Stage Master 1
Mathématiques appliquées à l'ingénierie, l'industrie et l'innovation
Année 2017/2018

Exploration du comportement d'un simulateur individu-centré de fonctionnement du troupeau caprin laitier : analyse de sensibilité globale

---

Vergès Vanessa

Enseignant référent :
F. Gamboa

Maîtres de stage :
P. Chabrier
V. Picheny
L. Puillet

Stage effectué du 16 avril au 16 août 2018,
à l'INRA (département MIA, unité MIAT),
24 Chemin de Borde Rouge, 31320 Auzeville-Tolosane, France

Remerciements

Je remercie tout d'abord mes encadrants, Victor, Patrick et Laurence qui m'ont donné l'occasion de réaliser cette merveilleuse expérience en m'accordant leur confiance. Ils m'ont donné de leur temps et m'ont guidée tout au long de ce stage. Ils m'ont tous les trois apporté un regard critique sur mon travail et m'ont permis d'avancer dans les meilleures conditions. C'est entièrement à eux que je dois l'aboutissement de ce stage pour leur encadrement et leur suivi.
Plus particulièrement, je remercie Patrick, pour m'avoir fait découvrir et m'avoir formée sur le cluster de calculs, mais aussi pour avoir gentillement complété mes connaissances sur linux.
Je remercie Victor pour avoir suivi de près chacun de mes résultats, pour m'avoir partagé son expertise mathématique dans ce domaine et avoir guidé mes recherches.
Je remercie ma troisième encadrante, Laurence, pour m'avoir partagé son expertise de modélisatrice ainsi que ses connaissances dans le domaine de l'élevage caprin. Je tiens aussi à la remercier pour le séjour à Poitiers, elle m'y a fait découvrir les ficelles d'un projet ainsi que toute la chaleur humaine qui pouvait y avoir derrière.
Je tiens au passage à remercier Hugues Caillat, responsable du dispositif Patuchev [1] et Nicole Bossis (Institut de l'Elevage), pour m'avoir écoutée et avoir partagé leur expérience du terrain afin de confirmer mes résultats.
Je tiens encore à remercier Hugues, pour son accueil et pour ses visites guidées de la plateforme Patuchev ainsi que de la ville de Poitiers.

Je souhaite remercier aussi les membres du département MIA, les personnes de l'administration, les chercheurs, les doctorants ainsi que les stagiaires que j'ai pu rencontrer, pour leur contribution à mon intégration et pour la chaleur de nos déjeunés partagés.

Finalement, je souhaiterais remercier Fabrice Gamboa qui a accepté d'être mon enseignant référent et qui a aussi très gentillement accepté de suivre ma soutenance à distance.

1. Unité expérimentale Fourrages, environnement, ruminants de Lusignan. Le dispositif expérimental Patuchev a pour but de concevoir des systèmes d'élevage caprins autonomes et économes. Il compare trois troupeaux conduits avec deux périodes de reproduction différentes et alimentés avec de l'herbe soit fauchée, soit pâturée.

Introduction

Ce stage a été effectué au sein de l'INRA de Toulouse sous la supervision de Patrick Chabrier, Victor Picheny (Unité MIAT, Toulouse) et de Laurence Puillet (Unité MoSAR, Paris), dans le cadre du projet FLECHE.

L'INRA

L'Institut National de la Recherche Agronomique est un organisme public de recherche scientifique, fondé en 1946 et réunissant plus de 10 000 agents sur l'ensemble du territoire national. L'INRA est le premier institut de recherche agronomique en Europe. Ses recherches concernent les domaines comme l'agriculture, l'alimentation et la sécurité des aliments, l'environnement et la gestion des territoires, avec un accent tout particulier en faveur du développement durable.
Le siège de l'INRA est situé à Paris mais de nombreux centres de recherche sont implantées partout en France. Comme annoncé précédemment, ce stage a été effectué dans le centre de recherche de Toulouse au sein de l'unité MIAT [2], Mathématiques et Informatique Appliquées de Toulouse. Cette unité fait partie du département de Mathématiques et Informatique Appliquées, le MIA.
Le MIAT a pour mission scientifique de développer et mettre en oeuvre des méthodes mathématiques et informatiques pertinentes afin de résoudre certains problèmes. L'unité comporte deux équipes de recherche (MAD [3] et SAB [4]) et trois équipes de service (Plateformes GENOTOUL [5], RECORD [6] et SIGENAE [7]). L'unité MoSAR, Modélisation Systémique Appliquée aux Ruminants, est quant à elle située à Jouy-en-Josas. Elle a pour objectifs de comprendre, caractériser, et prédire les relations entre l'animal d'élevage et son environnement alimentaire afin de développer des outils pour augmenter l'efficacité d'utilisation des ressources alimentaires par une réalisation optimale des performances, des capacités d'adaptation, et du bien-être.

2. anciennement unité de Biométrie et Intelligence Artificielle
3. Modélisation des Agro-écosystèmes et Décision
4. Statistiques et Algorithmique pour la Biologie
5. Plateforme bioinformatique du GIS GENOTOUL - Génopole Toulouse Midi-Pyrénées
6. REnovation et COoRDination de la modélisation des cultures pour la gestion des agro-écosystèmes : Plateforme de modélisation et de simulation des agro-écosystèmes
7. Systèmes d'Information des GENomes des Animaux d'Elevage : Plateforme Systèmes d'information des génomes des animaux d'élevage

Projet FLECHE : Fromages et Laits issus d'Élevages de Chèvres conduites à l'HErbe

Le projet FLECHE a été initié en 2016 pour une durée de 4 ans. Il vise à renforcer la durabilité des filières caprines laitières du Grand Ouest, et l'enjeu est majeur : en effet, ces filières représentent actuellement 46% de l'effectif national de chèvres et 64% du lait livré en France.
Il s'appuie sur un dispositif original de recherche et de développement en Grand Ouest [8] : les dispositifs expérimentaux Inra Patuchev (Lusignan) et de Méjussaume (Le Rheu), le réseau INOSYS (Institut de l'élevage et Chambres d'Agriculture), ainsi que le Réseau d'expérimentation et de Développement Caprin (REDCap) porté par le Bureau Régional de l'Interprofession Laitière Caprine (BRILAC).

Le projet FLECHE a pour objectifs de comprendre la place, le niveau actuel ainsi que le potentiel de valorisation de l'herbe dans les systèmes caprins, d'en déterminer de manière objective les avantages et les freins techniques et sociologiques, et d'apporter des références scientifiques et techniques pour permettre d'accroître significativement l'utilisation de l'herbe dans les systèmes caprins du Grand Ouest, accroissant ainsi la durabilité de la filière. Ce projet est financé par le programme PSDR Grand Ouest [9]. Parmi les nombreux objets de recherche de ce programme, ce stage s'inscrit dans la partie portant sur l'évaluation des résultats technico-économiques (consommations des aliments par le troupeau et production de lait). Ces résultats seront étudiés pour trois systèmes d'alimentation, que l'on appelera scénarios : l'ensilage de maïs, le foin de luzerne et le foin de graminées.

Intérêt du stage

L'évaluation des performances technico-économiques de différents systèmes alimentaires, valorisant plus ou moins d'herbe dans la ration du troupeau, s'appuie sur un modèle de simulation existant, SIGHMA *SImulation of Goat Herd MAnagement*, par L.Puillet, 2010 [4]. Le but de ce stage est de réaliser une exploration de ce dernier afin d'appréhender son comportement global et de vérifier la cohérence des simulations. Une analyse de sensibilité sera réalisée afin d'identifier les paramètres qui ont le plus d'influence sur les sorties d'intérêt, à savoir celles qui conditionnent les performances technico-économiques (production laitière et consommation d'aliments).
L'enjeu de cette analyse de sensibilité est d'identifier les leviers du fonctionnement du troupeau qui affectent les performances pour pouvoir orienter par la suite les pratiques des éleveurs et leurs choix autour de leur système d'alimentation.

8. Nouvelle-Aquitaine, Pays de la Loire, Bretagne, Normandie
9. Pour et Sur le développement Régional du Grand Ouest

Table des Matières

Chapitre 1

Modèle SIGHMA

Dans ce chapitre nous verrons le fonctionnement du troupeau caprin laitier et ensuite l'organisation du simulateur qui le représente : SIGHMA.

1.1 Un troupeau comment ça marche ?

Tout d'abord, étudions le fonctionnement d'un troupeau caprin laitier.

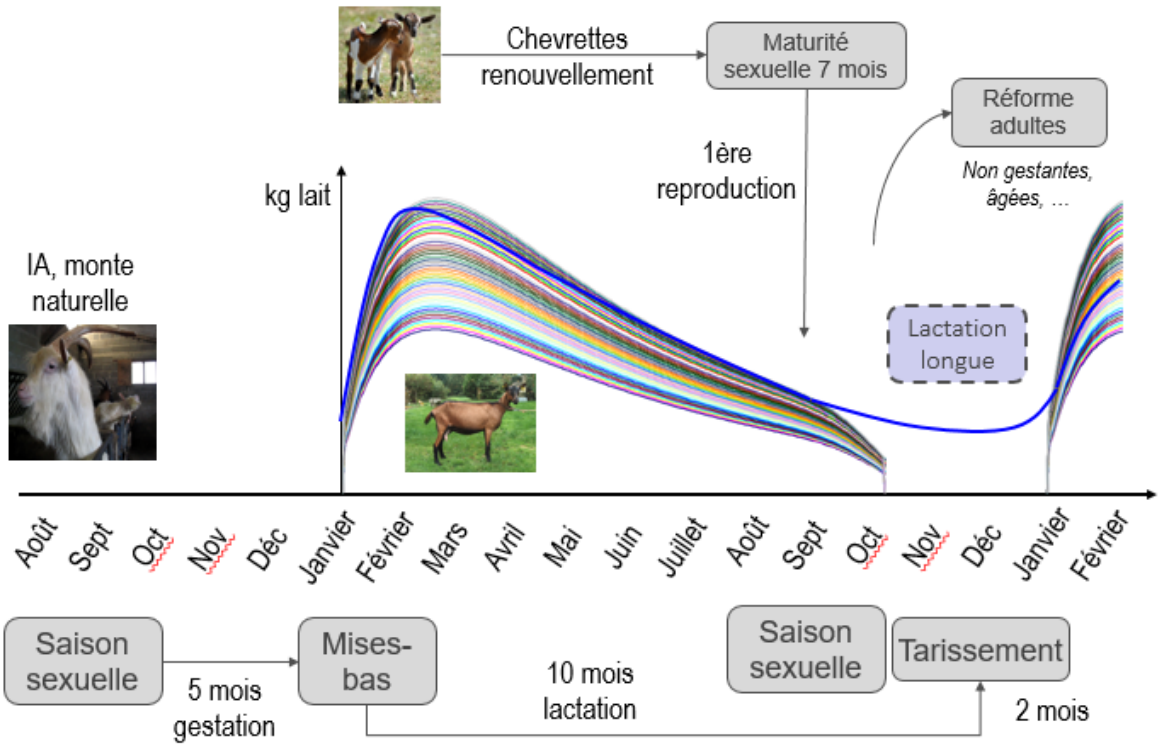


FIGURE 1.1 – Cycle de production laitière caprine

L'espèce caprine présente une activité sexuelle saisonnière. La saison sexuelle s'étale d'août à octobre. En conséquence, les mises-bas sont groupées de janvier à mars. A noter que dans la représentation précédente, figure 1.1, les courbes de mises-bas sont regroupées en janvier afin de simplifier la lisibilité du graphique. La gestation dure environ cinq mois (153 jours en moyenne). L'existence d'une saison sexuelle marquée entraîne une production laitière saisonnière maximale deux mois après la mise-bas, vers fin février, début mars. Les chevrettes sont conduites à la reproduction à partir de sept mois, donc début août, en même temps que le début de la saison sexuelle des chèvres déjà présentes dans le troupeau. Ces chèvres continuent de produire durant cette saison sexuelle.
A la fin du mois d'octobre un bilan est dressé : les chèvres en gestation sont automatiquement gardées au sein du troupeau et elles sont alors taries [1] ; la question se pause en revanche pour les chèvres ayant échoué à la reproduction. Ces dernières sont alors soit gardées au sein du troupeau et placées en lactation longue, c'est-à-dire que l'on continue de les traire, soit sorties du troupeau. Les critères utilisés pour cette selection sont des critères d'âge et/ou de génétique (notamment le potentiel laitier de celles-ci).

Passons maintenant à la modélisation de ce fonctionnement de troupeau.

1.2 Description générale du modèle

Ce modèle est dit individu-centré : il est composé d'un ensemble de modèles individuels (chacun d'eux représentant le fonctionnement biologique d'une chèvre, de sa naissance jusqu'à sa sortie du troupeau). Ces modèles individuels sont basés sur un formalisme d'équations différentielles et d'intégration numérique Runge Kutta 4.

Voici le fonctionnement d'un modèle individuel : une chèvre ingère de l'énergie par la consommation de ration alimentaire, cette énergie est ensuite allouée à quatre fonctions biologiques : la croissance, l'entretien, la lactation et la gestation, en fonction du stade physiologique de l'animal (gestante ou en lactation).
Ceci est représenté un peu plus en détails sur le graphique suivant :

1. Le tarissement consiste à interrompre la lactation en arrêtant la traite.

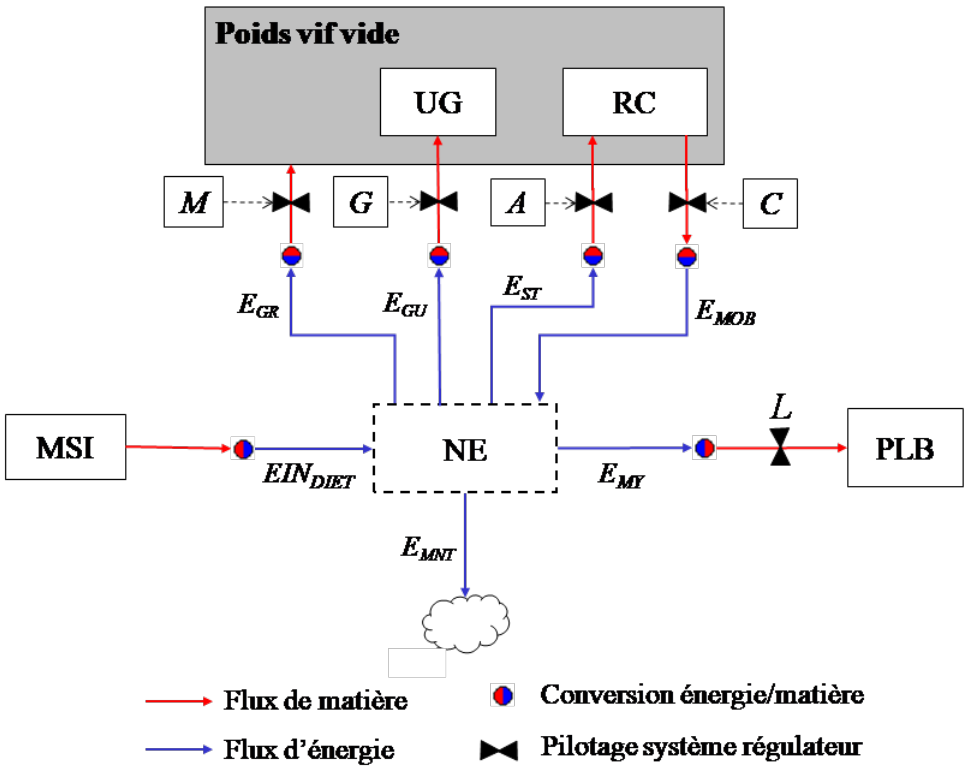


FIGURE 1.2 – Représentation schématique du sous-système représentant la répartition d'énergie chez la chèvre laitière - *UG : utérus gravide ; RC : réserves corporelles ; NE : compartiment zéro-pool d'énergie nette ; MSI : matière sèche ingérée ; PLB : production laitière brute ; EGR : flux d'énergie associé à la croissance ; EGU : flux d'énergie associé à la croissance de l'utérus gravide ; EST : flux d'énergie associé à la reconstitution ; EMB : flux d'énergie associé à la mobilisation ; EINDIET : flux d'énergie ingérée ; EMY : flux d'énergie exportée dans le lait ; EMNT : flux d'énergie associé à l'entretien ; M, G, A, C, L : compartiments du sous-système régulateur représentant respectivement les hormones théoriques de croissance, de gestation, d'anabolisme, de catabolisme et de lactation.*

En ce qui concerne les actions de l'éleveur, elles sont représentées par des évènements aléatoires discrets. On en distingue trois types : la gestion de l'effectif du troupeau, la gestion de la reproduction et la gestion de l'alimentation.

1.3 Les entrées

Pour effectuer une simulation, SIGHMA nécessite un fichier de management contenant entre autres, le lieu d'enregistrement des résultats, certains paramètres fixés ainsi que sept valeurs : les sept variables d'entrée.
Deux de ces variables servent au tirage aléatoire dans une loi triangulaire du potentiel laitier d'une chèvre attribué à la naissance de celle-ci : la moyenne (*POT_mean*) et l'écart-type (*POT_sd*) de cette loi triangulaire.
Une autre de ces variables représente la durée de la période de reproduction (*BSL*), une autre la probabilité journalière de réussite à la reproduction (*Repro_Success*) et enfin, trois autres variables intervenant dans le renouvellement du troupeau : le taux de renouvellement (*Replace_rate*), le critère de selection pour la mise en

lactation longue (*POT_extlac*) et enfin l'intensité de selection des femelles à réformer sur leur potensiel laitier et leur âge (*POT_cull*).
Voici les plages de variation de ces paramètres :

Paramètre	Valeurs
POT_mean	3.5 - 5.5
POT_sd	0.205 - 0.612
BSL	63 - 126
Repro_Success	0.0143 - 0.0309
POT_extlac	1.99 - 7.01
Replace_rate	0.15 - 0.45
POT_cull	0 - 1

TABLE 1.1 – Tableau récapitulatif des paramètres et des variations possibles

1.4 Les sorties

Le modèle génère un grand nombre de sorties, au niveau des individus simulés (variables biologiques de chaque chèvre du troupeau) et au niveau agrégé du troupeau. Dans le cadre du stage, nous nous intéressons seulement au niveau troupeau et nous avons centré l'analyse sur six variables d'intérêt : celle concernant les quantités de matières produites (la production laitière, *Prod*), celles représentant les quantités de matières consommées (la consommation de trois fourrages et d'un concentré, *Conso1, Conso2, Conso3* et *ConsoC*), ainsi que l'effectif de chèvres en lactation, *Eff*.

1.5 La structure du code

SIGHMA a tout d'abord été codé avec ModelMaker, puis traduit en Python et C++ avec un langage passerelle entre les deux : le Cython.
(D'ailleurs, l'un des enjeux de ce stage est aussi d'explorer le comportement du modèle sous Python pour détecter d'éventuels problèmes de programmation par rapport à l'ancienne version sous ModelMaker.)

Comme décrit un peu plus haut, SIGHMA est un simulateur individu centré, c'est-à-dire que le comportement de chaque individu appartenant à la population étudiée est modélisé individuellement. Une des conséquences de cette structure est que nous devons le faire tourner sur une longue durée, c'est pourquoi SIGHMA est étudié sur 20 ans. Une autre conséquence de cette caractéristique est que c'est un simulateur coûteux en temps de calcul : une simulation dure environs 20 minutes . Nous avons donc dû trouver des moyens pour palier à cela.

Chapitre 2

Exploration du modèle : Etude qualitative des sorties

Une première approche a été d'observer le comportement des cinq sorties d'intérêt du modèle en fonction de chaque entrée. Cette partie a permis d'emettre certaines hypothèses mais surtout d'explorer le modèle et ainsi de détecter certaines anomalies ou cas non prévus.

2.1 Scripts de génération des fichiers de management

Une première étape de ce stage a consisté à la mise en place de scripts pour générer des fichiers de management. Globalement, il s'agit de créer des copies du fichier de management de base contenant des *placeholders*, en trouvant et en complétant ces emplacements par les valeurs des paramètres d'entrée, puis de créer **s** copies de chaque fichier en y ajustant la ligne contenant la destination d'enregistrement de la simulation.

2.2 Générer les simulations

Nous avons réalisé quelques simulations, pour cela nous avons opté pour lancer des simulations en parallèle sur la machine. Nous pouvions ainsi lancer 10 simulations en même temps.
Puis, nous savions que nous aurions besoin de faire de nombreuses simulations pour réaliser une exploration du modèle. Pour par exemple faire 3500 simulations (nous verrons dans la partie suivante pourquoi 3500), nous avons besoin de 3500×20 minutes, soit 70 000 minutes, c'est-à-dire un peu plus de 1167 heures. En lançant le tout en parallèle on ne divise le temps de calcul que par 10.

C'est ainsi qu'est venu la solution : le cluster de calculs de la plateforme GenoToul (évoquée page 3). Grâce à ce dernier, les 3500 simulations sont faîtes en seulement 4 à 5 heures [1].

1. temps variable en fonction de la disponibilité des CPUs

2.3 Etude de la variabilité du modèle. Support : nombre de simulations pour obtenir une stabilisation des sorties

Etant donné que SIGHMA est un modèle stochastique, et dans le but d'obtenir des résultats significatifs, nous avons travaillé avec les moyennes des différentes sorties calculées sur **s** simulations. Comme nous l'avons vu dans la partie précédente, SIGHMA est un modèle assez coûteux. Ainsi, chaque simulation compte. Nous avons donc tout d'abord étudié quel serait le nombre de simulations nécessaires par combinaison de paramètres. Nous avons démontré que $\mathbf{s} = 10$ simulations suffisaient.

2.4 Plan d'expérience : Hypercube Latin Optimisé

Afin d'explorer le modèle, nous avons besoin de générer de nombreuses observations. Pour cela, nous devons déterminer les points en lesquels nous allons explorer. C'est ensuite à partir des observations réalisées en ces points que l'on pourra approcher le résultat en tout point de l'espace.
Ces points sont de dimension $d = 7$, ils représentent une combinaison des 7 paramètres d'entrée (cf tableau 1.1). Pour choisir au mieux ces points nous avons utilisé une méthode appelée Hypercube Latin Optimisé.

Tout d'abord nous allons définir ce qu'est un hypercube latin.
Soit N le nombre de points et d le nombre de paramètres, un hypercube latin de dimensions $N \times d$ est une matrice (de taille $N \times d$), dont chaque colonne représente une permutation de l'ensemble {1,...,N}. Ainsi, pour un plan hypercube latin (on dit aussi plan LHS, Latin Hypercube Sampling), à N points, chaque dimension de l'espace à observer sera découpée en N intervalles et un point par intervalle sera choisi.
Voici un exemple sur un espace de dimension $d = 2$, avec $N = 5$ points à déterminer.

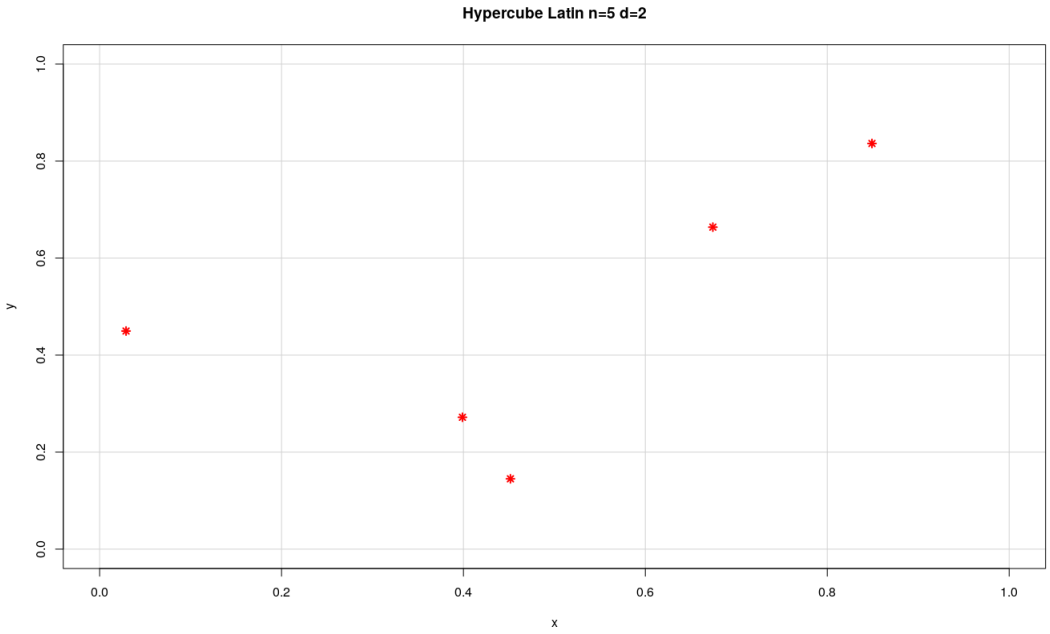


FIGURE 2.1 – Echantillongae par hypercube latin avec 5 points en dimension 2

L'utilisation de cette méthode permet d'obtenir une projection uniforme sur les marginales.
L'inconvenient de cette dernière est qu'elle peut s'avérer peu performante dans le sens où certaines zones de l'espace peuvent être inexplorées, comme on peut le voir sur la figure 2.1 . C'est pour cela que nous utiliserons plutôt la méthode d'hypercube latin optimisé.

Il existe plusieurs critères d'optimisation pour les hypercubes latins, mais les deux plus connus sont le minimax et le maximin. Ces critères sont basés sur des distances entre les points de l'espace à explorer E. Par exemple, pour construire un plan d'expérience $X = (x_i)_{i=\{1,...,N\}}$ (avec chaque point $x_i = (x_{i,1}, ..., x_{i,d})$), on peut vouloir maximiser la distance minimale séparant toute paire de points du plan. On note $d(x_i, x_j) = ||x_i - x_j||_{L_p}$.

Le maximin sert à espacer le plus possible les points du plan les uns des autres afin de recouvrir un maximum de surface.
Le maximin est déterminé en maximisant la distance minimale entre deux points du plan d'expériences, il minimise ainsi le nombre de paires de points exactement séparés par la distance minimale.
Le maximin est :

$$\max_{x_1,...,x_N} \Big(\min_{i,j \in \{1,...N\},\ i \neq j} d(x_i, x_j) \Big).$$

Le critère maximin est à maximiser afin d'espacer les plus possible les points du plan d'expériences.

Le minimax quant à lui permet de rapprocher le plus possible tout point du plan X de tout point de l'espace à explorer E.
Le minimax représente la distance maximale entre un point $\tilde{x}$ du domaine à explorer E et un point x_i du plan X.
Le critère minimax est donc :

$$\max_{\tilde{x}\in E}(\min_{i} \in \{1,...N\}d(\tilde{x},x_i)).$$

Une valeur peu élevée du minimax pour un plan d'expériences donné signifie que chaque point du domaine n'est jamais trop distant d'un point du plan. Le critère minimax est donc à minimiser.
Le calcul de ce critère passe par le calcul des distances entre tous les points du domaine et tous les points du plan. En pratique, la discrétisation du domaine permet de calculer une approximation de ce critère mais la méthode devient très coûteuse en dimension supérieure à 3. Ce critère n'est donc pas utilisé pour la planification d'expériences numériques. Ainsi nous n'avons utilisé que le critère du maximin.

Voici un exemple sur le même espace de dimension $d = 2$, avec toujours $N = 5$ points à déterminer mais avec cette fois un hypercube latin optimisé.

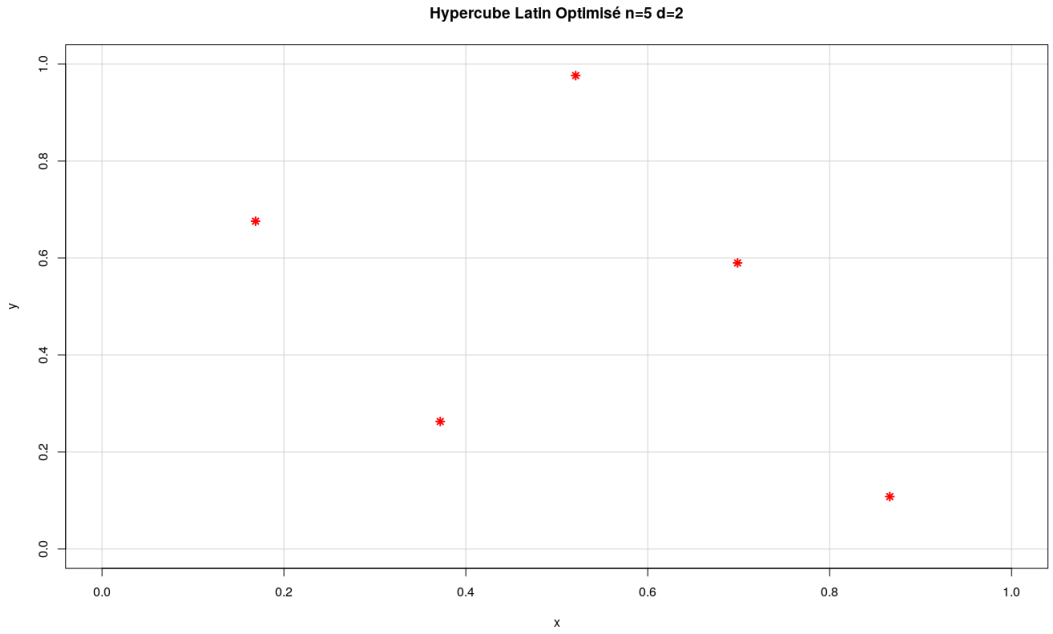


FIGURE 2.2 – Hypercube Latin Optimisé avec 5 points en dimension 2

2.5 Etude des sorties en fonction de chaque entrée

Nous avons en amont étudié le comportement de différentes variables présentes dans le modèle et de différentes sorties en fonction de chaque entrée, dans le but d'explorer le comportement du modèle et afin d'émettre des hypothèses sur les entrées qui semblent être les plus influentes.

Suite à certains résultats obtenus pendant l'exploration, le modèle a été revu par l'équipe de l'UMR MoSAR avant de lancer les analyses finales. Certaines anomalies ont ainsi été détectées et corrigées dans le code. Par exemple une anomalie portant sur la consommation des chevrettes qui était la même que la consommation des

chèvres a été corrigée. Nous avons aussi, grâce à cette exploration, observé que certains des résultats obtenus présentent des caractéristiques atypiques, telles que des effectifs de chèvres en lactation trop faibles. La diminution du nombre d'apparition de ces résultats permetterait d'affiner la réponse du modèle. En effet, dans la réalité, le nombre de chèvres n'est jamais trop bas : l'éleveur rachète automatiquement des chèvres afin que son cheptel, et donc sa production, ne chutte pas. C'est ainsi que l'on peut justifier l'intêret pour ces points.
Notre idée a été d'identifier les caractéristiques communes des entrées correspondant à ces sorties. Pour cela nous avons tout d'abord réalisé des arbres de classification supervisée.
Nous en avons ensuite déduit de nouvelles bornes pour les entrées :

Paramètre	Valeurs
POT_mean	3.5 - 5.5
POT_sd	0.205 - 0.612
BSL	~~63~~ 84 - 126
Repro_Success	~~0.0143~~ 0.019 - 0.0309
POT_extlac	1.99 - 7.01
Replace_rate	~~0.15~~ 0.33 - 0.45
POT_cull	0 - 1

TABLE 2.1 – Tableau récapitulatif des paramètres

2.5.1 Arbre de classification supervisée

La construction d'un arbre de classification peut être vue comme un partitionnement récursif. Ces arbres font partie des arbres de décision. Un arbre de décision est un enchaînement hiérarchique de règles logiques construites de manière automatique à partir d'un ensemble de départ : un ensemble de couples (x, y) où x représente la description et y la classe d'appartenance. La construction de l'arbre de décision consiste à utiliser les descripteurs, les x, pour subdiviser progressivement l'ensemble de couples en sous-ensembles de plus en plus fins. Dans chaque sous-ensemble, une nouvelle évaluation est faite, celle-ci va permettre un nouveau découpage. Les ensembles terminaux sont appelés feuilles et les ensembles intermédiaires sont appelés noeuds.

Ce qui caractérise les arbres de classification, est que le schéma de séparation est construit de façon à minimiser le taux d'erreur de classification, contrairement aux arbres de regression pour lesquels le schéma de séparation vise à maximiser la variance inter-classes (avoir des sous-ensembles dont les valeurs de la variable soient les plus dispersées possibles). On parle de classification supervisée lorsque le critère de partitionnement est fixé.

Théorie

Pour un problème de classification, il existe différents critères par lesquels l'erreur sur un noeud peut être minimisée mais quatre d'entre eux sont les plus connus.

Le premier indice est l'erreur lié à une mauvaise classification. Il est simplement la proportion de points dans le noeud qui ne font pas partie de la classe majoritaire de ce noeud.

Comme indice de l'erreur de classification à minimiser, on a l'indice de Gini. Supposons qu'il existe K classes. Soit $\hat{p}_{mk}$ la proportion de points de classe k dans le noeud m. L'indice de Gini est

$$\sum_{k=1}^{K} \hat{p}_{mk}(1-\hat{p}_{mk}).$$

Cette mesure est souvent utilisée et est plus sensible que l'erreur de classification aux changements dans la probabilité de noeud.

L'indice d'entropie, aussi appelé l'entropie croisée peut être écrit comme

$$\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$$

Cet indice est aussi plus sensible que l'erreur de classification aux changements.

Et enfin, il existe un dernier critère conçu pour les problèmes mélant plusieurs classes. Cette approche favorise la séparation entre les classes plutôt que hétérogénéité des noeuds. Chaque split multiclass est traité comme un problème binaire. Les partitionnements qui conservent les classes liées sont favorisés. L'approche offre l'avantage de révéler les similitudes entre les classes.

Pour créer nos arbres de classification, nous avons utilisé la fonction *rpart*, de la bibliothèque du même nom. Nous avons ainsi pu créer des arbres CART (Classification And Regression Trees). Avec ces derniers, c'est l'indice de Gini qui est utilisé.

Résultats et interprétation

Nous avons utilisé cette méthode pour un traitement individuel sur les différentes sorties (comme on le montre en exemple dans le paragraphe suivant pour l'effectif) et aussi pour certaines combinaisons de sorties (comme par exemple dans le paragraphe d'après avec la production et la consommation du fourrage 3).

Effectif Nous avons d'abord créé des groupes (des classes) de points : les points représentant une moyenne des effectifs plus élevée ou plus basse que la valeur de référence, et parmis les points ayant une moyenne élevée, ceux ayant une variance élevée. Voici la répartition trouvée :

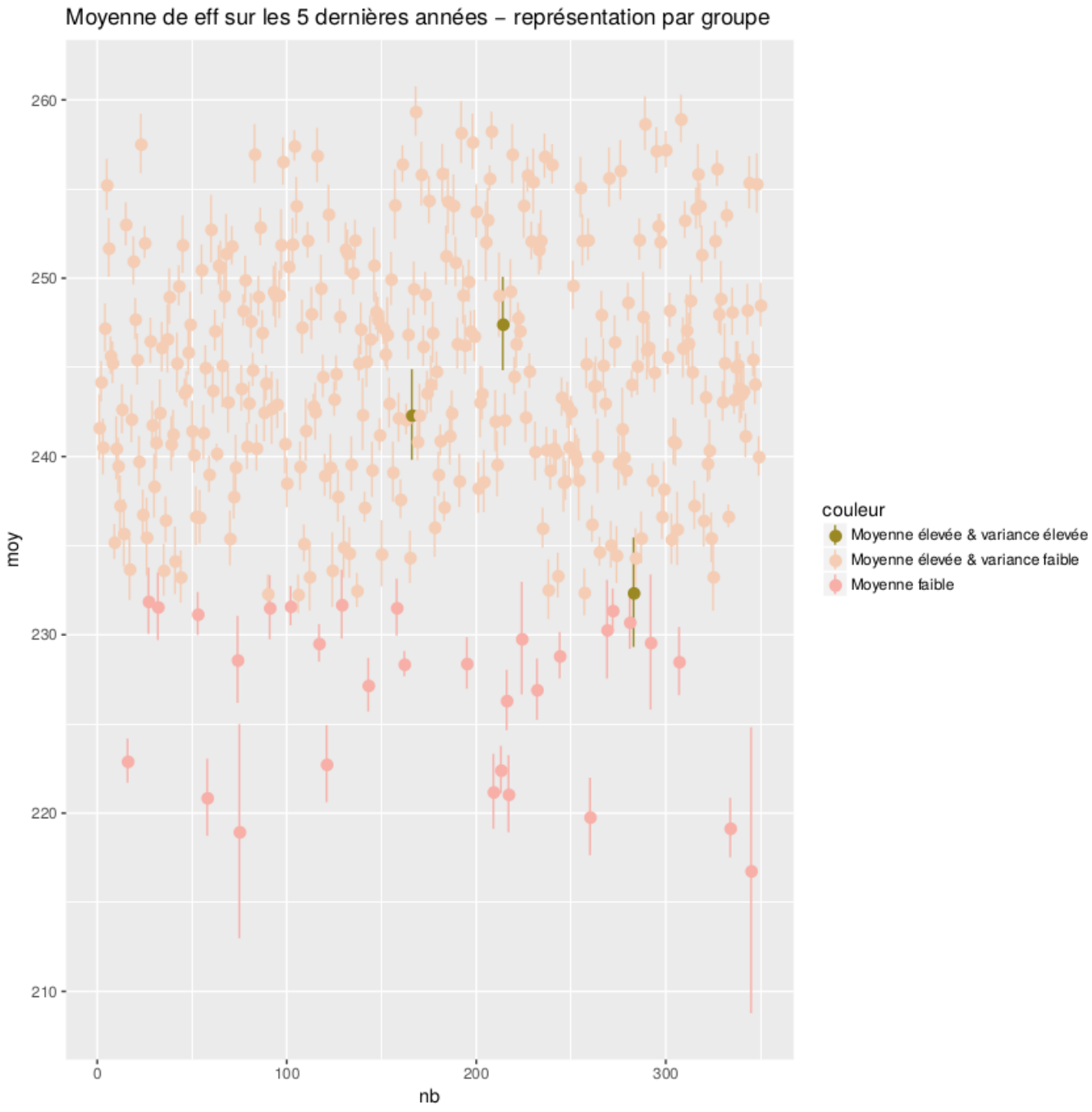


FIGURE 2.3 – Groupe pour l'effectif de chèvres en lactation - moyenne de référence à 232

A partir de cela nous avons créé l'arbre de classification correspondant. Etant donné que les groupes ont été déterminés au préalable, on parle de classification supervisée.

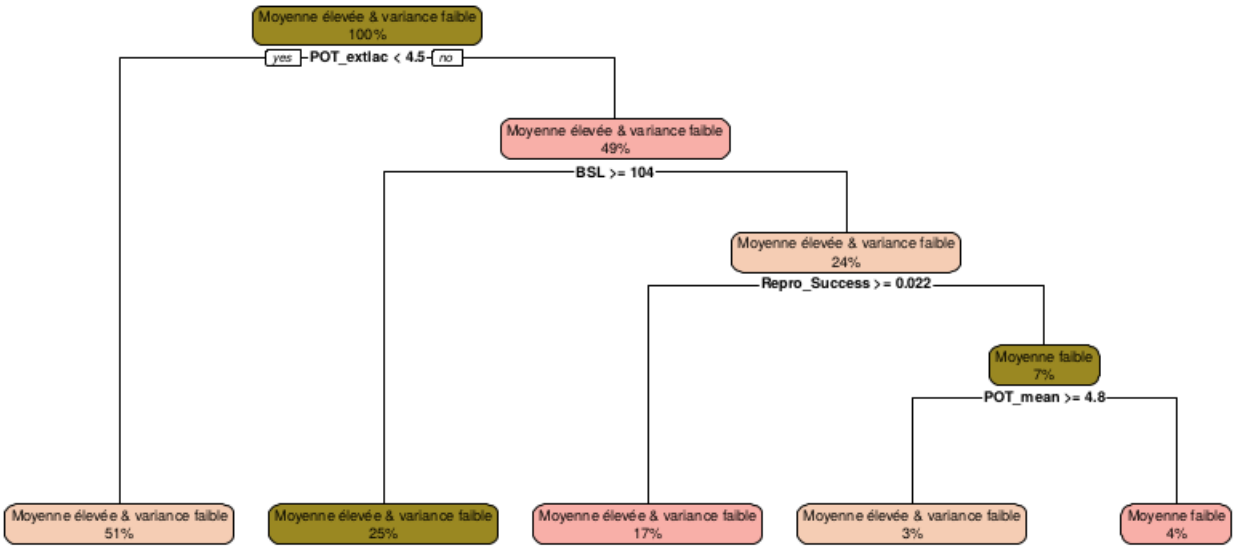


FIGURE 2.4 – Classification pour l'effectif de chèvres en lactation

Nous pouvons déduire de cet arbre qu'un peu plus de 50% des points présentent un effectif élevé et une variance faible et que ce résultat découle entièrement de l'influence du critère de sélection pour la mise en lactation longue. Ce résultat paraît cohérent avec la réalité. Il pourrait être expliqué comme ceci : si ce critère de selection est faible, pratiquement toutes les chèvres sont gardées au sein du troupeau et ainsi l'effectif de chèvre en lactation augmente.
On voit ensuite que le second paramètre influant est la durée de la période de reproduction. Si, en plus d'avoir un critère de sélection un peu strict ($= 4.5$) pour les chèvres non gestantes, cette durée est supérieure ou égale à 104 jours, alors là encore on aura un effectif élevé.
On peut continuer l'interprétation ainsi.

Production et consommation du fourrage 3 Voici la répartition réalisée en fonction des valeurs de production laitière et de consommation du fourrage 3.
Les frontières ont été définies par rapport aux valeurs de référence.

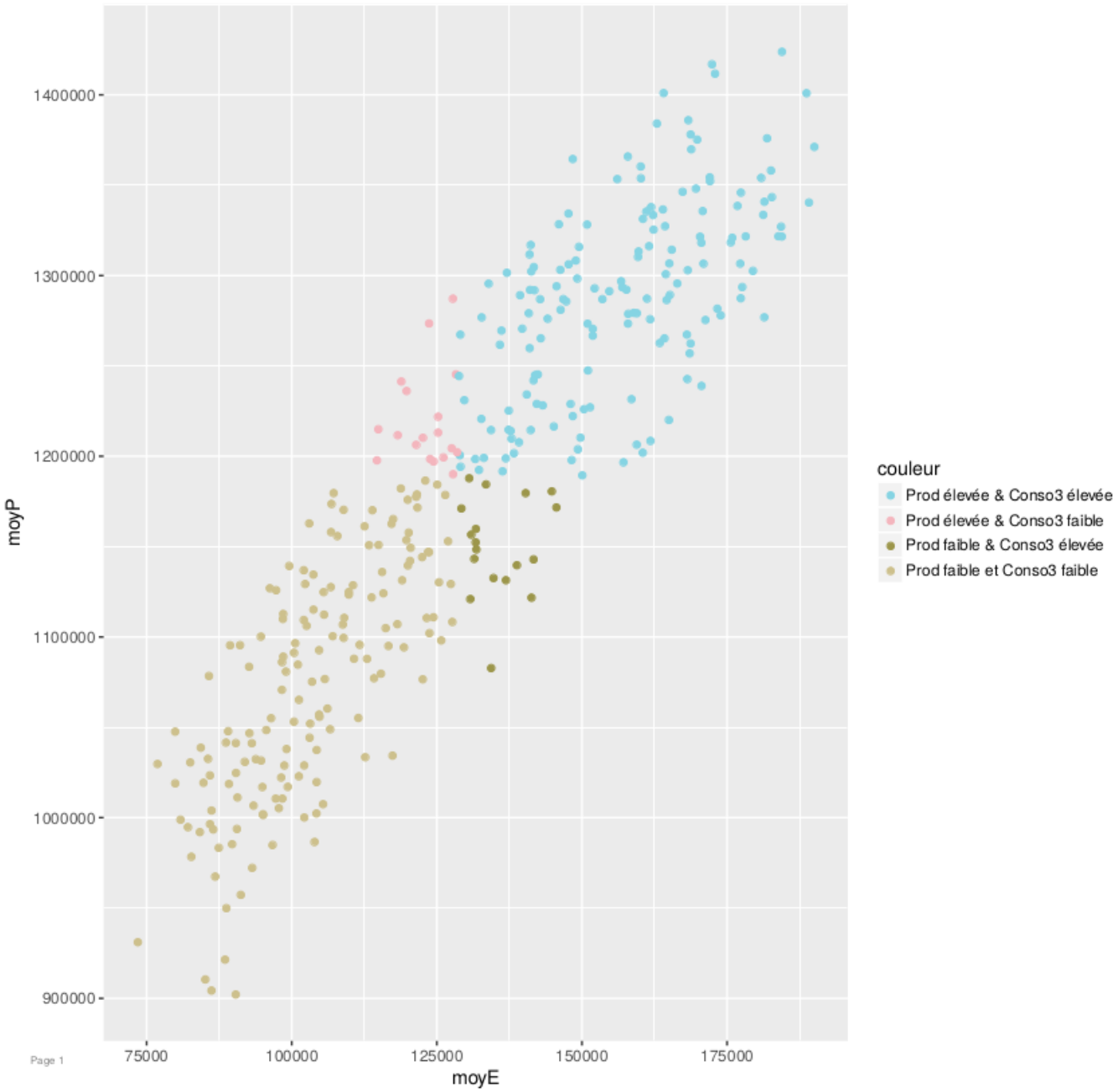


FIGURE 2.5 – Groupe pour la production et la consommation du fourrage 3 - moyenne de référence production à 1.190.000 moyenne de référence conos3 à 128.000

A partir de cela nous avons créé l'arbre de classification correspondant.

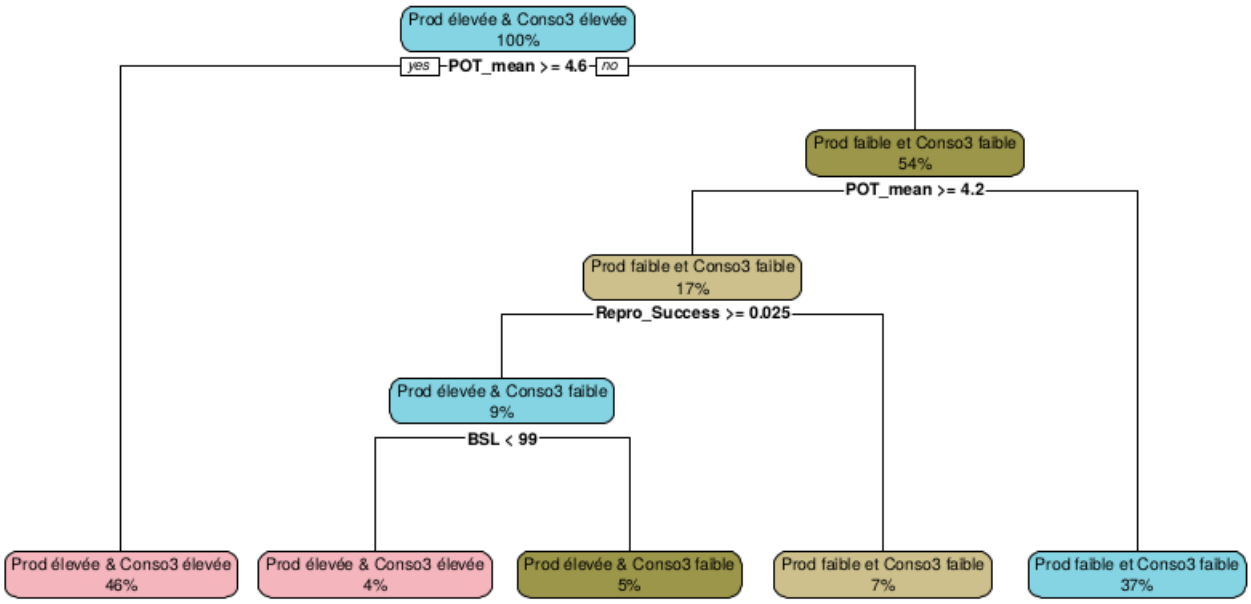


FIGURE 2.6 – Classification pour la production et la consommation du fourrage 3

En ce qui concerne l'interprétation que l'on peut apporter, on constate que le potentiel laitier est un paramètre décisif, s'il est supérieur ou égale à 4.6 on aura une production et une consommation de ce fourrage qui seront élevées, en revanche s'il est inférieur à 4.2, on aura des quantités faibles. Entre les deux, ce sont les paramètres correspondants à la probabilité de réussite à la reproduction et la durée de la période de reproduction qui rentent en jeu.

Il existe des inconvénients liés à l'utilisation de ces arbres. Le premier que nous avons remarqué est que la classification faite est intégralement liée à la définition des seuils de référence donc elle est sensible à la moindre perturbation ce ces derniers. Pour notre premier exemple concernant l'effectif, la valeur de seuil "Moyenne élevée / moyenne faible" est fixé à 232 chèvres. En modifiant que très légèrement ce seuil on pourrait avoir des résultats très différents. Un deuxième inconvénient est que la méthode peut être peu robuste car elle sélectionne pas à pas les variables. En prenant par exemple ces dernières dans l'ordre inverse on aurait là aussi pu avoir un résultat complétement différent.

2.6 Corrélation entre les sorties

Nous nous sommes aussi demandé si certaines sorties étaient fortement corrélées.

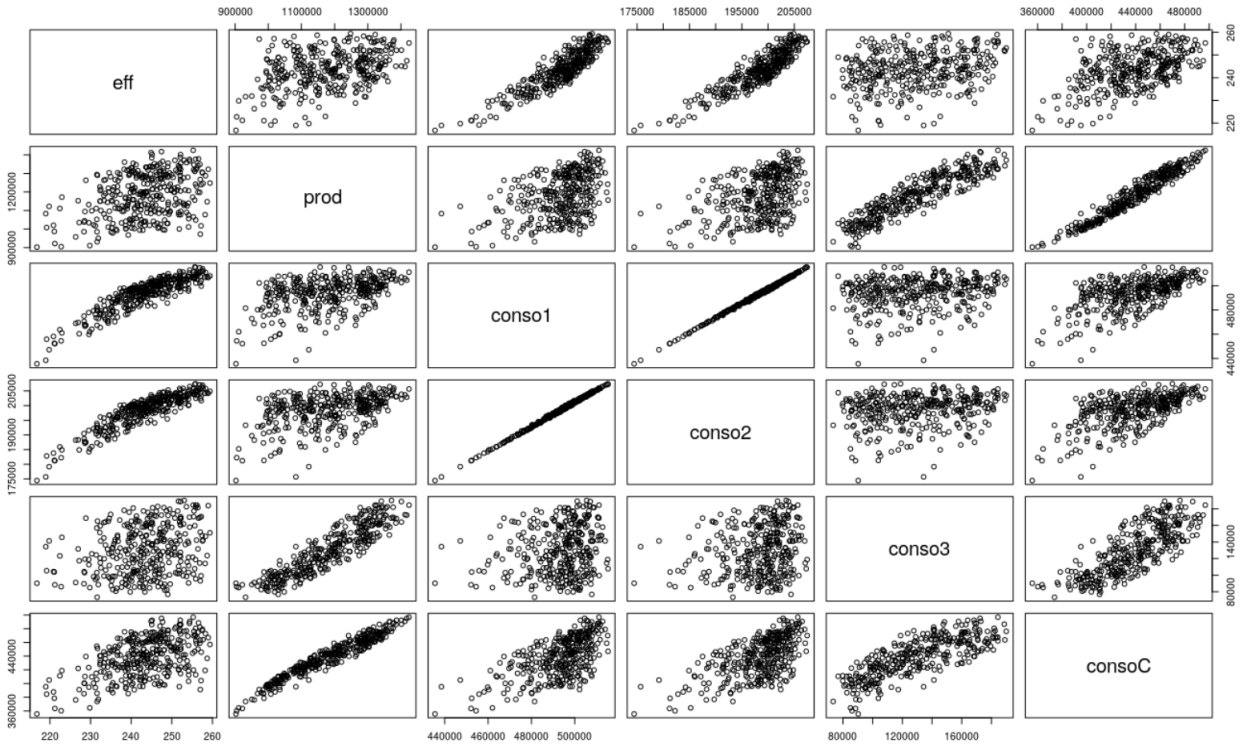


FIGURE 2.7 – Représentation des sorties les unes en fonction des autres

A partir du graphique 2.7, on en déduit que les consommations des fourrages 1 et 2 sont très fortement corrélés. A partir de cette remarque, on peut s'attendre à obtenir des résultats de l'analyse de sensibilité très proches pour ces deux sorties.

Chapitre 3

Analyse de sensibilité

La première partie concernant l'exploration du modèle a permi entre autre de modifier les plages de variation des entrées. C'est avec ces nouvelles valeurs que sera faite l'analyse de sensibilité.

Rappelons les notations introduites précédemment :

- E est le domaine à explorer :
  $E = [3.5;5.5] \times [0.205;0.612] \times [84;126] \times [0.019;0.0309] \times [1.99;7.01] \times [0.33;0.45] \times [0;1]$

- d = 7 est le nombre de paramètres étudiés

- On note X le plan d'expériences de dimensions $N \times d$, $(X \subset E)$

$$X = (x_1, ..., x_N)^T = \begin{bmatrix} x_{1,1} \cdots x_{1,d} \\ \vdots \ddots \vdots \\ x_{N,1} \cdots x_{N,d} \end{bmatrix}$$

- Y est le vecteur contenant les résultats des simulations (Y pourra être le vecteur des sorties de production, des effectifs ou des différentes consommations).

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

  On a donc y_i la sortie correspondant au point x_i.

3.1 Définition

Le but de cette partie est d'identifier les facteurs (les entrées) auxquels le modèle est sensible (cela peut être des effets directs de certains facteurs ou des interactions entre plusieurs facteurs).

Il existe deux types d'analyse de sensibilité :

- l'analyse de sensibilité **locale**, qui évalue quantitativement comment de petites perturbations autour d'une valeur de variable d'entrée se répercutent sur la sortie
- l'analyse de sensibilité **globale**, qui concerne l'intégralité du domaine de variation des entrées. Elle peut être utilisée pour plusieurs raisons : mieux comprendre la relation entre les variables d'entrée et de sortie, identifer les variables d'entrée les plus et les moins influentes, déterminer les variables d'entrée qui interagissent entre elles.

C'est cette dernière qui nous intéresse.

Parmi les méthodes d'analyse de sensibilité globale, on peut en distinguer deux groupes : les **méthodes de criblage** et les **méthodes de décomposition de la variance**. Ce sont toutes les deux des analyses de la variabilité de la sortie par rapport aux variables d'entrée. La première est une analyse qualitative et la seconde est une analyse quantitative.

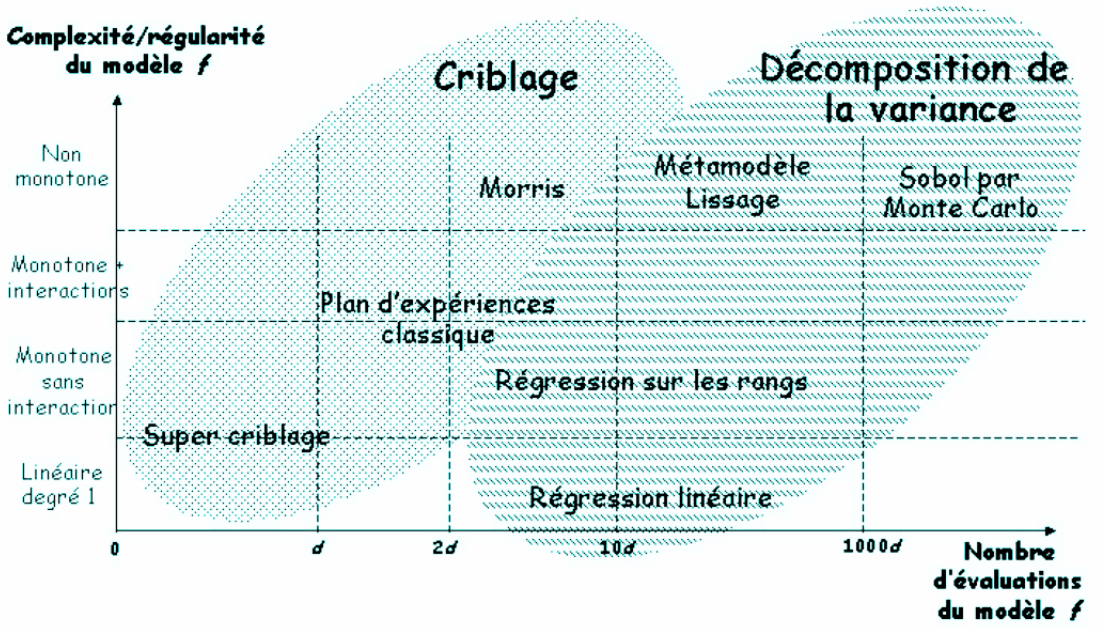


FIGURE 3.1 – Classification des méthodes d'analyse de sensibilité globale [3]

Nous avons à notre disposition une base de tests de 350 évaluations du modèle (obtenue à la suite de la première partie de l'étude) et un modèle qui est non monotone. En se référant à la figure 3.1, on voit assez rapidement que l'on va procéder à la mise en place d'un métamodèle. Ceci sera abordé un peu plus tard, dans la partie 4.

Restons juste sur le fait que nous allons utiliser une méthode de décomposition de la variance et qu'en tant que méthode d'analyse quantitative, elle vise à déterminer des indices de sensibilité.

3.2 Indices de sensibilité

3.2.1 Définition

Pour identifier les facteurs auxquels le modèle est sensible, on calcule des indices de sensibilité, compris entre 0 et 1. Ces indices sont définis comme une mesure de l'importance d'une variable d'entrée du modèle sur la variabilité de la sortie du modèle. Ils représentent la proportion de variance expliquée. Il existe différents types d'indices de sensibilité, nous présenterons ici les indices qui ont été utilisés dans cette étude : les indices principaux et les indices totaux . Ce sont les indices de sensibilité les plus courants.

- **Indice principal** (ou indice du premier ordre)
  Les indices de sensibilité du premier ordre sont associés aux effets principaux. En effet, ils ne prennent en compte la sensibilité de la sortie Y que sous l'effet de la variable V_i seule.
  Ces indices s'écrivent :

$$SI_i = \frac{Var[E(Y|V_i)]}{Var(Y)} = 1 - \frac{E[Var(Y|V_i)]}{Var(Y)}, \ i \in \{1,...,d\}.$$

  Plus cet indice est proche de 1, plus la variable V_i aura de l'effet sur Y. Ces indices permettent donc de classer les variables par ordre d'influence sur Y.

- **Indice total**
  Les indices de sensibilité totaux prennent en compte les intéractions de la variable V_i avec les autres variables V_j ($j \neq i$) ainsi que l'influence de la variable V_i seule sur Y.
  Ces indices s'écrivent :

$$TSI_i = \frac{E[Var(Y|V_{-i})]}{Var(Y)} = 1 - \frac{Var[E(Y|V_{-i})]}{Var(Y)}, \ i \in \{1,...,d\},$$

  où V_{-i} représente l'ensemble des variables à l'exclusion de V_i : $V_{-i} = \{V_1,...,V_{i-1},V_{i+1},...,V_d\}$.
  Les indices de sensibilité totaux permettent d'identifier les variables qui peuvent être fixées arbitrairement sans changer de beaucoup le comportement du modèle. Les indices totaux correspondant à ces variables sont très faibles.

Il est souvent difficile de déterminer analytiquement ces indices, ils sont donc généralement estimés de façon numérique.

3.2.2 Estimation des indices de sensibilité

Il existe différentes méthodes d'estimation des indices de sensibilité. Les plus connues sont les méthodes de Sobol, de Morris et FAST. Ces méthodes diffèrent notamment par les hypothèses et le nombres de simulations qu'elles nécessitent.
N'ayons que très peu d'informations sur notre modèle, nous avons opté pour une des méthodes de Sobol. En effet, la seule hypothèse requise par ces dernières est que la variance et l'espérance de la sortie soient finies.

- Estimation des indices de sensibilité par *SobolJansen* .
  Voici les grandes lignes de la méthode de Sobol évoquée précédemment :

  1. Tirage aléatoire de deux matrices A et B de taille $N \times d$ (pour notre étude, elles ont été déterminées avec une méthode d'échantillonnage par hypercubes latins (présentée section 2.4)) :

$$A = \begin{bmatrix} A_{1,1} \cdots A_{1,d} \\ \vdots \ddots \vdots \\ A_{N,1} \cdots A_{N,d} \end{bmatrix}, \quad B = \begin{bmatrix} B_{1,1} \cdots B_{1,d} \\ \vdots \ddots \vdots \\ B_{N,1} \cdots B_{N,d} \end{bmatrix}$$

  2. Création de d nouvelles matrices par mélange des matrices A et B :

$$C_i = \begin{bmatrix} A_{1,1} & \cdots & A_{1,i-1} & B_{1,i} & A_{1,i+1} & \cdots & A_{1,d} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ A_{N,1} & \cdots & A_{N,i-1} & B_{N,i} & A_{N,i+1} & \cdots & A_{N,d} \end{bmatrix}, \; i \in \{1,...,d\}$$

     C_i correspond à la matrice A dans laquelle la i^{eme} colonne a été remplacée par la i^{eme} colonne de la matrice B.

  3. Calcul des sorties pour les $(d+2)$ matrices créées aux étapes précédentes. On obtient donc $N \times (d+2)$ sorties, représentées dans $(d+2)$ vecteurs :

$$Y^A = \begin{bmatrix} {y^A}_1 \\ \vdots \\ {y^A}_N \end{bmatrix}, \; Y^B = \begin{bmatrix} {y^B}_1 \\ \vdots \\ {y^B}_N \end{bmatrix}, \; Y^{C_i} = \begin{bmatrix} {y^{C_i}}_1 \\ \vdots \\ {y^{C_i}}_N \end{bmatrix}, \; i \in \{1,...,d\}.$$

  4. Calcul de la moyenne empirique $\hat{m}$ du vecteur $(Y^A, Y^B, Y^{C_1}, ..., Y^{C_N})$ et de la moyenne $\hat{\sigma}^2$ des variances empiriques de Y^A,de Y^B,de Y^{C_1},...,de Y^{C_N}
     On a donc :

$$\hat{m} = \frac{1}{d+2}(Y^A + Y^B + Y^{C_1} + ... + Y^{C_N}),$$

$$\hat{\sigma}^2 = \frac{1}{d+2}[Var(Y^A) + Var(Y^B) + Var(Y^{C_1}) + ... + Var(Y^{C_N})].$$

  5. Pour chaque entrée V_i , $i \in \{1,...,d\}$, estimation de l'indice principal et de l'indice total :

$$\hat{SI}_i = \frac{\frac{1}{N} < Y^A, Y^{C_i} > - \hat{m}}{\hat{\sigma}^2},$$

     avec $< u, v >= \sum_{j=1}^{N} u_j v_j$

$$\hat{TS}_i = \frac{||Y^A - Y^{C_i}||^2}{2N\hat{\sigma}^2},$$

     avec $||u-v||^2 = \sum_{j=1}^{N} (u_j - v_j)^2$

- Bootstrap : précision des estimations. La méthode de bootstrap permet de déterminer la précision de l'estimation des indices de sensibilité. En R, elle est représentée par le paramètre *nboot* de la fonction d'estimation des indices de sensibilité *sobolJansen* [2].

Nous avons vu que la méthode de *sobolJansen* est assez simple d'utilisation : elle nécessite un faible nombre d'hypothèses et la précision de ses prédictions peut être calculée. Mais en contrepartie, cette méthode exige un grand nombre d'appels au modèle ($N(d+2)$ simulations, avec N au alentours de 10.000) pour qu'elle soit précise. Or, comme vu dans la section 1.5, notre modèle est très coûteux. Nous ne pouvons donc pas appliquer directement cette méthode à notre modèle.
En réalité, ceci n'est pas un problème puisque d'après la section 3.1 nous allons passer par un métamodèle.

Chapitre 4

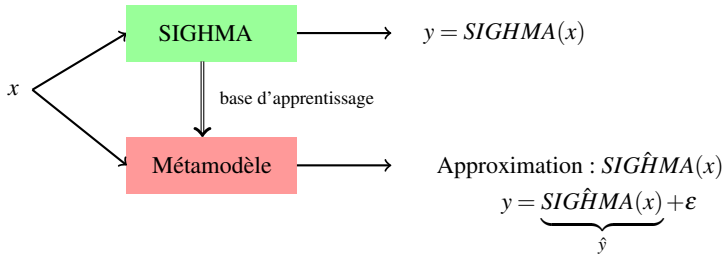
Métamodélisation - Krigeage

4.1 Définition

Un métamodèle (ou surface de réponse) peut être vu comme un modèle du modèle initial. A partir de ce nouveau modèle (plus simple), une prédiction de la réponse du modèle initial peut être obtenue en tout point de l'espace, beaucoup plus rapidemment.
Pour construire ce métamodèle, on génère un plan d'expériences X, de taille N, puis on simule les N sorties $(y_1,...,y_N) = Y$ de ces points par notre modèle initial.

$$X = \begin{bmatrix} x_{1,1} \cdots x_{1,d} \\ \vdots \ddots \vdots \\ x_{N,1} \cdots x_{N,d} \end{bmatrix}, Y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

A partir de cette matrice et de ce vecteur, qui forment notre base d'apprentissage, le métamodèle est construit.



Auparavant pour un point x, nous avions une sortie simulée par notre modèle, y , mais avec un temps de calcul très lent. Maintenant nous avons grâce à notre métamodèle, une sortie prédite, $\hat{y}$, avec un temps de calcul très court.
La sortie c'est cette prédiction majorée d'une erreur d'approximation ε.

4.2 Pourquoi choisir un métamodèle dans notre cas

Nous utilisons les métamodèles dans le cas où le modèle initial ne présente pas d'expression analytique simple, ou encore lorsque le temps de simulation par ce dernier est trop lent, mais aussi, comme vu dans la figure 3.1, pour réaliser une analyse de sensibilité. Avec notre modèle nous sommes dans tous ces cas, mais nous allons surtout utiliser un métamodèle dans le but de réaliser une analyse de sensibilité.

4.3 Principe

4.3.1 Différents types de métamodèle

Il existe différents types de métamodèle, comme par exemple les modèles de regression paramétriques, les modèles de regression non paramétriques et les méthodes de prédiction par processus gaussiens.

Les modèles linéaires paramétriques sont souvent utilisés lorsque l'on dispose d'une hypothèse forte de linéarité sur le domaine de variation des facteurs, mais aussi avec très peu d'intéractions entre ces derniers.
D'après la partie sur l'exploration du modèle (chapitre 2), il semblerait que ça ne soit pas notre cas.

En ce qui concerne les métamodèles basés sur la régression non-paramétrique, ils permettent de suivre des comportements locaux particuliers. Cependant ils nécessitent un plan d'expériences de grande taille. Comme vu dans la partie 1.5, notre simulateur est très coûteux en temps de calcul, ainsi mettre en place un gros plan d'expériences serait donc compliqué.

Concernant les métamodèles basés sur les processus gaussiens, ils sont très efficaces pour de petits plans d'expériences et la seule hypothèse pour la construction de ces métamodèles repose sur une corrélation entre les comportements en les points du domaine. Au vu de ces caractéristiques, nous avons choisi ce type de métamodèle pour notre étude.

Le principe des processus gaussiens est de considérer la valeur de la sortie d'un nouveau point x' comme une combinaison linéaire des valeurs simulées des points $x_1, ..., x_N$ appartenant au plan d'expériences.
La formule de prédiction est de la forme :

$$\hat{y} = \sum_{i=1}^{N} \beta_i(x') \; SIGHMA(x_i)$$

On obtient donc :

$$\hat{y} = \hat{SIGHMA}(x')$$

avec $\hat{SIGHMA}$ notre modèle de prédiction : c'est-à-dire notre métamodèle.

Les paramètres β_i, qui correspondent au poids de x_i dans la prédiction de la valeur de la sortie pour le nouveau point x', sont estimés de façon à ce que la prédiction soit sans biais (1) et que la variance de prédiction soit minimale (2) quelque soit x'. C'est-à-dire que :

$$E[\hat{SIGHMA}(x') - SIGHMA(x')] = 0 \quad (1)$$

$$Var[\hat{SIGHMA}(x') - SIGHMA(x')] = \min_{\beta} Var[\sum_{i=1}^{N} \beta_i(x')\ SIGHMA(x_i) - SIGHMA(x')] \quad (2)$$

Pour cela, on s'intéresse aux corrélations entre les observations en deux points distincts. La modélisation par processus gaussien exprime ce lien par le biais d'une structure de covariance entre les valeurs observées. La prédiction en tout point x' se fait par une formule d'interpolation exacte avec la formule du meilleur prédicteur linéaire sans biais (BLUP, best linear unbiased predictor) construit à partir des observations et de la structure de covariance estimée (section 4.3.2).
Le modèle de base du krigeage s'écrit :

$$SIGHMA(x) = \mu(x) + \delta(x),$$

où μ est la fonction moyenne et δ un processus gaussien centré stationnaire de variance σ^2 et de fonction de corrélation $R : Cov[\delta(x_i), \delta(x_j)] = \sigma^2 R(x_i - x_j)$

4.3.2 Fonctions de corrélation

Comme évoqué précédemment, les processus gaussiens sont entièrement caractérisés par leur moyenne μ et leur fonction de corrélation *R*. La fonction de covariance (aussi appelée noyau) s'exprime comme un facteur σ^2 multiplié par la fonction de corrélation R. Ainsi, parmi les métamodèles basés sur les processus gaussiens, là encore il existe de nombreuses possibilités et le point le plus important réside dans le choix de cette fonction de corrélation.

Comme nous le verrons dans la partie 4.4.2, différentes fonctions d'auto-corrélation ont été testées. Nous présenterons ici l'exponentielle, la matern 5/2, la matern 3/2 et la gaussienne.
La principale différence entre ces fonctions réside dans leur régularité.

- Gaussienne : Cette fonction est C^∞.

$$k(x,y) = \sigma^2 \exp(-\frac{1}{2}||x-y||_\theta^2)$$

- Matern 3/2 : Cette fonction d'auto-corrélation est très courante. Elle est dérivable deux fois.

$$k(x,y) = \sigma^2(1+\sqrt{3}||x-y||_\theta)\exp(-\sqrt{3}||x-y||_\theta)$$

- Matern 5/2 : Cette fonction d'auto-corrélation est la plus courante. Elle est aussi dérivable deux fois.

$$k(x,y) = \sigma^2(1+\sqrt{5}||x-y||_\theta + \frac{5}{3}||x-y||_\theta^2)\exp(-\sqrt{5}||x-y||_\theta)$$

- Exponentielle : Cette fonction est continue est non dérivable.

$$k(x,y) = \sigma^2 \exp(-||x-y||_\theta)$$

avec :
$||x-y||_\theta = (\sum_{i=1}^{d} \frac{(x_i-y_i)^2}{\theta_i^2})^{\frac{1}{2}}$,
σ^2 = paramètre de variance,
θ = paramètre de portée.

Pour déterminer celle qui correspondrait le mieux à notre étude, nous les avons toutes mises en place, puis nous avons réalisé différents tests de validations, qui seront présentés dans la partie 4.4.2.

Les paramètres σ^2 et θ sont déterminés par maximum de vraisemblance.
Cette estimation est réalisée directement dans la fonction *km* qui sert à construire le métamodèle. Le paramètre *multistart* de cette fonction correspond au nombre d'itérations réalisées en partant d'une nouvelle graine 0à chaque fois. Les valeurs des paramètres σ^2 et θ sont ensuite déterminées en prennant le meilleur résultat. Cette fonction utilise une optimisation par méthode BFGS [1].

4.3.3 Tendance

Comme nous l'avons vu, un des paramètres qui entre en jeu lors de la création d'un métamodèle est la tendance, ou encore appelée fonction moyenne, μ. La tendance correspond à la valeur autour de laquelle oscille le processus.
Cette tendance détermine le krigeage (construction d'un métamodèle par processus gaussien) qui sera utilisé. Il existe différents types de krigeage.

- le krigeage simple : il s'utilise lorsque la tendance n'existe pas ou est une constante connue : $\mu(x) = m$
- le krigeage ordinaire : il s'emploie lorsque la tendance est une constante inconnue : $\mu(x) = \mu$
- le krigeage universel : on se sert de celui-ci lorsque la tendance s'exprime comme une combinaison linéaire de fonctions $(f_1, ..., f_d)$: $\mu(x) = \sum_{j=1}^{d} f_j(x)\beta_j$

4.3.4 Estimation des indices de sensibilité

Maintenant que le métamodèle est en place nous pouvons réaliser l'estimation des indices de sensibilité. Nous avons désormais deux façons de le faire : avec la fonction *sobolJansen* (vue section 3.2.2) ou encore, maintenant que nous travaillons avec un métamodèle par processus gaussiens, la fonction *sobolGP*.

- Estimation des indices de sensibilité par *sobolGP*
  La méthode utilisée par la fonction *sobolGP* est la même que pour *sobolJansen*. Cette fonction prend en compte en plus le fait que l'on travaille avec un métamodèle par processus gaussiens et donc calcule aussi les erreurs liées à ce métamodèle. Le but de cette fonction est d'effectuer une analyse de sensibilité globale basée sur le krigeage en prenant en compte à la fois le métamodèle et les erreurs de Monte-Carlo. En effet, les indices de Sobol sont estimés avec une intégration de Monte-Carlo et le modèle est remplacé par un métamodèle : le modèle de krigeage.

1. La méthode de Broyden-Fletcher-Goldfarb-Shanno (BFGS) est une méthode permettant de résoudre un problème d'optimisation non linéaire sans contraintes.

- Bootstrap : précision des estimations. La méthode de bootstrap vue précédemment (dans la section 3.2.2), est aussi utilisée avec cette fonction et est toujours prise en compte par le paramètre *nboot* de la fonction d'estimation des indices de sensibilité *sobolGP* [1].

4.4 Mes résultats

4.4.1 Base d'apprentissage

Comme vu précédemment, nous avons utilisé un métamodèle par processus gaussiens et ces modèles sont contruits par apprentissage.
La construction de cette base d'apprentissage se fait en deux étapes.

La première consiste en la création d'un ensemble de points. Nous avons aussi vu que cette méthode ne nécéssite qu'une petite base d'apprentissage. Nous avons donc généré un ensemble de 350 points par la méthode des Hypercubes Latins optimisés. Nous avons donc obtenu une matrice de taille 350×7.

Nous avons ensuite simulé la réponse de SIGHMA pour ces 350 points.

$$X = \begin{bmatrix} x_{1,1} \cdots x_{1,7} \\ \vdots \ddots \vdots \\ x_{350,1} \cdots x_{350,7} \end{bmatrix}, Y = \begin{bmatrix} y_1 \\ \vdots \\ y_{350} \end{bmatrix}$$

Ceci forme notre base d'apprentissage : c'est sur cette base que le métamodèle va être construit.

4.4.2 Comparaison des différents métamodèles

Nous avons construit plusieurs métamodèles pour la prédiction des différentes sorties. Nous présenterons ici les résultats concernant la sortie "production de lait". Nous avons soumis les différents métamodèles à trois types de validation : une validation graphique, une validation Q2 sur la base de tests et une validation Q2 par validation croisée.

- Validation graphique.
  Sur les graphes suivants, nous pouvons voir les observations prédites (en rouge), les observations simulées (en bleu) ainsi que leurs intervalles de confiance respectifs. Ce n'est bien sûr pas une méthode quantitative et ce n'est pas à partir de ces résultats que les décisions ont été prises, mais cela a servi de première approche de vérification afin de voir si globalement cela semblait cohérent.

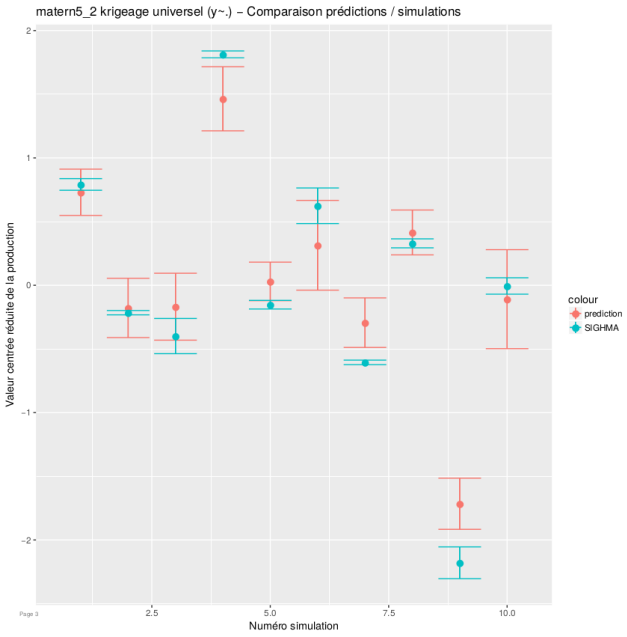


FIGURE 4.1 – Comparaison pédiction Matern5/2 et simulation par SIGHMA

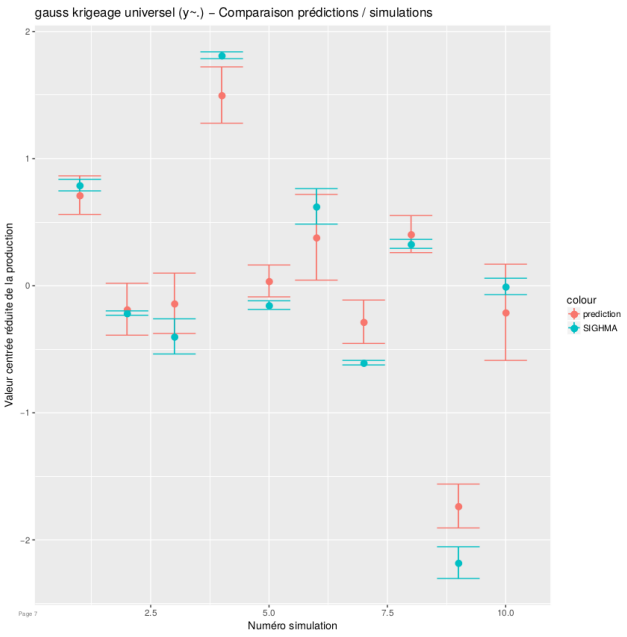


FIGURE 4.2 – Comparaison pédiction Gauss et simulation par SIGHMA

En terme de convenance, sur ces deux graphiques, les deux prédictions semblent équivalentes. On distingue que de très légères différences, comme par exemple que les intervalles de confiance de la prédiction faite avec Gauss sont légèrement plus petits que ceux issus de la prédiction avec Matern5/2.

- Validation Q2 sur la base de test.
  La validation Q2 est une validation quantitative, elle consiste au calcul d'un coefficient de prédictivité.

  Voici la formule de la validation Q2 :

$$Q^2 = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}$$

  avec y la simulation sur SIGHMA et $\hat{y}$ la prédiction faite avec $\hat{SIGHMA}$, le métamodèle.

  Ce coefficient approxime 1 - l'écart quadratique normalisé entre les prédictions et les valeurs réelles.

  Plus cette quantité est proche de 1, plus l'écart est quant à lui petit et donc meilleur est l'ajustement du métamodèle aux observations.

  Pour la validation Q2, on considère qu'une valeur supérieure à 0.90, est signe d'un modèle précis, et supérieure à 0.95 d'un modèle très précis. Il est à noter que puisque nos données présentent du bruit (du fait de travailler sur les moyennes de production de lait [dans ce cas précis] et non sur la production directement), on ne peut que se rapprocher d'un Q2 valant 1, mais nous ne pouvons pas atteindre cette valeur.

- Indicateurs Q2 par validation croisée.
  La méthode de validation croisée est une méthode d'estimation de fiabilité d'un modèle fondée sur une

technique d'échantillonnage.
Le principe de cette méthode est que l'on divise l'échantillon de départ en k échantillons, puis on sélectionne un de ces k échantillons comme ensemble de validation et les $k-1$ autres échantillons constituent l'ensemble d'apprentissage. On calcule le score de performance, puis on répète l'opération en sélectionnant un autre échantillon de validation parmi les k-1 échantillons qui n'ont pas encore été utilisés pour la validation du modèle. L'opération se répète ainsi k fois pour qu'à la fin chaque sous-échantillon ait été utilisé exactement une fois comme ensemble de validation. La moyenne des k erreurs quadratiques moyennes est enfin calculée pour estimer l'erreur de prédiction.

En ce qui concerne notre étude, nous avons utilisé une méthode de validation croisée LOO (Leave One Out), c'est-à-dire que l'échantillon de validation ne contient qu'un seul point et donc l'ensemble d'apprentissage contient tout le reste de l'échantillon de départ, c'est-à-dire 349 points.

Voici un tableau regroupant les résultats énoncés précédemment : avec y˜1 correspondant au krigeage simple

	Matern 5/2		Exp		Gauss	
	y˜1	y˜.	y˜1	y˜.	y˜1	y˜.
Validation graphique	correcte	correcte	correcte	correcte	correcte	correcte
Validation Q2	0.9307	0.9344	0.8839	0.8965	0.9190	0.9366
Validation croisée + Q2	0.9626	0.9707	0.9334	0.9469	0.9578	0.9700

TABLE 4.1 – Tableau de comparaison des métamodèles créés

et y˜ représentant le krigeage universel.

Les valeurs des indicateurs Q2 calculés sur la base de tests ne permettent pas de prendre une décision puisqu'ils ont été calculés sur un échantillon de seulement dix simulations. En revanche les indicateurs Q2 calculés par validation croisée sont eux significatifs.

Nous remarquons que les résultats obtenus en utilisant les méthodes de Gauss et Matern5/2, sont presque équivalents. La méthode de Gauss est légèrement meilleure en ce qui concerne les validations Q2. En revanche, pour des raisons de stabilité numérique, il serait recommandé de plutôt utiliser la méthode Matern 5/2, mais dans notre cas, nos observations sont bruitées, cet argument ne rentre donc pas dans nos critères de selection.
Nous avons ainsi choisi le métamodèle avec fonction de correlation de Gauss pour nos calculs d'indices de sensibilité.

4.4.3 Indices de sensibilité

Comme vu dans la partie 4.3.4, nous avons utilisé deux méthodes pour estimer ces indices : la fonction *sobolJansen* et *sobolGP*.
Voici les résultats obtenus pour la sortie de production laitière :

Nous avons d'abord utilisé la fonction *sobolJansen*.

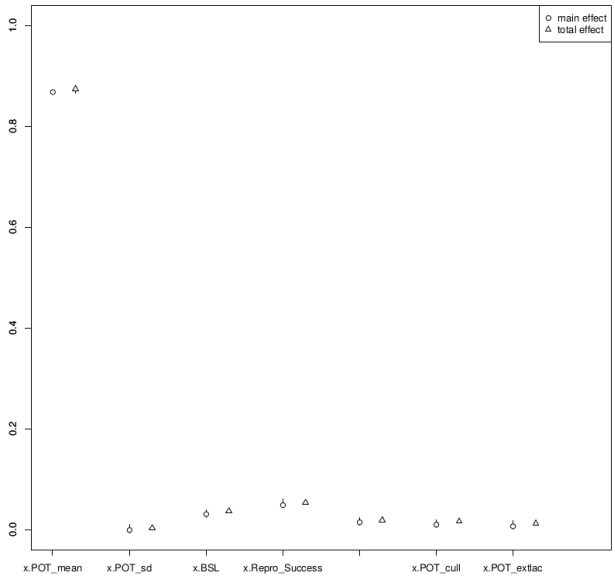


FIGURE 4.3 – Indices de sensibilité pour la sortie Production laitière

Sur ce graphique, les symboles en cercle représentent les effets principaux et les symboles en triangle représentent les effets totaux évoqués dans la partie 3.2.1.

Cette méthode ne donne pas d'explication sur les erreurs des indices (représentées par les bâtons). C'est pour cela que nous avons ensuite utilisé la fonction *sobolGP*.

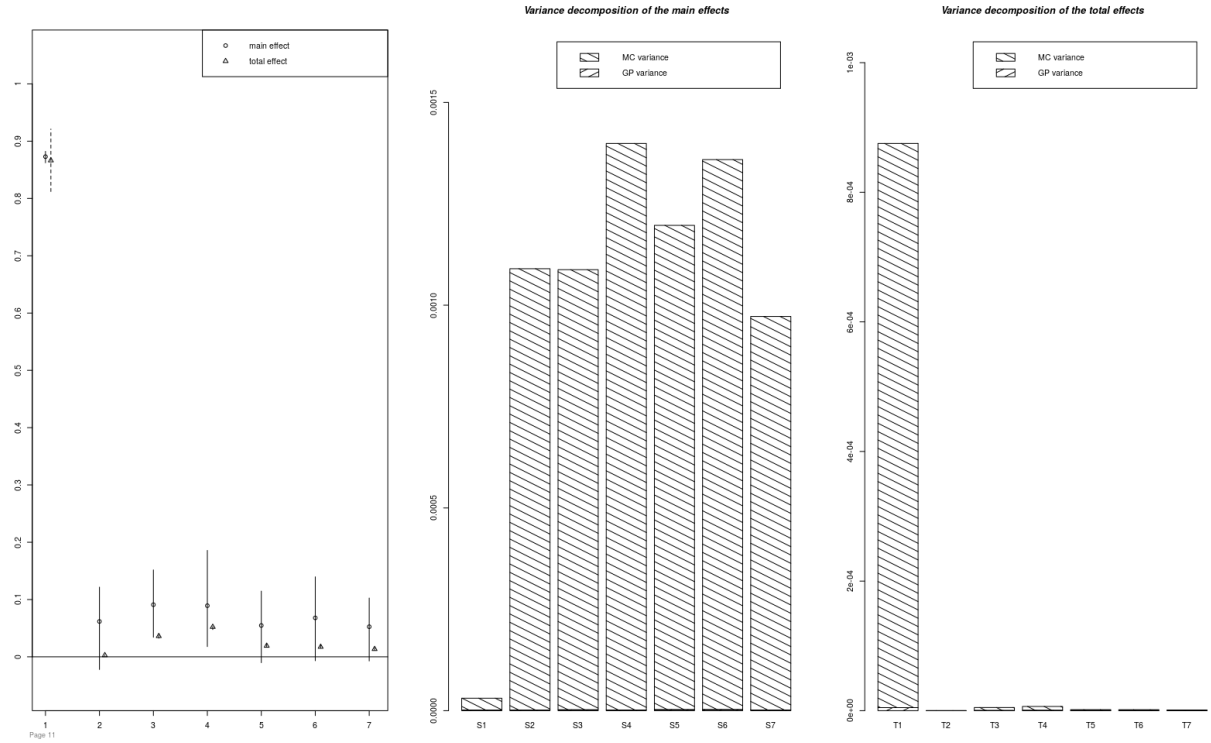


FIGURE 4.4 – Indices de sensibilité (avec prise en compte des erreurs des processus gaussiens) pour la sortie Production laitière

Cette fonction renvoie 3 résultats : les indices de sensibilité, tout comme *sobolJansen*, et deux graphiques expliquant respectivement l'origine des erreurs des indices principaux puis totaux. Cette erreur peut provenir des calculs du métamodèle par processus gaussiens et aussi des calculs par méthode de Monte Carlo.

Ayant réussi à minimiser l'erreur due au métamodèle (comme on peut le voir sur la figure 4.4), nous nous sommes concentré uniquement sur les résultats donnés par la méthode *sobolJansen*, en considérant que l'intégralité des erreurs présentées sur ce graphique était due aux calculs par méthode de Monte Carlo.

Chapitre 5

Conclusion, ouvertures et perspectives

L'intégralité des scripts créés lors de ce stage sont réutilisables afin étudier d'éventuelles sorties supplémentaires, d'autres fichiers de management et d'autres scénarios.

Nous avons généré différents résultats pour différentes sorties, sous différentes formes, pour des publics différents.

L'exploration de modèle a permis d'apporter différentes améliorations au modèle.Comme par exemple changer un des paramètres en continu, relever et corriger un problème d'effectif de chèvres en lactation trop faible, adapter les bornes des entrées et relever un problème de consommation des chèvres adultes vs consommation chevrettes.

En ce qui concerne l'analyse de sensibilité, voici une forme de représentation des résultats trouvés. Représentation des indices de sensibilité sous forme de camemberts (Une autre représentation est disponible dans l'annexe .1). Nous avons choisi ce type de graphique afin d'être les plus explicites possible et de faciliter la communication avec les gens du terrain que nous avons été ammenés à rencontrer lors du déplacement sur la plateforme Patuchev afin de présenter nos résultats.

Production laitière

La production laitière est principalement influencée par le potentiel laitier moyen du troupeau (POT_mean) qui explique 87% de la variabilité. Ce résultat peut s'expliquer : plus un troupeau a des animaux avec un niveau génétique élevé, plus la production sera élevée. Les paramètres de reproduction (BSL et repro_success) expliquent quant à eux deux 8% de la variabilité de la production laitière. Ce résultat est lié au fait que la reproduction pilote le nombre de jours en lactation des animaux et inversement le nombre de jours improductifs sans production. Il n'est donc pas surprenant que ces paramètres expliquent une partie de la variabilité de la production laitière globale du troupeau.

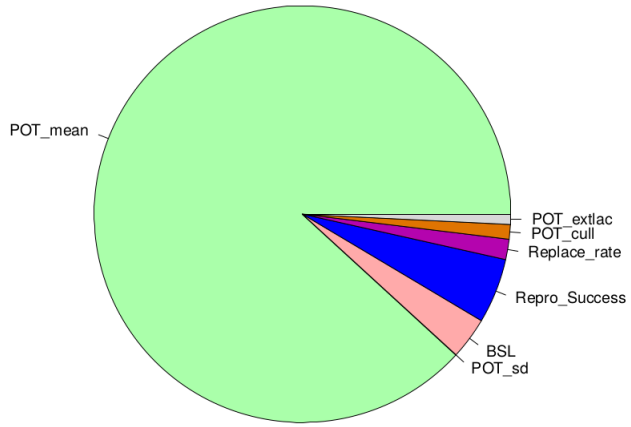


FIGURE 5.1 – Représentation des indices de sensibilité principaux de la production sous forme de diagramme en camembert

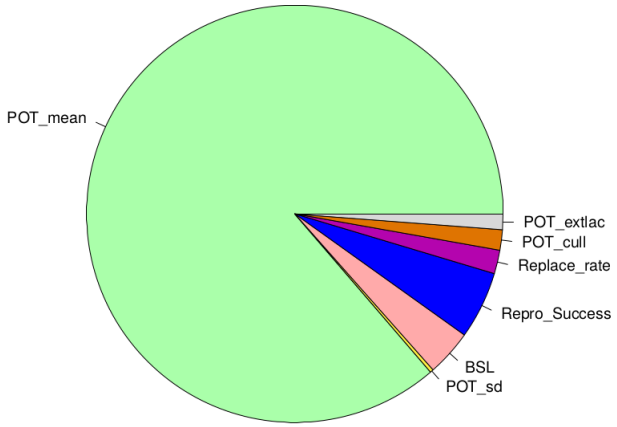


FIGURE 5.2 – Représentation des indices de sensibilité totaux de la production sous forme de diagramme en camembert

Consommation fourrage 3

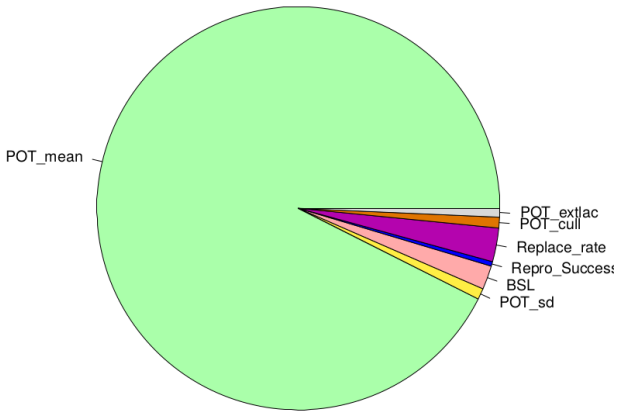


FIGURE 5.3 – Représentation des indices de sensibilité principaux de la consommation de fourrage 3 sous forme de diagramme en camembert

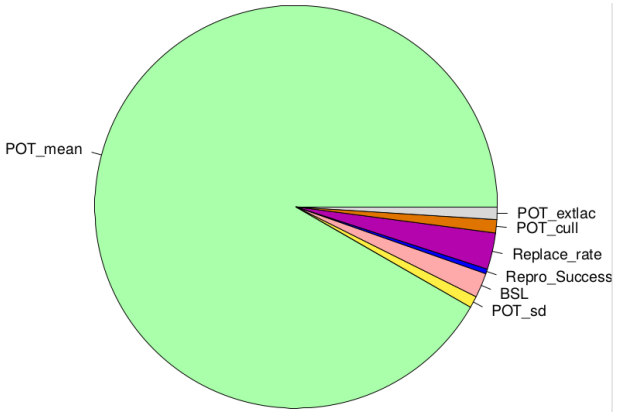


FIGURE 5.4 – Représentation des indices de sensibilité totaux de la consommation de fourrage 3 sous forme de diagramme en camembert

En ce qui concerne les résultats liés à la consommation de fourrage 3, on peut remarquer que certaines caractéristiques ressemblent à celle de la production laitière. La consommation de ce fourrage est principalement influencée par le potentiel laitier moyen du troupeau (POT_mean) qui explique 93% de la variabilité.
En effet, cette consommation reflète la consommation d'un fourrage distribué en quantité non limitée. Le niveau de consommation d'un animal est déterminé par ses besoins énergétiques qui sont eux pilotés par la production laitière de l'animal, elle même fortement déterminée par son potentiel laitier (donc par POT_mean).

Consommation fourrage 1 et consommation du fourrage 2

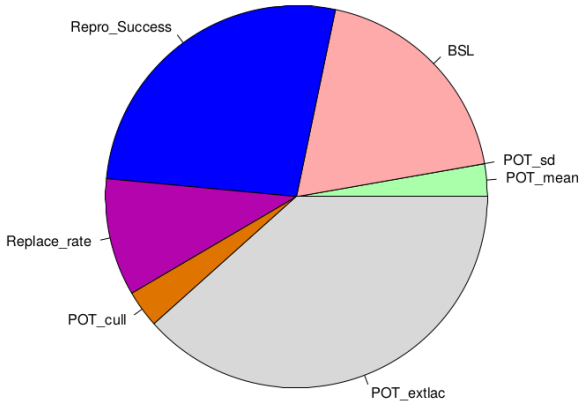


FIGURE 5.5 – Représentation des indices de sensibilité principaux de la consommation de fourrage 1 sous forme de diagramme en camembert

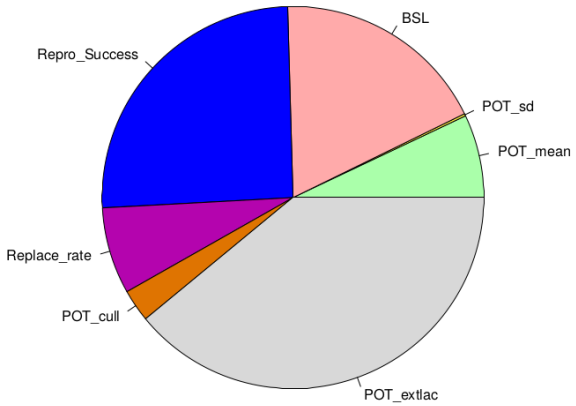


FIGURE 5.6 – Représentation des indices de sensibilité totaux de la consommation de fourrage 1 sous forme de diagramme en camembert

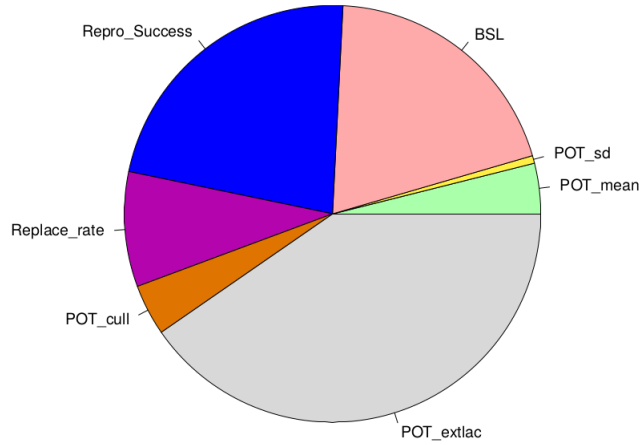


FIGURE 5.7 – Représentation des indices de sensibilité principaux de la consommation de fourrage 2 sous forme de diagramme en camembert

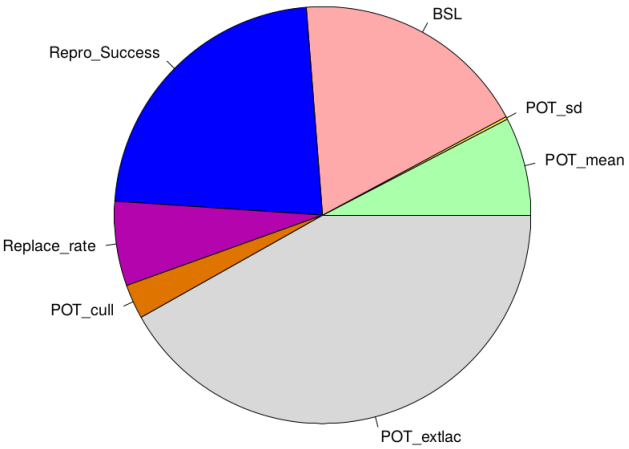


FIGURE 5.8 – Représentation des indices de sensibilité totaux de la consommation de fourrage 2 sous forme de diagramme en camembert

En ce qui concerne la consommation des fourrages 1 et 2 qui sont distribués en quantité fixe à chaque animal. Pour ces deux variables, on obtient pratiquement les mêmes résultats, ce qui est tout à fait cohérent puisque le niveau de consommation ne dépend pas des besoins physiologiques de chaque individu mais plutôt des effectifs et de leur stade de lactation.
On en conclu donc la même chose : les paramètres expliquant la variabilité de la consommation de ces fourrages sont : *POT_extlac* (qui explique un peu plus de 30% de la variabilité de cette consommation),*Repro_Success* (qui en explique presque 30%), *BSL* (qui en explique 20%) et enfin *Replace_rate* (qui en explique 8%).

Effectif de chèvres en lactation

On en déduit que les paramètres leviers du système pour l'effectif des chèvres en lactation sont le taux de

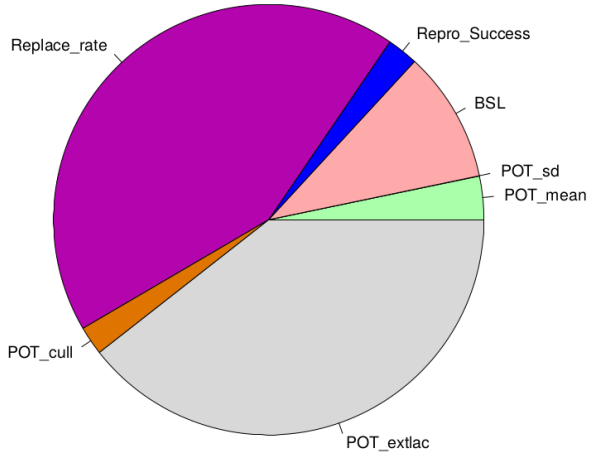


FIGURE 5.9 – Représentation des indices de sensibilité principaux des effectifs de chèvres en lactation sous forme de diagramme en camembert

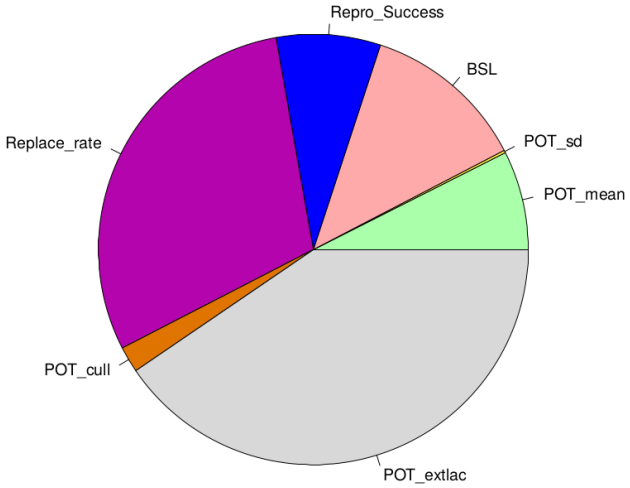


FIGURE 5.10 – Représentation des indices de sensibilité totaux des effectifs de chèvres en lactation sous forme de diagramme en camembert

remplacement (*Replace_rate*), le critère de sélection pour la mise en lactation longue (*POT_extlac*), mais aussi la valeur moyenne de la distribution dans laquelle est tiré le potentiel laitier attribué à une chèvre à la naissance (*POT_mean*), la probabilité journalière de réussité à la reproduction (*Repro_Success*) et enfin la durée de la période de reproduction (*BSL*).

Consommation concentré

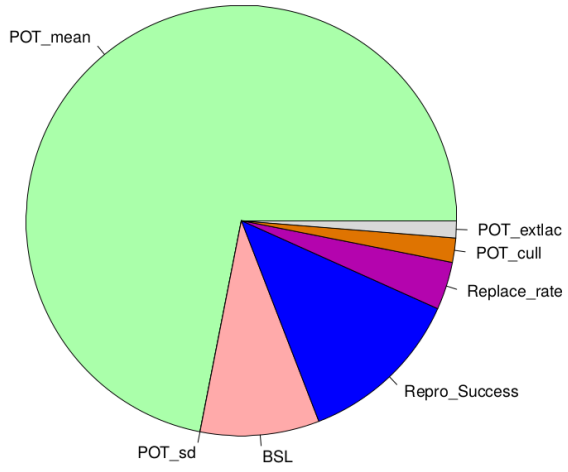


FIGURE 5.11 – Représentation des indices de sensibilité principaux de la consommation de concentré sous forme de diagramme en camembert

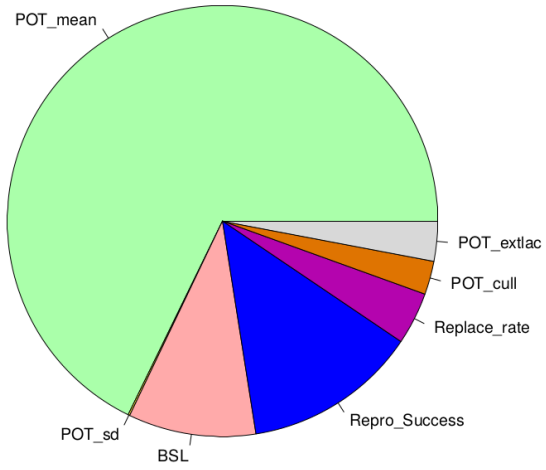


FIGURE 5.12 – Représentation des indices de sensibilité totaux de la consommation de concentré sous forme de diagramme en camembert

Et enfin, c'est *POT_mean* qui a un effet de levier pour ce qui est de la consommation de concentré, cette variable explique à lui seul presque 70% de la variabilité de consommation de concentré, cet effet est soutenu par*Repro_Success* (expliquant 13 % de variabilté) et enfin *BSL* (avec 8 %).

Comme on peut le constater, le paramètre *POT_sd* n'intervient que très peu et uniquement sur la production laitière, en intéraction avec un autre paramètre. Ce paramètre, *POT_sd* peut donc être fixé.

Faire partie d'une équipe et travailler au sein de celle-ci

5.1 L'organisation, les outils

Au cours de ce stage j'ai pu découvrir GitLab, une plateforme permettant d'héberger et de gérer des projets. C'est sur cette dernière que mon travail, mais aussi de la documentation et mes compte-rendus ont été partagés tout au long de ce stage.
Chaque semaine une réunion était organisée de façon à partager et faire le point sur les derniers résultats trouvés et les pistes en cours. Cette réunion s'appuyait sur des compte-rendus rédigés en amont.
Ce cadre m'a permis d'affiner ma pratique d'organisation, de consultation et de restitution.

Une autre expérience de restitution s'est présentée lors de la journée des stagiaires de l'unité MIAT. Comme son nom l'indique, cette journée est dédiée aux stagiaires, et a pour but de permettre à ces derniers de présenter leur cadre de travail et leurs avancées. A la suite de chaque présentation des discussions sont lancées et cet échange permet de diversifier les points de vue et éventuellement d'apporter de nouvelles pistes de recherches pour l'étudiant, mais aussi de faire découvrir ou redécouvrir des domaines d'intérêt actuels aux chercheurs.

5.2 La même étude, vue dans différents domaines

Comme présenté au début de ce rapport, l'encadrement de ce stage a été réalisé par trois chercheurs, tous issus de domaines différents. J'ai ainsi pu bénéficier d'un encadrement optimal : pour une nouvelle approche ou un nouveau résultat, trois points de vue pouvaient être discutés. J'ai aussi pu ainsi approcher des jargons quelques peu différents et m'intéresser à différents domaines.

Bibliographie

[1] Kriging-based sensitivity analysis. `https://www.rdocumentation.org/packages/sensitivity/versions/1.14.0/topics/sobolGP`. Accessed : 2018-07-05.

[2] Monte carlo estimation of sobol' indices (improved formulas of jansen (1999) and saltelli et al. (2010)). `https://www.rdocumentation.org/packages/sensitivity/versions/1.14.0/topics/soboljansen`. Accessed : 2018-07-05.

[3] B. Iooss. Revue sur l'analyse de sensibilité globale de modèles numériques. Journal de la Societe Française de Statistique, Societe Française de Statistique et Societe Mathematique de France, 2011.

[4] L. Puillet. An individual-based model simulating goat response variability and long-term herd performance. The Animal Consortium, 2010.

[5] R.Faivre, B.Iooss, S.Mahévas, D.Makowski, and H.Monod. Analyse de sensibilité et exploration de modèles. Editions Quae, 2013.

Chapitre 6

Annexes

Annexe .1 : Indices de sensibilité - sobolJansen et sobolGP

Annexe .1.1 : Effectif

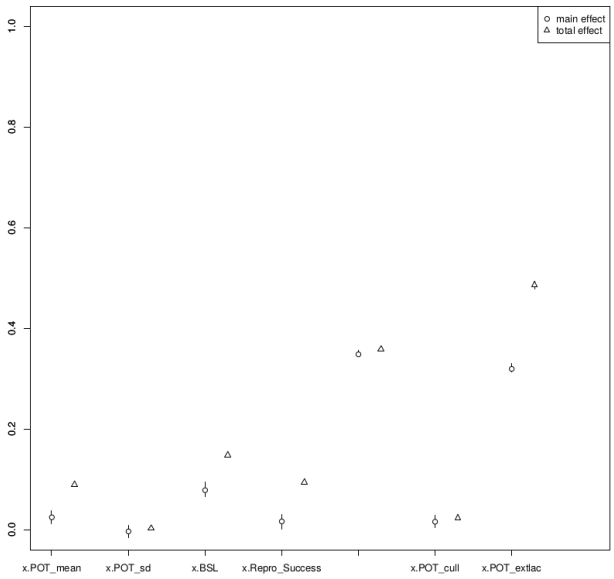


FIGURE 1 – Indices de sensibilité pour la sortie Effectif de chèvres en lactation

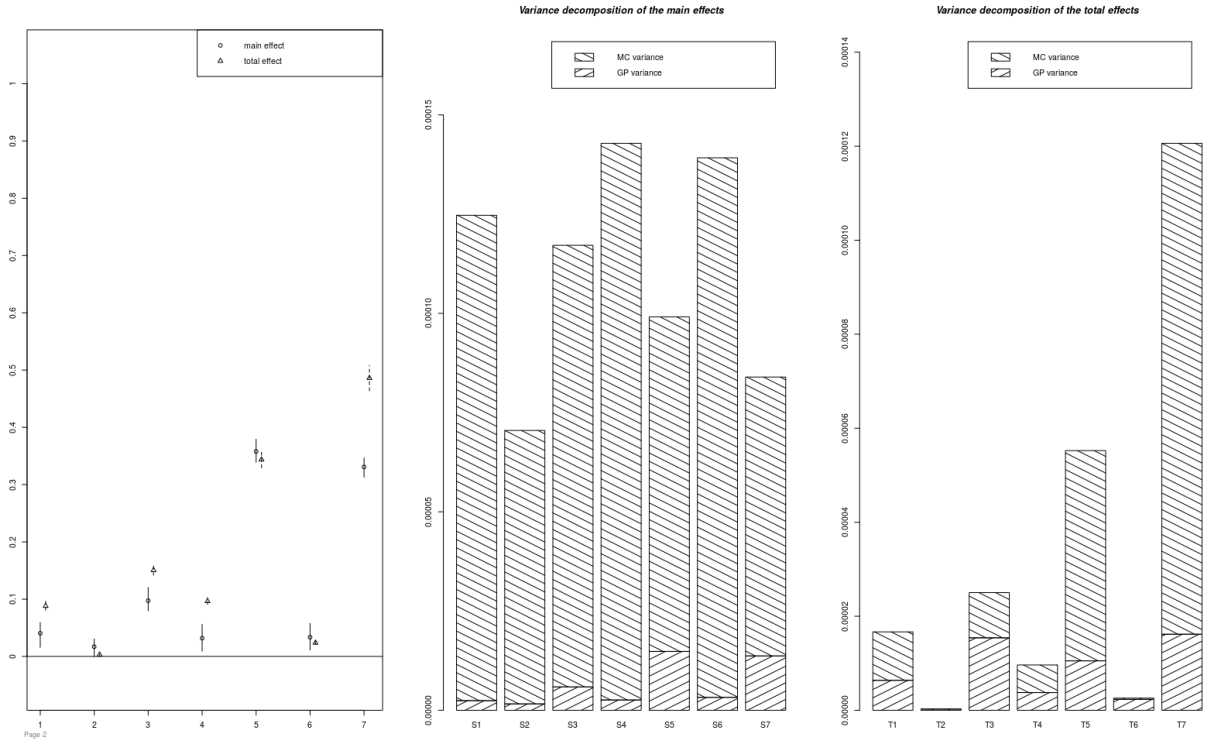


FIGURE 2 – Indices de sensibilité (avec prise en compte des erreurs des processus gaussiens) pour la sortie Effectif de chèvres en lactation

Annexe .1.2 : Production

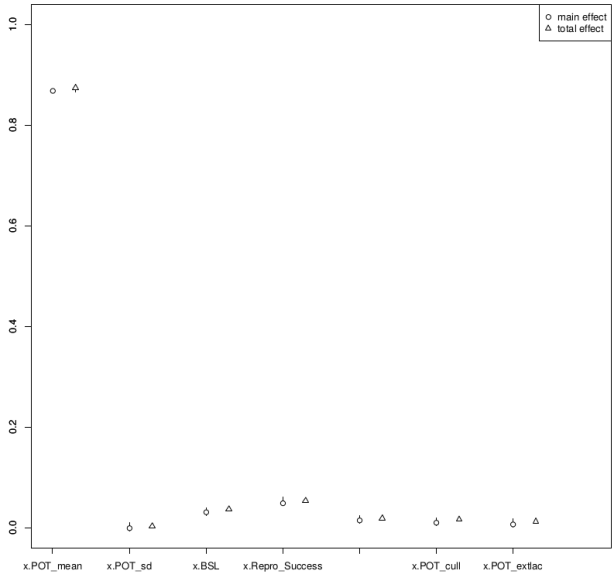


FIGURE 3 – Indices de sensibilité pour la sortie Production laitière

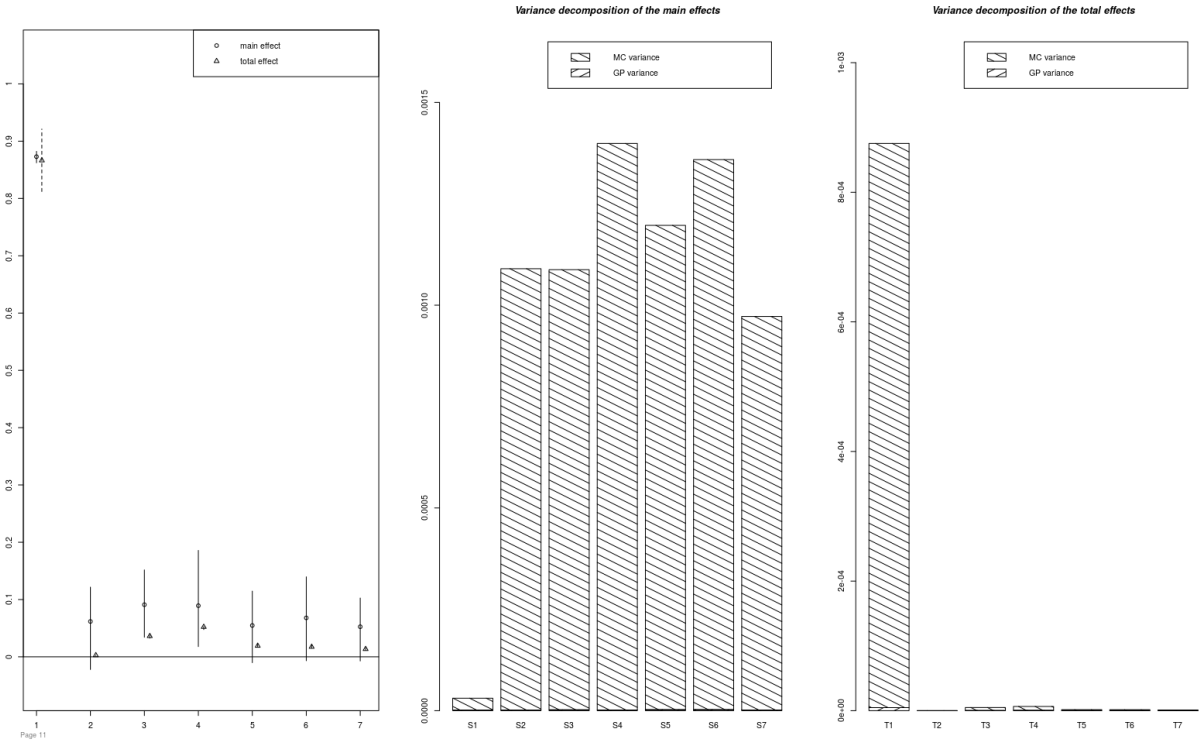


FIGURE 4 – Indices de sensibilité (avec prise en compte des erreurs des processus gaussiens) pour la sortie Production laitière

Annexe .1.3 : Conso1

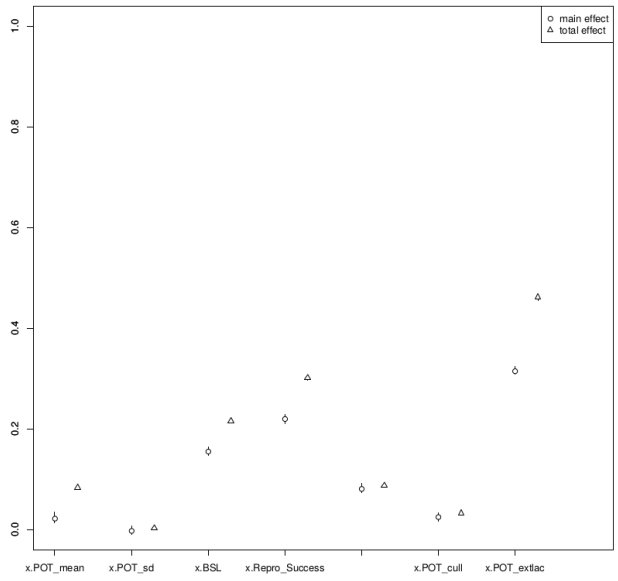


FIGURE 5 – Indices de sensibilité pour la sortie Consommation de fourrage 1

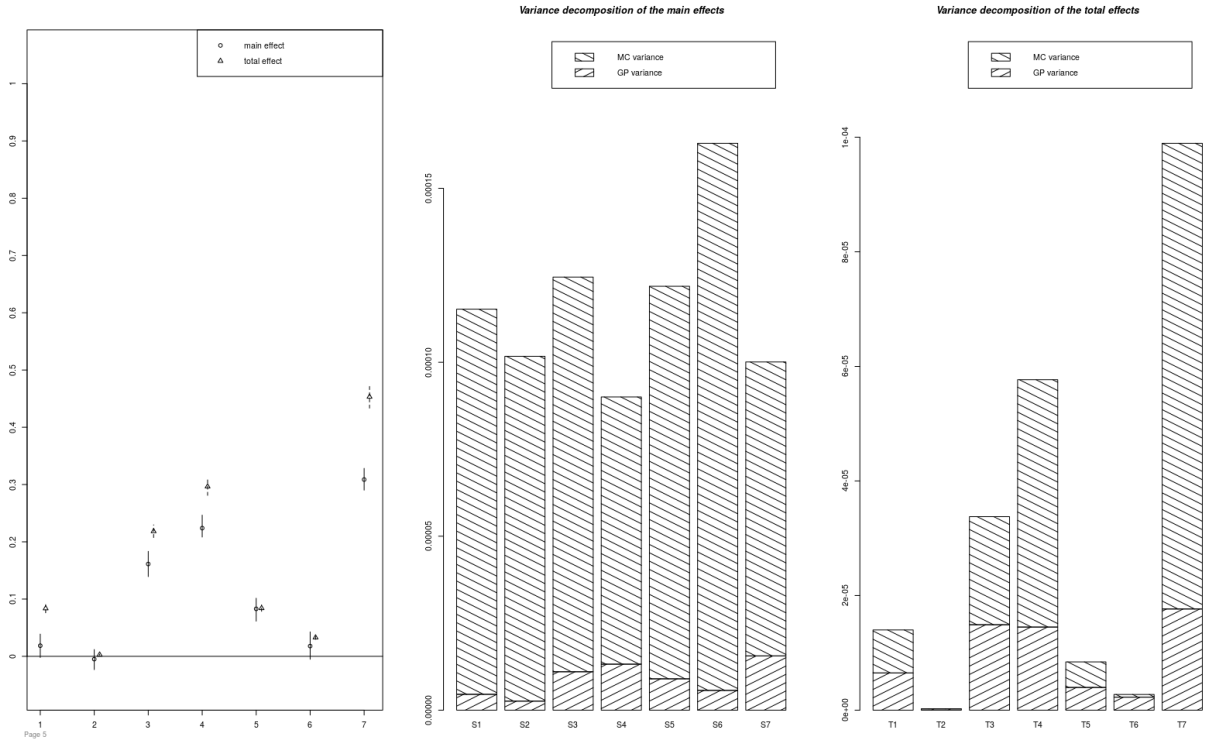


FIGURE 6 – Indices de sensibilité (avec prise en compte des erreurs des processus gaussiens) pour la sortie Consommation de fourrage 1

Annexe .1.4 : Conso2

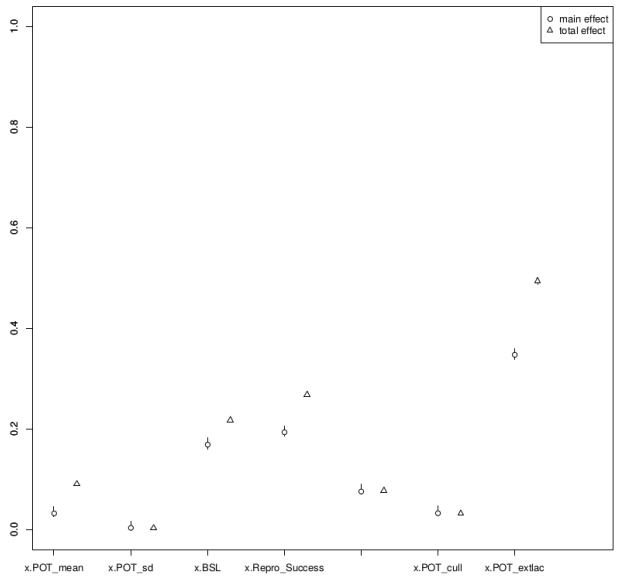


FIGURE 7 – Indices de sensibilité pour la sortie Consommation de fourrage 2

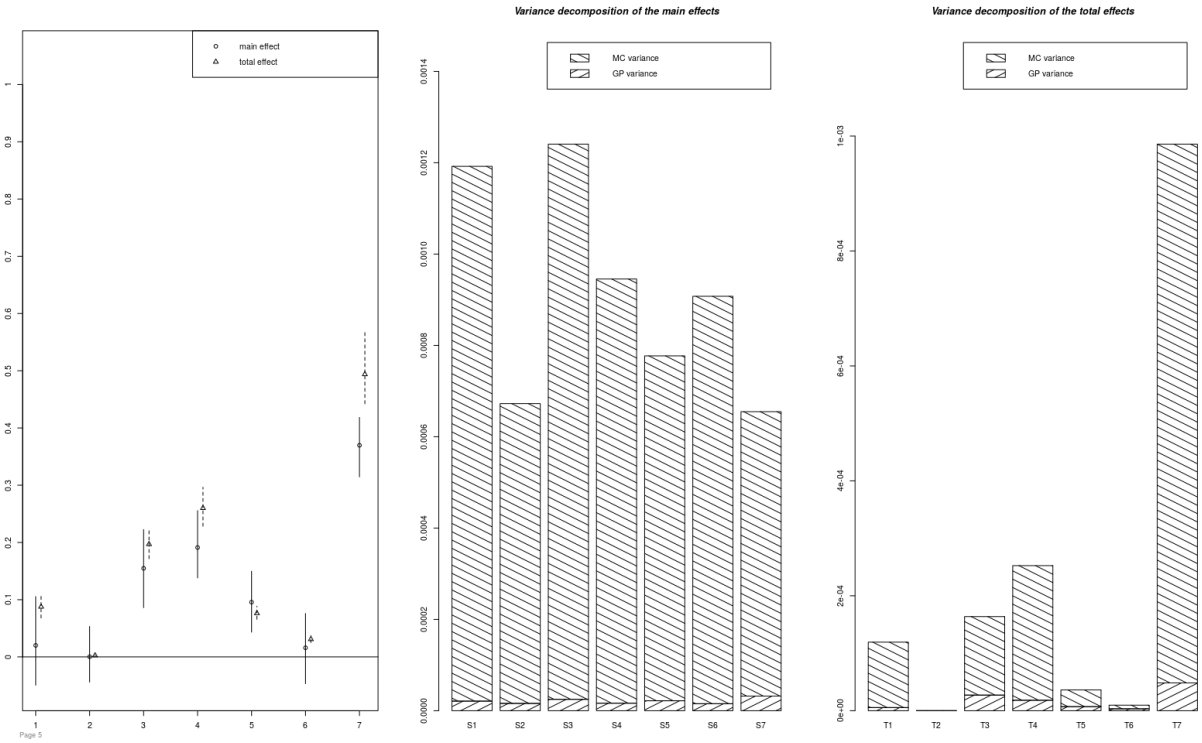


FIGURE 8 – Indices de sensibilité (avec prise en compte des erreurs des processus gaussiens) pour la sortie Consommation de fourrage 2

Annexe .1.5 : Conso3

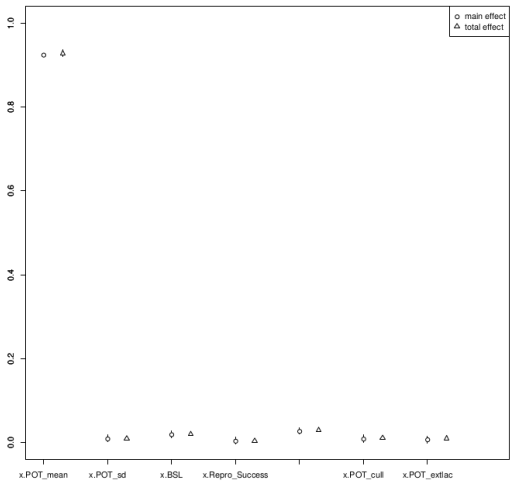


FIGURE 9 – Indices de sensibilité pour la sortie Consommation de fourrage 3

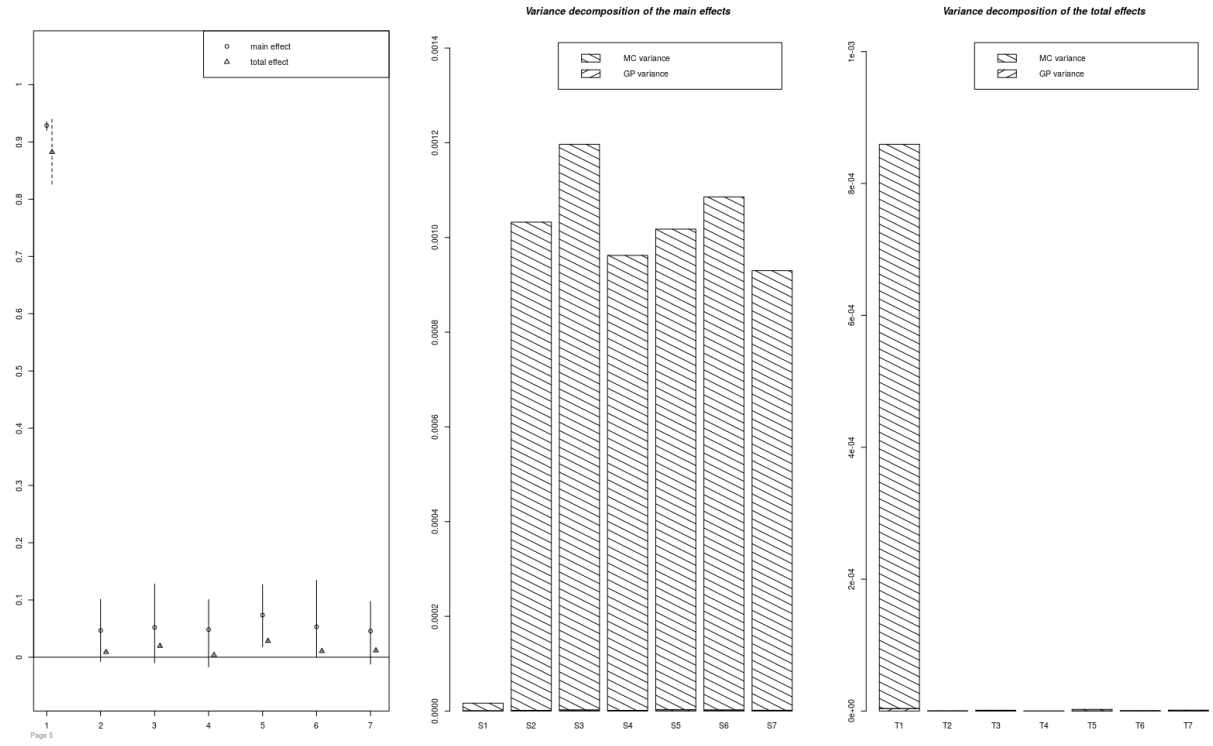


FIGURE 10 – Indices de sensibilité (avec prise en compte des erreurs des processus gaussiens) pour la sortie Consommation de fourrage 3

Annexe .1.6 : ConsoC

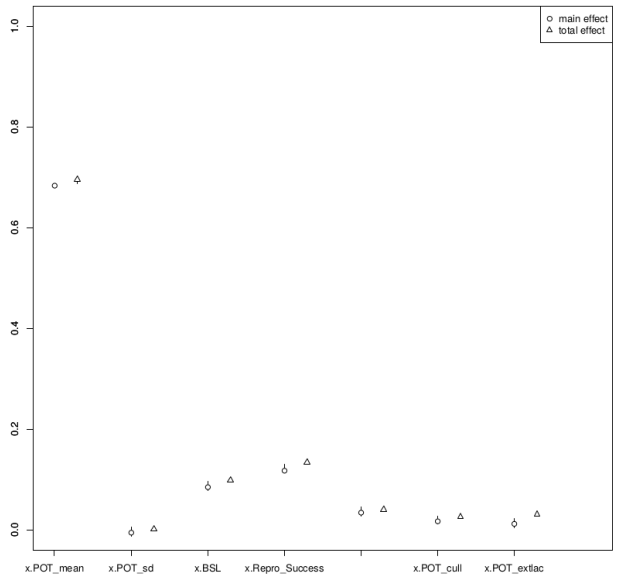


FIGURE 11 – Indices de sensibilité pour la sortie Consommation de concentré

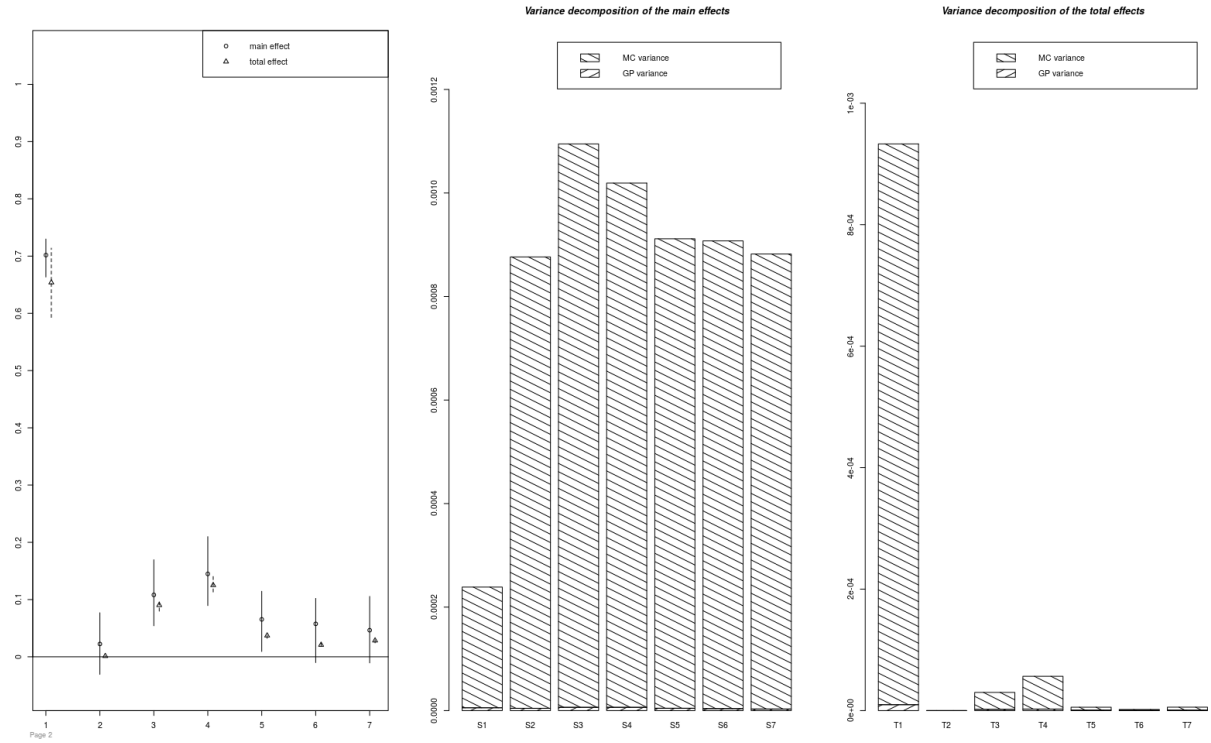


FIGURE 12 – Indices de sensibilité (avec prise en compte des erreurs des processus gaussiens) pour la sortie Consommation de concentré