Certified H-1B Visa Locations in Southern California

[redacted]

Abstract

This project used spatial statistics methods and techniques to study the geographical locations of certified H-1B visa cases in the Southern California area in 2016. The main goal of the study is to analyze the clusters of the locations of the events, as well as fitting varied spatial statistics models to further analyze the data. The findings revealed significant clustering in the locations; the fitness of different statistical models were also demonstrated.

Introduction

H-1B Visa Background

The H-1B visa is a non-immigrant visa in the United States. It allows U.S. employers to temporarily employ foreign workers in specialty occupations. The visa provides a foreign worker a duration of staying and working in the United States up to 6 years. H-1B work-authorization is strictly limited to employment by the sponsoring employer.

The procedure of obtaining a H-1B visa for each individual involves being sponsored by the employer, filing for H-1B visa petition, and being approved by the U.S. Citizenship and Immigration Service. Due to the large amount of population applying for H-1B visa in the past decade, a lottery system was implemented since 2008. In 2016, there were over 230,000 applicants for the H-1B. Currently, processing for all H-1B visa petitions is temporarily suspended and the U.S. government is working on reforming the H-1B visa system.

Dataset

The data was originally collected by The Office of Foreign Labor Certification (OFLC), and it included all the H-1B petition cases in 2011-2016, in total 3002458 observations. The dataset provided information such as the job title of the individual, the location of the worksite (longitude and latitude), the employer name, the case status, and so on. Due to the nature of H-1B visa application procedure, the case status has four different levels: "certified", "denied", "withdrawn", and "certified-withdrawn". For the purpose of this project, only "certified" cases filed in 2016 were examined. Since there was no exact date or time provided for each petition case, the data was purely spatial. As shown in Plot 1-1, each point stands for one certified H-1B visa case in 2016, and the location of each point represents the location of the worksite of the individual. As the points roughly formed the shape of the United States, it is obvious to observe clustering in the major cities throughout the U.S. continent.

Due to the area of interest for this project, only the cases in the Southern California area was examined, as shown in Plot 1-2.

Analysis and Results

Spatial Clustering

In order to study the overall clustering of the data, Kernel Smoothing and K,L Functions were performed.

Kernel Smoothing

As shown in Plot 2-1, obvious clustering can be observed. The part in the plot with the highest clustering rate represents the Los Angeles area, and it can also be seen that the regions surrounding Los Angeles also have relatively higher rates of clustering. Another strongly clustered region can be found to the lower right of Los Angeles, which represents the location of San Diego.

The above findings were not a surprise because Los Angeles and San Diego are the two most important cities in providing employment opportunities in the Southern California area. In other words, foreign employees would have a higher chance to obtain H-1B visas through working in such areas.

K,L Function

To further examine the significance of the clustering observed as above, K and L functions were performed (Plot 2-2). The red dashed lines in the plot represents the theoretical stationary Poisson process, and the black dashed lines are the confidence bounds obtained through simulation. It can be seen from the plot that for both K and L function, the fitted lines (solid black lines) are very far out of the confidence bounds, which implies that the clustering in the data was very significant.

Model Fitting

Pseudo-likelihood Model

The first model fitted to the data was a Pseudo-likelihood model using $\lambda(z|z_{1,\ldots,}z_k) = \mu + \alpha x + \beta y + \gamma \sum_{i=1}^{k} a_1 e^{-a_1 D(z_i,z)}$, where $z = (x, y)$, and D means distance. The parameter estimates and the corresponding standard errors are shown in Table-1. The plots of the fitted background rate and λ are shown in Plot 2-3. From Table-1, it can be seen that the only statistically significant parameters fitted with this model were γ and a_1, the two interaction parameters. Also, γ has a positive value, confirming that there is indeed clustering in the data. The background rate plot seems not capturing the nature of the data very well. The plot of the fitted λ shows a better fit since it looks very similar to the Kernel Smoothing plot (Plot 2-1), which can be explained by the fact that the fitted λ depends on the two interaction parameters γ and a_1, which are the better fitted parameters with this model. Therefore, it can be concluded that this Pseudo-likelihood model is not a good fit for the data although it is able to capture the interactions between the points.

Strauss Model

Several different Strauss models were also fitted, but the results were not satisfying. For Strauss models, the interaction parameter gamma must be less than or equal to 1 to be valid. However, for this dataset, regardless what interaction radius r was used in the model, the fitted interaction parameter gamma was always greater than 1, which is out of the bound of the model, implying that the Strauss models are not valid choices for this dataset. Table 2-4 shows the output from one of the Strauss models fitted with $r = 0.05$, and it can be seen that the gamma is indeed out of bound.

Poisson Model

From Plot 2-2, it can be seen that the point process in the data was not following a stationary Poisson process. Therefore, when fitting Poisson models, only non-stationary models were used.

Model 1: $\lambda = exp(\alpha + \beta\sqrt{(x^2 + y^2)})$

The first Poisson model fitted has the conditional intensity calculated as above. Table 2-5 shows the parameter estimates with this model and it can be seen that both coefficients are statistically significant. However, the fitted conditional intensity is not a good reflection of the actual data; as shown in Plot 2-6, the brightest colored area (right upper corner) does not actually have a high intensity of points.

Superthinning was then conducted to check if the model was a good fit. In Plot 2-7, the black points stand for the superthinned points while the red points are the original points from the data. The plot shows some big gaps in the right upper corner of the frame, meaning that the conditional intensities are overestimated in those regions, and the model is not a perfect fit. In order to confirm such findings, K and L functions were performed on the model after superthinning (Plot 2-8). It can be seen from Plot 2-8 that there is still clustering even after conducting superthinng, and it can be concluded that this model did not fit well.

Model 2: $\lambda = exp(\alpha_1 + \alpha_2 x + \alpha_3 y + \alpha_4 x^2 + \alpha_5 xy + \alpha_6 y^2)$

The second Poisson model fitted has the conditional intensity calculated as above. Table 2-9 shows the parameter estimates with this model and it can be seen that all the coefficients are statistically significant. The fitted conditional intensity with this model (Plot 2-10) seems much closer to the actual data; Plot 2-10 accurately captured the area with the highest rate of clustering, Los Angeles, and it also highlighted the secondary clustering region in the San Diego area.

Superthinning was then conducted to check if the model was a good fit, and the results were not very satisfying (Plot 2-11). It can be seen from Plot 2-11 that there is a big gap surrounding the original points, which means that the model overestimated the conditional intensity in this area. K and L function were then performed on the supperthinned points to further examine the model (Plot 2-12). In Plot 2-12, the fitted K function line almost overlaps with the theoretical stationary Poisson process line; although the fitted L function line still suggests clustering, the scale of the y-coordinate, approximately 0.05, of this plot is much smaller than the scale of the L function plot of the previous Poisson model, which is approximately 0.12. Therefore, it can be concluded that although this Poisson model is still not an ideal fit, but it is a better model than the previous one.

Conclusion and Future Improvements

In this study, the locations of the certified H-1B visa cases in Southern California were analyzed with spatial statistics methods and models. It can be concluded that the points are significantly clustered, with Los Angeles and San Diego being the areas with the strongest intensity rates. It was also found that neither Strauss models or Poisson models are the ideal choices for such dataset, very likely due to the lack of homogeneity.

As for potential improvement, different spatial models that could handle heavily clustered data should be used. Including relevant information as covariates, such as the city sizes, population sizes, or income levels in the Southern California area, can also potentially improve the results.

Plots and Tables

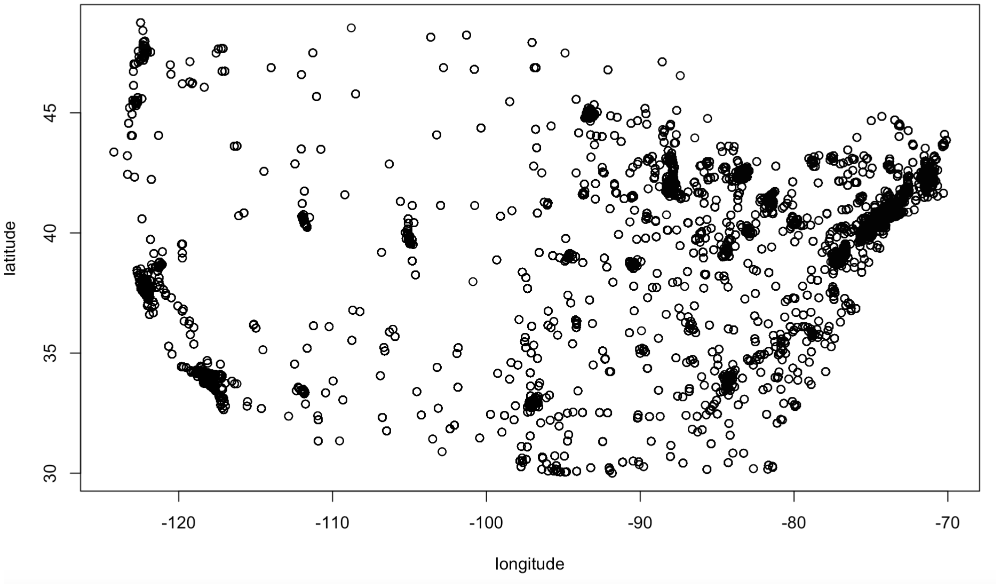


Figure 1: Plot 1-1

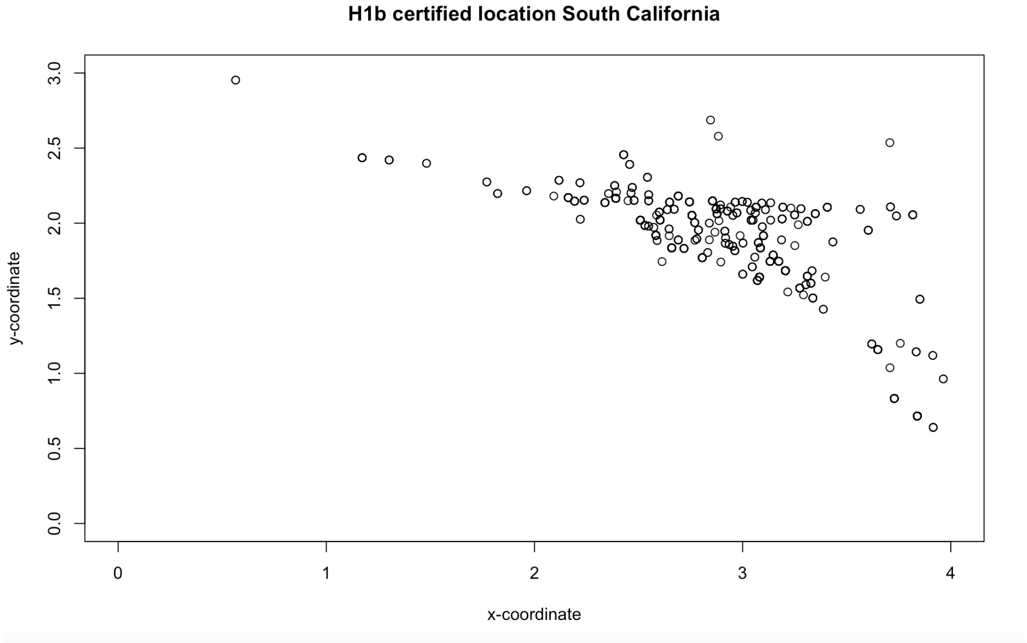


Figure 2: Plot 1-2

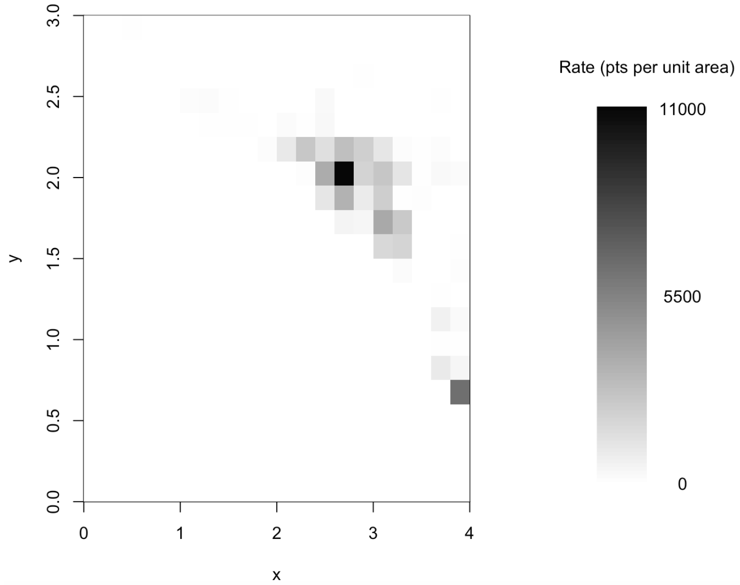


Figure 3: Plot 2-1

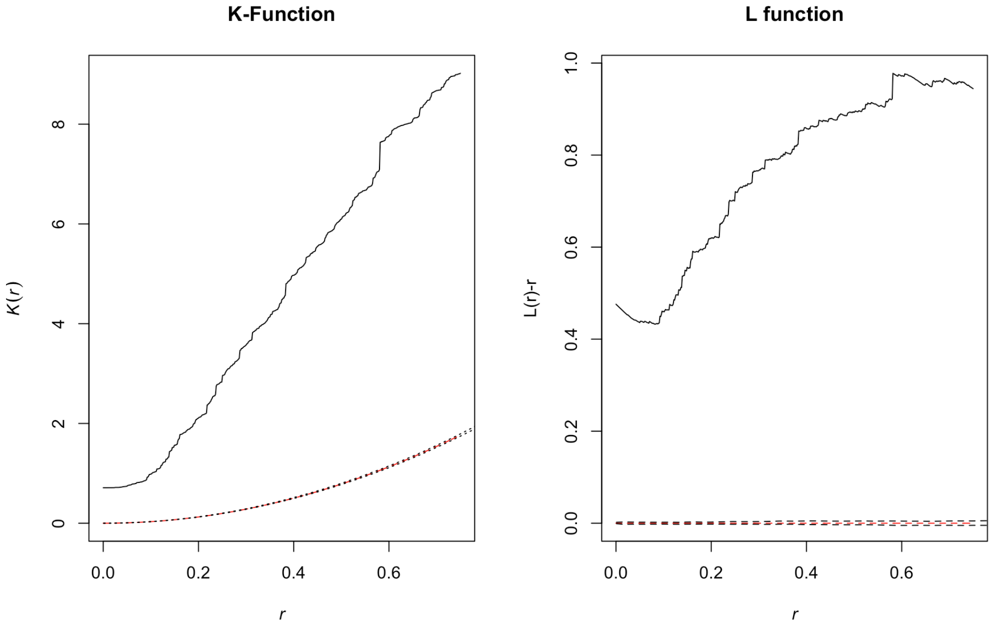


Figure 4: Plot 2-2

Parameter	μ	α	β	γ	a_1
Estimate	22.875	-4.062	-18.803	0.958	24.809
SE	242.406	337.149	224.329	0.053	12.774

Table 1: Table 1

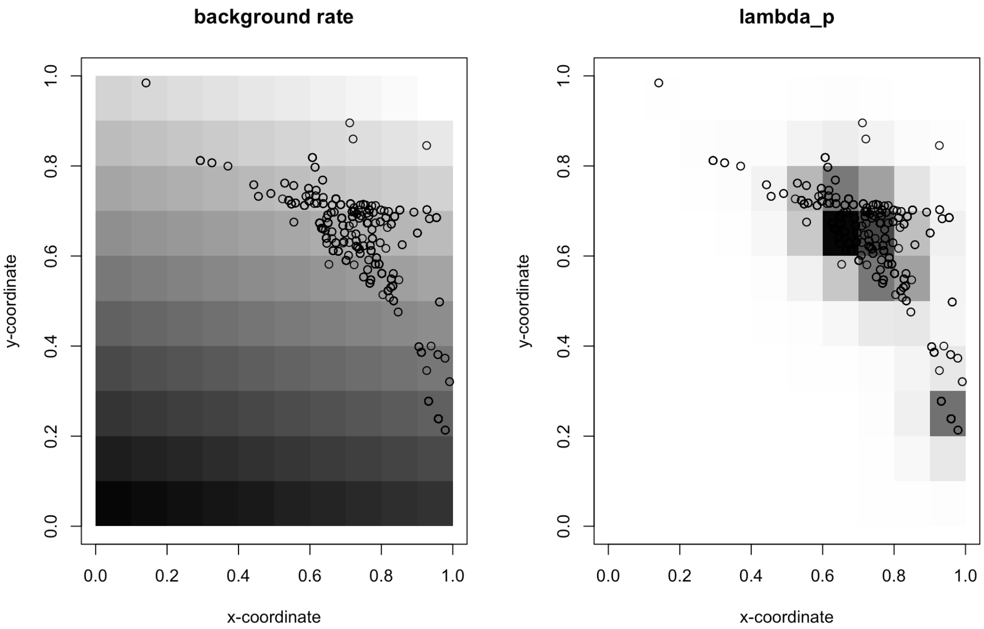
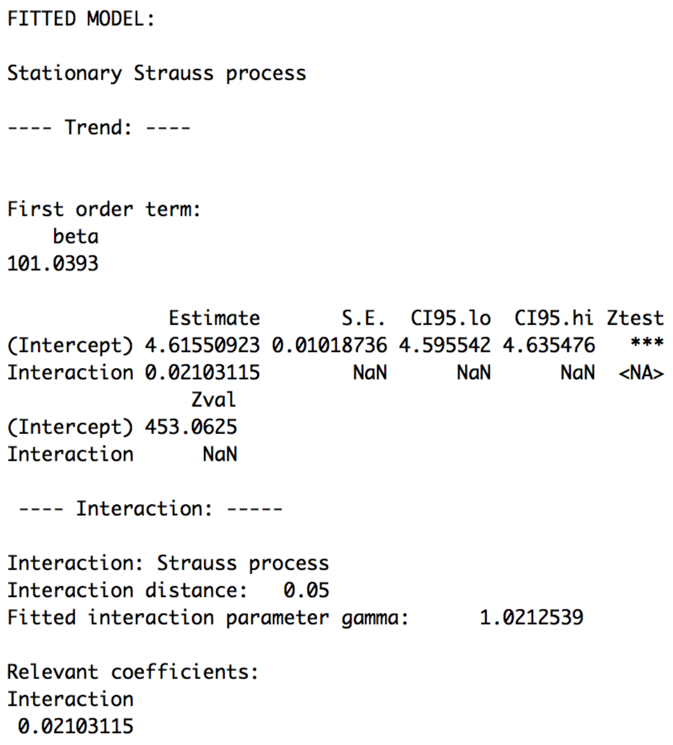
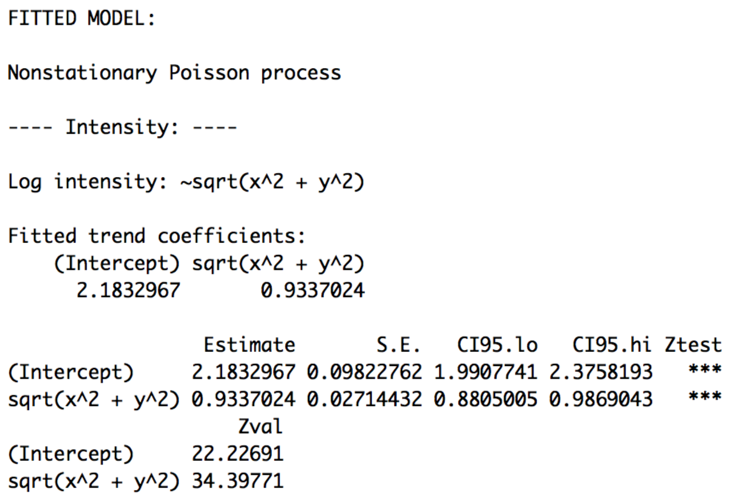


Figure 5: Plot 2-3



```
FITTED MODEL:

Stationary Strauss process

---- Trend: ----


First order term:
    beta
101.0393

              Estimate       S.E.  CI95.lo  CI95.hi Ztest
(Intercept) 4.61550923 0.01018736 4.595542 4.635476   ***
Interaction 0.02103115        NaN      NaN      NaN  <NA>
                Zval
(Intercept) 453.0625
Interaction      NaN

 ---- Interaction: -----

Interaction: Strauss process
Interaction distance:   0.05
Fitted interaction parameter gamma:      1.0212539

Relevant coefficients:
Interaction
 0.02103115
```

Figure 6: Table 2-4



```
FITTED MODEL:

Nonstationary Poisson process

---- Intensity: ----

Log intensity: ~sqrt(x^2 + y^2)

Fitted trend coefficients:
    (Intercept) sqrt(x^2 + y^2)
      2.1832967       0.9337024

                 Estimate       S.E.   CI95.lo   CI95.hi Ztest
(Intercept)     2.1832967 0.09822762 1.9907741 2.3758193   ***
sqrt(x^2 + y^2) 0.9337024 0.02714432 0.8805005 0.9869043   ***
                    Zval
(Intercept)     22.22691
sqrt(x^2 + y^2) 34.39771
```

Figure 7: Table 2-5

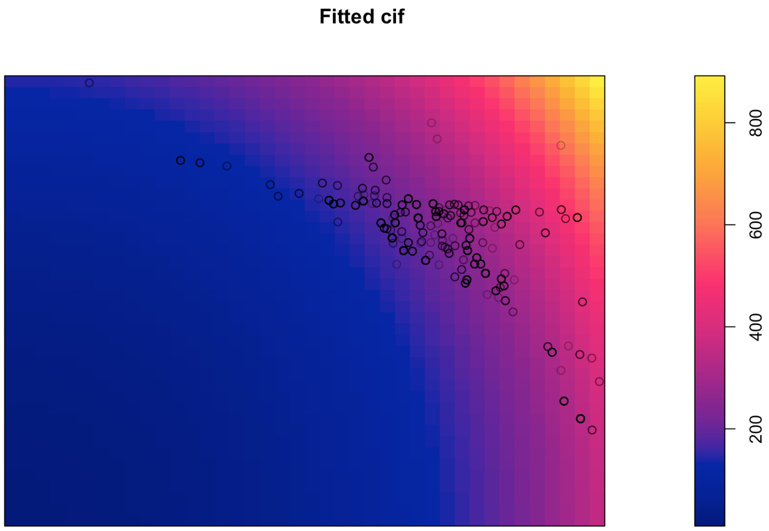


Figure 8: Plot 2-6

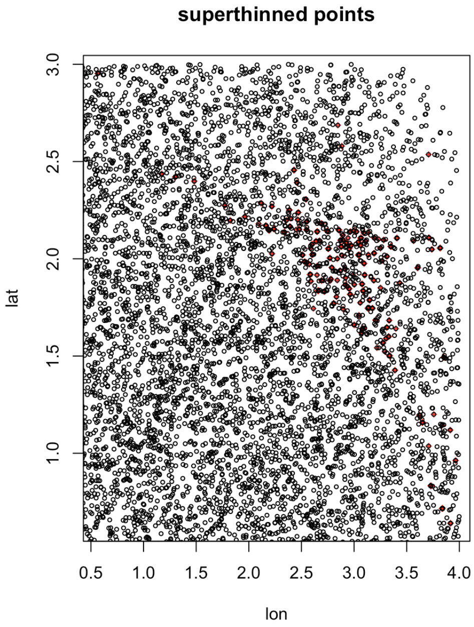


Figure 9: Plot 2-7

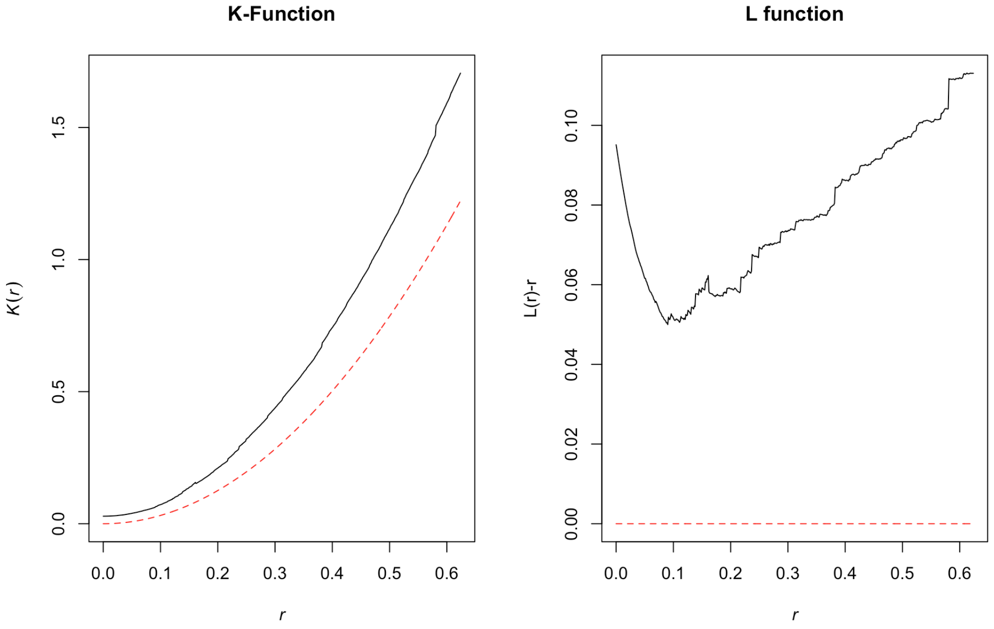


Figure 10: Plot 2-8

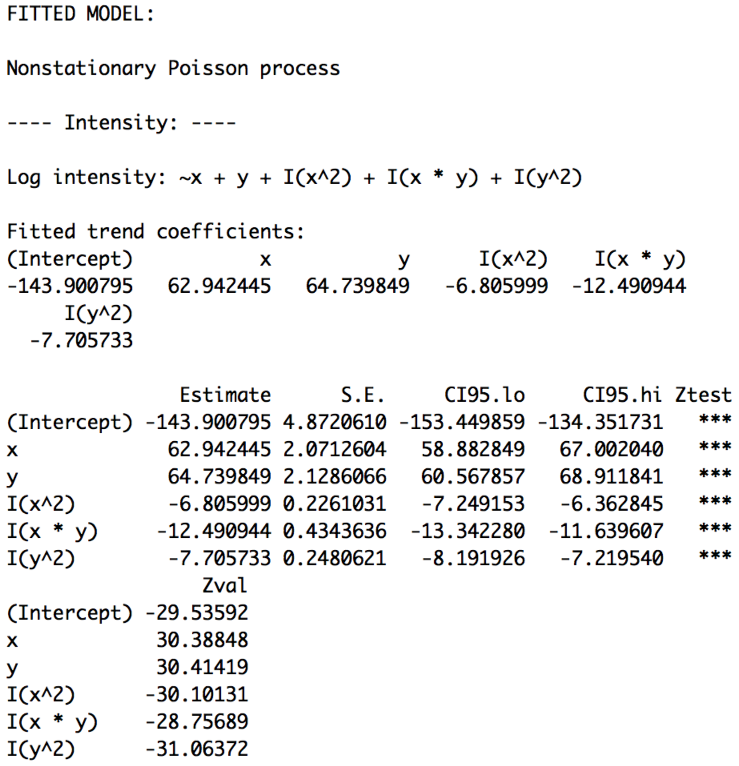

```
FITTED MODEL:

Nonstationary Poisson process

---- Intensity: ----

Log intensity: ~x + y + I(x^2) + I(x * y) + I(y^2)

Fitted trend coefficients:
(Intercept)           x           y      I(x^2)    I(x * y)
-143.900795   62.942445   64.739849   -6.805999  -12.490944
     I(y^2)
  -7.705733

               Estimate      S.E.     CI95.lo     CI95.hi Ztest
(Intercept) -143.900795 4.8720610 -153.449859 -134.351731   ***
x             62.942445 2.0712604   58.882849   67.002040   ***
y             64.739849 2.1286066   60.567857   68.911841   ***
I(x^2)        -6.805999 0.2261031   -7.249153   -6.362845   ***
I(x * y)     -12.490944 0.4343636  -13.342280  -11.639607   ***
I(y^2)        -7.705733 0.2480621   -8.191926   -7.219540   ***
                 Zval
(Intercept) -29.53592
x            30.38848
y            30.41419
I(x^2)      -30.10131
I(x * y)    -28.75689
I(y^2)      -31.06372
```

Figure 11: Table 2-9

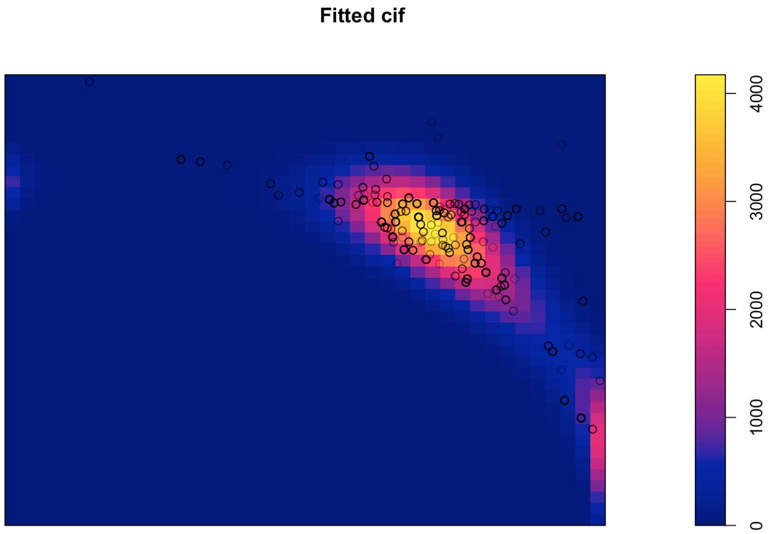


Figure 12: Plot 2-10

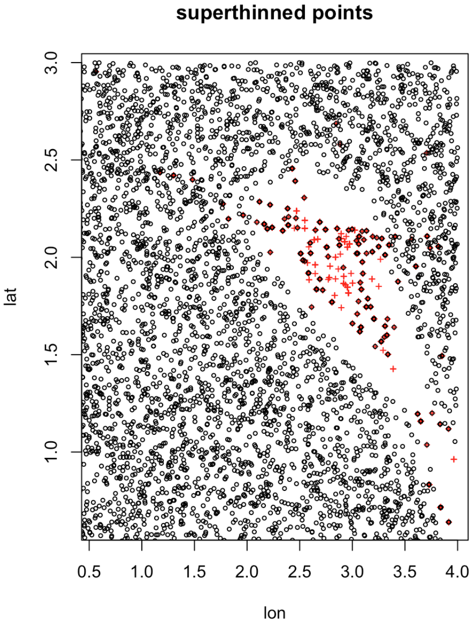


Figure 13: Plot 2-11

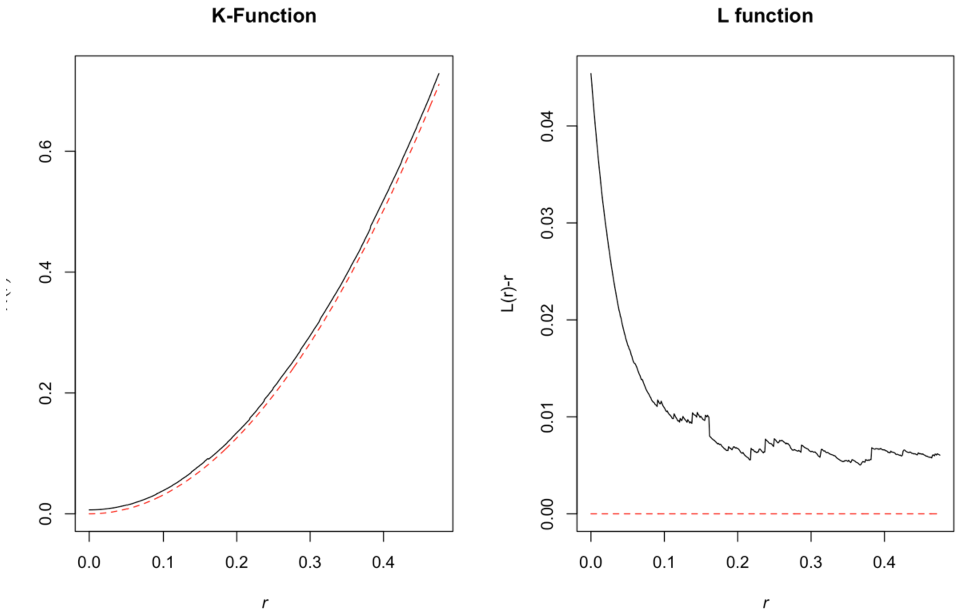


Figure 14: Plot 2-12