Impact of Weather Covariates on Wildfire in Tanjung Puting National Park

Esa Eslami[1], Akane Nishimura[2], and Frederic Paik Schoenberg[1]

Running Title: Impact of Weather Covariates on Wildfire

Keywords: forest fires, separability, kernel regression, South-East Asia, Borneo.

[1] Department of Statistics, University of California, Los Angeles, CA 90095–1554, USA.

phone: 310-794-5193

fax: 310-206-5658

email: frederic@stat.ucla.edu

Postal address: UCLA Dept. of Statistics

8142 Math-Science Building

Los Angeles, CA 90095–1554, USA.

[2] Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA 90095–1554, USA.

Abstract

This paper explores wildfire modeling based on meteorological variables for Tanjung Puting National Park, located on the island of Borneo. Based on the point process models developed in other papers to describe wildfires in Los Angeles County, a separable, or entirely multiplicative, model is developed and each individual component is estimated using kernel smoothing and maximum likelihood methods. The data are shown to be largely compatible with a separable model, suggesting that the impact on wildfire burn area of a particular weather variable does not appear to vary significantly depending on the values of other weather variables.

1 Introduction

The island of Borneo has suffered severe deforestation and forest degradation over the past two decades, with fire acting as a significant factor (Langner et al. 2007). Located on the southern coast of the island's Indonesian territory known as Kalimantan, Tanjung Puting National Park covers over 450,000 hectares and is susceptible to forest fires year round. The park contains a variety of habitats, including lowland rainforest, seasonal swamp forest as well as other agricultural areas, and is well known as the home of Camp Leakey, a world renowned center for the study and rehabilitation of orangutans (Galdikas et al. 1994).

Accurate estimation of wildfire hazard is very important in aiding National Park officials to prepare supplies and staff in preventing, combatting, and controlling large wildfires. One way to increase the park's estimates of wildfire hazard would be to produce a statistical model that utilizes weather variables such as mean humidity, mean temperature, and precipitation to predict their impact on wildfire incidence in the national park.

Recently, separable point process models have been used to estimate wildfire hazard in Southern California, as a function of weather variables (Schoenberg et al. 2007). Schoenberg et al. (2007) explores the fit of such models to Tanjung Puting National Park. Using weather variables as covariates, components of a purely multiplicative model can readily be estimated individually if the assumption of separability is satisfied (Chang and Schoenberg 2008). 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. While Schoenberg et al. (2007) found separable models to fit rather well to wildfire data in Southern California, a question posed was whether these types of models could fit adequately in other regions. Here, we explore the use of kernel smoothing and semi-parametric approaches in estimating separable point process models for fire incidence in Tanjung Puting National Park. The purpose of fitting such a model is not only for the accurate estimation of wildfire hazard on a given day, but also in order to simulate realistic wildfire behavior given conditions in the National Park.

A description of the weather and fire data for Tanjung Puting National Park used in this paper can be found in Section 2. Kernel smoothing techniques as well as several bandwidth selection methods are explored in Section 3. The definition of separability is also reviewed in Section 3, and the different distributions explored in order to simulate fires for testing separability are described. Results of the methods chosen in Section 3 are then detailed and explained in Section 4. A summary and conclusions are given in Section 5.

2 Data

There are over 160 weather stations located among the islands of Indonesia. Situated in Pangkalan Bun within the boundaries of Tanjung Puting National Park ($-2\,°7'$, $111\,°7'$, elevation 25 meters) weather station 966450 (WRBI) records a variety of daily meteorological variables. We focus here on temperature, sea level pressure, humidity, precipitation, visibility, and wind speed, collected from January 2001 to January 2007. The data are gathered from the meteorological service of the United States, as presented on Tutiempo.net, which bases its data summaries on data exchanged under the World Meteorological Organization (WMO) World Weather Watch Program according to WMO Resolution 40 (Cg-XII) (Tutiemp.net 2007).

The MODIS Rapid Response System utilizes a contextual fire detection algorithm that incorporates a combination of an absolute threshold test and a series of contextual tests that look for the characteristic signature of an active fire using two 4μm wavelength bands and an 11μm wavelength band (Giglio et al. 2003). The algorithm further uses cloud and water masking, as well as several false alarm rejection tests such as sun glint rejection to verify the existence of detected wildfires. On-board the satellites Terra and Aqua, the MODIS sensor passes over Borneo four times a day, ensuring accurate and thorough coverage of fire activity on the Island (Miettinen et al. 2007). The MODIS sensor is a well-established system used to recognize fires at a spatial resolution of 1 km (Justice et al. 2002). All fires detected within the region of Tanjung Puting National Park from January 2001 to January 2007 by the MODIS sensor on both the Terra and Aqua satellites, whose total area exceeded 9600 km^2, were used for this analysis.

Data were missing for one or more weather variables on certain days over the time range considered here. We restrict our attention to the 1533 days where temperature, visibility, wind speed, sea level pressure, humidity, and precipitation were all recorded. On these days, there were 329 days on which fires were recorded, with $793km^2$ being the largest amount of area burned on any single day during this 6 year period.

3 Methods

Spatial-temporal marked point process models are used to represent observations of rare events such as wildfires or earthquakes. For a thorough treatment of point processes and related constructs, see Daley and Vere-Jones (2003). A few important details are summarized here. A point process N is a random collection of points in some metric space χ. In modeling the occurrence of wildfires, for example, one may identify with each event a point $(t, \mathbf{x}, m) \in \mathbf{R}^5$, where t represents the time of the event's origin, $\mathbf{x}$ the corresponding three-dimensional location, and m a real-valued measure of its size. The basic construct of a point process model is the *conditional intensity* (CI), $\lambda(t, \mathbf{x}, m)$, which one can interpret as the limiting expected rate at which points of mark m amass around any location $(t, \mathbf{x})$ of space-time, conditional on the history of the process prior to time t.

In order to model the incidence of wildfires in Tanjung Puting National Park one technique would be to create a model based on the point process models developed in other papers to describe wildfires in Los Angeles County. As suggested by Schoenberg (2004), a model that is purely multiplicative, or *separable* in the terminology of Cressie (1993), may be appropriate. Typically, in such models, each component of the model may be estimated

individually. Schoenberg et al. (2007) considered a model of the form

$$\lambda(t, \mathbf{x}, m) \quad = \quad f_1(P(t))f_2(V(t))f_3(H(t))f_4(S(t))f_5(T(t))f_6(W(t))\mu(\mathbf{x})g(m), \qquad (1)$$

where $P(t), V(t), H(t), S(t), T(t)$, and $W(t)$ represent precipitation, visibility, humidity, sea level pressure, temperature, and wind speed, respectively, for day t, $\mu(\mathbf{x})$ represents the spatial background rate, and $g(m)$ represents the distribution of wildfire areas.

In estimating each of the individual component functions f_i in (1), one approach is to use a non-parametric method such as kernel smoothing Silverman (1986). That is, if y represents the corresponding weather variable in equation (1), then the component $f_i(y)$ may be estimated using kernel regression via $\hat{f}(y) = \sum_i m_i K(x - x_i); h) / \sum_i K(x - x_i; h)$, where m_i represents the area burned in wildfires on day i, and x_i is the value of the weather variable 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 represents the *bandwidth*, which controls the degree of smoothing.

There are several different methods for automatically choosing a bandwidth for kernel smoothing. Silverman's "rule of thumb" bandwidth selection technique is a common method used for automatically choosing a bandwidth for kernel smoothing, where the bandwidth $h = 0.9min\{s, IQR\}n^{-1/5}/1.34$, with s the sample standard deviation, IQR the inter-quartile range, and n the number of observations of the variable being smoothed (Silverman 1986). The bandwidth chosen by Silverman's rule, however, often is too small when the covariate under consideration is not normally distributed (Silverman 1986; Schoenberg et al. 2009).

Another method commonly used in bandwidth selection is the likelihood cross validation (LCV) technique (Silverman 1986). This approach temporarily removes each point x_i in

the dataset and then calculates the estimate of the kernel smoothed function at that point using an initial bandwidth h. This value, $\hat{f}(x_i; h)$, is then used to calculate the distance $d(x_i; h) = |\hat{f}(x_i; h) - f(x)|$ from the observed quantity $f(x_i)$. The bandwidth h that minimizes $\sum_i log(d(x_i; h))$ is then chosen as the optimal bandwidth. LCV bandwidth selection is not optimal, however, when used to estimate the relationship between a particular weather variable and observations of rare events such as fire incidence (Schoenberg et al. 2009). In particular, when the covariate has many repetitions of identical values, bandwidths estimated by LCV tend to be too small. This is the case for the observed weather variables studied, where over 58% of mean temperature observations, for example, are exactly the same on ten or more days.

In light of the shortcomings of likelihood cross validation, (Schoenberg et al. 2009) suggests a modified version of LCV bandwidth selection that will result in a smoother estimate. In modified likelihood cross validation, 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. Thus, rather than removing one observation at a time, the modified LCV approach removes one small portion of the x-axis at a time. As with LCV, the bandwidth h that minimizes $\sum_i log(d(x_i; h))$ is then chosen as the optimal bandwidth.

One may consider several choices for the size distribution $g(m)$ in model (1). Cumming (2001) and Schoenberg et al. (2003) have suggested that the overall wildfire size distribution tends to be well-approximated by the Pareto or tapered Pareto distributions. However, the size distribution may depend substantially on the weather variables, as noted in Schoenberg

(2004). Hence one may consider replacing $g(m)$ in the model (1) with the modification

$$g_t(m) = j(m; P(t), V(t), H(t), S(t), T(t), W(t)), \tag{2}$$

for some density function, j. We considered several different choices for the density j, including not only the Pareto and tapered Pareto distributions but also the Poisson, truncated normal, and exponential forms.

The model (1) is purely multiplicative, and one may wish to test whether such a model, which is called *separable* in the terminology of Cressie (1993), may be appropriate. Several statistics for testing separability in point process models were proposed in Schoenberg (2004), and extended in Chang and Schoenberg (2008) to the case of multi-dimensional point processes with covariates. The method described in Schoenberg (2004) involves selecting a pair of covariates, and comparing a bivariate kernel smoothing $\hat{\lambda}$, smoothed with respect to both covariates, with the product $\tilde{\lambda}$ of two univariate kernel estimates, smoothed with respect to each of the covariates individually. The statistics suggested by Schoenberg (2004) and Chang and Schoenberg (2008) to be most powerful in detecting departures from separability is the integrated squared difference between these two kernel estimates, i.e.

$$S_3 = \int_0^T \int_{\mathbf{R}^d} \int_{\mathbf{R}} [\hat{\lambda}(t, \mathbf{x}, m) - \tilde{\lambda}(t, \mathbf{x}, m)]^2 dm \, d\mathbf{x} \, dt. \tag{3}$$

In order to produce p-values for these test statistics, simulations of separable kernel estimates each with CI equal to $\tilde{\lambda}(t, \mathbf{x}, m)$ may be used. In addition, one may assess the fit of the resulting separable model by computing its mean squared error in predicting daily wildfire area burned, and comparing with a simple alternative such as a homogeneous Poisson model.

4 Results

Wildfire incidence in Tanjung Puting National Park appears to depend critically on weather variables such as precipitation, temperature, humidity, and atmospheric pressure. For instance, the solid curve in Figure 1 shows a smoothed estimated of the relationship between daily area burned and sea level pressure, obtained by kernel regression using a Gaussian kernel function and bandwidth selected by modified LCV. While the scatter about the curve is considerable, one can discern that as atmospheric pressure increases, so too does the average daily burn area. (Note that in Fig. the y-axis has been truncated to highlight the smoothed curve, but as a result not all points are shown in the Figure.)

Figure 2a shows the smoothed estimate of the relationship between daily burn area and visibility. As visibility increases, the mean area burned in wildfires decreases rapidly. In fact, on days where mean visibility exceeds 5 kilometers, the mean area burned becomes infinitesimal. This kernel regression plot of mean visibility and number of fires per day suggests an exponential form for the function f_2 in the model (1). Similar kernel regression plots of number of daily fires against each of the other four weather variables suggest exponential forms for f_3, f_4, f_5, and f_6, whereas a linear model appears preferable for f_8.

The assumption of separability in the model (1) should be tested to ensure that a separable model is in fact appropriate for the data. Figure 4 shows the nonseparable and separable CI estimates $\hat{\lambda}$ and $\tilde{\lambda}$ as a function of temperature and mean sea level pressure. Both CI estimates show that when mean sea level pressure is high expected area burned is high, though the two estimates have obvious discrepancies, especially when both temperatures and atmospheric pressures are highest. Nevertheless, Figure 6a shows that the difference

between the nonseparable and separable CI estimates shown in Figure 4 are not statistically significant. The estimated p-value of S_3 using 100 simulations is 0.22, suggesting that a separable model for mean temperature and mean sea level pressure may be reasonable for wildfire incidence in Tanjung Puting National Park.

Similar to Figure 4, Figure 5 shows the nonseparable and separable CI estimates $\hat{\lambda}$ and $\tilde{\lambda}$ as a function of humidity and precipitation. The two CI estimates in Figure 5 appear to agree generally. Both the nonseparable and separable CI estimates in Figure 5 are high when humidity is between 58% and 68%, and precipitation is low. The nonseparable CI estimate predicts a high amount of area burned when precipitation is below 25 millimeters, while the separable CI estimate expects a high amount of area burned when precipitation is below 10 millimeters. Figure 6b shows that the difference between the nonseparable and separable CI estimates shown in Figure 5 is not statistically significant. The estimated p-value of S_3 using 100 simulations is 0.35, suggesting that a separable model for mean humidity and precipitation may be reasonable for wildfire incidence in Tanjung Puting National Park. Similar tests of separability were conducted for all possible combinations of weather variables and their p-values are presented in Table 1.

	$W(t)$	$T(t)$	$S(t)$	$H(t)$	$V(t)$
$P(t)$	0.54	0.54	0.38	0.35	0.52
$V(t)$	0.37	0.43	0.65	0.79	
$H(t)$	0.46	0.88	0.46		
$S(t)$	0.26	0.22			
$T(t)$	0.61				

Table 1: Estimated p-values of S_3 using 100 simulations for testing the separability of each

of the weather covariates in the model (1). $P(t), V(t), H(t), S(t), T(t)$, and $W(t)$ represent precipitation, mean visibility, mean humidity, mean sea level pressure, mean temperature, and mean wind speed, respectively, for day t.

Table 1 shows that a separable, or purely multiplicative form for the model (1) may be reasonable in light of the fact that the difference between the nonseparable and separable CI estimates for any two covariates c_i and c_j are not statistically significant. The implication is that the relationship between wildfire area burned and one covariate such as temperature, for example, does not appear to change significantly depending on the values of the other covariates.

One may wonder to what extent the weather variables recorded by the MODIS system and used in model (1) result in improved predictions of daily wildfire burn area. Compared to the best-fitting homogeneous Poisson model with constant intensity over all days, using the separable CI estimate of mean temperature and mean visibility alone, the root mean squared error decreases from 38.2 km^2 to 29.2 km^2.

5 Discussion

Accurate wildfire prediction based solely on daily weather variables such as those considered in the model (1) is inherently limited. Weather is only one of several factors relating to wildfire occurrence and spread in Tanjung Puting National Park. In addition to obvious human interactions with wildfire activity such as arson, fire prevention policies, and fire suppression activities, *slash-and-burn* techniques, the preferred method of land clearing in

Indonesia where fire is used as a tool to clear land, can rapidly spread fire if conducted in a negligent fashion or during periods of drought (Tomich et al. 1998). Nevertheless, the use of weather variables for gaining a better knowledge of when Tanjung Puting National Park is most susceptible to wildfire activity would be very valuable to park management and officials. The weather variables are easily attainable for park officials, and thus the use of current weather or immediate future weather information could be used in a model such as that discussed in this paper to inform park officials when they should prepare supplies and staff for containing or fighting particularly large fires. Further, models such as those explored here may readily be used to simulate realistic wildfire patterns and to explore their dependence on local weather conditions.

The separability of model (1) has not been shown to be significantly violated for the dataset considered here. Were we to suggest this model for use by officials at Tanjung Puting National Park we must also note the model (1) is quite simplistic and its fit could no doubt be improved by using more complicated functional forms for each of the terms, as well as considering different interactions between the variables. Furthermore, a homogeneous Poisson model is not an ideal baseline with which to compare the mean squared prediction error, and in future research, actual forward prediction should be used to assess the validity of the model, using data obtained separately from that used in model fitting.

In addition to these shortcomings, many important variables are excluded from the model. Only six weather variables are used, while other important factors such as vegetation, land use, and other various human interaction variables are not included in the model. Nevertheless, the model (1-2) could potentially be used as a starting point for aiding in fire prediction for Tanjung Puting National Park.

References

Chang, C., and Schoenberg, F.P. (2008). Testing separability in multi-dimensional point processes with covariates.

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

Cumming, S.G. (2001). Parametric models of the fire-size distribution. Can. J. For. Res. 31:1297-1303.

Daley, D. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes*, 2nd edition. New York: Springer.

Galdikas, B. and Shapiro, G. (1994). *A Guidebook to Tanjung Puting National Park.* Orangutan Foundation International, U.S.A.

Giglio L., Descloitres J., Justice C.O., Kaufman Y.J. (2003). An enhanced contextual fire detection algorithm for MODIS. *Remote Sensing of Environment 87*, 273-282.

Justice C.O., Giglio L., Korontzi S. et al. (2002). The MODIS fire products. *Remote Sensing of Environment*, 83, 244-262.

Langner A., Miettinen J. and Siegert F. (2007). Land cover change 2002-2005 in Borneo and the role of fire derived from MODIS imagery. *Global Change Biology 13*: 2329-2340.

Miettinen J., Langner A. and Siegert F. (2007). Burnt area estimation for the year 2005 in Borneo using multi-resolution satellite imagery. *International Journal of Wildland Fire 16*: 45-53.

Schoenberg, F.P., Peng, R., and Woods, J. (2003). On the distribution of wildfire sizes. Environmetrics , 14(6), 583–592.

Schoenberg, F.P. (2004). Testing separability in multi-dimensional point processes. *Biometrics 60*: 471-481.

Schoenberg, F.P., Chang, C., Keeley, J., Pompa, J., Woods, J., and Xu, H. (2007). A Critical Assessment of the Burning Index in Los Angeles County, California. *International Journal of Wildland Fire*, to appear.

Schoenberg, F.P., Pompa, J.L., and Chang, C. (2009). A note on non-parametric and semi-parametric modeling of wildfire hazard in Los Angeles County, California. *Environmental and Ecological Statistics 16* (2-3).

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Tomich T.P., Fagi A.M., de Foresta H., Michon G., Murdiyarso D., Stolle F. and van Noordwijk M. (1998). Indonesia's fires: smoke as a problem, smoke as a symptom. *Agroforestry Today*.

Tutiempo.net (accessed 2007). *Historical Weather: Pangkalan Bun*.
http://www.tutiempo.net/en/Climate/Pangkalan_Bun_Iskandar/966450.htm

Vere-Jones, D. (1992). Statistical methods for the description and display of earthquake catalogs. In *Statistics in Environmental and Earth Sciences*, A. Walden and P. Guttorp (eds), 220-236. London: Edward Arnold.

Figure Captions

Figure 1: Estimate (solid line) of total area burned (km^2) per day by mean sea level pressure, smoothed using a Gaussian kernel smoother and bandwidth of 0.8 millibars, calculated by modified likelihood cross-validation. 95% confidence limits (dotted lines) of the smoothed estimate are also shown.

Figure 2: Estimate (solid line) of total area burned (km^2) per day by mean visibility, smoothed using a Gaussian kernel smoother and bandwidth of 0.9 kilometers, calculated by modified likelihood cross-validation. (a) The 95% confidence limits (dotted lines) of the smoothed estimate are shown; (b) the exponential fit (dashed line) is added to the plot.

Figure 3: (a) Total daily area burned ordered chronologically; (b) comparison of observed area burned distribution with area burned distribution simulated from Poisson distribution; (c) comparison of observed area burned distribution with area burned distribution simulated from truncated normal distribution; (d) comparison of observed area burned distribution with area burned distribution simulated from models (1) - (2).

Figure 4: CI estimates of mean temperature and mean sea level pressure: (top) non-separable kernel CI estimate; (bottom) separable kernel CI estimate.

Figure 5: CI estimates of mean humidity and precipitation: (top) non-separable kernel CI estimate; (bottom) separable kernel CI estimate.

Figure 6: Histogram of 100 simulated values of S_3 for testing the separability of (a) mean temperature and mean sea level pressure and (b) mean humidity and precipitation. The dashed vertical line represents the value of S_3 calculated from the observed data.

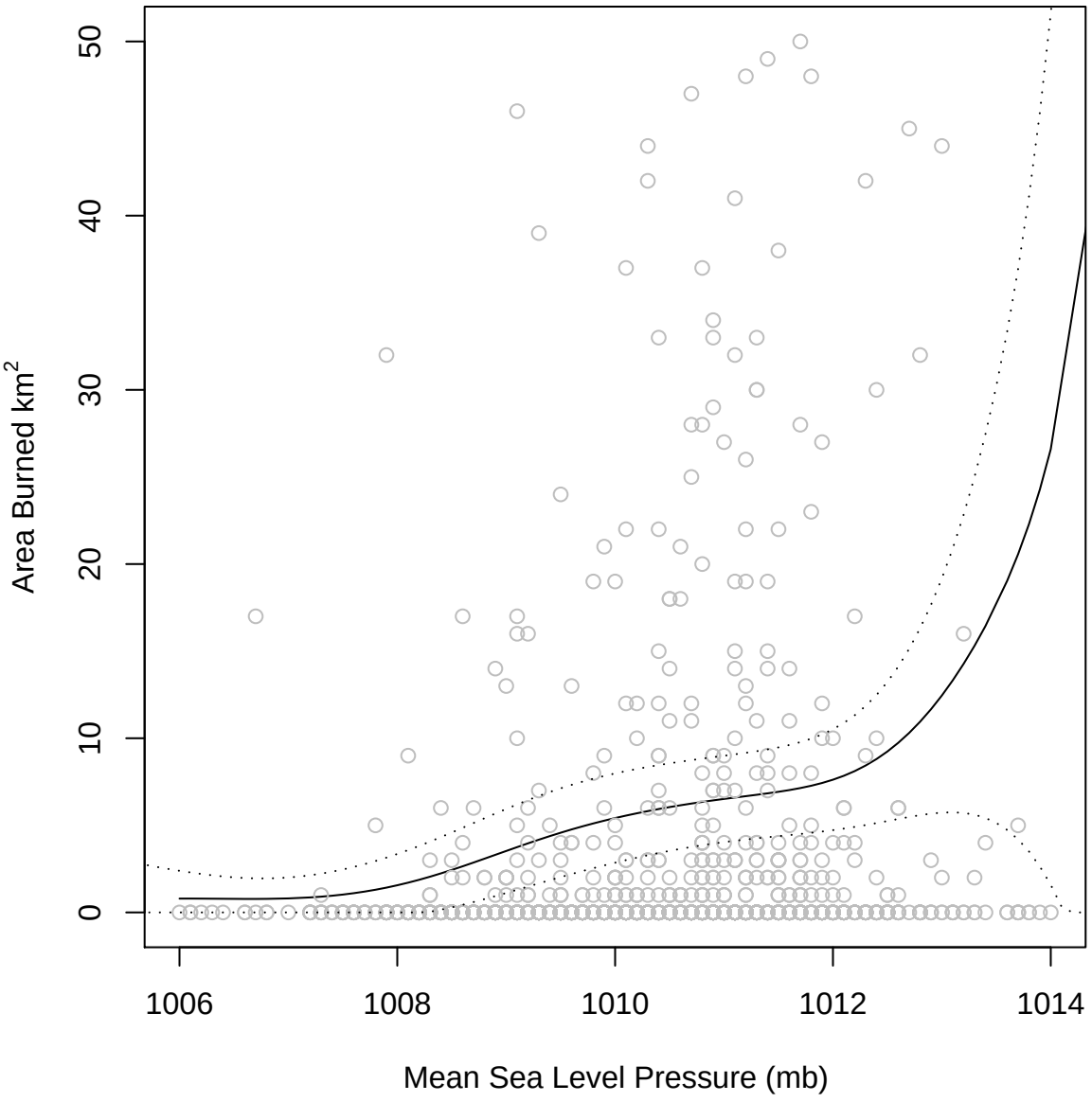
Area Burned km^2
0
10
20
30
40
50
1006
1008
1010
1012
1014
Mean Sea Level Pressure (mb)

Figure 1:

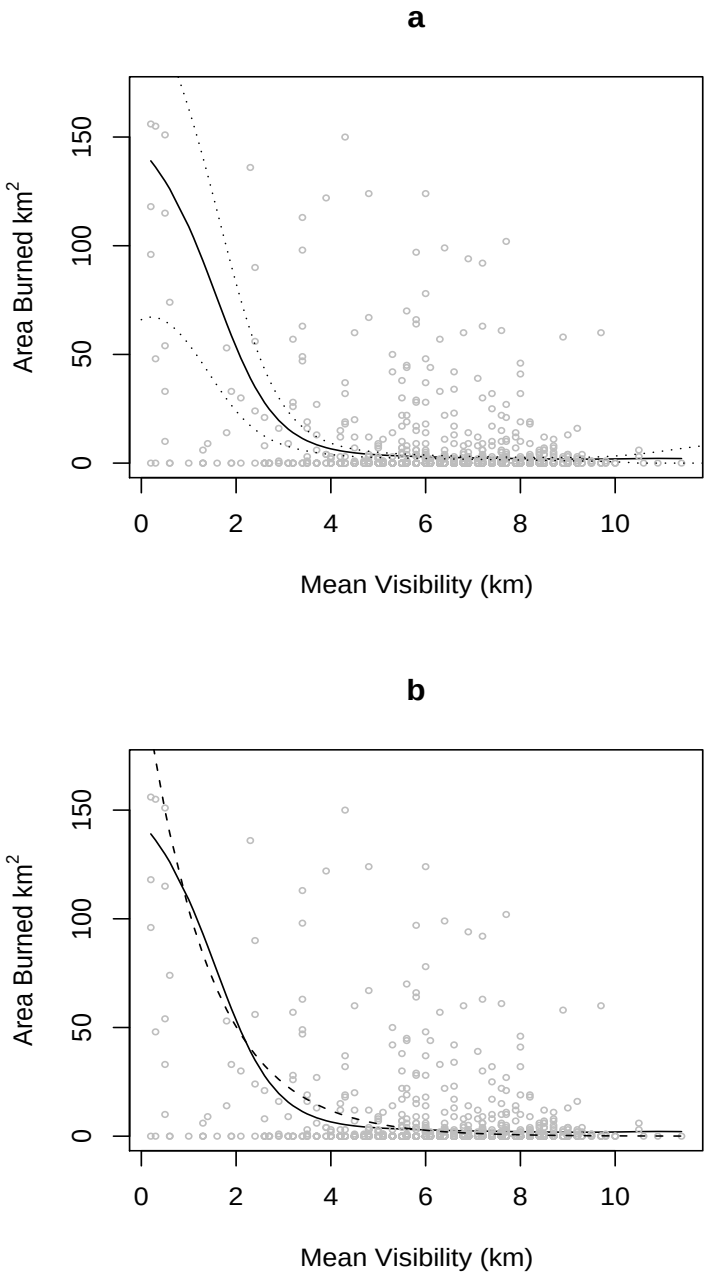
a
Area Burned km^2
0
50
100
150
0
2
4
6
8
10
Mean Visibility (km)
b
Area Burned km^2
0
50
100
150
0
2
4
6
8
10
Mean Visibility (km)

Figure 2:

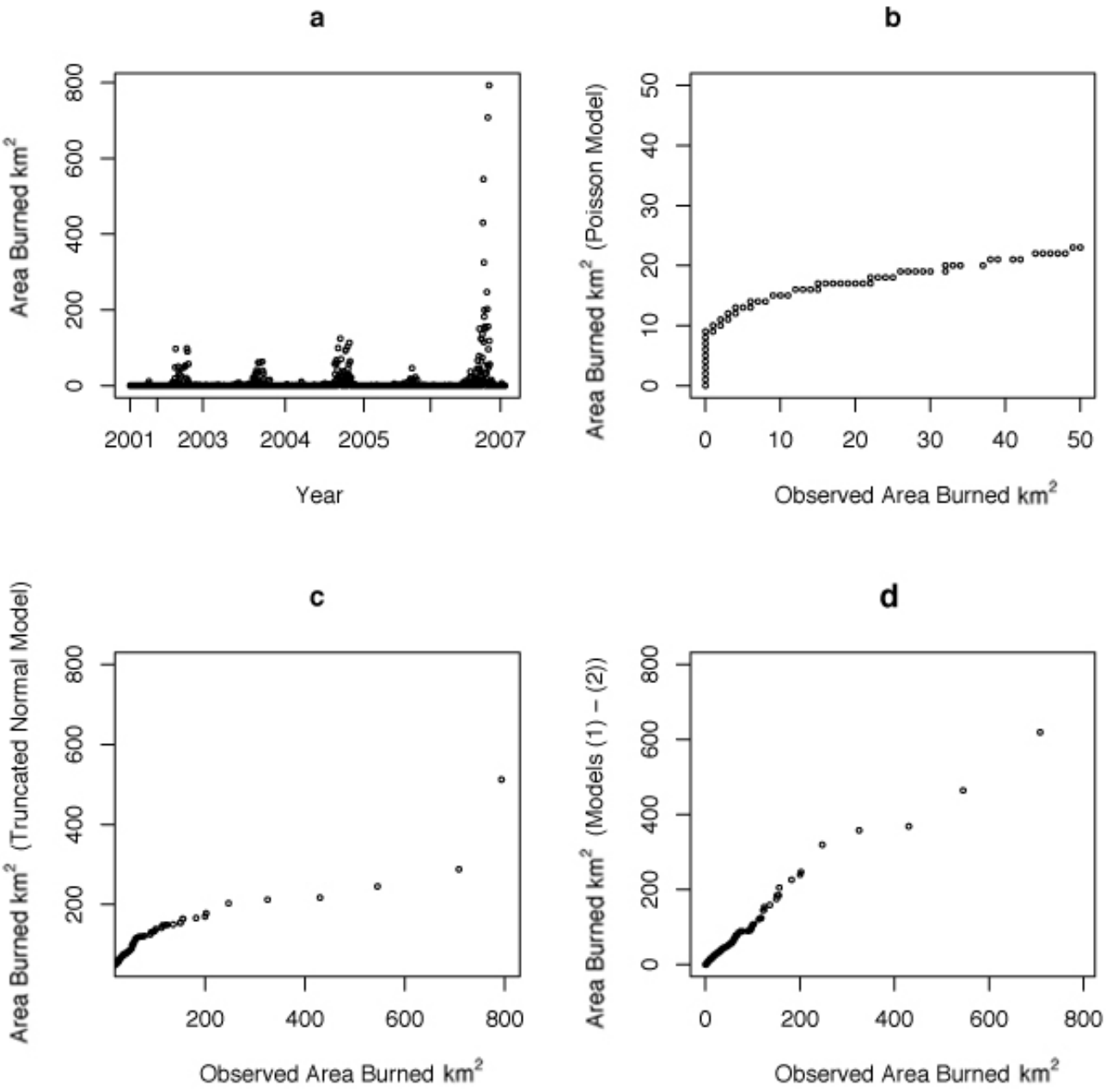
a
Area Burned km^2
0
200
400
600
800
2001
2003
2004
2005
2007
Year
b
Area Burned km^2 (Poisson Model)
0
10
20
30
40
50
Observed Area Burned km^2
c
Area Burned km^2 (Truncated Normal Model)
200
400
600
800
Observed Area Burned km^2
d
Area Burned km^2 (Models (1) – (2))
Observed Area Burned km^2

Figure 3:

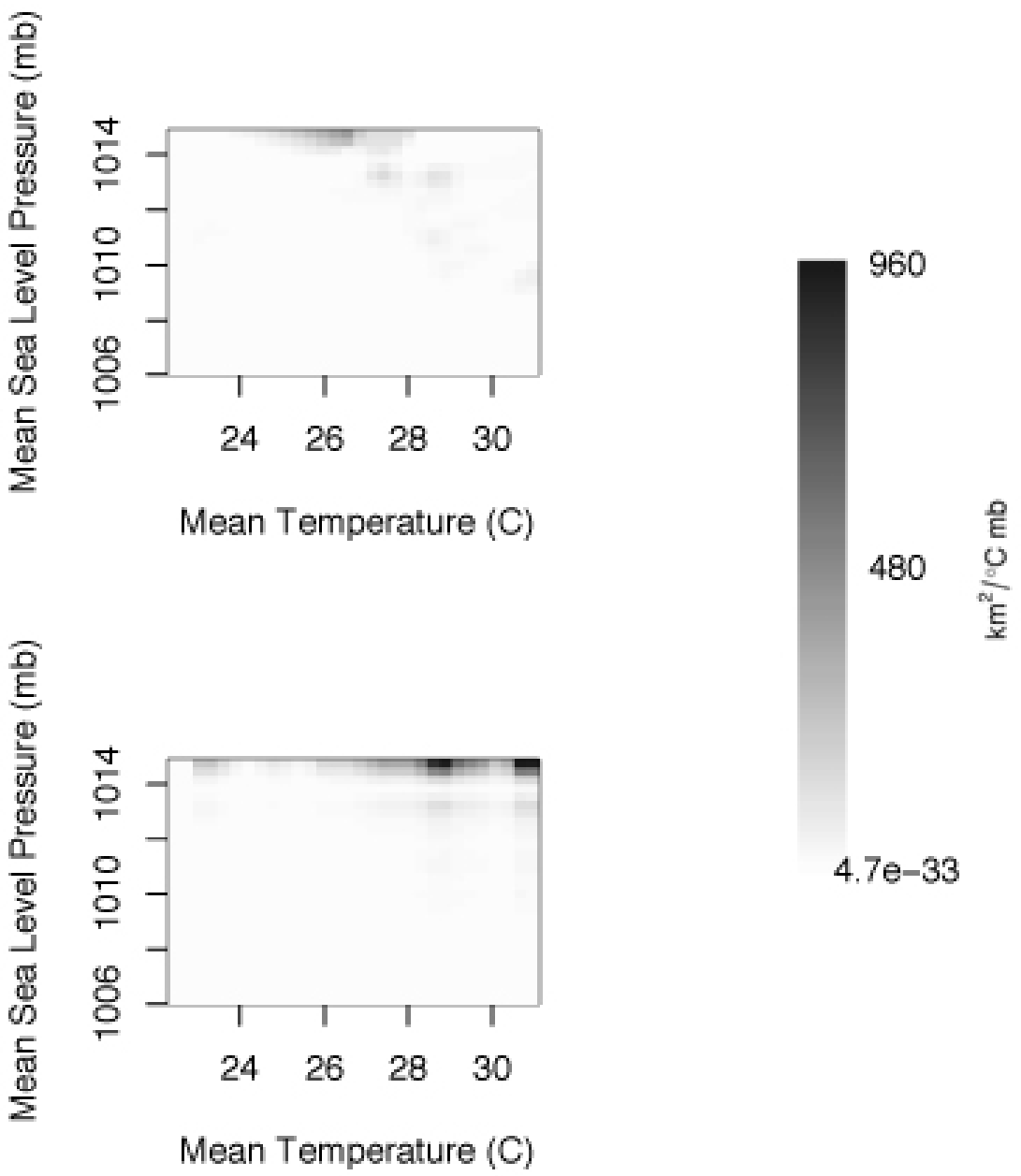
Mean Sea Level Pressure (mb)
1006
1010
1014
24
26
28
30
Mean Temperature (C)
960
480
4.7e−33
km²/°C mb
Mean Sea Level Pressure (mb)
1006
1010
1014
24
26
28
30
Mean Temperature (C)

Figure 4:

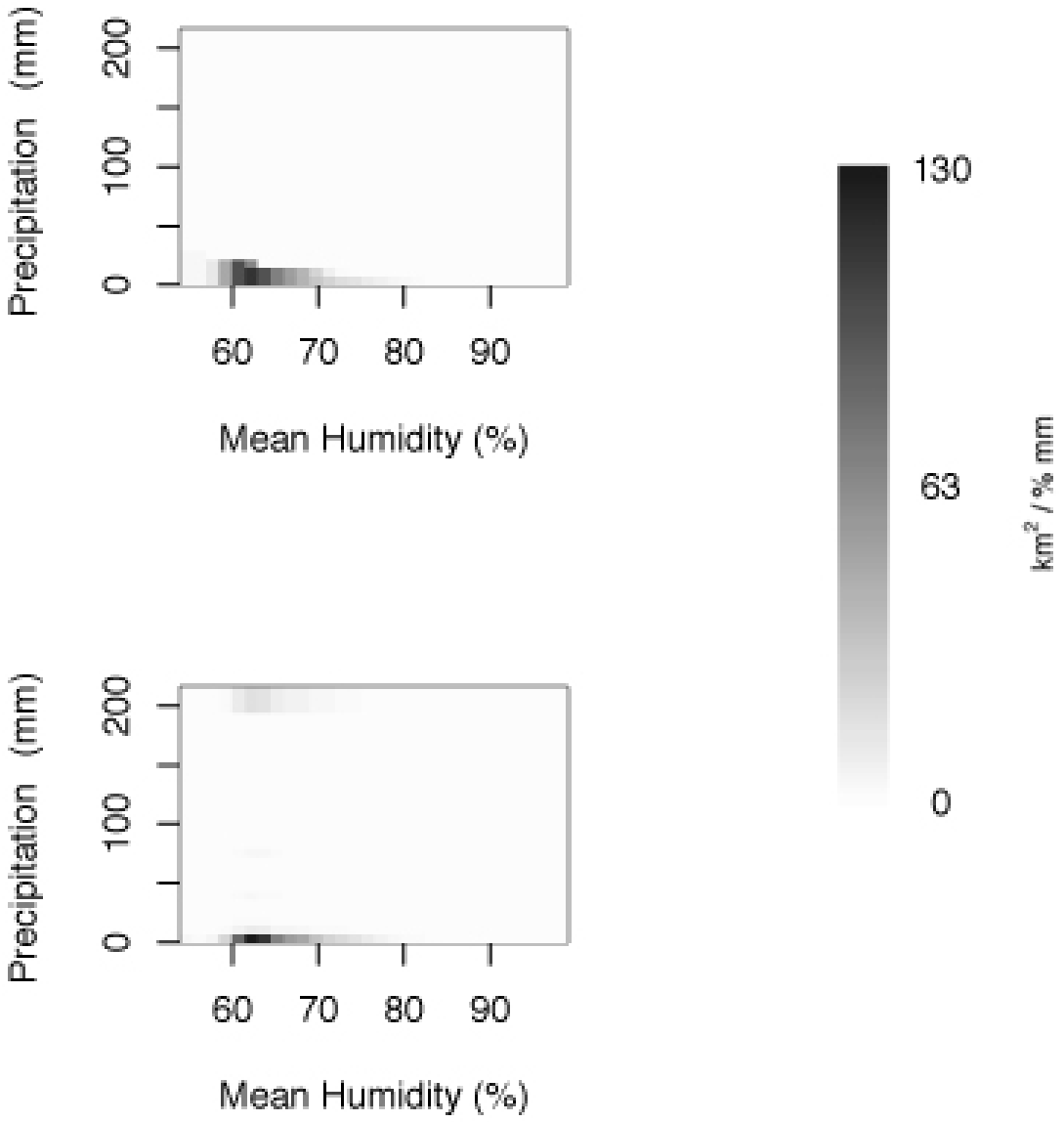
Precipitation (mm)
0
100
200
60
70
80
90
Mean Humidity (%)
Precipitation (mm)
0
100
200
60
70
80
90
Mean Humidity (%)
130
63
0
km² / % mm

Figure 5:

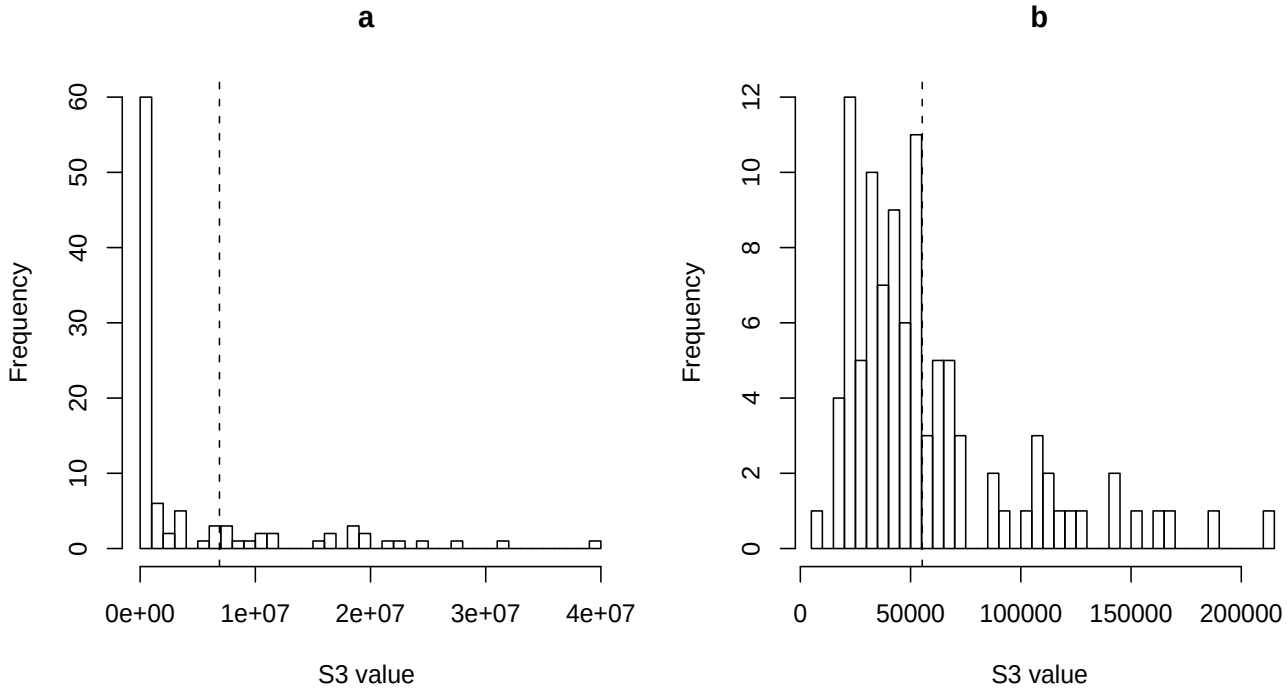
a
Frequency
0
10
20
30
40
50
60
0e+00
1e+07
2e+07
3e+07
4e+07
S3 value
b
Frequency
0
2
4
6
8
10
12
0
50000
100000
150000
200000
S3 value

Figure 6: