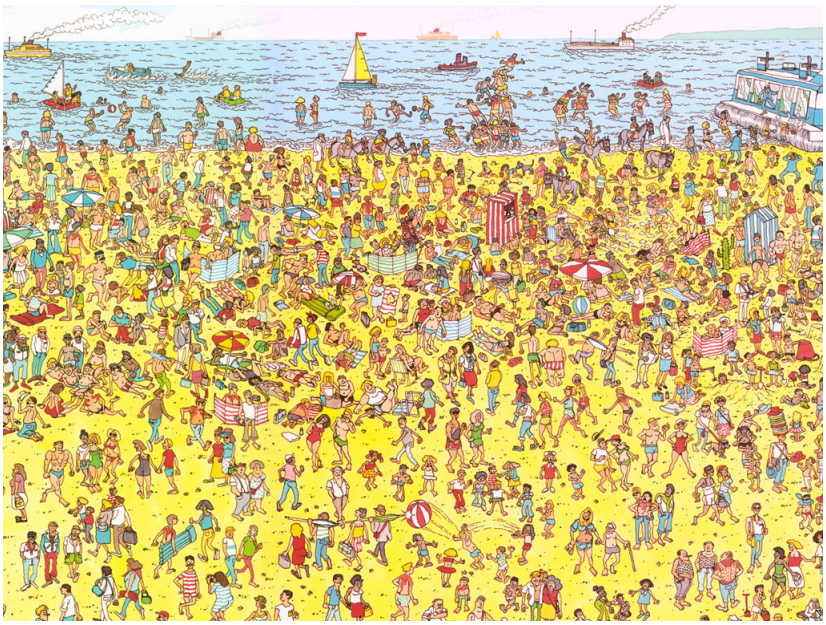
INSA Lyon
20, avenue Albert Einstein
69621 Villeurbanne Cedex

Livrable de projet

Fouille de données
« Flickr – Découverte de points d'intérêt »

du 24 février au 24 mars 2014



Les Martins :
Aline Martin
Martin Wetterwald

Enseignants :
Jean-François Boulicaut
Mehdi Kaytoue

Année scolaire 2013-2014

Sommaire

1. Préparation des données

2. Visualisation

2.1 Introduction

Pour la visualisation géographique des données, nous avons utilisé les technologies suivantes :

— Python ;
— Flask (serveur web) ;
— Scikit-learn (librairie de data mining) ;
— Google Maps API.

Avant de commencer toute tentative de représentation des données sous forme de cluster, nous avons commencé par afficher les points sur la carte. Étant donné qu'il y environ 80 000 points, nous ne pouvons en afficher qu'une partie, en raison de la lenteur observée sur les navigateurs lors du rendu de la carte.

Nous avons opté pour le rendu d'un pourcent des points. La figure 2.1 présente le rendu d'un échantillon des points.

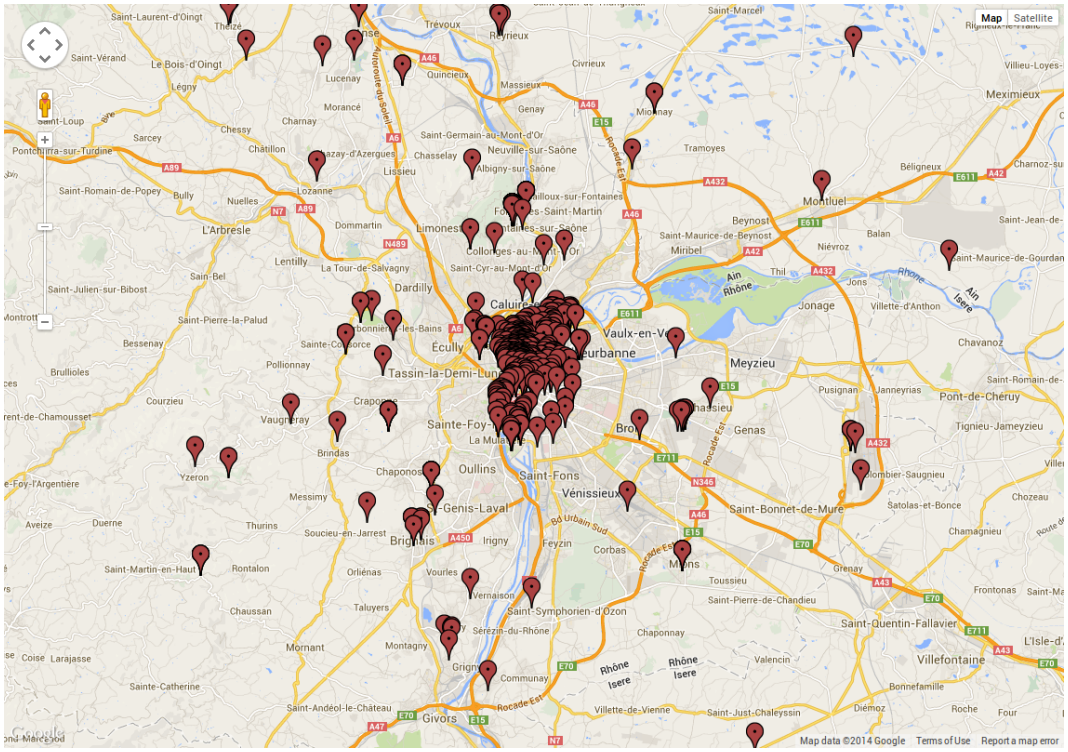


FIGURE 2.1 – Affichage d'un échantillon des points

Cela nous donne déjà un premier aperçu. Nos points sont répartis autour de la ville de Lyon, et on constate une forte densité de points dans la ville.

Mais zoomons entre les deux fleuves de Lyon (figure 2.2).

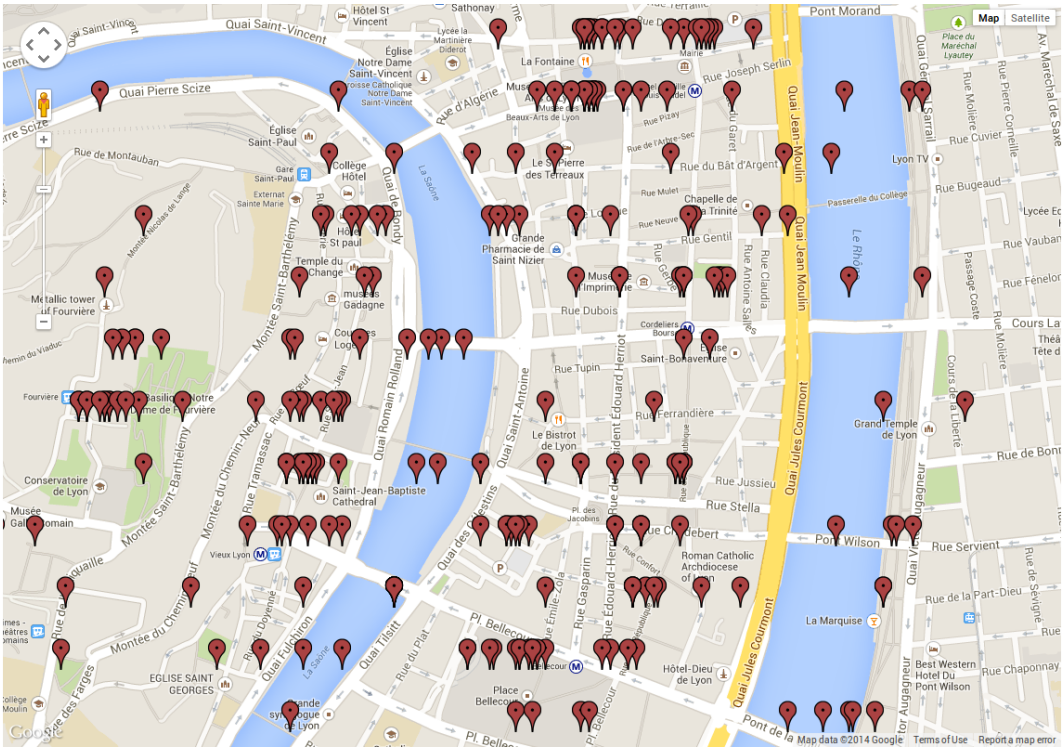


FIGURE 2.2 – Imprécision détectée lors du zoom

On constate que tous les points se trouvent sur des lignes dont la séparation est très nette. Cela est dû à un manque de précision dans nos données de départ. Dans le fichier CSV initial, nous remarquons en effet que les coordonnées GPS les plus précises que nous avons ne comportent que 4 chiffres après la virgule, ce qui explique la répartition étrange des points sur des lignes. Nous savons donc que cette erreur aura un impact sur la qualité de nos futurs clusters.

2.2 Meanshift

Nous avons choisi d'appliquer l'algorithme de clustering **Meanshift**, pour tenter de trouver des clusters géographiques intéressants.

De manière très haut niveau, l'algorithme de clustering Meanshift peut être résumé comme suit :

— fixer une fenêtre autour de chaque point ;
— calculer la moyenne (le barycentre) des points à l'intérieur de cette fenêtre ;
— déplacer (*shift*) la fenêtre sur la moyenne (*mean*) et répéter ces étapes jusqu'à atteindre une convergence.

Nous avons choisi de représenter les clusters par des cercles dont l'aire est proportionnelle au nombre de points contenus dans le cluster. Chaque cercle est coloré aléatoirement,

et son centre est le centre du cluster. Le rayon du cercle en mètres est de : $r = 10 \times \sqrt{n}$, où n est le nombre de points présents dans le cluster. L'usage de la racine carrée nous permet d'éviter d'avoir une taille de cercle trop importante dans le cas de gros clusters.

La figure 2.3 montre le résultat d'un premier lancement de l'algorithme Meanshift avec un quantile de 0.2 et un nombre minimal de points par cluster de 15. Nous n'avons pas affiché les points pour plus de clarté.

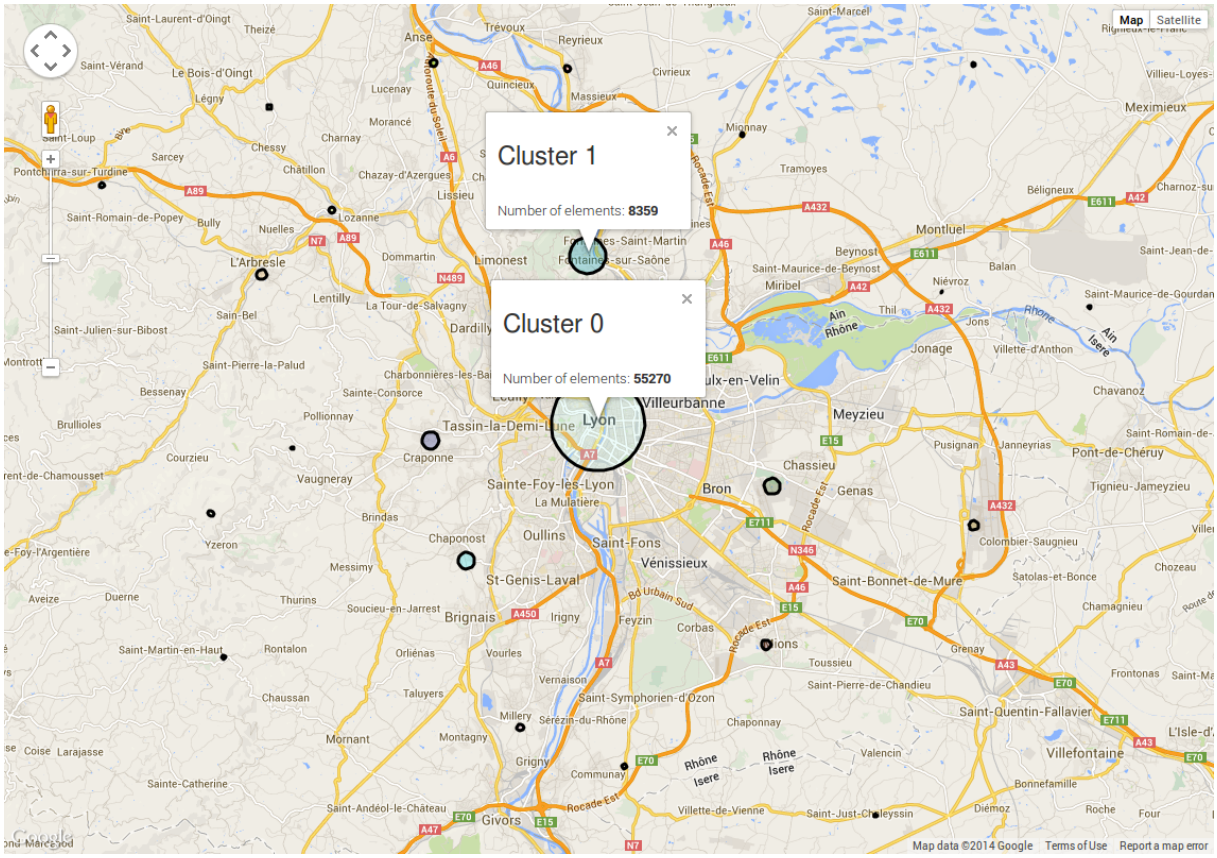


FIGURE 2.3 – Meanshift, $q = 0.2$ et $n = 15$

On remarque sans surprise qu'un cluster sort nettement du lot : il s'agit du centre ville de Lyon, endroit où les photos Flickr sont les plus concentrées. Ce cluster comporte **55 270** photos, soit 69 % des photos prises.

Le deuxième cluster est loin derrière, avec **8359** photos (soit 10.5 % des photos prises), mais il reste néanmoins intéressant de constater que son centre se trouve exactement à *La Demeure du Chaos* (Musée l'Organe). Il s'agit d'un ancien relais de poste qui a été transformé en musée d'art contemporain. Cela explique donc la présence d'un grand nombre de clichés à cet endroit (photos touristiques lors de la visite du musée).

Le paramétrage choisi ici est encore grossier. Il ne nous permet que de détecter les clusters sur de grandes étendues. Par exemple, nous ne pouvons pas nous contenter de ce paramétrage pour tenter de découvrir les différentes zones touristiques de Lyon (un seul cluster a été généré pour la ville de Lyon toute entière). Il nous faut pour cela diminuer le quantile.

La figure 2.4 page suivante tente d'illustrer les zones touristiques de Lyon en diminuant le quantile et en affichant également les points en plus des centres des clusters (la couleur

choisie pour les points est celle de leur cluster d'appartenance).

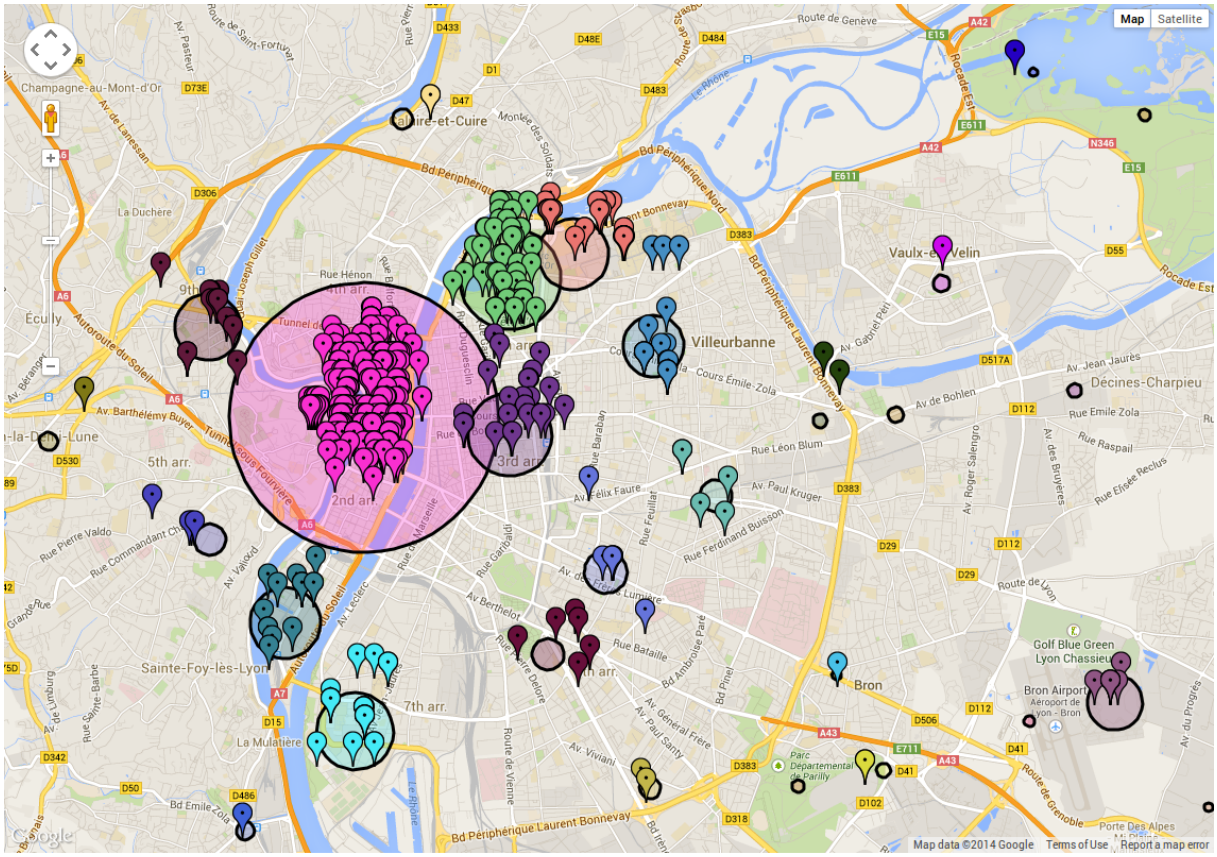


FIGURE 2.4 – Meanshift, $q = 0.05$ et $n = 15$

Cette représentation est déjà beaucoup plus intéressante. Le plus gros cluster (en rose-violet sur la figure 2.4) est toujours situé au même endroit (centré entre la place Bellecour et l'Hôtel de Ville) et il représente le cœur touristique de Lyon.

On voit également apparaitre d'autres zones d'intérêt, comme le cluster vert, représentant avec certitude l'attrait des touristes pour le parc de la Tête d'Or, le cluster violet foncé, situé autour de la gare de la Part-Dieu, le cluster cyan représentant l'attrait pour le quartier de Gerland, ainsi que le *Pôle de Commerces et de Loisirs de Confluence* (cluster vert foncé entre les deux fleuves), dont l'architecture très particulière incite à prendre des photos. On devine même le centre de ville de Villeurbanne (cluster bleu entre l'arrêt de métro République et Gratte Ciel).

Notons que si certains cercles (centres de clusters) n'ont aucun point à proximité, c'est parce que nous n'affichons qu'un pourcent de tous les points, alors que le calcul des clusters se fait lui sur l'ensemble de tous les points.

2.3 TODO