

# Données, modèles et application *Shiny* pour une classification statistique des vignobles de Côte-d'Or

Jean-Sauveur AY                      Mohamed HILAL  
< [jean-sauveur.ay@inra.fr](mailto:jean-sauveur.ay@inra.fr) >      < [mohamed.hilal@inra.fr](mailto:mohamed.hilal@inra.fr) >

Unité Mixte de Recherche CESAER  
AgroSup / INRA / Univ. Bourgogne Franche-Comté  
26 boulevard du docteur Petitjean 21000 DIJON

*Data paper* version 1.0 du Vendredi 12 juillet 2019

## Résumé

Cet article présente la construction d'une base de données au niveau des parcelles cadastrales pour étudier les relations entre les caractéristiques biophysiques (topographie, géologie, pédologie) et les appellations d'origine contrôlée (AOC) viticoles. Sur les 31 communes de la Côte-d'Or qui forment la côte de Beaune et la côte de Nuits, nous proposons une modélisation statistique qui permet de classer l'ensemble des parcelles sur une échelle de qualité continue exclusivement à partir de leurs caractéristiques biophysiques. Nous montrons en particulier la persistance d'effets communaux significatifs issus d'éléments historiques non reliés aux caractéristiques des parcelles. Les données, méthodes et prédictions du modèle sont disponibles sous licence GNU GPL V3 sur <https://data.inra.fr/> et sont consultables par le biais d'une application *Shiny* sur <http://github.com/jsay/geoInd/>.

**Mots-clés:** Économie viti-vinicole ; signes de qualité ; recherche reproductible ; système d'information géographique ; modélisation économétrique.

## Table des Matières

<b>1</b>	<b>Introduction</b>	<b>2</b>	<b>4</b>	<b>Modèle statistique</b>	<b>14</b>
			4.1	Estimation du modèle	14
			4.2	Variables biophysiques	15
			4.3	Effets communaux	17
			4.4	Prédiction continue	18
			4.5	Agrégation par lieux dits	19
<b>2</b>	<b>Construction des données</b>	<b>3</b>	<b>5</b>	<b>Application <i>Shiny</i></b>	<b>22</b>
2.1	Les AOC à la parcelle cadastrale	3	5.1	Cartographie dynamique	22
2.2	Enrichissement de la topographie	5	5.2	Lancer l'application localement	22
2.3	Enrichissement de la géologie	5	5.3	Utilisation	22
2.4	Enrichissement de la pédologie	6	5.4	Développements à venir	22
2.5	Enrichissement des AOC de 1936	7			
2.6	Enrichissement des lieux dits	7	<b>6</b>	<b>Conclusion</b>	<b>23</b>
2.7	Enregistrement de la base	8	<b>7</b>	<b>Bibliographie</b>	<b>23</b>
<b>3</b>	<b>Statistiques descriptives</b>	<b>10</b>	<b>A</b>	<b>Annexes</b>	<b>25</b>
3.1	Sélection des données	10			
3.2	Distribution des AOC	10			
3.3	La pyramide des AOC	12			
3.4	Tableau des variables utilisées	13			

# 1 Introduction

Les appellations d'origine contrôlée (AOC) viticoles en Bourgogne résultent de processus historiques complexes au cours desquels les parcelles ont été classifiées selon leurs caractéristiques biophysiques au grès des rapports économiques, politiques et sociaux en vigueur (Garcia, 2011; Wolikow and Jacquet, 2011). La classification actuelle est issue de plusieurs siècles de culture de la vigne, de production de vin et de négociation sur les dénominations. Ces trois ensembles de pratiques forment les usages loyaux et constants selon la doctrine de l'institut national de l'origine et de la qualité (INAO) pour définir, reconnaître et gérer les AOC (Capus, 1947; Humbert, 2011). La complexité des informations contenues dans la référence au lieu de production et leurs évolutions dans le temps sont à la fois des forces et des faiblesses pour les AOC. Elles permettent de simplifier les nombreux déterminants biophysiques de la qualité des vins au risque d'une faible pertinence et d'une opacité croissante pour les acteurs du marché du vin.

La question de la transmission de l'information entre les producteurs et les consommateurs sur la qualité des biens échangés fait l'objet d'une littérature économique abondante (Coestier and Marette, 2004). L'asymétrie d'information y est typiquement décrite comme une défaillance de marché qui porte préjudice aux parties prenantes de la relation commerciale. Par contre, le recours aux indications géographiques apparaît comme une solution partielle qui peut segmenter artificiellement la production et générer des rentes non justifiées pour les producteurs au détriment des consommateurs. Nous nous concentrons dans cet article sur la capacité des AOC à simplifier l'information sur les caractéristiques des lieux de production, en laissant de côté la question de la pertinence de cette information pour le marché. Il s'agit de quantifier dans quelle mesure la hiérarchie des AOC bourguignonnes en 5 catégories, Coteaux bourguignons < Bourgogne régional < Bourgogne village < Premier cru < Grand cru, permet de résumer les nombreux déterminants biophysiques de la qualité des vins (topographie, géologie, pédologie). Cette analyse statistique permet de mettre en évidence des déterminants historiques de la hiérarchie des AOC non reliés aux caractéristiques individuelles des parcelles mais reliés à la commune à laquelle elles appartiennent. Les résultats issus de cette modélisation sont consultables par le biais d'une application *Shiny* dont l'utilisation est présentée dans cet article.

Le travail sur les données consiste à apparier les informations biophysiques des parcelles cadastrales aux AOC par l'utilisation d'un système d'information géographique. La Section 2 présente en détail la construction des données, avec en particulier les références des fichiers sources et le dictionnaire des variables produites. Ce premier travail aboutit à la construction d'une base de données spatialisée librement disponible sous licence GNU GPL V3 sur <https://data.inra.fr/>. La parcelle cadastrale est l'unité géographique de base qui permet l'appariement de variables altimétriques du RGE ALTI® 5m (IGN), d'un modèle d'occupation du sol (Hilal et al., 2018)??, de variables géologiques de Charm-50 (BRGM), de variables pédologiques du Référentiel Pédologique de Bourgogne (Gis Sol), de variables historiques sur les AOC en 1936 (MSH Dijon) et de variables sur les lieux dits issus du Plan Cadastral Informatisé (DGFIP). Les données construites concernent actuellement les 31 communes qui constituent la côte de Beaune et la côte de Nuits, soient l'ensemble des vignobles du département de la Côte-d'Or à l'exception des hautes côtes et du Châtillonnais (voir Figure 1). Cette base de données permet de relier finement les AOC aux caractéristiques des parcelles dont les vins sont issus, et peut ainsi être utilisée pour d'autres questions de recherche. Par contre, l'ensemble des données initiales ne sont pas diffusables à l'état brut. C'est ainsi que la Section 2 présente les traitements effectués, avec les codes R reportés en Annexe, sans que les données qui servent à les alimenter. La partie reproductible de cet article commence avec les principales statistiques descriptives relatives à ces données qui sont présentées dans la Section 3.

La Section 4 contient ensuite le détails de l'estimation du modèle statistique dont la spécification est décrite plus extensivement dans un article associé (Ay, 2019). Le principe de la modélisation est d'utiliser la structure hiérarchique qui existe entre les différents niveaux des AOC pour simplifier le rôle

des caractéristiques biophysiques au travers d’une variable latente de qualité des vignes. Cette variable continue représente un niveau ordinal de qualité des vignes tel que révélé par la hiérarchie des AOC actuelles. Nous présentons l’estimation d’un modèle ordonné additivement semi-paramétrique (OGAM, [Wood et al., 2016](#)) qui permet de prédire correctement près de 90% des niveaux d’AOC par un lissage spatial fin. Ce modèle permet d’estimer non-paramétriquement l’effet de chaque variable biophysique sur la hiérarchie des AOC. Il permet également d’identifier des effets communaux indépendants des variables biophysiques, potentiellement issus de facteurs humains tels que la réputation de la commune, la proximité à la ville centre (Dijon) ou l’antériorité des syndicats de producteurs ([Jacquet, 2009](#)). L’estimation d’effets communaux significatifs permet de hiérarchiser chaque commune de la zone en fonction de la probabilité qu’une de ses parcelles soit plus haute dans la hiérarchie qu’une parcelle similaire prise au hasard sur la zone ([Agresti and Kateri, 2017](#)). Cela nous permet également de corriger ces effets communaux dans la variable latente de qualité des vignes déduite des AOC, ces effets étant *a priori* non reliés à la qualité des vins produits.

Les prédictions issues du modèle statistique sont reportées, corrigées et non corrigées des effets communaux, dans la base de donnée issue de cette recherche. Ces données alimentent également une application *Shiny* dont le code et le manuel d’utilisation sont présentés dans la Section 5. Cette application peut être utilisée localement (moyennant la présence du [logiciel R](#) sur l’ordinateur de l’utilisateur) ou consultable sur internet à l’[adresse dédiée](#). Par le biais de cette application, l’utilisateur peut saisir une référence de vin à partir d’informations typiquement disponibles sur l’étiquette de la bouteille afin d’identifier géographiquement le lieu de production du raisin et son niveau de qualité (avec ou sans correction) sur une échelle normalisée de 0 à 100. Cette information permet une évaluation plus précise que la hiérarchie des AOC en 5 niveaux, et permet en outre de situer ce vin par rapport aux autres vins du même niveau hiérarchique. Elle permet ainsi d’augmenter l’information disponible au niveau du consommateur pour effectuer ses choix de vin. Il est important de mentionner que la classification obtenue se base exclusivement sur les AOC actuelles et à ce titre ne contient aucune appréciation subjective sur l’importance des différents facteurs biophysiques ou olfactif sur la qualité. La classification obtenue est déduite de la hiérarchie actuelle des AOC qui, au regard de la hiérarchie de prix qu’elle produit, semble crédible pour les acteurs du marché.

Une part de subjectivité persiste toutefois dans la classification proposée. Elle est relative à la spécification du modèle statistique, et le fait d’avoir favorisé des variables biophysiques pour décrire la relation entre les lieux de production et la qualité des vins. Le débat sur l’articulation des facteurs humains et des facteurs naturels de la qualité vin existe depuis plus d’un demi-siècle et produit des discussions toujours d’actualité ([Dion, 1952](#); [Delay and Chevallier, 2015](#)). C’est pour alimenter ce débat que les analyses présentées dans cet article sont reproductibles à partir de la base de donnée produite. Les codes pour l’estimation du modèle sont reportés dans l’article, afin de permettre l’estimation de modèles alternatifs, sachant que l’appariement de données supplémentaires pour relier les AOC aux caractéristiques des lieux est également possible. La transparence des analyses permet aux résultats d’être discutés et contestés afin de faire l’objet d’une appropriation par les chercheurs, les décideurs, les professionnels du secteur, ou les amateurs de vin.

## 2 Construction des données

### 2.1 Les AOC à la parcelle cadastrale

L’unité géographique de base est la parcelle cadastrale dont la géométrie est issue de la BD parcellaire de l’IGN version 2014 pour la Côte-d’Or téléchargée le XX/XX/2018. Ces données sont disponibles gratuitement pour la recherche, elle ne peuvent pas faire l’objet d’une diffusion à l’état brut. Trois traitements ont été effectués au préalable et ne sont pas reportés en détail ici. Nous avons calculé avec un système d’information géographique les caractéristiques géométriques (surface, périmètre, et distance maximale

entre deux sommets). Nous avons ensuite créé un identifiant pour appairier les parcelles avec les données du modèle numérique de terrain présenté dans la sous-section suivante. Nous avons enfin apparié les délimitations parcellaire des AOC Viticoles de l'INAO disponible à l'adresse <https://www.data.gouv.fr/fr/datasets/delimitation-parcellaire-des-aoc-viticoles-de-linao> sous licence ouverte. Les codes R correspondants à ces opérations sont reportés en Annexe.

L'objet Geo.Cad contient 110 350 parcelles et 16 variables que la Table 1 suivante présente plus en détails. L'information brute issue de la superposition avec la couche INAO est présente dans les variables PAOC à GCRU. Les 49 718 valeurs manquantes qui apparaissent correspondent aux parcelles hors AOC. Nous avons retravaillé l'information brute des données INAO dans les trois variables AOC, AOCtp et AOC1b qui sont plus opérationnelles pour l'analyse statistique. Dans la doctrine de l'INAO, les parcelles d'un niveau hiérarchique supérieur peuvent toujours être revendiquées dans un niveau inférieur (c'est le principe des replis). La superposition des couches de l'INAO conduit à la présence de plusieurs AOC sur une même parcelle, ce qui entre en contradiction avec une autre partie de la doctrine de l'INAO, à savoir qu'il est interdit de revendiquer des AOC différentes pour un même produit. Dans les faits, les producteurs revendiquent très souvent l'AOC maximale à laquelle ils peuvent prétendre. La variable AOC représente cette AOC pour chacune des parcelles de la base, elle est codée 0 pour les parcelles hors AOC, 1 pour les Coteaux bourguignons, 2 pour les Bourgognes régionaux, jusqu'à 5 pour les Grands crus. Par contre, les informations présentes sur l'étiquette des vins peuvent être des appellations ou des dénominations au sein du système des AOC (même si cette distinction n'est pas toujours claires pour les consommateurs, nous utilisons AOC comme le terme générique qui englobe les deux en précisant lorsque c'est nécessaire). Le libellé AOC1b contient donc généralement le nom de l'appellation maximale de la parcelle, sauf pour les Bourgognes régionaux (ou la dénomination "Bourgogne côte d'or" est plus haute dans la hiérarchie mais peu utilisée du fait de sa faible antériorité, 2015) et les "Premiers Crus" (qui ont chacun une dénomination qui permet de les distinguer).

Table 1: **Nom, type et description des variables disponibles au niveau des parcelles cadastrales**

NOM	TYPE	DESCRIPTION
IDU	<i>Caractère</i>	Identifiant de la parcelle cadastrale (14 caractères)
CODECOM	<i>Caractère</i>	Code INSEE de la commune d'appartenance (5 caractères)
AREA	<i>Numérique</i>	Surface calculée de la parcelle cadastrale (en mètres carrés)
PERIM	<i>Numérique</i>	Périmètre calculé de la parcelle cadastrale (en mètres)
MAXDIST	<i>Numérique</i>	Distance maximale calculée entre deux sommets (en mètres)
PAR2RAS	<i>Numérique</i>	Identifiant pour appariement avec le modèle numérique de terrain
PAOC	<i>Indicatrice</i>	1 si la parcelle est dans au moins une AOC
BGOR	<i>Indicatrice</i>	1 si la parcelle est dans le niveau Coteaux Bourguignon
BOUR	<i>Indicatrice</i>	1 si la parcelle est dans le niveau Bourgogne Régional
VILL	<i>Indicatrice</i>	1 si la parcelle est dans le niveau Bourgogne Village
COMM	<i>Indicatrice</i>	1 si la parcelle est dans le niveau Bourgogne Communal
PCRUC	<i>Indicatrice</i>	1 si la parcelle est dans le niveau Premier Cru
GCRUC	<i>Indicatrice</i>	1 si la parcelle est dans le niveau Grand Cru
AOC	<i>Numérique</i>	Rang de la parcelle dans la hiérarchie des AOC (entre 0 et 5)
AOCtp	<i>Caractère</i>	Appel si le libellé est une appellation, Denom pour dénomination
AOC1b	<i>Caractère</i>	Libellé de l'appellation ou de la dénomination selon la variable AOCtp

## 2.2 Enrichissement de la topographie

Les données sur la topographie sont issues du modèle numérique de terrain de l'IGN RESOLUTION, SITE, sous licence XX. Un premier traitement non reporté a été l'attribution de l'identifiant PAR2RAS aux cellules du raster par superposition avec la géographie du parcellaire présentée ci-dessus. Nous avons ensuite enrichi les données raster d'un mode d'occupation des sol (SOURCE) et d'une perméabilité calculée (SOURCE). Nous avons enfin calculé les variables topographiques que sont l'altitude, la pente, l'exposition et les radiations solaires (détails en Annexe). À partir des plus de 14 millions de cellules pour 13 variables, le code ci-dessous permet l'agrégation des variables raster au niveau des parcelles. Nous calculons des moyennes à l'échelle des parcelles, sachant que d'autres méthodes d'agrégation ont été utilisées sans différences sur les résultats.

Le détails des variables issue du fichier raster est disponible dans la Table 2 ci-dessous. Nous obtenons 2 096 valeurs manquantes pour lesquelles le code PAR2RAS des parcelles ne s'apparie à aucune cellule raster. Ces parcelles sont de très petites parcelles avec des géométrie particulières et font penser à des "erreurs" du cadastre. Nous les enlèverons de l'analyse sachant que cela revient à enlever 2.7 ha, moins de 0.01 % de la surface totale. Nous n'utilisons qu'un sous ensemble du MOS principalement afin de distinguer le non agricole.

Table 2: **Nom, type et description des variables issues des données raster**

NOM	TYPE	DESCRIPTION
XL93	<i>Numérique</i>	Latitude du centroïde de la parcelle (système Lambert 93)
YL93	<i>Numérique</i>	Longitude du centroïde de la parcelle (système Lambert 93)
NOMOS	<i>Numérique</i>	Part de la parcelle hors du mode d'occupation des sol (entre 0 et 1)
URBAN	<i>Numérique</i>	Part de la parcelle en usage urbain selon le MOS (entre 0 et 1)
FOREST	<i>Numérique</i>	Part de la parcelle en usage forestier selon le MOS (entre 0 et 1)
WATER	<i>Numérique</i>	Part de la parcelle en eau selon le MOS (entre 0 et 1)
DEM	<i>Numérique</i>	Altitude moyenne de la parcelle selon le MNT (en mètres)
SLOPE	<i>Numérique</i>	Pente moyenne de la parcelle selon le MNT (en degrés)
ASPECT	<i>Numérique</i>	Exposition moyenne de la parcelle selon le MNT (en degrés)
SOLAR	<i>Numérique</i>	Radiation solaire moyenne sur la parcelle (en Joules)
PERMEA	<i>Numérique</i>	Perméabilité des sols moyenne (entre 0 et 4)

## 2.3 Enrichissement de la géologie

Les données géologiques sont issues de la Bd Charm-50 du BRGM à l'échelle 1/50 000 disponible sur le site <http://infoterre.brgm.fr> sous licence Ouverte. Nous utilisons ici une extraction du fichier GEO050K\_HARM\_021\_S\_FGEOL\_CGH\_2154 effectuée en avril 2019 pour le département de la Côte d'Or. Le seul travail non reporté sur ces données est une sélection des variables bien renseignées et qui contiennent une variance non nulle sur la zone considérée. La faible taille moyenne des parcelles sous AOC (moins de 0.2 ha de moyenne) permet de s'assurer de la validité de cette procédure.

Les détails des 16 variables géologiques issues de la procédure sont disponibles dans la Table 3 suivante. La description des variables manque de détails car les données géologiques ne possèdent pas encore de dictionnaire (une demande est en cours auprès du BRGM). Ce manque de détails n'est pas fondamental pour l'analyse économétrique (il peut l'être pour d'autres usages des données) car les variables géologiques ne sont utilisées qu'au travers d'effets fixes qui permettent de s'affranchir de la nécessité de spécifier les

relations entre les variable géologiques et les AOC. Cette méthode est par ailleurs la plus générale pour contrôler l'hétérogénéité associée à la géologie. Comme nous le voyons ci-dessus, les parcelles non appariées qui produisent des valeurs manquantes sont peut nombreuses (entre 31 et 862 selon les variables) et seront négligées dans l'analyse économétrique sans conséquences. Intitulés en Annexe.

Table 3: **Nom, type et description des variables issues des données géologiques**

NOM	TYPE	DESCRIPTION
CODE	<i>Caractère</i>	Code de la géologie (31 modalités)
NOTATION	<i>Caractère</i>	Notation géologie (31 modalités)
DESCR	<i>Caractère</i>	Description géologie (31 modalités)
TYPEGEOL	<i>Caractère</i>	Type superficiel (4 modalités)
APLOCALE	<i>Caractère</i>	Colluvions, Eboulis, etc. (28 modalités)
TYPEAP	<i>Caractère</i>	Type de formation (7 modalités)
GEOLNAT	<i>Caractère</i>	Nature Géologique (3 modalités)
ISOPIQUE	<i>Caractère</i>	Faciès des couches (4 modalités)
AGEDEB	<i>Caractère</i>	Age de la couche (24 modalités)
ERADEB	<i>Caractère</i>	Céno ou Méso (2 modalités)
SYSDEB	<i>Caractère</i>	Age autre (5 modalités)
LITHOLOGIE	<i>Caractère</i>	Litho (16 modalités)
DURETE	<i>Caractère</i>	Dureté (3 modalités)
ENVIRONMT	<i>Caractère</i>	Environnement (9 modalités)
GEOCHIMIE	<i>Caractère</i>	Géochimie (5 modalités)
LITHOCOM	<i>Caractère</i>	Litho détaillée (30 modalités)

## 2.4 Enrichissement de la pédologie

Les données pédologiques sont extraites du Référentiel Pédologique de Bourgogne : Régions naturelles, pédopaysage et sols de Côte d'Or (étude 25021) à l'échelle 1/250 000, compatible avec la base de données nationale DoneSol, sous licence XX (Chrétien, 1998). La localisation des types de sol et l'appariement avec le cadastre s'opèrent par les 194 Unités Cartographiques de Sols de la zone, qui sont des polygones plutôt homogènes en termes de paysage mais qui contiennent différents types de sols. Ces derniers, regroupés en unités typologiques, ne peuvent pas être localisés plus précisément (Ay, 2011). En l'absence de données plus fines spatialement, les données parcellaires seront enrichies du code des unités cartographiques et les valeurs de l'unité typologique dominante, c'est-à-dire celle qui est la plus étendue au sein de chaque unité cartographique. Comme pour la géologie, les données pédologiques seront utilisées par des effets fixes au niveau des unités cartographiques, ce qui fait que cette procédure n'est pas limitante (elle peut cependant l'être pour d'autres usages). Les intitulés des unités cartographiques reportés en Annexe 3 sont obtenus par un travail manuel à partir du site <https://bourgogne.websol.fr/carto>.

Les détails des 13 variables pédologiques issues de la procédure sont disponibles dans la Table 4 suivante. Les valeurs manquantes associées aux parcelles non couvertes par la pédologie sont 14 645, soit XX %. Ces parcelles correspondent visuellement aux espaces urbanisés bien que cela ne se retrouve pas vraiment à partir du MOS. A DECIDER.



Table 4: **Nom, type et description des variables issues des données pédologiques**

NOM	TYPE	DESCRIPTION
NOUC	<i>Caractère</i>	Numéro de l'unité cartographique (2 caractères)
SURFUC	<i>Numérique</i>	Surface de l'unité cartographique (en hectares)
TARG	<i>Numérique</i>	Taux d'argile de l'unité typologique dominante (pourcentage)
TSAB	<i>Numérique</i>	Taux de sable de l'unité typologique dominante (pourcentage)
TLIM	<i>Numérique</i>	Taux de limons de l'unité typologique dominante (pourcentage)
TEXTAG	<i>Caractère</i>	Classes de textures agrégées en 9 modalités (voir Ay, 2011)
EPAIS	<i>Numérique</i>	Épaisseur des sols de l'unité typologique dominante (centimètre)
TEG	<i>Numérique</i>	Taux d'éléments grossiers de l'unité typologique dominante (pour mille)
TMO	<i>Numérique</i>	Taux de Matière organique de l'unité typologique dominante (pourcentage)
RUE	<i>Numérique</i>	Réserve Utile par excès de l'unité typologique dominante (millimètre)
RUD	<i>Numérique</i>	Réserve Utile par défaut de l'unité typologique dominante (millimètre)
OCCUP	<i>Numérique</i>	Part de l'unité typologique dominante dans l'unité carto (entre 0 et 1)
DESCRp	<i>Caractère</i>	Libellé de la classe pédologique en 33 modalités

## 2.5 Enrichissement des AOC de 1936

Les AOC en vigueur en 1936 à la création de l'INAO ont été obtenues de la Maison des Sciences de l'Homme de Dijon (Licence?? avec l'aide de Florian Humbert). Un travail préalable a été effectué sur les AOC de 1936 afin de compiler les différentes années de 1936 à 1940. La localisation est effectuée par le centroïde des parcelles cadastrales car la géométrie des polygones ne correspond pas parfaitement (à la fois par la numérisation et parce que le cadastre a changé). Encore une fois, la faible taille des parcelles permet d'avoir confiance en cette procédure d'appariement.

Nous obtenons des aires sensiblement plus réduites que les actuelles, 27% au lieu de 55% trouvés ci-dessus. Hormis le creux de 1938, entre 10 et 15% des parcelles sont classées chaque année, sachant qu'il y a du double compte. Les premiers crus n'apparaissent pas car ils n'existaient pas à l'époque (création en 1948). le décret instaurant les Premiers Crus ne fut toutefois adopté qu'en 1943. Deux classements historiques servirent de principales références à la désignation de ces ceux-ci: celui de Jules Laval de 1855 et le Classement du Comité d'Agriculture et de Viticulture de l'Arrondissement de Beaune de 1860. Ces données sur les AOC de 1936 ne sont pas utilisées dans la suite de l'article.

Table 5: **Nom, type et description des variables issues des AOC de 1936**

NOM	TYPE	DESCRIPTION
AOC361ab	<i>Caractère</i>	Libellé de l'appellation en 1936 (56 modalités)
AOC361v1	<i>Caractère</i>	Rang de la parcelle dans la hiérarchie des AOC (entre 0 et 5)

## 2.6 Enrichissement des lieux dits

Il s'agit ici d'inclure de l'information cadastrale à partir des sources [data.gouv.fr](https://data.gouv.fr). Nous utilisons le Plan Cadastral Informatisé Vecteur (Format EDIGÉO, <https://cadastre.data.gouv.fr/datasets/plan-cadastral-informatise>) téléchargé pour la Côte d'Or (21) le <2019-01-13 dim.>. License ouverte

Etalab. La difficulté avec les lieux dit est qu'ils doivent être croisés avec les communes car un même nom lieu dit peut être présent sur plusieurs communes. Comme la géométrie des lieux dits et des parcelles colle parfaitement, nous pouvons enrichir les données parcellaires directement par le centroïde. Ajout <2019-01-23 mer.>, des données communales, nous extrayons également les coordonnées des chefs-lieux pour calculer une distance à vol d'oiseaux, la population (peuvent être des sur-identifications sur le land use) et la distinction Côte de Beaune / Côtes de Nuits. Nous enregistrons également une shapefile MapCom qui permet de cartographier les contours communaux dans les figures.

Pour 4% des parcelles, aucun lieu dit n'a été apparié. Ces parcelles se concentrent sur les communes de Chenôve, Marsannay-la-Côte et Beaune (Corgoloin dans une moindre mesure). Ces "trous" apparaissent déjà dans le fichier source et ne sont donc pas un résultat de l'appariement. Ils semblent être des espaces bâtis sur la carte, mais ce n'est toujours pas confirmé par le MOS.

Table 6: **Nom, type et description des variables issues des lieux dits**

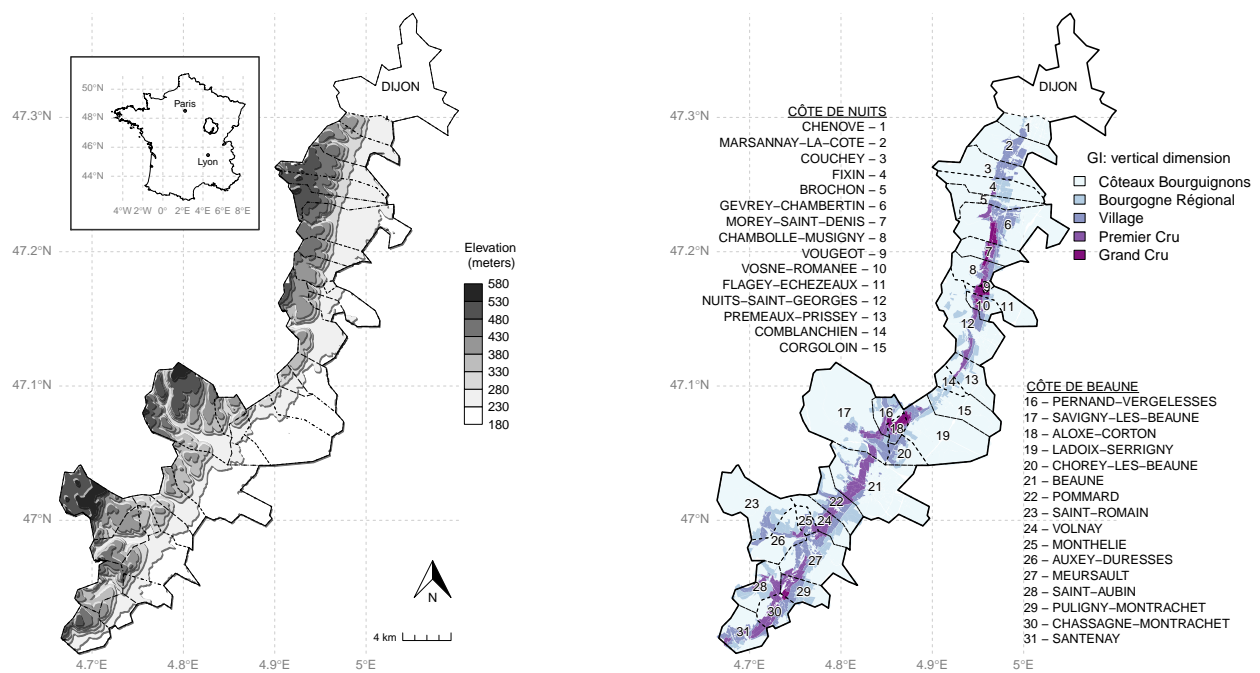
NOM	TYPE	DESCRIPTION
LIEUDIT	<i>Caractère</i>	Libellé du lieu dit de la parcelle (2691 modalités)
CLDVIN	<i>Caractère</i>	Identifiant du lieu dit de la parcelle (2691 modalités)
LIBCOM	<i>Caractère</i>	Libellé de la commune de la parcelle (31 modalités)
XCHF	<i>Numérique</i>	Latitude du chef-lieu de la commune (système Lambert 93)
YCHF	<i>Numérique</i>	Longitude du chef-lieu de la commune (système Lambert 93)
ALTCOM	<i>Numérique</i>	Altitude du point culminant de la commune (mètre)
SUPCOM	<i>Caractère</i>	Superficie de la commune de la parcelle (hectare)
POPCOM	<i>Numérique</i>	Population de la commune de la parcelle en 2015 (millier d'hab)
CODECANT	<i>Caractère</i>	Identifiant du canton d'appartenance (2 caractères)
REGION	<i>Caractère</i>	Region viticole (CDB Côte de Beaune, CDN Côte de Nuits)

## 2.7 Enregistrement de la base

Pour l'instant, on est à moins de 500 Mo. C'est le résultat de tous ces traitements que l'on va mettre sur le *dataverse* de l'INRA. La partie reproductible du data paper commence ici.



Figure 1: Vignobles de la Côte d'Or, topographie et appellations d'origine contrôlées  
blabla.



## 3 Statistiques descriptives

### 3.1 Sélection des données

Nous commençons l'analyse par le chargement du fichier `GeoRas.Rda` issue du serveur data de l'INRA que l'utilisateur doit télécharger puis placer dans un répertoire `Inter/` dans le répertoire de travail de R (soit le répertoire renvoyé par la commande `getwd()`). La première procédure à exécuter est présentée ci-dessous, elle consiste à :

- Recoder les codes communaux selon le gradient Nord-Sud
- Calculer la distance de chaque parcelle au chef lieu de sa commune
- Centrer et réduire la variable sur les rayonnements solaires
- Recoder la variable exposition en 8 catégories
- Re-projeter les coordonnées dans le système WGS84
- Enlever les valeurs manquantes de la base de données

---

```
library(sp) ; load("Inter/GeoRas.Rda")
tmp <- unique(Geo.Ras$LIBCOM[order(Geo.Ras$YCHF, decreasing= TRUE)])
Geo.Ras$LIBCOM <- factor(Geo.Ras$LIBCOM, levels= tmp)
Geo.Ras$DISTCHF <- sqrt((Geo.Ras$XL93- Geo.Ras$XCHF* 100)^2
+ (Geo.Ras$YL93- Geo.Ras$YCHF* 100)^2)
Geo.Ras$RAYAT <- (Geo.Ras$SOLAR- mean(Geo.Ras$SOLAR, na.rm= TRUE))/
sd(Geo.Ras$SOLAR, na.rm= TRUE)
Geo.Ras$EXPO <- cut(Geo.Ras$ASPECT,
breaks= c(-2, 45, 90, 135, 180, 225, 270, 315, 360))
GR84 <- spTransform(Geo.Ras, CRS("+proj=longlat +ellps=WGS84"))
dd <- coordinates(GR84) ; Geo.Ras$X= dd[, 1] ; Geo.Ras$Y= dd[, 2]
dim(Reg.Ras <- subset(Geo.Ras, !is.na(AOC1b) & !is.na(DEM) & !is.na(DESCR)
& !is.na(RUD) & !is.na(AOC361ab) & !is.na(REGION)))
```

---

```
[1] 59113      73
```

Nous obtenons une base de données avec 59 113 observations. Le principal critère de sélection des parcelles provient de la limitation aux parcelles ayant au moins une AOC. Sur la zone, nous avons 60 632 (= 110 350 – 49 718) parcelles dans ce cas, ce qui signifie que le retrait des valeurs manquantes cause la perte de seulement 1 519 parcelles (= 60 632 – 59 113).

### 3.2 Distribution des AOC

Nous pouvons désormais présenter plus en détails la distribution des AOC sur les parcelles de la zone. Il y a en particulier 2 informations typiquement reportées sur les étiquettes des bouteilles de vins. Une première information est de type verticale, elle consiste à mentionner le niveau de l'AOC dans la hiérarchie régionale. Cette information est contenue dans la variable `AOC`. La deuxième information est de type horizontale, avec la mention de la commune d'appartenance de la parcelle, sans qu'il n'y ai de hiérarchie entre les communes. Cette information est contenue dans la variable `LIBCOM`. Le code ci-dessous permet de reproduire la Figure XX de Ay (2019) avec les pourcentages de chaque niveau d'AOC au sein de chaque commune.

```

library(lattice) ; library(RColorBrewer)
fig.dat <- aggregate(model.matrix(~0+ factor(Reg.Ras$AOC))*
  Reg.Ras$AREA/ 1000, by= list(Reg.Ras$LIBCOM), sum)
names(fig.dat) <- c("LIBCOM", "BGOR", "BOUR", "VILL", "PCRU", "GCRU")
fig.dat$LIBCOM <- factor(fig.dat$LIBCOM, lev= rev(levels(fig.dat$LIBCOM)))
fig.crd <- t(apply(fig.dat[, -1], 1, function(t) cumsum(t)- t/2))
fig.lab <- round(t(apply(fig.dat[, -1], 1, function(t) t/ sum(t)))* 100)
my.pal <- brewer.pal(n= 9, name = "BuPu")[ 2: 8]
barchart(LIBCOM~ BGOR+ BOUR+ VILL+ PCRU+ GCRU, xlim= c(-100, 10200),
  xlab="Surfaces sous appellation d'origine contrôlée (hectare)",
  data= fig.dat, horiz= T, stack= T, col= my.pal, border= "black",
  par.settings= list(superpose.polygon= list(col= my.pal)),
  auto.key= list(space= "top", points= F, rectangles= T, columns= 5,
    text=c("Coteaux b.", "Bourgogne",
      "Village", "Premier cru", "Grand cru")),
  panel=function(x, y, ...) {
    panel.grid(h= 0, v = -11, col= "grey60")
    panel.barchart(x, y, ...)
    ltext(fig.crd, y, lab= ifelse(fig.lab> 0, fig.lab, ""))})

```

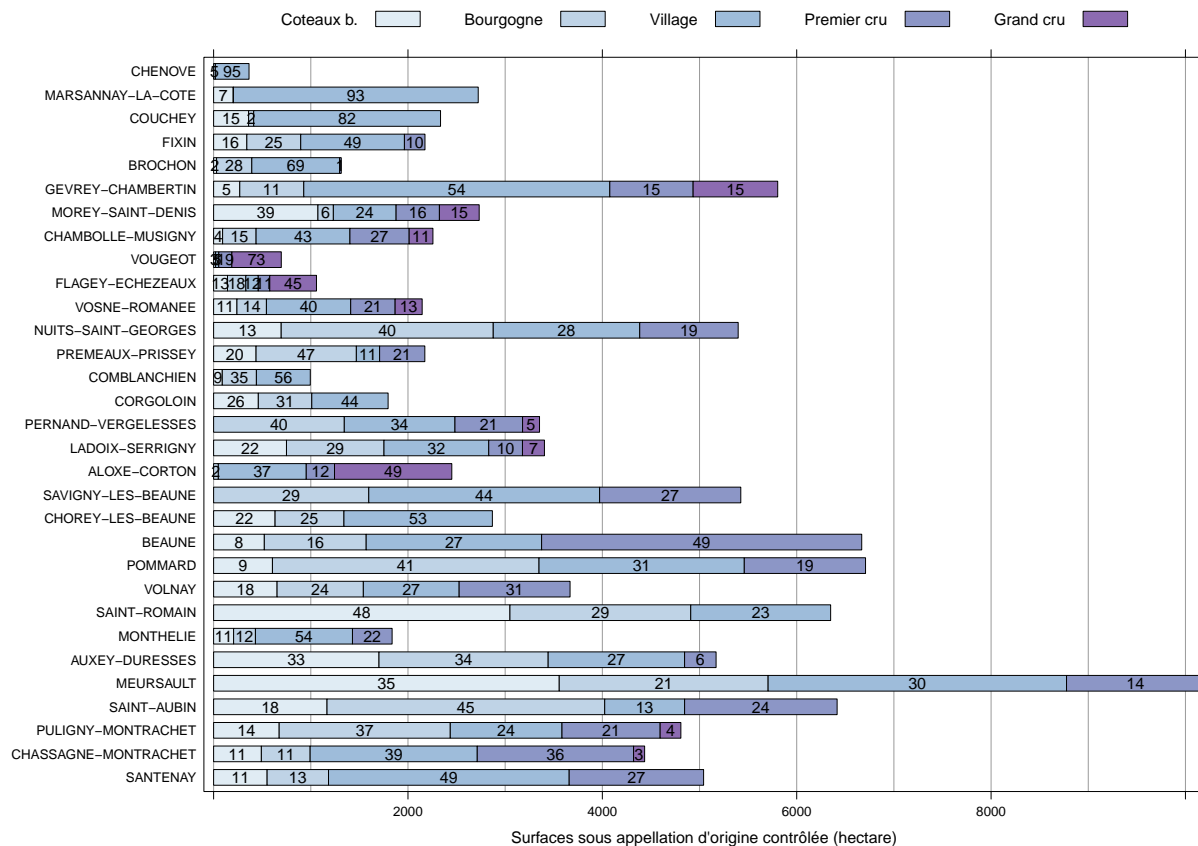


Figure 2: Distribution des niveaux des AOC entre les communes

### 3.3 La pyramide des AOC

Les AOC en Bourgogne sont souvent représentés sous forme pyramidale (voir par exemple <https://www.vins-bourgogne.fr/plan-de-site/classification-des-appellations,2314,12208.html>) en accord avec le principe qu'en montant dans la hiérarchie les surfaces de vigne concernées deviennent moins importantes. Cette structure pyramidale est *a priori* observable à l'échelle régionale mais pas à l'échelle départementale. Étant donné que nous travaillons sur un sous-échantillon des vignes de Bourgogne, limité au département de la Côte d'Or, nous n'obtenons pas cette structure pyramidale comme en atteste la Figure ci-dessous. Il apparaît que les AOC inférieures (Bourgogne régional, Coteaux Bourguignons) sont sous-représentées, à la fois dans la Côte de Beaune et la Côte de Nuits.

```
ddd <- aggregate(Reg.Ras$AREA/ 10000,
                 by= list(Reg.Ras$AOC, Reg.Ras$REGION), sum, na.rm= TRUE)
names(ddd) <- c("AOC", "REGION", "SURFACES")
ddd$SURFACES[ddd$REGION== "CDB"] <- -ddd$SURFACES[ddd$REGION== "CDB"]
library(ggplot2)
ggplot(ddd, aes(x= AOC, y= SURFACES, fill= REGION))+
  geom_bar(data= subset(ddd, REGION== "CDB"), stat= "sum")+
  geom_bar(data= subset(ddd, REGION== "CDN"), stat= "sum")+
  coord_flip()+ theme_bw()+ ylab("Surfaces en hectares")+
  xlab("Niveau d'Indication Géographique")
```

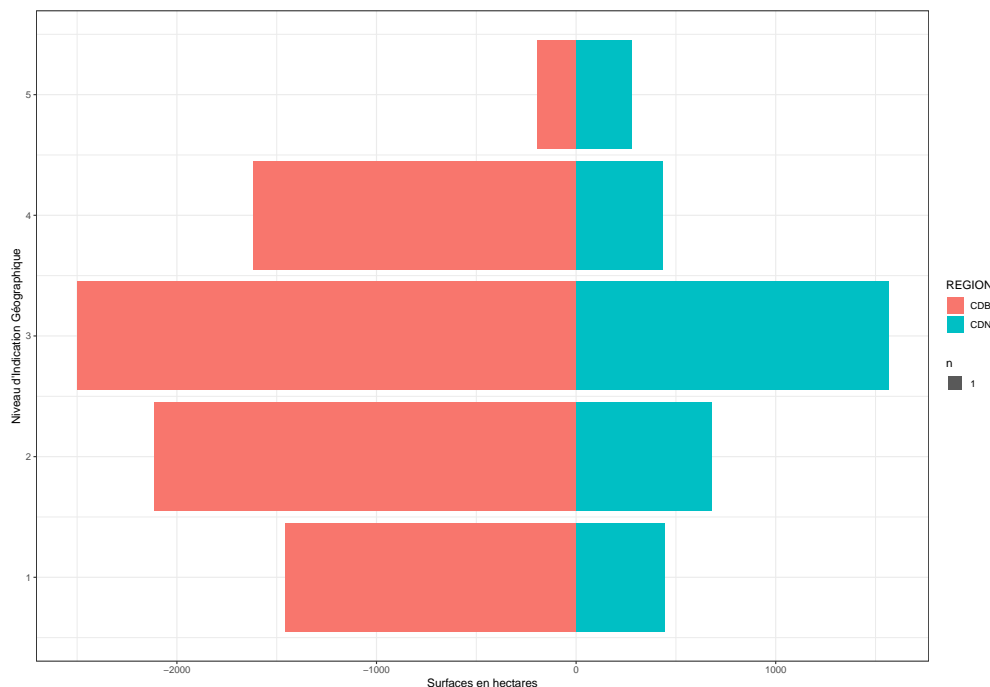


Figure 3: Surfaces des différents niveaux d'AOC au sein de la Côte d'Or

Nous observons des distributions presque symétriques avec un niveau village majoritaire. Notons que la Côte de Nuits est relativement privilégiée par rapport à la Côte de Beaune en termes de grands crus mais compte moins de surfaces totales.

### 3.4 Tableau des variables utilisées

Le détails des variables utilisées dans l'analyse économétrique est détaillé dans ce qui suit. Nous observons des surfaces faible (moyenne de 0.2 ha), des altitudes comprises entre 200 et 500 m (moyenne à 286), des pentes entre 0 et 37 (moyenne à 5.75) et des radiations solaires entre 581 000 et 1.2 millions de Joules. Les vignobles sous AOC sont globalement orientés à l'Est.

```
Stat.Ras <- data.frame(Reg.Ras@data, model.matrix(~0+ factor(Reg.Ras$AOC)),
                      model.matrix(~ 0+ factor(Reg.Ras$EXPO)))
names(Stat.Ras)[74: 78] <- paste0("AOC", 1: 5)
names(Stat.Ras)[79: 86] <- paste0("EXPO", 1: 8)
Stat.Ras$AREA <- Reg.Ras$AREA/ 1000
Stat.Ras$DEM <- Reg.Ras$DEM/ 1000
Stat.Ras$SOLAR <- Reg.Ras$SOLAR/ 1000000
lab <- c(AREA= "Surface [1000 m$^2$]", DEM= "Altitude [1000 m]",
        SLOPE= "Pente [degrés]", SOLAR= "Radiation solaire [millions J]",
        X= "Longitude [degrés]", Y= "Latitude [degrés]",
        AOC1= "Niveau AOC Coteaux", AOC2= "Niveau AOC Régional",
        AOC3= "Niveau AOC Village", AOC4= "Niveau AOC Premier Cru",
        AOC5= "Niveau AOC Grand Cru",
        EXPO1= "Exposition [$0-45$]", EXPO2= "Exposition [$45-90$]",
        EXPO3= "Exposition [$90-135$]", EXPO4= "Exposition [$135-180$]",
        EXPO5= "Exposition [$180-225$]", EXPO6= "Exposition [$225-270$]",
        EXPO7= "Exposition [$270-315$]", EXPO8= "Exposition [$315-360$]")
library(stargazer)
stargazer(Stat.Ras[, names(lab)], covariate.labels=lab, font.size= "small",
          column.sep.width= "0pt", float= T, digit.separate= c(0, 3),
          title= "\\textbf{Statistiques descriptives des variables utilisées}")
```

Table 7: **Statistiques descriptives des variables utilisées**

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Surface [1000 m <sup>2</sup> ]	59113	1.908	3.399	0.000	0.517	2.178	177.200
Altitude [1000 m]	59113	0.286	0.056	0.210	0.241	0.319	0.505
Pente [degrés]	59113	5.772	5.478	0.000	1.556	8.747	36.970
Radiation solaire [millions J]	59113	1.060	0.049	0.581	1.048	1.076	1.230
Longitude [degrés]	59113	4.837	0.104	4.665	4.740	4.955	5.003
Latitude [degrés]	59113	47.060	0.110	46.900	46.980	47.170	47.300
Niveau AOC Coteaux	59113	0.164	0.370	0	0	0	1
Niveau AOC Régional	59113	0.229	0.420	0	0	0	1
Niveau AOC Village	59113	0.428	0.495	0	0	1	1
Niveau AOC Premier Cru	59113	0.147	0.354	0	0	0	1
Niveau AOC Grand Cru	59113	0.032	0.177	0	0	0	1
Exposition [0 – 45]	59113	0.046	0.210	0	0	0	1
Exposition [45 – 90]	59113	0.186	0.389	0	0	0	1
Exposition [90 – 135]	59113	0.362	0.481	0	0	1	1
Exposition [135 – 180]	59113	0.212	0.409	0	0	0	1
Exposition [180 – 225]	59113	0.100	0.300	0	0	0	1
Exposition [225 – 270]	59113	0.044	0.206	0	0	0	1
Exposition [270 – 315]	59113	0.030	0.170	0	0	0	1
Exposition [315 – 360]	59113	0.021	0.142	0	0	0	1

## 4 Modèle statistique

### 4.1 Estimation du modèle

Nous abordons désormais l'estimation du modèle statistique, dans le processus de spécification est présenté plus extensivement dans [Ay \(2019\)](#). Il s'agit d'un modèle ordonné additivement semi-paramétrique (OGAM) qui prend en compte la structure hiérarchique des AOC de la zone. Nous supposons en effets que les désignations des AOC suivent une règle de décision basée sur une variable latente de qualité biophysique des vignes qui franchit des seuils différents selon les communes d'appartenance des parcelles. Le modèle est estimé avec la fonction `gam` du package `mgcv` comme décrit récemment par [Wood et al. \(2016\)](#). Le manuel d'utilisation de l'auteur du package [Wood \(2017\)](#) contient de nombreux détails méthodologiques.

Notons qu'en préalable nous opérons un regroupement des unités géologiques et pédologiques afin de pouvoir les inclure comme des variables binaires dans le modèle. En effet, un nombre trop faible d'observation au sein d'une unité géologique ou pédologique diminue sensiblement la précision de l'estimation et peut poser des problèmes de convergence de l'algorithme de régression pénalisée. Nous choisissons de ne retenir que les unités qui contiennent plus de 1 000 parcelles, d'autres valeurs ont été testées sans que cela change les résultats. Notons également que le modèle que nous estimons opère un lissage spatial fin (avec un nombre maximal de degré de liberté effectifs de 900 sur les termes spatiaux) ce qui implique une procédure estimation assez longue (environ 9 heures avec un processeur Intel Core i7-7820HQ CPU 2.90 GHz x8 et 64 Go of RAM). D'autres modèles plus parcimonieux sont également présents dans l'objet `gamod.Rda` téléchargeable sur le serveur data de l'INRA. L'ensemble des modèles est présenté en détails dans [Ay \(2019\)](#).

---

```
Reg.Ras$NOTATION <- factor(Reg.Ras$NOTATION)
tmp <- table(Reg.Ras$NOTATION)< 1000
Reg.Ras$GEOL <- factor(
  ifelse(Reg.Ras$NOTATION %in% names(tmp[ tmp]), "0AREF",
    as.character(Reg.Ras$NOTATION)))
Reg.Ras$NOUC <- factor(Reg.Ras$NOUC)
tmp <- table(Reg.Ras$NOUC)< 1000
Reg.Ras$PEDO <- factor(
  ifelse(Reg.Ras$NOUC %in% names(tmp[tmp]), "0AREF",
    as.character(Reg.Ras$NOUC)))
library(mgcv) ; load("Inter/gamod.Rda")
## system.time(
##   gam900 <- gam(AOC~ 0+ LIBCOM+ EXPO+ GEOL+ PEDO
##               + s(DEM)+ s(SLOPE)+ s(RAYAT)+ s(X, Y, k= 900)
##               , data= Reg.Ras, family= ocat(R= 5))
## )
## utilisateur      système      écoulé
##   32271.43        93.78    32366.00
## save(gam900, file= "Inter/gam900.Rda")
anova(gamod$gam900)
```

---

Family: Ordered Categorical(-1,5.34,14.01,20.99)

Link function: identity

Formula:

AOC ~ 0 + LIBCOM + EXPO + GEOL + PEDO + s(DEM) + s(SLOPE) + s(RAYAT) +  
s(X, Y, k = 900)

Parametric Terms:

	df	Chi.sq	p-value
LIBCOM	31	1363	<2e-16
EXPO	7	131	<2e-16
GEOL	14	441	<2e-16
PEDO	13	388	<2e-16

Approximate significance of smooth terms:

	edf	Ref.df	Chi.sq	p-value
s(DEM)	8.81	8.98	867	<2e-16
s(SLOPE)	7.72	8.61	190	<2e-16
s(RAYAT)	7.33	8.38	531	<2e-16
s(X,Y)	841.42	870.01	86597	<2e-16

Les résultats présentés ci-dessous sont directement accessibles dans le fichier XX. Ils concernent la significativité statistique des différentes variables au regard des statistiques de  $\chi^2$ . Les coordonnées géographiques sont les variables explicatives les plus importantes, suivies des indicatrices communales, de l'altitude, du rayonnement solaire, de la géologie, de la pédologie, de la pente et enfin de l'exposition. Ce modèle produit 90% de bonnes prédictions avec un pseudo  $R^2$  égal à 0.75, comme le code ci-dessous le fait apparaître.

---

```
pcgp <- function(x){
  sum(diag(table(cut(x$line, c(-Inf, x$family$getTheta(TRUE), Inf)),
    x$model[, 1])))/ nrow(x$model)* 100
}
pcgp(modMontr$gam8)
summary(modMontr$gam8)
```

---

[ 1 ] YOP

## 4.2 Variables biophysiques

Nous pouvons alors représenter les effets marginaux des variables biophysiques sur la variable latente de qualité continue des parcelles de vigne. La fonction `plot` par défaut du package `mgcv` permet de représenter graphiquement ces effets en fixant toutes les autres variables explicatives du modèle à leurs moyennes. Notons également dans le graphique ci-dessous que les effets ont une moyenne normalisée à 0 car le niveau des effets n'est pas identifiable semi-paramétriquement.

---

```
plot(gamod$gam700, page= 1, scale= 0)
```

---

Des Figures plus détaillées qui contiennent en particulier les effets associés aux autres modèles sont reportées dans l'article associé. La structure des effets reste cependant robuste à la spécification, elle reste proche de ce qui est observé dans la Figure XX. L'altitude et la pente ont des effets en U inversé, le rayonnement solaire a un effet linéaire à proximité de sa moyenne (la présence de valeur extrême entraîne de forte non linéarité qui concernent que très peu d'observations). Enfin, les effets spatiaux semblent se structurer dans une relation de centre/ périphérie.



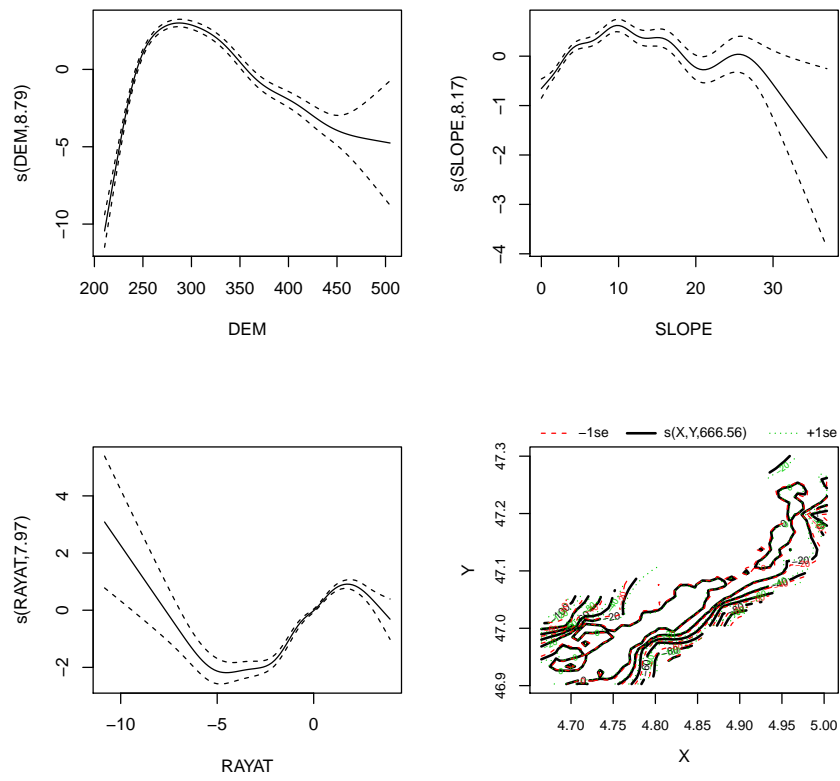


Figure 4: Effets semi-paramétriques de la topographie et de la localisation

### 4.3 Effets communaux

Les coefficients associés aux effets fixes communaux sont d'un intérêt particulier car ils correspondent à la partie historique des désignations AOC, ils représentent la partie de la variable latente de qualité qui n'est pas expliquée par les caractéristiques biophysiques des vignes. Comme présenté dans [Ay \(2019\)](#), la finesse du lissage spatial permet de contrôler les effets de *terroir* potentiellement non captés par les variables topographiques, géologiques et pédologiques. Cette interprétation des effets fixes communaux fait écho à certains travaux historiques pour lesquels nos résultats offre une sorte de confirmation statistique. En effet, Christophe Lucand dans [Wolikow and Jacquet \(2011\)](#) évoque l'existence d'une hiérarchie implicite des communes "qui ne détermine cependant en rien la réalité des zones d'approvisionnement concernées. Il s'agit plutôt d'identifications commerciales communes, investies d'un plus ou moins grand capital symbolique hérité. Ce capital symbolique hérité attribut un prestige plus ou moins grand à certaines communes ou propriétaires particulier." Les effets fixes que nous estimons peuvent alors être vus comme des mesures de ce capital symbolique. De manière complémentaire, [Jacquet \(2009\)](#) étudie la structuration des syndicats de viticulteurs qui s'opère quasi-exclusivement à l'échelle communale et mentionne le fait que (p.193) "plus l'appellation requise se calque sur le syndicat qui la défend, plus elle a de chance d'émerger et d'être délimitée strictement". Les effets fixes communaux peuvent donc également mesurer l'action des syndicats, qui apparaît ainsi avoir une forte inertie historique.

---

```
library(latticeExtra) ; yop <- summary(gamod$gam900)
plogi <- function(x) exp(x/ sqrt(2))/( 1+ exp(x/ sqrt(2)))
cf <- yop$p.coeff[ 4: 31]- mean(yop$p.coeff[ 4: 31])
se <- yop$se[ 4: 31]
names(gamod$gam900$coef[ 4: 31])
zz <- data.frame(LIBCOM= substr(names(gamod$gam900$coef[ 4: 31]), 7, 30),
                REGION= c(rep("tomato", 12), rep("chartreuse", 16)),
                OS= 2* plogi(cf)- 1,
                OSi= 2* plogi(cf- 1.5* se)- 1,
                OSa= 2* plogi(cf+ 1.5* se)- 1)

foo_key <- list(x = .35, y = .95, corner = c(1, 1),
               text = list(c("Côte de Beaune", "Côte de Nuits")),
               rectangle = list(col = c("chartreuse", "tomato")))

segplot(reorder(factor(LIBCOM), OS)~ OSi+ OSa,
        length= 5, draw.bands= T,
        key= foo_key,
        data= zz[order(zz$OS), ], center= OS, type= "o",
        col= as.character(zz$REGION[order(zz$OS)]),
        unit = "mm", axis = axis.grid, col.symbol= "black", cex= 1,
        xlab= "Mesure de supériorité ordinale et intervalles à 10 %")
```

---

Nous traduisons les coefficients estimés en mesures de supériorité ordinale comme cela est présenté dans [Agresti and Kateri \(2017\)](#). Les valeurs obtenues s'interprètent alors plus intuitivement : de telles mesures en abscisses correspondent aux probabilités qu'une parcelle de la commune reportée en ordonnées soit mieux classée qu'une parcelle aux caractéristiques biophysiques identiques mais localisée dans une commune au hasard. Les communes relativement favorisées par les AOC apparaissent en haut de la Figure et les communes relativement défavorisées en bas. Les intervalles de confiance qui encadrent les valeurs moyennes sont différents de ceux reportés dans [Ay \(2019\)](#). Au lieu de représenter l'incertitude quand à la spécification des effets spatiaux, ils représentent ici l'incertitude associée à l'estimation des effets fixes par la procédure statistique. L'ordre de grandeur reste toutefois similaire.

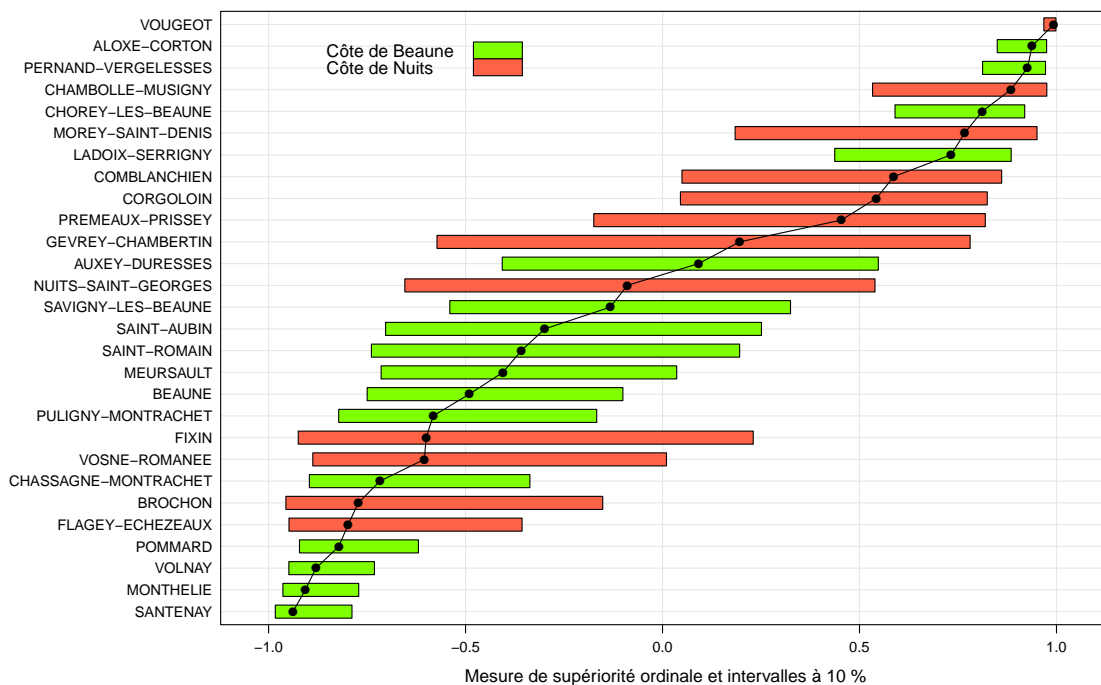


Figure 5: Effets communaux sur la classification AOC des parcelles

#### 4.4 Prédiction continue

En complément de l'estimation d'effets communaux, un intérêt important de cette modélisation économétrique est la possibilité de prédire un score continue pour chaque parcelle selon ses caractéristiques biophysiques. Cette possibilité de classification continue est vraiment l'originalité du travail présentée dans cet article, peu développé dans l'article académique. Notons que cette classification continue des parcelles est directement issue des AOC qui existent aujourd'hui et ne se base donc pas sur des appréciations subjectives sur ce qui fait la qualité d'une vigne ou d'un vin. Les appréciations subjectives se limitent au choix du modèle économétrique (le lissage spatial en particulier, mais aussi la spécification des effets des autres variables explicatives). Le principe de reproductibilité devrait permettre l'utilisation de méthodes alternatives pour produire des classifications alternatives.

Nous proposons ici simplement deux sorties principales du modèle économétrique : la valeur latente de la qualité avec les effets fixes communaux (ce qui signifie que les prédictions ne sont pas corrigées) et la valeur latente de la qualité sans les effets fixes communaux (les moyennner permet de corriger les prédictions). Le code suivant présente la mise en oeuvre des prédictions et leur normalisation pour qu'ils soient distribués entre 0 et 100. Nous les appellerons alors des scores, bruts ou corrigés.

```
Prd.Ras <- subset(Geo.Ras, !is.na(AOC1b))
Prd.Ras$GEOL <- ifelse(Prd.Ras$NOTATION %in% levels(Reg.Ras$GEOL),
  as.character(Prd.Ras$NOTATION), "0AREF")
Prd.Ras$PEDO <- ifelse(Prd.Ras$NOUC %in% levels(Reg.Ras$PEDO),
  as.character(Prd.Ras$NOUC), "0AREF")
## prd <- predict(gamod$gam900, newdata= Prd.Ras@data, type= "terms")
Prd.Ras$LTraw <- rowSums(prd, na.rm= TRUE)
```

```

Prd.Ras$LTcor <- mean(prd[, 1], na.rm= T)+ rowSums(prd[, -1], na.rm= T)
unini <- function(x) (x- min(x))/( max(x)- min(x))
Prd.Ras$UFraw <- round(unini(Prd.Ras$LTraw)* 100, 2)
Prd.Ras$UFcor <- round(unini(Prd.Ras$LTcor)* 100, 2)

```

---

Pour le graphique, on a besoin des codes en annexe qui sont également disponibles dans le fichiers myFcts.R disponible sur le dataverse de l'INRA. Annexe 4.

balbla clearpage

balbla clearpage

balbla clearpage

balbla clearpage

---

```

library(ggplot2) ; library(plyr) ; source("./myFcts.R")
NVA <- c("Coteaux b.", "Bourgogne", "Village", "Premier cru", "Grand cru")
names(NVA) <- 1: 5
cc <- rbind(
  data.frame(AOC= revalue(factor(Prd.Ras$AOC), NVA),
    Score= Prd.Ras$UFraw,
    Pr= "Non corrigé : maintient des effets communaux"),
  data.frame(AOC= revalue(factor(Prd.Ras$AOC), NVA),
    Score= Prd.Ras$UFcor,
    Pr= "Corrigé : suppression des effets communaux"))
ggplot(cc, aes(factor(AOC), Score, fill= Pr))+
  geom_split_violin()+
  ylab("Score de qualité des vignes (échelle de 0 à 100)")+
  ylim(40, 100)+ theme_minimal()+ xlab("")+
  geom_split_violin(draw_quantiles = c(0.25, 0.5, 0.75))+
  theme(legend.justification=c(0, 1), legend.position=c(0, 1),
    legend.title = element_blank())

```

---

Un premier élément du graphique précédent est que les caractéristiques biophysiques permettent de bien discriminer les différents niveaux d'AOC car les distributions se chevauchent peu. Un deuxième élément tient de la correction qui ne produit pas les même conséquences selon le niveau d'AOC. Pour les Coteaux bourguignons, les Villages et les Grands Crus, la correction maintien globalement les scores médians, et leur variance hormis pour le niveau Village où elle augmentent sensiblement. À l'inverse, pour les Bourgogne régionaux et les Premiers crus, les scores médians sont revus à la hausse, ce qui indique la présence de parcelles dans ces niveaux qui sont sous-estimées du fait de leur communes d'appartenance. Cet effet est particulièrement important pour les Premier cru comme on peut le voir dans le déplacement du mode de la distribution.

## 4.5 Agrégation par lieux dits

Pour faciliter la consultation des résultats de la modélisation, nous agrégeons les scores prédits, ce qui passe par un recodage des délimitations. Nous utilisons les lieux dits administratifs pour localiser plus précisément les parcelles en niveaux d'AOC Coteaux bourguignon, Bourgogne et Village pour lesquels la mention de la parcelle n'est pas reporté sur l'étiquette. Par contre les producteurs ont cette information, et les usagers pourraient la demander. Il s'agit également ici de renommer les premiers crus pour qu'ils soient plus lisibles. (On pourrait le mettre ce travail en préalable, à voir).

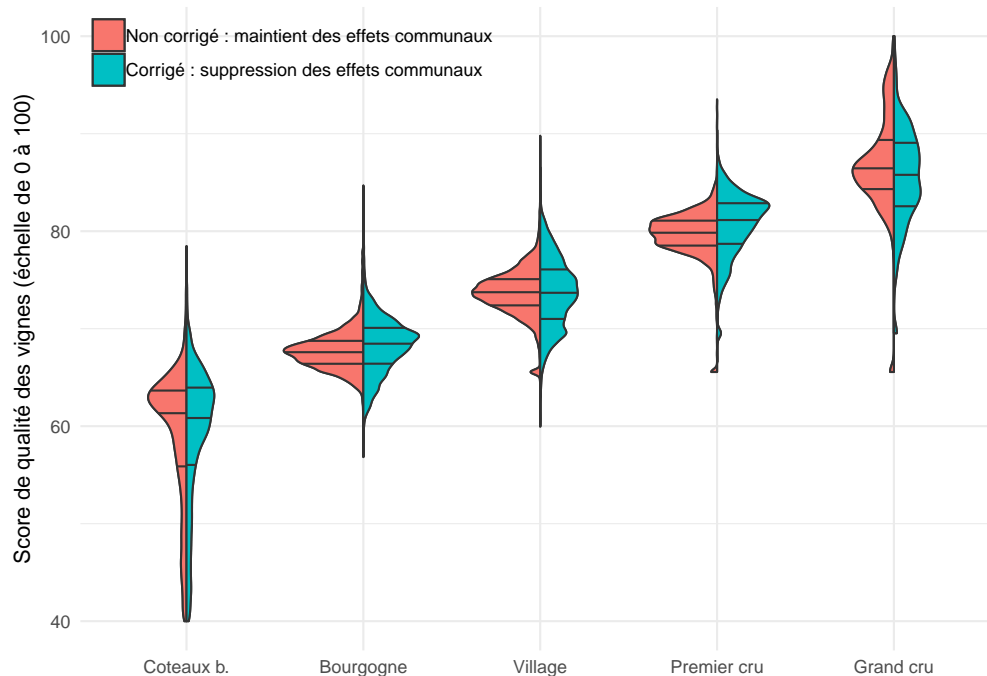


Figure 6: **Distribution des qualités prédites avec et sans correction**

```
Prd.Ras$NIVEAU <- as.character(revalue(factor(Prd.Ras$AOC), NVA))
Prd.Ras$NAME <- ifelse(Prd.Ras$AOC== 5, as.character(Prd.Ras$AOC1b),
                      ifelse(Prd.Ras$AOC< 4, as.character(Prd.Ras$LIEUDIT), NA))
for (i in 1: nrow(Prd.Ras)){
  if (is.na(Prd.Ras$NAME[ i])){
    Prd.Ras$NAME[ i] <- substr(Prd.Ras$AOC1b[ i],
                              regexpr(" cru+", Prd.Ras$AOC1b[ i],
                                         perl= T)+ 5,
                              nchar(as.character(Prd.Ras$AOC1b[ i])))
  } else {(Prd.Ras$NAME[ i])}
}
Prd.Ras$Concat <- paste0(Prd.Ras$AOC, Prd.Ras$LIBCOM, Prd.Ras$NAME)
length(unique(Prd.Ras$Concat))
```

---

[1] 2391

Ainsi, des quelques 60 000 parcelles cadastrales utilisées dans la modélisation, nous obtenons environ 2 400 localités, qui correspondent peu ou prou aux lieux dits administratifs.

Nous allons désormais regrouper la géographie des parcelles selon la variable Contat tout juste créée. Les scores sont reportés au niveau des nouvelles localités par moyenne pondérée par la surface de chaque parcelle qui l'a compose. Nous calculons également la position de chaque localité dans la hiérarchie continue issue de la modélisation, ce qui permet de présenter les 10 localités les mieux notées. Nous utilisons pour cela les packages XX.

---

```

library(data.table) ; Prd.Dtb <- data.table(Prd.Ras@data)
Dat.Ldt <- Prd.Dtb[, .(LIBCOM= LIBCOM[ 1], NOM= NAME[ 1],
                     NIVEAU= NIVEAU[ 1],
                     SURFACE_ha= round(sum(AREA)/ 1e5, 2),
                     SCORE_brut= round(weighted.mean(UFraw, AREA), 2),
                     SCORE_corrige=round(weighted.mean(UFcor, AREA), 2)),
                     by= Concat]
library(rgdal) ; library(rgeos) ; library(maptools)
tmp_geo <- gBuffer(Prd.Ras, byid= TRUE, width= 0)
Poly.ldt <- unionSpatialPolygons(tmp_geo, Prd.Ras$Concat)
Poly.ldt$Concat <- as.character(row.names(Poly.ldt))
Poly.Ras <- merge(Poly.ldt, Dat.Ldt, by= "Concat")
Poly.Ras$RANG_brut<- round(rank(Poly.Ras$SCORE_brut)/ nrow(Poly.Ras)*100,2)
Poly.Ras$RANG_corrige <- round(rank(Poly.Ras$SCORE_corrige)/
                                nrow(Poly.Ras)*100,2)
head(Poly.Ras@data[order(Poly.Ras$RANG_corrige, decreasing= T), c(3, 4, 6, 7)], n= 10)
Poly.Ras$NIVEAU <- factor(Poly.Ras$NIVEAU, levels= NVA)

```

---

	NOM	NIVEAU	SCORE_brut	SCORE_corrige
2364	Chambertin	Grand cru	94.22	94.11
2363	Grands-Echezeaux	Grand cru	87.73	90.76
2384	Montrachet	Grand cru	88.72	90.69
2381	Bâtard-Montrachet	Grand cru	87.73	89.68
2361	Montrachet	Grand cru	87.05	89.58
2362	Echezeaux	Grand cru	86.13	89.12
2369	Latricières-Chambertin	Grand cru	88.73	88.53
2371	Mazoyères-Chambertin	Grand cru	88.71	88.50
2359	Bâtard-Montrachet	Grand cru	85.80	88.30
2010	La Combe d'Orveau Premier	cru	91.01	87.83

Sans surprise, les Grands Crus arrivent en haut de la hiérarchie autant de la Côte de Nuits (Chambertin, Grands-Echezeaux) que de la Côte de Beaune (Montrachet, Bâtard-Montrachet). Notons tout de même qu'un Premier Cru arrive en 10<sup>ième</sup> position, ce qu'il veut dire qu'il dépasse les 2/3 des Grands crus. Le Premier Cru La Combe d'Orveau se trouve sur la commune de Chambolle Musigny qui n'apparaît pas pourtant si désavantagé selon la Figure XX. Cela indique que la haute classification de ce premier cru (en particulier au-dessus du Grand cru Musigny situé sur la même commune) provient des caractéristiques biophysiques et non de la correction communale. Plus étonnant, la Romanée-conti qui apparaît souvent parmi les vins les plus chers du monde (<https://www.wine-searcher.com/most-expensive-wines>) n'apparaît qu'en 26<sup>ième</sup> position (elle est tout de même dans les 2 % meilleurs). On peut penser que la situation de monopole peut expliquer la fort prix indépendamment des caractéristiques biophysiques.

Nous enregistrons ensuite les résultats dans une base de données géographique de type sf qui pourra directement être utilisée dans l'application *Shiny*. Ces résultats sont accessibles sur le serveur data de l'INRA à l'adresse XX. Attention la ligne sur les prédictions est longue (5 minutes) à tourner.

---

```

library(sf)
Poly.Ras <- st_as_sf(Poly.Ras)
Poly.Ras <- st_transform(Poly.Ras, crs= 4326)
save(Poly.Ras, file= "Inter/PolyRas.Rda")

```

---

Sauvegarde disponible sur le serveur de l'INRA.

## 5 Application *Shiny*

### 5.1 Cartographie dynamique

Ceux qui sont intéressés par l'utilisation de l'application en local peuvent commencer ici.

---

```
library(RColorBrewer) ; library(mapview) ; library(sf)
load("Inter/PolyRas.Rda")
AocPal <- brewer.pal(5, "BuPu")
mapviewOptions(basemaps= c("Esri.WorldImagery", "OpenStreetMap",
                           "OpenTopoMap", "CartoDB.Positron"),
               raster.palette= colorRampPalette(brewer.pal(9, "Greys")),
               vector.palette= colorRampPalette(brewer.pal(9, "YlGnBu")),
               na.color= "magenta", layers.control.pos = "topleft")
map <- mapview(Poly.Ras, zcol= "NIVEAU", label= Poly.Ras$NOM,
               layerId= Poly.Ras$Concat, alpha.regions= .5,
               col.regions = AocPal, color= "white", legend.opacity= .5,
               popup = popupTable(Poly.Ras, feature.id= FALSE,
                                   zcol= names(Poly.Ras)[ -1]))
## mapshot(addLogo(map, "http://www7.inra.fr/fournisseurs/images/logo.jpg",
##                  width = 200, height = 100),
##          url = paste0(getwd(), "/Application/CotePrd.html"),
##          file = paste0(getwd(), "/Application/CotePrd.png"),
##          remove_controls = c("homeButton", "layersControl"))
```

---

### 5.2 Lancer l'application localement

Le détails des calculs est reporté en Annexe XX, fichier sur Github. `runApp()`, sinon sur le serveur <https://geoind.shinyapps.io/application/>

---

```
library(shiny) ; library(shinydashboard) ; library(shinyjs)
library(leaflet) ; library(maptools) ; library(ggplot2)
Pts.Crd <- st_centroid(Poly.Ras)

source("ui.R")
source("server.R")

shinyApp(ui, server)
```

---

### 5.3 Utilisation

Que l'application soit lancée localement ou à distance, le fonctionnement est identique.

Détails des codes reportés en Annexe XX, fichier sur Github

Utilisation.

### 5.4 Développements à venir

- Régler le LIBCOM = NA sur la commune de Marsannay (pres des echzeaux)



- Diminuer la taille du graphique pour fitter le smart phone
- Faire un onglet pour des croisements exhaustifs, réfléchir à

l'interaction avec la carte et le graphique

- Pour le rank, mettre la position plutôt que la part
- Réfléchir à l'emplacement d'enregistrement de la base

<https://www.shinyapps.io/admin/#/dashboard><http://shiny.rstudio.com/articles/shinyapps.html>

---

```
library(rsconnect)
rsconnect::setAccountInfo(name='geoind',
  token='82F46B13B55046680E20FD09364E805A',
  secret='XXQkLNx5bGEOAiAsuBcv9Sz5dYjojlo/KIiag1G')
rsconnect::deployApp('./Application')
```

---

## 6 Conclusion

Le chiffres d'affaire des signes de qualité c'est 32 milliards d'euros et le budget de l'INAO 32 millions d'euros, c'est un millièmme du chiffre d'affaires.

---

```
sessionInfo()
```

---

## 7 Bibliographie

### References

- Agresti, A. and Kateri, M. (2017). Ordinal probability effect measures for group comparisons in multinomial cumulative link models. *Biometrics* 73: 214–219.
- Ay, J.-S. (2011). Hétérogénéité de la terre et rareté économique. Theses, Université de Bourgogne.
- Ay, J.-S. (2019). The informational content of geographical indications. *AAWE Working Paper n XXX*.
- Capus, J. (1947). *L'Évolution de la législation sur les appellations d'origine. Genèse des appellations contrôlées*. L. Larmat.
- Coestier, B. and Marette, S. (2004). *Economie de la qualité*. La découverte.
- Delay, E. and Chevallier, M. (2015). Roger dion, toujours vivant! *Cybergeog: European Journal of Geography*.
- Dion, R. (1952). Querelle des anciens et des modernes sur les facteurs de la qualité du vin. *Annales de géographie* 61: 417–431.

- Garcia, J.-P. (2011). *Les climats du vignoble de Bourgogne comme patrimoine mondial de l'humanité*. Ed. Universitaires de Dijon.
- Humbert, F. (2011). L'INAO, de ses origines à la fin des années 1960: genèse et évolutions du système des vins d'AOC. Ph.D. thesis, Université de Bourgogne.
- Jacquet, O. (2009). *Un siècle de construction du vignoble bourguignon. Les organisations vitivinicoles de 1884 aux AOC*. Editions Universitaires de Dijon.
- Wolikow, S. and Jacquet, O. (2011). *Territoires et terroirs du vin du XVIIIe au XXIe siècles*. Éditions Universitaires de Dijon.
- Wood, S. N. (2017). *Generalized additive models: An introduction with R*. Chapman and Hall/CRC, second edition.
- Wood, S. N., Pya, N. and Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association* 111: 1548–1563.

## A Annexes

### Annexe 1 : Construction des données

Blabla introductif, disponible auprès des auteurs.

- *Données parcellaires*

Le résultat de ces traitements se trouve dans le fichier /Carto/GeoCad.shp (disponible auprès des auteurs sur demande) utilisé dans le code suivant :

---

```
library(sp) ; library(rgdal)
Geo.Cad <- readOGR("./Carto", "GeoCad")
sapply(Geo.Cad@data, function(x) sum(is.na(x)))

table(Geo.Cad$CODECOM, Geo.Cad$BGOR)

plot(Geo.Cad[Geo.Cad$CODECOM== "21150", ])

head(Geo.Cad@data[Geo.Cad$CODECOM== "21150" & , ])

plot(Geo.Cad[Geo.Cad$AOC1b%in% "Chassagne-Montrachet premier cru Morgeot", ],
     add= TRUE, col= "red")
```

---

OGR data source with driver: ESRI Shapefile  
Source: "/home/jsay/geoInd/Carto", layer: "GeoCad"  
with 110350 features  
It has 16 fields

IDU	CODECOM	AREA	PERIM	MAXDIST	PAR2RAS	PAOC	BGOR
0	0	0	0	0	0	49718	49718
BOUR	VILL	COMM	PCRUR	GCRUR	AOC	AOCtp	AOC1b
49718	49718	49718	49718	49718	0	49718	49718

- *Données raster*

Le fichier Data/DatRas appariés aux données du cadastre peut être obtenu auprès des auteurs.

---

```
library(data.table)
dim(Dat.Ras <- fread("./Data/DatRas.csv"))
Cad.Ras <- Dat.Ras[, lapply(.SD, mean),
                  by= list(PAR2RAS), .SDcols= names(Dat.Ras)[ -c(1, 4)]]
Geo.Ras <- merge(Geo.Cad, Cad.Ras, by= "PAR2RAS")
sapply(Geo.Ras@data[, 17: 26], function(x) sum(is.na(x))); rm(Dat.Ras)
```

---

data.table 1.11.4 Latest news: <http://r-datatable.com>

[1] 14253070 13

XL93	YL93	NOMOS	URBAN	FOREST	WATER	DEM	SLOPE	ASPECT	SOLAR
2096	2096	2096	2096	2096	2096	2096	2096	2096	2096

- *Données géologiques polygone*

Nous apparions les 13 960 polygones géologiques présent dans /Carto/GeolMap (disponible sur demande) sur la base du centroïde des parcelles cadastrales, comme présenté dans le code suivant.

---

```
Geol.Map <- readOGR("./Carto/", "GeolMap")
Pts.Cad <- SpatialPoints(Geo.Ras, proj4string= CRS(proj4string(Geol.Map)))
Geo.Ras@data <- cbind(Geo.Ras@data, over(Pts.Cad, Geol.Map))
sapply(Geo.Ras@data[, 28: 43], function(x) sum(is.na(x)))
```

---

OGR data source with driver: ESRI Shapefile  
Source: "/home/jsay/geoInd/Carto", layer: "GeolMap"  
with 13960 features  
It has 16 fields

CODE	NOTATION	DESCR	TYPEGEOL	APLOCALE	TYPEAP
31	31	31	31	862	862
GEOLNAT	ISOPIQUE	AGEDEB	ERADEB	SYSDEB	LITHOLOGIE
31	31	31	31	31	31
DURETE	ENVIRONMT	GEOCHIMIE	LITHOCOM		
69	31	31	69		

- *Données pédologiques polygone*

---

```
Pedo.Map <- readOGR("./Carto", "PedoMap")
Geo.Ras@data <- cbind(Geo.Ras@data, over(Pts.Cad, Pedo.Map))
Geo.Ras@data[, c(45: 48, 50: 55)] <-
  apply(Geo.Ras@data[, c(45: 48, 50: 55)], 2, as.numeric)
sapply(Geo.Ras@data[, 44: 56], function(x) sum(is.na(x)))
```

---

OGR data source with driver: ESRI Shapefile  
Source: "/home/jsay/geoInd/Carto", layer: "PedoMap"  
with 194 features  
It has 13 fields

NOUC	SURFUC	TARG	TSAB	TLIM	TEXTAG	EPAIS	TEG	TMO	RUE
14645	14645	14645	14645	14645	14645	14645	14645	14645	14645
RUD	OCCUP	DESCRp							
14645	14645	14645							

- *Données AOC historiques (polygones)*

---

```
Hist.Aoc <- readOGR("Carto/", "Aoc1936")
Geo.Ras@data <- cbind(Geo.Ras@data, over(Pts.Cad, Hist.Aoc))
sapply(Geo.Ras@data[, 57: 58], function(x) sum(is.na(x)))
```

---

OGR data source with driver: ESRI Shapefile  
Source: "/home/jsay/geoInd/Carto", layer: "Aoc1936"  
with 56 features  
It has 2 fields  
AOC36lab AOC36lvl  
70 70

- *Données Lieux dits (polygones)*

---

```
Lieu.Dit <- readOGR("./Carto/", "LieuDit")
Geo.Ras@data <- cbind(Geo.Ras@data, over(Pts.Cad, Lieu.Dit[, -1]))
sapply(Geo.Ras@data[, 59: 68], function(x) sum(is.na(x)))
```

---

OGR data source with driver: ESRI Shapefile  
Source: "/home/jsay/geoInd/Carto", layer: "LieuDit"  
with 3285 features  
It has 11 fields

LIEUDIT	CLDVIN	LIBCOM	XCHF	YCHF	ALTCOM	SUPCOM
4494	4494	4494	4494	4494	4494	4494
POPCOM	CODECANT	REGION				
4494	4494	4494				

- *Enregistrement de la base*

---

```
dim(Geo.Ras)
save(Geo.Ras, file= "Inter/GeoRas.Rda")
writeOGR(Geo.Ras, "Carto/", "GeoRas", driver= "ESRI Shapefile")
```

---

[1] 110350 68

## Annexe 2 : Intitulés géologiques

---

```
trans_geol <- data.frame(
  GEOL= Reg.Ras$GEOL[!duplicated(Reg.Ras$GEOL)],
  GEOf= c(
    "Calcaires massifs de \"Comblanchien\" (Bathonien sup.)",
    "Marnes et calcaires divers (Callovien inférieur)",
    "Marnes et calcaires argileux (Oxfordien moyen)",
    "Eboulis ordonnés cryoclastiques et colluvions diverses",
    "Oolithe ferrugineuse (Oxfordien moyen-sup)",
    "Calcaires hydrauliques de Molesmes et Noiron (Oxfordien sup.)",
    "Colluvions diverses",
    "Dépôts argilo-limoneux, sables et graviers du Villafranchien",
    "Calcaires de Tonnerre, Oisellemont et calcaires á Astartes",
    "Eboulis et glissements de terrains",
    "Calcaires grenus bicolores (Bathonien terminal)",
    "Terrasse argilo-limoneuse de Saint-Usage",
    "Formation de Saint-Cosme (marnes fluviolacustres varvées)",
    "Alluvions anciennes indifférenciées, argilo-limoneuses",
    "Calcaires bioclastiques, graveleux, á oolites (Bathonien inf.)"
  ),
  GEOe= c(
    "Massive limestones from \"Comblanchien\" (upper Bathonian)",
    "Various marls and limestones (lower Callovian)",
    "Marls and argillaceous limestones (middle Oxfordian)",
    "Ordered cryoclastic scree and various colluviums",
    "Ferruginous Oolite (middle-upper Oxfordian)",
    "Hydraulic limestones of Molesmes and Noiron (upper Oxfordian)",
    "Various colluviums",
    "Clay-silt deposits, sand and gravel from Villafranchien",
    "Limestones of Thunder, Oisellemont and limestones in Astartes",
    "Screes and landslides",
    "Two-tone gray limestones (terminal Bathonian)",
    "Clay-silty terrace of Saint-Usage",
    "Formation of Saint-Cosme (varnished fluviolacustrine marls)",
    "Undifferentiated ancient alluvium, clay-silty",
    "Bioclastic limestones, gravelly, with oolites (lower Bathonian)"
  )
)
```

---

### Annexe 3 : Intitulés pédologiques

---

```
trans_pedo <- data.frame(  
  PEDO= Reg.Ras$PEDO[!duplicated(Reg.Ras$PEDO)],  
  PEDf= c(  
    "Vignoble de la Côte de de Beaune",  
    "Cônes de déjection du pied de Côte",  
    "Côteaux viticoles des Hautes Côtes de Nuits",  
    "Courtes pentes marneuses des plateaux plio-pléistocène",  
    "Piedmont de la côte viticole",  
    "Versants pentus des Hautes Côtes de Beaune",  
    "Sommets des collines des Hautes Côtes de Beaune",  
    "Alluvions récentes calcaires des vallées (Ouche, Tille, Meuzin)",  
    "Pentes liasiques du Haut-Auxois",  
    "Basses terrasses gravelo-caillouteuses des plaines alluviales",  
    "Basses terrasses argileuses des plaines alluviales",  
    "Terrasse argilo-limoneuse de Saint-Usage",  
    "Vignoble de la Côte de Nuits",  
    "Rebord oriental des plateaux calcaires dominant la Côte viticole"  
  ),  
  PEDe= c(  
    "Vineyard of the Côte de Beaune",  
    "Coot footing cones",  
    "Wine hills of Hautes Côtes de Nuits",  
    "Oxfordian limestone-marly trays of the Hautes Côtes",  
    "Short marly slopes of Plio-Pleistocene plateaus",  
    "Piedmont of the vineyard of the Côte",  
    "Sloping slopes of the Hautes Côtes de Beaune",  
    "Summits of the hills of the Hautes Côtes de Beaune",  
    "Recent alluvial limestone valleys (Ouche, Tille, Meuzin)",  
    "Liastic slopes of Haut-Auxois",  
    "Gravelo-stony low terraces of alluvial plains",  
    "Low clay terraces of alluvial plains",  
    "Vineyard of the Côte de Nuits",  
    "Eastern edge of the limestone plateaus overlooking the Côte"  
  )  
)
```

---



## Annexe 4 : Code pour le graphique en violon

---

```
GeomSplitViolin <- ggproto("GeomSplitViolin", GeomViolin,
  draw_group = function(self, data, ..., draw_quantiles = NULL) {
    # Original function by Jan Gleixner (@jan-glx)
    # Adjustments by Wouter van der Bijl (@Axeman)
    data <- transform(data, xminv = x - violinwidth * (x - xmin), xmaxv = x + violinwidth * (xmax - x))
    grp <- data[1, "group"]
    newdata <- plyr::arrange(transform(data, x = if (grp %% 2 == 1) xminv else xmaxv), if (grp %% 2 == 1) y else -y)
    newdata <- rbind(newdata[1, ], newdata, newdata[nrow(newdata), ], newdata[1, ])
    newdata[c(1, nrow(newdata) - 1, nrow(newdata)), "x"] <- round(newdata[1, "x"])
    if (length(draw_quantiles) > 0 & !scales::zero_range(range(data$y))) {
      stopifnot(all(draw_quantiles >= 0), all(draw_quantiles <= 1))
      quantiles <- create_quantile_segment_frame(data, draw_quantiles, split = TRUE, grp = grp)
      aesthetics <- data[rep(1, nrow(quantiles)), setdiff(names(data), c("x", "y")), drop = FALSE]
      aesthetics$alpha <- rep(1, nrow(quantiles))
      both <- cbind(quantiles, aesthetics)
      quantile_grob <- GeomPath$draw_panel(both, ...)
      ggplot2:::ggname("geom_split_violin", grid::grobTree(GeomPolygon$draw_panel(newdata, ...), quantile_grob))
    }
    else {
      ggplot2:::ggname("geom_split_violin", GeomPolygon$draw_panel(newdata, ...))
    }
  }
)

create_quantile_segment_frame <- function(data, draw_quantiles, split = FALSE, grp = NULL) {
  dens <- cumsum(data$density) / sum(data$density)
  ecdf <- stats::approxfun(dens, data$y)
  ys <- ecdf(draw_quantiles)
  violin.xminvs <- (stats::approxfun(data$y, data$xminv))(ys)
  violin.xmaxvs <- (stats::approxfun(data$y, data$xmaxv))(ys)
  violin.xs <- (stats::approxfun(data$y, data$x))(ys)
  if (grp %% 2 == 0) {
    data.frame(
      x = ggplot2:::interleave(violin.xs, violin.xmaxvs),
      y = rep(ys, each = 2), group = rep(ys, each = 2)
    )
  }
  else {
    data.frame(
      x = ggplot2:::interleave(violin.xminvs, violin.xs),
      y = rep(ys, each = 2), group = rep(ys, each = 2)
    )
  }
}

geom_split_violin <- function(mapping = NULL, data = NULL, stat = "ydensity", position = "identity", ...,
  draw_quantiles = NULL, trim = TRUE, scale = "area", na.rm = FALSE,
  show.legend = NA, inherit.aes = TRUE) {
  layer(data = data, mapping = mapping, stat = stat, geom = GeomSplitViolin, position = position,
    show.legend = show.legend, inherit.aes = inherit.aes,
    params = list(trim = trim, scale = scale, draw_quantiles = draw_quantiles, na.rm = na.rm, ...))
}
```

---

## Annexe 5 : Interface utilisateur de l'application

---

```
ui <- dashboardPage(  
  dashboardHeader(  
    titleWidth= 550,  
    title= "Classification statistique des vignobles de la Côte d'Or"),  
  dashboardSidebar(disable = TRUE),  
  dashboardBody(  
    fluidRow(  
      box(width= 5, height= 670,  
        column(width= 4,  
          selectInput(  
            "niveau", label= "Niveau de l'appellation",  
            choices=  
              c(as.character(unique(Poly.Ras$NIVEAU))),  
            selected= 1)),  
        column(width= 4,  
          selectInput(  
            "commune",  
            label= "Commune de la parcelle",  
            choices= c(  
              as.character(unique(Poly.Ras$LIBCOM)),  
              "TOUTES"), selected= 1)),  
        column(width= 4,  
          selectInput(  
            "nom",  
            label= "Lieu dit de la parcelle",  
            choices= c(  
              as.character(unique(Poly.Ras$NOM)),  
              "TOUS"), selected = 1)),  
        plotOutput("miplot", width='100%')  
      ),  
      box(width= 7,  
        column(width = 12,  
          leafletOutput("mymap", height= 645),  
          fluidRow(verbatimTextOutput("mymap_shape_click"))  
        )  
      )  
    )  
  )  
)
```

---

## Annexe 6 : Partie serveur de l'application

---

```
server <- function(input, output, session) {
  ## Reactive values
  values <- reactiveValues(niveau= NULL, commune= NULL, nom= NULL)
  ## Initialisation reactive values
  observe({
    if (is.null(values$niveau)) values$niveau <- input$niveau
    if (is.null(values$commune)) values$commune <- input$commune
    if (is.null(values$nom)) values$nom <- input$nom
  })
  ## MAJ des reactive values apres un click sur un polygone
  observeEvent(input$mymap_shape_click,{
    values$niveau <- Pts.Crd$NIVEAU[Pts.Crd$Concat==
                                     input$mymap_shape_click$id]
    values$nom <- Pts.Crd$NOM[Pts.Crd$Concat==
                              input$mymap_shape_click$id]
    values$commune <- Pts.Crd$LIBCOM[Pts.Crd$Concat==
                                     input$mymap_shape_click$id]
  })
  ## MAJ des reactive values apres un choix dans menus deroulants
  observeEvent(c(input$commune, input$niveau, input$nom),{
    if (values$niveau != input$niveau) {
      values$niveau <- input$niveau
      values$commune <- Pts.Crd$LIBCOM[Pts.Crd$NIVEAU==
                                       values$niveau][ 1]
      values$nom <- Pts.Crd$NOM[Pts.Crd$LIBCOM==
                               values$commune][ 1]
    }
    else if (values$commune != input$commune) {
      values$commune <- input$commune
      values$nom <- Pts.Crd$NOM[Pts.Crd$LIBCOM== values$commune][ 1]
    }
    else if (values$nom!=input$nom){
      values$nom<-input$nom
    }
  })
  ## MAJ menus deroulants
  observeEvent(c(values$commune, values$niveau, values$nom),{
    updateSelectInput(session, "niveau",
                      choices= c(as.character(
                                unique(Poly.Ras$NIVEAU))),
                      selected=values$niveau)
    updateSelectInput(session, "commune",
                      choices= c(as.character(
                                unique(Poly.Ras$LIBCOM[Poly.Ras$NIVEAU %in%
                                                         values$niveau]))),
                      selected=values$commune)
    updateSelectInput(session, "nom",
                      choices= c(as.character(
                                unique(Poly.Ras$NOM[Poly.Ras$LIBCOM %in%
                                                         values$commune &
                                                         Poly.Ras$NIVEAU %in%
                                                         values$niveau]))),
                      selected=values$nom)
  })
  ## Subset donnees
  getPts <- reactive({
    Pts.Crd[Pts.Crd$NIVEAU %in% values$niveau &
```

```

        Pts.Crd$LIBCOM %in% values$commune &
        Pts.Crd$NOM %in% values$nom, ]})
## Carte de base
output$mymap <- renderLeaflet({
  map@map
})
## Rafraichissement carte
observe({
  gg <- getPts()
  if (nrow(gg)== 0) return(NULL)
  else {
    bound_box <- as.numeric(st_bbox(Poly.Ras[Poly.Ras$Concat %in%
                                     gg$Concat,]))
    leafletProxy("mymap") %>%
      clearMarkers() %>%
      fitBounds(lng1= bound_box[ 3], lng2= bound_box[ 1],
                lat1= bound_box[ 4], lat2= bound_box[ 2]) %>%
      addCircleMarkers(data= (getPts()))}
})
## Violon Plot de base
output$miplot <- renderPlot({
  yop <- getPts()$SCORE_corrige
  if (length(yop)==0) return(NULL)
  top <- round(100-
               aggregate(I(Poly.Ras$SCORE_corrige< yop)* 100,
                           by= list(Poly.Ras$NIVEAU), mean)[, 2])
  ggplot(Poly.Ras, aes(x= factor(NIVEAU),
                       y= SCORE_corrige, fill= factor(NIVEAU)))+
  geom_violin(trim= FALSE)+ theme_minimal()+ ylim(40, 100)+
  geom_boxplot(width=0.1, fill= "white")+
  annotate("text", x= 1: 5, y= 100,
            label= paste("Top", top, "%"), col= "red", size= 7)+
  labs(title= "Comparaison avec les autres parcelles",
        x= "\n source: jean-sauveur ay @ inra cesaer, voir https://github.com/jsay/geoInd/",
        y= "Niveau sur une échelle de 1 à 100")+
  scale_fill_manual(values= AocPal)+
  theme(legend.position= "none",
        plot.title = element_text(hjust = 0, size = 16),
        axis.text.x = element_text(size= 14),
        axis.title.x = element_text(hjust= 0, size= 14),
        axis.title.y = element_text(size= 14))+
  scale_x_discrete(expand= expand_scale(mult= 0, add= 1),
                    drop= T)+
  geom_hline(yintercept= yop, lty= 2, col= "red")+
  annotate("text", x= 0.35, y= yop+ 2,
            label= round(yop, 2), col= "red", size= 8)
}, height = 575, width = 700)
## Pour debugguer, ca permet de voir la valeur des input en direct
output$table <- renderTable({
  data.frame(inp= names(unlist(reactiveValuesToList(values))),
             val= unlist(reactiveValuesToList(values)))
})
}

```