# *Mobility in Smart Cities Module*

## Practical exercises to observe and analyse mobility data

The aim of the practical work is to familiarise you with the complete process of analysing a corpus of data reflecting mobility flows, from the raw data to their interpretation.

During this session, your work includes two distinct parts that can be done separately.

As a prerequisite, you will need to download and install QGIS 3.10 using this link: https://www.qgis.org/fr/site/forusers/download.html

# *Part I*

The aim of this exercise is to identify the major traffic hubs regarding origins and destinations of taxi trips in the city of Porto, Portugal, in a time slot of a given day. This exercise is divided into two sections: data pre-processing and data visualisation.

You can download the raw data here: http://www.geolink.pt/ecmlpkdd2015-challenge/data/Porto_taxi_data_training.csv .

The dataset description can be found here: http://www.geolink.pt/ecmlpkdd2015-challenge/dataset.html .

Note that the dataset is a 11MB sample of the total dataset and only covers taxi trajectories from July 1st and 2nd, 2013. All the following operations, except those regarding data visualisation, will be performed using Python.

### 1. Data pre-processing

To prepare your data before any processing, it is essential to work the dataset to obtain consistent data that will not distort your analyses.

#### a. Import the data

- Import *pandas* as *pd* and *csv* libraries and load data sample of 7,733 rows as a pandas dataframe.
- Import *datetime* library and convert Unix timestamp to human friendly, python readable, format. You can use *pd.to_datetime.*

### b. Clean the data

- To clean the data, we will remove instances with erroneous POLYLINE attribute. For that purpose:
  - ✓ Delete rows where the attribute MISSING_DATA equals TRUE.
  - ✓ Delete rows with less than 4 coordinates in the POLYLINE attribute (it means that the destination is missing) using String formatting and eventually StringIO library.

To make sure the erroneous instances have been deleted, print the number of lines before and after cleaning the data. 125 instances should have been removed.

### c. Select the data

- Select data from July 1st, 2013.
- Import matplotlib.pyplot and plot the number of trips on July 1st hour by hour
- Notice that there is a peak in traffic from 9:00 am to 9:59 am.
- Select data representing trips starting on July 1st from 9:00 am to 9:59 am. In the following, we will work with this sub dataset only.

### d. Format the data for QGIS

- Extract the origins and destinations coordinates using String formatting and eventually StringIO library, and add columns origin_longitude, origin_latitude, destination_longitude, destination_latitude to dataframe.
- Export dataframe with origin and destination coordinates to csv file: "origins_destinations.csv".

## 2. Data visualisation

Visualizing the results is an important step in this line of work. You are asked to visualize the data through QGIS tool by following the steps below.

### a. Create the project and add a raster map

- Open QGIS and create a new project. Save it as "mobility.qgz" in your preferred folder.
- Add an Open Street Map layer by selecting in the menu: Web>QuickMapServices>OSM>OSM Standard.
- You can manage your layers in the layers panel on the left.

### b. Visualize the origins or departures locations

- Select Layer>Add Layer>Add Delimited Text Layer…
- Select the "origins_destinations.csv" file, write "origins" as the Layer name and select the corresponding file format.
- In Geometry definition, check Point coordinates and choose ORIGIN_LONGITUDE as X field and ORIGIN_LATITUDE as Y field.
- Make sure the Geometry CRS is EPSG:4326 – WGS 84.
- Click Add, then close the window.
- Zoom in on the map on Porto, Portugal using the magnifying glass tool in the top toolbar

### c. Display origins or departures heatmap

- In the layer panel, right click on the "origins" layer and select Properties.
- Choose Symbology in the left panel and select Heatmap in the top drop-down menu
- For Colour ramp, choose Spectral, and then, again in Colour ramp, click Invert Colour Ramp.
- Expand the Layer Rendering tab and select a 50% opacity.
- Click OK.
- Notice that there are two important departure hubs. Using the magnifier in the bottom toolbar, identify the hubs as the Porto-Campanhã and Porto-São Bento railway stations.

### d. Visualize the destinations or arrivals locations and display their heatmap

- Same steps as before but be careful to choose this time DESTINATION_LONGITUDE as X field and DESTINATION_LATITUDE as Y field, in Geometry definition, when adding the delimited text layer.
- Notice that there are two important destination hubs. Using the magnifier in the bottom toolbar, identify the hubs as Porto city centre and Porto airport.

---

## *Part II*

---

The aim of this exercise is to perform a clustering, using K-means, of origins and destinations of taxi trips in the city of Porto, Portugal, in a time slot of a given day. This exercise is divided into three sections: data pre-processing, data processing and data visualisation.

### 1. Data pre-processing

You can proceed with the same pre-processed data of part I.

### 2. Data processing

Once your data is ready, you can do any analysis you want. In this scenario, we will be interested in the characterization of taxi trips.

### a. Extract origins and destinations

- Extract the origins and destinations coordinates using String formatting and eventually StringIO library, and add columns origin_longitude, origin_latitude, destination_longitude, destination_latitude to dataframe

### b. Convert origin and destination longitudes and latitudes into cartesian coordinates

- Install pyproj and from pyproj, import Proj, transform
- Transform origin and destination coordinates from EPSG:4326 to EPSG:5018 by adding new columns ORIGIN_X, ORIGIN_Y, DESTINATION_X, DESTINATION_Y to the dataframe. Running this operation takes some time.

### c. Clustering with K-means

- To cluster the origins, create a new dataframe, origins, with ORIGIN_X and ORIGIN_Y as its columns
- From sklearn.cluster, import KMeans. Set K-means parameters as follows: "kmeans = KMeans(n_clusters=14, init='random', n_init=100, max_iter=300)" with n_clusters being the number of clusters, init the method of initialization for initial centroids, n_init the number of time the k-means algorithm will be run with different centroid seeds and max_iter the maximum number of iterations for a single run.
- Use kmeans.fit_predict to add a column with the cluster ID for each origin:
  " origin_clusters = origins
  origin_clusters["cluster_ID"] = kmeans.fit_predict(origins) "
- Save as csv file "origin_clusters.csv"
- Repeat the same steps for the destinations.

## 3. Data visualisation

Here again, you are asked to visualize the data through QGIS tool by following the steps below.

### a. Create the project and add a raster map

- Open QGIS and create a new project. Save it as "mobility.qgz" in your preferred folder.
- Add an Open Street Map layer by selecting in the menu:
  Web>QuickMapServices>OSM>OSM Standard.
  You can manage your layers in the layers panel on the left.

### b. Visualize the origins or destinations locations

- Select Layer>Add Layer>Add Delimited Text Layer…
- Select the "origin_clusters.csv" file and choose the corresponding file format.
- In Geometry definition, check Point coordinates and choose origin_X as X field and ORIGIN_Y as Y field.
- Make sure to choose EPSG:5018 as the Geometry CRS by clicking the Select CRS button and using the search field.
- Click Add, then close the window.
- Zoom in on the map on Porto, Portugal using the magnifying glass tool in the top toolbar
- Repeat the same steps for destinations

### c. Visualize the clusters

- In the layer panel, right click on the "origin_clusters" layer and select Properties.
- Choose Symbology in the left panel and select Categorized in the top drop-down menu
- For Value, select cluster_ID in the drop-down list
- Click on Classify button on the bottom left to display the clusters' symbols and then click OK.
- Repeat the same steps for destinations.