Notes on *Ecological inference in empirical software engineering* [1]

Akond Rahman Manish Singh Bennett Narron

August 24, 2015

Abstract

AbstractSoftware systems are decomposed hierarchically, for example, into modules, packages and files. This hierarchical decomposition has a profound influence on evolvability, main- tainability and work assignment. Hierarchical decomposition is thus clearly of central concern for empirical software engineering researchers; but it also poses a quandary. At what level do we study phenomena, such as quality, distribution, collaboration and productivity? At the level of files? packages? or modules? How does the level of study affect the truth, meaning, and relevance of the findings? In other fields it has been found that choosing the wrong level might lead to misleading or fallacious results. Choosing a proper level, for study, is thus vitally important for empirical software engineering research; but this issue hasnt thus far been explicitly investigated. We describe the related idea of ecological inference and ecological fallacy from sociology and epidemiology, and explore its relevance to empirical software engineering; we also present some case studies, using defect and process data from 18 open source projects to illustrate the risks of modeling at an aggregation level in the context of defect prediction, as well as in hypothesis testing.

1 Thesis

Summary of the thesis being investigated along with research questions, goals, and hypotheses. This paper focuses on the importance of ecological inference in software engineering and the risk of ecological fallacy as a

result of mistaken ecological inference. Given that large systems consist of complex software resulting in hierarchical organization of the systems, the teams and the processes to develop the systems, it becomes important to study the ecological inference done at different levels of aggregation and how good the findings hold at other aggregated/dis-aggregated levels. The author emphasizes the importance of selecting the right level of aggregation to conduct empirical studies to measure observable outcomes such as quality and productivity. The paper discusses the various risks that accompany the ecological inference done at the aggregated levels and which result in ecological fallacy when there is discrepancy between the findings at the aggregated and disaggregated levels. The paper also discusses the various factors - sample size, zonation and class imbalance , that contribute to the risk of ecological fallacy.

The goal of this paper is to have a conceptual framework of ecological inference risk in software engineering and empirically demonstrate the existence of this risk , by building prediction models at different aggregation levels and comparing inferences drawn from these models. While pursuing this goal,this paper tries to answer the following research questions:

1. What is the right level of study?
2. How does the level of study affect the truth, meaning and relevance of the findings?
3. Are prediction models subject to ecological inference risk?
4. Is hypothesis testing subject to ecological inference risk?
5. What are the effects of aggregation on model quality?
6. Do inferences drawn from models built at aggregated levels transfer to disaggregated levels used to build the aggregations?

2 Contributions

Summary of the authors' contributions, either implicit or explicit, to ASE

3 Investigation Methods

Summary of the authors' investigation methods or experimental design.

4 Results

Summary of the authors' results

4.1 "Power"

Notes on the impact of the authors' results

4.2 Applicability

Notes on the applicability of the authors' results

5 Technical Development

Summary of the authors' technical development

5.1 Examples

Details of any examples to clarify technical developments

References

[1] D. Posnett, V. Filkov, and P. Devanbu. Ecological inference in empirical software engineering. In *Proceedings of the 2011 26th IEEE/ACM International Conference on Automated Software Engineering*, pages 362–371. IEEE Computer Society, 2011.