[redacted]

A Point Process Analysis of San Francisco Property Fires

Introduction

Property fires occur in the United States at an average rate of once every second (NFPA 2016). When a property fire occurs, it has the potential to grow into a highly volatile and damaging event. As of 2014, house fires caused an estimated 12 billion dollars in property losses (USFA 2016). A fire's impact can also lead to extreme economic hardship due to temporary loss of a home and in more rare cases, death from the fire itself.

Firefighters, researchers, and public officials make great efforts to improve response times to fires and minimize the impact of urban fires. However, unlike wildfires, addressing the cause of property fires can be a difficult issue as they can be started from a great variety of ways ranging from kitchen accidents to damaged electrical wiring. As society strives to address the issues property fires cause, new insights may come from statistical analysis.

Many individuals and companies could benefit from a better understanding of the causes of property fires. If a home is in a neighborhood with a high propensity for property fires, a city could allocate more firefighting resources to that area to save homes and lives. A home buyer may also interested in understanding the risks of fire in a neighborhood as it will impact his/her decision purchase a home or apartment. Insurance companies can also benefit from improving their understanding of property fires. A home insurance company could charge a more competitive rate for homes in areas that are believed to be less risky for property fires.

There is a wide body of research surrounding property fires and their causes, but no research has yet been done from a point process perspective. The goal of this project is to conduct a point process analysis of San Francisco fire data to provide a probabilistic characterization of property fires in space and time. If this analysis provides meaningful and interpretable insights, it could lead to useful tools for city governments, home buyers, and insurance companies to save lives, property, and money.

Data and Pre-processing

The data was obtained from the San Francisco city website and contained all fire department calls during the 2016 calendar year. The raw data contained 31,554 calls for fire department services. A clear majority of the data contained calls for fire department service that did not pertain to an actual fire, such as rescue operations and medical/EMT services (5,300 calls) or false/malfunctioning alarms which were subsequently canceled (9,000 calls). These types of incidents were removed, leaving 1,053 instances of calls for actual fires that required extinguishment or assistance.

Data was also filtered to remove fires that resulted in minor damages, where minor damages are defined as less than $10,000. The threshold of $10,000 was appropriate for separating minor and major damages, as it is an order of magnitude above most home insurance plan deductibles (typically $500 - $1,000) so damages above this amount could have a significant impact on insurance pricing schemes. Once those data points were removed, 145 points remained for property fire calls which had estimated damages at or above $10,000.

A final data cleaning consideration was how to treat the locations of the points. The original data were in Latitude and Longitude format. This poses interpretation problems, because the background rate coefficients that are calculated in fitting a point process will correspond to the Latitude/Longitude window of San Francisco which is roughly (-122.5, -122.4) x (37.7, 37.8). It is an extra mental step to convert and understand the fit of the background rate in a meaningful fashion with the traditional coordinates. To allow for easier analysis, the locations of the data were normalized to fit on a (0,1) x (0,1) unit square. This conversion is much easier to interpret because it is quick to see that (0.5, 0.5) is the centroid of the city, and certain regions such as Northwest or Southeast San Francisco can be identified by comparing normalized locations to this centroid. The point process analysis for this project was conducted on the 145 points with major fire damage at their normalized locations.

Analysis – Inhomogeneous Poisson

The first avenue of research on this data was to determine if there was any evidence of clustering. To achieve that, the data was first visualized in a kernel density plot which is shown in Figure 1. The density plot seems to indicate some clustering around districts such as Hunter's Point, Potrero Hill, and the Downtown. A map of San Francisco's neighborhoods is shown in Figure 2 for reference. With only 145 points, based on this density plot it is difficult to tell if these fires occurred by a homogenous Poisson process or some other inhomogeneous process.

To assess if the data were generated by a homogeneous Poisson process, the fitted F, G, J, K, and L functions were plotted alongside the theoretical fit of a homogeneous Poisson process. Figure 3 shows the F, G, and J functions. The F function, or empty space function, seems to be well below the homogeneous Poisson F function which indicates an overall lower than expected spacing of points in the region of study. The G function, or nearest neighbor function, is well above the homogeneous Poisson G function which indicates higher than expected density around neighboring points. Both functions, along with the J function, indicate evidence of clustering that would not normally be seen in a homogeneous Poisson process.

The K and L functions are shown in Figure 4 and further support evidence of clustering. Both functions provide evidence of clustering, but the L function shows this particularly well. The simulated and theoretical 95% confidence bounds for a homogeneous Poisson process are plotted, as well as a line at 0 for the perfect homogeneous Poisson fit. The L function is well above the 95% bounds for all distances shown on the plot, which indicates enough evidence to reject the hypothesis of a homogeneous process in favor of a more clustered inhomogeneous process.

After evidence of clustering was established, an inhomogeneous Poisson model was fit to the data by maximum likelihood estimation. The equation of the model is $\lambda(z|z_1, \dots, z_k) = \mu + \alpha x + \beta y + \gamma \sum_{i=1}^{k} \frac{a_1 e^{-a_1 D(z_i ,z)}}{2\pi D(z_i ,z)}$ where $z = (x, y)$. The results of the fit are shown in Table 1 and Figure 5. Table 1 contains the parameter estimates and standard errors, and Figure 5 shows the background rate and total density rate (lambda) for the space. The density plot of the model lines up relatively well with the density of the data, but the standard errors, especially for the parameters of the background rate, are extremely high. Because of the high parameter variance this model does not appear to be very suitable for the data and a different model was fitted.

The next model that was fitted to the data was a marked inhomogeneous Poisson model of the form $\lambda(s|s_1, \dots, s_k) = \mu + \alpha x + \beta y + \gamma \sum_{i=1}^{k} a_1 e^{-a_1 z_i D(s_i ,s)}; s = (x, y),\ z = damage\ (\$1{,}000s)$. The results of the fit are shown in Table 2 and Figure 6. Table 2 contains the parameter estimates and standard errors, and Figure 6 shows the background rate and total density rate (lambda) for the space. Figure 6 also contains the data locations with point sizes scaled to the log of the data to aid in visualization. The density plot of the model is slightly different than the previous model, but the fit is still similar to the initial density of the data. The standard errors of the model decreased but are still relatively large compared to the overall average rate. Just like the prior fit, due to the high variance this model does not appear to be very suitable for the data and a more complex model accounting for time was fitted.

Analysis – Hawkes Model

The final model that was fitted in this analysis was a basic Hawkes model, which takes the general form $\lambda(\mathrm{t,x,y}) = \mu(\mathrm{x,y}) + \mathrm{k}\int_{t'<t} g(t-t',\ x-x',\ y-y')dN(t',\ x',\ y')$. The results of the fit are shown in Table 3 where α and β are parameters pertaining to the triggering density. It can be seen in Table 3 that the standard errors around the parameter estimates are much smaller than before,

indicating that the Hawkes model may provide the best fit. However, the value of the triggering function is approximately 0.8, which seems highly suspicious for fires. A triggering function of 0.8 would imply that on average, an additional 4 points are generated from each background fire. This seems highly unlikely unless there is a massive arson problem in San Francisco or some other significant causal linkage relating the appearances of these fires.

Despite the concerns over the triggering function, the model was evaluated further using superthinning. The original points and superthinned points are displayed in Figure 7. The superthinned points appear to follow a homogeneous Poisson process, indicating that the Hawkes model was a good fit. The superthinned points were also evaluated using the K and L functions as shown in Figure 8. The K and L functions are both well within the 95% bounds of a homogeneous Poisson which provides additional evidence that the Hawkes model was appropriate in describing the data.

Conclusion and Next Steps

The two inhomogeneous Poisson models that were fitted showed similar results, but overall the standard errors of the parameter estimates were too large for any substantive interpretations as to whether the fitted model accurately describes the data. The Hawkes model fit the data very well as indicated by the superthinning analysis, but the high triggering value of 0.8 is very suspicious and perhaps too high to characterize property fire data. Adding more data from additional years such as 2014 and 2015 may provide lower variance estimates for the inhomogeneous models, and may also lead to a smaller triggering value in the Hawkes model. With more time an exploration of these models with additional data could provide more substantive results.

In addition to gathering more data, adding covariates to the model could also improve or change the fit. In the initial density plot of the data, the three areas of high density (Hunter's Point, Potrero Hill, Downtown) are all areas of high crime, population density and poverty. This suggests that more poor, dense, and high-crime areas tend to experience more fires and would be worthwhile covariates to add to the model. Other potential covariates could include neighborhood or zip code, median property value, dates of home construction, and distance from the nearest fire station. A final area of further exploration would be expanding the analysis to other cities, which may indicate how generalizable the models are to other urban areas throughout the country.

References

National Fire Protection Association (NFPA). *The Consequences of Fire.* 2016. Web. 7 June 2017.

United States Fire Administration (USFA). *U.S. Fire Statistics*. 2016 Web. 7 June 2017.

Figures

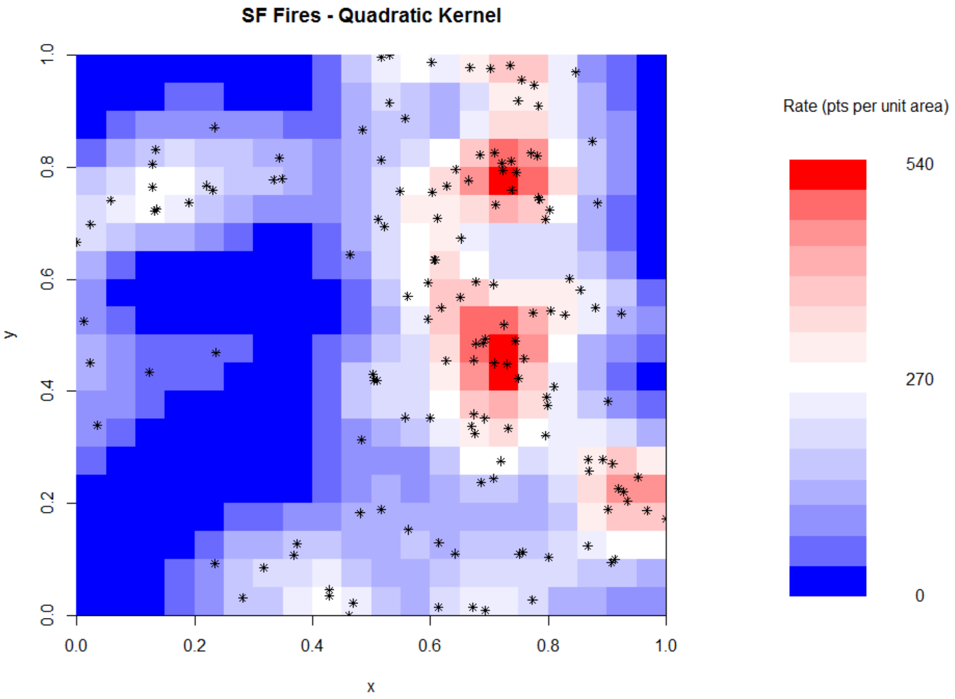

Figure 1 – Initial kernel density plot of significant property fires in San Francisco

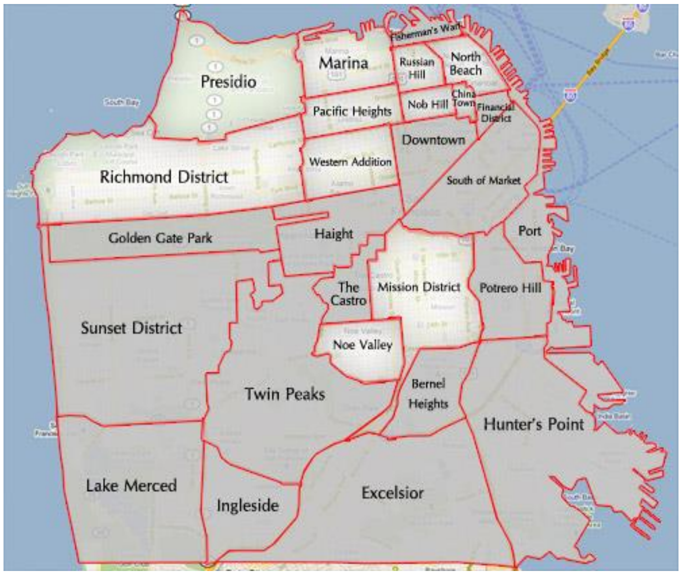

Figure 2 – San Francisco neighborhood map

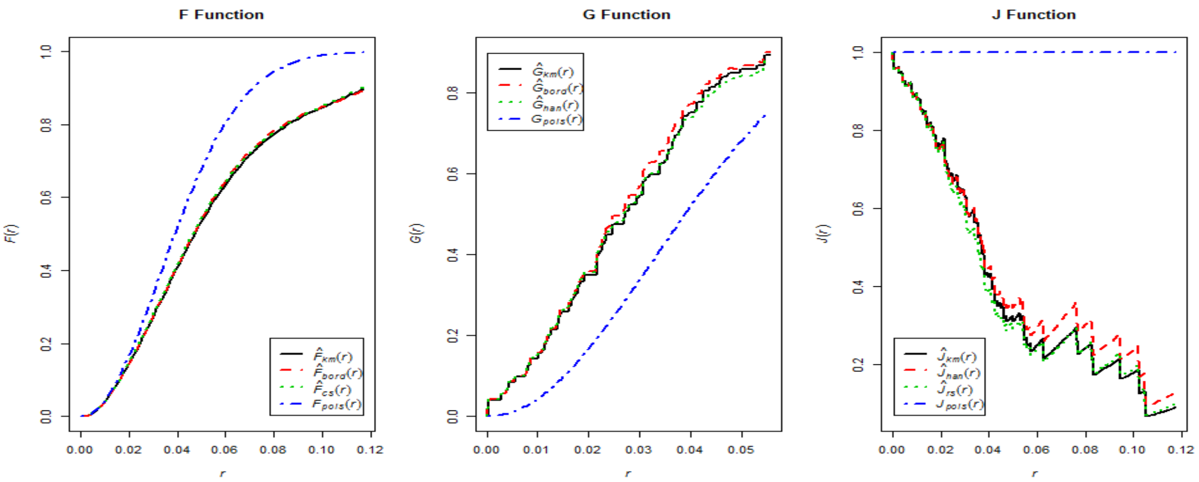

Figure 3 – F, G, and J functions for the data

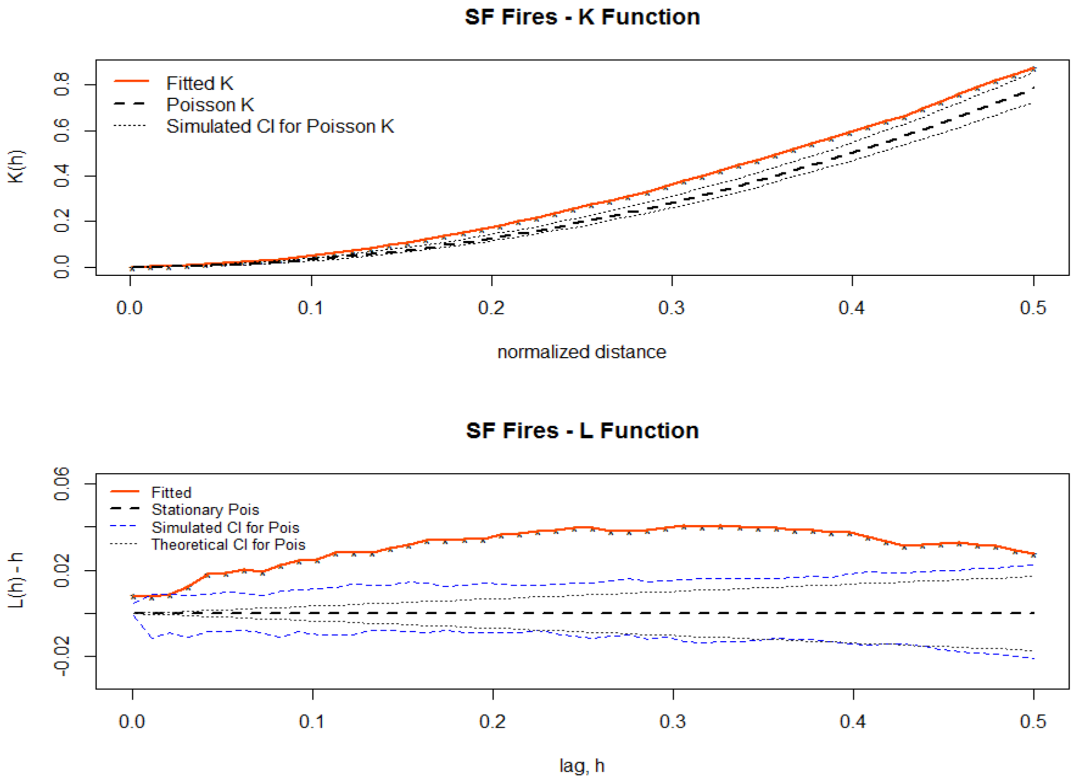

Figure 4 – K and L function of the data with 95% confidence bounds for a homogeneous Poisson

Parameter	μ	α	β	γ	a1
Estimate	40.0677	46.3967	-25.5232	0.6668	22.9478
SE	23.5418	40.2591	35.5214	0.1382	5.5774

Table 1 – Parameter estimate and SE's for inhomogeneous Poisson model

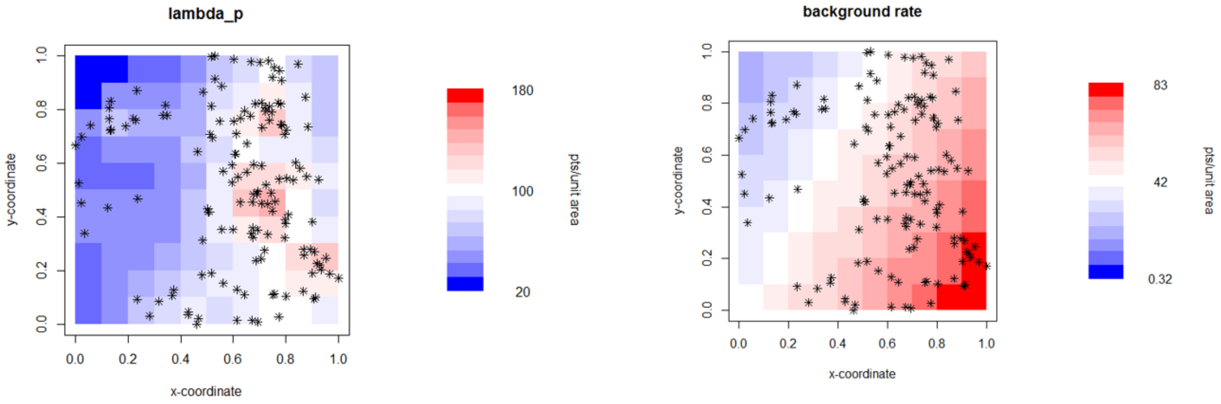

Figure 5 – Fitted inhomogeneous lambda (left) and fitted background rate (right) with data points

Parameter	μ	α	β	γ	a1
Estimate	9.6308	8.3263	-5.4937	0.8870	0.0949
SE	10.8862	12.1349	15.7862	0.1033	0.0136

Table 2 – Parameters and SE's of fitted inhomogeneous marked Poisson process

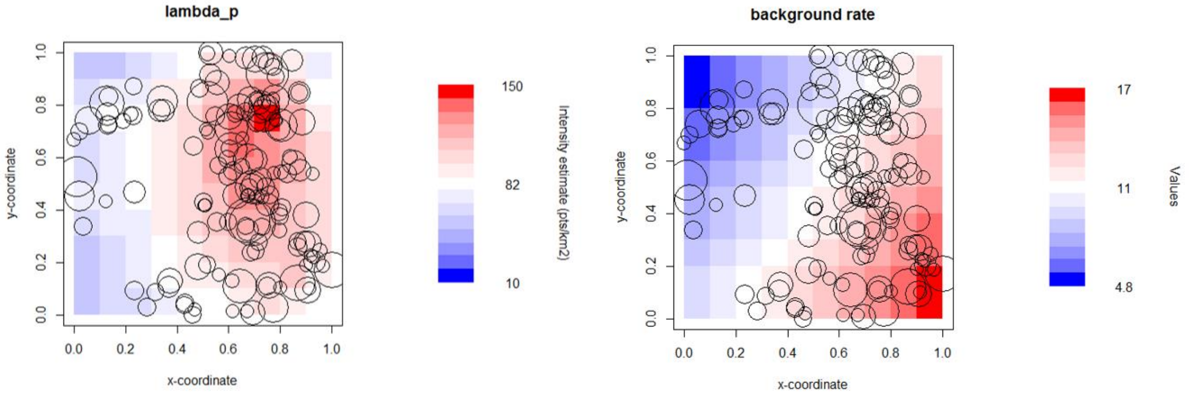

Figure 6 – Fitted marked lambda (left) and marked background rate with scaled data point sizes

Parameter	μ	K	α	β
Estimate	0.0180	0.8187	29.8745	0.0345
SE	0.0078	0.0807	5.3949	0.0068

Table 3 – Fitted Hawkes model parameters and SE's

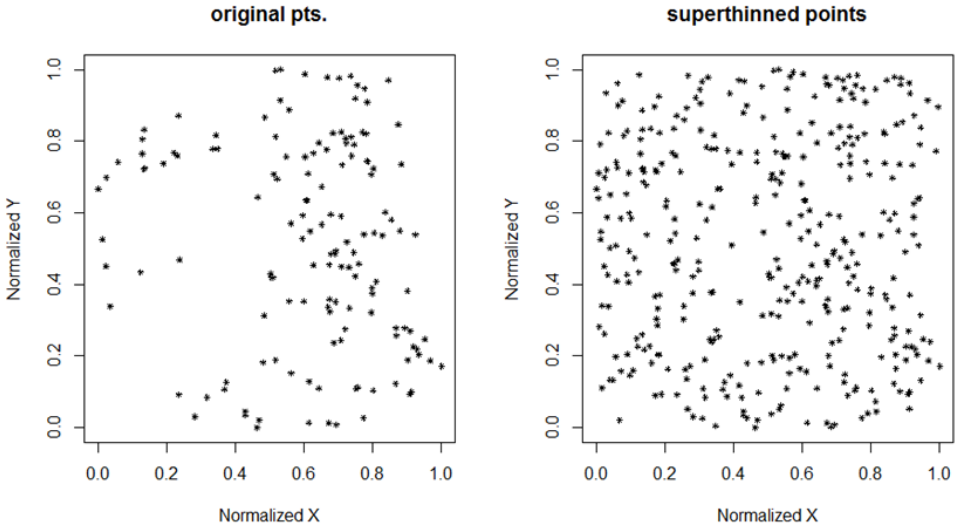

Figure 7 – Original points and superthinned points using the Hawkes model

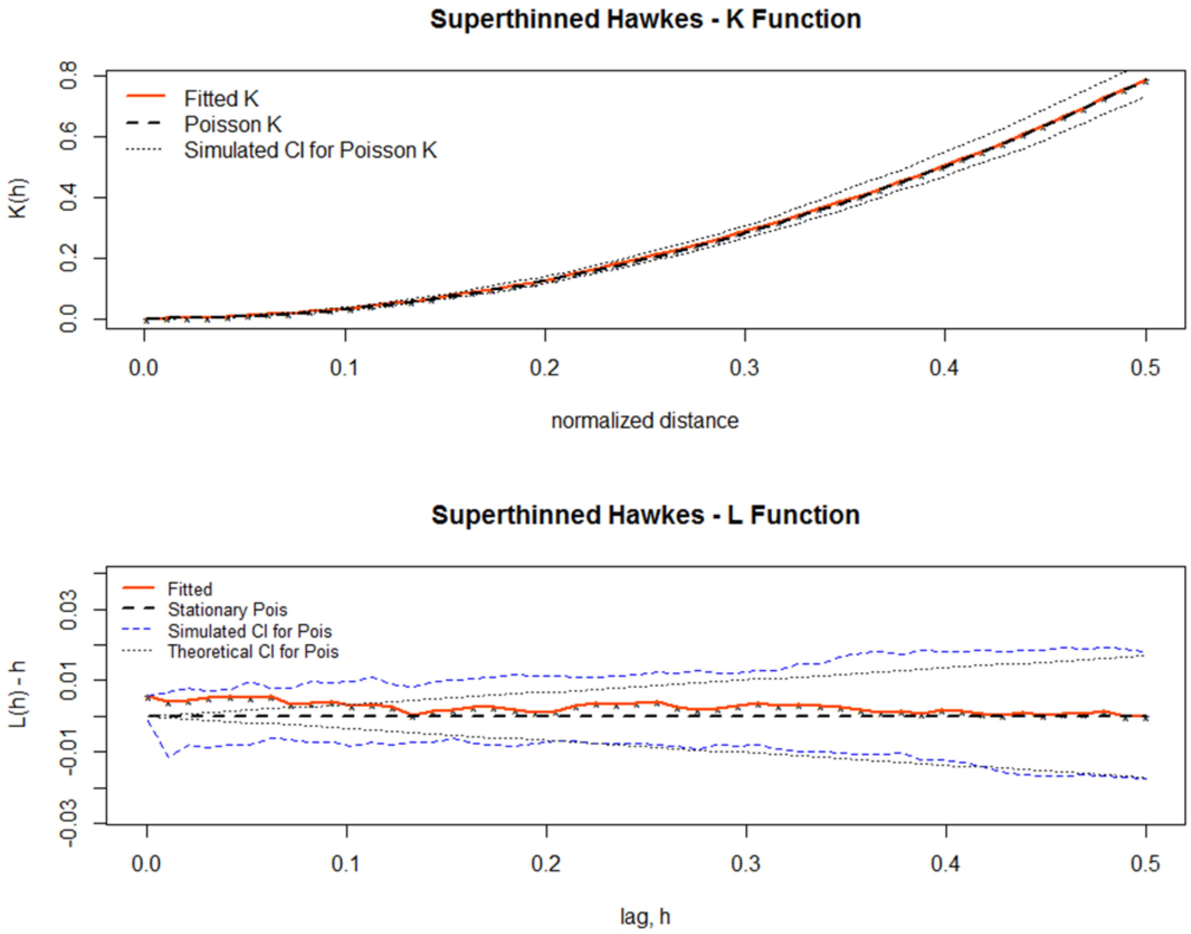

Figure 8 – K and L functions for the superthinned points under Hawkes process