Wildfire Chances and Probabilistic Risk Assessment

D. R. Brillinger[1], H. K. Preisler[2] and H. M. Naderi[1]
Statistics Department[1]
University of California
Berkeley, CA, 94720-3860
Pacific Southwest Research Station[2]
USDA Forest Service
Albany, CA, 94710

Keywords: biased sampling, false discovery rate, forest fires, generalized mixed model, penalized quasi-likelihood, risk

Abstract

Forest fires are an important societal problem in many countries and regions. They cause extensive damage and substantial funds are spent preparing for and fighting them. This work applies methods of probabilistic risk assessment to estimate chances of fires at a future time given explanatory variables. One focus of the work is random effects models. Questions of interest include: Are random effects needed in the risk model? If yes, how is the analysis to be implimented? An exploratory data analysis approach is taken employing both fixed and random effects models for data concerning the state of Oregon during the years 1989-1996.

1. Introduction

In simplest terms this work seeks probabilities associated with the occurrence of wildfies, for example the probability that a fire might occur in a specified region during some given day, week or month. Explanatory variables such as elevation and fire indices are available. In particular one may wish to estimate

$$Prob\{fire\ in\ particular\ region\ and\ time\ period \mid explanatories\}$$

where the (future) time period may be short or long term. Various sources of variability arise. One follows from the response, Y, being binary and Bernoulli distributed. Another follows from the year to year changes. This last is dealt with here by introducing a random effect for year. One then wishes an estimate of

$$E_I\{Prob\{fire\ in\ particular\ region\ and\ time\ period \mid explanatories,\ year\}\}$$

Spatial and daily (fixed) effects are included in the model as smooth additive functions $g_1(x,y)$ and $g_2(d)$ respectively. The year effect is also additive to the linear predictor. Two estimates are considered for estimating $var\{year\}$. One estimate is the sample variance of fixed year effects estimates obtained by fitting as if they were constant. The second estimate is obtained via quasi-likelihood estimation. In the case at hand the results are similar.

The study presented involves fire data for the Federal Lands in Oregon during the period 1989 to 1996. Further details of the data and work may be found in Brillinger et al (2003) and Preisler et al (2004).

2. Statistical background The model employed here is a particular case of the following generalized mixed effects model. The vector Y, and the matrices X and Z are given. The linear predictor takes the form

$$\eta \ = \ X\alpha \ + \ Zb$$

with the vector α fixed and the random vector b normal with mean 0 and covariance matrix R. Further there exists an inverse link function h such that

$$E\{Y|b\} \ = \ h(\eta)$$

It is further assumed that

$$var\{Y|b\} \ = \ v(\eta)$$

with v a known function. These assumptions may be used to set down a penalized quasi-likelihood function, Green (1987), Schall (1991), Breslow and Clayton (1993) and an iterative scheme may be set up to evaluate the parameters. The function glmmPQL() of Venables and Ripley (2002), which employs the linear mixed effects function lme() of Pinheiro and Bates (2000), is used in the computations presented below. The estimates obtained may be used in turn to estimate $E\{b|Y\}$.

A common alternative to attacking the model directly is to employ a fixed effects procedure and take the random effects as parameters and then compute the sample variance of the estimates obtained.

There is an added aspect to the present circumstance; the model proposed includes an offset term. With the quasi-likelihood approach this does not cause any difficulty.

In Section 4.1 below the False Discovery Rate (FDR) is used to describe the certainty of an estimated spatial effect. A null hypothesis that there is no spatial effect is considered. The FDR is defined as the fraction of false rejections of the null hypothesis among all rejections. It is useful here because hypotheses are being examined for each spatial pixel. There are various FDR procedures for controlling the indicated rate. The one used here is due to Benjamini and Yekutieli (2001). The FDR values presented below were computed via the Splus function listed in "www.math.tau.ac.il/ roee/FDR_Splus.txt".

3. The case of Oregon

The fires studied occurred in Oregon Federal lands during the period 1989-1996. These lands make up much of the state. Data for the months of April through September were employed. Spatial pixels that were 1km by 1km were employed. This meant that the data set was very large, 578,192,400 voxels. Because of this only a sample of the locations where no fires occurred was employed. The number of fires was 13,834 and all the voxels with fires were included. The chosen sampling fraction of the voxels with no fires was $\pi = .00012$ leading to a total of 41,279 no-fire cases. Further details, including maps, may be found in Brillinger et al. (2003) and Preisler et al (2004).

Having in mind the problem and the data available a number of models may be set down. Set $Y = 1, 0$ according as to whether there has been a fire or not in a particular region and time period. Suppose the fixed explanatories employed are location, (x, y), and day of the year, d. The further explanatory year, I, is taken as fixed in Model II and random in Model III.

In the models g_1 and g_2 are respectively splines for location and day of year.

Model I. With Y binary-valued and (x, y) and d fixed

$$logit\, Prob\{Y = 1|\ x,\ y,\ d\} \ =\ g_1(x, y)\ +\ g_2(d) \tag{1}$$

Model II. With I a fixed factor for year

$$logit\, Prob\{Y = 1|\ x,\ y,\ d,\ I\} \ =\ g_1(x, y)\ +\ g_2(d)\ +\ I \tag{2}$$

Model III. With I a factor whose effects are independent normals with mean 0 and variance σ^2

$$logit\, Prob\{Y = 1|\ x,\ y,\ d,\ I\} \ =\ g_1(x, y)\ +\ g_2(d)\ +\ I \tag{3}$$

The distribution of Y in Model III is logit-normal. Writing

$$\eta \ =\ g_1(x, y) + g_2(d)$$

for Model III, with the normality of I, the probability of a fire is

$$Prob\{Y\ =\ 1\ |\, explanatories\} \ =\ \int exp\{\eta + \tau z\}/(1 + exp\{\eta + \tau z\})\phi(z)dz$$

The distribution of Y in Model III is logit-normal.

An added detail of the current set up is that the no-fire cases were sampled. Interestingly, with the logit link, one has a generalized linear model with an offset of $log\ (1/\pi)$. The new logit is simply $logit\ p' \ =\ logit\, p\ +\ log(1/\pi)$, i.e. an offset. See Maddala (1992), page 330. This meant that standard generalized linear model computer programs could be used for the analysis. The offset is not indicated in expressions (1), (2) or (3).

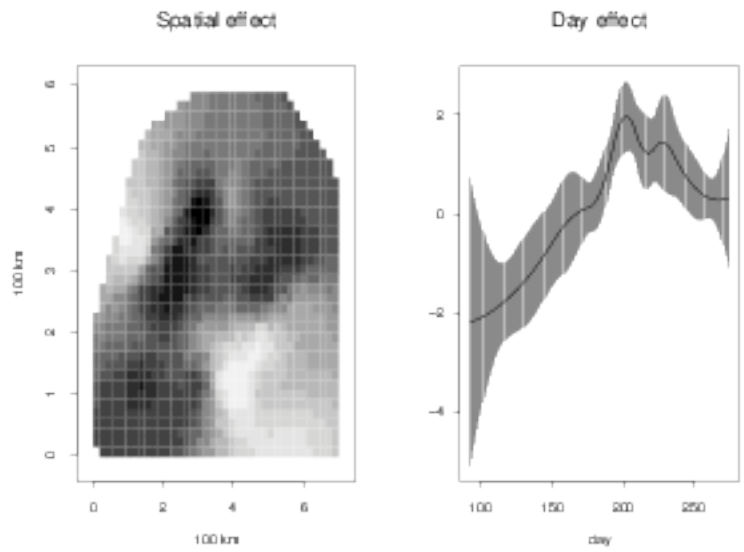

Figure 1: The estimated and daily effects for Model I. The lefthand panel provides an image plot of the estimated spatial effect. The darker values correspond to increased fire risk. In the righthand panel the vertical lines provide approximate 95% confidence limits about a smoothed version of the solid line.

4. Results

4.1 Fitting and assessing Models I and II

Model I involves fitting the sum of smooth functions, specifically splines, of (x, y) and d. Figure 1 shows the fitted spatial and day effects. The spatial effect is seen to be less in the SE corner of the state. This part of Oregon is desert-like. The righthand panel of Figure 1 shows the fitted effect of day as a solid curve. The effect is seen to peak around day 200 and be small for April and September. The vertical lines about the curve provide approximate 95% marginal confidence limits graphed about a smoothed form of $\hat{g}_2(d)$.

Figure 2, lefthand panel, provides the estimated spatial effect in contour. A map outlining Oregon has been superposed. Again one notes the reduced risk in th SE part of the state. The righthand panel shows the results of controlling the overall False Discovery Rate at level .05 . There is strong evidence for a spatial effect around much of the region of study.

Projections are of basic interest in this work, for example the expected fire count for a given region and months. To obtain these one sums probabilities over pixels of the selected region, and days of the year, specifically one computes

$$\sum_i exp\{\hat{\eta}_i\}/(1 + exp\{\hat{\eta}_i\}) \tag{4}$$

where i sums over the pixels in the region, the days of the selected month and

$$\hat{\eta}_i \;=\; \hat{g}_1(x_i, y_i) \;+\; \hat{g}_2(d)$$

As an example, predictions are presented for the Umatilla Forest, a region of size 1100 km^2 in the Federal Lands of Oregon. It is shown on a map in Brillinger et al (2003). Figure 3 provides estimates of the expected count of fires for Umatilla each month derived via expression (1). The vertical lines are approximate marginal 95% confidence limits derived via a jackknife dropping single years successively. One sees the expected count peaking, just below a level of 6 fires, in July. The distribution of a future count can be approximated by a Poisson with the indicated estimated expected count.

A residual plot, see Figure 7 in Section 4.3, suggested that a year effect needed to be included in the model. When Model II was fit the change in deviance resulting from doing so is 1139.7 with 7 degrees of freedom. Years are included in the next model.

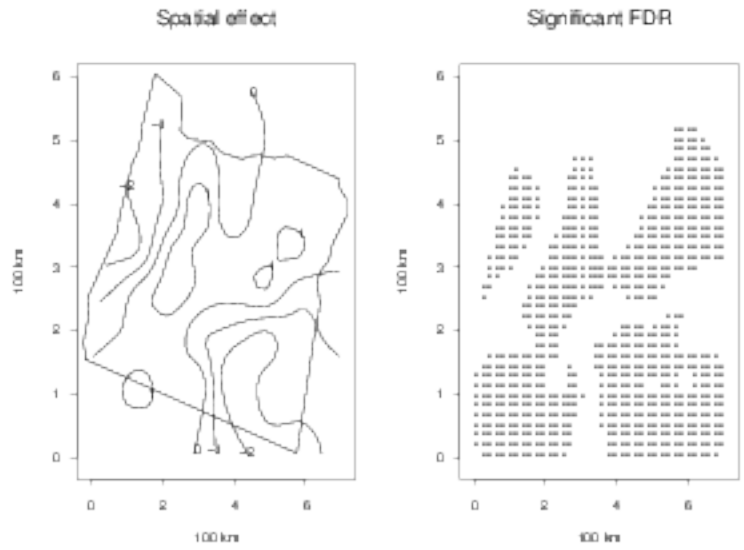

Figure 2: The lefthand panel provides the estimated spatial effect in contour form. An outline of Oregon has been superposed. The righthand shows the region found significant by the FDR analysis.

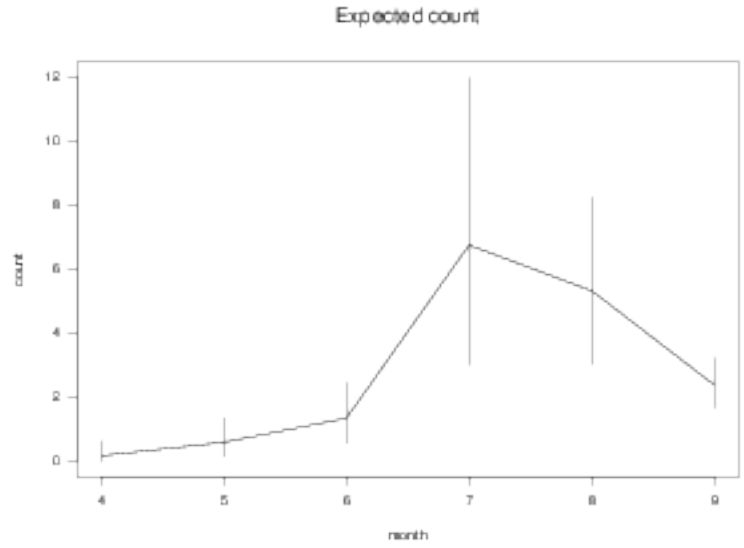

Figure 3: Model I predictions of expected counts for Umatilla Forest based on expression (4). The vertical lines provide approximate 95% marginal confidence intervals.

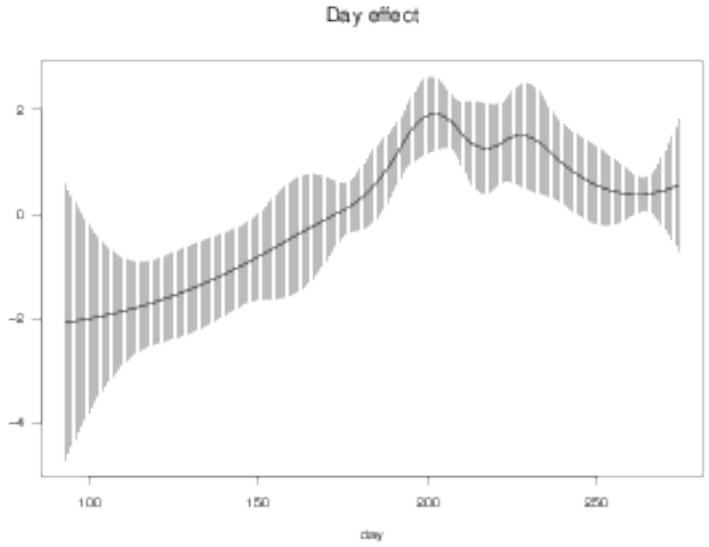

Figure 4: The estimated spatial and daily effects for Model III.

4.2 Results of fitting Model III

Next consideration turns to Model III which includes a random effect for year. The fitting is carried out via penalizied quasi-likelihood implimented as the function glmmPQL(), Venables and Ripley (2002). Doing so one obtains for the standard error of the year effect $\hat{\tau} = .429$. The program output also includes $X/\sqrt{n} = 1.000$, where X is square root of the estimate of the dispersion statistic. This value suggests that the logistic-normal model is fitting the data rather well.

Figure 4 shows the estimated daily effects. There are not great changes from the Model I results presented in Figure 1.

An estimate of the expected number of fires for some region and future occasion is provided by

$$\sum_i \int exp\{\hat{\eta}_i + \hat{\tau}z\}/(1 + exp\{\hat{\eta}_i + \hat{\tau}z\})\phi(z)dz \tag{5}$$

with i again labelling the pixels of the region of concern and the days of the month, and $\hat{\eta}_i = \hat{g}_1(x_i, y_i) + \hat{g}_2(d_i)$. The integral in (5) is evaluated numerically and the results are given in the lefthand panel of Figure 5. Again the uncertainties are derived via the jackknife.

In some circumstances an estimate of

$$Prob\{At\ least\ one\ fire\ in\ a\ particular\ region\ and\ month\}$$

is desired. With an approximate Poisson process of intensity of fires $\mu(x', y', d')$, and a region M this probability is given by

$$1 - exp\{-\int_M \mu(x', y', d')dx'dy'dd'\}$$

whose integrand may be approximated by expression (4). The results are given in the righthand panel of Figure 5. The lefthand panel may be compared with Figure 3.

Figure 6 gives estimates of the random year effects, that is $E\{I|data\}$ for each of the of the years 1989 - 1996. Such effects are sometimes referred to as shrunken effects.

Another way to estimate the variance of random effects is to fit them as if they were fixed, i.e. under Model II, and then to compute the sample variance of the values obtained. This leads to a standard error of the fixed effect year values as $\hat{\tau} = .464$, which is suprisingly close to the glmmPQL() produced value.

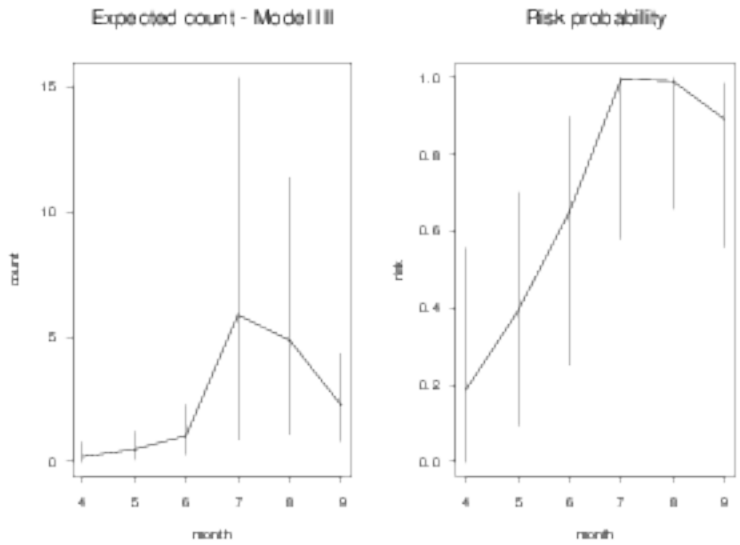

Figure 5: Model III predictions - expected counts and risks - for Umatilla Forest based on expression (2). The vertical lines are approximate 95% confidence intervals.

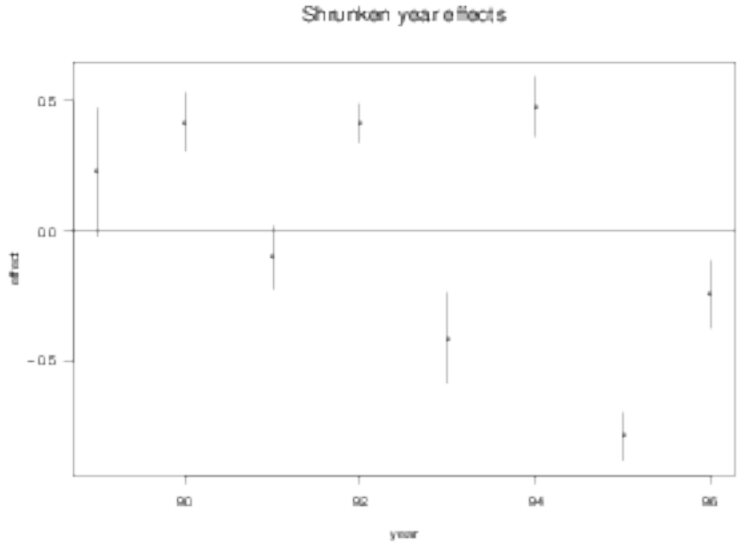

Figure 6: Shrunken year effects for Model III.

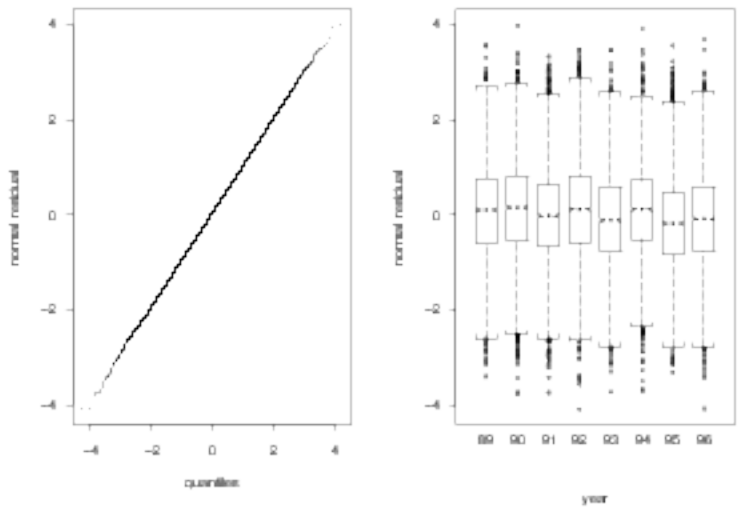

Figure 7: Model assessment results for Model I. The lefthand panel is a normal probability plot of the normal residuals, the righthand shows notched boxplots for each year's normal residuals.

4.3 Model assessment

The response is binary, $Y = 0,\ 1$ and so various of the classical model effect procedures are not particularly effective. However uniform residuals are an aid to assessing goodness of fit, in various such nonstandard cases, see Brillinger and Preisler (1983) and Brillinger (1996). In the binary case these may be computed as follows: suppose $Prob\{Y = 1|explanatories\} = \gamma$ and that U_1 and U_2 denote independent uniforms on the intervals $(0, 1-\gamma)$, $(1-\gamma, 1)$, respectively then the variate

$$U = U_1 * (1 - Y) + U_2 * Y$$

has a uniform distribution on the interval $(0,1)$. Whereas when $Prob\{Y = 1|explanatories\} = \gamma_0$, then $E\{U\} = (1 + \gamma - \gamma_0)/2$ for example. We refer to U, when an estimate of $\hat{\gamma}$ is employed, as a uniform residual and write $\hat{U}$. We refer to $\Phi^{-1}(\hat{U})$ as a normal residual. Working with the normal residuals has the advantage of spreading the values out. Various traditional residual plots may now be constructed, e.g. normal probability plots involving the normal residuals $\Phi^{-1}(\hat{U})$ or plots of $\Phi^{-1}(\hat{U})$ versus explanatories.

The righthand panel of Figure 7 shows notched boxplots of the normal residuals of Model I against the year. The graph suggests that year needs to be in the model as an explanatory. The lefthand panel provides a normal probability plot and there is not much to be concerned with.

Figure 8 provides similar plots for Model III, but now the explanatory in the righthand panel is elevation. Again one has an indication that a further explanatory should be included. The lefthand shows little evidence against the distributional assumptions made.

5 Summary and discussion.

The work that has been presented is preliminary and exploratory. It is meant to lead to possible approaches in the full modelling effort.

Two distinct models, one fixed effect and the other random effect, have been set down and studied. Their fit has been assessed by probability and residual plots.

One question in the work was whether a random effect was in fact needed. In answer to this is: Yes, the context of the situation and the desire for estimates of future probabilities more or less guarantees so. It was noted how close the estimate of the random effect variance derived from fixed effect modelling was to that based on random effects modelling.

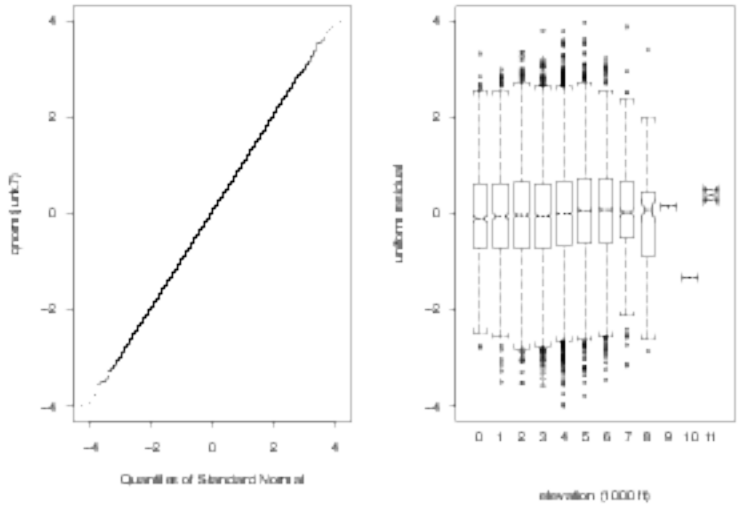

Figure 8: Model assessment results for Model III.

ACKNOWLEDGEMENTS

Alastair Scott commented on the necessity of investigating whether the offset trick really worked in the random effects case. Professor David Martell alerted us to the existence of the Maddala reference.

The work of DRB was supported by the NSF Grant DMS-02-03921 and the U.S. Forest Service Contract 02-JV-11272165-02. The work of HMN was supported by the VIGRE grant awarded to the Statistics Department, University of California, Berkeley.

REFERENCES

Benjamini, Y. and Yekutieli, D. (2001). The control of false discovery rate in multiple testing under dependency. *Ann. Statist.* 29, 1165-1188.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear models. *J. Amer. Statist. Assoc.* 88, 9-25.

Brillinger, D. R. (1996). An analysis of an ordinal-valued time series. *Lecture Notes in Statistics, Vol. 115*. Springer, New York.

Brillinger, D. R. and Preisler, H. K. (1983). Maximum likelihood estimation in a latent variable problem. *Studies in Econometrics, Time Series and Multivariate Statistics*. Academic Press, New York 31-65.

Brillinger, D. R., Preisler, H. K. and Benoit, J. (2003). Risk assessment: a forest fire example. *Statistics and Science: A Festschrift for Terry Speed*. Volume 40 IMS Lecture Notes.

Green, P. F. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review* 55, 245-259.

Maddala, G.S. (1992). *Introduction to Econometrics, Second Edition*. MacMillan, New York.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-effects Models in S-PLUS*. Springer, New York.

Preisler, H.K., Brillinger, D.R., Burgan, R.E., and Benoit, J.W. (2004). Probability based models for estimating wildfire risk. *International Journal of Wildland Fire*. To appear.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika* 78, pp. 719-727.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S-PLUS, 4th edition*. Springer, New York.