



UNIVERSITÀ DEGLI STUDI ROMA TRE
DIPARTIMENTO DI ECONOMIA

**PARADATA AND BAYESIAN NETWORKS: A TOOL FOR
MONITORING AND TROUBLESHOOTING THE DATA
PRODUCTION PROCESS**

Marco Ballin, Mauro Scanu, Paola Vicard

Working Paper n° 66, 2006



UNIVERSITÀ DEGLI STUDI ROMA TRE
DIPARTIMENTO DI ECONOMIA

Working Paper n° 66, 2006

Comitato Scientifico

J. Mortera
M. Barbieri
C. Conigliani

- I “Working Papers” del Dipartimento di Economia svolgono la funzione di divulgare tempestivamente, in forma definitiva o provvisoria, i risultati di ricerche scientifiche originali. La loro pubblicazione è soggetta all’approvazione del Comitato Scientifico.
- Per ciascuna pubblicazione vengono soddisfatti gli obblighi previsti dall’art. 1 del D.L.L. 31.8.1945, n. 660 e successive modifiche.
- Copie della presente pubblicazione possono essere richieste alla Redazione.

REDAZIONE:

Dipartimento di Economia
Università degli Studi Roma Tre
Via Ostiense, 139 - 00154 Roma
Tel. 0039-6-57374003 fax 0039-6-57374093
E-mail: dip_eco@uniroma3.it

UNIVERSITÀ DEGLI STUDI ROMA TRE
DIPARTIMENTO DI ECONOMIA

**PARADATA AND BAYESIAN NETWORKS: A TOOL FOR
MONITORING AND TROUBLESHOOTING THE DATA
PRODUCTION PROCESS**

Marco Ballin^{*}, Mauro Scanu^{**} and Paola Vicard^{***}

* Istat, via A.Ravà 150,00100 Roma Italy

** Istat, via C.Balbo 16,00184 Roma Italy (scanu@istat.it)

*** Dipartimento di Economia, Università Roma Tre, via Ostiense 139, 00154 Roma Italy (vicard@uniroma3.it)

1. INTRODUCTION	1
1.1 DESCRIPTION	1
2. INFLUENCE DIAGRAMS	2
2.1 WHAT IS AN INFLUENCE DIAGRAM	2
2.2 HOW INFLUENCE DIAGRAMS WORK	5
3. INFLUENCE DIAGRAMS AS A DIAGNOSTIC TOOL FOR RESPONSIVE DESIGNS	7
3.1 A TOY EXAMPLE	7
3.2 THE INFLUENCE DIAGRAM FOR THE MOTIVATING EXAMPLE	8
4. FURTHER DEVELOPMENTS AND EXTENSIONS	13
REFERENCES	15

Abstract. *The problem of monitoring and managing the data production process by means of process flow indicators is presented in a decision theory framework. Here it is shown how to represent and solve the decision problem via influence diagrams, i.e. Bayesian network supporting decisions. An illustrative example is provided.*

KEY WORDS: Expected utility, graphical models, probability update, responsive survey design.

1. INTRODUCTION

1.1 Description

As stated in Heeringa and Groves (2004): “The ability to continually monitor the streams of process data and survey data creates the opportunity to alter the design during the course of data collection in order to improve cost efficiency and achieve more precise, less biased estimates.” The designs that are adaptive to the flow of process data are usually named *responsive designs*. There are a number of indicators that describe the process stream based on interviewer, housing unit, respondent, attempt, and response characteristics. These indicators are usually named *paradata*. The use of paradata in responsive designs may be described in the following way:

1. the process of altering the design is a *decision* procedure;
2. each possible decision is associated with costs/benefits;
3. each decision can be taken via an optimisation procedure, *i.e.* by finding the decision which minimizes the *expected cost*;
4. the expected costs are updated by the flow of paradata.

This procedure shows that the notion of *expected value* is central not only in planning the survey or in evaluating the final results (e.g. variance or mean

square error) but in the management of the processes flow too. This expected value has to be applied to the cost associated to the different survey phases.

In order to deduce the expected costs, it is important to understand how paradata interact, or in other words, to find a suitable multivariate model for paradata and an easy scheme for model parameter updating when some paradata are observed.

The different costs are those related to the actions (or decisions) of the survey manager. Hence the previous problem is actually a decision problem. Statistical decision theory is widely applied in the most diverse settings, but it is still not widely used for survey planning and management.

An easy way to model this problem is offered by graphical models known as *Bayesian networks* (BN), Cowell et al. (1999). When BNs are used in a decision context, they are usually named *Influence diagrams* (ID).

The paper is organised as it follows. An introduction to IDs and their usefulness in the responsive design context are presented in Section 2. An explanation of how a responsive survey design can be modelled via an ID is given in Section 3. Further aspects to investigate are given in Section 4.

2. Influence Diagrams

2.1 What is an Influence Diagram

An *influence diagram* (ID) is a graphical and mathematical representation of a decision problem. In order to define an ID, it is necessary to consider (i) its structure and (ii) its quantitative specification (Jensen, 2001).

Structure of an ID - The structure of an ID consists of:

- chance nodes: these nodes are random variables, possibly latent (represented by circles);

- decision nodes: these nodes represent the set of possible decisions (represented by rectangles);
- utility nodes: these nodes represent utilities or costs associated with each decision; they are represented by rhombuses;
- directed arcs connecting the nodes in such a way that the resulting graph does not contain any cycle (directed acyclic graph);
- decision nodes are included in a directed path (describing the chronological or logical order of the decisions).

Parents, children, ancestors and descendants are defined as in Lauritzen (1996). The meaning of the arcs (arrows connecting two nodes) is different according to the characteristics of the node to which the arrow is directed (receiver). If the receiver is a chance node, the arc is named *conditioning arc*. Its meaning is that the chance node is independent of all its ancestors given its parents. If the receiver is a decision node, the arc is named *information arc*, and the parents and ancestors of the decision node are assumed to be known before the decision is taken. When the receiver is a utility node, the arc specifies the functional dependence between the utilities and the values of its parents. Utility nodes cannot have children.

Quantitative specification of an ID - it is assumed that both chance and decision nodes have a finite set of mutually exclusive states (the utility nodes have no states). Furthermore, each chance node is given a conditional probability distribution given its parents, while each utility node is associated with a real valued function over the states of its parents.

IDs are a compact representation of *decision trees*, and for this reason are widely used in decision analysis. One of its most important characteristics is that an ID is an extension of a Bayesian network (BN, also known as probabilistic expert systems, see Cowell *et al.*, 1999). A BN is obtained from an ID removing the information arcs. The BN is a graph representing the multivariate probabilistic structure of chance nodes (decision nodes are just conditional states). Its most important feature is the propagation of

evidence by fast algorithms for updating the joint distribution of all variables in the network. Propagation is performed when the state of some variables becomes known, or when external knowledge on some marginal distributions should be included in the multivariate distribution. These models, as well as their extension to decision problems with IDs, have been successfully applied in different settings, as artificial intelligence, medical diagnostics and forensic genetics (see Neapolitan, 2004).

Although an ID is a compact representation of a decision tree, it may still have computational problems. In decision analysis it is required that each decision depends on all the relevant past observed variables (an assumption known as *no forgetting* or *perfect recall*). Lauritzen and Nilsson (2001) introduced a modification of ID, known as LIMIDs (limited memory influence diagrams), that do not need the perfect recall constraint of an ID.

They built the LIMID modelling the decision problem of a pig breeder in a period of a few months, before pigs are sold. A pig may be with or without a disease, affecting its final market value. Once a month, the pig is tested in order to detect the presence of the disease. The problem is that the test is subject to error, with certain specified probabilities. Once the test is performed, the pig breeder decides whether to treat the pig by injecting a drug, with a specified cost. The injection changes the probability of the health status of the pig in the next months. The result is the ID in Figure 1. Note that Decision 2 does not depend on the first test on health status, which instead is necessary in an ID.

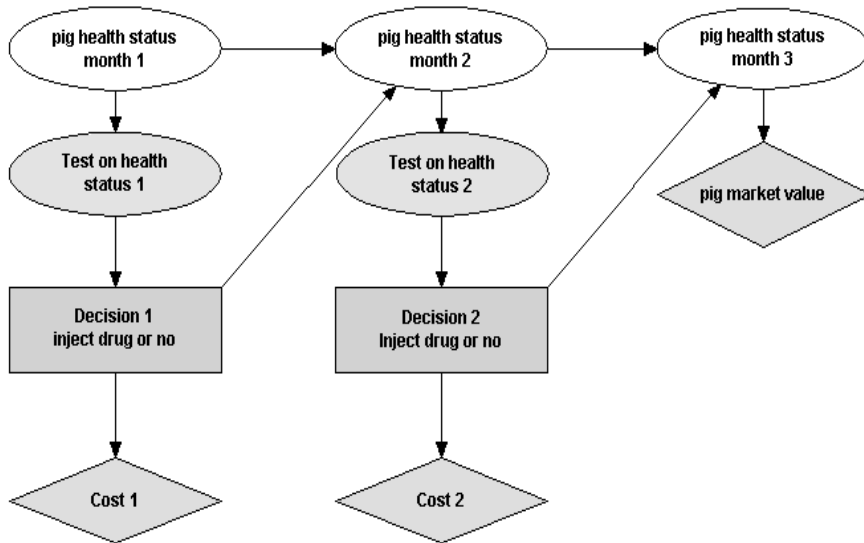


Figure 1 – Pig breeder decision problem (Lauritzen and Nilsson, 2001)

2.2 How Influence Diagrams work

As already stated in Section 2.1, IDs join the BN feature of an easy propagation of evidences with the search of the solution for the decision problem. In this section we show how these two aspects interact (see also Jensen, 2001, Neapolitan, 2004).

In decision analysis, the objective is the definition of a *policy*, *i.e.* of a set of decisions, one for each decision node. The policy is chosen by an optimisation procedure, *i.e.* maximizing the expected utility associated to the policy itself. The expected utility of each policy is found summing the expected utilities of each utility node. For instance, a policy in the ID of Figure 1 may be (Decision 1: do not inject; Decision 2: inject), and its associated expected utility is given by the sum of the expected utilities resulting from the three utility nodes, namely Cost 1, Cost 2 and pig market value. Note that the first and second utility nodes do not depend on random variables, but are fixed once a decision is taken (e.g. the utility of no injection is 0, while the utility of an injection is the negative cost of the

injection). The final utility node (the pig market value) depends on a chance node, which cannot be decided by the decision maker. Policy makers should choose policies that maximize the chance of having a healthy pig. If $u(\text{healthy})$ and $u(\text{ill})$ are the final market values of a healthy and an ill pig respectively, the expected utility is computed with respect to the probability distribution of the pig health status in month 3.

In general, let U be a utility node, \mathbf{D} and \mathbf{C} be its parents partitioned respectively in decision and chance nodes. Let \mathbf{d} be a policy, and \mathbf{c} be a set of states for \mathbf{C} given its parents $pa(\mathbf{C})$, with probability $P(\mathbf{C}=\mathbf{c}|pa(\mathbf{C}))$. Further, let $u(\mathbf{d},\mathbf{c})$ be the utility of U corresponding to each state \mathbf{c} of its parents. Then, the expected utility for U is

$$\sum_{\mathbf{c}} u(\mathbf{d},\mathbf{c})P(\mathbf{C} = \mathbf{c} | pa(\mathbf{C})),$$

the overall expected cost of a policy \mathbf{d} is

$$EU(\mathbf{d}) = \sum_U \sum_{\mathbf{c}} u(\mathbf{d},\mathbf{c})P(\mathbf{C} = \mathbf{c} | pa(\mathbf{C})),$$

and the objective is to find that policy \mathbf{d} for which $EU(\mathbf{d})$ reaches its maximum. For the ID in Figure 1, this corresponds to finding the policy that maximizes the expected final market value of the pig minus the cost of the two decisions about injection.

The updating properties of BNs are essential in finding a solution, because some chance nodes may become manifest during the decision process. For instance, the result of the “Test on health status 1”, once known, influences the probability distribution of the “Pig health status in month 1” which changes the next probability distributions on the pig health status, thus influencing the final expected market value of the pig. For the algorithms used to update IDs, see Jensen (2001). In the following section, these aspects are applied in a responsive design.

The networks are built and analyzed with Hugin 6.4 (see Madsen *et al.*, 2003).

3. Influence Diagrams as a Diagnostic Tool for Responsive Designs

3.1 A toy example

Heeringa and Groves (2004) show that the analysis of paradata may suggest changes in the survey process in many different ways. In what follows, we consider the following simplified situation.

The final quality of a survey can be classified as high (if it meets the target) or low (if it does not). We split the survey in two distinct phases in which an indicator of the process flow is observed: response rate or paradata useful to evaluate the presence of bias due to possible missingness, etc. Although paradata can be the result of a continuous variable, the behaviour of survey managers is often based on a finite set of “warning levels”. In the following we will restrict to three warning levels, from warning level 1 (low probability of obtaining this level when the survey quality is low) to warning level three (high probability of obtaining this level when the survey quality is low).

The observation of the paradata may suggest to continue the survey without any change, or to change the survey plan, e.g. increasing the sample size or changing the contact phase to a more precise (and expensive) approach. For the sake of simplicity, we will consider these two situations. After having observed paradata in the first phase, the survey planner may decide

- to add 0 sampling units (Action A: cost=0);
- to add 100 sampling units (Action B: cost=50) or;
- to add 200 sampling units (Action C: cost=200).

After having observed paradata in phase two, the survey planner may still decide:

- to add 0 sampling units (Action D: cost=0);
- to add 100 sampling units (Action E: cost=100) or;
- to add 200 sampling units (Action F: cost=200).

The final quality of the survey may still be low or high. If it is low, the survey is not of use unless an additional survey costing 800 is made.

The decision that the survey planner must take in the two phases of the survey can be described as a decision problem, and formalized by an influence diagram.

3.2 The Influence Diagram for the motivating example

The previous example can be easily modelled as a LIMID, and in particular, its structure coincides with that of Figure 1.

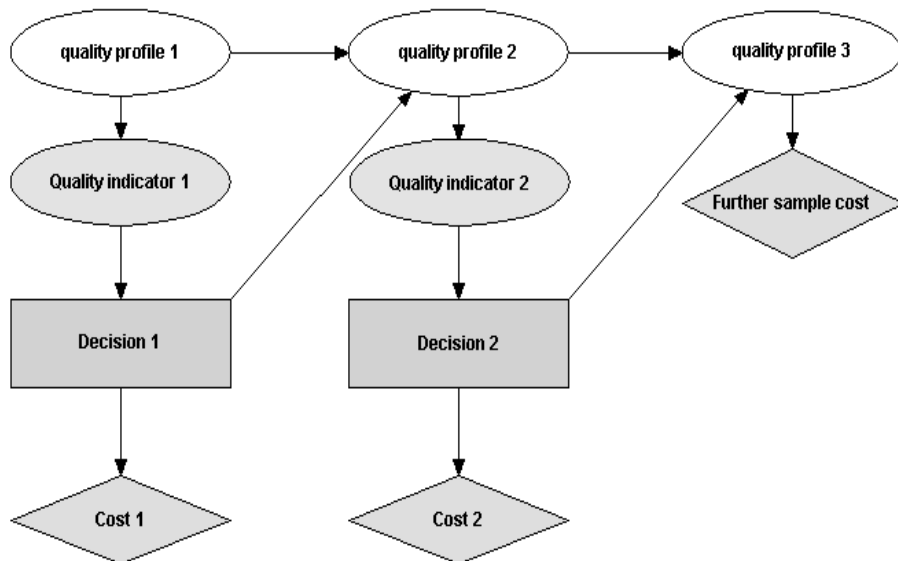


Figure 2 - Influence diagram for the example of Section 3.1

The chance nodes are the quality profile and the quality indicator. These nodes are indexed by the different survey phases.

- “quality profile 1” distribution reflects the feelings, opinions or knowledge of the survey planner about the final quality profile at initial stage;
- “quality profile 2” distribution is that induced by the decision in the first phase;
- “quality profile 3” distribution is that induced by the decision in the second phase.

Note that both “quality profile 2” and “quality profile 3” also depend on the status of the quality profile in the previous phase. The probability distributions for these variables are given in Figure 3. Note that when 100 or 200 more units are added to the sample it is more probable that the quality profile is high.

“Quality indicator 1” and “quality indicator 2” represent the warning level detected in the two phases. These variables depend only on the corresponding quality profile.

The costs associated with the decisions are shown in Figure 4.

quality profile 1

Low	0.99
High	0.01

quality profile 2

Decision	A		B		C	
	Low	High	Low	High	Low	High
Low	0.95	0.1	0.25	0.01	0.05	0.0
High	0.05	0.9	0.75	0.99	0.95	1.0

quality profile 3

Decision 2	D		E		F	
	Low	High	Low	High	Low	High
Low	0.95	0.1	0.25	0.01	0.05	0.0
High	0.05	0.9	0.75	0.99	0.95	1.0

Quality indicator 1

quality profil:	Low	High
warning lev 1	0.05	0.8
warning lev 2	0.3	0.15
warning lev 3	0.65	0.05

Quality indicator 2

quality profil:	Low	High
warning lev 1	0.05	0.8
warning lev 2	0.3	0.15
warning lev 3	0.65	0.05

Figure 3 - Probability distributions of the chance nodes of ID in Figure 2

Cost 1

Decision	A	B	C
Utility	0.0	-50.0	-200.0

Cost 2

Decision 2	D	E	F
Utility	0.0	-100.0	-200.0

Further sample cost

quality profil:	Low	High
Utility	-800.0	0.0

Figure 4 - Costs associated to the decision nodes and to the final quality result

The expected overall cost is given by the sum of the cost of decisions 1 and 2 plus the expected “Further sample cost”, which depends on the final quality profile. The situation in the absence of any evidence is given in Figure 5.

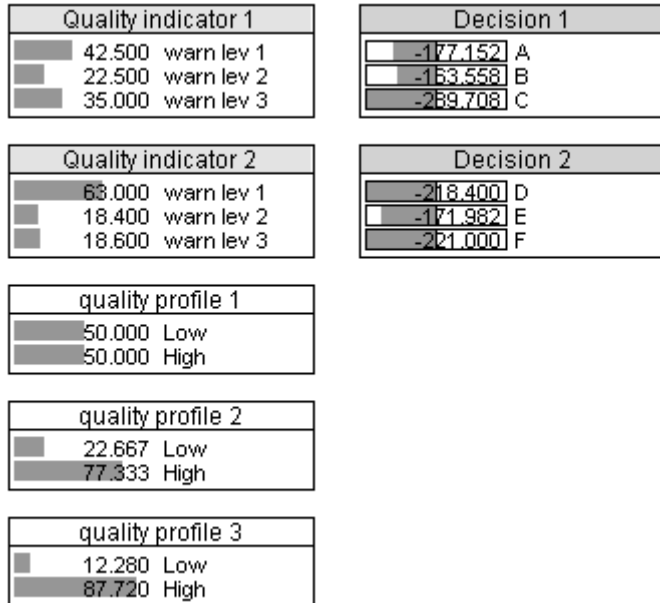


Figure 5 - expected costs in the absence of any evidence

Note that, in the absence of any evidence, and for uniformly distributed “quality profile 1”, the decision is to add 100 units in both phase 1 (corresponding to B for decision 1) and phase 2 (corresponding to E for decision 2). Changes in the initial distribution naturally lead to different results. Expected costs are lower if “quality profile 1” is not uniformly distributed but give higher probability to initial high quality profile.

We now show the optimal decisions when paradata are available.

Assume that “Quality indicator 1” is at warning level 1. This evidence propagates through the network updating the distributions and modifying the probability of “Quality profile 3”. Hence, the expected cost also changes. In this case (Figure 6), it seems better to avoid increasing the

sample size in the first phase with an expected overall cost (for A of Decision 1) of 119.27 (whereas in phase 2 Decision 2 remains E).

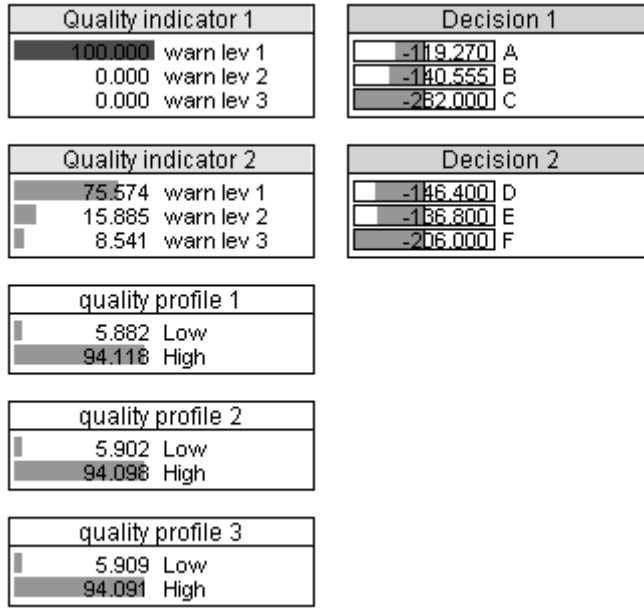


Figure 6 - expected costs when “Quality indicator 1” shows warning level 1 – dark bar corresponds to conditioning.

Now assume that the warning level of “Quality indicator 1” is 3, that in the first phase the decision taken is to add 100 units (*i.e.* action A of Decision 1). If the warning level of “Quality indicator 2” is still 3, the best choice is to add 200 more units (*i.e.* action F of Decision 2), with an expected overall cost of 281.91 (Figure 7).

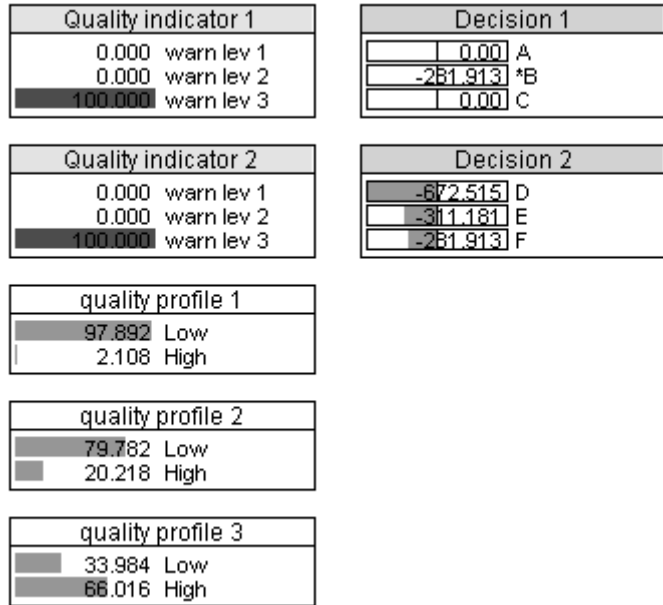


Figure 7 - expected costs when quality indicators 1 and 2 show warning level 3, and when the first decision is to add 100 sample units; dark bar corresponds to conditioning.

4. Further Developments and Extensions

The previous example is just a toy example, and needs further developments in order to be practically used in real cases.

First of all, we have restricted to just one aspect of the survey quality profile, while quality has many different interrelated aspects: missingness, bias, efficiency and timelines. Both these quality aspects and the corresponding paradata studied in order to detect possible problems should be defined in an appropriate multivariate model.

Furthermore the specification of all the elements in the ID for a real case should be appropriately investigated. For instance the effect of discretization

of indicators, the form of the utility function and the sensitivity to initial distributions should be analysed.

Moreover, the possibility to estimate all the distributions in an ID from previous survey processes should be investigated.

Acknowledgements

The authors are grateful to Julia Mortera for fundamental suggestions and comments. This work has been partially supported by MIUR grant PRIN05 “Statistical analysis of complex problems in the presence of incomplete information: statistical methodologies and applications” and MIUR grant PRIN2005 “The statistical information in agriculture: present needs and future developments”.

References

- Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999): *Probabilistic Networks and Expert Systems*, Heidelberg: Springer.
- Hansen, S. E., and Maher, P. (2005), “Using Process Data for Responsive Design”, paper presented at the XXII International Methodology Symposium, Ottawa, Canada, October 25-28 2005.
- Heeringa, S. G., and Groves, R. M. (2004), “Responsive Design for Household Surveys”, *Proceedings of the Survey Research Methods Section, American Statistical Association*.
- Jensen, F. V. (2001), *Bayesian Networks and Decision Graphs*, New York: Springer.
- Lauritzen, S. L. (1996), *Graphical models*, Oxford: Oxford University Press.
- Lauritzen, S. L., and Nilsson, D. (2001), “Representing and Solving Decision Problems with Limited Information”, *Management Science*, 47, pp. 1235-1251
- Madsen, A., Lang, M., Kjaerulff, U., and Jensen, F. (2003), “The Hugin tool for learning Bayesian Networks” *Proceedings of the 7th European Conference, ECSQARU 2003 - Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, Nielsen T. e Zhang N., eds., pp. 594–605, Heidelberg: Springer.
- Neapolitan, R. E. (2004), *Learning Bayesian Networks*, Upper Saddle River (NJ): Prentice Hall