Montpellier University
Faculty of Science
MSc Statistics & Data Science

The contribution of climatic conditions to wine quality.

Multivariate Data Analysis

-

Two tables methods

Author :
Anas Zakroum

March 9, 2023

Contents

Introduction

We have 34 observations of wines from Bordeaux presenting three levels of quality: good, medium, and poor described by:

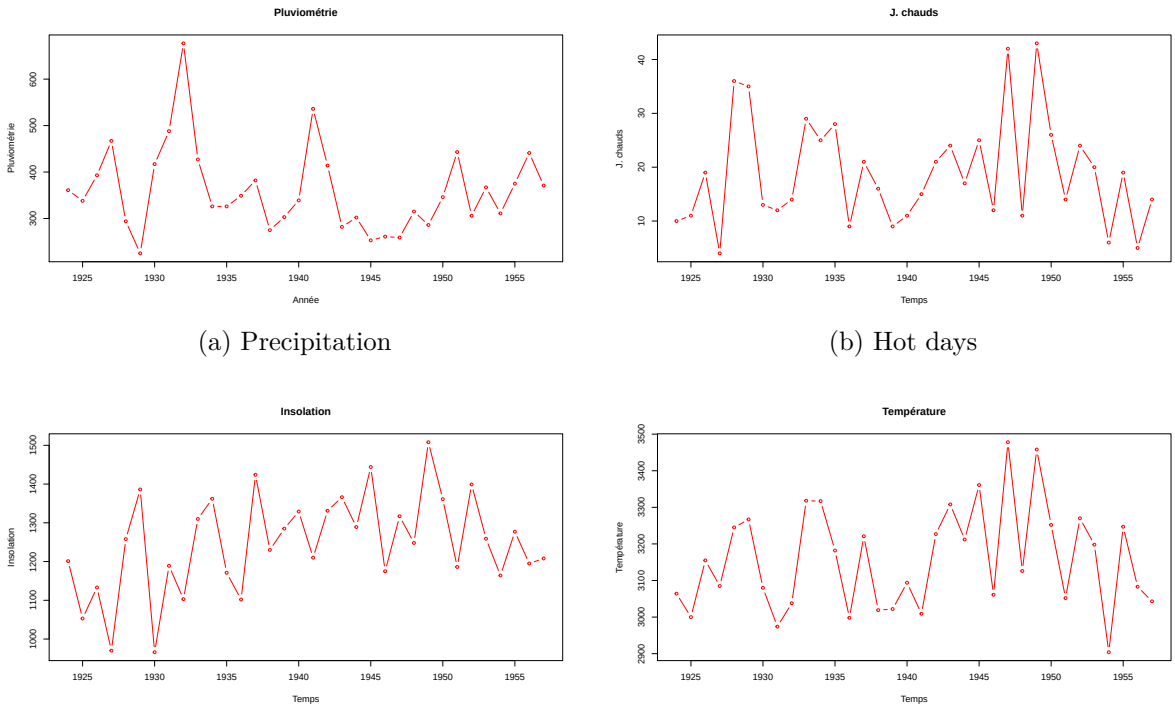
- The year of harvest
- Sum of the average daily temperatures over the year ($C^{\check{r}}$)
- Duration of sunshine (in hours)
- Number of very hot days
- Precipitation (in millimeters)

The goal of this study is to extract from the collected data, the characteristics that allow the best discrimination among the three levels of wine quality as well as to investigate where this difference lies, and finally, to construct a model suited for the task of classification of wine quality.

1 Description of the data

Temporal changes of the descriptors

The wine data was collected between 1924 and 1957. The shape of the time series of the descriptors in Figure 1 is non-stationary and does not appear to visually exhibit linear or particular trends.



(a) Precipitation

(b) Hot days

(c) Insolation

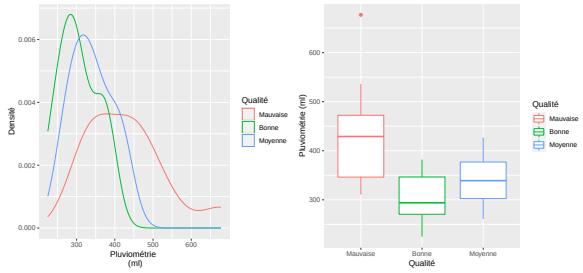
(d) Temperature

Figure 1: Temporal evolution of the descriptors.

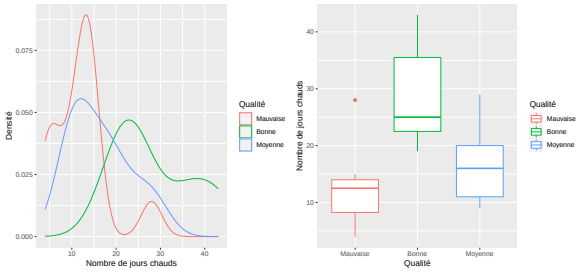
Descriptive statistics of the wine qualities.

The 34 wines are distributed according to their quality as follows: 12 of poor quality, 11 of medium quality, and 11 of good quality. We first propose to provide a description of the distributions of the measured variables according to the wine qualities.

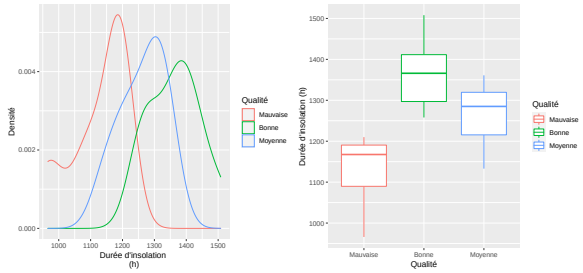
At first glance, we observe from Figure 2 that the majority of the distributions have approximately Gaussian shapes. The medians of the descriptors seem to be clearly differentiated between the three wine qualities. However further investigations need to be carried to look for statistical significance.



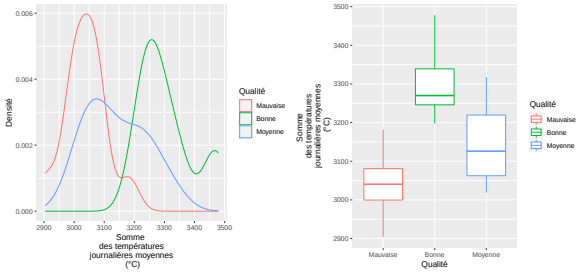
(a) Precipitation



(b) Hot Days



(c) Insolation



(d) Temperature

Figure 2: Distributions of observed data according to wine quality.

However, it appears that good quality wines stand out during hot years, when heat descriptors (insolation, temperature, hot days) are at their highest. The trend is reversed for rainfall; the distribution of rainfall for poor quality wines is distinctly different from those of medium and good quality wines. The difference between poor and medium quality wines is less clear when it comes to some heat descriptors.

Furthermore, by looking at other statistical descriptors (mean and standard deviation), we observe that (i) the sum of daily average temperatures and (ii) the duration of sunshine have a minimal variance compared to rainfall and the number of hot days.

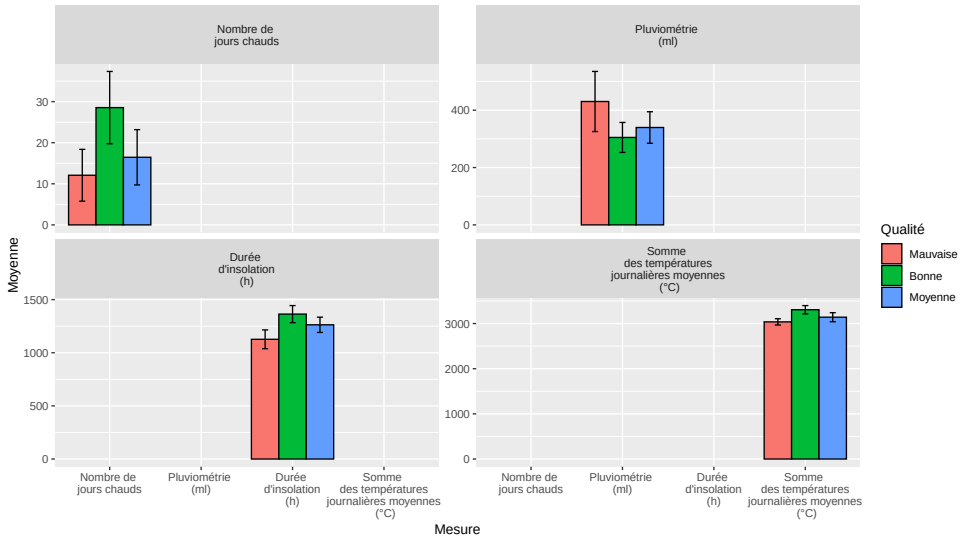


Figure 3: Means and standard deviations according to wine quality.

We can distinguish marked differences between the qualities based on empirical means; the average number of hot days allows differentiation between the group of good quality wines and the groups of medium and poor quality wines. On the other hand, the average annual rainfall allows differentiation between the group of poor quality wines and the groups of good and medium quality wines, but this measure does not allow differentiation between the latter.

2 Preliminary analyses of variance

Let's now look from an ANOVA modeling perspective which variables appear to be the most discriminative of wine quality when taken one at a time.

2.1 Checking the ANOVA assumptions

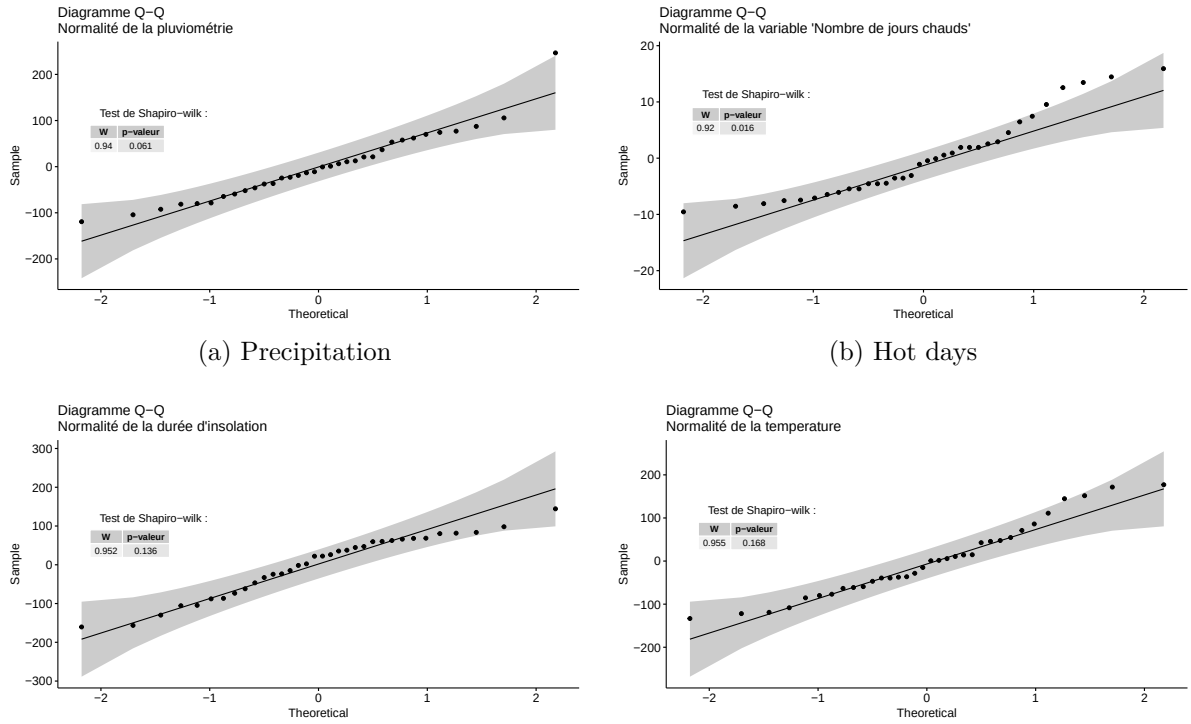
None of the descriptors contain extreme outliers. Outliers can have a significant impact on the quality of the ANOVA model, as it assumes a linear relationship between the dependent and independent variables. Extreme outliers can distort this relationship and result in a poor fit of the model.

It is common practice to check the assumptions of a model before proceeding with its fitting to avoid conducting erroneous analyses. ANOVA has assumptions about data distributions and homogeneity of variances that are rarely met in practice. Thus, the purpose of this verification is rather to determine whether the data available do not exhibit severe violation of the model assumptions.

The ANOVA model puts each quantitative variable in an equation as a function of a constant μ augmented by the group effect A_i to which a measurement error ϵ is added and is assumed to be Gaussian.

$$Y_{i,j} = \mu + A_i + \epsilon_{i,j} \quad , \quad \text{with} \quad \epsilon_{i,j} \sim \mathcal{N}(0, \sigma^2)$$

Normality of residuals To verify the normality assumption of the residuals, we conduct a visual diagnostic in addition to the Shapiro-Wilk test, which, under the null hypothesis, assumes normality of the residuals as presented in Figure 4.



(a) Precipitation

(b) Hot days

(c) Insolation

(d) Temperature

Figure 4: Visual investigation of the normality of residuals for each descriptor.

The critical values present an argument against rejecting the null hypothesis only for rainfall, temperature, and sunshine at a risk of 5%, that is there is not strong evidence that these samples are drawn from a non-normal distribution. However, the plots of all residuals do not show too strong deviations from the theoritical quantiles.

Homoscedasticity Another assumption of ANOVA is the equality of variances across groups. The Levene's test states under the null hypothesis that the variances across groups are equal.

$$H_0 : \quad \sigma_{Med} = \sigma_{Bad} = \sigma_{Good} \quad \forall x^j, j \in \{1, \ldots, p\},$$

where p is the number of variables.

At a risk level of 5%, the test associates critical probabilities (see Table 1) in favor of accepting the hypothesis of equality of variances across groups for all wine descriptors.

Table 1: Test de l'égalité des variances pour les différentes qualités de vin.

Variable	W	Prob. Critique
N Jours chauds	0.642	0.533
Pluviométrie	2.336	0.114
Insolation	0.079	0.924
Température	0.917	0.41

Thus, the wine descriptors do not show overall extreme violations of the ANOVA assumptions

2.2 Discriminatory power and analysis of variance.

We are now interested in identifying the discriminatory powers of the wine quality descriptors available. For this purpose, we use two criteria: the correlation coefficient R^2 of the ANOVA and the Wilks' Lambda Λ.

$$\Lambda = \frac{\det(A)}{\det(A+B)}$$

where A is the within-class dispersion matrix

$$A = \frac{1}{n}\sum_{k=1}^{q} n_k \Sigma_{X|_{Y=k}}$$

and B is the between-class disperion matrix

$$B = \frac{1}{n}\sum_{k=1}^{q} n_k (\overline{X|_{Y=k}} - \overline{X})^T (\overline{X|_{Y=k}} - \overline{X}),$$

with n is the number of observations and q the number of categories of variable Y (i.e the wine qualities).

In discriminant analysis, Wilks' Lambda is used as a measure of the discriminative power of a variable. It takes values in the range $[0, 1]$, where a value of 00 indicates perfect discrimination, a value of 1 indicates no discrimination, and intermediate values represent the intensity of the discriminatory power. The statistical significance of Λ in contributing to the discrimination of a variable is measured by an F-test.

Table 2: Analyse des variances et du pouvoir de discrimination des variables

Variables	R^2	Wilks' Lambda	Stat. F	P-value
Year of harvest	0.029	0.971	0.467	**0.631**
Temperature	**0.639**	**0.361**	27.389	10^{-7}
Insolation	**0.618**	**0.382**	25.061	3×10^{-7}
N of hot days	0.497	0.503	15.334	2×10^{-5}
Precipitation	0.353	0.647	8.44	0.001

The F-test statistic evaluates the ratio of between-class variance to the within-class variance, providing a measure of the dispersion between the centroids of each class (here, the classes being the wine qualities).

Thus, reading the F-test tells us that at a significance level of 5%, the most important differences between the centroids of the wine qualities are found in temperature, insolation, the number of hot days, and precipitation.

Combining Wilks' Lambda with the F-test indicates that

- Temperature and insolation are the most discriminatory.
- The number of hot days and rainfall have moderate discriminatory power.
- The year of harvest has no discriminatory power, and the difference between the centroids of each class is not statistically significant.

As we can see in Figure 5, the wine qualities do not suggest any temporal trend.

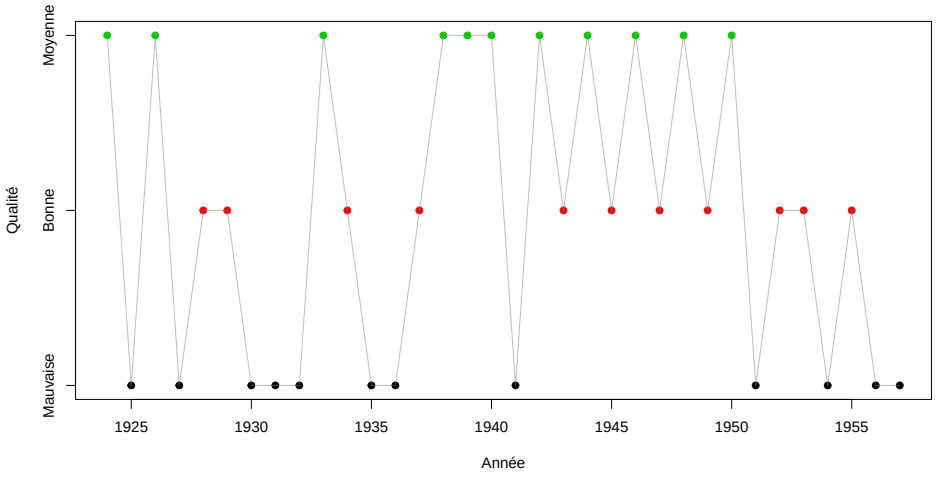


Figure 5: Wine quality over time.

3 Classical Discriminant Analysis

We now seek, from a linear combination of the original variables X, a new representation of the feature space that allows for better discrimination between the categories of the nominal variable Y, representing the three qualities of wines, by maximizing the following criterion.

$$\max_{f \in \langle X \rangle} \quad \cos^2_W(f, \langle Y \rangle). \tag{1}$$

The year of harvest variable, which has no discriminatory power, was excluded from the discriminant analysis. Since there are $q = 3$ classes, discrimination between the different wine qualities is carried out in a $q - 1 = 2$ dimensional space.

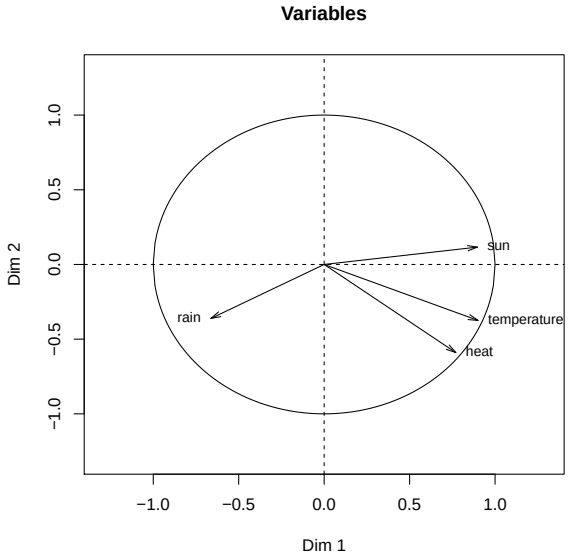


Figure 6: Correlation between the descriptors and the discriminatory axes.

The first and second factorial axes have discriminatory powers of $\eta_1 = 0.73$ and $\eta_2 = 0.12$ respectively, and generate a discriminant subspace with a power of

$$\eta_{E_{q-1}} = \frac{1}{q-1} \sum_{k=1}^{q-1} \eta_k = 0.42.$$

Since the second axis has weak discriminant power, it does not allow for differentiation between wine qualities. Thus, it will not be interpreted.

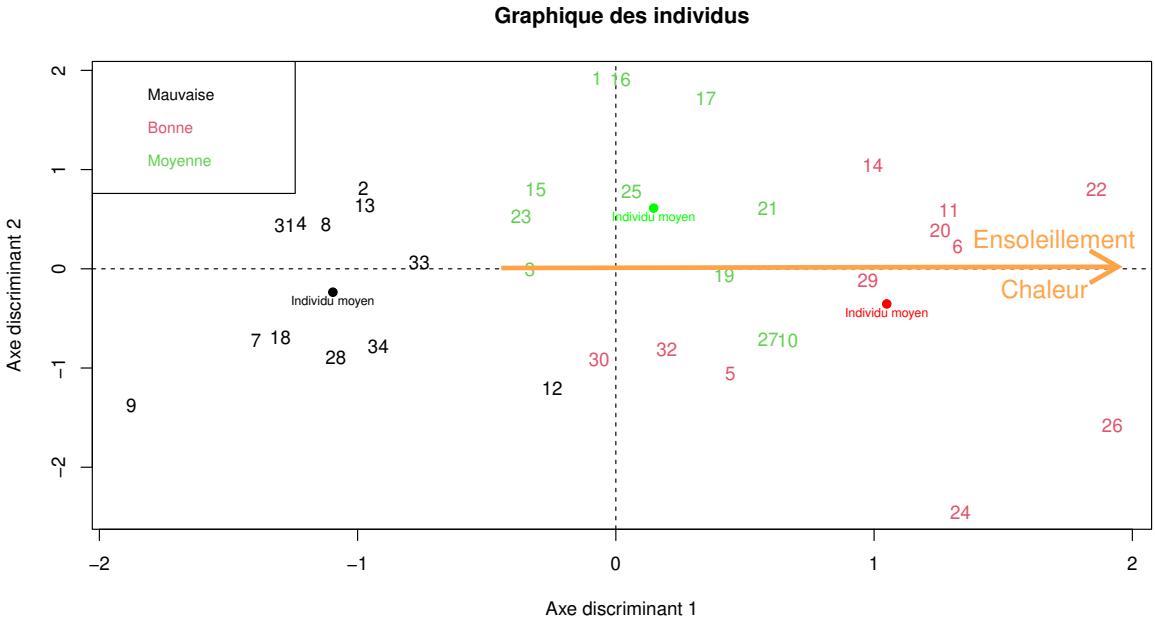


Figure 7: Projection of the observations onto the resulting factorial space.

3.1 Interpretation of the plots

Based on the individuals' graph in Figure 7, a clear distinction can be observed on the first discriminant axis between wines of poor quality and those of good and medium quality. However, the difference between wines of medium quality and those of good quality is less marked.

On the other hand, the variables graph in Figure 6 shows the correlations of the variables with the discriminant axes. The variables that are strongly correlated with the subspace formed by these axes have a good representation on it. Thus, the axes carry out the quality of the variables with which they are most strongly correlated ($\rho^2 > 0.7$).

The joint reading of the two discriminant analysis graphs indicates that wines of poor quality are mainly characterized by heat indicators, namely low insolation, low cumulative temperatures, and low numbers of hot days. Conversely, wines of good quality are associated with hugh heat indicators and low precipitation.

4 PLS Discriminant Analysis

In the above discriminant analysis, although the discriminative power of the first factorial axis seems quite good, not all variables are strongly correlated with it. Since the discriminant analysis does not take into account the structural relevance of the components (Equation 1), the adjustment can be made on noise, and as a result, some variables that could participate in discrimination end up being poorly represented in the discriminant subspaces. Partial Least Squares discriminant analysis (PLS-DA) can address this issue by maximizing the following criterion:

$$\max_{f=Xu\ ,\ u'u=1} \quad \cos^2_W(f, \langle Y \rangle)\ ||f||^2_W, \tag{2}$$

that takes into account the structural relevance of explanatory components $||f||^2_W$.

Maximizing the criterion in Equation 2 allows the components to be separated from the noise by bringing them closer to the original variables, enabling them to have a stronger structural relevance while they remain uncorrelated (orthogonal).

The PLS-DA model was fitted on two components using cross-validation with a leave-one-out (LOO) approach.

4.1 Interpretation of the plots

Table 3 reports the correlations between wine descriptors and discriminant components of PLS-DA, and those of classic discriminant analysis.

Table 3: Correlations **in absolute values** of the wine descriptors with classical DA and PLS-DA component.

	F1		F2	
Variables	AD-PLS	AD	AD-PLS	AD
Temperature	0.91	0.9	0.28 ↙	0.36
Insolation	0.87	0.89	0.18	0.13
Hot days	**0.89** ↖	0.77	**0.39** ↙	0.58
Precipitation	0.65	0.66	**0.62** ↖	0.37

In PLS-DA, heat indicators are more strongly captured by the first component, while the second component moves away from them and get closer to precipitation. The variables have a better representation in the factorial subspace $\langle 1, 2 \rangle$. We report them on the individuals' plot.

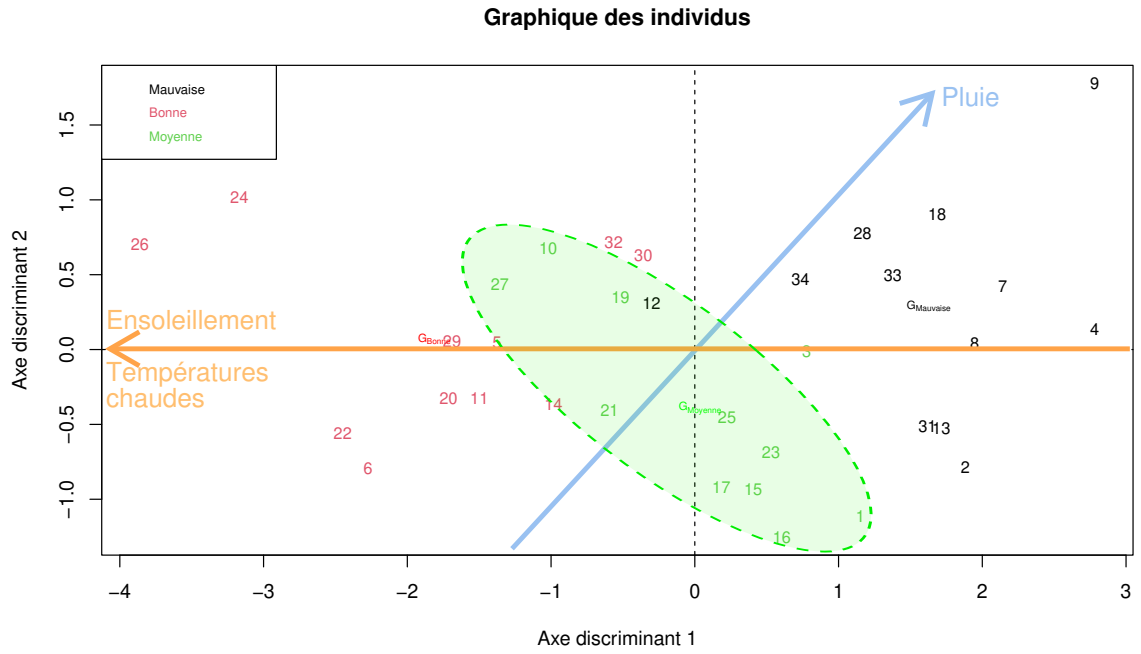


Figure 8: Bi-plot of the PLS-DA.

PLS-DA allows for a clearer partitioning between wine qualities by containing the dispersion of wines of medium quality in a tighter zone at the origin of the factorial plane (see ellipse in Figure 8), whereas classical discriminant analysis attributed them with greater dispersion. This allows for better discrimination, especially between wines of poor quality and those of good quality. Once again, low heat indicators combined with high precipitation characterize wines of poor quality, while it is the opposite for wines of good quality. Thus, the years in which these descriptors are average may result in wines of medium quality.

4.2 Model evaluation

The PLS-DA model was fitted on two components using leave-one-out (LOO) cross-validation. The overall error rate is 20%.

The confusion matrix is reported below, and the performance of the model is evaluated using the following classification metrics:

$$\text{Sensitivity (Recall)} = \frac{TP}{TP+FN} \quad , \qquad \text{Specificity} = \frac{TN}{TN+FP}$$
$$\text{Pos. Pred. Value (Precision)} = \frac{TP}{TP+FP} \quad , \qquad \text{Neg. Pred. Value} = \frac{TN}{TN+FN}.$$

		Truth		
		Bad	Good	Medium
Pred.	Bad	10	0	2
	Good	1	11	3
	Medium	1	0	6
	Total	12	11	11

Table 4: Confusion matrix of the PLS-DA model.

	Quality		
Metrics	Bad	Good	Medium
Sensitivity (Recall)	0.83	**1**	0.545
Specificity	**0.91**	0.82	**0.95**
PPV (Precision)	**0.83**	0.73	0.85
NPV	**0.9**	**1**	0.81

Table 5: Classification metrics of the PLS-DA model.

Interpretation of the metrics The results of the classification reveal that the model has difficulties recognizing wines of medium quality, which results in a low sensitivity for this wine quality $\sim 54\%$). The learned decision rule of the model thus mistakes wines of medium quality for those of good and bad quality. On the other hand, when the model is presented with wines of a quality other than medium (i.e. good or bad), it only attributes medium quality to them in

$$1 - \text{Specificity} = 1 - 95\% = 5\% \quad \text{of the cases.}$$

And while the model recognizes wines of good quality perfectly when they are presented to it (Sensitivity of 100%), it happens that it incorrectly labels wines as good quality in around

$$1 - \text{Précision} = 1 - 73\% = 27\% \quad \text{of the cases,}$$

when they are in fact of average or poor quality.

However, upon examining the confusion matrix, it can be observed that this confusion occurs more often with wines of medium quality than with wines of poor quality.

Furthermore, out of all non-bad quality wines, the model recognizes them as non-bad quality in 91% of cases (Specificity) with a NPV of 90% which indicates that the confusion happens more between good quality wines and those of medium quality.

When the model is presented with wines of bad quality, it recognizes them in 83% of the cases, which is a decent score.

Concluding, the wines of medium quality represent a gray area for which the model can mistake them for either good or bad quality, but it is rare that the confusion occurs between wines of good quality and those of bad quality.

4.3 Discussion and contextualization

The strength of this model is that when faced with wines of good quality, the sensitivity and NPV are optimal which provide a rate of false negatives of zero ($|FN| = 0$) for this quality of wine while wines of poor quality are attributed to wines of good quality only rarely.

This indicates that when the heat indicators for the year are average overall, special attention should be paid to the production of that year in order to properly distinguish the good and bad quality wines.
The opposite is also true, i.e in years with high temperatures or heavy rainfall, the harvest tends to respectively be rather good or rather bad.

However, the scores should be interpreted according to the intended use of the model, namely as a fully autonomous annotation tool or as an annotation aid tool.

The Precision and Recall scores could be effectively used and oriented according to the stakes in which the user's needs for the model reside.
The Precision score, for example, informs us about the relevance of the model in not making mistakes when predicting a particular quality, while the Recall score informs us about the model's ability to recognize a certain quality when it is presented.

These are very different concepts, and although we would like to have a model that performs well in both, they only make sense when the stakes related to the classification are taken into account and contextualized within the scope of identified needs.

One metric alone is far from sufficient in many cases. For example, a high precision score and a very low recall score would mean that, despite the model's tendency not to make mistakes when predicting a certain quality, its overall success rate remains low due to its inability to recognize the quality presented to it in the majority of cases.

Thus, in our opinion, the complementarity of classification metrics is striking, and valuing one over the other must be carefully considered while taking into account the context in which they are to be used, as well as their potential consequences.