Master 2 Ingénierie Statistique

Modèle de Markov caché pour la surveillance de la diarrhée virale bovine

Auteur :
Trotreau Matthieu

Professeure référente :
Mme. Philippe Anne

Encadrants :
M. Madouasse Aurélien
M. Galharret Jean-Michel

Avril - Septembre 2024

Remerciements

Je souhaiterais remercier en premier lieu mes deux encadrants de stage Aurélien MADOUASSE et Jean-Michel Galharret. Cela a été un réel plaisir de travailler sous votre supervision. Merci pour cet encadrement qui m'a permit d'évoluer dans une environnement sain et stimulant durant toute la durée de ce stage.

Merci aussi à l'équipe pédagogique du master Ingénierie Statistique de Nantes Université, en particulier madame Lise BELLANGER, madame Anne PHILIPPE et monsieur Aymeric STAMM. Ces deux années n'auraient pas pu avoir lieu sans le temps que vous avez consacré à l'enseignement et la gestion du master.

Merci également à madame Valentine BOUR, doctorante à Nantes Université, et l'équipe du laboratoire de psychologie des Pays de la Loire qui m'a agréablement accueilli lors de mon premier stage dans le monde de la recherche.

Enfin, merci à toute l'équipe PEPS de l'UMR Bioepar pour votre bonne humeur et votre accueil chaleureux au sein du laboratoire. Avec un petit clin d'œil à l'équipe des stagiaires et thèses vétérinaires sans qui ces 6 mois ne seraient pas passés à une telle vitesse.

Je tiens maintenant à remercier mes parents, Catherine et Jean-Marc, merci pour votre dévouement et votre soutien depuis toutes ces années. Pour avoir accepté mes changements de parcours sans sourciller et surtout pour votre présence à mes côtés à chaque instant.

Je ne peux pas évoquer mes parents sans un mot pour mon frère Samuel avec qui la relation a grandement évoluée ces quelques dernières années. Malgré une erreur de parcours avec l'arrêt des mathématiques en fin de lycée je ne peux que te souhaiter du courage dans les études que tu entreprends aujourd'hui.

Enfin, merci à mes amis, à ceux du lycée encore présents aujourd'hui, à ceux plus récents qui sont du monde des maths appliquées ou de la mécanique, ainsi que ceux avec qui des liens forts se sont crées pour parfois se dénouer.

Table des sigles et des abréviations

1. **BLCM** : Modèle Bayésien en Classe Latente
2. **BVD** : Diarrhée Virale Bovine
3. **ELISA** : Enzyme-Linked Immunosorbent Assay
4. **HMC** : Hamiltonian Monte Carlo
5. **HMM** : Hidden Markov Model
6. **IPI** : Infecté Persistant Immunotolérant
7. **IT** : Infecté Transitoire
8. **MCMC** : Monte Carlo Markov Chain
9. **ODR** : Optical Density Ratio
10. **ROC** : Receiver Operating Characteristic
11. **Se** : Sensibilité
12. **Sp** : Spécificité
13. **STOC free** : Surveillance Tool for Outcome-based Comparison of FREEdom from infection
14. **UMR** : Unité Mixte de Recherche

Table des figures

Liste des tableaux

Table des matières

Introduction

Je présente dans ce rapport les différents travaux de recherche réalisés durant mon stage de fin d'étude au laboratoire BIOEPAR, unité mixte de recherche Oniris VetAgroBio Nantes - INRAE. Ce stage était financé par la coopérative de service en élevages Innoval qui gère notamment le programme de maîtrise de l'infection par le virus de la diarrhée virale bovine (BVD) en Bretagne.

La BVD est une maladie virale très contagieuse des bovins ayant fait son apparition au milieu du XX$^{\text{ième}}$ siècle. Elle est rapidement devenue l'une des maladies les plus problématiques au sein des élevages, tant sur le plan économique (Houe, 1999) que sur le plan médical.

Ce stage avait pour objectif la poursuite d'un travail de recherche réalisé au sein du laboratoire BIOEPAR dans le cadre du projet européen STOC free (https://stocfree.eu/). Le projet STOC free (Surveillance Tool for Outcome-based Comparison of FREEdom from infection) est né en 2017 de la volonté de rendre comparable les résultats des différents programmes de surveillance et contrôle des maladies infectieuses chez les bovins. C'est donc naturellement que la BVD a été incluse dans le projet en tant que cas d'étude principal des maladies considérées.

Aujourd'hui les approches dans la gestion de l'infection par le virus de la BVD diffèrent grandement selon les pays. Les plans de lutte sont divisés en deux catégories, les plans de luttes *systématiques* et les plans de lutte *non-systématiques*. Les premiers portent l'objectif d'une diminution généralisée de la prévalence de la maladie sur le territoire. Les seconds reposent sur une gestion au cas par cas et la vaccination des individus. On s'intéresse ici aux plans systématiques qui requièrent la mise en place d'une phase de surveillance. Cette phase permet la détection précoce des nouvelles infections et l'évaluation des mesures mises en places. On trouve alors deux grandes approches qui sont les plans d'éradication et les plans de contrôle (Metcalfe, 2019). Ces approches, principalement empiriques, reposent sur des étapes décrites par Lindberg et Houe en 2005 (Lindberg et Houe, 2005). Les 3 étapes importantes sont la mise en place de mesures de biosécurité, l'élimination du virus, et la surveillance des troupeaux afin de détecter et prévenir les nouvelles infections. Les différents modèles construits à

l'aide de ce schéma font intervenir des tests sérologiques si l'objectif est la détection d'anticorps, ou des tests virologiques si c'est le virus qui est recherché.

L'existence de nombreux tests différents, le choix pouvant varier selon l'approche et l'échelle considérée, entraîne des difficultés dans la comparaison des statuts d'animaux de différents pays. De plus, il est commun de voir les mesures continues de ces tests être discrétisées pour faciliter la prise de décision. La variation des seuils appliqués, avec parfois une absence de seuil réglementaire, peut rendre non identifiable le processus de classification des animaux.

En recherche épidémiologique et vétérinaire on parle de test gold standard pour qualifier le meilleur test existant à un moment donné et faisant office de référence. En l'absence d'un tel test, les modèles bayésiens en classe latente (BLCM) se sont imposés comme des méthodes fiables et efficaces pour l'estimation des seuils optimaux (Olsen et al., 2022), des prévalences de maladies, ainsi que des sensibilités et spécificités des tests utilisés (Hui et Walter, 1980 ; C. I. McAloon et al., 2024 ; C. G. McAloon et al., 2016). Cependant, la grande majorité des études sont encore réalisées à l'aide de résultats de tests dichotomisés malgré l'absence fréquente de gold standard. Ces dernières années des articles illustrant l'utilisation des données de mesures continues voient le jour avec des résultats en faveur de leur conservation (Wang et al., 2024 ; Yang et al., 2022). De nouvelles possibilités complétant les BLCM, comme l'utilisation des modèles de Markov cachés (HMM), sont régulièrement proposées. Les HMM sont des modèles en classes latentes permettant la prise en compte des évolutions d'un système dans le temps. Ceux-ci ont fait preuve de leur efficacité et pertinence mais restent encore peu utilisés en recherche épidémiologique malgré une modélisation adaptée à la surveillance des maladies (Le Strat et Carrat, 1999 ; Watkins et al., 2009).

L'équipe de l'UMR Bioepar participant au projet STOC free a fait le choix de la mise en place d'un HMM pour la surveillance épidémiologique de la BVD (Madouasse et al., 2022). Le modèle initial ayant été construit à l'aide d'observations discrètes (séropositif/séronégatif) l'objectif premier du stage était de répondre à la question suivante :

La conservation des données de mesures continues améliore-t-elle les performances du modèle ?

Parmi les raisons justifiant l'abandon de la dichotomisation nous pouvons citer la perte importante d'information lors de la discrétisation des données, notamment lorsque le résultat du test se trouve proche du seuil utilisé. De plus, l'utilisation de données continues est supposée plus appropriée du fait du caractère dynamique du HMM, le passage entre un état infecté et non infecté étant alors moins abrupte que dans le cas discret. Enfin, la conservation des données continues permettrait la détermination des distributions des mesures de test selon l'état infectieux de l'élevage.

Ce mémoire est structuré en 3 grandes parties. Dans un premier temps nous détaillons le contexte du sujet de stage en étayant la description de la BVD et les objectifs poursuivis. Dans une seconde partie nous présentons les données dont nous disposons ainsi que les détails des modèles et algorithmes mis en œuvre. Enfin, une dernière partie mettra en exergue les performances des modèles sur données simulées et les résultats obtenus sur données réelles.

Chapitre 1

Contexte et objectifs

L'impact de la BVD sur la santé des bovins et l'économie des exploitations étant considérable un arrêté ministérielle, du 31 juillet 2019, impose la détection et l'assainissement des troupeaux infectés dans toute la France. De nombreux pays ont eux aussi mis en place des programmes de détection de la maladie avec des objectifs allant jusqu'à l'éradication du virus.

1.1 L'unité mixte de recherche BIOEPAR

Le stage s'est déroulé au sein de l'unité mixte de recherche (UMR) BIOEPAR (https://bioepar.angers-nantes.hub.inrae.fr/). Celle-ci est placée sous la tutelle de l'INRAE et l'école Oniris VetAgroBio. L'école forme des ingénieurs dans les domaines de l'agroalimentaire et des biotechnologies ainsi que des vétérinaires. L'unité BIOEPAR se situe sur le site de la Chantrerie s'occupant de la formation vétérinaire.

Les missions de l'UMR, dirigée par Nathalie BAREILLE, s'organisent en 3 axes :

- Réduction de la consommation de médicaments anti-infectieux par les animaux
- Prévention sanitaire dans les élevages, territoires et filières
- Adaptation aux évolutions des systèmes d'élevage pour la gestion de la santé animale

Les objectifs du stage s'insèrent en particulier dans la seconde mission de prévention sanitaire au sein des troupeaux.

1.2 Épidémiologie de l'infection par le virus de la BVD

1.2.1 Modes de transmission du virus

La BVD est une maladie virale des bovins observée pour la première fois en 1946 par le canadien Childs. Elle sera décrite la même année aux États-Unis par Peter Olafson (Olafson et al., 1946). Dans un premier temps le virus se multiplie dans les voies respiratoires, puis il s'introduit dans les voies sanguines et se propage à tous les organes. Cette propagation entraîne une grande variété des voies d'excrétion.

La transmission de la maladie est dite horizontale lorsqu'elle intervient lors d'un contact direct entre un individu infecté et un individu sensible. Elle est dite verticale lorsqu'elle est transmise par la mère à sa progéniture. Dans le cas de la BVD ces différents modes de transmissions entraînent la présence de deux catégories d'individus infectés, les infectés persistants immunotolérants (IPI) et les infectés transitoires (IT). Les premiers sont des veaux ayant survécu à l'infection de leur mère lors de la gestation, en particulier lors d'une période où le virus est reconnu comme part intégrante de l'individu ne déclenchant pas de réponse immunitaire. Les seconds sont les individus infectés par transmission horizontale.

La transmission verticale joue un rôle majeur dans la dynamique d'infection de cette maladie. En effet, les IPI sont les principaux vecteurs de la maladie au sein des troupeaux. Ils excrètent en permanence le virus, que ce soit dans l'urine, les matière fécales ou encore le lait. La majorité des infections a lieu lors du contact d'un IPI avec un animal sensible.

1.2.2 Conséquences des infections transitoires et persistantes

La plupart des infections transitoires donnent des cas de faible gravité, mais des symptômes plus importants voire létaux peuvent apparaître. Parmi eux on trouve des diarrhées sanguinolentes, de la fièvre ou encore des ulcères sur les muqueuses. Il peut aussi apparaître des troubles digestifs généralement bénins chez l'adulte mais pouvant être mortels chez le nouveau-né. Cette maladie provoquant une immunodépression on la retrouve fréquemment chez des animaux atteints de troubles respiratoires causés par d'autres pathologies. Des troubles reproductifs peuvent apparaître chez le mâle comme la femelle avec, respectivement, une baisse de la qualité du sperme dans les premiers jours de l'infection et des perturbations du fonctionnement ovarien.

Les conséquences d'une infection par le virus de la BVD sont plus problématiques lorsqu'une vache gestante est contaminée. En effet, il est très fréquent qu'une vache malade avorte précocement si l'infection survient lors des 3 premiers mois de gestation. Les avortements peuvent cependant continuer à se produire jusqu'au terme de la

période des 9 mois. Ces avortements ont un impact considérable sur l'économie d'un élevage ainsi que le bien-être des animaux. De plus, si la vache tombe malade entre le premier et le cinquième mois de gestation, en particulier lors du premier mois, cela peut entraîner la naissance d'un animal IPI, caractérisation que nous avons évoquée dans la partie 1.2.1. Le reste des différents effets d'une infection chez une vache enceinte peut être résumé dans la figure 1.1 suivante (Grooms, 2004) :

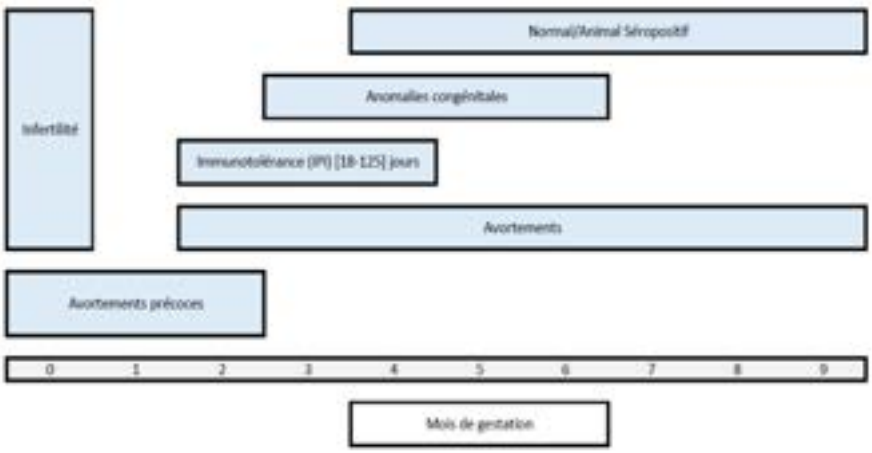


Figure 1.1 – Conséquences d'une infection par le virus de la BVD chez une vache gestante

1.2.3 Protocole de détection de la BVD en Bretagne.

Les approches pour la détection de l'infection peuvent varier selon les pays ou les régions (van Roon et al., 2020). En Bretagne il a été choisi de réaliser ce que l'on appelle des tests ELISA (enzyme-linked immunosorbent assay). Rechercher le virus au sein des troupeaux en testant le statut virologique de chaque animal est coûteux et ne présente que peu d'intérêt. En effet, peu de vaches excrètent le virus à un moment donné ce qui augmente le risque de détecter des faux négatifs. Il a été décidé d'appliquer ces tests ELISA, qui sont des tests sérologiques, sur des mélanges de lait de tank. Cette méthode permet de disposer d'un mélange recensant l'ensemble des laits du troupeau dans lequel on recherche la présence d'anticorps indiquant si au moins l'une des vaches a déjà rencontré le virus. Le principal défaut de cette méthode étant que la présence d'anticorps peut signifier qu'une vache est actuellement infectée ou déjà immunisée du fait d'une infection antérieure. Pour des raisons de moyens disponibles les données dont nous disposons proviennent de deux tests différents seulement réalisés tous les 3 à 6 mois (à l'origine tous les 6 mois puis tous les 3 mois après apparition du second test). De plus, nos résultats (x) sont standardisés à partir des mesures de témoins positifs (x^+) et négatifs (x^-) à partir de la formule :

$$\frac{x - x^-}{x^+ - x^-} \times 100$$

Cette standardisation permet généralement l'obtention de mesures entre 0 et 100, mais celles-ci peuvent être négatives ou supérieures à 100 selon les mesures des témoins présents sur la plaque de test. Ces résultats sont appelés odr, pour Optical Density Ratio, et sont utilisés pour l'apprentissage de nos modèles.

1.3 Objectifs

Le modèle de Markov caché initial a été élaboré afin d'estimer la dynamique de la BVD à partir de résultats binarisés. L'objectif premier du stage était de modifier le modèle pour qu'il intègre les mesures de tests continues, les estimations obtenues par ces deux modèles seront comparées sur données simulées puis données réelles.

Une des éventualités après application du modèle était de retrouver les sensibilités et spécificités des tests à partir des distributions estimées, avec notamment la possibilité de proposer un seuil de discrétisation optimal. Cela peut paraître contradictoire avec l'objectif initial d'utilisation de données continues. Cependant, la présence d'un seuil de référence optimal pourrait uniformiser et faciliter l'utilisation des tests sur le terrain.

Enfin, l'objectif final du stage est de mettre à disposition un outil opérationnel pour la surveillance de l'infection notamment sous la forme d'un package R.

Chapitre 2

Matériels et méthodes

Nous débutons cette partie avec la présentation des données réelles dont nous disposons. Cela nous permettra notamment de justifier certaines modélisations présentées dans la seconde moitié de cette partie, avec au préalable une description détaillée des concepts généraux des modèles de Markov cachés.

2.1 Description et manipulation des données réelles

Il convient donc de commencer cette partie avec la présentation et l'analyse descriptive du matériel sans lequel la création de méthodes d'apprentissage perdrait un grand intérêt. Nous avons donc un jeu de données original de dimensions 798593×5, la suppression d'un certains nombre de données manquantes nous donne un jeu exploitable de 793296 observations. Nous pouvons expliciter nos 5 variables :

- *ede* [Chaîne de caractères] : L'identifiant unique du troupeau. Les 2 premiers chiffres indiquent le département et les 3 suivants la commune associée. Ces 5 premiers chiffres forment ainsi le numéro INSEE de la commune.
- *test_date* [Date] : La date de la mesure du test sur les laits de tank. La plage de dates s'étend de 2000 à 2021.
- *test_id* [Chaîne de caractères] : L'identifiant du test utilisé. C'est un facteur à deux modalités, LGMCAT et LGMVEA.
- *odr* [Réel] : La variable répertorie les résultats continus des mesures de test.
- *t_res* [Entier] : Résultats binaires des tests, obtenus par seuillage de la variable odr. Facteur à deux modalités 0 et 1.

- *herd_dep* [Entier] : Département du troupeau, obtenu à partir de l'identifiant *ede*. Facteur à 4 modalités 22 (Côtes-d'Armor), 29 (Finistère), 35 (Ille-et-Vilaine) et 56 (Morbihan).

Il est important d'indiquer qu'il n'y a pas une unique mesure *odr* par troupeaux mais plusieurs mesures réalisées à intervalles réguliers, on parle alors de données longitudinales. Le graphique en figure 2.1 permet d'illustrer le fait que les tests sont réalisés tous les 6 mois puis tous les 3 mois après apparition du test LGMVEA.

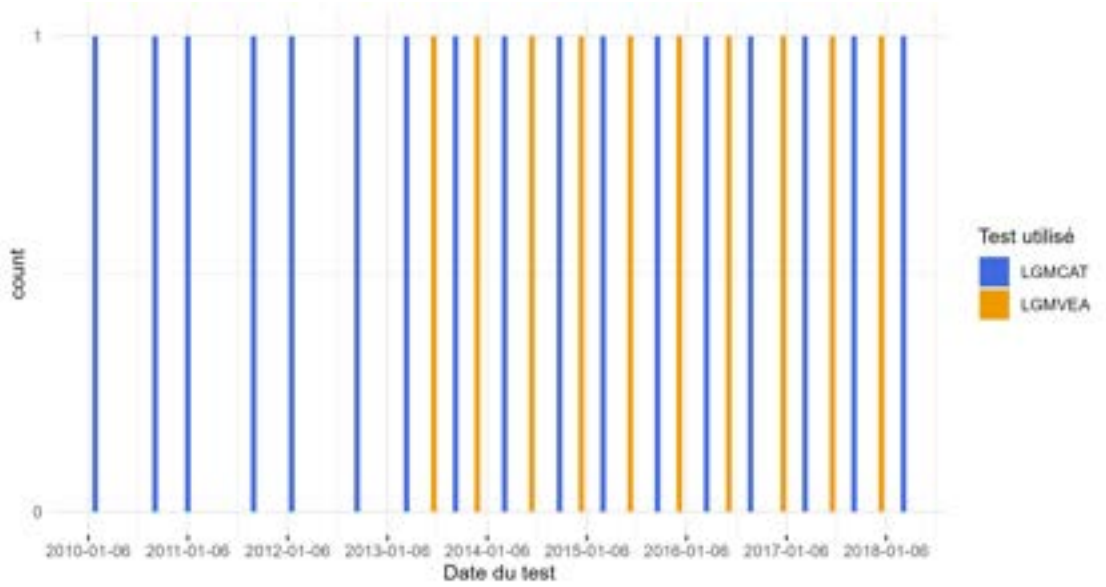


FIGURE 2.1 – Histogramme des tests réalisés sur un troupeau issu du département 22, entre 2010 et 2021.

C'est notamment cette caractéristique longitudinale qui justifie l'utilisation d'un HMM, et plus généralement d'un processus de Markov, comme nous pouvons le voir en partie 2.2.1.

Pour continuer la description des données nous réalisons un graphique des données continues, pour chaque test, mettant en avant les résultats binaires.

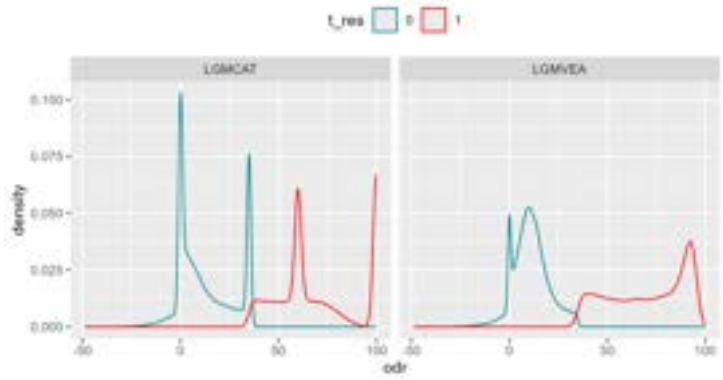


FIGURE 2.2 – Distribution des résultats des deux tests (LGMCAT à gauche et LGMVEA à droite) en fonction du statut des troupeaux (non infecté en bleu et infecté en rouge).

La figure 2.2 fournit la distribution des résultats de tests en fonction du statut retenu après discrétisation. Elle est obtenue après suppression de quelques données manquantes et aberrantes. On remarque une saturation des données en 0 et 100 ainsi qu'autour de 40 et 60. Un mélange gaussien à deux composantes paraît envisageable pour les deux tests, avec l'idée d'une troisième composante pour la censure en 0.

Après une analyse plus détaillée des données on trouve que les 4 valeurs de saturation sont exactement 0, 35, 60 et 100. On relève notamment que le seuil de discrétisation des données continues est justement fixé à une valeur d'odr de 35. Une possible explication serait la recatégorisation de certains résultats autour de ces 4 seuils, possiblement à certaines dates ou par certains départements. Il apparaît en réalité que les données provenant du département 35 sont grandement responsables des différents excès de valeurs. Si on décide de supprimer ce département de notre jeu on obtient finalement les distributions du graphique 2.3.

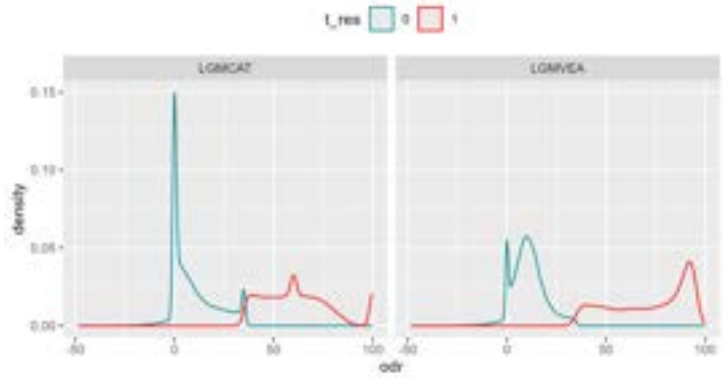


FIGURE 2.3 – Distribution des mesures de test après traitement des données.

Nous allons maintenant extraire des données plus propres de notre jeu afin d'appliquer les modèles explicités en parties 2.2.4 et 2.2.5 et dont nous présentons les résultats sur données réelles en 3.2.

L'échantillon retenu permettra la mise en œuvre de plusieurs cas d'application assez simples pour l'illustration des modèles sur données réelles. Le jeu de données complet a été séquencé en des périodes de 3 ans, les graphiques correspondants sont disponibles en annexe A.1. Il a été décidé de conserver la dernière période s'étendant de 2018 à 2021, dont nous avons ensuite extrait uniquement les 300 premiers troupeaux du département 29 avec le plus de mesures disponibles et ce pour le test LGMVEA. L'idée était de de choisir des données avec une distribution relativement propre en évitant une saturation de 0. Le jeu final est donc composé de 6 variables pour 2100 observations et est illustré par le graphique en figure 2.4.

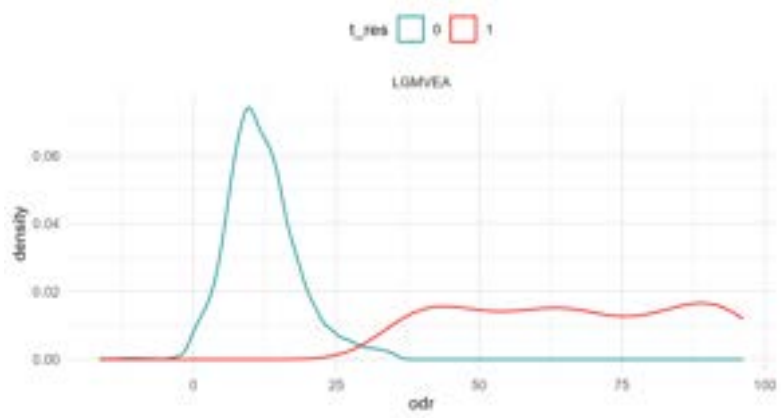


FIGURE 2.4 – Distribution de l'échantillon de données réelles sélectionné.

Enfin, nous présentons en partie 3.2 les résultats d'une dernière expérience. Celle-ci consiste à entraîner le modèle mis au point lors de ce stage sur une plage de temps scindée en 28 périodes, avec une prédiction de la probabilité d'infection à la fin de chacune de ces périodes. La table de ces 28 intervalles de temps est donnée en annexe D.1. Nous disposons d'une période complète de 10 ans s'étendant de début 2011 à fin 2020. On prédit la probabilité d'infection 4 fois par ans à partir de janvier 2014, la prédiction étant réalisée à partir de 3 ans d'historique des données.

Nous avons détaillé notre jeu de données réelles ainsi que les différentes parties de celui-ci qui serviront à l'entraînement de nos modèles. Ces modèles sont exposés à la suite de cette partie avec les considérations générales et théoriques en préliminaires.

2.2 Modèle de Markov caché

L'un des éléments clés de ce stage était l'utilisation d'un modèle de Markov caché afin de représenter la dynamique d'infection des troupeaux par le virus de la BVD. Il est donc nécessaire d'en présenter les fondements afin de correctement saisir, les raisons qui rendent cette modélisation pertinente dans le cas d'une maladie infectieuse, ainsi que la construction des modèles appliqués à notre problématique.

2.2.1 Concept général

Un modèle de Markov caché, ou hiden Markov model (HMM) est un processus stochastique qui repose sur le principe de Markov qui suppose que l'état futur de la chaîne dépend uniquement de son état présent et est indépendant des états précédents. Dans le cas classique nous disposons d'un espace d'états discret E auquel sont associées des probabilités de transition entre états. Une chaîne de Markov peut être à temps discret comme continu mais nous nous intéressons ici seulement au temps discret avec $T \subset \mathbb{N}$ l'espace de temps. La chaîne de Markov est alors une séquence de n variables aléatoires à valeur dans l'espace d'état $X_1, X_2, ..., X_n$ avec la propriété suivante :

$$\forall t \leq n, \forall (i_1, ..., i_t, j) \in E^{t+1},$$
$$\mathbb{P}(X_{t+1} = j \mid X_t = i_t, X_{t-1} = i_{t-1}, ..., X_1 = i_1) = \mathbb{P}(X_{t+1} = j \mid X_t = i_t).$$

On parle de chaîne homogène lorsque la probabilité de transition est indépendante du temps t :

$$\mathbb{P}(X_{t+1} = j \mid X_t = i_t) = \mathbb{P}(X_t = j \mid X_{t-1} = i_t).$$

On représente sur la figure 2.5 un exemple de dynamique pour une chaîne de Markov homogène à deux états.

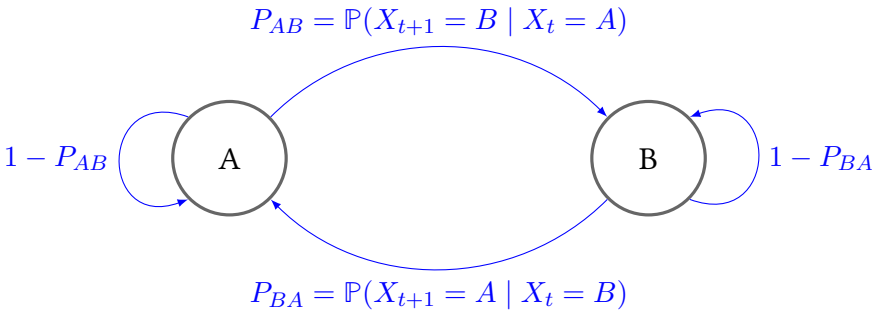


FIGURE 2.5 – Graphe d'une chaîne de Markov à deux états

Dans le cas d'une chaîne de Markov classique nous connaissons l'état au temps t et pouvons déterminer sa probabilité d'atteindre chacun des autres états au temps $t+1$ à l'aide des probabilités de transitions. Le modèle de Markov caché intervient lorsque la position de la chaîne à l'instant t n'est pas connue explicitement. Le seul élément dont nous disposons est une observation, ou émission, qui peut être observée dans n'importe quel état avec une certaine probabilité. Nous disposons ainsi de l'espace d'états E mais aussi d'un espace d'observations O. Nous pouvons reprendre la chaîne de la figure 2.5 en la représentant dans le cas d'un HMM simple.

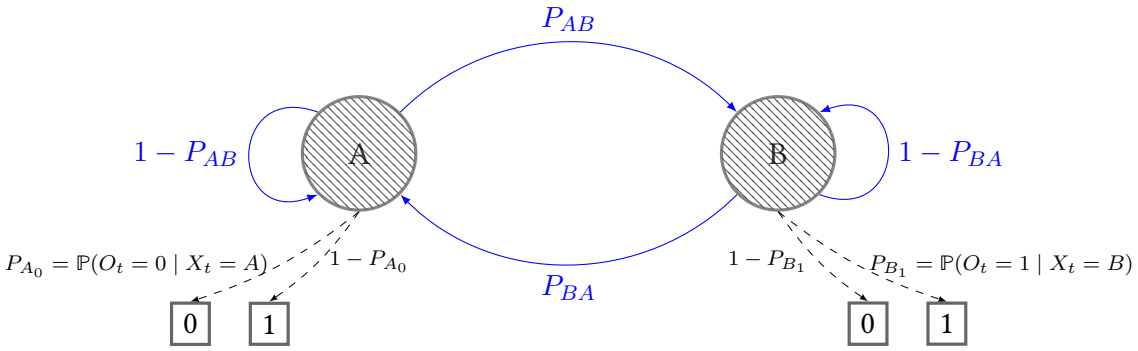


FIGURE 2.6 – Graphe d'un modèle de Markov caché à deux états et deux émissions.

L'exemple de la chaîne de Markov classique en figure 2.5 peut produire la chaîne $(X_1 = A, X_2 = B, X_3 = B)$ de manière unique. Tandis que dans l'exemple de HMM en figure 2.6 cette même suite d'états peut produire les suites d'observations $(O_1 = 1, O_2 = 0, O_3 = 1)$ et $(O_1 = 0, O_2 = 1, O_3 = 1)$ avec des probabilités différentes. De la même manière, ces séquences d'observations auraient pu provenir de séquences d'états différentes. En réalité une séquence d'états cachés de longueur T peut générer $|O|^T$ séquences d'observations différentes. Si on ajoute à ça le nombre d'états possibles on obtient un maximum de $|E|^T \times |O|^T$ associations possibles entre séquences d'états et séquences d'observations. On comprend ainsi aisément que les statuts réels de nos variables aléatoires $X_{i=1,\dots,T}$ deviennent rapidement non identifiables. De ce fait, il existe différents algorithmes permettant l'estimation des paramètres d'un HMM. Nous allons maintenant présenter l'algorithme forward qui est l'un des plus utilisé.

2.2.2 Algorithme Forward - Backward

Lorsque nous travaillons avec un modèle de Markov caché nous pouvons souhaiter répondre à différentes questions. En réalité 3 problèmes majeurs s'offrent à nous :

- *Evaluation problem* : Quelle est la probabilité qu'une séquence d'observations $(O_i)_{i=1,\dots,T}$ ait été produite par notre modèle.

- *Uncovering problem* : Quelle est la séquence d'états $(X_i)_{i=1,...,T}$ la plus probable de produire la séquence d'observations $(O_i)_{i=1,...,T}$.
- *Learning problem* : Comment ajuster les paramètres du modèle pour obtenir une probabilité maximale d'obtenir la séquence $(O_i)_{i=1,...,T}$.

Nous ne cherchons pas tout à fait à répondre à l'une de ces 3 questions. Cependant la méthode de résolution du premier problème va nous permettre d'atteindre nos objectifs. Une manière naïve d'envisager ce problème serait de lister toutes les séquences d'états de longueur T possibles. Cependant, cette solution n'est pas envisageable du fait de coûts informatiques bien trop importants la rendant impossible à mettre en œuvre. Le *problème d'évaluation* peut être résolu à l'aide d'une procédure que l'on nomme *Procédure Forward - Backward* (Rabiner, 1989) et que nous présentons dans cette sous-partie.

On considère un espace d'états X discret avec $X = \{s_1, ..., s_n\}$ et on écrira $x_t(i)$ pour $x_t = s_i$. Dans un premier temps on pose la *variable forward* $\alpha_{t+1}(i)$ correspondant à la probabilité jointe $p(x_{t+1}(i), o_{1:t+1})$ d'être dans l'état s_i au temps $1 \leq t+1 \leq T$ et d'avoir observé une chaîne $o_1, ..., o_{t+1}$. Nous avons donc,

$$\begin{aligned}\alpha_{t+1}(i) &= p\left(x_{t+1}(i), o_{1:t+1}\right) \\ &= \sum_{j=1}^{n} p\left(x_{t+1}(i), x_t(j), o_{1:t+1}\right).\end{aligned}$$

On rappelle la chain rule en probabilité qui nous dit :

Theorème (Chain rule)**.** *Soit* $(\Omega, \mathcal{A}, \mathbb{P})$ *un espace de probabilité,* $A_1, ..., A_k \in \mathcal{A}$ *alors :*

$$\mathbb{P}(A_1, ..., A_k) = \mathbb{P}(A_1) \prod_{j=2}^{k} \mathbb{P}(A_j \mid A_{j-1}, ..., A_1).$$

On peut donc écrire :

$$p(x_{t+1}(i), x_t(j), o_{1:t+1}) = p(x_t(j), o_{1:t}) p(x_{t+1}(i) \mid x_t(j), o_{1:t}) p(o_{t+1} \mid x_{t+1}(i), x_t(j), o_{1:t}).$$

or, o_{t+1} est conditionnellement indépendant de tout sauf x_{t+1} et x_{t+1} dépend uniquement de x_t donc :

$$\begin{aligned}\alpha_{t+1}(i) &= p\left(o_{t+1} \mid x_{t+1}(i)\right) \sum_{j=1}^{n} p\left(x_{t+1}(i) \mid x_t(j)\right) p(x_t(j), o_{1:t}) \\ &= p(o_{t+1} \mid x_{t+1}(i)) \sum_{j=1}^{n} p\left(x_{t+1}(i) \mid x_t(j)\right) \alpha_t(j).\end{aligned}$$

Dans le cas d'un HMM dont les observations sont discrètes $p\left(o_{t+1} \mid x_{t+1}(i)\right)$ correspond à la probabilité $\mathbb{P}\left(O_{t+1} = o_{t+1} \mid x_{t+1}(i)\right)$, tandis que dans le cas d'observations continues c'est l'évaluation de la densité de probabilité au point o_{t+1} connaissant l'état x_{t+1} (Damiano et al., 2017). La grandeur $p\left(x_{t+1}(i) \mid x_t(j)\right)$ est donnée par les paramètres de la dynamique du modèle. On peut donc entièrement calculer $\alpha_{t+1}(i)$ à partir des $\alpha_t(j)$.

L'initialisation de l'algorithme $\alpha_1(i) = p\left(o_1 | x_1(i)\right) p\left(x_1(i)\right)$ est donnée par les distributions a priori sur les paramètres du modèle. Ensuite, à l'aide de la distribution jointe $\alpha_{t+1}(i)$ on obtient facilement $\alpha_{t+1} = p(o_{1:t+1})$:

$$\alpha_{t+1} = p(o_{1:t+1}) = \sum_{i=1}^{n} p\left(x_{t+1}(i), o_{1:t+1}\right) = \sum_{i=1}^{n} \alpha_{t+1}(i).$$

À terme on obtient $p\left(x_{t+1}(i) \mid o_{1:t+1}\right) = \frac{\alpha_{t+1}(i)}{\alpha_{t+1}}$ et l'estimation de l'état au temps $t+1$ avec $\hat{x_{t+1}} = \underset{i}{\operatorname{argmax}}\left(p\left(x_{t+1}(i) \mid o_{1:t+1}\right)\right)$.

Nous venons de détailler l'algorithme forward qui nous permet d'obtenir les grandeurs $p\left(x_{t+1}(i) \mid o_{1:t+1}\right)$, $\hat{x_{t+1}}$ et α_{t+1}, cette dernière répondant au *problème d'évaluation* lorsque $t+1 = T$. La seconde possibilité est d'implémenter la procédure backward qui repose sur la *variable backward* $\beta_t(i)$ qui correspond aux probabilités conditionnelles des chaînes d'observations partielles $\{o_{t+1}, ..., o_T\}$ connaissant l'état x_t :

$$\begin{aligned}\beta_t(i) &= p\left(o_{t+1:T} \mid x_t(i)\right) \\ &= \sum_{j=1}^{n} p\left(x_{t+1}(j) \mid x_t(i)\right) p\left(o_{t+1} \mid x_{t+1}(j)\right) \beta_{t+1}(j).\end{aligned}$$

L'initialisation est généralement arbitraire avec $\beta_T(i) = 1$ pour tout i. On peut alors déterminer la quantité d'intérêt $p(o_{1:T})$ en fonction de la *variable backward* :

$$
\begin{aligned}
p(o_{1:T}) &= \sum_{i=1}^{n} p\left(o_{1:T}, x_1(i)\right) \\
&= \sum_{i=1}^{n} p\left(o_{1:T} \mid x_1(i)\right) p\left(x_1(i)\right) \\
&= \sum_{i=1}^{n} p\left(o_1 \mid x_1(i)\right) p\left(o_{2:T} \mid x_1(i)\right) p\left(x_1(i)\right) \\
&= \sum_{i=1}^{n} p(o_1 \mid x_1(i)) \beta_1(i) p\left(x_1(i)\right).
\end{aligned}
$$

L'association des deux procédures nous permet de déterminer la probabilité de chacun des états au temps t connaissant la séquence complète des observations :

$$
p\left(x_t(i) \mid o_{1:T}\right) = \frac{p\left(o_{1:T}, x_t(i)\right)}{p(o_{1:T})}.
$$

À l'aide de la chain rule et de l'indépendance conditionnelle des $o_{t+1}, ..., o_T$ avec les $o_1, ..., o_t$ on peut écrire :

$$
\begin{aligned}
p\left(x_t(i) \mid o_{1:T}\right) &= \frac{p\left(o_{1:t}, x_t(i)\right) p\left(o_{t+1:T} \mid x_t(i)\right)}{p(o_{1:T})} \\
&= \frac{\alpha_t(i)\beta_t(i)}{p(o_{1:T})} \\
&= \frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{n} \alpha_t(i)\beta_t(i)}.
\end{aligned}
$$

Nous venons de détailler les deux procédures permettant de répondre au premier problème. Il est important de relever que la procédure forward répond à elle-seule au *problème d'évaluation*. De plus, nous nous intéressons plus particulièrement aux probabilités $(p\left(x_T(i) \mid o_{1:T}\right))_{i=1,..n}$ pour lesquelles l'algorithme forward est nécessaire et suffisant. De ce fait, seul celui-ci a été implémenté dans les modèles bayésiens que nous allons maintenant présenter.

2.2.3 Une dynamique commune aux deux modèles

Avant toute chose nous présentons la dynamique du modèle de Markov caché, celle-ci ne dépendant pas de l'utilisation d'observations discrètes ou continues. La

représentation de l'infection des troupeaux par le virus de la BVD provient de l'idée de considérer l'état réel des troupeaux, séronégatif (0) ou séropositif (1), comme étant inconnu et donc latent. Cela implique un espace d'états latents discret de cardinal 2 qui ne varie pas selon le type d'observation considéré. On peut représenter la dynamique du modèle à l'aide de la figure suivante (Madouasse et al., 2022).

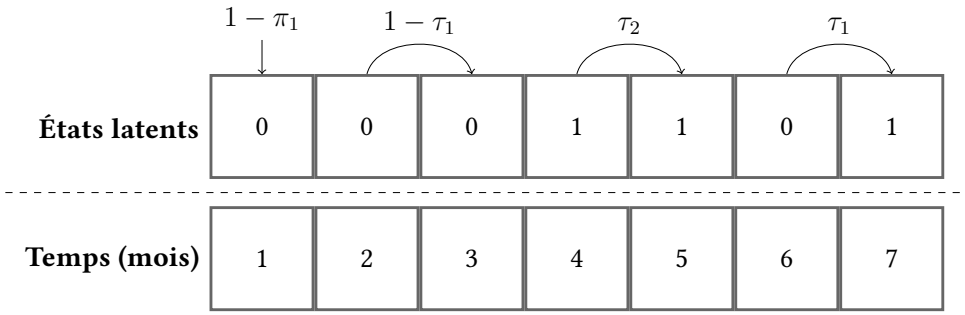


Figure 2.7 – Représentation de la dynamique du modèle de Markov caché

Notre espace d'états latents étant discret on peut représenter les probabilités de transition dans la matrice P, avec P_{ij} la probabilité de passer de l'état i à l'état j :

$$\mathrm{P} = \begin{bmatrix} 1-\tau_1 & \tau_1 \\ 1-\tau_2 & \tau_2 \end{bmatrix}.$$

À noter que nous sommes bien dans une chaîne de Markov homogène, les probabilités de transition étant indépendantes du temps.

Le vecteur des probabilités initiales est donné par $\Pi = (1-\pi_1, \pi_1)$. Dans notre cas π_1 correspond à la probabilité qu'un troupeau soit considéré infecté et représente donc la prévalence de troupeaux infectés par la BVD. Pour une prévalence à l'équilibre on dispose de l'équation suivante :

$$\pi_1(1-\tau_2) = (1-\pi_1)\tau_1. \tag{2.1}$$

L'équation (2.1) peut être réécrite de la manière suivante :

$$\begin{aligned} & \pi_1(1-\tau_2) = \tau_1 - \pi_1\tau_1 \\ \iff \quad & \pi_1(1+\tau_1-\tau_2) = \tau_1 \\ \iff \quad & \pi_1 = \frac{\tau_1}{(1+\tau_1-\tau_2)}. \end{aligned} \tag{2.2}$$

On peut montrer que l'on retrouve bien ce résultat en déterminant le vecteur de mesure invariante pour notre chaîne de Markov. Il existe le théorème suivant :

Theorème. *Soit* P *une matrice stochastique irréductible. Alors,* P *admet une unique probabilité invariante* π *et on a* $\pi > 0$.

Pour déterminer cette probabilité invariante il nous suffit de déterminer un vecteur propre associé à la valeur propre 1 de la matrice ${}^{t}\mathrm{P}$ et de le normaliser par la somme de ses composantes.

Soit $\lambda \in \mathbb{R}$ et I_2 la matrice identité d'ordre 2,

$$\begin{aligned} | \ {}^{t}\mathrm{P} - \lambda \mathrm{I}_2 \ | &= (1-\tau_1-\lambda)(\tau_2-\lambda) - (1-\tau_2)\tau_1 \\ &= \lambda^2 + \lambda(\tau_1 - \tau_2 - 1) + \tau_2 - \tau_1 \\ &= (\lambda - 1)\left(\lambda - (\tau_2 - \tau_1)\right). \end{aligned}$$

On a bien 1 comme valeur propre de la matrice ${}^{t}\mathrm{P}$. On trouve l'espace propre associé de la manière suivante :

$$\begin{aligned} \left(\ {}^{t}\mathrm{P} - I_2 \ \right) \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} &\iff \begin{cases} -\tau_1 x + (1-\tau_2)\, y = 0 \\ \tau_1 x + (\tau_2 - 1)\, y = 0 \end{cases} \\ &\iff \{\tau_1 x + (\tau_2 - 1)\, y = 0 \\ &\iff \begin{cases} y = \dfrac{\tau_1}{1-\tau_2} x \ . \end{cases} \end{aligned}$$

On obtient donc le vecteur propre $\left(1, \frac{\tau_1}{1-\tau_2}\right)$ associé à la valeur propre 1. On le normalise pour obtenir le vecteur de probabilité invariante :

$$\Pi = \left(\frac{1-\tau_2}{1-\tau_2+\tau_1}, \frac{\tau_1}{1-\tau_2+\tau_1} \right).$$

On retrouve bien l'expression précédente de la prévalence à l'équilibre obtenue en (2.2). Nous passons maintenant à la présentation des spécificités du premier modèle avec observations discrètes.

2.2.4 Modèle initial - Observations discrètes

Nous avons présenté en partie 1.2.3 la détection du virus à l'aide d'analyse de lait de tank. Les mesures de test obtenues sont ensuite discrétisées à l'aide d'un seuil h

pour établir un statut séronégatif (0) ou séropositif (1). Ce dernier est alors considéré "observé" et non "réel" car ces tests ne sont pas des *gold standards* et ne permettent donc pas d'assurer l'exactitude du statut déterminé. On note Y la variable aléatoire pour les mesures de test, S et O les variables aléatoires respectives des états réel et observé.

$$O = \begin{cases} 0 \; si \; Y < h, \\ 1 \; sinon. \end{cases} \tag{2.3}$$

Cette discrétisation s'accompagne des notions de sensibilité (Se) et spécificité (Sp). La première correspond à la probabilité de détecter des vrais positifs, la seconde correspond à la probabilité de détecter des vrais négatifs.

$$\begin{cases} Se = \mathbb{P}(O = 1 \mid S = 1) \\ Sp = \mathbb{P}(O = 0 \mid S = 0). \end{cases}$$

Ces quantités feront office de probabilités d'émission dans le cadre de ce premier HMM à observations discrètes.

Nous pouvons illustrer le lien entre état latent et état observé avec la figure 2.8.

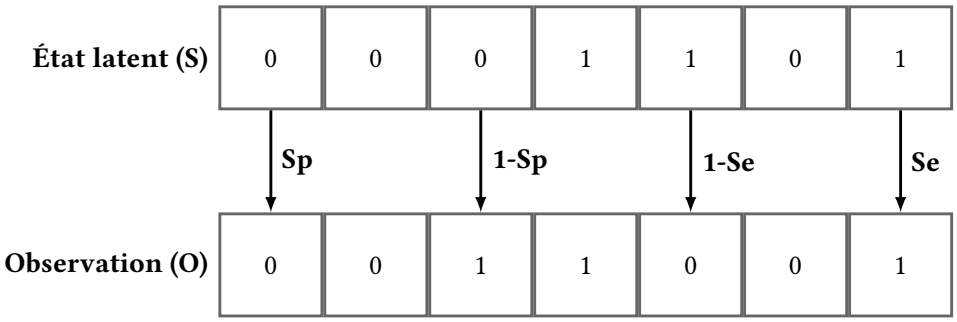


FIGURE 2.8 – Représentation de la relation entre états latents et résultats de tests

Le mélange posé pour ce premier HMM est alors caractérisé par les probabilités suivantes :

$$\begin{cases} \mathbb{P}(O = 0) = (1 - \pi_1) \times Sp + \pi_1 \times (1 - Se) \\ \mathbb{P}(O = 1) = (1 - \pi_1) \times (1 - Sp) + \pi_1 \times Se. \end{cases} \tag{2.4}$$

Nous pouvons résumer le HMM avec observations discrètes à l'aide du graphe de la figure 2.9.

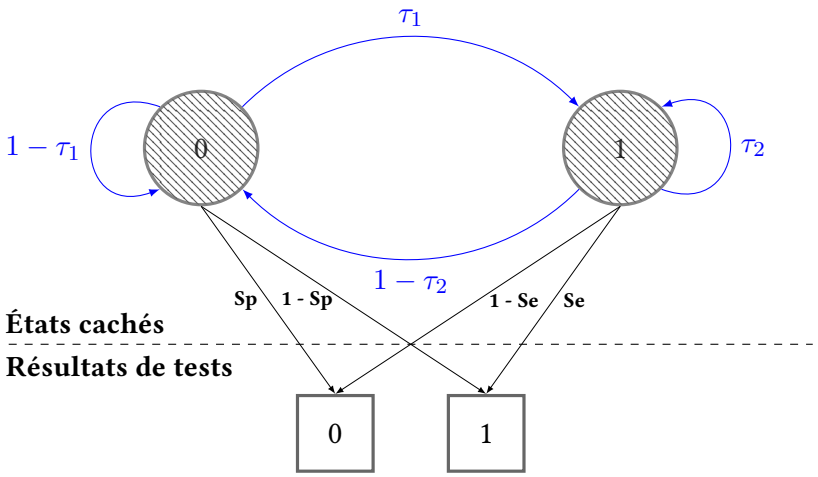


FIGURE 2.9 – Graphe de notre HMM avec espace d'observations discret

Le modèle étant implémenté dans un cadre bayésien nous présentons maintenant les différentes possibilités de lois a priori pour les paramètres à estimer.

Les paramètres Se et Sp étant des probabilités une bonne manière de poser leurs lois a priori est d'utiliser des lois Beta.

$$\begin{cases} Se \sim \text{Beta}(\alpha_{Se}, \beta_{Se}) \\ Sp \sim \text{Beta}(\alpha_{Sp}, \beta_{Sp}). \end{cases}$$

Ensuite une première possibilité implémentée dans le package STOC free est de poursuivre avec des lois Beta sur le reste des paramètres.

$$\begin{cases} \pi_1 \sim \text{Beta}(\alpha_{\pi_1}, \beta_{\pi_1}) \\ \tau_1 \sim \text{Beta}(\alpha_{\tau_1}, \beta_{\tau_1}) \\ \tau_2 \sim \text{Beta}(\alpha_{\tau_2}, \beta_{\tau_2}). \end{cases}$$

Une seconde possibilité est d'utiliser des gaussiennes échelonnées entre 0 et 1 à l'aide d'une transformation logit.

$$\begin{cases} \theta_{\pi_1} \sim \mathcal{N}(\mu_{\pi_1}, \sigma^2_{\pi_1}) \\ \theta_{\tau_1} \sim \mathcal{N}(\mu_{\tau_1}, \sigma^2_{\tau_1}) \\ \theta_{\tau_2} \sim \mathcal{N}(\mu_{\tau_2}, \sigma^2_{\tau_2}), \end{cases} \qquad \begin{cases} \pi_1 = \text{logit}(\theta_{\pi_1}) \\ \tau_1 = \text{logit}(\theta_{\tau_1}) \\ \tau_2 = \text{logit}(\theta_{\tau_2}). \end{cases}$$

Pour terminer cette présentation du premier modèle nous explicitons en table 2.1 les expressions des $(\alpha_{t=1,\dots,T}(i))_{i=1,2}$ qui interviendront dans l'algorithme forward.

		o_{t+1}	
		0	1
s_{t+1}	0	$Sp\,((1-\tau_1)\alpha_t(1)+(1-\tau_2)\alpha_t(2))$	$(1-Sp)\,((1-\tau_1)\alpha_t(1)+(1-\tau_2)\alpha_t(2))$
	1	$(1-Se)\,(\tau_1\alpha_t(1)+\tau_2\alpha_t(2))$	$Se\,(\tau_1\alpha_t(1)+\tau_2\alpha_t(2))$

TABLE 2.1 – Tableau des $(\alpha_{t+1}(i))_{i=1,2}$ lorsque $2 \leq t+1 \leq T$, pour l'algorithme forward du HMM avec observations discrètes

On dresse le tableau 2.2, similaire au précédent, pour l'initialisation des $(\alpha_1(i))_{i=1,2}$.

		o_1	
		0	1
s_1	0	$Sp \times (1-\pi_1)$	$(1-Sp) \times (1-\pi_1)$
	1	$(1-Se) \times \pi_1$	$Se \times \pi_1$

TABLE 2.2 – Tableau des $(\alpha_1(i))_{i=1,2}$ pour l'algorithme forward du HMM avec observations discrètes

Les différentes variables aléatoires sont initialisées manuellement en fixant une constante ou en échantillonnant une réalisation de leur loi a priori. Maintenant que le modèle original a été correctement explicité nous pouvons passer à la présentation du nouveau modèle qui intègre des densités d'émissions continues.

2.2.5 Nouveau modèle - Observations continues

L'objectif initial de ce stage était d'implémenter la possibilité de conserver les mesures continues dans le modèle de Markov caché. Pour ce faire il a été envisagé de modéliser les distributions des mesures de tests à l'aide d'un mélange gaussien. Ce type de mélange associé à une dynamique de type markovienne est aussi appelé Markov Switching Models (HAMILTON, 1989). Le mélange présenté en (2.4) est alors entièrement défini par la fonction de densité suivante :

$$f(y) = (1 - \pi_1)f_0(y) + \pi_1 f_1(y) \tag{2.5}$$

Nous avons f_0 la densité gaussienne de paramètres (μ_0, σ_0) représentant les troupeaux sains et f_1 la densité gaussienne de paramètres (μ_1, σ_1) représentant les troupeaux infectés. Dans le cas d'un HMM avec observations continues les probabilités d'émissions sont remplacées par l'évaluation ponctuelle des densités en l'observation, le graphe de la figure 2.9 devient alors :

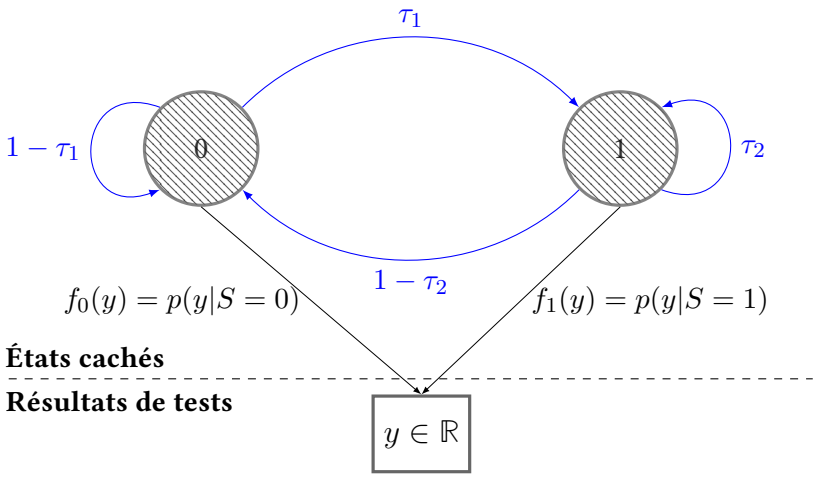


FIGURE 2.10 – Graphe de notre HMM avec espace d'observations continu

Sur la figure 2.10 les émissions discrètes ont été remplacé par une infinité d'émissions continues. Les probabilités d'émissions sont quant à elles remplacées par l'évaluation en l'observation des densités des deux composantes du mélange.

Pour ce nouveau modèle la possibilité de poser des lois a priori à l'aide de la transformation logit n'a pour le moment pas été implémentée. De ce fait, seules des lois Beta sont actuellement possibles pour les paramètres de la dynamique, on retrouve donc :

$$\begin{cases} \pi_1 \sim \text{Beta}(\alpha_{\pi_1}, \beta_{\pi_1}) \\ \tau_1 \sim \text{Beta}(\alpha_{\tau_1}, \beta_{\tau_1}) \\ \tau_2 \sim \text{Beta}(\alpha_{\tau_2}, \beta_{\tau_2}). \end{cases}$$

Cependant, les grandeurs Se et Sp ayant disparues au profit des paramètres des gaussiennes il nous faut maintenant poser 2 lois a priori supplémentaires. Nous avons choisi

d'utiliser des lois normales pour μ_0 et μ_1. Concernant σ_0 et σ_1 nous souhaitions utiliser des inverses gamma afin de disposer d'un support sur $[0, +\infty]$ en imitant une uniforme sur celui-ci. Cependant, du fait de soucis rencontrés dans l'implémentation, nous avons décidé de ne pas préciser de lois a priori pour ces deux paramètres. Dans cette situation Stan pose de lui même une loi de Laplace sur le support de la variable.

$$\begin{cases} \mu_i \sim \mathcal{N}(\mu_{\mu_i}, \sigma_{\mu_i}) \\ \sigma_i \sim \mathcal{U}(0, +\infty) \end{cases}, \quad i = 0, 1.$$

L'une des possibilités envisagées pour les lois a priori des $(\mu_i)_{i=0,1}$ est d'utiliser les données réelles pour diriger la position des gaussiennes. En effet, nos données étant des mesures de test à l'origine dichotomisées à l'aide du seuil h nous avons jugé bon de proposer, pour un jeu de données avec N mesures de Y, la paramétrisation suivante :

$$\begin{cases} \mu_{\mu_0} = \dfrac{\min_n(y_n) + h}{2} \\ \mu_{\mu_1} = \dfrac{\max_n(y_n) + h}{2} \end{cases}, \quad \begin{cases} \sigma_{\mu_0} = \left(\dfrac{h - \min_n(y_n)}{8}\right)^2 \\ \sigma_{\mu_1} = \left(\dfrac{\max_n(y_n) - h}{8}\right)^2. \end{cases} \tag{2.6}$$

Il est important d'être vigilant lorsque l'on utilise de l'information provenant des données d'apprentissage. Cependant de telles lois a priori sur les moyennes des distributions permettent seulement d'indiquer que les troupeaux non-infectés se situent à gauche du seuil et les infectés à droite. Il est toutefois obligatoire ici de connaître ou d'indiquer une valeur pour h qui ne peut pas être fixée par défaut.

Enfin, h nous permet d'évoquer la manière avec laquelle nous mettons en parallèle les modèles avec émissions continues et discrètes. En effet, ce nouveau modèle nous permet lui aussi de déterminer des valeurs de Se et Sp. On pose F_0 et F_1 les fonctions de répartition associées à f_0 et f_1.

$$\begin{cases} Sp = \mathbb{P}(O = 0 \mid S = 0) & = \mathbb{P}(y < h \mid S = 0) = F_0(h) \\ Se = \mathbb{P}(O = 1 \mid S = 1) & = \mathbb{P}(y > h \mid S = 1) = 1 - F_1(h). \end{cases} \tag{2.7}$$

Lorsque nous aurons à évaluer les performances des deux modèles nous pourrons alors comparer les estimations de tous les paramètres présents dans le modèle original.

Nous avons présenté dans cette sous-partie les généralités du modèle avec observations continues. Cependant, nous avons vu en partie 2.1 que nos données sont considérablement saturées en 0. La suppression de ces données en 0 représentant une perte d'information conséquente nous avons décidé d'inclure une nouvelle composante dans le mélange.

$$f(y) = (1 - \pi_1)\left(w f_w(y) + (1 - w) f_0(y)\right) + \pi_1 f_1(y). \tag{2.8}$$

Cette nouvelle composante, définit par la fonction de densité f_w, nous permet de considérer ces observations comme des mesures provenant de troupeaux sains tout en débruitant la réelle distribution de ces troupeaux qui pourrait se retrouver biaisée du fait d'un trop grand nombre de mesures égales à 0. Cependant, seul le paramètre w vient s'ajouter au modèle bayésien, la fonction f_w étant la densité d'une loi normale $\mathcal{N}\left(0, (0.01)^2\right)$. De la même manière que pour π_1 nous utilisons une loi beta comme loi a priori sur w.

$$\{w \sim \text{Beta}(\alpha_w, \beta_w)\ .$$

Comme pour le modèle discret nous concluons la présentation du modèle avec le résumé des $(\alpha_{t+1}(i))_{i=1,2}$ dans le tableau 2.3.

		y_1	y_{t+1}
s_1, s_{t+1}	0	$(1 - \pi_1) \times f_0(y_1)$	$f_0(y_{t+1})\left((1 - \tau_1)\alpha_t(1) + (1 - \tau_2)\alpha_t(2)\right)$
	1	$\pi_1 \times f_1(y_1)$	$f_1(y_{t+1})\left(\tau_1\alpha_t(1) + \tau_2\alpha_t(2)\right)$

TABLE 2.3 – Tableau complet (initialisation en première colonne) des $(\alpha_{t+1}(i))_{i=1,2}$ pour l'algorithme forward du HMM avec observations continues

2.2.6 Implémentation des modèles

Dans cette partie nous présentons très brièvement l'implémentation des deux modèles à l'aide du langage Stan (CARPENTER et al., 2017). Les deux programmations étant très similaires nous en évoquons uniquement les grandes lignes, elles sont inspirées du papier de DAMIANO et al., 2017.

Stan fait office d'interface en permettant le lien entre R et un algorithme d'échantillonage de Monte Carlo Markov Chain (MCMC). En l'occurence la méthode proposée par Stan est une variation de l'algorithme de Monte Carlo Hamiltonien (HMC).

Les probabilités calculées dans l'algorithme forward décroissants de manière exponentielle, afin d'éviter des soucis d'underflow ce sont les $(\log(\alpha_{t+1}(i)))_{i=1,2}$ qui sont codés. Ceux-ci sont calculés en deux parties à l'aide d'une matrice 1×2 qui stocke les valeurs et qui passe ensuite dans la fonction *log_sum_exp* définie par l'égalité suivante :

$$\text{LogSumExp}(x_1, ..., x_k) = \log\left(\sum_{i=1}^{k} e^{x_i}\right).$$

Cette fonction permet de correctement effectuer la somme des deux éléments afin d'obtenir les $(\alpha_{t+1}(i))_{i=1,2}$ tout en restant à l'échelle logarithmique.

C'est le bon moment pour préciser une spécificité des $(\alpha_{t+1}(i))_{i=1,2}$ de nos deux HMM. En effet, nous ne disposons pas de données à chaque pas de temps mais uniquement de mesures de tests tous les 3 à 6 mois. De ce fait une majorité des $p(s_{t+1}, o_{1:t+1})$ sont en réalité égales à $p(s_{t+1}, o_{1:t})$. Il nous faut donc trouver une manière correcte d'écrire ces probabilités tout en conservant la cohérence de l'algorithme forward. Nous pouvons montrer qu'il suffit simplement de reprendre les expressions des tableaux 2.1 et 2.3 en supprimant les probabilités d'émissions conditionnelles, c'est à dire l'évaluation des densités conditionnelles en l'observation. Dans le cas d'une absence d'observation au temps $t+1$ on peut écrire :

$$\begin{aligned}
\alpha_{t+1}(i) &= p\left(s_{t+1}(i), o_{1:t+1}\right) \\
&= p\big(s_{t+1}(i), o_{1:t}\big) \\
&= \sum_{j=1}^{n} p\big(s_{t+1}(i), s_t(j), o_{1:t}\big) \\
&= \sum_{j=1}^{n} p\big(s_{t+1}(i) \mid s_t(j), o_{1:t}\big) p\left(s_t(j), o_{1:t}\right) \\
&= \sum_{j=1}^{n} p\big(s_{t+1}(i) \mid s_t(j)\big)\alpha_t(j).
\end{aligned}$$

Ces quelques lignes terminent de justifier l'écriture des $(\alpha_{t+1}(i))_{i=1,2}$ dans nos deux modèles de Markov cachés avec la présence d'une différenciation entre présence et absence d'émissions.

Pour finir, l'ensemble des grandeurs de l'algorithme forward étant déterminées à l'échelle logarithmique c'est la transformation softmax qui permet d'obtenir les $\big(p(s_{t+1}(i) \mid o_{1:t+1})\big)_{i=1,2}$ dont l'expression est donnée en partie 2.2.2.

$$p(s_{t+1}(i) \mid o_{1:t+1}) = \frac{e^{\alpha_{t+1}(i)}}{\sum_{j=1}^{2} e^{\alpha_{t+1}(j)}}.$$

Les différents modèles et leurs implémentations ayant été présentés nous passons maintenant à la description des possibilités pour la prédiction des probabilités d'infection.

2.3 Prédiction de la probabilité d'infection

2.3.1 Dans le cadre bayésien

Les deux modèles permettent de donner une estimation de la probabilité d'infection du troupeau au dernier pas de temps. En effet, nous avons vu en partie 2.2.2 que l'algorithme forward nous permet de déterminer la probabilité d'être dans chacune des classes à chacun des pas de temps t et ce connaissant la chaîne des observations jusqu'au temps t. Ainsi, nous conservons la probabilité d'être dans la classe séropositif lors de chacune des itérations de l'algorithme MCMC mais uniquement pour la dernière observation. Nous obtenons ainsi une distribution a posteriori pour $\mathbb{P}(S_T = 1 \mid y_{1:T})$ et pouvons approcher cette probabilité par l'estimateur de Bayes $\mathbb{E}\left[\mathbb{P}(S_T = 1 \mid y_{1:T})\right]$

Nous pouvons illustrer ces distributions a posteriori à l'aide d'un graphique obtenu dans le cas d'une application aux données réelles que nous présentons en partie 3.2. Nous segmentons une plage de temps en 28 périodes et appliquons le modèle bayésien afin qu'il prédise la probabilité d'infection pour la dernière date de chaque période. Pour un premier troupeau sur la $24^{\text{ième}}$ période nous obtenons par exemple :

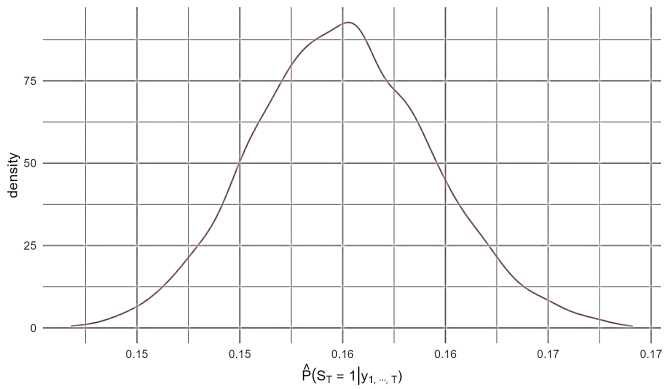


Figure 2.11 – Distribution a posteriori de la probabilité d'infection estimée à la $24^{\text{ième}}$ période, pour un troupeau des Côtes-d'Armor (22).

Nous pouvons compléter ce graphique de densité avec un intervalle de crédibilité de niveau 95% pour la probabilité estimée et ce sur chacune des 28 périodes.

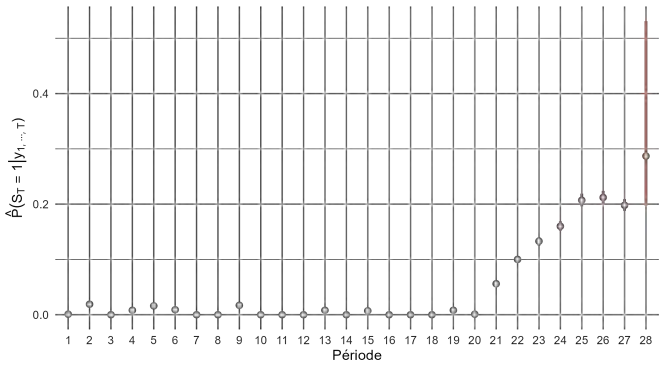


FIGURE 2.12 – Intervalles de crédibilité pour la probabilité d'infection d'un troupeau des Côtes-d'Armor (22) pour les 28 périodes

Nous détaillons et commentons les résultats de cette application dans la partie 3.2 dédiée.

Le coût de calcul du modèle bayésien étant relativement conséquent nous présentons maintenant une manière complémentaire de prédire les probabilités d'infection des troupeaux.

2.3.2 Dans le cadre déterministe

Il apparaît peu pertinent de relancer le modèle sur l'ensemble des données dès l'apparition de nouvelles observations notamment à cause d'un temps de calcul important dans le cas de jeux de données conséquents.

Une première possibilité après estimation des paramètres par le modèle bayésien est de conserver les estimateurs de Bayes des probabilités prédites et les utiliser pour l'initialisation d'une simple chaîne de Markov. En effet, nous pouvons récupérer $\hat{\tau_1}$ et $\hat{\tau_2}$ qui permettent de déterminer la matrice de transition de la chaîne et appliquer cette dynamique aux vecteurs de probabilités de manière à obtenir une prévision rapide ne nécessitant aucune observation sur les quelques mois à venir.

Ensuite, lorsque nous disposons d'une unique nouvelle mesure de test et souhaitons déterminer le statut probable du troupeau nous pouvons déterminer la classe et les probabilités à la manière d'un algorithme EM. Pour ce faire nous utilisons la propriété répondant au nom de *optimal Bayes rule*. La probabilité conditionnelle d'appartenance à une classe est notamment définie par :

$$\mathbb{P}(S = s \mid y) \propto \hat{\rho}\hat{f}_s(y) \quad , \quad \hat{\rho} = \begin{cases} 1 - \hat{\pi_1} \; si \; s = 0, \\ \hat{\pi_1} \; si \; s = 1. \end{cases}$$

L'*optimal Bayes rule* intervient alors pour nous donner la classe la plus probable que l'on nomme alors le maximum a posteriori :

$$S* = \underset{s\in\{0,1\}}{\operatorname{argmax}} \left(\mathbb{P}(S = s \mid y)\right).$$

De cette façon nous obtenons une estimation de la classe la plus probable. Cependant, pour une observation y_0, si nous souhaitons déterminer la probabilité "ponctuelle" une manière correcte de le faire est la suivante (Nguyen, 2016) :

$$\begin{aligned}
\mathbb{P}(S = s \mid y_0 - \epsilon < y < y_0 + \epsilon) &= \frac{\mathbb{P}(S = s, y_0 - \epsilon < y < y_0 + \epsilon)}{\mathbb{P}(y_0 - \epsilon < y < y_0 + \epsilon)} \\
&= \frac{\mathbb{P}(S = s) \int_{y_0-\epsilon}^{y_0+\epsilon} \hat{f}_s(y)}{\int_{y_0-\epsilon}^{y_0+\epsilon} \hat{f}(y)} \\
&= \frac{\hat{\omega} \int_{y_0-\epsilon}^{y_0+\epsilon} \hat{f}_s(y)}{\int_{y_0-\epsilon}^{y_0+\epsilon} \hat{f}(y)} \\
&= \frac{\hat{\rho}\left(\hat{F}_s(y_0 + \epsilon) - \hat{F}_s(y_0 - \epsilon)\right)}{\hat{F}(y_0 + \epsilon) - \hat{F}(y_0 - \epsilon)}.
\end{aligned}$$

Avec $\hat{f}$ la densité du mélange complet déterminée à partir des différents paramètres estimés par le modèle et $\hat{F}$ la fonction de répartition associée. La constante ϵ est généralement fixée à 0.1. Enfin, cette méthode nous permet de déterminer U que l'on peut nommer *incertitude*.

$$U(y_0) = 1 - \max_{s\in\{1,\dots,n\}} \mathbb{P}(S = s \mid y_0 - \epsilon < y < y_0 + \epsilon) \in \left[0, \frac{1}{n}\right].$$

On considère alors la classification incertaine si $U(y_0) > H_u$ où H_u est un seuil légèrement inférieur à $\frac{1}{n}$. Cette seconde méthode peut être illustrée avec le graphique en figure 2.13.

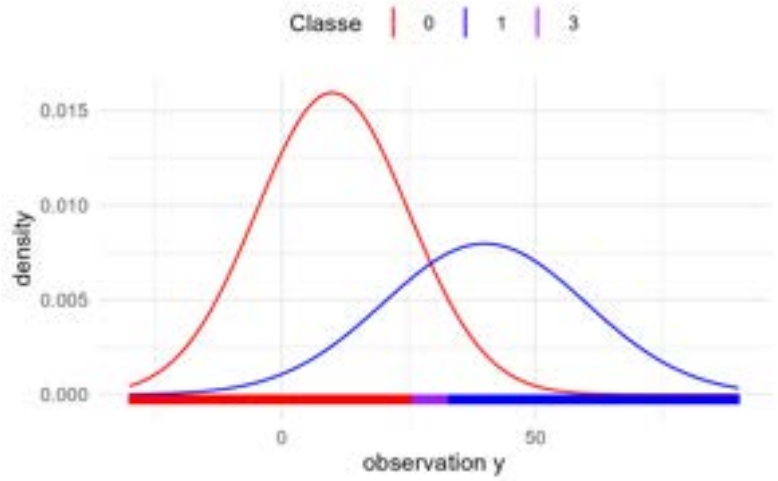


FIGURE 2.13 – Graphique de distribution illustrant la méthode de prédiction à l'aide de l'*optimal Bayes rule*.

Les paramètres des distributions de la figure 2.13 sont $\mu_0 = 10$, $\mu_1 = 40$, $\sigma_0 = 15$, $\sigma_1 = 20$, $w = 0$,$\pi_1 = 0.4$. La classe 3 correspond aux points dont l'incertitude est supérieure à $H_u = 0.4$.

Enfin, une dernière possibilité, plus complète, permet la prédiction des probabilités pour un lot de nouvelles observations. Elle consiste simplement en l'application des procédures *Forward* et *Backward* dans un cadre déterministe. Il suffit de disposer d'un vecteur de nouvelles observations pour un ou plusieurs troupeaux, ainsi que de l'ensemble des estimations obtenues à l'aide du modèle bayésien, afin d'obtenir les estimations des probabilités et donc des statuts infectieux pour chaque mois. Si plusieurs troupeaux sont utilisés il est alors possible d'avoir un aperçu de la prévalence pour chaque mois en appliquant un seuil sur les probabilités et en déterminant le ratio $\frac{\text{Nombre de troupeaux infectés}}{\text{Nombre total de troupeaux}}$.

Afin de tester cette méthode de prédiction nous avons simulé 100 jeux de données de 20 troupeaux avec 100 observations chacun. Les paramètres du mélange gaussien utilisés sont les mêmes que ceux de la figure 2.13, nous ajoutons à ceux-là les paramètres de la dynamique $\pi_1 = 0.4$, $\tau_1 = 0.1$ et $\tau_2 = 0.85$. À la différence du modèle bayésien les probabilités déterminées à partir de la variable *backward* ont été implémentées car le coût de calcul est ici largement inférieur et elles étaient supposées améliorer les performances de la prédiction.

Nous représentons en figure 2.14 l'évolution des sensibilités et spécificités selon le seuil considéré. Il est important de comprendre que ces sensibilités et spécificités diffèrent des précédentes car elles proviennent d'un seuillage sur les probabilités prédites et non d'un seuillage sur les mesures d'un test sérologique. Ici elles permettent d'évaluer la

prédiction tandis que précédemment elles définissaient les caractéristiques des tests utilisés.

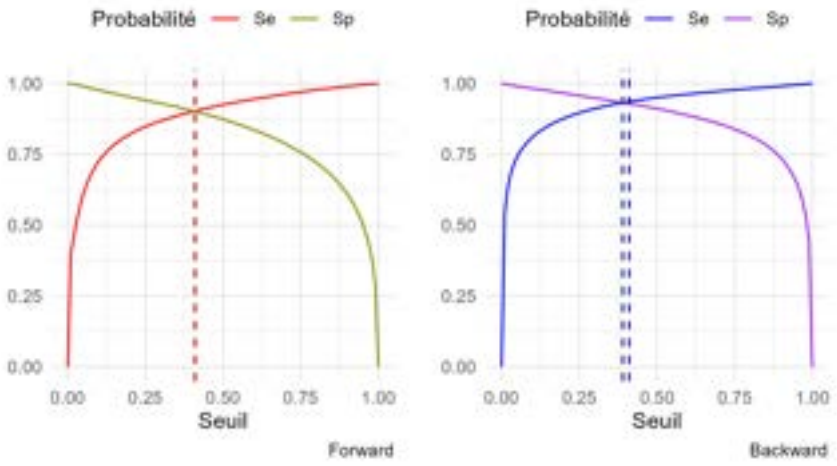


FIGURE 2.14 – Sensibilités et spécificités des procédures en fonction du seuil s.

Le tracé gauche de la figure 2.14 correspond à celui de la procédure *Forward*, le second celui de la procédure *Backward*.

Il existe différents critères d'optimisation de ces grandeurs, nous pouvons par exemple chercher le seuil $\mathrm{H}_{=}$ qui donne $Se = Sp$, une autre alternative est l'optimisation du J_Y de Youden qui se définit par $\mathrm{J}_Y = Se + Sp - 1$.

		Procédure	
		Forward	Backward
Critère	$\mathrm{H}_{=}$	0.41	0.39
	J_Y	0.41	0.41

TABLE 2.4 – Tableau des seuils optimaux pour la prédiction des statuts infectieux par l'algorithme Forward-Backward déterministe.

Les seuils optimaux répertoriés dans la table 2.4 sont représentés sur la figure 2.14 à l'aide des lignes pointillées. Ces lignes sont confondues pour la procédure forward.

Nous pouvons accompagner cette analyse d'une courbe ROC résumant les performances des deux procédures.

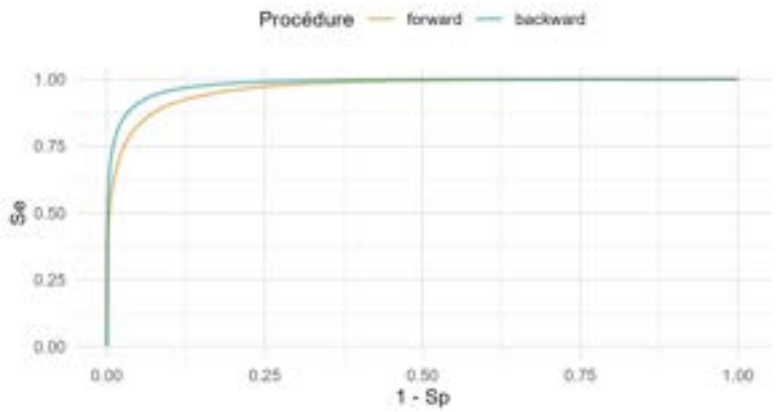


FIGURE 2.15 – Courbe ROC pour les procédures *Forward* et *Backward* déterministes

La figure 2.15 illustre bien la meilleure prédiction de l'état du troupeau à partir des probabilités définies par la variable *backward*. Quel que soit le critère d'optimisation choisit nous obtenons la combinaison $(Se = 0.9, Sp = 0.9)$ dans le cas de la procédure *forward* et $(Se = 0.93, Sp = 0.93)$ pour la seconde. Enfin, la distribution des prévalences des deux algorithmes est donnée figure 2.16.

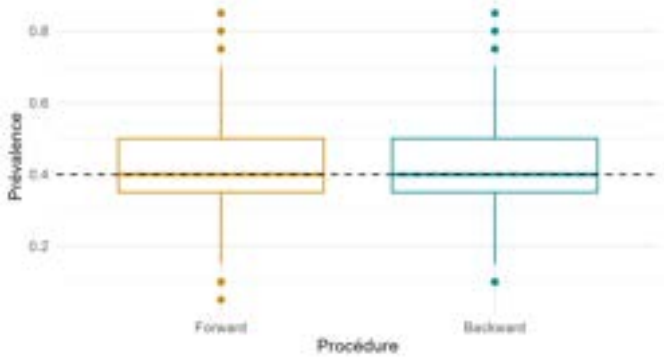


FIGURE 2.16 – Boxplot des prévalences pour les procédures *forward* et *backward* déterministes.

Concernant les prévalences nous obtenons des résultats identiques pour les deux procédures. Chacune d'elles fait preuve de très bonnes performances, du moins dans le cas de paramètres de distribution et dynamique relativement favorables. Nous observons un léger avantage pour la méthode *backward*, un résultat espéré du fait de probabilités déterminées à chacun des pas de temps en connaissant l'ensemble des observations.

Pour conclure cette partie sur la prédiction des probabilités d'infection nous pouvons évoquer le fait que malgré une utilisation en pratique pertinente des prédictions

dans le cadre déterministe il convient de rester prudent sur leur utilisation. En effet, les distributions considérées connues proviennent en réalité d'estimateur de bayes déterminé à partir des lois a posteriori du modèle. Il est difficile de confirmer la validité de l'utilisation des lois a posteriori de cette manière. Enfin, le cadre déterministe nous fait perdre la capacité des statistiques bayésiennes à évaluer l'erreur commise notamment à l'aide des intervalles de crédibilité.

Nous avons, dans l'ensemble de cette partie, évoqué des données simulées nous permettant de juger nos capacités prédictives. Il convient maintenant de présenter notre méthode de simulation ainsi que le plan d'expérience mis en œuvre afin de justifier de l'avantage des observations continues sur les observations discrètes.

2.4 Simulation des données

Afin de juger des qualités et limites d'un modèle il est important de l'appliquer sur données simulées dans un premier temps. De cette manière nous contrôlons l'ensemble des paramètres que nous souhaitons estimer. Dans un premier temps nous détaillons certains éléments de la fonction de simulation. Une seconde partie permettra de poser le plan d'expérience dont nous présentons les résultats en partie 3.

2.4.1 Méthode de simulation

Afin de simuler nos données il a été créé puis perfectionné une fonction *sim_mast_BVD*. Nous commençons en exhibant les arguments importants en entrée de cette fonction.

La possibilité est donnée de choisir le nombre de troupeaux qui correspond alors au nombre d'individus de l'expérience. Ensuite il est nécessaire d'indiquer le nombre de mesures souhaité, qui va correspondre au nombre de pas de temps pour chacun des troupeaux. L'écart entre chaque résultat de test est d'un mois par défaut mais il est donné la possibilité de le modifier pour coller aux données réelles. De plus, la fonction prend en entrée la matrice de transition permettant notamment de préciser les valeurs de τ_1 et τ_2. Elle est accompagnée des matrices des moyennes et écarts-types des composantes du mélange ainsi que de la valeur du w introduit en fin de partie 2.2.5. Si l'on désire simuler des observations discrètes il faut remplacer les matrices des paramètres $(\mu_i)_{i=0,1}$ et $(\sigma_i)_{i=0,1}$ par une matrice spécifiant les sensibilités (Se) et spécificité (Sp) des tests.

La fonction de simulation démarre avec la récupération des valeurs de τ_1 et τ_2 afin de déterminer π_1 à l'aide de la formule (2.2) donnée en partie 2.2.3. Nous simulons ainsi des chaînes dans un état déjà stationnaire. La séquence des dates permettant simplement de mimer les données réelles elle débute à la date du jour. Ensuite, il nous faut l'état initial

qui est généré à l'aide d'une multinomiale à deux classes $\mathcal{M}(1; 1-\pi_1, \pi_1)$ afin d'obtenir l'état sous forme de one hot encoding. Afin de générer les états suivants on utilise le vecteur des états précédents pour déterminer la ligne de la matrice de transition à conserver et on génère une nouvelle fois à l'aide d'une multinomiale employant cette fois-ci les probabilités de transitions sélectionnées.

Après obtention de l'ensemble des états il nous faut créer la variable des observations qui servira pour l'apprentissage du modèle de Markov caché. Dans le cas discret nous utilisons simplement une $\mathcal{B}(1, Sp)$, soit une Bernoulli de probabilité Sp, si nous sommes dans l'état 0 et une $\mathcal{B}(1, Se)$ pour l'état 1. Cette méthode de simulation nous permet de générer des données discrètes sans discrétiser à partir de résultats initialement continus. Concernant ces derniers, dans le cas des individus sains (0) nous réalisons une $\mathcal{B}(1, w)$ qui induit une mesure égale à 0 (saturation en 0) si c'est un succès et une réalisation d'une $\mathcal{N}(\mu_0, \sigma_0)$ dans le cas contraire. Pour le second état nous tirons simplement une réalisation d'une $\mathcal{N}(\mu_1, \sigma_1)$. En pratique nous nous contentons simplement de générer des données continues qui sont ensuite discrétisées si l'on souhaite générer des données discrètes. La fonction se termine sur le remplacement des observations par des données manquantes s'il a été demandé un intervalle de temps supérieur à 1 mois entre plusieurs émissions.

L'ensemble de cette fonction nous permet d'obtenir des jeux de données artificiels mimant les données réelles en maîtrisant l'ensemble des paramètres estimés par le modèle bayésien. Elles joueront un rôle important pour la mise en exécution du plan d'expérience proposé dans la partie qui suit.

2.4.2 Plan d'expérience

Nous avons débuté les tests en vérifiant si les paramètres des distributions étaient correctement estimés dans le cas d'un modèle simple sans dynamique et donc sans algorithme forward. Les résultats avaient l'avantage d'être obtenus rapidement et ont été concluants. Nous nous contentons de cela pour l'évocation de cette phase, celle-ci étant seulement présente pour justifier la poursuite de notre approche.

Ensuite, nous avons soumis le HMM avec observations continues à différents contrôles. L'objectif était de vérifier que l'ajout de la dynamique, et notamment de paramètres supplémentaires à estimer, ne dégradait pas les résultats. De plus, nous souhaitions détecter les limites du modèle, en particulier lorsque les composantes du mélange commencent à se confondre. Nous l'avons donc testé sur des jeux de données simulées dont la variation des paramètres est résumée dans le tableau 2.5.

Combinaisons	π_1	σ_1	σ_2
1	0.2	0.05	0.05
2	0.2	0.05	0.2
3	0.2	0.2	0.05
4	0.2	0.2	0.2
5	0.4	0.05	0.05
6	0.4	0.05	0.2
7	0.4	0.2	0.05
8	0.4	0.2	0.2

[1] Pour chacunes des combinaisons $w = 0.6$, $\mu_0 = 1$, $\mu_1 = 2$, $\tau_1 = 0.1$, $\tau_2 = 1 - \frac{\tau_1(1-\pi_1)}{\pi_1}$.

TABLE 2.5 – Plan d'expérience pour les paramètres π_1, σ_1, σ_2.

L'écart entre les différentes valeurs d'un même paramètre étant relativement faible nous nous permettons d'utiliser les mêmes paramétrages pour les lois a priori des 8 combinaisons. Celles-ci sont données dans le tableau 2.6.

Paramètre	Prior
w	Beta(4, 8)
π_1	Beta(4, 8)
τ_1	Beta(2, 10)
τ_2	Beta(9, 3)

[1] Pour μ_0 et μ_1 nous utilisons le paramétrage présenté en (2.6) en partie 2.2.5.

TABLE 2.6 – Table des lois a priori posées sur les paramètres pour la première phase du plan d'expérience.

Pour chacune des combinaisons nous générons 1000 tableaux de 400 observations pour un seul troupeau.

Suite à cela nous avons souhaité observer le comportement du modèle sur des jeux de données plus proches du réel, tout en évaluant les performances du nouveau modèle par rapport à celles de son prédécesseur. Pour ce faire nous avons essayé 6 combinaisons de paramètres avec 1500 jeux de 500 observations pour un seul troupeau. Cependant, ici nous supprimons le paramètre w qui permet essentiellement de débruiter les données et qui n'est pas tout à fait conventionnel. De cette manière nous testons le premier mélange présenté en partie 2.2.5 sur des données propres sans pic de densité en 0. Les résultats binaires nécessaires au modèle original sont créés en discrétisant la variable

continue à l'aide d'un seuil h. Cette discrétisation nous donne des valeurs de Se et Sp déterminées à partir des égalités (2.7). Nous fixons $\pi_1 = 0.4$, $\sigma_1 = 15$, $\sigma_2 = 20$ tandis que le reste des paramètres est résumé dans le tableau 2.7.

Combinaisons	τ_1	τ_2	μ_1	μ_2	Seuil (h)	Se	Sp
1	0.1	0.85	25	75	50	0.95	0.95
2	0.1	0.85	25	50	37.5	0.79	0.79
3	0.1	0.85	30	50	37.5	0.79	0.69
4	0.1	0.85	25	45	37.5	0.69	0.79
5	0.2	0.7	25	75	50	0.95	0.95
6	0.4	0.4	25	75	50	0.95	0.95

[1] Les combinaisons 5 et 6 sont identiques à la première excepté sur les valeurs de τ_1 et τ_2 afin de faire varier la dynamique du HMM.

TABLE 2.7 – Résumé des paramètres pour la simulation des données utilisées pour comparer les deux modèles bayésiens.

Ce tableau a été déterminé avec l'objectif de comparer les deux modèles sur des composantes du mélange gaussien bien différenciées dans un premier temps. Ensuite nous avons diminué les sensibilités et spécificités en augmentant la superposition de ces composantes. La spécificité a alors été de nouveau diminuée puis nous avons interverti les valeurs de Se et Sp. Enfin, les deux dernières combinaisons permettaient d'étudier l'influence possible des paramètres de la dynamique sur les estimations.

Nous précisons qu'une première étape a consisté en la réalisation de 3 inférences pour la première combinaison en faisant décroître la quantité d'information fournie par les lois a priori. Les résultats obtenus étant positifs nous nous sommes fixés sur le tableau 2.7 avec les lois a priori qui suivent.

Pour μ_1 et μ_2 nous réutilisons le paramétrage précédent introduit en partie 2.2.5. Ensuite, nous avons 3 variations pour Se et Sp, on rappelle que ces valeurs étant des probabilités nous utilisons des lois Beta.

Se,Sp	Prior
0.69	Beta(22.8, 19.3)
0.79	Beta(19.7, 10.7)
0.95	Beta(13.26, 3.27)

TABLE 2.8 – Table des lois a priori posées sur les paramètres Se et Sp pour la deuxième phase du plan d'expérience.

Le tableau 2.8 répertorie les lois a priori utilisées pour cette seconde phase du plan d'expérience pour Se et Sp. Les valeurs des α et β sont fixées de telles sortes que l'espérance des trois lois Beta est égale à $Se - 0.15$ (respectivement $Sp - 0.15$). De cette manière nous évitons de biaiser nos comparaisons en apportant des informations plus importantes pour certaines combinaisons.

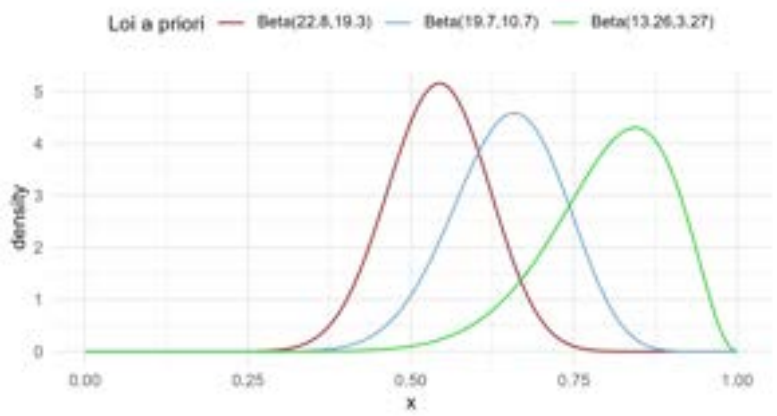


FIGURE 2.17 – Distribution des lois a priori des paramètres Se et Sp

La loi a priori sur π_1 est une Beta(4,8). Comme pour Se et Sp nous répertorions les lois a priori sur τ_1 et τ_2 en table 2.9.

τ_1	Prior
0.1	Beta(2, 10)
0.2	Beta(3, 9)
0.4	Beta(4, 8)

τ_2	Prior
0.4	Beta(4, 8)
0.7	Beta(8, 4)
0.85	Beta(9, 3)

TABLE 2.9 – Lois a priori des paramètres τ_1 et τ_2 pour la seconde phase du plan d'expérience.

L'ensemble du plan que nous appliquons sur données simulées étant maintenant établi il est temps de passer à la présentation des résultats.

Chapitre 3

Résultats

3.1 Application des modèles sur données simulées

3.1.1 Évaluation du modèle avec observations continues

La première phase du plan d'expérience consiste à vérifier que le nouveau modèle fonctionne correctement et d'en déterminer les limites. Nous allons uniquement nous intéresser aux estimations des moyennes et de la prévalence ainsi que du paramètre w, ce dernier étant une spécificité de notre modèle avec observations continues. Le reste des boxplots est à retrouver en annexe (B.1).

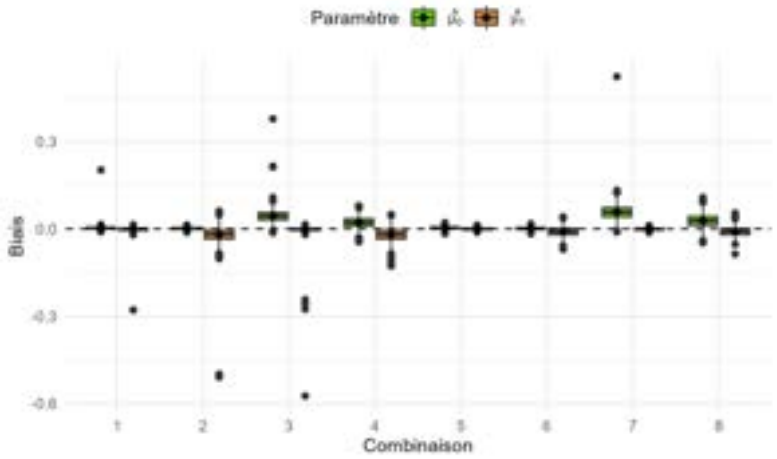


Figure 3.1 – Boxplot du biais des estimations pour les moyennes μ_0 et μ_1 du mélange.

On rappelle que la figure 3.1 a été obtenue en réalisant 1000 simulations différentes. On remarque un biais en moyenne proche de 0, avec cependant la présence de quelques

estimations aberrantes sur les combinaisons 3,4,7 et 8. C'est sur ces mêmes combinaisons que l'on devine un biais un peu plus important.

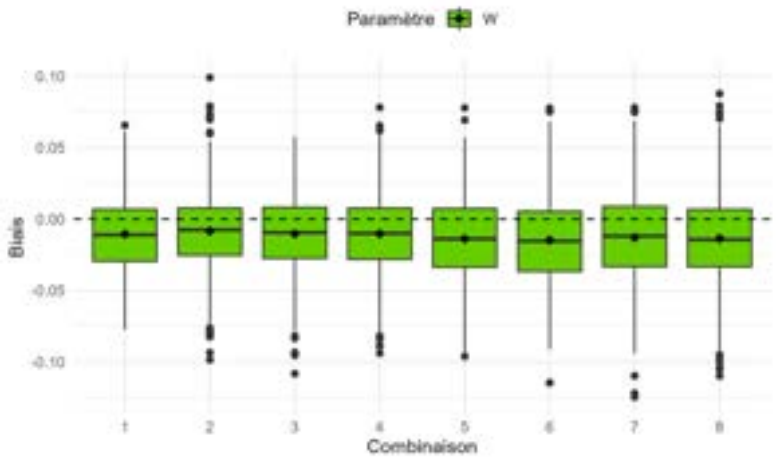


FIGURE 3.2 – Boxplot du biais des estimations pour le paramètre w.

Ensuite, on pouvait craindre une estimation difficile pour w mais finalement on découvre un biais très proche de 0 et tout à fait acceptable pour une valeur réelle du paramètre de 0.6. Cela semble justifier que cette composante a toute sa place dans le mélange et peut correctement aider à débruiter les données.

On termine cette partie en effectuant un zoom sur l'estimation de π_1.

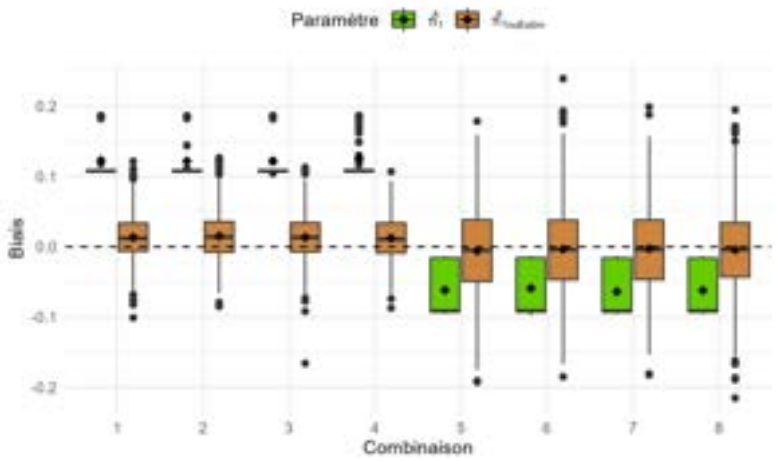


FIGURE 3.3 – Biais des estimations pour les paramètres π_1 et $\pi_{1_{noEstim}}$.

Sur cette figure 3.3 apparaît en vert le biais de $\hat{\pi_1}$, il se trouve que dans les 4 premières

combinaisons le paramètre est sans arrêt sur-estimé et est ensuite sous-estimé dans les 4 suivantes. Il est possible que π_1 intervenant uniquement dans l'initialisation de l'algorithme forward la loi a priori utilisée jouent un rôle prédominant dans l'estimation finale. En effet, on rappelle que la loi a priori utilisée ici est une Beta$(4, 8)$ et est donc de moyenne 0.33 avec π_1 qui vaut 0.2 dans les 4 premières combinaisons et 0.4 dans les 4 dernières. Nous avons alors envisagé la possibilité de déterminer la prévalence à l'aide des chaînes de τ_1 et τ_2. On rappelle que les données sont simulées à l'aide d'une chaîne de Markov qui se trouve déjà dans un état stationnaire, de ce fait nous disposons de l'égalité (2.2). Le graphique illustre que l'on obtient une meilleure estimation de π_1 avec $\hat{\pi}_{1_{noEstim}}$. Ceci était attendu du fait des bonnes estimations de τ_1 et τ_2 (Annexe B.1).

On peut résumer l'ensemble de ces observations en utilisant les pourcentages de couverture des intervalles de crédibilité, on donne ces pourcentages pour les intervalles de niveau 95% mais on trouve aussi les niveaux 90% et 99% en annexe B.2.

	Scénarios							
Variable	1	2	3	4	5	6	7	8
μ_0	0.92	0.94	0.34	0.78	0.89	0.96	0.27	0.75
μ_1	0.94	0.88	0.95	0.90	0.95	0.91	0.95	0.92
$\hat{\pi}_{1_{noEstim}}$	0.96	0.96	0.97	0.96	0.93	0.94	0.94	0.94
σ_0	0.96	0.95	0.92	0.94	0.95	0.95	0.93	0.94
σ_1	0.95	0.94	0.96	0.96	0.94	0.96	0.95	0.95
τ_1	0.95	0.96	0.95	0.96	0.96	0.95	0.94	0.95
τ_2	0.94	0.94	0.95	0.95	0.94	0.95	0.95	0.95
w	0.95	0.93	0.94	0.94	0.93	0.93	0.92	0.92

TABLE 3.1 – Pourcentages de couverture pour l'ensemble des paramètres du modèle continu // niveau 95%

Ce tableau confirme une faiblesse du modèle sur le paramètre μ_0 pour les combinaisons 3,4,7 et 8. Ce résultat était envisageable puisque ce sont les combinaisons pour lesquelles les distributions sont moins différenciées. Afin de justifier ces mauvais taux de couverture on peut ajouter les écarts-types des estimations $\hat{\mu_0}$ et $\hat{\mu_1}$.

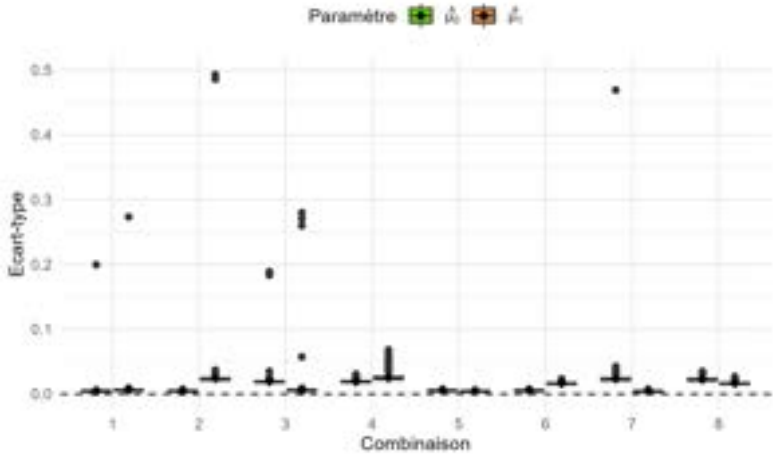


FIGURE 3.4 – Écarts-types des estimations pour les paramètres μ_0 et μ_1.

La figure 3.4 indique que les écarts-types, malgré des valeurs extrêmes sur certaines combinaisons, sont extrêmement faibles et entraîne donc une valeur réelle exclue des intervalles de crédibilité dès qu'un léger biais est présent (figure 3.1).

Cependant, il est notable que les pourcentages sont particulièrement moins bons pour les combinaisons 3 et 7 alors que les distributions sont plus différenciées que pour les scénarios 4 et 8. On peut faire une dernière remarque, le même comportement se répète dans les deux groupes de 4 combinaisons, le seul élément différenciant ces deux groupes étant la valeur de π_1. Il semble donc que la valeur du paramètre n'ait pas une grande influence sur les estimations données par le modèle. En réalité, son influence se restreint essentiellement à la quantité de données dans chacun des groupes (sains/malades) et donc influence directement le nombre de données disponibles pour l'apprentissage du modèle.

Le reste des graphiques disponibles en annexe B.1 permet de valider les bonnes performances du HMM dans le cas d'observations continues. Nous pouvons donc poursuivre avec la comparaison de ce nouveau modèle à son prédécesseur.

3.1.2 Comparaison des deux modèles

Dans cette sous-partie nous nous intéressons aux résultats de la seconde phase du plan d'expérience. Tous les graphiques ne seront pas donnés ici non plus, ils serviront simplement à illustrer et commenter les performances des deux modèles.

Puisque nous souhaitons mettre en parallèle les deux modèles il convient de trouver un point de comparaison. Celui-ci se situe dans les paramètres Se et Sp. En effet, ils constituent la spécificité du modèle avec observation discrètes mais peuvent aussi être

déterminés de manière déterministe dans le cas du modèle avec observations continues. On donne uniquement les boxplots biais de $\hat{\mathrm{Sp}}$ mais les observations pour $\hat{\mathrm{Se}}$ sont identiques.

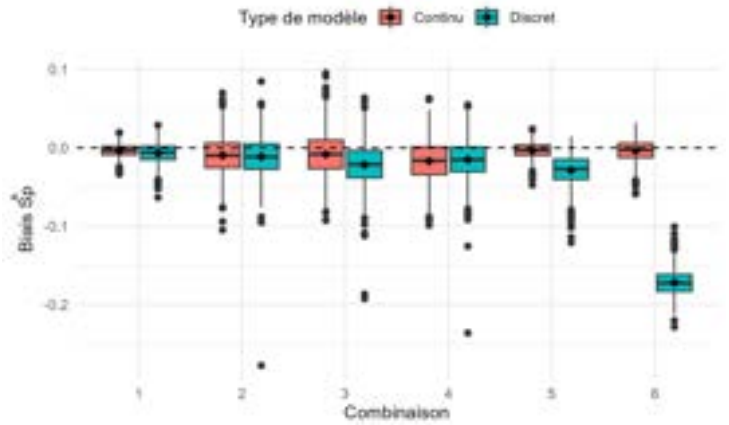


FIGURE 3.5 – Boxplots des biais des estimations du paramètre Sp par les deux modèles bayésiens.

Pour le modèle continu les sensibilités et spécificités ont encore été établies à partir des expressions (2.7), en utilisant les chaînes de Markov de μ_0 et μ_1 estimées par le HMC. On remarque rapidement en figure 3.5 un biais non négligeable pour le modèle discret dans le cas de la sixième combinaison. Pour le reste des scénarios les deux modèles présentent un biais proche de 0, avec un léger avantage pour le modèle à émissions continues. Un autre avantage est que les estimations sont plus précises comme le montre la figure 3.6.

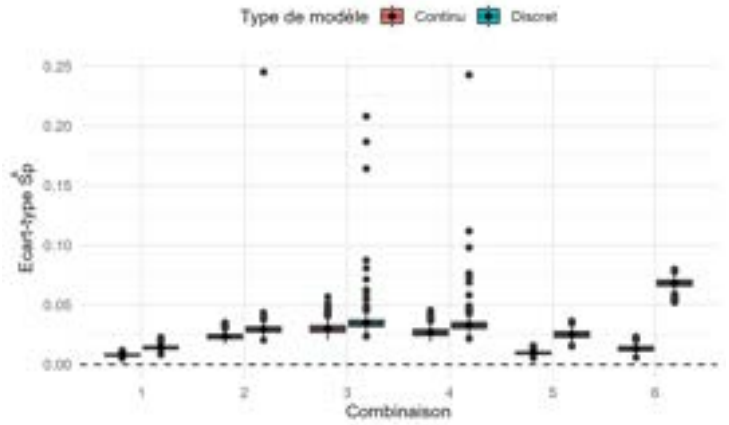


FIGURE 3.6 – Boxplots des écarts-types des estimations du paramètre Sp par les deux modèles bayésiens.

L'ensemble des graphiques en annexe C viennent confirmer nos observations avec une estimation biaisée du modèle discret sur la dernière combinaison et le modèle continu qui semble globalement plus pertinent. De plus, cette fois-ci aussi $\hat{\pi_1}_{noEstim}$ permet d'améliorer les estimations, et ce pour les deux modèles (Annexe C pour le biais de $\hat{\pi_1}$). L'avantage pour le nouveau modèle persiste sur l'estimation de π_1 comme illustré sur la figure 3.7.

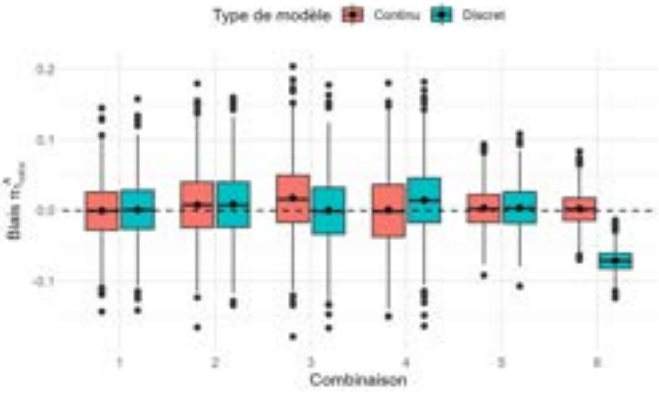


FIGURE 3.7 – Boxplots des biais des estimations du paramètre π_1 à partir des chaînes de τ_1 et τ_2.

Enfin, les taux de couverture des intervalles de crédibilité permettent de résumer efficacement les résultats des modèles pour l'ensemble des paramètres.

		Combinaisons					
Variable	Type de modèle	1	2	3	4	5	6
Se	Continu	0.94	0.91	0.87	0.94	0.93	0.93
	Discret	0.96	0.96	0.98	0.95	0.89	0.70
Sp	Continu	0.94	0.94	0.96	0.92	0.96	0.93
	Discret	0.97	0.98	0.97	0.99	0.91	0.00
$\pi_{1_{noEstim}}$	Continu	1.00	1.00	1.00	1.00	1.00	1.00
	Discret	0.95	0.99	1.00	0.99	0.99	1.00
τ_1	Continu	0.94	0.97	0.98	0.98	0.94	0.95
	Discret	0.95	0.97	0.99	0.99	0.87	0.99
τ_2	Continu	0.95	0.96	0.96	0.98	0.94	0.96
	Discret	0.96	0.97	0.99	0.98	0.82	1.00

TABLE 3.2 – Pourcentages de couverture des intervalles de crédibilité de niveau 95% pour les paramètres communs aux deux modèles, obtenus sur $B = 1500$ simulations.

Dans l'ensemble les taux de couverture de la table 3.2 sont très bons et approchent, voire dépassent, les 95%. Excepté pour le modèle original dans quelques situations visiblement défavorables.

Cette partie était consacrée à la démonstration des bonnes performances des modèles sur données simulées, et en particulier aux bénéfices du nouveau modèle sur l'original. Ceci étant fait nous terminons avec l'examen des résultats du modèle continu sur notre jeu de données présenté en 2.1.

3.2 Application aux données réelles

3.2.1 Échantillon de données entre 2018 et 2021

La première étape sur données réelles avait pour objectif de débuter tranquillement avec une application sur un fragment de nos données favorable à l'utilisation du modèle. On peut commencer par montrer que l'échantillonnage par l'algorithme MCMC se fait convenablement. Pour le modèle continu sans w nous avons par exemple les graphiques en figure 3.8 pour les moyennes et écarts-types des distributions.

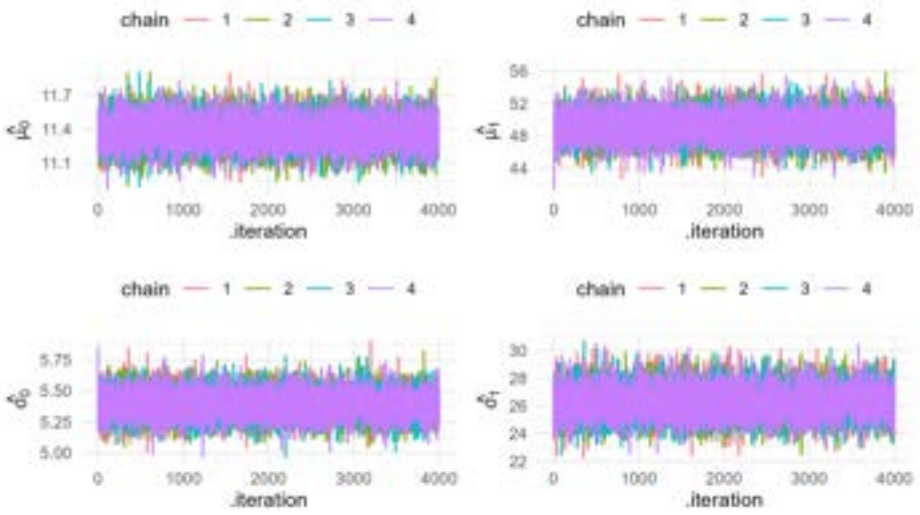


Figure 3.8 – Mélange des chaînes de Markov par l'algorithme MCMC du modèle continu (sans w) pour les paramètres des distributions.

Les résultats sont similaires pour le modèle avec le poids w, cependant le modèle discret peut présenter des soucis dans le mélange. En effet, il arrive que certaines chaînes mélangent en des valeurs différentes des autres chaînes.

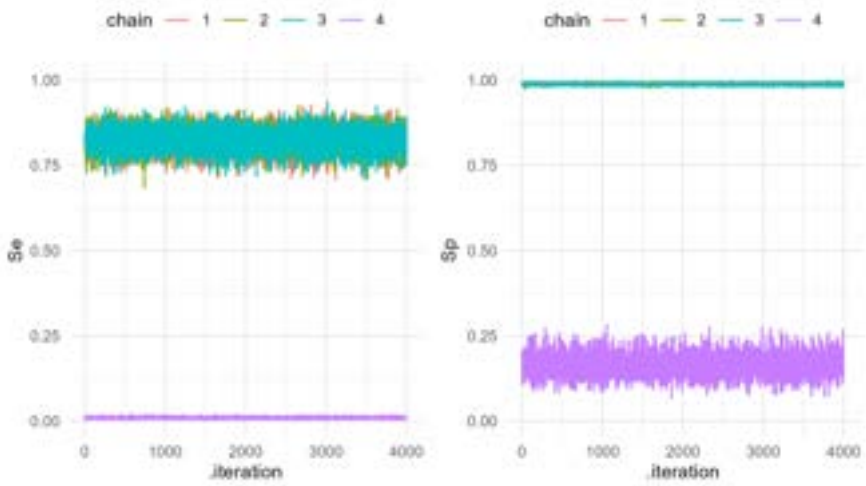


FIGURE 3.9 – Mélange des chaînes de Markov par l'algorithme MCMC du modèle discret pour Se et Sp.

On voit sur la figure 3.9 que la quatrième chaîne mélange systématiquement en des valeurs qui semblent erronées. Il est possible que ce modèle soit plus sensible au faible nombre de données disponibles dans le groupe des troupeaux infectés puisque la discrétisation entraîne une perte d'information certaine.

Les 3 modèles ayant été appliqué à notre échantillon de données nous pouvons comparer les résultats à l'aide des intervalles de crédibilité sur les estimations. Par exemple pour τ_2 nous avons le graphique suivant.

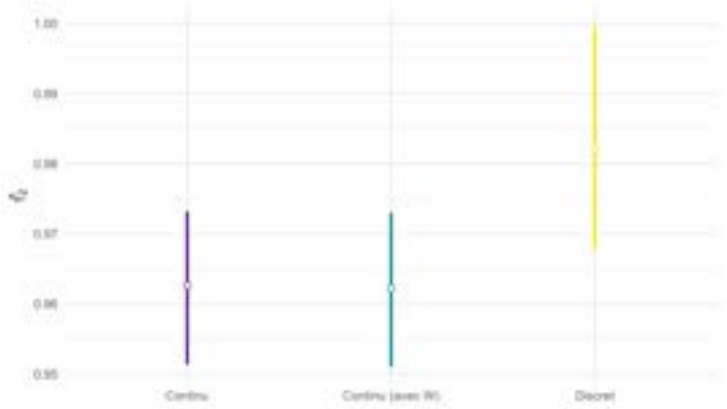


FIGURE 3.10 – Intervalles de crédibilité à 95% de l'estimation de τ_2, par les 3 modèles sur notre échantillon de données réelles.

D'après la figure 3.10 les deux modèles continus donnent des estimations identiques tandis que le modèle discret indique une probabilité de rester malade plus élevée. Peu observable ici le phénomène de mauvais mélange chez le modèle discret peut cependant venir fausser les intervalles de crédibilité quand les deux modèles continus seront eux certains de leurs estimations respectives. Il est intéressant de regarder les estimations de μ_0 par ces deux modèles continus.

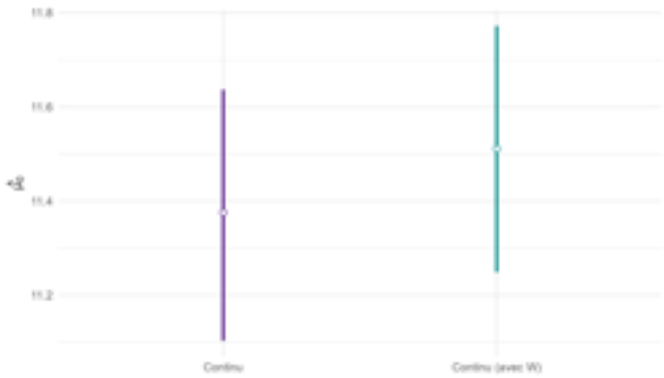


FIGURE 3.11 – Intervalles de crédibilité à 95% de l'estimation de μ_0, par les 2 modèles avec observations continues sur notre échantillon de données réelles.

Nous supposions qu'une saturation des données en 0 pouvait entraîner une sous-estimation de μ_0. Comme espéré la figure 3.11 indique que l'estimation de la moyenne du groupe sain est un peu plus élevée dans le cas du second modèle. La différence est légère car nous avons fait en sorte de sélectionner un échantillon peu saturé en 0. L'ajout de la composante en 0 paraît donc judicieuse pour aider à détecter la réelle distribution des mesures de tests pour les troupeaux sains.

Nous pouvons maintenant superposer les graphiques de densité des lois normales estimées aux distributions réelles, dans un premier temps sur la distribution complète. Nous choisissons d'utiliser les estimations du modèle incluant la composante w mais en représentant un mélange à seulement deux composantes. En effet nous considérons le paramètre w comme un élément nous permettant de détecter les distributions non-bruitées et non comme un poids réellement présent dans le mélange.

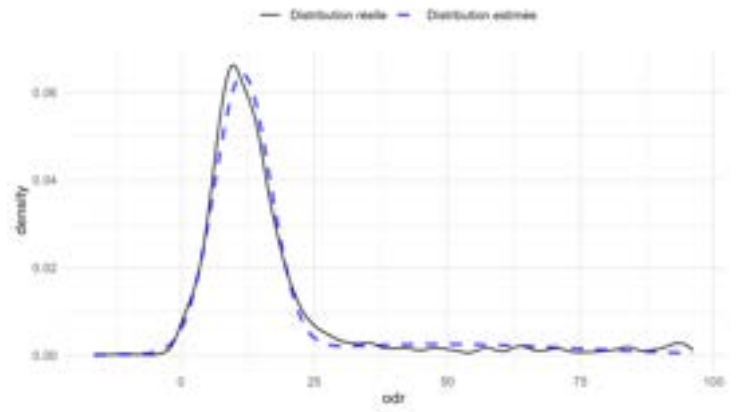


FIGURE 3.12 – Distributions réelles et estimées pour l'échantillon de données sélectionné.

Pour accompagner la figure 3.12 nous pouvons ajouter le graphique des distributions des groupes sains et infectés.

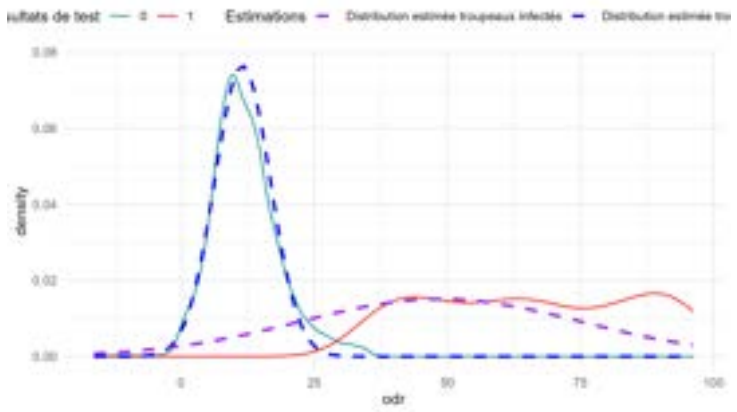


FIGURE 3.13 – Distributions réelles et estimées, groupes sains et infectés, pour l'échantillon de données sélectionné.

Que ce soit sur la figure 3.12 ou la 3.13 les estimations paraissent assez fidèles aux données réelles. On rappelle que la variable t_{res} indiquant les 0 et 1 des données réelles ne correspond pas aux statuts exacts. Elle est déterminée à partir du seuil $odr = 35$. Les distributions estimées étant superposées entre, environ, 0 et 25 cela indique une probabilité non négligeable de détecter des faux négatifs et ce quel que soit le seuil utilisé.

Pour terminer cette sous-partie il peut être intéressant de regarder les prédictions de la probabilité d'infection que rendent les 3 modèles pour quelques troupeaux au dernier pas de temps.

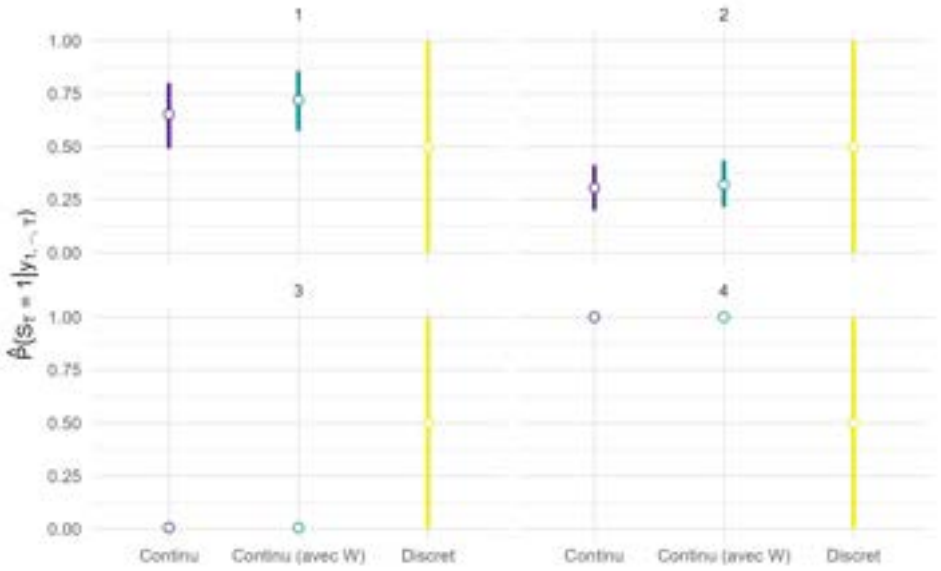


FIGURE 3.14 – Intervalles de crédibilité de niveau 95% pour les prédictions des probabilités de séropositivité de 4 troupeaux.

Les deux modèles utilisant les mesures continues produisent bien des résultats similaires en terme de prédiction de la probabilité de séropositivité. Cependant, on remarque bien sur la figure 3.14 que le souci de convergence du modèle discret impacte grandement cette prédiction. Il est fort probable que ce problème apparaisse du fait d'un faible nombre de mesures pouvant être attribuées à des troupeaux infectés et d'un déséquilibre entre les deux groupes. Nous l'observons notamment dans la suite de cette partie sur le modèle continu. Le modèle discret paraît toutefois bien plus sensible à ce phénomène.

3.2.2 Découpage des données en 28 périodes

Nous avons déjà évoqué à deux reprises le séquençage de nos données en 28 périodes. Le modèle continu incluant la composante en 0 a permis de déterminer l'ensemble des estimations souhaitées. Pour le paramètre w il est suffisant de dresser un tableau répertoriant les différentes espérances des lois a posteriori (Annexe D.2). On relève essentiellement une bien plus forte présence de saturation en 0 dans le cas du test LGMCAT. Les distributions des différentes lois a posteriori ont été obtenues pour chacune des variables du modèle et sont présentées sur la figure 3.15.

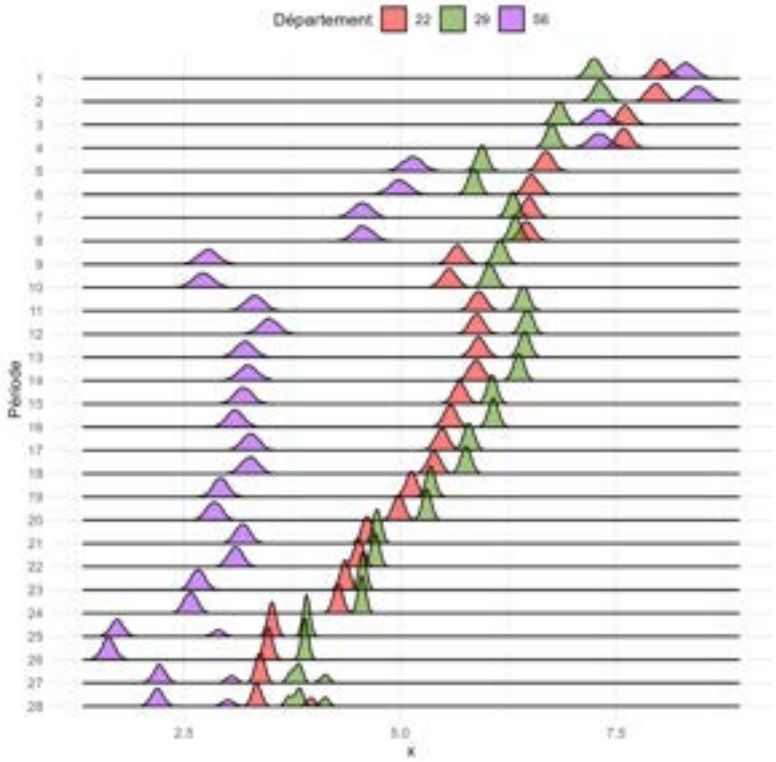


FIGURE 3.15 – Distributions a posteriori de μ_0 pour le test LGMCAT sur les 28 périodes.

On pourrait espérer des distributions centrées en des valeurs similaires pour les 3 départements avec une constance dans le temps. Cependant, on observe une décroissance de la moyenne tout au long des 28 périodes avec une différence entre les départements en particulier avec le Morbihan (56). Toutefois, on relève que ces distributions restent concentrées entre environ 2.5 et 7.5. On peut toutefois évoquer la présence de plusieurs modes sur les distributions des dernières périodes.

Si on s'intéresse aux paramètres de la dynamique on peut faire quelques remarques intéressantes. On donne cette fois-ci les graphiques des intervalles de crédibilité.

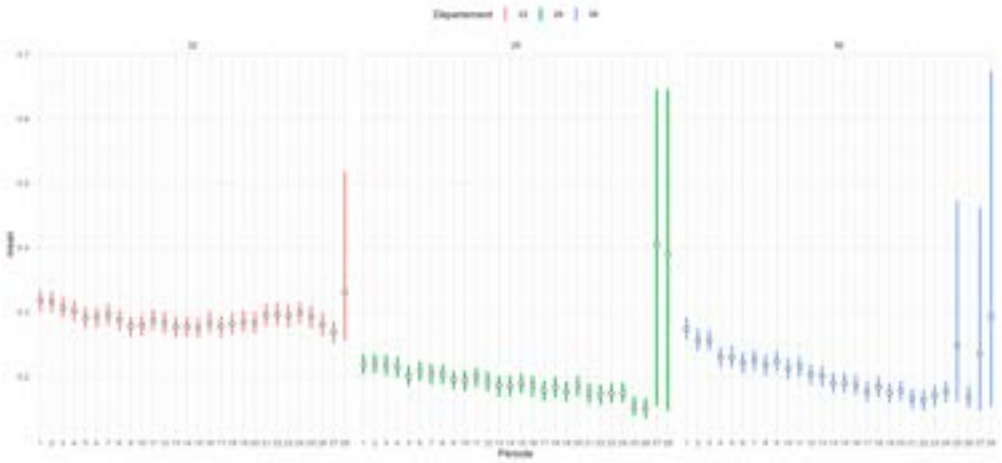


FIGURE 3.16 – Intervalles de crédibilité pour l'estimation de π_1 sur les 28 périodes.

Les intervalles de la figure 3.16 illustrent une décroissance de π_1 au fil du temps, soit une prévalence de la maladie qui semble diminuer quel que soit le département. On peut aussi comparer $\hat{\pi_1}$ à $\hat{\pi_1}_{noEstim}$ pour se rendre compte que les graphiques sont tout à fait similaires (Annexe D.4). Cette seconde prévalence étant déterminée à l'aide de l'expression à l'équilibre cela semble signifier que la BVD se trouve dans un état stationnaire à chaque pas de temps. On accompagne π_1 de la probabilité τ_1 de passer d'un état sain à infecté.

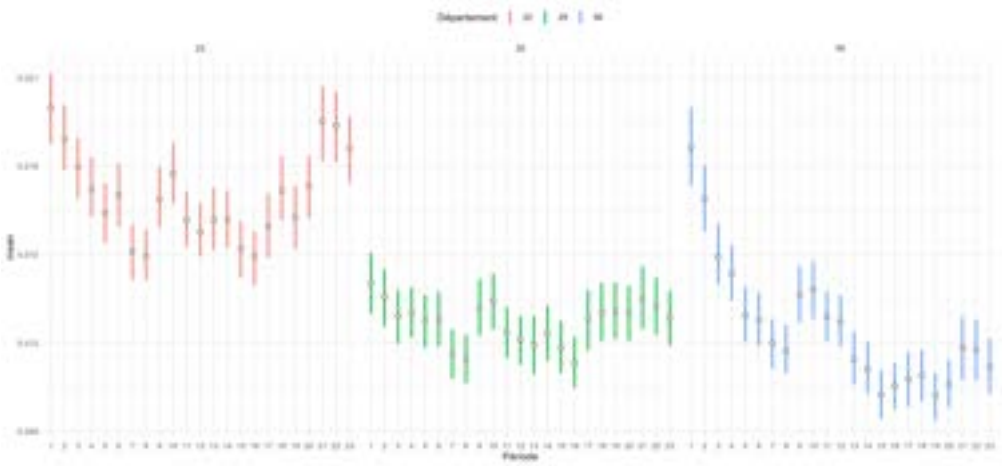
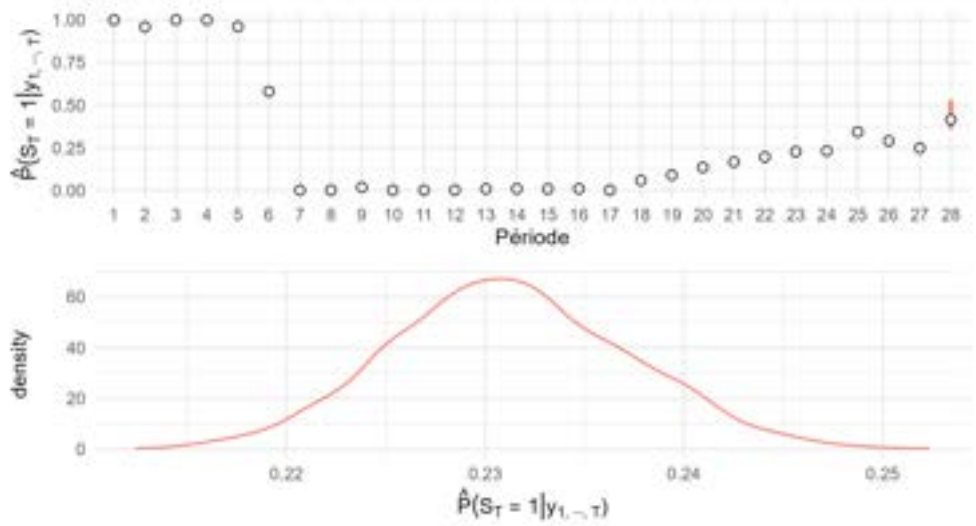


FIGURE 3.17 – Intervalles de crédibilité pour l'estimation de τ_1 sur les 24 premières périodes.

Pour la figure 3.17 nous avons supprimé les dernières périodes qui dégradaient la visualisation de l'évolution des estimations. Ainsi, on peut voir que la probabilité d'infection semble constante voire décroissante au fil du temps. L'important étant

surtout qu'elle ne soit pas croissante. Cette diminution vient supporter l'observation précédente concernant une baisse de la prévalence dans le temps. Les graphiques pour τ_2 (Annexe D.4) semblent aussi abonder dans la direction d'une baisse de la prévalence avec une probabilité de rester infecté qui paraît diminuer. Tout cela soutient le fait que les mesures mises en place pour la surveillance et l'éradication du virus sont efficaces. Il nous faut cependant mentionner les soucis sur les dernières périodes. En effet, la présence de plusieurs modes sur les distributions a posteriori pose problème dans l'utilisation de la moyenne a posteriori en tant qu'estimateur de Bayes. De plus, les intervalles de crédibilités sont complètement faussés. Il est, encore une fois, probable que ce défaut provienne d'un nombre d'observations trop faibles de troupeaux possiblement infectés (Annexe D.3).

Enfin, nous avons donné des exemples de graphiques pour la prédiction des probabilités en partie 2.3.1 et pouvons maintenant ajouter quelques remarques importantes. Nous illustrons ici les probabilités prédites dans le cas d'un troupeau du département 22.



[1] Les intervalles de crédibilité ne sont pas optimaux mais déterminés à partir des quantiles à 2.5% et 97.5%.
[2] La distribution a posteriori est donnée uniquement pour la période 24.

FIGURE 3.18 – Intervalles de crédibilité à 95% et distribution a posteriori de l'estimation de la probabilité d'être infecté pour un troupeau des Côtes-d'Armor (22).

Le premier graphique de la figure 3.18 nous permet de relever quelques éléments. Premièrement le troupeau semble être infecté sur les premières périodes puis devient sain tout à coup. Il semble ensuite y avoir une probabilité que le troupeau soit de

nouveau infecté qui augmente graduellement dans le temps. La volonté d'utiliser les données continues provient notamment de l'idée que la transition entre l'état sain et l'état infecté serait plus facilement observable, avec une augmentation progressive de l'odr et donc de même pour la probabilité $\mathbb{P}(S_T = 1 \mid y_{1,\dots,T})$. L'augmentation entre les périodes 7 et 28 que l'on observe en figure 3.18 semble aller dans le sens de notre hypothèse. Cependant, la transition entre les périodes 5 et 7 est relativement abrupte à la manière d'une discrétisation binaire. Une possible explication serait le fait que nous ne disposons pas de mesures de tests à chaque pas de temps mais uniquement tous les 3 à 6 mois. De ce fait, le troupeau a tout a fait le temps de passer d'un état infecté à sain, et inversement, sans que l'on ait pu détecter ce basculement.

Nous venons de détailler les résultats sur données simulées puis sur données réelles. Nous passons maintenant à leur discussion afin de conclure ce rapport.

Discussion et conclusion

Premièrement la partie 3.1.1 a permis de confirmer que l'estimation bayésienne fonctionne pour l'estimation des paramètres d'un HMM. S'ajoutent à cela des résultats qui semblent répondre par la positive à notre problématique de départ qui consistait à déterminer si la conservation des données continues améliorait les estimations. L'amélioration n'est pas drastique car le modèle original fonctionnait déjà correctement malgré quelques imperfections, cependant la marge d'amélioration reste importante avec par exemple des lois a priori sur les écarts-types du mélange qui sont actuellement très peu informatives.

Les résultats sur données réelles, en plus d'illustrer l'avantage du modèle continu sur le modèle discret, permettent d'émettre quelques observations. La composante du mélange caractérisant le groupe des troupeaux séropositifs semble un peu plus difficile à estimer. C'est probablement une conséquence directe de la valeur de prévalence qui réduit le nombre des mesures constitutives de ce groupe. De plus, le résultat obtenu pour la première application sur données réelles indique des distributions qui ne sont pas parfaitement différenciées. Cela signifierait un taux de faux négatifs significativement supérieur à 0 au moins pour le test LGMVEA. L'expérience sur les 28 périodes semble quant à elle témoigner de l'efficacité des mesures de surveillances mises en place avec une baisse de la prévalence dans le temps et de la probabilité d'infection, en particulier dans le département du Morbihan (56).

Un élément intéressant à ajouter serait la mise en perspective des probabilités prédites avec la détection d'individus IPI dans les troupeaux. En effet, les sauts dans la prédiction des probabilités entre différentes périodes peuvent provenir d'états qui ne sont pas observés entre les 3 mois séparant deux tests, mais ils peuvent aussi être la conséquence de l'apparition d'individus IPI dans le troupeau et la mise en place de mesures afin d'éliminer toute présence de la maladie.

Nous pouvons discuter des pistes d'amélioration du modèle bayésien. À l'origine il était souhaité l'ajout de facteurs de risque au modèle. Ceux-ci sont inclus dans le modèle d'origine par le biais d'une régression logistique sur le paramètre τ_1 (probabilité pour le troupeau de passer d'un état séronégatif à séropositif). Par manque de temps cela n'a pas pu être accompli pour le nouveau modèle. Les principaux facteurs envisagés sont

l'introduction de nouveaux individus dans le troupeau ainsi que la taille de ce dernier. On pourrait aussi y ajouter une dimension géographique avec la proximité d'autres troupeaux. Par ailleurs, un travail de recherche sur les lois a priori utilisées pourrait être réalisé. On peut notamment penser à l'ajout de lois normales avec une transformation logit et des lois plus informatives pour σ_0 et σ_1. Enfin, une idée évoquée au cours du stage était l'utilisation dans le mélange d'une loi de Pareto pour la composante de gauche afin de représenter la censure des données en 0.

Pour terminer, on peut rapidement parler des données disponibles pour une telle modélisation. La présence de saturation en différents endroits dans les données oblige la mise en place d'un modèle plus complexe et rend plus fastidieuse l'estimation des paramètres étudiés. Les performances du modèle sont aussi directement impactées par des mesures de test réalisées au minimum tous les 3 mois. L'implémentation du modèle continu semblant faire ses preuves il pourrait être intéressant de disposer de mesures plus fréquentes afin de faciliter les estimations et améliorer leur précision.

Bibliographie

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan : a probabilistic programming language. *Journal of Statistical Software*, *76*, 1-32. https://doi.org/10.18637/jss.v076.i01

Damiano, L., Peterson, B., & Weylandt, M. (2017). A Tutorial on Hidden Markov Models using Stan. https://github.com/luisdamiano/stancon18

Grooms, D. L. (2004). Reproductive consequences of infection with bovine viral diarrhea virus. *The Veterinary Clinics of North America. Food Animal Practice*, *20*(1), 5-19. https://doi.org/10.1016/j.cvfa.2003.11.006

Hamilton, J. D. (1989). A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle [Publisher : [Wiley, Econometric Society]]. *Econometrica*, *57*(2), 357-384. https://doi.org/10.2307/1912559

Houe, H. (1999). Epidemiological features and economical importance of bovine virus diarrhoea virus (BVDV) infections. *Veterinary Microbiology*, *64*(2), 89-107. https://doi.org/10.1016/s0378-1135(98)00262-4

Hui, S. L., & Walter, S. D. (1980). Estimating the Error Rates of Diagnostic Tests [Publisher : [Wiley, International Biometric Society]]. *Biometrics*, *36*(1), 167-171. https://doi.org/10.2307/2530508

Le Strat, Y., & Carrat, F. (1999). Monitoring epidemiologic surveillance data using hidden Markov models. *Statistics in Medicine*, *18*(24), 3463-3478. https://doi.org/10.1002/(sici)1097-0258(19991230)18:24<3463::aid-sim409>3.0.co;2-i

Lindberg, A., & Houe, H. (2005). Characteristics in the epidemiology of bovine viral diarrhea virus (BVDV) of relevance to control. *Preventive Veterinary Medicine*, *72*(1), 55-73. https://doi.org/10.1016/j.prevetmed.2005.07.018

Madouasse, A., Mercat, M., van Roon, A., Graham, D., Guelbenzu, M., Santman Berends, I., van Schaik, G., Nielen, M., Frössling, J., Ågren, E., Humphry, R., Eze, J., Gunn, G., Henry, M. K., Gethmann, J., More, S. J., Toft, N., & Fourichon, C. (2022). A modelling framework for the prediction of the herd-level probability of infection from longitudinal data. *Peer Community Journal*, *2*. https://doi.org/10.24072/pcjournal.80

McAloon, C. I., McAloon, C. G., Barrett, D., Tratalos, J. A., McGrath, G., Guelbenzu, M., Graham, D. A., Kelly, A., O'Keeffe, K., & More, S. J. (2024). Estimation of sensitivity and specificity of bulk tank milk PCR and 2 antibody ELISA tests for herd-level diagnosis of *Mycoplasma bovis* infection using Bayesian latent class analysis. *Journal of Dairy Science*. https://doi.org/10.3168/jds.2023-24590

McAloon, C. G., Doherty, M. L., Whyte, P., O'Grady, L., More, S. J., Messam, L. L. M., Good, M., Mullowney, P., Strain, S., & Green, M. J. (2016). Bayesian estimation of prevalence of paratuberculosis in dairy herds enrolled in a voluntary johne's disease control programme in ireland. *Preventive Veterinary Medicine*, *128*, 95-100. https://doi.org/10.1016/j.prevetmed.2016.04.014

Metcalfe, L. (2019). An Update on the Status of BVD Control and Eradication in Europe Veterinary Science & Medicine. *Journal of Veterinary Medical Science*, *7*.

Nguyen, L. (2016). Continuous Observation Hidden Markov Model. *Revista Kasmera*, *44*, 65-149.

Olafson, P., MacCALLUM, A. D., & Fox, F. H. (1946). An apparently new transmissible disease of cattle. *The Cornell Veterinarian*, *36*, 205-213.

Olsen, A., Nielsen, H. V., Alban, L., Houe, H., Jensen, T. B., & Denwood, M. (2022). Determination of an optimal ELISA cut-off for the diagnosis of *Toxoplasma gondii* infection in pigs using Bayesian latent class modelling of data from multiple diagnostic tests. *Preventive Veterinary Medicine*, *201*, 105606. https://doi.org/10.1016/j.prevetmed.2022.105606

Rabiner, L. (1989). A tutorial on hidden Markov models and selected applications in speech recognition [Conference Name : Proceedings of the IEEE]. *Proceedings of the IEEE*, *77*(2), 257-286. https://doi.org/10.1109/5.18626

van Roon, A. M., Santman-Berends, I. M. G. A., Graham, D., More, S. J., Nielen, M., van Duijn, L., Mercat, M., Fourichon, C., Madouasse, A., Gethmann, J., Sauter-Louis, C., Frössling, J., Lindberg, A., Correia-Gomes, C., Gunn, G. J., Henry, M. K., & van Schaik, G. (2020). A description and qualitative comparison of the elements of heterogeneous bovine viral diarrhea control programs that influence confidence of freedom. *Journal of Dairy Science*, *103*(5), 4654-4671. https://doi.org/10.3168/jds.2019-16915

Wang, Y., Vallée, E., Compton, C., Heuer, C., Guo, A., Wang, Y., Zhang, Z., & Vignes, M. (2024). A novel Bayesian Latent Class Model (BLCM) evaluates multiple continuous and binary tests : A case study for *Brucella abortus* in dairy cattle. *Preventive Veterinary Medicine*, *224*, 106115. https://doi.org/10.1016/j.prevetmed.2024.106115

Watkins, R. E., Eagleson, S., Veenendaal, B., Wright, G., & Plant, A. J. (2009). Disease surveillance using a hidden markov model. *BMC Medical Informatics and Decision Making*, *9*(1), 39. https://doi.org/10.1186/1472-6947-9-39

Yang, D. A., Xiao, X., Jiang, P., Pfeiffer, D. U., & Laven, R. A. (2022). Keeping continuous diagnostic data continuous : Application of Bayesian latent class models in veterinary research. *Preventive Veterinary Medicine*, *201*, 105596. https://doi.org/10.1016/j.prevetmed.2022.105596

Annexes

Annexe A

Recherche d'un échantillon propre dans les données réelles

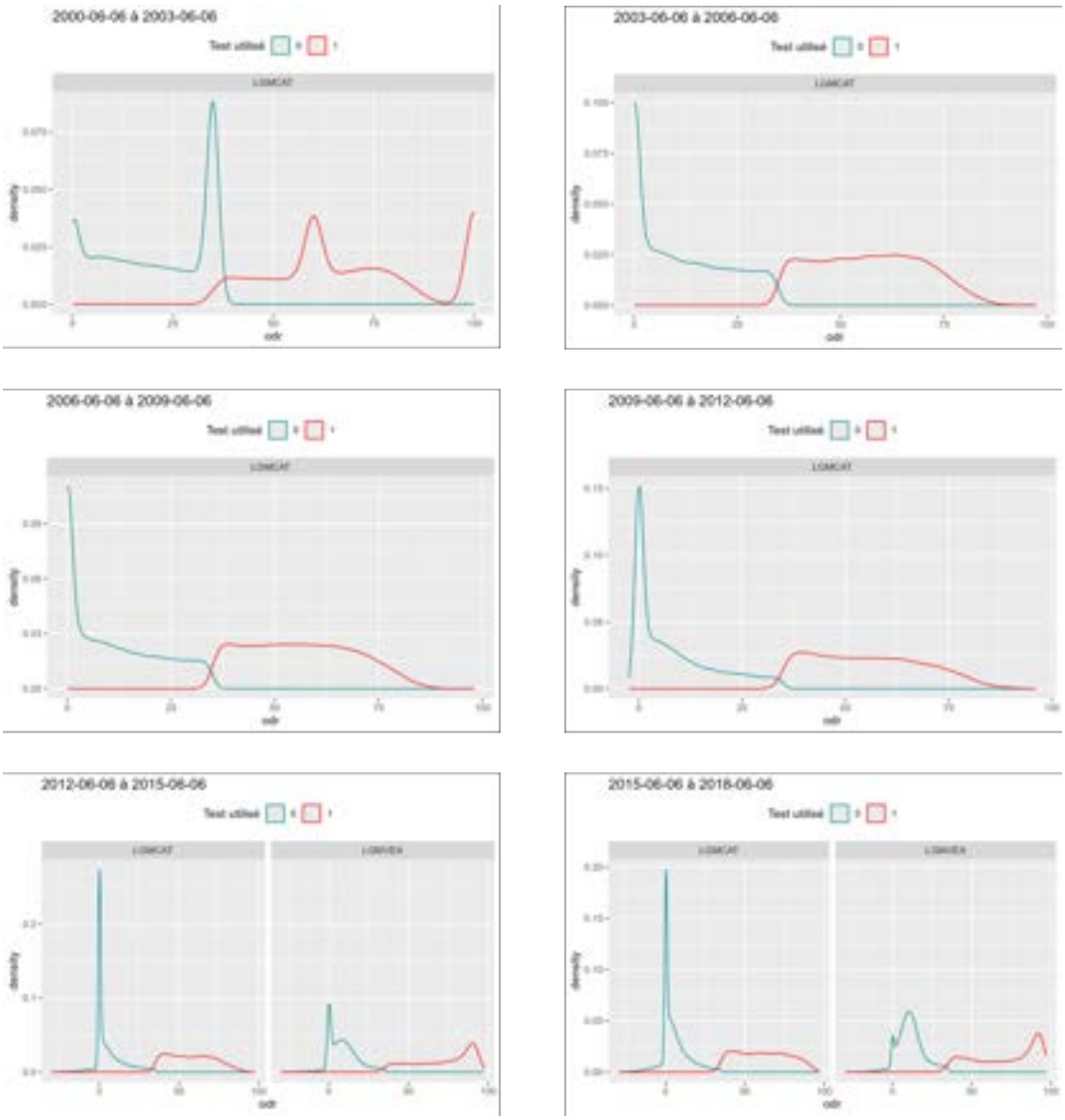


FIGURE A.1 – Distributions des données réelles sur des périodes de 3 ans, de 2000 à 2018

Annexe B

Plan d'expérience - première phase

B.1 Boxplots

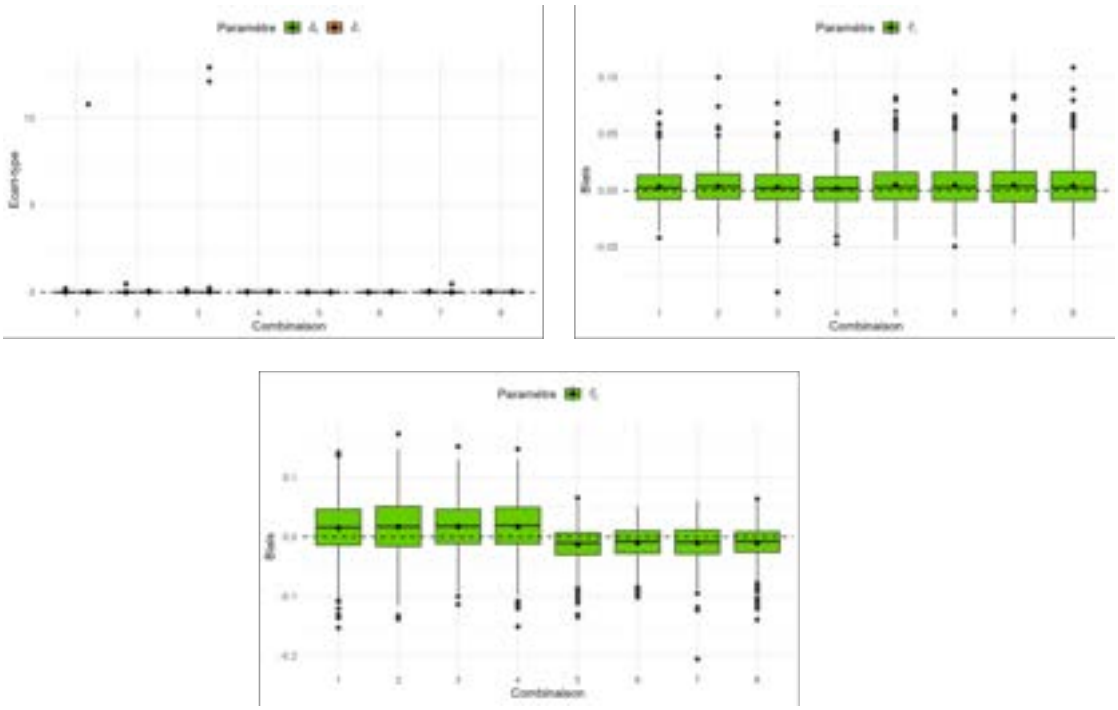


Figure B.1 – Boxplots des biais des estimations pour la première phase de plan d'expérience, pour σ_0 et σ_1 puis τ_1, τ_2.

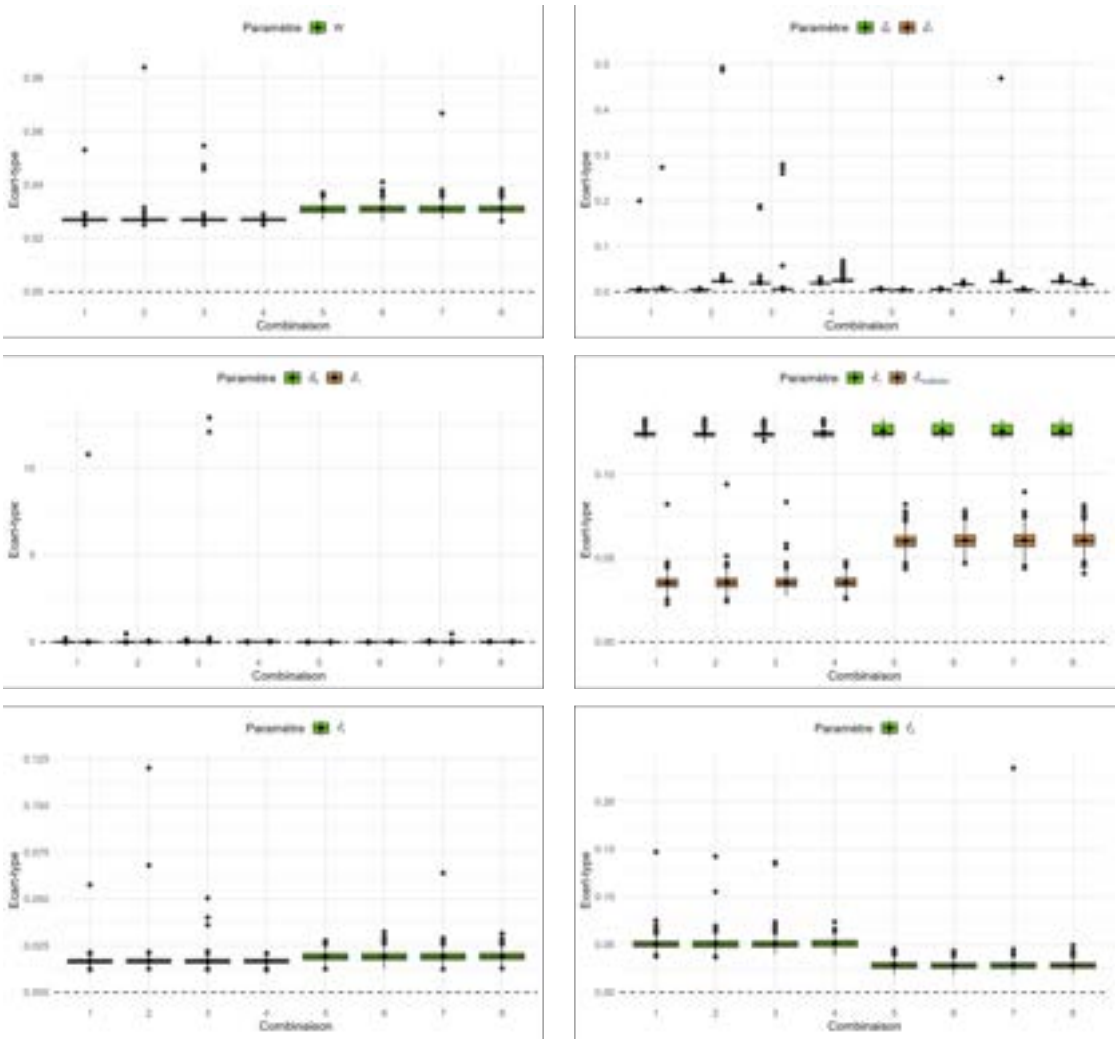


FIGURE B.2 – Boxplots des écarts-types des estimations pour la première phase du plan d'expérience.

B.2 Taux de couverture

	Scénarios							
Variable	1	2	3	4	5	6	7	8
μ_0	0.84	0.89	0.23	0.70	0.83	0.91	0.17	0.63
μ_1	0.89	0.79	0.89	0.82	0.89	0.84	0.89	0.85
$\hat{\pi}_{1_{noEstim}}$	0.92	0.91	0.92	0.94	0.87	0.89	0.88	0.87
σ_0	0.90	0.89	0.87	0.89	0.90	0.90	0.87	0.89
σ_1	0.90	0.89	0.91	0.91	0.90	0.92	0.90	0.90
τ_1	0.91	0.91	0.90	0.92	0.90	0.91	0.88	0.91
τ_2	0.88	0.88	0.91	0.90	0.89	0.92	0.90	0.90
w	0.88	0.90	0.90	0.88	0.88	0.85	0.87	0.87

Table B.1 – Pourcentages de couverture pour l'ensemble des paramètres du modèle continu // niveau 90%

	Scénarios							
Variable	1	2	3	4	5	6	7	8
μ_0	0.98	0.99	0.57	0.92	0.96	0.99	0.50	0.90
μ_1	0.99	0.96	0.99	0.97	0.99	0.98	0.99	0.98
$\hat{\pi}_{1_{noEstim}}$	0.99	0.99	0.99	1.00	0.98	0.99	0.99	0.98
σ_0	0.99	0.99	0.98	0.99	0.99	0.99	0.97	0.99
σ_1	0.99	0.99	0.99	1.00	0.98	0.99	0.99	0.99
τ_1	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
τ_2	0.99	0.98	0.99	0.99	0.99	0.99	0.99	0.99
w	0.99	0.98	0.99	0.98	0.99	0.99	0.98	0.98

Table B.2 – Pourcentages de couverture pour l'ensemble des paramètres du modèle continu // niveau 99%

Annexe C

Plan d'expérience - seconde phase

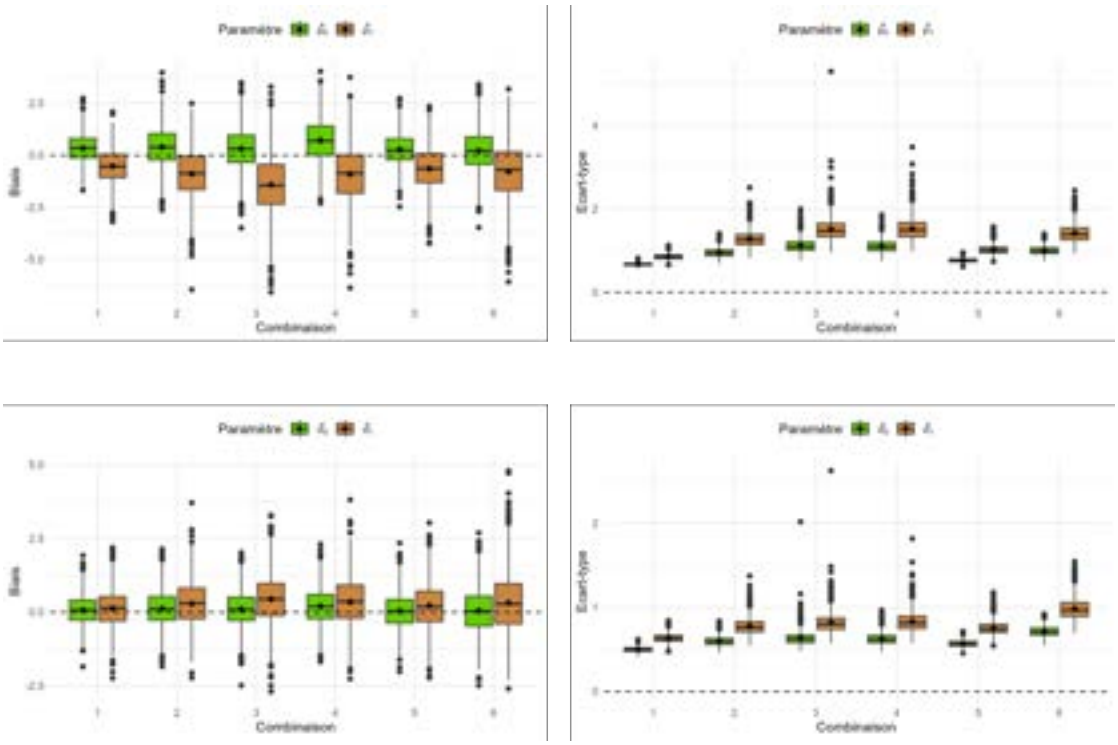


Figure C.1 – Boxplots des biais et écarts-types des estimations pour la seconde phase du plan d'expérience, pour μ_0, μ_1, σ_0, σ_1.

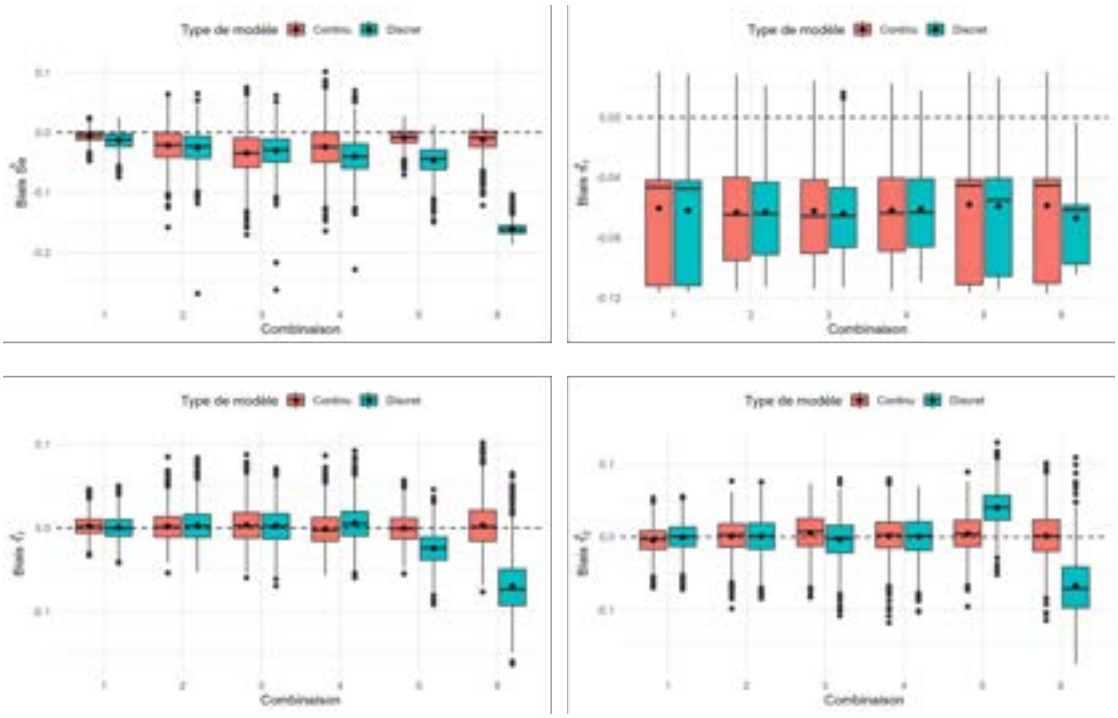


Figure C.2 – Boxplots des biais des estimations pour la comparaison des modèles // seconde phase du plan d'expérience.

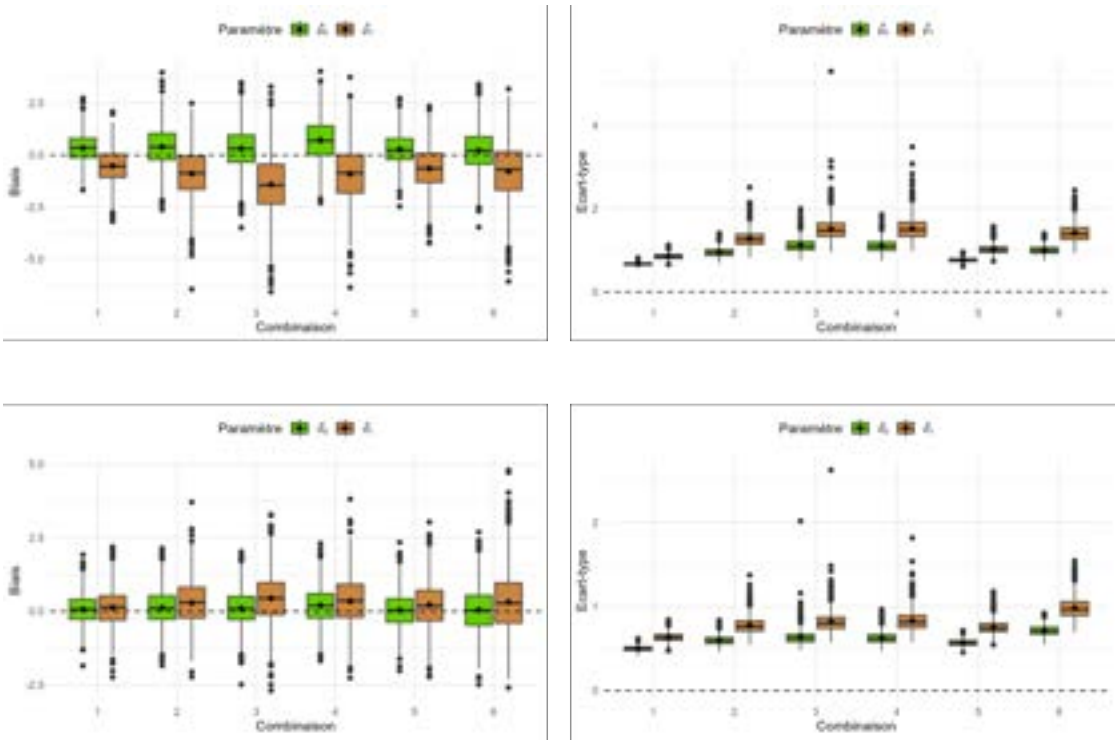


FIGURE C.3 – Boxplots des écarts-types des estimations pour la comparaison des modèles // seconde phase du plan d'expérience.

Annexe D

Application des modèles sur les 28 périodes

D.1 Table des périodes

Période	Date de début	Date de fin
1	2011-01-01	2014-04-01
2	2011-01-01	2014-07-01
3	2011-01-01	2014-10-01
4	2011-01-01	2014-12-31
5	2012-01-01	2015-04-01
6	2012-01-01	2015-07-01
7	2012-01-01	2015-10-01
8	2012-01-01	2015-12-31
9	2013-01-01	2016-04-01
10	2013-01-01	2016-07-01
11	2013-01-01	2016-10-01
12	2013-01-01	2016-12-31
13	2014-01-01	2017-04-01
14	2014-01-01	2017-07-01
15	2014-01-01	2017-10-01
16	2014-01-01	2017-12-31
17	2015-01-01	2018-04-01
18	2015-01-01	2018-07-01
19	2015-01-01	2018-10-01
20	2015-01-01	2018-12-31
21	2016-01-01	2019-04-01
22	2016-01-01	2019-07-01
23	2016-01-01	2019-10-01
24	2016-01-01	2019-12-31
25	2017-01-01	2020-04-01
26	2017-01-01	2020-07-01
27	2017-01-01	2020-10-01
28	2017-01-01	2020-12-31

Table D.1 – Table des 28 périodes utilisées pour la seconde application sur données réelles.

D.2 Table des estimations de w

	Numéro du département	Période	W moyen
LGMCAT			
1	22	1	0.50
2		2	0.49
3		3	0.49
4		4	0.48
5		5	0.46
6		6	0.46
7		7	0.46
8		8	0.46
9		9	0.47
10		10	0.47
11		11	0.45
12		12	0.45
13		13	0.36
14		14	0.36
15		15	0.36
16		16	0.37
17		17	0.37
18		18	0.38
19		19	0.36
20		20	0.37
21		21	0.34
22		22	0.34
23		23	0.34
24		24	0.34
25		25	0.38
26		26	0.37
27		27	0.32
28		28	0.35
29	29	1	0.53
30		2	0.53
31		3	0.52
32		4	0.52
33		5	0.49
34		6	0.49
35		7	0.49

36		8	0.49
37		9	0.51
38		10	0.52
39		11	0.49
40		12	0.49
41		13	0.39
42		14	0.40
43		15	0.42
44		16	0.42
45		17	0.41
46		18	0.41
47		19	0.41
48		20	0.41
49		21	0.38
50		22	0.39
51		23	0.37
52		24	0.37
53		25	0.40
54		26	0.40
55		27	0.58
56		28	0.58
57	56	1	0.48
58		2	0.47
59		3	0.45
60		4	0.44
61		5	0.40
62		6	0.40
63		7	0.39
64		8	0.38
65		9	0.37
66		10	0.36
67		11	0.35
68		12	0.34
69		13	0.22
70		14	0.22
71		15	0.23
72		16	0.23
73		17	0.22
74		18	0.22
75		19	0.22

76		20	0.22
77		21	0.20
78		22	0.20
79		23	0.19
80		24	0.19
81		25	0.26
82		26	0.21
83		27	0.20
84		28	0.25
LGMVEA			
85	22	1	0.19
86		2	0.20
87		3	0.20
88		4	0.21
89		5	0.20
90		6	0.20
91		7	0.20
92		8	0.16
93		9	0.16
94		10	0.15
95		11	0.15
96		12	0.14
97		13	0.13
98		14	0.11
99		15	0.11
100		16	0.10
101		17	0.06
102		18	0.05
103		19	0.05
104		20	0.05
105		21	0.04
106		22	0.03
107		23	0.03
108		24	0.03
109		25	0.02
110		26	0.01
111		27	0.01
112		28	0.01
113	29	1	0.16
114		2	0.22

115		3	0.21
116		4	0.23
117		5	0.22
118		6	0.22
119		7	0.22
120		8	0.18
121		9	0.19
122		10	0.18
123		11	0.18
124		12	0.16
125		13	0.16
126		14	0.14
127		15	0.14
128		16	0.13
129		17	0.07
130		18	0.06
131		19	0.06
132		20	0.06
133		21	0.04
134		22	0.04
135		23	0.04
136		24	0.04
137		25	0.02
138		26	0.02
139		27	0.02
140		28	0.02
141	56	1	0.20
142		2	0.17
143		3	0.17
144		4	0.15
145		5	0.15
146		6	0.15
147		7	0.15
148		8	0.13
149		9	0.12
150		10	0.11
151		11	0.12
152		12	0.11
153		13	0.08
154		14	0.07

155		15	0.07
156		16	0.06
157		17	0.04
158		18	0.04
159		19	0.04
160		20	0.03
161		21	0.02
162		22	0.02
163		23	0.02
164		24	0.02
165		25	0.01
166		26	0.01
167		27	0.01
168		28	0.01

Table D.2 – Tableaux des estimations du paramètre w pour les 28 périodes selon le département.

D.3 Table répartition des données

	LGMCAT			LGMVEA		
Résultats de test	**22**	**29**	**56**	**22**	**29**	**56**
0	9,075	7,860	7,296	8,494	7,270	6,722
1	950	296	407	2,530	894	1,113

Table D.3 – Répartition des mesures de tests en fonction de la discrétisation des mesures d'odr, de mai 2019 à mai 2021.

D.4 Intervalles de crédibilité pour l'estimation de paramètres supplémentaires.

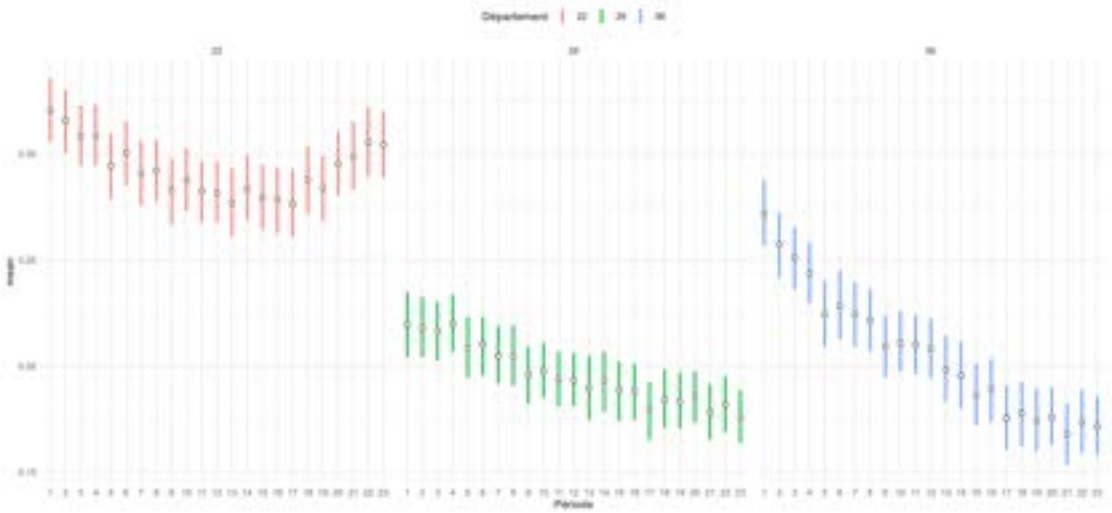


Figure D.1 – Intervalles de crédibilité pour l'estimation de π_1 sur les 23 premières périodes à partir des chaînes de τ_1 et τ_2

Figure D.2 – Intervalles de crédibilité pour l'estimation de τ_2 sur les 23 premières périodes