INSPECTING THE QUALITY OF ITALIAN WINE THROUGH CAUSAL REASONING

Eugenio Brentari [1], Maurizio Carpita [1] and Silvia Golia [1]

[1] Department of Economics and Management, University of Brescia (e-mail: eugenio.brentari@unibs.it, maurizio.carpita@unibs.it, silvia.golia@unibs.it)

KEYWORDS: Sensory analysis, Bayesian Networks, Altroconsumo Global Score

1 Introduction

Understanding the mechanisms by which some variables came to take on the values they take and predicting what the values of those variables would be under outside manipulations are the major goals of many sciences. Finding answers to questions about these mechanisms or predicting the value of a variable after an intervention, are characteristic of the causal inference. In this paper the causal reasoning (Spirtes et al., 2000) is applied to the sensory analysis field (Lawless, 2013) in order to study the factors that have a direct influence in determining the quality of the Italian wine. The most frequently used causal models are the Bayesian Networks (BNs) and the Structural Equation Models (SEMs). A BN specifies, for each variable in the network, a density function as a function of the values of its causes, whereas SEM specifies, for each variable, the values of the variable as a function of the values of its causes including some unmeasured noise term; these two models are closely linked, as shown in Spirtes et al. (2000). A BN is given by the pair (G,P), where G is a directed acyclic graph (DAG), and P is a probability distribution which factorizes according to G. The DAG G is composed by a set of nodes (or vertices) V, which correspond to a set of random variables X_V indexed by V, and a set E of directed links (or edges) between pairs of nodes in V. The joint probability distribution P over the set of variables X_V is factorized as

$$P(X_V) = \prod_{\nu \in V} P(X_\nu | X_{pa(\nu)}) \tag{1}$$

where $X_{pa(\nu)}$ denotes the set of parent variables of variable X_ν for each node $\nu \in V$. So a BN can be described in terms of a qualitative component, that is the DAG, and a quantitative component, consisting of the joint probability distribution (1). The construction of a BN runs in two steps. First, one identifies the causal relations among the variables generating a DAG, then the joint

probability distribution has to be specified in terms of the set of conditional probability distributions $P(X_v|X_{pa(v)})$. To find the causal structure represented as a DAG is a problem impossible to solve with observational data only; nevertheless, under suitable assumptions, such as causal sufficiency, causal Markov condition and causal faithfulness condition, causal structures can be retrieved at least up to some equivalence class to which the true DAG belongs. The DAG can be derived either manually or automatically from data, including also partial knowledge about the BN structure. In order to automatically find the BN structure, several algorithms have been proposed in the literature, falling under three broad categories: constraint-based, score-based, and hybrid algorithms. The method used in this paper to derive the DAG belongs to the class of constraint-based algorithms and it is the PC algorithm (Spirtes et al., 2000) implemented in the Tetrad 5.1.0-6 program provided by Spirtes et al. (2010). The PC algorithm is the first practical application of the Inductive Causation algorithm which learns the structure of a BN using conditional independence tests. Given that the dataset under study contains both continuous and ordinal variables, the discretization task for continuous variables has to be taken into account. Discretization must be made properly because it may impact the quality of the learned structure, given that it can in general generate spurious dependencies among variables. The discretization can be made using an "expert" approach, where the choice of the cut points is driven by the knowledge of an expert, or a "statistical" approach, where the choice of the cut points is driven by the characteristics of the probability distribution.

2 The Altroconsumo database

In order to study the determinants that have an influence on the quality of the Italian wines, the following dataset was analyzed. The database was created using the data produced by Altroconsumo, an Italian independent consumer's association, from 2006 to 2012 for its annual publication *Guida Vini*. Each year, about 280 wines were bought and some chemical and sensory characteristics were measured (Brentari and Vezzoli, 2015). The wines were chosen in order to represent the variety of Italian vineyards, producers and region of origin. For each year, different vineyards and producers were considered, so that the observations could be considered as independent. The chemical characteristics were evaluated using the following continuous variables: the wine's verified alcoholic strength (*Verif.Alcohol*), the residual sugar (*Sugar*), the total and the volatile acidity (*Acidity.Tot* and *Acidity.Vol*), the total sulphur dioxide (*SO2.Tot*) and the ratio between free and total sulphur dioxide from

which the free sulphur dioxide (*SO2.Free*) was obtained. The sensory characteristics were evaluated with the help of Brescia's *Centro Studi Assaggiatori* using a panel of experts balanced for age and sex and their median score was attributed for each characteristic to each wine. The resulting variables were ordinal variables on a 0-9 scale. The sensory variables were divided in four groups, representing the visual characteristics, which considered colour saturation (*Col.Satur*), violet/green reflections (*Violet.Reflect*), orange/gold reflections (*Orange.Reflect*) and attraency (*Attraency*), the olfactory characteristics, which involved olfactory intensity (*Olfact.Inten*), floral (*Floral*), fruit (*Fruit*), spicy (*Spicy*), vegetal (*Vegetal*) and olfactory frankness (*Olfact.Frank*), the gustatory characteristics, which comprised structure (*Structure*), spherical perception (*Spherical.Perc*), acidity (*Acidity*), bitterness (*Bitterness*), aromatic richness (*Arom.Richness*) and gustatory harmony (*Gustatory.Harmony*), and the intense aromatic persistence, which took int account persistence (*Persistence*) and aftertaste frankness(*Aftertaste.Frank*). Moreover, four descriptive variables were taken into account, that is the year of production (*Year*), which could be seen as a proxy of the weather conditions which could have an impact on the wine production in terms of quantity and quality, the region of production (*Area*), the type of wine (red, white ect.) (*Type*) and the designation of origin (*Design*). As indicator of the overall quality of the wine, Altroconsumo produced a Global Score of Quality (*Global.Score*), which attributes a score ranging from 0 (lowest quality) to 100 (highest quality).

3 Preliminary data analysis

The chemical variables are considered as continuous as well as discrete, after a suitable discretization. Great attention is put on the discretization procedure, in order to avoid spurious relationships between the variables that were not present in the original continuous ones. With regards to the sensory variables, due to a sparse distribution of the observations, they are merged together in order to create a 3 or 4 points scale for each variable. The Altroconsumo Global Score of Quality is considered as continuous, when used jointly with the continuous chemical variables, and discretized in a 5-points scale, when used jointly with the ordinal variables. Regardless of the analysis considered, in searching the DAG underlying the BN, some knowledge is included, inhibiting arches from *Global.Score* to all the other variables, from all the chemical and sensory variables to the descriptive variables *Year*, *Area*, *Type* and *Design* and some arches connecting the descriptive variables with themselves. Between the chemical variables, preliminary results show a direct impact of total

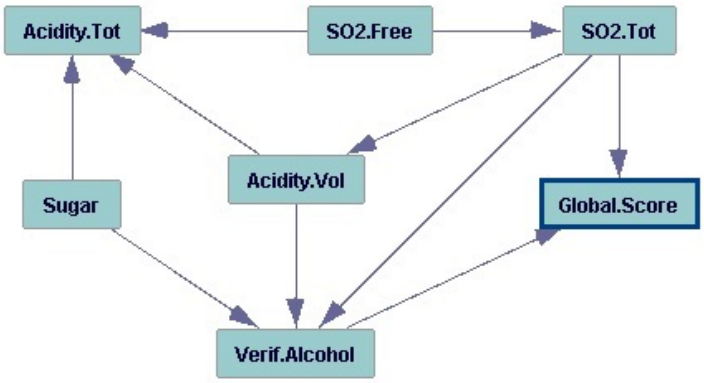

Figure 1. *Preliminary DAG for the continuous chemical variables*

sulphur dioxide on the Altroconsumo global score, as shown in other studies involving the same dataset but different statistical techniques (Brentari et al., 2012; Brentari and Vezzoli, 2015). Considering only the continuous chemical variables, a preliminary DAG is shown in Figure 1, where the total sulphur dioxide and the wine's verified alcoholic strength are the direct causes of the Altroconsumo global score.

References

BRENTARI E., CARPITA M., & VEZZOLI M. 2012. CRAGGING: a novel approach for inspecting Italian wine quality. In *: Proc. of the XXII European Symposium on Statistical Methods for the Food Industry. Paris*.

BRENTARI E., & VEZZOLI M. 2015. Evaluating Italian wine quality by cross-aggregating multiple regression trees. In *: Proceeding of the 143-rd Joint EAAE/AAEA Seminar on Consumer Behavior in a Changing World: Food, Culture and Society*.

LAWLESS H.T. 2013. *Quantitative Sensory Analysis*. Wiley-Blackwell.

GUIDA VINI. 2006-2012. Milano: Altroconsumo Edizioni.

SPIRTES P., GLYMOUR C., & SCHEINES R. 2000. *Causation, Prediction, and Search, 2nd edition*. Cambridge, Massachusetts: The MIT Press.

SPIRTES P., SCHEINES R., RAMSEY J., & GLYMOUR C. 2010. The TETRAD project: Causal models and statistical data. www.phil.cmu.edu/projects/tetrad/current