



UNIVERSITÉ DE MONTPELLIER  
FACULTÉ DES SCIENCES  
M2 STATISTIQUES ET SCIENCES DES DONNÉES

---

# L'apport des conditions climatiques à la qualité des vins

---

Analyse et modélisation multivariée

-

Méthodes à deux tableaux

*Auteurs:*  
Anas ZAKROUM

*Enseigné par:*  
PR. Xavier BRY

27 février 2023

# Table des matières

<b>1</b>	<b>Description des données</b>	<b>2</b>
<b>2</b>	<b>Quelques analyses de variance préliminaires</b>	<b>4</b>
2.1	Vérification des hypothèses . . . . .	4
2.2	Pouvoir discriminatif et analyse de la variance . . . . .	6
<b>3</b>	<b>Analyse discriminante classique</b>	<b>7</b>
3.1	Interpretation . . . . .	9
<b>4</b>	<b>Analyse discriminante PLS</b>	<b>9</b>
4.1	Interpretation des graphiques . . . . .	9
4.2	Évaluation du modèle . . . . .	10
4.3	Discussion et mise en situation . . . . .	11

# Introduction

Nous disposons de 34 observations de vins de bordeaux présentant trois niveaux de qualités : bonne, moyenne et mauvaise décrits par :

- L'année de récolte
- Somme des températures moyennes quotidiennes sur l'année ( $C^r$ )
- Durée d'insolation (h)
- Nombre de jours très chauds
- Pluviométrie (mm)

L'objectif de cette analyse est d'extraire à partir des données récoltées, les caractéristiques permettant de discriminer entre les trois niveaux de qualité des vins.

## 1 Description des données

### Evolution temporelle des descripteurs

Les données sur les vins ont été récoltées entre 1924 et 1957. L'allure des séries temporelles des descripteurs dans la Figure 1 est non stationnaire et ne semble pas présenter visuellement de tendances linéaires ou quadratiques.

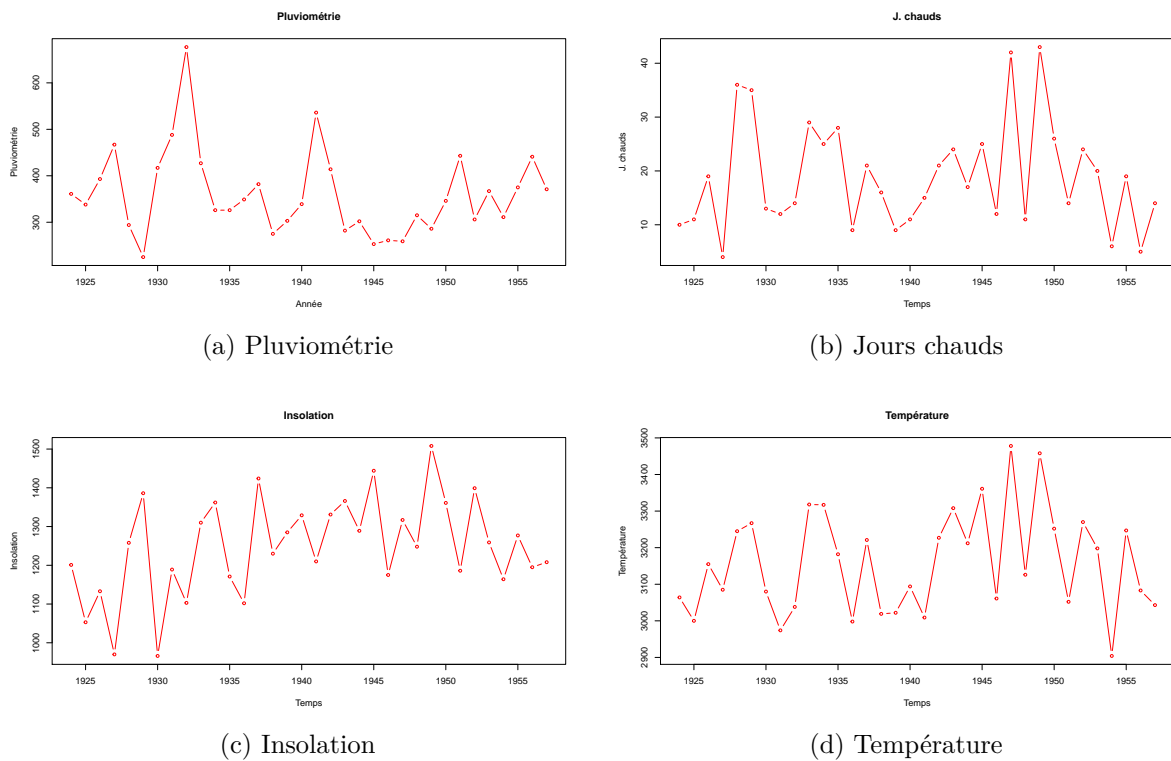


FIGURE 1 – Evolutions temporelles des descripteurs.

## Statistiques descriptives des qualités des vins

Les 34 vins sont répartis selon leur qualité de la manière suivante : 12 de qualité mauvaise, 11 de qualité moyenne et 11 de bonne qualité. Nous proposons dans un premier temps de fournir une description des distributions des variables mesurées selon les qualités de vins.

A premier abord, nous observons à partir de la Figure 2 que la majorité des distributions des descripteurs possèdent des allures approximativement gaussiennes. Les médianes des descripteurs sont nettement différenciées entre les trois qualités de vins. Nous ne pouvons pas dire la même chose sur le reste des quartiles.

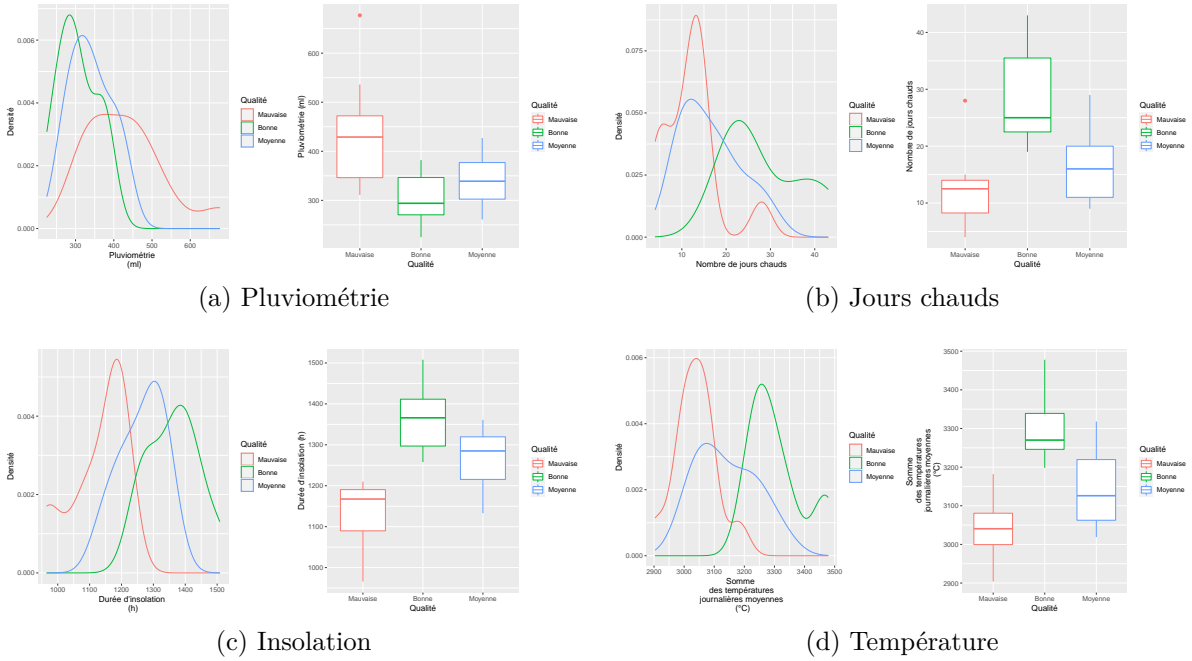


FIGURE 2 – Distributions des mesures prises des vins

Il semble toutefois que les vins de bonne qualité ressortent pendant les années chaudes, années où les descripteurs de chaleur (Insolation, Température, J. chauds) sont au plus haut. La tendance est inversée pour la pluviométrie; la distribution de la pluviométrie pour les vins de mauvaise qualité se différencie nettement de celles des vins de qualité moyenne et mauvaise. La différence entre les vins de qualité mauvaise et moyenne est moins nette lorsqu'il s'agit de certains descripteurs de chaleur.

D'autre part, en regardant d'autres descripteurs statistiques (la moyenne et l'écart-type), nous observons que (i) la somme des températures quotidiennes moyennes et (ii) la durée d'insolation sont de variance minimale comparées à la pluviométrie et le nombre de jours chauds.

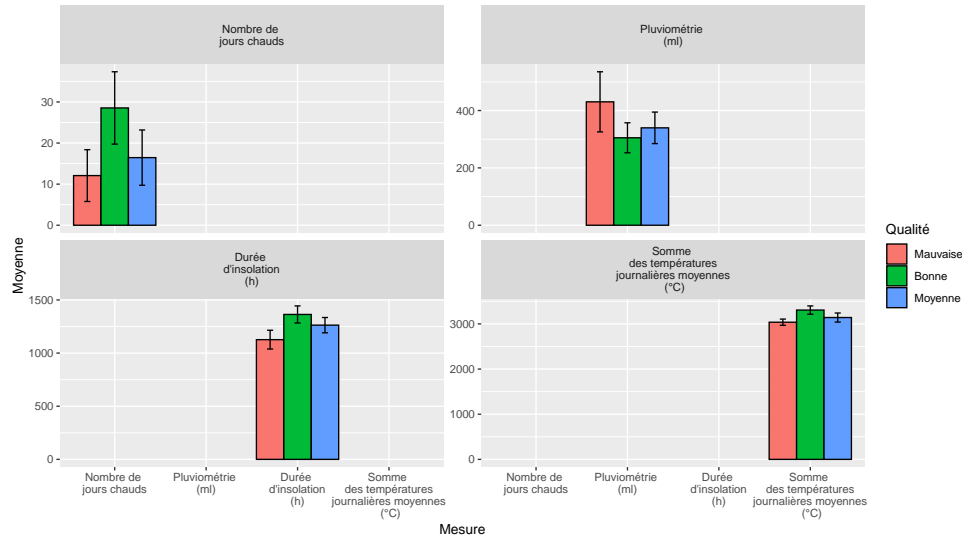


FIGURE 3 – Moyennes et écarts-types des descripteurs par qualité

Nous pouvons distinguer des différences marquées entre les qualités sur les moyennes empiriques ; le nombre de jours chauds moyen permet la distinction entre le groupe de vin de bonne qualité et les groupes de vin de moyenne et mauvaise qualité. D'autre part, la pluviométrie annuelle moyenne permet de différencier entre le groupe des vins de mauvaise qualité et les groupes de vin de bonne et moyenne qualité, mais cette mesure ne permet pas la différenciation entre ces dernières.

## 2 Quelques analyses de variance préliminaires

Regardons maintenant d'un point de vue de la modélisation d'ANOVA quelles sont les variables apparaissant être les plus discriminatives de la qualité de vin lorsqu'elles sont prises une à une.

### 2.1 Vérification des hypothèses

Premièrement, l'ensemble des descripteurs ne comportent pas de valeur atypique extrême. Les valeurs atypiques extrême sont connues pour impacter la qualité de l'ajustement de l'ANOVA qui, étant un modèle linéaire est sensible aux valeurs extrêmes.

**Hypothèse de normalité des descripteurs** Il est de pratique courante de vérifier les hypothèses d'un modèle avant de procéder à son ajustement de risque à conduire des analyses erronées. L'ANOVA possède des hypothèses sur les distributions des données et sur l'homogénéité des variances qui sont en pratique atteintes de manière peu fréquente. Ainsi le but de cette vérification est plutôt de chercher si les données à disposition ne présentent pas une violation sévère des hypothèses du modèle. Le modèle met en équation

chaque variable quantitative en fonction d'une valeur  $\mu$  augmentée de l'effet de groupe  $A_i$  plus une erreur de mesure  $\epsilon$  supposée gaussienne.

Pour vérifier l'hypothèse de normalité des résidus, nous procédons à un diagnostic visuel en plus du test de Shapiro-Wilk qui, sous l'hypothèse nulle suppose la normalité des résidus comme présenté dans la Figure 4.

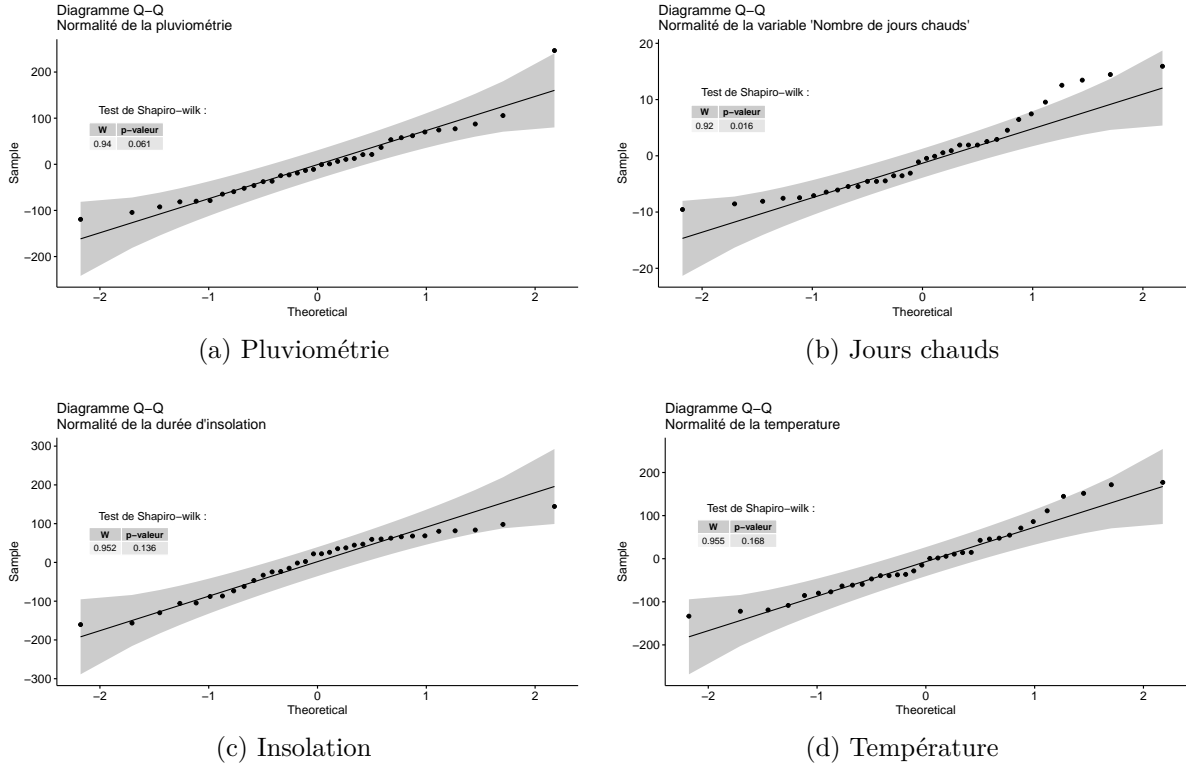


FIGURE 4 – Diagnostic de la normalité des résidus des modèles d'ANOVA pour chaque descripteur.

Les valeurs critiques présentent un argument en défaveur du rejet de l'hypothèse nulle seulement pour la pluviométrie, la température et l'insolation au risque de 5%. Toutefois, les graphiques de l'ensemble des résidus des descripteurs ne montrent pas de trop fortes déviations.

**Homoscédasticité** Une autre hypothèse de l'ANOVA est l'égalité des variances à travers les groupes. Le test de Levene stipule sous l'hypothèse nulle que les variances à travers les groupes sont égales.

$$H_0 : \sigma_{Moy} = \sigma_{Mauv} = \sigma_{Bon} \quad \forall x^j, j \in \{1, \dots, p\}$$

Au risque de 5%, le test associe des probabilités critiques (Voir Table 1) en faveur de l'acceptation de l'hypothèse d'égalité des variances à travers les groupes, et ce pour tous les descripteurs de vins

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

Ainsi les descripteurs de vins ne présentent pas globalement des violations extrêmes des hypothèses de l'ANOVA.

## 2.2 Pouvoir discriminatif et analyse de la variance

Nous nous intéressons maintenant à connaître les pouvoirs discriminatifs de la qualité de vin des descripteurs à disposition. Nous utilisons pour cela deux critères, le coefficient de corrélation  $R^2$  de l'ANOVA et le Lambda de Wilks  $\Lambda$ .

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

où A la matrice de dispersion intra-classe

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

et B est la matrice de dispersion inter-classe

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

avec  $n$  le nombre d'individus et  $q$  le nombre de modalités de la variable  $Y$ .

En analyse discriminante le Lambda de Wilks est utilisé comme une mesure de la capacité discriminative d'une variable. Il est à valeur dans  $[0, 1]$  où la valeur 0 indique une discrimination totale et 1 indique aucune discrimination et une valeur intermédiaire représente l'intensité du pouvoir discriminatif. La significativité de  $\Lambda$  dans l'apport à la discrimination d'une variable est mesurée par le test F.

TABLE 2 – Analyse des variances et du pouvoir de discrimination des variables

Variables	$R^2$	Wilk's Lambda	Stat. F	P-valeur
Année	0.029	0.971	0.467	<b>0.631</b>
Température	<b>0.639</b>	<b>0.361</b>	27.389	$10^{-7}$
Insolation	<b>0.618</b>	<b>0.382</b>	25.061	$3 \times 10^{-7}$
N jours chaud	0.497	0.503	15.334	$2 \times 10^{-5}$
Pluviométrie	0.353	0.647	8.44	0.001

La statistique de test F évalue le rapport de l'inertie inter-classe avec l'inertie intra classe induisant une mesure de la dispersion entre les barycentres de chaque classe (ici les classes étant les qualités de vin).

Ainsi la lecture du test F nous dit qu'au seuil de 5% les différences entre les barycentres des qualités de vin les plus importantes se trouvent au niveau de la température, l'insolation, le nombre de jours chauds et la pluviométrie. La lecture conjointe du Lambda de Wilks et du test F nous indique que

- La température et l'insolation sont les plus discriminantes.
- Le nombre de jours chauds et la pluviométrie sont à pouvoir discriminatif moyen.
- L'année des vendanges possède un pouvoir discriminatif nul et la différence entre les barycentres de chaque classe n'est pas statistiquement significative.

Comme nous pouvons le voir dans la Figure 5 les qualités de vin ne suggèrent pas d'évolution temporelle.

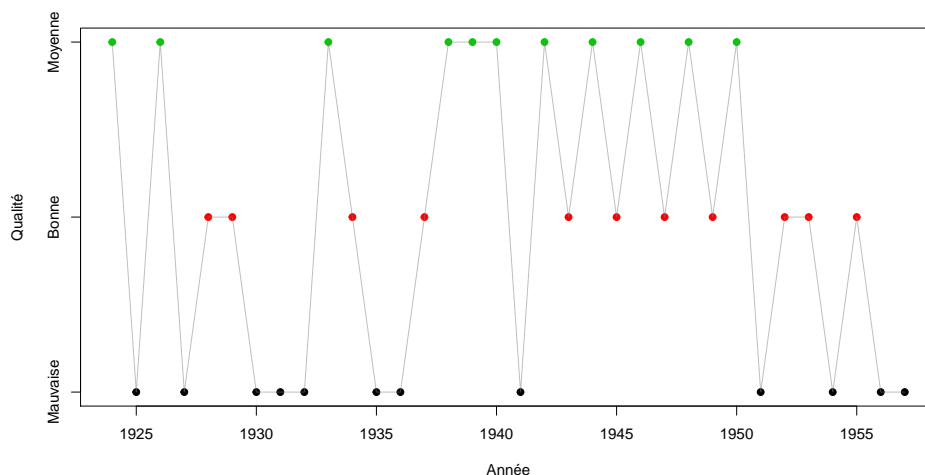


FIGURE 5 – Qualité des vins par année

### 3 Analyse discriminante classique

Nous cherchons maintenant, à partir d'une combinaison linéaire des variables d'origine  $X$  une nouvelle représentation de l'espace des variables qui permette une meilleure discrimination entre les modalités de la variable nominale  $Y$  représentant les trois qualités des vins en maximisant le critère suivant

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

L'année de récolte des raisins, possédant un pouvoir discriminatif nul a été omise de l'analyse discriminante. Le nombre de classes étant au nombre de  $q = 3$ , la discrimination entre les différentes qualités des vins se fait sur un espace à  $q - 1 = 2$  dimensions.



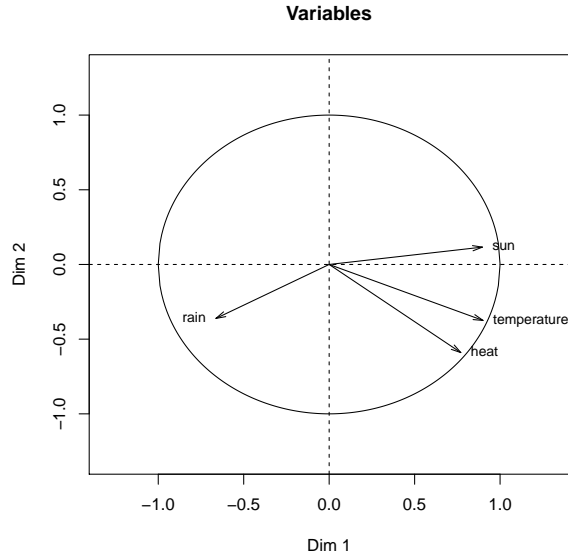


FIGURE 6 – Corrélation des variables avec les axes discriminants.

Les premier et deuxième axes factoriels possèdent respectivement les pouvoirs discriminatifs  $\eta_1 = 0.73$  et  $\eta_2 = 0.12$  et engendrent un sous-espace discriminatif de pouvoir

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

. Le deuxième axe étant de pouvoir discriminatif faible, et donc ne permettant pas la différenciation entre les qualités des vins ne sera pas interprété.

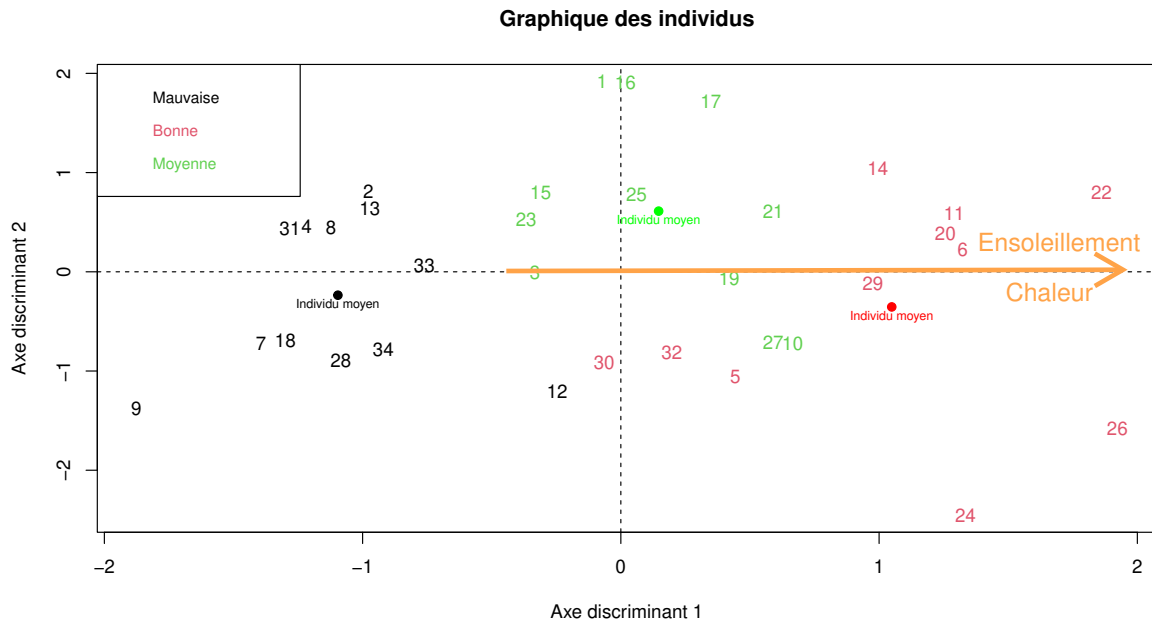


FIGURE 7 – Graphique des individus sur le sous-espace discriminant  $\langle 1, 2 \rangle$

### 3.1 Interprétation

Au regard du graphique des individus dans la Figure 7, la distinction sur le premier axe discriminant est marquée entre les vins de mauvaise qualité et ceux de bonne et de moyenne qualité. En revanche, la différence entre les vins de moyenne qualité et ceux de bonne qualité est moins marquée.

D'autre part, le graphique des variables dans la Figure 6 indique les corrélations des variables avec les axes discriminants. Les variables les plus fortement corrélées avec le sous-espace formé par ces axes possèdent donc une bonne représentation sur celui-ci. Ainsi, les axes portent en eux la qualité des variables avec lesquelles ils sont le plus fortement corrélés ( $\rho^2 > 0.7$ ).

La lecture jointe des deux graphiques de l'analyse discriminante nous indique alors que les vins de mauvaise qualité se caractérisent surtout à l'aide des indicateurs de chaleur à savoir une insolation, des températures cumulées et un nombre de jours chauds faibles. A l'inverse les vins de bonne qualité sont associés à des indicateurs de chaleur marqués et une pluviométrie faible.

## 4 Analyse discriminante PLS

Dans l'analyse discriminante ci-dessus, bien que le pouvoir discriminatif du premier axe factoriel semble assez correct, pas toutes les variables  $y$  sont fortement corrélées. L'analyse discriminante ne prenant pas en compte pas la force structurelle des composantes (Equation 1), l'ajustement peut se faire sur du bruit et par conséquent certaines variables pouvant participer à la discrimination se retrouvent mal représentées sur les sous-espaces de discrimination.

L'analyse discriminante PLS permet de remédier à ce problème en maximisant le critère suivant

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

qui prend en compte la force structurelle des composantes  $\|f\|_W^2$ .

La maximisation du critère dans l'Equation 2 permet d'éloigner les composantes du bruit en les rapprochant des variables originelles leur permettant d'avoir une force structurelle plus forte.

Le modèle de l'AD-PLS a été ajusté sur deux composantes par validation croisée sur une observation tournante (LOO).

### 4.1 Interprétation des graphiques

La Table 3 reporte les corrélations des descripteurs des vins les composantes de l'AD-PLS et celles de l'AD classique.

TABLE 3 – Corrélations **en valeur absolue** des composantes AD-PLS et AD classique avec les variables.

Variables	F1		F2	
	AD-PLS	AD	AD-PLS	AD
Température	0.91	0.9	0.28	✓ 0.36
Insolation	0.87	0.89	0.18	0.13
J. Chauds	<b>0.89</b>	↖ 0.77	<b>0.39</b>	✓ 0.58
Pluviométrie	0.65	0.66	<b>0.62</b>	↖ 0.37

Dans l'AD-PLS, les indicateurs de chaleur sont davantage capturées par la première composante, tandis que la deuxième composante s'en éloigne et se rapproche de la pluviométrie. Les variables possèdent une meilleure représentation dans le sous-espace factoriel  $\langle 1, 2 \rangle$ . Elles sont alors reportées sur le graphique des individus.

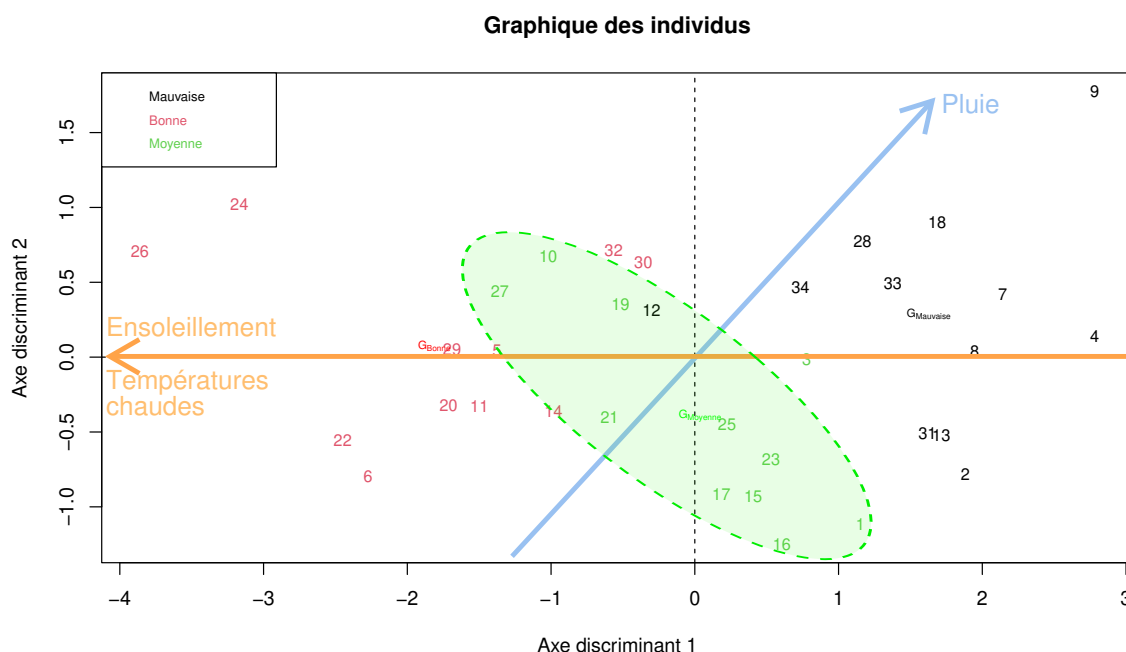


FIGURE 8 – Graphique des individus sur le plan factoriel  $\langle 1, 2 \rangle$

L'AD-PLS permet un partitionnement plus clair entre les qualités des vins en contenant la dispersion des vins de moyenne qualité sur une zone plus serrée à l'origine du plan factoriel (voir ellipse dans la Figure 8), alors que l'AD leur attribuait une plus grande dispersion. Cela permet une meilleure discrimination surtout entre les vins de mauvaise qualité et ceux de bonne qualité. Encore une fois, le faible ensoleillement couplé à une pluviométrie forte caractérisent les vins de mauvaise qualité et l'inverse, les vins de bonne qualité. Les années où ces descripteurs sont moyens résultent en des vins de qualité moyenne.

## 4.2 Évaluation du modèle

Le modèle de l'AD-PLS a été ajusté sur deux composantes par validation croisée sur une observation tournante (LOO). Le taux d'erreur global est de 20%.

La matrice de confusion est reportée ci-dessous et la performance du modèle est évaluée à l'aide des métriques de classification suivantes

$$\begin{aligned} \text{Sensibilité} &= \frac{TP}{TP + FN} \quad , \quad \text{Spécificité} = \frac{TN}{TN + FP} \\ \text{Val. Pred. Pos.} &= \frac{TP}{TP + FP} \quad , \quad \text{Val. Pred. Nég.} = \frac{TN}{TN + FN} . \end{aligned}$$

		Vérité		
		Mauvaise	Bonne	Moyenne
Préd.	Mauvaise	10	0	2
	Bonne	1	11	3
	Moyenne	1	0	6
	Total	12	11	11

TABLE 4 – Matrice de confusion du modèle de l’AD-PLS

		Qualité		
Métriques		Mauvaise	Bonne	Moyenne
Sensibilité (Recall)		0.83	<b>1</b>	0.545
Spécificité		<b>0.91</b>	0.82	<b>0.95</b>
VPP (Precision)		<b>0.83</b>	0.73	0.85
VPN		<b>0.9</b>	<b>1</b>	0.81

TABLE 5 – Métriques de classification du modèle de l’AD-PLS

**Interprétation des métriques** La lecture des résultats de la classification relève que le modèle présente des difficultés à correctement reconnaître les vins de moyenne qualité (sensibilité faible pour cette classe). De plus quand le modèle reconnaît très bien les vins de qualité non-moyenne quand ils lui sont présentés (spécificité forte pour cette classe  $\sim 0.95$ ).

De tous les vins de qualité non mauvaise, le modèle les reconnaît dans 91% des cas avec des VPP de 85% et 0.83% respectivement pour les vins de moyenne et de mauvaise qualité.

Toutefois, quand le modèle classe un vin comme de bonne qualité, il se trompe dans  $1 - 73\% = 27\%$  des cas, avec une VPN de 90% pour les vins de mauvaise qualité i.e le modèle a tendance dans ses prédiction à plus prendre des vins de moyenne qualité pour des vins de bonne qualité que de prendre des vins de mauvaise qualité pour des vins de bonnes qualité. Ce qui présente à notre sens un coût moins grave.

### 4.3 Discussion et mise en situation

Le fort de ce modèle est que face à des vins de bonne qualité, le taux de faux négatifs est nul, ce qui lui procure une sensibilité et une VPN optimale pour cette qualité de vin et que les vins de mauvaise qualité ne sont pas attribués aux vins de bonne qualité.

Ceci indique que quand les indicateurs de chaleur pour l’année sont moyen dans la globalité, une attention particulière doit être portée pour la production de cette année là pour pouvoir correctement faire ressortir les vins de bonne et de mauvaise qualité. L’inverse est valable aussi, i.e les années de fortes chaleur ou de forte pluie, la récolte aura tendance à être plutôt bonne ou plutôt mauvaise respectivement.