California Unplanned Power Outage

XUAN HUYNH

University of California, Los Angeles

Department of Statistics

Abstract

There are many things that cause unplanned outages: traffic accidents, natural disasters, strong winds, heavy rain, poor equipment maintenance, or even metallic balloons caught in overhead wires. When an unplanned power outage happens, it implies that there is a downed electric power pole or fallen power line, which in turn can ignite wildfires. Knowing about where the outage would be more likely to happen not only helps preventing economic loss to power companies but also is useful in wildfire prevention. Within the scope of an introductory spatial statistics course, this report hopes to employ several techniques that were introduced in the course to gain some preliminary knowledge about power outages in California. Thus, this report serves solely as a learning tool, and by no means can its results be used for scientific purposes. However, given that more complete data are available and with more advanced spatial-temporal analyzing and assessing skills, power outages can be an interesting study topic.

1 Introduction

The cause of Kincade Fire, the most recent deadly fire in California, which started at 9:24pm on October 23, 2019 and scorched more than 77,000 acres of Sonoma County, has not been determined. However, the Pacific Gas & Electric (PE&G) reported that it was aware of a transmission line failed near the point of the fire origin. In addition, four months before the Kincade Fire happened, after meticulous investigations, PG& E was accused to be responsible for the Camp Fire in Butte County, which also destroyed roughly 19,000 structures and killed 85 civilians. Further investigations also concluded that the company's failure to properly inspect and upkeep its transmission lines started the lethal blaze.

Unplanned power outage does not directly imply a high chance of having a wildfire, yet, understanding where and when power outages happen can give insights to power companies on their strategy of apparatus maintenance, which in

turn would lower the risk of igniting a fire. This report analyzes the unplanned power outages happened in California using a varieties of point process methods The report proceeds as follows. Section 2 describes the power outage data that are used in the analysis. The details of model used are presented in Section 3. Results and model assessment are discussed in Section 4, followed a brief discussion in Section 5. All figures mentioned in this report are contained in the Appendix.

2 Data

California Open Data Portal is a statewide open data portal website which is sponsored by California Government Operations Agency. Its sub-portal, the California State Geoportal offers geographic data and applications from a multitude of California state entities including land use, education, economy transportation, energy, etc. The dataset used in this report is obtained from Power Outage Incidents API under the Energy section.

The data obtained with this API is updated every 15 minutes and does not keep old records. Hence, the dataset was collected on daily basis, at 7:00PM , starting from April 5, 2020 to May 6, 2020. The data contains 13 features, some are pertinent to the subject of interest: locations, date and time, number of impacted customers and the others are not very helpful: estimated restoration date and time, outage type and outage status color (helpful for making maps), utility company names, outage cause. Each reported incident has a unique ID, which was used to combine daily collected data from the API. The final dataset contains unique unplanned power outage incidents.

There are several concerns about this dataset. In particular, it does not keep track of the power outages of all power companies that are operating throughout California. Large providers, namely, Pacific Gas and Electric (PGE), Southern California Edison (SCE), and San Diego Gas and Electric (SDGE) are included while smaller local electric providers are not present in the data. However, large com-

panies aforementioned provide power service to a significantly large area of the state and have a much greater number of customers, which allows us to, hopefully, draw some conclusions that are applicable to the majority of Californians. Secondly, since the data was collected everyday at a fixed time while the API updates every 15 minutes and may omit some of formerly recorded outages, the data obtained at the fixed time may be incomplete. Lastly, the *StartDate* column in the dataset is not in Pacific Time. The time zone information is missing on the data source page, but it appears to be in the GMT (UTC+0). Nonetheless, this is merely an assumption and should be reevaluated when possible, since it is a crucial factor in spatio-temporal data analysis.

3 Data Analysis & Models

3.1 Clustering Analysis

It is reasonable to expect that clustering exists in the dataset, i.e., power outages occur more frequently in some locations such as urban areas than in the countryside. To testify this expectation, the kernel smoothing density (Fig.1) and L function plots (Fig.2) are used.

The kernel smoothing method used for this dataset employs Gaussian kernel function together with Silverman's rule-of-thumb bandwidth (Silverman, 1986, Section 3.2).

3.2 Purely Spatial Hawkes Model

A simple purely spatial Hawkes model is fitted to the data:

$$\lambda(x, y) = \mu + \alpha x + \beta y + \kappa \sum_i \frac{a_1 \mathrm{e}^{-\mathrm{a_1 D(z,z_i)}}}{2\pi D(z, z_i)}$$

Here, we assume that the background rate is a constant μ add a linear combination of the point location (x,y). κ is productivity, the expected number points triggered by each point whose density function is the triggering function $g(z, z_i) = \frac{a_1 e^{-a_1 D(z_i,z)}}{2\pi D(z_i,z)}$.

z is the location (x, y) scaled to [0,1], and $D(z, z_i)$ is the Euclidean distance between point z and z_i. This is an overly simple model, but within the scope of this introductory course, it serves as a learning tool.

The parameters $\theta = (\mu,\ \alpha,\ \beta,\ \kappa,\ a_1)$ are estimated by maximizing the log-likelihood function, or equivalently, minimizing the negative log-likelihood function:

$$\begin{aligned} -\log L(\theta) &= \int_0^1 \int_0^1 \lambda(x, y, \theta)dxdy - \sum_i log(\lambda(z_i, \theta)) \\ &= \mu + \frac{\alpha}{2} + \frac{\beta}{2} + \kappa N - \sum_i log(\lambda(z_i, \theta)) \end{aligned}$$

3.3 ETAS Model

The R package ETAS was utilized to fit the data. The ETAS model assumes that each event is either a background (spontaneous) event or triggered by a previous event (Ogata 1998), and in this package, the model formula is:

$$\lambda_{\beta,\theta}(t, x, y, m|\mathcal{H}_t) = \{\beta \exp(-\beta(m - m_0))\} \left\{ \mu\rho(x, y) + \sum_{i:t_i<t} \kappa_{A,\alpha}(m_i) g_{c,p}(t - t_i) f_{D,\gamma,q(x-x_i,y-y_i;m_i)} \right\}$$

Details about each term in the formula can be obtained from the package vignette. The parameters to be estimated are β, and θ = ($\mu, A, \alpha, c, \rho, D, \gamma, q$)

Since our data do not have magnitude information, the number of impacted customers is used instead. This number was scaled to range [0,6]. Due to the adverse effect of outliers, only outages that have magnitude between 0.1 and 2.5 are used to best mimic an exponential distribution. The reason is because the Gutenberg-Richter law says that $log_{10}(N_m) = a - bm$ for some a, b, or equiva-

lently, assumes that magnitude follows an exponential distribution (Otaga, 1988). N_m is the number of events that have magnitude greater than or equal to m.

4 Results

4.1 Clustering Analysis

In Figure 1, it can be clearly observed that, power outage incidents cluster around two biggest metropolitan areas of California: San Francisco and Los Angeles. The plot in Figure 2 also confirms this finding: the curve $L(r) - r$ increases along with r, while as if the data indeed follows a homogeneous Poisson process, it should be a horizontal line at zero (the red dashed line). The density plot using kernel method with data scaled to [0,1] range is also shown(Fig.3). Notice that with the same method and bandwidth, the result may look different from one the density plot with the original data. With the scaled data, both x and y axes are scaled to the same range [0,1] even though they should not be (consider the geographic shape of California). Hence, the distances between the points are not scaled appropriately, and this will affect the density plot.

4.2 Purely Spatial Hawkes Model

The following table reports the estimated parameters using the maximizing likelihood method. Standard errors are obtained from the hessian matrix.

θ	μ	α	β	κ	a_1
$\hat{\theta}$	31.73	-20.54	-10.71	0.97	88.53
$\hat{se}$	57.74	50.34	62.43	0.04	6.79

Noticeably, the standard errors of the MLE's used for the background rate are much higher than the estimates themselves. This reflects the overly simple background rate we set up for this model. However, the κ estimate has a reasonable

standard error. Since $\hat{\kappa}$ is positive and very close to 1, it confirms again the presence of clustering in our data. Overall, this model does not seem to be a good fit since most of the estimates have unreasonably large standard errors.

4.3 ETAS Model

The following table reports the estimated parameters returned by the R package ETAS:

θ	β	μ	A	α	c	ρ	D	γ,	q
$\hat{\theta}$	3.4632	1.0218	1.4641	0.0102	0.7548	1.0172	0.0000	1.0053	0.0003
$\hat{\text{se}}$	0.0397	0.0139	0.0361	0.1908	0.0405	0.0007	29.2903	0.0009	10.9680

With the ETAS model, only three out of nine estimates have relatively high values for their standard errors. This signifies that the model seems to be plausible.Fig.4 includes location of epicenters (top left panel), logarithm of frequency by magnitude (bottom left panel), cumulative frequency over time (bottom middle panel) and latitude, longitude and magnitude against time (right panels). It can be seen that the magnitude seems to follow an exponential distribution and time stationary assessment is shown with the N_t vs time plot. The next figure, Fig.5, shows that the convergence of the estimated parameters is obtained after 11 steps. The estimated background rate and other related plots are included in Fig.6. To further assess how well ETAS model fits the data, one can use a function in the package to visualize the residuals as in Fig.7. The model appears to be unable to explain the variation of the time variable in the data as the raw temporal residuals plot and the Q-Q plot show that the temporal residuals are not uniform. The Kolmogorov-Smirnov test for goodness-of-fit is carried out and yields a significant p-values, which confirms the above. There is also clear evidence that this model does not seem to adequately explains the variation in spatial data. The intensity rate of the Bay area and the Greater Los Angeles, where the power outages happen the most

frequently, are consistently underestimated. These unpleasant results point out some shortages in the input data and will be discussed in section 5.

5 Discussion

As discussed in Section 4.1, although scaling data to [0,1] range and use it for kernel smoothing may be convenient for both computing and visualizing tasks, it should be used with cautions. At least, conclusions should not be made solely based on one density plot. It is recommended to experiment different bandwidths and also compare with the map of points with their original geographic locations. In section 4.2, we obtained estimates whose true values can actually be zero (large standard errors). It appears that with this model, we do not have any background rate. By using ETAS model, it was assumed that each power outage can trigger other outages with certain magnitude. This is obviously not true in general. Furthermore, the dataset does not have a feature that can be considered as magnitude. Here, the number of affected customers of each outage was assumed to be its magnitude. With all of these poor assumptions, surprisingly, most of the estimates generated by this model have relatively small standard errors. However, by examining the residual plots, we can conclude that this model still does not perform that well on our dataset. The raw temporal residuals do not follow the uniform distribution of range (0,1). Similarly, the raw spatial residuals indicates that the model consistently under-fits the areas where outages happened the most. Even though the models examined in this report performs poorly, they still give us preliminary insights about our data. Moreover, this report examines several point process analysis techniques and serves as a learning tool rather than to yield some significant results. Given that enough data can be collected and more advanced point process analyzing tools are available in the future, unplanned outages is still an interesting topic one can study about to offer helpful results in multiple ways.

Appendix

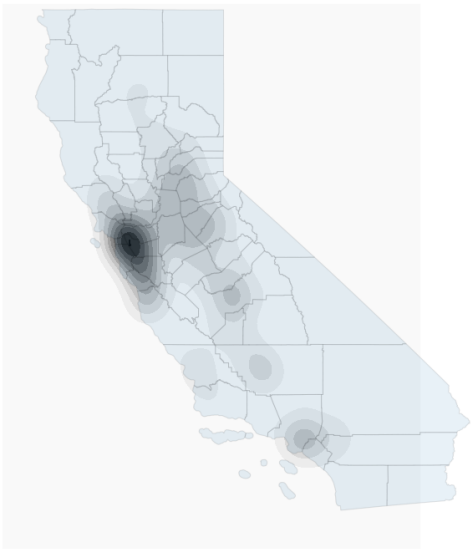

Figure 1. Density Plot Using Kernel Smoothing

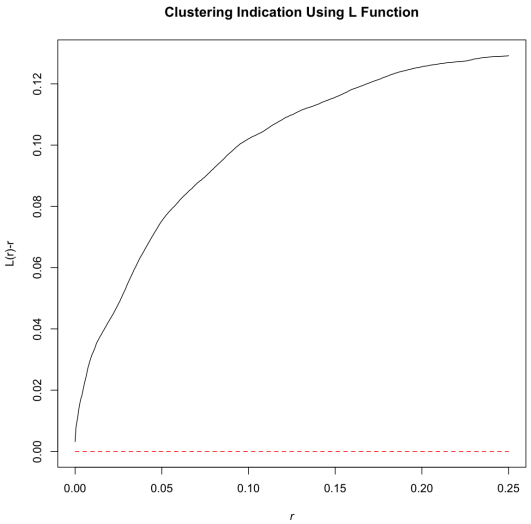

Figure 2.Clustering Indication Using L Function

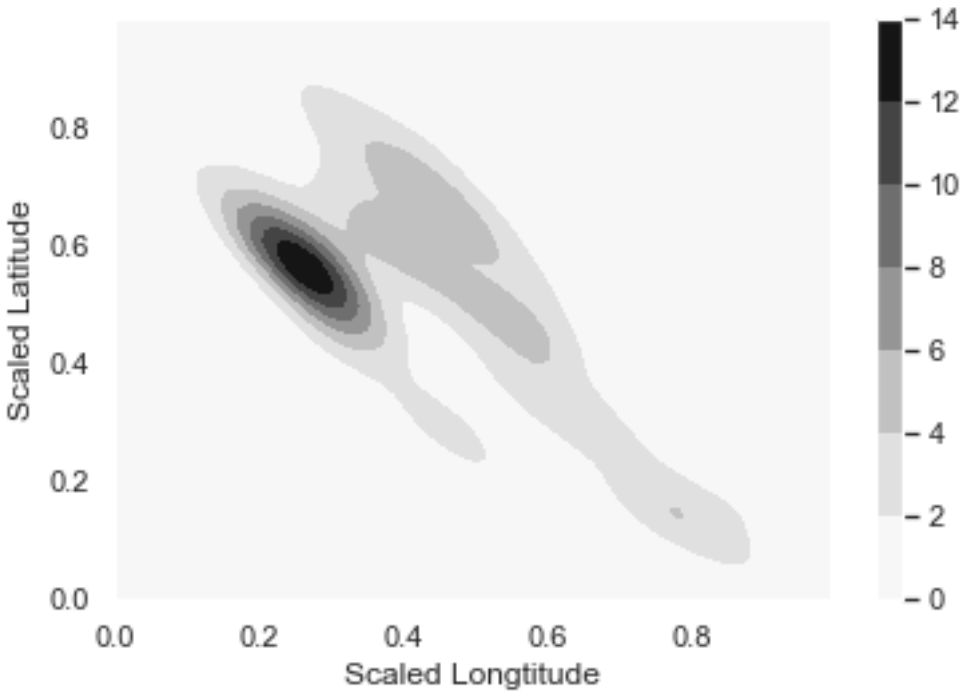

Figure 3. Density Plot Using Kernel Smoothing with scaled data.

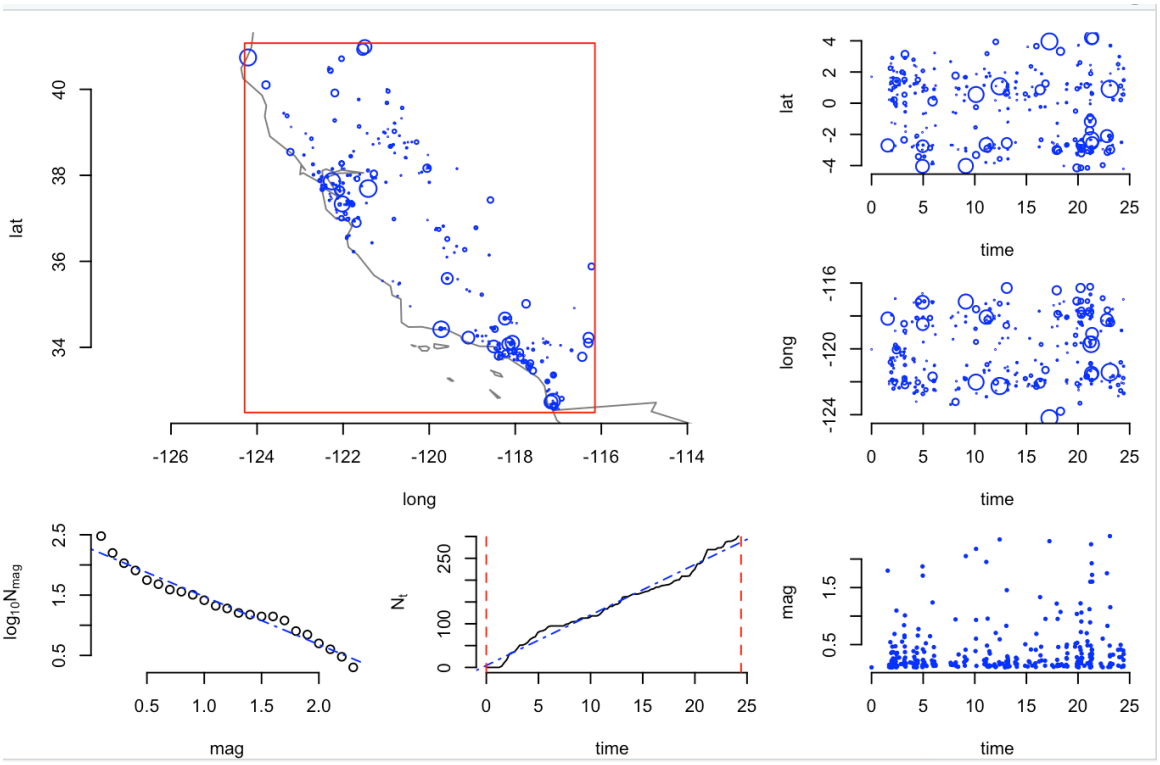

Figure 4. Descriptive plots from ETAS package.

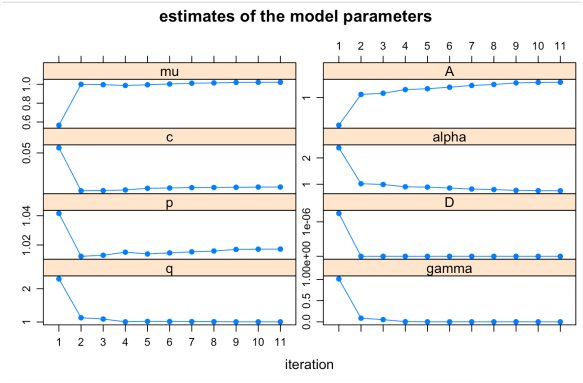

Figure 5. Convergence of ETAS estimated parameters.

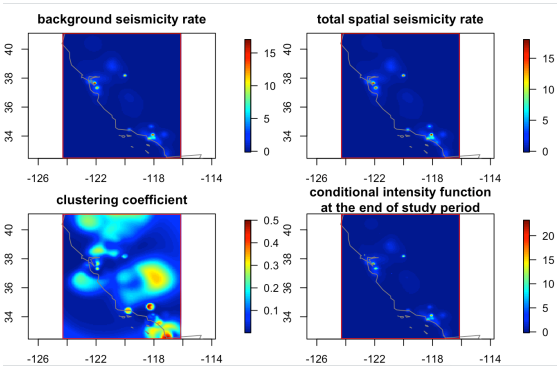

Figure 6. Plots of estimates.

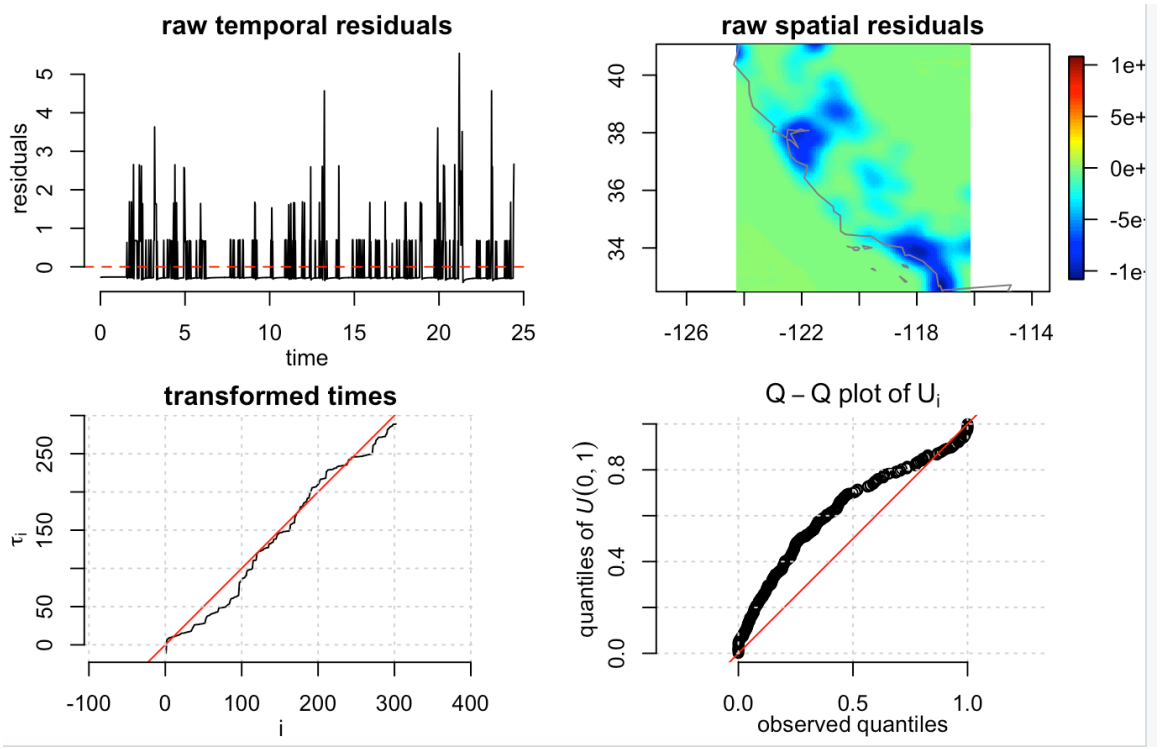

Figure 7. Residual plots from ETAS package.

References

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

Ogata Y (1998). "Space-Time Point-Process Models for Earthquake Occurrences." The Annals of the Institute of Statistical Mathematics, 50(2), 379–402.