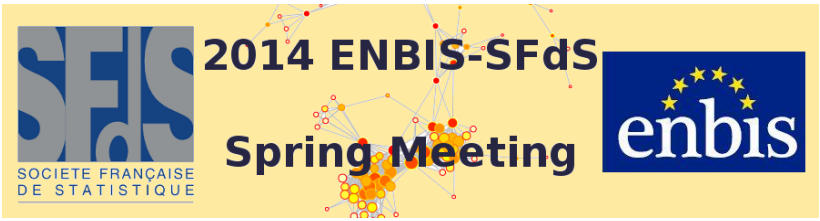
Consumers' satisfaction with railway transport: a Bayesian Network approach

Giovanni Perucca and Silvia Salini

Paris, 10th April 2014



Research motivation

In the last years the interest for consumers' experience with Services of General Interest (SGI) largely increased. Several surveys have been conducted in order to observe how consumers' satisfaction differs across EU countries.

Judgements are based on personal perceptions and evaluations which, in turn, depend on several unobservable and subjective factors, such as individual characteristics, group-specific features and social norms, as pointed out by several contributions (McFadden *et al.,* 2005).

Many studies (Fiorio and Florio 2010) already focused on the connections between consumers' satisfaction with SGI, respondents' characteristics and socio-economic indicators. This literature usually makes use of econometric models for categorical dependent variable.

Our work tries to analyse the same issue through Bayesian Networks (Kenett and Salini 2009) and to compare the results of these two methodologies.

The data

Eurobarometer surveys provides every semester large data sets which analyse and investigate individual attitudes and perception across all European countries.

In 2000, 2002 and 2004 a survey has been devoted (among other issue) to railway transport: a sample of respondents has been asked to state their judgements about prices (defined as excessive, unfair, fair) and service quality (defined as very unsatisfied, fairly unsatisfied, fairly satisfied, very satisfied).

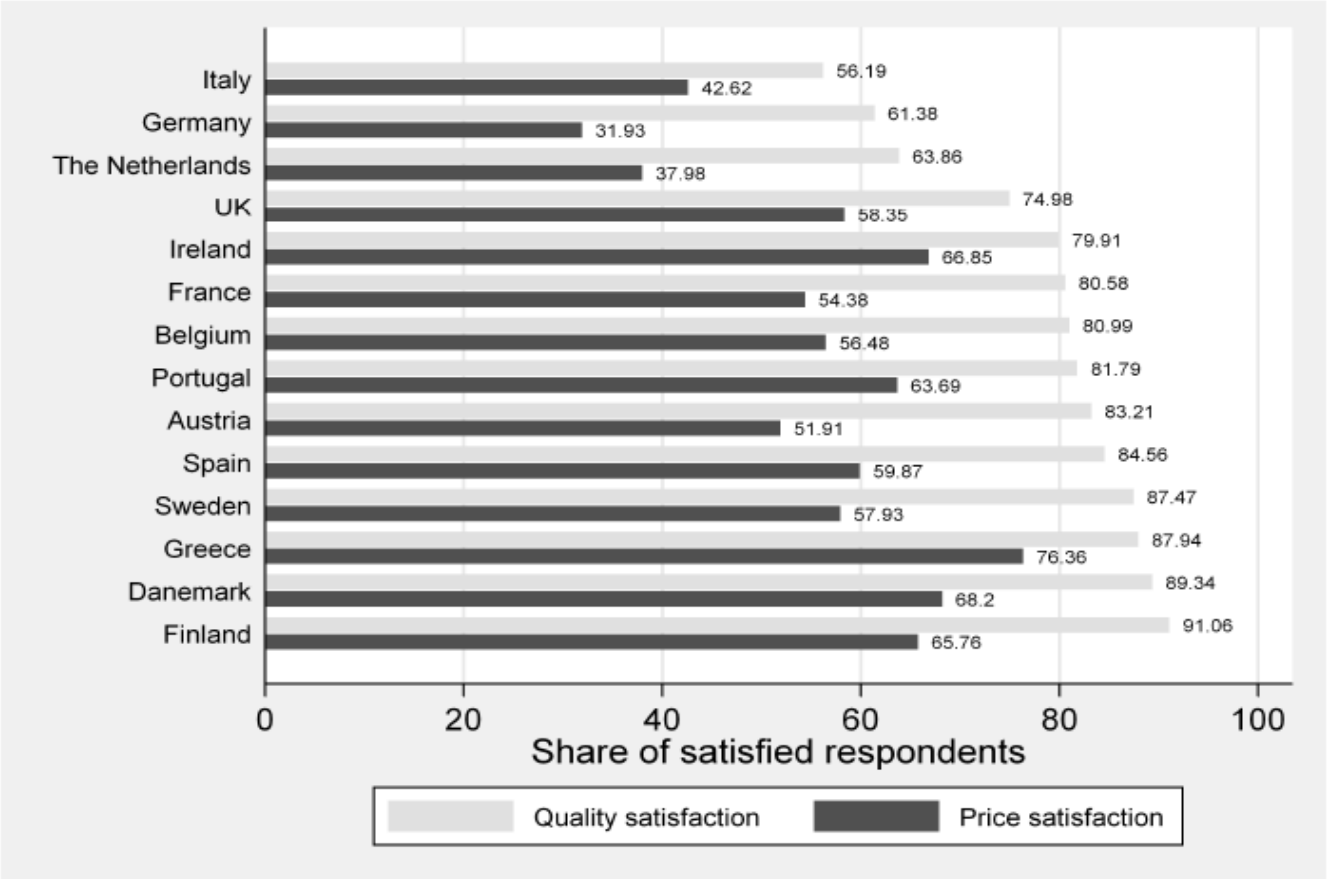
Moreover, Eurobarometer surveys provide us with a large amount of information about respondents' characteristics (such as age, gender, education) and about their attitudes toward railway transport.

Our sample is made up by 19,458 observations from 14 European countries.

We split our data set in two sub-samples. We estimated our models on the first one (90% of the observations) and we used the second sample (10% of the observations) for testing the results.

The data

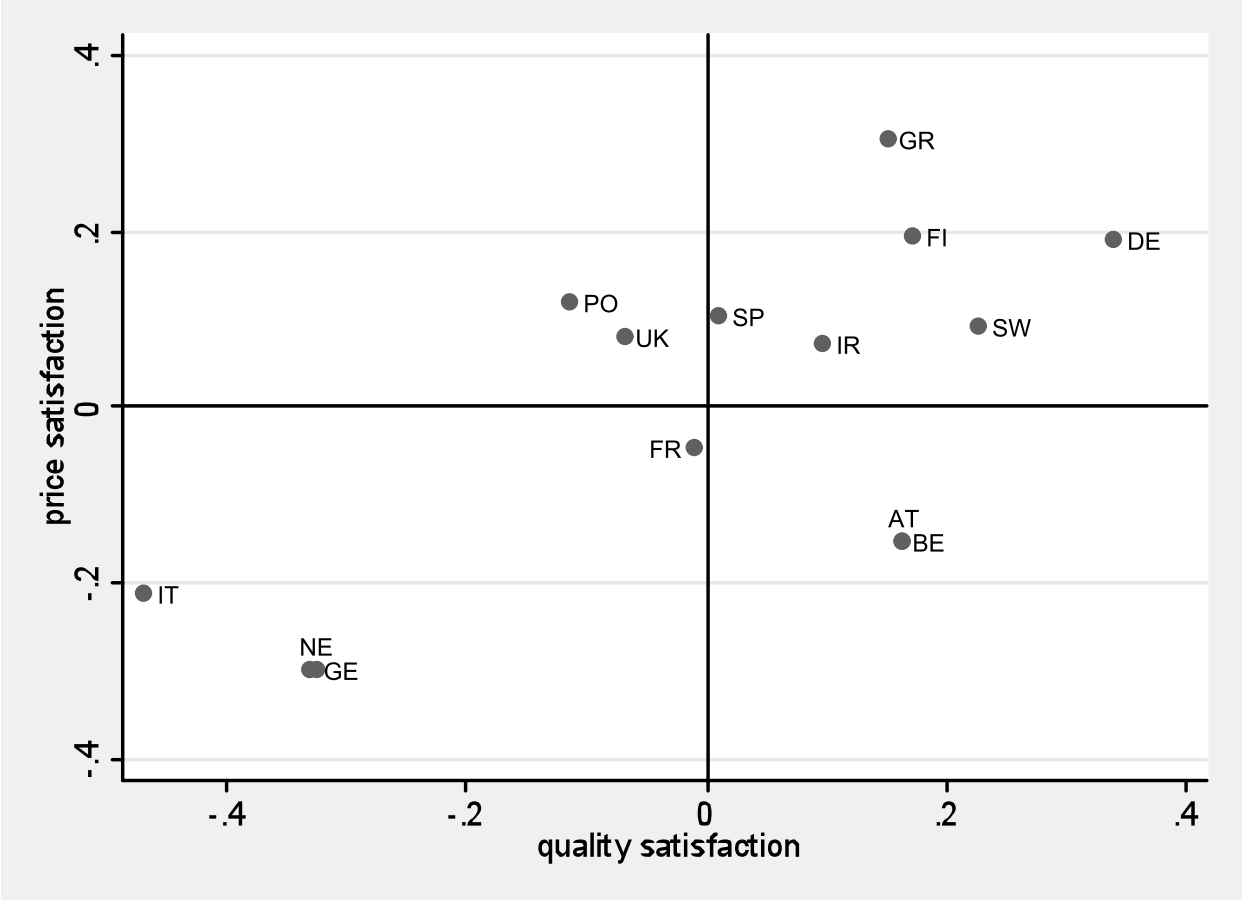
Share of satisfied respondents in EU countries (2000, 2002 and 2004).



Source: our elaborations on Eurobarometer data

The data

Patterns of satisfaction significantly differ across countries.



Departures from the mean of price and quality satisfaction in EU countries.

Ordered logistic regression

As a first step, following previous literature on the topic (Fiorio and Florio, 2010), we analyse the problem by applying econometric models.

The true level of satisfaction can be represented as a continuous, unobserved latent variable , which is equal to:

$$Y^* = x\beta + e$$

Instead of *Y** we observe *Y* , our ordinal variable which can be seen as a collapsed version of *Y** whose value depends on whether or not the continuous variable crossed a particular, unknown, threshold

$$Y_i = 0 \quad \text{if} \quad Y_i^* \le \alpha_1$$
$$Y_i = 1 \quad \text{if} \quad \alpha_1 \le Y_i^* \le \alpha_2$$
$$\vdots$$
$$Y_i = M \quad \text{if} \quad Y_i^* > \alpha_M$$

Ordered logistic regression

Then, we can estimate the probability that Y will take on any particular value through standard **ordered logistic regression**

$$P(Y_i > j) = g(x\beta) = \frac{\exp(\alpha_j + x_i\beta)}{1 + \left[\exp(\alpha_j + x_i\beta)\right]}, j = 1, 2, ..., M - 1$$

Model I

y_i= price satisfaction

Model II

y_i= quality satisfaction

socio-demographic characteristics
(age, gender, education, job, marital status, political views)

macroeconomic indicators at regional level (NUTSII)
(per capita income, unemployment rate, population density)

"Objective" quality indicators
(fares, demand, railway length)

Country dummies

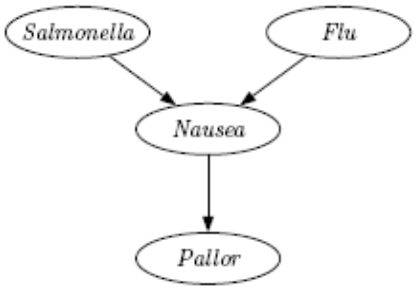
Ordered logistic regression

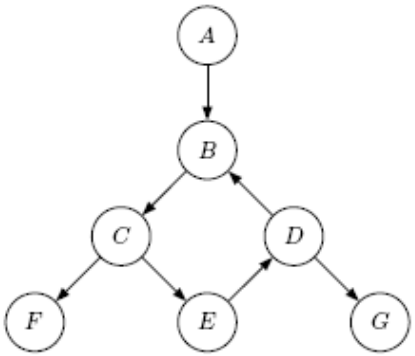
	Dep. var.: price satisfaction			Dep. var.: quality satisfaction		
	1	2	3	4	5	6
Female	0.813***	0.812***	0.813***	1.065**	1.066**	1.068**
Age	1.007***	1.007***	1.007***	1.005***	1.005***	1.005***
Low education	0.888	0.885	0.877	1.269**	1.273**	1.258**
Medium education	1.054	1.052	1.046	1.120	1.127	1.121
High education	1.037	1.038	1.036	0.970	0.973	0.972
Separated/divorced	0.849***	0.852***	0.854***	1.017	1.013	1.017
Unmarried	0.933	0.938	0.941	0.970	0.970	0.975
Widow	1.144*	1.150*	1.149*	0.941	0.941	0.942
Unemployed	0.770***	0.773***	0.780***	0.994	0.994	0.996
Manual worker	0.862**	0.863**	0.861**	0.943	0.943	0.941
House person	0.907	0.912	0.911	1.102	1.105	1.107
Self employed	0.908	0.912	0.911	0.935	0.937	0.939
Employee	0.859**	0.861**	0.864**	0.866**	0.865**	0.871**
Vote intention: centre	1.186***	1.187***	1.185***	1.084**	1.084**	1.083**
Vote intention: right	1.136***	1.138***	1.133***	1.058	1.059	1.058
Information: clear	2.105***	2.103***	2.102***	2.882***	2.876***	2.873***
Complaint in the last 12 months	0.785***	0.782***	0.781***	0.483***	0.474***	0.475***
Easy access: yes	1.344***	1.352***	1.354***	3.721***	3.725***	3.742***
Quality satisfaction	2.284***	2.277***	2.273***			
Price satisfaction				2.434***	2.430***	2.425***
Railway indicators						
Railway fares		0.993***	0.995*		1.077***	1.078***
Demand		1.002**	1.001*		1.000	1.000
Railway length		1.039***	1.041***		1.006**	1.008***
Macroeconomic indicators						
Unemployment			0.980***			0.988**
Income			0.973**			0.968***
Population density			1.000			1.000
Year: 2002	1.147***	1.049	1.063	0.991	0.833***	0.852***
Year: 2004	1.283***	1.148**	1.188**	1.053	0.809***	0.848***
Country dummies	yes	yes	yes	yes	yes	yes
Constant	yes	yes	yes	yes	yes	yes
Observations	17,529	17,529	17,529	17,529	17,529	17,529

Bayesian networks

Another way to look at the same issue involves Bayesian networks (Pearl, 2000).

A Bayesian network is a graphical model that encodes the joint probability distribution (physical or Bayesian) for a large set of variables.





A Bayesian network consists of the following (Jensen and Nielsen, 2007):

1. A set of variables and a set of directed edges between variables
2. Each variable has a finite set of mutually exclusive state
3. The variable together with the direct edges form an acyclic directed graph; a directed graph is acyclic if there is no directed path $A_1 \rightarrow \ldots \rightarrow A_n$ so that $A_1 = A_n$
4. To each variable A with parents $B_1, \ldots B_n$, a conditional probability table $P(A \mid B_1, \ldots, B_n)$ is attached.

Bayesian networks in Customer Surveys

An early attempt to apply Bayesian Networks for the analysis of customer surveys was presented in Kenett and Salini (2009) and Salini and Kenett (2009).

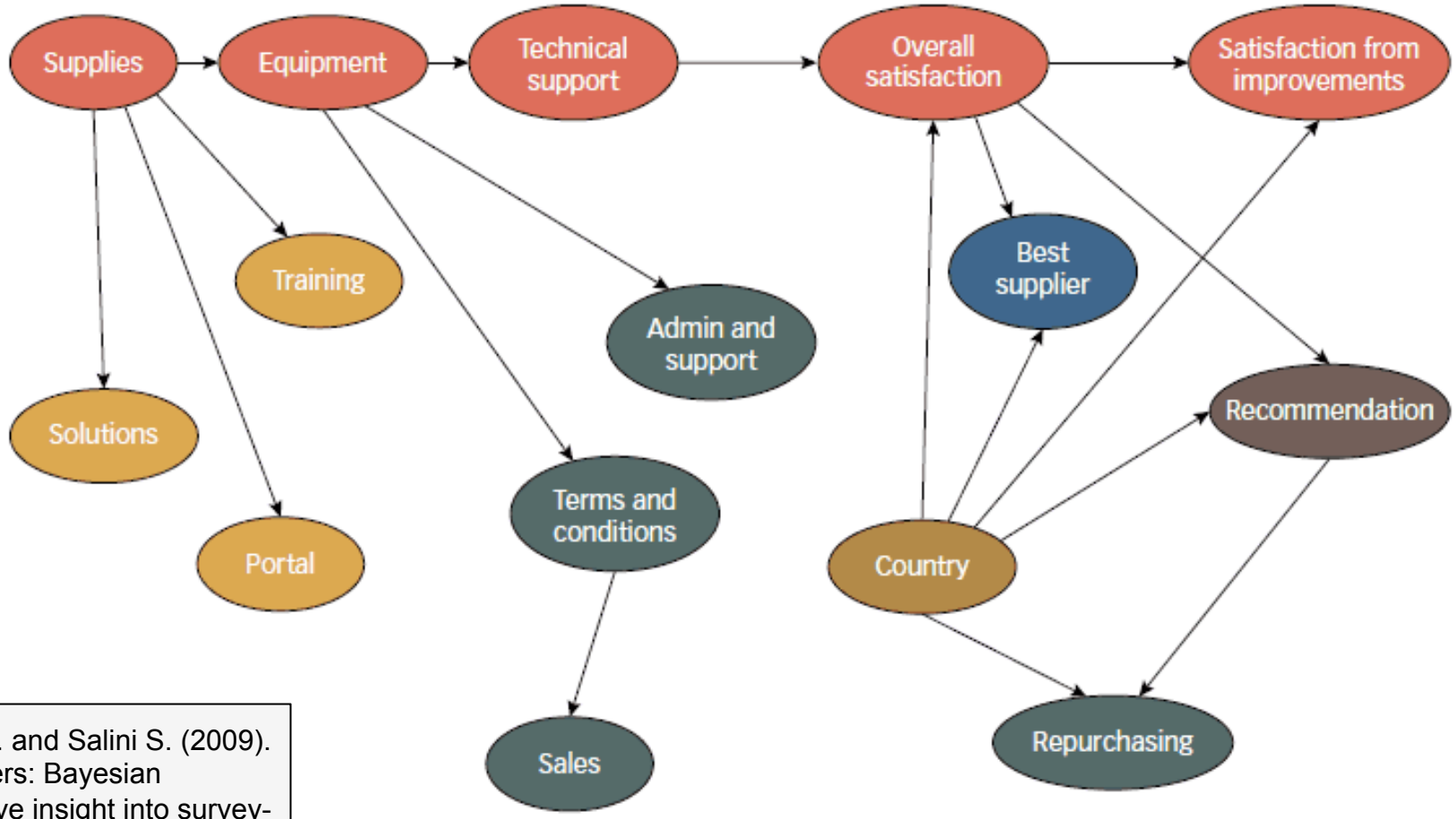
A survey with n questions produces responses that can be considered as random variables, X1, . . . , Xn.

Some of these variables, q of them, are responses to questions on overall satisfaction, recommendation or repurchasing intention, that are considered target variables.

Responses to the other questions, X1,. . . , Xk , k = n-q , can be analyzed under the hypotheses that they are positively dependent with the target variables.

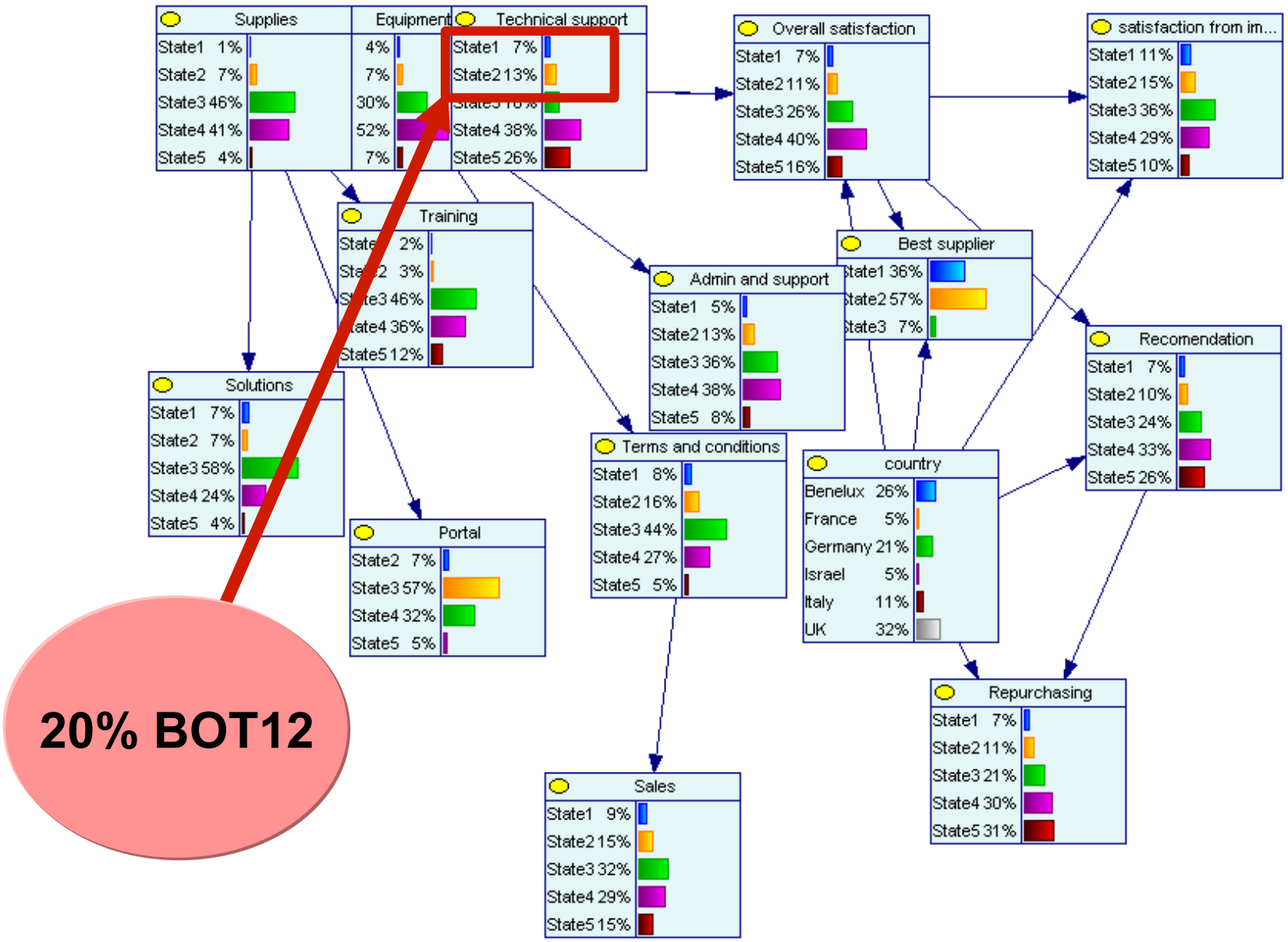
The combinations (Xi, Xj), Xi $\in$ X1, . . . , Xn-q, Xj $\in$ Xn-q+1, . . . , Xn , are either positive dependent or independent, for each pair of variable (Xi, Xj), $i \leq n-q$, $n-q < j \leq n$.

Bayesian networks in Customer Surveys



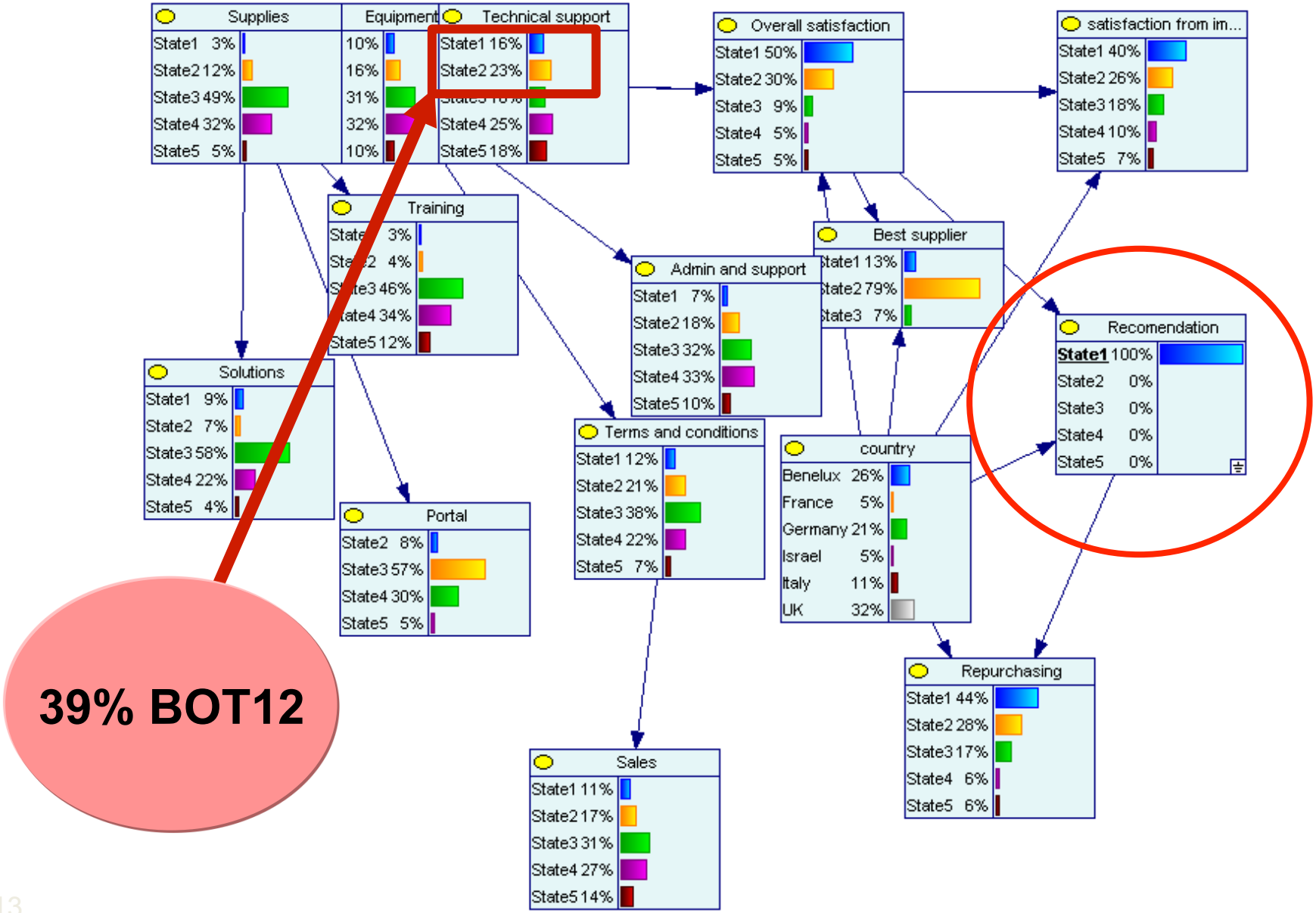
Kenett, R.S. and Salini S. (2009). New Frontiers: Bayesian networks give insight into survey-data analysis, *Quality Progress*, pp. 31-36, August.

Bayesian networks in Customer Surveys



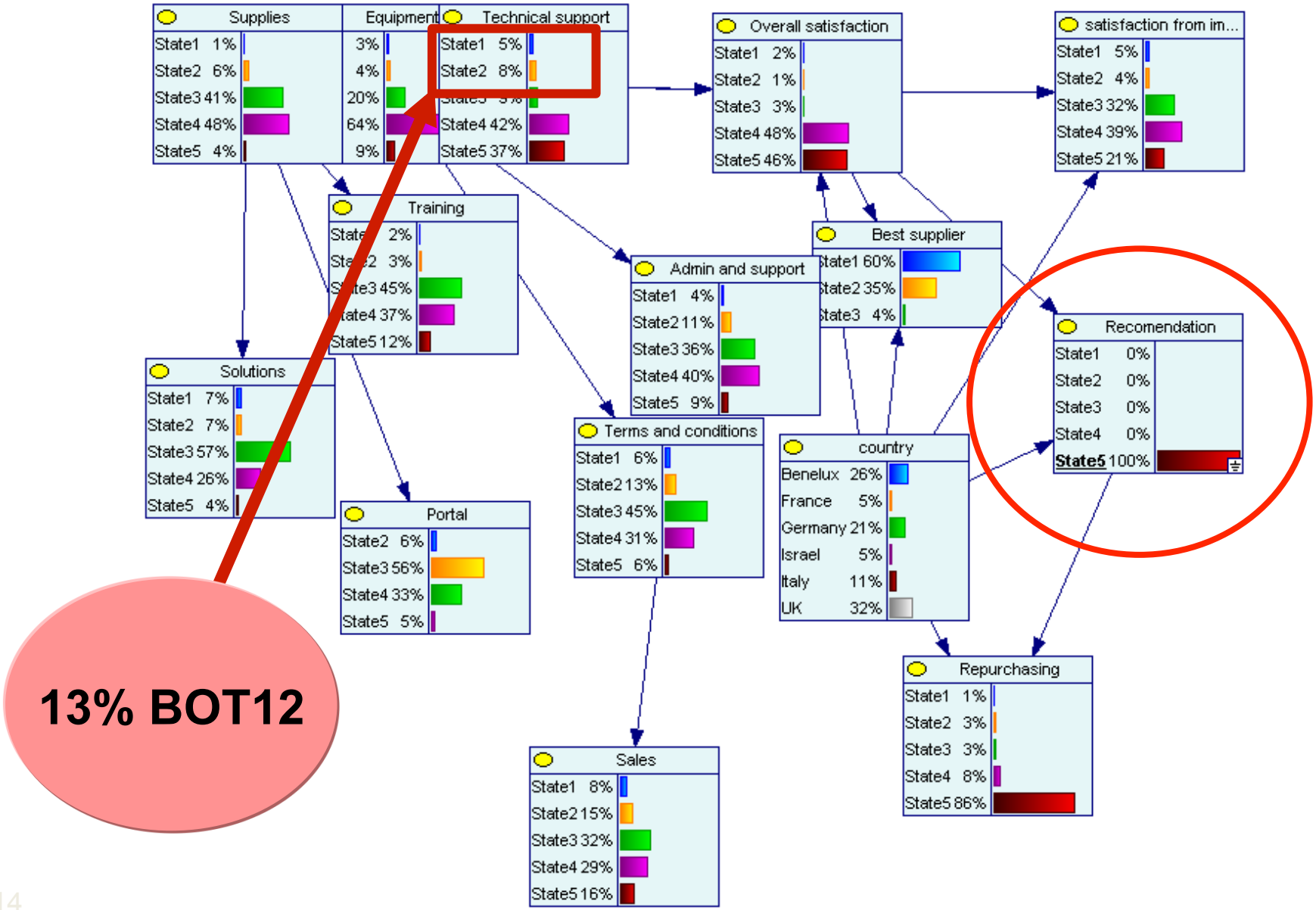
12

Bayesian networks in Customer Surveys



13

Bayesian networks in Customer Surveys



14

Bayesian networks: the package

The *R* package provides a set of tools devoted to the analysis and estimation of Bayesian networks.
In this analysis we used the library *bnlearn* (Scutari, 2010), which allows to test several learning algorithm on our data.

The disadvantage of this analysis relies on the impossibility to mix in our data set continuous variables (such as age) and factors. Even if some libraries (*deal*, Bottcher and Dethlefsen, 2007) handle networks with mixed variables, their learning procedure is and hardly applicabile to complex models.

Hence, we recoded continous variables (age and macro indicators) as factors.

Bayesian networks: robustness issues

We implemented in our analysis all the 11 algorithms available in *bnlearn*. Results significantly differ according to the algorithm implemented.

Hence, we tried to choose the best network, also based on the table (partly) reported below.

		Score based Alg.				Constrained based Alg.				Hybrid Algorithms			TOT.
Nodes (fathers)	Nodes (children)	HC (bic)	HC (aic)	TABU (bic)	TABU (aic)	GS	IAMB	FAST IAMB	INTER IAMB	MMHC (bic)	MMHC (aic)	PHM	
Age	Educ	1	1	1	1								4
	Occupation	1	1	1	1	1	1		1	1	1	1	10
	Marital	1	1	1	1	1	1	1	1	1	1	1	11

Some algorithms tend to overestimate the number of arcs in the network.

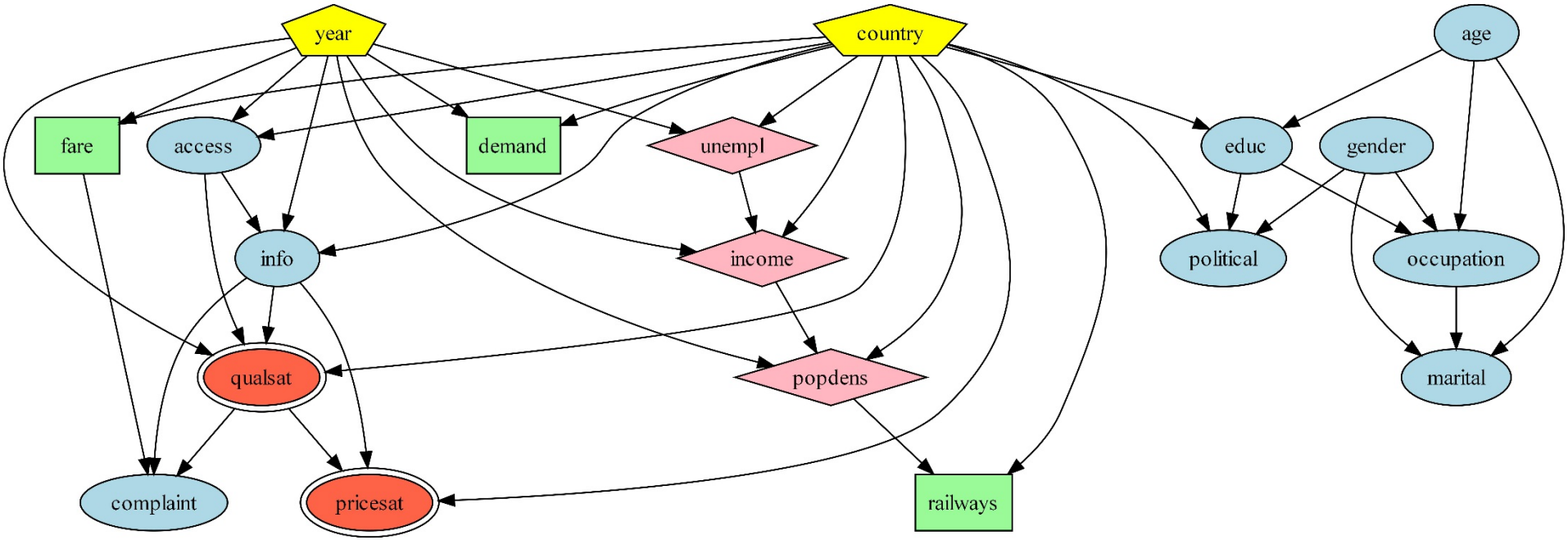
We chose the BN including all the most robust connections between nodes and minimizes the number of "weak" arcs.

Bayesian networks: the model

We select the Hill-Climbling learning algotithm

Learning algorithm	Nodes	Arcs	Undirected arcs	Directed arcs	BIC score
Hill-Climbing	19	31	0	31	-353662.1
Tabu Search	19	31	0	31	-379004.8
Grow-Shrink	19	29	3	26	n.a
Incremental Association	19	32	3	29	n.a
Fast Incremental Association	19	27	2	25	n.a
Interleaved Incremental Association	19	32	3	29	n.a
Max-Min Hill-Climbing	19	33	0	33	-400703.4
Two-Phase Restricted Maximization	19	23	0	23	-405967.7

Bayesian networks: main results



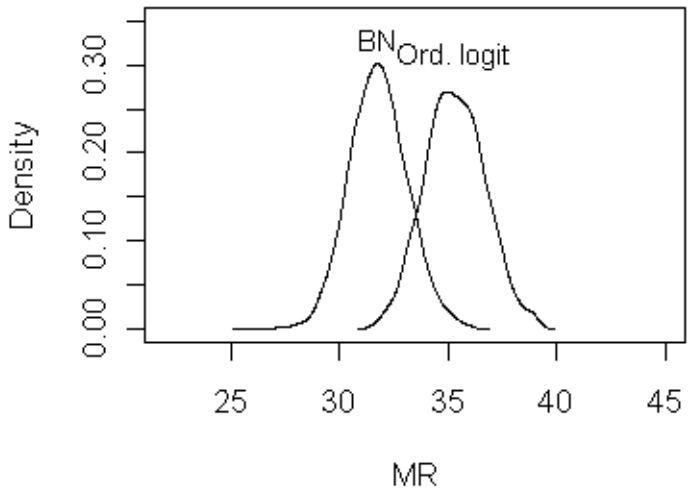
Ordered logistic regression vs BN

The predictive performance of the models has been tested by applying cross-validation.

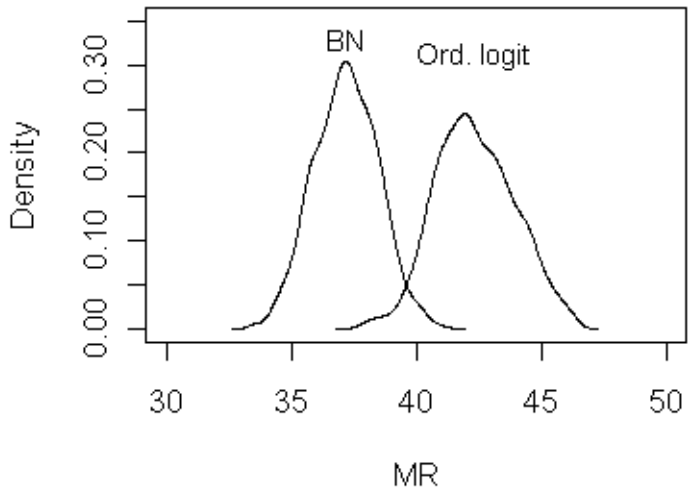
A bootstrap resampling procedure on the initial data set was performed. We generated 1,000 random subsets, each of them made up by 1,000 observations.

Then, the misclassification rate (MR) of the two models (ordered logit and BN) was estimated on each random sample.

MR distribution for price satisfaction



MR distribution for quality satisfaction



In terms of predictory capability, the two methods almost lead to the same results.

Ordered logistic regression vs BN

However, from the point of view of a policy the messages conveyed by these two methods significantly differ.

The **ordered logit model** suggests that price and quality satisfaction are related to a number of items, including individual characteristics and some features of the service provided, but it does not help us understand where the action of the policy maker should be focused.

It associates the variables to each other, pointing out the sign and the strength of these linkages, without providing an extensive explanation of the cause/effect relationships between them.

Ordered logistic regression vs BN

What are the causes of satisfaction?

Finding an answer to this question is the most relevant information from the planner's perspective.

Even if the issue concerning "explanatory" vs. "predictive" modelling is often ignored in the economic and social science literature, its relevance is straightforward, especially in an empirical study not supported by a robust theoretical framework (Shmueli, 2010).

Ordered logistic regression vs BN

The **BN** provides more information with regard to clarifying this direction:

-For example, that price satisfaction is directly affected by quality satisfaction, while the reverse does not hold.

-A reduction of fares would not be accompanied by an increase in quality satisfaction.

-A policy aimed at increasing the quality of the service (such as improving access to the railway facilities or enhancing the information provided to the travellers) would be more effective, since it would allow for generating positive spill-overs on both price and quality satisfaction.

Working Progress and Further research

- BN in the framework of IPA approach and Sensitivity Analysis (with F. Cugnata and R. Kenett)
- R function to conduct "what if" sensitivity scenarios given a BN (with F. Cugnata)
- Bootstrap confidence intervals for performance indicators in the value-added models in Education (with F. Cugnata and G. Perucca)

Importance-Performance Sensitivity Analysis

How are the factors in the model affecting the target variable levels?

Which factors affect customer satisfaction and how

How are factors affecting the distance between conditioned distributions and observed data?

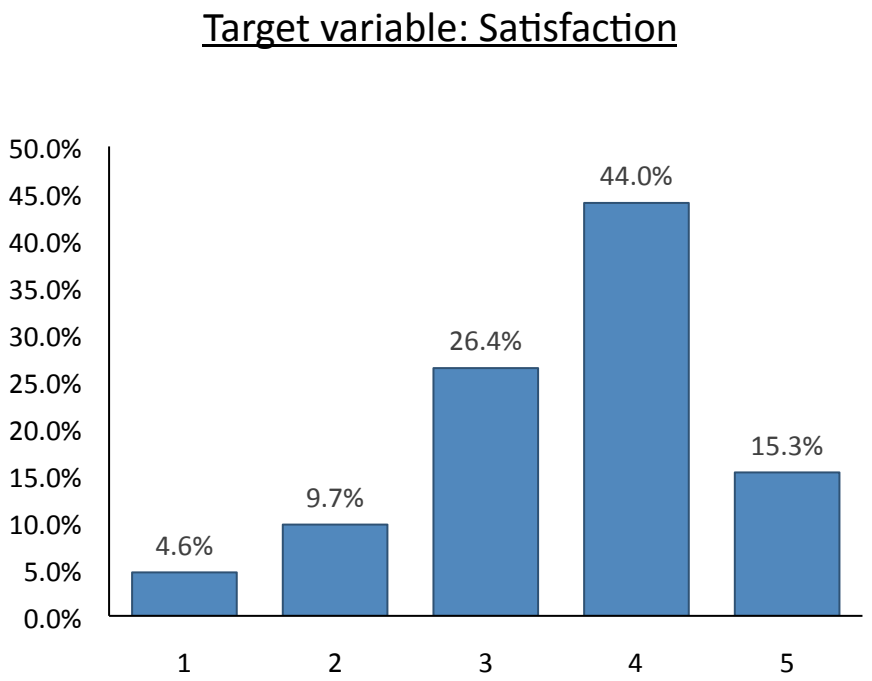
Which factors will have the largest impact if changed

Cugnata, F., Kenett, R.S. and Salini S. (2014). Bayesian Network Applications to Customer Surveys and InfoQ, *to appear in Procedia Economics and Finance*

Importance-Performance Sensitivity Analysis

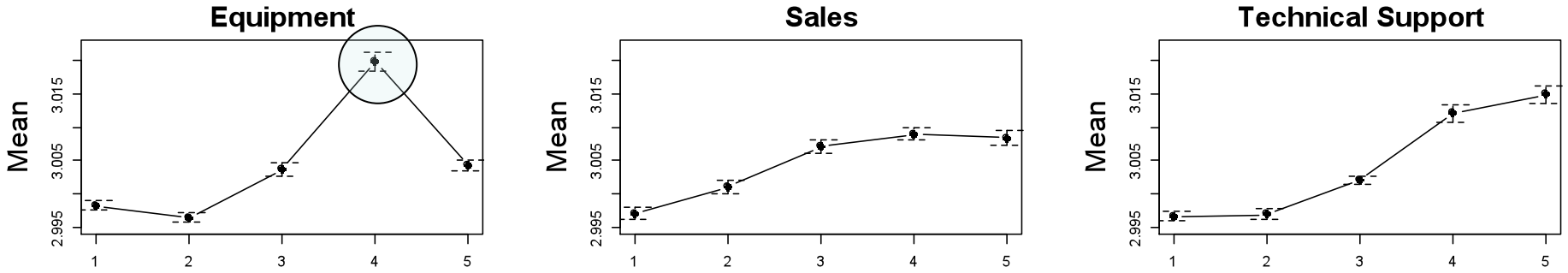
Generate a full factorial experiments based on driver combinations.

Analyse the target variable thus generated (overall satisfaction) with respect to the observed one, in order to study the effect that each driver combinations has on its variability.

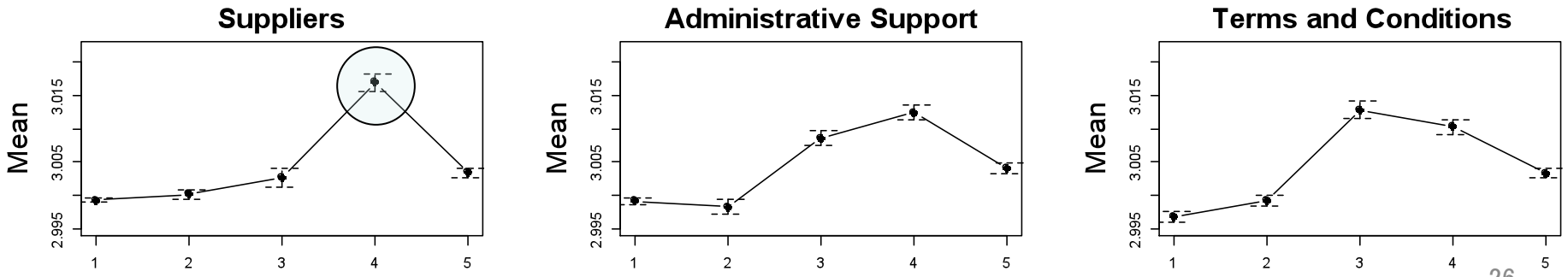


Importance-Performance Sensitivity Analysis

1000 simulated target variable using the BN estimated on the observed dataset.
Distribution of simulated target variable for each level of each variable



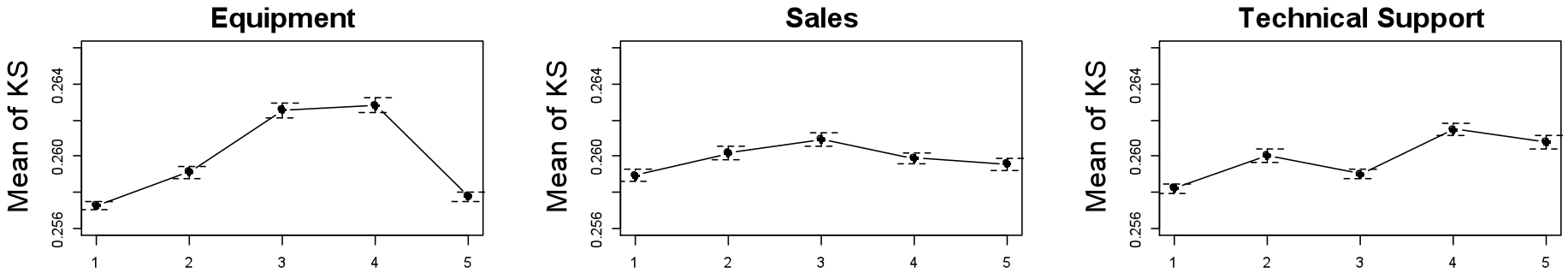
Impact of factors on target variable



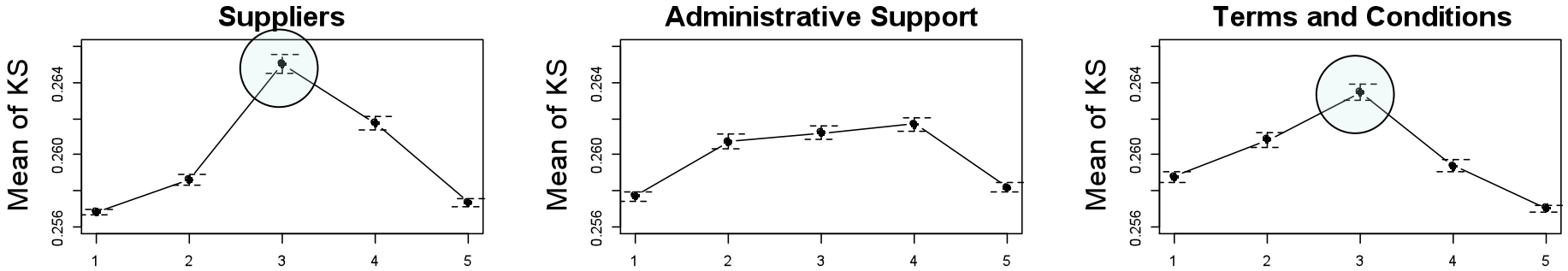
Mean of simulate target variable for each level of each variable (ABC)

Importance-Performance Sensitivity Analysis

Compare the observed distribution with the simulated one (1000 runs) using Kolmogorov-Smirnov test



Impact of factors on conditioned target variable distribution



Mean of Kolmogorov-Smirnov for each level of each variable (ABC)

Thank you for your attention