**The SEERA Dataset: ReadMe File**

| | | |
|---|---|---|
| Name of dataset | SEERA | |
| Version | 1 | |
| Total number of attributes/variables | 76 | |
| Total number of records | 120 | |
| Date of collection | From June 2019 to February 2020 | |
| Donors of dataset | Emtinan I.Mustafa<br>Rasha Osman | |
| Data collection method | Questionnaire | |
| Purpose of data collection | • To provide up to date dataset with traditional cost attributes in addition to socio-economic and organizational attributes.<br>• The dataset projects represent constrained technical and economic software development environments.<br>• 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.<br>• 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. | |
| Country of contribution | Sudan | |
| Organizations from which data was collected | Types of organizations:<br>• Public and private software development companies<br>• Federal ministries<br>• Federal directorates<br>• Public universities<br>• Freelancers<br>• Corporate IT departments<br>• Telecommunication companies | |
| Files | SEERA cost estimation dataset.xlsx | The main 76 attributes and their scores. |
| | SEERA dataset original raw data.xlsx | Includes the complete set of attributes, sub-attributes, their original values and scores/ratings. Each category of attributes is in separate Excel sheet within the file. |
| | SEERA dataset attribute formulas.pdf | Includes the formulas used to derive some of the main attributes. |
| | SEERA dataset (sub)attribute codes.pdf | Details the codes corresponding to the options of the categorical/options attributes in the SEERA dataset. |
| | Project Programming Languages.pdf | Distribution of the programming languages implemented within the dataset projects. |
| | PROMISE2020 Paper.pdf | The paper discussing the dataset. |

| | |
|---|---|
| Dataset Format | • The dataset is in Excel format.<br>• Column headings represent attributes/sub-attributes.<br>• Rows represent projects.<br>• Missing values are represented by “?”<br>• The names of sub-attributes not included in the formulas of their main attributes are preceded with a (-).<br>• For values that depend on the answers of conditional questions: if the condition is not met the value is encoded as “N/A” and is not considered a missing value. |
| Information included within the dataset | • ID of organizations that contributed data<br>• Identifiers of projects in each group<br>• Customers type<br>• Start/end date of projects |
| Dataset information included in research paper | • Attributes with outliers<br>• Method/tool used to identify outliers<br>• Number and proportion of incomplete data<br>• List of problems encountered during data collection<br>• Redundant data and reason for redundancy |