Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal. Vol. 4, No 2 (2011), 265 – 276

The impact of chemical and sensory characteristics on the market price of Italian red wines

Abstract: *The hedonic price analysis presented in this paper is carried out on a dataset containing observable, sensory and chemical characteristics of a sample of Italian red wines. The study starts from the commonly observed evidence that in general the market price can be explained by the objective characteristics appearing on the label of the bottle and not by the wine's quality. The aim of the analysis is to discover how quality matters. This objective is pursued by means of the construction of latent sensory and chemical factors, whose implicit value, quantified using Random Forest variable importance measures, turns out to be appreciably high.*

Keywords: *hedonic price of wine, canonical correlation, Random Forest*

1. Introduction

The hedonic price method aims at analyzing the relationship between a product's price and its quality. In general, the method consists in a (possibly nonlinear) regression analysis of the price on the characteristics of the product. The implicit value of a characteristic is then given by the importance of the product attribute in the prediction of price, according to the above regression model.

The implicit value of a given feature can be different from what we mean to be its role in the quality of the product, as the former is influenced by market mechanisms, while the latter is concerned with the domain of intrinsic quality. The analysis of hedonic price is able to highlight these differences, thus assessing the market effect on the definition of price. For a long time the hedonic price method has been applied mainly to durable goods, but in the recent literature several applications take into account also different categories of goods.

This paper deals with the analysis of wine from the point of view of the relation between its intrinsic quality and its price. In the wine market, whatever the value of wine characteristics, the

decision of purchasing is mainly affected by previous experience and knowledge of the product, objective information described on the label, and the price itself. Consumers lacking previous experience and knowledge for a given wine, can rely only on objective information and on collective reputation associated with the production region, and brand names. In addition, since it can be argued that the price of a wine embodies characteristics that differentiate the product, many consumers use price as a signal of quality. In this context the use of hedonic price methods can help to understand the extent to which the quality of a wine with a given price meets consumers' expectations.

The first examples of studies on the hedonic pricing of wine date back to some seminal papers by Oczkowski (1994), Nerlove (1995), Combris et al. (1997, 2000), Schamel (2003), Schamel and Anderson (2003) and Bombrun and Sumner (2003). In particular, in the paper of Combris et al. (1997), a dataset including both observable (vintage year, vineyard region, grape variety, …) and sensory (taste, texture, odour, …) characteristics of wine is used for the first time. The interesting evidence found by Combris and his co-authors about the Bordeaux wine is that, while its quality, as measured by a jury grade assigned by professional wine tasters, can be explained primarily by the wine's sensory characteristics, the market price can be explained by the objective characteristics appearing on the label of the bottle. Apparently, many variables that are important in explaining quality do not play a role in the determination of the market price. The authors explain this evidence by stating that the hypothesis of perfect information, usually assumed in economic studies, is not realistic for wine market, where the objective characteristics (those mentioned on the label) are much easier to identify by consumers than the sensory ones.

A similar study concerning a sample of red Italian wines has been carried out by Brentari et al. (2007), with nearly the same conclusions of Combris et al. (1997). An important difference regards the dataset, including also chemical features of wines, which turned out to be completely uninfluent on price, too.

In this paper a development of that study is proposed. Firstly, a canonical correlation analysis is carried out between the set of sensory and chemical features, in order to define possibly correlated chemical and sensory latent factors, accounting for the intrinsic quality of wine. After the extraction of these latent factors, they are used as covariates in the hedonic price model, together with the original sensory and chemical variables and the objective characteristics. The

important result is that the latent factors emerge among the most important features determining the price of wine, so that their implicit value turns out to be appreciably high.

From the statistical point of view, the Random Forest method (Breiman, 2001), a very recent kind of nonlinear regression based on the theory of ensemble learning (Breiman, 1996; Friedman, 2003, 2006; Friedman and Popescu, 2003, 2005) is used. The implicit value of wine characteristics is evaluated by means of variable importance measures, a tool for discovering important predictors, introduced in the context of machine learning.

The paper is organized as follows: Section 2 contains a very brief recall about learning ensembles, with special reference to Random Forest and variable importance measurement. Section 3 and 4 describe the main characteristics of the dataset and the results of the analysis, respectively. Section 5 concludes.

2. Learning ensembles, Random Forest and variable importance measurement

In the context of hedonic price analysis a regression of the price on the characteristics of the product is performed in order to quantify the implicit value of each characteristic, given by its importance in the prediction model. Hence we are not really interested to prediction itself but to the extrapolation of the role played by covariates in the ability of the model of providing good predictions.

This is the main reason why researchers in this field rarely take advantage of powerful nonlinear regression techniques such as, for example, neural networks, which, in change of a significant accuracy, are impenetrable black-boxes.

Recent advances in data mining have tried to overcome this drawback and new prediction tools have been developed, able to generate, together with predictions, variable importance (VI) measures identifying the most important predictors of the response variable within the set of covariates. These powerful algorithms have been proposed in the framework of learning ensembles (Breiman, 1996; Friedman, 2003, 2006; Friedman and Popescu, 2003, 2005), and are particularly well suited to datasets composed by many predictors (the most part of them often

redundant or unrelated to the response variable) and characterized by complex relationships among the variables.

Learning ensembles are sequences of ensemble members. Each ensemble member is given by a different function of the input covariates; predictions are obtained by a linear combination of the prediction of each member. Learning ensembles can be built using different prediction methods, i.e. different base learners as ensemble members. The most interesting proposals use decision trees (more specifically CART, Classification And Regression Trees - Breiman *et al.*, 1984) as base learners and are called tree-based learning ensembles. Popular examples are the Random Forest technique (RF - Breiman, 2001) or the tree-based Gradient Boosting Machine (Friedman, 2001). Both these algorithmic techniques identify the most important predictors within the set of covariates, by means of the computation of some VI measures.

In this paper we are specifically interested to RF. Random Forest with randomly selected inputs are sequences of trees grown by selecting at random at each node a small group of F input variables to split on. This procedure is often used in tandem with bagging (Breiman, 1996), i.e. with a random selection of a subsample of the original training set at each tree. This simple and effective idea is founded on a complete theoretical apparatus analytically described by Breiman (2001) in his seminal work. The RF prediction is computed as an average of the single trees predictions. This successfully neutralizes the well-known instability of decisions trees.

In addition, two main measures of variable importance are available in order to identify informative predictors (Breiman, 2002):

1. *Mean Decrease in Accuracy* (MDA): at each tree of the RF all the values of h-th covariate are randomly permuted. New predictions are obtained with this dataset, where the role of h-th covariate is completely destroyed. The prediction error provided by this new dataset is compared with the prediction error of the original one: the MDA measure for h-th variable is given by the difference of these two errors.
2. *Total Decrease in Node Impurities* (TDNI): at each node z in every tree only a small number of variables is randomly chosen to split on, relying on some splitting criterion given by a variability/heterogeneity index such as the MSE for regression and the Gini index or the Shannon entropy for classification. Let $d(h,z)$ be the maximum decrease (over all the possible cutpoints) in the index allowed by variable X_h at node z. X_h is used to split at node z if $d(h,z)>d(w,z)$ for all variables X_w randomly chosen at node z. The

TDNI measure is calculated as the sum of all decreases in the RF due to *h*-th variable, divided by the number of trees.

Some recent studies have shown that the TDNI variable importance measures are affected by a bias in favour of variables with a higher number of possible cutpoints (for example numerical variables or nominal variables with a high number of categories) or having more missing values (see for example Strobl, 2005; Sandri and Zuccolotto, 2008, 2009). When TDNI is used, a preliminary bias-correcting procedure is thus recommended: some recent proposals are present in literature (Strobl *et al*., 2007a, 2007b; Sandri and Zuccolotto, 2008, 2009). The correction is less essential when the variance explained by the RF regression is high.

In section 4 RF and its VI measures will be used in the hedonic price analysis.

3. Data description

In this study we work on the dataset that Altroconsumo (an Italian Independent Consumers' Association) uses for its guide (Guida Vini 2006-2008)[1]. This dataset contains information on several characteristic of the wines, which we have grouped into different categories: *label characteristics*, *chemical characteristics*, *sensory characteristics* and *information about the price charged in different channels.*

- *Label variables* are the wine appellation (***App***: Merlot, Cabernet, …), the different geographical origin marking (***GOM***: DOC, DOCG, IGT), the Region of production (***Region***), the declared alcoholic strength, the ***Awards*** and the sales channel (***Channel***: large distribution or wine shops). Some frequency distributions are displayed in Figure 1.
- *Chemical variables* include the verified alcoholic strength (***Alcohol***), the residual sugar (***Chem1***), the total (***Chem2***) and the volatile (***Chem3***) acidity, the ratio between free and total sulphur anhydrides (SO_2) (***Chem4***) and the total sulphur anhydrides (***Chem5***). Finally, there is a chemical overall evaluation (***ChemG*:** chemical grade).

1 For further information about the dataset used in this work, see Brentari and Levaggi (2010) and Brentari *et al*. (2007). The authors thank Luigi Odello, chairman of Centro Studi Assaggiatori of Brescia, who supplied data, and Altroconsumo for the allowance to use them.

- *Sensory variables* list the average evaluations of a panel of experienced judges about the most important sensory characteristics of wine. Judges are required to express their vote about the following sensory variables: *visual characteristics* (***V1***, ***V2***, …: colour, violet reflections, …; ***ATT***: attractiveness or attraency), *olfactory characteristics* (***O1***, ***O2***, ...: intensity of the bouquet, floral, fruit, vegetal and spicy perfumes, olfactory intensity perception, harmonization of aromas, …), *gustatory characteristics* (***G1***, ***G2***, …: structure, harmony of different component, acidity, bitterness, astringency, aromatic richness) and Intense Aromatic Persistence (***PAI***: after-taste clean and quality). There are also the evaluations of overall olfactory characteristics (***QO***) and overall gustatory characteristics (***QRO***). The sensory overall evaluation is ***Grade***. Further wine quality indices are obtained as a (possibly weighted) average of some of the above sensory variables (***IIE***, ***IZOB***, …). The perception of each descriptor has been registered using a 0-9 scale where 0 denotes the lowest and 9 the highest score.

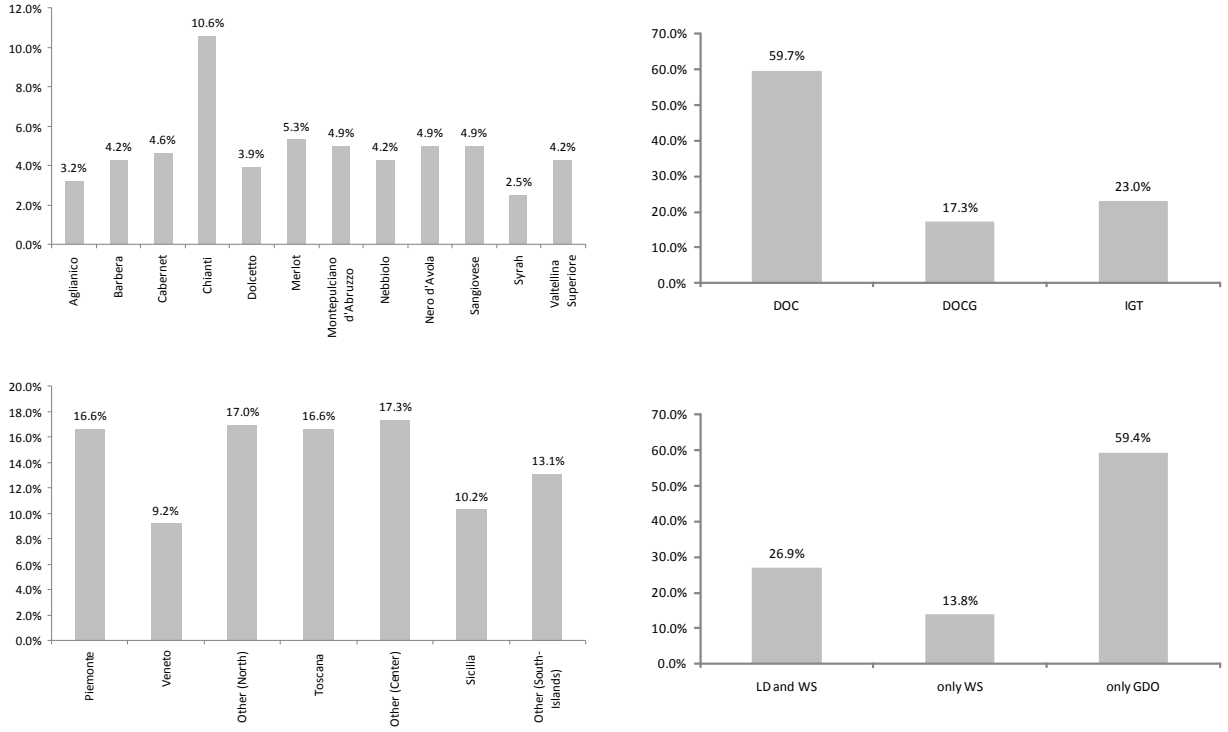

Figure 1. Frequency distribution of main wine appellations (*top left*), geographical origin marking (*top right*), Region of production (*bottom left*), sales channel (*bottom right*).

In our study we consider only medium-low priced red wines (the frequency distribution of prices is shown in Figure 2, left). The database consists of 283 observations concerning the period 2007-2008. For each bottle of wine sold both in the large distribution (LD) and in wine shops (WS), the average price is recorded. It can be interestingly noticed that the prices of wines sold in both channels tend to be higher in WS than in LD (Figure 2, right).

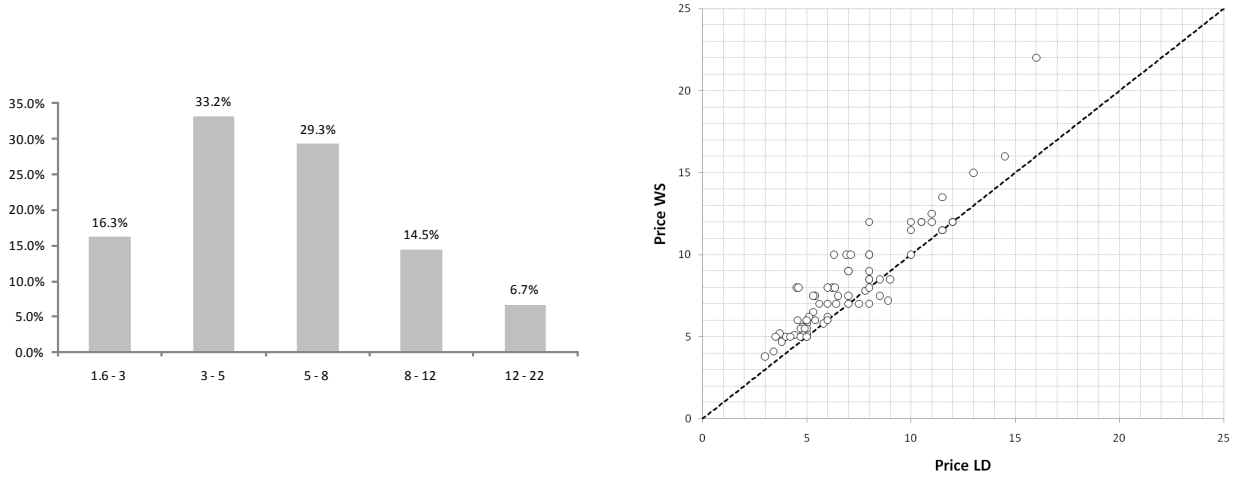

Figure 2. *Left*: frequency distribution of average prices (€). *Right*: scatterplot of the price in LD against the price in WS, for wines sold in both the channels.

4. Analysis of hedonic price

The analysis starts from the results presented in Brentari *et al.* (2007), obtained using the same dataset described in section 3. The main remarks are concerned with the fact that variables explaining (sensory and chemical) quality apparently do not play a role in the determination of the market price.

Although observed also in similar studies (see Combris *et al.*, 1997), this evidence is quite disappointing. It is our opinion that the variables explaining quality should necessarily be somehow determinant on price. Maybe the effect of each single variable is weak or highly disturbed by noise and this determines that its importance remains hidden in the regression models used for hedonic price analysis.

Following this idea we have tried to construct sensory and chemical latent factors able to summarize the effect of the corresponding variables. This should hopefully preserve the effects of the single variables on price, simultaneously eliminating noise. Latent factors can be obtained using a variety of dimensionality reduction techniques. We decided to carry out Canonical Correlation Analysis between the set ***C*** and ***S*** of chemical and sensory variables, respectively. We argue that forcing chemical and sensory latent factors to be correlated to each other should hopefully result in latent factors globally correlated to the wine's quality.

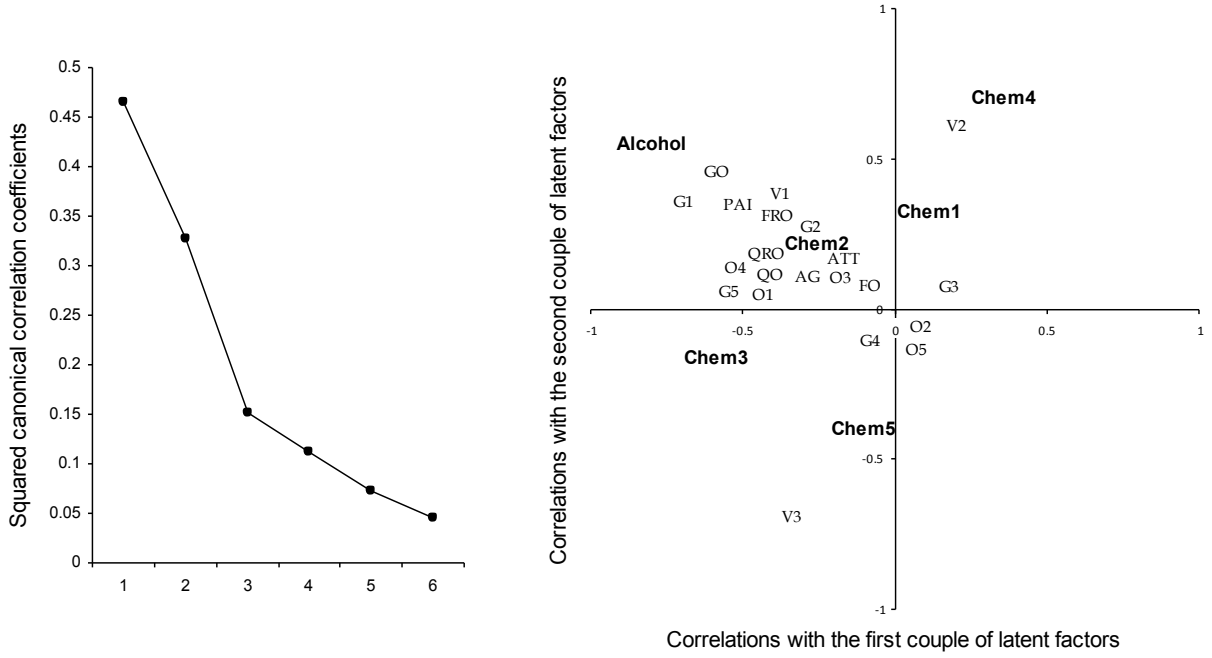

Figure 3. Scree plot of the squared canonical correlation coefficients (left) and plot of the correlations between the original variables and the corresponding latent factors (right)

The left part of Figure 3 displays the scree plot of the squared canonical correlation coefficients, showing that the canonical correlation significantly decreases after the first two couples of latent factors. For this reason we chose to draw the following analysis using only the first and the second chemical and sensory latent factors, denoted by LF_C1, LF_C2, LF_S1, LF_S2. The corresponding canonical correlation coefficients are ρ(LF_C1,LF_S1) = 0.68 and ρ(LF_C2,LF_S2) = 0.57. Inspecting the correlations between the original chemical and sensory variables with the corresponding latent factors (right part of Figure 3) we find out that:

- the first sensory latent factors is negatively correlated with almost all the sensory variables;

- the second sensory latent factors is positively correlated with almost all the sensory variables;
- the correlations between the chemical variables and their latent factors are heterogeneous.

This means that the sensory factors can be considered as composite indices of sensory quality (inversely – the first one – and directly – the second one – correlated to quality) and that the chemical variables play different roles in the linear combination obtained by forcing the chemical latent factors to be correlated to the sensory quality. For example, we notice that the variable ***Alcohol*** is associated to high values of sensory quality, as measured by both the indices.

The existence of some association between latent factors and price is confirmed by the inspection of the scatterplots, completed by a nonparametric regression obtained with kernel smoothing (Figure 4). It is worthy noting that the price is positively associated to quality in all the four graphs (let us recall that the first latent factors resulted to be inversely associated to quality, as shown in Figure 3).

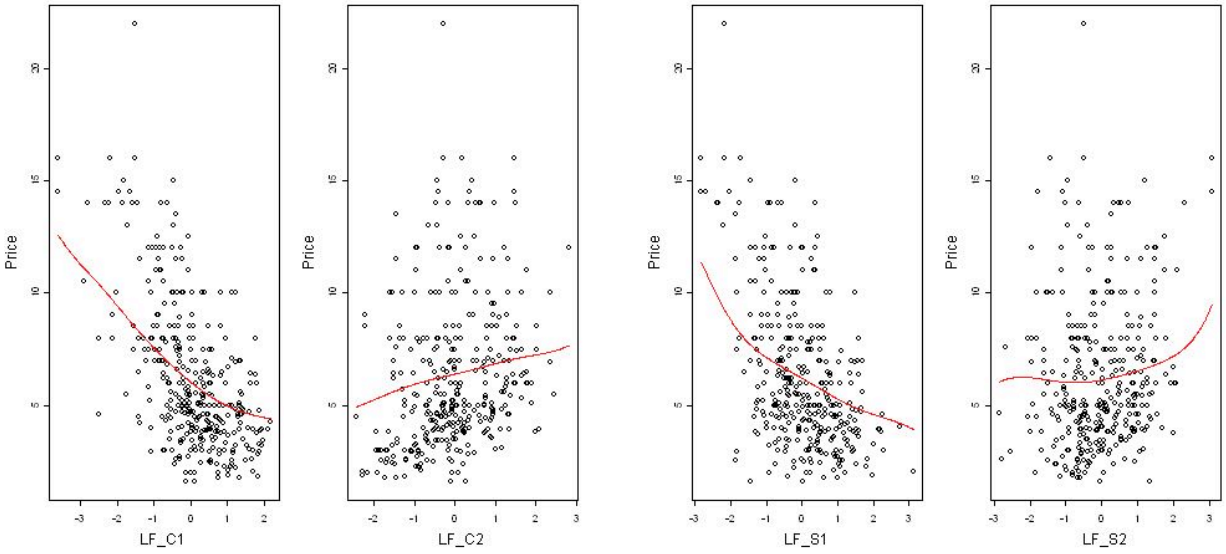

Figure 4. Chemical (left) and sensory (right) latent factors against price

In the second step, a RF regression with 8000 trees is carried out using as covariates the sensory and chemical variables, the latent factors and the label variables. The explained variance is 81.49%. The implicit values of wine's characteristics are quantified by means of VI measures (Figure 5). The high value of the explained variance allows the computation of the TDNI measure in its original version, without bias correction.

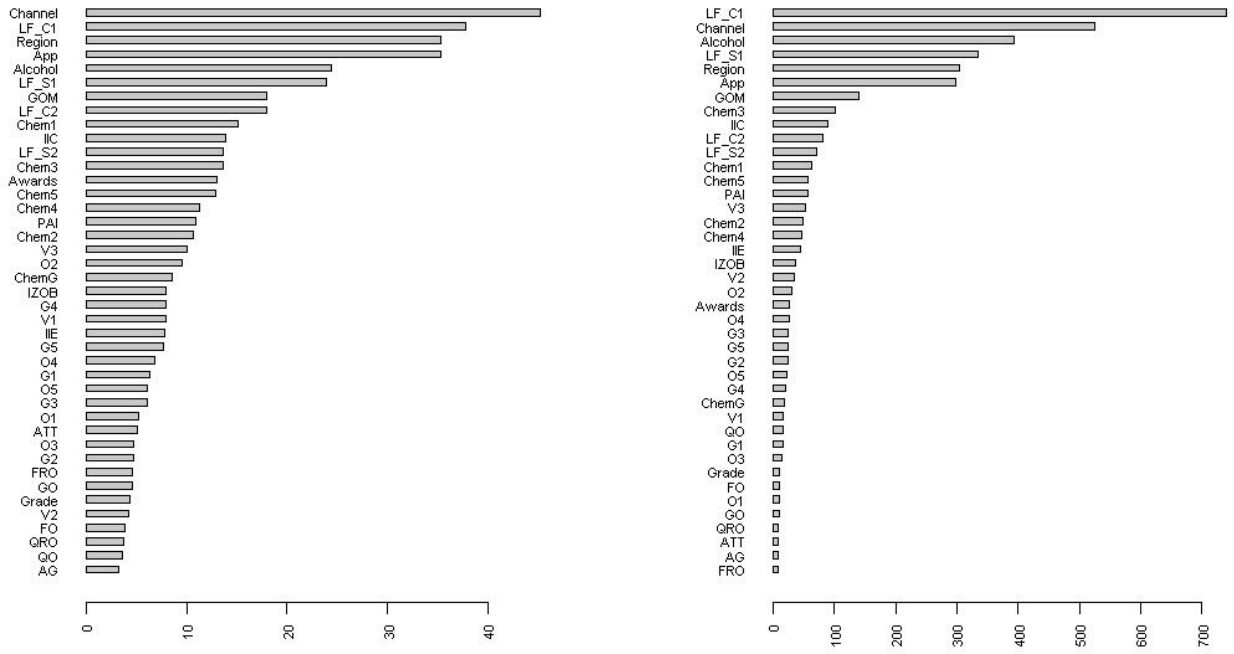

Figure 5. MDA (left) and TDNI (right) measures

The latent factors emerge among the most important variables in the RF regression. Thus chemical and sensory variables, if summarised through a proper composite index, exhibit an appreciable influence on market price.

4.1 Market segmentation based on chemical-sensory quality

The analysis can be completed by performing a market segmentation by means of a cluster analysis of wines based on the four latent factors denoting chemical and sensory quality. Figure 6 shows the pattern of the ratio DB/DT, where DB and DT denote respectively the between and the total deviance, as a function of the number of clusters (*k-means* algorithm). The relative increase of DB in the solution with k clusters with respect to the solution with $k-1$ clusters is reported in the same graph. We opted for the solution with 5 clusters (DB/DT=0.59).

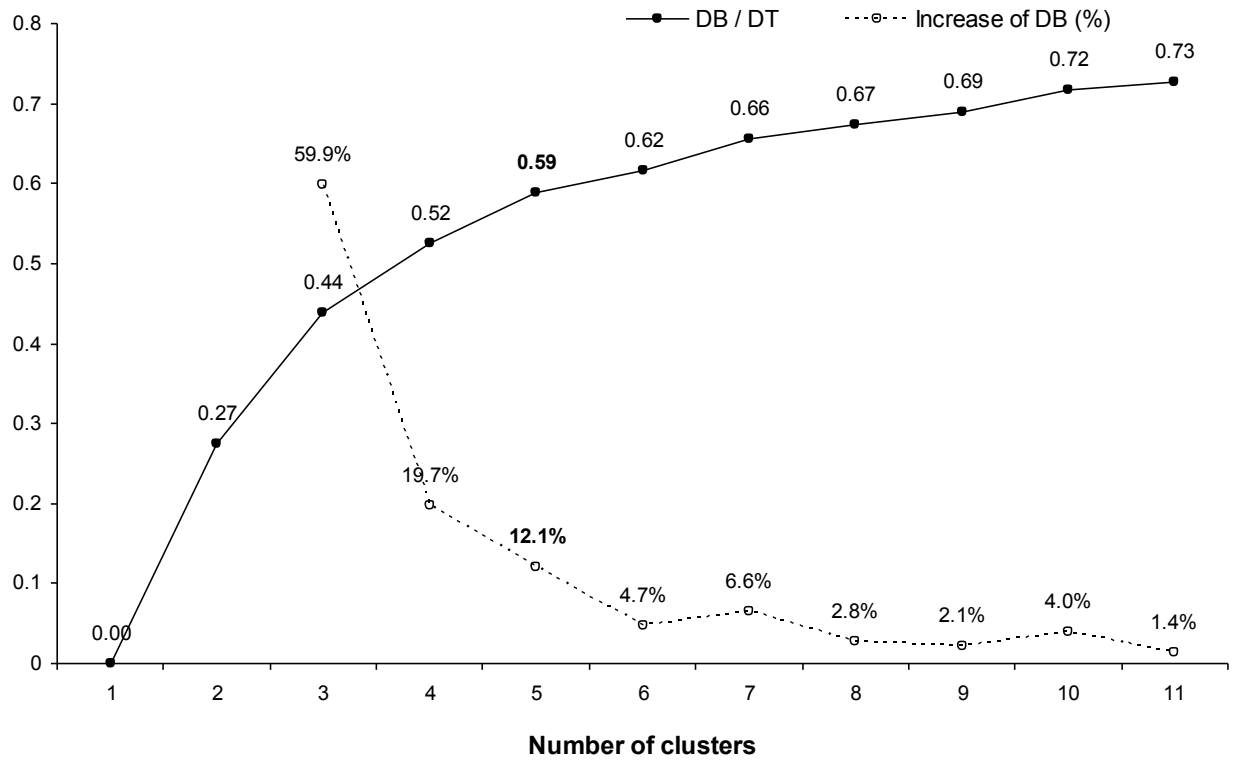

Figure 6. Ratio DB/DT and relative increase of DB against number of clusters

Thanks to the relatively high values of the canonical correlation coefficients obtained in the previous analysis, the average latent factors LF_1 = (−1)·(LF_C1+LF_S1)/2 and LF_2 = (LF_C2+LF_S2)/2 can be considered as indicators of the overall chemical-sensory quality (both positively associated to quality). The left part of Figure 7 shows the position of the 5 clusters according to the two overall quality indices. The average price within the clusters (reported in the same graph) clearly suggests that the segmentation based on chemical and sensory quality is convincingly reflected by the market price.

We can visibly distinguish two clusters of low priced and medium/low quality wines (C1 and C2), one cluster of medium priced and medium/high quality wines (C3), one cluster of medium priced and medium/low quality wines (C4) and one cluster of high priced and high quality wines (C5). The frequency distribution of the clusters is displayed in the right part of Figure 7. In other words, in clusters C1 and C2 we find wines addressed to the market of low-requiring consumers mostly interested to price, while C5 contains wines for demanding consumers and connoisseurs.

Interestingly, the medium price clusters C3 and C4 significantly differ each other from the point of view of quality in the second dimension.

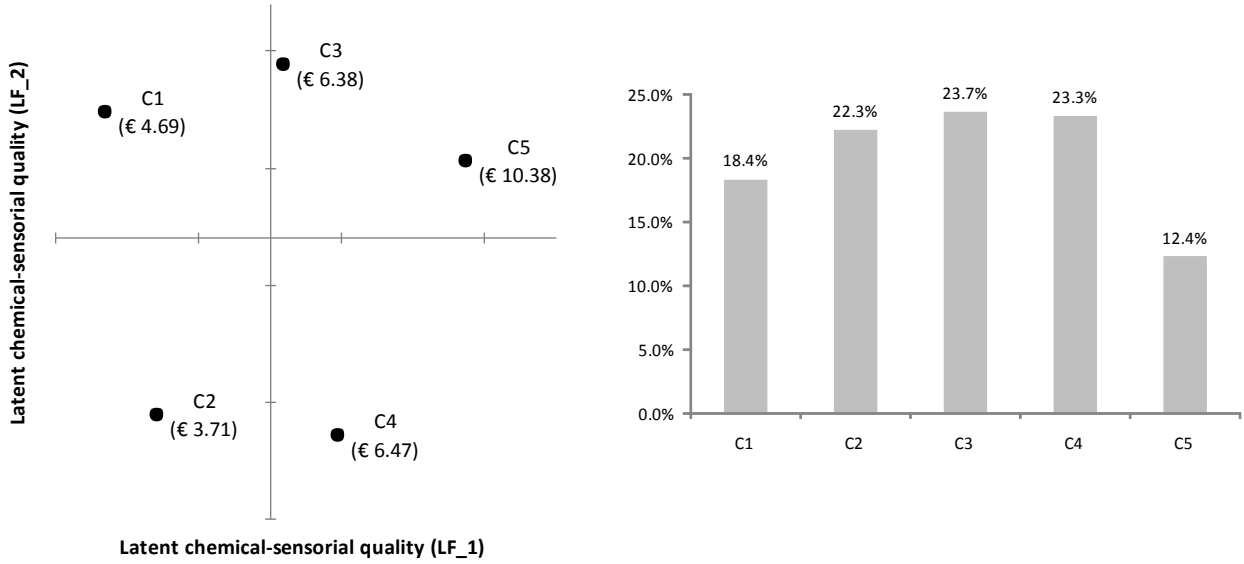

Figure 7. *Left*: positioning of the 5 clusters according to the overall chemical-sensory quality (average price within the cluster in parenthesis). *Right*: frequency distribution of the clusters.

From a market-oriented perspective, the profiles of the clusters can be described by means of the sales channel and the geographical origin marking distributions (Figures 8 and 9).

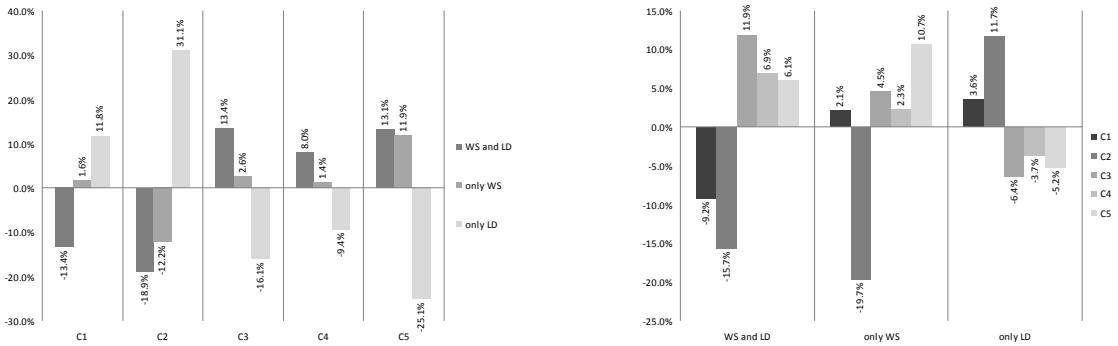

Figure 8. Cluster distribution by sales channel (differences respect to the corresponding proportions in the whole sample)

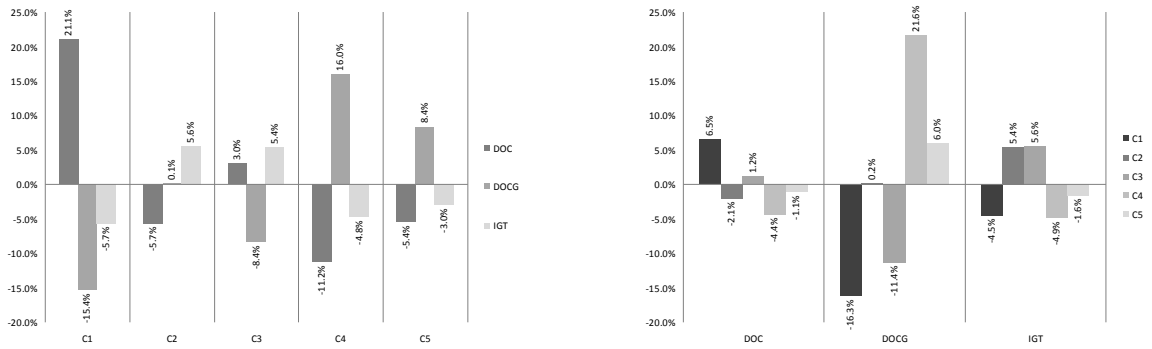

Figure 9. Cluster distribution by geographical origin marking (differences respect to the corresponding proportions in the whole sample)

As expected, the wines of clusters C1 and C2 tend to be mostly sold only in LD, while a clear prevalence of wines sold only in WS can be observed in cluster C5. The medium price clusters C3 and C4 tend to be present in both the channels. From the point of view of the geographical origin marking, DOC and DOCG wines are over-represented respectively in cluster C1 and C4, which are characterized by a medium/low overall quality level. A reasonable explanation of this disappointing evidence is that the geographical origin marking does not necessarily imply a high quality, due to the presence of several vineries with possibly different quality standard in their production processes. In the case of DOC wines, the market seems to be able to distinguish among wines of different quality levels, as demonstrated by the very low average price of cluster C1, despite the large amount of DOC wines within it. Quite the opposite, the marking DOCG seems to benefit of a premium price whatever the wine quality, which may justify the relatively high average price of the medium/low quality cluster C4.

5. Concluding remarks

Several authors have highlighted the existence of a relation between price and label variables of a wine, while sensory characteristics, although important in explaining quality, apparently do not play a role in the determination of the market price. This could discourage many wine producers

who, unable to rely on an important brand or to access costly advertising campaigns, could renounce to penetrate market.

In this paper the importance of chemical and sensory variables in determining market price is brought to light thanks to the joint application of a traditional statistical technique – the *Canonical Correlation Analysis* – and a recent ensemble learning algorithm – *Random Forest.* The analysis reveals that the role of chemical and sensory variables remains hidden when they are separately considered in the hedonic price model, while it appreciably comes to light if they are summarised through a proper composite index. Following this interpretation, a producer who pursues high quality, after a reasonable period of time should expect a positive effect on the wine reputation which will permit him to increase the market price.

In a second step of the analysis, the composite indices of the overall chemical and sensory quality are used in order to depict a market segmentation, able to point out the existence of different combinations of the pair quality/price, which continues to be partly affected by some label features, like the geographical origin marking.

References

[1]. Altroconsumo, Guida Vini (2006-2008), Altroconsumo Edizioni, Milano.

[2]. Bombrun H., Sumner A. (2003), What determines the price of wine? The value of grape characteristics and wine quality assessments, *Agricultural Issues Center* – University of California, 18, 1–6.

[3]. Breiman, L. (1996), Bagging predictors, *Machine Learning*, 24, 123-140.

[4]. Breiman, L. (2001), Random forests, *Machine Learning*, 45, 5-32.

[5]. Breiman L. (2002), *Manual on setting up, using, and understanding random forests v3.1*, http://oz.berkeley.edu/users/breiman.

[6]. Breiman L., Friedman J.H, Olshen R.A. and Stone C.J. (1984), *Classification and Regression Trees*, Chapman & Hall, New York.

[7]. Brentari E., Dancelli L., Zuccolotto P. (2007), Analisi dei fattori di maggior impatto sulla formazione del prezzo del vino: un esempio di utilizzo delle Random Forest, *Rapporti di ricerca del Dipartimento di Metodi Quantitativi*, Università di Brescia, n. 297.

[8]. Brentari E., Levaggi R. (2010), Hedonic price for the Italian Red Wine: a panel analysis, Agrostat 2010 Proceedings, Benevento.

[9]. Combris P., Lecocq S. e Visser M (1997), Estimation of a hedonic price equation for Bordeaux wine: does quality matter?, *The Economic Journal*, 107(441): 390-402.

[10]. Combris P., Lecocq S., Visser M. (2000), Estimation of a hedonic price equation for Burgundy wine, *Applied Economics*, 32 (8): 961–967.

[11]. Friedman J.H. (2001), Greedy function approximation: a gradient boosting machine, *Annals of Statistics*, 29, 1189-1232.
[12]. Friedman, J. H. (2003), Recent advances in predictive machine learning, *Stanford University, Department of Statistics, Technical report*.
[13]. Friedman, J. H. (2006), Separating signal from background using ensembles of rules, *Stanford University, Department of Statistics, Technical report*.
[14]. Friedman, J. H., and Popescu, B. E. (2003), Importance Sampled Learning Ensembles, *Stanford University, Department of Statistics, Technical Report*.
[15]. Friedman, J. H., and Popescu, B. E. (2005), Predictive learning via rule ensembles, Stanford University, Department of Statistics, Technical report.
[16]. Nerlove, M. (1995), Hedonic price functions and the measurement of preferences: the case of Swedish wine consumers, *European Economic Review,* 39, 1697-716.
[17]. Oczkowski, E. (1994), A hedonic price function for Australian premium wine, *Australian Journal of Agricultural Economics,* 38, 93-110.
[18]. Sandri M., Zuccolotto P. (2008), A bias correction algorithm for the Gini measure of variable importance, *Journal of Computational Graphical Statistics*, 17, 1-18.
[19]. Sandri, M., Zuccolotto, P. (2009), Analysis and correction of bias in Total Decrease in Node Impurity measures for tree-based algorithms, *Statistics and Computing, forthcoming.*
[20]. Schamel G. (2003), "A hedonic pricing model for German wine", *Agrarwirtschaft*, 52 (5), 247–254 .
[21]. Schamel G., Anderson K. (2003), "Wine Quality and varietal, regional and winery reputations: hedonic prices for Australia and New Zealand", *The Economic Record*, 79 (246), 357–369.
[22]. Strobl, C. (2005) Statistical Sources of Variable Selection Bias in Classification Trees Based on the Gini Index, *Technical Report*, SFB 386.
[23]. Strobl, C., Boulesteix, A.-L., Zeileis, A. and Hothorn, T. (2007a) Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution, *BMC Bioinformatics*, 8-25.
[24]. Strobl, C., Boulesteix, A.-L. and Augustin, T. (2007b), Unbiased split selection for classification trees based on the Gini Index, *Computational Statistics & Data Analysis*, 52 (1), 483-501.