Contribution au développement d'un outil de quantification du degré d'intensification d'exploitations laitières via les variables prédites sur base de la spectrométrie infrarouge

Auteur : Courtois, Guillaume
Promoteur(s) : Tedde, Anthony; Soyeurt, Hélène
Faculté : Gembloux Agro-Bio Tech (GxABT)
Diplôme : Master en bioingénieur : sciences et technologies de l'environnement, à finalité spécialisée
Année académique : 2021-2022
URI/URL : http://hdl.handle.net/2268.2/15177

LIÈGE université
Gembloux
Agro-Bio Tech

Contribution au développement d'un outil de quantification du degré d'intensification d'exploitations laitières via les variables prédites sur base de la spectrométrie infrarouge

Guillaume Courtois

TRAVAIL DE FIN D'ÉTUDES PRÉSENTÉ EN VUE DE L'OBTENTION DU DIPLÔME DE MASTER BIOINGÉNIEUR EN SCIENCES ET TECHNOLOGIES DE L'ENVIRONNEMENT

ANNÉE ACADÉMIQUE 2021 - 2022

PROMOTEURS : PR. HÉLÈNE SOYEURT & ANTHONY TEDDE

LIÈGE université
Gembloux
Agro-Bio Tech

Contribution au développement d'un outil de quantification du degré d'intensification d'exploitations laitières via les variables prédites sur base de la spectrométrie infrarouge

Guillaume Courtois

TRAVAIL DE FIN D'ÉTUDES PRÉSENTÉ EN VUE DE L'OBTENTION DU DIPLÔME DE MASTER BIOINGÉNIEUR EN SCIENCES ET TECHNOLOGIES DE L'ENVIRONNEMENT

ANNÉE ACADÉMIQUE 2021 - 2022

PROMOTEURS : PR. HÉLÈNE SOYEURT & ANTHONY TEDDE

Remerciements

Je tiens en premier lieu à remercier mes co-promoteurs, la professeure Hélène Soyeurt et Monsieur Anthony Tedde, pour leur suivi, leurs conseils et l'opportunité qu'ils m'ont donnée de travailler sur une thématique complètement différente de ce que j'ai pu étudier tout au long de mon master, à savoir le machine learning et grossièrement l'élevage laitier. Merci pour votre encadrement et votre patience nécessaire à la réalisation de ce travail. Je tiens plus particulièrement à remercier Anthony pour les heures qu'il m'a consacrées en présentiel ou à distance et les progrès considérables qu'il m'a permis de faire en R. Merci !

Je remercie aussi Charles Nickmilder et Sébastien Franceschini tant pour leur aide que pour leur présence et la bonne ambiance qu'ils pouvaient amener. Merci à toute l'unité de statistique, Yves, Mireille, Lionel et Dominique ainsi que les TFistes, Sébastien, Claire et Killian pour l'accueil et l'ambiance dans le service.

Ensuite, je remercie mes parents et ma soeur qui m'ont accompagné tout au long de ce parcours qui n'a pas toujours été facile. Merci d'avoir cru en moi et de m'avoir soutenu. Ca y est, on voit le bout du tunnel ! Je remercie aussi celle qui me supporte depuis quleques années et sans qui pas mal de choses n'auraient pas été possible, Merci Loli.

Pour finir, je tiens à remercier tous les gens que j'ai croisé durant ces 7 dernières années et plus particulièrement mes colocs, cokoteurs et amis, BBKot, Floppy, Sven, Lison, Mrs Short, La Nuit, Ping et le STU ! Je remercie aussi l'équipe de l'**A**lcoolisme **B**ien **I**rréprochable. J'ai perdu quelques mois/années de vie mais j'ai gagné des copains et des souvenirs mémorables. Dans la même optique, je remercie le comité AG avec qui le covid a été fort peu respecté via nos réunions essentielles à la survie du folklore AGéen. Globalement, on a fort bien rigolé ! Merci aussi à la team STE pour ce master de folie et surtout au Rems, Anne-K, la Zouz, Nat et Tom pour ces derniers mois.

Abstract

In order to allow farmers to quantify the economic and environmental interests according to the management method of their dairy farm, a decision support tool is developed within the framework of the *SIMBA* project. This tool uses simulations linked to a life cycle analysis and agent-based models for which some input data are not available on a large scale. Thus, one aspect of the project is to exploit existing data through a machine learning approach to define new variables, including the "green consciousness" of farmers. In this context, knowledge of the intensification of milk production could help improve the assignment of this variable, which is currently very theoretical. Indeed, previous works have shown that the green actions implemented were different according to the intensification of the farm.

Therefore, the aim of this work is to determine an intensification gradient characterising Walloon dairy farms on a large scale. To that end, a working hypothesis was formulated and consists of assuming that intensification is seen in the fine composition of milk. The innovative aspect of the present work consists in predicting this intensification gradient on the basis of 112 predicted variables resulting from the average infrared analysis of tank milk and four variables communicated by the dairy laboratory which were averaged over the year.

To achieve this objective, two approaches were tested. The first one is based on prediction. An intensification gradient was estimated from the score on the first principal component of 15 technico-economic variables from the farm accounts by means of a partial least squares (PLS) regression. The second one is based on clustering, predicting the clusters found by a partial least squares discriminant analysis (PLS-DA) and using their membership probabilities. While the first approach uses data mainly limited to the Liège grassland region (1 776 observations collected from 2009 to 2018 within 285 farms), the second approach uses all available dairy data (42 110 observations collected from 2009 to 2021 within 4 375 farms) making it more representative.

The performance of the PLS model obtained is moderate with R^2_v varying from 0.49 to 0.51. In the second approach, three clusters were identified. The reliability of their prediction by the PLS-DA method varied from 0.69 to 0.70. The correlation between the probability of observations belonging to one of these clusters and the intensification gradient was -0.51. The latter shows that even though the analysis is unsupervised, it was able to highlight different feeding patterns. Indeed, the first cluster can be attributed to extensive farms.

To sum up, the two approaches have their advantages and disadvantages (supervised but not representative vs. unsupervised and representative). Consequently, it is difficult to conclude from the use of one model rather than the other, but both models were in the same direction. However, it is necessary to mention that the milk variables are partly able to highlight the degree of intensification of a farmer's behaviour, which was the objective.

Résumé

Dans le cadre du projet *SIMBA*, un outil d'aide à la décision est développé afin de permettre aux agriculteurs de quantifier les intérêts économiques et environnementaux selon la méthode de gestion de leur exploitation laitière. Cet outil utilise des simulations liées à une analyse de cycle de vie et des modèles basés sur les agents dont certaines données d'entrée ne sont pas disponibles à large échelle. Ainsi, un aspect du projet consiste à exploiter des données déjà existantes via une approche d'apprentissage automatique pour définir ces nouvelles variables dont notamment la "conscience verte" des agriculteurs. Dans ce cadre, la connaissance de l'intensification de la production laitière pourrait permettre d'améliorer l'assignement de cette variable qui est actuellement très théorique. En effet, des travaux antérieurs ont montré que la mise en place d'actions vertes était différente selon l'intensification de l'exploitation.

L'objectif de ce travail vise donc à déterminer un gradient d'intensification caractérisant les exploitations laitières wallonnes à large échelle. Pour ce faire, une hypothèse de travail a été formulée et consiste à supposer que l'intensification se voit dans la composition fine du lait. Ainsi, l'aspect innovant de ce travail consiste donc à prédire ce gradient d'intensification sur base de 112 variables prédites issues de l'analyse moyen infrarouge du lait de tank et de quatre variables communiquées par le laboratoire laitier qui ont été moyennées à l'année.

Pour atteindre cet objectif, deux approches ont été testées. La première est basée sur la prédiction, au moyen d'une régression des moindres carrés partiels (PLS), d'un gradient d'intensification estimé à partir du score sur la première composante principale de 15 variables technico-économiques issues des comptabilités agricoles. La seconde approche repose sur un clustering, la prédiction des clusters trouvés par une analyse discriminante des moindres carrés partiels (PLS-DA) et l'utilisation de leurs probabilités d'appartenance. Comparé à la première approche dont les données sont limitées majoritairement à la région herbagère liégeoise (1 776 observations collectées de 2009 à 2018 au sein de 285 exploitations), la seconde utilise l'entièreté des données laitières disponibles (42 110 observations collectées de 2009 à 2021 au sein de 4 375 exploitations) la rendant plus représentative.

Les performances du modèle PLS obtenues sont modérées avec R^2_v variant de 0,49 à 0,51. Dans le cadre de la seconde approche, trois clusters ont été identifiés. La fiabilité de leur prédiction par la méthode PLS-DA variait de 0,69 à 0,70. La corrélation entre la probabilité des observations d'appartenir à un de ces clusters et le gradient d'intensification était de -0,51, montrant que même si l'analyse est non supervisée, elle a pu mettre en avant des modes d'alimentation différents. En effet, le premier cluster peut être attribué à des exploitations extensives.

Les deux approches apportent donc leurs lots d'avantages et d'inconvénients (supervisé mais non représentatif vs. non-supervisé et représentatif). Il est donc difficile de conclure de l'utilisation d'un modèle plutôt qu'un autre mais les deux modèles allaient dans le même sens. Cependant, il est important de mentionner que les variables issues du lait sont en partie capables de mettre en lumière le degré d'intensification du comportement d'un agriculteur ce qui était l'objectif poursuivi.

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

ACP Analyse en composantes principales

AG Acide gras

$RMSE_V$ Racine de l'erreur quadratique moyenne en validation

ACV Analyse de cycle de vie

Bio Biologique

MG Matière grasse

MIR Moyen infra rouge

PLS Moindres carrés partiels

PLS-DA Analyse discrimination des moindres carrés partiels

R_v^2 Coefficient de détermination en validation [−]

R^2 Coefficient de détermination [−]

RMSE Racine de l'erreur quadratique moyenne [−]

SIMBA Simulating economic and environmental impacts of dairy cattle management using Agent Based Models

Chapitre 1

INTRODUCTION

De tout temps, l'Homme a eu recours à l'utilisation des ressources mises à disposition sur Terre, notamment pour l'alimentation. Au cours de son développement, et aujourd'hui plus encore, la question de la gestion de ces ressources se pose. La production agricole et l'activité qui s'y associe sont essentielles pour nourrir chaque être humain. En effet, sans l'intervention de l'Homme au niveau de la gestion des ressources, ces dernières fourniraient de quoi nourrir 500 millions de personnes. Cela représente 6,25 % des huit milliards d'êtres humains peuplant aujourd'hui la Terre (Mazoyer and Roudart, 2006).

L'agriculture remplit plusieurs fonctions. La première est d'assurer la sécurité alimentaire de la population. Cette fonction est d'autant plus importante que la population mondiale croit de manière exponentielle. L'agriculture assure aussi des fonctions économiques, sociales et environnementales. Les deux dernières fonctions ont évolué sur les 30 dernières années. Une prise de conscience est observée vis-à-vis de l'impact des agriculteurs sur les ressources naturelles. Les méthodes de production d'après seconde guerre mondiale, très intensives, ont causé de plus en plus de dommages. Ces impacts telles que la pollution azotée, l'utilisation de pesticides ou la chute de biodiversité sur l'environnement ont de plus en plus été étudiés (Ryden et al. (1984), Wauchope (1978)). Les premières notions du développement durable sont d'ailleurs apparues en 1987 avec le rapport du Brudhlandt (Brundtland et al., 1987). Depuis, l'enjeu environnemental s'est considérablement renforcé via l'étude des problématiques et l'application de règlementations les enrayant. Parallèlement à cela, des enjeux sociaux sont apparus tels que le bien-être animal et la transparence des systèmes de production. Les consommateurs attachent de plus en plus d'importance quant à la qualité du produit qui leur est fourni. Il en va de même pour les méthodes d'approvisionnement de ceux-ci. Cela va du mode d'élevage des animaux au conditionnement des produits avec un critère d'éthique toujours plus important.

C'est dans ce contexte, et plus particulièrement celui de l'élevage laitier, que s'inscrit le projet *SIMBA* (Simulating economic and environmental impacts of dairy cattle management using Agent Based Models). Les méthodes de gestion des troupeaux exercent une grande influence sur les impacts environnementaux associés. *SIMBA* vise à développer un outil d'aide à la décision associant l'Analyse de Cycle de Vie (ACV [1]) et des modèles basés sur les agents (simulation des effets des choix des agriculteurs). Selon différents modes de gestion des troupeaux, les impacts économiques et environnementaux sont quantifiés. Pour cela, le projet aborde les pistes suivantes (GxABT & LIST, 2018) :

— 1. Étudier l'influence potentielle de la gestion des troupeaux et des systèmes de production sur les impacts environnementaux.
— 2. Étudier les mécanismes améliorant la durabilité des systèmes d'exploitation et comment les évaluer via une analyse de cycle de vie.
— 3. Améliorer la compréhension des décisions prises par les agriculteurs et les aider dans leur choix afin de maximiser leur revenu et minimiser leurs impacts.

1. L'ACV regroupe les informations sur les effets environnementaux d'un système (`https://www.belgium.be/fr/environnement/consommation_durable/labels_ecologiques/analyse_du_cycle_de_vie?fbclid=IwAR3esxS56S9DT5IN1EcNZYHX0lwdeP_Ls1pyaCxi_Oea3tJBwkNHW0YqCM0`, consulté le 14/08/2022)

Les données nécessaires pour effectuer cette simulation ne sont pas collectées à large échelle. Ainsi, le projet vise à exploiter les données déjà existantes via une approche de machine learning afin de définir de nouvelles variables, nécessaires au développement des simulations. Une de ces variables est l'influence du comportement de l'agriculteur dans le modèle. Dans ce cadre, des travaux ont été menés pour un indice de "conscience verte" des agriculteurs (Marvuglia et al. (2022), Marvuglia et al. (2017)). Cet indice est basé sur l'hétérogénéité du comportement d'agriculteurs. C'est à dire sur l'importance que chacun décide d'attribuer à la durabilité environnementale vis-à-vis de la stratégie agricole entreprise. Pourtant, des travaux récents montrent que le degré d'intensification de la production engendre des développements "verts" différents. Les exploitations intensives iront plus facilement vers l'installation d'un récupérateur de chaleur, la mise en place de panneaux photovoltaïques ou d'unités de biométhanisation. Elles s'adaptent à une économie d'échelle en s'inscrivant dans une optique de haute productivité. Les exploitations plus extensives vont, quant à elles, plus se diriger vers la mise en place d'un système diversifié et de mesures favorisant la biodiversité. Elles tendent à réduire leurs impacts sur l'environnement et leur production est destinée au marché local. Or, Dalcq (2020) a montré qu'il était possible de prédire ce degré d'intensification à partir des données technico-économiques collectées dans les comptabilités agricoles. Malheureusement, ce type de données n'est pas disponible pour toutes les exploitations.

Par conséquent, l'objectif de ce travail vise à estimer à large échelle ce gradient d'intensification à partir de variables prédites notamment sur base du spectre du moyen infrarouge (MIR). L'hypothèse de ce travail repose sur le fait que des liens indirects existent entre la composition fine du lait et le degré d'intensification de la production laitière.

Après une introduction, un état de l'art justifiant l'hypothèse de travail sera réalisé, ensuite les objectifs, matériels et méthodes seront abordés. Finalement les résultats seront présentés et discutés afin de formuler les conclusions et les perspectives relatives à ce travail.

1.1 État de l'art

L'intensification peut se définir de plusieurs façons : l'augmentation de l'utilisation d'intrants (énergie, fertilisant, irrigation, ...) pour produire plus de lait pour une même surface de terre (Zealand, 2004) ou la maximisation de la productivité du facteur le plus rare, traditionnellement la surface agricole (García-Martínez et al., 2009). Dans l'optique de ce travail, une exploitation laitière intensive a tendance à être spécialisée dans la production de lait standard. Elle se base généralement sur les nouvelles technologies et tend à se fournir en aliments importés. L'alimentation utilisée par l'éleveur va accroître la production via l'utilisation de concentrés. Leur production est vouée au marché mondial. Les exploitations extensives se caractérisent par une production diversifiée, pouvant combiner viande et lait. Leur production de lait peut être de qualité différenciée. Elles ne sont pas basées particulièrement sur les nouvelles technologies et développent un marché local. Elles se caractérisent aussi par la tendance à optimiser l'alimentation herbeuse en bonne saison et d'un point de vue plus large, à optimiser l'utilisation de ressources locales.

1.1.1 Composition du lait

Comme mentionné précédemment, l'hypothèse qui sous-tend ce travail est que le degré d'intensification de la production laitière peut se voir dans la composition fine du lait étant donné que celle-ci va varier.

En fonction des aliments ingérés par la vache, la majorité des constituants du lait sont synthétisés par la glande mammaire sur base des nutriments bruts absorbés sélectivement dans le sang (McDonald et al., 2010). D'autres constituants, tels que des vitamines, des minéraux ou certaines protéines sont directement transférés au lait par cette même glande. Pour obtenir les nutriments bruts absorbés dans le sang, la digestion s'effectue sur plusieurs niveaux.

Les matières azotées sont dégradées dans le rumen pour produire de l'ammoniac. L'azote non protéique est directement transformé en ammoniac alors que les protéines sont d'abord transformées en acides aminés et ensuite en ammoniac via une fermentation. Les micro-organismes du rumen utilisent cet ammoniac pour synthétiser leurs propres protéines microbiennes. Les protéines (alimentaires ou microbiennes) qui arrivent dans la caillette sont dégradées par voie enzymatique en acides aminés et ces derniers sont absorbés par le sang (Waghorn and Clark (2004) et Cuvelier et al. (2020)). Les acides aminés sont ensuite absorbés par la glande mammaire et permettent la synthèse des protéines du lait. Environ 95 % de l'azote du lait se retrouve sous forme de protéines. Cette fraction protéique est dominée par les caséines et ensuite par la β-lactoglobuline (McDonald et al., 2010).

Les vitamines et composés inorganiques issus de l'alimentation sont absorbés au niveau du tube digestif. Ces composants du lait sont ensuite absorbés du sang par la glande mammaire (McDonald et al., 2010).

Les lipides issus de l'alimentation sont en majorité composés de triglycérides. Ces derniers subissent une lipolyse dans le rumen et sont dégradés en AG libres, saturés et insaturés. Certains AG insaturés peuvent subir le processus d'hydrogénation ruminale avant leur absorption, les autres AG sont absorbés dans les intestins. L'hydrogénation de certains AG insaturés a pour conséquence la diminution de la teneur en AG insaturés absorbés au profit des teneurs en AG saturés (Doreau and Chilliard, 1997). Les acides gras retrouvés dans le lait proviennent de quatre sources. Ils peuvent provenir de la lipolyse, de l'hydrogénation, mais aussi de la synthèse *de novo* dans la glande mammaire ou sont d'origine bactérienne. Les lipides bactériens sont originaires de sources exogènes (aliments riches en AG longue chaîne) et endogènes (synthèse *de novo*). La contribution de chaque source dépend de la teneur en lipides du régime alimentaire et des espèces bactériennes du rumen (Jenkins, 1993). Les AG représentent près de 98 % de la MG du lait et leur diversité en termes de nombre avoisine approximativement 400 (Riuzzi et al., 2021).

Sous l'action des enzymes hydrolytiques microbiennes, les glucides sont hydrolysés dans le rumen. Le résultat de ce processus de dégradation est le glucose, qui va à son tour se transformer en acide pyruvique via des fermentations microbiennes. L'acide pyruvique se dégrade en AG volatils sous la forme d'acide acétique, d'acide propionique et d'acide butyrique. Ces AG volatils sont ensuite absorbés à travers la paroi du rumen. Les ruminants sont caractérisés par un mécanisme leur permettant de synthétiser du glucose sur base d'acide propionique. Une deuxième source de glucose provient de la digestion ruminale de l'amidon. Le lactose est presque le seul sucre dans le lait. Il est produit de l'union d'une molécule de glucose et une molécule de galactose. Le galactose dérive du glucose et est synthétisé dans la glande mammaire (Cuvelier et al. (2020), McDonald et al. (2010)). La Figure 1.1 représente synthétiquement l'origine des constituants majeurs du lait.

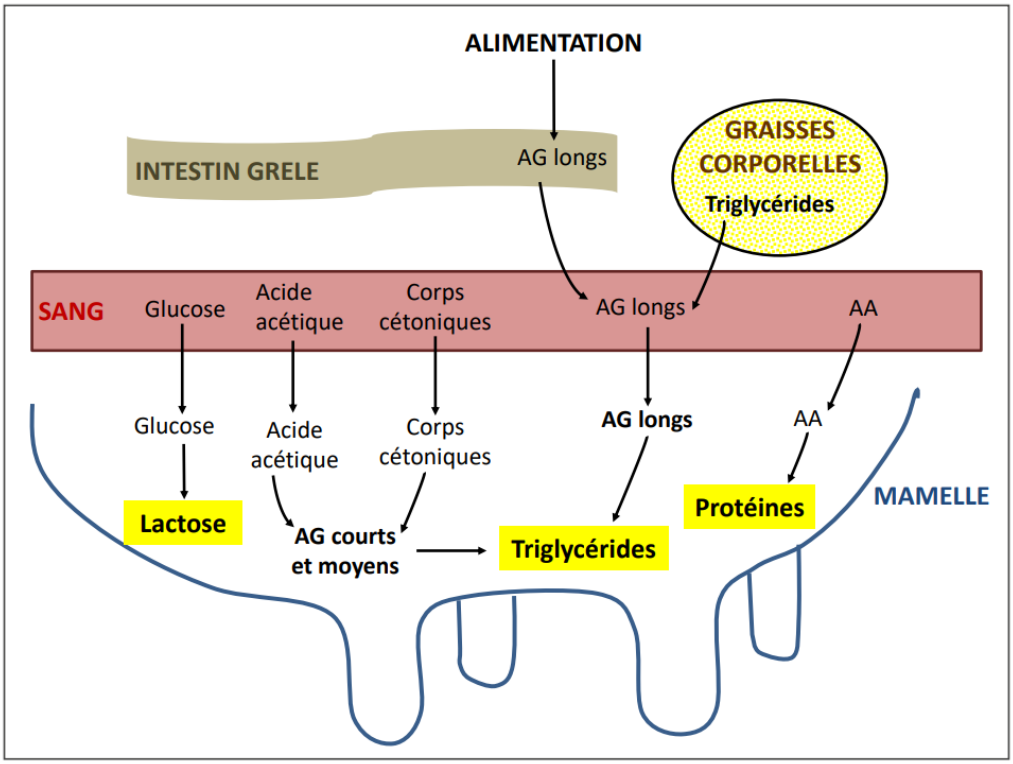


FIGURE 1.1 – Métabolisme permettant la synthèse du lait au sein de la mamelle (Cuvelier et al., 2020).

1.1.2 Alimentation des vaches

La connaissance des différents types d'aliments donnés aux vaches laitières est donc essentielle à la compréhension des différences en termes de composition, de production du lait et de mode de gestion d'exploitation. Les différents types de fourrages, d'ensilages, de concentrés et de mélanges minéraux sont abordés ci-dessous.

Les fourrages verts L'herbe pâturée peut constituer une ration complète. Les prairies gérées par l'Homme se composent d'un assortiment de légumineuses et de graminées. En général, deux espèces cohabitent et dominent l'espace. Cependant, augmenter la diversité d'espèces fourragères tend à augmenter la stabilité et la productivité de la prairie (Tracy and Sanderson, 2004). La composition de la pâture, et donc sa valeur alimentaire est influencée par son stade de végétation. Les jeunes pousses sont caractérisées par une haute teneur en matière azotée totale et en sucres solubles avec peu de cellulose (Cuvelier et al., 2020). Avec le temps, les teneurs en fibres et matière sèche augmentent et les teneurs en matières azotées totales et en sucres diminuent.

Au niveau de sa composition, l'herbe en début de croissance a une teneur en matière sèche de 15 %. Cette teneur peut atteindre les 25 % à la fin du mois de mai (Demarquilly et al., 1964). La teneur en cellulose brute représente 18 à 37 % de cette matière sèche. La teneur en matières azotées représente, quant à elle, 17 à 21 % de la matière sèche (Demarquilly et al., 1964). Les teneurs en minéraux tels que le calcium, le potassium, le phosphore, le sodium ou le magnésium représentent respectivement entre 0,5 et 0,9 %, 2,5 et 3,8 %, 0,32 et 0,49 %, 0,012 et 0,038 % et 1,5 et 2,5 % de la matière sèche (Schlegel et al., 2016). Les lipides représentent 4 à 12 % de la matière sèche dont 45 à 65 % sont des AG. Le reste est constitué de lipides complexes et de pigments (Morand-Fehr and Tran, 2001). Les AG se composent majoritairement d'AG insaturés. En effet, l'acide linolénique représente plus de 50 % des AG et l'acide linoléique entre 10 et 20 % (Rouille and Montourcy, 2010).

Les ensilages L'ensilage est une technique de conservation qui permet de stabiliser les fourrages via une diminution de pH par fermentation anaérobique. La teneur en sucres solubles est quasi nulle du fait de leur consommation par les bactéries. Les aliments qui sont les plus utilisés en ensilage sont l'herbe, les maïs, les céréales immatures et les dérivés de betteraves. Les compositions dépendent donc des aliments initialement utilisés mais aussi de la qualité d'ensilage.

Un bon ensilage d'herbe se compose d'un haut taux en matières azotées et d'un pauvre taux en fibre. La composition en AG est proche de celle du fourrage vert. Il en va de même pour les autres types d'ensilage. Cela est vrai s'il n'a pas subi de préfanage (Dewhurst and King, 1998). L'oxydation et les transformations des AG polyinsaturés dues au fanage tendent à diminuer la teneur en lipides. L'ensilage de maïs est, de manière générale, un fourrage pauvre en nutriments mais riche en énergie. En termes d'AG, sa composition est proche de celles des grains. Le maïs est riche en acide linoléique et oléique et pauvre en acide linolénique. Il a une teneur en MG assez faible (Rouille and Montourcy, 2010).

Les concentrés Les concentrés sont riches en énergie, en matière sèche (88 %) et présentent une teneur en matières azotées variant de 16 à 18 %. Les graines protéagineuses et oléagineuses possèdent une teneur élevée en protéines. Il existe deux types de concentrés : les simples et les composés. Les premiers sont des matières premières alors que les seconds résultent d'un mélange d'aliments concentrés simples. Leur fonction principale est d'équilibrer les rations de base. Ils sont aussi utilisés afin de soutenir la production laitière. Ils proviennent soit du commerce, soit ils sont produits sur place (Cuvelier et al., 2020).

1.1.3 Influence de l'alimentation sur le lait

Différentes variables sont abordées pour identifier les types de comportements des agriculteurs en termes d'intensivité. Ces variables sont utiles pour observer le comportement du gradient les caractérisant.

La production laitière La quantité de lait produite par vache peut être utilisée comme indicateur d'intensivité. En effet, il a été démontré qu'une augmentation du degré d'intensification d'une exploitation induisait une augmentation de la production laitière. C'est le cas dans le cadre de l'étude menée par Chobtang et al. (2017). Ces derniers ont construit leur index d'intensification sur base du taux de stockage (unité de bétail/ha), de la quantité totale d'aliments importés et de la quantité totale de fertilisants azotés utilisées. Les résultats de l'étude démontrent que la quantité de lait produite par vache augmente avec le degré s'intensification de l'exploitation. Dans le même ordre d'idée, la durabilité d'exploitations laitières a été évaluée. Les exploitations avec une faible densité de bétail, une faible intensité de peuplement et une bonne autosuffisance alimentaire sont performantes environnementalement parlant (Penati et al., 2011). La variable de production de lait semble donc intéressante pour déceler les comportements d'exploitations plus ou moins intensives.

Influence du système d'exploitation (bio/conventionnel) Les exploitations agricoles sont caractérisées par un système d'exploitation, biologique ou conventionnel. Ce type de système peut apporter une indication quant au mode de fonctionnement de l'exploitation. Une exploitation en bio a tendance à se rapprocher d'une exploitation de type extensif par son fonctionnement. En effet, ces exploitations se différencient de celles en conventionnel sur plusieurs plans. Elles diminuent la charge animale à l'hectare, réduisent leur surface de culture, privilégient la production végétale locale et l'utilisation de cette production aux concentrés dans l'alimentation des vaches, etc. (Ollivier and Guyomard, 2013). Les exploitations bio sont donc plus autonomes en termes de matières sèches et matières azotées. Elles se caractérisent comme fertilisant peu et achetant peu de concentrés (Paccard et al., 2003).

Influence sur les teneurs en MG et AG du lait La teneur en MG du lait et sa composition en acide gras peut varier selon plusieurs facteurs liés au milieu (alimentation, saison, traite) ou zootechniques (état sanitaire, génétique, stade physiologique). Ces derniers représentent les principaux moyens de variation utilisés par l'éleveur car il n'a que peu d'influence sur les facteurs liés au milieu. Au sein des facteurs zootechniques, l'alimentation est le plus utilisé et agit rapidement (Hoden and Coulon, 1991).

Le mode d'alimentation des vaches va donc impacter directement la teneur en MG et les AG la composant (Rouille et al., 2011). Une exploitation qualifiée d'extensive cherche à être autonome en ce qui concerne l'alimentation. Elle privilégie le pâturage en bonne saison et l'utilisation d'aliments locaux ou issus des cultures de l'exploitation. Les exploitations intensives ont tendance à utiliser des aliments concentrés importés. Leur autonomie alimentaire est inférieure à celle des exploitations extensives (Paccard et al., 2003).

La mise au pâturage se caractérise par une diminution du taux de MG ($kg/jour$) ainsi qu'une diminution de quantité de lait produit (Kelly et al. (1998), Morand-Fehr and Tran (2001)). Cette diminution en MG s'accompagne d'une augmentation du rapport AG insaturés/AG saturés. Le pâturage induit une augmentation des proportions en AG polyinsaturés et linolénique conjugué (Elgersma, 2015). En effet, un taux en AG polyinsaturés quatre fois supérieur est observé pour un lait de vache en pâture. L'herbe fraiche est un aliment plus riche en acide linolénique (Chilliard et al., 2000). Les concentrations en acide linolénique dans la matière grasse laitière sont doublées lorsque les vaches sont en pâturage par rapport à celles alimentées avec une ration totale mélangée. Le lait de vache au pâturage présente donc une réduction de la teneur en AG synthétisés *de novo* (C4 à C14) et une augmentation de la teneur en AG à longues chaînes et en acide linolénique (Kelly et al., 1998). En d'autres mots, ses teneurs en AG longues chaînes augmentent et ses teneurs en AG courtes et moyennes chaînes diminuent (Frelich et al., 2012).

1.2 Objectif

L'objectif général de ce travail est de développer un outil d'estimation du degré d'intensification d'une exploitation agricole laitière via des variables issues de l'analyse infrarouge du lait de tank moyennées à l'année. Ces variables sont prédites sur base des données issues du spectre MIR du lait. Les données spectrales sont mesurées via les données d'échantillons de lait de tank, récoltés dans le cadre du paiement du lait en Wallonie par le Comité du Lait (Battice, Belgique).

Le spectre MIR représente l'absorption des rayons infrarouges à des fréquences corrélées à la vibration des liaisons chimiques spécifiques dans une molécule. Depuis les années 1980, la spectrométrie dans le moyen infrarouge (MIR) permet la détermination des teneurs en protéines et en matières grasses du lait (Hammami et al., 2015). Pour ce faire, des équations de prédiction sont développées pour estimer ces teneurs sur base du spectre des échantillons de lait. Un nombre croissant de composantes du lait et de biomarqueurs du statut métabolique des vaches peut être estimé via cette technique, avec les performances les caractérisant. Par exemple, des équations de prédictions ont été développées

dans le domaine de la quantification de différents AG (Soyeurt et al. (2006), Rutten et al. (2009)), la proportion de lactoferrine (Soyeurt et al., 2012), la composition en protéines (De Marchi et al., 2009), la teneur en minéraux (Soyeurt et al., 2009), etc.

Les principaux avantages de la spectrométrie MIR sont la facilité et la vitesse de la prise de mesures des composantes du lait ainsi que des biomarqueurs en routine. Elle peut être utilisée sur les données collectées pour le paiement du lait ou individuellement par exemple lors du contrôle laitier.

Deux approches sont utilisées afin de répondre à l'objectif, une première approche utilisant un modèle de régression et une seconde utilisant une méthode de regroupement hiérarchisé.

La première méthode consiste en une régression PLS (moindres carrés partiels, "partial least square" en anglais) utilisant les variables prédites du lait et un gradient d'intensification mis au point par Dalcq (2020). Il prédit directement un nouveau gradient d'intensification réalisable sur l'ensemble de la base de données laitières.

La deuxième méthode consiste en l'élaboration d'un clustering hiérarchisé sur base duquel une régression PLS-DA (analyse discriminante des moindres carrés partiels, "partial least square - discriminant analysis" en anglais) est implémentée. La régression reprédit l'appartenance de chaque observation sur base des données lait et du clustering. Les coefficients d'appartenance de chaque cluster sont calculés et conservés.

Chapitre 2

MATÉRIEL ET MÉTHODES

2.1 Logiciel utilisé

Le langage R et l'environnement *RStudio* via le logiciel R version 4.1 (R Core Team, 2016) ainsi que différents packages : dplyr (Wickham et al., 2022), data.table (Dowle and Srinivasan, 2021), tidyverse (Wickham et al., 2019), magrittr (Kuhn, 2022), FactoMineR (Lê et al., 2008), factoextra (Kassambara and Mundt, 2020), ggplot2 (Wickham, 2016), pls (Liland et al., 2021) et caret (Kuhn, 2022) sont utilisés dans le présent travail dans la création de bases de données fonctionnelles, la création de regroupement hiérarchisé et l'élaboration de modèles de prédiction.

2.2 Données

Les données laitières utilisées dans ce travail proviennent de la convention *FuturoSpectre*. Cette convention lie l'Association Wallonne de l'Élevage (*AWé*, Ciney, Belgique), le Centre Wallon de Recherches Agronomiques (*CRA-W*, Gembloux, Belgique), le Comité du Lait (Battice, Belgique) et le campus de Gembloux Agro-Bio Tech de l'Université de Liège (Gembloux, Belgique). Les données spectrales utilisées dans le présent travail sont issues de l'analyse du lait effectuée par un spectromètre MIR de type Foss MilkoScan FT6000, FT4000 et FT+ (Hillerod, Danemark). Elles ont été collectées entre 2009 et 2022 auprès de 5 283 exploitations wallonnes pour un total de 49 227 observations. Tous les échantillons ont été collectés dans le cadre du paiement du lait, il s'agit donc de lait de tank.

Au total, 112 équations ont été appliquées sur ces données laitières afin de prédire 112 variables permettant d'apprécier la composition fine du lait, la quantité de méthane éructé ou encore la quantité de lait produite. La précision de la plupart de ces différentes équations est accessible dans l'Annexe 7.1. Les teneurs en AG initialement prédites sont exprimées en g/dl de lait ont été converties en g/100g de MG pour mieux refléter les processus métaboliques responsables de leur production. Pour ce faire, ces teneurs ont été divisées par le taux en MG prédit directement par le spectromètre. Toutes ces variables sont ensuite moyennées à l'année et par exploitation afin de s'aligner aux données technico-économiques collectées sur une base annuelle.

Les données technico-économiques ont, quant à elles, été collectées entre 2006 et 2018 auprès de 1 018 exploitations wallonnes pour un total de 8 364 observations, situées majoritairement à l'est de la Région Wallonne. Ces données ont été communiquées par Elevéo (*AWé*, Ciney, Belgique).

2.2.1 Données lait

Après l'effacement des enregistrements avec données manquantes et d'après les seuils communiqués par l'International Committee for Animal Recording (ICAR, 2017), les observations dont les teneurs en MG ne sont pas comprises entre 1,5 % et 9 % ont été écartées. La composition laitière varie d'une année à l'autre à cause des conditions climatiques. Par conséquent, cet effet temporel a été gommé

via une normalisation annuelle de chaque variable. Pour cela, la moyenne annuelle de chaque variable a été calculée et chaque observation a été soustraite de la moyenne correspondante.

Ensuite, Les données extrêmes ont été détectées via le calcul de la distance de Mahalanobis standardisée, ou distance GH pour chaque observation. Pour ce faire, une analyse en composantes principales (ACP) est nécessaire. Cette ACP est effectuée sur l'ensemble des prédictions MIR du set de données laitières nettoyé des données manquantes et selon le seuil ICAR. Suite à cette ACP, 22 composantes principales sont définies en fonction de la part de variabilité qu'elles expliquent, à savoir 99 %. Cette analyse a été réalisée en utilisant le package *FactoMinR* (Lê et al., 2008). Les distances GH sont alors calculées en utilisant le package *Stats* (R Core Team, 2016). La distance GH se calcule via l'Équation 2.1 (Zhang et al., 2021) :

$$GH = \frac{(\bar{x} - \bar{\mu})^T \; S^{-1} \; (\bar{x} - \bar{\mu})}{nCP} \tag{2.1}$$

Où :
— $\bar{x}$ est le vecteur de composantes principales de l'observation ;
— $\bar{\mu}$ est le vecteur des moyennes de chaque composante principale ;
— S est la matrice de variance-covariance de ces mêmes composantes principales ;
— nCP est le nombre de ces composantes principales.

Les observations caractérisées par un GH supérieur à 5 sont considérées comme valeurs extrêmes et ont été supprimées (Zhang et al., 2021). Ces données supprimées sont au nombre de 278 et représentent 0,6 % des observations initiales. Au total, 8,8 % des observations ont été supprimées.

2.2.2 Données technico-économiques et estimation du gradient d'intensification

Les mêmes prétraitements que ceux utilisés par Dalcq (2020) ont été appliqués en vue de calculer le gradient d'intensification des fermes. Ces prétraitements sont détaillés à la Figure 2.1. les variables suivantes ont été utilisées par Dalcq (2020) pour calculer le gradient d'intensification : vaches traites par hectare de surface fourragère (VACTRAHASF), quantité de lait produite par hectare de surface fourragère (LITLAIHASF), nombre d'unités de bétail par hectare de surface fourragère (UGBTOTHASF), pourcentage de pâturage dans la surface fourragère (PCPRAIRISF), pourcentage de culture de maïs (PCMAISSF) et pourcentage d'ensilage d'herbe (PCSCONSSF), pourcentage de première coupe de foin (PCFAUCH1C), pourcentage d'autres coupes de foin (PCFAUCHAC), pourcentage de première coupe d'ensilage (PCENS1C), pourcentage d'autres coupes d'ensilage (PCENSAC), nombre d'unités azotées appliquées par hectare de pâture (NHAPRE), surface totale de pâture (AREUGBPRE) et d'ensilage de maïs (AREUGBMAI) par unité de bétail, équivalents de concentrés par vache traite (EQCONCVT), quantité de lait produit par vache traite (LITVACHEVT).

```
#<<---------------------- Filtation of variables ACD ---------------------------

# if MkcowsHAFoA=0.01 or MkcowsHAFoA=0.02  then MkcowsHAFoA = MkcowsHAFoA*100
x <- (data$VACTRAHASF == 0.01)
y <- (data$VACTRAHASF == 0.02)
tmp <- x | y
data[tmp,'VACTRAHASF'] <- data[tmp,'VACTRAHASF'] * 100

# if UMOFAMIL = 0.1 then UMOFAMIL = 1
tmp <- (data$UMOFAMIL == 0.1)
data[tmp,'UMOFAMIL'] <- 1
# if UMOFAMIL < 0.1 then UMOFAMIL = UMOFAMIL*100
tmp <- (data$UMOFAMIL < 0.1)
data[tmp,'UMOFAMIL'] <- data[tmp,'UMOFAMIL'] * 100
# if UMOFAMIL =0 then UMOFAMIL=.
tmp <- (data$UMOFAMIL == 0)
data[tmp,'UMOFAMIL'] <- NA

# if UMOTOTAL < 0.1 then UMOTOTAL = UMOTOTAL*100
tmp <- (data$UMOTOTAL < 0.1)
data[tmp,'UMOTOTAL'] <- data[tmp,'UMOTOTAL'] * 100
# if UMOTOTAL =  0.1 then UMOTOTAL = UMOTOTAL*10
tmp <- (data$UMOTOTAL == 0.1)
data[tmp,'UMOTOTAL'] <- 1
# if UMOTOTAL =  0.5 then UMOTOTAL = 1.25
tmp <- (data$UMOTOTAL == 0.5)
data[tmp,'UMOTOTAL'] <- 1.25
# if UMOTOTAL =  0.25 then UMOTOTAL= .
tmp <- (data$UMOTOTAL == 0.25)
data[tmp,'UMOTOTAL'] <- NA

# if grazedareaLU < 1 and grazedareaLU > 0 then grazedareaLU = grazedareaLU*100
x <- (data$AREUGBPRE < 1)
y <- (data$AREUGBPRE > 0)
tmp <- x & y
data[tmp,'AREUGBPRE'] <- data[tmp,'AREUGBPRE'] * 100

# if cornsilageLU <= 0.1 and cornsilageLU > 0 then cornsilageLU = cornsilageLU*100
x <- (data$AREUGBMAI <= 0.1)
y <- (data$AREUGBMAI > 0)
tmp <- x & y
data[tmp,'AREUGBMAI'] <- data[tmp,'AREUGBMAI'] * 100

# if CFERME='902124' then  PCosilagecutFoA = .
tmp <- (data$CFERME == 902124)
data[tmp,'PCENSAC'] <- NA
# if CFERME='902124' then PCsilagegrassFoA = .
tmp <- (data$CFERME == 902124)
data[tmp,'PCSCONSSF'] <- NA
# if CFERME='902124' then PC1haycutFoA=. OR if PC1haycutFoA > 100 then PC1haycutFoA = .
x <- (data$CFERME == 902124)
y <- (data$PCFAUCH1C > 100)
tmp <- x | y
data[tmp,'PCFAUCH1C'] <- NA

# if CFERME='902124' then PCohaycutFoA=.
tmp <- (data$CFERME == 902124)
data[tmp,'PCFAUCHAC'] <- NA
# if CFERME='902124' then PC1silagecutFoA=. OR if PC1silagecutFoA > 100 then PC1silagecutFoA=.
x <- (data$CFERME == 902124)
y <- (data$PCENS1C > 100)
tmp <- x | y
data[tmp,'PCENS1C'] <- NA

# if VACHHAPAT < 0.1 then VACHHAPAT = VACHHAPAT*100
tmp <- (data$VACHHAPAT < 0.1)
data[tmp,'VACHHAPAT'] <- data[tmp,'VACHHAPAT'] * 100
# if VACHHAPAT=0 then VACHHAPAT = .
tmp <- (data$VACHHAPAT == 0)
data[tmp,'VACHHAPAT'] <- NA

# if BIO=' ' then BIO='0'
tmp <- (data$BIO == '')
data[tmp,'BIO'] <- 0

# if MilkHAFoA=0 then MilkHAFoA = .
tmp <- (data$LITLAIHASF == 0)
data[tmp,'LITLAIHASF'] <- NA

# if LUHAFoA=0.01 or LUHAFoA=0.02 or LUHAFoA=0.03 or LUHAFoA=0.04 then LUHAFoA = LUHAFoA*100
x <- (data$UGBTOTHASF == 0.01)
y <- (data$UGBTOTHASF == 0.02)
z <- (data$UGBTOTHASF == 0.03)
zz <- (data$UGBTOTHASF == 0.04)
tmp <- x | y | z | zz
data[tmp,'UGBTOTHASF'] <- data[tmp,'UGBTOTHASF'] * 100

#if milkMkCow < 900 then  milkMkCow = .
tmp <- (data$LITVACHEVT < 900)
data[tmp,'LITVACHEVT'] <- NA

# if exercice = '6' then delete
tmp <- (data$exercice == 6)
data[tmp,'exercice'] <- NA
```

FIGURE 2.1 – Prétraitements appliqués par Dalcq (2020) sur le set de données technico-économiques.

Une première ACP a été réalisée sur ces 15 variables pour évaluer la variabilité du set de données afin de détecter les valeurs extrêmes. Un seuil de 100 % variabilité expliquée est utilisé ici pour définir le nombre de composantes principales à utiliser pour calculer la distance GH des observations car le nombre de variables est assez faible. Les observations dont la distance GH excédait cinq ont été écartées du jeu de données.

La méthode développée par Dalcq (2020) a été scrupuleusement suivie afin de reproduire le plus rigoureusement possible le gradient d'intensification étant donné que nos données couvraient une année supplémentaire (i.e., 2017 à 2018). Le calcul du gradient a consisté à extraire les valeurs de la projection des enregistrements de la base technico-économique sur la première composante principale calculée à partir d'une ACP réalisée sur les 15 variables susmentionnées sur le jeu de données nettoyé.

2.3 Prédiction sur base du gradient

Afin de permettre une prédiction à large échelle de ce gradient d'intensification, une équation de prédiction du gradient est développée à partir des données laitières. Pour son développement, les deux sets de données, laitières et technico-économiques, sont joints l'un à l'autre. Cela est rendu possible par l'utilisation d'un fichier fourni par l'*AWé* permettant de connecter les identifiants des exploitations présentes dans le set de données technico-économiques et celles présentes dans le set de données laitières.

2.3.1 Division des jeux de données

Afin de permettre une validation externe, le set de données fusionné, appelé "account milk", est divisé en deux sous-jeux de données. La division est réalisée de sorte qu'une exploitation ne se retrouve pas dans le set d'entraînement et de validation pour éviter une complaisance du jeu de validation vis-à-vis de celui d'entraînement. Ainsi, la division s'est effectuée aléatoirement de la manière suivante : 80 % des exploitations en entraînement et le reste en validation.

2.3.2 Modèle de prédiction

La méthode PLS est utilisée dans le but de prédire un gradient d'intensification sur base des variables composant le set de données lait et du gradient initial (Dalcq, 2020). La méthode PLS est une méthode de régression dont l'objectif est d'atteindre un compromis entre la maximisation de la variance expliquée par l'axe X (principe de l'ACP) et la maximisation de la variance expliquée par l'axe Y. Ce modèle est intéressant pour le traitement de nombreuses variables corrélées et la réduction de la dépendance interne des variables explicatives tout en augmentant le pouvoir de prédiction (Wold et al., 2004). L'utilisation de ce modèle est réalisée par le package pls (Liland et al., 2021) via l'interface fournie par le package caret (Kuhn, 2022). Le nombre de variables latentes utilisées dans le modèle est fixé sur base de la validation croisée. Ainsi, à l'aide de la fonction "groupKFold" du package caret (Kuhn, 2022), la validation croisée a divisé aléatoirement le set d'entraînement en 10 sous-jeux de données, toujours en respectant l'indépendance des exploitations. Le modèle a ensuite été validé en appliquant le modèle obtenu sur le set de validation (validation externe).

Les performances du modèle sont ensuite calculées pour évaluer la qualité de ses prédictions. Pour cela, plusieurs paramètres statistiques sont utilisés tels que le coefficient de détermination (R^2) et la racine de l'erreur quadratique moyenne ($RMSE$). Le R^2 est la proportion de la variance de la variable dépendante expliquée par le modèle. La $RMSE$ permet, quant à elle, d'évaluer la variabilité de la qualité de l'estimation. Ces deux paramètres ont été estimés sur base de la validation croisée et de la validation externe.

Afin d'éliminer des échantillons extrêmes du jeu d'entraînement qui pourrait impacter négativement les performances de prédiction, une analyse des résidus issus des prédictions sur le jeu d'entraînement est utilisée. Cette analyse est basée sur la valeur moyenne des résidus (r). Ainsi, les observations sont

éliminées du set d'entraînement selon la règle communiquée à l'Équation 2.2. Dans cette dernière, mean et SD représentent respectivement la moyenne et l'écart-type du résidu.

$$|r| = |grad_{prédit} - grad_{observé}| > |mean \pm (3 \times SD)| \tag{2.2}$$

Ce mécanisme est répété trois fois pour étudier l'évolution du R^2 en validation externe après chaque tour de détection.

2.3.3 Méthode d'analyse du gradient prédit via une variable

Des variables permettant de vérifier le comportement du gradient d'intensification prédit sont utilisées. Lorsqu'elles sont de l'ordre du quantitatif, une corrélation est calculée entre le gradient et la variable. Dans le même temps, quatre classes d'intensification sont élaborées. Elles représentent chacune 25 % des observations. La première représente les observations extensives, la deuxième représente les observations peu intensives, la troisième, les observations intensives et la quatrième, les observations très intensives. Le comportement des observations dans chaque catégorie est étudié vis-à-vis de la variable. Lorsque la variable est qualitative, le comportement du gradient est étudié selon les classes imposées par la variable.

2.4 Prédiction sur base du regroupement

Pour définir le degré d'intensification de la ferme, au lieu d'utiliser le gradient d'intensification développé par Dalcq (2020) estimable uniquement sur base de données technico-économiques, une hypothèse peut être émise : la composition du lait pourrait directement refléter ce gradient. Ainsi, la seconde approche développée dans ce travail a consisté à regrouper les données laitières partageant des caractéristiques communes pour tenter d'observer l'intensification. Le regroupement (clustering en anglais) est donc un apprentissage non supervisé car aucune observation ne se voit attribuer une cible à atteindre.

2.4.1 Regroupement

Afin d'uniformiser les unités, le set de données laitières est en premier lieu standardisé en soustrayant de chaque observation la moyenne de la variable et en divisant ce résultat par l'écart-type de la variable. Ensuite, la distance euclidienne entre chaque couple d'observations a été calculée sur les données laitières nettoyées standardisées. Cette matrice de distance a permis de réaliser le clustering qui s'est basé sur une méthode hiérarchisée agglomérative. Cela implique que chaque observation est considérée initialement comme un cluster. Ainsi, les clusters les plus semblables sont fusionnés jusqu'à ce qu'il n'en reste plus qu'un.

La méthode utilisée ici est la méthode "WARD" (Landau et al., 2011). L'algorithme de "WARD" cherche à minimiser la variance inter-cluster. Dans ce travail, le nombre de clusters choisi est fixé sur base du dendrogramme. Les groupes sont choisis en fonction de leur hauteur afin d'en faciliter l'interprétation. Une augmentation graduelle du gradient d'intensification pourrait apparaître pour les observations selon leur classement dans chaque cluster. Les observations classées dans le premier groupe pourraient être caractérisées par un gradient à tendance extensive alors que les observations classées dans le troisième groupe auraient un gradient à tendance intensive. Ensuite, afin de prédire ces clusters à large échelle, un modèle prédictif a été développé.

2.4.2 Division du jeu de données

Afin de permettre une validation externe, le jeu de données est divisé en deux sous-jeux de données selon les proportions d'exploitations suivantes : 90 % en entraînement et 10 % en validation. Le plus grand nombre de données de la base de données laitières permet l'augmentation du ratio entraînement/validation par rapport au modèle précédent développé sur la base de données technico-économiques.

2.4.3 Modèle de prédiction

Une analyse discriminante utilisant la méthode des moindres carrés partiels (PLS-DA) est utilisée pour prédire l'appartenance de chaque observation à un cluster issu du clustering. Cette prédiction se base sur les mêmes variables que lors de la prédiction du gradient, à savoir les variables issues du set de données laitières. Le nombre de facteurs PLS pris en considération est fixé sur base des résultats issus de la validation croisée sur 10 groupes, effectuée de manière indépendante de l'exploitation. Les performances du modèle, basées sur la qualité de prédiction, sont évaluées via l'utilisation des paramètres statistiques suivants : la justesse et le Kappa.

La justesse se définit comme étant le taux d'observations correctement classées. Il est très intéressant de compléter ce paramètre avec le Kappa qui permet de corriger la justesse des prédictions du modèle par la précision attendue par hasard. Par conséquent, la valeur de l'indice Kappa sera toujours inférieure à celle de la justesse globale (Cohen, 1960). Une valeur du Kappa inférieure à 0,4 est considérée comme pauvre, comprise entre 0,41 et 0,6 comme modérée, entre 0,61 et 0,80 comme bonne et supérieure à 0,81 comme excellente (Landis and Koch, 1977).

Afin d'étudier la pertinence des clusters pour mettre en lumière le degré d'intensification, le modèle développé est appliqué aux observations possédant des données technico-économiques et donc un gradient d'intensification. Une matrice de corrélation est alors estimée entre le gradient initial calculé par la méthode développée par Dalcq (2020) et la probabilité d'appartenir aux différents clusters.

Chapitre 3

RÉSULTATS ET DISCUSSION

3.1 Traitement des bases de données

3.1.1 Base de données laitières

Les données manquantes sont en premier lieu éliminées. Elles représentent 4,7 % des données totales. Ces valeurs se situent exclusivement dans les variables communiquées par le laboratoire, en particulier la teneur en lactose. Une ACP est ensuite effectuée sur l'ensemble des données lait prédites (N = 46 896). Le graphique des individus révèle deux groupes d'observations distincts. Ces deux groupes se différencient sur l'axe de la première composante principale. Une variable sort du lot selon le poids qu'elle exerce sur cette composante, la MG. La présence de si faibles teneurs en MG implique que des données de lait écrémé sont présentes. Afin de rendre la base de données fonctionnelle, un nettoyage est appliqué sur base des normes ICAR. L'année 2022 est écartée suite au fait qu'elle soit incomplète.

Après normalisation, une ACP est alors de nouveau effectuée sur le set de données épuré (N = 42 388). Un total de 23 composantes principales explique 99 % de la variabilité du set de données. Ces composantes sont utilisées pour calculer la distance de Mahalanobis (GH) pour chaque observation. 278 observations ont une distance GH plus grande que cinq et ont donc été effacées. Le set de données contient 42 110 observations après nettoyage. Les statistiques descriptives de ce jeu de données sont présentées dans la Table 3.1.

TABLE 3.1 – Statistiques descriptives des 116 variables à partir du jeu de données laitières nettoyé (nombre d'observations (N) = 42 110).

Variables	Moyenne	Écart-type	Min	Médiane	Max
MG_Labo	-0,002	0,25	-1,549	0,012	1,895
Prot_Labo	-0,028	1,19	-6,41	-0,02	9,348
Urea_Labo	-0,634	47,202	-179,739	-0,188	322,092
Lactose_Labo	0,34	3,876	-91,253	0,125	81,591
Taux_MG	-0,003	0,251	-1,56	0,01	1,887
Lait	0,014	1,35	-7,893	0,152	6,225
CH4_Ame	0,254	16,324	-99,865	1,885	67,173
Lactose	0,002	0,062	-0,533	0,007	0,222
Lactofer_avEMR	-0,584	33,035	-213,23	-4,55	334,717
Lactofer_ssEMR	-1,067	30,039	-153,583	-4,142	215,63
Milk_Lactofer_avEMR	-0,66	33,37	-173,913	-3,335	289,999
Meth_breed	0,277	17,313	-110,274	1,988	86,548
Acidite_titrable_new	-0,004	0,487	-3,05	0,009	2,635
Rdmt_From_Frais_new	-0,001	0,04	-0,215	0	0,323
Rdmt_From_Sec_new	-0,047	1,967	-10,744	0,019	14,386

Taux_Mat_NProt_new	-0,199	8,976	-45,945	-0,201	72,83
Bodyweight	-0,244	6,289	-33,989	-0,34	43,723
Taux_Mat_NProt_OptiMIR	-0,02	0,927	-4,713	-0,017	7,737
Caseine_As1	-0,009	0,388	-1,891	-0,009	3,295
Caseine_As2	-0,008	0,35	-1,732	-0,002	2,92
Caseine_B	-0,006	0,414	-3,165	0,006	3,167
Lactalbumine	0	0,033	-0,25	0,002	0,127
Lactoglobuline	0,002	0,098	-0,568	0,005	0,733
Caseine_TOT	-0,027	1,157	-6,389	-0,016	9,952
Prot_Lactoserum	0,002	0,124	-0,615	0,007	0,75
Taux_Mat_NProt_Levicek	-0,025	1,064	-5,844	-0,027	8,664
Caseine_TOT_Kjeldahl	-0,002	0,093	-0,527	-0,002	0,758
AG_C4	0,015	0,878	-4,778	0,025	4,631
AG_C6	-0,008	0,725	-5,019	0,09	3,486
AG_C8	-0,015	0,662	-4,308	0,073	3,153
AG_C10	-0,063	2,198	-11,581	0,117	10,131
AG_C12	-0,089	2,764	-13,172	0,082	12,568
AG_C14	-0,161	5,748	-26,287	0,265	28,777
AG_C14.1	-0,021	0,556	-3,006	-0,008	2,461
AG_C16	-0,185	14,574	-78,613	0,697	71,54
AG_C16.1_c	-0,017	0,771	-3,15	-0,115	5,351
AG_C17	-0,003	0,168	-0,773	-0,022	1,328
AG_C18	0,096	5,095	-22,332	0,032	32,619
AG_C18.1_t	0,078	3,818	-15,003	-0,254	26,997
AG_c18.1_c9	0,103	15,959	-49,711	-1,788	87,86
AGtot_C18.1_c	0,113	16,916	-53,327	-1,86	93,292
AGtot_C18.2	-0,001	0,978	-5,004	-0,085	5,446
AG_C18.2.c12	-0,012	0,723	-3,294	-0,057	5,232
AG_C18.2.c15	0,002	0,363	-1,721	-0,019	2,318
AG_C18.2.t11	0,012	1,224	-5,515	-0,087	8,826
AG_sat	-0,372	20,628	-97,318	1,61	123,538
AG_monoinsat	0,089	19,44	-62,303	-1,983	116,317
AG_polyinsat	0,013	2,557	-11,951	-0,21	16,993
AG_insat	0,094	21,49	-71,373	-2,165	130,607
AG_cc	-0,072	3,945	-26,428	0,487	18,703
AG_mc	-0,512	22,972	-113,164	1,154	116,842
AG_lc	0,28	26,381	-82,663	-2,247	157,093
AG_isoanteiso	0,003	0,646	-4,029	-0,014	3,995
AG_O3	0,004	0,441	-2,15	-0,025	2,908
AG_O6	-0,009	1,007	-5,291	-0,1	5,851
Odd_FA_TOT	0	0,009	-0,06	0	0,059
AG_TOT_t	0,074	4,614	-19,465	-0,34	32,876
AG_TOT_18.1	0,12	19,532	-62,261	-1,982	111,99
Na	-0,372	17,527	-64,723	-2,858	149,828
Ca	-0,704	31,901	-173,319	-1,833	260,493
P	-0,5	30,915	-164,287	-1,447	193,677
Mg	-0,026	2,683	-11,886	-0,126	20,496
K	0,807	28,74	-184,498	2,554	131,317
Citrates	0,01	0,476	-2,494	-0,001	2,333
Acetone	0	0,008	-0,028	0	0,061
B_Hydroxybutiric_acid	0,223	12,498	-56,725	-0,468	100,566
pH	0	0,021	-0,105	0	0,163

Acidite_titrable	0,003	0,492	-3,102	0,02	2,964
Rdmt_Beurre_Lait	-0,002	0,316	-1,958	0,018	2,199
pH_Beurre	0,001	0,162	-0,819	0,012	0,811
Rennet_Coag	0,002	0,114	-0,54	0,005	0,86
RCT_K20_JG	0,002	0,093	-0,44	0,005	0,66
RCT_K20_JR	2,286	80,875	-399,894	5	595,377
K20_JR_LOG2	0,002	0,057	-0,407	0,005	0,324
A30_racine	-0,045	0,972	-6,821	-0,063	5,191
Rdmt_From_Sec	-0,052	3,605	-19,53	0,152	25,046
Rdmt_From_Frais	-0,036	1,918	-8,892	-0,043	16,315
Rdmt_From_Sec_2	-0,011	1,876	-9,166	0,086	13,294
Rdmt_From_Atelier	-0,01	0,848	-5,787	0,048	4,018
Cheese_Curd	-0,01	0,848	-5,787	0,048	4,018
Cheese_Solid	-0,007	0,308	-1,523	-0,003	2,391
Solid_recov	-0,042	3,25	-19,042	0,077	23,229
Fat_Recov	-0,139	3,833	-27,811	0,149	26,797
Prot_Recov	-0,054	1,527	-10,188	0,017	10,227
RCT_CRM	1,589	53,163	-283,345	0,923	467,918
K20_CRM	0,263	9,136	-61,538	0,096	55,527
RCT_K20_CRM	1,57	55,825	-277,931	0,294	473,592
A30_CRM	-0,038	0,792	-6,681	-0,028	5,13
pH_Yaourt	0	0,019	-0,133	0,001	0,141
Activite_Yaourt	0	0,011	-0,064	0,001	0,04
Mat_seche_Yaourt	-0,004	0,287	-1,921	-0,003	2,319
Synerese_Yaourt	0,02	2,161	-11,695	0,119	11,292
Text_Yaourt	0	0,003	-0,024	0	0,018
Test_Lactofer	-0,835	38,442	-185,817	-5,836	320,736
Energy_balance	0,007	1,434	-6,712	0,033	7,919
Prot_Efficiency	-0,01	0,713	-2,472	-0,061	2,94
Milk_Gluc_6P	0	0,01	-0,057	0	0,056
Milk_Gluc_Free	0,001	0,018	-0,123	0,002	0,074
Milk_Bohb	0,156	7,415	-35,05	0,248	53,075
Milk_IsoK	0	0,008	-0,061	0	0,05
Milk_Urea	-0,012	0,474	-2,54	-0,011	3,183
Milk_NAG	-0,001	0,329	-1,644	0,001	2,445
Milk_LDH	-0,014	0,533	-2,382	-0,063	3,702
Milk_UA	0,101	6,239	-30,343	0,053	28,384
Milk_Progest	-0,002	0,182	-1,017	0,019	1,094
Blood_IGF1	0,256	15,565	-100,031	1,996	46,171
Blood_Glucose	0,001	0,097	-0,475	0,008	0,357
Blood_Urea	-0,006	0,358	-1,72	0,003	2,184
Blood_Cholest	0,004	0,156	-1,035	0,014	0,684
Blood_Fructosamine	0,003	2,7	-14,498	0,273	11,699
Blood_Bohb_Log10	0,001	0,046	-0,319	0,002	0,202
Blood_NEFA	0,828	62,665	-257,44	-2,006	338,848
Blood_Progest	-0,002	0,315	-1,774	0,007	1,422
Dry_Matter_Intake	-0,015	1,237	-6,291	0,103	5,604
RFI_1	0,006	0,884	-4,397	0,026	4,744
RFI_2	0,008	1,498	-7,052	-0,057	9,974

3.1.2 Base de données technico-économiques

Les prétraitements utilisés par Dalcq (2020) sont d'abord effectués. Le set de données composé initialement de 1 018 exploitations pour un total de 8 364 observations passe alors à 670 exploitations pour un total de 4 872 observations. Le set de données comprend des observations s'étalant de 2007 à 2018. Initialement, le set comprenait l'année 2006 mais cette dernière est éliminée suivant les recommandations de Dalcq (2020).

La distance GH est ensuite calculée pour chaque observation avec 15 composantes principales et celles dont la distance dépasse cinq sont éliminées. Après ce nettoyage, le set de données se compose de 664 exploitations pour un total de 4 789 observations. Les statistiques descriptives de ce jeu de données sont présentées dans la Table 3.2.

TABLE 3.2 – Statistiques descriptives des 15 variables à partir du jeu de données nettoyé (nombre d'observations (N) = 4 872)

	Moyenne	**Écart-type**	**Min**	**Q1**	**Médiane**	**Q3**	**Max**
VACTRAHASF	1,224	0,388	0,1	0,99	1,25	1,47	3,09
PCPRAIRISF	89,68	10,43	43	82	91	100	100
PCMAISSF	9,732	9,953	0	0	8	17	57
PCSCONSSF	46,88	16,47	0	37	47	57	112
PCFAUCH1C	6,515	9,037	0	0	3	10	73
PCFAUCHAC	7,097	13,078	0	0	0	10	108
PCENS1C	51,98	24,50	0	40	57	69	100
PCENSAC	81,08	59,73	0	37	76	116	400
NHAPRE	97,77	62,98	0	52	97	139	477
AREUGBPRE	35,61	12,45	8,73	27,36	32,98	41,22	108,3
AREUGBMAI	3,483	3,585	0	0	2,93	5,84	19,44
LITLAIHASF	8310	3397	51	5791	8407	10637	27335
UGBTOTHASF	2,767	0,755	0,92	2,25	2,74	3,22	6,41
EQCONCVT	1693	737	0	1207	1654	2109	5970
LITVACHEVT	6687	1365	936	5888	6823	7612	13491

Le gradient d'intensification est alors calculé par l'utilisation de la première composante principale issue de l'ACP (Figure 3.1) réalisée sur le set de données.

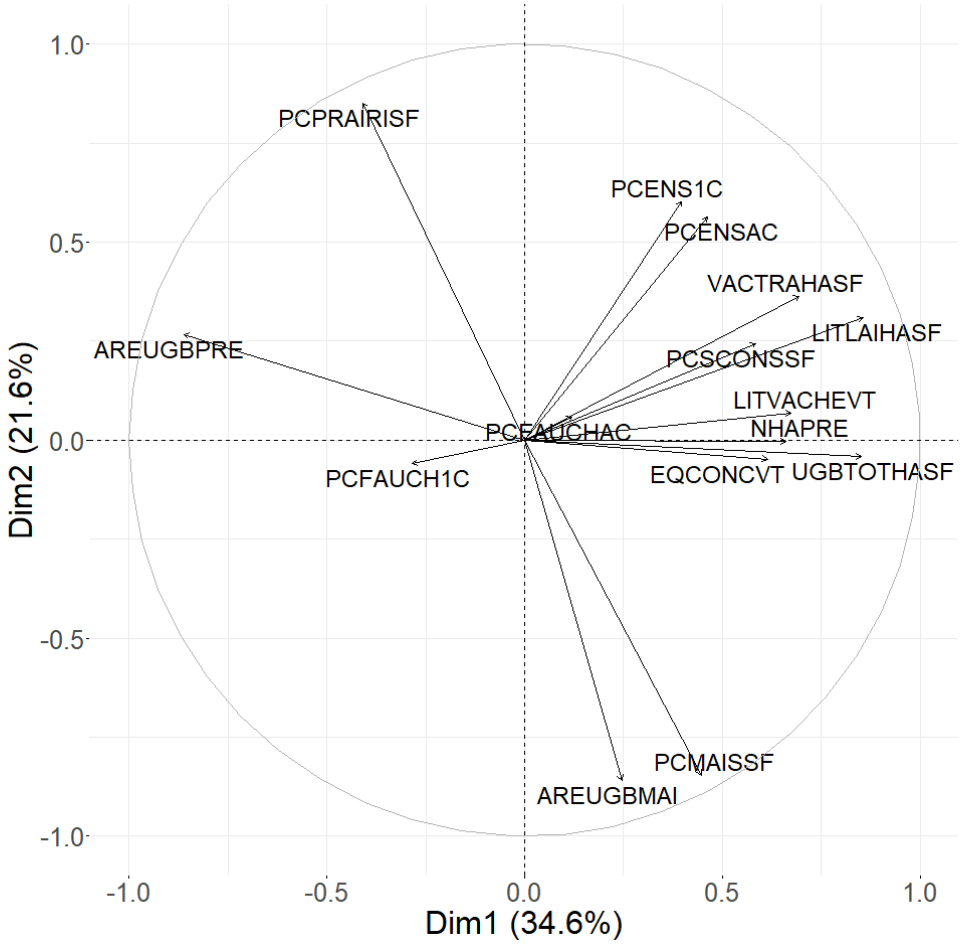


FIGURE 3.1 – Représentation des 15 variables d'intensification étudiées sur les deux premières composantes principales.

La première composante principale explique 34,6 % de la variabilité du set de données. Elle est corrélée positivement aux variables d'intensification telles que les vaches traites par hectare de surface fourragère, la quantité de lait produite par hectare de surface fourragère, le nombre d'unité de bétail par hectare de surface fourragère et le pourcentage d'ensilage d'herbe. Elle est dans le même temps négativement corrélée avec des variables d'extensification telles que la surface totale de pâture par unité de bétail (Figure 3.1).

3.2 Prédiction sur base du gradient

La fusion entre la base de données laitières et la base de données technico-économiques est réalisée : 285 exploitations sont conservées à partir de 2009 jusque 2018 pour un total de 1 776 observations. Deux sets de données sont alors créés, un set de calibration, composé de 1434 observations issues de 228 fermes et un set de validation, composé de 342 observations issues de 57 exploitations. Leur dispersion est visible à la Figure 3.2. Cette dernière permet d'observer la dispersion des deux sous-jeux de données. Le jeu de validation est assez représentatif du jeu de calibration.

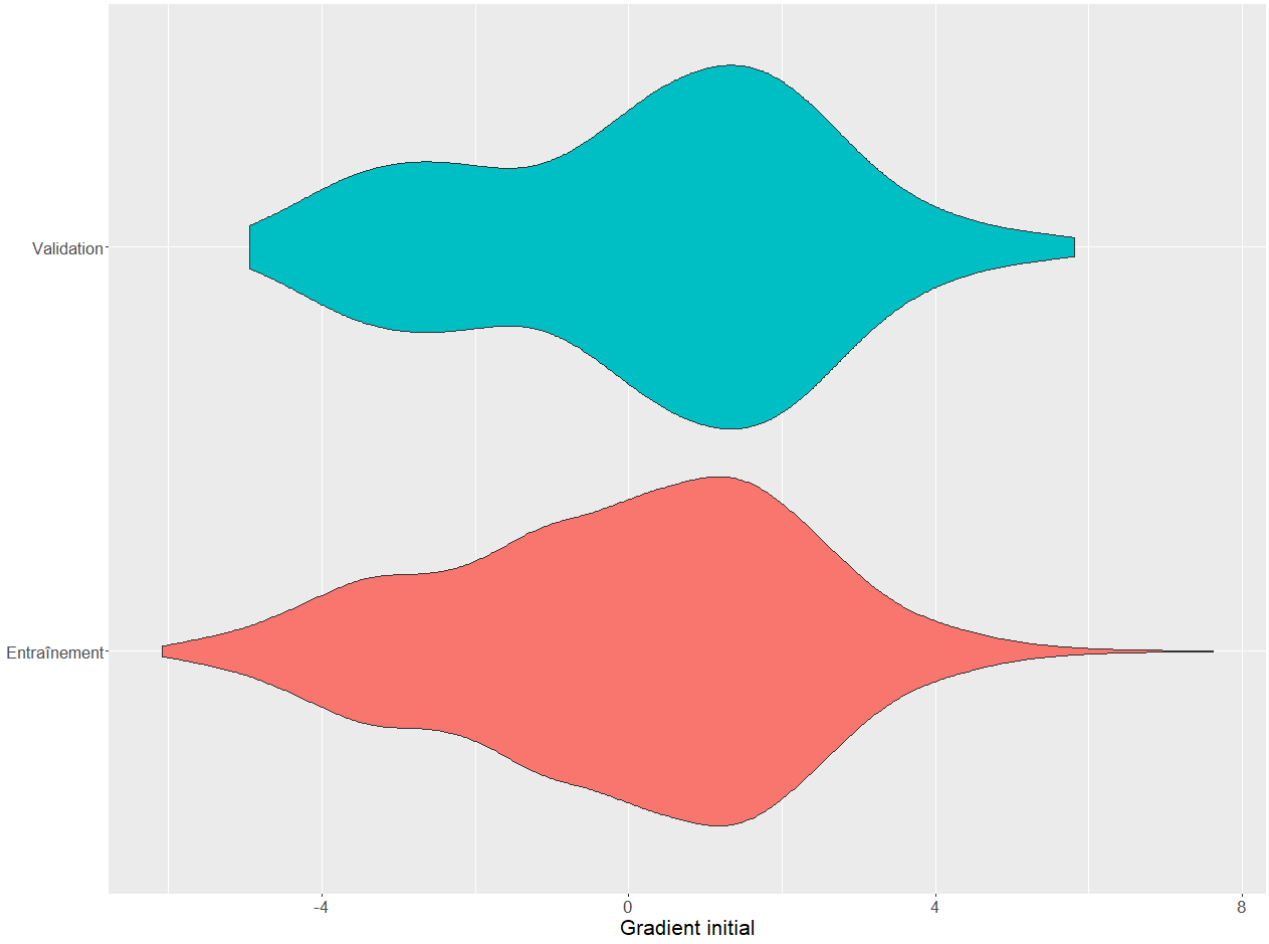


FIGURE 3.2 – Dispersion des sets d'entraînement et de validation.

Sur base du jeu d'entraînement, un modèle PLS est développé dans le but de prédire le gradient d'intensification sur base du gradient initial démontré par Dalcq (2020). Le nombre de variables latentes prises en compte dans le modèle est de cinq. Les variables explicatives utilisées dans ce modèle sont issues de prédictions à partir de l'information spectrale MIR. Par conséquent, ces variables présentent une incertitude plus ou moins grande selon le caractère étudié (l'Annexe 7.1 montre la précision). Ainsi, afin de détecter la présence d'observations potentiellement aberrantes dans le set d'entraînement, une analyse des résidus est effectuée. Chaque observation dont la contrainte sur le résidu explicitée à l'Equation 2.2 est satisfaite, a été retirée du jeu d'entraînement. Ces observations "potentiellement aberrantes" signifient qu'elles sont soit effectivement aberrantes soit qu'elles peuvent tout simplement se retrouver dans les marges des prédictions du modèle sans pour autant être aberrantes. Ce procédé est réalisé à trois reprises.

La figure en Annexe 7.2 illustre l'évolution du nombre de données retirées du jeu d'entraînement. Il y est observable que le nombre d'échantillons identifiés comme potentiellement aberrants diminue rapidement avec le nombre de tours de correction jusqu'à être nul après le troisième tour. Les observations identifiées comme potentiellement aberrantes proviennent de 11 exploitations différentes sur un laps de temps allant de 2010 à 2018. Trois de ces exploitations sortent du lot avec 4, 5 et 7 années les concernant. La majorité de ces observations a tendance à sous-estimer le gradient initial. Le nombre d'observations potentiellement aberrantes et supprimées est acceptable, il représente moins de 2,1 % du set initial.

Les performances de la PLS sur les données brutes (PLS1) et les données corrigées (PLS2) sont reprises dans la Table 3.3. Le nombre de variables latentes prises en compte dans le modèle PLS1 et

PLS2 est de 5 et 33, respectivement. Au travers des résultats affichés dans la Table 3.3, il apparaît clairement que les performances sont meilleures quand le modèle est développé sur les données nettoyées. En effet, comparé à PLS1, le R^2_V atteint 0,51 avec une $RMSE_V$ de 1,68. Il est intéressant de noter que les performances du modèle en validation sont proches de celles en validation croisée suggérant une bonne robustesse du modèle.

TABLE 3.3 – Performance de prédiction du gradient d'intensification estimé par Dalcq (2020) sur les données laitières (R^2 = coefficient de détermination, RMSE = racine carrée de l'erreur quadratique).

	N train	**Cross validation**		**Calibration**		**Validation**		**Tout**	
		R^2	**RMSE**	R^2	**RMSE**	R^2	**RMSE**	R^2	**RMSE**
PLS1	1434	0,38	1,78	0,43	1,72	0,49	1,68	0,44	1,71
PLS2	1405	0,51	1,56	0,63	1,37	0,51	1,68	0,49	1,66

La performance de prédiction du gradient d'intensification est faible avec un R^2 ne dépassant pas 0,70. Cela pourrait peut-être s'expliquer par une variabilité annuelle qui ne serait pas considérée. La Table 3.4 reprend le nombre d'exploitations, le R^2, le $RMSE$ et le gradient moyen prédit estimés chaque année.

TABLE 3.4 – Performances annuelles du modèle de prédiction de gradient

	N	R^2	$RMSE$	**Gradient moyen prédit**
2009	227	0,56	1,33	- 0,003
2010	225	0,43	1,67	- 0,034
2011	210	0,49	1,59	- 0,111
2012	203	0,47	1,57	- 0,223
2013	168	0,48	1,64	- 0,086
2014	171	0,55	1,65	0,017
2015	165	0,39	1,93	- 0,270
2016	140	0,57	1,69	- 0,090
2017	136	0,53	1,8	- 0,149
2018	131	0,44	1,91	- 0,274

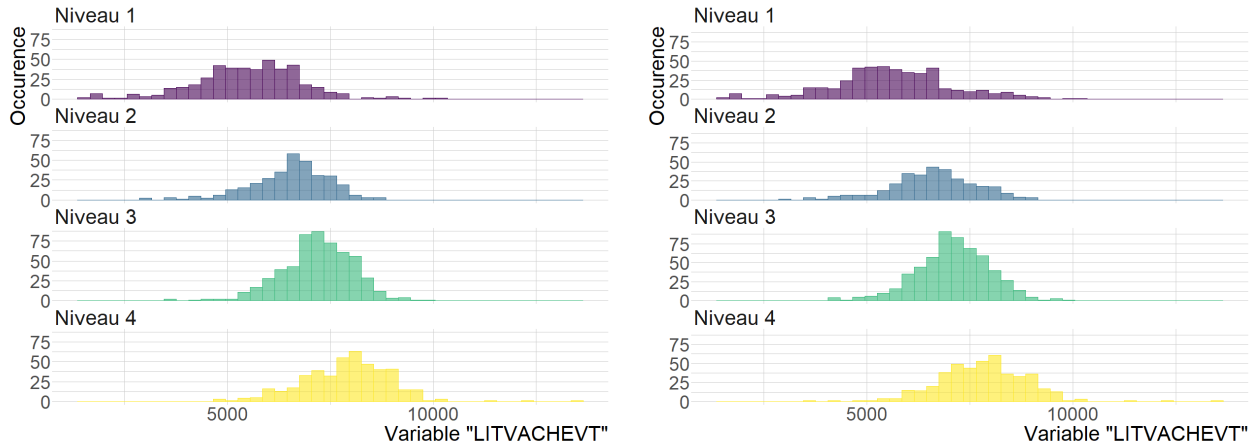
L'année la mieux prédite est 2009 avec un R^2 de 0,56 et une $RMSE$ de 1,33. L'année la moins bien prédite est 2015 avec un R^2 de 0,39 et une $RMSE$ de 1,93. Entre ces deux extrêmes, les performances sont relativement similaires entre les années et la performance de prédiction reste faible. La corrélation entre le $RMSE$ et le nombre d'exploitations par année est de - 0,73 suggérant un effet du nombre d'exploitation par année. Cet effet du nombre d'exploitations pourrait s'expliquer par un effet annuel sur la détermination du gradient d'intensification. La faible performance est attendue dans le sens où le gradient bouge en fonction des années un tout petit peu ainsi que la composition du lait.

Afin de mieux interpréter ces prédictions, le gradient prédit va être confronté à des variables qui sont connues pour être des éléments significatifs pointant l'intensification ou l'extensification d'une production laitière.

3.2.1 Analyse de la variable "*Lait*" (Production)

La production moyenne annuelle de lait par vache et par an peut se révéler être un indicateur d'intensification d'une exploitation. En effet, les exploitations dites intensives ont tendance à optimiser le rendement en production de lait par vache via, par exemple l'utilisation de concentrés importés. Les exploitations dites extensives quant à elles, tentent d'optimiser l'utilisation des ressources produites sur place et du pâturage (Chobtang et al., 2017). La variable "*Lait*" ici utilisée provient du set de données technico-économiques. Ainsi, 1 837 exploitations issues de 287 exploitations sont utilisées.

Les graphiques visibles à la Figure 3.3 montrent la tendance qu'a le gradient prédit de mettre en évidence les exploitations produisant plus de lait par vache comme intensive et vice versa. Les corrélations entre cette variable et les gradients initial/prédit viennent appuyer la tendance avec des valeurs de 0,67 et de 0,61 respectivement. Ainsi, même si la prédiction semble être d'une faible qualité, l'information véhiculée par ce biais a du sens. En effet, la production de lait de la majorité des exploitations tend à augmenter graduellement en fonction du gradient. Cependant, en comparant les répartitions entre le gradient prédit et le gradient initial (de référence), des comportements différents sont visibles. Cela peut être dû à plusieurs facteurs.



(a) Relatif au gradient initial.

(b) Relatif au gradient prédit.

FIGURE 3.3 – Évolution de la distribution de la production de lait établie selon quatre niveaux croissants d'intensification des gradients initial et prédit sur la base de données "account milk".

Premièrement, cette analyse est réalisée pour toutes races de vaches laitières confondues. Une différence de production peut donc apparaître entre deux exploitations qui utilisent en majorité deux races différentes et qui ont le même mode de fonctionnement. Par exemple, les vaches Holstein produisent le plus de lait et les Normandes en produisent le moins (Dillon et al., 2003). Cette omission de la race dans le modèle induit une erreur proportionnelle au poids que la variable "*Lait*" occupe dans celui-ci. Des exploitations à caractère plus extensif utilisant une race productrice se verront attribuer un gradient légèrement plus intensif qu'attendu. Dans le cas contraire, une exploitation à caractère plus intensif utilisant des races moins productives se verra attribuer un gradient légèrement plus extensif qu'attendu.

De plus, toutes les exploitations ne sont pas exclusivement vouées à la production laitière. Une race bien connue en Belgique, la Blanc Bleu Belge-mixte, permet l'association de la production de lait à la production de viande (Bastin et al., 2007). Certaines exploitations à caractère intensif auront tendance à se retrouver classées en tant que plus extensives vis-à-vis du fait qu'elles produisent moins de lait au profit de la viande. Dans le même sens, une exploitation extensive suivant le même type d'élevage se verra attribuer un gradient plus extensif qu'attendu.

Une analyse semblable peut être effectuée avec la quantité de lait prédite par la spectrométrie MIR inclue dans la base de données laitières (Annexe 7.2). Cependant, les performances du modèle de prédiction de cette variable ne sont pas optimales. Le choix d'analyse et d'interprétation s'est donc porté sur la variable issue de la base technico-économique.

3.2.2 Analyse de la variable "*BIO*" (système d'exploitation)

Un autre moyen indirect de mettre en lumière l'intensification d'une exploitation concerne le type d'exploitation. Le système d'exploitation est ici utilisé pour évaluer le comportement du gradient prédit. La variable "*BIO*" utilisée provient du set de données technico-économiques. L'ensemble des données du set de données lait n'a donc pas pu être utilisé. 1 776 exploitations issues de 285 exploitations sont utilisées. La Figure (3.4) illustre le comportement du gradient prédit selon le système d'exploitation caractérisant les exploitations agricoles. La Figure (3.5) illustre, quant à elle, le comportement du gradient initial vis-à-vis du système d'exploitation caractérisant les mêmes exploitations agricoles.

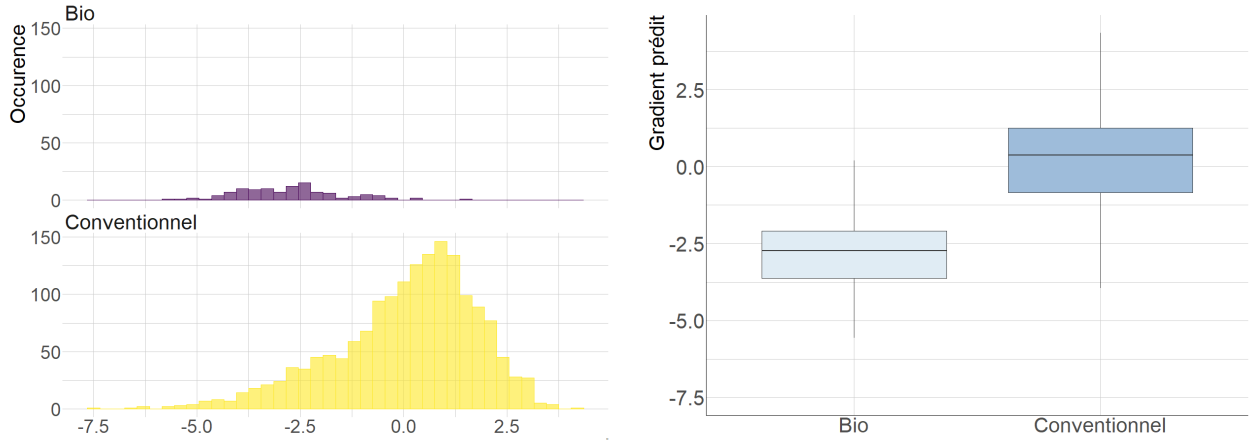


FIGURE 3.4 – Caractérisation des exploitations selon leur gradient **prédit** et leur système d'exploitation sur l'ensemble de la base de données laitières.

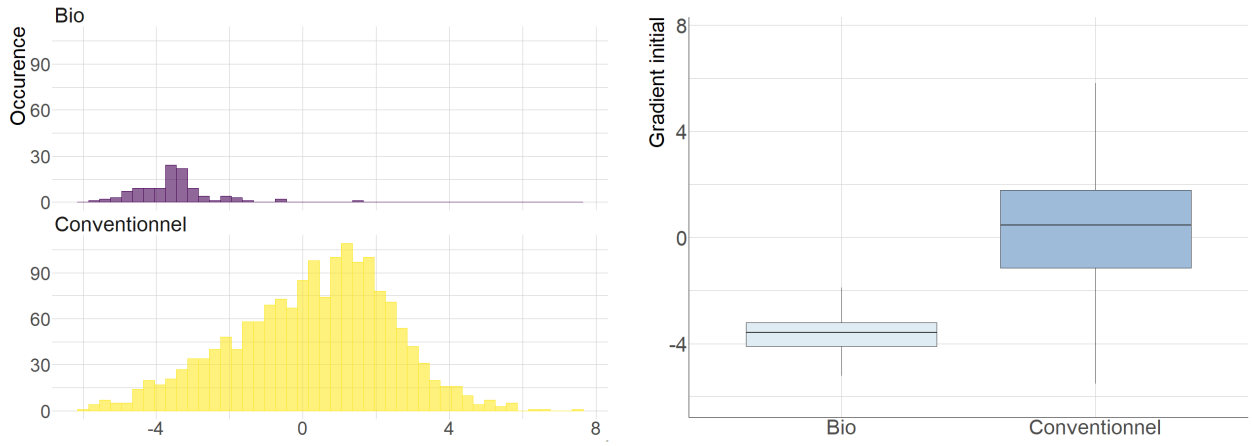


FIGURE 3.5 – Caractérisation des exploitations selon leur gradient **initial** et leur système d'exploitation sur l'ensemble de la base de données laitières.

Premièrement, les deux figures permettent de comparer l'occurrence des exploitations selon les gradients prédit et initial qui est leur attribuée. Il est intéressant de remarquer que le gradient prédit des exploitations conventionnelles tend à être moins intensif par rapport au gradient initial de Dalcq (2020). Les gradients prédits ne dépassent pas 4,3 alors que les gradients initiaux atteignent des valeurs de plus de 7,2. La différence est moindre mais présente pour les gradients très extensifs. Les gradients initiaux les plus extensifs atteignent des valeurs de près de -7,5 alors que les gradients prédits ne

dépassent pas -6,1. Le modèle a donc tendance à attribuer des gradients moins extrêmes et à recentrer l'ensemble des exploitations. Des différences sont visibles, telles que l'attribution de gradients plus intensifs à certaines exploitations bio. La séparation entre les exploitations en bio et en conventionnel selon le gradient est moins frappante avec l'utilisation du gradient prédit qu'avec le gradient initial.

Comme annoncé, les exploitations en bio adoptent pour la plupart un comportement extensif. La majorité des exploitations en conventionnel se voit singulier la majorité attribuer un gradient à tendance extensive. Une différence notable entre les deux distributions est la présence d'exploitations à tendance extensive dans le groupe des conventionnels. Cela s'explique par le fait qu'une exploitation conventionnelle n'a pas forcément un système d'exploitation opposé à celui des bios. Le système bio est un système labellisé qui impose certaines règles à ceux qui le suivent. Les exploitations en bio auront moins tendance à se retrouver dans les extrêmes. Les exploitations conventionnelles ont le choix du mode d'exploitation. Une graduation de l'intensification va être retrouvée dans cette catégorie. Cependant la majorité va tendre vers un mode d'exploitation plus intensif, voire très intensif. Deux caractéristiques des exploitations en bio sont leur diversification et l'utilisation de ressources produites sur site (Sautereau and Penvern, 2011).

3.2.3 Variable "*MG_Labo*" (Teneur en matière grasse)

Une tendance à la hausse de la matière grasse avec l'augmentation du gradient prédit peut être légèrement observée (Annexe 7.3). En effet, le pâturage tend à diminuer la teneur en MG du lait. La Figure 3.6 montre à son tour la tendance qu'a le gradient prédit à classer les exploitations extensives comme ayant une teneur en AG linolénique la plus élevée et vice versa. La corrélation entre cette variable et le gradient prédit vient appuyer la tendance avec une valeur de -0,44. Le gradient respecte les tendances attendues au niveau des différentes teneurs en acides gras composant le lait selon le degré d'intensification de l'exploitation.

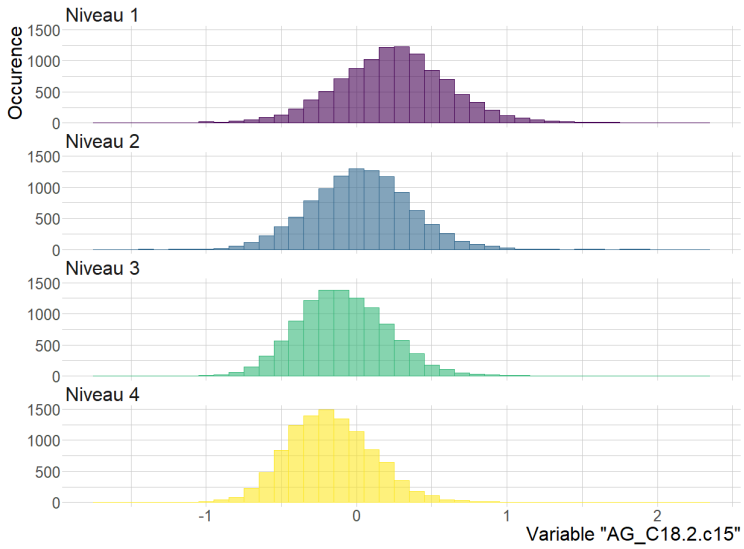


FIGURE 3.6 – Évolution de la distribution des teneurs en acide linolénique selon quatre niveaux croissants d'intensification sur l'ensemble de la base de données lait normalisée.

Quelques précisions doivent néanmoins être apportées. De multiples facteurs n'ont pas été pris en compte dans le modèle de prédiction. Cela induit le mauvais classement de certaines exploitations.

Pour commencer, l'état de santé des vaches dans les exploitations n'est pas connu. Le facteur santé peut affecter la production et les composantes du lait. La mammite (principale maladie étudiée) va,

par exemple, diminuer la teneur en matière grasse (Stokes et al., 2000). L'effet reste toutefois très limité. Toutes les vaches d'une exploitation ne tombent pas malade en même temps. De plus, les données utilisées sont moyennées à l'année et cela atténue encore plus l'effet.

Ensuite, le stade de lactation n'est pas connu non plus. L'effet de ce facteur se caractérise par une augmentation des AG à courtes chaînes dans les cinq premiers mois. Les AG à chaînes moyennes ont tendance à diminuer vers la fin de lactation. Les AG à longues chaînes ont tous tendance à augmenter en début de lactation. La majorité diminue ensuite en milieu de lactation pour réaugmenter en fin de lactation. L'acide linolénique continue d'augmenter progressivement durant toute la lactation (Karijord et al., 1982).

De plus, le facteur race n'a été pris en compte. La composition en AG du lait varie en fonction de la race. Comme le démontrent Lawless et al. (1999) et Kelsey et al. (2003), une différence claire existe entre les races dans la composition en AG du lait. Cela peut avoir comme conséquence de stigmatiser une exploitation comme extensive, composée majoritairement de Jersey, ou comme intensive, composée d'Holstein. La première situation a du sens. Il existe très peu d'exploitations intensives utilisant ce type de race. La deuxième situation est plus discutable. L'Holstein est connue pour sa rentabilité. Il est certain de retrouver cette race dans les exploitations intensives. Cependant, l'utilisation de cette race n'est pas gage d'intensification. Des exploitations se verront attribuer un gradient à tendance intensive sans pour autant l'être vis-à-vis de ces variables.

3.3 Clustering

Le clustering hiérarchisé est effectué sur l'ensemble du set de données laitières nettoyées. Le dendrogramme ainsi que les groupes (les clusters) sont visibles à la Figure 3.7. Le choix du nombre de clusters s'est porté sur trois par rapport à leur hauteur de scission. Le cluster 1 se compose de 10 385 observations, le cluster 2 de 14 004 observations et le cluster 3 se compose de 17 721 observations. Ces dernières sont visibles en Annexe 7.4.

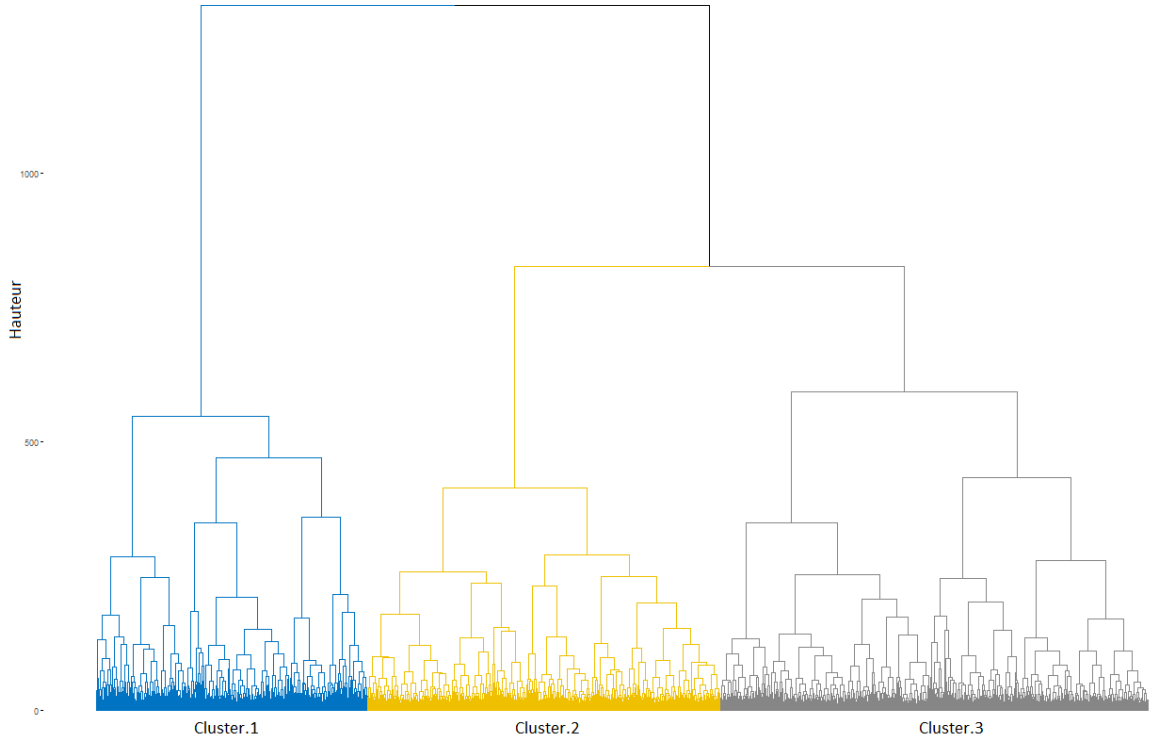


FIGURE 3.7 – Représentation du dendrogramme issu du clustering.

Une visualisation de la pertinence de l'utilisation du clustering comme moyen de classement des exploitations selon leur intensivité est nécessaire. Les exploitations présentes à la fois dans le set de données technico-économiques et dans le set de données lait sont isolées. Cela permet d'observer le comportement du gradient prédit selon les groupes auxquels ces exploitations sont attribuées (Figure 3.8).

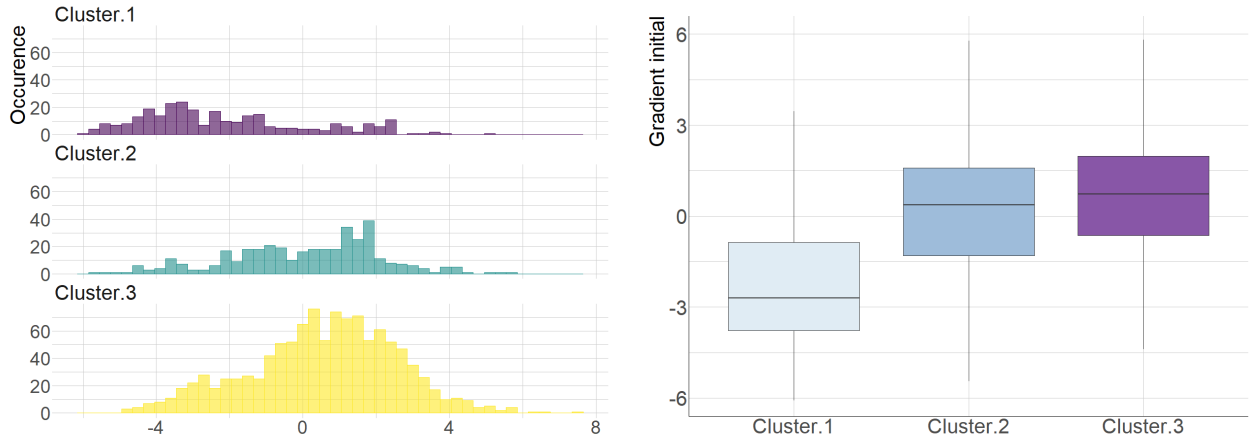


FIGURE 3.8 – Évolution de la distribution du gradient initial selon les clusters (issus du clustering sur l'ensemble de la base de données laitières) sur la base de données laitières fusionnée à la base de données technico-économiques.

Une tendance de classement des exploitations est visible. Les exploitations présentes dans le cluster 1 tendent, en majorité, vers un petit gradient suggérant une exploitation à tendance extensive. Les exploitations du cluster 3 tendent vers un gradient plus grand indiquant une exploitation à tendance intensive. Une grande scission apparait initialement entre le cluster 1 et les clusters 2 et 3. Les cluster 2 et 3 ont donc plus de similarités. C'est ce qu'on observe aussi avec les valeurs de gradient prédit (Figure 3.8).

3.3.1 Modèle de prédiction

Afin de prédire à large échelle le cluster d'appartenance, un modèle PLS-DA est développé sur base des mêmes variables utilisées pour la prédiction du gradient. Le jeu d'entraînement contient 3 719 exploitations pour un total de 35 787 observations. Le jeu de validation contient 656 exploitations pour un total de 6 323 observations. Le nombre de variables latentes utilisées par le modèle est de 107. Les performances de prédiction du clustering sont visibles à la Table 3.5.

TABLE 3.5 – Performances du modèle de prédiction dans l'attribution des clusters.

	N	**Justesse**	**Kappa**
Cross-validation	35 787	0,80	0,69
Calibration	35 787	0,80	0,70
Validation	6 323	0,80	0,69
Tout le set	42 110	0,80	0,70

Les performances en termes de justesse et de Kappa sont bonnes et très semblables, que ce soit en validation croisées, en calibration ou en validation. Le modèle a une bonne aptitude à prédire l'apparetenance des observations aux trois clusters. La matrice de confusion entre les clusters prédits et observés par le clustering est présentée à la Table 3.6. Lorsque les données sont incorrectement prédites, elles sont prédites en majorité dans le cluster 2 qui agit comme un tampon entre les clusters 1 et 3. Le cluster 2 est le moins bien prédit.

TABLE 3.6 – Matrice de confusion entre les clusters prédits par la PLSDA et calculés par le clustering sur l'ensemble du set de données lait.

Prédiction/référence	**Cluster.1**	**Cluster.2**	**Cluster.3**
Cluster.1	8 870	1 037	478
Cluster.2	1 414	10 465	2 125
Cluster.3	574	2 632	14 515

3.3.2 Calcul des corrélations

Les corrélations entre les probabilités d'appartenance à un cluster et le gradient initial ou prédit sont calculées. Ces corrélations permettent de mesurer l'intensité de la relation entre les deux variables. Dans un premier temps, le gradient initial est utilisé. Les exploitations présentes au sein des sets de données laitières et technico-économiques sont donc sélectionnées. Dans un second temps, l'ensemble des exploitations du set de données est employé via l'utilisation du gradient prédit. Les résultats sont visibles à la Table 3.7.

TABLE 3.7 – Corrélation entre la probabilité d'appartenance à un cluster et le gradient initial/prédit.

	N	**Cluster 1**	**Cluster 2**	**Cluster 3**
Gradient initial	1 776	- 0,51	0,01	0,37
Gradient de prédiction	1 776	- 0,74	0,05	0,52
Gradient de prédiction	42 110	- 0,63	0,26	0,39

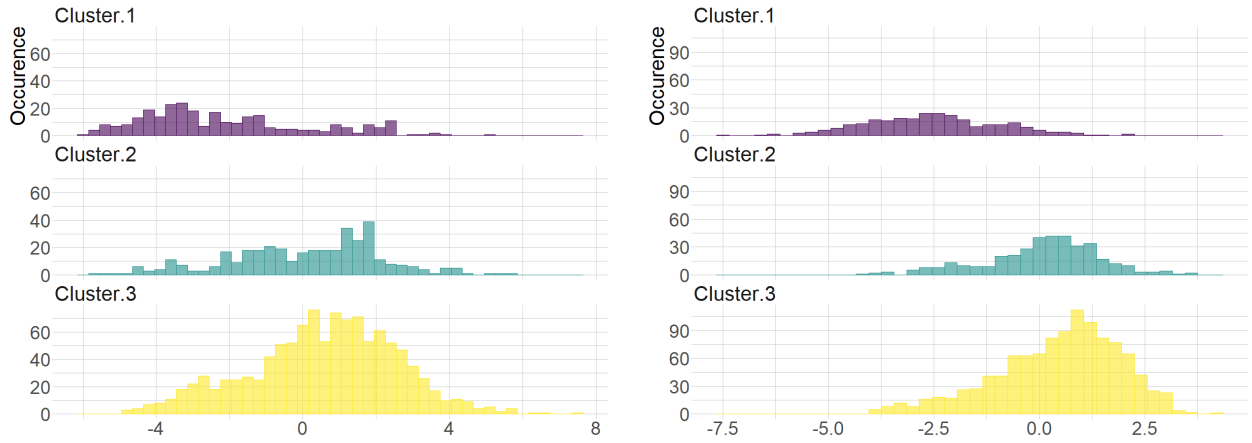
Gradient initial Premièrement, en ce qui concerne le gradient initial sur base des 1 776 observations, les résultats sont intéressants. Une corrélation négative est observable pour le cluster 1. La corrélation négative indique une relation inversément proportionnelle entre les probabilités d'appartenance au cluster 1 et le gradient. Cela indique la tendance qu'a le modèle à prédire une probabilité élevée d'appartenance au cluster 1 pour un gradient qui est faible et vice versa. Ensuite, la corrélation nulle du second cluster n'est pas étonnante. En effet, cela signifie que le gradient des observations avec une haute probabilité d'appartenance peut tendre à la fois vers une faible valeur comme vers une haute valeur de gradient. Il en va de même lorsque la probabilité d'appartenance est faible. Une interprétation possible est de partir du principe que le cluster 2 représente les observations semi-intensives. Il n'y a pas de lien entre la probabilité d'appartenance au cluster 2 et le gradient puisque les observations avec une haute probabilité d'appartenance à ce cluster peuvent aussi bien être extensives qu'intensives. Les exploitations avec une faible probabilité d'appartenance peuvent aussi bien être très intensives que très extensives. Enfin, la corrélation entre la probabilité d'appartenance au cluster est positive. Cela indique une relation proportionnelle entre les probabilités d'appartenance au cluster 3 et le gradient. Le modèle a tendance à attribuer une probabilité d'appartenance au cluster 3 élevée lorsqu'elles sont caractérisées par un gradient élevé (intensif).

Gradient prédit Deuxièmement, en ce qui concerne le gradient prédit, deux sets de données ont été utilisés : le premier comprend les mêmes observations que pour le gradient initial et le second comprend l'entièreté des données laitières. Dans un premier temps, les corrélations entre gradient initial et de prédiction dont les observations sont communes sont comparables. Il est observable que la relation entre le gradient prédit et les probabilités d'appartenance aux clusters 1 et 3 sont renforcées par rapport au gradient initial. Cette augmentation est essentiellement due au fait le modèle de prédiction de gradient que le clustering et le modèle de prédiction des clusters utilisent les mêmes variables de prédictions issues du même set de données. Dans un second temps, une perte de corrélation est observable pour les clusters 1 et 3 pour le gradient de prédiction. Les corrélations passent respectivement de -0,74 à -0,63 et de 0,52 à 0,39. Il est à supposer que si la prédiction du gradient a du mal à extrapoler en dehors de son domaine initial alors, les prédictions sont moins bonnes et cela peut affecter la relation qui existe entre elles et les probabilités d'appartenance aux clusters 1 et 3. Parallèlement, la corrélation

du cluster 2 passe de 0,05 à 0,26. L'équation PLS qui a été entraînée sur des fermes plutôt extensives a potentiellement tendance à associer une valeur trop élevée aux *extensives-proche-intensives* et aux *intensives-proche-extensives*. Cette équation peut donner ainsi une valeur plus intensive qu'attendue. Dès lors, le cluster 2, regroupant les faiblement intensives et extensives se voit maintenant inclure des fermes plus intensives que prévu à cause de la prédiction biaisée vers l'intensivité.

3.3.3 Comportement du gradient prédit et du gradient initial selon le clustering

Le clustering permet dans le même temps d'observer le comportement du gradient prédit par rapport au gradient initial selon l'attribution des observations dans les trois clusters. La Figure 3.9 met côte à côte l'occurrence des observations selon leur gradient (initial et prédit) et selon le cluster auquel ces observations appartiennent.



(a) Relatif au gradient initial.

(b) Relatif au gradient prédit.

FIGURE 3.9 – Évolution de la distribution des gradients selon les clusters (issus du clustering sur l'ensemble de la base de données laitières) sur la base de données laitières fusionnée à la base de données technico-économiques.

Il a été vu que le modèle de prédiction du gradient avait tendance à centrer les observations. Cela est visible au niveau de l'attribution de gradients moins extensifs pour les observations extensives et moins intensifs pour des exploitations intensives par rapport au gradient initial. L'effet est également bien visible dans ce cas-ci. Cependant, il est intéressant de voir que le gradient prédit réagit mieux au clustering que le gradient initial. Les observations présentes dans le cluster 1 ont plus tendance à se voir attribuer un gradient prédit extensif. Les observations présentes dans les clusters 2 et 3 se voient attribuer un gradient prédit intensif. Le gradient prédit fait moins de différences entre les clusters 2 et 3 que le gradient initial. Ce dernier, quant à lui, réagit moins bien au clustering et cela est visible au niveau des trois clusters. Une différence est observable entre les trois clusters quant à leur "degré d'intensification" mais est moins marquée. Le gradient initial a tendance à différencier les clusters 2 et 3, ce que le gradient prédit ne faisait pas. Le gradient initial est donc plus sensible à la classification du clustering sans pour autant y réagir mieux.

L'analyse relative au gradient prédit vis-à-vis du clustering peut être transposée non plus sur l'utilisation des bases de données laitières et technico-économiques fusionnées mais bien sur l'ensemble de la base de données laitières. La Figure relative à cela se trouve en Annexe 7.6.

Chapitre 4

CONCLUSION

Deux approches ont été vues dans le cadre de ce travail pour quantifier le degré d'intensification caractérisant une exploitation. Ces deux approches ont utilisé les variables prédites issues du set de données lait. La première approche est basée sur l'utilisation du gradient d'intensification initialement développé par Dalcq (2020) et la seconde sur un clustering, la re-prédiction du clustering et l'utilisation des probabilités d'appartenance de chaque observation à chaque cluster.

La première approche est une méthode de régression linéaire de type PLS utilisant les variables prédites issues du set de données laitières et le gradient initial. Cette méthode est intéressante car elle est supervisée en utilisant comme variable cible le gradient d'intensification développé par Dalcq (2020). Les performances du modèle sont modérées. Des inconvénients subsistent autour de cette modélisation. En effet, il y a un manque de représentativité car les observations étaient issues majoritairement d'une même région géographique, à savoir, la région herbagère liégeoise. L'ensemble des exploitations n'est donc pas représentatif de la diversité du gradient d'intensification rencontrée en Wallonie. Le modèle extrapole donc dans le cas d'exploitations présentant des variabilités non comprises dans le modèle pouvant engendrer un biais dans la prédiction.

La deuxième approche est une méthode non linéaire basée sur un regroupement utilisant, elle aussi, les variables prédites issues du set de données laitières. Une régression discriminante de type PLS-DA est ensuite utilisée pour reprédire les clusters sur base des variables issues du set de données lait et du clustering. Les probabilités d'appartenance de chaque observation à chaque cluster sont conservées. Cette méthode a l'avantage d'utiliser l'ensemble du set de données laitières. Cela évite le manque de représentativité. De plus, les performances de reprédictions des clusters sont satisfaisantes. La corrélation des coefficients d'appartenance au cluster 1 et le gradient initial indique une tendance qu'a ce cluster à attribuer une valeur de gradient élevée aux observations extensives et invérsement, ce qui était souhaité.

Les deux approches apportent donc leur lot d'avantages et d'inconvénients. Il est donc difficile de conclure quant à l'utilisation d'un modèle plutôt qu'un autre. Diverses pistes d'amélioration peuvent être envisagées et testées à l'avenir ainsi qu'un processus de validation pour les deux modèles. Cependant, il est important de mentionner que les variables issues du lait sont capables de mettre en partie en lumière le degré d'intensification du comportement d'un agriculteur ce qui était l'objectif poursuivi.

Chapitre 5

PERSPECTIVES

En termes de perspectives, différents niveaux sont à prendre en compte. Premièrement, les variables issues du set de données laitières pourraient être épurées selon la pertinence qu'elles occupent dans le modèle. En effet, certaines variables n'apportent que très peu d'informations utiles au modèle. De plus, d'autres variables pourraient être écartées selon les performances de leur modèle de prédiction. En effet, certaines variables pertinentes contribuent à l'apport de bruits dans le modèle de prédiction du gradient via l'utilisation d'observations elles-mêmes mal prédites par le modèle de prédiction de ladite variable. Les performances du modèle de prédiction de gradient en sont directement impactées car ces observations erronées avec une plus ou moins grande intensité sont prises en compte de manière égale à leurs homologues bien prédites. Cela est valable dans l'établissement du clustering et du modèle de prédiction des clusters également. Au-delà des performances des modèles de prédiction des variables, certaines informations supplémentaires pourraient s'avérer nécessaires. Les données de lait de "tank" ne tiennent pas compte de la race de vache dont est issu le lait de l'exploitation. La race peut être un facteur déterminant dans l'utilisation de certaines variables pertinentes pour le modèle de prédiction du gradient ainsi que pour le clustering et modèle qui en découle.

Deuxièmement, en ce qui concerne le modèle de prédiction du gradient, il pourrait s'avérer intéressant, dans un premier temps, d'augmenter le nombre d'exploitations pour la base de données technico-économiques. Le gradient initial est calculé sur base des 15 variables présentes dans ce set de données. Or, le nombre d'exploitations participant à l'établissement de ce set de données ne cesse de décroître au fur et à mesure du temps. Au-delà d'augmenter le nombre d'exploitations, il s'avérerait pertinent, dans un second temps, d'augmenter la diversité des exploitations. Au plus le modèle serait confronté à de la diversité en termes de gradient d'intensification, au plus le risque d'extrapolation est réduit et au mieux les observations seront prédites. Dans les deux cas, le modèle aurait plus de données pour prédire le gradient.

Ensuite, en ce qui concerne le clustering, il serait intéressant de l'étudier plus en profondeur. Le clustering présenté dans ce travail n'a été étudié que par rapport à sa capacité à catégoriser les observations selon leur degré d'intensification. Cela a été réalisé sur trois clusters. Il aurait été intéressant d'imposer un nombre de clusters plus élevé et de visionner le comportement d'autres variables caractéristiques, par exemple de l'état de santé du troupeau ou de la race des vaches. Cela n'est pas directement lié au but recherché ici mais aurait apporté des informations supplémentaires sur l'influence qu'ont certaines variables par rapport à d'autres et sur l'intérêt réel du clustering. Le fait d'imposer un nombre plus grand de clusters au clustering permettrait aussi d'observer le comportement du gradient initial. Si le clustering permet vraiment le classement des observations selon leur gradient d'intensification, le classement devrait se détailler avec le nombre de clusters. Il aurait aussi été pertinent d'augmenter le nombre d'exploitations issues de la base de données technico-économiques dans ce cas-ci dans le même but qu'expliquer ci-avant. Cela aurait permis de mettre en exergue une tendance plus forte de classement mais qui est actuellement masquée par le nombre d'observations peu élevé.

L'intérêt principal de l'outil est l'utilisation de données rapidement et facilement disponibles, à savoir le set de données laitières. Une dimension différente de ce dont il a été question jusqu'ici (relatif aux données ou aux modèles) pourrait s'avérer intéressante. Les exploitations extensives se caractérisent par l'utilisation d'alimentation produite sur l'exploitation ou du moins localement. Elles se caractérisent aussi par l'optimisation de l'utilisation d'alimentation herbeuse en bonne saison. Les effets météorologiques pourraient donc être pris en compte. En effet, la météo va impacter radicalement les cultures localement. Les agriculteurs pourraient se tourner vers d'autres sources d'alimentation à connotation plus intensives. Ces exploitations se verraient donc attribuer un gradient plus intensif qu'attendu. Les exploitations plus intensives ont recours à de l'alimentation moins sujette à ce type d'effet météorologique local. Dans cette voie, d'autres facteurs macroscopiques pourraient aussi être pris en compte tels que le climat politique et économique. L'année 2022 en est un exemple au niveau de la situation politique en Ukraine ou de la situation économique générale. Ces climats de "tension" entravent la disponibilité de produits (fertilisants ou concentrés par exemple) auxquels des exploitations à tendance intensive auraient eu recours. La question est de savoir si l'objectif est de prédire un gradient objectif ou réel. Cela pourrait expliquer aussi en partie pourquoi la performance de prédiction est si faible.

Enfin, un processus de validation est nécessaire via l'utilisation de nouvelles données technico-économiques de nouvelles exploitations. Il serait intéressant de sélectionner aléatoirement des exploitants d'accord de participer.

BIBLIOGRAPHIE

Bastin, C., Mayeres, P., Bertozzi, C., Michaux, C., and Gengler, N. (2007). Produire de la viande et du lait avec la blanc-bleu belge de type mixte.

Brundtland, G. H. et al. (1987). Notre avenir à tous (rapport brundtland). *Oslo : Nations Unies. Consulté le*, 4(16) :2015.

Chilliard, Y., Ferlay, A., Mansbridge, R. M., and Doreau, M. (2000). Ruminant milk fat plasticity :nutritional control of saturated, polyunsaturated, trans and conjugated fatty acids. *Annales de Zootechnie*, 49(3) :181–205.

Chobtang, J., Ledgard, S. F., McLaren, S. J., and Donaghy, D. J. (2017). Life cycle environmental impacts of high and low intensification pasture-based milk production systems : A case study of the waikato region, new zealand. *Journal of Cleaner Production*, 140 :664–674.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1) :37–46.

Cuvelier, C., Hornick, J.-L., Beckers, Y., Froidmont, E., Istasse, L., and Dufrasne, I. (2020). L'alimentation de la vache laitiere physiologie et besoins. page 67.

Dalcq, A.-C. (2020). Caractérisation de la diversité des stratégies conçues par les producteurs laitiers wallons : déterminants socio-économiques et dynamique d'adaptations face aux enjeux passés, présents et futurs.

De Marchi, M., Bonfatti, V., Cecchinato, A., Di Martino, G., and Carnier, P. (2009). Prediction of protein composition of individual cow milk using mid-infrared spectroscopy. *Italian Journal of Animal Science*, 8(sup2) :399–401.

Demarquilly, C., Jarrige, R., and Boisseau, J.-M. (1964). Valeur alimentaire de l'herbe des prairies temporaires aux stades d'exploitation pour le paturage. i. - composition chimique et digestibilitÉ. *Annales de zootechnie*, 13(4) :301–339.

Dewhurst and King (1998). Effects of extended wilting, shading and chemical additives on the fatty acids in laboratory grass silages. *Grass and Forage Science*, 53(3) :219–224.

Dillon, P., Buckley, F., O'Connor, P., Hegarty, D., and Rath, M. (2003). A comparison of different dairy cow breeds on a seasonal grass-based system of milk production : 1. milk production, live weight, body condition score and dm intake. *Livestock Production Science*, 83(1) :21–33.

Doreau, M. and Chilliard, Y. (1997). Digestion and metabolism of dietary fat in farm animals. *British Journal of Nutrition*, 78(1) :S15–S35.

Dowle, M. and Srinivasan, A. (2021). *data.table : Extension of 'data.frame'*. R package version 1.14.2.

Elgersma, A. (2015). Grazing increases the unsaturated fatty acid concentration of milk from grass-fed cows : A review of the contributing factors, challenges and future perspectives. *European Journal of Lipid Science and Technology*, 117(9) :1345–1369.

Frelich, J., Šlachta, M., Hanuš, O., Špička, J., Samková, E., Węglarz, A., and Zapletal, P. (2012). Seasonal variation in fatty acid composition of cow milk in relation to the feeding system*. *Animal science papers and reports*, 30 :219–229.

García-Martínez, A., Olaizola, A., and Bernués, A. (2009). Trajectories of evolution and drivers of change in european mountain cattle farming systems. *Animal*, 3(1) :152–165.

GxABT & LIST (2018). Research project (PDR) - SimBa. Technical report.

Hammami, H., Vandenplas, J., Vanrobays, M. L., Rekik, B., Bastin, C., and Gengler, N. (2015). Genetic analysis of heat stress effects on yield traits, udder health, and fatty acids of walloon holstein cows. *Journal of Dairy Science*, 98(7) :4956–4968.

Hoden, A. and Coulon, J.-B. (1991). Maîtrise de la composition du lait : influence des facteurs nutritionnels sur la quantité et les taux de matières grasses et protéiques (1). *INRA Productions Animales*, 4(5) :361–367.

ICAR (2017). Dairy cattle milk recording working group.

Jenkins, T. C. (1993). Lipid metabolism in the rumen. *Journal of Dairy Science*, 76(12) :3851–3863.

Karijord, O., Standal, N., and Syrstad, O. (1982). Sources of variation in composition of milk fat. *Zeitschrift fuer Tierzuechtung und Zuechtungsbiologie (Germany, F.R.)*.

Kassambara, A. and Mundt, F. (2020). *factoextra : Extract and Visualize the Results of Multivariate Data Analyses*. R package version 1.0.7.

Kelly, M. L., Kolver, E. S., Bauman, D. E., Van Amburgh, M. E., and Muller, L. D. (1998). Effect of intake of pasture on concentrations of conjugated linoleic acid in milk of lactating cows. *Journal of Dairy Science*, 81(6) :1630–1636.

Kelsey, J. A., Corl, B. A., Collier, R. J., and Bauman, D. E. (2003). The effect of breed, parity, and stage of lactation on conjugated linoleic acid (cla) in milk fat from dairy cows1. *Journal of Dairy Science*, 86(8) :2588–2597.

Kuhn, M. (2022). *caret : Classification and Regression Training*. R package version 6.0-92.

Landau, S., Leese, M., Stahl, D., and Everitt, B. S. (2011). *Cluster Analysis*. John Wiley Sons.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1) :159–174.

Lawless, F., Stanton, C., L'Escop, P., Devery, R., Dillon, P., and Murphy, J. J. (1999). Influence of breed on bovine milk cis-9, trans-11-conjugated linoleic acid content. *Livestock Production Science*, 62(1) :43–49.

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR : A package for multivariate analysis. *Journal of Statistical Software*, 25(1) :1–18.

Liland, K. H., Mevik, B.-H., and Wehrens, R. (2021). *pls : Partial Least Squares and Principal Component Regression*. R package version 2.8-0.

Marvuglia, A., Bayram, A., Baustert, P., Gutiérrez, T. N., and Igos, E. (2022). Agent-based modelling to simulate farmers' sustainable decisions : Farmers' interaction and resulting green consciousness evolution. *Journal of Cleaner Production*, 332 :129847.

Marvuglia, A., Rege, S., Navarrete Gutiérrez, T., Vanni, L., Stilmant, D., and Benetto, E. (2017). A return on experience from the application of agent-based simulations coupled with life cycle assessment to model agricultural processes. *Journal of Cleaner Production*, 142 :1539–1551.

Mazoyer, M. and Roudart, L. (2006). *A History of World Agriculture : From the Neolithic Age to the Current Crisis*. NYU Press. Google-Books-ID : vt0VCgAAQBAJ.

McDonald, P., Edwards, R., Greenhalgh, J., Morgan, C., Sinclair, L., and Wilkinson, R. (2010). *Animal nutrition seventh edition*. Pearson Education Limited Harlow, United Kingdom.

Morand-Fehr, P. and Tran, G. (2001). La fraction lipidique des aliments et les corps gras utilisés en alimentation animale.

Ollivier, G. and Guyomard, H. (2013). *Les performances sociales de l'Agriculture Biologique*, volume 1, page 106–130.

Paccard, P., Capitain, M., and Farruggia, A. (2003). Autonomie alimentaire des élevages bovins laitiers. page 4.

Penati, C., Berentsen, P. B. M., Tamburini, A., Sandrucci, A., and de Boer, I. J. M. (2011). Effect of abandoning highland grazing on nutrient balances and economic performance of italian alpine dairy farms. *Livestock Science*, 139(1) :142–149.

R Core Team (2016). R : A language and environment for statistical computing.

Riuzzi, G., Davis, H., Lanza, I., Butler, G., Contiero, B., Gottardo, F., and Segato, S. (2021). Multivariate modelling of milk fatty acid profile to discriminate the forages in dairy cows' ration. *Scientific Reports*, 11(11) :23201.

Rouille, B. and Montourcy, M. (2010). Influence de quelques systèmes d'alimentation sur la composition en acides gras du lait de vache en france. *Institut de l'Elevage, CNIEL, DTEQ, Collections Résultats, ISSN, France. 33p.*

Rouille, B., Peyraud, J.-L., Hurtaud, C., and Brunschwig, P. (2011). La composition en acides gras du lait de vache : les possibilités d'action par l'alimentation. *La composition en acides gras du lait de vache : les possibilités d'action par l'alimentation, Institut de l'Elevage(2011) - 978-2-36343-130-1.*

Rutten, M. J. M., Bovenhuis, H., Hettinga, K. A., van Valenberg, H. J. F., and van Arendonk, J. a. M. (2009). Predicting bovine milk fat composition using infrared spectroscopy based on milk samples collected in winter and summer. *Journal of Dairy Science*, 92(12) :6202–6209.

Ryden, J. C., Ball, P. R., and Garwood, E. A. (1984). Nitrate leaching from grassland. *Nature*, 311(59815981) :50–53.

Sautereau, N. and Penvern, S. (2011). Pluralité de l'ab & recherche-formation-développement. construire les systèmes agricoles et alimentaires de demain. page 114.

Schlegel, P., Wyss, U., Arrigo, Y., and Hess, H. D. (2016). Mineral concentrations of fresh herbage from mixed grassland as influenced by botanical composition, harvest time and growth stage. *Animal Feed Science and Technology*, 219 :226–233.

Soyeurt, H., Bastin, C., Colinet, F. G., Arnould, V. M. R., Berry, D., Wall, E., Dehareng, F., Nguyen, H. N., Dardenne, P., Schefers, J., Vandenplas, J., Weigel, K., Coffey, M. P., Theron, L., Detilleux, J., Reding, E., Gengler, N., and McParland, S. (2012). Mid-infrared prediction of lactoferrin content in bovine milk : potential indicator of mastitis. Accepted : 2013-05-08T14 :41 :57Z.

Soyeurt, H., Bruwier, D., Romnee, J.-M., Gengler, N., Bertozzi, C., Veselko, D., and Dardenne, P. (2009). Potential estimation of major mineral contents in cow milk using mid-infrared spectrometry. *Journal of Dairy Science*, 92(6) :2444–2454.

Soyeurt, H., Dardenne, P., Dehareng, F., Lognay, G., Veselko, D., Marlier, M., Bertozzi, C., Mayeres, P., and Gengler, N. (2006). Estimating fatty acid content in cow milk using mid-infrared spectrometry. *Journal of Dairy Science*, 89(9) :3690–3695.

Stokes, S. R., Jordan, E. R., Looper, M., and Waldner, D. (2000). Managing milk composition : Normal sources of variation. Accepted : 2009-07-20T22 :25 :43Z.

Tracy, B. and Sanderson, M. (2004). Productivity and stability relationships in mowed pasture communities of varying species composition. *Crop Science*, 44.

Waghorn, G. and Clark, D. (2004). Feeding value of pastures for ruminants. *New Zealand Veterinary Journal*, 52(6) :320–331.

Wauchope, R. D. (1978). The pesticide content of surface water draining from agricultural fields—a review. *Journal of Environmental Quality*, 7(4) :459–472.

Wickham, H. (2016). *ggplot2 : Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43) :1686.

Wickham, H., François, R., Henry, L., and Müller, K. (2022). *dplyr : A Grammar of Data Manipulation*. R package version 1.0.9.

Wold, S., Eriksson, L., Trygg, J., and Kettaneh, N. (2004). The pls method – partial least squares projections to latent structures – and its applications in industrial rdp (research, development, and production). page 44.

Zealand, N. (2004). Parliamentary commissioner for the environment. farming trends, in growing for good : Intensive farming, sustainability and new zealand's environment. *Wellington : Parliamentary Commissioner for the Environment*, pages 31–53.

Zhang, L., Li, C., Dehareng, F., Grelet, C., Colinet, F., Gengler, N., Brostaux, Y., and Soyeurt, H. (2021). Appropriate data quality checks improve the reliability of values predicted from milk mid-infrared spectra. *Animals*, 11(22) :533.

Chapitre 7

ANNEXES

7.1 Annexe 1

TABLE 7.1 – Variables prédites et R^2 de prédiction correspondant disponibles à partir du spectre MIR du lait.

Variable	**Unité**	**N**	**R^2**	**Erreur standard**
Taux en matière grasse	g/dl lait	1799	0.9999	0.0086
Production laitière journalière	kg/jour	457	0.6888	3.4797
Lactoferrine	mg/l lait	2442	0.7102	50.5539
Acidité titrable	degré Dornic	459	0.8028	0.8003
Taux en matière azotée protéique	g/l lait	4305	0.9976	0.2004
Caséines alpha s1 (Levicek)	g/l lait	135	0.7398	0.6324
Caséines alpha s2 + kappa (Levicek)	g/l lait	135	0.7665	0.4050
Caséine Beta (Levicek)	g/l lait	133	0.6854	1.2657
Lactalbumine (Levicek)	g/l lait	138	0.3243	0.1528
Lactoglobulines (Levicek)	g/l lait	134	0.7076	0.3195
Caséines totales (Levicek)	g/l lait	133	0.7930	1.7906
Protéines du lactosérum (Levicek)	g/l lait	126	0.7184	0.3448
Matière Azotée Protéique (Levicek)	g/l lait	133	0.8041	1.9213
Caséine totale mesuree par Kjeldahl	g/100g lait	976	0.9498	0.0755
Acide gras C4	g/dl lait	1780	0.9258	0.0079
Acide gras C6	g/dl lait	1786	0.9079	0.0063
Acide gras C8	g/dl lait	1789	0.9083	0.0042
Acide gras C10	g/dl lait	1785	0.9122	0.0103
Acide gras C12	g/dl lait	1784	0.9237	0.0117
Acide gras C14	g/dl lait	1781	0.9338	0.0314
Acide gras C14 :1	g/dl lait	1781	0.6836	0.0079
Acide gras C16	g/dl lait	1789	0.9428	0.0926
Acide gras C16 :1 cis	g/dl lait	1783	0.7255	0.0128
Acide gras C17	g/dl lait	1785	0.7971	0.0034
Acide gras C18	g/dl lait	1782	0.8411	0.0565
Acides gras totaux C18 :1 trans	g/dl lait	1784	0.7863	0.0261
Acide gras C18 :1 cis9	g/dl lait	1781	0.9472	0.0617
Acides gras totaux C18 :1 cis	g/dl lait	1787	0.9522	0.0638
Acides gras totaux C18 :2	g/dl lait	1786	0.6878	0.0147
Acide gras C18 :2 cis9 cis12	g/dl lait	1783	0.7243	0.0116
Acide gras C18 :3 cis9 cis12	g/dl lait	1773	0.6774	0.0044
Acide gras C18 :2 cis9 trans11	g/dl lait	1768	0.7351	0.0103

Acides gras saturés	g/dl lait	1790	0.9904	0.0719
Acides gras monoinsaturés	g/dl lait	1793	0.9705	0.0581
Acides gras polyinsaturés	g/dl lait	1799	0.7729	0.0214
Acides gras insaturés	g/dl lait	1788	0.9698	0.0648
Acides gras à courtes chaînes	g/dl lait	1791	0.9334	0.0246
Acides gras à moyennes chaînes	g/dl lait	1781	0.9673	0.1048
Acides gras à longues chaînes	g/dl lait	1779	0.9510	0.1104
Acides gras isoanteiso	g/dl lait	1785	0.7495	0.0124
Acides gras Omega 3	g/dl lait	1779	0.6649	0.0056
Acides gras Omega 6	g/dl lait	1784	0.7182	0.0149
Total od odd FA	g/dl lait	1777	0.8278	0.0162
Acides gras totaux trans	g/dl lait	1783	0.8049	0.0302
Acides gras totaux C18 :1	g/l lait	1790	0.9628	0.0610
Na	mg/kg lait	1019	0.4445	50.9780
Ca	mg/kg lait	1094	0.8167	53.3776
P	mg/kg lait	1083	0.7472	58.7148
Mg	mg/kg lait	1124	0.7171	6.5349
K	mg/kg lait	1090	0.5461	88.1391
Citrates	mmol/L lait	498	0.8917	0.7197
Acétone (log10)	mmol/l lait	201	0.6236	0.2101
β-Hydroxybutyric acid (log10)	micromole/l lait	419	0.7081	0.1363
pH	[-]	1152	0.633	0.070
Acidité titrable	degree Dornic	930	0.683	1.04
Rendement Beurre Lait	g beurre/100g lait	97	0.729	0.5787
pH beurre	[-]	63	0.3394	0.4973
RCT JR (log2)	secondes	413	0.7029	0.2561
RCT + K20 en JG (log2)	secondes	663	0.6320	0.2107
RCT + K20 en JR	secondes	347	0.679	177.94
K20 en JG (log2)	secondes	791	0.437	0.182
A30 en racine	mm	185	0.50	3.04
Rendement fromager en sec	g MS/100g MS lait	344	0.812	3.188
Rendement fromager en frais	g/100 g lait	352	0.7184	3.8652
Rendement fromager en sec	g MS/100g MS lait	367	0.7782	3.5156
Rendement fromager atelier	g / 100g lait	45	0.507	1.99
Cheese Yield Curd	g / 100g lait	45	0.507	1.99
Cheese Yield Solide	g / 100g lait	51	0.685	0.63
Solid Recovery	g / 100 g	49	0.599	4.77
Fat Recovery	g / 100 g	28	0.475	16.63
Protein Recovery	g / 100 g	41	0.475	7.518
RCT	seconde	935	0.639	148.0
K20	seconde	874	0.511	29.2
RCT + K20	seconde	854	0.629	159.9
A30	mm	547	0.422	3.22
pH Yaourt	—	102	0.12	0.12
Activité Yaourt	Degre dornic	98	0.226	0.0475
Matière sèche Yaourt	g/100g gr yaourt/100g	70	0.550	0.81
Synérèse Yaourt	surnageant	72	0.472	5.40
Texture Yaourt (PAS BON)	N	51	0.075	0.019
Milk lactoferrin	—	2189	0.656	130.4
Energy Balance	—	1010	0.432	5.038
Protein Efficiency	—	1093	0.5188	3.17

Glucose 6P in milk	mmoles/l milk	2129	0.517	0.O549
Milk Glucose free	mmoles/l milk	2129	0.5014	0.0743
Milk BOHB	µmole/l milk	2129	0.40	47.31
Milk IsoC	mmoles/l milk	2129	0.5742	0.0334
Milk urea	mmoles/l milk	2129	0.50	0.93
Milk NAGase	units/l milk	2129	0.426	1.55
Milk LDH	units/l milk	2129	0.333	2.79
Milk UA	µmoles/l milk	2128	0.346	43.36
Milk Progesterone	ng/ml milk	2128	0.09	2.599
Blood IGF-1	—	371	0.550	39.3
Blood Glucose	mmoles/l plasma	369	0.402	0.366
Blood urea	mmoles/l plasma	374	0.491	0.807
Blood Cholesterol	mmoles/l plasma	373	0.464	0.82
Blood Fructosamine	µmoles/l plasma	368	0.224	15.31
Blood BOHB log10	mmoles/l plasma	199	0.662	0.1380
Blood NEFA	µEq/l plasma	230	0.373	347.114
Blood Progesteron	ng /ml plasma	367	0.142	2.377
Dry Matter Intake	kg/d	1020	0.450	3.31
Residual Feed Intake 1	kg/d	1008	0.478	2.84
Residual Feed Intake 2	kg/d	1013	0.426	2.89

7.2 Annexe 2

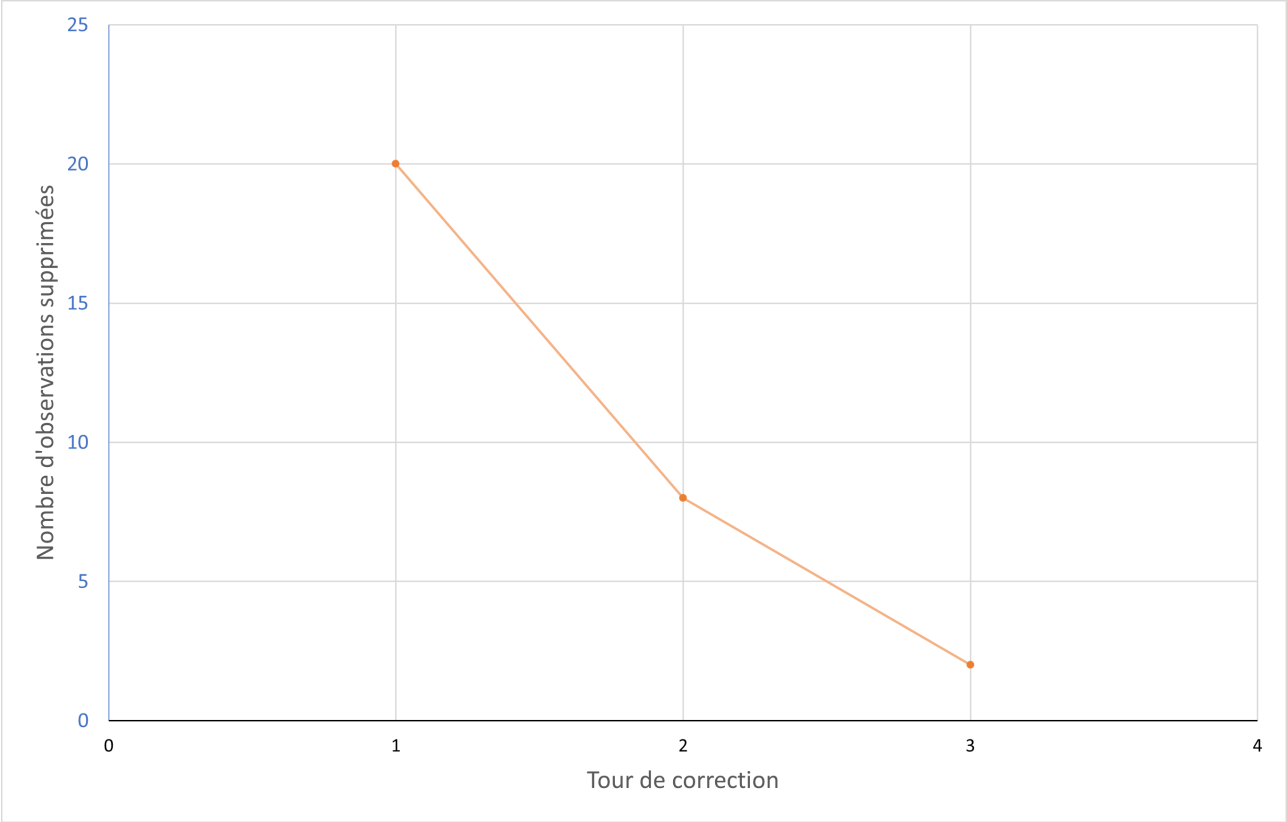


FIGURE 7.1 – Nombre d'échantillons supprimés par tour de correction via l'analyse des résidus.

7.3 Annexe 3

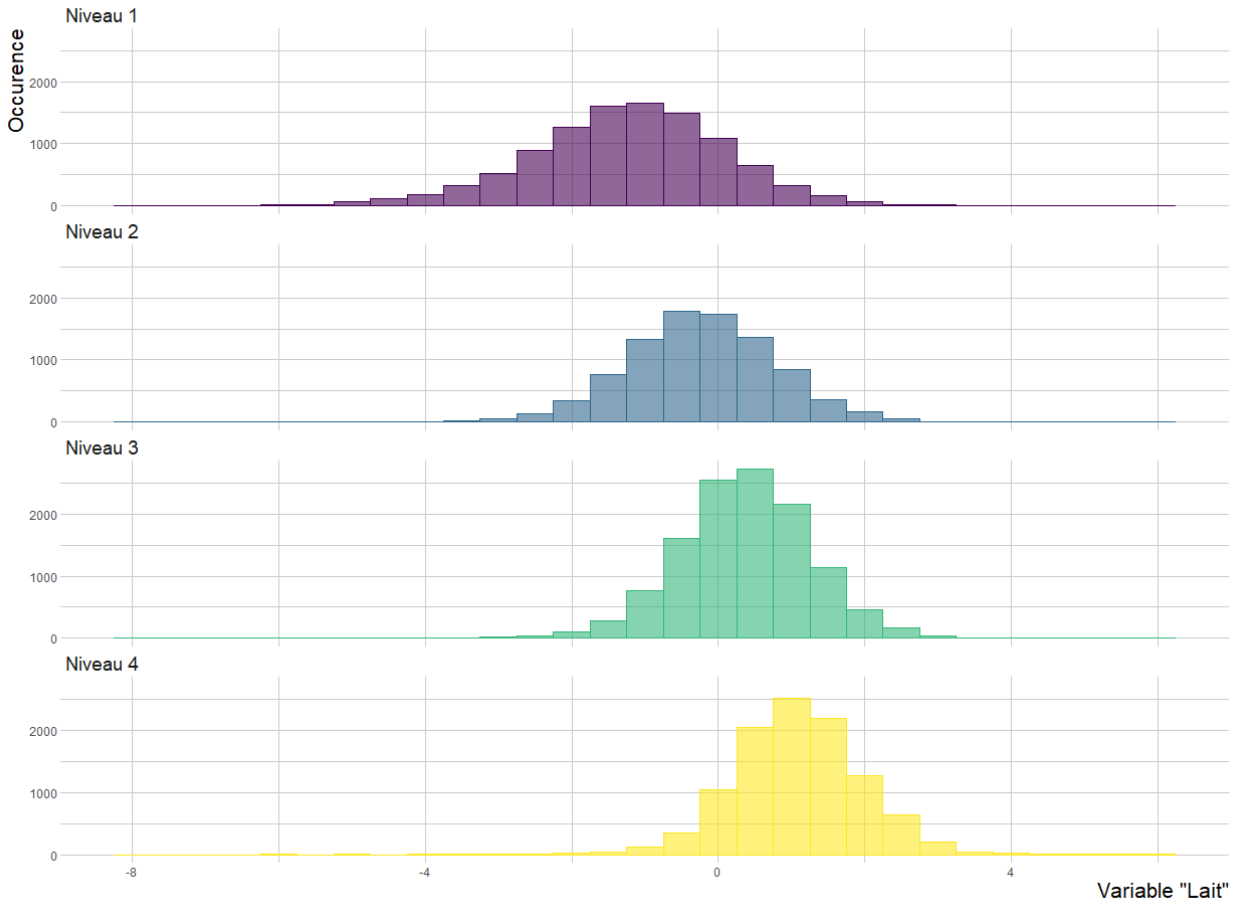


FIGURE 7.2 – Évolution de la distribution de la production de lait normalisée établie selon quatre niveaux croissant d'intensivité sur l'ensemble de la base de données lait.

7.4 Annexe 4

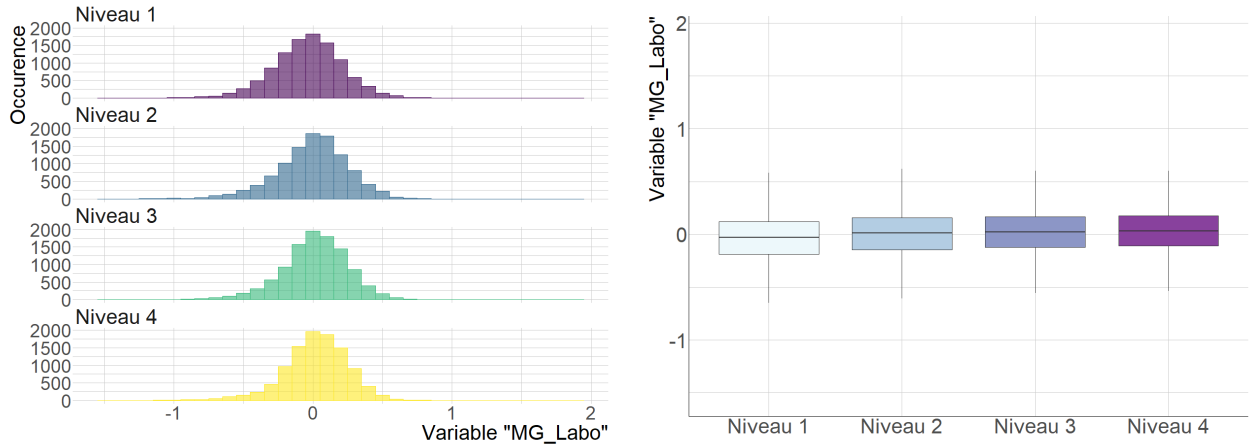


FIGURE 7.3 – Évolution de la distribution des teneurs en MG selon quatre niveaux croissants d'intensivité sur l'ensemble de la base de données lait normalisées.

7.5 Annexe 5

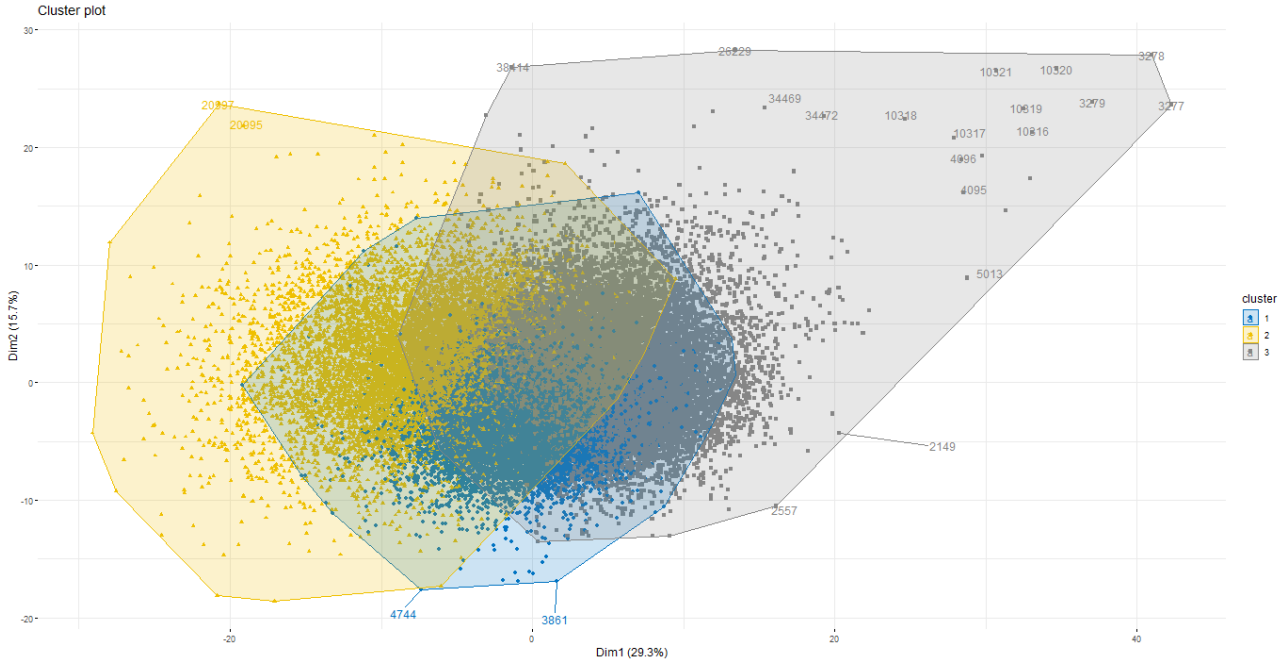


FIGURE 7.4 – Représentation du clustering selon les deux premières composantes principales issues d'une ACP.

7.6 Annexe 6

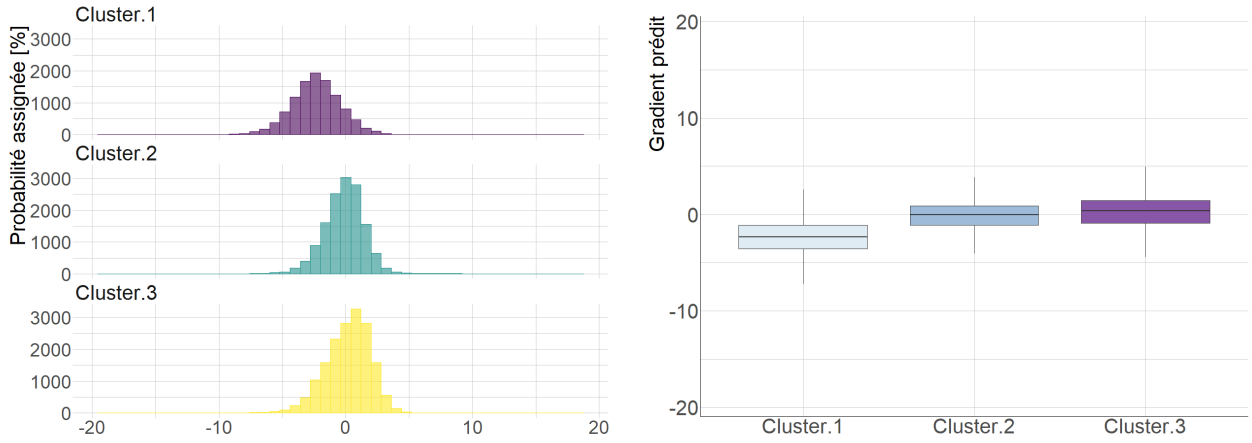


FIGURE 7.5 – Évolution de la distribution du gradient prédit selon les clusters issus du clustering sur l'ensemble de la base de données lait.