**Assignment: AI-Enhanced wildlife Corridor**

**SUBJECT:** Introduction to AI/ML

**Assistant Professor:** Ramavath Ganesh

**Student Name:** Ankit Singh Rawat

**Problem Context: Habitat fragmentation threatens biodiversity by blocking natural wildlife movement. Ecologists need AI-based tools to predict movement corridors, detect barriers, and recommend optimal connectivity paths between habitats.**

**This project applies AI + ML techniques (computer vision, clustering, graph algorithms, predictive modelling, reinforcement learning) to solve this problem efficiently**

**Section 1: A Taxonomy of Missingness in Telemetry Data**

The phenomenon of "missing data" in GPS telemetry is not monolithic. The cause, pattern, and duration of data gaps have profound implications for both the choice of treatment method and the validity of the final analysis. Understanding the underlying mechanism of missingness is the first and most crucial step.

* 1. **Mechanical or Logistical Failures**

This category includes catastrophic or predictable hardware issues. A GPS collar's battery may expire, the unit may detach from the animal, or the animal itself may perish. This typically results in a **terminal data gap**, where the time series for an individual abruptly ends.

**1.2 Environmental Obstruction**

GPS technology relies on a clear line of sight to multiple satellites. Dense forest canopies, deep canyons, or other complex terrain can temporarily block these signals, preventing a successful location fix.

This is a primary cause of **intermittent, short-to-medium duration gaps**. The pattern of missingness is often correlated with specific landscape features, meaning it is not completely random.

**1.3 Behavioural Occlusion**

An animal's own behaviour can directly cause data loss. For example, an animal denning in a cave or rock crevice will be shielded from GPS signals. A bear hibernating in a den for several months will produce a predictable, long-duration seasonal gap. This type of missingness is classified as **Missing Not At Random (MNAR)**, as the probability of data being missing is directly related to the unobserved behaviour (e.g., denning) and location.

**1.4 Comparative Overview of Missing Data Types**

|  |  |  |  |
| --- | --- | --- | --- |
| **Type of Missingness** | **Common Causes** | **Typical Gap Pattern** | **Implication for Analysis** |
| **Mechanical Failure** | Battery death, collar drop-off | Terminal; data stream ends permanently. | Reduces the total observation period for an individual. |
| **Environmental Obstruction** | Dense canopy, canyons, urban structures | Intermittent; short-to-medium gaps, often spatially clustered. | Can introduce spatial bias; analysis might underrepresent habitats with poor satellite reception. |
| **Behavioural Occlusion** | Denning, hibernation, sheltering | Intermittent; can be long-duration and seasonal. | **High risk of bias**. Naively ignoring these gaps can lead to the exclusion of critical habitats (e.g., dens) from the analysis. |

**Section 2: Diagnostic Analysis: Quantifying and Visualizing Data Gaps**

Before any intervention, a thorough diagnostic phase is required to understand the scope and nature of the missing data.

* **Gap Duration Histogram:** The first step is to calculate the time difference between every consecutive point for each animal. A histogram of these time differences (or "gaps") immediately reveals the data's temporal structure. It helps distinguish between the expected sampling interval (e.g., 1 hour) and anomalous long gaps that require special treatment.
* **Spatial Distribution of Missingness:** Mapping the start and end points of data gaps can reveal environmental correlations. If gap start-points consistently cluster in areas of dense forest or steep topography, it strongly suggests environmental obstruction is the primary cause.
* **Temporal Patterns of Missingness:** Plotting the frequency of missing points by the time of day or month of the year can identify behavioural patterns. For instance, a spike in missing data during nocturnal hours might suggest the use of specific resting sites (e.g., dens) that block GPS signals.

**Section 3: A Methodological Framework for Handling Missing Values**

The choice of method depends entirely on the duration of the gap, the likely cause of missingness, and the ultimate goal of the analysis.

**3.1 Deletion (Case Deletion)**

This involves simply ignoring the gaps and treating the remaining points as a continuous (but irregular) time series.

* **Pros:** Simple to implement.
* **Cons:** Severely biases any time-based analysis (e.g., calculation of speed, turn angles). It effectively assumes the animal teleported between points. **This method is generally unacceptable for movement analysis.**

**3.2 Single Imputation: Simple Interpolation Methods**

These methods "fill in" the gaps with a single, estimated value or path.

**Linear Interpolation:** This method connects the two points on either side of a gap with a straight line and places new points along that line at the regular sampling interval.

**Pros:** Simple, fast, and suitable for very short gaps (e.g., 1-2 missing points) where an animal is likely moving in a consistent direction.

**Cons:** Biologically unrealistic for longer gaps. It assumes constant speed and zero deviation, artificially straightening paths and underestimating distance travelled.

**3.3 Advanced Imputation: Model-Based & Stochastic Methods**

These methods are more computationally intensive but provide more biologically realistic and statistically robust solutions.

* **Correlated Random Walk (CRW):** This is a movement model that can be used to simulate a likely path within a gap. It assumes that the speed and direction of a step are correlated with the previous step, but also includes a random component. This produces a more realistic, non-linear path compared to simple interpolation.
* **Multiple Imputation (MI):** This is considered the gold standard for handling missing data in statistical modelling. Instead of filling the gap once, MI creates multiple (e.g., 5-10) plausible imputed datasets. The desired analysis (e.g., a habitat selection model) is run on *each* of these datasets, and the results are then pooled together.

**Pros:** The key advantage of MI is that it **accounts for the uncertainty** of the imputed values. The variation in results across the different imputed datasets provides a more honest and robust final estimate.

**Cons:** Complex to implement and requires more computational resources.

**3.4 Decision Framework for Method Selection**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Gap Duration** | **Likely Cause** | **Analysis Goal** | **Recommended Method** | **Rationale** |
| **Short (1-3 intervals)** | Environmental | Path visualization, simple speed calculation | **Linear Interpolation** | The introduced bias is minimal over very short periods of directed travel. |
| **Medium (4-24 intervals)** | Environmental | Formal habitat modelling, home range analysis | **Correlated Random Walk** | Provides a more biologically plausible path without the complexity of MI. |
| **Long (>24 intervals)** | Behavioural / Environmental | Formal habitat modelling | **Multiple Imputation** | Essential for robustly accounting for the high uncertainty associated with long gaps. |
| **Any Duration** | Behavioural (e.g., Denning) | Identifying critical habitats | **Do not impute.** Analyse the gap itself as a data point (e.g., "denning period"). | The missingness itself contains the most important behavioural information. Imputing would erase this signal. |

**Section 4: Critical Considerations**

* **The "No Free Lunch" Principle:** No imputation method is perfect. Every method makes assumptions about the movement process. The goal is to choose the method whose assumptions are most reasonable for the specific context.
* **Impact on Downstream Analysis:** Be aware that the chosen method will influence results. For example, linear interpolation will artificially decrease home range size estimates and straighten movement paths, potentially causing a model to miss the importance of small-scale exploratory behaviours.
* **Transparency and Reporting:** It is an absolute requirement of scientific rigor to clearly document and report the methods used to handle missing data, including the rationale for the chosen approach and the amount of data that was imputed.

**Part II: Assignment Questions**

**Instructions:** Based on the source material provided above, answer the following questions. Your responses should be clear, concise, and demonstrate a thorough understanding of the concepts presented.

1. **Classifying Missingness:**

An elk's GPS collar fails to record locations whenever it is in a specific, densely forested canyon. Separately, a bear's collar fails to record locations for four months each winter. Using the terminology from the guide, classify these two types of missingness and explain why this distinction is critical for how you would interpret the data.

1. **Diagnostic Workflow:**

You have just been given a new, unfamiliar GPS dataset. What are the first *two* diagnostic analyses you would perform to understand its missing data problem? For each analysis, describe what it is and what it would tell you.

1. **Methodological Trade-offs:**

Consider a single, two-hour gap in a dataset that is otherwise recorded hourly. Compare and contrast the use of **Linear Interpolation** versus a **Correlated Random Walk** to fill this gap. What is the primary advantage of the random walk approach, even for a short gap?

1. **Advanced Concepts:**

According to the guide, what is the single most important advantage of using Multiple Imputation (MI) over any single imputation method (like linear interpolation or a single random walk)?

1. **Implications for AI Modelling:**

The guide's decision framework suggests that for long gaps caused by denning behavior, one should "Do not impute." Explain how naively "filling in" a four-month hibernation gap with an interpolated path could negatively bias the final output of your AI-driven wildlife corridor model.