Données et Modèles pour une Classification Statistique des Vignobles de Côte-d'Or

Jean-Sauveur AY
< jean-sauveur.ay@inra.fr >

Mohamed HILAL
< mohamed.hilal@inra.fr >

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

Data paper version 1.2 du Mercredi 4 décembre 2019

Résumé

Cet article présente la construction d'une base de données au niveau des parcelles cadastrales pour étudier les relations entre leurs caractéristiques biophysiques (topographie, géologie, pédologie) et leurs 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, ces données permettent d'estimer un modèle statistique expliquant la position des parcelles dans la hiérarchie des AOC à partir des caractéristiques naturelles et humaines. Les prédictions issues du modèle permettent de préciser la hiérarchie des AOC en positionnant chaque parcelle sur une échelle de qualité continue uniquement selon ses attributs biophysiques. Les données, modèles et prédictions sont disponibles sous licence GNU GPL V3 sur le serveur https://data.inra.fr/ et sont consultables par le biais d'une application hébergée à l'url https://cesaer-datas.inra.fr/geoind/. Les codes R permettant de reproduire l'intégralité des résultats sont également fournis.

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

Table des Matières

1 Introduction **2**

2 Présentation des données **4**
2.1 Les AOC actuelles 4
2.2 Enrichissement des AOC historiques 6
2.3 Enrichissement des lieux dits . . . 7
2.4 Enrichissement de la topographie . 8
2.5 Enrichissement de la géologie . . 9
2.6 Enrichissement de la pédologie . . 11
2.7 Statistiques descriptives 12

3 Modèle statistique **13**
3.1 Estimation du modèle 13
3.2 Effets des variables biophysiques . 15
3.3 Effets communaux 16
3.4 Prédiction de la qualité continue . 18
3.5 Agrégation par lieux dits 19

4 Application *Shiny* **21**
4.1 Cartographie dynamique 21
4.2 Lancer l'application localement . . 22
4.3 Exemple d'utilisation 22

5 Conclusion **23**
5.1 Remerciements 23

A Annexes **26**

1 Introduction

Les appellations d'origine contrôlée (AOC) viticoles de Bourgogne résultent de processus historiques complexes au cours desquels les parcelles ont été classifiées selon leurs caractéristiques biophysiques et selon les rapports économiques, politiques et sociaux en vigueur (Garcia, 2011; Wolikow and Jacquet, 2011). Ainsi, 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 la complexité de 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 prix d'une perte d'information et d'une certaine opacité pour les acteurs des marchés du vin.

La référence au lieu de production permet de donner une certaine indication composite sur la qualité des vins lors des échanges. Une abondante littérature économique (synthétisée par Coestier and Marette, 2004) montre qu'en diminuant l'asymétrie d'information entre les vendeurs et les acheteurs, les AOC peuvent limiter cette défaillance de marché préjudiciable aux deux parties prenantes. La question de la nature de l'information contenue dans les AOC se pose alors, en particulier la distinction de la part relative aux processus naturels de la part relative aux processus humains. Cette séparation est déterminante pour identifier les facteurs naturels, immobiles et non-reproductibles, qui justifient réellement la référence au lieu de production (Ay, 2019). Nous présentons dans cet article la constitution de données et l'estimation de modèles qui permettent d'opérer statistiquement cette distinction. Nous montrons la présence d'une hiérarchie implicite entre les communes de la zone qui biaise l'information transmise par les AOC sur les caractéristiques biophysiques des parcelles. Nous utilisons les prédictions de cette modélisation pour classifier l'ensemble des parcelles sur une échelle continue de qualité (entre 0 et 100) tout en corrigeant les effets communaux issus de l'histoire. Nous présentons cette information par le biais d'une application cartographique libre.

La Section 2 présente la construction de la base de données géographique disponible sous licence GNU GPL V3 sur le serveur https://data.inra.fr/. La parcelle cadastrale est l'unité élémentaire d'observation qui permet l'appariement des variables sur les AOC actuelles (produites par l'INAO), sur les AOC de 1936 (produites par la MSH de Dijon), sur les lieux dits par le Plan Cadastral Informatisé (produit par la DGFiP), sur l'altimétrie par le RGE ALTI® à 5 mètres (produit par l'IGN), sur l'occupation du sol (produite par Hilal et al., 2018), sur la géologie par Charm-50 (produit par le BRGM) et sur la pédologie par le Référentiel Pédologique de Bourgogne (produit par le Gis Sol). Les données ainsi constituées concernent l'ensemble des parcelles des 31 communes inclues dans 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 (Figure 1). Cette base de données permet de relier finement les AOC aux caractéristiques biophysiques des parcelles dont les vins sont issus, et possède ainsi une utilisation plus large que celle présentée ici.

La Section 3 présente l'estimation des modèles statistiques dont les spécifications sont décrites plus extensivement dans un article associé (Ay, 2019). Le principe est d'utiliser la structure hiérarchique des AOC (Coteaux bourguignons < Bourgogne régional < Villages < Premiers crus < Grands crus) pour les relier aux caractéristiques biophysiques des parcelles par une variable latente de qualité des vignes. Nous montrons que cette variable continue non observable peut être déduite des AOC actuelles de manière flexible. Les estimations s'effectuent par des modèles ordonnés additifs généralisés (OGAM pour *ordered generalized additive model*, Wood et al., 2016) qui prédisent correctement près de 90 % des niveaux AOC actuels. Ils permettent également d'estimer semi-paramétriquement l'effet de chaque variable biophysique sur la hiérarchie, ainsi que des effets communaux issues de l'histoire. Ces estimations permettent de corriger les effets communaux pour prédire la qualité des vignes uniquement à partir des caractéristiques biophysiques.

La Section 4 présente le codage et l'utilisation de l'application *Shiny* (Chang et al., 2019) qui permet de consulter la classification continue des parcelles de vignes, telle que prédite par la modélisation statistique. L'utilisateur peut ainsi saisir les informations typiquement disponibles sur les étiquettes des bouteilles de vin de Bourgogne (niveau de l'AOC dans la hiérarchie, commune de production, et lieux dit de la parcelle) pour identifier géographiquement l'ensemble des parcelles et leur niveau de qualité sur une échelle de 0 à 100 (avec ou sans correction des effets communaux). Cette information permet une évaluation plus précise de la qualité des vins que la hiérarchie actuelle des AOC en 5 niveaux, sans introduire de facteurs subjectifs exogènes. Cela permet en outre d'améliorer l'information disponible pour les consommateurs à partir d'informations déjà présentes sur les étiquettes. Chaque vin identifié peut alors être comparé aux autres vins du même niveau hiérarchique ou aux vins d'autres niveaux hiérarchiques afin d'évaluer sa qualité relative.

Ce document contient les codes R (R Core Team, 2019) qui permettent de reproduire l'ensemble des tables et des figures à partir des données disponibles sur le serveur https://data.inra.fr/. La version du logiciel et des packages utilisés au moment de la rédaction de ce document sont reportés en Annexe 1. L'intégralité du code relatif à l'application *Shiny* est également reportée en Annexes 4, 5 et 6, afin qu'elle puisse être lancée localement, voire modifiée, par les utilisateurs. La version la plus récente des différents codes reportés dans ce document est accessible sur un répertoire distant à https://github.com/jsay/geoInd.

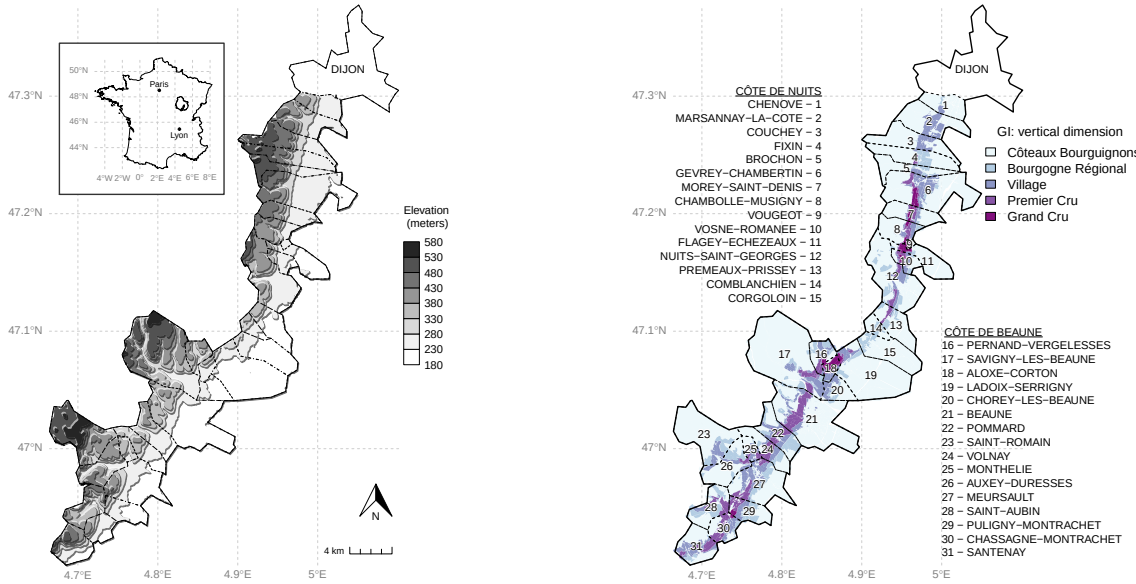

Figure 1: **La zone de la *Côte d'Or* étudiée, sa topographie et ses appellations d'origine contrôlées**

Source : INRA / INAO / MSH / DGFiP / IGN / BRGM / Gis Sol.

Lecture : La carte de gauche représente l'altitude des parcelles, catégorisée en 8 classes de 50 mètres d'amplitude. Situant la ville de Dijon au Nord de la zone, elle permet de faire apparaître la topographie de la Côte viticole majoritairement orientée à l'Est mais dont la présence de vallées sèches (appelées combes) produit des variations localisées. La carte de droite présente les 31 communes de la zone, qui constituent la dimension horizontale des AOC viticoles (il n'y a pas de hiérarchie explicite entre les communes). La dimension verticale est représentée par la hiérarchie en 5 niveaux reportée sur cette même carte. Ces deux cartes sont reproductibles à partir des données présentées dans cet article, les codes utilisés sont consultables à https://github.com/jsay/geoInd.

2 Présentation des données

2.1 Les AOC actuelles

Pour la construction des données, l'unité géographique de base est la parcelle cadastrale des 31 communes viticoles de la zone présentée dans la Figure 1. La géométrie des parcelles est issue de la BD parcellaire de l'IGN dans sa version 2014 pour la Côte-d'Or (téléchargement le 09/10/2015). Nous l'avons enrichie d'attributs décrivant la géométrie des parcelles avec l'ajout de la surface, du périmètre et de la distance maximale entre deux sommets pour chaque polygone cadastral (Conrad et al., 2015). Les polygones cadastraux ont ensuite été appariés par jointure aux délimitations parcellaires des AOC viticoles produites par l'INAO disponibles à l'adresse https://www.data.gouv.fr/fr/datasets/delimitation-parcellaire-des-aoc-viticoles-de-linao sous licence ouverte (téléchargement le 21/08/18).

Le résultat de cette étape et des autres étapes ci-dessous relatives à la construction des données sont disponibles sur le serveur https://data.inra.fr/. Le code ci-dessous permet aux utilisateurs de `R` de charger directement la version la plus récente des données par l'utilisation du package `dataverse` disponible sur le CRAN (Leeper, 2017). Nous constatons que la version actuelle de la base compte 110 350 parcelles et 63 variables. Les variables issues de cette première étape sont présentes dans les colonnes 2 à 16.

```
library(dataverse) ; library(sp)
Sys.setenv("DATAVERSE_SERVER" = "data.inra.fr")
GeoRasRaw <- get_file("GeoRas.Rda", "https://doi.org/10.15454/ZZWQMN")
writeBin(GeoRasRaw, "GeoRas.Rda") ; load("GeoRas.Rda")
dim(Geo.Ras) ; names(Geo.Ras)[ 2: 16]
```

```
[1] 110350     67

 [1] "IDU"     "CODECOM" "AREA"    "PERIM"   "MAXDIST" "PAOC"
 [7] "BGOR"    "BOUR"    "VILL"    "COMM"    "PCRU"    "GCRU"
[13] "AOC"     "AOCtp"   "AOClb"
```

L'objet `Geo.Ras` téléchargé est un objet de la classe `SpatialPolygonsDataFrame`, définie par le package `sp` (Bivand et al., 2013) que l'on charge au préalable. Le dictionnaire de ces 15 variables relatives aux parcelles cadastrales est reporté dans la Table 1 ci-dessous. L'information brute issue de la superposition de la couche cadastrale avec la couche INAO sur les AOC actuelles est présente dans les variables `PAOC` à `GCRU`. Ces variables contiennent la valeur 1 lorsque que le niveau AOC est revendicable sur la parcelle correspondante et 0 sinon. Les 49 718 observations manquantes qui apparaissent pour chacune de ces 7 variables correspondent aux parcelles cadastrales hors du périmètre des AOC.

Nous avons ensuite recodé cette information dans les trois dernières variables `AOC`, `AOCtp` et `AOClb` qui sont plus opérationnelles pour l'analyse statistique. Selon le principe des replis (issu de la doctrine de l'INAO), les parcelles d'un niveau hiérarchique supérieur peuvent toujours être revendiquées dans un niveau inférieur. La superposition des couches de l'INAO conduit donc à la présence de plusieurs niveaux d'AOC sur une même parcelle, ce qui entre en contradiction avec une autre 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` que nous avons créée représente cette valeur pour chacune des parcelles, elle est codée `0` pour les parcelles hors AOC, `1` pour les Coteaux bourguignons, `2` pour les Bourgognes régionaux et jusqu'à `5` pour les Grands crus. De plus, les informations présentes sur les étiquettes des vins peuvent correspondre soit à des AOC soit à des dénominations au sein du

système (le plus souvent sans que cette distinction soit claire pour le consommateur). Les modalités prises par la variable `AOClb` est une combinaison du nom des appellations et des dénominations (la variable `AOCtp` code cette combinaison). Les modalités correspondent souvent au nom de l'AOC maximale revendicable. Pour les Bourgognes régionaux, nous n'utilisons pas la dénomination "Bourgogne Côte d'Or" plus haute dans la hiérarchie que l'AOC Bourgogne mais peu connue du fait de sa faible antériorité (la dénomination a été crée en 2015). D'ailleurs, l'analyse se limite à la Côte d'Or où les délimitations "Bourgogne Côte d'Or" et "Bourgognes régionaux" sont très proches. C'est principalement pour les Premiers Crus que la variable `AOClb` contient les dénominations géographiques, car l'AOC ne fait référence qu'au niveau village alors que les dénominations permettent d'identifier plus précisément les lieux dits des parcelles.

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

NOM	TYPE	DESCRIPTION
`IDU`	*Caractère*	Identifiant de la parcelle cadastrale (14 caractères)
`CODECOM`	*Caractère*	Code INSEE de la commune d'appartenance (5 caractères)
`AREA`	*Numérique*	Surface calculée de la parcelle cadastrale (en mètres carrés)
`PERIM`	*Numérique*	Périmètre calculé de la parcelle cadastrale (en mètres)
`MAXDIST`	*Numérique*	Distance maximale calculée entre deux sommets (en mètres)
`PAOC`	*Indicatrice*	1 si la parcelle est dans au moins une AOC
`BGOR`	*Indicatrice*	1 si la parcelle est dans le niveau Coteaux bourguignon
`BOUR`	*Indicatrice*	1 si la parcelle est dans le niveau Bourgogne régional
`VILL`	*Indicatrice*	1 si la parcelle est dans le niveau Bourgogne village
`COMM`	*Indicatrice*	1 si la parcelle est dans le niveau Bourgogne communal
`PCRU`	*Indicatrice*	1 si la parcelle est dans le niveau Premier cru
`GCRU`	*Indicatrice*	1 si la parcelle est dans le niveau Grand cru
`AOC`	*Numérique*	Rang de la parcelle dans la hiérarchie des AOC (entre 0 et 5)
`AOCtp`	*Caractère*	`Appel` si le libellé est une appellation, `Denom` pour dénomination
`AOClb`	*Caractère*	Libellé de l'appellation ou de la dénomination selon la variable `AOCtp`

La distribution de l'ensemble des parcelles de la zone entre la dimension horizontale (entre les communes) et verticale des AOC (entre les niveaux hiérarchiques) est présentée dans la Figure 2 suivante, dont le code est reporté ci-dessous. Pour la clarté du code, les objets et fonctions de configuration graphique `my.lab`, `my.pal`, `my.par`, `my.key` et `my.pan` sont reportés dans l'Annexe 2. Ces objets doivent être chargés en préalable pour le fonctionnement du code en local, en plus des packages `lattice` et `RColorBrewer`.

```
tmp <- unique(Geo.Ras$LIBCOM[order(Geo.Ras$YCHF, decreasing= TRUE)])
Geo.Ras$LIBCOM <- factor(Geo.Ras$LIBCOM, levels= tmp)
Geo.Fig <- subset(Geo.Ras, !is.na(AOClb))
fig.dat <- aggregate(model.matrix(~ 0+ factor(Geo.Fig$AOC))*
                     Geo.Fig$AREA/ 1000, by= list(Geo.Fig$LIBCOM), sum)
names(fig.dat) <- c("LIBCOM", "BGOR", "BOUR", "VILL", "PCRU", "GCRU")
fig.dat$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)
barchart(LIBCOM~ BGOR+ BOUR+ VILL+ PCRU+ GCRU, xlim= c(-100, 10500),
         xlab="Surfaces sous appellation d'origine contrôlée (hectare)",
         data= fig.dat, horiz= T, stack= T, col= my.pal, border= "black",
         par.settings= my.par, auto.key= my.key, panel= my.pan)
```

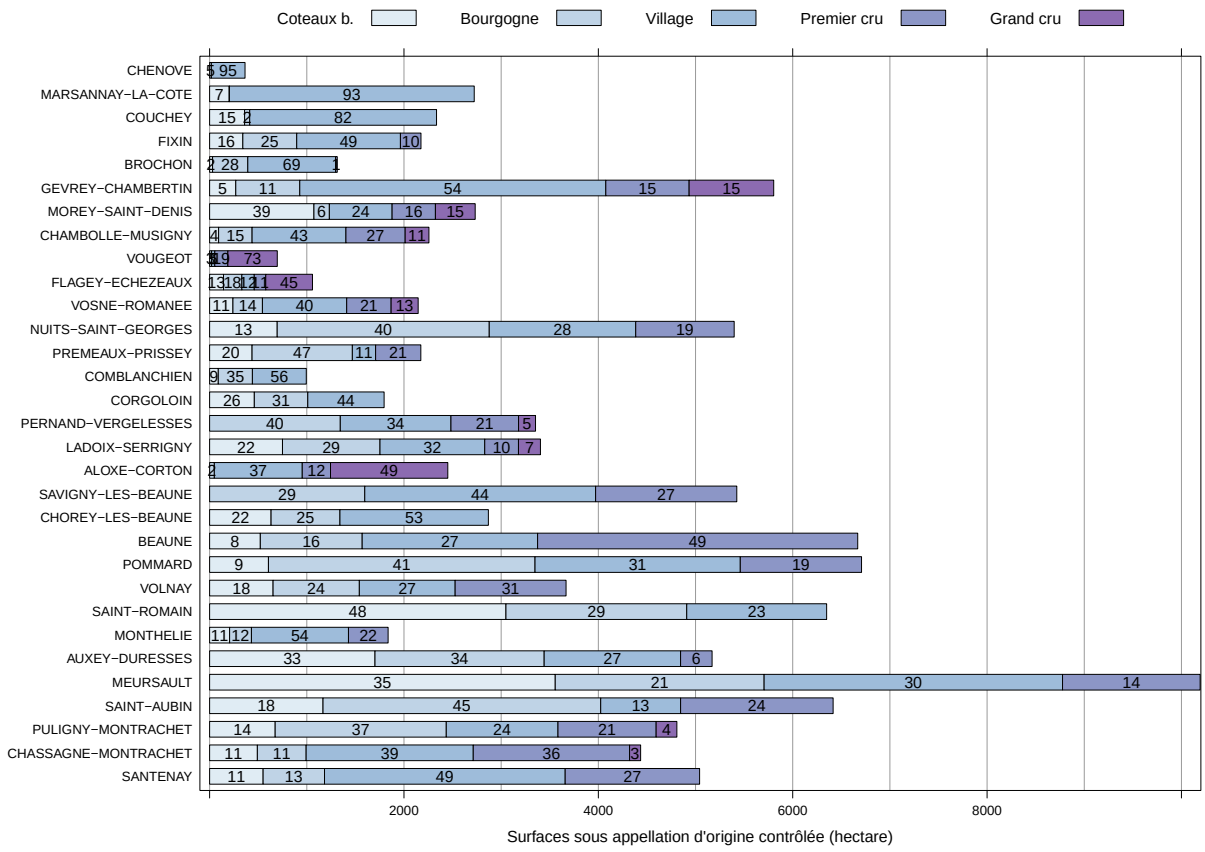

Figure 2: **Croisement entre les communes (dimension horizontale) et les niveaux (dimension verticale).**

Source : INRA / INAO / MSH / DGFiP / IGN / BRGM / GisSol.

Lecture : Pour chacune des 31 communes reportées en ordonnées, le graphique présente les surfaces de chacun des 5 niveaux hiérarchiques en absisses. Les pourcentages intra-communaux sont reportés à l'intérieur du graphique. Le niveau d'AOC de chaque parcelle est sélectionnée par la méthode du maximum revendicable telle que codée dans la variable `AOC` de l'objet `Geo.Ras`.

2.2 Enrichissement des AOC historiques

Nous enrichissons ensuite ces données parcellaires de variables sur des classifications historiques obtenues auprès de la Maison des Sciences de l'Homme de Dijon. Alors que l'INAO a été créé en 1936, la première délimitation officielle des parcelles viticoles s'est opérée entre 1936 et 1940 sur la zone d'intérêt. Elle fut basée sur deux classements antérieurs non officiels: celui de Jules Lavalle de 1855 (Lavalle, 1855) et le Classement du Comité d'Agriculture et de Viticulture de l'Arrondissement de Beaune de 1860 (Wolikow and Jacquet, 2011). Nous compilons ces différentes classifications pour obtenir une hiérarchie des parcelles en 3 niveaux: Régional < Village < Grand Cru que nous considérons comme les niveaux d'AOC en 1936. Cette classification historique est moins détaillée que l'actuelle (3 niveaux au lieu de 5) car l'AOC Coteaux bourguignons n'existait pas encore (les niveaux ordinaires et grands ordinaires qui la précédèrent n'étaient pas délimités) tout comme les Premiers Crus qui ont été instaurés par décret en 1943 (Lucand, 2017). L'appariement s'effectue par le centroïde des parcelles cadastrales car la géométrie des polygones ne correspond pas parfaitement (à la fois par la numérisation des cartes historiques et parce que le cadastre a changé). La faible taille des parcelles (0.2 ha en moyenne) permet de faire confiance en cette procédure d'appariement, confirmée par de nombreuses vérifications manuelles. La base parcellaire est ainsi enrichie des 2 variables `AOC1936lab` et `AOC36lvl` présentées dans la Table 2.

Table 2: **Nom, type et description des variables issues des AOC historiques**

NOM	TYPE	DESCRIPTION
AOC36lab	*Caractère*	Libellé de l'appellation en 1936 (56 modalités)
AOC36lvl	*Caractère*	Rang de la parcelle dans la hiérarchie des AOC de 1936 (0, 3 ou 5)

Ces deux nouvelles variables correspondent aux colonnes 56 et 57 de la base Geo.Ras téléchargée sur le serveur. Le croisement de la hiérarchie des AOC de 1936 avec la hiérarchie des AOC actuelles dans le code ci-dessous montre que les surfaces sous AOC étaient sensiblement plus réduites. Elles représentaient 27% des parcelles de la zone au lieu de 55% actuellement. Près de 165 000 parcelles hors AOC en 1936 le sont actuellement (tous niveaux confondus) alors que seulement 2 610 sont dans le cas inverse. La majorité des parcelles en niveaux Village et Grand cru actuellement l'étaient déjà en 1936, les Premiers Crus actuels étaient principalement en Village, et les Coteaux bourguignons et les Bourgogne niveau régional étaient hors AOC.

```
names(Geo.Ras)[ 56: 57]
table(Geo.Ras$AOC36lvl, Geo.Ras$AOC)
```

```
[1] "AOC36lab" "AOC36lvl"
```

```
       0     1     2     3     4     5
  0 47056  9832 13337 10554   593    44
  3  2586    15   565 15529  8226   266
  5    24     0     1    14     3  1635
```

2.3 Enrichissement des lieux dits

Nous utilisons également le Plan Cadastral Informatisé vecteur présent sur le site officiel data.gouv.fr à l'adresse https://cadastre.data.gouv.fr/datasets/plan-cadastral-informatise téléchargé pour la Côte-d'Or (21) le 13/01/2019. Ces données sont en licence ouverte Etalab, elles nous permettent d'obtenir les lieux dits pour les parcelles viticoles de la zone, en plus de certaines variables communales agrégées présentées dans la Table 3. Une attention particulière est portée sur les lieux dits dont les intitulés doivent être croisés avec le nom des communes pour être uniques (un même nom de lieu dit peut être présent sur plusieurs communes). Comme la géométrie des lieux dits et des parcelles de l'IGN colle parfaitement, nous pouvons enrichir les données parcellaires directement par la jointure des polygones.

Les variables issues de cette étape se retrouvent dans les colonnes 58 à 66 de l'objet Geo.Ras téléchargé sur le serveur. Comme reporté dans le code ci-dessous, nous pouvons calculer la distance euclidienne (à vol d'oiseau) entre les centroïdes de chaque parcelle et le centroïde du chef lieu de la commune d'appartenance. Nous obtenons une distance moyenne de 1 km 200, en cohérence avec la taille des communes de la zone (environ 2.5 km^2). Notons que 6 426 parcelles de la BD parcellaire sont absentes du Plan Cadastral Informatisé. Elles correspondent à environ 4 % de la base initiale et n'ont donc pas été appariées (des valeurs omises sont reportées pour ces variables). Ces parcelles sont pour la plupart hors AOC et se concentrent sur les communes les plus urbanisées, telles que Chenôve, Marsannay-la-Côte et Beaune. Ces valeurs manquantes semblent correspondre à des espaces bâtis qui ne peuvent pas être classés en AOC, mais des vérifications

Table 3: **Nom, type et description des variables issues des lieux dits**

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

manuelles n'ont pas suffit pour statuer définitivement sur ce point. Ces valeurs manquantes sont exclues de l'analyse statistique, elles ne sont pas décisives pour les estimations présentées par la suite.

```
Geo.Ras$DISTCHF <- sqrt(((Geo.Ras$XL93- Geo.Ras$XCHF* 100))^2
                        + ((Geo.Ras$YL93- Geo.Ras$YCHF* 100))^2)
names(Geo.Ras)[ 58: 66] ; summary(Geo.Ras$DISTCHF)
```

```
[1] "LIEUDIT"  "CLDVIN"   "LIBCOM"   "XCHF"     "YCHF"     "ALTCOM"
[7] "SUPCOM"   "POPCOM"   "CODECANT"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
      0     595    1049    1230    1679    6314    6425
```

2.4 Enrichissement de la topographie

L'enrichissement des variables biophysiques sur les parcelles cadastrales commence avec l'appariement d'informations raster issues d'un modèle numérique de terrain (MNT) et d'une couche d'occupation du sol, opéré par la rastérisation à 5 mètres de la couche vectorielle du parcellaire cadastral. Les données raster sur l'altimétrie sont issues du MNT RGE ALTI® 5 m, sous licence IGN "Recherche", les données d'occupation du sol proviennent du modèle développé par Hilal et al. (2018) à partir de la BD TOPO®, du registre parcellaire graphique et de Corine Land Cover. De la couche altimétrique du MNT ont été dérivées 3 autres couches raster supplémentaires, toujours à 5 m de résolution: la pente, l'exposition et les radiations solaires. Ces attributs ont été calculés en utilisant le logiciel ArcGis (Rich and Fu, 2000). Les 5 couches raster (altitude, pente, exposition, radiation solaire et occupation du sol) ont été alors converties en tables au format XYZ, avec X et Y les coordonnées Lambert 93 du centre de chaque pixel et Z la variable d'intérêt de chacune des couches. Ces tables sont regroupées dans une même table qui contient un seul XY et les 5 attributs Z issus chacun des tables de départ. Cette table est ensuite appariée avec la table XYZ issue de la rastérisation des parcelles cadastrales pour récupérer l'identifiant PAR2RAS qui servira à l'appariement avec la topographie. Nous obtenons ainsi une base contenant plus de 14 millions de lignes, une pour chaque pixel de 5 m. Pour les 8 variables issues de cette procédure, nous calculons des moyennes à l'échelle des parcelles cadastrales, sachant que d'autres méthodes d'agrégation ont été utilisées sans différences notables sur les résultats présentés. Le dictionnaire de ces variables est reporté dans la Table 4.

Table 4: **Nom, type et description des variables topographiques à la parcelle**

NOM	TYPE	DESCRIPTION
PAR2RAS	*Numérique*	Identifiant pour appariement entre vecteurs et raster
XL93	*Numérique*	Latitude du centroïde de la parcelle (système Lambert 93)
YL93	*Numérique*	Longitude du centroïde de la parcelle (système Lambert 93)
NOMOS	*Numérique*	Part de la parcelle hors du mode d'occupation des sol (entre 0 et 1)
URBAN	*Numérique*	Part de la parcelle en usage urbain selon le MOS (entre 0 et 1)
FOREST	*Numérique*	Part de la parcelle en usage forestier selon le MOS (entre 0 et 1)
WATER	*Numérique*	Part de la parcelle en eau selon le MOS (entre 0 et 1)
DEM	*Numérique*	Altitude moyenne de la parcelle selon le MNT (en mètres)
SLOPE	*Numérique*	Pente moyenne de la parcelle selon le MNT (en degrés)
ASPECT	*Numérique*	Exposition moyenne de la parcelle selon le MNT (en degrés)
SOLAR	*Numérique*	Radiation solaire moyenne sur la parcelle (en Joules)

Ces variables parcellaires sont disponible dans les colonnes 17 à 26 de l'objet `Geo.Ras`, comme présenté dans le code ci-dessous. Pour des raisons d'unité de mesure qui se poseront lors de l'analyse statistique, nous centrons et réduisons la variable `SOLAR` sur les rayonnements solaires. Toujours pour des raisons de spécification statistique, nous discrétisons la variable `ASPECT` sur l'exposition moyenne des parcelles en 8 classes de 45 degrés d'amplitude. Nous obtenons 2 096 valeurs manquantes pour lesquelles le code `PAR2RAS` des parcelles ne s'apparie à aucune cellule raster. Ces parcelles sont de faible taille avec des géométries particulières et font penser à des "erreurs" du cadastre. Nous les enlèverons au moment de l'analyse statistique sachant que cela revient à enlever 2.7 ha, soit moins de 0.01 % de la surface totale de la zone. Sur les variables issues du MOS, nous de conservons que les catégories relatives aux modes d'occupation non agricoles (urbain, forêt, eau), afin de pouvoir les exclure si besoin (non utilisé actuellement).

```
names(Geo.Ras)[ c(1, 17: 26)]
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))
```

```
[1] "PAR2RAS" "XL93"    "YL93"    "NOMOS"   "URBAN"   "FOREST"
[7] "WATER"   "DEM"     "SLOPE"   "ASPECT"  "SOLAR"
```

2.5 Enrichissement de la géologie

Les données géologiques utilisées dans ce travail sont issues de la BD harmonisée Charm-50 produite par le BRGM à l'échelle du 1/50 000. Cette base est disponible sur le site http://infoterre.brgm.fr sous licence Ouverte. Nous utilisons une extraction des formations géologiques, nommée `GEO050K_HARM_021_S_FGEOL_CGH_2154`, dont le téléchargement a été effectué le 25/04/2019 pour le département de la Côte-d'Or. L'appariement des polygones géologiques avec le parcellaire cadastral enrichi s'effectue par le centroïde des parcelles, la faible taille des parcelles permet de s'assurer de la validité de cette procédure, vérifiée manuellement par ailleurs. Le dictionnaire associé aux 16 variables sur la géologie est disponible dans la Table 5. La description des variables est peu précise actuellement car les données du BRGM sont disponibles depuis peu et ne possèdent

pas encore, à notre connaissance, de dictionnaire exploitable. Ce manque de précision n'est pas limitant pour l'analyse statistique que nous présentons ici car les variables géologiques sont utilisées de manière discrètes, par le biais d'indicatrices qui ne nécessitent pas de spécification explicite. Cela peut néanmoins être différent pour d'autres utilisations de la base de données.

Table 5: **Nom, type et description des variables issues des données géologiques**

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

Nous nous concentrons donc sur la variable NOTATION qui découpe la zone d'intérêt en 31 formations géologiques homogènes. Sa distribution spatiale et ses intitulés sont présentés dans la Figure 5 en Annexe 3. Les parcelles non appariées produisant des valeurs manquantes sont peu nombreuses (entre 31 et 862 parcelles selon les variables), elles seront enlevées au moment de l'analyse statistique sans conséquence sur les résultats. Pour diminuer la multi-colinéarité lors de l'utilisation de ces indicatrices qui codent la géologie des parcelles et pour s'assurer d'estimations précises, nous regroupons les unités géologiques qui comptent moins de 1 000 parcelles dans une modalité de référence codée 0AREF. Il reste ainsi 17 unités géologiques qui pourront être utilisées dans la modélisation statistique de la hiérarchie des AOC.

```
names(Geo.Ras)[27: 42]
Geo.Ras$NOTATION <- factor(Geo.Ras$NOTATION)
tmp <- table(Geo.Ras$NOTATION)< 1000
table(Geo.Ras$GEOL <- factor(
          ifelse(Geo.Ras$NOTATION %in% names(tmp[ tmp]), "0AREF",
                 as.character(Geo.Ras$NOTATION))))
```

```
 [1] "CODE"       "NOTATION"   "DESCR"      "TYPEGEOL"   "APLOCALE"
 [6] "TYPEAP"     "GEOLNAT"    "ISOPIQUE"   "AGEDEB"     "ERADEB"
[11] "SYSDEB"     "LITHOLOGIE" "DURETE"     "ENVIRONMT"  "GEOCHIMIE"
[16] "LITHOCOM"

0AREF     C     E    Fu    Fx    Fy    Fz    GP  j1-2    j3   j3a
```

```
 5487 29040  2683  1653  9321 10006  7951 11181  1359  1848  3785
  j3b   j4a   j5a   j5b   j6a  p-IV
 2887  2934  5201  5301  4827  4855
```

2.6 Enrichissement de la pédologie

Les données pédologiques utilisées sont extraites du Référentiel Pédologique de Bourgogne : Régions naturelles, pédopaysage et sols de Côte-d'Or à l'échelle 1/250 000 (étude 25021 dans le référentiel Gis Sol). Ces données sont compatibles avec la référence nationale DoneSol. La localisation des types de sol et l'appariement avec le parcellaire cadastral s'opèrent par le biais des 194 Unités Cartographiques de Sols qui composent la zone. Les UCS sont des polygones construits pour être homogènes en termes de pédo-paysages (topographie, climat, géologie). Elles sont typiquement utilisées pour représenter cartographiquement les caractéristiques des sols, elles peuvent néanmoins contenir différents types de sols. Ces derniers, regroupés en unités typologiques, ne peuvent pas être localisés plus précisément que les unités cartographiques. Cette imprécision dans la localisation des données est une limite importante pour leur usage statistique à l'échelle parcellaire (Ay, 2011). En l'absence de données plus fines spatialement, les données parcellaires du cadastre sont enrichies du libellé de l'UCS et des 11 variables correspondantes à l'unité typologique de sol dominante, c'est-à-dire celle qui est la plus étendue au sein de chaque UCS. Ce choix ne change pas les résultats obtenus. Le dictionnaire des 13 variables pédologiques issues de cette procédure est disponible dans la Table 6.

Table 6: **Nom, type et description des variables issues des données pédologiques**

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

Comme pour les variables sur la géologie, les variables pédologiques sont intégrées dans le modèle par des indicatrices qui correspondent aux UCS. Le détails des 11 variables pédologiques est maintenu dans les données constituées pour ne pas limiter les autres usages qui peuvent en être faits. Les libellés des unités cartographiques reportés dans la variable `DESCRp` sont obtenus par une saisie manuelle à partir du site https://bourgogne.websol.fr/carto. Les valeurs manquantes associées aux parcelles non couvertes par la couche pédologique sont assez nombreuses : 14 645 parcelles cadastrales, soient environ 4.25% des surfaces de la zone. Ces parcelles non couvertes sont en revanche peu désignées en AOC (moins de 1% des AOC ont des variables pédologiques manquantes), il s'agit donc dans de rares cas de vignes et principalement de parcelles construites au coeur des villages. Une explication intuitive de ces valeur manquantes est l'absence de données pédologiques sur les sols artificialisés, cette interprétation étant corroborée manuellement. La

faible précision spatiale de ces données pédologiques peut s'illustrer par comparaison avec les variables du MOS sur l'artificialisation. Les UCS pour lesquelles les variables pédologiques manquent peuvent regrouper des modes d'occupation du sol très différents. D'une série initiale de 33 modalités (présentées dans la Figure 6 en Annexe 3), le code ci-dessous opère le regroupement des modalités peu présentes (moins de 1 000 parcelles) et reporte la distribution des 19 modalités restantes.

```
names(Geo.Ras)[43: 55]
Geo.Ras$NOUC <- factor(Geo.Ras$NOUC)
tmp <- table(Geo.Ras$NOUC)< 1000
table(Geo.Ras$PEDO <- factor(
          ifelse(Geo.Ras$NOUC %in% names(tmp[tmp]), "0AREF",
                 as.character(Geo.Ras$NOUC))))
```

```
 [1] "NOUC"   "SURFUC" "TARG"   "TSAB"   "TLIM"   "TEXTAG" "EPAIS"
 [8] "TEG"    "TMO"    "RUE"    "RUD"    "OCCUP"  "DESCRp"

0AREF    10    13    14    26    27    28    29    30    32    34
 3265  2074  3770 23472  4750  1348 11641  7636  6983  3072  2469
   35    36    38     5    61    69     7     8
 8356  1602  2198  4767  1605  2116  1445  3136
```

2.7 Statistiques descriptives

Nous avons donc compilé les variables de 7 sources différentes, le code suivant effectue les derniers traitements, à savoir la re-projection des coordonnées des centroïdes dans le système WGS84 utilisé pour l'application *Shiny*, la suppression des valeurs manquantes sur certaines variables, le codage des indicatrices (pour les AOC et l'exposition), et la normalisation des unités de mesure. L'objet `tb.lab` qui contient l'intitulé des variables, nécessaire dans le code ci-dessous, est défini en Annexe 2. Le package `stargazer` doit être chargé en préalable pour construire la Table 7.

```
GR84 <- spTransform(Geo.Ras, CRS("+proj=longlat +ellps=WGS84"))
dd <- coordinates(GR84) ; Geo.Ras$X= dd[, 1] ; Geo.Ras$Y= dd[, 2]
Reg.Ras <- subset(Geo.Ras, !is.na(AOClb) & !is.na(DEM) & !is.na(DESCR)
                  & !is.na(RUD) & !is.na(AOC36lab) & !is.na(REGION))
Stat.Ras <- data.frame(Reg.Ras@data, model.matrix(~0+ factor(Reg.Ras$AOC)),
                       model.matrix(~ 0+ factor(Reg.Ras$EXPO)))
names(Stat.Ras)[75: 79] <- paste0("AOC", 1: 5)
names(Stat.Ras)[80: 87] <- 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
stargazer(Stat.Ras[, names(tb.lab)], covariate.labels=tb.lab, float= F,
          font.size= "small", column.sep.width= "0pt", digit.separate= c(0, 3))
```

Nous obtenons une base de données qui contient 59 113 observations utilisables pour estimer le modèle statistique. 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). Nous observons des surfaces parcellaires faibles (0.2 ha de moyenne), des altitudes comprises entre 200 et 500 m (286 m de moyenne), des pentes entre 0 et 37 degrés (5.75 degrés de moyenne) et des radiations solaires

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

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Surface [1000 m^2]	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

Notes: L'échantillon est composé de 59 113 parcelles (pas nécessairement en vignes) sous AOC dans la zone considérée. Le tableau reporte la moyenne, l'écart-type, le minimum, les quartiles et le maximum pour les principales variables biophysiques issues du processus d'appariement présenté.

comprises entre 581 000 et 1.2 millions de Joules (1 millions de Joules en moyenne). Le niveau village de la hiérarchie des AOC regroupe 42% des parcelles, les niveau régionaux et coteaux bourguignons respectivement 23% et 16.5%, alors que les niveaux premier et grand cru respectivement 15% et 3%. Les vignobles sous AOC sont globalement orientés à l'Est, avec 55% des observations qui ont une orientation comprise entre 45 et 135 degrés.

3 Modèle statistique

3.1 Estimation du modèle

Nous abordons désormais l'estimation du modèle statistique, dont le processus de spécification est présenté plus extensivement dans un article associé (Ay, 2019). Il s'agit d'estimer un modèle ordonné additivement semi-paramétrique (OGAM) qui prend en compte la structure hiérarchique des AOC de la zone, notée $y \in \{1, 2, 3, 4, 5\}$ par ordre croissant de qualité. Nous supposons que les désignations des AOC suivent une règle de décision basée sur une variable latente non observable de qualité des vignes qui franchit des seuils différents selon la commune à laquelle elle appartient. Notons X_i le vecteur des caractéristiques biophysiques de la parcelle de vigne i (avec $i = 1, \ldots, N$) et C_i le vecteur de dimension 31 qui a pour élément générique c_{ih} égal à 1 si i se situe dans la *commune* h et 0 sinon. L'hypothèse d'une distribution logistique de la partie aléatoire de la variable latente produit un modèle de logit ordonné classique (Agresti and Kateri, 2017) :

$$\text{Prob}(y_i > j \mid X_i, C_i) = \Lambda[B(X_i)^\top \beta + C_i^\top \mu - \alpha_j], \tag{1}$$

où Λ est la fonction cumulative de la loi logistique. Les déterminants humains qui ont impacté la classification AOC au cours de l'histoire sont pris en compte par les effets fixes communaux notés μ. En l'absence d'*a priori* théoriques sur l'effet de chaque variable biophysique X_i, nous les spécifions au travers d'une série de transformations additives *B-splines* que nous notons $B(\cdot)$ avec β le vecteur des coefficients associés. Ce modèle de désignation peut alors être 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 sur le processus de pénalisation semi-paramétrique des effets des variables continues.

Afin de contrôler les effets du terroir qui ne seraient pas pris en compte par les variables biophysiques présentées précédemment (par des effets de variables omises ou d'erreurs de mesure), nous incluons également les coordonnées géographiques des centroïdes des parcelles comme variables explicatives. Cela permet d'améliorer sensiblement les capacités prédictives du modèle et de proposer une estimation non biaisée des effets communaux (Ay, 2019). Nous estimons des modèles à des degrés divers de lissage spatial au travers du lissage des coordonnées géographiques. L'objet `gamod.Rda` téléchargeable sur https://data.inra.fr/ contient 10 modèles de désignation des AOC actuelles qui vont du moins lissé `gam50` au plus lissé `gam900`.

```
GamModRaw <- get_file("gamod.Rda", "https://doi.org/10.15454/ZZWQMN")
writeBin(GamModRaw, "gamod.Rda") ; load("gamod.Rda") ; names(gamod)
```

```
[1] "gam50"  "gam100" "gam200" "gam300" "gam400" "gam500" "gam600"
[8] "gam700" "gam800" "gam900"
```

Pour la reproductibilité des analyses, nous reportons ci-dessous le code pour l'estimation du modèle qui sera utilisé dans l'application, qui est celui avec le lissage le plus fin et qui présente les meilleures prédictions. Les coordonnées spatiales sont lissées avec un nombre maximal de degré de liberté effectifs de $k = 900$, ce qui est très flexible et implique une longue procédure estimation (environ 9 heures avec un processeur Intel Core i7-7820HQ CPU 2.90 GHz x8 et 64 Go de RAM). Le lecteur peut accéder directement au résultat de cette estimation avec l'objet `gamod$gam900` téléchargé précédemment sur le serveur.

```
library(mgcv)
## 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
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
```

Nous obtenons avec la fonction `anova` la significativité statistique des différentes variables inclues dans le modèle au regard des statistiques de χ^2. Les variables indicatrices et les effets fixes sont dans la partie paramétrique (`Parametric Terms`, reportée en premier) alors que les variables continues sont dans la partie lissée (`Approximate significance of smooth terms`, reportée en second). Toutes les variables introduites dans le modèle sont significative aux seuils habituellement utilisés, ce qui conforte notre hypothèse d'un modèle de désignation des AOC basé sur une variable latente de qualité des vignes. Dans ce modèle avec un lissage spatial très flexible, les coordonnées géographiques apparaissent les variables explicatives les plus importantes au sens du χ^2, 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 avec un lissage spatial fort produit près de 90% de bonnes prédictions des niveaux d'AOC pour un pseudo R^2 (au sens de McFadden) égal à 0.76. Les mêmes statistiques peuvent être obtenues localement pour les autres modèles présents dans l'objet `gamod`, moins lissés spatialement, mais ne sont pas reportées ici. Les résultats sur la significativité des variables et la forme des effets sont globalement robustes au degré de lissage des coordonnées géographiques.

```
sum(diag(table(cut(gamod$gam900$line,
                   c(-Inf, gamod$gam900$family$getTheta(TRUE), Inf)),
                   gamod$gam900$model[, 1])))/ nrow(gamod$gam900$model)*100
1- (logLik(gamod$gam900)/ logLik(update(gamod$gam900, . ~ + 1)))
```

```
[1] 89.48
'log Lik.' 0.7565 (df=964)
```

3.2 Effets des variables biophysiques

Nous pouvons alors représenter les effets marginaux de chaque variable biophysique sur la variable latente de qualité des vignes, soient les fonctions $B(\cdot)$ dans l'équation (1). La fonction `plot` définie par le package `mgcv` permet de représenter graphiquement chacun de ces effets additivement séparables.

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

La Figure 6 présente des effets en U inversés pour l'altitude et la pente, avec les vignes les mieux classées en termes d'AOC qui sont à environ 300 mètres d'altitude et 10 degrés de pente. L'effet du rayonnement solaire est plus linéaire, contrairement à ce que le troisième quadrant de la Figure peut le laisser apparaître. En effet, la plupart des parcelles sous AOC ont un rayonnement solaire centré-réduit compris entre -2 et 2,

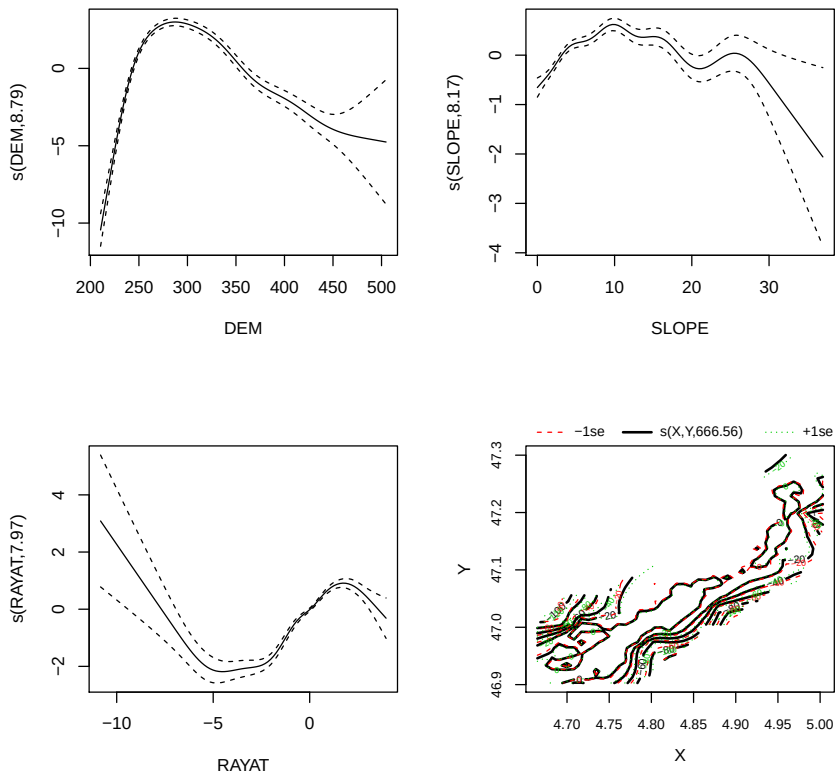

Figure 3: **Effets de la topographie et de la localisation sur la qualité des vignobles.**

Source : INRA / INAO / MSH / DGFiP / IGN / BRGM / GisSol.

Lecture : Les trois premières fenêtres représentent les effets de l'altitude, de la pente, et du rayonnement solaire en fixant toutes les autres variables explicatives du modèle à leurs moyennes de l'échantillon. Les effets ont une moyenne normalisée à 0 car leurs niveaux moyens n'est pas identifiable semi-paramétriquement (Wood, 2017). La dernière fenêtre en bas à droite représente l'effet joint de la longitude et de la latitude par le biais de lignes de niveau et de leurs intervalles de confiance.

soit la partie linéaire de la courbe représentée. Enfin, les effets spatiaux en bas à droite semblent se structurer dans une relation de centre/ périphérie par rapport aux altitudes intermédiaires. Des figures plus détaillées, qui contiennent en particulier les effets associés aux autres modèles moins lissés spatialement, sont reportées dans l'article associé (Ay, 2019). Le lecteur peut aussi reproduire ces effets pour d'autres modèles avec la fonction `plot` du package `mgcv`. La structure des effets reste cependant robuste à la spécification des effets et au lissage spatial, elle reste proche de ce qui est observé ici pour le modèle le plus lissé.

3.3 Effets communaux

Les coefficients associés aux effets fixes communaux sont d'un intérêt particulier, ils correspondent à la partie historique des désignations AOC actuellement en vigueur, soit la partie de la variable latente de qualité des vignes qui n'est pas expliquée par les caractéristiques biophysiques des vignes. Cette interprétation des effets fixes communaux fait écho à certains travaux d'historiens pour lesquels nos résultats offre une confirmation statistique. En effet, Christophe Lucand dans Wolikow and Jacquet (2011) évoque l'existence d'une hiérarchie implicite des communes comme des "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" (p. 68). 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 aux XIXe et XXe siècles, 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.

Pour faciliter l'interprétation des effets fixes communaux, nous traduisons les coefficients estimés en mesures de supériorités ordinales γ_A pour la commune A par rapport à la commune moyenne de la zone (Agresti and Kateri, 2017). Par définition,

$$\gamma_A = \Lambda\left[(\mu_A - \overline{\mu})/\sqrt{2}\right] \tag{2}$$

où μ_A représente l'effet fixe de la commune A et $\overline{\mu}$ la moyenne des effets fixes sur la zone d'intérêt. Ainsi, cette mesure de supériorité ordinale comprise entre -1 et 1 représente l'écart de probabilité qu'une parcelle de la commune A soit mieux classée qu'une parcelle aux caractéristiques biophysiques identiques mais localisée dans une commune au hasard. Des valeurs positives indiquent des communes avantagées et des valeurs négatives des communes désavantagées par les désignations AOC. Le code suivant calcule ces mesures pour l'ensemble des communes de la zone et les représente graphiquement dans la Figure 5. Les objets `plogi` et `mso.key` requis pour l'évaluation du code sont définis en Annexe 2.

```
library(latticeExtra) ; resum900 <- summary(gamod$gam900)
cf <- resum900$p.coeff[ 4: 31]- mean(resum900$p.coeff[ 4: 31])
dat.fig <- data.frame(LIBCOM=substr(names(gamod$gam900$coef[ 4: 31]),7,30),
                      REGION= c(rep("tomato", 12), rep("chartreuse", 16)),
                      OS= 2* plogi(cf)- 1,
                      OSi= 2* plogi(cf- 1.5* resum900$se[ 4: 31])- 1,
                      OSa= 2* plogi(cf+ 1.5* resum900$se[ 4: 31])- 1)
segplot(reorder(factor(LIBCOM), OS)~ OSi+ OSa,
        length= 5, draw.bands= T, key= mso.key,
        data= dat.fig[order(dat.fig$OS), ], center= OS, type= "o",
        col= as.character(dat.fig$REGION[order(dat.fig$OS)]),
        unit = "mm", axis = axis.grid, col.symbol= "black", cex= 1,
        xlab= "Mesure de supériorité ordinale et intervalles á 10 %")
```

Le communes relativement favorisées par la classification des AOC apparaissent en haut de la Figure 5 et les communes relativement défavorisées en bas. Les intervalles de confiance qui encadrent les valeurs moyennes sont différents de ceux reportés dans Ay (2019). Ils représentent ici l'incertitude associée à l'estimation des effets fixes communaux plutôt que l'incertitude associée à la spécification du lissage spatial. Les ordres de grandeur obtenus pour ces deux sources d'incertitude sont toutefois similaires. Nous observons que certaines mesures de supériorité ordinale suivent la hiérarchie des dotations brutes en AOC telles que présentées dans la Figure 2 de la Section 2.1, où les communes privilégiées sont celles qui possèdent les plus grosses proportions d'AOC en haut de la hiérarchie. Mais cette relation n'est pas systématique, certaines communes peu dotées en hauts niveaux d'AOC apparaissent également privilégiées. Parmi les 5 communes les plus privilégiées par la classification AOC, les communes de Vougeot et d'Aloxe-Corton sont relativement bien dotées en Premiers et Grands Crus, alors que ce n'est pas le cas pour les communes de Pernand-Vergelesses et de Chorey-les-Beaunes. À l'inverse, Chassagne-Montrachet ou Vosnes-Romanée possèdent de fortes proportions de Premiers et Grands Crus, sans que cela semble venir d'un traitement préférentiel dans la classification.

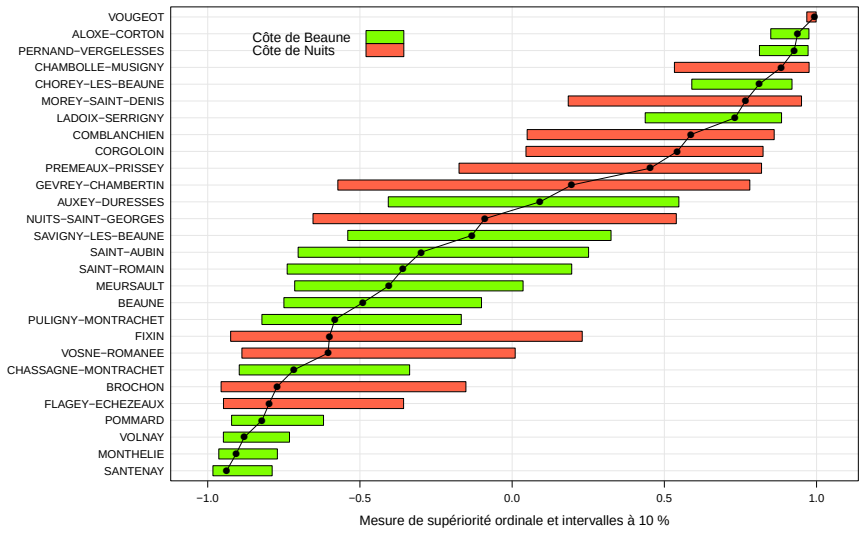

Figure 4: **Classification des communes selon les mesures de supériorité ordinale.**

Source : INRA / INAO / MSH / DGFiP / IGN / BRGM / Gis Sol.

Lecture : Pour chacune des communes de la zone (en ordonnées), la figure reporte la mesure de supériorité ordinale de la désignation des AOC par rapport à la commune moyenne, c'est-à-dire l'écart de probabilité qu'une même parcelle soit mieux classée dans la commune considérée que dans une commune de la zone prise au hasard. Les intervalles de confiance représentent l'incertitude associée à l'estimation des effets fixes communaux dans la modélisation statistique (à 90%).

3.4 Prédiction de la qualité continue

Les prédictions issues du modèle OGAM de désignation vont représenter les valeurs estimées de la variable latente de qualité des vignes pour chacune des parcelles de la zone. Nous obtenons ainsi un score continue pour chaque parcelle uniquement selon ses caractéristiques biophysiques et, selon que l'on prenne ou pas en compte sa commune d'appartenance, le traitement préférentiel dont elle a fait l'objet au cours de l'histoire. Notons que cette classification statistique des parcelles est directement issue des AOC qui existent aujourd'hui et ne se base pas sur des appréciations subjectives exogènes sur ce qui fait la qualité d'une vigne ou d'un vin. Le code suivant présente le calcul des prédictions et leur normalisation pour qu'ils soient distribués entre 0 et 100 (avec la fonction `unini`), pour l'ensemble des parcelles dans la base `Prd.Ras`. Notons que la ligne sur les prédictions, commentée, est assez longue (5 minutes) à évaluer dans R.

```
Prd.Ras <- subset(Geo.Ras, !is.na(AOClb))
Prd.Ras$GEOL <- ifelse(Prd.Ras$NOTATION%in%levels(gamod$gam900$model$GEOL),
                       as.character(Prd.Ras$NOTATION), "0AREF")
Prd.Ras$PEDO <- ifelse(Prd.Ras$NOUC %in% levels(gamod$gam900$model$PEDO),
                       as.character(Prd.Ras$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)
lapply(Prd.Ras@data[, c("UFraw", "UFcor")], summary)
```

```
$UFraw
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0    67.0    72.5    70.6    75.6   100.0

$UFcor
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0    67.4    71.7    70.8    76.6   100.0
```

Les prédictions sont donc disponibles avec (`UFcor`) et sans (`UFraw`) correction des effets communaux afin que, en plus de pouvoir consulter le classement corrigé des effets communaux, l'utilisateur puisse apprécier le niveau du biais dans la classification actuelle des AOC. Nous avons normalisé ces deux variables pour produire un score de classification des parcelles selon des valeurs comprises entre 0 et 100, avec des distributions qui apparaissent aplaties à gauche (les médianes sont supérieures aux moyennes).

3.5 Agrégation par lieux dits

Pour faciliter la consultation des résultats de la modélisation dans l'application *Shiny*, nous agrégeons les scores prédits sur la base d'un recodage des dénominations et des références aux lieux dits. Nous utilisons pour cela les lieux dits administratifs qui permettent en outre de localiser plus précisément les parcelles en niveaux Coteaux bourguignons, Bourgogne régional et Village pour lesquels la mention de la parcelle n'est pas reportée systématiquement sur l'étiquette des bouteilles de vin (cette pratique est néanmoins de plus en plus fréquente). Il s'agit également ici de renommer les dénominations géographiques complémentaires associées aux premiers crus pour qu'ils soient plus lisibles dans l'application.

```
Prd.Ras$NIVEAU <- as.character(revalue(factor(Prd.Ras$AOC), NVA))
Prd.Ras$NAME <- ifelse(Prd.Ras$AOC== 5, as.character(Prd.Ras$AOClb),
                ifelse(Prd.Ras$AOC< 4, as.character(Prd.Ras$LIEUDIT), NA))
for (i in 1: nrow(Prd.Ras)){
    if (is.na(Prd.Ras$NAME[ i])){
        Prd.Ras$NAME[ i] <- substr(Prd.Ras$AOClb[ i],
                                   regexpr(" cru+", Prd.Ras$AOClb[ i],
                                           perl= T)+ 5,
                                   nchar(as.character(Prd.Ras$AOClb[ 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, à partir des 60 000 parcelles cadastrales utilisées pour estimer le modèle statistique de classification, nous obtenons environ 2 400 localités, qui correspondent aux lieux dits pour les niveaux Coteaux Bourguignons, Bourgogne régional, et Village; ils correspondent aux dénominations retravaillées pour les Premiers crus; et aux appellations pour les Grands Crus. Cela fait une moyenne de 25 parcelles par localité.

Nous allons désormais regrouper la géographie des parcelles selon la variable `Contat` tout juste créée pour agréger les scores prédits. Les scores alors reportés au niveau des nouvelles localités seront calculés par moyennes pondérées par la surface de chaque parcelle qui les compose. Nous calculons également la position de chaque localité dans la hiérarchie continue issue de la modélisation par rapport à l'ensemble des

localités de la zone, ce qui permet de présenter en sortie du code ci-dessous les 10 localités les mieux notées sur la base des scores corrigés.

```
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)/ 1e4, 2),
                       SCORE_brut= round(weighted.mean(UFraw, AREA), 2),
                       SCORE_corrigé=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$RANG_corrigé <- round(rank(Poly.Ras$SCORE_corrigé)/
                               nrow(Poly.Ras)*100,2)
head(Poly.Ras@data[order(Poly.Ras$RANG_corrigé, decreasing= T), c(3, 4, 6, 7)], n= 10)
Poly.Ras$NIVEAU <- factor(Poly.Ras$NIVEAU, levels= NVA)
```

	NOM	NIVEAU	SCORE_brut	SCORE_corrigé
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

Comme on pouvait s'y attendre avec les pourcentage de bonnes prédictions obtenues, les Grands Crus arrivent en haut de la classification statistique, qu'ils appartiennent à la côte de Nuits (comme Chambertin et Grands-Echezeaux) ou à la côte de Beaune (comme Montrachet et Bâtard-Montrachet). Notons tout de même qu'un Premier cru arrive en dixième position, ce qui signifie qu'il dépasse les 2/3 des Grands crus de la zone. Ce Premier cru qui a pour dénomination "La Combe d'Orveau" se trouve sur la commune de Chambolle-Musigny, qui n'apparaît par ailleurs pas relativement désavantagée selon la Figure 5. 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. Également étonnant, le Grand cru 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 26ième position. Le lieu dit est tout de même dans les 2 % meilleures localités de la zone. Les résultats amènent à penser que la situation de monopole du domaine de la Romanée-Conti qui exploite ce climat peut expliquer le fort prix observé des bouteilles, indépendamment des caractéristiques biophysiques.

Nous enregistrons ensuite les résultats agrégés de la prédiction dans un objet de type `sf` défini par le package du même nom (Pebesma, 2018). Cette objet nommé `Poly.Ras` pourra directement être utilisée dans l'application *Shiny* pour consulter, lancer, ou modifier le classement statistique. Les résultats issus du recodage des dénominations et de l'agrégation des scores prédits sont accessibles sur le serveur data de

l'INRA à l'adresse https://data.inra.fr/.

```
library(sf) ; Poly.Ras <- st_as_sf(Poly.Ras)
Poly.Ras <- st_transform(Poly.Ras, crs= 4326)
save(Poly.Ras, file= "Inter/PolyRas.Rda")
st_write(Poly.Ras, "/home/jsay/geoInd/Application2/Inter/PolyRas.shp",
         delete_layer= TRUE)
```

Sauvegarde disponible sur le serveur de l'INRA, qui peut être aussi chargée comme précédemment avec le package `dataverse`.

4 Application *Shiny*

Il y a deux manières d'utiliser l'application *Shiny* (Chang et al., 2019) pour consulter les résultats de cette recherche. Le lecteur peut simplement consulter l'application par le biais d'un explorateur internet à l'adresse https://cesaer-datas.inra.fr/geoind/, ou alors il est possible de lancer l'application en local sur la base d'une version récente de R, des packages spécifiés ci-dessous et des données et codes téléchargés sur le serveur https://data.inra.fr/. Pour la première solution, le lecteur peut se reporter directement à l'exemple d'utilisation à la sous-section 4.3 suivante. Nous présentons dans les deux sous-sections suivantes le détails de la construction de l'application pour un exécution en local.

4.1 Cartographie dynamique

Une première étape pour générer l'application localement consiste à définir la cartographie dynamique de type `Leaflet` par le package `mapview` (Appelhans et al., 2018) qui sera intégrée ensuite à l'application. Cela nécessite la présence de l'objet `Poly.Ras` issu des traitements précédents, il peut être directement télécharger du serveur avec le package `dataverse`.

```
library(RColorBrewer) ; library(mapview) ; library(sf)
Poly.Ras <- st_read("Inter/PolyRas.shp")
Poly.Ras$NIVEAU <- factor(Poly.Ras$NIVEAU,
                          levels= c("Coteaux b.", "Bourgogne", "Village",
                                    "Premier cru", "Grand cru"))
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]))
```

L'objet `map` créé permet de faire apparaître sur un navigateur internet une cartographie dynamique pour visualiser les différentes parcelles viticoles de la zone et faire apparaître le score (corrigé des effets communaux ou pas) suite à un clic sur la parcelle. Il apparaît alors le nom de la commune d'appartenance (`LIBCOM`), le nom du lieu-dit (`NOM`), le niveau dans la hiérarchie des AOC (`NIVEAU`), la surface du lieu-dit

(SURFACE), le score brut (SCORE_b) et corrigé (SCORE_c), ainsi que la position du lieu dit dans la hiérarchie générale de la région (RANG_br pour la version brute et RANG_cr pour la version corrigée).

4.2 Lancer l'application localement

Une fois la cartographie dynamique effectuée, le code précédent et le code suivant doivent être enregistrés dans un fichier global.R en accord avec la structuration des applications *Shiny*. Les deux autres objets ui.R et server.R nécessaire pour lancer localement l'application *Shiny* sont reportés en Annexes 4 et 5. Ils contiennent respectivement le paramétrage de l'interface utilisateur et les calculs qui sont effectués sur le serveur pour l'interactivité de l'application. Le code ci-dessous source ces deux fichiers qui sont également disponibles sur un répertoire distant de la plateforme github.

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

source("ui.R") ; source("server.R")
## source("Application/ui.R") ; source("Application/server.R")
## source("Application2/ui.R") ; source("Application2/server.R")

enableBookmarking(store = "url")
shinyApp(ui,server)
```

La commande shinyApp(ui,server) lance l'application dans le navigateur internet par défaut, et permet d'obtenir localement le même résultat que la version en ligne de l'application.

4.3 Exemple d'utilisation

L'application est structurée en trois parties, avec en haut à gauche trois zones de saisie pour renseigner directement une localité (à partir des informations disponibles sur les étiquettes d'un vin ou à partir des connaissances personnelles de l'utilisateur); en bas à gauche un graphique qui positionne la qualité de la localité sélectionnée dans la distribution générale des qualité (où les niveaux d'appellation sont différentiés); et à droite la cartographie dynamique qui permet de faire apparaître la localité sélectionnée. L'utilisateur peut sélectionner un vin par une référence mais peut aussi se déplacer librement sur la zone d'étude, dans les parcelles voisines d'une référence par exemple. Les fonctions de la cartographie dynamique précédente sont maintenues, ainsi un clic sur une parcelle permet de faire apparaître ses caractéristiques, et les prévisions du modèle corrigées et non corrigées en particulier.

Prenons l'exemple d'un vin d'un niveau Premier cru, sur la commune de Flagey-Echezeaux, qui a pour nom de lieu dit (dénomination géographique complémentaire en l'occurrence) "Les Rouges". Suite à la saisie de ces caractéristiques dans les zones dédiées en haut à gauche, le score corrigé prédit pour ce vin égal à 83.82 apparaît dans le graphique en bas à gauche de l'application, et nous observons que ce score est supérieur à l'ensemble des Coteaux bourguignons, des Bourgognes régionaux et des villages de la zone. Ce vin de niveau Premier cru est parmi les 10% de Premiers crus qui ont les plus hauts scores et il dépasse même 30% des Grands crus de la zone. La partie à droite de l'application a zoomé sur cette zone, un clic sur le lieu dit en question permet de faire apparaître la différence entre les prédictions brutes et corrigées. Ainsi, le score non corrigé de la localité est sensiblement plus bas (80.92), ce qui implique que suite à la correction des effets communaux, le vin passe du top 7% au top 3% sur l'ensemble de la zone étudiée. Ce résultat est consistant

avec les effets communaux reportés dans la Figure 5, où la commune de Flagey-Echezeaux apparaît comme relativement défavorisée dans la hiérarchie. Nous pouvons également que ce Premier est mitoyen du Grand cru Echezeaux qui se trouve tout juste à l'Est.

5 Conclusion

Nous avons présenté la construction de données exhaustives sur les parcelles de 31 communes viticoles de la Côte-d'Or et l'utilisation de ces données pour une modélisation statistique des classifications AOC de la zone. Nous avons alors présenter une application cartographique de type *Shiny* qui permet de consulter ces résultats de manière interactive. Il est important de mentionner que la classification statistique obtenue, qui permet de préciser la hiérarchie des AOC, se base exclusivement des relations entre les caractéristiques biophysiques des parcelles et ne contient aucune appréciation subjective exogène sur l'importance des différents facteurs biophysiques ou olfactif censés influer sur la qualité d'un vin. L'approche statistique permet en outre de corriger les effets communaux qui sont montré comme ayant une influence sur les classement sans que ce soit justifié du point de vue des caractéristiques biophysiques des parcelles. La classification obtenue est donc directement 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.

Par contre, le modèle proposé n'est pas déterministe, et une incertitude persiste toutefois dans la classification obtenue. La hiérarchie des parcelle dépend de 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 peut faire débat, par opposition aux facteurs humains qui ont un effet indéniable sur 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 totalement 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. L'appariement de données supplémentaires pour relier les AOC aux caractéristiques des lieux est également possible par la géolocalisation des parcelles cadastrales. 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.

5.1 Remerciements

Nous tenons à remercier Pauline Mialhe, Vincent Larmet, et Guillaume Royer pour leur aide sur le développement de l'application.

References

Agresti, A. and Kateri, M. (2017). Ordinal probability effect measures for group comparisons in multinomial cumulative link models. *Biometrics* 73: 214–219.

Appelhans, T., Detsch, F., Reudenbach, C. and Woellauer, S. (2018). mapview: Interactive Viewing of Spatial Data in R. R package version 2.6.3.

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. *en révision* .

Bivand, R. S., Pebesma, E. and Gomez-Rubio, V. (2013). *Applied spatial data analysis with R, Second edition*. Springer, NY.

Capus, J. (1947). *L'Évolution de la législation sur les appellations d'origine. Genèse des appellations contrôlées*. L. Larmat.

Chang, W., Cheng, J., Allaire, J., Xie, Y. and McPherson, J. (2019). shiny: Web Application Framework for R. R package version 1.4.0.

Coestier, B. and Marette, S. (2004). *Economie de la qualité*. La découverte.

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., Wehberg, J., Wichmann, V. and Böhner, J. (2015). System for automated geoscientific analyses (saga) v. 2.1. 4. *Geoscientific Model Development* 8: 1991–2007.

Delay, E. and Chevallier, M. (2015). Roger Dion, toujours vivant! *Cybergeo: 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.

Hilal, M., Joly, D., Roy, D. and Vuidel, G. (2018). Visual structure of landscapes seen from built environment. *Urban Forestry & Urban Greening* 32: 71–80.

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.

Lavalle, J. (1855). *Histoire et statistique de la vigne et des grands vins de la Côte d'Or*. Daumier, Dijon.

Leeper, T. J. (2017). dataverse: R Client for Dataverse 4. R package version 0.2.0.

Lucand, C. (2017). *Le vin et la guerre: Comment les nazis ont fait main basse sur le vignoble français*. Armand Colin.

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal* 10: 439–446, doi:10.32614/RJ-2018-009.

R Core Team (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rich, P. M. and Fu, P. (2000). Topoclimatic habitat models. *Proceedings of the Fourth International Conference on Integrating GIS and Environmental Modeling* .

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 : Configuration logicielle

```
sessionInfo()
```

```
R version 3.6.0 (2019-04-26)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 18.04.2 LTS

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.7.1
LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.7.1

locale:
 [1] LC_CTYPE=fr_FR.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=fr_FR.UTF-8        LC_COLLATE=fr_FR.UTF-8
 [5] LC_MONETARY=fr_FR.UTF-8    LC_MESSAGES=fr_FR.UTF-8
 [7] LC_PAPER=fr_FR.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=fr_FR.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods
[7] base

other attached packages:
[1] latticeExtra_0.6-28 lattice_0.20-38     mgcv_1.8-28
[4] nlme_3.1-140        rgdal_1.3-6         sf_0.7-2
[7] mapview_2.6.3       RColorBrewer_1.1-2  sp_1.3-1

loaded via a namespace (and not attached):
 [1] Rcpp_1.0.0        compiler_3.6.0    later_0.7.5
 [4] base64enc_0.1-3   class_7.3-15      tools_3.6.0
 [7] digest_0.6.18     satellite_1.0.1   viridisLite_0.3.0
[10] png_0.1-7         Matrix_1.2-17     shiny_1.2.0
[13] DBI_1.0.0         crosstalk_1.0.0   e1071_1.7-0
[16] raster_2.8-4      htmlwidgets_1.3   stats4_3.6.0
[19] classInt_0.3-1    leaflet_2.0.2     grid_3.6.0
[22] webshot_0.5.1     R6_2.4.0          magrittr_1.5
[25] scales_1.0.0      codetools_0.2-16  promises_1.0.1
[28] htmltools_0.3.6   units_0.6-2       splines_3.6.0
[31] mime_0.6          xtable_1.8-3      colorspace_1.3-2
[34] httpuv_1.4.5      munsell_0.5.0
```

Annexe 2 : Configuration graphique

```
## Pour Figure 2
my.lab= c(BGOR= "Coteaux b.", BOUR= "Bourgogne", VILL= "Village",
          PCRU= "Premier cru", GCRU= "Grand cru")
library(RColorBrewer)
my.pal= brewer.pal(n= 9, name = "BuPu")[ 2: 8]
library(lattice)
my.key= list(space= "top", points= F, rectangles= T, columns= 5, text= my.lab)
my.par= list(superpose.polygon= list(col= my.pal))
my.pan= 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, ""))}
## Pour Tableau 7
library(stargazer)
tb.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 [$315-360$]")
## Pour Figure 4
plogi <- function(x) exp(x/ sqrt(2))/ (1+ exp(x/ sqrt(2)))
mso.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")))
```

Annexe 3 : Unités géologiques et pédologiques

Figure 5: **Distribution des 31 formations géologiques de la zone.**

Source : INRA / BRGM.

Lecture : La carte représente les 31 unités géologiques qui regroupent des sous-sols homogènes selon la formation géologique renseignée dans la base de données. Ces unités seront simplifiées avant l'inclusion dans le modèle statistique.

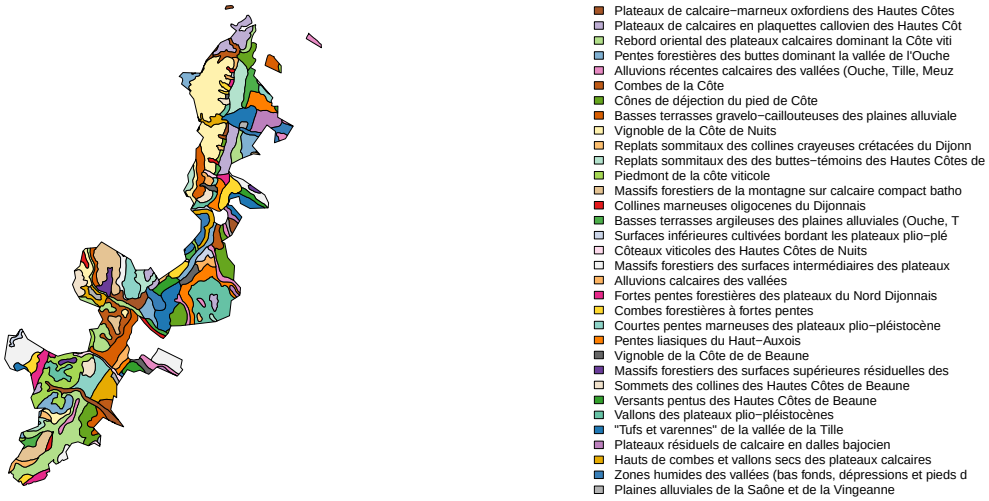

Figure 6: **Distribution des unités cartographiques de sols et leurs intitulés.**

Source : INRA / Gis Sol.

Lecture : La carte représente les 33 unités cartographiques de sols qui correspondent à des pédo-paysages homogènes en termes de climat, topographie et géologie. Ces unités seront simplifiées avant l'inclusion dans le modèle statistique.

Annexe 4 : Interface utilisateur de l'application

```
ui <- dashboardPage(
  dashboardHeader(
    titleWidth= 400,
    title= "Classification des vignobles de la Côte d'Or"),
  dashboardSidebar(disable = TRUE),
  dashboardBody(
    fluidRow(
      box(width= 2, height= 375,
          selectInput("niveau", label= "Niveau de l'appellation",
                      choices=
                          c(as.character(unique(Poly.Ras$NIVEAU))),
                      selected= 1),
          selectInput("commune", label= "Commune de la parcelle",
                      choices= c(
                          as.character(unique(Poly.Ras$LIBCOM)),
                          "TOUTES"), selected= 1),
          selectInput("nom", label= "Lieu dit de la parcelle",
                      choices= c(
                          as.character(unique(Poly.Ras$NOM)),
                          "TOUS"), selected = 1)),
      box(width= 3, height= 645, plotOutput("miplot")),
      box(width= 3,
          column(width = 12,
                 leafletOutput("mymap", height= 645),
                 fluidRow(verbatimTextOutput("mymap_shape_click"))
          )
      )
    )
  )
)
```

```
ui <- function(request){
    fluidPage(sidebarLayout(
        sidebarPanel(
            width= 2,
            selectInput("niveau",label= "Niveau de l'appellation",
                        choices=
                            c(as.character(unique(Poly.Ras$NIVEAU))),
                        selected= 1),
            selectInput("commune", label= "Commune de la parcelle",
                        choices= c(
                            as.character(unique(Poly.Ras$LIBCOM)),
                            "TOUTES"), selected= 1),
            selectInput("nom", label= "Lieu dit de la parcelle",
                        choices= c(
                            as.character(unique(Poly.Ras$NOM)),
                            "TOUS"), selected = 1),
            bookmarkButton()
        ),
        mainPanel(
            width= 3,
            tabsetPanel(id = "inTabset",
                        tabPanel(title = "Carte",
                                 leafletOutput("mymap", height= 645)),
```

```
                    tabPanel(title = "Figure", plotOutput("miplot")),
                    tabPanel(title = "Détails",
                             includeMarkdown("README.md"))
                    )
    )
))}
```

Annexe 5 : 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()$SCORE_corrigé
        if (length(yop)==0) return(NULL)
        top <- round(100-
                     aggregate(I(Poly.Ras$SCORE_corrigé< yop)* 100,
                               by= list(Poly.Ras$NIVEAU), mean)[, 2])
        ggplot(Poly.Ras, aes(x= factor(NIVEAU),
                             y= SCORE_corrigé, 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 á 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 = 400)}
```

```
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_c
    if (length(yop)==0) return(NULL)
    top <- round(100-
                 aggregate(I(Poly.Ras$SCORE_c< yop)* 100,
                           by= list(Poly.Ras$NIVEAU), mean)[, 2])
    ggplot(Poly.Ras, aes(x= factor(NIVEAU),
                         y= SCORE_c, 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, "%"), col= "red", size= 5)+
        labs(title= "Comparaison avec les autres parcelles",
             x= "",
             y = "Niveau sur une échelle de 1 á 100")+
        scale_fill_manual(values= AocPal)+
        theme(legend.position= "none",
              plot.title = element_text(hjust = 0, size = 16),
              axis.text.x = element_text(size= 12),
              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= 5)
}, height = 500, width = 400)}
```
