Data sets and Variables for the Systematic Literature Review

Mike A. Marin

February 14, 2017

1 Introduction

This document describes the different data-sets and variables used during the Process Models Complexity Metrics – Systematic Literature Review (SLR). Each data-set corresponds to a tab in one of the two Microsoft Spreadsheets used ('report(full).xls' or 'Analysis.xlsx'). Those tabs have been saved as csv files with data-set name.

1.1 Report data-set (in.slr.raw.report.csv)

Data frame report.df: (from 'paper' tab in 'report(full).xls'). Each row represents a paper in the first phase of the review.

Id id assigned by the StArttool.

Search A number indicating the search that produced the document, as follows:

- (0) ACM
- (1) Web of Science
- (2) Scopus
- (3) IEEE
- (4) IEEE
- (5) Science Direct
- (6) Springer
- (7) Google Scholar
- (-1) Manual (backward snowballing)

Status.Selection Status of the selection phase: ACCEPTED, DUPLICATED, or REJECTED.

Status.Extraction Status of the extraction phase: UNCLASSIFIED, ACCEPTED, DUPLICATED, or REJECTED.

Year Year the paper was published.

I.one describe at least one complexity metric.

I.model related to workflow or BPM modeling.

I.survey survey of BPM or workflow complexity metrics.

E.no.complex does not define any complexity metrics.

E.no.modeling it is not a BPM or workflow modeling paper.

E.no.paper unable to access the paper.

E.no.new no new metrics defined (in most cases the paper include the validation of previously defined metrics).

E.no.english paper is not in English.

E.no.process.metrics paper may contain metrics but these metrics are not at the process level (for example metrics for activities).

E.runtime these are either runtime or repository based metrics (which are no considered in this review).

bibtex bibtex key of the reference.

1.2 Papers data-set (in.slr.raw.papers.csv)

Data frame papers.df: (from the 'papers' tab in 'Analysis.xlsx'). Each row represents a paper in the last phase of the review.

id id assigned by the StArttool.

Author bibtex key of the reference.

Year year the paper was published.

StArt.type corresponds to report.df$Status.Extraction.

Type classification of the paper as: Primary, Secundary, Duplicated, Rejected, Survey, or Uses.

New.metrics number of new metrics.

New.no.PLC.metric number of complexity metrics that are not at the process level (plc), and so they are not considered in this review.

Reused.metrics number of metrics this paper reuses from other previusly published papers.

Undefined.metrics number of metrics mentioned in the paper, but are not formally defined.

Notation the main notation used in the paper.

Primary.source is this paper the primary source for some of the metrics?

Complexity.metrics ere the metrics in the paper considered complexity metrics by the author?

Classification Does the paper contains a classification of process metrics?

Sum.Subjects maximun number of subjects used for validation mentioned in the paper (this is the addition of subjects in all experiments described in the paper).

Sum.Models maximun number of models used for validation.

Theoretical.val does the paper includes a theoretical validation?

Empirical.Validation does the paper includes an empirical validation?

Implemented were the metrics implemented in a tool?

Non.validation.experiment does the paper describes an experiment, but it is not designed to validate the metrics (the author is interested in other aspects and not in validation).

Survey.of.metrics does the paper includes a survey of process model metrics?

Comments commenst about the paper.

Concepts main concepts explored in the paper.

Other.concepts secondary concepts explored in the paper.

Reused.from bibtex key of the paper that originally defined the metrics used in this paper. There may be multiple primary papers being mentioned.

1.3 Metrics data-set (in.slr.raw.metrics.csv)

Data frame metrics.df: (from the 'metrics' tab in 'Analysis.xlsx'). Each row describes a single metric as it was described in the 'primary' paper.

id id assigned by the StArttool.

Author bibtex key of the reference containing the metric.

Year Year the paper was published.

Metric the symbol of the metric. This is the master value (if the symbol needs to be change during the review, I just need to change it here)

Name the name of the metric. This is the master value.

Label is the primary key for this data-set (and it must be unique). Label is also the key that should be used in other data sets to recover the metric and its name.

Notation used to define or validate the metric.

Category the type of metric:

- Algorithm
- Average
- Percentage
- Calculated
- Counter
- Ratio
- Weighted

Applicable.to.CMMN Yes or No.

Reason This variable describe the reason for Applicable.to.CMMN.

Cluster.reason if Applicable.to.CMMN is yes, then this variable contains a cluster of similar reasons.

1.4 Duplicated metrics data-set (in.slr.raw.dup-metrics.csv)

Data frame dup.metrics.df: (from the 'dup-metrics' in 'Analysis.xlsx'). Each row represents a duplicate metric.

id.orig id of the original paper (assigned by the StArttool). The original paper is the one that first defined the metric.

Original.author bibtex key of the original paper.

Original.Label label of the original metric.

id.dup id of the paper defining the duplicate metric (assigned by the StArttool).

Dup.Author bibtex key of the paper defining the duplicate metric.

Dup.Label label of the duplicate metric.

1.5 Validation data-set (in.slr.raw.validation.csv)

Data frame validated.df: (from 'validation' tab in 'Analysis.xlsx'). Each row represents an experiment. A paper may have more than one experiment, and each experiment will have a row in this data-set.

id id assigned by the StArttool

Author the bibtex label of the paper containing the experiment

Experiment.number Some authors have done series of experiments, in which case this field indicates the experiment in the series (as numbered by the author). In other cases a paper contained more than one experiment, in which case they were numbered by the order they were presented by the author. We use the word *experiment* in here very loosely to mean any type of validation

Type The following types were identified:

- **M** Metric correlation
- **E** Anecdotal
- **I** Intuition
- **CS** Comparing against measuring stick
- **CM** Comparing against other metrics
- **Hw** Within-subjects
- **H** human validation
- **HS** Online-survey
- **S** Software [Error prediction]
- **T** Theoretical validation
- **N** No validated

Design if the study presents measured variables, measurement criteria, treatments, the number of subjects, and sampling.

- **1** Good description of the study including most of the information required to understand what was done
- **0.5** A deficient description of the study used to validate the metric
- **0** No description of the study

Hypothesis for papers using statistics:

- **1** formal null hypothesis is presented, or clearly stated research questions
- **0.5** informal hypothesis.
- **0** no hypothesis or clearly stated research questions

Threats Discuss all threats to validity (internal, external, conclusion and construct). Threats to validity:

- **4** all four threats to validity are present and clearly identified (internal, external, conclusion and construct)
- **3** three threats to validity are discussed
- **2** two threats to validity is discussed
- **1** one threat to validity is discussed, or validity of the results are discussed
- **0** No discussion of threats to validity or validity

Validity is calculated based on the Threats variable, as follows:

- **1** All four threats to validity are discussed
- **0.5** Two to three threats to validity are discussed
- **0** Less than two threats are discussed

Research Based on research or lessons learned or based on expert opinion

- **R** research
- **L** Lesson learned
- **E** Expert opinion

Power sample size justified or power analysis conducted:

- **1** Yes
- **0** No

Normality.test Are the assumptions of the statistical procedure taken into account, in particular normality tests?

- **1** Yes
- **0** No

iv.id are the independent and dependent variables clearly identified?

- **1** clearly identified independent and dependent variables
- **0.5** mention only one type of variable (either dependent or independent variables)
- **0** no mention of independent or dependent variables

Subjects for experiments including humans, this indicates the number of subjects as indicated by the author.

Models the number of models involved in the experiment

Subject.num is a calculated variable based on the variables subjects and models. For human experiments, where subjects are use the Subject.num will have the same value as the Subjects variable. For software experiments, where only models are used, the Subject.num will have the number of models.

Scale is a calculated variable based on Subject.num, as follows:

1 Subject.num greater or equal than 20

0.5 Subject.num greater or equal to than 10 (and less than 20)

0 Subject.num less than 10

Clear.Findings Clear statement of findings

1 clearly stated findings (clear list of the metrics that were validated or not).

0 no clearly stated findings include statements like "we believe", or there is contradiction between the results and the conclusion.

Control.group uses groups to either compare results (control versus experimental groups), or account for fatigue.

Y Uses control groups

G multiple groups, but no control group

U seems to use a control group, but not stated by the authors (implicit)

O the subjects received the material in different order.

N No control group (or not stated)

Control is a calculated variable based on Control.group, as follows:

1 Control.group = `Y`

0.5 Control.group = `G` or `U` or `O`

0 Control.group = `N`

Instruments as follows:

1 Sample of some of the instruments are included (normally in an appendix, or a URL to supplemental material)

0 No sample instruments are included.

Shows.statistics could be:

S show p-values and discusses significance of the results.

p show p-values

y show statistics being used but no p-values. Or did not accounted for multiple comparisons.

N No statistical values being shown (although claim the use of stats)

X No statistical values (author did not used any statistics in the paper)

Stats is a calculated variable based on the values of Shows.statistics, as follows:

1 Shows.statistics = `S`

0.5 Shows.statistics = `p`

0.2 Shows.statistics = `y`

0 Shows.statistics = `N` or `X`

Sampling can be:

R described by author as random

Rx random (not stated)

C described by author as by convenience

Cx by convenience (not stated)

N No mention and not easy to guess

X Author made up the sample (normally for models, where the author created the models)

Sample is a calculated variable based on Sampling as follows:

1 Sampling = `R` or `Rx`

0.5 Sampling = `C` or `Cx`

0 Sampling = `N` or `X`

Subject.type The values are:

S Students

E Experts

P Professionals

U Professors and students

A Academics, practitioners, and students

Sm Student models (models created by students)

Mx Models, but no explanation of the source of these models

Mr Models from real industrial settings.

Ma Artificial models (created for the paper or by a different author).

Subject.val is a calculated variable based on Subject.type, as follows:

1 Subject.type = `P` or `Mr`

0.5 Subject.type = `E` or `A`

0 Subject.type = `S` or `U` or `Sm` or `Mx` or `Ma`

Experiment.dsg The author describes the experimental design used.

- **E** Describes a well know experimental design
- **W** Within-Subjects (described by author)
- **Wx** Within-Subjects, but not stated in the paper
- **N** No mention the type of the experiment being used.

Experiment is a calculated variable based on the value of experiment.dsg, as follows:

- **1** experiment.dsg = `E` or `W`
- **0.5** experiment.dsg = `Wx`
- **0** experiment.dsg = `N`

Secondary.data one of the following:

- **Y** Data from a previous experiment is being used.
- **N** Data was created specifically for this experiment.
- **U** It is unclear is the data is being reused from a previous experiment.

Concept the main concept being validated.

Statistics the statistical analysis used for the validation.

Comments basic comments.

iv the independent variables (if any) used for the validation.

dv the dependent variables (if any) used for the validation.

1.6 Theoretical validation data-set (in.slr.raw.theor.vali.csv)

Data frame T.validation.df: (from the 'theor.vali' tab in 'Analysis.xlsx'). Each row describes a theoretical validation. If Theoretical.validation == Weyuker, then properties one to nine corresponds to the nine Weyuker properties. If Theoretical.validation == Briand, then property one to five, corresponds to the five Brian property (which in turn are based on the type of metric – size, length, etc.)

id id assigned by the StArttool

Author doing the theoretical validation

Metric being validated. Note that this is not the master metric value (this is just to keep the spreadsheet readable), but in R if the Metric column is required, then a merge with metrics.df should be done using Label (because the master Metric variable is in that data set).

Theoretical.validation

id.orig id assigned by the StArttool

Original.author this is the author that defined the metric

property.1 Weyuker or Briand's property one (pass Y' or failed 'N')

property.2 Weyuker or Briand's property two (pass Y' or failed 'N')

property.3 Weyuker or Briand's property three (pass Y' or failed 'N')

property.4 Weyuker or Briand's property four (pass Y' or failed 'N')

property.5 Weyuker or Briand's property five (pass Y' or failed 'N')

property.6 Weyuker's property six (pass Y' or failed 'N')

property.7 Weyuker's property seven (pass Y' or failed 'N')

property.8 Weyuker's property eight (pass Y' or failed 'N')

property.9 Weyuker's property nine (pass Y' or failed 'N')

1.7 Empirical validation data-set (in.slr.raw.validated-metrics.csv)

Data frame E.validation.df: (from the 'validated-metrics' tab in 'Analysis.xlsx'). Each row corresponds to an experiment to validate a single metric, and validated.by.[is—Author] indicates the author that conducted the experiment. Experiment in this data-set means any type of claimed empirical validation.

id.orig id of the original paper that defined the metric (assigned by the StArttool).

Original.Author bibtex key of the original paper that defined the metric.

Year Year the paper was published.

Delete.metric Indicates if the metric is:

- **D** is a deleted metric (it does not meet some of the inclusion criteria).
- **Dx** is a deleted metric, but should keep the definition of the metric because it is used as apart of the calculation for another metric.
- **C** have changed the name of the metric.
- **Dup** the metric is duplicated.

Metric.obsolete being validated. Note that this is not the master metric value (this is just to keep the spreadsheet readable), but in R if the Metric column is required, then a merge with metrics.df should be done using Label (because the master Metric variable is in that data set).

Name.obsolete of the metric being validated. Note that this is not the master metric name (this is just to keep the spreadsheet readable), but in R if the name column is required, then a merge with metrics.df should be done using Label (because the master name variable is in that data set).

Label label of the metric being validated.

Notation Notation used for the validation of the metric.

Category The type of validation being performed.

- **M** Metric correlation
- **E** Anecdotal
- **I** Intuition
- **CS** Comparing against measuring stick
- **CM** Comparing against other metrics
- **H** human validation
- **HS** Online-survey
- **Hw** Within-subjects experiment using humans
- **S** Software [Error prediction]
- **T** Theoretical validation
- **N** No validated

Validated.by.id id of the paper that validated the metric (assigned by the StArttool).

Validated.by.Author bibtex key of the paper that validated the metric.

Experiment an experiment id. Authors that conducted a set of experiments, normally will indicate the first (1) experiment, or the second (2) experiment. In some cases an experiment may have sub-experiments, for example experiment five will have two versions (51 and 52).

Subjects the number of human subjects used for the validation.

Models the number of process models used in the validation.

Concept the concept being tested in the validation.

Statistics.test.used the statistical test used for the validation.

Result one of the following values:

- **validated** The metric is considered empirically validated by the author (based on statistical tests).
- **No** The metric failed to be validated (based on statistical test or the author claim)

partial The metric seems to provide some explanatory power, but it is not needed to explain the errors (used by Mendling)

intuitive The author categorizes the result of the validation as the metric is: 'reasonable', 'the best metric', 'effective and rational', 'intuitive', or 'institutional'. This value also indicates a metric that is claimed to be validated using an anecdotal (Category = 'E'), or intuition (Category = 'I') validation.

Comments short comment when needed.