University of Information Technology - VNU-HCM

Clustering Animal Migration Trajectories for Conservation Planning

Supervisor: Vo Nguyen Le Duy

Group members:

Hoang Cong Chien - 22520155
Phan Thanh Dang - 22520193
Tran Dinh Khanh Dang - 22520195
Duong Dinh Phuong Dao - 22520202
Tran Quang Dat - 22520236
Nguyen Huu Duc - 22520270

May 9, 2025

Contents

0.1 Motivation

The Importance of Studying Animal Migration. Migration is a vital natural behavior for many animal species, allowing them to access more suitable habitats according to seasonal changes, such as areas with milder climates, abundant food sources, and safe breeding grounds. However, migratory patterns are being significantly altered by factors such as climate change, habitat loss, and human activities. These disruptions not only threaten the survival of individual animals but also impact the structure and functioning of entire ecosystems. Gaining a thorough understanding of migration routes and identifying critical stopover sites enables conservationists to develop more effective habitat protection strategies, particularly for areas that play a key role in supporting migration.

The Urgency of Conservation. Stopover and congregation sites of migratory species are often strategically important—serving as places for resting, foraging, and energy replenishment. If these habitats are encroached upon or degraded, the entire migratory chain can be disrupted. Therefore, identifying these critical areas through clustering analysis of flight trajectory data is a crucial step in guiding conservation efforts and ensuring the long-term sustainability of migratory routes.

0.2 Related Problems

0.2.1 Trajectory Clustering

Trajectory clustering refers to the task of grouping multiple movement trajectories (typically sequences of GPS points) based on spatial and temporal characteristics. In the context of animal migration, this technique allows researchers to identify typical migration paths, group individuals with similar movement behaviors, and detect intersection or stopover points where multiple individuals gather.

One of the main advantages of trajectory clustering is its ability to highlight ecological "hotspots" that are critical for conservation efforts. It also supports monitoring changes in migration routes caused by climate change or habitat destruction. However, this method may be sensitive to data noise and requires careful preprocessing to align trajectories of different lengths and sampling rates. Additionally, the choice of clustering algorithms and similarity metrics can significantly affect the outcome, requiring domain expertise to tune appropriately.

0.2.2 Movement Pattern Mining

Movement pattern mining focuses on identifying recurring movement behaviors, such as seasonal migrations, repeated stopovers, or correlations between movement and environmental variables. This approach helps in uncovering predictable behavioral patterns that are consistent over time or across individuals.

The strength of movement pattern mining lies in its potential to predict future migration behaviors and analyze the ecological role of specific locations within the migration chain. It is particularly useful for long-term monitoring and planning. However, its effectiveness relies heavily on the availability of large, high-quality historical data. Additionally, distinguishing meaningful patterns from random variations or noise can be challenging, especially in dynamic environments.

0.2.3 Anomaly Detection

Anomaly detection aims to identify unusual or unexpected movement behaviors in migration data. Examples include individuals deviating from the main group, taking entirely new paths, or pausing for abnormal durations at certain locations. These anomalies may indicate ecological threats, navigational errors, or health issues.
The primary benefit of anomaly detection is its capacity for early warning, allowing conservationists to quickly respond to potential threats such as poaching, entrapment, or environmental degradation. It supports proactive intervention and habitat assessment. On the downside, anomaly detection can be prone to false positives, especially when natural variation in behavior is high. It also requires clearly defined baselines of "normal" behavior, which may not always be available for all species or regions.

0.3 Problem: Clustering Animal Migration Trajectories for the Greater White-fronted Goose

0.3.1 Input and Output Description

The input to the Clustering Animal Migration Trajectories problem consists of real-world GPS data collected from the migration journeys of animals. Each trajectory is typically represented as a time-ordered sequence of latitude and longitude points, possibly enriched with timestamps and additional contextual information such as altitude or speed. The output of the system is one or more clusters of migration trajectories, where each cluster groups similar movement patterns based on spatial and temporal similarity. These clusters represent common migratory routes or behavioral patterns, enabling researchers to identify key pathways and prioritize areas for conservation.

0.3.2 Practical Applications for the Greater White-fronted Goose (Anser albifrons)

Understanding and analyzing the migration trajectories of animals, especially birds like the Greater White-fronted Goose, is essential for wildlife conservation in the face of rapid environmental changes. Clustering these migration trajectories allows researchers to group similar movement patterns, uncover hidden structures in the data, and extract biologically meaningful insights. By identifying groups of similar routes, we can better understand how animals interact with their environment over space and time. This analysis plays a critical role in shaping conservation strategies, anticipating future changes, and maintaining ecological balance. Some key real-world applications of migration trajectory clustering include:

- **Identifying Common Migration Routes.** By grouping similar paths, researchers can detect major flyways used by many individuals, which are vital for large-scale conservation planning.
- **Detecting and Protecting Critical Stopover Sites.** Trajectory clusters often highlight regions where geese regularly rest and refuel. These hotspots are essential for survival and thus become priority zones for habitat protection.
- **Monitoring Migration Shifts Due to Climate Change.** Comparing trajectory clusters across years helps detect shifts in migration behavior that may be driven by rising temperatures, altered landscapes, or seasonal disruptions.

- **Supporting Cross-border Conservation Efforts.** Since the migration paths of Greater White-fronted Geese span multiple countries, trajectory clustering supports international cooperation and policymaking to protect key regions along the routes.
- **Simulating and Predicting Future Migration Patterns.** Historical clusters can serve as a foundation for predictive modeling, enabling simulation of how geese might migrate under changing environmental scenarios.
- **Early Detection of Abnormal Migration Behavior.** Outlier trajectories—those that don't fit into any known cluster—can signal distress, disorientation, or human-induced threats, prompting early intervention.

0.4 Common Algorithms for the Clustering Animal Migration Trajectories Problem

0.4.1 K-Means Clustering

K-Means is a widely-used partitioning algorithm that divides data into K clusters by minimizing the squared Euclidean distance between each point and the center of its assigned cluster. It is particularly effective for datasets where the clusters are spherical and well-separated.
Advantages and Disadvantages. K-Means is fast, easy to implement, and works well with large datasets that contain convex-shaped clusters. However, it requires the number of clusters K to be defined beforehand, making it less flexible in exploratory analysis. It is also sensitive to noise and outliers, and struggles with identifying non-linear or curved migration paths, which are common in animal trajectory data.

0.4.2 DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

DBSCAN is a density-based clustering algorithm that groups points closely packed together while marking sparse regions as noise. It identifies clusters based on local density, without requiring the number of clusters in advance.
Advantages and Disadvantages. DBSCAN excels at detecting clusters of arbitrary shapes, such as curved or branching migration routes, and is robust against noise and outliers. However, its performance heavily depends on the careful tuning of parameters like neighborhood radius and minPts (minimum number of points to form a dense region). Additionally, it may perform poorly on datasets with varying densities.

0.4.3 HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise)

HDBSCAN is an extension of DBSCAN that builds a hierarchy of clusters and selects the most stable ones based on density. It allows for more flexibility in handling clusters with varying densities and structures.
Advantages and Disadvantages. HDBSCAN automatically determines the number of clusters and supports complex patterns, such as migration paths that vary seasonally. It also enables soft clustering by assigning membership probabilities to data points. However, this algorithm requires careful parameter tuning and is computationally more intensive than standard DBSCAN, especially for large trajectory datasets.

0.4.4 Reason for Choosing DBSCAN

DBSCAN is particularly well-suited for clustering animal migration trajectories due to its ability to detect clusters of arbitrary shapes and its robustness against noise and outliers. Unlike K-Means, which assumes spherical clusters, DBSCAN can effectively identify non-linear and curved migration paths that are common in real-world GPS data. Moreover, the migration routes of animals like the Greater White-fronted Goose often include branching or irregular stopover points, which DBSCAN can capture without requiring a predefined number of clusters. Its density-based approach also makes it resilient to minor GPS errors or sparse observations, which are typical in wildlife tracking. These strengths make DBSCAN a practical and adaptive choice for uncovering ecologically significant patterns in migration data.

0.5 Application

0.5.1 Dataset

This dataset contains high-resolution GPS tracking data of Greater White-fronted Geese (Anser albifrons) across their migration routes in the Western Palearctic region. Collected using satellite telemetry, the dataset captures detailed spatiotemporal movement patterns of individual geese, enabling the study of migration behavior, habitat usage, and environmental influences. Each row in the dataset represents a recorded location at a specific timestamp. The main columns include:

Feature	**Meaning**
event ID	Unique identifier for each time-location record.
visible	Indicates if the event is visible or marked as an outlier.
timestamp	Date and time when sensor measurement was taken.
longitude (decimal degree)	Geographic longitude of the recorded location.
latitude (decimal degree)	Geographic latitude of the recorded location.
ground speed	Estimated ground speed at the recorded point (m/s).
heading	Direction of movement in degrees clockwise from north (0–360).
height above mean sea level	Estimated height above sea level (meters).
sensor type	Type of sensor used (e.g., GPS, barometer, acceleration).
taxon	Scientific name of the species with the deployed tag.
tag ID	Identifier for the tag device.
animal ID	Identifier for the individual animal.
study	Name of the study containing the data in Movebank.

Table 1: Description of features in the dataset

The following plots 1,2, 3, 4 provide some basic visualizations of the dataset to better understand the migration patterns.

0.5.2 Data Preprocessing

- **Dropped unnecessary columns.** Removed: event-id, sensor-type, individual-taxon-canonical-name, tag-local-identifier, individual-local-identifier, study-name, visible.

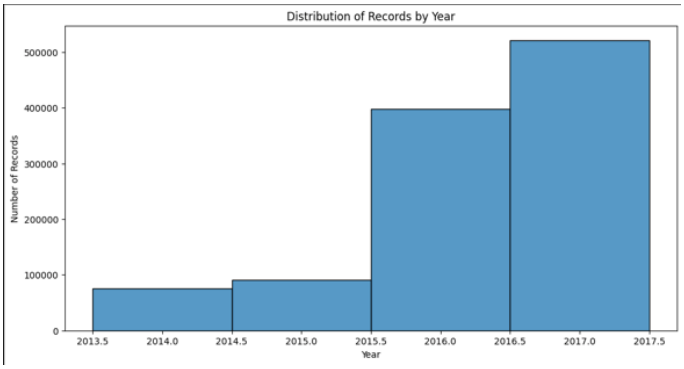

Figure 1: The distribution of records by year from 2014 to 2017 shows a significant increase in data collection over time, especially during 2016–2017.

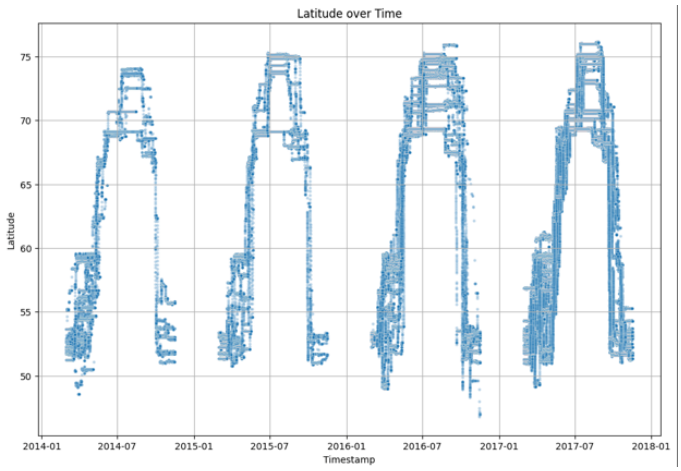

Figure 2: The chart illustrates the change in latitude positions of GPS-tagged animals over time. This provides a clear visual representation of seasonal migration behavior.

- **Handled missing values.** Filled missing values in ground-speed, heading, and height-above-msl with their respective column means.
- **Removed invalid/outlier values.** Kept only rows where height-above-msl is between 0 and 5000; Replaced heading values outside the range [0, 360] with NaN.
- **Dropped remaining missing values.** Removed rows with NaN in heading or height-above-msl.
- **Datetime conversion and feature extraction.** Converted timestamp to datetime format; Extracted features: year, month, day, hour.

0.5.3 Data augmentation with GAN and VAE for Migration Flight Path Generation

GAN (Generative Adversarial Networks)

- **Real Flight Data.** This is the actual flight path data of the geese (preprocessed and normalized), containing features such as: Longitude, latitude, speed, altitude, heading, time (hour, month, day of the year).

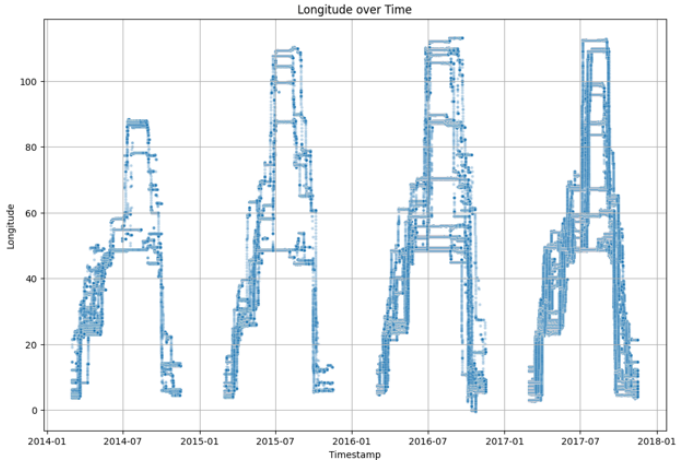

Figure 3: The chart shows the change in longitude positions of individuals (possibly GPS-tagged animals) over time.

- **Noise Vector (z).** This is a random vector from the latent space (typically of size 100) used as input for the Generator. The purpose is to generate a variety of new, realistic-looking flight paths.
- **Generator.** The Generator receives the noise vector (z) and generates a fake flight path that matches the format and characteristics of the real flight data.
- **Discriminator.** The Discriminator receives two sources of data: Real data from the training set; Fake data generated by the Generator. Its task is to predict whether each sample is real (1) or fake (0).
- **Training Loop.** The Generator and Discriminator are trained simultaneously in a competitive game: The Generator tries to deceive the Discriminator; The Discriminator attempts to accurately distinguish real from fake data. After multiple epochs, the Generator learns to produce flight paths very similar to the real ones.
- **Generated Flights.** After training, the Generator can create new, plausible future flight paths based on the learned distribution.

VAE (Variational Autoencoders)

- **Input Data.** These are the feature vectors at each point in the flight journey of the geese: longitude, latitude, speed, heading, altitude, and time (year, month, hour, day of the year). The data is normalized using MinMaxScaler and then input into the model.
- **Encoder.** The Encoder compresses the input feature vector into two vectors: μ (mean) and $\log \sigma^2$ (log-variance). These vectors represent the probabilistic distribution of the data. The Objective is to learn a meaningful latent space representation.
- **Latent Representation.** From μ and $\log \sigma^2$, the model generates a latent vector z using the formula:

$$z = \mu + \sigma \cdot \epsilon \quad \text{where} \quad \epsilon \sim \mathcal{N}(0, 1)$$

This technique enables training via gradient descent, since sampling is not differentiable.

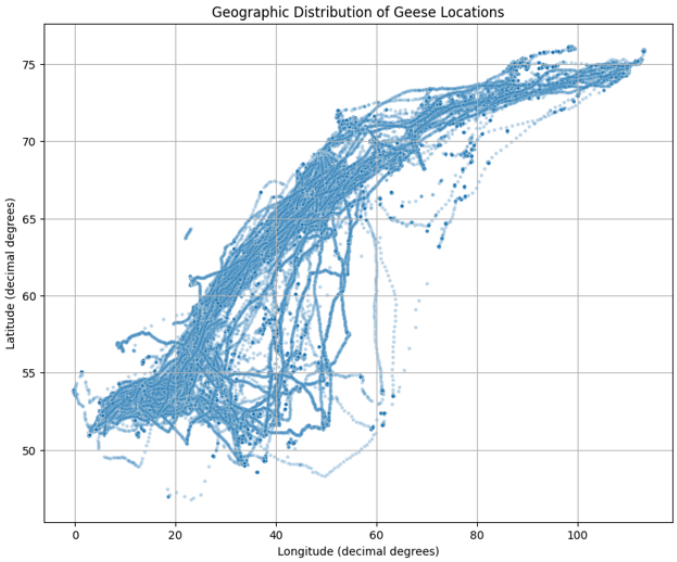

Figure 4: This chart combines longitude (x-axis) and latitude (y-axis) to illustrate the geographical movement trajectories of individual geese.

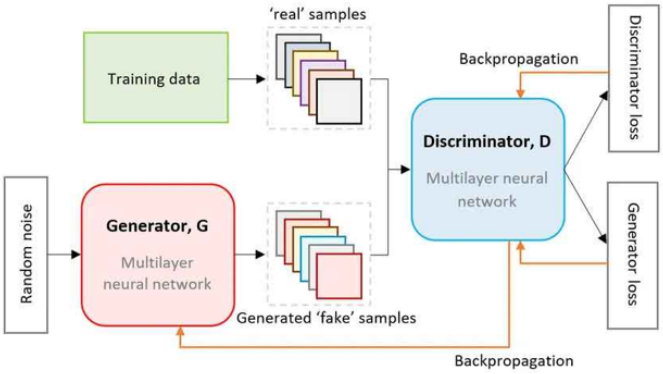

Figure 5: Architecture of GAN

- **Decoder.** The decoder receives the latent vector z and reconstructs the original feature vector. The objective is to generate data points that closely resemble the real data.
- **Output.** After training, we can sample from the latent space $z \sim \mathcal{N}(0, I)$ to generate new flight paths that are statistically consistent with past flight behaviors.

0.5.4 Justification and Selection Process of DBSCAN for Clustering Goose Migration Behavior

In the task of analyzing the migration behavior of the Greater White-fronted Goose (Anser albifrons), selecting an appropriate clustering algorithm plays a vital role in identifying typical movement patterns, important stopover sites, and potential behavioral anomalies. After reviewing several popular clustering techniques such as K-Means, MeanShift, and Hierarchical Clustering, the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm was chosen due to its distinctive advantages that align well with the characteristics of this problem.

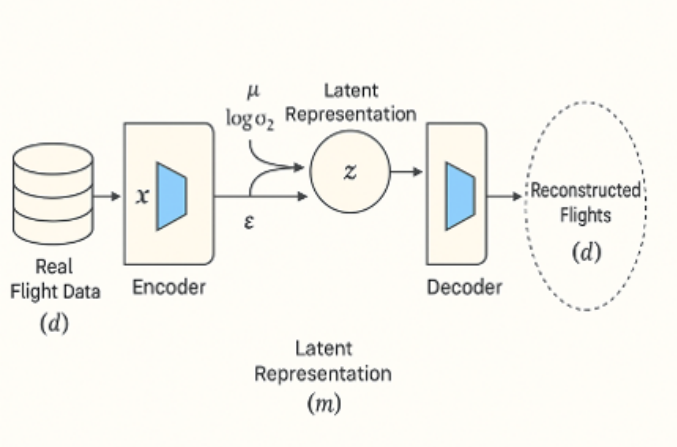

Figure 6: Architecture of VAE

One of the key reasons for selecting DBSCAN is its non-reliance on a predefined number of clusters—a particularly important feature when working with natural movement data, which often lacks clear or fixed group structures. Furthermore, DBSCAN is well-suited for spatial data and is capable of identifying clusters with arbitrary shapes, unlike algorithms such as K-Means that assume spherical or uniform clusters. This flexibility is crucial since migration trajectories tend to be irregular, with scattered and non-uniform stopover points.
Another major advantage is DBSCAN's ability to detect noise points, which do not belong to any cluster. These outliers can signify abnormal behaviors, such as migration path deviations or prolonged stops in uncommon areas—providing valuable inputs for further anomaly detection and real-time monitoring applications.
The selection process involved experimenting with different values of the two key DBSCAN parameters: **epsilon** – neighborhood radius and **minPts** (minimum number of points required to form a dense region). Through iterative tuning and evaluation, DBSCAN demonstrated stable clustering performance, effectively grouping common migration paths while highlighting potentially significant outlier behaviors.
In summary, DBSCAN was selected due to its strong alignment with the spatio-temporal complexity of migration data, its ability to handle irregular cluster shapes, and its integration potential with downstream tasks such as anomaly detection and real-time analysis.

0.6 Conclusion

The study of migration patterns of the Greater White-fronted Goose using clustering algorithms like DBSCAN provides valuable insights into their spatial behavior and seasonal movement across vast geographic regions. By identifying common migratory routes and stopover points, this approach supports ecological conservation, environmental monitoring, and data-driven decision-making. However, as ecological systems grow increasingly dynamic due to climate change and human impact, it becomes essential to enhance the system's capabilities. Future developments should focus on real-time data integration, the adoption of more advanced clustering methods, and the implementation of anomaly detection techniques to improve the accuracy, responsiveness, and research value of migration behavior analysis.