

주요 6개 항구 기준

AIS CHOI ET 기반 Waiting Area 추출

DIMA4 5조

김묘경, 곽현수, 김정수, 문시아, 유지민

팀원 소개

1



김정수
메인 코더
자료조사



곽현수
서브 코더
자료조사
발표



김묘경
조장
메인 코더
자료조사



문시아
자료조사
발표



유지민
서브 코더
자료조사



목차

01

프로젝트 개요

주제 설명

02

데이터 전처리

전처리 과정

사용 데이터

대기 선박 조건 지정

03

군집화

Waiting Area 지정

모델 선정

군집별 특성 분석

04

결론

의의와 한계

개선 사항, 기대 효과

Q&A

01 프로젝트 개요

주제와 목적, 기대 효과

주제

AIS 데이터 기반 대기 구역 추출

목적

선박 대기 구역 분석을 통한 환경 조사

기대 효과

안전한 대기란 무엇인지 파악

02 데이터 전처리

데이터 처리 방법과 순서, 대기 선박 조건 설정

AIS (Automatic Identification System)

- = 선박의 항해 안전 및 보안 강화를 위해 선박의 위치, 속력, 진로, 선박 식별 ID 등의 정보를 무선 통신(VHF 주파수)를 통해 자동 송수신 할 수 있는 항해 장비
 - 국제해사기구(IMO) 선박안전법에 따르면, 총 300톤을 초과하는 선박은 AIS 설치가 의무, AIS 정보는 이동 속도에 따라 2초-3분 간격으로 지상 수신기 또는 위성 수신기로 전송



AIS 데이터 = 선박의 이동 정보

사용 데이터 - 1) AIS 신호 데이터

- 범위: 3개국 6개항 (인천, 부산(신), 상해, 홍콩, 싱가폴)
- 기간: 2022.09 ~ 2023.05



ais_for_waiting_area.csv

FROM_PORT_CD	TO_PORT_CD	VSL_ID	VSL_TIMESTAMP	VSL_SPEED	VSL_HEADING
출발항	도착항	선박 ID (IMO)	AIS 데이터 수신 시간	선박 속도	선수 각도
AEJEA	CNSHA	27d8a671-a29d-355b-a1b9-07582c37e466	2023-03-09 23:50:02	12.0	321.0
AEJEA	CNSHA	27d8a671-a29d-355b-a1b9-07582c37e466	2023-03-09-23:55:06	12.7	313.0
AEJEA	CNSHA	27d8a671-a29d-355b-a1b9-07582c37e466	2023-03-10 00:00:03	12.8	313.0
LAT	LON	VSL_DRAFT	VSL_COURSE	NAVIGATIONAL_STATUS	
위도	경도	선박이 바다에 잠긴 깊이	선박 항로	항해 상태	
25.033032	55.034568	999.0	326.3	Under way using engine	
25.045750	55.021503	8.1	313.6	Under way using engine	
25.058315	55.007398	999.0	313.7	Under way using engine	

사용 데이터 - 2) 선박 정보 데이터

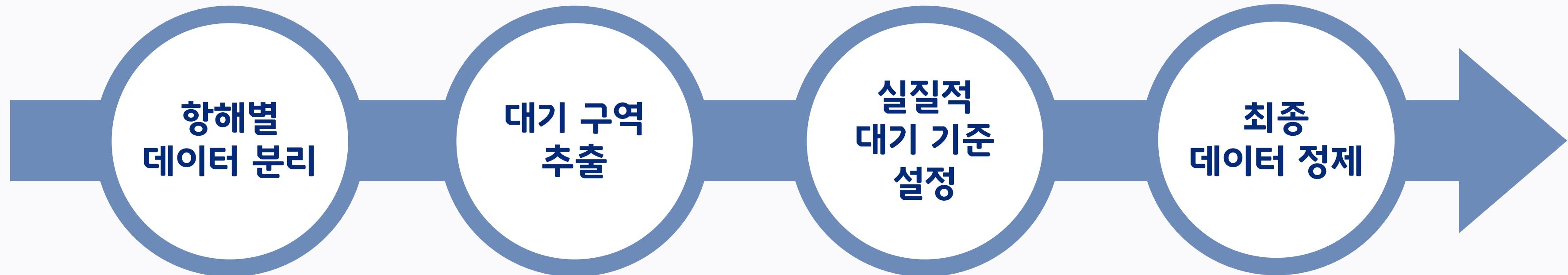
- 선박 개수: 1,334개



vsl_info.csv

VSL_ID	VSL_NAME	VSL_MMSI	VSL IMO	CALL_SIGN	SHIP_TYPE	VSL_LENGTH
선박 ID	선박명	선박 MMSI	선박 IMO 번호	호출 부호	선박 종류	선박 길이
00260274-dbeb-3672-a688-34cf851acc	KOTA PAHLAWAN	563033800	9786712.0	9V5293	CONTAINER	330.0
003756d7-a298-391e-ba9b-d336079996a8	XIN WEN ZHOU	414736000	9523043.0	BPKJ	CONTAINER	255.0
00a3b1cf-589e-39b4-8c26-fc7975e53188MILD	MILD SONATA	477390900	9666857.0	VRNF6	CONTAINER	148.0
VSL_WIDTH	DIMENSION_A	DIMENSION_B	DIMENSION_C	DIMENSION_D	GROSS_TONNAGE	
선박 넓이	선박 치수 A	선박 치수 B	선박 치수 C	선박 치수 D	총톤수	
48.0	107.0	223.0	30.0	18.0	113976.0	
37.0	199.0	56.0	13.0	24.0	47917.0	
23.0	130.0	18.0	8.0	15.0	9994.0	

데이터 전처리 단계



Step 1. AIS 신호 데이터 + 선박 정보 데이터 결합

VSL_ID 기준으로 MERGE

AIS 신호 (1,548,952) + 선박 정보 (1,334) → 1,253,626

Step 2. 선박 정보 없음, VSL_DRAFT 결측치 제거

문제 1. 선박 정보 불분명, 정보 수집 不可

선박 정보 데이터 中 결측치 존재



VSL_MMSI, VSL IMO, CALL_SIGN 기준,
선박 추적 사이트에서 검색 · 채움

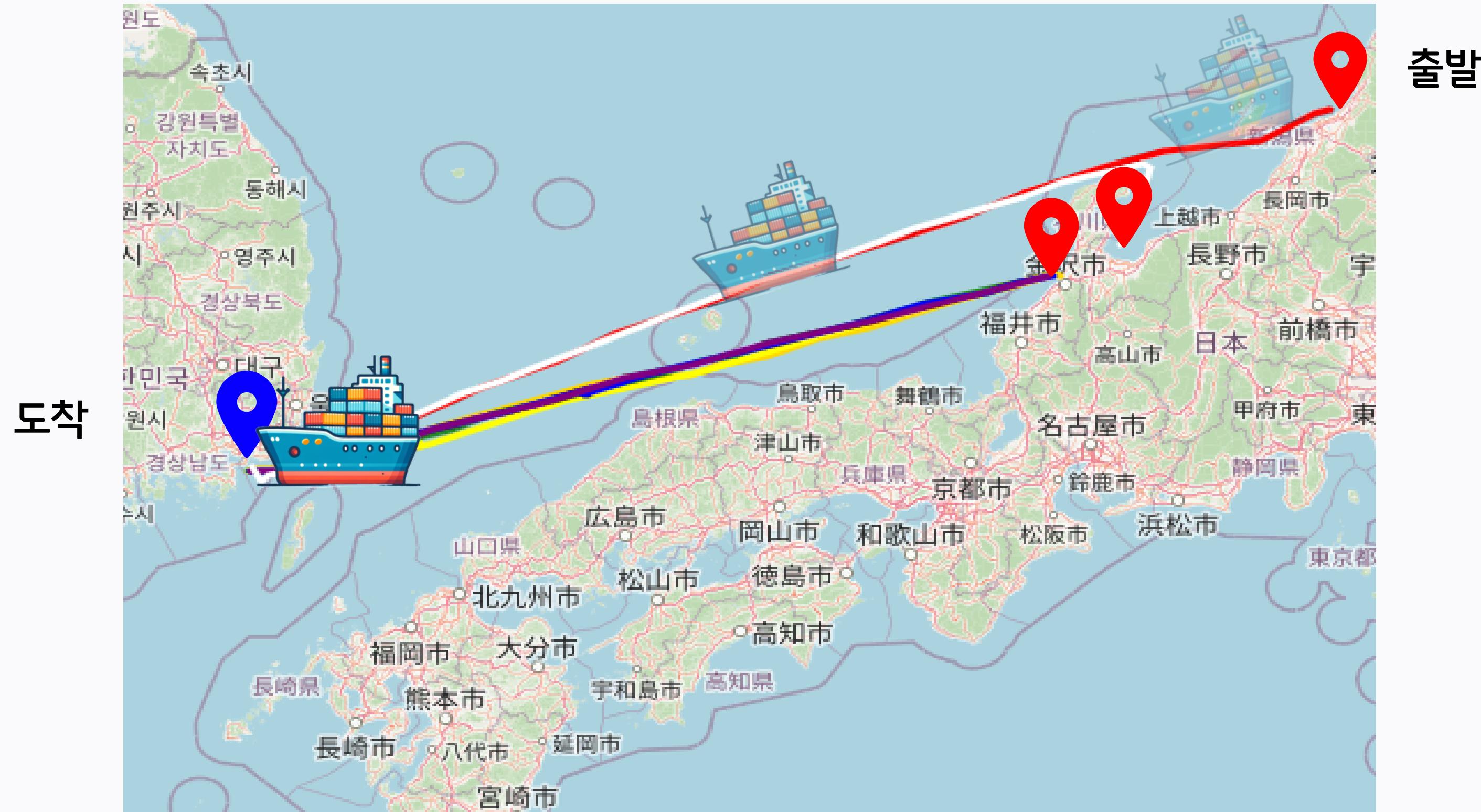
→ 남은 결측치 제거 + 각 선박별로 상이해 유사/평균값 대체 불가능한 결측치 제거

FROM_PORT_CD	TO_PORT_CD	VSL_ID
출발항	도착항	선박 ID (IMO)
AEJEA	CNSHA	27d8a671-a29d-355b-a1b9-07582c37e466
JPKNZ	KRPUS	736779af90f964-a74f-3875-a75b-c06ee10503b4
AEJEA	CNSHA	27d8a671-a29d-355b-a1b9-07582c37e466
JPKNZ	KRPUS	736779af90f964-a74f-3875-a75b-c06ee10503b4
AEJEA	CNSHA	27d8a671-a29d-355b-a1b9-07582c37e466

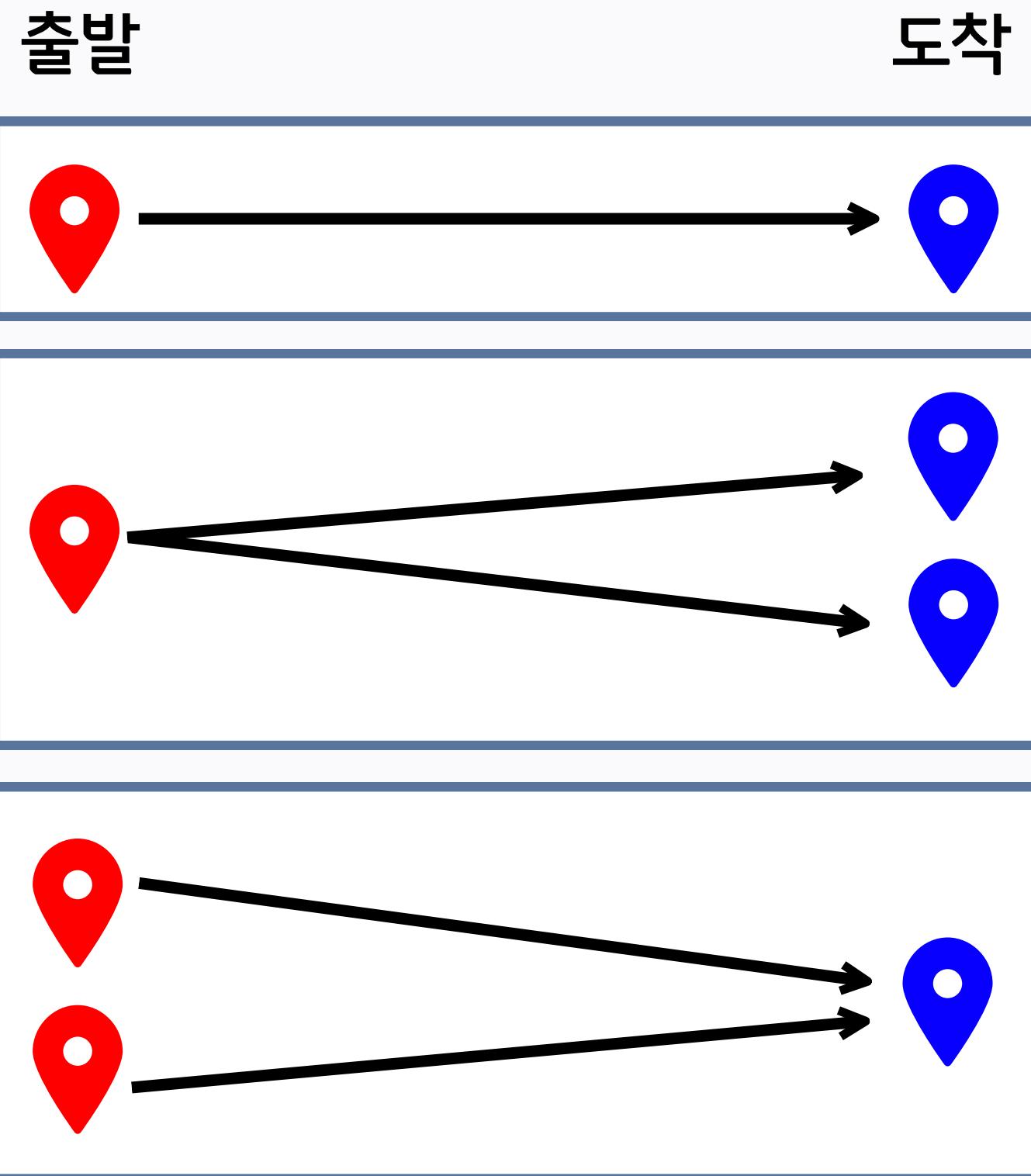
항해 1

항해 2

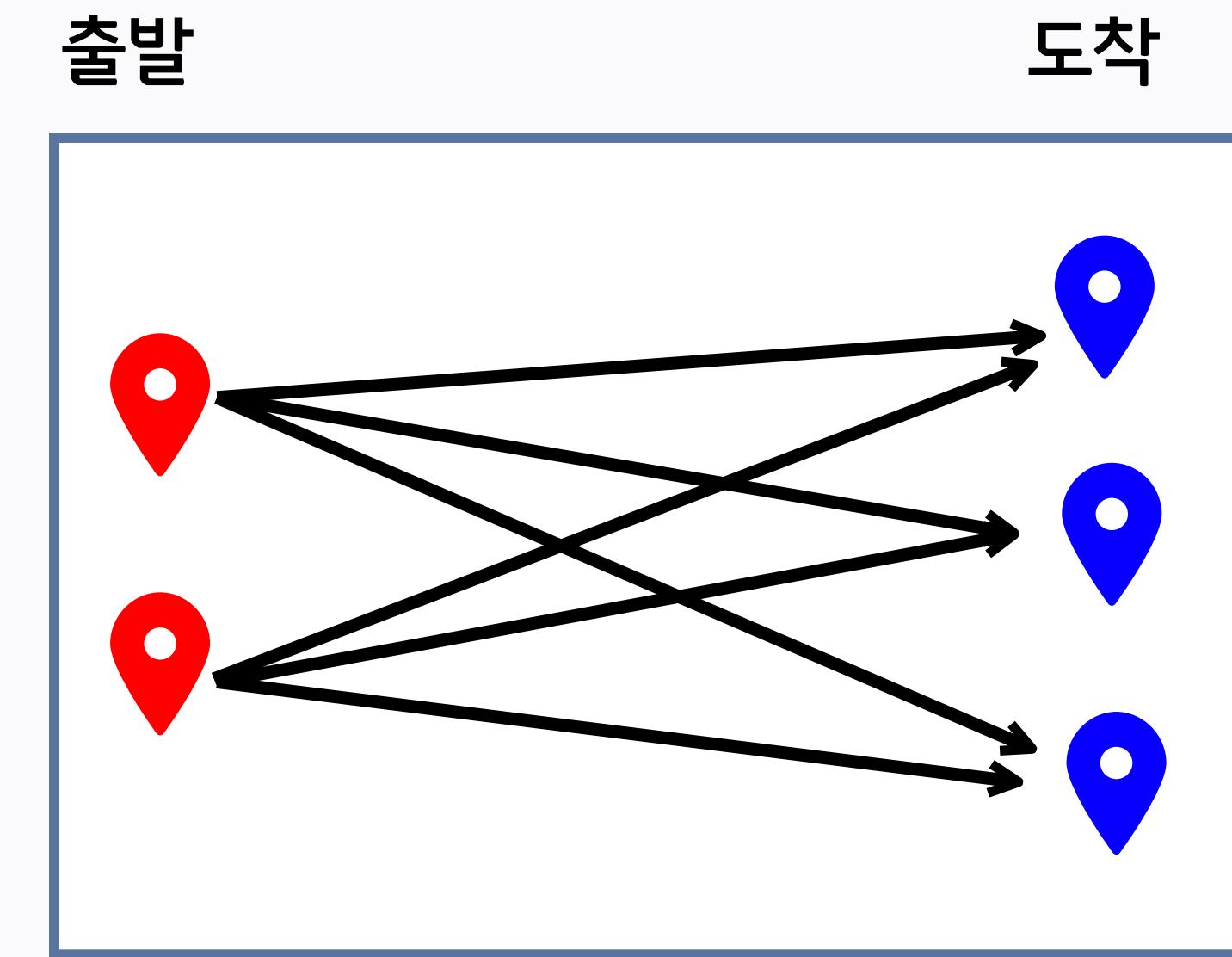
Step 2-1. 특정 선박 1개의 모든 데이터를 지도에 표기



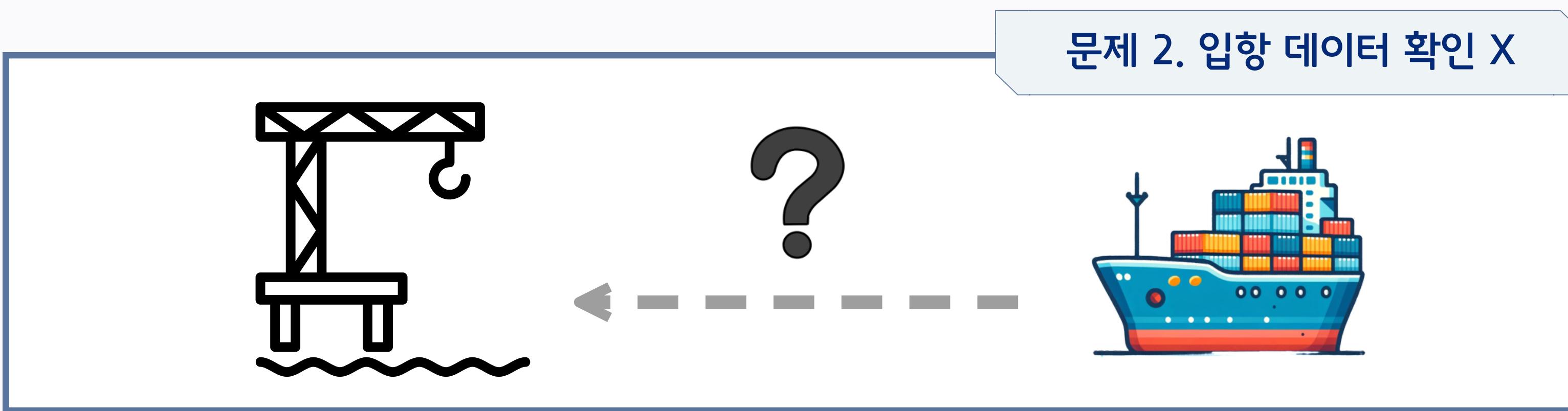
Step 2-1. 특정 선박 1개의 모든 데이터를 지도에 표기



문제 1. 여러 항해가 혼재



Step 2-1. 특정 선박 1개의 모든 데이터를 지도에 표기



→ 각 항해별 분리 필요

Step 3. 항해별 데이터 분리



Step 3. 항해별 데이터 분리



Step 4. 근접 항구 간 항해 제거

IF. 출발항과 도착항이 가까운 경우,

선박이 도착항 근처에 있는지, 출발항 근처인지 파악 불가

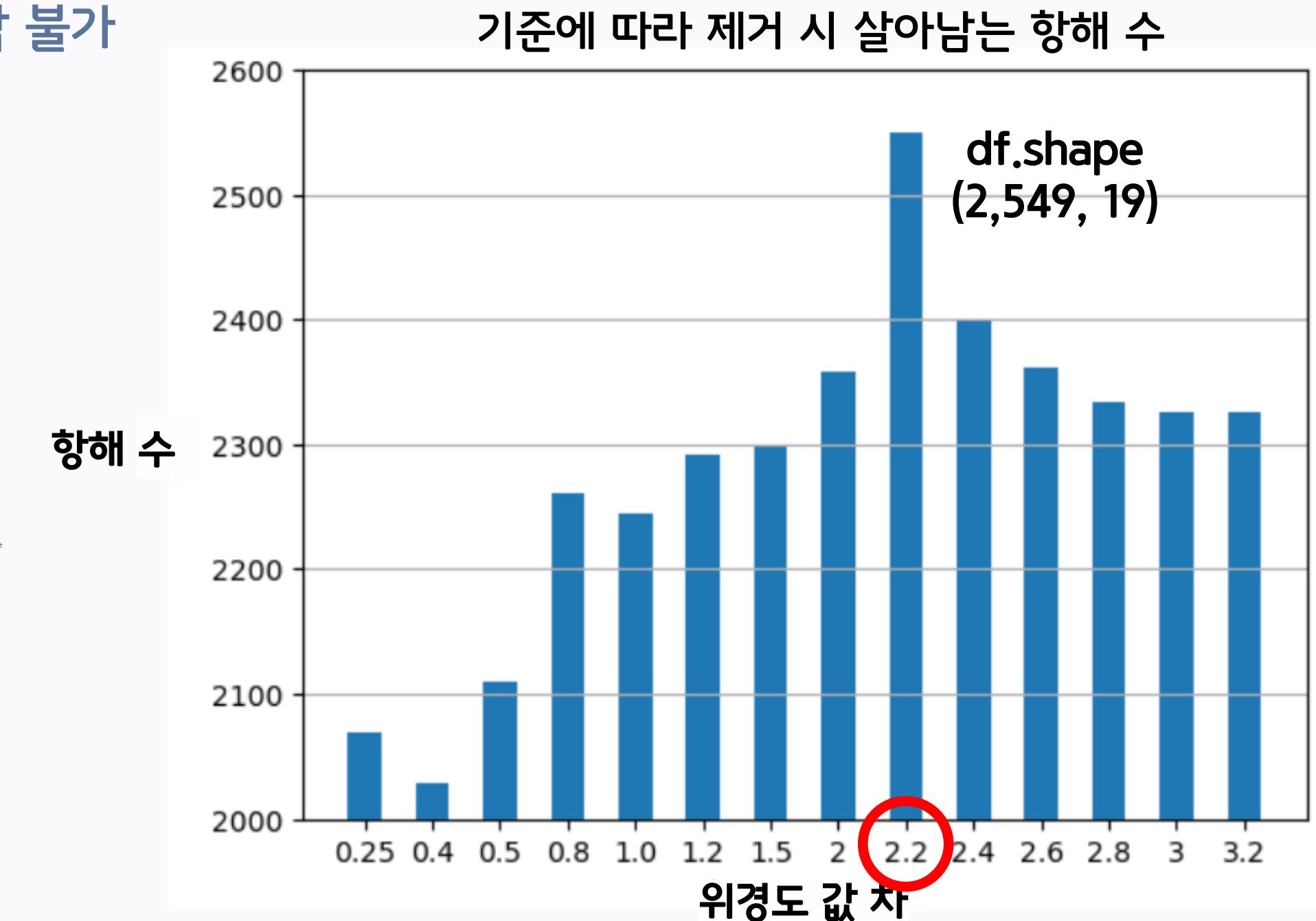
▶ 가까운 거리에 위치한 항해 제거 필요

가까움의 기준

$|도착항 위도 - 출발항 위도| \leq 2.2$

&

$|도착항 경도 - 출발항 경도| \leq 2.2$



Step 5. 신호 누락 제거

- = 앞뒤로 오는 두 신호 사이의 시간 간격 > 도착항별 평균 대기 시간
- 편도 항해로 이루어진 AIS 데이터만으로는 장시간의 신호 부재가
1) 대기 상태인지, 2) 회항 중인지 알 수 없음

→ 제거

Step 6. 선박의 대기 상태를 판단하는 기준 설정



대기 중 속도

3 노트 (5.556 km/h) 이하 유지



대기 지점의 범위

도착항의 위경도 $\pm 121\text{km}$ 이내

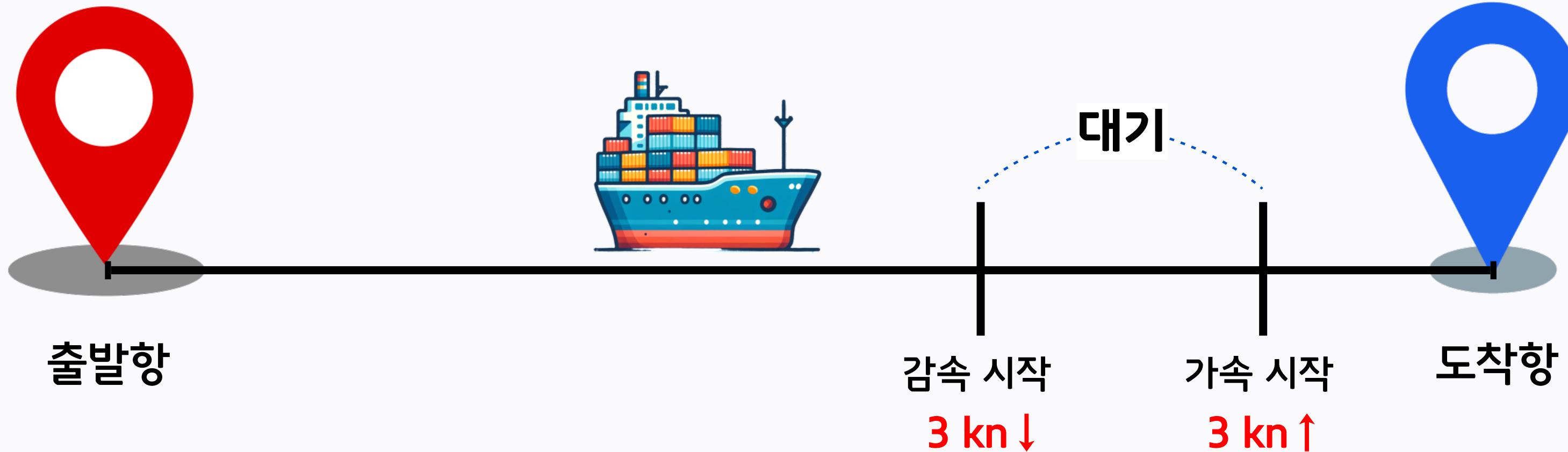


실질적 대기 선별

1번의 항해에서 대기가 여러 번 발생한다면,
가장 오래 대기한 위치와 시간 선택

→ 선박 대기 사유를 알 수 없기 때문

Step 6. 선박의 대기 상태를 판단하는 기준 설립



$$\text{대기 시간} = \text{가속 시작 시점} - \text{감속 시점의 차}$$

→ 단 1회의 대기 신호만 존재할 경우, 0분인 문제 有

소지한 AIS 데이터의 평균 신호 간격(13분) 으로 대체

Step 7. 군집화를 위한 데이터 처리

대기별 중심 좌표 산출

선박의 대기 = 낯을 내린 정박(묘박) + 표류 상태

- 선박의 위치가 조금씩 바뀔 수 있어 AIS 신호 상 좌표 변동 多
- 대기당 1개의 좌표 추출 필요 = 중심 좌표

1) 신호 1개 = 그 신호의 좌표값

2) 신호 2개 = 두 좌표값의 평균

3) 신호 3개 이상 = KMeans Clustering 사용

Step 8. 각 항해별 기상 데이터 수집 및 정리

「 특성 분석 위한 데이터 수집 → 좌표별 기상 데이터 (수심, 파고, 풍속, 해류) 」

- 수집 방법: 웹스크래핑 (전세계 기상 데이터 제공 사이트 earth.nullschool.net)

↳ 선박의 모든 AIS 신호에 대해 수집 진행할 경우, 예상 경과 시간 = 7日 이상



최초 대기 시점, 최후 대기 시점에 대한 데이터만을 수집

전처리 결과

정제 후 데이터 수



1,053 개

선박 평균 정보



총톤수 - 4238.01 ton

드래프트 - 9.32 m

대기 시간 - 7H 42M

항구별 평균 대기 시간



항구	시간	항구	시간
부산	7H 28M	싱가폴	4H 21M
부산신항	8H 18M	인천	5H
상해	21H 11M	홍콩	6H 35M

03 군집화

모델 선정 및 군집 특성 분석

1. 가설 설정



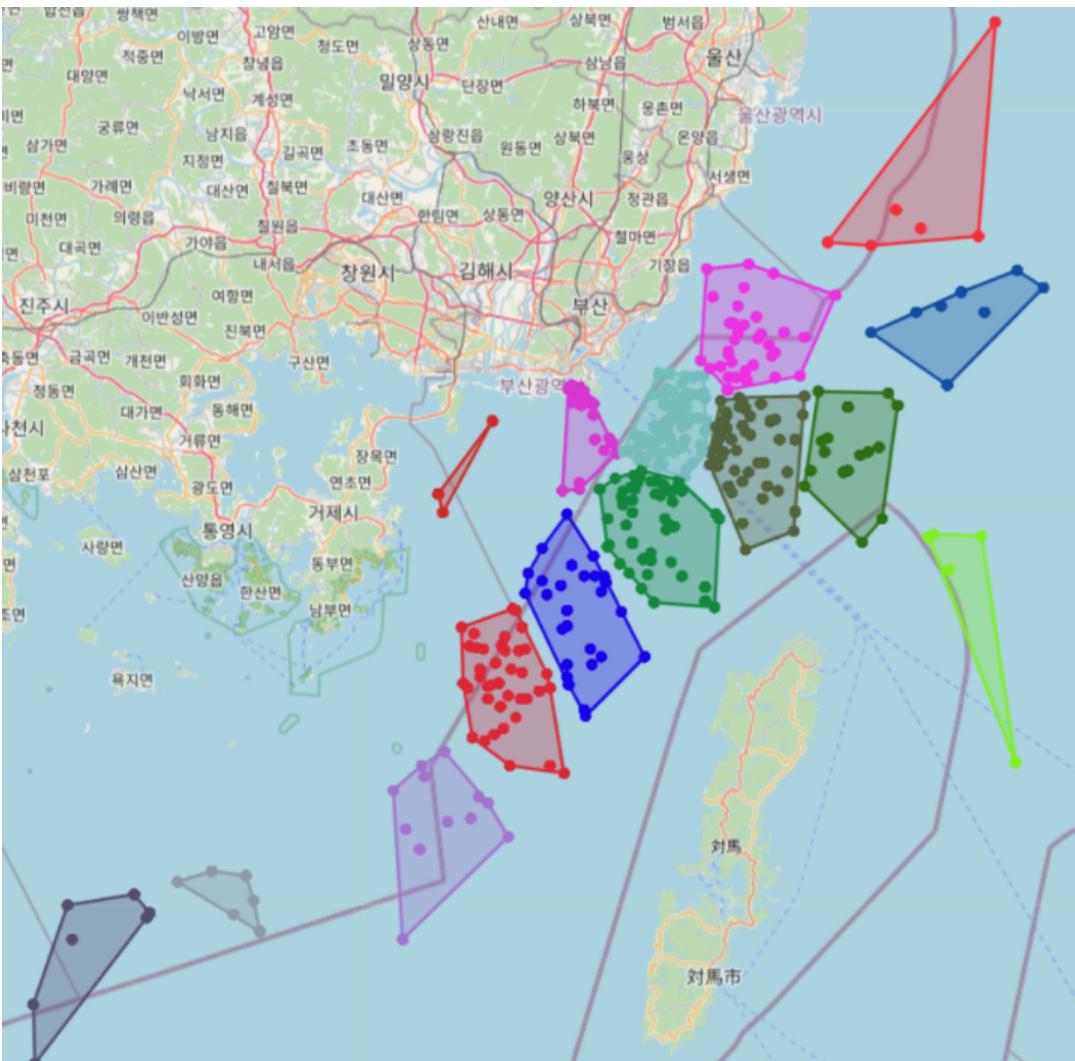
군집별로 환경, 선박의 특성 차이가 있을 것이다

2. 모델 선정

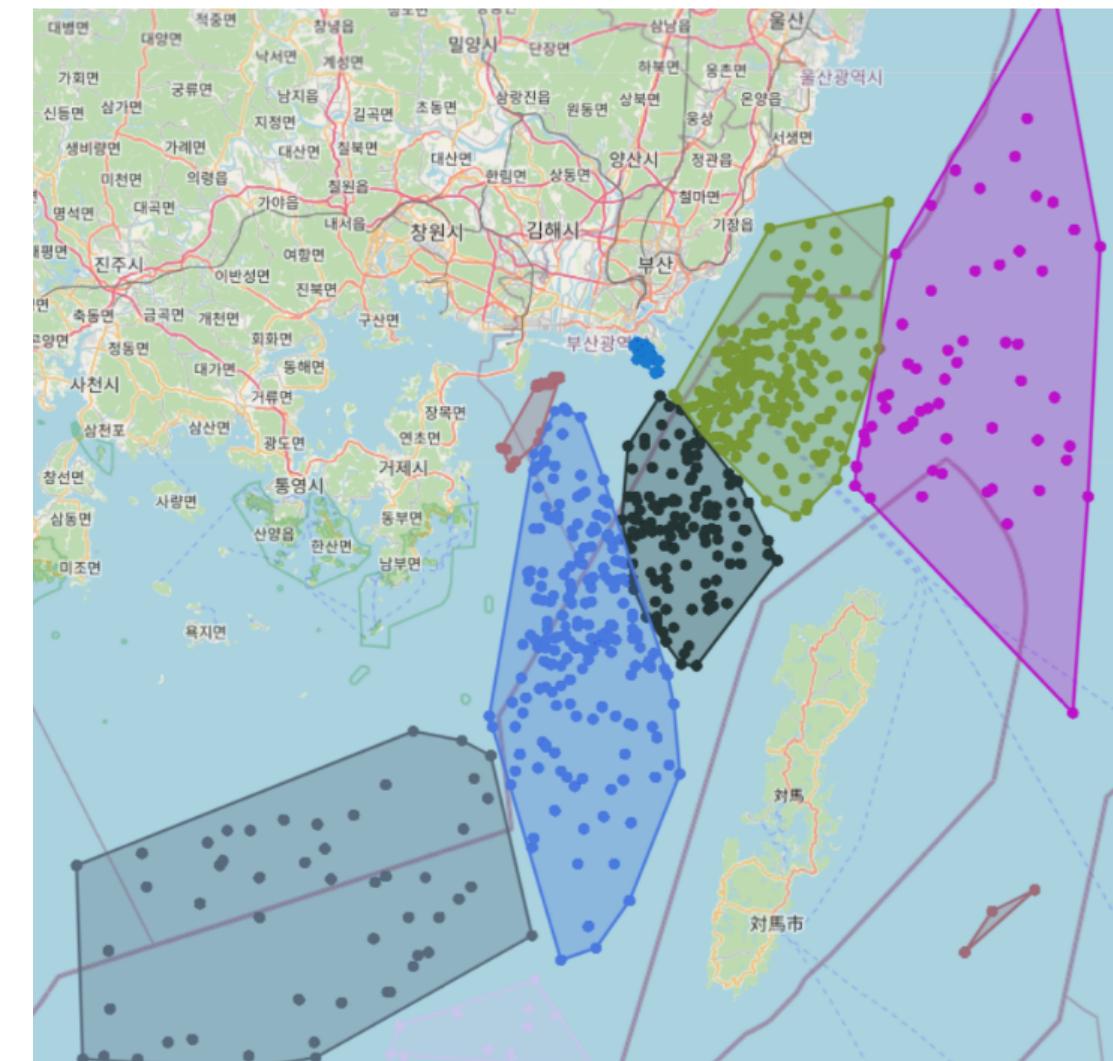
	K-Means	GMM	Mean-Shift	DBSCAN
파라미터	- 군집 수 - 초기화 방식 - 최대 반복 수	- 군집 수 - 반복 횟수 - 공분산 유형	- 대역폭 - 최대 반복 횟수	- eps (최대 반경) - min_samples
특징	- 구형 군집에 적합	- 타원형 군집에 적합 - 확률 기반	- 임의의 군집 모양	- 밀도 기반 군집화 - 노이즈 처리 용이
적합성	- 밀도 및 거리 기반 군집화가 아님 - 노이즈 감지 어려움	- 노이즈 감지 어려움		✓

2. 모델 선정

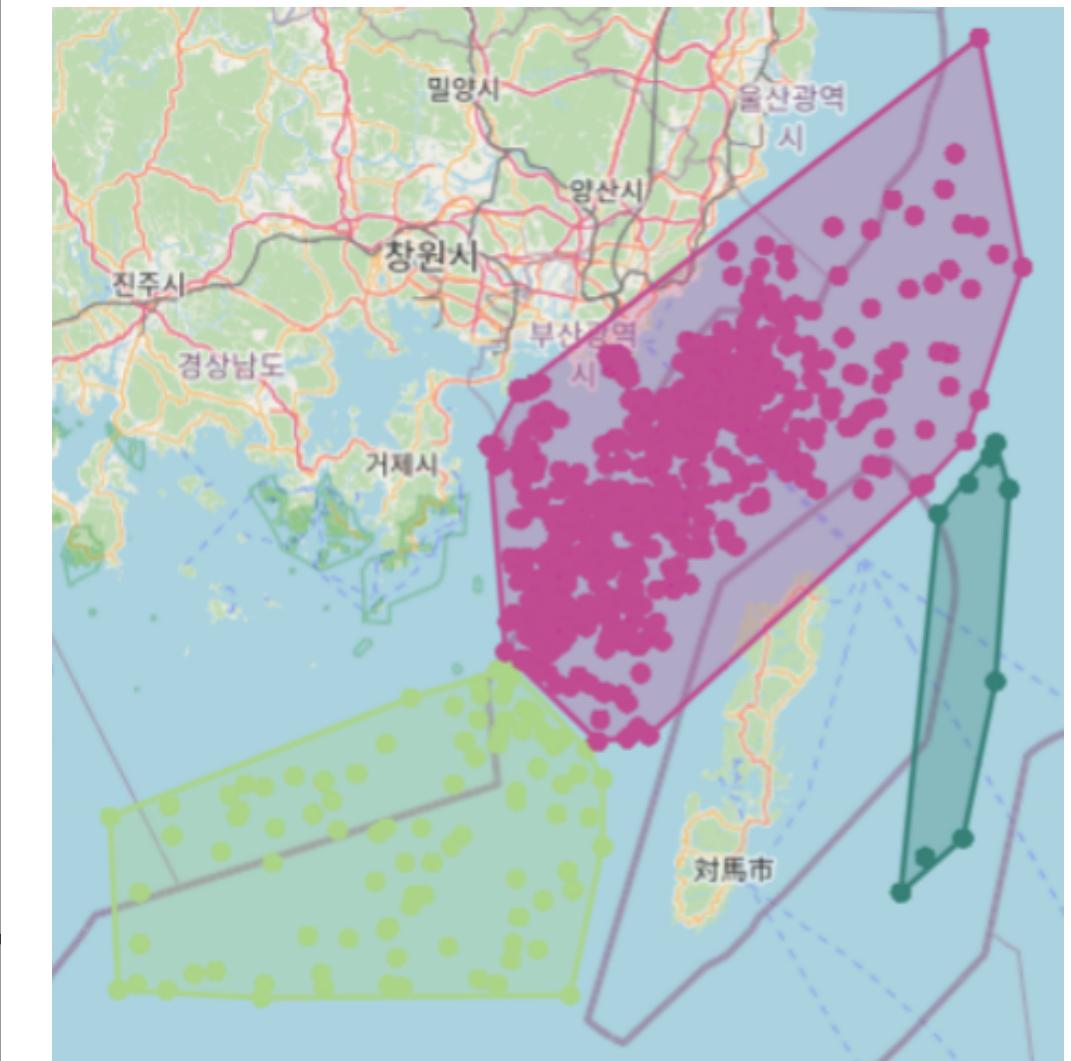
K-Means



GMM



Mean-Shift

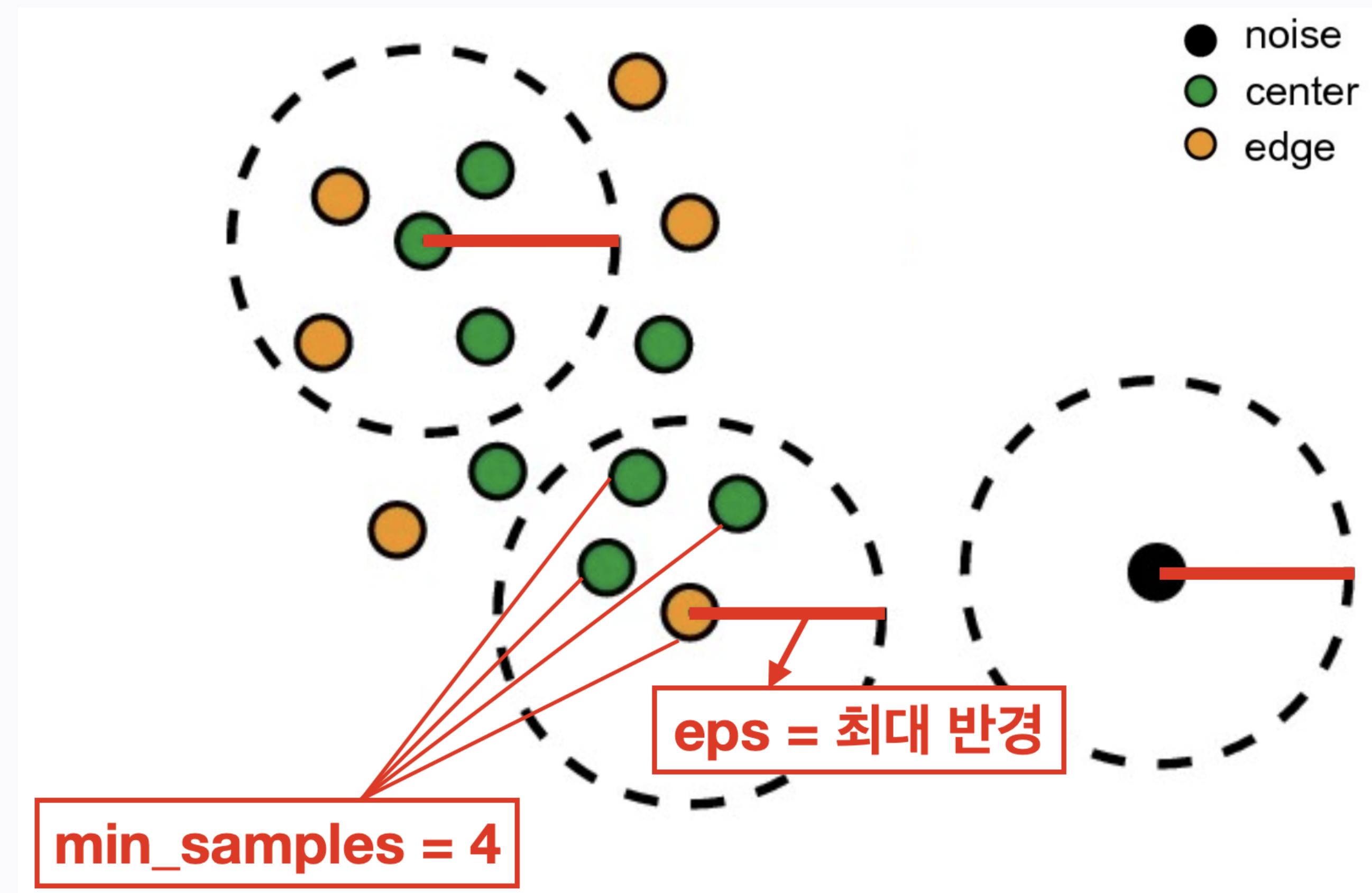


이상치가 포함된 군집

2. 모델 선정

	K-Means	GMM	Mean-Shift	DBSCAN
파라미터	- 군집 수 - 초기화 방식 - 최대 반복 수	- 군집 수 - 반복 횟수 - 공분산 유형	- 대역폭 - 최대 반복 횟수	- eps (최대 반경) - min_samples
특징	- 구형 군집에 적합	- 타원형 군집에 적합 - 확률 기반	- 임의의 군집 모양	- 밀도 기반 군집화 - 노이즈 처리 용이
적합성	- 밀도 기반 군집화가 아님 - 노이즈 감지 어려움		- 노이즈 감지 어려움	

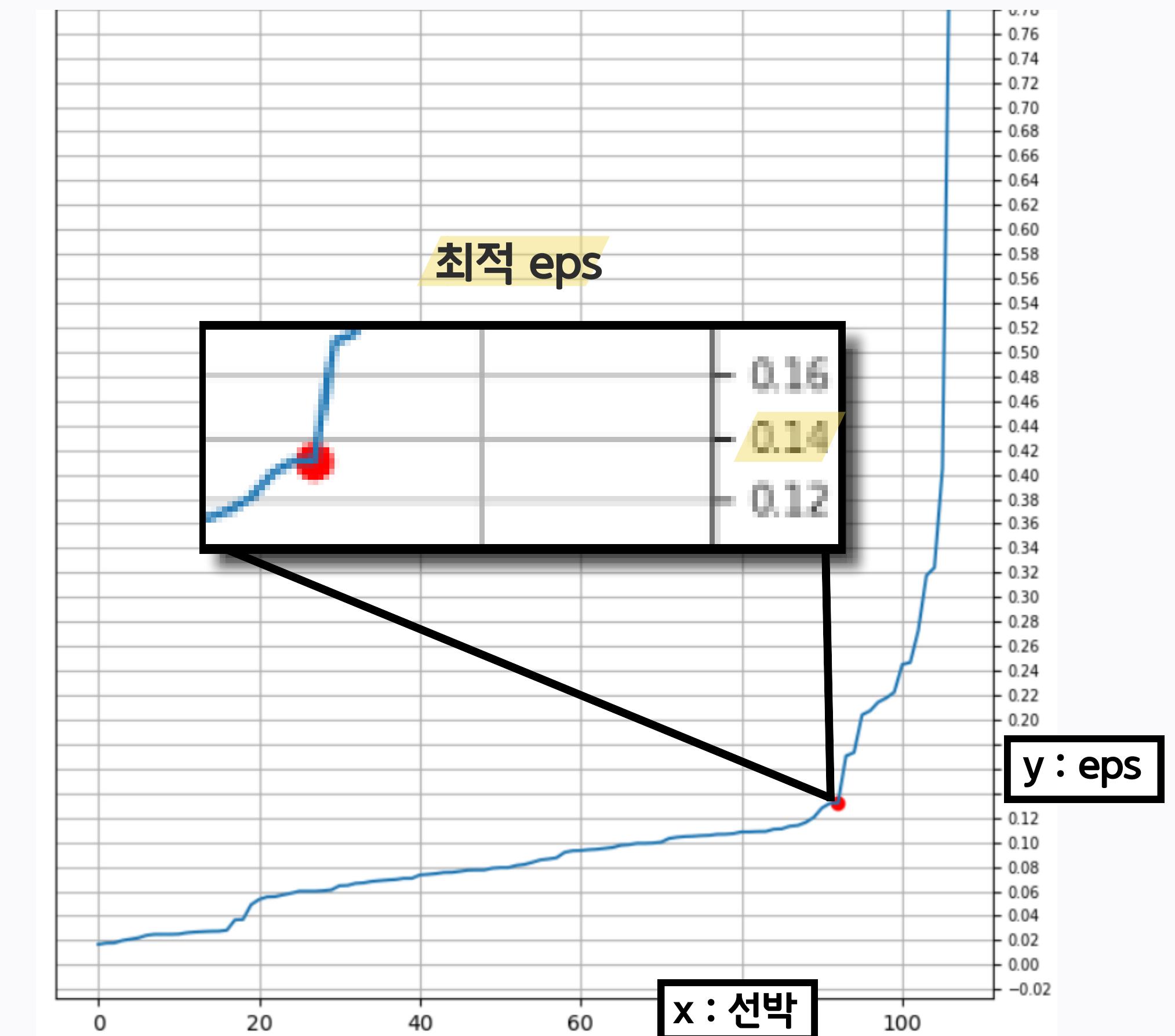
3. DBSCAN



3. DBSCAN

최적의 eps 찾기
: elbow method

elbow point = 기울기가 급격히 변화하는 지점



4. 평가 지표

군집 품질 평가 지표

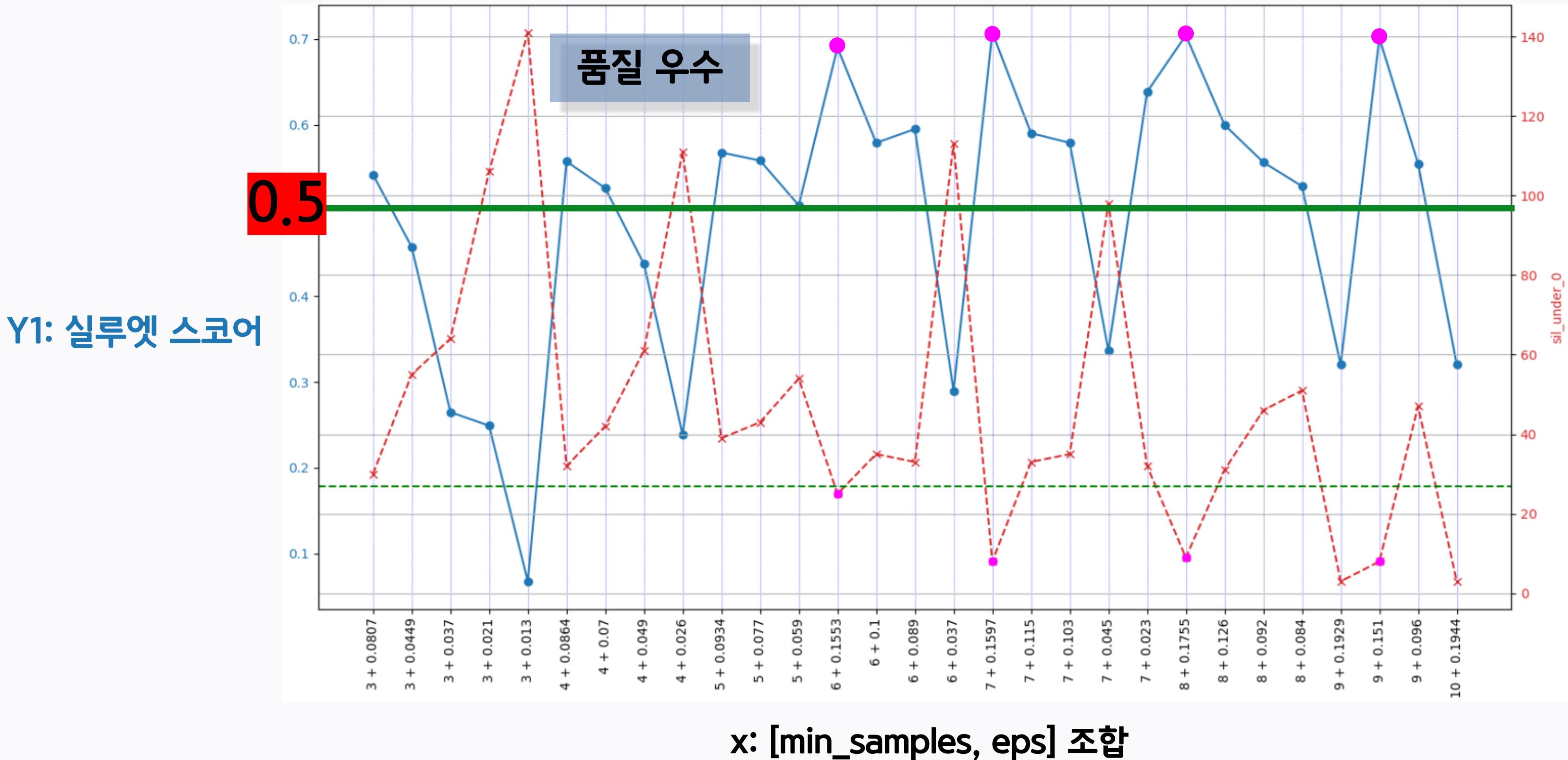
- 실루엣 계수: 각 데이터 포인트가 제대로된 군집에 속해 있는지 판단
- 실루엣 스코어: 실루엣 계수의 평균 = 모델의 군집 정확도

음수일 경우, 부적절

범위 [-1, 1]

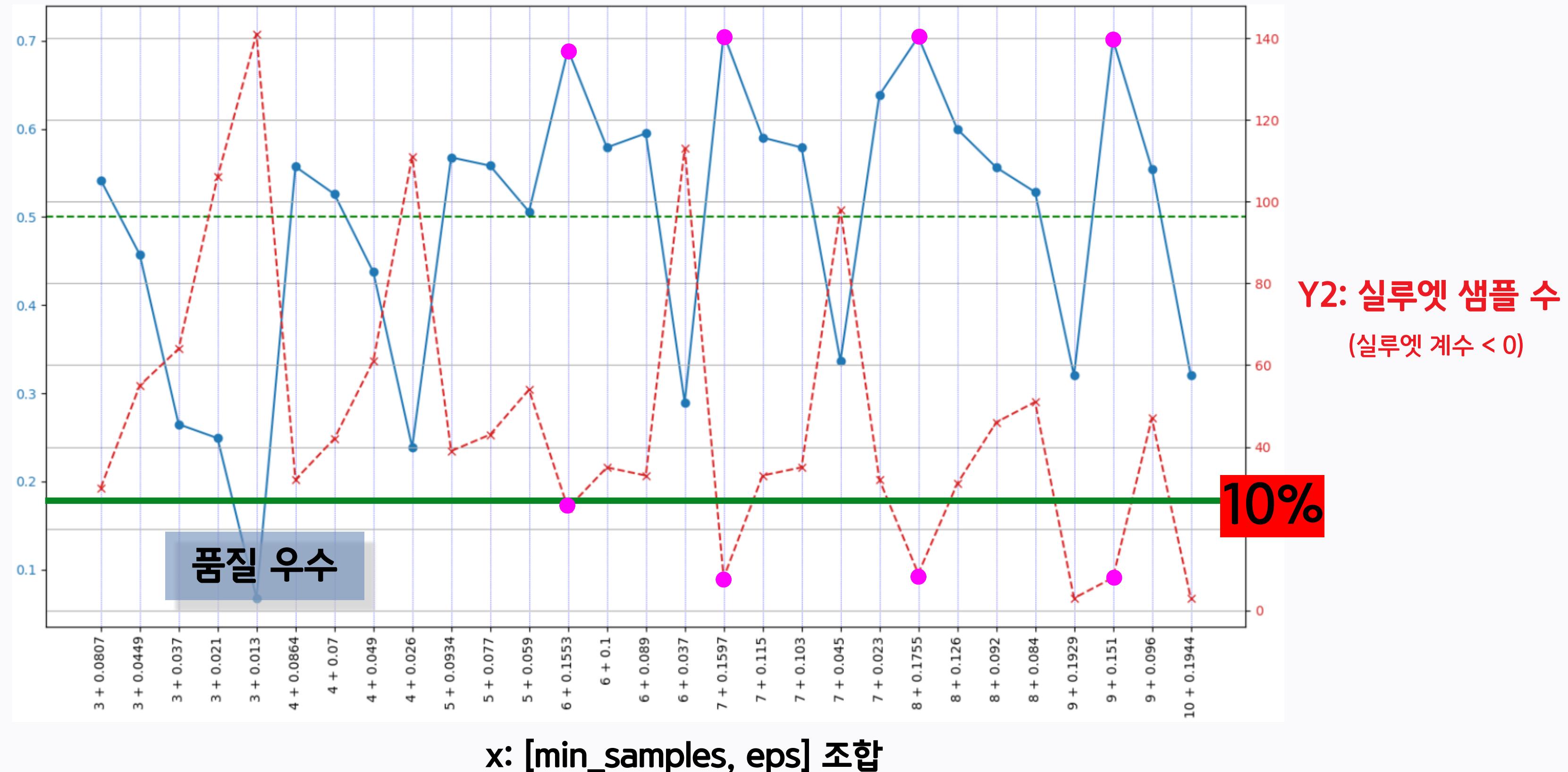
5. 최종 모델 선정 기준

— — — (실루엣 계수 < 0)인 샘플 수
● 실루엣 스코어 ● 우수 모델



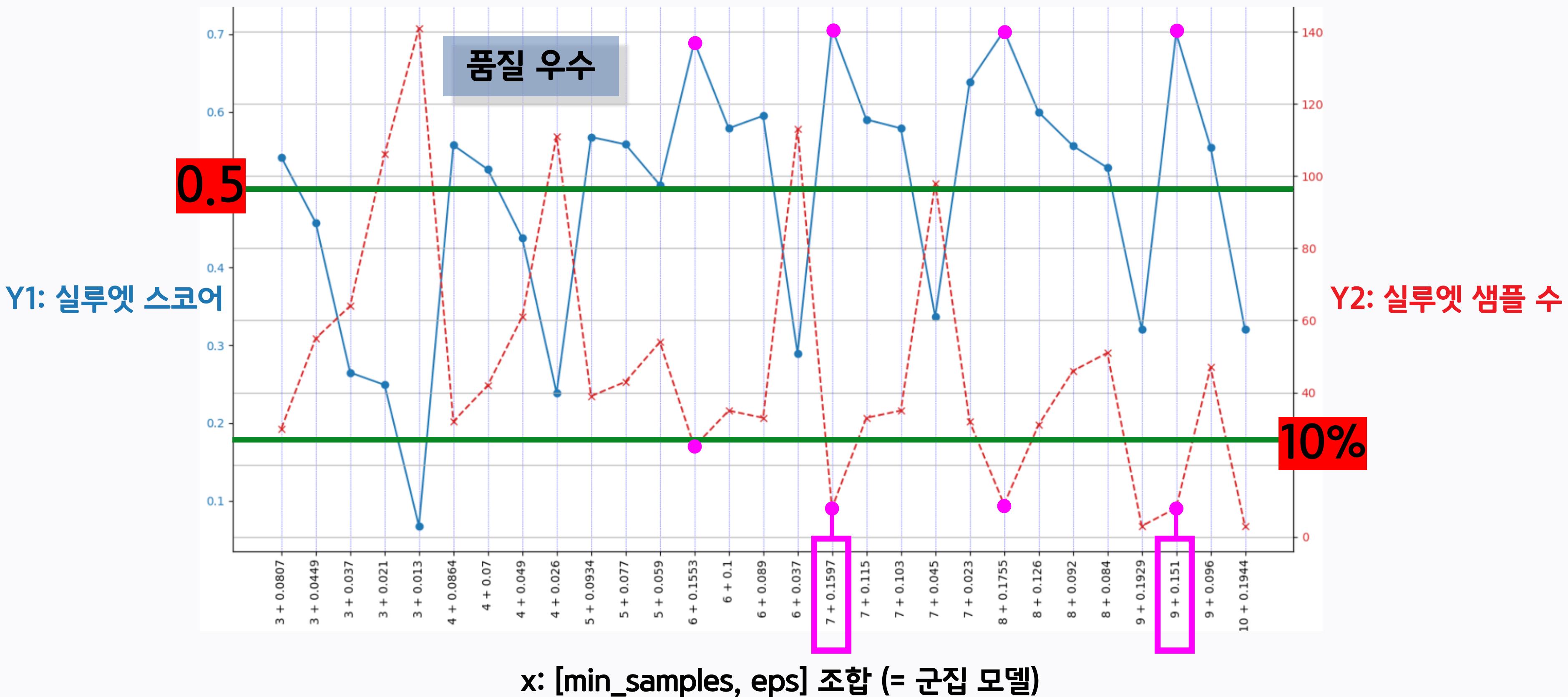
5. 최종 모델 선정 기준

— — — (실루엣 계수 < 0)인 샘플 수
● 실루엣 스코어 ● 우수 모델



5. 최종 모델 선정 기준

(실루엣 계수 < 0)인 샘플 수
실루엣 스코어 ● 우수 모델 □ 최적 모델



6. 군집화 완료

6개 도착항에 대한 모델 각각 생성



Clustered_싱가폴.csv

shape (275, 30)



Clustered_상해.csv

shape (88, 30)



Clustered_홍콩.csv

shape (107, 30)



Clustered_인천.csv

shape (14, 30)



Clustered_부산.csv

shape (260, 30)

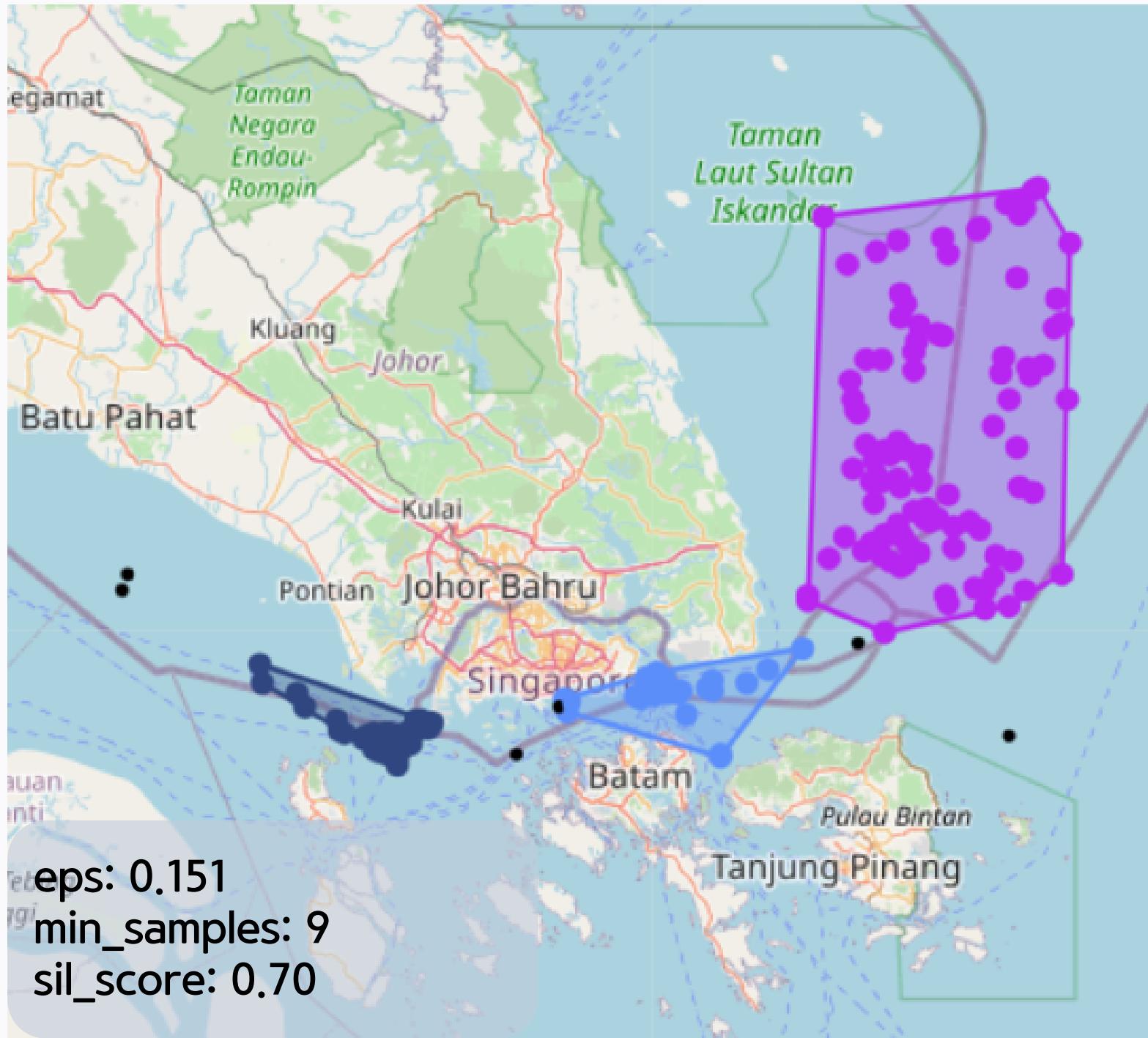


Clustered_부산신항.csv

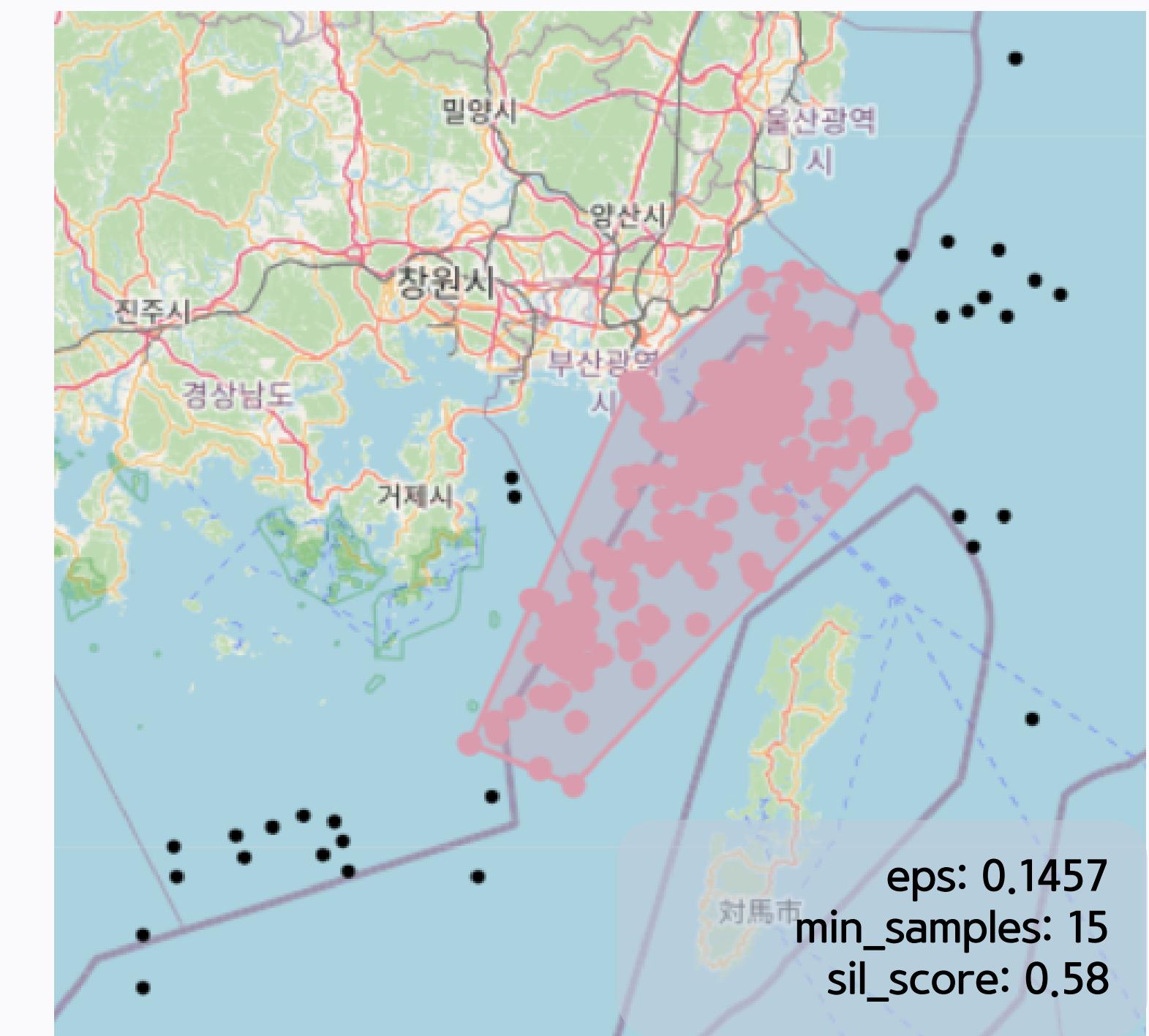
shape (309, 30)

7. 결과 시각화

싱가포르항

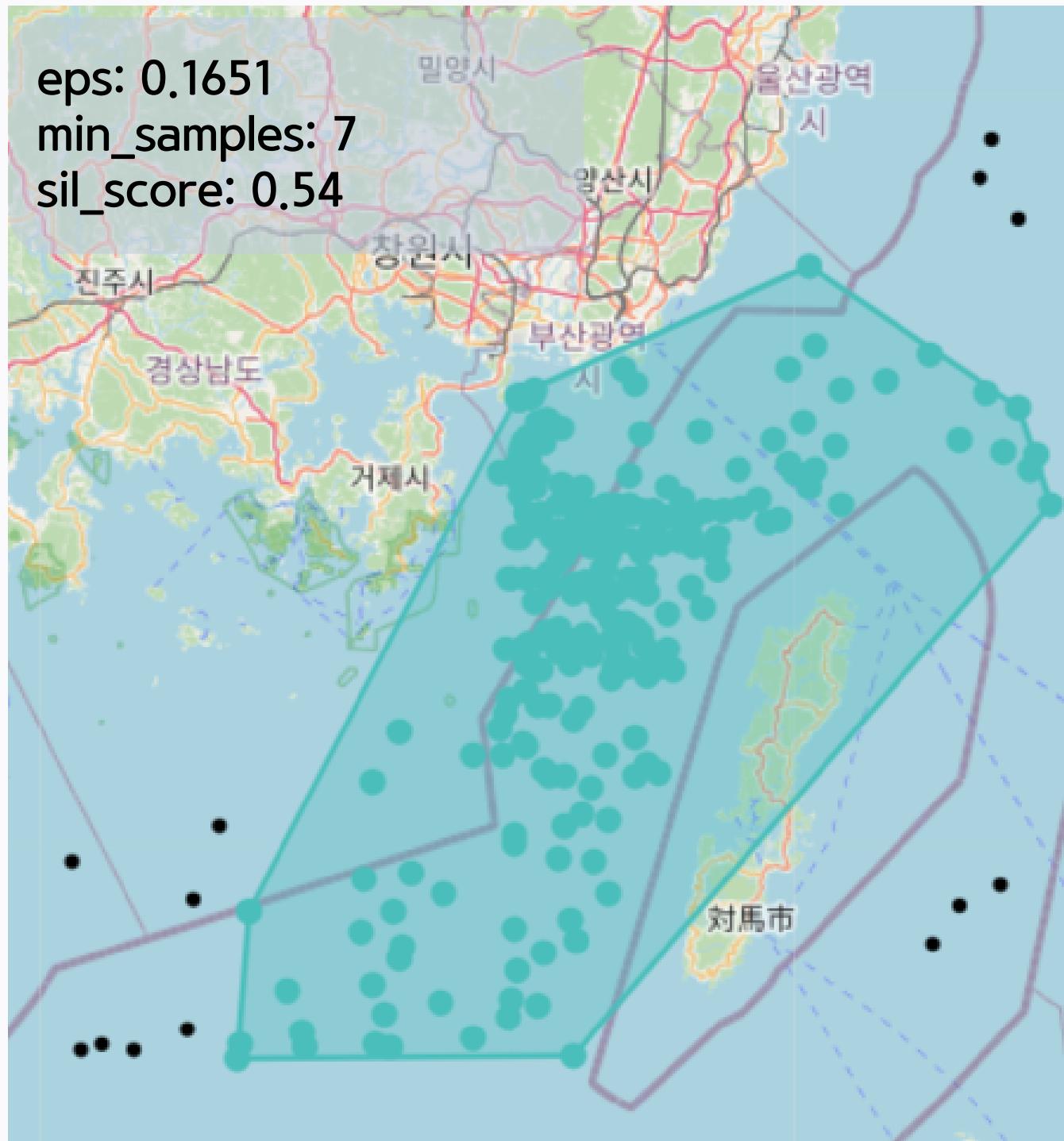


부산항

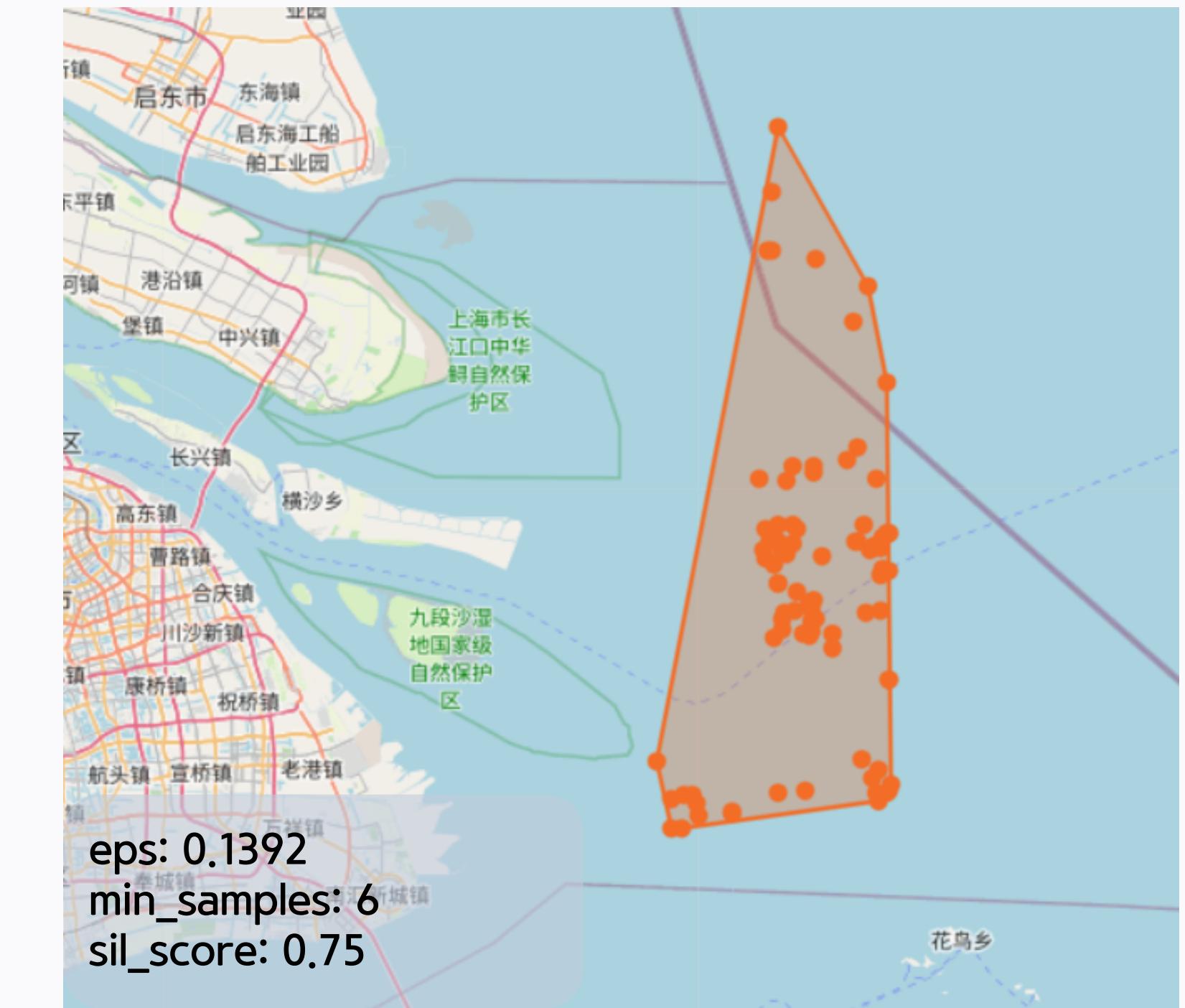


7. 결과 시각화

부산신항

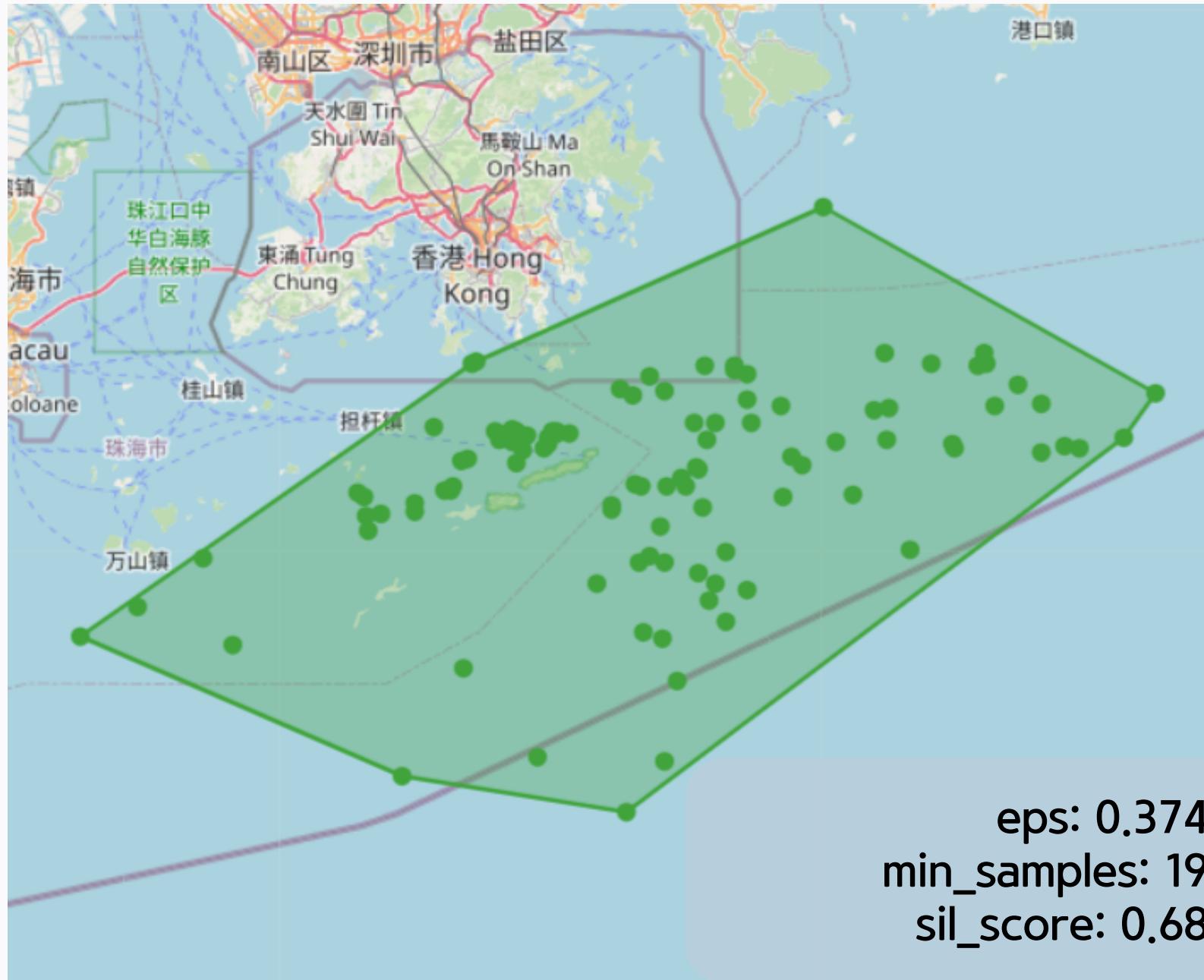


상해항

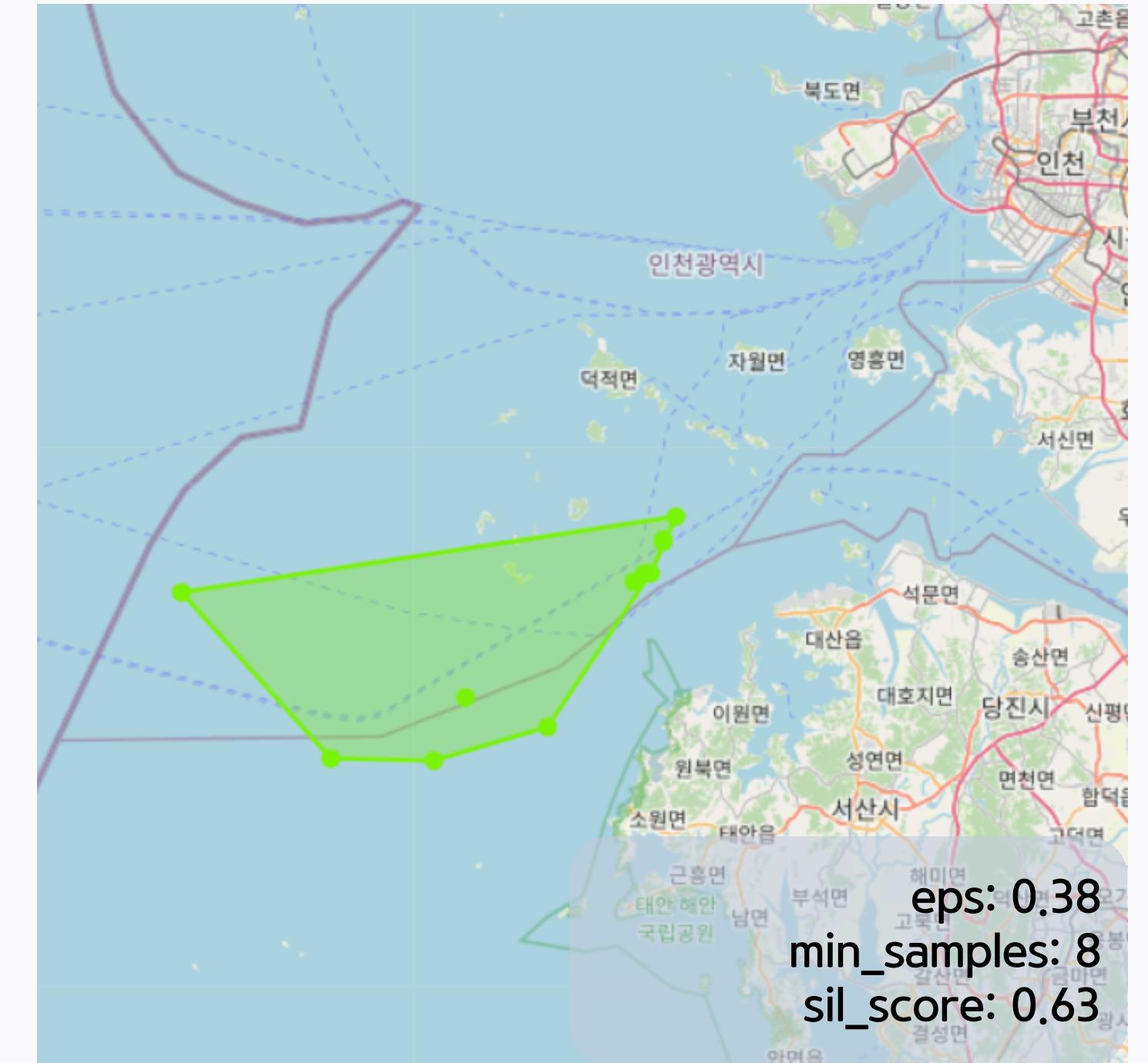


7. 결과 시각화

홍콩항



인천항



8. DBSCAN 모델링 결과

싱가폴 제외 5개 항구에 단일 군집이 만들어진 이유

→ 군집 수가 많아질 경우, 실루엣 스코어 $0.5 \downarrow$

= 대기 구역 사용 위치가 일정하다



군집 간 차이가 아닌
공통점과 동질성 파악을 목적으로 설정

03-1 군집별 특성 분석

1. 최종 데이터 생성

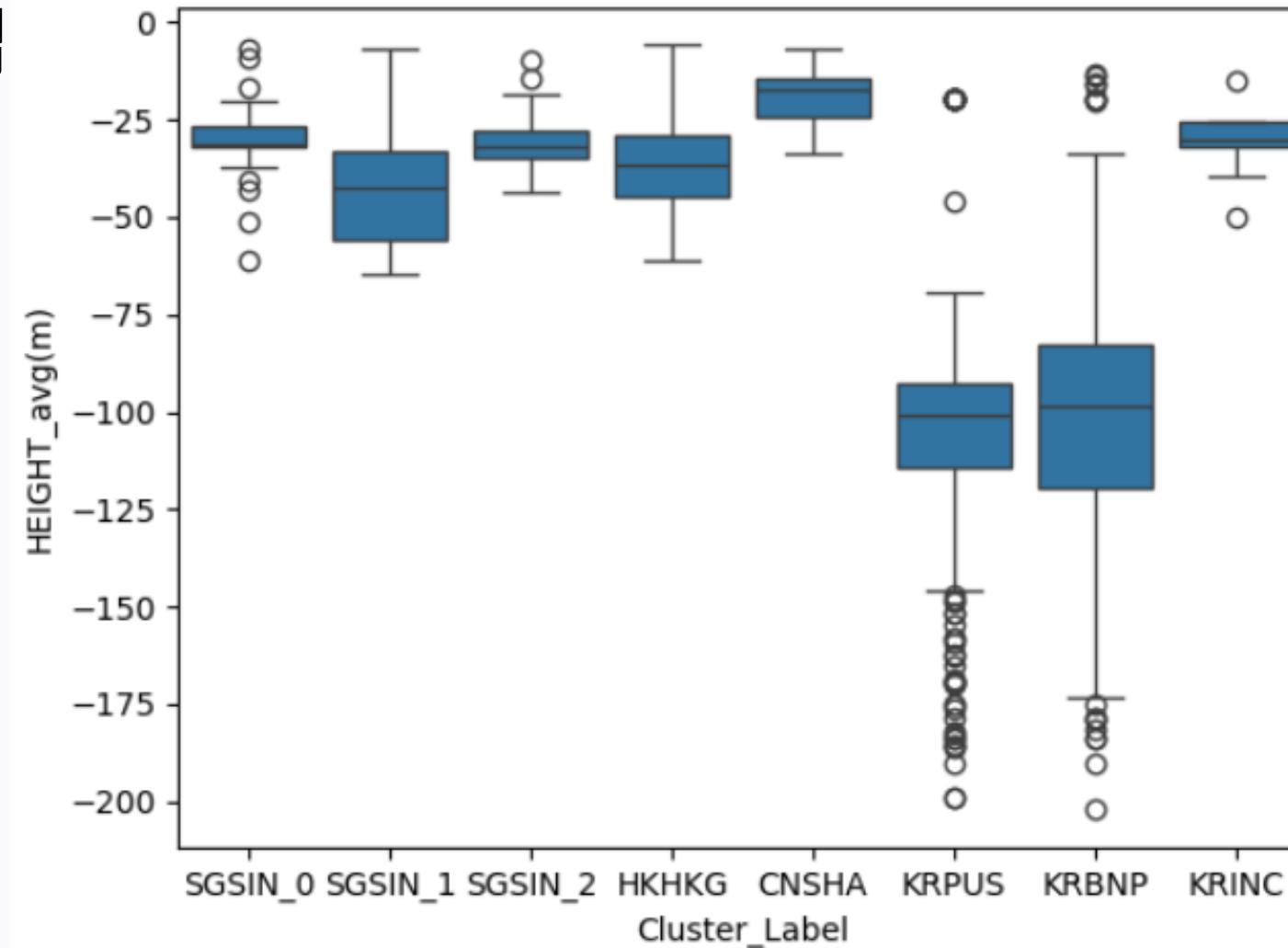
군집화 결과 + 선박 정보 + 환경 정보

xlsx

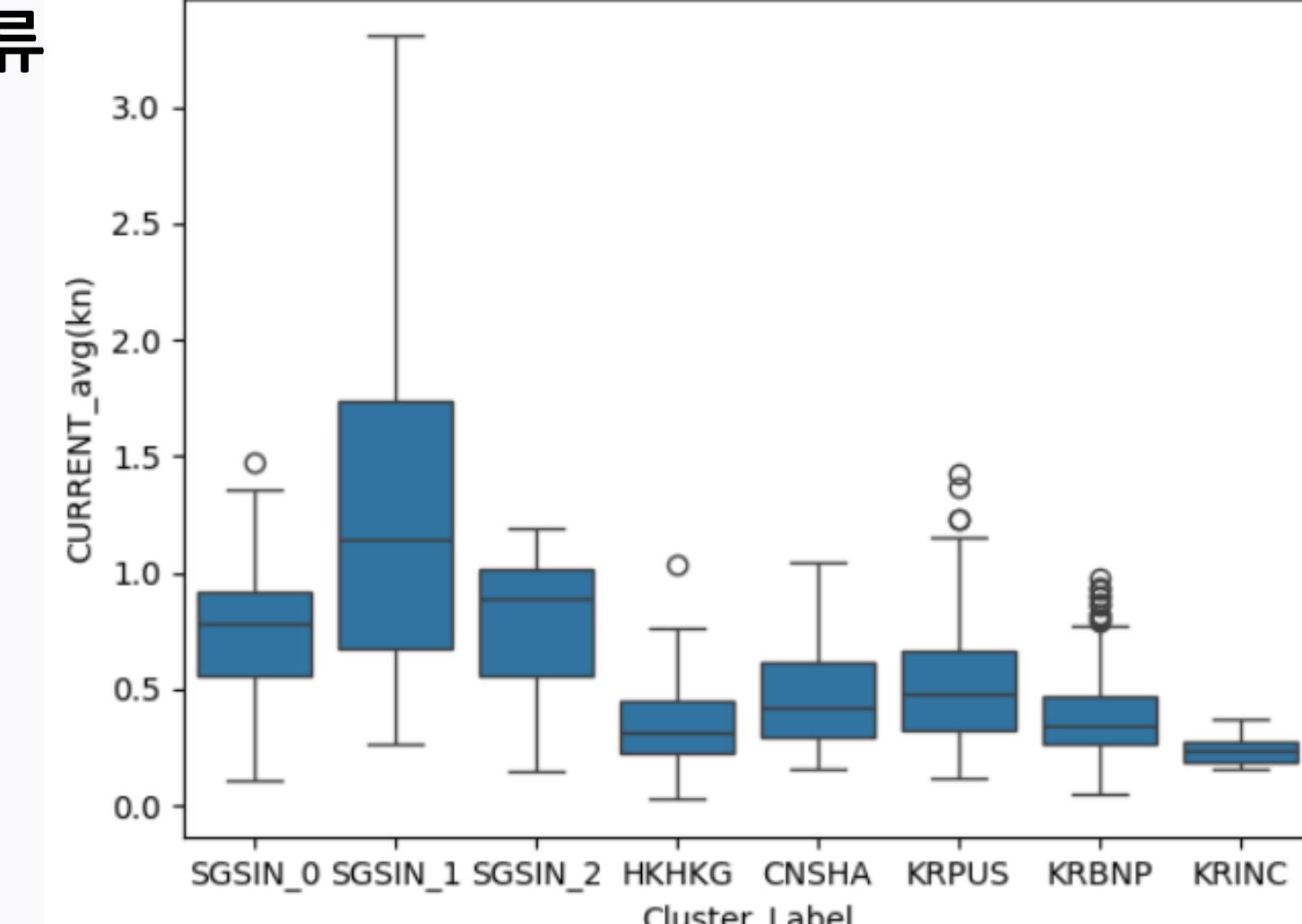
THE_LAST_DF.csv

TO_PORT_CD	VSL_ID	TRIP	WAITING_TIME	WAIT_LAT	WAIT_LON	SHIP_TYPE	GROSS_TONNAGE	Cluster_Label
도착 항구 코드	선박 고유 ID	항해 번호	대기 시간	대기지점 위도	대기지점 경도	선박 종류	총 톤수	군집 식별자
SGSIN	97966ffe-bd6b-3f06-8988-5d14d024d40c	0	0 days 00:20:30	1.189	103.505	CONTAINER	41678	SGSIN_0
SGSIN	31980a8d-0b50-3606-9799-fbac49a269b1	1	1 days 05:21:58	1.795	104.572	CONTAINER	151963	SGSIN_1
SGSIN	2f99a722-9d17-3e58-981c-d13cd722178e	2	0 days 01:27:58	1.154	103.513	CONTAINER	10308	SGSIN_0
WAVE_avg(m)	WIND_avg(kn)	CURRENT_avg(kn)	HEIGHT_avg(m)	DRAFT(m)	SQUARE(m ²)			
평균 파고(m)	평균 풍속(노트)	평균 해류 속도(노트)	평균 수심(m)	흘수(m)	면적(m ²)			
0.5	10.799	0.979	-28.5	11.0	8512			
0.455	6.7494	0.718	-49.5	15.0	20298			
0.091	5.939	0.435	-32.0	7.6	3300			

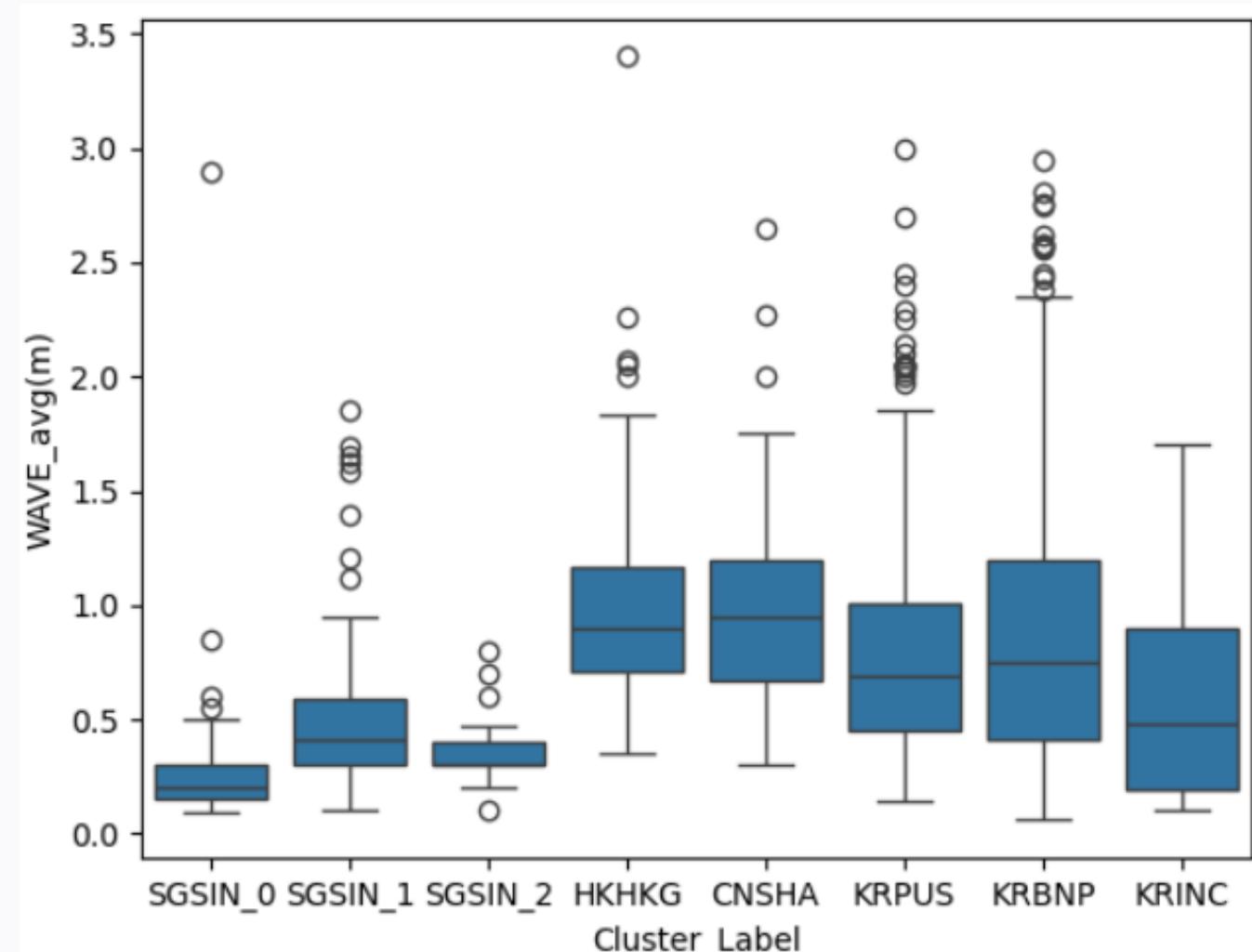
수심



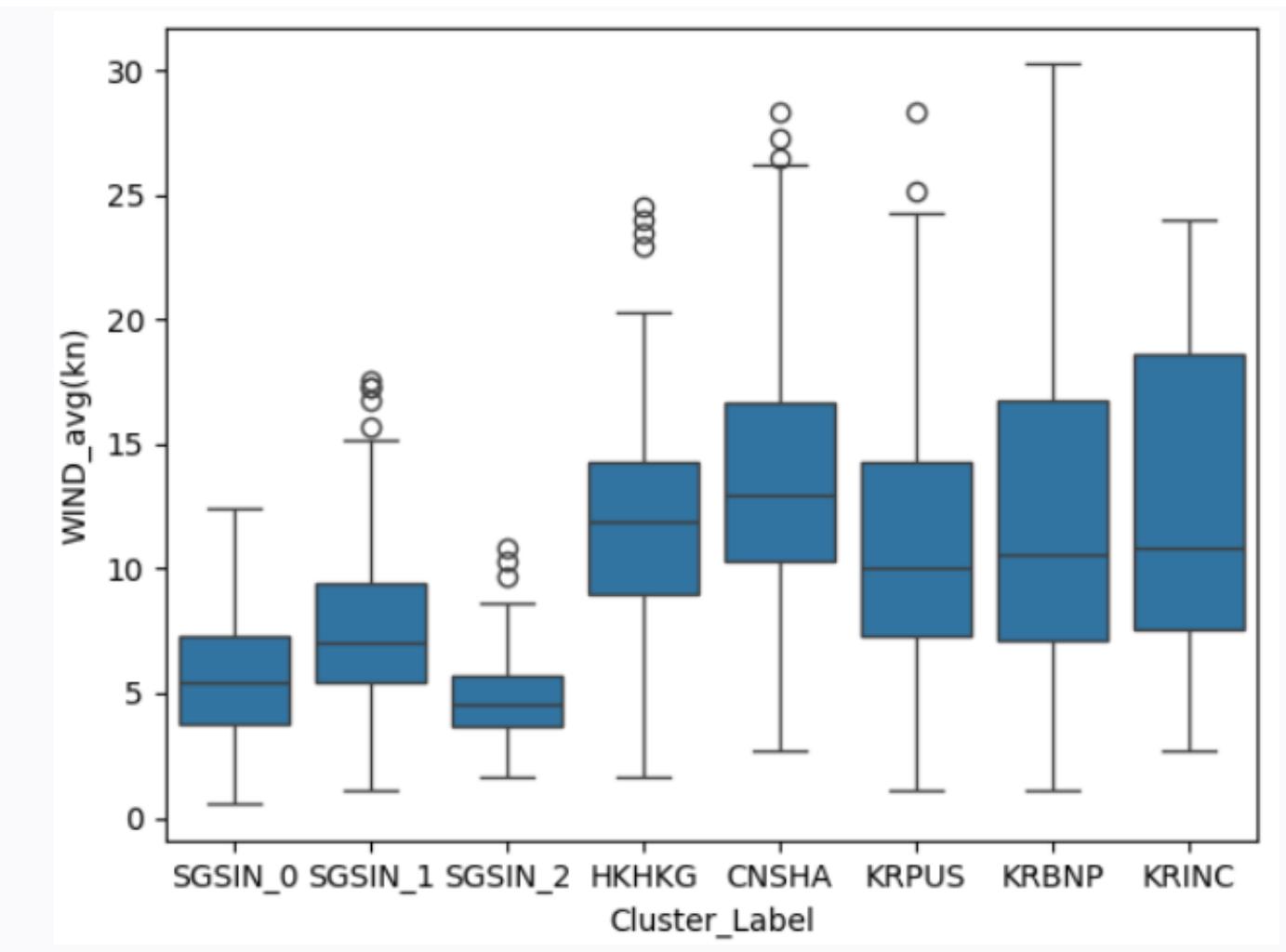
해류



파고



풍속



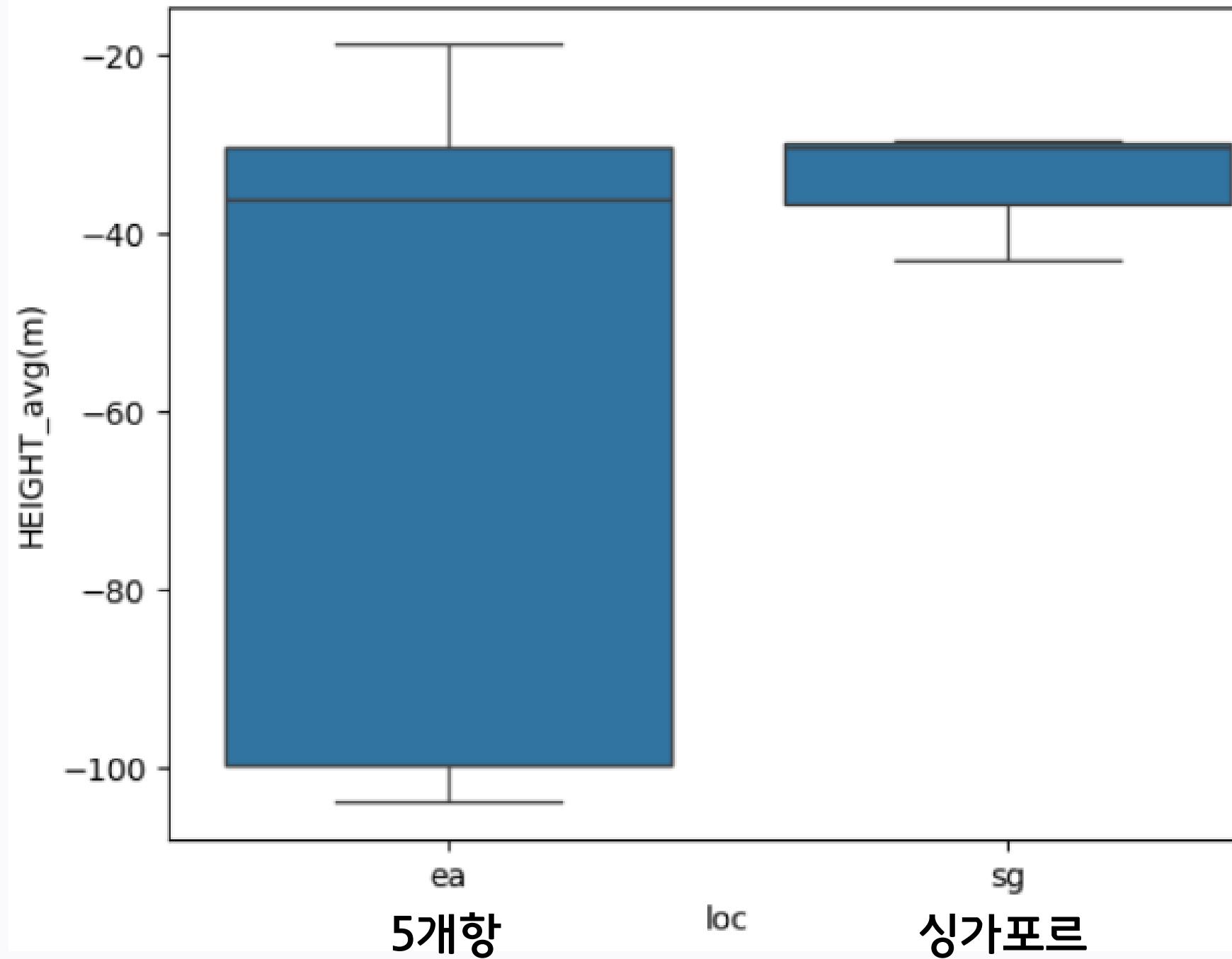
2. 항구별 분석 결과

군집	평균 파고(m)	평균 풍속(노트)	평균 해류속도(노트)	평균 수심(m)
	평균	평균	평균	평균
상해	0.99	13.67	0.47	-18.75
홍콩	1.0	11.91	0.34	-36.08
인천	0.65	12.35	0.24	-30.39
부산 _新	0.91	12.15	0.38	-99.75
부산	0.82	11.01	0.52	-103.8
싱가폴_0	0.25	5.61	0.76	-29.45
싱가폴_1	0.5	7.72	1.3	-43.04
싱가폴_2	0.35	5.08	0.78	-30.28

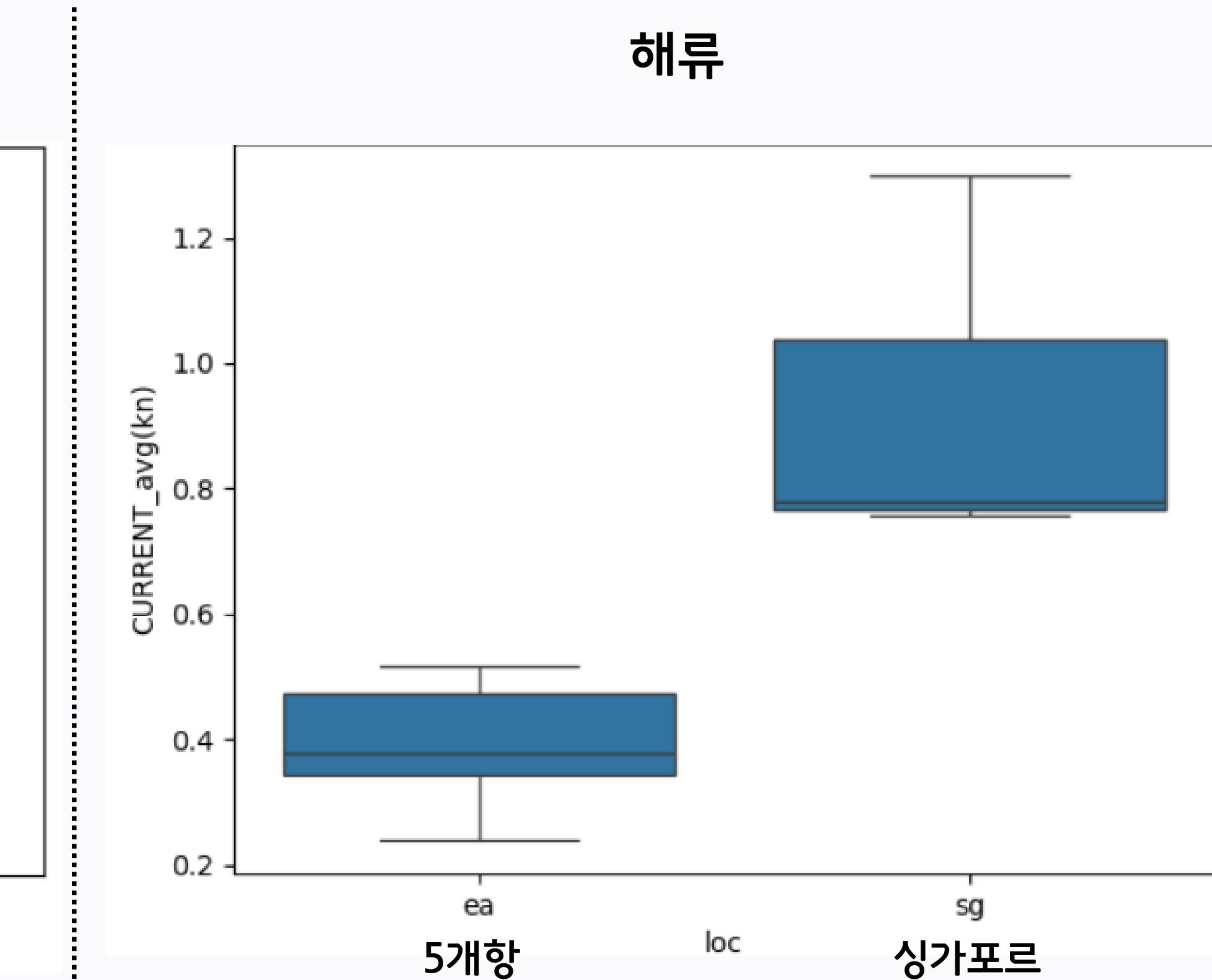
high
low

4. 권역별 분석 결과

수심

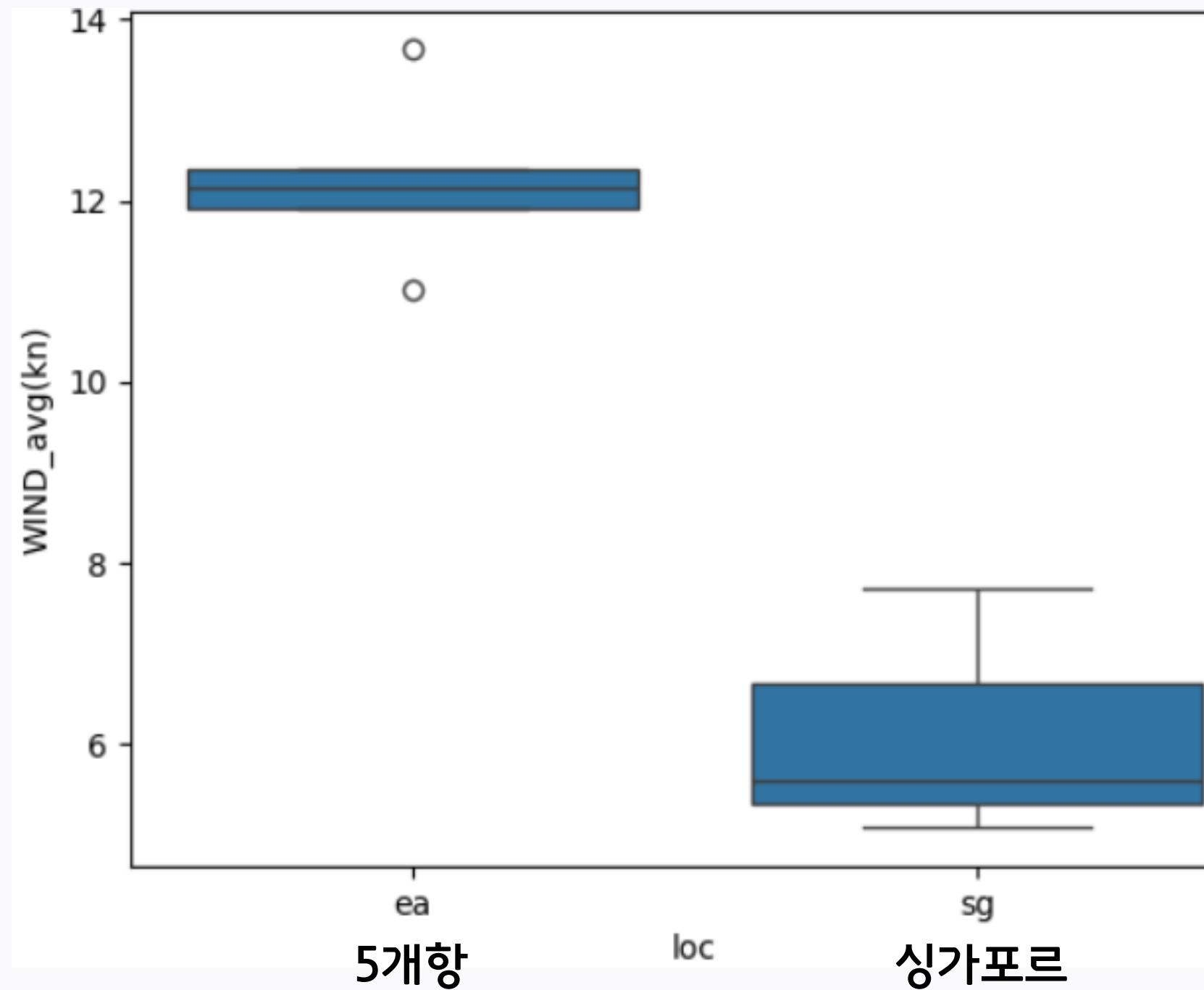


해류

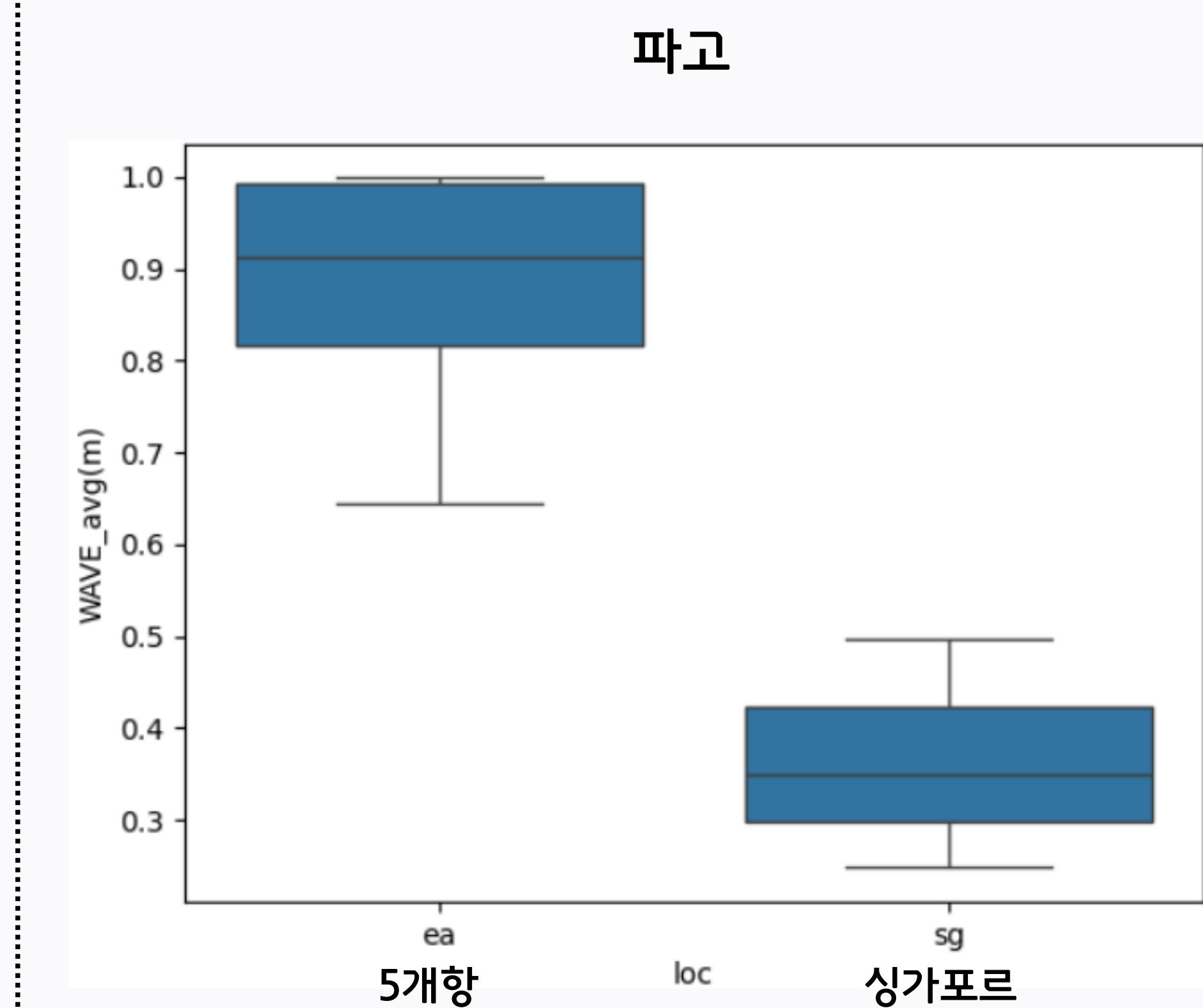


4. 권역별 분석 결과

풍속



파고



4. 권역별 분석 결과

1. 싱가포르를 제외한 항구들의 기상 정보 및 해양 지형의 특성 차이 발견

2. 싱가포르가 다른 결과를 보이는 이유

- 1) 섬이 많은 지역에 위치
- 2) 상승 기류, 적도무풍대
- 3) 좁은 구역에서 많은 선박이 이동하여 빨라진 해류

4. 권역별 분석 결과

그러나, 실질적으로 싱가포르-다른 5개항 사이 큰 폭의 차이 X

	파고(m)		풍속(kn)		해류(kn)	
	평균	최대	평균	최대	평균	최대
동아시아	0.87	3.4	12.2	30.2	0.39	1.4
싱가폴	0.36	2.9	6.1	17.5	0.94	3.3
평균	0.615	3.15	9.15	23.85	0.665	2.35

- 풍속: 6 kn 이하
- 해류: 1 kn 이하
- 파고: 1m 이하



각 대기 구역 간 환경적 동질성이 있다

결론

『선박의 종류에 따른 군집의 차이는 없으며,
도착항의 환경적 특성이 대기 구역에 영향을 준다』

→ 항구의 대기 구역별로 선박이나 대기 시간에 있어 큰 차이를 보이지 않음

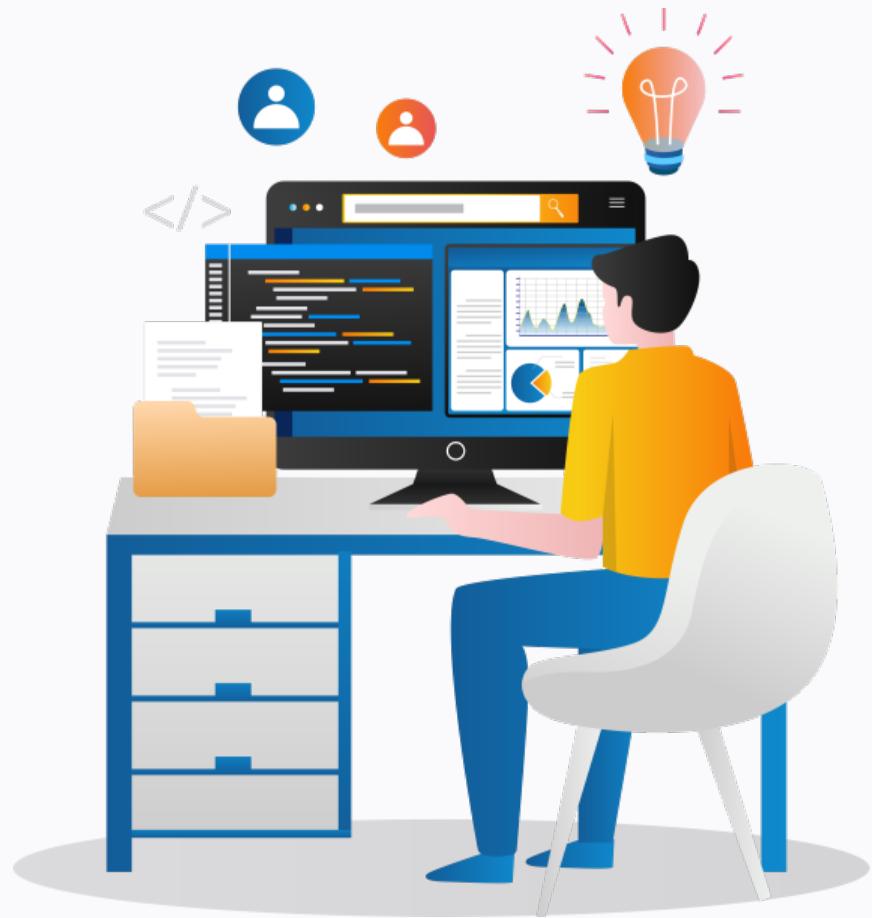
그래서, 안전한 대기 조건이란 무엇인가?

- 대기 구역의 수심 > 선박의 드래프트
- 풍속 23.8 kn 이하
- 파고 3.1 m 이하
- 해류 2.3 kn 이하

04 총평

한계 및 기대 효과

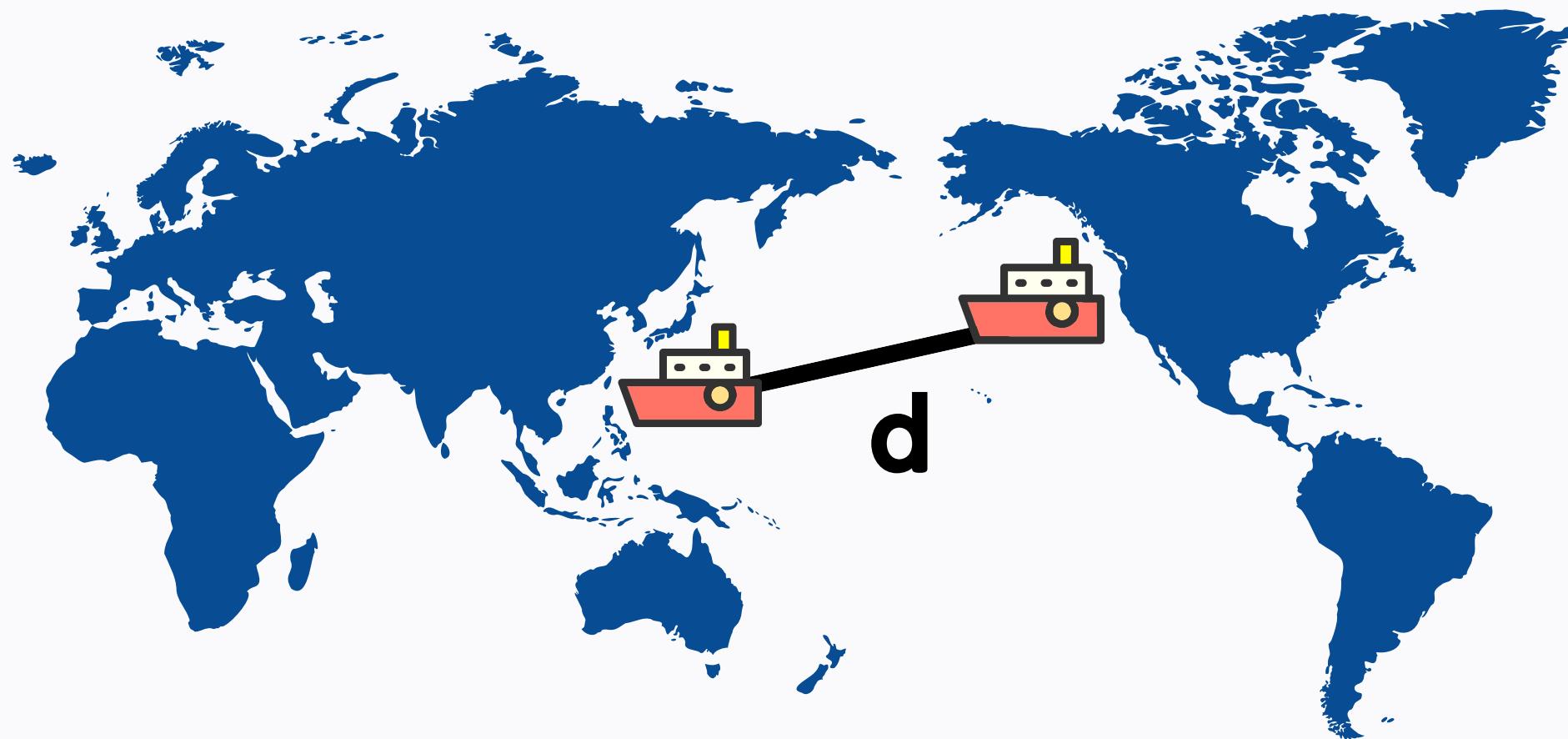
한계점



데이터 수 부족
전처리 과정 중 시행착오 多
→ 군집 간 명확한 차이점,
특이점 도출 不可

선박 상태 확인 불가
- 난파, 수리, 연료 부족 etc.

개선 사항



2차원 직교좌표계

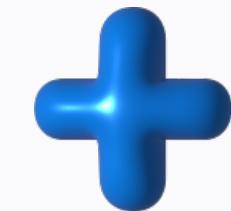


WGS84 좌표계

→ 곡률 반영

기대 효과

보다 많은 도착항



선박 상태 파악



모델 개선

군집 개수 증가

모델 정확도 상승

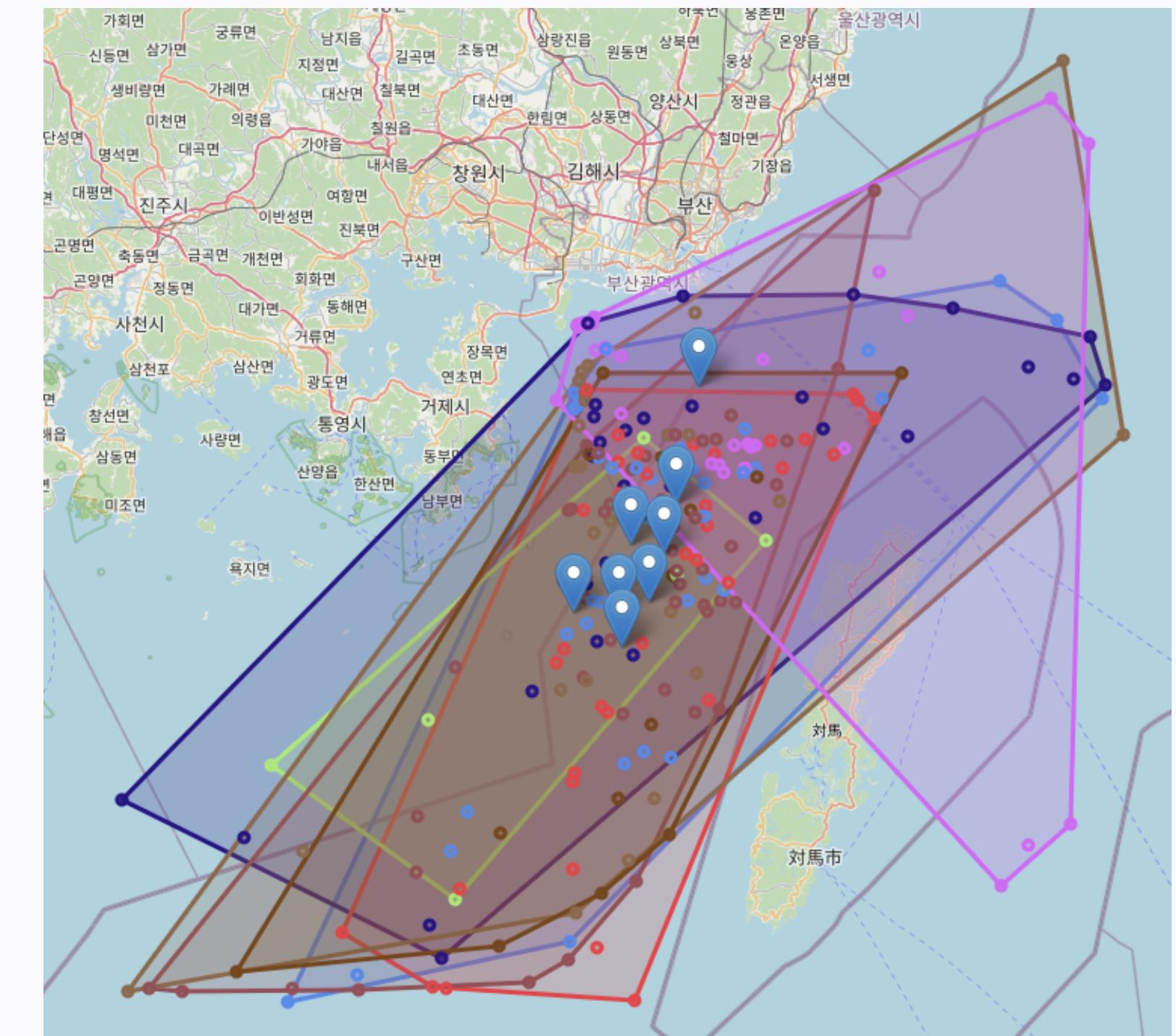
감사합니다

Q & A

모델링에 좌표만 사용한 이유

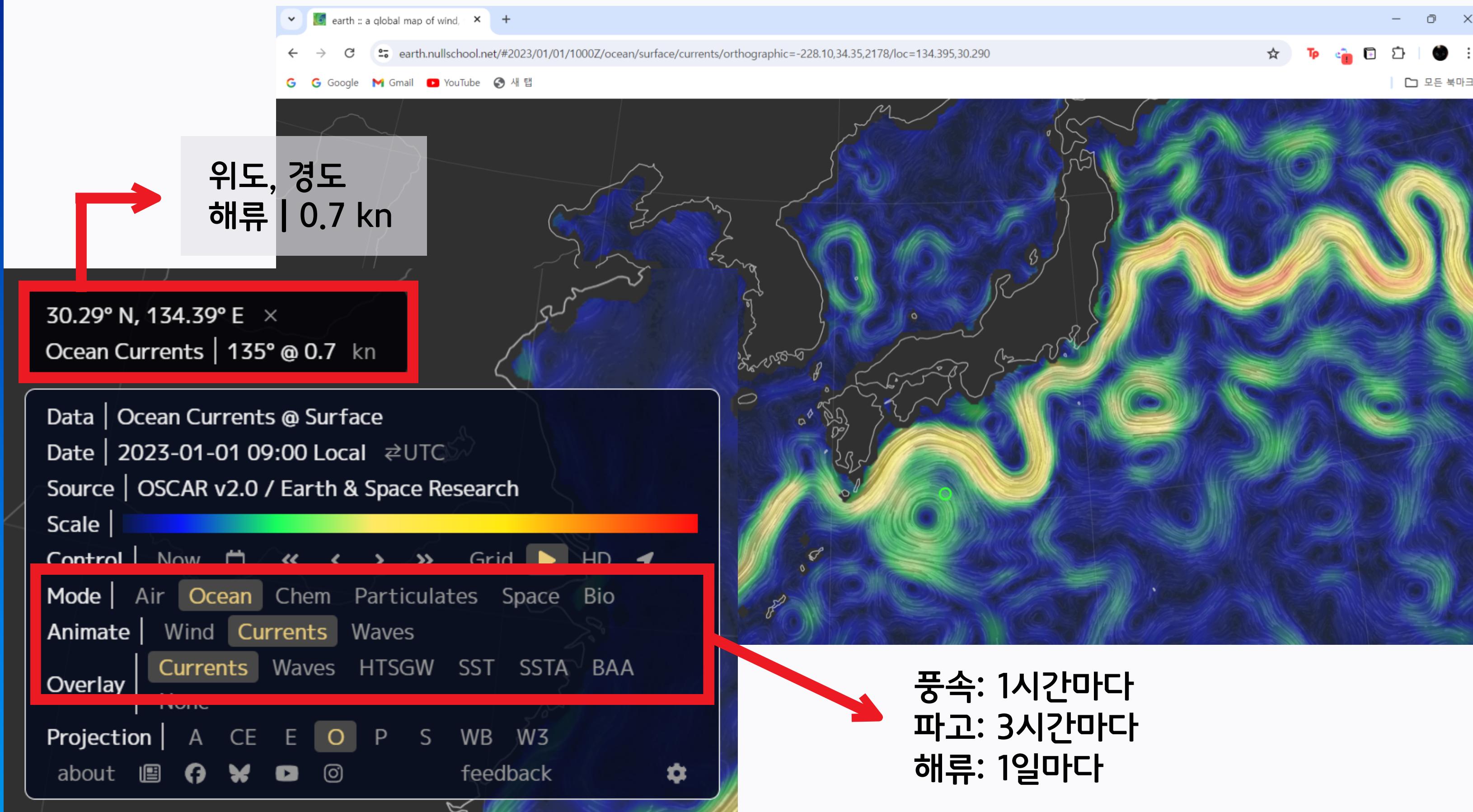
선박 정보, 기상 정보 등을 함께 넣고 모델링한 결과,

1. 군집이 겹쳐서 형성되는 경우 발생
2. 평가 지표 결과(실루엣 스코어)가 0.5 미만
3. 데이터 스케일링 + 상관계수 높은 컬럼 제거
→ 정상적 진행 X, 좌표만으로 군집화 수행

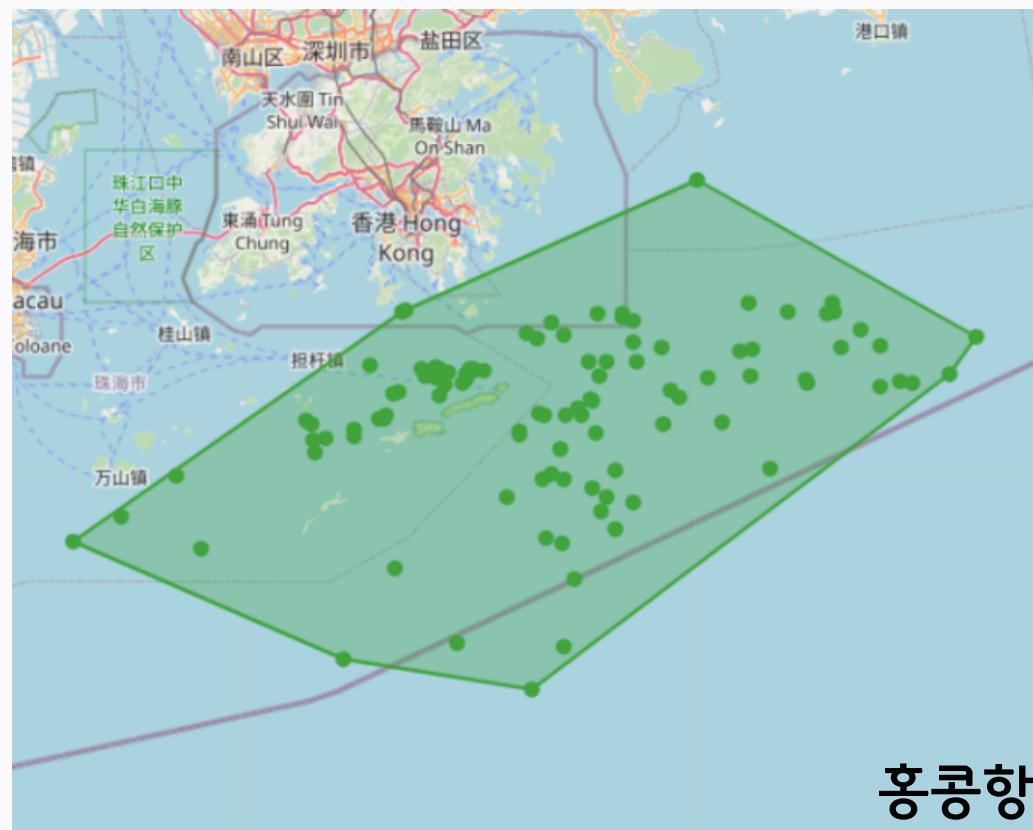
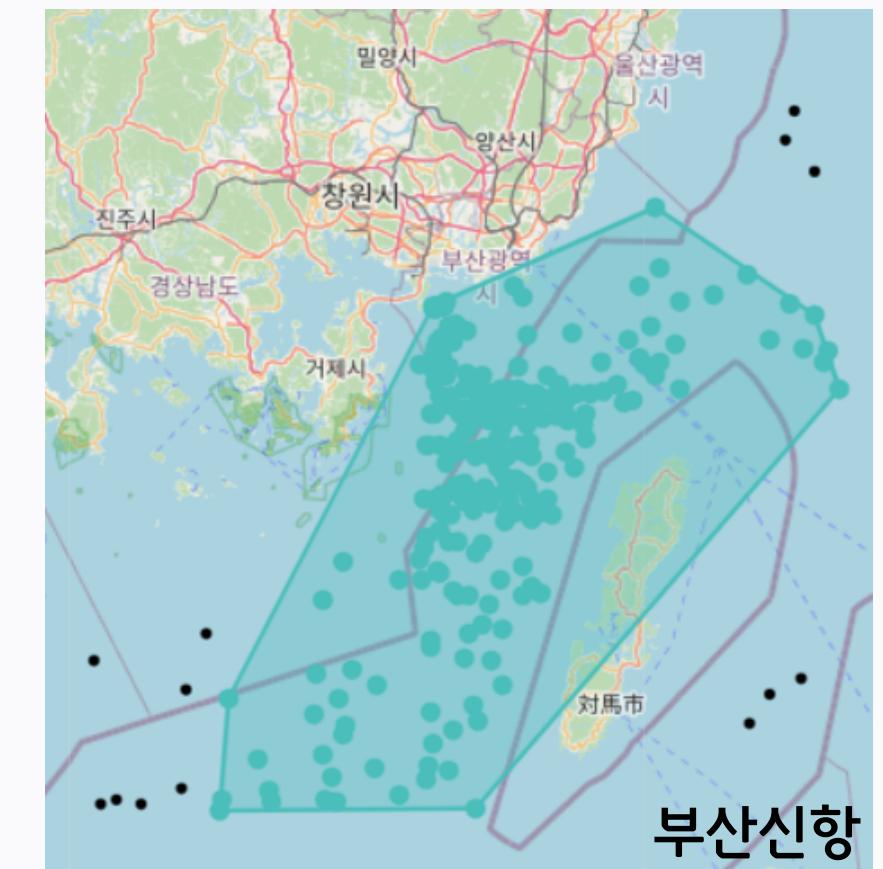
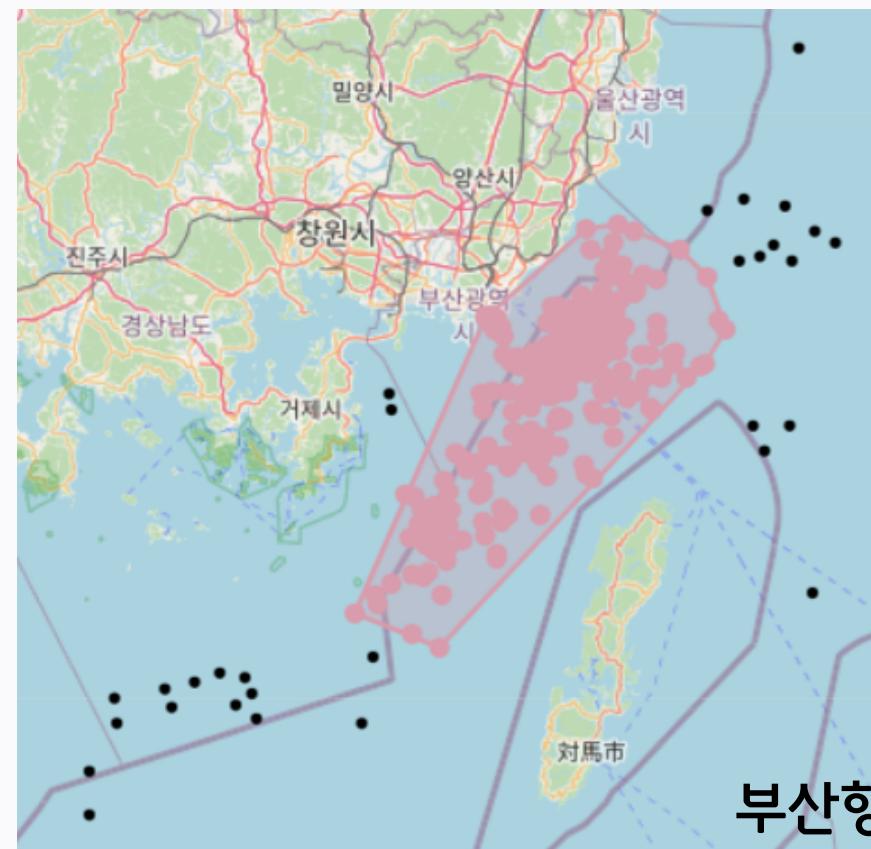
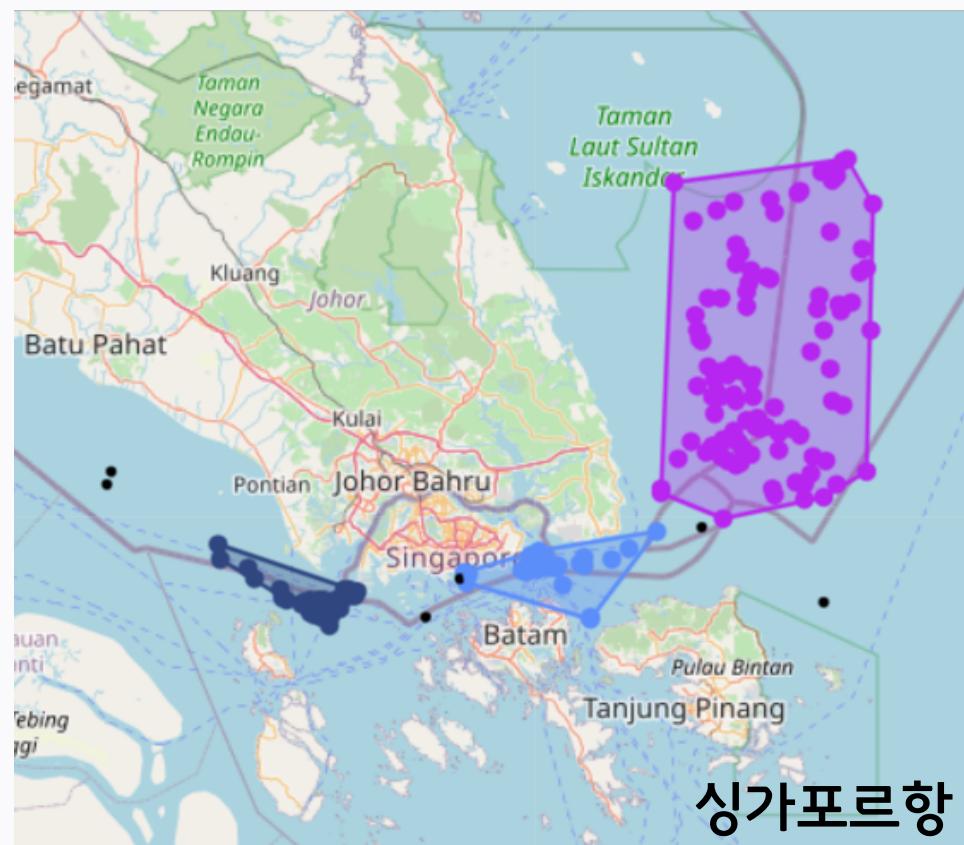


웹스크래핑 사이트

<https://earth.nullschool.net>



군집화 결과 모음



하나의 선박이 대기 중 얼마나 멀어질 수 있는가?

START_LAT	END_LAT	START_LON	END_LON
34.176315	34.791668	128.86574	129.28072
	값 차 : 0.615		값 차: 0.414

최종 모델링 데이터 수

- 부산항(KRPUS): 309 개
- 싱가포르항(SGSIN): 275 개
- 홍콩항(HKHKG): 107 개
- 부산신항(KRBNP): 260 개
- 상해항(CNSHA): 88 개
- 인천항(KRINC): 14 개

항구별 군집 내 가장 먼/가까운 거리

항구	가장 먼 거리(km)	가장 가까운 거리(km)
부산항(KRPUS)	110.19	55.60
부산신항(KRBNP)	182.80	0.29
인천항(KRICN)	40.46	0.31
홍콩항(HKHKG)	93.43	0.85
상해항(CNSHA)	97.02	0.34
싱가포르항(SGSIN) 0	38.83	4.39
싱가포르항(SGSIN) 1	118.50	6.37
싱가포르항(SGSIN) 2	28026	4.06

항구에서 가장 먼/가까운 선박

항구	가장 먼 거리(km)	가장 가까운 거리(km)
부산항(KRPUS)	85.35	24.62
부산신항(KRBNP)	139.38	23.93
인천항(KRICN)	93.80	53.70
홍콩항(HKHKG)	85.20	29.13
상해항(CNSHA)	82.60	83.14
싱가포르항(SGSIN) 0	26.60	25.63
싱가포르항(SGSIN) 1	168.32	102.86
싱가포르항(SGSIN) 2	45.92	31.50

Q. 군집이 1개인 것이 유의미한 결과인가?

A. DBSCAN은 최적의 파라미터와 평가 지표를 기반으로 나온 모델이기 때문에 군집의 개수 자체에서 의미를 찾기보다는 군집이 왜 1개로 나왔는지를 볼 필요가 있습니다.

군집이 1개라는 것은 어떤 의미일까요? 항구 근처에서 밀집 지역이 그만큼 넓게 형성되어 있다고 해석할 수 있겠죠? 다만 개선 사항에서 말씀드린 대로 지구의 곡률이 반영되지 않아 적도에서 멀어질수록 선박 간의 거리값에 오차가 발생했기 때문에 밀집도에도 오차가 생겼을 것입니다. 좌표값에 해수면 높이값을 추가한다면 더욱 정확한 거리값을 얻을 수 있을 것이고, 개선된 밀집도를 반영해서 군집화를 수행한다면 군집의 개수도 바뀔 수 있겠습니다.