Évaluation de l'influence du vécu hydrique de la vigne sur la qualité des raisins

Cheikh Moustapha Diakhate

Université de Strasbourg

UFR de Mathématiques et Informatique

MASTER 2 MATHEMATIQUES ET APPLICATIONS
PARCOURS BIOSTATISTIQUE ET STATISTIQUES INDUSTRIELLES

RAPPORT DE STAGE

Evaluation de l'influence du vécu hydrique de la vigne sur la qualité du raisin à la vendange

effectué à

L'Institut National de la Recherche Agronomique

Unité Mixte de Recherche de Mathématiques, Informatique et STatistique pour l'Environnement et l'Agronomie

du 03 Février au 04 Août 2014

par

Cheikh Moustapha DIAKHATE

Maitres de stage de l'entreprise
Nadine HILGERT et Brigitte CHARNOMORDIC

Directeur de stage de l'Université
Nicolas POULIN

Date de soutenance : 28 Août 2014

Remerciements

Je commence par remercier M. Pascal Neveu, directeur de l'UMR MISTEA pour avoir accepté de m'accueillir avec bienveillance au sein de sa structure.

Je souhaite vivement remercier mes encadrantes Nadine HILGERT et Brigitte CHARNOMORDIC pour leur disponibilité, leur patience et toute l'aide continue qu'elles m'ont apportée tout au long de mon stage. Je les remercie aussi pour leurs critiques, leurs corrections et leurs suggestions, ce qui m'a permis de toujours aller au delà de mes possibilités et de redoubler d'efforts.

Je remercie également M. Nicolas POULIN, enseignant et encadrant de ce stage à l'Université de Strasbourg, pour sa disponibilité, ses conseils et directives complémentaires.

Je remercie l'ensemble des chercheurs de l'UMR MISTEA, avec qui les partages et les réflexions pendant la traditionnelle "pause-café" ont toujours été enrichissantes.

Je remercie également les partenaires du projet notamment Thibault SCHOLASCH de Fruition Sciences, Aurélie VIALARET de Nyseos et Nicolas SAURIN de PechRouge, avec qui j'ai collaboré tout au long de mon stage et qui n'ont ménagé aucun effort pour le bon déroulement de mon travail.

De manière plus globale, je tiens à remercier tous ceux que j'ai eu à côtoyer à l'UMR MISTEA pour leur accueil et leurs qualités humaines qui m'ont permis d'effectuer ce stage dans un cadre totalement détendu notamment Véronique Sals-Vettorel, l'assistante de direction ddu laboratoire, Malika NASSIF, Nicolas SUTTON et Alexandre MAIRIN.

Je tiens à remercier également mes parents et tous les membres de ma famille en particulier ma grande soeur Fanta DIAKHATE, et tous ceux qui de près ou de loin ont constitué un soutien moral et financier pour une réussite totale dans mes études.

Table des matières

Préambule

Dans le cadre de la deuxième année du Master Mathématiques et Applications, Parcours Biostatistique et Statistiques industrielles de l'Université de Strasbourg, j'ai effectué un stage de six mois au sein de l'INRA (Institut National de Recherche Agronomique). J'ai été accueilli dans l'UMR MISTEA (Mathématiques, Informatique et STatistique pour l'Environnement et l'Agronomie). L'objectif de mon stage a été d'évaluer l'impact du vécu hydrique de la vigne pendant la saison sur la qualité des raisins à la vendange.

Chapitre 1

Introduction

1.1 Présentation de l'INRA

L'institut National de la Recherche en Agronomique (INRA) est un organisme public français de recherche en agronomie qui a été fondé en 1946 et placé sous le statut d'Etablissement Public à caractère Scientifique et Technique (EPST). L'INRA est sous la double tutelle du ministère chargé de la recherche et celui de l'Agriculture et de la pêche. Classé premier institut de recherche agronomique en Europe et deuxième dans le monde, les recherches menées par l'INRA portent sur 3 principaux domaines : l'alimentation, l'agriculture et l'environnement.

L'institut se compose de près de 500 unités couvrant l'ensemble du territoire :

- 14 départements de recherche ;
- 19 centres régionaux ;
- 213 unités de recherche dont 112 unités mixtes de recherche ;
- 49 unités expérimentales.

Les effectifs dénombrent près de 8 500 agents titulaires dont environ 1828 chercheurs, 2427 ingénieurs, 4249 techniciens et administratifs répartis dans 18 centres régionaux et 13 départements scientifiques. Et près de 3000 stagiaires, doctorants et autres chercheurs étrangers qui sont également accueillis chaque année.

Le cap scientifique de l'institut est défini sur quatre chantiers prioritaires :

- Améliorer les performances économiques, sociales et environnementales de l'agriculture ;
- Assurer des systèmes alimentaires sains et durables ;
- Valoriser la biomasse ;
- Atténuer le réchauffement climatique et s'y adapter.

1.2 Structure et campus de Montpellier

Parmi les 18 centres régionaux, celui de Montpellier est composé de 8 implantations dans le Languedoc-Roussillon et accueille 1170 agents dont 720 titulaires et 450 non titulaires.

Le campus de la Gaillarde est composé de 20 unités mixtes de recherche avec comme partenaires principaux : Montpellier SupAgro, le Cirad, le CNRS et l'IRD.

Les principaux axes scientifiques du centre de Montpellier sont :

– Biologie intégrative et biologie du développement ;
– Agro-environnement pour un développement durable ;
– Biodiversité ;
– Sciences et technologies intégrées des produits alimentaires ou non ;
– Société, économie et décision.

1.3 La structure d'accueil : l'UMR MISTEA

1.3.1 Présentation générale

J'ai effectué mon stage dans l'une des 20 Unités Mixtes de Recherche de Montpellier Supagro qu'est l'UMR MISTEA (Mathématiques, Informatique et STatistique pour l'Environnement et l'Agronomie) au sein de l'équipe GAMMA (Gestion, Analyse et Modèles pour les Masses de données en Agronomie). Les travaux de recherche menés par ce laboratoire, relèvent principalement de l'analyse et du contrôle de systèmes dynamiques d'intérêt biologique, agronomique, agro-alimentaire ou environnemental. Ils sont organisés en trois grandes activités, associant chacune recherche et valorisation :

– Mathématiques de la gestion des ressources renouvelables (forêt, eau) ;
– Mathématiques de la conduite des procédés des agro-eco-systèmes ;
– Représentation et gestion de connaissances pour les gestion des agro-eco-systèmes.

Ces activités de recherche portent sur le développement d'outils mathématiques et informatiques utiles pour l'observation, l'analyse, la modélisation et le contrôle des différents types de systèmes dynamiques rencontrés dans ce cadre.

1.3.2 Structure de l'unité

L'UMR MISTEA est structurée en deux équipes réparties sur deux bâtiments : MODEMIC (Modélisation et Optimisation des Dynamiques des Écosystèmes MICrobiens) et GAMMA (Gestion, Analyse et Modèles pour les Masses de données en Agronomie). Le laboratoire est actuellement formé de 19 personnes permanents combinant les disciplines des mathématiques appliquées (théorie du contrôle et automatique, statistique, analyse numérique, recherche opérationnelle) et de l'informatique (raisonnement qualitatif, gestion de données) et pouvant apporter des regards complémentaires sur le même problème de recherche.

1.4 Contexte du stage : Le projet PILOTYPE

Depuis quelques années, la crise touche la filière viticole et la concurrence s'étoffe avec des vins provenant des pays du nouveau monde. La France qui était il y a 10 ans le 1^{er} exportateur en volume se retrouve aujourd'hui à la 3^{e} place derrière l'Italie et l'Afrique du Sud.

A ce recul conjoncturel, s'ajoute l'évolution des modes de consommation. Les consommateurs préfèrent aujourd'hui les vins bénéficiant d'une bonne image, garantissant un goût homogène, régulier et affichant un prix correct. La segmentation française, très attachée à la provenance territoriale des vins et peu à la notion de marque, n'est pas très lisible à l'export et ne répond que partiellement aux attentes des consommateurs.

C'est dans ce contexte qu'a été initié le projet PILOTYPE en 2010 pour fournir à la filière viticole, régionale et nationale, des outils d'aide à la décision lui permettant de piloter la qualité du vin et ainsi améliorer sa compétitivité à l'export et de permettre son maintien de façon durable. La réflexion qui sous-tend la démarche n'est plus basée sur la qualité des vins mais sur leur valeur, c'est-à-dire le rapport entre leur qualité et leur coût de production. Pour cela, le comité du projet veut mettre en place un outil industriel pour suivre la valeur du vin des premières opérations en vigne, à la vente de la bouteille sur le marché français et international. L'objectif est de doter la filière viticole de technologies innovantes intégrées, de la vigne à la cave jusqu'à la mise en bouteille et la commercialisation.

Le projet PILOTYPE est porté par un consortium regroupant des acteurs majeurs à tous les niveaux majeurs de la filière viticole :

- Les Grands Chais de France (CGF), 1^{er} exportateur de vin français dans le monde ;

- NYSEOS, société de conseil et d'analyse d'arômes, porteur du projet et FRUITION SCIENCES qui utilisent des capteurs de flux de sève pour la gestion de l'état hydrique de la vigne ;

- Des structures de production : la cave coopérative d'Ouveillan et les Celliers Jean d'Alibert ;

- Des organismes d'expérimentation et de recherche : l'Institut Français de la Vigne et du Vin (IFV) en sous traitance, l'Institut National de Recherche Agronomique (INRA) et Montpellier Supagro (en sous-traitant de l'INRA).

1.5 Objectif du stage

Au cours du projet PILOTYPE, deux modules d'Outils d'Aide à la Décision (OAD) devront être développés autour d'un système d'information. Un module "plante" qui, grâce à l'identification des liens entre le parcours hydrique de la vigne, l'environnement cultural et la qualité finale du raisin, permettra de garantir un profil de matière première optimal. Un module "cave" qui va intégrer les données du module "plante" et préconiser les opérations œnologiques à tenir en cave afin d'obtenir un vin de bonne qualité.

Le stage s'inscrit dans la phase de développement du module "plante" et a pour objectif d'étudier l'impact de l'état hydrique de la vigne tout au long de la saison sur la qualité des raisins à la vendange. Cette qualité est caractérisée par des analyses physico-chimiques (concentration en sucre dans les raisins, poids des baies, taux d'azote assimilable, composés aromatiques) et dépend de plusieurs facteurs, à la fois agronomiques et œnologiques. Les données relatives à la vigne et aux plantations recueillies en 2012 et 2013 se composent de mesures temporelles pendant la saison et de données effectuées à la récolte. La connaissance du domaine viticole est également disponible et a été implémentée dans une ontologie [1] [11] conçue dans le cadre du projet. Toutefois, l'état hydrique de la vigne pendant la saison ne peut être mesuré directement et il est nécessaire au préalable de l'estimer sur chaque parcelle du protocole.

La relation vécu hydrique de la vigne - qualité du raisin n'est également pas facile à établir en raison d'une part de la forte variabilité du matériel biologique et d'autre part de l'estimation difficile et complexe de l'état hydrique de la vigne qui nous sert de variable explicative. En effet, cette estimation met en œuvre des données temporelles complexes liées à la plante et une expertise agronomique basée sur un modèle écophysiologique. De plus, elle passe par une longue et délicate phase de sélection de données en raison de la nature variable des capteurs de flux de sève placés sur les vignes pour recueillir les données de transpiration.

1. Une ontologie est la formalisation d'un ensemble de concepts et de relations entre ces concepts liés à un domaine d'expertise donné.

1.6 Démarche adoptée

Dans ce rapport, nous utilisons des méthodologies mettant en œuvre des données fonctionnelles (les courbes de vécu hydrique) notées $Ks(t)$ et des connaissances du domaine présentées dans le chapitre 2, dans le but d'étudier un phénomène agronome complexe qu'est l'impact du vécu hydrique de la vigne estimé sur la saison ou une partie de la saison, sur la qualité des raisins.

Les mesures analytiques des variables de qualité du raisin sont recueillies à la récolte et donc disponibles. En revanche, le vécu hydrique de la vigne ne peut être mesuré directement. La procédure d'estimation de cette courbe temporelle du vécu hydrique est développée dans le chapitre 3. Cette phase nécessite en partie l'expertise agronomique disponible via un modèle écophysiologique ainsi que des règles agronomiques et œnologiques définies dans l'ontologie.

La relation vécu hydrique de la vigne - qualité finale du raisin a été déjà étudiée auparavant mais de tels travaux se limitaient à l'étude de la consommation d'eau de la vigne représentée par des mesures occasionnelles, voir des Gachons et al. [2], Koundouras et al. [10]. Dans ce rapport, nous allons considérer le vécu hydrique de la vigne $Ks(t)$ sous forme d'une courbe discrétisée et estimée avec un pas de temps donné. Notre objectif est de le mettre en relation avec les indicateurs de qualité du fruit (concentration en sucre dans les baies, poids des baies, précurseurs d'arômes, etc...) en utilisant de nouvelles techniques d'analyses de données fonctionnelles et des outils d'aide à la décision. Cependant, compte tenu de l'incertitude associée aux données, travailler avec des modèles prédictifs ne serait pas réaliste. Nous parlons d'incertitude pour tenir compte de la variabilité du matériel expérimental notamment les capteurs de flux de sève renvoyant les mesures de transpirations des vignes, et de l'erreur commise dans l'estimation des courbes de vécu hydrique. Nous nous sommes donc orientés vers des modèles explicatifs notamment avec deux démarches que nous présentons dans le chapitre 4.

La première démarche consiste à extraire de la courbe temporelle $Ks(t)$, des variables numériques égales à l'aire sous la courbe de périodes délimitées par les trois principales étapes de croissance de la vigne (nouaison, véraison, maturité) en plus de la récolte. Ces variables sont représentatives du cumul de vécu hydrique de la plante dans ces périodes et sont utilisées en entrée d'arbres de régression de même que d'autres prédicteurs comme la variété ou l'année viticole, dans le but de déterminer les périodes d'influence du vécu hydrique sur les variables de qualité du raisin.

La deuxième démarche consiste à considérer la courbe $Ks(t)$ comme le prédicteur d'un modèle de régression linéaire pour données fonctionnelles. Nous avons utilisé la méthode FLRTI (Functional Linear Regression That's Interpretable) [9], une méthode interprétable d'analyse de données fonctionnelle, pour évaluer le lien de $Ks(t)$ avec la qualité du fruit.

Comme en régression linéaire classique, il est nécessaire de construire la matrice des covariables constituée des courbes discrétisées $Ks(t)$ de dimension $n * p$, où n est le nombre de parcelles suivies sur chaque année et p le nombre de points de mesure de chaque courbe de vécu hydrique. Cependant, les courbes $Ks(t)$ sont obtenues sur une échelle de temps propre à chaque parcelle. Il est donc impératif de toutes les recaler sur une échelle de temps commune en faisant correspondre chaque date de stade majeur de croissance des vignes qui leur sont associées, à une date fixée à l'avance. Ce recalage de courbe est fait sur la base d'un algorithme détaillé dans la suite.

Nous avons également utilisé des procédure de tests de significativité [7] d'une relation linéaire entre la variable fonctionnelle $Ks(t)$ et chacune des variables de qualité du raisin afin de s'assurer de la pertinence du modèle de régression linéaire pour données fonctionnelles. Dans le chapitre 5, nous présentons les résultats et interprétations ressortant des analyses statistiques, et enfin des conclusions et quelques perspectives sont mentionnées dans le chapitre 6.

Chapitre 2

Présentation des données

2.1 Le système d'information Silex-VitiOeno

Silex-VitiOeno Pilotype est un Système d'Informations (SI) permettant de collecter, classifier, traiter et diffuser des données expérimentales sur la vigne, le raisin et le vin et comportant une interface web. Il a été développé au sein de l'UMR MISTEA dans le cadre du projet PILOTYPE. Silex-VitiOeno utilise une base de données PostgreSQL. Le modèle de cette base est générique et peut s'adapter à de nombreux besoins grâce aux notions d'objet (sur lequel s'articule le modèle de la base), de variables (température, pluviométrie, transpirations) et de groupes de variables (données météo, données de maturité). Un objet représente une entité culturale organisée de manière hiérarchique (domaine, parcelle, sous-parcelle, zone) et il est toujours issu d'un objet parent, sauf le domaine qui est le plus haut dans la hiérarchie.

Le SI possède plusieurs fonctionnalités, principalement l'ajout et la modification de données par des utilisateurs enregistrés dans la base. L'ajout peut se faire par formulaire ou par fichier Excel. Les doublons sont gérés de manière automatique, la donnée ne sera pas rajoutée si elle est déjà présente dans la base. Une donnée équivalente à une autre, recueillie à la même date, à la même heure, au même endroit sera modifiée si la valeur de la mesure a changé. Les utilisateurs non-membres peuvent consulter les mesures hors lignes grâce à une recherche en fonction de la date, de l'objet, du groupe de variables ou de la variable auxquels la mesure est rattachée. Les résultats de la requête de recherche peuvent être exportés sous forme de fichier Excel.

La connexion au SI à partir de R se fait de façon simple à l'aide du package "RPostgreSQL" et de lignes de commandes génériques nécessitant au préalable un nom d'utilisateur et un mot de passe fournis par l'administrateur de la base.

```
library(RPostgreSQL) (chargement de la librairie)
drv<-dbDriver("PostgreSQL")
con<-dbConnect(drv, user="diakhate", password="xxxxxxx", dbname="vinnotec")
```

L'import des données journalières est effectué à l'aide de requêtes génériques utilisant la connexion *con* établie auparavant. A titre d'exemple, la commande ci-dessous importera dans le data-frame *meteo* toutes les variables météorologiques des sites suivis, ainsi que leurs valeurs pour des dates allant du 1^e janvier au 30 Novembre 2012.

```
meteo <- dbGetQuery (con, "select objet, date, variable, valeur" from mesures_hors_lignes
where groupe_variables ="Meteorologie_j" and date between "2012-01-01" and "2012-11-30")
```

2.2 Dispositif expérimental

Les données disponibles proviennent d'une expérience multi-site dans le sud de la France. Le même protocole expérimental a été mis en place dans sept sites de la région Languedoc-Roussillon afin de tester les effets du déficit en eau sur la qualité des raisins et des vins dans des conditions environnementales différentes. Il s'agit notamment de La Baume, de l'Ouveillan, du Rieux, du Piolenc, du Saint Sauvignon, du Saint Gervasy et de PechRouge. Au total, 24 parcelles de vigne ont été suivies en 2012 et 33 en 2013 , chacune plantée avec l'une des variétés suivantes : Merlot, Cabernet- Sauvignon, Syrah, Grenache, Chardonnay.

Pour obtenir un plus large éventail d'états hydriques de la vigne au cours de la saison, un traitement d'irrigation a été appliqué pendant deux ans sur chacune des sept combinaisons de sites-variétés. Le traitement d'irrigation est défini suivant les modalités répertoriées dans le tableau 2.1 et répété deux fois sur chaque parcelle.

Irrigation	i_0	i_1	i_2	$iPreV$	$iPostV$
Modalités	Non irriguée	Irriguée	Très irriguée	Avant véraison	Après véraison
Site	Toutes parcelles	Toutes parcelles	Ouveillan	Piolenc	Piolenc

TABLE 2.1 – Tableau explicatif des traitements d'irrigation effectuées sur les parcelles

Dans les parcelles non irriguées, les plantes ont reçu uniquement des précipitations naturelles pendant la saison, tandis que sur les parcelles irriguées, elles reçoivent régulièrement des quantités en eau via des lignes de goutteurs. Sur chaque parcelle, sont suivies 4 vignes distantes de 25 mètres et sur chacune d'elles, ont été placés des capteurs de flux de sève renvoyant chacun une donnée de transpiration toutes les 15 minutes.

Plusieurs types de données, recueillies selon ce protocole expérimental sont disponibles : les données météorologiques locales, les données phénologiques des vignes de chaque parcelle (caractéristiques des étapes de croissance de la vigne), les données relatives au vécu hydrique des vignes notamment les mesures de flux de sève (transpirations), le potentiel de base, l'apex, l'irrigation et les données de qualité du raisin recueillies à la récolte.

2.3 Description des données utilisées

2.3.1 Les données météorologiques

2.3.1.1 Les données météorologiques brutes

Il s'agit des données météorologiques horaires extraites des stations locales de chaque site : la vitesse du vent (en $km.h^{-1}$), la température minimale, maximale et moyenne de l'air (en °C), la radiation solaire (en $W.m^{-2}$), l'humidité relative de l'air (en %), et les quantités de précipitations (en mm). Les données météorologiques journalières ont été obtenues en appliquant une intégration par trapèze sur les données horaires.

2.3.1.2 Les données météorologiques transformées

Le *VPD* horaire (Vapor Pression Deficit) et l'évapotranspiration de référence ET_{ref} sont définis selon les méthodes visées à la FAO-56, voir Allen et al. [1].

La notion d'évapotranspiration potentielle ETP est couramment opposée à l'évapotranspiration réelle. Cette dernière est l'eau réellement dissipée dans l'atmosphère sous forme de vapeur et il est impossible de la mesurer à l'échelle d'une parcelle ou d'une région. A l'opposé de l'évapotranspiration potentielle qui se détermine à partir de formules mathématiques. Ces deux valeurs caractérisant ainsi l'évaporation au niveau du sol et la transpiration au niveau des plantes, sont calculées sur chaque site car elles sont utiles et nécessaires pour étudier les bilans de circulation de l'eau et pour mesurer les besoins en eau des cultures. L'évapotranspiration de référence ET_{ref} (en $mm.d^{-1}$) est une valeur d'évapotranspiration sur une végétation choisie, permettant ensuite d'en déduire l'évapotranspiration pour d'autres couverts végétaux.

Le temps thermique de la vigne (GDD), c'est à dire l'accumulation croissante en degrés-jours depuis le 1^{er} avril, est une mesure utilisée pour calculer l'accumulation de chaleur et sert à caractériser la durée d'un développement biologique, ici la croissance de la vigne. En l'absence d'événements perturbateurs (hiver rude, sécheresse), le développement de la vigne est fortement influencée par la température ambiante. Par exemple, les viticulteurs peuvent grâce au calcul des degrés-jours faire des analyses et estimer la date de floraison pour savoir quand apporter de l'engrais à la plante.

2.3.2 Les données phénologiques

Les phases phénologiques autrement dit les étapes de croissance de la vigne sont mesurées visuellement dans chaque parcelle expérimentale lorsque 50% des vignes ont atteint le stade en question. Les principales étapes retenues sont le débourrement (éclosion des bourgeons), la floraison (ouverture des capuchons floraux), la nouaison (les étamines flétrissent mais restent souvent fixées à leur point d'attache) et la véraison (lorsque 50% du fruit se colore en rouge).

En revanche, la maturité est mesurée et est définie soit suivant une certaine concentration en sucre ou un degré d'alcool probable, présents dans les baies. Ces valeurs seuils sont fixées par les agronomes et varient suivant les cépages. En choisissant le degré d'alcool probable comme référence, la règle est fixée sur le ratio du seuil fixé pour un cépage avec un taux de conversion statique égal à 16.83. A titre d'exemple, pour les cépages de vin rouge comme le Merlot, on considère que la maturité est atteinte si la concentration en sucre dans les baies dépasse 220 g/l.

2.3.3 Les données du vécu hydrique de la vigne

Le vécu hydrique de la vigne est observé avec trois grands indicateurs d'état hydrique : les mesures discrètes du potentiel hydrique foliaire de base avant l'aube, les mesures continues de flux de sève ou de transpirations et les mesures de l'apex.

2.3.3.1 Le potentiel hydrique foliaire de base

Les mesures sont effectuées chaque semaine avant l'aube (entre 3h et 5h du matin), à partir de fin juin jusqu'en mi-Août avec une chambre à pression encore appelée bombe de Scholander. En fin de nuit, alors que la transpiration est négligeable et que la plante a reconstitué ses réserves en eau, on considère que la tension de sève dans le végétal est en équilibre avec le potentiel hydrique du sol. La mesure du potentiel foliaire à cet instant, à l'apparition d'humidité sur la section du faisceau ligneux de la feuille cueillie, renseigne sur la disponibilité en eau du sol et fournit une information sur l'état hydrique dans lequel se trouve le végétal, en raison d'une plus faible variabilité des conditions du milieu. Par définition, la plante est déja en état de "stress" hydrique avec un potentiel de base inférieur ou égal à $-3.5Kpa$ (kilo pascal).

2.3.3.2 Les mesures de flux de sève

Le flux de sève mesure le niveau de transpiration, c'est-à-dire le volume d'eau qui circule dans la plante, depuis les racines, jusqu'aux feuilles en passant le tronc. Elle varie avec l'humidité du sol. Comme le flux de sang permet de diagnostiquer la santé d'une personne, le flux de sève décrit directement la santé de la vigne. FRUITION SCIENCES, partenaire du projet, fournit les capteurs de flux de sève destinés à mesurer les transpirations des plantes.

Sur chaque vignoble, sont choisies 2 parcelles expérimentales irriguée et non irriguée (avec les cas particuliers : très irriguée sur le Site de l'Ouveillan, irriguée avant et après la véraison sur le site de la Piolenc en 2013) et dans chaque parcelle, une rangée de plantes est sélectionnée. Sur chaque rangée, 2 plantes, distantes de 25 mètres chacune, sont équipées chacune d'un capteur qui mesure le flux de sève tous les quarts d'heure. Une rangée de 2 plantes correspond à une répétition de mesure sur chacune des 2 plantes. Pour s'assurer de la fiabilité des données recueillies, des indices de confiance et des pourcentages de fiabilité ont été attribués à chacune d'elles, sur la base de la valeur de transpiration mesurée.

2.3.3.3 La méthode dite des apex

Observé visuellement, l'apex constitue une alternative à la mesure du potentiel hydrique foliaire de base car celle-ci n'est pas évidente à mettre en place concrètement dans certains vignobles. La Chambre d'Agriculture de l'Hérault travaille depuis plusieurs années sur cette méthode d'estimation du "stress" hydrique dite "méthode des apex" afin de pouvoir déclencher l'irrigation.

2.3.3.4 Procédure de sélection des données de flux de sève fiables

Un capteur renvoie une mesure de transpiration tous les quarts d'heure et seules les données diurnes entre 7h à 22h sont considérées, soit 64 données par jour. A chaque mesure est associée un indice de fiabilité allant de 0 (Donnée bonne) à 3 (donnée non fiable), cf Figure (2.1).

EN 2012, l'indice journalier de fiabilité du capteur est défini sur la somme journalière des indices horaires S_{ID_h}. Un capteur est considéré comme fiable (ID=2) si $S_{ID_h} < 3$, moyennement fiable (ID=1) si $3 < S_{ID_h} < 20$ et pas du tout fiable (ID=0) si $S_{ID_h} > 20$. Au final, un certain nombre de données ont été sélectionnées sur la base de ces seuils de confiance journaliers.

En 2012, nous avons retenu les vignes (capteurs) ayant un indice de confiance journalier de 2 soit 97% de données horaires fiables. En 2013, le seuil a été fixé à la médiane des pourcentages de fiabilité journaliers attribués sur toute la saison, soit 50%. Nous avons ainsi retenu 4166 données de transpiration fiables sur un total de 7781 données recueillies en 2012, soit 46% de données considérées comme non fiables. En 2013, sur 9045 mesures de transpiration recueillies, 6896 données fiables ont été retenues, soit 24% de données non fiables. Il faut remarquer que ce choix du seuil de confiance optimal a été moins rigoureux en 2013, dans un souci de conserver le maximum de vignes (capteurs) pour assurer la robustesse des analyses statistiques.

Ces données de transpiration mesurées sur 2012 et 2013, ont été ensuite représentées sur chaque parcelle (combinaison site-variété-irrigation) avec un pas de temps de 2 jours afin de voir l'allure des courbes. Les points de mesure correspondent aux ronds sur les courbes.

La figure 2.2 représente l'évolution des mesures de transpirations de chacune des 4 plantes de la parcelle RIEUX-Grenache-i1 en 2012. Le but de cette représentation est de sélectionner les plantes de la parcelle ayant le moins de données manquantes sur plus de 4 jours (3 points d'affilée).

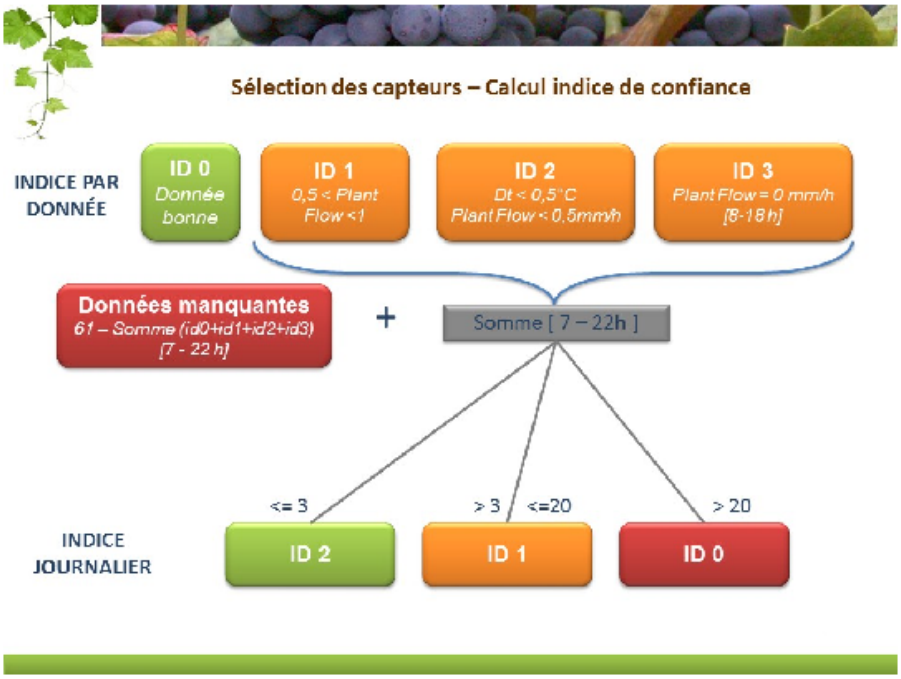


FIGURE 2.1 – Calcul des indices de confiance journaliers et sélection des capteurs fiables

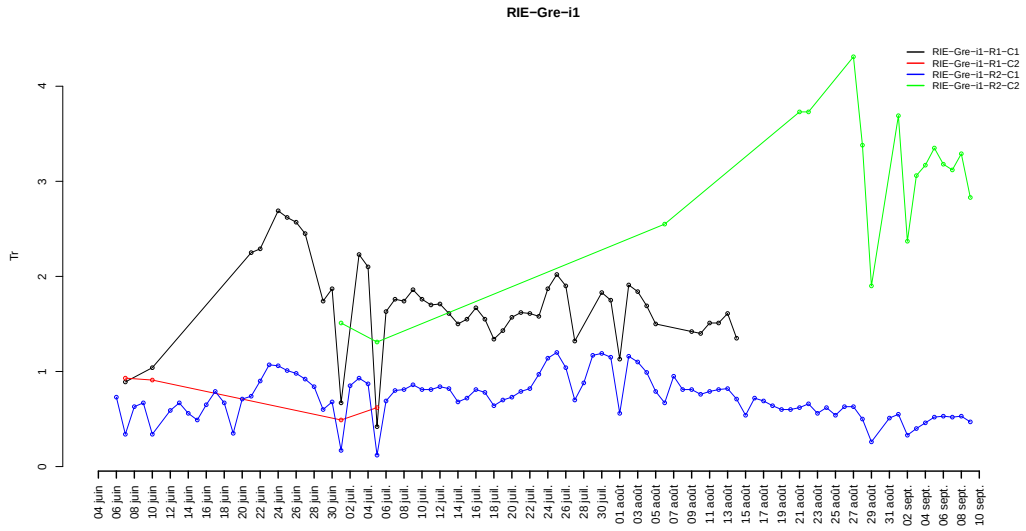


FIGURE 2.2 – Graphiques des transpirations sur les plantes de la parcelle *RIE-Grenache-i1*

Ceci pour que la courbe de transpiration moyenne $\overline{T}(t)$ des mesures de chaque plante de la parcelle soit définie sur le plus de dates possibles, en particulier sur les dates d'atteinte des stades phénologiques majeurs notamment la nouaison, la véraison et la maturité.

A titre d'illustration, sur ce graphique, nous n'avons retenu que la courbe bleue. En considérant la rouge, la courbe $\overline{T}(t)$ ne serait plus définie après le 5 juillet 2012 (on n'aurait pas eu les données jusqu'à la véraison qui survient le 14 Août), ni après le 15 Août 2012 en considérant la noire, et entre le 5 juillet et le 24 Août 2012 en ayant considéré la courbe verte (on n'aurait pas eu de données de véraison ni de maturité sur cette parcelle).

Cette règle de sélection a été reconduite sur toutes les parcelles expérimentales en 2012 et 2013. Au final, 25 plantes réparties sur 15 parcelles ont été retenues en 2012, et 33 plantes en 2013 réparties sur 18 parcelles de vignes. Nous obtenons ainsi sur 2012 (resp. 2013) 15 (resp. 18) courbes de transpirations moyennes représentatives du vécu hydrique des vignes sur chaque parcelle.

Par ailleurs, on peut constater que la courbe bleue choisie présente des valeurs de transpiration très basses et ne reflète pas totalement la moyenne réelle sur cette parcelle. Pour élargir les possibilités de sélection et écarter le moins de plantes possible, une alternative aurait été d'extrapoler les extrémités des courbes de transpiration ne comportant pas beaucoup de données manquantes au cours de la saison. Cela consiste à compléter la courbe en fin et début de saison par la dernière valeur mesurée. Par exemple, sur la parcelle *RIE-Grenache-i1*, la courbe noire serait complétée à partir du 15 août avec la valeur de $T(t)$ à cette date. Les courbes rouge et verte comportant trop de données manquantes et surtout en milieu de saison ne seraient toutefois pas prises en compte.

2.3.4 Les données relatives à la qualité du raisin

Elles sont recueillies à la date de récolte qui reste par ailleurs variable selon les parcelles. Elles sont constituées des données liées à la composition du raisin telles que la concentration en sucre, le poids des baies, les anthocyanes, l'azote assimilable (Nass), etc... et des données relatives au potentiel aromatique du raisin comme le PDMS, le cys3MH, le G3MH ou le GSH. Ces dernières sont obtenues au terme d'analyses chimiques exclusivement menées par NYSEOS et elles ont été choisies en priorité pour mener les analyses statistiques durant ce stage, car étant des facteurs déterminants de la couleur, la composition et le goût du vin final produit.

Saison	Nombre parcelles	Sucre	Nass	Poids des baies	PDMS	cys3MH	G3MH	GSH
2012	15	n=14	n=14	n=14	n=14	n=14	n=14	n=14
2013	18	n=17	n=17	n=17	n=16	n=16	n=16	n=16

TABLE 2.2 – Nombre de mesures des variables de qualité sur les parcelles suivies en 2012 et 2013

Le tableau 2.2 fait état du nombre de mesures relatives à la qualité du raisin disponibles sur les parcelles suivies. En 2012, nous disposons des données de qualité sur 14 parcelles pour un total de 15 parcelles retenues : aucune récolte n'a été faite sur la parcelle StGer-Mer-CH0. Pour la même raison, en 2013, nous disposons des données de composition du raisin sur 17 parcelles et des données d'arômes sur 16 parcelles pour un total de 18 parcelles : les analyses chimiques n'ont pas été effectuées sur les raisins de la parcelle PIO-Gre-iPrev en plus de la parcelle StGer-Mer-CH0.

Au final, les méthodes statistiques sont appliquées sur les données de qualité des raisins mesurées sur 14 parcelles en 2012. Et sur celles de 17 parcelles en 2013 (resp. 16 parcelles) sur lesquelles sont mesurées les données relatives à la composition du fruit (resp. les données relatives au potentiel aromatique).

Les graphiques 2.3 et 2.4 ci-dessous présentent les résultats de l'analyse descriptive de trois variables de qualité du raisin sur 2012 et 2013. L'analyse est faite à l'aide de boxplots représentant la valeur de la variable à la récolte en fonction du cépage. Il s'agit de la concentration en sucre dans les baies, du poids des baies et d'un composé aromatique du raisin : le G3MH. Les jeux de données complets sont joints en annexes (annexes A & B).

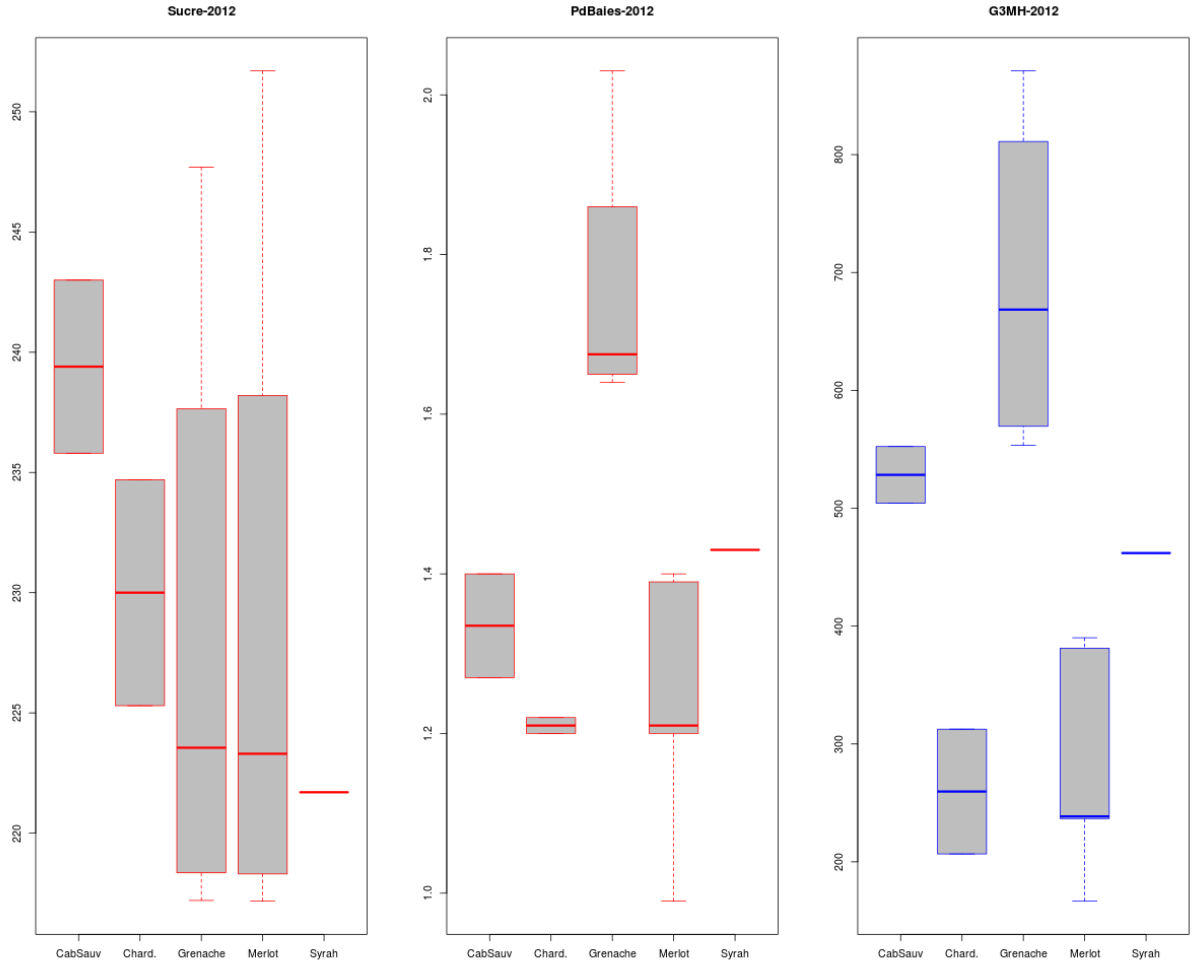


FIGURE 2.3 – Boxplot de l'évolution de la concentration en sucre, du poids des baies et du G3MH en fonction des cépages plantés sur 2012

En 2012 La concentration en sucre est plus élevée dans les raisins sur les cépages de Merlot et de Grenache avec une valeur maximale de 251.7 g/l sur la parcelle Ouveillan-Merlot-i_2. Une valeur atypique, vu sa représentation à l'extrémité de la barre hachurée du boxplot. Il subsiste également une forte disparité entre les concentrations en sucre dans les raisins des parcelles de Merlot et de Grenache.

Le boxplot montre que le poids des baies est nettement plus élevé sur les parcelles de Grenache. Les plus petites baies sont produites dans les parcelles de Merlot et de Chardonnay.

Pour le G3MH, sa teneur est plus faible dans les raisins du Merlot et du Chardonnay que ceux de la Grenache et du Cabernet-Sauvignon.

En 2013

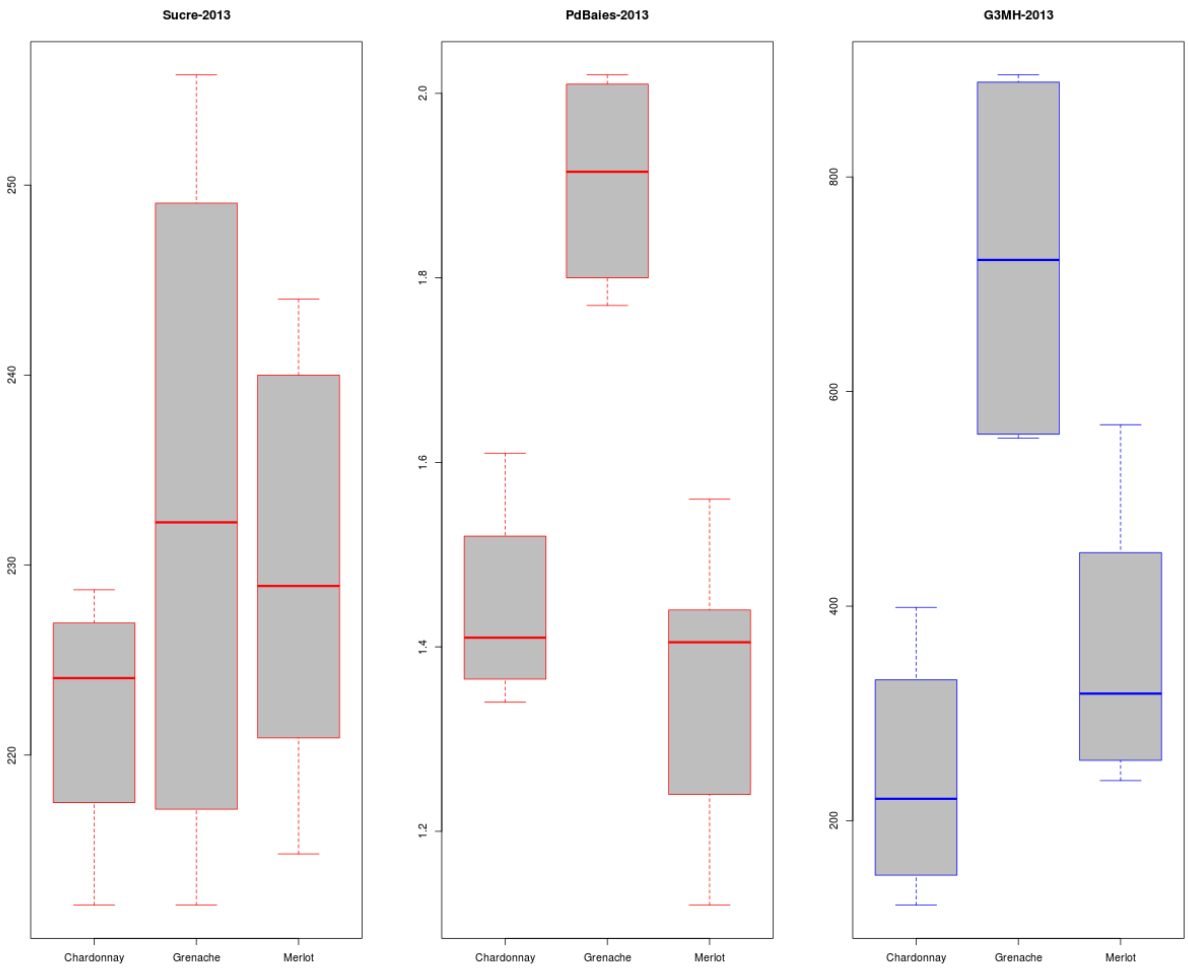


FIGURE 2.4 – Boxplot de l'évolution de la concentration en sucre, du poids des baies et du G3MH en fonction des cépages plantés sur 2013

Comme en 2012, nous pouvons constater que les concentrations en sucre dans les raisins issus des parcelles de Grenache et de Merlot sont les plus élevées de la saison.

Le poids des baies est nettement supérieur sur les parcelles de Grenache et moins élevées sur celles de Merlot et de Chardonnay comme en 2012.

Pour le G3MH, même constat qu'en 2012 avec une teneur plus faible dans les raisins des parcelles de Merlot et de Chardonnay.

En dernier, nous avons calculé les corrélations entre les variables de qualité, voir en annexes les tableaux C.1 et C.2. Et au final, les liens mis en exergue sur les deux saisons sont entre le G3MH et le poids des baies (68% en 2012 et 69 % en 2013). Plus le raisin est lourd, plus la teneur en G3MH est élevée. Un lien fort existe également entre la concentration en sucre et l'azote assimilable qui, elles, évoluent en sens inverse (-65% en 2012 contre -58% en 2013). Il semble que plus la concentration en sucre est élevée dans les baies, moins il y aura d'azote assimilable.

Chapitre 3

Estimation de la courbe du vécu hydrique de la vigne

Le vécu hydrique Ks est défini comme étant le rapport entre la transpiration observée sur la plante $T(t)$ et la transpiration maximale $T_{max}(t)$:

$$Ks(t) = \frac{T(t)}{T_{max}(t)} \tag{3.1}$$

Il représente la baisse de la consommation d'eau de la vigne en raison d'un déficit hydrique du sol et correspond également au niveau d'eau utilisée par la vigne par rapport au niveau maximal.

$Ks = 1$ correspond à la situation où le niveau maximal d'eau pouvant être utilisé par la vigne est entièrement atteint. Lorsque $Ks < 1$, la consommation d'eau maximale de la vigne n'a pas été atteinte, autrement dit la consommation journalière en eau est limitée et la valeur du Ks indique un certain déficit en eau. Nous qualifierons cette situation de "stress" hydrique de la vigne.
$Ks = 0$ correspond donc à une situation de "stress" extrême où la plante n'a pas consommé d'eau à ces dates. Des méthodes générales d'estimation du vécu hydrique Ks ont déjà été proposées dans la littérature. Ferreira et al. [5] ont rapporté des résultats montrant des variations spécifiques du Ks pour des vignes soumises à différents régimes d'humidité du sol.

Dans le contexte de la vigne, cf équation (3.1), $T(t)$ est la transpiration journalière (moyenne des transpirations horaires) mesurée par les capteurs de flux de sève installés sur les vignes suivies. T_{max}(t) est la transpiration maximale obtenue quand l'humidité du sol est non limitative, et a été définie comme suit par Allen et al. [1] :

$$T_{max}(t) = Kc_B(t) * ET_{ref}(t) \tag{3.2}$$

où ET_{ref} est l'évapotranspiration de référence mesurée et $Kc_B(t)$, un coefficient linéaire à déterminer spécifiquement sur chaque combinaison parcelle-variété afin de tenir compte des différences entre les écosystèmes et les densités de plantations.

3.1 Détermination du coefficient de culture $Kc_B(t)$

Le coefficient cultural $Kc_B(t)$ est spécifique à la vigne et varie en fonction de plusieurs facteurs : les caractéristiques de la culture, le développement de la surface foliaire, les conditions climatiques en particulier au début de la croissance de la plante, le VPD, le rythme des pluies et des irrigations.

Il donne une estimation de la transpiration maximale de la plante lorsqu'il est multiplié par l'évapotranspiration de référence ET_{ref}. Ce qui correspond au volume d'eau normalement utilisé par la plante en absence de déficit d'humidité du sol, cf équation (3.2).

Nous allons essentiellement utiliser les relations et concepts implémentés dans l'ontologie OVWS (Ontology of Vine Water Stress) basée sur l'expertise agronome et viticole. Elle a été mise en place dans le cadre du projet pour formaliser la connaissance du domaine.

La trajectoire du $Kc_B(t)$ peut être divisée en deux principaux stades de croissance : L_{dev} et L_{mid} comme reporté par Allen et al.[1], voir Figure 3.1. L_{dev} correspond à la période où l'indice de surface foliaire (LAI) croît rapidement en temps thermique sous forme linéaire et L_{mid} la période où elle stagne et atteint un plateau constant. L'indice de surface foliaire (LAI) étant le ratio entre la surface totale supérieure des feuilles et la surface du sol sur laquelle la végétation se développe.

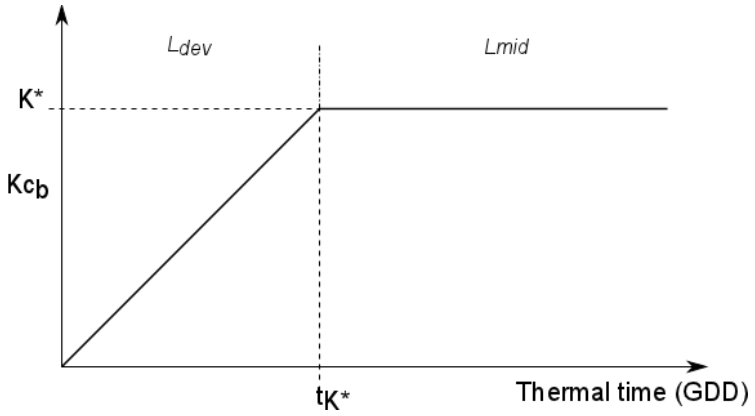


FIGURE 3.1 – Courbe théorique de l'évolution du Kc_B au cours de la saison [1].

Pour déterminer $Kc_B(t)$, deux hypothèses sur la forme de la courbe sont faites :

$$Kc_B(t) = f(t) \text{ for } t < t_{K^*} \tag{3.3}$$

$$Kc_B(t) = K^* \text{ for } t \geq t_{K^*} \tag{3.4}$$

où $f(t)$ est supposé être linéaire en t et t_{K^*} correspondant au point d'arrêt à partir duquel la courbe $Kc_B(t)$ atteint le plateau K^*, voir Figure 3.1. En se rapportant à l'équation (3.2), nous faisons l'hypothèse qu'en absence de déficit hydrique, K^* est défini commme suit :

$$K^* = \frac{T(t_{K^*})}{ET_{ref}(t_{K^*})} \tag{3.5}$$

La partie clé est alors de trouver t_{K^*} (ou indifféremment K^*). Ceci en opérant différentes règles de sélection implémentées dans l'ontologie OVWS, afin d'obtenir un ensemble restreint de t_{K^*} potentiels. L'intérêt d'avoir ces règles dans l'ontologie est double : *i*) ils sont bien explicités et *ii*) ils peuvent évoluer indépendamment des procédures numériques.

3.1.1 Sélection basée sur la phénologie

Une relation de nature linéaire existe entre les variations du $Kc_B(t)$ et l'indice de surface foliaire ou avec la fraction de recouvrement du sol par la vigne, voir Ferreira et al.[5]. Nous supposons alors que le pic t_{K^*} (ou K^*) est atteint lorsque l'indice de surface foliaire cesse de croître. Par conséquent, cette période est définie entre le débourrement et la véraison.

3.1.2 Sélection basée sur le potentiel foliaire hydrique de base

Les conditions du déficit hydrique maximal du sol peuvent être déduites des mesures du potentiel hydrique foliaire de base associées à un intervalle de confiance dérivé du VPD mesuré. La règle générée dans ce sens est que le K^* est atteint avant la première date d'apparition de "stress" hydrique (arrêt de l'allongement des pousses de la vigne). Cette situation correspond à une valeur de potentiel hydrique foliaire de base de -3.5 kPa. Par conséquent, les dates possibles d'atteinte du t_{K^*} sont celles avant l'apparition des premiers signes de déficit hydrique et s'il n'y en a pas, au plus tard à la véraison.

3.1.3 Sélection basée sur la météorologie

Les mesures de transpiration via les capteurs de flux de sève dépendent énormément des conditions climatiques, principalement la lumière et le *VPD*. Pour tenir compte de la sensibilité des transpirations au $VPD(t)$, une règle de filtrage a été définie pour éliminer les potentiels t_{K^*} obtenus dans les périodes de fortes chaleurs, sauf en cas d'irrigations ou de précipitations. Ces périodes sont caractérisées par une valeur de $VPD(t)$ supérieure à 3.5 kPa.

3.1.4 Sélection basée sur l'apex

La règle imposée sur la valeur de l'apex est d'éliminer tous les t_{K^*} potentiels avant la première date d'une valeur d'apex inférieure ou égale à 1.5.

3.1.5 Sélection finale basée sur la forme de la courbe

Par définition, t_{K^*} est atteint lorsque le ratio $\frac{T(t)}{ET_{ref}(t)}$ sur une parcelle a atteint son maximum pendant quelques jours. Les options possibles pour t_{K^*} sont donc les points sur lesquels la dérivée première du ratio est nulle et la dérivée seconde, négative. Cela revient à détecter les points d'inflexion, c'est-à-dire les points de changement de concavité de la courbe $\frac{T(t)}{ET_{ref}(t)}$.
Cette courbe est interpolée finement sur la saison avec la fonction $splinefun$ de R afin de contourner la présence de points manquants (données non fiables) à certaines dates.

Les courbes de dérivées sont obtenues avec les paramètres *deriv* de la fonction $splinefun$. Les points d'inflexion sont détectés simultanément comme les points de changement de signe de la dérivée première de $\frac{T(t)}{ET_{ref}(t)}$ et de dérivée seconde négative.

Ainsi, l'analyse de cette courbe $\frac{T(t)}{ET_{ref}(t)}$ associée à toute la connaissance agronomique conduit à la proposition d'un petit ensemble de t_{K^*} candidats. Cet ensemble de points est soumis à l'expertise de l'agronome pour le choix final car il est le plus au courant des pratiques agronomiques menées et de certains événements qui pourraient avoir interféré avec la croissance de la vigne (irrigation, fortes pluies, système de treillis...) et donc avec la courbe $Kc_B(t)$.

3.2 Bilan des procédures d'estimation de la courbe Ks(t)

Cette phase d'estimation des courbes de vécu hydrique a été particulièrement minutieuse et laborieuse. Elle a nécessité dans un premier temps l'implémentation de toutes les règles de sélection agronomiques et œnologiques décrites plus haut afin d'obtenir les ensembles de points potentiels t_{K^*} sur les courbes $\frac{T(t)}{ET_{ref}(t)}$ de chaque parcelle.

Ensuite, il a fallu continuellement interagir avec les partenaires du projet notamment FRUITION SCIENCES pour le choix final du point t_{K^*} sur la parcelle, et aussi avec les informaticiens du SI pour d'éventuelles précisions sur les données recueillies. Le choix final du t_{K^*} se fait simultanément en regard de l'évolution du potentiel de base et des données climatiques enregistrées aux points candidats t_{K^*}. En dernier, nous avons pu retracer sur chaque parcelle le coefficient cultural $Kc_B(t)$ à partir duquel va se déterminer la courbe finale de vécu hydrique de la vigne $Ks(t)$. Les courbes de vécu obtenues par année et par cépage sont présentées dans le tableau 3.1.

Saison	Cépage	Nombre de courbes Ks(t)
2012	Merlot	5
	Cabernet Sauvignon	2
	Grenache	4
	Syrah	1
	Chardonnay	2
2013	Merlot	8
	Grenache	5
	Chardonnay	4

TABLE 3.1 – Distribution des courbes de vécu hydrique Ks(t) par cépage sur 2012 et 2013

3.3 Variables scalaires du cumul de vécu hydrique

Les courbes temporelles $Ks(t)$ obtenues peuvent être résumées par des paramètres scalaires. Ces derniers sont obtenus en appliquant une intégration par trapèze sur les régions en dessous des courbes à des périodes d'intérêt bien définies.

En premier, trois périodes sont définies suivant le temps phénologique de la vigne : la période de pré-véraison allant de la nouaison à la véraison, la période post-véraison qui va de l'étape de la véraison à la récolte et toute la saison de croissance : de la nouaison à la récolte. Ensuite, la période de post-véraison a été divisée en tenant compte du stade de maturité des baies, ce qui apporte en complément une quatrième période supplémentaire allant de la véraison à la maturité.

Nous disposons ainsi de quatre variables scalaires sur chaque parcelle (NouVer, VerRec, VerMat, NouRec) quantifiant le vécu hydrique de la vigne cumulé sur des périodes phénologiques majeures. Ces variables vont être utilisées comme prédicteurs dans des méthodes d'arbres de régression que l'on met en œuvre dans la suite pour expliquer la qualité du raisin récolté.

3.4 Analyse descriptive des scalaires de vécu hydrique

La variable d'intérêt qu'est le vécu hydrique représente la consommation d'eau de la vigne par rapport au niveau maximal. Ce ratio est compris entre 0 et 1. Par définition, plus il est proche de 1, plus le déficit en eau de la vigne est faible et donc le stress hydrique est faible. Autrement dit, le vécu hydrique évolue toujours à l'opposé du stress hydrique de la vigne.

Dans cette section, nous nous intéressons à l'évolution des variables scalaires du cumul de vécu hydrique calculées entre les stades phénologiques majeurs, sur chacune des saisons et de façon générale sur 2012 et 2013.

3.4.1 Cumul de vécu hydrique entre la nouaison et la véraison

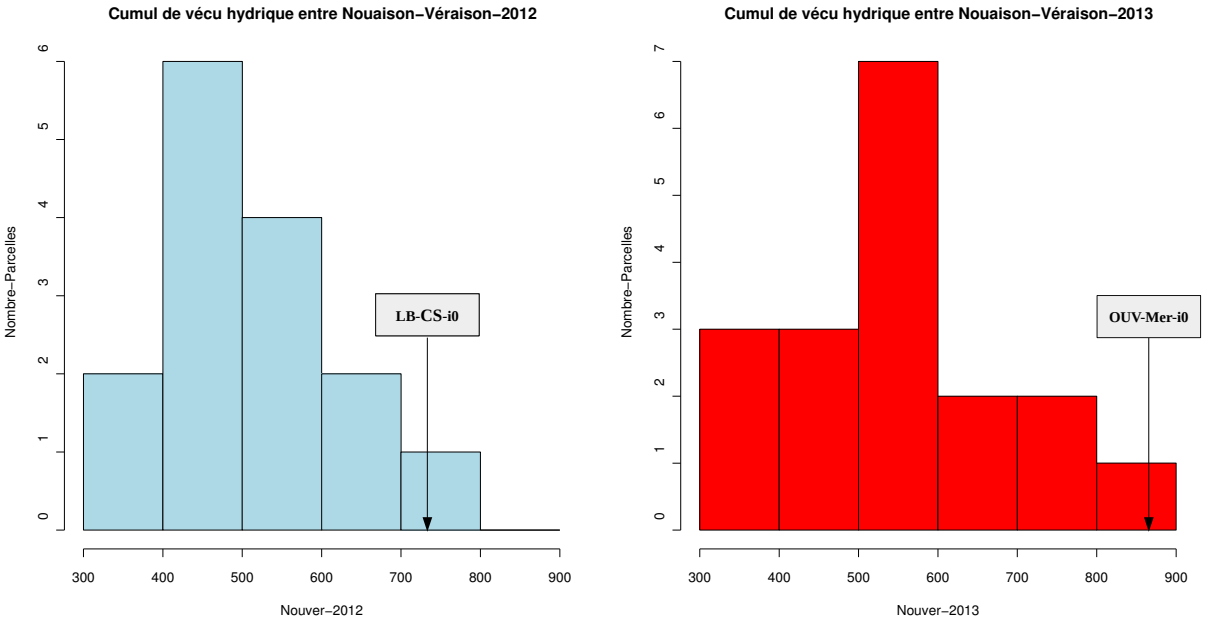


FIGURE 3.2 – Cumul de vécu hydrique entre la nouaison et la véraison sur 2012 et 2013

Le vécu hydrique s'étend sur des valeurs comprises entre 300 et 900. Par rapport à la médiane de la distribution, égale à 600, 3 parcelles sur 15 (resp. 5 parcelles sur 18) ont été les plus irriguées en 2012 (resp. en 2013), que ce soit par précipitations naturelles ou par traitement d'irrigation. La proportion de parcelles dans le régime sec et irrigué est la même sur les deux saisons.

Cependant, des parcelles non irriguées comme la LB-CS-i0 (2012) et la OUV-Mer-i0 (2013) sont parmi les plus irriguées de la saison avec des cumuls hydriques respectifs de 730 et 869. Ceci peut être dû à la pluviométrie importante enregistrée sur ces sites de LaBaume et de l'Ouveillan au cours de la pré-véraison.

Ce qui porte à croire que le traitement d'irrigation qui à l'origine a été mis en place pour renforcer la consommation hydrique des vignes ne permet pas toujours d'obtenir un vécu hydrique supérieur à celui des vignes sur les parcelles non irriguées. Les pluies sur ces dernières ont grandement perturbé l'un des principaux objectifs de l'étude qui était de pouvoir séparer les effets régime sec/irrigué/bien irrigué sur les vignobles car des parcelles non irriguées ont été plus hydratées que des parcelles irriguées sur les deux saisons.

3.4.2 Cumul de vécu hydrique entre la véraison et la maturité

Les parcelles de vignes sont réparties de façon homogène entre le régime sec et hydraté en 2012 par rapport à la médiane de la distribution, voir figure 3.3. En 2013, les parcelles ont toutefois absorbé plus d'eau avec 12 parcelles aux cumuls de vécu hydrique supérieurs au cumul médian. Ceci s'explique par le fait que l'été a été plus pluvieux en 2013 en sachant que la post-véraison se situe généralement entre début Août et Fin Septembre.

En revanche, on retrouve toujours cet effet inversé d'une part avec les vignes de parcelles irriguées qui ont des cumuls hydriques très faibles : la Piolenc-Grenache (en 2012) avec une valeur de 144.337 ou la SaintGervasy-Merlot (en 2013) avec 213.467. Et d'autre part, les vignes sur des parcelles non irriguées qui ont enregistrées les cumuls hydriques maximaux sur la post-véraison

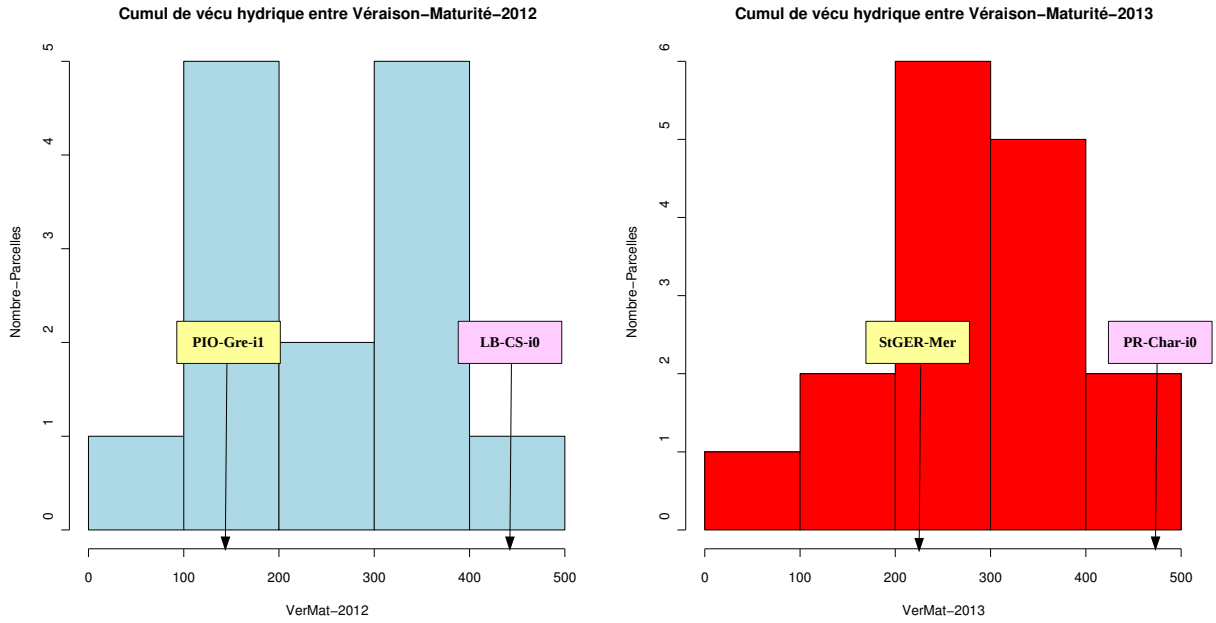


FIGURE 3.3 – Cumul de vécu hydrique entre la véraison et la maturité sur 2012 et 2013

comme LaBaume-CabernetSauvignon (en 2012) pour une valeur de 442.435 et la PechRouge-Chardonnay (en 2013) pour 475.561. Cette situation inversée s'explique par les fortes précipitations sur les parcelles non irriguées. En effet, le site PechRouge non irrigué a enregistré la pluviométrie maximale avec 50 mm au soir du 07 septembre. Le traitement d'irrigation que l'on a voulu mettre en avant pour séparer les niveaux de consommation d'eau des vignes sur les parcelles irriguées et non irriguées a été perturbé par les précipitations notées sur la période post-véraison. Ainsi, le facteur d'irrigation ne peut donc plus être pris en compte en terme d'interprétation de l'influence du vécu hydrique sur la qualité des raisins à la vendange.

3.4.3 Cumul de vécu hydrique entre la nouaison et la récolte

Cet histogramme, voir figure 3.4, relate le cumul de vécu hydrique des vignes tout au long de la saison de culture. Les tendances relevées sur les périodes phénologiques intermédiaires se confirment. Les parcelles sont proportionnellement réparties suivant le régime sec et irrigué, de la même manière que sur 2012 et 2013. On relève toutefois plus de parcelles dans le régime sec en référence à la valeur de cumul de vécu hydrique médian (9 sur 15 en 2012 et 11 sur 18 en 2013).

Par ailleurs, on note plus de parcelles bien irriguées en 2013 qu'en 2012, sans doute en raison de la forte pluviométrie enregistrée pendant l'été, dans la période post-véraison. L'irrigation n'a pas donné les résultats escomptés sur la saison globalement : les vignes suivies sur LaBaume Cabernet Sauvignon et non irriguées sur toute la saison enregistrent le vécu hydrique maximal de même que sur l'Ouveillan-Merlot en 2013.

En résumé, le vécu hydrique qu'on a cherché à contrôler dans cette étude avec le traitement d'irrigation sur certaines parcelles a été grandement perturbé par les précipitations reçues sur les sites. En effet, sur certaines parcelles non irriguées qui ont enregistré une importante pluviométrie, surtout pendant l'été, les vignes présentent un cumul hydrique supérieur à ceux de parcelles irriguées tout au long de la saison. L'irrigation n'est donc plus un facteur prédominant pour modéliser la consommation d'eau de la plante.

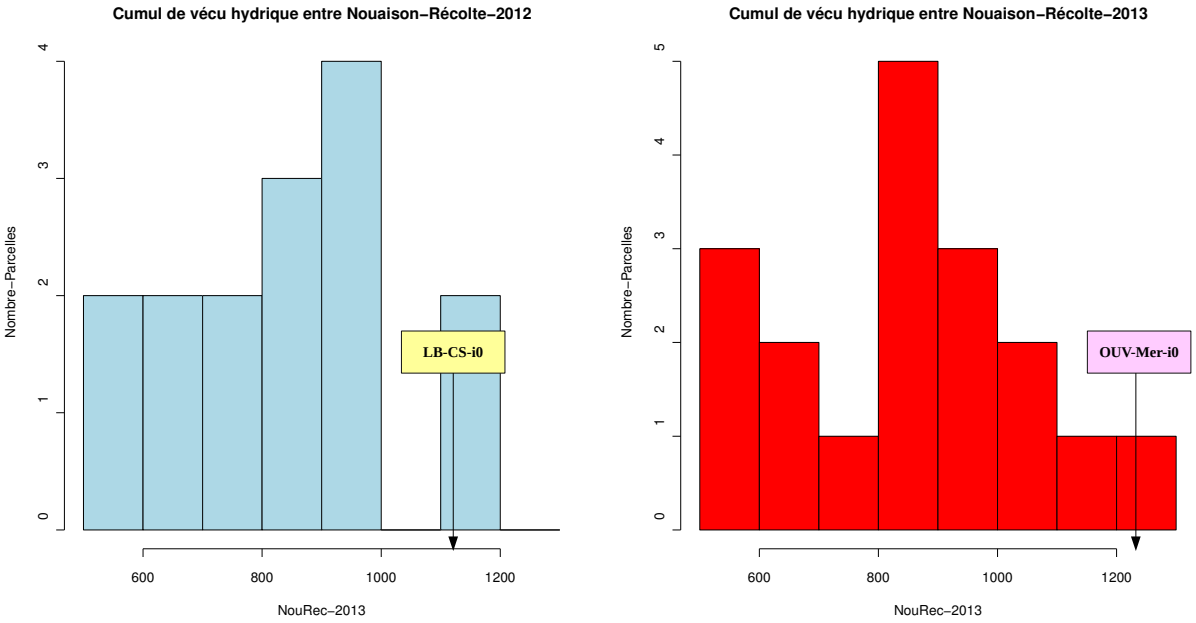


FIGURE 3.4 – Cumul de vécu hydrique entre la nouaison et la récolte sur 2012 et 2013

Dans la suite, nous utilisons ces variables du cumul de vécu hydrique comme variables explicatives, en plus du cépage, de méthodes d'arbres de régression pour expliquer l'influence du vécu hydrique sur la qualité du raisin à la récolte.

Chapitre 4

Méthodes statistiques utilisées

4.1 Les Arbres de régression

4.1.1 Présentation générale de la méthode

On suppose que sur n individus nous disposons d'une variable cible Y à expliquer (ou à prédire) qualitative ou quantitative, ainsi que de variables explicatives quantitatives et/ou qualitatives X_1,X_2,....X_p.

Les n individus sont situés dans l'espace engendré par les variables explicatives. L'objectif est de faire un partitionnement récursif de l'espace engendré par les variables, autrement dit de diviser cet espace en rectangles (multidimensionnels) qui ne se chevauchent pas. Les individus à l'intérieur de chacun des rectangles devant être les plus homogènes possibles. On parle d'homogénéité si :

- Les individus sont de même classe si Y est qualitative ;
- Les individus ont des valeurs de Y "proches" si Y est quantitative.

Ce partitionnement devient difficile si l'espace engendré par les prédicteurs est de dimension supérieure à 2, d'où l'intérêt d'utiliser des arbres. On parle d'arbre de régression si la variable réponse Y est quantitative et d'arbre de classification si elle est qualitative.

L'échantillon complet est au sommet de l'arbre, et chaque division correspond à une condition à vérifier. Pour chacun des individus, si la condition est vérifiée la lecture de l'arbre se fait par la gauche, sinon par la droite. L'arbre est constitué de la racine, des branches, de feuilles ou nœuds intermédiaires et de nœuds terminaux. On peut aussi parler de noeuds parents et de nœuds fils.

Au final, d'un point de vue explicatif ces arbres permettent de savoir quelles variables expliquent le plus l'hétérogénéité entre les individus concernant Y, et d'un point de vue prédictif ils permettent de disposer d'une règle de décision pour affecter de nouveaux individus à l'un des nœuds terminaux.

4.1.2 Méthode générale de construction d'un arbre de régression

Nous nous intéressons uniquement à la construction d'un arbre de régression, compte tenu du caractère quantitatif des variables Y de qualité du raisin récolté que nous avons à étudier.

4.1.2.1 Liste des divisions possibles

Seules des divisions binaires sont envisagées, c'est-à-dire qu'un noeud se pourra pas se diviser en plus de 2 noeuds fils. Si k valeurs sont prises par les n individus, on considère qu'on a une variable qualitative ordinale à k modalités, et il y a $k-1$ divisions possibles.

4.1.2.2 Choix de la meilleure division possible

Une division est meilleure qu'une autre si elle aboutit à des nœuds terminaux qui sont aussi homogènes que possible (ou le moins hétérogènes possibles). L'idée est de calculer la valeur d'une fonction d'hétérogénéité permettant de quantifier l'hétérogénéité d'un nœud donné A :

$$SS(A) = \sum_{i \in A} (y_i - \overline{y}_A)^2 \tag{4.1}$$

Elle est nulle si le nœud A est parfaitement homogène (tous les individus du nœud ont la même valeur de Y) et sa valeur augmente si les valeurs de Y des individus du nœud se dispersent. Lorsque qu'un nœud est divisé en deux nœuds A_G et A_D, son hétérogénéité se réduit. La réduction d'hétérogénéité d'un nœud A due à la division d est :

$$\Delta(A, d) = SS(A) - ((SS(A_G) + SS(A_D)) \tag{4.2}$$

La meilleure division possible d'un nœud peut être encore défini comme celle réduisant l'hétérogénéité du nœud parent.

4.1.2.3 Critère d'arrêt

Un nœud ne se divise plus, devient donc un nœud terminal, s'il est pur, ou s'il n'y a pas de division admissible ou s'il a un effectif plus petit qu'un seuil fixé. Des critères portant sur l'arbre entier peuvent être aussi utilisés tels que la profondeur de l'arbre atteignant une limite fixée, la qualité de l'arbre n'augmente plus de façon sensible, etc...

4.1.2.4 Affectation des nœuds terminaux

Une fois que l'arbre est construit, on doit affecter une valeur à chaque nœud terminal. Le nœud prend la valeur moyenne de Y calculée sur les individus du nœud.

4.1.3 Élagage des arbres de régression

Si l'arbre est très court (peu profond), l'hétérogénéité est importante au sein des nœuds terminaux. Et s'il est très long (très profond) avec beaucoup de divisions et beaucoup de nœuds terminaux, alors il est peu fiable. En effet, l'arbre est trop spécifique aux données d'apprentissage et on parle de "sur-apprentissage". La règle de décision est non reproductible si l'échantillon est un peu modifié. Il faut alors procéder à l'élagage de l'arbre obtenu.

Élaguer un arbre revient à supprimer certains sous-arbres. Le but est de garder des sous-arbres ayant de petits taux d'erreur. Dans le cas d'une variable quantitative, on s'intéresse à l'hétérogénéité qui subsiste au sein des nœuds terminaux une fois l'arbre construit pour évaluer son taux d'erreur. Celui ci est égal à la somme des hétérogénéités des nœuds terminaux, soit :

$$c(T) = \sum_{A \in terminaux} SS(A) = \sum_{A \in terminaux} \sum_{i \in A} (y_i - \overline{y}_A)^2 \tag{4.3}$$

Pour faire ce choix de sous-arbre faisant le compromis entre un petit taux d'erreur (moindre coût) et une faible complexité (pas trop de nœuds), on utilise une fonction de coût-complexité.
Soit un sous-arbre T d'un arbre plus grand T_0, son nombre de noeuds terminaux est noté |T|. Le coût-complexité de cet arbre est donné par :

$$c_\alpha(T) = c(T) + \alpha|T| \tag{4.4}$$

avec α un paramètre de complexité, qui permet de pénaliser la complexité d'un arbre. Plus α est grand, plus on pénalise la complexité d'un arbre. Pour α =0, le sous-arbre de coût-complexité minimum est l'arbre maximal T_{max}.

4.1.4 Mise en œuvre : La procédure CART

Nos travaux sont effectués à partir des arbres CART (Classification And Regression Tree) introduits par Breiman et al. [3], un des principaux algorithmes d'apprentissage par arbres de décision, souvent utilisés par les statisticiens. Nous utilisons le logiciel R pour implémenter des arbres de décision avec le package *rpart* décrit dans R Development Core Team (2009).

L'idée à la base de la procédure CART est de construire un arbre maximal n'ayant que des noeuds terminaux homogènes (ou plus aucune division admissible) puis d'élaguer l'arbre a posteriori. La procédure est la suivante :

1. On construit l'arbre maximal T_{max}, aussi noté T_0 associé au paramètre de complexité α=0 ;
2. On augmente progressivement α pour tomber sur un seuil α_1 pour lequel un sous-arbre T_1 de T_0 sera préférable à T_0. T_1 est donc le sous-arbre de T_0 avec le plus petit α possible ;
3. On réitère le procédé : on part de T_1, et on augmente progressivement α (en partant de α_1). On va arriver à un seuil $\alpha_2 > \alpha_1$ pour lequel un sous-arbre de T_1 noté T_2 sera préférable à T_1. T_2 est donc le sous-arbre de T_1 avec le plus petit α possible ;
4. On réitère le procédé jusqu'à arriver à l'arbre ne contenant que la racine.

On obtient donc une séquence d'arbres emboités $T_0 \supset T_1 \supset T_2 \ ... \supset T_{racine}$, en élaguant au fur et à mesure, associée à une séquence de seuils de complexité $0 < \alpha_1 < \alpha_2 \ ... \ \alpha_{racine}$. Les seuils de complexité α sont donnés par le logiciel R dans la colonne CP (Complexity Parameter) du tableau obtenu avec la commande *printcp*.

Au final, pour sélectionner un sous-arbre, l'idée est de se donner un paramètre de complexité α que l'on juge acceptable et de conserver le sous-arbre de coût-complexité minimum associé à ce paramètre. Pour chaque sous-arbre de notre séquence d'élagage $T_0 \supset T_1 \supset T_2 \ ... \supset T_{racine}$, l'idée est d'estimer par validation croisée son taux d'erreur, puis de choisir le seuil α (et donc le sous-arbre) qui minimise cette erreur.

Choix du paramètre de complexité : la procédure de validation croisée

Notre échantillon initial est découpé en V sous-échantillons ($V = 5$ ou 10). A partir de ces V sous-échantillons, on peut obtenir V couples (jeu d'apprentissage, jeu test). Pour chacun de ces V couples, on construit l'arbre maximal sur le jeu d'apprentissage qu'on élague au fur et à mesure comme vu précédemment. Puis on calcule sur le jeu test le coût d'erreur de chacun des arbres emboîtés de la séquence d'élagage. La procédure d'obtention du taux d'erreur d'un arbre T_h de la séquence d'élagage par validation croisée est la suivante :

1. Dans chacune des V séquences d'élagage obtenues par validation croisée, on choisit le sous-arbre le plus proche de T_h en terme de coût-complexité (avec une règle portant sur les seuils de complexité).

2. On a alors V estimations du coût de T_h.
3. En faisant la moyenne de ces V estimations, on obtient une estimation par validation croisée du coût de T_h : $c^{vc}(T_h)$.
4. On obtient un écart-type de $c^{vc}(T_h)$ par : $\sqrt{\frac{c^{vc}(T_h)(1-c^{vc}(T_h))}{n}}$.

R fournit les taux d'erreur obtenus par validation croisée ainsi que leurs écart-type dans les colonnes *xerror* et *xstd* du tableau renvoyé par la commande *printcp*. On peut également utiliser un graphique renvoyé par la commande *plotcp* représentant le taux d'erreur obtenu par validation croisée en fonction du paramètre de complexité.

Le meilleur choix du paramètre de complexité *cp* pour élaguer l'arbre maximal est celui qui est associé au taux d'erreur minimum minimal dans la procédure de validation croisée. Ou encore à partir de la représentation graphique des taux d'erreur en fonction des paramètres de complexité associées, le choix suggéré dans la documentation est la valeur du *cp* la plus à gauche en dessous de la ligne. Une fois que la valeur de cp est choisie, on peut récupérer l'arbre correspondant par la commande *prune* dans R.

4.2 Modèle de régression linéaire pour données fonctionnelles : la méthode FLRTI

4.2.1 La régression linéaire sur une variable explicative fonctionnelle

La statistique fonctionnelle a connu un très important développement ces dernières années. Cette branche de la statistique vise à étudier des données qui, de par leur structure et le fait qu'elles soient collectées sur des grilles de points très fines, peuvent être assimilées à des courbes, en général fonctions du temps et connues sous le nom de données fonctionnelles.

Bien qu'elle ait les mêmes objectifs que les autres branches de la statistique (analyse de données, estimation,...), les données mises en œuvre prennent cependant leurs valeurs dans des espaces de fonctions. Par conséquent, les méthodes usuelles de la statistique multivariée sont ici dès lors mises en défaut, cf. Crambes et al. [4].

Considérons que l'on dispose des observations de n courbes en p points de discrétisation, ces courbes étant utilisées comme prédicteurs d'une autre variable dans un modèle. Si l'on regroupe ces données notées x_{ij} $(i = 1, \ldots, n$ et $j = 1, \ldots, p)$ dans une matrice de taille $n * p$:

$$\mathbf{X} = \begin{pmatrix} x_{11} & \ldots\ldots\ldots & x_{1p} \\ x_{n1} & \ldots\ldots\ldots & x_{np} \end{pmatrix},$$

la méthode des moindres carrés ordinaires (MCO) peut donner de très mauvais résultats puisqu'elle amène à l'inversion de la matrice X^tX qui peut se révéler difficile voire impossible pour deux raisons. La première est que p est généralement grand (parfois même $p > n$) et donc la matrice X^tX devient non inversible. La seconde raison est qu'il y a de fortes chances d'avoir une colinéarité importante entre les p prédicteurs du fait qu'ils sont les points de mesure d'une même fonction. Pour contourner ce problème, des solutions ont été envisagées dont :

- la "ridge regression", cf. Hoerl et Kennard. [8] qui consiste à ajouter un terme de pénalisation dans le critère des moindres carrés. Cela amène à inverser $(X^tX + \lambda I_p)$ au lieu de X^tX avec λ, un réel strictement positif et I_p matrice identité de taille p ;
- la régression sur composantes principales qui consiste à réduire la dimension p en utilisant les k premières composantes principales issues de l'ACP du tableau X ;

– la régression "partial least squares", cf. Helland. [6] qui est une méthode algorithmique basée à chaque étape sur la régression par moindres carrés ordinaires sur les résidus de l'étape précédente.

Il est également possible de définir un modèle de régression linéaire dans le cas où la variable explicative est une variable fonctionnelle. C'est par ailleurs le cadre de ce stage. La courbe X(t) est une variable aléatoire fonctionnelle, c'est-à-dire une variable aléatoire à valeurs dans un espace de fonctions et le modèle de régression linéaire s'écrit comme suit :

$$Y_i = \beta_0 + \int_{t_0}^{t_f} X_i(t)\beta(t)dt + \varepsilon_i \quad i = 1, \ldots, n \tag{4.5}$$

où t_0 et t_f représentent respectivement les instants,de début et de fin, communs à toutes les trajectoires des courbes X_i(t).

L'objectif est d'estimer la fonction de coefficients $\beta(t)$ dans le modèle 4.5 sur la base des observations indépendantes $(X_i(t), Y_i)_{i=1,\ldots,n}$. Les hypothèses faites sur ε_1, ..., ε_n diffèrent suivant les situations envisagées.

Comme pour la régression linéaire standard, les régions où $\beta(t) \neq 0$ correspondent aux périodes de temps où il y a une relation entre $X(t)$ et Y. Alternativement, les régions où $\beta(t)$ =0 correspondent aux régions où $X(t)$ n'a aucun effet sur Y. Si $\beta(t)$ est constant sur une région donnée, alors l'effet de $X(t)$ sur Y demeure constant à l'intérieur de cette région. Finalement, $\beta(t)$ est exactement linéaire dans les régions où l'effet de $X(t)$ sur Y est constant. Implicitement, l'interprétation de la relation prédicteur-réponse devient de plus en plus difficile lorsque la forme de la courbe $\beta(t)$ est instable avec de fortes ondulations. Ainsi, en terme d'interprétation, il est souhaitable pour une méthode d'estimation donnée de pouvoir produire des estimateurs de $\beta(t)$ d'allure simple et qui soient exactement nulles sur les régions où il n'y a pas de relation apparente.

4.2.2 Tests de significativité du modèle de régression linéaire fonctionnelle

Ici, nous présentons brièvement des procédures pour tester la nullité de la fonction de coefficients $\beta(t)$ dans le modèle (4.5). Ce qui est équivalent à tester la significativité globale du modèle. Les hypothèses testées :

$$H_0 \textbf{ : } "\beta = 0" \textbf{ contre } H_1 \textbf{ : } "\beta \neq 0"$$

La documentation sur les tests de nullité de la fonction β dans un modèle de régression linéaire pour données fonctionnelles est plutôt rare. La plupart des procédures de tests proposées dans la littérature sont basées sur des approches déja développées pour estimer β. On peut citer la technique de minimisation d'un critère des moindres carrés pénalisé par un terme λ strictement positif, ou encore l'estimation de β dans un espace de dimension finie engendré par les k premiers axes associés aux k premiers vecteurs propres de l'opérateur de covariance empirique de X.

Nous avons utilisé les procédures introduites par Hilgert et al. [7], basées sur des procédures de test multiple et des projections aléatoires des covariables $X_i(t)$ sous forme d'une ACP fonctionnelle. Ces procédures ont la particularité de ne faire aucun a priori sur la fonction β.

La première statistique de test $T_1(\alpha)$ est obtenue à partir d'une correction de Bonferroni et la deuxième $T_2(\alpha)$, à partir de simulations de Monte-Carlo. Implémentés sous R, d'un point de vue pratique, nous retenons juste que la règle de décision est telle qu'on ne rejette pas H_0 si la statistique de test $T_1(\alpha)$ est négative, et on accepte H_1 si elle est positive. La p-value est fournie par les simulations de Monte-Carlo dans l'évaluation de la statistique $T_2(\alpha)$.

4.2.3 Motivation de la méthode d'estimation FLRTI

Supposons que la variable fonctionnelle $X_i(t)$ a pour support l'intervalle de temps $[0, 1]$. L'estimation de $\beta(t)$ se fait généralement suivant deux approches.

La première approche appelée approche "base" propose de projeter $\beta(t)$ dans une base de fonctions $B(t) = [b_1(t), b_2(t), ..., b_p(t)]$ de dimension p tel que : $\beta(t) = B(t)^T\eta$ où p est assez grand pour capturer les variations de $\beta(t)$ et assez petit pour avoir un ajustement régulier. Le modèle (4.5) peut alors être réécrit comme $Y_i = \beta_0 + \mathcal{X}_i^T\eta + \varepsilon_i$, où $\mathcal{X}_i = \int X_i(t)B(t)dt$, et η pouvant être estimé par la méthode des moindres carrés ordinaires.

La deuxième approche appelée approche "pénalisation" propose une procédure d'estimation par la méthode des moindres carrés avec un critère de pénalisation pour réduire la variabilité dans $\beta(t)$. Un critère standard sur la forme de $\int \beta^{(d)}(t)^2dt$ avec d =2 devient un choix commun. Dans ce cas, on peut trouver le $\beta(t)$ qui minimise : $\sum_{i=1}^{n}(Y_i - \beta_0 - \int X_i(t)\beta(t)dt)^2 + \lambda \int \beta^{(d)}(t)^2dt$, λ>0, où λ est le terme de pénalisation.

Malheureusement, ces approches "base" et "pénalisation" génèrent des courbes d'estimation de $\beta(t)$ de forme complexe et qui ne sont linéaires ou constantes sur aucune région. De plus, ces courbes estimées sont rarement nulles sur des régions de temps. Elles ont plutôt la forme de vaguelettes instables autour de 0.

La méthode FLRTI (Functional Linear Regression That's Interpretable), développée par James et al. [9] produit des courbes d'estimations de $\beta(t)$ beaucoup plus flexibles sur la forme, interprétables, en plus de disposer de bonnes propriétés théoriques de convergence. Cette méthode emprunte des idées aux approches "base" et "pénalisation" sans pour autant leur être similaire.

4.2.4 Présentation de la méthode FLRTI

La méthode est essentiellement basée sur le choix d'une fonction de base particulière et des techniques de sélection de variables. En premier, la période de temps est divisée en une fine grille de points $(t_j)_{j=1,...,p}$ sur $[0, 1]$.

La fonction β est supposée être nulle sur certaines régions temporelles et linéaire sur d'autres : l'emplacement exacte de ces régions étant par ailleurs inconnu. La première hypothèse est justifiée par le fait que les prédicteurs $X_i(t)$ n'ont pas tous le même "poids" pour expliquer la variable réponse Y_i, ceci dès le moment où si $\beta_j = 0$, $X_i(t_j)$ n'a pas d'impact sur Y_i. Ainsi, la fonction β est supposée être parcimonieuse.

La deuxième hypothèse de linéarité est faite afin d'obtenir une fonction β facilement interprétable, ce qui revient à supposer que la dérivée seconde de $\beta(t)$ est nulle dans ces zones de linéarité de β. Cette dérivée seconde est donc supposée être parcimonieuse.

Ces hypothèses apportent ainsi des contraintes sur l'estimation de β dans le modèle de régression linéaire (4.5). Cette situation correspond alors à une régression pénalisée dans des modèles parcimonieux avec un nombre de points de mesures p largement supérieur au nombre d'observations n. Pour estimer la fonction β en chaque point t_j, il est nécessaire de minimiser l'erreur du critère des moindres carrés sous une contrainte de régularité. Le sélecteur de Dantzig est la solution choisie par les auteurs de la méthode FLRTI dans leurs programmes, pour les bons résultats empiriques qu'il produit pour des modèles avec de grandes valeurs de p.

Les fonctions d'ajustement de la méthode FLRTI sont implémentées dans R et disponibles en ligne sur la page web de Gareth JAMES [1]. Il est toutefois nécessaire d'installer le package *lpSolve*.

1. http ://www-bcf.usc.edu/ gareth/research/flrtidoc.pdf

La principale fonction d'ajustement est $flrti$ et en complément, sont également disponibles la fonction $flrti.boot$ qui produit des intervalles de confiance pour $\beta(t)$ par bootstrap et $predict.flrti$ fournissant une prédiction de la réponse Y à partir d'un nouvel ensemble de prédicteurs $X(t)$.

Au final, deux paramètres de réglage doivent être choisis : un terme de pénalisation σ et un poids ω. Le terme de pénalisation fait partie de la procédure du sélecteur de Dantzig. Plus il est grand, plus la contrainte liée à la forme est renforcée. Le poids ω est utilisé pour agir sur le nombre relatif de points égals à 0 dans la fonction β. Un poids de 0 indique que seule l'hypothèse de linéarité est respectée et aucune hypothèse n'a été faite sur la parcimonie de β.

Une procédure de validation croisée a été proposé par les auteurs pour optimiser ces choix de σ et ω. Elle a pour but d'estimer des valeurs optimales de σ et ω à partir de deux ensembles de valeurs possibles pour $(\sigma_k)_k$ et $(\omega_l)_l$. Le principe est de diviser l'échantillon initial en N_f sous-groupes (généralement 10). Tous les sous-groupes, un à la fois, sont utilisés pour effectuer le processus d'estimation avec une combinaison (σ_k, ω_l). Chaque sous-groupe isolé va servir à valider le modèle estimé, fournissant ainsi un taux d'erreur pour chaque combinaison (σ_k, ω_l) utilisée. La procédure est répétée jusqu'à ce que chaque sous-groupe soit utilisé comme échantillon de validation. Au final, nous obtenons N_f taux d'erreur dont la moyenne produit une erreur de validation-croisée associée à chaque (σ_k, ω_l). Le meilleur choix pour σ et ω est alors le couple (σ_k, ω_l) associé au plus petit taux d'erreur obtenu. La fonction $flrti.cv$ dans R effectue cette validation croisée à partir de deux séquences de valeurs possibles pour σ et ω générées par l'utilisateur.

4.2.5 Limites dans le modèle de régression linéaire pour données fonctionnelles

L'ajustement du modèle (4.5) présente certains points limites. Nous ne sommes pas en mesure d'inclure un autre prédicteur qualitatif dans le modèle en plus de la courbe explicative $X(t)$. Par exemple, pour expliquer la qualité du raisin récolté, il serait pertinent de considérer des facteurs qualitatifs comme le cépage de la parcelle, l'année viticole, etc... Mais à ce jour, les techniques d'estimation interprétables dans les modèles de régression pour données fonctionnelles ne sont pas implémentées avec plus d'un facteur explicatif.

4.3 Complémentarité entre les deux méhodes

Les deux méthodes utilisées ont l'avantage d'utiliser la variable explicative fonctionnelle $Ks(t)$ de deux façons possibles. Le modèle de régression linéaire (4.5) traite la courbe $Ks(t)$ dans son ensemble, sans avoir besoin d'en extraire des paramètres représentatifs. Cependant, la limite de cette méthode réside dans le fait que les méthodes existantes à ce jour ne nous permettent pas d'intégrer des facteurs explicatifs supplémentaires dans le modèle. C'est là que les arbres de décision représentent une alternative, car en plus de la courbe $Ks(t)$, résumée en des variables scalaires quantifiant le cumul de vécu hydrique dans des régions d'intérêt, ils peuvent intégrer d'autres facteurs qualitatifs ou quantitatifs comme le cépage ou l'année viticole.

Chapitre 5

Résultats et interprétations

5.1 Présentation des résultats des arbres de régression

Dans cette section, nous présentons les sorties de la mise en œuvre d'arbres de régression pour expliquer deux variables de qualité du raisin en fonction du vécu hydrique de la vigne : le poids des baies et le précurseur d'arôme G3MH. Le G3MH fait partie des facteurs déterminants de la qualité du vin final obtenu en contribuant fortement au goût, à la couleur et à l'astringence. [1]

5.1.1 Le poids des baies

Sur 2012 :

Le tableau de la figure 5.1 est un résumé des résultats de la procédure d'élagage de l'arbre maximal planté. La fonction *rpart* fournit une procédure de validation croisée permettant de choisir le paramètre de complexité *cp* associé au meilleur sous arbre en terme de coût et de complexité. Les paramètres de complexité sont répertoriés dans la colonne CP, le risque estimé de chaque sous-arbre est répertorié dans la colonne $xerror$, et l'écart type de l'estimation du risque est présentée dans la colonne $xstd$. On prend en général la première valeur de CP (i.e. la plus grande) qui est à moins de un écart-type du minimum de $xerror$. Une fois que la valeur de *cp* est choisie, on peut récupérer l'arbre correspondant par la commande *prune*. Nous avons ainsi choisi un *cp* égal à 0.1 et associé à l'erreur minimale.

L'arbre obtenu montre que le poids des baies semble être affecté par le cépage. La Grenache produit les plus gros raisins et on obtient des poids plus faibles sur les autres cépages : Cabernet-Sauvignon, Merlot, Syrah, Chardonnay. Les mêmes constats ont été faits dans l'analyse descriptive de la variable réponse. Ce résultat est d'ailleurs bien connu en viticulture.

1. Ce sentiment de sécheresse, de rudesse et de rugosité en bouche, au moment de la dégustation du vin. Cette impression provient directement de la présence des tanins du vin qui réagissent avec la salive et les protéines buccales.

	CP	nsplit	rel error	xerror	xstd
1	0.71	0.00	1.00	1.08	0.41
2	**0.10**	1.00	0.29	0.71	0.23
3	0.04	2.00	0.19	1.29	0.36
4	0.03	3.00	0.15	1.16	0.29
5	0.03	4.00	0.12	1.15	0.29
6	0.02	5.00	0.09	1.17	0.29
7	0.01	6.00	0.08	1.19	0.29

FIGURE 5.1 – Arbres de régression sur le poids des baies en 2012 et résumé de la procédure d'élagage

Sur 2013 :

	CP	nsplit	rel error	xerror	xstd
1	0.80	0.00	1.00	1.12	0.29
2	**0.06**	1.00	0.20	0.26	0.07
3	0.05	2.00	0.15	0.37	0.09
4	0.03	3.00	0.10	0.41	0.09
5	0.03	4.00	0.07	0.56	0.14
6	0.01	5.00	0.05	0.58	0.16
7	0.01	6.00	0.04	0.59	0.14
8	0.01	7.00	0.02	0.59	0.14
9	0.01	8.00	0.01	0.61	0.15

FIGURE 5.2 – Arbres de régression sur le poids des baies en 2013 et résumé de la procédure d'élagage

D'après la figure 5.2, l'arbre optimal correspond à un *cp* de 0.06. Comme sur 2012, les résultats font ressortir un effet du cépage sur le poids des baies. Les baies semblent être plus grosses sur les parcelles de Grenache que sur celles de Chardonnay et de Merlot comme nous l'avions constaté dans l'analyse descriptive de cette variable. Nous avons également appliqué les arbres de régression sur les données relatives à chaque cépage afin d'étudier l'effet du vécu hydrique par cépage.

Les données ont été agrégées sur 2012/2013 pour avoir un nombre de parcelles assez grand pour pouvoir appliquer la méthode. Au final, nous avons disposé de peu d'individus sur le Cabernet-Sauvignon (2 parcelles), la Syrah (1 parcelle) : un arbre de régression n'aurait pas donné de résultats significatifs sur de tels échantillons. Et sur les parcelles de Merlot, Grenache et Chardonnay, la procédure de validation croisée pour le choix du *cp* n'a pas donné d'arbres significatifs. Les arbres complets sont élagués à la racine. Par conséquent, nous considérons que le vécu hydrique de la vigne n'a aucun impact sur le poids des baies à la récolte et ceci, quelque soit la variété.

5.1.2 Le précurseur d'arôme G3MH

Sur 2012 :

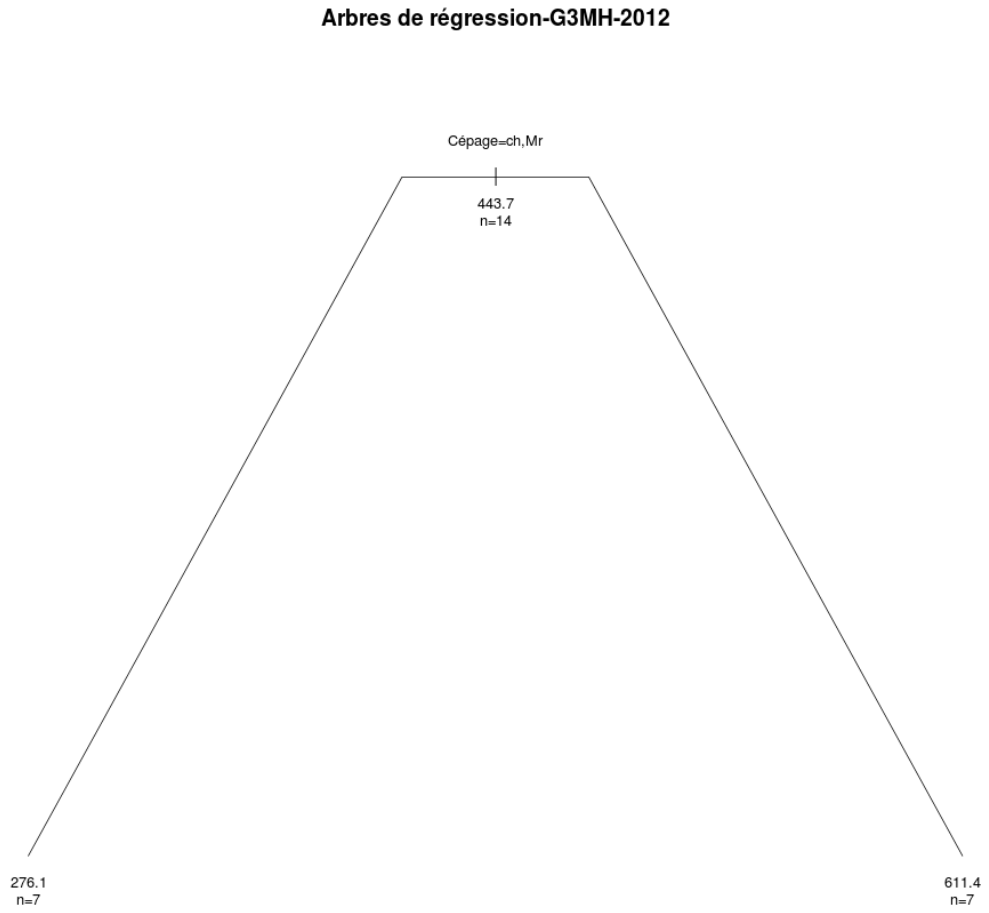
Avec un paramètre de complexité *cp* égal à 0.1, l'arbre optimal correspond au sous-arbre à un seul niveau ($nsplit$=1) associé au taux d'erreur minimal, voir figure 5.3. Le G3MH semble être expliqué par le cépage sur la saison, le vécu hydrique ne ressort pas dans les variables discriminatoires de l'arbre. On observe une teneur dans les raisins plus forte sur les parcelles de Grenache, Cabernet-Sauvignon, Syrah que sur celles de Chardonnay ou le Merlot. Des résultats qui ont été obtenus lors de l'analyse descriptive des données.

Sur 2013 :

Nous choisissons un paramètre de complexité égal à 0.02 et associé au sous-arbre de 4 splits. Ce paramètre est préféré au *cp* de 0.01 qui est associé à l'erreur minimale car il fournit un sous-arbre avec un plus petit nombre de noeuds et associé à un taux d'erreur proche de l'erreur minimale.

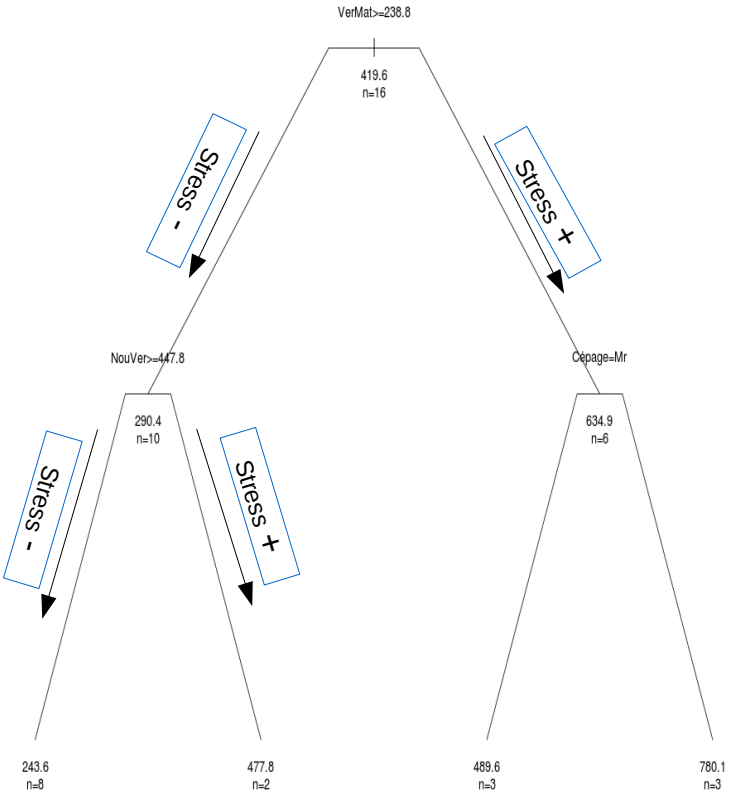
Le G3MH semble être expliqué par le vécu hydrique sur la période post-véraison, autrement dit entre la véraison et la maturité. Comme le montre le graphique 5.4, sa teneur est plus forte en situation de stress hydrique élevé sur cette période avec des valeurs supérieures à 635 notamment sur les parcelles de Grenache et Chardonnay. L'arbre de régression montre également un effet du vécu hydrique sur la période pré-véraison, de la nouaison à la véraison. Si la vigne est beaucoup stressée dans cette période, le G3MH est fortement présent dans les raisins avec des valeurs supérieures à 300 et inférieures dans la situation inverse.

En résumé, sur 2013, le vécu hydrique et le cépage sont des facteurs d'influence de la teneur en G3MH dans les raisins. Plus la vigne est stressée aussi bien sur la pré-véraison que sur la post-véraison, plus il y a de G3MH dans les baies . Les raisins récoltés sur les parcelles de Grenache et Chardonnay ont toutefois une teneur en G3MH encore plus importante que ceux du Merlot.



	CP	nsplit	rel error	xerror	xstd
1	0.69	0.00	1.00	1.16	0.37
2	**0.10**	1.00	0.31	0.43	0.15
3	0.06	3.00	0.10	0.66	0.16
4	0.01	4.00	0.04	0.64	0.12
5	0.01	5.00	0.03	0.77	0.17

FIGURE 5.3 – Arbres de régression sur le G3MH en 2012 et résumé de la procédure d'élagage



	CP	nsplit	rel error	xerror	xstd
1	0.56	0.00	1.00	1.18	0.40
2	0.16	1.00	0.44	1.45	0.57
3	0.11	2.00	0.28	1.45	0.57
4	0.09	3.00	0.17	1.06	0.30
5	**0.02**	4.00	0.08	0.89	0.30
6	0.02	5.00	0.06	0.90	0.29
7	0.01	6.00	0.04	0.86	0.29

FIGURE 5.4 – Arbres de régression sur le G3MH en 2013 et résumé de la procédure d'élagage

5.2 Mise en œuvre et présentation des sorties de la régression linéaire sur la variable fonctionnelle de vécu hydrique

5.2.1 Procédure de distorsion des courbes de vécu hydrique

Cette procédure est une étape préliminaire à l'ajustement d'un modèle de régression linéaire sur la variable fonctionnelle Ks(t). L'objectif est de recaler toutes les courbes Ks(t) estimées sur le même espace de temps phénologique. En effet, il s'agit de les définir le long de la saison de culture, à partir de la même date d'atteinte de nouaison, sur une même date de véraison et en dernier sur une date commune de maturité ou de récolte si la maturité n'a pas été atteinte sur le site. Le but étant, d'une part, d'avoir le même nombre de points sur toutes les courbes Ks(t). Et d'autre part, en terme d'interprétation de la fonction de coefficients $\beta(t)$ à estimer, nous voulons avoir un espace de temps phénologique fixé et commun à toutes les courbes de vécu hydrique pendant la saison. Ceci va nous permettre de définir de facon générale les périodes de la saison où le vécu hydrique a de l'influence sur les variables de qualité du raisin dans une parcelle donnée.

Notons U(t) l'axe de référence choisi et sur lequel sont définies les dates phénologiques de référence U_0, U_1, U_2,...,U_f. Notre but est de faire correspondre chacune des dates de stades phénologiques T_{i0}, T_{i1}, T_{i2},...,T_{if} aux dates phénologiques de référence U_0, U_1, U_2,...,U_f.
Les $(T_{ij})_{j=1,\dots,f}$, étant les dates phénologiques propres à chaque courbe de vécu hydrique Ks_i(t) de la i-ème parcelle.

L'axe $U(t)$ de référence peut être choisi aléatoirement parmi les espaces de temps phénologique de la 1ère, 2ème,... ou n-ième parcelle expérimentale. Rappelons que l'on dispose des courbes de vécu hydrique sur 15 parcelles suivies en 2012 et sur 18 parcelles en 2013.

On suppose que la relation existant entre les deux axes U et T_i est de la forme :

$$T_i = \phi_i(U) + \delta$$

Nous sommes alors en présence d'un modèle de régression non-paramétrique, les fonctions ϕ_i étant inconnues et δ la marge d'erreur accordée. Nous choisissons de les estimer dans R par interpolation spline avec la fonction $splinefun$, de même que leurs dérivées ϕ_i'.

A partir de la relation entre les deux axes, le changement de variable $t = \widehat{\phi}_i(u)$ est fait de sorte à garder la même aire sous la nouvelle courbe de vécu hydrique et donc :

$$\int_{T_{i0}}^{T_{if}} Ks_i(t)dt = \int_{U_0}^{U_f} Ks_i(\widehat{\phi}_i(u))\widehat{\phi}_i'(u)du \tag{5.1}$$

On note $\tilde{Ks}_i(u) = Ks_i(\widehat{\phi}_i(u))\widehat{\phi}_i'(u)$ et ainsi, nous avons le nouveau jeu de données $(\tilde{Ks}_i, Y_i)_{i=1,\dots,n}$ avec lequel nous pouvons réécrire le modèle 4.5 comme suit :

$$Y_i = \beta_0 + \int_{U_0}^{U_f} \tilde{Ks}_i(u)\beta(u)du + \varepsilon_i\,, \quad i = 1, \dots, n \tag{5.2}$$

où les $\tilde{Ks}_i(u)$ sont les nouvelles courbes de vécu hydrique "distordues" sur l'axe phénologique de référence choisie, et obtenues à partir d'une interpolation spline de ϕ et ϕ'.

A titre d'illustration, nous présentons ici les résultats en 2013 du recalage de la courbe de vécu hydrique (en noire) du site Saint Sauveur non irrigué planté avec du Chardonnay (StSAU-Char-i0), sur le temps phénologique du site de LaBaume non irrigué planté avec du Cabernet-Sauvignon (LB-CS-i0). Ce dernier a été pris comme axe phénologique de référence pour le recalage

de toutes les courbes de vécu hydrique. La courbe rouge est celle obtenue après recalage sur l'axe phénologique de référence. La nouaison, la véraison et la maturité étant les stades phénologiques d'intérêt sur la saison, représentées par les barres rouges, bleues et vertes.

Si la maturité n'a pas été atteinte sur une parcelle, la récolte est alors considérée comme dernière étape dans le processus de développement de la vigne. Avant la nouaison, peu de données sont recueillies notamment entre le débourrement et la floraison. La première région de la courbe noire à recaler (entre 170 et 430 degrés jours) n'a donc pas été prise en compte dans le recalage car se situant avant la nouaison qui est enregistrée à partir de 430 degrés-jours.

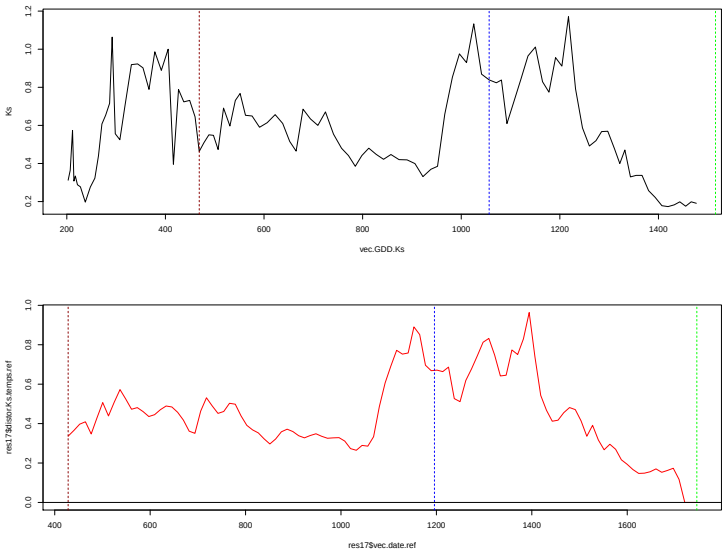


FIGURE 5.5 – Recalage de la courbe de vécu hydrique Ks de la parcelle StSAU-Char-i0 sur le temps phénologique de la parcelle LB-CS-i0, de la Nouaison à la Maturité.

Un moyen de validation de cette technique de distorsion de courbe $Ks(t)$ a été de s'assurer que les aires en dessous de la courbe recalée sont égales aux aires sous la courbe initiale de vécu hydrique, sur les périodes délimitées par les stades phénologiques considérées. Les aires ont été calculées de la même facon que les scalaires de cumul du vécu hydrique, section 3.3, en appliquant une intégration par trapèze sous les courbes dans les régions en question.

5.2.2 Mise en oeuvre des tests de significativité des modèles de régression linéaire sur des données fonctionnelles

Avant la mise en œuvre de la méthode FLRTI pour l'estimation de la fonction de coefficients $\beta(t)$ dans le modèle (5.2), nous nous sommes intéressés à la significativité globale de ce modèle. Ceci en appliquant les procédures de tests de nullité de la fonction de coefficients $\beta(t)$.

Les courbes $Ks(t)$ étant recalées sur le même espace de temps phénologique avec un pas de temps donné, on dispose sur chaque saison de 111 points de mesures du vécu hydrique obtenus sur 15 (resp. 18) parcelles en 2012 (resp. en 2013). Les données de qualité du fruit et des précurseurs d'arômes dans les raisins sont disponibles sur ces mêmes parcelles aux dates de récolte.

Par ailleurs, pour reproduire la même démarche que pour les arbres de régression, on a appliqué les tests sur uniquement les données des parcelles de Merlot agrégées sur 2012 et 2013. Ceci toujours dans le but d'étudier un éventuel effet du vécu hydrique par cépage.

Nous avons également séparé les mesures de $Ks(t)$ sur toute la saison en deux ensembles : sur la pré-véraison et sur la post-véraison afin de tester l'impact du vécu hydrique sur chacune ces deux phases phénologiques majeurs de la vigne. Nous disposons ainsi sur chaque parcelle de 65 points de mesures du vécu hydrique sur la pré-véraison et de 46 points à la post-véraison.

Au final, la méthode d'estimation FLRTI est appliquée sur les modèles de régression linéaire où l'effet de la courbe $Ks(t)$ est significatif à un seuil de 5%. Cependant, il s'agit de la significativité d'une relation linéaire entre le vécu hydrique et la variable de qualité en question. Par conséquent, une statistique de test négative même associée à une p-value très faible peut être due au fait que la relation est non linéaire ou que l'on ne dispose pas de suffisamment d'individus pour conclure.

En 2012

Variables	Saison entière p = 110	Pré-véraison p=65	Post-véraison p=45
Sucre n = 15	$T_2(\alpha)$ = -0.6 p.value = 0.06	$T_2(\alpha)$ **= 5.67**	$T_2(\alpha)$ = -6.4 p.value = 0.74
Poids des Baies n = 15	$T_2(\alpha)$ = -5.83 p.value = 0.66	$T_2(\alpha)$ = -6.33 p.value = 0.33	$T_2(\alpha)$ = -6.31 p.value = 0.72
Azote Assimilable n = 15	$T_2(\alpha)$ = -4.83 p.value = 0.11	$T_2(\alpha)$ = -2.08 p.value = 0.09	$T_2(\alpha)$ = -6.67 p.value = 0.18
PDMS n = 15	$T_2(\alpha)$ = -6.41 p.value = 0.33	$T_2(\alpha)$ = -6 p.value = 0.17	$T_2(\alpha)$ = -6.54 p.value = 0.95
cys3MH n = 15	$T_2(\alpha)$ = -1.82 p.value = 0.08	$T_2(\alpha)$ = -6.5 p.value = 0.17	$T_2(\alpha)$ **= 3.37**
G3MH n = 15	$T_2(\alpha)$ = -6.38 p.value = 0.32	$T_2(\alpha)$ = -6.19 p.value = 0.4	$T_2(\alpha)$ = -6.7 p.value = 0.91
GSH n = 15	$T_2(\alpha)$ = -5.63 p.value = 0.19	$T_2(\alpha)$ = -3.57 p.value = 0.2	$T_2(\alpha)$ = -2.5 p.value = 0.09

TABLE 5.1 –
Sorties des tests de significativité du modèle (5.2) pour expliquer les variables de qualité du raisin sur 2012 en fonction des courbes de vécu hydrique sur 3 périodes : la saison entière, la pré-véraison et la post-véraison.

– Au vu des statistiques de test $T_2(\alpha)$ négatives sur 2012, l'hypothèse d'une relation linéaire entre le vécu hydrique mesuré sur toute la saison et les variables de qualité du raisin est rejetée.

– Sur la pré-véraison, l'hypothèse H_1 est acceptée sur le modèle expliquant la concentration en sucre dans les baies. Il semblerait que le vécu hydrique présente une relation linéaire avec la concentration en sucre. La méthode FLRTI peut alors être appliquée pour localiser les régions de la pré-véraison où $\beta(t) \neq 0$, autrement dit, les régions exactes de l'influence du vécu hydrique sur la concentration en sucre.

– De la véraison à la maturité, le test sur le modèle linéaire avec le cys3MH est significatif. Par ailleurs, le cumul de vécu hydrique dans cette période est ressortie comme variable discriminatoire dans l'arbre de régression appliqué sur cette variable (Annexe D) avec une teneur en cys3MH élévée dans les raisins en situation de stress élevé.

En 2013

– Les statistiques de tests sont positives sur les deux modèles de régression expliquant le poids des baies et le G3MH avec comme prédicteurs les courbes $Ks(t)$ mesurées sur toute la saison.

Variables	Saison entière p = 110	Pré-véraison p = 65	Post-véraison p = 45
Sucre n=18	$T_2(\alpha)$ = -5.69 p.value = 0.58	$T_2(\alpha)$ = -5.94 p.value = 0.56	$T_2(\alpha)$ = -6.54 p.value = 0.63
Poids des Baies n=18	$T_2(\alpha)$ **= 1.03**	$T_2(\alpha)$ **= 7.17**	$T_2(\alpha)$ **= 5.65**
Azote Assimilable n=18	$T_2(\alpha)$ = 2.65 p.value = 0.05	$T_2(\alpha)$ = -6.33 p.value = 0.2	$T_2(\alpha)$ = -6.9 p.value = 0.47
PDMS n=18	$T_2(\alpha)$ = -3.35 p.value = 0.16	$T_2(\alpha)$ = -0.91 p.value = 0.07	$T_2(\alpha)$ **= 1.98**
cys3MH n=18	$T_2(\alpha)$ = -1.35 p.value = 0.07	$T_2(\alpha)$ **= 0.1**	$T_2(\alpha)$ = -1.18 p.value = 0.07
G3MH n=18	$T_2(\alpha)$ **= 9.65**	$T_2(\alpha)$ **= 14.48**	$T_2(\alpha)$ **= 6.9**
GSH n=18	$T_2(\alpha)$ = -6.05 p.value = 0.79	$T_2(\alpha)$ = -6.35 p.value = 0.82	$T_2(\alpha)$ = -4.65 p.value = 0.36

TABLE 5.2 –
Sorties des tests de significativité du modèle (5.2) pour expliquer les variables de qualité du raisin sur 2013 en fonction des courbes de vécu hydrique sur 3 périodes : la saison entière, la pré-véraison et la post-véraison

Ces deux variables évoluent de façon linéaire avec le vécu hydrique et l'arbre obtenu sur le G3MH à la figure 5.4 montre un impact de la courbe $Ks(t)$ dans la période post-véraison.

– Entre nouaison et véraison, ces deux mêmes modèles de régression linéaire sont significatifs.

– Entre véraison et maturité, on retrouve une significativité des modèles sur ces 2 variables mais aussi sur le PDMS. Sur l'arbre de régression obtenu sur le PDMS, annexe E, le vécu hydrique ressort comme première variable discriminatoire sur la période post-véraison et nous voyons que le PDMS évolue en sens inverse du stress de la vigne. Il semble alors qu'il est positivement corrélé au vécu hydrique entre la véraison et la maturité.

5.2.3 Présentation des résultats de la méthode FLRTI

Sur le G3MH

Ici, nous appliquons la méthode FLRTI pour estimer la fonction de coefficients $\beta(t)$ dans le modèle de régression (5.2) expliquant le G3MH par le vécu hydrique en 2013. Nous avons vu que ce modèle est significatif avec une courbe de vécu hydrique $Ks(t)$ mesurée sur toute la saison et la post-véraison. Nous verrons dans la suite dans quels intervalles de temps le vécu hydrique impacte sur ce précurseur d'arôme.

Modèle de régression avec un vécu hydrique mesuré sur toute la saison

σ	ω	$\widehat{\beta}_0$	R^2
0.7	0.7	749.0097	0.92

TABLE 5.3 – Estimation des paramètres de l'ajustement du modèle (5.2) sur le G3MH

Les zones d'influence sont les intervalles de temps où $\beta \neq 0$. Le graphique 5.6 ci-après représente la courbe $\widehat{\beta}$ estimée, sur laquelle se base toute l'analyse de l'influence de la courbe $Ks(t)$.

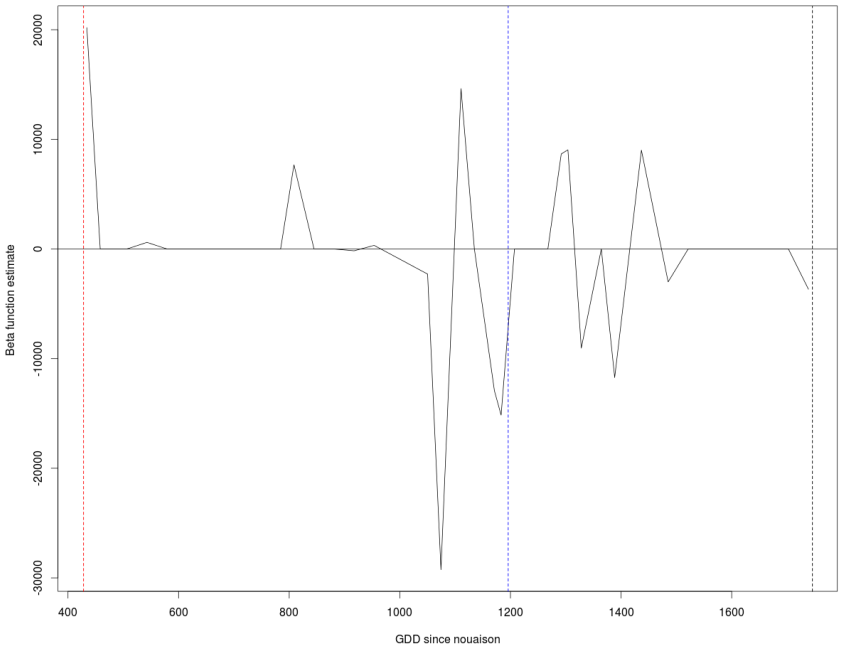


FIGURE 5.6 – Estimation de β dans le modèle expliquant le G3MH avec le vécu hydrique mesuré sur toute la saison 2013.

Les sorties du modèle de régression linéaire sont mentionnées dans le tableau 5.3 pour un couple optimal de paramètres (σ,ω) égal à (0.7,0.7). Ce couple de paramètres est obtenu par validation croisée (cf. section 4.2.4) à partir d'une fine grille de valeurs possibles comprises entre 0 et 1, que nous avons générée. Le R^2 estimé du modèle est égal à 0.92. La valeur de $\widehat{\beta}_0$ est à comparer avec la moyenne $\overline{Y}$ (419.61), ce qui donne un poids non négligeable à l'intégrale.

Sur la période de pré-véraison, délimitée par la nouaison (en rouge) et la véraison (en bleue), le vécu hydrique n'a pas d'influence sur le G3MH avant 700 GDD, compte tenu des périodes où la fonction $\widehat{\beta}$ est nulle. Sur de petites régions entre 790 et 840 GDD, ou entre 1100 GDD et 1150 GDD, on note toutefois une influence positive. L'influence négative est cependant plus marquée avec le pic élevé de la courbe située entre 880 et 1100 GDD. Cette influence n'est pas ressortie dans les résultats des arbres.

Sur la post-véraison, on note beaucoup plus de régions d'influence du vécu hydrique sur le G3MH, ceci de 1270 GDD à 1520 GDD. Au delà, le vécu hydrique n'a plus d'impact sur le précurseur d'arôme. L'influence négative est cependant plus marquée sur la période notamment entre 1320 et 1400 GDD et de 1500 à 1520 GDD, ce qui correspond à une relation linéaire entre les 2 variables sur environ 25 jours, sachant qu'on a 46 jours de la véraison à la maturité sur le site de LaBaume pris comme axe de référence phénologique pour recaler les courbes de vécu hydrique. Nous décidons donc d'approfondir l'analyse sur cette période post-véraison où l'impact du vécu hydrique sur le G3MH semble être plus fort que sur la pré-véraison.

Modèle de régression avec le vécu hydrique sur la post-véraison

σ	ω	$\widehat{\beta}_0$	R^2
0.5	0.6	856.5619	0.91

TABLE 5.4 – Estimation des paramètres du modèle (5.2) sur le G3MH avec un $Ks(t)$ post-véraison

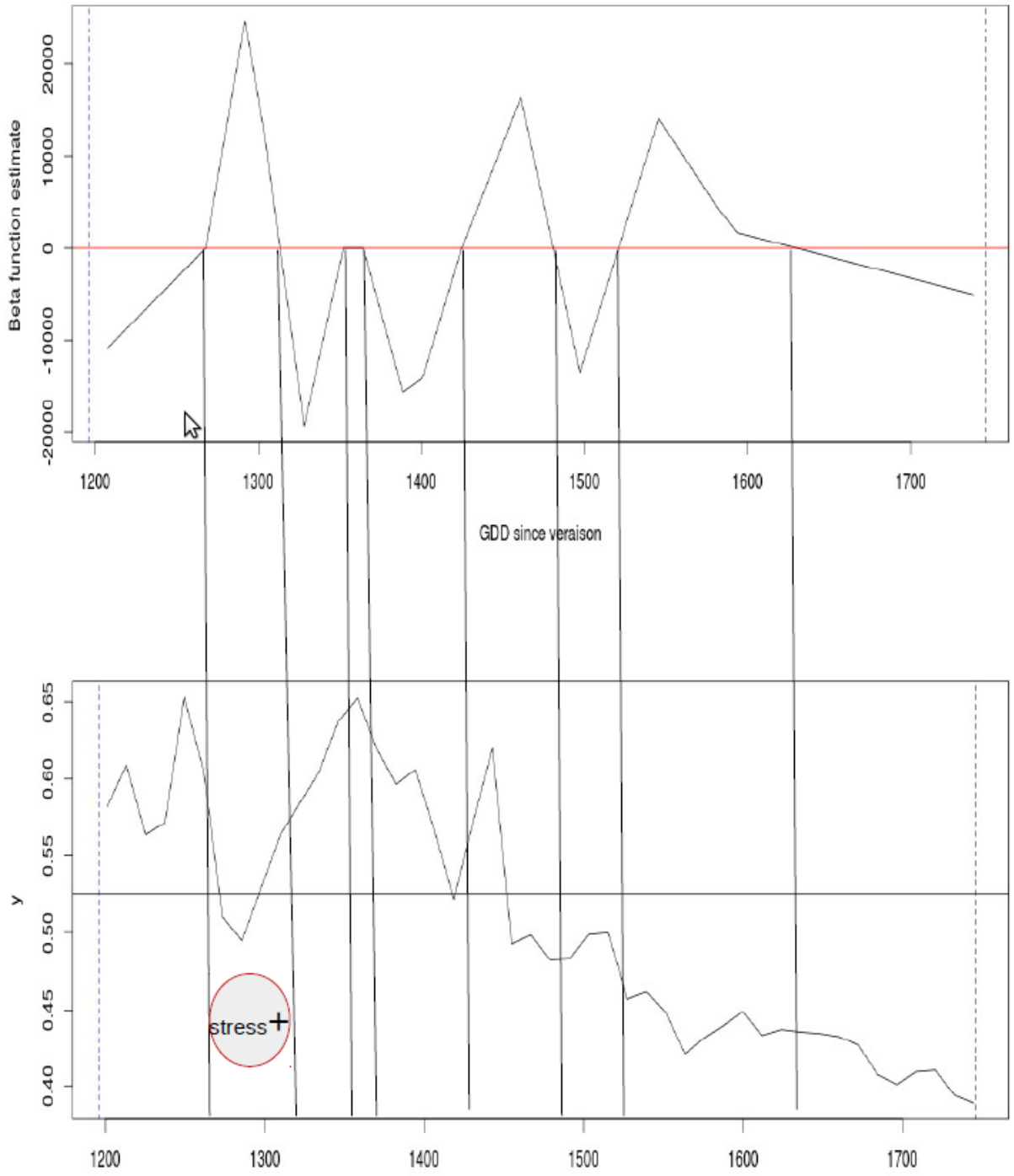


FIGURE 5.7 – Estimation de β dans le modèle expliquant le G3MH avec le vécu hydrique post-véraison.

Le 1^{er} graphique de la figure 5.7 représente l'estimation de β dans le modèle et le 2^e est une représentation de la courbe moyenne de vécu hydrique obtenue à partir des 16 courbes $Ks(t)$ mesurées sur la post-véraison. Les deux graphiques sont mis en regard afin de voir simultanément la forme de $\widehat{\beta}(t)$ par rapport à l'évolution du vécu hydrique moyen. On peut déjà voir que la courbe $\widehat{\beta}(t)$ se présente sous la même forme que celle du graphique 5.6 à la post-véraison. C'est donc en quelque sorte un zoom qui nous permet de mieux analyser l'impact du vécu hydrique sur le G3MH dans la période post-véraison.

Comme mentionné plus haut, l'influence du vécu hydrique est plus ressentie sur la post-véraison. Sur les régions où le vécu hydrique a un effet négatif sur le G3MH (juste après véraison, de 1310 à 1340 GDD ou de 1355 à 1425 GDD), le vécu hydrique moyen sur la saison prend des valeurs élevées en référence à la moyenne de ses points représentée par la ligne horizontale noire. A l'opposé, sur les régions où l'influence est positive, la courbe moyenne prend des valeurs faibles. Par conséquent, il semble que le G3MH évolue en sens inverse du vécu hydrique post-véraison autrement dit, il est positivement corrélé au stress de la vigne.

Ce résultat est par ailleurs obtenu sur l'arbre de régression expliquant le G3MH en 2013 dans la section 5.2. Le cumul de vécu hydrique sur la post-véraison est ressorti comme $1^{ère}$ variable discriminatoire de l'arbre et plus la vigne est stressée, plus la valeur du G3MH augmente.

Le poids des baies

Nous avons vu que la relation linéaire entre le poids des baies et le vécu hydrique sur toute la saison en 2013 est également significatif. La méthode FLRTI peut dans ce cas être utilisé pour déterminer les périodes exactes de la saison où le vécu hydrique impacte le poids des baies. Toutefois, l'arbre de régression obtenu avec cette variable montre uniquement un effet du cépage. La méthode FLRTI ne pouvant intégrer de facteur qualitatif dans le modèle, le cépage ne pourra donc pas être pris en compte. Ajuster le modèle (5.2) malgré cela pourrait alors constituer une importante source d'erreur. L'effet du cépage serait alors dans les résidus du modèle.

Le cys3MH

Les périodes d'influence du vécu hydrique sur cette variable sont nombreuses sur la période post-véraison en 2012, voir Annexe F. Nous observons une influence négative du vécu hydrique de 1300 GDD à 1380 GDD, puis de 1490 à 1570 GDD. Cette influence est encore plus apparente entre 1610 et 1660 GDD avec ce pic négatif de grande amplitude.

L'influence positive est également marquée en milieu de saison et surtout à quelques jours d'atteinte de la maturité, au delà de 1660 GDD. Comparé au vécu hydrique moyen (sur le 2e graphique), nous voyons que le stress de la vigne était croissant sur cette période. Le même constat est fait au niveau de la première région d'influence négative du vécu hydrique (1300 GDD-1380 GDD) où nous notons un stress de plus en plus faible de la vigne (la valeur du vécu hydrique est supérieure à la moyenne).

Compte tenu de ces analyses, il semble que le cys3MH évolue dans le sens du stress hydrique de la vigne pendant la post-véraison . Autrement dit, sa teneur est plus importante dans les raisins en cas de stress élevé. Ce résultat est par ailleurs confirmé avec l'arbre de régression expliquant ce précurseur d'arôme en 2012, annexe D.

Chapitre 6

Conclusions et perspectives

Dans ce stage, nous avons eu pour objectif de mesurer l'impact du vécu hydrique de la vigne pendant la saison sur la qualité des raisins à la récolte. Les résultats présentés aux différents partenaires dont NYSEOS et FRUITION SCIENCES vont dans le sens des hypothèses pressenties.

La plus grande partie de mon travail a été consacrée à l'estimation du vécu hydrique des vignes, variable de type courbe. Cela a été une phase difficile mettant en œuvre des données et règles agronomiques complexes mais surtout ayant nécessité une longue étape de programmation structurée pour tenir compte de toutes ces règles. C'était aussi une étape où il a fallu interagir continuellement avec les agronomes et autres experts du domaine auxquels revenait, entre autres, le choix final de la date d'atteinte du coefficient cultural $Kc_B(t)$.

Nous avons pu traiter ensuite la principale variable explicative, la courbe $Ks(t)$, sous deux angles différents. Dans un premier temps, elle a servi à extraire des variables scalaires représentatives du cumul de vécu hydrique dans des périodes phénologiques majeures et dans un deuxième temps, elle a été considérée comme une variable fonctionnelle évaluée sur toute ou une partie de la saison. Les méthodes d'analyses utilisées sont très complémentaires car les arbres de décision prennent le pas sur la méthode FLRTI qui ne peut prendre en compte plus d'un facteur qualitatif. De plus, nous nous sommes rendus compte que le traitement d'irrigation n'a pas servi à séparer le régime sec du régime irrigué sur les parcelles à cause des précipitations. En terme d'interprétation du vécu hydrique, l'irrigation ne peut donc pas être considéré comme un facteur explicatif. Au final, nous avons observé une influence du vécu hydrique sur des indicateurs de qualité du raisin à la maturité comme le poids des baies, le sucre dans la période pré-véraison, les précurseurs d'arôme comme le PDMS, le cys3MH pendant la post-véraison et le G3MH sur toute la saison. Le cépage reste l'autre facteur explicatif de la qualité des baies.

En terme de perspectives, la même étude pourrait être menée sur un cépage donné afin d'étudier l'impact du vécu hydrique sur les raisins par variété. Il serait également bénéfique d'arriver à isoler les états de consommation en eau des vignes (régime sec, régime irrigué, régime très irrigué) afin de mieux analyser l'influence du vécu hydrique sur la qualité des raisins.

Sur un plan personnel, ce stage m'a permis de perfectionner mes compétences techniques en programmation statistique avec le langage *R* que je n'avais jamais utilisé auparavant pour interagir avec un système d'information. Mais également, le projet PILOTYPE prenant fin en Décembre 2014, j'ai du respecter des délais imposés par le cahier de charges, ce qui m'a forcé à mieux m'organiser et optimiser mon temps de travail pour pouvoir rendre les résultats à temps.

Bibliographie

[1] Allen, R. G., Pereira, L. S., Raes, D., Smith, M., et al. Crop evapotranspiration-guidelines for computing crop water requirements. *Irrigation and drainage paper 56* (1998), 377–384.

[2] Ameri, K., and Harris, A. L. Activating transcription factor 4. *The International Journal of Biochemistry & Cell Biology 40*, 1 (2008), 14–21.

[3] Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. *Classification and regression trees*. CRC press, 1984.

[4] Crambes, C., Kneip, A., and Sarda, P. Smoothing splines estimators for functional linear regression. *The Annals of Statistics 37*, 1 (2009), 35–72.

[5] Ferreira, M. I., Silvestre, J., Conceição, N., and Malheiro, A. C. Crop and stress coefficients in rainfed and deficit irrigation vineyards using sap flow techniques. *Irrigation Science 30*, 5 (2012), 433–447.

[6] Helland, I. S. Partial least squares regression and statistical models. *Scandinavian Journal of Statistics 17* (1990), 97–114.

[7] Hilgert, N., Mas, A., Verzelen, N., et al. Minimax adaptive tests for the functional linear model. *The Annals of Statistics 41*, 2 (2013), 838–869.

[8] Hoerl, A. E., and Kennard, R. W. Ridge regression—1980 : Advances, algorithms, and applications. *American Journal of Mathematical and Management Sciences 1*, 1 (1981), 5–83.

[9] James, G. M., Wang, J., and Zhu, J. Functional linear regression that's interpretable. *The Annals of Statistics 37*, 5 (2009), 2083–2108.

[10] Koundouras, S., Tsialtas, I. T., Zioziou, E., and Nikolaou, N. Rootstock effects on the adaptive strategies of grapevine under contrasting water status : Leaf physiological and structural responses. *Agriculture, Ecosystems & Environment 128*, 1 (2008), 86–96.

[11] Thébaut, A., Scholash, T., Charnomordic, B., and Hilgert, N. A modeling approach to design a software sensor and analyze agronomical features-application to sap flow and grape quality relationship. *arXiv preprint arXiv :1309.5316* (2013).

Annexe A

Mesures des variables de qualité du raisin sur les parcelles (2012)

	Sucre	Nass	PdBaies	PDMS	cys3MH	G3MH	GSH	Cépage
OUV-Mer-i0	238.20	68	1.21	621.71	102.85	166.63	0.91	Merlot
StSAU-Char-i1	234.70	226	1.20	128.50	2.61	206.70	0.90	Chardonnay
OUV-Mer-i2	251.70	89	1.40	797.95	99.04	236.67	0.74	Merlot
PR-Mer-i0	217.17	241	1.20	1516.31	20.89	238.52	2.65	Merlot
StSAU-Char-i0	225.30	245	1.22	239.00	2.86	312.57	2.27	Chardonnay
StGER-Mer-i0	218.30	133	0.99	791.27	174.70	381.25	1.19	Merlot
PR-Mer-i1	223.30	182	1.39	1098.19	17.23	390.11	3.16	Merlot
LB-Sy-AgeJ	221.70	106	1.43	158.14	13.14	461.94	6.39	Syrah
LB-CS-i1	243.00	91	1.27	514.18	35.08	504.29	3.32	Cabernet-Sauvignon
LB-CS-i0	235.80	80	1.40	467.86	43.45	552.33	1.85	Cabernet-Sauvignon
PIO-Gre-i0	227.60	151	1.66	268.14	107.65	553.28	6.71	Grenache
PIO-Gre-i1	247.70	77	2.03	91.91	151.02	585.72	5.58	Grenache
RIE-Gre-i1-Chm	219.50	183	1.69	64.61	93.38	751.27	1.46	Grenache
RIE-Gre-i1	217.20	204	1.64	114.55	167.30	871.01	1.25	Grenache

Annexe B

Mesures des variables de qualité du raisin sur les parcelles (2013)

	Sucre	Nass	PdBaies	PDMS	C3MH	G3MH	GSH	Cépage
StGER-Mer-i0-Chm	237.60	140	1.41	158.85	50.75	569.09	5.75	Merlot
RIE-Gre-i1-Chm	222.20	249	1.77	66.62	168.41	881.58	4.21	Grenache
StGER-Mer-i1	244.00	102	1.40	499.02	106.89	506.53	1.87	Merlot
PIO-Gre-i0	255.80	136	2.00	54.81	165.44	895.12	11.16	Grenache
StSAU-Char-i0	222.90	169	1.39	46.84	13.29	398.83	5.44	Chardonnay
OUV-Mer-i2	228.90	36	1.42	856.82	37.52	237.38	4.96	Merlot
OUV-Mer-i1	242.40	60	1.29	764.82	127.17	367.99	4.31	Merlot
StSAU-Char-i1	225.20	154	1.61	74.26	7.24	263.69	4.25	Chardonnay
PR-Mer-i0	226.40	209	1.19	557.80	56.08	268.96	5.93	Merlot
PIO-Gre-i1	242.30	116	2.02	28.99	88.29	556.69	7.94	Grenache
OUV-Char-i1	212.10	254	1.43	81.39	6.65	177.16	17.62	Chardonnay
PR-Mer-i1	214.80	185	1.56	604.32	49.28	258.15	9.28	Merlot
OUV-Mer-i0	228.90	70	1.12	649.75	82.13	254.35	1.21	Merlot
PR-Char-i0	228.70	135	1.34	71.20	3.62	121.29	7.11	Chardonnay
RIE-Gre-i0	212.10	213	1.83	58.69	122.04	563.71	5.20	Grenache
LB-Mer-i0	215.40	146	1.46	485.16	54.08	393.28	2.64	Merlot

Annexe C

Corrélations entre les variables de qualité du raisin sur 2012 et 2013

	Sucre	Nass	PdBaies	PDMS	cys3MH	G3MH	GSH
Sucre	1.00	**-0.65**	0.20	-0.12	0.04	-0.27	-0.01
Nass	-0.65	1.00	-0.20	0.10	-0.33	-0.02	-0.19
PdBaies	0.20	-0.20	1.00	-0.51	0.37	**0.68**	0.49
PDMS	-0.12	0.10	-0.51	1.00	-0.19	-0.53	-0.21
cys3MH	0.04	-0.33	0.37	-0.19	1.00	0.44	-0.09
G3MH	-0.27	-0.02	0.68	-0.53	0.44	1.00	0.24
GSH	-0.01	-0.19	0.49	-0.21	-0.09	0.24	1.00

TABLE C.1 – Corrélations entre les variables descriptives de la qualité du raisin en 2012

	Sucre	Nass	PdBaies	PDMS	C3MH	G3MH	GSH
Sucre	1.00	**-0.58**	0.19	0.04	0.45	0.43	-0.11
Nass	-0.58	1.00	0.25	-0.58	-0.02	0.18	0.47
PdBaies	0.19	0.25	1.00	-0.59	0.47	**0.69**	0.30
PDMS	0.04	-0.58	-0.59	1.00	0.01	-0.40	-0.39
cys3MH	0.45	-0.02	0.47	0.01	1.00	0.82	-0.20
G3MH	0.43	0.18	0.69	-0.40	0.82	1.00	-0.06
GSH	-0.11	0.47	0.30	-0.39	-0.20	-0.06	1.00

TABLE C.2 – Corrélations entre les variables descriptives de la qualité du raisin en 2013

Annexe D

Arbre de régression sur le cys3MH (2012)

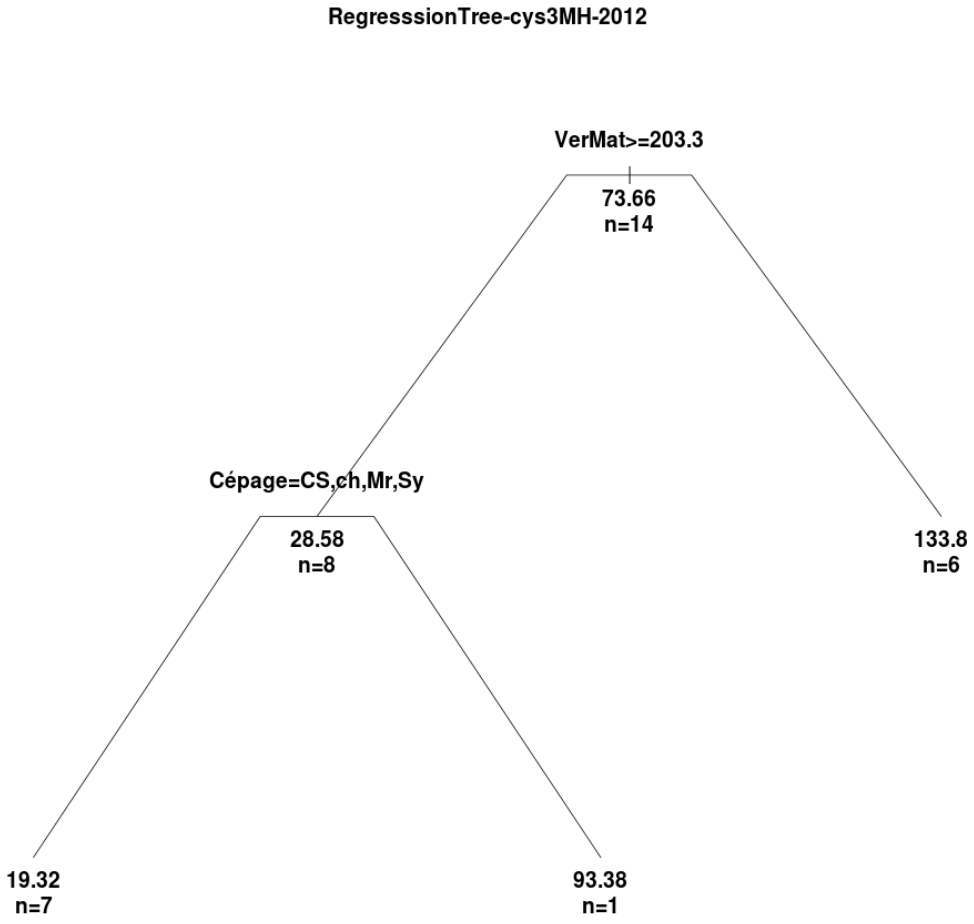


FIGURE D.1 – Arbre de régression sur le cys3MH en 2012

Annexe E

Arbre de régression sur le PDMS en 2013

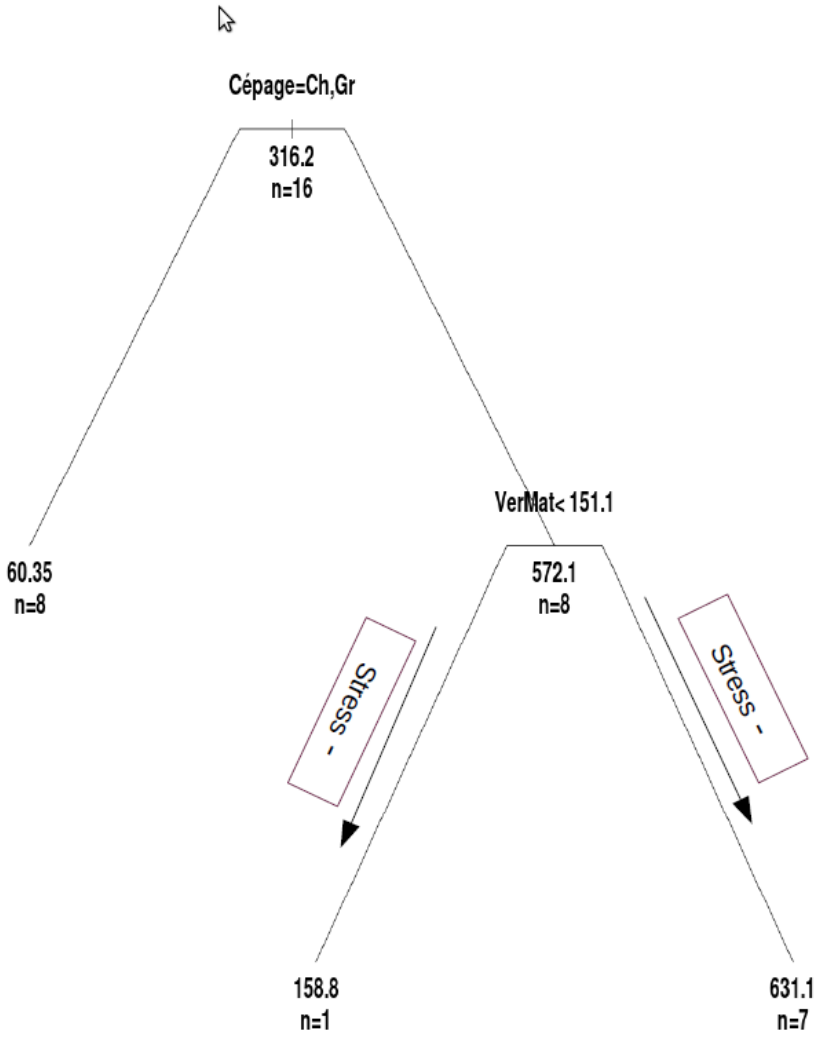


FIGURE E.1 – Arbre de régression sur le PDMS en 2013

Annexe F

Mesure de l'influence du vécu hydrique post-véraison sur le cys3MH en 2012

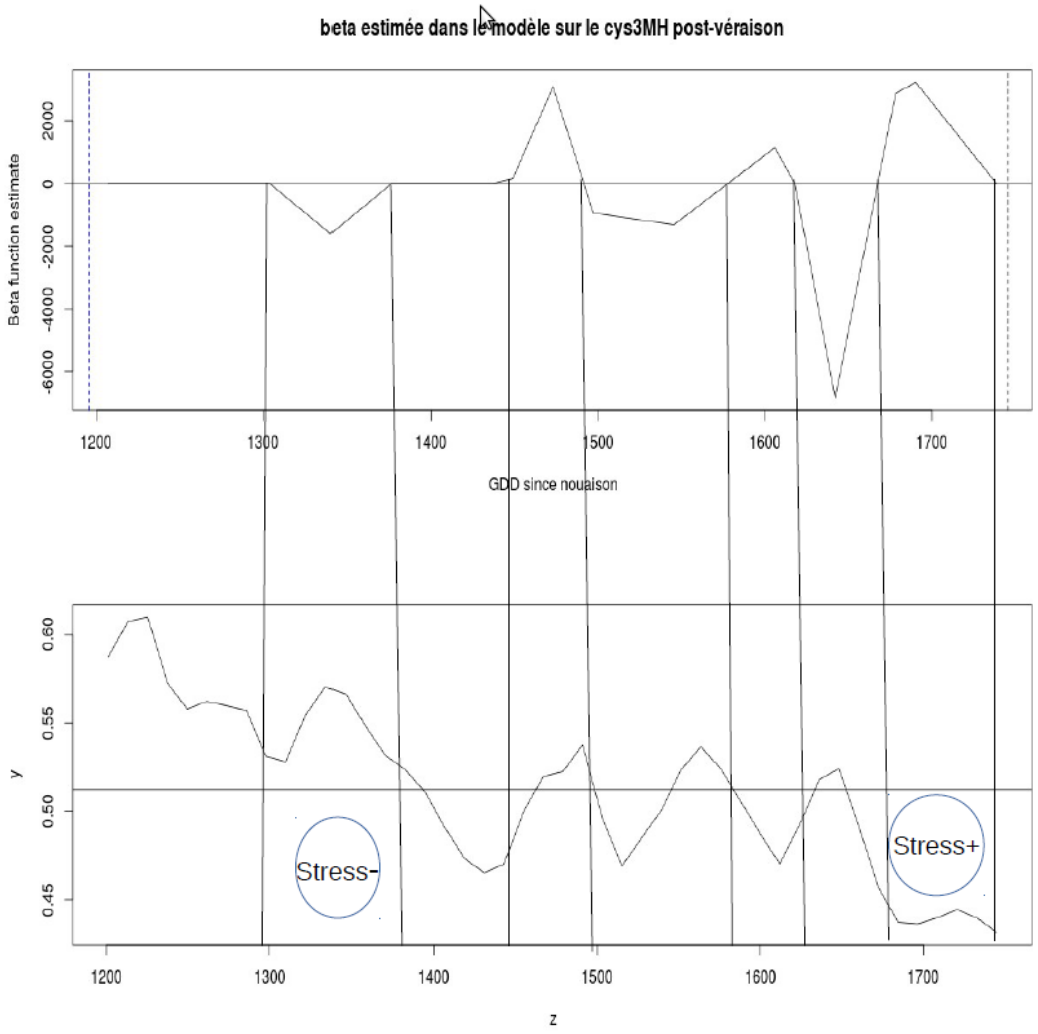


FIGURE F.1 – Estimation de β dans le modèle expliquant le cys3MH en 2012 avec le vécu hydrique post-véraison