Dependence of Lottery Winners in Perfectia on Income: A Project for Stat 201B

Ray Luo
rayluo@ucla.edu
UCLA Neurobiology Department
Los Angeles, CA 90095-176318

March 22, 2007

1 Introduction.

The pattern of lottery winners in Perfectia appears to be distributed unevenly according to the x and y coordinates of the winner locations. We will examine whether the dependence of lottery winning on location also depends on the income of individuals who win, and whether the point process that governs the winner locations can be adequately described by a stationary Poisson process. I will begin with a summary of the data, an examination of the K, J, and other functions that tell us whether a stationary Poisson process is adequate, and various models that attempt to fit the point process, including log linear Poisson, Poisson with covariate, quadratic Poisson, and clustered linear models. I will end with a discussion of the goodness of fit of these models, which well examine via AIC, Baddeley residuals, and data thinning. I found that clustering is helpful in accounting for the data I observe, and that using income as a covariate greatly increases the models goodness-of- fit. However, better models may be possible in view of the revelation of clustering patterns in the thinned data generated from a clustered model.

2 Results.

Figure 1 shows the distribution of 2961 lottery winners in Perfectia since 2002 as x and y coordinates of where the winners lived. Notice that there appears to be some regions of empty space where no winners live, and that there appears to be few lottery winners below the fourth latitude (y). The income covariate is shown in figure 2, in which cyan indicates low income areas and progressively more maroon areas denote where higher income Perfectians live. Thus the graph is a contour plot of a 10 by 10 array of numbers indicating the level of the average income at each point. It isnt a 10 by 10 grid because its a contour plot. The

colors indicate that there are spots of high income below the fourth latitude, as well as a region near (2, 6). This may provide an explanation of the pattern of lack of winners below the fourth latitude found in figure 1.

Figure 3 shows a Kernel smoothed contour that estimates the number of points within a unit area of Perfectia as a continuous density estimate for the intensity of the point process. Here, red indicates low density, and progressively lighter colors area areas of high density. I chose a bandwidth of 1.5, because below this value, one tends to see a pattern that has a possibly spurious peak near (7, 6). I elected to have as smooth a density estimate as possible that provides a sensible summary of the main trends. The figure shows that the highest density of points appear to be at the upper right hand corner Perfectia, and that more lottery winners are found at higher latitude.

Figure 4 attempts to examine whether our simple point process can be sufficiently described using a stationary Poisson process with space-independent conditional intensity, which would provide a model without clustering and inhibition. The Ripley K function provides a measure of the expected number of points at distance r away from a given point, not including the point itself. The function that we use already account for the boundary effect. As the first plot shows, the K function always increases with r, but at the interest of showing just the major trends, and to tractably run the code, Ive only looked at K(r) up to $r = 0.3$. Since the estimated (solid) K(r) is greater than the theoretical (dotted) K(r) at $r > 0.05$, I surmise that the stationary Poisson process model may not be sufficient. If the actual K(r) is large for some r, that implies that there are more points at distance r away and below than expected, indicating the possibility of clustering below r. To see this better, we plotted in figure 5 $\sqrt{\frac{K(r)}{\pi}} - r$, which should be zero for all r, since K(r) should scale as πr^2. Indeed, we see that this L function has values above zero at all $r > 0$, indicating clustering. The L function reaches a peak at about $r = 0.1$, which provides a measure for the sizes of the clusters.

The other plots in figure 4 provide the information in a different form. The F, or empty space function, gives the area around each point before hitting its nearest neighbor. As expected, this function is smaller than expected at a given r, indicating the smaller space between points compared to a stationary Poisson process. The G function shows the distribution of nearest neighbor distances at each r, so that at a high enough r (which I see as 0.3), all nearest neighbor distances fall within r (1.0 for a cumulative distribution function). Again, higher than expected estimated G(r) at some r indicates that more points are within r than expected under a stationary Poisson process. Finally, the J function of Van Lieshout and Baddeley is given by $\frac{1-G(r)}{1-F(r)}$, which should be one for a uniformly random Poisson process, since there should be the same amount of empty space for each nearest neighbor distance. Since we see that J(r) tends to zero as r increases, we conclude that theres clustering thats not accounted for by the stationary Poisson model.

The next step is to fit a model to the data. As a first approximation, we try a log linear Poisson model with the intensity related to the coordinates as

$$\log \lambda(x, y) = \beta_0 + \beta_1 x + \beta_2 y,$$

with the trend surface and contour plot shown in figure 6. As the surfaces show, the model tries to pick up on the increase in number of winners with y (latitude) and also slightly with x (longitude). The fitted model gives an intercept $\beta_0 = 2.67$, while $\beta_1 = 0.018$ for x and $\beta_2 = 0.114$ for y. The latter βs tell us that when x (or y) is increased by one unit, the rate λ is multiplied by e^{β} unit, so that $\beta > 0$ implies that the rate increases with both x and y, although much more so with y. The contour plots show the same information in two dimensions, indicating that the model predicts a higher rate at high latitudes. The density of points is labeled in the last figure. Notice that the clustering and income information are not picked up by the model. In the latter case, the income is high at around (2, 6) and (8, 2), and generally has a nonlinear looking distribution. The fact that our model is linear in the log of the rate means that we may not be able to pick this up. Also, theres no mechanism for explaining clustering in this simplest model. It turns out the AIC is -14456 for this log linear Poisson model.

In figure 7, the fit of a log quadratic Poisson model is shown:

$$\log \lambda(x, y) = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 y^2.$$

Notice that the model tries to pick up on the pattern of high density at the upper right corner and low density at the lower left corner by having a quadratic trend surface. The AIC given two extra parameters is now -14501, which is slightly better than the linear model. However, neither the clustering nor the income covariates are taken into account, and theres little reason to suggest why the winner distribution should depend on the squared terms of each coordinate. The βs for the quadratic terms are both less than zero (-0.007 and -0.0167), so we see that the increase with x and especially y are greater at some central regime, and the quadratic terms ($e^{\beta} < 1$) refines this area.

In figure 8, a smoothed contour for a log linear Poisson model with income as a covariate is used, specified by

$$\log \lambda(x, y) = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 inc,$$

where *inc* is the income level at coordinate (x, y). As the figure shows, the income level tends to push the model from figure 6 into a more refined fit that incorporates the income distribution seen in figure 5. As expected, $\beta_3 < 0$ for income, so that as income increases (the lower latitudes), the number of lottery winners decrease. Thus, the lottery may be a mechanism for redistribution of wealth. Figure 9 shows the contour for a log quadratic Poisson model with formula

$$\log \lambda(x, y) = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 y^2 + \beta_5 inc.$$

Notice how the quadratic fit rotates the contour in a sense, allowing the income-influenced trend curve take on a more curved shape around the upper right and lower left corners. The βs for all by the x and y coordinate predictors are less than zero, as expected from individual previous models that use each of the components on its own. The AIC for the log linear covariate Poisson model is -15020; the AIC for the quadratic linear covariate Poisson model is -15029, both significant improvements over the corresponding models that dont use the

covariate. Notice, however, that the fit for the quadratic model is not that much better than the linear one.

Next, we decided to try using maximum likelihood estimation (MLE) to find parameters for the models detailed above. This gives us an additional level of flexibility in specifying the model and displaying the results. We minimize the negative log likelihood

$$\int_A \log \lambda(x,y) dN - \int_Z \lambda(x,y) dx dy$$

in each of the following cases, but using different expressions of the log likelihood depending on whether were doing a Poisson model and whether were using covariates. Because we call an iterative optimization function that looks for parameters estimates that lower the negative log likelihood, I compared the estimate that we began the iteration with and the estimate at the conclusion of the call in the case of the Poisson model with covariates. Notice that the covariate model found by MLE has more areas of twists and turns as seen in the intensity plot when compared with the simple Poisson model. With this added flexibility, we were able to fit a cluster model with covariates to the data. Here, instead of using just x, y, and income, we also included a term that penalizes large distances before seeing a point. As seen in figure 10, the cluster model with covariate has a smoother pattern than the one with the Poisson model. Also in the figure, the original starting parameter gives a terrible model. The darker cells are points of increased intensity, and the lighter cells have lower intensity. The AIC for the Poisson model estimated using MLE is similar to the one estimated by using spatstat, and is -14488. The Poisson model with covariate has AIC -14492. However, the cluster model has the best fit and the lowest AIC at -15310. This indicates that clustering likely pervades the data and the best model takes advantage of this effect.

The AIC measures the goodness-of-fit of a model to the data, but discounts the number of parameters in its comparison. It is given by

$$2p - 2\log(likelihood),$$

where p is the number of parameters in the model. Thus, increasing the likelihood and decreasing the number of parameters both have the effect of lowering the AIC. A plot of the AICs for each model is shown in the upper left corner of figure 11. In general, the models with covariates (the first two from the left) have much lower AIC than the models without covariates. Also, the cluster model is even better than the Poisson covariate models in terms of having lower AIC. Thus, in terms of trading off maximizing likelihood and minimizing number of parameters, the cluster model is the best model we have examined.

The rest of figure 11 shows the Baddeley residuals for the Poisson log linear, Poisson log linear with covariate, and clustered linear with covariate models. Notice that none of the three plots show any obvious pattern, so we can suggest that the model assumption of constant variance is likely to be valid. The Baddeley residuals compare the number of points within the cell to the expected number of points. Although weve used an arbitrary 50 by 50 grid here for the computation, we do note that there appears to be a higher number of points at the center of the region than expected based on the model. This may be due to the lack

of negative income correlation with lottery winning at the center of Perfectia. But since this would have no bearing on the log linear Poisson model without covariates, it is more likely that a linear model does not take into account a sizable number of lottery winners at the casino, where lottery tickets may have been sold, or where people are more likely to gamble regardless of their backgrounds. The models we have considered does not take into account this increase in lottery winners near the location of the casino.

Another way to evaluate the goodness-of-fit of a model is to thin the data according to the probability of being in a region over the intensity at that point. The resulting thinned data set will be close to a simple stationary Poisson process if the intensity is indeed well modeled by the model under consideration. In figure 12, the thinned data and their resulting K and J functions are shown for both the clustered and Poisson log linear models. Notice that we still have the estimated K(r) greater than the theoretical K(r), and the estimated J(r) below zero, both sighs that the data is clustered, even after removal according to our algorithm. This may mean that even the clustered models fit does not completely take out the clustering factor in the design, but closer inspection also reveals that the departure from zero of J(r) is less pronounced, especially in the clustered models case. It is likely that calling nlm a few more times with the estimated parameter as the starting parameter may alleviate the situation by finding the more optimal MLE, although this thinning technique does reveal an inadequacy in the fit of our models.

3 Discussion.

The number of lottery winners within a region of Perfectia is explained by geographical location and income level. The number of winners correlates with latitude (y) and slightly with longitude (x), while also negatively correlating with income level. Fitting linear and quadratic Poisson and clustered models using spatstat and optimization of MLE show that the best model is the clustered model that includes a term for the income covariate, and that other models with the covariate are better than the models without the covariate term. We suggest that the distribution of lottery winners in Perfectia is not uniformly random as expected under a stationary Poisson process, since the K, J, and other functions also express deviations from theoretical values.

The distribution of income suggests that the pattern of lottery winners can potentially be explained by a decrease in income with an increase in latitude. The models that utilize income as a covariate had much lower AICs, and provided better fits. This suggests that Perfectians who life in Moller park and Baddeley housing projects may purchase tickets and win the lottery more frequently, possibly due to their need for money and motivation for simple get-rich projects. Higher income individuals also may have little motivation for acquire more wealth. Moreover, recreational centers like gold courses and tennis courts may have provide facilities for selling lottery tickets. Income is an important predictor of lottery winning.

The clustered model, and especially the Poisson models, do not take into account the presence of a casino at the center of Perfectia. Looking at the kernel smoothed density of the

lottery winners, we notice a small bulge at the center of Perfectia, especially at bandwidths less than 1.5. This may reflect the more numerous locations for selling lottery tickets at the casino and the possibility that people who live near the casino are more risk-prone, and like to purchase lottery tickets, hence increasing their chances of winning. The clustered model and Poisson model evaluated under the Baddeley residuals appear to be adequate, but the finer data thinning scheme reveals that neither are wholly accurate. Moreover, as the J function shows, the maximum discrepancy between the model and the actual data occurs at around $r = 1.5$, and the kernel smoothed version of the data shows a bulge around (6, 5), which is the location of the casino, while the models all show fairly even ascent in terms of intensity going from low to high latitude. Thus the presence of the casino may be a factor on top of the clustering and income distribution variables that we used to come up with our best models.

Future analysis of the data may include the refinement of the MLE technique to ensure that we only stop changing parameters to descend the negative log likelihood when the likelihood no longer changes due to a nudge in the parameters. This would allow us to more accurately determine the MLE for a model and improve the AIC. This may have been a problem with the nlm function arguments. Another suggestion would involve the use of a general clustered process model like the Strauss model or Neyman-Scott model. The latter was seen to generate points that look qualitatively similar to our data set when the radius is kept low (0.1 or so). Unfortunately, my computational hardware could not estimate either model using spatstat due to memory requirements. One drawback in my analysis is the inability to distinguish between clustering of lottery sellers and clustering of the population. The increased number of lottery sellers in an area may be due either to the clustering of neighborhoods around some nonresidential (empty) areas, or to the clustering of shops that sell lottery tickets. Another limitation of our technique is the lack of ability to account for structures and institutions within Perfectia, such as the casino at the center of the region. These institutions will differentially possess differing numbers of lottery sellers. The population at each area is also not specified. It could be that more people live at the higher latitudes, and hence are more likely to win lotteries. The income data would in that case not be an explanatory variable, and its correlation with the number of winners may be coincidental.

4 Figures.

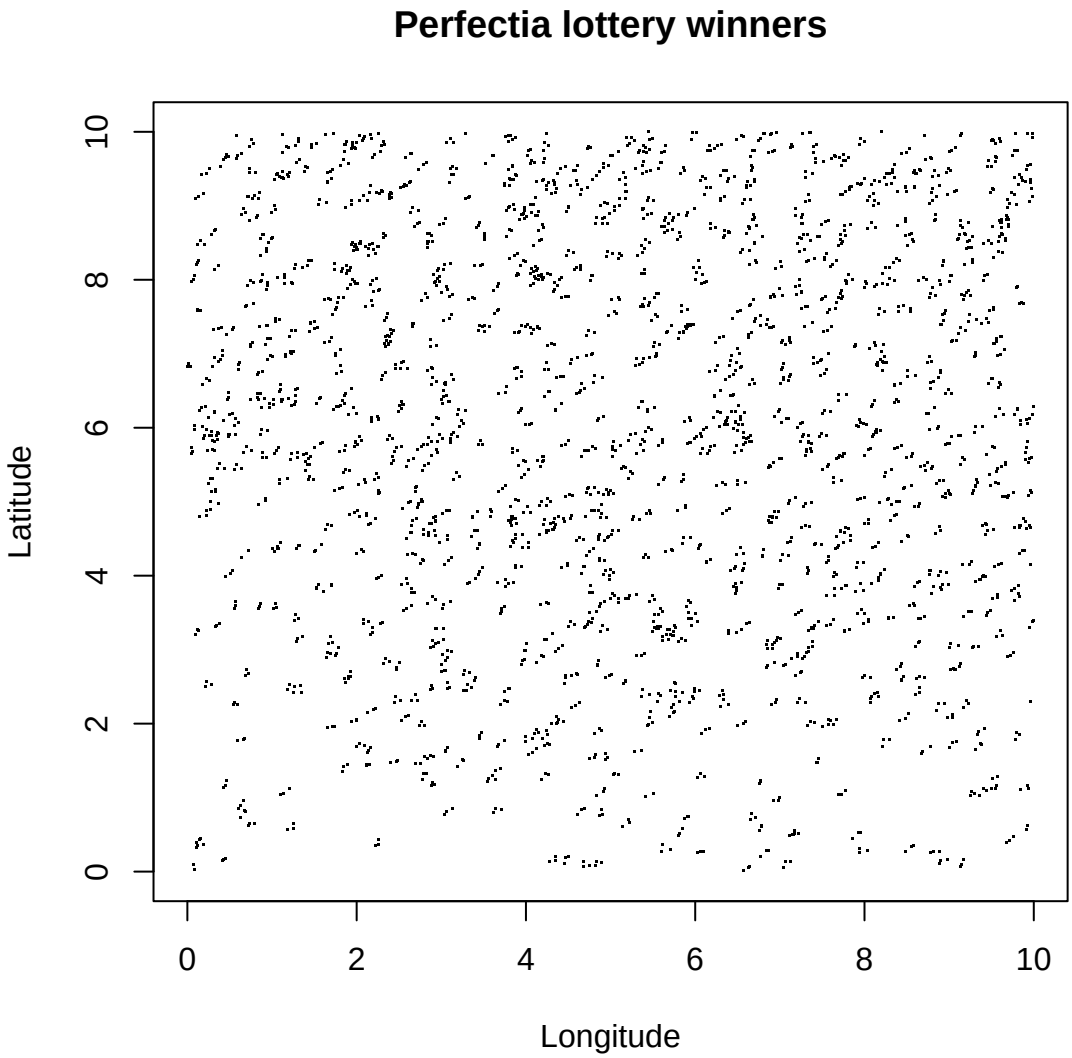

Figure 1: Graphical summary of lottery winners in Perfectia.

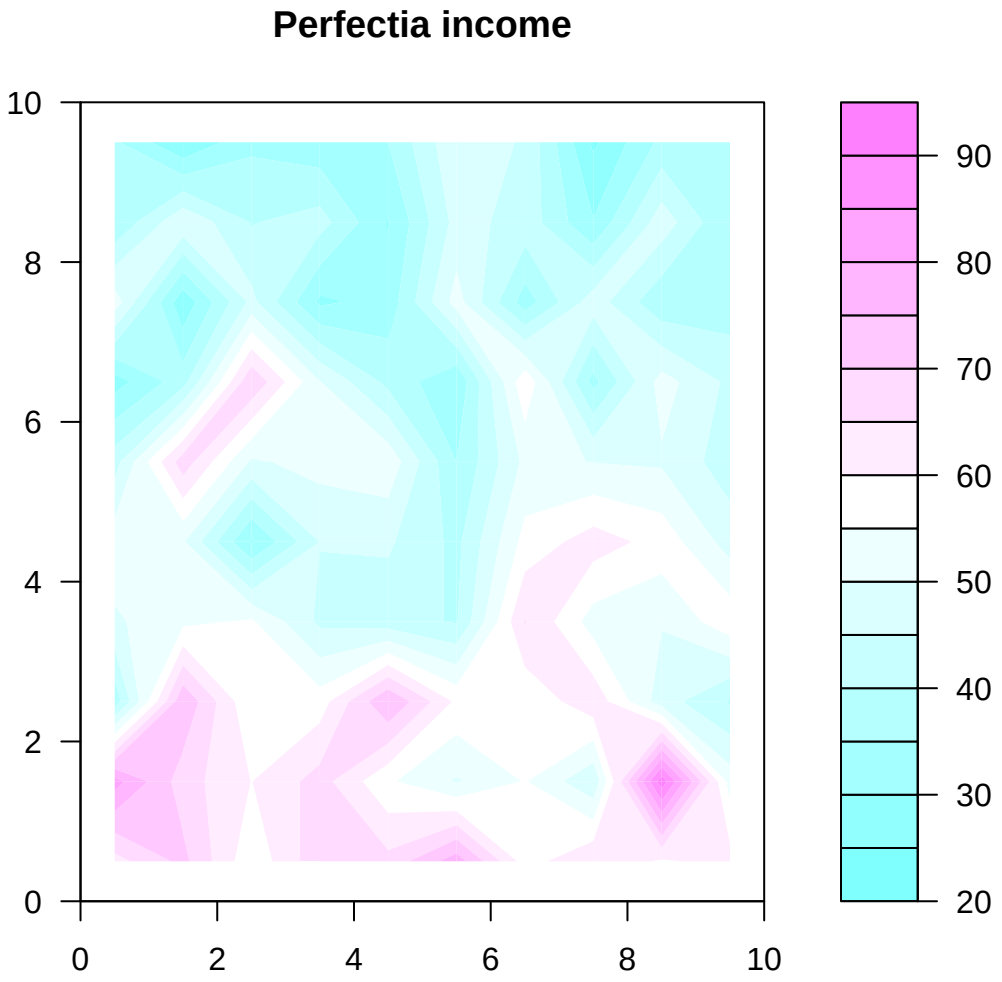

Figure 2: Graphical summary of income covariate in Perfectia.

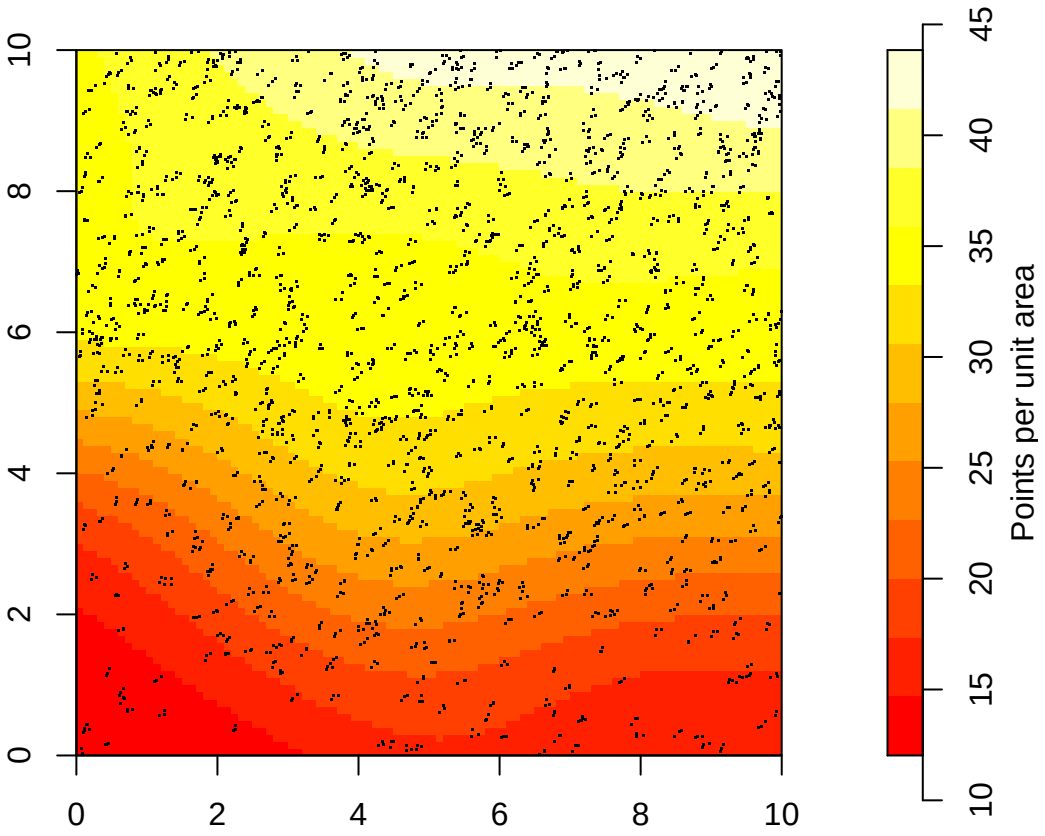

Figure 3: Kernel smoothed winners data using a bandwidth of 1.5.

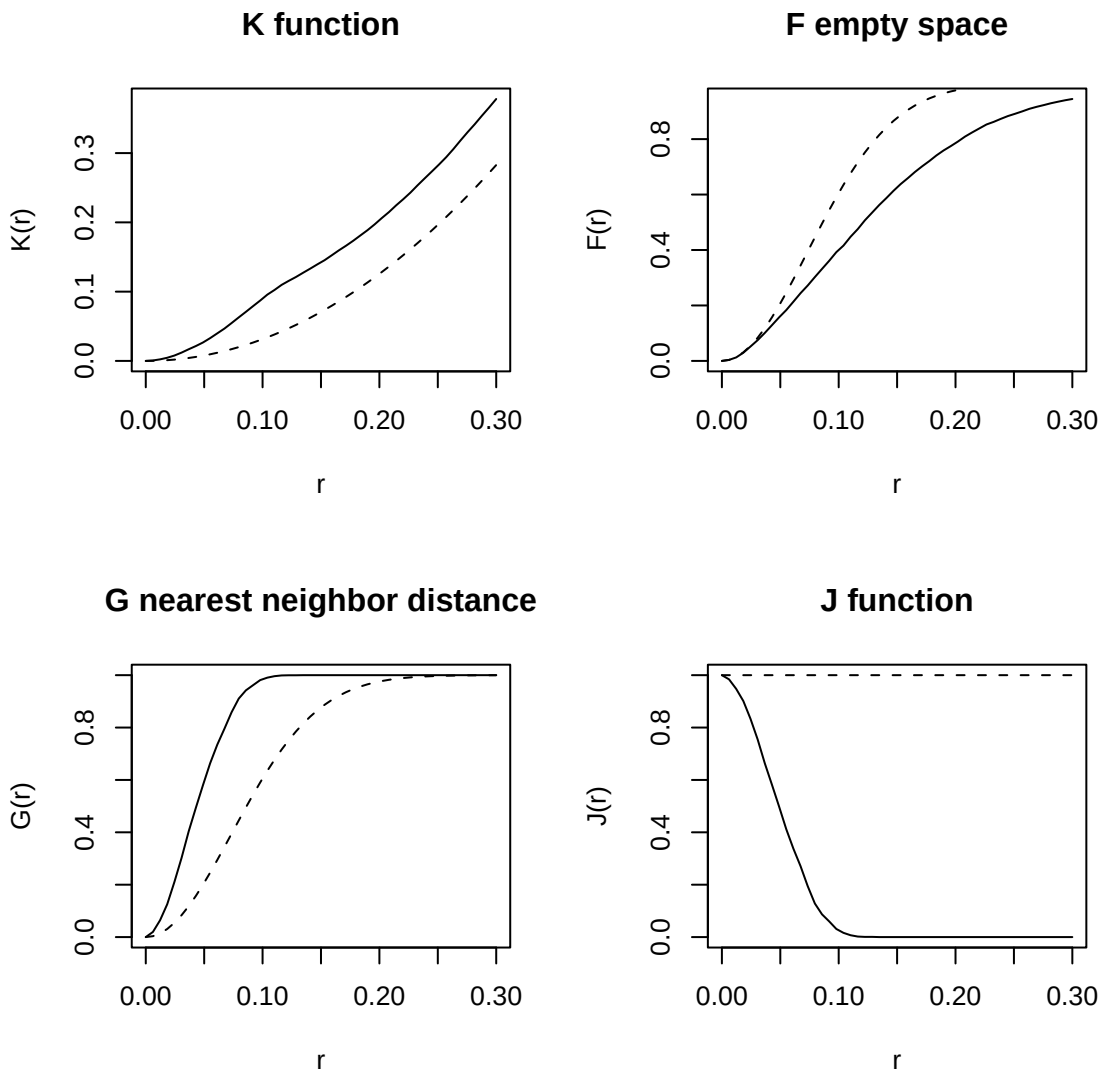

Figure 4: Theoretical (dotted) and estimated (solid) K, F, G, and J functions for the Perfectia winners data.

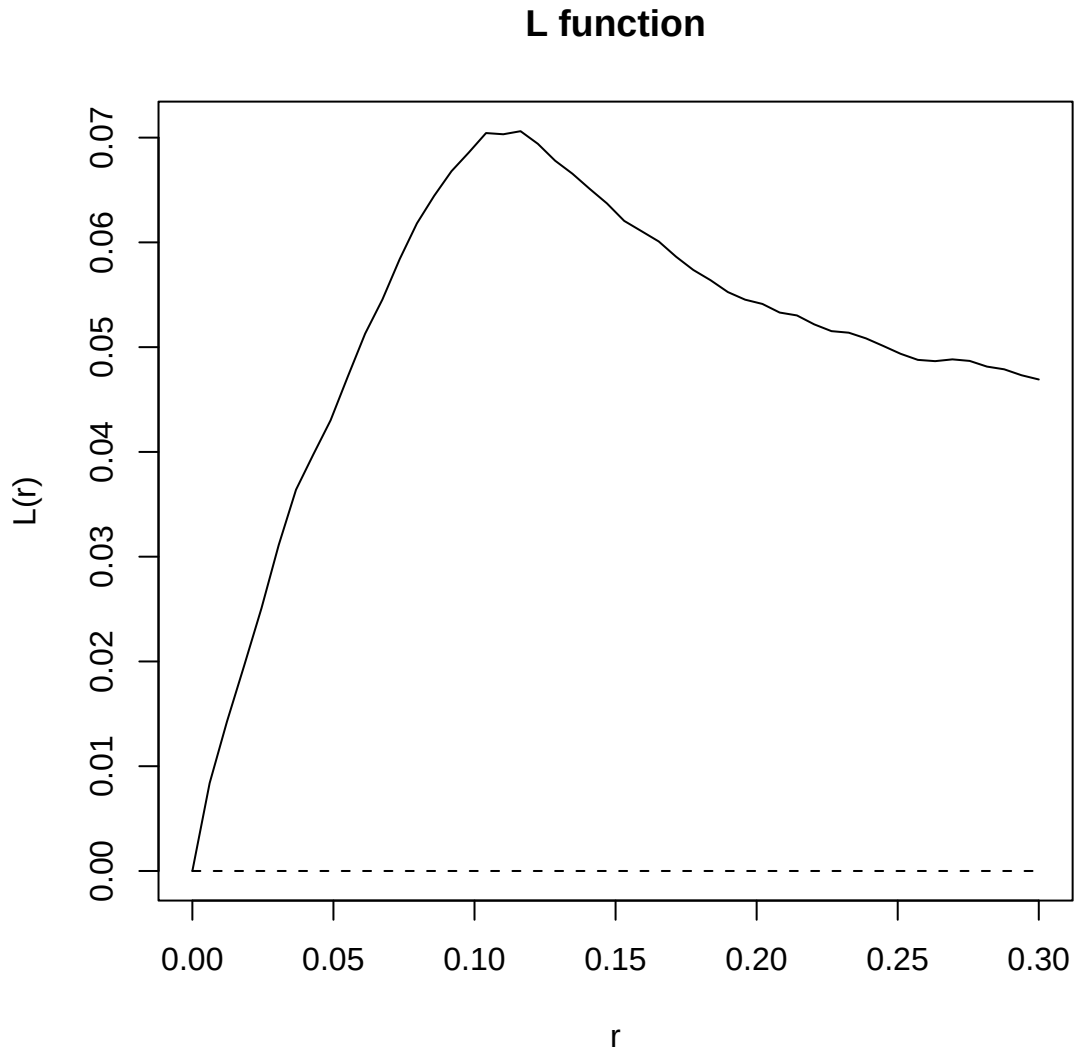

Figure 5: L function for the Perfectia data, calculated as $\sqrt{\frac{K(r)}{\pi}} - r$, which is expected to be constantly zero if the process is Poisson.

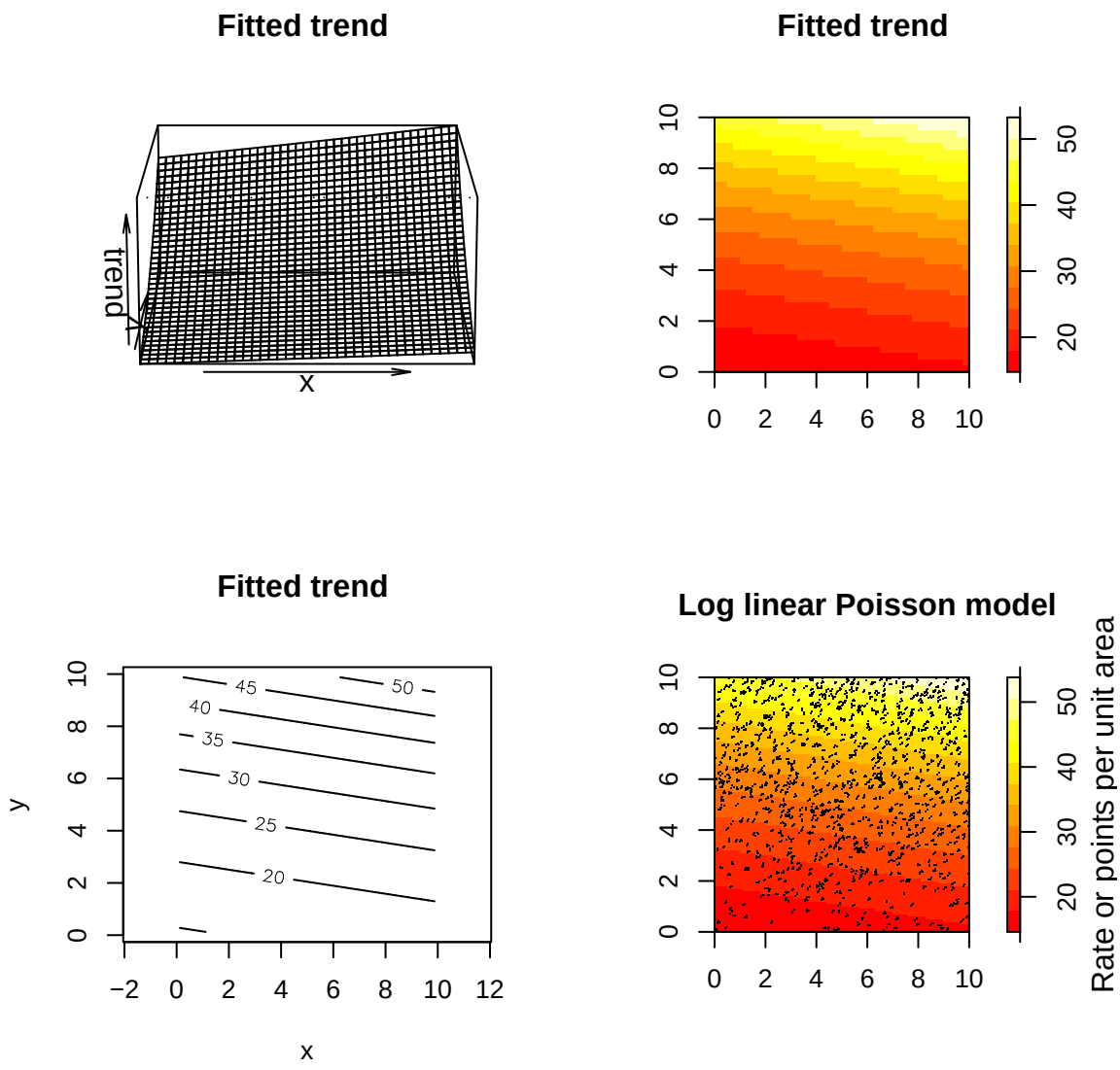

Figure 6: Log linear Poisson model fit to winners data, including trend surface and contour plot, points superimposed in the last plot.

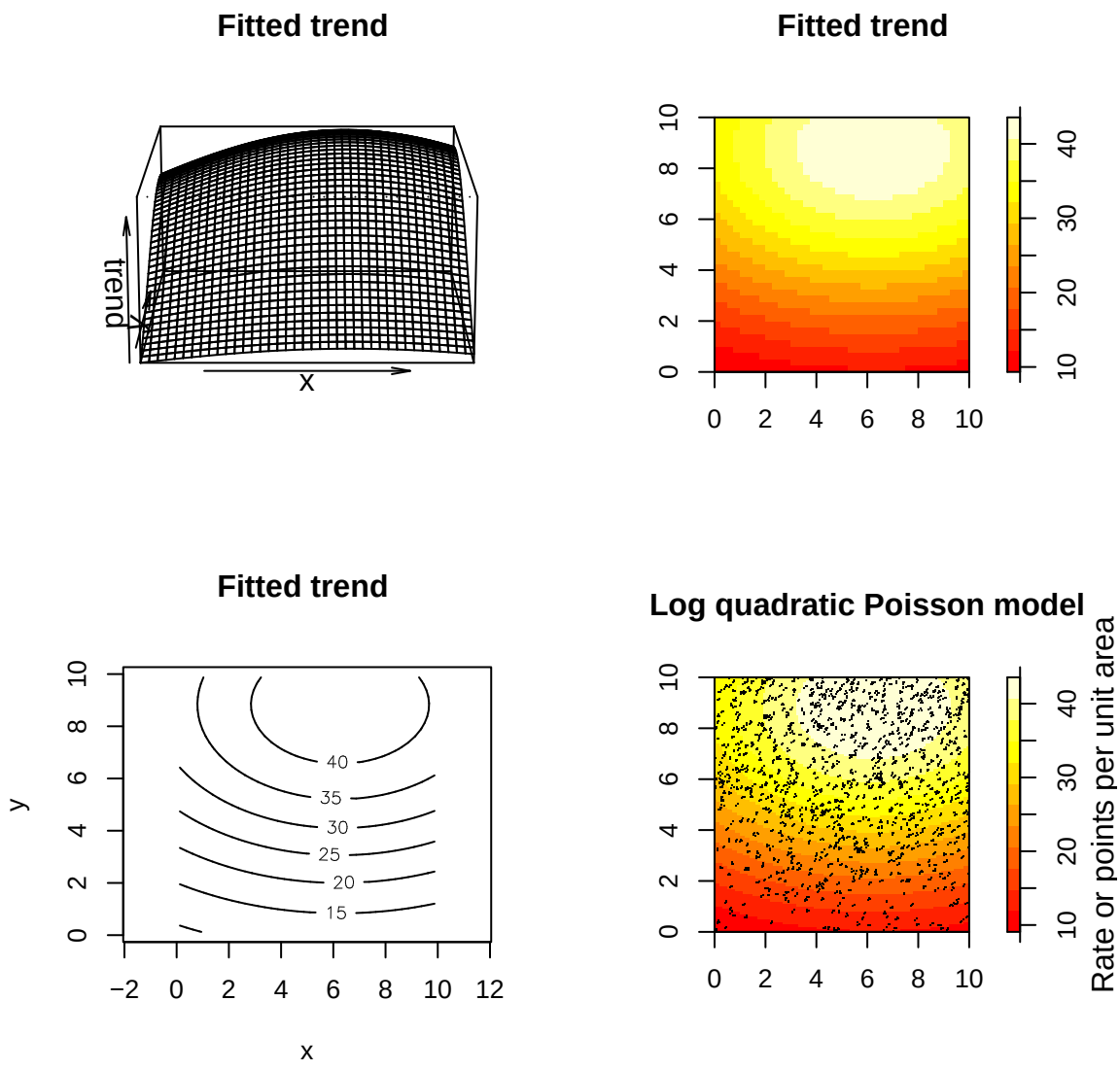

Figure 7: Log quadratic Poisson model fit to winners data, including trend surface and contour plot, points superimposed in the last plot.

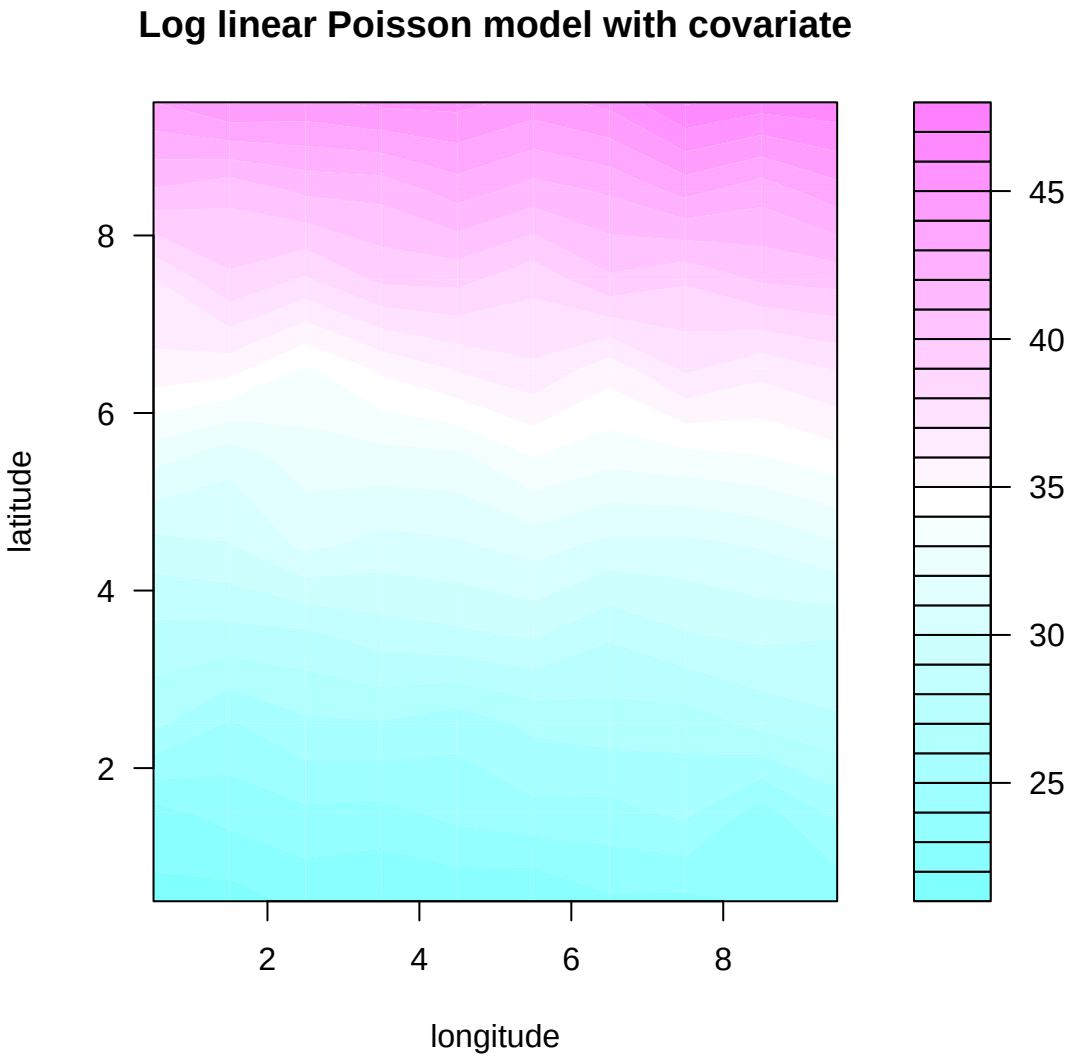

Figure 8: Contour plot of log linear Poisson model with income covariate.

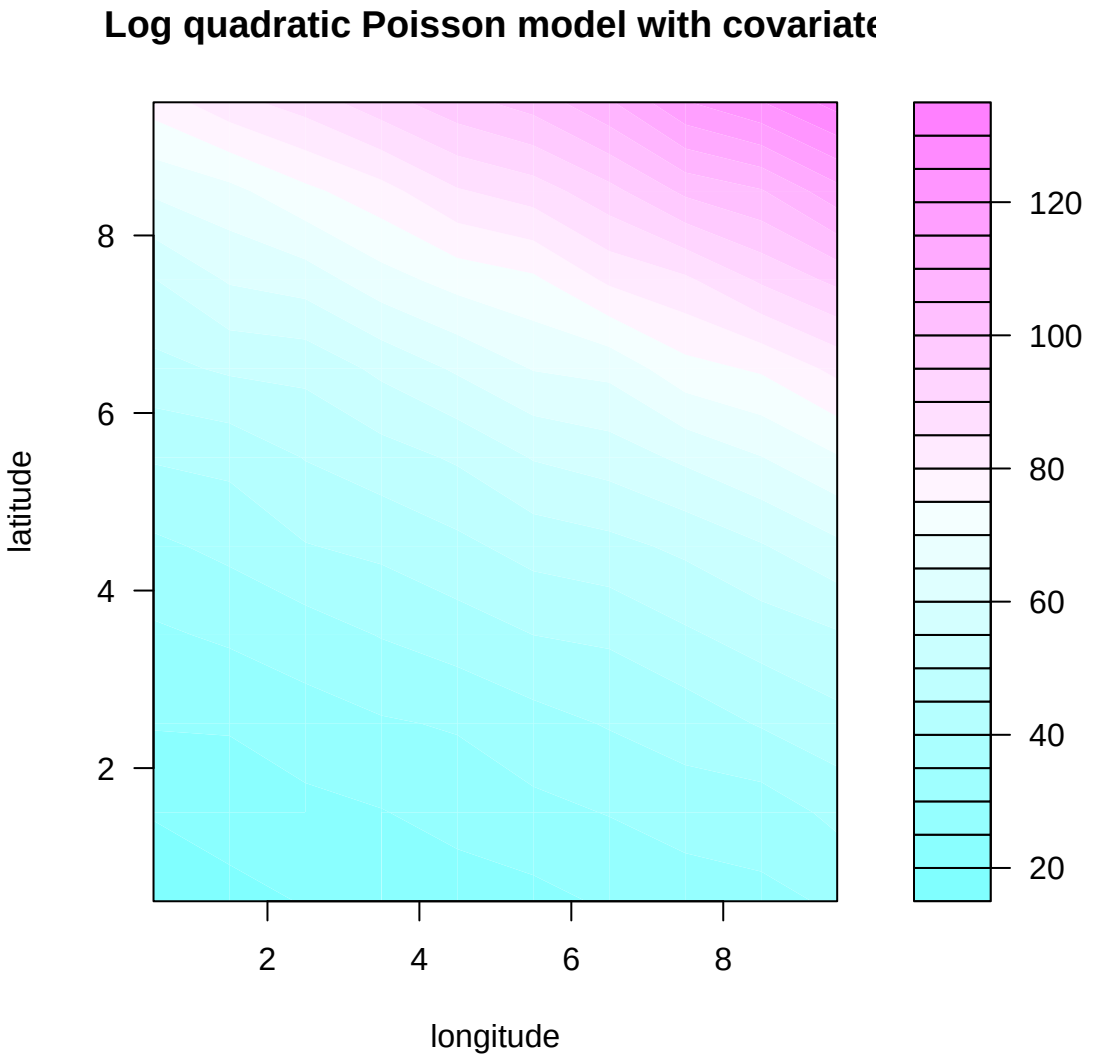

Figure 9: Contour plot of log quadratic Poisson model with income covariate.

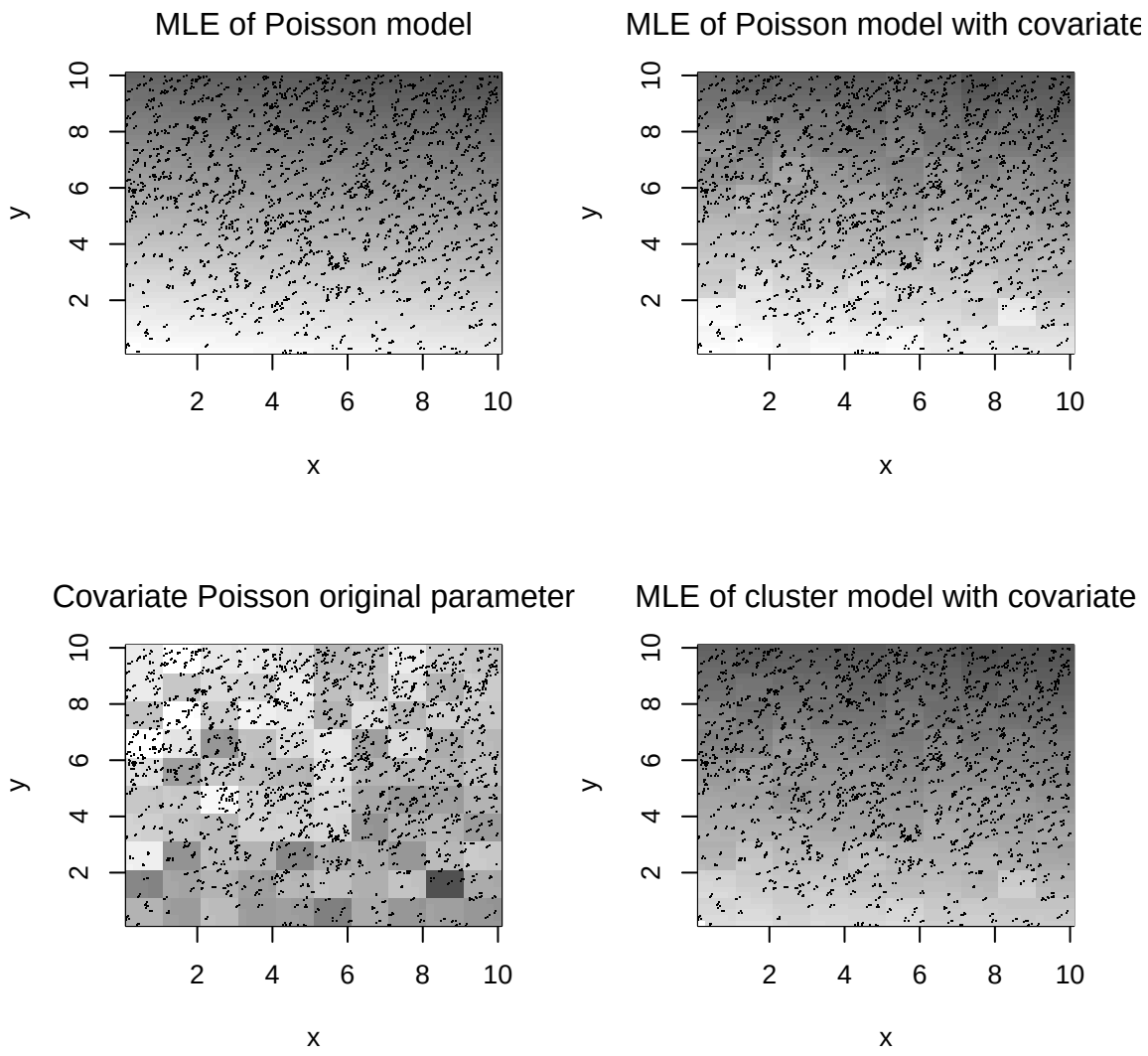

Figure 10: MLE implementations of log linear Poisson (upper left), Poisson with covariate (upper right), and cluster model with covariate (lower left) models, with original parameters for the Poisson with covariate before calling nlm as comparison.

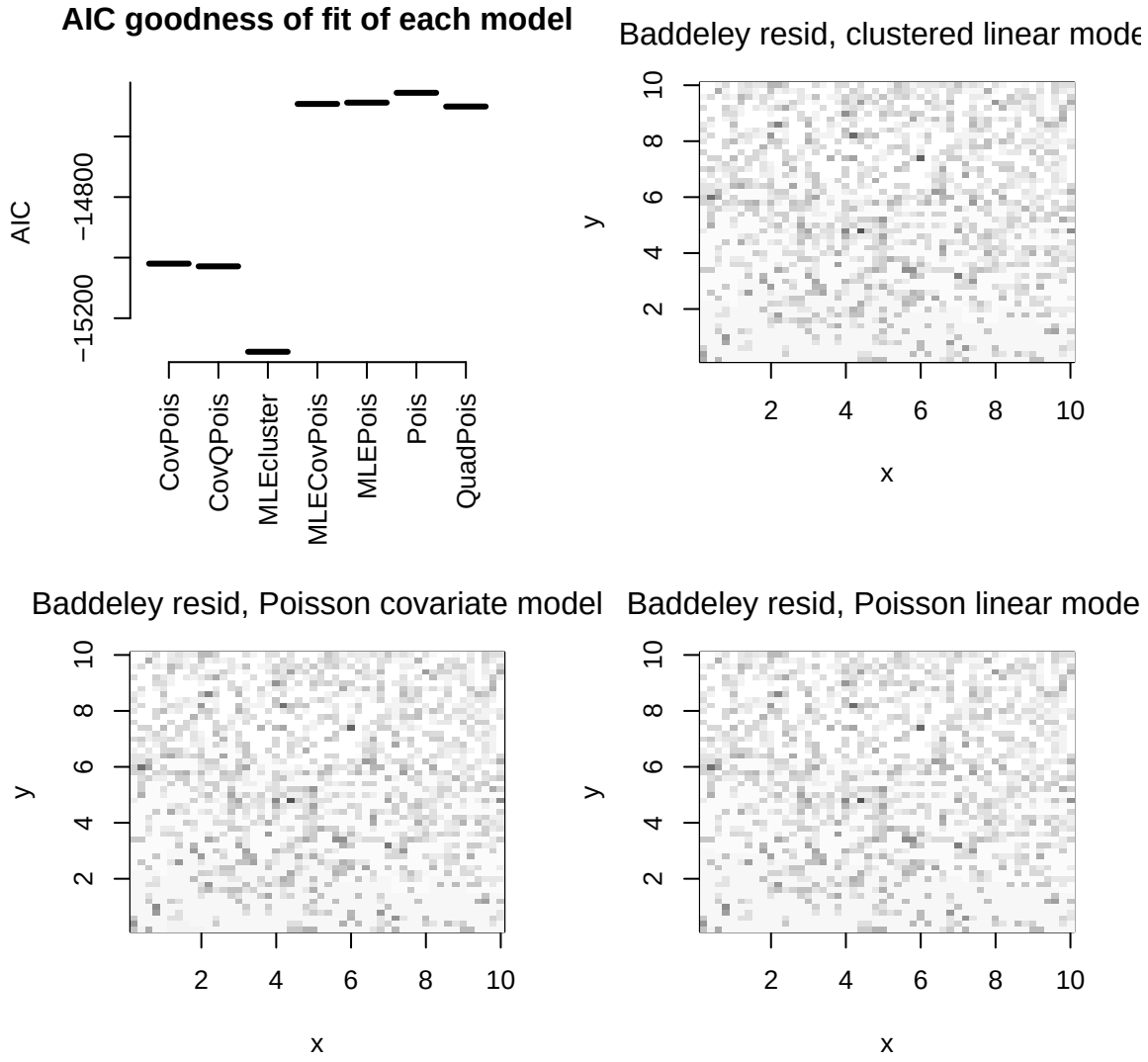

Figure 11: Goodness of fit assessment for each model. AIC (upper left) for each model, from left to right, the models are Poisson with covariate (-15020), quadratic Poisson with covariate (-15029), MLE estimation of cluster model (-15210), MLE estimation of covariate Poisson model (-14492), MLE estimation of Poisson model (-14488), log linear Poisson model (-14456), log quadratic Poisson model (-14501). Baddeley residuals (other plots) for each of the MLE estimated models.

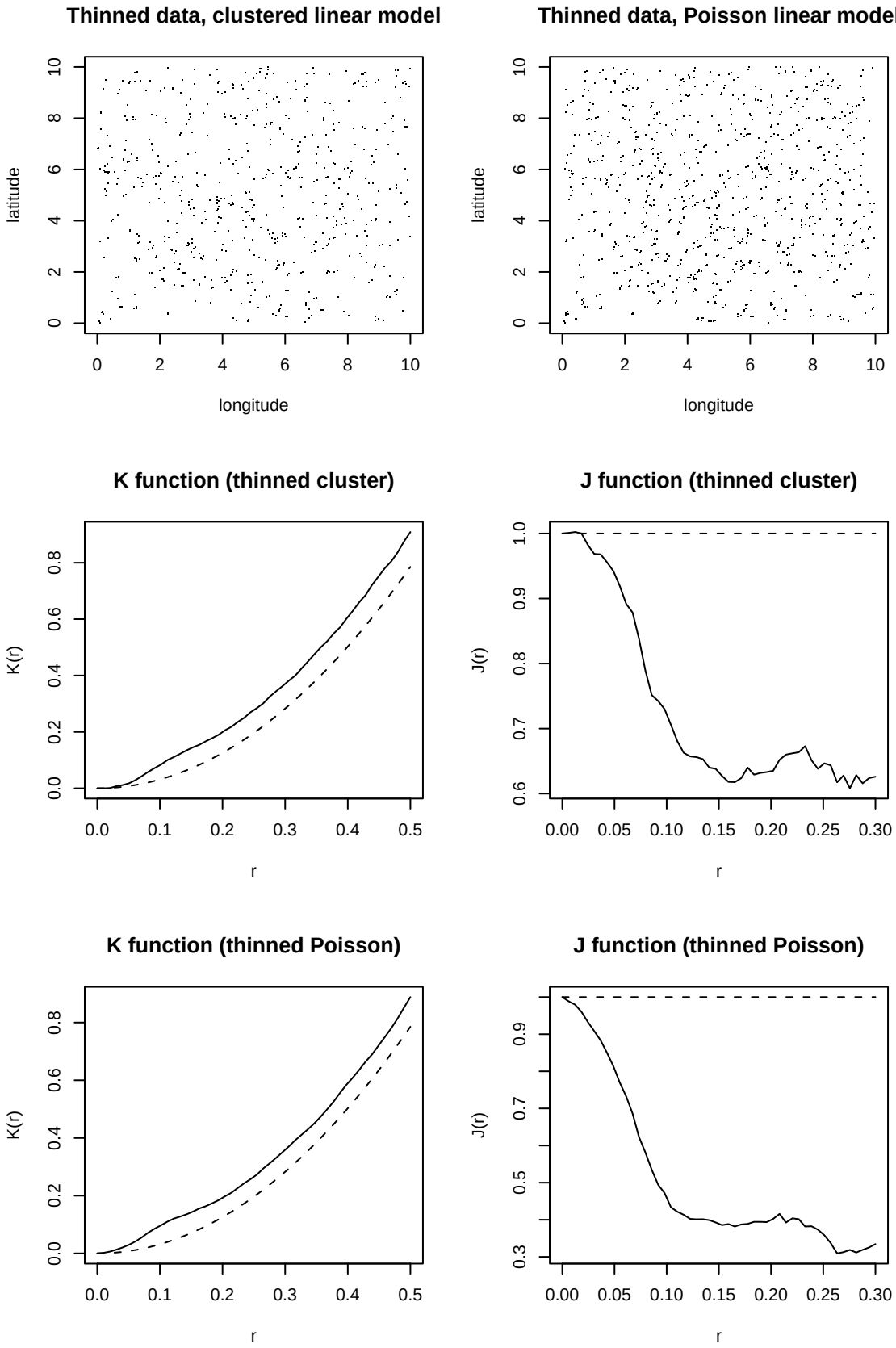

Figure 12: Thinned data using each model and their K and J functions.