Analysis of UFO Sightings in California

Greg Johnson

Math 531T - Spatial Statistics

Abstract

We examine the point process of UFO reports in California in the 6 year period 2010 to 2015. Data were webscraped from the National UFO Reporting Center database. A time-series plot alludes to possible clustering in the process however upon examination of a month prototype and Ripley's K and H functions, the process appears to be inhomogeneous Poisson i.e. it has spatially dependent intensity rates which we suspect are related to population density. Our final model is a kernel density estimate of the probability of a UFO report being generated from a certain location in California.

Keywords: Conspiracy, Aliens, Probing

1. Introduction

The subject of the existence of alien civilizations is a hotly contested one, even in hardcore scientific circles where intelligent conversations about extraterrestrial life lead to thoughtful discussions on topics like the Fermi Paradox or the Great Filter. Seemingly on the opposite side, there are those who chase lights in the sky (aka UFOs) and avidly watch the show Ancient Aliens; they are more like pseudo-conspiracy theorists than anything. This article pursues the subject of the spatial distribution of UFO reports without making judgments about what such phenomena may imply for the existence of aliens. This is partly because no matter how the data is analyzed and what the resultant statistics say, both camps can interpret it as evidence against or for aliens. Also, the author doesn't care - it's probably a deactivated missile from a Navy submarine anyway.

2. Data

The National UFO Reporting Center (NUFORC) has been collecting UFO reports since at least 1935. Somewhat recently they moved operations to a website with a submission form that people could self-report UFO encounters with. NUFORC makes the database of self-reports (dating back almost 100 years) freely available so long as one can scrape from html tables - which we did, specifically reports from California during the 6 year period from 2010 to 2015. We ended up with a sample size of 4,312 reports.

It is important to note that the data are from the point process of reports generated, not the point process of UFOs themselves (whatever they may actually be). This is important to keep in mind because the former is a poor proxy for the latter. For example, a UFO incident may occur but the witness may not know about the NUFORC reporting system or even care enough to report it (in rare cases of extreme probing, the witness may suffer from amnesia and not be able to remember what happened; read: sarcasm). Alternatively, an incident may occur where it streaks across the sky and a cluster of people see it from different locations and report it. Finally, there may be hoax/joke reports though the people who run NUFORC assure website visitors that they have a reliable vetting process for submitted reports.

The reports themselves consist of event date and time, location (often city but can be street intersections, national forests, etc.), qualitative description of the event, characteristics of the UFO (was it shiny?), and details of any close encounters (did it make the animals go crazy?). For our analysis we just needed date-time and location though the other variables may be an interesting mark for future analyses.

In its raw form, the data consist of text locations. We used a geocoding function in the ggmaps R package to convert the location variable to latitude and longitude. The function uses the Google Maps API to geocode. Any data points with latitude and longitude outside of California were removed. Another issue was that the data weren't exactly point process data because they failed the requirement of simplicity i.e. uniqueness of points. This was due to people who put counties or large cities as their location e.g. "I was touched by an alien in San Bernardino." These nonunique points were removed which reduced our final sample size down to 4,128 reports with unique locations.

3. Methodology

Our analysis consisted of classifying the point process with various diagnostic/exploratory techniques, and then creating an appropriate model based on the classification. As a purely exploratory step, the data were aggregated over the spatial dimension, providing UFO report counts for each month over the six year period. The resulting time series plot (see Appendix, plot 1) shows a relatively constant amount of UFO reports with maybe a bit of seasonality. A separate analysis might look at changes in the time series versus the occurrence of alien-related cultural phenomena (like when Jaden Smith asked Obama if aliens were real). What really stands out is a gigantic spike in November. The majority of these reports come from November 7th, 2015 when a glowing blue light could be seen all over Southern California and some of Nevada. It would seem that the underlying point process of UFO incidents generated a point at this location, which in turn generated a cluster of reports all around it. This may point to some sort of clustering effect in UFO reports.

Our next stop was to look at spatial data at certain time-frames. We assume independence of the spatial points between years and between months so we could treat the spatial data at each year and each month as a separate realization. Looking at these plots, it became clear that the clumps of reports in our data are not from a clustering effect but rather inhomogeneity in the reporting process - people were just more likely to file reports about incidents in some areas than not. A prototype (see Appendix, plot 2) based on monthly realizations confirms inhomogeneity - in fact the higher intensity rates seem to be near urban areas. As a final nail in the coffin for our suspicion of clustering, we looked at Ripley's K and H functions which also confirmed inhomogeneity.

4. Results

Since our methods point towards an inhomogeneous Poisson process, our final spatial model was a kernel density estimate (see Appendix, plot 3) of the probability of reporting a UFO at any one location in California.

5. Conclusions and Future Directions

Treating years and months as independent realizations of the UFO reporting process, we can visually see there are clumps of reports in the places - this

is corroborated by our prototype of a month of UFO reporting. The spatial patterns don't really change over time - they consistently clump in the same places (though they is still random variation in where they appear). Ripley's K and H functions provide another source of evidence that the process is inhomogeneous. It appears that the intensity (expected number) of UFO reports is a function of location - higher in more densely populated areas and lower in less densely populated areas. Thus we conclude that the point process of UFO reports in California is an inhomogeneous Poisson process. Our final spatial model is a kernel density estimation.

Intensity looks to be partly a function of the local population density. One way of testing this hypothesis would be to convert the point process data into geostatistical data using some sort of lattice - probably by county or zip code. Then the sum of reportings for each section of the lattice could be normalized by the appropriate county/zip code population. Then we could start looking at the number of UFO sightings per one thousand people (or whatever amount is more interpretable) and whether the geostatistical data are stationary, and/or isotropic. Then we could remove the background rate and create a kriging model.

Another point of conversation is the presence of clustering. The event on November 7, 2015 suggests that there may be some clustering mechanism at work - the underlying process (of UFO phenomenona) may generate a point that a cluster of people see who then create reports for it. Looking at yearly and monthly aggregates, if these clusters do occur, they are not apparent. Perhaps future analyses can factor in the dimension of time and look at a smaller time scale. Maybe clustering happens on the order of days where someone sees a UFO and tells their friends who are then psychologically primed to see a UFO. Maybe clustering happens on the order of weeks where a popular alien film (ala Cloverfield or Signs) comes out and movie-goers are sensitized to lights/unusual phenomena in the sky.

Appendix

Figure 1: Time Series Plot of Data Aggregated over Spatial Dimensions

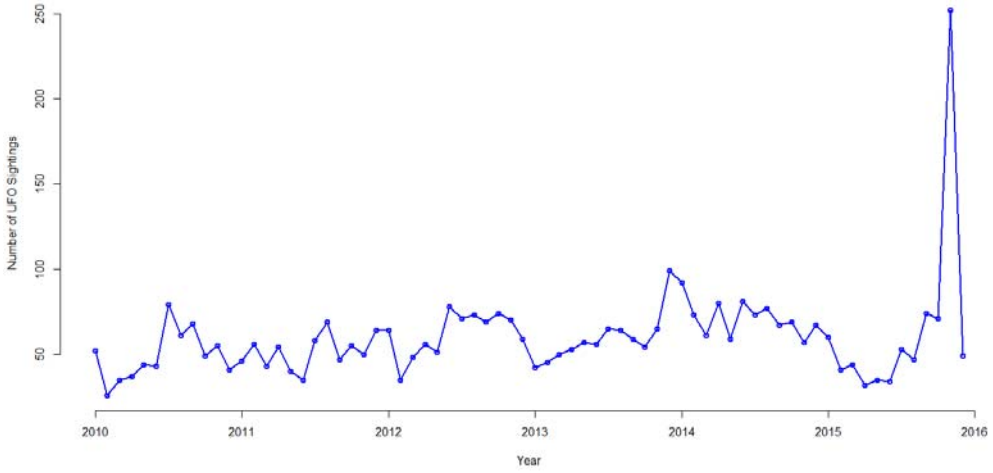

Figure 2: Prototype Treating Monthly Spatial Data as Realizations

Figure 3: Kernel Density Estimate Over 6 Years