# Assessment of the chain dependence relationships between wine grape quality, soil, and geological substrate using a metric generalization of PLS regression

*Pietro Amenta*
*Department of Analysis of Economic and Social Systems,*
*University of Sannio, Via delle Puglie 3, 82100, Benevento, Italy*
*amenta@unisannio.it*

*Antonio P. Leone*
*CNR-ISAFoM, via Patacca, 85, 80056 Ercolano, Italy*
*a.leone@isafom.cnr.it*

*Antonello D'Ambra*
*Department of Analysis of Economic and Social Systems,*
*University of Sannio, Via delle Puglie 3, 82100, Benevento, Italy*
*andambra@unisannio.it*

**Abstract:** *A study was conducted to assess the influence of soil properties and underlying geology on the characteristics of Falanghina wine grape in a viticultural area of southern Italy. Soil, along with climate, is one of the main physical-environmental factors affecting the composition of wine grape. Soil properties reflect to a more o less extent the characteristics of underlying geology, depending on the nature of geological substrate and the age of soil. Therefore, if relationships can be found between soil properties and grape characteristics, an indirect influence of geological substrate on grape composition should be expected. The structure of geology-soil-grape data is then characterized by a chain of dependence relationships between the three sets of variables. As such, it can evaluated using a single theoretical approach based on the metric generalization of PLS regression.*

**Keywords:** wine grape quality, soil properties, geology, Generalized PLS Regression.

## 1. Introduction

There is a widespread agreement that wine grape characteristics result from the combined effects of a number of factors, such as genetic, anthropogenic and physical-environmental. While anthropogenic and genetic factors can be modified, the physical-environmental factors are substantially stable and poorly adjustable, so that they represent crucial features, determining the specificity and distinctiveness of the wine made from grape juice fermentation.

The influence of physical-environmental factors on grape characteristics has enjoyed research attention. Many wine scientists, particularly from New World countries emphasised the central importance of climate on wine grape production, relegating the role of soil as secondary to climate. Conversely, the majority of scientists from European countries (especially from France) associated grape quality with the type of soil from which the grapes are produced. Against the above controversy, a need exists to improve knowledge on the effects of climate and soil on grape and wine quality. To contribute to this need, a research activity has recently been undertaken by the CNR-ISAFoM in the Telesina Valley, an important viticultural area of southern Italy. As a first step in this research, the influence of basic soil properties on the Falanghina wine grape quality was investigated (Leone et al., 2006). Falanghina (from which a namesake wine is produced) is one of the most celebrated wine grape cultivar of the study area, and, more generally, of the southern Italy.

Soil properties may be significantly affected by the underlying geology, depending on the nature of geological substrate from which soils originate and the age of soil (i.e., the time of soil for-

 

mation). Therefore, if relationships can be found between soil properties and grape characteristics, an indirect influence of geological substrate on these characteristics should be expected. The structure of geology-soil-grape data is, then, characterized by a chain of dependence relationships between three sets of variables: the grape characteristics are influenced by the soil properties, which, in turn, are affected by nature of geological substrate.

This kind of problem can be statistically tackled in a "two-stages least squares" (TSLS) framework. If we are not in the presence of quasi collinearity among variables or in situations where the number of variables is large enough with respect to the number of samples, then a TSLS regression can be performed. TSLS regression refers to a first stage in which new dependent variables are created to substitute for the original ones regressing soil properties onto the geological substrate. A second stage is then performed, in which the grape characteristics are regressed onto the newly created variables. Following the same strategy, a two-stage PLS regression can be suggested to overcome the quasi collinearity conditions among variables or when the number of variables is large with respect to that of samples. However, we should remark that in both cases the newly created variables are computed without taking into account the grape characteristics and not in a single theoretical approach. The main objective of this paper is to evaluate the dependence chain relationships between geology, soil and grape characteristics in the Telesina Valley within a single theoretical approach without considering a two-stage strategy.

## 2. Notation and structure of the data

Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}^T$ be a matrix of order $n \times p$ collecting the values taken by $n$ statistical units on $p$ explanatory variables. We consider the notation of the statistical study (triplet) $(\mathbf{X}, \mathbf{Q}_\mathbf{X}, \mathbf{D})$ (Escoufier, 1987) to describe the data and their use. The triplet allows to present the factorial methods in a single theoretical framework by using suitable choices for the metrics $\mathbf{Q}_\mathbf{X}$ and $\mathbf{D}$, where $\mathbf{D} = diag(d_1, ..., d_n)$ specifies the weights metric in the vectorial space $\Re^n$ of variables with $\sum_{i=1}^{n} d_i = 1$ (herafter $d_i = 1/n$) and $\mathbf{Q}_\mathbf{X}$ defines the metric measuring the distance between the data vectors $\mathbf{x}_j$, $\mathbf{x}_k$ of two statistical units *j*, *k* in $\Re^p$ given by $(\mathbf{x}_j - \mathbf{x}_k)^T \mathbf{Q}_\mathbf{X} (\mathbf{x}_j - \mathbf{x}_k)$. We assume that $\mathbf{X}$ is mean centred with respect to $\mathbf{D}$ ($\mathbf{1}_n^T \mathbf{D} \mathbf{X} = \mathbf{0}$ with $\mathbf{1}_n$ unitary column vector). Moreover, let $(\mathbf{Y}, \mathbf{Q}_\mathbf{Y}, \mathbf{D})$ and $(\mathbf{Z}, \mathbf{Q}_\mathbf{Z}, \mathbf{D})$ be two statistical studies associated with the matrices $\mathbf{Y}$ and $\mathbf{Z}$ of order $(n \times q)$ and $(n \times r)$, respectively, collecting additional sets of $q$ and $k$ criterion variables observed on the same $n$ statistical units. $\mathbf{Q}_\mathbf{Y}$ and $\mathbf{Q}_\mathbf{Z}$ are the $(q \times q)$ and $(r \times r)$ metrics of the statistical units in $\Re^q$ and $\Re^r$, respectively. Finally, we highlight that the set $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_q\}$ plays the role of explanatory or criterion variables with respect to the sets of variables collected in $\mathbf{Z}$ and $\mathbf{X}$, respectively, such that to be the ring of the chain of the dependence relationships between $\mathbf{Z}$ and $\mathbf{X}$.

## 3. Few words on the Generalized Partial Least Squares

We recall the PLS definitions according to Tenenhaus (1998), which have been generalized by Cazes (1997) (GPLS). The PLS regression of $(\mathbf{Y}, \mathbf{Q}_\mathbf{Y}, \mathbf{D})$ with respect to $(\mathbf{X}, \mathbf{Q}_\mathbf{X}, \mathbf{D})$ uses two statistical studies (triplets) observed on the same statistical units and weighted by $\mathbf{D}$. The PLS regression is an iterative method maximizing the objective function $\text{cov}(\mathbf{Y}\mathbf{Q}_\mathbf{Y}\mathbf{c}, \mathbf{X}\mathbf{Q}_\mathbf{X}\mathbf{w})$ with constraints on the axes $\| \mathbf{c} \|^2_{\mathbf{Q}_\mathbf{Y}} = \| \mathbf{w} \|^2_{\mathbf{Q}_\mathbf{X}} = 1$. At each step $s$ this objective function is maximized by replacing $\mathbf{Y}$ and $\mathbf{X}$, respectively, with the residual matrices $\mathbf{Y}^{(s-1)}$ and $\mathbf{X}^{(s-1)}$ obtained by the $\mathbf{D}$-

 

orthogonal projections of $\mathbf{Y}$ and $\mathbf{X}$ onto the subspace spanned by the $s-1$ PLS components of $\mathbf{X}$ of inferior order $(\mathbf{t}_k = \mathbf{X}^{(k-1)}\mathbf{Q}_\mathbf{X}\mathbf{w}_\mathbf{k}; k = 1,\ldots,(s-1))$. At step $s = 1$, we have $\mathbf{X}^{(0)} = \mathbf{X}$ and $\mathbf{Y}^{(0)} = \mathbf{Y}$. The axis $\mathbf{w}_s$, associated with $\mathbf{X}$, is given by the eigenvector corresponding to the highest eigenvalue $\lambda^2$ of $\mathbf{X}^{(s-1)T}\mathbf{D}\mathbf{Y}^{(s-1)}\mathbf{Q}_\mathbf{Y}\mathbf{Y}^{(s-1)T}\mathbf{D}\mathbf{X}^{(s-1)}\mathbf{Q}_\mathbf{X}$ and the objective function maximum equals to $\lambda$. By permuting $\mathbf{X}$ and $\mathbf{Y}$ ($\mathbf{Q}_\mathbf{X}$ and $\mathbf{Q}_\mathbf{Y}$) into previous equation, it is possible to compute the axis $\mathbf{c}_s$. Alternatively, axis $\mathbf{c}_s$ can be obtained by the transition formula $\mathbf{c}_s = (1/\lambda)\mathbf{Y}^{(s-1)T}\mathbf{D}\mathbf{X}^{(s-1)}\mathbf{Q}_\mathbf{X}\mathbf{w}_s$ with an analogous formula to pass from $\mathbf{c}_s$ to $\mathbf{w}_s$.

## 4. The method

In order to evaluate the dependence chain relationships between geology, soil and grape characteristics within a single theoretical approach, we propose to consider the Cazes's generalized PLS method with a suitable choice of the metrics, taking into account the relationships between the sets of variables. We call this approach "B-Crossed PLS Regression". In this context, as the set $\mathbf{Y} = \{\mathbf{y}_1,\ldots,\mathbf{y}_q\}$ plays the role of ring of the chain of the dependence relationships between $\mathbf{Z}$ and $\mathbf{X}$, we can directly study the influence of the geological substrate ($\mathbf{X}$) onto the grape characteristics ($\mathbf{Z}$) by using the metric $\mathbf{B}\mathbf{B}^T$ as $\mathbf{Q}_\mathbf{X}$ where the matrix $\mathbf{B}$, of order $(p \times pq)$, is given by

$$\mathbf{B} = \begin{pmatrix} \mathbf{b}_{(1)}^T & & 0 \\ & \ddots & \\ 0 & & \mathbf{b}_{(p)}^T \end{pmatrix} \tag{1}$$

with $\mathbf{b}_{(j)}^T = (\mathbf{x}_j^T\mathbf{x}_j)^{-1}\mathbf{x}_j^T\mathbf{Y} = [(\mathbf{x}_j^T\mathbf{x}_j)^{-1}\mathbf{x}_j^T\mathbf{y}_1,\ldots,(\mathbf{x}_j^T\mathbf{x}_j)^{-1}\mathbf{x}_j^T\mathbf{y}_q]$ vector of the simple regression coefficients of each response $\mathbf{y}_l$ $(j = 1,\ldots,q)$ on each explanatory variable $\mathbf{x}_j$ $(j = 1,\ldots,p)$. It is evident that each vector $\mathbf{b}_{(j)}^T$ provides information on the dependence relationships between $\mathbf{Y}$ and $\mathbf{X}$. If the column variables of $\mathbf{X}$ are also standardized ($\tilde{\mathbf{x}}_j$) then each above quantity is given by the scalar products of $\tilde{\mathbf{x}}_j$ with each $\mathbf{y}_l$ $(l = 1,\ldots,q)$, that is $\mathbf{b}_{(j)}^T = \tilde{\mathbf{x}}_j^T\mathbf{Y} = [\tilde{\mathbf{x}}_j^T\mathbf{y}_1,\ldots,\tilde{\mathbf{x}}_j^T\mathbf{y}_q]$.

In literature several authors proposed suitable uses of the matrix $\mathbf{B}$ within multidimensional data approaches. For instance, Garthwaite (1994) shows that the univariate PLS latent variables $\mathbf{t}_k$ can be obtained as the weighthed averages of the simple regressions of response variable $\mathbf{y}$ on each explanatory variable $\mathbf{x}_j$, that is $\mathbf{t}_k \propto \sum_{j=1}^{p}(\mathbf{x}_j^T\mathbf{x}_j)\mathbf{x}_j(\mathbf{x}_j^T)^{-1}\mathbf{x}_j^T\mathbf{y} = (\mathbf{x}_j^T\mathbf{x}_j)\mathbf{P}_{\mathbf{x}_j}\mathbf{y}$ where $\mathbf{P}_{\mathbf{x}_j}$ is the orthogonal projection operator onto the subspace spanned by $\mathbf{x}_j$. If $\mathbf{X}$ matrix is also standardized then the PLS latent variable $\mathbf{t}_k$ amounts to the simple average of the $p$ vectors $\tilde{\mathbf{x}}_j\tilde{\mathbf{x}}_j^T\mathbf{y}$ and it can be expressed as $\mathbf{t}_k \propto \tilde{\mathbf{X}}\tilde{\mathbf{X}}^T\mathbf{y} = \hat{\mathbf{Y}}\mathbf{1}_p$ with $\hat{\mathbf{Y}} = \tilde{\mathbf{X}}\mathbf{B}$ and where $\mathbf{B}$ is a diagonal matrix with diagonal elements equal to $\{\tilde{\mathbf{x}}_j^T\mathbf{y}\}$. An analogous approach was also proposed for multivariate PLS by the above author. Merola and Abraham (2000) suggested an alternative method to PLS called "Principal Components of Simple Least Squares" in which the latent variables $\mathbf{t}_k$ are solutions of the generalized principal components problem $\min_{\mathbf{t}_k^T\mathbf{t}_{k'}=\delta_{kk'}} \left\|\hat{\mathbf{Y}} - \mathbf{t}_l\mathbf{t}_k^T\hat{\mathbf{Y}}\right\|^2$ with $\hat{\mathbf{Y}} = \tilde{\mathbf{X}}\mathbf{B}$ of order $(n \times pq)$ according to (1). This approach amounts to a weighted principal component analysis of the explanatory variables using the coefficients of determination of the simple regressions as weights. The latent variables $\mathbf{t}_k$

are then the weighted principal components of $\tilde{\mathbf{X}}$ and the coefficients are given by $\mathbf{BB}^T\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\mathbf{w} = \lambda\mathbf{w}$. Finally, D'Ambra et al. (2005) proposed a strategy (Crossed regression approach) to investigate the dependence structure between $\mathbf{X}$ and $G$ sets of response variables $\mathbf{Y}^g$ $(g = 1,...,G)$ of order $(n \times q_g)$ observed on the same $n$ satistical units. The Crossed regression approach is performed using the $\tilde{q} \times J$ $(\tilde{q} = \sum_{g=1}^{G} q_g)$ simple linear regressions of each generic $j$-th column of the $g$-th matrix $\mathbf{Y}^g$ against each $\mathbf{x}_j$. In this way $\tilde{q}$ matrices $\hat{\mathbf{Y}}^g = \mathbf{XB}_g$ are obtained, where $\mathbf{B}_g$ is a diagonal matrix with the weighted regression coefficients $\mathbf{b}_{(j)}^g$ according to (1). The authors perform then a Multiple Co-Inertia Analysis (Chessel and Hanafi, 1996) of the $\tilde{q}$ matrices $\hat{\mathbf{Y}}^g$ in order to analyze the common structure. It is evident that all the authors correspond then a vector of "weights" $\mathbf{b}_{(j)}^T$ to each explanatory variable $\mathbf{x}_j$ collected in the matrix $\mathbf{B}$, in order to take into account the dependence relationships between both sets of variables. So it could be interesting to study the influence of the explanatory variables $\mathbf{X}$ onto the response variables $\mathbf{Z}$ by weighting the former with the relationships with the ring variables $\mathbf{Y}$. This last set of variables will enter in the PLS analysis by the matrix of weights $\mathbf{B}$.

B-Crossed PLS Regression (B-C-PLSR) is then defined as the PLS analysis of the triplets $\{\mathbf{Z}, \mathbf{I}_k, \mathbf{D}\}$ and $\{\mathbf{X}, \mathbf{BB}^T, \mathbf{D}\}$ where $\mathbf{B}$ is defined according to (1). It is easy to show that the B-C PLSR axis $\mathbf{w}_s$ of order $s$ maximizes the criteria $\max_{\mathbf{w}_s, \mathbf{c}_s} \text{cov}^2(\mathbf{X}^{(s)}\mathbf{BB}^T\mathbf{w}_s, \mathbf{Z}^{(s)}\mathbf{c}_s)$ with the constraints $\|\mathbf{w}_s\|^2_{\mathbf{BB}^T} = 1$, $\|\mathbf{c}_s\|^2 = 1$. The B-C-PLSR solution of order $s$ is then given by the vector $\mathbf{w}_s = (\mathbf{BB}^T)^{-1/2}\tilde{\mathbf{w}}_s$ where the eigenvector $\tilde{\mathbf{w}}_s$ is linked to the higher eigenvalue $\lambda$ of the eigen-system $(\mathbf{BB}^T)^{1/2}\mathbf{X}^{(s-1)T}\mathbf{DZZ}^T\mathbf{DX}^{(s-1)}(\mathbf{BB}^T)^{1/2}\tilde{\mathbf{w}}_s = \lambda\tilde{\mathbf{w}}_s$ with $\mathbf{t}_s = \mathbf{X}^{(s)}\mathbf{BB}^T\mathbf{w}_s$.

Finally, we highlight that the B-Crossed PLS Regression is also equivalent to the PLS analysis of the triplets $\{\mathbf{Z}, \mathbf{I}_k, \mathbf{D}\}$ and $\{\hat{\mathbf{Y}}, \mathbf{I}_{pq}, \mathbf{D}\}$ with $\hat{\mathbf{Y}} = \mathbf{XB}$, which allows a graphical representation of all set of variables. In this case, the B-C-PLSR solution of order $s$ is given by $\mathbf{w}_s = \lambda^{-1/2}\mathbf{B}^T\mathbf{X}^T\mathbf{DZc}_s$ where $\mathbf{c}_s$ is the eigenvector linked to the higher eigenvalue $\lambda$ of the eigen-system $\mathbf{Z}^T\mathbf{D}\hat{\mathbf{Y}}\hat{\mathbf{Y}}^T\mathbf{DZc}_s = \lambda\mathbf{c}_s$. For this B-C-PLSR, it would be better to use this last eigen-system instead of $\mathbf{B}^T\mathbf{X}^{(s-1)T}\mathbf{DZZ}^T\mathbf{DX}^{(s-1)}\mathbf{Bw}_s = \lambda\mathbf{w}_s$ cause of the lower computational complexity of the former.

## Bibliography

Cazes P. (1997), Adaptation de la régression PLS au cas de la régression aprés analyse des correspondances multiples, Revue de Statistique Appliquées, XLV(2): 89-99.

Chessel D., Hanafi M. (1996), Analyse de la co-inertie de K nuages de points, Revue Statistique Appliquée, XLIV, 35-60.

D'Ambra L., Amenta P., Gallo M. (2005), Dimensionality Reduction Methods, Metodološki zveski, vol. 2, 1: 115-123.

Escoufier Y. (1987), The duality diagram: a means of better practical applications, in: Legendre P, Legendre L. (Eds.) Development in numerical ecology. NATO advanced Institute, Springer Verlag, Berlin.

Garthwaite P.H. (1994), An interpretation of partial least squares, JASA, 89: 122-127.

Leone A.P., Buondonno A., Basile A., Letizia A., Masotta G. (2007). Influeza delle proprietà dei suoli sulle caratteristiche dell'uva Falanghina nel comprensorio viticolo della Valle Telesina (BN). In SISS "Bollettino Atti del Convegno Nazionale Suolo Ambiente Paesaggio", 66-74

Merola G.M., Abraham B. (2000), Principal Components of Simple Least Squares: A New Weighting Scheme for Principal Component Regression, Technical Reports IIQP RR-00-06, University of Waterloo.

Tenenhaus M. (1998), La Règression PLS. Thèorie et pratique, Editions Technip, Paris.

 