A Network Tour of Data Science

Authors:
Baptiste Hériard-Dubreuil
Jean-Baptiste Membrado
Guilhem Noiraud
Amaury Véron

Supervisor:
Rodrigo Pena

Professor:
Pierre Vandergheynst
Pascal Frossard

Retrieving the continent labels from the air routes structure

---

1 Context

This project is based on the dataset "Flight Routes" supervised by Rodrigo PENA. Following our work on the four milestones, we will pursue our analysis of the structure of the air routes structure linking the airports.

Our idea for this project is to retrieve the value of the *Continent* parameter from the structure of the graph. The assumption forming the basis of our work is that the aerial network is heavily linked to the proximity of the airports between each other. We think that if airports are close, it is more likely that an air route exists between those two than between airports that are further away from each other.

From the results we had in the milestones, and the tools seen in class we will predict the value of the *Continent* on the airports of our dataset.

1.1 Air routes

The basic element of our graph is the air route : it is an undirected edge between two airports corresponding to the existence of a flight linking those two airports. Those edges are first unweighted : the edges only state that the airports are linked, but do not transmit any information on the duration of the flight for example. Later we chose to add a weight on the edges by computing the Geodesic distance between the airports locations. Those two ways of weighting the edges will lead to two distinct matrices : the adjacency matrix for the binary way of weighting the edges and the weighted matrix for the Geodesic distance weighting.

1.2 Problematic

There are eight different labels if we neglect the single airport of Arctic and the unlabelled nodes, which represent about 0.7% of the graph. If we just use the distances between airports, how well can we reconstruct the continents?

2 Creating the basic structures

There are two approaches to compute the matrix that describes the connections between airports. The first idea that come to mind is to consider only the existence of a direct route between two airports. The weights will be 0 or 1. The second approach is to create weights in order to have more representative connections. We decided to use both approaches. We will compare both of them.

2.1 Adjacency matrix and weighted matrix

To create both the adjacency and the weighted matrices, we first need to identify the interesting data. Since some airports was not consistent (because of a lack of latitude, longitude) and some connections between airports miss the departure or the arrival, we had to filter the data to be used. The result has given several clusters, we kept only the biggest one. Finally, over the initial 3333 airports, we get a coherent cluster of 3154 airports.

The adjacency matrix is formed only with 0 and 1 according to the existence (or not) of a direct path between two airports. The weighted matrix is created according to the following formula:

$$W_{i,j} = exp\Big(-\frac{d_{i,j}}{\sigma}\Big)$$

Where $d_{i,j}$ is the distance between the airport i and the airport j and σ is an arbitrary parameter.

2.2 Laplacian: eigenvectors and eigenvalues

Before using the K-means algorithm on our graph, we need to embed our graph on a Laplacian Eigenmap. We will detail here the process of this embedding.

First, we compute the graph Laplacian. Simply, the Laplacian is $L = D - W$ where D is the diagonal matrix with the degrees of the nodes being the diagonal elements of the matrix, and W is the weighted matrix. We compute the eigenvectors and eigenvalues of the generalised problem:

$$Lu = \lambda Du$$

These eigenvectors will serve for the K-means algorithm. We chose these ones rather than the eigenvectors of the Laplacian as we obeserved that they returned better results.

The aim of the Laplacian embedding is to simplify our graph to its most important features. The eigenvectors with the lowest associated eigenvalues form a good simplification of our graph. This is why we sort the eigenvalues by increasing value, and then the eigenvectors with the same order. For the K-means algorithm, we will keep only some of the eigenvectors, amongst the lowest eigenvalue ones. If we want to run the algorithm for k clusters, we will keep as many eigenvectors to represent our graph.

2.3 Clustering

Since we decided to neglect the labels corresponding to Arctic and unlabelled nodes, we have eight remaining clusters. The K-means method has been chosen to do the clustering. K-means can take initial centres as input, so we initialised K-means with specific points to improve the clustering. The result of the clustering is then displayed over the two first eigenvectors.

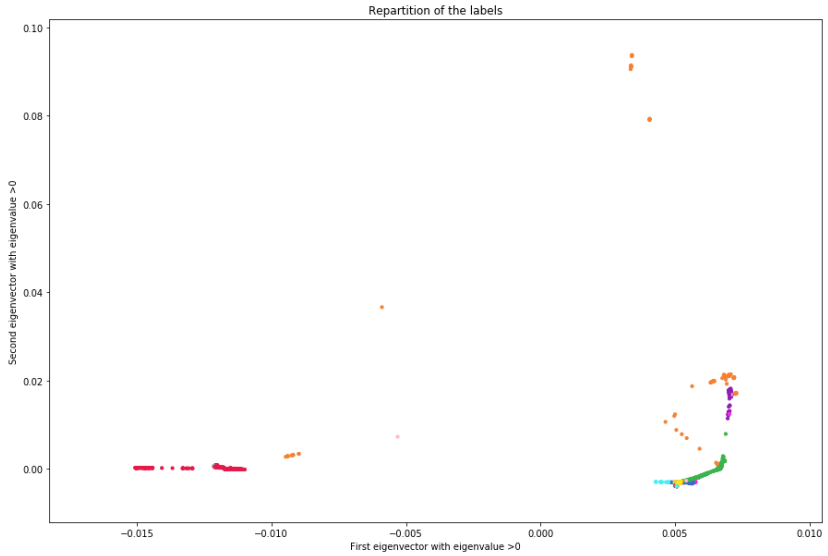


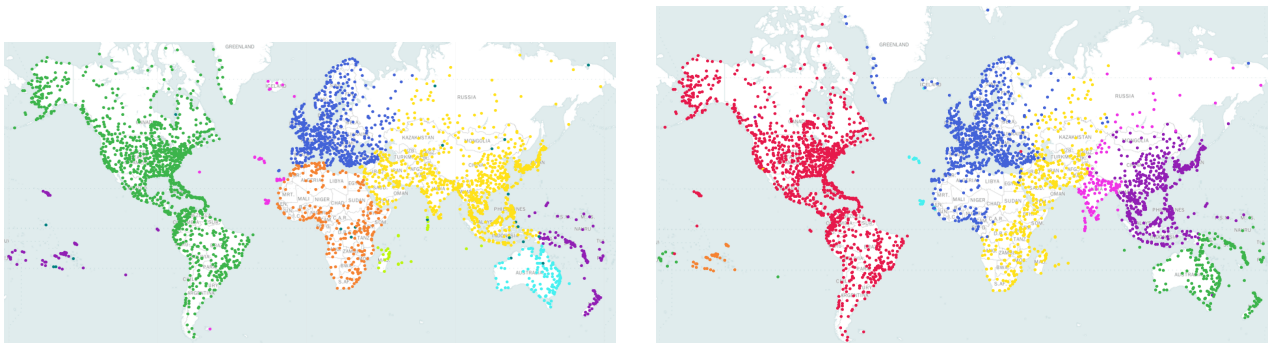
Figure 1: Clustering representation over the first two eigenvectors with non-zero associated eigenvalue

3 Results and discussion

This section only reports the results for the weighted matrix since the adjacency matrix give very bad results. Although, the results are shown in the appendix (figure 7).

3.1 The continents according to the flight routes

Figure 2 presents the resulting world map after labelling with two eigenvectors used. Since the labelling with two eigenvectors gives the best visual results, it has been kept for the following discussion. Other maps are available in appendix (figure 8).



(a) Ground truth

(b) Labelling with two eigenvectors

Figure 2: Comparison between the labelling and the ground truth with two eigenvectors used

3.2 Evaluation of the clustering

The figure 3 presents the percentage of well labelled airports for each continent. The global score is 80%.

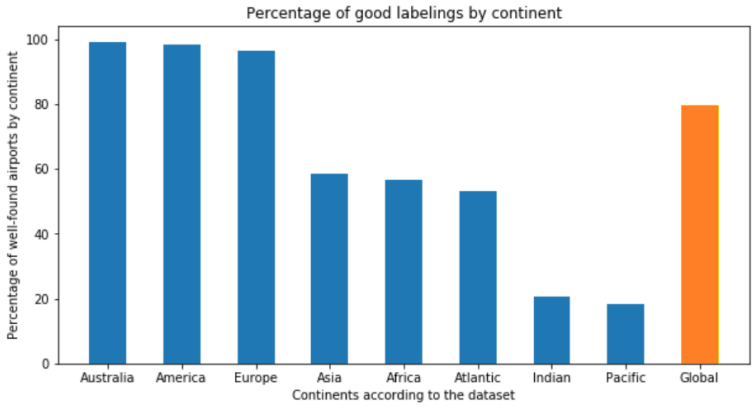


Figure 3: Percentage of well labelled airport per continent

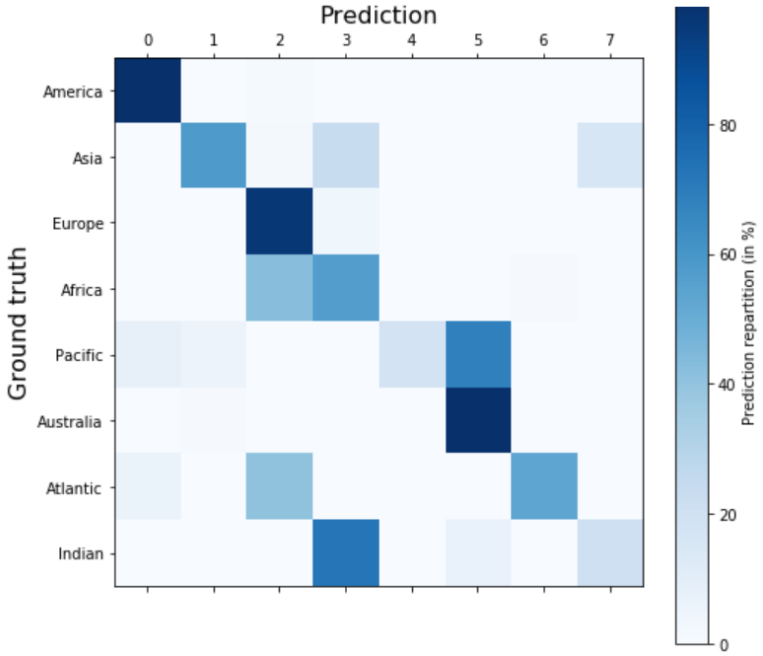


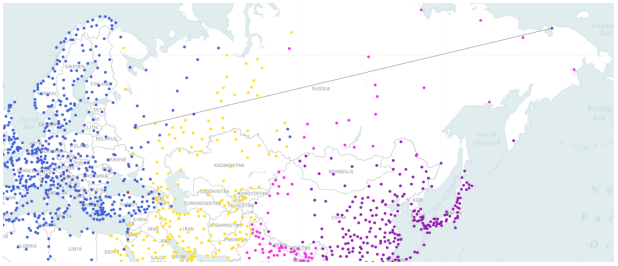
Figure 4: Confusion matrix of continent clustering. For a given continent, you can see how many of its airports have been attributed to which cluster (in percent).

3.3 Isolated labelled airports

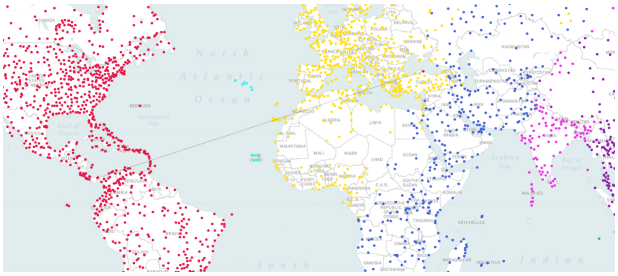
The labelling divides the world airports into eight pieces. Surprisingly, four isolated airports have been noticed. They are presented in figure 5. We found the explanation for all of them.

Pevek is an isolated airport in the North of Russia. There is only one airline to Moscow which belongs to Europe. This is why Pevek is labelled as a European airport.

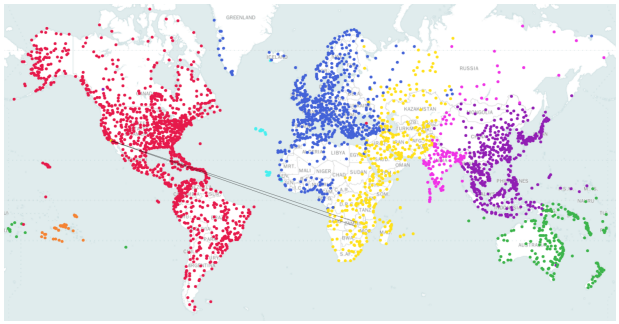
Sinop, in Turkey, is labelled as an American airport. Its single airline links to Panama City in Panama, but some researches on the Internet present nowadays connections to Istanbul. The history of the airport permits to explain what happened. Until March 2013 Sinop airport was an American military base, which is approximately the same year of the last update of our data. The new Sinop airport IATA code is 'NOP' and the American Sinop airport had 'SIC' as IATA code, as in our dataset. And, since 2012, the 'SIC' IATA code has been attributed to San Jose Island Airport, near Panama City. This confusion even makes some travel reservation



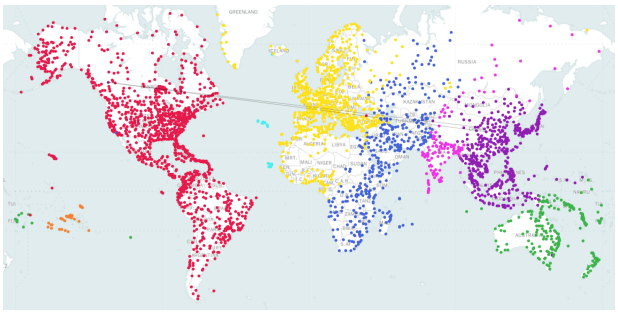
(a) Connections with Pevek airport, Russia



(b) Connections with Sinop airport, Turkey



(c) Connections with Los Alamitos Army Air Field, United States



(d) Connections with Mackenzie airport, United States

Figure 5: Airlines to some specific destination

website[1] locate the San Jose airport in Turkey.

The most interesting case is the one of Los Alamitos Army Air Field in the United States. In the dataset *airports*, this airport has its coordinates in the United States and is labelled in the United States, but is localised in Solwesi, Zambia. The dataset *routes* gives two airlines between this airports and other cities in Zambia, this explains why our code has labelled this airport as an African airport. After some researches, we found the explanation. The Los Alamitos Army Air Field has not IATA code, but a ZAA code which is 'SLI'. A IATA code or a ZAA code is unique, but it is possible to find an identifier that exists in both lists. In fact, Solwesi aiport has 'SLI' as IATA code. This confusion must have brought errors in the airport database : the IATA code that corresponds to our airport is wrong, since it does not exist. This also explains the airlines in Zambia.

The last airport of interest is the Mackenzie airport in the United States. The whole data is coherent, but the airport has connections with airports in China. Just like Los Alamitos Army Air Field, the airport has not IATA code. This code corresponds to Zhangye Ganzhou airport, China.

4 Conclusion

In conclusion, we managed to fit 80% of the labels to the ground truth. The strong connection between North Africa and Europe compared to the rest of the world explains the fact that we did not manage to split very well the two continents. We were also able to explain the isolated points of our graph. If we merge some labels, we find a map that fits quite well to the real continents (figure 9), with 87% of good answers. Even the border between Indonesia and Papua New Guinea is right. Among all of this, this project allowed us to improve our geographic skills by discovering many cities, countries and properties of the global air routes.

[1]See *Skyscanner*, https://www.skyscanner.com/airports/sic/san-jose-island-airport.html

Appendices

A Ground truth without unlabelled airports

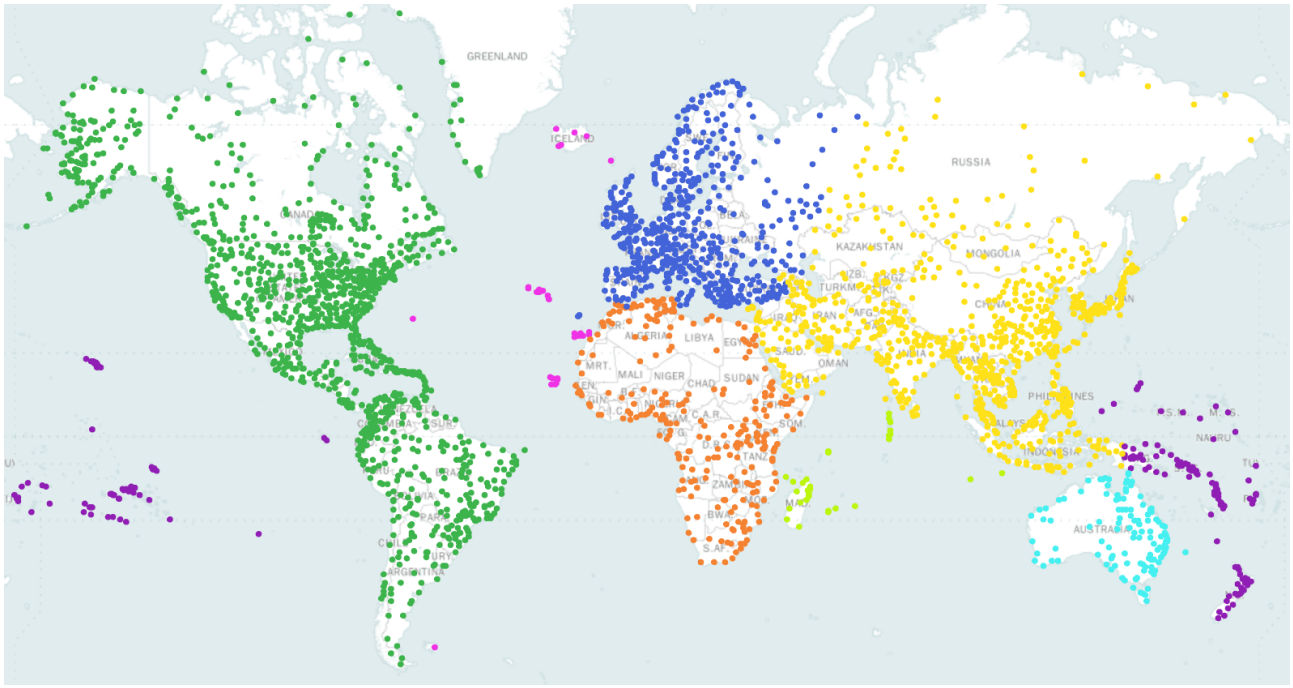
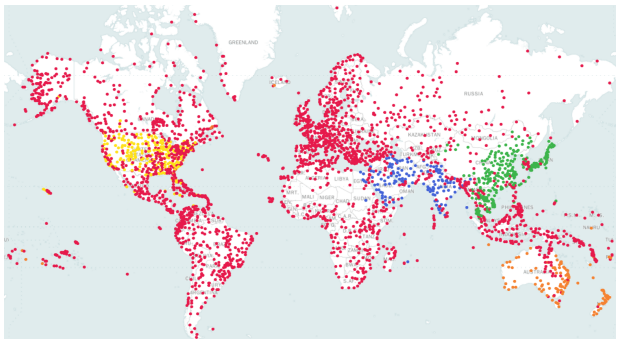
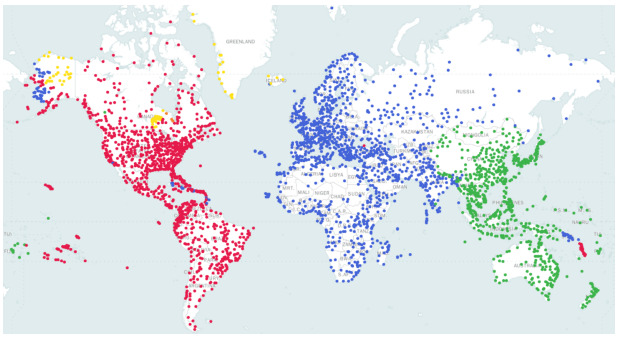


Figure 6: Ground truth Continent labels

B Clustering with adjacency matrix



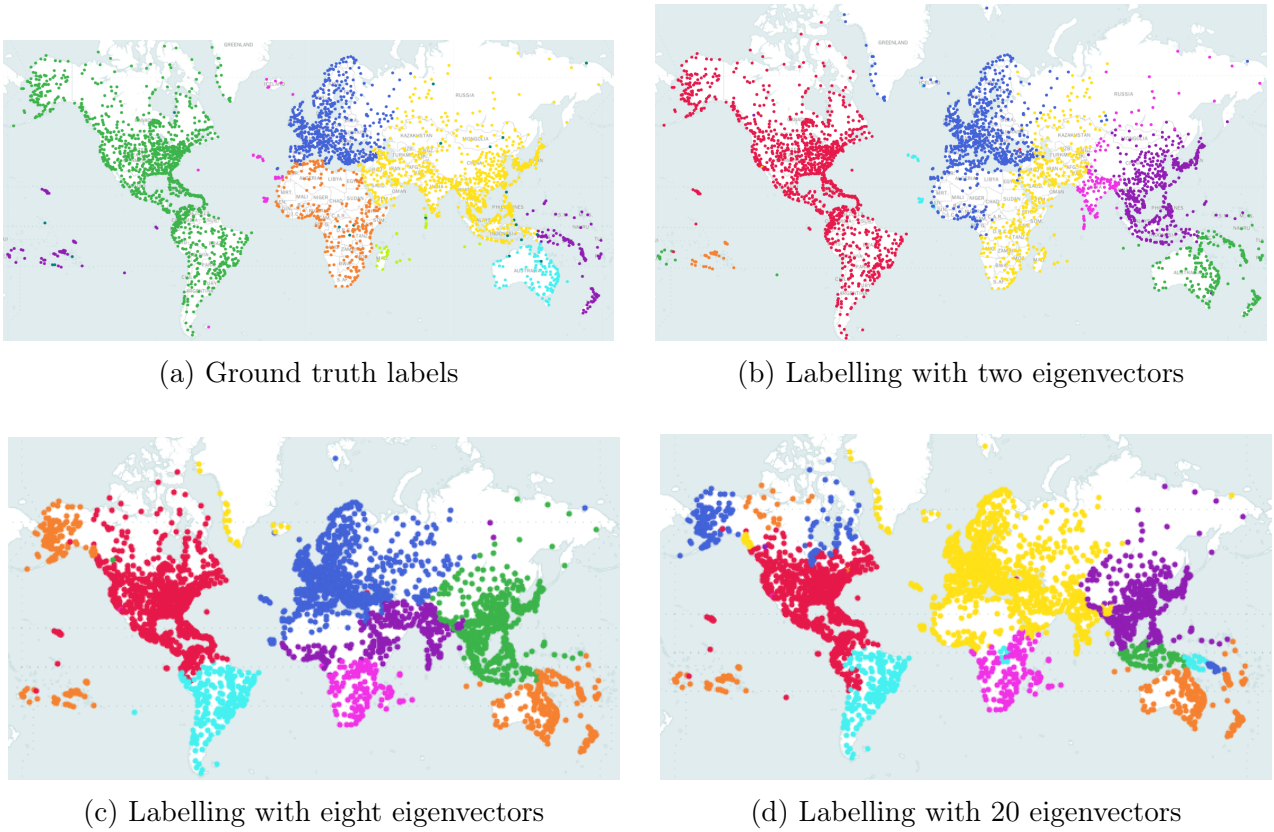
(a) K-means with five clusters and random initialisation



(b) Five clusters with initialisation.

Figure 7: Clustering map with only the adjacency matrix as graph data

C Different labelling for the heat kernel



(a) Ground truth labels

(b) Labelling with two eigenvectors

(c) Labelling with eight eigenvectors

(d) Labelling with 20 eigenvectors

Figure 8: Comparison between the labelling and the ground truth for different numbers of eigenvectors

D Labelled map and true continents

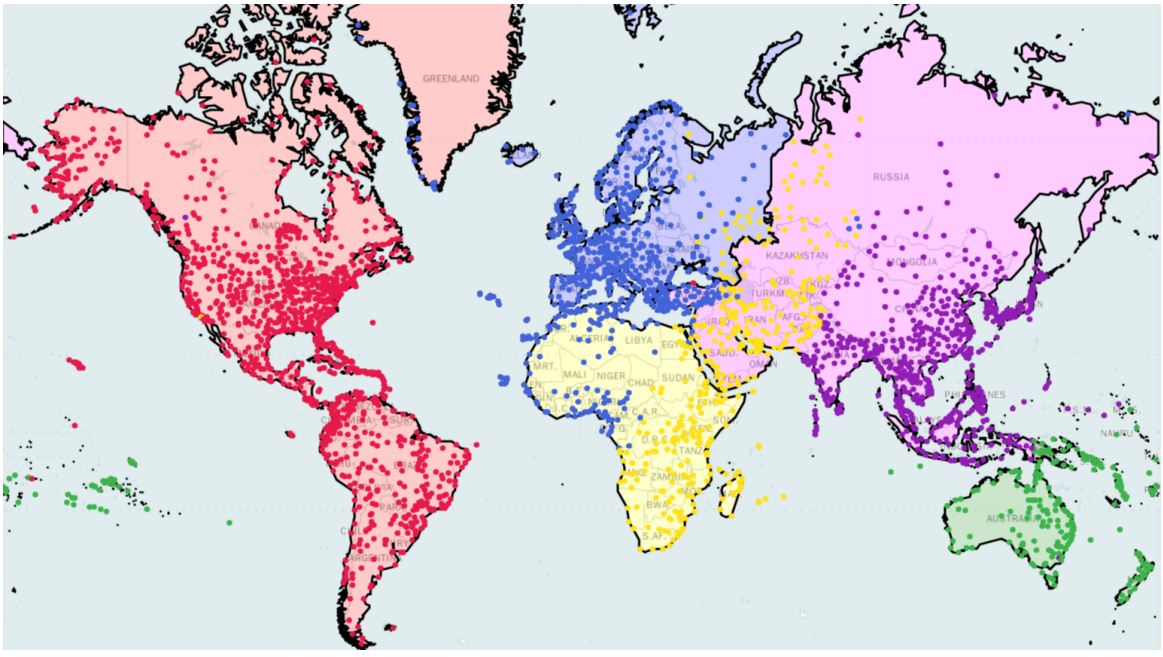


Figure 9: Labelled map over the geographic continent map where several labels have been merged. About 87% of good labels.