Journal of Wine Economics, Volume 3, Number 1, Spring 2008, Pages 30–50

https://doi.org/10.1017/S1931436100000547 Published online by Cambridge University Press

Analyzing Wine Demand with Artificial Neural Networks

Margherita Gerolimetto[a], Christine Mauracher[b] and Isabella Procidano[c]

Abstract

In this paper we analyse wine demand in Italy with microdata. Instead of estimating a parametric model, we study the demand following a non parametric approach by means of Artificial Neural Networks. The input set includes the usual economic variables (price and income) and some socio-demographic factors that are also shown to be relevant for demand analysis. We compute price elasticities using two different nonparametric procedures. (JEL Classification: C14, C21, Q11, Q13)

I. Introduction

This paper is aimed at modeling the behaviour of Italian wine consumers with the Artificial Neural Network[1] in order to estimate the demand curve and elasticity, not only for the entire sample, but also for various consumer groups. The input variable set includes the usual economic factors (price and income) as well as some socio-demographic variables (e.g. the demographic structure of the population).

Unlike parametric demand systems models, e.g., the AIDS or the Rotterdam model (Deaton and Muellbauer, 1980; Theil, 1965), the neural network approach is non parametric and more flexible because of its capability of capturing complex non linear interactions between input and output variables. While the parametric framework allows to compute elasticities directly from the model parameters, in the Artificial Neural Network (henceforth ANN) approach, elasticities are computed in a different manner. To our knowledge, prior published studies contain only a few proposals to compute elasticities in a non

[a] Department of Statistics and Section of Agricultural Economics and Politics – Ca' Foscari University of Venice, E-mail: margherita.gerolimetto@unive.it
[b] Department of Statistics and Section of Agricultural Economics and Politics – Ca' Foscari University of Venice, E-mail: maurache@unive.it
[c] Department of Statistics and Section of Agricultural Economics and Politics – Ca' Foscari University of Venice, E-mail: isabella@unive.it

[1] See, for example, Bishop (1995), Fine (1999), Hagan et al.(1999), Haykin (1999) and White (1992).

© The American Association of Wine Economists, 2008

parametric manner. In this paper, we focus on two methods developed by Subba Rao et al. (1998), hereafter SR, and by Gruca and Klemz (1998), hereafter GK.

The paper is organized as follows: In the second section we describe the data set; in the third section we give some basic notions about ANN; in the fourth section we illustrate the SR and GK methods to compute elasticities; in the fifth section we illustrate the ANN results for the Italian wine market; the sixth section summarizes the main findings.

II. The Data Set

This paper draws on household expenditure microdata for the year 2003. The data set, obtained from Osservatorio AcNielsen, consists of 4245 households that are representative for the national population. A sample of households was randomly selected using sampling techniques that ensure household representations of socio-demographic variables (e.g. household income, education) and geographic location.

As for the socio-demographic characteristics, representativeness is ensured for five microareas (primary units of sampling) that are comparatively homogeneous with respect to demographic and economic factors. As for the geographic location, representativeness is ensured for four areas and different kinds of towns, taking into account the urbanization level and the presence of large distribution structures.

The survey does not consider so-called "away from home expenditures" (hotels, restaurants, and coffee) and "special source expenditures" (for example, when wine is purchased directly from the producer). In other words, we are working with a data set that only comprises expenditures that occur through large distribution, cash and carry, and traditional commerce channels.

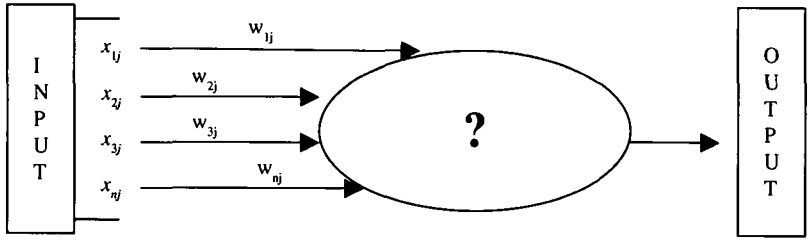
The data have been collected by a "home scanning" method that registers (using bar codes) all products bought by a household, providing for a detailed set of information that includes, e.g., the color of wine (red, white etc.) or the specific type of wine (prosecco, chardonnay, malvasia, etc.).

Since the data are available by geographical region we can analyze the wine consumption of wine in the four main Italian territorial repartitions, the North-West (Piemonte, Valle d'Aosta, Lombardia and Liguria), the North-East (Veneto, Friuli, Trentino and Emilia), the Center (Toscana, Umbria, Marche, Lazio and Sardegna) and the South (Abruzzo, Campania, Calabria, Basilicata, Molise, Puglia and Sicilia).

In addition, the data set provides information, such as education level, social position (employed, unemployed, retired etc.), about the reference person of the household (hereafter R.P.). The Appendix reports the main socio-demographic characteristics of the sample and some aspects of wine consumption connected to those factors.

https://doi.org/10.1017/S1931436100000547 Published online by Cambridge University Press

Figure 1
The generic *j-th* neuron



III. Neural Network Basics

In this section the ANN theory is introduced from an intuitive point of view. For further formal details see Appendix B.

ANN (see for example Haykin, 1999) were originally developed in mathematical neurobiology. However, during the last few years several economics and finance applications have been explored.

The artificial neuron, called "node" in neural network literature, is the analog of the biological neuron. A very simple ANN consists of four elements: a set of inputs (independent variables), a weight vector, a processing step, and an output (dependent variable). Intuitively, a node takes a set of inputs and computes an output according to the transfer function, as in Figure 1. More specifically, the node generates the output by weighting, summing up and transforming input variables (Figure 1).

Usually, the input variable and intermediate nodes are fully connected. That is, every input variable is weighted and combined (summed up and transformed) by every output node.

For a given neuron $j, j = 1,\ldots,N$, the relationship between the input vector $\mathbf{X}_j = (X_{1j},\ldots,X_{ij},\ldots X_{nj})$ and the output Y_j is given by equations (1) and (2):

$$P_j = \sum_{i=1}^{n} w_{ij} X_{ij} \tag{1}$$

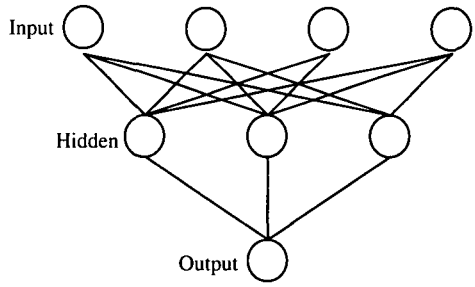
$$Y_j = f(P_j) \tag{2}$$

Where $w_j = (w_{1j}, w_{2j}, \ldots . w_{mj})$ is a weight vector and P_j is sometimes called net potential. As for the choice of function f (transfer function) in equation (2), a variety of functions may be considered (for example product, minimum, maximum), but the logistic function is often used because of its desirable analytical properties.[2]

[2] If a linear transformation function is used, this simple neural network model coincides with the linear regression model.

https://doi.org/10.1017/S1931436100000547 Published online by Cambridge University Press

Figure 2
Topological Structure



An ANN can model a large number of complicated input-output patterns by means of its structure, called the architecture, consisting of an input layer, one or more intermediate (hidden) layers and one or more output layers (Figure 2).

The weight vectors can be estimated in a variety of ways. The most common is a learning algorithm called "back-propagation", which essentially is a two-phase, gradient, descent optimization algorithm. The goal is to minimize the global error defined as follows:

$$E = \sum_{j=1}^{N} \left(Y_j - \hat{Y}_j \right)^2 \tag{3}$$

where, for the j-th neuron Y_j is the real output and $\hat{Y}_j$ is the output estimated by the network. The objective of the estimation process is to iteratively adjust the weight vectors in order to minimize E. If E is minimized the training was successful.

In the training stage, only a part of the data set is presented to the network. The holdback sample (i.e. the part that has not been used for training) is employed afterwards to check the performance of the neural network in reproducing the output. This validation phase is necessary to use the neural network model with any degree of confidence.

In summary, the steps involved to train the network are:

a) present an example of input data;
b) compute a set of outputs using the current connection weights;
c) compare the estimated output with the real output;
d) update the connection weights in order to minimize the global error E.

The global error reduction can be displayed by plotting the global error E against the training iterations (called "epochs" in the ANN literature).

Preprocessing the data is necessary in the ANN approach (as it is in the parametric approach) to avoid the interference induced by possibly different unit measures of the input variables. One possible technique to do this is to normalize the input patterns to the values

https://doi.org/10.1017/S1931436100000547 Published online by Cambridge University Press

https://doi.org/10.1017/S1931436100000547 Published online by Cambridge University Press

in the range [0,1].[3] For more details on the normalization within the ANN framework see Azoff (1994).

To summarize, the main advantages of the ANN approach are:

1) Ability to use data that are incomplete or possibly affected by errors.
2) Robustness to possible contaminations of the data.
3) Flexibility, i.e., capability of operating in complex situations that may vary over time.
4) Capability of generalizing, i.e., producing outputs that correspond to unseen input patterns.

The disadvantages include:

1) ANN is a black box, i.e., no theoretical justification of the results is given.
2) A large number of input/output pairs are necessary to suitably train the ANN.
3) There are no rigorous methods for constructing the best ANN for a given data set.

IV. Elasticity

The impact of price and income (input) changes is usually expressed in terms of elasticities. Given a pair of goods s and l, there are two forms of price elasticity to consider. The first is the own-elasticity which measures the effects of a variation of the price of s on the quantity of s. The second is the cross-price elasticity which measures the effects of a variation of the price s on the quantity of l.

Traditionally, income and price elasticity are determined with closed formulas after the estimation of econometric models. Here we compute the elasticity following different methods that suit the non parametric framework. To our knowledge, prior publications contain only a few of such techniques, and we focus on two of these, both based on the definition of elasticity itself: SR (Subba Rao et al., 1998) and GK (Gruca and Klemz, 1998).

The SR method identifies elasticity by studying how the normalized output Y_j^* varies when one of the input variable varies (in particular, price or income). The steps are as follows:

1. run the ANN for the training sample by increasing the price (income) by different percentages between 1% and 100%;
2. let Y_{ij}^* be the normalized output, obtained after P% change in the i-th input variable (price or income);
3. use the following formula to compute a series of individual elasticities by increasing the i-th input variable (price or income) between 1% to 100%;

[3] Of course, for binary variables the normalization is not needed.

$$SR\varepsilon_{i,j} = \frac{\dfrac{Y_{ij}^* - Y_j^*}{Y_j^*}}{\dfrac{P}{100}} \tag{4}$$

4. calculate the weighted average of these individual elasticities and obtain the demand elasticity.

Thus, by taking the weighted average of the individual elasticities it is possible to obtain the sample demand curve with regards to the entire sample as well as for specific socio-demographic groups of households (e.g., those living in a specific area or possessing a certain education level).

The GK methods, on the other hand, computes elasticities as follows (Gruca and Klemz, 1998):

1. set all input variables, except the i-th price or income, to their mean levels;
2. for every observed value of X_{ij} (price or income), obtain $\hat{Y}_j^*$ with the ANN;
3. change X_{ij} and construct the demand curve;
4. estimate the elasticity using the standard elasticity formula;

$$GK\varepsilon_{ij} = \frac{\Delta \hat{Y}_j^*}{\Delta X_{ij}} \frac{X_{ij}}{\hat{Y}_j^*} \tag{5}$$

Once more, by averaging the individual elasticities across all sample units (or groups of them) one can obtain the sample's demand curve.

V. Empirical Evidence

In this section we present some results obtained by applying the ANN approach to the data set described in Section 2.

First of all, some preprocessing of the data is necessary before presenting the input patterns to the neural network. This usually involves normalizing the input variables which we did by dividing the data by the respective maximum values.

The ANN employed here consists of 1 input layer (22 input variables, listed in Table 1), 1 hidden layer, and 1 output layer[4] which is the quantity of red and white wine consumed by each household.

[4] A 3-layers architecture is adept to model most real situations. Generally, there is no need to include more than one hidden layer.

https://doi.org/10.1017/S1931436100000547 Published online by Cambridge University Press

https://doi.org/10.1017/S1931436100000547 Published online by Cambridge University Press

Table 1
Input Variables Presented to the Artificial Neural Network

Variable	*Name of the input variable*	*Description*
Sex	Gender	*0- Female* and *1- Male*
e1, e2, e3, e4	Age	Dummy: *up to 25 years old* *from 25 to 44 years old* *from 45 to 60 years old* *over 60 years old*
High, medium, low	Education level	Dummy: *high* (post graduate degree, short degree diploma) *medium* (high school) *low* (middle school, primary school, no diploma)
E, O, C, S	Geographic area	Dummy: *North West* *North East* *Centre* *South*
Empl, Unempl, Other	Social position	Dummy: *employed* *unemployed* *other*
r1, r2, r3, r4	Economic status	Dummy: *income less than 260 €* *income between 260 and 420 €* *income between 460 and 620 €* *income over 620 €*
Pb	$Price_i$	Price of the analyzed good (e.g. white wine)
Pr	$Price_j$	Price of the other good (e.g. red wine)
expend	Income	Total wine expenditure is considered an approximation of the income.

The transfer function we use is a linear positive function taking values in [0,1].

To train the network, we followed the back-propagation scheme; the weights have been determined with a Levemberg-Marquardt algorithm (Marquardt, 1963) that seemed the fastest.

To show the performance of the ANN we report a sequence of graphs (Figures 3 to 5). In the training phase (Figure 3), the global error decreases at a rate that tends to decline until it becomes almost constant. The global error variance in the training phase (Figure 4) equals 0.0077 (after 38 training epochs), whereas in the validation phase (Figure 5) it is 0.016. As expected, the global error variance is bigger in the latter case since the ANN uses previously unknown data.

Figure 3
Mean Global Error in the Training Phase

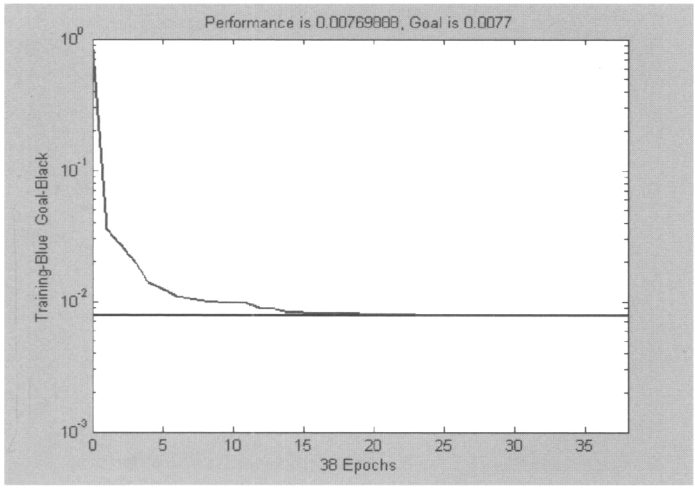
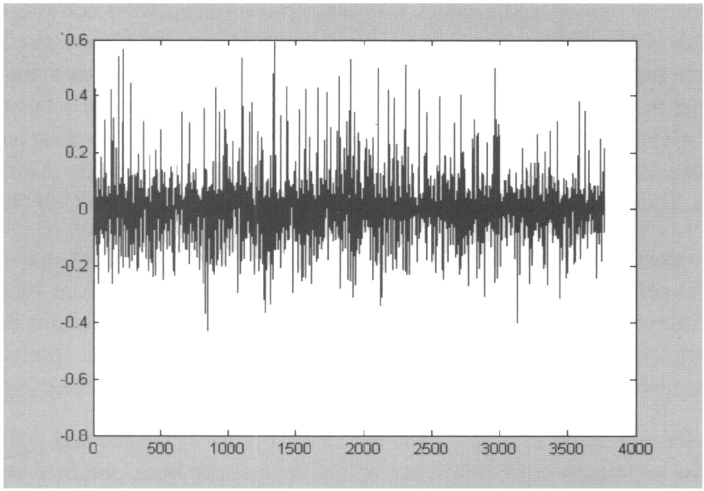


Figure 4
Error in the Training Set

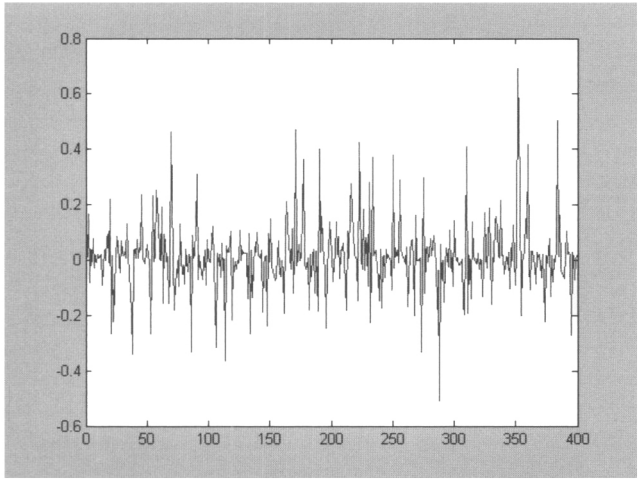


var error_in	0.0077
std dev error_in	0.0874
mean error_in	0.0075

https://doi.org/10.1017/S1931436100000547 Published online by Cambridge University Press

https://doi.org/10.1017/S1931436100000547 Published online by Cambridge University Press

Figure 5
Error in the Validation Set



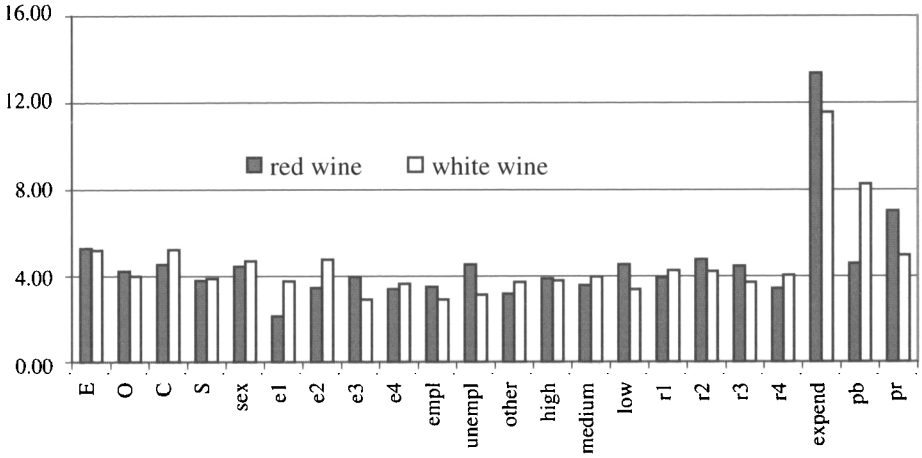
var of error_valid	0.0162
std dev of error_valid	0.1273
mean of error_valid	0.0116

Before estimating the demand curve and elasticities, we analyze the marginal effects of various, especially demographic, input variables on household wine consumption. Following Subba Rao et al. (1998), we employ a partitioning-of-weights algorithm (Garson, 1991) to interpret the input parameters. The connection weights from the input to the hidden layer and those from the hidden to the output layer can be used to evaluate the relative influence of each input variable on the estimated output. The method essentially involves partitioning the hidden output connection weights of each hidden neuron into components associated with the input neuron. The computation process to partition the weights follows Goh (1994).

Figure 6 reports the relative influences of the variables for our data. As expected, income and price ("expend" and "pr", "pb", respectively) are the most relevant for both red and white wine. Nevertheless, demographic variables appear to be important in some cases (e.g., the territorial repartition North East (E) and Center (C)). These results suggest that demographic factors should also be taken into account to study consumers' behaviour.

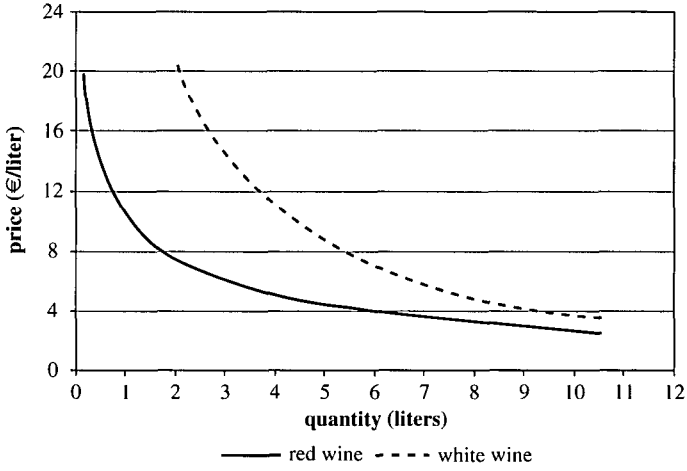
We then determine the sample's demand curve (Figure 7) for red and white wine as an average of the estimated individual curves. The latter have been obtained on the basis of the output produced with the trained ANN by changing the prices by increasing percentages, keeping all the other input variables constant (SR method).

Figure 6
Relative Influence of the Variables for White and Red Wine



E: north-east; O: north-west; C: center; S: south; e1: up to 25 years old; e2: 25 to 44; e3: 46-60; e4: over 60; r1: less than 260 €; r2: 260-420 €; r3: 460-620 €; r4: over 620 €; pb: white wine price; pr: red wine price.

Figure 7
White and Red Wine Demand Curves



https://doi.org/10.1017/S1931436100000547 Published online by Cambridge University Press

In a consumer-oriented market it is certainly more useful to evaluate the demand curve for consumer groups that are homogenous with respect to some socio-demographic variables. In Figures 8 to 11, the red wine demand curves are estimated relative to gender, age, educational and income levels. Examining Figure 9 in particular, note that consumers belonging to the third age class (over 60) are less price-responsive than those belonging to the other age classes.

Figure 8
Red Wine Demand Curve by Gender

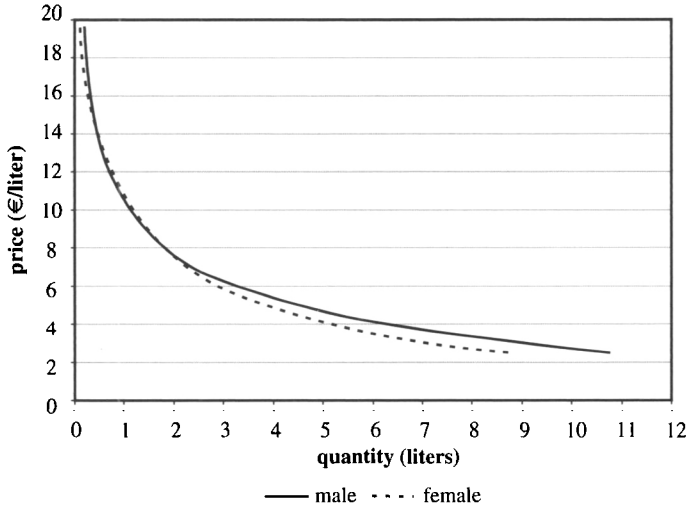
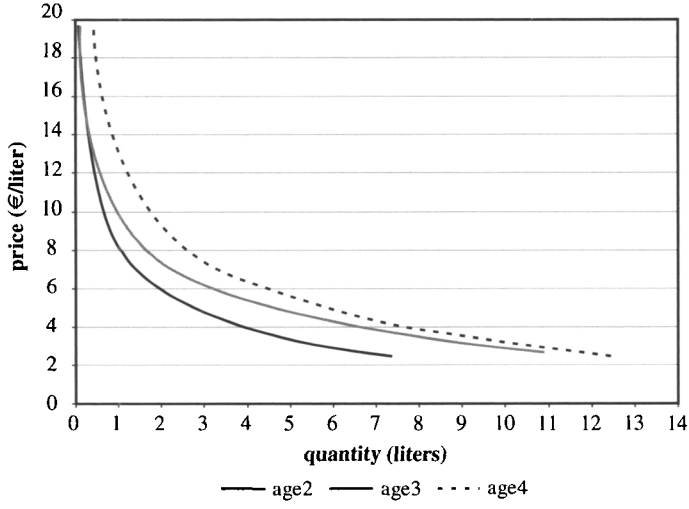


Figure 9
Red Wine Demand by Age



https://doi.org/10.1017/S1931436100000547 Published online by Cambridge University Press

Figure 10
Red Wine Demand by Educational Level

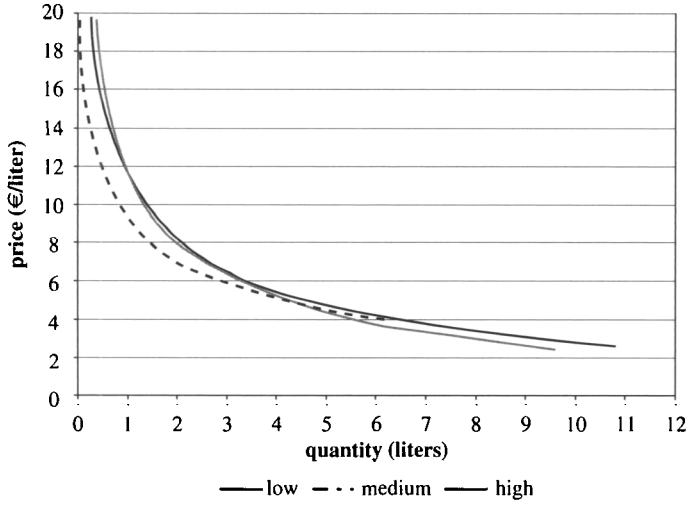
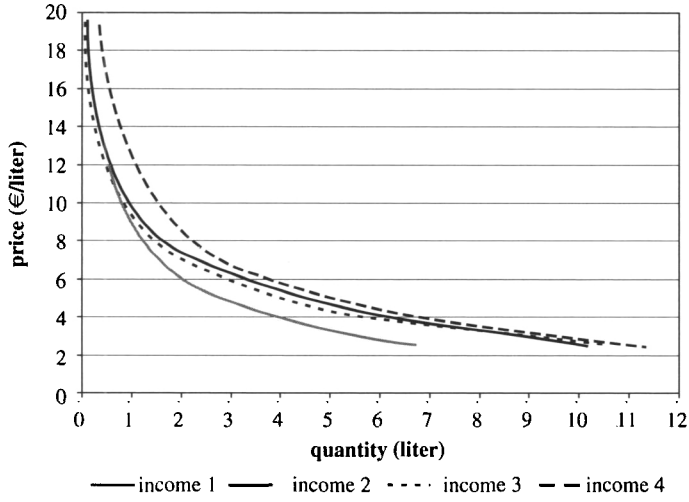


Figure 11
Red Wine Demand Curve – Income



https://doi.org/10.1017/S1931436100000547 Published online by Cambridge University Press

https://doi.org/10.1017/S1931436100000547 Published online by Cambridge University Press

We also determine the demand curve for groups of consumers that are simultaneously homogenous with respect to more than one socio-demographic variable. In particular, focusing on households whose R.P. is male and employed, we identify two profiles: consumer 1 with a low education level, first age class, and low income and consumer 2 with a high education level, third age class, and high income (Figure 12).

In Table 2, we report the weighted average price elasticities independent of socio-demographic family characteristics.

The elasticity values are all smaller than one suggesting a comparatively price-inelastic demand. In Figures 13 to 16, we report selected elasticities computed for each demographic variable with the SR method.

Figure 12
Red Wine Demand by Consumer Group

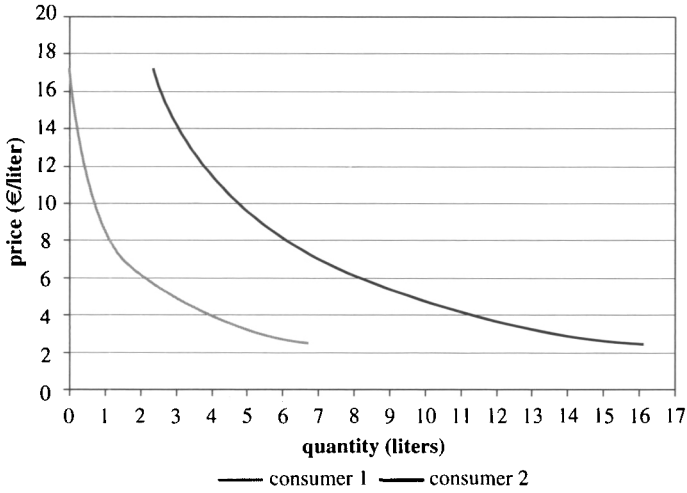


Table 2
Weighted Average Price Elasticity per Typology of Wine

	Price Elasticity	
	own price elasticity	*cross price elasticity*
White wine	–0.67	0.66
Red wine	–0.87	0.85

Figure 13
Own Price Elasticity for White and Red Wine by Geographic Area

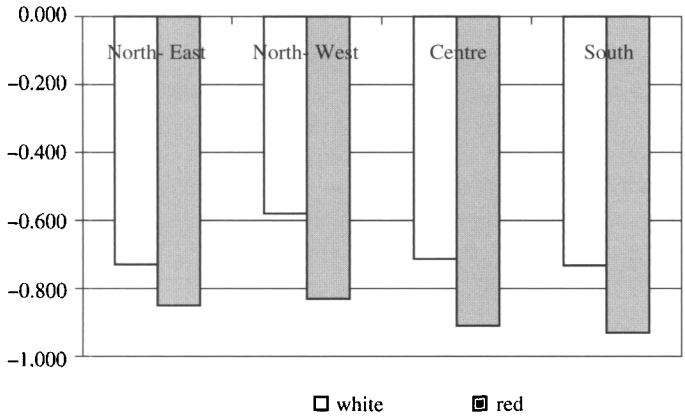
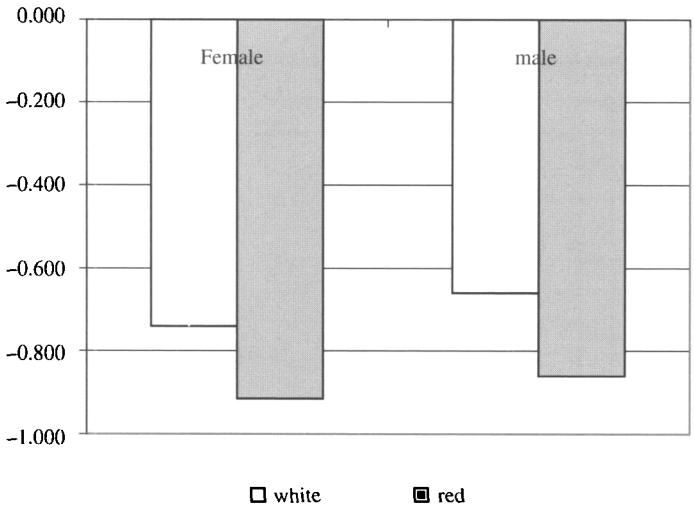


Figure 14
Own Price Elasticity for White and Red Wine by Gender



https://doi.org/10.1017/S1931436100000547 Published online by Cambridge University Press

https://doi.org/10.1017/S1931436100000547 Published online by Cambridge University Press

Figure 15
Own Elasticity for White and Red Wine by Age

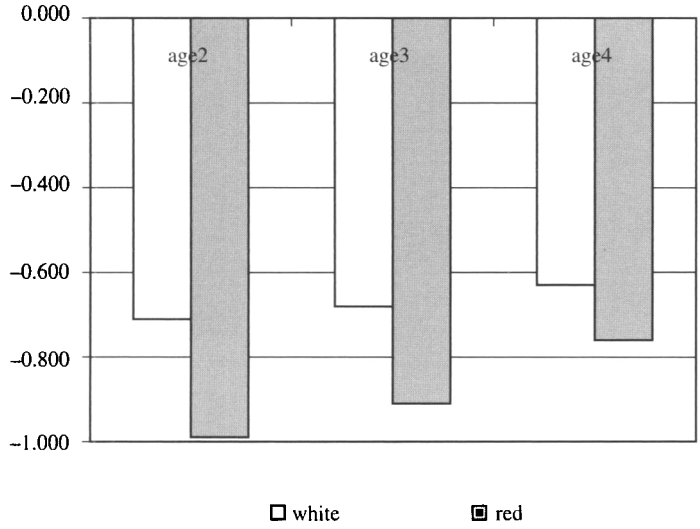


Figure 16
Own Elasticity for White and Red Wine by Income Class

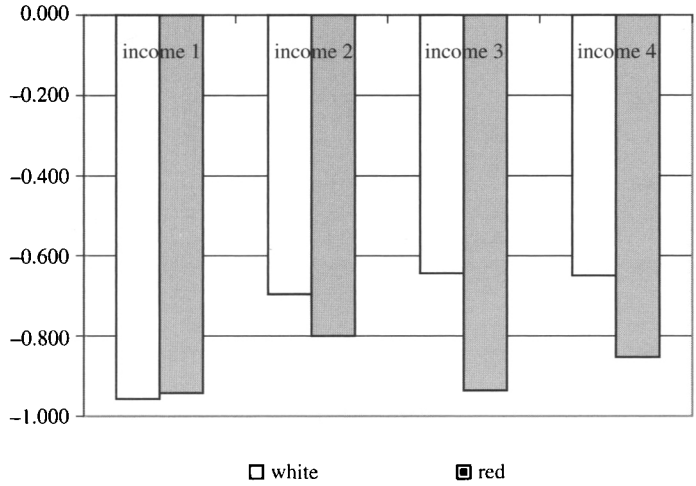


Table 3
Own Price Elasticity (White Wine) by Consumer Profile

Consumer profile	*Gruca Method*	*Subba Rao Method*
Consumer 1	−3.21	−0.97
Consumer 2	−1.18	−0.67

Apparently, wine consumers from the North-West differ from those in other regions. The gender variable also affects elasticity. In fact, women appear significantly more price responsive than men. As expected, the lowest income classes exhibit the highest price elasticities.

With respect to white wine consumption, we analyzed the joint effects of several demographic variables on the elasticity using SR and GK methods in order to compare their performance. For this purpose, we refer to the two consumer profiles previously presented (*consumer 1* and *consumer 2*, respectively).

As indicated in Table 3, the computed elasticity values are considerably different in size. Although the two approaches refer to the usual definition of elasticity, the results should be deemed complementary. The GK method, based on the demand curve estimated from the neural network, computes the price and income elasticity at every point of the demand curve and for the average of the sample ("empirical" elasticity). With the SR method, it is possible to construct individual demand curves from a theoretical point of view by measuring the reaction of a single consumer to different price variations. The results express the average reaction of consumers to a 1% increase in price.

VI. Conclusions

The main objective of this paper is to model the Italian wine demand market using the non parametric approach of ANN. Especially with respect to the food and beverage sector, the ANN approach is relatively novel.

Our findings are the following. First, even with microdata, ANN performs well in terms of generalizing from presented examples to unseen cases. Second, ANN can be used to estimate the demand curve and elasticity not only for the entire sample but also for various consumer groups.

To our knowledge, there has been no ANN analysis using microdata. However, our results are consistent with those obtained by some applied works using time series data (Moro and Canali, 1996). Our analysis emphasizes the relevance of various socio-demographic variables that influence consumer choices along with traditional economic variables (income and price). The simulations for different consumer profiles quantify the influence of specific characteristics on the respective demand. From a marketing point of view, our study suggests that there are distinct consumer groups at whom specific price policies should be aimed.

https://doi.org/10.1017/S1931436100000547 Published online by Cambridge University Press

https://doi.org/10.1017/S1931436100000547 Published online by Cambridge University Press

In general, the differences in price elasticities can be utilized for customer targeted price policies. Absolute elasticities smaller than one may suggest increasing prices and mark-ups. However, this conclusion has to be taken very carefully for two reasons. First, price increases on a wine market that is shrinking may accelerate the negative trend. Second, our price elasticities refer to the overall wine market. Due to the presence of many close substitutes, the price elasticity of each particular wine or wine brand will be substantially higher.

References

Azoff, E.M. (1994). *Neural Network Time Series Forecasting of Financial Markets*. New York: Wiley Financial Edition.

Bishop, C.M. (1995). *Neural Networks for Pattern Recognition*. Oxford: Clarendon Press.

Canali, G. and Moro, D. (1997). La domanda di bevande in Italia: un'analisi della sostituibilità e dei cambiamenti strutturali. *Annali della Facoltà di Agraria*. Università Cattolica del Sacro Cuore, Milano.

Deaton, A. and Muellbauer, J. (1980). An almost ideal demand system. *American Economic Review*, 70(3), 312–326.

Fine, T.L. (1999). *Feedforward Neural Network Methodology*. Berlin and New York: Springer Verlag.

Garson, D.G. (1991). Interpreting neural-network connection weights. *AI Expert*, 6(4), 46–51.

Goh, A.T.C. (1994). Seismic liquefaction potential assessed by neural networks. *Journal of Geotechnical Engineering*, 120(9), 1467–1479.

Gruca, T.S. and Klemz, B.R. (1998). Using neural networks to identify competitive market structures from aggregate market response data. *Omega International Journal of Management Science*, 26 (1), 49–62.

Hagan, M.T., Demuth, H.B. and Beale, M. (1999). Neural Network Design. Boston: PWS Publishing Company.

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*. New York: Macmillan.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*. Berlin and New York: Springer Verlag.

Marquardt, D.W. (1963). An algorithm for least squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 2, 431–441.

Mauracher, C. (2005). Le variabili sociodemografiche nel consumo di vino: un'analisi mediante Reti Neurali Artificiali. Ph.D. thesis, Padova University.

Subba Rao, P.V., Sikdar, P.K., Krishna Rao, K.V. and Dhiringra, S.L. (1998). Another insight into artificial neural networks through behavioural analysis of access mode choice. *Computational Environmental and Urban Systems*, Vol. 22, pp. 485–496.

Theil, H. (1965) The information approach to demand analysis. *Econometrica,* Vol. 33, pp. 67–87.

White, H. (1992). Artificial Neural Networks, Springer Verlag, New York.

https://doi.org/10.1017/S1931436100000547 Published online by Cambridge University Press

Appendix A – The sample

In this Appendix we briefly summarize the main socio-demographic characteristics of the sample and some aspects of wine consumption that are connected to the socio-demographic variables.

The "high level education" class includes people who obtained at least a short graduation diploma (*laurea breve*), whereas those who obtained a high school and middle school diploma belong to the "medium and low level education" classes, respectively.

The "Other" category includes housewives, students, people unable to work, retirees and army enlistees.

Table A.1

Socio-Demographic Characteristics of the Sample

		Abs. frequency	*Percentage*
Sex	male	3.709	87.37%
	female	536	12.63%
Age	Up to 44 years old	1.110	26.15%
	from 45 to 60 years old	1.825	42.99%
	over 60 years old	1.310	30.86%
Educational level	low	1.748	41.18%
	medium	1.941	45.72%
	high	556	13.10%
Geographic area	North East	708	16.68%
	North West	1.266	29.82%
	Centre	840	19.79%
	South	1.431	33.71%
Social position	employed	2.871	67.64%
	unemployed	60	1.41%
	other	1.314	30.95%
Economic status	less than 260 €	365	8.60%
	between 260 and 420 €	972	22.90%
	between 420 and 620 €	1.241	29.23%
	Over 620 €	1.667	39.27%

https://doi.org/10.1017/S1931436100000547 Published online by Cambridge University Press

Table A.2
Average Red and White Wine Consumption per Socio-Demographic Characteristic

		white wine	*red wine*
Sex	male	11.50	12.19
	female	10.81	10.97
Age	Up to 44	8.90	8.86
	from 45 to 60	12.13	12.58
	over 60	12.51	13.93
Educational level	low	10.74	11.51
	medium	12.11	12.71
	high	11.07	11.23
Geographic area	North East	11.17	11.07
	North West	13.63	15.82
	Centre	13.95	13.36
	South	8.12	8.45
Social position	employed	11.43	11.95
	unemployed	11.45	12.91
	other	9.39	13.71
Economic status	less than 260 €	7.61	8.34
	between 260 and 420 €	10.43	11.12
	between 420 and 620 €	11.87	12.14
	over 620 €	12.47	13.27

Appendix B – Methodological Note on Artificial Neural Network

ANN are a wide class of learning methods used to carry out two-stage regression or classification models. The main idea of learning methods is to extract linear combinations of the inputs as derived features, and then model the output as nonlinear function of these features.

In a very extensive ANN setting[5] the derived features Z_m are obtained from linear combinations of the inputs $\mathbf{X}$ and then the output Y_k is modeled as a function of linear combinations of the Z_m:

$$\begin{aligned} Z_m &= \sigma(\alpha_{0m} + \alpha_m^T \mathbf{X}), \ m = 1,...,M, \\ T_k &= \beta_{0k} + \beta_k^T \mathbf{Z}, \ k = 1,...,K, \\ Y_k &= f_k(\mathbf{X}) = g_k(\mathbf{T}), \ k = 1,...,K \end{aligned} \tag{B.1}$$

[5] In this Appendix the notation follows Hastie et al. (2001) where also more formal details on the ANN topic can be found.

https://doi.org/10.1017/S1931436100000547 Published online by Cambridge University Press

where $\mathbf{Z} = (Z_1, Z_2, ..., Z_M)$, $\mathbf{T} = (T_1, T_2, ..., T_K)$, $\sigma(\cdot)$ is called the activation function, and g_k is called the output function. The function obtained by the composition of $g_k(\cdot)$ and $\sigma(\cdot)$ is called the transfer function.

The units in the middle of the network are called hidden units because their values are not directly observed. In general, there can be more than one hidden layer. The neural network has unknown parameters, often called weights. Their values are determined so that the model fits the training data well. The complete set of weights, denoted by θ, consists of:

$$\begin{aligned} &\{\alpha_{0m}, \alpha_m;\ m = 1, ..., M\} \quad M(p+1) \text{ weights}, \\ &\{\beta_{0k}, \beta_k;\ k = 1, ..., K\} \quad K(M+1) \text{ weights} \end{aligned} \tag{B.2}$$

As a measure of fit (loss function) the sum of squared errors is often used:

$$R(\theta) = \sum_{k=1}^{K} \sum_{i=1}^{N} (y_{ik} - f_k(x_i))^2 \tag{B.3}$$

The usual approach to minimizing $R(\theta)$ is by gradient descent, called back propagation, which can easily be derived using the chain rule for differentiation because of the compositional form of the model. Let us write $R(\theta)$ as follows:

$$R(\theta) = \sum_{i=1}^{N} R_i = \sum_{i=1}^{N} \sum_{k=1}^{K} (y_{ik} - f_k(x_i))^2 \tag{B.4}$$

The derivates are:

$$\begin{aligned} \frac{\partial R_i}{\partial \beta_{km}} &= -2(y_{ik} - f_k(x_i)) g_k'(\beta_k^T z_i) z_{mi}, \\ \frac{\partial R_i}{\partial \alpha_{ml}} &= -\sum_{k=1}^{K} 2(y_{ik} - f_k(x_i)) g_k'(\beta_k^T z_i) \beta_{km} \sigma'(\alpha_m^T x_i) x_{il} \end{aligned} \tag{B.5}$$

A gradient descent update at the $(r+1)st$ iteration has the form

$$\begin{aligned} \beta_{km}^{(r+1)} &= \beta_{km}^{(r)} - \gamma_r \sum_{i=1}^{N} \frac{\partial R_i}{\partial \beta_{km}^{(r)}}, \\ \alpha_{ml}^{(r+1)} &= \alpha_{ml}^{(r)} - \gamma_r \sum_{i=1}^{N} \frac{\partial R_i}{\partial \alpha_{ml}^{(r)}}, \end{aligned} \tag{B.6}$$

where γ_r is called learning rate. Now it is possible to write:

$$\begin{aligned} \frac{\partial R_i}{\partial \beta_{km}} &= \delta_{ki} z_{mi}, \\ \frac{\partial R_i}{\partial \alpha_{ml}} &= s_{mi} x_{il} \end{aligned} \tag{B.7}$$

The quantities δ_{ki} and s_{mi} are "errors" relative to the model at the output and hidden layers, respectively. From their definition, these errors comply with

$$s_{mi} = \sigma'(\alpha_m^T x_i)\sum_{k=1}^{K}\beta_{km}\delta_{ki} \tag{B.8}$$

known as the back-propagation equations.

The updating process (formula B.6) can be implemented by means of a two-pass algorithm. In the forward pass the current weights are fixed and the estimated output $\hat{Y}_k$ are calculated from formula B.2. In the backward pass, first the errors δ_{ki} then the errors s_{mi} are obtained (from B.8). On the basis of δ_{ki} and s_{mi}, the gradient is computed to finally implement the updating process.

https://doi.org/10.1017/S1931436100000547 Published online by Cambridge University Press