Available online at www.sciencedirect.com

SCIENCE DIRECT®

Forest Ecology and Management 221 (2006) 123–132

Forest Ecology and Management

www.elsevier.com/locate/foreco

A better alternative to Wald's test-statistic for simple goodness-of-fit tests under one-stage cluster sampling

Steen Magnussen [a,*], Michael Köhl [b]

[a] *Natural Resources Canada, Canadian Forest Service, 506 West Burnside Road,*
Victoria, BC, Canada V8Z 1M5
[b] *University of Hamburg, Department of Wood Science, Section World Forest,*
Leuschnerstr. 91, D-21031 Hamburg, Germany

Received 16 May 2005; received in revised form 12 September 2005; accepted 13 September 2005

Abstract

Significance levels of the popular Wald's Chi-squared statistic for simple goodness-of-fit (GOF) tests under one-stage cluster sampling are often unreliable. A large number of alternatives to Wald's GOF test with Type-I error rates more closely matching the nominal level of significance have been proposed but not yet found their way into applied statistics. Type-I error rates with Wald's test-statistic in cluster sampling from 10 actual forest cover-type maps from 5 sites and 81 sample designs are compared to the error rates of 11 alternatives. The effects of site, sampling design, evenness of cover-type class proportions, and intra-cluster correlation on Type-I error rates are quantified with logistic regressions for Wald's statistic and five promising alternatives. Our proposed second-order bias correction of Finney's [Finney, D.J., 1971. Probit Analysis, vol. 3. Cambridge University Press, p. 350] and Brier's [Brier, S.S., 1980. Analysis of contingency tables under cluster sampling. Biometrika 67, 591–596] method of moments correction of Pearson's Chi-squared test statistic emerged as the overall best alternative in this study. It was the least sensitive to design and cluster effects. Test power was investigated for the alternative simple hypothesis of equality of cover-type proportions in two site-specific maps. The proposed alternative test statistic had slightly (3%) less power than Wald's test for designs with a power of 80% or greater, yet a consistently better odds ratio of a correct test decision.

© 2005 Elsevier B.V. All rights reserved.

Keywords: Chi-squared statistic; Type-I error rate; Type-II error rate; Test power; Design-effect; Intra-cluster correlation; Method of moments; Bias correction; Logistic regression; Forest cover-types

1. Introduction

The goodness-of-fit (GOF) test of a simple hypothesis about a categorical attribute is frequently entertained in forest ecology and management. The test arises naturally when one wish to compare a sample-based estimate of the distribution of class values of a categorical variable against, for example, a benchmark distribution. Tests concerning distributions of age-classes, forest cover-types, soil-types, and land-use are but a few examples.

Categorical forest attribute data are often obtained from sampling (inventory). For practical and cost reasons, it is customary to make several observations at each sample location (Shiver and Borders, 1996). A sample plot with more than one observational unit is called a cluster in statistical terminology (Cochran, 1977). In forestry a unit can be, for example, a tree, a specified portion of a plot, or a sub-plot. Units within a cluster often tend to be more similar in terms of class attribute values than units sampled at random (Ridout et al., 1999). When this is the case, we say that there is a positive intra-cluster correlation of class attribute values (Gilliland et al., 2002; Hall and Severini, 1998; Stoner and Leroux, 2002). A statistical consequence of a positive intra-cluster correlation (cluster effect) is that the variation among clusters with, say, m units in each cluster is larger than the variance among groups of m randomly selected units. Note, the intra-cluster correlation is class-specific and may vary among classes. A positive intra-cluster correlation may or may not be of consequence in the context of statistical testing and inference (Legendre, 1993). It depends on the hypothesis and the assumptions accompanying the statistical test (Dale and Fortin, 2002; Gregoire, 2004).

* Corresponding author. Tel.: +1 250 363 0712; fax: +1 250 363 0775.
E-mail address: Steen.Magnussen@nrcan.gc.ca (S. Magnussen).

0378-1127/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.foreco.2005.09.013

The most popular GOF test statistics are Pearson's Chi-squared statistic (χ^2), the likelihood ratio statistic (G^2), and Wald's generalized (χ^2_{W}) statistic (Agresti, 1992; Lloyd, 1999). It is well known that both χ^2 and G^2 are seriously inflated by "cluster effects" when data come from a sample design employing clusters or strata (Bedrick, 1983; Cerioli, 2002a, 1997; Clifford et al., 1989; Cohen, 1976; Holt et al., 1980; Rao and Scott, 1981). An inflated test statistic increases the risk of falsely rejecting a null hypothesis (Type-I error rate). The inflation of the Type-I error rate will depend on the class-specific intra-cluster correlation coefficients (Rao and Scott, 1981; Cerioli, 2002b).

Wald's Chi-squared test statistic generalizes readily to a complex sampling design (Koch et al., 1975). However, it relies on asymptotic expectations and a design-consistent estimator of the covariance matrix under the null hypothesis. In practice sample sizes may be too small to justify asymptotic expectations, and a sample-based estimate of the covariance matrix is used in place of a design-consistent estimate of the covariance matrix under the null hypothesis (Rao and Scott, 1981). Wald's test-statistic is, therefore, quite sensitive to design settings (sample size, cluster size) and data characteristics such as the number of classes, the proportions themselves, and the intra-cluster correlation. Typically the false rejection rate increases rapidly with the number of classes for a fixed cluster and sample size (Thomas et al., 1997). As a consequence, the analyst should always be careful when drawing inference from a Wald's statistic (Agresti and Caffo, 2000; McCullagh and Nelder, 1989; Pawitan, 2000).

Several proposed corrections and transformations of χ^2 and G^2 have a lower Type-I error rate than Wald's test-statistic (Brier, 1980; Fay, 1979; Fellegi, 1980; see Miao and Gastwirt, 2004 and references therein; Thomas and Rao, 1987). A first- and a second-order correction of χ^2, and G^2 for 'cluster effects' achieved Type-I error rates much closer to the nominal levels of significance than possible with χ^2_{W}. Thomas et al. (1997) confirmed these results in an extensive comparison of 16 test statistics. Interestingly, the intuitively simple and appealing method of moments correction of χ^2 first proposed by Finney (1971) and later by Brier (1980) still awaits a formal evaluation.

Despite documented problems with χ^2_{W} and the availability of potentially better alternatives, Wald's test-statistic remains popular. A recent scan of over 2000 ecological journal articles published in 2004 found no less than 36 applications of Wald's GOF test under one-stage cluster sampling. One can be tempted to surmise that the number of reported significant results may have been skewed. The alternatives do not seem to have found their way into mainstream applied statistics yet.

Given the widespread use of one-stage cluster sampling and the importance of GOF hypothesis testing in forest ecology and forest management, we think it is timely to demonstrate again the problems with Wald's test and to suggest a better alternative. We narrow the choices for the applied statistician by conducting an extensive assessment of 11 alternative GOF test statistics under the null hypothesis and data from cluster sampling of forest cover-types in five sites. Finney's (1971) and Brier's (1980) method of moments correction and our proposed second-order bias correction of their test statistic are included. Since a test statistic must also exhibit a good power to reject the null hypothesis when it is false, we also assessed test power (Lehmann, 1983). The GOF test statistic that improved the odds of making a correct decision from a simple hypothesis test at a nominal significance level $\alpha = 0.05$ is recommended for practical use.

2. Material and methods

2.1. Sample data

Ten forest cover-type maps, two from each of five sites, were used to simulate one-stage simple random cluster sampling of forest cover-types. Each cover-type map was completely tessellated into a regular array of N 30 m × 30 m units, each assigned to a specific cover-type. Cover-type maps from a single site were the result of a classification of a Landsat TM image or an interpretation of aerial photography. Two maps from a single site would differ due to the method of classification or due to temporal land-use changes. Sites are referred to as A, B, C, D, and E while map types within a site are labelled I and II, respectively.

Site A – near Prince George British Columbia (Canada) – cover-type data are from a classification of $N = 121{,}104$ units in a 348 × 348 array (10,899 ha) of Landsat TM image pixels from 1990 (map A.I) and 1999 (map A.II) to 15 cover-type classes (Wulder et al., 2002).

Site B – near Hinton, Alberta (Canada) – data are from a classification of $N = 129{,}600$ units in a 360 × 360 array (11, 664 ha) of Landsat TM image pixels from 1985 (map B.I) and 1990 (map B.II) to 16 cover-type classes (Goodenough et al., 2000).

Site C – near Latium, Viterbo (Italy) – data are from a photo-interpretation (map C.I) and a classification of $N = 119{,}133$ units (10,720 ha) in an array of Landsat-7 ETM+ image pixels (map C.II) to six cover-types classes (Corona et al., 2002).

Site D – near Sussex, New Brunswick (Canada) – data are from a photo-interpretation (map D.I) and a classification (map D.II) of $N = 135{,}806$ units (12,223 ha) Landsat TM pixels to 11 cover-type classes (Magnussen et al., 2004).

Site E – near Kuala Lumpur (Malaysia) – data are from a 1989 (map E.I) and a 1999 (map E.II) classification of $N = 166{,}467$ units (14,982 ha) in a Landsat TM image to eight cover-type classes (Suratman et al., 2004).

Site specific cover-type classes were collapsed to simplified cover-type maps with three, four, or five broadly defined cover-types. Thus, a total of 5 × 2 × 3 = 30 cover-type maps were available for simulated one-stage cluster sampling.

2.2. Sample strategy

Simple random one-stage cluster sampling with n plots (clusters) composed of a square array of m units (m = 9, 16, 25, or 36) was simulated for each cover-type map. Sampling on sites A, B, and E was with plots of size 9, 16, and 25 (Magnussen et al., 2004) while plots of size 9, 16, and 36 were employed in sites C and D (Magnussen, 2004).

Sample sizes (n) were 20, 40, ..., 100, 150, ..., 300. Sampling under each design (s) of plot size (m) × sample size (n) × number of cover-type classes, $K = (3, 4, 5)$ was repeated 2000 times. A total of $3(K) \times 3(m) \times 9(n) = 81$ designs were employed for each of the 5 × 2 cover-type maps to a total of 810 designs.

For the ith sample plot ($i = 1, \ldots, n$) the vector $\mathbf{y}_i = \{y_1, y_2, \ldots, y_K\}$ of cover-type class counts was noted ($K = 3, 4, 5$). A sample-based estimate of the cover-type class proportions under design s was obtained as:

$$\hat{\mathbf{p}}_s = \frac{1}{n_s}\sum_{i=1}^{n_s}\mathbf{p}_i = \frac{1}{n_s \times m_s}\sum_{i=1}^{n_s}\mathbf{y}_i \qquad (1)$$

where $\mathbf{p}_i$ is the K length vector of class proportions in the ith plot. Suffixes identifying sites (A, ..., E) and maps (I and II) have been suppressed for notational convenience. The sample-based estimate of the covariance matrix $\hat{\mathbf{\Sigma}}_s$ of $\hat{\mathbf{p}}_s$ was obtained by standard calculus (Lehmann, 1983) and adjusted for finite population size by a factor $f_s = 1 - n_s \times m_s \times N^{-1}$.

Map-wide cover-type class proportions $\mathbf{p}_s$ for $K = 3$, 4, and 5 are in Table 1. For $K = 3$, the most prevalent cover-type occupies between 39% and 69% of a study site, and the least common type between 3% and 23%. With $K = 5$, the corresponding figures were 21–67% and 1–17%, respectively. Within-site class-specific map differences averaged 5.2% (range 0–11%).

2.3. Test statistics

A simple GOF test under the true null hypothesis $H_0 : \hat{\mathbf{p}}_s = \mathbf{p}_s$ was carried out for each of the 1,620,000 estimates of $\mathbf{p}_s$ (10 maps × 81 design × 2000 replicates).

Wald's test-statistic (χ^2_{W}) and 11 alternatives were computed for each hypothesis test. Computational details are only provided for χ^2_{W} and the five most promising alternatives (see below). Nine of the 11 alternatives have been detailed, among others, by Rao and Scott (1981), Thomas and Rao (1987), and Thomas et al. (1997). The nine previously tested alternatives were: G_F an F-ratio transform of the likelihood ratio test statistic G^2, G^2_{JK} a jackknifed estimate of G^2, F_{W} an F-ratio transform of Pearson's Chi-squared test statistic χ^2, χ^2_{RS1} a first-order Rao–Scott correction of χ^2, χ^2_{RS2} a second-order correction of χ^2, F_{RS1} an F-ratio transform of χ^2_{RS1}, F_{RS2} an F-ratio transform of χ^2_{RS2}, χ^2_F Fellegi's Chi-squared statistic, and χ^2_{RW} Singh's robust version of χ^2_{W}. The last two have not been formally evaluated. They were χ^2_{C} Finney's (1971) and Brier's (1980) method of moments corrected Chi-squared statistic, and χ^2_{BC} our proposed second-order bias correction of χ^2_{C}.

Computational details are only provided for Wald's statistic and five of the most promising alternatives (see below). Wald's GOF test statistic (Wald, 1941) is:

$$\hat{\chi}^2_{\mathrm{W}} = n_s \times (\hat{\mathbf{p}}^*_s - \mathbf{p}^*_s)'\hat{\mathbf{\Sigma}}_s^{-1}(\hat{\mathbf{p}}^*_s - \mathbf{p}^*_s) \qquad (2)$$

where $\mathbf{p}^*$ denotes the first $K - 1$ elements of $\mathbf{p}$ and $\hat{\mathbf{\Sigma}}_s^{-1}$ denotes the inverse of the sample-based estimate of the covariance matrix of $\mathbf{p}^*_s$. Under the null and a consistent estimate of the covariance matrix $\hat{\chi}^2_{\mathrm{W}}$ is asymptotically distributed as a Chi-squared random variable with $K - 1$ degrees of freedom.

To correct for 'cluster effects' a first-order correction factor λ to χ^2 has been proposed by several authors:

$$\hat{\chi}^2_{\mathrm{RS1}} = \hat{\chi}^2_s \times \hat{\lambda}^{-1} \quad \text{and} \quad \hat{\lambda} = \mathrm{tr}(\hat{\mathbf{P}}_s^{-}\hat{\mathbf{\Sigma}}_s) \times (K-1)^{-1} \qquad (3)$$

where $\hat{\chi}^2_s$ is Pearson's Chi-squared test statistic

$$\hat{\chi}^2_s = n_s \times m_s \times \sum_{k=1}^{K}\frac{(\hat{p}_k - p_k)^2}{p_k} \qquad (4)$$

Table 1
Cover-type proportions (%) in site specific (A–E) cover-type maps (I and II)

	A.I	A.II	B.I	B.II	C.I	C.II	D.I	D.II	E.I	E.II
$K = 3$										
1	39	50	65	67	63	63	69	63	62	53
2	38	28	18	16	34	29	21	20	23	14
3	23	23	17	17	3	7	10	18	15	33
$K = 4$										
1	39	50	65	67	62	59	51	43	42	41
2	23	23	17	17	34	29	21	20	23	14
3	22	19	15	7	3	7	18	20	20	12
4	15	8	3	9	1	4	10	18	15	33
$K = 5$										
1	23	26	65	67	62	59	27	26	27	21
2	23	23	15	7	34	29	24	17	23	14
3	22	8	10	11	2	7	21	20	20	12
4	16	24	7	5	1	4	18	20	15	33
5	15	19	3	9	1	1	10	18	14	21

All entries have been rounded to the nearest integer percent value.

$\hat{p}_k$ is the estimated proportion of the kth class, and $\hat{\mathbf{P}}_s = D(\hat{\mathbf{p}}_s) - \hat{\mathbf{p}}_s\hat{\mathbf{p}}_s'$ the multinomial covariance matrix of $\hat{\mathbf{p}}_s$, and tr is the trace operator yielding the sum of the elements on the diagonal in a square matrix. $\hat{\lambda}$ is an estimate of the average (across classes) 'cluster effect'. Under H_0 and no among-class variation in the intra-cluster correlations $\hat{\chi}^2_{\mathrm{RS1}}$ is distributed asymptotically as a Chi-squared random variable with $K - 1$ degrees of freedom.

Fellegi (1980) proposed an F-ratio transform of the Chi-squared GOF test statistic. F-ratio transforms of Chi-squared statistics are more conservative (fewer rejections of the null hypothesis) especially for smaller sample sizes ($n < 80$) and $K > 3$. Application of the F-transform to $\hat{\chi}^2_{\mathrm{RS1}}$ yields:

$$\hat{F}_{\mathrm{RS1}} = \frac{n_s - K + 1}{(K-1)(n_s - 1)} \times \hat{\chi}^2_{\mathrm{RS1}} \tag{5}$$

where under H_0 $\hat{F}_{\mathrm{RS1}}$ is asymptotically distributed as an F-ratio variable with $K - 1$ and $n_s - K + 1$ degrees of freedom, respectively.

Holt et al. (1980) and Rao and Scott (1981) argued for a second-order adjustment of λ to capture the effect of variation among classes in the intra-cluster correlations. Their Satterthwaite-type correction yields:

$$\hat{\chi}^2_{\mathrm{RS2}} = \hat{\chi}^2_{\mathrm{RS1}} \times (1 + \hat{a}^2)^{-1} \tag{6}$$

where

$$\hat{a}^2 = \frac{\sum_k \sum_{k'} \hat{\sigma}_{kk'} (\hat{p}_k \hat{p}_{k'})^{-1} - K \times \hat{\lambda}^2}{(K-1)\hat{\lambda}^2} \tag{7}$$

where $\hat{\sigma}_{kk'}$ is the sample estimate of the covariance of p_k and $p_{k'}$ (i.e. the elements of $\hat{\boldsymbol{\Sigma}}_s$). Under H_0 $\hat{\chi}^2_{\mathrm{RS2}}$ is asymptotically distributed as a Chi-squared random variable with ν degrees of freedom with $\nu = Rnd[(K-1)/(1+\hat{a}^2)]$.

Finney (1971) and later Brier suggested a simple intuitively appealing method of moments correction of Pearson's Chi-squared statistics $\hat{\chi}^2_C = \hat{\chi}^2 \times \hat{C}^{-1}$ with

$$\hat{C} = \frac{1}{(n_s - 1)(K - 1)} \sum_{i=1}^{n_s} \frac{(\mathbf{y}_i - m_s \times \hat{\mathbf{p}}_s)^2}{m_s \times \hat{\mathbf{p}}_s} \tag{8}$$

Intuitively appealing because C is the average Chi-squared statistic for testing equality of the n_s cluster-specific probability vectors $\mathbf{p}_i$ divided by the degrees of freedom ($K - 1$). In absence of cluster effects the expected value of C is one. C is, asymptotically, free of any unknown parameters. The allowed range for $\hat{C}$ is between 1 and m_s, values of $\hat{C}$ outside this range are truncated to the nearest allowed value.

A second-order Taylor-Series expansion of (8) around $m_s \times \hat{\mathbf{p}}_s$ will show that the correction factor C is downward-biased. We, therefore, propose the test statistic:

$$\hat{\chi}^2_{\mathrm{BC}} = \hat{\chi}^2 \times (\hat{C} + \hat{\omega})^{-1} \quad \text{with} \quad \hat{\omega} = \frac{1}{K}\sum_{k=1}^{K} \frac{(\hat{p}_k^2 + \hat{\sigma}_k^2)\hat{\sigma}_k^2}{n_s \times m_s \times \hat{p}_k^2} \tag{9}$$

Again, the allowed range for $\hat{C} + \hat{\omega}$ is also between 1 and m_s, with values outside of this range truncated as above. Note, a bootstrap estimation of C would also correct for this bias (Efron and Tibshirani, 1993).

2.4. Type-I error

The performance of each test statistic is measured in terms of the Type-I error rate, viz. the rate of rejection ($\hat{\alpha}_s$) of a true H_0 at a nominal significance level of α. Here, $\alpha = 0.05$ as it remains the most common level in applied statistics. For each test statistic and the 810 design combinations we counted the number of times the true null hypothesis was falsely rejected in 2000 replications of a sample design (#reject) and computed $\hat{\alpha}_s = \text{reject} \times 2000^{-1}$. With 2000 replications, the standard error of $\hat{\alpha}_s$ is approximately 0.005 (Serfling, 1980). Results with $\alpha = 0.01$ and $\alpha = 0.10$ were very similar (not shown).

A ranking of the observed error rates of the 12 test statistics was obtained for each of the 810 designs. The ranking criterion was $\mathrm{Abs}[0.05 - \hat{\alpha}_s]$. A test statistic with a ranking of 6 or higher in 75% of the 810 design settings was eliminated from further consideration. Only the above five alternative test statistics were retained after this screening. Wald's statistic would have been eliminated by this screening, but it was retained for reason of comparison.

2.4.1. Predicting Type-I error rates

The Type-I error rate of an ideal test statistic under the null should closely match the nominal significance level α and it should not depend on 'nuisance' effects like sample size, plot size, number of classes, 'site' factors, or 'cluster effects'. To summarize the 'nuisance' effects on observed Type-I error rates we fitted a logistic regression model (Hosmer and Lemeshow, 1980) with $\hat{\alpha}_s$ as the dependent variable and an intercept (constant), design factors (n_s and m_s), an index of evenness of $\mathbf{p}_s$, and the average intra-cluster correlation as independent predictors. It is known that Chi-squared tests are sensitive to the 'evenness' of class proportions. We used Simpson's index of evenness of proportions (Patil, 1982) which is computed as $\mathbf{p}_s'(1 - \mathbf{p}_s)$. Simpson's index reaches a maximum when all class proportions are equal and a minimum when class proportions tend towards the extremes of 1 and 0. Simpson's index is restricted to values between 0 and 1.

A logistic regression model for an ideal test statistic would have an intercept of $\log(0.05) = -2.99573$ and no statistically significant 'nuisance' effects. A comparison of estimated logistic regression coefficients allows us to quantify the sensitivity of a test statistic to 'nuisance' factors. In this regression context, the two cover-type maps per site are regarded as within-site replications.

2.5. Type-II errors

In practice the analyst will, of course, not know whether a simple test hypothesis is true or not. Provided the test statistic is consistent the analyst controls Type-I errors through the choice of α. It is, of course also important not to accept a hypothesis when it is false (Type-II error). For a given design and significance level the best test statistic is the one that minimizes the Type-II error rate β while keeping the Type-I error rate at the nominal level. A minimum Type-II error rate provides maximum test power $(1-\beta)$ for discrimination between hypotheses at a fixed α.

To evaluate β, we formulated the alternative simple hypothesis H_1 stating that a sample-based estimate of cover-type proportions for map I was equal to the known cover-type proportions of map II (and vice versa). Tests of this nature arise naturally in applied settings where one might, for example, compare estimates from remotely sensed images to estimates from ground-based sampling. Here, the alternative hypothesis H_1 is false (Table 1). The Type-II error rate β of a test statistic is the number of times H_1 is accepted divided by the number of replicate tests of the hypothesis.

2.5.1. *Odds of a correct test decision*

For a given α and a given sample design, a test statistic presents a trade-off between Type-I and -II errors. The decision to reject or accept a hypothesis is made on the significance of the test statistic. The chance of a correct acceptance or rejection of a hypothesis is captured by the (log) odds ratio of a correct decision ψ (Fleiss, 1981)

$$\psi = \log\left(\frac{\begin{array}{c}\text{correct acceptance of } H_0 + \\ \text{correct rejections of } H_1\end{array}}{\begin{array}{c}\text{false acceptance of } H_0 + \\ \text{false rejections of } H_1\end{array}}\right) \qquad (10)$$

where the logarithm provides a convenient scaling. We computed ψ for all 810 design settings.

3. Results

Site and map specific average Type-I error rates ($\hat{\alpha}_s$) of Wald's statistic and the five best alternatives are in Table 2. For Wald's statistic, the error rate was significantly above the nominal 5% level on all sites ($P < 0.05$). In addition, the variability of Type-I error rates obtained with Wald's statistics is an order of magnitude larger than expected from random variation due to sampling, and also about four to five times larger than the variability observed with the five best alternatives. Strongly inflated error rates were concentrated in designs with small sample sizes and five cover-types.

Type-I error rates of the five best alternatives to Wald's GOF test statistics are clustered much more closely around the nominal level of 5%. Not only are the rates much closer to 5% but their variability is also substantially less. An F-transform of $\hat{\chi}^2_{\mathrm{RS1}}$ improves slightly the performance of the first-order correction of Pearson's Chi-squared statistic. A second-order correction (χ^2_{RS2}) generates a conservative test with fewer rejections than at the nominal significance level. Finney's (1971) and Brier's (1980) method of moments correction of Pearson's Chi-squared test statistics appears overall attractive in terms of average absolute deviation from α, consistency across sites (maps), and a low variability of error rates. Our proposed bias-correction of this statistic (χ^2_{BC}) mostly improves performance. An improvement that was approximately equal to the effect of an F-transform of a Chi-squared test statistic. Type-I error rates of χ^2_{BC} were most often (52%) the alternative closest to the nominal level and only occasionally (9%) the alternative furthest away. Only one in seven of the observed error rates for χ^2_{BC} differed significantly from the nominal level. The corresponding figure for χ^2_{W} was two in three. The runner up to χ^2_{BC} was χ^2_{C} with a ratio of one in five. χ^2_{BC} is not uniformly better than χ^2_{W}. However, in the 14% of the 810 test scenarios where the Type-I error rate of Wald's was closer to the nominal level their differences were trivial (average 0.3%) and never exceeded 0.5%. A graphical summary of the Type-I error rates in Fig. 1 captures the erratic performance of Wald's test and the improved performance of the alternatives, in particular that of χ^2_{BC}.

The sensitivity of Type-I error rates to 'nuisance' effects is captured by the logistic regression coefficients in Table 3. An ideal test statistic would show no sensitivity to these factors. All error rates were significantly influenced by 'site' but $\hat{\alpha}_s(\hat{\chi}^2_{\mathrm{C}})$ and $\hat{\alpha}_s(\hat{\chi}^2_{\mathrm{BC}})$ were the least affected with an average absolute site coefficient of just 0.06 compared to 0.15 for Wald's, and a surprisingly high average of about 0.5 for $\hat{\chi}^2_{\mathrm{RS1}}$, $\hat{\chi}^2_{\mathrm{FRS1}}$, and $\hat{\chi}^2_{\mathrm{RS2}}$. Effects of K were non-significant for $\hat{\alpha}_s(\hat{\chi}^2_{C})$ and $\hat{\alpha}_s(\hat{\chi}^2_{\mathrm{BC}})$ but highly significant for the others. Type-I error rates of $\hat{\chi}^2_{W}$, $\hat{\chi}^2_{\mathrm{RS1}}$, $\hat{\chi}^2_{\mathrm{FRS1}}$, and $\hat{\chi}^2_{\mathrm{RS2}}$ increased at a rate of about 20–40% for every increase in the number of cover-type

Table 2
Average Type-I error rates ($\hat{\alpha}_s$) in % of Wald's statistic and the best five alternatives across 81 sample designs in five sites (A–E) and two cover-type maps per site (I and II)

Site.Map	$\hat{\chi}^2_{\mathrm{W}}$	$\hat{\chi}^2_{\mathrm{RS1}}$	$\hat{F}_{\mathrm{RS1}}$	$\hat{\chi}^2_{\mathrm{RS2}}$	$\hat{\chi}^2_{\mathrm{C}}$	$\hat{\chi}^2_{\mathrm{BC}}$
A.I	7.5 (3.7)	5.5 (0.8)	5.2 (0.7)	4.4 (0.6)	5.6 (0.9)	5.3 (0.7)
A.II	7.6 (3.8)	5.7 (1.0)	5.4 (0.8)	4.5 (0.6)	5.7 (0.9)	5.4 (0.7)
B.I	11.7 (7.7)	5.9 (1.6)	5.6 (1.4)	4.9 (1.0)	4.6 (0.5)	4.5 (0.6)
B.II	11.4 (8.4)	6.3 (1.6)	6.0 (1.4)	4.7 (1.0)	5.3 (0.8)	5.0 (0.7)
C.I	7.4 (4.2)	6.0 (3.0)	5.7 (2.8)	5.4 (2.3)	5.4 (1.1)	5.0 (1.0)
C.II	9.0 (6.0)	4.3 (2.0)	4.3 (1.9)	4.0 (1.6)	5.9 (0.9)	5.6 (0.7)
D.I	8.3 (4.3)	4.3 (0.7)	4.0 (0.6)	4.0 (0.6)	5.2 (0.6)	4.9 (0.6)
D.II	7.9 (3.4)	3.3 (0.8)	3.2 (0.7)	3.0 (0.6)	5.5 (0.6)	5.2 (0.5)
E.I	8.4 (4.4)	5.0 (0.7)	4.7 (0.6)	4.7 (0.6)	5.0 (0.6)	4.8 (0.6)
E.II	9.0 (4.9)	5.3 (0.8)	5.0 (0.6)	4.7 (0.6)	5.3 (0.8)	5.0 (0.6)
All	8.8 (5.5)	5.2 (1.7)	4.9 (1.5)	4.4 (1.2)	5.4 (0.9)	5.1 (0.8)

Standard deviations of error rates are in brackets. Nominal error rate is 5%.

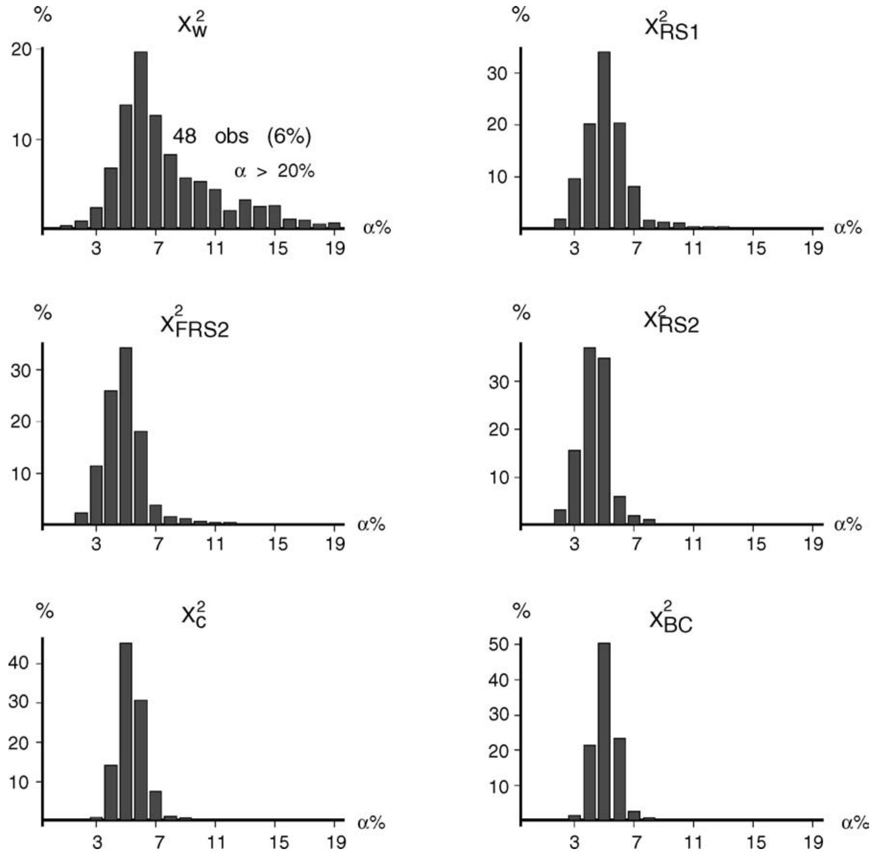

Fig. 1. Histograms of Type-I error rates (α) for Wald's GOF test statistic χ^2_{W} and five of the best alternatives (see text for details). Nominal significance level (α) = 5%. Relative frequencies are in % of 810 estimates of α each based on 2000 replicates of a specific combination of sample design, site, and cover-type map.

Table 3
Maximum likelihood estimates of the regression coefficients in the logistic model predicting the Type-I error rate of Wald's test-statistic $\hat{\alpha}_s(\chi^2_{\mathrm{W}})$ and of five alternatives ($\hat{\alpha}_s(\chi^2_{\mathrm{RS1}}), \ldots, \hat{\alpha}_s(\chi^2_{\mathrm{BC}})$)

Predictor	$\hat{\alpha}_s\left(X^2_W\right)$	$\hat{\alpha}_s\left(X^2_{RS1}\right)$	$\hat{\alpha}_s\left(X^2_{FRS1}\right)$	$\hat{\alpha}_s\left(X^2_{RS2}\right)$	$\hat{\alpha}_s\left(X^2_C\right)$	$\hat{\alpha}_s\left(X^2_{BC}\right)$
μ	0.228	-0.720	-0.989	-1.058	-2.925	-3.112
δ_B	-0.131	-0.488	-0.477	-0.475	-0.090	-0.069
δ_C	-0.210	-0.776	-0.781	-0.676	0.097	0.115
δ_D	0.046	-0.697	-0.727	-0.599	-0.008	-0.003
δ_E	0.203	-0.333	-0.360	-0.252	-0.036	-0.037
K	0.395	0.241	0.249	0.186	0.011	0.001
m	-0.000	0.003	0.003	0.001	-0.000	0.000
n_s	-0.041	-0.001	-0.001	-0.001	-0.001	-0.000
Simpson	-5.779	-5.014	-4.869	-4.641	0.472	0.568
ρ	-4.066	-2.472	-2.291	-2.007	0.140	0.145
RMSE	0.009	0.004	0.004	0.004	0.003	0.003
$R^2_{adj.}$	0.61	0.64	0.63	0.54	0.28	0.18

δ_{site} is an indicator variable taking the value of 1 if observations are from *site* and 0 otherwise (site = A, ..., E). Site A is taken as the reference for site estimates, i.e. $\delta_{\mathrm{A}} \equiv 0$. RMSE = root mean square error of model. $R^2_{\text{adj.}}$ = adjusted squared correlation coefficient. Shaded table entries are not statistically significant ($P \geq 0.05$).

classes ($P \leq 0.05$). Plot size (m) had only a minor effect on error rates. Wald's test was the only test statistic with a significant effect of sample size. Its error-rate would decrease about 0.4% for every cluster (plot) added to the sample. Evenness of cover-type proportions (Simpson's index) had a very strong and significant impact on error rates of $\hat{\chi}^2_W$, $\hat{\chi}^2_{\mathrm{RS1}}$, $\hat{\chi}^2_{\mathrm{FRS1}}$, and $\hat{\chi}^2_{\mathrm{RS2}}$. Least and non-significantly impacted were $\hat{\alpha}_s(\hat{\chi}^2_C)$ and $\hat{\alpha}_s(\hat{\chi}^2_{\mathrm{BC}})$. A similar pattern were seen in the effect of the average intra-cluster correlation coefficient. Finally, one should notice that the only two logistic models with a constant term close to the nominal target of 2.996 for $\alpha = 0.05$ are those for $\hat{\alpha}_s(\hat{\chi}^2_C)$ and $\hat{\alpha}_s(\hat{\chi}^2_{\mathrm{BC}})$.

Type-II error rates for the alternative test of equality of a sample-based estimate of map I cover-type proportions and map II proportions (and vice versa) for Wald's test were, as expected, somewhat lower than for any of the five best alternatives. However, for the 65 site and map specific designs with a test power of 80% or better (i.e. Type-II errors are less than 20%) the differences were in the 0–5% range with an average of 2.7%. Typical results exemplified by sites D and E are in Fig. 2. For designs with a good test power (>90%) the Type-II errors of the five best alternatives were within 1% (average 0.3%) of Type-II errors incurred with Wald's test.

A simultaneous consideration of Type-I and -II errors shows that the bias-corrected version of Finney's (1971) and Brier's (1980) corrected Chi-squared test statistic provides the overall best odds of correctly accepting or rejecting a simple test hypothesis. For χ^2_{BC}, the log-odds ratio ψ was consistently higher (range 0.7–2.2, mean 1.1) than the log-odds ratio for χ^2_{W} and also consistently the highest among the five best alternatives. Typical results exemplified by sites D and E are in Fig. 2.

4. An example of application

The practical consequence of switching from Wald's test-statistic $\hat{\chi}^2_{\mathrm{W}}$ to $\hat{\chi}^2_{\mathrm{BC}}$, apart from an expected general lowering of the Type-I error rates, depends, of course, on the decision riding on the statistical inference. If it is a simple matter of either non-rejection or acceptance of a null hypothesis the switch will lower the rate of rejections when the null hypothesis is true or we have low statistical power to reject the null when it is false. Often, however, the two test statistics will lead to the same conclusion. For example, when testing ($\alpha = 0.05$) the null hypothesis of equality of a sample-based estimate ($n = 40$, $m = 16$)) of relative cover-type frequencies ($K = 4$) for map I and a census result for map II (i.e. $\hat{\mathbf{p}}_{\mathrm{I}s} = \mathbf{p}_{\mathrm{II}}$) the null hypothesis was rejected jointly by the two test statistics in 1806–1992 cases out of 2000 on sites A, B, D, and E. On site C, the squared distance between $\hat{\mathbf{p}}_{\mathrm{I}s}$ and $\mathbf{p}_{\mathrm{II}}$ is considerably smaller than on any other site and the agreement is accordingly weaker (60%). Our example will be from site C with the above design settings. To appreciate the example we should mention that the design, despite an appreciable sampling effort of 640 units, only has a

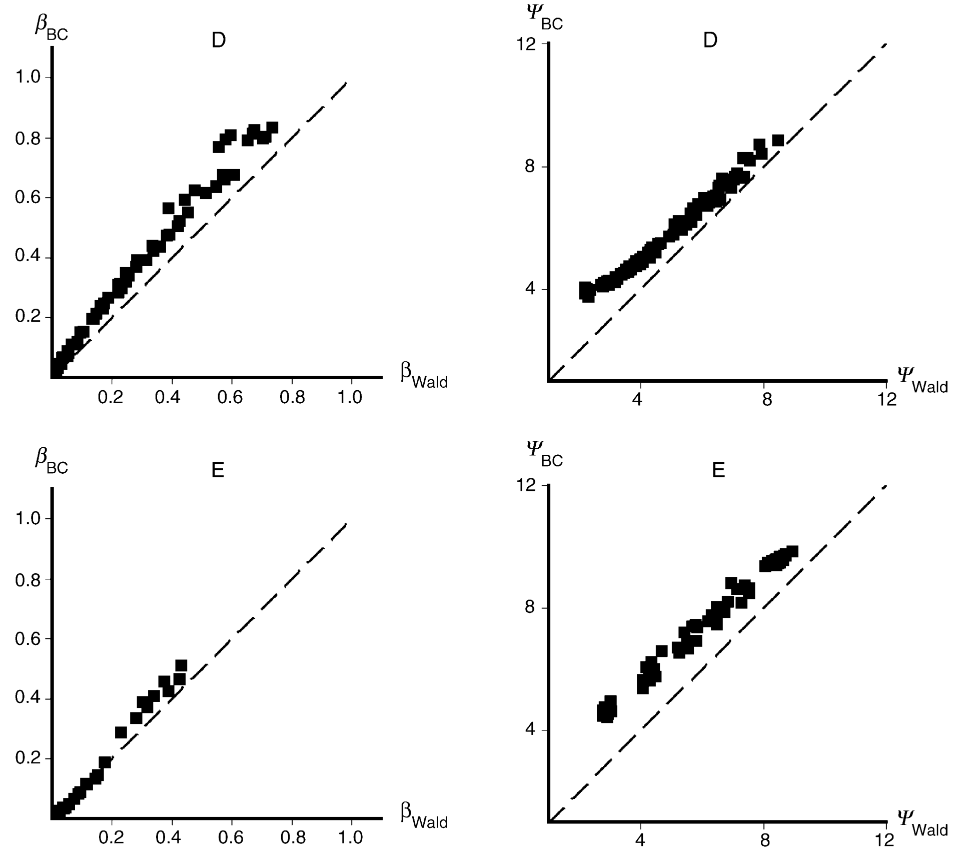

Fig. 2. Type-II error rates ($\hat{\beta}_s$) and log-odds of a correct test decision (ψ) of $\hat{\chi}^2_{\mathrm{BC}}$ in sites D and E plotted against those for $\hat{\chi}^2_{\mathrm{W}}$. Estimates are from 81 sample designs and 2000 replications per design.

statistical power of about 22% to declare an average difference of 5% in relative cover-type frequencies significant at the 5% level of significance. A claim of statistical significance should, therefore, a priori, be carefully scrutinized.

Sample data for the example are in the Appendix A. The sample mean of cover-type proportions for map I is $\hat{\mathbf{p}}_{\mathrm{I}s} = \{0.640, 0.005, 0.333, 0.022\}$ while the census estimate for map II is $\mathbf{p}_{\mathrm{II}} = \{0.592, 0.041, 0.294, 0.073\}$. Note that two sample-based estimates are close to zero. As mentioned earlier, this has negative implications for the performance of Wald's test. The sample variance-covariance matrix of relative cover-type proportions was (rounded to nearest 0.001) was:

$$\hat{\boldsymbol{\Sigma}}_s = \begin{pmatrix} 0.165 & 0.001 & -0.150 & -0.004 \\ 0.001 & 0.001 & -0.002 & -0.000 \\ -0.150 & -0.002 & 0.152 & -0.005 \\ -0.004 & -0.000 & -0.005 & 0.009 \end{pmatrix} \tag{11}$$

Wald's test-statistic after dropping results for the second cover-type (Eq. (2)) is hereafter:

$$\hat{\chi}^2_{\mathrm{W}} = 40 \times (0.048, 0.04, -0.051)' \begin{pmatrix} 61.363 & 47.472 & 61.220 \\ 47.742 & 1153.67 & 59.361 \\ 61.220 & 59.361 & 67.796 \end{pmatrix} (0.048, 0.04, -0.051) = 18.49 \tag{12}$$

The probability of obtaining a Chi-squared statistics with $4 - 1$ degrees of freedom larger than 18.49 under the null hypothesis is 0.0004. Had we chosen Wald's test-statistic we would conclude that there were significant differences between the sample (from map I) and the census from map II. Pearson's Chi-squared statistic (Eq. (4)) for the same data is $\hat{\chi}^2_s = 49.00$ and Brier's correction factor (Eq. (8)) came to $\hat{C} = 6.91$. Our bias correction of $\hat{C}$ (Eq. (9)) amounted to $\hat{\omega} = 0.32$. Accordingly, $\hat{\chi}^2_{\mathrm{BC}} = 6.78$ with $P(\chi^2_3 \geq \hat{\chi}^2_{\mathrm{BC}}) = 0.08$. Hence, if we adopt Brier's bias-corrected test statistic (Eq. (9)) and test at the 5% level of significance, we would have some support of the null hypothesis and would unlikely outright reject it. In light of the low statistical power of the chosen sample design the latter result is more reasonable than a rejection of the null hypothesis. The near null results for two cover-types should, everything else equal, signal problems with Wald's test.

5. Discussion and conclusion

Our assessment of Type-I error rates of Wald's GOF test statistic under the simple null hypothesis and one-stage cluster sampling of categorical data confirmed its sensitivity to factors like the number of classes, evenness of class proportions, intra-cluster correlation, and 'site'. This dependency is a nuisance to the applied statistician since an observed significance level is likely skewed by one or more of these 'nuisance' effects. The logistic regression models derived from this study may serve to gauge their impact. Of the many available alternatives to Wald's test-statistic five appears to offer consistent and significant improvements in Type-I error rates and the odds of a correct decision to either reject or accept a simple GOF hypothesis. These benefits come at the expense of a small loss of test power, least for designs with a high test power (>90%), more for designs with low or medium test power (<80%).

Six test statistics evaluated by Thomas and Rao (1987) and Thomas et al. (1997) as reasonable alternatives to Wald's statistic performed too erratically in this study to be of much practical value. The simulations by Thomas and Rao (1987) and later Thomas et al. (1997) were simplified to equal probable classes and a constant (among classes) intra-cluster correlation of 0.12. Class proportions are rarely equal in ecological data and the intra-cluster correlation of categorical classes is often stronger, at least when the area of a 'patch' with a single categorical class value is a multiple of the plot size (Cerioli, 1997; Ferguson and Bester, 2002; Garcia-Gigorro and Saura, 2005; Lichstein et al., 2002; Magnussen, 2001; Reed and Burkhart, 1985). The demonstrated sensitivity of most test statistics to 'evenness' of class proportions, intra-cluster correlation, and 'site' confirms the importance of assessing GOF test statistics with realistic data.

Type-I error rates of the five best alternative GOF test statistics were also dependent on one or more 'nuisance' effects. Our proposed bias-correction of Finney's (1971) and Brier's (1980) method of moments correction of Pearson's Chi-squared test statistic was, however, by far the least sensitive. In consequence it showed the most consistent performance in this study. In light of these results it may seem surprising that Finney's and Brier's method of moments correction has not before been formally evaluated. Both Finney and Brier mention that their correction performed well in some non-specified Monte Carlo simulations but no results were given. We surmise that the necessary simplifications of artificially generated data unwittingly 'masked' the robust property of this statistic, which, as far as we can tell, is the primary attribute responsible for its improved performance vis-à-vis alternatives based on Taylor-Series approximation or Jackknifing. The truncation of the correction factor to the allowed range of values achieves the 'robustness'. Our proposed bias-correction simply improves the overall performance in a non-negligible and consistent way.

The form of Finney's and Brier's test statistic lend itself seamlessly to a bootstrap procedure. In an earlier study (unpublished), we demonstrated a simple implementation of such a procedure. A bootstrapped version of Brier's and Finney's test statistic would eliminate the need for our proposed bias-correction.

We restricted our study to sampling with plots with a complete set of m units. In practice units will be missing from some plots due to, for example, straddling of population boundaries. Brier's and Finney's method of moments correction adapts easily to data with missing observations. A correction factor is computed separately for plots with 0, 1, …, $m - 1$ missing observations and then combined into a weighted average. In a separate study (unpublished), we found this scheme to work well for data with 5%, 10%, and 15% of the units in a plot missing at random. No simple recourse to address the impact of missing data is readily available for Wald's test-statistic or any other of the studied alternatives.

Continued use of Wald's test for a simple GOF null hypothesis under one-stage cluster sampling with significant design effects should be discouraged. A better alternative is proposed here. An alternative with only slightly less test power for designs with good test power (>80%) but with a consistent improvement in the odds of making a correct decision to either reject or accept a simple hypothesis. We reiterate that this conclusion is limited to simple GOF hypotheses. Studies by, among others, Thomas et al. (1997) indicate that the need for an alternative to Wald's GOF test involving a non-linear function of class proportions, as in the test of independence of rows and columns in a contingency table, is less clear.

Acknowledgements

We are grateful to Drs. P. Corona, D. Goodenough, M.N. Suratman, and M.D. Wulder, and the New Brunswick Department of Natural Resources and Energy, Forest Management Branch for making the forest cover-type maps available to us. Two anonymous reviewers made helpful suggestions for improvements.

Appendix A

Sample data (y_{ij}) for detailed example. Site C: sample size $n = 40$, plot-size $m = 16$ units, number of cover-types $K = 4$. Sampling from photo-interpreted forest cover-type map. y_{ij} = number of cover-type j units in sample plot i. Note samples 21 and 22 have fewer than 16 units. They are both located near the edge of the cover-type map I and have, respectively, eight and four units outside the sampling frame.

Sample	Cover type			
	1	2	3	4
1	16	0	0	0
2	0	0	16	0
3	1	0	15	0
4	13	0	0	3
5	14	0	2	0
6	2	0	14	0
7	16	0	0	0
8	13	0	3	0
9	2	0	12	2
10	6	0	10	0
11	4	0	12	0
12	10	0	6	0
13	16	0	0	0
14	0	0	16	0
15	16	0	0	0
16	6	0	10	0
17	16	0	0	0
18	16	0	0	0
19	16	0	0	0
20	16	0	0	0
21	0	0	8	0
22	0	0	12	0
23	4	0	12	0
24	14	0	2	0
25	16	0	0	0
26	13	3	0	0
27	16	0	0	0
28	16	0	0	0
29	16	0	0	0
30	4	0	12	0
31	16	0	0	0
32	16	0	0	0
33	7	0	0	9
34	16	0	0	0
35	7	0	9	0
36	6	0	10	0
37	0	0	16	0
38	16	0	0	0
39	4	0	12	0
40	16	0	0	0
Mean	10.05	0.08	5.23	0.35

References

Agresti, A., 1992. A survey of exact inference for contingency tables. Stat. Sci. 7, 131–177.

Agresti, A., Caffo, B., 2000. Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. Am. Stat. 54, 280–288.

Bedrick, E.J., 1983. Adjusted Chi-squared tests for cross-classified tables of survey data. Biometrika 70, 591–595.

Brier, S.S., 1980. Analysis of contingency tables under cluster sampling. Biometrika 67, 591–596.

Cerioli, A., 1997. Modified tests of independence in 2 × 2 tables with spatial data. Biometrics 53, 619–628.

Cerioli, A., 2002a. Tests of homogeneity for spatial populations. Stat. Prob. Let. 58, 123–130.

Cerioli, A., 2002b. Testing mutual independence between two discrete-valued spatial process: a correction to Pearson Chi-squared. Biometrics 58, 897.

Clifford, P., Richardson, S., Hemon, D., 1989. Assessing the significance of the correlation between two spatial processes. Biometrics 45, 123–134.

Cochran, W.G., 1977. Sampling Techniques. Wiley, New York, p. 380.

Cohen, A.C., 1976. The distribution of the Chi-squared statistics under clustered sampling from contingency tables. J. Am. Stat. Assoc. 71, 665–669.

Corona, P., Chirici, G., Marchetti, M., 2002. Forest ecosystem inventory and monitoring as a framework for terrestrial natural renewable resource survey programmes. Plant Biosys. 136, 69–82.

Dale, M.R.T., Fortin, M.J., 2002. Spatial autocorrelation and statistical tests in ecology. Ecoscience 9, 162–167.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall, Boca Raton, p. 436.

Fay, R.E., 1979. On adjusting the Pearson Chi-squared statistic for cluster sampling. In: Proceedings of the Social Statistics Section, American Statistical Association, pp. 665–670.

Fellegi, I.P., 1980. Approximate tests of independence and goodness-of-fit based on stratified multistage samples. J. Am. Stat. Assoc. 75, 216–268.

Ferguson, J.W.H., Bester, M.N., 2002. The treatment of spatial autocorrelation in biological surveys: the case of line transect surveys. Antarctic Sci. 14, 115–122.

Finney, D.J., 1971. Probit Analysis, vol. 3. Cambridge University Press, p. 350.

Fleiss, J.L., 1981. Statistical Methods for Rates and Proportions. Wiley, New York, p. 313.

Garcia-Gigorro, S., Saura, S., 2005. Forest fragmentation estimated from remotely sensed data: Is comparison across scales possible? For. Sci. 51, 51–63.

Gilliland, D., Schabenberger, O., Liu, H., 2002. Intercluster correlations for binomial data: an application to seed testing. J. Agric. Biol. Ecol. Stat. 7, 95–106.

Goodenough, D.G., Bhogal, A.S., Dyk, A., 2000. Determination of above-ground carbon in Canada's forests. In: IGARSS 2000, IGARSS, Honolulu, HI.

Gregoire, T.G., 2004. Editorial: special issue on statistical methods and techniques for analyzing spatial and temporal-spatial data. Env. Ecol. Stat. 11, 353–354.

Hall, D.B., Severini, T.A., 1998. Extended generalized estimating equations for clustered data. J. Am. Stat. Assoc. 93, 1365–1375.

Holt, D., Scott, A.J., Ewings, P.D., 1980. Chi-squared tests with survey data. J. R. Stat. Soc. A 143, 303–320.

Hosmer Jr., D.W., Lemeshow, S., 1980. Applied Logistic Regression. Wiley, New York.

Koch, G.G., Freeman Jr., D.H., Freeman, J.L., 1975. Strategies in the multivariate analysis of data from complex surveys. Int. Stat. Rev. 43, 59–78.

Legendre, P., 1993. Spatial autocorrelation: trouble or new paradigm? Ecology 74, 1659–1673.

Lehmann, E.L., 1983. Theory of Point Estimation. Wiley, New York, p. 506.

Lichstein, J.W., Simons, T.R., Shriner, S.A., Franzreb, K.E., 2002. Spatial autocorrelation and autoregressive models in ecology. Ecol. Monogr. 72, 445–463.

Lloyd, C.J., 1999. Analysis of Categorical Variables. John Wiley, New York, p. 468.

Magnussen, S., 2001. Fast pre-survey computation of the mean spatial autocorrelation in large plots composed of a regular array of secondary sampling units. Math. Model. Sci. Comput. 13, 204–217.

Magnussen, S., 2004. Prediction of 2 × 2 tables of change from repeat cluster sampling of marginal counts. Can. J. For. Res. 34, 1703–1713.

Magnussen, S., Stehman, S.V., Corona, P., Wulder, M.A., 2004. A Pòlya-urn resampling scheme for estimating precision and confidence intervals under one-stage cluster sampling: Application to map classification accuracy and cover-type frequencies. For. Sci. 50, 810–822.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, second ed. Chapman & Hall, London, p. 511.

Miao, W., Gastwirt, J.L., 2004. The effect of dependence on confidence intervals for a population proportion. Am. Stat. 58, 124–130.

Patil, G.P., 1982. Diversity as a concept and its measurement. J. Am. Stat. Assoc. 77, 548–561.

Pawitan, Y., 2000. A reminder of the fallibility of the Wald statistic: likelihood explanation. Am. Stat. 54, 54–56.

Rao, J.N.K., Scott, A.J., 1981. The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. J. Am. Stat. Assoc. 75, 221–230.

Reed, D.D., Burkhart, H.E., 1985. Spatial autocorrelation of individual tree characteristics in loblolly pine stands. For. Sci. 31, 575–587.

Ridout, M.S., Demetrio, C.G.B., Firth, D., 1999. Estimating intraclass correlation for binary data. Biometrics 55, 137–148.

Serfling, R.J., 1980. Approximation Theorems of Mathematical Statistics. J. Wiley Sons, New York, p. 364.

Shiver, B.D., Borders, B.E., 1996. Sampling techniques for forest resource inventory. Wiley, New York, p. 368.

Stoner, J.A., Leroux, B.G., 2002. Analysis of clustered data: a combined estimating equations approach. Biometrika 89, 567–578.

Suratman, M.N., Bull, G.Q., Leckie, D.G., LeMay, V.M., Marshall, P.L., Mispan, M.R., 2004. Prediction models for estimating the area, volume, and age of rubber (*Hevea brasiliensis*) plantations in Malaysia using Landsat TM data. Int. For. Rev. 6, 1–12.

Thomas, D.R., Rao, J.N.K., 1987. Small-sample comparisons of level and power for simple goodness-of-fit statistics under cluster sampling. J. Am. Stat. Assoc. 82, 630–636.

Thomas, D.R., Singh, A.C., Roberts, G.R., 1997. Tests of independence on two-way tables under cluster sampling: an evaluation. Int. Stat. Rev. 64, 295–311.

Wald, A., 1941. Asymptotically most powerful test of statistical hypotheses. Ann. Math. Stat. 12, 1–19.

Wulder, M.A., Cranny, M., Dechka, J., 2002. An Illustrated Methodology for Land Cover Mapping of Forests with Landsat-7 ETM+ data: Methods in Support of EOSD Land Cover. Version 2, Canadian Forest Service, Victoria, B.C., unpublished report, pp. 1–39.