2022 48th Euromicro Conference on Software Engineering and Advanced Applications (SEAA) | 978-1-6654-6152-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/SEAA56994.2022.00044

Effort Prediction with Limited Data: A Case Study for Data Warehouse Projects

Hüseyin Ünlü
Computer Engineering Department
Izmir Institute of Technology
Izmir, Turkey
huseyinunlu@iyte.edu.tr

Ali Yıldız
Computer Engineering Department
Izmir Institute of Technology
Bilgi Grubu
Izmir, Turkey
ali.yildiz@bg.com.tr

Onur Demirörs
Computer Engineering Department
Izmir Institute of Technology
Izmir, Turkey
onurdemirors@iyte.edu.tr

***Abstract*— Organizations may create a sustainable competitive advantage against competitors by using data warehouse systems with which they can assess the current status of their operations at any moment. They can analyze trends and connections using up-to-date data. However, data warehouse projects tend to fail more often than other projects as it can be tough to estimate the effort required to build a data warehouse system. Functional size measurement is one of the methods used as an input for estimating the amount of work in a software project. In this study, we formed a measurement basis for DWH projects in an organization based on the COSMIC Functional Size Measurement Method. We mapped COSMIC rules on two different architectures used for DWH projects in the organization and measured the size of the projects. We calculated the productivity of the projects and compared them with the organization's previous projects and DWH projects in the ISBSG repository. We could not create an organization-wide effort estimation model as we had a limited number of projects. As an alternative, we evaluated the success of effort estimation using DWH projects in the ISBSG repository. We also reported the challenges we faced during the size measurement process.**

Keywords—data warehouse, size measurement, effort estimation, COSMIC, ISO/IEC 19761

I. INTRODUCTION

Today, data is one of the essential cores of modern systems for decision-making. Organizations create a sustainable competitive advantage against competitors by using data with which they can assess the current status of their operations at any moment and with which they can analyze trends and connections using up-to-date data [1]. Data warehouse systems (DWH) are used for this purpose. Inmon, who is one of the pioneers of DWH, describes a data warehouse as "*a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process*"[2].

Although the term "data warehouse" appeared during the late 80s, the official year of birth of DWs is considered to be 1992, when Inmon defined DWH [3]. After a few years, Kimball introduced star schema as the primary solution for modeling multidimensional data on relational DWHs [4]. Since then, data warehouse systems have progressively emerged in both industry and academy.

For a software project to be considered successful, it is expected not only to meet the needs of the customers but also to have it finalized on time and within the expected budget [5]. However, it can be tough to estimate the effort required to build a data warehouse system [6]. Thus, data warehouse projects tend to fail more often than other projects.

Functional size measurement is one of the methods used as an input for estimating the amount of work in a software project. COSMIC [7] is one of the functional size measurement methods, which is ISO certified as well, that has been widely used to size different types of software projects [8]–[11].

In 2018, COSMIC proposed a "Guideline for sizing Data Warehouse and Big Data Software" [12]. This guideline describes the measurement process over two types of DWH architectures: Kimball and Inmon.

In this study, we performed a case study to size DWH projects using the COSMIC DWH Guideline. We mapped the rules described in the manual to two different DWH architectures applied in the organization. We calculated the productivity of the projects using the recorded effort and measured size. Then, we compared the productivity among the previous projects of the organization and data warehouse projects included in the ISBSG repository [13]. We also reported the challenges we faced during the size measurement process.

Organizations may not have enough historical data to construct an organization-wide effort estimation model. Alternatively, organizations may use ISBSG data to build estimation models [14]. However, the culture and experience of each organization differ. Therefore, organizations should not rely only on the estimations based on the ISBSG repository in critical projects. It will be helpful for effort estimation to use the ISBSG repository in cases where the organization does not have effort data about past projects. In our case, the number of DWH projects was not enough to construct an effort estimation model; thus, we evaluated the success of effort estimation using DWH projects in the ISBSG repository as an alternative.

The remaining of this paper is structured as follows: Section II summarizes the COSMIC DWH manual and mentions the related work in the literature. Section III explains the implementation of the measurement model. Section IV describes the mapped measurement guideline for two different architectures of the organization. Section V gives the measurement results. Section VI discusses our findings, and Section VII concludes the study by stating the further studies.

II. BACKGROUND AND RELATED WORK

This section summarizes the measurement rules mentioned in the "Guideline for Sizing Data Warehouse and Big Data Software" [1] by COSMIC and then summarizes the related work in the area.

978-1-6654-6152-8/22/$31.00 ©2022 IEEE
DOI 10.1109/SEAA56994.2022.00044

A. Background

"Guideline for Sizing Data Warehouse and Big Data Software" [12] describes two data warehouse architectures: Inmon and Kimball. Inmon type architecture includes 5 ETL (Extract, Transform and Load) sub-systems: data source area, data staging area, data warehouse area, data mart area, and business intelligence area. On the other hand, the Kimball type includes four sub-systems, not the data mart area. In the Kimball type, a data warehouse system is a collection of data marts.

"ETL" are the sub-systems of a data warehouse system that Extract, Transform, and Load data from one stage to the next. ETL sub-systems extract data from data sources, cleanse the data, perform data transformations, load the target data warehouse, and load the data marts.

There are two types of processing for ETLs: stream and batch. Stream-processing is the simplest. Each ETL sub-system receives data as Entry data movements, process them, and passes on the processed data as Exits to the next ETL sub-system. With batch processing, there are two differences from stream processing. First, assuming the data available at any stage to an ETL sub-system is stored persistently, the ETL sub-system will obtain this data by reading data movements, processing them, and then making the data available for the next stage by writing data movements. Second, every functional process must be started by a triggering Entry, regardless of the processing mode. Further explanations can be found in the guideline.

B. Related Work

In the literature, a few studies attempt to measure the size of DWH projects. Santillo [15] offered to apply Function Point Analysis (FPA) for DWH projects. He emphasized the relevant counting principles to measure the size of DWH projects. However, he stated that COSMIC FPA would perform better because of the layer concept and that the size of individual functions is not cut-off by the maximum size of a function.

Heeringen [6] proposed to use COSMIC for sizing DWH projects. The measurement method includes three phases: preparation, mapping, and measurement, similar to COSMIC manuals. However, the study does not have any case to measure the size of a DWH project.

Rasool and Malik [16] developed an effort estimation model using Forward Stepwise Regression with a dataset comprising 220 industrial ETL projects from five different software houses. They identified six variables: number of different types of sources used for data extraction, number of tables used for storing data, prior experience in developing similar ETL projects, level of difficulty in transforming source data, the degree of documentation, and suitability of source data for target systems, and the number of hierarchies representing levels of detail in the data. After eliminating 20 outliers, they achieved 0.16 MMRE and 81.16% PRED(25) over the estimation.

Literature review showed us that the number of studies attempting to measure DWH projects' size is limited. On the other hand, COSMIC provides a detailed guideline for measuring the size of DWH projects. However, we could not find any practical usage of this manual. Thus, in this study, we aim to apply this manual to the measurement of DWH projects and report our experience during the process.

III. Implementation of the Measurement model

In this study, our goal was to form a basis for the size measurement of DWH projects in the organization. First, we analyzed the requirement analysis and design documents of DWH projects. We had a total of four documents from two different directorates of the same organization. Directorate X developed projects A and B, and Directorate Y developed Project C, Phase 1, and Phase 2. We observed that the documents lacked the technical information required for the direct measurement. Consequently, we organized a meeting with the organization, including Directorate X and Directorate Y teams. We presented them with two different architectures (Inmon and Kimball) in the manual to understand which architecture they adopted. However, we analyzed that the architectures of the DWH projects are not precisely the same as the architectures described in the manual. Thus, we have to further explore the architectures the directorates used in their projects.

Next, we modified the rules on the COSMIC DWH Manual regarding the architectures applied by the directorates (see Section IV). Then, we performed the measurements based on the modified rules. After finalizing the measurements, we organized another meeting with the organization to discuss our measurement. We also took the recorded effort of the projects during the meeting. After the meeting, we calculated the productivity of the projects based on measured size and recorded effort. As the number of projects was limited, the effort prediction model was limited to productivity averages. We also filtered the ISBSG repository with DWH projects to compare the organization's productivity with the ISBSG average productivity. We organized another meeting to discuss our findings at the final stage with the organization. During the meeting, we agreed that Project C Phase 1 and Phase 2 should be evaluated together, and thus we updated the productivity calculations. We evaluated the outliers with the organization and suggested possible future steps to form the measurement and effort estimation basis of DWH projects.

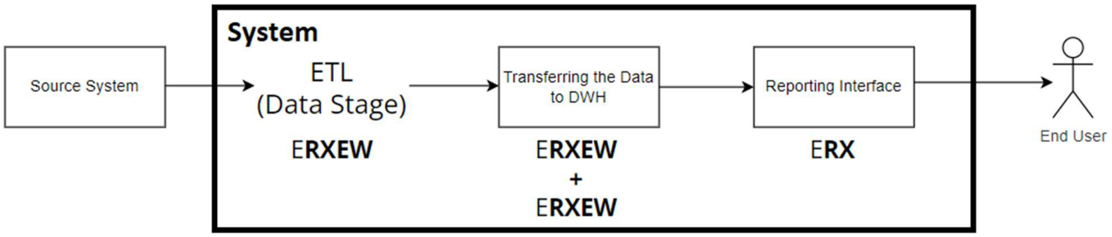

Fig. 1.. The architecture of Directorate X and measurement guideline

IV. Creating Measurement Guidelines for the Organization

We analyzed the DWH architectures of two directorates in the organization and formed measurement guidelines for each architecture. As we mentioned before, the COSMIC Guideline includes two DWH architectures: Kimball and Immon. The measurement process in the guideline follows these two architectures. However, the architectures provided by the directorates are not directly based on Kimball and Immon type. They include different modules and steps. The architectures and modified measurement rules for Directorate X and Directorate Y are as follows:

A. Mearument in Directorate X

The measurement guideline modified based on the COSMIC DWH Manual for the architecture provided by Directorate X is shown in Fig. 1. In the meetings, we analyzed that 1 ETL is included while receiving data from the Source System, and 2 ETLs are included during the data transfer to DWH. Also, the directorate adopted a clock-triggered batch process in their architecture.

Data movements shown in **bold** in Fig. 1 are calculated as one per transaction for each object of interest, and data movements shown in not bold are calculated as one for each process. The explanations for the data movements of an ETL are as follows:

E – Triggering entry

R – Read from the table

XE – Metadata management

W – Write to the table

We also analyzed that there is no ETL process in the Reporting Interface. In this case, the normal COSMIC Functional Size measurement is performed on the requirements for the Reporting Interface.

B. Measurement in Directorate Y

The measurement guideline modified based on the COSMIC DWH Manual for the architecture provided by Directorate Y is shown in Fig. 2. Data movements shown in bold in Fig. 2 are calculated as one per transaction for each object of interest, and data movements shown in not bold are calculated as one for each process.

From Fig. 2, it can be seen that there are 4 ETL processes in the architecture used by the directorate. The data used for the ETL processes in Source B, Source C, ODS, and Data Warehouse, are stored in tables and files, while in Source A, data comes from the service.

The explanations for the data movements of an ETL (reads data from tables or files) are as follows:

E – Triggering entry

R – Read from the table

XE – Metadata management

W – Write to the table

The explanations for the data movements of an ETL (data comes from the service) are as follows:

E – Data comes from a service connection (response)

XE – Metadata management

W – Write to the table

Module A and Module B consume the data and write back to ODS. The explanations for the data movements are as follows:

E – Triggering entry

R – Read from the table (ODS)

W – Write to the table (ODS)

X – Confirmation/error message

Module C and Module D take the data stored in ODS by a service connection and serve it to the end-user. The explanations for the data movements are as follows:

E – Triggering entry

R – Read from the table (ODS)

XE – Service request and response

X – Serve data to the end-user

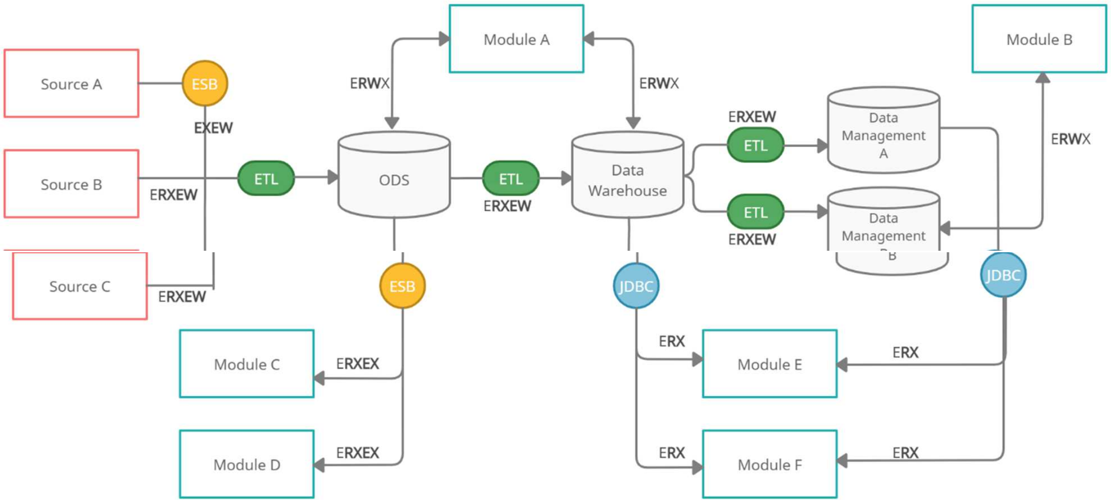

Fig. 2. The architecture of Directorate Y and measurement guideline

Module E and Module F directly read the data stored in Data Management A and B databases to serve it to the end-user. The explanations for the data movements are as follows:

E – Triggering entry

R – Read from the table (Data Management A/B)

X – Serve data to the end-user

V. Results

As an output of this study, we first calculated the productivity (person-day per COSMIC Function Point) of the projects based on the measured size and actual effort gathered from the organization.

The measured size and the calculated productivity of the projects are given in Table I. Upon the meeting with the organization, we decided to evaluate Phase 1 and 2 of Project C together to calculate productivity. There are two main reasons behind this: (1) the recorded effort of these two phases includes similar project management activities, and (2) the project team spent more time on Phase 1 learning the terminology required for this project. As a result, the average productivity of the measured three projects is calculated as 1.60 person-day/CFP, where a person-day consists of 8 person-hours.

The productivity results show deviations in the productivity of three projects (see Table I). The reasons behind the variations can be listed as follows:

- Deploying new technologies and lack of knowledge in these new technologies
- Development involving extensive algorithm development
- Development including simple and repeatedly performed operations

More specifically, we obtained that the development of Project A included extensive algorithm development, and they deployed new technologies in this project. In opposite, the development of Project B had simple and repeatedly performed operations. The productivity of Project A is relatively larger than the productivity of Project B. Thus, the organization reported to us that Project A and Project B could be potential outlier projects for them in terms of productivity.

TABLE I. Measured Projects and Related Efforts

Project	Functional Size (CFP)	Actual Effort (person-day)	Productivity (person-day/CFP)
A	18	78.13	4.34
B	130	78.38	0.60
C	116	267.00	2.30
Total	264	423.50	1.60 (average)

We performed the second productivity comparison with the ISBSG [13] repository projects. For this purpose, we filtered the repository. We selected “A” and “B” for the “Data Quality Rating,” “Data Warehouse System” for the “Application Type,” “New Development,” and “Enhancement” for the “Development Type.” As a result, we obtained 80 projects. The “Programming Language” of these projects was “Java,” and the “Count Approach” was “IFPUG.” We needed to perform IFPUG to COSMIC mapping using the COSMIC size measurement method. The previous studies show a 1-1 mapping between IFPUG and COSMIC [17]. Thus, the average productivity of 80 projects from the ISBSG repository is calculated as 1.15 person-day/CFP (see Table II). This productivity rate is close to the average productivity of the measured three projects (1.60 person-day/CFP). However, as Project A and B were evaluated as outliers, the productivity of Project C is much more than the average productivity of ISBSG projects.

TABLE II. ISBSG 2019 Productivity Values

Resource	ISBSG-2019
Number of Projects	80
Data Quality Ranking	A&B
Application Type	Data Warehouse System
Development Type	New Development & Enhancement
Programming Language	Java
Count Approach	IFPUG
Average Productivity (person-day/CFP)	1.15
Median Productivity (person-day/CFP)	1.04
Productivity Standard Deviation	0.98

The average productivity measure gives an insight to the organizations to compare their productivity with other organizations or projects. However, it does not indicate any clue to estimate the effort. For this purpose, a regression-based effort estimation model using the ISBSG repository could help organizations with a limited number of projects form an effort estimation model. Thus, as a third step, we performed regression analysis to evaluate the success of the effort estimation models established with ISBSG data. We constructed the effort estimation model for productivity calculations using the same ISBSG DWH projects.

Fig. 3 shows the effort estimation model for 80 projects extracted from the ISBSG repository. We applied linear regression for the projects. According to the effort estimation model, we observed that the R^2 value of the effort estimation model was very low (R^2=0.1436). In other words, it can be said that ISBSG data explain the 14.36% change in effort.

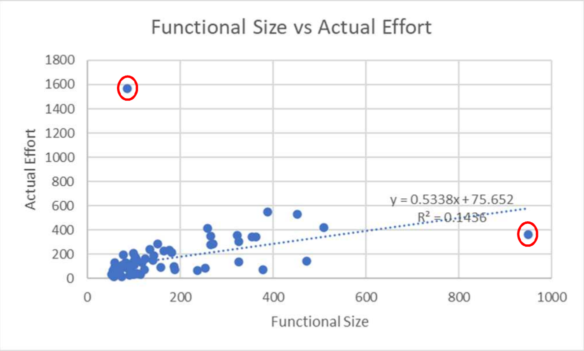

Fig. 3. Effort estimation model before eliminating outliers

The Mean Magnitude of Relative Error (MMRE), the Median Magnitude of Relative Error (MdMRE), and PRED(30) metrics were used to evaluate the success of the effort estimation model. These results are presented in Table III.

TABLE III. ANALYSIS RESULTS (BEFORE ELIMINATING OUTLIERS)

Number of Projects	80
R^2	0.1436
MMRE	1.04
MdMRE	0.50
PRED (30)	0.34

In the estimation model (see Fig. 3), we have seen two possible outlier projects (circled in red). Although we could not discuss them with the organization, we identified these projects as outliers and repeated the process of eliminating these two projects. The effort estimation model after eliminating outliers can be seen in Fig. 4.

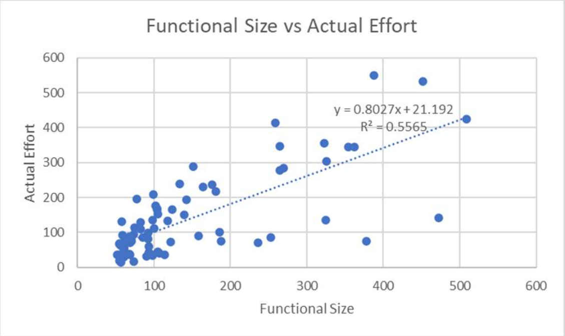

Fig. 4. Effort estimation model after eliminating outliers

After eliminating the possible two outlier projects, we improved our analysis results (see Table IV).

TABLE IV. ANALYSIS RESULTS (BEFORE ELIMINATING OUTLIERS)

Number of Projects	78
R^2	0.5565
MMRE	0.72
MdMRE	0.37
PRED (30)	0.42

As the last step, we used the equation obtained from the model in Fig. 4 to estimate the effort of 3 measured projects of the organization. Using the ISBSG data, we estimated the effort of projects A, B, and C with 0.54, 0.60, and 0.27 MRE, respectively (see Table V).

TABLE V. ESTIMATED EFFORT FOR THE MEASURED PROJECTS

Project	Functional Size (CFP)	Actual Effort (person-day)	Estimated Effort (person-day)	MRE
A	18	78.13	35.64	0.54
B	130	78.38	125.54	0.60
C	116	267.00	114.31	0.57

VI. DISCUSSION

In this study, we modified the COSMIC DWH Measurement Guideline based on the two different architectures of the projects. We then measured the size of three projects with the modified method. As the number of the projects was limited, we could not construct an effort estimation model using the organization's DWH project data; however, we calculated the productivity of these projects.

We compared the productivity of DWH projects with the productivity calculated in our previous measurements within the organization. Previously, we measured the size of 18 business application projects in the same organization and calculated the average productivity of these projects as 2.66 person-day/CFP. In this study, if we evaluate Project A and Project B as outliers, it can be seen that the productivity of Project C (2.30 person-day/CFP) is close to the previously calculated average productivity (2.66 person-day/CFP).

We should note that the average productivity was calculated from a limited number of DWH projects (3 projects). With the increase in the number of measurements, the productivity value will be determined more precisely, and it will be possible to compare the productivity based on project characteristics.

The primary aim of this study was to form a basis for the size measurement and effort estimation of DWH projects in the organization. For this purpose, we first analyzed the COSMIC Measurement Manual for DWH Projects. However, we have seen that the architecture given in the manual cannot be applied directly in the organization as there are differences in the DWH architecture of the organization. Thus, we modified the measurement guideline based on the organization's architecture. This process was not straightforward in comparing the measurement process projects such as Business Applications etc. The effort in the measurement process was much more than a typical COSMIC measurement process. We organized several meetings with the organization to discuss the modified manual based on their architecture. Thus, the COSMIC Manual on DWH projects should be adjusted according to project architecture.

Commonly, many organizations may not have a set of projects, including recorded effort, to construct an effort estimation model. In our case, we have previously built an effort estimation model for the organization. However, the organization had limited historical data on DWH projects. Thus, we could not construct an effort estimation model specifically for these projects. As an alternative, we evaluated the success of effort estimation using ISBSG DWH project data. We chose Data Quality as "A" and "B, "meaning that the project data is reliable. We also eliminated two possible outliers and updated the model. However, the calculated MMRE value was high to perform a reliable estimation. With the model, we estimated the effort of the projects between 0.54 and 0.60 MMRE.

Different filters can be used for the ISBSG data, and new estimation models can be constructed to have a more predictive estimation. For this purpose, we tried different filters on ISBSG data. First, we aimed to have projects with "A" data quality. However, only 2 out of 78 projects had "A" data quality. We then also checked the "Development Type" and "Organization Type" attributes of the data. Similarly, all projects were Enhancement projects, and the Organization Type of the majority was Telecommunications.

Thus, we can conclude that effort estimation using ISBSG data may give an insight into the organizations. However, the culture and experience of each organization differ. Therefore, organizations should not rely only on the estimations based on the ISBSG repository in critical projects. However, they should not rely on only ISBSG-based estimations. Project managers should be aware of the risks of these estimations and evaluate the possible outliers carefully.

As a result, we have achieved a basis for the size measurement and effort estimation of DWH projects. However, there is a need to measure new projects as it occurs to construct an effort estimation model and perform more precise comparisons.

VII. Conclusion and Future Work

In this study, we formed a functional size measurement basis for DWH projects in an organization. For this purpose, we identified the project features and adopted the COSMIC Measurement Manual for DWH projects based on these features, performed functional size measurements of DWH projects from the related analysis and design documents, and calculated productivity for these projects. We also evaluated the success of the effort estimation model using DWH projects in the ISBSG repository.

The measurement guidelines are modified based on the architecture of the organization in this study. The modified measurement guideline may not be relevant to other organizations. Organizations should modify the COSMIC Guideline based on their DWH architectures. However, the effort estimation model constructed with ISBSG data can be used by organizations with limited effort data. However, they should be aware of the risks of these estimations, and organizations should not rely only on the estimates based on the ISBSG repository in critical projects.

References

[1] A. Samuel, A. K. Pandey, and V. K. Sharma, “Estimation of Functional Size of a Data Warehouse System using COSMIC FSM Method,” presented at the International Conference on Advances in Computer Science and Application 2013, 2013.

[2] W. H. Inmon, *Building the data warehouse*. John wiley & sons, 2005.

[3] M. Golfarelli and S. Rizzi, “From Star Schemas to Big Data: 20+ Years of Data Warehouse Research,” in *A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years*, S. Flesca, S. Greco, E. Masciari, and D. Saccà, Eds. Cham: Springer International Publishing, 2018, pp. 93–107. doi: 10.1007/978-3-319-61893-7_6.

[4] R. Kimball and M. Ross, *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons, 2011.

[5] “Standish Grup Chaos Report,” T.S.G. International, 1995.

[6] H. van Heeringen, “Measuring the functional size of a data warehouse application using COSMIC-FFP”.

[7] “COSMIC Measurement Manual Version 4.0.2,” The Common Software Measurement International Consortium, Dec. 2017. [Online]. Available: https://cosmic-sizing.org/publications/measurement-manual-v4-0-2/

[8] H. Ünlü, T. Hacaloglu, O. Leblebici, and O. Demirörs, “Effort Prediction for Microservices: A Case Study,” presented at the 2021 Turkish National Software Engineering Symposium (UYMS), Nov. 2021.

[9] M. Haoues, A. Sellami, H. Ben-Abdallah, and O. Demirors, “Evaluating Software Security Change Requests: A COSMIC-Based Quantification Approach,” in *2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, Aug. 2019, pp. 268–275. doi: 10.1109/SEAA.2019.00049.

[10] C. Commeyne, A. Abran, and R. Djouab, “Effort estimation with story points and cosmic function points-an industry case study,” *Software Measurement News*, vol. 21, no. 1, Art. no. 1, 2016.

[11] M. Salmanoglu, T. Hacaloglu, and O. Demirors, “Effort estimation for agile software development: comparative case studies using COSMIC functional size measurement and story points,” in *Proceedings of the 27th International Workshop on Software Measurement and 12th International Conference on Software Process and Product Measurement*, Gothenburg, Sweden, Oct. 2017, pp. 41–49. doi: 10.1145/3143434.3143450.

[12] “Guideline for sizing Data Warehouse and Big Data Software.” The COSMIC Functional Size Measurement Method, Dec. 2018.

[13] “ISBSG Repository,” *ISBSG*. https://www.isbsg.org/ (accessed Oct. 18, 2021).

[14] H. Ünlü *et al.*, “Software Effort Estimation Using ISBSG Dataset: Multiple Case Studies,” in *2021 15th Turkish National Software Engineering Symposium (UYMS)*, Nov. 2021, pp. 1–6. doi: 10.1109/UYMS54260.2021.9659655.

[15] L. Santillo, “Size & Estimation of data warehouse systems,” 2001.

[16] R. Rasool and A. A. Malik, “Effort estimation of ETL projects using Forward Stepwise Regression,” in *2015 International Conference on Emerging Technologies (ICET)*, Dec. 2015, pp. 1–6. doi: 10.1109/ICET.2015.7389209.

[17] J. J. Cuadrado-Gallego, L. Buglione, M. J. Domínguez-Alda, M. F. de Sevilla, J. Antonio Gutierrez de Mesa, and O. Demirors, “An experimental study on the conversion between IFPUG and COSMIC functional size measurement units,” *Information and Software Technology*, vol. 52, no. 3, pp. 347–357, Mar. 2010, doi: 10.1016/j.infsof.2009.12.001.