School of Mathematical
and Physical Sciences

University of Reading

Department of Mathematics and Statistics

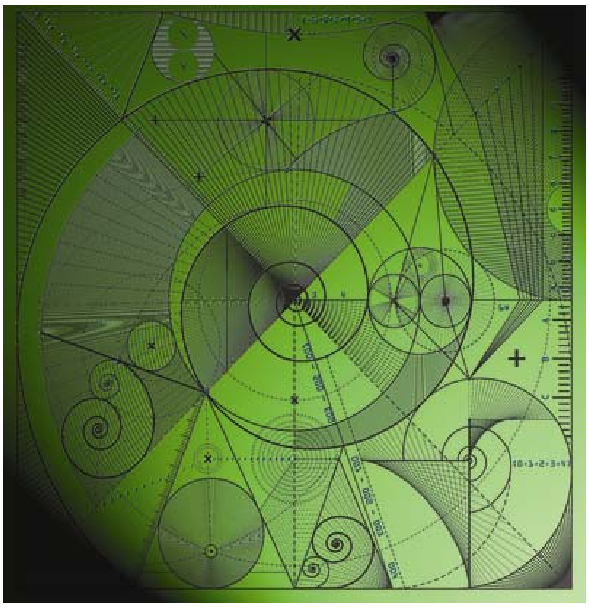
Preprint MPS_2010-30

17 August 2010

Capture-Recapture Estimation by Means of Empirical Bayesian Smoothing with an Application to the Geographical Distribution of Hidden Scrapie in Great Britain

by

Dankmar Böhning, Ronny Kuhnert and Victor Del Rio Vilas



Capture-Recapture Estimation by Means of Empirical Bayesian Smoothing with an Application to the Geographical Distribution of Hidden Scrapie in Great Britain

Dankmar Böhning
Department of Mathematics and Statistics
School of Mathematical and Physical Sciences
University of Reading, Whiteknights
Reading, RG6 6BX, UK
email: d.a.w.bohning@reading.ac.uk

Ronny Kuhnert
Robert Koch-Institute, FG 23, Berlin, Germany
email: KuhnertR@rki.de

Victor Del Rio Vilas
Department for Environment, Food and Rural Affairs (Defra)
London, SW1P 3JR, UK
email: victor.delriovilas@defra.gsi.gov.uk

August 17, 2010

Abstract

This note discusses population size estimation on the basis of a frequency distribution of zero-truncated counts and is motivated by a study on the geographical distribution of hidden scrapie in Great Britain. Aggregation of scrapie cases is considered at the county level and results in sparse zero-truncated count distributions which make the application of conventional capture-recapture procedures for estimating the hidden part of the scrapie affected population difficult. We suggest a smoothed generalization of Zelterman's estimator of population size which overcomes the overestimation bias of the conventional Zelterman estimator and instead produces a lower bound, typically larger than Chao's lower bound estimator. The estimator uses an empirical Bayes approach with various choices for the prior distribution including a parametric choice of the Gamma distribution as well as various non-parametric ones. A simulation study investigates the performance of the new estimators, also in comparison to conventional estimators. The empirical Bayes estimator with a nonparametric mixture model as prior performs well and the boundary problem of the conventional nonparametric discrete mixture model estimator leading to spurious population size is avoided. In the application on hidden scrapie in Great Britain the new estimators lead to scrapie maps of observed–hidden ratios as well as completeness of the current surveillance system.

Some key words: capture-recapture, empirical Bayes, nonparametric mixture model, geographical analysis

1 Introduction

For integer N, we consider a sample of counts $x_1,\ x_2,\ ...,\ x_N \in \{0, 1, 2, ..., \}$ arising from a count random variable X having a mixture probability density function

$$p_x = \int_0^\infty p(x|\lambda)q(\lambda)d\lambda \tag{1}$$

with unspecified mixing density $q(\lambda)$ and a mixture kernel $p(x|\lambda)$ which needs to be specified. In this paper, a typical choice for the mixture kernel is the Poisson $p(x|\lambda) = Po(x|\lambda) = \exp(-\lambda)\lambda^x/x!$ though other choices are possible as well. Whenever $x_i = 0$ unit i remains unobserved, so that only a zero-truncated sample of size $n = \sum_{j=1}^m f_j$ is observed, where f_j is the frequency of counts with value $x = j$ and m is the largest observed count. Hence, f_0 and consequently N are unknown. The purpose is to find an estimate of the size N. Since frequently the count variable X represents repeated identifications of an individual in an observational period, the problem at hand is a special form of the capture-recapture problem (see Bunge and Fitzpatrick (1993) for a review on the topic).

The sample of counts $x_1,\ x_2,\ ...,\ x_N$ can occur in several ways. A target population which might be difficult to count consists out of N units. This population might be a wildlife population, a population of homeless people, drug addicts, software errors or animals with a specific disease. Furthermore, let an identification device (a trap, a register, a screening test) be available that identifies unit i at occasion t where $t = 1, .., T$. Let the binary result be x_{it} where $x_{it} = 1$ means that unit i has been identified at occasion t and $x_{it} = 0$ means that unit i has not been identified at occasion t. The indicators x_{it} might be observed or not, but it is assumed that $x_i = \sum_{t=1}^T x_{it}$ is observed if at least one $x_{it} > 0$ for $t = 1, ..., T$. Only if $x_{i1} = x_{i2} = ... = x_{iT} = 0$ and, consequently $x_i = 0$, the unit i remains *unobserved.* In this kind of situation the *clustering* occurs by repeated identifications of the same unit.

In another setting, which is also the basis for this work, the clustering occurs by means of a grouping variable such as herds, holdings, households, or villages. In this case, x_{it} denotes if the $t-$th element in cluster i is identified ($x_{it} = 1$) or not ($x_{it} = 0$). In the example given in the next section the clusters are holdings of sheep and x_{it} informs about the disease status of the $t-$th animal in holding i. Note that $x_i = \sum_t x_{it}$ is observed only if it is positive. In other examples the cluster corresponds to villages or households, one of the earliest applications of this kind is the cholera-outbreak in a community in India studied by McKendrick (1926) in which the cluster corresponds to households in a village. A more recent example involves cholera occurrence in rural East Pakistan where the cluster structure consists of villages (see also Mosley *et al.* (1972)).

The paper is organized as follows. The next section 2 introduces the data on scrapie in Great Britain. In section 3 we review some of the existing approaches in the capture-recapture methodology for the setting of interest. Section 4 describes the development of a new set of empirical Bayes estimators which are then further evaluated by means of a simulation study. The application of the empirical Bayes estimator to the spatial data on scrapie in Great Britain, including the development of maps at county level of completeness and observed–hidden ratio, ends the paper in section 5.

2 The data of scrapie in Great Britain

We now consider as a specific case study the spatial distribution of scrapie in Great Britain. Classical scrapie, a neurological fatal disease of small ruminants is endemic in Great Britain (see Del Rio Vilas *et al.* (2006) for more details). There is ample evidence to support the occurrence of under-reporting affecting the clinical notification of scrapie cases (Hoinville *et al.* (2000), Del Rio Vilas *et al.* (2005), Böhning *et al.* (2008)). Although not established to date, there is reason to believe that, reflecting population and surveillance related heterogeneities, under–reporting presents an uneven distribution across Great Britain. The spatial analysis presented in the following uses county-specific disease data from the Scrapie Notifications Database (SND) (see Vilas *et al.* (2006) for more details), more specifically the number of confirmed clinical cases. Table 1 shows the frequency distribution f_x of the count of confirmed clinical cases X for $x = 1, 2, 3, ...$ by county. Evidently, there is a considerable range in the number of scrapie-affected holdings per county, ranging from counties with only 1 affected holding to counties with a large number of affected holdings, the largest number occurring in county 37 with 75 affected holdings.

Our main interest in the following analysis is to investigate the performance of the SND surveillance stream as measured in the *observed–hidden ratio* (the larger the ratio the better the system) as well as in the *completeness rate*, defined as the proportion of observed affected holdings among observed and hidden scrapie affected holdings. If the case count per holding is collapsed over all counties we find the distribution as given at the bottom of Table 1. With $f_1 = 298$, $f_2 = 89$ and $f_3 = 42$ most of the distribution is concentrated on counts of ones, twos and threes with the largest count occurring at 29.

3 Background on capture-recapture estimation

Before we go into the details of the suggested novel approach we give a brief review of the existing capture-recapture methodology for the setting of interest.

3.1 Heterogeneity

The importance of the mixture $p_x = \int_0^\infty p(x|\lambda)q(\lambda)d\lambda$ can be seen in the fact that it is a natural model for the population heterogeneity. There appears to be consensus (see for example Pledger [24] for the discrete mixture model approach and Dorazio and Royle (2005)for the continuous mixture model approach) that a simple model $p(x|\lambda)$ is not flexible enough to capture the variation in the re-capture probability for the different members of most real life populations. There has also been recently a debate on the identifiability of the binomial mixture model (see Link (2003, 2006) and Holzmann *et al.* (2006). Furthermore, using the nonparmatric maximum likelihood estimate (NPMLE) $\hat{q}(\lambda)$ of the mixing density $q(\lambda)$ in constructing an estimate of the population size $\hat{N} = n/[1 - \int_0^\infty \exp(-\lambda)\hat{q}(\lambda)d\lambda]$ leads to the *boundary problem*. This results in often unrealistic high values for the estimate of the population size (Wang and Lindsay (2005), Wang and Lindsay (2008)). Hence, a renewed interest has re–emerged in the lower bound approach for population size estimation suggested by Chao (1987).In this approach there is neither need to specify a mixing distribution, nor is there need to estimate it. In this sense it is completely non-parametric. To give some details of the lower bound approach consider the Poisson mixture kernel $\exp{(-\lambda)}\lambda^x/x!$. It follows from the Cauchy-Schwarz inequality that

$$\left(\int_0^\infty \exp{(-\lambda)}\lambda q(\lambda)d\lambda\right)^2 \leq \int_0^\infty \exp{(-\lambda)}q(\lambda)d\lambda \int_0^\infty \exp{(-\lambda)}\lambda^2 q(\lambda)d\lambda,$$

or equivalently, $p_1^2 \leq p_0(2p_2)$. Replacing the theoretical probabilities p_j by their sample estimates f_j/N for $j = 0, 1, 2$, the Chao lower bound estimate $f_1^2/(2f_2)$ for f_0 follows (see Chao [7], [8]) since the unknown denominator N cancels out. The estimate for the population size N is $\hat{N}_C = n + f_1^2/(2f_2)$. Since the Chao estimator uses only frequencies with counts of 1 and 2, a truncated sample consisting only of counts of ones and twos might be considered. This truncated sample leads to a binomial log-likelihood $f_1 \log(p_1) + f_2 \log(p_2)$ which is uniquely maximized for $\hat{p}_2 = 1 - \hat{p}_1 = f_2/(f_1 + f_2)$. Since $p_2 = \lambda/(\lambda + 2)$ and $p_1 = 2/(\lambda + 2)$ in a Poisson that truncates all counts except ones and twos, the estimate $\hat{\lambda} = 2f_2/f_1$ for the Poisson parameter λ suggested by Zelterman

(1988) arises. In the approach of Zelterman the homogeneous Poisson serves only as a working model and it was suggested by Zelterman that the estimate $\hat{N}_Z = \frac{n}{1-\hat{p}_0} = \frac{n}{1-\exp(-\hat{\lambda})}$ is more robust against misspecifications of the Poisson model than the usual maximum likelihood estimate.

3.2 A re-analysis of Zelterman estimation

We are interested in developing a generalization of the Zelterman estimator.Consider the Horvitz-Thompson-type estimate of the population size suggested by Zelterman (1988):

$$\hat{N}_Z = \frac{n}{1-\exp(\frac{-2f_2}{f_1})}. \tag{2}$$

Although the estimator (2) is popular among practitioners there are two disadvantages of the estimator:

- it uses only the frequencies f_1 and f_2 and ignores f_3 to f_m.
- it can experience *severe overestimation bias.*

The first issue is evident and results in large variance. The second issue is less evident but becomes clear with the following theorem.

Theorem 1 *Let $p_x = qp(x|\lambda)+(1-q)p(x|\mu)$ a discrete, two-component mixture with $p(x|\theta) = Po(x|\theta)$ being the Poisson kernel and $0 < q < 1$. Then,*

$$E(\hat{N}_Z) \approx N \frac{1-[q\exp(-\lambda)+(1-q)\exp(-\mu)]}{1-\exp\left(-\frac{q\exp(-\lambda)\lambda^2+(1-q)\exp(-\mu)\mu^2}{q\exp(-\lambda)\lambda+(1-q)\exp(-\mu)\mu}\right)}$$

$$\to_{\mu\to\infty} N\frac{1-q\exp(-\lambda)}{1-\exp(-\lambda)} \geq N$$

Proof. The theorem is proved by replacing sample frequencies by their theoretical values. □

Note that the biasing factor $\frac{1-q\exp(-\lambda)}{1-\exp(-\lambda)}$ can become arbitrarily large since it is a monotone decreasing function of q and λ. But even for realistic values of p and λ the factor can be considerably larger than 1. For example if $q = 0.5$ and $\lambda \leq 0.4$ the factor is larger than 2, so that the Zelterman estimate would overestimate severely. The question arises as to what is the source of this overestimation bias. We approach this question in the next theorem which states that the Zelterman estimator uses the *wrong* expected value in predicting f_0.

Theorem 2 *i) Let* $\log L(\lambda) = f_1 \log(p_1) + f_2 \log(p_2)$ *with* $p_1 = e^{-\lambda}\lambda/(e^{-\lambda}\lambda + e^{-\lambda}\lambda^2/2) = 2/(\lambda+2)$ *and* $p_2 = e^{-\lambda}\lambda^2/2/(e^{-\lambda}\lambda + e^{-\lambda}\lambda^2/2) = \lambda/(\lambda+2)$ *being the Poisson probabilities truncated to counts of ones and twos. Then* $\log L(\lambda)$ *is maximized for*

$$\hat{\lambda} = 2f_2/f_1.$$

ii)

$$E(f_0|f_1, f_2; \hat{\lambda}) = f_1^2/(2f_2), \text{ for } \hat{\lambda} = 2f_2/f_1.$$

Proof. For the first part, it is clear that $f_1 \log(p_1) + f_2 \log(p_2)$ is maximal for $\hat{p}_1 = f_1/(f_1 + f_2)$, which is attained for $\hat{\lambda} = 2f_2/f_1$. For the second part, we see that with $e_x = E(f_x|f_1, f_2; \lambda) = Po(x|\lambda)N$:

$$e_x = Po(x|\lambda)N = Po(x|\lambda)N = Po(x|\lambda)(e_0 + f_1 + f_2 + \sum_{j=3}^{\infty} e_j)$$

so that

$$e_0 + e_3^+ = [1 - Po(1|\lambda) - Po(2|\lambda)](e_0 + e_3^+) + [1 - Po(1|\lambda) - Po(2|\lambda)](f_1 + f_2)$$

with $e_3^+ = \sum_{j=3}^{\infty} e_x$. Hence

$$e_0 + e_3^+ = \frac{1 - Po(1|\lambda) - Po(2|\lambda)}{Po(1|\lambda) + Po(2|\lambda)}(f_1 + f_2)$$

and

$$\begin{aligned} e_0 = Po(0|\lambda)(f_1 + f_2 + e_0 + e_3^+) \quad &= Po(0|\lambda)(f_1 + f_2) + Po(0|\lambda) \\ &= \tfrac{1 - Po(1|\lambda) - Po(2|\lambda)}{Po(1|\lambda) + Po(2|\lambda)}(f_1 + f_2) \end{aligned}$$

$$= \frac{Po(0|\lambda)}{Po(1|\lambda) + Po(2|\lambda)}(f_1 + f_2) = \frac{f_1 + f_2}{\lambda + \lambda^2/2}.$$

Plugging in the maximum likelihood estimate $\hat{\lambda} = 2f_2/f_1$ for λ yields the desired result. □

Theorem 2 establishes a close connection between the approach by Zelterman and Chao's estimator. It shows that Zelterman's estimator of the Poisson parameter λ arises when all counts are truncated except counts of ones and twos and when the resulting likelihood is maximized. If the correct prediction for f_0 is used, namely the conditional expectation for the truncated Poisson model, the Chao estimator arises. Hence the strong overestimation of the original Zelterman estimator stems from using a *wrong* conditional expectation.

3.3 Comparing some conventional estimators in a simulation

Before we continue developing the generalized, adjusted version of the Zelterman estimator, we consider the performance of Chao and Zelterman estimators in a small simulation study. In the case of a homogeneous Poisson the maximum likelihood estimate is found by maximizing the likelihood of zero-truncated Poisson observations in λ:

$$\prod_{j=1}^{m} \left(\frac{p_j}{1 - p_0} \right)^{f_j} = \prod_{j=1}^{m} \left(\frac{1}{1 - \exp(-\lambda)} \exp(-\lambda)\lambda^j/j! \right)^{f_j},$$

or equivalently, in solving the following equation in $\hat{N}_{\text{hom}}$:

$$\hat{N}_{\text{hom}} = n \left(1 - \exp(-\frac{S}{\hat{N}_{\text{hom}}}) \right)^{-1}.$$

We have to maximize the zero-truncated Poisson mixture likelihood in Q to find the nonparametric maximum likelihood of the mixing distribution

$$L(Q) = \prod_{j=1}^{m} \left(\frac{p_j}{1 - p_0} \right)^{f_j} = \prod_{j=1}^{m} \left(\sum_{\ell=1}^{k} \frac{Po(j|\lambda_\ell)q_\ell}{1 - \sum_i \exp(-\lambda_i)q_i} \right)^{f_j}$$

where Q is the discrete mixing distribution giving k weights q_j to Poisson parameters λ_j:

$$Q = \begin{pmatrix} \lambda_1 & \lambda_2 & ... & \lambda_k \\ q_1 & q_2 & ... & q_k \end{pmatrix}.$$

Note that we have to maximize $L(Q)$ in terms of $\lambda_1, ..., \lambda_k$ and $q_1, ..., q_k$ but also in k to find the NPMLE. This is typically done in a step-wise manner by fixing k to be 1,2,3,..., and conditionally upon k using the EM algorithm for finding the MLE. Alternatively, using results from convex optimization a direct, gradient-function based algorithm might be employed. For details see Böhning and Kuhnert (2006). Occasionally, we might be interested in comparing mixture models with different number of components k by means of BIC-based selection criteria.After the NPMLE $\hat{Q}$ of Q has been identified, we can define

$$\hat{N}_{\text{NPMLE}} = \frac{n}{1 - \sum_{j=1}^{k} \exp(-\hat{\lambda}_j)\hat{q}_j}. \tag{3}$$

If the number of components k in the mixture model has been identified using the *Bayesian Information Criterion* we will denote the associated population size estimate by $\hat{N}_{\text{BIC}}$.

To illustrate the performance of these estimators we consider the following simulation experiments. Samples of counts $X_1, ...X_N$ were drawn from a two-component mixture of Poisson densities: $X \sim 0.5Po(1) + 0.5Po(\lambda)$, evidently with equal weights $q_1 = q_2 = 0.5$. The population size was set to $N = 100$ and $10,000$ replications used. Ignoring zero counts the estimators of Chao and Zelterman were determined as well as the maximum likelihood estimator under homogeneity and the nonparametric maximum likelihood estimator under heterogeneity. The results can be found in Table 2. When heterogeneity increases the Zelterman estimator overestimates whereas the MLE under homogeneity underestimates – both as expected. The Chao lower bound estimator does well under heterogeneity – again as expected. Most dominant in Table 2 is the drastic failure of the NPMLE which leads to spurious overestimating values.

4 A new empirical Bayes estimator of population size

Although it is clear that $2f_2/f_1$ estimates the Poisson parameter in the case that $p_x = Po(x|\lambda)$, it is not clear what it estimates when there is a mixing distribution present instead of Poisson homogeneity. Here, a Bayesian perspective is helpful. We think of the mixing distribution $q(\lambda)$ as a prior distribution on λ so that

$$E(\lambda|x) = \int_0^\infty \lambda \frac{Po(x|\lambda)q(\lambda)}{\int_0^\infty Po(x|\theta)q(\theta)d\theta} d\lambda \tag{4}$$

is the *posterior mean* w.r.t the prior $q(\lambda)$ and Poisson likelihood for observation x. Note that (4) can be further simplified to

$$\lambda_x = E(\lambda|x) = \frac{\int_0^\infty \lambda Po(x|\lambda)q(\lambda)d\lambda}{\int_0^\infty Po(x|\lambda)q(\lambda)d\lambda}$$

$$= (x+1)\frac{\int_0^\infty Po(x+1|\lambda)q(\lambda)d\lambda}{\int_0^\infty Po(x|\lambda)q(\lambda)d\lambda} = (x+1)\frac{p_{x+1}}{p_x},$$

where p_x is the marginal density (1). Before we continue on the ways to estimate the ratio of marginals we point out an important monotonicity property.

Theorem 3

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_m.$$

Proof. Consider

$$p_j = \int_0^\infty \exp(-\lambda)\lambda^j/j!q(\lambda)d\lambda$$

with unknown $q(\lambda)$ for $\lambda > 0$. Then, by means of the *Cauchy-Schwarz inequality* for random variables X and Y:

$$[E(XY)]^2 \leq E(X^2)E(Y^2)$$

we have that

$$\left(\int_0^\infty \overbrace{\sqrt{\exp(-\lambda)}\lambda^{(j-1)/2}}^{X}\overbrace{\sqrt{\exp(-\lambda)}\lambda^{(j+1)/2}}^{Y} d\lambda\right)^2$$

$$\leq \int_0^\infty \overbrace{\exp(-\lambda)\lambda^{(j-1)}}^{X^2} d\lambda \int_0^\infty \overbrace{\exp(-\lambda)\lambda^{(j+1)}}^{Y^2} d\lambda$$

or,

$$(j!\ p_j)^2 \leq (j-1)!\ p_{j-1}(j+1)!\ p_{j+1},$$

or, finally $\frac{jp_j}{p_{j-1}} \leq \frac{(j+1)p_{j+1}}{p_j}$. $\square$

Theorem 3 has an important application. Since under heterogeneity we have that $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_m$, we expect that the graph $x \to \hat{\lambda}_x = (x+1)f_{x+1}/f_x$ shows a monotone increasing pattern if heterogeneity is present. Hence we can develop a *diagnostic device* for the presence of heterogeneity by plotting $(x+1)f_{x+1}/f_x$ against x, which we call the *ratio plot.* The ratio plot for the SND data of the years 2002–2006 is presented in Figure 1. There is a clear evidence for a monotone increasing trend, hence a mixture model coping with the presence of heterogeneity appears appropriate.

Since $1/(1-\exp(-\lambda))$ is monotone non-increasing in λ we have the following corollary which we state without further proof.

Corollary 1

$$\sum_{x=1}^{m} \frac{f_x}{1-\exp(-\lambda_x)} \leq \frac{\sum_{x=1}^m f_x}{1-\exp(-\lambda_1)} = \frac{n}{1-\exp(-2p_2/p_1)}. \qquad (5)$$

Note that the Zelterman estimator occurs on the right-hand side of (5) if p_2/p_1 is replaced by its sample version f_2/f_1. Hence we expect that the overestimation bias of the Zelterman estimate is reduced if λ_x on the left-hand side of (5) is appropriately estimated. Furthermore, if $\frac{f_1}{1-\exp(-\lambda_1)} = f_1 + \frac{f_1}{\exp(\lambda_1)-1}$ is replaced by its first-order Taylor expansion $f_1 + \frac{f_1}{\lambda_1}$ and again $\lambda_1 = 2p_2/p_1$ estimated by $2f_2/f_1$, we find that

$$\frac{f_1}{\hat{\lambda}_1} = f_1 + \frac{f_1^2}{2f_2},$$

the lower bound estimator of Chao (1987, 1989). Hence we expect that the left-hand side of (5) provides an improved lower bound estimator if λ_x is estimated appropriately since

$$\sum_{x=2}^{m} \frac{f_x}{1-\exp(-\lambda_x)} \geq \sum_{x=2}^{m} f_x.$$

We are now considering ways of doing so.

The marginal density p_x can be estimated by the relative, empirical frequency f_x/N so that

$$\widehat{E(\lambda|x)} = \hat{\lambda}_x = (x+1)\frac{f_{x+1}}{f_x}$$

provides an estimate of the posterior mean $E(\lambda|x) = \lambda_x$ using the fact that the unknown denominators N cancel out. Hence, the Zelterman estimate occurs as a special case of the nonparametric, empirical Bayes estimator for observation x (Robbins (1955), Carlin and Louis (1997)).

The understanding of Zelterman's original estimator of λ as $\hat{\lambda}_1 = 2f_2/f_1$ as empirical Bayes estimator for observation $x = 1$ is useful, since it helps to find ways to eliminate the overestimation bias. We need to define a Horvitz-Thompson estimator that takes into account the different counts $x = 1, 2, ..$ separately. This can be accomplished by defining

$$\hat{N}^* = \frac{f_1}{1-\exp(-\hat{\lambda}_1)} + \frac{f_2}{1-\exp(-\hat{\lambda}_2)} + ... + \frac{f_m}{1-\exp(-\hat{\lambda}_m)}. \tag{6}$$

The question arises as to which way the estimator $\hat{\lambda}_x$ should be constructed. A naive estimator would follow the Robbins-type estimation to arrive at

$$\hat{N}_R = \frac{f_1}{1-\exp(-2f_2/f_1)} + \frac{f_2}{1-\exp(-3f_3/f_2)} + ... + \frac{f_{m-1}}{1-\exp(-mf_m/f_{m-1})} + f_m, \tag{7}$$

where we define

$$\frac{f_j}{1-\exp(-(j+1)f_{j+1}/f_j)} = \begin{cases} 0, & \text{if } f_j = 0; \\ f_j, & \text{if } f_{j+1} = 0. \end{cases}$$

Although the estimator (7) is intuitively attractive, it has some considerable difficulties. Not only is it unclear what to do with the largest count m (in (7) it is not up-weighted), but also various counts could have frequencies zero which would leave some of the frequencies f_x unweighted. More importantly, most of the observed count data will lie on the lower counts resulting in highly unstable estimates for larger counts.

It is more attractive to consider a *smoothed* version of the Bayes estimator. This can be accomplished by constructing an estimate of the marginal distribution $p_x = \int_0^\infty p(x|\lambda)q(\lambda)d\lambda$ using a discrete, finite mixture

$$p_x = \sum_{j=1}^{k} Po(x|\lambda_j)q_j,$$

where $\lambda_j > 0$ and the non-negative weights q_j sum up to 1. Estimates can be constructed by means of the EM algorithm or using some gradient-type algorithm. For details see Böhning and Kuhnert (2006). Some attention needs to be given to the question of the number of components k. Two strategies will be looked at:

- The number of components is determined by the nonparametric maximum likelihood estimator (NPMLE).
- The mixture model is chosen on the basis of the Bayesian Information Criterion (BIC) defined as $-\log L(Q) + (2k-1)\log(n)$.

In both cases we arrive at some estimate of the marginal distribution

$$\hat{p}_x = \sum_{j=1}^{k} Po(x|\hat{\lambda}_j)\hat{q}_j \tag{8}$$

leading to smoothed estimates of the population size

$$\hat{N} = \sum_{\ell=1}^{m} \frac{f_\ell}{1 - \exp(-(\ell+1)\frac{\hat{p}_{\ell+1)}}{\hat{p}_\ell})}, \tag{9}$$

where we attach a subscript NPMLE to the population size estimate N_{NPMLE} in (9) if the first strategy is used and we use the notation N_{BIC} if the second strategy is used.

We will also consider two further ways of estimating the mixing distribution $q(\lambda)$ in $\int_0^\infty Po(x|\lambda)q(\lambda)d\lambda$. The first estimator is based upon the idea of using the empirical distribution itself as an estimator of the mixing distribution. To accomplish this task we have to consider the appropriate transformation of the observed frequencies. Let $\tilde{q}_i = f_i/n$ denote the relative frequencies of the observed, zero-truncated sample. According to Böhning and Kuhnert [1] the associated relative proportions of the zero-truncated mixture are given as

$$\hat{q}_i = \frac{\tilde{q}_i/[1 - Po(0|x_i)]}{\sum_{\ell=1}^{n} \tilde{q}_i/[1 - Po(0|x_\ell)]},$$

so that $\hat{p}_x = \sum_{j=1}^{m} Po(x|x_j)\hat{q}_j$ and

$$\hat{N}_{\text{EDF}} = \sum_{\ell=1}^{m} \frac{f_\ell}{1 - \exp(-(\ell+1)\frac{\hat{p}_{\ell+1)}}{\hat{p}_\ell})},$$

where the index EDF associates with the empirical distribution function. The benefit of this approach is that the estimate of the mixing distribution is readily available without any computational effort. The second additional estimator is building upon the Γ-distribution for $q(\lambda)$ in $p_x = \int_0^\infty Po(x|\lambda)q(\lambda)d\lambda$, namely $q(\lambda) = \theta^r \lambda^{r-1} e^{-\theta\lambda}/\Gamma(r)$ with shape parameter $r > 0$ and rate parameter $\theta > 0$. The parameters r and θ can be estimated by a weighted least squares estimator as suggested in Rocchetti *et al.* (2010). The associated population size estimator is denoted by $\hat{N}_G$.

In the following we continue the simulation study and provide evidence that the suggested empirical Bayes estimator performs better than the conventionally used estimators $\hat{N}_C$ and, in particular, $\hat{N}_Z$. Besides these conventional two estimators we will consider the nonparametric estimator $\hat{N}_R$ and the smoothed mixture model version $\hat{N}_{\text{BIC}}$. More details are available in supplementary material which also studies the estimators $\hat{N}_{\text{EDF}}$ and $\hat{N}_G$ which we have excluded here since their performance is less satisfactory than $\hat{N}_R$ and $\hat{N}_{\text{BIC}}$. The design of the simulation corresponds to the one used previously. Samples of counts $X_1, ...X_N$ were drawn from a two-component mixture of Poisson densities: $X \sim 0.5Po(1) + 0.5Po(\lambda)$, evidently with equal weights $q_1 = q_2 = 0.5$. The population size was set to $N = 100$ and $1,000$ replications used. Here, we will concentrate on the main findings. More details are available in the supplementary material Böhning *et al.* (2010). We see from Table 2 that both empirical Bayes estimators perform better with respect to their standard error and root mean square error than the other estimators adjusting for heterogeneity. If we compare the two empirical Bayes estimators it appears that the one based upon the nonparametric mixture model as smaller variance which is reflected also in a better mean squared error.

5 Application to spatial analysis of scrapie in Great Britain

Following the results of the previous section we will concentrate on using the NPMLE of the mixing distribution as the smoothed empirical Bayes estimate of the prior distribution for further analysis, in particular

$$\hat{p}_x = \sum_{j=1}^{k} Po(x|\hat{\lambda}_j)\hat{q}_j, \tag{10}$$

as derived in (8). In a first step, this will be done using the entire SND data, unstratified by county. Once an estimate for the mixing distribution has been achieved, a smoothed *county-specific* estimate of the population size can be developed as follows:

$$\hat{N}_{\text{NPMLE},i} = \sum_{\ell=1}^{m} \frac{f_{\ell,i}}{1 - \exp(-(\ell+1)\frac{\hat{p}_{\ell+1)}}{\hat{p}_\ell})}, \tag{11}$$

where $f_{\ell,i}$ is the frequency of holdings with ℓ cases in the $i-$th county and $\hat{p}_\ell$ is taken from (10).

5.1 Determining the NPMLE for the SND data

We have seen in section 4 using the ratio plot that there is strong evidence for heterogeneity captured by a mixing distribution. We consider the marginal distribution over all counties as available from Table 1: $f_1 = 298$, $f_2 = 89$, $f_3 = 42$,..., $f_{29} = 2$. We are using this (truncated) count distribution to determine the maximum likelihood estimators for the various possible mixture models. The results are provided in Table 3. For each number of components k, starting with the homogeneous case $k = 1$, the estimated mixture model $\hat{Q}$ is provided, the Poisson parameters $\hat{\lambda}_j$ and associated component weights $\hat{q}_j$. This is followed by the log-likelihood $\log L(\hat{Q})$ and the BIC-value $-2 \log L(\hat{Q}) + (2k-1) \log(n)$. Note that there are two estimates of the population size of scrapie-affected holdings given. One is based upon the direct computation using the mixture model estimated as provided in (3), the other is the empirical Bayes estimate using the estimated mixture as prior distribution (10). It is evident from columns 6 and and 7 in Table 3, that the empirical Bayes estimate of the population size is less sensitive to the choice of the number of components. Furthermore, the empirical Bayes estimates is not prone to spurious estimates as is the conventional mixture model based estimator. We have already mentioned that Figure 1 supports that there is considerable evidence for a monotone increasing pattern. In addition, the estimate of the posterior mean based upon the estimated mixture model with 4 components (this is what the BIC suggests) shows that this monotone pattern is met. Note that columns 6 and and 7 in Table 3 contain also (in brackets) an estimate of the standard error of the repsective population size estimate. This was achieved by applying the nonparametric bootstrap as adapted to capture–recapture situations by van der Heijden *et al.* (2003) and Böhning (2008). It is evident from columns 7 in Table 3 that the conventional mixture model based estimator is prone to extreme variance inflation when the number of components become large.

5.2 Estimating the number of hidden scrapie-affected holdings per county

We now apply these results to the individual counties using (11). Note that we are using the same mixture distribution in (11) estimated from the entire SND data. This is necessary since the county specific case distributions are frequently very sparse. Take for example county 1 in Table 1: we find $f_{1,1} = 2$, $f_{2,1} = 1$, $f_{3,1} = 1$, so $n_1 = 4$. It is clear that a reliable estimation of a mixing distribution is not possible from this count distribution. Hence we use the mixing distribution estimated from the entire data set and assume that the heterogeneity found for the entire data set is also valid in each county. Then we compute the *predicted* number of scrapie-affected holdings by applying the weight $(1 - \exp[-(\ell+1)\frac{\hat{p}_{\ell+1)}}{\hat{p}_\ell}])^{-1}$ to the frequency $f_{\ell,i}$ of count ℓ in the $i-$th county and summing up over all observed frequencies $f_{\ell,i}$ leading to

$$\hat{N}_i = \sum_{\ell=1}^{m} \frac{f_{\ell,i}}{1 - \exp(-(\ell+1)\frac{\hat{p}_{\ell+1)}}{\hat{p}_\ell})}.$$

This process is very similar to indirect standardization used in epidemiologic methodology (see Waller and Gotway (2004, p. 17). The results are provided in Table 4. In addition, two further measures are computed. The *observed–hidden ratio* defined as $n_i/(\hat{N}_i - n_i)$ and the *completeness* measure defined as $n_i/\hat{N}_i$, provided as columns 4 and 5 in Table 4. The completeness ranges between 48% and 99%. Figure 2 shows a scatterplot of the completeness against the observed count (on log-scale) of scrapie-affected holdings. There is no evidence for a specific pattern, though the variation of completeness seems to decrease with increasing observed count of scrapie-affected holdings. Median observed–hidden ratio is 1.29 with 95% nonparametric CI (1.11, 1.43) and completeness is 56.36 with 95% nonparametric CI (52.62%, 58.83%).

Figure 3 shows the geographical distribution of county–specific completeness and observed–hidden ratios. Completeness is fairly stable with most counties in the 50-59% category and fewer counties in the upper completeness categories. Note that as well as providing completeness and observed/hidden ratios, we can also estimate adjusted measures of disease occurrence for each county. However, for our particular case, this would not have a clear biological interpretation as annual data was pooled to increase the power of our analyses.

6 Discussion

As described in section four and five, providing theoretical evidence and empirical support respectively, $\hat{N}_i = \hat{N}_{BIC,i}$ represents a lower bound of the population size in each county i. Hence, the estimated completeness $n_i/\hat{N}_i$ in county i will be an upper bound for n_i/N_i, so that the estimated values for completeness will be too large on average. Consequently, since the observed values already have an upper limit of almost 100%, it is expected that only the observed minimum for completeness of 48% will be in fact a bit lower. Similarly, we expect that the observed-hidden ratios are overestimated. Typically, we have seen in the simulation study that N is underestimated by $5-10\%$, never more than 20%.

The maps are based upon an estimated size of the scrapie population in county i, given as

$$\hat{N}_i = \sum_{\ell=1}^{m} \frac{f_{\ell,i}}{1-\exp(-(\ell+1)\frac{\hat{p}_{\ell+1)}}{\hat{p}_\ell})} = \sum_{\ell=1}^{m} \hat{w}_\ell f_{\ell,i},$$

where $\hat{p}_\ell$ is found from (10) with an estimated BIC-selected nonparametric mixing distribution. Since the estimated weights $\hat{w}_\ell = 1/[1-\exp(-(\ell+1)\frac{\hat{p}_{\ell+1)}}{\hat{p}_\ell})]$ do not depend on the county index i we have that

$$\sum \hat{N}_i = \sum_i \sum_{\ell=1}^{m} \hat{w}_\ell f_{\ell,i} = \sum_{\ell=1}^{m} \hat{w}_\ell \sum_i \hat{f}_{\ell,i} = \sum_{\ell=1}^{m} \hat{w}_\ell f_\ell = \hat{N},$$

where $f_\ell = \sum_i f_{\ell,i}$, so that the margin (over counties) of the county-specific estimates of the size of the scrapie population and the estimate of the size of the scrapie population, unstratified by county, coincide.

Finally, note that it is also possible to derive estimates for the standard errors of $\hat{N}_i = \sum_{\ell=1}^m \hat{w}_\ell f_{\ell,i} = \hat{\mathbf{w}}^T \mathbf{f}_i$. The variance conditional upon $\hat{\mathbf{w}}$ is simply $\hat{\mathbf{w}}^T Cov(\mathbf{f}_i)\hat{\mathbf{w}}$ with $Cov(\mathbf{f}_i) = \Lambda_{\mathbf{f_i}} - \mathbf{f}_i\mathbf{f}_i^T/n_i$, where $n_i = \sum_{\ell=1}^m f_{\ell,i}$ and $\Lambda_{\mathbf{f_i}}$ the diagonal matrix with elements $f_{\ell,i}$. $\ell = 1, ..., m$, on the diagonal. This variance estimate is dependent on the vector $\mathbf{f}_i$ and will be different for each county, but it is conditional upon $\hat{w}_\ell = \frac{1}{1-\exp(-(\ell+1)\frac{\hat{p}_{\ell+1)}}{\hat{p}_\ell})}$ for $\ell = 1, ..., m$ which is identical for each county. Although a conditional variance estimate seems appropriate for comparison of variation within the county strata, it might be sometimes desirable to provide an unconditional variance estimate. This can be achieved by adding an additional variance component due to the random error involved in the estimate $\hat{\mathbf{w}}$ (for more details on variance computations in the capture–recapture setting see Böhning (2008)), so that the unconditional variance estimate becomes

$$\widehat{Var(\hat{N}_i)} = \hat{\mathbf{w}}^T Cov(\mathbf{f}_i)\hat{\mathbf{w}} + \mathbf{f}_i^T Cov(\hat{\mathbf{w}})\hat{\mathbf{f}}_i,$$

where $Cov(\hat{\mathbf{w}})$ is the covariance matrix for the vector $\mathbf{w}$. This needs to be determined only once for the entire data set, but will depend on the estimator used to estimate $\hat{p}_\ell$ in $\hat{w}_\ell = \frac{1}{1-\exp(-(\ell+1)\frac{\hat{p}_{\ell+1)}}{\hat{p}_\ell})}$ and it is best done using the nonparametric bootstrap mentioned in section 4.

In conclusion, we would like to point out that the technique developed here to estimate county-specific population sizes of scrapie can be used for any stratified situation where there is interest in providing stratum-specific estimates of population size and the data per stratum are potentially sparse. Examples of such strata could be the laboratories involved in determining the disease or particular time windows of interest. Hence the suggested techniques has a general characteristic and is by no means limited to geographical applications.

References

[1] Böhning, D. and Kuhnert, R. (2006). The Equivalence of Truncated Count Mixture Distributions and Mixtures of Truncated Count Distributions. *Biometrics* **62**, 1207-1215.

[2] Böhning, D., Del Rio Vilas, V.J. (2008). Estimating the hidden number of scrapie affected holdings in Great Britain using a simple, truncated count model allowing for heterogeneity. *Journal of Agricultural, Biological and Environmental Statistics* **13**, 1–22.

[3] Böhning, D. (2008). A simple variance formula for population size estimators by conditioning. *Statistical Methodology* **5**, 410–423.

[4] Böhning, D., Kuhnert, R. and Del Rio Vilas, V.J. (2010). Supplementary material to *Population size estimation based upon ratios of recapture probabilities.*

[5] Bunge, J. and Fitzpatrick, M. (1993). Estimating the Number of Species: a Review. *Journal of the American Statistical Association* **88**, 364–373.

[6] Carlin, B. P. and Louis, T. A. (1997). Bayes and Empirical Bayes Methods for Data Analysis. *Monographs on Statistics and Applied Probability, London, Chapman* & *Hall.*

[7] Chao, A. (1987). Estimating the Population Size for Capture-Recapture data with Unequal Catchability. *Biometrics* **43**, 783–791.

[8] Chao, A. (1989). Estimating Population Size for Sparse Data in Capture-Recapture Experiments. *Biometrics* **45**, 427–438.

[9] Chao A., Tsay P.K., Lin S.H, Shau W.Y, Chao D.Y. (2001). Tutorial in Biostatistics: The Applications of capture-recapture models to epidemiological data. *Statistics in Medicine* **20**, 3123–3157.

[10] Dawson M., Del Rio Vilas V.J. (2008) The control of classical scrapie in sheep in Great Britain. *In Practice* **30**, 330–333.

[11] Del Rio Vilas, V.J., Sayers, R., Sivam, K., Pfeiffer, D.U., Guitian, J., Wilesmith, J.W. (2005). A case study of capture-recapture methodology using scrapie surveillance data in Great Britain. *Preventive Veterinary Medicine* **67**, 303–317.

[12] Del Rio Vilas, V.J., Guitian, J., Pfeiffer, D.U., Wilesmith, J.W. (2006). Analysis of data from the passive surveillance of scrapie in Great Britain between 1993 and 2002. *Veterinary Record* **159**, 799–804.

[13] Dorazio, R.M. and Royle, J.A. (2005). Mixture models for estimating the size of a closed population when capture rates vary among individuals. *Biometrics* **59**, 351–364.

[14] Hay, G. and Smit, F. (2003). Estimating the number of drug injectors from needle exchange data. *Addiction Research and Theory*, **11** 235–243.

[15] Van Hest, N.A.H., De Vries, G., Smit, F., Grant, A.D., and Richardus, J.H. (2008). Estimating the Coverage of Tuberculosis Screening among Drug Users and Homeless Persons with Truncated Models. *Epidemiology and Infection* **136**, 14–22..

[16] Van der Heijden, P. G. M., Cruyff, M., van Houwelingen, H. C. (2003). Estimating the Size of a Criminal Population from Police Records Using the Truncated Poisson Regression Model. *Statistica Neerlandica* **57**, 1–16.

[17] Van der Heijden, P. G. M., Van Putten, W., Van Rongen, R. (2006). A Comparison of Zelterman's and Chao's Estimators for the Size of an Unknown Population by Capture-Recapture Frequency Data. Personnel Communication with P.v.d. Heijden.

[18] Hoinville, L.J., Hoek, A.R., Gravenor, M.B., McLean, A.R. (2000). Descriptive epidemiology of scrapie in Great Britain: results of a postal survey. *Veterinary Record* **146**, 455–461.

[19] Holzmann, H., Munk, A., and Zucchini, W. (2003). On identifiability in capture-recapture models. *Biometrics* **62**, 934–939.

[20] Link, W.A. (2003). Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics* **59**, 1123–1130.

[21] Link, W.A. (2003). Response to a paper by Holzmann, Munk and Zucchini. *Biometrics* **62**, 936–939.

[22] McKendrick, A.G. (1926): Application of Mathematics to Medical Problems. *Proceedings of the Edinburgh Mathematical Society* **44**, 98-130.

[23] Mosley, W.H., Bart, K.J., and Sommer, A. (1972). An epidemiological assessment of cholera control programs in rural East Pakistan. *International Journal of Epidemiology* **1**, 5–11.

[24] Pledger, S. A. (2005). The performance of mixture models in heterogeneous closed population capture-recapture. *Biometrics* **61**, 868–876.

[25] Roberts, J.M. and Brewer, D.D. (2006). Estimating the prevalence of male clients of prostitute women in Vancouver with a simple capture-recapture method. *Journal of the Royal Statistical Society (Series A)* **169**, 745–756.

[26] Rocchetti, I., Bunge, J. and Böhning, D. (2010). Population size estimation based upon ratios of recapture probabilities. *submitted for publication.*

[27] Robbins, H. (1955). An empirical Bayes approach to statistics. *In Proc.3rd Berkeley Symp. on Math Statist. and Prob.,1, Berkeley, CA: University of California Press*, 157–164.

[28] Waller, L.A., Gotway, C.A. (2004). *Applied Spatial Statistics for Public Health Data.* Hoboken, NJ, Wiley.

[29] Wang, J.-P. and Lindsay, B.G. (2005). A penalized nonparametric maximum likelihood approach to species richness estimation. *Journal of the American Statistical Association* **100**, 942–959.

[30] Wang, J.-P. and Lindsay, B.G. (2008). An exponential partial prior fro improving nonparametric maximum likelihood estimation in mixture models. *Statistical Methodology* **5**, 30–45.

[31] Wilson, R.M. and Collins, M.F. (1992). Capture-recapture estimation with samples of size one using frequency data. *Biometrika* **79**, 543–553.

[32] Zelterman, D. (1988). Robust estimation in truncated discrete distributions with applications to capture-recapture experiments. *Journal of Statistical Planning and Inference* **18**, 225–237.

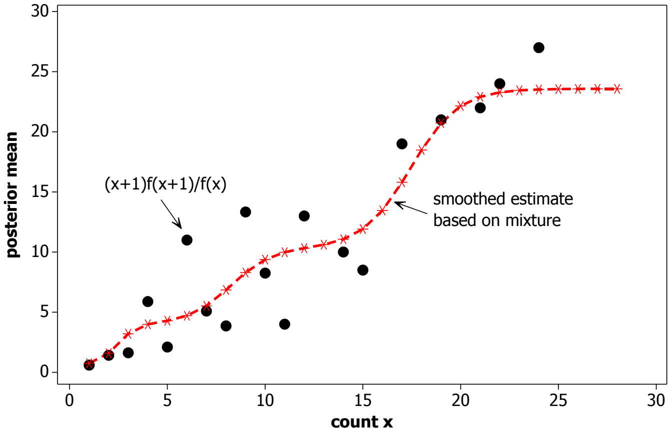


Figure 1: *Ratio plot for SND data 2002-2006, unstratified by county, for Robbins estimate of posterior mean as well as the discrete mixture (4 components) based empirical Bayes estimate of the posterior mean*

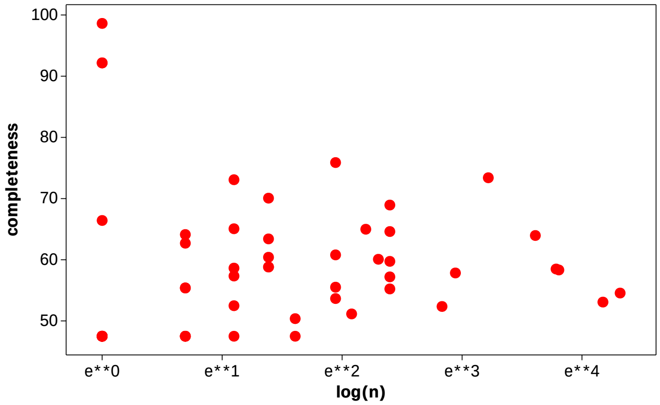


Figure 2: *Scatterplot of completeness of surveillance stream per county against observed count of scrapie affected holdings per county*

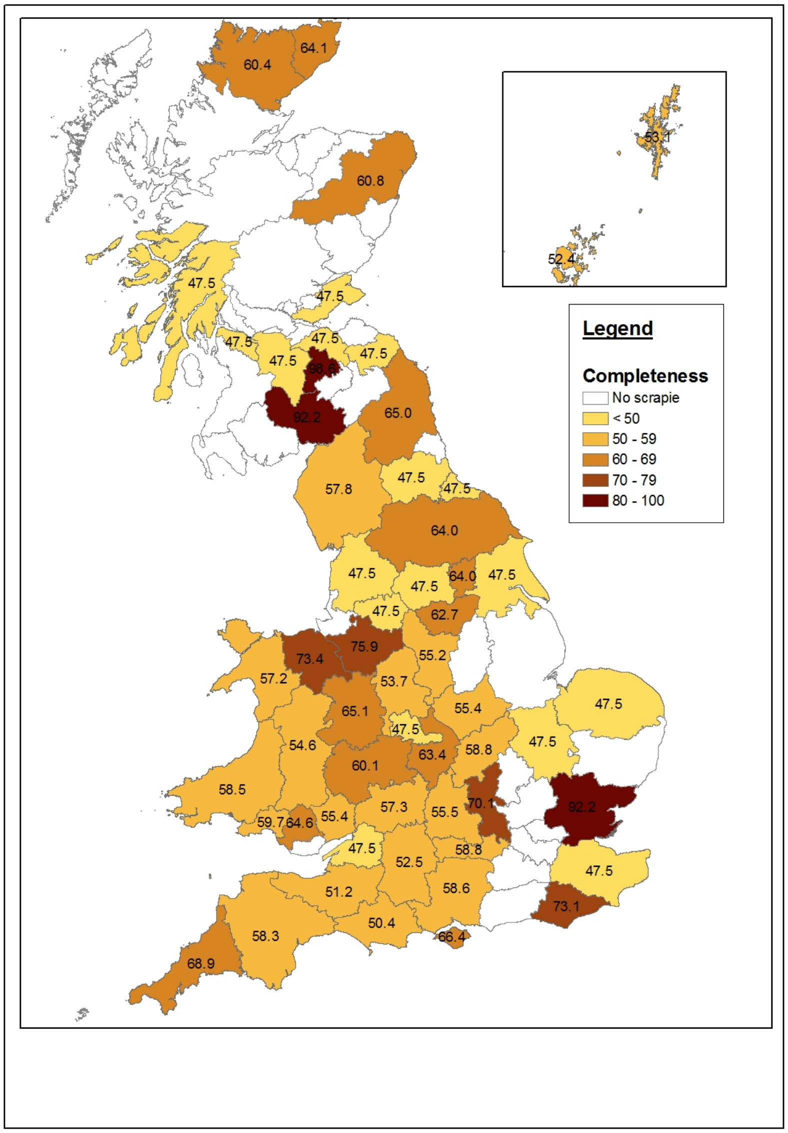


Figure 3: *Map of estimated completeness on county level for SND data 2002-2006*

Table 1: *Distribution of confirmed scrapie–affected holdings from the SND database 2002–2006 by county*

county	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10+}	n
1	2	1	1	0	0	0	0	0	0	0	4
2	1	1	1	0	1	0	0	0	0	0	4
3	1	0	0	0	0	0	0	0	0	0	1
4	1	0	0	0	0	0	0	0	0	0	1
5	2	0	0	0	1	0	0	1	0	3	7
6	4	1	0	1	0	1	0	1	0	3	11
7	12	1	0	2	3	0	0	0	0	1	19
8	7	2	2	0	0	0	0	0	0	0	11
9	25	8	5	1	1	1	2	0	0	2	45
10	4	1	0	0	0	0	0	0	0	0	5
11	1	0	0	0	0	0	0	0	0	0	1
12	0	0	1	0	0	0	0	0	0	0	1
13	2	0	0	1	0	0	0	0	0	0	3
14	1	2	0	0	0	0	0	0	0	0	3
15	0	1	0	0	0	0	0	0	0	0	1
16	5	2	1	1	0	0	0	1	0	0	10
17	1	0	0	0	0	0	0	0	0	0	1
18	5	0	0	0	0	0	0	0	0	0	5
19	1	1	0	0	0	0	0	0	0	0	2
20	1	0	0	0	0	0	0	0	0	0	1
21	2	1	1	0	0	0	0	0	0	0	4
22	3	3	0	0	1	0	1	0	1	0	9
23	5	0	1	0	0	0	0	0	0	1	7
24	2	0	0	0	0	0	0	0	0	0	2
25	1	1	0	0	0	0	0	0	1	0	3
26	6	2	0	0	0	0	0	0	0	0	8
27	5	1	0	0	1	0	0	0	0	0	7
28	1	0	0	0	0	0	0	0	0	2	3
29	2	0	1	0	1	0	0	0	0	0	4
30	1	0	0	0	0	0	0	0	0	0	1
31	2	1	0	0	0	0	0	0	0	0	3
32	1	0	0	0	0	0	0	0	0	0	1
33	1	0	1	0	0	0	0	0	0	0	2
34	14	10	3	1	3	0	2	1	0	3	37
35	2	0	0	0	0	0	0	0	0	0	2

... continued on next page ...

... continued from previous page ...

county	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10+}	n
36	2	0	0	0	0	0	0	0	0	0	2
37	51	11	5	5	0	1	1	1	0	0	75
38	6	3	1	0	0	0	0	0	1	0	11
39	24	9	1	1	3	1	2	1	0	2	44
40	6	4	4	1	2	1	2	1	0	4	25
41	3	5	2	0	1	0	0	0	0	0	11
42	6	1	3	1	0	0	0	0	0	0	11
43	1	1	0	0	0	0	0	0	0	0	2
44	4	0	1	0	0	1	0	0	0	1	7
45	1	0	0	0	0	0	0	0	0	0	1
46	3	0	0	0	0	0	0	0	0	0	3
47	1	0	0	0	1	0	0	0	0	0	2
48	0	0	1	0	0	0	0	0	0	0	1
49	1	0	0	0	0	0	0	0	0	0	1
50	2	0	0	0	0	0	0	0	0	0	2
51	1	0	0	0	0	0	0	0	0	0	1
52	13	2	1	0	0	0	1	0	0	0	17
53	0	0	0	0	1	0	0	0	0	0	1
54	1	0	0	0	0	0	0	0	0	0	1
55	47	10	5	2	0	1	0	0	0	0	65
56	1	3	0	0	0	0	0	0	0	0	4
All	298	89	42	17	20	7	11	7	3	22	516

Table 2: *Simulation using $X \sim 0.5Po(1) + 0.5Po(\lambda)$ and $N = 100$; provided are estimates of $E(\hat{N})$, $Var(\hat{N})^{1/2}$ and $[E(\hat{N} - N)^2]^{1/2}$ as mean, SD and RMSE*

λ	estimator	mean	SD	RMSE
1	MLE-hom	102	13	13
	NPMLE	484	12,098	20,028
	Chao	104	19	19
	Zelterman	105	21	22
	EB-NPMLE	105	15	15
	EB-Robbins	108	21	22
2	MLE-hom	94	7	9
	NPMLE	4599	35	21,328
	Chao	99	12	12
	Zelterman	101	16	16
	EB-NPMLE	98	8	9
	EB-Robbins	102	12	12
3	MLE-hom	88	5	13
	NPMLE	12,517	52,425	23,955
	Chao	97	10	11
	Zelterman	102	15	16
	EB-NPMLE	93	7	10
	EB-Robbins	96	9	10
4	MLE-hom	85	4	16
	NPMLE	11,715	54,501	23,114
	Chao	97	10	10
	Zelterman	108	20	20
	EB-NPMLE	92	7	11
	EB-Robbins	95	9	10
5	MLE-hom	84	4	17
	NPMLE	4,657	33,069	17,373
	Chao	98	10	1017
	Zelterman	115	23	27
	EB-NPMLE	92	8	11
	EB-Robbins	95	9	10

Table 3: *Estimated mixture models for 1, 2, 3 , 4 and 5 (NPMLE) number of components with associated estimator of the size of the scrapie–affected population of holding from the unstratified SND database 2002–2006*

					discrete mixture model based	
k	$\hat{\lambda}_j$	$\hat{q}_j$	$\log L(\hat{Q})$	BIC	$\hat{N}_{\text{NPMLE}}$ (10), (SE)	$\hat{N}_{\text{NPMLE}}$ (3), (SE)
1	2.33	1.00	-1,279.0	2,561.4	572 (9.4)	572 (9.4)
2	0.97 9.80	0.88 0.12	-865.4	1,740.8	776 (32.4)	793 (34.6)
3	0.67 5.46 19.10	0.80 0.17 0.03	-807.8	1,632.4	869 (44.8)	946 (65.8)
4	0.56 4.03 10.35 23.58	0.75 0.19 0.05 0.01	-802.3	1,628.2	896 (48.0)	1,036 (60,102)
5	0.01 1.08 5.13 11.76 23.98	0.27 0.54 0.14 0.03 0.01	-801.2	1,632.7	916 (25.5)	528,694 (419,663)

Table 4: *Observed and hidden scrapie-affected counts of holdings by county, observed–hidden ratio and completeness of surveillance stream*

county	n	$\hat{N}$	o/h	completeness
1	4	7	1.4	59
2	4	6	2.3	70
3	1	2	0.9	48
4	1	2	0.9	48
5	7	9	3.1	76
6	11	16	2.2	69
7	19	33	1.4	58
8	11	20	1.2	55
9	45	77	1.4	58
10	5	10	1.0	50
11	1	2	0.9	48
12	1	1	11.8	92
13	3	5	1.3	57
14	3	5	1.4	59
15	1	2	2.0	66
16	10	17	1.5	60
17	1	2	0.9	48
18	5	11	0.9	48
19	2	4	1.2	55
20	1	2	0.9	48
21	4	7	1.4	59
22	9	14	1.9	65
23	7	13	1.2	56
24	2	4	0.9	48
25	3	5	1.9	65
26	8	16	1.0	51
27	7	13	1.2	54
28	3	4	2.7	73
29	4	6	1.7	63
30	1	2	0.9	48
31	3	6	1.1	52
32	1	2	0.9	48
33	2	3	1.7	63
34	37	58	1.8	64
35	2	4	0.9	48
36	2	4	0.9	48

... continued on next page ...

county	n	$\hat{N}$	o/h	completeness
... continued from previous page ...				
37	75	137	1.2	55
38	11	19	1.3	57
39	44	75	1.4	58
40	25	34	2.8	73
41	11	17	1.8	65
42	11	18	1.5	60
43	2	4	1.2	55
44	7	12	1.6	61
45	1	2	0.9	48
46	3	6	0.9	48
47	2	3	1.8	64
48	1	1	11.8	92
49	1	2	0.9	48
50	2	4	0.9	48
51	1	2	0.9	48
52	17	32	1.1	52
53	1	1	71.6	99
54	1	2	0.9	48
55	65	122	1.1	53
56	4	7	1.5	60