Environmental Data Analytics
Project 2 — Birds Occurrence Data in Kenya Analysis

Anton Zaitsev, Othmane Mahfoud — University of Luxembourg

November 2, 2024

Introduction

The Birds Occurrence Data in Kenya dataset documents bird sightings across the Republic of Kenya, collected from June 1, 2017, to May 16, 2018. The dataset provides valuable insights into bird diversity across Kenya. The data is curated and preprocessed by experts at the National Museums of Kenya to ensure data quality. This analysis focuses on identifying clusters of bird sightings, regardless of bird species, and creating a bird species richness map to visualize the geographic distribution of bird diversity across Kenya, with the aim of identifying potential biodiversity hotspots and areas with lower species richness in Kenya.

Data

Species Data

The bird species occurrence data was downloaded from the GBIF dataset page. We used the following columns for this analysis: SPECIES, SPECIESKEY, DECIMALLATITUDE, DECIMALLONGITUDE, DAY, MONTH, YEAR, INDIVIDUALCOUNT, and OCCURRENCESTATUS.
The data was filtered to include only records where OCCURRENCESTATUS is marked as PRESENT, ensuring that all entries represent confirmed bird sightings.

Kenya Boundary Data

The Kenya boundary shapefile was downloaded from GeoBoundaries to overlay the species richness map on Kenya's geographic area. After loading the shapefile, we confirmed its projection in WGS84 (EPSG:4326) to ensure compatibility with the species data's latitude and longitude coordinates (see Figure 1).

Clusters of Bird Sightings

The goal of this section is to analyze bird sighting data in Kenya, identify significant clusters of sightings, and visualize these clusters on a map. By clustering bird sightings, we aim to highlight regions with higher bird activity, potentially identifying ecologically important areas or hotspots for bird populations. The clustering was performed using DBSCAN, a density-based clustering algorithm, due to its suitability for geographic data with varying densities.

Data Pre-processing

The dataset contained alongside bird sightings, geospatial data (latitude and longitude) of each sighting. In order to prepare our data to be be used by our DBSCAN model we did the following:

1. Extract decimalLatitude and decimalLongitude and remove any missing data.
2. Convert the data type to numeric which is suitable for our algorithm.

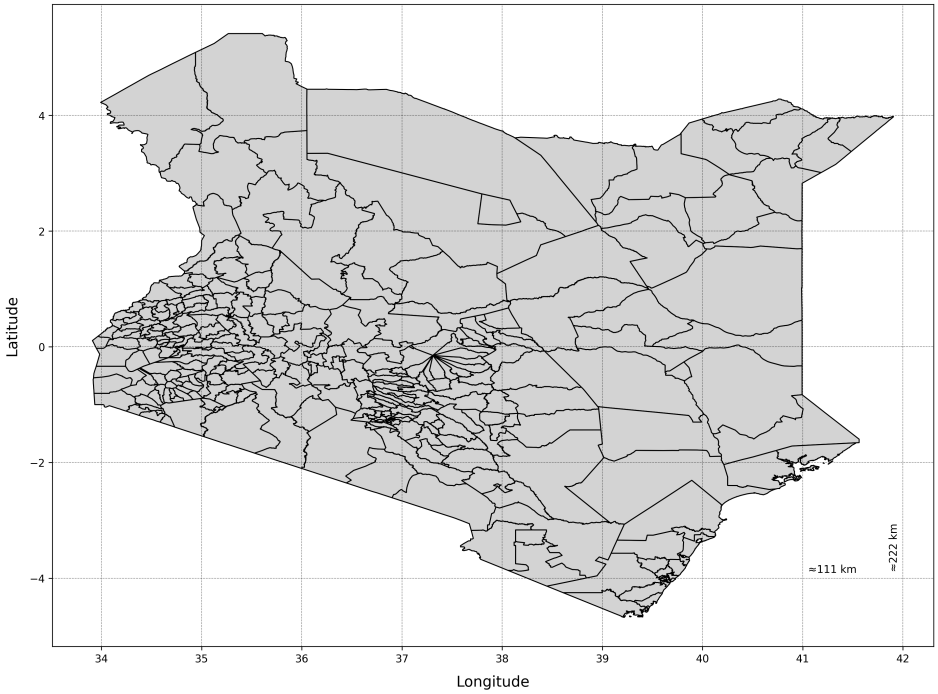

Figure 1: Kenya's Boundary Map.
Map includes county and district boundaries.

Model and Hyper-Parameter Selection

Since we do not have a pre-established idea on the number of clusters in our data, we opted for DBSCAN (Density-Based Spatial Clustering of Applications with Noise) which does not require setting a number of clusters before running it unlike K-Means where you need to provide the number K of clusters. DBSCAN is also perfectly optimized to identify clusters based on density, which is our goal in this exercises.

Instead of K, DBSCAN require to set two hyper-parameters:

1. **Epsilon** The maximum distance between two points to be considered in the same neighborhood.
2. **Minimum Samples** The minimum number of samples to form a cluster.

For minimum samples, we randomly settled on 50 as a fair minimum number to consider a cluster, while we optimized for epsilon. To do this we used the **K-distance Plot** method which consists of calculating for each point the distance to its k-th nearest neighbour (k being the minimum number of samples for a cluster), sort and plot these distances and find an elbow (like in the elbow method for K-Means) where the distance sharply increases which represents the ideal value for epsilon. In our case the optimal value of epsilon was around 0.1 (see Figure 2).

Applying DBSCAN and Mapping the Clusters

Applying DBSCAN with the selected values for our hyper-parameters we were able to efficiently cluster bird sightings and plotting them on Kenya's map.

To achieve the result in Figure 3 we did the following:

1. We used GeoPandas and Matplotlib to display the map, loading a shape file containing Kenya's administrative boundaries.
2. Each cluster was represented with a unique color, and noise points were plotted in black to distinguish them from the main clusters.

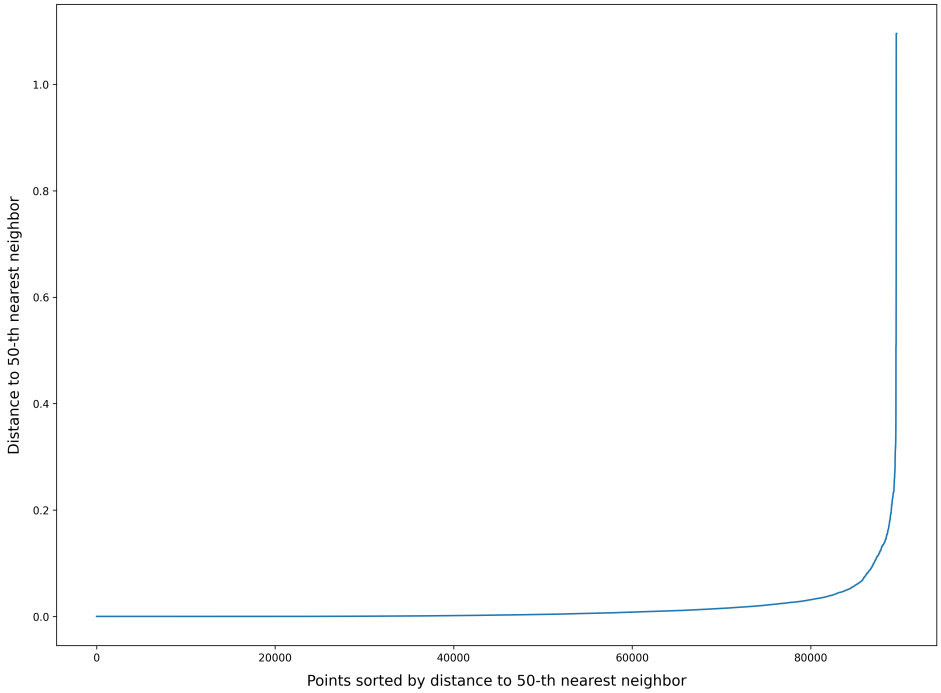

Figure 2: K-Distance Plot
For optimal selection of the "eps" DBSCAN parameter.

Conclusion

The resulting clusters' map allowed us to identify the orange labeled cluster area as a significant bird activity zone in Kenya. Regions like these with dense clusters may indicate areas with high biodiversity or habitats that support large bird populations, potentially guiding conservation efforts or further ecological studies.

Bird Species Richness Heatmap

A species richness map counts unique species at a location, rather than individual sightings. To interpolate the count of the species across the whole map we can use, for example, spatial interpolation techniques, such as kriging. Ordinary kriging is a geostatistical interpolation method that estimates unknown values at specific locations based on spatial autocorrelation. The idea that nearby points are more similar to each other than distant points. However, in a species richness map, simpler interpolation methods like linear, cubic, or nearest-neighbor interpolation should be sufficient. Kriging is more appropriate when we have strong spatial autocorrelation in the data, which should not be the case for species richness. We thus used cubic interpolation, which also produces smoother interpolation values compared to linear and nearest-neighbor interpolation techniques.

The species richness map was created using the following steps:

1. **Data Definition**: We loaded species data and geographical area data.
2. **Grid Creation**: We defined a grid across the study area by dividing Kenya into evenly spaced cells based on latitude and longitude ranges, ensuring coverage of the full geographic area.
3. **Species Count Aggregation**: For each cell, we initialized a set to store unique species within that cell. As we iterated through each data point in the species data, we added each species sighting to the corresponding cell's set, resulting in a count of unique species for each cell.
4. **Interpolation**: The species counts were then interpolated across the grid using cubic interpolation. This smooths out abrupt changes in species richness, providing a more visually cohesive map. To further smooth our the interpolation heatmap we applied Gaussian filter with $\sigma = 3$.

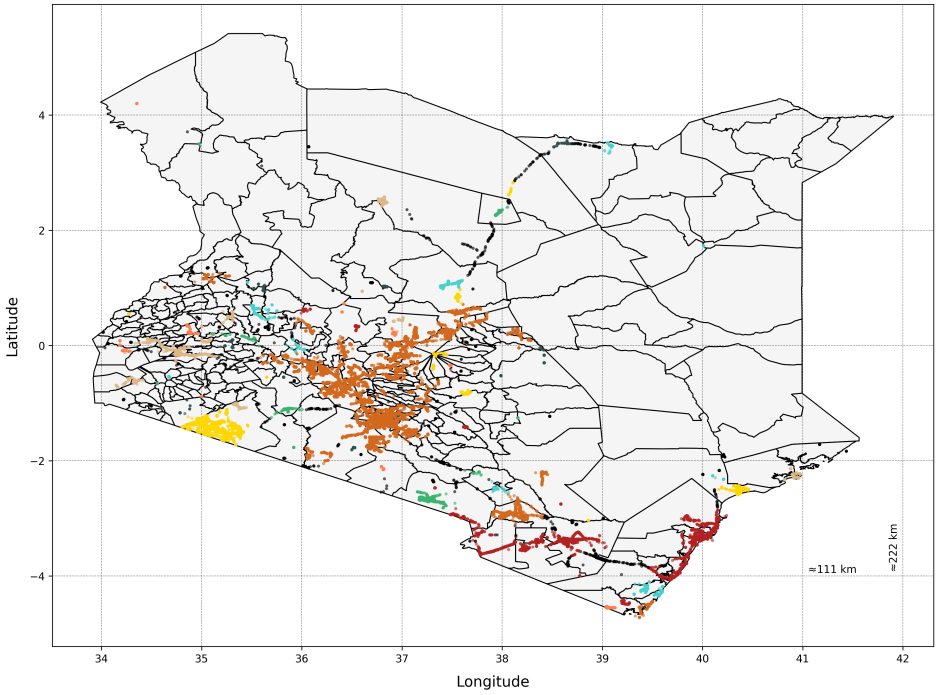

Figure 3: Bird Sighting Clusters in Kenya.
The most prominent cluster is shown in orange.

5. **Masking**: The interpolated data that is outside Kenya's boundary was masked with NaN so that the heatmap values are not shown outside the Kenya's boundary.

6. **Overlay and Visualization**: We overlaid the interpolated species richness map on Kenya's geographic boundaries to provide spatial context.

Heatmap Analysis

The bird species richness heatmap (see Figure 4) reveals patterns in biodiversity distribution across Kenya, with notable regional variations:

- Nairobi (Capital City) and Nanyuki: The highest species richness is concentrated in and around Nairobi and Nanyuki, with values ranging from 300 to over 500 unique species. This concentration may be attributed to the high level of human activity and the presence of conservation and monitoring efforts, which could facilitate more frequent bird sightings and better-recorded observations.

- Eldoret and the Tanzania Border: Elevated species richness is also observed near the city of Eldoret and along the border with Tanzania, with values around 300 species. This pattern may indicate that diverse habitats along this boundary support a wide range of bird species or reflect targeted conservation and study efforts in these areas.

- Coastal Regions (e.g., Mombasa): Coastal cities like Mombasa show high species counts, approximately 300 species, likely due to the unique coastal and marine ecosystems, which attract both local and migratory bird species.

- Inland Regions: Significant species richness is observed further inland, although with lower counts than major urban and coastal areas. Regions like Kakuma (bordering Uganda and South Sudan), Lake Turkana, Marsabit, and Wajir support around 100 species, showing the ecological diversity of Kenya's interior landscapes.

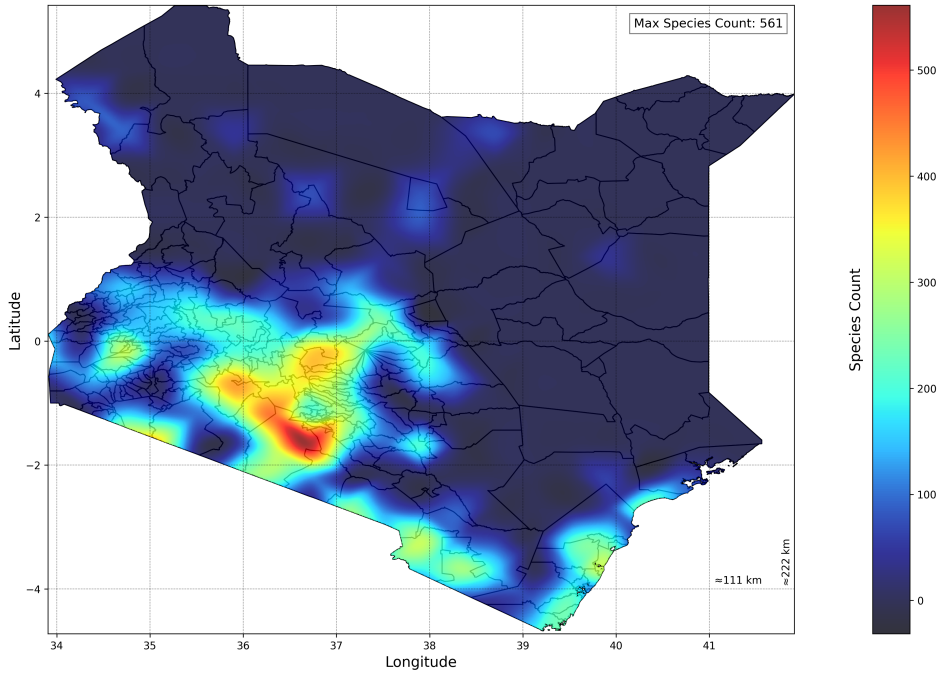

Figure 4: Kenya Bird Species Richness Heatmap.
Higher species richness values are shown in dark red, lower values are shown in dark blue.