Exploring White Stork Migration Patterns with Dynamic Bayesian Network Analysis

Jana Nikolovska
Master's Degree in Artificial Intelligence, University of Bologna
jana.nikolovska@studio.unibo.it

March 11, 2024

Abstract

The aim of this study is to develop a Dynamic Bayesian Network (DBN) to understand the migration patterns of the white stork, leveraging climate and vegetation indicators to uncover the underlying relationships. The model is constructed using the pgmpy library in Python, fitting it with time-series data gathered from various sources after integration and preprocessing. The fundamental concepts are explored, including Markov blanket, Independence, Sampling, and Inference within the context of the DBN, along with calculated Conditional Probability Distributions (CPDs) and Discretized Factors.

Introduction

Domain

The subject of study in this project is the Eastern flyway routes of the white storks (Ciconia ciconia).[1] This avian phenomenon encompasses the spring migration of adult white storks, commencing their journey from wintering grounds in sub-Saharan Africa and traversing diverse landscapes en route to breeding grounds in northern Europe. This annual migration underscores the adaptive strategies of migratory birds, shaped by various environmental cues. To explore this phenomenon, we utilize the dataset [1] compiled by Rotics et al. in 2018 [2], which comprises time-series records for 35 adult white storks over a five-year period (2012-2016). This dataset is augmented with climate information sourced from the Open-Meteo API[3] and Normalized Difference Vegetation Index (NDVI) derived from Google Earth Engine.

Aim

The aim of the project is to model and analyse the relationship between the climate and vegetation indices and the precise geospatial information detailing the locations of the white storks throughout their migration. Additionally, we try to understand the dependency of the indicators, observe how different indicators influence the migration process and try to predict based on observed data for the indicators whether a white stork is in process of migration or not.

[1] https://en.wikipedia.org/wiki/White$_{s}tork$

Method

To reach the goal we set, the data is modeled using the Dynamic Bayesian Network model implemented with the pgmpy library in Python. The functions the library offers are taken into advantage to do a deeper analysis of the network, looking into Markov blankets, Independence, Sampling and Inference. We asses the Conditional Probability Distribution learned from fitting the data into the model and derive conclusions

Results

The analysis indicates that the model struggles to accurately predict whether a white stork is migrating, with predictions resembling random guesses. This suggests that the Bayesian network structure may not effectively capture the relationships between variables or that the selected variables and data lack relevance for the problem. Hence, exploring alternative modeling approaches that better capture the system's dynamics is warranted.

Model

Data pre-processing

As stated previously, for the purposes of the project we obtained Eastern flyway migration for white storks data[1][2]. During the pre-processing, most of the attributes of the original dataset were removed, keeping only the stork's identification number, timestamp, longitude and latitude information. Additionally, records with missing data were removed. The total number of records in the processed dataset reached approximately 1 035 000.

The input data is then augmented with a new feature called NDVI, or Normalized Difference Vegetation Index. Using coordinates and timestamps from the input data, we retrieve Landsat imagery from Google Earth Engine API[4], filter it to the specified time range, and calculate the average NDVI value for each location.[2]

To further enrich the dataset, we incorporated climate indicators sourced from the Open-Meteo API[3] for

[2] Code for this function was taken from `https://gis.stackexchange.com/`[5]

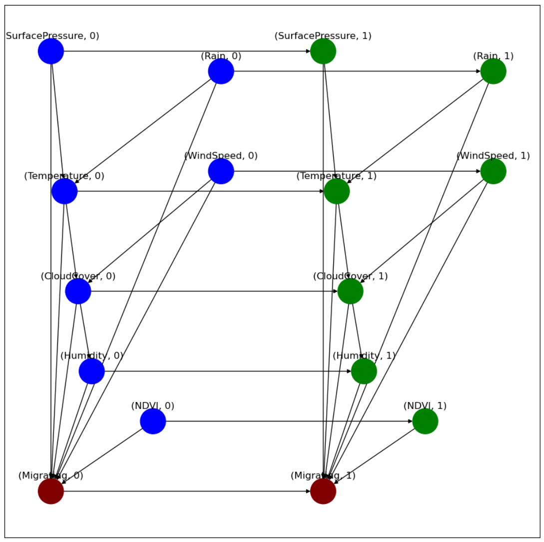

Figure 1: Proposed Dynamic Bayesian network

each location and timestamp. The climate indicators include Relative Humidity, Cloud Coverage, Temperature, Wind Speed, Rain(mm) and Surface Pressure.

In the next step we discretize the the previously mentioned climate indicators, categorizing them into 2 bins or, only in the case of the Temperature indicator, into 3 bins. Since there is no stated or suggested way to discretize based on the domain knowledge, a categorization is proposed such that the bins(categories) contain the same amount of samples. The longitude and latitude features are grouped together and discretized by using the Machine Learning clustering technique KNeareast-Neighbours with preset value for number of clusters to 5.

The data is afterwards grouped for every stork identification number and in a time span of a week, calculating the distance flown in the selected time span. By discretizing, similar to the climate indicators, this final metric we get the attribute Migrating. After preprocessing the input data size is reduced to approximately 800 records.

Modelling the data in a Dynamic Bayesian Network

Since the data is in the format of time-series the chosen model is a Dynamic Bayesian Network. The proposed model (Figure 1) is derived by analysing the domain. However, this is just a proposed structure and one of the many possible for the obtained data. It is important to note that in this Bayesian Network, each variable represents a key meteorological or environmental concept, and their relationships are closely intertwined, influencing one another in complex ways, therefore, the network's design can be adapted and refined to capture different facets of these interactions. For the purpose of this project, we will proceed only with the above mentioned model. For the implementation of the model and visualizations and analysis the Python libraries pgmpy[6] and networkx[7] are used.

Once the DBN model is initialized using pgmpy with the variable dependencies represented as causality edges, the data is fitted to the model.

Analysis

Experimental setup

The model, described in the previous section, is then analyzed from the following aspects:

- Calculated Conditional Distribution Probabilities (CPDs) for each of the variables and Discretized Factors.
  These metrics are calculated using the functions provided in the pgmpy library. Once computed, it can be observed that the discretized factors have around 0.5 value, suggesting that the model's predictions are close to random guessing. In contrast to what we were hoping to achieve, this model exhibits uncertainty or lack of strong correlation between variables.
- Calculated Independencies and Markov blankets.
  These metrics are also calculated using the functions provided in the pgmpy library with the intent to better understand the proposed network.
- Simulating new records.
  Again, using the functions provided in the pgmpy library, we sample timeseries data from the previously specified model as a demonstration. This approach can be useful for solving other problem in this domain.

Results

Going over the results of the analysis, especially the computed discretized factors, regarding the model's ability to predict whether a white stork is in the process of migrating it is concluded that the model's predictions are close to random guessing. Unfortunately, due to the lack of strong correlation between variables, there is limited predictive power in the DBN.

Conclusion

Despite initial expectations of correlation between climate, vegetation indices, and white stork migration along Eastern flyway routes, our attempt to model these dependencies using a Dynamic Bayesian Network fell short. The data collected and analyzed did not effectively capture the underlying relationship between these factors. The findings suggest potential challenges either with the proposed Bayesian network structure or with the relevance of the selected variables and gathered data for addressing the problem at hand. This prompts consideration for an alternative approach, one that may better align with the underlying dynamics and dependencies within the system under study.

Links to external resources

- Code availability: https://github.com/jananikolovska/StorkMigrationPatternsWithDBN
- The pre-processed data used as an input to the network, as well as the intermediate files, is not available on the gihub repository. If needed contact jana.nikolovska@studio.unibo.it

References

1 Publicly available data: https://www.movebank.org MovebankID = 560041066

2 Rotics S, Kaatz M, Turjeman S, Zurell D, Wikelski M, Sapir N, Eggers U, Fiedler W, Jeltsch F, Nathan R. 2018. Early arrival at breeding grounds: causes, costs and a trade-off with overwintering latitude. J Anim Ecol. 87(6):1627-1638. https://doi.org/10.1111/1365-2656.12898

3 Open-Meteo API https://open-meteo.com/en/docs/

4 Google Earth Engine https://earthengine.google.com/

5 GIS stackexchange community https://gis.stackexchange.com/questions/360278

6 Pgmpy documentation: https://pgmpy.org/models/dbn.html

7 NetworkX documentation https://networkx.org/