Modélisation statistique de la qualité en viticulture par la méthode PLS Spline

Modelling grape and wine quality through PLS Spline statistical method

CLAVERIE M.*, PRUD'HOMME PY., MONGENDRE J., ZABOLLONE E., RAYNAL M., COULON T. [1], DURAND J.F. [2], MAZEIRAUD JF., RIVES C. [3], LAVAL C. [4], LAPORTE R. [5], FORGET D. [6]

[1] Institut Français de la Vigne et du Vin (ENTAV-ITV France), Station régionale Aquitaine, 39 rue Michel Montaigne, Blanquefort, France,
[2] Laboratoire de Probabilités et Statistiques, Université de Montpellier II, Montpellier, France
[3] Chambre d'Agriculture de Lot-et-Garonne, 271 rue de Péchabout, Agen, France
[4] Chambre d'Agriculture de Dordogne, CRDA du Bergeracois, Monbazillac, France
[5] Chambre d'Agriculture des Landes, Mont de Marsan, France
[6] INRA Domaine expérimental de Couhins, Villenave d'Ornon, France

***Corresponding author**: marion.claverie@itvfrance.com

Abstract

Started in 1994, this project intends to explain quality of grapes and wines using data of soil, climate and vineyard that are currently used in field trials. Firstly set at a national scale, it has been transferred to the Aquitaine region in 2000. The work has been conducted by the ITV institute thanks to many other partners. 2 cultivars have been considered: cvs. Merlot and cabernet sauvignon.
A set of data has been collected using different years and plots showing varied environnemental and cultural situations. Data mining used *PLS Spline* method. 4 models have been produced: sugar and total acids in musts, colour intensity and total polyphenolic compounds in wines. These models point out the variables that are most influent on quality and order them. A validation with plots that have not been used to build the models has been done in 2006. The prediction is of correct level and gives a potential-like result. At the same time, the models have been integrated into a better convenient tool called SPQV 1.1 software. It is aimed to farmers's advisors.
The models do not give any prediction during the year the grapes are produced, because it uses post-harvest variables. Nevertheless they can be a helpful tool for potential zoning, plots selection or planting advising.

Key Words: vine, quality, model.

Introduction

En 1994, s'est constitué autour de l'ITV et de l'INRA de Montpellier un réseau de parcelles d'essais en agronomie à partir de partenariats techniques et scientifiques issus de nombreuses régions françaises, afin de « modéliser » la qualité des raisins et des vins. A travers ce vaste objectif, se cachait le souhait de caractériser et de hiérarchiser les grandeurs pédo-climatiques et agronomiques qui influençaient le plus l'établissement de la qualité d'un raisin et d'un vin.

Les objectifs initiaux visaient alors « la recherche de relations simples entre les facteurs de production et les principaux paramètres de la qualité des vendanges et des vins ». Relations qui, dans une second temps, doivent permettre de concevoir « des modèles simples en tant qu'outils d'aide à la décision » (CR d'activités 1994- Ch. RIOU). Le projet Modèle Qualité doit permettre, à travers la description d'un grand nombre de situations contrastées de parcelles et de millésimes, d'expliquer et de modéliser de manière statistique les variables de la qualité du raisin et du vin en utilisant les descripteurs agro-climatiques couramment utilisés en expérimentation.

Le suivi de ces sites, le renseignement des bases de données correspondantes puis les traitements statistiques de ces données entre 1994 et 1999 ont constitué le cœur de la phase dite « nationale » du projet « Modèles Qualité ». A partir de 2000, l'échelle du projet a été revue, donnant lieu à la phase

« régionale » du projet. La région retenue pour l'étude est l'Aquitaine. Deux cépages sont alors étudiés séparément : merlot et cabernet sauvignon.

Au-delà de la réponse à ces objectifs, le projet « Modèle Qualité » a permis au fil des ans d'acquérir et d'expérimenter une méthodologie adaptée au traitement des dispositifs multi-variables de type « Observatoire ».

Matériel et méthodes

Démarche

La démarche du projet « Modèles Qualité » est la suivante :

1- collecte de données d'essais permettant de renseigner une base de données de variables explicatives de la qualité (sol, climat, plante au sens large) et de variables dites « à expliquer » décrivant cette qualité (caractéristiques analytiques des raisins, des moûts, des vins, description organoleptique des vins) ;

2- traitements statistiques successifs de ces données ; c'est-à-dire modélisation d'une variable « à expliquer » à l'aide des variables explicatives (phase d'apprentissage); ajout de données annuelles et actualisation des modèles ;

3- obtention de modèles validés ; initiation de la validation « externe » des modèles qui consiste à les éprouver sur des sites et/ou des millésimes n'ayant pas pris part à la construction des modèles ;

4- développement d'un outil de diffusion des modèles permettant une manipulation facilitée par les utilisateurs finaux (conseillers de terrain).

Sites suivis pour la construction des modèles

L'objectif de départ était de valoriser des données d'essais existants. En conséquence, des données émanant de plusieurs sites d'essai suivis entre 1994 et 1999 ont été intégrés à l'étude. Il s'agit d'essais d'enherbement, de systèmes de taille ou d'éclaircissage. A partir de 2000, une série de sites d'essais de type « rognage-éclaircissage » a été également intégrée à la base de données.

Du fait de la mise en place progressive des sites, toutes les années ne sont pas identiquement représentées sur chacun. De la même façon, les sites n'ont pas été explicitement choisis pour rassembler toute la diversité des situations régionales. Ils doivent représenter des situations suffisamment contrastées pour proposer une gamme de variation des variables explicatives la plus large possible.

Ces sites d'essais suivis sont localisés sur différents secteurs viticoles :

- En merlot : Libournais, Entre-Deux-Mers, Bergeracois, Buzet et Marmandais ;
- En cabernet sauvignon : Médoc, Graves, Entre-Deux-Mers et Tursan dans les Landes.

7 sites d'essai suivis entre 1994 et 2003 (92 individus) ont servi à bâtir les modèles du merlot ; en cabernet sauvignon, ce sont 8 sites suivis entre 1994 et 2004 (71 individus). Les densités de plantation varient de 2380 à 6060 ceps à l'hectare en merlot et de 2380 à 8700 ceps à l'hectare en cabernet sauvignon. Le type de taille est variable : le guyot double, simple, arcure et, en retrait, le cordon bilatéral.

Sites de validation externe des modèles

La validation a été obtenue à partir des données issues de parcelles dites « tout-venant » suivies sur 2004 et 2005. Choisis en Aquitaine, de préférence hors des zones viticoles qui avaient servi à bâtir les modèles, ces sites n'ont fait l'objet d'aucun dispositif d'essai, ils étaient suivis à la façon du viticulteur. Ils présentent une diversité de conduites, densités, sols intéressantes dans une optique de validation. Respectivement 15 et 16 parcelles ont été suivies en merlot et en cabernet sauvignon.

Variables utilisées

Le suivi des parcelles d'essai porte sur l'environnement climat/sol, la phénologie, le feuillage, la vigueur, la maturation, la récolte, les moûts et les vins.

Les parcelles sont suivies par chaque partenaire selon un protocole commun et pilotées en réseau par l'IFV (réunion annuelle et fichier de saisie commun). A la récolte, 50 kg de raisin sont vinifiés selon

un protocole de vinification commun sur 3 centres différents. Ces vinifications séparées constituent indéniablement un biais plus ou moins important dont il faut tenir compte dans l'interprétation des résultats concernant les données sur vins. La dégustation de l'ensemble des vins est, elle, intégralement centralisée.

Les variables « à expliquer » choisies pour décrire la qualité concernent la teneur en sucre (en g/l) et l'acidité totale (en g/l H_2SO_4) des moûts, ainsi que l'intensité colorante (=DO 420nm+ 520nm+ 620nm) et l'Indice de Polyphénols Totaux (IPT = DO280 nm) des vins. Ces deux dernières mesures ont été réalisées sur vins jeunes avant mise en bouteille, dans une période de temps allant de février à juin suivant la vinification.
La liste des variables explicatives figure au tableau 1.

Variables explicatives		*Unité*
Wth	Bilan hydrique théorique $_{\text{du 1 avr. au 30 sept.}}$ (W_0=200mm, k=0.5)	*mm*
West	Bilan hydrique estimé $_{\text{du 1 avr. au 30 sept.}}$ (W_0, estimé, k calculé/végétation)	*mm*
IH	Indice de Huglin =Σ K*(Tmax(base10)+Tmoy(base10))/2 $_{\text{1 avr. au 30 sept.}}$	*°C*
STVR	Somme des Températures Véraison-récolte =Σ (Tmoy(base10))$_{\text{véraison-récolte}}$	*°C*
MD	Date de mi-débourrement (50% bourgeons stade C)	*N° julien*
MF	Date de mi-floraison (50% inflorescences stade I)	*N° julien*
MV	Date de mi-véraison (50% des baies vérées)	*N° julien*
DR	Date de récolte	*N° julien*
SFEp	Surface foliaire exposée (méthode Carbonneau)	*m²/m² de sol*
SCV	Gabarit de végétation =(2*hauteur+largeur) * (1-% discontinuités)	*m²/m² de sol*
C	Charge en bourgeons à l'hectare	*eff./ha*
N	Nombre de rameaux à l'hectare	*eff./ha*
NF	Nombre de feuilles principales par rameau primaire	*effectif*
PBT	Poids des bois de taille par hectare	*kg/ha*
V	Vigueur =PBT par cep/nombre de sarments par cep	*g*
PBT/H	Indice de vigueur corrigée =PBT/hauteur de feuillage	*g/cm*
P	Puissance =0.5*PBT + 0.2*PR	*Kg/m² de sol*
NG	Nombre de grappes par cep à la récolte	*eff./cep*
PR	Poids de récolte à l'hectare	*kg/ha*
PMG	Poids moyen d'une grappe	*g*
SFE/PR	Rapport feuille/fruit (avec SFE)	*m²/kg*
SECV/PR	Rapport feuille/fruit (avec SCV)	*m²/kg*

Tableau 1 Liste des variables explicatives utilisées dans le traitement des données Modèle Qualité

Traitement statistique des données

Après plusieurs tâtonnements lors de la phase nationale du projet, la méthode choisie pour le traitement des données est *Partial Least Square* (PLS), selon un programme mis au point par JF.Durand (Univ. Montpellier II). Les traitements sont faits sous le logiciel R 2.0.1, utilisant les versions 10.0 et 10.2 du programme de JF Durand.

PLS est une méthode de régression sur composantes principales. Elle permet d'expliquer une (ou plusieurs) variables résultat (Y) par une combinaison linéaire de composantes issues des variables explicatives (X). Dans le cas de la méthode PLS linéaire, les composantes principales sont issues de la maximisation de la covariance entre les X et le (ou les) Y. PLS Spline, qui est une variante de PLS linéaire, permet de capturer des influences polynomiales entre les X et les Y. La matrice des X est préalablement transformée en une base B de polynômes par morceaux. La régression linéaire est alors faite sur les composantes principales issues de la maximisation de la covariance entre B et Y.

La modélisation consiste à rechercher un modèle de bonne qualité et « qui a du sens » pour l'agronome. Elle est appelée « validation interne », et la sanction est le PRESS : cet indicateur, dépendant du nombre de composantes, est obtenu par modélisation sur les données auxquelles une fraction des individus a été ôtée (généralement 5-10%, fraction établie par le modélisateur). Chaque

individu est donc ôté une fois puis estimé : le meilleur PRESS est obtenu quand la somme des écarts à la valeur réelle est la plus faible. Le PRESS, associé à une composante est affecté d'un R^2, qui représente la qualité de l'ajustement de l'estimation à la valeur réelle. L'influence du modélisateur est grande dans l'établissement d'un modèle : c'est lui qui décidera à partir de quand un modèle qui présente un bon PRESS et des courbes interprétables peut être arrêté.

Le principal intérêt de PLS est de s'affranchir de variables explicatives très corrélées entre elles, ce qui est couramment le cas dans ce type d'étude. PLS Spline permet quant à elle de capturer des relations non-linéaires entre 2 variables et de s'affranchir d'éventuelles données extrêmes.

Résultats

Gamme de variation des variables explicatives (climat/sol, plante)

Un modèle statistique n'étant valable que sur l'intervalle qui a servi à le bâtir, il est important de vérifier que les données d'apprentissage décrivent bien un large choix de situations.
A titre d'exemple :
- des indices de Huglin variant de 1750°C à 2640°C ;
- des bilans hydriques de –200 à 300 mm ;
- des stades phénologiques variables : débourrement de mi mars à fin avril, véraison de fin juillet à fin août ;
- des SFEp de végétation de 0.3 à 0.9 ha/ha de sol ;
- des charges en bourgeons de 20 000 à 80 000 bourgeons/ha ;
- des poids de récolte de 3 500 à plus de 20 000 kg/ha :
- des rapports SFEp/PR de 0.3 à 2,5 m²/kg.

Modèles- Validation externe

Les figures 1 et 2 présentent respectivement les 4 modèles sucre et AT des moûts, intensité colorante et IPT des vins pour le merlot puis le cabernet sauvignon. Les valeurs sont centrées-réduites.

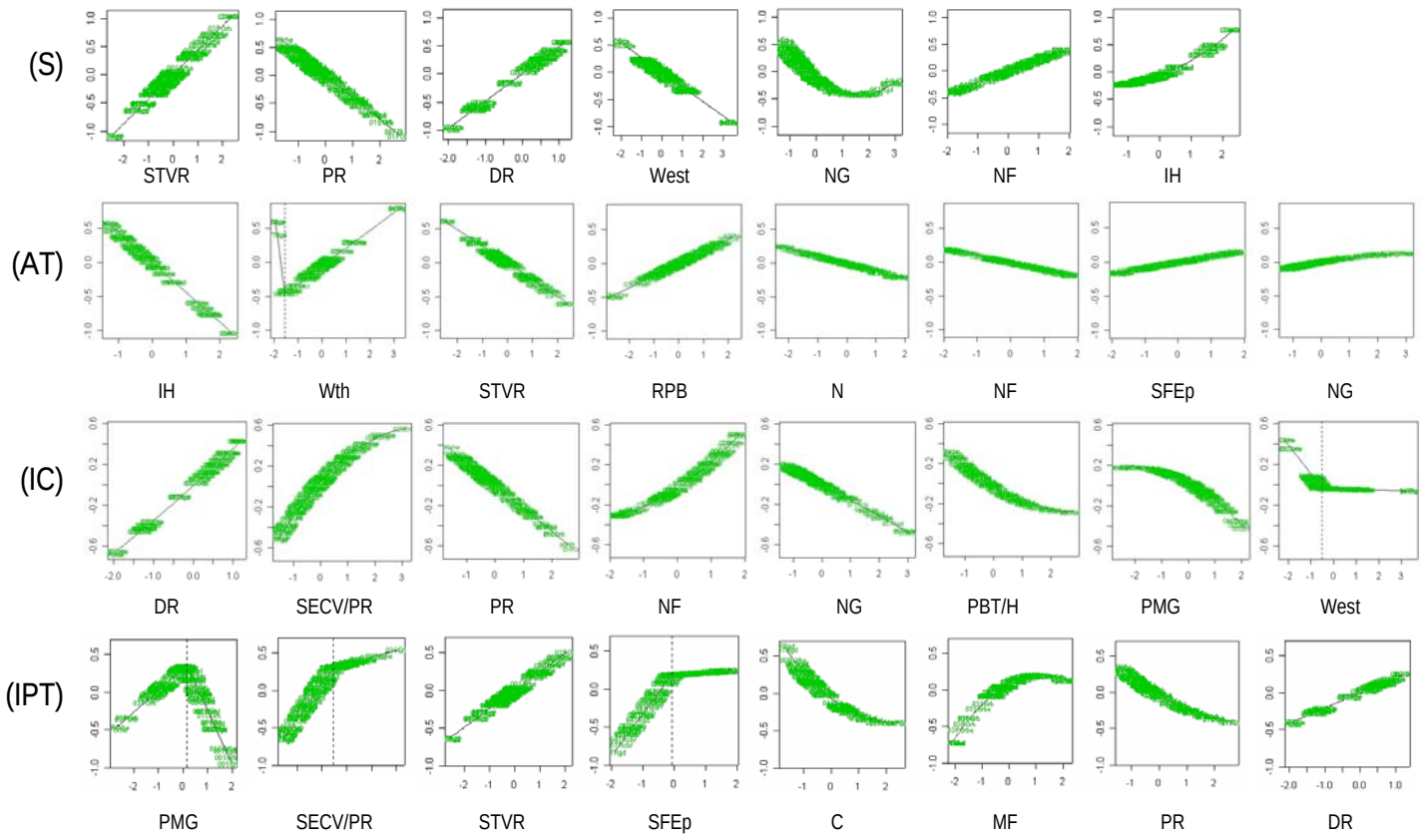

Figure1 Variables influentes pour les modèles sucre (S) et acidité totale (AT) des moûts, intensité colorante (IC) et indice de polyphénols totaux (IPT) des vins. MERLOT.

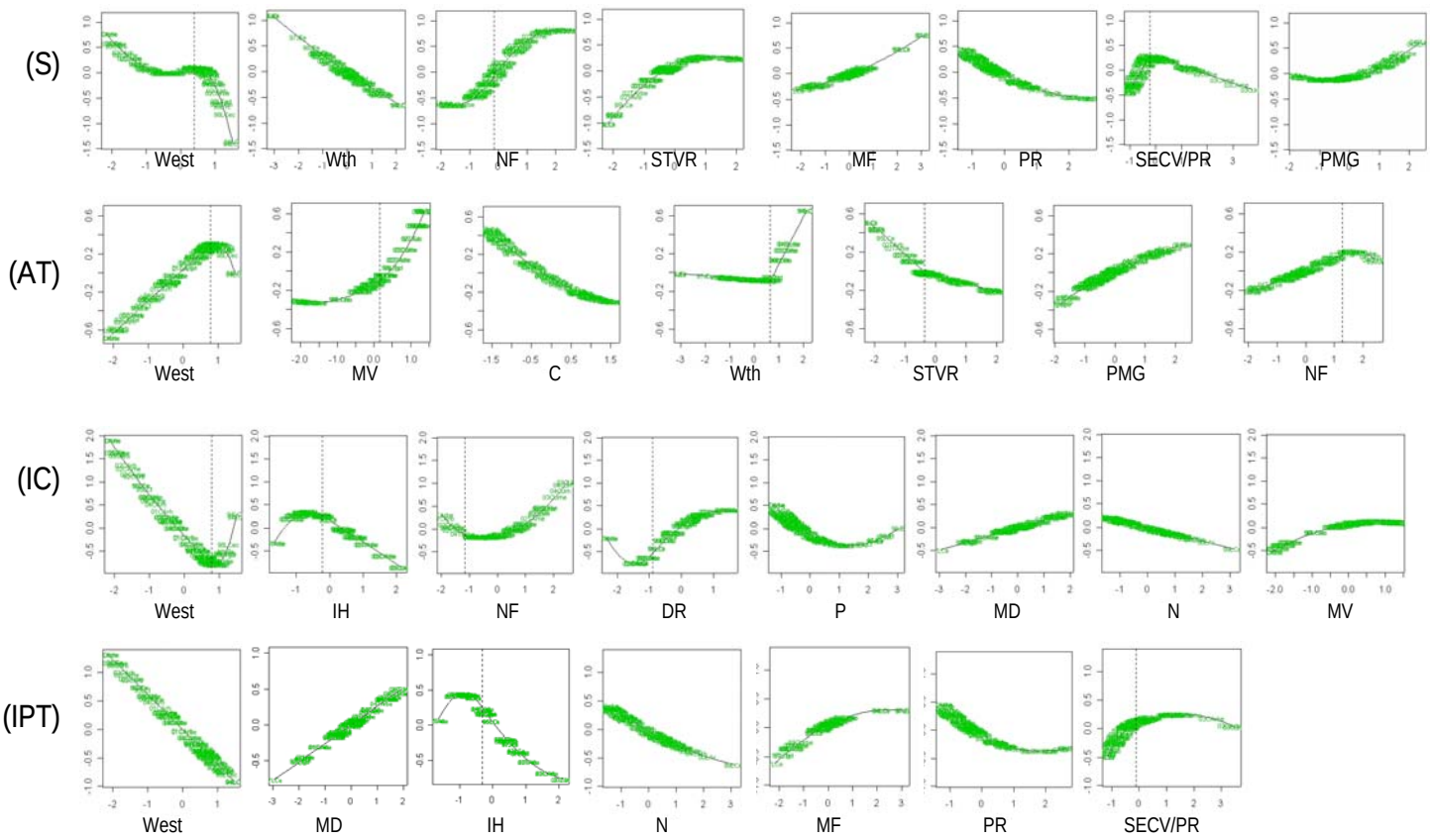

Figure 2 Variables influentes pour les modèles sucre (S) et acidité totale (AT) des moûts, intensité colorante (IC) et indice de polyphénols totaux (IPT) des vins. CABERNET SAUVIGNON.

Commentaires des modèles- Interprétation

Qualité des modèles, précision de la prédiction

Les 8 modèles présentés sont de qualité correcte : PRESS moyens à bons (0.2-0.4), R^2 convenables (0.75 à 0.88, 1 modèle moins bon IC merlot à 0.66), peu de dimensions (3 ou 4).

La qualité de la validation externe (non présentée ici) est variable : dans l'ensemble, les sites-test sont correctement ordonnés, et l'ordre de grandeur est appréhendé. Quelques sites font exception toutefois. La précision à attendre de l'estimation est d'environ 0.5%vol. pour le sucre, 0.5 g/l (H2SO4) pour l'AT, 1 à 1.5 pt pour l'IC et 5 points pour l'IPT.

Déterminisme du sucre, de l'AT, de l'IC et de l'IPT- Différences entre cépages

Pour les 2 cépages, on voit que le déterminisme du **sucre** est fortement influencé par les conditions de maturation, le bilan hydrique, la quantité de récolte et de végétation. Pour le merlot, les conditions pendant la maturation ainsi que la quantité de récolte sont de loin les grandeurs les plus influentes : maturations longues ou chaudes et récoltes faibles expliquent à elles seules une quantité plus importante de sucre. Pour le cabernet sauvignon en revanche, ce sont les influences du bilan hydrique et de la quantité de feuillage qui jouent le plus, les conditions de récolte et de maturation restant importantes mais en retrait.

Pour l'**acidité totale**, les 2 cépages se comportent de manière plus proche : c'est d'une part le bilan hydrique et donc la taille des baies, et d'autre part les conditions thermiques du millésime (notamment des maturations courtes) qui déterminent le niveau d'acidité dans les moûts.

Pour l'**intensité colorante** des vins, on retrouve pour les 2 cépages l'influence forte de la date de récolte et de la quantité de feuillage (positives). Cependant les autres grandeurs influentes en merlot sont la production et la vigueur (toutes deux influencent négativement le potentiel couleur) quand pour

le cabernet sauvignon il s'agit du bilan hydrique et des conditions thermiques du millésime (influence négative des millésimes chauds, ce qui ne ressort pas sur merlot).

Enfin, le déterminisme de la quantité de **polyphénols totaux** dans les vins (à conditions de vinification constante) diffère d'un cépage à l'autre : quand pour le merlot l'IPT dépend surtout des conditions de maturation (maturation longue), de quantité de récolte et de feuillage, pour le cabernet sauvignon c'est surtout une question climatique (alimentation hydrique faible et de millésime frais sont favorables) ainsi que, plus en retrait, de production.

Des spécificités se dessinent pour chacun des 2 cépages étudiés. On remarque que bilan hydrique est une grandeur très importante pour le cabernet sauvignon. On le retrouve toujours très influent dans les modèles. Or cette variable varie dans une gamme aussi large en merlot qu'en cabernet sauvignon. De même on a pu vérifier que ce comportement ne masquait pas un effet site en cabernet sauvignon. Tout au plus peut on dire qu'en merlot les valeurs de bilan hydrique estimé sont plus resserrées autour de la médiane quand elles décrivent mieux l'ensemble de la gamme en cabernet sauvignon. Le raisonnement de l'implantation de ce cépage vis-à-vis du type de sol est donc déterminant pour la réussite de la production ultérieure.
Le merlot quant à lui est bien moins regardant, les modèles font ressortir une forte influence des conditions de maturation et de rendement, et de feuillage, cette dernière surtout en ce qui concerne les composés phénoliques. Ce dernier cépage semble d'ailleurs bien plus sensible au rendement que le cabernet sauvignon. Ceci est visible même sans modèle, sur les graphes bivariés reliant sucre, intensité colorante ou IPT au poids de récolte à l'hectare. Et ce avec une gamme de poids de récolte très large (de 5 à plus de 25 t/ha) et tout à fait superposable pour l'un et l'autre cépage.

Les modèles sont additifs ; en l'état ils ne prennent pas en compte les interactions entre variables. Une telle approche a été testée sur le jeu de données grâce une programme de PLS Spline prenant en compte de ces interactions. Certaines interactions ressortent effectivement, notamment impliquant la variable de bilan hydrique, mais elles n'améliorent pas suffisamment les modèles pour mériter d'être prises en compte.

Développement- Outil d'aide à la décision

L'objectif final du projet « Modèles Qualité » est d'intégrer de ces relations statistiques dans un outil d'aide à la décision dont la cible serait la communauté technique, en particulier les conseillers techniques en Aquitaine.

Des travaux de construction d'un logiciel en vue d'une prise en main des modèles ont débuté en 2005. Le programme, baptisé SPQV (Scénarios pour la Prévision de la Qualité en Viticulture), fonctionne sous R (logiciel en libre accès) et utilise des commandes rappelant celles sous Windows (barre de taches, menus). Il permet, moyennant la saisie des valeurs des variables explicatives, de prévoir un potentiel en sucre des moûts et en polyphénols dans le vin fini (via des conditions de vinifications qui sont celles de l'essai).

L'application sous forme de logiciel d'utilisation de ces modèles a été transférée en 2007 auprès des professionnels intéressés (conseillers de terrain, chefs de culture) de la région Aquitaine. Un retour est attendu afin de vérifier si les relations donnent effectivement satisfaction au quotidien.

Conclusion

Le projet de modélisation de la qualité de la récolte initié en Aquitaine en 2000 a abouti à 4 modèles validés permettant de prédire un potentiel de maturité, de couleur ou de polyphénols d'une parcelle de cabernet sauvignon ou de merlot.

Les modèles obtenus confortent nos connaissances sur les facteurs influents sur les 4 variables de qualité considérées, mais au-delà, leur intérêt réside dans le fait qu'ils nous permettent de quantifier et combiner ces influences pour aboutir à un résultat. Ils utilisent des variables qui sont celles de l'expérimentateur ; de ce fait, ils s'adressent plus à un conseiller qu'à un producteur.

Ces modèles étant statistiques, ils ne sont utilisables que sur l'intervalle de validité des variables utilisées ; cependant cette gamme est large aussi bien en termes de millésimes que de sites.

Le transfert des modèles via un logiciel d'utilisation a été initié en 2007 auprès d'un petit groupe de professionnels intéressés. Ce n'est qu'à l'issue de la prise en main de l'outil par ces techniciens sur le plus de secteurs possibles que les modèles pourront réellement être jugés validés.

Un tel outil peut venir en appui d'un conseil plantation ou d'une sélection parcellaire. Dans le cas où le climat joue un rôle important, le modèle peut aussi servir de base à une cartographie des potentialités.

Enfin au-delà de cela, cette étude a permis d'expérimenter une méthode de traitement statistique de dispositifs multi-factoriels en réseau.

Bibliographie

DURAND J.F. 2001. Local polynomial additive regression through PLS and splines : PLSS. *ELSEVIER Chemometrics and intelligent laboratory systems*. **58** (2001) 235-246.