I spent the final two weeks at Level Data Analytics Bootcamp working on a capstone project in partnership with EMC, a cloud computing and data storage company that provides a unified data platform for storing bird migration data gathered by citizen scientists, that is, non-professional scientists. I used two citizen science datasets: *Hawk Watch,* consisting of counts of raptors for three months in the fall at seven fixed sites in the northeastern continental U.S. going back as far as 1952, and the *ebird* dataset, consisting of counts of land birds, shore birds, and raptors (totaling over 7 million birds) since 1995 gathered at many locations across the continental U.S. throughout each entire year. *Hawk Watch*, on the other hand, gives larger counts at each site, but there are seven exact locations, so *ebird* gets much broader spatial coverage.

The *ebird* locations are not pre-specified, so at any precise latitude/longitude coordinate there will be few observations. Thus, the *ebird* observations must be aggregated; I chose 0.25 degree latitude (69.04 miles) by 0.25 degree longitude (49.63 miles) “squares” (which I will from now on call squares). Because all coordinates are rounded to the nearest 0.25 degrees latitude and longitude, all coordinates in a square are converted to the coordinates of the square’s center.

The main goal of this EMC group is to study changes in bird migration over time due to climate change, and many aspects of this are being studied. I wanted to study something that has not been examined, and through discussions with team member Richard Feldman I decided on the following problem: the extent to which *ebird* and *Hawkwatch* together can correctly predict known within-year migration patterns. In particular, raptors are known to migrate west-southwest in the fall in Maine, without any different major routes. With perfect data, this should be detectable, so I wanted to find out whether this was apparent in the data. My project was essentially a “reality check” on the data to assess its precision and predictive power.

There were two separate metrics I used: first, the date of peak fall migration (i.e. the day in the fall with the largest count of raptors) at each *ebird* square in Maine and at the *Hawk Watch* site at Acadia National Park (the only *Hawk Watch* site in Maine); second, the date on which 50% of each respective location’s total fall raptor count was observed (I will call this “50% date)—for example, if at the square with coordinates rounded to 45.5 degrees latitude, 68.5 degrees longitude, 200 raptors were observed throughout the entire fall (defined as September 1st to December 31st), the 50% date is the date on which the 100th raptor that fall was observed. The evidence I sought for west-southwest fall migration was, within each year, a consistent pattern of locations that are further southwest having peak dates and 50% dates occurring later in the year. I will refer to this phenomenon as “motion of peaks” or “motion of 50% dates.” In order to visualize this “motion” of peak and 50% date occurrences, I created a map for each location (an *ebird* square in Maine, or the Acadia *Hawk Watch* site) with a popup showing the date of the peak or 50% count—separate maps for the two—with a circle centered at the location with an area proportional to the precise count of raptors on the peak or 50% day at that location. If I then put these maps in order of time, then evidence of the pattern I sought would be a discernible motion of these circles over time in the west-southwest direction each year. It turns out that I could not detect this motion, and I will discuss this result later on.

There were many steps required before I could create these maps. First of all, I needed to figure out what constitutes a valid “peak” (the 50% date is formulaic and unambiguous). For each year, for each location with a nonzero raptor count, I used the *dplyr* package in R to aggregate together raptor observations that are on the same day to create the “birdcount” metric, and make a table with the birdcount value for each day. I then created a scatter plot for each location each year. A “peak” would simply be a local maximum that is also the maximum count for the entire fall. In practice, I accepted a date as a peak if the date of maximum raptor count in the fall had other large, but smaller, counts on days soon before and soon after. Since the *ebird* data are noisy, there are sometimes enormous outliers, i.e. enormous counts for which there are no large, but smaller, counts soon before or after. I simply omitted locations with outliers, or locations that simply do not have any discernible peak. Some of the sites did not have nearly enough observations to be reliable, so I omitted these sites also. By trial and error, I settled on the criterion for a location having “enough data”: having at least 20 days in the year with a nonzero raptor count.

That a location had fewer than 20 days with raptor observations does not necessarily mean that there were not many raptors at these locations, but could perhaps mean that few people went out to gather data at these locations. The counts of raptors are much larger in coastal Maine than in the interior, which could partly be explained by a real ecological pattern, but could just as easily be explained by the fact that more people live near the coast in Maine and hence more people go out to observe birds. This is clearly a bias problem, and it is an extremely difficult one to resolve. I attempted dividing each day’s count by that day’s number of observers, in order to normalize the data. I carried this out and got nonsensical results, so I have not included the R code or maps from this attempt. What I could do, however, was consider only the peak (or 50% count) *dates* and not the actual number of birds counted on those dates (at each site that I deemed to have the minimum total amount of data). I decided to make the circles in the maps proportional to absolute bird count, but only to provide information: I did not analyze absolute counts. The important information given by each map is the *location* of the center of the circle.

As I briefly discussed earlier, for each year, for each included location, I created a detailed map (one for peak date, one for 50% date) and placed a circle centered at the location, with an area proportional to the count on the peak dates or 50% dates, and I also placed a popup displaying the date on which the peak or 50% count occurred. I did all of this using the R package *rstudio/leaflet*, which creates HTML widgets with live connections to the internet, thus allowing the maps to be zoomed and shifted even after they are created; the maps are interactive.Within each year, I displayed the maps in order of time. I then used the R package *ImageMagick* to create a GIF, one for each year, stringing the separate maps together into a single animation. I did this by manually saving each map as a PNG file and entering all of these PNG files into an *ImageMagick* function. Unlike the maps created by *rstudio/leaflet*, the GIFs created through *ImageMagick* cannot be zoomed or shifted. There may be a way to create interactive time-sequenced animations of the maps, and this is something I would like to look into in the future. Creating fixed GIFs with *ImageMagick*, however, was more feasible given that I had only two weeks.

It was never obvious that the imperfect *ebird* and *Hawk Watch* data could demonstrate west-southwest “motion” of peak dates and 50% dates using the methods I devised, and it turned out that the pattern I sought could not be discerned. For years before 2010, and for 2014, there weren’t enough locations with enough data, so I have not included animations or code for those years. Looking at the GIFs for each year from 2010 to 2013, for both peak date and 50% date separately, there was no clear west-southwest motion over time. For the peak date GIFs, the motion of circles was seemingly random. For the 50% date GIFs, the circles moved less randomly, in the sense that one circle was usually followed by a somewhat nearby circle, although this was often in a direction other than west-southwest. There were, however, still some large jumps in circle location.

The GIFs for peak dates are on my Github page in the section titled “ebird (raptors) fall peak dates 2010-2013” and the GIFs for 50% dates are in the section titled “ebird (raptors) date on which 50% of total fall migration has been observed 2010-2013.” It still seems plausible to me that the pattern I sought exists and would be detectable in perfect data, based on the intuition that if there is a huge number of raptors at location A, and the raptors tend to be moving west-southwest, then the huge number of raptors should move in that direction and be detected again in huge numbers at some location B, west-southwest of location A. If so, the failure to detect the pattern must be attributed to some other cause. Perhaps the negative result was due to flaws in the data, mistakes in my analysis, insufficiently sophisticated methods on my part, or a combination of these factors.

In an attempt to shed some light on the source of the negative result, I repeated the same analysis of the 50% date for the *setophaga* genus (warblers) in Maine, since there were about 279,000 of these observed since 1995, but only about 48,000 raptors observed in that time. I only looked at 50% dates, since I had gotten better results for these in my analysis on raptors in *ebird*. The GIFs looked similar to the 50% GIFs for *ebird* raptors, as they did not jump around randomly, but they did not follow a consistent west-southwest motion (or in any direction) either—the GIFs are on my Github page in the section titled “ebird (setophaga) date on which 50% of total fall migration has been observed 2010-2013”. This seems to imply that the negative results for raptors were not a coincidence. Either the methods I used simply are not good enough to be able to demonstrate motion of peaks, the data are not strong enough to demonstrate the pattern, or perhaps, my intuition that this pattern should exist simply is not correct. It could be that birds fly in many directions during their migration in addition to purely in the direction of the average motion. If this were the case, it would give a pattern more complicated than pure west-southwest motion

One possible cause of the negative result is that the spatial scale of Maine simply is not large enough; the pattern might have been covered by random fluctuations in the data. To address this, I repeated my analysis, except this time on six of the *Hawk Watch* sites (one didn’t have enough data), which are spread out across the northeastern U.S. In the future, I am interested in carrying this out on *ebird* data throughout the northeast, but I did not have time to address this question, so I simply looked at the *Hawk Watch* sites. Further, if the problem is an insufficient *quantity* of data as well as too small of a spatial scale, then only analyzing the *Hawk Watch* sites is a good next step.

For the *Hawk Watch* data, I only carried out the peak date analysis, and not the 50% date analysis, since *Hawk Watch* consistently has clearly discernible peaks, unlike *ebird* squares. I did not include Pilgrim Heights, which almost exclusively contains observations from the spring. The animations are linked on my Github page in the section titled “Hawk Watch peak fall dates 2002-2013.” Unfortunately, there once again was no discernible west-southwest motion of peaks. This negative result casts further doubt on whether my methods are capable of making fine distinctions in the data and perhaps on whether the pattern I am looking for exists.

I was pleased that I had successfully run a fairly complex analysis (with a negative result), but I wanted to find at least one positive result. I decided to directly analyze the issue of compatibility between the *Hawk Watch* and *ebird* datasets. I did so by comparing, within each year, the peak date at the *Hawk Watch* site at Acadia in Maine with the peak date of the *ebird* square physically containing Acadia. I looked only at 2010 to 2013, as these were the only years that the *ebird* site had sufficient data. I found an R2 value of 0.994 between the peak dates at the *Hawk Watch* and *ebird* site. The scatter plot is on my Github page in the section titled “Correlation between peak dates for Acadia Hawkwatch and nearby ebird locations.” A possible interpretation of the high correlation I found is that on very small spatial scales, the *ebird* and *Hawk Watch* data are consistent. However, either the data are unable to make any meaningful predictions across space, or the “motion of peaks” pattern simply does not exist.

The high correlation could have occurred by chance, since I only looked at four different years, but this result at least suggests that the datasets agree well at this specific location, and perhaps that they generally agree well across the northeastern U.S. A good way to test this, which I did not find time to do, would be to compare peak dates at each *Hawk Watch* site with peak dates of their respective surrounding *ebird* squares. If the correlations are also very high, it would be strong evidence that the two datasets are consistent.

For my final piece of analysis, I decided to look separately at the peak dates at the six *Hawk Watch* locations over time. That is, I did not try to track motion over space (since I couldn’t find this): I wanted to directly explore whether, at each site, peak dates are getting later over the years, an effect of climate change that has been observed by other means. The multiple R2 values for peak date vs year were 0.4944 at Mount Peter in Warwick, New York State, 0.04084 at Waggoner’s Gap in Carlisle, Pennsylvania, 0.107 at Pilgrim Heights in Truro, Massachusetts, 0.04314 at Allegheny Front in Shade Township, Pennsylvania, 0.4498 for Franklin Mountain in Oneonta, New York, and 0.02763 for Acadia in Mount Desert, Maine. I did not include the site at Wachusett Mountain in Westminster, Massachusetts, since there is not enough data for a correlation calculation to be reliable. The multiple R2 values at Franklin Mountain and perhaps at Pilgrim peak are appreciable, although it is not clear whether taken together these six R2 values constitute a negative or positive result. On the whole it is safer not to make any conclusions from these results. The scatter plots are on my Github page in the section titled “Fall Peak Dates over time at six Hawk Watch locations.”