数字中国建设峰会
DIGITAL CHINA SUMMIT

DCIC
数字中国创新大赛
DIGITAL CHINA INNOVATION CONTEST

培育数字经济新动能 助推数字中国新发展

2020数字中国创新大赛

Digital China Innovation Contest, DCIC 2020

数字政府赛道——智慧海洋建设

海洋智慧监管系统

队伍：Pursuing the Past Youth

中国 · 福州 FUZHOU.CHINA

数字中国建设峰会 DIGITAL CHINA SUMMIT | DCIC 数字中国创新大赛 DIGITAL CHINA INNOVATION CONTEST

刘世欢

单位/职位：
哈啰出行/数据挖掘工程师

曾获奖项：
- 天池合肥心电亚军
- 2018 融360天机大赛二等奖
- 天池杭州地铁人流量预测季军
- 拍拍贷第四届魔镜杯Rank4
- 天池第二届阿里巴巴大数据智能云上编程大赛Rank4

陈欣

单位/职位：
网易游戏/数据挖掘工程师

曾获奖项：
- Kaggle-IEEE Fraud Detection 金牌
- JData2019 亚军
- KDD2019 Rank10
- 拍拍贷第四届魔镜杯Rank4
- 天池第二届阿里巴巴大数智能云上编程大赛Rank4

姚程栋

单位/职位：
华南理工大学/本科生二年级

曾获奖项：
- 天池合肥心电亚军
- 2018广东工业智造大数据创新大赛Rank6

黄成斌

单位/职位：
华东师范大学/研究生二年级

曾获奖项：
- 天池合肥心电亚军
- 首届心电智能大赛二等奖
- 2019数字中国创新大赛消费者人群画像Rank3

数字中国建设峰会
DIGITAL CHINA SUMMIT

DCIC
数字中国创新大赛
DIGITAL CHINA INNOVATION CONTEST

A

算法赛

目录
Contents

01 赛题分析
02 特征工程
03 模型搭建
04 其他尝试

数字中国建设峰会
DIGITAL CHINA SUMMIT

DCIC
数字中国创新大赛
DIGITAL CHINA INNOVATION CONTEST

01

赛题分析

赛题任务

本赛题为智能算法赛，主题是**基于位置数据对海上目标进行智能识别和作业行为分析**，要求选手通过分析渔船北斗设备位置数据，得出该船的生产作业行为，具体判断出是**拖网作业**、**围网作业**还是**流刺网作业**。

赛题难点

① 位置信息存在上报不准确的情况

② 数据有丢失或频繁上报等异常情况

③ 存在部分数据的标签和实际作业方式不符的情况

数字中国建设峰会
DIGITAL CHINA SUMMIT

DCIC
数字中国创新大赛
DIGITAL CHINA INNOVATION CONTEST

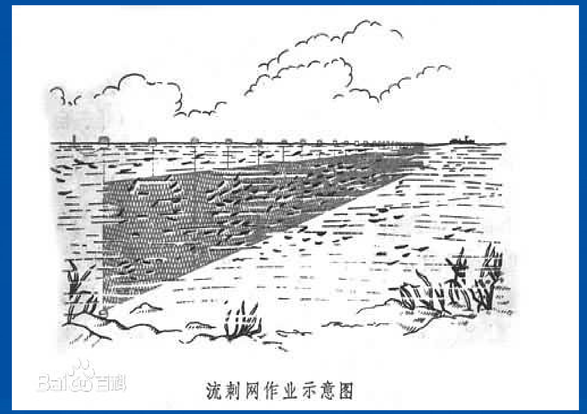
02

特征工程

渔船作业方式介绍

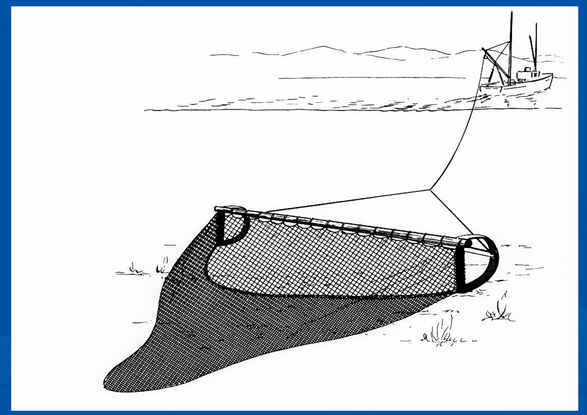
刺网

将**长带形的网列敷设于水域中**，使鱼刺入网目或被网衣缠络后加以捕捞。



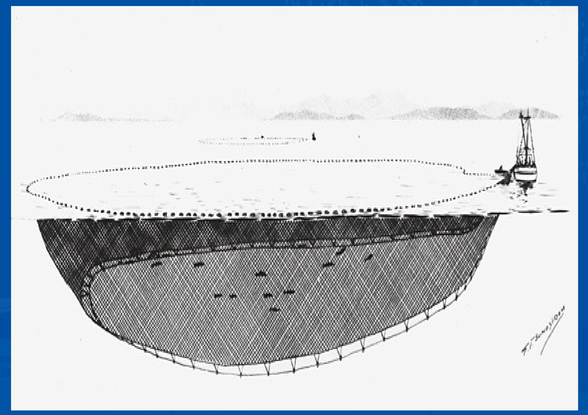
拖网

拖网是利用船舶的运动，**拖曳渔具在海底或海水中前进**，迫使渔具经过水域中的鱼虾蟹等捕捞对象进入网囊，达到捕捞的目的。



围网

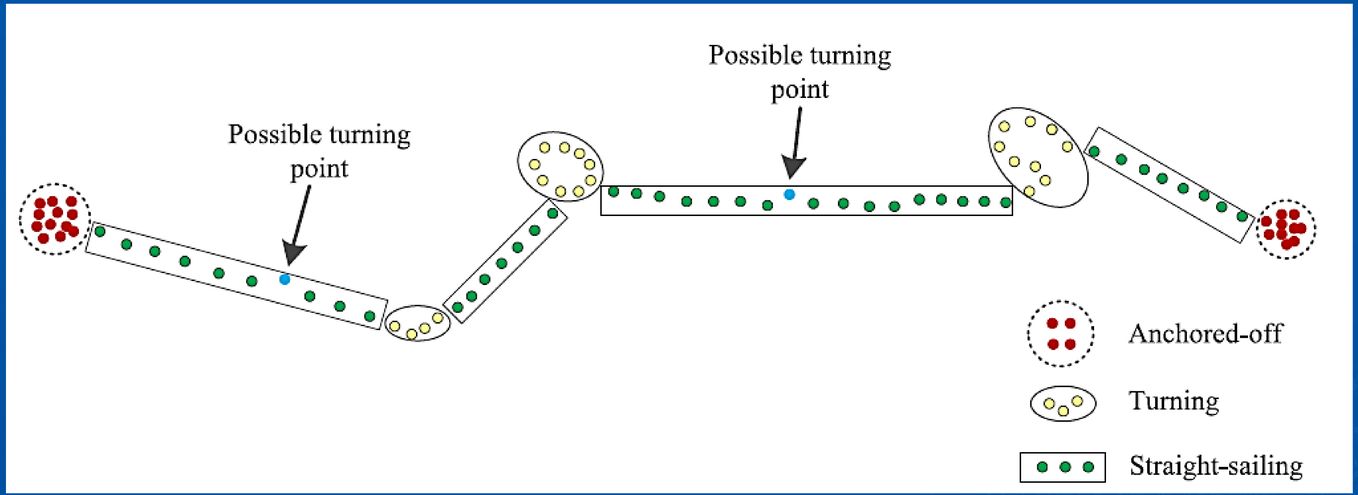
以长带形或一囊两翼形网具**包围鱼群进行捕捞**的作业方式。为了完成这一系列作业，围网渔船需要有灵活的船身，这样方便快速的**兜圈子**。



渔船作业行为状态

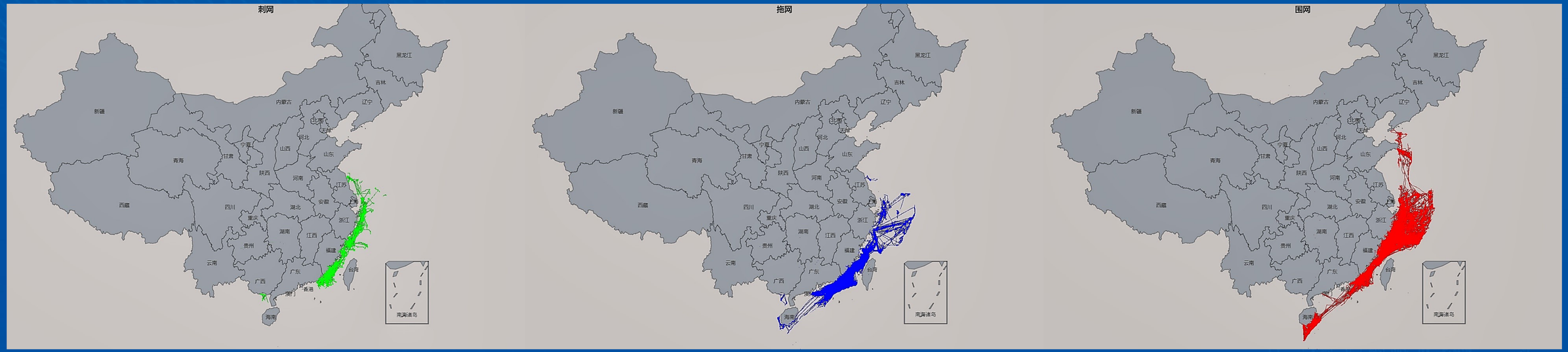
船舶的行为状态主要可以分为三类[1]：

1. **锚点**：船舶休息（停靠）点
2. **转弯**：船舶转弯点
3. **直行**：船舶直行点



[1] Sheng, K., Liu, Z., Zhou, D., He, A., & Feng, C. (2018). Research on Ship Classification Based on Trajectory Features. Journal of Navigation, 71(1), 100-116. doi:10.1017/S0373463317000546

渔船作业方式的位置分布



刺网渔船：

由于刺网作业方式的限制，其主要分布在我国的东南沿海的近海区域

拖网渔船：

拖网渔船相对于刺网来说，活动范围稍广，个别会航行在远洋区域

围网渔船：

围网渔船一般在远洋作业，需要的捕鱼周期较长

基于上述情况，可以得出不同渔船**活动的海域位置**是识别其作业方式的重要特征

基于位置（经纬度）的特征工程

1. 常规统计特征

均值、方差、中位数、分位数、偏度、峰度等

```
for col in ['x', 'y']:
    group_df = data.groupby(['id'])[col].agg({col+'_mean': 'mean',
                                              col+'_std': 'std',
                                              col+'_max': 'max',
                                              col+'_min': 'min',
                                              col+'_skew': 'skew',
                                              col+'_kurt': kurtosis,
                                              col+'_q25': lambda x: np.quantile(x, 0.25),
                                              col+'_q75': lambda x: np.quantile(x, 0.75),
                                              col+'_first': 'first',
                                              col+'_last': 'last'}).reset_index()
```

2. 众数特征

渔船出现最多次的经纬度及其频次

```
mode_df = data.groupby(['id', 'x', 'y'])['time'].agg({'mode_cnt':'count'}).reset_index()
mode_df['rank'] = mode_df.groupby('id')['mode_cnt'].rank(method='first', ascending=False)
for i in range(1, 4):
    tmp_df = mode_df[mode_df['rank']==i]
    del tmp_df['rank']
    tmp_df.columns = ['id', 'rank{}_mode_x'.format(i), 'rank{}_mode_y'.format(i), 'rank{}_mode_cnt'.format(i)]
    data_ = data_.merge(tmp_df, on=['id'], how='left')
```

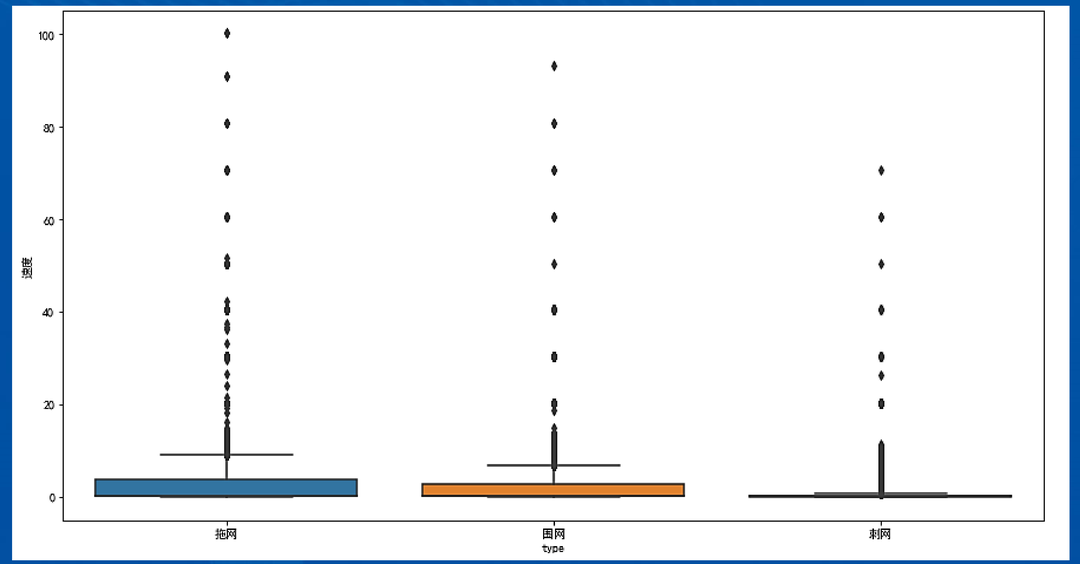
3. 特殊状态

① 速度小于某阈值时的经纬度统计（锚点）

② 速度大于0时的经纬度统计（运行）

```
for col in ['lat', 'lon']:
    print(col)
    group_df = data[data['速度']==0].groupby(['渔船ID'])[col].agg({'stop_'+col+'_mean': 'mean',
                                                                   'stop_'+col+'_max': 'max',
                                                                   'stop_'+col+'_min': 'min',
                                                                   'stop_'+col+'_nuniq': 'nunique',
                                                                   'stop_'+col+'_q1': lambda x: np.quantile(x, 0.10),
                                                                   'stop_'+col+'_q2': lambda x: np.quantile(x, 0.20),
                                                                   'stop_'+col+'_q3': lambda x: np.quantile(x, 0.30),
                                                                   'stop_'+col+'_q4': lambda x: np.quantile(x, 0.40),
                                                                   'stop_'+col+'_q5': lambda x: np.quantile(x, 0.50),
                                                                   'stop_'+col+'_q6': lambda x: np.quantile(x, 0.60),
                                                                   'stop_'+col+'_q7': lambda x: np.quantile(x, 0.70),
                                                                   'stop_'+col+'_q8': lambda x: np.quantile(x, 0.80),
                                                                   'stop_'+col+'_q9': lambda x: np.quantile(x, 0.90)}).reset_index()
    features = features.merge(group_df, on=['渔船ID'], how='left')
```

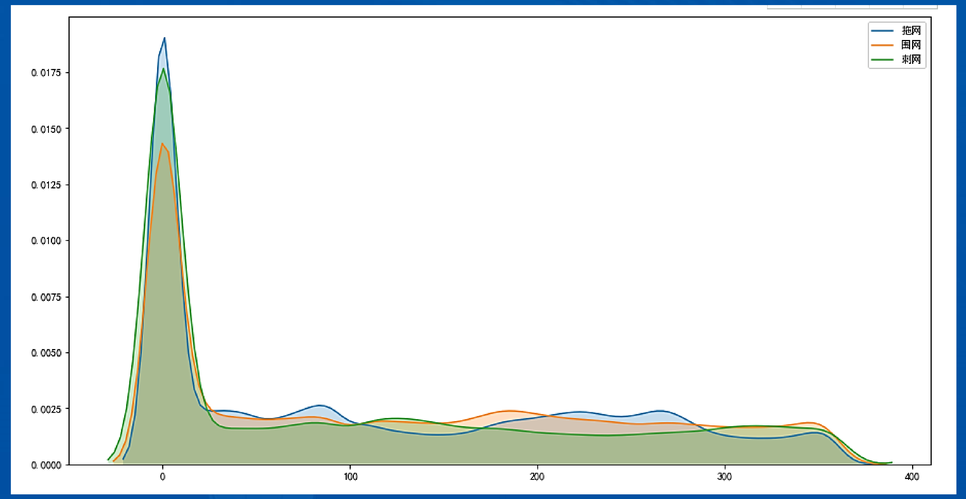
基于速度的特征工程——直行



1. 速度的均值、方差、偏度、峰度等
2. 速度>0时的统计特征（去除0对于统计的影响）
3. 速度一阶差分统计特征

```
group_df = data[data['velocity']>0].groupby(['id'])['velocity'].agg({
    'velocity_mean_new': 'mean',
    'velocity_STD_new': 'std',
    'velocity_SKEW_new': 'skew',
    'velocity_kurt_new': kurtosis,
    'velocity_q10_new': lambda x: np.quantile(x, 0.10),
    'velocity_q20_new': lambda x: np.quantile(x, 0.20),
    'velocity_q30_new': lambda x: np.quantile(x, 0.30),
    'velocity_q40_new': lambda x: np.quantile(x, 0.40),
    'velocity_q50_new': lambda x: np.quantile(x, 0.50),
    'velocity_q60_new': lambda x: np.quantile(x, 0.60),
    'velocity_q70_new': lambda x: np.quantile(x, 0.70),
    'velocity_q80_new': lambda x: np.quantile(x, 0.80),
    'velocity_q90_new': lambda x: np.quantile(x, 0.90),
}).reset_index()
```

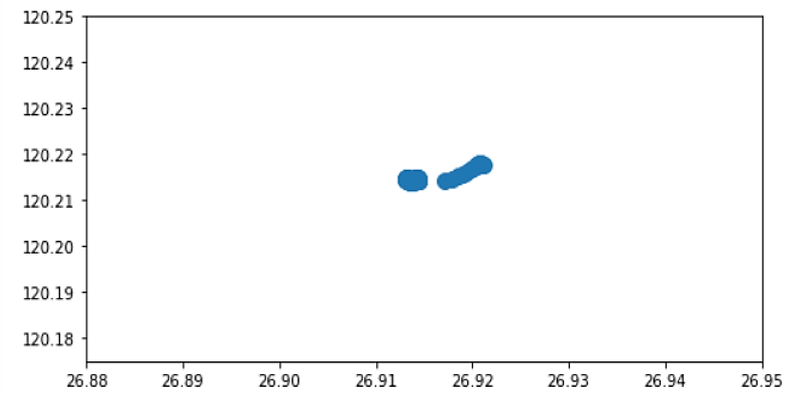
基于方向的特征工程——转弯

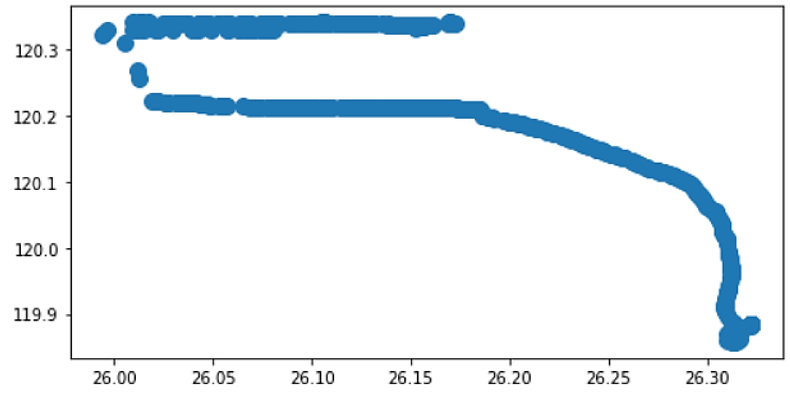


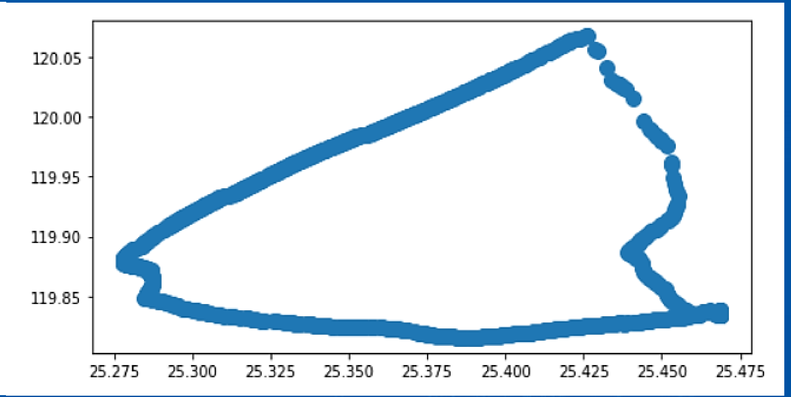
1. 方向统计特征（方差）
2. 方向一阶差分统计特征
3. 方向一阶差分绝对值大于0的速度统计特征
4. 方向离散化后的占比特征

```
data['degree'] = data['orientation'].apply(lambda x:(int(round(x/30)))*30)
degree_df = data.pivot_table(index='id',columns='degree',values='x', dropna=False, aggfunc='count').fillna(0)
degree_df.columns = [str(f)+'_orientation_count' for f in degree_df.columns]
degree_df.reset_index(inplace=True)
data_ = data_.merge(degree_df, on='id', how='left')
```

渔船作业方式的轨迹特点







刺网渔船：

通常需要将网置于某处，等待即可，刺网捕鱼时船往往会静止或随着海风或海流缓慢前进，移动范围较小

拖网渔船：

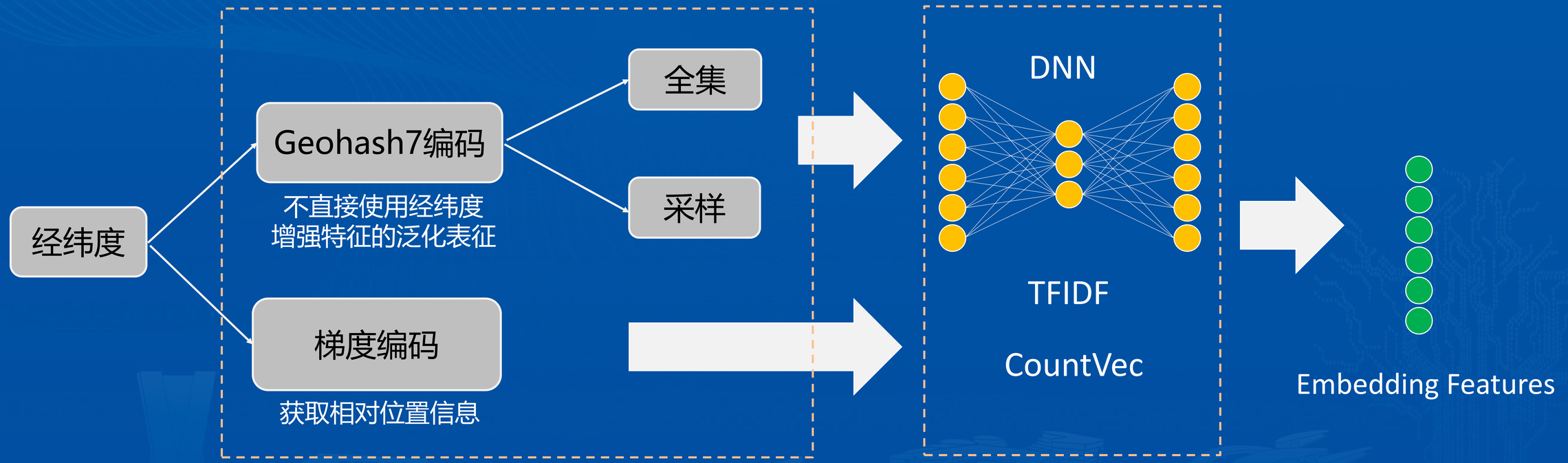
需要依靠渔船拖曳渔具进行捕鱼，大部分时间处于直行状态，转弯次数较少

围网渔船：

利用长带形或一囊两翼的网具包围鱼群，捕鱼的过程中通常需要兜圈子，转弯幅度较大

基于上述情况，可以得出不同渔船轨迹的特点也是识别其作业方式的重要特征

基于轨迹序列绝对和相对位置的复合向量编码



数据处理层

特征编码层

DNN

Word2Vec

① 首先将经纬度利用geohash将经纬度进行编码

② 将每艘渔船各时刻的位置转化为文本向量

③ 利用文本向量训练Word2Vec模型

④ 利用得到的Word2Vec模型计算各个位置的embedding特征

⑤ 取每艘渔船所有时刻位置embedding特征的均值

Node2Vec [2]

① 首先将经纬度利用geohash将经纬度进行编码

② 构造节点与边数据，位置为节点，边为两个节点间的连接关系

③ 利用节点边数据训练Node2Vec模型

④ 取每艘渔船所有时刻节点特征的均值

[2] Scalable Feature Learning for Networks. A. Grover, J. Leskovec. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2016.

数字中国建设峰会
DIGITAL CHINA SUMMIT

DCIC
数字中国创新大赛
DIGITAL CHINA INNOVATION CONTEST

03

模型搭建

模型训练及融合

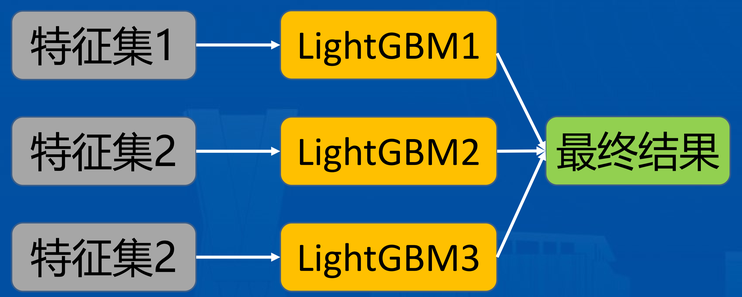
单模型

5折LightGBM

0.9026 Rank1

排名	参与者	组织	score	最优成绩提交日
1	Pursuing the Past You...	OTTO	0.9040	2020-03-12
2	Champion Chasing Boy	武汉大学	0.9011	2020-03-12
3	liu123的航空母舰	交控科技	0.8995	2020-03-12
4	天才海神号	北京航空航天大学	0.8994	2020-03-12
5	大白	湖南大学	0.8990	2020-03-12
6	lazji	电子科技大学	0.8917	2020-03-12

模型融合



0.9040 Rank1

数字中国建设峰会
DIGITAL CHINA SUMMIT

DCIC
数字中国创新大赛
DIGITAL CHINA INNOVATION CONTEST

04

其他尝试

其他尝试

① 将轨迹转化为二维图像，利用resnet识别（效果在0.80左右）

② 将轨迹转化为一维信号，利用resnet识别（效果在0.82左右）

③ 取出resnet最后的FC层为特征与其他特征合并一起

④ 取出resnet分类的概率与其他特征合并一起

⑤ 自己制作数据利用神经网络识别轨迹是否包含圈

⑥ ……

数字中国建设峰会
DIGITAL CHINA SUMMIT

DCIC
数字中国创新大赛
DIGITAL CHINA INNOVATION CONTEST

B

可视化

目录
Contents

01 RS-HCD算法
02 可视化平台
03 船舶行为画像
04 海域监管

数字中国建设峰会
DIGITAL CHINA SUMMIT

DCIC
数字中国创新大赛
DIGITAL CHINA INNOVATION CONTEST

01

RS-HCD算法

抽样-哈希碰撞算法

① 对纬经度四舍五入到两位小数，以克服不同测量体系造成的系统误差

② 对于每个设备聚合得出相应的路径点集合，以及按时间倒排的轨迹向量

③ 将设备根据时序相似性，用LSH算法分成$\lfloor\sqrt{N}\rfloor$个桶，再根据空域相似性进一步剪枝，以减小搜索空间

④ 对同个桶中的不同设备，随机抽样部分路径点，检测哈希碰撞的概率

⑤ 剔除碰撞概率低于一定值(如0.8)的设备组合，余下的即匹配结果候选集

⑥ 若候选集有多个组合，则用FDTW算法计算轨迹向量间的最短L_1距离，取最小的一组作为最终结果

传统方法

× 需遍历完整条轨迹，**耗时长**

匹配完全部渔船轨迹需要~17小时

× 时间复杂度$O(N^2L + NL)$

N为待匹配轨迹量，L为平均轨迹长度

VS

RS-HCD

√ 在抽样中动态剪枝，**耗时短**

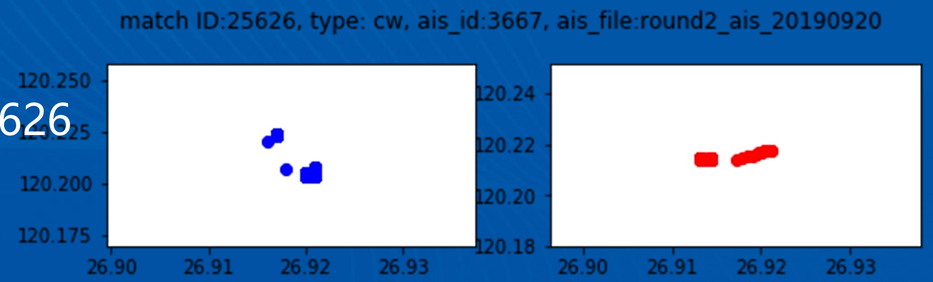
匹配完全部渔船轨迹只需~1分钟

√ 时间复杂度$O(N + NL)$

N为待匹配轨迹量，L为平均轨迹长度

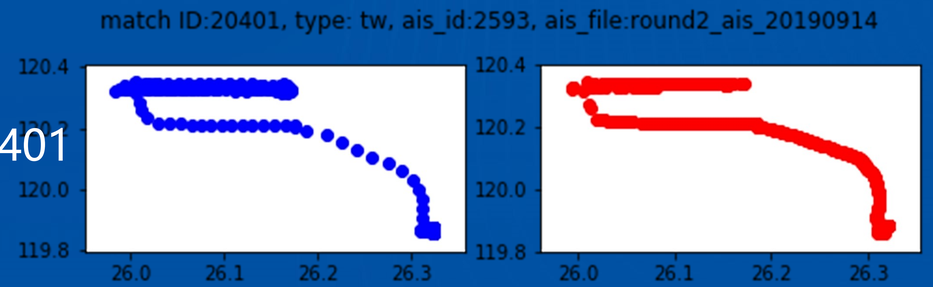
部分匹配结果展示

渔船ID: 25626



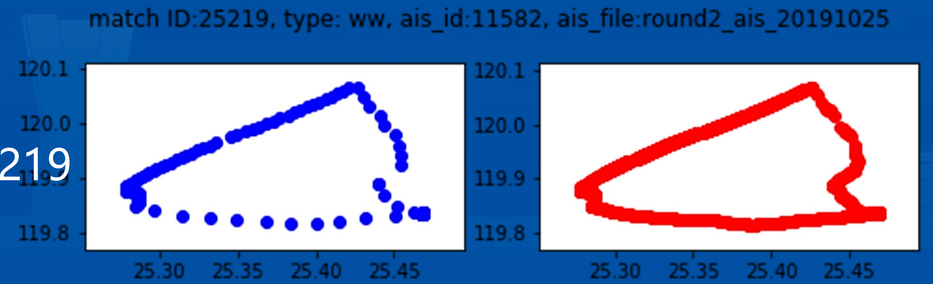
AIS ID: 3667

渔船ID: 20401



AIS ID: 2593

渔船ID: 25219



AIS ID: 11582

匹配之后的好处：

- √ 补充部分渔船缺失的位置数据
- √ 利用AIS数据辅助修正北斗位置信息
- √ 根据匹配出来的AIS ID追踪渔船的历史作业记录

数字中国建设峰会
DIGITAL CHINA SUMMIT

DCIC
数字中国创新大赛
DIGITAL CHINA INNOVATION CONTEST

02

可视化平台

数字中国建设峰会 DIGITAL CHINA SUMMIT

DCIC 数字中国创新大赛 DIGITAL CHINA INNOVATION CONTEST

系统架构

算法平台

Flink　Spark　TensorFlow

实时　离线

- 渔船作业类型分类
- AIS渔船匹配
- 渔船作弊风控识别
- 渔船航线规划
- 海域安全系数画像
- 轨迹缺失填充

特征平台

上层宽表　定时调度

Hive 统计特征

Spark Embedding特征

Extract-Transform-Load

分布式存储

Hadoop　MongoDB　Redis

ZooKeeper　HBase　Kafka

可视化平台

ECHARTS　spring boot

海域监管

- 高频捕捞海域监控
- 船舶调度管理
- 船舶监管安排
- 船舶安全提醒

渔船行为画像

- 停泊地点分布
- 船舶位置缺失
- 群体作业行为
- AIS渔船匹配

阿里云

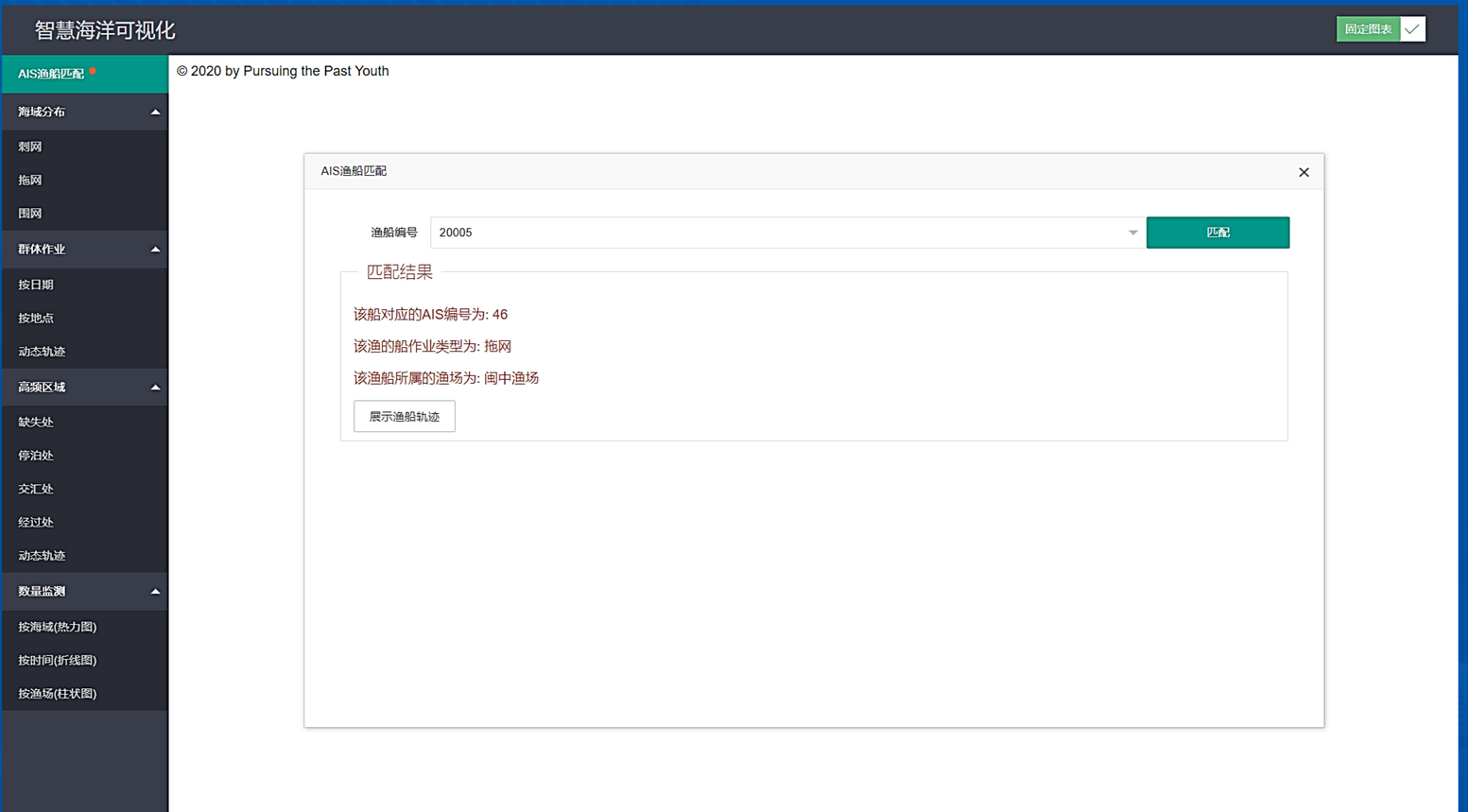
数字中国建设峰会
DIGITAL CHINA SUMMIT

DCIC
数字中国创新大赛
DIGITAL CHINA INNOVATION CONTEST

03

船舶行为画像

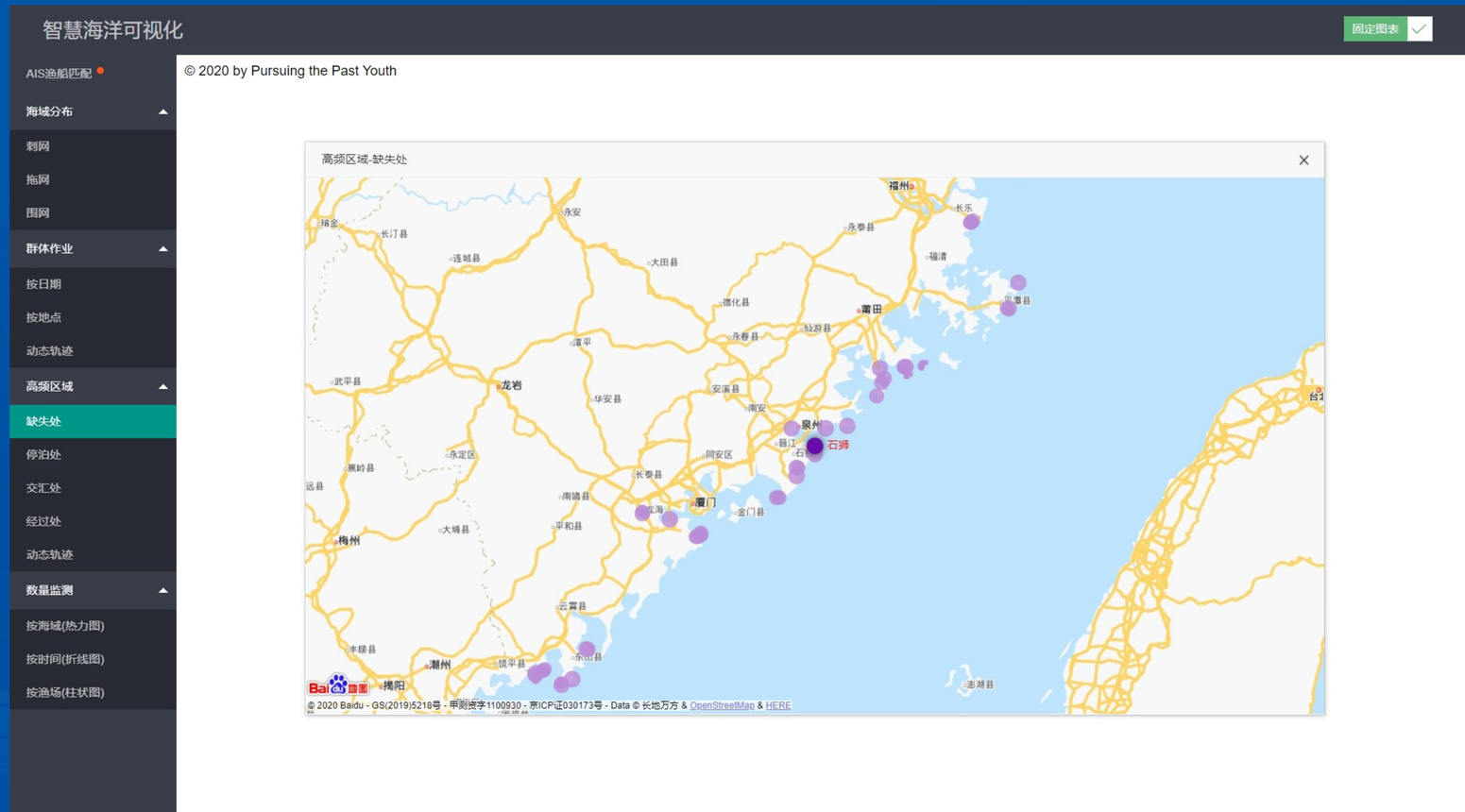
船舶行为画像——AIS渔船匹配



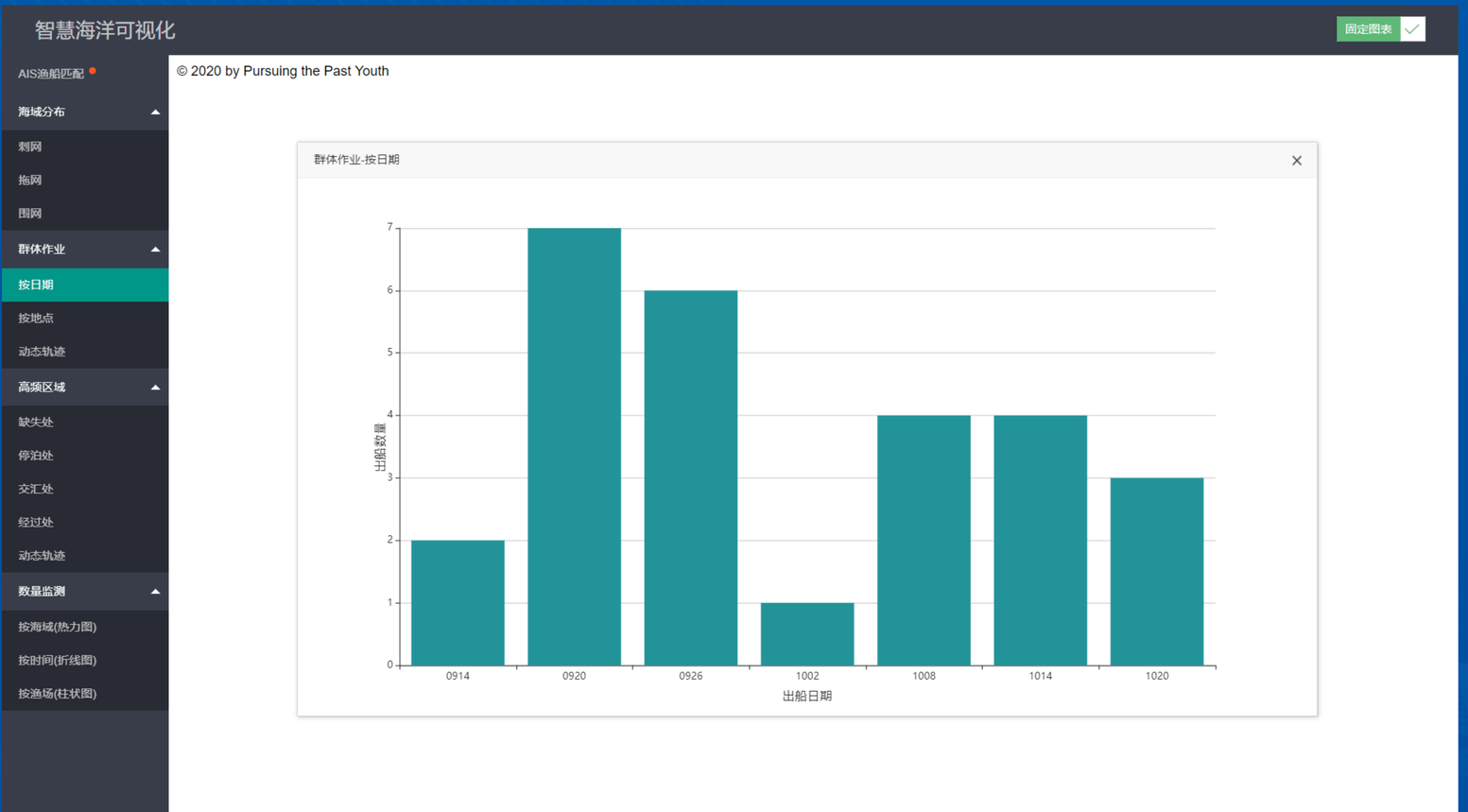
在我们的可视化平台，输入渔船号即可自动匹配对应的AIS号，并通过算法阶段的模型来判断作业类型。还可以查看所在的渔场名称以及在北斗和AIS中记录到活动轨迹。

船舶行为画像——船舶位置缺失

由于位置缺失的现象会给相关渔船的跟踪和监管带来困难，所以需要通过数据进行分析。在我们的可视化平台，可查看易发生位置缺失的区域分布情况。此处可以改提高监管力度，呼吁船员经过该区域时注意安全等。



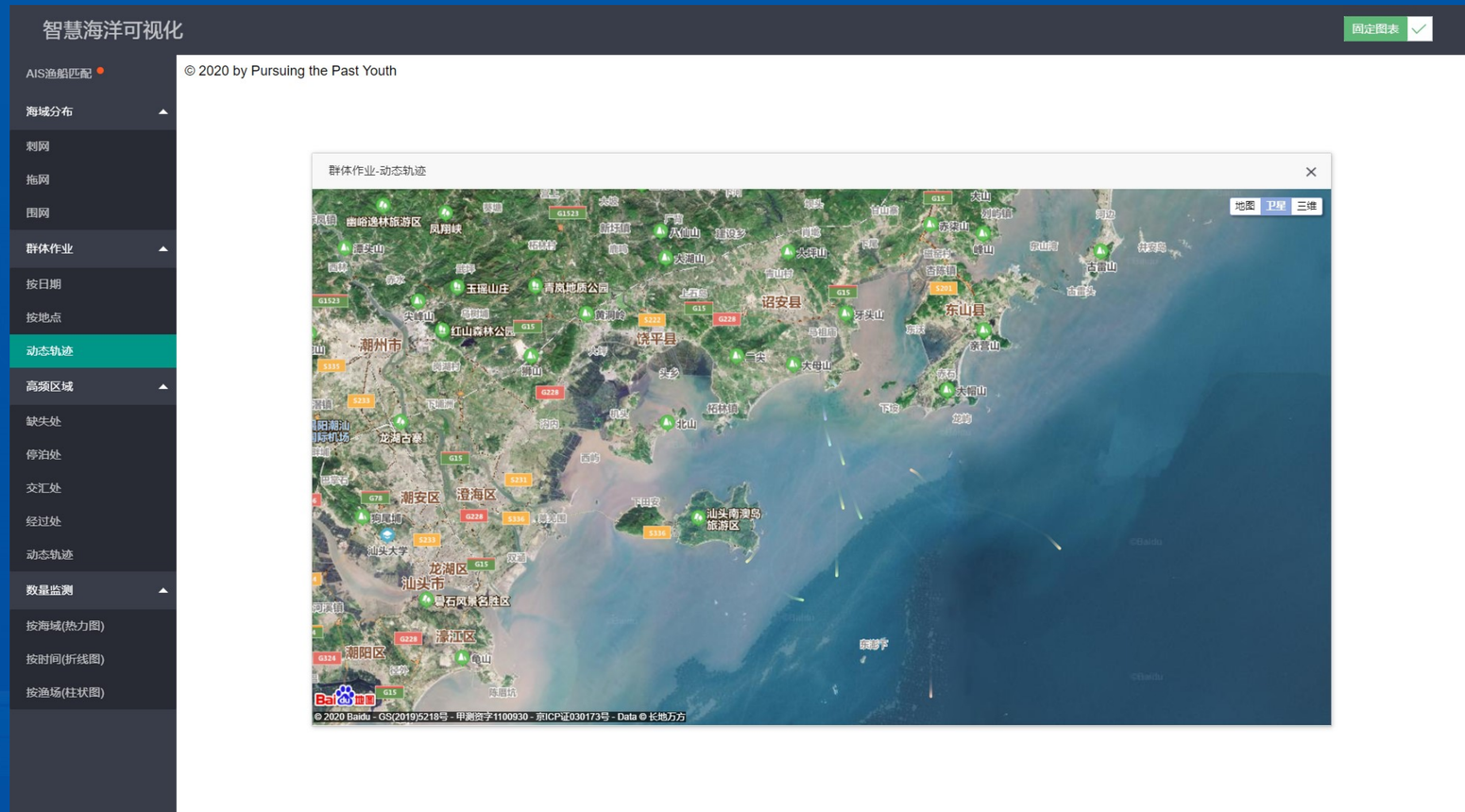
船舶行为画像——群体作业行为



我们认为，群体作业需满足以下条件：在同一时段同一地点出海、以相同方式作业、完成后大多返回出发地、有稳定的持续时间和间隔周期。在我们的可视化平台，可按日期和地点查看群体作业行为的规模。

船舶行为画像——群体作业行为

在我们的可视化平台，还可以查看各批次群体作业渔船的船舶轨迹动态图。



数字中国建设峰会
DIGITAL CHINA SUMMIT

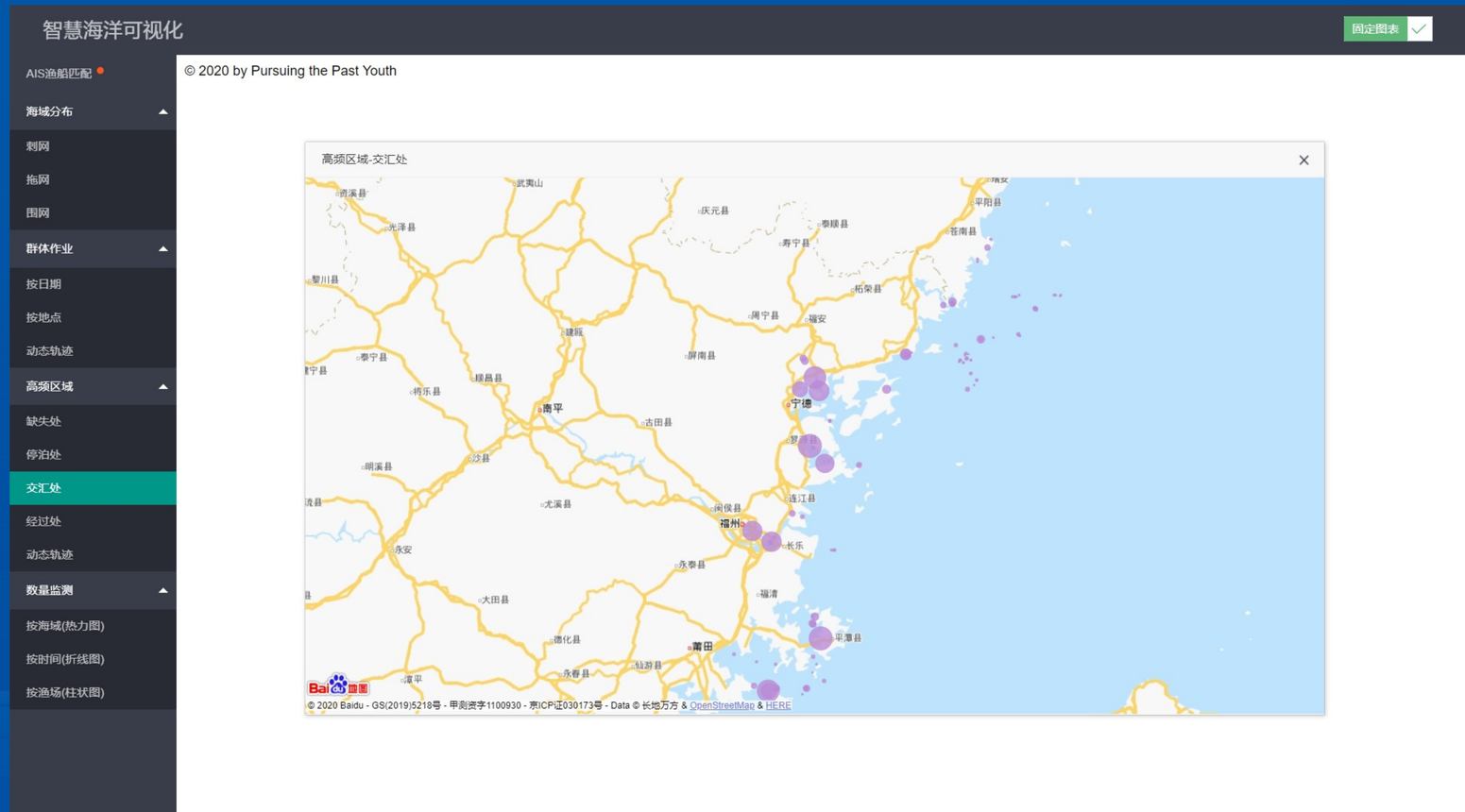
DCIC
数字中国创新大赛
DIGITAL CHINA INNOVATION CONTEST

04

海域监管

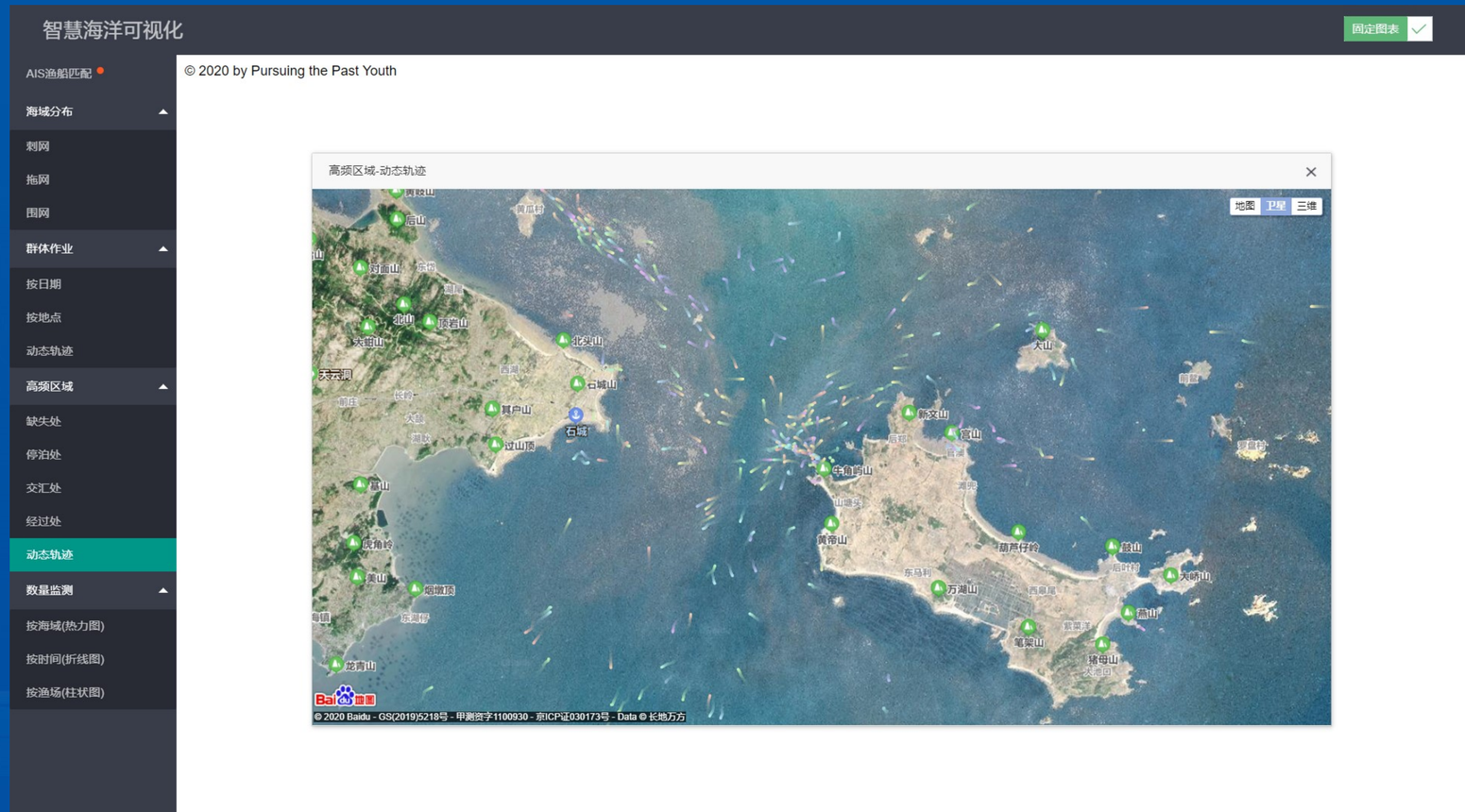
海域监管——船舶安全提醒

在航线交汇多的地方，一般容易发生事故。在我们的可视化平台，可查看高频交汇地点的区域分布情况。然后提醒船员注意安全，减少碰撞等事故发生并帮助监管人员有效的进行海上交通调度。

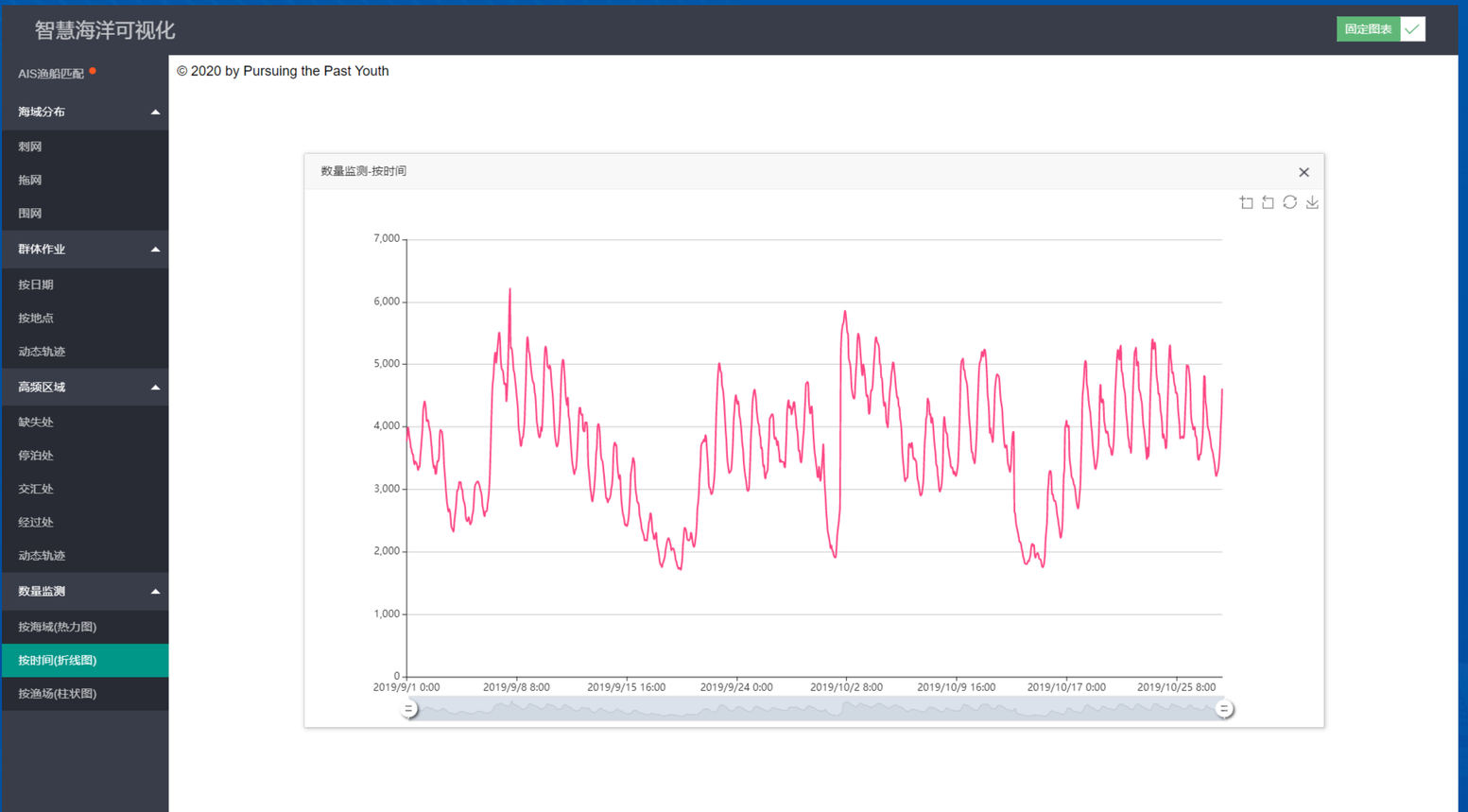


海域监管——船舶调度管理

在我们的可视化平台，还可以查看某一时段、某一海域的船舶轨迹动态图。



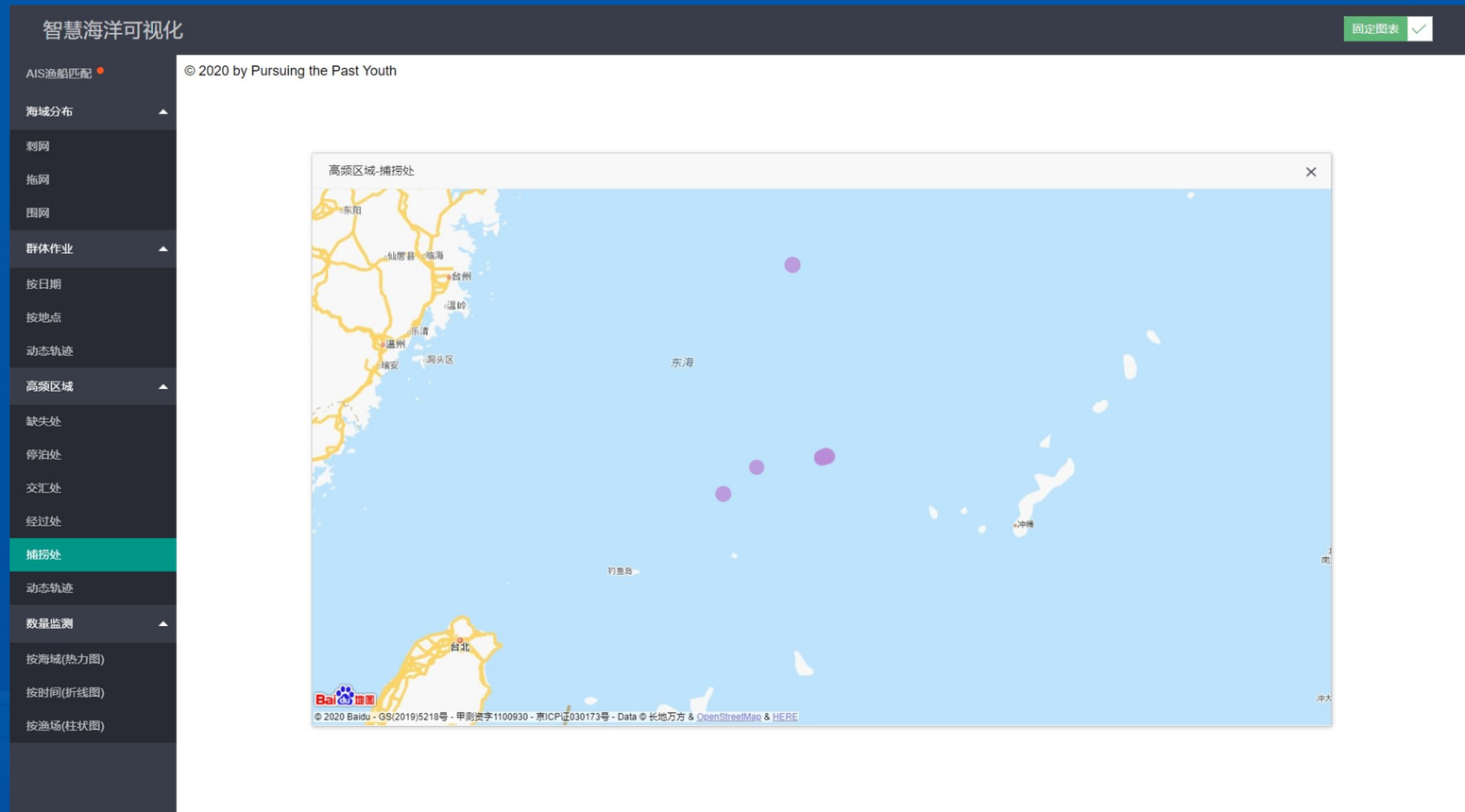
海域监管——船舶监管安排



渔船的出海时间具有很强的周期性。在我们的可视化平台，可查看不同时间段监测到的船舶数量。参考这部分信息可掌握渔船出海的规律，动态地分配工作人员进行海上作业监管。

海域监管——高频捕捞海域监控

船速小的时候，收网的可能性较大。在我们的可视化平台，可查看高频捕捞地点的区域分布情况。通过这项监测，限制渔船的捕捞频率，可以防止鱼类资源枯竭。



数字中国建设峰会
DIGITAL CHINA SUMMIT

DCIC
数字中国创新大赛
DIGITAL CHINA INNOVATION CONTEST

THANK YOU