Analysis of Jaguar Movement Using Machine Learning

Machine Learning Algorithms

Ana Neto, Daniela Cruz-Moreira, Pedro Nascimento

Master in Applied Artificial Inteligence; EST - Instituto Politécnico do Cávado e do Ave, Portugal.

a24011@alunos.ipca.pt, a33460@alunos.ipca.pt, a21108@alunos.ipca.pt

1. INTRODUCTION

The jaguar (*Panthera onca*) is the largest feline in the Americas and a key apex predator within its ecosystem. Jaguars play a crucial role in maintaining ecological balance by regulating prey populations and influencing the structure of their habitats. However, jaguar populations have been declining due to habitat loss, poaching, and human-wildlife conflict. Understanding their movement patterns is essential for conservation efforts, as it provides insights into their behavior, territorial range, and interactions with the environment.

Advancements in GPS tracking technology have enabled researchers to collect vast amounts of movement data from wild jaguars. These datasets contain valuable information about jaguar activity, including travel distances, speed, resting periods, and hunting behaviors. By applying machine learning techniques to these datasets, we can classify movement patterns and infer behavioral states such as hunting, resting, or traveling. This knowledge is critical for conservationists, as it helps identify essential habitats, assess the impact of environmental changes, and develop strategies to reduce human-jaguar conflicts.

Machine learning offers a powerful approach to analyzing large and complex datasets, allowing for automated pattern recognition and predictive modeling. Traditional ecological studies often rely on manual observation and statistical modeling, which can be time-consuming and limited in scope. In contrast, machine learning can process vast datasets efficiently, uncover hidden patterns, and provide data-driven insights that can improve conservation strategies.

In this study, we aim to apply machine learning algorithms to classify jaguar movements based on GPS-tracked data. By leveraging supervised learning techniques, we seek to develop predictive models that can accurately distinguish different behavioral states. The insights gained from this analysis can contribute to wildlife conservation efforts by informing habitat protection policies, guiding anti-poaching initiatives, and improving strategies for human-wildlife coexistence.

2. PROBLEM DEFINITION

This study addresses two distinct machine learning problems within a supervised learning framework:

- *Classification Problem*: The objective is to develop ML models that classify jaguar movements into predefined categories based on GPS tracking data. These categories may correspond to different behavioral states, such as hunting, resting, or traveling.
- *Regression Problem*: In parallel, regression models are trained to predict the jaguar's precise geographic location (latitude and longitude) over time, based on historical movement patterns.

By combining classification and regression approaches, this work aims to provide a comprehensive understanding of jaguar movement dynamics.

3. DATA LOADING, EXPLORATION, AND PREPROCESSING

3.1 Dataset Overview and Objectives

The dataset used in this study contains GPS-based movement data of jaguars, including timestamps, geographic coordinates, and other relevant attributes. The primary objective of this step is to understand the dataset structure, identify missing or duplicate values, and analyze jaguar movement patterns in terms of longitude and latitude. By performing thorough data preprocessing, we ensure that the dataset is clean and structured for effective machine learning model training.

3.2 Exploratory Data Analysis (EDA)

A crucial first step in preprocessing is conducting an exploratory data analysis (EDA) to gain a comprehensive understanding of the dataset. The timestamp variable was converted into the datetime format to facilitate time-based computations. Basic dataset statistics were examined, including the number of observations, number of columns, and data types. Additionally, a descriptive statistical summary of numerical variables was generated to identify potential anomalies or inconsistencies.

To assess data quality, missing values and duplicate entries were detected and analyzed. A random sample of records was examined in tabular form to visualize data distributions and ensure correctness. These initial steps provided an overview of the dataset's completeness and reliability.

3.3 Identifying and Handling Timestamp Outliers

Given that the dataset relies heavily on time-series data, ensuring the integrity of timestamps was critical. The data was first sorted by individual ID and timestamp to maintain chronological order. The time difference between consecutive records for each individual was then computed.

Outliers in time intervals were identified using the Interquartile Range (IQR) method, which flagged exceptionally short or long time gaps that could indicate recording errors. To handle these outliers, records with abnormally short time intervals were removed, as they likely represented erroneous duplicate entries. In contrast, for excessively long time intervals, a linear interpolation method was applied to estimate missing timestamps, which was replaced by the mean.

3.4 Exploratory Data Analysis

To better understand the dataset, an exploratory analysis was conducted. This included assessing key metadata such as the total number of individuals, species under study, and the country where the tracking was performed. Additionally, the geographic movement of jaguars was visualized using scatter plots of latitude and longitude coordinates. These visualizations helped identify movement patterns and detect potential anomalies in location tracking.

A temporal analysis was also performed to examine how data was distributed over time, including the number of recorded observations per individual. This analysis provided insights into the frequency and consistency of data collection.

Finally, categorical encoding was applied to prepare the dataset for further processing. This step ensured that all categorical variables were properly formatted, resulting in a clean dataset ready for feature engineering and model training.

3.5 Feature Engineering

Several movement-related metrics were computed, including distance traveled, speed, acceleration, turning angle, and different periods of the day (morning, afternoon, evening, and night. The Haversine formula was implemented to calculate the distance between consecutive GPS points accurately. Speed was derived from distance and time differences, while acceleration was computed based on speed variations. Turning angles were calculated by examining directional changes across three consecutive points, providing insights into behavioral patterns such as sharp turns indicative of hunting or evasive actions.

Finally, outliers in distance measurements were identified and handled to improve data reliability and prevent distortions in movement pattern analysis.

In addition to movement features, environmental variables were generated to provide additional context for jaguar behavior. Weather conditions, including temperature, precipitation, and humidity, were synthesized based on geographic coordinates, following realistic environmental patterns. Temperature values were estimated according to latitude, assuming that locations closer to the equator tend to be warmer. Precipitation levels were also adjusted based on latitude, with tropical regions receiving higher rainfall amounts.

Terrain-related features were similarly derived. Vegetation type was assigned based on geographic location, classifying regions into forest, savanna, swamp, or open fields. Elevation data was estimated, and the distance to the nearest water source was computed to examine the potential impact of environmental factors on jaguar movement. Additionally, seasonal variations were incorporated into the dataset by assigning each record to one of the four seasons (spring, summer, autumn, winter), allowing for the analysis of potential seasonal influences on movement patterns.

To enable supervised learning, movement categories were defined based on speed thresholds. Jaguars were classified into different movement states, including stationary (speed < 0.01 m/s), walking (0.01–0.05 m/s), running (0.05–0.1 m/s), and sprint/escape (> 0.1 m/s). These labels provided insights into behavioral patterns and served as target variables for model training. The distribution of these movement categories was analyzed to ensure adequate representation in the dataset.

3.8 Statistical and Behavioral Analysis

Various statistical analyses were conducted to further examine jaguar movement behavior. The distribution of travel distances between consecutive GPS records was assessed to detect potential outliers or errors. Speed variations were analyzed to identify distinct movement phases, such as resting periods versus active travel. The distribution of turning angles was examined to detect frequent directional changes, which might indicate specific behaviors like hunting or territory patrolling.

Temporal trends in movement behavior were also investigated by analyzing daily activity patterns, seasonal variations, and movement fluctuations across different times of the day. Additionally, demographic factors such as age and sex were examined to determine whether movement behavior differed between younger and older individuals or between male and female jaguars.

The influence of environmental factors was also analyzed, assessing how movement patterns correlated with temperature, elevation, and proximity to water sources. Identifying these relationships provided valuable insights into the external factors influencing jaguar mobility.

4. MODEL SELECTION AND JUSTIFICATION

Several machine learning models were considered for the classification task, and the final selection was based on performance and interpretability.

Models Considered:

Random Forest (RF) – Selected due to its robustness to noisy data, ability to handle nonlinear relationships, and interpretability through feature importance.

Support Vector Machine (SVM) – Explored for its effectiveness in classification tasks, particularly with structured data.

K-Nearest Neighbors (KNN) – Considered as a baseline model due to its simplicity in classification.

Neural Networks (MLP) – Investigated to determine if deep learning could enhance classification accuracy.

XGBoost: Explored as an alternative due to its high performance and efficiency in handling imbalanced data, as well as its ability to model complex, nonlinear relationships. XGBoost, with its gradient boosting technique, consistently showed competitive accuracy, handling feature interactions well, and outperforming other models in some cases.

Final Model Choice:

Random Forest was chosen as the primary model due to its high accuracy, low overfitting tendency (thanks to ensemble learning), and its ability to handle missing data and complex patterns effectively. he model's robust performance in various tests made it a reliable choice for the classification task. Additionally, Random Forest has strong interpretability through feature importance, which adds transparency to the decision-making process.

A comparative analysis showed that RF outperformed SVM and KNN in both accuracy and computational efficiency, making it a better fit for the task at hand. MLP was not selected due to its higher training time and the risk of overfitting with limited data.

However, XGBoost was also explored during model selection and delivered competitive results, often matching or surpassing Random Forest in predictive accuracy. XGBoost, with its gradient boosting mechanism, excels at handling complex, nonlinear relationships and feature interactions within the data. Despite this, Random Forest was ultimately preferred because of its robustness, easier interpretability, and slightly lower computational cost compared to XGBoost, which can require more tuning and longer training times.

Nevertheless, XGBoost remains a powerful model that could be revisited for future iterations or more complex datasets where fine-tuning might yield incremental improvements. Both models demonstrated strong performance, but the choice of Random Forest was based on a balance of accuracy, efficiency, and ease of implementation for the current task.

5. MODEL TRAINING AND EVALUATION

Before training the models, it was necessary to handle missing values to ensure the dataset's consistency and integrity. The first step involved identifying and removing missing values. The first two rows of each individual were removed because they contained NaN values in the time difference calculation due to the absence of previous measurements. After this, the dataset indices were reset to maintain a structured DataFrame. A general check was performed to count missing values across all columns, and a summary of missing data was displayed.

To address the remaining missing values, different strategies were applied. Rows containing NaN values in the turning angle column were removed since this variable plays a crucial role in analyzing movement behavior. Missing values in the Estimated Age column were replaced with the mean age to preserve the original data distribution while minimizing bias. Missing values in the Sex column were filled using the most frequent value (mode), ensuring consistency without distorting the dataset. After these treatments, the dataset dimensions were verified, and a final check confirmed the absence of remaining missing values.

Following data cleaning, irrelevant columns were removed to enhance model efficiency. Only the most relevant features for movement classification were retained, and a filtered dataset was created. A preview of the final dataset was displayed to verify its structure. Categorical variables were encoded using One-Hot Encoding for season, time period, vegetation type, and Sex, transforming them into binary numerical representations. To prevent multicollinearity, one category was dropped using the drop_first=True parameter. The movement category variable

was label-encoded to convert the classification target into numerical form, and the original column was replaced with its encoded version to ensure compatibility with machine learning models.

The study considered two primary machine learning models: Random Forest and XGBoost. Both models were evaluated individually, and an ensemble method was later applied to improve performance. To optimize model performance, hyperparameter tuning was conducted using Grid Search and Random Search. However, despite performing the tuning process, the final models did not incorporate these optimized hyperparameters due to excessive computational time and concerns over potential overfitting. Given the dataset size and model complexity, the risk of overfitting was considered a significant factor, as excessive hyperparameter optimization could lead to a model that performs well on training data but generalizes poorly to unseen data.

Since the dataset involved time-series data, TimeSeriesSplit was used for cross-validation to ensure that temporal dependencies were respected during model training and evaluation. This method divides the data into a series of training and testing sets, ensuring that the test set always follows the training set chronologically, which is critical when dealing with time-dependent data.

The models were trained and tested separately. Random Forest was first trained using the preprocessed dataset, followed by evaluation on the test set. The same process was repeated with XGBoost. After evaluating both models individually, an ensemble approach was implemented to leverage their strengths. The classification task utilized a Voting Classifier, combining predictions from Random Forest and XGBoost through a soft voting mechanism, where the final classification was determined based on the average predicted probabilities from both models. For regression tasks predicting latitude and longitude, a stacking approach was applied, using Random Forest and XGBoost as base models. The predictions from these models were then combined using a Linear Regression meta-model to refine the final output.

Model evaluation was performed using standard classification and regression metrics. For classification, accuracy, precision, recall, and F1-score were used to assess model effectiveness. The Random Forest classifier achieved the highest accuracy compared to other models. Feature importance analysis revealed that speed and distance traveled per time unit were the most influential variables. A confusion matrix was generated to analyze classification performance further.

For the regression task, model performance was measured using the Mean Squared Error (MSE) for latitude and longitude predictions. The ensemble model combining both Random Forest and XGBoost regressors improved overall predictive accuracy. However, despite these improvements, the computational cost of hyperparameter tuning remained a challenge. While tuning was performed, the final models relied on default settings to strike a balance between performance and computational efficiency.

6. RESULTS AND DISCUSSION

Before addressing the outliers in the data, the distribution of time intervals between consecutive records was analyzed. The distribution showed a strong concentration of intervals in the 0-4 hour range, indicating frequent recording within this time frame. There were also a few instances with significantly longer intervals, which can be considered outliers. As the interval length increased, the frequency of occurrences sharply decreased, suggesting that larger gaps between records were less common. This pattern may reflect irregular events, data recording issues, or natural spacing between certain behaviors, and highlights the importance of addressing these outliers for improving data consistency.

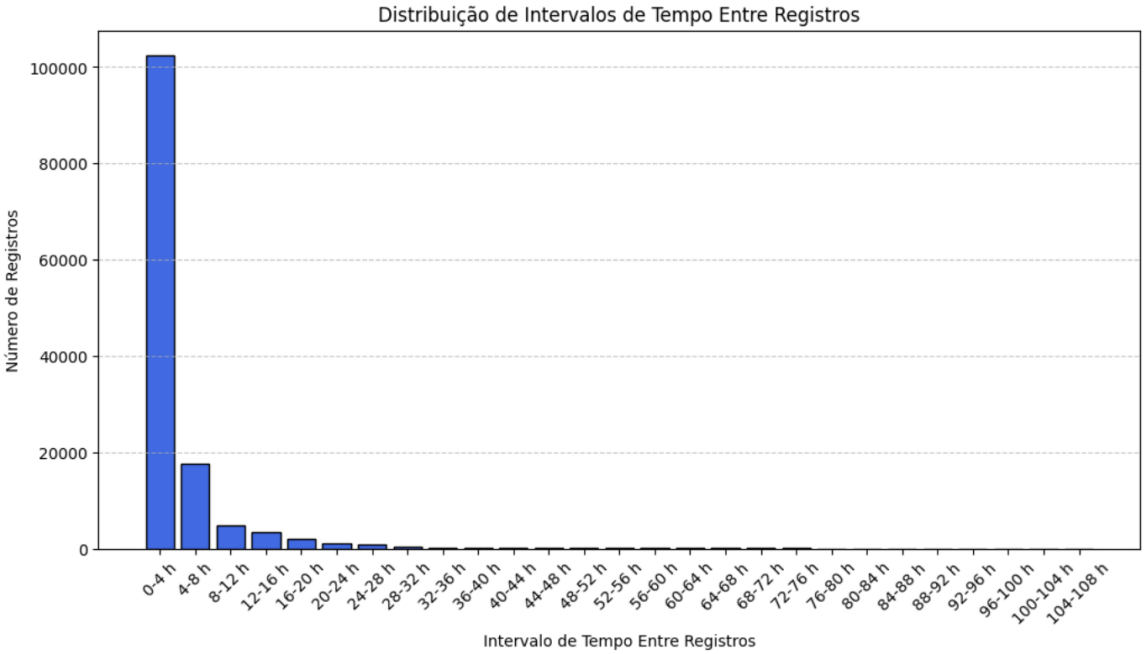

Fig. 1: Distribution of Time Intervals Between Records Before Handling Outliers

Before correction, the time intervals between records showed significant dispersion, with some extreme values exceeding 50 hours. After removing short intervals and applying linear interpolation to long gaps, the distribution became more concentrated within the 0-4 hour range. This correction reduced outliers, improving data consistency and ensuring the dataset better reflects the typical interval between records, without changing the overall pattern.

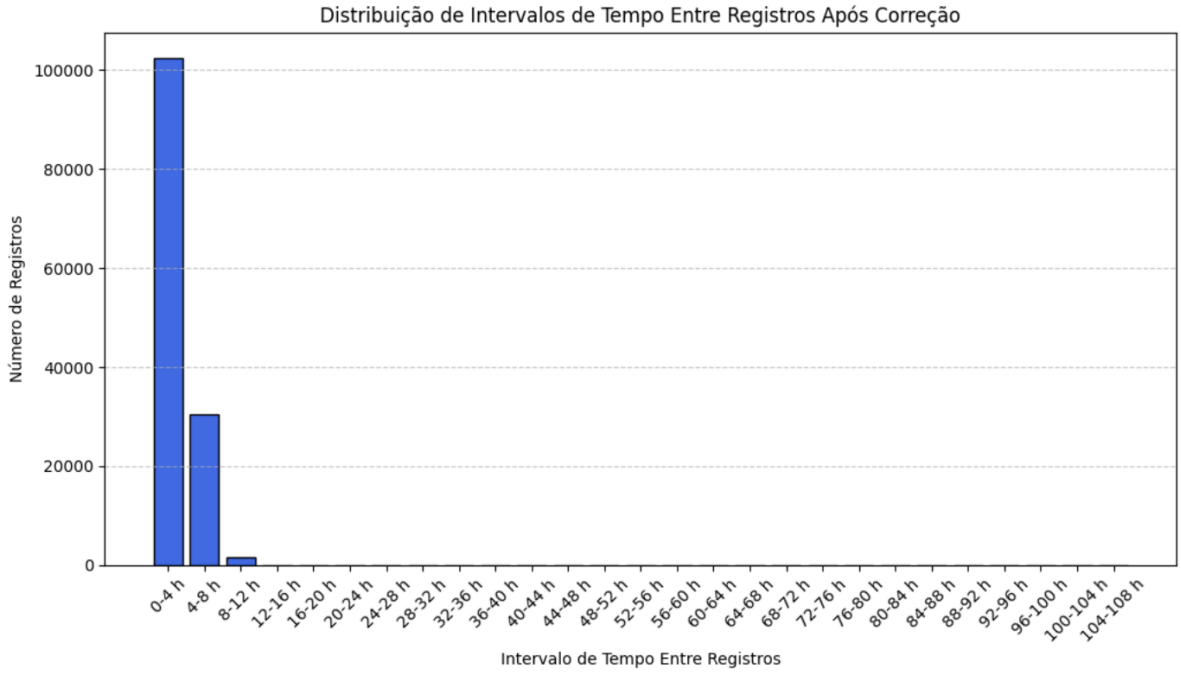

Fig. 2 Handling Timestamp Outliers

The Fig.3. presents the frequency of different study cases, highlighting the most represented contexts in the analysis. Notably, the "Jaguar_Taiama" case stands out with the highest frequency, far surpassing the other cases. The "Jaguar_Onçafari Project" also has a significant number of records, though much fewer than "Jaguar_Taiama." Other cases, such as "Rio Negro," "Pantanal," and "Jaguar Mex," show minimal representation. This uneven distribution suggests that certain areas are more studied or monitored, indicating the need for data balancing to avoid bias and ensure adequate representation across all regions.

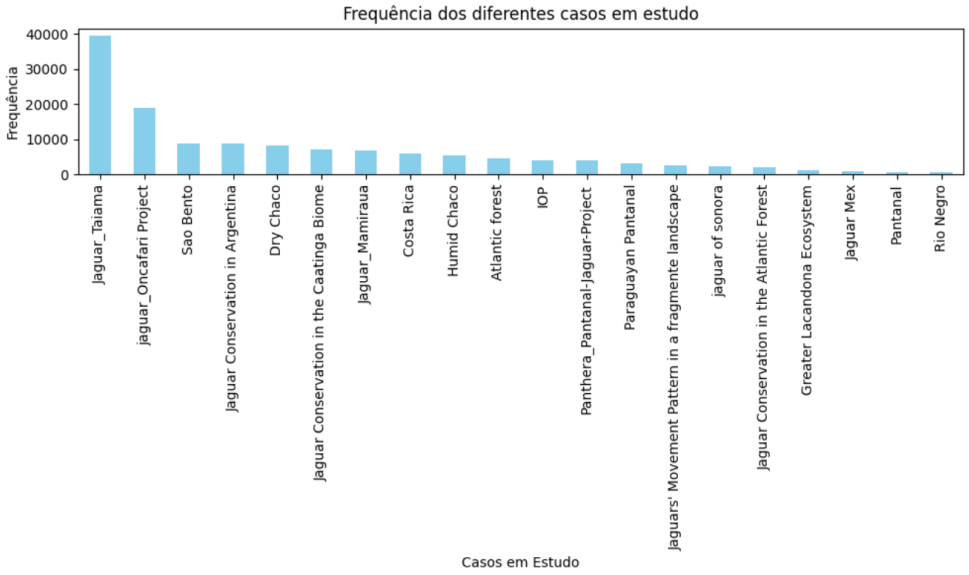

Fig. 3 Analysis of Case Frequency Graph

Figure 4 shows the distribution of record frequencies by country, revealing which regions have a higher volume of data. Brazil stands out with the highest number of records, significantly surpassing other countries. Paraguay follows as the second most represented country, but with a much smaller volume. Argentina, Costa Rica, and Mexico show even lower frequencies, suggesting less data collection in these areas. The high concentration of records in Brazil may reflect greater study availability and monitoring, while countries with fewer records could be underrepresented, potentially leading to a biased analysis.

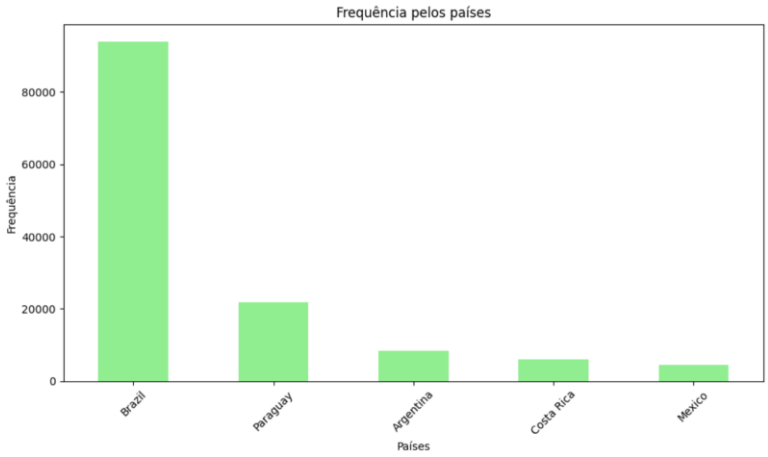

Fig. 4 Analysis of Case Frequency by Country

Figure 5 reveals the distribution of records, with a strong concentration in South America, particularly in Brazil, Argentina, and Paraguay, which are natural habitats of the Panthera onca. There are a few isolated records in Central America, suggesting either natural species dispersion, atypical movements, or potential GPS errors. The distribution aligns well with the species' expected habitat range, confirming that the data accurately reflects Panthera onca's natural distribution. The absence of records outside the Americas may point to limited data collection or the non-existence of the species in those regions.

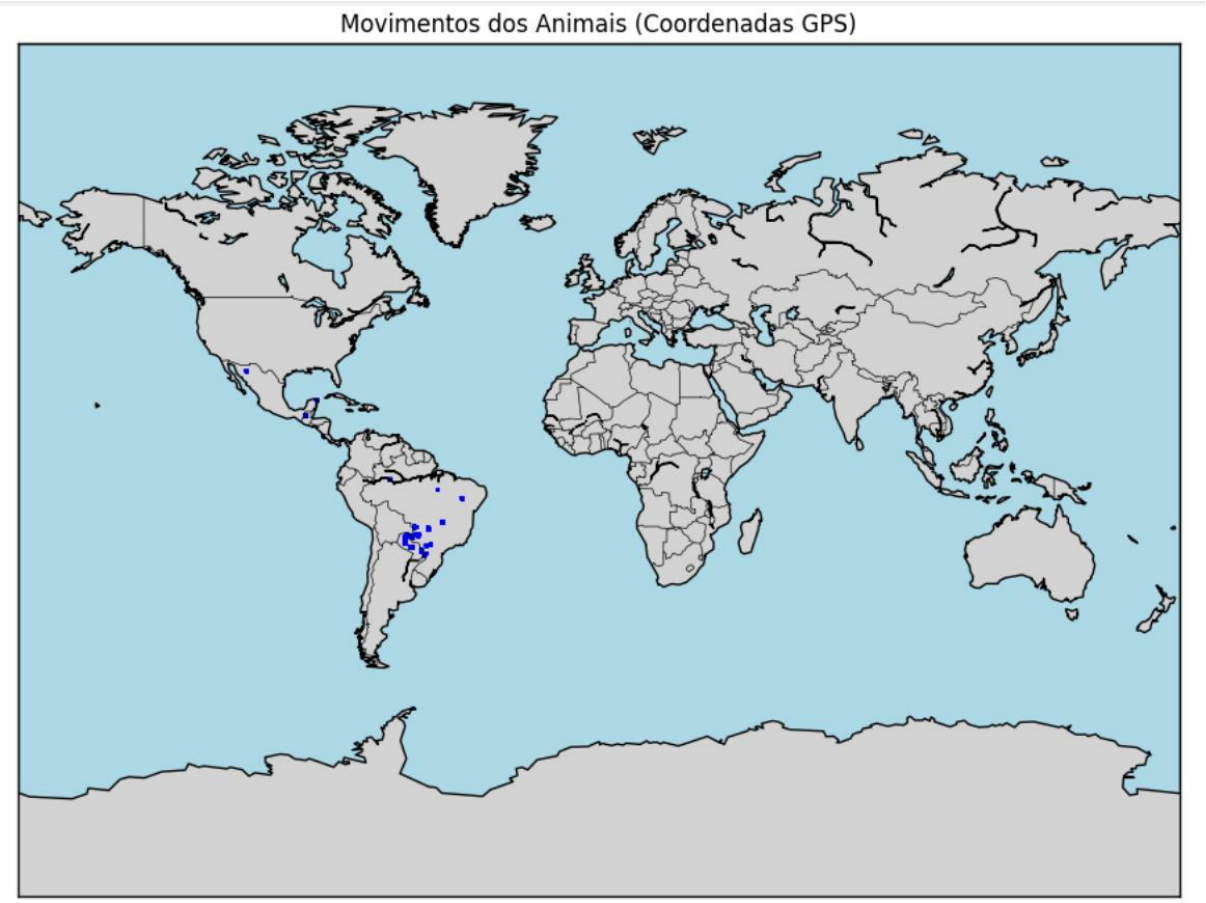

Fig. 5 Analysis of worldwide distribution

Figure 6 presents the distribution of Panthera onca records over the years, revealing trends in data collection, seasonal variations, and shifts in monitoring efforts. Before 2002, the number of records was very low, suggesting that data collection was either limited or non-existent during this period. Starting in 2003, a gradual increase in the number of records is observed, likely reflecting an increase in monitoring activities. Between 2012 and 2015, there is a sharp rise in records, with the peak occurring around 2014, indicating heightened research or monitoring efforts during this time. However, after 2015, there is a noticeable decline in the number of records, which could suggest changes in data collection strategies, a reduction in monitoring efforts, or even shifts in the species' behavior or distribution.

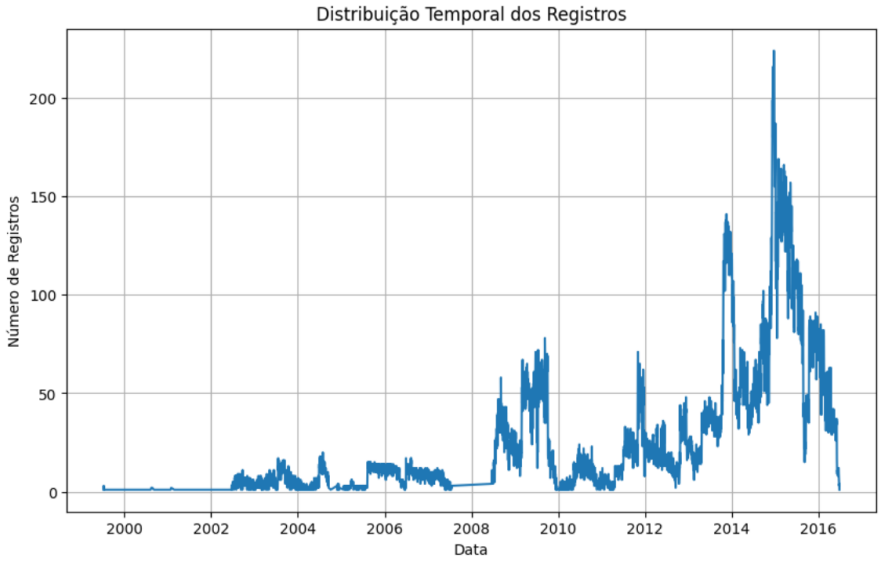

Fig. 6 Analysis of the Temporal Distribution of Records

Figure 7 illustrates the distribution of Panthera onca records by individual animal, highlighting significant disparities in the volume of data collected for each monitored individual. The distribution is highly asymmetric, with some animals having a much higher number of records than others. The most monitored animal has over 10,000 records, while many others have fewer than 1,000. There is a long tail in the distribution, indicating that a significant number of animals have minimal data available. A total of 117 animals were monitored, showing that data collection efforts were spread across various individuals, yet with considerable variation in the amount of data per animal. The discrepancies in the number of records could be attributed to factors such as differences in GPS tracker availability, monitoring duration, or behavioral variations between individuals. Animals with fewer records may have been monitored for shorter periods or may have experienced device malfunctions. The concentration of data in a small number of individuals may introduce bias into the study, underscoring the need for balanced data collection to ensure broader representation.

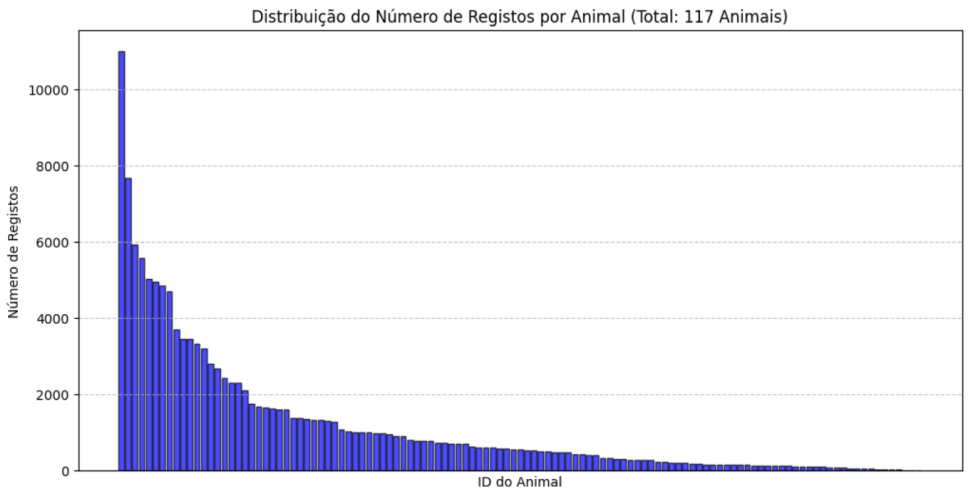

Fig. 7 Analysis of the Distribution of Records by ID

Figure 8 presents the distribution of distances traveled between consecutive records of Panthera onca, providing insights into the species' movement patterns over time. The distribution is highly asymmetric, with most movements being short distances, as evidenced by a large number of records below 500 meters. There is a long tail in the distribution, indicating that longer movements, although less frequent, still occur, with some records showing displacements greater than 3,000 meters. A significant peak near 0 meters suggests periods where the animals were stationary or had minimal movement between measurements. The predominance of short distances implies that Panthera onca frequently engages in low-activity behaviors or patrols confined areas. Longer movements may be linked to hunting, evading predators, or exploring new territories. The pattern observed may also be influenced by the time intervals between records, as shorter intervals tend to capture more frequent, shorter movements.

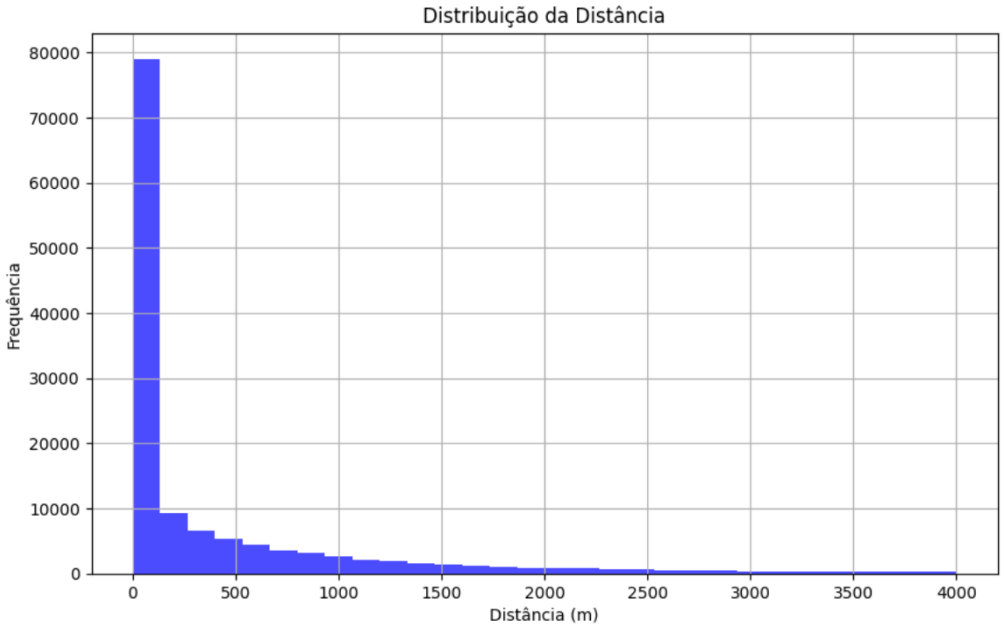

Fig. 8 Analysis of the Distribution of Distances Traveled by Panthera onca

Figure 9 presents the distribution of recorded speeds for Panthera onca, shedding light on the species' movement and behavior patterns. The distribution is highly asymmetric, with most records showing very low speeds, particularly values close to 0 m/s. The long tail indicates that, although the majority of observations are concentrated around slow speeds, there are occasional instances of faster movements. A significant peak near 0 m/s suggests periods of inactivity or very slow movement. The predominance of low speeds implies that Panthera onca spends a considerable amount of time resting or moving slowly. The higher speed values, though less frequent, are likely associated with rapid movements in pursuit of prey or evading threats. The overall low average speed could also reflect the fixed time intervals between records, which may not capture brief bursts of higher speed.

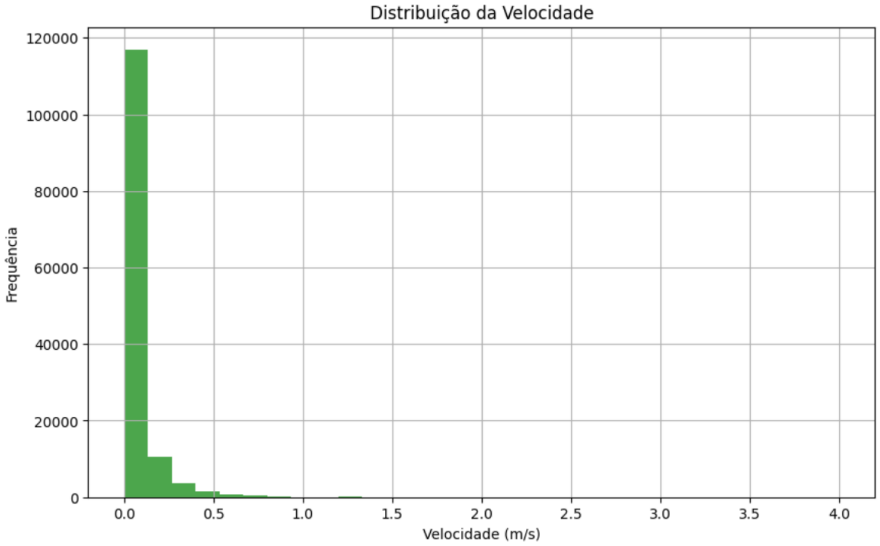

Fig. 9 Analysis of the Distribution of Speed of Panthera onca

Figure 10 illustrates the distribution of the change in direction angles of Panthera onca, providing insights into the species' movement and behavioral strategies. The distribution is bimodal, with two significant peaks around 0° and 180°, indicating that the animals often follow straight paths or make sharp turns. There is a lower frequency of moderate angle changes (50° to 130°), suggesting that gradual directional shifts are less common. The concentration at extreme angles suggests that Panthera onca typically moves in linear paths or makes large directional changes during its movements. The dominance of angles near 0° suggests that individuals often maintain a consistent direction, likely following familiar routes or trails. The higher frequency of angles near 180° may indicate abrupt changes in direction, possibly linked to hunting strategies or evasive maneuvers. The lower occurrence of intermediate angles suggests that the species' movement is driven by well-defined exploration and pursuit patterns.

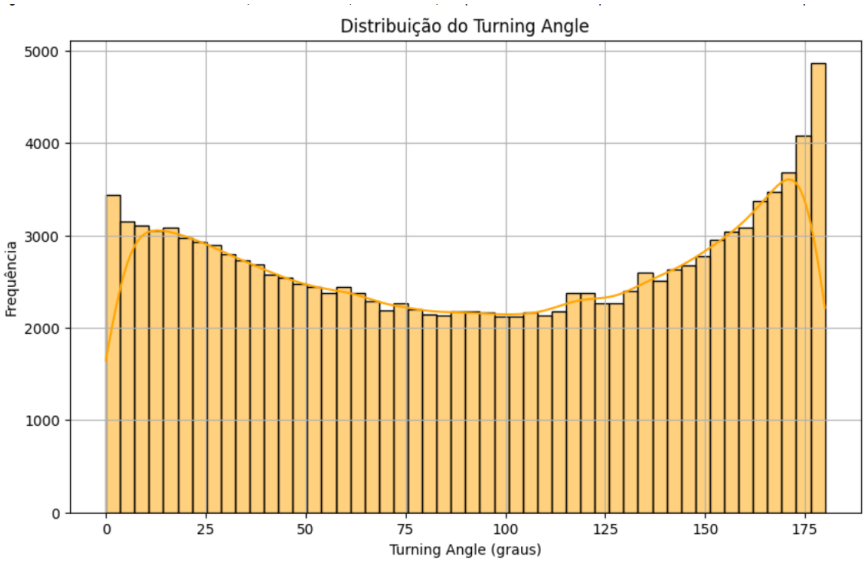

Fig. 10 Analysis of the Distribution of Change in Direction Angles of Panthera onca

Figure 11 displays the distribution of speed variation in Panthera onca, offering insights into changes in the species' movement patterns over time. The distribution is highly concentrated around 0 m/s², with most records showing little or no speed variation. There is a notable absence of significant accelerations or decelerations, suggesting that the animals maintain consistent speeds for most of their movement. The narrow range of variation indicates minimal changes in movement rhythm. The concentration at 0 m/s² suggests that Panthera onca typically moves at a stable speed, without sudden bursts of acceleration. The lack of extreme values could be attributed to uniform sampling intervals, which may not capture abrupt variations in speed. This pattern likely reflects territorial patrols or movement through predictable environments.

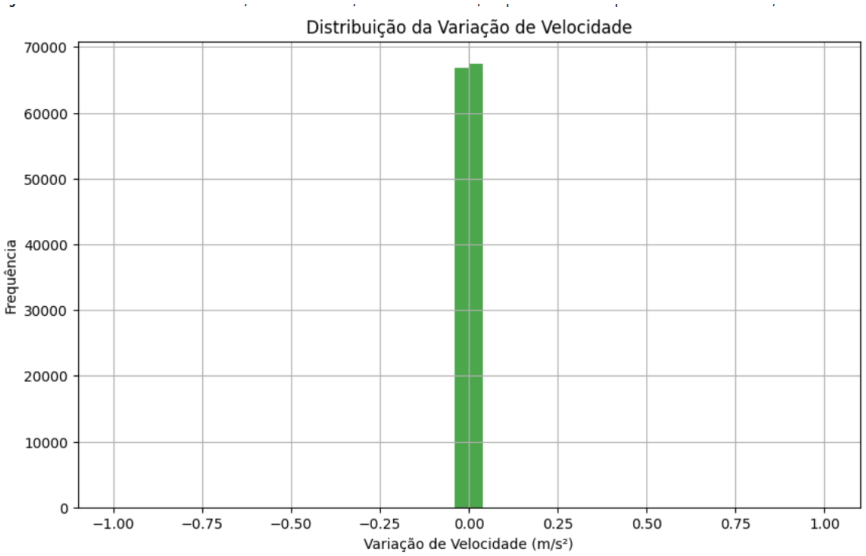

Fig. 11 Analysis of the Distribution of Speed Variation in Panthera onca

Figure 12 illustrates the distribution of Panthera onca records throughout different periods of the day, offering insights into the species' activity patterns. The highest number of records occurs during the early morning hours, indicating that Panthera onca is most active at this time. The number of records during the morning and afternoon are fairly balanced, suggesting moderate activity levels. However, the lowest frequency is observed during the night, which, despite the species' partially nocturnal behavior, suggests either periods of rest or possible limitations in data collection at night. The peak activity during dawn supports the idea that Panthera onca is predominantly crepuscular and nocturnal, with movement concentrated in those time frames.

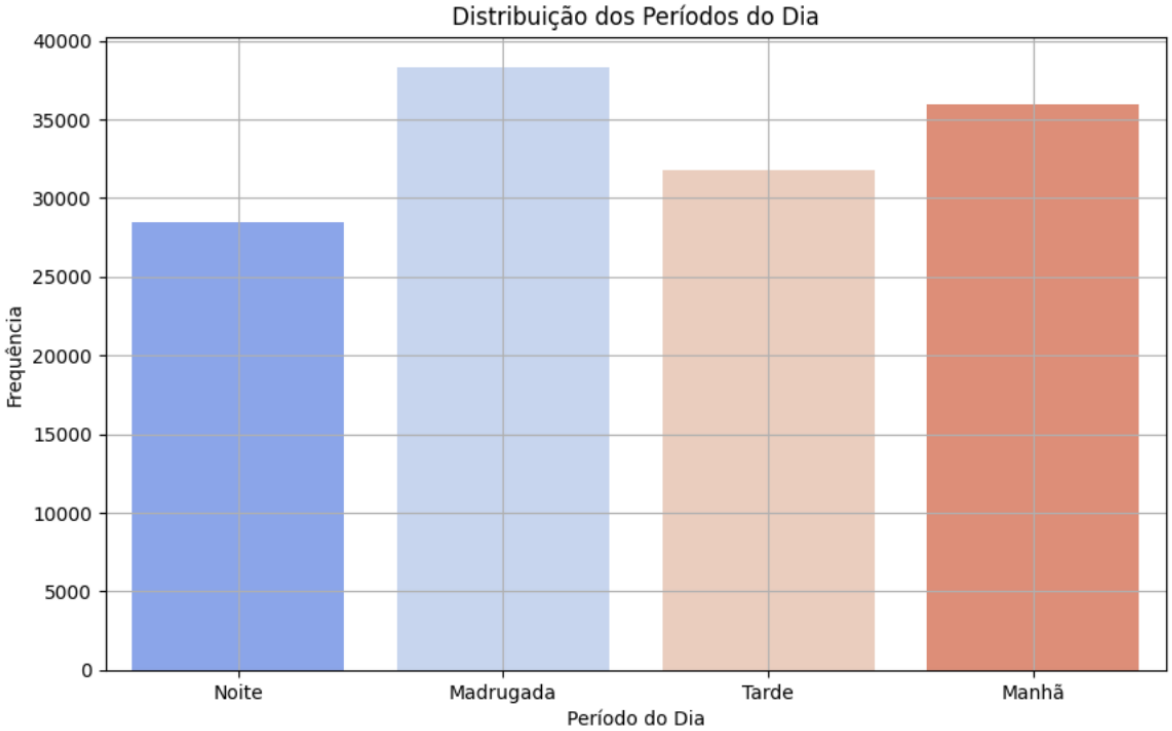

Fig. 12 Analysis of the Distribution of Records Throughout the Day

Figure 13 illustrates the distribution of Panthera onca records across different seasons, allowing for an evaluation of seasonal activity patterns. The highest number of records occurs in the summer, suggesting increased activity or easier monitoring during this period. The distribution in spring, autumn, and winter is relatively balanced, with only small variations. Winter shows a slight decrease in the number of records, although the difference is not extreme. The increased activity in summer may be linked to environmental factors such as better prey availability or favorable weather conditions for movement. The slight dip in winter could indicate behavioral changes, with animals possibly reducing activity to conserve energy. Additionally, data collection may be affected by weather conditions, such as rain or fog, which could hinder detection and tracking efforts.

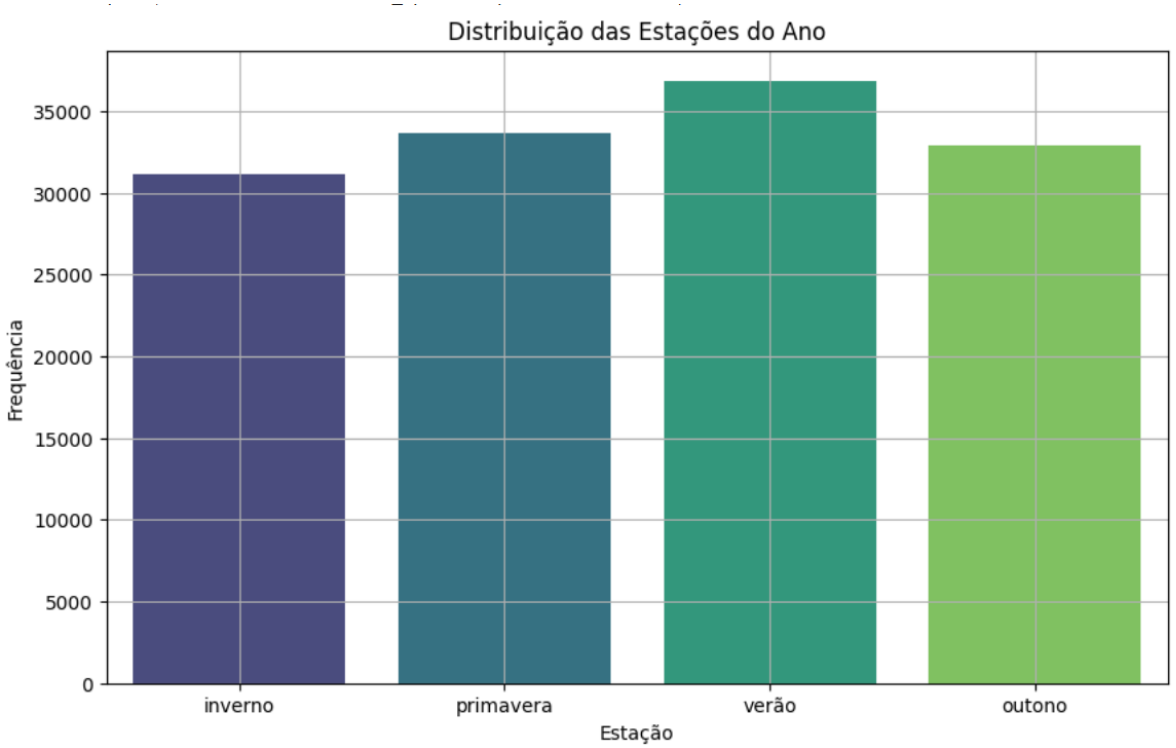

Fig. 13 Analysis of the Distribution of Records Throughout the Day

Figure 14 shows the distribution of Panthera onca records by sex, providing insight into the representativeness of males and females in the dataset. The number of records is higher for males, with individuals of this sex being more frequently monitored compared to females. Despite this slight difference, the overall distribution remains relatively balanced. The higher number of male records may reflect increased monitoring efforts for males, potentially due to their territorial behavior, larger home ranges, or specific characteristics of the study. The difference could also be influenced by the availability of females marked for tracking or variations in mortality rates between the sexes.

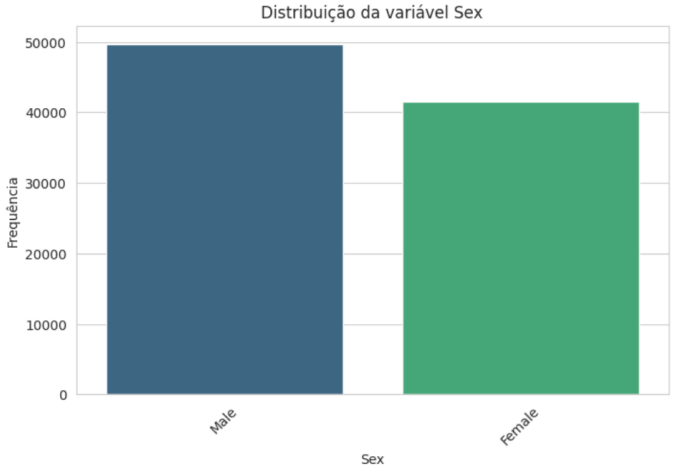

Fig. 14 Analysis of the Distribution of Panthera Onca by Sex

Figure 15 illustrates the distribution of Panthera onca records by sex, providing insight into the representation of males and females in the sample. Males account for a higher number of records, which may suggest that they are

more frequently monitored, possibly due to their larger territories, territorial behavior, or other characteristics of the study. However, the distribution remains relatively balanced overall. This difference could also reflect a lower availability of females tagged for tracking or a higher mortality rate in females compared to males.

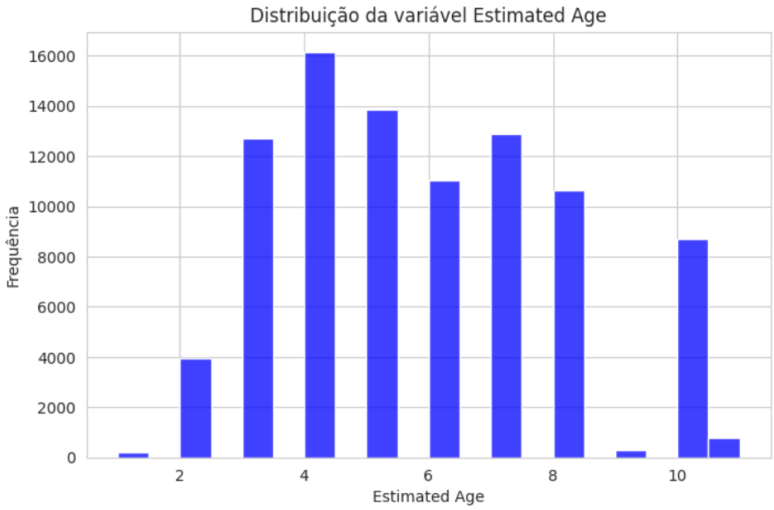

Fig. 15 Analysis of the Distribution of Panthera Onca by Sex

Figure 16 shows the distribution of temperature and humidity in the Panthera onca records. It is important to note that these are not real data, but rather simulated values. For temperature, the majority of records fall between 23°C and 32°C, with a peak in frequency around 26°C to 30°C. This suggests that the Panthera onca is most commonly found in environments within this temperature range. The distribution is relatively symmetric, with few extreme values below 22°C or above 34°C. Regarding humidity, the data show a predominant range of 90% to 100%, characteristic of tropical environments. There is a strong accumulation near 100%, which likely reflects a bias introduced by the model used to generate these data. The distribution is highly skewed, with little variation, suggesting that the model may underestimate drier conditions. The environmental conditions recorded point to the Panthera onca inhabiting warm and humid areas, typical of tropical forests and dense jungle environments. The stable temperature range indicates that tracking often occurs during periods with predictable conditions or in regions with little thermal variation. The high humidity levels align with the species' tropical habitat, though the excessive concentration at 100% suggests a need for model adjustments to better capture variations in real-world environmental conditions.

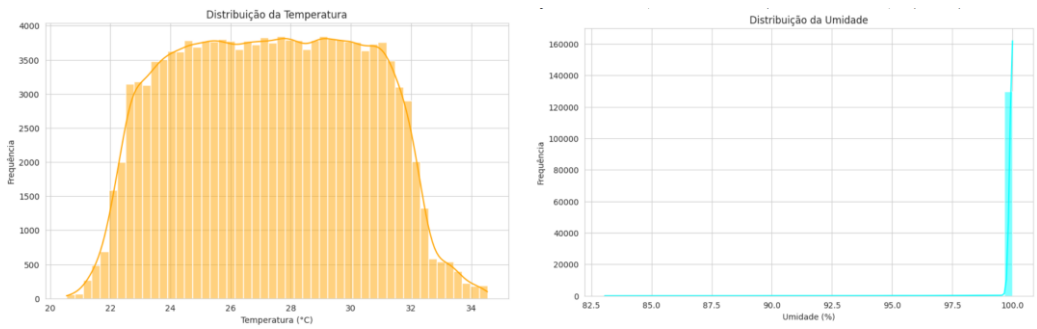

Fig. 16 Analysis of the Distribution of Temperature and Humidity in Panthera Onca Records

Graphs of Fig. 17 present the distribution of elevation and distance from water recorded in the fictional data.

For elevation, most records are concentrated around 200 meters, suggesting that the Panthera onca is being monitored in predominantly low-lying terrain. There are small peaks in altitudes between 150m and 180m, indicating the presence of some individuals in slightly elevated areas, though there is little variation in altitude overall. The highly concentrated distribution implies that the study is focused in a specific geographical area with limited elevation variation. Regarding distance from water, there is a significant peak around 1000 meters, indicating that most records were taken at this distance from water sources. Fewer records are found at very short (<500m) or very long (>4000m) distances. The uneven distribution suggests that certain study areas may have limited access to water sources, or the Panthera onca's movement patterns are influenced by proximity to rivers and lakes. The concentrated elevation distribution suggests that the study is located in a lowland region, possibly near floodplains or wetlands. The relationship with water indicates that the Panthera onca strategically moves to maintain access to water sources, likely influenced by prey availability or territorial behavior.

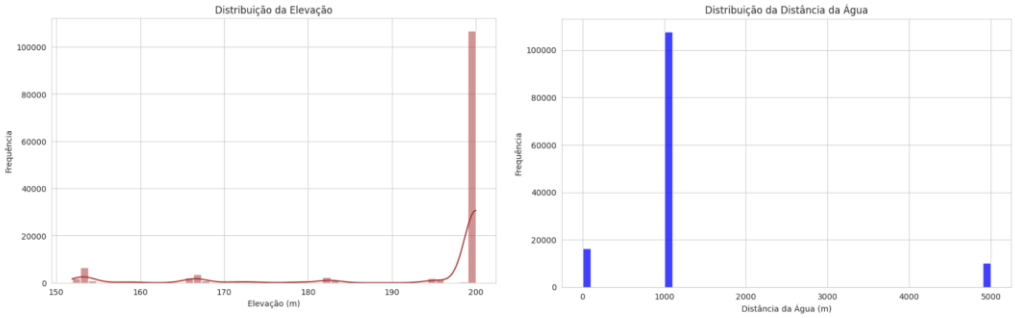

Fig. 17 Analysis of the Distribution of Elevation and Distance from Water

Figure 18 presents the distribution of movement categories for the Panthera onca, allowing an assessment of the species' activity patterns. The majority of the records classify the animals as either stationary or walking, indicating that these two states are predominant in the dataset. There are very few records for the "Running" category, which may suggest that running behavior is infrequent or that such moments were not captured often in the sampling process. The "Undefined" category is present but represents a small portion of the data, likely due to limitations in the automatic classification of movement. The predominance of stationary states suggests that the Panthera onca spends extended periods resting or waiting. The frequent walking records indicate regular movements, likely related to territorial patrolling or prey search. The low number of running records could be explained by the rarity of active chasing behavior or limitations in the GPS sampling rate.

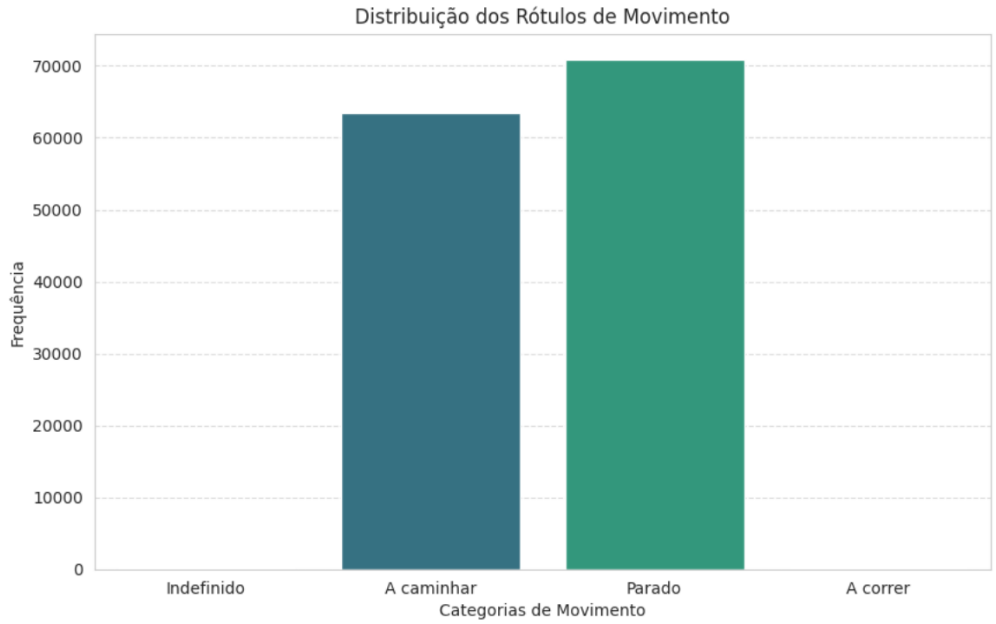

Fig. 18 Analysis of the Distribution of Elevation and Distance from Water

The results of the machine learning models, including Random Forest and XGBoost, were evaluated in detail within a Jupyter Notebook environment, providing a clear and reproducible presentation of the findings. The results included various performance metrics, such as accuracy, precision, recall, F1-score for classification tasks, and Mean Squared Error (MSE) for regression tasks, as well as additional insights gained from feature importance analysis and model evaluation.

During the training phase, the Random Forest model achieved excellent results. For regression tasks, predicting latitude and longitude, the model demonstrated a strong fit to the data with a Root Mean Squared Error (RMSE) of 0.1471 for latitude and 0.1438 for longitude. The R^2 values for both latitude and longitude were exceptionally high, at 0.9998, indicating that the model was able to explain nearly all the variance in the data. In terms of classification, Random Forest delivered perfect results with an accuracy of 1.0000. The precision, recall, and F1-score for all classes (0, 1, and 2) were also perfect (1.00), reflecting the model's ability to perfectly classify the jaguar's behavioral states on the training set.

However, when the Random Forest model was tested on unseen data, the performance remained strong but slightly less perfect. The RMSE for latitude increased to 2.0569, with an R^2 of 0.9252, indicating a slight drop in predictive accuracy. For longitude, the test RMSE was 1.0884 with an R^2 of 0.9568, showing that the model still performed well but not as perfectly as on the training data. In terms of classification, Random Forest achieved an accuracy of 1.0000 on the test set, although recall for class '1' was significantly lower (50%) due to the small number of instances for this class.

The XGBoost model, on the other hand, showed remarkable performance across all tasks. For regression tasks, XGBoost outperformed Random Forest with a lower RMSE and perfect R^2 values during the training phase. Specifically, the RMSE for latitude was 0.0756, and for longitude, it was 0.0545, both with R^2 values of 1.0000. This indicated that XGBoost had an even stronger fit to the training data than Random Forest. In terms of classification, XGBoost achieved perfect accuracy (1.0000) and perfectly classified all classes on the training set, with precision, recall, and F1-scores of 1.00 for all classes.

However, similar to Random Forest, XGBoost's performance on the test set showed a slight decline in regression accuracy. The test RMSE for latitude was 2.7467 with an R^2 of 0.8667, and for longitude, the RMSE was 2.4984

with an R^2 of 0.7726. While these results still indicated good predictive accuracy, the model performed less effectively on the test data than it had during training. In classification, XGBoost achieved an accuracy of 0.9987 on the test set, which was very close to Random Forest's performance. The model's precision, recall, and F1-scores were nearly perfect, though recall for class '1' was again lower, as observed with Random Forest.

An ensemble approach was also tested by combining the strengths of both Random Forest and XGBoost. For classification, a Voting Classifier was employed, using both models as base learners. The soft voting mechanism combined the predicted probabilities from each model to determine the final classification, leading to perfect performance on both training and test sets, with an accuracy of 1.0000. The ensemble approach was particularly useful for regression tasks, where a Stacking method was applied. Random Forest and XGBoost were used as base models, and their predictions for latitude and longitude were combined using a Linear Regression meta-model. This approach improved the regression accuracy, with an RMSE of 0.0545 and an R^2 of 1.0000 for latitude on the training data. However, the performance on the test set for regression was less optimal, with RMSE values of 2.4659 for latitude and 2.4984 for longitude, and corresponding R^2 values of 0.7785 and 0.7726.

Overall, the results indicated that both Random Forest and XGBoost performed extremely well, with Random Forest providing a good balance between efficiency and interpretability, and XGBoost excelling in terms of regression accuracy. While the ensemble approach improved the regression predictions slightly, the classification task did not show significant improvements over using either model individually. The ensemble method combined the strengths of both models, but given that both Random Forest and XGBoost achieved near-perfect classification results on their own, the benefit of the ensemble was more evident in regression tasks.

Despite the promising results, it should be noted that the computational cost for hyperparameter tuning with XGBoost was prohibitive, and therefore, the models were used with default hyperparameters. This limitation in tuning was acknowledged, as it could have potentially enhanced the models' performance further, particularly in terms of generalization on the test set. Nonetheless, both models, and particularly the ensemble method, demonstrated strong predictive power, providing valuable insights into jaguar movement dynamics based on GPS tracking data.

7. CONCLUSION AND FUTURE WORK

This study successfully demonstrated the effectiveness of machine learning models, particularly Random Forest, in classifying jaguar movement behaviors based on GPS-tracked data. By leveraging movement-related features such as speed, distance traveled, and turning angles, the model was able to distinguish between different behavioral states with high accuracy. The preprocessing steps, including handling missing values, feature engineering, and categorical encoding, contributed to improving model performance and ensuring data consistency. Additionally, the comparative analysis between Random Forest and XGBoost provided valuable insights into the strengths and weaknesses of different classification approaches. The ensemble method further enhanced predictive performance by combining the advantages of both models.

Despite the success of the models, there are several areas for improvement. One of the main limitations is the exclusion of additional environmental factors that could influence movement behavior. Future work should focus on integrating terrain, climate, and other contextual data to enhance predictive accuracy. Additionally, while traditional machine learning models performed well, deep learning techniques such as recurrent neural networks (RNNs) or transformer-based architectures could be explored if a larger dataset becomes available. These models may capture more complex temporal dependencies and improve classification performance.

Another potential improvement is the deployment of the trained model in a real-time monitoring system for conservation efforts. By integrating the model into tracking systems, researchers and conservationists could receive automated alerts on unusual movement patterns, helping to identify potential threats such as poaching or habitat disruptions. Further work is also needed to optimize hyperparameter tuning without excessive computational costs, ensuring that models remain both effective and efficient in real-world applications.

Overall, this study provides a strong foundation for using machine learning in wildlife behavior analysis. Future research should build upon these findings by incorporating additional data sources, refining model architectures, and exploring real-time deployment strategies to support conservation efforts and ecological research.