TEAM #87
THE FREE BIRDS

CSE 6242 TEAM POSTER

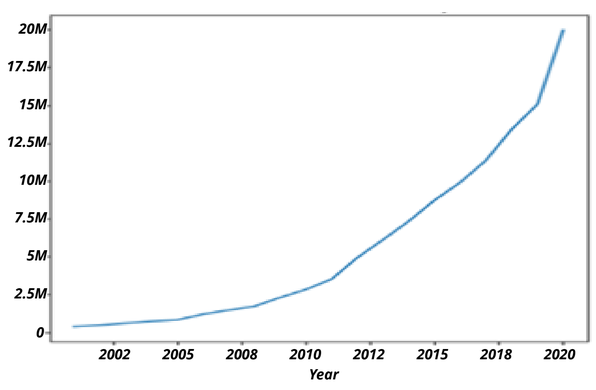
Andrew Del Grosso
Anthony Menninger
Jonathan Wright
Nick Venditti
Ty Martz

Motivation / Introduction

Temperature change has important ecological impacts, but quantifying temperature effects can be difficult. Our team performed a data-driven analysis against temporal eBird bird sightings (*world's largest bird sighting database*) & NOAA weather station temperature readings across North America to answer:

- *Do bird sightings in North America show changes over time that correlate with temperature changes?*
- *Can a citizen sourced dataset be effectively used for analysis?*

eBird Observations by Year



The eBird database of observations has been growing exponentially. Effective analysis would create a significant new tool.

Data

eBird Sighting Data
Cornell Lab Ornithology
Citizen Gathered Data
1 billion+ rows
Downloaded CSV File

Temperature Data
70 years Daily Data
65,000 Weather Stations
Downloaded Fixed Files

amazon web services S3
amazon web services EC2

Loaded into Database
65 species, US Only
210 million rows
188 million rows

MySQL

County, Week, Year Aggregation
Removes Weekday Bias
27 million rows
Manages Weather Station transitions
9 million rows

pandas
GeoPandas
matplotlib

GeoSpatial and Temporal Functions Merge Datasets

Analysis

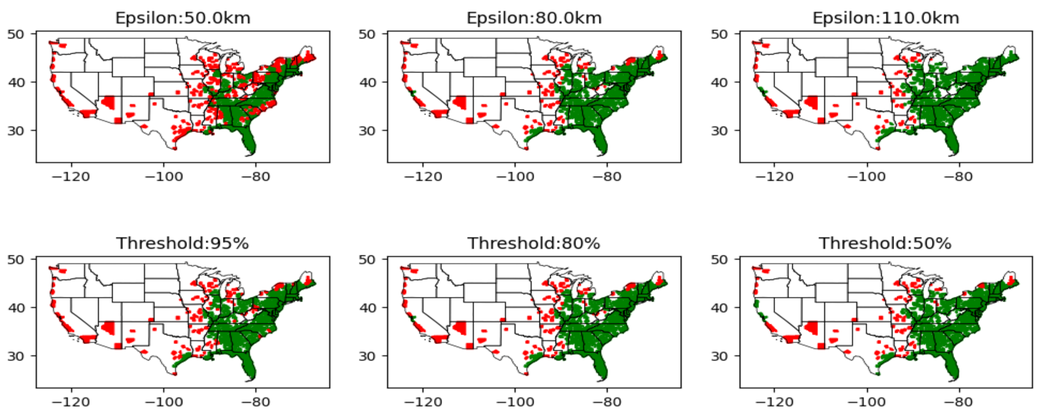
tableau
Scikit learn

Approaches

Noise Reduction |sklearn DBSCAN

A key consideration using a citizen-based data set was the potential noise from irregular observations. We ran outlier analysis to identify and remove noisy data points using the DBSCAN algorithm, weighting data points by total bird sightings / # of observers. For 65 species, we then created a single North America centroid by year and week, as shown to the right.
The DBSCAN algorithm required tuning two key parameters. Epsilon is the maximum distance between nodes in a cluster. Using a Haversine distance calculation for use with latitude and longitude, 80km returned the best result, as seen in below. We also determined the minimum number of points in a cluster using a threshold, also shown below.

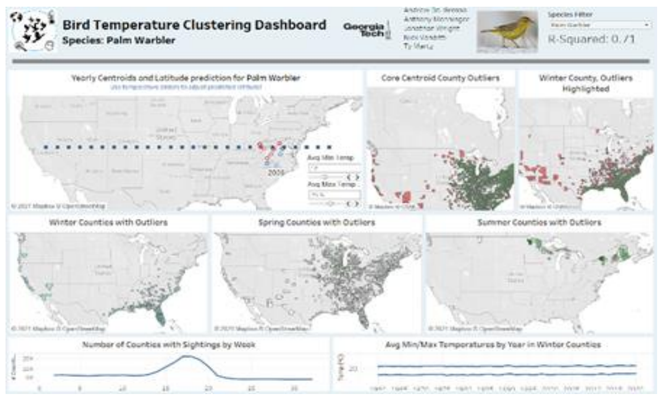
Palm Warbler 2013 DBSCAN Parameter Review



Interactive Visualization

We fed our data into a Tableau visualization dashboard, creating species drill-down capabilities to view centroids with R-squared results, choropleth migration examples, noise charts, temperature trends, and location prediction.

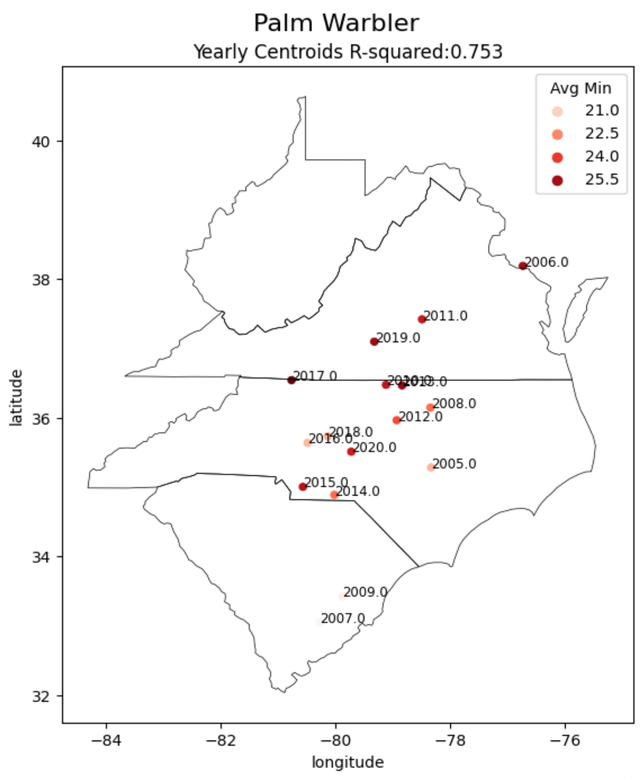
Interactive visualization provides the ability to see predicted latitude, by species, based on min and max temperatures



Newness in Analysis Approach

Our approach used aggregation to remove day of week bias from the eBird sightings and normalized changing weather stations over time. DBSCAN noise reduction significantly improved analysis correlation. We were effective in using a citizen sourced database for many species.

Experiment



Annual Centroids for the Palm Warbler with Temperature. With the final parameters and features, there was a 0.753 R-squared, showing particularly good correlation.

Linear Regression | sklearn Linear Model

We ran linear regression to fit our centroids to determine the correlation between temperature and the northern progress (latitude) in migration. We considered 6 features, with 4 ultimately creating the highest R-squared correlation.

You can see the darker, warmer years are farther north in the chart to the left.

Below are the results for the 65 species reviewed.

Results

R-Squared	Species Count	Percent
> 0.50	18	28%
> 0.25, < 0.50	29	45%
< 0.25	12	18%
Not Enough Data	6	9%

- The analysis was successful, showing good correlations across a range of species. Warmer temperatures cause most birds to migrate earlier.
- Many other factors could impact migration, such as wind, rain, snow, drought, crop conditions
- Lack of effective data was an issue for some species, which is to be expected in a citizen gathered data set

Method Comparison

Our ability to use citizen-gathered data points allowed us to determine insights for a broader geographic region across a broader set of species versus many previous experimental methods.
It is ideal for broad surveys and trend identification but may not be useful for some specific species or time frames due to the variability of the data collection.