Evaluation of Statistical Models for Analysis of Insect, Disease and Weed Abundance and Incidence Data

G. Sileshi

World Agroforestry Centre (ICRAF), SADC-ICRAF Agroforestry Programme, Chitedze Agricultural Research Station, P O Box 30798, Lilongwe, Malawi
E-mail: sgwelde@yahoo.com

Abstract: Analysis of variance (ANOVA) has been a fundamental method used for analysis of abundance and incidence data. However, abundance and incidence data often violate the assumptions of ANOVA. Researchers often ignore ANOVA assumptions, transform the data using arbitrarily chosen functions and then fail to evaluate whether or not the transformation actually corrected the problem. The statistical power of the tests used is also seldom reported. Therefore, the objectives of this paper are to demonstrate (1) implications of using arbitrarily chosen transformations and ANOVA to the validity of statistical inference on pest abundance and incidence and (2) the application of LMMs and GLMs for efficient analysis of such data. Abundance data were analyzed assuming normal, Poisson and negative binomial error distributions. Incidence data were analyzed assuming normal and binomial error distributions. Among the data transformation functions, logarithmic transformation gave better description of abundance data compared with square root. Working logits were better than angular or square root transformation of incidence data. The study has also demonstrated that the choice of transformation can influence the statistical significance and power of test. Transformation of either abundance or incidence data did not necessarily ensure normality or variance homogeneity. According to the Akaike information criterion (AIC), a GLM assuming negative binomial error distribution was better for description of most abundance datasets compared with a GLM assuming Poisson error distribution or LMM. LMM based on working logits also gave a better description of the data than a GLM assuming binomial distribution. It is concluded that LMMs and GLMs simultaneously consider the effect of treatments and heterogeneity of variance and hence are more appropriate for analysis of abundance and incidence data than ordinary ANOVA.

Key words*:* Mixed Models; Generalized Linear Models; Statistical Power

1. Introduction

Abundance refers to the number of individuals per unit area. It is a fundamental ecological parameter and a critical consideration when making management and conservation decisions. In most work in entomology, pathology and weed sciences, counts are used as proxies of abundance. In the case of counts a substantial proportion of the values are zero, and the remainder have a skewed distribution (Fletcher *et al*., 2005; Martin *et al*., 2005; Warton, 2005). The term incidence refers either to the number of plants (or plant units) that is visibly diseased or affected by an insect (out of a given total number) (Madden and Hughes, 1995). Pathologists and entomologists often collect incidence data because in many instances, notably with plant diseases caused by viruses, it is impractical to assess diseases on the basis of pathogen abundance (McRoberts *et al*., 1996). Similarly, with small arthropods such as mites, thrips, aphids, psyllids and leafhoppers, presence or absence is often easier to establish than estimating abundance by counting individuals. Recently, a positive relationship between incidence and abundance of a species has been demonstrated (Gaston *et al*., 2000). Based on such relationships, Sileshi *et al.* (2006a) have demonstrated that insect abundance can be estimated from incidence or *vice versa*.

By their nature, abundance and incidence are not normally distributed (Madden and Hughes, 1995; Garrett *et al*., 2004). Abundance is quantified by discrete variables, and can be described well by the Poisson or negative binomial distributions. The Poisson distribution is described by one parameter, θ, or the mean. In the Poisson distribution the variance is equal to the mean (Johnson and Kotz, 1969). The negative binomial distribution (NBD) is more convenient model for analyzing insect and weed counts or pathogen density with over-dispersion (Anscombe, 1949; McRoberts *et al*., 1996; Sileshi *et al*., 2006a; b). The NBD is related to several distributions. According to Johnson and 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. This supports one of the four derivations of the NBD (Anscombe, 1950). 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 $k \rightarrow \infty$ (Saha and Paul, 2005).

Incidence is a binary variable because each observed individual plant is either visibly affected or not, or damage symptoms are present or absent (Madden,

ISSN 1992-0407

2002). Hence, it is characterized by a binomial or beta-binomial distribution (Madden and Hughes, 1995; Collette, 2002). Despite the many advantages of using the binomial distribution (Collette, 2002), this distribution only occasionally describes actual disease incidence data. Diseased individuals typically are clustered in nature, resulting in greater heterogeneity of disease incidence than would be expected for a random pattern (Madden, 2002). More typically, the variance is larger and the observed frequency of diseased individuals is more skewed than that predicted by the binomial distribution (Hughes and Madden, 1995). The variance is a function of the mean in both incidence (Hughes and Madden, 1995) and abundance (Taylor, 1961) data.

Statistical inference based on abundance and incidence using conventional statistical methods such as analysis of variance (ANOVA) poses several challenges. There is a wide range of situations where the assumptions of normality and homogeneity of variance are not met for insect or weed abundance data (Sileshi and Mafongoya, 2002; 2003; Sileshi *et al*., 2002; 2006b). ANOVA models focus on null hypothesis testing based on mean tendencies in the data. These tests typically assume that the errors (after fitting the model) are independent and identically distributed as normal random variables with constant variance. These techniques were developed, and to some extent derive their validity, from the randomization underlying designed experiments (Fisher, 1935). However, a large proportion of entomological and pathological research consists of observational studies in which the goal is to explain a pattern relative to a series of explanatory variables.

The standard methodology in ANOVA has been to use a transformation of the response variable that results in a variable that is approximated by a normal distribution. In a sense, this is forcing the data to fit a model that was developed for analysis of continuous variables, rather than using an appropriate statistical model for the data at hand (Hughes and Madden, 1995; Garrett *et al*., 2004; Madden *et al*., 2002). Furthermore, variance-stabilizing transformations may not, in fact, fully stabilize variances in count (McArdle and Anderson, 2004) or incidence data when some of the means are close to 0 or 100% (Madden, 2002). It is well known that departures from the assumption of homogeneity can result in inflated error rates (Cochran, 1947). Tests of significance, standard errors, and contrasts of the means can be affected if ANOVA is used for discrete and binary data.

In ANOVA, coefficients are computed using ordinary least square (OLS) methodology which minimizes the sum of squared distances of data points to the parameter estimate. An alternative to OLS is provided by the Restricted Maximum Likelihood (REML) and maximum likelihood (ML) estimation methods (Littell, 2002). REML is used in linear mixed models (LMM), while ML can be used in both LMM and generalized linear models (GLMs) (Collett, 2002). The LMM is an extension of ANOVA, and it still assumes normality (Littell, 2002). However, it extends the ANOVA model by allowing for both correlation and heterogeneous variances. Wolfinger (1993) and Piepho *et al*. (2003) provide detailed information on LMMs. Better still are GLMs, which are more appropriate for analyzing discrete and binary data (McCullagh and Nelder, 1989; Collett, 2002; Madden, 2002; Hughes and Madden, 1995; Garrett *et al*., 2004; Turechek and Madden, 2002). In GLMs, the response is assumed to possess a probability distribution of the exponential form such as the Poisson, NBD and binomial. In GLM coefficients are computed using ML, which maximize the odds that a dependent variable equals a given value. Here, a function of the expected value of Y is modelled as a linear function of the variables of interest (Collett, 2002). This function can be written as $g(\mu)$, where μ is the expectation of $Y[\mu=E(Y)$, and is known as the link function. This is quite different from the regular normal distribution-based approach of transforming Y to produce $g(Y)$ and then fitting a model to $g(Y)$. The reader is referred to McCullagh and Nelder (1989) for detailed information on GLMs, and to Hartley and Rao (1967) and Harville (1977) for information on REML and ML.

Despite the recent developments on LMM and GLM methodology (Garrett *et al*., 2004; Madden, 2002; Madden *et al*., 2002) and wider availability of computer software, they have been little used by entomologists, pathologists and weed scientists. Researchers still use arbitrarily chosen data transformations and apply OLS ANOV. Very few actually are aware of the power of these tests (Thomas and Krebs, 1997). The statistical power of a significance test is the long-term probability (given the population effect size, alpha, and sample size) of rejecting a false null hypothesis. While power analysis is a vital tool for study planning, it has been largely ignored in entomology, pathology and weed research. The objectives of this paper are to demonstrate (1) implications of using arbitrarily chosen transformations and ANOVA to the validity of statistical inference on pest abundance and incidence and (2) the application of LMMs and GLMs for efficient analysis of such data.

2. Materials and Methods

2.1. Source of Data

The data reanalysed in this study included abundance of witch weeds (*Striga asiatica*), grass and broad leafed weeds in maize, and two insect species, namely the leucaena psyllid (*Heteropsylla cubana*) and *Exosoma* sp. The data on witch weed (*Striga asiatica*) comes from Sileshi *et al.* (2006b) but is restricted to only one of the experiments described in that paper. The experiment was established in December 1991 and consisted of maize grown in a mixed intercropped system with the tree legumes *Caliandra callothyrsus*, *Flemingia macrophylla*, *Gliricidia sepium*, *Leucaena leucocephala*, *Senna siamea*, *Sesbania sesban*. Details of the treatments, plot layout, randomization and

management of this experiment have been described in Sileshi *et al.* (2005; 2006b). The abundance of witch weed was monitored in 1995, 1996 and 1997 cropping season, and the effect of treatment and year of sampling was analyzed. The abundance of arable weeds was assessed in a legume fallow experiment established in the year 2000 at Msekera sites. The treatments in this experiment consisted of maize planted in pure-species fallows of *Gliricidia sepium*, *Acacia anguistissima*, *Leucaena collinsi*, *Calliandra calothyrsus*, *Senna siamea* and maize monoculture with and without fertilizer. Assessment was made by counting the number of grass weeds (all weeds of the family Graminae) and broad leafed weeds (all non-grass weeds) in an area measuring 1 m by 1 m, and the effect of treatments on abundance of grass and broad leaved weeds was analysed.

Populations of the leucaena psyllid, *Heteropsylla cubana* (Homoptera: Psyllidae), a pest of the tropical agroforestry tree *Leucaena leucocephalla* were monitored in April-May 2005 in four experiments established in 1991, 1992, 1997 and 1999 at Msekera. These experiments have been described in detail in Sileshi *et al*. (2005). In all the trials, trees were cut to a height of 30 cm above ground after three years of growth and allowed to 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 numbers of psyllids on a randomly selected shoot per stump were recorded. The effect of site of establishment on the abundance of psyllids was analyzed using the different statistical models.

The datasets on abundance of *Exosoma* came from studies reported elsewhere by the author (Sileshi and Mafongoya, 2002; and 2006b). Abundance data were collected from various agroforestry treatments involving *Sesbania sesban* at Msekera, eastern Zambia in 2002. In each treatment the numbers of *Exosoma* on 10 randomly selected plants were recorded and effect of treatment on abundance analysed.

The incidence data used in this study included foliar diseases (a complex of fungal diseases) of the indigenous fruit Uapaca (*Uapaca kirkiana*), and termite damage in maize reported by the author elsewhere (Sileshi *et al*., 2005). The study on Uapaca foliar diseases involved a randomized and replicated experiment consisting of a factorial combination of three potting mixtures (unsterilized forest soil, sterilized forest soil and forest soil + saw dust), soil applied fertilizer (with and without compound D), and a foliar applied fertilizer (with and without). Data on foliar disease incidence were collected in July and October 2002. The incidence of the insects and diseases was determined by observing the disease status of single whole plants used as the sampling unit. Incidence constituted the proportion of plants in a row showing foliar disease symptoms.

Termite damage was assessed in 2002 and 2003 in two experiments established in 1991 and 1992 at Msekera. The treatments consisted of maize grown after of *Caliandra callothyrsus*, *Flemingia macrophylla*, *Gliricidia sepium*, *Leucaena leucocephala*, *Senna siamea* and monoculture maize grown with and without the recommended rate of fertilizer. Details of the treatments, plot layout, randomization and management of this experiment have been described in Sileshi *et al*. (2005). In both experiments, damage was assessed by recording the number of lodged plants per plot in 2002 and 2003, and the effect of fallow length, year of sampling and treatment on termite incidence was analyzed.

2.2. Statistical Analyses of the Data

For analysis of abundance data, the normal, Poisson and negative binomial distribution models were used. The normal distribution model applies ordinary least square (OLS) ANOVA on transformed insect counts. The probabilistic model using OLS assumes that the underlying errors of the transformed data are all uncorrelated with homogeneous variance, and hence follow an approximate log-normal distribution (McArdle and Anderson, 2004; Warton, 2005). In this study, count data were transformed using natural logarithms and square root functions because of the popularity of these transformations. However, various other types of transformation are available for count data (Taylor, 1961; McArdle and Anderson, 2004). Incidence data were transformed using the angular (arcsine), square root functions and the working logit (Cox, 1970) given by

$$Z = \ln\left(\frac{R + ½}{n - R + ½}\right)$$

where Z is the working logit, R is the number responding (e.g., infested plants) and n is the number observed. The working logit was tested because this transformation has the advantage of being able to take 0 and 100% response data into account.

Then tests for normality and homogeneity of variance were conducted. Shapiro-Wilk statistic and the Kolmogorov-Smirnov D statistic were used for testing normality. The assumption of equality of variance in the transformed data was tested using Bartlett's and Levene's tests of homogeneity of variance via the GLM procedure of the SAS system (SAS/STAT, 2003). ANOVA was conducted on the transformed data using the GLM procedure of the SAS. The statistical power of the ANOVA was calculated using the GLMPOWER procedure of SAS system.

LMM was fitted to the transformed abundance and incidence data using MIXED procedure of the SAS system. All interaction effects were considered to be random effects in the LMM. The MIXED procedure was used because in most cases the experimental units on which the data were recorded were grouped into clusters (e.g. replications, rows etc.), and it was assumed that data from a common cluster were correlated. The Poisson and negative binomial distributions were used to analyse the abundance data.

A GLM to relate the mean abundance (μ) to the explanatory variables (X_i), the following linear probability model was used:

$$Log(\mu) = a + b_1 X_1 + b_2 X_2 + ... + b_n X_n$$

(Equation 1)

where a is the random intercept, X_1, X_2 … X_n are covariates and b_1, b_2 … b_n are parameters to be estimated for the n^{th} covariate. The *log* is the canonical link for the Poisson and negative binomial distributions.

Incidence data were analyzed using the GLMs by assuming binomial distribution of diseased individuals. A GLM to relate the binomial parameter (p) of incidence to the explanatory variables (X_i) a linear probability model of the following form was used:

$$Logit(p_i) = a + b_1 X_1 + b_2 X_2 + ... + b_n X_n$$

(Equation 2)

where the *logit* function is the canonical link for the binomial distribution. For the incidence of the foliar diseases of Uapaca, X_1 , X_2, X_3 and X_4 stand for potting mixture, foliar fertilizer, soil-applied fertilizer and month of sampling, respectively. When over-dispersion was noted, a dispersion parameter was introduced using the ratio of the deviance to its associated degrees of freedom (McCullagh and Nelder, 1989). Parameters of equations 1 and 2 were estimated by the ML method using the GENMOD procedure of SAS systems. For GLMs, the residual deviance is of central importance for determining goodness-of-fit of a model. Therefore, the residual deviance divided by its degrees of freedom (RD/DF) was used to detect goodness-of-fit to the models. Values of RD/DF greater than 1 indicated over-dispersion while values less than 1 indicated under-dispersion. Evidence of over-dispersion or under-dispersion was used as an indication of inadequate fit of the statistical model to the data. Akaike information criterion (AIC) was used for comparing the statistical models (Burnham and Anderson, 2002) and transformations. The second-order Akaike information criterion (AIC_c) correcting for small sample size (Hurvich and Tsai, 1989) was computed from the log likelihood (LL) estimates as:

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

(Equation 3)

where K is the number of parameters in the model and n is the sample size. The "smaller AIC_c is better" approach was used for comparisons among models. Among the models under consideration, the one with the smallest AICc has the smallest expected loss of information, and was interpreted as the best.

3. Results

The transformations did not normalize the abundance data except for grass and broad leafed weeds. Transformations also failed to normalize the incidence data. Levene's and Bartlett's tests indicated heterogeneity of variance across years in both raw and the transformed witch weed abundance data. When treatment was considered, Levene's test indicated homogeneous variance in the raw data, while Bartlett's test indicated heterogeneity. Both Levene's and Bartlett's test indicated variance heterogeneity across treatments in the raw as well as the transformed data on grass weed, broad leafed weed and *Exosoma* abundance (Table 1).

Table 1. Probability levels for Levene's and Bartlett's tests of homogeneity of variance in abundance before and after transformation of the data using logarithmic and square root (SQRT) functions

Pest	Fixed	Levene's test			Bartlett's test		
group	effect	Before	Logarithmic	SQRT	Before	Logarithmic	SQRT
Witch weeds	Year	0.006	<0.001	<0.001	<0.001	<0.001	<0.001
	Treatment	0.299	0.020	0.107	<0.001	0.044	<0.001
Grass weeds	Treatment	0.004	0.003	0.003	0.006	0.004	0.113
Broad leafed weeds	Treatment	<0.001	0.038	<0.001	<0.001	0.065	<0.001
Leucaena psyllid	Site	0.053	0.681	0.118	0.014	0.672	0.214
Exosoma sp.	Treatment	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

On the other hand, variances of leucaena psyllid abundance were homogeneous across sites after transformation compared with the raw data. Levene's and Bartlett's test indicate heterogeneity of variance in incidence of Uapaca leaf disease across months, potting mixtures and soil applied fertilizers before and after angular and square root transformation. However, foliar fertilizer treatment had homogeneous variance before and after transformation (Table 2). Incidence of termite damage in maize had heterogeneous variance across fallow length and year of sampling while treatment had homogeneous variance before and after angular and square root transformation according to the Levene's test. Bartlett's test indicated variance heterogeneity across treatment in the raw data and angular transformed termite incidence, but homogeneity in the square root transformed data. Transformation using working logits homogenized variance across fallow length and years in termite incidence according to both Levene's and Bartlett's tests (Table 2).

Table 2. Probability levels for Levene's and Bartlett's tests of homogeneity of variance in incidence data before and after transformation using angular (Arcsine), square root (SQRT) and working logits (Logit)

Pest	Fixed	Levene's test				Bartlett's test			
group	effect	Before	Arcsine	SQRT	Logit	Before	Arcsine	SQRT	Logit
UFD	Month	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
	Potting mix	<0.001	<0.001	<0.001	0.008	<0.001	<0.001	<0.001	0.024
	Soil applied	0.033	0.038	0.041	0.034	0.017	0.090	<0.001	0.045
	Foliar fert	0.901	0.687	0.354	0.836	0.889	0.735	0.064	0.843
Termites	Fallow	0.002	0.002	0.003	0.419	<0.001	<0.001	<0.001	0.312
	Year	0.005	0.087	0.037	0.412	<0.001	<0.001	0.012	0.291
	Treatment	0.559	0.478	0.685	0.151	<0.002	0.717	0.499	0.008

Under the normal distribution assumption the fixed effects were not significant before transformation of witch weed, grass weed and *Exosoma* abundance (Table 3).

Table 3. Significance (P values) of effects on abundance of pest groups using different data transformation and distribution assumptions

Pest	Fixed	LMM (Normal distribution)			GLM	
group	effect	Before	Logarithm	SQRT	Poisson	NBD
Witch weed	Year	0.144ns	0.021	0.164	<0.001	<0.001
Witch weed	Treatment	0.364ns	0.124ns	0.195	<0.001	<0.001
Grass weeds	Treatment	0.161ns	0.018	0.056	<0.001	<0.001
Broad leafed weeds	Treatment	<0.001	<0.001	<0.001	<0.001	<0.001
Leaucaen psyllid	Site	0.025	0.049	0.025	<0.001	0.070ns
Exosoma sp.	Treatment	0.384	0.320	0.342	<0.001	<0.001

ns= variances not significantly different within fixed effect

Significant effects were indicated after logarithmic transformation of witch weed and grass weed abundance at the 5% level. On the other hand, the Poisson and negative binomial models indicated highly significant (P<0.001) effects for all pest groups except the leucaena psyllid (Table 3). The statistical power of ANOVA for the various fixed effects was sufficiently high (>0.90) for most abundance data except for the raw data on witch weed abundance (Table 5).

If one were to analyze the raw data, one would require twice the number of observations to achieve the desired statistical power of 0.90. However, when data were transformed using the logarithmic function, the desired statistical power was achieved using the same sample size. Under both the normal and binomial distribution assumptions, the incidence of Uapaca foliar disease significantly differed with month, potting mixture and foliar application of fertilizer. The statistical power of test for the effect of month, potting mixture and foliar application of fertilizer were sufficiently high (>0.90) (Table 4).

Table 4. Significance (P values) of effects on and incidence of pest groups using different data transformation and distribution assumptions

Pest	Fixed	LMM (Normal distribution)				GLM (Binomial distribution)
group	effect	Before	Arcsine	SQRT	Logit	
UFD	Month	<0.001	<0.001	<0.001	<0.001	<0.001
	Potting mix	<0.001	<0.001	<0.001	<0.001	<0.001
	Soil applied	<0.001	<0.001	<0.001	<0.001	0.007
	Foliar fertilizer	0.083	0.068	0.231	0.162	0.258
Termites	Fallow length	0.003	0.007	<0.001	<0.001	<0.001
	Year	0.002	0.004	<0.001	<0.001	<0.001
	Treatment	0.799	0.892	0.467	0.504	0.582

Table 5. Statistical power of test for ANOVA before and after logarithmic and square root (SQRT) transformation of abundance of pest groups and additional samples required to achieve the desired statistical power

Pest group	Fixed effect	Sample size	Transformation		
			Before	Logarithmic	SQRT
Witch weed	Year	96 plots	0.77 (192)	>0.90 (0)	0.66 (96)
	Treatment	96 plots	0.66 (192)	>0.90 (0)	0.89 (96)
Grass weeds	Treatment	63 plots	>0.90 (0)	>0.90 (0)	>0.90 (0)
Broad leafed weeds	Treatment	63 plots	>0.90 (0)	>0.90 (0)	>0.90 (0)
Leaucaen psyllid	Site	360 shoots	>0.90 (0)	>0.90 (0)	0.72 (720)
Exosoma sp.	Treatment	270 plants	>0.90 (0)	>0.90 (0)	>0.90 (0)

Figures in parenthesis indicate additional sampling units required to achieve the desired statistical power of 0.90

However, incidence did not differ with soil application of fertilizer. Statistical power analysis showed that the lack of significance was due to the inadequacy of the sample size used. If a meaningful conclusion is to be drawn about the effect of soil application of fertilizer, at least 5-11 times more observations (or 2260-4972 Uapaca plants) would be required (Table 6).

Table 6. Statistical power of test for ANOVA before and after transformation of incidence data using the angular (arcsine), square root (SQRT) functions and working logits and additional samples required to achieve the desired statistical power

Pest group	Fixed effect	Sample size	Transformations			
			Before	Arcsine	SQRT	Logit
UFD	Month	452 plants	>0.90 (0)	>0.90 (0)	>0.90 (0)	>0.90 (0)
	Potting mix	452 plants	>0.90 (0)	>0.90 (0)	>0.90 (0)	>0.90 (0)
	Soil applied	452 plants	0.80 (452)	0.80 (904)	0.81 (452)	0.84 (906)
	Foliar fertilizer	452 plants	0.31 (1808)	0.33 (2260)	0.18 (4520)	0.22 (3624)
Termites	Fallow length	128 plots	>0.90 (0)	0.66 (246)	>0.90 (0)	>0.90 (0)
	Year	128 plots	>0.90 (0)	0.51 (369)	>0.90 (0)	>0.90 (0)
	Treatment	128 plots	0.21 (640)	0.07 (4059)	0.25 (512)	0.24 (612)

Figures in parenthesis indicate additional sampling units required to achieve the desired statistical power of 0.90

Similarly, ANOVA showed lack of treatment effects on the incidence of termite damage in maize. Power analysis indicated that this was due to inadequate sample size. To make a valid conclusion about the effect of treatments 6-8 times more observations than the current sample (or 768-984 plots) would be needed.

The DEV/DF values were greater than unity indicating that the Poisson assumption of random distribution did not hold for all the abundance data. The AIC scores (Table 7)

Table 7. Second-order Akaike Information Criteria (AICc smaller is better) for selection of transformation functions and statistical models appropriate for analysis of abundance and incidence of different pests

Pest group	Linear Mixed Model					Generalized Linear Models		
	Before	SQRT	Log	Arcsine	Logit	Poisson	NBD	Binomial
A. Abundance								
Witch weed	1227.4	593.5	212.6	NAPP	NAPP	-59170.6	**-70089.5**	NAPP
Grass weeds	622.2	270.3	6.8	NAPP	NAPP	-58671.9	**-60142.8**	NAPP
Broad leafed weeds	537.7	233.5	20.4	NAPP	NAPP	-11042.7	**-11455.2**	NAPP
Leucaena psyllid	2848.1	1389.2	1046.4	NAPP	NAPP	-19828.0	**-22376.6**	NAPP
Exosoma sp.	623.0	88.3	**-253.2**	NAPP	NAPP	304.2	264.2	NAPP
B. Incidence								
UFD	3922.9	1511.2	NAPP	3819.7	**1176.6**	NAPP	NAPP	1547.8
Termites	1132.9	519.4	NAPP	913.8	**388.1**	NAPP	NAPP	846.7

Bold entries indicate the best transformation or model
NAPP = not applicable

showed that the negative binomial model is better for description of the abundance data than the Poisson or normal distribution models. The only exception was abundance of *Exosoma* sp (Table 7). In the case of *Exosoma* sp, LMM (based on log transformation) was adequate for analysis of the data. Among the data transformation functions, logarithms gave the best description of the data (smallest AICc). According to AIC analysis of Uapaca foliar disease and termite incidence without transformation gives poorer description of the data than transformation. For Uapaca foliar disease and termite incidence, the best transformation was working logits. LMM based on working logits also gave a better description of the data than logistic regression (Table 7).

4. Discussion

The results presented indicate that transformation of either abundance or incidence data do not necessarily ensure normality. This is in agreement with the growing body of literature on the subject matter in ecology (Fletcher *et al.*, 2005; McArdle and Anderson, 2004; Martin *et al*., 2005; Warton, 2005). Even if approximate normality is indicated by goodness-of-fit tests on the transformed data, if the data come from some other distribution than the normal then the significance tests may be misleading. For instance, the Chi-square test of normality is a non-specific test, in that the test criterion is directed against no particular type of departure from normality (Snedecor and Cochran, 1989). Examples occur in which the data are noticeably skew, although the goodness-of-fit test does not reject the null hypothesis. For small sample sizes, power of test is also low for detecting larger departures from normality that may be important. It is only with larger sample sizes that increasingly smaller departures from normality can be detected (Snedecor and Cochran, 1989).

The study has also demonstrated that transformation of either abundance or incidence data do not necessarily ensure homogeneity of variances, and that transformation functions differed in their ability to ensure homogeneity. Close scrutiny of the tests of homogeneity of variance revealed that the two tests differed in their sensitivity in detecting variance heterogeneity in abundance and incidence. It is well known that ANOVA is less robust to violations of homogeneity of variance than normality. Homogeneity of variance is essential for the valid application of parametric ANOVA. A transformation used to normalize the data may lead to heterogeneity of variance. This is because one transformation might be best for ensuring homogeneity of variance, while another might be best for ensuring normality. In practice, only one of these two transformations can be used, so all the statistical requirements cannot be met with linear models (Garrett *et al.*, 2004). Transforming the data to rectify the problem can result in apparently grossly inflated type I errors, altering the model under test and affect the spatial scale of the hypothesis (McArdle and Anderson, 2004). Adding 1 to the zero counts during logarithmic transformations can also result in strange distributions, which has led some workers to model the zeros separately for count data (McArdle and Anderson, 2004; Martin *et al*., 2005). Among the data transformation functions used in this study, logarithmic transformation gave better description of abundance data compared with square root. Working logits were better than angular or square root transformation of incidence data. The study has demonstrated that the choice of transformation can influence the statistical significance and power of test. However, during statistical analyses, researchers all too often ignore the assumptions, transform the data and then fail to evaluate whether the transformation corrected the problem (McArdle and Anderson, 2004). To test for homogeneity variances, the Bartlett's and Levene's tests are often used. However, as indicated by the results in Table 2 the sensitivity of these tests differ. While Bartlett's test has accurate Type I error rates and optimal power when the underlying distribution of the data is normal, it can be very inaccurate if the distribution is even slightly nonnormal (Box, 1953).

Researchers some times use nonparametric methods as alternatives to parametric tests for analyses of abundance and incidence when the data violate the assumptions of ANOVA (Sileshi and Mafongoya, 2002; 2003). Until recently (Brunner and Puri, 2001; Turecheck, 2004) the use of nonparametric approach had been limited because these tests are less powerful than parametric methods. Secondly, they could only be used in one-way analysis as there had been no satisfactory theoretic foundation for analysing data in factorial designs and repeated measures (Shah and Madden, 2004). Unlike parametric ANOVA and nonparameteric tests, GLMs enable appropriate analyses of skewed frequency or binary data. In addition, with GLMs, the properties of data from discrete distributions such as the Poisson and negative binomial distribution (counts) and binomial distribution (proportions) can be accounted for (Hughes and Madden, 1995; Collett, 2002). For example, the GLMs used in this study tested whether the abundance distribution was random (Poisson) or spatially aggregated (negative binomial). The GLM also demonstrated that the negative binomial model is considerably more robust for analysis of the abundance data compared with the LMM or the Poisson (Table 7). Using the GLMs it was possible to simultaneously consider the effect of treatments and variance heterogeneity.

While common parametric approaches, such as ANOVA are well known and convenient, their assumptions may not always be met in contexts studied by plant pathologists, entomologists and weed biologists. For example, if ANOVA shows lack of statistical significance, it may be because there is no effect or because the study design makes it unlikely that a biologically real effect would be detected. When the sample size is small and variance is high as is common in abundance and incidence data, biologically interesting phenomena may be missed because ANOVA is unlikely to yield significant results (e.g. Tables 5 and 6). Under such situations computation of statistical power is as important as significance testing. Power analysis can

distinguish between these alternatives, and is therefore a critical component of designing experiments and testing results (Thomas and Krebs, 1997). For abundance and incidence data, LMMs and GLMs offer tremendous opportunities for improvement of statistical inference. Just as standard ANOVA has been expanded to LMMs, recent research has expanded GLMs to generalized linear mixed models (GLMMs) (Garrett *et al*., 2004). While biologists have traditionally stressed hypothesis testing as a statistical approach, emphasis has shifted in recent years towards information theoretic approaches (Burnham and Anderson, 2002). Information criteria such as AIC provide a more objective way of determining which model among a set of models is most appropriate for analyses of the data at hand. Often one has no *a priori* reason for selecting a specific data transformation to normality. The AIC may be used as a potentially valuable tool for selecting functions for data transformation. The major limitation in using the methods described is that they are computationally intensive. However, software that handle such computations with relative ease are appearing.

5. Acknowledgements

I am grateful to the Canadian International Development Agency (CIDA), Swedish International Development Agency (SIDA) and World Agroforestry Centre (ICRAF) for their financial support for this work.

6. References

Anscombe, F.J. 1949. The analysis of insect counts based on the negative binomial distribution. *Biometrics* 5: 165-173.

Anscombe, F.J. 1950. Sampling theory of the negative binomial and logarithmic series distributions. *Biomentriaca* 37: 358-382.

Box, G.E. 1953. Non-normality and tests on variance. *Biometrika* 40: 318 - 335.

Brunner, E. and Puri M.L. 2001. Nonparametric methods in factorial designs. *Statatistics Papers* 42: 1-52.

Burnham, K. P. and Anderson, D. R. 2002. *Model Selection and Multimodel Inference: a practical information-theoretic approach*, 2nd edition. Springer-Verlag, New York.

Cochran, W.G. 1947. Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics* 3: 22–38.

Collett D. 2002. *Modelling Binary Data.* 2nd edition. CRS Press, Boca Raton, FL.

Cox, D.R. 1970. *Analysis of Binary Data.* Chapman and Hall, London.

Fisher, R.A. 1935. *Design of experiments*. Oliver and Boyd, Edinburgh, UK.

Fletcher, D., MacKenzie, D. and Villouta, E. 2005. Modelling skewed data with many zeros: a simple Approach combining ordinary and logistic regression. *Environmental and Ecological Statistics* 12: 45-54.

Garrett, K.A., Madden, L. V., Hughes, G. and Pfender, W.F. 2004. New applications of statistical tools in plant pathology. *Phytopathology* 94: 999-1003.

Gaston, K., Blackburn, T.M., Greenwood, J.J.D., Gregory, R. Quinn, R.M. and Lawton, J.H. 2000. Abundance-occupancy relationships. *Journal of Applied Ecology* 37: 39-59.

Hartley, H.O. and Rao, J.N.K. 1967. Maximum likelihood estimation for mixed analysis of variance models. *Biometrika* 54: 93-108.

Harville, D.A. 1977. Maximum likelihood approaches to variance component estimation and to related problems. *Journal of American Statistical Association* 72: 320-340.

Hughes, G., and Madden, L. V. 1995. Some methods allowing for aggregated patterns of diseases incidence in the analysis of data from designed experiments. *Plant Pathology* 44: 927-943.

Hurvich, C.M. and Tsai, C.L.1989 Regression and time series model selection in small samples. *Biometrika* 76: 297-307.

Johnson, N.I. and Kotz, S. 1969. *Discrete Distributions*. Houghton Mifflin Company, Boston.

Littell, R.C. 2002. Analysis of unbalanced mixed model data: A case study comparison of ANOVA versus REML/GLS. *Journal of Agricultural, Biological and Environmental Statistics* 7: 472-490.

Madden, L. V., and Hughes, G. 1995. Plant disease incidence: Distributions, heterogeneity, and temporal analysis. *Annual Review of Phytopathology* 33: 529-564.

Madden, L.V., Turechek, W.W. and Nita, M. 2002. Evaluation of generalized linear mixed models for analyzing disease incidence data in designed experiments. *Plant Disease* 86: 316-325.

Martub, T.G., Wubtek, B.A., J.R., Kuhnert,Field,P.M., S.A., Low-Choy, S.J., Tyre,A., Possingham, H.P. 2005. Aero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecology Letters* 8: 1235-1246.

McArdle, B.H and Anderson, M.J. 2004. Variance heterogeneity, transformations, and models of species abundance: a cautionary tale. *Canadian Journal of Fisheries and Aquatic Sciences* 61: 1294-1302.

McCullagh, P. and Nelder, J.A. 1989. *Generalized linear models*, 2nd edition , Longo, Chapman and Hall.

McRoberts, N., Hughes, G. and Madden, L.V. 1996. Incorporating spatial variability into simple disease progress models for crop pathogens. *Aspects of Applied Biology* 46: 1-8.

Peopho, HP., Büchse,A. and Emrich, K. 2003. A hitchhiker's guide to mixed for randomized experiments. *Journal of Agronomy and Crop Science* 189: 310-322.

SAS Institute Inc. 2003. *SAS/STAT*, Release 9.1, Cary, NC: SAS Institute Inc.

Shah, D.A. and Madden, L.V. 2004. Nonparametric analysis of ordinal data in designed factorial experiments. *Phytopathology* 94: 1022-1026.

Saha,K. and Paul, S. 2005. Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics* 61: 179–185.

Sileshi, G. and Mafongoya, P.L. 2002. Incidence of insect pests in mixed species fallows with special emphasis on *Mesoplatys ochroptera* Stål (Coleoptera: Chrysomelidae) on *Sesbania sesban* in eastern Zambia. *Agrofoerey Systems* 56: 225-231.

Sileshi, G., Baumgaertner, J., Sithanantham,S. and Ogol, C.K.P.O. 2002. Spatial distribution and sampling plans for *Mesoplatys ochroptera* Stål (Coleoptera: Chrysomelidae) on sesbania. *Journal of Economic Entomology* 95: 499-506.

Sileshi, G. and Mafongoya, P. L. 2003. Effect of rotational fallows on abundance of soil insects and weeds in maize crops in eastern Zambia. *Applied Soil Ecology* 23: 211-222.

Sileshi, G., Mafongoya, P. L., Kwesiga F. and Nkunika, P. 2005. Termite damage to maize grown in agro forestry systems, traditional fallows and monoculture on Nitrogen-limited soils in eastern Zambia. *Agricultural and Forest Entomology* 7: 61-69.

Sileshi, G., Girma, H. and Mafongoya, P.L. 2006a. 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

Sileshi, G., Mafongoya, P.L. and Kuntashula, E. 2006b.The effect of agro forestry practices on parasitic and arable weeds of maize in Zambia. *Zambia Journal of Agriculture* 9: (in press)

Snedecor, G.W. and Cochran, W.G. 1989. *Statistical* Methods, 8th edition, Iowa State University, Ames.

Taylor, L.R. 1961. Aggregation, variance and the mean. *Nature* 189: 732-735.

Thomas, L. and Krebs, C.J. 1997. A review of statistical power analysis software. *Bulletin of the Ecological Society of America* 78: 126-139.

Turecheck, W.W. 2004. Nonparametric tests in plant disease epidemiology: Characterizing disease associations. *Phytopathology* 94: 1018-1021.

Turecheck, W.W. and Madden, L.V. 2002. A generalized linear modelling approach for characterizing disease incidence in spatial hierarchy. *Phytopathology* 93: 458-466.

Warton, D. I. 2005. Many zeros does not mean zero Inflation : comparing the goodness-of-fit of parametric models to multivariate abundance data. *Environ metrics* 16: 275-289.

Wolfinger, T.D. 1993. Covariance structure selection in general mixed models. *Communications in Statistics-Simulation and Computation* 22: 1079-1106.