引言

研究背景及意义

研究现状

主要研究内容和贡献

论文组织结构

相关工作与基础

主要技术&方法

本章小结

研究

本章小结

系统实现

本章小结

结果分析

本章小结

总结与展望

总结

展望

参考文献

基于渔船轨迹数据的海洋捕捞规律挖掘

摘要

随着渔业船舶监控系统（VMS）在全球范围部署的日趋完善，它能够提供的服务在不断地丰富，不仅包括基本的渔船定位、碰撞检测，还涉及到了渔业经济与渔业资源分布的量化分析预测，对渔业生产的可持续发展意义重大。

前人研究侧重于捕捞区的识别分析，揭示捕捞区的时空分布及变化规律。本文在此基础上，利用渔船在捕捞区之间的航行轨迹信息，设计了针对捕捞区域的投票分类算法，实现了对捕捞区域的合理划分。一方面，这样处理能够对渔业资源的时空变化规律进行更细致的量化分析；另一方面，路径信息可以用于分析渔民的捕捞行为，进而可以作为渔业管理的参考依据。

参照单船拖网渔船作业的特点，本文完成了以下工作。首先，利用捕捞行为识别系统（Fishing Activity Recognition System，FAR）完成港口识别、航次划分和捕捞区域识别；其次，利用捕捞区域之间的航行轨迹，设计了投票分类算法，对捕捞区域进行划分，并根据分类结果，对渔业资源变化和渔民捕捞行为进行分析；最后，设计实现了VMS数据分析可视化系统，将研究成果进行整合，并对系统性能进行评估。

关键字:VMS；分类；渔业资源；行为分析

Mining the regular pattern of marine fishing based on the data of fishing vessel trajectory

Abstract

As the worldwide deployment of the Vessel Monitoring System (VMS) becoming more and more perfect, the services it can provide are constantly enriched, including not only the basic functions such as fishing location and collision detection, also relates to the quantitative analysis and prediction of the fishery resources distribution and the fishery production, which is significant to sustainable development of fishery economy.

The previous research focused on identifying and analyzing the fishing area, aiming to reveal the spatial-temporal pattern of the fishing area. Based on this, we use the trajectory data between fishing areas to design a voting classification algorithm for fishing area. It achieves a reasonable division of fishing area. On the one hand, it can make more detailed quantitative analysis of the temporal-spatial pattern of fishery resources. On the other hand, the trajectory information can be used to analyze the fishing behaviors, which can be used as a reference for fishery management.

According to the operation characteristics of single trawler, the following works have been accomplished. First of all, by using the Fishing Activity Recognition System(FAR), achieving the port identification, track segmentation and fishing areas identification. Secondly, a voting classification algorithm is proposed exploiting the trajectory between fishing areas, to divide the fishing areas into fine-grained degree. And then, the fishery resources distribution and the fishing behaviors are analyzed after the algorithm. Finally, an application system is achieved to integrate the above algorithms and result. There is also an evaluation about it.

Keywords: VMS; Classification; Fishery Resources; Behavior Analysis

[引言 6](#_Toc504590828)

[研究背景与意义 6](#_Toc504590829)

[研究现状 7](#_Toc504590830)

[主要研究内容和贡献 8](#_Toc504590831)

[渔船行为识别系统（FAR） 10](#_Toc504590832)

[FAR系统介绍 10](#_Toc504590833)

[VMS原始数据预处理 10](#_Toc504590834)

[港口定位与航次划分 11](#_Toc504590835)

[捕捞区识别与统计 12](#_Toc504590836)

[FAR评估 12](#_Toc504590837)

[本章小结 13](#_Toc504590838)

[投票分类算法 14](#_Toc504590839)

[选票的产生 14](#_Toc504590840)

[投票分类算法 16](#_Toc504590841)

[海上路径 19](#_Toc504590842)

[本章小结 20](#_Toc504590843)

# 引言

目前的船舶行为分析普遍基于船舶监控系统，研究者们利用终端采集的信息进行数据分析和挖掘工作。本章首先简单介绍了船舶轨迹数据及其研究背景；其次，概述了基于VMS数据的国内外研究现状,以及轨迹数据研究趋势；之后，总结了本文的研究内容、贡献和难点；最后，对全文的组织架构进行了简要说明。

## 研究背景与意义

我国位于亚洲东部，太平洋西岸，大陆海岸线1.8万公里，海洋国土面积约300万平方公里，南北纬横跨近40度。这样的海域环境蕴含了丰富的渔业资源，不仅丰富了食物的来源，更增加了劳动收入。但是近年来，我国近海渔业资源日趋枯竭，并且渔民收入下降难以维持收支平衡，这对矛盾越来越受到重视。主要原因是：一方面，在利益驱使下“竭泽而渔”，严重破坏了海洋生态平衡；另一方面，渔民缺乏科学捕捞知识，往往依靠口口相传的经验。为了规范渔业捕捞，实现海洋资源的可持续性利用，中国农业部出台了一系列举措。单以伏季休渔为例，在2013年延长各海域不同类型渔船禁渔期为2至3个月之后，2017年初，再次调整禁渔期，且普遍延长一个月。此外，

2017年初规定，在“十三五”期间控制渔船船数和功率数，淘汰老旧木制渔船和过度捕捞船型，如船底拖网、三角虎网等，并为每个省份制订了指标和进度计划。

在这样的形势下，船舶监控系统(Vessel Monitoring System，VMS)越来越受到重视。它起源于葡萄牙，最初是为了保障远洋航运安全，如今已经发展为集全球卫星定位系统、地理信息系统、电子海图、计算机网络通讯和数据库技术于一体的综合应用系统。主要由船载终端与地面基站两部分构成，通过卫星进行通信。船载终端负责对船舶进行GPS定位，并利用传感器采集船舶运行状态的一系列数据。地面基站收集、存储数据，并向船舶反馈实时信息，如警示与其他船只存在碰撞危险等。

我们经常称VMS数据为渔业船舶轨迹数据。它的每条记录除了时间、经纬度坐标信息之外，还包含瞬时速度、船艏向等多个字段。相比于国外VMS系统数据采集间隔长达2小时，我国基于北斗卫星导航的VMS系统能够达到平均3分钟一条记录的采样频率，且传输可靠性高。这为科学研究提供了便利，不仅能够统计不同海域的捕捞强度，更能够揭示渔业资源的时空变化，分析渔民的捕捞策略。

目前，通过定位系统对人或车辆进行跟踪，来分析人的轨迹、行为，进一步对地理区块之间的关系进行分析，一直是受到研究者们关注的热点问题。它可以揭示个体行为的潜在规律，以及地理区块关系的时空变化。但是，海洋上没有统一的道路约束，船舶航行相对，如何分析区域之间的时空变化有待进一步研究。

综上所述，借助于VMS数据对渔业资源时空变化进行定量化分析，对于我国渔业管理有着重要的意义。另一方面，基于渔业轨迹数据的分析研究已经具备一定基础，且可以借鉴人和车辆的轨迹分析研究，进行知识迁移。所以，渔业船舶轨迹数据具备进一步挖掘的潜力，本文将以此为基础，对海洋捕捞规律进行研究。

## 研究现状

本小节从VMS数据研究和轨迹研究两个方面简要介绍相关领域的研究成果。

在VMS数据上的研究集中于如何可视化分析，可以分为以下三种类型：

直接可视化。它将原始轨迹数据按时间序列进行绘制。这种方法借助于计算机将数据映射到经纬度图像上，特点是计算量小，最大程度的保留了原始数据的原貌，而分析大部分依靠人的观察，一般只能对规律进行定性的描述。因此，工作效率低，不够系统，得出的结果也难以评价。

聚集可视化。从轨迹数据生成聚集数据，并对其可视化分析。但这种方法一方面容易丢失信息，另一方面需要对数据有很深的理解，有目的性的进行计算，否则就是盲目的尝试，即使生成的聚集数据具有规律性，还是无法解释它的物理意义。适合于数据自身规律行差的情况。

特征可视化。先计算出轨迹的特征，然后通过直接或者聚集的方法绘制这些特征，但这种方法除一般需要大量的编程才能实现外，特征计算丢失了与特征无关的大量信息，换言之就是从某个维度或几个维度观察数据。同样需要对数据有较深的理解。

在轨迹数据挖掘方面，根据Yu Zheng等的总结，主要有以下几方面研究：

处理不确定轨迹（Trajectory Uncertainty）。在被监测物体的持续移动过程中，由于设备、信号情况等的不同，往往在相邻两个位置点之间的时间段内，存在物体位置、停留时间等不确定的一段时间，在这段时间内物体的移动情况一般通过插值等方法来进行确认，从而使整个轨迹变得相对完整。

轨迹模式挖掘(Trajectory Pattern Mining)。通过分析不同物体移动轨迹的相似性，找出移动模式之间的关联，从而为后面聚类等挖掘分析工作打下基础。

轨迹分类(Trajectory Classification)。利用监督学习，根据移动轨迹或轨迹片段所代表的行为或交通方式，将数据进行分类。

轨迹轮廓探测(Trajectory Outlier Detection)。被探测物体即便种类相同（物种、交通方式、移动方式等），物体与物体之间的移动轨迹的轮廓可能存在很大不确定性，而这种不同往往可以揭示出特定事件（如车祸）等重要信息。

## 主要研究内容和贡献

本课题以浙江省温州市单拖渔船为例，深入挖掘VMS数据信息，分析渔业资源的时空演变规律。主要研究内容如下：

识别捕捞区。设计了渔船捕捞区识别系统（FAR），具体来说是通过原始数据预处理、港口识别、航次划分等一系列前期处理，利用数学形态学进行捕捞区识别，实现快速准确的捕捞区判断。

对捕捞区域进行分类。通过一个投票分类算法，利用渔民自身捕捞经验，设定分类标准，对捕捞区进行细粒度合理划分。

海洋捕捞规律挖掘。一方面，基于捕捞区的分类结果，分析区块间的联系。（抽象出捕捞区之间的通路，统计中心区块、边缘区块和主干道、次干道的变化规律。）另一方面，分析渔民捕捞行为，并提取特征进行分类。

将研究内容整合为渔业轨迹数据分析系统，并从运行效率和准确率两个角度进行权衡和优化。主要是针对影响运行效率的参数进行调整，对损失的准确率通过算法优化、替换等方式进行补偿。此外，对于系统实际应用中可能遇到的问题采取了针对性的措施。

主要贡献

之前的研究多关注于对捕捞行为的分析，忽视了航行行为。本文从航行轨迹数据入手，分析捕捞区域之间的联系。

本文利用渔民自身经验设计了一个投票分类算法，实现了对捕捞区的细粒度划分，并在此基础上分析捕捞区的变化情况。

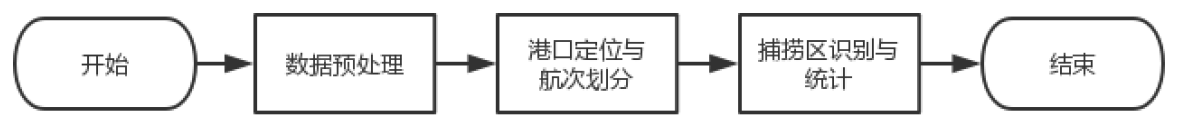
通过渔民的行为对渔业资源的时空变化进行定量评估，而不是单纯通过观察捕捞区的变化进行分析。

# 渔船行为识别系统（FAR）

对于捕捞区域的准确识别有助于分析渔业资源的分布情况，也有助于对渔业经济进行预测。然而，我国渔民缺乏记录航海日志的习惯，难以对VMS轨迹数据进行状态标记。为了完成识别任务，我们设计了渔船行为识别系统（FAR）。本章首先简要介绍了FAR的基本流程，之后对每个环节进行详细介绍，最后对捕捞区的识别结果进行评估。

## FAR系统介绍

针对单船拖网渔船的作业特点，设计了渔船行为识别系统。该系统包含三个主要步骤：数据预处理，港口定位与航次划分，捕捞区识别与统计，如图所示：



数据预处理：通过对原始数据进行格式转换、数据清洗等一系列操作，实现对异常值的剔除，同时对特殊格式的数据进行转化。

港口定位与航次划分：在缺少航海日志、港口信息的前提下，仅根据VMS数据自身的特征进行分析，确定港口坐标，并依此对数据进行航次划分。

捕捞区识别：以航次数据为单位，利用数学形态学实现捕捞区的识别，并按时间段统计捕捞区的整体分布和变化。

## VMS原始数据预处理

国际普遍采用基于全球定位系统（GPS）的VMS系统，平均采样频率为两小时一条记录。国内的VMS系统基于北斗卫星导航系统（BDS），平均采样间隔仅有3分钟。主要字段包括渔船ID、时间、经纬度、瞬时速度、船艏向等。

目前，VMS的主流研究集中于分析单船拖网渔船，一方面是因为这种类型的渔船数量多，且已经完成了大规模的VMS终端部署；另一方面，相比于其他类型的渔船，单船拖网船的捕捞轨迹更加的稳定且规则。本文沿用这一思路，针对该船型进行分析，图为单船拖网渔船示意图。

本文处理的数据来自于浙江省温州市海洋渔业安全救助信息中心，样例数据包含了自2014年4月1日起至2016年6月30日止共31条渔船2140288条记录。单船最多144413条记录，最少15799条记录。

通过将GPS时间戳和GPS经纬度转换为我们常用的时间及经纬度格式，我们利用公式计算渔船在采样点AB间的平均速度。由于我国东海海域渔船航速小于30kn，因此设阈值，剔除时间、经纬度异常值。

其中，D表示采样点AB间的球面距离，R表示地球半径，A点GPS坐标，B点GPS坐标。

国外学者在处理基于GPS的VMS数据时常进行插值操作对数据进行补充，由于国内VMS系统采样频率高，较真实地还原了渔船的实际轨迹，考虑到插值操作还会引入误差，因此本文不再进行插值。

## 港口定位与航次划分

完成航次划分的最有力依据就是航海日志，它能够从渔民角度判断船舶的运行状态。相比之下，记录在案的港口信息往往只能覆盖大型港口、码头，对渔船经常使用的离家近的小型港口或沿海锚泊地难以统计。同时，渔船捕捞作业周期一般为10天左右，在海上同样存在着锚泊和靠近海岛休整的情况，这对于从数据角度进行港口定位造成了干扰。

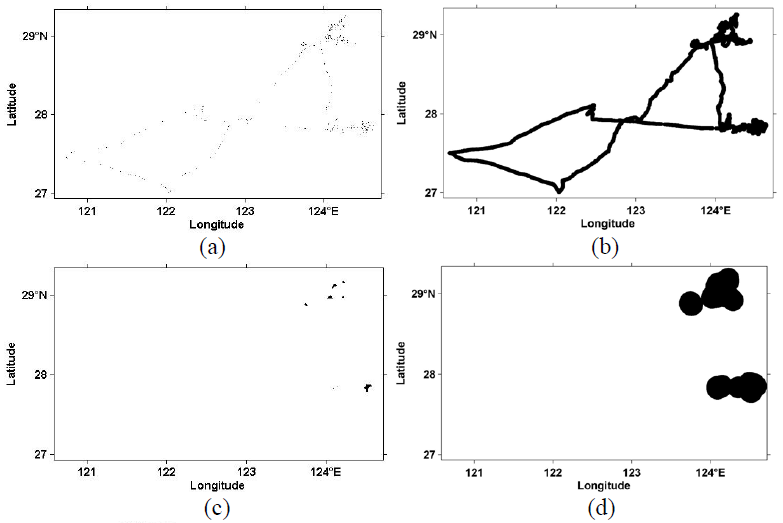
经过对比发现，停船时瞬时速度和船艏向数值为零。考虑进出港船速较慢，采样点密度在地图上局部增大。整合这些特点，对港口进行定位。具体来讲，一方面，利用网格法将地图按1′经度×1′纬度，约1平方海里进行划分，统计区块的数据点密度。另一方面，根据前人的研究成果，将 [0,1.5]kn和[-pi/3,pi/3]作为停船时的平均速度、船艏向。最终，完成对港口的识别。

取港口定位结果的经纬度信息，对数据进行分段。当一个采样点与港口的欧氏距离小于两海里，我们就认为该船的状态为在港。本文把离开港口-出海捕捞-返回港口的过程，称为一个航次。然而，我国沿海岛屿多，海岸线复杂，对应的锚泊地点多。按照上述划分的航次的方法，渔船路过港口A到达港口B的轨迹有可能被视为先抵达港口A，后抵达港口B，从而将AB之间的轨迹作为一个航次。这是由于网格法以及港口阈值等误差累积等因素造成的。不过没有关系，这个问题将会在下一个环节处理。

## 捕捞区识别与统计

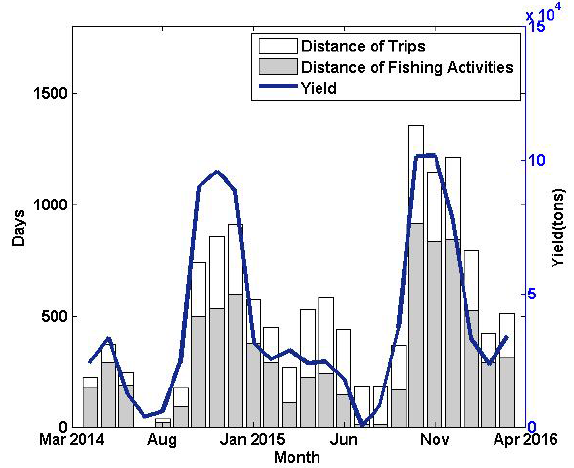
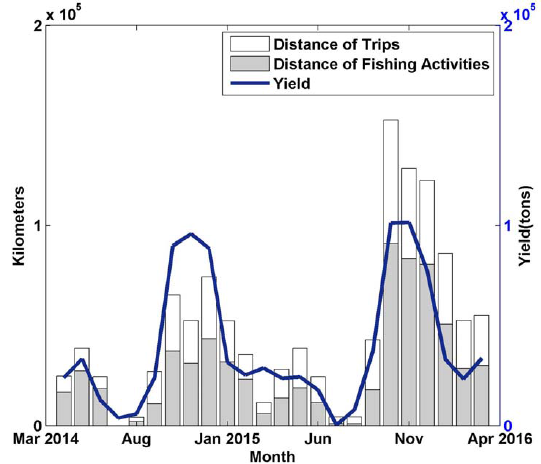
在以往的研究中，判断捕捞区这一环节主要采用监督学习或隐马尔可夫模型（HMM）等方法。它们都需要进行模型训练，因此需要大量的人工标定数据，且计算开销大。

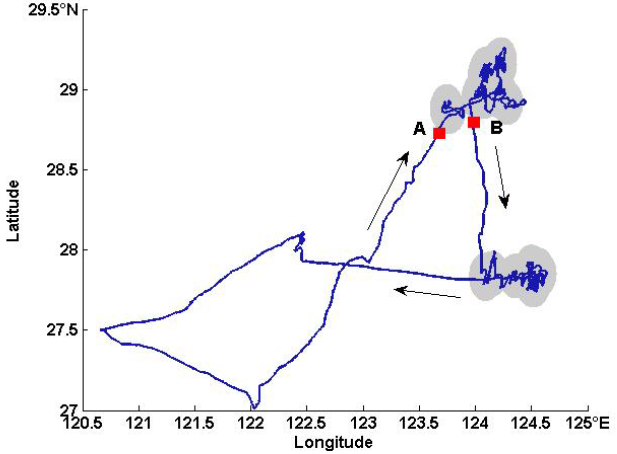
本文提出利用数学形态学进行捕捞区识别。它的计算开销取决于图片的像素数量，计算精度取决于结构元素（Structuring Element）的形状和大小。该方法的本质是利用了单拖船的捕捞作业特点：通过低速拖曳渔网在小范围内迂回航行——局部数据点密度较大。明显区别于渔船高速航行状态——数据点密度较小。通过“扩张”与“腐蚀”操作，剔除航行状态的数据点，保留捕捞状态的数据点，如图



## FAR评估

经过人工检查，绝大部分识别的捕捞区与人工判断相符，为了进一步评估结果的准确性，本文将捕捞时间和捕捞距离与官方公布的渔业经济数据进行对比。



图中，蓝色折线为2014年4月至2016年3月浙江省渔业产量，灰色条形图为捕捞区内的（1）总航行时间（2）总航行距离。捕捞区内的总航行时间是统计对应月份全部渔船在捕捞区内的累计总时间，捕捞区内总航行距离的计算方式类似。捕捞作业总时间与渔业产量的Pearson相关系数为89.80%，捕捞作业总距离与渔业产量的Pearson相关系数为82.11%。由于总航行时间和总航行距离的计算本身就有偏差，最终得到的强相关性一定程度上验证了捕捞作业行为被准确识别。

## 本章小结

本章提出了渔船活动识别系统FAR，并按照系统实现的顺序，依次介绍了数据预处理、港口定位与航次划分、捕捞区识别与统计三个环节。通过与雨夜经济数据对比可以看出，FAR的输出结果较准确的将VMS数据划分为在港、航行和捕捞三种状态。其中，在港状态在港口定位环节完成识别，捕捞状态在捕捞区识别环节完成，其余数据为航行状态。

# 投票分类算法

捕捞区的识别有助于对渔业资源的时空分布进行整体分析，我们通过比较同一时期不同年份的捕捞区，发现其有东移趋势。但是，捕捞区域是如何变化的？区域之间又有怎样的联系？这些问题我们难以从FAR的输出结果中判断。究其原因，这种变化是动态的、独立的、不均匀的。具体来说，虽然捕捞区形如图，但局部可能随时间扩大或缩小，某些热点区域可能会东移，也可能保持不动。这就要求我们对识别出的捕捞区进行合理的划分。

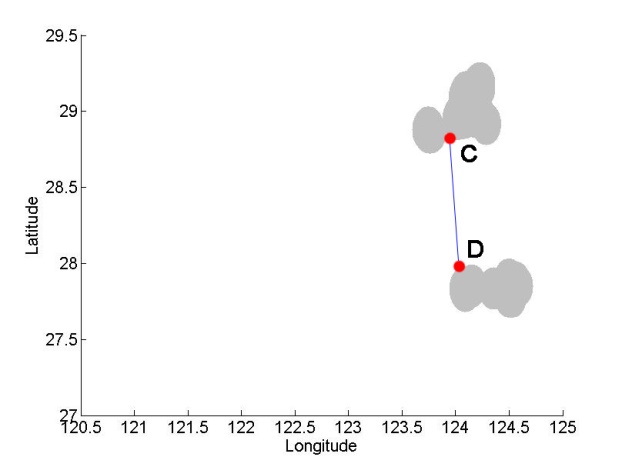
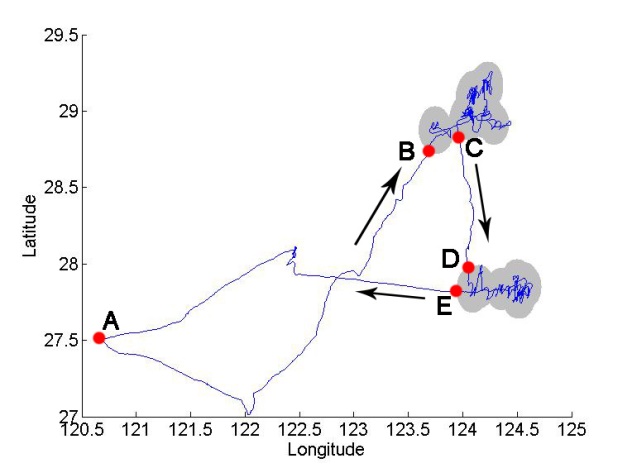
我们提出了一个投票分类算法（Voting Classification Algorithm，VCA），具体流程如图所示。首先，将渔船在捕捞区之间的航行状态数据抽象成线段，并作为选票；其次，对捕捞区热度图进行分层，通过比较不同层次间捕捞区的合并关系，对局部进行投票，得到判断结果。

本章将按照上述顺序介绍VCA，并对该算法输出的捕捞区域分类结果进行分析。

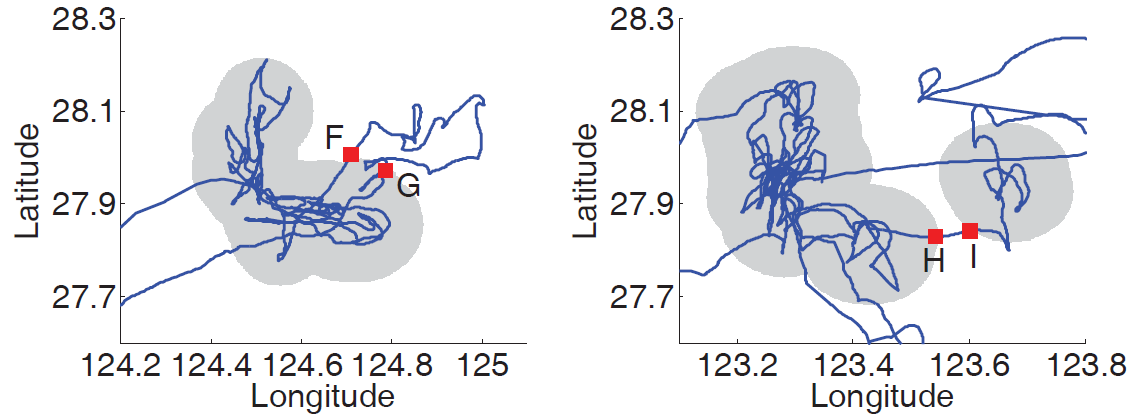
## 选票的产生

首先要明确的一点是，要想统计分析真实的渔业资源分布及变化情况，必然是要通过数据分析、遥感、气象学、生物学等一系列跨学科手段相结合才能完成的任务。但是我们可以通过分析VMS轨迹数据来近似分析渔业资源，更确切的说，是根据渔民的行为分析他们眼中的渔业资源分布情况。

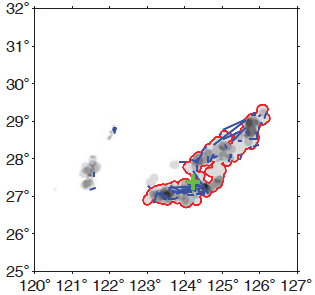
以往的VMS数据研究中，关注重点多是与捕捞相关的内容，比如识别捕捞行为、识别捕捞区域、分析捕捞区域变化。相较而言，对航行状态的数据关注不多。但是，作为渔民的一种行为，高速航行必然蕴含着渔民的自身经验。投票分类算法（VCA）最初的思路就是要从这部分数据中挖掘出对捕捞区分析有用的信息。

通过渔船行为识别系统（FAR），我们得到了每个航次的捕捞区域分布情况。将这部分信息与轨迹数据结合，如图所示，可以得到渔船在一个航次中对捕捞区的访问顺序和出入位置。我们只保留捕捞区域间的轨迹信息，并将其抽象为一条线段，如图。渔民将这两个捕捞区视为两个独立的个体。我们用线段CD来描述渔民的这种经验：线段的两端应该分属不同的捕捞区。我们称这种线段为航行线段。

在统计航行线段时，我们忽略了下面两种情况：图（a）中，渔船在从F点离开捕捞区之后，又通过G点返回了原来的捕捞区，线段FG将会把该捕捞区视为两个个体，与我们定义的线段含义相悖，所以不考虑这种情况；图（b）中，虽然点H和点I分属于两个不同的捕捞区，事实上，完全可以把两个捕捞区视为一个整体。这种由于渔船行为识别系统（FAR）识别精度造成的误差，我们通过设定阈值，忽略直线距离小于5海里的航行线段。



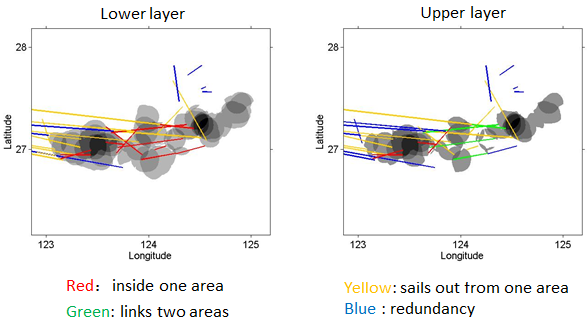
图将2014年第二季度统计的所有航行线段映射到捕捞热度图上进行分析。首先，根据连通性，我们把捕捞区粗略地分为七个部分，可以看出，航行线段不仅仅分布在不同部分之间，还分布于它的内部。以红色轮廓勾勒的区域为例，它是一个面积较大的捕捞区域，内部包含了大量的航行线段。回顾我们对航行线段的定义可知，该捕捞区应当被继续分割为多个小区域。接下来介绍如何进行分割。



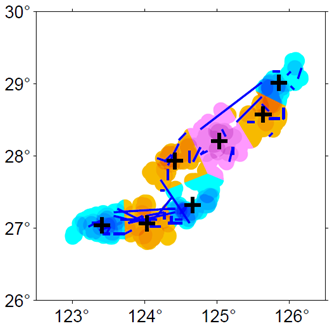
## 投票分类算法

投票分类算法的输入是航行线段和捕捞区热度图。

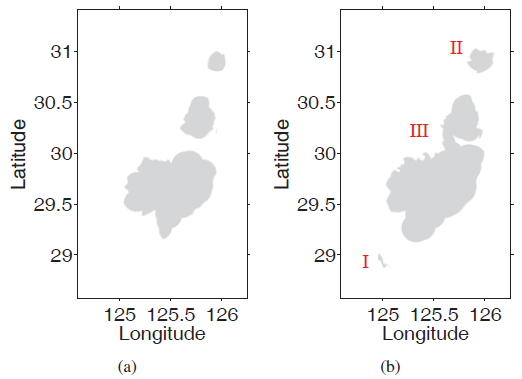
航行线段对应算法中的“选票”。根据线段两端落点的不同，共有四种不同的选票，如图所示：两端在同一个捕捞区内的选票用红色表示，称为反对票，记作-2票；两端在不同捕捞区内的用绿色表示，称为支持票，数值记作+2；只有一端在捕捞区内的用黄色表示，至少是对其中一个捕捞区分类的支持，记作+1；两端均不在捕捞区内的用蓝色表示，称为弃权票，记作0。所以，对于一个分类方案进行投票，就是根据落点判断选票的类型，将对应数字相加求和。对于不同的两个方案，比较投票结果，保留得票数高的方案。



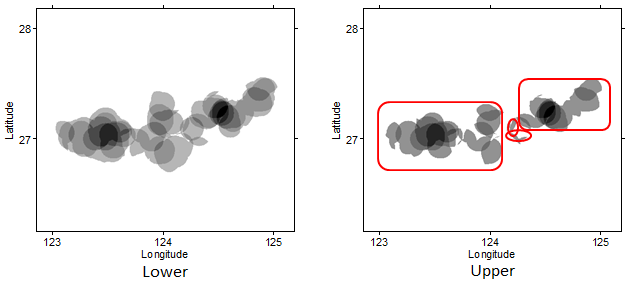
对于捕捞区域热度图，首先要明确的一点是，直接对它进行划分缺乏实践意义。举例来说，如果对图中红色轮廓包裹的区域进行如K-means算法的分类，如图所示，那么对于渔民而言分类前后没有区别，因为在两个相邻捕捞区之间捕捞的唯一区别就是跨过一条分界线。所以我们的目的是要将捕捞区划分为多个互不连通的小区域。



我们采用对捕捞区热度图分层的方法，对捕捞区热度图进行粗略分类。设热度阈值θ=0，当热度值大于θ，可以保留图上全部的捕捞区信息；随着θ增大，保留的信息减少，且逐渐以高热度值为核心相互独立。通过分析对比，当θ取值为热度最大值的25%时，效果最好。

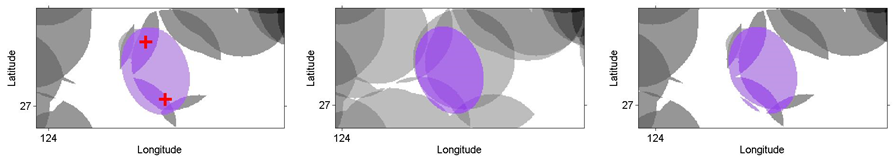


图为2015年第三季度捕捞区热度图的局部，通过令θ=热度最大值\*25%作为热度图的切入点，观察到的相邻两个图层（a）上一层（b）下一层。可以看出，在相邻图层间捕捞区存在三种变化方式：（Ⅰ）浮现，上一层的空白位置对应到下一层是一个独立的捕捞区域；（Ⅱ）扩张，上一层中的某个捕捞区对应到下一层范围变大，但仍保持独立性；（Ⅲ）合并，上一层的两个或多个捕捞区域在下一层构成了一个连通的捕捞区。和前面一样，按照捕捞区的存在与否，把（Ⅰ）（Ⅱ）两种情况视为独立的捕捞区，对（Ⅲ）中捕捞区合并的情况单独分析。



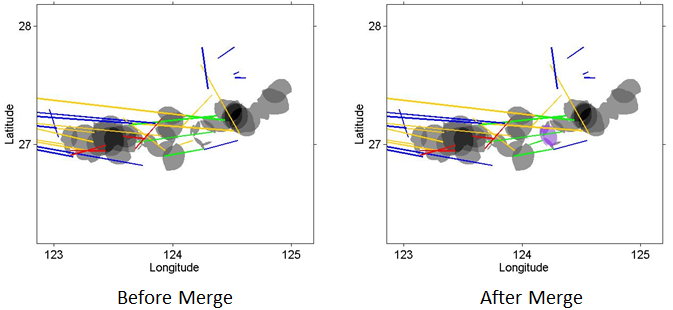
图展示了2014年第二季度捕捞区的局部，符合上述情况（Ⅲ）：上一层中四个独立的捕捞区在下一层中合并为一个整体。问题是：这些区域是否应当合并？或者保持某些部分的独立性？比如，图中一种可能的分类方式是保持左侧捕捞区域的独立，将右侧一大两小三个捕捞区合并为一个整体。具体方法是对任意组合的合并结果进行投票，根据得票数做出判断。

虽然明确了分类方法，但是又引入了新的问题：如何组合？如何合并？由于需要对全部组合进行投票，组合数将随着捕捞区数量上升呈指数型增长。为了简化模型降低计算量，我们设计了由近及远的、不可逆的方案。具体步骤为：取各个捕捞区的几何中心点两两连线，按线段长度排序；根据排序顺次进行合并、投票；若合并后得票数较高，则将两个区域视为一个整体，不再分开讨论其他分组情况；当尝试合并的两个几何中心点对应的捕捞区归属于同一个整体，则不再进行尝试。对于n个捕捞区域，这种处理方案最少只需要进行（n-1）次比较，即n个捕捞区域合并为一个整体；最坏情况只进行(n-1)n/2次比较，对应着n个捕捞区域全部独立的情况。在算法实现过程中，利用并查集算法简化判断归属的时间开销，且并没有出现时间开销过大的情况。这是因为当n比较大时（最多为14个），其中包含了大量的形如图中间部分的小捕捞区，这些区域往往合并前后得票相同。我们倾向于尽量合并为一个整体，于是将得票数相同判断为合并。



合并的具体方法如图。取长度最短的线段对应的两个捕捞区域进行分析，以两个捕捞区的几何中心点作为椭圆焦点，延长线较短一侧作为长轴，作椭圆连接两个捕捞区如图（1）。将椭圆与上一层捕捞区取交集如图（2），计算结果即为两捕捞区域的连接方式如图（3）。这种合并方法虽然存在与其他捕捞区域产生交集的可能，但胜在计算简便，基本能够满足合并的需求。

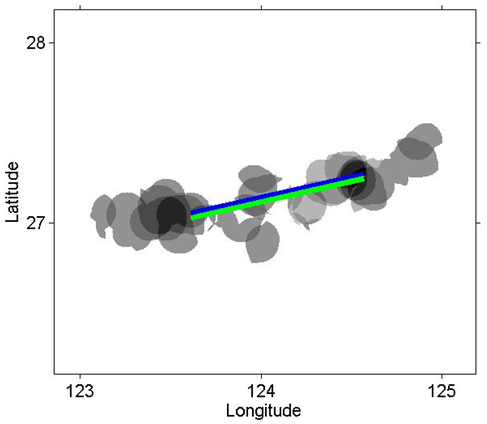
比较合并前后的得票数，如图所示，得票数相等，则确定合并两个捕捞区域。接下来，继续按照排序比较两两捕捞区合并前后的得票情况。最终得到分类方案。



## 海上路径

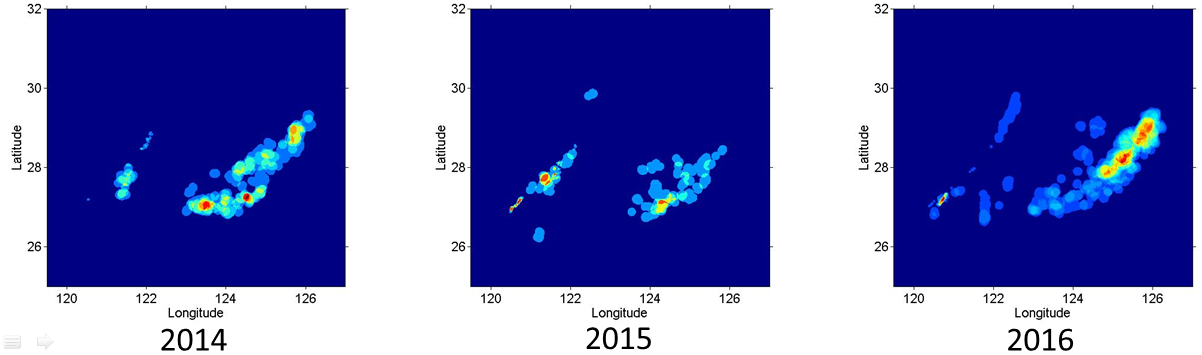
通过投票分类算法（VCA），实现了对捕捞区的细粒度划分。在此基础上，通过统计各个捕捞区的出入度和转移关系，得到了对于海上路径的分析。

如图所示，之前讨论的四个捕捞区通过VCA最终得到两个部分，统计渔船先后在这两个部分捕捞的轨迹，即选票中的“绿色支持票”，作图如图上蓝绿色线段。蓝色表示自西向东，绿色表示自东向西，线段粗细表示船次，也可以成为流量。图上两个捕捞区自西向东有三个船次，自东向西有三个船次。

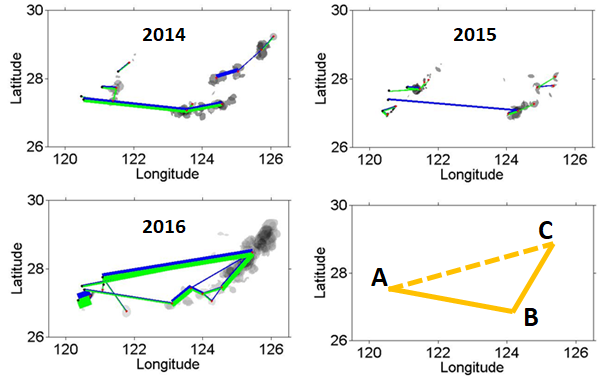


通过构建形如上图的海上道路，就可以构筑与陆地上的路网相似的路经网络，从而通过渔民的捕捞作业行为分析渔业资源的变化情况。

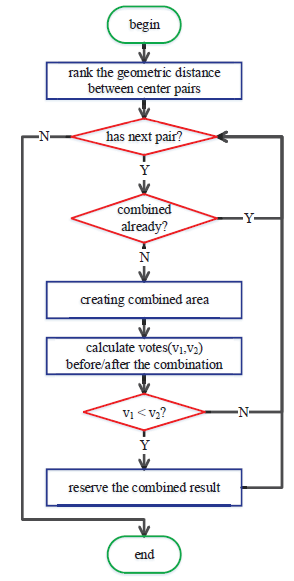
比如下图是2014年至2016年第二季度的捕捞热度图，通过观察可以发现捕捞区有向东北移动的趋势，还可以通过捕捞区的几何中心位置描述这种变化现象。但是这种刻画方式仍然太粗放。



通过投票分类算法（VCA）对捕捞区进一步划分，并构筑路网，如图所示，在2014年，大部分的船只出海沿着AB边航行，而到了2016年，只有少量船只沿AB边捕鱼，大部分的船次都选择直接沿AC边捕捞。选择更远的C点进行捕捞，这充分说明B点渔业资源减少。同时，2014年便开始在C点捕捞的渔民冒险和开拓意识较强，而2016年仍在B点捕捞的渔民相对而言比较保守。



## 本章小结

本章提出了一个投票分类算法（VCA）。它的输入包含两组参数：一个是捕捞区之间航行轨迹抽象的“航行线段”，另一个是渔船行为识别系统（FAR）的输出——捕捞区热度图。航行线段作为算法中的“选票”，表示线段两端点分属于不同的捕捞区，并由于线段端点的落点位置，分为+2、+1、0、-1四种不同的票型。捕捞区热度图作为分类的对应，从相邻两层间的变化进行分析。重点研究在上一层是多个捕捞区，而在下一层合并为一个整体的情况，称这种情况为“合并”。对于会在下一层合并的多个捕捞区，通过如图所示的流程，执行投票分类算法。算法的思路是尝试捕捞区的各种组合方式，通过投票的方法比较不同组合方式的得票数，取最高得票数对应的方案作为分类结果。遍历整个捕捞区，重复上述步骤，得到最终的分类结果。

我们把渔民出海捕捞的轨迹作为一种经验，融入到投票分类算法的设计中，完成了对捕捞区的细粒度分类。在此基础上，对渔船捕捞路径进行抽象，得到了捕捞活跃区域以及海上路径信息。

从长远来看，我们可以借鉴城市路网的分析经验，将知识迁移到投票分类算法的结果上，对海上活跃区域、路径规划进行分析。将它们与海洋信息（洋流、大气信息）相结合，这有助于从新的角度揭示渔业资源的时空变化规律，同时，为渔业管理部门制定计划、合理开发提供了依据。

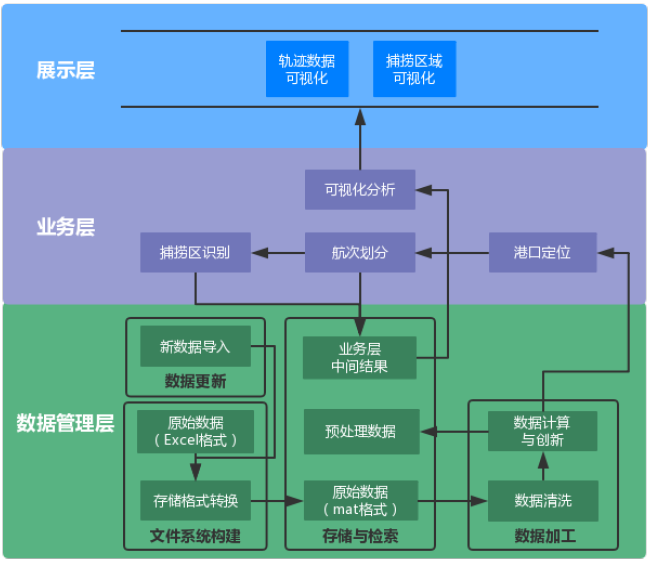
# 船舶轨迹数据分析系统

前面两章介绍了渔船行为识别系统（FAR）和投票分类算法（VCA），前者完成了捕捞区的识别任务，后者对捕捞区进行了分类，为分析渔业资源的变化规律和渔民行为提供了新的依据。

在此基础上，我们设计并实现了船舶轨迹数据分析系统。通过MATLAB的GUI编程将FAR封装为一个可视化系统，便于分析和操作。虽然在前面提到了FAR的设计思路，但在系统实现过程中，针对实际问题，同时为了提高运行效率，我们对算法进行了调整和优化，主要体现在利用空间换时间的设计思路上。本章将从系统架构入手，介绍系统的设计流程及功能实现。

## 系统架构

如下图所示，我们将船舶轨迹数据分析系统分为三个层次：数据管理层、业务层和展示层。数据管理层设计了整个系统的文件存储结构，此外完成了对原始数据的预处理操作。业务层承担FAR的主要算法内容，包括港口定位、航次划分和捕捞区识别。展示层主要从轨迹数据可视化和捕捞区域可视化两个角度展示系统运算结果



## 数据管理层

数据管理层主要完成了两部分内容：定义了文件系统结构，完成了原始数据的预处理操作。

### 文件系统结构

系统主界面通过Sys.fig和Sys.m两个文件组织实现功能，两个子界面分别通过 FishingBoatNumber.fig / FishingBoatNumber.m 和TrackNumber.fig / TrackNumber.m 实现。系统运行过程中调用的函数存放在与上述文件同一级的文件夹下。

原始数据以EXCEL形式组织，每个EXCEL文件对应一艘渔船的轨迹数据。这些数据存放在“\Excel\”目录下。

指定“\Excel”的上级目录为根目录，如“\Data\Excel\”中，“\Data\”为根目录。在实际系统运行中，根目录通过手动选择设定。

系统中间过程中产生的数据存放在与“\Excel\”同级的“\Rawdata\”目录中，如“\Data\Rawdata\”

### 数据预处理

我们使用的原始数据来自于浙江省温州市海洋渔业安全救助信息中心，共31艘单拖渔船轨迹数据，包含渔船编号、时间、经纬度、瞬时速度、船艏向等字段。数据在采集过程中由于通信质量等问题存在误差，同时为了向业务层提供可靠的数据输入，于是在数据管理层将数据预处理功能封装为一个独立的模块。

（1）数据格式转换。由于原始数据是Excel格式，且存在大量无用字段，所以数据预处理的第一步就是保留有效数据并存储为Matlab对应的.mat文件。通过更大规模数据的观察发现，瞬时速度和船艏向的数据波动性较大且缺失记录较多，所以实际只保留时间和经纬度三个字段。关于忽略瞬时速度和船艏向如何进行港口定位的问题将在业务层进行分析。又因为北斗卫星导航系统记录的时间和经纬度与我们常规意义上的不同，通过转换公式将其转换为我们熟知的形式。比如VMS原始数据中的时间记录是自1970年1月1日0时0分起至记录时刻的秒数。

（2）数据清洗。数据异常现象普遍存在于传感器采集的数据中。通过对轨迹数据的可视化图像进行观察，发现经纬度存在归零的情况，或者是某条记录的经纬度坐标明显偏离上下文。对于这样的现象，提出以下几点约束规则：1）通过查阅文献，单船拖网渔船的最高航速不高于30节，我们通过两点之间的球面距离和时间差计算平均速度与该阈值比较进行筛选；2）在已知我国渔船不会出现在经纬度<0,0>点的前提下，删除经纬度同时为零的记录。在此基础上得到的数据仍然在局部范围内存在异常，但我们认为这是数据精度造成的问题，为了保持数据的真实性，不再对原始数据进行平滑处理。此外，原始数据中存在数据缺失的情况，这主要是由于海上通信环境恶劣造成的。由于后面要分析捕捞区的变化规律，对于这部分数据，我们不进行插值或其他处理。

由于我们的渔船监控系统（VMS）基于北斗卫星导航系统，平均采样频率达到了三分钟一次，相较于国外两小时的采样间隔，足以刻画渔船在海上的行为信息，所以不再考虑进行插值操作。

## 业务层

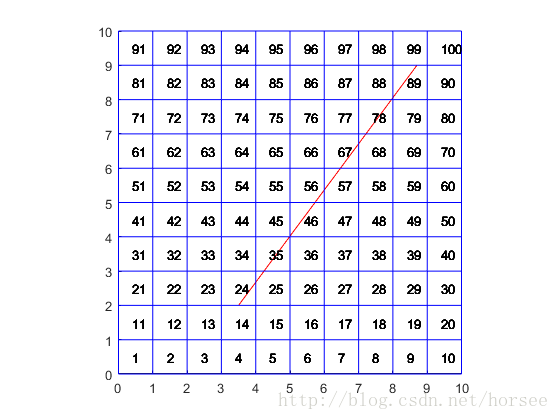
业务层是整个系统的算法核心，是在前文提到的渔船行为识别系统（FAR）上的改进，包括了港口定位、航次划分和捕捞区识别三个模块，前一个模块的输出作为后一个模块的输入。港口定位模块的输入即为经过预处理的数据。

### 港口定位

由于原始数据中没有港口位置的标识，又缺少额外的辅助信息，如航海日志，所以港口位置无法从数据中直接获得。又因为渔船的行为分析以航次为单位，而航次之间的结点即为港口，所以我们把港口定位作为第一个任务。

前文在介绍FAR时提到同时从点密度和<瞬时速度，船艏向>=<0,0>两方面确定港口的方法。然而在实际处理中，我们忽略了瞬时速度和船艏向字段，只保留了时间和经纬度信息。此外，还需要注意以下两个问题：a.停船现象不是港口独有，海上同样存在这些特征；b.停船后存在关机现象，不再采集数据。基于这些原因，为了寻找港口独有的、仅基于经纬度坐标的特征，我们提出五种思路。

（1）累计停留时间

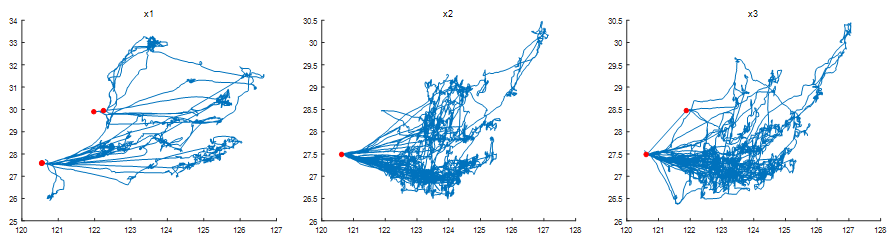
 利用网格法对地理位置进行分割，根据线段长度统计渔船在每个方格内的停留时间，累计停留时间长的区域对应着港口。这个方法的本质是利用渔船在港口这一确定位置停留，而在海上的停留位置随机的特点进行分析。优点是能够区分港口停泊与海上锚泊。缺点是：a.只适合统计进出频繁的港口，对于访问次数较少的港口易与海上锚泊混淆。b.对数据量要求大，数据越多港口区域的累计停留时间与其他区域差距越大。c.计算量大，一方面，网格越密越能区分港口停泊与海上锚泊，但相邻两点之间往往跨过更多的区域，计算量呈指数级增长；另一方面，只有极少数的数据对判断捕捞区产生贡献，无效计算量大。

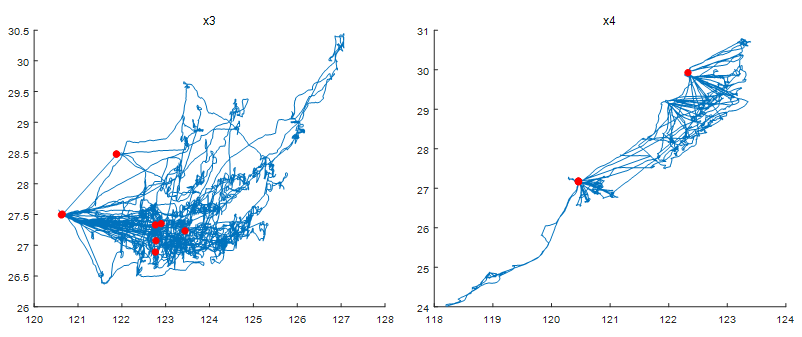
（2）相邻坐标重复

相邻两条记录的经纬度坐标值相同，表明船舶近似静止。一方面，海上风浪对渔船影响比港口大，这一现象在港口发生的概率远大于海上发生的概率，另一方面，同样是因为港口固定而海上锚泊位置随机，在港口对应区域发生相邻坐标重复的频度远远高于海上。优点是判定条件简单，缺点是：a.系统关机对判定结果有决定性影响，渔民为了省电甚至可能会在入港前提前关闭系统；b.难以识别访问次数少的港口,易与海上锚泊混淆。

对于某艘渔船（编号x2）数据进行分析，共存在48次坐标重复情况，其中港口38次（下图）。同样通过网格法，设置阈值判断港口。这里，网格大小影响着港口判断的经度，定为1’ ×1’，相当于1平方海里的误差。对于不同渔船数据的验证见下图。对于港口点较少的x1/x3而言，判断准确；但是对于多港口的x4，只能判断出个别主要港口。为了识别更多港口，放宽阈值限制会导致错误的把x3中海上锚泊点识别为港口——难以平衡阈值与港口识别准确率的关系。究其原因还是由于渔船在部分港口停船时间短，访问次数少。





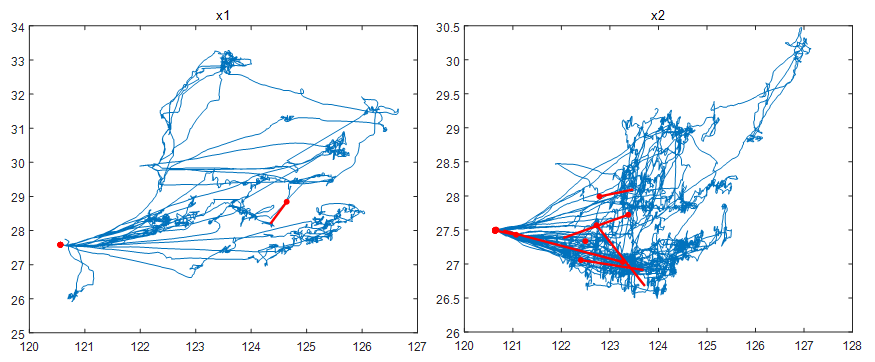


（3）超低速

考虑到两个航次之间存在关机现象（数据采集停止），将对港口的判断产生严重影响。假设关机后船舶保持不动，利用关机前最后一条记录和开机后第一条记录计算得到的平均速度应该极小。即对于关机行为，渔船的位移距离短且时间间隔长，以此作为依据判断港口。优点是：把关机这一干扰因素纳入考虑范围。缺点是：一旦渔船不关机，或渔船在关机状态下发生了运动，都将无法识别。

在对数据x2的研究中，有两种方法确定超低速的阈值：动态方法，将每艘渔船10倍的最低速作为阈值；静态方法，用常数作为所有渔船的阈值。经过分析发现，最低速基本发生在7到9月，对应着禁渔期。时间间隔短则半个月，长则接近两个月，对应计算出的最小速度差别较大，无法用动态方法全部囊括港口停船行为。所以，最后确定用静态常量作为阈值。

在寻找超低速的阈值时遇到的另一个问题是数据缺失。下图中的红色线段是以相邻数据间隔大于3天为标准筛选得到的，而正常的数据间隔为3分钟。观察左侧x1的数据可以发现，在两个明显的港口中，上方的港口不存在相隔三天的数据点，有两种可能：船舶始终开机；关开机间隔不超过3天。前一种可能性比较小，而后一种则意味着船舶靠港时间不超过三天，易与海上锚泊混淆，为“超低速”这种方法带来了困难。



为了避免这些干扰，将阈值设置偏小。虽然这样一来部分港口就难以识别，但可以借鉴其他渔船的港口识别结果，即不再单独识别一艘渔船对应的港口，而是将全部渔船的港口识别结果叠加，维护一个公共港口信息表。通过不同阈值判断港口确定了最终阈值设为2\*1e5节，如下图。对于多艘渔识别到了同一个港口，这里使用0.1’×0.1’的网格进行去重——落在同一个网格中的多个港口点视为同一个港口，最终确定了43个港口点。



（4）识别网次

拖网渔船的捕捞行为可以通过速度变化进行识别。通过分类可以区分渔船航行和捕捞行为，进而的可以通过航行的公共点识别港口。缺点是：需要计算量大，且需要人工标记数据对模型进行训练；捕捞行为识别的准确度无法保证。

（5）轨迹周期性

渔船捕捞行为的起始点和终止点都是位于港口，每个航次呈现出明显的周期性，可以利用轨迹的这种周期性直接识别航次，而航次之间的交点即为港口。这种方法同时考虑的因素较多，计算量和误差大；单拖渔船一个航次的时间一般十天左右，整个数据集的包含的航次数少则不足十个，多则二三十个，不足以训练模型。

前三种方法通过数据自身的特征进行分析，后两种方法本质是分类问题，需要训练模型，经过比较最后采用第三种方法“超低速”作为判断的依据。需要注意的是，方法二的判断条件最简单，计算量最小。目前准确率不足的情况会随着数据的累积逐渐改善，未来可以作为主要的判定方法。此外，在时间允许的情况下，五种方法可以交叉使用互相验证，提高准确度。

### 航次划分

由于采用“超低速”的方法识别港口，维护了一个公共的港口列表。在划分航次的过程中就需要将当前记录与每个港口坐标进行比较，判断是否在某个港口的一定范围内。从渔船轨迹数据中截取航次的方法就是，标记每条记录的在港状态，两个港口之间的即为一个航次。如下图所示，对该渔船轨迹共划分29个航次。



**+**

**=**



然而在实际运行时，31艘渔船数据划分航次的时间开销是189.65s，大大超出了可以接受的范围。假设港口信息表有n个港口坐标，渔船数据m条记录，共需计算nm次才能完成该艘船的航次划分。这其中包含了大量的冗余计算，而且随着数据量的增加，n和m都会增大，时间开销呈指数增长。为了解决这个问题，我们引入一个01矩阵（网格法）。首先，矩阵初始清零，将港口信息坐标按照映射关系f投影到矩阵上，对应矩阵元素置1；其次，轨迹数据中的每条记录通过同样的映射关系f进行查询，元素为1则为在港状态，为0则为离港状态，流程如下图所示：



矩阵



通过这样的变换，在对一艘船进行航次划分时共需（n+m）次计算，同时额外需要一个矩阵记录状态。经过统计，31艘渔船数据划分航次的时间开销是0.7s。

需要注意的是，对于上面的例子，共划分了99个航次。前后两种方法划分的航次数相差较大，这是因为网格法扩大了港口范围，尤其是将港口附近的轨迹切分成多段（前一种方法识别出的航次同样存在这样的问题）。已知平均数据采样间隔是3分钟，经过统计，99个航次中有66个航次小于20个数据点（约1小时），77个航次小于60个数据点（约3小时），81个航次小于120个数据点（约6小时）。我们认为6个小时内无法完成一个航次，将少于120个数据点的航次剔除保留结果如下，作图为99个航次的叠加，右图为剔除后剩余18个航次叠加，基本保留了全部的信息。



### 捕捞区识别

我们在利用形态学进行捕捞区识别时，通过“腐蚀”和“扩张”操作，消除掉航行轨迹的同时，保留捕捞行为对应的轨迹信息。其中存在两个问题：a.时间开销巨大，这是由于轨迹所映射图像的分辨率过高导致的。具体来讲，下图中蓝色标记区域不是捕捞区，但是由于轨迹交叉造成了局部数据点密集程度增强，在进行形态学处理时可能会错误识别为捕捞区。为了避免这种错误，采用了提高图像清晰度的方法，增大了蓝色标记区域数据点之间的像素数。不仅时间开销大，其中还包含了大量的冗余计算冗余计算

