SEERA: A software cost estimation dataset for constrained environments

Emtinan I. Mustafa
Faculty of Mathematical Sciences
University of Khartoum
Khartoum, Sudan
eiomustafa@uofk.edu

Rasha Osman
Faculty of Mathematical Sciences
University of Khartoum
Khartoum, Sudan
rosman@ieee.org

ABSTRACT

The accuracy of software cost estimation depends on the relevancy of the cost estimation dataset, the quality of its data and its suitability for the targeted software development environment. Software development cost is impacted by technical, socio-economic and country-specific organizational and cultural environments. Current publicly available software cost estimation datasets represent 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 current cost estimation datasets, is augmented with metadata and the original raw data. This paper describes the data collection process, submitting organizations and project characteristics. In addition, we give a general analysis of the dataset projects to illustrate the impact of local factors on software project cost and compare the data quality of the SEERA dataset to public datasets from the PROMISE repository. The SEERA dataset fills a gap in the diversity of current cost estimation datasets and provides researchers with an opportunity to evaluate the generalization of previous and future cost estimation methods to constrained environments and to develop new techniques that are more suitable for these environments.

CCS CONCEPTS

• Software and its engineering • Software libraries and repositories

KEYWORDS

Software effort estimation, datasets, data quality, cost attributes, constrained environments, socio-economic factors, Africa

WOODSTOCK'18, June, 2018, El Paso, Texas USA
© 2018 Copyright held by the owner/author(s). 978-1-4503-0000-0/18/06...$15.00
https://doi.org/10.1145/1234567890

ACM Reference format:

Emtinan I. Mustafa and Rasha Osman. 2020. SEERA: A software cost estimation dataset for constrained environments. In *Proceedings of the 16th International Conference on Predictive Models and Data Analytics in Software Engineering (PROMISE'20), Nov 8-9, 2020, Sacramento, California, USA. ACM, New York, NY, USA, 10 pages.* https://doi.org/10.1145/3416508.3417119

1 INTRODUCTION

Software cost estimation methods utilize historical datasets of previous projects to estimate project costs and effort. These methods must encompass the ever-changing software development landscape [1] and the impact of country-specific environmental and cultural factors on software development practice [2]. Research in empirical software engineering and cost estimation has recognized that the relevance and the representativeness of the dataset are important for accurate and realistic cost estimation [3-6]. There should be a correspondence between the datasets, the models and the environment in which the results are to be applied. Nonetheless, most research in cost estimation relies on outdated datasets that may be unsuitable for current software development environments [3, 4].

Research in 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 practice. Research into cost estimation datasets concentrated on surveys of datasets (e.g., [11-13]), quality evaluations of datasets (e.g., [4, 14]) and addressing specific quality issues of current datasets (e.g., [15, 16]). Unfortunately, the number of datasets collected from the year 2000 and later do not exceed 25% of the available public cost estimation datasets [4, 12]. Moreover, all these datasets have been collected and evaluated from software development environments representative of North American and European settings, with no dataset of African or South American origin [12]. There were minimal factors reflecting the impact of the organizational environment and no factors reflecting socio-economic and cultural factors [12]. This limits the application of these datasets in countries with dissimilar societal and organizational norms, i.e., developing countries with constrained technical and economic environments. Furthermore, it questions the generalization of cost estimation methods to such constrained environments.

There are limited empirical investigations of software engineering practices in Africa [17]. Surveys have shown that the African software industry is locally focused with the majority of companies developing bespoke local applications or customizing imported applications [18, 19]. Previous studies of African software industries have emphasized the importance of socio-economic factors on the success of the software industry [19, 20]. Other studies have shown a difference in the ranking of critical success factors for software projects between African experts (e.g., ranking requirements solicitation and team capability highest) and European and US experts (e.g., ranking project management and cost estimation issues highest) [20, 21]. For these economically and technically constrained environments research and implementation of software cost estimation and scheduling should be consistent with the locally recognized factors and the relevant information and experience that impact software development costs. The level of granularity of cost attributes, how they are defined and collected should be consistent with the realities of these constrained environments.

This paper introduces the **SEERA** (**S**oftware engin**EER**ing in Sud**A**n) cost estimation dataset [22]. The SEERA dataset fills the current gap in cost estimation datasets in that (1) it provides a current dataset with traditional cost attributes in addition to socio-economic and organizational attributes. (2) The dataset projects represent constrained technical and economic software development environments, thus providing the international software engineering research community with a recent and diverse dataset to evaluate the generalization of previous and future costing models. (3) The dataset fills an urgent gap for the Sudanese and African research community with a more relevant cost estimation dataset that includes factors more aligned with the realities of their software industries. (4) The SEERA dataset overcomes the current limitations in dataset quality and transparency by augmenting the cost estimation dataset with the original raw data before coding/scaling. This allows researchers the flexibility in creating new cost estimation datasets (e.g., a COCOMO-style dataset from the original SEERA data), sub-datasets (e.g., excluding attributes to represent different environments) or rescaling of the attributes from the original SEERA data. This allows for the replicability of results and the general application of the dataset for international research.

The main contributions of this work are as follows:

- The SEERA dataset. A curated set of software development projects from a technically and economically constrained environment. The dataset contains detailed attributes and factors that are more suitable to the cost and schedule information available within these environments. The SEERA dataset is publicly available in the Zenodo repository [22].
- We provide a detailed description of the characteristics of the projects within the dataset in terms of submitting organization, development type and application domain. We describe the attributes of the dataset, their categories and types.
- We explain the local attributes reflecting the constrained technical and economic environment and their relationship with the international attributes and categories. We illustrate the importance of these local attributes through an analysis of project data which shows the impact of these attributes on project cost and duration.
- We provide a data quality assessment of the SEERA dataset in comparison to the PROMISE repository cost estimation datasets [23]. The SEERA dataset includes the original raw data before coding/scaling and is augmented with metadata thus increasing its transparency and trustworthiness in comparison to the PROMISE datasets.

The rest of this paper is organized as follows. Section 2 details the study design and the SEERA dataset collection. The characteristics of the dataset in terms of: submitting organizations, project characteristics, attribute descriptions and project analysis are in Section 3. We compare the data quality of the SEERA dataset to that of the PROMISE repository datasets in Section 4. Conclusions and future work are discussed in Section 5.

2 SEERA DATASET COLLECTION

2.1 Study Design

To facilitate the collection of software project data we designed a questionnaire that incorporated international software project costing factors (adapted from [12]) and factors specific to the software development industry in developing countries (adapted from [12, 19-21, 24]). We designed the questions following the convention in [25, 26], i.e., multiple questions are used to collect data relating to one cost factor. The questions were focused on local issues and factors as described in [20, 21], e.g., we asked about developer hiring and incentive policies within the organization. Most practitioners are not familiar with costing terminology and therefore the questionnaire reflected local realities without stating the explicit attributes and rating scales. For example, most systems are database/information systems and thus we asked directly for the *number of screens/reports* without referring to the *object points* attribute. In addition, the questionnaire was prepared in the Arabic language. The questionnaire was nine pages long and contained three sections as follows.

- **Respondent information**. Name, employer and contact details. This is to contact the respondent in case of any missing or ambiguous answers within the questionnaire. The idea for this section was adapted from [27].
- **Project general information**. Information relating to the software project: developing organization size, customer organization, estimated and actual schedule, application domain and size, development and methodology type, team size and team hiring policies. This section consisted of 20 short open ended questions, 11 single-select and two multiple-select multiple choice questions. Six questions relating to product size and project schedule were adapted from [26].
- **Factors affecting software product development**. This is the largest section of the questionnaire

containing six subsections. Table 1 details each subsection and the number of questions. For each subsection the questions were a mix of single-select or multiple-select multiple choice questions, Likert-type questions, closed end questions and a few short open-ended questions.

The subsections of the *Factors affecting software product development* reflect the local environmental factors that impact software development cost in developing countries. For example, the *organization environment* subsection has attributes and related questions that were derived from factors identified in [20, 21] which differ from attributes of current software cost estimation datasets [12]. In addition, the *Users* and *Project management* subsections were based on local factors with only two and three cost attributes, respectively, adapted from [26] but with different localized questions. In contrast, technical issues relating to software development were based on international attributes, e.g., all the cost factors of the *Team*, *Product* and *Product complexity* subsections were adapted 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 Study Execution

To identify the targeted software development organizations we referred to the organizations identified in [19], the organizations recognized by the Sudanese National Information Center [28], and organizations suggested by experts in the Sudanese software industry [24]. This resulted in identifying 70 organizations. The questionnaire was distributed and collected from June 2019 to February 2020. Data entry, reviewing and analysis were conducted in parallel.

Organizations were contacted through direct visits (57%) or by telephone, email or chatting (43%). Printed questionnaires were provided during direct visits and an electronic version was emailed to the contact person within the targeted organization. Table 2 details the number of distributed and returned questionnaires and the total number of collected projects. The total number of organizations responding to the questionnaire (42 organizations) represented 60% of the overall targeted institutions, with the organizations that were directly visited representing 76% of the respondents. We note that we received responses from individuals from some visited organizations who implemented projects as freelancers. For simplicity, we have assumed each freelancer as a different entity, i.e., organization.

On the outset, each contacted organization pledged to provide information for a certain number of projects of which we calculated the *expected number of projects* in Table 2. However, during collection, 40% of the organizations did not respond or decided not to participate further in the study. The organizations that did respond only provided on average 50% of the pledged number of projects. Thus the collected number of projects was only 30% of the anticipated number and 85% of these projects were collected through direct visits to the organizations.

We note that organizations did not return the questionnaires by the agreed dates which required multiple visits to some institutions and contact persons. In some cases, this was due to the lack of documentation which required the respondent to refer to the original contracts, previous team members or the software product itself. Each collected questionnaire was revised for ambiguous answers and missing data. This resulted in another round of contacting individuals to complete the data for 80% of the returned questionnaires. A further quality evaluation of the collected projects led to the exclusion of 10 projects: four duplicated projects, two projects with incomplete schedule information, two projects that were submitted by users and two projects that were abandoned before completion. Thus the total number of projects within the SEERA dataset is 120 projects. The next section will detail the characteristics of these projects.

3 SEERA DATASET CHARACTERISTICS

3.1 Organization 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 within public and private institutions. Table 3 provides the details of the organizations contributing project data. The public sector represents 26% of the organizations with a contribution of 37% of the projects. Only public sector software companies developed software for customers, the rest of the public organizations provided in-house software projects developed by their respective IT departments.

Private software companies contributed 52% of the total projects and 82% of the projects contributed by the private sector. However, the average contribution of each private software company is one to three projects with one company contributing 11 projects. This is in contrast to the public software companies in which two companies contributed 16 and 8 projects and one company contributed two projects. To reflect the heterogeneity of the projects, the dataset includes attributes for the *type of organization*, *sector* and *organization id*.

The respondents were asked to provide the organization and software development department sizes (in number of employees) during the project lifetime. For organizations in which in-house software development was conducted by the IT department, the size of the IT department was provided. Table 4 and Table 5 detail this information. From Table 4, the majority of projects (61%) were implemented by relatively small organizations with less than 100 employees, with 23% of the projects implemented by organizations of less than 10 employees. The majority of the projects (91%) implemented by organizations of less than 50 employees have been developed by private software companies. The largest organizations (>500 employees) are the federal ministries, the universities, the telecommunication companies, one corporate company and one 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 development/IT departments, as 60% of the projects were implemented by departments of less than 10 employees. The majority (90%) of projects developed by private software companies had software development departments of less than 10 employees. The largest IT departments (> 50 employees) belong to the corporate and telecommunication companies and one public software company. The dataset reflects a software development industry dominated by small to medium size organizations with limited staff.

Respondents were asked to specify their roles during the lifetime of the project. Table 6 details the distribution of roles within the dataset. Project managers represented 63% of the respondents, followed by developers (27%) who worked on the projects. We note that the same respondent may have had the same role for more than one project. 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 recorded the roles as project manager and disregarded 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 Characteristics

In this section, we give an overview of the characteristics of the software development projects within the SEERA dataset. Table 7 shows the year of development of the projects. We note that the respondents were asked to provide: the *contract date*, *contract software delivery date* and the *actual software development start date* from which the value of the *year of project* was inferred. The dataset contains relatively recent projects with the majority of projects (67%) initiated less than ten years ago and 43% of projects beginning after the year 2015. A minority (7%) of projects began before the year 2000 of which six are banking systems. In regards to the software development methodology, Table 8 shows that an equal percentage of projects (~34%) applied a hybrid or waterfall methodology. We note that the distribution of projects over time for both methodologies is similar; with more than 80% of projects beginning after 2004. Agile methodologies were applied in less than 25% of the

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%	

projects, with 80% of agile projects starting after 2016. A minority of projects (3%) reported applying no methodology.

Table 10 details the development type and application domain of the SEERA project dataset. A majority (75%) of these projects are new software development projects which include all the banking system, web application and mobile application projects. In addition, 26% of the new software development projects are based on open source software; this includes all types of application domains except the banking systems. Less than 10% of the projects are customized from imported software, of these, 91% are based on open source software (only one bespoke application is closed source). Projects that are upgrades or modification of existing systems each represent 8% of the projects and 22% and 60% of their projects, respectively, are based on open source software. In general, projects based on open source software represent 34% of the projects within the SEERA dataset.

Irrespective of the application domains presented in Table 10, all the projects implement a DBMS. Table 9 show the popularity of four DBMSs, with the Oracle DBMS being the preferred technology for 43% of the projects. In regard to the programming languages implemented within the projects, there were 13 distinct programming and scripting languages, and in agreement with Table 9, Oracle Developer is the interface technology for 31% of the projects, with the Python programming language being the second popular (20% of the projects) and the Java programming language the third (18% of the projects). Eight projects reported implementing a combination of programming languages within one application. Complete details of the distribution of programming languages are available in [22].

3.3 Dataset Attributes

Figure 1 details the SEERA dataset attributes. The attributes are divided into eight categories: six (*general information*, *size*, *users*, *developers*, *project* and *product*) are based on the categories described in [12] and two are *effort* representing the estimated and actual effort attributes (calculation formulas provided in [22]) and *environment* representing local attributes derived from [20, 21]. The dataset categories and attributes mostly correspond to the sections of the questionnaire described in Section 2. The developers and project categories each represent 21% of the attributes, followed by the general information (17%), product (13%), users (11%) and environment (11%) categories.

Due to the lack of implementation of cost estimation methods within the Sudanese software industry, we followed the convention in the ISBSG questionnaire [27] in which respondents answered questions without prior knowledge of the rating scale. This is in contrast to the convention of the COCOMOII questionnaire [26]. Therefore, the initial dataset that directly reflects the questionnaire contained 176 attributes. Then some of the 176 attributes were grouped into categories and the remaining attributes were grouped to create new attributes. This resulted in the 76 attributes represented within the eight categories in Figure 1. These attributes represent 52 attributes from the original 176 attributes and 24 new attributes in which their score is derived from the remaining attributes (which we will refer to as *sub-attributes*). We refer to the main dataset that encompasses the 76 attributes as the SEERA dataset. Figure 1 shows some examples of attributes that are derived from sub-attributes. These are illustrated with an arrow from the main attribute to the set of sub-attributes, e.g., *year of project* (*general information*) is derived from three date attributes. Attributes that are preceded with an *(*)* are reversed scored and sub-attributes that are preceded with a *(-)* are not included in the scoring of their main attributes.

The SEERA dataset contains three types of attributes: (1) international attributes whose names and questions are adapted from international datasets, e.g., all attributes of the *product* category. (2) International attributes with localized questions, i.e., the scoring of the international attribute is derived from questions (reflected as sub-attributes in Figure 1) based on local issues found to impact the evaluation of these attributes, e.g., *user resistance* and *team cohesion* are international attributes but their sub-attributes are localized. (3) Localized attributes with localized questions (sub-attributes), i.e., attributes not included in international datasets but were identified as important factors in local software development cost impact, e.g., all the attributes/sub-attributes of the *environment* category. Localized categories, attributes and sub-attributes are denoted with a Δ in Figure 1.

Catering to the local software development environment has led to some redundant sub-attributes that have not been included in the scoring of the main attributes, e.g., the attribute *economic instability impact* is scored based on only one of its three sub-attributes. Another example is the size attributes in which we considered *object points* to be the main method of sizing due to its simplicity [25], which resulted in two extra attributes as a catch-all for other sizing methods (if no other sizing method is reported these attributes are assigned the value: *N/A* (not applicable) and are not considered a missing value). In regards to scoring, we have followed the COCOMO rating levels in which a higher impact on cost results in a higher score. Details of the attribute scoring methods and formulas are provided with the dataset [22].

The SEERA dataset includes the complete set of sub-attributes, their original values and scores/ratings in addition to the dataset with the main attributes only. Our intention is to facilitate further research in software cost estimation, i.e., using the attributes/sub-attributes to create new datasets, to research different scaling/rating methods or groupings of attributes and researching the generalization of different costing models. Furthermore, the sub-attributes can be used in other empirical software engineering studies, especially given the limited availability of similar datasets.

3.4 Project Analysis

In this Section, we highlight some of the software development realities within the projects of the SEERA dataset. We aim to give general insights to identify some of the impacts that constrained environments have on the success of software development projects. Our aim is to encourage further research in cost estimation modelling and evaluations for such environments. Table 11 details the overall means and percentages of *actual responses* to the questionnaire. The naming conventions and results in Table 11 do not necessarily reflect the attributes/coding of the SEERA cost estimation dataset.

In regard to project economics, the average estimated project duration is less than 6 months; however, the actual durations exceeded the estimate by an average of 86%. The impact of economic instability was reported for 93% of the projects, this included inflation and high developer turn over. The software product size reflects the application domains (Table 10) and universal DBMS implementation (Table 9). However, previous research suggests practitioners viewed that software size has minimal impact on cost and schedule [21]. Some level of software reuse is evident in 63% of projects, however, outsourcing and the incorporation of open source software is not prevalent. These issues have been reflected in project financial losses, as the overall loss was 4%. This includes 25% of projects reporting zero gains, 12% of projects reporting losses with an average of 47% loss and only 16% of projects reporting gains with an average of 24% gain. Five projects reported losses of 100% and above. Given this reality, we would expect costing and scheduling to be an important activity within a project, however, 33% of the projects did not report project price and incurred costs and 14% stated that they do not calculate actual incurred costs.

Underestimation of project duration may be related to the lack of adequate management procedures. From Table 11, about 50% of projects adhere to a project schedule; have set working hours and policies for dealing with lack of productivity. Limited use of standards and lack of tools and developer training does not correspond to the fact that 54% of projects were new to the organizations. The majority of projects reported that the customer environment was adequate with 40% reporting changes in requirements during different development phases. Team experience (~75%) may reflect the 25% part-time and national service/training team members, as national service personnel are new graduates with no previous experience who are hired with minimal wages.

The majority of organizations in this study have small to medium sized software development departments. Nonetheless, it seems there is a tendency to lower running costs through temporary contracts and the hiring of inexperienced developers. This analysis corresponds with previous work that has identified the lack of expenditure in personnel development and training in the Sudanese software industry [19]. This reality coupled with the lack of training and the limited availability of suitable software development tools has most likely led to the overrun of these projects and unfortunately resulting in the limited financial gains. Further research is required to determine the relationships between these attributes and their impact on other cost factors.

3.5 Threats to Validity

In regard to the threat of selection bias, i.e., it is possible that the submitted data reflects the projects that the respondent was familiar with, thus excluding projects implemented by other personnel or different time periods. This threat was difficult to overcome, as most organizations assigned one senior project manager/director to participate in the project. There is a threat of the reliance on the recollection of the respondents, however 80% of the projects relied on documentation, 10% relied on documentation and recollection and only 10% relied on recollection only. Given the size of the study, the quality review of all submitted projects and the lack of previous work, we believe that these limitations are reasonable and do not jeopardize the credibility of the dataset.

4 SEERA DATASET EVALUATION

Bosu and Macdonell [29] introduced the *data quality taxonomy* to enable researchers to evaluate and compare the suitability and relevance of datasets for empirical software engineering research. The taxonomy groups quality issues into three classes [29]: accuracy, relevance and provenance. *Accuracy* is concerned with the correctness of data and is assessed through the identification of: noise, outliers, inconsistency, incompleteness (missing values) and redundancy within a dataset. *Relevance* is concerned with evaluating the suitability of the data and is based on: heterogeneity (diversity), amount of data and timeliness (age) of the dataset. *Provenance* is concerned with the origin of the dataset and is assessed based on: commercial sensitivity (evidence of data anonymization or transformation), accessibility (public availability) and trustworthiness (documented source and ownership of the dataset).

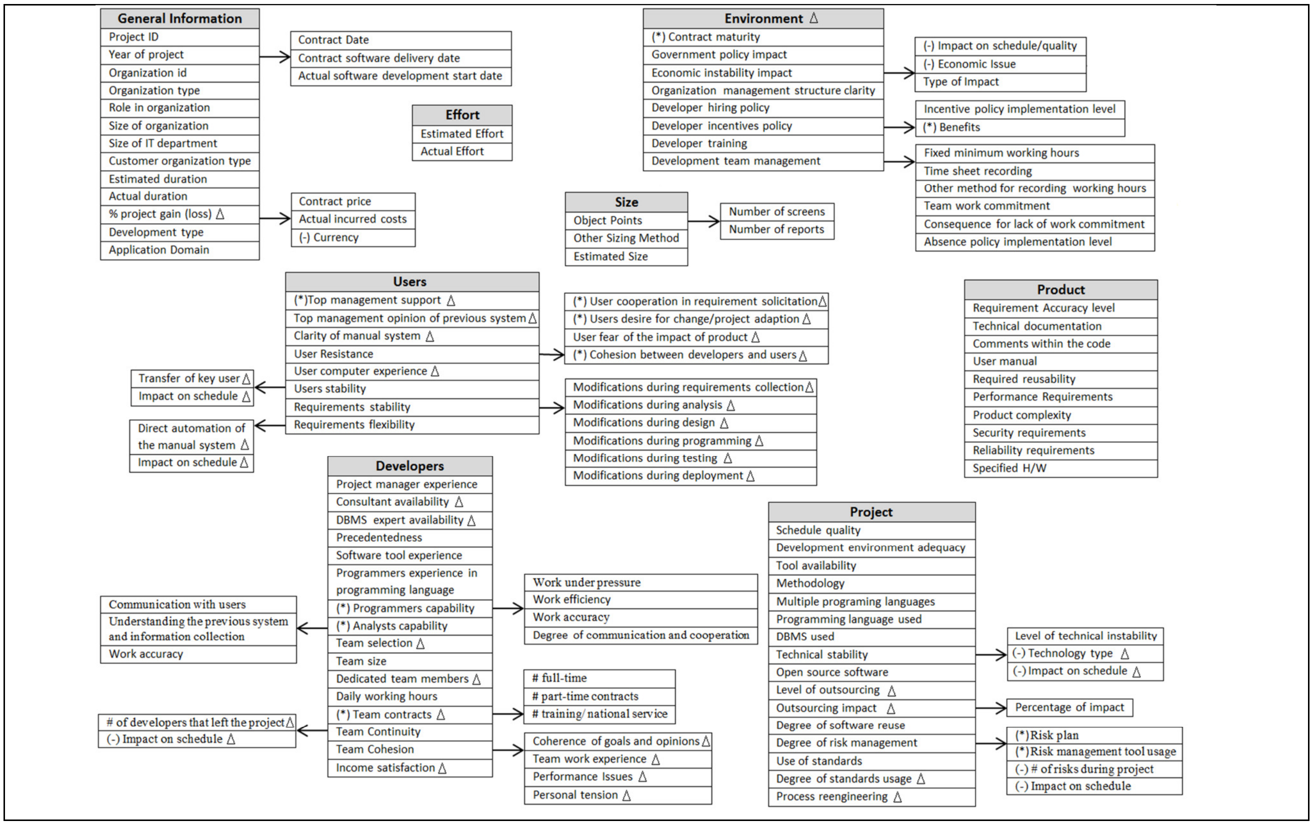

Figure 1 SEERA dataset attributes by category. (Attributes preceded with an (*) are reversed scored and sub-attributes preceded with a (-) are not included in the scoring of their main attributes. Localization is denoted by Δ: e.g., for a category this indicates that all its attributes are localized, 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	Commercial 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	44	1%	No	Yes: 42 (3)	Yes: 4 + 2	Yes	1993 -2019

P: date based on first publication date

Bosu and Macdonell [4] applied the data quality taxonomy to evaluate 13 publicly available software cost estimation datasets from the PROMISE [23] repository. In this section, we evaluate the quality of the SEERA dataset based on this taxonomy and compare our results with the data quality evaluation results for the PROMISE datasets presented in [4]. We note that the SEERA dataset was reviewed based on the quality characteristics in [30] before conducting this comparison. Table 12 presents the general characteristics of the datasets and the evaluation results for the provenance quality class. Table 13 summarizes the evaluation for the accuracy and relevance quality classes.

Bosu and Macdonell [4] utilized data classification (a decision tree algorithm) for noise identification, i.e., incorrect classification of a record represents a proxy measure for noise. However, a noisy record may not be *erroneous*, as real world data may contain exceptional records [31]. Noisy instances were identified for all 13 datasets [4]. The same method identified noisy instances in the SEERA dataset. Research has shown that outliers are common in empirical software datasets [32]. The identification of outliers and the reasons for their presence allows for implementing suitable methods for inclusion/exclusion of outliers and for better model method selection [4]. Table 13 shows that all the PROMISE datasets had outliers. However, no dataset provided justifications for the presence of these outliers.

For the SEERA dataset, as in [4], to determine outliers we excluded all categorical and limited range attributes and thus we tested five attributes. Figure 2 shows three of these attributes: *estimated duration*, *actual duration* and *team size*. The percentage of outliers for the estimated duration, actual duration and team size attributes is 3%, 9% and 7% respectively. Outliers for the estimated and actual duration were related to the application domain as 70% were closed source systems. Outliers

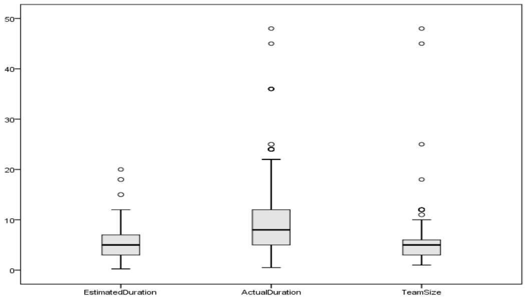

Figure 2 Boxplot showing outliers of *estimated duration*, *actual duration* and *team size* attributes.

for team size were for projects that utilized part-time and training/national service team members (ranging on average from 20% to 90% of team size). We also determined outliers for the *object points* attribute with a percentage of 3%, all three are open source projects and two of the three projects had outliers for *team size*. For the *% project gain (loss)* attribute the percentage of outliers was 34% due to the skewness in the data, as 47% of those reporting monetary cost had *% project gain (loss)* of zero.

No redundancy was found in any of the 13 datasets and only two datasets had minor inconsistencies [4] and to the best of our knowledge the SEERA dataset has no inconsistencies. However, as discussed in Section 3 there exist a few redundant sub-attributes. Table 13 shows the average percentage of missing values per attribute based on the results of the evaluation in [4]; only five PROMISE datasets had missing attribute values. Table 14 shows the attributes with missing values for the SEERA dataset. There are 44 attributes with missing values, 86% have two or less missing values (due to space considerations only their distribution by category is shown) and 8% have three or four missing values. The *team contracts* and *% project gain (loss)* have 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 regard to missing values per project, Table 15 details the percentage of missing values within the projects showing that the majority of projects (86%) have none or one missing value.

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

Table 13 compares the datasets in terms of heterogeneity and timeliness. In regard to timeliness, only five of the 13 datasets and the SEERA dataset included attributes for project start/finish dates. Only four datasets (including SEERA) have projects implemented after the year 2000, which questions the suitability of older datasets for accurate cost estimation modelling [4]. From Table 13, five of the 13 datasets were collected from multiple organizations, i.e., *organizational heterogeneity*; however, no attributes exist within the datasets to distinguish the origins of each project [4]. In contrast, the SEERA dataset includes attributes to distinguish the origins and characteristics of the submitting organization: *organization id*, *organization size*, and *IT department size*.

We further investigated other factors of heterogeneity that can form data subsets within these datasets. This was based on a subset of the grouping attributes of the ISBSG16 [33] and Kitchenham datasets, i.e., *submitting organization type*, *industry sector*, *development type*, *application domain*, *application type*, *customer/client id* and *programming language*. From Table 13, six of the 13 datasets did not include any of the previous grouping attributes and six included only one or two attributes. The ISBSG16 dataset included all attributes except the *customer/client id* and the SEERA dataset included all attributes except *application type* and *customer/client id*; however, it included the *customer organization type* attribute. The ISBSG16 dataset included three extra grouping attributes: *development platform*, *language type* and *count (project sizing) approach*. The SEERA dataset included an extra attribute: *submitter role in 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%	10	Size: 1, Environment: 3, Developers: 1, Project: 2, 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		44	

Table 15 Projects with missing values in the SEERA dataset.

# of missing values	# of projects	% of projects
0	63	53%
1	40	33%
2	13	11%
3	2	2%
9	1	1%
30	1	1%
	120	100%

In conducting the above comparison, the SEERA dataset provides recent heterogeneous project data with rich attributes that can be applied for different empirical research questions. The SEERA dataset overcomes the current limitations in dataset transparency through providing detailed original raw data (sub-attributes) and coding formulas which allows researchers to create new cost estimation datasets or rescale current attributes from the original data. This allows for the replicability of results and the verification of the data. All this combined raises the quality, flexibility and trustworthiness of the SEERA dataset.

5 CONCLUSIONS AND FUTURE WORK

This paper presented the SEERA cost estimation dataset: a dataset for technically and economically constrained environments. It is the result of the collection of 120 software development project data from 42 organizations in Sudan. The SEERA dataset contains 76 attributes and, unlike current cost estimation datasets, is augmented with metadata and sub-attributes containing the raw data before coding. This dataset fills the current gap in providing data more relevant to developing countries and software industries in constrained environments. Moreover, it provides recent diverse data that will allow researchers to compare the applicability of international methods to constrained environments and develop new techniques that are more suitable for these environments. We plan to further analyze the dataset to investigate the impact of environmental and socio-economic factors on technical cost factors. Moreover, we will compare the prevalence and magnitude of cost factors to those of relevant PROMISE datasets. In addition, we will develop a cost estimation model using the SEERA dataset and compare with known software costing models. We anticipate that the SEERA cost estimation dataset will lead to more diverse software cost estimation research.

REFERENCES

[1] Barry Boehm. 2017. Software cost estimation meets software diversity, in *39th ICSE'17*, May 2017, Buenos Aires, Argentina. 495–496.

[2] Ansgar Lamersdorf, Jürgen Münch, Alicia Fernandez-del Viso Torre, Carlos Rebate Sánchez, and Dieter Rombach. 2010. Estimating the Effort Overhead in Global Software Development, in *5th IEEE ICGSE*, Aug. 23-26, 2010, Princeton, NJ,. 267-276.

[3] Magne Jørgensen and Martin Shepperd. 2007. A systematic review of software development cost estimation studies. *IEEE Trans. Software Eng.,* vol. 33, 1, pp. 33-53, Jan. 2007.

[4] Michael F. Bosu and Stephen G. MacDonell. 2019. Experience: Quality Benchmarking of Datasets Used in Software Effort Estimation. *Journal of Data and Information Quality,* vol. 11, 4, Article 19, 38 pages, Aug. 2019.

[5] Peter Alexander Whigham, Caitlin A Owen, and Stephen G. MacDonell. 2015. A Baseline Model for Software Effort Estimation. *ACM Trans. Softw. Eng. Methodol.,* vol. 24, 3, Article 20, 11 pages, May 2015.

[6] Vahid Khatibi Bardsiri, Dayang Norhayati Abang Jawawi, Siti Zaiton Mohd Hashim, and Elham Khatibi. 2014. A flexible method to estimate the software development effort based on the classification of projects and localization of comparisons. *Empir. Softw. Eng.,* vol. 19, pp. 857–884, Aug. 2014.

[7] François Bergeron and Jean Yves St-Arnaud. 1992. Estimation of information systems development efforts: a pilot study. *Information and Management,* vol. 22, 4, pp. 239–254, April 1992.

[8] Parag C. Pendharkar and James A Rodger. 2007. An empirical study of the impact of team size on software development effort. *Information Technology and Management,* vol. 8, 4, pp. 253–262, Dec. 2007.

[9] Barbara Ann Kitchenham. 1992. Empirical studies of assumptions that underlie software cost-estimation models. *Information and Software Technology,* vol. 34, 4, pp. 211-218, April 1992.

[10] Ricardo Britto, Vitor Freitas, Emilia Mendes, and Muhammad Usman. 2014. Effort Estimation in Global Software Development: A Systematic Literature Review, in *9th ICGSE*, Aug. 18-21, 2014, Shanghai, China. 135-144.

[11] Carolyn Mair, Martin Shepperd, and Magne Jørgensen. 2005. An analysis of data sets used to train and validate cost prediction systems, in *PROMISE '05*, May 15, 2005, St. Louis, Missouri, USA. 1–6.

[12] Emtinan I. Mustafa and Rasha Osman. 2018. An Analysis of the Inclusion of Environmental Cost Factors in Software Cost Estimation Datasets, in *IEEE QRS Workshop on Conflicts and Synergies among Reliability, Security, and other Qualities*, July 16 – 19, 2018, Lisbon, Portugal. 623-630.

[13] Karel Dejaeger, Wouter Verbeke, David Martens, and Bart Baesens. 2012. Data mining techniques for software effort estimation: a comparative study. *IEEE Trans. Softw. Eng.,* vol. 38, 2, pp. 375-397, March 2012.

[14] Martin J. Shepperd, Qinbao Song, Zhongbin Sun, and Carolyn Mair. 2013. Data quality: Some comments on the NASA software defect datasets. *IEEE Trans. Softw. Eng.,* vol. 39, 9, pp. 1208–1215, Sept. 2013.

[15] Jason Van Hulse and Taghi M. Khoshgoftaar. 2014. Incomplete-case nearest neighbor imputation in software measurement data. *Information Sciences,* vol. 259, pp. 596–610, Feb. 2014.

[16] Wen Zhang, Ye Yang, and Qing Wang. 2011. Handling missing data in software effort prediction with naive Bayes and EM algorithm, in *7th Promise '11,* 1–10.

[17] Akinola S. Olalekan. 2005. Conducting empirical software engineering research in Nigeria: the posing problems, in *27th ICSE 2005*, Saint Louis, MO, USA.

[18] H. Abimbola Soriyan, Anja Mursu, Adebayo D Akinde, and Mikko Korpela. 2001. Information Systems Development in Nigerian Software Companies: Research Methodology and Assessment from the Healthcare Sector's Perspective. *EJISDC,* vol. 5, 4, pp. 1-18, May 2001.

[19] Saleh Alamdi and Rasha Osman. 2017. The Realities of the Software Industry in Sudan (in Arabic). *JECS,* vol. 8, 2, pp. 5-64, 2017.

[20] Anja Mursu, Kalle Lyytinen, H. Abimbola Soriyan, and Mikko Korpela. 2003. Identifying software project risks in Nigeria: an international comparative study. *Eur. J. Inf. Syst,* vol. 12, 3, pp. 182–194, 2003.

[21] Emtinan Mustafa and Rasha Osman. 2019. Identifying Critical Success Factors for the Sudanese Public Sector Software Projects (in Arabic). *JECS,* vol. 20, 2, pp. 2-19, 2019.

[22] Emtinan I. Mustafa and Rasha Osman. 2020. The SEERA Software Cost Estimation Dataset. Zenodo. http://doi.org/10.5281/zenodo.3987969

[23] The OpenScience tera-PROMISE Repository. 2019. Retrieved from: http://openscience.us/repo

[24] Saleh Alamdy and Rasha Osman. 2017. Software industry practice in Africa: case study Sudan, in *41st IEEE COMPSAC*, July 4-8, 2017, Turin, Italy.

[25] Barry Boehm, Bradford Clark, Ellis Horowitz, Chris Westland, Ray Madachy, and Richard Selby. 1995. Cost models for future software life cycle processes: COCOMO 2.0. *Ann. Softw. Eng.,* vol. 1, pp. 57–94, Dec. 1995.

[26] Center for Software Engineering. 1999. COCOMO II and COQUALMO Data Collection Questionnaire. Retrieved June 3, 2020 from: https://www.yumpu.com/en/document/view/32692060/cocomo-ii-and-coqualmo-data-collection-questionnaire

[27] ISBSG. 2012. IFPUG/NESMA Questionnaire. Retrieved June 3, 2020 from: https://isbsg.org/wp-content/uploads/2015/10/DC_IFPUGNESMA_Form.zip

[28] National Information Center (Sudan). 2017. Retrieved June 3, 2020 from: http://nic.gov.sd

[29] Michael F. Bosu and Stephen G .MacDonell. 2013. A taxonomy of data quality challenges in empirical software engineering, in *22nd ASWEC 2013*, Melbourne, Australia. 97–106.

[30] Martin Shepperd, Qinbao Song, Zhongbin Sun, and Carolyn Mair. 2013. Data Quality: Some Comments on the NASA Software Defect Datasets. *IEEE Trans. Softw. Eng.,* vol. 39, 9, pp. 1208-1215, Sept. 2013.

[31] K.-A. Yoon and D.-H. Bae. 2010. A pattern-based outlier detection method identifying abnormal attributes in software project data. *Information and Software Technology,* vol. 52, 2, pp. 137–151, Feb. 2010.

[32] Gernot A. Liebchen and Martin J. Shepperd. 2005. Software productivity analysis of a large data set and issues of confidentiality and data quality, in *11th IEEE METRICS'05*, 19-22 September 2005, Como, Italy. 3-46.

[33] ISBSG. 2016. ISBSG D&E Repository Field Descriptions. Retrieved June 12, 2020 from: http://isbsg.org/wp-content/uploads/2016/11/ISBSG-Release-2016-R1.1-Field-Descriptions.pdf