Back to Basics - The 4R's of Software Estimation

Barbara Kitchenham
Keele University
Aim

• To discuss the need for
  – Rigour, Reproducibility, Replication and Relevance
  – In the context of current software estimation research
• To identify limitations with current practice
• To suggest means of addressing those limitations
Definitions

• Rigour
  – Are scientific methods applied correctly?

• Reproducibility
  – Can an independent researcher verify the results published in a study?

• Replication
  – Are the results consistent across different data sets?

• Relevance
  – Do the study results address practitioner problems?
Rigour

• Many poor quality studies still published
• Researchers
  – Do not justify their choice of data set(s)
  – Don’t apply the same rigour to all methods
    • Ordinary regression without logarithmic transformation
  – Use invalid metrics
    • Cost estimation
      – All the relative error family (MRE, Balanced MRE etc)
    • Fault prediction
      – F-1 and AUC
Reproducibility

• Not considered important in SE papers
  – Reports of methodology insufficient
    • Machine learning papers seldom explicitly report their fitness function
      – Sometimes use different fitness function in wrappers
    • Use data sets that aren’t publically available
    • Build and verification subsets not specified
    • Prediction rather than goodness of fit not confirmed

• Cost Estimation
  – Whigham et al. (2015)
    • Unable to reproduce results of two studies

• Fault Prediction
  – Shepperd et al. (2014)
    • Analysed 42 papers
    • Different people using the same method on the same data set get different results
    • “It matters more who does the work than what is done.”
Replication

• The R most considered in SE research
  – Addressed by applying methods to
    • Multiple data sets
    • BUT alas, not always public data sets

• Even public data sets have problems
  – Different versions of the data set
  – Overlapping data sets
    • May be treated as independent but are not
  – Errors in the data sets
    • NASA fault prediction data sets
  – Assuming data set & dataset subsets provide independent evidence
    • Using COCOMO1 plus the 3 mode-based subsets does not mean you have 126 projects
Relevance

• Least considered R
• Typical SE estimation study justified because
  – “Poor quality cost estimation/residual defects cost the IT industry X billions of dollars per year”
• Few papers consider practical issues:
  – Most software development is evolution
    • Size of maintenance work hard to measure
    • Components differ wrt age & fault history
    • Difficult to find comparable items for model building
  – Practitioners want to know
    • How much to bid
    • If a project plan is realistic
    • If a product is in a suitable state to release
    • Our research doesn’t usually answer those questions
Relationships between the Rs

• Without Rigour
  – Reproducibility is pointless

• Without Reproducibility
  – Replication is valueless

• With Rigour, Reproducibility & Replication
  – We get good science

• Without Relevance
  – Don’t get good engineering science
  – We can’t influence practice
Is there really a problem?

- 2016 Statistics based on SCOPUS search
  - 36 cost/duration estimation comparative papers
    - 18 journal papers, 18 not journal papers
  - Evaluation criteria
    - MMRE
      - 25 papers, 12 journal papers
    - MAR (or MdMAR or SumMAR)
      - 16 papers, 10 journal papers
    - MMRE & MAR 6 papers
  - Data sets
    - More than 1
      - 16 papers (9 journal papers)
    - No data set publically available
      - 7 papers (4 used ISBSG only)
  - Identifiable problems
    - 8 papers (3 journal papers)
      - Predictions too good to be true, 5 papers
      - Used overlapping data sets as if independent, 2 papers
      - Reported negative absolute values
      - Procedia Computer Science, 3 papers
        » Elsevier electronic publishing of conference proceedings
Improving Rigour

• Improve the standard of reporting
  – Needs the support of the journals and conferences
    • Current reporting standards assume things are basically correct
      – Need to be better if rigour is to be confirmed
        » Need to confirm prediction is taking place
    • Ensure novel/rare techniques reviewed by a statistician/methodology expert
      – Otherwise poor use of methodology not detected
        » E.g. incorrect analysis of cross-over designs
  • Reject papers we review if we cannot be sure of study rigour
  • Do better ourselves
Improving Reproducibility

• Use open source languages
  – R for statistical analysis & simulation studies
  – Weka or OpenML for machine learning
  – Publish the algorithms rather than just pseudo code

• Make sure selection of build and verification subsets fully defined

• Need support from journals
  – ACM Transactions on Mathematical Software
    • Replicated Computational Results Initiative
    • Publish studies that have reproduced results
Improving Replication

• Justify the selection/omission of data sets
  – Define inclusion/exclusion criteria

• Reject papers that use data that isn’t public
  – Unless new data set important to demonstrate relevance and
    • Method confirmed on public data sets
    • Data & analysis process available for checking by other reviewer

The 4 R’s
Improving first 3 Rs

- **Benchmarking**
  - BUT, just making data available is not sufficient

- **Need to**
  - Agree a set of useful data sets
  - Confirm agreed versions of data for each data set
  - Have agreed build and verification subsets
  - Have reproducible results of applying standard methods to those data sets
    - Regression
    - Analogy
    - Genetic Algorithms
    - Etc.
  - Use unbiased accuracy statistics
  - Ensure prediction is taking place
    - E.g. Regression prediction must outperform mean
  - Reject papers advocating any new method that is not as good or better than standard methods on all of the data sets
  - Query papers with results that look too good
    - Probably goodness of fit NOT prediction

- **Psychology have just completed a major replication project**
  - Software Estimation needs one too
Improving Relevance

- Explaining how the technique fits with actual development practice, BUT, in industry
  - Components are usually all in different states
    - Consider data as a time series
  - Defect prediction
    - What group of i.i.d items are we going to build a model on?
      - Statistical models and machine learning assume that the past patterns reflect the future
    - What items are we going to apply the model to?
  - Cost estimation
    - Models still use data values only available and/or collected at the end of development to build models
      - Size (FP or Loc)
        » Need early phase estimates of size to build prediction model
      - Duration
        » Need early phase values & whether estimate or constraint
    - Ignore quality requirements

- Work with industry partners
  - Obtain more realistic datasets
  - BUT, don’t settle for commercially confidential data
Conclusions

• Software Estimation research
  – Concentrates on ever more complex algorithms
  – Based on aging and suspicious data sets
    • Delivering minor improvements
    • Irrelevant to industry

• We need to get back to basics
  – If we are genuinely an engineering science
    • Must embrace the reproducible science movement
  – Start doing reproducibility studies
    • Must agree basic standards
    • Good first step for post-grads
  – Develop trustworthy benchmarks
  – But must not forget Relevance

The 4 R’s