Causal reasoning applied to sensory analysis: the case of the Italian wine

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

[1] *Department of Economics and Management, University of Brescia, Italy*
E-mail : *silvia.golia@unibs.it; eugenio.brentari@unibs.it*; *maurizio.carpita@unibs.it*

Abstract
In this paper the causal reasoning is applied to the sensory analysis field in order to study the factors that have a direct influence in determining the quality of the Italian wine. Directed acyclic graphs, involving chemical as well as sensory variables, will be proposed in order to show the causal connections between these variables and the Altroconsumo's Global Score of Quality produced by the Italian independent consumer's association Altroconsumo for its annual publication *Guida Vini*. The analysis will be performed considering all the type of wines included in the database.
Keywords: Sensory analysis, Causal Graph, Altroconsumo Global Score of Quality

1. Introduction

The major goal of many sciences is to find the mechanisms that regulate the relations between variables in order to understand how the variables come to take on the values they take and to predict the values of the those variables under outside manipulations; this can reach by means of the causal inference. In this paper the causal reasoning is applied to the sensory analysis field in order to study the factors that have a direct influence in determining the quality of the Italian wine. To find a causal model for the available data allows one to highlight the relations between all the variables under study and to measure the direct and indirect effects that the variables have on each other. In the context of this paper the variables of interest are some variables that describe the wine and a measure of its quality, so the information that can be derived from a causal model is wide and includes the effects of chemical and sensory variables on the measure of the quality of the wine. The available data refer to the years 2006-2012 and were yearly collected by Altroconsumo, an Italian independent consumer's association, for its annual publication *Guida Vini* (Wines' Guide). This study was conducted within the project of sensory analysis developed by the *Data Methods and Systems Statistical Laboratory* (DMS StatLab) of the University of Brescia. The paper is organized as follows. Section 2 contains some theory related to the causal modeling, in section 3 the variables of interest are explained whereas section 4 reports preliminary analysis on the available data.

2. Causal models

A formal definition of causal model has been given by Pearl (2009); a causal model is defined by a tuple $CM = (\mathbf{U}, \mathbf{X}, \mathbf{F}, P(\mathbf{u}))$, where **U** contains background variables, **X** is a set of endogenous variables determined within the model, **F** is a set of structural functions specifying how each endogenous variable is determined by the other variables of the model and $P(\mathbf{u})$ is a joint probability distribution over **U**. The most frequently applied causal models are the Bayesian Networks (BNs), used mainly when the

variables are categorical, and the Structural Equation Models (SEMs), used when the variables are continuous. A BN specifies a factorization of the joint probability distribution of the variables under study whereas a SEM is composed by a set of structural equations with error terms. It is possible to demonstrate that these two models are closely linked (Spirtes et al., 2000).
The structure underlying both BN and SEM is represented by a Directed Acyclic Graph (DAG). A DAG $G=\{\mathbf{V}, \mathbf{E}\}$ is composed by a set of nodes (or vertices) $\mathbf{V}=\{V_1, V_2,\ldots, V_n\}$, which correspond to a set of random variables $\boldsymbol{X_V}$ indexed by **V**, and a set **E** of directed links (or edges) between pairs of nodes in **V**. A BN is defined by the pair (G, P), where G is a DAG and P is a probability distribution over the set of variables $\boldsymbol{X_V}$ which factorizes according to G as follows:

$$P(X_V) = \prod_{v \in V} P(X_v | X_{pa(v)})$$

where $X_{pa(v)}$ denotes the set of parent variables of variable X_v for each node V 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 reported above.
A linear SEM is a system of linear structural equations among a set of variables $\boldsymbol{X_V}$ such that each variable appears on the left hand side of at most one equation. Each equation is structural in the sense that it should be interpreted as an assignment process which express the causal relation between the dependent variable, which is on its left hand side, and its independent or explanatory variables. An error or disturbance term u is added to each equation, representing all factors omitted from the set of variables $\boldsymbol{X_V}$ that, together with the explanatory variables, determine the value of the dependent variable. This interpretation of the equations in linear SEM renders the equality sign in the equations non-symmetrical (Pearl, 2009). The set of equations for the SEM can be read from the corresponding DAG.
For both BN and SEM, as the first step one has to identify the (causal) relations among the variables generating a DAG, then for BN the joint probability distribution has to be specified in terms of the set of conditional probability distributions $P(X_v \mid X_{pa(v)})$, whereas for the linear SEM the system of linear equations must be read from the DAG.
To find the causal structure represented as a DAG is a problem impossible to solve with observational data only as in the case under study; 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 underlying structure. In order to automatically find the DAG, several algorithms have been proposed in the literature, the one used in this paper belongs to the class of constraint-based algorithms and it is the PC algorithm (Spirtes et al., 2000) and its Conservative version (CPC) (Ramsey et al., 2006) implemented in the Tetrad 5.1.0-6 program provided by Spirtes et al. (2010). The PC algorithm conducts a sequence of independence and conditional independence tests, and efficiently builds a DAG, or at least a Complete Partially DAG (CPDAG) that represents the equivalence class which contains the true DAG, from the results of those tests. CPC algorithm is a slight variation of the PC able to give a more conservative orientation of edges. From the obtained DAG it is possible to read the suitable causal model.

3. The dataset and the variables of interest

The database that will be analyzed in this study, was created using the data produced by Altroconsumo 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 Zuccolotto, 2011; Brentari et al., 2011, 2012; Brentari and Levaggi, 2014; Brentari and Vezzoli, 2015). The wines were chosen in order to represent the variety of Italian vineyards, producers and regions of origin. For each year, different vineyards and producers were considered, so that the observations could be considered as

independent. The variables considered measure chemical as well as sensory characteristics of a wine; the former are continuous whereas the latter are categorical. Moreover, Altroconsumo created a global score of quality for each wine; this is an indicator of the overall quality of the wine and assumes a score ranging from 0 (lowest quality) to 100 (highest quality). In the analysis that will follow, it will be considered as a continuous variable (*Global.Score*), when analyzed jointly with the chemical variables, as well as a categorical variable (*Global.Score.Cat*), when analyzed jointly with the sensory and discretized chemical variables. In order to obtain the global score in categorical form, the cut points considered were: 55, 60, 65 and 75. Lastly, three exogenous variables complete the set of variables that will be analyzed, that is the type of wine (*Type*), the designation of origin (*Denom*) and the region of production (*Region*).

The chemical variables considered by Altroconsumo are 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. Their distribution is not normal, with the exception of the total sulphur dioxide, which passes the Jarque-Bera normality test. In order to obtain the categorical version of these variables, cut points were identified with the help of experts.

The sensory characteristics considered by Altroconsumo can be divided in four groups representing visual, olfactory and gustatory characteristics of a wine and its intense aromatic persistence. The visual characteristics of a wine describe how a wine appears at a visual inspection and they are the intensity of the color (*Color.Int*) and how pleasant the aspect of the wine is (*Attraency*). The olfactory characteristics are related to the wine aroma and can be represented by the intensity of the bouquet (*Olfact.Int*), several fragrances that can be perceived in a wine, like floral (*Floral*), fruity (*Fruity*), spicy (*Spicy*) and vegetal (*Vegetal*), and the olfactory cleanness (*Olfact.Clean*) and quality (*Olfact.Qual*). The gustatory characteristics are connected to taste and mouthfeel of a wine which are described by its structure (*Structure*), the harmony of the different components measured by roundness (*Roundness*), gustatory harmony (*Gustatory.Harmony*), the aromatic richness (*Arom.Rich*) and the type of taste or mouthfeel sensation such as sourness (*Sourness*) and bitterness (*Bitterness*). Lastly, the intense aromatic persistence is described by the persistence of aromas (*Persistence*) and the aftertaste cleanness (*Aftertaste.Clean*) and quality (*Aftertaste.Qual*).

These variables were evaluated with the help of Brescia's *Centro Studi Assaggiatori*, the most advanced unit of sensory analysis in Italy. Each year, about 21 judges divided into three panels, evaluated the sensory characteristics of the wine already described, considering about 280 wines. The judges, for each wine analized, were asked to give a grade to each sensory variable considered, using a 0–9 scale where 0 denotes the lowest and 9 the highest score; the median score was the final score recorded in the Altoconsumo database. Due to the distribution of these sensory variables in the available dataset, it was necessary to properly merge the observed scores, obtaining binary variables for *Attraency*, *Bitterness*, *Gustatory.Harmony* and *Aftertaste.Clean*, and variables with three categories for the remaining ones.

4. The causal model for inspecting the quality of the Italian wines

As stated previously, in order to derive a causal model for the variables and data available, the first step consists in searching the causal DAG, or at least the CPDAG, underlying the variables. When possible, it is suitable to consider specific background information and assumptions during the searching step; this narrows down the various possible causal graphs found by the searching algorithms keeping them from exploring nonsensical graphs in which there are oriented arrows which show unrealistic causal connections. On the other hand, the imposition of constraints that represent background information and assumption has to be done very carefully, given that the constraints considerably condition the results of any searching algorithm. The following two subsections report the proposed CPDAG for the Altroconsumo database when the continuous chemical variables plus the global score of quality

(subsection 4.1) or the discretized chemical and sensory variables plus the discretized version of the global score of quality and the three exogenous variables (subsection 4.2) are considered.

4.1 Chemical variables versus global score

In this subsection the relations between the continuous chemical variables plus the global score of quality are analyzed. As specified at the beginning of section 4, if available, background knowledge has to be taken into account. In the present context, experts in the field entailed the constraints visualized in Figure 1, where the edges in the right hand side graph are forbidden edges, meaning that if a relationship between two variables connected by a forbidden edge exists, it is represented by an arrow with orientation opposite to the one expressed by the forbidden edge. These forbidden edges originate from the tiers ordering shown in the left-hand graph of Figure 1, which illustrates an ordering in the variables, meaning that variables in higher-numbered tiers can cause, but not be caused by, the variables in lower-numbered tiers.

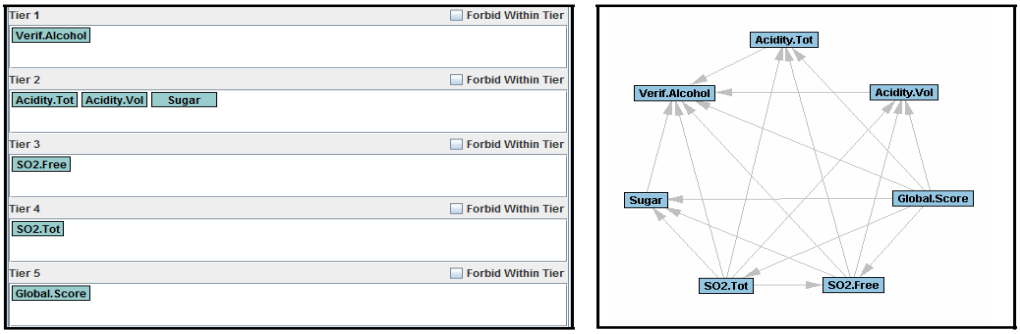

Figure 1: Constraints for the chemical variables plus the global score of quality

These constraints have been used by the PC algorithm in order to obtain a final CPDAG. Given that almost all the variables were not normal, the conditional independence test for non normal variables proposed by Ramsey (2014) was used. Moreover, the significance level alpha, which represents, in the context of the PC algorithm, a particular threshold for announcing that a certain link between variables in the causal model is significant, was set equal to 0.001. Figure 2 reports the final DAG for the chemical variables plus the global score of quality. One notes that the only one chemical variable that affects directly the Altroconsumo global score of quality is the level of total sulphur dioxide.

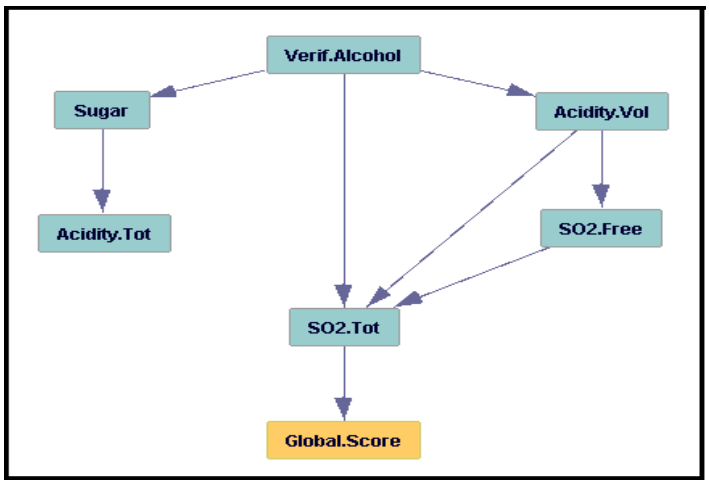

Figure 2: DAG for the chemical variables

4.2 Chemical and sensory variables versus global score

In this subsection the relations between the discretized chemical and sensory variables plus the discretized version of the global score of quality and the three exogenous variables are analyzed. The background knowledge used in the searching step is stated by the tiers ordering shown in Figure 3.

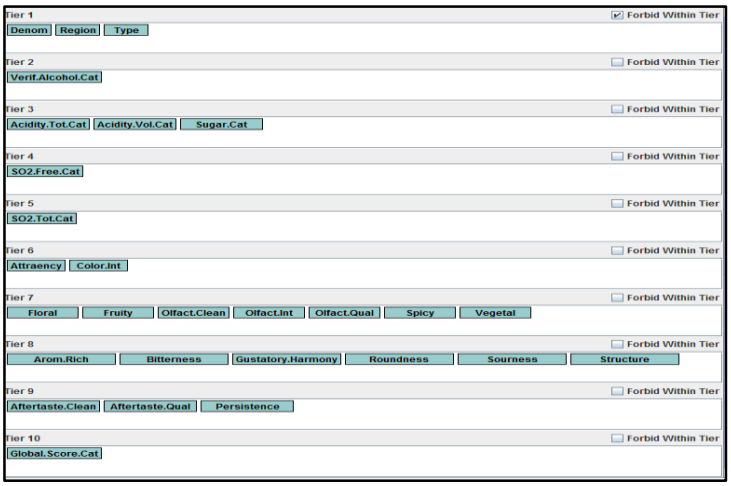

Figure 3: Constraints for the discretized global score, chemical, sensory and exogenous variables

These constraints have been used by the CPC algorithm in order to obtain the final CPDAG. The conditional independence test used in this analysis is based on the G^2 statistic (Spirtes et al., 2000) with alpha set equal to 0.001. Figure 4 shows the final CPDAG.

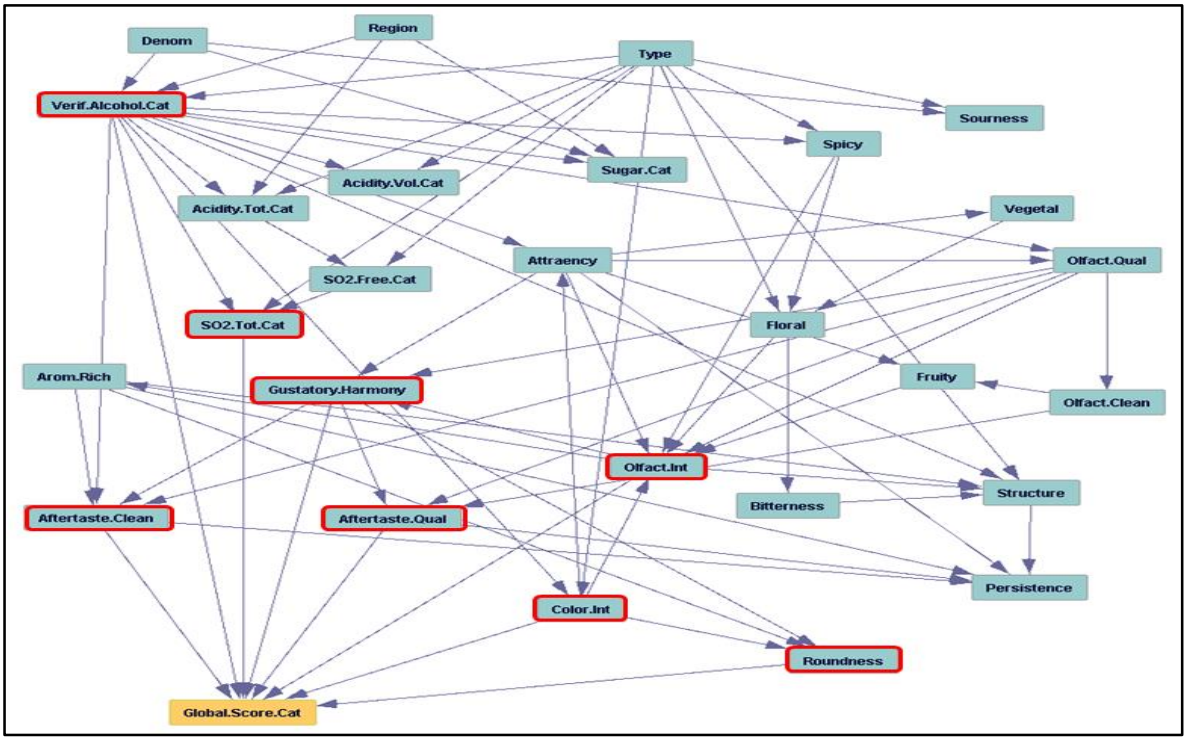

Figure 4: CPDAG for the discretized global score, chemical, sensory and exogenous variables

When all the chemical as well as the sensory variables are considered, one can note that the wine's verified alcoholic strength adds itself to the total sulphur dioxide as additional chemical variable that have a direct effect on the global score of quality. Between the sensory characteristics of a wine, it is possible to notice that the global score of quality is directly caused by two variables that express gustatory characteristics of a wine, such as roundness and gustatory harmony; the first is considered as objective whereas the second one is an hedonic indicator, that is linked to the pleasure of the judges. The other variables that have a direct impact to the global score are the color intensity, which is a visual characteristic, the intensity of the bouquet, which is an olfactory characteristic, and the aftertaste cleanness and quality, which are hedonic indicators that belong to the indicators of intense aromatic persistence.
To trace back to the most important variables that lead to a high global score of quality, is of great interest for those producers who want to score high in the Altoconsumo's *Guida Vini*.

References

Brentari, E., Carpita, M., & Vezzoli, M. (2012). *CRAGGING: a novel approach for inspecting Italian wine quality*. In: Proceedings AGROSTAT 2012, pp. 343-350.

Brentari, E., Levaggi, R., & Zuccolotto, P. (2011). Pricing strategies for Italian Red Wine. *Food Quality and Preference*, 22(8), 725-732.

Brentari, E., & Levaggi, R. (2014). The Hedonic Price for Italian Red Wine: Do Chemical and Sensory Characteristics Matter? *Agribusiness*, 30(4), 385-397.

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.

Brentari, E., & Zuccolotto, P. (2011). The impact of chemical and sensory characteristics on the market price of Italian red wines. *Electronic Journal of Applied Statistical Analysis*, 4(2), 265-276.

Pearl, J. (2009). *Causality : models, reasoning, and inference, 2nd edition*. Cambridge University Press.

Ramsey, J., Zhang, J., & Spirtes, P. (2006). *Adjacency-faithfulness and conservative causal inference*. In Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence. AUAI Press, Arlington, VA.

Ramsey, J.D. (2014). *A scalable conditional independence test for nonlinear, non-gaussian data*. arxiv.org/abs/1401.5031.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, Prediction, and Search, 2nd edition*. The MIT Press, Cambridge, Massachusetts.

Spirtes, P., Scheines, R., Ramsey, J., & Glymour, C. (2010). *The TETRAD project: Causal models and statistical data*. www.phil.cmu.edu/projects/tetrad/current.