Oiled Oceanic Ornithological Observations: Deepwater BP Oil Spill Data Investigation

Karsten Maurer Dai-Trang Le Li Wang Xiaoyue Cheng

April 27, 2011

1 Background

The Deepwater Horizon oil spill in the summer of 2010 is known as one of the largest manmade environmental catastrophes of all time. The spill stemmed from a gusher from a well on the oceans floor caused by an explosion on the Deepwater Horizon Oil Platform in the Gulf of Mexico, on April 20. The gusher expelled approximately an estimated 53,000 barrels of oil (8,400 m^3/d) per day until the well was capped on July 15. Over the course of the three months there was an estimated 4.9 million barrels (205.8 million gallons) of oil spilled into the gulf. The oil spill was detrimental to marine and wildlife habitats in the Gulf of Mexico and over the course of the oil spill thousands of animals were found dead or debilitated in the gulf.

2 Data Description

In this project we are looking at the bird observations from May 5 to Nov 6. The website for the data is http://gomex.erma.noaa.gov/erma.html#x=-90.04395&y=29.20012&z=7&layers=14581+5723. The dataset consists of 6 variables.

- Species: species of the bird found.
- Latitude, Longitude: location where the bird is found.
- Oiling: oiling condition on the bird. There are 3 levels: "Not Visibly Oiled","Unknown", and "Visibly Oiled".
- Condition: the bird condition, including two levels: "Dead" and "Live".
- Date: the date when the bird is found, from May 5 to Nov 6.

The main question we want to answer is: how did the oil spill affect the condition of the birds on the Gulf coast?

Associated questions of interest are:

1. What is the temporal trend of the bird death rate after the oil spill?
2. What is the spatial differences of the bird death rates?
3. Are there any unusual features in the data?

3 Suggested Analysis

Analysis Step	Reason for Use	Possible Queries Addressed
Data Restructuring - bin the locations to a few areas - bin the date by week and by month - aggregate the counts of death and live birds by time, area, and species	The count for most locations and days are small, so aggregating the data by grid and week is necessary.	- How to divide the shoreline into many spatial areas properly?
Summary Statistics	Discover location and scale for numerical data. Look at frequencies of observations for categorical variables. Examine the marginals for conditionals.	- Which type of birds has the highest death rate? - Is the location with the highest death rate the closest to the gusher?
Plots - maps - time series plots - animation	Demonstrate the variation of bird death by time and by location. Animate the change of death rates.	- Where is the killing zone for birds? - Has the oil spill impact decayed?
Modeling - spatial clustering	Numerically analyze the influence of time and location.	- Is there time dependency for the bird death rate? - Is spatial clustering related to the distance from gusher?

Table 1: Analysis Plan

4 Actual Results

4.1 Data Restructuring

To reveal the influence of the oil spill on the death of birds, one cannot only focus on the amount of dead birds, because the population base may be large. The death rate within an area during a time period would be a better variable which describes the living condition of birds. Also, the oiled rate in a location-by-time grid is another proper measurement. The comparison between oiled rate and death rate by time or by location will help us better understand the outbreak, spread, and decay of the oil, temporally and spatially.

The data are aggregated to calculate the death rate and oiled rate. Hence the time points and site coordinates need to be binned first. Considering the count of birds in each grid, we set each time interval to be one week and each area grid to be a rectangle with 0.5 degree of longitude and 1 degree of latitude. Figure 1 presents the aggregation.

In each of the bins, by the variable "Condition", we count the number of dead and live birds, and define a new variable,

$$\text{DeathPct} = \frac{\#\text{ of dead birds}}{\#\text{ of dead birds} + \#\text{ of live birds}} \times 100\%.$$

And with the count of birds on different oiling conditions, we define another variable,

$$\text{OiledPct} = \frac{\#\text{ of visibly oiled birds}}{\#\text{ of visibly oiled birds} + \#\text{ of not visibly oiled birds} + \#\text{ of unknown birds}} \times 100\%.$$

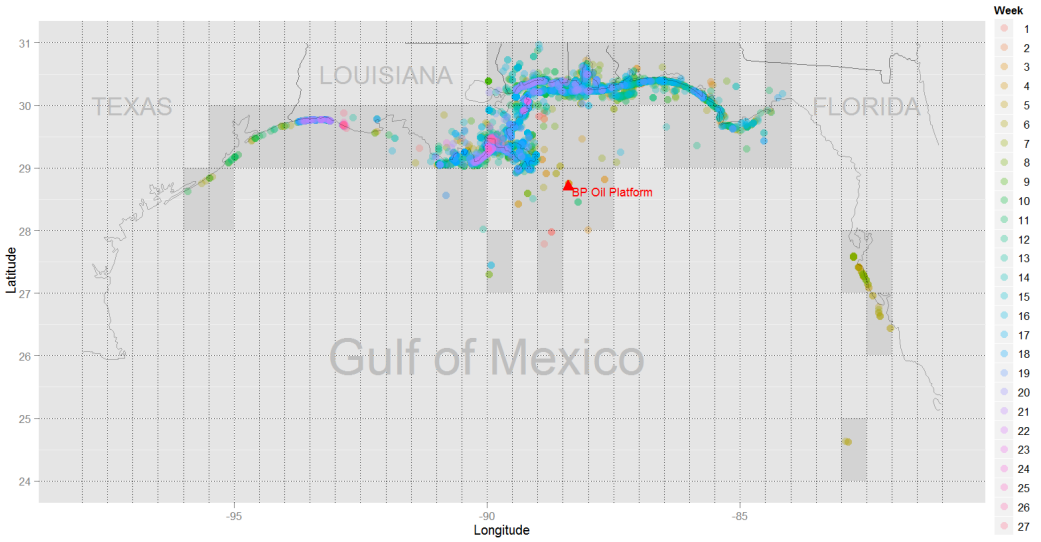

Figure 1: Map for the birds found after the oil spill. The records of birds are aggregated by week and by the spatial grids shown as dash lines on the map. There are totally 27 consecutive weeks and 47 not-all-adjoining areas.

Another issue is the species of birds. From the original data, the variable "Species" has 124 levels, including various gulls, pelicans, etc. To simplify the species, we mixed all different gulls to gull, all different terns to tern, and so forth. Then four biggest species are kept, and all the other species are combined into "other" species. Now all the birds are divided into five groups: gull, pelican, tern, gannet, other. For each of the species group, we aggregated the records and computed its own death rate and oiled rate.

4.2 Summary Statistics

Now the data are formed in two levels: case level and aggregated level. In the case level, we have longitude, latitude, gridx (grid indicator for longitude), gridy (grid indicator for latitude), date, week, month (derived from the date), species, group (mixed species group, as discussed in the last subsection), oiling, condition. In the aggregated level, we have gridx, gridy, week, # of birds, # of dead birds, # of live birds, DeathPct, # of visibly oiled birds, # of not visibly oiled birds, # of unknown birds, OiledPct, and these counts and rates for each group of species. Table 2 and 3 show the death and oiling situation for different species and different months. Table 4 reveals the relationship between bird death and oiling.

	Count	# Dead	DeathPct %	# Vis. Oiled	# Not Vis. Oiled	OiledPct %
Total	**7610**	**5783**	**75.99**	**2942**	**3907**	**38.66**
Gull	3544	2840	80.14	1379	1829	38.91
Pelican	877	516	58.84	374	378	42.65
Tern	766	580	75.72	356	343	46.48
Gannet	569	308	54.13	372	176	65.38
Other	1854	1539	83.01	461	1181	24.87

Table 2: General oiling and death situation for all birds and different species. Gulls and other species have higher death percentages, while the oiled rate of gannets is much higher than any other species.

	Count	# Dead	DeathPct%	# Vis. Oiled	# Not Vis. Oiled	OiledPct%
May	381	324	85.04	37	222	9.71
June	916	573	62.55	470	367	51.31
July	2366	1792	75.74	1222	985	51.65
August	2801	2104	75.12	997	1553	35.59
September	1080	924	85.56	206	739	19.07
October	57	57	100	9	33	15.79
Noverber	9	9	100	1	8	11.11

Table 3: Oiling and death situation by month. Most birds are observed on July and August. The death rate was the lowest in June and highest in Oct and Nov. Notice that the amount of birds found in Oct and Nov are quite small. The oiled percent increased in the first two months, reached the highest point in July, and then fell in the next few months.

	Dead	Live	DeathPct %
Not Visibly Oiled	3040	867	77.81
Unknown	759	2	99.74
Visibly Oiled	1984	958	67.44
OiledPct %	34.31	52.44	

Table 4: Contingency table for the oiling and condition of birds. The oiled percent of dead birds is lower than that of live birds. Except the unknown oiling situation, the death rate of visibly oiled birds is much lower than that of not visibly oiled birds.

4.3 Exploratory Plots

- Scatterplot Matrix

For the aggregated data, scatterplot matrix is an effective way to check the relationship among location, time, oiled condition and the death percent. From Figure 2, there seems to be weak association among the variables, but something interesting may be seen. For example, people began to find birds on the west coastline after 5 weeks of the oil spill, but most of birds found in the west were only slightly oiled but dead. Death percent and oiled percent are negatively connected which is beyond our imagine.The reason for it is probably that people would not intentionally catch the live but not-oiled birds, but this group may have large population. The lack of this group made the death and oil status negatively related.

- Temporal

The Deepwater Horizon oil spill was a long term event that impacted a large geographic area, and the data reflects this. Since the data is recorded with spatial and temporal components we want to investigate trends related to these measures of proximity. We aggregate all locations into 27 weekly totals and use this data to create timeplots to explore temporal trends. The percentage of birds found dead and the percentage of birds found visibly oiled are displayed in the timeplot in Figure 3. We see that of the recorded birds the percentage found oiled increases quickly until week 5 and stays around 60% until week 14 when it dropped considerably. Also of the birds recorded there is an initial spike in the percentage found dead then the percentage climbs steadily through the 27 weeks.

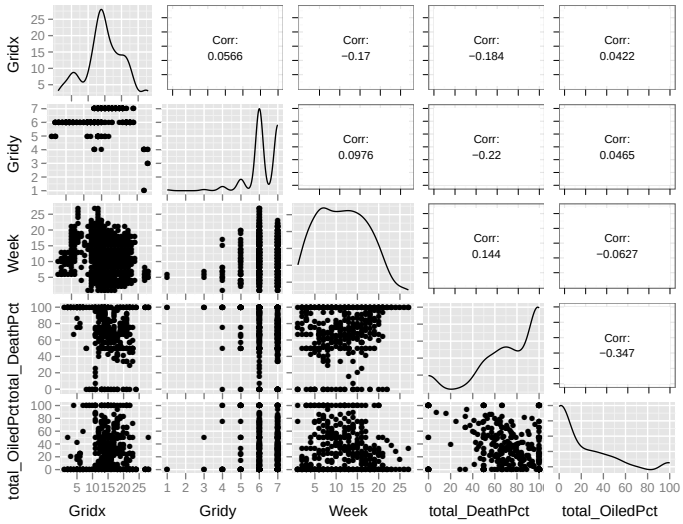

Figure 2: Scatter Plot Matrix of Location, Time, Percent of total death and Percent of total oil. Among these variables, percent of total death and total oil have the strongest correlation which is only -0.347.

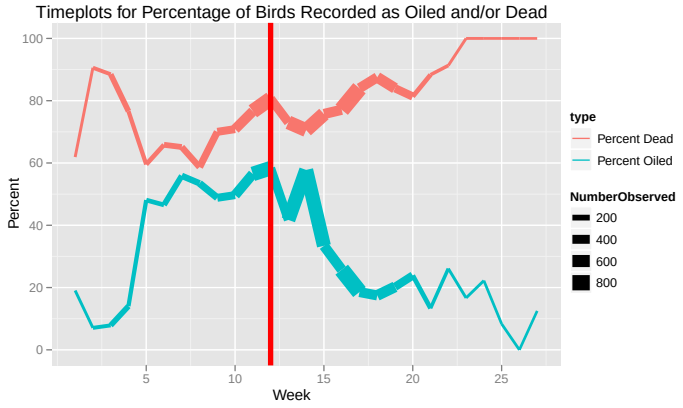

Figure 3: Timeplot of percentages of birds recorded as oiled and/or dead. The vertical red line indicates the time when the oil well was capped, we see the number of observations increase up until this point then begin to decrease significantly a few weeks after. The percentage oiled rises to about 60% then quickly drops after the well was capped, whereas the percent found dead spikes initially then grows consistantly.

- Spatial

To explore the spatial component we first aggregate the data over weeks and then use tiles to observe the variable values for a given area. The map in Figure 4 has tiles that are colored by the percentage of birds found dead. Of the reported birds there is a higher percentage as your move either east or west

from the Louisiana coast from which the oil was spreading. It seems odd that the percentage dead was higher further from the source of the spill, but perhaps this is because of a flaw in the way that the data was being recorded. Looking at the tiled map colored by the percentage of birds found oiled we see more of what we would expect from the situation (Figure 4). The areas closer to the source of the oil spill have higher percentages of oiled birds.

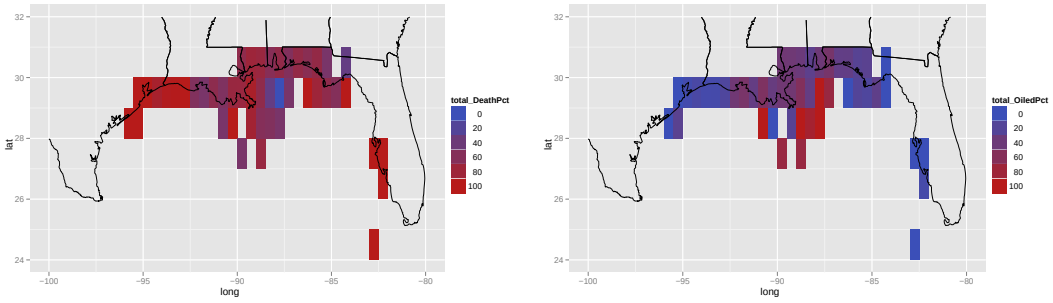

Figure 4: Tiled maps of average percentages of birds reported dead (left) and oiled (right) over the 27 week period. The reported dead percentages are lower around the Louisiana shore and grow higher further along the coast to the east and west. The reported oiled percentages are higher near the location of the Deepwater Horizon oil platform and are lower further away.

- Spatial/Temporal

Additionally, we can try to observe how these spatial trends change over time by creating an animated map that displays the slideshow of the 27 individual weekly tiled maps. The animated map of oiled percentage shows the same thing as the timeplot, where we see a larger percentage of oiled birds between weeks 5 and 14, but we also see that the oiled percentage increases off the Louisiana coast initially then the increase spreads across the coastline as the weeks go by.

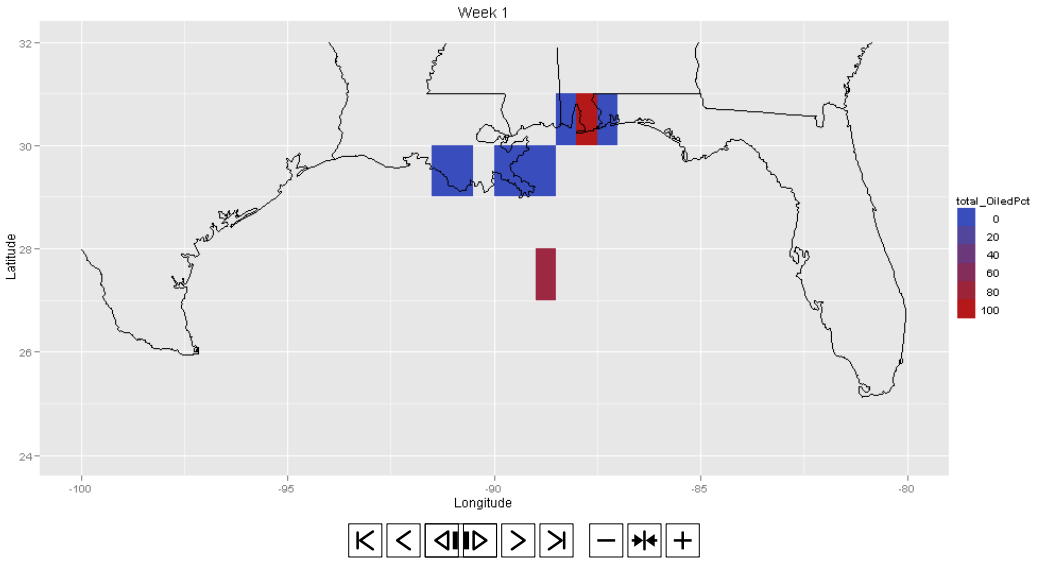

Figure 5: Animation for the oiled percentage during 27 weeks

The animated map of percentage of the recorded birds listed as dead shows that the cases were initially only recorded near Louisiana then recorded from Texas to Florida by week 6. The trend toward higher and higher death percentage in the observed birds is visible with more and more red showing up in the tiles. Again, this seems contradictory to what we see with the percentage of oiled birds, so perhaps as the time wore on the only birds the observers bothered to record were the dead ones.

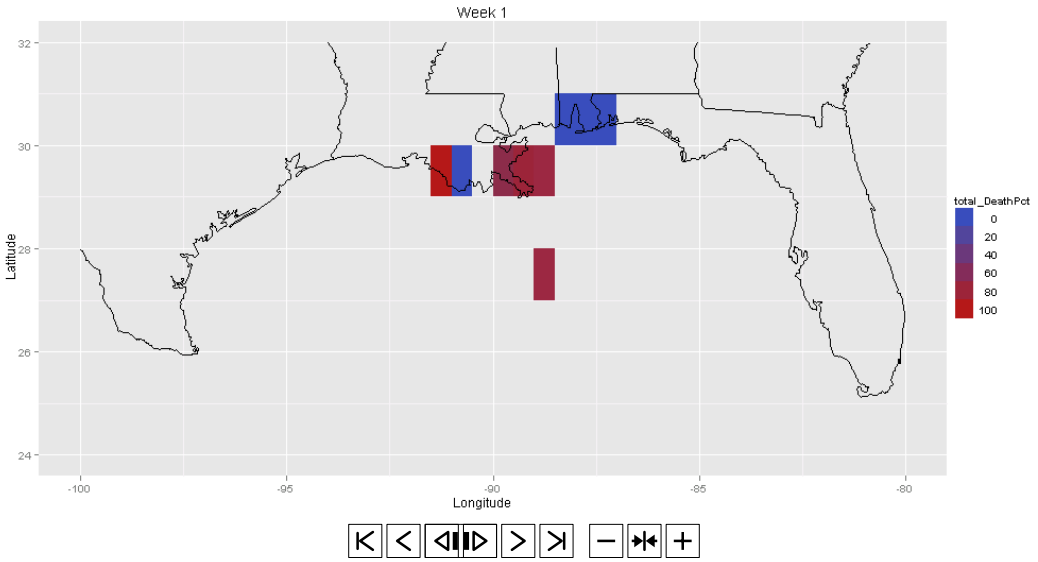

Figure 6: Animation for the death percentage during 27 weeks

5 Clustering Techniques

5.1 Determine the number of clusters

In this section we explore the spatial clustering for all the 47 grids to see whether the living and oiling conditions for birds are similar in some areas and different in others. In order to decide how many clusters exist in the data, we look at the scree-plot like pattern of the sum of squared errors for the 15 consecutive k-means models ranging from one to 15 clusters. The plot in Figure 7 shows that the sum of squared errors decrease rapidly from the first model to the fifth model then platau out from there all the way to the last model. This suggests that there may be five distinct groups in our data.

5.2 Hierachical Clustering

We then employ a variety of clusterinng techniques to separate our data. Among a range of hierarchical clustering approaches, Ward, single, complete, average, etc., Ward's linkage seems to create the most reasonable and readable plot. with five clusters. The plot in figure 8 displays the dendogram with five clusters enclosed in the red rectangles.

The Ward-linkage clusters are clearly divided and contain different sizes with two clusters having a small number of observations.

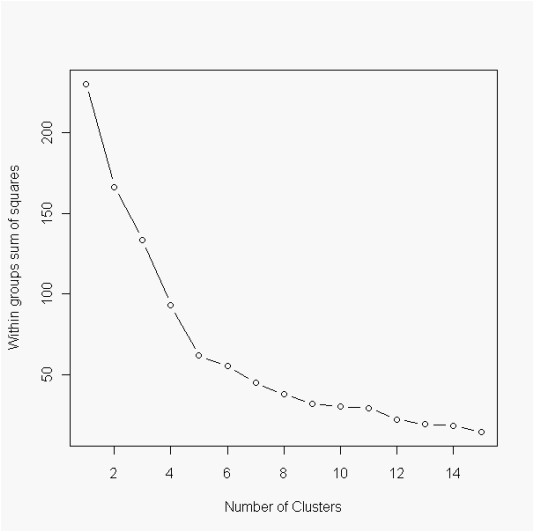

Figure 7: A plot of the within groups sum of squares by number of clusters extracted used to determine the appropriate number of clusters. The bend in the plot occuring at the fifth dots suggests to separate the data into five clusters.

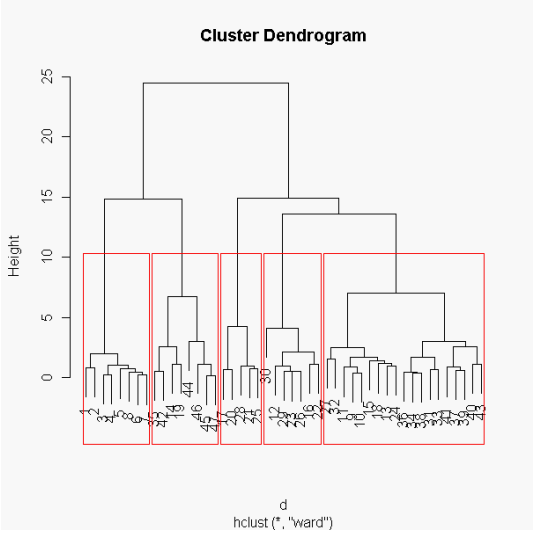

Figure 8: The dendogram suggests a seperation of five clusters enclosed in the red rectangles.

5.3 Model-Based Clustering

Using the model based-clustering method, we notice that model EEV with four components has the highest BIC. This suggest that the groups may have equal shape and sizes but with different orientation. The model based clustering EEV arrives to a different conclusion that the data is best separated into four groups based on the total death and total oiled percentages (Figure 9).

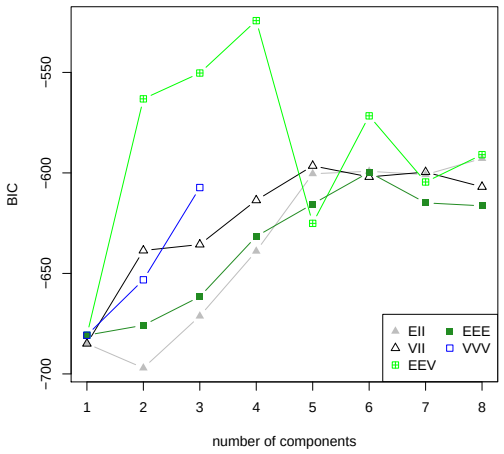

Figure 9: Model-based Clustering Plot. The EEV model with 4 clusters has the highest BIC.

5.4 Clustering Validation

Comparing the scores for each criterion between Ward's and model-based methods (Table 5), we notice that the Ward linkage technique performs better than the model-based clustering.

Criterion	Ward	model EEV	Better if
1	1121.62	1478.78	Lower
2	1.05	0.56	Higher
3	0.00	0.00	Lower

Table 5: Comparison of the three validation criteria for the two clustering methods above.

Consequently, we conclude that the spatial grids along the coastline of the five affected states can be segregated into five distinct clusters.

5.5 Spatial Mapping Result

Mapping the clusters based on the percent of oiled and the percent of dead birds found at each combination of longtitude and latitude on the coastline of the five affected states, we generate a spatial plot in Figure 10. The five clusters are separated by the different colors. Geographically, we can identify the clusters as, the area to the west of the oil well (in red), area to the east (in blue), the off-shore area closest to the well (in green), the shoreline area just a bit further away from the well (in purple) and the middle area on both sides of the platform (in yellow).

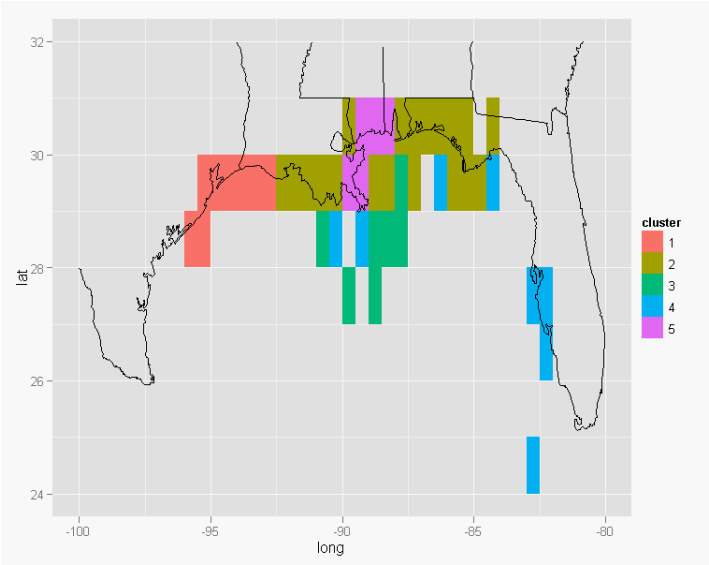

Figure 10: Five spatial clusters of bird species overlayed on the coastlines of the five affected states.

6 Conclusion

The data provided by NOAA allows for the exploration of the spatial and temporal effect of the Gulf oil spill on the wellbeing of the birds that inhabit the region. By restructuring and spatially binning the data, we are able to see these effects more clearly. The conditions that we are most interested in are the percentage of birds reported as oiled and the percentage reported as dead.

From summary statistics and timeplots we observe the temporal trends. The percent oiled increases to 60% by week 5 and remains near this level until two weeks after the well was capped (week 14), then drops quickly. The percent dead increases consistently to 100% over the 27 weeks, which indicates ill behaved data because we know that birds have not become extinct on the Gulf Coast. This peculiar characteristic of the death rate persists when viewing the spatial trends with higher death percentages observed further from the location of the Deepwater Horizon. The oiled percentages follow a more expected trend, with the higher proportion of birds found oiled on the Louisianna coast nearest the source of the spill. Clustering is used to group geographic regions of the coast that are effected in similar ways by the spill. The five clusters from Ward's linkage show fairly logical groupings: offshore near the Deepwater Horizon platform, Louisianna coastline nearest the platform, coastline to either side of the platform, far western Gulf coastline and far eastern Gulf coastline.

Overall, the data has a number of unexpected qualities and limitations. With no baseline data from a non-oil spill year we cannot argue whether the reported death rates are unusual. Also the proportion reported dead is obviously not a good measure of actually mortality rates in the Gulf Coast bird populations because there is a distinct lack of live, non-oiled birds being reported. Beyond that, data was gathered by many people with no guarantee of consistency in observation. Despite these data inadequacies, the spatial and temporal trends surrounding the oil spill proved interesting to explore.