arXiv:2602.11890v1 [cs.DB] 12 Feb 2026

用于船舶移动性分析的数据驱动的轨迹插补

Giannis Spiliopoulos
爱琴海大学与 Archimedes/Athena RC 希腊

Alexandros Troupiotis-Kapeliaris
爱琴海大学 & Archimedes/Athena RC 希腊

Kostas Patroumpas
Archimedes/Athena RC 希腊

Nikolaos Liapis 希腊海洋技术研究所 希腊

Dimitrios Skoutas
Athena RC 希腊

Dimitris Zissis 爱琴海大学 希腊

Nikos Bikakis 希腊地中海大学 & Archimedes/Athena RC 希腊

摘要

在海上对船舶活动进行建模，对于航线规划、运输物流、海事安全和环境监测等众多应用至关重要。在过去二十年中，自动识别系统（AIS）实现了对数十万艘船舶的实时监测，每天产生海量数据。使用AIS数据面临的一项主要挑战是船舶轨迹中存在大量缺口，这些缺口通常由覆盖限制或有意的传输中断造成。这些缺口会显著降低数据质量，导致分析结果不准确或不完整。最先进的插补方法主要用于处理车辆轨迹中的缺口，即使底层道路网络未被纳入考虑。然而，航行船舶的运动模式存在显著差异，例如平稳转弯、在港口附近操纵，或在恶劣天气条件下航行。在这篇应用论文中，我们提出了HABIT，即一种轻量、可配置的基于H3聚合的船舶轨迹插补框架。该数据驱动框架通过从历史AIS数据中提取、分析和索引运动模式，为插补缺失轨迹片段提供了有效手段。我们针对不同时间范围、密度和船舶类型的AIS数据开展了实证研究，结果表明，在准确度方面，HABIT生成的海事轨迹插补结果与基线方法相当；同时，在考虑船舶特征及其运动模式的情况下，其延迟表现更优。

关键词

轨迹插补、轨迹重建、移动性分析、海事监测、海上交通分析、AIS数据

1 引言

全球超过80%的贸易量通过海运完成，预计到 2030[1],年平均每年增长2%，这凸显了海事运营对世界经济、环境保护和地缘政治稳定的影响。目前，全球船舶上安装的自动识别系统（AIS）已达40万艘，为监测海事活动提供了宝贵手段。捕获船舶发送的AIS信号，不仅能够实现实时跟踪和安全航行；大量积累的数据还为各种作业带来了重要机遇，包括绘制详细的交通密度图、高效的港口管理以及经济高效的航线规划。

AIS数据的局限性。然而，AIS数据存在一些固有缺陷，不仅因为覆盖不完整、信号丢失以及AIS消息报告频率不规则，有时还因为跟踪设备被故意关闭或故意传输虚假信息（欺骗），从而引发安全隐患（例如碰撞）或引起怀疑（例如走私）。事实上，AIS位置消息可能记录得较为稀疏，导致出现缺口，在相当长的时间内（甚至数小时或数天）无法获知船舶的位置。因此，可能无法将船舶的完整轨迹重建为一系列带时间戳的位置，因为其确切航线并非始终可知。采样位置越稀疏，轨迹数据的准确度和质量就越低。尤其是，可能出现多个此类缺口，极大地降低这类数据用于可靠分析的价值。某些应用（例如交通监测、海事监视）需要高质量的轨迹，即更“密集”、没有缺口的轨迹，以便开展有效分析。

海事领域中的挑战。为弥补这一不足，轨迹插补 技术 [3, 6, 8, 12] 可用于沿这些缺口插入人工点，从而恢复轨迹片段的连续性。其目标是获得与船舶实际行驶路径相似的插补轨迹，并通过消除采样率差异来提升轨迹质量。现有大多数轨迹插补方法（又称 轨迹插值、轨迹补全、轨迹恢复、轨迹恢复 [13]）依赖于将轨迹与底层网络（通常是道路网络）进行匹配，因而默认此类网络可用且可靠。然而，由于船舶运动具有若干特点，这些方法并不特别适用于海上轨迹，例如：（1）尽管某些类型的船舶（如客船、油轮）遵循严格的航行规则和重复出现的模式，但不能假设海上存在固定网络；（2）变化不定且有时较低的 AIS位置消息采样率，使得专为频繁报告的 GPS轨迹（通常来自行驶中的车辆）设计的此类方法更难适用；以及

[1]https://unctad.org/publication/review-maritime-transport-2024

EDBT '26，坦佩雷（芬兰）© 2026 版权所有者/作者所有。本文依据ISBN 978-3-98318-104-9、系列 ISSN 2367-2005，发表于 OpenProceedings.org。本文可根据知识共享许可协议 CC-by-nc-nd 4.0 的条款进行分发。

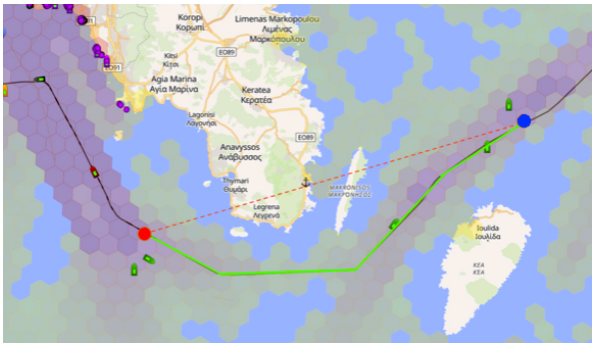


图1：插补船舶轨迹中的一个缺口。

（3）最重要的是，车辆的运动方式与海上船舶的运动模式有很大不同，后者的特点包括平滑转弯（大型船舶尤为如此）、港口或锚地附近的复杂操纵、恶劣天气条件下的临时航行等。

应用。插补轨迹可作为宝贵资源，服务于广泛应用场景中的关键业务，包括：

海事决策。 插补轨迹可在多种场景下支持更有效的海事监测，例如帮助主管部门优先处理拥堵区域或环境敏感区域的事务、根据天气状况预测生成备选航线 [17] 或协调搜索与救援行动中的应急响应 [2]。

船舶交通可视化分析。通过消除AIS数据中的缺口和噪声，可以生成、更新并维护更准确的密度图 。这类地图（图1）能够有效揭示船舶交通行为及其演变的丰富信息，并以直观的可视化形式呈现有价值的信息，帮助识别模式、异常及其他相关现象 [18]。

异常或可疑海事行为。分析完整轨迹中的模式，可能会揭示异常或可疑行为，例如欺诈或犯罪活动，如包裹接驳或船舶会合 [15]。

我们的方法。 我们介绍了 HABIT[2]，，这是一种专为船舶轨迹设计的 H3 聚合式插补框架。这是一种高效、数据驱动且可扩展的方法，依托在 H3[3] 六边形单元上计算的 AIS位置报告空间聚合数据，对海上轨迹中的缺口进行插补。当前最先进的插补方法主要用于处理车辆轨迹中的缺口，即使没有考虑底层道路网络也不例外。基于近期提出的生命活动模式框架 [16]，我们认为，从大量历史AIS数据中提取船舶运动模式，有助于更好地插补缺失轨迹片段。我们提出的数据驱动、基于网格的方法能够识别潜在的船舶移动关系和运动模式，还能保留每个单元的统计数据，例如船舶数目、单元间转移的次数和距离等。

[2]https://github.com/M3-Archimedes/HABIT [3] https://h3geo.org

这些特征建立索引后，便可通过查询来推荐可能的路径，以插补观测到的缺口。

HABIT 专注于填补由 AIS覆盖问题而非蓄意 AIS停用造成的缺口。覆盖不足导致的缺口会影响特定时间和区域内的大多数船舶，通常持续时间较短，且船舶会继续其日常活动。相比之下，蓄意停用会使跟踪系统无法识别船舶、获取其位置和活动信息。这类情况偏离典型行为，因此与某个区域的历史数据关联较小。我们假设，按固定时间间隔从近期历史AIS数据中计算出的区域统计数据，能够代表该区域的典型海上交通。因此，可以利用这些统计数据重建历史轨迹中缺失的任何片段。

插补路径必须是可航行的，这一点至关重要；也就是说，它们不能穿越海岸线、岛屿、保护区或其他障碍物，也不能包含频繁的急转弯。因此，船舶在实际航行中可以沿着建议的航线行驶，如图 1 中蓝点与红点之间以绿色线条标示的间隔所示；相比之下，通过线性插值得到的直线路径显然不可航行。还可以考虑船舶类型，因为大型船舶或吃水较深的船舶（例如客船或货船）可能无法通过狭窄海峡或浅水区。我们对时间跨度、空间覆盖范围和船舶类型各异的真实世界AIS数据开展了实证研究，结果证实，插补海上轨迹时必须考虑船舶及其运动模式的独特特征，从而显著提升准确度、性能和可扩展性。HABIT 具有实用优势，因为它基于轻量级库和框架，在构建框架和进行推断时所需的计算能力都显著更低。此外，通过纳入数据驱动的路径修正，可以实现更真实、更有依据的轨迹重建。

概述。 本文其余部分安排如下。第2节介绍船舶监测的必要背景，并综述与轨迹插补相关的工作。第3节介绍HABIT，这是我们用于从历史AIS数据中捕获船舶交通的框架。第4节报告评估研究的结果。最后，第5节总结全文，并提出未来工作的方向。

2 背景与相关工作

船舶监测。 通过自动识别系统（AIS）跟踪船舶自行报告的位置，可以对海上人类活动开展深入研究。船舶转发的AIS消息包括其唯一的海上移动业务识别码（*MMSI*）、当前坐标（*LON*经度、 *LAT*纬度）、对地速度（*SOG*）、对地航向（*COG*）等。需要特别指出的是，时间戳是在接收到消息时分配的。每天收集的大量数据需要借助适当的工具进行处理和建模，以便对船舶交通进行简洁描述 [16]。此外，还应制定专门的机制来确保所得数据的质量，因为AIS存在一些固有缺陷（主要源于覆盖范围和信号丢失问题）。

AIS集中式系统通常依靠地面接收器和卫星接收器的组合来获取数据 [22]。地面接收器的覆盖范围受无线电频率传输规范的限制，通常可扩展至

20海里，并受地面地形和天气条件的限制。另一方面，卫星接收器虽然具有大得多的接收覆盖范围，但也会受到覆盖问题的影响，主要原因是卫星的存在，即轨道中的物体，以及拥挤区域中的数据包碰撞。鉴于AIS是一种自报告系统，任何类型的接收器都可能因船载AIS遭到蓄意误用或恶意使用而接收到缺失或虚假数据。近期一项将AIS与地球观测检测结果进行比较的分析 [18] 表明，即使是在行业级的集中式AIS数据库中，某些区域也可能缺失大量AIS消息，说明经过这些区域的船舶轨迹中存在较大的缺口。

轨迹插补。 报告位置中出现的缺口并非船舶轨迹所独有，而是普遍存在于各类移动对象（车辆、商品、动物等）的轨迹中。因此，最先进的插补方法会尝试在真实点之间插入人工点，以适当地弥合这些缺口。基于网络的方法，如 [4]，利用底层网络（通常是道路）的结构，推断缺失片段可能经过的路径。

相比之下，无网络 插补技术基于位置数据。接下来，我们首先概述这类数据驱动方法，尤其是针对城市环境中GPS轨迹缺口的方法。TrImpute [6] 假设对底层道路网络一无所知。对于每个待插补的轨迹片段，该方法首先从附近的现有轨迹中收集候选点，然后选取能够构成该片段端点之间最短路径的点。秉持相同思路，GTI [8] 提出了一种改进的、与网络无关但计算开销较大的方法，该方法利用相邻轨迹的已有信息来加密目标轨迹。它根据稀疏轨迹创建一个连通的有向图，以便通过遍历图节点，使用 Dijkstra算法求出每两个连续位置之间的最短路径。

机器学习模型也已被应用于轨迹插补，但会带来较高的计算复杂度。例如，TERI [3]提出了一种轨迹恢复框架，无需任何关于待恢复位置的先验信息。假设这些位置之间存在不规则时间间隔，该框架首先检测这些位置出现的时间，然后对缺失的数据点进行插补。在每个阶段，它都部署了一个基于Transformer编码器架构、并集成可学习傅里叶编码模块的模型（RETE），以更好地建模空间和时间相关性，同时结合集体转移模式学习与轨迹对比学习，有效捕捉序列转移模式。此外，Kamel [12] 提出了一种可扩展的轨迹插补框架，通过插入额外的真实轨迹点来提高准确度。该方法将轨迹插补问题映射为自然语言处理中的缺失词预测问题，从而采用广泛使用的BERT模型，并进行了一些重要改进。它为空间感知的BERT操作增添了空间信息，使轨迹数据更接近语言数据的特性，并提供多点插补能力。PaLMTO [11] 探索了基于 N元组的概率语言模型在轨迹插补中的应用，即将前面的 $N-1$ 个词元作为上下文，预测轨迹中的下一个词元（即点）。该方法采用不依赖任何底层道路网络的基于网格的空间表示，并将轨迹点转换为与其所在网格单元相对应的词元，以训练不同规模的模型。

然而，所有上述技术都是针对城市环境中的GPS轨迹设计并进行测试的。正如我们上文所强调的，它们可能无法根据AIS数据生成可靠的路径，因此研究者提出了专门用于船舶轨迹的插补方法。DAISTIN [10] 根据原始AIS数据构建图，以专门表示油轮的运动。为填补给定轨迹上两个位置之间的缺口，它会返回最短路径。然而，该方法尚未在其他船舶类型上进行测试。在 [23], 中，研究者针对油轮和货船的长期缺口（最长可达10小时）提出了一种物理引导扩散概率模型。该模型利用条件信息从白噪声中重建轨迹，并在训练扩散模型时加入运动学约束，以增强轨迹平滑性。类似地，TrajDiff [20] 采用扩散模型，通过添加噪声生成潜在AIS数据表示，从而提升对轨迹交通分布和船舶运动模式的学习能力。去噪过程将已知轨迹点和运动特征作为条件，以便在持续时间较长的缺口（例如10小时）中生成更真实的轨迹片段。最近，MH-GIN [9] 构建了一种多尺度图网络，显式建模异构AIS属性（而不仅仅是位置）之间的依赖关系，通过图传播更准确地插补缺失值。相比之下，有些研究通过引入额外的数据源来解决轨迹插补问题。例如， [21]采用数据融合方法整合沿海摄像头检测结果，从而获得更完整的轨迹。

在这项工作中，我们借鉴了 [16], 中提出的概念，即采用 H3空间索引进行信息压缩并构建转移图。与其基于 Apache Spark 的分布式处理架构相比，我们提出了一种更简单的技术栈，能够利用 DuckDB[4], 这一现代分析型数据库，在内存中高效处理数据。最重要的是，我们专门研究了这些概念在轨迹插补问题中的应用。我们引入了一种数据驱动机制，以减轻使用 H3 网格造成的信息损失，并针对不同配置和数据集，将我们的方法与原始路径及最先进的方法进行分析和比较。

3 HABIT：基于H3聚合的船舶轨迹插补

本节介绍 HABIT（基于H3聚合的船舶轨迹插补框架）[2], 。该框架基于 DuckDB 和 NetworkX 等开源工具构建。HABIT 是一种数据驱动且可扩展的海上轨迹缺口插补方法，依托于在H3六边形单元上计算的 AIS位置报告空间聚合。

总体而言，我们的框架由四个主要阶段组成：
(1) 数据预处理与行程划分（第 3.1 节）：过滤错误数据，并按航次清理、分析和整理原始位置报告。(2) 图生成（第 3.2 节）：使用空间 H3 索引计算位置嵌入。首先将网格单元之间的转移整理为边列表，然后据此构建图。(3) 轨迹插补（第 3.3 节）：利用图统计信息，通过最短路径算法确定缺失轨迹片段上的点。此外，还采用数据驱动方法，更准确地完成 H3 网格单元与坐标之间的映射。

[4]https://duckdb.org

(4) 轨迹简化（第 3.4 节）：对插补轨迹进行平滑处理，以生成真实且可导航的路径。

3.1 数据预处理与航次分段

在初始阶段，我们将每艘船舶的完整轨迹划分为多个独立航次，并进行数据清洗。为了识别航次，我们会分析 AIS 原始数据，并考虑时间、位置、速度变化和航向等多个因素。通过这一分析，我们不仅能够识别各个航次，还能检测出各种运动事件，并为其添加语义标注。

在我们的方法中，航次 指两个连续停船或间隔之间的 AIS 位置子序列：

– 停船 表示船舶处于静止状态，即在一段时间内其速度为 < 0.5 节。这类停船的起始位置（通常位于港口，但有时也在海上）标志着当前航次的结束。反之，停船期间的最后一个位置则表明船舶刚刚启程开始新的航次。
– 通信间隔 是指最近一段时间内（例如过去 $\Delta T = 30$ 分钟）未收到 AIS 位置。如果出现此类间隔，当前航次会突然结束；待与船舶恢复通信后，系统会为其分配新的航次。请注意，提取出的航次中仍可能存在短于 ΔT 的间隔。

为识别此类停船和缺口，我们采用一个轨迹压缩框架 [7]，专门用于标注AIS数据[5]。通过考察给定船舶的运动模式（速度、航向）随时间的变化，该方法能够逐步标注代表多种移动事件的选定位置（不仅包括停船和缺口，还包括转向点、慢速移动和速度变化）。这一过程还会应用过滤器，消除AIS消息中固有的大量噪声，例如重复位置、无效坐标、会扰乱序列的延迟消息等。获得这些标注后，将噪声位置丢弃，并将船舶在两个连续停船或缺口之间的所有剩余位置归为一个航次，同时为每个位置分配唯一的航次_标识符。

最后，某个根据上述标准划定的航次可能仅位于给定分辨率下的一个或至多两个相邻H3网格单元内。由于这些记录主要涉及细微且非必要的局部位移（例如海上漂移），因此在我们的分析中将其排除，不再进一步考虑。

3.2 图生成

与短时间内可视为静态的陆路交通网络不同，海上交通网络具有动态性。港口之间的航线也会因天气、交通和地缘政治因素而迅速变化。为捕捉船舶的集体行为，我们将原始AIS数据投影到 H3 网格上，并将网格单元之间的局部转移进行分组。通过这一过程，可根据 AIS数据以图的形式构建数据驱动的海上网络，进而利用现有的基于图的路径算法。

数据处理与索引。在 DuckDB 中使用通用表表达式（CTE）处理 AIS消息，步骤如下：

(1) 读取初始阶段获取的航次数据和源文件，并定义 H3 空间分辨率。(2) 根据航次标识符（TRIP_ID）将消息划分为不同的组。(3) 在指定分辨率下，将每条消息分配到对应的 H3 网格单元（cl），并为其补充航次序列中的前一个 H3 网格单元（lag_cl）。(4) 对数据进行两次分组：一次按 H3 网格单元（cl）分组，另一次按（lag_cl,cl）元组分组。在这两种情况下，都会计算每个组的聚合统计数据。

索引步骤中一个具有挑战性且至关重要的问题，是确定 H3 网格的分辨率（r）。一些研究，尤其是在城市环境中，可能采用仅几米的分辨率；[12], 而 AIS 轨迹通常跨越更大的距离。如果网格过细，就难以提取可靠的描述性统计数据，因为每个网格单元中只包含少量 AIS 点。反之，选择较大的网格单元大小会降低细节程度，使生成的路径更贴合网格布局，而非实际的船舶轨迹。

统计计算。统计数据通过*DuckDB* 内置函数计算，以获得精确指标和近似指标。

对于每个 H3 网格单元组 cl，我们计算：

– $count(*)$：网格单元cl内 AIS 记录的总数。
– $apprx_count_distinct(VESSEL_ID)$：经过网格单元$cl$的不同船舶数量。– $median(lon)$和 $median(lat)$：分别为经度中位数和纬度中位数。– $median(sog)$和 $median(sog)$：分别为中位速度和对地航向。

然后，对于每个 H3 网格单元元组组（lag_cl,cl），我们计算：

– $apprx_count_distinct(TRIP_ID)$：从网格单元$lag_cl$到$cl$的转移次数。– $h3_grid_distance(lag_cl, cl)$：以 H3 网格单元数表示的转移距离。

图构建。H3 索引与计算所得的统计数据用于构建加权图[6]：

– 图的节点 由边列表中出现的相应H3网格单元标识，并与上一步计算得到的以下属性相关联：(a) 经度中位数；(b) 纬度中位数；(c) 消息总数；以及 (d) 不同船舶数量。– 基于其 H3 索引标识符（通常为 cl和 lag_cl）连接两个节点的边，由轨迹中发现的转移定义，条件为 $lag_cl \neq cl$。相应的边权重对应于转移的聚合数量。

3.3 轨迹插补

首先，我们将间隙起止点的坐标投影到 H3 网格中对应的单元上。为了能够利用底层图拓扑，起始和结束 H3网格单元都必须对应于图节点。如果这些单元不对应于图节点，则执行最近邻搜索，以找到确实对应的最近单元。因此，一个间隔的起止点被映射到两个图节点。

随后，A* 搜索通过将转移次数降至最低来识别最短路径，从而有效揭示

[5]https://github.com/M3-Archimedes/AIS-trajectory-annotation

[6]该图使用 Python 库 https://networkx.org 构建。

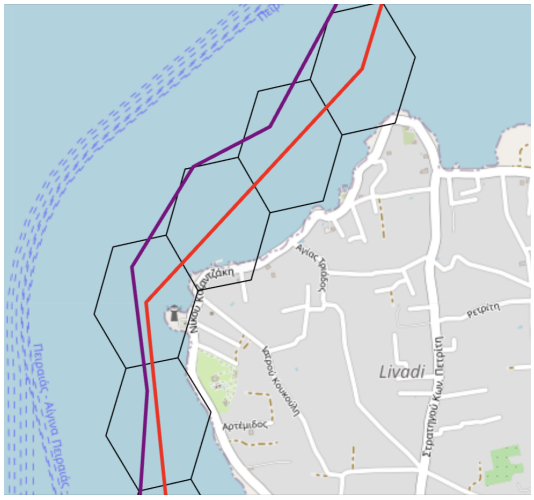


图2：使用（a）网格单元中心[red]；以及（b）数据驱动（中位数）方法 [purple]，将插补路径的网格单元序列逆投影回真实世界坐标。

根据历史数据得出的最常见路径。在此阶段，生成的插补路径由间隔对应的最常见节点（H3 网格单元）序列组成（按转移次数计算）。为提供真实世界坐标中的解决方案，我们将 H3 插补路径逆投影回经纬度元组。最简单直接的方法是使用每个网格单元的几何中心作为其坐标表示，但这样很容易导致位置落在船舶常见活动区域之外，甚至落在陆地区域。为解决这一问题，我们采用数据驱动的方法，在图构建过程中根据 AIS位置计算每个网格单元的坐标中位数。

使用坐标中位数而非平均值，可确保得出的位置有历史上有效的船舶位置作为依据，从而与真实运动模式保持一致（图2）。HABIT 可根据所选配置参数 p，支持对网格单元采用任一投影选项。我们将几何中心记为 c，将中位数记为 w。

3.4 轨迹简化

最后阶段会对插补轨迹进行简化，因为在海上航行中，生成逼真且可航行的路径至关重要。路径需要尽量避免频繁的大幅转弯，并且可以用航路点之间的一系列直线段来表示。这种对平滑性的要求有别于城市出行，源于大多数船舶，尤其是大型船舶，机动能力有限。基于网格的方法生成的路径天然遵循底层网格的结构，因此即便准确追踪了经过的网格单元序列，也常会引入多个突兀的“锯齿”。在此过程中，采用了 Ramer-Douglas-Peucker（RDP）算法 [19]，这是一种高效的轨迹简化与泛化技术。所选的容差 t 是决定 RDP 运作方式的关键参数，因为它规定了路径允许偏离给定路径的最大幅度。较大的阈值会生成更平滑的路径，但也可能导致路径偏离过多，以至于违反图结构约束或穿越陆地区域。相反，若选择较小的容差，可能无法实现令人满意的路径压缩，还可能保留相邻网格单元之间不理想的急转弯。

表1：AIS数据集的特征。

数据集	类型	大小（MB）	位置	航次	船舶
DAN	客船	786	4,384,003	1292	16
KIEL	客船	145	806,498	86	2
SAR	All	141	1,171,162	20,778	2579

4 实验分析

4.1 设置

为评估我们的框架，我们使用了来自 AIS 的真实船舶轨迹，数据覆盖欧洲两个不同的区域。评估过程分为两个步骤：（a）通过考察所提方法的参数对重建轨迹的影响来进行微调；（b）将选定的配置与基线方法进行比较。对于后一步分析，我们重点考察生成的插补结果的准确度，以及整体性能，包括查询延迟和内存需求。最后，我们还分析了 HABIT 在较长间隔时长下的准确度。

数据集。 我们使用来自两个数据源的 AIS数据进行了实验，它们涵盖了不同的配置和场景。 *丹麦海事局 AIS*[7]：我们提取了丹麦周边为期 3 个月（2024 年 1 月至 3 月）的所有客船 AIS位置。随后据此构建了两个数据集：**DAN**：一个扩展数据集，包含该地区（丹麦、瑞典、德国）10 个特定港口之间经过筛选的航次。此场景关注更广阔区域内特定船舶类型的选定航线。**KIEL**：往返于基尔（德国）与哥德堡（瑞典）港口之间的所有航次。此场景关注沿特定航线行驶的航程。

AegeaNET AIS[8]（**SAR**）：我们构建了一个为期一个月的数据集，覆盖比雷埃夫斯（希腊）港口附近更广阔的萨罗尼克湾区域，其中包含所有船舶类型的轨迹。此场景涵盖该海域内的全部交通，AIS 接收质量各不相同，且不对船舶类型或航线进行任何筛选。这些数据集的详细信息列于表 1。

为评估插补结果，我们引入了固定时长的合成缺口：60、120 和 240 分钟（默认值：60 分钟）。在每个航次中随机设置一个缺口。原始航次（无人工缺口）作为真实值。为评估我们的方法，70% 的航次用于构建后续推断所需的底层图结构。其余 30% 用于准确度和性能测试。

竞争方法。 我们将基于 H3 的方法 HABIT 与 GTI [8], 这一高精度、先进的插补方法进行比较。该方法在创建图时使用两个距离参数： rm（以米为单位的半径）和 rd（以度为单位的半径），用于筛选点之间的候选边。在实验中，我们采用了与船舶尺寸相关的取值，以反映船舶交通中的典型间距。最初，我们也将 PaLMTO [11] 纳入实验，因为它生成的模型大小与最精细的 HABIT 配置相当。然而，在推断过程中，生成的模型未能提供足够的插补结果，且经常超过时间限制并发生超时；因此，下文不报告 PaLMTO 的结果。最后，我们将简单的直线插值（SLI）作为基线，

[7]https://www.dma.dk/safety-at-sea/navigational-information/ais-data [8] https://smartmove.aegean.gr/aegeanet

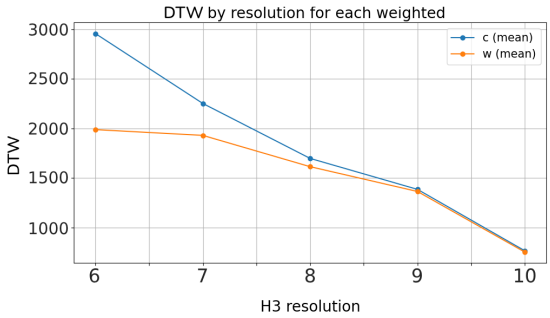


图 3： p **不同 H3 分辨率和选项（H3 网格单元中心 / 数据中位数）下 HABIT 的准确度（动态时间规整） [DAN 数据集]。**

它会直接用一条线段连接间隔的两个端点。

评估指标。 我们从准确度和性能两方面评估这些方法。对于准确度，我们使用动态时间规整 （DTW） [1]，该指标表示插补路径与原始路径之间的平均距离。为获得有意义的 DTW 测量结果，我们对插补轨迹进行插值，确保相邻位置之间的距离不超过 250 米。对于性能，我们测量插补查询的平均执行延迟，以及所构建框架的内存占用。实验在一台服务器上进行，该服务器配备第 13 代 Intel Core i9-13900K（32 核）处理器和 128GB 内存，运行 Ubuntu 22.04.5 LTS。

4.2 HABIT 的参数微调

所提出的基于 H3 的方法 HABIT 在很大程度上依赖一组参数来补全所需路径。这些参数用于设定精细的空间离散化程度和路径简化程度，并通过轻量级优化得出符合实际运行要求的解。我们在 DAN数据集上测试了多种配置，该数据集包含多个港口对之间的长距离航次（起讫点）。

在确定用于图构建（第 3.2 节）的 H3 网格分辨率（r）时，我们应考虑到商船的平均长度可达 300 米（某些情况下甚至更长），而其航次可能覆盖数百千米。因此，我们考察了从 6 到 10 的 r 值。正如预期，更高的分辨率能产生更忠实的结果（图 3），但同时也需要更多内存空间来维护这些图（表 2）。

此外，由于图补全结果（第 3.3 节）以 H3 网格单元的形式表示，我们根据投影配置参数（p）研究了逆投影过程。如图 3 所示，选择中位数中心能够得到更精细的结果，更接近船舶实际航行的真实路径。这在较粗的分辨率下尤为明显，因为网格单元内的位移要大得多。

最后，如第 3.4 节所述，采用 RDP 算法来解决插补路径的可航行性问题。我们研究了容差 t 对插补轨迹特征的影响，考察了每条路径的位置数、平均转向率和最大转向率（rot），以及超过 45^o 的转弯数量。通常，我们需要避免非常急转弯（超过 45^o），同时不能过度简化生成的路径。根据表 3，我们得出结论，容差范围为 100

表 2：框架存储大小（MB）。

方法	配置	KIEL	SAR
HABIT	$r = 6$	0.06	0.22
	$r = 7$	0.29	0.59
	$r = 8$	1.54	2.96
	$r = 9$	8.20	18.03
	$r = 10$	37.28	57.40
GTI	$rd = 10^{-4}$	50.24	115.19
	$rd = 5{\cdot}10^{-4}$	369.41	3541.89
	$rd = 10^{-3}$	1428.77	4844.12

表 3：简化对插补轨迹的影响：位置数量（*cnt*）、平均转向率（*Avgrot*）、最大转向率（*Max rot*）以及超过 45^o 的转弯数量。测量值是所有路径的平均值，采用不同的容差 t，在两个不同的 H3 分辨率 r [下针对 DAN数据集]进行测量。

r	t	cnt	平均旋转	最大旋转	$> 45°$
9	0	96.35	30.79	112.71	34.13
	100	51.76	54.92	112.31	33.78
	250	35.32	57.61	109.96	23.75
	500	14.57	44.89	84.03	6.11
	1000	6.93	34.32	56.05	1.64
10	0	198.31	30.64	119.07	62.37
	100	71.96	48.53	116.93	35.26
	250	21.03	33.85	77.01	4.43
	500	8.62	24.70	43.31	0.60
	1000	4.67	19.85	27.38	0.09
原始		595.63	6.55	110.79	33.84

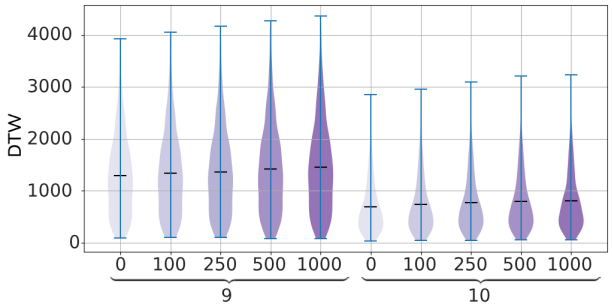


图 4：不同简化容差 $t = \{0{,}100{,}250{,}500{,}1000\}$**、分辨率**$r = \{9{,}10\}$ **[DAN数据集]上的 HABIT 准确度（动态时间规整）。**

简化至 250 米可在路径平滑度与压缩之间实现最佳权衡。此外，如图 4 所示，准确度似乎不受不同 t 值的影响。不过，H3 网格采用更精细的分辨率 r 似乎也不会影响准确度。

4.3 结果

接下来，我们报告在固定间隔为 60 分钟的情况下，对轨迹插补方法进行实证比较的结果。比较涵盖了我们提出的框架（HABIT）和所选先进方法（GTI）的多种配置，以及朴素的线性插值（SLI）。我们选择了 KIEL 和 SAR 数据集，以涵盖两种不同的应用场景。

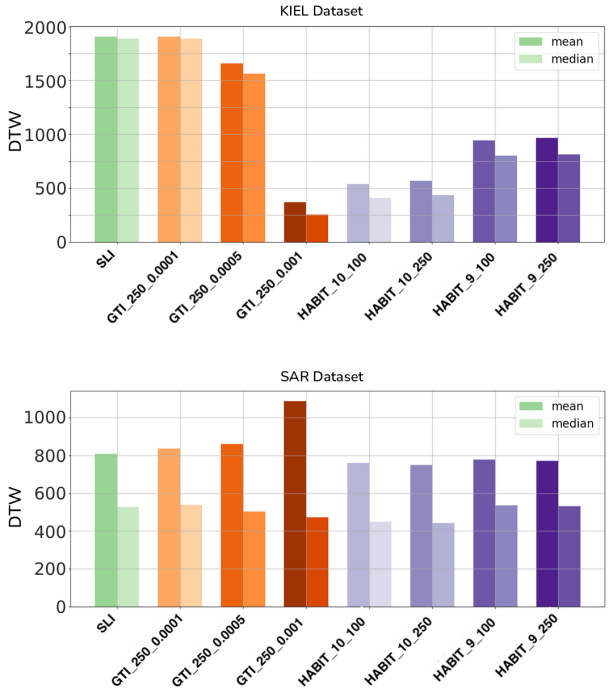


图5：GTI（rm, rd）与HABIT（r, t）在不同参数化设置下插补路径准确度（均值与中位数动态时间规整）的敏感性分析 [KIEL 与 SAR 数据集]。

由于 DAN 数据集规模过大，不适合用于生成 GTI 模型，因此未纳入分析；即使对其进行缩减（按每个航次每隔 1 或 5 分钟对位置进行重采样），结果也不稳定。

4.3.1 准确度。 如图5所示，所生成轨迹的质量对参数化非常敏感。总体而言，两种方法都能通过在图构建过程中利用原始AIS数据限制搜索空间，较好地建模局部交通模式（图 6 a–b）。

KIEL数据集中的轨迹往往遵循一条特定且受限的航线。鉴于这一特点，GTI与HABIT方法相对于原始路径的偏差都相对较低，尤其是与无法捕捉转向点的SLI基线相比。由于GTI被明确设计为跟随真实的历史航迹（在此场景中它们都属于同一航线），因此达到了更高的准确度。另一方面，HABIT会受到单元格内固有位置偏移的影响，即使H3序列预测正确，也会产生较大误差。

对于规模更大的 SAR 数据集（涵盖该区域内所有船舶类型和所有航次），两种方法呈现出不同的表现。这是因为训练数据中发现，两点之间存在多种可选航线。具体而言，GTI 的某些配置所取得的准确度与 SLI 相当，有时甚至略逊一筹。对于更精细的配置（$rm = 250m, rd = 0.001^o$），GTI 的内存需求使其无法完成计算。因此，我们采用了另一种训练配置（$rm = 500m, rd = 0.001^o$，在表 2 中以带下划线标出），但其表现不如前者，误差更大。相较之下，我们提出的框架 HABIT 在准确度方面往往保持相对稳定，

表4：HABIT（r, t）和 GTI（rm, rd）不同配置的平均与最大查询延迟（秒）。

数据集	方法配置		Avg	Max
KIEL（26 个缺口）	HABIT	$r = 9,\ t = 100$	0.024	0.041
		$r = 9,\ t = 250$	0.019	0.047
		$r = 10,\ t = 100$	0.071	0.121
		$r = 10,\ t = 250$	0.070	0.128
	GTI	$rm = 250,\ rd = 10^{-4}$	0.261	0.281
		$rm = 250,\ rd = 5{\cdot}10^{-4}$	0。300	0。431
		$rm = 250,\ rd = 10^{-3}$	0.402	0.931
SAR（1,871 个缺口）	HABIT	$r = 9,\ t = 100$	0.032	0.202
		$r = 9,\ t = 250$	0.031	0.186
		$r = 10,\ t = 100$	0.139	0.963
		$r = 10,\ t = 250$	0.139	0.866
	GTI	$rm = 250,\ rd = 10^{-4}$	0.492	0.550
		$rm = 250,\ rd = 5{\cdot}10^{-4}$	0.711	1.598
		$rm = 500,\ rd = 10^{-3}$	1.216	5.185

避免了会影响其平均得分的极端误差（图 5）。检查生成的路径后发现，这种现象很大程度上源于 GTI 选择了异常路径，这些路径并未直接朝目的地前进，尤其是在主要航道以外航行时（图 6c–d）。HABIT 通常倾向于沿着较简单的路径行进，在最小化总权重的同时，也遵循局部移动模式，而非单条采样轨迹；因此，在典型航行运动的情况下表现更佳。

为进一步评估所提出框架的适用性和效率，我们使用 HABIT 在更大间隔时长（2 小时和 4 小时）下筛选出的变体进行了额外实验。如图 7 所示，随着间隔大小增加，准确度会下降，这符合由于插值距离变长而产生的预期；然而，中位误差的增加并不与间隔长度成正比，凸显了该方法的稳健性。这种影响在 KIEL数据集中较为轻微，但在 SAR数据集中更加明显，其中略高的中位数伴随着显著的离群值。这些极端误差可能与 SAR数据集中存在的非商业船舶或客船的不规则轨迹有关。值得注意的是，不同变体的相对性能排名总体上在各种间隔大小下仍保持一致。

4.3.2 性能。 考察不同解决方案的实现要求可知，HABIT 提供了可扩展性和性能均十分出色的解决方案。如表 2 所示，基于我们提出的 HABIT 框架构建的框架，其规模会随着H3分辨率的提高而增大，但在所有区域中始终能够对船舶交通进行高度压缩的表示。相比之下，GTI 生成的模型规模要大得多（大一个数量级），尤其是与原始 AIS数据量相比时。GTI 模型的大小对稀疏性等 AIS数据特征高度敏感。范围较为集中的 KIEL数据集所需的框架占用空间相对较小，而 SAR数据集则需要更大、更复杂的结构。此外，GTI 在构建框架期间需要大量内存资源，有时甚至无法完成识别最佳拟合参数所需的计算。

关于查询响应时间，我们在不同方法配置下的插补测试中测量了平均延迟和最大延迟（表4）。针对同一海域（KIEL、SAR），所有方法配置均测试了相同的缺口（持续60分钟）。

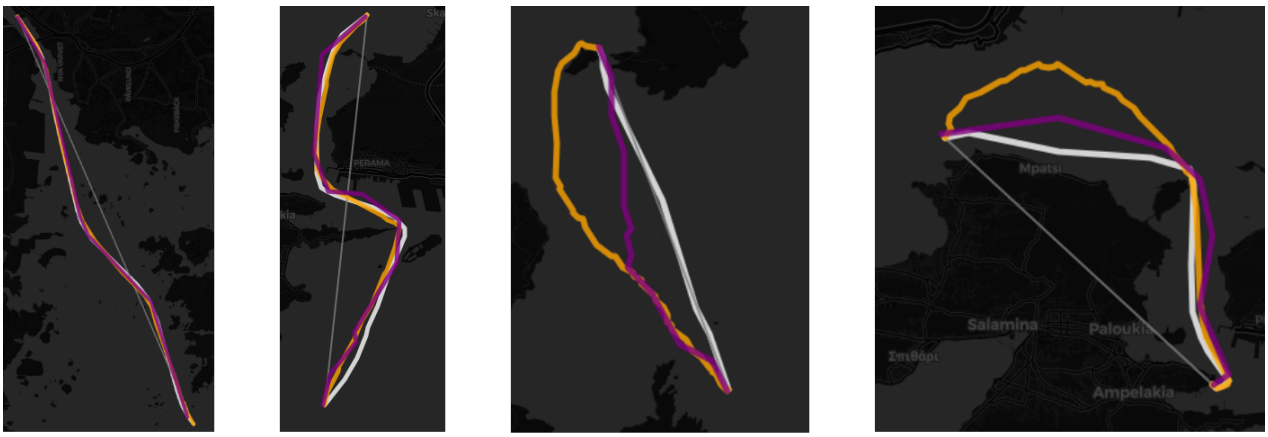


图 6：示例性插补结果：原始数据 [白色], HABIT [紫色], GTI [橙色], 和 SLI [灰色]。

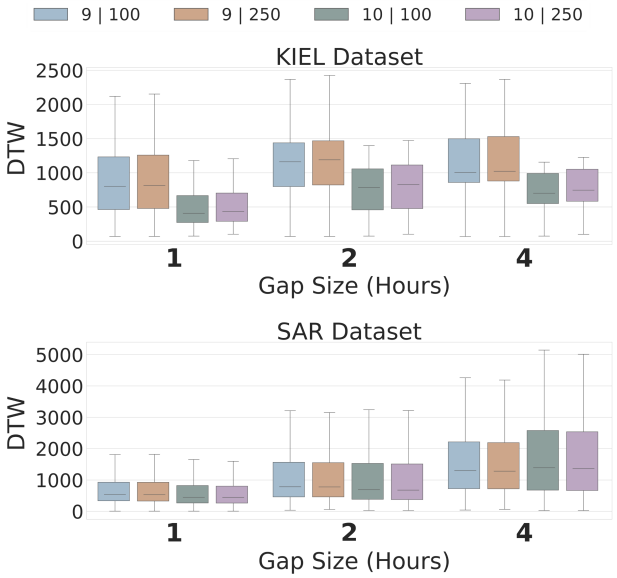


图 7：针对持续 1、2 和 4 小时的缺口，在不同分辨率和简化因子配置（$r|t$）[KIEL 与 SAR 数据集]下，HABIT 插补路径的准确度（动态时间规整）。

值得注意的是，HABIT 的延迟保持在亚秒级，且已计入其简化和重建开销。正如预期，H3分辨率越精细，延迟越高。不过，GTI 为获得更好的插补结果需要更长的响应时间，通常也比 HABIT 耗时更多；在更复杂的 SAR 数据集中，这一点尤为明显，因为该数据集包含多种船舶类型，且涉及 AIS 覆盖范围各异的区域。总体而言，我们提出的框架 HABIT 可扩展性更高，能够在不同海事场景中有效运行，同时减少计算资源消耗。

5 结论与未来工作

本应用论文提出了一种基于 H3graph 的轻量级船舶轨迹插补框架，该 H3graph 依据批量 AIS 统计数据对区域特定的运动模式进行建模。通过广泛的微调，包括轨迹简化以及在重建过程中纳入局部中位位置，该框架对真实船舶轨迹实现了较高的几何保真度。实证结果表明，其准确度与最先进的插补方法相当，在某些情况下甚至更优，同时展现出较强的可扩展性、亚秒级查询延迟以及显著更低的内存需求，即便在数据规模和稀疏性增加时亦是如此。

下一步，我们计划扩展当前方法，纳入与船舶状态（例如吃水）或航程背景（例如天气条件、航行月份、目的港）相关的特征。此外，我们将评估基于 Transformer 的模型 [5, 12, 14] 在大规模插补任务中的表现，旨在构建统一流水线，在生成区域特定分析（如密度图）之前提升数据质量。

致谢。本研究部分由希腊国家复苏与韧性计划 2.0资助的项目 MIS 5154714 提供支持。该计划由欧盟通过 NextGenerationEU 计划资助。

参考文献

[1] Donald J. Berndt 和 James Clifford。1994。使用动态时间规整发现时间序列中的模式。发表于*ACM* 国际知识发现与数据挖掘会议 *(KDD)*。[2] Konstantinos Chatzikokolakis、Dimitrios Zissis、Giannis Spiliopoulos 和 Konstantinos Tserpes。2018。挖掘船舶轨迹数据以发现搜索与救援模式。发表于大型移动数据分析研讨会 *(BMDA)*。[3] Yile Chen、Gao Cong 和 Cuauhtemoc Anda。2023。TERI：一种用于不规则时间间隔轨迹恢复的有效框架。*VLDB* 基金会 *(PVLDB)* (2023)。[4] Yuqi Chen、Hanyuan Zhang、Weiwei Sun 和 Baihua Zheng。2023。RNTra-jRec：利用时空 Transformer 增强道路网络轨迹恢复。发表于*IEEE* 国际数据工程会议 *(ICDE)*。[5] Nicolas Drapier、Aladine Chetouani 和 Aurélien Chateigner。2024。通过 H3 索引和因果语言建模 (CLM) 增强海事轨迹预测。*arXiv:2405.09596* (2024)。[6] Mohamed M. Elshrif、Keivin Isufaj 和 Mohamed F. Mokbel。2022。无网络轨迹插补。发表于*ACM* 国际地理信息系统进展会议 *(SIGSPATIAL)*。[7] Giannis Fikioris、Kostas Patroumpas、Alexander Artikis、Manolis Pitsikalis 和 Georgios Paliouras。2022。优化船舶轨迹压缩以实现海事态势感知。*Geoinformatica* 27, 3 (2022)。[8] Keivin Isufaj、Mohamed Mokhtar Elshrif、Sofiane Abbar 和 Mohamed Mokbel。2023。GTI：一种可扩展的基于图的轨迹插补方法。发表于*ACM* 国际地理信息系统进展会议 *(SIGSPATIAL)*。[9] Hengyu Liu、Tianyi Li、Yuqiang He、Kristian Torp、Yushuai Li 和 Christian S. Jensen。2025。MH-GIN：面向 AIS 数据的多尺度异构图插补网络。*VLDB* 基金会 *(PVLDB)* 19, 2 (2025)。

[10] Búgvi Benjamin Magnussen、Nikolaj Bläser 和 Hua Lu。2023。DAISTIN：一种数据驱动的 AIS 轨迹插值方法。发表于国际时空数据研讨会 *(SSTD)*。[11] Hayat Sultan Mohammed、Mario A. Nascimento 和 Denilson Barbosa。2024。使用简单概率语言模型进行有效的轨迹插补。发表于*IEEE*国际移动数据管理会议 *(MDM)*。[12] Mashaal Musleh 和 Mohamed Mokbel。2023。KAMEL 演示：一种可扩展的基于 BERT 的轨迹插补系统。发表于*ACM*国际数据管理会议 *(SIGMOD)*。[13] Mashaal Musleh 和 Mohamed F. Mokbel。2024。让我们聊聊轨迹：将自然语言处理模型用于轨迹分析任务的愿景。*ACM*空间算法与系统汇刊 *(TSAS)* (2024)。[14] Amirhossein Nadiri、Jing Li、Ali Faraji、Ghadeer Abuoda 和 Manos Papagelis。2025。TrajLearn：使用深度生成模型进行轨迹预测学习。*ACM*空间算法与系统汇刊 *(TSAS)* 11, 3 (2025)。[15] Kostas Patroumpas、Elias Alevizos、Alexander Artikis、Marios Vodas、Nikos Pelekis 和 Yannis Theodoridis。2017。基于移动船舶轨迹的在线事件识别。*Geoinformatica* 21, 2 (2017)。[16] Giannis Spiliopoulos、Marios Vodas、Georgios Grigoropoulos、Konstantina Bereta 和 Dimitris Zissis。2024。生命活动模式：海上移动模式全球清单。发表于国际数据库技术拓展会议 *(EDBT)*。[17] A. Troupiotis-Kapeliaris、G. Grigoropoulos、M. Vodas 和 K. Bereta。2025。利用大型 AIS 数据进行动态耐候船舶航线规划 [行业]。发表于 *ACM* 国际地理信息系统进展会议（*SIGSPATIAL*）。[18] Alexandros Troupiotis-Kapeliaris、Dimitris Zissis、Konstantina Bereta、Marios Vodas、Giannis Spiliopoulos 和 Giannis Karantaidis。2023。《大局观：一种改进的航运活动制图方法》。*Remote Sensing* 15，21（2023）。[19] Mahes Visvalingam 和 J. Duncan Whyatt。1990。《用于线简化的 Douglas–Peucker 算法：通过可视化重新评估》。载于*Computer Graphics Forum*，第9卷。[20] Wentao Wang、Wei Xiong、Luo Chen 和 Hao Chen。2025。《TrajDiff：一种利用重采样条件扩散模型进行船舶轨迹补全的方法》。*Results in Engineering* 28（2025）。[21] Song Wu、Kristian Torp、Alexandros Troupiotis-Kapeliaris、Dimitris Zissis、Esteban Zimányi 和 Mahmoud Attia Sakr。2025。《利用多台沿海摄像机进行有效的船舶轨迹补全》。载于*IEEE*国际移动数据管理会议（*MDM*）。[22] Dong Yang、Lingxiao Wu、Shuaian Wang、Haiying Jia 和 Kevin X. Li。2019。《大数据如何丰富海事研究：自动识别系统（AIS）数据应用的批判性综述》。*Transport Reviews*（2019）。[23] Zhiwen Zhang、Zipei Fan、Zewu Lv、Xuan Song 和 Ryosuke Shibasaki。2024。《基于物理引导扩散概率模型的长期船舶轨迹补全》。载于 *ACM* 国际知识发现与数据挖掘会议（*KDD*）。