La valeur implicite de la qualité chimique et sensorielle dans l'analyse hédonique des vins rouges italiens à bas prix

The implicit value of chemical and sensorial quality in the hedonic analysis of low-priced Italian red wines

Eugenio Brentari & Paola Zuccolotto

Department of Quantitative Methods, University of Brescia
E-mail : brentari@eco.unibs.it; zuk@eco.unibs.it

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 low-priced Italian red wines. The study starts for 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 sensorial and chemical factors, whose implicit price, quantified using Random Forest variable importance measures, turns out to be appreciably high.
Keywords: hedonic price of wine, canonical correlation, Random Forest

Résumé
L'analyse du prix hédonique présentée dans cet article est conduite sur une base de données sur des caractéristiques observables, sensorielles et chimiques d'un échantillon de vins rouges italiens à bas prix. L'étude se fonde sur l'évidence communément observée selon laquelle le prix de marché, en général, peut être expliqué par les caractéristiques objectives que l'on trouve sur l'étiquette de la bouteille et non par la qualité du vin. Le but de l'analyse est de découvrir l'influence de la qualité. Cet objectif est poursuivi à travers des moyens de construction de facteurs latents, sensoriels et chimiques, dont le prix implicite, calculé à travers les mesures Random Forest pour l'importance d'une variable, est sensiblement élevé.
Mots-clés: prix hédonique du vin, corrélation canonique, Random Forest

1. Introduction

The hedonic price method aims at analyzing the relationship between price and quality of a product. In general, the method consists in a (possibly nonlinear) regression analysis of the price on the characteristics of the product. The implicit price of a characteristic is then given by the importance of the product attribute in the prediction of price, according to the defined 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 difference, thus assessing the effect of market 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 are present also to 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 the hedonic price method can help in understanding 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), Shamel (2003) and Shamel and Anderson (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 Combis 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 sensory characteristics of the wine, the market price can be explained by the objective characteristics appearing on the label of the bottle. Many variables that are important in explaining quality apparently do not play a role in the determination of the market price. The authors explain this evidence stating that the hypothesis of perfect information, usually assumed in economic studies, is not realistic for the 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), nearly with the same conclusions of Combris et al. (1997). An important difference was that the dataset included also chemical features of wine, which turned out to be completely uninfluent on price, too.
In this paper a development of that study is proposed. First, 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 if wine. After the extraction of these latent factors, they are used as covariates in the hedonic price model. The important result is that the latent factors emerge among the important features in determining the price of wine, so that their implicit price turns out to be appreciably high.
From the statistical point of view, the method of Random Forest (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 price 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 price of each characteristic, given by its importance in the

prediction model. Hence we are not really interested in prediction itself, but in the extrapolation of the role played by covariates in the capacity of the model of providing good predictions.
Mainly for this reason 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 for 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 and predictions are obtained by a linear combination of the prediction of each member. Learning ensembles can be built using different prediction methods, that is 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. RF 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), that is 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 measure is 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). When TDNI is used, a preliminary bias-correcting procedure is thus recommended: some recent proposals are present in the 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. Description of data

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 *informations about the price charged in different channels.*

- The *label variables* are the *wine appellation* (***App***; Merlot, Cabernet, ...), the different geographical origin marking (***GOM***) as DOC (AOC), DOCG, ...; the Region of production (***Region***), the declared *alcoholic strenght* (***Alcohol***), the ***Awards*** and the *sales channel* (***Channel***; large distribution channel or wine shops).
- The *chemical variables* (***Chem1***, ***Chem2***, ...) include the *verified alcoholic strength*, the residual sugar, the volatile and the total acidity, the sulphur anhydrides (SO_2) and the ratio between free and total sulphur anhydrides. Finally, there is a chemical overall evaluation (**ChemG;** chemical grade).
- The *sensory variables* list the average evaluations of a panel of experienced judges about the most important sensory characteristics of wine. They must express their vote about the following sensorial variables: *visual characteristics* (**V1**, **V2**, ...: colour, violet reflections and attractiveness or attraency **ATT**), *olfactory characteristics* (**O1, O2, ...**: intensity of the bouquet; floral, fruit, vegetal and spicy perfumes; olfactory intensity perception; harmonization of aromas) and *gustatory characteristics* (**G1, G2, ...**: structure, harmony of different component, acidity, bitterness, astringency, aromatic richness, *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**. Other indexes of wine quality indices are obtained as an 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.

In our study we considered only red wines: the database is made up of 434 observations concerning the period 2006-2008. For each bottle of wine in the sample we have recorded the average **price** when sold in the large distribution and in wine shops.

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 (sensorial 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

[1] For more 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.

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 sensorial 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 sensorial variables, respectively. We argue that forcing chemical and sensorial latent factors to be correlated each other should hopefully result in latent factors globally correlated with the wine's quality.

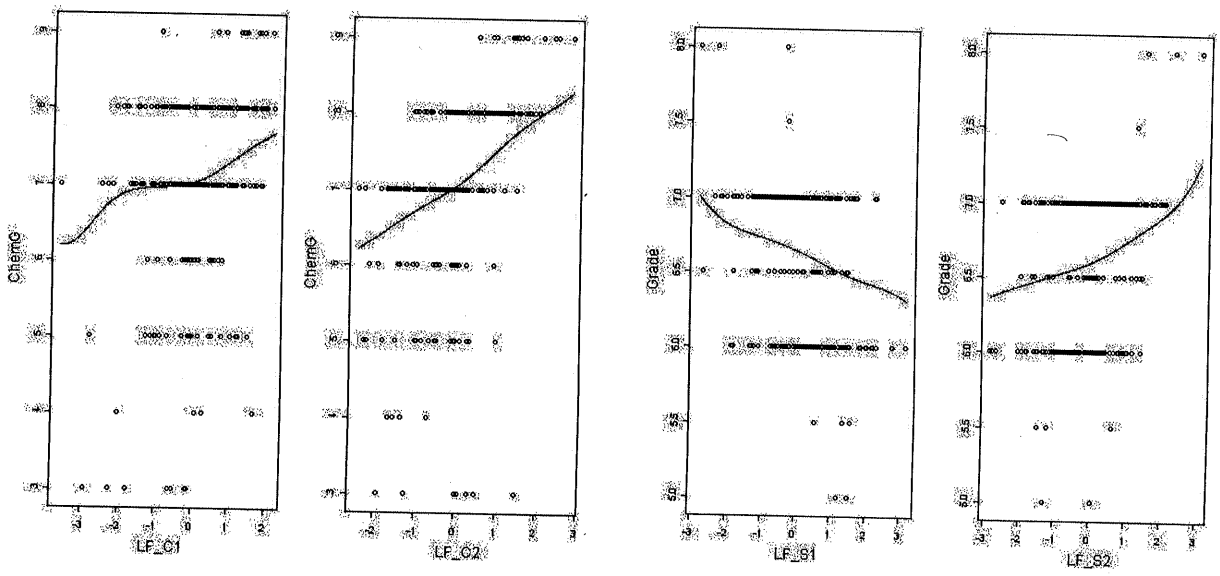

Figure 1. Chemical latent factors against chemical overall grade (left) and sensorial latent factors against sensorial overall grade (right)

Let LF_C1, LF_C2, LF_S1, LF_S2 be respectively the first and the second chemical and sensorial latent factors, the resulting canonical correlation coefficients are 0.68 and 0.57. The existence of some association of latent factors with the corresponding chemical and sensorial grades and with price is confirmed by the inspection of the scatterplots, completed by a nonparametric regression obtained with kernel smoothing (Figures 1 and 2). It follows that the latent factors are somehow related both with the wine's quality and with its price.
In the second step, the RF regression with 8000 trees is carried out including the latent factors within the set of covariates. The explained variance is 81.49%.
The implicit prices of wine's characteristics are quantified by means of VI measures (Figure 3). The high value of the explained variance allows the computation of the TDNI measure in its original version, without bias correction.

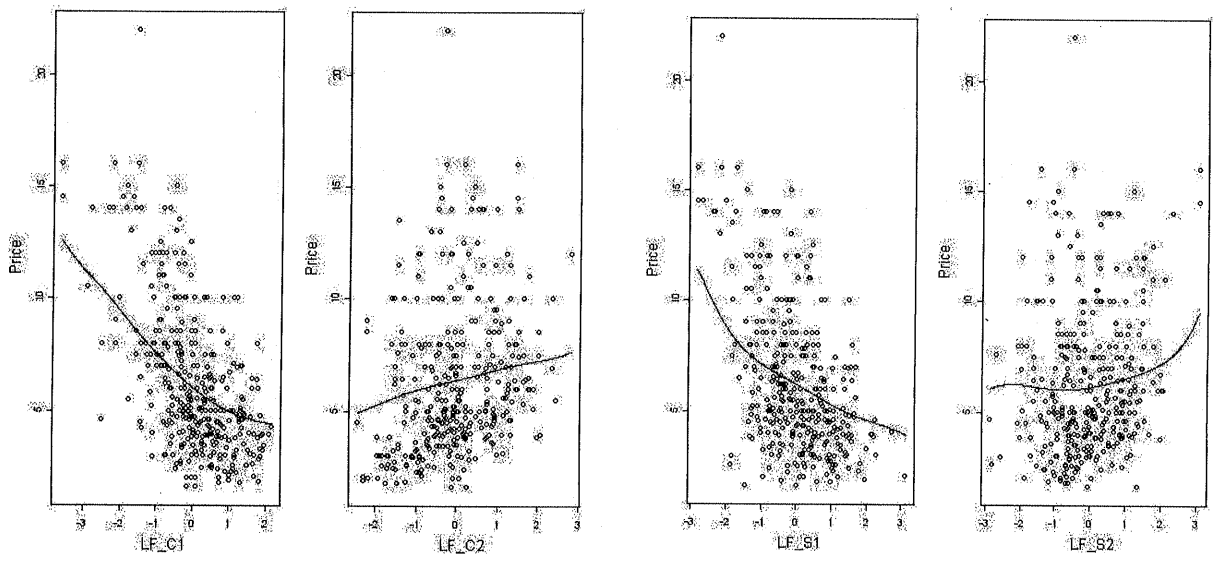

Figure 2. Chemical (left) and sensorial (right) latent factors against price

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

5. Concluding remarks

Several authors have evidenced the existence of a relation between price and label variables of a wine, while sensorial 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 sensorial 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 sensorial 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.

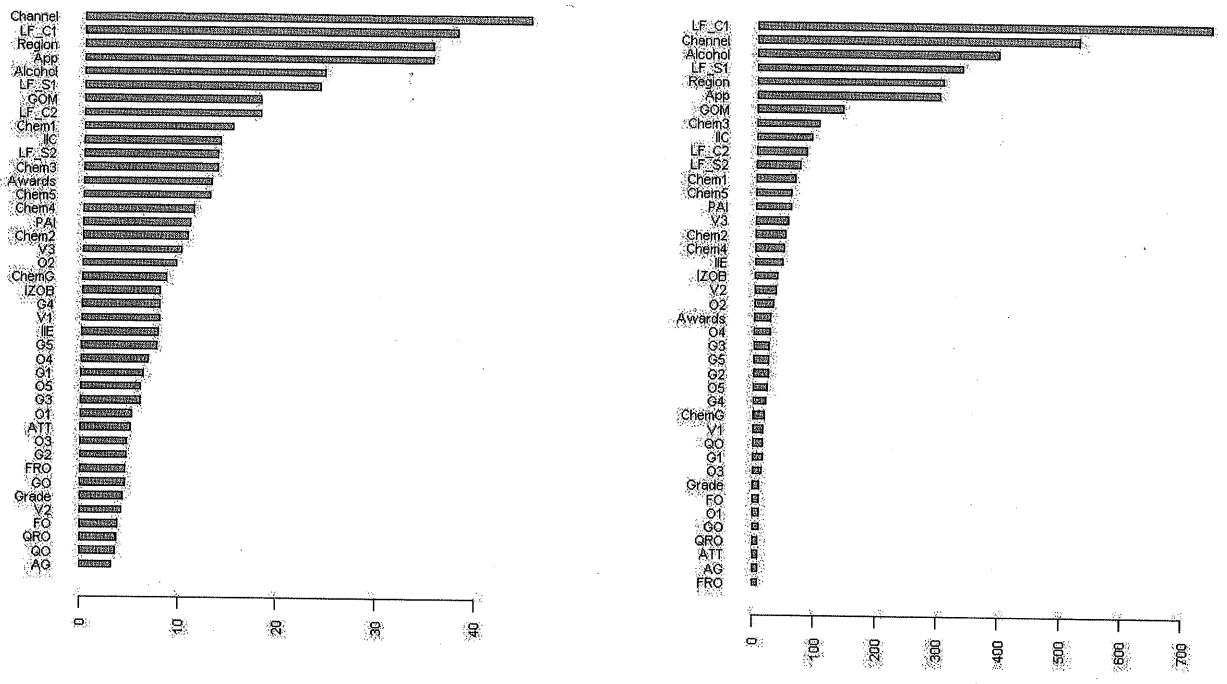

Figure 3. MDA (left) and TDNI (right) measures

Bibliography

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

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.

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

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

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

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

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.

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

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.

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

Friedman J.H. (2001), Greedy function approximation: a gradient boosting machine, *Annals of Statistics*, 29, 1189-1232.

Friedman, J. H. (2003), Recent advances in predictive machine learning, *Stanford University, Department of Statistics, Technical report.*

Friedman, J. H. (2006), Separating signal from background using ensembles of rules, *Stanford University, Department of Statistics, Technical report.*

Friedman, J. H., and Popescu, B. E. (2003), Importance Sampled Learning Ensembles, *Stanford University, Department of Statistics, Technical Report.*

Friedman, J. H., and Popescu, B. E. (2005), Predictive learning via rule ensembles, Stanford University, Department of Statistics, Technical report.

Nerlove, M. (1995), Hedonic price functions and the measurement of preferences: the case of Swedish wine consumers, *European Economic Review,* 39, 1697-716.

Oczkowski, E. (1994), A hedonic price function for Australian premium wine, *Australian Journal of Agricultural Economics,* 38, 93-110.

Sandri M., Zuccolotto P. (2008), A bias correction algorithm for the Gini measure of variable importance, *Journal of Computational Graphical Statistics*, 17, 1-18.

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.*

Schamel G. (2003), "A hedonic pricing model for German wine", *Agrarwirtschaft*, 52 (5), 247–254 .

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.

Strobl, C. (2005) Statistical Sources of Variable Selection Bias in Classification Trees Based on the Gini Index, *Technical Report*, SFB 386.

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.

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.