17 March 2025 · CC-BY 4.0

Qeios

Peer Review

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

Christin Rhen[1]

[1]. Underwater Systems, Saab (Sweden), Bromma, Sweden

The paper “Data Integrity vs. Inference Accuracy in Large AIS Datasets” is a clear illustration of some data integrity issues in AIS data. The paper is well-written and easy to follow, but could be strengthened by some deeper analysis of the results.

The introduction is clear, and the literature review is thorough, providing a good overview of the field. The presentation of data in Section 2 is clear, but could perhaps be strengthened. Tables with only numbers, like Table 1, can be hard to understand for the reader. It could be interesting to complement this with a graph that illustrates the proportion of the different statuses for each year, like a stacked bar chart or an area chart. Instead of the bar plot in Figure 4, I think a box plot would be more informative, showing also the statistical properties of the data. In connection to Figure 3, would it be possible to identify if there are any “usual suspects” among the ships observed several times in the data? I think it would be interesting to look at whether a large fraction of flawed data reports comes from a small number of repeat offenders, or if it is a broader problem. Finally, in Figure 5, not only is the highest reported speed a completely unreasonable 102 knots - this ship appears to be located in an inland lake! Is this a mistake in the graph, or could you discuss this fact? How would a tanker come to be located in such a place?

The conclusion is very brief. Could you discuss your results a bit more? What could be the implications of using incorrect AIS data in inference? You claim that the “analysis shows that error detection and correction techniques and data verification methods can significantly improve the quality of AIS data” - exactly what techniques and methods have you used, and are they included in the references? From how I read Section 2, you essentially look closely at data and point out some discrepancies/inaccuracies that must lead to the conclusion that the data is reported erroneously. This is in itself a valuable result, but

doi.org/10.32388/2MGORQ

does not support the community much in how to handle the problem. Could you suggest some ways to mitigate flawed data, once identified? Have you tested some methods in the references?

Finally, how would you like to continue this study? Are there any particular avenues of future work you see?

Declarations

Potential competing interests: No potential competing interests to declare.