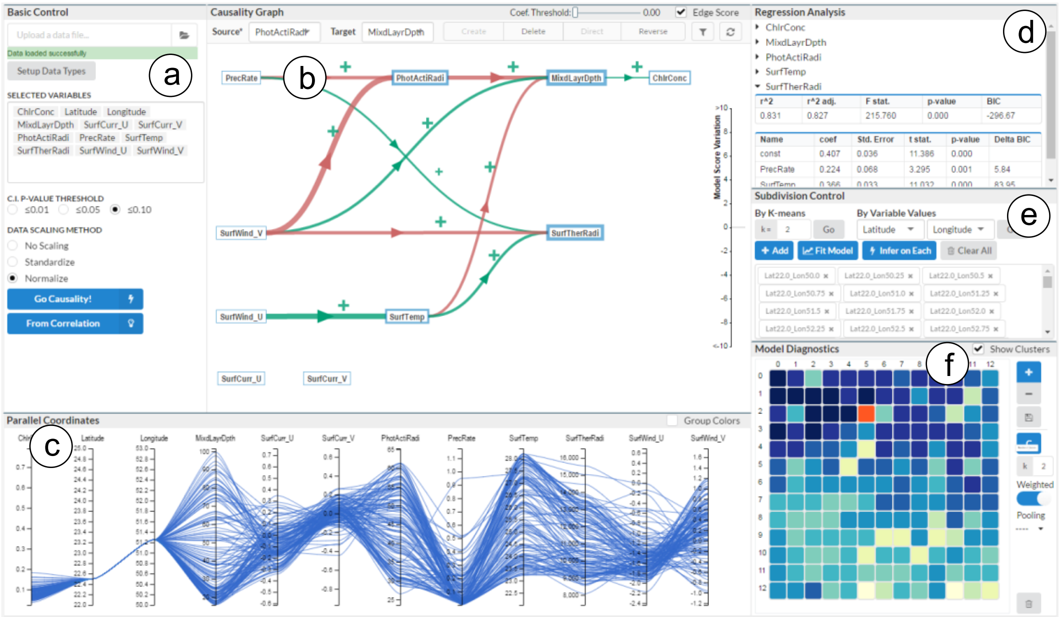
Proceedings of IEEE Visual Analytics Science and Technology (VAST), Phoenix, AZ, October 2017

Visual Causality Analysis Made Practical

Jun Wang and Klaus Mueller*

Visual Analytics and Imaging Lab, Computer Science Department, Stony Brook University



SELECTED VARIABLES: ChlrConc, Latitude, Longitude, MixdLayrDpth, SurfCurr_U, SurfCurr_V, PhotActiRadi, PrecRate, SurfTemp, SurfTherRadi, SurfWind_U, SurfWind_V

C.I. P-VALUE THRESHOLD: ○ ≤0.01 ○ ≤0.05 ● ≤0.10

DATA SCALING METHOD: ○ No Scaling ○ Standardize ● Normalize

Go Causality!

From Correlation

Regression Analysis: ChlrConc, MixdLayrDpth, PhotActiRadi, SurfTemp, SurfTherRadi

r^2	r^2 adj.	F stat.	p-value	BIC
0.831	0.827	215.760	0.000	-296.67

Name	coef	Std. Error	t stat.	p-value	Delta BIC
const	0.407	0.036	11.386	0.000	
PrecRate	0.224	0.068	3.295	0.001	5.84
SurfTemp	0.366	0.033	11.032	0.000	83.95

Subdivision Control: By K-means k= 2 Go; By Variable Values: Latitude, Longitude

+ Add | Fit Model | Infer on Each | Clear All

Lat22.0_Lon50.0, Lat22.0_Lon50.25, Lat22.0_Lon50.5, Lat22.0_Lon50.75, Lat22.0_Lon51.0, Lat22.0_Lon51.25, Lat22.0_Lon51.5, Lat22.0_Lon51.75, Lat22.0_Lon52.0, Lat22.0_Lon52.25, Lat22.0_Lon52.5, Lat22.0_Lon52.75

Fig. 1 The Causal Structure Investigator interface (a) Control panel for reading in data and setting inference parameters. (b) Interactive path diagrams for causal network visualization. (c) Parallel coordinates view for exploring data partitions. (d) Statistic coefficients tables of regressions associated with the causal model. (e) Data subdivision control, where a subdivision can be saved as a clickable tag. (f) Model diagnostic controls and the model heatmap, where users can examine learned models by clicking each tile colored by model scores.

Abstract

Deriving the exact casual model that governs the relations between variables in a multidimensional dataset is difficult in practice. It is because causal inference algorithms by themselves typically cannot encode an adequate amount of domain knowledge to break all ties. Visual analytic approaches are considered a feasible alternative to fully automated methods. However, their application in real-world scenarios can be tedious. This paper focuses on these practical aspects of visual causality analysis. The most imperative of these aspects is posed by Simpson' Paradox. It implies the existence of multiple causal models differing in both structure and parameter depending on how the data is subdivided. We propose a comprehensive interface that engages human experts in identifying these subdivisions and allowing them to establish the corresponding causal models via a rich set of interactive facilities. Other features of our interface include: (1) a new causal network visualization that emphasizes the flow of causal dependencies, (2) a model scoring mechanism with visual hints for interactive model refinement, and (3) flexible approaches for handling heterogeneous data. Various real-world data examples are given.

Keywords: Visual knowledge discovery, Causality, Hypothesis testing, Visual evidence, High-dimensional data

1 Introduction

The urge to find the causal explanations behind one or more observed phenomena is an inherent trait of human nature, and the massive growth of data can help satisfy this innate curiosity. While correlation has been widely used as evidence of causation, relations derived in this way can be ambiguous and often even spurious (A great many of such examples can be found at the website of spurious-correlations [1]). What is needed is a dedicated causality framework capable of measuring the dependency between two variables in the context of another set of controlled variables. While a number of algorithms have been devised for identifying causal relation in multivariate data, these algorithms typically cannot encode existing domain knowledge, or even common sense, to guide their analyses. This, in turn, leads them to hold strong assumptions on data distributions which can rarely be satisfied in practice. A remedy to overcome this significant shortcoming is to insert a human into the casual inference loop as a synergist partner.

This realization has led to efforts that use a visual analytics approach to casual inference, called *visual causality analysis*. It allows human experts endowed with domain knowledge and intuition to refute or propose causal links. We proposed a prototype of such an interface in an earlier paper [2], called the *Visual Causality Analyst*. It utilizes a 2D graph visualization of causal networks and a set of interactive tools that users can employ to examine the derived relations. While effective, this interface is nevertheless relatively simple and can only provide very basic functions of operating on a single model. Real world scenarios, however, incur many practical difficulties that such a simple tool cannot handle.

* Email: {junwang2, mueller}@cs.stonybrook.edu

The greatest practical challenge is posed by *Simpson's Paradox* [3] which states that a relation held in the general population may be altered in data sub-groups given proper partitions. A widely-used example for this phenomenon is the 1973 discovery of an apparent gender bias favoring male applicants in the graduate school admissions at UC Berkeley [4]. However, in fact, the gender bias was reversed when each department was considered separately – 6/85 departments appeared to favor females while only 4/85 appeared to favor males. This discrepancy was not deliberating but explainable by unrelated admission facts. For causality analysis, Simpson's paradox implies that possibly multiple causal models underlie a dataset, each for a certain subrange of the data across the factors. We propose a new set of tools to help analysts recognize where such decompositions might be appropriate and allow them to subdivide the data along certain dimensions or into clusters. In addition, we also provide facilities that allows analysts to compare between and extract credible relations from the derived multiple causal models via a pooling process that can either occur at the causal link level or at the model level.

Another challenge is that real-world problems often have a mix of numerical and categorical (ordinal, nominal) data. This stands at odds with current causality algorithms which can only handle either numerical or categorical variable, but not both. To make the data homogeneous, we can either bin all numeric variables into categorical ones, or use our method [2] which transformed the categorical variables into numerical ones using a global re-spacing and re-ordering scheme. The problem with this scheme was that the distribution of the levels remains to be sparse which adds complexity to the casual inferencing. We propose a novel level-enrichment scheme that absolves this problem, and along with it, we also devise a set of generalized inference algorithms with flexible options for handling heterogeneous data.

Finally, causal models are often drawn in form of general directed networks and graphs in which flows of causal dependencies are hard to recognize. This also impedes the practical use of causality analysis as an analytics platform for general use. We have devised a new and more appropriate visualization of causal networks in form of path diagrams laid out using spanning trees. We find that these path diagrams give causal flows an effective narrative structure.

In general, the major contributions of this paper include:

- A new visualization of causal networks that better exposes the flow of causal sequences;
- A scoring function along with corresponding visual hints that can be used to compare alternative causal models;
- An improved method for handling heterogeneous data in causal inference along with their experimental evaluation;
- Interactive facilities that allow users to explore data sub-divisions from which different models can be inferred;
- Mechanisms for diagnosing (or pooling) all derived models to recognize valuable causal relations and patterns.

And all of these techniques have been implemented into a novel visual interface we call the *Causal Structure Investigator* (CSI). The teaser image in Fig. 1 shows its individual components.

Our paper is structured as follows. Section 2 discusses related work. Section 3 briefly introduces the interface components and then presents detailed techniques used in the CSI framework for visual analysis of a single model. Section 4 discusses the impact of data partition on causal inference. Then two case studies, respectively, are presented in Section 5. Finally, Section 6 ends with conclusions and future work.

2 Related Work

Following the seminal work of Pearl [5][6] and Spirtes [7], theories of causation modeling and discovery on multivariate datasets have been widely studied. Visual causality analysis has also become a popular topic in the field of visual analytics (VA) in recent years.

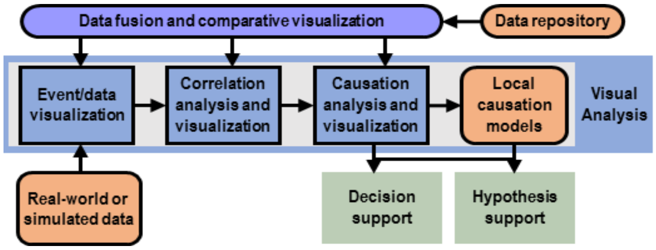

Data fusion and comparative visualization
Data repository
Event/data visualization
Correlation analysis and visualization
Causation analysis and visualization
Local causation models
Visual Analysis
Real-world or simulated data
Decision support
Hypothesis support

Fig. 2 The workflow of visual causality analysis by Chen et al. [21].

2.1 Causality Modeling and Inference

The set of causal relations between variables of a multidimensional dataset is usually depicted as a Directed Acyclic Graph (DAG) where variables are nodes and a directed edge between two nodes means the first causes the second. Algorithms learning the structure of such DAGs can be roughly classified into two categories – score-based algorithms and constraint-based algorithms. The former typically associate a DAG with a score function, e.g. the Bayesian Information Criterion (BIC) [8][9], and performs, for instance, a greedy search in the space of all possible DAGs. Examples are the GES algorithm [10] and the K2 algorithm [11]. Since the number of possible structures is super-exponential in the number of variables, such algorithms usually suffer from high search cost. In contrast, the constraint-based algorithms build causal networks according to the constraints of dependencies and conditional dependencies in the data. Some well-known algorithms are SGS[7], PC [7][12], IC [13], Total Conditioning [14], and others. These constraints are usually learned with conditional independence (CI) tests via partial correlation [15], G^2 statistics [16], or other techniques [17][18]. It is important to note that such algorithms are commonly based on several strong assumptions of data distributions which are rarely satisfied by real-world data. As a consequence, none can guarantee an exact model, especially when there are latent or nonlinearly related variables.

Several causal modeling methods can be used to parameterize the learned DAG. The two most common choices are Bayesian Networks (BN) [5][19] and Structural Causal Models (SCM) [6][20]. The former quantifies causal relations with conditional probability tables, and the latter with linear functions plus Gaussian noise, e.g. linear regression and logistic regressions. As the knowledge of data distribution required in BN is usually hard to acquire in practice, we will use the algorithm of Total conditioning and PC in this paper to infer causal structures and then parameterize them as SCM models.

2.2 Visual Causality Analysis

Fig. 2 pictures the workflow of visual causality analysis proposed by Chen et al. [21], aiming to provide decision support in a typical organization and aid hypothesis generation and evaluation in a scientific investigation. One of the earliest attempts of such a system is the *Growing-polygons* [22] scheme which captures causation at the *process* level, i.e. as a sequence of causal events. It uses animated polygon colors and sizes to signify causal semantics. The work of Vigueras and Botia [23] considers ordered events in a distributed system as causations and visualizes their dependencies as causal graphs. Focusing on the upstream-downstream relations of variables, *ReactFlow* [24] visualizes causal relations as pairwise pathways connecting duplicated variables in two columns. Some other efforts in the visual mining of causation include *OutFlow* [25] and *EventFlow* [26]. Both visualize temporal event sequences as alternative pathways and use event chains to explore embedded patterns. Liu et al. [27] visualize event streams as flows aligned by event types. However, none of these above systems leverages automated algorithms for causal discovery, and so they require significant user input to acquire such knowledge.

The first visual interface with the capability of automatic causal inference was proposed by us in previous work [2]. It visualizes causal networks as color-coded 2D graphs with force-directed layouts and offers a set of interactive tools for the user to examine the derived relations. The graph visualization we employed in this previous work has also been widely used in visualizing Bayesian belief networks [28], correlation networks [29], uncertainty networks [30], and many other graph-based analytic models [31][32]. The work in this paper is inspired by these methods but will provide a much-improved visualization and more comprehensive analytic capabilities that can handle many practical difficulties in real-world causality analysis.

3 Visual Inference of Single Causal Model

The design of the CSI interface (Fig. 1) fulfills the requirements of a causality VA system proposed by Chen et al. [21] (Fig. 2). More specifically, the parallel coordinates view (Fig. 1c) serves as the component for data visualization. Users have the option to start from either a causality model or a correlation graph (Fig. 1a). The path diagram view (Fig. 1b) and the regression analysis view (Fig. 1d) then allow the visual analysis of both causation and correlation. The analytics on local causation models are achieved through the data subdivision view (Fig. 1e) and the model heatmap (Fig. 1f), with which user can visually examine each model derived from a data subdivision as well as the pooled models, getting full support for decision making and hypothesis evaluation.

In this section, we will describe the various features of our framework in terms of a single model, which serves two major purposes: (1) communicate the automatically derived relations for the causal network and (2) allow users to examine their own proposed causal links as well as ones derived by algorithms. The next section will then expand it to analyze multiple models arising from data subdivisions.

3.1 Causal networks visualization

As mentioned, although force-directed graphs could be a feasible choice for demonstrating the overall structure of the network, they often suffer from a dense and unpredictable layout. With such layouts, local structures in causal sequences can become difficult to observe especially when they are part of more complex networks. However, these local structures can often be of great interest to domain users. For instance, Dang et al. [24] show that recognizing the upstream and downstream causal relation of variables is commonly required by biologists when examining relations between proteins and biochemical reactions. While their work succeeds in visualizing local causal relations as pairwise pathways, it is less successful in conveying global structures of the network.

We aimed to create a framework that would convey both local causal sequences as well as the overall network structure. For this, we devised a new approach that visualizes causal networks as path diagrams. In a causal path diagram, a causal relation is visualized as a straight or curved path from the cause to the effect variable denoted by named nodes. Such design is inspired by previous works using pathways to represent relation or event flows [24][25]. The arrow mark in the middle of a path signals the direction of the relation. To remit the clutter of local structures, i.e. sequences of causal relations, the path diagram is laid out using spanning trees of the network built with *Breadth-first Search*. More specifically, we first layout the nodes of the spanning trees to fit the canvas in a left-to-right manner regarding their parent-child relations, and then add back all edges during rendering. Variables not related to others shall be isolated at the bottom. By such, most paths of causal sequences will connect and direct from left to right, intuitively forming causal stories. Finally, although the generated diagrams are usually clear enough for demonstrating the causal paths, users are also allowed to adjust it manually by dragging each node.

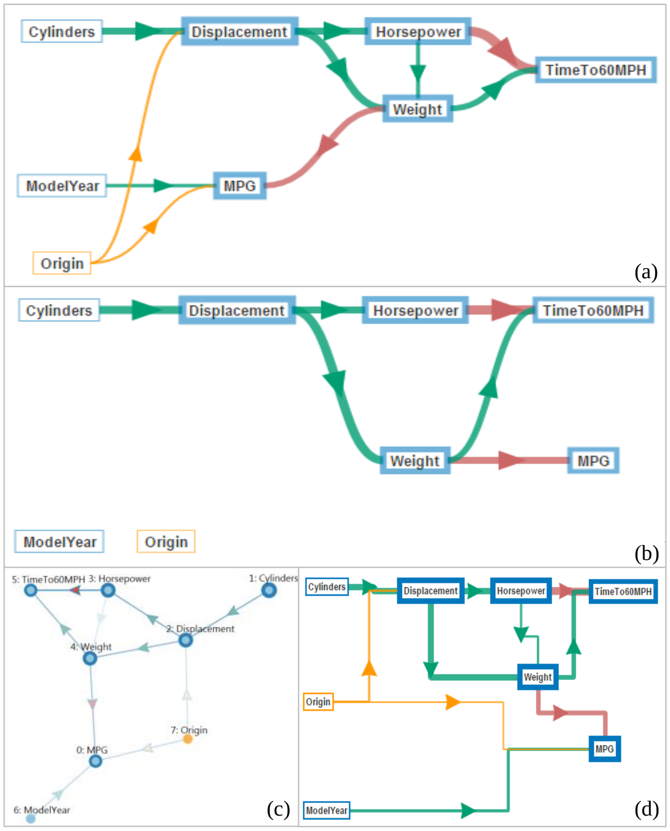


Fig. 3 Visualization of the causal network derived from the AutoMPG dataset [34]. (a) The path diagram visualization of the network. (b) The path diagram after setting an edge coefficient threshold of 0.3. (c) Visualization of the network as a force-directed graph from our earlier work [2]. (d) An orthogonal graph visualization of the network.

Besides the directional structure, parameterized relations also come with a set of statistical coefficients quantitatively measuring their strengths and significances. In our interface, the width of a path signifies the strength of the relation measured by linear (targeting numeric variables) or logistic (targeting categorical variables) regression coefficients. Using the color code for causal semantics we proposed in [2], green paths denote positive causal influence and red paths denote a negative influence. Compound relations between levels of categorical variables and other variables are colored yellow. Node colors indicate variable type – blue for numeric and yellow for categorical. A node’s border thickness suggests the goodness of fit of the variable’s regression model measured by r-squared (for linear regression) or McFadden’s pseudo r-squared (for logistic regression) coefficients [33], both have a value range of 0 to 1.

Fig. 3a shows a first application, using the causal network learned from the AutoMPG dataset [34]. We can observe that nodes are mostly positioned left to right in topological order following their dependencies. The flow of causations, especially those with strong relations, become even clearer after weak relations (narrow paths) have been filtered out (which is a function included in the CSI interface). For example, Fig. 3b shows the same network with a coefficient (path width) threshold of 0.3. Here we can observe several causal paths flowing from left to right. One of them is *Cylinder→Displacement→ Weight→MPG*, which indicates that it is weight rather than the size of the engine that is directly affecting a car's gas mileage. This can be a useful finding for a car company which now knows that it can counter-balance the adverse effect a big engine has on mpg by designing a car with a lighter chassis but designed for increased structural stability.

The force-directed graph used in our earlier work [2] is shown in Fig. 3c and an example for an orthogonal graph is shown in Fig. 3d where nodes are connected by orthogonal edges. Both demonstrate the AutoMPG network to facilitate a fair comparison. Compared to these two methods, we believe that our new path diagram exposes flow of causal sequences embedded in the network in a much more prominent way than the two competing methods. Future work will compare the three methods in a formal setting.

3.2 Visual Model Refinement with Model Scoring

According to Fig. 2, one of the major tasks of visual causality analysis is to provide visual evidence supporting a user's decision on refuting or accepting causal relations. This can be achieved by scoring each relation as well as the overall network with proper metrics. Although common statistics calculated from regression residuals, e.g. F-statistics and r-squared, are capable of measuring the model goodness of fit, they usually do not take model complexity into consideration. This implies that just by adding more relations into the model these statistics will mostly improve. However, this can potentially lead to overfitting, which means that the model is an extremely good fit for the dataset from which it was learned, but generates huge errors on any other dataset recorded from the same source. Hence, based on William of Occam's parsimony principle, models should be kept as simple as possible. The idea is that by adding new relations to a causal model we obtain an improvement in its fit to the data to some degree, but at the same time the model also becomes "worse" because it is harder to fit new data. So, the question is how complex should the model be for a given dataset.

The Bayesian Information Criterion (BIC) [8][9], applicable to both linear and logistic regressions, serves well in answering this question. It rewards the improvement in fit but also punishes for increasing model complexity. For a single regression model, it is formulated as

$$\mathrm{BIC} = -2\ln\hat{L} + k\ln(n) \qquad (1)$$

where $\hat{L}$ is the likelihood of the model, k is the number of independent variables, and n is the number of data points. The BIC of a linear regression can be computed from residuals following

$$\mathrm{BIC} = n\ln RSS/n + k\ln(n) \qquad (2)$$

where the *residual sum of squares* $RSS = \sum(y_i - \hat{y})^2$, in which $\hat{y}$ is the predicted value of the dependent variable given values of independent variables in a regression equation, and y_i is the actual observed value of the dependent variable. The likelihood of logistic regressions can be computed directly using logistic functions. Eq. 2 also suggests that a smaller BIC score with small residuals and less parameters implies a better regression model.

For each variable in a causal network, k in Eq. 2 is the number of incoming directed edges. Variables with no observed cause can be fitted with a null model (with only the error term, thus $k = 0$). As such, a causal edge is preferable only when it reduces the error term of the first part of Eq. 2 more than it increases the complexity term of the second part of the equation, i.e. it reduces the regression's BIC. Further, as suggested by Kass and Raftery [35], the difference of a regression's BIC with and without a certain independent variable can be interpreted qualitatively following Table 1. According to the table, if adding a causal edge causes the BIC of the regression model to be reduced by more than 10 points, the resulting model can be deemed as "very strongly" better and the edge should be favored.

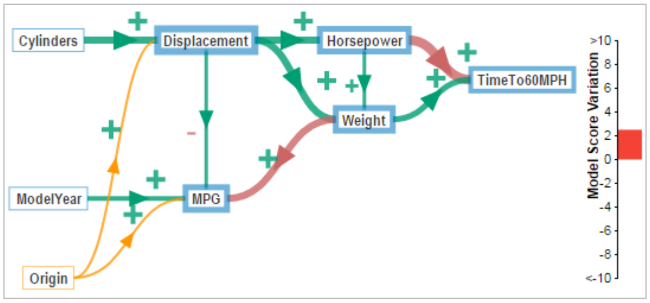


Fig. 4 The path diagram with model scores visualizing the AutoMPG network. A new relation from *Displacement* to *MPG* is added. However, the red minus glyph next to it and the red score bar on the right show that the relation is not valid and so should be removed.

Table 1 Qualitative interpretation of BIC score difference. Here $\boldsymbol{p}$ is a regression model with one extra independent variable added to $\boldsymbol{q}$.

$\lvert \mathbf{BIC}_p - \mathbf{BIC}_q \rvert$	**Evidence Against Model *q***
0 to 2	Not worth more than a bare mention
2 to 6	Positive
6 to 10	Strong
>10	Very Strong

Based on this fact, an automated analysis process can be applied whenever the DAG is parameterized by regressions. Since each node implies a variable regressed on its causes linked by all the incoming edges, we assign each edge a level of importance by calculating the regression's BIC change when the edge is removed while keeping all other causes. If the BIC score goes up after removing it, the edge should be recognized as valid and a green plus glyph is attached to it in the path diagram (Fig. 4). Otherwise, it is considered doubtful and a red minus glyph is placed. The size of the glyph encodes how much the score would change such that bigger glyphs indicate larger score changes. However, since changes larger than 10 points can all be classified into the "very strong" category, the maximum glyph size can be correspondingly fixed. As such, good causal relations, as well as false ones suggested by the data, can be visually recognized.

The sum of all the BIC calculated from these regressions can be used as the score of the overall causal network g, which is

$$F(g) = \sum_i BIC_i \qquad (3)$$

where BIC_i is the BIC of the regression model on variable v_i. Such a scoring strategy has also been adopted by many score-based inference algorithms to score potential causal structures [10][11].

Based on the model score, a colored bar is rendered whenever the user modifies the network, showing the impact of the modification on the overall model. A red bar means the overall model score is rising and a green bar stands for a score decreasing. The length of the bar encodes by how much the score has changed. With these visual hints, users can be intimately aware if they have made a good move in their quest of refining the model under study.

Fig. 4 illustrates an example where we added a path from *Displacement* to *MPG* to the causal network of Fig. 3a. While most relations are valid according to the green plus glyphs, the red minus next to the newly added edge indicates that it is increasing the BIC score of the regression of *MPG*, thus increasing the total model

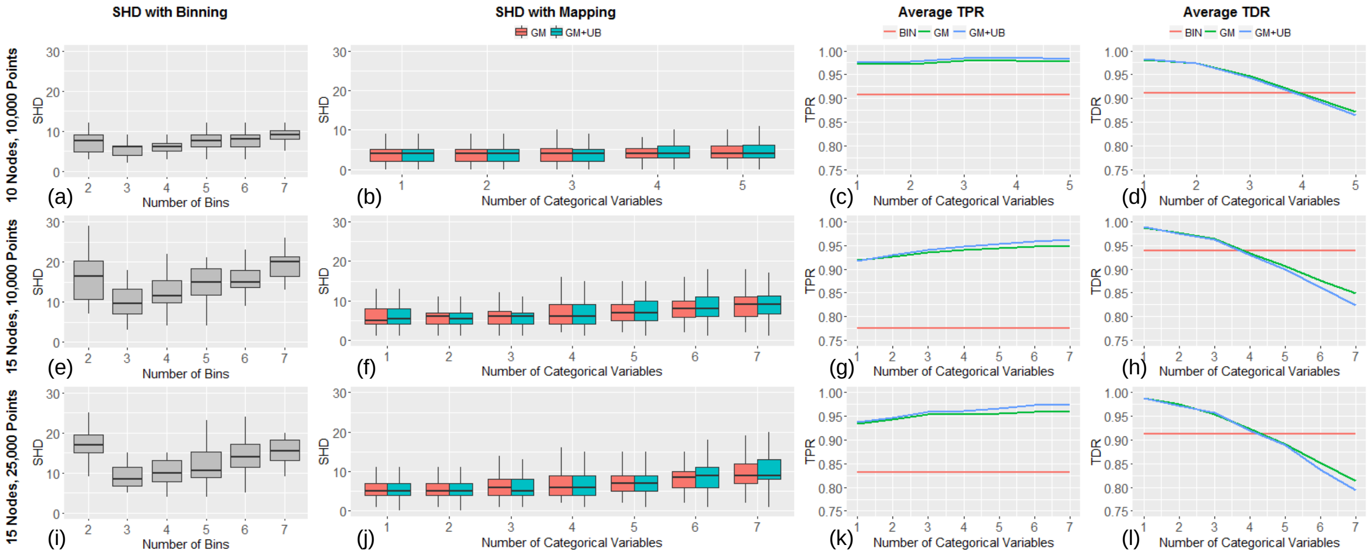


Fig. 5 Experimental evaluation of the impact of GM with/without UB in the causal inference of heterogeneous data, comparing to the strategy of simply binning. Charts in each row are from experiments running on the same simulated dataset. Charts in each column visualize the same metric. (a), (e) and (i) are the SHDs of rebuilt causal networks by binning numeric variables with different levels. (b), (f), and (j) are the SHDs from GM and GM+UB with different numbers of categorical variables included in the dataset. (c), (g), and (k) show the average TPR and (d), (h) and (l) show the average TDR of the reconstructed networks with the three strategies under different numbers of categorical variables.

score. The score bar shows the model score changed about 2 points ("Positive" according to Table 1), so it is suggested to be removed.

It is worth noting that the Akaike information criterion (AIC) [8], which is defined very similar to BIC but with a less stringent punishment for model complexity, is also a widely applied scoring strategy used in model selection. While the AIC can work exactly the same function as BIC and might be preferred in some circumstances, we choose BIC in our implementation as it is more often adopted in causality studies and emphasis more on solving the issue of overfitting [36].

3.3 Working with Heterogeneous Data

As mentioned, heterogeneous data containing both numeric and categorical variables are problematic when learning the structure of a causal DAG. It requires a CI test method capable of testing and conditioning on variables of arbitrary distributions. However, typical CI tests using partial correlation or the G^2 test can only handle either numeric or categorical data, and none can handle both. Simply binning all numeric variables and applying the G^2 test can be a plausible solution but it comes at the potential price of a significant information loss. With this approach, not only is there a loss in value scales, but also the order of bins will be ignored in the G^2 tests, both of which can introduce error relations in the result.

Another recently proposed solution is the Global Mapping (GM) strategy (see our earlier paper [2]), which re-orders and re-spaces categorical variables' levels so that Pearson's correlations involving categorical variables are generally maximized with respect to all numeric variables in the dataset. This allows the CI test via partial correlation to be applied to all, which also means a faster inference process since the G^2 test usually takes much longer. More specifically, the GM strategy assigns values to level j of categorical variable v_c according to the following formula:

$$v_c(j) \propto \sum_{i=1}^{D} \Theta_i \rho_i \mu\big(v_i(j)\big) \qquad (4)$$

where $\mu(v_i(j))$ is the average of numeric variable v_i corresponding to level j of v_c, ρ_i is the maximized Pearson's correlation between v_i and v_c, and Θ_i decides the sign of ρ_i by comparing the level orders of v_c regarding v_i and regarding the numeric variable most correlated with v_c, supposing there are D numeric variables in total.

A shortcoming of GM is that the mapped values are still discrete while CI tests via partial correlation assume they are continuous. To ease this issue, we add an *un-binning* (UB) process after GM in which mapped levels are converted to value ranges separated by the middle point of two levels. For example, if a three-level variable is mapped to values {0, 0.4, 1}, the converted ranges shall be {[-0.2, 0.2], [0.2, 0.7], [0.7, 1.3]}. Then data points are randomly assigned with values in the according range based on a Gaussian distribution. By such, categorical variables can be simulated to be continuous.

3.3.1 Experimental Evaluation

We evaluate the effectiveness of the GM with and without UB via three runs of experiments, comparing them to the strategy of equal-width binning of all numeric data. We use 100 randomly generated DAGs in each run as ground truth. A DAG has 10 nodes in the first run and 15 nodes in the second and third runs. A node in a DAG has a 0.2 probability to connect to any other nodes. Coefficients of graph edges are uniformly distributed within the range [0.1, 1], based on which 10000 data points are sampled for each DAG in the first two runs and 25000 in the third run. Some randomly selected variables are then converted into categorical ones in each run with equal-width binning. The three aforementioned strategies applied with the PC-stable algorithm [12] are tested under each setting, trying to reconstruct simulated DAGs from the sampled mixed-type data. All experiments were done with the R package *pcalg* [37].

The charts in each row of Fig.5 show the results of each run. The charts in the left most column of Fig. 5 (a, e, and i) visualize the *Structure Hamming Distance* (SHD) error of the causal models inferred with binning all variables into 2 to 7 levels, respectively. The SHD is defined as the minimum number of edge insertions, deletions, directions, and reversions needed to transform the estimated graph into the ground truth. In SHD, the deletion or the direction of an undirected edge is each counted as one error, while it counts as two errors if a directed edge needs to be reversed. In each of the three charts, we observe that the SHD increases both when there are too few levels (equivalent to a loss of value scale) as well as when there too many (ignorance of bin order). We also observe that the error increases when reconstructing a larger network (comparing Fig. 5a and e), but it drops when more data is available (Fig. 5e and i).

The charts in the second column of Fig. 5 (b, f, j) demonstrate the SHD from GM (red boxes) and GM+UB (blue boxes) under the situation that at most 50% of variables are categorical. While the error increases when more categorical variables are introduced, both of the two strategies outperform the best case from binning in all three runs (compare Fig. 5a, e, and i). A deeper inspection is offered when looking at charts in the right two columns of Fig. 5 (c, d, g, h, k, and l), which shows the average *True Positive Rate* (TPR, the number of correct edges out of ground truth edges) and *True Discovery Rate* (TDR, the number of correct edges out of all found edges) of the results. Edge directions are omitted here. We learn from Fig. 5c, g, and k that GM+UB (blue line) generally shows a better TPR than GM (green line), which means more correct relations are discovered. However, when looking at Fig. 5d, h, and l, the TDR from GM+UB drops much faster than the pure GM when there are more than 4 categorical variables in the first two runs and 5 in the third run, which means many error relations are falsely linked too. Also, both GM strategies tend to introduce more spurious relations than binning with more categorical variables in the dataset. We suspect that when the ratio of categorical variables is too large, the global re-ordering and re-spacing can no longer preserve the fidelity of the data.

Taking all of the experiment results into consideration, we suggest users take the GM strategy whenever no more than 30% of the variables in a dataset are categorical, while UB can further boost the inference accuracy. When there are more categorical variables, binning numeric variables could be a more plausible choice. Finally, we would like to stress that the strategy is only applied when learning the structure of causal networks. Conversely, in the subsequent parameterization, the original levels of the categorical variables are used as they can be well handled by logistic regressions. Our GUI allows users to choose any of the three methods when working with heterogeneous datasets.

4 Causality Analysis with Multiple Models

As mentioned, our framework also supports the visual investigation of multiple causal models underlying a dataset. We now present details of this mechanism, along with illustrative examples.

4.1 Causal Inference on Data Subdivisions

According to Simpson's Paradox, a relation found in the overall data may not hold in certain data subdivisions, and conflicting relations buried in some specific data ranges may cancel each other so that none can be observed in the general population. Such effect has often been observed in correlation analysis [29]. For example, by bracketing the price of a product to lower ranges one may see positive correlations with sales, while negative correlations come with a higher price range. What's more, causal relations with opposite directions may also exist as feedback loops. For instance, the price of a product will affect sales when sales are low, but a large number of sales can also reduce the cost and so lower the price. As a result, it is often the case that multiple causal models differing in both structure and regression parameters can arise from data partitions. Ignoring such facts and always learning the model using the whole dataset will potentially lead to faulty relations returned by inference algorithms. Without data partitioning, the regression model constructed will probably contain considerable large residuals. Seeing that the BIC of a model is computed from such residuals (Eq. 2), refining these miscalculated causal models based on their score change can also be difficult in this situation.

To eliminate or at least reduce such disturbances and reveal the different causal models hiding in the data, an interactive parallel coordinates interface (Fig. 1c) is employed in our CSI framework. Via the parallel coordinates, users can directly observe potentially attractive data subdivisions and partition the data by adjusting the brushed value range of variables. Conversely, data partitions can also be detected automatedly based on unique values of some variables or as data clusters recognized by clustering algorithms, using the interactive facilities shown in Fig. 1e.

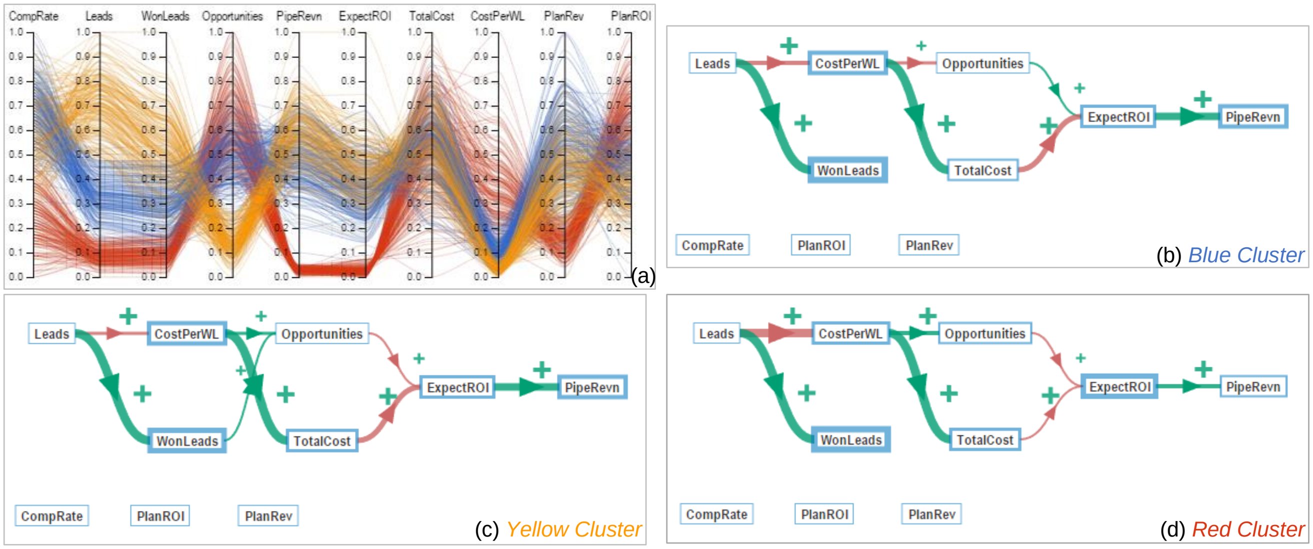
These interactive facilities also allow users to manage the recognized partitions. Users can save a partition as a tag, recall it in the parallel coordinates by clicking the tag, or fit it to a causal structure by hitting the "Fit Model" button. Most importantly, they can learn a causal model from each such data subdivision and refine it with the visual approaches introduced in Section 3.

4.1.1 Illustrative Example

We now demonstrate how to discover different causal models from data with the CSI interface through an illustrative example, leveraging the Sales Campaign dataset. This dataset contains 10 numerical variables and 600 records describing several important factors in sales marketing and their effects on a company's financials. Each sample in the dataset represents a sales person's sales behaviors. Three data clusters have been recognized by k-means clustering [38] and are colored blue, yellow and red, respectively (with interactive facilities shown in Fig. 1e). It is worth noting here that while we have implemented the k-means in the current version of the CSI interface for illustration, the proper choice of clustering algorithms may vary depending on the data. When constructing the causal model, we assume the following background knowledge. A sales pipeline starts with a *lead* generator developing prospective customers called *Leads*. When some leads return positive feedback, they become *WonLeads* and an increased sales pitch at cost of *CostPerWL* is invested in each of them, so that they might be further developed into real customers called *Opportunities*. The *TotalCost* reports the actual cost of each sales person. The goal of the entire efforts is to increase the expected return on investment (*ExpectROI*) and ultimately maximize the pipeline revenue (*PipeRevn*).

In our earlier work [2] we found several meaningful relations but these were conjunctive over the entire population of sales people in the dataset. However, when looking at the three clusters in the parallel coordinates in Fig. 6a it seems more meaningful to consider the three groups of sales people separately, as it is obvious that they are behaving very differently. It is likely that by doing so specific sales plans can be strategized for each of them. Hence, we click the "Infer on Each" button in Fig. 1e and three causality graphs are generated (see Fig. 6b, c, and d). They allow specified prescriptive analytics to be made for each sales group.

First, it is interesting to note that the three causality graphs have some structures that are similar, which is consistent with the background knowledge that there must be some marketing model guiding the sales behaviors. From the three graphs, one can see that *CompRate*, *PlanROI*, and *PlanRevn* are not related in the pattern and thus adjusting any of these variables will likely not affect revenue. A relation observed in all three graphs is that *ExpectROI* is directly affecting *PipeRevn* in a positive manner. This implies that the company's revenue prediction model seems to work well. *TotalCost* is consistently caused by *CostPerWL*, which is reasonable as investing in each customer represents the major costs in the pipeline. Further sound business facts realized by all groups are: (1) higher *TotalCost* will reduce *ExpectROI*, and (2) more Leads will require a reduction of *CostPerWL* (which is natural when the budget is fixed).



(a)

(b) *Blue Cluster*

(c) *Yellow Cluster*

(d) *Red Cluster*

Fig. 6 Causality analysis on the Sales Campaign dataset containing three sales groups. (a) The parallel coordinates view of the CSI interface displaying the three clusters of the dataset. (b), (c), (d) The path diagrams of causal networks generated from the corresponding sales groups. Both the structure and parameters of the three networks are somehow different, which implies different facts in sales behaviors.

However, the pathway *CostPerWL→Opportunity→ExpectROI* is somehow different for each model, implying distinct patterns in each group's sales behaviors. In the causality graph of the blue cluster (Fig. 6b), it is striking to see that more investment on each won lead is not bringing them more "opportunities" (referring to the negative effect of *CostPerWL* on *Opportunities*), i.e. they might have invested too much on each customer and probably inappropriately. But, the opportunities they get with their approach are profitably increasing *ExpectROI* and revenue, and so overall, they are successful. In contrast, the sales people in the yellow group (Fig. 6c) are gaining more opportunities from their investments (referring to the positive relation from *CostPerWL* to *Opportunities*), however, this is not bringing them more revenue, as *Opportunities* is not positively related to *ExpectROI*. Thus, they should work on increasing the profit of each closed deal. Finally, the sales group of the red cluster converts much less *ExpectROI* into *PipeRevn*, as indicated by the thinner green arrow between these two in Fig. 6d. Based on the negative causal relation from *Opportunities* to *ExpectROI*, this may have similar reasons than for the yellow cluster that their deals are not profiting, although their generous investment in *CostPerWL* does bring them many opportunities. They might better reduce the cost of each won lead and focus on increasing the profit.

Based on the different causal patterns observed, the analyst team may have many suggestions for each sales group. While discussing these specific strategies is beyond the scope of our research we believe that the case study presented here has shown that causality analysis with data partitioning can indeed reveal different causal facts hidden in the data.

4.2 Causal Model Visual Diagnostics

While causal inference on data subdivisions can result in multiple models revealing different causal patterns, diagnosing these models by investigating their similarities can often reveal interesting knowledge, especially when the data is bracketed into a large number of subsets and a corresponding number of models are learned. Meanwhile, doing so also brings the issue that the number of data points available to learn each model will be heavily reduced with more partitions added. This may potentially lower the statistical saliency of causal relations so that they may often be missed. Reducing p-value thresholds in CI tests could be a solution, however, it also results in more false relations and thus in less credible models.

To uncover the common causal patterns and extract reliable relations from all learned models, we propose a visual *pooling* process that can either occur at the causal link level or at the model level. In the following, we shall present the specific visual pooling strategies leveraging a real-world dataset.

4.2.1 Pooling at the Causal Model Level

The purpose of pooling at the causal model level is to recognize the possible grouping of causal models so that common causal relations can be summarized from models in the same group and different causal trends can be compared between models in different groups. To achieve this, we represent each causal graph as an adjacency matrix. Since a causal model features both its structure and parameters, we use the regression coefficient of each edge as the corresponding element in the matrix. Then, we can pool at the causal model level by clustering these adjacency matrices to uncover the different causal mechanisms embedded in them.

To demonstrate this method, we utilize the Ocean Chlorophyll dataset. The dataset was merged from several satellite data sources [39][40][41][42], monitoring the area of S22° ~ S25°, E50° ~ E53° (located at the south Madagascar sea). Each data source contains a particular physical property – ocean surface temperature, surface currents speed, wind speed, thermal radiation, precipitation rate, and water mixed layer depth, or a biological property – photosynthesis radiation activation and chlorophyll concentration. These satellite data come in different horizontal resolutions and were recoded into a 0.25-by-0.25-degree resolution in longitude and latitude. At each of the 169 geolocations, the time series spans 12 years (from 1998 to 2009) and were averaged in months (thus 144 data points). Partitioning data by each geolocation, 169 causal models are learned. Fig. 1f contains the heatmap of these models, where a darker tile denotes a model with a lower model score (thus better goodness) following the criterion in Section 3.2. Fig. 1b is the causal model denoted by the highlighted tile (that is colored in orange) in Fig. 1f.

To find possible groupings of the 169 models derived from the dataset, we apply k-medoids clustering [43], which is good at

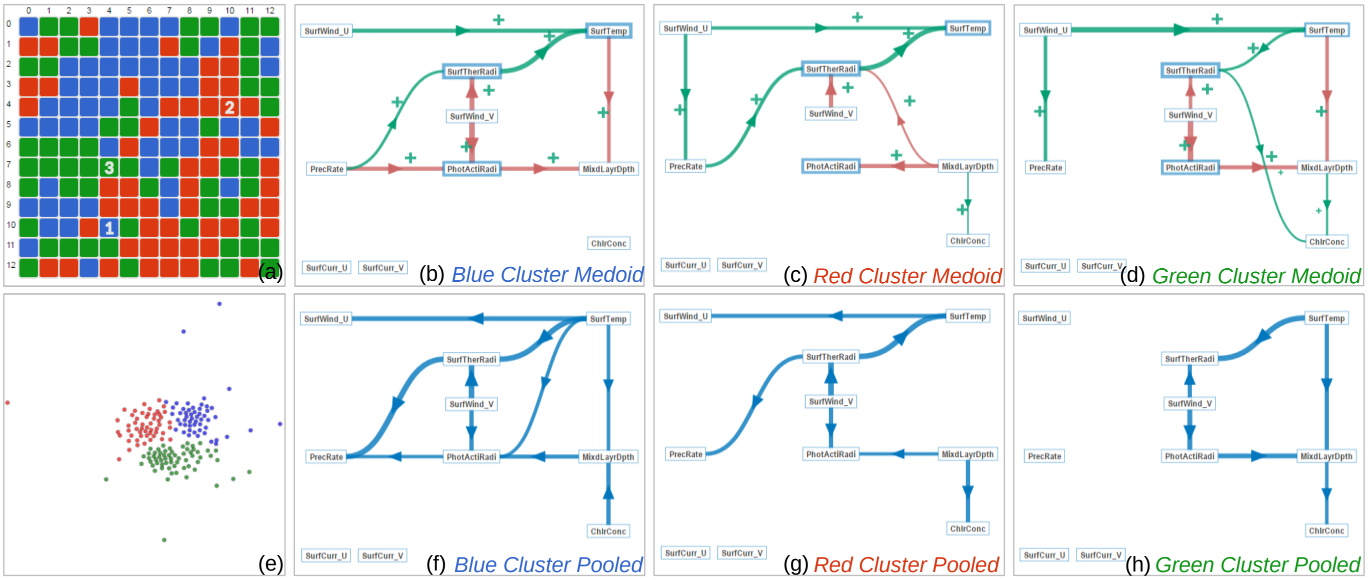


Fig. 7 Diagnostic of causal models learned from the Ocean Chlorophyll dataset by conditioning on each geolocation. (a) Heatmap of all models clustering into three clusters. (b), (c), and (d) are the representative models for the three clusters corresponding to the numbered tiles in (a). (e) is the t-SNE layout of these models' adjacency matrices in which we observe there are indeed three clusters. (f), (g), and (h) are pooled causal relations from the three clusters accordingly, with a credibility coefficient threshold of 0.5.

finding the representative objects among all. Here, by setting $k = 3$ with the controls in Fig. 1f, a new heatmap is generated in Fig. 7a. The three tiles marked with numbers denote the medoid models found by the clustering algorithm, i.e. the most representative model in each cluster. These three medoid causal models are visualized in Fig. 7b (blue cluster), c (red cluster), and d (green cluster). Here we place the nodes at the same location for each model to make comparisons easy for the analyst. As he has been trying to use this dataset to relate the unique cycle of the chlorophyll concentration variation with other variables, the most attractive difference for him could be that the *ChlrConc* is associated with other variables differently in the three representative models. Users can also examine other models by clicking on tiles of the heatmap. Also, we can cluster models into more groups with controls shown in Fig. 1f, although we observe there are indeed three dense areas in the t-SNE layout [44] of these models' adjacency matrices in Fig. 7e. The t-SNE layout is not included in the current CSI interface but can be easily incorporated in future extension.

4.2.2 Pooling at the Causal Links Level

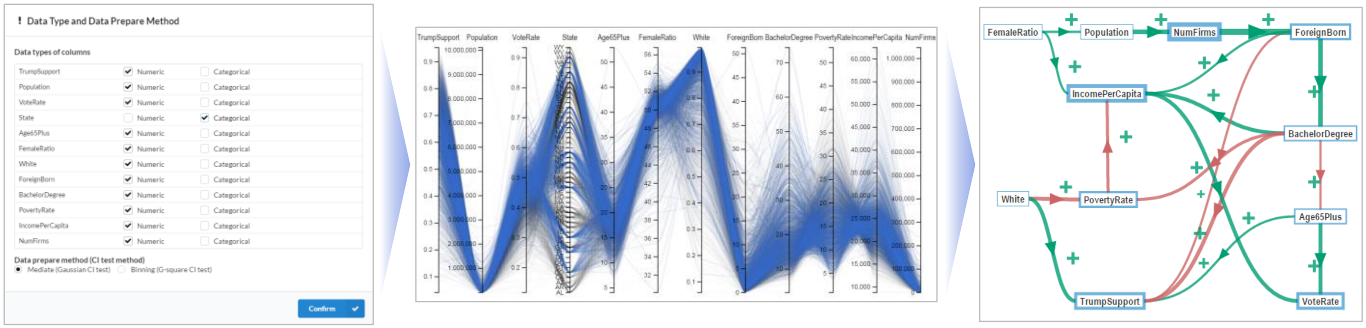
To summarize the common and credible relations from models in each cluster, we need to conduct pooling at the causal links level. The simplest pooling strategy that occurs at the causal link level is to count the frequency of each possible causal relation observed in all models. Then by setting thresholds on such statistics, only causal relations observed more than a certain number of times are returned, resulting in a combined model. A shortcoming of such strategy is that it equally considers all observed causal models, while they may actually have different levels of credibility. This might be fine for datasets in which all bracketed subsets enclose a sufficient number of records. But for other scenarios where the dataset is bracketed into a large number of subdivisions each containing only limited data samples, pooling by frequency may potentially enlarge the impact of the false relations found in low credibility models.

When a group of models is following similar causal processes, it is reasonable to infer that those true causal relations will be observed frequently in models with higher credibility so that they should be emphasized in pooling; while models with lower credibility can be considered random noise and thus should have a small weight. When a dataset is evenly partitioned (this is important as BIC is sensitive to sample numbers n in Eq. 2), the credibility of causal models learned from each data subset can be measured by their model scores. Then, as all possible causal relations form a complete graph, we assign each edge of the graph a normalized score calculated by summing up the credibility of all models in which the relation is observed. Specifically, the credibility score $C_e(e_j)$ for edge e_j is calculated as

$$C_e(e_j) = \frac{\sum_i \delta_{ij}\,(F_{max} - F_i)}{N(F_{max} - F_{min})} \tag{5}$$

where $\delta_{ij} = 1$ if e_j is included in model i, otherwise $\delta_{ij} = 0$; F_i is the score of model i, while F_{max} and F_{min} are the largest and the smallest score of all N models. By such, we consider edges with larger $C_e(e_j)$ are with higher credibility. Users can then work with a slider control to filter out edges with small scores, leaving only reliable relations.

We illustrate the effect of such pooling strategy by continuing the example of the Ocean Chlorophyll dataset. After clustering the causal models into three clusters, three combined models are pooled and shown in Fig. 7f, g, and h respectively. Here a credibility threshold of 0.5 is applied so that only strong credible causal relations are retrieved. Looking at the three models, there are seemingly some causal loops between environmental and biological variables in the whole area as causal relations with opposite directions between the same pair of variables are observed in different models. But one direction of the loop could be more dominating than the other in some sub-areas. For example, *MaxLayrDepth* is a good predictor of *PhotActiRadi* in the pooled models of the blue and the red clusters but the relation is reversed in the green cluster's model. Similarly, *MaxLayrDepth* is he only variable strongly associated with *ChlrConc* but the causal mechanisms are different in the three models. The scientific implication behind these findings could be rich but explaining them


! Data Type and Data Prepare Method

Data types of columns

TrumpSupport	☑ Numeric	□ Categorical
Population	☑ Numeric	□ Categorical
VoteRate	☑ Numeric	□ Categorical
State	□ Numeric	☑ Categorical
Age65Plus	☑ Numeric	□ Categorical
FemaleRatio	☑ Numeric	□ Categorical
White	☑ Numeric	□ Categorical
ForeignBorn	☑ Numeric	□ Categorical
BachelorDegree	☑ Numeric	□ Categorical
PovertyRate	☑ Numeric	□ Categorical
IncomePerCapita	☑ Numeric	□ Categorical
NumFirms	☑ Numeric	□ Categorical

Data prepare method (CI test method)

● Mediate (Gaussian CI test) ○ Binning (G-square CI test)

(a) (b) (c)

Fig. 8 Analyzing the Presidential Election dataset with the CSI framework. (a) The pop-up window where the user can select variable types and data preparation method. (b) The parallel coordinates visualizing the dataset. Counties of the 11 swing states are brushed as the election results in these areas are more decisive. (c) The derived causal network which uncovers many interesting facts behind the election results.

goes beyond the purpose of this paper. But the presented example has demonstrated the effectiveness of our pooling methods.

5 Case Studies

In this section, we further demonstrate the use of the CSI interface by analyzing two real-world datasets with all the techniques proposed in previous sections.

5.1 Case Study – Presidential Election Dataset

Donald Trump's unexpected triumph in the 2016 US Presidential Election has gathered worldwide attention and sparked extensive discussion. Since most polls and political analyses before the election failed to predict the win, there has been strong interest in finding the causes of what led to it. In an attempt to gain insight into this question, we have used our framework to conduct a causality analysis on the Presidential Election dataset. The dataset contains variables of the county-level election results and of each county's selected geographical features, i.e. population, vote rate, race ratios, income level, the level of education, etc., which are extracted from a more inclusive Kaggle data archive [45].

To analyze the dataset using our CSI interface, we first load the data and then select variable types (categorical or numeric) as well as data preparation method (GM with UB or equal-width binning) via the pop-up window shown in Fig. 8a. Then the data is visualized in the parallel coordinates as shown in Fig. 8b. Here data points corresponding to counties of the 11 swing states (according to the website Politico [46]) are brushed, as the election results in these areas are more decisive and Trump won in most of them. Then by clicking "Go Causality!" the causal network of Fig. 8c is returned.

We can observe many interesting causal relations in Fig. 8c. For example, *Age65Plus* and *White* (population percentage of those aged 65 or plus and those identified as White) are positively causing *TrumpSupport*, which is the supporting rate of candidate Trump in the county. This means that older people and Whites are mostly supportive for Trump. What's more, both of these two variables are positively causing *VoteRate* via different causal paths, implying Trump supporters are voting actively. On the other hand, those who were not preferring Trump are the immigrants and people with high education level, referring to the negative relation from *ForeignBorn* and *BachelorDegree* to *TrumpSupport*. However, the negative causal path *ForeignBorn*→ *BachelorDegree*→ *Age65Plus*→ *VoteRate* says that more immigrants and more people with Bachelor degree may indirectly hurt voting rate. Besides, when looking at the parallel coordinates, values on the axes of *ForeignBorn* and *BachelorDegree* are generally much smaller than values on axes of *Age65Plus* and *White*, suggesting the latter two are much bigger groups.

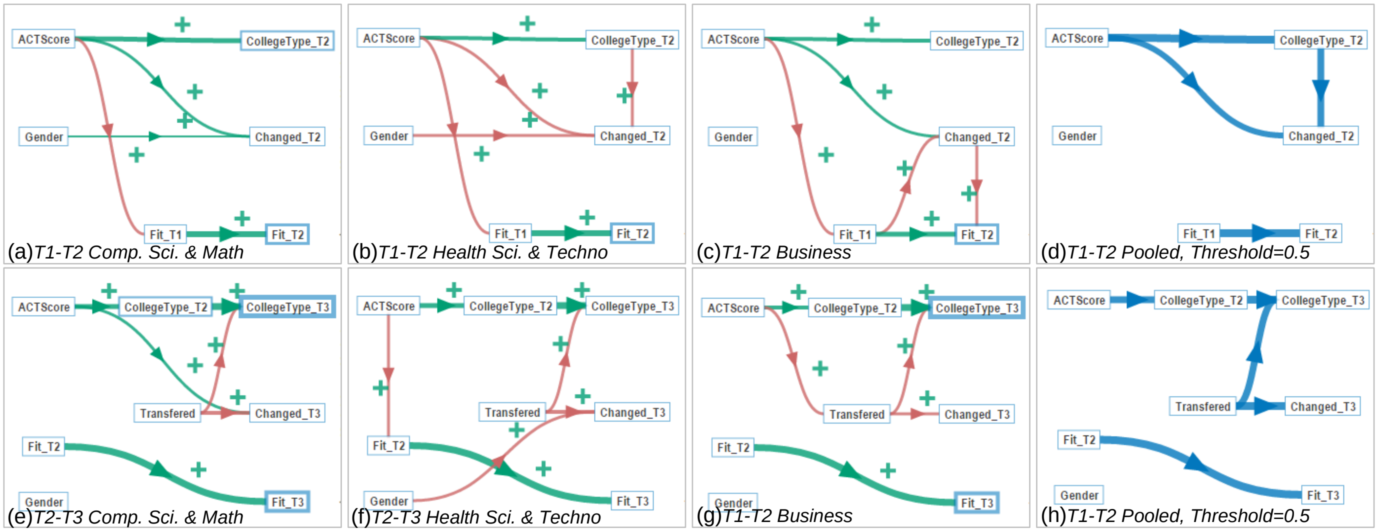
There are many more causal patterns we can observe that may entail various social facts. We cannot list them all here. While the presented analytics has explained the major reasons behind Trump's victory, we believe the causality analysis can also be applied to other political datasets, e.g. poll data, in a similar manner, which can potentially improve prediction accuracy.

5.2 The ACT Dataset

The original ACT dataset [47] was used to study why high school graduates change majors at college and has been modified so that its variables are more suited in a causality context. There are about 230,000 data points each represents a participated student. A student would report his/her college major three times in total – the expected one at the senior year of high school (*T1*) and the actual major at the first and second year of college (*T2* and *T3*). Majors are categorized into 18 fields. A test was also conducted at each point in time quantifying the student's fitness for his choice (*Fit_T1/T2/T3*). Other factors considered include a student's gender, ACT score, attended college type (2 or 4 years), and transfer between colleges.

Since there are two times at which a student may change majors (*T1* to *T2* and *T2* to *T3*), we arrange the variables into two different but overlapping groups, each corresponds to a sub-dataset. We then further subdivide the first sub-dataset based on students' major at *T1* and the second based on major at *T2*, so that students selecting different fields are studied separately, avoiding possible disturbances by Simpson's Paradox. Conditioning on these subdivisions, 36 causal networks (18 majors ×2 sub-datasets) are inferred and refined with our CSI framework. Some are visualized in Fig. 9. Again, we place the nodes at the same location for each model from the same sub-dataset to facilitate comparison.

Fig. 9a, b, and c are the causal models learned correspondingly from students who claimed at *T1* that they would take Computer Science and Math, Health Science, and Business in college. Here *Changed_T2* indicates whether the student entered a different major in the first year of college. There are some interesting observations when comparing the three figures. For example, in Fig. 9a, we see there is a gender bias indicated by the positive edge *Gender* → *Changed_T2*. As males are valued 1 in the binary variable *Gender*, this implies that they are more likely than females to major differently from what they expected earlier. Meanwhile, *ACTScore* is also playing as a positive motivation. However, the two relations become just the opposite in Fig. 9b, implying that a low ACT score would very likely make a girl, who initially wanted to take Health Science, attend another major. It also appears that students who wanted to enter Business schools are the only group among the three who considered their fitness to the major (referring



(a)*T1-T2 Comp. Sci. & Math* (b)*T1-T2 Health Sci. & Techno* (c)*T1-T2 Business* (d)*T1-T2 Pooled, Threshold=0.5*

(e)*T2-T3 Comp. Sci. & Math* (f)*T2-T3 Health Sci. & Techno* (g)*T1-T2 Business* (h)*T1-T2 Pooled, Threshold=0.5*

Fig. 9 Causal models inferred from the ACT dataset [47]. (a), (b), and (c) are causal networks explaining why students changed to other majors when entering college. (d) The model pooled from the first group of 18 models learned from data subdivisions. (e), (f), and (g) are causal networks explaining why students changed major in the first two years in college. (d) The model pooled from the second group of 18 models.

to *Fit_T1* → *Changed_T2* in Fig. 8c), even though they usually didn't get to change to a better fitting one (the negative edge *Changed_T2* → *Fit_T2*). As each data subdivision has a sufficient but different number of data points, the strategy of pooling by frequency is then applied. Fig. 9d shows the causal relations pooled from the 18 models with a frequency threshold of 0.5. We see that a student's decision for college major is generally affected by his ACT score and the type of college he had been admitted to, while the fitness score is seemingly irrelevant in most cases.

To see the motivation behind the major switch of a college student actually taking the above three majors at *T2*, the second data-subset variables are analyzed. Fig. 9e, f, and g are the corresponding causal networks and Fig. 9h is the pooled model with the frequency threshold of 0.5. From these visualizations, we can see that the transfer of college now becomes the most common reason for a student to change major, regarding the edge *Transferred* → *Changed_T3* in the three models as well as in the pooled model, while gender bias can only be observed in very few fields, e.g. the edge *Gender* →*Changed_T3* observed in Fig. 9f but not in Fig. 9e and 9g. Again, the fitness score is generally shown to be irrelevant.

Due to space limitations, we cannot list all inferred models here, but examining them comparably can surely lead to many more interesting findings. Nevertheless, the case study on the ACT dataset has demonstrated that different models underlying data subdivisions can be effectively uncovered with our CSI framework.

6 Conclusions

We have presented several new VA techniques for making visual causality analysis more practical for real-world applications. All of these new visual analytical methods were implemented in our CSI (Causal Structure Investigator) interface. They are general and applicable to a wide set of real world cases as demonstrated by examples and case studies presented in this paper.

In future work, we would like to compare different causal network visualizations with user studies, such that the most receptive one can be chosen accordingly. Further, we also plan to visualize the differential network so that two or more causal models can be compared visually in a single visualization.

A present limitation of our framework is that it does not support causality analysis on time series data, which would have many popular applications, such as finance, health, etc. A possible solution is to utilize the theory of logic-based causality, which can be capable of learning causes of certain events within time series. Another future work we like to explore is to gain the ability to build causal models utilizing data from different measurements and sources but generated by the same causal mechanism, which is called the data fusion problem [48] or integrative causal analysis [49]. A visual interface supporting such analytics would allow users to study scientific systems over a series of data collections.

Finally, as illustrated in this paper, causality analysis can serve as a starting point for prescriptive analytics. Automatic generation of such analytics is also a promising extension to our work, where specific actions could be recommended given a user's request.

Acknowledgement

This research was partially supported by NSF grant IIS 1527200 and the Ministry of Science, ICT and Future Planning, Korea, under the "IT Consilience Creative Program (ITCCP)" supervised by NIPA. Partial support was also provided by the US Department of Energy (DOE) Office of Science, Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences, and Biosciences. Some of this research was performed in the Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the DOE's OBER at Pacific Northwest National Laboratory (PNNL). PNNL is operated by the US DOE by Battelle Memorial Institute under contract No. DE-AC06-76RL0.

References

[1] T. Vigen, "Spurious Correlations." [Online]. Available: http://www.tylervigen.com/spurious-correlations. [Accessed: 01-Mar-2017].

[2] J. Wang and K. Mueller, "The Visual Causality Analyst: An Interactive Interface for Causal Reasoning," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 1, pp. 230–239, 2016.

[3] E. H. Simpson, "The Interpretation of Interaction in Contingency Tables," *Source J. R. Stat. Soc. Ser. B*, vol. 13, no. 2, pp. 238–241, 1951.

[4] P. J. Bickel, E. A. Hammel, and J. W. O'connell, "Sex bias in graduate admissions: data from berkeley.," *Science*, vol. 187, no. 4175, pp.

398–404, 1975.

[5] J. Pearl, *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.

[6] J. Pearl, “An Introduction to Causal Inference,” *Int. J. Biostat.*, vol. 6, no. 2, pp. 1–62, 2010.

[7] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*. New York, NY: Springer New York, 1993.

[8] K. P. Burnham and R. P. Anderson, “Multimodel Inference: Understanding AIC and BIC in Model Selection,” *Sociol. Methods Res.*, vol. 33, no. 2, pp. 261–304, 2004.

[9] G. Schwarz, “Estimating the Dimension of a Model,” *Ann. Stat.*, vol. 6, no. 2, pp. 461–464, 1978.

[10] D. M. Chickering, “Optimal structure identification with greedy search,” *J. Mach. Learn. Res.*, vol. 3, pp. 507–554, 2002.

[11] G. Cooper and E. Herskovits, “A Bayesian Method for the Induction of Probabilistic Networks from Data,” vol. 347, pp. 309–347, 1992.

[12] D. Colombo and M. H. Maathuis, “Order-independent constraint-based causal structure learning,” *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3741–3782, 2014.

[13] J. Pearl and T. S. Verma, “A theory of inferred causation,” *Stud. Log. Found. Math.*, vol. 134, pp. 789–811, 1995.

[14] J. P. Pellet and A. Elisseeff, “Using Markov Blankets for Causal Structure Learning,” *J. Mach. Learn. Res.*, vol. 9, pp. 1295–1342, 2008.

[15] K. Baba, R. Shibata, and M. Sibuya, “Partial correlation and conditional correlation as measures of conditional independence,” *Aust. New Zeal. J. Stat.*, vol. 46, no. 4, pp. 657–664, 2004.

[16] R. E. Neapolitan, “Chapter 10.3.1,” in *Learning Bayesian Networks*, Pearson, 2003, pp. 600–603.

[17] W. Bergsma, “Nonparametric testing of conditional independence by means of the partial copula,” *Arxiv Prepr. arXiv11014607*, pp. 1–14, 2011.

[18] K. Zhang, J. Peters, D. Janzing, and B. Schölkopf, “Kernel-based Conditional Independence Test and Application in Causal Discovery,” *27th Conf. Uncertain. Artif. Intell. (UAI 2011)*, pp. 804–813, 2011.

[19] N. Friedman and D. Koller, “Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks,” *Mach. Learn.*, vol. 50, no. 1–2, pp. 95–125, 2003.

[20] S. Bongers, J. Peters, B. Schölkopf, and J. M. Mooij, “Structural Causal Models: Cycles, Marginalizations, Exogenous Reparametrizations and Reductions,” *arXiv*, Nov. 2016.

[21] M. Chen *et al.*, “From Data Analysis and Visualization to Causality Discovery,” *Computer.*, vol. 44, no. 10, pp. 84–87, 2011.

[22] N. Elmqvist and P. Tsigas, “Animated visualization of causal relations through growing 2D geometry,” *Inf. Vis.*, vol. 3, no. 3, pp. 154–172, 2004.

[23] G. Vigueras and J. A. Botia, “Tracking Causality by Visualization of Multi-Agent Interactions Using Causality Graphs,” in *Programming Multi-Agent Systems*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 190–204.

[24] T. Dang, P. Murray, J. Aurisano, and A. Forbes, “ReactionFlow: an interactive visualization tool for causality analysis in biological pathways,” in *Proceedings of the 5th Symposium on Biological Data Visualization: Part 2*, 2015, vol. 9, no. Suppl 6.

[25] K. Wongsuphasawat and D. Gotz, “Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization,” *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 12, pp. 2659–2668, 2012.

[26] M. Monroe, R. Lan, H. Lee, C. Plaisant, and B. Shneiderman, “Temporal event sequence simplification,” *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 12, pp. 2227–2236, 2013.

[27] Z. Liu, Y. Wang, M. Dontcheva, M. Hoffman, S. Walker, and A. Wilson, “Patterns and Sequences: Interactive Exploration of Clickstreams to Understand Common Visitor Paths,” *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 1, pp. 321–330, 2017.

[28] J. Zapata-rivera, E. Neufeld, and J. E. Greer, “Visualization of Bayesian Belief Networks,” pp. 6–9, 2003.

[29] Z. Zhang, K. T. Mcdonnell, E. Zadok, and K. Mueller, “Visual Correlation Analysis of Numerical and Categorical Data on the Correlation Map,” *IEEE Trans. Vis. Comput. Graph.*, vol. 21, no. 2, pp. 289–303, 2015.

[30] C. Schulz, A. Nocaj, J. Goertler, O. Deussen, U. Brandes, and D. Weiskopf, “Probabilistic Graph Layout for Uncertain Network Visualization,” *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 1, pp. 531–540, 2017.

[31] J. Zhao, M. Glueck, S. Breslav, F. Chevalier, and A. Khan, “Annotation Graphs: A Graph-Based Visualization for Meta-Analysis of Data Based on User-Authored Annotations,” *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 1, pp. 261–270, 2017.

[32] Y. Wang *et al.*, “AmbiguityVis: Visualization of Ambiguity in Graph Layouts,” *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 1, pp. 359–368, 2016.

[33] D. McFadden, “Conditional logit analysis of qualitative choice behavior,” in *Frontiers in Econometrics*, 1974, pp. 105–142.

[34] K. Bache and M. Lichman, “UCI Machine Learning Repository,” *University of California, Irvine, School of Information*. 2013.

[35] R. E. Kass and A. E. Raftery, “Bayes Factor,” *Journal of American Statistical Association*, vol. 90, no. 430. pp. 773–795, 1995.

[36] J. J. Dziak, D. L. Coffman, S. T. Lanza, and L. Runze, “Sensitivity and specificity of information criteria,” 2012.

[37] M. Kalisch, M. Machler, D. Colombo, M. H. Maathuis, and P. Buhlmann, “Causal Inference Using Graphical Models with the R Package pcalg,” *J. Stat. Softw.*, vol. 47, no. 11, p. 26, 2012.

[38] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, “An efficient k-means clustering algorithm: analysis and implementation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, 2002.

[39] NASA, “SeaWiFS Project.” [Online]. Available: https://oceancolor.gsfc.nasa.gov/SeaWiFS/. [Accessed: 01-Mar-2017].

[40] NASA, “MODIS.” [Online]. Available: https://modis.gsfc.nasa.gov/. [Accessed: 01-Mar-2017].

[41] European Centre for Medium-Range Weather, “ECMWF.” [Online]. Available: http://www.ecmwf.int/. [Accessed: 01-Mar-2017].

[42] NASA, “Precipitation Measurement Missions.” [Online]. Available: https://pmm.nasa.gov/TRMM. [Accessed: 01-Mar-2017].

[43] H. S. Park and C. H. Jun, “A simple and fast algorithm for K-medoids clustering,” *Expert Syst. Appl.*, vol. 36, no. 2 PART 2, pp. 3336–3341, 2009.

[44] L. J. P. Van Der Maaten and G. E. Hinton, “Visualizing high-dimensional data using t-sne,” *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

[45] J. Wilson, “2012 and 2016 Presidential Elections,” *Kaggle*. [Online]. Available: https://www.kaggle.com/joelwilson/2012-2016-presidential-elections. [Accessed: 01-Mar-2017].

[46] POLITICO, “Battleground States Polling Average.” [Online]. Available: http://www.politico.com/2016-election/swing-states. [Accessed: 01-Mar-2017].

[47] “IEEE VGTC VPG International Data-Visualization Contest.” [Online]. Available: http://vacommunity.org/ieeevpg/viscontest/2015/index.html. [Accessed: 24-Mar-2017].

[48] E. Bareinboim and J. Pearl, “Causal inference and the data-fusion problem,” *Pnas*, vol. 113, no. 27, pp. 7345–7352, 2016.

[49] I. Tsamardinos, “Advances in Integrative Causal Analysis,” in *Proceedings of the UAI 2015 Conference on Advances in Causal Inference*, 2015, pp. 90–91.