How repeatable are systematic reviews?

Steve MacDonell

Based on work done by myself and:

Barbara Kitchenham
Emilia Mendes
Martin Shepperd
Guilherme Travassos

}

Disclaimer: none of these individuals have seen this work and so are blameless!

Motivation

- We are still (relatively) new to Systematic Reviews (SR) in SE
- We are rightly borrowing, adapting, learning as we go
- As part of this learning process we need to consider SR credibility, just as we assess the credibility of our empirical modeling
- In particular, how reliable are SRs?

mellow
mellow
mellow

Approach to the study (1 of 3)

- Two teams of researchers, addressing the 'same' research question:
  - Team KMT – Kitchenham, Mendes, Travassos
  - Team MS – MacDonell, Shepperd
  - KMT RQ: What evidence is there that cross-company estimation models are not significantly worse than within-company estimation models for predicting effort for software/Web projects?
  - MS RQ: What evidence is there that cross-company estimation models are at least as good as within-company estimation models for predicting effort for software projects?

mellow
mellow
mellow

Approach to the study (2 of 3)

- Teams negotiated and agreed on meta-level issues, specified in a 'meta-protocol'
- Among other things, this specified the basis for comparison:
  - Sources searched – where did the teams look for studies?
  - Search strategy – how was the search executed (automatically, manually)?
  - Terms used in searching – what fields or similar were considered in the search, over what period of time?

Approach to the study (3 of 3)

- Basis for comparison (ctd):
  - Papers found – retrieved or located as being potentially relevant
  - Papers discarded – those studies not selected, and why
  - Papers included – those selected as primary studies
  - Analysis approach – issues considered and steps followed
  - Analysis outcomes – interpretations and conclusions drawn
  - *Effort expended in various activities – determining the protocol, data extraction, data aggregation, write-up of outcomes*

Same or different? (1 of 10)

- Sources searched:
  - SR1 – INSPEC, EI Compendex, Science Direct, Web of Science, IEEE Xplore, ACM DL; plus 7 specific journals and conference proceedings
  - SR2 – The databases listed for SR1 plus Blackwell/Synergy, EBSCOhost, Expanded Academic ASAP, ProQuest, Scholar.Google, Springer, Wiley Interscience, WoK Proceedings; no specific journals or conference proceedings
- Search time-span:
  - SR1 1999-2005; SR2 1995-2005

mellow
mellow
mellow

Same or different? (2 of 10)

- Search strategy:
  - Both SRs used a combination of automated database searching and manual citation analysis, but SR1 utilised multiple searchers whereas SR2 used a searcher/checker approach. SR1 also additionally asked for feedback from the authors of the primary studies regarding possible ongoing work.
- Terms used:
  - Same basic approach through aggregation from keywords and search piloting, but…
  - SR1 used a more comprehensive single concatenated string
  - SR2 used a combination of three strings (using wildcards where possible), dealing with the function, the object and the context

Same or different? (3 of 10)

- Terms used (ctd):
  - SR1 - (software OR application OR product OR Web OR WWW OR Internet OR World-Wide Web OR project OR development) AND (method OR process OR system OR technique OR methodology OR procedure) AND (cross company OR cross organisation OR cross organization OR cross organizational OR cross organisational OR crosscompany OR cross-organisation OR cross-organization OR cross-organizational OR cross-organisational OR multi company OR multi organisation OR multi organization OR multi organizational OR multi organisational OR multicompany OR multi-organisation OR multi-organization OR multi-organizational OR multi-organisational OR multiple company OR multiple organisation OR multiple organization OR multiple organizational OR multiple organisational OR multiple-company OR multiple-organisation OR multiple-organization OR multiple-organizational OR multipleorganisational OR within company OR within organisation OR within organization OR within organizational OR within organisational OR within-company OR within-organisation OR within-organization OR within-organizational OR within-organisational OR single company OR single organisation OR single organization OR single organizational OR single organisational OR single-company OR single-organisation OR single-organization OR single-organizational OR single-organisational OR company-specific) AND (model OR modeling OR modelling) AND (effort OR cost OR resource) AND (estimation OR prediction OR assessment)

- SR2 - (("cost model" OR "cost estimate" OR costimation OR "cost prediction" OR "effort prediction" OR "estimating cost" OR "estimating effort") AND ("software project" OR "software product" OR "software development" OR "web application" OR "web project" OR "web development")) AND ("company specific" OR "company external" OR "cross company" OR "individual company" OR "multi company" OR "multi organization" OR "multi organisation" OR "within company")

mellow
mellow
mellow

Same or different? (4 of 10)

- Papers found, discarded, included:
  - Ten/several papers 'known' beforehand
  - SR1 retrieved 772 papers, 24 matched the nine known, IEEE Xplore found 5
  - SR2 retrieved 185 papers, 38 matched the ten later identified, Compendex & Inspec found 9
  - SR1 discarded studies with low numbers of organisations in the cross-company data set or comparisons of single-organisation models to general cost-estimation models
  - SR2 discarded studies due to topic, treatment then credibility issues
  - Ten primary studies found per SR

mellow
mellow
mellow

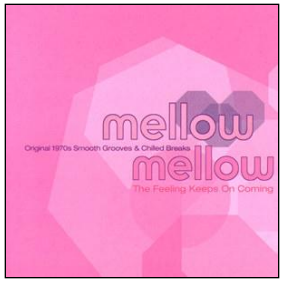
Same or different? (5 of 10)

- Primary studies:
  - Nine of the eleven identified primary studies were common to both SRs
  - One different primary study was identified in each SR:
    - SR1 included a paper unpublished but in press for 2005 publication, authored by Mendes, Lokan, Harrison and Triggs
    - SR2 included a 2003 paper published in the Proceedings of the International Symposium on Software Metrics, authored by Mendes, Mosley and Counsell

- mellow
- mellow
- mellow

Same or different? (6 of 10)

- Analysis approach:
  - The same questions were addressed…
    - Is the analysis process description complete? Is it clear how accuracy was measured? Is it clear what cross-validation method was used? etc
  - And the same high-level issues considered…
    - Data quality, data set size, number of projects per model, metrics used, etc



Same or different? (7 of 10)

- Analysis approach (ctd):
  - But the specific method used was quite different…
    - SR1 studies were randomly allocated to extractor, checker and adjudicator roles
    - SR2 used one extractor and one checker
    - SR1 focused strongly on aspects of data analysis validity
    - SR2 focused strongly on aspects of data sampling/splitting, data quality and diversity/representativeness

Same or different? (8 of 10)

- Analysis approach (ctd):
  - SR1 calculated and assigned a quality score to each primary study, based on its degree of adherence (or otherwise) to the quality criteria, followed by a qualitative assessment of the evidence
  - SR2 used a more qualitative approach to the quality analysis, assigning labels such as 'Low' to indicators, and then considered the proportion of total statistical comparisons favouring one modeling method over another

Same or different? (9 of 10)

- Analysis outcomes:
  - Over the nine primary studies common to the two SRs there was total agreement on the interpretation of results:
    - The same three studies were interpreted as favouring cross-company models
    - The same four studies were interpreted as favouring within-company models
    - The same two studies were interpreted as being inconclusive due to the absence of significance testing

mellow
mellow
mellow

Same or different? (10 of 10)

- Analysis outcomes (ctd):
  - Both SRs identified limitations in the primary studies, particularly in relation to data quality, model construction and experimental design
  - Both SRs identified primary study author reservations about the outcomes of their work
  - Some studies acknowledged that they were effectively pointing out which approach was "less bad" than the other
  - Both SRs noted the lack of strong evidence in the primary studies - individually and collectively - and the impossibility of meta-analysis

mellow
mellow
mellow

Summary

- The primary studies selected for analysis were (close to) the same in both SRs and…
- The conclusions reached were also the same, but…
- On almost every other dimension the SRs were different!

Acknowledgements

- The team members – Emilia, Martin, Barbara and Guilherme
- Martin as host
- The EPSRC and AUT University

mellow
mellow
mellow

Open questions

- Is the outcome more important than the means of achieving it?
- Are the differences just a function of the relative novelty of SRs in SE? Or weak protocols?
- Or is this to be expected – and even embraced – in these formative stages?