UCLA

Department of Statistics Papers

Title

A Note on Non-parametric and Semi-parametric Modeling of Wildfire Hazard in Los Angeles County, California

Permalink

<https://escholarship.org/uc/item/42r598rb>

Authors

Schoenberg, Frederic P
Jamie Pompa
Chang, Chien-Hsun

Publication Date

2007-01-29

A Note on Non-parametric and Semi-parametric Modeling of Wildfire Hazard in Los Angeles County, California.

Frederic Paik Schoenberg^{1,2}, Jamie Pompa¹, Chien-Hsun Chang¹.

Abstract: This paper explores the use of, and problems that arise in, kernel smoothing and parametric estimation of the relationships between wildfire incidence and various meteorological variables. Such relationships may be treated as components in separable point process models for wildfire activity. The resulting models can be used for comparative purposes in order to assess the predictive performance of the Burning Index.

Keywords: Estimation, kernel smoothing, model evaluation, point process, wildfires.

1 Introduction.

The United States National Fire Danger Rating System, the current version of which was implemented in 1978, outputs several daily indices including the Burning Index (BI) at each of various Remote Automated Weather Stations (RAWS) in the United States (Bradshaw et al. 1983; Pyne et al. 1996). The BI is widely used in the United States by federal, state, and local agencies, including Los Angeles County, as a predictor of overall daily wildfire activity

¹ Department of Statistics, University of California, Los Angeles, 90095-1554.

² Corresponding author : tel. 310-794-5193, fax 310-206-5658, email: frederic@stat.ucla.edu.

and as a tool for the allocation and preparation of fire suppression equipment and personnel (Andrews and Bradshaw 1997) and has been shown to be correlated with wildfire incidence and daily burn area for several different regions (Haines et al. 1983; Mees and Chase 1991).

However, there has been little research into the construction of reasonable alternative models to which indices such as the BI can be compared. One possibility is to estimate a separable point process model, using weather variables as covariates. The advantage of assuming separability is that, when this assumption is satisfied, each component in the model may be estimated individually (Rathbun 1996, Schoenberg 2004). In such cases, one may use a non-parametric method such as kernel smoothing in order to suggest a parametric form for each component in the model. Here, we explore issues that arise in the use of kernel smoothing and semi-parametric approaches in estimating separable point process models for wildfire incidence in a particular region.

The focus in the present paper is on Los Angeles County, where wildfires perennially pose a particularly serious threat, especially each Fall when temperatures are warm, Santa Ana winds blow hot, dry air from the North-East, and the chaparral, Los Angeles County's dominant vegetation type, is especially dry and prone to ignition (Keeley 2000, Keeley 2002). In addition, Los Angeles County Fire Department and Department of Public Works officials have collected and compiled a wealth of very detailed data on wildfire activity dating back over a century, making the region unusually well-suited to statistical analysis.

Note that the present work does not concern the problem of forecasting the spread of existing wildfires, which is a major concern for wildfire management and suppression. Numerous studies have investigated the empirical predictive performance of models for the

spread of wildfires (e.g. Rothermel 1991, Turner and Romme 1994). Instead, our attention is on the forecasting of wildfire incidence, i.e. the probability of a wildfire occurring in a given location on a certain day, given weather conditions and other information typically available in advance to fire managers and other planners.

Section 2 of this paper describes the Burning Index and the Los Angeles County dataset used for analysis. Following a discussion in Section 3 of kernel smoothing applied to an assortment of variables related to wildfire occurrence, a comparison is given in Section 4 between a model that uses the BI and a simple semi-parametric, separable point process model that uses weather variables which are inputs in the computation of the BI. This is followed by a discussion in Section 5.

2 Data.

There is a network of over 1200 fire weather stations in the United States. We focus here on data from the sixteen RAWS stations located within Los Angeles County, California. Each of the RAWS records information on a host of meteorological variables, including air temperature, relative humidity, precipitation, wind speed, and wind direction (Warren and Vance 1981). Daily summaries of these measurements are collected each afternoon at 1pm and transmitted via satellite to a central station for archiving. For some of the variables, such as relative humidity and temperature, not only is the current 1pm value recorded, but also its maximum and minimum values over the previous 24 hours.

The current version of the National Fire Danger Rating System (NFDRS) was implemented in 1978 (Deeming et al. 1977; Bradshaw et al. 1983) with additional options added

in 1988 (Burgan 1988). Fire danger indices are calculated from daily RAWS weather data and values assigned to each site including fuel model, slope class, and herbaceous fuel type. Fuel moisture values are calculated, which are then used to construct a Spread Component (SC) and an Energy Release Component (ERC); these two components are combined to produce the Burning Index, based on a suite of physics-based nonlinear dynamic equations involving heat transfer and moisture exchange, and whose original basis was to predict flame length in an existing fire (see Pyne et al. 1996, Andrews and Bradshaw 1997).

Daily BI values were computed from the RAWS station measurements using FireFamily Plus software, which is freely available from the Forest Service (Bradshaw and Brittain 1999; Bradshaw and McCormick 2000; Andrews et al. 2003). The *Southern California Chaparral* fuel model was used in the computation of BI for each of the weather stations. For several of the RAWS, data were missing on certain days in the time range considered here (1976-2000); proportions and potential impacts of this missing data for certain stations and months are discussed in Peng et al. (2005). In what follows, for any particular day we consider the BI values and RAWS measurements averaged across the number of RAWS which were available on that day, as in Mees and Chase (1991) and Peng et al. (2005).

Detailed data on Los Angeles County wildfires have been collected and compiled by several agencies, including the Los Angeles County Fire Department (LACFD), the Los Angeles County Department of Public Works (LACDPW), the Santa Monica Mountains National Recreation Area, and the California Department of Forestry and Fire Protection. These data include the origin dates and spatial burn patterns for wildfires dating back to 1878. The LACFD has been mapping fires as small as 0.00405 km² (1 acre) since 1950, and

the modern catalog includes records of wildfires whose area is as small as 858.1 m² (0.212 acres). However, there may be many small fires missing from the dataset. According to LACFD officials, the wildfire catalog before 1950 is believed to be complete for fires burning at least 0.405 km² (100 acres), and some evidence suggests that the 1976-2000 data may be complete only for fires burning at least 0.0405 km² (10 acres) (Schoenberg et al. 2003; Peng et al. 2005). We restrict our attention here to the 592 wildfires of at least 0.0405 km² (10 acres) recorded between January 1, 1976 and December 31, 2000. This represents 75.5% of the wildfire records for 1976-2000 which were made available to us, and these 592 fires account for 99.78% of the area burned in Los Angeles County wildfires recorded during this period. For further detail and images of the centroid locations of these wildfires, see Peng et al. (2005).

3 Kernel smoothing.

One way to develop an alternative to the BI is by estimating a point process model whose conditional rate $\lambda(t, \mathbf{z}, A)$ of a wildfire on day t of area A centered at location $\mathbf{z}$ depends on the meteorological variables measured by the RAWS stations. For simplicity, as an initial attempt one might assume the model is purely multiplicative, or in the terminology of Cressie (1993) is *separable*; in such cases each component of the model may be estimated individually (Rathbun 1996, Schoenberg 2004). For a thorough treatment of point processes and conditional intensities, see Daley and Vere-Jones (1988, 2003).

One way to approach the estimation of each individual component $f(x)$ in the conditional intensity of a point process N is by a non-parametric method such as kernel smoothing (see

e.g. Silverman 1986). Such a method may be appropriate when the relationship being estimated, such as that between temperature and the probability of fire occurrence, may reasonably be expected to be smooth. Using kernel smoothing, the estimate of such a component $f(x)$ is given by $\hat{f}(x) = \sum_i N_i K(x - x_i; h) / \sum_i K(x - x_i; h)$, where N_i represents the number of wildfires occurring on day i , and x_i is the temperature on day i . The function K is called the *kernel density* and typically obeys the constraint $\int K(y; h) dy = 1$. The parameter h , which governs the degree of smoothing, is called the *bandwidth*.

Generally the kernel density $K(y)$ is chosen to be symmetric around zero and monotonically decreasing in $|y|$, so that the observed values closer to x contribute more to the estimate of $f(x)$. When the explanatory variable x is confined to a finite interval $[a, b]$, some boundary correction method is typically necessary in order to ensure that $\int_a^b K(x - x_i; h) dx = 1$. One option proposed in Silverman (1986) and Venables and Ripley (2002) is to create an expanded dataset by reflecting each observed point across each of the interval's endpoints a and b and then kernel smoothing the expanded dataset. This reflection method is applied in all results in Section 3.1 below.

In kernel smoothing one must determine the kernel density K as well as the bandwidth h . While both affect the resulting estimates, the choice of a bandwidth is known to be far more critical than the choice of the kernel density (Silverman 1986). In general, larger bandwidths result in smoother estimates with lower variance but can introduce substantial bias.

3.1 Univariate kernel smoothing.

Various algorithms exist for automatically selecting a bandwidth for kernel smoothing. A commonly used option is Silverman's "rule of thumb": $h = 0.9 \min\{s, IQR\}n^{-1/5}/1.34$, where s is the sample standard deviation, IQR is the inter-quartile range, and n is the number of observations of the variable being smoothed (Silverman 1986). Unfortunately, this bandwidth often seems to be too small, and the resulting plot appears to be inadequately smoothed. This is particularly true in the case of estimating the relationship between fire incidence and either wind speed or maximum daily relative humidity.

For instance, Figure 1 shows a scatter plot of wind speed versus fire incidence along with a kernel-smoothed estimate calculated using a Gaussian density. Silverman's bandwidth is clearly too small for wind speeds greater than 20 miles per hour (mph), especially for higher wind speeds where the data is more sparse and the proportion of days with at least one fire is higher. The root cause of the small bandwidth is a large concentration of observed values of wind speed within a small range of values, which resulted in a very small inter-quartile range relative to the actual range of this variable. This is highlighted in Figure 2 which shows a bar plot of the distribution of the wind speed values separated into deciles.

Note that Silverman's formula does seem to provide reasonable bandwidths for cases where the observed values of the explanatory variable are mostly clustered within a small range of values, provided that the values of the response variable are all identically zero whenever the explanatory variable is outside this range. Such is the case when kernel smoothing the relationship between daily precipitation and number of fires, for instance, as shown in Figure 3. Although similar to wind speed in that most of the observations are within a

small range of values, those precipitation amounts where the data becomes sparser are days where there are no fires at all, producing a smooth kernel estimate. Most of the variation in the daily number of fires observed occurs for smaller amounts of precipitation, where there are far more observations, and hence the small bandwidth is quite appropriate for this case. Note that the commonly used variant of Silverman’s formula proposed by Scott (1992), which replaces the multiplying factor 0.9 with 1.06, yields very similar results.

An alternative method for selecting bandwidths is the likelihood cross validation (LCV) approach (Stone 1974; Stone 1984; Silverman 1986). Similar to cross validation methods used in regression, an initial bandwidth, h , is chosen and then, for each point x_i in the dataset, x_i is temporarily removed from the dataset and the estimate of the kernel smoothed function at x_i is calculated, given the bandwidth h . This value is stored, say as $\hat{f}(x_i; h)$, and then the distance $d(x_i; h) = |\hat{f}(x_i; h) - f(x)|$ from the observed quantity $f(x_i)$ is examined. The optimal bandwidth is determined by finding the bandwidth that minimizes $\sum_i \log(d(x_i; h))$.

It is well-known that when bandwidths are selected by LCV, the resulting kernel estimates can be inconsistent and highly sensitive to outliers under a variety of conditions (Scott and Factor 1981; Schuster and Gregory 1981). In estimating the relationship between a particular weather variable and fire incidence in Los Angeles County, the LCV bandwidths are even smaller for several of the weather variables than the bandwidths calculated using Silverman’s algorithm. Here, the problem seems to be due to the presence of multiple repetitions of exactly the same recorded value of the weather variable. For maximum relative humidity, for instance, 87.3% of the observations are duplicate values; in fact 18.6% of the days have a recorded maximum relative humidity of exactly 100%. Estimates based on LCV tend to

seriously overfit in such situations and the resulting bandwidths are far too small.

In light of the shortcomings of LCV, one might consider a modified likelihood cross validation method, where instead of only removing x_i in the prediction of the density at x_i , all observations with the same value as x_i are removed when predicting x_i . Again, the optimal bandwidth is determined by minimizing $\sum_i \log(d(x_i; h))$. This modified LCV approach essentially involves removing one small segment of the x-axis at a time, rather than one observation.

Several alternative data-driven methods for bandwidth selection are compared in Park and Marron (1990), including biased cross-validation (BCV) and plug-in (PI) estimates. For clustered data, such methods often result in larger bandwidths and thus smoother estimates compared to cross-validation (Park and Marron 1990; Park and Turlach 1992). However, these methods may still not be ideal for datasets with multiple repetitions of identical values. Indeed, the justification for these methods is based on the assumption that the density being estimated is smooth, which is essentially inconsistent with multiple observations that overlap exactly.

In the case of the Los Angeles County weather and wildfire data described in Section 2, the kernel estimates using the modified LCV approach appeared to be more appropriate for those weather variables whose bandwidths were too small when calculated via ordinary LCV or Silverman’s formula. For instance, Figure 4 shows the kernel estimates of maximum daily relative humidity for the likelihood cross validation method and the modified likelihood cross validation method. It is clear that the modified likelihood cross-validation approach produces more reasonable estimates in this case.

It is often useful to construct standard errors and confidence bounds for kernel-smoothed estimates. The 95% confidence bounds for the kernel smoothed estimate of $f(x_j)$ can be calculated via the formula $1.96\sqrt{\sigma^2(x_j)\|K\|_2^2/(N_jnh)}$, where $\sigma^2(x)$ is defined as $\sum_{i=1}^n W_{hi}(x)(N_i - \hat{m}_h(x))^2/n$, $W_{hi}(x) = [K((x - x_i)/h)]/[h\hat{f}_h(x)]$, $\hat{f}_h(x) = \sum_{i=1}^n K_h(x - x_i)/n$, $\hat{m}_h(x)$ is calculated as $\left[\sum_{i=1}^n K_h(x - x_i)N_i/n\right] / \left[\sum_{i=1}^n K_h(x - x_i)/n\right]$, $\|K\|_2^2 = \int_{-\infty}^{\infty} [K_h(x)]^2 dx$, n is the number of days, N_i is the number of wildfires observed on day i , x_i is the value of the variable x on day i , and h is the bandwidth, with $K_h(x) = K(x/h)/h$ (Härdle 1991). Note that when the selected bandwidth is too small, the corresponding standard error may reflect the high variance associated with the resulting estimate. This is highlighted in Figure 5 where the instability of the estimate due to the small bandwidth calculated from Silverman's formula is readily seen.

3.2 Bivariate kernel smoothing.

Since not all locations in Los Angeles County are equally likely to burn, one may wish to estimate a spatially-varying background rate of wildfires, $\mu(\mathbf{z})$. As with relationships with meteorological variables, such a function $\mu(\mathbf{z})$ can be estimated by kernel smoothing the observed wildfires. However, there is an especially serious risk of over-fitting if all fires in the dataset are used for this purpose. Hence some authors estimate such a background rate using only the largest events (Ogata 1998) or using a historical catalog of prior events (Peng et al. 2005). For instance, in Los Angeles County, since wildfire histories are available dating back to 1878, one option would be to kernel smooth the spatial locations of fires occurring before 1976 and use data from 1976 on to estimate the bandwidth. Since there are thought to be

severe missing data problems before 1950, another option is to smooth only the centroids of fires occurring between 1951 and 1975. That is, one may estimate a spatial background rate by smoothing the locations of the wildfires from 1950-1975 using kernel smoothing, with a bandwidth chosen so that the smoothing of the 1950-1975 wildfires most closely fits the data from 1976-2000.

However, one problem with simply kernel smoothing the centroids of historical fires is that the resulting estimate of μ will be non-zero for locations $\mathbf{z}$ outside of the boundary of Los Angeles County, including locations in the Pacific Ocean. One way to address this is to allow the kernel and/or bandwidth to vary spatially. For instance, the edge correction method suggested in Diggle (1985) involves dividing each term $K(\mathbf{z} - \mathbf{z}_i; h)$ by the integral $\int K(\mathbf{z} - \mathbf{z}'; h) d\mathbf{z}'$, where the integral is over all locations $\mathbf{z}'$ in Los Angeles County. Unfortunately, the computation of this integral of the kernel density over an irregularly shaped area such as Los Angeles County is non-trivial, and such an integral must be computed separately for each location $\mathbf{z}$ at which the kernel density is to be estimated. A simple alternative which gives each point in the historical dataset equal weight is to truncate the kernel used to smooth a historical fire whose centroid is at location $\mathbf{z}_j$. That is, $\hat{\mu}(\mathbf{z}) = \sum_i K_{(j)}(\|\mathbf{z}_j - \mathbf{z}\|; h)$ where $K_{(j)}(u) = 0$ for $u > r_j$, with r_j the shortest distance from $\mathbf{z}_j$ to the boundary of Los Angeles County, and rescaling K_j accordingly so that $\int K_{(j)}(\|\mathbf{z}_j - \mathbf{z}\|; h) d\mathbf{z} = 1$. Note that this spatial integral is over a circle of radius r_j and hence is far simpler to compute than the integral over all of Los Angeles County, and that this integral need only be computed for each point in the historical dataset, rather than for all locations at which the kernel density estimate is sought.

A complication arises from truncating the kernel in this way, as the kernel estimate of $\mu(\mathbf{z})$ will be exactly zero for those locations $\mathbf{z}$ that are closer to the boundary of Los Angeles County than to any other fire occurring between 1951 and 1975. The 23 fire centroids in the post-1975 data set for which the kernel estimate of μ would be zero are highlighted in Figure 6. It is easy to see that these fires are closer to the county boundary than to any of the pre-1976 fire centroids. One simple way to address this problem is to estimate $\mu(\mathbf{z})$ instead via $\hat{\mu}(\mathbf{z}) = c + \sum_j K_{(j)}(\|\mathbf{z}_j - \mathbf{z}\|; h)$, where c is a constant representing the minimum background rate.

One method of bandwidth selection for the spatially varying background fire rate is maximum likelihood estimation (Peng et al 2005; Ogata 1998). One may readily use a standard Newton-Raphson minimization routine to find a minimum of the negative log-likelihood, $\int \lambda(\mathbf{z}_i) d\mathbf{z} - \sum \log \lambda(\mathbf{z}_i)$, where the summation is over all points i between 1976 and 2000. For instance, if the exponential kernel $K(d; h) = h^2 \exp(-hd)/2\pi$ is used and truncated for each point at the boundary of Los Angeles County as described above, then $\int \lambda(\mathbf{z}_i) d\mathbf{z}$ may be calculated as $cL + \sum_i 2\pi \{1 - [\exp(-hr_j)(1 + hr_j)]\}$ and $\sum_i \log \lambda(\mathbf{z}_i) = \sum_i \log \left[\sum_j \left\{ c + (h^2/2\pi) \exp(-hd_{(i,j)}) \mathbf{1}_{\{\mathbf{d}_{(i,j)} < r_j\}} \right\} \right]$, where c is a constant representing the minimum background rate of wildfires, h is the bandwidth, L is the area of Los Angeles County, r_i is the minimum distance from the fire at location $\mathbf{z}_i$ to the county boundary line, and $d_{(i,j)}$ is the distance from the location $\mathbf{z}_i$ of fire i to the location $\mathbf{z}_j$ of fire j . As before, the index i is over all wildfires occurring between 1976 and 2000, and the index j is over all fires occurring between 1951 and 1975. As a starting value in the optimization routine, a reasonable initial choice for the constant c is the total number of fires occurring between

1976-2000 divided by the area of Los Angeles County.

The bandwidth estimated by maximum likelihood can be examined through a spatial plot of the kernel estimate, truncated for those points at the boundary of Los Angeles County as described above, with the data points overlaid. Figure 7 shows such a plot for the centroids of Los Angeles County wildfires with an exponential kernel; the bandwidth found via maximum likelihood estimation appears to be rather reasonable. The exponential kernel was chosen based on the computational convenience of computing the integral term in the negative log-likelihood equation. Areas in Figure 7 with a low background rate (lightly shaded pixels) correspond to areas in Los Angeles County with few or no historical fires, while darker pixels represent a higher historical fire rate for that location.

4 A semi-parametric alternative to the BI.

While the BI is correlated with wildfire incidence in Southern California (Mees and Chase 1991), one may wish to compare the strength of this relationship by examining the fit of simple competing models that, like the BI, rely primarily on the information recorded at the RAWS. For instance, one alternative model considered here is directly based on seven elementary variables: the maximum relative humidity over the previous 24 hours, 1pm wind speed, 1pm temperature, 1pm precipitation, the spatial pattern of historical fire centroids, the date within the year, and the cumulative precipitation over the previous 60 days. The first four of these variables are used directly in the computation of the BI. Cumulative precipitation variable is included as a simple proxy for the drought severity index which is also an element of the NFDRS (Pyne et al. 1996).

Since non-parametric estimates are especially prone to over-fitting, one may wish to consider parametric alternatives as reasonable choices for purposes of comparison and evaluation of the BI. As mentioned in Section 3, the assumption of separability simplifies the construction and estimation of point process models, as the functional form for each covariate can be investigated individually, and each of the model's components can be estimated by maximum likelihood, the results simply multiply together to form a complete model.

In general, the relationship between wildfire incidence and a single weather variable would not be well estimated by looking exclusively at those two variables; the presence of confounding factors could substantially bias the estimates. However, there are situations where such relationships can be estimated separately. For instance, if the model is purely multiplicative, then under some restrictions the estimates one obtains by estimating each of the components in the model by maximum likelihood may be consistent; some of these cases are investigated in Schoenberg (2006). While the results of applying the tests for separability proposed in Schoenberg (2004) to various components of the Los Angeles County wildfire and meteorological data are rather mixed and are the subject of ongoing investigation by the present authors, only a few of the relationships show a very strong violation of separability. While a separable point process model is clearly far from ideal, as an initial approximation such a model is not grossly inappropriate for this data set and may provide a suitable null model to which the BI may be compared.

The kernel smoothed estimates of the relationships between each of the meteorological variables and the number of fires per day suggests that each is approximately exponential, while the relationship with date within the year seems to be approximately parabolic. These

relationships suggest a point process model of the form:

$$\lambda(t, \mathbf{z}, A) = f(A)\mu(\mathbf{z})\beta_1 e^{(\beta_2 W(t) + \beta_3 R(t) + \beta_4 P(t) + \beta_5 L(t) + \beta_6 T(t) - \beta_7(\beta_8 - D(t))^2)}, \quad (1)$$

where $\lambda(t, \mathbf{z}, A)$ is the fire rate on day t at location $\mathbf{z}$, $f(A)$ follows the tapered Pareto distribution (Schoenberg et al. 2003), $\mu(\mathbf{z})$ is the spatial background fire rate at location $\mathbf{z}$, $W(t)$, $R(t)$, $P(t)$, $L(t)$, $T(t)$, and $D(t)$ represent the wind speed, maximum relative humidity over the previous 24 hours, precipitation, lagged precipitation over the previous 60 days, temperature for day t , and the day of the year for day t , respectively. The model is semi-parametric since the spatial background rate $\mu(\mathbf{z})$ is estimated non-parametrically; however, since this estimation is done by kernel-smoothing the pre-1976 fires as explained in Section 3, rather than kernel-smoothing the post-1976 fires themselves, the resulting estimates should not have the usual problems of overfitting generally associated with non-parametric estimates. The maximum likelihood estimates for the coefficients in model (1), using a Nelder-Mead optimization routine, are $\{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8\} = \{0.00149(0.0000220), 0.0749(0.00213), -0.0233(0.00293), -7.53(0.178), -0.416(0.0157), 0.0659(0.000138), 0.000138(2.61 \times 10^{-23}), 211(2.62)\}$. Estimates of the standard errors, given above in parentheses, are obtained using the square root of the reciprocal of the diagonal of the estimated Hessian of the loglikelihood at its maximum (see e.g. Wilks 1962, ch. 13.8).

A point process model for the BI to be used for comparison with model (1) is

$$\lambda(t, \mathbf{z}, A) = f(A)\mu(\mathbf{z})\beta_1 e^{(\beta_2 B(t))}, \quad (2)$$

where $\lambda(t, \mathbf{z}, A)$ represents the BI on day t at the point $\mathbf{z}$ in Los Angeles County, $f(A)$ again follows the tapered Pareto distribution, $\mu(\mathbf{z})$ is the historical fire rate at location $\mathbf{z}$, and $B(t)$ is the BI value for day t . Maximum likelihood estimates of the parameters $\{\beta_1, \beta_2\}$ in

model (2) are $\{0.0317(0.00130), 0.00851(0.000325)\}$. As with (1), model (2) was suggested by kernel regression on the relationship between number of fires and BI, and was fit by maximum likelihood using a Nelder-Mead optimization routine. A model linear in BI was also estimated, but provided substantially worse fit than the exponential model in (2).

One way to compare models (1) and (2) is with the Akaike Information Criterion (AIC) $2L(\hat{\theta}) + 2p$, where p is the number of fitted parameters and $L(\hat{\theta}) = \sum_i \log(\lambda(t_i, \mathbf{z}_i, A_i; \hat{\theta})) - \int \lambda(t, \mathbf{z}, A; \hat{\theta}) dt d\mathbf{z} dA$ is the log-likelihood of the data at the MLE, $\hat{\theta}$ (Akaike 1977). Under the standard large sample theory used to justify maximum likelihood estimation, the AIC is an estimate of the expected negative entropy and thus measures the distance between the true and estimated probability laws of the data (Ogata and Akaike 1982). Lower AIC indicates better fit, and as a rule of thumb differences in AIC greater than 2 are generally considered statistically significant; while the original justification of the AIC is based on the use of nested models where the difference in AIC is approximately χ^2 under the null model, the AIC has been shown to be useful for comparing non-nested models as well (Ogata 1988).

For the models (1) and (2), the AICs are 2162.0 and 2457.5, respectively, demonstrating that model (1) offers far superior fit. Figures 8 and 9 show the differences $\log \lambda_1(t_i, \mathbf{z}_i, A_i) - \log \lambda_2(t_i, \mathbf{z}_i, A_i)$, where λ_1 and λ_2 represent models (1) and (2), respectively. From Figure 8 one sees that the model (1) generally has a substantially higher rate than model (2) on days in the late summer and early autumn when fires occur, particularly on days when the maximum relative humidity over the previous 24 hours is moderate or low. This is important since the majority of large fires occur on such days. Figure 9 shows that λ_1 tends to be considerably higher than λ_2 for moderate wind speeds and high temperatures, which

are conditions under which the bulk of fires occur. The results suggest that the BI may be significantly underpredicting fire hazard on relatively warm, windy, dry autumn days and sometimes overpredicting fire hazard on cooler and wetter days in the winter and spring.

To verify that the differences in fit between models (1) and (2) are not due to overfitting, the data may be divided into two halves, with the data from 1976-1988 used to estimate each model, and the data from 1989-2000 reserved for model assessment. The results are quite similar to those above. The AIC (using only data from 1989-2000 to compute the log-likelihood) for models (1) and (2) are 1569.9 and 1697.5, respectively, for a difference of 127.5. This comparison is very similar to that using the entire dataset for both estimation and assessment: after dividing the dataset, the AIC difference per day is just 13.7% smaller than that obtained using the entire dataset. The maximum likelihood parameter estimates and estimated standard errors for model (1) were $\{0.00237(0.00009.19), 0.0690(0.00317), -0.0249(0.000837), -5.40(1.55), -0.421(0.0361), 0.0640(0.000624), 0.000133(8.16 \times 10^{-23}), 200(3.19)\}$. and those for model (2) were $\{0.0428(0.00218), 0.00785(0.000390)\}$. Results similar to those in Figures 8 and 9 are also obtained when the data are divided in this way. For instance, Figure 10 shows the differences $\log \lambda_1(t_i, \mathbf{z}_i, A_i) - \log \lambda_2(t_i, \mathbf{z}_i, A_i)$, as in Figure 9, where now the data used in the fitting are from 1976-1988 and the points plotted are from 1989-2000. The scale is narrower, but the results are essentially similar: $\lambda_1 - \lambda_2$ tends to be positive, especially on days when temperatures are high and windspeeds are at least 10 mph.

5 Discussion.

The improvement in the prediction of wildfire occurrence by model (1) over model (2) highlights some deficiencies of the BI. Model (1) is a simple model which is not suggested as a replacement for the BI, but is merely presented as a comparative model for the purpose of assessing the adequacy of the BI in predicting wildfire occurrence. Since model (1) uses covariates which are directly used as inputs into the BI, and since the BI also incorporates a host of complex relations based on physical experiments as well as local information (such as fuel models, fuel moisture readings, etc.) not used in model (1), one might expect model (2) to vastly outperform model (1) in terms of predicting wildfire incidence. Instead, the results suggest that the BI may not be an ideal predictor of wildfire activity, as the BI appears to be generally too low on days when wildfires occur in late summer and early autumn, especially when such days are relatively warm, dry, and windy.

It is clear that despite the inclusion of a number of factors that relate to a wildfire incidence, there are still many ways in which model (1) may be improved. Model (1) is completely separable, and it appears likely that more complicated models involving the interaction of various components in model (1) may offer substantially improved fit. Furthermore, there are many other factors not included in model (1) that contribute to wildfire occurrence, a primary one being human activity, which is known to be a large contributor to wildfire ignition, as approximately 95% of recent Southern California wildfires are caused by humans (Keeley 1982; Keeley and Fotheringham 2001). One direction for future research is to include the distance of a location from major roads and highways as a measure of accessibility to a given area. Of course, human interaction does not end with ignition; land use and fire

prevention and suppression decisions can also have substantial impact on wildfire incidence. Indeed, it may be argued (as in Mees and Chase, 1991) that discrepancies between the BI and the occurrence of large wildfires may simply indicate appropriate response and suppression activities by Fire Department personnel, rather than any sort of problem with the BI. On the other hand, this seems to contradict the justification for the use of the BI based on its apparent correlation with wildfire incidence.

Another variable of interest is wind direction, which may have a large impact on wildfire occurrence rates. The warm Santa Ana winds blowing from the North-East are associated with low relative humidity and other conditions conducive to wildfire ignition (Keeley et al. 1999; Hu and Liu 2003). Unfortunately, the influence on wildfire incidence of the wind directions recorded at the RAWS is rather difficult to analyze by standard methods and is a subject for future research.

Similarly, other variables known to affect wildfire incidence include fuel moisture, land use, topography, and fuel age. The focus of the present paper is on the use of the variables recorded at the RAWS and used directly in the computation of the BI. The model in (1), based solely on these measurements, seems to be a suitable null model to which the performance of more complex models, such as the BI, can be compared. Substantial improvement in the predictive performance of this model can no doubt be achieved by exploring the spatial variation of other weather measurements and by considering non-separable alternatives as well as more complex forms for the functional relationships between wildfire incidence and each of the weather variables. The simplicity of model (1) is attractive for purposes of comparison and estimation, and it is especially striking in light of this simplicity that model

(1) so vastly outperforms model (2) in terms of goodness-of-fit.

The comparison of these results with other regions is an important area for future research. The results here are exclusively for Los Angeles County, and it is likely that the relationships between meteorological variables and wildfire incidence may be much different in other regimes with different types of vegetation and climate. While our results suggest that the BI is not an ideal predictor of wildfire incidence in Los Angeles County using the Southern California Chaparral fuel model, the BI may perform quite differently in other regions and with different fuel models. In addition, in this analysis we have only considered wildfires burning greater than 0.0405 km^2 (10 acres); analysis of more precise and complete wildfire catalogs may yield different results. We nevertheless suggest that the non-parametric and semi-parametric techniques described here can be used in order to critically assess the predictive performance for other wildfire datasets in other areas.

Acknowledgements. Our sincerest gratitude to Larry Bradshaw at the USDA Forest Service for so generously providing us with the weather station data and helping us process it and generate the corresponding BI values. We also thank the members of the Los Angeles County Fire Department and the Los Angeles County Department of Public Works (especially Mike Takeshita and Frank Vidales, and Herb Spitzer) for sharing their data and expertise. Thanks also to Roger Peng, James Woods, Haiyong Xu, and both anonymous reviewers for their very useful suggestions.

Biographical Sketches. Frederic Paik Schoenberg is an Associate Professor of Statistics at UCLA. His main research interests are point processes and applications in fire ecology and seismology. Jamie Pompa received her Master's degree in Statistics from UCLA in July, 2006. She currently works for the Department of Defense. Chien-Hsun Chang is a Ph.D. student in Statistics at UCLA. Her research focuses on separability tests for multi-dimensional marked point processes.

References

- Akaike, H. 1977. On Entropy Maximization Principle. In Applications of Statistics, P.R. Krishnaiah, ed., North-Holland: Amsterdam, 27–41.
- Andrews, P.L., and L.S. Bradshaw. 1997. FIRES: Fire Information Retrieval and Evaluation System - a program for fire danger rating analysis. Gen. Tech. Rep. INT-GTR-367. Ogden, UT; U.S. Department of Agriculture, Forest Service, Intermountain Research Station. 64 p.
- Andrews, P.L., D.O. Loftsgaarden, and L.S. Bradshaw. 2003. Evaluation of fire danger rating indexes using logistic regression and percentile analysis. International Journal of Wildland Fire 12: 213-226.
- Bradshaw, L.S., and S. Brittain. 1999. FireFamily Pls: fire weather and fire danger climatology at your fingertips. in 3rd Symposium on Fire and Forest Meteorology, Long Beach, CA. American Meteorological Society, Boston.
- Bradshaw, L.S., J.E. Deeming, R.E. Burgan, and J.D. Cohen. 1983. The 1978 National Fire-Danger Rating System: Technical Documentation. United States Department of Agriculture Forest Service General Technical Report INT-169. Intermountain Forest and Range Experiment Station, Ogden, Utah. 46 p.
- Bradshaw, L.S., and E. McCormick. 2000. FireFamily Plus User's Guide, Version 2.0. USDA Forest Service.
- Burgan, R.E. 1988. 1988 revisions to the 1978 National Fire-Danger Rating System. USDA

Forest Service, Southeastern Forest Experiment Station, Research Paper SE-273.

Cressie, N.A. 1993. *Statistics for Spatial Data*, revised ed. Wiley, New York.

Daley, D. and D. Vere-Jones. 1998. *An Introduction to the Theory of Point Processes*. Springer, NY.

Daley, D. and D. Vere-Jones. 2003. *An Introduction to the Theory of Point Processes*, 2nd ed., Volume I. Springer, New York.

Deeming, J.E., R.E. Burgan, and J.D. Cohen. 1977. *The National Fire-Danger Rating System* – 1978. Technical Report INT-39, USDA Forest Service, Intermountain Forest and Range Experiment Station.

Diggle, P. 1985. A kernel method for smoothing point process data. *Applied Statistics* 34(2): 138–147.

Geisser, S. 1975. The predictive sample reuse method with applications. *Journal of the American Statistical Association* 70: 320-328.

Haines, D.A., W.A. Main, J.S. Frost, and A.J. Simard, A.J. 1983. Fire-danger rating and wildfire occurrence in the Northeastern United States. *Forest Science* 29(4): 679–696.

Hall, P. 1987. On Kullback-Leibler loss and density estimation. *Annals of Statistics* 15: 1492-1519.

Härdle, W. 1991. *Smoothing Techniques: With Implementation in S*. Springer, New York.

Hu, H., and W.T. Liu. 2003. Oceanic thermal and biological responses to Santa Ana winds. *Geophysical Research Letters* 30(11): 50-1 – 50-4.

- Keeley, J.E. 1982. Distribution of lightning and man-caused wildfires in California. pp. 431-437 in Proceedings of the symposium on dynamics and management of Mediterranean-type ecosystems. General technical report PSW-58, C.E. Conrad and W.C. Oechel, editors. Forest Service, Pacific Southwest Forest and Range Experiment Station, Berkeley, CA.
- Keeley, J.E. 2000. Chaparral. P. 203–253 in North American Terrestrial Vegetation, 2nd Edition, M.G. Barbour and W.D. Billings, (eds.). Cambridge University Press, N.Y.
- Keeley, J. 2002. Fire management of California shrubland landscapes. *Environmental Management* 29: 395-408.
- Keeley, J.E., and C. J. Fotheringham. 2001. History and management of crown-fire ecosystems: a summary and response. *Conservation Biology* 15(6): 1536–1548.
- Keeley, J.E., C.J. Fotheringham, and M. Morais. 1999. Reexamining fire suppression impacts on brushland fire regimes. *Science* 284: 1829–1832.
- Mees, R. and R. Chase. 1991. Relating burning index to wildfire workload over broad geographic areas. *International Journal of Wildland Fire* 1: 235-238.
- Ogata, Y. 1988. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association* 83(401): 9–27.
- Ogata, Y. 1998. Space-time point-process models for earthquake occurrences. *Ann. Inst. Statist. Math.* 50(2): 379–402.
- Ogata, Y., and H. Akaike. 1982. On linear intensity models for mixed doubly stochastic

- Poisson and self-exciting point processes. *Journal of the Royal Statistical Society, Series B* 44(1): 102–107.
- Diggle, P.J. 1985. A kernel method for smoothing point process data. *Applied Statistics* 34: 138–147.
- Park, B.U., and J.S. Marron. 1990. Comparison of data-driven bandwidth selectors. *Journal of the American Statistical Association* 85: 66–72.
- Park, B.U., and B.A. Turlach. 1992. Practical performance of several data driven bandwidth selectors. *Computational Statistics* 7: 251–270.
- Peng, R. D., F.P. Schoenberg, and J. Woods. 2005. A space-time conditional intensity model for evaluating a wildfire hazard index. *Journal of the American Statistical Association* 100(469): 26–35.
- Pyne, S.J., P.L. Andrews, and R.D. Laven. 1996. *Introduction to Wildland Fire*, 2nd ed. John Wiley & Sons, Inc., New York, NY.
- Rathbun, S.L. 1996. Asymptotic properties of the maximum likelihood estimator for spatiotemporal point processes. *Journal of Statistical Planning and Inference* 51: 55–74.
- Rothermel, R.C. 1991. Predicting behavior of the 1988 Yellowstone fires: projections versus reality. *Int. J. Wildland Fire* 1(1): 1–10.
- Schoenberg, F.P. 2004. Testing separability in multi-dimensional point processes. *Biometrics* 60: 471–481.

- Schoenberg, F.P., R. Peng, R., and J. Woods. 2003. On the distribution of wildfire sizes. *Environmetrics* 14(6): 583–592.
- Schoenberg, F.P. 2006. A note on the separability of multidimensional point processes with covariates. UCLA Preprint Series, No. 496.
- Schuster, E.F., and C.G. Gregory. 1981. On the non-consistency of maximum likelihood nonparametric density estimators. In *Computer Science and Statistics: Proceedings of the 13th Symposium on the interface*, W. Eddy, ed.
- Scott, D.W. 1992. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, New York.
- Scott, D.W., and L.E. Factor. 1981. Monte Carlo study of three data based nonparametric density estimators. *Journal of the American Statistical Association* 76: 9-15.
- Silverman, B. W. 1986. *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.
- Stone, C. 1984. An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics* 12: 1285-1297.
- Stone, M. 1974. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*(36): 111- 147.
- Turner, M.G. and W.H. Romme. 1994. Landscape dynamics in crown fire ecosystems. *Landscape Ecology* 9(1): 59–77.

- Venables, W.N., and B.D. Ripley. 2002. *Modern Applied Statistics with S*. Springer, New York.
- Wand, M.P., and M.C. Jones. 1995. *Kernel Smoothing*. Chapman and Hall, London.
- Warren, J. R., and D.L. Vance. 1981. *Remote Automatic Weather Station for Resource and Fire Management Agencies*. United States Department of Agriculture Forest Service Technical Report INT-116. Intermountain Forest and Range Experiment Station.
- Wilks, S.S. 1962. *Mathematical Statistics*. Wiley, New York.

Figure Captions:

Figure 1: Smoothed estimate of number of fires per day by wind speed, smoothed using a Gaussian kernel smoother and a bandwidth of 0.449 miles per hour, calculated using Silverman's rule. The data points used to calculate the estimate are also shown, with each point corresponding to one day.

Figure 2: Bar plot for wind speed measurements, separated into deciles.

Figure 3: Smoothed estimate of number of fires per day by precipitation, smoothed using a Gaussian kernel smoother and a bandwidth of 0.0425 inches, calculated using Silverman's rule.

Figure 4: Smoothed estimate of the relationship between number of fires per day and maximum daily relative humidity, smoothed using a Gaussian kernel smoother. The solid grey line corresponds to the smoothed estimate using the likelihood cross validation bandwidth of 0.0206% and the dashed black line to the smoothed estimate using the modified likelihood cross validation bandwidth of 8.86%.

Figure 5: Smoothed estimate of number of fires per day by wind speed, smoothed using a Gaussian kernel smoother and a bandwidth of 0.449 miles per hour, calculated using Silverman's rule. The dashed curves correspond to the 95% confidence limits of the smoothed estimate.

Figure 6: Centroids of Los Angeles County wildfires from 1951-1975 (*), those fires from 1976-2000 which are closer to the Los Angeles County boundary than a historical fire (+), and all other wildfires from 1976-2000 (*).

Figure 7: Estimate of the spatial fire rate based on historical fires from 1951 through 1975

using an exponential kernel smoother for Los Angeles County. Locations of the wildfire centroids are overlaid.

Figure 8: Plot of the differences $\log \lambda_1(x_i, y_i, t_i, a_i) - \log \lambda_2(x_i, y_i, t_i, a_i)$, where λ_1 and λ_2 represent models (1) and (2), respectively. This plot shows the differences with respect to the variables maximum relative humidity and date within the year on days when wildfires occurred, with darker pixilation representing a better fit by model (1) than model (2). Data from 1976-2000 are plotted here and are used in fitting each model.

Figure 9: Plot of the differences $\log \lambda_1(x_i, y_i, t_i, a_i) - \log \lambda_2(x_i, y_i, t_i, a_i)$, where λ_1 and λ_2 represent models (1) and (2), respectively. The plot shows the differences with respect to the variables wind speed and temperature on days when wildfires occurred, with darker pixilation representing a better fit by model (1) than model (2). Data from 1976-2000 are plotted here and are used in fitting each model.

Figure 10: Plot of the differences $\log \lambda_1(x_i, y_i, t_i, a_i) - \log \lambda_2(x_i, y_i, t_i, a_i)$, where λ_1 and λ_2 represent models (1) and (2), respectively. The plot shows the differences with respect to the variables wind speed and temperature on days when wildfires occurred between 1988-2000, with darker pixilation representing a better fit by model (1) than model (2). Data from 1976-1988 are used in fitting each model, and data from 1989-2000 are plotted here for the purpose of model evaluation.