Kyle Krakoski

TWEET DATA ABOUT DEER IN COLORADO- EXECUTIVE SUMMARY

HYPOTHESIS

In this analysis I will attempt to answer the question: "Are deer tweeted about more on one day of the week than other days in Colorado?" This analysis is a proof-of-concept test to see if a study done by environmental professors at the University of Gloucestershire ("Testing the potential of Twitter mining methods for data acquisition: Evaluating novel opportunities for ecological research in multiple taxa", Hart et al., https://doi.org/10.1111/2041-210X.13063) can be replicated for other animals. In their study, they graphed the number of tweets mentioning winged ants in the UK on Twitter over the course of 3 summers and compared it to official population estimate graphs for the same periods (Figure 1 at the bottom). Interestingly, they found the temporal pattern of the Twitter data graph mirrored the temporal pattern of the official estimate extremely well. It's my belief that this methodology could be used to determine population trends of other animals as well if given access to all tweet data (not just 7 days) and time to refine the search query. Eventually, this could be used by typically resource-stretched government wildlife/nature departments to determine if more thorough testing is necessary. However, for now, I believe my question and analysis will serve as a way to test methods used for a larger analysis.

DATA COLLECTION AND METHODOLOGY

For this test analysis I chose deer. I felt that deer, being fairly common, would have a decent number of opportunities to be seen, yet uncommon enough that someone might tweet about seeing one (especially in an unusual area such as a suburb, which could also be telling about a current population trend). Additionally, I started early to get as much data as I could, which was 14 days of counts. I filtered out retweets and unrelated phrases which greatly skewed the count (such as "Deer Valley", a popular guest ranch in Colorado, and "Promo Code, when a Colorado company used "deer" as a promo code.) Other "bad" tweets (tweets that mentioned the word "deer" without actually being about deer) were left in for simplicity on the Big Data principal that with enough data points, a few bad points shouldn't make a statistical difference. Finally, for location, I tried two methods: search for tweets also containing "Colorado," and search for tweets using the geographical coordinates of the middle of Colorado, with a radius to encompass all of Colorado. I found the second method got me a noticeably greater number of tweets.

Even though I chose deer, I made my Python program generic so it could take a number of different inputs and be personalized by animal, location (state, city, or parks/reserve, by keyword or coordinates), number of days to get counts for, and start date. My program also outputs a csv file containing a tweet count for each day dependent on the number of days chosen to analyze, to be used in R. However, since I could only get seven days of data at once, for my R analysis I used a csv file that I manually inputted 14 days of counts into.

CONCLUSION

In order to test if there was any statistical difference between the count of deer for each day of the week, I used a one-way ANOVA test, which compares the mean count of each day and determines if there is a statistical difference between them. Before doing this test, I assumed four requirements, all necessary for an ANOVA test: Independence of observations, no significant outliers, normality, and homogeneity of variances. Independence of observations might not be guaranteed- the same person could have tweeted about the same deer more than one day. Also, one day only had 17 counts (10 less than the next lowest), which could be considered an outlier. However, I found both these concerns to be negligible; I don't think they're prevalent enough to change the outcome significantly.

My null hypothesis for this ANOVA test is: there is no statistically significant difference between the mean tweet counts of each day of the week. Running the test, I get a p-value of 0.555 and fail to reject the null hypothesis. Therefore, there is no statistically significant difference in mean tweet counts by the day of the week. Had the null hypothesis been rejected, I would've performed a post-hoc Tukey HSD to find out which pairs of means were statistically different. Looking at a scatterplot of the raw data and a bar graph of the average count by day (Figure 2 and Figure 3), and how similar they are, failure to reject the null hypothesis is expected. I believe my major limiter is amount of data I had- 14 days with 2 counts per day is just not enough to form a significant analysis. With more data, I believe a statistical difference is possible.

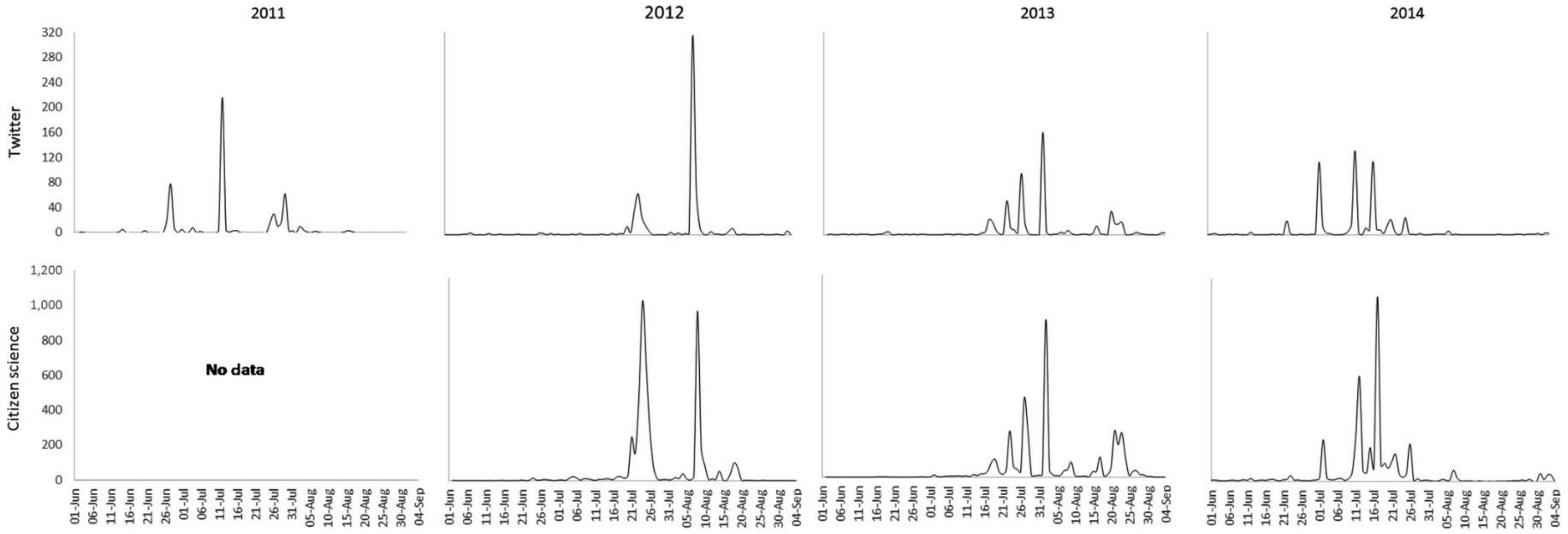
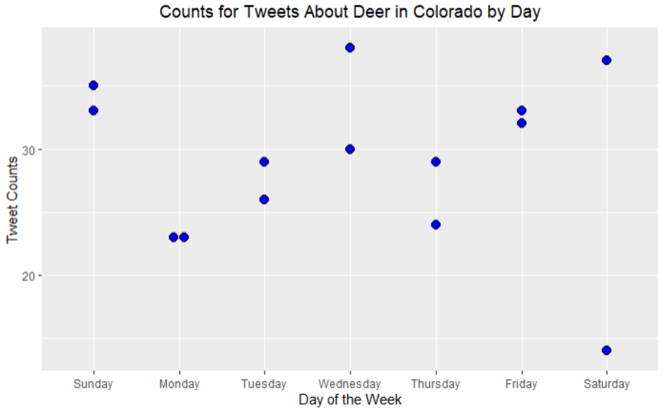


Figure 1- Data from UK-wide study of winged ant emergences derived from Twitter mentions (top row) and a citizen science study (Hart et al., 2017) (bottom row). This study is what I tried to model my data for, if given access to enough tweets.



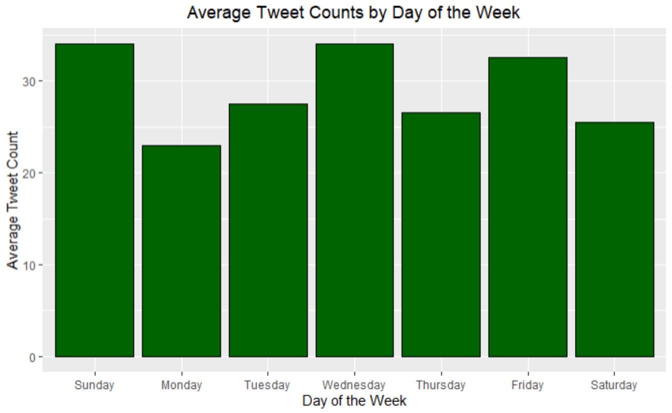


Figure 2- Scatterplot for tweet counts by day over two weeks.

Figure 3- Bar graph of average tweet counts by day