# Exploratory Data Analysis of Irregular Shark Data

## Assessing Observation Gaps, Missingness, and Spatial Heterogeneity Across Individuals

**Madelyne Ruijia Zhang**

## 1 Data Overview

The dataset consists of satellite telemetry records from an initial sample of 60 individual whale sharks. Each observation contains a shark identifier, geographic location given by longitude and latitude, a timestamp, as well as the associated Argos satellite system and error class. The data are inherently irregular in time, with observation intervals varying substantially both within and across individuals.

Given this structure, exploratory analysis is conducted at the shark level to examine how observation density varies across individuals. In particular, the number of recorded locations per shark differs markedly, with some individuals having dense trajectories and others represented by only a small number of observations.

## 2 Data Cleaning Summary

The raw telemetry data contain 15 observations with missing timestamps (date). These observations do not provide valid temporal information and were therefore excluded prior to exploratory analysis. No missing or invalid values were detected in the longitude or latitude coordinates.

After removing observations with missing timestamps, the data were examined for duplicate records and temporal consistency within individuals. One individual (ID 203645) was found to contain an exact duplicated timestamp, which resulted in non-increasing time order within the trajectory. This duplicated record was removed, and all trajectories were ordered to ensure strictly increasing timestamps. These removed observations represent less than 0.12% of the total dataset. After cleaning, the final dataset contains 12,157 observations from 60 individual sharks, spanning the full study period. All subsequent analyses are based on the cleaned dataset with valid coordinates, unique timestamps, and consistent temporal ordering.

## 3 Observation Density Across Individual Sharks

A small subset of individuals contributes a disproportionately large number of observations, as illustrated by the Lorenz curve of observation counts (Figure 2), which shows a strong deviation from the line of equality. Most sharks contribute only a small fraction of total observations, while a small number of individuals dominate the dataset.

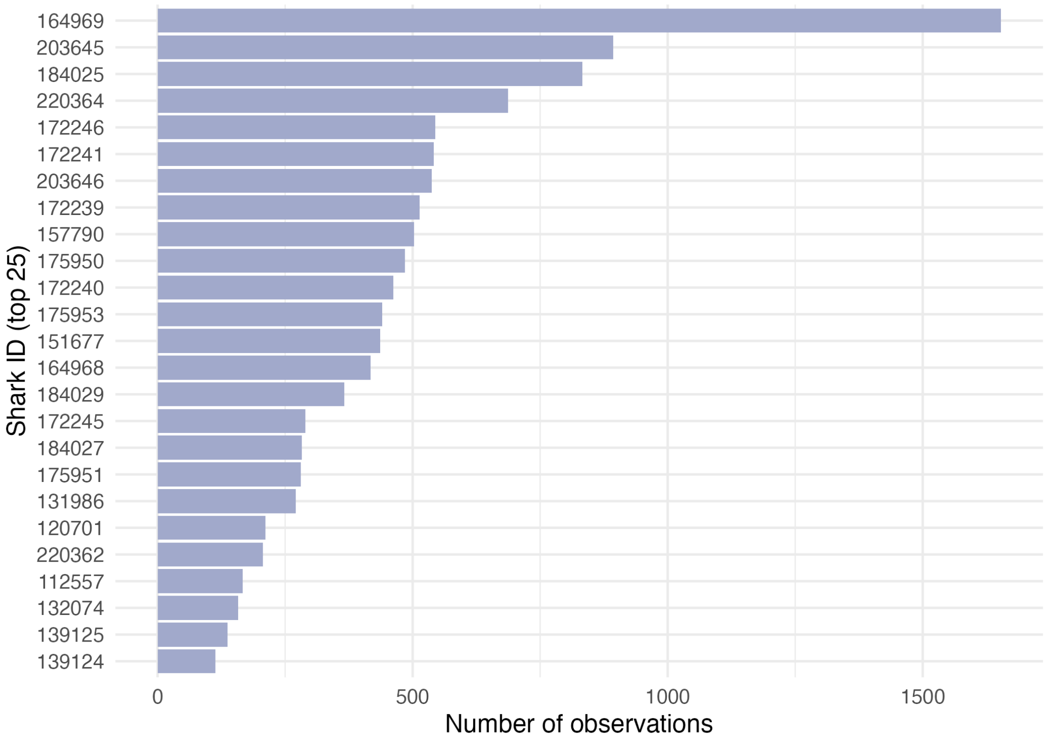

**Figure 1.** Number of observations for the 25 sharks with the largest number of recorded locations.

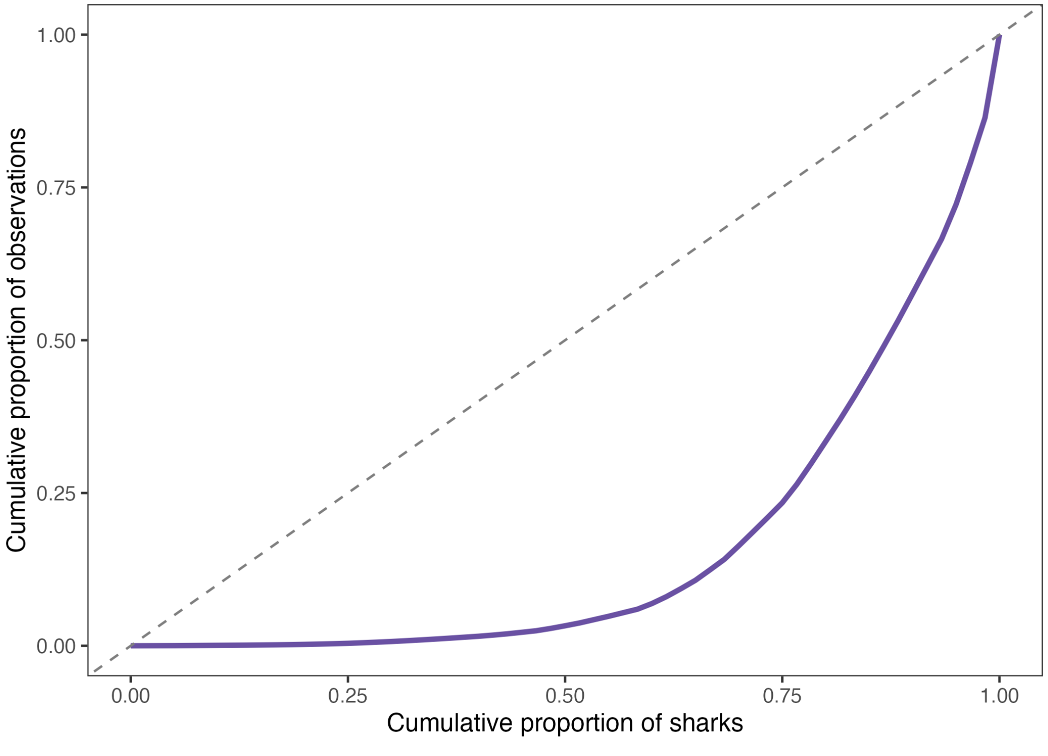

**Figure 2.** Lorenz curve of observation counts across sharks.

Table 1 summarizes the distribution of observation counts across all sharks. The median number of observations per shark is 54.5, whereas the mean is substantially higher (202.9), reflecting strong right skewness in the distribution. Observation counts range from as few as one recorded location to over 1,600 locations for the most frequently observed individual.

Table 1: Summary of Number of Observations per Shark

| Mean | Median | Min | Max |
|---|---|---|---|
| 202.9 | 54.5 | 1 | 1655 |

This heterogeneity in observation density is an important feature of the dataset and motivates careful consideration of individual-level data in subsequent analyses. In particular, individuals with very sparse observations may provide limited information about movement dynamics, while individuals with dense tracks are likely to dominate model fitting and inference.

In particular, a subset of sharks has extremely sparse tracks. Specifically, 8 individuals have fewer than three recorded locations, and 24 individuals have fewer than 20 observations overall. Because state-space movement models require at least three observations per individual to estimate movement parameters, sharks with fewer than three observations are excluded from subsequent modeling. The remaining individuals retain substantial variability in observation density, which may influence the relative contribution of different sharks to model estimation and uncertainty. The degree of heterogeneity implies that models assuming regular sampling or equal information across individuals may be inappropriate.

# 4 Temporal Structure and Observation Processes

## 4.1 Temporal Structure: Observation Gaps

The temporal spacing between consecutive observations varies substantially across and within individual sharks. Figure 3 shows the distribution of time gaps ($\delta_t$) between consecutive observations, measured in hours, on a logarithmic scale, highlighting strong irregularity in the sampling process.

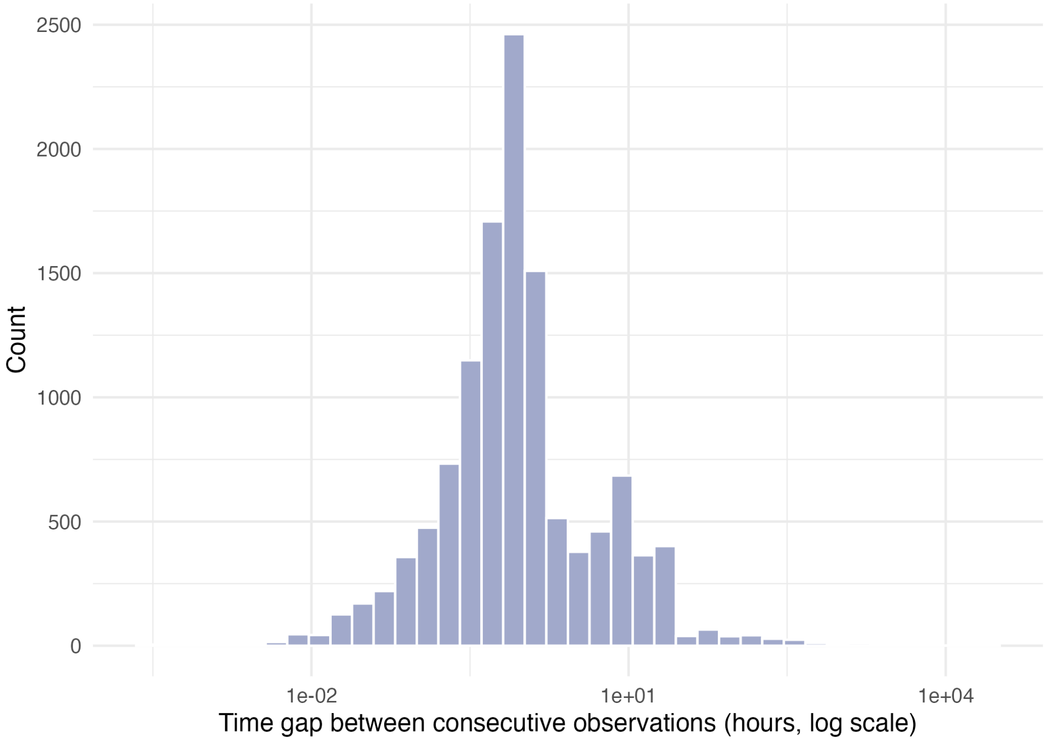

**Figure 3.** Distribution of time gaps between consecutive observations across all sharks.

Across all individuals, most consecutive observations are separated by relatively short time intervals, with the distribution concentrated around gaps on the order of one hour. However, the distribution is highly right-skewed, with a long tail indicating the presence of occasional long gaps between observations. These long gaps extend to several orders of magnitude larger than the typical sampling interval.

To summarize temporal irregularity at the individual level, individual-level median and maximum time gaps were calculated for each shark. The median of individual-level median gaps is approximately 0.93 hours, suggesting that for most sharks, typical observation spacing is close to hourly. In contrast, the maximum gap observed for at least one individual is 25,638.46 hours, which corresponds to approximately 2.93 years between recorded locations.

This combination of relatively dense short-term sampling and sporadic extreme gaps indicates that the telemetry data are highly irregular in time. Such temporal structure has important implications for movement modeling. In particular, long gaps may reflect extended periods without detection rather than continuous movement, and they complicate direct interpretation of displacement over time. These features motivate the use of state-space models that explicitly account for irregular observation times and measurement error instead of relying on regularly sampled trajectories or naive interpolation. Extended gaps suggest that discrete-time formulations with fixed step sizes would require further interpolation.

In addition to the marginal distribution of time gaps, the temporal irregularity also exhibits clear structure. For individual sharks, periods of frequent observations with short time gaps are often followed by extended intervals with no detections. Figure 4 illustrates this pattern for a representative individual (ID 164969), where short gaps cluster in time rather than occurring independently.

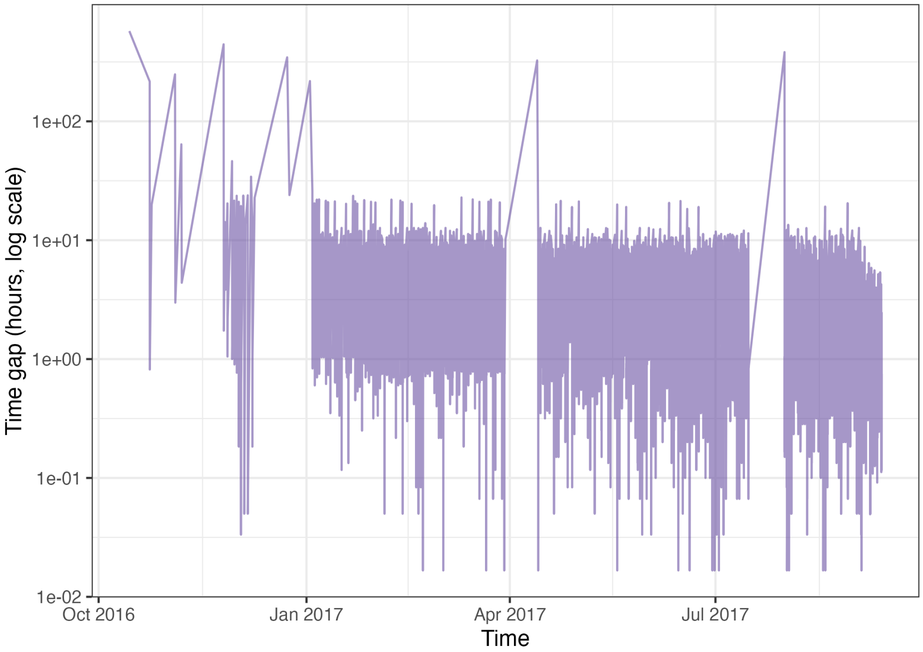

**Figure 4.** Time gaps between consecutive observations over time for a representative individual (log scale).

This structure is further reflected at the population level. Figure 5 shows the distribution of run lengths of short gaps across all sharks, where long sequences of consecutive short gaps are common. The presence of long runs indicates burst sampling behavior, rather than randomly varying observation intervals. Together, these patterns suggest that temporal irregularity in the telemetry data arises from structured sampling regimes, characterized by bursts of dense observations separated by silent periods.

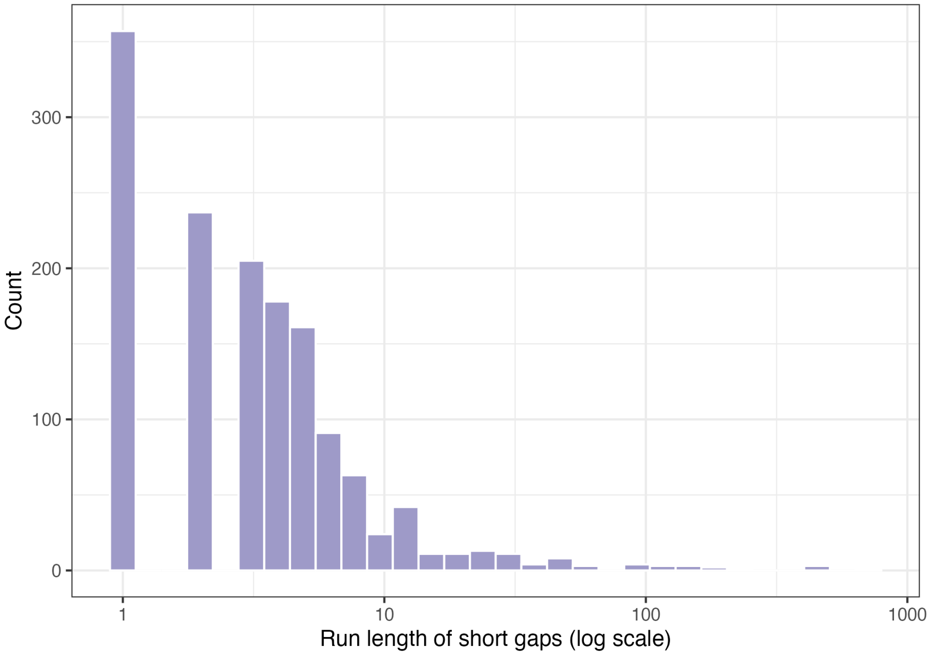

**Figure 5.** Distribution of run lengths of short time gaps across all individuals (log scale).

## 4.2 Argos Error Class and Step Characteristics

Although Argos error class is recorded for each location, it was not previously incorporated into the exploratory analysis. A descriptive check is therefore used to see whether measurement error is reflected in basic movement summaries. Figure 6 shows that observations are unevenly distributed across error classes, with most locations assigned to classes 1–3 and B, and relatively few observations in class Z.

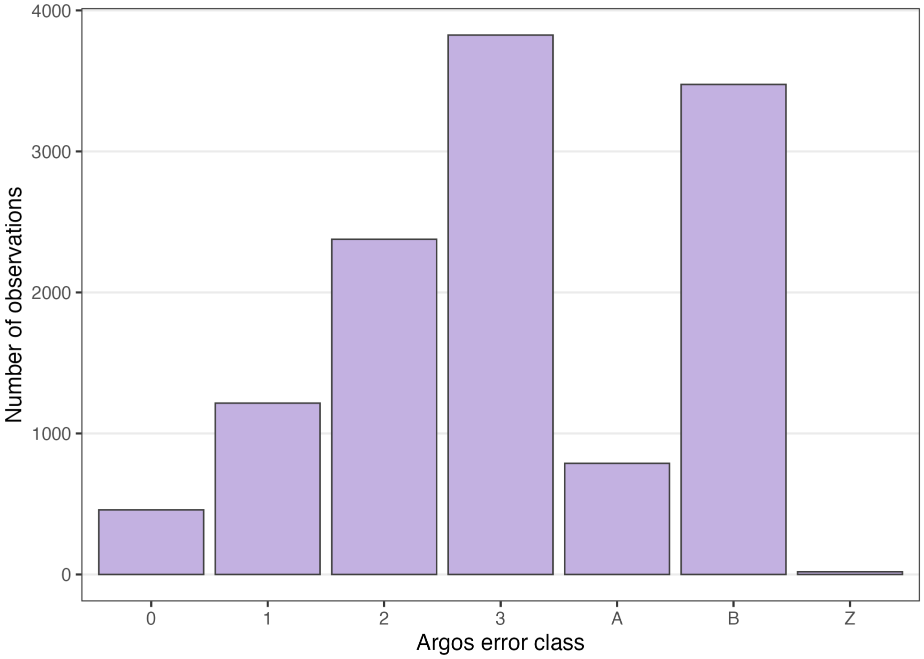

**Figure 6.** Number of observations by Argos error class.

Figure 7 reveals clear differences in step length distributions across error classes. Locations in classes 1–3 tend to produce more concentrated step lengths with fewer extreme displacements. In contrast, classes A and B show greater dispersion and a higher frequency of large steps, while class Z, though rare, exhibits the most extreme step lengths. These patterns indicate that many of the large step lengths seen in the raw tracks are likely driven by lower-quality location fixes, rather than reflecting actual movement.

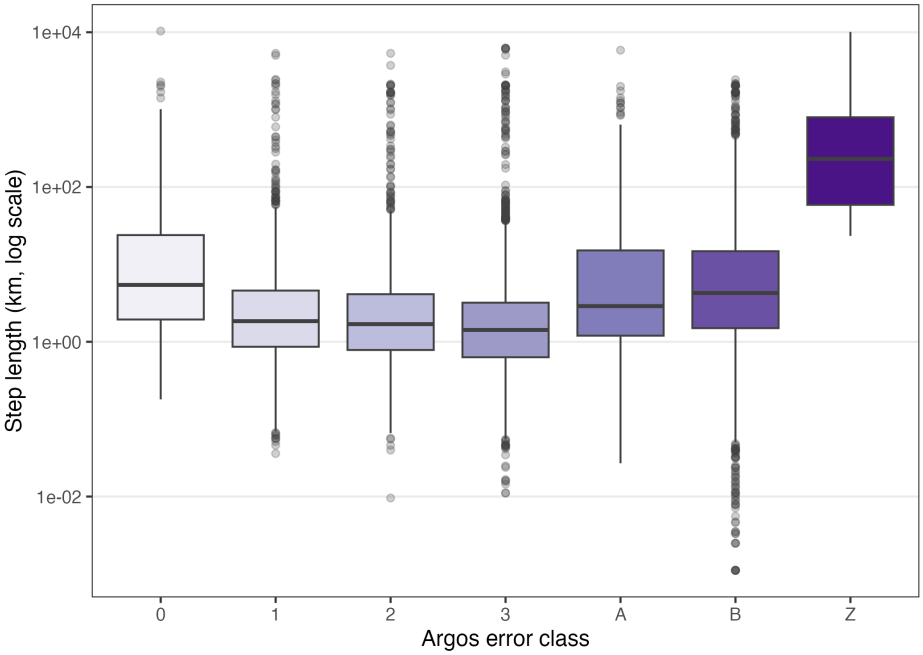

**Figure 7.** Distribution of step lengths between consecutive locations by Argos error class (log scale).

As a result, large step lengths in raw tracks are more likely to occur when observations belong to classes A, B, or Z. These extreme displacements are therefore unlikely to represent typical movement behavior and are unlikely to reflect continuous movement between locations. This finding motivates the next stage of analysis, in which state-space movement models are used to explicitly account for Argos measurement uncertainty when estimating underlying movement dynamics.

# 5 Spatial EDA: Raw Tracks

## 5.1 Overview of All Raw Shark Tracks by Telemetry Data

Raw movement tracks for all sharks are shown in Figure 8 by connecting consecutive recorded locations within each individual. This plot is intended to give a general picture of the spatial coverage and overall movement patterns in the dataset, without focusing on individual sharks.

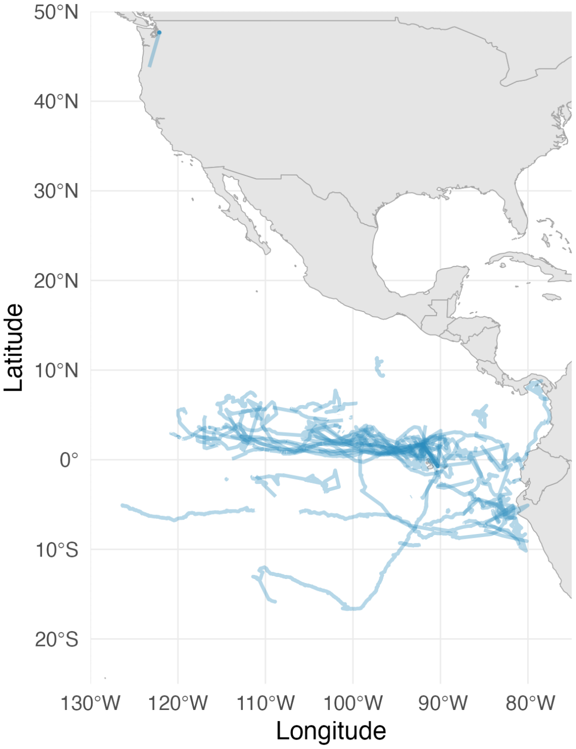

**Figure 8.** Overview of raw whale shark movement tracks across all individuals.

When consecutive observations are directly connected, long straight-line segments may appear due to extended time gaps between detections. For visualization, tracks were therefore split whenever the time gap between successive observations exceeded 720 hours (30 days) or when the spatial displacement between locations exceeded 500 km. This step was applied solely to reduce visually misleading long-distance connections that could be biologically implausible, while preserving local movement structure.

After applying these constraints, the resulting map summarizes the primary spatial domain covered by the telemetry data and highlights heterogeneity in movement extent across individuals. This overview suggests that subsequent analyses should examine individual-level movement patterns in greater detail.

## 5.2 Selected Individual Tracks: Representative Movement Patterns

Based on the faceted visualization of individual raw locations, four sharks were selected to illustrate key sources of heterogeneity in the telemetry data (Figure 9). These individuals differ markedly in the number of observations, temporal continuity, and apparent movement structure, providing useful contrast for subsequent modeling and diagnostics.

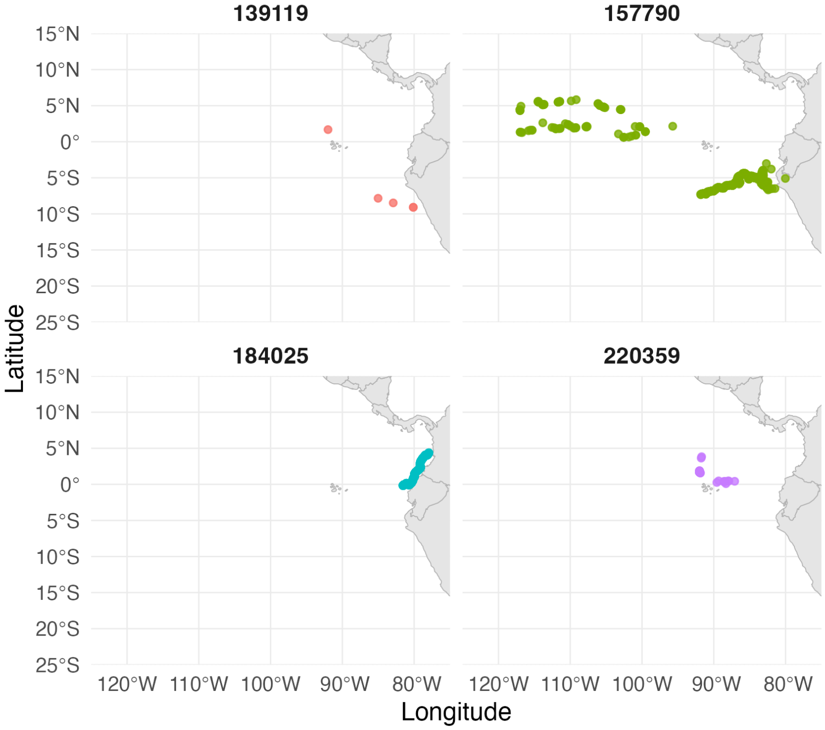

**Figure 9.** Raw location data for four selected whale sharks, shown as individual faceted maps.

Shark 139119 represents an extremely sparsely sampled case, with only five recorded locations. The limited number of observations provides minimal information about movement structure and serves as an example of trajectories for which detailed movement modeling would be unreliable. This individual highlights the lower bound of data availability present in the dataset.

Shark 157790 exhibits a moderate number of observations, but with a pronounced temporal gap between detections. In particular, there is at least one instance where two consecutive recorded locations are separated by a long interval, with one observation on 2018-03-28 followed by the next available observation on 2020-01-01. This pattern appears in Figure 9 as two clusters of points separated by a clear break, illustrating how long gaps in sampling can cause an individual's track to appear fragmented.

Shark 220359 contains 31 recorded locations. Although the total number of observations is relatively small, the spatial pattern remains interpretable when visualized, forming a coherent localized track. This individual represents cases with limited but usable data, where movement structure can still be visually assessed despite low sampling frequency.

In contrast, shark 184025 provides a well-sampled example, with 833 observations and relatively regular temporal spacing. The resulting trajectory appears continuous and spatially coherent, making it suitable as a reference case for typical movement behavior within the dataset. This individual serves as a baseline example for later modeling and comparison.

Together, these four sharks capture the heterogeneity of observation density and temporal irregularity present in the telemetry data. They provide a set of contrasting cases that inform both the interpretation of raw movement patterns and the methodological choices made in subsequent modeling steps.

# 6 EDA Summary and Implications for Modeling

The exploratory analysis shows that the telemetry data vary substantially across sharks in terms of observation density, temporal spacing, and spatial extent. Some individuals are represented by dense and nearly continuous tracks, while others have only a small number of locations or fragmented records separated by long gaps. This uneven data structure is a defining feature of the dataset.

Differences in observation density imply that sharks do not contribute equally to movement inference. Individuals with many observations are likely to dominate estimation, whereas sharks with sparse data provide limited information about movement behavior. This suggests that treating all sharks as equally informative may be unrealistic. Instead, individual-level structure may need to be acknowledged in modeling, for example by allowing for random effects or other hierarchical components that account for between-shark variability.

Temporal irregularity further affects how movement should be modeled. Short gaps are often clustered in time, while occasional gaps span months or even years. These long gaps are unlikely to represent continuous movement and instead reflect extended periods without detection. As a result, models that rely on fixed time steps or regular sampling may be poorly suited to the data. Continuous-time formulations or approaches that allow movement uncertainty to increase with gap length may be more appropriate. For individuals with exceptionally large gaps, it may also be preferable to split trajectories into separate segments rather than forcing a single continuous track.

Spatial EDA indicates that apparent long-distance movements in raw tracks are often driven by large temporal gaps or low-quality locations rather than true rapid displacement. This reinforces the need for modeling approaches that separate the observation process from the underlying movement process and explicitly account for measurement error. When environmental covariates are introduced at later stages, the strong individual-level heterogeneity observed here also suggests that stratification or individual-level random effects may be needed to avoid over-interpreting patterns driven by the small subset of sharks.

Overall, the EDA highlights that the main challenges in this dataset arise from irregular sampling and heterogeneous data quality across individuals. These features motivate modeling choices that adapt to variable observation density and time gaps, rather than relying on simplified assumptions of regular sampling or uniform information across sharks.

# 7 Appendix A. Software and Data Handling

All analyses were conducted in R. Spatial coordinates were handled as geographic coordinates (longitude and latitude, WGS84). Temporal variables were processed by converting timestamps to POSIXct format and computing individual-level time differences between consecutive observations. These time differences form the basis of all temporal gap summaries and visualizations presented in the main text.