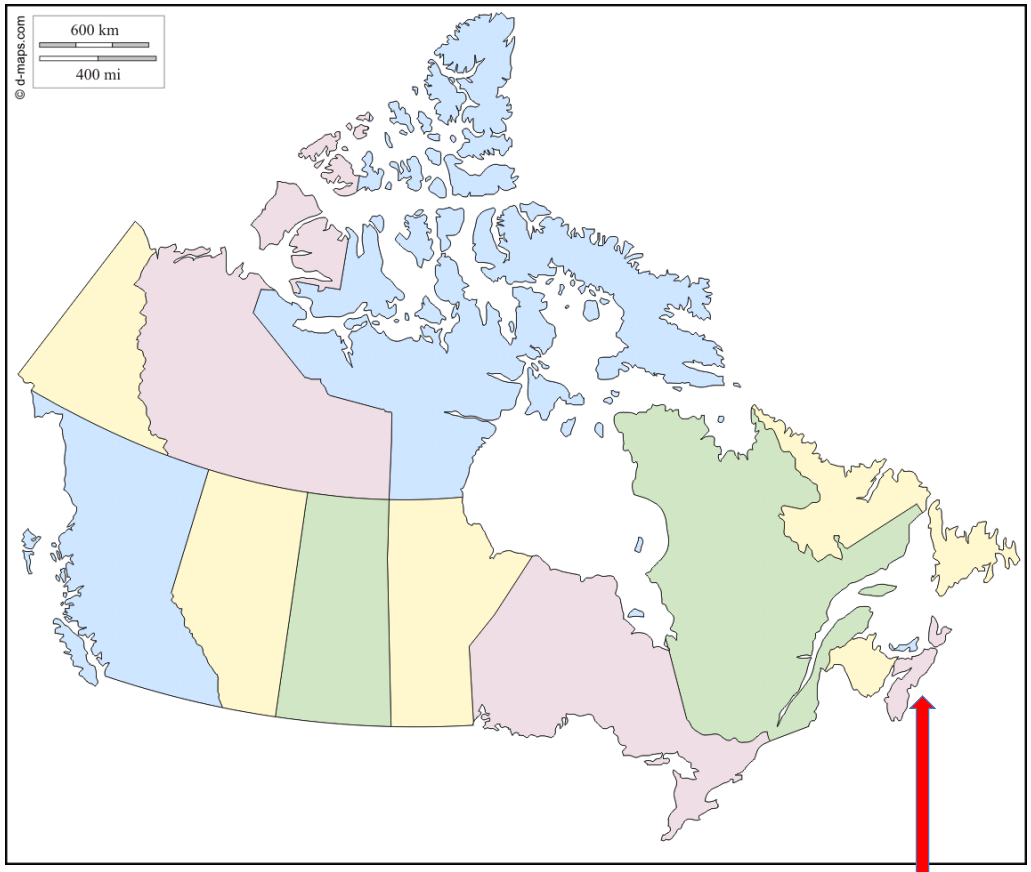
Spatio-Temporal Data Mining: methods and applications to ocean data.

Stan Matwin, CRC
Luis Torgo, CRC

Faculty of Computer Science
Institute for Big Data Analytics
Dalhousie University
Halifax, NS, Canada
stan@cs.dal.ca

ACM SIGIR/SIGKDD Africa Summer School on Machine Learning for Data Mining and Search
27-31 JANUARY 2020, CAPE TOWN, SOUTH AFRICA



Dalhousie University

Halifax

Plan : part I

- Spatio-temporal (STD) Data Mining
  - Definition of STD
  - STD properties
  - STD tasks
  - Ocean data – Automatic Identification System

Acknowledgement: parts of the material are based on **"Spatio-Temporal; Data Mining: A Survey of Problems and Methods"**, *G. Alturi, A. Karpante, V. Kiumar, ACM Computing Surveys Nov. 2017*

Plan : part II

Introduction to R and Spatio-Temporal Data in R

- Main packages and data structures
    - Importing spatiotemporal data into R
- Spatiotemporal data visualization using ggmap
- Interactive data visualization using leaflet
- Hands on with a case study

Definition of Spatio-Temporal Data (STD)

- Data capturing behavior of an entity in both time and space
- Presence of dependencies among instances introduced by the spatial and temporal dimensions
  [airplane trajectory example flightradar24.com]
- Consequences
- Examples
  - brain imaging
  - ship movements
  - wildfires

Properties

- Autocorrelation
  - Observations made at nearby locations and times are not independent, but are correlated with each other [Atluri]
  - Related to smoothness of change, eg SST temperature in nearby locations, changes in traffic from day to day
- Heterogeneity = non-stationarity [***not*** i.i.d.]
  - Data distribution changes with time (also known as concept drift): extra-terrestrial visitor example

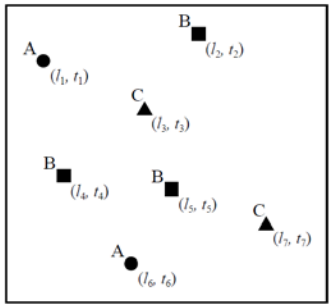
i.i.d.= independent, identically distributed

Main types of ST data

- Event data
- Trajectory data
- Point reference data
- Raster data

Event data

- Point in date and time, representing where and when an event occurred
  - A traffic accident, or a crime incident
  - A sick patient reported
  - …
- Collection of event points is a spatial point pattern



- Spatial point pattern
- types of events (A, B, C)
- granularity of “points” –
  eg polygons (forest fires)

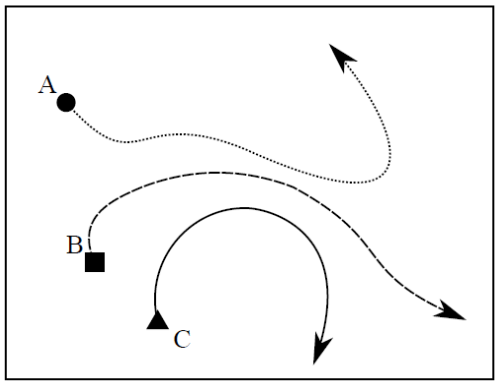
Trajectories

- Trajectory = path traced by a moving object in space and time
- $L = \{l_i = (x_i; y_i; t_i; o_i)\}$ is a set of all trajectory points
- definition of a raw trajectory τ_o ,where o is a moving object:

$$\tau_o = (l_0, l_1, \ldots, l_n)\ l_j = (x_j ; y_j ; t_j ; o_j)\ ;\ l_j \in L; 0 \le j \le n$$

where

$$\forall\ l_u, l_v \in \tau_o \text{ if } u \le v \text{ then } t_u \le t_v$$

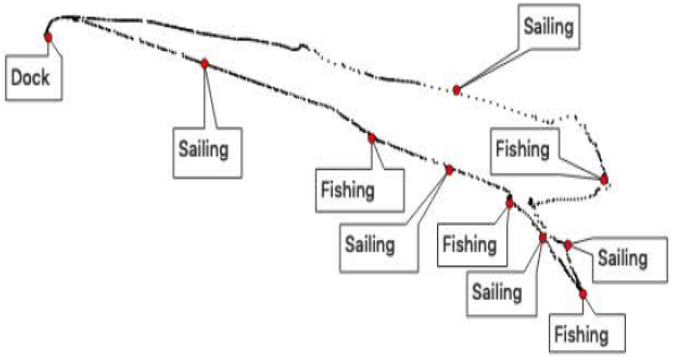


trajectories of three different moving objects: A, B and C

Trajectories

- Fundamental for mobility analytics
  - human – privacy; see eg [Gambs, Killijian, De-anonymization attack on geolocated data, IEEE TSP 2013]
  - vehicular
  - animal
  - semantic trajectories

human

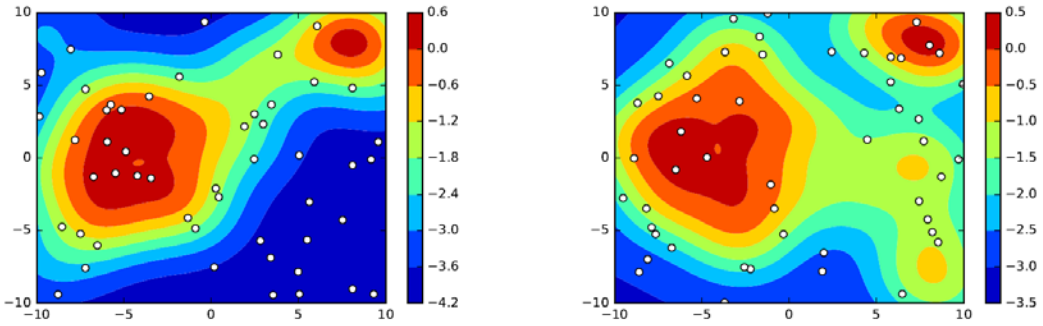


vehicle – fishing vessel

- trajectory data mining tutorial **CITE**

Point reference data

- Measurements of a continuous Spatio-Temporal field, e.g. temperature, vegetation, population, over a set of moving reference points in space and time. Locations and collection time stamps may differ
- Example a drone moving in the water (eg **ARGO project**); an underwater microphone array measu

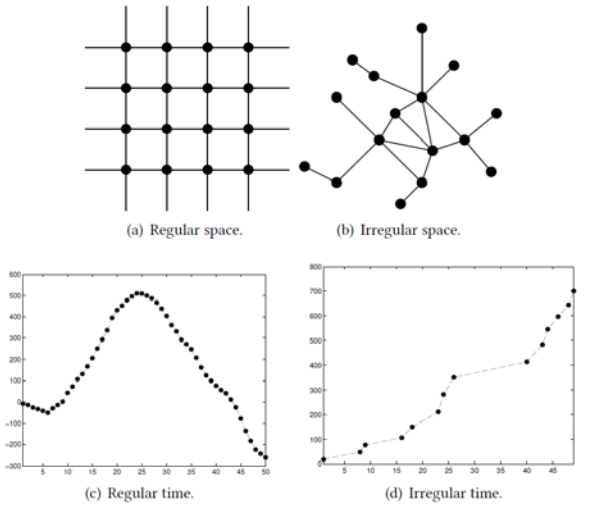


(a) Reference points on time stamp 1.

(b) Reference points on time stamp 2.

Raster data

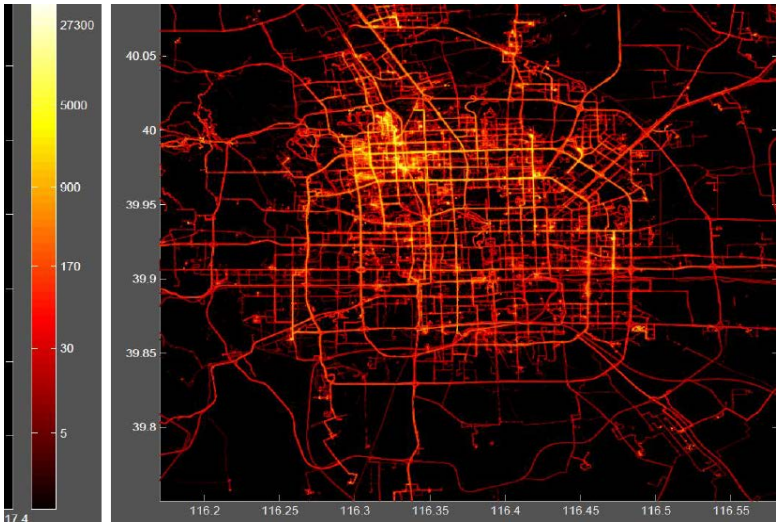
- Like point reference data, but locations and time stamps are the same.
- Locations and times may have coarser or finer granularity, eg a grid. Data aggregation is also common, eg fishing in an area over a month/year



Available trajectory dataset examples

- GeoLife
- AIS
- wildfires

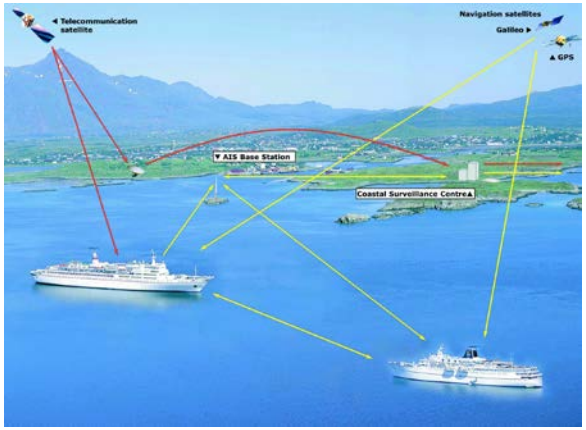
GeoLife dataset

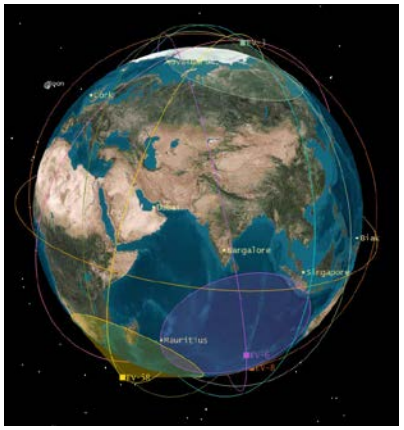


- GPS traces of people moving around a big city; has become a #1 benchmark dataset for mobility research
- Data openly available for download
- Collected in 2007-2011 mainly in Beijing, China, from 178 volunteers
- 90% of users recorded their coordinates very 1-5 sec.
- 69 users labeled their trajectories wrt transportation mode: walk, bike, bus, car/taxi, train, airplane, other
- 17K trajectories, 1.2M km, 48K hrs.

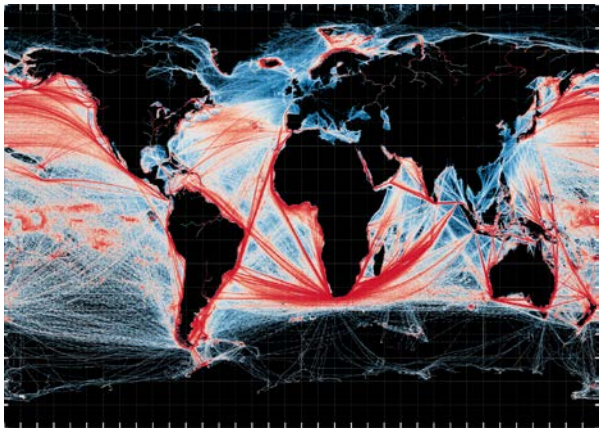
Automatic Identification System (AIS)

IMO/ITU standard





Courtesy of ExactEarth, Inc.



Institute for Big Data Analytics

400,000 ships
At least 100M records/day

Terrestrial *vs* satellite AIS

From weak to big signal

Wildfires dataset [event dataset]

- To be studies in the afternoon class
- 25,000 locations in Portugal, each described with a number of attributes about landcover, terrain, road density, and census data
- Over 10 years
- Whether an area was burnt or not

Tasks and methods: trajectories

- Segmentation of trajectories
- Clustering trajectories
- Classification
  - Segments
  - Trajectories
- Predictive learning
- Frequent pattern mining
- Anomaly detection

Trajectory segmentation

- The very idea: segment starts and ends where the trajectory changes rapidly (eg speed, or bearing, change)
- For a $\tau = (l_0; l_1, \ldots, l_n)$, we define a segment, S_i $i = 0,\ldots, k$, such that
- $\forall(S_i;S_{i+1})$ $S_i = (l_p; l_{p+1}; :::; l_{p+t})$; $S_{i+1} = (l_{p+t+1}, l_{p+t+2},\ldots, l_{p+t+u})$ and

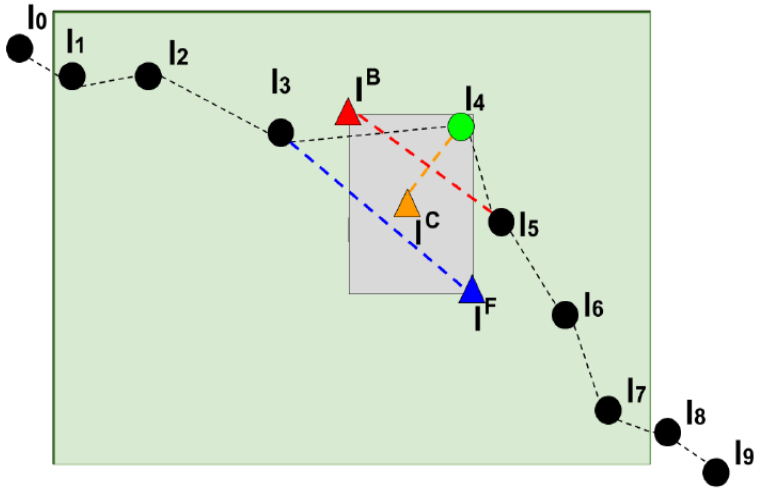
  $S_0 = (l_0; l_1 ,\ldots, l_x)$; $S_k = (l_y ,\ldots, l_n)$

Trajectory segmentation

- Fixed-size segment approaches
- Optimization-based approaches (eg optimizing Minimum Description length of a segment to increase its homogeneity; e.g. A. Soares et al. GRASP-UTS: an algorithm for unsupervised trajectory segmentation ,Int'L J. GIS, 2015)
- Clustering-based approaches (TRACLUS)
- Window-based approaches (e.g. M. Etemad et al., SWS: An unsupervised trajectory segmentation algorithm based on change detection with interpolation kernels, Geoinformatica, to appear)

Trajectory segmentation

- A window-based approach – forward and backward extrapolation
- Threshold-based



Clustering trajectories

- Clustering of a set of instances is partitioning of this set into groups, such that instances inside a group are **closer** to each other than to instances outside the group.
- Clustering locations has to take into account their spatial closeness
- Unsupervised learning

TRACLUS

An efficient trajectory clustering algorithm based on grouping similar trajectories **(Lee, Han, Wang, Trajectory Clustering: A Partition-and-Group Framework, SIGMOD 2007)**

Given: a set of trajectories T

- Segment elements of T
- Group similar segments with DBSCAN
- Find a representative trajectory for each cluster

Classification of trajectories

- Two kinds
  - Classification of segments [example - GeoLife]
  - Classification of entire trajectories [example - Gashin]
  - Learning a mapping from a set of labeled instances (segments+class) such that the mapping will perform well on unseen data
- Standard machine learning methods apply (eg HMM)

Classification of trajectories

- Of interest are methods that take autocorrelation into account:
  - Conditional Random Field, eg [B.Hu et al. : Identifying Fishing Activities from AIS Data with Conditional Random Fields. FedCSIS 2016 47-52]
  - Deep learning approaches, eg [X. Jiang et al TrajectoryNet: an embedded GPS trajectory representation for point-based classification using recurrent neural networks. CASCON 2017: 192-200]

Classification of trajectories

- For entire trajectories, instances can be treated as geometric shapes
- Then powerful, modern image classification methods (Convolutional Neural Networks) could apply …
- If there was enough data
  - M.Sc. Thesis by G. Ghzizadeh

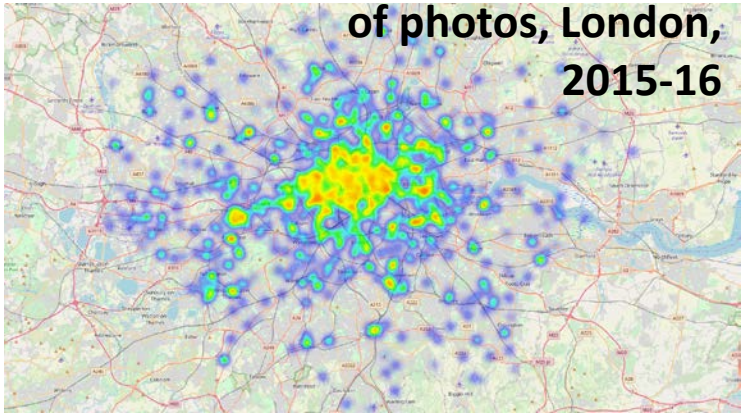
Predictive learning

- Time series prediction – next value
- Raster – use entire ST information to predict a scalar
  - Diagnosing mental disorder from fMRI or even audio recording
  - Predicting fishing catch from ST fishing history (more later)

Frequent pattern mining

- Co-occurrence pattern detection (Apriori-based, with a dedicated interestingnness measure)
- ST patterns in ST points:
  - *sequence of events*$_a$ → *sequence of events*$_b$
- Sequential patterns in trajectories: sequences of locations (or regions) that are visited by multiple moving objects in the same order, in multiple trajectories instances.
- eg tourists' behavior in a city, see e.g. [F. Vaziri et al., Discovering tourist attractions of cities using Flickr and OpenStreetMap data, Advances in Tourism, Technology and Smart Systems 2019, pp 231-241]

Spatial distribution of photos, London, 2015-16



Anomaly detection - trajectories

- Detecting anomalies (outliers) in ST domains can help identify interesting but rare phenomena(a ship in trouble, or making trouble, e.g. B. Hajsoleimani et al., Anomaly detection in maritime data based on geometrical analysis of trajectories, FUSION 2015 ☺)
- For instance, for ST points, such a point that breaks the “natural” ST-autocorrelation
- As always in anomaly detection, clustering is used to capture the “natural behaviour”, and events “significantly” distant from the cluster are likely anomalies

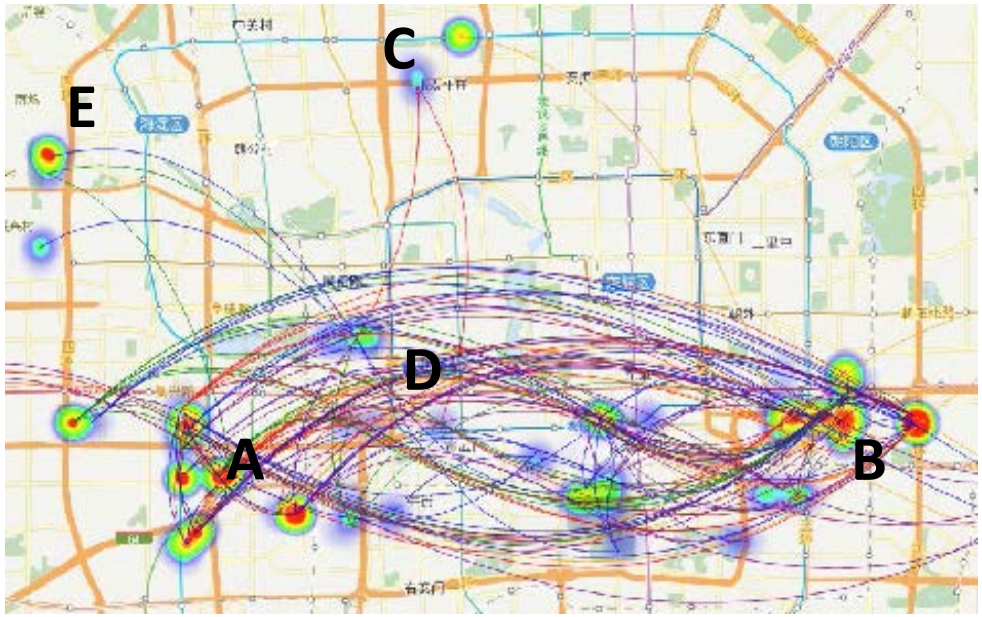
Finding… pickpockets in a subway/metro system

- Catch Me If You Can: Detecting Pickpocket Suspects from Large-Scale Transit Records, Du et al, KDD 2016
- Uses Automatic Fare Collection system

"behavioral differences are coined in the mobility footprints, separating suspects from regular passengers":

- traveling for an extended length of time,
- making unnecessary transfers, and/or
- wandering on certain routes while making random stops

- lots of passengers go from A to B
- Person going A → C → D → B is suspicious
- Person going E → D → B is an outlier – there are a few others

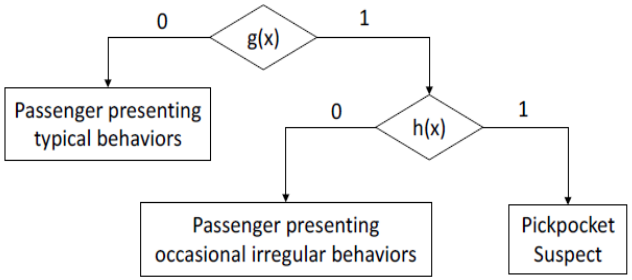


Challenges

- Highly imbalanced
- Need to avoid false positives
- What are the good features?
- Data: <ID, station, time stamp>; 6M passengers, 1.6B records
- ALSO known pickpocket incidents are made public on Weibo: **“At 7:40 a.m. on July 10th, a thief was caught at Route 349 East Chengzhuanglukou Station."**

Two-step approach

- Because of extreme imbalance, a two-step approach:



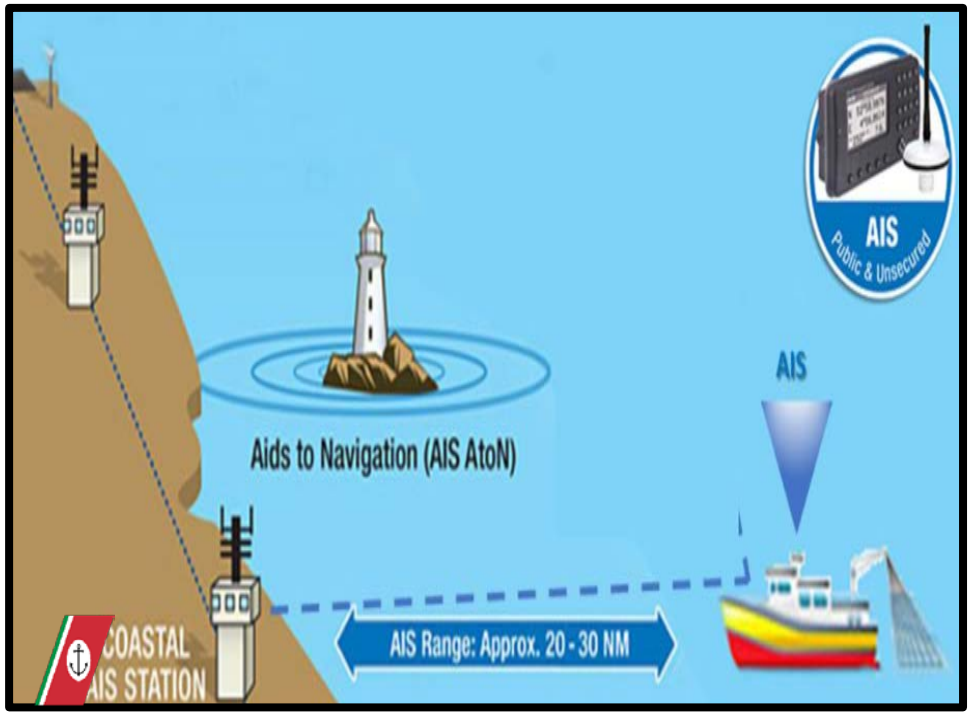
- Both *h* and *g* are implemented as one-class SVM, with the ame gaussian kernel. Training of *h* is on a much more balanced data than training of *g*

Results

- Measured with recall, precision, F-measure
- Recall: percentage of all positives classified as positives
- Precision: percentage of correctly identified positives
- Two-step approach produces precision ~7%, recall ~93%
- Low precision due to not all pickpockets being caught

AIS data

Collected by the local radio antennas



STATIC INFORMATION

- MMSI: Identification Number
- Name
- Call Sign (IRCS)
- Length
- Type Of Ship

DYNAMIC INFORMATION

- Ship's Position
- Time (UTC)
- Course Over Ground

Global fishing watch

- https://globalfishingwatch.org/
- Go see
- explain

Oceans

- Almost 70% of the surface of the planet
- **Source of 15% of protein for 4.5B people**
- livelihood supplement for **10%** to **12%** of world's population
- “the last frontier”
- quickly surveyed from space-borne and under-water sensors

AIS data – details and applications

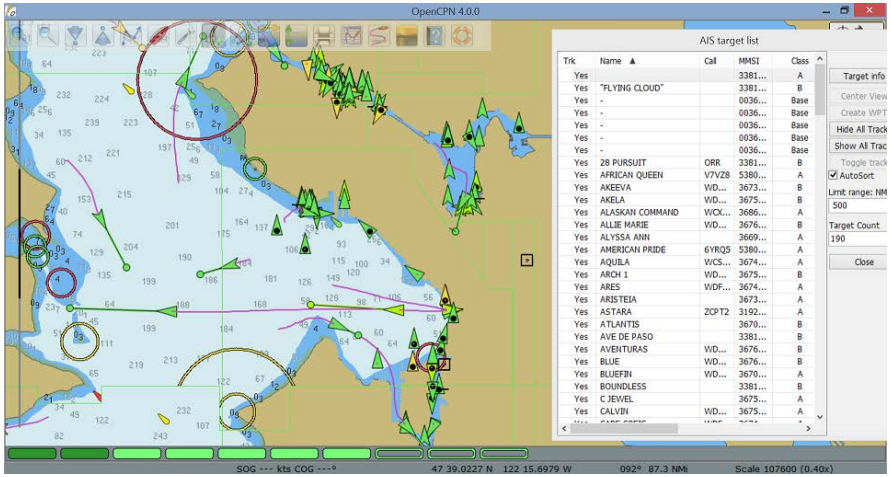
- You can become an AIS data collector and have access to global AIS coverage…if you live on the coast of a se or ocean
- All you need is a small investment (even a few hundred dollars) and an internet connection

Antenna and Hardware

- AIS antenna (Several options)
  - Shakespeare 5225XT Galaxy VHF Antenna, 8-Feet (Link)
- Raspberry-Pi 4 (Link)
- dAISy HAT
  - AIS Receiver for Raspberry Pi (Link)

Software and sharing

- OpenCPN (AIS receiver viewer)
  - Setup tutorial (Link)
- Decoder software (Link)
- **Share and Get**
  - MarineTraffic (Link)
    - Free antenna (Link)
  - AISHub (Link)
    - Dispatcher (Link)
      - Repeated and oversampled data

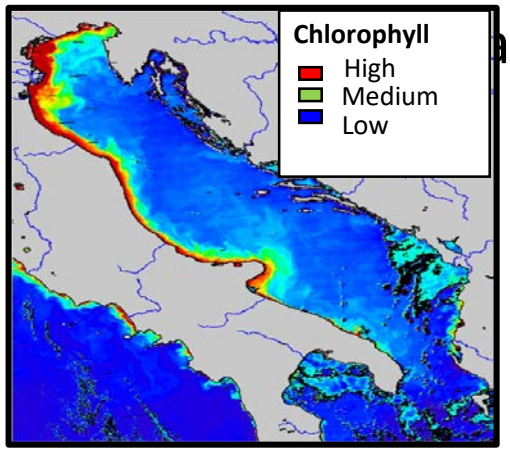


Predicting Fishing Effort and Catch
Using Semantic Trajectories
and Machine Learning, Adibi et al. ECML 2019 Workshop on mobility data

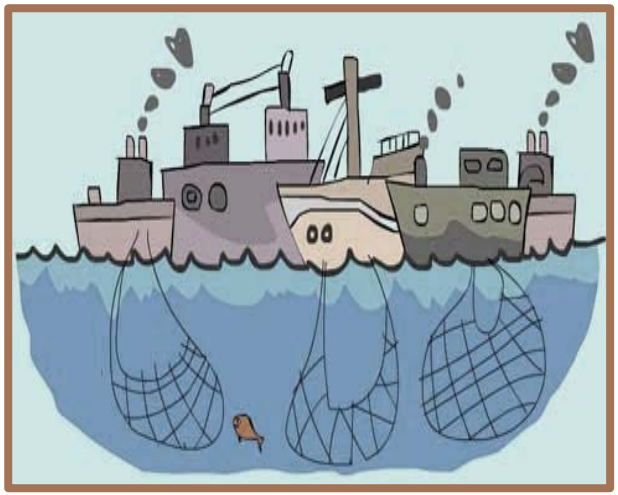
- Scenario & Motivations: Fishing in the Mediterranean
- Datasets: AIS data and Sold fish
- Methods
- Results

Scenario

The most productive area of the



Very high fishing pressure



One of the most intensively fished area in Europe

Multispecific & Multigear Fisheries

Northern and Central Adriatic Sea are Overexploited

Goals

- ✓ Improve the **knowledge of the spatio-temporal aspects** of the fishing activities in the Northern Central Adriatic Sea.
  - Spatial and temporal distribution of the **fishing effort**
  - Spatial and temporal distribution of the main **species**
  - Prediction of the Catch Per Unit Effort **CPUE**

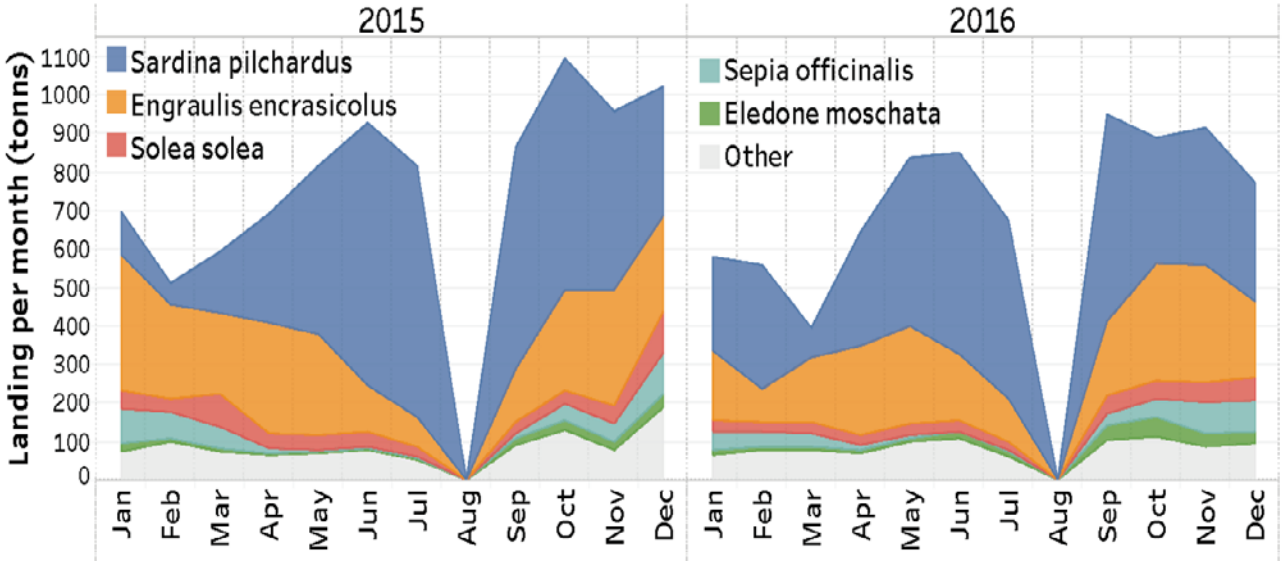
- Development of **effective fishery management** plans to reduce unsustainability of exploitation and ensure a productive and healthy ecosystem.

Datasets

- ✓ Terrestrial AIS data provided by the Italian Coast Guard for the Chioggia fleet of ships.
- ✓ Daily landing reports: amount of fish sold at the Chioggia Bulk Fish Market
- ✓ Environmental data:
  - Sea Surface Temperature (in Kelvin)
  - Sea Daily Chlorophyll-a Concentration (in mg/m^3)
  - Spectral significant wave height (in meters)

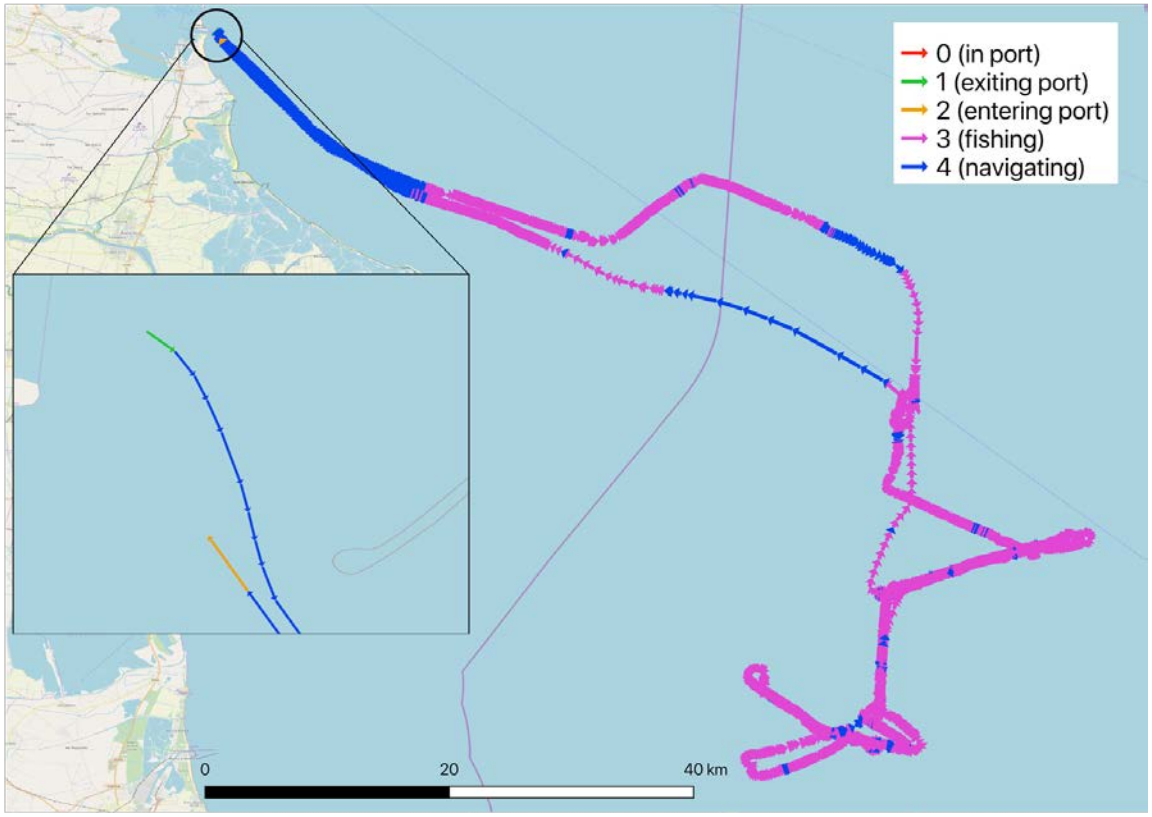
Landing data & catch distribution

- Daily landing reports for 2015-16 (amount of catch in Kg)
- 17921 fishing trips, 82 vessels
- Catch for a ship on a given day is assigned to the corresponding trip from AIS data
- Catch is distributed over the segments of the trip that are labeled as “fishing”
- Two types of distribution:
  - uniform
  - weighted



Trajectory construction & trip detection from AIS data

- Two year data (2015-16)
- Trajectory segments: straight lines between two consecutive AIS records
- Segments are labeled:
  - (0): in port
  - (1): exiting port
  - (2): entering port
  - (3): fishing
  - (4): navigating
- Trips are identified as the segments between exiting and entering the port



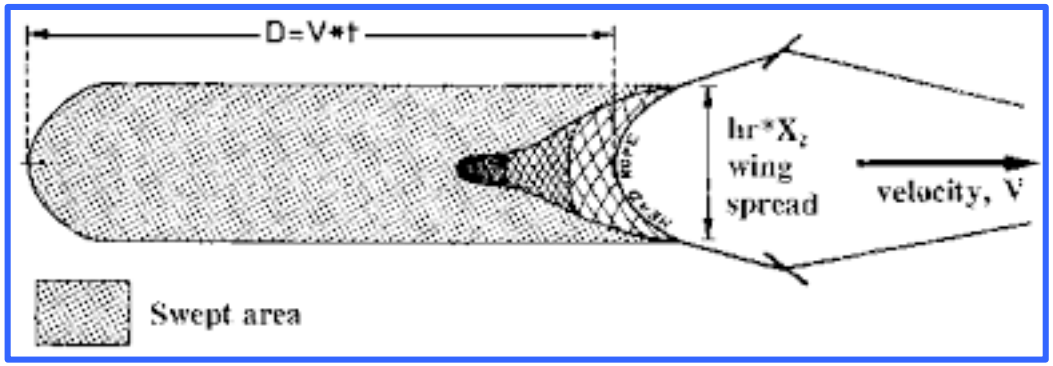
Fishing effort & Catch per unit effort

Area of interest partitioned into a square grid.

For a cell:

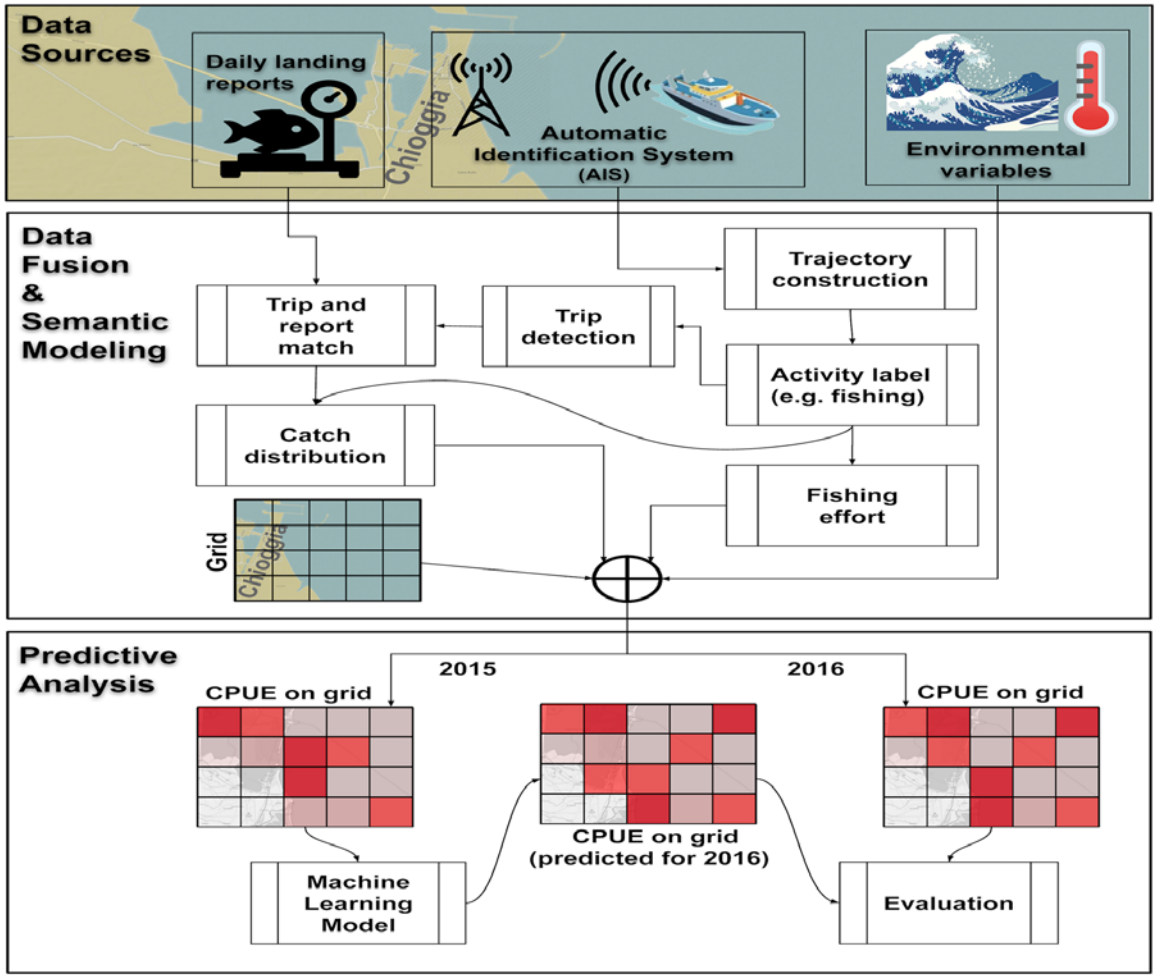
- Fishing effort = Swept area / area

trip length * net opening



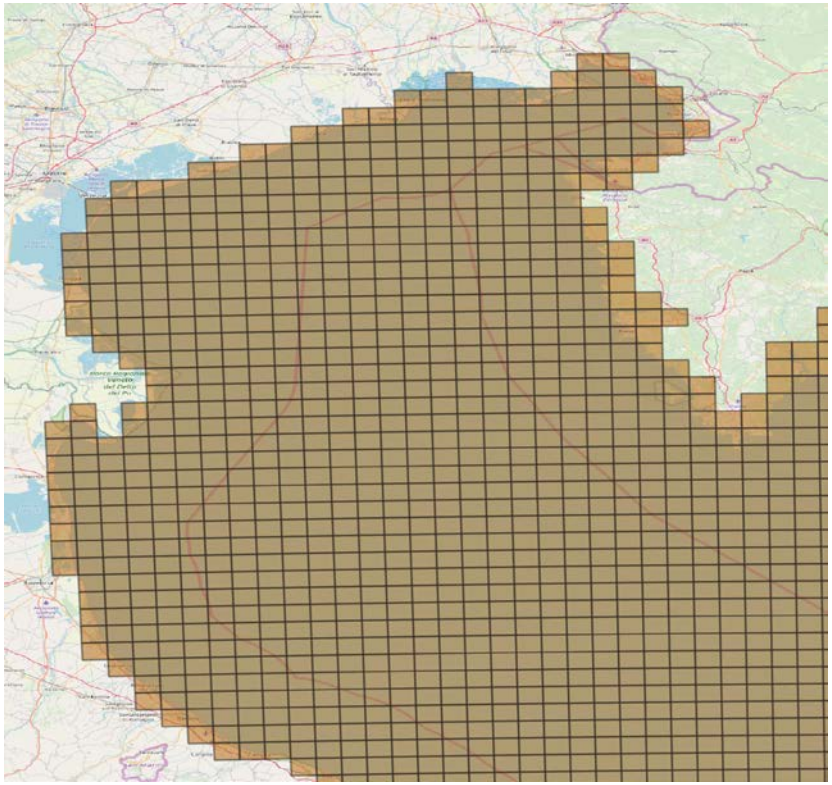
- Catch per unit effort (CPUE)= Caugth fish / Fishing effort

our solution - bird's-eyes view



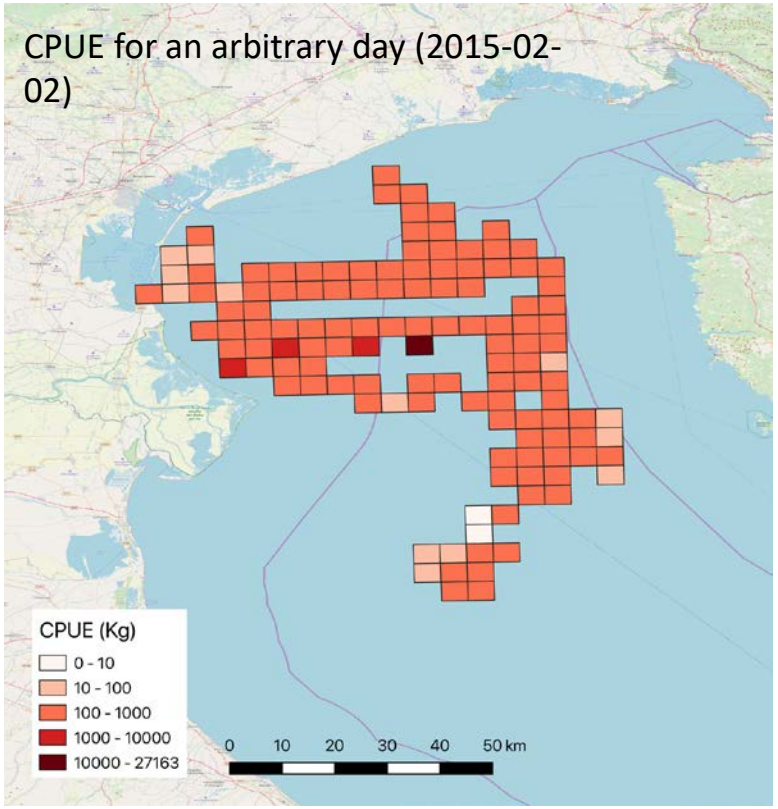
Spatial grid

- Data is mapped onto a 5x5Km spatial grid, which is used for
  - calculating fishing effort
  - weighted catch distribution
  - prediction modelling
- **Weighted catch distribution**: amounts of catch over each segment, derived using the uniform distribution, is weighted by the (daily) vessel counts in the cell containing the segment. Weights are normalized for each trip, so they add up to 1.



Modeling data

- Target attribute
  - CPUE (catch per unit effort) calculated per grid cell
- Predictive attributes
  - Environmental (sampled on grid cells)
    - chlorophyll concentration
    - wave height
    - sea surface temperature
  - Spatial and temporal attributes
    - cell centre coordinates
    - day of year, month, season, etc.



Evaluation method

- Relative Absolute Error (RAE): performance compared to baseline method
- For a given period $\boldsymbol{p}$

$$RAE_p = \frac{MAE_p}{MAE_p^*}$$

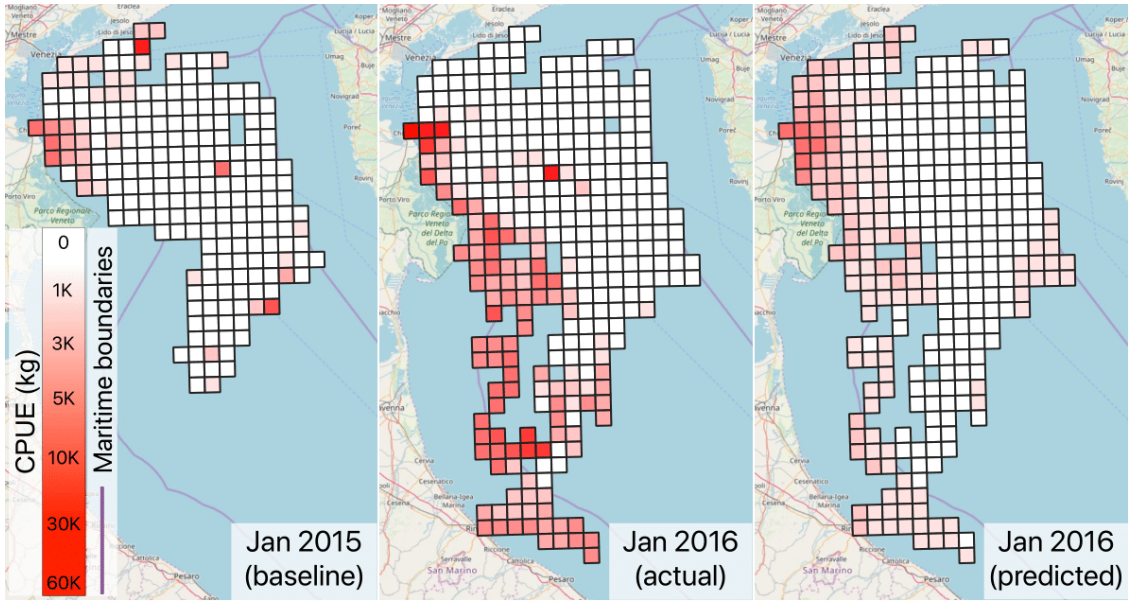
where $\boldsymbol{MAE_p}$ and $\boldsymbol{MAE_p^*}$ are respectively the *model* and *baseline* Mean Absolute Errors

- $\boldsymbol{RAE_{2016} = 0.87}$ (13% improvement compared to baseline)

Modeling & Results

- Random forest regression: trained on 2015, tested on 2016
- 13% improvement in *average monthly CPUE* compared to baseline prediction

- Baseline prediction: values from the previous year for that period; e.g. baseline forecast for Jan-2016 is values from Jan-2015



Limitations & challenges

- Temporal span: more data is needed
  - Seasonality is hard to capture by training on one year data
- No autocorrelation in space and time

Trajectory classification – fishing/no fishing detection [EN de Souza et al. Improving fishing pattern detection from satellite AIS using data mining and machine learning, PloS one, 2016]

- The problem
- The data
- Segmentation
- Classification
- Results

Fish ecology – AIS applications

TO MAP FISHING EXPLOITATION IN ALL THE AREAS OF THE WORLD

- Impact of marine protected areas on fish stocks
- Estimating total amount of fish taken out by large operators

For a Machine learnist:

develop models of fishing activities **from data**, especially:

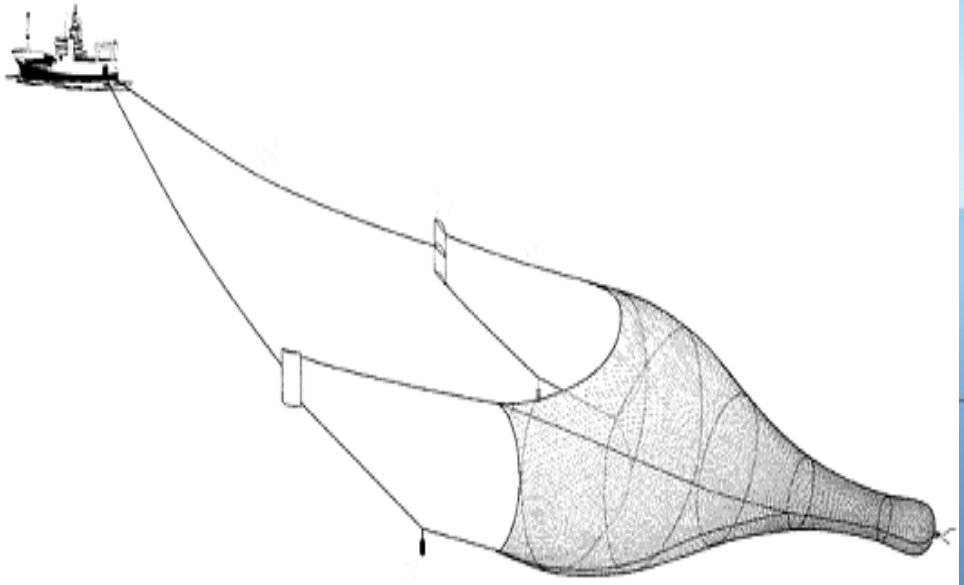
- classification of trajectories (AIS!) by ship type
- classification of trajectory segments by activity – fishing/no fishing

Ship trajectory representation is the key

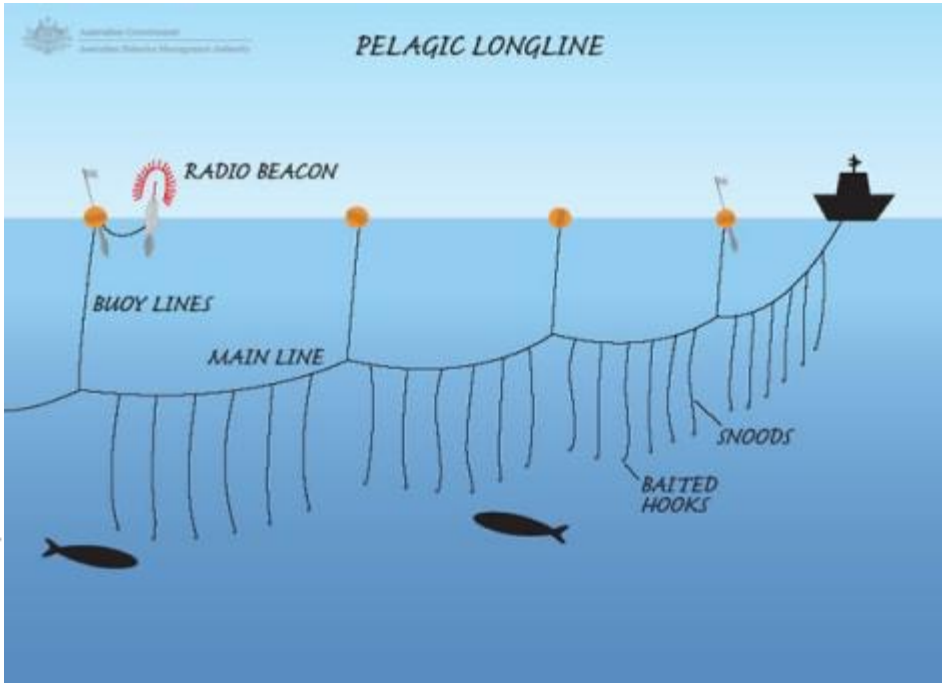
Trajectory representation and reasoning

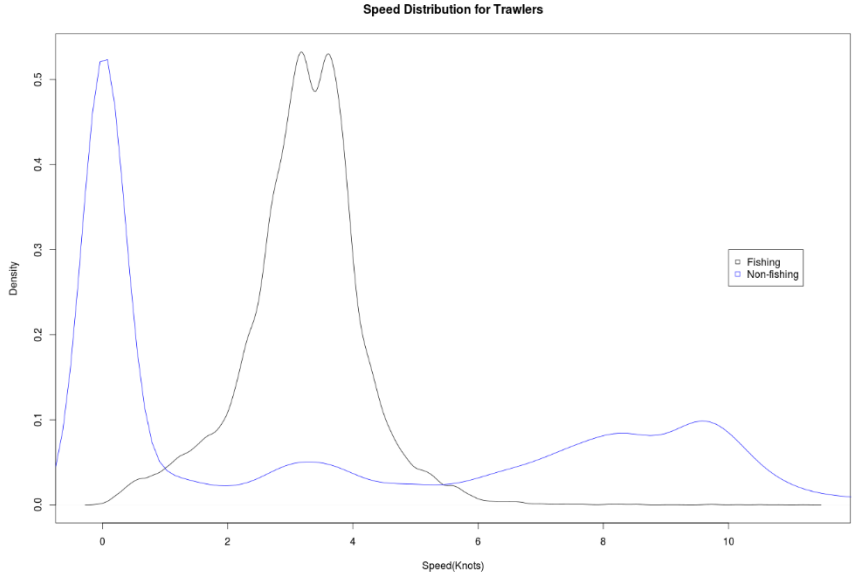
- Spaccapietra , et al. 2008
- GeoPKDD 2005-2008
- Palotta, Vespe and Bryan 2013
- ….
- Soares et al 2014

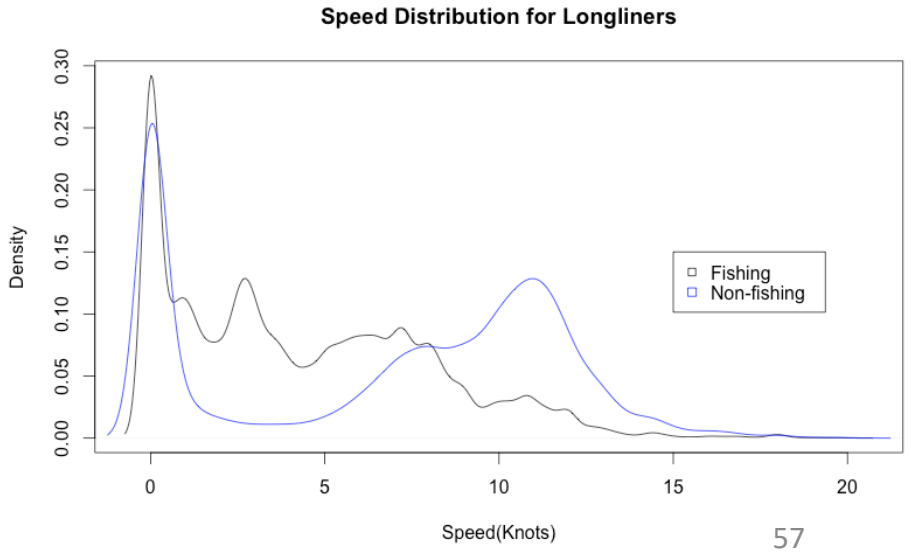
Trawler - 80% Not Fishing



Long Liner - 75% Fishing







Data

- Trawlers:
  - 83 vessels operating in the North Pacific
  - Training: 217,860 data points collected in July 2013.
  - Testing: seven vessels January 2011 - October 2015 across various ocean basins. 884,478 data points.
- Longliners: 16 vessels from June 2012 until December 2013,
  - 573,204 data points.
- Labeling: point-wise

Early results

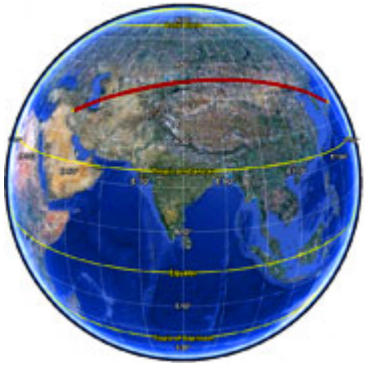
- **de Souza, Erico N., et al. “Improving Fishing Pattern Detection from Satellite AIS Using Data Mining and Machine Learning.” *PloS one* 11.7 (2016): e0158248**.
- Used Lavielle segmentation (habitat selection - “fishing ships move like predators”)

  Lavielle M. Using penalized contrasts for the change-point problem. Signal Processing. 2005;85(8):1501–1510

Basic navigational trigonometry….

- Bearing = angle between the true North and an object
- Because of the spherical shape of the earth, bearing changes as we move depending on longitude and distance

- if you were to go from say 35°N,45°E (≈ Baghdad) to 35°N,135°E (≈ Osaka), you would start on a heading of 60° and end up on a heading of 120°!).



- Moving from ϕ_1,λ_1 *to* ϕ_2,λ_2 *(*$\Delta\lambda$ is the difference in longitude) means starting on a bearing θ:

- $\theta = \text{atan2}(\sin \Delta\lambda \cdot \cos \phi_2 , \cos \phi_1 \cdot \sin \phi_2 - \sin \phi_1 \cdot \cos \phi_2 \cdot \cos \Delta\lambda)$

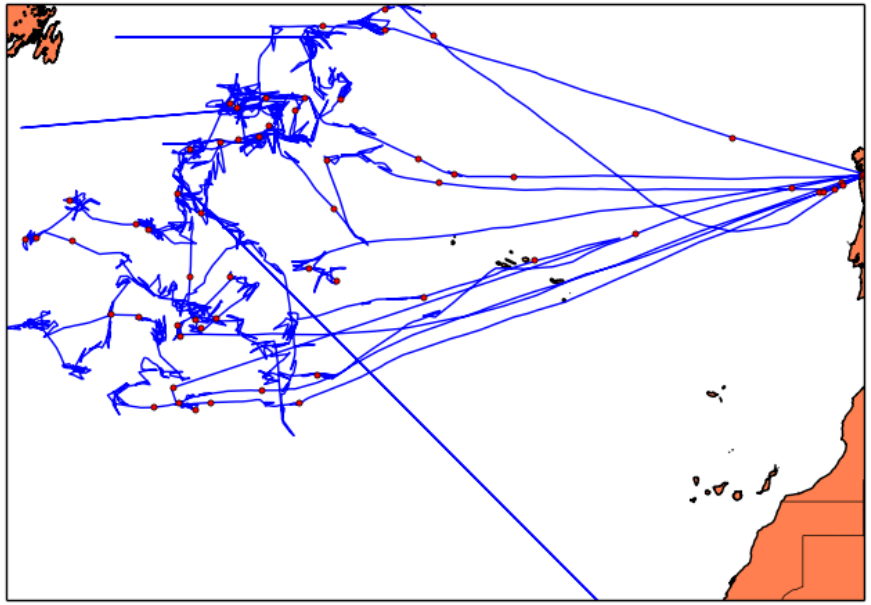
arctan with 2 arguments, to tell the quadrant from the sign

Segmentation of trajectories

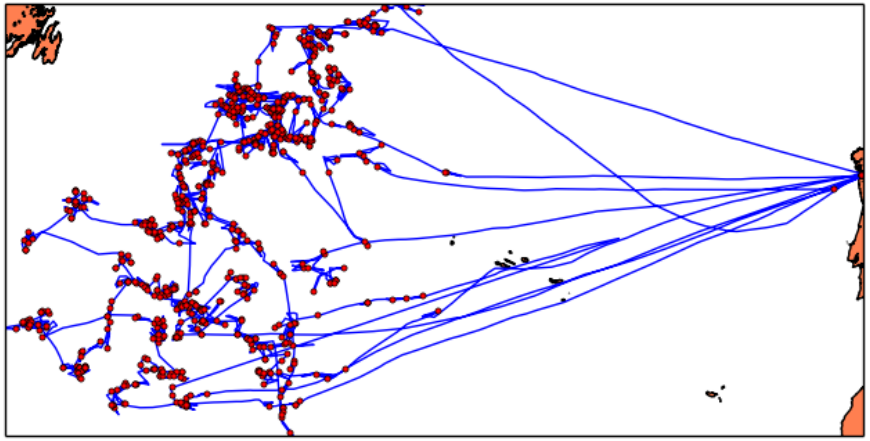
- Discretization of bearings into roughly 40 degree bins
- checking the bearing between each 2 points
- If they are in bins distant by more than 1, check the stochastic significance of the difference between θ with a t-test within a 5 point window, $p = 0.05$
- This process might create various incorrect segments; however, comparing with state-of-the-art Lavielle segmentation…

How does it translate to Segments?

Lavielle's Segmentation



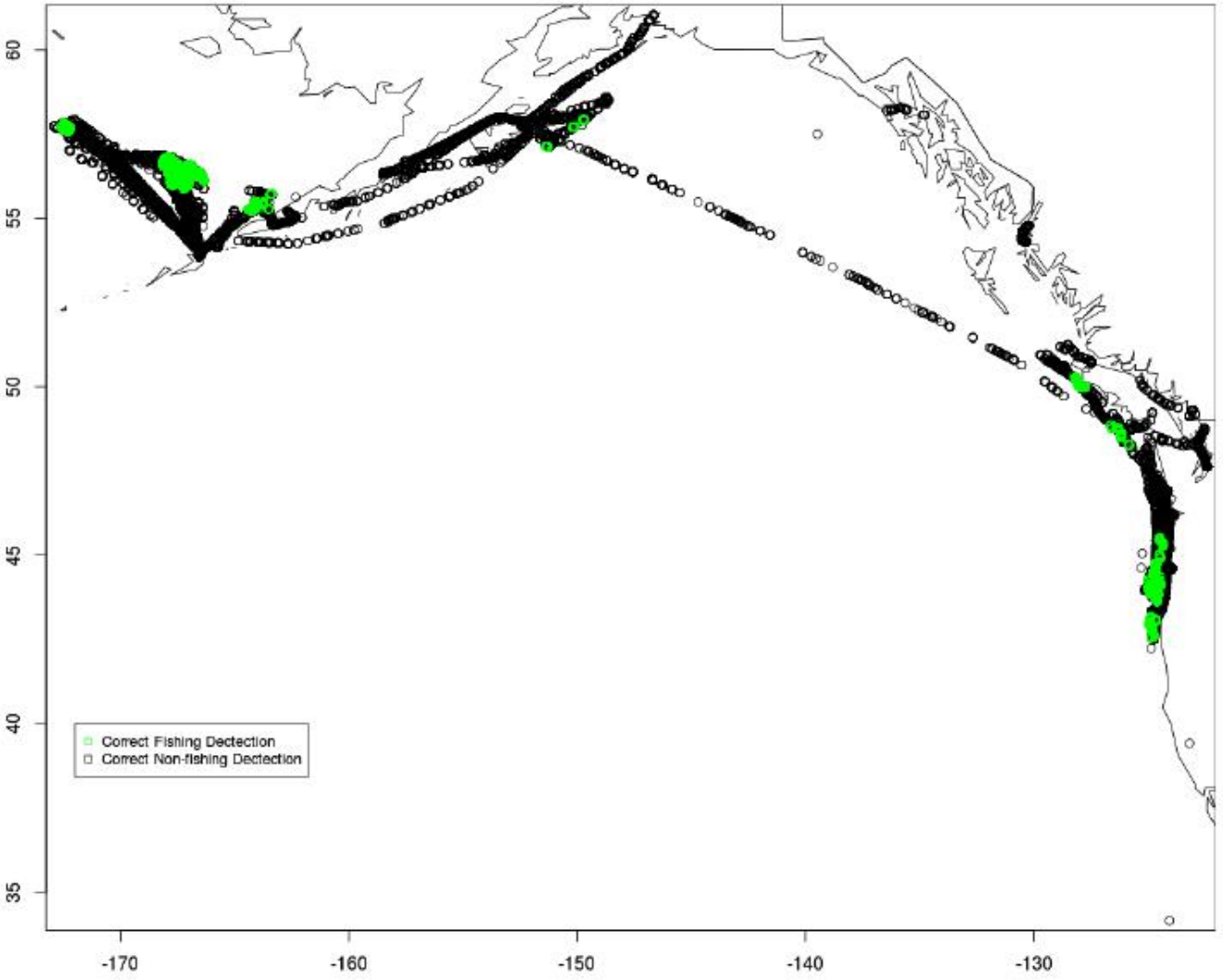
Proposed Approach



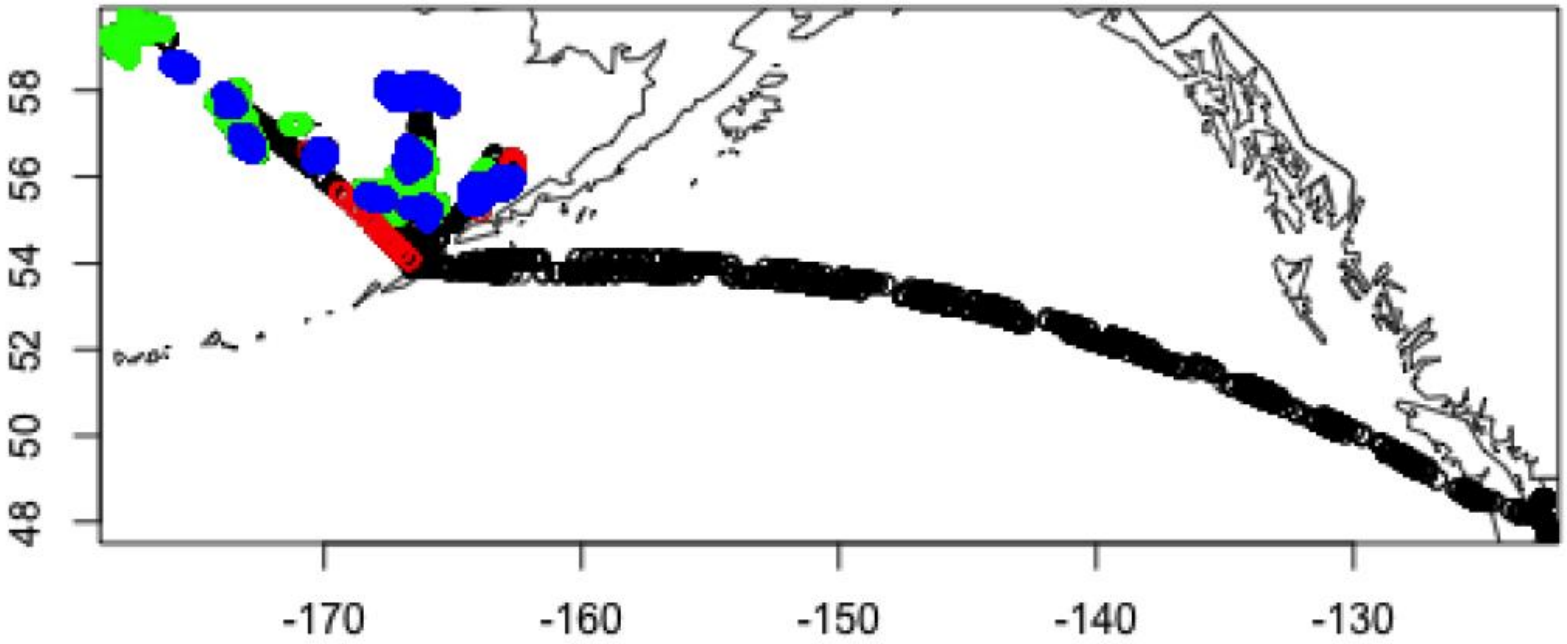
Machine Learning

- Points (position, speed, course) **PLUS** context info (averages of speed, acceleration, distance, etc. for the segment)
- Used ensembles of Decision Forests
- Long track trawler data limited, so
- Experiments on trawlers, but trained on six Long Liners with long trajectories
- With or without near shore (10 mi) information
- Results generally better than those reported in Plos

Trawler results visualisation



Long liner results visualization



(d) Resutls long liner vessel number 2 (Accuracy: 54%).

Fishing – expert and algorithm
Fishing – expert label
Fishing – algorithm label
Non-fishing - expert and algorithm

Future AIS work

- Data management issues (scaling of Big Data access, 1-3-6-15(?) TB…)
- Other machine learning approaches
  - Deep learning – RNN (see e.g. [X. Jiang et al., Improving point-based AIS trajectory classification with partition-wise gated recurrent units. IJCNN 2017: 4044-4051] for the first results)
  - Embeddings for trajectories [Gao, Q., et al. Identifying Human Mobility via Trajectory Embeddings, IJCAI 17]
- Visualization

Class summary so far:

- Spatio-temporal data:
  - Types
  - challenges
  - Typical tasks
- Trajectories
- Ocean vessel trajectories – AIS data source
- Two examples of ST applications of AIS data
- Next:
  - Spatio-temporal data in R
  - Visualization
  - Hands-on case study – forest fires data