

# Movement Behaviour Recognition for Water Activities

Mirco Nanni, Roberto Trasarti, Fosca Giannotti

ISTI CNR - KDD Lab - Pisa, Italy

**Abstract.** This work describes an analysis process for the movement traces of users during water activities. The data is collected by a mobile phone app that the Navionics company developed to provide to its users sea maps and navigation services. The final objective of the project is to recognize the prevalent activity types of the users (fishing, sailing, cruising, canoeing), in order to personalize services and advertising.

## 1 Introduction

Mobility data analysis has been a very hot topic in the last decade due to the wide diffusion of localization technologies – especially GPS devices on board of vehicles and location-based services for smartphones – as well as the large interest in understanding human mobility in urban settings, e.g. for smart-city applications.

Quoting one of its mottos, the main activity of the Navionics company<sup>1</sup> “starts where the roads end”, providing electronic charting to the ease and safety of navigation to around one million users worldwide. Nowadays, that is accomplished through a GPS-enabled smartphone app that also allows to collect the users’ tracks (under her consent) to improve the service.

This paper focuses on a specific analytical task on such data: inferring the main water activities of interest for the user, in particular distinguishing those interested in fishing, from those keen on sailing or canoeing, and those spending most of the time simply cruising around. The impact on the company is at least two-fold: on one hand that would lead to better customizing the navigation and mapping services to the user (e.g. highlighting on the map popular fishing spots only to fishermen); on the other hand, it would allow more effective target marketing initiatives.

The task presents several challenges, mainly due to the large heterogeneity of the users’ behaviours, due both to individual attitudes (a sailor obviously behaves very differently from a fisherman, but there are also several different ways to sailing, such as competitive vs. enjoyable, beginner vs. expert, etc.) and the geographical context (fishing in Australia might be very different from the US coasts or the Great lakes). That requires to abstract, to some extent, from the geography and take into consideration the existence of several different (and largely unknown a priori) behaviour profiles for the same activity type.

---

<sup>1</sup> [www.navionics.com](http://www.navionics.com)

The paper will describe the main phases of the process, highlighting the challenges faced. After discussing the state-of-art, we will introduce the data sources adopted and the preprocessing phases, which were the most time consuming steps. Next, the clustering-based selection of typical behaviors and their labeling based on expert’s input is described, followed by the construction of a decision tree-based classifier. The paper finishes with a study the results obtained with the application of the model on the whole dataset.

## 2 Related work

Several works exist in literature dealing with human mobility, trying to model and recognize users’ daily activities either by looking at their movement trajectories [6, 5] or at smartphone sensors, e.g. accelerometers, etc. [7].

Moving to boats and water activity, much less work is available. Most of it deals with AIS (automatic identification system) [1], i.e. a tracking system used for collision avoidance on ships and by vessel traffic services. Typical works try to study specific locations with issues [8] (crowded harbors, frequent maneuvers, etc.) or to detect some kinds of events [2] (e.g. collision risks). Also, some works are devoted to recognize very specific patterns, e.g. fishing in some specific (e.g. protected) areas [3].

The task we focus on is related to the analysis of individuals and their habits, which is poorly discussed in literature and there exist basically no suitable proposals for tackling the problem.

## 3 Data preparation

### 3.1 Input data sources

The analysis process is based on two types of data sources: the tracks recorded by users of the Navionics app and various geographical information of the areas where the users moved.

**Tracks** data record the trips performed by a user when the Navionics app is activated. Tracks are basically sequences of GPS points that allow to reconstruct where and when the trip took place.

Various sources of **geography** data provide information about the borders of *land* (used to remove points outside water, deemed not interesting for this project), a spatial description of *coastline* (used to asses the distance of the user from the nearest coast) and borders delimiting different *types of water*, the type of *bottom* in each point in water (represented as polygons of homogeneous bottom type), a measure of *water depth* at each geographical point.

Finally, additional information about **attractiveness of places** was available, in particular a (large but partial) database of *wrecks* localized by Navionics users, and a *popularity heatmap* of each geographical area, inferred from the track data as number of users that visited each place.

### 3.2 Pre-processing

After cleaning the raw track data from points outside water, which are not interesting for our purposes and were most likely collected by mistake (for instance the user forgot to switch the app off), we analyze the user’s trajectories to split them into subtrajectories, called *components*, that can belong to two categories: *movements* and *stops*.

The key operation is to define when a boat is still and when it is moving. Indeed, both the floating of a boat when it is not really moving (e.g. when anchored) and the fluctuations of GPS localization can create an apparent movement, therefore a total stop usually does not exist. For this reason, we operationally defined the status of a boat at a given moment as a stop if it remains very close to its initial position (w.r.t. a spatial threshold) for a significant amount of time (w.r.t. a minimum duration threshold). Both thresholds are parameters provided by the analyst. In the experiments shown later in the paper, they were empirically set to 50 meters and 10 minutes. Complementary, all the points that do not belong to a stop period represent movements.

Each single point of a component was characterized by several features derived either from the local movement performed (estimate speed and angle) or from the area around its geographical position. In particular, in this project we selected the following basic features: *speed*; *angle*, which represents the turn angle described by three consecutive points; *distance from the coast*; *sea bottom type* (sand, rock, etc.); *depth*; *water type* (salted, fresh); presence of *wrecks*; *sea bottom steepness*, estimated from depth information.

Based on them, component-level features were extracted. Several of them are obtained through simple aggregation, such as median value, inter-quartile values and range, or taken from the first point of the component, such as the initial timestamp and latitude. Others required more complex processes. In particular, we shortly describe here *acceleration periods* and *noodle shapes*, while other examples are omitted for lack of space. Detecting the **accelerations** of a boat in a stable way requires to evaluate the speed over a relatively long period of time. Here, the speed at each point of a track is compared against previous ones, in particular those that are more recent than a given temporal threshold (fixed to 2 minutes in our experiments). If the present speed is higher than the minimum previous speed in such interval by more than a fixed threshold (fixed to a very small 0.55 m/s) and more than a fixed percentage (fixed to 20%), then the current point is considered an acceleration point. **Noodle shapes** represent wandering movements around a fixed area, for instance waiting for the fish to bite. In terms of trajectories, that results in forming very entangled shapes. The heuristics we adopted consists in computing the ratio between the length of a track segment and the actual distance traveled by the boat – as the crow flies. Entangled trajectories tend to have a high ratio, and we experimentally set a threshold for this parameter to 2.5, i.e. the track between two endpoints is 2.5 times larger than their spatial distance. Clearly, noodles might appear only in small segments on a component, and therefore we need to look for appropriate intervals within the component. We scan all the points of the component, and

for each of them look ahead for the first point that gets farther than 2km from the starting point. This interval is then evaluated as described above, and if a noodle results, it is recorded and the scan proceeds from the endpoint (the noodle segment is skipped). Otherwise, the scan goes on. Notice that the 2km threshold (a input parameter) defines the granularity of noodles, i.e. the geographical scale at which look for noodle shapes. Small values (e.g. 50m) would capture very local wanderings, rather similar to stops. Large values (e.g. 50km) would miss fishing spots, while capturing larger exploration areas.

## 4 Modelling step 1: Components clustering

Our objective, here, is to build a catalog of representative user behaviours by grouping the components into homogeneous clusters, i.e. the components in a cluster are similar to each other.

The results will be used later for two purposes: first, clusters will be used to extract a stratified sample of data to be used as training set – after that domain experts analyzed and labeled them appropriately; second, the cluster representatives will be a key part of the activity recognition process, since it will indeed start with associating each component of the new tracks (to classify) to its most similar cluster.

### 4.1 Correlation-based features selection

In order to identify possible issues of hidden redundancy among features, which might affect the creation of clusters, we computed all pair-to-pair Pearson's correlation coefficients, and used the corresponding correlation graph to select a subset of redundant features to remove. Figure 1 reports the strongest relations discovered (cyan=between 75% and 90%, black=above 90%), either positive or negative, and the features discarded (red crosses).

### 4.2 K-means clustering

Among the several alternative clustering methods available in literature, we chose to adopt the simple K-means algorithm, which is efficient and provides results that are easy to interpret – both features were very important for this phase of the project.

K-means [4] seeks K clusters that contain objects as much homogeneous as possible, in such a way that all objects in a cluster are usually rather similar to each other, and rather different from those in other clusters. Similarity between objects is computed as Euclidean distance between the feature vectors for the objects. The algorithm also outputs a centroid for each cluster, i.e. a synthetic objects whose features average those of all the objects in the cluster.

The value of parameter K has to be provided by the analyst. Since the preliminary exploratory experimentation we performed did not let emerge any critical

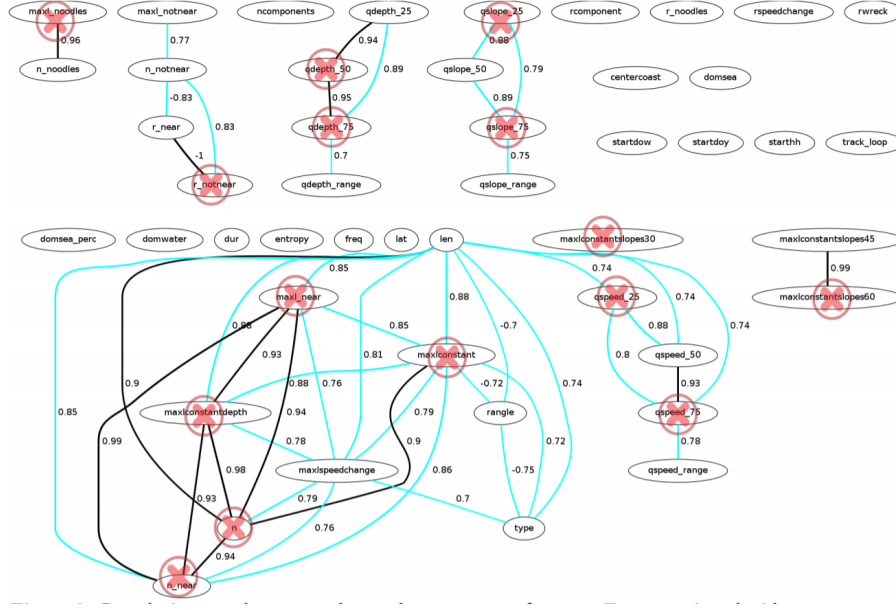


**Fig. 1.** Correlation graph computed over the component features. Features signed with a cross were later removed to keep only relevant and non-redundant information for the clustering step.

value of  $K$ , we chose to set  $K=100$  as a trade-off between our two main objectives: (i) have enough clusters to capture the different possible users' behaviours, and (ii) keep the number of clusters small enough to make it feasible, for a domain expert, to observe and label a reasonable number of sample components that belong to each cluster.

Experimental results on our data show a set of rather well-balanced clusters (see size distribution in Figure 2), with very few small ones and no *giant component*.

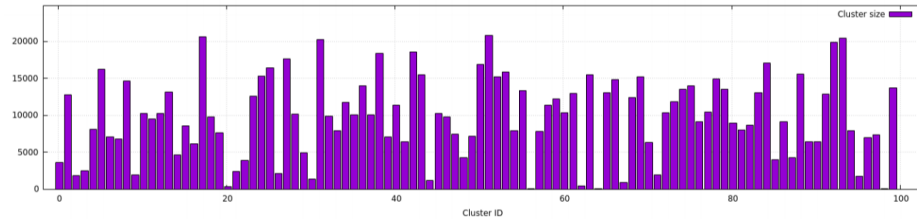


**Fig. 2.** Size of the 100 clusters obtained.

## 5 Modelling step 2: Component classification

The clusters found in the previous phase of the process represent significant behaviours of the users, yet it is still unknown to us exactly what type of activity corresponds to each cluster. This step involves the knowledge of a domain expert, which will be exploited to build the final classification model.

### 5.1 Training set construction and division in to world areas

Our objective, here, is to identify a reasonable set of tracks (and corresponding components) to be labeled by domain experts. In order to characterize our clusters, we start selecting a fixed number of components for each cluster (in our case set to 5), in particular focusing on components that are close to the representative centroid. That ensures both their representativeness and the fact to consider, in a limited amount, the variability of behaviours within each cluster.

Also, since similar behaviours might have different meaning or different weights in different areas of the world, the whole learning process will be differentiated over various sections of the world. In particular, the world has been divided into 6 macro-areas: USA East coast (USE), USA West coast (USW), Australia (Aus), Mediterranean sea (Med), Scandinavia (Scand), and United Kingdom (UK).

For each geographical area and each cluster, then, the 5 components closest to cluster centroid are selected, together with the complete track they belong to. This yields 6 different training sets to be labeled by the domain experts, each composed of around 500 tracks and their associated components. Our experiments involved Navionics experts of the areas above, who manually analyzed the track data and assigned an activity label to each component and track.

### 5.2 Clusters labeling

The information provided by the experts over the sample of components they labeled is propagated to the clusters they belong to in the following way.

For each cluster we compute a probability distribution over the set of possible activities, obtained mixing the two levels of information provided by the experts:

- component-level: the number of components with that specific activity label;
- track-level: the number of components that belong to a track having that activity as overall label.

For each activity, the two counts above are merged through a weighted sum. The optimal setting obtained empirically consists in a weight of 0.85 for track-level labels, and 0.15 for component-level ones. That means that the label of the overall track is the most important information to recognize activities, yet the labels of the components still contribute significantly to it.

The reason for keeping a distribution of probability for a cluster, instead of just picking one representative activity, is that often the secondary activities have a significant probability, and therefore discarding them would artificially

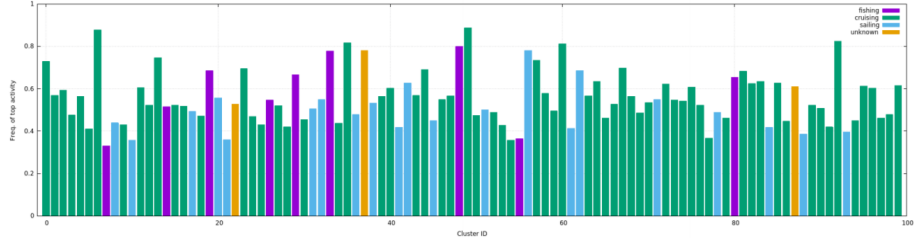


**Fig. 3.** Most frequent activity of each cluster and its frequency.

polarize clusters and loose information. This is supported by Figure 3, which shows for each cluster the dominant activity and its probability.

From the plot it is possible to see that cruising dominates in several clusters, meaning that it is an activity capturing a large variety of behaviours, whereas fishing and sailing tend to have a lesser variability. Also, we can see that while some clusters are almost pure (they contain close to 100% of a single activity), most of them are quite mixed, leaving a lot of space (in most cases around 50%) for the secondary activities. Finally, a few clusters (orange in the figure) mainly contain components of unknown activity. No cluster was dominated by the canoeing activity, which was expected due to its low presence in the training set provided by the domain experts.

The process has been performed separately for each geographical region, thus yielding a set of labeled clusters for USE, one for USW, etc.

### 5.3 Clustering-based meta-features for tracks

The labeled clusters obtained joining clustering and expert labeling provides a way to map any component to a set of potential activities performed. This operation can be performed by simply associating the component to the closest cluster, and then take its probability distribution.

Since our final objective is to learn to classify tracks, and not single components, we aggregate the distributions of all the components into a single track-level distribution. This is obtained by just averaging the probability of each activity label across all the components.

This final distribution might already be considered a classification of the track, the idea being to associate it to the predominant activity. Yet, in order to make our model flexible and to exploit all the information sources we have, we use this first "pseudo-"classification as an input for a later machine learning step that integrates also other information. These inputs, that we call "meta-features", include the following data:

- predominant *activity* label and probability distribution over all activities;
- predominant *boat type* (another information provided by domain experts labeling) and probability distribution over all boat types;

- predominant *water type* (salted, fresh, salt lakes) and probability distribution over all water types;
- predominant *area type* (also these were provided by domain experts: near-shore, off-shore, intra-coastal) and probability distribution over all areas.

#### 5.4 Training set: predictive features and target of classification

The final classification model is built over a set of predictive features composed of all the meta-features described in the previous sections, plus some additional attributes derived from them. These extra attributes have the purpose of providing the classification models a list of explicit comparative measures of (the probabilities of) some key activities that might otherwise be not detected automatically. This list includes:

- $P(\text{fishing}) - P(\text{cruising})$ : this allows to understand how much predominant is one activity over the other.
- $P(\text{fishing}) - P(\text{sailing})$ : same as above.
- $P(\text{anchored}) + P(\text{docked}) + P(\text{drifting})$ : this sums up a set of activities that were deemed significantly distinct by the domain experts, yet they all imply a stationary behaviour, virtually equivalent to a stop.

#### 5.5 Classification model

In this work we adopted the well-known and widely adopted Decision Tree model, consisting of a tree structure that starts from a root and follows some specific sequence of branches according to the values of some specific attributes.

A big advantage of using decision trees over other solutions is the readability of the model, which allows to check its sensibleness and better debug it.

The algorithm adopted for inferring a decision tree out of the training set is the standard C4.5 algorithm, which builds the tree in an incremental way, starting from a trivial root-only tree and expanding it by adding splitting nodes (basically the tests that create branches in the tree, such as is  $P(\text{fishing}) > 0.6?$ ). Also, in order to evaluate some performance indicators of the models constructed, we adopted a cross-validation approach, where (a large) part of the training set is used to build the model, while another (smaller) part is used to check how well the predictions of the model fit the actual labels.

The two main parameters of the algorithm, namely the minimum number of objects in each leaf and the confidence factor of leaves (i.e. how much should the dominant label of a leaf predominate on the others) have been set by a simple grid search, choosing the combination of values that maximize the accuracy computed through cross-validation.

#### 5.6 Performances of the models

Figure 4 summarizes some characteristics of the models obtained on each area of the world.

| Area  | min_leaf | missed classes         | Accuracy |
|-------|----------|------------------------|----------|
| USW   | 7        | (canoeing)             | 70.21%   |
| USE   | 7        | -                      | 71.35%   |
| UK    | 13       | ?, (canoeing)          | 66.11%   |
| Scand | 5        | ?, (canoeing)          | 65.81%   |
| Med   | 19       | ?, (canoeing)          | 64.53%   |
| AU    | 11       | sailing, ?, (canoeing) | 69.01%   |

**Fig. 4.** Summary table of the decision trees characteristics.

The table shows that the accuracy ranges between 64% and 71%, which looks reasonably good considering the high number of classes involved (five, including the “?” class). Also, in most cases the models were not able to capture the *canoeing* class simply because the domain experts did not provide a sufficient number of cases. In the Australia dataset, instead, *sailing* was not captured, although present in the experts’ labeled data. *USE* was the richest dataset, and indeed the model performs best.

## 6 Using the classifier

The ultimate objective of the project is to assign to all tracks of all users an activity label, exploiting the knowledge encoded in the classification models built in the previous sections. Then, through aggregation of the inferred labels we can assign an overall category to each user, based on what are the main activities (s)he performed.

### 6.1 Labeling new tracks

Given a dataset of new, unlabeled tracks, in order to to assign them to an activity, we basically need to follow the same preparation process performed during the model construction.

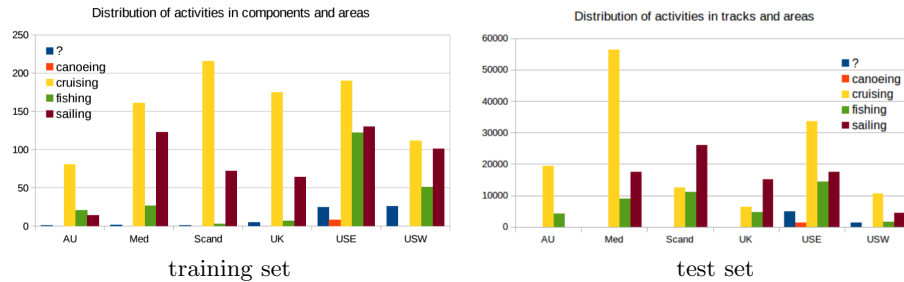


**Fig. 5.** Activity distribution: (i) training set (number of components assigned by domain experts to each activity); (ii) test set (number of tracks classified to each activity).

Figure 5(right) shows how the tracks were distributed across the different categories. That is compared against the distribution of components provided by the domain experts in the training phase, described in Section 5.1 (Figure 5(left)). As we can observe, the two distributions are rather similar. The main differences include the fact that *cruising* looks more present now in the USE, USW and Mediterranean areas, whereas it dropped dramatically in UK and Scandinavia. Also, as already noticed in previous sections, *sailing* completely disappeared in Australia, since its model did not capture that category.

An interesting view on the data can be obtained plotting the temporal distribution of the activities along the whole duration of the data we had access to, i.e. from May 2014 to April 2016, as shown below.

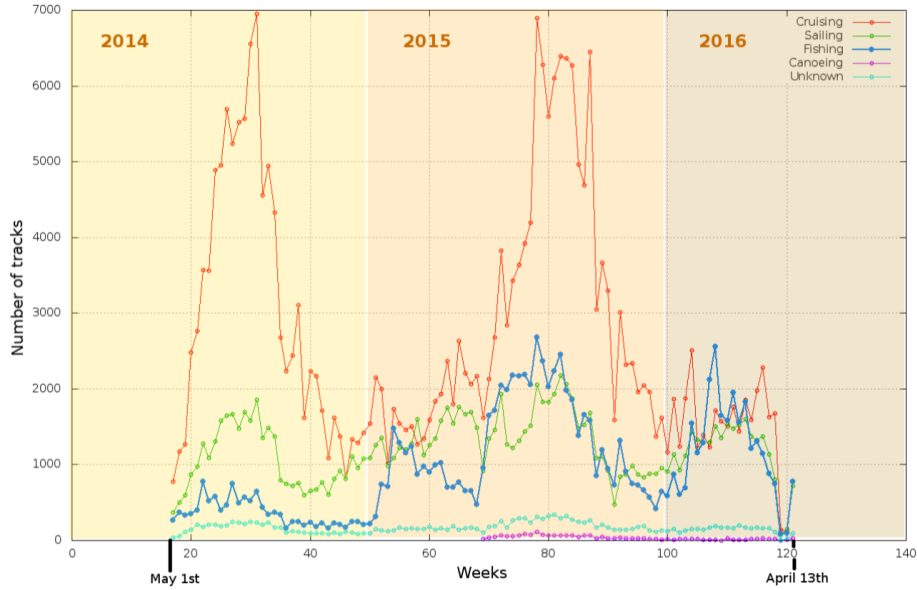


**Fig. 6.** Temporal distribution of activities along 2 years (mid-2014 to mid-2016).

In addition to the usual seasonal behaviours – overall increase of all activities in the summer months – we can observe that fishing increased sensibly its presence in the data during the last year. Possible causes might be an increased number of fishermen among Navionics users, or an increased propensity among fishermen to share their tracks, or a combination of the two.

## 6.2 Labeling users

The labels assigned to each single track can be simply aggregated (counted) to infer the distribution of activities for each user. The next step, then, consists

in selecting the activity – or activities – that represent(s) the user best. After some trials and evaluation of the results with the domain experts, the following approach was decided:

- If the user has a percentage of fishing tracks larger than 30%, we label the user as “fisherman”, since at the present fishing is considered a strategic segment of customers.
- Otherwise, the label with the largest percentage is selected, with no minimum thresholds.

### 6.3 Sample statistics of results

Figure 7 shows the distribution of users activities for each area. In the left plot we assign to each user the predominant activity, while in the second one we adopt the incentive mechanism for “fishing” mentioned above. Figure 8 summarizes the same information in tabular form and percent distribution.

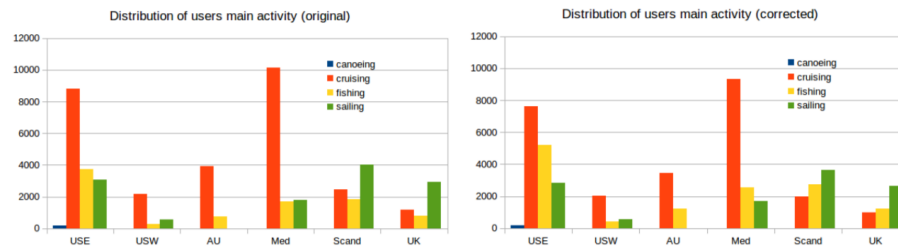


**Fig. 7.** Distribution of users’activity (left) and effects of fishing incentives (right).

| Original |        |        |        |        |        |        |       |
|----------|--------|--------|--------|--------|--------|--------|-------|
|          | USE    | USW    | AU     | Med    | Scand  | UK     |       |
| canoeing | 1.29%  | 0.00%  | 0.00%  | 0.00%  | 0.00%  | 0.00%  | 0.00% |
| cruising | 55.66% | 72.42% | 83.65% | 74.20% | 29.36% | 23.87% |       |
| fishing  | 23.70% | 8.50%  | 16.35% | 12.48% | 22.25% | 16.11% |       |
| sailing  | 19.36% | 19.08% | 0.00%  | 13.32% | 48.38% | 60.01% |       |

| Corrected |        |        |        |        |        |        |       |
|-----------|--------|--------|--------|--------|--------|--------|-------|
|           | USE    | USW    | AU     | Med    | Scand  | UK     |       |
| canoeing  | 1.15%  | 0.00%  | 0.00%  | 0.00%  | 0.00%  | 0.00%  | 0.00% |
| cruising  | 48.09% | 67.38% | 73.95% | 68.65% | 23.70% | 20.20% |       |
| fishing   | 32.84% | 14.52% | 26.05% | 18.88% | 32.72% | 25.07% |       |
| sailing   | 17.93% | 18.10% | 0.00%  | 12.48% | 43.58% | 54.73% |       |

**Fig. 8.** Percentage distribution of activities without and with fishing incentives.

As we can see, “cruising” remains the top activity in most areas, exceptions being Scandinavia and UK. Also, the small thresholding mechanism applied to fishermen yielded significant increases to their distribution.

## 7 Conclusions and Future Work

We proposed an analytical process to build a classification model able to infer the main water activities of an user having his recent traces. In particular the considered activities are: fishing, sailing, canoeing, and cruising. The methodology is presented and evaluated on a real use case with more than two years of worldwide data. Several challenges in managing such data are solved with general solutions making the process abstract and flexible. The final results can be used by the company to customize the navigation and mapping services to the user as well as propose a more effective target marketing initiatives. In the future we want to focus in obtaining better expert data in order to test the proposed methodology with an higher level of confidence in the different world areas. Moreover the original activities can be the initial nodes of an ontology of more specific behaviors which can be pushed in the system.

## References

1. Christophe Claramunt, Cyril Ray, Elena Camossi, Anne-Laure Jousselme, Melita Hadzagic, Gennady L. Andrienko, Natalia V. Andrienko, Yannis Theodoridis, George A. Vouros, and Loïc Salmon. Maritime data integration and analysis: recent progress and research challenges. In *Proceedings of the 20th International Conference on Extending Database Technology, EDBT 2017, Venice, Italy, March 21-24, 2017.*, pages 192–197, 2017.
2. Aldo Napoli Romain Gallen Alain Bouju Cyril Ray, Clement Iphar. Deais project: Detection of ais spoofing and resulting risks.
3. Erico N. de Souza, Kristina Boerder, Stan Matwin, and Boris Worm. Improving fishing pattern detection from satellite ais using data mining and machine learning. *PLOS ONE*, 11(7):1–20, 07 2016.
4. Xin Jin and Jiawei Han. *K-Means Clustering*, pages 563–564. Springer US, Boston, MA, 2010.
5. Lin Liao, Dieter Fox, and Henry Kautz. Extracting places and activities from gps traces using hierarchical conditional random fields. *The International Journal of Robotics Research*, 26(1):119–134, 2007.
6. Salvatore Rinzivillo, Lorenzo Gabrielli, Mirco Nanni, Luca Pappalardo, Dino Pedreschi, and Fosca Giannotti. The purpose of motion: Learning activities from individual mobility networks. In *International Conference on Data Science and Advanced Analytics, DSAA 2014, Shanghai, China, October 30 - November 1, 2014*, pages 312–318, 2014.
7. Charissa Ann Ronao and Sung-Bae Cho. Human activity recognition with smartphone sensors using deep learning neural networks. *Expert Systems with Applications*, 59:235 – 244, 2016.
8. Feixiang Zhu. Abnormal vessel trajectories detection in a port area based on ais data. In *ICTE 2015*, pages 2043–2049. 2015.