HW5: Non-Gaussian spatial data (generalized responses)

STAT 574E: Environmental Statistics

DUE: 11/21 11:59pm

Homework Guidelines

Please submit your answers on Gradescope as a PDF with pages matched to question answers.

One way to prepare your solutions to this homework is with R Markdown, which provides a way to include mathematical notation, text, code, and figures in a single document. A template `.Rmd` file is available through D2L.

Make sure all solutions are clearly labeled, and **please utilize the question pairing tool on Gradescope**. You are encouraged to work together, but your solutions, code, plots, and wording should always be your own. Come and see me during office hours or schedule an appointment when you get stuck and can't get unstuck.

I. Mathematical [6 pts]

The goal of the following exercise is to show that sometimes log-linear inhomogeneous Poisson processes and Poisson GLMs are equivalent probabilistic models. The equivalence occurs when the areal units used to aggregate a point pattern into counts (e.g., the central park squirrels example) match up with constant values of the predictors.

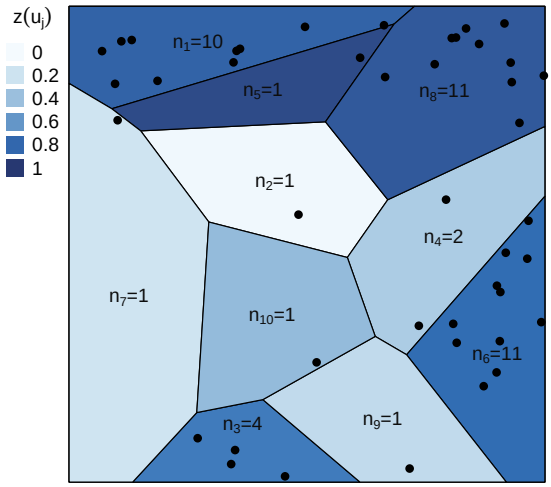


Figure 1: Example domain and partition. Points in the pattern $\mathbf{X}$ are marked by solid dots and counts of points in each subset, $n_j = n(u_j)$, are written at the center of each component of the partition. The color of each partition corresponds to the level of a predictor variable, $z(u_j)$.

(1) [6 pts] Consider a point process, $X(u)$, which takes place in a finite domain W that is partitioned into J compact sets. Let $u_j,\ j = 1, \dots J$ denote the jth set of the partition such that $\cup_{j=1}^{J} u_j = W$ and $u_j \cap u_{j^*} = \emptyset$ for $j \neq j^*$. Consider also the aggregated process $n(u_j) = \sum_{i=1}^{n} \mathbf{1}_{x_i \in u_j}$ that is the sum of points in any finite realization of $\mathbf{X} = \{x_1, \dots, x_n\}$ occurring in each of the J subsets of the partition.

Suppose a predictor of interest, $Z(u)$, is piece-wise constant with value z_j across each component of the partition such that $Z(u) = z_j \; \forall \; u \in u_j$. Figure 1 shows an example.

(Option 1) Show mathematically that the likelihoods for the log-linear inhomogeneous Poisson process for $\mathbf{X}$ given by $\log(\lambda(u)) = \beta_0 + \beta_1 Z(u)$ and the GLM for $\mathbf{n}$ with a Poisson response distribution, log link, and log-expected value $\log(\lambda(u_j)) = \beta_0 + z(u_j)\beta_1 + \log(|u_j|)$ are proportional to each other up to a constant that does not depend on $\boldsymbol{\beta} = (\beta_0 \; \beta_1)^\top$, which therefore implies that the MLE estimators of $\boldsymbol{\beta}$ are the same for each model (*hint: an expression for the likelihood of a log-linear IPP is provided in BRT Ch. 9.7.4*). Note, the extra term in the GLM, $\log(|u|)$ is not a predictor and has no coefficient. It is a known quantity sometimes called an "offset".

(Option 2) The file `HW5_1.RData` contains the realized point pattern `X` from Figure 1 above and a pixel-image `Z_pix` that provides the value of the predictor $z(u)$ at a fine grid of locations in the plane $u = (x, y)$. Use the objects to fit a log-linear model for the point pattern and estimate the coefficients β_0 and β_1. `HW5_1.RData` also contains the data frame named `aggregated` with columns `z` and `area` corresponding to $z(u_j)$ and $|u_j|$, respectively. Use the data frame to fit a GLM for the counts, $\mathbf{n}$, with a Poisson response distribution, log link function, and offset $\log(|u_j|)$. Verify these two approaches give approximately the same estimates for the coefficients β_0 and β_1.

Forest Inventory and Analysis

The data for this homework come from the US Forest Service's Forest Inventory and Analysis (FIA) program, which "collects, processes, analyzes, and reports on data necessary for assessing the extent and condition of forest resources in the United States" (**https://research.fs.usda.gov/programs/fia**). The data are publicly available, and access can be facilitated by the `rFIA` package. Two other resources I used to prepare the data for this assignment were (1) the USFS list of tree species and codes **https://usfs-public.box.com/v/FIA-TreeSpeciesList**), and (2) metadata about the meanings of the variables appearing in the FIA records. (**https://research.fs.usda.gov/understory/forest-inventory-and-analysis-database-user-guide-nfi**). You don't need to know about the data pre-processing to be able to complete this assignment, but I'm happy to share more details about what I did for those who are interested.

The file `FIA_AZ_Doulgas-Fir.csv` contains records about **Douglas Fir** trees in Arizona at the scale of FIA ground plots.

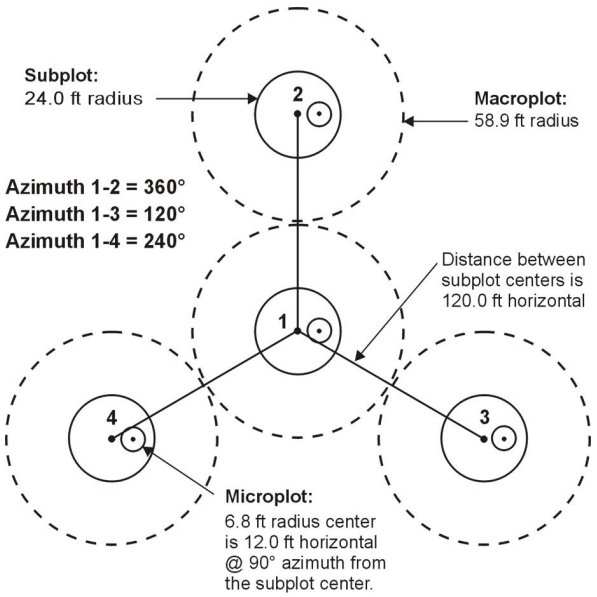


Figure 2: FIA sub-plot structure

Variable name	Definition
`PLOT`	Plot ID
`INVYR`	Inventory year. The year the data for the plot were collected.
`Common.Name`	Common name for tree species of interest: Douglas fir.
`DEAD`	Number of newly dead (since last site visit) Douglas fir trees recorded at plot.
`ELEV`	Elevation in feet at plot coordinates.
`LON, LAT`	Longitude and latitude of plot.

Figure 2 is taken from the metadata document referenced above and shows how four sub-plots are arranged within each plot. All the sub-plots associated with a given plot have the same recorded coordinates and elevation, and the rows in `FIA_AZ_Doulgas-Fir.csv` refer to observations aggregated to the plot level. A primary goal for this homework is to try and understand patterns in the deaths of Douglas Fir trees attributable to time and elevation.

II. Modeling counts [23 pts]

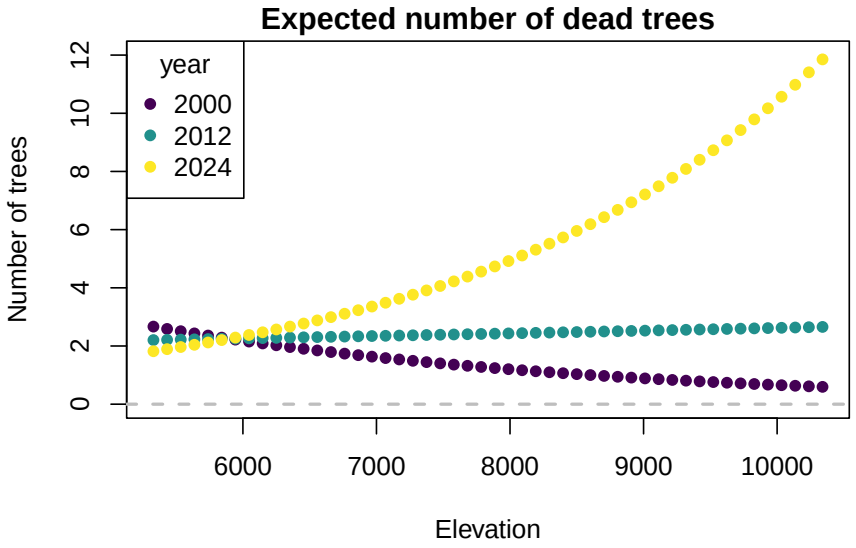
(2) [5 pts] Make a plot of the FIA plot locations with:

(i) map tiles in the background for context,
(ii) points colored by the number of dead trees recorded in a given year (you can include multiple inventory years for the same FIA plot as overlapping points), and
(iii) points scaled in size by the number of dead trees recorded. You may need to experiment with the relationship between point size and number of dead trees to get a reasonable figure, but make sure larger points correspond to more dead trees.

(3) [3 pts] Explain in your own words why a geostatistical model with a conditionally Gaussian response is inappropriate for the number of dead trees at each plot.

(4) [3 pts] Fit a generalized linear model (GLM) with a Poisson response distribution and log link function to the number of dead trees. Include linear relationships for elevation and year, as well as their interaction. Report 95% confidence intervals for each estimated effect. What is the predicted number of dead trees for a plot at an elevation of 7,000 ft in the year 2024?

(5) [5 pts] Create a figure like the one below that shows the predicted number of dead trees by plot elevation for the years 2000, 2012, and 2024. **PEC.** Explain in your own words how the effect of elevation on number of dead trees appears to have changed from 2000 to 2024.



For all following spatial models, use projected spatial coordinates for plot location according to a UTM Zone 12 coordinate reference system.

(6) [4 pts] Fit a spatial GLM to counts of dead trees with a Poisson response distribution and log link function and the same predictors as the non-spatial GLM. Assume the latent spatial random effect has a Gaussian-shaped covariance function. Report 95% CIs for the regression coefficients. What is the predicted number of dead trees at the plot with ID 82202 in the year 2024?

(7) [3 pts] Compute AIC for the models with and without a spatial random effect. In addition, use cross validation to estimate predictive bias mean-square prediction error for each model. Which model shows evidence of the best fit to the data based on predictive performance?

III. Modeling whether tree death occurred at all [11 pts]

Instead of modeling the counts of dead trees for each plot and inventory year, we could instead model whether *any* dead trees were recorded as a binary outcome (0 = no dead trees, 1 = at least one dead tree).

(8) [3 pts] Fit a new non-spatial logistic regression model for the binary outcome of "at least one dead tree recorded" as a function of elevation, year, and their interaction. Report 95% CIs for each regression coefficient. Qualitatively speaking, how much evidence is there for a non-zero interaction effect?

(9) [4 pts] Fit a spatial logistic regression model for the binary outcome of "at least one dead tree recorded" as a function of elevation and year and their interaction. Assume the latent spatial random effect has a Gaussian-shaped covariance function. Report 95% CIs for the regression coefficients. What is the predicted probability of recording at least one dead tree at the plot with ID 82202 in the year 2000? 2024?

(10) [4 pts] In your opinion, do you think it makes more sense to model counts of dead trees or whether any trees died at all? Explain your reasoning. Can you think of any situation in which the approach you *did not* identify might make more sense?