ENVIRONMETRICS
*Environmetrics* 2006; **17**: 623–633
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/env.768

# Probabilistic risk assessment for wildfires[‡]

D. R. Brillinger[1,*,†], H. K. Preisler[2] and J. W. Benoit[3]

[1] *Statistics Department, University of California, Berkeley, CA 94720-3860, U.S.A.*
[2] *Pacific Southwest Research Station, USDA Forest Service, Albany, CA 94710, U.S.A.*
[3] *Pacific Southwest Research Station, USDA Forest Service, Riverside, CA 92507, U.S.A.*

## SUMMARY

Forest fires are an important societal problem. They cause extensive damage and substantial funds are spent preparing for and fighting them. This work develops a stochastic model useful for probabilistic risk assessment, specifically to estimate chances of fires at a future time given explanatory variables. Questions of interest include: Are random effects needed in the risk model? and if yes, How is the analysis to be implemented? An exploratory data analysis approach is taken using both fixed and random effects models for data concerning the Federal Lands in the state of California during the period 2000–2003. Published in 2006 by John Wiley & Sons, Ltd.

KEY WORDS: biased sampling; false discovery rate; forest fires; generalized mixed model; penalized quasi-likelihood; risk

## 1. INTRODUCTION

The concern of the work is wildfires in the Federal Lands of California. A previous paper (Brillinger *et al.*, 2004) was concerned with the same problems but for the case of Oregon. For comparative purposes, this paper parallels it to a close extent. The goal is to estimate probabilities associated with the occurrence of wildfires, for example, the probability that a fire might occur in a specified region during some given day, week, or month of the year. Explanatory variables such as elevation and fire danger indices are available. In particular, one may wish to estimate

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

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

*Correspondence to: D. R. Brillinger, Statistics Department, University of California, 367 Evans, Berkeley, CA 94720-3860, U.S.A.

[†] E-mail: brill@stat.berkeley.edu

[‡] This article is a U.S. Government work and is in the public domain in the U.S.A.

Contract/grant sponsor: NSF; contract/grant number: DMS-02-03921.
Contract/grant sponsor: U.S. Forest Service; contract/grant number: 02-JV-11272165-02.

*Received 29 October 2004*
*Revised 14 July 2005*

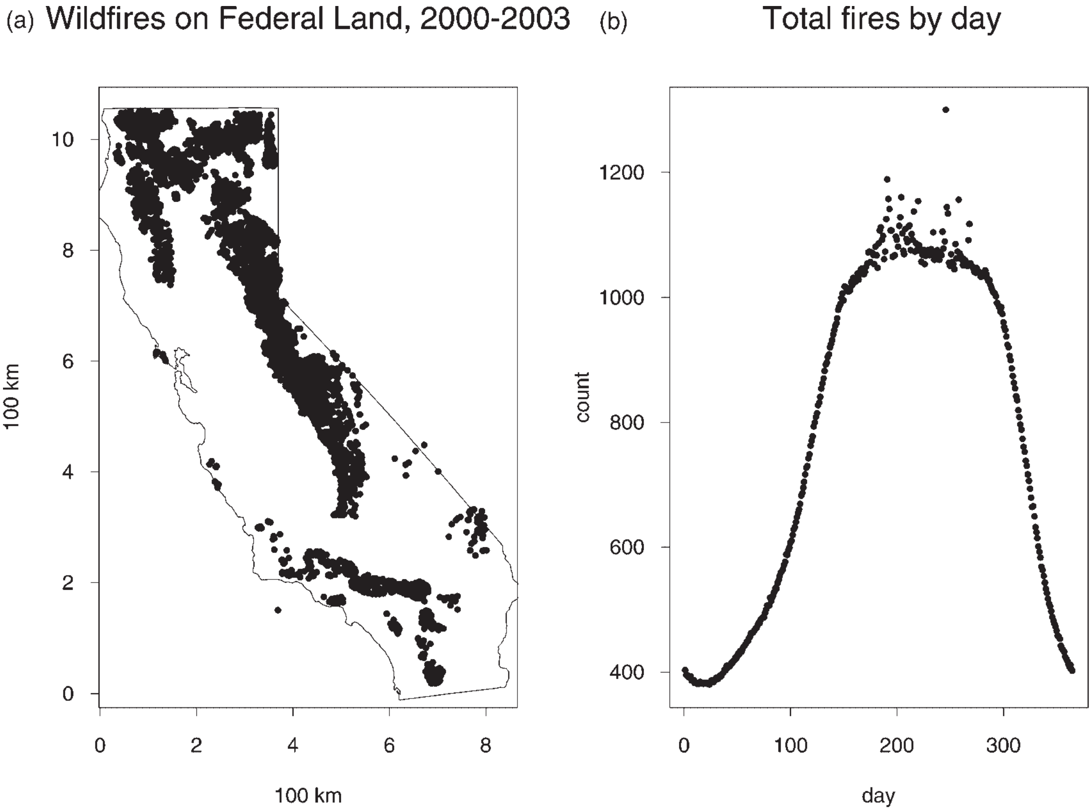

Figure 1. (a) Fire locations; the blocked off upper right area is the state of Nevada; (b) 4-year total for each day of the year

year changes. The latter is dealt with here by introducing a random effect, $I$, for year. One then wishes an estimate of

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

Spatial and daily (fixed) effects are included in the model as smooth additive functions $g_1(x, y)$ and $g_2(d)$, respectively of space $(x, y)$ and day of the year $d$. The year effect, $I$, is also additive to the linear predictor. An estimate of var$\{I\}$ is obtained via quasi-likelihood estimation.

Figure 1 shows the basic data in a derived form. Figure 1(a) gives the locations of all of the fires in the data set. Figure 1(b) gives the daily totals over the 4 years. The latter appears as a smooth curve plus quite a number of larger outliers in the summer fire season.

As an example of projecting daily counts the Yosemite National Forest is considered.

## 2. STATISTICAL BACKGROUND

The model used is a particular case of the generalized mixed effects model, for example (Breslow and Clayton, 1993). The vector $Y$, and the matrices $X$ and $Z$ are given. The linear predictor takes the form

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

with the vector $\alpha$ 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 also assumed that

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

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 uses 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 use a fixed effects procedure and take the random effects as parameters and then compute the sample variance of the estimates obtained.

In Sub-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).

## 3. THE CASE OF CALIFORNIA

The fires studied occurred in California Federal lands during the period 2000–2003. These lands make up an important part of the state. Data for the full year were used. Spatial pixels that were 1 km by 1 km were used. This meant that the data set was very large. Because of this only a sample of the locations where no fires occurred was used. All the space-time cells (voxels) with fires were included. For each day a sample of voxels with no fires was taken. The number of voxels sampled each day was proportional to the total number of fires on that day, proportional to a smoothed version of Figure 1b.

Having in mind the problem and the data available, a number of models may be considered. Set $Y = 1, 0$ according to whether there has been a fire or not in a particular spatial pixel and time period. Suppose, to begin, that the fixed explanatories used 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 below.

In the models, $g_1$ and $g_2$ are respectively smooth splines for location and day of the year.
*Model I*: With $Y$, binary-valued and $(x, y)$ and $d$ fixed

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

*Model II*: With $I$, a fixed factor for year

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

*Model III*: With $I$, a factor whose effects are independent normals with mean 0 and variance $\tau^2$

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

Writing

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

for Model III, and assuming $I$ to be normal, the probability of a fire is

$$\text{Prob}\{Y = 1|\text{explanatories}\} = \int \frac{\exp\{\eta + \tau z\}}{(1 + \exp\{\eta + \tau z\})} \phi(z)\text{d}z$$

with $\phi(.)$ the density function of the standard normal. 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 $\text{logit}\, p' = \text{logit}\, p + \log(1/\pi)$, i.e., an offset. One reference is Maddala (1992, p. 330). This meant that

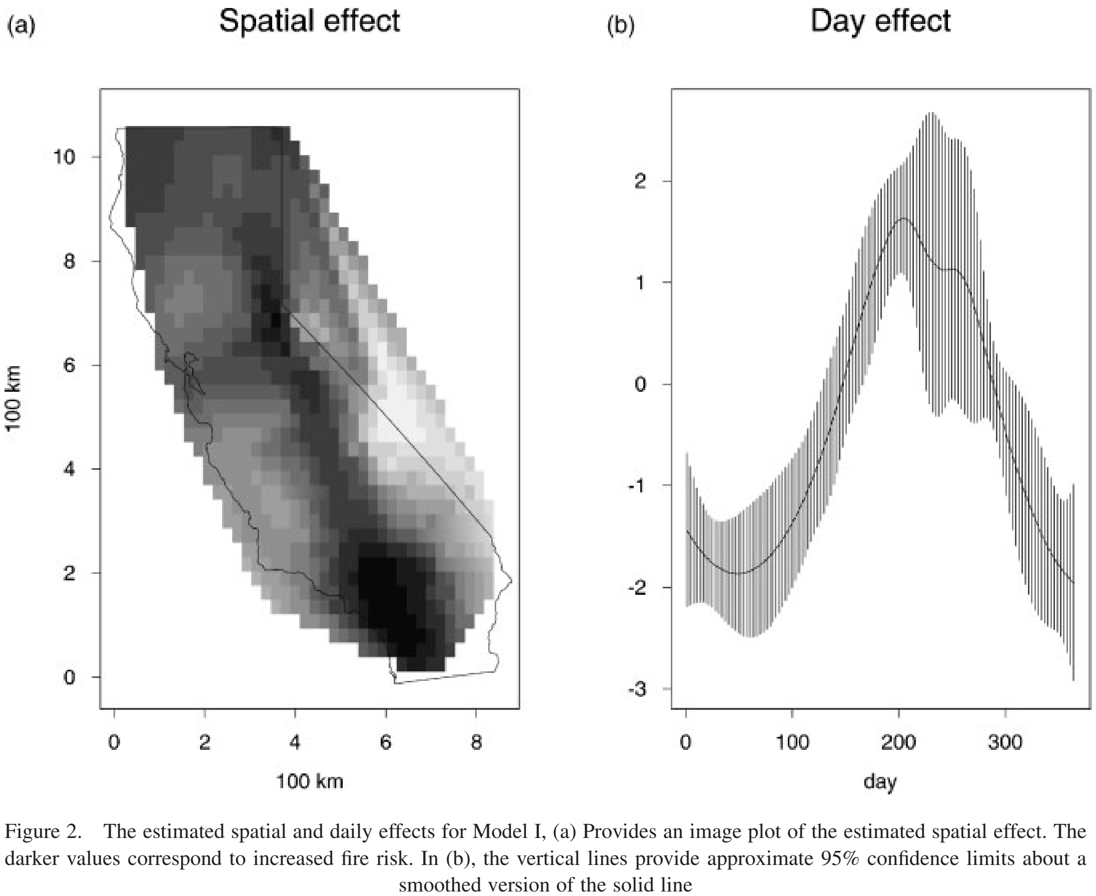

Figure 2. The estimated spatial and daily effects for Model I, (a) Provides an image plot of the estimated spatial effect. The darker values correspond to increased fire risk. In (b), the vertical lines provide approximate 95% confidence limits about a smoothed version of the solid line

standard generalized linear model computer programs could be used for the analysis. The offset is not indicated in expressions (1), (2), or (3) as they are the basic models.

## 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 2 shows the fitted spatial and day effects. Figures 2(a) and 2(b) look like smoothed version of Figure 1. Figure 2(b) shows the fitted effect of day as a solid curve. The effect is seen to peak around day 200 and be smaller for September through April. The vertical lines about the curve provide approximate 95% marginal confidence limits graphed about a smoothed form of $\hat{g}_2(d)$. The limits were computed via a jackknife dropping years in turn.

Figure 3(a) provides the estimated spatial effect in contour form. As in Figure 2(a) one notes reduced risk in the eastern and western parts of the state. Figure 3(b) shows the results of controlling the overall FDR at level 0.01. There is strong evidence for a spatial effect around much of the region of study.

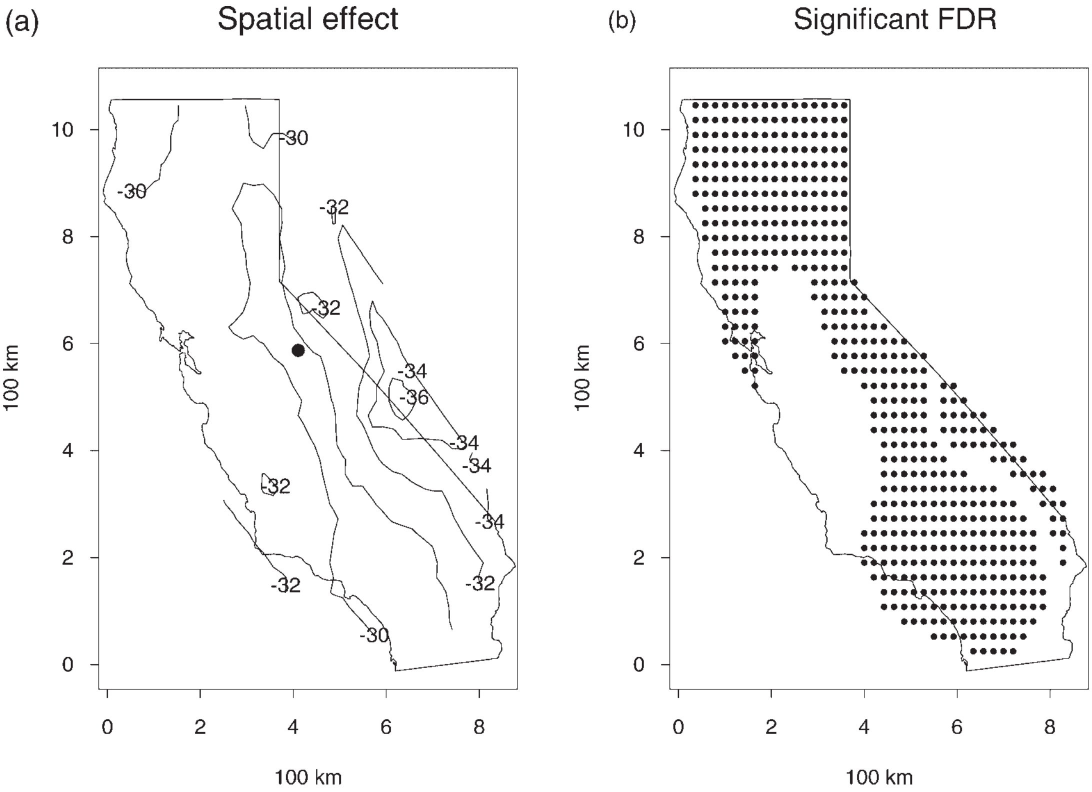

Figure 3. (a) Provides the estimated spatial effect in contour form, (b) shows the region found significant by the FDR analysis. The dot in (a) shows the location of the Yosemite National Forest

Projections of future counts 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

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

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

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

As an example, predictions are presented for the Yosemite National Park in the Federal Lands of California. Figure 4 provides estimates of the expected count of fires for Yosemite National Park each month derived via expression (4). (Yosemite is indicated by the dot in the left-hand panel of Figure 3(a).) One sees the expected count peaking, just below a level of 34 fires, in July. The distribution of a future count can be approximated by a Poisson with the indicated estimated expected count.

*4.2. Results of fitting Model III*

Next consideration turns to Model III, i.e., including a random effect for year. The fitting is carried out via penalized quasi-likelihood implemented as the function glmmPQL(), (Venables and Ripley, 2002).

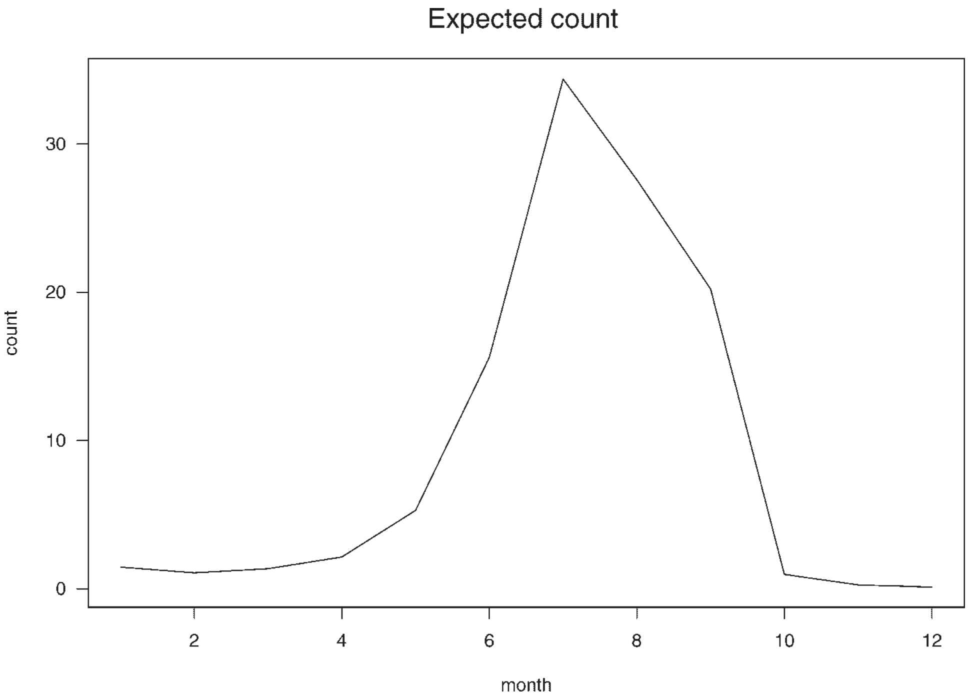

Figure 4. Model I predictions of expected counts by month for Yosemite National Park based on expression (4)

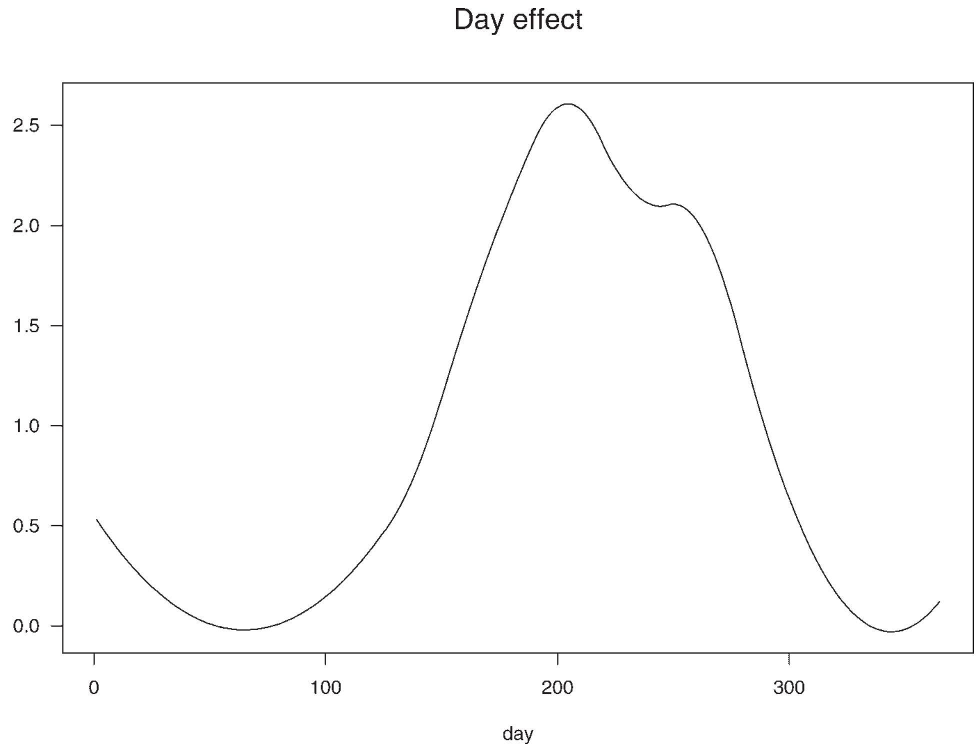

Figure 5. The estimated daily effects for Model III

Doing so, one obtains for the standard error of the year effect $\hat{\tau} = 0.157$. The program output also includes $X/\sqrt{n} = 1.062$, 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 5 shows the estimated daily effects. There are not great changes from the Model I results presented in Figure 2(a).

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

$$\sum_i \int \frac{\exp\{\hat{\eta}_i + \hat{\tau} z\}}{(1 + \exp\{\hat{\eta}_i + \hat{\tau} z\})} \phi(z) \mathrm{d}z \tag{5}$$

with $i$ again labeling 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 expression (5) is evaluated numerically and the results are given in Figure 6a. Uncertainties may be derived via the jackknife splitting on year.

In some circumstances an estimate of

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

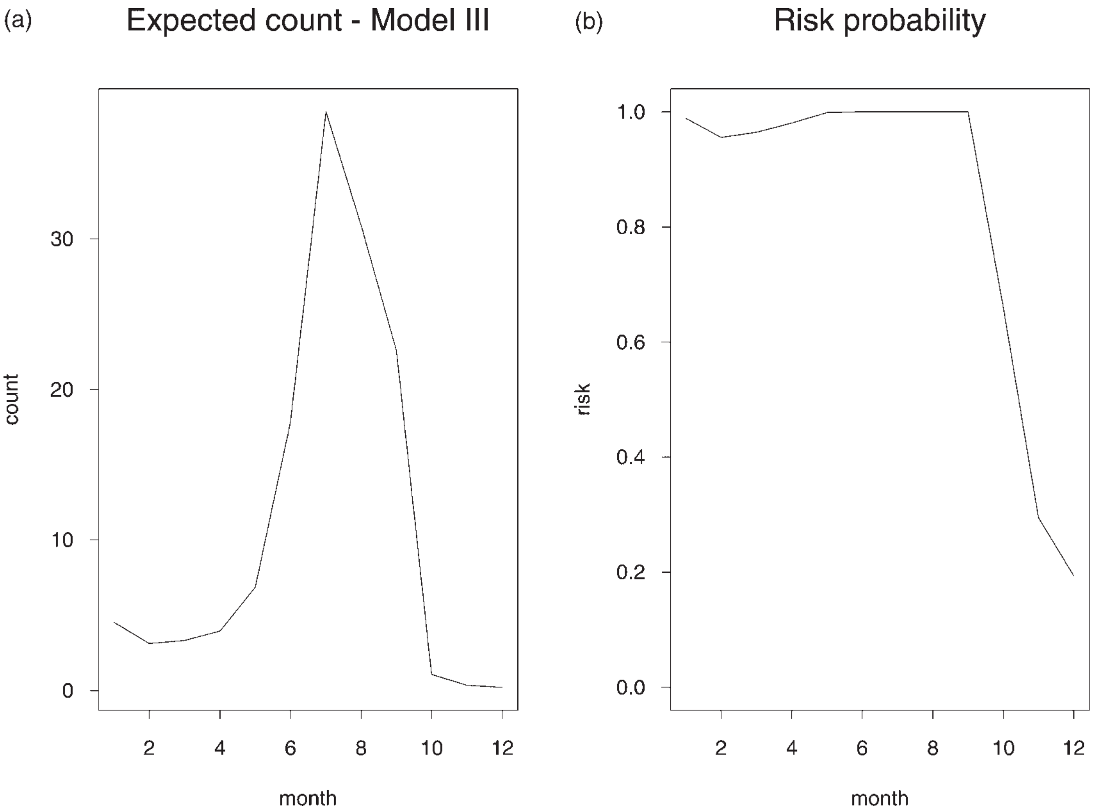

Figure 6. Model III predictions of expected monthly totals and corresponding risks for Yosemite National Forest based on expression (4). The risk computation assumes a Poisson distribution of the count

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

$$1 - \exp\left\{ -\int_M \mu(x', y', d')\mathrm{d}x'\mathrm{d}y'\mathrm{d}d' \right\}$$

whose integrand may be approximated by expression (4). The results are given in Figure 6(b). Figure 6(a) may be compared with Figure 4. The two are very similar. It may be noted that $\hat{\tau}$ is small, 0.157.

Figure 7 gives estimates of the random year effects, that is $E\{I|\text{data}\}$ for each of the of the years 2000–2003. Such effects are sometimes referred to as shrunken effects.

### 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

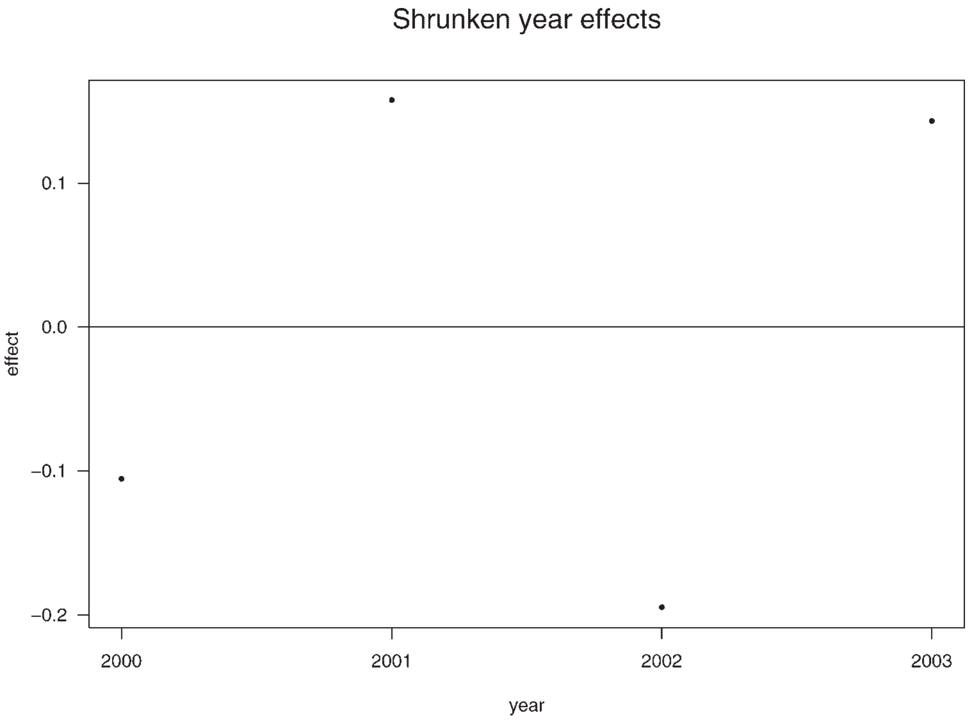

Figure 7. Shrunken year effects for Model III

such nonstandard cases, (Brillinger and Preisler, 1983; Brillinger, 1996). In the binary case, these may be computed as follows: suppose Prob$\{Y = 1|\text{explanatories}\} = \gamma$ and that $U_1$ and $U_2$ denote independent uniforms on the intervals $(0, 1 - \gamma),\ (1 - \gamma, 1)$, respectively, then the variate

$$U = U_1 \times (1 - Y) + U_2 \times Y$$

has a uniform distribution on the interval $(0, 1)$. Whereas when Prob$\{Y = 1|\text{explanatories}\} = \gamma_0$, then $E\{U\} = (1 + \gamma - \gamma_0)/2$ for example. We refer to $U$, when an estimate of $\hat{\gamma}$ is used, 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, for example normal probability plots involving the normal residuals $\Phi^{-1}(\hat{U})$ or plots of $\Phi^{-1}(\hat{U})$ versus explanatories.

The right-hand panel of Figure 8 shows notched boxplots of the normal residuals of Model I against the year. The graph suggests that year need not be in the model as an explanatory. The left-hand panel provides a normal probability plot and there appears not much to be concerned with.

Figure 9 provides similar plots for Model III, but now the explanatory in the right-hand panel is elevation. Again there is little indication that this explanatory need be included. The left-hand shows little evidence against the distributional assumptions made.

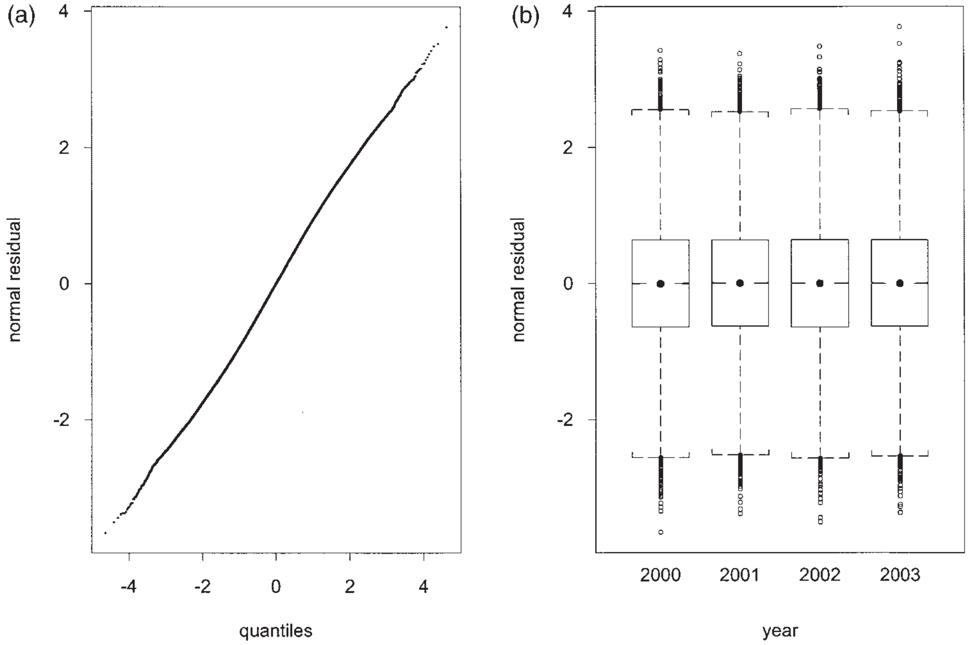

Figure 8. Model assessment results for Model I; (a) is a normal probability plot of the normal residuals, (b) shows notched boxplots for each year's normal residuals. The notches are very small here

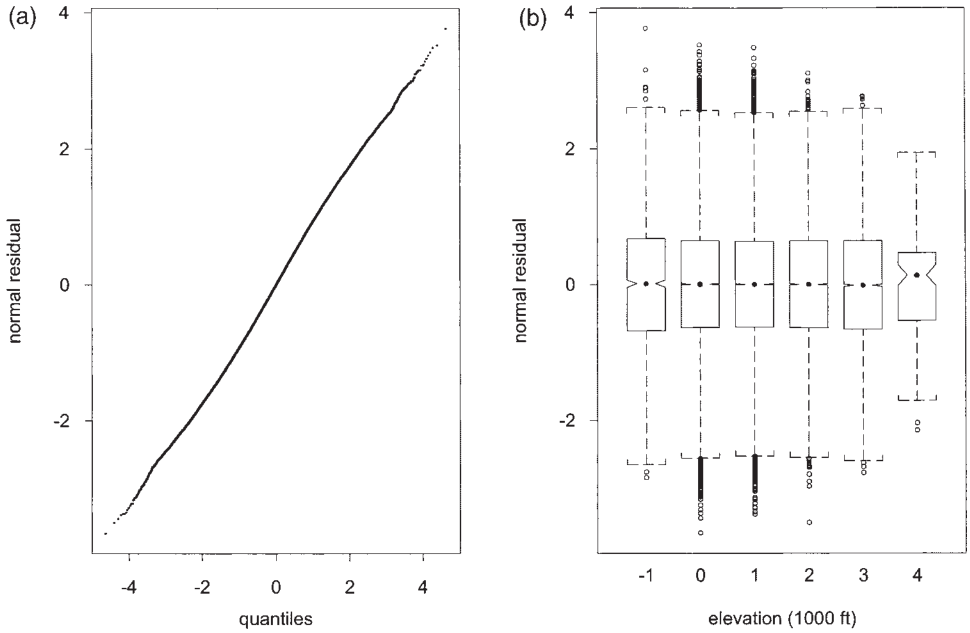

Figure 9. Model assessment results for Model III. The negative elevations in (b) come from Death Valley

## 5. SUMMARY AND DISCUSSION

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. An 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 modeling was to that based on random effects modeling.

Turning to the promised comparisons with the Oregon results , the same methodology appears to work equally well with both the California and Oregon data.

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

Further details of the approach may be found in Brillinger *et al*. (2003) and Preisler *et al*. (2004). This work parallels that of Brillinger *et al*. (2004) and Preisler and Benoit (2004) closely.

## REFERENCES

Benjamini Y, Yekutieli D. 2001. The control of false discovery rate in multiple testing under dependency. *Annals of Statistics* **29**: 1165–1188.

Breslow NE, Clayton DG. 1993. Approximate inference in generalized linear models. *Journal of American Statistical Association* **88**: 9–25.

Brillinger DR, Preisler HK, Benoit JW. 2003. Risk assessment: a forest fire example. *Statistics and Science: A Festschrift for Terry Speed*. Vol. 40 IMS Lecture Notes; 177–196.

Brillinger DR, Preisler HK, Naderi HM. 2004. Wildfire chances and probabilistic risk assessment. In *Proceedings of the Sixth International Symposium on Spatial Accuracy Assessment in Natural Resource and TIES*, Portland Maine, USA.

Brillinger DR, Preisler HK. 1983. Maximum likelihood estimation in a latent variable problem. *Studies in Econometrics, Time Series and Multivariate Statistics*. Academic Press: New York; 31–65.

Brillinger DR. 1996. An analysis of an ordinal-valued time series. In *Lecture Notes in Statistics*, Vol. 115. Springer: New York.

Green PF. 1987. Penalized likelihood for general semi-parametric regression models. *International Statistical Review* **55**: 245–259.

Maddala GS. 1992. *Introduction to Econometrics* (2nd edn). MacMillan: New York.

Pinheiro JC, Bates DM. 2000. *Mixed-effects Models in S-PLUS*. Springer: New York.

Preisler HK, Benoit JW. 2004. A state space model for predicting wildland fire risk. In *Proceedings of the Joint Statistics Meetings*, Toronto.

Preisler HK, Brillinger DR, Burgan RE, Benoit JW. 2004. Probability based models for estimating wildfire risk. *International Journal of Wildland Fire* **13**: 133–142.

Schall R. 1991. Estimation in generalized linear models with random effects. *Biometrika* **78**: 719–727.

Venables WN, Ripley BD. 2002. *Modern Applied Statistics with S* (4th edn). Springer: New York.