ENGAGE PhD Study

Machine Learning Techniques for Seamless Traffic Demand Prediction

Manuel Mateos (PhD Student)

Xavier Prats (Supervisor), Oliva Garcia (Co-supervisor)

3 December 2019

Contents

Research overview

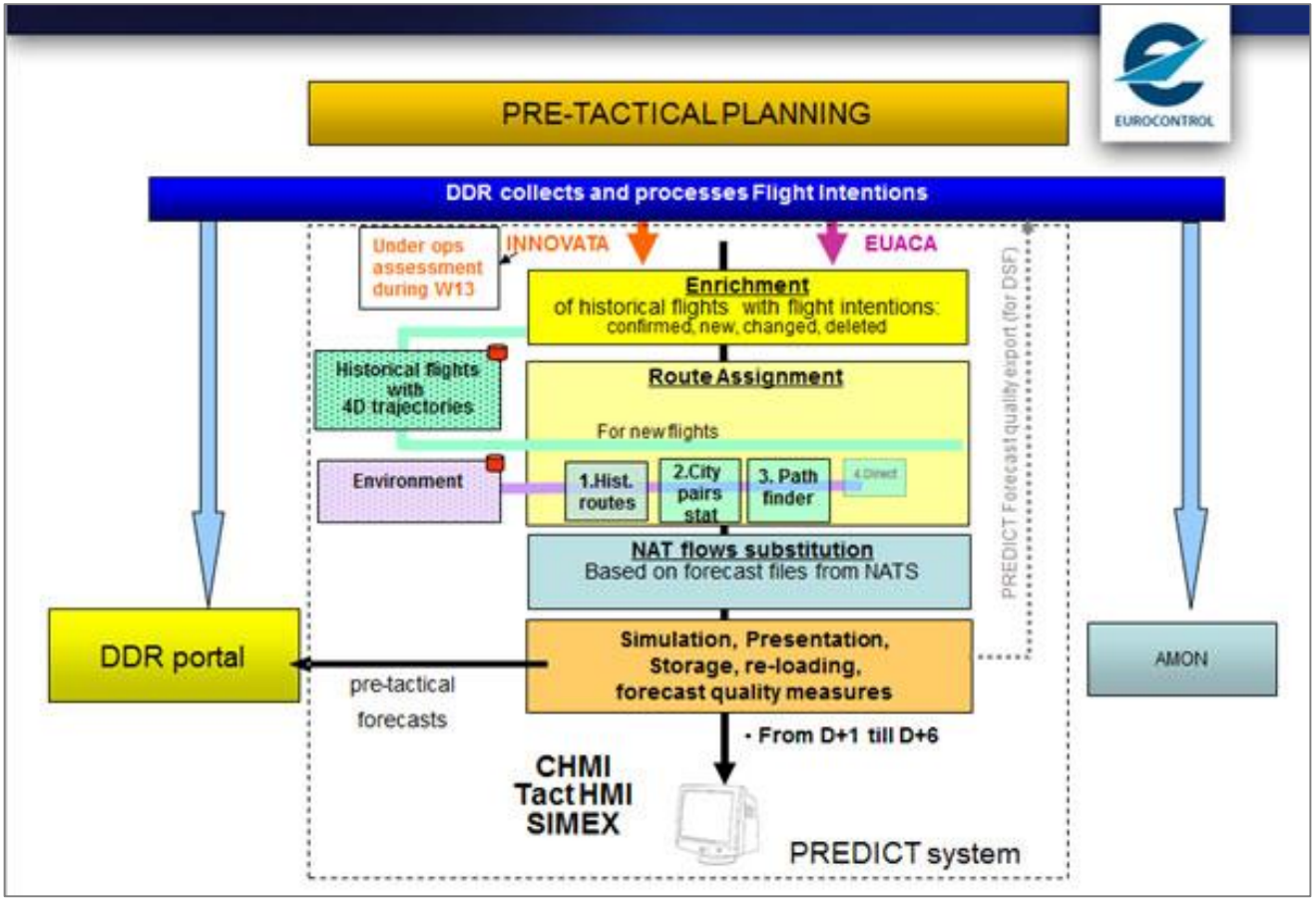
About the PhD

- Industrial PhD carried out by Nommon in collaboration with UPC
- Funded within “1st SESAR ENGAGE KTN Call for PhDs”
- Focused on improving demand prediction through machine learning techniques
- Supported by the Network Manager through data provision and review of results

Background and motivation

- Accurate demand forecast is one of the key enablers of ATFCM service provision
- Traditionally, trajectory prediction research has focused almost exclusively on short-term forecasts (tactical phase and operations), when FPLs are already available
- Current pre-tactical traffic forecast is based on a number of simple criteria about similarity with previous flights, overlooking certain relevant factors, such as meteorology and congestion
- Assumption: **demand forecasting methods for the ATFCM pre-tactical phase have significant margin for improvement**

Current pre-tactical traffic forecast: the PREDICT system



Proposed approach

The problem

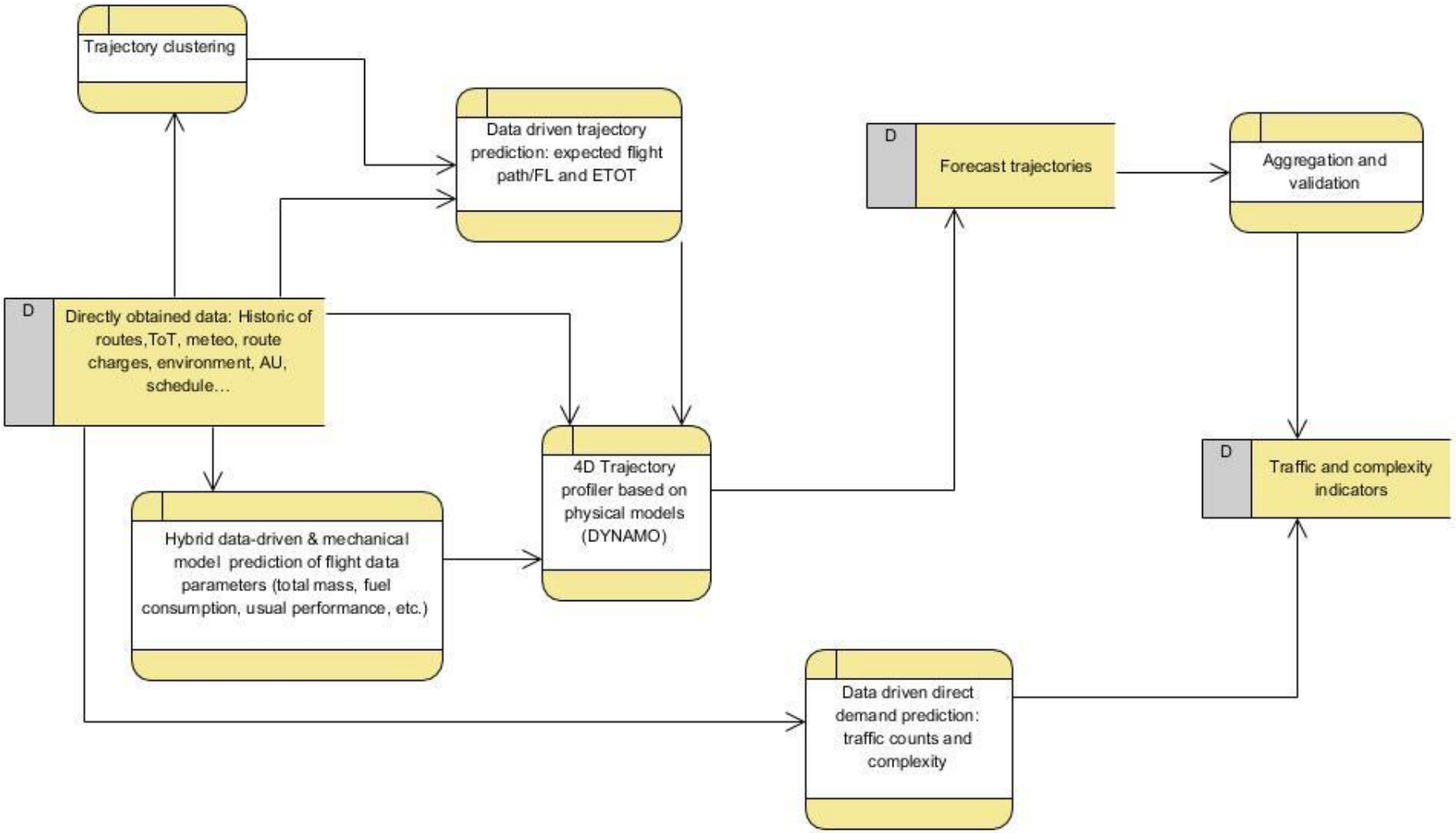
Objective:

Predict the First Filed Flight Plan before it is submitted by AUs

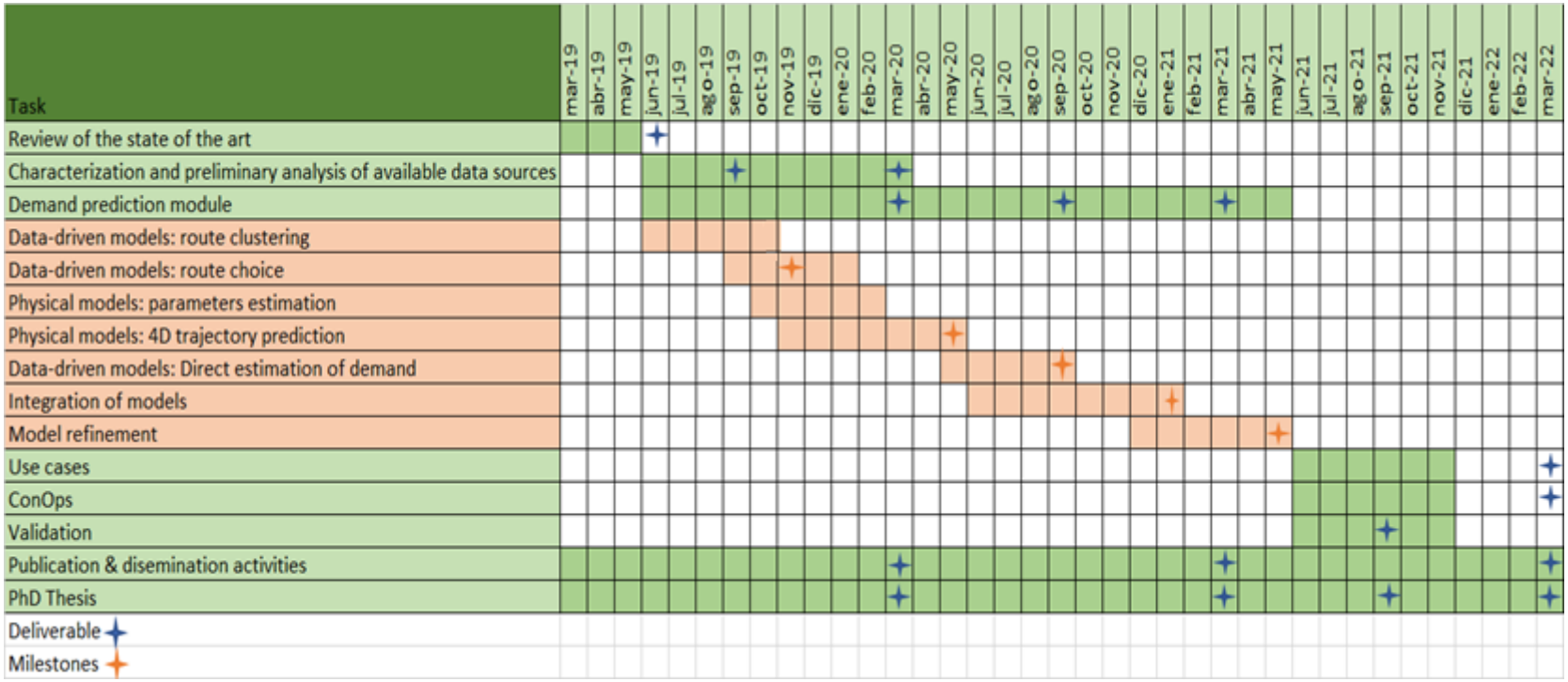
Approach

- Hybrid data-driven/physical model
- New sources of information and performance parameters will be included:
  - Meteorological prediction
  - Company preferences and configurations
  - Estimation of aircraft configuration
  - Route availability (Scenarios)
  - Disruptions/special events (e.g., strikes)
  - Optimal fuel (DYNAMO)
  - Optimal cost (DYNAMO)

Preliminary flow diagram



Work plan


Task
mar-19
abr-19
may-19
jun-19
jul-19
ago-19
sep-19
oct-19
nov-19
dic-19
ene-20
feb-20
mar-20
abr-20
may-20
jun-20
jul-20
ago-20
sep-20
oct-20
nov-20
dic-20
ene-21
feb-21
mar-21
abr-21
may-21
jun-21
jul-21
ago-21
sep-21
oct-21
nov-21
dic-21
ene-22
feb-22
mar-22
Review of the state of the art
Characterization and preliminary analysis of available data sources
Demand prediction module
Data-driven models: route clustering
Data-driven models: route choice
Physical models: parameters estimation
Physical models: 4D trajectory prediction
Data-driven models: Direct estimation of demand
Integration of models
Model refinement
Use cases
ConOps
Validation
Publication & disemination activities
PhD Thesis
Deliverable
Milestones

Initial results

Route choice modelling: initial results

1. Selection of a 6 relevant OD pairs for the experiments
2. Route clustering
3. Selection of the relevant variables
4. Feature engineering
5. Prediction experiments

1. OD pair selection

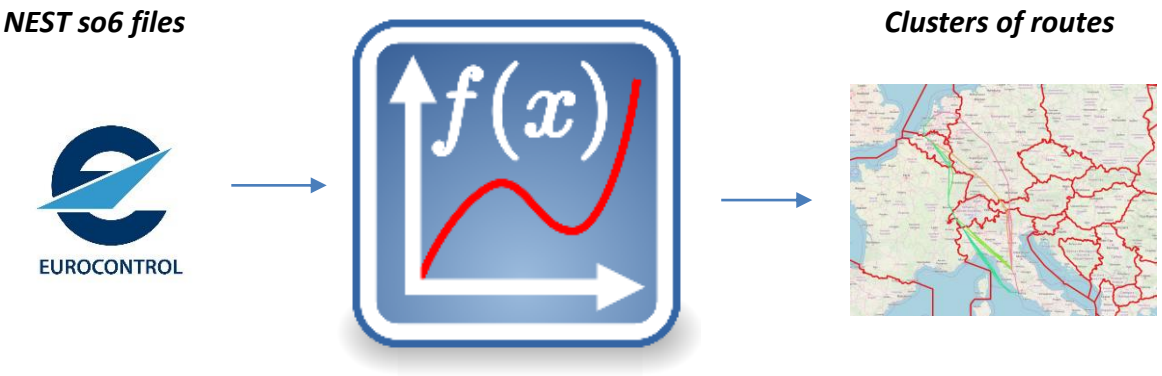
The relevant routes are selected taking into account the following requirements:

- High number of daily routes
- Geometrical variability on the routes
- Routes crossing several charging zones
- More than one airline flying the route
- Hourly distribution of the flights throughout the day

2. Route clustering

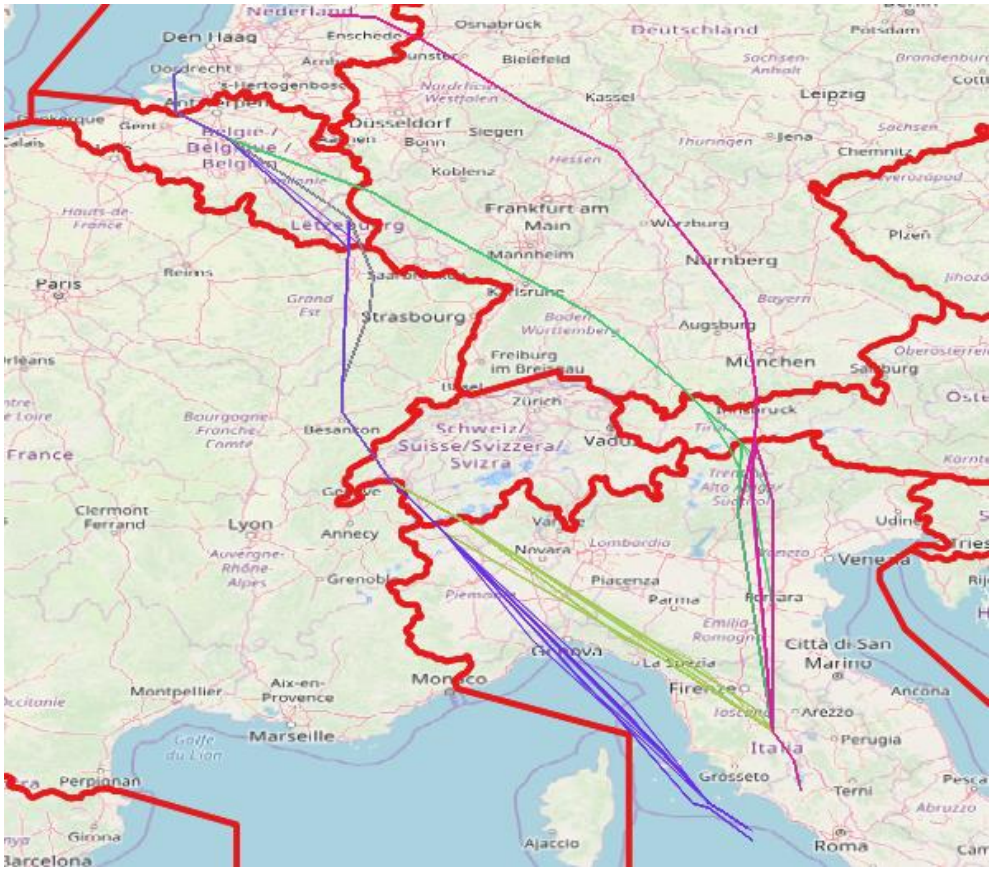
The route clustering problem is modelled by clustering algorithms (unsupervised machine learning)

The goal is to train a clustering scheme that better divides routes into categories, trying to convey similar geographical routes that travel similar distances and through the same charging zones



2. Route clustering

The clustering technique applied is the DBSCAN using as a metric the SSPD (pure geometric on 2D trajectories). Terminal area has been fitted to 40 NM:



Symmetrized Segment-Path Distance (SSPD)

SSPD is a geometry-based distance that does not take into account the time index of the trajectory. It compares trajectories as a whole, and it is less affected than other trajectory distances by incidental variation between trajectories.

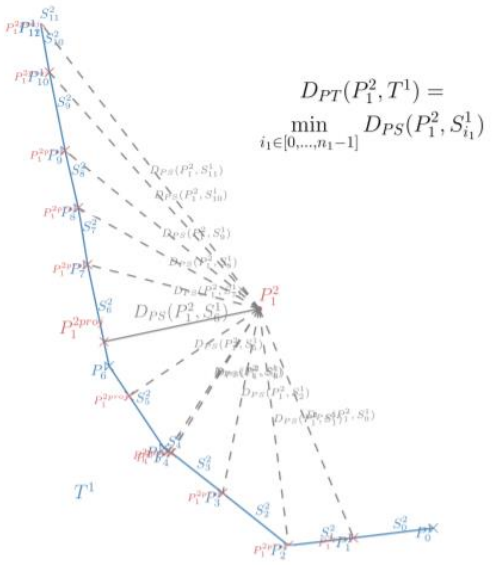


Fig. 4: Distance from point p_1^2 to trajectory T^1

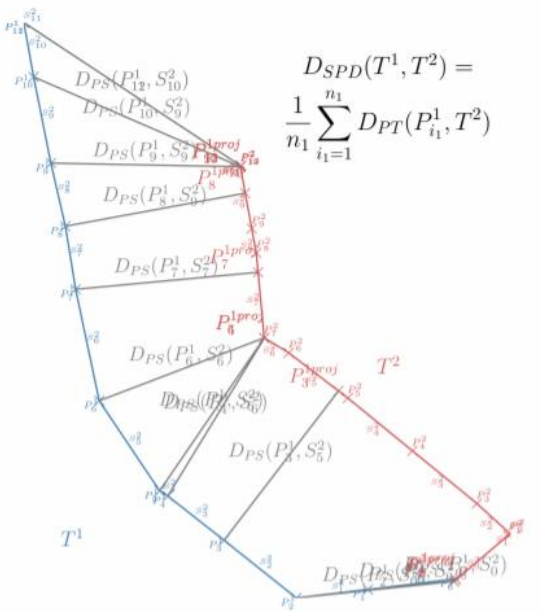
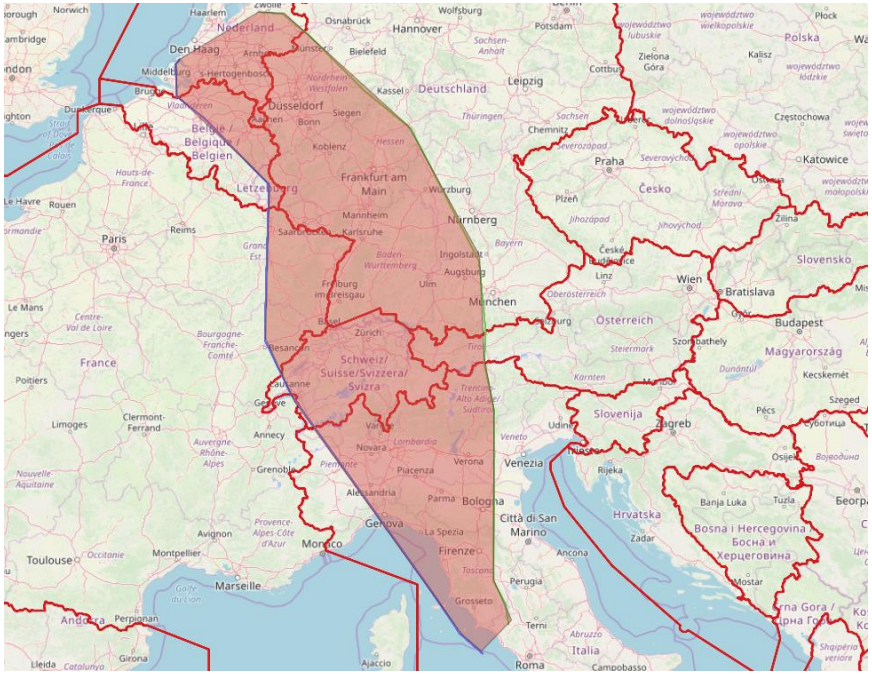


Fig. 5: SPD Distance from trajectory T^1 to trajectory T^2

Review & Perspective for Distance Based Clustering of Vehicle Trajectories, Philippe Besse et al.

Change in between-route distance metric

- SSPD distance works pretty well. But:
  - Dependent on the number of points of the route
  - Long processing times
- Propose the area inside the polygon formed by any two routes.
  - Geometric features only
  - Similar approach to SSPD
  - Much faster and efficient
- Using this metric, results are similar and the system is scalable.



Example for pair LIRF-EHAM

3. Selection of Relevant Variables

The following variables are going to be considered in this first proof of concept:

- **Route:** The route geometrical information is extracted from the pre-ops data. This route, obtained from the DDR pre-ops extraction (So6 files), is going to be the element to be predicted. Route charges and length are considered characteristics of each cluster.
- **Meteo:** Several meteorological indicators extracted form NOAA.
- **Airspace regulations**: Used as an indicative of the congestion affecting the route. Information from DDR.
- **General data obtained from the FPL**: Airline, aircraft, day of week and hour of day.

4. Feature engineering

The data has to be transformed into features which can be used as the problem predictors. The methodology is described below:

1. Understanding of the data source
2. Data interpolation: for time/space distributed data interpolation may be needed
3. Selection of the feature type: raw value, dummy values, functions of the data (e.g., sinusoidal)
4. Selection of the most statistically significant value(s) for each data source (e.g., maximum or average)

5. Prediction experiments

The first prediction experiments are performed taking into account the following considerations:

- The 2D route prediction is treated as a classification machine learning problem. In order to keep track of the decision value of the variables, multinomial regression is used
- Each OD pair is modelled as an independent problem
- No segmentation beyond OD pair is performed
- The benchmark to compare is the PREDICT software, codified in Python according to public EUROCONTROL's description

Variables considered

The variables considered in the model are currently the following ones (example for LIRF-EHAM pair, training 3 AIRACS, 6 different clusters detected):

- Cluster label: identifier of which cluster each flight belongs to
- General variables: Aircraft (MTOW), Airline (dummy variable for each one of the 5 airlines), DOW, cos(Start_time), sin(Start_time), cos(Day of Year), sin(Day of Year), dest_dir_wind, dest_spd_wind, origin_dir_wind, origin_spd_wind
- Variables assigned to each cluster at the time of the flight (one of each for cluster), e.g.: wind_factor, CAPE, CIN, Humidity, regulations_1, regulations_7, reg_delay_1, reg_delay_7, regulations_1_sum, regulations_7_sum, reg_delay_sum_7, reg_delay_sum_1
- Total example (15+12*6)= 87 different variables considered for this pair

Prediction

To perform a prediction, the tool extracts the variables (the already mentioned 87 variables) from the information contained in Flight Intentions

These variables are introduced in the model and it calculates the output, which looks like this for each one of the flight intentions:

Cluster	0	1	2	3	4	5
Probability of belonging to this cluster	0,02	0,66	0,1	0,21	0,001	0,009

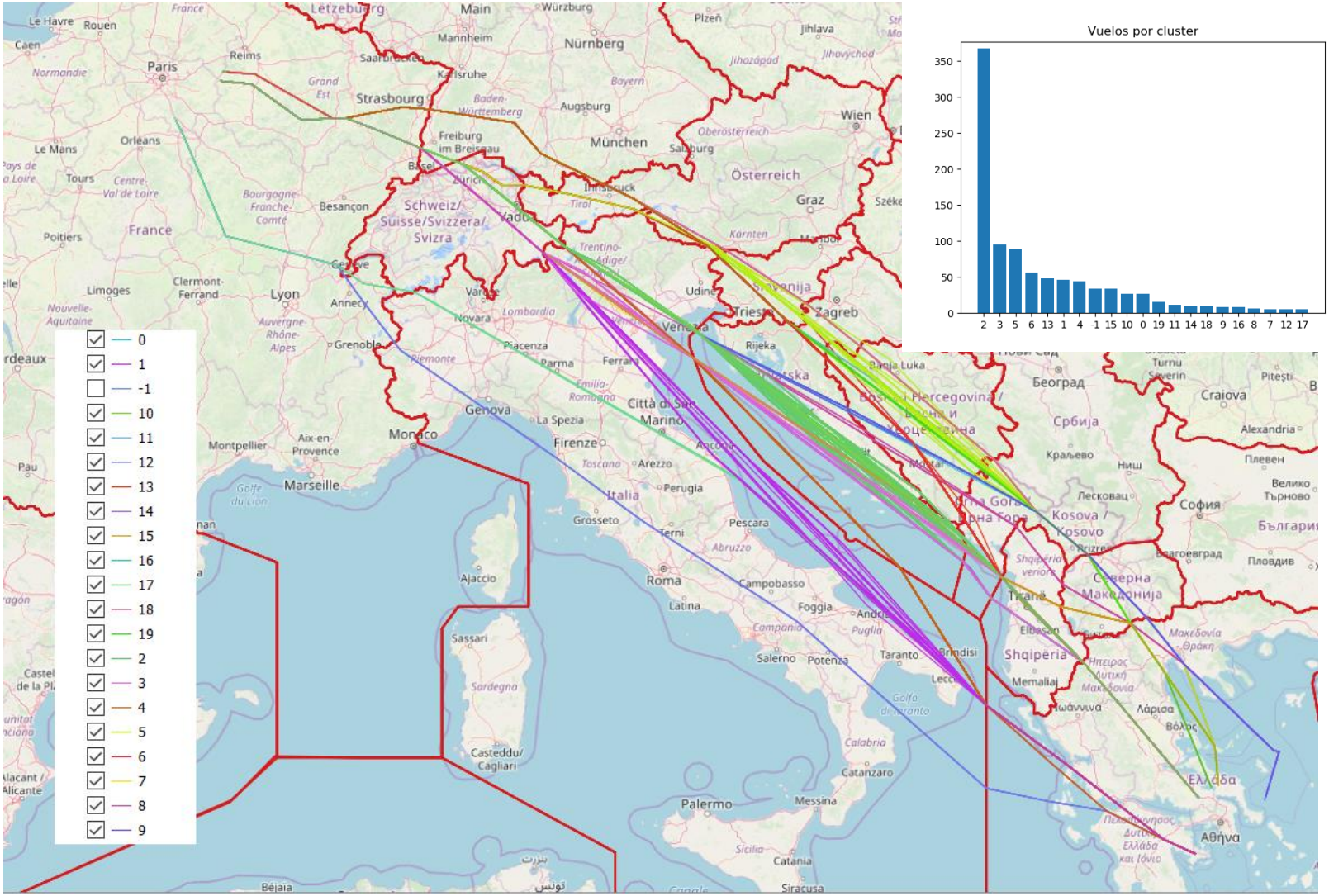
In this case the prediction will be that the predicted cluster for the flight considered will be cluster “1” with a probability of 66%.

Accuracy definition

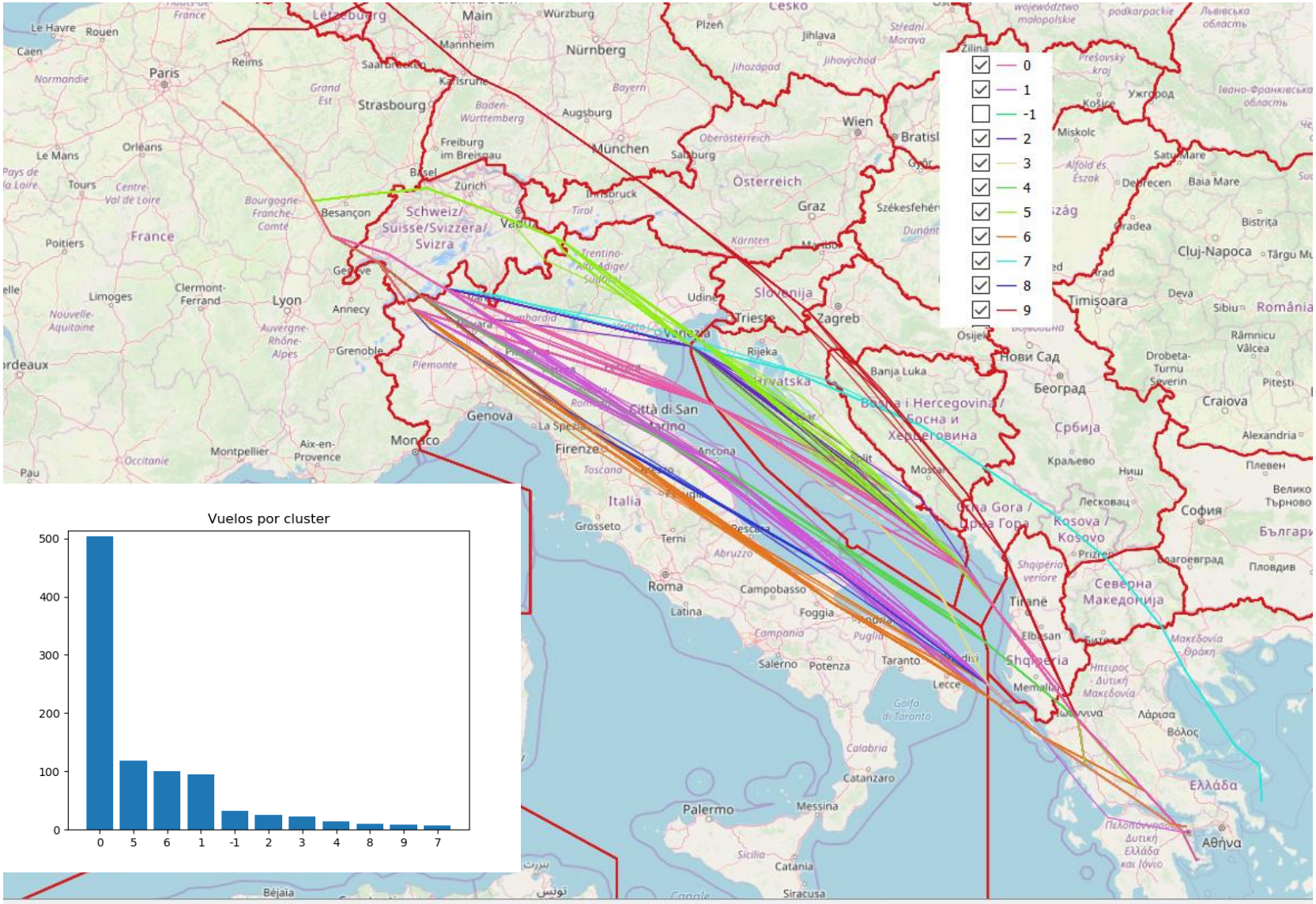
The only output of the prediction is the cluster label. The accuracy measures the relative number of predicted cluster labels that match with the ones from the pre-processing

- PREDICT: cluster labels are calculated using the PREDICT algorithm
- Model: cluster labels are calculated using the proposed multinomial regression model

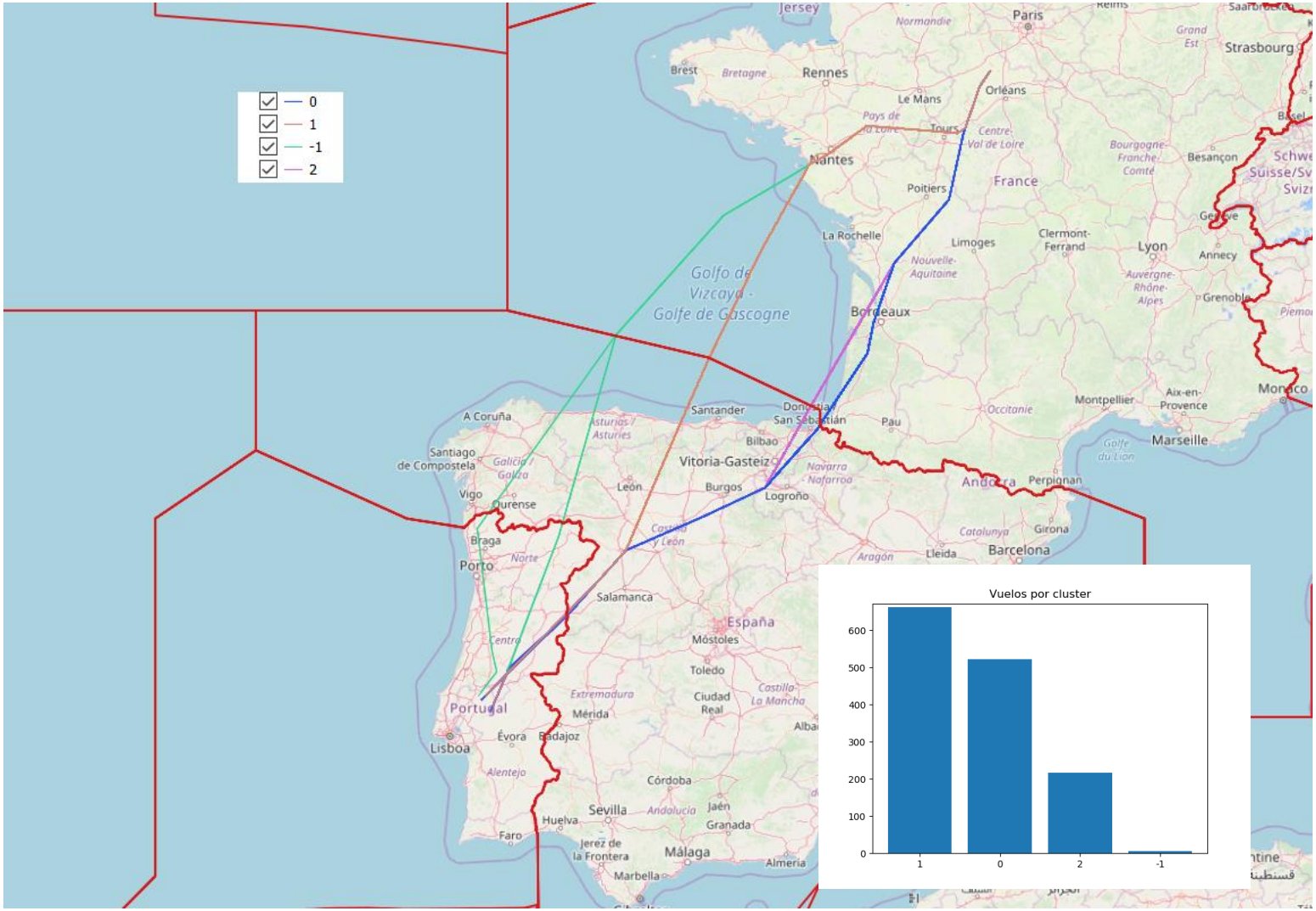
Classification example: LFPG-LGAV



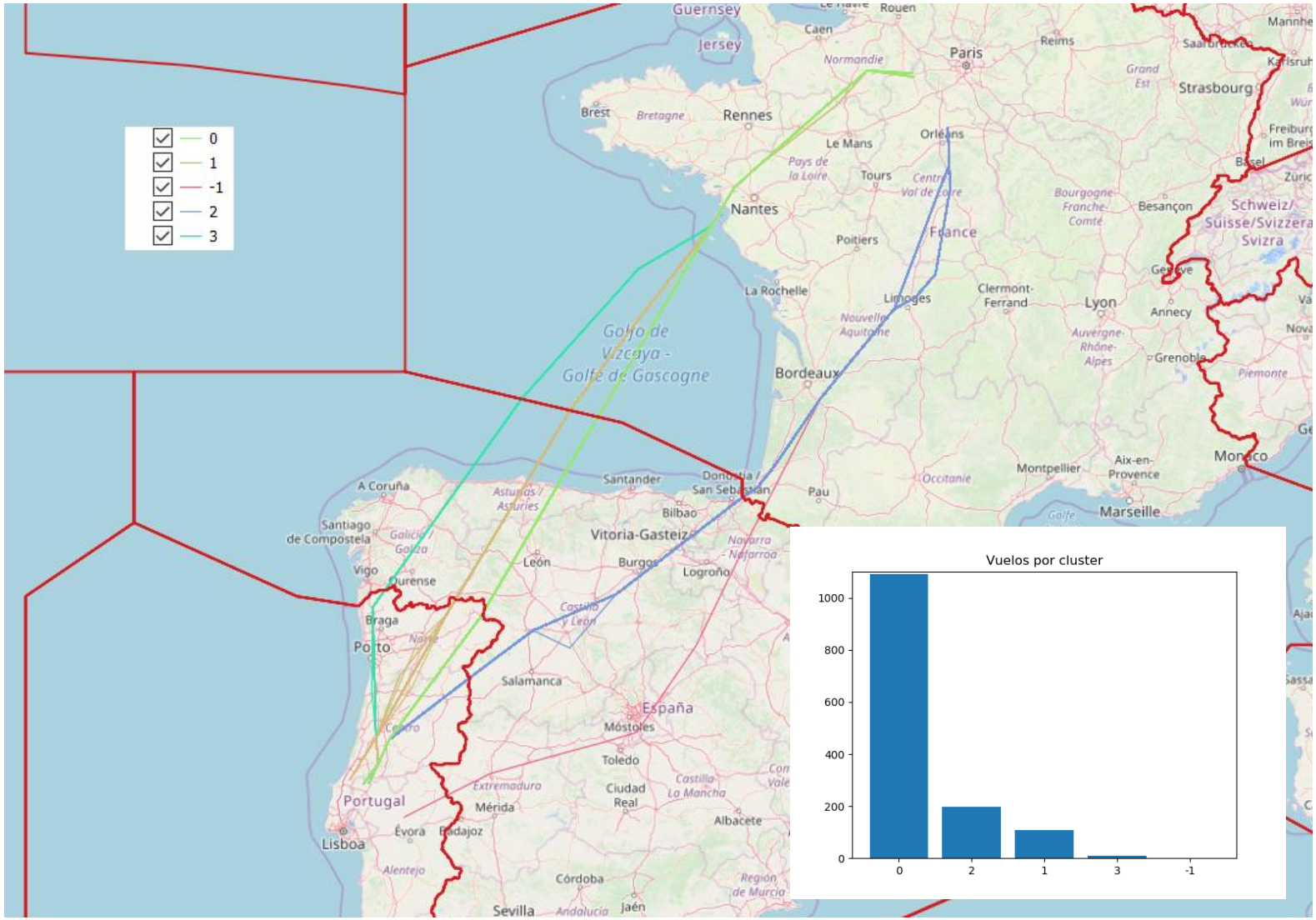
Classification example: LGAV-LFPG



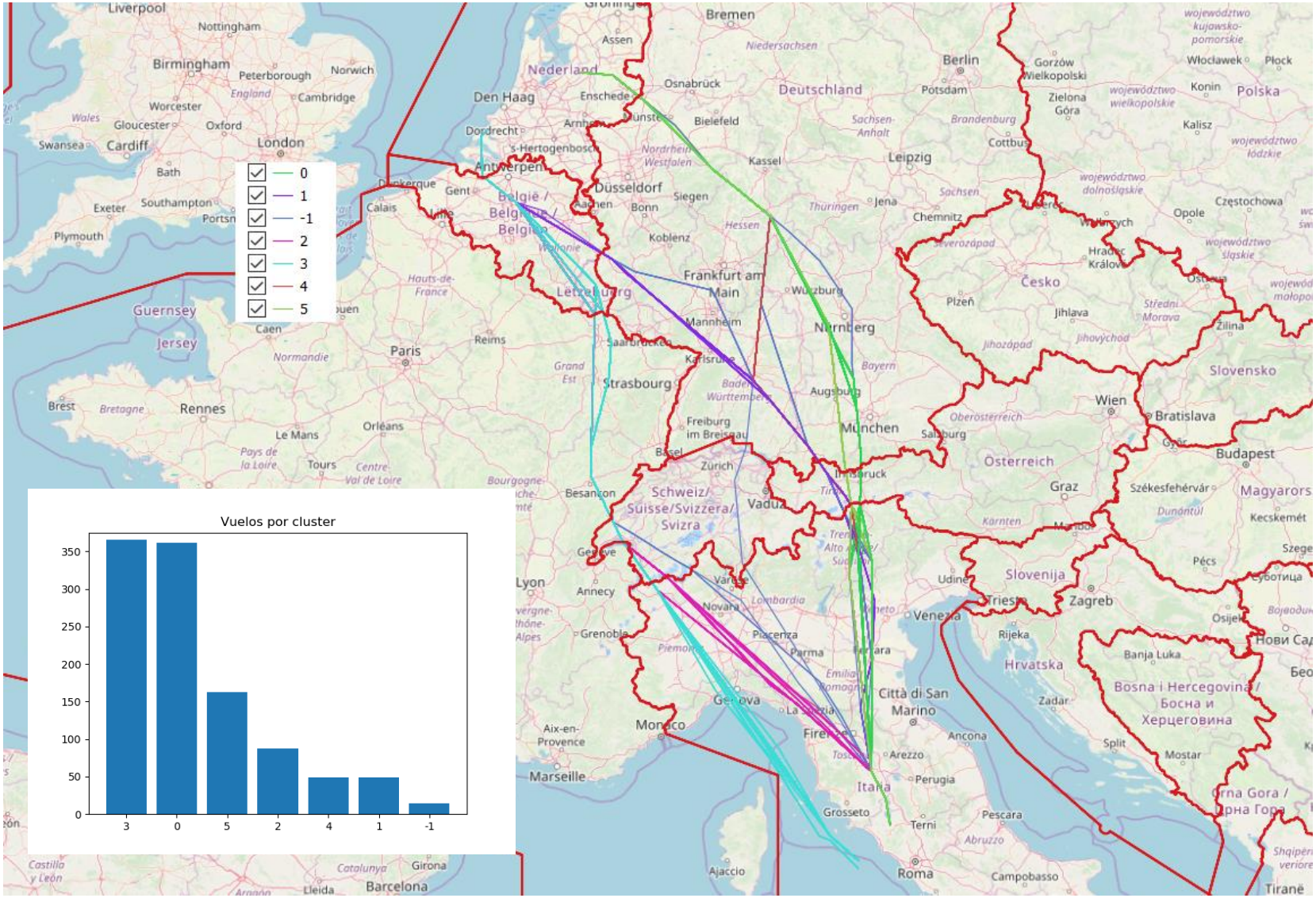
Classification example: LPPT-LFPO



Classification example: LFPO-LPPT



Classification example: LIRF-EHAM



General results

OD PAIR-Accuracy	EDDT-LEPA	LEPA-EDDT	LIRF-EHAM	EHAM-LIRF	LPPT-LFPO	LFPO-LPPT	LFPG-LGAV	LGAV-LFPG
NEW MODEL	0.841	0.907	0.610	0.916	0.823	0.624	0.520	0.654
PREDICT	0.780	0.859	0.544	0.875	0.554	0.557	0.253	0.471

Next steps

According to the observed progress, the next steps are foreseen:

- The inclusion of DYNAMO in the tool will be assessed
  - Hidden variables exploration
  - Route enrichment
- An integration attempt regarding direct cost related variables will be made (distance, charges, fuel cost, etc.) also considering DYNAMO capabilities
- Explore combinations of OD pairs (alternative modelling approaches)
  - Aggregating OD pairs would yield more pairs and enable more complex pattern-learning routines
- Continue modelling phase
  - Analyse new pairs
  - Try combinations of variables (and even automatic feature selection algorithms like RFE)

Any Questions?

Manuel Mateos (PhD Student)

Xavier Prats (Supervisor), Oliva Garcia(Co-supervisor)

3 December 2019