Preliminary Results of Spatial Modeling of Selected Forest Health Variables in Georgia

Brock Stewart[1], Chris J. Cieszewski[2], and Eric L. Smith[3]

Abstract.—Variables relating to forest health monitoring, such as mortality, are difficult to predict and model. We present here the results of fitting various spatial regression models to these variables. We interpolate plot-level values compiled from the Forest Inventory and Analysis National Information Management System (FIA-NIMS) data that are related to forest health. These data included information concerning mortality, trees killed by various causes of death (indicated by the FIA-NIMS variable AGENTCD), and species richness.

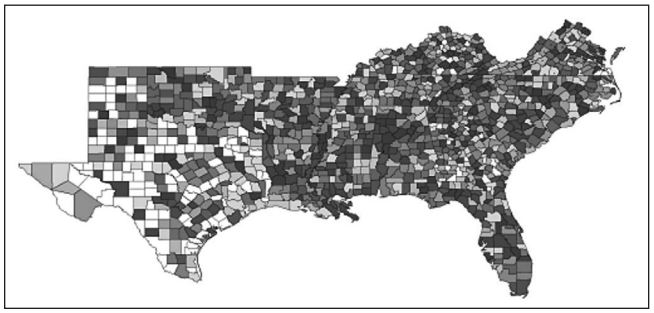
Introduction

In this study, we were interested in gathering information about forest health and mortality in the Southeast. The U.S. Department of Agriculture, Forest Service, Forest Inventory and Analysis (FIA) inventory data provide a good starting point for such analysis. The FIA data contain records of many variables, including information on stand structure with extensive coverage of large areas. We decided to use a means of displaying these estimates in a mapped context and/or testing for spatial trends that is more revealing and informative than a straightforward compilation of tabular estimates. The selection of appropriate resolution of mapping posed a challenge. While mapping State-level estimates is too course for our purposes, FIA data are intended to provide accurate information only at large scales, and county-level estimates can typically have large sampling errors. Kriging techniques have been used in abundance on FIA data.

The purpose of the current work was to investigate the fitting of various spatial models that would provide interpolation maps and significance testing. In this article, we focus predominantly on model fitting. That is, we attempt to find interpolation models with predictive value from phase 2 FIA data. Further research might include, e.g., incorporating Forest Health Monitoring (FHM) aerial surveys, which do not provide detailed information on stand structure but do provide large-scale trends of mortality and infection. Remote-sensing data such as this can provide information on large-area coverage of when and where trees are dying but cannot provide details of stand structure. On the other hand, plot-level data provide information on forest type, species, stand size, and density, etc., but are not suited well for detecting rare or sparse events. In the future, we would like to incorporate information at both scales.

In figure 1, we give an example of a FIA plot-level value mapped by county. Here, the value is the number of trees killed by insects. Our main interest was in detecting trends in mortality, and, if possible, mortality by specific causes such as insects or disease. We also examined species richness. We wish to test if plot-level quantities like these vary across the Southeast and if they vary by other covariates, such as forest type, species, age, size, etc. We also want to examine elevation, forest cover type, FIA unit, physiographic region, and ecoregion because these covariates can be determined at prediction points (i.e., where no FIA plots are present).

Figure 1.—*Relative county-level estimates of the number of trees killed by insects from newest Forest Inventory and Analysis plot-level data for each southeastern State as of August 2006.*



[1] Graduate Student, The University of Georgia, Warnell School of Forestry and Natural Resources, Athens, GA.

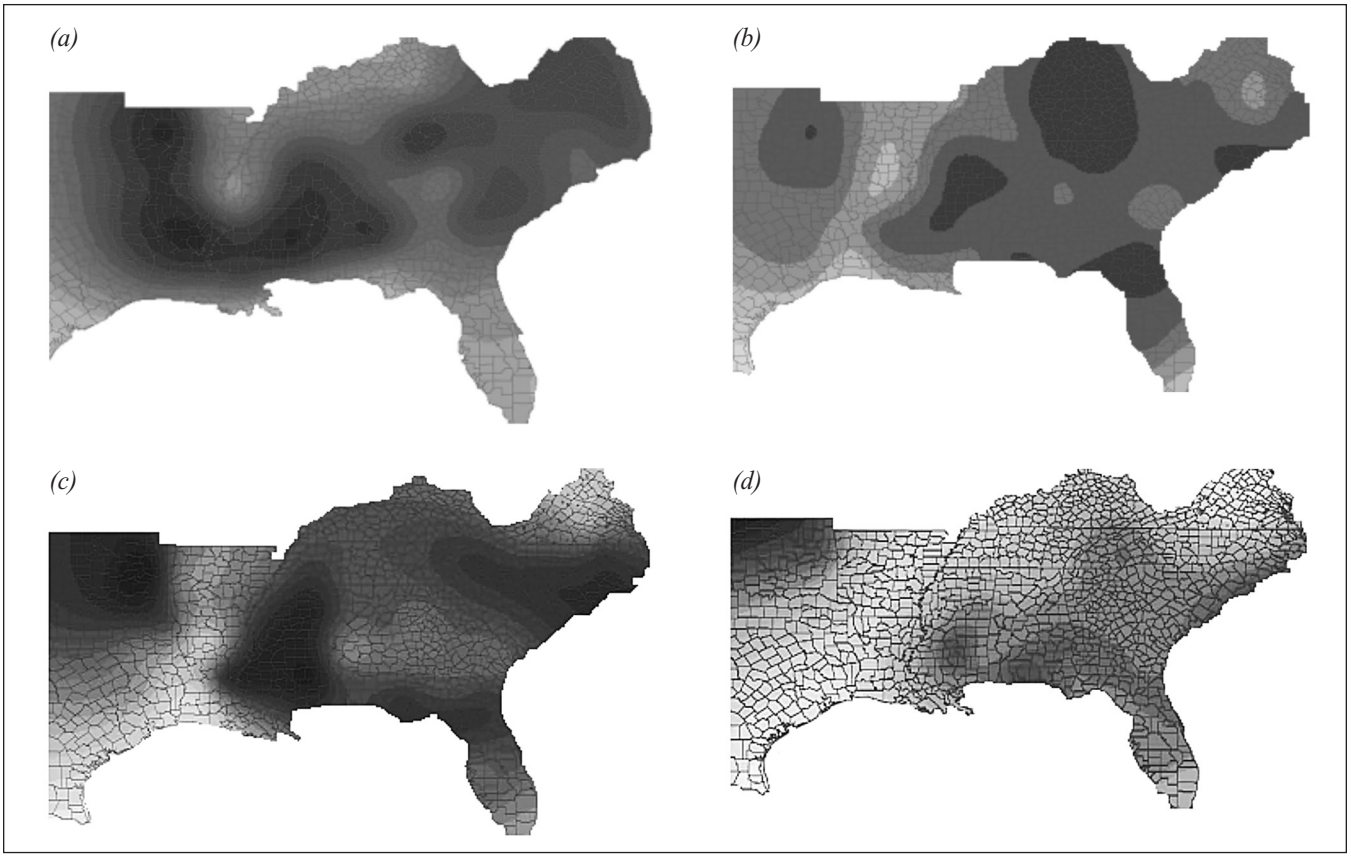
[2] Associate Professor The University of Georgia, Warnell School of Forestry and Natural Resources, Athens, GA.

[3] Quantitative Analysis Program Manager, U.S. Department of Agriculture, Forest Service, Forest Health Technology Enterprise Team, Fort Collins, CO. E-mail: elsmith@fs.fed.us.

Examples of maps produced from interpolating FIA data include those in figure 2. Although these maps were made after fitting semiparametric penalized spline models (SPSMs) (Ruppert *et al.* 2003, Shabenberger and Gotway 2004), model diagnostics for these models indicate inadequate fits. In the following text, we discuss the possible shortcomings of trying to model data like this across the Southeast. As a baseline case for spatial modeling from the FIA phase 2 data, we compare an ordinary kriging (OK) (Cressie 1993, Shabenberger and Gotway 2004) of the number of trees on forested conditions to the forest area map produced by Zhu and Evans (1994) in figure 3. We can see that the OK procedure was at least successful in detecting major, broad-level trends indicated in Zhu and Evans' (1994) map; e.g., lack of forests along the Mississippi River and southern FL. We can also notice a strong relationship between prediction error and which State the plot-level data is in. This relationship can be attributed to different sampling intensities between States and even within a State.

Several difficulties occur when trying to model data such as plot-level mortality and sources of this mortality. First of all, these data are rare and overdispersed. For example, over the whole Southeast, roughly 95 percent of the plots with a forested condition had zero trees killed by insects, with a sample variance-to-mean ratio of approximately 12.1. This is an extreme case of zero-inflated data. In situations in which the data exhibit inflation on one value (e.g., zero), transformations merely move this inflation to another value. Forest attributes, in general, tend to have large local variability. Also, in gathering the most recent FIA phase 2 data for each State in the Southeast, we had to use data for each State from different measurement years. Aside from any temporal differences that might actually exist in the data (e.g., one year with high mortality rates over the whole region), data collection methods may vary due to changes in FIA sampling procedures. In fact, data collection procedures may vary from State to State anyway, especially for more "obscure" plot values such as sources of mortality.

Figure 2.—*Trend maps from semiparametric penalized spline models predicting (a) forest area, (b) the number of trees on plots killed by insects,(c) the number of trees on plots killed by disease, and (d) the number of trees on plots killed by fire.*



To illustrate this, we provide figure 4. Here, we can see clear trends across States, which are due either to varying numbers of plots in the data, data collection procedures, or both. Even the newest data for SC did not contain information on sources of mortality (i.e., AGENTCD). Due to these difficulties, we resorted to fitting models from plots in only AL and GA. If the data were more consistent across States, we might hope to handle the situation of no data for one State by extrapolating into the State and keeping track of prediction error.

Figure 3.—*Zhu and Evans' (1994) (a) forest coverage map, (b) ordinary kriging prediction on the total number of trees on forested conditions in plots, and (c) prediction error.*

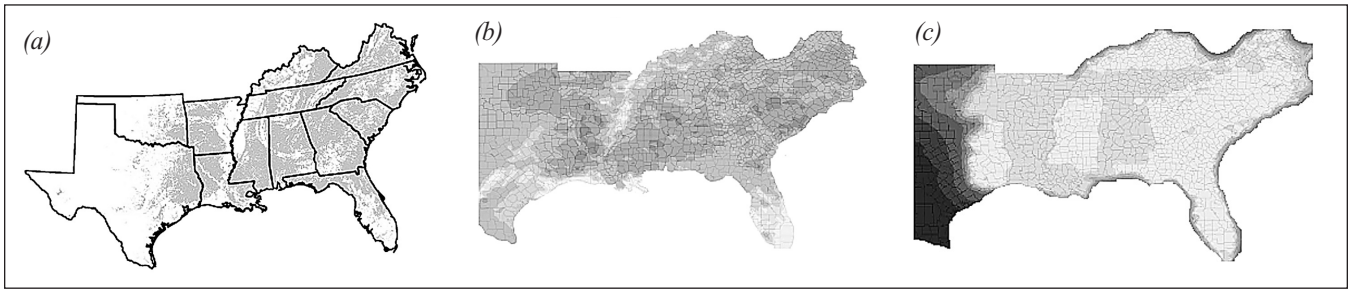
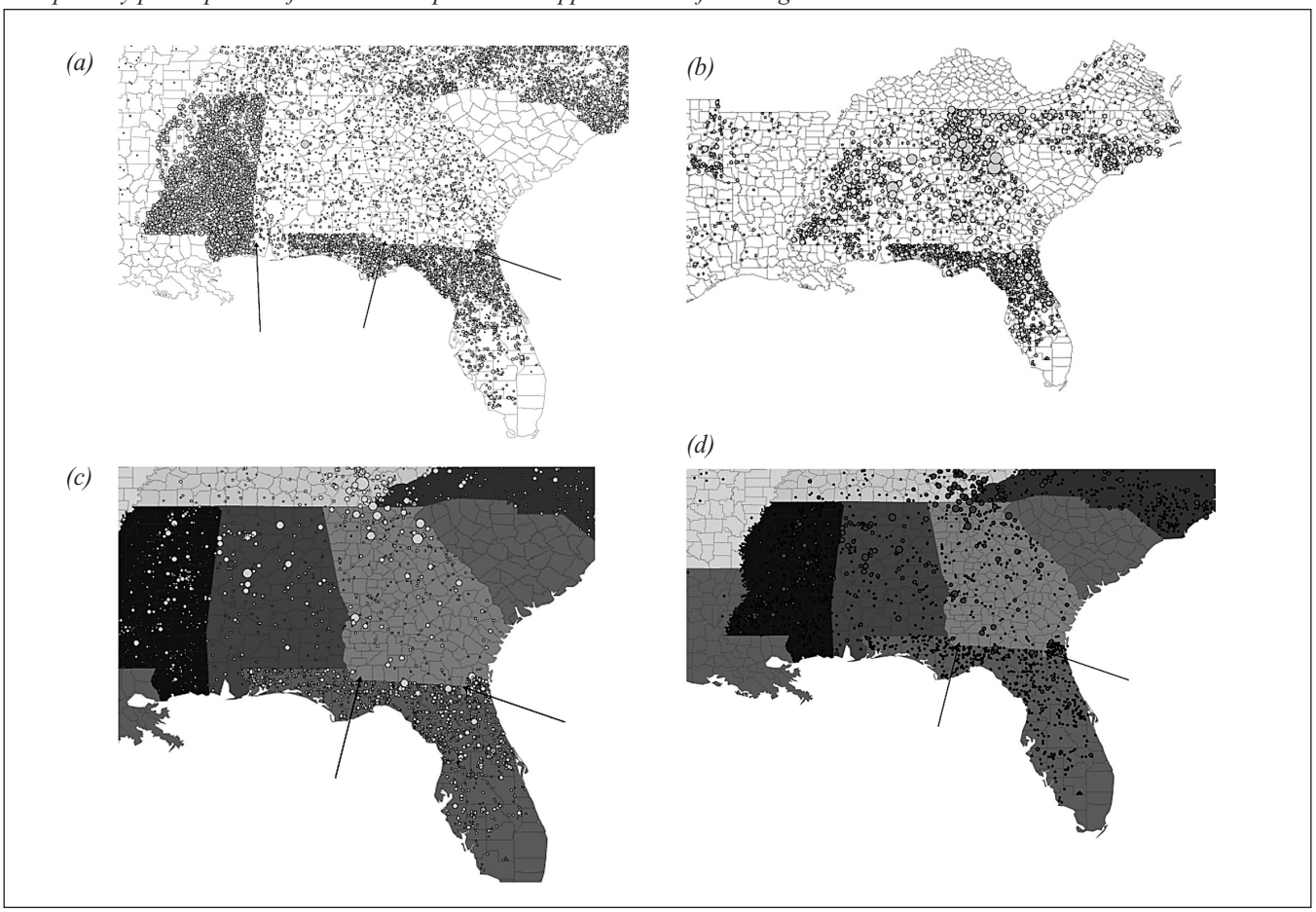


Figure 4.—*Forest Inventory and Analysis phase 2 plots with (a) symbol proportional to number of trees on plot killed by disease, (b) the number of trees on plots killed by insects, (c) a closer view of the same, and (d) the number of trees on plots killed by insects multiplied by plot expansion factor. Arrows point to an apparent trend following State boundaries.*



Materials and Methods

We fit several models to the number of trees on plots in AL and GA killed by insects. Instead of kriging, or geostatistical models, we try spatial regression models (Cressie 1993, Shabenberger and Gotway 2004). We chose these models for several reasons. A large number of FIA plots are present, even for one State, and standard Kriging techniques require inverting an n-by-n matrix, where n equals the number of plots. This may be prohibitively expensive computationally, and many authors have chosen ad hoc methods to deal with kriging on large data (e.g., over a moving window). One interesting method that we do not explore here is the fixed rank kriging method of Cressie (2006) and Johannesson and Cressie (2004). This method still uses all of the data, but necessitates inverting only smaller matrices. Also, the data we work with here exhibit extreme departures from the standard Gaussian distribution. Instead of fitting a trend to the data first and then kriging the residuals, we try direct models to the data. Although universal kriging and nonlinear kriging methods may work for these data, the problem of large data size still exists, and we do not explore them here. Hence, we chose to explore models that are low rank in that k knots are selected in the domain where k<<n. The resulting models are then spline functions connected at the knots. Theory and software is also readily available to extend these models to the generalized situation of various distributions assumed on the response.

Zero-inflated data are not rare in real-world data, and much effort has recently been applied to finding techniques for fitting models to them. Zero-inflated data, as the name implies, are data exhibiting a large number of zeros. These types of data can be found in many disciplines and often are the result of rare count data. Lambert (1992) provided techniques for modeling data with a zero-inflated Poisson (ZIP) model. Incorporating zero-inflated likelihoods in spatially explicit models is described in Agarwal *et al.* (2002), Barry and Welsh (2002), Fahrmeir and Echavarria (in press), Gschlobl and Czado (2006), Rathbun and Fei (2006), Rigby and Stasinopoulos (2005) and in general (not spatial) in Hall (2000), Lambert (1992), and Li *et al.* (1999).

We used data from GA cycle 08 and AL cycle 07. We counted the raw (i.e., not expanded) number of trees killed by insects in forested conditions (LANDCLCD=1), indicated by AGENTCD=10. Plot species richness was determined by counting the number of unique species codes (SPCD) for trees in forested conditions on each plot. We used the R Project for Statistical Computing (http://www.r-project.org) packages SemiPar and Generalized Additive Models for Location, Scale and Shape to fit the models here. The SPSMs we use through SemiPar have a smoothing parameter fit via restricted maximum likelihood. Knots were automatically selected in SemiPar via a space-filling algorithm (Ruppert *et al.* 2003) with the default number of 50 knots.

Results and Discussion

A histogram of the number of trees on plots killed by insects for only plots having at least one tree killed by insects is given in figure 5. In figures 6 and 7, we give histograms of residuals,

Figure 5.—*Histograms of (a) number and (b) proportion of trees killed by insects on forested conditions of plots in Alabama and Georgia, only for plots with at least one tree killed by insects in a forested condition.*

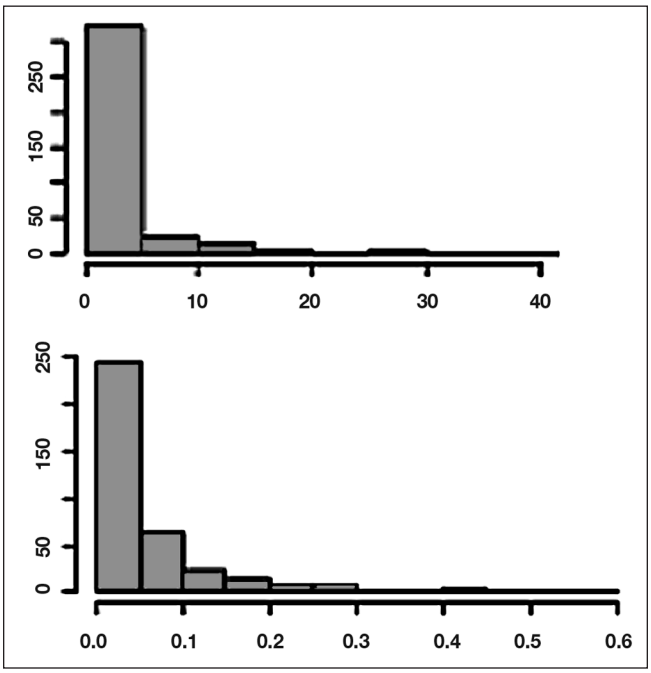


Figure 6.—*Semiparametric penalized spline model Poisson number of trees killed by insects: (a) residuals histogram,(b) Q-Q plot, and (c) histogram of fitted values.*

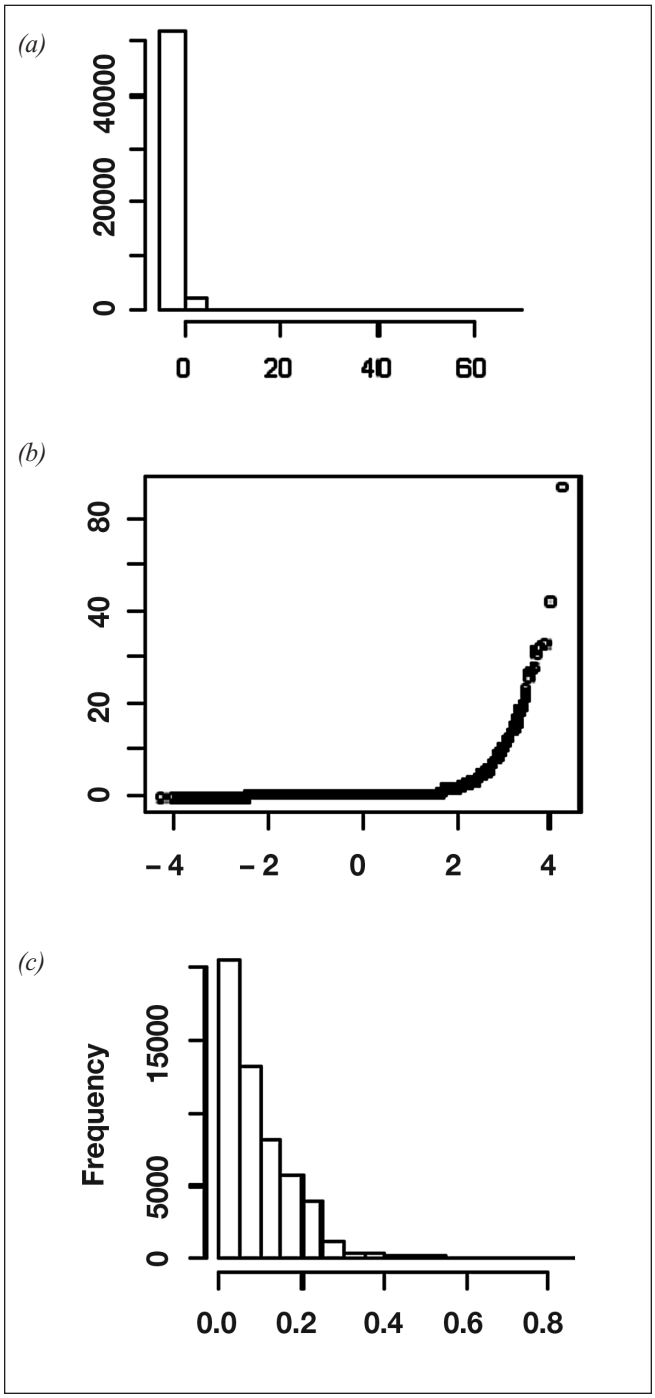
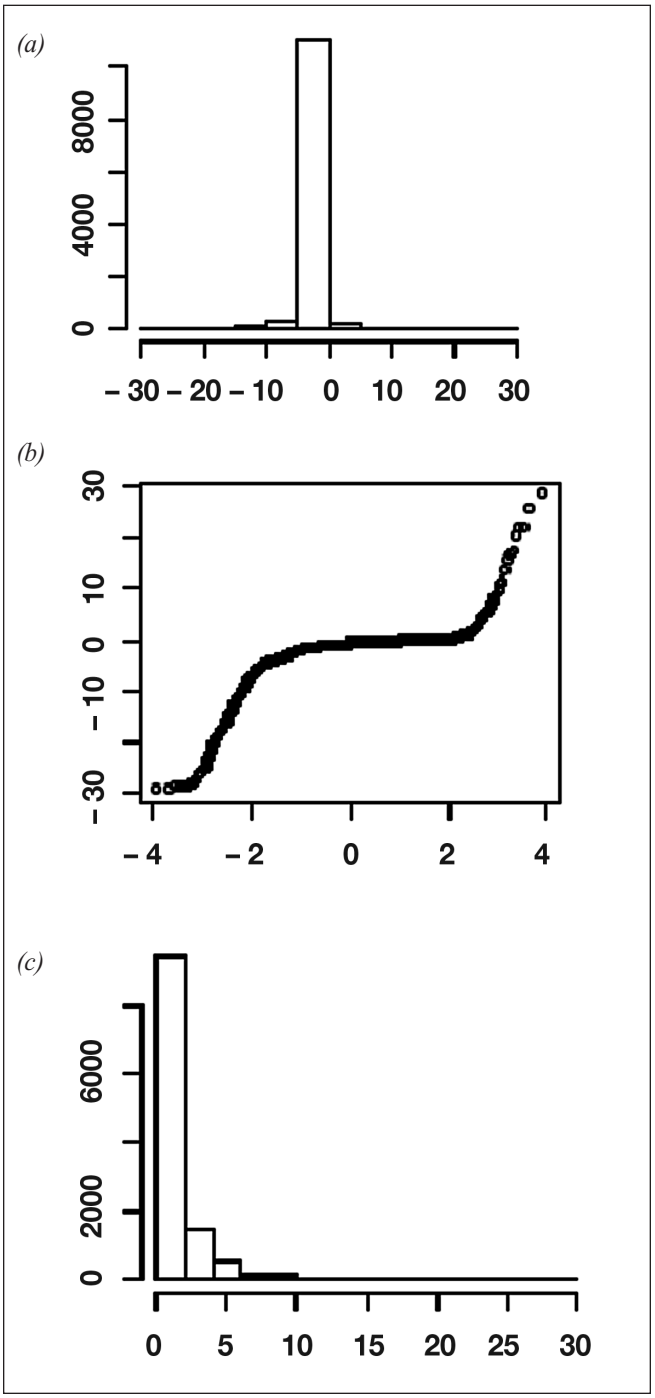


Figure 7.—*Semiparametric penalized spline model zero-inflated Poisson number of trees killed by insects: (a) residuals histogram. (b) Q-Q plot, and (c) histogram of fitted values.*



Q-Q plots of residuals, and histogram of predicted values for a SPSM fit to the number of trees killed by insects with distribution imposed on the response being Poisson and ZIP (Lambert 1992); we give these same plots for a local polynomial regression (LOESS) model fit to the number of trees killed by insects with distribution imposed on the response being ZIP and negative binomial (NB). Poisson was chosen first because these are count data. As we can see, however, the ZIP and NB performed better, which would be expected because the data are overdispersed. For the two ZIP models (figures 7 and 8), the SPSM and LOESS, it is hard to tell which one was best. The SPSM seemed to reach out to the extremes of the data better but fit worse in the middle range of the data. See figures 8 through 11.

We next fit models to species richness on phase 2 plots in AL and GA. The histogram of the number of species on plots in AL and GA is given in figure 12. These data had a sample variance-to-mean ratio of 4.1. We scaled species richness to [0,1] by the maximum number of species on plots, and we show a fitted lognormal and inverse Gaussian distribution to these species richness distribution. We included covariates of elevation, forest cover type, FIA unit, physiographic regions, and ecoregion. These covariates were chosen because they could be determined for prediction points where no FIA plots are present.

Figure 8.—*Local polynomial regression zero-inflated Poisson number of trees killed by insects: (a) residuals histogram and (b) Q-Q plot.*

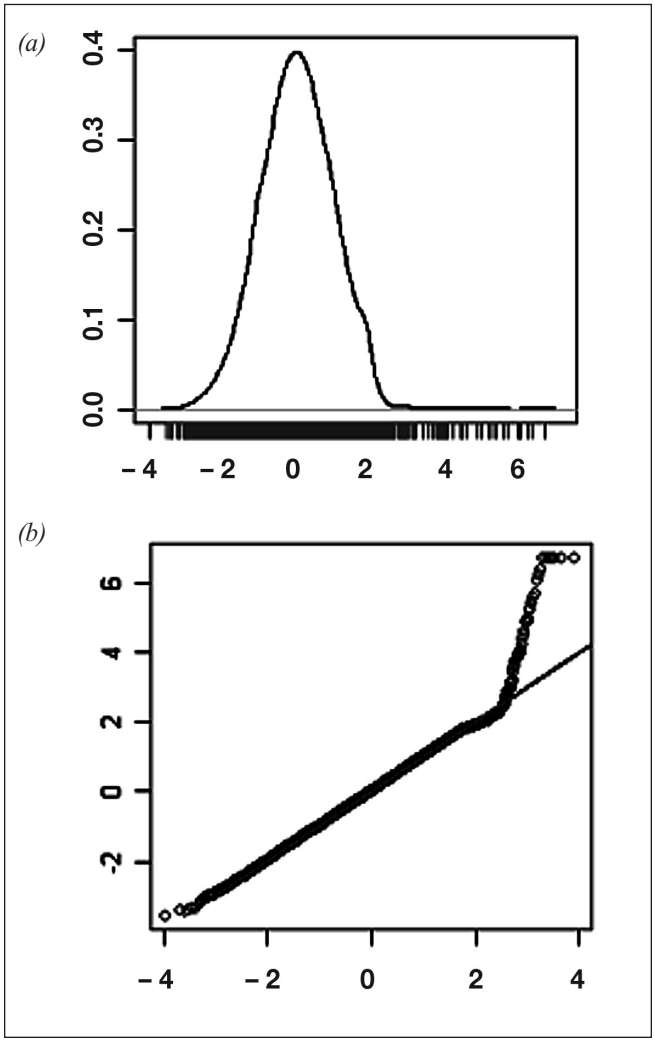


Figure 9.—*Local polynomial regression negative binomial number of trees killed by insects: (a) residuals histogram and (b)Q-Q plot.*

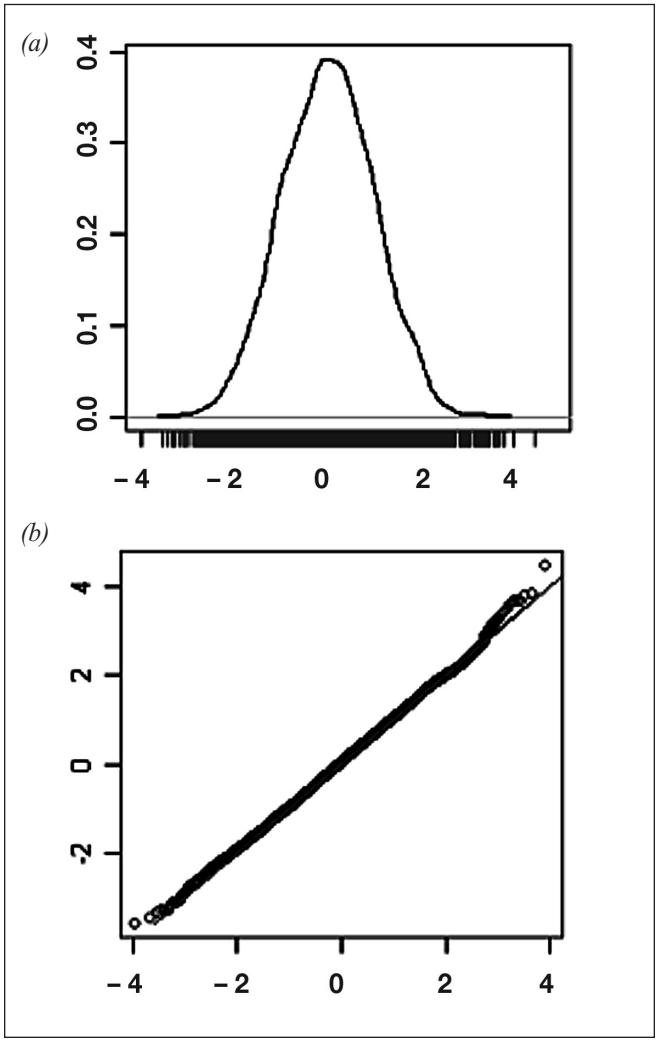


Figure 10.—*Local polynomial regression proportion of trees killed by insects: (a) residuals histogram, (b) Q-Q plot, and (c) histogram of fitted values.*

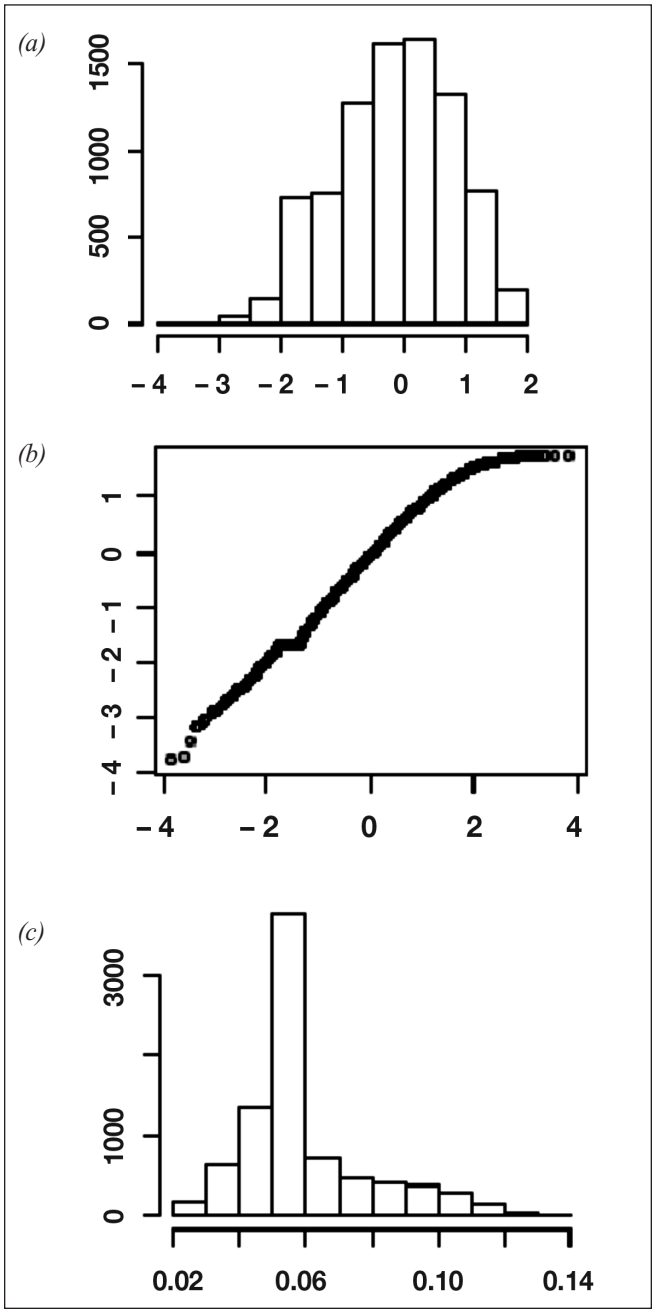


Figure 11.—*Local polynomial regression BI proportion of trees killed by insects: (a) residuals histogram, (b) Q-Q plot, and (c) histogram of fitted values.*

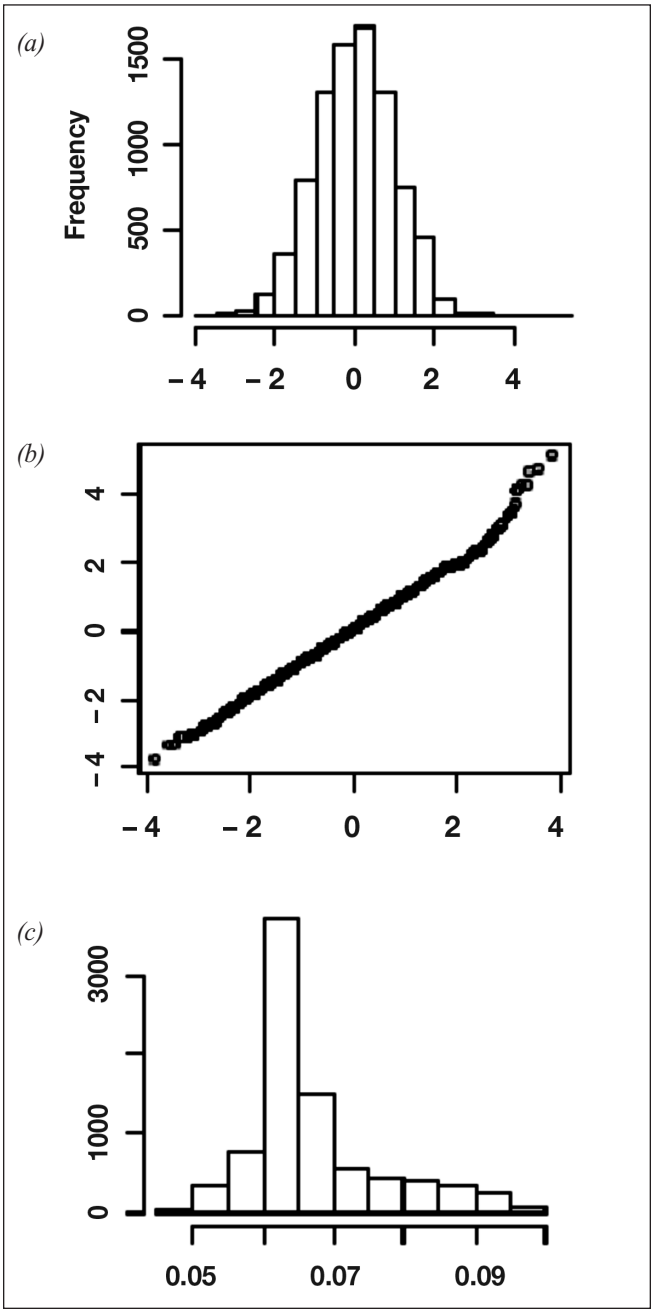
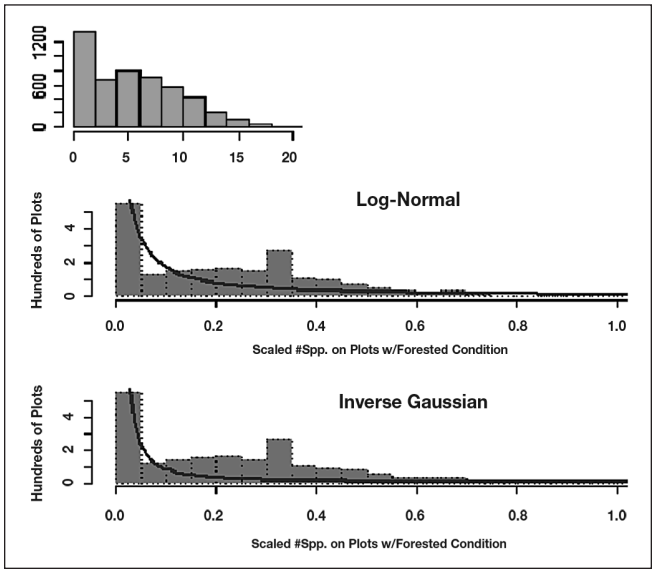


Figure 12.—*Top: histogram of species richness counts on plots in AL and GA with at least one forested condition. Middle and bottom: Two distributions fitted to the scaled species counts.*



In figures 13 and 14, we can see that models fit to the less dispersed plot-level value of scaled species richness, including covariates, have better residuals and histograms of fitted values more resembling the histogram of the measured values. Nevertheless, the models we chose so far, even though flexible, could not handle the extreme variation in the data we tried to model. We will continue with other spatial methods to test for significance. We will also include information from remote sensing, such as the FHM aerial surveys.

Figure 13.—*Local polynomial regression beta-inflated species richness, scaled to [0,1], only on plots in AL and GA with forested condition: (a) residuals histogram, (b) Q-Q plot, and (c) histogram of fitted values.*

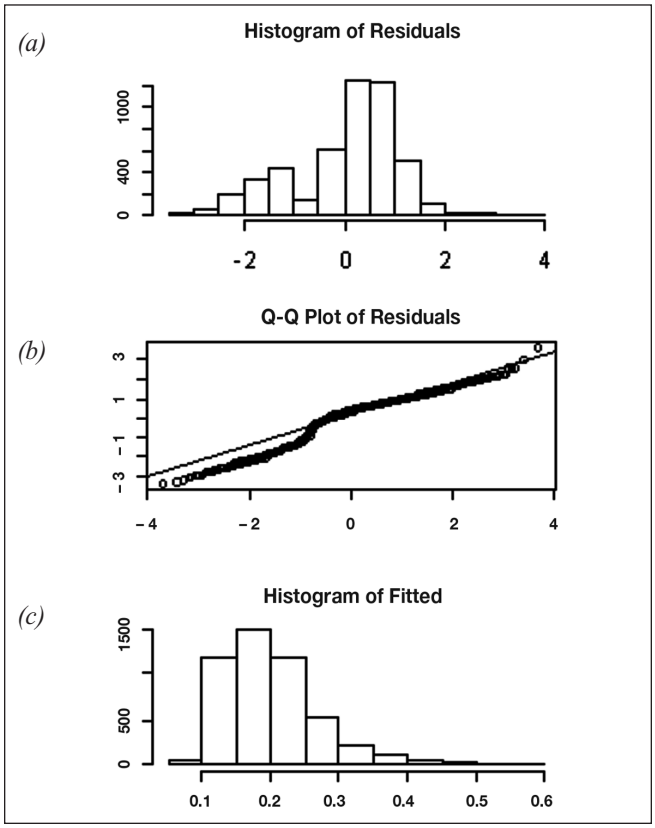
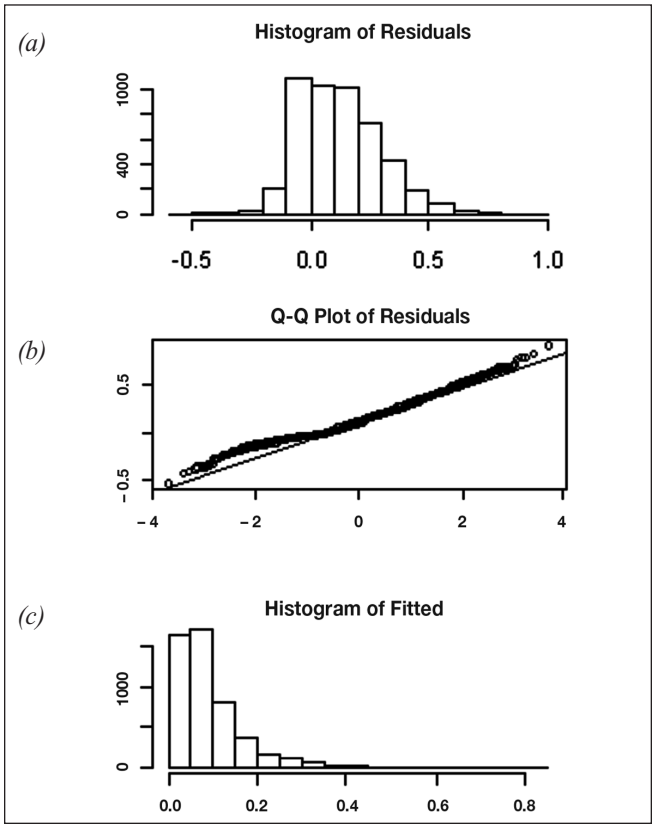


Figure 14.—*Local polynomial regression lognormal species richness, scaled to [0,1], only on plots in AL and GA with forested condition: (a) residuals histogram, (b) Q-Q plot, and (c) histogram of fitted values.*



Literature Cited

Agarwal, D.K.; Gelfand, A.E.; Citron-Pousty, S. 2002. Zero-inflated models with application to spatial count data. Environmental and Ecological Statistics. 9: 341-355.

Barry, S.C.; Welsh, A.H. 2002. Generalized additive modelling and zero inflated count data. Ecological Modelling. 157: 179-188.

Cressie, N.A. 1993. Statistics for spatial data. New York: John Wiley. 900 p.

Cressie, N. 2006. Spatial prediction for massive datasets. http://wwwrses.anu.edu.au/cadi/Whiteconference/papers/CressieMassiveData.pdf. (August).

Fahrmeir, L.; Echavarría, L.O. 2006. Structured additive regression for overdispersed and zero-inflated count data. Applied Stochastic Models Business Industry. 22: 351-369.

Gschlobl, S.; Czado, C. 2006. Modelling count data with overdispersion and spatial effects. http://www.stat.uni-muenchen.de/sfb386/papers/dsp/paper475.pdf. (August).

Hall, D.B. 2000. Zero-inflated Poisson and binomial regression with random effects: a case study. Biometrics. 56: 1030-1039.

Johannesson, G.; Cressie, N. 2004. Finding large-scale spatial trends in massive, global, environmental datasets. Environmetrics. 15: 1-44.

Lambert, D. 1992. Zero-inflated Poisson regression, with an application to defects in manufactoring. Technometrics. 34(1): 1-4.

Li, C.; Lu, J.; Park, J.; Kim, K.; Brinkley, P.A.; Peterson, J.P. 1999. Multivariate zero-inflated Poisson models and their applications. Technometrics. 41(1): 29-38.

Rathbun, S.L.; Fei, S. 2006. A spatial zero-inflated Poisson regression model for oak regeneration. Environmental and Ecological Statistics. 13: 409-426.

Rigby, R.A.; Stasinopoulos, D.M. 2005. Generalized additive models for location, scale and shape. Applied Statistics. 54(3): 507-554.

Ruppert, D.; Wand, M.P.; Carroll, R.J. 2003. Semiparametric Regression. New York: Cambridge University Press. 385 p.

Shabenberger, O.; Gotway, C.A. 2004. Statistical methods for spatial data analysis. Boca Raton, FL: Chapman & Hall. 488 p.

Zhu, Z.; Evans, D.L. 1994. U.S. forest types and predicted percent forest cover from AVHRR data. Photogrammetric Engineering & Remote Sensing. 60(5): 525-531.