Trajectory Prediction Using LSTM RNN for Jaguar Movement in the Amazon Region of Paraguay

Abstract

This study implements Recurrent Neural Network (RNN) models with Long Short-Term Memory (LSTM) to predict movement patterns of jaguars in the Amazon region of Paraguay. The jaguar movement database from Morato et al. (2018a) was used, complemented with daily climatological data from NASA (available since 2010). The integration of these sources, after data normalization, achieved an accuracy above 67% in route prediction, with an F1-score of 81%, outperforming other models. The insufficient number of records prevented the effective use of Uber's H3 polygons, although their relevance is proposed for future studies with larger datasets. Limitations in the NASA API were overcome by designing a script for collecting climatological data. The results provide a basis for deeper analyses and offer potential applications in the conservation of protected species, such as the prevention of risk events.

1 Introduction

The advancement in machine learning and deep learning has been remarkable in the last decade, driven by reduced storage and processing costs. The so-called deep learning explosion, popularized since 2010 by the works of Geoffrey Hinton, Yann LeCun, and Yoshua Bengio, has enabled the successful application of Recurrent Neural Networks (RNN) for processing sequential data. RNNs are designed to handle sequences (such as time series or audio) and have a "memory" that allows them to remember information from previous steps; however, traditional RNNs suffer from the vanishing gradient problem, making it difficult to capture long-term dependencies.

Long Short-Term Memory (LSTM) networks are a special type of RNN that address the vanishing gradient issue. LSTMs introduce "gates" that control the flow of information through the network, allowing them to remember relevant data over extended periods. In this study, LSTMs are employed to predict jaguar movement routes using data from Morato et al. (2018a) enriched with climatological and geolocation variables. Additionally, derived variables such as azimuth, speed, and distance traveled are calculated to enhance the performance of the proposed model.

2 Problem Description

The problem addressed in this study is how to accurately and in real time predict the future location of large felines based on their historical movements and environmental and temporal variables. In practical terms, the aim is to develop a model that integrates GPS data, climatological conditions (e.g., temperature, humidity, pressure, etc.), and temporal characteristics (time of day, season) to estimate in which polygon or geographical area a specific animal is most likely to be found at a given moment.

2.1 Key Aspects of the Problem

- **Localization Uncertainty:** The movements of felines are complex and influenced by multiple environmental and biological factors. Predicting their future position is essential for conservation and species management.
- **Integration of Multiple Variables:** The model must combine spatial (longitude and latitude), temporal (time, date), and climatological data, requiring advanced machine learning techniques to capture the nonlinear and stochastic dynamics of movement.
- **Practical Applications:** Accurate predictions allow for the design of ecological corridors, planning of conflict mitigation measures (e.g., avoiding high-traffic zones), and improved decision-making in conservation efforts.

In summary, the study is proposed as a solution to anticipate the presence of large felines in specific areas, contributing to better management and protection of these animals in increasingly human-impacted natural environments.

3 Data Description

3.1 Feline Movement Data

The study by Morato et al. (2018a) provides an unprecedented dataset containing 134,690 jaguar locations tracked by GPS in five countries, aimed at overcoming the scarcity of data on large carnivores. This freely accessible resource allows the investigation of various ecological questions, from landscape connectivity and usage to intra-species and predator-prey interactions, ultimately enhancing jaguar conservation through a better understanding of its movement ecology. Due to varying climatological and geographical conditions across countries, this study focuses on the "Humid Chaco" region in Paraguay. Data were recorded for 8 individuals, with a total of 5,436 records, including variables such as:

- **EVENT_ID:** Event identifier.
- **TIMESTAMP:** Date and time of the record.
- **LOCATION_LONG and LOCATION_LAT:** Geographic coordinates.
- **CANONICAL_TAXONOMY:** Feline taxonomy (in this study, all correspond to *Panthera onca*).
- **TAG_LOCAL_IDENTIFIER, ID_INDIVIDUAL_IDENTIFIER, STUDY_NAME, and COUNTRY.**

3.2 Climatological Data

NASA's POWER (Prediction Of Worldwide Energy Resources) portal is an online platform that provides free access to georeferenced climatological and meteorological data. The climatological data used in this study include:

- **T2M (Temperature at 2 Meters):** Air temperature at 2 meters above the surface.
- **RH2M (Relative Humidity at 2 Meters):** Relative humidity at 2 meters.
- **WS10M (Wind Speed at 10 Meters):** Wind speed at 10 meters.
- **PS (Surface Pressure):** Atmospheric pressure at the surface.
- **PRECTOTCORR (Precipitation Total Corrected):** Corrected total precipitation.

3.3 Uber H3 Polygons

Uber's H3 polygons are a geographic tessellation system that divides the Earth's surface into hexagonal cells of various sizes. This system offers:

- **Hexagonal Shape:** Cells that better approximate the Earth's spherical shape compared to rectangular grids.
- **Hierarchical Resolutions:** 16 levels, ranging from very large cells (level 0) to very small cells (level 15).
- **Unique Indexing:** Each hexagon has a unique identifier (H3 index), facilitating its use in databases and spatial analysis systems.
- **Computational Efficiency:** Optimized for operations such as neighbor cell lookup, data aggregation, and spatial interpolation.

3.4 Limitations

- The granularity of the climatological data for dates prior to 2020 is daily.
- The slow response time of the NASA server necessitated batch queries, reducing the geographic focus to the Humid Chaco region in Paraguay.
- Although H6 and H8 polygons were included, exploratory analysis revealed a low frequency of cases per polygon, which hindered model training. It is recommended to use smaller-diameter polygons in future studies with larger datasets.

4 Data Preparation

A script was developed to perform batch queries (batch size = 100) to avoid having the NASA Data site block the IP address of the extraction machine. Additionally, the script performs partial loading of each data block into a local CSV file. The notebook used for this purpose is `weather data and polygons.jpynb`, which contains the function `get_weather_data`. Uber polygons (H8 and H6) were extracted using the `h3` library and the function `latlng_to_cell`. Furthermore, derived variables were added:

- Distance traveled between measurements, calculated using the Haversine formula.

- Azimuth.
- Speed.
- Altitude above sea level.

5 Model Description

The notebook `routes_model.ipynb` presents the complete implemented solution. The components of the model are described below:

5.1 Data Reading

This section reads data from a CSV file, which must contain at least:

- **TIMESTAMP:** Date and time of the record.
- **ID_INDIVIDUAL_IDENTIFIER:** Unique identifier of the feline.
- **LOCATION_LONG and LOCATION_LAT:** Coordinates of the record.
- Climatological variables: `T2M`, `RH2M`, `WS10M`, `PS`, `PRECTOTCORR`, `HIGH`.

5.2 5.2 Data Cleaning

This stage calculates the hour of the day from the TIMESTAMP and orders the data by individual and timestamp, which is crucial for applying window functions and segmenting the data by individual.

5.3 Sequence Generator

The `SequenceGenerator` class takes a DataFrame containing data for individuals (e.g., location coordinates and other features) and generates fixed-length sequences (`seq_length`) from these records. These sequences serve as inputs (X) for the model, while the target values (y) are the coordinates to be predicted (e.g., the next location). Additional functionalities include:

- **Oversampling:** If the movement between the start of the sequence and the target is significant (above a specified threshold), the sequence is duplicated several times (using an oversample factor) to give more weight to these cases during training.
- **Feature Scaling:** Both features (X) and targets (y) are scaled to normalize the data and improve model performance.

5.4 Model Configuration

The `ModelConfig` class encapsulates the construction and configuration of an LSTM neural network for predicting coordinates (latitude and longitude).

- **Initialization:** The class receives the sequence length and the number of features per time step.

- **Custom Loss Function (weighted_mse):** This function computes a weighted mean squared error based on the displacement (Euclidean norm of the true values), penalizing errors more when movements are significant.
- **Model Construction:**
  - A sequential model is built with two LSTM layers (the first with 64 units and the second with 32 units) along with a Dropout layer (20%) to prevent overfitting.
  - A dense layer with 32 units and ReLU activation is used as an intermediate layer.
  - The output layer has 2 neurons to predict the latitude and longitude.
  - The model is compiled with the Adam optimizer and the custom loss function.

5.5 Trainer Class

The `Trainer` class manages the training process of the deep learning model and logs relevant information using MLflow for tracking and reproducibility. It logs parameters, metrics, and saves the trained model.

5.6 Evaluator Class

The `Evaluator` class is designed to evaluate the performance of the model in predicting geographic coordinates. It uses the Haversine formula to calculate the distance between the true and predicted coordinates and considers a prediction correct if the distance is within a specified threshold (e.g., 1 km). It computes metrics such as accuracy, F1 score, and recall.

5.7 Predictor Class

This class is essential for making predictions with the trained model.

- **predict_next:** Predicts the next step (coordinates) based on an input sequence.
- **predict_trajectory:** Simulates an extended future trajectory by iteratively updating the input sequence with the predicted values.
- It also includes a visualization function to inspect the actual trajectories of different individuals.

5.8 Pipeline

The pipeline integrates data reading, preprocessing, sequence generation, model configuration, training, evaluation, saving, and visualization. This automated workflow is essential for building and applying an LSTM-based trajectory prediction system for tracking felines.

6 Results and Conclusions

The trained model achieved the following performance metrics on the training data (with internal validation):

- **Accuracy:** 0.677

- **F1 Score:** 0.807
- **Recall:** 0.677

The model's performance, particularly the F1 score, is noteworthy considering the set of variables used (`LOCATION_LONG, LOCATION_LAT, hour, T2M, RH2M, WS10M, PS, PRECTOTCORR, HIGH`).

7 Conclusions

The feasibility and effectiveness (F1 ≈ 0.81) of implementing an LSTM-based RNN model for studying and predicting the movement routes of felines have been demonstrated.

- Custom classes and functions were developed to automate the process for future studies.
- Strategies such as normalization and oversampling were crucial in improving training performance.

8 Recommendations

- An evaluation on an independent test dataset is necessary, as this study only evaluated the model on the training data (with internal validation), posing a risk of overfitting.
- Additional variables, such as the gender and age of the individual, could be included provided adequate pre-processing is performed.
- This work serves as a starting point for further improvements to the model and its application in real-world animal movement research.

References

- Morato, R.G., Thompson, J.J., Paviolo, A., de La Torre, J.A., Lima, F., McBride-Ref ellipsis Jr, R.T., Ribeiro, M.C. (2018a). *Jaguar movement database: a GPS-based movement dataset of an apex predator in the Neotropics*. Ecology. Wiley Online Library.
- NASA POWER. NASA Open Data Portal. Retrieved from `https://power.larc.nasa.gov`.
- UBER H3. Uber's Hexagonal Hierarchical Spatial Index. Retrieved from `https://www.uber.com/en-PT/blog/h3/`.