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 Sweden[1]. 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.

PLEASE NOTE: If you find that the backgrounds or problem is not clearly explained, the target audience is not clearly stated, or the description of the data have some problem – please give me feedback on this regard. Thank you!