using the correlation coefficient. As explained in section II, we selected data points whose sizing method is IFPUG FP method.

The correlation coefficients of the relationships are shown in Table II. Application size has positive relationships to calculated maintenance hours and calculated support hours. The relationships are similar to the relationship between FP and effort on software development datasets. The relationships shown in Table II were not very strong (The correlation coefficients were around 0.6). That is, although larger application needs larger maintenance and support hours, compared with smaller application, the hors do not increase linearly. This is supposed to be one of characteristics of software maintenance and support.

Note that this is preliminary analysis, and the results are not surprising. However, we think this analysis is indispensable because there are few researches which analyzed software application maintenance and support, and hence the results were not necessarily obvious.

### *B. Maintenance and Support FP rates*

#### *1) Analysis Procedure*

We analyzed distributions of maintenance FP rate and support FP rate, considering application set. Intuitively, application set is an identifier which signifies where a data point was collected from. When data points were submitted from an organization at a certain point, same number is given to application set of them. So, when a value of application set is same between two data points, they are collected from same organization. Generally, an identifier like application set is not included in a cross-company dataset to keep anonymity of it.

To see the diversity of the dataset, we checked number of data points of each application set. The dataset was filtered based on data quality rating and sizing method, as explained in section II. The filtering may make number of cases of each application set biased. Although the dataset is cross-company dataset, i.e., there are 27 application sets collected from 19 organizations in the dataset, most of the filtered dataset may consist of few organizations' data.

To identify data points whose reliability is low, we compared the median of maintenance and support FP rate between application set. That is, when the median was very different for each application set, we regarded reliability of the application set as low. We assumed that reliability of the data point depends on organizations, because the data collection rule may be different for different organizations, and the rule affects the reliability. Hence, we considered reliability of data points by application set.

TABLE III. FP RATES ON EACH APPLICATION SET

| FP rate | Application set | Number of data points | Median |
|---|---|---|---|
| Maintenance | 5 | 2 | 0.55 |
| | 11 | 1 | 6.3 |
| | 12 | 4 | 21.55 |
| | 13 | 23 | 1.4 |
| | 15 | 2 | 30.1 |
| | 20 | 34 | 0.85 |
| | 30 | 31 | 0.5 |
| Support | 5 | 2 | 0.15 |
| | 11 | 1 | 1.7 |
| | 15 | 2 | 6.3 |
| | 20 | 29 | 0.4 |
| | 30 | 34 | 0.3 |

TABLE IV. ADJUSTED VARIANCE EXPLAINED OF APPLICATION SET

| FP rate | Application set | Number of data points | $\omega^2$ | p-value |
|---|---|---|---|---|
| Maintenance | all | 97 | 0.50 | *0.00* |
| | 5, 11, 13, 20, 30 | 91 | 0.35 | *0.00* |
| | 13, 20, 30 | 88 | 0.17 | *0.00* |
| Support | all | 68 | 0.12 | *0.02* |
| | 5, 11, 20, 30 | 66 | 0.10 | *0.03* |
| | 20, 30 | 63 | 0.02 | 0.18 |

To see influence of application set to maintenance and support FP rate, we calculated adjusted variance explained of application set. When the value is large, the application set affects maintenance and support FP rate. If so, when one analyzes relationships between an attribute and the rates, he/she should consider the influence of application set. Since the relationship may be spurious, and the application set may affect the relationship actually.

#### *2) Results on IFPUG FP Method*

We selected data points whose sizing method is IFPUG FP method. Table III shows number of data points and median of maintenance and support FP rate on each application set. On maintenance FP rate, on the subset of the dataset, application set 13, 20 and 30 were dominant. On support FP rate, application set 20 and 30 were dominant. That is, although the dataset is cross-company dataset, the filtering explained in section II made number of data points of each application set biased, and the subset had little diversity as a result.

On maintenance FP rate, compared with application set 13, 20 and 30, medians of maintenance FP rate on application set 12 and 15 were very large, although their data quality ratings were A or B, and their sizing methods were IFPUG FP method. Likewise, on support FP rate, median of support FP rate on application set 15 was larger than others. Based on the results, we regarded that reliability of application set 12 and 15 was not high.

We calculated adjusted variance explained ($\omega^2$) of application set for maintenance and support FP rate. The values were calculated using all data points, eliminating data points

TABLE IX. NUMBER OF DATA POINTS ON EACH SIZE APPROACH

| FP rate | Size approach | Number of data points | Median | Application set |
|---|---|---|---|---|
| Maintenance | FiSMA | 86 | 1.25 | 19 |
| | IFPUG | 97 | 1.00 | 5, 11, 12, 13, 15, 20, 30 |
| | NESMA | 12 | 0.80 | 3, 8 |
| Support FP | FiSMA | 29 | 0.30 | 19 |
| | IFPUG | 68 | 0.35 | 5, 11, 15, 20, 30 |
| | NESMA | 9 | 0.10 | 3, 8 |

whose reliability is not high (the set 12 and 15), and using application set in which number of data point is not small (the set 13, 20 and 30). The values are shown in Table IV. Focusing on maintenance FP rate, when application set 13, 20 and 30 were used, $\omega^2$ was the smallest. However, influence of application set was not ignorable, because p-value was smaller than 0.05, i.e., application set relates to maintenance FP rate significantly. Focusing on support FP rate, when application set 20 and 30 is used, influence of application set was ignorable, since p-value was larger than 0.05.

### 3) *Results on other FP Methods*

When sizing method is different between two data points, FP rate should not be compared. Since application size is denominator of FP rates, and same values do not mean same application sizes in the case. For example, it should be avoided to compare FP rate based on IFPUG FP method with FP rate based on FiSMA FP method. However, it is clean that a dataset is divided based on sizing method, and analyze relationships on each subset. This is because it does not compare FP rates based on different sizing methods.

Table V shows number of data points on each size approach. Note that FiSMA FP was calculated based on LOC, and it is often called as backfired. NESMA FP method had too small on both FP rates to analyze relationships between attributes. FiSMA FP method had 86 data points on maintenance FP rate, and 29 data points on support FP rate. They are enough to analyze relationships. So, in addition to IFPUG FP method, we used data points whose sizing method is FiSMA FP method, dividing dataset based on sizing method. On FiSMA FP method, application set was 19 only.

## C. *Industry Sector*

We analyzed relationships between industry sector and maintenance and support FP rates, considering application set. We did not use data points whose size approach is FiSMA FP method, because their industrial sectors were banking only. Table VI shows $\omega^2$ of industry sector for maintenance and support FP rate. The values were calculated using all data points, eliminating data points whose reliability is not high (the set 12 and 15), and using application set in which number of data point is not small (the set 13, 20 and 30). On support FP rate, when application set 20 and 30 were used, there were only one industry sector, and it is not able to calculate $\omega^2$. On maintenance FP rate, $\omega^2$ was significant in all cases, and on support FP rate, $\omega^2$ was not significant in all cases.

TABLE V. ADJUSTED VARIANCE EXPLAINED OF INDUSTRY SECTOR

| FP rate | Application set | Number of data points | $\omega^2$ | p-value |
|---|---|---|---|---|
| Maintenance | all | 64 | 0.30 | *0.00* |
| | 5, 11, 13, 20, 30 | 59 | 0.15 | *0.01* |
| | 13, 20, 30 | 56 | 0.07 | *0.03* |
| Support | all | 31 | 0.16 | 0.06 |
| | 5, 11, 20, 30 | 29 | 0.08 | 0.13 |
| | 20, 30 | 26 | - | - |

TABLE VI. CROSSTAB OF INDUSTRY SECTOR AND APPLICATION SET

| FP rate | Industry sector \ Application set | 5 | 11 | 12 | 13 | 15 | 20 |
|---|---|---|---|---|---|---|---|
| Maintenance | Communication | 0 | 1 | 0 | 0 | 0 | 0 |
| | Electronics & computers | 0 | 0 | 0 | 0 | 1 | 0 |
| | Financial | 0 | 0 | 4 | 23 | 0 | 0 |
| | Service industry | 2 | 0 | 0 | 0 | 0 | 0 |
| | Wholesale & retail | 0 | 0 | 0 | 0 | 0 | 33 |
| Support | Communication | 0 | 1 | - | - | 0 | 0 |
| | Electronics & computers | 0 | 0 | - | - | 2 | 0 |
| | Service industry | 2 | 0 | - | - | 0 | 0 |
| | Wholesale & retail | 0 | 0 | - | - | 0 | 26 |

TABLE VII. ADJUSTED VARIANCE EXPLAINED OF HARDWARE

| FP rate | Application set | Number of data points | $\omega^2$ | p-value |
|---|---|---|---|---|
| Maintenance | all | 59 | 0.168 | *0.025* |
| | 5, 11, 13, 20, 30 | 58 | 0.113 | 0.068 |
| | 13, 20, 30 | 58 | 0.113 | 0.068 |
| Support | all | 57 | 0.078 | 0.166 |
| | 5, 11, 20, 30 | 56 | 0.092 | 0.124 |
| | 20, 30 | 55 | 0.073 | 0.158 |

TABLE VIII. CROSSTAB OF HARDWARE AND APPLICATION SET

| FP rate | Hardware \ Application set | 11 | 15 | 20 | 30 |
|---|---|---|---|---|---|
| Maintenance | AS400 | 0 | - | 2 | 0 |
| | Combination | 0 | - | 1 | 0 |
| | Mainframe | 0 | - | 9 | 1 |
| | Midrange | 0 | - | 0 | 17 |
| | PC | 0 | - | 5 | 0 |
| | PC or client/server | 0 | - | 0 | 10 |
| | Sun | 1 | - | 0 | 0 |
| | Unix | 0 | - | 7 | 0 |
| | VMS | 0 | - | 6 | 0 |
| Support | AS400 | 0 | 0 | 1 | 0 |
| | Combination | 0 | 0 | 1 | 0 |
| | Mainframe | 0 | 0 | 7 | 1 |
| | Midrange | 0 | 0 | 0 | 17 |
| | PC | 0 | 0 | 6 | 0 |
| | PC or client/server | 0 | 0 | 0 | 14 |
| | Sun | 1 | 0 | 0 | 0 |
| | Unix | 0 | 0 | 4 | 0 |
| | VMS | 0 | 0 | 4 | 0 |
| | Websphere | 0 | 1 | 0 | 0 |

Next, we focused on relationships between industry sector and application set, using a crosstab. In Table VII, values in the cells indicate number of data points which were classified by industry sector and application set. Except for financial on maintenance FP rate, each industry sector has one application set. So, the relationship was very strong, and it is difficult to distinguish influence of industry sector to FP rates from application set. The reason why relationship between industry sector and support FP rate was weak would be most of industry sector was wholesale & retail.

### *D. Hardware*

We analyzed relationships between hardware and maintenance and support FP rates. We did not use data points whose size approach is FiSMA FP method, because their hardware were IBM only, except for two data points. Table VIII shows $\omega^2$ of hardware for maintenance and support FP rate. The values were calculated using all data points, eliminating data points whose reliability is not high, and using application set in which number of data point is not small. On maintenance FP rate, $\omega^2$ was significant when all application set were used, and on support FP rate, $\omega^2$ was not significant in all cases.

Next, we focused on relationships between hardware and application set. As shown in Table IX, application set which has many data points (the set 20 and 30) include different types of hardware. So, we can eliminate influence of application set from the analysis, stratifying the dataset by the set.

Table X shows $\omega^2$ of hardware stratified by the application set. It is regarded as the result obtained from two single company dataset. On maintenance FP rate, although $\omega^2$ was not significant on both application set, $\omega^2$ on application set 30 is relatively large. So, hardware may affect maintenance FP rate in some situations. On support FP rate, $\omega^2$ was not significant on both application set. The result suggests that hardware does not affect support FP rate.

### *E. Programming Language*

We analyzed relationships between programming language and maintenance and support FP rates. We did not use data points whose size approach is IFPUG FP method, because programming language is recorded on only eight data points when their sizing method is IFPUG FP method. Table XI shows $\omega^2$ of programming language for maintenance and support FP rate. We used data points whose sizing method is FiSMA FP method, and there is only one application set. That is, it is regarded as the result obtained from only one single company dataset, in spite of using cross-company dataset. On maintenance FP rate, $\omega^2$ was significant, although the value is not large. On support FP rate, $\omega^2$ was not significant.

Programming language often collates with hardware. As shown in Table XII, when size approach was FiSMA FP method, their hardware was almost IBM. That is, influence of hardware to FP rate is ignorable on the analysis. From the result, we think programming language affects maintenance FP rates to some extent.

### *F. Application Size*

We analyzed relationships between application size and maintenance and support FP rates, considering application set. Table XIII shows $\rho$ of application size to maintenance and support FP rate. The values were calculated using data points

TABLE X. ADJUSTED VARIANCE EXPLAINED OF HARDWARE STRATIFIED BY THE APPLICATION SET

| FP rate | Application set | Number of data points | $\omega^2$ | p-value |
|---|---|---|---|---|
| Maintenance | 20 | 31 | -0.10 | 0.79 |
| | 30 | 28 | 0.15 | 0.05 |
| Support | 20 | 23 | 0.14 | 0.18 |
| | 30 | 32 | -0.01 | 0.45 |

TABLE XI. ADJUSTED VARIANCE EXPLAINED OF LANGUAGE

| FP rate | Application set | Number of data points | $\omega^2$ | p-value |
|---|---|---|---|---|
| Maintenance | 19 | 86 | 0.06 | *0.05* |
| Support | 19 | 25 | 0.02 | 0.32 |

TABLE XII. CROSSTAB OF HARDWARE AND LANGUAGE

| Hardware \ Language | COBOL | EASY | JCL | TELON |
|---|---|---|---|---|
| IBM | 55 | 1 | 13 | 20 |
| S/88 | 0 | 0 | 0 | 1 |
| TAND | 1 | 0 | 0 | 0 |

TABLE XIII. CORRELATION COEFFICIENTS BETWEEN SIZE AND FP RATE

| FP rate | Application set | Number of data points | $\rho$ | p-value |
|---|---|---|---|---|
| Maintenance | 19 (FiSMA) | -0.46 | 86 | *0.00* |
| | all | -0.14 | 97 | 0.18 |
| | 5, 11, 13, 20, 30 | -0.03 | 91 | 0.75 |
| | 13, 20, 30 | -0.03 | 88 | 0.80 |
| Support | 19 (FiSMA) | -0.48 | 29 | *0.01* |
| | all | -0.39 | 68 | *0.00* |
| | 5, 11, 20, 30 | -0.41 | 66 | *0.00* |
| | 20, 30 | -0.41 | 63 | *0.00* |

whose size approach is FiSMA FP method (application set 19). The rest of the values were calculated using data points whose size approach is IFPUG FP method. Application set are classified based on their reliability and number of data points.

On maintenance FP rate, $\rho$ was significant when application set was 19, and on support FP rate, $\rho$ was significant in all cases. Based on the results, we concluded that application size is negatively affected support FP rate, even when application set is considered. Application size is sometimes negatively affected maintenance FP rate, but it has no relationship to it in most cases.

## IV. RELATED WORK

Some researches analyzed work efficiency factors on software maintenance. Jørgensen [9] analyzed software company dataset, and showed that work efficiency is not affected by the number of base modules and programming language. Ahn et al. [1] used variables which are similar to the productivity factors in a software maintenance effort estimation model. Although they used dataset collected from few companies, the number of organizations included in the datasets is less than ISBSG dataset.

There are few reports or researches which analyzed cross-company software maintenance dataset. Japan Users Association of Information Systems (JUAS) and Ministry of

Economy, Trade and Industry used the cross-company dataset, and showed work efficiency (maintenance cases per engineer) stratified by business sector [4]. Tsunoda et al. [17] analyzed relationships between attributes and work efficiency (number of modified modules per engineer), using analyzed a cross-company dataset. However, definitions of work efficiency in them are rough, compared with the definition in this research. That is, they do not use FP and work hour. So, more replicated study is needed to establish a benchmark of work efficiency for software maintenance.

Software support is the part of IT operations explained in ITIL (Information Technology Infrastructure Library) [3]. While there are some studies treating IT operations, they mainly focus on the process of IT operations. For instance, Pollard et al. [15] performed case study of the United States and Australian companies' IT operations, and identified some critical success factors (CSF) for successful ITIL implementations. Tsunoda et al. [18] analyzed relationships between attributes and number of staff, using analyzed a cross-company dataset. They do not analyze relationships to work efficiency defined by FP and work hour, and this is one of major differences of our research.

## V. CONCLUSIONS

This paper showed some pitfalls when one analyzes ISBSG software maintenance and support dataset. The dataset has attributes "application set." Intuitively, application set is an identifier which signifies where a data point was collected from. Application set had strong relationships to work efficiency of software maintenance and support. So, if one analyzes the dataset without considering application set, he/she may found spurious relationships. For example, as shown in this research, industry sector seemed to have strong relationships to maintenance FP rate when all data point were used (i.e., application set was ignored). However, it is not sure that industry sector has strong relationships to maintenance FP rate when application set was considered. In addition, there are some low reliable data points, although their data quality rating was high. They should remove before analyzing. One should analyze the dataset considering them.

## ACKNOWLEDGMENT

This research was partially supported by the Japan Ministry of Education, Science, Sports, and Culture [Grant-in-Aid for Scientific Research (C) (No. 25330090)].

## REFERENCES

[1] Y. Ahn, J. Suh, S. Kim, and H. Kim, "The software maintenance project effort estimation model based on function points," Journal of Software Maintenance: Research and Practice, Vol. 15, No. 2, pp. 71-85, 2003.

[2] J. Blackburn, G. Scudder, and L. Wassenhove, "Improving Speed and Productivity of Software Development: A Global Survey of Software Developers," IEEE Transactions on Software Engineering, Vol. 22, No. 12, pp. 875-885, 1996.

[3] J. Bon, ed., Foundations of IT Service Management: based on ITIL, Van Haren Publishing, 2005.

[4] Japan Users Association of Information Systems 2008. Software Metric Survey 2008. Japan Users Association of Information Systems.

[5] International Software Benchmarking Standards Group, https://www.isbsg.org/

[6] International Software Benchmarking Standards Group (ISBSG), ISBSG Estimating, Benchmarking and Research Suite Release 10, ISBSG, 2007.

[7] International Software Benchmarking Standards Group (ISBSG), ISBSG Maintenance & Support Data Suite Release 4, ISBSG, 2010.

[8] ISO/IEC 14764:2006, Software Engineering - Software Life Cycle Processes - Maintenance, International Organization for Standardization, 2006.

[9] M. Jørgensen, "Experience With the Accuracy of Software Maintenance Task Effort Prediction Models," IEEE Transactions on Software Engineering, Vol. 21, No. 8, pp. 674-681, 1995.

[10] C. Lokan, and E. Mendes, "Cross-company and single-company effort models using the ISBSG Database: a further replicated study," Proc. international symposium on Empirical software engineering (ISESE 2006), pp. 75-84, Rio de Janeiro, Brazil, 2006.

[11] C. Lokan, T. Wright, P. Hill, and M. Stringer, "Organizational Benchmarking Using the ISBSG Data Repository," IEEE Software, Vol. 18, No. 5, pp. 26-32, 2001.

[12] C. Lopez-Martin, C. Isaza, and A. Chavoya, "Software development effort prediction of industrial projects applying a general regression neural network," Empirical Software Engineering, Vol. 17, No. 6, pp. 738-756, 2012.

[13] K. Maxwell, L. Wassenhove, and S. Dutta, "Software Development Productivity of European Space, Military, and Industrial Applications," IEEE Transactions on Software Engineering, Vol. 22, No. 10, pp. 706-718, 1996.

[14] E. Mendes, and C. Lokan, "Replicating studies on cross- vs single-company effort models using the ISBSG Database," Empirical Software Engineering, Vol. 13, No. 1, pp. 3-37, 2008.

[15] C. Pollard, and A. Cater-Steel, "Justifications, Strategies, and Critical Success Factors in Successful ITIL Implementations in U.S. and Australian Companies: An Exploratory Study," Information System Management, vol.26, no.2, pp.164-175, 2009.

[16] D. Rodríguez, M. Sicilia, E. García, and R. Harrison, "Empirical findings on team size and productivity in software development," Journal of Systems and Software, Vol. 85, No. 3, pp. 562-570, 2012.

[17] M. Tsunoda, A. Monden, K. Matsumoto, and T. Oshino, "Analysis of Software Maintenance Efficiency Focused on Process Standardization," Proc. of International Workshop on Empirical Software Engineering in Practice (IWESEP), pp.3-8, Nara, Japan, 2011.

[18] M. Tsunoda, A. Monden, K. Matsumoto, A. Takahashi, and T. Oshino, "An Empirical Analysis of Information Technology Operations Cost," Proc. International Workshop on Software Measurement (IWSM/Metrikon/Mensura), pp.571-585, Stuttgart, Germany, 2010.

[19] B. Winer, D. Brown, and K. Michels, Statistical Principles in Experimental Design (3rd edition), McGraw-Hill, 1991.