A method for estimating insect abundance and patch occupancy with potential applications in large-scale monitoring programmes

G. Sileshi*

World Agroforestry Centre (ICRAF), SADC-ICRAF Agroforestry Programme, Chitedze Agricultural Research Station, P.O. Box 30798, Lilongwe, Malawi

Large-scale monitoring programmes often make inferences about insect abundance based on count data collected using some probability-based sampling technique. Unfortunately, it is rather difficult to obtain reliable estimates of insect abundance from count data unless the scale is very fine or localized. A major issue, that has to be explicitly addressed when estimating insect abundance, is the problem of false negatives. The objective of this paper is to demonstrate a simple approach to estimate insect abundance from occupancy data collected using presence–absence surveys. Using count data on the seed-feeding insect, *Eurytoma* sp. (Hymenoptera: Eurytomidae), and the alien invasive insect, *Heteropsylla cubana* (Homoptera: Psyllidae), this paper has demonstrated the application of (1) generalized linear models for modelling abundance and detection probability, (2) information criteria for model selection and (3) occupancy-abundance models for precise estimation of insect abundance. Potential applications of this approach in monitoring colonization of sites by alien invasive species and local extinction of species endangered by habitat fragmentation are also indicated.

Key words: alien invasive species, *Eurytoma, Heteropsylla cubana,* maximum likelihood.

INTRODUCTION

The number of individuals, or the abundance, of a species is a fundamental ecological parameter (Andrewartha & Birch 1954) and a critical consideration when making management and conservation decisions (Sileshi 2007). Most monitoring programmes use counts of insects as proxies of true abundance, and there are many kinds and levels of decisions that need to be made based on insect abundance (Sileshi 2006, 2007). An immediate decision might be to spray a field or not; to reject a plant shipment or not at a port of entry; and to declare an alien species, whether introduced insect biocontrol agent or pest as established or invasive. Other decisions may include targeting hot-spots of abundance for conservation of endangered species.

A key interest in large-scale monitoring of insect species lies in detecting spatial and temporal changes in abundance. However, large-scale monitoring studies often make inferences about large areas by collecting information from sample units selected by some probability-based sampling technique. Unfortunately, it is rather more difficult to obtain reliable estimates of abundance of a species unless the scale is very fine or localized (Speight *et al.* 1999; He & Gaston 2003). First, data resulting from simple count surveys can be biased to an unknown degree by heterogeneous and imperfect detection (Pollock *et al.* 2002). The second problem associated with monitoring small and often abundant insects such as psyllids, mites and aphids is the time required to process sufficient sampling units to obtain a reliable estimate of abundance (Sileshi 2006; 2007).

One potential approach to reducing effort in large-scale monitoring programmes involves a shift of interest from abundance to patch occupancy (probability of occurrence) (Freckleton *et al.* 2005; Gaston *et al.* 2000; He & Gaston 2003; Mackenzie & Nichols 2004; Royle & Nichols 2003). Patch occupancy is usually estimated using presence–absence data collected as part of biological surveys, ecological monitoring or pest management programmes. The results of such surveys are used to assess the efficacy of management actions, to look for species declines or reductions in range, or to model the habitat of a species (Tyre *et al.* 2003). Presence–absence sampling is a shortcut method that saves time, in particular, for small insects that occur at high densities and where the population has to be sampled several times during a season

*Postal address: POST DOT NET, P.O. Box X389, Cross Roads, Lilongwe, Malawi.
E-mail: sileshi@africa-online.net or sgwelde@yahoo.com

(Wilson & Room 1983; Nachman 1981). The method is non-destructive and thus interferes relatively little with the population being studied, and avoids long-term changes in the habitat or the insect population.

Estimation of occupancy rates and associated dynamics including extinction and colonization from presence–absence data is fundamental to many habitat models (Cabeza *et al.* 2004), metapopulation studies (Hanski & Gilpin 1997) and monitoring efforts (Joseph *et al.* 2006; Mackenzie & Nichols 2004). Interest in patch occupancy, in a population monitoring context (Joseph *et al.* 2006), has been motivated by the observation that the proportion of areas occupied by a species increases with its abundance among those areas, and again that this is manifest from micro- to macro-spatial scales both for a given species at different times or in different regions (Gaston *et al.* 2000; Kunin *et al.* 2000; He & Gaston 2003; Warren *et al.* 2003). Hence, occupancy can be used as a surrogate for abundance estimation (MacKenzie & Nichols 2004).

A significant aspect of the abundance–occupancy relationship is that it may be used to predict how the size of the total or regional population of a species changes with occupancy or local density (Gaston *et al.* 2000; Freckleton *et al.* 2005). It has been observed widely that the local abundance and regional distribution of species tend to be correlated positively (Hartley 1998), such that species with low abundance within sites (*i.e.* average numbers or densities of individuals) also tend to occupy fewer sites (*i.e.* the area or range of a species at a national or continental scale), while species with high abundance also tend to occupy a large number of sites (Gaston *et al.* 2000). Moreover, if local abundance is related to habitat quality, then the abundance–occupancy relationship can be used to predict how the total size of a population varies as a function of habitat quality.

Although occupancy-abundance models hold great potential for monitoring of invasive alien insects, conservation status of local species and performance of exotic biocontrol agents, it has not yet been formally applied in insect ecology. The problem has been that until recently no attempt has been made to incorporate detectability into occupancy-abundance relationships for insect species. Therefore, the objective of this paper is to demonstrate a simple approach to estimate insect abundance from occupancy data collected using presence–absence surveys. Using field data on the seed-feeding insect, *Eurytoma* sp. (Hymenoptera: Eurytomidae), and the alien invasive species, *Heteropsylla cubana* (Homoptera: Psyllidae), methods for valid estimation of abundance and occupancy parameters are demonstrated.

Table 1. Definition of parameters used in this paper.

Parameter	Definition
α	The intercept term in regression models
β	Parameter describing the influence of covariates in regression models
μ	Abundance parameter of the negative binomial distribution
λ	Abundance parameter of the Poisson. For the Poisson distribution $\lambda = \mu = \sigma^2$
k	Dispersion parameter of the negative binomial distribution
N_i	Abundance of a species in a set of spatial locations i
ψ	Occupancy (probability of occurrence)
p	Detection probability
θ	Number of parameters estimated for a given model
φ	Dispersion statistic; residual deviances divided by its degrees of freedom

METHODOLOGY

Characterizing abundance and occupancy

Abundance and occupancy surveys involve visiting sites multiple times within a season and examining relevant sampling units where the target species is either detected or not detected (MacKenzie *et al.* 2002; Royle & Nichols 2003; Royle *et al.* 2005). The goal is often to estimate parameters (Table 1), especially abundance, knowing the species is not always detected perfectly even when present (Bailey *et al.* 2004). Abundance data are collected by complete enumeration, counting all individuals, in well-defined sampling units. Occupancy data are collected by recording whether any individuals of interest are present or absent or more specifically detection or non-detection of the species. In practice, counts or some other ordinal measures of abundance may be observed, and such data can be reduced to the binary responses of presence or absence (Royle *et al.* 2005). In cases where animals are counted, the counts can be viewed as realizations of a binomial

random variable with index N_i (local abundance) and detection probability p (Royle & Nichols, 2003). At each visit, an effort has to be made to detect the insect species of interest, producing a detection history for each site. However, several design issues have to be considered, such as stratified random sampling, cluster sampling, random transects and multistage sampling (Leather & Watt 2004).

When estimating both abundance and occupancy, an issue that has to be explicitly addressed is the problem of false negatives or imperfect detection of a species (Gu & Swihart 2004; MacKenzie *et al.* 2003; Royle & Nichols 2003; Tyre *et al.* 2003). This problem arises when a visit to a site or examination of the chosen sampling unit fails to record the species of interest when it is in fact present. Failure to detect the species may be a consequence of either true absence or nondetection due to various factors (Gu & Swihart 2004). Detectability may vary among species, observer experience, survey methodology, and temporally due to seasonal behavioural patterns, or spatially due to site-specific habitat characteristic that affect abundance. Most ecological sampling methods may lead to false-negatives and underestimate abundances even of common insects (Speight *et al.* 1999) if they are mobile or concealed, *e.g.* soil-dwelling, seed-feeding and stem-boring. Even modest false-negative errors can have significant influence on ecological conclusions drawn from the data. The solution to this problem is to estimate the false-negative rate, and how it varies with habitat or other covariates. Several workers (MacKenzie *et al.* 2002; Gu & Swihart 2004; Royle & Nichols 2003; Royle *et al.* 2005; Tyre *et al.* 2003) have proposed modelling frameworks that provide for directly estimating the effect of factors that lead to variations in abundance and detection probability. This framework allows one to determine how abundance and occupancy change in response to environmental, habitat and other landscape characteristics.

For count data, the most obvious modelling approach is a Poisson (Lawless 1987; Cameron & Trivedi 1998) generalized linear model (GLM) explaining variation in the mean of local abundance (Ni). Poisson regressions involve explicitly modelling the distribution of counts assuming that the variance (σ^2) is proportional to the mean (λ); say $\sigma^2 = \phi\lambda$, where ϕ is a dispersion parameter (Cameron and Trivedi, 1998). The variance equals the mean when $\phi = 1$, while $\phi > 1$ indicates overdispersion in the Poisson model. In the GLMs, the dependent variable (N_i) is assumed to have a Poisson distribution with parameter λ which, in turn, depends on a vector of explanatory variables (or covariates X_i) as:

$$Log(\lambda) = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n \,, \quad (1)$$

where α is the intercept, X_i is the value of the n-th measurable covariate, and β_i is a parameter to be estimated for the n-th covariate.

For the Poisson model to apply, the assumption of randomness, mutual independence of individual insects, must exist. Hence, the Poisson regression model is too restrictive for count data (Cameron & Trivedi 1998; Saha & Paul 2005). The fundamental problem is that the distribution is parameterized in terms of a single scalar parameter (λ) so that all moments of y are functions of λ. In many insects a Poisson density predicts the probability of a zero count to be considerably less than is actually observed in a sample (Sileshi 2006, 2007). A second and more obvious way that the Poisson is deficient is that for count data the variance usually exceeds the mean (overdispersion), while the Poisson implies equality of variance of mean. Large overdispersion can lead to grossly deflated standard errors and inflated statistics in the usual maximum likelihood output (Cameron & Trivedi, 1998). One can use a conventional dispersion statistic to assess overdispersion and goodness-of-fit of the Poisson model.

The negative binomial distribution is more appealing in some instances because it allows the density of animals to vary spatially. It has also been shown to be more robust for modelling zero-inflated and overdipersed insect count data (Sileshi 2006). The negative binomial distribution is described by a mean parameters (μ) and dispersion parameter (k). The variance of the negative binomial distribution is equal to $\mu + k\mu^2$. According to Johnson & Kotz (1969) the NBD is a mixture of Poisson distributions such that the expected values of the Poisson distribution vary according to a gamma (Type III) distribution (Sileshi 2007). It has been shown that the limiting distribution of the NBD, as the dispersion parameter (k) approaches zero, is the Poisson. When k is an integer, the NBD becomes the Pascal distribution, and the geometric distribution corresponds to $k = 1$. The log series distribution occurs when zeros are missing and as $k \to \infty$ (Saha & Paul 2005). To relate trends in

abundance to the explanatory variables (X_i) equation 1was used. Accommodating trend as a parameter allows one to model it directly as a function of covariates, for instance, to test for temporal or habitat-specific differences. This approach allows one to directly incorporate temporal changes in abundance into the model as a parameter to be estimated.

In characterizing occupancy, the underlying distribution is assumed to be a binomial distribution. Here we are interested in models for detection probability (p), which can vary according to some of the values of explanatory variables (X_1, X_2, X_3, X_n). For such data, the standard method of estimation is a generalized linear model (GLM) (Sileshi 2007) implemented using a maximum likelihood procedure such as logistic regression. To relate the probability (p) of detecting occupancy to the n-th explanatory variables (X_i) we write a GLM of the following form:

$$Logit(p) = \alpha + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n \,, \qquad (2)$$

where the *logit* function is the canonical link for the binomial distribution, α is the intercept and β_i is a parameter describing the influence of covariate X_i.

Often data do not support only one model as clearly best for analysis (Burnham & Anderson, 2002). Therefore, there is always uncertainty about the operating model that has given rise to the observations because only a sample from the population is observed (Sileshi 2006). This raises the issue of comparing models to assess which of the models are adequate for the data and which one could be chosen as the basis for interpretation, prediction, or other subsequent use. Currently, there are two basic approaches to model selection: the classical generalized likelihood ratio test used for comparing nested models and the new approach based on information theoretic measures. Unlike likelihood ratio tests, information theoretic measures are more consistent and can be used in comparison of nested as well as non-nested models (Johnson & Omland 2004). Using information measures, one seeks a model that loses as little information as possible or a model that is nearest to the truth (Burnham & Anderson 2002).

Relating abundance with occupancy

A suite of empirical models, generally known as occupancy-abundance models, are widely employed to relate abundance with occupancy (Gaston *et al.* 2000). The fundamental relationship between abundance and occupancy has been recently elaborated by many workers (He & Gaston 2003; Royle *et al.* 2005). The simplest procedure relating abundance with occupancy may be derived by assuming that the individuals of the subject species are randomly and independently distributed in space (Wilson & Room 1983; Wright 1991). For this reason, the Poisson is a standard null model for the distribution of animals in many ecological studies (Royle *et al.* 2005). Under this assumption patch occupancy (ψ) and abundance (λ) can be predicted from the Poisson distribution (Wright 1991; Royle *et al.* 2005; Sileshi *et al.* 2006) as:

$$\psi = 1 - \exp^{-\lambda} \,, \qquad (3)$$

where the parameter $\lambda = \mu = \sigma^2$ (the variance) for the Poisson distribution.

For species that show aggregated spatial pattern, the model relating patch occupancy with abundance can be derived from the negative binomial distribution as (He & Gaston 2003; Royle *et al.* 2005; Sileshi 2006):

$$\psi = 1 - \left(1 + \frac{\mu}{k}\right)^{-k} \,, \qquad (4)$$

where k is a spatial aggregation parameter defined in the domain of $(-\infty, -\mu)$ and $(0, \infty)$. When $k < -\mu$, occupancy is derived from the positive binomial distribution that describes spatial regularity, and when $k > 0$, it is derived from the negative binomial distribution for spatial aggregation (He & Gaston 2003). There are different methods of estimating k (Saha & Paul 2005). However, extensions of the maximum likelihood method are probably more appropriate to the generalized linear regression situation (Lawless 1987; Saha & Paul 2005).

Application of the methods to Eurytoma *and psyllid data*

Eurytoma sp. (Hymenoptera: Eurytomidae) has been identified as one of the major insect limiting production of quality seeds of the multipurpose agroforestry tree, *Sesbania sesban* (Sileshi 2003). Though not tested yet as weed biocontrol agents, *Eurytoma* has the potential to limit seed production by *Sesbania* should it become invasive outside its native range (Sileshi 2003). Seed damage was monitored monthly in *Sesbania*-improved fallow fields and isolated natural stands during 1997–2000 and 2005. *Sesbania* pods were sampled once every month when they were available on the

plants. During each sampling occasion, 50 or more pods of different levels of maturity were collected randomly from 10 or more plants. The pods were then placed individually in Petri dishes for rearing of seed-feeding insects. Seeds were extracted from the pods and each seed carefully examined for exit holes and evidence of insect damage (Sileshi 2003). The number of pods and seeds per pod damaged by *Eurytoma* were recorded based on prior knowledge of their damage symptoms.

Since a single insect attacks a seed of *Sesbania,* the number of seeds damaged per pod was used as an index of abundance. The proportion of seeds damaged per pod was used as an index of occupancy. The effects of year, month of sampling and stand (either as pure fallow or isolated stand) on abundance of *Eurytoma* per pod were evaluated using a log-linear model as: $Log(\mu) = a + b_1Year + b_2Month + b_3Satnd$. Assuming either Poisson or negative binomial distributions, eight models were considered for *Eurytoma* abundance. The first model under either the Poisson (Poisson1) or negative binomial distribution (NBD1) assumption contained only the intercept, while models 2–8 contained the individual covariates or combinations of covariates (Table 2). Parameters (Table 1) of the various models were estimated using the GENMOD procedure of the SAS system (SAS 2003), which produces maximum likelihood estimates of the regression parameters for the Poisson and negative binomial distributions. The SAS codes used for each model are presented in Appendix 1. The residual deviances divided by their respective degrees of freedom (D/DF = φ) judge adequacy of a model. If the regression model is adequate, the expected value of φ will be close to unity. Otherwise, the validity of the model could be doubtful.

The count data were converted to the binary responses of detection (when count > 0) and non-detection (when count = 0). The effects of covariates on detection probability was also assessed using a logit-linear model as: $Logit(p) = a + b_1Year + b_2Month + b_3Satnd$. Assuming binomial error distribution of detection/non-detection, eight models were considered. The first model (Logistic 1) contained only the intercept, while models 2–8 contained the individual covariates or combinations of covariates (Table 2). Parameters, defined in Table 1, of the various models were estimated using the LOGISTIC procedure of the SAS (Appendix 2), which produces maximum likelihood estimates of the logistic regression parameters.

The leucaena psyllid, *Heteropsylla cubana* (Homoptera: Psyllidae), a pest of the tropical agroforestry tree *Leucaena leucocephalla* is probably most notable for its dramatic spread across continents (Hassan *et al.* 1994; Day 1999; Macdonald *et al.* 2003). Starting its journey from its native home in Central America in 1986, it reached Africa in 1991. Now it is found throughout most of Africa where *L. leucocephalla* is planted. As populations of the psyllid can increase rapidly creating an over-abundance in the tropics, estimating abundance even on a single tree with a reasonable degree of accuracy becomes almost an impossible task (Hassan *et al.,* 1994). The psyllid populations were monitored in April/May 2005 in four experiments established in 1991, 1992, 1997 and 1999 at different sites at Msekera (Sileshi 2007). These experiments have been described in detail in Sileshi (2007). In all the trials, trees were cut to a height of 30 cm above ground after three years of growth and allowed to coppices (re-sprout) in the subsequent years where the shoots were cut back to fertilize maize crops. A cluster of 10 adjacent stumps were selected in every replicate of each experiment, and the number of psyllids per shoot and number of infested shoots per stump were recorded. The sampling unit from one stump comprised of a randomly selected shoot with three fully expanded leaves. Here the shoot was defined as the growing point including the first unfurled leaf (Day 1999). The number of psyllids per shoot and proportion of infested shoots per stump constituted indices of local abundance and occupancy, respectively.

The potential effects of site on abundance of psyllids was assessed using a log-linear model as: $Log(\mu) = a + b_1 site$. After reducing the count data into detection/non-detection data, the effect of site on detection probability of psyllids was evaluated using a logit-linear model as $Logit(p) = a + b_1 site$. Alternative models were considered for both abundance and detection probability of psyllids. Model parameters for abundance and detection probability were estimated using SAS procedures as described for *Eurytoma.*

The alternative models were compared using the Akaike Information Criterion (AIC), which evaluates models based on their likelihood. The second-order Akaike Information Criterion (AIC_c) correcting for small sample size was used for comparing the models.

Table 2. Parameters, goodness-of-fit of abundance and occupancy models of *Eurytoma* and model selection based on second-order Akaike information criteria (AIC_c) and Akaike weights (AIC_w).

Variable	Model	Covariates in model	φ	λ, μ or p	k	AIC_c	AIC_w
Abundance	Poisson1	No covariate	6.1	8.3	–	–19325.0	0
	Poisson2	Year	5.4	10.5	–	–20027.6	0
	Poisson3	Month	5.3	7.2	–	–20155.4	0
	Poisson4	Stand	6.1	8.4	–	–19321.6	0
	Poisson5	Year Month	5.1	10.0	–	–20377.8	0
	Poisson6	Year Stand	5.4	9.9	–	–20053.0	0
	Poisson7	Month Stand	5.3	6.9	–	–20157.5	0
	Poisson8	Year Month Stand	5.0	9.3	–	**–20393.9**	**1.0**
	NBD1*	No covariate	1.2	8.3	0.894	–22350.2	0
	NBD2	Year	1.3	10.4	0.777	–22439.3	0
	NBD3	Month	1.3	7.2	0.741	–22460.9	0
	NBD4	Stand	1.2	8.4	0.894	–22346.8	0
	NBD5	Year Month	1.3	10.8	0.698	**–22496.4**	**0.83**
	NBD6	Year Stand	1.3	10.1	0.775	–22440.5	0
	NBD7	Month Stand	1.3	7.0	0.741	–22457.2	0
	NBD8	Year Month Stand	1.3	10.5	0.696	–22493.3	0.17
Occupancy	Logistic1	No covariate	10.6	0.81	–	1012.7	0
	Logistic2	Year	2.5	0.99	–	681.1	0
	Logistic3	Month	6.5	0.72	–	811.4	0
	Logistic4	Stand	10.5	0.86	–	1003.5	0
	Logistic5	Year Month	1.5	0.99	–	**648.6**	**0.89**
	Logistic6	Year Stand	2.5	0.77	–	681.1	0
	Logistic7	Month Stand	2.5	0.76	–	813.0	0
	Logistic8	Year Month Stand	1.5	0.99	–	652.7	0.11

φ, λ, μ, p and k are defined in Table 1.
*NBD = negative binomial distribution.

$$AIC_c = -2\,LL + 2\theta + \frac{2K(\theta+1)}{n-\theta-1}, \quad (5)$$

where LL is the log likelihood, θ is the number of parameters in the model and n is the sample size. In general, models with lower AICc scores are considered to be better candidates than those with higher scores. Akaike weights (AIC_w) were calculated from AIC_c. AIC_w indicates the probability that the model is the best among the whole set of candidate models. Therefore, it provides a measure of the strength of evidence for each model (Johnson & Omland 2004).

Maximum likelihood estimates of the dispersion parameter (k) of the negative binomial distribution models (NBD1-8) of *Eurytoma* (Table 2) and psyllid (Table 3) abundance were incorporated into Equation 4 to derive the occupancy-abundance relationship. Mean abundance values were calculated based on the best model from the GLMs (Tables 2, 3) for *Eurytoma* and psyllids. Hence, the predicted occupancy of *Eurytoma* was obtained by inserting the average number of damaged seeds per pod for each year and month (NBD5) and the corresponding value of k into Equation 4. For the sake of comparison, the occupancies predicted by the Poisson and the worst negative binomial distribution model without covariates (NBD1) were obtained for *Eurytoma* and psyllids.

The predicted abundances (λ or μ) were estimated from the proportion of damaged seeds computed based on the best model (Logistic5) for *Eurytoma* (Table 2). The proportion of shoots infested by psyllids was computed based on both Logistic1 and Logistic2 (Table 3). Then abundance was predicted by inserting the occupancy values and the dispersion parameter (k) thus computed into Equations 6 and 7 (Sileshi *et al.,* 2006) for the Poisson and negative binomial distribution, respectively:

$$\lambda = -\ln(1-\psi) \quad (6)$$

$$\mu = k((1-\psi)^{-1/k} - 1)\,. \quad (7)$$

Table 3. Parameters, goodness-of-fit of abundance and occupancy models of leucaena psyllid and model selection based on second-order Akaike information criteria (AIC_c) and Akaike weights (AIC_w).

Variable	Model	Covariates in model	φ	λ, μ or p	k	AIC_c	AIC_w
Abundance	Poisson1	No covariate	10.89	15.6	2	–19738.8	0
	Poisson2	Site	10.70	20.5	2	**–19828.0**	**1.0**
	NBD1*	No covariate	1.18	15.6	0.85	**–22377.6**	**0.62**
	NBD2	Site	1.19	20.5	0.79	–22376.6	0.38
Occupancy	Logistic1	No covariate	0.89	0.93	2	**193.8**	**0.97**
	Logistic1	Site	0.93	0.93	2	201.0	0.03

φ, λ, μ, p and k are defined in Table 1.
*NBD = negative binomial distribution.

RESULTS

Abundance

Examination of the dispersion statistic (φ) shows that the Poisson distribution assumption, *i.e.* spatial randomness, did not hold for the abundance of *Eurytoma* and psyllids. Values of φ were smaller for the models under the negative binomial distribution assumption than the Poisson (Tables 2, 3). Estimates of *Eurytoma* (Table 2) and psyllid abundance (Table 3) parameters (λ and μ) differed among models. For example, the model without covariates (Poisson1 and NBD1), month (Poisson3 and NBD3) or stand alone (Poisson4 and NBD4) underestimated *Eurytoma* abundance compared to the model with year alone or year and month (Table 2).

The AIC_c scores were larger in all models under the Poisson than the negative binomial distribution indicating that the negative binomial distribution is better than the Poisson for *Eurytoma* (Table 2) and psyllid (Table 3) abundance. Under the Poisson assumption, the smallest AIC_c score (–20393.9) was obtained in the model containing year, month and stand (Poisson8). This model also had the maximum AIC_w value indicating that it has 100 % likelihood of being the model that gives the best estimate of *Eurytoma* abundance under the Poisson assumption. Under the negative binomial distribution assumption, the smallest AIC_c (–22496.4) was recorded in the model consisting of year and month (NBD5), which had the largest likelihood of 83 % (AIC_w = 0.83). This model also had wider confidence interval for *Eurytoma* abundance and the smaller estimate of k (0.698) among all the models considered (Table 2). Among the single-covariate Poisson and negative binomial distribution models, only year and month had significant effect on the abundance of *Eurytoma* (Table 4). When stand was considered along with month or year, it had significant influence under the Poisson assumption but not under the negative binomial distribution (Table 4).

In the case of psyllids (Table 3), the smaller AIC_c (–19828) was found in the model with site (Poisson2), which had 100 % likelihood (AIC_w = 1.0) of being the best Poisson model. On the other hand, the negative binomial distribution model without covariates (NBD1) had 62 % likelihood compared to the model with site, which had 38 % likelihood. The psyllid abundance estimates under both Poisson and negative binomial distribution assumptions were higher in the model with site compared with the one with out covariate effect (Table 3). However, likelihood ratio test showed no significant effect of site on psyllid abundance.

Occupancy

Estimates of detection probability (p) differed among the models of *Eurytoma* (Table 2) and psyllid occupancy (Table 3). The largest p value for *Eurytoma* occupancy was found in the model with year and month (Logistic5), while detectability was smallest in the model with month alone (Logistic3). Logistic5 also had the smallest value of φ (1.5) and AIC_c (648.6), while the largest values of φ (10.6) and AIC_c (1012.7) were recorded in the model without covariates effects (Table 2). The model with year and month also had 89 % (AIC_w = 0.89) more likelihood of being the best model among those compared. When considered singly, year, month and stand influenced detection probability significantly (Table 4). However, stand did not significantly influence detection probability when considered along with month or year.

In the case of the psyllids, the model without

Table 4. Significance of covariates under Poisson, negative binomial and logistic model to abundance and occupancy of *Eurytoma* in *Sesbania*.

Variable	Distribution	Model	Covariates in model	$P(\chi^2)$
Abundance	Poisson	Poisson2	Year	<0.0001
		Poisson3	Month	<0.0001
		Poisson4	Stand	0.4703
		Poisson5	Year	<0.0001
			Month	<0.0001
		Poisson6	Year	<0.0001
			Stand	<0.0001
		Poisson7	Month	<0.0001
			Stand	0.0124
		Poisson8	Year	<0.0001
			Month	<0.0001
			Stand	<0.0001
	NBD	NBD2	Year	<0.0001
		NBD3	Month	<0.0001
		NBD4	Stand	0.8041
		NBD5	Year	<0.0001
			Month	<0.0001
		NBD6	Year	<0.0001
			Stand	0.2538
		NBD7	Month	<0.0001
			Stand	0.5570
		NBD8	Year	<0.0001
			Month	<0.0001
			Stand	0.3148
Occupancy	Logistic	Logistic2	Year	<0.0001
		Logistic3	Month	<0.0001
		Logistic4	Stand	0.0004
		Logistic5	Year	<0.0001
			Month	<0.0001
		Logistic6	Year	<0.0001
			Stand	0.8820
		Logistic7	Month	<0.0001
			Stand	0.1110
		Logistic8	Year	<0.0001
			Month	<0.0001
			Stand	0.8763

covariates (Logistic1) had a smaller φ (0.89), AIC_c (193.8) and 97 % likelihood of being a better model than the one with site. Likelihood ratio tests also showed that site does not have a significant influence on detectability of psyllids.

Relation between occupancy and abundance

Figure 1 shows the observed occupancy of *Eurytoma* and that predicted from abundance assuming spatial randomness (Poisson), and the worst (NBD1) and best (NBD5) negative binomial distribution models. The Poisson and NBD1 overestimated occupancy especially at lower densities relative to the observed (Fig. 1a). Under the Poisson assumption, *Eurytoma* occupancy saturated faster at higher densities (>6 wasps per pod), yielding little information about abundance. On the other hand, the negative binomial distribution model with year and month (NBD5) predicted occupancy much closer to the observed than Poisson. Saturation was not observed under the negative binomial distribution models even at the highest density. The Poisson and NBD1 also underestimated abundance at lower *Eurytoma* occupancy rates, while NBD5 predicted abundance much closer to the observed abundance (Fig. 1b).

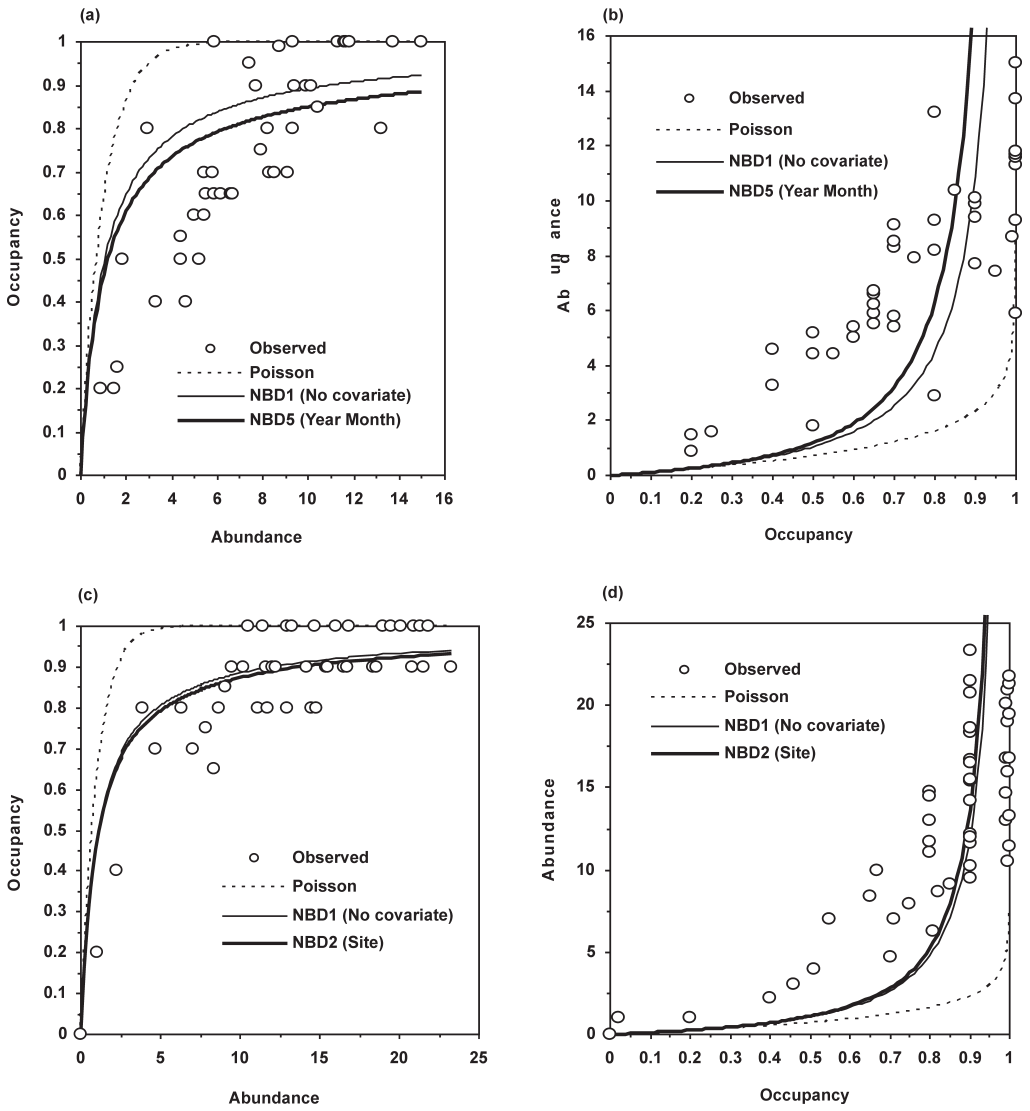

Fig. 1. Observed and predicted occupancy *versus* observed abundance (**a**) and observed abundance and abundance predicted from occupancy (**b**) of *Eurytoma* sp. (Hymenoptera: Eurytomidae) in *Sesbania* pods and observed and predicted occupancy *versus* observed abundance (**c**) and observed abundance and abundance predicted from occupancy (**d**) of psyllids on *Leucaena* shoots

The observed occupancy of the leucaena psyllid and that predicted from its abundance assuming Poisson and the negative binomial distribution models are presented in Fig. 1c,d. The Poisson overestimated occupancy, and the proportion of infested shoots saturated faster when psyllid densities are more than 5. The negative binomial distribution models (NBD1 and NBD2) predicted occupancy closer to the observed data compared with the Poisson (Fig. 1c). The Poisson also underestimated abundance at all levels of occupancy compared with the negative binomial distribution models (Fig. 1d). The occupancy and abundance of psyllids predicted by NBD1 and NBD2 only differed slightly.

DISCUSSION

Using *Eurytoma* and leucaena psyllid count data, this paper has demonstrated the application of (1)

GLMs for modelling abundance and detection probability, (2) information criteria for model selection and (3) occupancy-abundance models for precise estimation of insect abundance. GLMs indicated that the assumption of spatial randomness did not hold for both *Eurytoma* and psyllid data. Model selection criteria (AIC_c and AIC_w) (Table 2) and likelihood ratio statistics (Table 4) also showed that abundance and detectability of *Eurytoma* was significantly influenced by year and month when considered singly or in combination with other covariates. This indicates that temporal, both seasonal and annual, changes in *Eurytoma* populations induced heterogeneity that could not be adequately accounted by the Poisson assumption of spatial randomness.

Information criteria indicate that the negative binomial distribution models provide better descriptions of the abundance of both insects than did the Poisson. The negative binomial distribution has also been found to be better than the Poisson for description of several other insects (Sileshi 2006). Among the negative binomial distribution models, NBD5 had the highest likelihood of being the best model and gave a better estimate of *Eurytoma* abundance. This shows that covariate information further improved the precision of abundance estimates of *Eurytoma.* It also indicates that *Eurytoma* abundance varied as a function of year and month (Table 4) under the negative binomial distribution assumptions. Although site did not have a statistically significant effect on abundance, Poisson2 and NBD2 appear to give a better estimate of abundance of psyllids compared with Poisson1 and NBD1, which underestimated abundance.

Although the occupancy predicted by the negative binomial models was much closer to the observed than that predicted by the Poisson models (Fig. 1) the Poisson model generally overestimated occupancy and underestimated abundance relative to the observed. Under the Poisson assumption, occupancy of both insects studied also saturated faster at higher densities, yielding little information about their abundance (Fig. 1). It has also been shown in other insects (Sileshi *et al.* 2006) that the uncertainty associated with the prediction may be too large to ignore, as occupancy rates approach 1. The fact that many other distributions including the Poisson are related to the negative binomial distribution (Johnson & Kotz 1969; Sileshi 2007) suggests that the spatial distribution of both insects is likely to be approximated by this distribution. Hence, the occupancy-abundance models derived from the negative binomial distribution should be considered as more realistic approximations to the underlying process that generated the observed data. However, the negative binomial distribution may give an unreliable prediction if a constant k is used (Taylor *et al.* 1979; Sileshi *et al.,* 2006). As demonstrated in Table 2 and Fig. 1, k values and hence the predicted occupancies are greatly influenced by covariate structure in the data. This highlights the need for careful consideration of covariate effects and selection of models using objective criteria when constructing occupancy-abundance relationships.

In order to select an adequate model among the range of models considered (Table 2 and 3), information criteria are more appropriate than traditional likelihood statistics (Johnson & Omland 2004; Sileshi 2006). Once an appropriate model has been identified using such criteria, occupancy-abundance relationships can be established. Then, future monitoring could entirely be based on presence–absence sampling, and abundance estimates for a given area (belonging to the same statistical universe) may be obtained directly from the proportion of occupied sampling units observed in a random sample taken even from areas where abundance surveys could not be conducted.

This approach is particularly appealing in the monitoring of insects in conservation projects because it yields inferences about the status of a population based only on the presence or absence of individuals, data that can be relatively easily collected. The method also has a potential application for predicting large-scale population dynamics of invasive alien insects because it allows factors affecting fine-scale population dynamics, *e.g.* the effects of habitat quality, to be linked with the factors determining regional abundance and hence to be able to predict the total size of a regional population (Freckleton *et al.* 2005). Counting individual insects such as psyllids is not only time consuming and expensive but it can also be extremely difficult to obtain an unbiased estimate. Occupancy-abundance models have been shown to be cost effective (Joseph *et al.* 2006).

Occupancy-abundance models (He & Gaston 2003; Sileshi *et al.* 2006) have often been described assuming that abundance is observable without

error. On the other hand, approaches for estimating occupancy (*e.g.* MacKenzie *et al.* 2002; 2003) have been described without considering occupancy-abundance relationships. Only very few published studies (*e.g.* Royle *et al.* 2005) exist using the method described here. Therefore, further research is needed in this field to develop models that provide better estimates.

ACKNOWLEDGEMENTS

My sincere thanks are due to F. Phiri (Zambia) for data collection. Financial support for the various studies reported came from the Canadian International Development Agency (CIDA) and Swedish International Development Agency (SIDA) through the Zambezi Basin Agroforestry Project of the World Agroforestry Centre (ICRAF).

REFERENCES

ANDREWARTHA, H.G. & BIRCH, L.C. 1954. *The Distribution and Abundance of Animals.* University of Chicago Press, Chicago.

BAILEY, L.R., SIMONS, T.R. & POLLOCK, K.H. 2004. Estimating site occupancy and species detection probability parameters for terrestrial salamanders. *Ecological Applications* **14**: 692–702.

BURNHAM, K.P. & ANDERSON, D.R. 2002. *Model Selection and Multimodel Inference: A Practical Information-theoretic Approach,* 2nd Edition. Springer-Verlag, New York.

CABEZA, M., ARAUJO, M.B., WILSON, R.J., THOMAS, C.D., COWLEY, M.J.R. & MOILANEN, A. 2004. Combining probabilities of occurrence with spatial reserve design. *Journal of Applied Ecology* **41**: 252–262.

CAMERON, A.C. & TRIVEDI, P.K. 1998. *Regression Analysis of Count Data.* Cambridge University Press, New York.

DAY, R.K. 1999. *Integrated control of Leucaena psyllid.* Final Technical Report, DFID Crop Protection Programme, NRI, Greenwich, Kent, U.K.

FRECKLETON, R.P. GILL, J.A, NOBLE, D & WATKINSON, A. R. 2005. Large-scale population dynamics, abundance–occupancy relationships and the scaling from local to regional population size. *Journal of Animal Ecology* **74**: 353–364.

GASTON, K., BLACKBURN, T.M., GREENWOOD, J.J.D., GREGORY, R. QUINN, R.M. & LAWTON, J.H. 2000. Abundance-occupancy relationships. *Journal of Applied Ecology* **37**: 39–59.

GU, W. & SWIHART, R.K. 2004. Absent or undetected? Effects of non-detection of species occurrence on wildlife-habitat models. *Biological Conservation* **116**: 195–203.

HANSKI, I.A. & GILPIN, M.E. 1997. *Metapopulation Biology–Ecology, Genetics and Evolution.* Academic Press, San Diego.

HARTLEY, S. 1998. A positive relationship between local abundance and regional occupancy is almost inevitable (but not all positive relationships are the same). *Journal of Animal Ecology* **67**: 992–994.

HASSAN, S.T.S., RASHID, M.M., ARSHAD, M.A., BAKAR, R.A., HUSSEIN, M.Y. & SAJAP, A.S. 1994. Within-plant mainstem nodal distribution of the psyllid, *Heteropsylla cubana,* on leucaena, *Leucaena leucocephala,* plant. *Proceedings of 4th International Conference of Plant Protection in the Tropics, 28–31 March 1994, Kuala Lumpur.*

HE, F. & GASTON, K.J. 2003. Occupancy, spatial variance, and the abundance of species. *The American Naturalist* **162**: 366–375.

JOHNSON, N.I. & KOTZ, S. 1969. *Discrete Distributions.* Houghton Mifflin, Boston.

JOHNSON, J.B. & OMLAND, K.S. 2004. Model selection in ecology and evolution. *Trends in Ecology and Evolution* **19**: 101–108.

JOSEPH, L.N., FIELD, S.A., WILCOX, C. & POSSINGHAM, H.P. 2006. Presence–absence *versus* abundance data for monitoring threatened species. *Conservation Biology* **20**: 1679–1687.

KUNIN, W.E., HARTLEY, S. & LENNON, J. 2000. Scaling down: On the challenges of estimating abundance from occurrence patterns. *The American Naturalist* **156**: 560–566.

LAWLESS, J.F. 1987. Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics* **15**: 209–225.

LEATHER, S.R. & WATT, A.D. 2004. Sampling theory and practice. . In: Leather, S.R (Ed.) *Sampling in Forest Ecosystems.* 1–15. Blackwell Publishing, Oxford.

MACDONALD, I.A.W., REASER, J.K., BRIGHT, C., NEVILLE, L.E., HOWARD, G.W., MURPHY, S.J. & PRESTON, G. (Ed.) 2003. *Invasive Alien Species in southern Africa: National Reports and Directory of resources.* Global Invasive Species Programme, Cape Town, South Africa.

MACKENZIE, D.I. & NICHOLS, J.D. 2004. Occupancy as a surrogate for abundance estimation. *Animal Biodiversity and Conservation* **27**: 461–467.

MACKENZIE, D.I., NICHOLS, J.D. LACHMAN, G.B., DROEGE, S., ROYLE, J.A. & LANGTIMM, C.A. 2002. Estimating site occupancy rates when detection probabilities are less than one. *Ecology* **83**: 2248–2255.

MACKENZIE, D.I., NICHOLS, J.D., HINES, J.E., KNUTSON, M.G. & FRANKLIN, A.B. 2003. Estimating site occupancy, colonization and local extinction when a species is detected imperfectly. *Ecology* **84**: 2200–2207.

NACHMAN, G. 1981. A mathematical model of the functional relationship between density and spatial distribution of a population. *Journal of Animal Ecology* **50**: 453–460.

POLLOCK, K.H., NICHOLS, J.D., SIMONS, T.R. FARNSWORTH, G.L., BAILEY, L.L. & SAUER, 2002. Large-scale wildlife monitoring studies: statistical methods for design and analysis. *Environmetrics* **13**: 105–119.

ROYLE, J.A. & NICHOLS, J.D. 2003. Estimating abundance from repeated presence–absence data or

point counts. *Ecology* **84**: 777–790.
ROYLE, J.A., NICHOLS, J.D. & KÉRY, M. 2005. Modelling occurrence and abundance of species when detection is imperfect. *Oikos* **110**: 353–359.
SAHA, K. & PAUL, S. 2005. Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics* **61**: 179–185.
SAS Institute Inc 2003. SAS/STAT, Release 9.1. SAS Institute, Cary, NC.
SILESHI, G. 2003. Some insects feeding on seeds of *Sesbania sesban* and related species in southern Africa. *African Entomology* **11**: 134–137.
SILESHI, G. 2006. Selecting the right statistical model for analysis of insect count data by using information theoretic measures. *Bulletin of Entomological Research* **96**: 479–488.
SILESHI, G. 2007. Evaluation of statistical procedures for efficient analysis of insect, disease and weed abundance and incidence data. *East African Journal of Science* **1**: 1–9.
SILESHI, G., GIRMA, H. & MAFONGOYA, P.L. 2006. Occupancy-abundance models for predicting densities of three leaf beetles damaging the multipurpose tree *Sesbania sesban* in eastern and southern Africa. *Bulletin of Entomological Research* **96**: 61–69.
SPEIGHT, M.R., HUNTER, M.D. & WATT, A.D. 1999. *Ecology of Insects: Concepts and Applications*. Blackwell Science, Oxford.
TAYLOR, L.R., WOIWOD, I.P. & PERRY, J.N. 1979. The negative binomial as a dynamic ecological model and the density-dependence of *k*. *Journal of Animal Ecology* **48**: 289–304.
TYRE, A.J., TENHUMBERG, B., FIELD, S.A, NIEJALKE, D., PARRIS, K., & POSSINGHAM, H.P. 2003. Improving precision and reducing bias in biological surveys: estimating false-negative error rates. *Ecological Applications* **13**: 1790–1801.
WARREN, M., McGEOCH, M.A. & CHOWN, S.L. 2003. Predicting abundance from occupancy: a test for an aggregated insect assemblage. *Journal of Animal Ecology* **72**: 468–477.
WILSON, L.T. & ROOM, P.M. 1983. Clumping patterns of fruit and arthropods in cotton, with implications for binomial sampling. *Environmental Entomology* **12**, 50–54.
WRIGHT, D.H. 1991. Correlation between incidence and abundance area expected by chance. *Journal of Biogeography* **18**: 463–466.

Accepted 15 February 2007

Appendix 1. SAS procedures used for fitting models of *Eurytoma* abundance.

```
Data Eurytoma;
Input Year Month $ Stand $ Count;
If Count>0 then detect=1;
else detect=0;/*Converts counts into 1(=detection) or 0(=non-detection)*/
Cards;
1998    July     Fallow      6
1998    July     Isolat      20
1999    July     Fallow      4
1999    July     Isolat      6
1998    Aug      Fallow      6
1998    Aug      Isolat      0
1999    Aug      Fallow      0
1999    Aug      Isolat      0
1998    Dec      Fallow      10
1998    Dec      Isolat      0
1999    Dec      Fallow      0
1999    Dec      Isolat      4
.
.
;
Title1 'Poisson† regression model for Eurytoma abundance';
Proc genmod Data=Eurytoma;
Class Year Month Stand;
model Count=/dist=poisson link=log type3;/*Poisson1*/
Run;
Proc genmod Data = Eurytoma;
Class Year Month Stand;
model Count=Year/dist=poisson link=log type3;/*Poisson2*/
Run;
Proc genmod Data = Eurytoma;
Class Year Month Stand;
```

†For the NBD model, replace dist = Poisson with dist=nb in the model statement of all of the above.

```
model Count=Month/dist=poisson link=log type3;/*Poisson3*/
Run;
Proc genmod Data = Eurytoma;
Class Year Month Stand;
model Count=Stand/dist=poisson link=log type3;/*Poisson4*/
Run;
Proc genmod Data = Eurytoma;
Class Year Month Stand;
model Count=Year Month/dist=poisson link=log type3;/* Poisson5*/
Proc genmod Data = Eurytoma;
Class Year Month Stand;
model Count=Year Stand/dist=poisson link=log type3;/*Poisson6*/
Run;
Proc genmod Data = Eurytoma;
Class Year Month Stand;
model Count=Month Stand/dist=poisson link=log type3;/*Poisson7*/
Run;
Proc genmod Data = Eurytoma;
Class Year Month Stand;
model Count=Year Month Stand/dist=poisson link=log type3;/*Poisson8*/
Run;
```

Appendix 2. SAS procedures used for fitting various models of *Eurytoma* detection probability.

Title2 'Logistic regression model for *Eurytoma* detection probability';

```
Proc logistic Data=Eurytoma;
Class Year Month Stand;
model detect=/waldrl aggregate;/*Logistic1*/
Run;
Proc logistic Data = Eurytoma;
Class Year Month Stand;
model detect=Year/waldrl aggregate;/*Logistic2*/
Run;
Proc logistic Data = Eurytoma;
Class Year Month Stand;
model detect=Month/waldrl aggregate;/*Logistic3*/
Run;
Proc logistic Data = Eurytoma;
Class Year Month Stand;
model detect=Stand/waldrl aggregate;/*Logistic4*/
Run;
Proc logistic Data = Eurytoma;
Class Year Month Stand;
model detect=Year Month/ waldrl aggregate;/*Logistic5*/
Proc logistic Data = Eurytoma;
Class Year Month Stand;
model detect=Year Stand/waldrl aggregate;/*Logistic6*/
Run;
Proc logistic Data = Eurytoma;
Class Year Month Stand;
model detect=Month Stand/waldrl aggregate;/*Logistic7*/
Run;
Proc logistic Data = Eurytoma;
Class Year Month Stand;
model detect=Year Month Stand/waldrl aggregate;/*Logistic8*/
Run;
```