Paper One Summary

Matthew Neal, Joseph Sankar, and Alexander Sobran

September 3, 2015

Reference

Daryl Posnett, Vladimir Filkov, and Premkumar Devanbu. 2011. Ecological Inference in Empirical Software Engineering. Proceedings of the 2011 26th IEEE/ACM International Conference on Automated Software Engineering.

Important Keywords

Ecological inference The idea that an observation or finding at an aggregated level can apply at a disaggregated level. For example, the idea that a model of defects at the package level also applies. at the file level

Ecological fallacy An incorrect application of ecological inference. A discrepency between the findings at an aggregated level and disaggragated level. For example, an instance where a model of defects at the product level is incorrectly used to model defects at the file level.

Scale Size of an aggregated unit. The larger the scale the bigger and fewer the aggregated units. This can effect affect the quality of statistical models built for defect prediction.

Zonation The manner in which aggregation is performed. For example, gerrymandering. This can create internal validity threats and lead to ecological fallacies.

Aggregate Phenomena Certain phenomena may apply only at an aggregated level. The example given in the paper is that of inheritance. Inheritance only takes place at the class level and cannot be studied at a disaggregated level.

Cost Effectiveness A concept used within the paper to assess defect models. The baseline is that funds are given to investigate 10% of a project for defects, in 10% of the lines of code taken at random 10% of the defects should be found. Models are compared against this baseline.

Feature Extraction

- **Motivational Statements** Modern software applications are composed of many files. Thus, the files are usually organized into units, which may themselves be organized into broader units. While research could be focused at the most granular level, for example files, methods, even lines of code, modeling at these levels is costly and takes time. The authors motivation is determine if statistical models built at an aggregated level are valid at the dis-aggregated level. A second motivation is to asses the cost-effectiveness of predication models at varying aggregated levels.

- **Related Work** To improve model performance, Koru *et al.* recommend aggregating. Their methods to acheive improved performance include summing method-level measures into class-level measures and combining these with metrics from class-level features. They conclude that this approach overcomes the issues of asymmetry in defect data sets. Schröter *et al.* conclude that predicting at coarser granularities is less difficult in comparison to finer granularities but lead to worse results. Zimmerman *et al.* find different results where package level correlations and model correlations are higher than file level correlations and model correlations. Ambrose *et al.* argue against the use of aggregate level metrics, though without data, stating predictions are more cumbersome to investigate,classes are inherently self-contained, and information derived from the package-level cannot be applied to the class level.

- **New Results**: The authors grouped features in the models based on loss of significance(LS) or gain of significance(GS). Both of these can result in Ecological Fallacy. LS is a loss of significance when inferring from the aggregate level to the disaggregate level. GS is a gain of significance from the aggregate level to the file level. They found 28 instances of GS and 25 instances of LS out of 108 features over 68 models. The authors draw the conclusion that models built at the aggregated level should be undertaken with the knowledge of the underlying risk of Ecological Inference and Ecological Fallacy. The author's also provided data showing the advantages of AUCCE over AUCROC when comparing models at different levels of aggregation. These results lead to two final conclusions: aggregated prediction models can evaluate better than actual when applied to aggregated data and as a consequence models build at an aggregated level may not be applicable to the dis-aggregated level.

- **Future Work**:The authors bring up three sources of Ecological Inference risk: Zonation, Sample Size, and Class Imbalance. The authors believe that further study of the three sources is in order. To give a short background on how each applies:

  - Zonation: Aggregation runs the risk of creating distortion by the way that the data is aggregated.

  - Sample Size: As aggregation takes place inherently there is move to smaller and smaller sample sizes. As the move to those smaller sample sizes takes place there is a loss in Statistical power.
  - Class Imbalance: A situation where a set of samples have a particular attribute and a large proportion of that set have a particular value with a very small number having another significant but underrepresented value. The example in the paper is a situation where there are a very high number of projects without defects and a very lower number with defects.

- **Sampling Procedures**: Projects were taken from a set of 18 Apache Software Foundation Projects over 87 versions. The writers took JIRA issue reports and then mapped them to pull commits. Then, they took those commits and connected them to packages and files. They also pulled the number of lines of code as well as the number of developers per project. They were unable to look at past data sets due to their inability to correctly aggregate the data given that many data sets do not contain issue identifiers.

Possible Improvements

- There are some particularly distracting typos in the paper. For example, on pages 367-368 the authors failed to correctly reference their table correctly leaving "Table ??" in the middle of their text twice.

- The authors could have provided a better explanation of AUCROC and AUCCE and the meaning of their results with respect to those measures. The authors did a good job of comparing file and package measures, but did not explain why file AUCCE was lower than file AUCROC. Is this just the nature of AUCCE since it is including cost effectiveness? Why was there more variance in the AUCCE measures than AUCROC?

- At the end of the paper, the authors provide a brief note about threats to the external validity of the results. The authors only examined projects from one source: the Apache Software Foundation. Analyzing projects from other sources which employ different software engineering principles or which organize their projects differently would bolster their results and conclusions.