HW4: Random spatial index (point processes)

STAT 574E: Environmental Statistics

DUE: 11/7 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 [5 pts]

(1) [5 pts] Consider an inhomogeneous Poisson process over a region of the real plane that includes the whole unit circle. If the process has intensity $\lambda(u) = \kappa \exp\left\{\frac{-||u||^2}{2\alpha}\right\}$, $u \in \mathbb{R}^2$, what is the expected number of points in the unit circle, $\mathrm{E}[n(\mathbf{x})]$, when $\kappa = 10$ and $\alpha = 2$? You may provide your answer as a formula or a numerical value with three significant figures. You can answer the question using techniques from calculus, or take a numerical approach, or even a Monte Carlo approach.

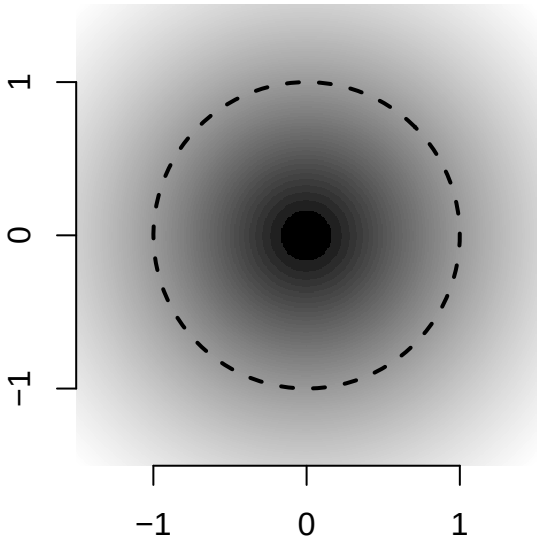


Figure 1: Intensity surface for question 1.

II. Exploration [17 pts]

For this lab, you will investigate a possible connection between earthquakes and oil & gas drilling in Kansas.

(2) [2 pts] Use the `tigris` package to create a simple features object named `kansas_sf` that represents the boundary of the state of Kansas. **Provide executable code** (**PEC**).

The following code reads data provided in the file `Kansas_Earthquake_Database.csv` by the **Kansas Geological Survey** into a data frame in `R` that contains the times, locations, and magnitudes of earthquakes detected in Kansas up through part of 2016.

```
1 library(sf)
2 earthquakes_df <- read.csv("Kansas_Earthquake_Database.csv")
3 earthquakes_df <- earthquakes_df[!duplicated(earthquakes_df[, c("Date", "Time",
4                                                                 "Longitude", "Latitude")]), ]
5 big_earthquakes <- earthquakes_df[which(earthquakes_df$Magnitude >= 3), ]
6 earthquakes_sf <- st_as_sf(big_earthquakes, coords = c("Longitude", "Latitude"),
7                            crs = "+proj=longlat")
```

(3) [2 pts] Explain in words what lines 3–7 do. **PEC** that replaces the original character string-class Date variable in `earthquakes_sf` with a new one that is a `Date`-class object (see `?as.Date()`). What is the most recent observation?

(4) [2 pts] Verify that there were 2 earthquakes of magnitude at least 3 in the year 1999. How many earthquakes were there in 2002? How many in 2012? 2015?

The file `ks_wells.RData` contains a data frame called `wells` with information about actively producing oil & gas wells in Kansas, also provided by the Kansas Geological Survey (see `?load()` for how to import it to your R environment). There are several attributes recorded for each well, including an individual identifier (`KID`), the date drilling began on the well (`SPUD_DATE`), the date the drilling was completed and production began (`COMPLETION` or `COMP_Date`), and the location of the well (`LONGITUDE`, `LATITUDE`).

(5) [2 pts] When was the most recent well completed? Which year had the largest number of wells completed? How many wells were completed that year? **PEC** that create a new variable called `wells_sf` that is a simple features object containing the information about wells in Kansas with an appropriate coordinate reference system.

(6) [2 pts] Make new versions of `kansas_sf`, `earthquakes_sf`, and `wells_sf` called `kansas_utm`, and `earthquakes_utm`, `wells_utm` that are projected to the Universal Transverse Mercator coordinate system, zone 13. **PEC**. Compare the bounding box for `kansas_utm` and the first few coordinate values in `earthquakes_utm` and `wells_utm` to the ones below to make sure you've transformed correctly (you might also want to try plotting all three objects on the same map).

```
8 st_bbox(kansas_utm)
```

```
##      xmin      ymin      xmax      ymax
##  751670.4 4098188.3 1425156.3 4473004.9
```

```
8 head(st_coordinates(earthquakes_utm))
```

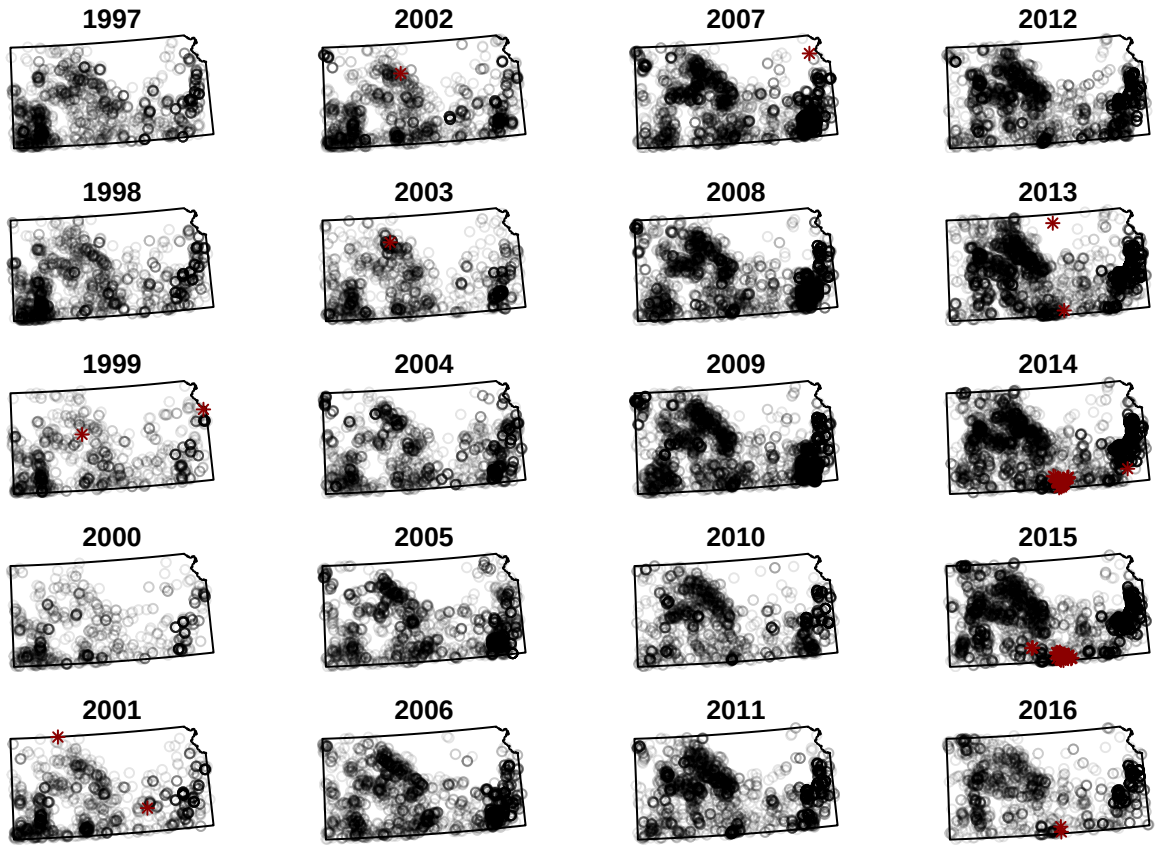
```
##            X       Y
## [1,] 1130987 4141142
## [2,] 1132613 4123701
## [3,] 1156619 4130853
## [4,] 1156767 4137308
## [5,] 1155940 4134256
## [6,] 1157185 4133917
```

```
8 head(st_coordinates(wells_utm))
```

```
##              X       Y
## [1,] 1054869 4131247
## [2,] 1052367 4301246
## [3,] 1058573 4284715
## [4,] 1058757 4288777
## [5,] 1067705 4288613
## [6,] 1068125 4281140
```

The following code plots the locations of all earthquakes detected in a given year (red asterisks) and all the locations of newly completed wells from the previous year (black circles). As you've probably already noticed in your explorations of these data, there was a dramatic increase in the number of earthquakes detected in Kansas in later years.

```
layout(matrix(1:20, 5, 4))
par(mar = c(1, 1, 1, 1))
for(year in 1996:2015){
  plot(kansas_utm$geometry, main = year + 1)
  plot(wells_utm$geometry[format(wells_utm$COMP_Date, "%Y") == year],
         col = scales::alpha("black", 0.1), add = T)
  plot(earthquakes_utm$geometry[format(earthquakes_utm$Date, "%Y") == year + 1],
         col = "darkred", pch = 8, add = T)
}
```



(7) [2 pts] From the explorations so far, do you anticipate any connection between the locations/numbers of new oil & gas wells and the location/numbers of earthquakes? Why or why not?

Moving forward, attention will focus on earthquakes recorded from 2014 onward and for wells completed between 2010–2013.

(8) [3 pts] **PEC** that creates an object called `earthquakes_ppp` using only earthquake records beginning in 2014. Verify your point pattern has 100 points. Compute the empirical $G(r)$ and $F(r)$ functions for the collection of all earthquake locations. Is it reasonable to conclude that the earthquakes arise according to a homogeneous Poisson process (HPP)? Why or why not?

(9) [2 pts] Compute the empirical $K(r)$ function for the collection of all earthquake locations. How does it compare to the theoretical function for a completely spatially random (CSR) process?

III. Models for earthquakes [18 pts]

To investigate whether there may be a connection between the locations of recently developed wells and the locations of earthquakes, one approach is to construct a model for the intensity function for the earthquake process as a function of the distance to the nearest well. The documentation for `?ppm()` explains that predictor variables in the trend formula can be specified in a few different ways, including as functions of spatial locations. You will create a function that represents a possible predictor of earthquakes and use it in a log-linear model for the intensity surface.

(10) [3 pts] Create a ppp object called `wells_ppp` that includes all wells completed anytime during 2010–2013. Verify there are 20099 records in the point pattern.

```
20 min_dist <- distfun(X = wells_ppp)
```

(11) [3 pts] Explain in your own words what the function `min_dist()` created in line 20 does. What value would `min_dist(wells_ppp[1])` return? Why?

(12) [3 pts] Fit a log-linear inhomogeneous Poisson process (IPP) model with `min_dist` as the sole predictor. Provide a 95% confidence interval for the effect of `min_dist` on the log-intensity surface. Do the values in the interval make sense to you? Why/why not? (Hint: if you get a warning about model fitting failing to converge, try using `method = "logi"` to use an alternative fitting algorithm.)

(13) [3 pts] Create simulation-based envelopes for the $G(r)$ and $F(r)$ functions implied by your fitted IPP from question (12) and compare them to the corresponding empirical functions for the observed point pattern in `earthquakes_ppp`. What do your figures suggest about the quality of the model fit?

(14) [3 pts] Fit a log-Gaussian Cox Process (LGCP) model to the earthquake point pattern with the same single predictor and an exponential covariance function. Report a 95% confidence interval for the effect of `min_dist`. How does your interval compare to the one from question (12)?

(15) [3 pts] Create simulation-based envelopes for the $G(r)$ and $F(r)$ functions implied by your fitted LGCP from question (14) and compare them to the corresponding empirical functions for the observed point pattern in `earthquakes_ppp`. What do your figures suggest about the quality of the model fit?