The Forest or the Trees?
Tackling Simpson's Paradox in Big Data Using Trees
Galit Shmueli, Indian School of Business, India
Inbal Yahav, Bar Ilan University, Israel

Abstract
Prediction and variable selection are major uses of data mining algorithms but they are rarely the focus in causal IS research. Because experiments are often impossible, unethical or expensive to perform, causal IS research often relies on observational data. A major challenge is to infer causality from such data. Simpson’s paradox can arise in such contexts, causing uncertainty regarding the right level of data aggregation, where different aggregation levels indicate opposite effects. Detecting a paradox is non-trivial with many potential confounders. This paper uses the predictive tool of Classification and Regression Trees for detecting Simpson's paradox. We introduce a new tree approach for detecting potential paradoxes in data that have either a few or a large number of potential confounding variables. Our approach relies on the tree structure and the location of the cause vs. the confounders in the tree. It is applicable to categorical and numerical outcomes and is efficient with large samples. We discuss theoretical and computational aspects of the approach and illustrate it using several real applications.

Macro-Decisioning, Micro-Decisioning and Simpson’s Paradox
With the growing availability of data at more granular levels, decision making has expanded from aggregate-level to personalized decisions. In medicine, we see a shift towards personalized medicine. In marketing, personalized offers and customer experiences are now common. A perplexing issue in the context of choosing the level of aggregation is Simpson's paradox (Simpson, 1951). The paradox describes the phenomenon where the direction of a cause on an effect appears reversed when examining the aggregate vs. disaggregates of a sample or a population. The practical decision making question that Simpson's paradox raises is choosing the level of data aggregation that presents the results of interest. This raises the challenge of identifying potential confounders and then establishing a criterion for deciding which (if any) of the potential confounders should influence the decision making.

One might think that with sufficiently large samples, it is always safer to use the disaggregate data, which are potentially more homogeneous. While this might be true for micro decisioning, it is not necessarily the case for macro decisioning, where the goal is to evaluate an overall effect. Pearl (2009) shows that in many cases it is the aggregated rather than the disaggregate data that gives the correct choice of action.

Pearl describes Simpson's paradox as “the phenomenon whereby an event C increases the probability of E in a given population p at the same time, decreases the probability of E in every subpopulation of p”. Pearl warns that Simpson's paradox can only be resolved when the observational data (frequency tables) are combined with a causal theory. He shows that the same data table can result from different causal paths and therefore the underlying causal structure is unidentifiable from the data alone. In other words, once the effect, cause, and potential confounding variable are singled out, a causal narrative is required for determining which level of aggregation to use for decision making.

The focus of this work is identifying potential confounding variables in a high-dimensional dataset that cause the Simpson's paradox. We take a data-driven approach that searches the terrain of possible relationships between the outcome of interest and the set of causes and potential confounders. A tree-based approach is applied to micro-level data and automatically identifies existing relationships and their structure. The result graphically displays potential confounders and the structure of confounding, allowing the researcher or decision maker to identify potential Simpson's paradox relationships to be further investigated with a causal toolkit such as Pearl's "back-door" test.
Trees for Detecting Simpson’s Paradox

Our use of trees in this explanatory modeling context differs from predictive modeling in a few ways. Most importantly, we are interested in the tree structure itself: not only which predictors are present, but also which predictors are absent, and importantly, what is the ordering of the splits. Second, unlike predictive modeling, we do not use the tree to predict new records. Third, in some cases we use fully grown tree which overfit the sample, for the purpose of identifying the tree structure. Fourth, to account for sampling variance, and when applicable, we prefer conditional-inference trees (Hothorn et al., 2006) where variable choice and splitting values are based on statistical tests of independence over trees that rely on cross-validation or holdout data pruning. And lastly, we develop a new stopping criterion for tree growth for identifying Simpson’s paradox.

Simpson’s paradox is classically displayed using contingency tables, where rows and columns are used for conditioning on X and Z (or vice-versa) and the cell values are counts, probabilities, percentages or numerical summaries such as averages of Y. The table then allows comparing the conditional values for different levels of Z, thereby conditioning on Z. The same information can be clearly displayed using a full-grown tree of Y on predictors X and Z. If we consider X and a single confounder Z, there are potentially five types of full-grown trees (see Figure 1). Trees of types 1, 2, and 3 exclude X and/or Z as splits and therefore the corresponding contingency tables would not exhibit Simpson’s paradox; Type 5 tree will also correspond to a no-paradox contingency table because the ordering of splits indicates a stronger X-Y relationship than a Z-Y relationship, whereas Simpson’s paradox requires the reverse (Schield, 1999). Hence, only type 4 trees can exhibit Simpson’s paradox.

For the case of a single potential confounder, we incorporate sampling error into the tree approach by considering conditional-inference trees in place of full trees. However, with more than a single confounder, the significance of tree splits no longer maps directly to the paradox’s significance. We therefore introduce a new tree, the X-terminal tree, offers a computationally effective solution for data with many potential confounders and large samples. The X-terminal trees are grown in accordance with detecting Type 4 tree structures.

Applications

We illustrate the usefulness of the proposed tree approach by examining potential confounders in causal relationships in two domains: (1) the effect of a new eGov service in India on police bribing and (2) the effect of patients’ race on waiting time for kidney transplant, for patients on the kidney transplant waitlist in the USA, maintained by the United Network for Organ Sharing.

References