# MODÈLES STATISTIQUES EXPLORATOIRES ET EXPLICATIFS EN ÉPIDÉMIOLOGIE ANIMALE

Patrick Gasqui

Unité d'épidémiologie animale, INRA, 63122 Saint-Genès-Champanelle, France

Résumé :
En épidémiologie animale, l'utilisation d'un modèle statistique est nécessaire quand on cherche à prendre en compte et/ou identifier des facteurs de risque à différents niveaux de perception, chez la vache laitière par exemple. La validité des tests réalisés à l'aide de ces modèles est assurée quand les hypothèses d'indépendance entre unités statistiques sont respectées, et quand d'une manière générale, l'ajustement réalisé ne fait pas apparaître de sur-dispersion par rapport aux données observées. Dans le cadre d'étude épidémiologique des mammites cliniques, après avoir identifié les principales sources de sur-dispersion, qui peuvent varier selon l'échelle considérée, les principales méthodes de modélisation statistique exploratoire (généraliste, empirique) envisagées par les auteurs sont présentées. L'apport d'une modélisation plus explicative (synthétique, intégrative) est ensuite illustrée à l'aide d'un modèle basé sur des mélanges de distributions paramétriques dans un cadre de modèle de survie, modèle statistique original construit à partir des concepts biologiques prédédemment identifiés. L'intérêt d'une approche explicative-intégrative, en complément des approches exploratoires et généralistes plus classiques dans le cadre de la modélisation statistique de phénomènes épidémiologiques est ainsi mis en avant.

Mots clés : modélisation / modèle de survie / épidémiologie / sur-dispersion

Abstract :
In animal epidemiology, modelling risk factors, as a key multifactorial disease in the dairy cow for example, requires statistical models. The type of model used depends on the choice of perception or the study level: herd, animal and lactation. The validity of the tests that are performed through these models is especially ensured when hypotheses of independence between statistical units are respected, and when the model adjustments do not involve overdispersion faced with the observed data. In clinical mastitis epidemiological study, the main sources of overdispersion are identified according to the different levels of perception of mastitis risk. The main generalist exploratory (empirical) statistical models used by the authors, are discussed. Then, the contribution of an synthetical explanatory statistical model to improve modelling accuracy and relevance is documented. This original statistical model are based on parametric distribution mixtures in survival model approach with biological concepts previously identified. The explanatory modelling approach to complement more classical generalist exploratory models, in statistical modelling of epidemiological events, are like this showed.

Key words : modelling / survival model / epidemiology / overdispersion

## 1. INTRODUCTION

En épidémiologie animale, l'utilisation d'un modèle statistique est nécessaire quand on cherche à prendre en compte et/ou identifier des facteurs de risque que ce soit à l'échelle de l'élevage, de l'animal ou de la lactation chez la vache laitière. La validité des tests réalisés à l'aide de ces modèles exploratoires est assurée quand les hypothèses d'indépendance entre unités statistiques sont respectées, et quand d'une manière générale, l'ajustement réalisé ne fait pas apparaître de sur-dispersion par rapport aux données observées (McCullagh, 1989 et Collett, 1991). Dans le cadre d'étude épidémiologique des mammites cliniques chez la vache laitière, une pathologie de la mamelle due à la pénétration de germes infectieux via le canal du trayon et susceptible de récidive

au cours de la lactation, différentes sources de surdispersion ont été identifiées pouvant varier selon l'échelle d'étude considérée. Divers modèles exploratoires généralistes sont utilisés aux différentes échelles permettant de prendre en compte cette sur-dispersion, afin de contrôler les résultats des tests des facteurs potentiellement intéressants, puisque l'objectif principal de ces études est l'identification de facteurs de risque. Ces modèles exploratoires sont vus comme un cadre méthodologique contraignant, nécessitant une stricte sélection des données utilisables afin d'être en adéquation avec les hypothèses (postulats) des modèles d'analyse statistique mis en oeuvre.
Au vu de la quantité d'information recueillie lors des enquêtes épidémiologiques, et pour résoudre des problèmes de confusion de facteurs ainsi que la prise en compte explicite des récidives et récurrences chez un même animal dans l'étude des mammites cliniques, il est important d'envisager, à partir d'un modèle "biologique" du phénomène, la construction d'un modèle explicatif-synthétique, qui puisse prendre en compte explicitement les caractéristiques du phénomène biologique. Le modèle explicatif auquel on a abouti (Gasqui, Coulon et Pons, 2003 et Pons et Gasqui, 2003), un modèle de mélange de distributions paramétriques dans un cadre de modèle de survie, a permis non-seulement de résoudre la majorité des problèmes précédemment rencontrés, mais a fourni aussi des pistes pour approfondir le phénomène biologique.

## 2. LE CADRE BIOLOGIQUE

Chez la vache laitière, les mammites cliniques (CMAST) représentent l'archétype de la maladie multifactorielle. En effet, non seulement les inflammations-infections mammaires des femelles laitières sont multifactorielles au niveau de leurs déterminants, mais également au niveau de leur étiologie. De très nombreux facteurs souvent imbriqués entre eux sont en cause dans la survenue des mammites, et sont susceptibles d'interagir aux différentes échelles : élevage, troupeau, animal, lactation voire quartiers d'une même mamelle. A l'échelle de l'animal, la persistance des germes au sein de la mamelle, suite à la non-guérison bactériologique d'une CMAST précédente, alors qu'il y avait eu guérison clinique, est clairement responsable de récurrences au sein d'une lactation, voire d'une lactation à l'autre. Ce phénomène est d'autant plus important à prendre en compte, qu'en cas de CMAST diagnostiqué en général par l'éleveur dans 95% des cas, un traitement antibiotique est appliqué par ce même éleveur en l'absence d'analyse bactériologique qui permettrait de connaître le type de germe responsable. Face à ces difficultés, les épidémiologistes ayant pour objectif l'identification des facteurs de risque de cette pathologie, ont tenté d'adapter leur méthode d'approche, en fonction de l'échelle de l'étude, du cadre d'enquête ou du type de suivi, afin d'essayer de tenir compte de manière optimale de ces différentes contraintes dans leur modèle d'analyse statistique.

## 3. LES MODELES EXPLORATOIRES POUR IDENTIFIER LES PRINCIPAUX FACTEURS

Le niveau de prédilection du domaine de l'épidémiologie est la population, lequel est constitué d'élevages, d'animaux, de lactations au sein d'élevages, voire de quartiers au sein d'une mamelle. En ce qui concerne la modélisation statistique des CMAST chez la vache laitière un article de synthèse publié par Gasqui en 2003, met en avant les différents modèles de type exploratoire mis en oeuvre par les auteurs pour identifier des facteurs de risque. Le degré de complexité des modèles d'analyse statistique de type exploratoire auquel les auteurs ont progressivement abouti est la conséquence directe de l'identification de ces différents niveaux d'application du risque, et de l'utilisation de méthodes d'analyse statistique "récemment" mises à disposition dans les outils logiciels statistiques pour tenter d'y répondre. Avec le logiciel SAS (SAS Software System, SAS Institute Inc., Cary, NC, USA), on a pu ainsi, par exemple, passer de la procédure CATMOD ("catégorical data analysis"), à la procédure LOGISTIC ("linear regression models for binary response as well as ordinal response data"), puis à la procédure GENMOD ("which fits generalized linear models and allows the response probability to be any number of an exponential family") et à la procédure MIXED ("including random effects models and a variety of mixed linear models to fit data") pour augmenter la capacité d'analyse en terme de prise en compte de sur-dispersion des données observées par rapport au modèle binomial ou poissonien théorique, et la souplesse de

l'analyse des données en limitant les pertes d'information induites par la sélection a priori des individus statistiques assurant leur indépendance mutuelle. Des contraintes sont ainsi imposées par les auteurs dans la sélection des individus introduits dans l'analyse statistique, entraînant une certaine perte d'information, au détriment de la précision de l'analyse, et finalement de sa capacité à mettre en évidence des facteurs de risque pertinents.
La solution la plus fréquemment utilisée a consisté à supprimer au sein d'une lactation d'un animal, les cas survenant après un intervalle de temps, fixé a priori, suivant le cas initial de CMAST. Comme aucune justification biologique ne permet de sélectionner une certaine durée de censure plutôt qu'une autre, on constate une variabilité importante dans les durées choisies qui vont de 0 à 90 jours selon les auteurs. Certains auteurs ont mis en évidence la variabilité de la réponse ainsi obtenue, en faisant varier cette durée de censure de 0 à 30 jours à partir des mêmes données initiales. La comparaison des résultats des différentes études est d'autant plus difficile que certains auteurs ne précisent même pas la valeur utilisée. L'ensemble des auteurs qui modélisent ainsi le nombre de CMAST, utilisent un modèle d'analyse basé sur une distribution de Poisson: régression de Poisson ou bien modèle de Poisson dans un cadre GLM ("Generalized Linear Models", McCullagh, 1989).
Ayant identifié le problème du choix de cette durée minimale, certains auteurs ont censuré les données de façon encore plus importante, en se contentant de la première occurrence de CMAST dans la lactation. Une partie de ces auteurs intègrent dans leur modèle d'analyse, une réponse de type binaire: lactation indemne contre lactation atteinte par au moins un cas de CMAST. Selon la même approche, d'autres limitent leur étude à l'expression de cette réponse dichotomique sur une sous-période fixée de la lactation : de 7 jours avant vêlage à 7 jours après, les 5, 14, 30, 60 ou les 100 premiers jours après vêlage voire une année calendaire à la place de la lactation. L'intérêt de ces travaux est que les résultats sont plus facilement comparables entre lactations, par rapport aux travaux dont la réponse est observée sur l'ensemble d'une lactation, sans tenir compte de la durée réelle des lactations. Les modèles d'analyse sont essentiellement basés sur une distribution binomiale du phénomène: régression logistique, ou modèle binomial dans un cadre GLM. D'autres auteurs s'orientent vers l'étude de la date d'occurrence du premier événement dans la lactation, en prenant la date de vêlage comme instant initial, au travers de modèles de survie, un modèle de Cox en général.
Pour tenir compte d'une éventuelle dépendance entre lactations successives, certains auteurs ont été amenés, en plus des considérations précédentes, à ne considérer dans leur étude qu'une seule lactation par animal : lactation tirée au hasard mais plus généralement en ne conservant que la première lactation de l'animal. Certains auteurs ont aussi étudier l'influence des résultats de la première lactation sur les résultats de la deuxième.
Les études abordant l'échelle de l'élevage sont par contre confrontées à une source principale de sur-dispersion qui a pour origine, excepté le cas ou les mêmes élevages sont suivis au cours du temps, la non prise en compte dans le modèle, de facteurs d'hétérogénéité importants entre élevages, facteurs le plus souvent inconnus et non facilement identifiables. Les modèles utilisés tentent alors de tenir compte de cette éventuelle sur-dispersion, en l'estimant à partir d'une distribution théorique de Poisson, ou en utilisant une distribution théorique incluant une sur-dispersion à l'aide d'une binomiale négative ou d'une bêta-binomiale. D'autres modèles permettent d'estimer une variance intra-élevages, dans le cas où les mêmes élevages sont suivis dans le temps, via une extension des modèles GLM : les modèles GEE ("Generalized Estimating Equations", Diggle, 1994). Mais de façon plus générale, un effet élevage fixe ou aléatoire est intégré dans le modèle, en plus des effets fixes correspondant aux autres variables de contrôle ou explicatives. Dans le cadre d'enquête concernant un grand nombre d'élevage, le choix s'oriente vers l'introduction d'un effet aléatoire élevage, afin d'éviter de consacrer un trop grand nombre de degré de liberté à ce facteur, au détriment de l'estimation de la variance ( ou de la déviance) résiduelle dont la bonne estimation conditionne la précision des tests des facteurs d'intérêt. La comparaison des résultats issus de différentes études est d'autant plus difficile, que les variables de contrôles nécessairement conservés dans le modèle au cours du processus itératif de sélection de variable, ne

sont pas systématiquement les mêmes. A vrai dire, le problème est d'autant plus difficile à prendre en considération de manière optimale que le type de réponse analysée au niveau d'un élevage, est construit à partir des réponses plus ou moins censurées aux divers niveaux sous-jacents, essentiellement animal, lactation et cas successifs. L'étude du niveau élevage dépend aussi en grande partie des choix qui sont faits dans le but de définir l'indice qualifiant la variable réponse (incidence, taux d'incidence) et caractérisant un élevage à un moment donné ou au cours d'une période donnée. A ce jour, au moins 12 définitions différentes du taux d'incidence ont été proposées. Dans la littérature on trouve ainsi des indices calculés en tant que pourcentages de lactations atteintes par rapport à un nombre de lactations totales ou à risque, des nombres d'occurrences (plus ou moins censurées) observés sur différentes périodes et ramenés à un nombre de jours, de semaines ou de mois à risque. Par ailleurs, le calcul effectué à partir du nombre total de lactations comportant au moins une CMAST est ensuite rapporté, selon les auteurs, au nombre total de vaches-lactation à risque, ou au total de mois de lactations indemnes. Quand le calcul est fait à partir du nombre total de cas, il est rapporté au nombre total de vaches-jour à risque, à la période dans la lactation, voire à un nombre de vaches-semaine ou de vaches-mois. Quand le nombre de jours à risque est considéré, l'indice dépendra évidemment de la valeur choisie a priori qui permet de décider de l'existence d'un nouveau cas, valeur on l'a vu qui peut varier de 0 à 90 jours.
Ayant constaté les différences non négligeables existant, au niveau des travaux publiés afin d'évaluer le risque de CMAST au sein d'un élevage, il apparaît très difficile de pouvoir comparer avec pertinence ces études, et encore moins de réaliser des méta-analyses afin de les regrouper. Ayant identifié, en analysant les diverses études disponibles, des pertes d'information relativement importantes et variées, il apparaît peu probable qu'il soit possible d'envisager par ces méthodes généralistes (exploratoires), une identification de facteurs de risque encore plus fine que celle à laquelle on a aujourd'hui abouti. On peut même aller jusqu'à émettre des réserves sur la réalité de certains facteurs de risque mis en évidence par ces approches, sachant qu'elles offrent peu de certitudes quand à l'élimination réelle des divers niveaux de dépendance. Une alternative à l'utilisation de ces méthodes exploratoires, au vu des connaissances qu'elles ont quand même participé à élaborer, est le développement de modèles plus explicatifs, basés sur une approche plus intégrative et plus causale du phénomène.

## 4. UN MODELE EXPLICATIF POUR FAIRE LA SYNTHESE DES CONNAISSANCES BIOLOGIQUES

A partir des connaissances biologiques élaborées à partir des diverses études réalisées sur les CMAST, il a été possible de construire un modèle biologique prenant explicitement en compte les notions de récidive et de réccurrence au niveau d'une lactation, et d'une lactation à l'autre pour un même animal, en tenant compte de l'éventuelle inefficacité des traitements antibiotiques appliqués systématiquement en cours de lactation en cas de CMAST diagnostiqués, ou au moment du tarissement de la vache laitière. Ce modèle biologique a permis le développement d'un modèle statistique synthétique-explicatif basé sur des éléments facilement observables, comme la standardisation des symptômes cliniques ou les dates d'occurrence des divers évenements clés : CMAST, vêlage, tarissement, ... Ce modèle statistique est basé sur des mélanges de distributions paramétriques dans un cadre de modèles de survie. Les propriétés théoriques des estimateurs du modèle ont été étudiées et publiées par Pons et Gasqui en 2003. L'élaboration et la validation de ce modèle appliqué à l'étude des CMAST est détaillée dans l'article de Gasqui, Coulon et Pons en 2003. Cette approche, effectuée à l'échelle de la mamelle, intègre les niveaux lactation et animal, et permet la réalisation de tests de facteurs de risque individuels ou d'élevage. Cette approche permet d'étudier à la fois la distribution du nombre d'événements par lactation, et la distribution des instants d'occurrence de ces événements. L'objectif est d'intégrer prochainement la variabilité due aux divers facteurs de risque mise en évidence au niveau élevage.
Avec ce modèle, il est par exemple possible de prendre en compte la situation de l'animal à l'étable et au pâturage, ou de façon plus générale toute variable individuelle dépendante du temps. L'association de cette variable à la prise en compte d'un taux d'infection au vêlage, permet par

exemple de ne plus avoir à prendre en compte dans la modélisation, le facteur stade de lactation, mis précédemment en évidence comme étant une des variables structurelles gouvernant le risque mammite. D'une façon plus générale l'approche a permis de résoudre les principaux problèmes liés à la confusion de facteurs de risque à laquelle on est classiquement confronté quand on travaille à partir de données issues d'enquêtes, pour lesquelles il n'est pas possible de contrôler a priori les facteurs de variation à l'aide d'un plan d'expérience adapté. De même la dépendance entre CMAST successives au sein d'une lactation, a pu être explicitement prise en compte dans le modèle au travers d'un taux de guérison bactériologique.
La connaissance précise de la réponse mammaire aux différents types de germes, pourra permettre dans l'avenir de porter l'analyse au niveau du quartier de la mamelle, ce qui permettrait sans doute d'affiner ce type d'approche, en prenant en compte le type de germe présent par exemple. D'un point de vue pratique, ces modèles explicatifs peuvent aussi aider à définir des critères optimum vis à vis de la définition d'un taux d'incidence caractéristique au niveau de l'élevage, voire à fournir une justification au délai à considérer quand on veut éliminer les ré-occurrences successives au cours d'une même lactation. Le modèle développé par l'intermédiaire de fonctions de survie permet aussi de justifier le fait de ne s'intéresser qu'à la première occurrence d'une lactation, ou de baser l'analyse sur une réponse dichotomique au niveau d'une lactation puisqu'en l'absence de récurrence, on ne devrait pas avoir plus d'une mammite par lactation dans le contexte des élevages étudiés. Les étapes de validation du modèle explicatif généré a permis de constater par exemple qu'il n'était pas besoin d'intégrer dans le modèle explicatif une notion d'indépendance entre animaux d'un même troupeau, pour obtenir de très bons ajustements. Cela pourrait signifier d'un point de vue biologique, que l'état de CMAST n'est pas un état transmissible d'un animal à l'autre, ce qui n'est pas en général le cas lors d'une infection sub-clinique : des caractéristiques propres à l'animal feraient qu'à un moment donné, une infection sub-clinique évoluerait en stade clinique, indépendamment de ce qui se passe pour les autres animaux du troupeau.
Par contre, le développement d'une modélisation explicative spécifique demande un investissement scientifique plus important que celui nécessaire à la mise en oeuvre d'un modèle exploratoire généraliste. L'apport du nombre très important d'études épidémiologiques et expérimentales réalisées en matière de CMAST a permis la mobilisation des connaissance biologiques préalables à ce type de développement.

5. CONCLUSION

L'approche de modélisation explicative ou synthétique présentée, a l'avantage d'être une voie de synthèse ayant pour but de modéliser les connaissances élaborées à partir des résultats des travaux basés sur des modèles exploratoires généralistes. Ces deux optiques de modélisation ( exploratoire versus explicative), dans le cadre de la modélisation épidémiologique de phénomènes pathologiques, non seulement ne sont donc pas à opposer l'une à l'autre, mais se complètent mutuellement. Les articles de Cox en 1990 et de Lehmann en 1990 présentent ces différentes approches de modélisation statistique avec une perspective historique, tout en cherchant a identifier la méthodologie qui aboutie à la construction d'un modèle statistique. Cette approche synthétique permet non seulement de construire un modèle permettant de tester plus finement d'éventuels facteurs d'intérêt, d'aboutir à un modèle prédictif vu comme une première étape vers un éventuel système d'aide à la décision, mais aussi de générer de nouvelles hypothèses biologiques sur le phénomène pathologique lui-même. Le développement d'une telle approche de modélisation rend donc plus concrète et surtout montre l'intérêt d'une réelle collaboration entre des biologistes et des modélisateurs statisticiens.

## Bibliographie

[1] Collett D. (1991) Modelling Binary Data, Chapman & Hall, Londres.
[2] Cox D.R. (1990) Role of Models in Statistical Analysis, Statistical Science, 5, 169-174.
[3] Diggle P.J., Liang K., Zeger S.L. (1994) Analysis of longitudinal Data, Clarendon Press,

Oxford.
[4] Gasqui, P. et Barnouin, J. (2003) Statistical modelling for clinical mastitis in the dairy cow: problems and solutions, “Review article”, Veterinary Research, 34, 493-505.
[5] Gasqui, P., Coulon, J.B. et Pons, O. (2003) An individual modelling tool for within and between lactation consecutive cases of clinical mastitis in the dairy cow: an approach based on a survival model, Veterinary Research, 34, 85-104.
[6] Lehmann E.L. (1990) Model Specifications: The Views of Fisher and Neyman, and Later Developments, Statistical Science, 5, 160-168.
[7] McCullagh P., Nelder J.A., (1989) Generalized Linear Models, 2nd éd., Chapman & Hall, Londres.
[8] Pons, O. et Gasqui, P. (2003) A mixture point process for repeated failure times, with an application to a recurrent disease, Biometrical Journal, 45, 7, 798-811.