Analyses Factorielles et Régressions Logistiques réalisées à partir de données récoltées dans le cadre de l'Observatoire National des Maladies du Bois de la Vigne

Frédéric Bertrand, Myriam Maumy, Lionel Fussler, Nathalie Kobes, Serge Savary, Jacques Grosman

ANALYSES FACTORIELLES ET RÉGRESSIONS LOGISTIQUES RÉALISÉES À PARTIR DE DONNÉES RÉCOLTÉES DANS LE CADRE DE L'OBSERVATOIRE NATIONAL DES MALADIES DU BOIS DE LA VIGNE

Frédéric Bertrand[(1)], Myriam Maumy[(1)], Lionel Fussler[(2)], Nathalie Kobes[(2)], Serge Savary[(3)], Jacques Grosman[(4)]

[(1)] Institut de Recherche Mathématique Avancée, Université Louis Pasteur.
Email : fbertran@math.u-strasbg.fr, mmaumy@math.u-strasbg.fr

[(2)] DRAF-SRPV Alsace.
Email : nathalie.kobes@agriculture.gouv.fr, lfussler@wanadoo.fr

[(3)] UMR Santé Végétale, Centre INRA de Bordeaux.
Email : ssavary@bordeaux.inra.fr

[(4)] DRAF-SRPV Rhône.
Email : jacques.grosman@agriculture.gouv.fr

RÉSUMÉ.

L'objectif de l'Observatoire National des Maladies du Bois de la Vigne est de faire un état des lieux de la répartition, de la fréquence et de l'intensité de l'expression des symptômes foliaires des maladies du bois, pour répondre objectivement à la question de leur progression éventuelle dans le vignoble français. En effet, suite à l'interdiction de l'utilisation de l'arsénite de soude le 8 novembre 2001, les viticulteurs ne disposent plus d'aucune méthode de lutte chimique « curative » contre les maladies du bois que sont l'esca, l'eutypiose et le black dead arm. Cet observatoire génère ainsi, chaque année, depuis sa création en 2003 par le ministère chargé de l'agriculture, un ensemble de données cohérentes susceptibles d'être abordées par différentes méthodes d'analyse statistique.

Le jeu de données pour cette étude est complexe. En effet, il comporte des variables quantitatives et qualitatives qui évoluent au fil du temps, étant donné que les questionnaires ont été récoltés de 2003 à 2005. La problématique de l'étude est de dégager les grandes tendances en matière d'épidémiologie végétale présentes dans le jeu de données afin de déterminer quelles sont les mesures prophylactiques adaptées à mettre en œuvre collectivement et à grande échelle. Pour aborder le problème, nous avons utilisé des techniques descriptives et inférentielles.

Le travail statistique s'est donc articulé en trois points. D'abord, nous nous sommes efforcés de mettre en évidence des relations entre les différentes variables de l'étude, puis nous avons utilisé des méthodes factorielles, à savoir l'analyse des correspondances multiples et l'analyse factorielle de données mixtes. Puis, afin de préciser les relations que nous avons pu mettre en évidence, notre choix s'est porté sur la construction et l'évaluation de modèles de régressions logistiques binaire et ordinale pour l'étude desquels nous avons utilisé des techniques bootstrap.

Enfin, pour tenir compte du facteur temps, nous avons employé des techniques d'analyse factorielle de tableaux multiples, chaque tableau représentant une année de données. Ces différentes techniques seront rappelées ici, ainsi que les résultats des analyses réalisées.

Mots clés : Analyse des correspondances multiples, Analyse factorielle de données mixtes, Régression logistique binaire, Régression logistique ordinale, Bootstrap, Analyse factorielle de tableaux multiples.

ABSTRACT. Analysing the data of the Grapevine Wood Diseases National Observatory using factorial analyses and logistic regressions.

The Grapevine Wood Diseases National Observatory yields a cohesive and large data set which may be dealt with using different approaches. In our study, we have to deal with complex data, as it is composed of quantitative and qualitative variables which evolve with time, since data for three successive years are available. The objective was to produce of the largest possible amount of information from this data set, in order to highlight main trends. To this aim, we used both descriptive and inferential techniques. Our study thus developed in three points. First, relationships between the different variables are identified using factorial methods, namely multiple correspondence analysis and factor analysis of mixed data. Then, binary and ordinal logistic regressions were used to specify the relationships we highlighted; confidence intervals for the parameters were constructed using bootstrap techniques. Lastly, we used factorial analysis of multi-tables, each table representing a year, in order to account for the successive years of data. Those different techniques will be explained here, as well as our results.

Keywords: Multiple correspondence analysis, Factor analysis of mixed data, Binary logistic regression, Ordinal logistic regression, Bootstrap, Factor analysis of multi-tables.

1 Historique de la situation

Depuis l'année 2001 et l'interdiction, en France puis en Europe1, de l'usage de l'arsénite de soude en agriculture, il n'existe plus de méthode de lutte curative homologuée contre les maladies du bois de la vigne, à savoir l'eutypiose, l'esca et le black dead arm (BDA). L'arsénite de soude est un composé chimique à base d'arsenic dont il n'est plus possible de se servir pour des raisons de santé publique puisque les risques cancérigènes pour les viticulteurs sont importants et avérés. Suite à l'interdiction de ce produit phytosanitaire, la profession viticole a exprimé des craintes par rapport à la progression potentielle de ces maladies de dépérissement au sein du vignoble français, dont la conséquence ultime est bien souvent la mort des plants. De plus, le manque de références techniques sur l'épidémiologie de ces maladies ne permettait pas d'envisager une solution rapide de remplacement.

Afin de coordonner l'ensemble des dispositifs de recherche et d'acquisition de références sur ces maladies, un groupe technique national a été constitué en 2001. C'est dans le cadre de ce groupe technique national qu'a été décidée la création de l'observatoire national pluriannuel des maladies du bois de la vigne qui fait l'objet de cette étude. Ce projet est unique au monde du fait de la taille du dispositif et du nombre de structures impliquées : il est coordonné par le service de la protection des Végétaux, sous l'égide de l'ONIVINS et associe différents partenaires techniques de la filière. Ce type d'étude n'est réalisé généralement que sur une région ou sur un cépage, alors qu'ici l'ensemble des régions viticoles françaises est considéré. D'une durée de six années, de 2003 à 2009, ce dispositif a pour objectif de produire un état des lieux de la situation des maladies au plan national, d'évaluer l'impact économique réel de l'arrêt des traitements à l'arsénite de soude et de tenter d'identifier les facteurs prédominants permettant d'expliquer les variations observées de ces maladies.

2 Protocole technique de l'observatoire

Le protocole technique a été défini par la note de service DGAL/SDQPV/N2003-8085 du 19 mai 2003[2]. Nous en rappelons ici les principales caractéristiques :

[1] Le règlement 2076/2002/CE du 20 novembre 2002 impose aux États membres de l'Union Européenne d'interdire l'utilisation des préparations à base d'arsénite de soude au plus tard le 31 décembre 2003.

2 Cette note est disponible à l'adresse suivante : http://www.agriculture.gouv.fr/spip/IMG/pdf/dgaln20038085.pdf.

- Toutes les régions viticoles françaises présentent au moins l'une des maladies et justifient donc de participer à la base de sondage. En pratique seules 11 d'entre elles ont été enquêtées : Alsace, Aquitaine, Beaujolais, Bourgogne, Centre, Diois, Jura, Languedoc-Roussillon, Provence-Alpes-Côte d'Azur, Poitou-Charentes, Pays de Loire.
- Un minimum de 25 parcelles par vignoble et par cépage devait être observé, pour faire un état des lieux satisfaisant de la répartition, de la fréquence et de l'intensité de ces maladies pour un cépage donné, dans un vignoble donné.
- Pour chaque parcelle choisie aléatoirement à partir des réseaux d'observation, 300 ceps sont marqués et repérés, répartis en dix placettes de 30 ceps choisies également aléatoirement.
- Les observations se font sur les mêmes placettes, donc les mêmes ceps, pendant les trois années d'observation.
- Les maladies sont observées chaque année à deux périodes particulières du stade de la vigne : la floraison pour l'eutypiose et la véraison pour l'esca et le BDA. À l'occasion de ces relevés est également quantifiée la mortalité des plants présents sur la parcelle.
- Enfin l'unique critère de choix de la parcelle est le cépage retenu.

Ce protocole a été proposé par le ministère chargé de l'agriculture et validé par le groupe technique national auquel appartiennent l'ITV[3] et l'INRA. Il reprend les ordres de grandeur des nombres d'observations qui sont habituellement réalisées lors des expérimentations du même type.

Quant au critère de choix de la parcelle, le cépage constitue un critère d'identification simple et donc adapté à un dispositif de cette ampleur comportant de nombreux intervenants, puisque plus de 40 structures différentes effectuent les relevés. D'autre part, des différences de vulnérabilité importantes ont été observées entre les cépages, et ce même au sein d'une même région viticole. De ce fait, une hypothèse ressentie par la communauté de la santé végétale est que le cépage est un des facteurs explicatifs principaux qu'il est donc nécessaire de prendre en compte lors de la planification de l'enquête.

En complément, des informations sur les caractéristiques de la parcelle et les pratiques culturales ont été recueillies par enquête auprès des exploitants viticoles dont les parcelles avaient été choisies aléatoirement.

3 Etude statistique réalisée

L'esca et le BDA ayant les mêmes symptômes apparents, ils sont traités comme une unique maladie que nous notons esca/BDA. Nous disposons de données sur trois années 2003, 2004, 2005, concernant les deux maladies du bois de la vigne à savoir l'eutypiose et l'esca/BDA. Pour expliquer les variations des intensités observées de ces deux maladies dans une population de parcelles cultivées, ainsi que celle de la mortalité des plants, les variables suivantes ont été prises en compte dans l'étude statistique : la région, le cépage, l'âge de la parcelle, la densité de plantation, le type de porte-greffe utilisé, mais encore le devenir des sarments, l'enlèvement des bois morts, le prétaillage ou non des vignes, le type de taille ainsi que les dates de début et de fin de taille et bien entendu le nombre de traitements à l'arsénite de soude effectués sur la période 1999-2001 c'est-à-dire juste avant l'interdiction de son utilisation. D'autres données telles que la vigueur de la vigne, le type de sol ou encore la superficie de la parcelle ont été également renseignées. Cependant le manque de réponses ou de fiabilité de celles-ci nous a amené à ne pas les prendre en compte dans le cadre de cette étude.

Au total, nous disposons d'un jeu de données représentant 701 parcelles observées dans 11 régions viticoles, correspondant à 26 cépages, et rassemblées suivant un protocole précis de saisie, et ce pour chaque année de l'étude. Il s'agit des parcelles pour lesquelles toutes les informations ont

[3] Institut Technique de la Vigne et du Vin.

été renseignées. En effet, les parcelles présentant des données manquantes ou aberrantes ont été systématiquement éliminées de la base de données.

Pour répondre à l'ensemble de questions posées préalablement, plusieurs méthodes statistiques ont été envisagées.

Afin de pouvoir appliquer dans les meilleures conditions ces méthodes statistiques, une analyse exploratoire des données a été réalisée dans un premier temps afin de pouvoir dégager les premières tendances et les liaisons entre les différentes variables. Celle-ci a essentiellement mis en évidence l'importance de l'âge de la parcelle notamment dans l'expression des symptômes d'esca/BDA.

Rappelons que l'incidence d'une pathologie est définie comme le pourcentage de pieds atteints par la pathologie parmi l'ensemble des pieds d'une parcelle, et que la mortalité des plants est estimée en cumulant pour chaque parcelle les ceps morts, manquants et les jeunes complants.

Par la suite, des tests bivariés ont été effectués, comme les tests du χ^2 d'indépendance détaillés dans le livre de Agresti (1990), les tests non paramétriques de Kruskal-Wallis, exposés dans le livre de Siegel et de Castellan (1988) et les tests exacts de Fisher-Freeman-Halton d'indépendance, présentés dans l'article de Freeman et de Halton (1951). Ceux-ci ont permis de restreindre l'ensemble des variables à considérer dans l'analyse statistique. Ainsi, seules les variables exprimant des liens avec les variables à expliquer, c'est-à-dire les incidences des deux maladies et le taux de mortalité des plants ont été retenues pour la suite de l'étude.

Suite à ces analyses préliminaires, le but de la première partie de l'analyse est de dégager des grandes tendances en termes d'épidémiologie végétale. Pour cela, nous avons utilisé des techniques d'analyse factorielle sur le tableau moyen[4] des trois années (2003, 2004, 2005) : l'analyse factorielle des correspondances multiples, décrite dans le livre de Escofier et Pagès (1998), et l'analyse factorielle de données mixtes, exposée dans l'article récent de Pagès (2004).

Dans la deuxième partie de l'analyse, le jeu de données a été modélisé en utilisant des techniques de statistique inférentielle. Compte tenu de la nature binaire de la réponse observée, la plupart du temps la mortalité des plants, notre choix s'est porté sur les modèles de régressions logistiques, aussi bien binaires qu'ordinaux, tous deux détaillés dans le livre de Davison (2003). Etant donné les associations complexes existant entre les deux variables « région » et « cépage », nous avons été amenés à introduire des modèles hiérarchiques. Enfin, nous avons utilisé des méthodes bootstrap pour établir des intervalles de confiance pour les paramètres des modèles de régressions logistiques. Ces méthodes sont discutées dans le livre de Davison et de Hinkley (1997).

Enfin, afin d'appréhender l'évolution temporelle des deux maladies du bois de la vigne et de la mortalité des plants, mais aussi pour tenir compte de la variabilité inter annuelle des conditions climatiques dont il est fortement supposé qu'elle ait eu une influence sur l'expression des deux maladies du bois de la vigne[5], des analyses factorielles de tableaux multiples, résumées dans l'article de Cazes (2004), ont été réalisées. Chaque tableau de données représente alors une année d'observation.

[4] Les valeurs des incidences de l'eutypiose, de l'esca/BDA et les pourcentages de mortalité utilisés correspondent aux moyennes des valeurs observées sur les trois années 2003, 2004 et 2005. Les valeurs des autres variables n'évoluent pas au cours du temps, aucune transformation les concernant n'est donc nécessaire.

[5] Nous rappelons que l'année 2003 a été marquée par un été caniculaire.

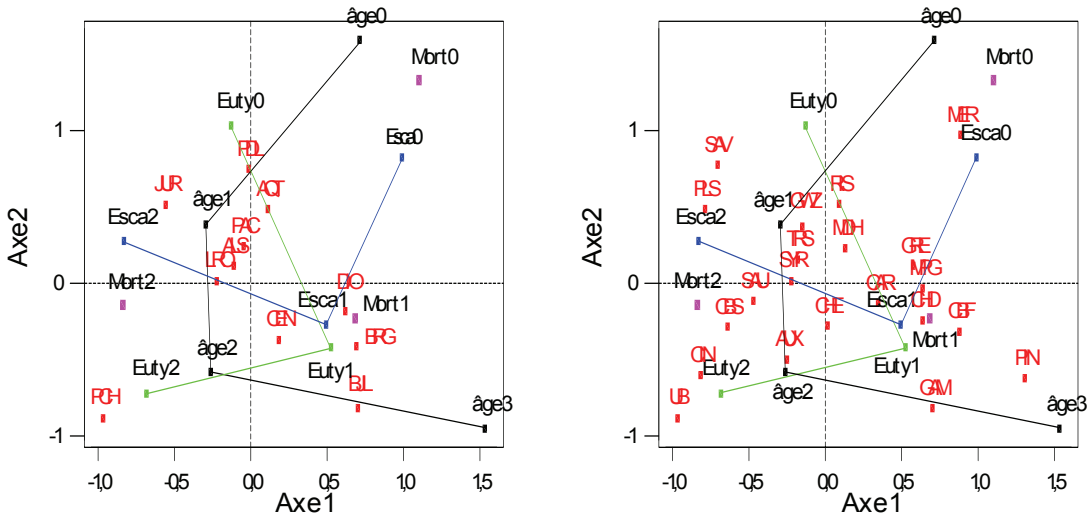

Figure 1

Représentations graphiques des deux premiers axes factoriels de l'analyse factorielle des correspondances multiples des variables « Eutypiose », « Esca/BDA », « âge » et « Mortalité » préalablement codées[6] avec respectivement la région puis le cépage mis en supplémentaire

4 Conclusion

Cette étude de cas montre la complémentarité de différentes méthodes statistiques. Son objectif est d'élaborer une méthodologie statistique performante pour traiter des données quantitatives, qualitatives et dynamiques. En effet, d'un point de vue biologique, les relations entre les variables étudiées ont fait l'objet de peu de travaux expérimentaux permettant d'affirmer leur association. Il s'agit donc d'analyser en profondeur le jeu de données, en portant une attention plus particulière aux effets de l'arsénite de soude sur les maladies du bois de la vigne et à l'évolution de celles-ci au cours du temps.

Ainsi, l'analyse des correspondances multiples nous a permis d'établir un panorama des associations entre les variables les plus importantes. L'analyse factorielle de données mixtes nous a confirmé les résultats de l'analyse des correspondances multiples et a dégagé d'autres relations entre les variables. La régression logistique nous a permis de modéliser ces relations, nos variables à expliquer étant de type binaire ou ordinal. Enfin, l'analyse des correspondances multiples multi-tableau a rendu possible la prise en compte du facteur temps et ainsi de juger de l'évolution des maladies du bois de la vigne et de la mortalité des ceps. Avec l'apport des données des trois années à venir de nouvelles techniques pourront être mises en œuvre, notamment pour mieux appréhender la question de l'évolution de ces trois maladies.

Un premier résultat essentiel dégagé par ces analyses est que ces maladies présentent des variations importantes, significatives, selon les cépages et donc selon certaines régions viticoles. Par la suite, nous avons pu remarquer que, de manière individuelle, l'eutypiose évolue essentiellement et parallèlement avec l'âge de la vigne, alors que l'esca/BDA touche particulièrement les plants jeunes et est lié aux traitements à l'arsénite de soude dans sa phase d'intensification, c'est-à-dire le passage des incidences d'esca/BDA de faibles à fortes, ce qui n'est pas le cas pour l'eutypiose. En ce qui concerne la mortalité des plants, les facteurs prépondérants sont l'eutypiose, l'esca/BDA et l'âge de la vigne. Quant aux traitements à l'arsénite de soude, ils n'apportent pas de contribution significative à la description de l'évolution de la mortalité.

[6] La signification des codes sont donnés dans la section 5 Annexes.

5 Annexes

Le questionnaire distribué aux viticulteurs qui ont participé à l'enquête est téléchargeable à l'adresse suivante : http://www-irma.u-strasbg.fr/~fbertran/recherche/QuestionnaireBois.pdf.

Variable	Symbole	Définition des catégories	Unité
Maladies et mortalité			
Eutypiose	Euty	Euty0 : Euty = 0 ; Euty1 : 0 < Euty ≤ 2 ; Euty2 : 2 < Euty	%
Esca/BDA	Esca	Esca0: Esca= 0 ; Esca1: 0 < Esca ≤ 3; Esca2: 3 < Esca	%
Mortalité	Mort	Mort0: 0 ≤ Mort < 3; Mort1: 3 ≤ Mort < 10; Mort2: 10 ≤ Mort	%
Variables explicatives			
Âge de la parcelle	âge	age0: 0 ≤ age < 15; age1: 15 ≤ age < 25; age2: 25 ≤ age < 40; age3: 40 ≤ age	années
Complémentaires			
Région	-	ALS : Alsace ; AQT : Aquitaine ; BJL : Beaujolais ; BRG : Bourgogne ; CEN : Centre ; DIO : Diois ; JUR : Jura ; LRO : Languedoc-Roussillon ; PAC : PACA ; PCH : Poitou-Charentes ; PDL : Pays de Loire	aucune
Cépage	-	AUX : Pinot Auxerrois; CAR : Carignan ; CBF : Cabernet Franc ; CBS : Cabernet Sauvignon; CHD : Chardonnay ; CHE : Chenin ; CIN : Cinsault ; GAM : Gamay ; GRE : Grenache ; GWZ : Gewurztraminer ; MDH : Muscat De Hambourg ; MEL : Melon ; MER : Merlot ; MPG : Muscat Petits Grains ; PIN : Pinot Noir ; PLS : Poulsard ; RIS : Riesling ; SAU : Sauvignon ; SAV : Savagnin ; SYR : Syrah ; TRS : Trousseau ; UB : Ugni Blanc	aucune

Bibliographie

[1] Agresti, A. (1990) *Categorical Data Analysis*, John Wiley & Sons, New York.

[2] Cazes, P. (2004) Quelques méthodes d'analyse factorielle d'une série de tableaux de données. *Revue de MODULAD*, n°31, Paris.

[3] Davison, A.C. (2003) *Statistical Models*, Cambridge University Press, New York.

[4] Davison, A.C. and Hinkley, D.V. (1997) *Bootstrap Methods and their Applications*, Cambridge University Press, New York.

[5] Draper, N.R. and Smith, H. (1998) *Applied regression analysis*, 3rd Edition, Wiley series in probability and statistics, New York.

[6] Escofier, B. et Pagès, J. (1998) *Analyses factorielles simples et multiples*, 3ème édition, Dunod, Paris.

[7] Freeman, G.H. and Halton, J.H. (1951) Note on an exact treatment of contingency, goodness of fit and other problems of significance, *Biometrika*,38, 141-149.

[8] Gordon, A.D. (1999) *Classification*, 2nd Edition, Chapman & Hall, New York.

[9] Hand, D.J. (1981) *Discrimination and Classification*, John Wiley & Sons, New York.

[10] Hosmer, D.W. and Lemeshow, S. (2000) *Applied logistic regression*, 2nd Edition, John Wiley & Sons, New York.

[11] Mosteller, F. and Tuckey, J.W. (1977) *Data analysis and regression*, Addison-Wesley, Boston.

[12] Pagès, J. (2002) Analyse factorielle multiple appliquée aux variables qualitatives et aux données mixtes, *Revue de. Statistique Appliquée*, L (4) 5-37.

[13] Pagès, J. (2004) Analyse factorielle de données mixtes, *Revue de Statistique Appliquée*, LII (4) 93-111.

[14] Siegel, S. and Castellan, N.J. (1988) *Nonparametric statistics for the behavioural sciences*, 2nd Edition, McGraw-Hill, New York.