HW6: Spatio-temporal data

STAT 574E: Environmental Statistics

DUE: 12/5 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.

Los Angeles Bikeshare

For this lab, you will study an expanded version of the bikeshare data we saw in class. Instead of looking at a single day, we'll look at a whole week broken up into two-hour intervals.

I. Exploring the data [26 pts]

(1) [6 pts] First load the `spacetime` and `sp` packages and then load the `departures_STFDF` object stored in `bike_share_HW6.RData`. What is the maximum number of bikes that departed a single station in any single two-hour window during the week of January 5-11, 2020? Where is the station at which this occurred? On what date and during which time interval did the maximum occur? (*Hint*: the spatial, temporal, and covariate information inside `departure_STFDF` can be accessed using `@sp`, `@time`, and `@data`, respectively.)

(2) [4 pts] What are the coordinates for the three stations that had the largest total number of departures over the one week period? How many total departures did each one have?

(3) [4 pts] Overlay all 187 bikeshare locations on a map of Los Angeles and mark the three locations you found in question 3. Include your map in your submission.

(4) [4 pts] What is the name of the major train station near these busy bikeshare stations? What is one other feature in the area that you think people might decide to bike to/from?

(5) [4 pts] Add a new column to the `data` in `departures_STFDF` called `weekend` that has a value of `TRUE` for time intervals that fall on a Saturday or Sunday, and `FALSE` otherwise. Double check that you have $2 \times 12 \times 187 = 4488$ `TRUE` values and 11220 `FALSE` values.

(6) [4 pts] Do you expect the number of departures on the weekend to be higher or lower than the number during the week? Why?

II. A spatio-temporal model for counts [14 pts]

(7) [4 pt] Create a new STFDF object called departures_STFDF_proj that projects the coordinates using zone 11 of the Universal Transverse Mercator coordinate system. Rename the coordinates x and y instead of lon and lat. Provide executable code (**PEC**; hint: you'll need to use the projection function from the sp package, not sf).

```
head(departures_STFDF_proj@sp@coords)
```

```
##             x       y
## [1,] 383287.3 3768543
## [2,] 383616.7 3768240
## [3,] 384816.1 3768118
## [4,] 383839.4 3768248
## [5,] 383685.0 3767721
## [6,] 383192.6 3767032
```

(8) [6 pts] Use the mgcv package to fit a GAM for a Poisson response with log link function in which weekend is a fixed effect. Model the spatio-temporal random effect using flexible basis functions of your choice over space and hour of the day. Make sure you use the projected spatial coordinates. **PEC**.

(9) [4 pts] What is the point estimate for the effect of weekend? Does it agree with your expectations?

III. OPTIONAL: Two more spatio-temporal models for counts

For those interested and with extra time on their hands (which I realize may be no one, but just in case), here are a few extra questions to stretch your GAM skills a bit further.

(10) Use your fitted model to predict the number of departures from each location and time interval. What is the mean predicted number of departures? What is the maximum predicted number of departures?

(11) Overlay the predicted number of departures for each station between 4:00–6:00pm on January 10, 2020 on a map of Los Angeles using an appropriate color scale.

By definition, the mean and variance of Poisson random variables are equal. Two alternatives are the **zero-inflated Poisson** distribution and the **negative binomial** distribution. The zero-inflated Poisson distribution is useful when the data contain more zeros than we would expect for a conditionally Poisson random variable. The negative binomial distribution is often used when the mean and variance for the response distribution are not equal, which is sometimes called over- or under-dispersion. Diagnostics to check for zero-inflation and over-/under-dispersion can be difficult to produce, so one way to "check" for these characteristics is to fit models that accommodate them and see if they outperform simpler alternatives.

Consider the following models and pick the one that best predicts the values in a test set of locations and times. Let the test set be the number of departures from each station on each day between 4:00–6:00pm. The training set is be everything else.

(12) Fit the Poisson model to all data except those for 4:00–6:00pm on each day (hint: use the [, timeIndex] sub-setting method on departures_STFDF_proj[,]). Use your fitted model to predict the numbers of departures for the test set. Compute the mean squared error between the predicted values and the observed departure counts.

(13) Fit two new models to the training set with the same linear effects but utilizing the **zero-inflated Poisson** and **negative binomial** distributions. Make predictions for the test set and compute the mean squared error for each.

(14) Which of your three models has the smallest mean squared error? Is this what you expected?