1. Summary of SEERA: A Software Cost Estimation Dataset for Constrained Environments

1.1. Goal

The paper introduces the SEERA (Software enginEERing in SudAn) cost estimation dataset. 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 for the replicability of results and the general application of the dataset for international research.

The paper describes the dataset in detail, including the collection process, data preprocessing, and exploratory data analysis. The authors also provide some initial results from applying machine learning algorithms to the dataset for software cost estimation.

1.2. Dataset collection and analysis

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.

Figure 1 Questionnaire subsection

1.2.1 SEERA Dataset collection

The study designed a questionnaire to collect software project data in developing countries. It contained three sections: respondent information, project general information, and Factors affecting software product development. The questionnaire was prepared in the Arabic language and contained three sections: respondent information, project general information, and Factors affecting software product development. The questions were based on local environmental factors that impact software development cost in developing countries, while technical issues relating to software development were based on international attributes. The questionnaire was distributed and collected from June 2019 to

February 2020. Data entry, reviewing and analysis were conducted in parallel.

1.2.2. SEERA DATASET CHARACTERISTICS

1.2.2.1. Organization Characteristics

The SEERA dataset is a heterogeneous dataset from 42 organizations representing the public and private sectors in Sudan. The public sector contributed 37% of the projects, while private software companies contributed 52% and 82% respectively. The dataset includes attributes for the type of organization, sector and organization id. The dataset reflects a software development industry dominated by small to medium size organizations with limited staff. Table 5 shows that 60% of projects were implemented by departments of less than 10 employees, with the majority of projects developed by private software companies having software development departments of less than 10 employees. Respondents were asked to specify their roles during the lifetime of the project.

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%

Figure 2 Organization Characteristic 1

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%

Figure 3 Organization Characteristic 2

1.2.2.2. Project Characteristics

The SEERA dataset contains software development projects with a majority of projects initiating less than ten years ago and 43% beginning after 2015. A hybrid or waterfall methodology was applied in less than 25% of the projects, while agile methodologies were applied in less than 25%. Table 10 details the development type and application domain of the projects, with a majority of new software development projects based on open source software. All projects implement a DBMS, with Oracle Developer being the preferred technology for 43%. Programming languages were 13 distinct languages, with Oracle Developer being the interface technology for 31% of the projects, Python being the second popular (20%) and Java being the third (18%). Eight projects reported implementing a combination of programming languages within one application.

1.2.3. Dataset Attributes

The SEERA dataset attributes are divided into eight categories: general information, size, users, developers, project and product, effort,

and environment. 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. The initial dataset contained 176 attributes, but some were grouped into categories and the remaining attributes were grouped to create new attributes. The SEERA dataset contains three types of attributes: international attributes whose names and questions are adapted from international datasets, and international attributes with localized questions. Localized attributes with localized questions (sub-attributes) were identified as important factors in local software development cost impact.

The initial dataset contained 176 attributes, which were grouped into categories and created new attributes. The SEERA dataset includes 76 attributes, 52 of which are derived from sub-attributes. Examples of attributes derived from sub-attributes are shown in Figure 1.

1.2.4. Project Analysis

The SEERA dataset highlights the impacts of constrained environments on software development projects. The average estimated project duration exceeded the estimate by 86%, and economic instability was reported for 93% of the projects. Software size had minimal impact on cost and schedule, and some level of software reuse was evident in 63% of projects. Five projects reported losses of 100% and above. This study found that 33% of projects did not report project price and incurred costs, 14% did not calculate actual incurred costs, and 50% adhered to a project schedule.

The majority of projects reported that the customer environment was adequate, with 25% part-time and national service/training team members. There is a tendency to lower running costs through temporary contracts and hiring of inexperienced developers. Further research is needed to determine the relationships between these attributes and their impact on other cost factors.

1.2.5. Threats to Validity

Selection bias and reliance on recollection are threats to validity, but these limitations are reasonable due to the size of the study, quality review, and lack of previous work.

1.3. SEERA DATASET EVALUATION

Bosu and Macdonell introduced the data quality taxonomy to evaluate and compare datasets for empirical software engineering research. The taxonomy groups quality issues into three classes: accuracy, relevance and provenance. Threats to validity include selection bias and reliance on recollection, but these limitations are reasonable and do not jeopardize the credibility of the dataset. Bosu and Macdonell [4] used the data quality taxonomy to evaluate 13 publicly available software cost estimation datasets from the PROMISE repository. The SEERA dataset was reviewed based on the quality characteristics in [30].

The SEERA dataset excluded all categorical and limited range attributes and tested five attributes. Outliers were found for estimated duration, actual duration, team size, object points, % project gain (loss) and % project gain (loss). No redundancy was found in any of the 13 datasets and only two had minor inconsistencies. Table 13 shows the average percentage of missing values per attribute and Table 14 shows the attributes with missing values. The overall average percentage of missing values for each attribute is 1%, with 87% of projects having none or one missing value.

Provenance is public access, but only six have provenance/trustworthiness metadata. Relevance is measured in terms of number of

records and attributes, and timeliness is measured in terms of timeliness. The SEERA dataset includes attributes to distinguish the origins and characteristics of the submitting organization, which is important for accurate cost estimation modelling.

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%

Figure 4 Missing Attributes in SERRA dataset

1.4. Conclusion

This paper presented the SEERA cost estimation dataset, a dataset for technically and economically constrained environments. It contains 76 attributes and is augmented with metadata and subattributes containing the raw data before coding. It fills the current gap in providing data more relevant to developing countries and software industries in constrained environments. The authors plan to further analyze the SEERA dataset to investigate the impact of environmental and socio-economic factors on technical cost factors, compare the prevalence and magnitude of these cost factors to those of relevant PROMISE datasets, and develop a cost estimation model using the SEERA dataset and compare with known software costing models.

2. Summary of An Analysis of the Inclusion of Environmental Cost Factors in Software Cost Estimation Datasets

2.1 Goal

The importance of software affordability in early phase software design and its direct impact on software development and maintenance costs. Software cost estimation methods are employed to quantify these costs and provide stakeholders with estimates to aid in decision making. However, inaccurate cost estimation can impact software quality and development processes. The paper highlights the need for software cost estimation methods to evolve to accommodate the currently varied software development landscape and country-specific environmental and cultural aspects that may directly affect the software development effort. The paper evaluates current cost estimation datasets and identifies their suitability for application within countries with dissimilar cultural, societal, and technical constraints. The paper analyzes the characteristics of 31 software cost estimation datasets between 1981 and 2017 and compares their attributes to demonstrate the need for explicit representation of environmental factors within future datasets.

2.2. Dataset

The study used 31 datasets for algorithmic cost estimation models, which were a combination

of datasets from the Open Science tera-PROMISE repository, datasets surveyed by Mair et al., and datasets in the study by DE jaeger et al. Private datasets with publicly available attributes were excluded. The date and sector of the datasets were determined using various criteria, and datasets with comprehensive attributes were chosen. The characteristics of all the datasets used are detailed in Table I in chronological order.

2.2.1. Dataset Characteristics

The majority of datasets used for software cost estimation were collected during the 1980s and 1990s, and only a few recent datasets are available. 48% of the datasets are from Europe, 31% from North America and 21% from Asia and Australia. No datasets represent projects from Africa or South America, and no low-to-middle income countries from Asia are represented. The datasets come from various sectors, but there is a lack of details on the comprehensiveness of the dataset application areas. The number of attributes within datasets is varied, with an overall average of 13 attributes.

However, the diversity of projects and attributes makes comparative costing studies difficult, and most studies report inconclusive results. Most datasets represent outdated project development environments and exclude environments with dissimilar cultural and societal practices, which is a limitation given the significance of the software industries in developing and emerging economies.

2.2.2. Cost Attributes

The text describes a study on the attributes of 30 datasets related to project cost estimation in software engineering. The authors categorized the attributes into six categories: general information, users, developers, size, project, and product. They combined attributes with similar meanings and reduced the total number of attributes to 48. The text reports the distribution of attributes in each category for all datasets and compares the mean coverage of each category between datasets collected before and after the year 2000.

The study found that product size was an important attribute included in all datasets, but only a few datasets contained attributes from the users category, which limits their applicability in low-to-middle income countries. The developers category is an important cost factor in technically-constrained environments, but the team continuity attribute is not considered in most datasets. The KSLOC and function points attributes were the most common attributes within the size category, and the programming language and tool availability attributes were the most common attributes in the project category. The product complexity and reliability attributes were the most common attributes in the product category.

2.2.3. Importance of Environmental Factor

Accounting for country-specific factors and environmental differences is essential for accurate software cost estimation. When using generic datasets or attributes at an international level, it is important to consider environmental factors and their impact on cost estimation. A method is needed to determine their impact on other cost drivers.

2.3. Solution

The paper discuss the importance of accurate software cost estimation in the context of internationalization, outsourcing, and a growing international software market. It notes that while most studies use cost estimation datasets to build and compare different software cost estimation methods, there is limited research on the impact of cost drivers on the accuracy of cost estimation methods. It also highlights the

need to consider environmental factors within cost drivers and datasets, including factors that differ from country to country, such as cultural and organizational norms. It provides examples of studies that have examined the impact of environmental factors on software cost estimation, including those related to global software development. Further research is needed to identify and impact the identification and impact of these factors.

2.4. Conclusion

The paper analyzed publicly available software cost estimation datasets to determine their inclusion of environmental factors that impact software development costs. It found that most datasets represented US and European organizational environments, and the datasets from 2000 and later included more diverse attributes and increased coverage of organizational environmental factors. However, none of the datasets considered country-specific environmental cost factors. Further research is needed to identify and quantify the impact of environmental factors on software cost estimation, including case studies within different environments and countries.