Big Data Analysis - Assignment 1

Davide Griffon

Vilnius University

March 26, 2025

Project Overview

Anomaly Types:

- **Vessel movement anomalies**
  - Speed inconsistencies
  - AIS transmission time gaps
- **Location anomalies**
  - Position conflicts between vessels

Technical Approach:

- Spatiotemporal binning for efficient processing
- Parallel computing with multiprocessing
- Vectorized operations (pandas/numpy)
- Unit testing with pytest
- Structured development with Taskfile

Vessel Movement Anomaly Detection

- **Two primary indicators:**
  - **Speed anomalies**: Vessels reporting unrealistic speeds
  - **AIS gaps**: Unusually long periods between transmissions
- **Implementation approach:**
  - Group data by vessel MMSI
  - Calculate distances between consecutive points using geopy
  - Vectorized operations for time differences and speed calculations
  - Flag vessels exceeding a threshold (default 50.0 miles/h)
  - Flag transmission gaps exceeding a threshold (default 1.0 hour)
- **Optimization:**
  - Process each vessel independently (perfect for parallelization)
  - Use of NumPy for vectorized distance calculations

Location Anomaly Detection - Binning

Spatiotemporal Binning:

- Challenge: Efficiently finding vessels at same location/time
- Solution: Group data into spatial and temporal bins
  - latitude "edge" = 0.01° (~1km)
  - longitude "edge" = 0.01° (varies)
  - temporal bin = "1min"
- Boundary handling:
  - Points near bin boundaries are placed in multiple bins
  - Prevents missing conflicts at bin edges

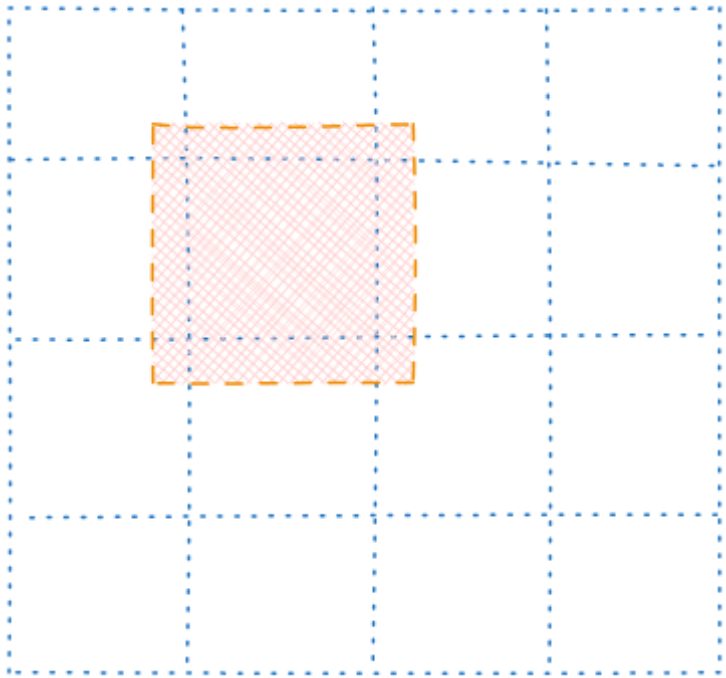

Illustration of spatial binning with overlap

Location Anomaly Detection - Finding Conflicts

Conflict Detection Process:

1. Preprocess data
   - Apply spatial binning
   - Apply temporal binning
   - Group by lat/lon/time bins
2. Process each chunk/bin in parallel
   - Create position-time hash for efficient lookup
   - Find exact position and time matches
   - Identify vessels with different MMSIs at identical positions
3. Combine results and remove duplicates
   - Keep only one conflict per vessel pair (smallest distance)
   - Sort by distance for easier analysis

Performance Results

Dataset Characteristics:

- 5,901 unique vessels (MMSI) analyzed
- 36,583 spatiotemporal bins processed
- Full day of AIS data (aisdk-2024-06-30.csv)

Performance Comparison:

- **Single-process:** 780 seconds
- **Multi-process:** 230 seconds
- **Speedup:** 3.4x

Conclusion: Parallel processing provides significant performance improvement for spatiotemporal anomaly detection in vessel tracking data.