# L'UTILISATION DES METHODES UNILISTES DE CAPTURE-RECAPTURE EN SURVEILLANCE DE MALADIES ANIMALES : APPLICATION AUX DONNEES FRANCAISES DE TREMBLANTE CLASSIQUE*

**Timothée Vergne**[1,2], **Vladimir Grosbois**[2], **Géraldine Cazeau**[3], **Didier Calavas**[3], **Benoit Durand**[4] et **Barbara Dufour**[1]

**RESUME**

Initialement développées en écologie afin d'estimer la taille de populations sauvages, les méthodes de capture-recapture ont été appliquées en épidémiologie humaine afin d'estimer le nombre d'individus affectés par une maladie donnée et non détectés par les systèmes de surveillance. Ce n'est qu'en 2005 que les premières applications en épidémiologie vétérinaire sont apparues. Parmi ces méthodes de capture-recapture, les modèles unilistes, aussi appelés modèles tronqués, sont ceux d'application la plus récente. Cet article a pour objectif de présenter ces modèles unilistes, et de les appliquer aux données françaises de surveillance de la tremblante ovine classique en 2006 et 2007 afin d'estimer, pour chaque année, le nombre d'élevages ovins qui, bien qu'infectés par l'agent de la tremblante, ne sont pas détectés par le dispositif de surveillance. L'accent est mis principalement sur l'importance du choix des données destinées aux analyses. Dans l'état actuel des connaissances et du fait des caractères rare et peu contagieux de la maladie, il s'avère que ces méthodes ne semblent pas encore entièrement adaptées à la tremblante.

**Mots-clés** : capture-recapture, uniliste, surveillance, tremblante.

**SUMMARY**

Initially, capture-recapture methods were developed in ecology to evaluate the size of wild populations. Then they were applied to human epidemiology in order to determine the number of humans affected by a given disease and who were not detected by the surveillance system. The first applications in the veterinary field appeared in 2005. Among these methods, the unilist approach is the most recent one. This article presents these unilist methods and applies them to the French conventional scrapie surveillance data for 2006 and 2007 in order to evaluate the number of affected holdings that were not detected by the French surveillance system.

…/..

* Texte de la communication orale présentée au cours des Journées scientifiques AEEMA, 21 mai 2010

1 EpiMAI/ ENVA-Anses Maisons-Alfort, 23 avenue du Général de Gaulle, Maisons Alfort cedex, F94701, France

2 Centre de coopération internationale en recherche agronomique pour le développement (CIRAD), Département ES, UR22, TA C22/E, Campus international de Baillarguet, 34398 Montpellier cedex 5, France

3 Anses-Lyon, Unité Epidémiologie, 31 avenue Tony Garnier, 69364 Lyon Cedex 07, France

4 EPI-Anses LERPAZ, Site de Maisons-Alfort, 23 avenue du Général de Gaulle, Maisons-Alfort cedex, F94706, France

…/..

The importance of properly choosing the data to be analyzed is highlighted. In the current state of knowledge, and because of the scarcity and the low contagiousness of the disease, it appears that these methods are not yet fully suitable for diseases such as scrapie.

**Keywords**: Capture-recapture, Unilist, Surveillance, Scrapie.

## I - INTRODUCTION

En France, la tremblante classique, maladie neuro-dégénérative progressive contagieuse des petits ruminants, fait l'objet d'une surveillance clinique depuis 1990. Il s'agissait à l'origine d'un réseau de surveillance passif limité au sud de la France, fondé sur la détection de symptômes nerveux chez les petits ruminants âgés de plus d'un an. A l'instar de la surveillance clinique de l'ESB, ce réseau s'est élargi à l'ensemble du territoire français lorsque la tremblante est devenue une maladie à déclaration obligatoire en 1996 [Calavas *et al.*, 1999]. Au vu d'études européennes mettant en évidence la faible exhaustivité de la détection des cas de tremblante [Baumgarten *et al.*, 2002 ; Hoinville *et al.*, 2000], un programme européen a été mis en place, imposant une surveillance active de la tremblante dans les abattoirs et à l'équarrissage. C'est en 2002 que ce programme a été appliqué en France. Depuis cette date, le protocole général de la surveillance de la tremblante en France peut être décomposé en deux étapes : détection des « cas initiaux », et détection des « cas additionnels ». La détection des cas initiaux se fait selon un des trois protocoles de surveillance (surveillance clinique et réalisation de tests en abattoir et à l'équarrissage). Une fois les cas initiaux découverts, l'éleveur concerné choisit l'abattage total de son troupeau avec testage de tous les animaux, ou le génotypage de l'ensemble du troupeau avec abattage et testage des seuls animaux génétiquement sensibles à la tremblante classique. Cette deuxième phase permet la découverte des cas additionnels. Il y a donc deux phases de détection permettant respectivement l'identification des cas initiaux et des cas additionnels. Cette surveillance en deux étapes concerne uniquement les cas de tremblante classique qui seule nous intéressera ici. Parce que les tremblantes classique et atypique diffèrent tant sur les plans du schéma épidémiologique que de la police sanitaire, nous avons choisi de ne pas considérer la tremblante atypique afin de garder une cohésion dans la signification épidémiologique des résultats. Par ailleurs, pour des raisons de très faible prévalence et de police sanitaire différente de chez les ovins, les cas de tremblante classique chez les caprins n'ont pas non plus été pris en compte. Dans la suite de l'article, le terme « tremblante » se réfère à la tremblante classique ovine. Les résultats obtenus sont donc relatifs à la population des élevages ovins atteints de tremblante classique.

Il est reconnu que la plupart des dispositifs de surveillance, tant dans le domaine vétérinaire que dans celui de la santé publique, souffrent du phénomène de sous-identification/sous-notification des cas [Hook et Regal, 1995]. Dans un objectif de surveillance et de suivi de la prévalence, comme c'est le cas pour la tremblante du mouton en France et plus généralement en Europe, il est important d'apprécier le plus précisément possible l'étendue de cette sous-estimation de manière à estimer au plus près la prévalence réelle de la maladie considérée.

Cette estimation du nombre de cas non détectés, devenue assez commune en médecine humaine tend à se développer dans le domaine de la surveillance épidémiologique vétérinaire, au travers des méthodes de « capture-recapture ». Supposons qu'un système de surveillance identifie $N_{obs}$ cas d'une maladie donnée. Il convient alors de

savoir combien de cas *N* il y a réellement dans la population (ou combien de cas n'ont pas été détectés par le système de surveillance). Si on considère que les cas sont détectés avec une probabilité $1-p_0$ (avec $p_0$ la probabilité de ne pas détecter un cas), on a donc $N = Np_0 + N(1-p_0)$ soit $N = Np_0 + N_{obs}$. Cette équation peut simplement être résolue pour obtenir l'estimateur d'Horvitz-Thompson $N = N_{obs} / (1-p_0)$. Si $p_0$ est connu, alors l'équation est directement résolue et un estimateur de *N* peut être calculé. Malheureusement, la plupart du temps $p_0$ est inconnu et doit être estimé. Pour estimer $p_0$, il est généralement admis que les données collectées par le système de surveillance sont structurées de telle manière qu'elles peuvent être modélisées. Pour estimer des populations de taille inconnue, deux méthodes sont classiquement utilisées : la méthode de capture-recapture multiliste (au moins deux listes) [Hook et Regal, 1995 ; IWGDMF, 1995], et la méthode de capture-recapture uniliste.

La méthode de capture-recapture uniliste s'intéresse à la fréquence de détection de l'unité épidémiologique étudiée. Un des exemples historiques a consisté à estimer le nombre de taxis dans la ville d'Edimbourg. Carothers [1973] a « capturé » et « recapturé » chaque taxi qu'il observait durant 5 jours selon un protocole d'échantillonnage déterminant aléatoirement les moments et les lieux d'observation, afin de s'assurer d'une bonne représentativité de son échantillon. Il a donc eu accès à la fréquence d'observation de chaque taxi « visible » de la capitale écossaise. On appelle ce type de données, des données de *comptage répété*. On a ainsi accès à la fréquence $f_1$ des unités observées une fois uniquement, à la fréquence $f_2$ des unités observées deux fois uniquement, *etc*. La fréquence $f_0$ des unités d'intérêt jamais détectées constitue l'information manquante et a besoin d'être estimée. En termes statistiques, on parle de *distribution tronquée en 0*. Les principaux développements de ces méthodes ont été apportés par Zelterman [1988] et Chao [1988].

Dans le cas de l'application aux données de tremblante en France, l'unité épidémiologique d'intérêt est un élevage ayant au moins un cas de tremblante détectable. On va donc chercher à estimer le nombre d'élevages infectés n'ayant pas été détectés par le système de surveillance. Les données de comptage répété sont issues de l'observation du nombre d'animaux infectés détectés au sein des élevages infectés : les « occasions de capture » sont représentées par les animaux infectés au sein de l'élevage (chaque animal infecté dans un élevage est considéré comme un révélateur potentiel de l'infection), et les « captures » sont donc représentées par la détection de ces animaux infectés. En comptabilisant le nombre d'animaux infectés détectés dans chaque élevage où au moins un animal infecté a été détecté, on va donc pouvoir construire une distribution tronquée en 0, *f*, telle que

$f_x$ = *nombre d'élevages infectés dans lesquels x cas de tremblante ont été identifiés*

à laquelle on va pouvoir appliquer la méthode de capture-recapture uniliste pour estimer $f_0$, le nombre d'élevages infectés dans lesquels aucun cas de tremblante n'a été identifié. Ces méthodes, déjà appliquées à la tremblante en Grande-Bretagne, ont utilisé comme données l'ensemble des cas initiaux et des cas additionnels qui ont été regroupés pour obtenir un nombre total d'animaux infectés détectés par élevage détecté. Nous pensons que l'utilisation des seuls cas initiaux permettrait une analyse par les méthodes unilistes plus adaptée aux conditions de surveillance et aux caractéristiques de la police sanitaire. La démarche suivie a donc consisté à calculer le nombre total d'élevages infectés selon les deux approches, puis à utiliser des données simulées pour tenter d'expliquer les phénomènes statistiques observés.

## II - MATERIEL ET METHODE

### 1. LES DONNEES DE « COMPTAGE REPETE » DE LA TREMBLANTE

En France, l'ensemble des cas initiaux, détectés à l'abattoir, à l'équarrissage et lors de la surveillance clinique, est centralisé par l'Agence française de sécurité sanitaire des aliments (Afssa). La surveillance à l'abattoir et à l'équarrissage est réalisée respectivement sur un échantillonnage des animaux abattus et morts. L'obex de ces animaux est prélevé et la présence de la protéine pathogène est recherchée par un test agréé (par exemple le

test Prionics WB® ou le test Biorad®). La surveillance clinique concerne tout animal suspect présentant des signes nerveux. L'animal est abattu, l'obex est prélevé et analysé. En 2006 et en 2007, près de 80% des cas initiaux ont été détectés à l'équarrissage, la détection en abattoir et celle sur les signes cliniques se répartissant les 20% restants (figure 1). A ces cas initiaux s'ajoutent les cas additionnels, découverts lors de la police sanitaire, résultat du protocole de détection mis en place dans les élevages où au moins un cas initial a pu être préalablement identifié par un des trois protocoles décrits plus haut. Pour chaque année, nous avons donc deux jeux de données différents, l'un composé des cas initiaux uniquement, et l'autre de l'ensemble des cas (cas initiaux et cas additionnels).

La répartition du nombre de cas initiaux détectés dans chaque élevage infecté en 2006 et en 2007 a été consignée dans le tableau 1. La répartition du nombre de cas totaux (initiaux et additionnels) détectés dans chaque élevage infecté en 2006 et 2007 a été consignée dans le tableau 2. En 2006, un total de 141 élevages avec au moins un cas de tremblante a pu être identifié. En 2007, ce nombre a diminué pour atteindre 67.

**Figure 1**

**Répartition des détections des cas initiaux de tremblante classique entre l'équarrissage, l'abattoir et la surveillance clinique pour 2006 et 2007**

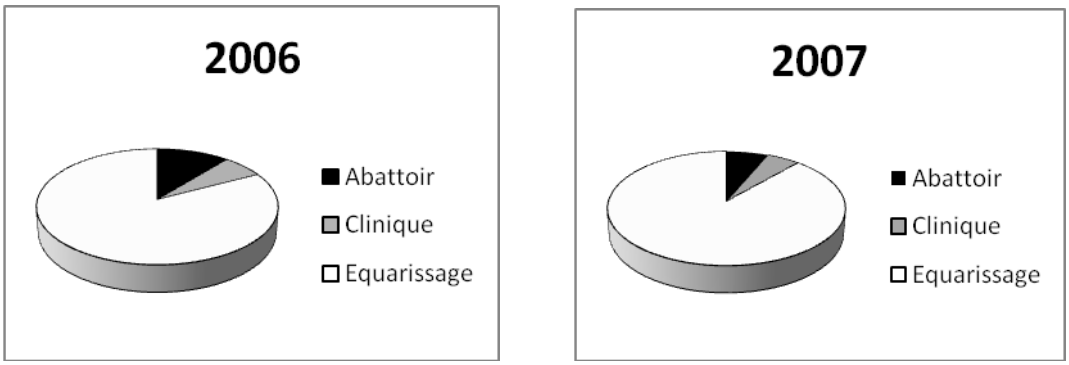



**Tableau 1**

**Distribution du nombre de cas initiaux de tremblante classique ovine détectés dans les élevages infectés identifiés par le système de surveillance en 2006 et en 2007**

| Nombre de cas détectés dans les élevages | 0 | 1 | 2 | 3 | 4 | Total |
|---|---|---|---|---|---|---|
| Nb élevages détectés en 2006 | - | 121 | 13 | 5 | 2 | 141 |
| Nb élevages détectés en 2007 | - | 59 | 5 | 2 | 1 | 67 |

**Tableau 2**

**Distribution du nombre de cas totaux (cas initiaux + cas additionnels) de tremblante classique ovine dans les élevages infectés identifiés par le système de surveillance en 2006 et 2007**

| Nombre de cas détectés dans les élevages | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10+ | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Nb élevages détectés en 2006 | - | 72 | 17 | 14 | 5 | 8 | 1 | 8 | 3 | 1 | 12 | 141 |
| Nb élevages détectés en 2007 | - | 40 | 10 | 6 | 3 | 2 | 0 | 0 | 1 | 0 | 5 | 67 |

## 2. LES CONDITIONS D'APPLICATION DES METHODES UNILISTES

L'utilisation des méthodes unilistes de capture-recapture nécessite trois conditions principales. La première est que la population étudiée soit fermée pendant la période d'observation. Tirée de l'écologie, cette hypothèse signifiait que chaque individu avait une probabilité non nulle d'être capturé. Appliquée au contexte de la tremblante, en considérant que chaque animal infecté correspond à une « occasion de capture », cette hypothèse signifie alors que chaque élevage a le même nombre d'animaux infectés. La deuxième condition est que les occasions de capture soient indépendantes : le fait qu'une unité soit détectée lors d'une certaine occasion de capture ne doit pas changer la probabilité d'être détectée lors des autres occasions. Enfin, la troisième condition implique que toutes les unités épidémiologiques soient détectées de manière équiprobable. Dans le contexte de la surveillance de la tremblante, ceci implique que tous les élevages infectés ont la même probabilité d'être détectés. Certains estimateurs permettent de s'affranchir de la dernière condition d'utilisation, qui est rarement remplie. La validité de ces hypothèses dans le cas de la surveillance de la tremblante en France est discutée plus loin.

## 3. ESTIMATIONS ET HETEROGENEITE

L'approche conventionnelle considère que la probabilité d'être « capturé » $x$ fois est déterminée par une loi de Poisson de paramètre $\lambda$ :

$$p_x(\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$$

avec $\lambda$ identique pour toutes les unités et pour toutes les occasions de capture. Une estimation du paramètre $\lambda$ peut alors être obtenue en maximisant la vraisemblance du modèle par rapport aux données observées, ce qui permet alors de calculer la probabilité de ne pas être capturé ($p_0$) et ainsi d'estimer la taille totale de la population étudiée par l'intermédiaire de l'estimateur d'Horvitz-Thompson. Cette approche n'est valable que dans le cas où toutes les conditions d'application sont respectées.

La plupart du temps en épidémiologie, l'hétérogénéité de la population est telle que les probabilités de détection des unités épidémiologiques ne sont pas égales pour toutes les unités. Par exemple, on détectera plus facilement les cas sévères d'une maladie que les cas plus bénins, ou une certaine classe de la population plutôt qu'une autre. De l'hétérogénéité apparaît donc dans la population d'étude si les probabilités de détection des unités épidémiologiques ne sont pas identiques. Ces probabilités de capture peuvent varier du fait de facteurs mesurables (ex. : taille du troupeau) ou de facteurs non mesurables ou non disponibles (ex. : sensibilité génétique générale du troupeau). Dans le cas d'hétérogénéité due uniquement au rôle de facteurs mesurables, il est possible de stratifier la population selon ces facteurs, en strates à l'intérieur desquelles on pourra considérer que les probabilités de détection sont homogènes. Avec suffisamment de données, il sera alors possible de revenir, pour chaque strate, à une situation d'homogénéité [Bohning et Del Rio Vilas, 2008]. Il est cependant important de noter que cette stratégie occasionnera une forte augmentation de la variance de l'estimateur ; c'est pourquoi, la stratification n'est envisageable que dans le cas où de nombreuses données sont disponibles.

L'hétérogénéité des probabilités de capture, occasionnant des observations ne suivant pas exactement une loi de Poisson, peut être diagnostiquée en calculant pour chaque fréquence d'observation $i$ le ratio $r_i$ défini par

$$r_i = \frac{(i+1)\, p_{i+1}(\lambda)}{p_i(\lambda)}$$

qui se calcule en se mettant sous la forme $r_i = (i+1)f_{i+1}/f_i$. Dans le cas d'homogénéité, $r_i$ est constant et vaut $\lambda$. En revanche, en présence d'hétérogénéité, il a été montré que quel que soit $i$, $r_{i+1} \geq r_i$, $r_i$ croit donc lorsque $i$ augmente [Chao, 1987]. Un graphique montrant les variations de $r_i$ en fonction de $i$ est un outil rapide et facile à mettre en place, permettant la mise en évidence de la présence d'hétérogénéité.

En présence d'hétérogénéité des probabilités de capture, il existe des solutions alternatives permettant d'utiliser des estimateurs relativement peu biaisés. Parmi ces estimateurs, les plus utilisés sont l'estimateur de Zelterman et l'estimateur de Chao. Ces estimateurs sont facilement calculables et on les considère comme relativement robustes à l'hétérogénéité car ils ne font intervenir que les fréquences observées des cas notifiés une fois et deux fois ($f_1$ et $f_2$). Zelterman [1988] a observé que $\lambda = (x+1)p_{x+1}/p_x$, ce qui l'a amené à proposer toute une série d'estimateurs de $\lambda$

de la forme $\hat{\lambda} = (i+1)f_{i+1}/f_i$. Partant du principe que les unités jamais détectées ressemblent plus aux unités rarement détectées qu'à celles très souvent observées, Zelterman proposa que le meilleur estimateur est celui calculé pour $i = 1$ ce qui conduit l'estimateur d'Horvitz-Thompson à se mettre sous la forme de l'estimateur de Zelterman $\widehat{N_Z}$ tel que

$$\widehat{N_Z} = \frac{N_{obs}}{1 - e^{\left(-\frac{2f_2}{f_1}\right)}}$$

Un autre estimateur très usité a été développé par Chao [1987 ; 1989] qui a proposé une borne inférieure à l'estimation. Observant qu'en présence d'hétérogénéité, $r_{i+1} \geq r_i$, on a donc en particulier $2p_2/p_1 \geq p_1/p_0$ ce qui permet d'obtenir une borne inférieure de $f_0$ telle que $\widehat{f_0} \geq f_1^2/2f_2$ ce qui conduit à l'estimateur de Chao tel que $\widehat{N_C} = N_{obs} + f_1^2/2f_2$, puis à l'estimateur de Chao corrigé tel que

$$\widehat{N_{Ccorr}} = N_{obs} + \frac{f_1(f_1 - 1)}{2f_2 + 2}$$

Ces estimateurs sont dits robustes, car ne faisant intervenir que les première et deuxième observations, ils restent exacts même en présence d'hétérogénéité dans les comptages supérieurs à deux [Wilson et Collins, 1992]. Pour les appliquer, il suffit donc que les observations 1 et 2 se comportent comme une loi de Poisson. Cependant, la principale critique de l'estimateur de Zelterman est qu'il est toujours associé à une large variance, ce qui fait que l'estimateur de Chao lui est très souvent préféré. Dans ce travail, seul l'estimateur corrigé de Chao qui est le plus robuste et celui associé à l'intervalle de confiance le plus étroit [Wilson et Collins, 1992] a été utilisé. Le calcul de l'estimateur a été réalisé grâce au logiciel SPADE [Chao et Shen, 2006].

## 4. LES PROBLEMES POSES PAR LA TREMBLANTE CLASSIQUE

Les méthodes de capture-recapture sont depuis peu utilisées pour estimer le nombre de cas de tremblante et l'exhaustivité des systèmes de surveillance pour cette maladie. La première application a utilisé une approche multiliste considérant chaque protocole de surveillance (abattoir, équarrissage et suspicion clinique) comme une source d'information [Del Rio Vilas *et al.*, 2005]. Le problème auquel se sont heurtés les auteurs est que dans le cas de la tremblante, et de la majorité des maladies animales réglementées, lorsqu'un cas est détecté, les mesures d'intervention sont telles que la probabilité de recapture par une autre source devient très faible. Par conséquent, les différentes sources ne se recouvrent quasiment pas. L'approche uniliste, qui permet une utilisation différente des données a alors été envisagée [Bohning et Del Rio Vilas, 2008 ; Del Rio Vilas et Bohning, 2008]. Dans ces deux études, l'ensemble des données de la surveillance a été utilisé pour calculer les estimateurs non paramétriques de Zelterman et de Chao. L'analyse a été affinée grâce à l'utilisation de co-variables (taille de l'élevage et origine géographique). L'effet de ces co-variables s'est avéré non significatif. Enfin, une dernière étude a utilisé, pour prendre en compte l'hétérogénéité de détection, les modèles de mélange de Poisson, appliqués encore une fois à l'ensemble des données issues de la surveillance en Grande Bretagne [Kuhnert *et al.*, 2008]. L'hétérogénéité dans la détection ne peut sans doute pas être mieux prise en compte dans l'état actuel des connaissances et compte tenu des données disponibles. Cependant, l'utilisation de l'ensemble des données issues de la surveillance (cas initiaux + cas additionnels) n'était probablement pas la mieux adaptée aux analyses réalisées comme cela est expliqué plus loin.

L'application des méthodes de capture-recapture unilistes aux données de tremblante classique en France pose un triple problème. Considérons le paramètre $\lambda$ caractérisant la loi de Poisson qui est suivie par le nombre de détections de cas de tremblante dans un troupeau infecté. De par l'approximation de la loi binomiale par la loi de Poisson, ce paramètre peut être considéré comme le produit de $Nt$ (le nombre d'animaux dans le troupeau) et d'une probabilité individuelle (à l'échelle de l'animal) d'être détecté comme infecté par la tremblante, conditionnellement à l'infection du troupeau (*i.e.* à l'infection d'au moins un animal dans le troupeau). Appelons cette dernière probabilité $Pdi|it$[5].

[5] di : détection individuelle ; it : infection troupeau.

Cette probabilité peut elle-même être considérée comme le produit de la probabilité individuelle d'infection conditionnellement à l'infection du troupeau (*Pii*|*it*)[6] et de la probabilité individuelle de détection de l'infection conditionnellement à l'infection individuelle (*Pdi*|*ii*).

$$\lambda = Nt^{*}(Pdi|it)$$

avec $Pdi|it=(Pii|it)^{*}(Pdi|ii)$

Les modèles unilistes les plus simples peuvent être utilisés à la condition que le même paramètre *λ* s'applique à tous les troupeaux infectés par la tremblante et à tous les animaux de chaque troupeau infecté par la tremblante. Cette condition n'est pas remplie, dès lors que la taille des troupeaux infectés varie fortement (*i.e. Nt* varie fortement entre troupeaux), et/ou que (*Pii*|*it*) varie au sein ou entre les troupeaux, et/ou que (*Pdi*|*ii*) varie au sein ou entre les troupeaux.

Dans le cas de la tremblante, les facteurs faisant potentiellement varier les différents paramètres de *λ* sont multiples. Tout d'abord, la taille des élevages infectés (*Nt*) est en effet fortement variable. Ensuite, la probabilité d'infection individuelle conditionnellement à l'infection du troupeau (*Pii*|*it*) varie selon différents paramètres individuels comme l'âge de l'animal et la génétique de l'animal (génotypes résistants), ou selon différents paramètres plus généraux comme la pression d'infection dans le troupeau voire la souche de tremblante. Enfin, la probabilité de détection individuelle conditionnellement à l'infection de l'individu (*Pdi*|*ii*) varie aussi selon différents facteurs. En premier lieu, les sensibilités des tests de dépistage (Biorad® et Prionics®), pourraient être différentes. Ainsi, un animal infecté testé avec le test Biorad® aurait une plus grande probabilité d'être détecté, donc la probabilité de détection individuelle varierait d'un élevage à l'autre (en faisant l'hypothèse que tous les animaux d'un même élevage sont testés avec un test identique).

En second lieu, on peut remarquer que les taux de sondage diffèrent d'un département à l'autre, provoquant aussi des variations de la probabilité de détection individuelle entre les départements. Enfin, l'échantillonnage non aléatoire aussi bien en équarrissage qu'à l'abattoir, influe également sur les variations de la probabilité de détection entre élevages [Morignat *et al*., 2006]. L'association de tous ces facteurs produit donc des variations du paramètre *λ* entre les élevages (*hétérogénéité inter-troupeaux*) et même à l'intérieur des élevages (*hétérogénéité intra-troupeau*). Il existe des outils permettant de prendre en compte ces hétérogénéités (*cf*. *supra*). Enfin, un dernier facteur important à considérer provoque des variations de la probabilité de détecter un animal infecté (*Pdi*|*ii*). Ce facteur résulte de la conception même du protocole de surveillance de la tremblante en Europe. En effet, dès lors qu'un (ou des) animal (animaux) est (sont) détecté(s) infecté(s) par un des trois protocoles de surveillance (abattoir, équarrissage, surveillance clinique), le sous-ensemble des animaux génétiquement sensibles du reste du troupeau est testé pour identifier l'ensemble des cas additionnels présents dans l'élevage. Il ne s'agit plus ici d'un problème d'hétérogénéité de détection des cas au sein d'un troupeau, mais bien d'un problème de dépendance positive de détection : ainsi, la probabilité de détecter un cas additionnel sachant qu'un cas initial a été identifié, est supérieure à la probabilité de détecter un cas initial. Dans le domaine des capture-recaptures en écologie, on parle de « *trap-happiness* » : la probabilité pour un animal de se faire recapturer est supérieure à la probabilité de se faire capturer. Les méthodes classiques de gestion de l'hétérogénéité présentées précédemment ne peuvent pas prendre en compte ces phénomènes de dépendance entre les observations successives. Une solution consiste à ignorer la détection des cas additionnels et à ne se focaliser que sur le comptage des cas initiaux : on peut considérer qu'au sein d'un élevage, les probabilités pour les animaux infectés d'être détectés en tant que cas initiaux sont relativement indépendantes.

[6] ii : infection individuelle.

# III - RESULTATS

## 1. RESULTATS DES ANALYSES DES DONNEES DE TREMBLANTE

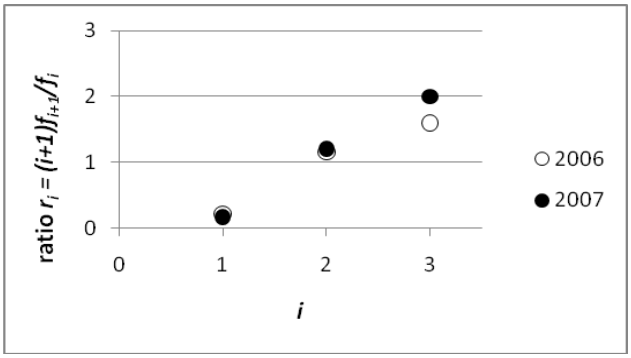
### 1.1. MISE EN EVIDENCE DE L'HETEROGENEITE

Pour illustrer l'hétérogénéité très probable de la probabilité de détection des cas initiaux entre élevages infectés, le ratio $r_i = (i+1)f_{i+1}/f_i$ a été calculé pour i allant de 1 à 3, et une représentation graphique de $r_i$ en fonction de $i$ est présentée en figure 2. Bien que le graphique ne comporte que trois points par année, il apparaît que le ratio $r_i$ croît, témoignant de la présence d'hétérogénéité dans les données. Il semble alors peu judicieux d'utiliser l'estimateur basé sur le maximum de vraisemblance du paramètre $\lambda$ de la loi de Poisson, pour estimer le nombre total d'élevages infectés par la tremblante classique. L'estimateur non paramétrique de Chao est donc le plus approprié et a été retenu pour cette étude.

Il est possible de démontrer que cette linéarité de $r_i$ peut être mise en relation avec une distribution du paramètre $\lambda$ selon une loi Gamma. Cependant, la faible quantité de données utilisées et l'observation de seulement trois valeurs par année impose de la prudence quant à l'interprétation de ce graphique. Néanmoins, il confirme la présence d'hétérogénéité dans la détection. Ainsi, de très nombreux facteurs de variation interviennent dans la probabilité d'un élevage infecté d'être détecté. Bien que n'ayant que peu de données, nous avons essayé de stratifier l'analyse selon un critère géographique. En effet, les taux de sondage à l'abattoir et à l'équarrissage varient d'un département à l'autre (variations de 0,10% à 5,85% à l'équarrissage par exemple pour l'année 2007). En 2006 et en 2007, près de 80% des cas initiaux ont été détectés à l'équarrissage, nous avons donc choisi de stratifier l'analyse uniquement selon le taux de sondage à l'équarrissage. Nous avons ainsi pu observer l'influence du taux de sondage en équarrissage sur la sensibilité générale de la détection.

**Figure 2**

**Mise en évidence de l'hétérogénéité de la détection des cas de tremblante classique ovine, en représentant le ratio $r_i = (i+1)f_{i+1}/f_i$ en fonction de la fréquence d'observation $i$ pour les années 2006 et 2007 ($f_i$ est le nombre d'élevages infectés détectés $i$ fois)**




### 1.1. ESTIMATIONS

Le tableau 3 présente l'estimateur de Chao calculé pour chacune des deux approches ainsi que l'intervalle de confiance associé. Les deux types d'approche (prise en compte des cas initiaux uniquement et prise en compte de l'ensemble des cas) sont représentés en colonne, et les différents jeux de donnés (2006 et 2007) sont représentés en ligne. Nous avons choisi d'illustrer l'analyse stratifiée uniquement pour l'année 2006 car les cas issus de la surveillance de 2007 sont trop peu nombreux pour fournir des résultats importants. Ainsi en 2006, pour les données

non stratifiées de la surveillance des cas initiaux, l'estimateur corrigé de Chao conduit à un total de 659,6 [422,9 ; 1095,1] élevages infectés par la tremblante ce qui correspond à une sensibilité de détection de 21%. Concernant l'analyse stratifiée, la strate 1 est composée des élevages détectés infectés qui sont situés dans les départements où le taux de sondage à l'équarrissage est supérieur à 1%. La strate 2 est composée des autres élevages. Seulement 137 élevages ont pu être inclus dans l'analyse stratifiée car la situation géographique de quatre élevages n'a pas pu être définie avec certitude.

Le tableau 3 illustre que, selon l'approche considérée, les estimateurs varient de manière assez importante. Ainsi pour l'année 2006, en considérant la population non stratifiée, les deux estimateurs ressortent significativement différents. Pour essayer de mieux comprendre ce phénomène, nous avons utilisé une approche par simulation mimant la surveillance de la tremblante en France.

**Tableau 3**

**Estimation du nombre (et de l'intervalle de confiance à 95%) d'élevages français infectés par la tremblante en 2006 et 2007 selon le type de données utilisées (cas index uniquement ou cas totaux)**

| | | | Nombre d'élevages détectés | Prise en compte des cas index uniquement | | Prise en compte des cas totaux (cas index + cas additionnels) | |
|---|---|---|---|---|---|---|---|
| | | | | Estimateur de Chao corrigé | Sensibilité de la détection* | Estimateur de Chao corrigé | Sensibilité de la détection* |
| 2006 | Population non stratifiée | | 141 | **659,6 [422,9 ; 1095,1]** | 0,21 [0,13 ; 0,33] | **272,0 [204,2 ; 401,7]** | 0,52 [0,35 ; 0,69] |
| | Population stratifiée | Strate 1** | 74 | 263,1 [162,7 ; 477,1] | 0,29 [0,16 ; 0,45] | - | - |
| | | Strate 2** | 63 | 371,0 [188,0 ; 822,2] | 0,17 [0,08 ; 0,34] | - | - |
| | | Total*** | 137 | 634,1 [350,7 ; 1299,3] | 0,22 [0,11 ; 0,39] | - | - |
| 2007 | Population non stratifiée | | 67 | **352,2 [187,9 ; 739,7]** | 0,19 [0,10 ; 0,36] | **132,9 [92,7 ; 225,5]** | 0,51 [0,30 ; 0,72] |

* La sensibilité de la détection est calculée comme le rapport entre le nombre d'élevages détectés et l'estimateur corrigé de Chao.

** La strate 1 est composée des élevages situés dans les départements où le taux de sondage à l'équarrissage est supérieur à 1%. La strate 2 est composée des autres élevages.

*** La ligne « total » correspond à la somme des estimateurs et des intervalles de confiance calculés dans chacune des strates.

## 2. SIMULATION : MISE EN EVIDENCE D'UNE SOUS-ESTIMATION DANS LE CAS D'UNE DEPENDANCE POSITIVE

La dépendance positive dans la situation de surveillance de la tremblante en France a été modélisée le plus simplement possible. Pour cela, la taille des cheptels a été considérée comme étant la même pour tous les élevages (N=100), et la probabilité de détecter un individu infecté au sein d'un élevage infecté comme identique pour tous les individus infectés de tous les élevages infectés. Il n'y a donc pas d'hétérogénéité de détection individuelle. La distribution du nombre réel d'animaux infectés par élevage a été modélisée par une loi géométrique de paramètre $p = 0,5$ sur un nombre total de 1000 élevages. A ces élevages, un protocole de surveillance unique dont la probabilité de détection individuelle est $p_1 = 0,2$ a été appliqué. Ce premier protocole permet de détecter les cas initiaux. Si dans un des élevages infectés, au moins un animal est détecté par le protocole de surveillance des cas initiaux, un protocole de détection des cas additionnels est alors appliqué au reste du troupeau. Pour ce deuxième protocole, trois sensibilités individuelles croissantes ($p_2$) ont été utilisées (0, 0,2, et 0,9). Les valeurs des probabilités de détection ont été choisies arbitrairement car l'objectif de cette simulation est d'illustrer le phénomène de dépendance

positive. On peut considérer que plus la probabilité de détection des cas additionnels augmente, plus la dépendance positive entre les détections est importante. Grâce au logiciel R [R-Development-Core-Team, 2008], mille simulations ont été réalisées dans chaque situation, et à chaque fois, les nombres d'élevages dans lesquels un cas, deux cas, trois cas,… ont été détectés ont permis de calculer l'estimateur de Chao.

La moyenne et l'écart-type de la distribution de cet estimateur dans chaque situation sont présentés dans le tableau 4. Le biais de l'estimateur a été calculé comme le quotient de la différence entre l'estimateur de Chao et la valeur réelle, par la valeur réelle.

**Tableau 4**

**Moyennes et écart-types de l'estimateur de Chao dans chacune des situations de dépendance positive croissante**

| $p_1$ | $p_2$ | $N_C$ [écart-type] | Biais de l'estimateur |
|---|---|---|---|
| 0,2 | 0 | 632,3 [130,9] | +25% |
| | 0,2 | 321,1 [39,5] | -37% |
| | 0,9 | 200,2 [14,3] | -60% |

$p_1$ : probabilité individuelle de détection des cas index ;
$p_2$ : probabilité individuelle de détection des cas additionnels ;
$N_C$ : estimateur de Chao.

L'examen du tableau 4 montre nettement que lorsque la dépendance positive entre les captures augmente, la sous-estimation du nombre total d'unités infectées s'accentue. On peut remarquer que même si la probabilité de détection des cas additionnels reste faible ($p_2 = 0,2$), la sous-estimation est si importante que la vraie valeur (500) n'appartient même plus à l'intervalle de confiance à 95%. Cette simulation montre donc bien que si l'on construit l'analyse avec l'ensemble des cas détectés par le système de surveillance (cas initiaux et cas additionnels), on risque de sous-estimer de manière importante le nombre de cas réels.

Enfin, on peut remarquer que, dans le cas où l'on ne prend en compte que la détection des cas initiaux ($p_2 = 0$), c'est-à-dire dans le cas où il n'y a pas de dépendance positive, les estimateurs tendent à surestimer la vraie valeur. Ceci est dû au fait qu'il existe une certaine part de dépendance négative non négligeable dans le cas de la détection d'une maladie à faible prévalence intra-troupeau comme la tremblante. En effet, statistiquement, dans le cas d'une faible prévalence intra-troupeau, la probabilité de détecter un animal infecté est supérieure à la probabilité de détecter un animal infecté sachant qu'un animal infecté a déjà été détecté. Cette dépendance négative existe toujours mais elle est plus marquée pour les faibles prévalences intra-troupeau. Donc en considérant que la tremblante est une maladie à faible prévalence intra-troupeau, l'application des méthodes unilistes de capture-recapture aux données de surveillance tend à surestimer le nombre de cas réels. On a donc seulement accès à une borne supérieure de la vraie valeur.

## IV - DISCUSSION

L'objectif de l'étude était d'étudier l'importance du choix des données dans les analyses de capture-recapture, en faisant l'hypothèse que l'utilisation de l'ensemble des données de tremblante (cas initiaux et cas additionnels) provoquait une forte sous-estimation du nombre réel d'élevages infectés. Comme cela a été montré par les résultats de l'analyse des données de tremblante, l'estimation du nombre total d'élevages infectés calculé en prenant en compte l'ensemble des cas (cas initiaux + cas additionnels), est effectivement largement

inférieure à celle calculée en prenant en compte uniquement les cas initiaux. Les résultats de la simulation nous ont permis de conclure à une sous-estimation importante de la vraie valeur du nombre d'élevages infectés, dans la situation où l'on considère l'ensemble des cas. Nous avons donc été amenés à éliminer tous les cas additionnels pour ne focaliser l'analyse que sur la détection des cas initiaux.

L'hypothèse d'homogénéité de détection des élevages a été mise en question par le rôle de facteurs liés à l'épidémiologie de la maladie (âge moyen des troupeaux, génétique des animaux, taille des troupeaux, pression d'infection, souche) et par le rôle de facteurs liés à la surveillance de la maladie (hétérogénéité du taux de sondage entre départements, échantillonnage non aléatoire à l'abattoir et à l'équarrissage, sensibilité des tests). Cette hétérogénéité, mise en évidence par la représentation graphique de l'évolution du ratio $r_i$ en fonction de $i$, représentant l'évolution du rapport de deux observations successives, nous a conduits à ne pas utiliser l'estimateur fondé sur le maximum de vraisemblance du paramètre de Poisson, qui aurait sous-estimé la vraie valeur que l'on cherchait à estimer. L'estimateur que nous avons considéré comme le plus adapté à la situation, est l'estimateur de Chao, dont nous avons choisi d'utiliser la forme corrigée pour compenser le biais introduit par l'estimateur simple. Cependant, il est probable que l'ensemble de l'hétérogénéité de la population n'ait pas pu être pris en compte par l'utilisation de l'estimateur de Chao qui reste relativement grossier. D'autres solutions plus complexes existent pour compenser plus finement l'hétérogénéité de la population, comme l'utilisation de l'estimateur non paramétrique de Chao-Bunge qui permet de prendre en compte une hétérogénéité de type Gamma [Chao et Bunge, 2002], ou comme la modélisation des données par des lois de type mélange de Poisson [Kuhnert et Bohning, 2009] ou de type négatif binomial [Cruyff et Van Der Heijden, 2008]. Ces méthodes alternatives qui tendent à se développer en épidémiologie ont pour avantage de très bien s'adapter aux données [Bohning *et al.*, 2004 ; Cruyff et Van Der Heijden, 2008]. Cependant, le problème est qu'elles nécessitent un assez grand nombre de données pour être considérées. Dans le cas de la tremblante, *a fortiori* si l'on n'utilise que les données issues des cas initiaux, les cas et les observations sont trop peu nombreux, limitant l'application de ces méthodes alternatives prometteuses. Dans notre cas, l'estimateur corrigé de Chao apparait comme le plus adapté.

L'hypothèse de population fermée est indiscutablement remise en cause dans le contexte de la tremblante, car en considérant que les cas additionnels donnent une représentation fiable du nombre d'individus infectés dans un élevage infecté, il apparait alors clairement que ce nombre varie d'un élevage à l'autre, ne validant pas l'hypothèse sous-jacente. Cette situation est très certainement à l'origine d'une surestimation du nombre d'élevages infectés du fait d'un déplacement artificiel des données de distribution du nombre de cas dans les élevages vers les faibles valeurs.

Même en n'utilisant que les données issues de la détection des cas initiaux, il semble que l'hypothèse d'indépendance des détections successives ne soit pas complètement validée. En effet, du fait de la mise en place des mesures de police sanitaire dès la découverte d'un cas initial, la durée qui pourrait permettre la découverte de nouveaux cas initiaux est raccourcie entrainant une dépendance négative des détections successives, d'où une possible surestimation.

Enfin, comme cela peut être observé dans les résultats de la simulation, l'utilisation de la très commune distribution de Poisson pour modéliser les données de comptage répété est aussi à discuter. En effet, la loi de Poisson n'est qu'une approximation de la loi binomiale pour des grandes tailles de population (on parle généralement de tailles supérieures à 30 individus). Dans le cas de la tremblante, les tailles de population sont représentées par les nombres d'animaux infectés dans chacun des élevages infectés. Dans notre application, près de 92% des élevages ont certainement moins de dix animaux infectés (tableau 2). La modélisation des données par la loi de Poisson n'est donc sans doute pas la plus adaptée, et conduirait à une surestimation du nombre réel d'élevages infectés. Cependant, l'utilisation du modèle binomial donne des résultats incohérents du fait de la très faible quantité de données.

Il est donc important de garder en mémoire que lorsque l'on s'intéresse au nombre de cas d'une maladie détectée dans des élevages, dans l'objectif d'appliquer des modèles unilistes de capture-recapture, il existe un risque de surestimer le nombre réel d'élevages infectés, *a fortiori* si la maladie est présente dans les troupeaux avec une faible prévalence (*cf*. III.2.). Les estimations de 659,6 [422,9 ;

1095,1] et de 352,2 [187,9 ; 739,7] élevages infectés par la tremblante respectivement en 2006 et 2007, déterminées par le calcul de l'estimateur corrigé de Chao, peuvent donc être raisonnablement considérées comme légèrement surestimées. Ces estimations conduisent au calcul de la sensibilité de la détection du système de surveillance qui semble relativement constante entre les deux années (21% en 2006 et 19% en 2007). De plus, comme on pouvait s'y attendre même si le résultat n'est pas significatif, il semble que l'analyse stratifiée pour l'année 2006 montre que dans les départements où le taux de sondage à l'équarrissage était supérieur à 1%, les élevages infectés ont été mieux détectés que dans les autres départements (sensibilité de la détection de 35% contre 25%).

## V - CONCLUSION

Dans le cas de la tremblante ovine, le principal problème rencontré est que cette maladie est rare et relativement peu contagieuse, ce qui limite le nombre d'élevages infectés (d'où le faible nombre d'élevages détectés) ainsi que le nombre d'animaux infectés au sein d'un élevage infecté (d'où le faible nombre d'animaux détectés). Le faible nombre d'élevages infectés empêche une stratification efficace des données, et le faible nombre d'animaux détectés par élevage infecté limite l'utilisation des modèles complexes (modèles de mélange et modèles mixtes). La tremblante laisse donc peu de libertés d'analyse. Malgré ces limites, les méthodes unilistes de capture-recapture, sont des méthodes très prometteuses pour estimer les tailles de populations infectées non détectées. En épidémiologie vétérinaire, elles semblent beaucoup mieux adaptées que les méthodes multilistes car le faible recouvrement des sources empêche la bonne application de ces dernières. Cependant, il apparait que ces méthodes nécessiteraient encore de nombreuses recherches concernant les applications vétérinaires. Tout d'abord, il est évident que le fait d'avoir supprimé de l'analyse les cas additionnels pour s'affranchir de la dépendance positive de la détection des cas au sein d'un élevage, fait perdre de l'information quant à la description des élevages détectés, et surtout des élevages peu détectés auxquels sont censés le plus ressembler les élevages non détectés [Zelterman, 1988]. Une future piste de recherche serait d'adapter à l'épidémiologie les modèles appliqués en écologie, ces derniers permettant de concilier une hétérogénéité de la probabilité de capture avec une auto-corrélation des captures successives [Chao, 2001 ; Pledger et Phillpot, 2008]. Enfin, une correction de l'estimateur de Chao aurait besoin d'être apportée pour compenser la surestimation liée à la notion de détection des cas différents au sein d'un même troupeau. Ces recherches sont en cours.

## BIBLIOGRAPHIE

Baumgarten L., Heim D., Fatzer R., Zurbriggen A., Doherr M.G. - Assessment of the Swiss approach to scrapie surveillance. *Vet. Rec.*, 2002, **151**, 545-547.

Bohning D., Del Rio Vilas V. - Estimating the hidden number of scrapie affected holdings in Great Britain using a simple, truncated count model allowing for heterogeneity. *J. Agr. Biol. Envir. Stat.*, 2008, **13**, 1-22.

Bohning D., Dietz E., Kunhert R., Schön D. - Mixture models for capture-recapture count data. *Statistical Methods and Applications*, 2005.

Bohning D., Suppawattanabodee B., Kusolvisitkul W., Viwatwongkasem C. - Estimating the number of drug users in Bangkok 2001: a capture-recapture approach using repeated entries in one list. *Eur. J. Epidemiol.*, 2004, **19**, 1075-1083.

Calavas D., Philippe S., Ducrot C., Schelcher F., Andreoletti O., Belli P., Fontaine J., Perrin G., Savey M. - Bilan et analyse de trente mois de fonctionnement

du Réseau français d'épidémiosurveillance de la tremblante des petits ruminants. *Epidémiol. et santé anim.*, 1999, **35**, 43-50.

Carothers A. - Capture-recapture methods applied to a population with known parameters. *J. Anim. Ecol.*, 1973, **42**, 125-146.

Chao A. - Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 1987, **43**, 783-791.

Chao A. - Estimating animal abundance with capture frequency data. *J. Wildl. Manage.*, 1988, **52**, 295-300.

Chao A. - Estimating population size for sparse data in capture-recapture experiments. *Biometrics*, 1989, **45**, 427-438.

Chao A. - An overview of closed capture-recapture models. *Journal of agricultural, biological and environmental statistics*, 2001, **6**, 158-175.

Chao A., Bunge J. - Estimating the number of species in a stochastic abundance model. *Biometrics*, 2002, **58**, 531-539.

Chao A., Shen T. - User's guide for program SPADE. *Available online at: http://chao.stat.nthu.edu.tw/softwareCE.html*, 2006.

Cruyff M.J., van der Heijden P.G. - Point and interval estimation of the population size using a zero-truncated negative binomial regression model. *Biom. J.*, 2008, **50**, 1035-1050.

Del Rio Vilas V.J., Bohning D. - Application of one-list capture-recapture models to scrapie surveillance data in Great Britain. *Prev. Vet. Med.*, 2008, **85**, 253-266.

Del Rio Vilas V.J., Sayers R., Sivam K., Pfeiffer D., Guitian J., Wilesmith J.W. - A case study of capture-recapture methodology using scrapie surveillance data in Great Britain. *Prev. Vet. Med.*, 2005, **67**, 303-317.

Hoinville L.J., Hoek A., Gravenor M.B., McLean A.R. - Descriptive epidemiology of scrapie in Great Britain: results of a postal survey. *Vet. Rec.*, 2000, **146**, 455-461.

Hook E.B., Regal R.R. - Capture-recapture methods in epidemiology: methods and limitations. *Epidemiol. Rev.*, 1995, **17**, 243-264.

IWGDMF. - Capture-recapture and multiple-record systems estimation I: History and theoretical development. International Working Group for Disease Monitoring and Forecasting. *Am. J. Epidemiol.*, 1995, **142**, 1047-1058.

Kuhnert R., Bohning D. - CAMCR:Computer-Assisted Mixture model analysis for Capture-Recapture count data. *AStA Adv. Stat. Anal.*, 2009, **93**, 61-71.

Kuhnert R., Del Rio Vilas V.J., Gallagher J., Bohning D. - A bagging-based correction for the mixture model estimator of population size. *Biom. J.*, 2008, **50**, 993-1005.

Morignat E., Cazeau G., Biacabe A.G., Vinard J.L., Bencsik A., Madec J.Y., Ducrot C., Baron T., Calavas D. - Estimates of the prevalence of transmissible spongiform encephalopathies in sheep and goats in France in 2002. *Vet. Rec.*, 2006, **158**, 683-687.

Pledger S., Phillpot P. - Using mixtures to model heterogeneity in ecological capture-recapture studies. *Biom. J.*, 2008, **50**, 1022-1034.

R-Development-Core-Team. - R: A language and environment for statistical computing. *R. Foundation for Statistical Computing, Vienna, Austria.*, 2008, **ISBN 3-900051-07-0**, URL http://www.r-project.org.

Wilson R., Collins M. - Capture-recapture estimation with samples of size one using frequency data. *Biometrika*, 1992, **79**, 543-553.

Zelterman D. - Robust estimation in truncated discrete distributions with application to capture-recapture experiments. *J. Statistic Plan. Inf.*, 1988, **18**, 225-237.