ANALYSIS METHODS FOR SITE-SPECIFIC TARNISHED PLANT BUG SAMPLING DATA

Jeffrey L. Willers
USDA-ARS
Genetics and Precision Agriculture Research Unit
Mississippi State, MS
Yongfeng Zhao
Department of Statistics
Mississippi State, MS
Jixiang Wu
Mississippi State University
USDA-ARS
Mississippi State, MS
Kenneth Hood
Hood Farms & Gin
Gunnison, MS
John R. Bassie
Bassie Ag Service
Cleveland, MS

Abstract

Geo-referenced field samples for the tarnished plant bug (TPB) (*Lygus lineolaris* Miridae:Heteroptera) from the third week of July during the 2004 cotton production season were analyzed by several statistical methods. Geo-referenced imagery of a complex of cotton fields was processed into a map of three cotton habitat categories (HABITAT) and used by two different samplers (OBSERVER) to select sample locations. Analyses of these data by traditional methods such as confidence intervals or ordinary least squares regression were compared to several regression methods specifically designed to analyze counts. It is demonstrated that the count models better describe the TPB sample counts as functions of the categorical explanatory variables (OBSERVER and HABITAT) than the two traditional approaches. It is concluded that count model regressions, which are based upon a generalized linear model approach, provide a valuable suite of tools for the research or industrial entomologist.

Introduction

Pielou (1977) apportioned the ecological study of organisms into two study areas: (1) population dynamics (the study of changes in population density … over time) and (2) the spatial pattern of populations. This work demonstrates applications of count models to analyze cotton insect sample counts for the purpose of (1) estimating insect population density and (2) determining if pest density geographically differs among distinct cotton habitats at a discrete point in time. The cotton pest insect of interest in this study is the tarnished plant bug (*Lygus lineolaris* [Palisot de Beauvois]).

The first kind of analysis approach was based upon confidence interval statistics. The second kind of analysis applied to the same data set was ordinary least squares (OLS) regression. The third kind of analysis was to examine the sample counts by the application of count model regression techniques. Results from the three different analysis techniques demonstrated that choice of analysis method affects conclusions.

The first assumption of this study is that the target pest insect population within a specific cotton habitat category is randomly dispersed (Willers et al. 2005; Willers et al. 2006). These habitat categories are extracted from the processing of a multispectral image of the cotton field and represent different geographic regions where the cotton plant-insect community interactions may differ. A second assumption is that the sample unit size is comprised of a small area of ground that contains several adjoining cotton plants (Willers et al. 2005), rather than sample unit sizes of single plants, numbers of sweeps, or drop cloth samples. Furthermore, since the location of every sample can be geo-referenced, the emphasis of an analysis must shift toward the comparison of the number of insects found per unit area within and among different geographic regions of the cotton field. Traditionally, the focus has been upon comparing a summarization of a field average to a threshold without consideration of the geographic location of the

samples. Therefore, with the ability to geographically place sample sites within various cotton habitat categories, the purpose of sampling is to determine if there is evidence of differences in cotton insect abundance among different regions of the field and not whether the field average is different from a threshold value.

Material and Methods

Image Processing and Field Data

Multispectral imagery (Jensen, 2000) used in this study was acquired by InTime, Inc. (Cleveland, MS, http://www.gointime.com) on 10 July 2004 and provided by the courtesy of Perthshire Farms (Gunnison, MS). The imagery was subset and processed for a group of fields called the Fisher Place. The spatial resolution of the imagery was 2 m and registered to earth coordinates in the UTM projection, for Zone 15, using the WGS84 datum (Bugavesky and Snyder, 1995). Image classification was similar to the process described by Willers et al. (2005), except that all three bands (red, green and near infrared) were used during the unsupervised classification step to derive 26 classes, where class 0 is the background. A color map (hardcopy) based upon these classifications was provided to two observers. Each chose their sample locations using their judgment and without knowledge of how the other chose their site.

Each tarnished plant bug (TPB) sample was based upon large sized sample units as described by Willers et al. (2005). The most common sampling method (to determine TPB abundance from a small area at a geographic location) was the sweep net, where each sample was comprised of 33 sweeps. A few samples were collected with a drop cloth, where counts were made from 4 contiguous rows. There was no evidence of differences in numbers of adults or nymphs observed based upon method of sampling in these data, nor of a difference in numbers between the sample dates of 19 and 22 July. Therefore, the numbers of adults and/or nymphs observed per sample were summed for each sample into a count variable named 'TOTAL'. Both observers logged the geographic coordinates of each of their sample sites using two different, low cost Global Positioning Systems (GPS) GARMIN® receivers.

Tables 1 and 2 contain the sampling information for two observers (the farm consultant is called 'Observer 1' and the researcher is 'Observer 2') during the third week of July 2004. These tables show the TPB numbers observed per sample, the coordinate location and time of each sample, and the cotton habitat class assigned to each, derived from the focal maximum (Theobald, 2003) of the unsupervised classification value of the pixels found within a buffer surrounding each sample location (Figure 1). For data analysis, the 25 unsupervised classes (1-25) were regrouped into cotton habitat categories, using supervised classification methods (Willers et al. 2005), that were labeled as Marginal, Good and Best (Appendix I).

The TPB samples were regrouped as follows. The first set, named 'Non-stratified' represents the estimate of the overall field average of TPB. Using further processing, the sample counts were summarized into an output table of other sets that tallied how many samples had the particular count values according to habitat category (Marginal, Good, and Best) and by the observer (Table 3).

For the confidence interval (CI) analyses, the various sets of samples were pooled across the two observers. For the regression analyses by OLS and the various count models, the observers and the habitats were both used as effects.

Confidence Interval Analysis

The samples were first analyzed by constructing 100(1-α) CIs. To test if the mean of a particular set of samples was equivalent to zero (Zar, 1981), the CI used was:

$$\bar{X} \pm t_{\alpha/2,n-1} * \left(s / \sqrt{n} \right). \quad (1)$$

To test if the difference of two means (Schenker and Gentleman, 2001) was equivalent to zero, the constructed 100(1-α) CI was:

$$\left(\bar{X}_1 - \bar{X}_2 \right) \pm t_{\alpha/2,\upsilon} * \sqrt{s_1^2 / n_1 + s_2^2 / n_2}, \quad (2)$$

where the effective df, ν [See Steel and Torrie, 1980, p.106.], is computed by:

$$\upsilon = (s_1^2 / n_1 + s_2^2 / n_1) / [(s_1^2 / n_1)^2 / (n_1 - 1)] + [(s_2^2 / n_2)^2 / (n_2 - 1)]. \quad (3)$$

For all CI, the selected significance level was $\alpha = 0.05$. The MEANS procedure in SAS® Ver. 9.1 obtained the summary statistics necessary to build the confidence intervals for all the TPB samples without stratification and for the three different habitats and to make various comparisons of means.

Ordinary Least Squares (OLS) Regression

The TPB counts (Table 3) were next analyzed using OLS (Myers and Montgomery, 1997) with the GLM procedure in SAS® Ver. 9.1 (Appendix I-B).

Poisson Regression and Variants

Statistical tests to determine which count model best fits these TPB counts is an advanced topic not fully discussed here (Long, 1997; Stokes et al., 2000; Cunningham and Lindenmayer, 2005). In this paper, the inspection of graphical plots was the approach that determined which count model best fit these TPB counts. The OLS fit and residuals were compared to several maximum likelihood models specifically developed for the analysis of counts (Long, 1997; Myers and Montgomery, 1997). There are several count models described in the literature (Long, 1997; Cunningham and Lindenmayer, 2005; Horton et al. 2007). Four of these are (1) Poisson regression, (2) Negative Binomial regression, (3) Zero-inflated Poisson regression and (4) Zero-inflated Negative Binomial regression. The latter three are variants of the Poisson regression model (Long, 1997).

The COUNTREG procedure available in Service Pack 4 for SAS® Ver. 9.1 fit these four count models (Appendix I-C) to the TPB counts of Table 3. The SAS® GENMOD procedure (Appendix I-D) was employed to determine the significance of mean differences between observer TPB counts (pooled over habitat) or among habitats (pooled over observers). Details of these models and of the software to fit them can be found in the texts by Agresti (1996), Long (1997), Allison (1999), Stokes et al. (2000), Schabenberger and Pierce (2002) and in papers by Johnston (1993), Myers and Montgomery (1997), Seavy et al. (2005), Slymen et al. (2006) and Horton et al. (2007). Software and additional information for the SAS® system can be found at: http://support.sas.com/kb/26/161.html.

Results and Discussion

CI Test for Equivalency to Zero

Here, the sample information is pooled to form a set labeled as 'Non-stratified' (Table 4), which is used to estimate the field average for TPB. An assumption (Zar, 1981; Thompson, 1992) of this kind of analysis is that these samples have been collected from a common TPB population, whose parametric mean and variance are unknown, but are estimable from the samples. One question of interest from the sample information was to determine if the estimated mean significantly differs from zero. By constructing a $100(1 - \alpha)$ % CI using Eqn. 1, this interval (Table 4) does not include 0, which indicates that the field average mean is significantly different from 0. Similar CIs show (Table 4) that the three means from the stratified habitat TPB samples (Best, Good, and Marginal) are significantly different from 0.

Equality of Paired Mean Differences

To test if the mean (or field average) estimated by the non-stratified set of samples is different from the mean of any of the stratified samples, three CIs were constructed (Eqn. 2). This CI of mean differences is examined for whether it excludes or includes 0. The intervals for [Non-stratified – Best], [Non-stratified – Good], and [Non-stratified – Marginal] were [-1.15, 0.27], [-0.42, 0.55], and [0.18, 0.98], respectively. The first two intervals indicate that there is no difference between the Best and Good habitat means and the field average since the include 0. However, the hypothesis that the mean of the Marginal habitat is equivalent to the field average is rejected.

To test if the means of the samples stratified by habitat categories are equivalent, the CIs of mean differences (Eqn. 2) were examined. The 95 % CI for the mean difference between the [Best – Good] cotton habitat was [-0.25, 1.25], therefore, the null hypothesis for these means was not rejected. The 95 % CI for the mean difference between [Best – Marginal] was [0.32, 1.72] and between [Good – Marginal was [0.05, 0.98]; thus, the null hypothesis for these two pairs was rejected.

These findings, based upon a traditional analysis approach, show evidence of a spatial pattern in TPB abundance. Some questions to examine further at this time, is if there are better methods to determine evidence for spatial patterns, and if so, do they obtain different results.

Ordinary Least Squares Regression

Figure 2 compares the fit of OLS regression to the observed probability of counts of TPB between 0 and 5. Figure 3 presents the graphs of the residuals. Both figures show that the OLS provides the poorest fit to these data.

Count Model Analyses

Figures 2 and 3 indicate that neither of the zero-inflated models fit any better than the Poisson or Negative Binomial regression, but all four of them fit the counts better than OLS. The fit of the Poisson and the Negative Binomial model are similar; leaving the choice of which one of the two is better to the preferences of the analyst.

Using PROC GENMOD, we chose to use the Poisson model to test for differences among the observers and habitats (Table 5). Table 5 indicates that the mean counts of the two observers were significantly different, suggesting that conclusions of the CI analyses described above, where the observer effect was discounted, needed to be examined in a different way. One of the simplest ways to do this was to examine the least significant mean comparisons derived from the Poisson regression analysis. These comparisons (Table 5) indicated that the Marginal and Good habitat counts were non-significant, while the two remaining comparisons were highly significant. The CI analyses resulted in opposing conclusions. The difference can be explained, in part, because the CI analyses were based upon assumptions about normality and these TPB sample counts were not normally distributed. Therefore, a count model was more appropriate for the analyses of these samples than either CI or OLS. In a future paper, the count model analysis methodology is further developed. Based upon the examination of the residual plot presented here (Figure 3) that paper will emphasize the Poisson model. That paper will compare the Poisson model and complete enumeration analyses with a correlation analysis.

Poisson regression is an appropriate application of a generalized linear model to count data (Long, 1997; Piegorsch and Bailer, 2005). Generalized linear models are defined as a refinement of traditional linear models where the mean of a population depends on a linear predictor through a nonlinear link function and where the probability distribution of the response is a member of the exponential family of distributions (Nelder and Wedderburn, 1972; Johnston, 1993). Johnston (1993) remarks that problems not appropriate for analyses by traditional linear models arise when (1) it is not reasonable to assume that the data are normally distributed as often occurs when modeling count data, (2) when the mean of the data is naturally restricted to a range of values, and (3) it is not appropriate to assume that the variances of the data are constant for all observations. For insect sampling data (particularly whenever samples can be geographically stratified using remote sensing information) Johnston's first and third points definitely apply.

Conclusion

While the sampling literature for insect pests of agricultural crops is quite large, few references at this time jointly employ remote sensing and the analyses of insect counts by count model regression methods. The results of this study indicate that researchers and industrial investigators would benefit from applications of count model methods to analyze insect sample counts. These count models are improvements over confidence intervals or ordinary least squares regression methods whenever large numbers of zeros occur, the data exhibit a skewed distribution, or are not otherwise normally distributed. Choice of which count model to use is dependent upon whether or not the sample counts are over-dispersed or contain excessive numbers of zeros.

Acknowledgements

The authors thank Dr. Blake Layton, Mississippi State University and Dr. JT Vogt, USDA-ARS and Mr. Ronald Britton, USDA-ARS for their time to read through the manuscript and suggest improvements. Additional financial support provided by Advanced Spatial Technologies for Agriculture (ASTA-322-298), Mississippi State University.

Disclaimer

Mention of trade names of commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendations or endorsement by the US Department of Agriculture.

References

Agresti, A. 1996. *An introduction to categorical data analysis*. (John Wiley & Sons, New York, USA) 290p.

Allison, P. D. 1999. *Logistic Regression Using the SAS® System: Theory and Application*. (SAS Institute, Inc., Cary, NC, USA). 304 pp.

Bugayevskiy, L.M., and J. P. Snyder. 1995. *Map Projections. A Reference Manual*. (Taylor and Francis, Ltd., Philadelphia, PA, USA) 328p.

Cunningham, R. B., and D. B. Lindenmayer. 2005. Modeling count data of rare species: Some statistical issues. Ecology 86(5): 1135-1142.

Horton, N. J., E. Kim, and R. Saitz. 2007. A cautionary note regarding count models of alcohol consumption in randomized controlled trials. BMC Medical Research Methodology 7:9 [On-line]. Available at: http://www.biomedcentral.com/1471-2288/7/9 (Verified 19 November 2007).

Jensen, J.R. 2000. *Remote sensing of the environment: An earth resource perspective*. (Prentice-Hall, Inc., Upper Saddle River, NJ, USA) 544p.

Johnston, G. 1993. SAS® software to fit the General Linear Model. SUGI 13, SAS® Institute, Cary, NC. 8 pp. [On-line]. Available at: http://support.sas.com/rnd/app/papers/papers_da.html (Verified 24 August 2007).

Long, J. S. 1997. *Regression Models for Categorical and Limited Dependent Variables*. (Sage Publications, Inc. Thousand Oaks, CA. USA) 297p.

Myers, R. H. and D. C. Montgomery. 1997. A tutorial on generalized linear models. Journal of Quality Technology 29(3), 274-291.

Nelder, J. A. and R. W. M. Wedderburn. 1972. Generalized Linear Models. Journal of the Royal Statistical Society, Series A (General), 135(3), 370-384.

Piegorsch, W. W. and Bailer, A. J. 2005. *Analyzing Envrionmental Data*. (Wiley and Sons, Ltd. Chichester. West Sussex, UK) 496 p.

Pielou, E. C. 1977. *Mathematical Ecology*. (John Wiley & Sons, New York) 384p.

Schabenberger, O., and F. J. Pierce. 2002. *Contemporary Statistical Models for the Plant and Soil Sciences*. (CRC Press, Boca Raton, FL, USA) 738 p.

Seavy, N. E, S. Quader, J. D. Alexander, and C. J. Ralph. 2005. Generalized linear models and point count data: Statistical considerations for the design and analysis of monitoring studies. USDA Forest Service Gen. Tech. Rep. PSW-GTR-191.

Schenker, N., and J. F. Gentleman. 2001. On judging the significance of differences by examining the overlap between confidence intervals. The American Statistician 55(3): 182-186.

Slymen, D. J., G. X. Ayala, E. M. Arredondo, and J. P. Elder. 2006. A demonstration of modeling count data with an application to physical activity. Epidemiologic Perspectives & Innovations **3**, 3 [On-line]. Available at: http://www.epi-perspectives.com/content/3/1/3 (Verified 8 November 2007).

Steel, R G. D. and J. H. Torrie. 1980. Principles and Procedures of Statistics: A Biometrical Approach. McGraw-Hill Publishing, New York. 633 pp.

Stokes, M. E., Davis, C. S. and Kock, G. G. 2000. *Categorical Data Analysis Using the SAS® System, 2nd edn*. (SAS Institute, Inc, Cary, NC, USA) 626p.

Theobald, D. M. 2003. *GIS Concepts and ARCGIS Methods*, First edn. (Conservation Planning Technologies, Fort Collins, CO, USA) 334 p.

Thompson, S.K. 1992. *Sampling*. (Wiley-Interscience, New York, USA) 343p.

Willers, J. L., J. N. Jenkins, W. L. Ladner, P. D. Gerard, D. L. Boykin, K. B. Hood, P. L. McKibben, S. A. Samson, and M. M. Bethel. 2005. Site-specific approaches to cotton insect control. Sampling and remote sensing analysis techniques. Precision Agriculture 6: 431-452.

Willers, J. L., J. M. McKinion, and J. N. Jenkins. 2006. Remote Sensing, Sampling, and Simulation Applications in Analyses of Insect Dispersion and Abundance in Cotton, pp. 879-885. In: Aguirre-Bravo, C., Pellicane, P. J., Burns, D. P., and Draggan, S. (Eds.). RMRS-P-42CD: Monitoring Science and Technology Symposium: Unifying Knowledge for Sustainability in the Western Hemisphere, 2004 Sept. 20-24, Denver, CO. Paginated CD or available at http://www.fs.fed.us/rm/pubs/rmrs_p042.pdf (Verified 30 Jan. 2007).

Zar, J. H. 1981. Power of statistical testing: Hypotheses about means. Am. Lab. June. Pp. 102-107.

Table 1. Scouting site observations by the farm consultant (Observer 1) on 19 and 22 July 2004.

SITE ID	DATE/TIME	TPB TOTAL/ SAMPLE	EASTING	NORTHING	HABITAT CLASS
42	19-JUL-04 13:48	0	698075.286237	3769139.300920	Best
43	19-JUL-04 13:50	2	698074.025755	3769199.390910	Best
44	19-JUL-04 13:54	0	697912.046940	3768795.413040	Good
45	19-JUL-04 13:57	0	697783.831192	3768789.751070	Best
46	19-JUL-04 14:02	0	697678.619068	3768797.666310	Marginal
47	19-JUL-04 14:02	1	697634.055826	3768796.138890	Best
48	19-JUL-04 14:04	1	697569.134474	3768583.478460	Marginal
49	19-JUL-04 14:07	1	697674.467807	3768617.229650	Best
50	19-JUL-04 14:10	2	697663.921286	3768694.981350	Best
51	19-JUL-04 14:12	0	697748.973855	3768703.309410	Best
52	19-JUL-04 14:16	1	697898.608765	3768561.806240	Best
53	19-JUL-04 14:19	1	697790.961092	3768520.267080	Good
54	19-JUL-04 14:22	1	698073.424075	3768637.491810	Best
55	19-JUL-04 14:25	1	697975.962112	3768724.731810	Marginal
95†	22-JUL-04 13:32	0	692858.212619	3761424.137690	Best
96†	22-JUL-04 13:35	4	698198.591028	3768387.150520	Best
97†	22-JUL-04 13:37	0	698258.909753	3768369.964320	Best
98†	22-JUL-04 13:40	0	698332.661691	3768303.062250	Best
99†	22-JUL-04 13:40	1	698220.538636	3768285.828170	Best
100†	22-JUL-04 13:58	3	698197.296242	3768283.555150	Best
101†	22-JUL-04 13:59	2	698086.213878	3768641.926480	Best
102†	22-JUL-04 14:00	0	698087.138828	3768668.730850	Best
103†	22-JUL-04 14:11	1	698086.272917	3768662.760960	Good
104†	22-JUL-04 14:16	0	697983.175488	3769160.573610	Best
105†	22-JUL-04 14:19	0	697998.551400	3769254.353830	Good

† Drop cloth sample of 4 contiguous rows (see Willers et al. 2005).

Table 2. Summary table for the 22 July 2004 scouting sites selected by the researcher (Observer 2).

SITE ID	DATE/TIME	TPB TOTAL/ SAMPLE	EASTING	NORTHING	HABITAT CLASS
1	22-JUL-04 14:26	0	698108.51223	3769610.21125	Marginal
2†	22-JUL-04 14:29	0	698104.65405	3769581.56997	Good
3	22-JUL-04 14:32	0	698105.27466	3769575.63015	Marginal
4	22-JUL-04 14:35	0	698101.17947	3769558.28265	Good
5	22-JUL-04 14:44	2	698026.50006	3769457.31527	Good
6	22-JUL-04 14:48	0	698000.66398	3769437.13074	Good
7	22-JUL-04 14:52	1	697950.94274	3769421.80249	Good
8	22-JUL-04 14:56	2	697939.43987	3769426.91833	Good
9	22-JUL-04 15:01	1	697904.02800	3769438.08057	Good

10†	22-JUL-04 15:05	2	697901.65170	3769433.26961	Good
11	22-JUL-04 15:10	1	697865.34517	3769463.46008	Good
12	22-JUL-04 15:16	0	697765.35778	3769435.17501	Good
13	22-JUL-04 15:22	0	697560.68719	3769508.26911	Good
14	22-JUL-04 15:26	2	697558.63842	3769558.81909	Good
15	22-JUL-04 15:30	0	697552.48590	3769663.44887	Good
16	22-JUL-04 15:34	3	697524.10201	3769717.61536	Good
17	22-JUL-04 15:37	1	697530.54435	3769788.58037	Marginal
18	22-JUL-04 15:42	0	697579.31221	3769896.74059	Marginal
19	22-JUL-04 15:46	0	697629.75697	3769948.38941	Good
20	22-JUL-04 15:51	1	697777.03066	3769989.56853	Good
21	22-JUL-04 15:54	1	697810.12082	3769923.59678	Marginal
22	22-JUL-04 15:58	0	697819.02925	3769876.76107	Marginal
23	22-JUL-04 16:00	0	697799.99872	3769886.47159	Marginal
24	22-JUL-04 16:02	0	697791.22444	3769950.58055	Marginal
25	22-JUL-04 16:06	1	697790.73713	3769808.31376	Marginal
26	22-JUL-04 16:15	0	698064.24521	3769712.86080	Good
27	22-JUL-04 16:17	0	698063.97413	3769702.14111	Marginal
28	22-JUL-04 18:31	0	698179.17001	3768415.31328	Marginal
29	22-JUL-04 18:34	4	698236.94608	3768424.85859	Best
30	22-JUL-04 18:37	1	698292.75310	3768433.76819	Good
30	22-JUL-04 18:40	1	698343.49010	3768471.73670	Marginal
31	22-JUL-04 18:44	1	698327.17721	3768564.24847	Best
32	22-JUL-04 18:47	5	698231.79747	3768552.12791	Best
33	22-JUL-04 18:57	4	698125.06400	3768679.64564	Best
34	22-JUL-04 19:02	2	698076.04675	3768654.21363	Best
35	22-JUL-04 19:10	2	697871.79966	3769179.08007	Good
36	22-JUL-04 19:12	0	697812.82644	3769179.63086	Best
37	22-JUL-04 19:16	2	697697.28060	3769184.35328	Good

† Drop cloth sample of 4 contiguous rows (see Willers et al. 2005).

Table 3. A contingency table (Schabenberger and Pierce, 2002; p. 318) for Tarnished Plant Bug (TPB) samples from cotton fields at the Fisher Complex during 19 and 22 July 2004.

No. of TPB per Sample	Observer 1			Observer 2			
	Marginal	Good	Best	Marginal	Good	Best	Total
0	1	2	8	8	8	1	28
1	2	2	5	4	5	1	19
2	0	0	3	0	6	1	10
3	0	0	1	0	1	0	2
4	0	0	1	0	0	2	3
5	0	0	0	0	0	1	1
Total	3	4	18	12	20	6	63

Table 4. Confidence intervals for the original samples, arranged into four groups, without partitioning by OBSERVER.

Label	Mean	Confidence Interval (95%)
Non-stratified	0.98	0.68 – 1.28
Best	1.42	0.76 – 2.08
Good	0.92	0.52 – 1.31
Marginal	0.40	0.12 – 0.68

Table 5. Least square differences of means for the HABITAT categories based upon the application of a Poisson regression model.

Differences of Least Squares Means						
Effect	Mean Difference	Estimate	Standard Error	df	Chi-Square	Pr > ChiSq
OBSERVER	Consultant - Researcher	-0.7942	0.3391	1	6.20	0.0128
HABITAT	Marginal - Good	-0.8090	0.4768	1	2.88	0.0898
HABITAT	Marginal - Best	-1.6779	0.4827	1	12.08	0.0005
HABITAT	Good - Best	-0.8690	0.3239	1	7.20	0.0073

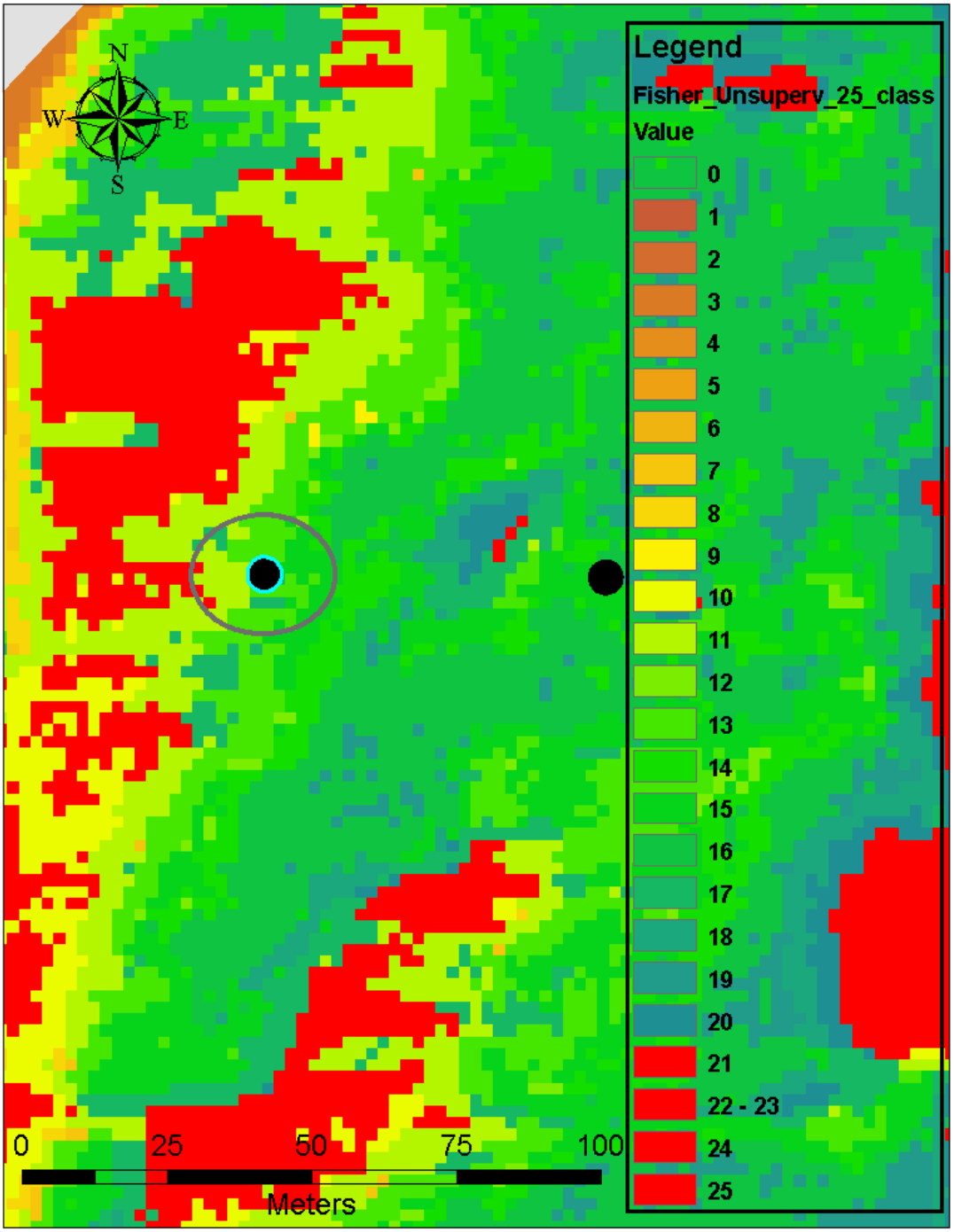

Figure 1. Illustration of the creation of a buffer around a sample site to extract the focal maximum of the pixel values within the buffer. Sample ID 36 is buffered, while sample ID 35 is shown just to the right. Both sites were collected by Observer 2.

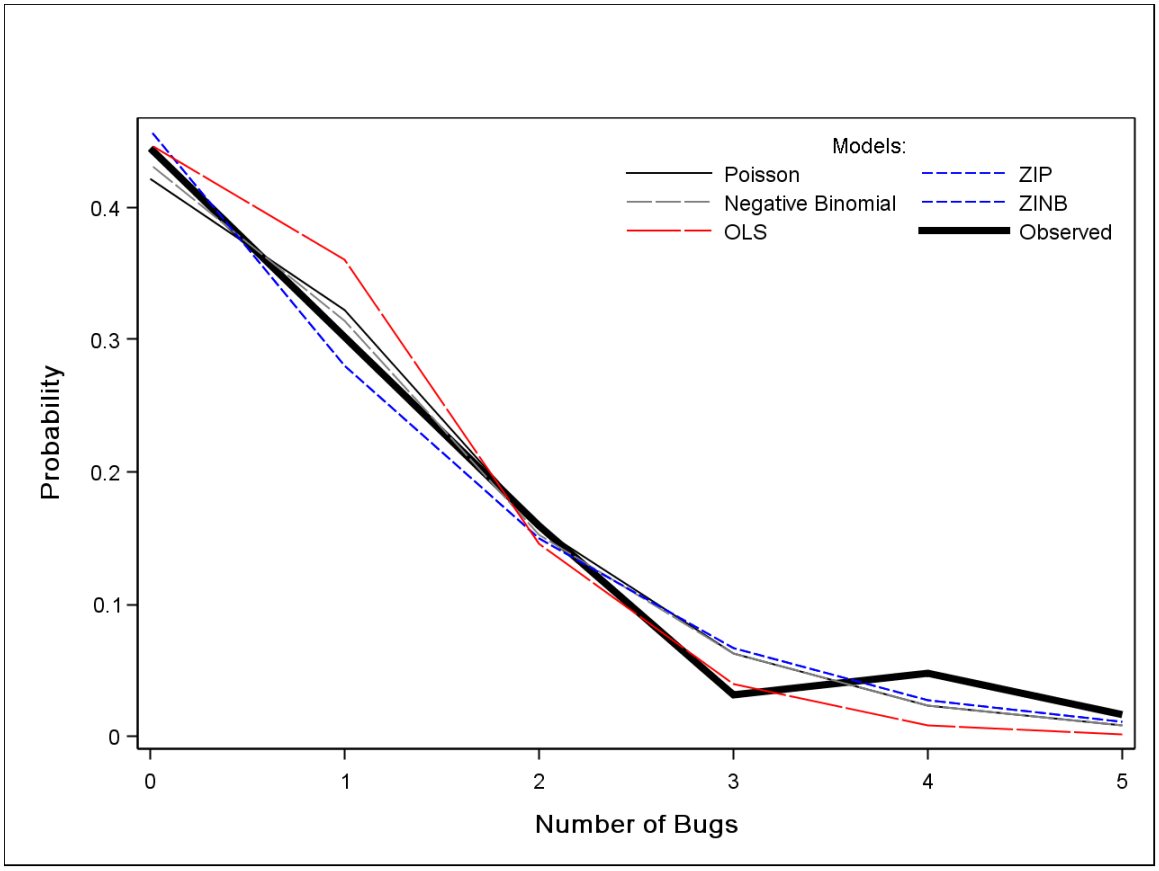

Figure 2. Predicted fits of several count models [Poisson, Negative Binomial, Zero-inflated Poisson (ZIP), and Zero-inflated Negative Binomial (ZINB)] and the ordinary least squares (OLS) regression in comparison to the observed marginal distribution (Allison, 1999; p. 219) values.

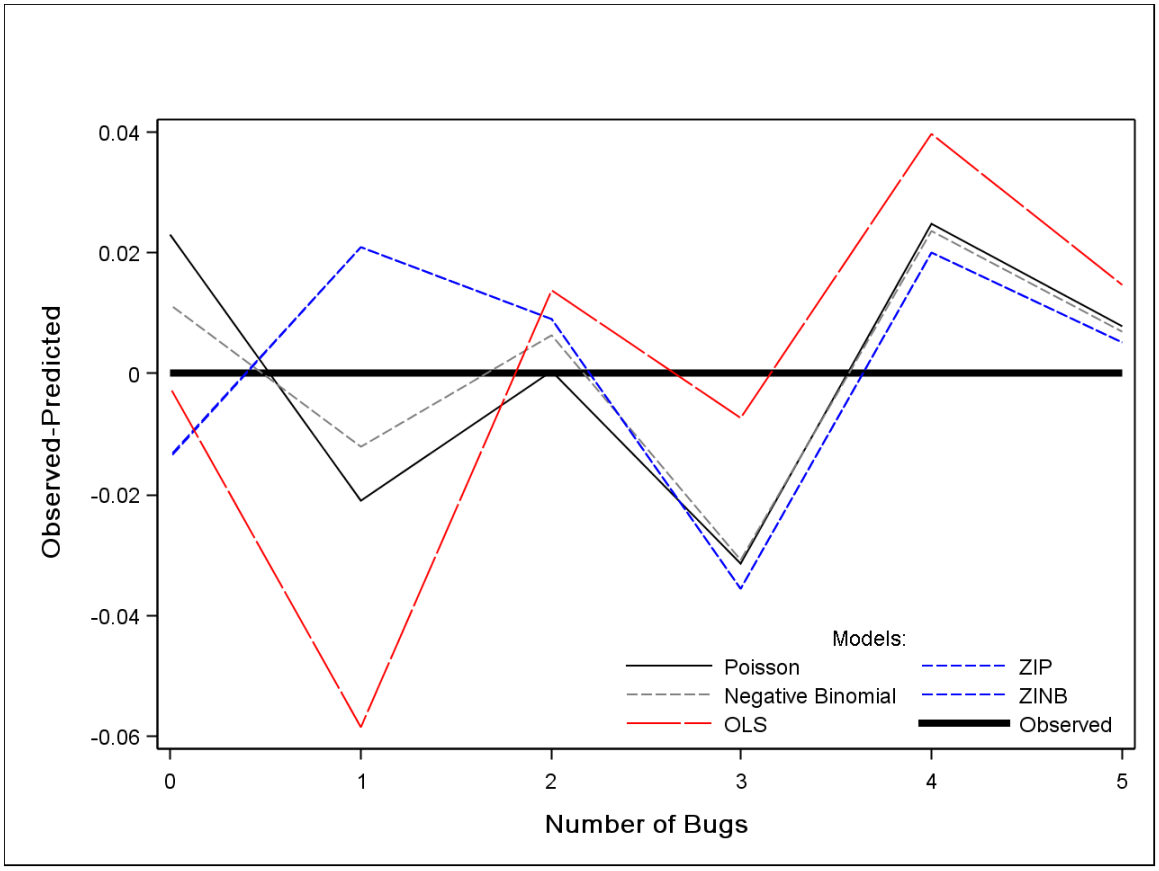

Figure 3. Residual plots for the Poisson, Negative Binomial, Zero-inflated Poisson (ZIP) and Zero-inflated Negative Binomial (ZINB) count models in comparison to ordinary least squares (OLS) regression.

Appendix I: Example SAS statements

A. For the consultant and researcher sample data, these SAS® statements recoded (Willers et al., 2005) the focal maximum (Theobald, 2003) of the unsupervised classes of the image pixels with the buffer of each sample site into three supervised classification categories (Marginal, Good and Best):

```
title 'Consultant sites classed by max_class';
data d;
set b;
if max_class LT 9 then h_class = 'Marginal';
if max_class GT 20 then h_class = 'Marginal';
if max_class GE 9 and max_class LT 17 then h_class = 'Good';
if max_class GE 17 and max_class LE 20 then h_class = 'Best';
keep ident h_class;
run;
```

B. The following SAS statements fit the OLS regression model to the frequencies of TPB counts (Table 3):

```
proc glm data=all;
model tpb = obsvr habitat / solution;
ods output ParameterEstimates=glmpe;
run;
```

The class statement is not employed in order to estimate the OBSERVER and HABITAT parameters similar in interpretation to those provided by the count models.

C. The following SAS statements fit four count models to the data of Table 3:

```
title 'Poisson Model';
proc countreg data=all type=poisson;
model tpb = obsvr habitat;
ods output ParameterEstimates=pe;
run;
title 'Negative Binomial Model';
proc countreg data=all type=negbin method=qn;
model tpb = obsvr habitat;
ods output ParameterEstimates=pe;
run;
title 'ZIP Model';
proc countreg data=all type=zip;
model tpb = obsvr habitat /
zi(var=obsvr habitat);
ods output ParameterEstimates=pe;
run;
title 'ZINB Model';
proc countreg data=all type=zinb method=qn;
model tpb = obsvr habitat /
zi(var=obsvr habitat);
ods output ParameterEstimates=pe;
run;
```

D. The GENMOD statements tested for differences in means between OBSERVERS and HABITATS:

```
proc genmod data=all;
class obsvr habitat;
model tpb = obsvr habitat / dist= Poisson pscale
                    link = log type1 type3;
lsmeans obsvr habitat / pdiff;
run;
```

See the SAS® documentation and cited references for more information about all of these code fragments .