# ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

# DEPARTMENT OF INFORMATICS

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΜΕΡΟΣ Β ΕΡΓΑΣΙΑΣ Γ.Π.Σ

ΟΝ/ΝΥΜΟ:ΜΑΡΚΟ ΠΛΑΚΟΥ
Α.Μ:Π17107
ΗΜ/ΝΙΑ:06/09/2021


# ΓΕΩΓΡΑΦΙΚΑ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

# ΠΡΟΛΟΓΟΣ

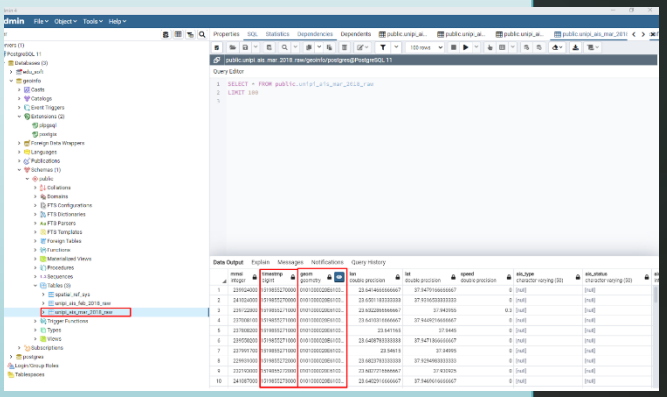



Στην εργασία αυτή:

- Παρουσιάζονται τα μέρη 1,2,3 της εργασίας.
- Επεξηγείται η απάντηση κάθε ερωτήματος.
- Χρησιμοποιήθηκαν τα εργαλία Postgres,Qgis,Python3 (Jupyter NB)

# 1. Φόρτωση δεδομένων (data loading)

- Χρησιμοποιήθηκε η βάση δεδομένων **Saronic Gulf Maritime AIS Dataset**.
- Η εισαγωγή της στήλης **point** έγινε μέσω της Postgres με **SIRD 4326**. Τα σήματα αποθηκεύτηκαν σε διαφορετικούς πίνακες για κάθε μήνα.

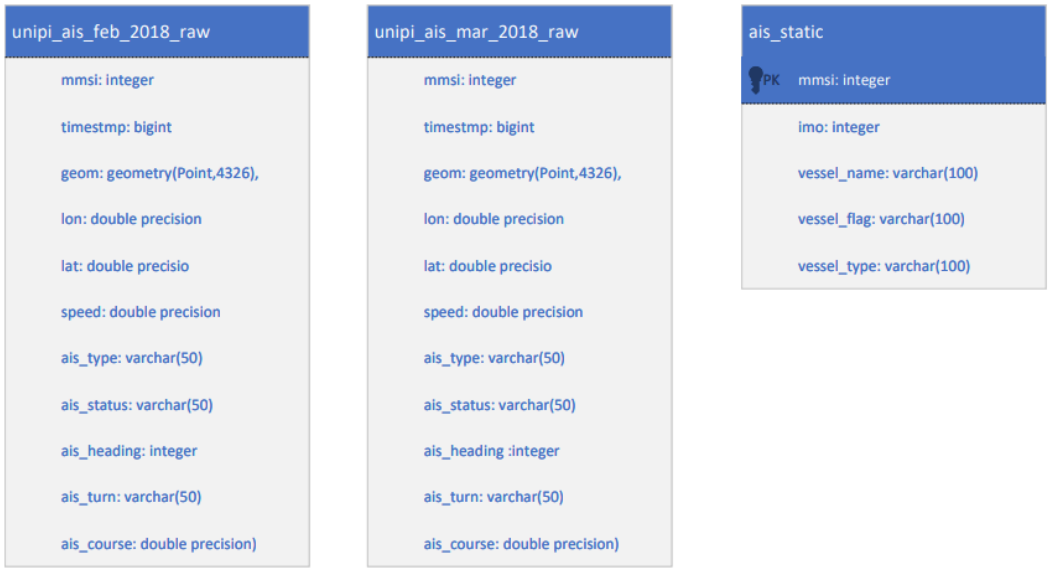


```
mar_geodf.head()
```

| | mmsi | timestmp | geom | lon | lat | speed | ais_type | ais_status | ais_heading | ais_turn | ais_course |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 239569 | 636013190 | 1519855200000 | POINT (23.53934 37.88567) | 23.539338 | 37.885675 | 1.4 | None | None | 44.0 | None | 15.5 |
| 239571 | 237991700 | 1519855201000 | POINT (23.54615 37.94995) | 23.546150 | 37.949950 | 0.0 | None | None | NaN | None | 268.0 |
| 239572 | 239722800 | 1519855201000 | POINT (23.63220 37.94397) | 23.632205 | 37.943968 | 0.0 | None | None | NaN | None | 0.0 |
| 239573 | 239550200 | 1519855201000 | POINT (23.64085 37.94716) | 23.640850 | 37.947157 | 0.0 | None | None | 264.0 | None | 142.1 |
| 239570 | 241024000 | 1519855201000 | POINT (23.65010 37.93166) | 23.650097 | 37.931657 | 0.0 | None | None | 332.0 | None | 178.1 |

# 1.ΣΧΕΔΙΑΓΡΑΜΜΑ ΤΗΣ ΒΑΣΗΣ

**PART B:**
**B1.GEOINFO UML DATABASE SCHEMA**



# 2.Γνωριμία με τα δεδομένα και προετοιμασία (data acquaintance and preprocessing)

**2.1** ΚΑΘΑΡΙΣΜΟΣ ΤΩΝ ΔΕΔΟΜΕΝΩΝ

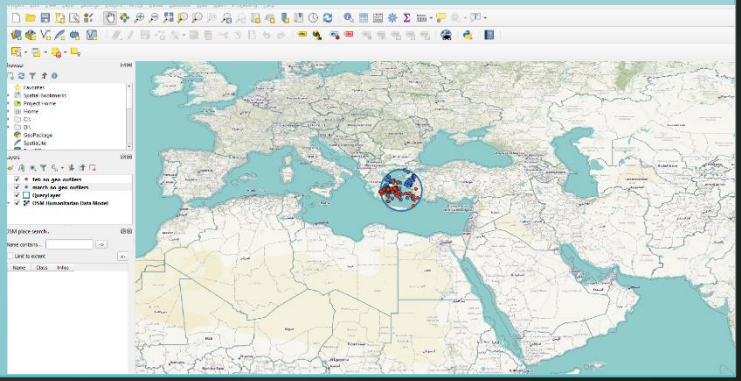
- Αφαιρέθηκαν οι διπλότυπες εγγραφές, και αυτές που είχαν NaN τιμές στις στήλες timestmp, speed.
- Αφαιρέθηκε ο χωρικός θόρυβος.
- Έμειναν μόνο οι εγγραφές που το mmsi τους υπήρχε στα στατικά δεδομένα και αυτές των οποίων η ταχύτητα είναι μικρότερη των 50 κόμβων.
- Υπολογίστηκε η επιτάχυνση.

**ΜΕΤΑ ΤΟΝ ΚΑΘΑΡΙΣΜΟ**

Το πλήθος δεδομένων του Φεβρουαρίου μειώθηκε κατά **1.145.481** εγγραφές.

```
len(final_feb_geodf) #the final length of the February dataset.
```

```
4960765
```



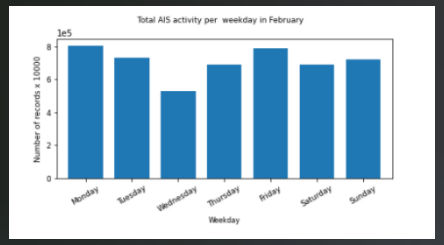
Το πλήθος του Μαρτίου μειώθηκε κατά **1.587.841**.

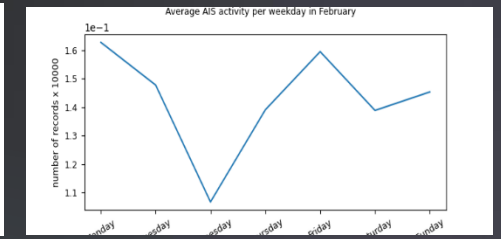
```
len(final_mar_geodf)#the final length of the March dataset.
```

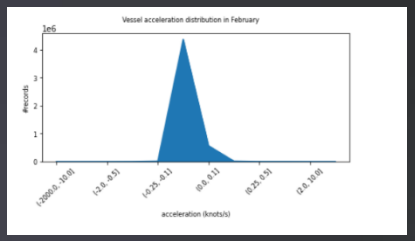
```
5116283
```

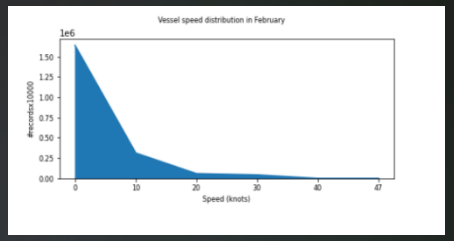
# 2.2-2.4

- **2.2-2.3** Τα διαγράμματα δειγματοληψίας και ταχύτητας για τον μήνα  Φεβρουάριο:



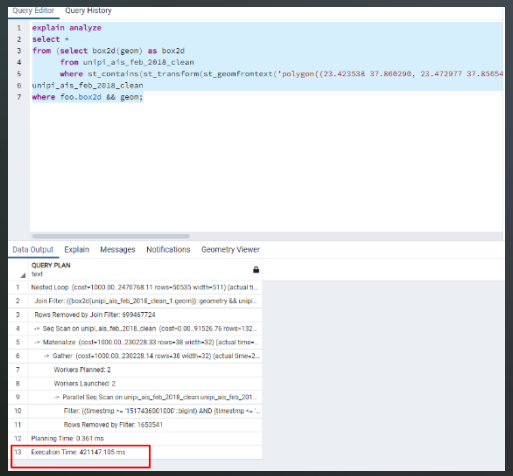



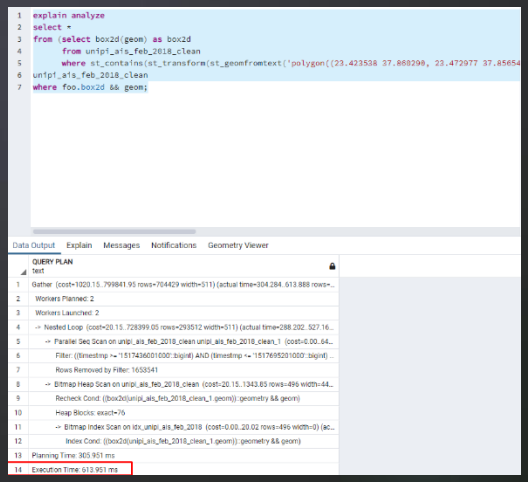




- **2.4** Η ύπαρξη του R-tree ευρετηρίου μείωσε κατά πολύ τον χρόνο εκτέλεσης :

ΠΡΙΝ(**421147.105 ms**)

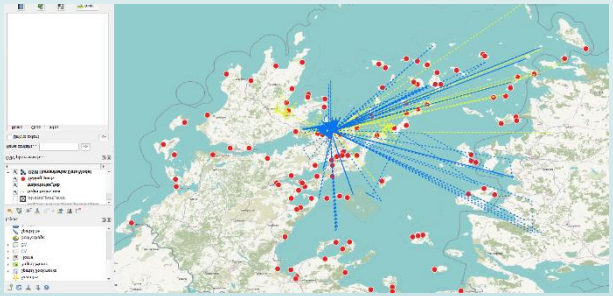



ΜΕΤΑ(**613.951 ms**)



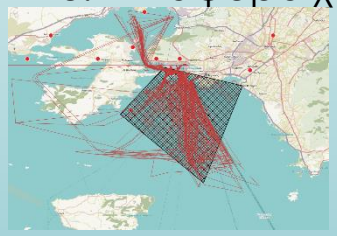
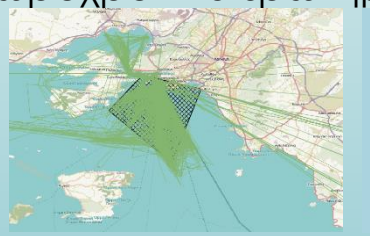
# 3.Επεξεργασία και αναλυτική των δεδομένων (data processing and analytics)

- **3.1** Για το ερώτημα αυτό εκτελέστηκε ένα query στο QGIS (συνδεδεμένο με Postgres) . Οι τροχιές κατηγοριοποιήθηκαν ανά mmsi και ανά μήνα. Το segmentation έγινε με την helper.



```
(Initial) Number of segments: 1
(Final-Useful) Number of port-based segments produced: 1
(Initial) Number of port-based segments: 1
(Intermediate) Number of temporal-gap-based segments: 6
(Final-Useful) Number of trips produced: 5
```

- **3.2** Διάφορα χωροχρονικά ερωτήματα πραγματοποιήθηκαν:

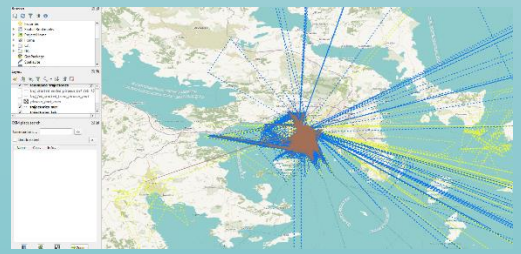
- **3.3** Το temporal alignment και η αναδειγματοληψία έγινε με την helper.

**3.3 Temporal alignment and Resampling Using Helper**

```
In [8]: #TEMPORAL ALIGNMENT AND RESAMPLING
        resampled_and_aligned=helper2.temporal_alignment_v2(trajectories_seg)

In [9]: resampled_and_aligned=resampled_and_aligned.rename(columns={'datetime':'ts'})

In [10]: len(resampled_and_aligned) #the length of the resampled dataset.
Out[10]: 83459
```

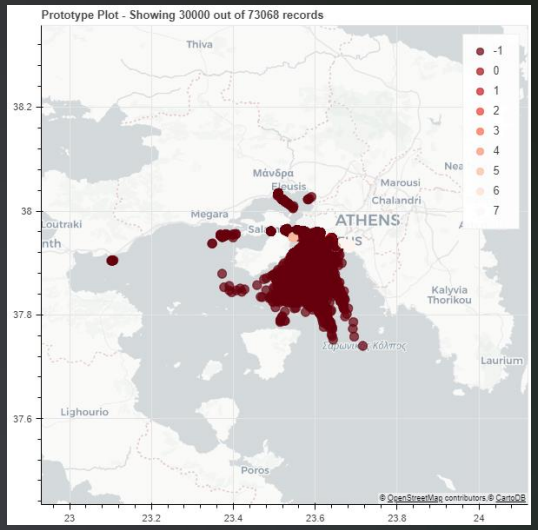


# 3.4-3.5

- **3.4** Υλοποιήθηκε ο OPTICS και βρήκε 9 συστάδες:

```
In [14]: gdf_radians_resampled_and_aligned = np.radians(ra_df [['lat', 'lon']]) #creating the radians
         optics_resampled_and_aligned  = OPTICS(max_eps=1/6371, min_samples=len(ra_df )//50, metric='haversine').fit(gdf_radians_resample
         set(optics_resampled_and_aligned .labels_)
         helper2.get_clusters_centers(gdf_radians_resampled_and_aligned , optics_resampled_and_aligned .labels_)

Out[14]: array([[37.94691256, 23.64038312],
                [37.94731144, 23.63740902],
                [37.9427454 , 23.64046713],
                [37.94420766, 23.63208578],
                [37.95457064, 23.55113343],
                [37.94993253, 23.54616344],
                [37.9316622 , 23.65010351],
                [37.93275919, 23.6812189 ]])
```




- **3.5** Υλοποιήθηκε ο Evolving Clusters και βρήκε 14 evolving cluster αποτελούμενα από 1 πλοίο:

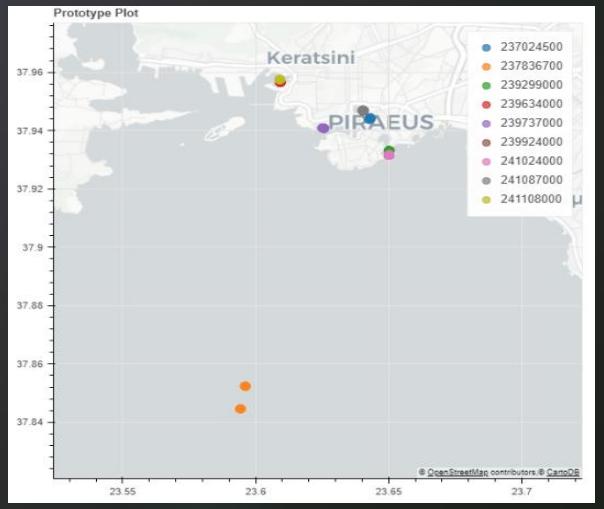
**Implementing Evolving Clusters Algorithm**

```
In [12]: [res_mcs, res_mc] = evolving_clusters(final, distance_threshold=1000, min_cardinality=1, time_threshold=1, disable_progress_bar=F

100%|██████████| 71910/71910 [06:03<00:00, 197.94it/s]

In [13]: for row in res_mcs.itertuples(): # display the clusters found
             print(row)

Pandas(Index=0, clusters=(241024000,), st=Timestamp('2018-02-01 01:07:00'), et=Timestamp('2018-02-01 01:08:00'))
Pandas(Index=0, clusters=(241024000,), st=Timestamp('2018-02-06 01:23:00'), et=Timestamp('2018-02-06 01:24:00'))
Pandas(Index=0, clusters=(241087000,), st=Timestamp('2018-02-14 18:37:00'), et=Timestamp('2018-02-14 18:38:00'))
Pandas(Index=0, clusters=(241024000,), st=Timestamp('2018-02-17 00:24:00'), et=Timestamp('2018-02-17 00:25:00'))
Pandas(Index=0, clusters=(241024000,), st=Timestamp('2018-02-17 18:06:00'), et=Timestamp('2018-02-17 18:07:00'))
Pandas(Index=0, clusters=(237836700,), st=Timestamp('2018-02-18 13:57:00'), et=Timestamp('2018-02-18 13:58:00'))
Pandas(Index=0, clusters=(239299000,), st=Timestamp('2018-03-03 14:13:00'), et=Timestamp('2018-03-03 14:14:00'))
Pandas(Index=0, clusters=(241024000,), st=Timestamp('2018-03-06 18:26:00'), et=Timestamp('2018-03-06 18:27:00'))
Pandas(Index=0, clusters=(239634000,), st=Timestamp('2018-03-13 00:34:00'), et=Timestamp('2018-03-13 00:36:00'))
Pandas(Index=0, clusters=(241108000,), st=Timestamp('2018-03-15 07:56:00'), et=Timestamp('2018-03-15 07:57:00'))
Pandas(Index=0, clusters=(239737000,), st=Timestamp('2018-03-19 22:08:00'), et=Timestamp('2018-03-19 22:09:00'))
Pandas(Index=0, clusters=(239737000,), st=Timestamp('2018-03-20 01:34:00'), et=Timestamp('2018-03-20 01:35:00'))
Pandas(Index=0, clusters=(237024500,), st=Timestamp('2018-03-28 09:52:00'), et=Timestamp('2018-03-28 09:53:00'))
Pandas(Index=0, clusters=(239924000,), st=Timestamp('2018-03-29 03:54:00'), et=Timestamp('2018-03-29 03:56:00'))
```




# Data Story και Κατακλείδα

Η εργασία αυτή ήταν εξαιρετικά ενδιαφέρουσα αφού έθεσε πραγματικά ερωτήματα και επίσης χρησιμοποιήθηκαν αληθινά δεδομένα.

Τα **συμπεράσματα** πού εξήχθησαν και 'διηγούνται' την **ιστορία των δεδομένων** είναι:

- Μειωμένη κίνηση των πλοίων τις Τετάρτες του Φεβρουαρίου και τις Κυριακές του Μαρτίου.
- Δεν υπήρχαν πλοία με ακριβώς κοινή τροχιά.
- Από την εύρεση των hot-spots , είναι εμφανές ότι η πλειοψηφία των σημάτων λαμβάνεται σε κοντινή απόσταση από το λιμάνι του Πειραιά.

**Εν κατακλείδι.**

- Έχοντας ως 'πρώτη ύλη' τα ακατέργαστα δεδομένα με θόρυβο που δεν παρέχουν σχεδόν καμία χρήσιμη πληροφορία και γνώση , με την κατάλληλο 'καθαρισμό' και προ επεξεργασία και με την χρήση τεχνικών αναλυτικής δεδομένων καταλήγουμε σε χρήσιμη πληροφορία που μπορεί να αξιοποιηθεί κατάλληλα από τους ειδικούς (domain experts).