

Applied Data Science Capstone Project

1. Introduction

1.1. Background

Birdwatching is becoming one of the main recreational activities in the UK, Netherlands, Denmark, France and Sweden¹. In the USA, the activities related to birdwatching and bird photography have been found the most persistent growth rate form of recreational activity (from 1982 to 2009 287%)². For 2016, the number of bird-watching people in the USA has grown up to 45.1 million (including trips devoted to birdwatching, 16.3 million). Also, economic sectors in Europe and North America, which are related with birdwatching services, ~~in~~ have been showing one of the strongest and steadiest growth rates during last decades.

1.2. Problem

One of the main objectives of birdwatchers is to see during their trips as many species as possible. Due to this, the aim of this capstone project is to provide an analytical overview of the main target locations of birdwatchers in England and Wales at counties level and to analyse whether the pattern of the species observed differs from one another.

The gain of this exercise will be to assist birdwatchers' to plan their birdwatching trips to see as many different species as possible by choosing the trip road through the maximum different counties (or choosing the trip to the county that is most different comparison their home county). Thereby achieving the best satisfactory result from their trip.

Another goal is to provide an overview of whether the species pattern changes with connection with the increase in the number of observations.

1.3. Target audience

In the context of this exercise, the target audience is either the British birdwatchers, or birdwatchers from other countries who are planning the birdwatching trips to England or Wales.

¹ CBI Ministry of Foreign Affairs. (2013). CBI Product Fact Sheet.

² H.K.Cordell, G.T. Green, and C.J. Betz (2009). Long-Term National Trends in Outdoor Recreation Activity Participation. USDA Forest Service.

2. Data

2.1. Data sources

Instead of Foursquare, this work relies on very similar the xeno-canto API. There are observations instead of a venue and instead of a venues category there are name of species in the xeno-canto. The more precise description is available on the web page: <https://www.xeno-canto.org/article/153>.

Reason why the different API is used is to understand and acquire the course material better (and likewise the example work added to 5th week ("Predicting the Improvement of NBA players", Zhenfeng Liu, October 18, 2018) doesn't use also Foursquare).

A geographic information system ArcGIS is also used – to retrieve England and Wales counties names that correspond the coordinates retrieved from the xeno-canto.

For marking counties borders, the GeoJSON file from the web page <https://data.gov.uk/dataset/d6f97a1a-25dc-485c-9af3-0e5681465d77/counties-and-unitary-authorities-december-2016-full-clipped-boundaries-in-england-and-wales> have been used.

From xeno-canto database during the current analysis the following features/data fields have been used:

- **id:** the catalogue number of the recording on xeno-canto
- **en:** the English name of the species
- **loc:** the name of the locality
- **lat:** the latitude of the recording in decimal coordinates
- **lng:** the longitude of the recording in decimal coordinates
- **time:** the time of day that the recording was made
- **date:** the date that the recording was made

2.2. Data cleaning

The original database (were restriction as United Kingdom as country was replaced), retrieved from the xeno-canto have sample with 26706 observations. For this data, it was possible to find a county match through the ArcGIS in the case of 18319 observations. For some reason, the ArcGIS was unable to locate the location counties in the rest of the observations.

The next stage was data extraction from the xeno-canto *loc* field to find additional data about location county and matching results with England and Wales counties list. Thereupon after removing missed values a final sample remained for analysis consist of 23782 observations.

3. Methodology

Data visualisation and k-means clustering from the machine learnings side have been used in the current exercise to follow the course materials.

Visual data exploratory data analysis is presented in paragraph 4. To analyse difference between the counties bird species pattern and look whether these differences are rather due from number of

observations in counties or for example geographical location, result of k-means clustering on the counties level are presented and geographically visualised in paragraph 5.

4. Data Description

There are 266 unique species in a final sample. As expected, most observations (recordings) have been made in the morning (see following Figure 1).

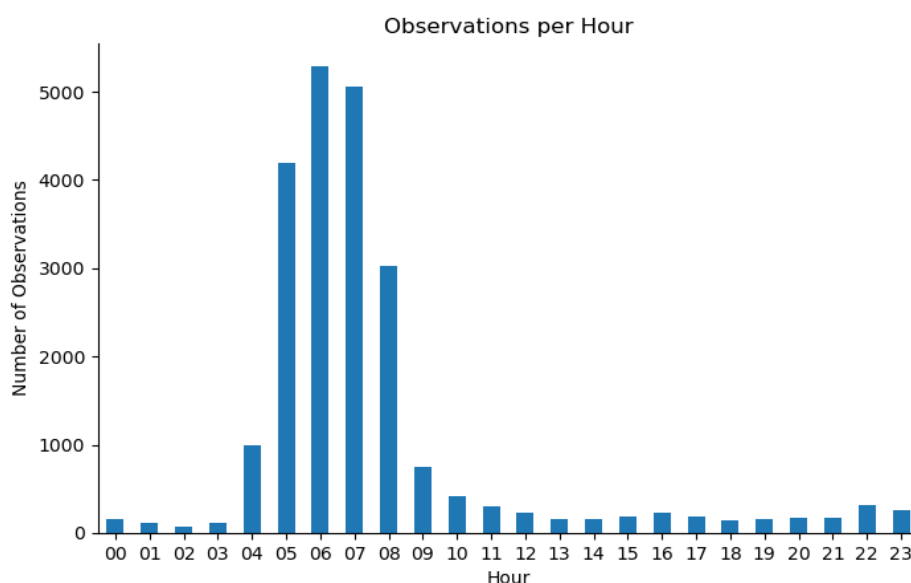


FIGURE 1 DAILY DISTRIBUTION OF OBSERVATIONS (RECORDINGS)

Also, in line with expectations, most observations (recordings) have been made in the spring (see following Figure 2).

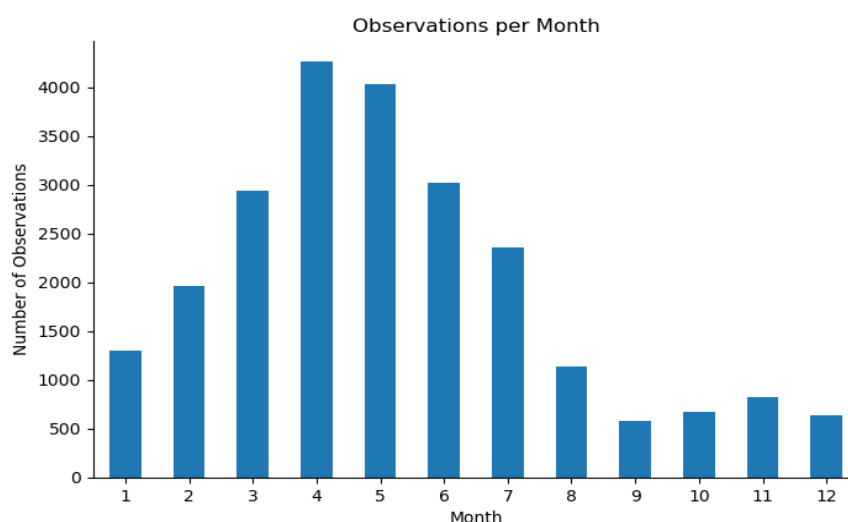


FIGURE 2 YEARLY DISTRIBUTION OF OBSERVATIONS (RECORDINGS)

However, the distribution of observations between counties is extremely uneven (see following Figure 3 and 4). Unfortunately, that makes the problem solution of the current exercise (marked in section 1.2) quite fragile and questionable.

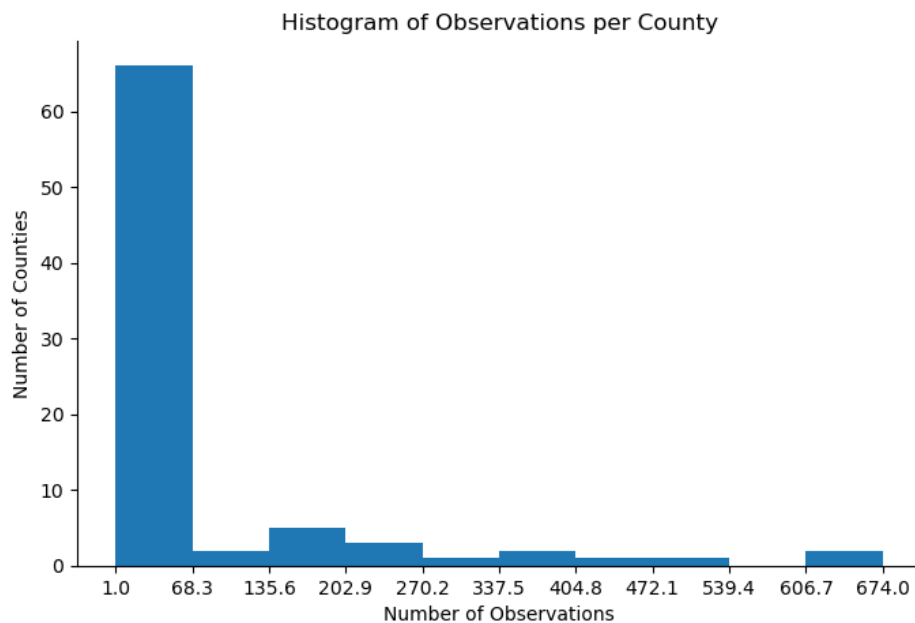


FIGURE 3 HISTOGRAM OF OBSERVATIONS (RECORDINGS) PER ENGLAND AND WALES COUNTIES. NORTH YORKSHIRE COUNTY ARE EXCLUDED FROM HISTOGRAM AS IT HAVE 17781 OBSERVATIONS OUT OF 23782 (TOTAL NUMBER) AND SO SIGNIFICANTLY OUT OF RANGE.

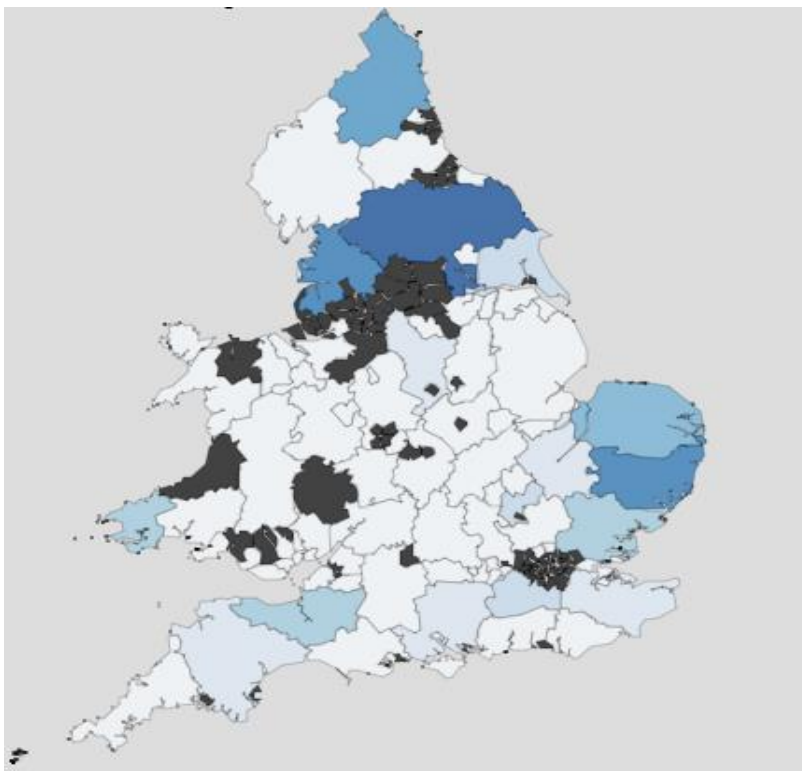


FIGURE 4 GEOGRAPHICAL DISTRIBUTION OF OBSERVATIONS (RECORDINGS) PER ENGLAND AND WALES COUNTIES. DARKER BLUE MARKED MORE RECORDINGS PER COUNTY. BLACK MARKED COUNTY WITH MISSING OBSERVATIONS.

5. Results

To analyse difference between the counties bird species pattern and look whether these differences are rather due from number of observations in the counties or for example the geographical location, the England and Wales counties divided into five clusters using k-means clustering. A result is presented in the following figure 5.

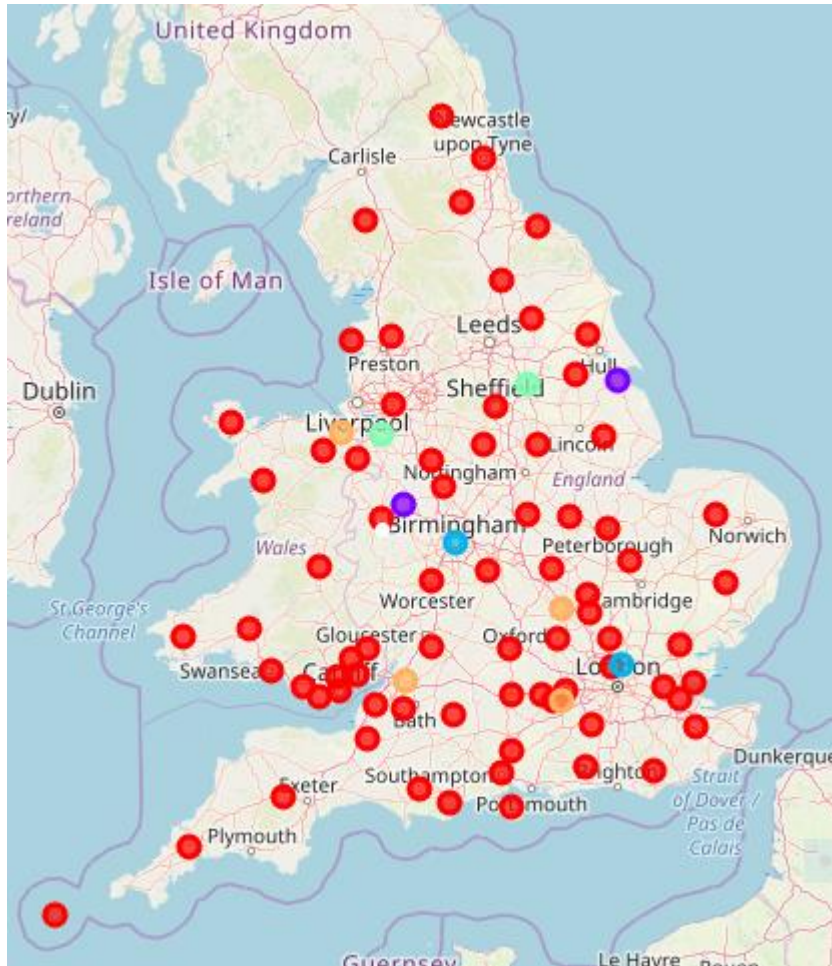


FIGURE 5 GEOGRAPHICAL PRESENTATION THE ENGLAND AND WALES COUNTIES DIVIDED ON FIVE CLUSTERS ON THE BASIS OF OBSERVED (RECORDED) BIRD SPECIES PATTERN.

As shown in figure 5, there is no remarkable distinction between the counties (in recorded bird pattern), based on geographical location doesn't emerged.

If to compare figure 5 and figure 4 it can be seen that no remarkable correlations between the number of recording per county and the observed bird pattern doesn't also emerged.

74 of the 84 counties (where we have a data) belonged to the first cluster. Two counties belonged to the second, third and fourth cluster and four counties belonged to the fifth cluster. The counties which belonged to the second or fifth cluster having only one observation (recording) per county. Therefore, it is possible to mark them as the outliers. A similar assessment is also valid for the third and fourth cluster, where is only two or three observations (recording) per county.

If to look the cluster species pattern (see appendix), then all clusters described similarly by birds which are broadly rather common to cultural landscape.

Reason of that kind results is probably quite uneven distribution of the observation per counties which is described in figure 3.

6. Discussion

The main result of current exercise is that despite representative overall number of recordings in xeno-canto database according England and Wales, the distribution of these recordings on counties level is too uneven to analyse difference between those counties.

Therefore, the main recommendation emerged of the current exercise is to use in further analyses about difference between England and Wales (or as the United Kingdom as whole) counties birds species pattern other public science open databases (which is also possible to find). The other recommendation is if to limit this exercise only in the counties level, it can be restrictive. If it possible to find a geoJSON file which describe rather borders between natural/ecological landscapes, it will be probably more useful.

7. Conclusion

The gain of this exercise was to assist birdwatchers' to plan their birdwatching trips, by analysing difference between bird species pattern between the England and Wales counties. Another goal was to analyse whether the species pattern changes with connection with the increase in the number of observations.

For this purpose, the public science open database xeno-canto API was used. The GeoJSON file from the web page <https://data.gov.uk/dataset/d6f97a1a-25dc-485c-9af3-0e5681465d77/counties-and-unitary-authorities-december-2016-full-clipped-boundaries-in-england-and-wales> was used for marking counties borders.

After data cleaning a final sample remained for analysis consisted of 23782 observations (recordings). Data visualisation and k-means clustering from machine learnings side was used to follow course materials in current exercise

As exercise result, no remarkable distinction between counties (in recorded bird pattern), based on geographical location doesn't appeared. Also, no remarkable correlations between the number of recording per county and the observed bird pattern doesn't emerged. If counties divided into clusters, vast majority of counties belonged to one cluster. Other clusters describing rather sample outliers. If to look the cluster species pattern, then all clusters described similarly by birds which are broadly rather common to the cultural landscape. Reason of that kind results is probably quite uneven distribution of the observation per counties which is described in figure 3.

The main result of the current exercise is that despite representative overall number of recordings in the xeno-canto database according England and Wales, the distribution of these recordings on the counties level is too uneven to analyse difference between those counties.

PLEASE NOTE: If you find that the methodology or results are not clearly stated and explained, the discussion or conclusion have some problem or similar – please give me feedback on this regard. Thank you!

Appendix: England and Wales counties observed bird pattern clusters and most common birds of them

[illegible]