The Forest or the Trees? Tackling Simpson's Paradox with Classification and Regression Trees

Galit Shmueli

and

Inbal Yahav

http://eprints.exchange.isb.edu/74

Working Paper

Indian School of Business

2014

The Forest or the Trees? Tackling Simpson's Paradox with Classification and Regression Trees

Galit Shmueli

e-mail: galit_shmueli@isb.edu

and

Inbal Yahav

e-mail: inbal.yahav@biu.ac.il

Abstract: Prediction and variable selection are major uses of data mining algorithms but they are rarely the focus in social science research, where the main objective is causal explanation. Ideal causal modeling is based on randomized experiments, but because experiments are often impossible, unethical or expensive to perform, social science research often relies on observational data for studying causality. A major challenge is to infer causality from such data. This paper uses the predictive tool of Classification and Regression Trees for detecting Simpson's paradox, which is related to causal inference. 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. The approach relies on the tree structure and the location of the cause vs. the confounders in the tree. We discuss theoretical and computational aspects of the approach and illustrate it using several real applications.

1. Simpson's Paradox and macro decisioning

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 customized customer experiences are now common. Data-driven decision making is important at different aggregation levels: population aggregates are needed for overall policy making ("macro decisioning"), while micro-data are used for personalized decisioning ("micro decisioning"[1]). An important question is therefore, *what data are needed for what level of decision making*?. This question relates to the concept of Information Quality, or InfoQ (Kenett and Shmueli (2014)), which is the potential of a dataset to achieve a specific goal using a given empirical analysis method. One of the dimensions of InfoQ is *data resolution*, which refers to the measurement scale and aggregation level of the data. This dimension is the focus of this paper. In particular, we ask which data aggregation level is needed in the context of macro decisioning.

[1] The terms macro and micro decisioning were coined by Divyabh Mishra from CrowdANALYTIX.com

A perplexing issue in the context of choosing the level of aggregation for macro decisioning 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 macro decision making question that Simpson's paradox raises is which level of data aggregation presents the results of interest. This practical question raises the challenge of identifying potential confounders and then establishing a criterion for deciding if and which 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 (personalized decision making), 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 partitioned data that gives the correct choice of action.

Pearl (2009) 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". Some authors have described the paradox in non-causal language as "the reversal of an association between two variables after a third variable (a confounding factor) is taken into account" (Schield (1999)), or "the result that a marginal association can have a different direction from each conditional association is called *Simpson's paradox*" (Agresti (2012), p.51). However, Pearl (2009) 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 literature on Simpson's paradox has focused on explaining the phenomenon, quantifying its magnitude (Blyth (1972); Zidek (1981)), the conditions where it disappears (Bishop et al. 1975) and its frequency (Pavlides and Perlman (2009)). Explanations range from mathematical arguments (*Simpson's Reversal of Inequalities* looks at the paradox algebraically in terms of ratios of numbers; see Stanford Encyclopedia of Philosophy) to 'group inhomogeneity' explanations (subpopulations of different sample sizes) to 'confounding factor' explanations. Yet, there still remains the question of which result to act on in the presence of contradicting directions. As Schield (1999) stated, "even if Simpson's Paradox were readily understood, it is not easily anticipated. There is no test for determining whether an association is spurious". Pearl (2009) notes that "there is no statistical criterion that would warn the investigator against drawing the wrong conclusion or would indicate which table represents the correct answer".

The focus of this paper is identifying potential confounding variables in a high-dimensional dataset that cause the Simpson's paradox. High-dimensional datasets are now common in many areas and are the basis for important policy

decisions. Our focus is on the pre-paradox scenario, where the goal is to identify potential confounding variables among a broad set of variables. The suspect confounders are those that would lead to Simpson's paradox. Once the potential confounders are identified, the researcher can investigate plausible causal narratives (such as self-selection) to determine the adequate data aggregate level to use.

The challenge of identifying confounding variables is encountered in fields where data from multiple sources are combined to obtain richer information for the purpose of stratifying records into more homogeneous groups. For example, actuaries gather data from many sources on states, companies and years of experience for computing insurance rates (Stenmark and Wu (2004)). These data are then aggregated across years, states, etc. for decision making purposes. Our approach is useful in such cases.

We take a data-driven approach that searches the terrain of possible relationships between the outcome of interest and the set of cause 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.

The paper proceeds as follows: In Section 2 we briefly describe classification and regression trees in the context of prediction and discuss several differences when using trees for an explanatory task. We then introduce a novel approach of using trees in the context of Simpson's paradox. Section 3 describes the use of full-grown trees and conditional inference trees for single confounders. In Section 4 we introduce a new type of tree, the X-terminal tree, for detecting Simpson's paradox in datasets with many potential confounding variables. The paper concludes with a discussion in Section 5.

2. Classification and Regression Trees for Prediction vs. Explanation

2.1. Trees in Predictive Modeling

Classification and Regression Trees ("trees") are a popular machine learning algorithm for predicting an outcome from a set of predictors. Trees use the notion of recursive partitioning, where at each step the sample is partitioned into subsamples to create the most homogeneous subsamples in terms of the outcome. Partitioning is obtained by searching (heuristically) across all the predictors and all their values for the best split. The best split is the one that splits the data into the most homogeneous sub-samples in terms of the outcome variable. The result of a single step is therefore a split on one predictor at a particular value (or in the case of a categorical predictor, a split into groups of categories). Given a partition from one step, the algorithm seeks the next (conditional) split. The end result is a set of rules based on predictor thresholding ("if Age<20 and Employed=Yes") that splits the data into non-overlapping

subsamples. Tree algorithms differ in terms of the criteria for variable selection, splitting rules, and avoiding over-fitting the data. Classic machine learning algorithms such as CART, C4.5 and C5.0 use "node impurity" measures to select variables and splits for generating the full tree, followed by a pruning step that cuts back partitions that do not improve predictive power. In contrast, CHAID and conditional-inference (CI) trees (Hothorn et al., 2006a) use statistical tests of association of the predictors with Y. In this scheme of recursive partitioning, at each stage predictors are ordered according to the significance of their (conditional) association with Y, and the strongest associated predictor is chosen for splitting. However, the split is only performed if the significance level exceeds a pre-determined threshold. Tree growth is stopped when the significance threshold is not met, thereby implying insufficient association strength.

In either case, the resulting tree is used for predicting the outcome for new records.

Similar to stepwise selection techniques in regression, trees are also a useful tool for variable selection and dimension reduction. Important predictors are those chosen by the algorithm for splitting the sample into homogeneous subsamples. The presence or absence of a predictor from a tree is therefore informative of its importance. Due to the recursive nature of the algorithm, predictors that appear at the top of the tree (generating early splits) are more likely to have stronger predictive power.

An advantage of trees over other predictive algorithms is that a tree is interpretable in terms of the relationship between the predictors and the outcome, and can be translated to a set of easily understood IF-THEN rules.

2.2. Trees for Explanatory Modeling

While trees are common in predictive analytics, they are nearly absent from causal modeling. In predictive modeling, trees are used for micro-decisioning, where a prediction is required for each new record. In this paper we consider an approach for using trees in macro decisioning in the context of explanatory modeling, and in particular, for automatically identifying confounders that might cause Simpson's paradox.

Our use of trees in this explanatory modeling context differs from predictive modeling in a few ways. First, 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 ("score") new records. Third, in some cases we use full grown trees ("full trees") which overfit the sample[2], 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. (2006a)) 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

[2]the full tree favors bias reduction over variance reduction

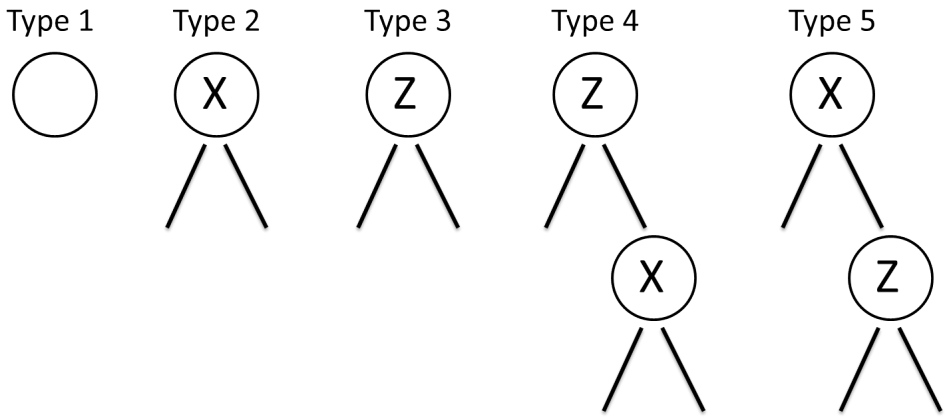

FIG 1. *Schematic of the five types of trees for a single causal variables and a single confounding variable*

potential confounders in high-dimensional data. Each of these approaches are motivated, described, and evaluated in the following sections.

3. Simpson's Paradox and Trees: Single Confounder Case

Consider a dataset with an outcome (effect) of interest (Y), a cause variable (X) and a set of potential confounding variables ($\underline{Z}$). Our tree approach starts as follows: Fit a tree with the outcome Y and predictor set that includes X and $\underline{Z}$. If Y is categorical, we fit a classification tree. If Y is numerical, we fit a regression tree. The structure of the resulting tree will then yield the potential confounders and their confounding behavior. In particular, we examine the presence and absence of each of X and $\underline{Z}$ variables as splitting variables in the tree, as well as the splitting sequence.

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):

Type 1 Tree: no splits ("stub"). A stub indicates no association between Y and X, nor between Y and Z. Hence, splitting by either does not create more homogeneous subsamples.

Type 2 Tree: only splits on X. This structure indicates an association between X and Y, with no confounding effect of (the absent) Z.

Type 3 Tree: only splits on Z. The absence of X indicates no association between X and Y. The presence of Z indicates that Y is associated with Z.

Type 4 Tree: split first on Z then on X. This tree indicates a relationship between X and Y that is confounded by Z.

Type 5 Tree: split first on X then on Z.

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, as one might expect. The more surprising result is that the type 5 tree will also correspond to a no-paradox contingency table. The reason is that the ordering of splits, where X occurs earlier, indicates that the X-Y relationship is stronger than the Z-Y relationship. According to Cornfield's condition, Simpson's paradox will only occur if Z has the strength - the effect size - necessary to nullify or reverse an observed association between X and Y (Schield (1999)).

Among the five tree types, therefore, only the type 4 tree corresponds to a contingency table that can exhibit Simpson's paradox. The ordering of the splits with Z earlier than X assures that the reversal is possible according to Cornfield's condition. Note that a type 4 tree does not guarantee a paradox. A type 4 tree can also correspond to a contingency table exhibiting a paradox that is statistically insignificant, a partial paradox where the effect is reversed only for some of the subgroups, or no paradox, where the effect for different subgroups differs in magnitude but not direction.

The result regarding tree structure and Simpson's Paradox is general to full trees grown using statistical independence tests (CHAID and conditional-inference trees) as well as impurity measures (e.g., entropy and Gini index). See Appendix 5 for proofs for both cases.

For the case of a single potential confounder, we can easily incorporate sampling error into the tree approach by considering conditional-inference (CI) trees in place of full trees. CI trees incorporate sampling error by using statistical inference for stopping tree growth. For a single confounding variable, CI trees therefore have the advantage of providing information about the statistical significance of the Paradox in terms of generalizing to the population. Setting the statistical significance threshold of the paradox to a preferred value will therefore limit findings to potential paradoxes that are significant at least at that level.

In Sections 3.1 and 3.2, we illustrate the full-tree approach for two classic examples of Simpson's Paradox. We show the advantage of the full tree and conditional tree approach over contingency tables and compare with a logistic regression approach (Agresti, 2012). Section 3.3 describes a two-confounder example where the tree type eliminates the possibility of a paradox. The example also illustrates the challenge encountered with using CI trees with more than one confounder, thereby motivating the need for a new type of tree, which is introduced and illustrated in Section 4.

TABLE 1
Death Sentence Rate, by Defendant's Race

Defendant's Race	**Death Sentence Rate**
Black	10.24%
White	**11.88%**

TABLE 2
Death Sentence Rate, by Defendant's Race and Victim's Race

Defendant's Race	**Victim's Race**	**Death Sentence Rate**
Black	Black	**5.83%**
	White	**17.46%**
White	Black	0.00%
	White	12.58%

3.1. Example 1: Death Sentence Rates

We use the example of death sentence rates from Agresti (2012). The data include information on 326 murder cases. In each case, data is available on the race of the defendant (X={white,black}), whether the outcome was a death sentence (Y={yes,no}), and the race of the victim (Z={white,black}). The question of interest is whether the defendant's race affects the probability for a death sentence. The potential confounder is the race of the victim.

Examining the contingency tables of the aggregate and disaggregate data indicates Simpson's paradox. The aggregate table (Table 1) indicates that white defendants are more likely to get the death sentence than black defendants. In contrast, the contingency table disaggregated by victim race (Table 2) indicates that white defendants are less likely to get the death sentence, when the victim is black as well as when the victim is white.

Figure 2 shows the full-grown classification tree of death on predictors X (defendant's race) and Z (victim's race). As expected, the resulting tree is of type 4, with an initial split on Z followed by splits on X. The tree also clearly displays a higher death sentence rate for black defendants (nodes 3,5) compared to white defendants (nodes 4,6). Figure 3 shows a conditional inference tree with statistical significance threshold set to 5%. In contrast to the full tree, the CI tree has a single split on victim's race (tree type 3), thereby indicating that the Simpson's paradox observed in the contingency table (and full-grown tree) is not generalizable to the population. In other words, the death sentence rate is unrelated to the race of the defendant (at 5% significance).

Lastly, we compare to an alternative approach that accounts for sampling error: fitting a logistic regression. A logistic regression of death sentence on victim race, defendant race and their interaction results in a statistically significant effect only for the victim's race (see Table 3; sequential elimination of insignificant covariates results in a model with only the victim's race), indicating that the confounding effect of the defendant's race seen in the sample is not generalizable

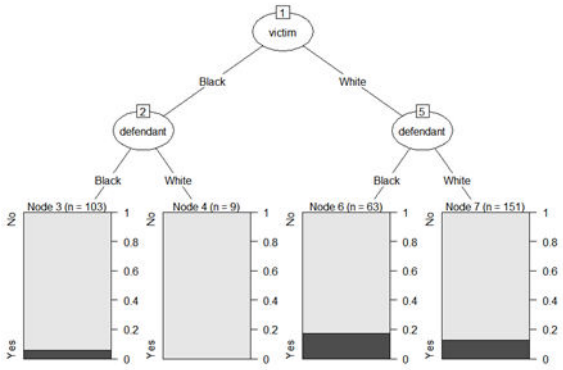

FIG 2. *Full-grown classification tree based on the death sentence disaggregate data, with defendant and victim race as predictors*

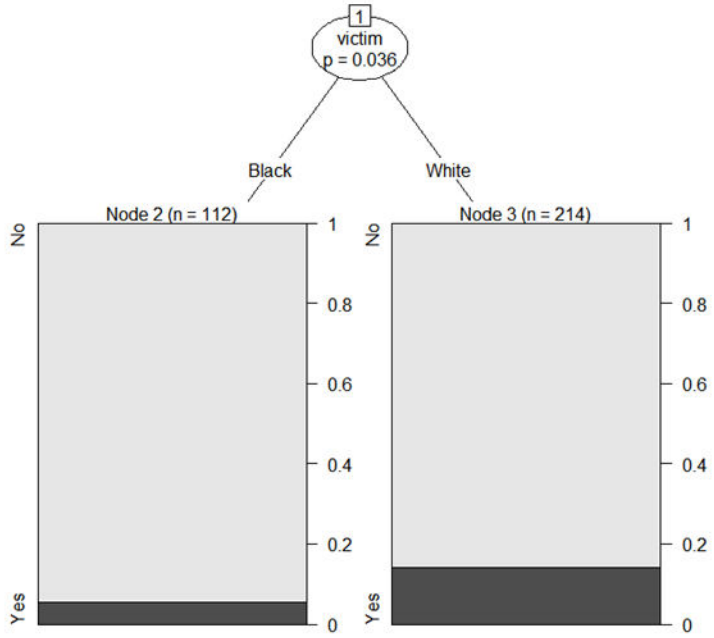

FIG 3. *Conditional-inference classification tree for the death sentence data*

to the population. Hence, in this example where the paradox does not generalize to the population, the same result is indicated by the logistic regression and the CI tree.

TABLE 3
Logistic regression model for death sentence with predictors defendant race, victim race, and their interaction

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.7830	0.4207	-6.62	0.0000
defendant	-13.7831	799.8483	-0.02	0.9863
victim	1.2296	0.5358	2.29	0.0217
defendant×victim	13.3981	799.8485	0.02	0.9866

3.2. Example 2: Berkeley Admissions

Probably the most famous example of Simpson's paradox is the Berkeley admissions case. Given data on admissions to the different departments at Berkeley, and given the gender of each applicant, the aggregate data indicated a lower rate of admissions for women (see Table 4). However, when broken down by department, admission rates were higher for women in almost every department. The question that arose was the existence of a gender bias in admissions.

The full-grown classification tree of the data (Figure 4) is, as expected, of type 4, where splits on department (Z) precede splits on gender (X).

Table 5 shows the estimated logistic regression model for gender, department and their interaction. In this case, gender, several department dummy variables and almost all interaction terms are statistically significant. However, this does not guarantee that the Simpson's paradox occurs, because it is unclear whether the effect of department is stronger than that of gender.

This example highlights two advantages of the tree over the logistic regression approach. First and most importantly, the hierarchical nature of the tree allows us to determine whether X appears before or after Z which is required for meeting Cornfield's criterion (as in tree type 4). While a logistic regression tells us only whether an interaction term is statistically significant, it does not tell us whether the $Z - Y$ relationship is stronger than the $X - Y$ relationship. In contrast, Figure 5, showing the tree for the Berkeley admissions data, is a type 4 tree thereby indicating a potential Simpson's Paradox that generalizes to the population at a 5% significance level.

Second, the tree automatically explores interaction terms, and in cases of multi-category or numerical covariates the binning is automated (see grouping of departments in Figure 5).

3.3. Example 3: Injury in Automobile Accidents and Seat-Belt Use

To further compare the tree approach and regression approach, and to illustrate the challenge of CI trees in the case of more than one confounder, consider the following example with two confounders. Agresti (2012) analyzes a

TABLE 4
1970's Berkeley Admissions, by gender and department

		Gender	
Department		Female	Male
A	Avg. Admitted-yes	82%	62%
	Number of Records	108	825
B	Avg. Admitted-yes	68%	63%
	Number of Records	25	560
C	Avg. Admitted-yes	34%	37%
	Number of Records	593	325
D	Avg. Admitted-yes	35%	33%
	Number of Records	375	417
E	Avg. Admitted-yes	24%	28%
	Number of Records	393	191
F	Avg. Admitted-yes	7%	6%
	Number of Records	341	373
Grand Total	Avg. Admitted-yes	30%	44%
	Number of Records	1,835	2,691

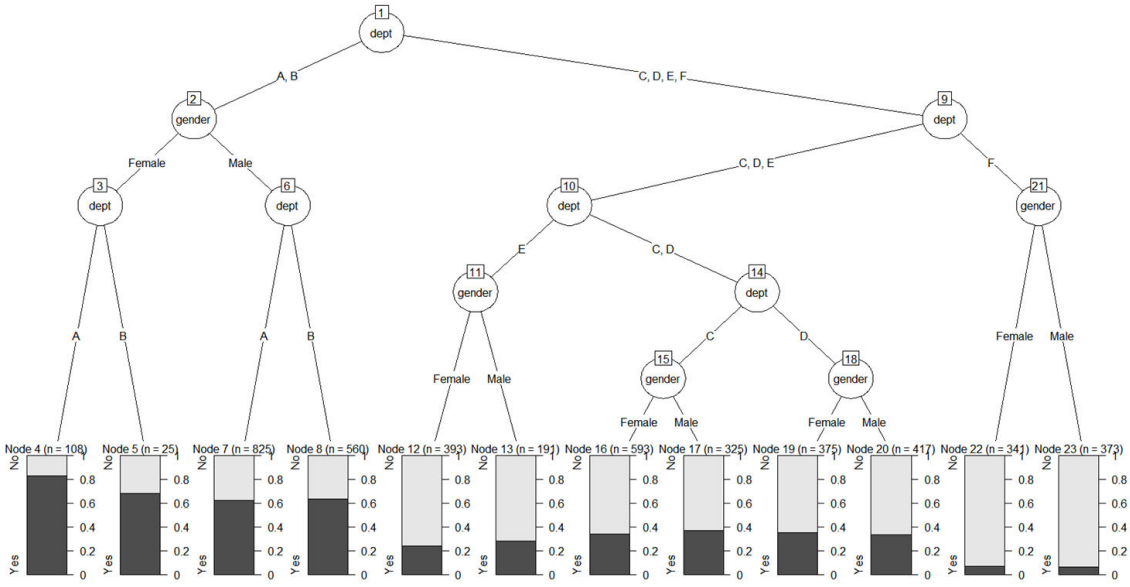

FIG 4. *Full-grown classification tree based on Berkeley admissions data, with gender and department as predictors*

TABLE 5
Logistic regression model for Berkeley admissions with predictors gender and department and their interaction terms

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.5442	0.2527	6.11	0.0000
deptB	-0.7904	0.4977	-1.59	0.1122
deptC	-2.2046	0.2672	-8.25	0.0000
deptD	-2.1662	0.2750	-7.88	0.0000
deptE	-2.7013	0.2790	-9.68	0.0000
deptF	-4.1250	0.3297	-12.51	0.0000
genderMale	-1.0572	0.2627	-4.02	0.0001
deptB×genderMale	0.8372	0.5104	1.64	0.1009
deptC×genderMale	1.1821	0.2995	3.95	0.0001
deptD×genderMale	0.9752	0.3026	3.22	0.0013
deptE×genderMale	1.2574	0.3303	3.81	0.0001
deptF×genderMale	0.8683	0.4027	2.16	0.0310

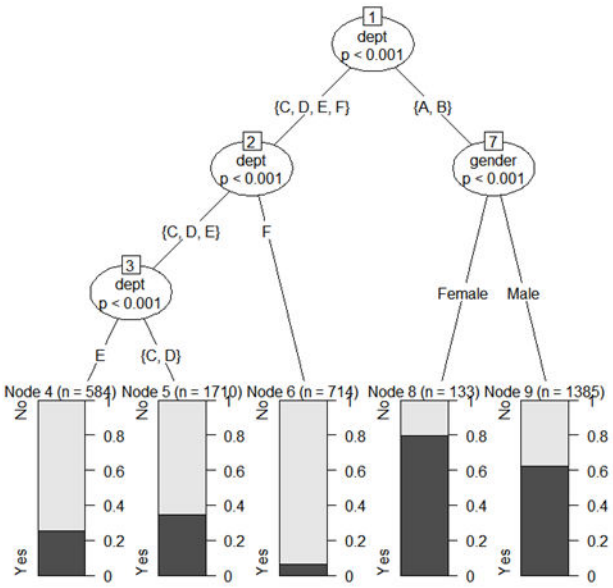

FIG 5. *Conditional-inference classification tree for the Berkeley admissions data*

dataset on 68,694 passengers in auto and light trucks involved in accidents during one year in the state of Maine. For each accident, information is available on whether the passenger used a seat-belt (yes/no) and whether the passenger was injured (yes/no). In addition, information is available on the passenger's gender (male/female) and the accident location (rural/urban). The relationship of interest is between the use of seat-belts (X) and the chance of injury (Y). Passenger gender and accident location are potential confounding variables.

A logistic regression with all three predictors and interaction terms between seat-belt use and the other covariates shows statistical significance for the main effects and for the interaction of seat-belt with location (see Table 6). The significant interaction indicates that Simpson's paradox will not manifest when the data are disaggregated by gender, but it *might* manifest when the data are broken down by location. In contrast, the tree approach gives conclusive results.

Figure 6 shows the full classification tree of injury (Y) on predictors seat-belt use (X), gender (Z_1), and location (Z_2). We see that the first split is on seat-belt, followed by splits on location and gender. Hence, this tree is of type 5, and therefore contingency tables of injury on seat-belt use will not display Simpson's paradox when disaggregated by either gender or location. This result can be seen in Table 7, which shows that the injury rate is higher without a seat-belt, both overall as well as when the sample is broken down by gender and/or by location.

TABLE 6
Estimated logistic regression model for seat-belt data, with interaction terms between seat-belt use and the other covariates (gender=1 for Male, location=1 for Urban, seatbelt=1 for Yes)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.1962	0.0311	-38.45	0.0000
seatbelt	-0.8654	0.0463	-18.69	0.0000
gender	-0.5422	0.0350	-15.51	0.0000
location	-0.8013	0.0349	-22.93	0.0000
seatbelt×gender	-0.0095	0.0559	-0.17	0.8647
seatbelt×location	0.1074	0.0550	1.95	0.0506

Note that in this two-confounder case we use the full tree (significance threshold=1) rather than setting a threshold to the significance level of interest. While

TABLE 7
Percent injuries by seat-belt use, gender and location

Gender	Location	Without Seat-belt	With Seat-belt
Female	Urban	12%	6%
Female	Rural	23%	11%
Male	Urban	7%	3%
Male	Rural	15%	7%
Total Female		16%	8%
Total Male		10%	5%
Total Urban		9%	5%
Total Rural		18%	9%
Overall Total		13%	6%

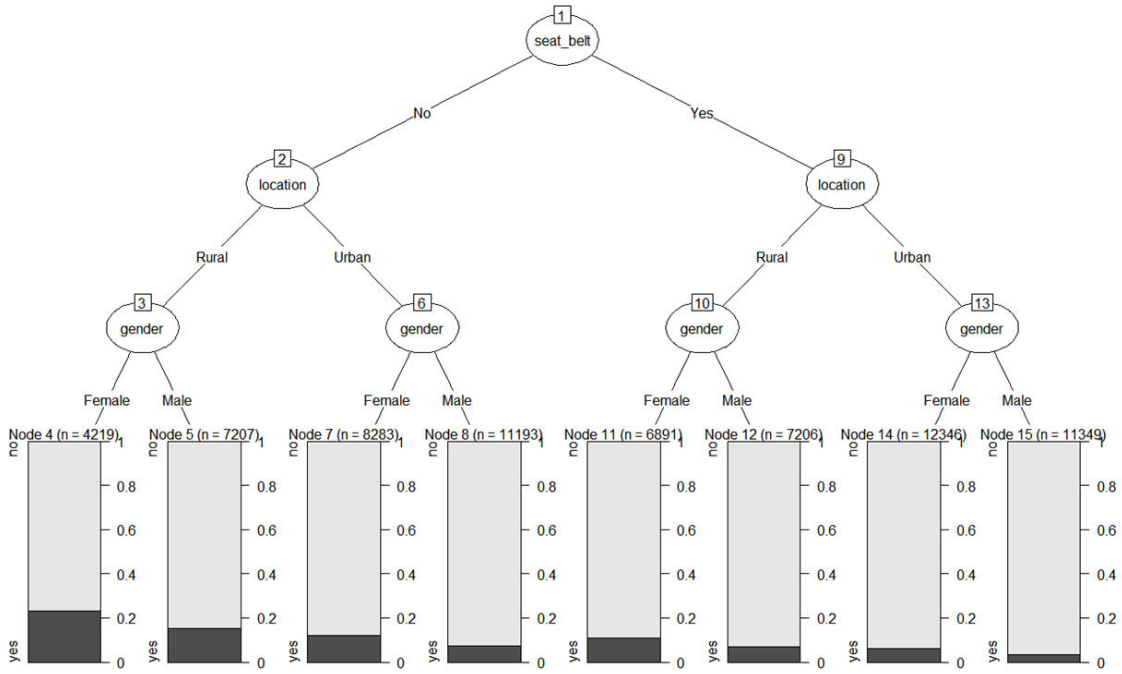

FIG 6. *Full-grown classification tree based on accidents and seat-belt use data, with seat-belt use, gender and accident location as predictors*

either tree is useful for determining whether it is a type 4 tree or not, the statistical significance threshold in a type 4 tree with more than two layers of splits does not directly map to the significance threshold imposed on the paradox. For example, consider a tree with a top split by Z_1, followed by a split by Z_2, followed by a split by X. This is a type 4 tree indicating a possible paradox. The paradox can be with respect to Z_1 or with respect to $Z_2|Z_1$. Hence, the tree's significance threshold which applies separately to each split[3] does not provide a suitable threshold for the conditional paradox.

To summarize, the advantage of CI trees over a regression approach is that trees can easily identify no-paradox situations based on the tree type. Their limitation is the ability to locate only statistical significant paradoxes, when the paradox is in fact conditional on a combination of confounders. In such a case, the full tree must be used, and statistical tests performed for each of the candidate conditional paradoxes.

4. Multiple Potential Confounders: X-Terminal Trees

Our result regarding tree structure applies not only to a single confounder Z, but to multiple confounding variables $\underline{Z}$. In this case, the full grown tree will contain X and the Z_i $(i = 1, \ldots, k)$ splits. As in the single Z case, the ordering of the splits is the key to detecting a possible paradox. Trees of types 1, 2, and 3, which exclude X and/or $\underline{Z}$, will correspond to a no-paradox contingency table. Trees of type 5, where X is the top split and one or more Z_i's appear lower in the tree, also correspond to a no-paradox contingency table. The reason, again, lies in the strength of the X-Y relationship, compared with the Z_i-Y relationships for the Z_i's in the tree.

Simpson's paradox can only occur in a type 4 tree, in which $Z_i \in \underline{Z}$ is the top split and X is a split that appears somewhere in the tree. Simpson's paradox can exist in any of the Z_i-X relationships, in which Z_i precedes X, and will not exist in all other relationships. A contingency table filtered by Z splits that precede the X split might therefore exhibit Simpson's paradox.

With multiple potential confounding variables, the advantage of using trees over scanning all possible contingency tables is clear: the tree approach easily scales up using automated variable selection, thereby detecting confounding variables that might display Simpson's paradox. Another advantage of the tree over contingency tables is that it easily generalizes to variables X and Z that are not categorical. In such cases, creating a contingency table requires binning the continuous variables, yet how to bin those variables is unclear. In contrast, the tree will automatically choose the best binning in terms of the strongest Z-Y association and if that association exists, it will be displayed in the tree as splits based on those bins.

[3]Conditional inference trees, as implemented in the R packages *party* and *partykit*, correct for multiple testing using either Bonferonni or minimum p-value resampling; a possible improvement would be using the False Discovery Rate approach (Benjamini and Hochberg (1995)), which is useful especially when a small subset of the effects does exist.

Compared to the regression approach, the tree approach is more computationally efficient. A stepwise regression that searches all interaction terms between X and the potential covariates can be computationally slow and suffer from pitfalls of spurious collinearity, among other challenges (see Agresti, 2012, p. 279). In comparison, the tree solution is extremely fast.

Yet, with more than a single potential confounder, we can no longer use CI trees since the statistical significance threshold on the complete paradox cannot be easily mapped to the threshold to use at each split. In addition, with high dimensional data, a full tree can become difficult to navigate due to the many splits and terminal nodes. Moreover, it is wasteful to construct the full tree, when all we need are paths that end in a split by the cause X. We therefore introduce a tree with a new stopping rule: the X-terminal tree, where growth is stopped when an X split is encountered. We describe this tree next.

4.1. X-Terminal Trees for Detecting Simpson's Paradox

In data with a large number of potential confounders, the resulting full tree might be overwhelming in size. For detecting confounders that might lead to Simpson's paradox, we have a sequential process: First detect tree type. If not type 4, then no paradox and stop here. If type 4, then examine only tree paths that lead to a split on X. This constrained space of information can be achieved by growing trees so that when an X split is encountered that branch is no longer grown. Using this process, the resulting tree will either be a non-paradox full-grown tree (of type 1,2,3 or 5), or else it will be a type 4 tree with at least one terminal node at X.

Algorithm 1 summarizes the process of growing an X-terminal tree for identifying Simpson's paradox in data with a large number of potential confounding variables. When a type 4 tree is encountered, we search for a paradox using the Z_i's that appear above every X split. These Z_i's are inspected conditionally, so that each Z is conditional on Z_i's from higher splits. A conditional paradox is equivalent to filtering the data by confounders from higher splits.

In the following, we use apply the tree approach with the proposed algorithm to two studies with multiple potential confounders. The goal is to identify paradoxes, and especially those that are statistically significant. In both examples, we find a type 4 tree that indicates a potential paradox. Recall that a type 4 tree can also imply a partial paradox, or a non-paradox. The examples illustrate these different possible findings. We illustrate through these examples the types of information that our proposed tree approach reveals.

4.2. Example 3: Impact of eGov Services in India

This example illustrates the usefulness of our approach in the case of high-dimensional data, where there is a large number of potential Z variables.

In the early 2000s, The Government of India (GoI) launched a national eGov plan to deliver government information and services to its citizens. The plan

Construct full-grown tree T using $\{Y, X, \underline{Z}\}$
Trim the tree at X nodes. Alternatively, stop tree from growing when an X node is reached
if $X \in T$ **then**
  if X is the top split of T **then**
    */*Tree Type 2 or 5*/*
    No significant paradox
  else
    */*Tree Type 4*/*
    Full or partial paradox might exist
    for all terminal X node **do**
      Examine terminal X nodes for paradox
      if Paradox detected **then**
        Paradox exists in path $p = \bigcup_i Z_i$ **to** X
        */*Note: Paradox might be insignificant*/*
        Examine $\bigcup_j Z_j \subseteq p$ for additional paradoxes
      else
        No significant paradox
      end if
    end for
  end if
else
  */*Tree Type 1 or 3*/*
  No significant paradox
end if

Algorithm 1: Algorithm for detecting Simpson's paradox in high-dimensional data

TABLE 8
Performance outcomes, as defined by the Government of India (Table 1.1 from Government of India (2008))

Dimension of Impact	Indicator
Cost of Availing Service (Measured Directly)	Number of trips made for the service
	Average travel cost of making each trip
	Average waiting time in each trip
	Estimate of wage loss due to time spent in availing the service
	Total time elapsed in availing service
	Amount paid as bribe to functionaries
	Amount paid to agents to facilitate service
Overall Assessment	Preference for manual versus computerized system
	Composite Score measured on 5-point scale factoring in the key attributes of delivery system seen to be important by users
Quality of Service	Interaction with staff, complaint handling, privacy, accuracy measured on 5-point scale
Quality of Governance	Transparency, participation, accountability, corruption measured on a 5-point scale

included creating an income tax portal, online delivery of passport services, and online delivery of services provided by the Ministry of Company Affairs (MCA), among others. Following the implementation of these online services, an impact study was carried out. The purpose of the study was to evaluate the success of the new electronic services in terms of several intended outcomes. For that purpose, in 2006 GoI commissioned a satisfaction survey among users of the new (online) and old (offline) services. The Indian Institute of Management, Ahmedabad designed the survey and overall assessment framework while eleven independent market research firms were empanelled to conduct the survey across the country. The survey queried users about various dimensions of their service experience including transaction costs, quality of service, and overall satisfaction. The survey also provided information on characteristics of individual respondents and the latter's ex ante perceptions of various service parameters such as clarity of rules and procedures or convenience of service facilities (see Government of India, 2008).

We focus on the assessment of one of the eGov portals, namely, the online delivery of passport services. Responses to the survey of passport services constitute a representative sample of 13 passport offices selected from different regions of the country. A sample of 9500 users was drawn from cities/towns where these offices were located. The study was designed as a quasi-experiment, where a large group of offline users was selected to match a group of online users in terms of their geography and demographics (age group, education level, etc.). Table 8 presents the outcome variables (Y's) of interest, as defined by the government. Table 9 describes the key survey questions.

The main question is whether the introduction of the online system (X =

TABLE 9
Survey questions that are potential Z variables

Variable	Variable Description	Scale
Awareness	Awareness of electronic services provided by the Government of India	Binary, 1='aware'
Availability	Ease of availability and accessibility of information pertaining to the service	5-point Likert scale
Experience	Indicator of prior usage of any other e-Gov application	Binary, 1='prior usage'
Clarity of Processes	Extent to which the processes and procedures characterizing the e-Gov application are clear and simple	5-point Likert scale
Clarity of Rules and Procedures	Extent to which the rules and procedures characterizing the e-Gov application are stated clearly without ambiguity and mistakes	5-point Likert scale
Convenience of Hours	Extent to which the working hours of the passport center or office are perceived as convenient	5-point Likert scale
Convenience of Location	Extent of satisfaction with the present location of the passport center or office	5-point Likert scale
Form Design	Extent of satisfaction with the design and layout of the application forms	5-point Likert scale
Service Area Facilities	Extent of satisfaction with the service area facilities	5-point Likert scale

TABLE 10
Police Bribes Rate, by Online/Offline Groups

Group	**Bribes Rate**
Offline	42.43%
Online	48.08%

{Online, Offline}) affected the outcomes of interest.

There are two sources of potential confounding variables: the demographic information (which was designed to be balanced across the online/offline groups) and questions in the survey that might be informative of confounding variables that reverse the impact of the online system on the outcome. For example, perhaps people who live far from a passport office are affected by an online passport service differently from people who live close to an office. The approach is then to fit an X-terminal tree with X = {Online, Offline} and $\underline{Z}$ = {demographic variables; survey questions}, separately for each outcome Y.

We focus on one of the important measures of the service's success: the reduction in police bribes. A naive summary for answering this question compares the police bribe rate in the online vs. offline groups. Table 10 shows the results, which indicate that the online system leads to an average of 6% higher bribe rates to the police (the difference is statistically significant; p-value$\approx$0).

Applying an X-terminal classification tree with X and $\underline{Z}$, and stopping growth at X terminal nodes (for a type 4 tree), we obtain the tree in Figure 7. For better visibility, we do not plot terminal nodes for paths that do not contain X splits. For example, the left most path in the trimmed tree, where $Trust \leq 2$, is removed from the figure. The next path, where $Trust > 2$, contains an X split, and therefore was not removed. The bar charts in the terminal nodes reflect the police bribe rates for the online (left) and offline (right) groups. Terminal nodes that exhibit Simpson's paradox (higher bribes for the offline group) are circled. The resulting tree is of type 4, with online/offline appearing as a split, but not as the top split.

We examine the tree to detect potential Simpson's paradox in the dataset. Looking at the tree, there are several sub-zones in the dataset in which the paradox exists. As explained earlier, the paradox might be insignificant even if the corresponding split is significant.

Tables 11-13 summarize three paths in the tree that end at an X split, illustrating the three possible results in terms of a paradox:

1. The right-most path in the tree correspond to the contingency table in Table 11, which exhibits a *significant* Simpson's paradox (p-value=0.003).
2. The fourth-from-right circled terminal node corresponds to the contingency table in Table 12, which exhibits an *insignificant* Simpson's paradox (p-value=0.16).
3. The second-left terminal node (not circled) corresponds to the contingency table in Table 13, and both exhibit *no* Simpson's paradox. Although there is an X split in this path, it does not lead to Simpson's paradox.

In paths that do not terminate at an X split, there should be no paradox.

TABLE 11
Police bribes rate by Online/Offline, filtered by $\underline{Z}$ factors from right-most path in the tree

Filters	
City	Cochin
Occupation	*Not* Businessman, Cultivators, Dependent, Others
Convenience of Hours	≥ 3
HH Role	Family Member
Clarity of Process	≥ 4
Group	**Bribes Rate**
Offline	87.93%
Online	52.38%

TABLE 12
Police bribes rate by Online/Offline, filtered by $\underline{Z}$ middle path in the tree

Filters	
City	Delhi
Awareness	Yes
Experience	Yes
Trust	> 4
Education	Diploma, Higher Secondary, Illiterate, Literate without Education, Matric
House	*Not* Temporary
Age	> 36
Convenience Hours	≥ 3
Group	**Bribes Rate**
Offline	100%
Online	92.5%

Two possibilities are a path that appears in the full tree but does not lead to an X split, and a path that does not appear in the tree altogether. The first case is shown in Table 14, which presents a path with no X split[4]. As expected, the corresponding contingency table does not exhibit Simpson's paradox. Table 15 illustrates the second case, of a path that is not in the tree. The corresponding contingency table, as expected, does not exhibit Simpson's paradox.

In addition to searching for a paradox right next to X splits (by filtering the data by all the Zs on the path), we use Algorithm 1 to also examine partial paths. A strong paradox in a particular area of the data might still appear even when we look at larger sub-zones of the data. Table 16 presents an example of this case, where a partial path exhibits a significant paradox.

4.3. Example 4: Kidney Transplant Waitlist

The last study illustrates the application of our tree approach to a large dataset with multiple potential confounders, in which we obtain a type 4 tree, but uncover no Simpson's paradox. In this example, the outcome of interest is continuous and therefore we use a regression X-terminal tree.

Acute renal failure, also referred to as acute kidney failure, is a medical condition in which the kidneys are no longer able to remove waste from the

[4]Note that the actual path is not visible in the tree. All paths that do not contain an X split were removed from the tree for better visibility.

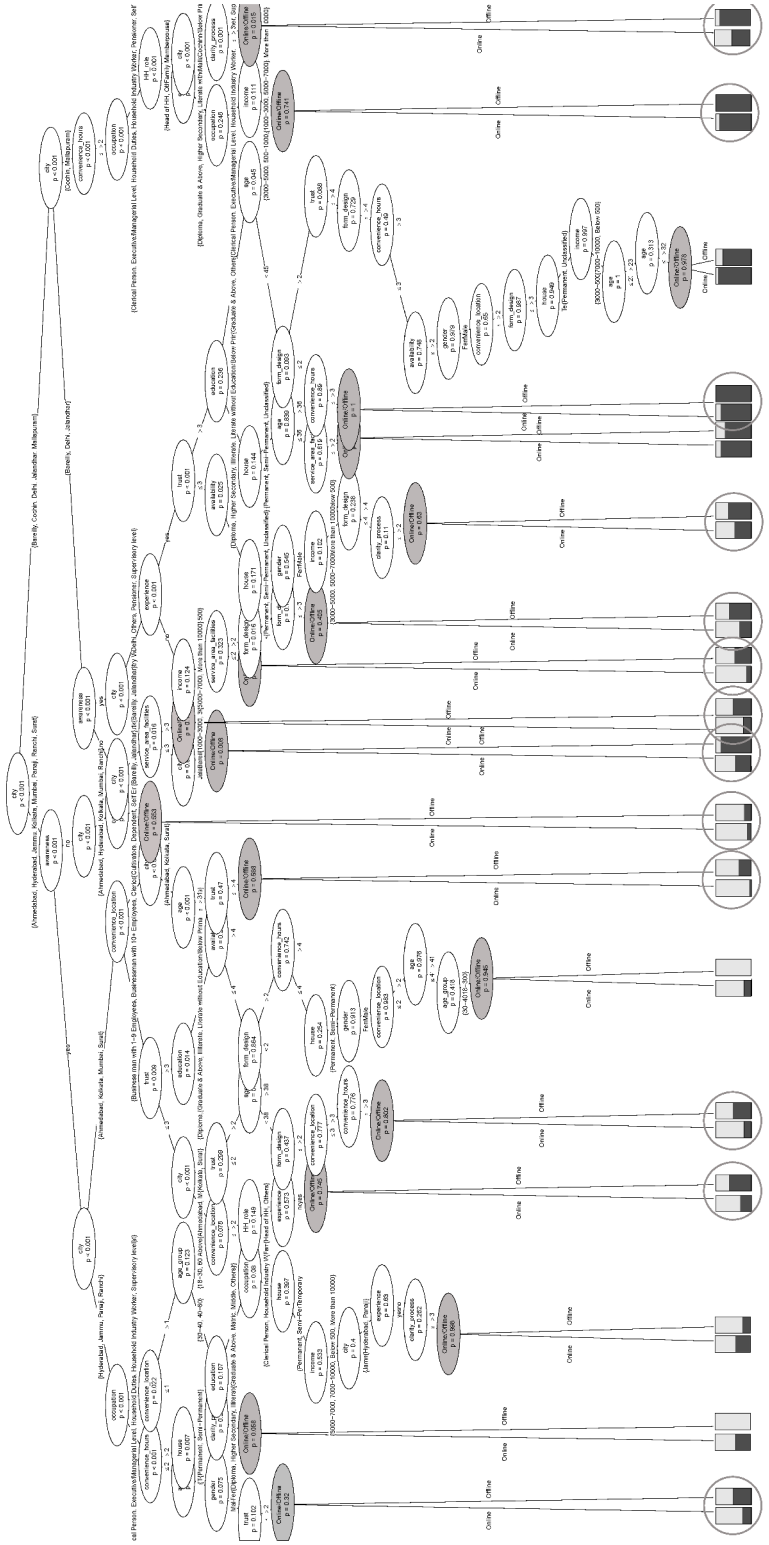

FIG 7. *X-terminal Classification tree for eGov data with group, demographic, and survey variables as predictors; bar charts show bribe rate for offline (right) and online (left) groups.*

TABLE 13
Police bribes rate by Online/Offline, filtered by $\underline{Z}$ factors from second-left path in the tree

Filters	
City	Hyderabad, Jammu, Panaji, Ranchi
Occupation	Businessman, Cultivators, Dependent, Others, Pensioner, Self Employed, Student
Convenience of Hours	> 2
House	Permanent, Semi-Permanent
Clarity of Process	≤ 2

Group	**Bribes Rate**
Offline	11.22%
Online	40.00%

TABLE 14
Police bribes rate by Online/Offline, filtered by $\underline{Z}$ factors from the complement of the right-most path in the tree (no Simpson's Paradox)

Filters	
City	Cochin
Occupation	*Not* Businessman, Cultivators, Dependent, Others
Convenience of Hours	≥ 3
HH Role	Family Member
Clarity of Process	< 4
Group	**Bribes Rate**
Offline	87.02%
Online	95.14%

TABLE 15
Police bribes rate by Online/Offline, filtered by $\underline{Z}$ factors that do not appear in the tree (no Simpson's Paradox)

Filters	
City	Jalandhar
Awareness	Yes
service_area_facilities	≤ 3

Group	**Bribes Rate**
Offline	43.59%
Online	53.68%

TABLE 16
Police bribes rate by Online/Offline, filtered by subset of $\underline{Z}$ factors from right-most path in the tree

Filters	
City	Cochin
Clarity of Process	*ge*4
Group	**Bribes Rate**
Offline	88.41%
Online	78.69%

TABLE 17
Waitind time in days, by Patient's Race

Race	Waiting Time
Black	716.86
White	472.49
Total	545.38

blood and control the level of fluid in the body. Treatments for acute kidney failure include dialysis and a kidney transplant. Dialysis treatment is available to everyone. However, it is not the preferred option by most patients, as it significantly reduces quality of life. In contrast, for kidney transplant, there is a continual supply shortfall: approximately 13,500 transplant options are offered annually in the USA, compared to 30,000 new patients with renal failure.

The United Network for Organ Sharing (UNOS, `www.unos.org`) is a private, non-profit organization that manages the organ transplant system in the USA under contract with the federal government. UNOS manages patients waitlists per organ and an allocation policy that determine the order in which candidates are offered an organ, when one becomes available. Under the current kidney allocation system in the United States, kidneys are allocated to patients primarily through a combination of tissue matching, sensitization level (the level of sensitization to donor antigens, measured by Panel Reactive Antibody), and waiting time.

A question of interest is whether waitlist patients' race (X) affects the time they wait for a kidney (Y). A naive summary for answering this question compares recipients' waiting time (in days) by race. Table 17 shows the results, where black recipients appear to wait longer than white recipients.

This difference is statistically significant (p-value$\approx$0), and more importantly[5], it appears practically significant. The waitlist dataset[6] includes many variables with information about waitlist patients, which are potential confounders ($\underline{Z}$):

INIT_YEAR Year placed on waiting list
MED_COND_TCR Recipient medical condition at registration
FUNC_STAT_TCR Recipient functional status at registration
DGN_TCR Primary diagnosis at registration
ON_DIALYSIS Indicator of whether recipient is on dialysis
DIAB Indicator of whether recipient has diabetes
ABO Blood type
AGE Recipient's age at registration
ORGAN Organ that the patient is listed for: kidney (KI), pancreas (PA), or both (KP)
HLAMIS Mean patient's antigen match with donor's pool

The question is whether Simpson's paradox will manifest when any of these variables are included. To address this question, we apply an X-terminal regression tree with $Y = Wait-time$, $X = RACE$ and $\underline{Z}$. The resulting tree is plotted in

[5]Statistical significance is not surprising, given the very large sample.
[6]For detailed information about the dataset and its context, see Yahav and Shmueli (2014).

Figure 8 and is of type 4 (RACE appears in the tree, but not as the top split). Following Algorithm 1, we examine all paths that terminate with RACE. It turns out that none of them exhibit Simpson's paradox. We therefore conclude that there is no paradox in the dataset, that is, black recipients wait longer for kidney transplants compared to white recipients across all the subgroups that we examined[7].

5. Discussion

We have introduced and adapted the use of a popular predictive algorithm, classification and regression trees, to a context where it is rarely used: explanatory modeling and in particular, identifying potential confounders that lead to Simpson's paradox. Using trees in this different context warrants several conceptual differences in its use as well as adaptations to tree growth criteria. First, we are mostly interested in the presence or absence of X and Z as splitting variables in the tree and especially in the sequence of splitting by X and Z. In predictive modeling, the main interest is in the terminal nodes, which provide the predicted value or class. The absence of presence of predictors in the tree are useful as a secondary goal of variable selection.

Second, in predictive modeling the main concern is over-fitting the tree to the training data, so that it is unable to generalize to new records, thereby having low predictive power. Hence, there exist various approaches to avoiding overfitting, the most common approach by pruning the tree using a holdout dataset or cross-validation. In the Simpson's paradox scenario, we are most interested in detecting the tree splitting sequence to determine whether the tree is of type 4 or not. With a single potential confounder, we can use CI trees and set the splitting threshold equal to the statistical significance required for the paradox. With multiple potential confounders, the significance threshold of the splitting criterion no longer coincides with the overall paradox significance. Hence, we cannot use CI trees. While we could potentially use full-grown trees, they are typically too large and complicated, thereby providing an inefficient solution for detecting tree type and potential paradox paths. We therefore introduce an efficient algorithm that stops tree growth whenever encountering an X split. The resulting X-terminal tree is smaller than the full tree and helps eliminate many paths that necessarily lead to non-paradox terminal nodes.

The tree approach helps the decision maker explore four types of questions related to Simpson's paradox:

1. Is there indication of an $X - Y$ relationship at all? If an association exists, we expect to find X as a splitting variable in the tree. Hence, trees of types 1 and 3 indicate no effect while types 2, 4 and 5 indicate an effect.
2. Does the variable Z confound the cause-effect relationship? The presence of Z before X (tree type 4) indicates a possibility of such confounding.

[7] Other explanations to this phenomena may apply, such as distribution of donors' antigen levels or religious beliefs

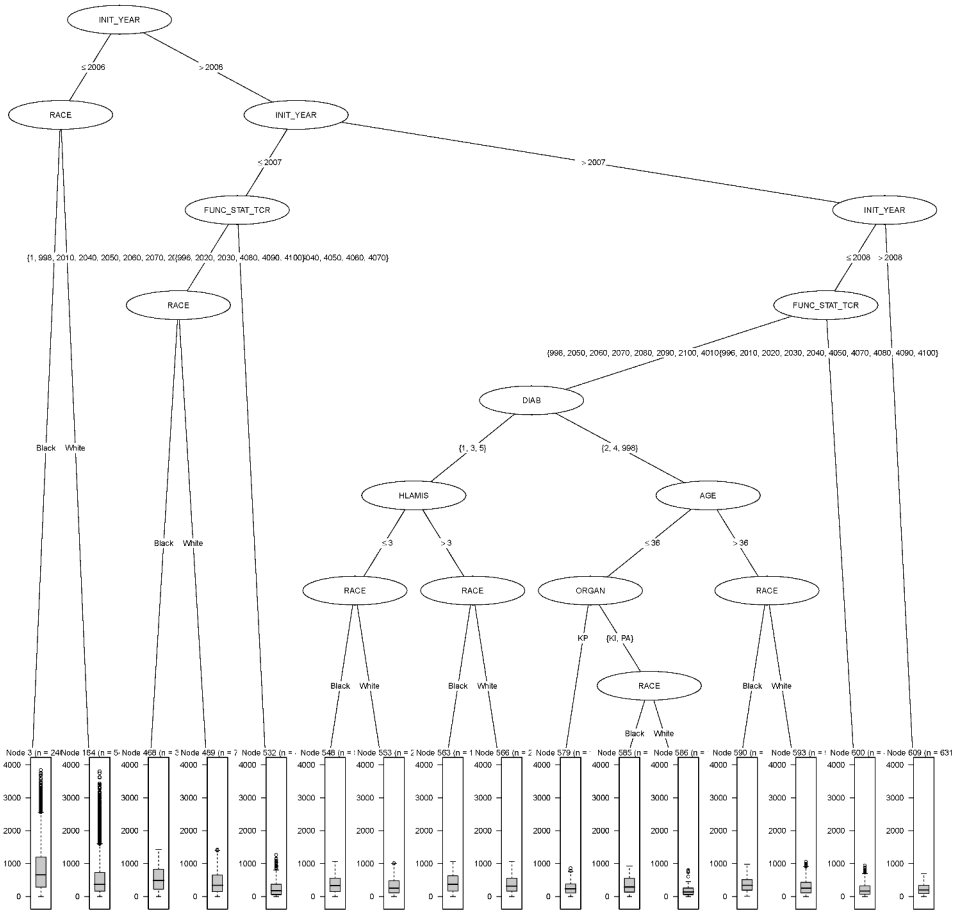

FIG 8. *Conditional-inference classification tree based on kidney waitlist data, with multiple patient variables as predictors*

3. What is the form of the confounded relationship? Given tree type 4, the actual splits on Z give an indication of the structure of the confounding. For example, for a multi-category confounding variable, the tree automatically identifies the particular grouping of categories that should be treated separately. Similarly, for a continuous confounding variable, the tree automatically identifies the splitting value for that variable.
4. What is the magnitude of the effect of X on Y? Given that X is present in the tree (tree types 2, 4 and 5), the corresponding terminal nodes of the tree can be compared to quantify the effect. In tree types 2 and 5, where an aggregate decisioning is warranted, each terminal node corresponds to the a category (or category groups) of X that differ form other categories in terms of Y. In tree type 4, where a disaggregate level decisioning is warranted, effect magnitudes should be compared separately within each level of Z.

Generating an exact classification or regression tree is a computationally difficult problem that involves exhaustive search, and practically, infeasible. The reason is that searching the optimal splits in continuous variables involves ranking, and hence is of order $n \log n$. For categorical variables with K categories, the number of possible splits is $2^{K-1} - 1$. Therefore, a fast heuristic is typically used to generate a sub-optimal tree for big data.

The conditional-inference tree algorithm used in this paper, is the common algorithm by Strasser and Weber (1999). This algorithm is a 'greedy' heuristic to construct CI trees, in which splits are selected based on local minimization criteria. Moreover, only a small subset of possible splits is evaluated at each step. Specifically, the algorithm works as follows[8]:

1. (a) Test the null hypothesis of independence between each input variable ($X, \underline{Z}$) and the response (Y). The test can be multivariate if X and/or $\underline{Z}$ are not binary. Stop if this hypothesis cannot be rejected. Otherwise select the input variable with strongest association to the response, measured by the univariate p-value. This selection avoids selection bias favoring input variables with multiple possible splits.
   (b) After selecting a split, compute its p-value. For non-binary input variables, Monte Carlo simulation is used to generate random permutations. P-values are computed for each permutation.
2. Split the tree by the selected input variable, using a binary split.
3. Recursively repeat steps 1 and 2.

The implication of the heuristic to our approach is that weak paradoxes and paradoxes in small sub-samples might not be detected. For example, consider a non-binary confounder Z (either continuous, or categorical with $K > 2$ levels). Assume also that the $X - Y$ relationship is stronger than the overall (multivariate) $Z - Y$ relationship, yet for a specific cutoff c ($Z_c = \{Z < c, Z \geq c\}$) that *reverses* the $X - Y$ relationship, the $Z_c - Y$ relationship is actually stronger

[8] For a detailed description of the methodology see Hothorn et al. (2006a,b)

than the $X-Y$ relationship. In this case, the algorithm described above will not detect the Simpson's paradox. Another example is a case where the multivariate relationship $Z-Y$ is in fact stronger than the relationship $X-Y$, yet the cutoff c was not tested in step 1(b) of the heuristic.

An alternative to our tree-based method that might overcome this challenge is a forest-based approach. Here, a relatively large number of bootstrap samples are used to generate multiple CI trees. It is likely that at least one of the trees in the forest will detect the weak paradox described above. However, due to sampling error, it is also possible that the forest will detect paradoxes in subset(s) of the data that do not occur in the entire dataset. The forest therefore might lead to over-detection of both real and false paradoxes.

References

Agresti, A. (2012) *Categorical Data Analysis*, Third Edition, Wiley and Sons.

Alin, A. (2010). "Simpson's paradox". *Wiley Interdisciplinary Reviews: Computational Statistics*, vol 2(2), pp. 247–250.

Benjamini, Y. and Hochberg, Y, "Controlling the false discovery rate: a practical and powerful approach to multiple testing". *Journal of the Royal Statistical Society*, Series B, vol 57 (1), pp. 289–300.

Blyth, C. R. (1972). "On Simpson's paradox and the sure-thing principle". *J. American Statistical Association*, vol 67, pp. 364–366.

Department of Information Technology, Ministry of Communications and Information Technology, Government of India and Indian Institute of Management, Ahmedabad (2008), "Impact Assessment of e-Governance Projects", Report, `www.iimahd.ernet.in/egov/documents/impact-assessment-of-egovernance-projects.pdf`

Donald P. Green and Holger L. Kern, "Modeling heterogeneous treatment effects in large-scale experiments using Bayesian Additive Regression Trees", `andrewgelman.com/movabletype/mlm/Green%20and%20Kern%20BART.pdf`

Hothorn, T., Hornik, K. and Zeileis, A. (2006), "Unbiased Recursive-Partitioning: A Conditional Inference Framework", *Journal of Computational and Graphical Statistics*, vol 15(3), pp. 651–674.

Hothorn, T., Hornik, K., Van De Wiel, M. and Zeileis, A. (2006), "A lego system for conditional inference", *The American Statistician*, vol 60(3), pp. 257-263.

Kenett, R. S., and Shmueli, G. (2014), "On Information Quality", *Journal of the Royal Statistical Society*, Series A, vol 177(1), pp. 3–38.

Lipovetsky, S and Conklin WM (2006), "Data aggregation and Simpsons paradox gauged by index numbers", *European journal of operational research*, vol 172(1), pp. 334–351.

Pagano, M. and Gauvreau, K. (2000), *Principles of biostatistics*, Duxbury Pacific Grove.

Pavlides, M. G. and Perlman, M. D. (2009), "How Likely is Simpson's Paradox?", *The American Statistician*, vol 63(3), pp. 226–233.

Judea Pearl (2009). *Causality: Models, Reasoning, and Inference*, Cambridge University Press, 2nd edition.

Radelet, M. (1981), "Racial characteristics and imposition of the death penalty", *American Sociological Review*, 46, 918–927.

Schield, M. (1999), "Simpson's Paradox and Conrfield's Conditions", in *Proceedings of the Section on Statistical Education, Joint Statistical Meeting of the American Statistical Association*, pp. 106111.

hmueli, G., Bruce, P. B., and Patel, N. (2010), *Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner*, Wiley and Sons, 2nd edition.

Simpson, E.H. (1951), "The Interpretation of Interaction in Contingency Tables." *Journal of the Royal Statistical Society B* 13: 238-241.

Strasser, H. and Weber, C. (1999), "On the asymptotic theory of permutation statistics." *SFB Adaptive Information Systems and Modelling in Economics*

and Management Science, WU Vienna University of Economics and Business.

Stenmark, J. A. and Wu, C.-S. P. (2004), "Simpsons Paradox, Confounding Variables and Insurance Ratemaking", *Proceedings of the Casuality Actuarial Society Annual Meeting.*

Yahav, I. and Shmueli, G. (2014), "Outcomes Matter: Estimating Pre-Transplant Survival Rates of Kidney-Transplant Patients Using Simulator-Based Propensity Scores", *Annals of Operations Research*, DOI 10.1007/s10479-013-1359-7, forthcoming.

Zidek, J. (1984), "Maximal Simpson-disaggregations of 2× 2 tables", *Biometrika*, 71(1), 187-190.

Appendix: Proof of Relationship between Type 4 Tree and Simpson's Paradox

We prove the correspondence between a Simpson's paradox and a Type 4 tree for the case of binary events (X and Y) with a single confounder Z. According to Alin (2010), Simpson's paradox is defined mathematically for three events and their complements: $Y = \{A, \bar{A}\}$, $X = \{B, \bar{B}\}$, and $Z = \{C, \bar{C}\}$, for which the following relationship holds:

$$\begin{aligned} &P(A|B) < P(A|\bar{B}), and \\ &P(A|BC) > P(A|\bar{B}C), and \\ &P(A|B\bar{C}) > P(A|\bar{B}\bar{C}). \end{aligned} \tag{1}$$

These inequalities can also be encountered in the form where the symbols $<$ and $>$ are reversed.

According to Blyth (1972), the probabilities $P(A|B)$ and $P(A|\bar{B})$ are equal to the following weighted averages, respectively:

$$\begin{aligned} P(A|B) &= P(C|B)P(A|BC) + P(\bar{C}|B)P(A|B\bar{C}) \\ P(A|\bar{B}) &= P(C|\bar{B})P(A|\bar{B}C) + P(\bar{C}|\bar{B})P(A|\bar{B}\bar{C}). \end{aligned} \tag{2}$$

Similarly, the probabilities $P(A|C)$ and $P(A|\bar{C})$ are equal to:

$$\begin{aligned} P(A|C) &= P(B|C)P(A|BC) + P(\bar{B}|C)P(A|\bar{B}C) \\ P(A|\bar{C}) &= P(B|\bar{C})P(A|B\bar{C}) + P(\bar{B}|\bar{C})P(A|\bar{B}\bar{C}). \end{aligned} \tag{3}$$

Simpson's paradox is only possible when B and C are dependent.

Corollary 1: The relationship $P(A|BC) > P(A|B)$ holds if and only if $P(A|B\bar{C}) < P(A|B)$ (and vice versa). Similarly, the relationship $P(A|\bar{B}C) > P(A|\bar{B})$ holds if and only if $P(A|\bar{B}\bar{C}) < P(A|\bar{B})$ (and vice versa).

Without loss of generality, let us assume that $P(A|B) < P(A|\bar{B})$, and $P(A|BC) > P(A|B)$.

Observation 1: if $P(A|BC) > P(A|B)$, then under Simpson's paradox $P(A|\bar{B}\bar{C}) < P(A|\bar{B})$ and $P(A|\bar{B}C) > P(A|\bar{B})$.

Proof of Observation 1:

1. Under the assumption and the definition of Simpson's paradox in Equation (1), $P(A|B) < P(A|\bar{B})$. This implies $P(A|BC) > P(A|\bar{B}C)$ and $P(A|B\bar{C}) > P(A|\bar{B}\bar{C})$.
2. Under the assumption, $P(A|BC) > P(A|B)$ and hence $P(A|B\bar{C}) < P(A|B)$.
3. Given (1) and (2), $P(A|\bar{B}\bar{C}) < P(A|B\bar{C}) < P(A|B) < P(A|\bar{B})$.
4. Given (3), $P(A|\bar{B}\bar{C}) < P(A|\bar{B})$, and hence $P(A|\bar{B}C) > P(A|\bar{B})$

5.1. CART: Impurity-Based Trees

Proof objective: we will show that the impurity of splitting the tree by Z is smaller than that of splitting the tree by X. Hence, in the presence of Simpson's paradox, an impurity based tree will necessarily be of type 4 (X appears in the tree, but is not as the top split).

Let us consider an impurity-based classification tree for the events $Y = \{A, \bar{A}\}$, $X = \{B, \bar{B}\}$, and $Z = \{C, \bar{C}\}$.

Definition 1: An impurity function ϕ is defined on all K-tuples of numbers $(P_1, ..., P_K)$ satisfying $P_j \geq 0$ for all values of j, and $\sum_j P_j = 1$. The properties of impurity functions are:

1. ϕ achieves its minimum when there exists j for which $P_j = 1$ and $P_i = 0$ for all $i \neq j$,
2. ϕ achieves its maximum only for the uniform distribution, that is, when all P_j are equal, and
3. ϕ is symmetric with respect to its arguments $(P_1, ..., P_K)$.

Definition 2: Given an impurity function ϕ, we define the impurity measure, denoted as I(T), of node T as follows:

$$I(T) = \phi(P(1|T), P(2|T), ..., P(K|T)) \tag{4}$$

The most commonly used impurity measures are Entropy and the Gini Index (see, e.g., **?**, Chap. 9.3), described in Equations (5) and (6), respectively.

$$Entropy(\cdot) = -\sum_{i=1}^{K} P_i \log_2(P_i) \tag{5}$$

$$GI(\cdot) = 1 - \sum_{i=i}^{K} P_i^2, \tag{6}$$

where K is the number of classes or events (in our case, the events of interest are $Y = \{A, \bar{A}\}$, and therefore $K = 2$)

For simplicity, we restrict our analysis to the case where $K = 2$. The analysis can be easily extended to cases where $K > 2$. We further abbreviate the notation in Equation (4):

$$I(T) = \phi(P(1|T)) \tag{7}$$

We note that for the case $K = 2$, an impurity function is *concave*, that is, the line segment between any two points on the graph of the impurity function lies below the graph. More formally, impurity function ϕ satisfies:

$$\phi(px_1 + (1-p)x_2) \geq p\phi(x_1) + (1-p)\phi(x_2) \tag{8}$$

for every $0 < p < 1$, and $x_1 \neq x_2$.

An impurity of a split in a classification or regression tree is defined by the weighted sum of the impurities of its children. Splits in the tree are then selected according to the ordering of their impurity (lowest impurity selected first).

Theorem 1: Given the condition in Equation (1) (Simpson's paradox) and a concave impurity function $\phi(p)$, the following relationship holds: $I(Z) \leq I(X) \leq I(Y)$.

Proof of Theorem 1: Let us consider the ordering of $P(A), P(A|B), P(A|\bar{B}), P(A|C)$, and $P(A|\bar{C})$.

First, observe that $P(A)$ can be rewritten as a weighted average of either $P(A|B)$ and $P(A|\bar{B})$, or $P(A|C)$ and $P(A|\bar{C})$:

$$\begin{aligned} P(A) =& P(B)P(A|B) + P(\bar{B})P(A|\bar{B}), \\ P(A) =& P(C)P(A|C) + P(\bar{C})P(A|\bar{C}). \end{aligned} \tag{9}$$

Second, from observation 1, we get $P(A|\bar{B}\bar{C}) < P(A|B\bar{C}) < P(A|B)$. Therefore, according to Blyth (1972):

$$P(A|\bar{C}) = P(B)P(A|B\bar{C}) + P(\bar{B})P(A|\bar{B}\bar{C}) < P(A|B). \tag{10}$$

Similarly, $P(A|BC) > P(A|\bar{B}C) > P(A|\bar{B})$, and therefore

$$P(A|C) = P(B)P(A|BC) + P(\bar{B})P(A|\bar{B}C) > P(A|\bar{B}). \tag{11}$$

Corollary 2: The ordering of $P(A), P(A|B), P(A|\bar{B}), P(A|C)$, and $P(A|\bar{C})$, is as follows:

$$P(A|\bar{C}) < P(A|B) < P(A) < P(A|\bar{B}) < P(A|C) \tag{12}$$

From concavity of the impurity function:

$$\begin{aligned} I(Y) &= \phi(P(A)) \\ &= \phi(P(B)P(A|B) + P(\bar{B})P(A|\bar{B})) \\ &\geq P(B)\phi(P(A|B)) + P(\bar{B})\phi(P(A|\bar{B})) \\ &= I(X) \end{aligned} \tag{13}$$

Let us now consider a new concave function $\widetilde{\phi}(\cdot)$, obtained by "trimming" the impurity function with a linear line drawn between $\{P(A|\bar{B}), \phi(P(A|\bar{B}))\}$ and $\{P(A|B), \phi(P(A|B))\}$, as illustrated in Figure 9, and given by:

$$\widetilde{\phi}(p) = \begin{cases} p\phi(p) + (1-p)\phi(1-p) & \text{if } P(A|B) \leq p \leq P(A|\bar{B}) \\ \widetilde{\phi}(p) = \phi(p) & \text{otherwise} \end{cases} \tag{14}$$

From concavity of the trimmed impurity function, we get:

$$\widetilde{\phi}(P(A)) = P(B)\phi(P(A|B)) + P(\bar{B})\phi(P(A|\bar{B})) = I(X) \tag{15}$$

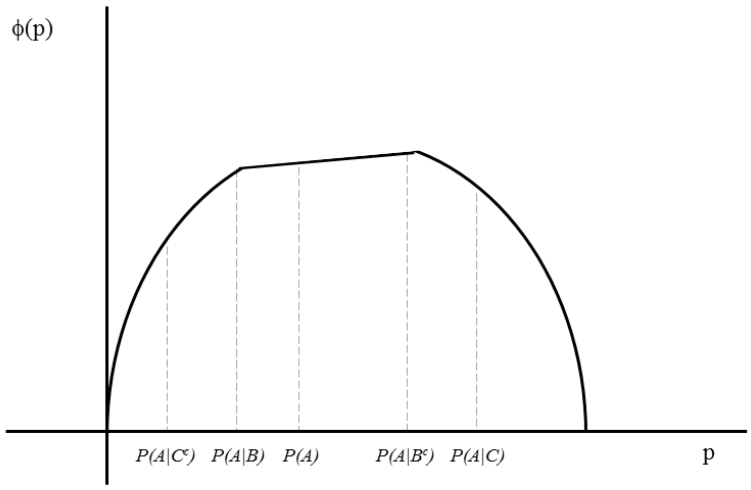

FIG 9. *Illustration of Concave Function* $\widetilde{I}(\cdot)$ *obtained by "trimming" the impurity Function with a Linear Line Drawn Between* $\{P(A \mid \bar{B}), \phi(P(A \mid \bar{B}))\}$ *and* $\{P(A \mid B), \phi(P(A \mid B))\}$

and

$$\begin{aligned} \widetilde{\phi}(P(A)) &\geq P(C)\widetilde{\phi}(P(A|C)) + P(\bar{C})\widetilde{\phi}(P(A|\bar{C})) \\ &= P(C)\phi(P(A|C)) + P(\bar{C})\phi(P(A|\bar{C})) \\ &= I(Z) \end{aligned} \tag{16}$$

Hence

$$\Longrightarrow I(Z) \leq I(X) \leq I(Y) \tag{17}$$

The proof is illustrated for the Entropy function in Figure 10.

5.2. Conditional Inference Trees: χ^2 Statistic-Based Trees

Proof objective: we will show that the p-value of splitting the tree by Z is smaller than that of splitting the tree by X. Hence, in the presence of Simpson's paradox, a χ^2 statistic-based tree will be of type 4 (X appears in the tree, but not as the top split).

Let us consider the 2×2 contingency tables of splitting Y by either X or Z (top split in the tree). The contingencies tables are given in Table 18.

The χ^2 statistics of splitting Y by X can be computed by (Pavlides and Perlman (2009), Ch. 15, pp. 342-352):

$$\chi^2_X = \frac{(p(AB)p(\bar{A}\bar{B}) - p(\bar{A}B)p(A\bar{B}))^2 \times N}{p(A)p(\bar{A})p(B)p(\bar{B})}, \tag{18}$$

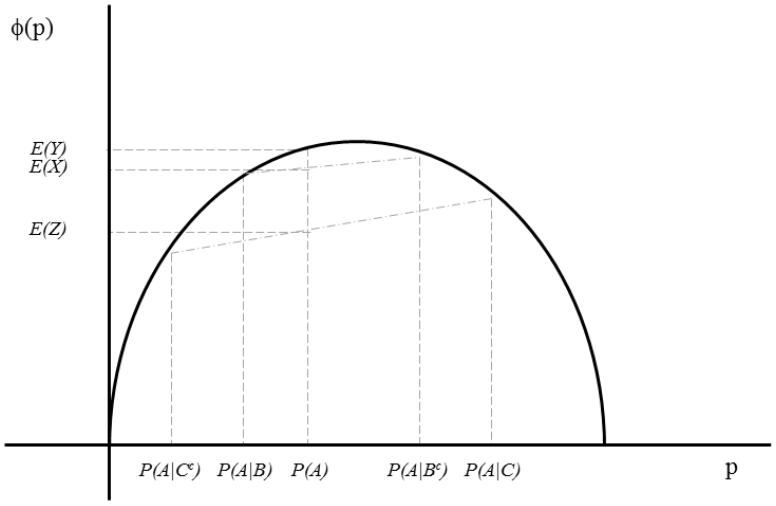

FIG 10. *Illustration of the proof for the Entropy function.*

TABLE 18
Contingency Table for Splitting Y by X (left) and Z (right)

	A	$\bar{A}$	**Total**
B	$p(AB)$	$p(\bar{A}B)$	$p(B)$
$\bar{B}$	$p(A\bar{B})$	$p(\bar{A}\bar{B})$	$p(\bar{B})$
Total	$p(A)$	$p(A)$	1

	A	$\bar{A}$	**Total**
C	$p(AC)$	$p(\bar{A}C)$	$p(C)$
$\bar{C}$	$p(A\bar{C})$	$p(\bar{A}\bar{C})$	$p(\bar{C})$
Total	$p(A)$	$p(A)$	1

Where N is the number of observations.

Using Bayes' rule, we obtain the following relationship:

$$P(AB) = P(A|B)P(B). \tag{19}$$

similar terms can be obtained for $p(\bar{A}\bar{B}), p(\bar{A}B)$ and $p(A\bar{B})$.

Therefore, the value of χ^2 can be rewritten as:

$$\begin{aligned}
\chi^2_X &= \frac{p(B)^2 p(\bar{B})^2 (p(A|B)p(\bar{A}|\bar{B}) - p(\bar{A}|B)p(A|\bar{B}))^2}{p(A)p(\bar{A})p(B)p(\bar{B})} \times N \\
&= \frac{p(B)p(\bar{B})(p(A|B)p(\bar{A}|\bar{B}) - p(\bar{A}|B)p(A|\bar{B}))^2}{p(A)p(\bar{A})} \times N \\
&= \frac{p(B)p(\bar{B})(p(A|B)(1-p(A|\bar{B})) - (1-p(A|B))p(A|\bar{B}))^2}{p(A)p(\bar{A})} \times N \\
&= \frac{p(B)p(\bar{B})(p(A|B) - p(A|\bar{B}))^2}{p(A)p(\bar{A})} \times N
\end{aligned} \tag{20}$$

We next compute the term $p(B)p(\bar{B})$ from Equation (20). Rewriting $p(A)$ as a weighted sum of $P(A|B)$ and $P(A|\bar{B})$, we can write $p(B)$ as a function of $p(A), p(A|B)$, and $p(A|\bar{B})$:

$$\begin{aligned}
P(A) &= P(B)P(A|B) + P(\bar{B})P(A|\bar{B}) \\
&= P(B)P(A|B) + (1 - P(B))P(A|\bar{B}) \\
&= P(B)P(A|B) + P(A|\bar{B}) - P(B)P(A|\bar{B}) \\
\Rightarrow p(B) &= \frac{p(A) - p(A|\bar{B})}{p(A|B) - p(A|\bar{B})}
\end{aligned} \tag{21}$$

Similarly, $p(\bar{B})$ equals to:

$$p(\bar{B}) = \frac{p(A) - p(A|B)}{p(A|\bar{B}) - p(A|B)} \tag{22}$$

Given the terms in Equations (21) and (22), the term $p(B)p(\bar{B})$ can be written as:

$$p(B)p(\bar{B}) = -\frac{(p(A) - p(A|B))(p(A) - p(A|\bar{B}))}{(p(A|B) - p(A|\bar{B}))^2} \tag{23}$$

Plugging the term in Equation (23) back into Equation (20) we get:

$$\chi^2_X = \frac{(p(A) - p(A|B))(p(A|\bar{B}) - p(A))}{p(A)p(\bar{A})} \times N \tag{24}$$

Finally, we compare the χ^2 statistics of splitting event Y by X and splitting event Y by Z. Essentially, the terms compared are the products absolute distance of the relevant conditional probabilities ($p(A|B)$ and $p(A|\bar{B})$ for computing χ^2_X statistic, $p(A|C)$ and $p(A|\bar{C})$ for computing χ^2_Z statistic) and from

the probability of the event $Y = A$ (see Equation (24)). Following Corollary 2 (Equation 12), we get

$$\chi^2_X \leq \chi^2_Z \tag{25}$$

and hence,

$$\Rightarrow p\_value(\text{splitting Y by X}) \geq p\_value(\text{splitting Y by Z}) \tag{26}$$