Bayesian Networks in Survey Data: Robustness and Sensitivity Issues

FEDERICA CUGNATA

University Centre for Statistics in the Biomedical Sciences, Vita-Salute San Raffaele University, Milan, Italy

RON S. KENETT

KPA Ltd., Raanana, Israel and University of Turin, Italy

SILVIA SALINI

Department of Economics Management and Quantitative Methods, University of Milan, Italy

Bayesian networks (BN) implement a graphical model structure known as a directed acyclic graph (DAG) that is popular in statistics, machine learning, and artificial intelligence. They enable an effective representation and computation of a joint probability distribution (JPD) over a set of random variables. The paper focuses on the selection of a robust network structure according to different learning algorithms and the measure of arc strength using resampling techniques. Moreover, it shows how 'what-if' sensitivity scenarios are generated with BN using hard and soft evidence in the framework of predictive inference. Establishing a robust network structure and using it for decision support are two essential enablers for efficient and effective applications of BN to improvements of products and processes. A customer-satisfaction survey example is presented and R scripts are provided.

Key Words: Bayesian Network; Do Calculus, Hard and Soft Evidence; Importance Performance Analysis; Information Quality (InfoQ); Survey Data; What-If Scenario.

1. Introduction

IMPROVEMENTS are generated from ideas that are implemented and proven effective. In order to identify such ideas, statistical analysis can be used to assess the impact of change in specific variables on target variables. The paper is focused on modeling data with Bayesian networks (BN), which are both mathematically rigorous and intuitively understandable data analytic tools. A BN, under appropriate conditions, can be used to study how systems respond to hypothetical interventions and to diagnose what caused a specific outcome. These effects are graphically represented in a network structure where the nodes are joined together by a set of arcs. Such a graphical structure is determined by the application of one or more types of data-driven learning algorithms. Choosing the best BN, among many different structures obtained through different algorithms, can be based on various optimality criteria. Moreover, a BN can be used as a decision-support tool by policy makers for determining which predictor variables are important on the basis of their effect on target variables. The goals of choosing an adequate BN structure and of using a BN as a decision-support tool are addressed by this paper using a new approach. Specifically, we develop a method to eval-

Dr. Cugnata is a is Research Fellow at the University Centre for Statistics in the Biomedical Sciences (CUSSB). Her email is cugnata.federica@hsr.it.

Dr Kenett is the Chairman and CEO of the KPA Group, Israel; a Research Professor at the University of Turin, Italy; and an International Professor, NYU Tandon School of Engineering, Center for Risk Engineering, New York, USA. His email is ron@kpa-group.com

Dr. Salini is an Associate Professor of Statistics in the Department of Economics, Management and Quantitative Methods. She is the corresponding author. Her email is silvia.salini@unimi.it

uate and characterize the properties of BN-derived information predictions and diagnostic analysis.

A first issue treated here is the selection of a robust-network structure for predictive and diagnostic analysis.

Following the application of different learning algorithms to constructing a BN structure, some arcs in the network are recurrently present and some are not. As a basis for designing a robust BN, we determine how often an arc is present, across various algorithms, with respect to the total number of networks examined. The robust structure reappears with specific arcs, in most learned networks. For these variables, the link connection does not depend on the learning algorithm and the derived prediction and is therefore considered robust. The predictive performance of the selected network is evaluated through misclassification rates using Monte-Carlo replications.

We use resampling techniques to compute arc strength when a target node is selected. It is possible to obtain by this approach a measure of importance for the node.

After selection of a robust network, we consider sensitivity scenarios with a "what-if" analysis. To achieve this, we conduct computer experiments on a BN by conditioning on specific variable combinations and predicting the target variables using empirically estimated networks. We then analyze the effect of variable combinations on target distributions in order to study the effect of each variable on the target. To implement these two methods, we developed dedicated R functions that are available for download at links listed at the end of the paper.

The approach outlined above provides an essential complement to BN analysis that enhances the efficiency and effectiveness of quality-improvement initiatives. This type of BN assessment emphasizes the role of graphical models in predictive and diagnostic applications and provides a new approach for determining sensitivity and robustness of BN-derived estimates. Application of these two approaches are presented in the paper. General examples of applications of BNs in various areas such as healthcare, biotechnology, and management are presented in Kenett (2016).

The paper is organized as follows: Section 2 introduces BN and Section 3 presents the proposed method. Section 4 is an application example. Section 5 presents conclusions and directions for future work. Supplementary material, with R code software, is provided separately. The proposed methods are new contributions to applications of BN in quality-improvement initiatives. The provided R code supports a concrete implementation path.

2. Introduction to Bayesian Networks

Bayesian networks (BN) implement a graphical model structure known as a directed acyclic graph (DAG) that enables an effective representation and computation of joint probability distributions (JPD) over a set of random variables (Pearl (2000)). The structure of a DAG is defined by the set of nodes and the set of directed arcs (arrows). The nodes represent random variables and are drawn as circles labeled by variable names. The arcs represent direct dependencies among the variables and are represented by arrows between nodes. In particular, an arc from node X_i to node X_j represents a statistical dependence between the corresponding variables. Thus, the arrow indicates that a value taken by variable X_j depends on the value taken by variable X_i. Node X_i is then referred to as a 'parent' of X_j and, similarly, X_j is referred to as the 'child' of X_i. An extension of these genealogical terms is often used to define the sets of 'descendants', i.e., the set of nodes from which the node can be reached on a direct path.

The DAG guarantees that there is no node that can be its own ancestor (parent) or its own descendent. Such a condition is of vital importance to the factorization of the joint probability of a collection of nodes. Although the arrows represent direct causal connection between the variables, the reasoning process can operate on a BN by propagating information in any direction. A BN reflects a simple conditional-independence statement, namely that each variable, given the state of its parents, is independent of its nondescendants in the graph. This property is used to reduce, sometimes significantly, the number of parameters that are required to characterize the JPD of the variables. This reduction provides an efficient way to compute the posterior probabilities given the evidence present in the data (Lauritzen and Spiegelhalter (1988), Pearl (2000)). In addition to the DAG structure, which is often considered to be the qualitative part of the model, one needs to estimate the quantitative parameters of the model. These parameters are derived by applying the Markov property, where the conditional probability distribution at each node depends only on its parents. For discrete

random variables, this conditional probability is represented by a table, listing the local probability that a child node takes on each of the feasible values—for each combination of values of its parents. The joint distribution of a collection of variables can be determined uniquely by these local conditional probability tables. Formally, a BN, B, is an annotated graph that represents a joint-probability distribution over a set of random variables, $\mathbf{V}$, (Ben-Gal (2007)). A network is defined by a pair $B = \langle G, \Theta \rangle$, where G is the DAG whose nodes $X_1, X_2, \ldots, X_n$ represent random variables and whose arcs represent the direct dependencies between these variables. The graph G encodes independence assumptions, where variable X_i is independent of its nondescendents given its parents in G. This set of parents is denoted generically as π_i. The second component, Θ, denotes the set of parameters of the network. This set contains the parameter $\theta_{x_i|\pi_i} = P_B(x_i \mid \pi_i)$ for each realization x_i of X_i conditioned on π_i, the set of parents of X_i in G. Accordingly, B defines a unique joint-probability distribution over $\mathbf{V}$, namely,

$$P_B(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P_B(X_i \mid \pi_i) = \prod_{i=1}^{n} \theta_{X_i|\pi_i}.$$

For simplicity of representation, we omit the subscript B. If X_i has no parents, its local probability distribution is said to be *unconditional*; otherwise, it is *conditional*. If the variable represented by a node is *observed*, then the node is said to be an evidence node; otherwise, the node is said to be *hidden* or *latent*. The complexity of a domain may be reduced by models and algorithms that describe an approximated reality. When variable interactions are too intricate to apply in an analytic model, we can still represent current knowledge about the problem, such as including a cause generating at least one effect where the final effect is the target of the analysis (Pearl (2000)).

In order to fully specify a BN, and thus fully represent the joint-probability distribution it represents, it is necessary to specify for each node X the probability distribution for X conditional on X's parents. The distribution of X, conditional on its parents, may have any form. Sometimes only constraints on a distribution are known. One can then use the principle of maximum entropy to determine a single distribution, i.e., the one with the greatest entropy given the constraints (Gruber and Ben-Gal (2012)).

Often these conditional distributions include parameters that are unknown and must be estimated from data, for example, using the maximum-likelihood approach. Direct maximization of the likelihood (or of the posterior probability) is often complex when there are unobserved variables. An approach to this problem in the context of BN is the expectation-maximization (E-M) algorithm, which alternates computing expected values of the unobserved variables, conditional on observed data, with maximizing the complete likelihood assuming that previously computed expected values are correct. Under mild regularity conditions, this process converges on maximum-likelihood (or maximum posterior) values for the parameters (Heckerman (1995)).

A more fully Bayesian approach to parameter inference is to treat parameters as additional unobserved variables and to compute a full posterior distribution, over all nodes, conditional on observed data, and then to integrate out the parameters. This approach can be computationally expensive and leads to large-dimension models, so in practice, classical parameter-estimation approaches are more common (Neapolitan (2003)).

BN can be specified by expert knowledge indicating arcs that are imposed due to prior knowledge and first principles and arcs that should not be included (white lists and black lists, respectively). The BN can also be partially determined by expert knowledge with network structure learned from data accounting for the white and black lists. The parameters of the local distributions can be learned from data, priors elicited from experts, or both. Learning the graph structure of a BN requires a scoring function and a search strategy. Common scoring functions include the posterior probability of the structure given the training data, the Bayesian information criteria (BIC), or Akaike information criteria (AIC) (Scutari (2010)). When fitting models, adding parameters increases the likelihood, which may result in over fitting. Both BIC and AIC address this problem by introducing a penalty term for the number of parameters in the model, the penalty term being usually larger in BIC than in AIC. An exhaustive search, returning back a structure that maximizes the score, produces a very large number of variables. A local search strategy makes incremental changes aimed at improving the score of the structure. A global search algorithm, like Markov chain Monte Carlo, can avoid being trapped in local minima. For more on BN-structure learning, see (Gruber and Ben-Gal (2012)).

BN, like other statistical models, can be used to answer questions about the nature of the data that go

beyond the mere description of the observed sample. Techniques used to obtain answers based on *new evidence* are known in general as *inference*. For BN, the process of answering these questions is also known as *probabilistic reasoning* or *belief updating*, while the questions themselves are called *queries*.

In practice, probabilistic reasoning on a BN has its roots embedded in Bayesian statistics and focuses on the computation of posterior probabilities or densities. In terms of Bayesian inference, an evidence function that assigns a zero probability to all but one state is often said to provide *hard* evidence; otherwise, it is said to provide *soft*, sometimes also called "*virtual*," evidence.

The domain knowledge allows experts to draw an arc to a variable from each of its direct causes (i.e. "visiting Africa" may cause "tuberculosis"). Given a BN that specifies the JPD in a factored form, one can evaluate all possible inference queries by marginalization, i.e., summing out over 'irrelevant' variables. Two types of inference support are often considered: *predictive support* for node X_i, based on evidence nodes connected to X_i through its parent nodes (also called *top-down reasoning*), and *diagnostic support* for node X_i, based on evidence nodes connected to X_i through its children nodes (also called *bottom-up reasoning*).

When a BN is given a causal interpretation, the interpretation of queries and evidence changes as well. Just as the arcs in the network describe causal relationships instead of probabilistic dependencies, queries evaluate the probability of known causes given their effects or vice versa.

In this setting, posterior probabilities are not interpreted in terms of beliefs changing according to some observed evidence but rather as measures of the effects of *interventions* on the causal structure. For more on this, see Buhlmann (2013) and Maathuis et al. (2009).

The next section discusses a proposed approach for choosing the BN structure and a related sensitivity analysis methodology.

3. Selection and Sensitivity Analysis of Bayesian Networks

3.1. Selection of a Robust Bayesian Network

In practical applications, one is faced with choosing which network to use after deriving different network structures by applying different learning algorithms. In particular, it is important to check whether the chosen structure and, therefore, its arcs are influenced by the presence of outliers or groups of observations. The choice of a robust BN is a complex problem with no easily derived analytic solution. For a similar analysis in the context of CART and random forests, see Bar Hen et al. (2015).

The approach in this paper is based on the selection of a robust-network structure using computer-intensive methods. The proposed approach selects a network that contains the most common number of arcs in the networks produced by different algorithms. Software programs for calculating BN structure typically include about a dozen such algorithms; see, for example, Scutari (2010). The second step in the approach consists of changing the network learning-algorithm parameters (for example, by applying different scoring functions) and again selecting the network that presents the largest number of repeated arcs. The third step is based on a bootstrap resampling procedure of the initial dataset. A network is estimated from each bootstrap sample and the arcs that are absent in a preset proportion of cases are removed.

The development of techniques for assessing the statistical robustness of network structures learned from data has been limited (Scutari and Nagarajan (2011)). Structural learning algorithms are commonly studied by measuring differences from the true (known) structure of a small number of reference data sets. Because the true structure of their probability distribution is unknown, the usefulness of such an approach in investigating networks learned from real-world data sets is limited. It is possible to interpret the proportion of arcs present in each network as "arc strength". A more systematic approach to model sensitivity and, in particular to the problem of identifying statistically significant features in a network, has been developed by Friedman et al. (1999) using bootstrap resampling and model averaging. In the proposed methodology, the "arc strength" obtained with bootstrap resampling is used to derive the importance of the dimensions on a target node.

3.2. Distance-Weighted Influence

In the framework of importance-performance analysis in customer surveys (Martilla and James (1977), Kenett and Salini (2011)), the objective is to identify and understand the dimensions considered of high importance by customers with low perceived quality or satisfaction. These dimensions are primary

candidates for focused improvement initiatives. The level of importance can be assessed directly by asking the customers to rate the level of importance of individual items, for example, on a three-point scale with "1" (low importance), "2" (neutral) and "3" (high importance). For a sample questionnaire with such a scale, see Kenett and Salini (2011). If a direct assessment of importance is not included in a questionnaire, implicit or indirect assessment of importance can be assessed using statistical models.

Albrecht et al. (2014) proposed a metric called *distance-weighted influence* that ranks the influence of query nodes based on the structure of the network. This ranking tends to reflect the structural properties in the network: the longer the path from a node to the target node, the lower the influence of that node, while the influence increases with the number of such paths.

Following Albrecht et al. (2014) the *distance-weighted influence* of X on Y, $\mathrm{DWI}(X, Y; w)$, is defined as

$$\mathrm{DWI}(X, Y; w) = \sum_{s \in S(X,Y)} w^{|s|}, \qquad (1)$$

where $S(X, Y)$ is the set of simple paths in the Bayesian network that join the nodes X and Y, $|s|$ is the length of the simple path s, and w is the path weight, and this measure is interpretable as node importance with respect to a fixed target node.

A node x is an ancestor of the target node if there is a path from x to the target node. For nodes that are not ancestors of the target, the index of importance is equal to zero. The path weight w is defined as the product of the strengths of all arcs of the path.

3.3. "What If" Sensitivity Scenario in Bayesian Networks

Determining causality has been traditionally based on applications of randomized trials, where the design of the trial aims at identifying the effect of an intervention, such as the application of a specific treatment, versus a placebo treatment. In general, causality has been studied from two main different points of view, the "probabilistic" view and the "mechanistic" view. Under the probabilistic view, the causal effect of an intervention is judged by comparing the evolution of the system when the intervention is and when it is not present. The mechanistic view focuses on understanding the mechanisms, determining how specific effects come about. The interventionist and mechanistic viewpoints are not mutually exclusive. For example, when studying biological systems, scientists carry out experiments where they intervene on the system, for instance, by adding a substance or by knocking out a gene. However, the effect of a drug product on the human body cannot be decided only in the laboratory. A mechanistic understanding, based on pharmacometric models, is needed in order to determine if a certain medication ought to work. The concept of potential outcomes is present in the work on randomized experiments by Fisher and Neyman in the 1920s and was extended by Rubin in the 1970s to nonrandomized studies and different modes of inference (Rubin (2008), Meali et al. (2011)). In Rubin's work, causal effects are viewed as comparisons of potential outcomes, each corresponding to a level of the treatment and each observable, had the treatment taken on the corresponding level with at most one outcome actually observed, the one corresponding to the treatment level realized. In addition, the assignment mechanism needs to be explicitly defined as a probability model for how units receive the different treatment levels. With this perspective, a causal inference problem is viewed as a problem of missing data, where the assignment mechanism is explicitly modeled as a process for revealing the observed data. The assumptions on the assignment mechanism are crucial for identifying and deriving methods to estimate causal effects; see, for example, Frosini (2006). The term 'causal inference' denotes different ways to approach causal aspects of statistical analysis. Causal Bayesian networks are BN where the effect of any intervention can be defined by a 'do' operator that separates intervention from conditioning. The basic idea is that an intervention breaks the influence of a confounder so that one can make a true causal assessment. The established counterfactual definitions of direct and indirect effects depend on the ability to manipulate mediators.

Following Pearl (2015), the mathematical operator called 'do' simulates physical interventions by deleting certain functions from the model, replacing them with a constant $X = x$, while keeping the rest of the model unchanged. For example, to emulate an intervention $\mathrm{do}(x)$ that holds X constant (at $X = x$) in the network $N1$ in Figure 1 (left), we replace X with $x = x_0$, and obtain a new network $N2$ represented in Figure 1 (right).

The joint distribution associated with the modified network, denoted by $P(zy \mid \mathrm{do}(x))$, describes the post-intervention distribution of variables Y and Z (also called "controlled" or "experimental" distri-

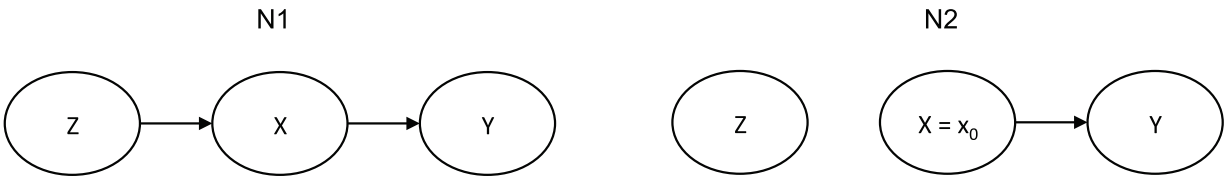

FIGURE 1. Network Pre and Post Intervention.

bution), to be distinguished from the preintervention distribution, $P(xyz)$, associated with the original network estimated from observed data. For example, if X represents a treatment variable, Y a response variable, and Z some covariate that affects the amount of treatment received, then the distribution $P(zy \mid \mathrm{do}(x))$ gives the proportion of individuals that would attain response level $Y = y$ and covariate level $Z = z$ under the hypothetical situation in which treatment $X = x_0$ is administered uniformly to the population. In words, the post-intervention distribution of outcome Y is defined as the probability that N_2 assigns to each outcome level $Y = y$. From this distribution, which is readily computed from any fully specified combination, we are able to assess treatment efficacy by comparing aspects of this distribution at different levels of x_0.

The 'do' operator makes it possible to conduct 'what-if' scenarios even if counterfactuals cannot be directly tested, such as in the presence of nonexperimental data. The intervention can correspond to 'hard' or 'soft' *new evidence* according to Bayesian inference; in the BN framework, the questions themselves are called *queries*.

Two type of queries can be considered:

1. *Conditional-probability query*, where conditions are on the distribution of one or more variables, but the probabilistic dependencies are left intact. The phenomenon is investigated as it was observed from the data and, therefore, the conditioning propagates to all other variables.
2. *Counterfactual query*, where the distribution of one or more variables is completely controlled, so the probabilistic dependencies of those nodes (e.g., incoming arcs) are removed from the BN. This is because an alternate scenario than that observed from the data is considered and the conditioning propagates only to variables downstream (the "effects", not the "causes").

The next section presents an application in the context of a customer-survey data. An early attempt to apply BN for the analysis of survey data was presented in Kenett and Salini (2009) and Salini and Kenett (2009); see also Gasparini et al. (2012). A survey with n questions produces responses that can be considered as random variables, $X_1, \ldots, X_n$. Some of these variables, q of them, are considered target variables. Responses to the other questions, $X_1, \ldots, X_k$, $k = n - q$, are analyzed under the hypotheses that they are potentially affecting the target variables.

Below is a summary of the procedure followed in customer survey the application:

Step (1) Estimate different network structures and select the more robust one.

Step (2) Measure the arch strength using bootstrap.

Step (3) Define a target node.

Step (4) Calculate DWI.

Step (5) Use soft and hard evidence and do calculus to obtain "what-if" sensitivity scenario.

4. Application

The example consists of a typical customer-satisfaction questionnaire filled out by passengers of airline companies to evaluate their experience on specific flights. The questionnaire contains questions on the passengers' satisfaction from their overall experience and from six specific dimensions of the service (*departure, booking, check-in, cabin environment, cabin crew, meal*). The evaluation of each item is based on a four-point scale (from 1= extremely dissatisfied to 4 = extremely satisfied). Additional information on passengers was also collected, such as gender, age, nationality, and the purpose of the trip. Results analyzed are based on responses in $n = 9{,}720$ valid questionnaires. The goal of the empirical application is to evaluate the importance of these six dimensions on satisfaction from the overall experience, taking into account the interdependencies between the degree of satisfaction from different aspects of the service. Clearly, these cannot be assumed to be independent of each other and, therefore, a BN analysis presents a particularly well-suited tool for this kind of analysis.

TABLE 1. BN with Proportion of Occurrence of Each Arc in the Bootstrap Replicates

	hc-bic	hc-aic	tabu-bic	tabu-aic	gs	iamb	fiamb	intamb	mmhc-bic	mmhc-aic	rsmax	tot
o_booking o_checkin	1.0	1.0	1.0	1.0	1.0	0.5	0.5	0.5	1.0	1.0	1.0	9.5
o_cabin o_crew	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	6.0
o_cabin o_departure	1.0	1.0	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	0.0	9.0
o_cabin o_experience	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	9.0
o_cabin o_meal	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	11.0
o_crew o_booking	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	6.0
o_crew o_departure	1.0	1.0	1.0	1.0	0.0	0.0	1.0	0.0	1.0	1.0	0.0	7.0
o_crew o_experience	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	9.0
o_departure o_booking	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	6.0
o_departure o_checkin	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	7.0
o_departure o_experience	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	0.0	9.0
o_meal o_crew	1.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	7.0
o_cabin o_booking	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	3.0
o_crew o_checkin	0.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	4.0
o_meal o_departure	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	3.0
o_meal o_experience	0.0	1.0	0.0	1.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	6.0
o_booking o_departure	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	0.0	0.0	1.0	5.0
o_crew o_meal	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	0.0	0.0	0.0	4.0
o_departure o_meal	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	2.0
o_checkin o_booking	0.0	0.0	0.0	0.0	0.0	0.5	0.5	0.5	0.0	0.0	0.0	1.5
o_checkin o_departure	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0	1.0	4.0
o_booking o_experience	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	2.0
o_checkin o_experience	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	2.0
o_checkin o_crew	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0
o_departure o_cabin	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0

The analysis shown below was performed using the R statistical language. Several R packages implement algorithms and models for constructing BN. We use here the *bnlearn* library of Scutari (2010) and Nagarajan et al. (2013) and the *gRain* library of Højsgaard (2012). We developed R functions to select the robust network, to obtain an measure of importance, and to analyze graphically the 'what-if' sensitivity scenarios. Algorithms for belief updating can be characterized either as exact or approximate. The *bnlearn* library implements approximate inference via rejection sampling (called in this setting logic sampling).

The first step of the analysis is the choice of the network. The data is analyzed with 11 algorithms implemented in the R package *bnlearn*: two scored-based learning algorithms (hill-climbing with score functions BIC and AIC and TABU with score functions BIC and AIC), five constraint-based learning algorithms (grow-shrink, incremental association, fast incremental association, interleaved incremental association, max-min parents and children), and two hybrid algorithms (max-min hill-climbing [MMHC] with score functions BIC and AIC, phase-restricted maximization). Table 1, based on Kenett et al. (2011), reports the occurrence of an arc between two nodes in implementation of each algorithm. A value of 1 indicates that the two nodes are linked by a directed arc, a value of 0.5 indicates that the two nodes are linked by an undirected arc. The last column reports the total score of each arc; this can be interpreted also as arc strength. We choose the BN that has the most arcs with scores equal to or higher than seven. This threshold is arbitrarily selected for this example; ideally it corresponds to more than 70% of the occurrences. The values of 1 and 0.5 represent a qualitative weight that, together with the cut-off criteria of 7, was tested empirically. Optimizing these weights and the-cut off value was beyond the scope of this paper.

Figure 2 (left panel) shows the BN obtained with the hill-climbing algorithm with score functions AIC. The total score on all 11 algorithms is reported for each arc. The red (light) arcs have a score equal to or higher than seven. To analyze the robustness of the chosen network, we also perform a bootstrap resam

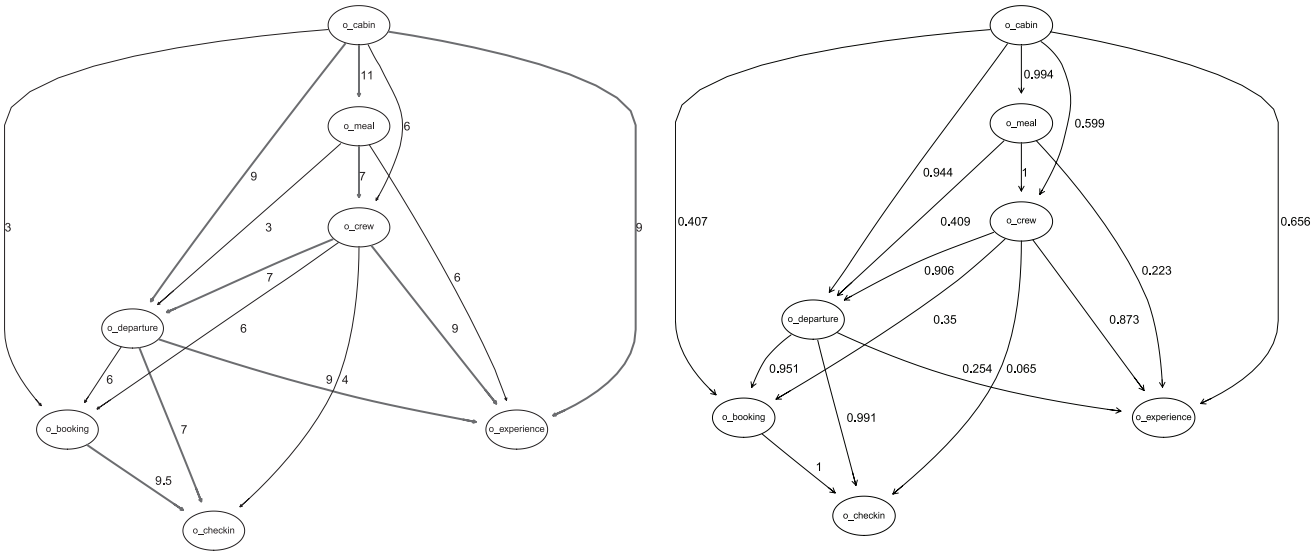

FIGURE 2. (left) BN Structure with Most Robust Arcs According to Different Structure Learning Algorithms; (right) BN with Proportion of Occurrence of Each Arc in the Bootstrap Replicates.

pling procedure on the initial dataset. We generate 1,000 random subsets, each of them with 1,000 observations. Then, the BN parameters are estimated for each bootstrapped sample. Figure 2 (right panel) shows the proportion of occurrence of each arc in the bootstrap replicates.

On the basis of the network, we evaluate the importance of the six dimensions on the overall experience-satisfaction target node. The index of importance DWI is based on all the paths from the considered nodes to the target node. The importance depends on the weight of each path and the length of the path from specific nodes to the target node, according to Equation (1). The weight of each path is equal to the product of the strengths of the arcs in the path.

Figure 3 shows a heatmap based on the DWI measure, with red denoting the target node (experience) and the intensity of the green (shade of grey) on the remaining nodes being proportional to the importance, i.e., paler means less influence. *Booking* and *Check in* aren't ancestors of the overall experience and their importance is therefore set to zero.

Figure 4 (left panel) shows the importance-performance analysis action grid. Each dimension is represented by its index of importance (x-axis) and its proportion of '4' ratings (very satisfied respondents), labeled TOP4 (y-axis). *Cabin environment* is the dimension with highest importance but lowest satisfaction. This dimension represents a key area that needs to be improved with top priority. For more on such an analysis, see (Kenett and Salini (2011)). In order to have a benchmark, we report in Figure 4 (right panel) the action grids obtained on the same data with other statistical models, in particular the performance indicator is obtained through the Rasch model (RM) and the importance indicator is obtained through the nonlinear principal-component

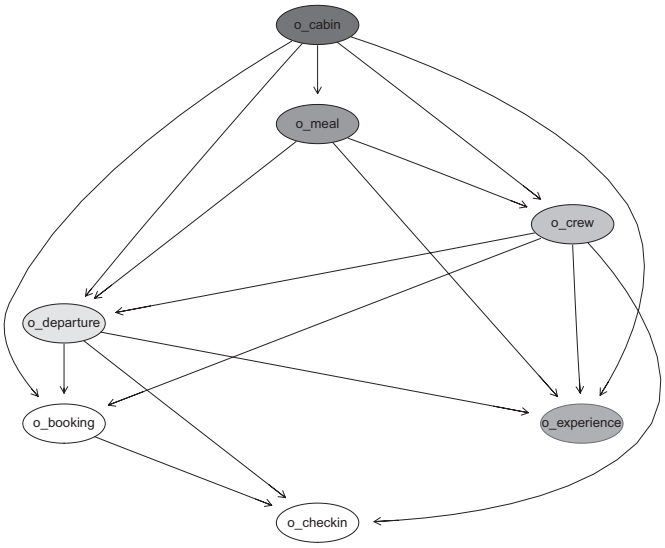

FIGURE 3. Heatmap Based on the DWI Measure, Target Node *Overall Experience*.

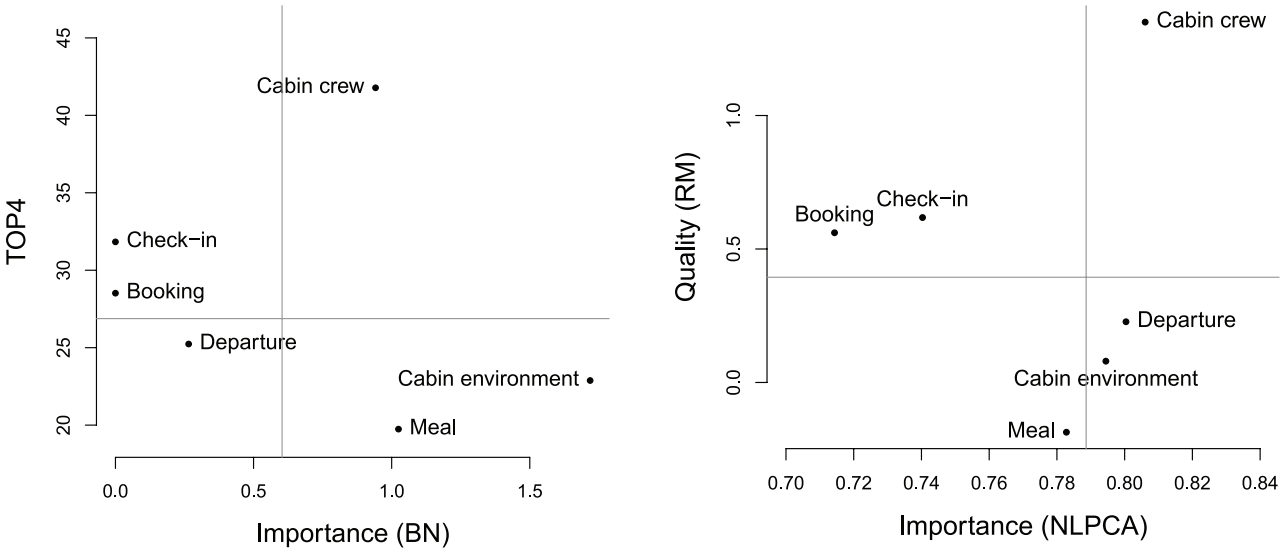

FIGURE 4. Importance-Performance Analysis: (left) BN Action Grid; (right) NLPCA-RM Action Grid.

analysis (NLPCA) (Cugnata and Salini (2013)). The results shown in the two figures are reasonably consistent. This provides additional evidence for the robustness of the findings. In the supplementary material, tables with the importance-performance values plotted in Figure 4 are reported.

The choice of the most robust network does not guarantee that the selected network is also the most efficient predictor. Between networks with the largest number of arcs, we choose the one with the lowest misclassification rate. To study the generalization aspect of the network, we train the net on a training sample and test it on a test sample. Figure 5 shows the misclassification rate resulting from 1,0000 Monte-Carlo replications of the procedure in the training and in the test set. The two paths represent different splitting percentage. The performance of the BN is in line with the classical model for ordinal data. As a benchmark, we consider an ordered logit model as in Pearl (2016) and the resulting misclassification rate is around 30% of cases as in the BN. The misclassification rate is similar in the test and in the training set, so there is no overfitting.

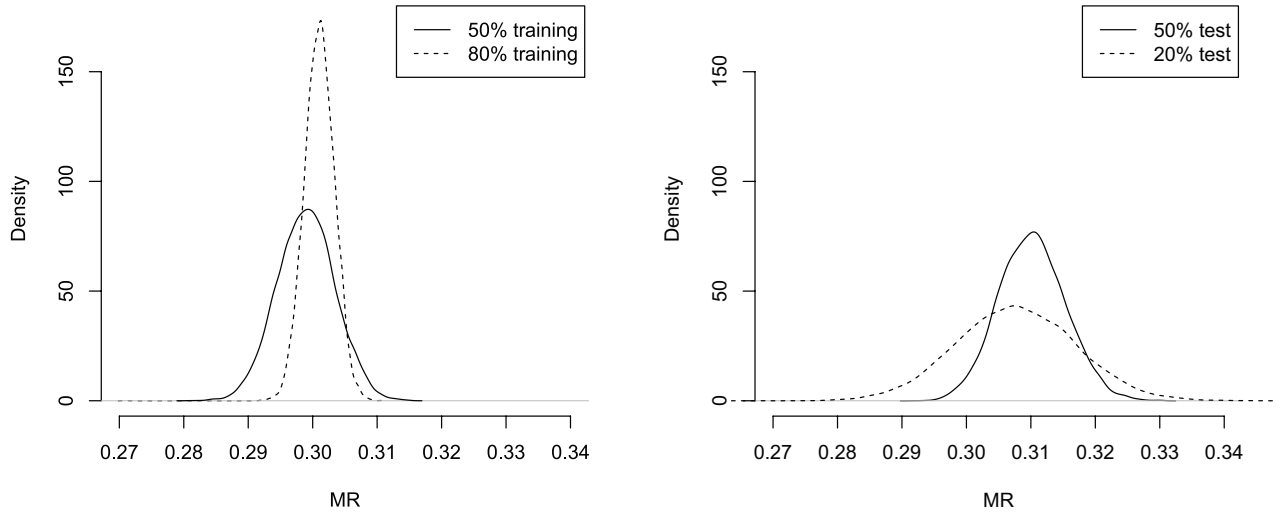

FIGURE 5. Misclassification Rate for the BN for Training Sets and Test Sets. Ten thousand Monte-Carlo replications.

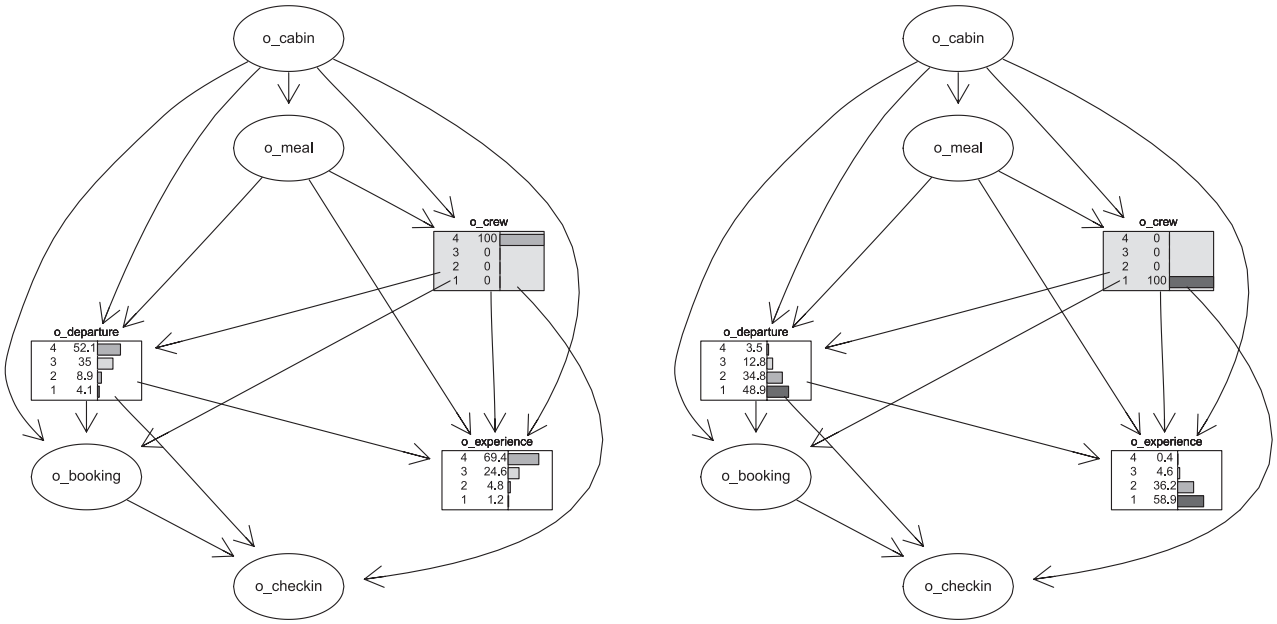

FIGURE 6. "What-If" Sensitivity Scenarios with Hard Evidence.

The network performance obviously increases when the training set is larger. There is more variability with new data (right panel) but the mean is a little smaller.

In the supplementary material, we provide the R script used to analyze the data with the algorithms implemented in the R package *bnlearn* and identify the BN with the most robust arcs and the lowest misclassification rate. It is also possible to get the R script that produces the index of importance.

On the basis of the network structure, as mentioned in Section 3.3, it is possible to perform various diagnostic checks to investigate the effects of evidence on the distribution of the target variable using "what-if" sensitivity scenarios and do calculus to test soft and hard evidence; see also Cugnata et al. (2014). The available R libraries implement algorithms for belief updating but do not incorporate an intuitive graphical representation of the results. We developed specific R functions using the *Rgraphviz* package to easily produce a BN plot with highlighted evidence and the consequent conditional probabilities of the target (see supplementary material). Hard evidence is an intervention of one or more variables in the network to a specific value. Figure 6 presents the actual results on the target variable of entering different types of hard evidence (evidence is shown as a gray box). In all cases, the structure of the BN remains fixed. Figure 6 (left) and Figure 6 (right) show distributions of overall experience conditioned on the *cabin crew* being at its highest level and its lowest level. The probability of being satisfied or very satisfied from the overall experience originally was equal to 70%; it is now 94% if satisfaction of the *cabin crew* is equal to four and it is 5% if satisfaction of the *cabin crew* is equal to one.

To investigate the effects of driver combinations, it is possible to use multiple evidence. Moreover, in case of uncertainty in setting the evidence, we can also test a soft-evidence hypothesis related to the explanatory variables.

Figure 7 shows an example of soft multiple evidence on *cabin crew* and *departure*. In particular, we assume 70% of extremely satisfied for both. Now the target variable probability is 95% instead of 70%.

In the supplementary materials there are other examples of 'what-if' sensitivity scenarios with multiple soft and hard evidence.

The airline dataset is a classic example of a customer-satisfaction survey performed to assess quality of a service. The dataset is composed of response variables (overall satisfaction, repurchasing intention, and recommendation) and dimensions that describe the service and are evaluated separately. In this type of application, the objective is to measure

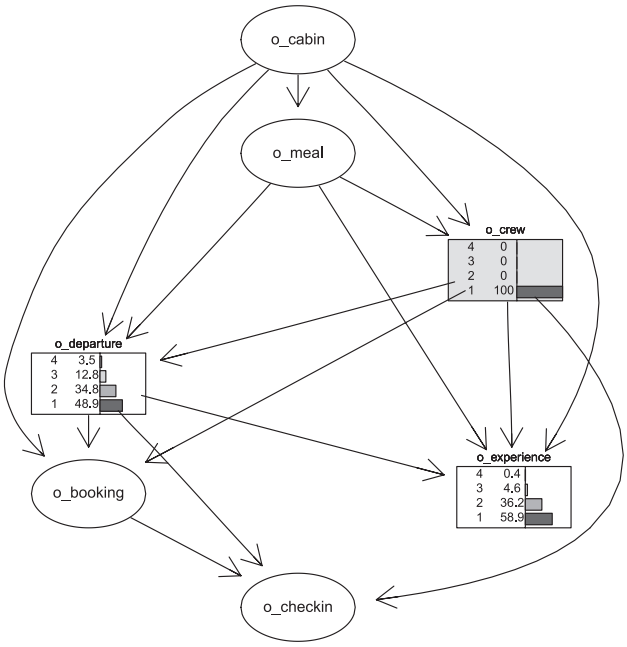

FIGURE 7. "What-If" Sensitivity Scenarios with Multiple Soft Evidence.

the quality of service as a whole and understand what dimensions are the most important in the opinion of the users of the service. A BN provides a measure of the importance of the dimensions on overall satisfaction and helps understand how the various categories of variables and their combinations change the JPD of the overall customer satisfaction. The ability to assume different interventions and to predict their effect are essential capabilities of decision-support systems for policy makers, engineers, and managers (Kenett and Salini (2011)). The graphical display of the scenarios makes the instrument operationally effective and very easy to explain and communicate.

5. Discussion

In the summary of his seminal 1995 paper, Pearl states "... graphical models can be used as a mathematical language for integrating statistical and subject-matter information ... diagrams can be queried to produce mathematical expressions for causal effects in terms of observed distributions" (Pearl (1995)). The development of BN, as a nonparametric model of causality effects in nonexperimental data, is providing new examples of applications in a range of domains. In Cornalba et al. (2007) and Cugnata et al. (2014), it is proposed to investigate the sensitivity of network estimates using experimental design conditioning of the driving variables. In this paper, this approach is expanded and integrated in a comprehensive framework investigating properties of both the structure of the network and the derived estimates. More important, an R application is provided to operationalize the proposed approach, which has been demonstrated using a case study. In the airline survey, the objective is to determine specific improvement actions. The relationships between variables, determined by a BN, have significant consequences. In-depth analysis of the robustness of these relations is therefore contributing to the quality of information provided by them. A similar proposal, in the context of reproducible research, is presented in Djulbegovic and Hozo (2014). Overall, this work is about increasing the quality of information provided by a BN. A general approach to determine information quality (InfoQ) is proposed by Kenett and Shmueli (2014). Additional research is needed to expand this analysis and extend the information quality provided by BN methods. Robustifying the BN analysis and identifying the importance of specific links enhances the value of decision-support systems based on a BN. This work provides theory, examples, and software for achieving these goals.

Acknowledgments

We thank MIUR PRIN MISURA (multivariate models for risk assessment) for financial support of the project.

Supplementary Material

- Customer-satisfaction survey data: http://users.unimi.it/salini/RSBN/data.zip.
- Complete R script used to derive all the tables and figures reported in the paper and R functions developed by the authors: http://users.unimi.it/salini/RSBN/Rcodes.zip.

References

Albrecht, D.; Nicholson, A. E.; and Whittle, C. (2014). "Structural Sensitivity for the Knowledge Engineering of Bayesian Networks". In *Probabilistic Graphical Models*, pp. 1–16. Basel, Switzerland: Springer International Publishing.

Bar Hen, A.; Gey, S.; and Poggi, J. M. (2015). "Influence Measures for CART Classification Trees". *Journal of Classification*, to appear.

Ben-Gal, I. (2007). "Bayesian Networks". In *Encyclopedia of Statistics in Quality and Reliability*, Ruggeri, F.; Kenett, R. S.; and Faltin F., eds. Chichester, UK: John Wiley and Sons.

Buhlmann, P. (2013). "Causal Statistical Inference in High Dimensions". *Mathematical Methods of Operations Research* 77, pp. 357–370.

Cornalba, C.; Kenett, R. S.; and Giudici, P. (2007). "Sensitivity Analysis of Bayesian Networks with Stochastic Emulators". Conference on Computer Experiments Versus Physical Experiments. Turin ENBIS-DEINDE.

Cugnata, F.; Kenett, R. S.; and Salini, S. (2014). "Bayesian Network Applications to Customer Surveys and InfoQ". *Procedia Economics and Finance* 17, pp. 3–9.

Cugnata, F. and Salini, S. (2013). "Model-Based Approach for Importance-Performance Analysis". *Quality and Quantity* 48(6), pp. 3053–3064.

Djulbegovic, B. and Hozo, I. (2014). "Effect of Initial Conditions on Reproducibility of Scientific Research". *Acta Informatica Medica* 22, pp. 156–159.

Friedman, N.; Goldszmidt, M.; and Wyner, A. (1999). "Data Analysis with Bayesian Networks: A Bootsotrap Approach". In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, pp. 196–205.

Frosini, B. (2006). "Causality and Causal Models: A Conceptual Perspective". *International Statistical Review* 74, pp. 305–334.

Gasparini, M.; Pellerey, F.; and Proietti, M. (2012). "Bayesian Hierarchical Models to Analyze Customer Satisfaction Data for Quality Improvement: A Case Study". *Applied Stochastic Models in Business and Industry* 28, pp. 571–584.

Gruber, A. and Ben-Gal, I. (2012). "Efficient Bayesian Network Learning for System Optimization in Reliability Engineering". *Quality Technology & Quantitative Management* 9(1), pp. 97–114.

Heckerman, D. (1995). "A Tutorial on Learning with Bayesian Networks". Tech. Rep. MSR-TR-95-06, http://research.microsoft.com.

Højsgaard, S. (2012). "Graphical Independence Networks with the gRain Package for R". *Journal of Statistical Software* 46(10), pp. 1–26.

Kenett, R. S. and Salini, S. (2009). "New Frontiers: Bayesian Networks Give Insight into Survey-Data Analysis". In *Quality Progress*, pp. 31–36.

Kenett, R. S. and Salini, S. (2011). *Modern Analysis of Customer Satisfaction Surveys*. Chichester, UK: John Wiley and Sons.

Kenett, R. S.; Perucca, G.; and Salini, S. (2011). "Bayesian Networks". In *Modern Analysis of Customer Satisfaction Surveys*, Kenett, R. S. and Salini, S., eds. Chichester, UK: John Wiley and Sons.

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

Kenett, R. S. (2016). "On Generating High InfoQ with Bayesian Networks". *Quality Technology and Quantitative Management* 13(3), to appear.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988). "Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems". *Journal of the Royal Statistical Society, Series B* 50(2), pp. 157–224.

Martilla, J. A. and James, J. C. (1977). "Importance-Performance Analysis". *The Journal of Marketing* 41(1), pp. 77–87.

Maathuis, M. H.; Kalisch, M.; and Buhlmann, P. (2009). "Estimating High-Dimensional Intervention Effects from Observational Data". *Annals of Statistics* 37, pp. 3133–3164.

Mealli, F.; Pacini, B.; and Rubin, D. B. (2011). "Statistical Inference for Causal Effects". In Kenett, R. S. and Salini, S., eds, *Modern Analysis of Customer Surveys*. Chichester, UK: John Wiley and Sons.

Neapolitan, E. R. (2003). *Learning Bayesian Networks*. Upper Saddle River, NJ: Prentice Hall.

Nagarajan, R.; Scutari, M.; and Lèbre, S. (2013). *Bayesian Networks in R with Applications in Systems Biology*. Use R! series. New York, NY: Springer.

Pearl, J. (1995). "Causal Diagrams for Empirical Research". *Biometrika* 82(4), pp. 669–688.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge, UK: Cambridge University Press

Pearl, J. (2016). "Trygve Haavelmo and the Emergence of Causal Calculus". *Economic Theory* 31, pp. 152–179.

Perucca, G. and Salini, S. (2014)." Travellers' Satisfaction with Railway Transport: A Bayesian Network Approach". *Quality Technology & Quantitative Menagment*. 11(1), pp. 71–84.

Rubin, D. (2008). "For Objective Causal Inference, Design Trumps Analysis". *The Annals of Applied Statistics*, pp. 808–840.

Salini, S. and Kenett, R. S. (2009). "Bayesian Networks of Customer Satisfaction Survey Data". *Journal of Applied Statistics* 36(11), pp. 1177–1189.

Scutari, M. (2010). "Learning Bayesian Networks with the bnlearn R Package". *Journal of Statistical Software*, pp. 1–22.

Scutari, M. and Nagarajan, R. (2011). "On Identifying Significant Edges in Graphical Models". In *Proceedings of the Workshop "Probabilistic Problem Solving in Biomedicine" of the 13th Artificial Intelligence in Medicine (AIME) Conference*, pp. 15–27.