Mustafa2018.pdf

Analysis of the inclusion of environmental cost factors in software costs

Estimated datasets

Summary— Many factors affect estimating the cost of a software project. Country-specific environmental factors, cultural, social or technical factors that impact software development costs. In this paper, we analyzed the characteristics of 31 cost estimation datasets to determine the combination of environmental factors in their cost attributes. Analyzing the heterogeneous attributes of the dataset, we combined all attributes in six categories and 48 representative attributes. We observed that the majority of the dataset represented the organizational environments of Europe and North America. Furthermore, recent datasets have provided more diverse attributes and increased coverage of organizational elements (users, developers, and project properties). However, environmental factors, i.e. cultural and social norms, were not represented. This limits the application of these datasets in environments where these factors have a major impact on software development costs. This article emphasizes the need for further research into cultural and environmental factors and their impact on software cost estimation.

Keyword—software effort estimates, datasets, attribute cost, environmental factors.

I. INTRODUCE

Software affordability has become an important factor in early-stage software design[1] and is directly affected by software development and maintenance costs[2]. Cost estimation methodology software is used to quantify costs and provide stakeholders with estimates to assist in the decision-making process. Furthermore, improperly estimated costs impact software quality, e.g. later stages of quality assurance activities and verification phases can be cut to minimize costs and exceed schedules [3]. It has been recognized that software cost estimation methods must evolve to accommodate software that currently diverges the evolving landscape of processes, products, assets, and personnel [4]. This setting, when considered on an international scale, must respond to country-specific environmental and cultural aspects that can directly affect software development efforts. Since, for example, in technically constrained countries, estimates of the impact of certain cost factors (i.e., team continuity) on the software development process will differ from those of a less restrictive environment. Model estimates estimate formal software effort and software project costs based on dataset project history [5]. The accuracy of these methods depends on the characteristics of the underlying dataset and the relevance of the attributes in the dataset [6-8]. These attributes represent software development cost drivers, i.e., factors that affect the cost and schedule of the software development process. The cost estimation dataset is small in size, lacks value, heterogeneity and project obsolescence that does not fit into the current software development context [7, 9]. Most of the attributes included in the dataset are based on specific assumptions made by the author [10], cost model assumptions [11] or the availability of information from historical projects [12]. Furthermore, most datasets cannot be applied to overhead cost estimates because they are only relevant to the environment from which the project data was collected [13, 14]. However, these datasets continue to be used to formulate, validate, and compare software cost estimation models as demonstrated in recent research, e.g., [7, 8, 15, 16]. Determining the degree of influence of cost drivers depends on an understanding of the project environment and the effect of attributes on costs [17]. Most studies have looked at the impact of the environmental organization [10, 13, 14, 18] where the magnitude of the impact varies depending on work settings, internal policies, etc. [13]. Some authors ([13, 14, 19]) have used the term environmental factors to refer to those factors. However, organizational factors influenced by country-specific cultural, social, and technical differences can directly affect software development costs. For example, the importance of being punctual is not considered the cost of driver software. However, cultural differences in perceptions of the importance of punctuality will influence estimating the impact of known software cost drivers, e.g., project management and team productivity. In this paper, we identify environmental factors such as country-specific culture, social or technical factors that directly or indirectly impact the cost of the software development process. This paper aims to evaluate current cost estimates of datasets and determine their suitability for application in countries with limited cultural, social, and technical differences. Specifically, we studied the overall characteristics of datasets and their detailed attributes for their coverage of overhead cost drivers and specific factors influenced by country-specific environmental factors. We show that more recent datasets provide more than granular attributes; However, they do not sufficiently represent properties that are better suited to software development costs in dissimilar environments. The contributions of this report are as follows. • We analyzed the characteristics of 31 estimated dataset software costs from 1981 to 2017. We show that the majority of the datasets represent European and North American environmental organizations.

To compare the attributes of these datasets, we identified six types of costs that represent: general information, users, developers, scale, projects, and product elements. We then adjusted and defined 48 cost attributes in these categories as a basis for comparison. We demonstrate that more recently datasets have more diverse attributes and represent more cost drivers than older datasets. • We point out that recent dataset properties represent factors related to users, developers, and project properties, i.e., factors that may be affected by environmental factors. However, environmental factors are clearly not represented. This limits the applicability of these datasets in organized environments and sociocultures. The rest of this article is organized as follows. Section II presents the work involved, and Part III describes the methods used to select the cost estimate dataset. IN Part IV, we analyze the characteristics of the dataset, compare their properties, and demonstrate the need for a clear representation of environmental factors in the future dataset. Conclusions and future work are discussed in Section V.

II. RELATED WORK

Most studies use cost estimation datasets to build and compare different software cost estimates methods, for example, [8, 15]. Miar et al. [20] Classification of 42 companies and separate datasets published between 1979 and 2003. The author categorized the datasets by age, availability for use, number of projects, missing data value, project source, and country of origin. The work presented the number of attributes per dataset but did not provide details of the attributes. Dejaeger et al. [5] Use multiple datasets to assess machine learning suitability Software cost estimation methods Several software cost estimation datasets from two previous studies are publicly available in OpenScience tera-PROMISE[21] software engineering research data repository. Limited work examines the impact of cost drivers on the accuracy of cost estimation methods. Bergeron and Arnaud [17] project managers surveyed identified 19 driver costs that they considered important for estimating costs, e.g., user availability and organizational development of policy. Using identified cost drivers, the authors collected a dataset to investigate the drivers that had the highest level of impact on costs. They found that user-related cost drivers had the highest impact of all software development stages. Pendharkar et al. [10] Examined the impact of team size, tool usage, programming languages, and functional point attributes on employee effort. They found that increasing product size and team size increased overall effort. The authors identified a relationship between attributes and their impact on total costs, e.g., large groups, and the use of tools reduced overall effort. Research involving the impact of environmental factors on software cost estimation mainly examines factors related to the organizational environment. Ham Kitchen[14] considers the impact of environmental organizational factors (team experience, project time, and environmental development) on employee productivity. The author found that improving the development environment improves productivity, e.g., increasing team experience and using productivity-boosting tools. Global software development cost estimates include cost drivers that reflect multi-location, multicultural aspects of the development process, for example, communication, time zone and language differences. Lamersdorf et al. [22] conducted a survey of practitioners in which they identified the most influential cost drivers on global software development and assessed the impact among cost drivers. The authors found that most practitioners identified: process maturity, process experience, and technical knowledge had the most impact on the effort. When determining the impact of cost drivers, the authors found that language and cultural differences were identified as having the greatest impact on other cost drivers. Britto et al. [23] conducted a systematic literature review of research cost estimates on global software development. The authors found that the majority of studies included time zones and languages and cultural differences as cost drivers in evaluating efforts in global software projects. Both studies looked at culture as a cost driver influencing the global cost of software development, however, no precise definition of what the term meant was provided. In the context of internationalization of software engineering, outsourcing and a growing international software market, the ability to accurately estimate software costs is of high importance. Accurate software cost estimation requires consideration of environmental factors, not only within an organization, but in the context of a country, its work culture and social norms. In this paper, we examine the characteristics of publicly available software cost datasets and show the need for further research addressing the identification and impact of environmental factors in terms of software cost estimation.

III. DATASET SELECTION METHOD

The datasets in this study are public datasets with known attributes that match the cost algorithm of the estimating models. As the goal of this work relates to attributes, we've included private datasets with publicly available properties. All published datasets without attributes have been excluded. To determine the characteristics of the dataset, in addition to some of the characteristics provided in [20], we included the date and region of the dataset. To determine the date, we used: year of publication of the dataset, year of the last project in the dataset, year of the first study to use the dataset, or year of donating the dataset to a public repository, respectively. Datasets with similar attributes from the same country are represented as the most recent dataset. For example, the US COCOMO II dataset: coc81[24] and nasa93[21] are represented by nasa93. In contrast, cocomosdr [21], a Turkish COCOMO II dataset was included. For collections of datasets collected by the same authors, we included datasets with the most comprehensive attributes, e.g. Miyazaki et al. [25] represents three other datasets by the same authors. Table I details the 31 datasets used in this study. The datasets in this study are a combination of the cost estimation dataset in OpenScience tera 624 PROMOTION [21], the dataset surveyed by Mair et al. [20] and the datasets in the study by Dejaeger et al. [5]. The tera-PROMISE OpenScience repository includes 14 cost estimate datasets. We have excluded a dataset non-algorithmic tool. As stated earlier, coc81 and nasa93 are represented by the nasa93 dataset. The cosmic dataset and ISBSG have the same properties and are therefore represented by the ISBSG (issue 10)[26] dataset. So we've included 11 datasets from the tera-PROMISE repository. Miar et al. [20] reviewed 42 datasets collected from literature. For each dataset, we have referred to the articles included in [20] to define the characteristics of the dataset. We exclude the following: 10 datasets are already included in the PROMISE Repository, 8 datasets for which we cannot identify cost attributes, 2 non-algorithmic datasets, and cocomo datasets because it has been replaced by COCOMO Model Properties II. We represent both Dolado datasets of Dolado-academic [27] and BT housing software, BT systems and ICL datasets represented by BT home software [28]. So in this study, we included 18 datasets from [20]. Dejaeger et al. [5] the study contained 9 datasets: four words[20] and three from PROMISE warehouse. From the other two datasets, we excluded a separate dataset, which included only the ESA dataset [29]. The characteristics of all datasets in this study are detailed in Table I. The datasets are arranged in ascending chronological order

IV. DATASET ANALYSIS

A. Dataset characteristics

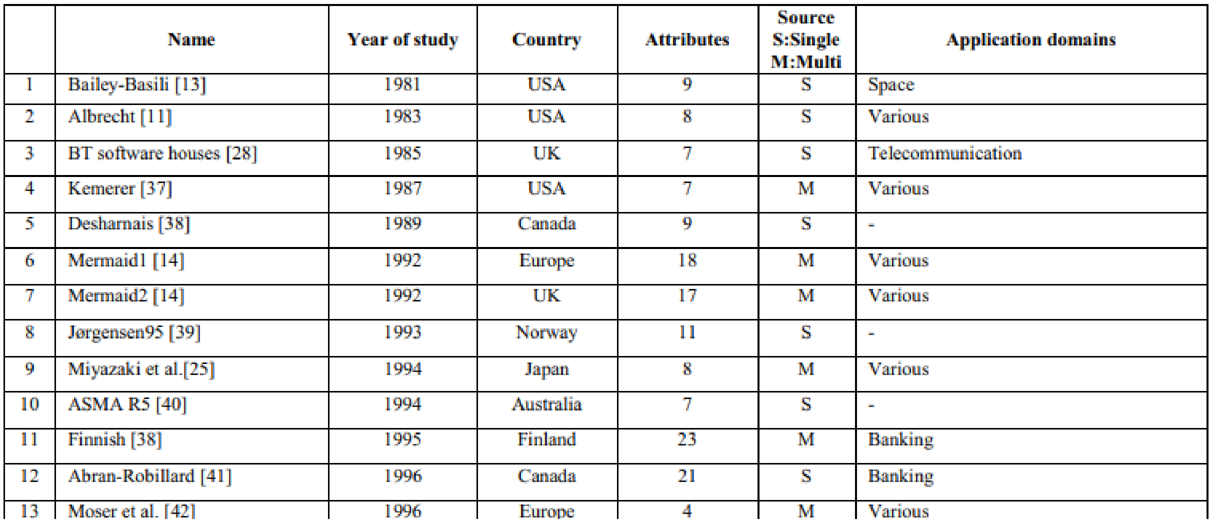
For each dataset, Table I presents details of the projects collected in terms of country of origin, number of cost attributes, project areas and whether the projects originated from one or more sources. Most datasets date the first year of the study using datasets. China's Day [21] and benchmark dataset use case points [19] are contribution dates to the PROMISE Repository. For benchmark use cases, we were unable to find information regarding the country of origin. Also, we were unable to determine the country of the ISBSG dataset. However, the ISBSG's website states that ISBSG's partner countries are: China, Finland, Italy, Japan, Mexico, the Netherlands, Spain, Switzerland, the United States, and Canada.[26] From Table I, the majority of datasets (64%) were collected in the 1980s and 1990s. The dataset collected at the turn of the millennium accounted for only 25% of the total dataset. The most recent datasets are ISBSG, Openeffort, and benchmark use cases. ISBSG attributes require pretreatment because some do not match the effort estimate [30]. Openeffort-based software costs estimate the amount of developer commitment in an open source repository [31]. Benchmark use cases use use case agents as dimensional measures [19]. Prior work had determined that older datasets were not suitable for representing current software development realities. However, these recent datasets are limited by their application project domains and are therefore suitable for overhead cost estimates. As for the country of origin, for the 29 datasets, 48% of the datasets came from Europe, followed by 31% from North America and 21% from both Asia and Australia. NO datasets were collected representing projects from Africa or South America. Moreover, no low to middle-income countries from Asia are represented. Studies have shown that these countries have an impact on internationally developed software, for example, [32-34]. In addition, a number of factors affecting software costs and schedules are directly related to the country's specific environments [35]. Most datasets reported that their projects originated from different sectors (38%), however, the authors of the dataset did not detail these industries. This was followed by datasets from the banking sector (16%) and the telecommunications sector (10%). The lack of detail makes it difficult to assess the comprehensiveness of the application field dataset. Although most datasets reported projects are collected from multiple organizations (55%), the ambiguity of descriptions hinders the adoption of datasets in dissimilar environments. The number of attributes in the dataset varied with an overall average of 13 attributes. The ISPSG dataset is an exception; However, most of its properties are not specific to cost estimates. The variety of projects and attributes makes comparative cost research difficult [13, 20] with most studies reporting inconclusive results. This is further complicated because the majority of datasets represent outdated project development environments. Furthermore, since most datasets reflect their respective developments, organizations or industrial environments they exclude environments with cultural differences and social practices. This is a major limitation due to the importance of the growing software industry and emerging economies [34].

B. Cost attributes

We examined the properties of 30 datasets from Table I. The total number of unique attributes for all datasets is approximately 139. Most attributes are present in multiple datasets, and some are unique to specific datasets, for example, use case benchmark datasets have a single point attribute of developer dynamics. To compare attributes between the datasets, we categorized them into six different cost categories, four adapted from the COCOMO II model [36] and two from [17]. The categories are chosen to represent important elements in the cost estimation project. Our category is detailed as follows. • General information: represents attributes related to project characteristics, e.g. year, duration, and application type. • Users: this category represents cost attributes related to user impact, e.g. user availability and user resistance. • Developer: these are attributes related to the development team's cost contribution, e.g. team experience and team size. • Dimensions: are attributes that measure the dimensions of a software product, e.g. KSLOC or attribute function points. • Project: are attributes that represent the details of the project. project development environment, e.g. availability tools and hardware and software platforms. • Product:

demonstrates the technical characteristics of the software, e.g. security constraints or reusability

TABLE I. DATASET CHARACTERISTICS



	Name	Year of study	Country	Attributes	Source S:Single M:Multi	Application domains
1	Bailey-Basili [13]	1981	USA	9	S	Space
2	Albrecht [11]	1983	USA	8	S	Various
3	BT software houses [28]	1985	UK	7	S	Telecommunication
4	Kemerer [37]	1987	USA	7	M	Various
5	Desharnais [38]	1989	Canada	9	S	-
6	Mermaid1 [14]	1992	Europe	18	M	Various
7	Mermaid2 [14]	1992	UK	17	M	Various
8	Jørgensen95 [39]	1993	Norway	11	S	-
9	Miyazaki et al.[25]	1994	Japan	8	M	Various
10	ASMA R5 [40]	1994	Australia	7	S	-
11	Finnish [38]	1995	Finland	23	M	Banking
12	Abran-Robillard [41]	1996	Canada	21	S	Banking
13	Moser et al. [42]	1996	Europe	4	M	Various

To categorize 139 attributes, we initially combine all attributes with similar meanings into one attribute. Because for example, tool experience, analytics experience, developer experience, and OOP experience are represented by the attribute: team experience. We then combine multiple attributes that represent the same cost attribute, for example, the function score attribute represents all related attributes to size the function score. The effort attribute has been excluded because it represents a dependent cost attribute that can be inferred from other attributes. Therefore, we reduced the total number of attributes to 48 as shown in Tables II and III. The number of datasets in Tables II &; III corresponds to the order in Table I. We excluded the nasa93 dataset because it has the same properties of the more recent cocomosdr than the dataset. Figure 1 summarizes the distribution of attributes in each type for all data sets in Tables II &; III. Only two datasets (Meramaid2 and use case benchmarks) contain attributes in all six categories. All datasets include attributes from the size category, reflecting importance in relation to product size. However, the six datasets contain only attributes in the size category. Therefore, regardless of the characteristics of the dataset, product size is considered an important attribute included in all datasets. An interesting observation is that only five (16%) datasets contained properties from user categories. This limits the applicability of the majority of datasets in software industries in low- to middle-income countries, as user factors, e.g. stability requirements, represent a major challenge [35]. Product categories were represented in only 36% of the dataset. The remaining categories are represented on average by 55% of the datasets.

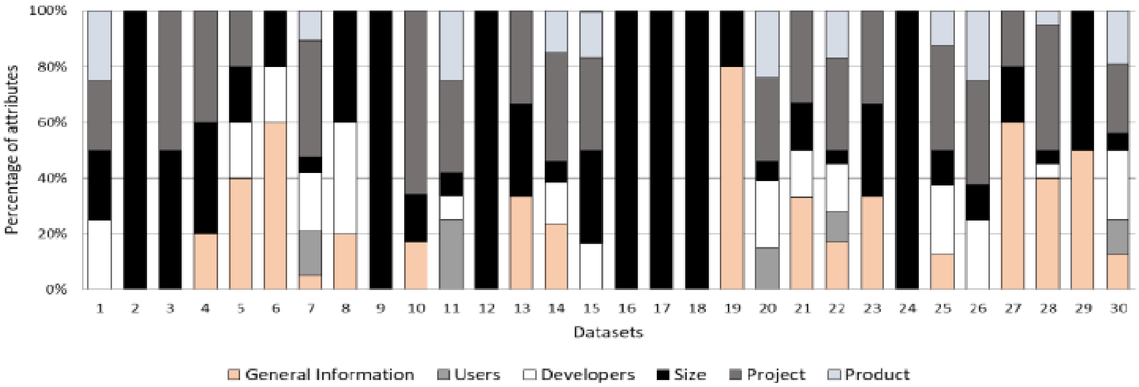


Fig.1 Percentage of dataset attributes within each category.

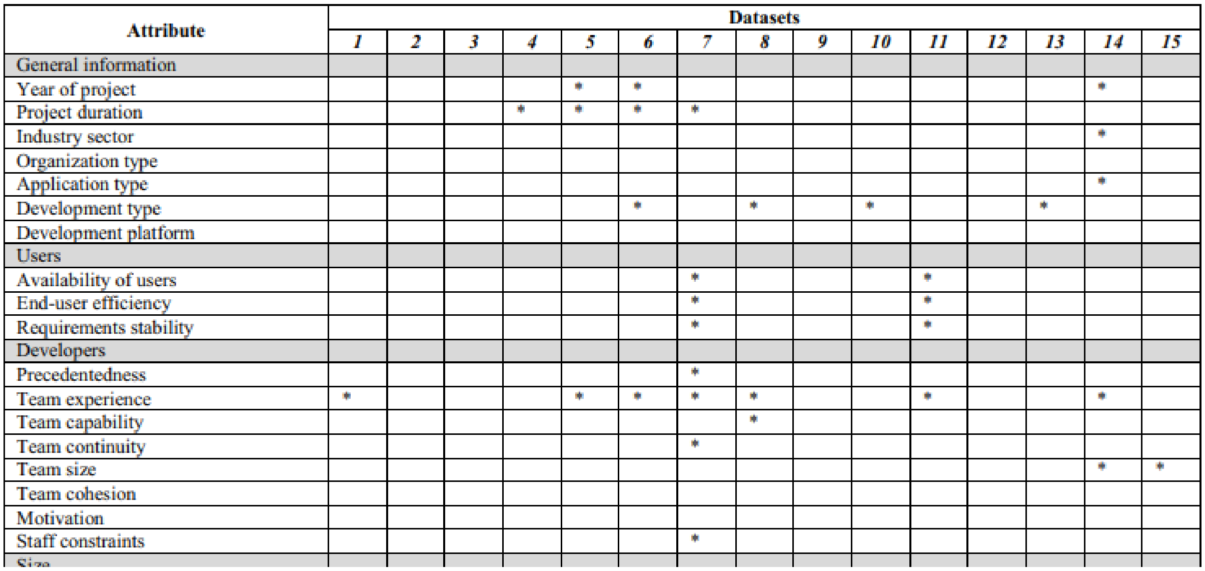
From Tables II & III, we analyze the coverage of attributes in each data set compared to the total number of attributes in each type. We then compared the average relevance of each category between datasets obtained before 2000 (pre-2000) and since 2000 (post-2000). For general information categories, project duration, app type, and property development platform were the most common attributes included in almost all 10 datasets. The ISSBG dataset covers all informational properties. Post-2000 datasets were mentioned on average 34% of general information attributes, in contrast to 26% for pre-2000 datasets. Five datasets with attributes in the user catalog comprise the majority of the attributes in the catalog. For developer category attributes, the experience attribute group is the most common attribute included in 12 datasets. The remaining attributes are representative in fewer than five datasets. Team size, continuous and cohesive attributes are not represented in the majority of datasets. Based on our previous research [35], developer portfolio is an important cost factor in technically restrictive environments. For example, in [33], the authors identified high developer turnover in the Sudanese Software Industry. So, in similar environments, the continuity attribute of the team will need to be considered in order to estimate software development costs accurately. However, portfolio developers' maximum coverage is 50% of properties in just three datasets:

Mermaid2, cocomosdr, and benchmark use cases. The mean attribution range for pre-2000 datasets was 21% and post-2000 datasets were 32%. The KSLOC attribute and function score are the most common attributes in the size category expressed in an average of 60% of the dataset. The average coverage of this category was similar between the post- and pre-2000 datasets. Because of project portfolios, programming languages (43%) and tool availability attributes (33%) are the most common attributes. Only one dataset, Mermaid2, accounts for 50% of these categories. The full attribute of the development environment is the environmental factor; However, it was only included in three datasets. Post-2000 datasets included more attribution (25%) than pre-2000 datasets' 18% relevance. For product categories, the most common attributes are product complexity and reliability attributes included in eight and six datasets, respectively. The Maxwell and cocomosdr datasets mentioned 57% of the total properties of this type. The average attribute coverage for pre-2000 datasets was 32% and post-2000 datasets were 26%. For all 48 attributes, the Little Mermaid Datasets 2 and Maxwell had the most attributes coverage (42%). Overall, post-2000 datasets had nearly double (21%) the coverage of pre-2000 datasets (11%). In terms of organizational environmental factors, post-2000 datasets show the extent to which elements are covered in categories of users, developers, and projects. The Baily-Basili Benchmarks, Meramaid2 and use cases clearly state the importance of environmental attribute organization [13, 14, 19]. Mermaid 2 organizational elements are included in our users, developers, projects, and product categories. Use cases Organizational factor benchmarks are included in users, developers, and project portfolios. The Baily Basili dataset considers all its properties to be organization of environmental factors except for dimensional measures. C. Importance of environmental factors The six categories and 48 attributes identified in the previous section represent technical cost attributes. A number of organization-specific elements can be taken into account within these attributes. Attributes that reflect country-specific factors are not adequately represented in these datasets. This is due to the fact that the majority of these datasets represent similar cultural and organizational norms. A common set of attributes or a common dataset will not fully represent the impact of country-specific factors on the cost and effort of the software product. When considering common datasets or attributes at an international level, consideration of environmental factors and their impact on software cost estimates is necessary. The implications of such differences on software development costs require their representation in cost drivers and datasets. Furthermore, as these factors vary from country to country, methods for determining the impact of environmental factors on other cost drivers are necessary.

IV. CONCLUSION AND FUTURE DIRECTION

This paper reviewed the characteristics of publicly available software cost estimation datasets in terms of including environmental factors associated with estimated software costs. Environmental factors are peculiar to the nation cultural, social or technical factors that directly or indirectly affect the cost of the software development process. We analyzed the characteristics of 31 estimated dataset publishing costs Datasets that vary in size, date, country of origin, application domain, and number of attributes. The majority of the datasets represent the United States and the European Organizational Environment. This limits the accurate

TABLE II. CATEGORIES OF COST DRIVERS AND THEIR DISTRIBUTION WITHIN THE DATASETS



Attribute	Datasets														
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
General information															
Year of project					*	*								*	
Project duration				*	*	*	*								
Industry sector														*	
Organization type															
Application type														*	
Development type						*		*		*			*		
Development platform															
Users															
Availability of users							*				*				
End-user efficiency							*				*				
Requirements stability							*				*				
Developers															
Precedentedness							*								
Team experience	*				*	*	*	*			*			*	
Team capability								*							
Team continuity							*								
Team size														*	*
Team cohesion															
Motivation															
Staff constraints							*								
Size															

application of even the most recent datasets on estimated research costs in countries with dissimilar organizations and social norms. The datasets include heterogeneous attributes (139 unique attributes in total). To analyze those attributes, we categorized them into six types of adaptations of previous works. We then summarize all the attributes in a representative set of 48 cost attributes. Our analysis shows that datasets from 2000 and beyond provide more diverse attributes and increased coverage of organizational environment factors (users, developers, and project properties). All datasets do not consider country-specific environmental cost factors. This is limited primarily to the nature of the dataset and research objectives and cost estimation methods.

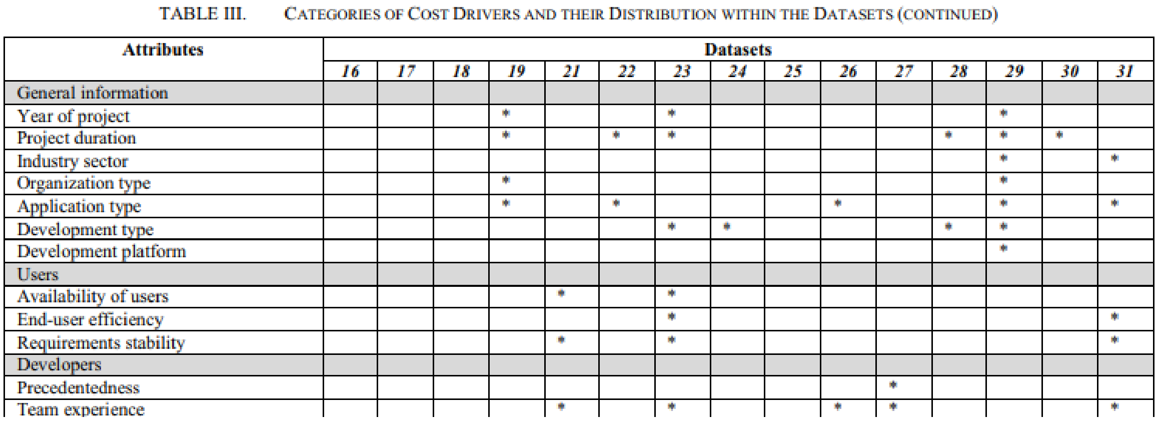


TABLE III. CATEGORIES OF COST DRIVERS AND THEIR DISTRIBUTION WITHIN THE DATASETS (CONTINUED)

Attributes	Datasets														
	16	*17*	*18*	*19*	*21*	*22*	*23*	*24*	*25*	*26*	*27*	*28*	*29*	*30*	*31*
General information															
Year of project				*			*						*		
Project duration				*		*	*					*	*	*	
Industry sector													*		*
Organization type				*									*		
Application type				*		*				*			*		*
Development type							*	*				*	*		
Development platform													*		
Users															
Availability of users					*		*								
End-user efficiency							*								*
Requirements stability					*		*								*
Developers															
Precedentedness											*				
Team experience					*		*			*	*				*

More work is needed to fully account for related and environmental cultural and environmental factors in the country cost estimation attributes. This will require identifying specific environmental factors that affect software development costs. In addition, quantifying the impact of environmental factors on known software cost attributes will require case studies in different environments and countries.

Mustafa2020.pdf

SEERA: SOFTWARE COST ESTIMATION DATASET FOR CONSTRAINED ENVIRONMENTS

ABSTRACT

The accuracy of software cost estimates depends on the relevance of the cost estimate dataset, the quality of the data, and its suitability for the targeted software development environment. Software development costs are influenced by the organizational environment and specific technical, socioeconomic, and national culture. The current publicly available software cost estimates datasets represent the environments of North America and Europe, thus limiting their application in technically and economically constrained software industries. In this paper, we introduce the SEERA (Software enginEERing in SudAn) cost estimation dataset, a dataset of 120 software development projects representing 42 organizations in Sudan. The SEERA dataset contains 76 attributes and, unlike the current cost estimate dataset, is supplemented with metadata and original raw data. This paper describes the data collection process, the organization of submissions, and the characteristics of the project. In addition, we provide a general analysis of dataset projects to illustrate the impact of local factors on software project costs and compare the data quality of the SEERA dataset with that of public datasets from the PROMISE repository. The SEERA dataset fills gaps in the diversity of the current cost estimation dataset and provides researchers with the opportunity to assess the generalization of previous and future cost estimation methods for constrained environments and develop new techniques that are better suited to these environments.

CCS CONCEPT

• Software and its engineering • Software libraries and repositories

KEYWORDS

Software effort estimation, datasets, data quality, cost attributes, constrained environment, socioeconomic factors, Africa

1. INTRODUCTION

Software cost estimation methods use historical datasets of previous projects to estimate project costs and efforts. These methodologies must include the ever-changing context of software development[1] and the impact of country-specific cultural and environmental factors on software development practices [2]. Research in experimental software engineering and cost estimation has recognized that the relevance and representativeness of datasets is critical for accurate and realistic cost estimation [3-6]. There should be a correspondence between the datasets, models, and environments in which the results apply. However, most research on cost estimation is based on outdated datasets that may not be suitable for the current software development environment [3, 4].

Research on cost estimation has focused on technical methods [3]. Limited work investigated the impact of organizational (e.g., [7-9]) and cultural (e.g., [2, 10]) factors on other cost attributes and their relevance to industrial software practices. Research on cost estimation datasets focuses on surveying datasets (e.g., [11-13]), assessing the quality of datasets (e.g., [4, 14]), and addressing specific quality issues of existing datasets (e.g., [15, 16]). Unfortunately, the number of datasets collected from 2000 onwards does not exceed 25% of the existing public cost estimate datasets [4, 12]. Furthermore, all of these datasets were collected and evaluated from software development environments representing North American and European settings, none of which originated from Africa or South America [12]. There are few factors reflecting the impact of the organizational environment and no factors reflecting socio-economic or cultural factors [12]. This limits the application of these datasets to countries with dissimilar organizational and social norms, i.e. developing nations with limited economic and technical environments. Moreover, it calls into question the generalization of cost estimation methods for such restrictive environments.

There are limited empirical investigations of software engineering practice in Africa [17]. Surveys have shown that the African software industry is locally focused with the majority of companies developing tailor-made local applications or custom imported applications [18, 19]. Previous studies of the African software industry have highlighted the importance of socioeconomic factors to the success of the software industry [19, 20]. Other studies have shown differences in the ranking of success factors important to software projects between African professionals (e.g., highest solicitation requirements ratings and team competencies) and European and U.S. professionals (e.g., ranking of project management issues and highest cost estimation) [20, 21]. For these technically and economically constrained environments, research and implementation of software cost estimates and scheduling should be consistent with locally recognized factors as well as relevant information and experience impacting software development costs. The level of detail of cost attributes, how they are defined and collected should be consistent with the realities of these confined environments.

This article introduces the SEERA (Software enginEERing in SudAn) cost estimation dataset [22]. The SEERA dataset fills the current gap in the cost estimation dataset in that (1) it provides the current dataset with traditional cost attributes in addition to organizational and socioeconomic attributes. (2) Dataset projects represent limited technical and economic software development environments, thus providing the international software engineering research community with a diverse and recent dataset to assess the generalization of past and future cost models. (3) The dataset fills the critical gap for Sudan and Africa's research communities with a more relevant cost estimate dataset that includes elements that are more relevant to the realities of their software industry. (4) The SEERA dataset overcomes the current limitations of dataset quality and transparency by supplementing the cost estimate dataset with the original raw data prior to encryption/scaling. This gives researchers the flexibility to create new cost estimate datasets (e.g., COCOMO-style datasets from the original SEERA data), subsets (e.g., excluding attributes to represent different environments), or changing the size of attributes from the original SEERA data. This allows for the possibility of replication of the results and general application of the dataset for international research.

The main contributions of this work are as follows:

- SEERA dataset. A set of software development projects managed from a technically and economically constrained environment. Datasets contain detailed attributes and elements that better match the cost and schedule information available in these environments. The SEERA dataset is publicly available in the Zenodo repository [22].

- We provide detailed descriptions of the characteristics of projects in the dataset about the submitting organization, type of development, and application domain. We describe the properties of datasets, categories and their types.

- We explain local properties that reflect the limited technical and economic environment and their relationship to international properties and categories. We illustrate the importance of these local attributes through analysis of project data that shows the impact of these attributes on the cost and duration of the project.

- We provide a data quality assessment of the SEERA dataset against the PROMISE repository cost estimate dataset [23]. The SEERA dataset consists of raw data initially prior to encryption/scaling and is enhanced with metadata, thus increasing transparency and reliability over the PROMISE dataset.

The rest of this article is organized as follows. Part 2 details the study design and dataset SEERA collected. The characteristics of the dataset about: submission organization, project characteristics, attribute description, and

project analysis is in Part 3. We compare the data quality of the SEERA dataset with that of the PROMISE repository dataset in Part 4. The conclusions and future work are discussed in Section 5.

2 SEERA DATA COLLECTOR

2.1 Study design

To facilitate the collection of software project data, we designed a questionnaire that combined international software project cost factors (adjusted from [12]) and factors specific to the software development industry in developing countries (adjusted from [12, 19-21, 24]) . We design the questions according to convention in [25, 26], meaning that multiple questions are used to collect data related to a single cost factor. The questions focused on local issues and factors as described in [20, 21], for example, we asked about developer recruitment and incentive policies within the organization. Most practitioners are unfamiliar with the term cost and therefore questionnaires reflect local realities without clearly stating attributes and assessment scales. For example, most systems are database/information systems and therefore we asked directly for the number of screens/reports without mentioning the object score attribute. In addition, the questionnaire was prepared in Arabic. The questionnaire is nine pages long and has three sections as follows.

- Respondent information. Name, employer and contact details. This is to contact respondents in case there are any missing or ambiguous answers in the questionnaire. The idea for this part was adapted from.[27]

- General information about the project. Information related to the software project: organization size development, client organization, estimated and actual schedules, domain and application size, development type and methodology, team size, and team recruitment policy. This section consists of 20 short open-ended questions, 11 single-choice questions, and 2 multiple-choice questions. Six questions related to product scale and project progress have been adjusted since [26].

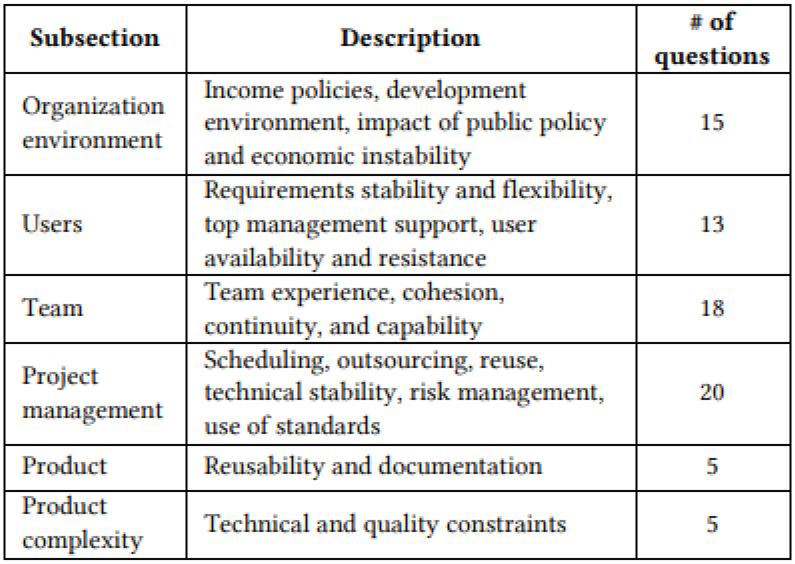
- Factors affecting the software product development process.

This is the largest part of the questionnaire consisting of six subsections. Table 1 details each subsection and the number of questions. For each subsection, the questions are a combination of single-choice or multiple-choice questions, type Likert

Closed questions and some short open-ended questions.

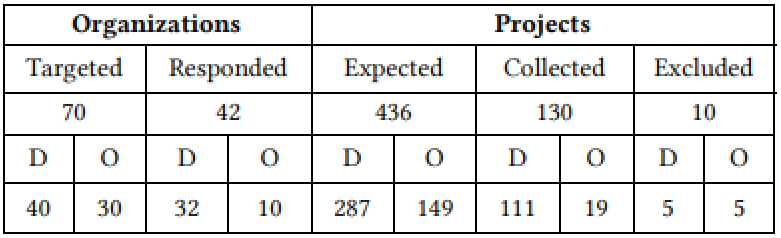
The subsections of Factors Influencing Software Product Development reflect local environmental factors that affect software development costs in developing countries. For example, the organizational environment subsection has related properties and questions derived from the factors identified in [20, 21] that are different from those of the current software cost estimate dataset [12]. In addition, the User and Project Management subsections based on local factors have only two and three corresponding cost attributes, adjusted from [26] but with different localized questions. In contrast, technical issues related to software development are based on international attributes, e.g. all cost elements of the Group, Product and Product Complexity subsections are adjusted from [26], albeit using different questions.

Table 1 Questionnaire subsections: Factors affecting software product development.



Subsection	Description	# of questions
Organization environment	Income policies, development environment, impact of public policy and economic instability	15
Users	Requirements stability and flexibility, top management support, user availability and resistance	13
Team	Team experience, cohesion, continuity, and capability	18
Project management	Scheduling, outsourcing, reuse, technical stability, risk management, use of standards	20
Product	Reusability and documentation	5
Product complexity	Technical and quality constraints	5

Table 2 Questionnaire distribution.



Organizations				Projects					
Targeted		Responded		Expected		Collected		Excluded	
70		42		436		130		10	
D	O	D	O	D	O	D	O	D	O
40	30	32	10	287	149	111	19	5	5

D: direct visit. O: contact by telephone, email or chat messages.

2.2 Conduct research

To identify targeted software development organizations, we mentioned those identified in [19], organizations accredited by the Sudanese National Information Center[28], and organizations recommended by experts in the Sudanese software industry [24]. This led to the identification of 70 organizations. The questionnaire was distributed and collected between June 2019 and February 2020. Data entry, review, and analysis are conducted in parallel.

Organizations were contacted through in-person visits (57%) or by phone, email, or chat (43%). Pre-printed questionnaires were provided during in-person visits and an electronic version was emailed to contacts in the targeted organization.

Table 2 details the number of questionnaires handed out and returned as well as the total number of projects collected. The total number of organizations responding to the questionnaire (42 organizations) accounted for 60% of all organizations targeted, with organizations with direct access accounting for 76% of respondents. We note that we have received feedback from individuals from several visiting organizations who have worked on projects as freelancers. For simplicity, we consider each freelancer to be a different entity, i.e. organization.

From the outset, each contacted organization undertakes to provide information for certain projects for which we have calculated the projected number of projects in Table 2. However, during the collection process, 40% of organizations did not respond or decided not to participate further in the study. The organizations that responded only delivered an average of 50% of the committed projects. As a result, the number of projects collected was only 30% of what was expected, and 85% of these projects were collected through face-to-face visits to institutions.

We note that the organizations did not return the questionnaire on the agreed date, which required multiple visits to several institutions and contact persons. In some cases, this is due to a lack of documentation that requires respondents to refer to initial contracts, previous team members, or the software product itself. Each collected question is modified for vague and missing answers. This led to another round of contacts with individuals to complete the data for 80% of the questionnaires returned. A further quality assessment of the collected projects led to the exclusion of 10 projects: four duplicate projects, two with incomplete schedule information, two projects submitted by users, and two projects that were abandoned before completion. Therefore, the total number of projects in the SEERA dataset is 120 projects. The next section will detail the characteristics of these projects.

3 SEERA DATASET CHARACTERISTICS

3.1 Organizational characteristics

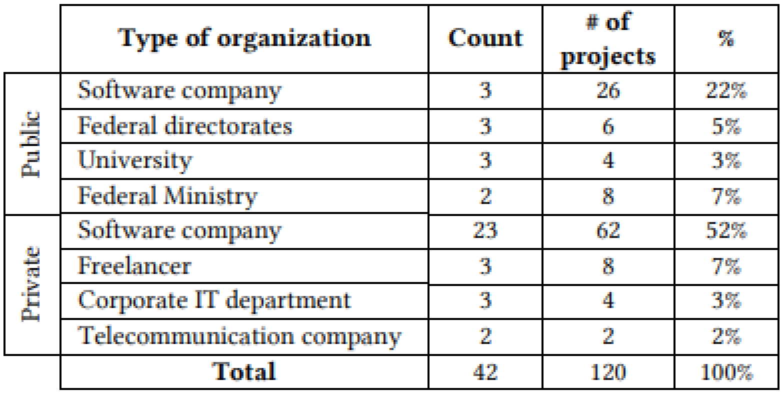
The SEERA dataset is a heterogeneous dataset from 42 different organizations representing the public and private sectors in Sudan. These organizations range from software development companies, to freelancers, to IT departments in public and private organizations. Table 3 provides detailed information about the organizations that contributed data to the project. The public sector represents 26% of organizations with the contribution of 37% of projects. Only public sector software companies develop software for customers, the rest of public organizations provide in-house software projects developed by their respective IT departments.

Private software companies contribute 52% of all projects and 82% of projects are contributed by the private sector. However, the average contribution of each private software company is between 1 and 3 projects with one company contributing 11 projects. This is in contrast to public software companies, where two companies contribute 16 and 8 projects and one contributes two. To reflect the heterogeneity of projects, the dataset includes attributes for the organization type, sector, and organization id.

Respondents were asked to provide the size of the organization and software development department (number

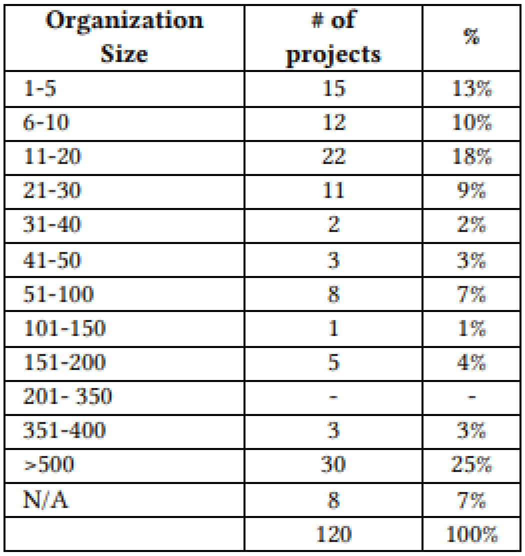
employee) throughout the life of the project. For organizations where the IT department conducts internal software development, the size of the IT department is already provided. Tables 4 and 5 present this information in detail. From Table 4, the majority of projects (61%) were carried out by relatively small organizations with fewer than 100 employees, with 23% of projects carried out by organizations with less than 10 employees. The majority of projects (91%) undertaken by organizations with less than 50 employees have been developed by private software companies. The largest organizations (>500 employees) are federal ministries, universities, telecommunications companies, a consortium company and a public software company.

Table 3 Type of organizations.



	Type of organization	Count	# of projects	%
Public	Software company	3	26	22%
	Federal directorates	3	6	5%
	University	3	4	3%
	Federal Ministry	2	8	7%
Private	Software company	23	62	52%
	Freelancer	3	8	7%
	Corporate IT department	3	4	3%
	Telecommunication company	2	2	2%
	Total	42	120	100%

Table 4 Organization size (in # of employees) during project implementation.



Organization Size	# of projects	%
1-5	15	13%
6-10	12	10%
11-20	22	18%
21-30	11	9%
31-40	2	2%
41-50	3	3%
51-100	8	7%
101-150	1	1%
151-200	5	4%
201- 350	-	-
351-400	3	3%
>500	30	25%
N/A	8	7%
	120	100%

Table 5 reflects the relatively small size of software/IT development departments, as 60% of projects are carried out by departments of less than 10 employees. The majority (90%) of projects are developed by private software companies with software development departments of less than 10 employees. The largest IT departments (> 50 employees) belong to corporations and telecommunications companies and a public software company. The dataset reflects the software development industry dominated by small and medium-sized organizations with limited staff.

Respondents were asked to define their roles throughout the life cycle of the project. Table 6 details the allocation of roles in the dataset. Project managers represent 63%

Respondents, followed by developers (27%) who worked on the projects. We note that the same respondent can have the same role for multiple projects. For 21 projects (18%) respondents reported multiple roles for the same project, e.g. project manager and developer or project manager, developer, and company manager. In these cases, we documented the role as the project manager and ignored the rest.

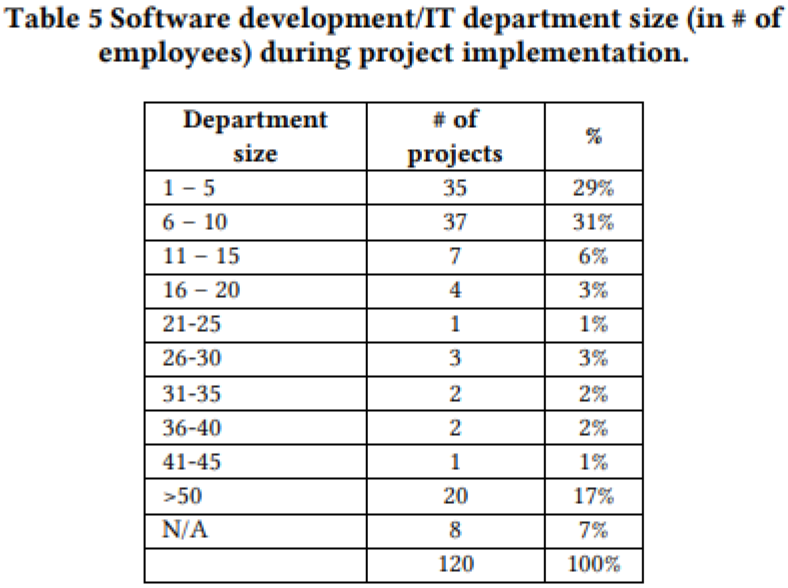


Table 5 Software development/IT department size (in # of employees) during project implementation.

Department size	# of projects	%
1 – 5	35	29%
6 – 10	37	31%
11 – 15	7	6%
16 – 20	4	3%
21-25	1	1%
26-30	3	3%
31-35	2	2%
36-40	2	2%
41-45	1	1%
>50	20	17%
N/A	8	7%
	120	100%

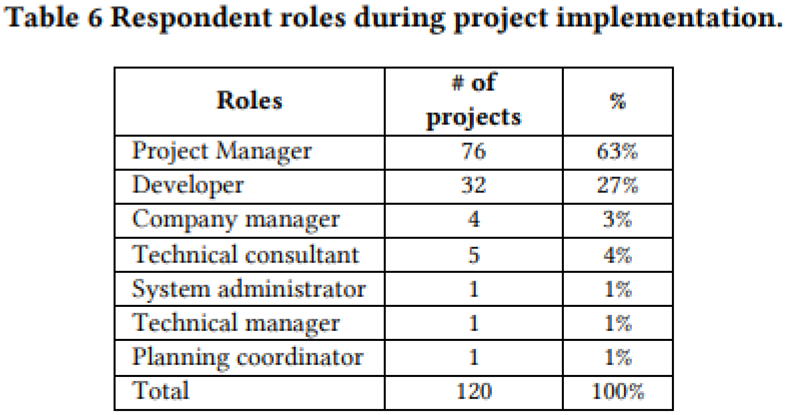


Table 6 Respondent roles during project implementation.

Roles	# of projects	%
Project Manager	76	63%
Developer	32	27%
Company manager	4	3%
Technical consultant	5	4%
System administrator	1	1%
Technical manager	1	1%
Planning coordinator	1	1%
Total	120	100%

3.2 Project Features

In this section, we provide an overview of the characteristics of software development projects in the SEERA dataset. Table 7 shows the year of development of the projects. We note that respondents are asked to provide: the date of the contract, the date of delivery of the contract software, and the actual start date of software development from which the value of the project year is inferred. The dataset contains relatively recent projects with the majority of projects (67%) started less than ten years ago and 43% of projects started after 2015. A small number (7%) of projects started before 2000, of which six were bank systems. With regards to software development methodology, Table 8 shows that an equal percentage of projects (~34%) have adopted a hybrid or waterfall approach. We note that the delivery of projects over time for both methods is similar; with more than 80% of projects started after 2004. Flexible methodologies are applied in less than 25% of projects, with 80% of flexible projects starting after 2016. A small number of projects (3%) reported taking no approach.

Table 10 details the development type and application domain of the SEERA project dataset. The majority (75%) of these projects are new software development projects including:

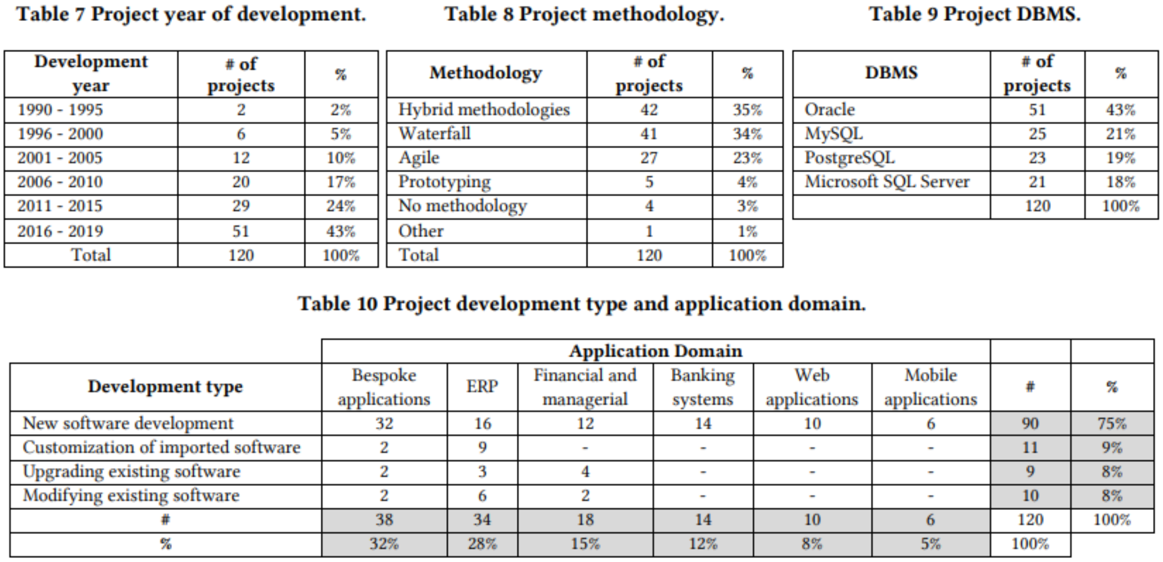


Table 7 Project year of development.

Development year	# of projects	%
1990 - 1995	2	2%
1996 - 2000	6	5%
2001 - 2005	12	10%
2006 - 2010	20	17%
2011 - 2015	29	24%
2016 - 2019	51	43%
Total	120	100%

Table 8 Project methodology.

Methodology	# of projects	%
Hybrid methodologies	42	35%
Waterfall	41	34%
Agile	27	23%
Prototyping	5	4%
No methodology	4	3%
Other	1	1%
Total	120	100%

Table 9 Project DBMS.

DBMS	# of projects	%
Oracle	51	43%
MySQL	25	21%
PostgreSQL	23	19%
Microsoft SQL Server	21	18%
	120	100%

Table 10 Project development type and application domain.

	Application Domain							
Development type	Bespoke applications	ERP	Financial and managerial	Banking systems	Web applications	Mobile applications	#	%
New software development	32	16	12	14	10	6	90	75%
Customization of imported software	2	9	-	-	-	-	11	9%
Upgrading existing software	2	3	4	-	-	-	9	8%
Modifying existing software	2	6	2	-	-	-	10	8%
#	38	34	18	14	10	6	120	100%
%	32%	28%	15%	12%	8%	5%	100%	

All banking system projects, web applications and mobile applications. In addition, 26% of new software development projects are based on open source software; This includes all types of application domains except the banking system. Less than 10% of projects are customized from imported software, of these, 91% are based on open source software (only one bespoke application is closed source). Projects that upgrade or modify existing systems, each representing 8% of projects and 22% and 60% of their projects, respectively, are based on open source software. In general, projects based on open source software account for 34% of projects in the SEERA dataset.

Regardless of the application areas presented in Table 10, all projects implement DBMS. Table 9 shows the popularity of the four database management systems, of which Oracle is the preferred technology for 43% of projects. With regards to the programming languages implemented in projects, there are 13 distinct programming languages and scripts, and in accordance with Table 9, Oracle Developer is the interface technology for 31% of projects, with Python being the second most popular language (20% of projects) and Java programming language being third (18% of projects). Eight projects reported the implementation of combining programming languages in one application. Full details of the distribution of programming languages are available in [22].

3.3 Dataset properties

Figure 1 details SEERA dataset properties. Properties are divided into eight categories: six (general information, scale, user, developer, project, and product) based on the categories described in [12], and two are effort that represents estimated and actual effort properties (the calculation formula is provided in [22]) and environment represents locally derived properties from [20, 21]. The categories and dataset attributes mainly correspond

to the sections of the questionnaire described in Section 2.

Each developer and project category represents 21% of attributes, followed by general information categories (17%), products (13%), users (11%), and environments (11%).

Due to the lack of implementation of cost estimation methods in Sudan's software industry, we followed the convention in the ISBSG questionnaire [27] in which respondents answered questions without prior knowledge of the ranking scale. This is contrary to the convention of the COCOMOII questionnaire [26]. Thus, the initial dataset directly reflected the questionnaire had 176 attributes. Some of the 176 properties are then grouped into categories, and the remaining properties are grouped to create new properties. This results in 76 attributes shown in eight categories in Figure 1. These properties represent 52 properties from the original 176 and 24 new properties where their scores are derived from the remaining properties (which we will call sub-attributes). We call the main dataset consisting of 76 attributes the SEERA dataset. Figure 1 shows some examples of attributes derived from sub-attributes. They are illustrated by an arrow from the main property to the set of sub-attributes, for example, the year of the project (general information) is taken from the three date properties. Attributes preceded by a preceding (*) are scored inverted, and sub-attributes preceded by a preceding (-) are not counted toward the scores of the parent attributes.

The SEERA dataset contains three types of attributes: (1) international attributes with names and questions adapted from international datasets, for example, all attributes of product categories. (2) International properties with localized questions, i.e., scores of international properties derived from questions (reflected as sub-attributes in Figure 1) based on local issues found to have an impact on the evaluation of these attributes,   For example, user objections and cohesion groups are international attributes but their sub-attributes are localized. (3) Localized attributes with localized questions (sub-attributes), i.e. attributes that are not present in international datasets but identified as significant factors in the impact on local software development costs, e.g. all

Environmental catalog attributes/sub-attributes. Localized categories, attributes, and sub-attributes are indicated by Δ IN Figure 1.

Catering to the local software development environment has resulted in some redundant sub-attributes not being included in the scoring of key attributes, for example, the economic instability impact attribute is only scored based on one of its three sub-attributes. Another example is the dimension attributes where we consider object points as the primary sizing method due to its simplicity [25], which results in two additional properties as a whole to other sizing methods (if no other sizing methods are reported these attributes are assigned values: N/A (not applicable) and is not considered a missing value). With regards to grading, we followed COCOMO's ratings, where more impact on cost will lead to higher scores. Details of attribute scoring methods and formulas are provided with the dataset [22].

The SEERA dataset includes the full set of sub-attributes, their initial values, and their score/ratings in addition to a dataset with only key attributes. Our aim is to facilitate further research into software cost estimation, that is, to use attributes/sub-attributes to create new datasets, to study different scaling/rating methods or groups of attributes, and to study generalizations of different cost models. Furthermore, sub-attributes can be used in other empirical software engineering studies, especially due to the limited number of similar datasets.

3.4 Project analysis

In this Section, we highlight some software development practices in SEERA dataset projects. We aim to provide general insights to identify some of the impacts that restrictive environments have on the success of software development projects. Our aim is to encourage further research into cost estimation and assessment modeling for such environments. Table 11 details the overall means and percentage of actual responses to the questionnaire. The naming conventions and results in Table 11 do not necessarily reflect the attributes/encoding of the SEERA cost estimate dataset.

Regarding the economics of the project, the estimated project duration is less than 6 months on average; however, the actual duration exceeded estimates by an average of 86%. The impact of economic instability was reported on 93% of projects, which included inflation and high developer switching rates. Software product dimensions reflect application domains (Table 10) and universal DBMS deployments (Table 9). However, previous research shows that practitioners assume that software size has minimal impact on cost and schedule [21]. Some degree of software reuse is evident in 63% of projects, however, outsourcing and combinations of open source software are not common. These problems were reflected in the financial losses of the project, as the overall loss was 4%. This includes 25% of projects reporting zero profits, 12% of projects reporting losses with an average loss of 47%, and only 16% of projects reporting profits with an average increase of 24%. There are 5 projects with losses of 100% or more. Given this fact, we consider that costing and planning is an important activity in a project, however, 33% of projects do not report project prices and costs incurred and 14% say they do not calculate actual costs incurred.

The underestimation of project timelines may be associated with a lack of adequate management procedures. From Table 11, approximately 50% of projects comply with the project schedule; has established work hours and policies to deal with productivity shortages. The limited use of standards and lack of tools and developer training do not correspond to the fact that 54% of projects are new to organizations. The majority of projects report that the customer environment is suitable with 40% reporting changes in requirements during various stages of development. Team experience (~75%) may reflect 25% of part-time and part-time national training/service team members, as national service workers are recent graduates with no previous experience and hired at minimum wage.

The majority of organizations in this study have small and medium-sized software development departments. However, there seems to be a tendency to reduce operating costs through temporary contracts and hiring inexperienced developers. This analysis corresponds to previous work that has identified a lack of expenditure in personnel development and training in Sudan's software industry [19]. This fact coupled with the lack of training and limitations on proper software development tools has most likely led to these projects being overwhelmed and unfortunately resulting in limited financial benefits. More research is needed to determine the relationship between these properties and their impact on other cost factors.

3.5 Threats to value

With regard to the risk of selection bias, i.e., it is possible that the submitted data reflect projects with which the respondent is already familiar, thus excluding projects undertaken by other employees or different time periods. This threat is difficult to remedy because most organizations appoint a senior project manager/manager to participate in the project. There is a risk of dependence on the respondent's memory, however 80% of projects rely on documents, 10% rely on documents and recollections and only 10% rely on recollections alone. Given the scale of the study, the quality assessment of all submitted projects, and the lack of previous work, we believe these limitations are reasonable and do not jeopardize the reliability of the dataset.

4 SEERA DATASET REVIEW

Bosu and Macdonell [29] introduced data quality classification to allow researchers to evaluate and compare the suitability and suitability of datasets for experimental software engineering research. The classification groups quality issues into three classes [29]: accuracy, relevance and origin. Accuracy refers to the accuracy of data and is evaluated through identifying: noise, exceptions, inconsistencies, incompleteness (missing values), and redundancy in the dataset. Relevance refers to assessing the relevance of data and is based on: heterogeneity (diversity), quantity of data, and timeliness (age) of the dataset. Provenance refers to the origin of a dataset and is evaluated based on: commercial sensitivity (evidence of anonymization or data transformation), accessibility (public availability), and reliability (source and ownership of the dataset).

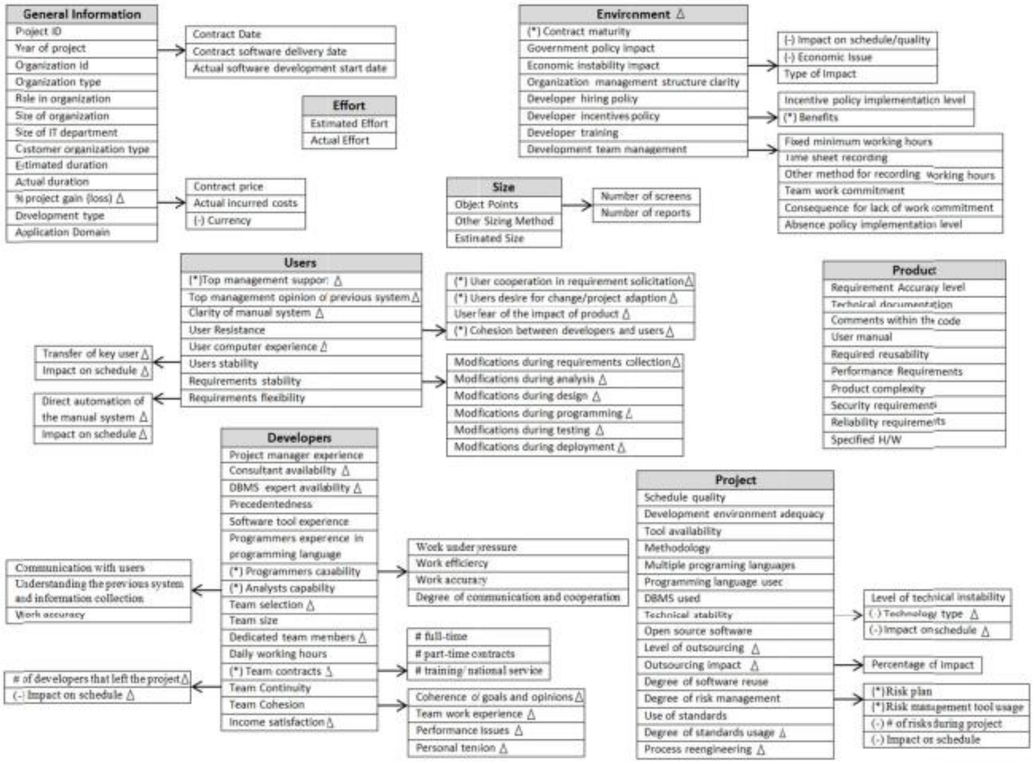


Figure 1 SEERA dataset attributes by category. (Attributes preceded with an (*) are reversed scored and sub-attributes preceded with a (-) are not included in the scoring of their main attributes. Localization is denoted by Δ: e.g., for a category this indicates that all its attributes are localized, for an attribute it indicates the attribute and its sub-attributes are localized and for sub-attributes this indicates that only the sub-attribute is localized.)

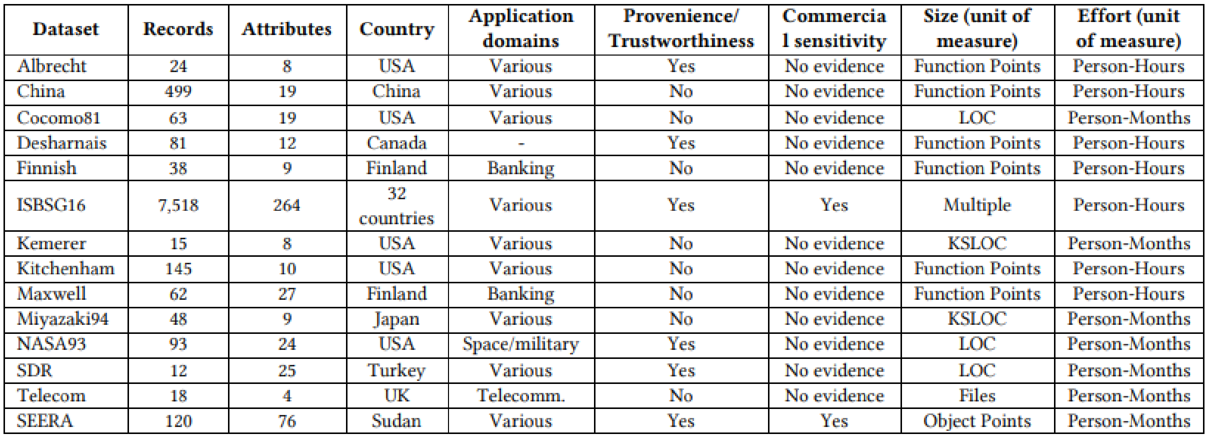
Table 11 SEERA dataset project analysis showing overall means.

Project Economics						
Duration		**Environment**		**Project Gain/Loss**		
estimated	5.7 months	% (-) impact economic instability	93%	overall mean loss		-4%
actual	10.4 months	% (-) impact government policy	52%	zero gain	25%	
Size		**Reuse**		+/- gain	28 %	gain = 24%
# of screens	99.09	% of open source	34%			loss = -47%
# of reports	106.18	% of outsourcing	10%	N/A	14%	
		evidence of software reuse	62%	missing	33%	

Project Management		**Team Composition**				
Adherence to schedule	53%	Team selection based on available developers		70%		
Adequate development environment	93%	Team contracts	Full-time	76%	Mean team size	5.7
Fixed minimum working hours	49%		Part-time	14%	% Dedicated	74%
Consequence for lack of work	50%		National service/ training	10%	Team cohesion	strong
Developer training	46%	Team is committed		90%		
Lack of tools	Yes	% Team continues to project completion		89%		
Use of standards	18%					

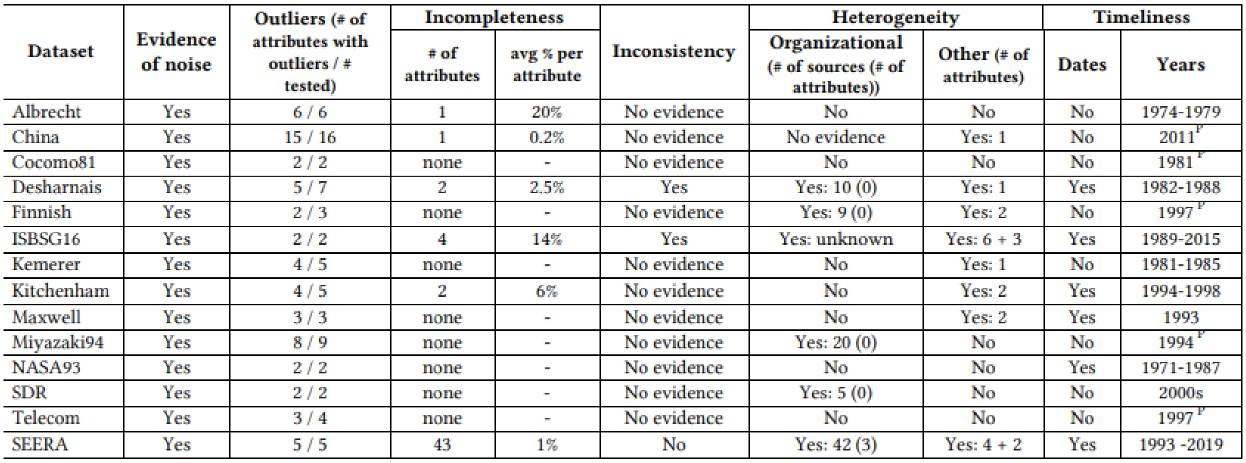
Customer Environment		**Team Experience**	
Top management support	supportive	Project manager previous experience in similar systems	68%
User stability	stable	Programmers are capable	75%
User resistance level	weak	Analysts are capable	77%
Requirements flexibility	flexible	Precedentedness	54%
Technical stability	very stable		
Requirements were stable	63%		

Table 12 Characteristics of the datasets. (PROMISE dataset data adapted from [4, 12] evaluation data adapted from [4]).



Dataset	Records	Attributes	Country	Application domains	Provenience/ Trustworthiness	Commercia l sensitivity	Size (unit of measure)	Effort (unit of measure)
Albrecht	24	8	USA	Various	Yes	No evidence	Function Points	Person-Hours
China	499	19	China	Various	No	No evidence	Function Points	Person-Hours
Cocomo81	63	19	USA	Various	No	No evidence	LOC	Person-Months
Desharnais	81	12	Canada	-	Yes	No evidence	Function Points	Person-Hours
Finnish	38	9	Finland	Banking	No	No evidence	Function Points	Person-Hours
ISBSG16	7,518	264	32 countries	Various	Yes	Yes	Multiple	Person-Hours
Kemerer	15	8	USA	Various	No	No evidence	KSLOC	Person-Months
Kitchenham	145	10	USA	Various	No	No evidence	Function Points	Person-Hours
Maxwell	62	27	Finland	Banking	No	No evidence	Function Points	Person-Hours
Miyazaki94	48	9	Japan	Various	No	No evidence	KSLOC	Person-Months
NASA93	93	24	USA	Space/military	Yes	No evidence	LOC	Person-Months
SDR	12	25	Turkey	Various	Yes	No evidence	LOC	Person-Months
Telecom	18	4	UK	Telecomm.	No	No evidence	Files	Person-Months
SEERA	120	76	Sudan	Various	Yes	Yes	Object Points	Person-Months

Table 13 Dataset data quality evaluation. (PROMISE dataset evaluation adapted from [4]).



Dataset	Evidence of noise	Outliers (# of attributes with outliers / # tested)	Incompleteness		Inconsistency	Heterogeneity		Timeliness	
			# of attributes	avg % per attribute		Organizational (# of sources (# of attributes))	Other (# of attributes)	Dates	Years
Albrecht	Yes	6 / 6	1	20%	No evidence	No	No	No	1974-1979
China	Yes	15 / 16	1	0.2%	No evidence	No evidence	Yes: 1	No	2011^{P}
Cocomo81	Yes	2 / 2	none	-	No evidence	No	No	No	1981^{P}
Desharnais	Yes	5 / 7	2	2.5%	Yes	Yes: 10 (0)	Yes: 1	Yes	1982-1988
Finnish	Yes	2 / 3	none	-	No evidence	Yes: 9 (0)	Yes: 2	No	1997^{P}
ISBSG16	Yes	2 / 2	4	14%	Yes	Yes: unknown	Yes: 6 + 3	Yes	1989-2015
Kemerer	Yes	4 / 5	none	-	No evidence	No	Yes: 1	No	1981-1985
Kitchenham	Yes	4 / 5	2	6%	No evidence	No	Yes: 2	Yes	1994-1998
Maxwell	Yes	3 / 3	none	-	No evidence	No	Yes: 2	Yes	1993
Miyazaki94	Yes	8 / 9	none	-	No evidence	Yes: 20 (0)	No	No	1994^{P}
NASA93	Yes	2 / 2	none	-	No evidence	No	No	Yes	1971-1987
SDR	Yes	2 / 2	none	-	No evidence	Yes: 5 (0)	No	No	2000s
Telecom	Yes	3 / 4	none	-	No evidence	No	No	No	1997^{P}
SEERA	Yes	5 / 5	43	1%	No	Yes: 42 (3)	Yes: 4 + 2	Yes	1993 -2019

Bosu and Macdonell [4] applied data quality classification to evaluate 13 publicly available software cost estimate datasets from the PROMISE repository [23]. In this section, we evaluate the quality of the SEERA dataset based on this classification and compare our results with the results of the data quality assessment for the PROMISE dataset presented in [4]. We note that the SEERA dataset was reviewed based on quality characteristics in [30] before conducting this comparison. Table 12 presents the general characteristics of the datasets and the results of the quality grade assessment of origin. Table 13 summarizes the assessment of accurate and appropriate quality grades.

Bosu and Macdonell [4] used data classification (decision tree algorithms) for noise recognition, i.e., the classification of records that incorrectly represented a measure representing noise. However, a noise record may be error-free, as real-world data may contain special records [31]. Noisy cases were identified for all 13 datasets [4]. The same method identified noisy cases in the SEERA dataset. Research has shown that

Exceptions are common in experimental software datasets [32]. The identification of exceptions and the reasons for their presence allows appropriate methods to be included / excluded and for better model method selection [4]. Table 13 shows that all PROMISE datasets have exceptions. However, no dataset gives justification for the presence of these exceptions.

For the SEERA dataset, as shown in [4], to identify outlier values, we excluded all taxonomic and limit scope attributes, so we tested five attributes. Figure 2 shows three of these attributes: estimated duration, actual duration, and group size. The exception rates for the estimated duration, actual duration, and group size attributes are 3%, 9%, and 7%, respectively. Exceptions at estimated and actual time intervals are related to application domains because 70% are closed source systems. Group size exceptions are for projects employing part-time team members and national training/service (on average

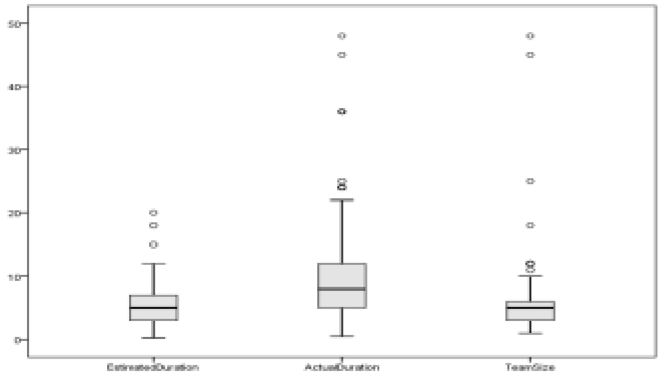


Figure 2 Boxplot showing outliers of *estimated duration*, *actual duration* and *team size* attributes.

from 20% to 90% of the group size). We've also identified exceptions for the object score attribute with a percentage of 3%, all three are open source projects and two of the three projects have team-size exceptions. For the % profit (loss) attribute of the project, the percentage of outlier values is 34% because the data is biased, since 47% of those who reported monetary costs had a % gain (loss) of the project equal to 0.

No redundancy was found in any of the 13 datasets, and only two had minor inconsistencies[4] and to the best of our knowledge, the SEERA dataset had no contradictions. Nevertheless

as discussed in Section 3, there exist a few redundant sub-attributes. Table 13 shows the average percentage of missing values per property based on the evaluation results in [4]; only five sets of PROMISE data are missing attribute values. Table 14 shows the attributes with missing values for the SEERA dataset. There are 43 attributes that are missing values, 86% that have less than two missing values (which, considering the space, only have their distribution according to the category shown), and 8% have three or four missing values. Group contracts and % profit (loss) of the project had the highest number of missing values, 11 and 39 missing values, respectively. The overall average percentage of missing values for each attribute is 1% as shown in Table 13. In terms of missing values for each project, Table 15 detailing the percentage of missing values in projects shows that the majority of projects (87%) have no value at all or one value is missing.

In terms of provenance, SEERA and all datasets are publicly accessible. However, from Table 12, only six (including SEERA) have origin metadata/reliability with source defined and dataset ownership. Commercial susceptibility is only evident in the ISBSG16 dataset and in SEERA, where organizations are represented by only one organization id. With regard to the relevance of datasets, Table 12 compares the amount of data for all datasets in terms of the number of records and attributes. Six of the 13 datasets had fewer than 50 records. The SEERA dataset has 120 records ranking fourth in size behind ISBSG16, China and Kitchenham datasets. The amount of data is an important characteristic when considering modeling assumptions for effort estimation [4].

Table 13 compares datasets in terms of heterogeneity and timeliness. In terms of timeliness, only five of the 13 SEERA datasets and datasets include attributes for project start/end dates. Only four datasets (including SEERA) have projects undertaken after 2000, which raises questions about relevance

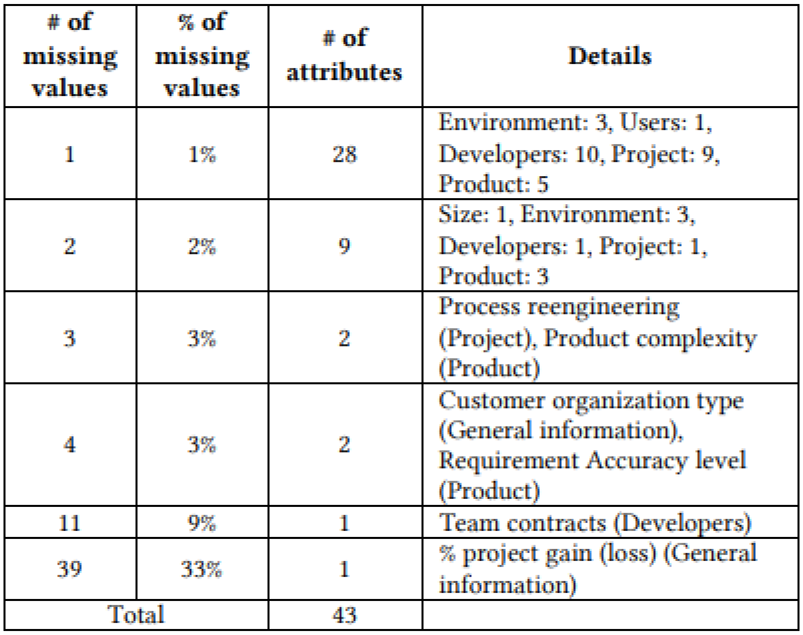
of older datasets for accurate cost estimation modeling [4]. From Table 13, five of the 13 datasets were collected from multiple organizations, i.e., organizational heterogeneity; however, no attributes exist in the dataset to distinguish the origin of each project [4]. In contrast, the SEERA dataset includes attributes to distinguish the origin and characteristics of the submitting organization: organization id, organization size, and IT department size.

We investigated additional heterogeneity factors that can form subsets of data in these datasets. This is based on a subset of the group attributes of the ISBSG16[33] and Kitchenham datasets, i.e. submit organization type, industry, development type, application domain, application type, client/client id, and programming language. From Table 13, six of the 13 datasets did not include any previous group attributes, and six included only one or two attributes. The ISBSG16 dataset includes all attributes except client/client id, and the SEERA dataset includes all attributes except application type and client/client id; However, it does include the Customer entity type attribute. The ISBSG16 dataset includes three additional group attributes: development platform,

language type and quantity approach (project size). SEERA

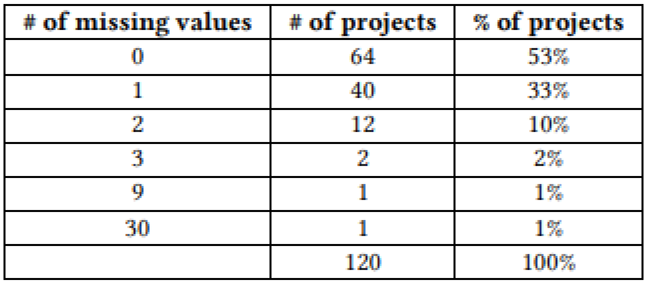
The dataset includes an additional attribute: the sender's role in the organization.

Table 14 Attributes with missing values in the SEERA dataset.



# of missing values	% of missing values	# of attributes	Details
1	1%	28	Environment: 3, Users: 1, Developers: 10, Project: 9, Product: 5
2	2%	9	Size: 1, Environment: 3, Developers: 1, Project: 1, Product: 3
3	3%	2	Process reengineering (Project), Product complexity (Product)
4	3%	2	Customer organization type (General information), Requirement Accuracy level (Product)
11	9%	1	Team contracts (Developers)
39	33%	1	% project gain (loss) (General information)
Total		43	

Table 15 Projects with missing values in the SEERA dataset.



# of missing values	# of projects	% of projects
0	64	53%
1	40	33%
2	12	10%
3	2	2%
9	1	1%
30	1	1%
	120	100%

When conducting the above comparison, the SEERA dataset provides recent heterogeneous project data with rich attributes that can be applied to various empirical research questions. The SEERA dataset overcomes current limitations in dataset transparency through providing detailed raw data (sub-attributes) and coding formulas that allow researchers to create new cost estimate datasets or scale existing attributes from the original data. This allows for the ability to replicate the results and verify the data. All of this combined increases the quality, flexibility, and reliability of SEERA datasets.

5 CONCLUSION AND FUTURE WORK

This paper presented the SEERA cost estimation dataset: a dataset for technically and economically constrained environments. It is the result of collecting data of 120 software development projects from 42 organizations in Sudan. The SEERA dataset contains 76 attributes and, unlike the current cost estimate dataset, is supplemented with metadata and sub-attributes containing raw data prior to encoding. This dataset fills the current gap in providing more relevant data to developing countries and the software industry in restrictive environments. Furthermore, it provides recently diverse data that allows researchers to compare the applicability of international approaches to confined environments and develop new techniques better suited to these environments. We plan to further analyze the SEERA dataset to investigate the impact of environmental and socioeconomic factors on technical cost factors. In addition, we will compare the prevalence and magnitude of these cost factors with the cost factors in the relevant PROMISE dataset. In addition, we will develop a cost estimation model using the SEERA dataset and compare it with known software cost calculation models. We anticipate that the SEERA cost estimation dataset will lead to more diverse software cost estimation research.