

Peer Review

Review of: "Data Integrity vs. Inference Accuracy in Large AIS Datasets"

Hsin-Hsiung Huang¹

1. Statistics and Data Science, University of Central Florida, United States

Review Report

Title of the Manuscript: *Data Integrity vs. Inference Accuracy in Large AIS Datasets*

Authors: Adam Kiersztyn, Dariusz Czerwinski, Aneta Oniszcuk-Jastrzabek, Ernest Czermanski, Agnieszka Rzepka

Journal: Preprint on Qeios

Manuscript ID: 8PUBHN

Summary

This manuscript explores the impact of data integrity on inference accuracy in large Automatic Identification System (AIS) datasets. The authors discuss error detection and correction techniques to improve AIS data reliability, emphasizing the implications for maritime traffic safety, operational efficiency, and environmental protection. The paper presents a structured analysis of AIS data quality, anomaly detection, and data verification methods.

The study is timely and addresses a crucial issue in maritime data analysis. The discussion on AIS data integrity and its implications is well-organized, and the findings contribute to the field of maritime analytics. However, some methodological aspects require clarification, and a stronger comparative analysis with recent advancements in anomaly detection and statistical modeling would enhance the study's impact.

Major Comments

1. Methodological Rigor and Reproducibility

- The manuscript outlines key methodologies for data integrity assessment but lacks a deeper discussion of the theoretical properties of the proposed detection and correction methods.

Expanding on the statistical foundations behind these methods would improve the rigor of the study.

- To enhance transparency, consider including a reproducibility section or publicly available code to facilitate validation of the results. Providing access to AIS data preprocessing scripts or algorithms used for anomaly detection would be beneficial.

2. Comparisons with Recent Work

- The paper references several existing methods for AIS data integrity assessment but does not sufficiently compare them with other contemporary approaches in maritime data analysis. A more detailed comparison with state-of-the-art Bayesian and deep learning models for anomaly detection in AIS data is recommended.
- The authors should consider referencing and discussing insights from the study "*Unsupervised Vessel Trajectory Reconstruction*" (Frontiers in Applied Mathematics and Statistics, <https://doi.org/10.3389/fams.2023.1124091>). This paper discusses statistical modeling (unsupervised trajectory methods) for vessel trajectories and anomaly detection and would provide a broader context to the discussion.

3. Data Selection and Processing

- The focus on tanker vessel AIS data is reasonable, but the rationale for selecting this category should be explicitly stated. Are tankers particularly prone to AIS anomalies, or was this choice based on data availability?
- More details on preprocessing steps, particularly on handling missing data and outlier correction, would be valuable. If multiple preprocessing methods were tested, discussing their comparative effectiveness would strengthen the study's robustness.

4. Impact on Inference Accuracy

- The manuscript claims that improving data integrity enhances inference accuracy, but additional quantitative evaluation metrics (e.g., precision, recall, RMSE) should be included to quantify the impact of the data cleaning techniques.
- A statistical comparison of inference accuracy before and after data integrity adjustments would reinforce the conclusions. Consider including statistical tests or confidence intervals to demonstrate the significance of improvements.

Minor Comments

1. Clarity and Readability

- Some technical terms and notation should be clearly defined to improve readability, ensuring consistency in terminology throughout the manuscript.
- Figures and tables should include more explicit captions explaining their relevance to the analysis to help readers interpret them more effectively.

2. Citations and Literature Review

- The discussion on AIS data quality is thorough but could benefit from additional references to statistical modeling techniques and anomaly detection in large datasets.
- As suggested, citing the *Frontiers in Applied Mathematics and Statistics* article on vessel trajectory reconstruction would be beneficial in positioning the study within the broader research landscape.

Recommendation: Minor Revision

The paper presents an important discussion on AIS data integrity and inference accuracy. However, it would benefit from additional methodological details, a stronger comparative analysis with recent research, and improved quantitative validation of its results. Specifically, the authors should:

1. Provide a more rigorous explanation of the statistical models used for data integrity assessment.
2. Compare their approach with recent state-of-the-art methods, particularly Bayesian and deep learning models.
3. Justify the choice of tanker vessel data and provide more details on preprocessing and anomaly detection methods.
4. Include quantitative metrics to demonstrate the impact of data integrity improvements on inference accuracy.

I appreciate the authors' contributions to this important topic and look forward to seeing the revised manuscript incorporating these suggestions.

Attachments: available at <https://doi.org/10.32388/5F7LVC>

Declarations

Potential competing interests: No potential competing interests to declare.