Identification de profils de pratiques en élevages porcins

Solène Le Manac'H

Master 2 Data Science pour la biologie

Rapport de stage présenté par :
Solène Le Manac'h

Identification de profils de pratiques en élevages porcins

Maîtres de stage :
Arnaud Buchet (Cooperl) et Mathieu Emily (Agrocampus Ouest)

Année : 2019-2020

REMERCIEMENTS

Je suis très reconnaissante envers l'entreprise Cooperl Arc Atlantique qui m'a accueillie pendant ces 6 mois de stage au sein de sa structure de Plestan et m'a permis d'acquérir une expérience professionnelle très enrichissante.

Tout d'abord, je tiens à remercier Anne Lacoste de m'avoir accueillie au sein du service R&D Transversal.

Je remercie Arnaud Buchet, mon tuteur de stage de m'avoir apporté ses connaissances métier pour mener à bien la mission confiée. Et d'avoir été disponible pour m'épauler et me guider tout au long de ce stage.

Je tiens également à remercier Bastien Riera, chef de projets Data d'avoir toujours été disponible pour répondre à mes questions et d'avoir apporté son expertise statistique sur ce projet.

Je remercie également Mathieu Emily, tuteur pédagogique et Maître de conférences à l'Agrocampus Ouest pour son aide tout au long de ce stage.

Merci également à François Husson d'avoir pris le temps de répondre à mes questions.

Merci à tous les membres de la plateforme Cooperl Nutrition d'avoir permis mon intégration au sein du service.

Je remercie toutes personnes ayant contribué de près ou de loin au bon déroulement de mon stage.

Identification de profils de pratiques en élevages porcins

Rédigé par Solène Le Manac'h
Sous la direction de Arnaud Buchet, tuteur Cooperl
et Mathieu Emily, tuteur Agrocampus Ouest

RÉSUMÉS

Les performances techniques d'un élevage porcin sont communément mesurées à l'aide de plusieurs indicateurs : l'âge à 115 kg, l'indice de consommation global et la productivité annuelle. Seulement, ces indicateurs ne prennent pas en compte les données de la filière. La Cooperl dispose d'une présence suffisamment large sur toute la filière porcine pour pouvoir collecter les informations tout au long du cycle de vie du porc allant de la naissance du porcelet chez l'éleveur à l'abattage en passant par l'alimentation, la gestion technico-économique du troupeau, l'achat des produits vétérinaires, l'achat de petits équipements, le type de reproduction de l'élevage, le type d'installations de l'élevage ainsi que les contrôles sanitaires réalisés à l'abattage. Toutes ces données représentent une mine d'informations inédites dans ce secteur. C'est pourquoi nous nous intéressons à ce jeu de données pour donner une vision plus complète de la performance des élevages que ce qui existe déjà aujourd'hui. L'objectif principal étant d'identifier des profils de pratiques dans les élevages. Ce travail se fait en deux phases principales : l'analyse des liens pouvant exister entre les variables de la base à l'aide de statistiques descriptives et l'identification de profils de pratique d'élevages à l'aide d'une Analyse Factorielle Multiple (AFM) et d'une Classification Hiérarchique sur Composantes Principales (HCPC).

Mots clés : Performance, Analyse macrofilère, Jeu de données, R, Analyse Factorielle Multiple, Classification Hiérarchique sur Composantes Principales

ABSTRACT

The performances of a pig farm are commonly measured with several indicators like the age at 115 kg, the global feed conversion rate and the number of piglets produced per present sows per year (PWSY). However, most of them don't take into account the entire chain datas. Cooperl has an enough huge presence in the entire pork industry to collect informations throughout the pork life cycle ranging from the piglet birth at farm to slaughter by way of the alimentation, the herd's technico-economic management, the purchase of veterinary products, the purchase of small equipments, the reproduction type in farm, the type of installations in farm as well as controls carried out at slaughter. All these datas represent a mine of new information in this sector. This is why we are interested in this dataset to give a more complete view of the performance of the farms than what already exists today. The main objective of this is to identify farm practice profiles. This work is done in two main phases : analysis of the possible links between the base variables using descriptive statistics and the identification of farm practice profiles using Multiple Factor Analysis (MFA) and Hierarchical Clustering on Principal Components (HCPC).

Key words : Performance, Macro-sector analysis, Dataset, R, Multiple Factor Analysis, Hierarchical Clustering on Principal Components

LISTE DES SIGLES ET ABRÉVIATIONS

A115KS : Age à 115 kilogrammes

ACM : Analyse des Correspondances Multiples

ACP : Analyse en Composantes Principales

AFM : Analyse Factorielle Multiple

ALIM : Base de données relatives à l'aliment et ses supplémentations

CAH : Classification Ascendante Hiérarchique

CEDEV : CEntre DE Valorisation des déchets

COOPERL : Coopérative des Eleveurs de la Région de Lamballe

FAF : Fabrication d'Aliment à la Ferme

FP : Farmapro

GTE : Gestion Technico-économique de l'Elevage

GTTT : Gestion Technico-économique du Troupeau de Truies

HCPC : Hierarchical Clustering on Principle Components (ou Classification Hiérarchique sur Composantes Principales)

ICG : Indice de Consommation Global

KPI : Key Performance Indicator

PRODAN : PRODuctivité ANnuelle (nombre de porcs produits par truie présente par an)

PSA : Porcs élevés Sans Antibiotiques

REPRO : Base de données relatives aux mâles reproducteurs (source : service de Génétique)

SVH : Service Vétérinaire Hyovet : Base de données relatives aux produits vétérinaires (antibiotiques, vaccins, antiparasitaires) achetés chez le cabinet vétérinaire partenaire Hyovet

TMP : Taux de Muscle par Pièce

SOMMAIRE

INTRODUCTION

Cooperl Arc Atlantique est une coopérative agroalimentaire française spécialisée dans la production porcine. Depuis sa création en 1966, le groupe n'a cessé d'innover et de se diversifier pour maîtriser aujourd'hui tous les métiers du porc et toutes les étapes de production : de l'abattage jusqu'à la vente en passant par la découpe, l'élaboration et la salaison/charcuterie. La Cooperl est organisée en différentes branches, cette organisation est le fruit de l'histoire de la coopérative de par ses acquisitions industrielles et ses fusions. Cet agencement par filières porte une vision stratégique pour atteindre la maîtrise et l'excellence de chacun des métiers du porc.

Je réalise mon stage de Master 2 Data Science pour la biologie (Agrocampus Ouest) à la Cooperl sur le site de Plestan pendant 6 mois du 10 Février au 7 Août 2020.

Dans le cadre de ce stage je suis accueillie dans le service Recherche et Développement Transversal, les travaux conduits dans ce service visent une montée en gamme du groupe par une différenciation permanente. Les missions du service consistent à coordonner des projets transversaux (sur toute la filière), définir et prioriser la stratégie de la R&D au niveau du groupe mais aussi à développer des réseaux scientifiques.

La présence de la Cooperl sur plusieurs branches génère une quantité importante de données provenant de sources variées. Ces données s'étendent de la génétique à l'abattage en passant par l'alimentation, les produits vétérinaires et autres. Ces bases de données ne se réfèrent pas toujours au même individu statistique (un élevage, un animal par exemple). Celles-ci sont croisées par jointure pour créer une unique table de données "macrofilière" dans laquelle l'individu statistique est un élevage à un semestre. Cette table finale mettant en relation autant d'informations relatives à l'élevage est inédite à la Cooperl. Celle-ci va nous permettre de répondre à des questions plus précises ou plus globales sur les performances des élevages adhérents.

La performance technique d'un élevage porcin se mesure principalement par trois critères synthétiques : l'âge à 115 kilos (A115ks[1]), indicateur de la croissance moyenne des porcs dans l'élevage, l'Indice de Consommation Global (ICG[1]), indicateur de l'efficacité alimentaire au sein de l'élevage et la productivité annuelle (PRODAN[1]), indicateur de la capacité de l'élevage à produire et vendre des porcs.

Peut-on identifier différents profils de pratiques d'élevages à partir des profils de performances ? Quels types de pratiques sont majoritaires chez les adhérents ?

Le principal objectif du stage est d'établir une typologie des pratiques d'élevages d'un point de vue "macrofilière" à l'aide d'une analyse descriptive des données et notamment des liens pouvant exister entre variables. Nous classifions les élevages à l'aide de la combinaison des méthodes d'Analyse Factorielle Multiple (AFM) et de Classification Ascendante Hiérarchique (CAH). Ces analyses nous permettent également de repérer les incohérences liées à la construction de cette table que nous corrigeons par itérations au fur et à mesure des analyses. Le livrable du projet "macrofilière" sera donc une aide à la décision très précieuse sur le terrain.

[1] (Voir liste des sigles et des abréviations)

Nous nous pencherons tout d'abord sur la présentation de la coopérative, puis celle du stage avant de présenter la démarche statistique utilisée ainsi que les principaux résultats obtenus. Nous finirons par une discussion et une conclusion.

1. PRÉSENTATION DE LA COOPERL

1.1. Son histoire

Cooperl est un groupe coopératif agroalimentaire français spécialisé dans la production porcine. C'est l'initiative de 24 éleveurs du secteur de Lamballe menés par Sébastien Coupé, Président fondateur de la Cooperl, qui a fait naître cette coopérative en 1966 dont l'objectif premier était de pouvoir vivre de leur métier.

"Permettre à un maximum d'hommes et de femmes de vivre décemment dans leur région du fruit de leur travail." *Sébastien Coupé, Président fondateur de la Cooperl, 1969.*

Aujourd'hui, plus de 2700 éleveurs sont adhérents de la Cooperl et donc propriétaires de son capital social et l'objectif principal de la coopérative reste le même.

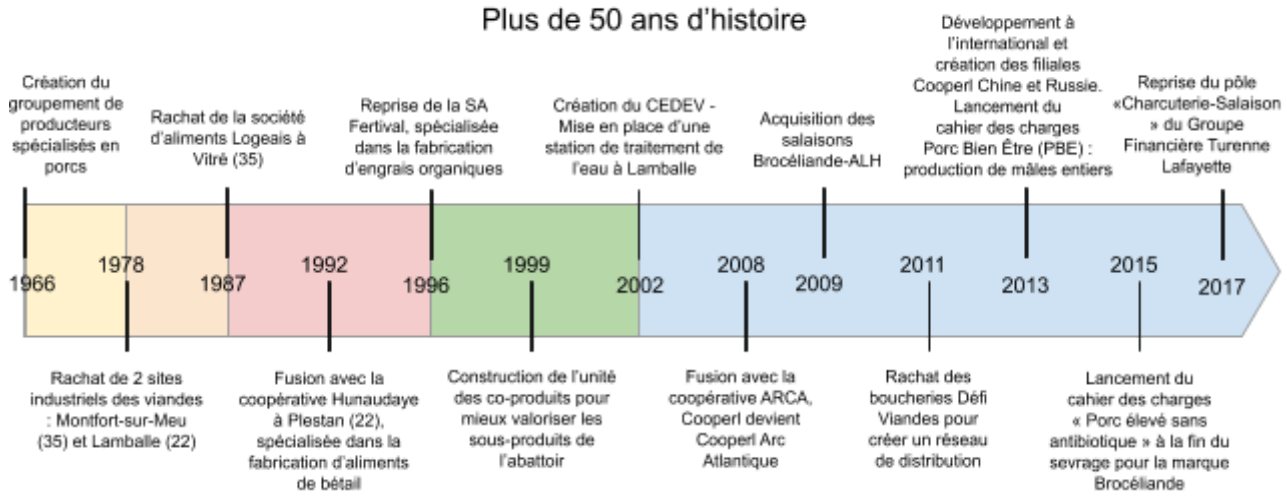


Figure 1. L'histoire de la Cooperl depuis sa création à nos jours

L'histoire de la Cooperl a fait qu'elle possède aujourd'hui un panel de marques avec des identités et des orientations différentes :

Figure 2. Les marques de la Cooperl

Brocéliande est la marque pionnière de la vente de "jambon bien élevé" respectant le strict cahier des charges du porc élevé sans antibiotiques dès la naissance. Madrange voue une véritable passion au goût en ne mettant dans ces produits que "l'essentiel". Montagne Noire incarne le plaisir de bien manger du Sud Ouest. Paul Prédault est une des marques de charcuterie les plus renommées du rayon coupe des grandes surfaces. VériTable est une jeune marque garantissant l'origine France et la traçabilité des produits.

1.2. Ses activités

Aujourd'hui la Cooperl est investie sur les métiers de :

- La nutrition et la santé animale

- La conception et les conduites d'élevages
- Le traitement des effluents et co-produits des élevages et des sites industriels
- La transformation agro-alimentaire de viandes et produits de charcuteries-salaisons
- La distribution spécialisée alimentaire (boucheries – charcuteries-traiteurs) et non alimentaire (matériel d'élevage)

Ces différents métiers sont organisés en 7 filières autour de la production porcine. La coopérative regroupe un ensemble d'activités complémentaires, de l'amont à l'aval avec plusieurs branches :

Figure 3. Les différentes filières de la Cooperl

Les objectifs principaux de la Cooperl sont alors d'améliorer les conditions d'élevages, le bien-être animal, la qualité des produits finis tout en garantissant une rémunération toujours plus juste pour les éleveurs-coopérateurs du groupe.

En 2020, Cooperl, ce sont ...

7000 salariés et 600 métiers
2700 adhérents
5 600 000 porcs produits

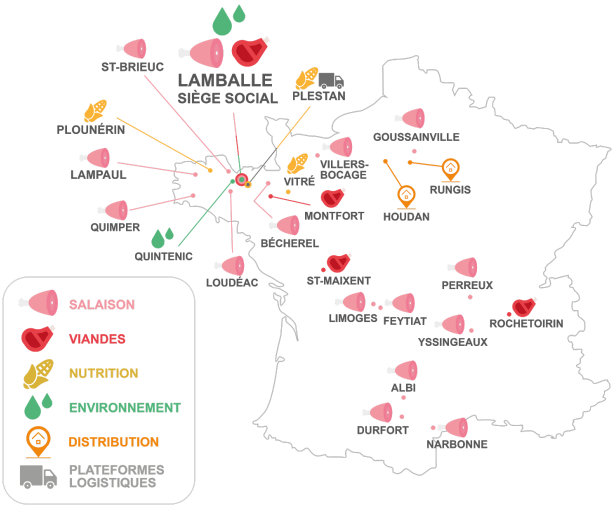
... mais aussi, en France et dans le monde :

N° 1 en France en production porcine
20 % de part d'activité abattage en France
13 millions de consommateurs dans le monde / jour

1.3. Ses implantations

Avec une implantation d'origine à l'Ouest, la Cooperl est aujourd'hui présente sur la majeure partie du territoire français mais également à l'international grâce aux exportations de ses produits et à ses sites de production (en Chine notamment).
Le groupe compte 9 zones de groupement de producteurs, 4 usines de viandes, 15 usines de salaisons, 3 usines d'aliment et 7 magasins Calipro.

Figure 4. Cartographie de la présence de la Cooperl en France



La Cooperl est présente sur une grande partie du territoire français grâce à ses acquisitions d'entreprises. Mais la Cooperl est aussi présente à l'international avec un site de production en Chine et plus de 165 000 tonnes de viande exportée dans le monde dont plus de 50 pays sont destinataires.

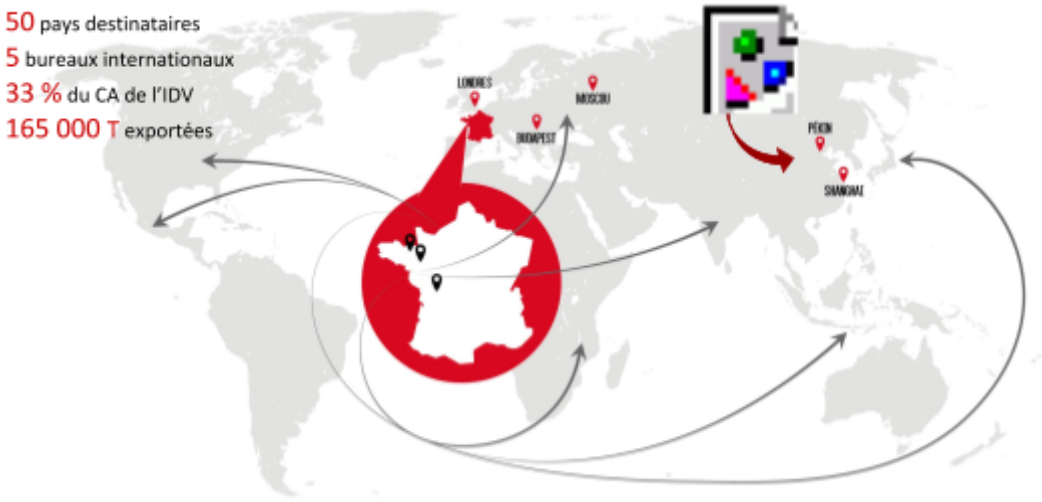


Figure 5. Cartographie des sites de la Cooperl à l'international

La stratégie générale du groupe est la segmentation du marché (porcs élevés sans antibiotiques, porcs nourris sans OGM, ...) et la montée en gamme. En effet, plus de 50 % de la production est garantie sans antibiotiques aujourd'hui, les élevages sont notés sur le bien-être animal. Il y a également un développement de filières spécifiques comme le Label Rouge, les porcs élevés sur paille, l'arrêt de la caudectomie, les cultures sans pesticides, les porcs nourris sans OGM ou encore biologiques.

2. PRÉSENTATION DU STAGE

2.1. Contexte

La Cooperl convient d'une présence suffisamment large sur toute la filière porcine pour pouvoir collecter des informations tout au long du cycle de vie du porc. C'est pourquoi la coopérative dispose d'une masse importante de données provenant de sources diverses et variées à partir de ses différentes branches. En effet, nous disposons de données relatives à l'abattage, aux équipements d'élevage, aux modes de fonctionnement de l'élevage, aux produits Cooperl achetés par l'élevage, aux données de gestion technico-économique du troupeau, aux quantités d'aliment achetées (si l'éleveur est client aliment à la Cooperl), aux produits vétérinaires achetés avec le cabinet vétérinaire partenaire de la Cooperl, aux produits de petit équipement achetés chez Farmapro, etc... Ces données provenant de multiples sources ont été croisées, pour la première fois, dans une unique table de données.

En collaboration avec un partenaire extérieur, Cooperl a mis en place le projet "macrofilière" visant à identifier des profils de pratiques d'élevages. Ce projet initié en avril 2019 débute actuellement sa phase 3. La phase 1 concerne la construction et le nettoyage de la table de données, la phase 2 concerne la modélisation des performances et la phase 3 consiste à identifier des profils de pratiques d'élevages. Ils ont ainsi développé un modèle pour chaque indicateur de performance (KPI) : pour l'âge à 115 kgs (A115KS), l'Indice de Consommation Globale (ICG) et la PRODuctivité ANnuelle (PRODAN). Cooperl et le partenaire extérieur ont préalablement fait une sélection des variables en analysant les corrélations, puis ils ont testé plusieurs modèles sur les variables sélectionnées (LASSO, combinaison de régression stepwise et algorithme génétique).

Ce projet a également permis de mettre en place un tableau de bord permettant de visualiser toutes les analyses statistiques réalisées dans chaque phase du projet. Cet outil regroupe

énormément d'informations que l'équipe n'a pas eu le temps d'analyser. Il n'est pas forcément orienté métier, l'objectif final étant d'apporter un support technique aux équipes terrain.

2.2. Problématique du stage

Les performances technico-économiques d'un élevage sont principalement expliquées par les pratiques de l'éleveur. Notre table de données "macrofilière" nous permet maintenant de retracer toutes les pratiques d'un élevage.

Cette table constitue une nouvelle manière de mesurer les performances des élevages adhérents en tenant compte des facteurs de toute la filière porcine.

Les adhérents peuvent avoir des pratiques diverses comme des pratiques qui se "ressemblent". Nous allons donc chercher à identifier des profils de pratiques d'élevages.

Tout d'abord, nous souhaitons repérer les incohérences liées à la construction de la table de données : comme des valeurs aberrantes. Ensuite, nous voulons identifier des profils de pratiques d'élevage à partir de profils de performance. Jusqu'alors les performances d'un élevage étaient caractérisées uniquement par les indicateurs bien connus du terrain (l'âge à 115kg, l'indice de consommation globale et la productivité annuelle). Cette table va nous permettre d'identifier plus largement les profils types de pratiques qui vont influencer les performances des élevages.

Aujourd'hui les techniciens n'ont pas d'outil simple et compréhensible pour justifier et illustrer leurs conseils auprès des éleveurs. Nous voudrions pouvoir quantifier et justifier leurs suggestions à l'aide d'un outil fiable et concret pour eux et pour appuyer leur argumentaire auprès des éleveurs. Nous disposons d'une grande table de données unique retraçant toutes les informations relatives à la filière porcine de laquelle nous pouvons ressortir une grande quantité d'informations, c'est pourquoi il est essentiel de trouver des pistes pour l'exploiter.

Finalement, ce projet vise à maximiser les performances des élevages des adhérents. Mais également à enrichir les connaissances du service R&D Transversal.

Notre objectif principal est donc d'établir une typologie des pratiques des élevages adhérents, c'est à dire identifier des groupes d'élevages qui se ressemblent. Quels sont les facteurs qui déterminent la mise en groupes des élevages et qui peuvent donc être considérés comme très importants pour expliquer les écarts de performances entre élevages.

Un deuxième objectif du stage, induit du précédent, est d'identifier les liens entre les facteurs des indicateurs de performance afin d'identifier les incohérences dans les données.

Nous intégrons des "variables de références" dans l'analyse, nous nous en servons comme référence car ce sont des variables très utilisées par ailleurs et que l'on connaît très bien.

La table de données "macrofilière" est constituée des 11 bases de données ci-dessous :

- ABAT : Base de données abattoir : quantité + qualité de la carcasse
- GTE : Base de données Gestion Technico-économique de l'Elevage : efficacité alimentaire, productivité
- GTTT : Base de données Gestion Technico Économique du Troupeau de Truies
- ODOR : Mâle entier odorant (détection des carcasses odorantes à l'abattoir)
- ALEA : indicateur synthétique de l'utilisation d'antibiotiques ($\frac{Qté\ Kg\ traités}{Qté\ Kg\ total\ de\ viande}$)
- CNP : Contrôle nez et poumons (ces contrôles sont un outil de diagnostic des pathologies respiratoires en élevage)
- PL : Produits Libres / petit matériels (aiguilles, gants, tatouage, ...)
- FP : Farmapro : Biosécurité (désinfectant / dégraissant), dératisation
- SVH : Produits vétérinaires : antibiotiques, vaccin, antiparasitaires

- REPRO : Vente de reproducteurs (verrats, cochettes) (hors éleveurs clients à l'extérieur et éleveurs en autorenouvellement[2])
- ALIM : Base de données aliment + supplémentations (antibiotiques, antiparasitaires, nutritionnelles (antioxydants, vitamines))

Ces données ont été nettoyées et croisées au sein d'une table de données "macrofilière", au départ elles n'avaient pas la même granularité selon les bases de données d'origine (certaines avaient pour individu statistique la truie, d'autres l'élevage, ...) car elles proviennent de sources différentes (élevages, abattoirs, vétérinaires,Farmapro, ...).

Nous disposons donc d'un tableau de données composé de 6104 individus et 802 variables. Un individu est un élevage à un semestre d'une année donnée.
Les données sont relatives à la période du premier semestre de 2015 jusqu'au premier semestre de 2019. Elles concernent 739 élevages, 9 semestres, 11 bases de données et 802 variables.
Pour les variables quantitatives, une moyenne par semestre a donc été estimée pour chaque individu. Pour les variables qualitatives, la modalité la plus fréquente a été conservée.

Afin d'identifier les différents profils de pratiques, nous analysons tout d'abord les liens entre les variables significatives des KPI[3] (âge à 115 kg, indice de consommation globale, productivité annuelle), mais aussi entre celles-ci et les variables de références. Puis nous classifions les élevages à l'aide de ces mêmes variables en réalisant une AFM puis une CAH.

3. DÉMARCHE STATISTIQUE

Les étapes suivantes ont été réalisées pour tous les facteurs significatifs des trois KPI[3] en y ajoutant les variables de références, variables que les experts métiers ont l'habitude d'analyser, celles-ci représentent donc une aide à l'interprétation. Une analyse globale a aussi été réalisée avec tous les facteurs significatifs des KPI[3] ainsi que toutes les variables de références.

3.1. Validation des données

La majeure partie du projet vise à expliquer les performances des élevages. En se faisant, nous nous apercevons que plusieurs incohérences y sont présentes.
Du fait de la taille conséquente de la table de données, de sa construction par jointure et du fait que certaines variables habituellement non utilisées aient été recalculées avec des problématiques de cohérence de données dans les données sources, plusieurs incohérences ont pu s'y infiltrer. Donc une partie conséquente du stage consiste à s'assurer que les données sont cohérentes. Pour caractériser les erreurs, la vision d'un expert métier est nécessaire.
Afin de repérer les incohérences dans la table de données nous nous penchons tout d'abord sur l'analyse des liens entre facteurs, mais aussi entre des variables considérées comme de références par les experts métiers pour chaque KPI[3].
Des incohérences peuvent encore subsister puisqu'au vu du nombre de variables dans la table de données, et de la présence de certains indicateurs synthétiques il est difficile d'analyser en totalité le tableau de données.

[2] Autorenouvellement : Production de cochettes avec les truies présentes dans le troupeau de l'élevage
[3] KPI : Key Performance Indicator (voir liste des sigles et abréviations)

En effet, les données manquantes représentent une part importante des données. La figure n°6 représente le nombre de données manquantes dans les variables influençant l'indicateur "âge à 115 kg". Les données manquantes représentées en rouge sur ces graphiques prennent une part importante de la totalité des données. La combinaison la plus fréquente est celle pour laquelle les variables prix des porcelets vendus à 25 kg (gte_SUP_pxptv), note moyenne pleuresie (cnp_MOYNOTEPLEUR) et part de mauvais poumons au contrôle (cnp_PERCPOUMAUV) sont des données manquantes. La variable comportant le plus de données manquantes est gte_SUP_pxptv, il s'agit d'un prix moyen de porcelets vendus à 25 kgs. Peu d'éleveurs vendent des porcelets à cet âge là aujourd'hui, c'est pour cette raison que le critère est peu renseigné.
Les données "abattoir" sont celles qui ont le moins de données manquantes, ce qui est logique puisqu'un éleveur adhérent a l'obligation de vendre ses animaux à Cooperl, c'est une des conditions de l'adhésion. Les bases GTE, GTTT, CNP ont plus de valeurs manquantes. Ceci est dû au fait que tous les éleveurs ne remplissent pas la GTE et la GTTT (suivi du troupeau) et pour la base CNP ceci est dû au fait que les contrôles nez poumons ont lieu minimum une fois par an pour les éleveurs en cahier des charges PSA[4] et pour les autres éleveurs, ce contrôle est réalisé uniquement en cas de problème sanitaire.
La combinaison pour laquelle toutes les données sont renseignées, ligne où toutes les cases sont bleues, concerne seulement 140 individus soit 2.29 % des individus.

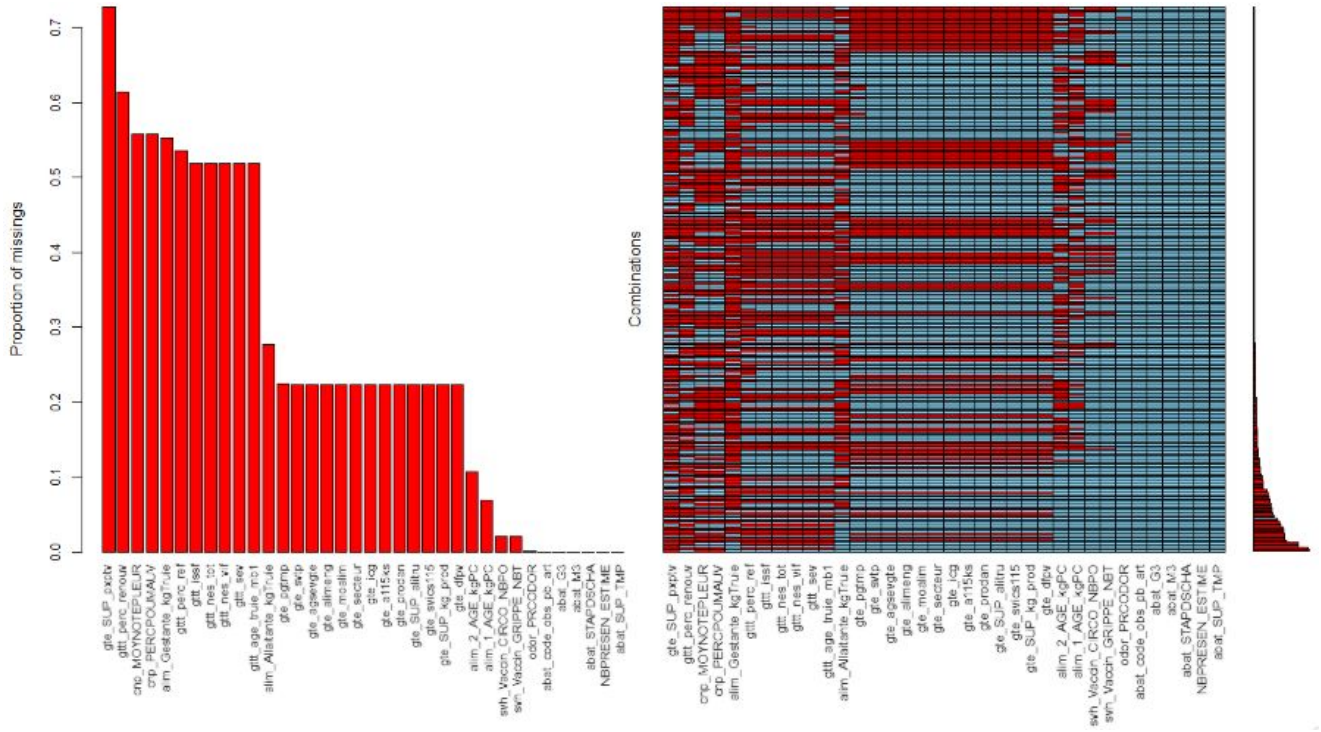


Figure 6. Effectifs des données manquantes par variables significatives de l'âge à 115 kg et par variables de références

Les valeurs manquantes n'ont pas la même signification selon la base de données d'origine dans laquelle elles se trouvent. Par exemple, une valeur manquante dans la base GTE[5] signifie que l'éleveur a déjà rempli la GTE[4] dans le passé mais qu'il ne la remplit pas à la période où il y a une donnée manquante. Une donnée manquante dans la base ALIM[6] signifie que l'élevage ne commande pas l'aliment à la Cooperl au semestre concerné. Une donnée manquante dans la base FP[7] signifie également que l'élevage n'est pas client au semestre

[4] PSA : Porcs élevés Sans Antibiotiques (voir liste des sigles et abréviations)
[5] GTE : Gestion Technico-économique de l'Elevage (voir liste des sigles et abréviations)
[6] ALIM : Base de données aliment (voir liste des sigles et abréviations)
[7] FP : Base de données Farmapro (voir liste des sigles et abréviations)

donné pour le produit concerné. Nous verrons par la suite que ces données manquantes influencent énormément les résultats de nos analyses.

Nous construisons de nouvelles variables permettant de prendre en compte le nombre de données manquantes de chaque base de données d'origine dans l'analyse. Nous créons ainsi l'indicateur suivant :

$$pourcentage\ de\ NAs\ dans\ la\ BDD\ = \frac{nombre\ de\ \ NAs\ dans\ la\ BDD}{nombre\ total\ de\ variables\ dans\ la\ BDD} \times 100$$

, avec NAs qui représente les données manquantes dans la base de données d'origine.

Nous intégrons ces indicateurs dans la classification, cela nous permettra de voir l'impact du nombre de données manquantes de chaque base de données d'origine sur la construction des classes et surtout sur les performances des élevages.

Variable	Part moyenne de NAs (%)
Aliment	15 %
Abattoir	0 %
ALEA (utilisation d'antibiotiques)	27 %
Contrôle Nez Poumons	41 %
Farmapro	42 %
GTE	29 %
GTTT	55 %
Reproduction	37 %
Mâles entiers odorants	0 %
Produits libres	41 %
Service Vétérinaire Hyovet	5 %

Table 1. Répartition moyenne de données manquantes par période et par base de données d'origine

La base de données d'origine regroupant le plus de données manquantes est la GTTT[8] avec une moyenne de 55 % de données manquantes par semestre d'élevage. Cela nous semble logique étant donné que les éleveurs ne sont pas obligés de la remplir et que cette manipulation prend du temps donc la majeure partie d'entre eux ne le font pas.
Nous voyons également que les bases de données ABAT[9] et ODOR[10] sont celles qui comptent le moins de données manquantes. Cela nous semble également cohérent car les données de ces deux bases sont construites automatiquement à l'abattoir et servent de base au paiement des éleveurs : classement des carcasses à partir d'un appareil "Image Meater", les mesures des carcasses sont récoltées par l'association Uniporc chaque jour et pour chaque carcasse. Les données de la base ODOR[10] sont également récoltées automatiquement à l'abattoir lorsque des mâles entiers y sont abattus.

3.2. Liens entre variables

Tout d'abord, nous estimons les corrélations entre les variables significatives pour chaque indicateur de performance afin de mettre en lumière les liens potentiels entre celles-ci. Nous ajoutons également les variables de références (déterminées par les experts métiers) à l'analyse.
Ces analyses nous permettent de confirmer certaines hypothèses connues que nous faisions à l'avance d'un point de vue métier. En effet, les experts métiers voient sur le terrain quels types de pratiques sont souvent liées aux performances et nos analyses le confirment.

[8] GTTT : Gestion Technico économique du Troupeau de Truies (voir liste des sigles et abréviations)
[9] ABAT : Base de données abattoir (voir liste des sigles et abréviations)
[10] ODOR : Base de données mâles entiers odorants (voir liste des sigles et abréviations)

Cela nous permet également de mettre en lumière des liens entre variables jusqu'alors méconnus mais fortement supposés par les experts métier.

Les liens entre variables quantitatives sont calculés à l'aide de la matrice des corrélations de Pearson en omettant les données manquantes car cette fonction ne peut les prendre en compte, mais nous y ajoutons les variables de pourcentage de données manquantes. Cette matrice de corrélation est ensuite testée avec le test de Pearson, celui-ci mesure la dépendance linéaire entre deux variables et permet d'évaluer le niveau de significativité des corrélations. En effet, seules les corrélations significatives peuvent être interprétées.
Les hypothèses du test des corrélations sont les suivantes :

H0 : pas de corrélation entre les variables ($\rho = 0$)

H1 : il existe une corrélation entre les variables ($\rho \neq 0$)

Sa statistique de test sous H0 est :

$$t = \frac{|r|}{\sqrt{\frac{(1-r^2)}{n-2}}}$$

, avec r le coefficient de corrélation entre deux variables et n le nombre d'observations.

Nous analysons les liens entre variables quantitatives et qualitatives à l'aide d'une Anova avec un test ad-hoc de Tukey, ce test fait une comparaison multiple des moyennes. En effet, pour chaque variable qualitative les moyennes des variables quantitatives sont testées deux à deux avec ce test et celui-ci attribue une lettre différente aux variables dont les moyennes sont testées significativement différentes (méthode décrite plus précisément dans la partie 3.3.2).

Pour analyser les liens entre variables qualitatives nous réalisons une comparaison des effectifs à l'aide du test du Chi2. Ce test se base sur un tableau de contingence. Le principe de ce test est de calculer l'écart entre la distribution observée et une distribution théorique que l'on aurait si les deux variables étaient indépendantes.

Ces analyses sont toujours suivies par une analyse critique d'un expert métier sur la cohérence de ces liens ou non.

3.3. Classification des élevages

3.3.1. Analyse Factorielle Multiple (AFM)

L'Analyse Factorielle Multiple (AFM) permet de visualiser un nuage de points dans l'espace. Un ensemble d'individus est décrit par des variables quantitatives et qualitatives, celles-ci étant structurées par groupes.

L'AFM va donc chercher à résumer et visualiser un tableau de données complexe. Un même groupe contient plusieurs variables de même type. Selon ce type, une ACP ou une ACM est réalisée sur le groupe de variables puis pondérée par la première valeur propre des différents tableaux. En effet, cette méthode attribue à chaque variable du groupe un poids égal à l'inverse de la première valeur propre de l'analyse effectuée sur ce groupe. Cela permet que les groupes ayant beaucoup de variables ne pèsent pas plus dans l'analyse.

Les groupes de variables qualitatives sont transformés en tableaux disjonctifs complets puis une Analyse des Correspondances Multiples (ACM) est réalisée sur ces tableaux. Une Analyse en Composantes Principales (ACP) est réalisée sur les groupes de variables quantitatives.

L'AFM convient ici car nous avons une diversité importante des données, celles-ci sont tout de même structurées en groupes (un groupe est composé des variables d'une même source d'information, ici d'une même base de données d'origine).

Nous avons des variables quantitatives et des variables qualitatives. Ces variables viennent de sources différentes, c'est pourquoi nous centrons et réduisons (standardisons) les données quantitatives. L'AFM permet de faire un "prétraitement" avant de faire la classification, en effet cette méthode va permettre d'**équilibrer l'influence des groupes** de variables. Cette technique permet également de supprimer l'information contenue sur les dernières dimensions, laquelle peut être considérée comme du bruit, afin de rendre la classification plus robuste.

Le fait de combiner les deux méthodes (AFM + CAH) permet d'avoir une visualisation plus complète : graphique des variables, des groupes, des individus (dans notre cas, nous n'allons pas visualiser les individus car ils sont trop nombreux).
Nous ajoutons nos trois indicateurs de performance techniques (A115KS[11], ICG[12], PRODAN[13]) dans un même groupe illustratif dans l'AFM. En effet, nous ne voulons pas que leurs valeurs soient prises en compte dans l'analyse car ce sont ces indicateurs que nous cherchons à analyser.

Nous testons plusieurs critères permettant de choisir au mieux le nombre de dimensions à interpréter dans l'AFM:

- Le critère du coude, cette méthode se base sur une appréciation visuelle de l'utilisateur. En effet, cette méthode consiste à identifier un point d'inflexion (une décroissance importante de l'inertie contenue dans les dimensions). Nous pouvons dire qu'il est purement subjectif et propre à l'oeil de l'utilisateur.

- Le critère de Cattell (G. Saporta), il s'agit de la version analytique du critère du coude. Ce critère consiste à identifier un changement de signe dans les différences secondes. Nous faisons la différence entre les valeurs propres successives, nous appellerons ces différences δ_i ici. Puis nous refaisons la différence entre ces deltas, ainsi nous obtenons les différences secondes entre les valeurs propres des dimensions, nommées Δ_i .

$$\delta_i = \lambda_i - \lambda_{i+1}$$
$$\Delta_i = \delta_i - \delta_{i+1}$$

  , avec i la variable correspondante.

- Le critère de Kaiser, celui-ci consiste à conserver uniquement les axes dont l'inertie est supérieure à l'inertie moyenne (formule ci-dessous). Comme nous nous situons dans le cas centré réduit, nous retiendrons les axes associés à une valeur propre supérieure à 1.

$$nb\ dim = \frac{I}{p}$$

  , avec I l'inertie totale et p le nombre de variables dans l'AFM.

Finalement, dans la pratique nous gardons surtout le nombre d'axes que l'on sait interpréter.

Pour construire une classification à partir des résultats de l'AFM, nous conservons les dimensions de l'AFM qui regroupent entre 75 et 80 % de l'inertie totale du nuage de points.

[11] A115KS : Age à 115 kgs (voir liste des sigles et abréviations)
[12] ICG : Indice de Consommation Globale (voir liste des sigles et abréviations)
[13] PRODAN : PRODuctivité ANnuelle (voir liste des sigles et abréviations)

Ceci va nous permettre de nous affranchir de l'information contenue dans les dernières dimensions de l'AFM que nous considérons comme du "bruit".

3.3.2. Classification Ascendante Hiérarchique (CAH)

La Classification Ascendante Hiérarchique (CAH) est une méthode de classification itérative permettant de partitionner la population en différentes "classes". L'objectif étant que les individus d'un même groupe soient les plus "proches" possible (variabilité intra-classe faible) et que les différents groupes soient les plus "éloignés" possible (variabilité inter-classe maximale). Cette méthode se base donc sur une matrice de distances. Les distances étant calculées entre chaque individu pris deux à deux. Plus les observations de deux individus sont équivalentes plus la distance entre ces deux individus est faible. Inversement, plus les observations entre deux individus sont dissemblables et plus ces individus sont éloignés.

Le choix de la distance dépend des données utilisées, ici nous avons des coordonnées des individus sur les composantes principales de l'AFM, donc nous choisissons la distance euclidienne pour construire la matrice des distances, cette méthode de calcul permet de représenter au plus juste la distance "réelle" entre individus.

Nous choisissons la méthode d'agrégation de Ward fondée sur l'inertie donc classiquement utilisée dans le cadre d'un enchaînement analyse factorielle - CAH. Cette méthode consiste à rassembler les individus qui font le moins varier l'inertie intra-classe.

La méthode HCPC permet également de combiner les méthodes factorielles et la CAH avec un partitionnement en k-moyennes (algorithme des K-Means). Ce partitionnement consiste en une ré-affectation individuelle au centre de classe le plus proche après découpage en classes.
Nous testerons également cette méthode avec une consolidation des classes par K-means afin de comparer la robustesse des deux types de classifications.
La CAH est un processus itératif :

- Première étape : chaque individu représente une classe
- A chaque étape l'algorithme agrège les deux classes les plus proches (ici n=6104 individus), donc en tout l'algorithme fait n-1 itérations (6104-1 = 6103 itérations) jusqu'à obtenir une seule classe. Cette série d'itérations est représentée par un arbre que l'on nomme le dendrogramme. Celui-ci nous permet de choisir le nombre de classes a posteriori.

Nous effectuons ainsi une classification hiérarchique sur les coordonnées des individus sur les composantes principales de l'AFM. Cette méthode identifie plusieurs groupes d'élevages. La CAH combinée à l'AFM permet donc d'obtenir une classification plus robuste puisque nous enlevons les dernières dimensions de l'AFM pouvant être considérées comme du "bruit".
Pour ce faire nous utilisons la fonction HCPC() du package FactoMineR. Cette fonction permet de recevoir les résultats d'une analyse factorielle et d'y réaliser une CAH.

Une fois le dendrogramme tracé, nous choisissons le nombre de classes à créer selon plusieurs critères :

- Il est possible d'utiliser les mêmes critères que pour l'AFM mais c'est surtout visuel : on coupe l'arbre quand la perte d'inertie devient trop élevée.
- La fonction HCPC() du package FactoMineR préconise une partition en un nombre de classes ayant la plus grande perte d'inertie relative entre classes consécutives. En effet, le nombre de classes proposé est calculé en regardant le saut relatif minimum entre 2 classes consécutives.

En général nous sélectionnons le nombre de classes suggéré par l'algorithme, tout en gardant un oeil sur le graphique des pertes d'inertie entre classes.

Afin de décrire les groupes selon les variables quantitatives, nous réalisons une Anova avec un test de Tukey (test de comparaisons multiples). Cette technique nous permet de comparer les moyennes deux à deux en attribuant une lettre identique à chaque couple de moyennes significativement égales.
Le test de Tukey envisage les moyennes des échantillons (ici des classes) deux par deux et calcule la statistique de test suivante :

$$Q_{ij} = \frac{\left| M_{k_i} - M_{k_j} \right|}{\sqrt{M_{intra}/n}}$$

, où M_{intra} est la moyenne des inerties intra-groupes (donc le quotient de l'inertie intra par le nombre de degrés de liberté intra) et n est le nombre d'observations dans chaque échantillon.

Pour décrire les classes selon les variables qualitatives, nous visualisons les effectifs de celles-ci à l'aide d'un histogramme empilé pour chaque classe. Les sorties de la CAH sur R permettent une caractérisation très précise des classes, par ses variables, ses modalités, ses individus ou encore par les axes factoriels.

Nous nous intéressons aux mouvements des élevages, quelle est la classe qui "reçoit" le plus d'élevages? Quelle est la classe qui "perd" le plus d'élevages? Ceci va traduire les évolutions de pratiques en élevages.

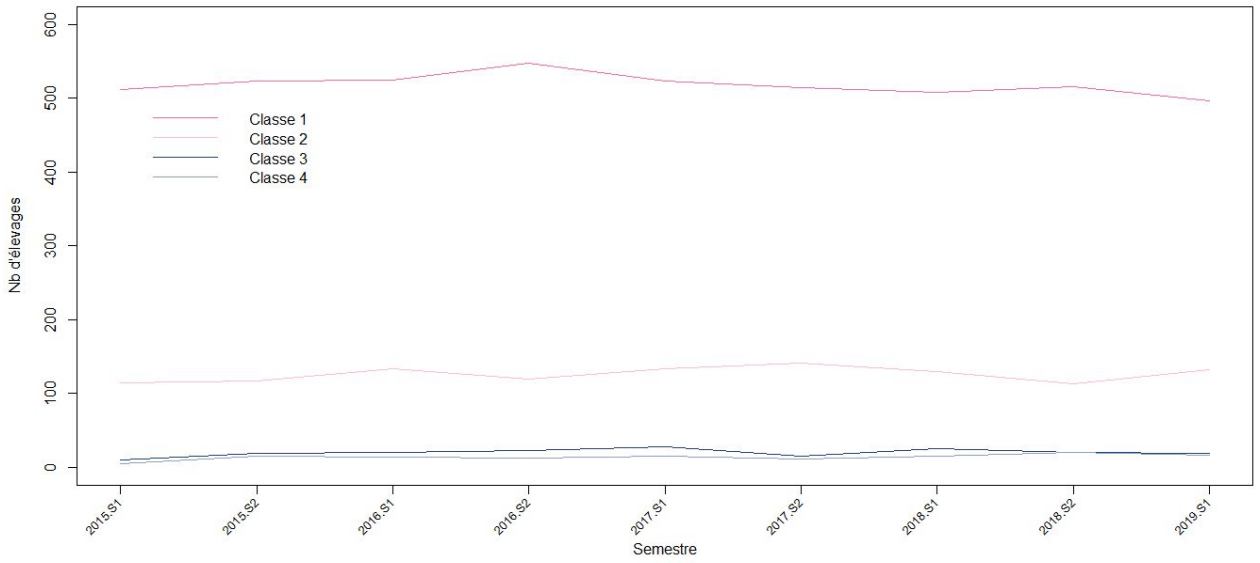


Figure 7. Evolution du nombre d'élevages par classe

Nous construisons ensuite la matrice de transition moyenne (Figure n°13) afin d'avoir une visualisation plus précise des flux entre classes selon la période. Pour ce faire, nous calculons tout d'abord les matrices de transitions entre chaque période (9 semestres au total), une matrice de transition correspond au nombre d'élevages qui ont changé de classe entre une période et la suivante. Nous moyennons toutes les matrices de transitions et nous utilisons cette matrice de transitions pour faire le schéma des flux avec une épaisseur du flux proportionnelle à son importance qui correspond à la probabilité conditionnelle d'être dans la classe j à la période 2 sachant que l'élevage était dans la classe i à la période 1.

Afin d'identifier différents profils de pratiques d'élevages, nous aurions pu envisager la méthode des k-means, cependant celle-ci ne peut prendre en compte que des variables quantitatives. Donc nous ne l'appliquons pas ici car nous disposons de variables quantitatives

et qualitatives, mais nous testons la méthode HCPC avec une consolidation par k-means (partie 4.3).

La méthode de la CAH admet certaines limites, en effet les résultats peuvent différer selon la paramétrisation : critère de distance, choix d'agrégation. Cette méthode peut également être lourde en calcul dès lors que l'on a un nombre de données important.

4. PRINCIPAUX RÉSULTATS

4.1. Analyse des liens

Nous prenons ici l'analyse des facteurs significatifs de la productivité annuelle (PRODAN) pour exemple.
Nous analysons tout d'abord les liens entre variables quantitatives du modèle. Ci-dessous nous avons les corrélations entre l'indicateur de performance (PRODAN) et quelques uns de ses facteurs.

	Corrélation avec PRODAN
Intervalle entre sevrage saillie fécondante	-0.041
Taux de réforme	-0.099
Qté moyenne d'aliment par truie et par an	0.180
Âge à 115 kgs	-0.151
Indice de Consommation Globale (ICG)	-0.472
Taux de conversion des aliments standardisé	-0.263
Qté de viande produite par truie et par an	0.886
Taux de mortalité entre 8 et 115 kgs	-0.333

Table 2. Extrait des corrélations significatives entre quelques variables quantitatives et la PRODAN

Plus la productivité annuelle est élevée, plus la quantité de viande produite par truie et par an est élevée . Tous ces liens sont logiques et ne nous surprennent pas.
Les autres liens entre variables quantitatives sont très nombreux, c'est pourquoi nous sélectionnons les liens dont les corrélations sont significatives et supérieures à 0.5.

Variable 1	Variable 2	Corrélation
Qté de viande produite / truie / an	Productivité annuelle (porcelets/truie/an)	0.925
Indice de Consommation Globale (kg/kg)	Indice de Consommation 8-115 kg	0.901
Nombre de nés totaux / portée	Nombre de nés vifs / portée	0.876
Nombre de sevrés / portée	Productivité annuelle (porcelets/truie/an)	0.762
Nombre de sevrés / portée	Qté de viande produite / truie / an	0.694
Nombre de nés vifs / portée	Nombre de sevrés / portée	0.671
Âge à 115 kg (jours)	Indice de Consommation 8-115 kg	0.612
Âge à 115 kg (jours)	Indice de Consommation Globale (kg/kg)	0.584
Nombre de nés totaux / portée	Nombre de sevrés / portée	0.565
Part de NAs dans la base GTTT (%)	Taux de réforme (%)	0.519
Nombre de nés vifs / portée	Productivité annuelle (PRODAN)	0.504
Poids moyen de carcasse (kg)	Taux de muscle au point P3 (%)	0.501

Table 3. Corrélations supérieures à 0.5 entre variables significatives de la PRODAN

Nous observons un lien positif entre la part de données manquantes dans la base GTTT et le taux de réforme, plus on a de données manquantes dans la GTTT plus le taux de réforme est élevé. Les éleveurs qui font la GTTT ont un meilleur regard sur l'élevage, ils vont plutôt choisir leurs truies à réformer en faisant en sorte de réduire le taux de réforme, et auront donc moins de réformes "subies". Ce lien traduit donc un comportement de l'éleveur.

L'analyse des liens entre les variables quantitatives et les variables qualitatives sont également visualisées dans un tableau présentant les résultats de l'Anova avec un test ad HOC de Tukey. Par exemple, nous voyons ci-dessous (Table 4) que la productivité annuelle n'est pas significativement différente selon le type d'aliment truie. Cependant, nous pouvons dire que les deux autres KPI (âge à 115 kg et indice de consommation globale) sont meilleurs avec l'aliment COMPLET qu'avec l'aliment FAF[14]. Ceci peut être dû au fait que ce lien cache autre chose. Notamment le fait que l'on approche la performance de l'élevage. Si un élevage fait son aliment lui-même, il ne va pas contrôler de manière quotidienne les valeurs nutritionnelles de ses matières premières comme c'est fait à l'usine d'aliment. En effet l'aliment COMPLET (fabriqué à l'usine) va avoir des valeurs nutritionnelles beaucoup plus précises, qui colleront donc mieux aux besoins des animaux.

Variable	COMPLET	FAF	Moyenne globale	Signif
Âge à 115 kg (A115KS) (jours)	177	180	177	***
Indice de Consommation Globale (ICG) (kg/kg)	2.73	2.78	2.74	***
Productivité annuelle (PRODAN) (porcelets/truie/an)	23.1	22.9	23.1	ns

Table 4. Comparaison des moyennes selon le type d'aliment pour truie

En ce qui concerne les liens entre variables qualitatives, nous avons ci-dessous l'histogramme des effectifs croisés entre les modalités des variables type d'aliment engraissement et la certification ainsi que le tableau des effectifs croisés correspondant.

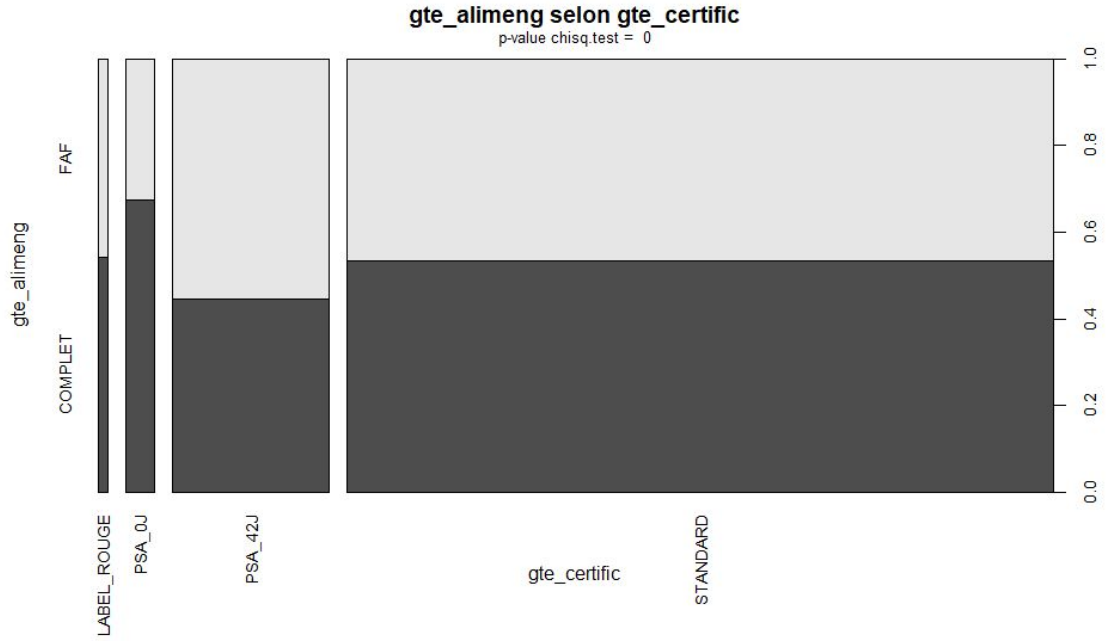


Figure 8. Répartition du type d'aliment engraissement selon le cahier des charges

[14] FAF : Fabrication d'Aliment à la Ferme (voir liste des sigles et abréviations p.0)

	LABEL ROUGE	PSA 0J	PSA 42J	STANDARD	Total
COMPLET	54 %	68 %	45 %	54 %	52 %
FAF	46 %	32 %	55 %	46 %	48 %
Total	1 %	3 %	18 %	78 %	

Table 5. Effectifs des différents types d'aliment engraissement selon le cahier des charges

Nous voyons que 54 % des élevages standards sont en aliment Complet. 68 % des élevages PSA 0J sont en aliment Complet et seulement 32 % d'entre eux sont en FAF. Nous avons donc autant d'élevages en FAF qu'en Complet.

4.2. Analyse Factorielle Multiple (AFM)

Dans les deux parties suivantes nous allons présenter les analyses visant à classifier les élevages en tenant compte de tous les facteurs influençant les 3 KPI, les variables de références mais également les indicateurs constituant l'approche de la composante du statut sanitaire. Nous l'appelons AFM globale avec le statut sanitaire.
Le statut sanitaire d'un élevage n'est pas trivial à estimer puisqu'il dépend de beaucoup de critères. C'est pourquoi, à l'aide du partenaire extérieur lié à ce projet, Cooperl a construit un premier indicateur pour approcher au mieux la notion de statut sanitaire d'un élevage.
Ce premier indicateur prend en compte les 4 aspects suivants :

- La gestion de la santé dans l'élevage avec :
  - Performances zootechniques
  - Performances sanitaires
- La façon dont l'éleveur "maîtrise" son statut sanitaire :
  - par l'utilisation d'antibiotiques
  - par l'utilisation de vaccins

Un deuxième indicateur est créé pour préciser cette notion de "maîtrise" du statut sanitaire. Cette notion a été remplacée par la notion de "stabilité" du statut sanitaire. Est-ce que les élevages ayant de bonnes performances de santé et zootechniques utilisent beaucoup d'antibiotiques ou non? Inversement, est-ce que les élevages ayant plutôt un mauvais statut sanitaire utilisent beaucoup d'antibiotiques ou non?

Nous obtenons l'éboulis des valeurs propres ci-dessous (figure 9), à l'aide du critère du coude nous choisissons d'interpréter 3 dimensions, avec le critère de Cattell nous choisissons d'interpréter 2 dimensions et le critère de Kaiser nous dit d'interpréter 10 dimensions. Finalement, comme nous conservons les dimensions que l'on saura interpréter, ce sont donc les 3 premières dimensions que nous choisissons.

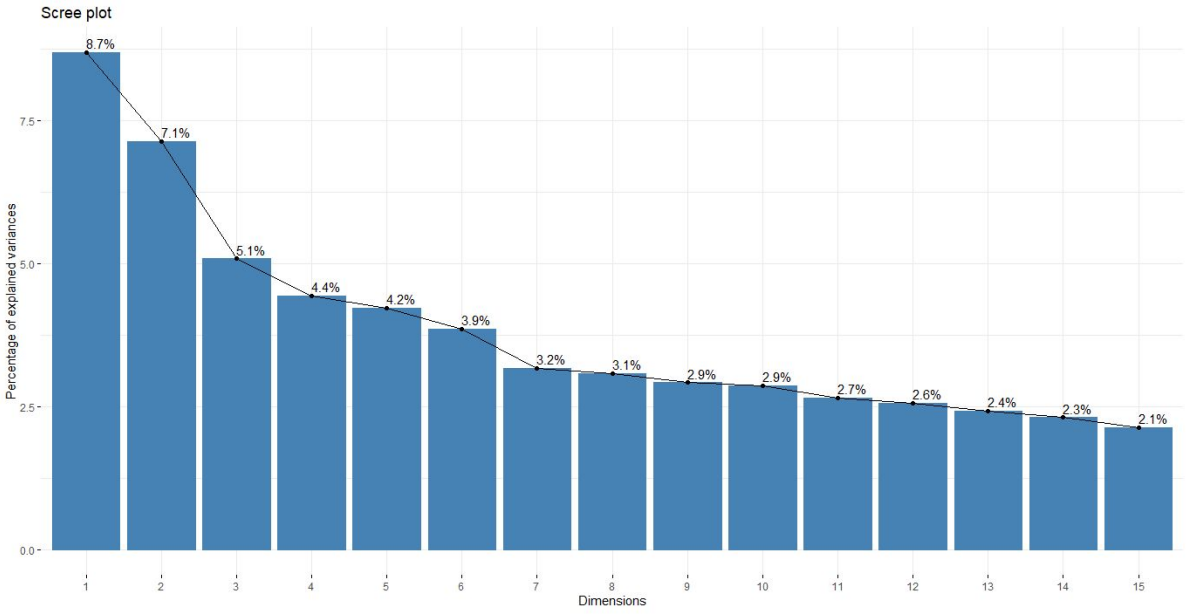


Figure 9. Eboulis des valeurs propres de l'AFM globale avec le statut sanitaire

Nous interprétons uniquement l'AFM sur le plan composé des dimensions 1 et 2 (Figure 10).

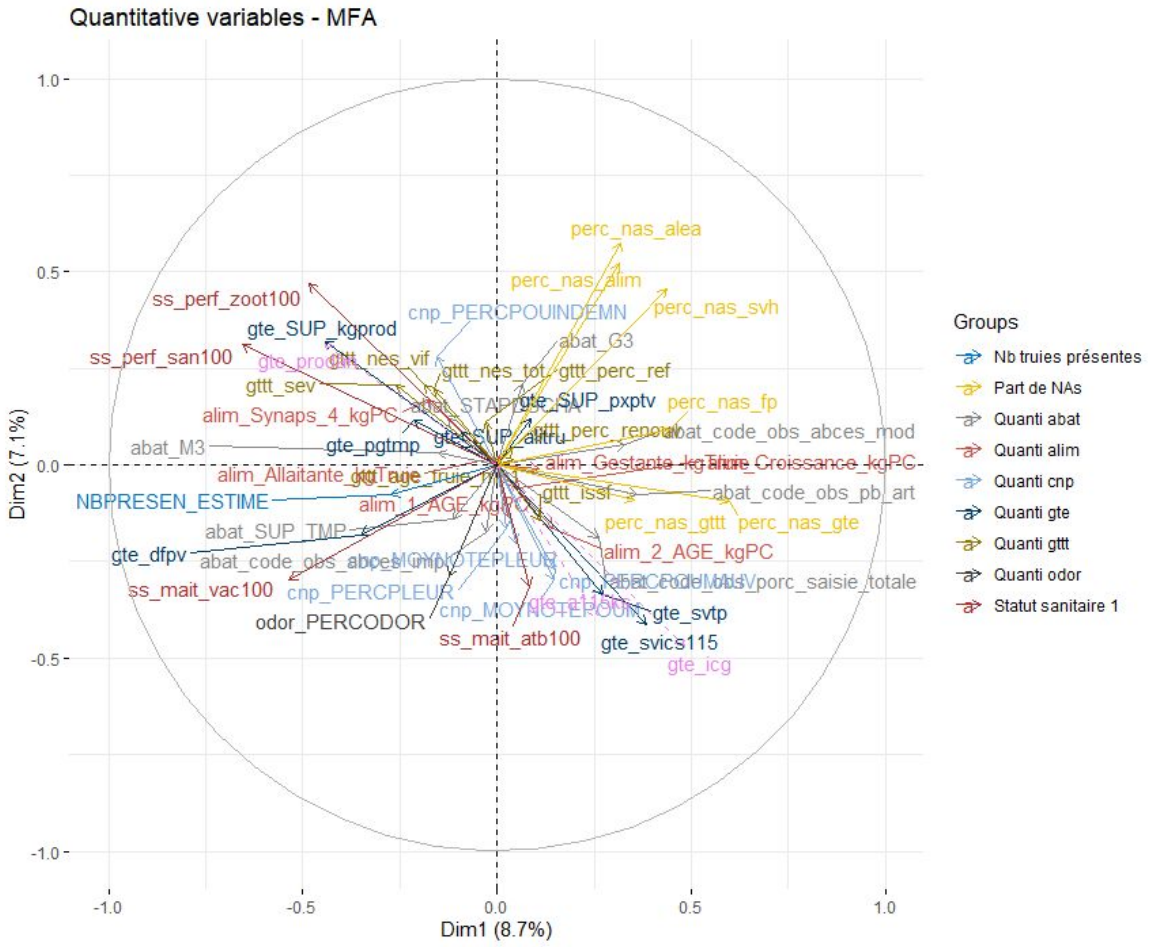


Figure 10. Cercle des corrélations de l'AFM pour la productivité annuelle (plan 1,2)

La dimension 1 oppose des élevages ayant beaucoup de données manquantes dans la GTE, FP et SVH aux élevages ayant de bonnes performances sanitaires et zootechniques avec une maîtrise de son statut sanitaire par vaccins. En effet, la dimension 1 est caractérisée par les variables part de données manquantes dans la GTE(perc_nas_gte), part de données

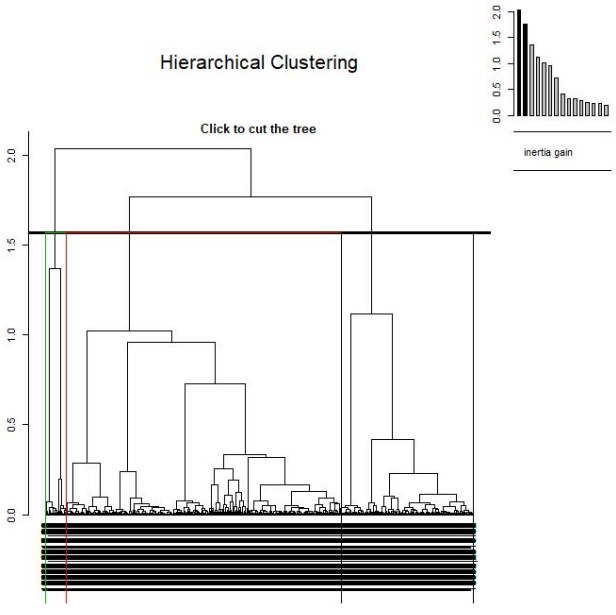
manquantes dans la base Farmapro (perc_nas_fp), la part de données manquantes dans la base de données référençant les ventes de produits vétérinaires (perc_nas_svh), performances sanitaires en base 100 (ss_perf_san100), maîtrise du statut sanitaire par vaccin (ss_mait_vac100), performances zootechniques (ss_perf_zoot100).
La dimension 2 oppose les élevages ayant un forte taux de conversion des aliments standardisé (gte_svics115 élevé) et beaucoup de mâles entiers odorants (odor_PERCODOR élevé) aux élevages utilisant peu d'antibiotiques (perc_nas_alea élevé), ayant de bonnes performances zootechniques (ss_perf_zoot100 élevé), non clients aliment et Hyovet (perc_nas_alim et perc_nas_svh élevés).

Nous voyons donc que les données manquantes et le statut sanitaire sont les variables qui orientent majoritairement l'analyse. En effet ce sont ces variables qui contribuent le plus à la construction des dimensions 1 et 2. Cela a donc une signification très forte dans nos analyses car cela signifie que la performance des élevages est considérablement liée au fait que les éleveurs soient clients ou non à la Cooperl (pour les produits vétérinaires, Farmapro), mais aussi que les éleveurs remplissent ou non la GTE/GTTT. Mais nous voyons également que les performances sont liées au statut sanitaire de l'élevage.

4.3. Classification des élevages

Nous utilisons ensuite les résultats de l'AFM à l'aide de la fonction *HCPC()* du package FactoMineR afin de faire une CAH sur les composantes principales de l'AFM sélectionnées. Le dendrogramme est calculé avec la méthode de Ward et la distance euclidienne. Ainsi, nous obtenons le dendrogramme ci-dessous (Figure 11).

Figure 11. Dendrogramme de la CAH globale



	Nb de semestres d'élevages	Part du nb de semestres d'élevages
Classe 1	1881	30.8 %
Classe 2	3921	64.2 %
Classe 3	302	4.9 %

Table 6. Effectifs des classes de la CAH

La fonction *HCPC()* nous suggère une partition en 4 classes, cette partition est celle qui minimise le saut relatif entre deux classes consécutives.
Afin de décrire les classes, nous réalisons une Anova des variables quantitatives significatives du modèle expliquant la productivité annuelle avec test ad-HOC de Tukey. Les résultats sont représentés dans le tableau suivant (Table 7).

	Variable	Classe 1	Classe 2	Classe 3	Moy globale	Signif
	Effectifs	**1881**	**3921**	**302**		
Exemples de variables	Poids moyen de carcasse (kg)	96.04b	95.57a	95.75ab	95.73	***
	Taux de muscle / pièce (%)	61.17c	61.12b	60.91a	61.12	***
	Qté aliment finition Synaps / porc (kg)	**18.73b**	5.59a	4.10a	9.76	***
	Score contrôle poumons	1.51b	1.98c	1.21a	1.75	***
	Qté de viande prod / truie / an (kg)	**2889.85c**	2639.42a	2728.96b	2742.26	***
Données manquantes	Part de NAs dans la base vétérinaire (SVH)	2.97a	3.16a	**43b**	05.07	***
	Part de NAs dans la base aliment (%)	11.17a	11.84b	**74.23c**	14.72	***
	Part de NAs dans la base GTE (%)	8.27a	**37.33b**	**42.98c**	28.66	***
	Part de NAs dans la base GTTT (%)	40.49a	**61.86b**	56.85b	55.03	***
	Part de NAs dans la base Farmapro (%)	36.35a	44.15b	**51.21c**	42.3	***
	Part de NAs dans la base ALEA (%)	21.23a	28.44b	**54.11c**	27.49	***
Indicateurs de productivité	Âge à 115 kgs (A115KS) (jours)	174a	**180b**	174a	178	***
	Indice de Consommation Globale (ICG) (kg/kg)	2.65a	**2.8c**	2.74b	2.74	***
	Productivité annuelle (PRODAN) (porcelets/truie/an)	**24.24c**	22.26a	22.88b	23.07	***
Composante de santé en élevage	Score de Performance sanitaire	**116.58b**	93.59a	96.02a	100.8	***
	Score de Performance zootechnique	**115.27c**	94.53a	103.42b	103.1	***
	Maîtrise du statut sanitaire par vaccin	**108.18c**	100.35b	74.32a	101.47	***
	Maîtrise du statut sanitaire par antibiotique	92.85b	96.39c	86.33a	94.91	***

Table 7. Extrait de l'Anova avec test de Tukey par classe sur les variables quantitatives de la CAH globale

Nous mettons ces résultats en relation avec les sorties de la fonction *HCPC()* liées aux variables quantitatives.

Pour chaque classe, nous effectuons une analyse des effectifs des variables qualitatives afin de voir quelles variables sont les plus caractéristiques de la classe. Nous avons l'exemple de la classe 1 ci-dessous (Figure 12).

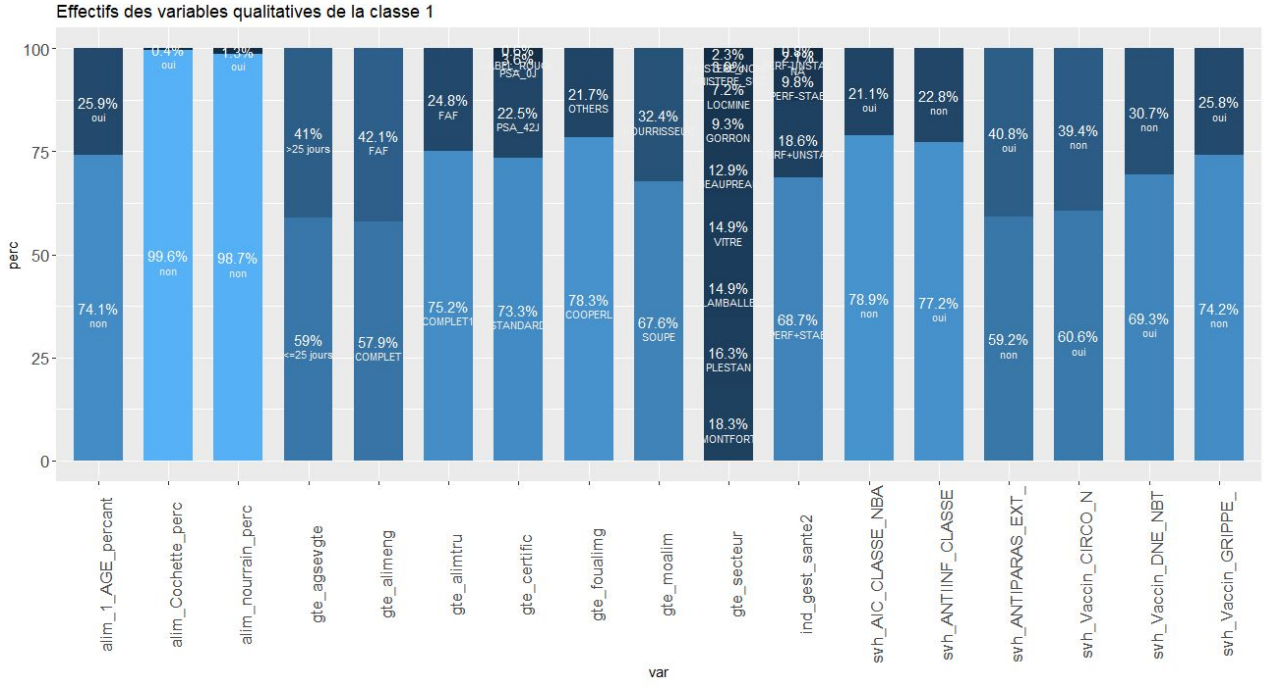


Figure 12. Effectifs de chaque variables qualitatives de la classe 1 de la CAH globale

La classe 1 regroupe des élevages productifs avec un bon statut sanitaire maîtrisé par vaccins. Ce sont des élevages plutôt en nourrisseur de type Soupe, en aliment pour truie et aliment d'engraissement Complet, clients aliment à la Cooperl. Ces élevages sèvrent les porcelets avant 25 jours, ils suivent le cahier des charges Standard et en PSA 42J, des secteurs de Lamballe et Plestan.
La classe 2 regroupe des élevages peu performants (ICG élevé, A115KS élevé, taux de perte élevé) avec un statut sanitaire plutôt mauvais (élevages ayant le statut sanitaire PERF-STAB ou PERF-UNSTAB).
La classe 3 regroupe des élevages non clients à la Cooperl pour l'aliment, non clients avec le cabinet vétérinaire partenaire de la Cooperl ni client Farmapro. Ils ont beaucoup d'abcès petits à modérés. Ils ne remplissent pas la GTE/GTTT. Ce sont donc des élevages qui ont peu d'activités avec la Cooperl.

Nous avons ci-dessous la matrice de transition moyenne liée à cette classification (Figure 13), cela nous permet de voir les changements de pratiques les plus fréquents.

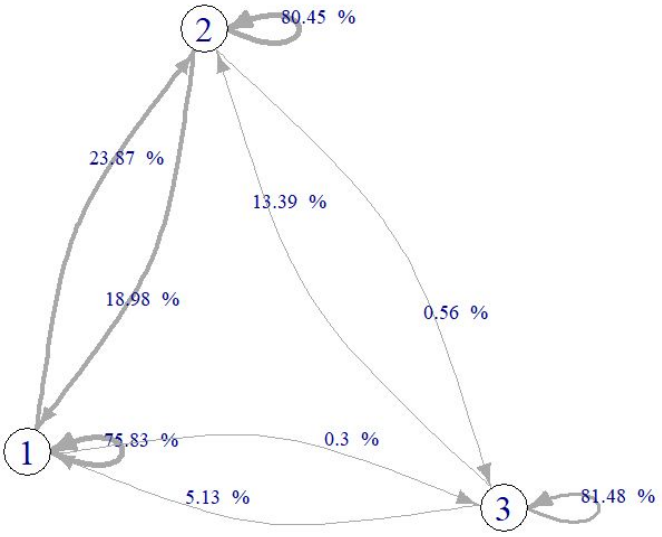


Figure 13. Stabilité des classes au cours des semestres

D'une manière générale, les élevages ont tendance à rester dans leur classe entre deux périodes consécutives.
La matrice de transitions moyenne entre classes permet d'identifier les classes les plus "réceptives" et les plus "réfractaires".
La classe 2 est la plus "réceptive", cette classe est aussi celle qui "transmet" le plus vers les autres classes. Cette dernière est constituée des élevages peu performants avec un mauvais statut sanitaire donc nous pouvons penser qu'elle est une "classe de passage" dans laquelle les élevages peuvent se retrouver pour quelques semestres. Ces résultats sont donc rassurants puisque cela signifie que les élevages peu performants le restent peu de temps.
Tandis que les classes 1 et 3 conservent en moyenne plus leurs élevages d'un semestre à l'autre. C'est-à-dire que les élevages performants (classe 1) le restent pour la plupart et les élevages ayant peu d'activités avec la Cooperl (classe 3) conservent également ce statut.

On ne constate aucun mouvement entre la classe 3 et la classe 1, cela signifie que les élevages ayant peu d'activités avec la Cooperl ne deviennent pas des élevages performants. Ce qui semble cohérent car les données manquantes présentes dans les élevages de la classe 3 traduisent un comportement des éleveurs, ceux-ci n'ont pas beaucoup d'activités à la Cooperl et ne comptent certainement pas en avoir plus.

Si nous réalisons la CAH avec une consolidation par k-means, nous obtenons les effectifs suivants :

	Nb de semestres d'élevages	Part de semestres d'élevages
Classe 1	2527	41.4 %
Classe 2	3273	53.6 %
Classe 3	304	5 %

Table 8. Effectifs des classes de la CAH avec consolidation par k-means

Après ré-affectation par k-means, la classe 1 possède 646 individus en plus, la classe 2 compte 648 individus en moins et la classe 3 compte 2 individus en plus. Nous obtenons également les mêmes variables décrivant les classes.

DISCUSSION

La construction de la base, qui s'est faite par jointure, a engendré plusieurs incohérences que l'on a pu relever au fil des analyses, l'analyse des liens notamment. La correction de ces erreurs n'a fait qu'augmenter le temps passé sur les analyses de données (analyse des liens mais également la classification). C'est pourquoi, nous avons dû faire des choix et nous avons préféré nous concentrer sur la mise en cohérence des données afin d'avoir des résultats interprétables, mais également sur une application rigoureuse des méthodes statistiques sur ces données. La mise en évidence de données manquantes nous a aussi contraints à faire un choix, nous avons créé des variables référençant la part de données manquantes par base de données d'origine car ces données ont une signification importante, elles traduisent un comportement humain et est donc à prendre en compte dans les analyses. Nous voyons finalement que ces variables orientent en grande partie nos analyses. L'ajout du statut sanitaire aux analyses a également permis de voir que celui-ci influence considérablement l'analyse.

L'interprétation des résultats nécessite une forte interaction avec les experts métier. C'est aussi pour cette raison que je n'étais pas en capacité de repérer les incohérences dans la table de données sans une connaissance pointue des nombreux indicateurs présents dans celle-ci.

L'objectif initial de ce stage consistait à rendre interprétable l'outil de visualisation créé par Cooperl et le partenaire extérieur, de mettre en place une segmentation permettant de décrire différents "groupes d'élevages" qui se ressemblent au niveau de leurs pratiques à partir de leurs performances. Pour ensuite, mettre en place un moyen d'évaluer les gains liés à d'éventuels changements de pratiques. Ceci constituant un appui pour les techniciens dans leur travail mais également une information importante pour la R&D. Finalement, la qualité des données a fait que le stage s'est orienté vers une démarche d'amélioration de la qualité de la donnée afin d'être en capacité d'en dégager des profils de pratiques d'élevages. Nous avons donc su faire preuve d'une nécessaire agilité dans ce projet comme il est nécessaire de le faire dans tout projet selon la réalité.

Nous obtenons finalement plusieurs profils de pratiques en élevages à partir des profils des facteurs influençant la performance de ces élevages. Si je devais aller plus loin dans l'analyse, je ferais des recherches sur d'autres méthodes de classification non supervisée applicables à un jeu de données composé de variables quantitatives et qualitatives et prendrais en compte la dimension temporelle du jeu de données. Puis analyser les gains liés aux changements de pratiques à partir des coefficients des modèles réalisés et des différents profils identifiés.

CONCLUSION

Ce stage s'est déroulé en plusieurs étapes, la première a consisté à se familiariser avec ce grand jeu de données, notamment en réalisant l'analyse des liens entre les variables significatives de chaque KPI et les variables de références. Une deuxième étape a consisté à dresser une typologie des élevages selon leurs pratiques et leurs performances afin de répondre à la problématique posée : peut-on identifier des profils de pratiques en élevages à partir des performances des élevages ?

Une étape préalable d'analyse des liens entre variables a permis de repérer les incohérences dans la base de données mais aussi d'identifier les corrélations entre variables significatives des KPI mais aussi entre ceux-ci et les variables de référence. Puis la correction de celles-ci nous a permis d'établir la classification des élevages et ainsi obtenir des profils de pratiques en élevages.
Les méthodes statistiques mises en place durant ce stage ont donc été principalement les critères de corrélations pour la première partie des analyses. Notamment à l'aide de la matrice de corrélation de Pearson pour les liens entre variables quantitatives, d'une Anova avec un test ad-hoc de Tukey pour les liens entre variables quantitatives et qualitatives et d'une comparaison des effectifs à l'aide du test du Chi2 pour les liens entre variables qualitatives. Pour identifier différents profils de pratiques en élevages nous réalisons une Analyse Factorielle Multiple (AFM) puis une Classification Ascendante Hiérarchique (CAH) sur les composantes principales de l'AFM. Ces deux méthodes combinées permettent d'équilibrer les groupes mais également de s'affranchir de l'information contenue dans les dernières dimensions de l'AFM pouvant être vue comme du "bruit", ce qui rend la classification plus robuste.

Nous obtenons ainsi une classification prenant en compte les facteurs significatifs des trois KPI (âge à 115 kilos, Indice de Consommation Globale et productivité annuelle), les variables de références, le statut sanitaire et la part de données manquantes dans chaque base de données d'origine.
Cette segmentation permet d'identifier trois principaux types de "pratiques en élevages". Nous avons un premier groupe qui correspond aux élevages productifs avec un bon statut sanitaire maîtrisé par vaccins. La classe 2 regroupe des élevages peu performants avec un statut sanitaire plutôt mauvais. La classe 3 regroupe des élevages ayant globalement peu d'activité avec la Cooperl.

Ce stage a permis de confirmer l'intérêt d'exploiter les données "macrofilière" d'une nouvelle manière, nous avons pu en extraire des profils de pratiques d'élevages à partir des profils de performances des élevages.

Pour ma part, ce stage a été une expérience très enrichissante me permettant d'analyser une grande table de données composée de données très variées puisqu'elle met en relation des informations jamais croisées auparavant. Mais également d'apprendre de nouvelles méthodes statistiques, notamment pour la gestion des données manquantes. Ce stage m'a appris à confronter un travail statistique à un point de vue métier, chose que nous n'apprenons pas à l'école.
Cette expérience de 6 mois m'a également permis de confirmer mon intérêt pour le domaine d'application de la data science à la science animale et à l'agroalimentaire.

RÉFÉRENCES BIBLIOGRAPHIQUES

B. Escofier, J. Pagès (1984), L'analyse Factorielle Multiple, Cahiers du Bureau de recherche opérationnelle. Série Recherche, tome 42, p. 3-68

G. Saporta, Inférence sur les valeurs propres et autres indices en ACP, AFC et ACM

F. Husson, 2018, R pour la statistique et la science des données, p. 213-262

M. Desgraupes (2012-2013), Université Paris Ouest Nanterre La Défense - UFR SEGMI, MATHS/STATS Document 6 : Exemple d'ANOVA

Raymond B. Cattell (1966) The Scree Test For The Number Of Factors, Multivariate Behavioral Research, p. 245-276

TABLE DES FIGURES

TABLE DES TABLEAUX