



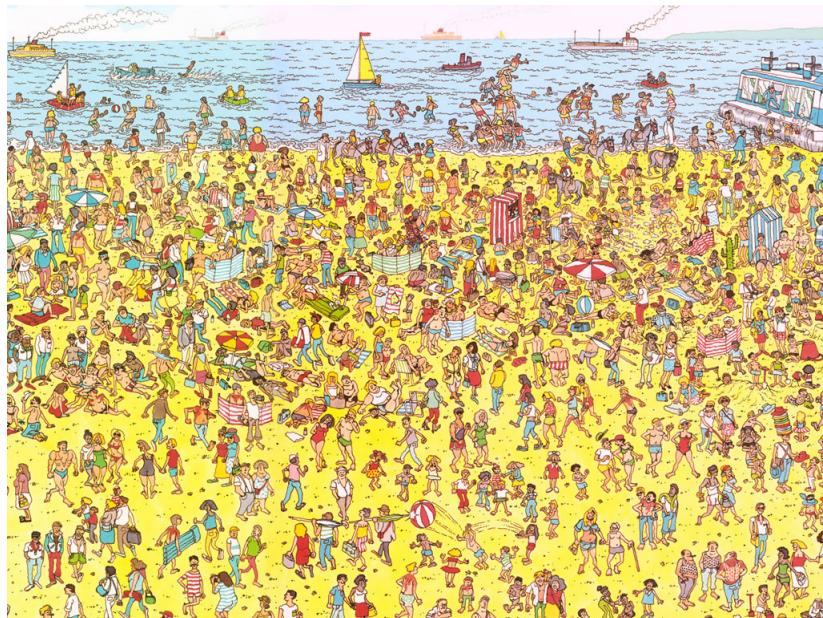
INSA Lyon
20, avenue Albert Einstein
69621 Villeurbanne Cedex

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

Sommaire

1	Préparation des données	1
1.1	Lecture du fichier sur Knime	1
1.2	Les informations « troll »	1
1.3	Les données particulières	1
1.4	Les mauvaises informations utilisateur	2
2	Visualisation	3
2.1	Introduction	3
2.2	Meanshift	4
2.3	Kmeans	7
3	résultats, axe d'amélioration et limites	9
3.1	interprétation des résultats	9
3.2	axes d'amélioration	9
3.3	limites	9
4	conclusion	11

1. Préparation des données

Ce projet nous a confrontés à plusieurs types de problèmes de données auxquels il a fallu trouver une solution ou faire un choix.

Parmi les données à problème, il y en avait trois types :

- les informations volontairement mauvaises et aberrantes, insérées par une personne bienveillante soucieuse de nous inciter à porter un regard critique sur le jeu de données qui nous a été confié ;
- les données particulières de type "Chine" qui perturbent les échelles en constituant des clusters très à part des autres ;
- les données involontairement aberrantes ou incomplètes insérées par les utilisateurs et transmises sous des formats variés.

1.1 Lecture du fichier sur Knime

Avant même de pouvoir commencer à regarder les données sur Knime pour faire une première préparation, il a fallu nettoyer les données de manière inattendue. En effet, les données commençant par des balises HTML entraînaient un plantage de la lecture du fichier pour Knime. Il a donc fallu encadrer de guillemets sur le même modèle que d'autres lignes fonctionnelles, les balises fautives.

1.2 Les informations « troll »

Nous avons fait le choix d'écartier sans remord les données appartenant à ce type. Il s'agissait de données où les dates et heures de prise ou d'upload de photos étaient impossibles (par exemple, le 89^{ème} jour du mois, ou encore avant l'invention de la photographie), mais aussi une ligne qui ne possédait ni longitude ni latitude. Ces lignes volontairement aberrantes étaient pour la plupart facilement identifiables comme des lignes « trolls », et non comme des erreurs utilisateur, grâce aux légendes et tags qui les composaient.

1.3 Les données particulières

Une donnée a particulièrement été repérée comme étant à l'écart des autres. Il s'agissait d'une photo prise en latitude et longitude 0. Ne sachant pas si cette ligne s'inscrivait dans la série des aberrations insérées volontairement ou s'il s'agissait d'une photo du véritable jeu de données, nous l'avons sérieusement considérée et c'est pourquoi elle est évoquée

dans cette catégorie à part. Nous avons néanmoins fait le choix d'écarter cette ligne tout comme les informations troll car elle perturbait à la fois les calculs de clusters, mais aussi les échelles, tant sous Knime pour la pré-analyse, que sous Google Map lors de l'affichage des clusters trouvés.

D'autres données auraient pu nous perturber si nous avions essayé de clusteriser en fonction de mots présents dans les tags ou les légendes, car certains d'entre eux ont été écrits dans des langues n'utilisant pas les caractères latins (chinois, japonnais, arabe, etc.), ces lignes auraient pu générer des erreurs si nous avions tenté d'analyser leur contenu textuel. Puisque nous n'avons pas eu le temps d'étudier les textes pour clusteriser, car cela aurait demandé une base de connaissance des synonymes au moins en français, nous avons laissé les lignes en langues étrangères.

1.4 Les mauvaises informations utilisateur

Enfin, les problèmes les plus importants viennent des informations ou justement du manque d'information utilisateur.

Il aurait difficile de clusteriser les éléments textuels car tous les utilisateur ne taggent ou ne légèdent pas leurs photos. De plus, au sein des textes, il n'est pas dit que les utilisateur parlent le même vocabulaire ou du même intérêt pour évoquer un lieu. Et quand bien même le feraient-ils, resterait encore le problème de gérer les fautes d'orthographe qui créent de nouveaux mots ou l'importance des majuscules/minuscules dans la comparaison de deux mots.

De plus un phénomène étrange a été observé sur environ 70 photos. Celles-ci auraient été uploadées avant même d'être prises. Même si l'erreur semble au premier abord être une information « troll » insérée artificiellement, elle se reproduit suffisamment pour pouvoir être considérée comme un comportement utilisateur atypique mais acceptable. Dans la mesure où ce comportement ne perturbait pas le calcul des clusters, nous avons choisi de le tolérer dans le jeu de 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.

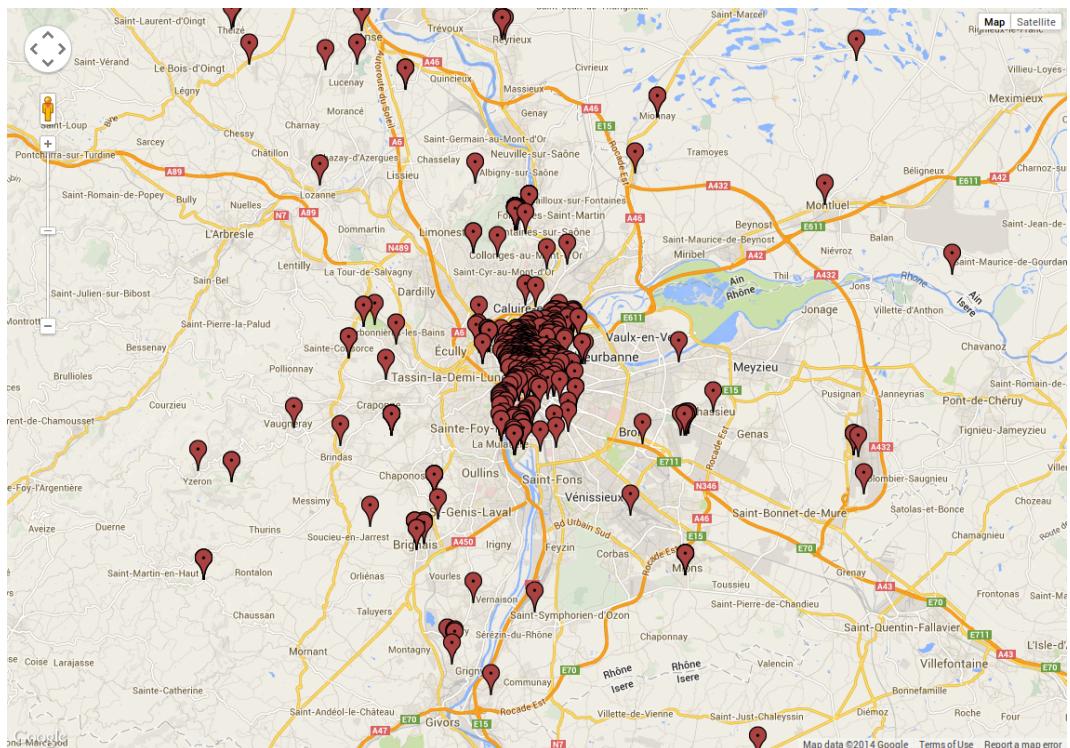


FIGURE 2.1 – Affichage d'un échantillon des points

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).

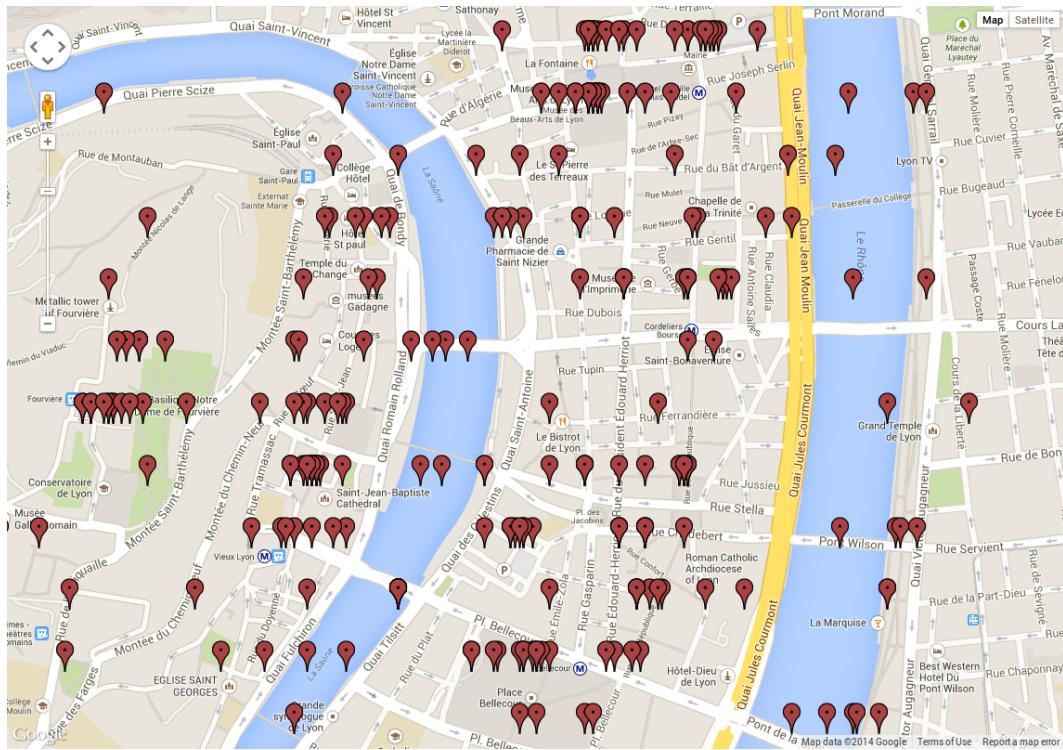


FIGURE 2.2 – Imprécision détectée lors du zoom

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.

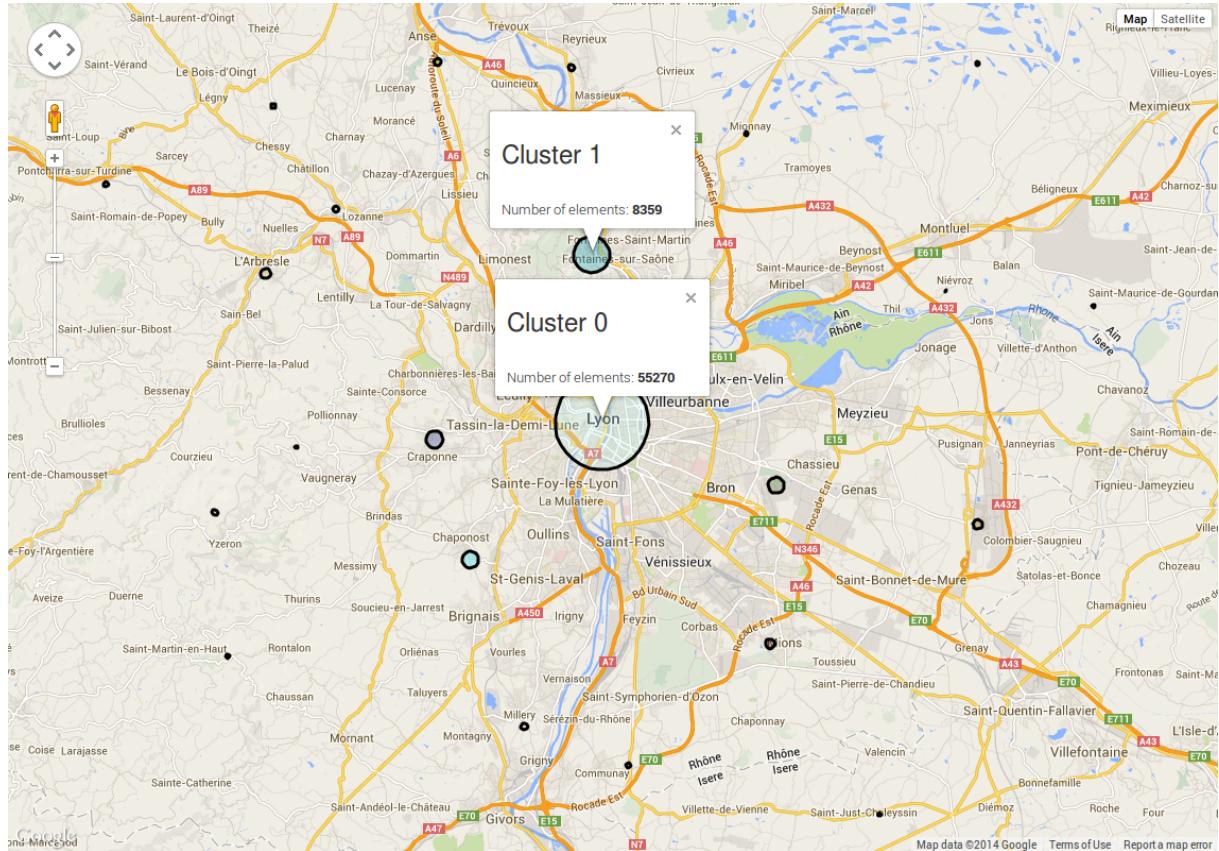
De manière très haut niveau, l'algorithme de clustering Meanshift peut être résumé comme suit :

- 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.



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

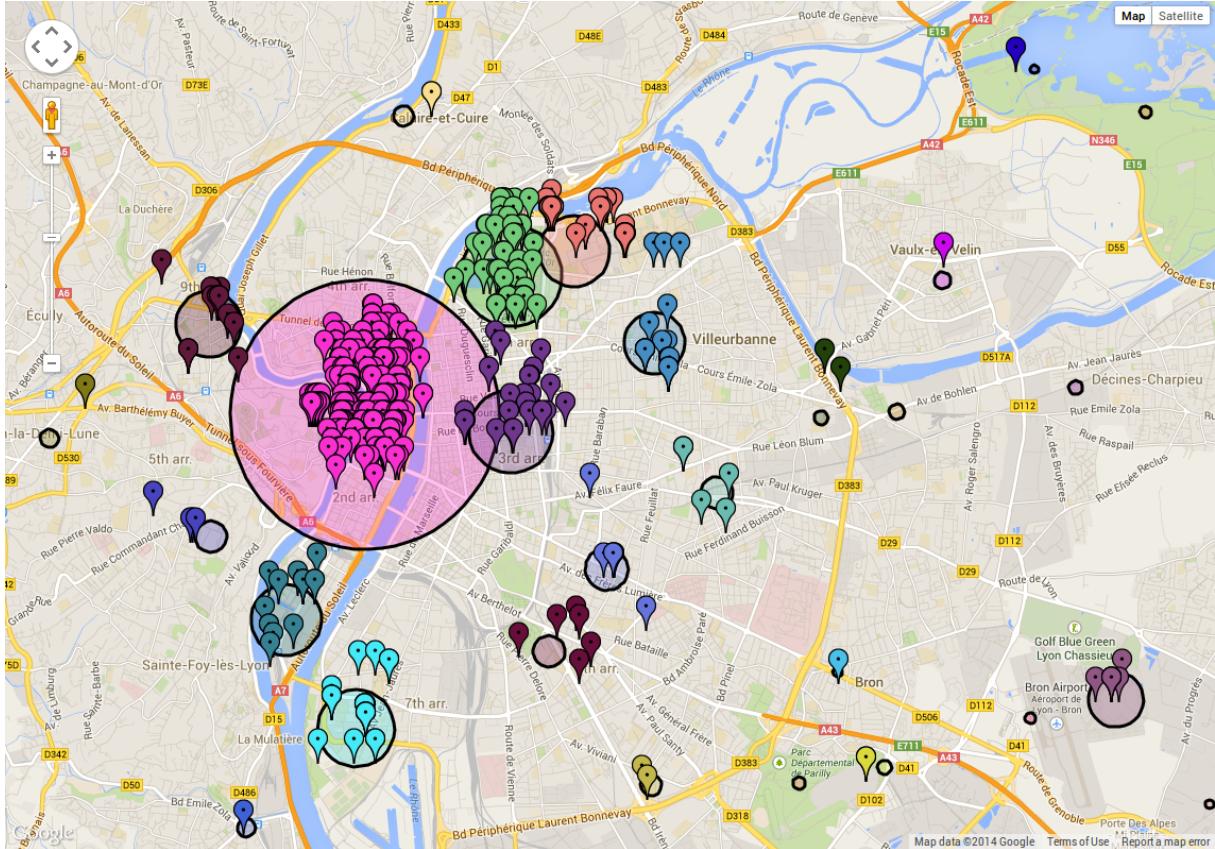


FIGURE 2.4 – Meanshift, $q = 0.05$ et $n = 15$

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 apparaître 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 Kmeans

Pour comparer nos résultats à ceux trouvés par d'autres méthodes de clustering, nous avons choisi d'utiliser la méthode de référence par excellence : Kmeans.

Cette méthode est une bonne proposition alternative car :

1. elle est relativement "simple" à mettre en place, elle ne demande comme paramètres qu'un nombre de clusters à trouver et les données à partir desquelles les générer.
2. c'est une méthode qui est utilisée par beaucoup de chercheurs pour comparer leurs résultats. Le fait que Kmeans soit familier de la plupart des chercheurs en data mining permet à ces personnes de plus facilement détecter les performances demandées par le jeu de données présenté.
3. c'est une méthode qui a quelques défauts, ce qui permet, par comparaison, de montrer l'intérêt de Meanshift.

Nous avons donc essayé de reproduire avec Kmeans les mêmes conditions que les deux exemples avec Meanshift en lançant un Kmeans avec un nombre de clusters voulus à peu près équivalent à ceux trouvés ci-dessus.

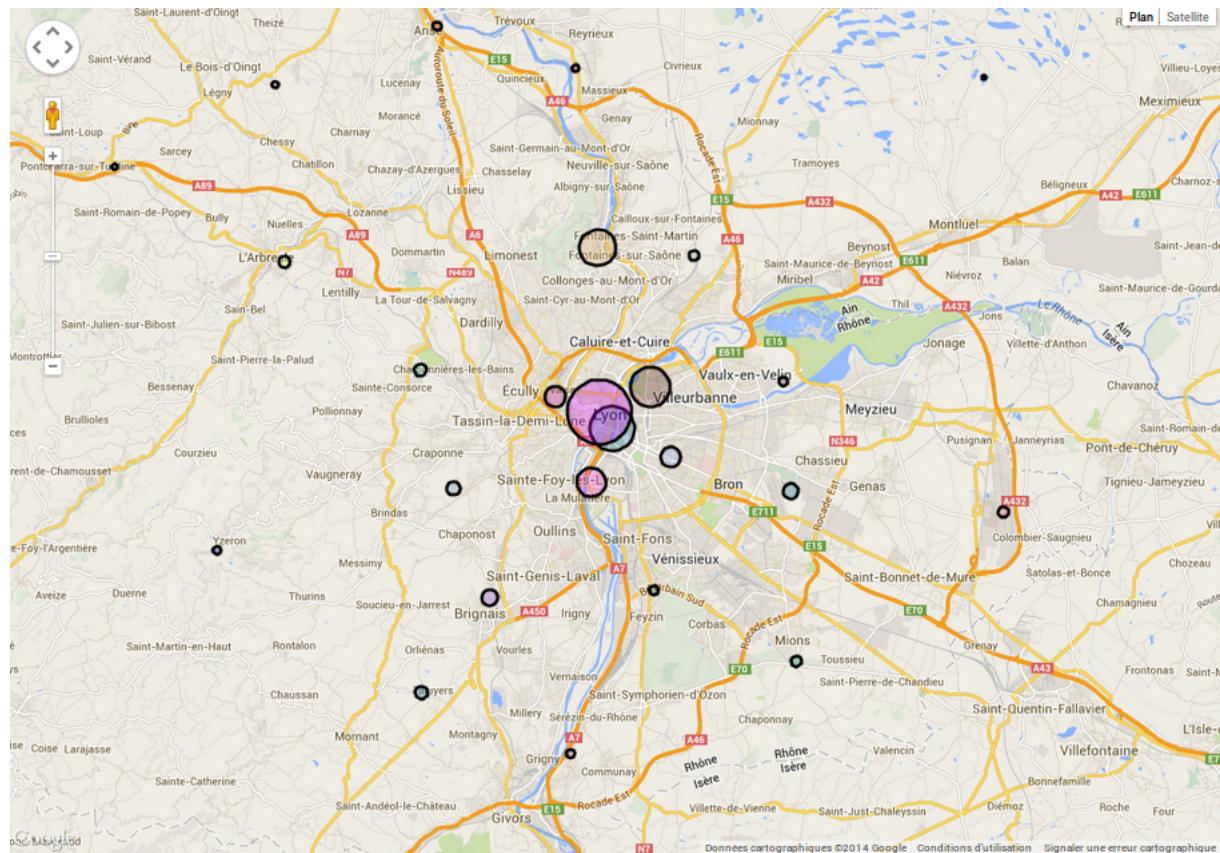
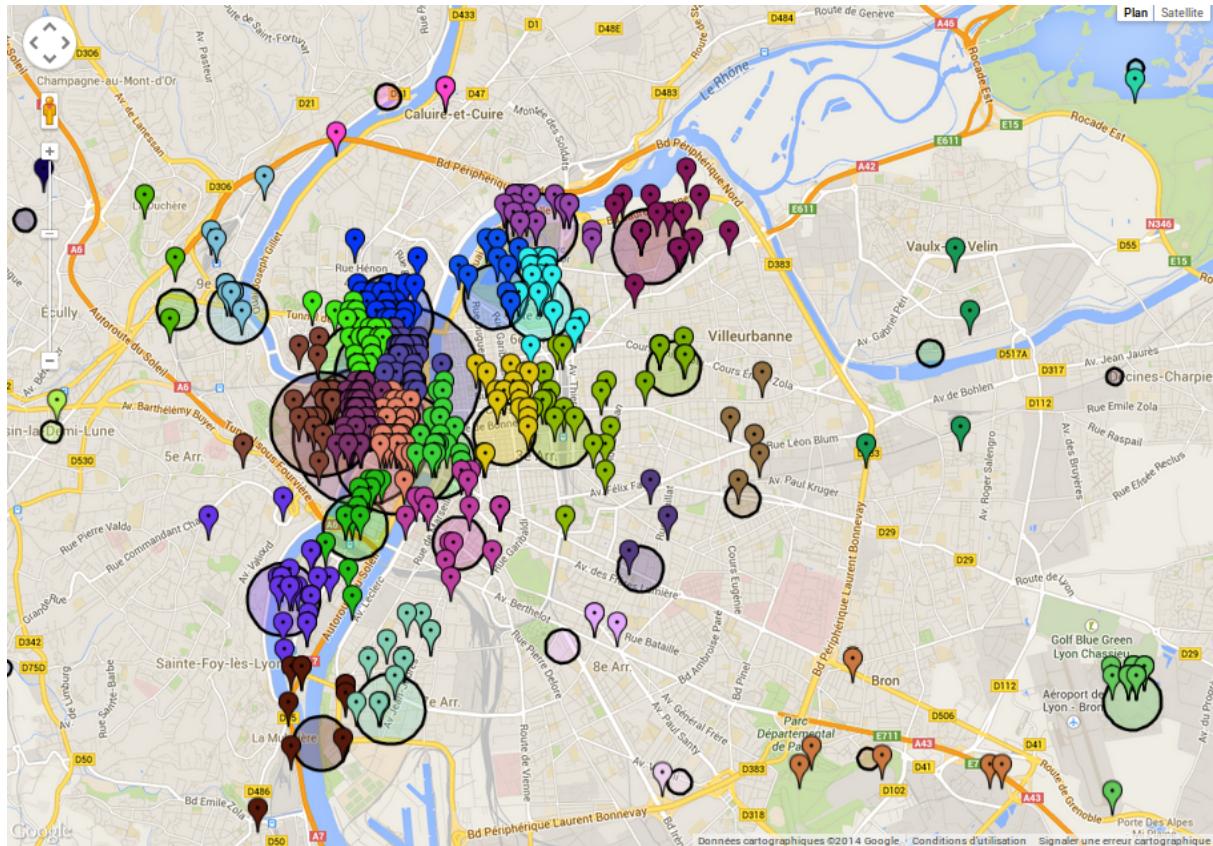


FIGURE 2.5 – K means, $nbClusters = 25$

On remarque que Kmeans trouve, comme Meanshift, les petites villes dispersées autour de Lyon. Cependant il parvient à couper le centre ville en plus de clusters. Il semble avoir déjà repéré le parc de la Tête d'Or et commencé à différencier les quartiers de la Part-Dieu de la presqu'île.

FIGURE 2.6 – K means, $nbClusters = 100$

Sur cette deuxième figure, Kmeans est dépassé en centre ville. Il découpe la presqu'île et le parc de la tête d'or de manière on peut dire arbitraire sous la forme d'un damier. Par contre les zones isolées comme l'aéroport de Lyon-Bron sont encore bien identifiées comme des clusters à part entière. On peut en déduire que Kmeans devient inefficace dans des zones à trop forte densité de données tandis que Meanshift ne sert pas à grand chose sans un minimum de précision.

3. résultats, axe d'amélioration et limites

3.1 interprétation des résultats

Notre étude est une clusterisation des espaces en fonction de la localisation géographique.

Elle peut déjà permettre à la commune du grand Lyon de repérer quels espaces ont tendances à être lieux de photos et quels autres ne le sont absolument pas et ainsi choisir d'appuyer les transports les espaces les plus pris en photo ou au contraire d'essayer de faire découvrir des endroits intéressants méconnus du public.

3.2 axes d'amélioration

En liant les dates avec la localisation, il pourrait être possible de détecter quels lieux auraient tendance à être plus pris en photo à une certaine période de la journée, du mois ou de l'année.

En liant les utilisateurs avec la localisation, il pourrait être possible de détecter des comportements de photos et ainsi de sugérer à un photographe d'autres endroits où ceux qui ont déjà photographié les mêmes lieux ont également été.

En liant dates, utilisateur et localisation, il pourrait être possible de déterminer des itinéraires de photos et pourquoi pas d'adapter les transports à ceux-ci.

Les analyses d'image, de légende et de tag pourraient également être utiles pour identifier ce qui est réellement pris en photo (enlever les selfies hors sujet, séparer les photos de bâtiments juste à côté de vues au loin, etc.).

3.3 limites

Le plus grand frein au projet a avant tout été le temps. Nous avons disposé de deux séances de 4 heures suivies de deux semaines sans séance pour travailler et rendre le compte rendu de projet. Travaillant en binôme nous avions donc un temps cumulé en

séance de 16 heures à nous deux. Malgré la possibilité de travailler hors séance nous n'avons pas non plus pu étendre beaucoup le temps passé sur le projet, car d'autres projets se présentaient également à l'ordre du jour.

Le deuxième frein a été les moyens. Pour pouvoir exploiter l'ensemble des données dont nous disposions il aurait fallut pouvoir analyser de manière sémantique les mots employés dans les textes et analyser les informations à partir des photos données. De même les données floues dans les positionnements empêchent de clusteriser de manière optimale, puisque ce manque de précision entraîne un fort décalage entre l'échelle des latitudes et des longitudes. Même l'interprétation des différents champs année mois jour heures minutes en type date, bien qu'elle ne soit pas aussi complexe qu'une analyse de texte demande du temps. Il faut trouver la bonne forme sous laquelle convertir et calculer, mais aussi la façon de convertir. Il n'est pas infaisable de faire tout cela, mais encore le temps nous a manqué.

Le troisième frein a été les données. Bien qu'ayant obtenu 84000 lignes d'information et ayant déjà du mal à toutes les afficher, c'était finalement assez peu pour en extraire des vraies tendances et il manquait des informations parfois.

Enfin le dernier frein, qui cette fois est pourtant positif, est que nous avions choisi de faire le sujet avec des outils différents de la plateforme knime à laquelle nous avions été formés (bien que nous l'ayons exploitée en début de projet pour visualiser la forme des données). Il a donc fallut intégrer dans notre projet la formation aux nouveaux outils employés (Web Api, google map, scikit sous python, méthode meanshift)

4. conclusion

Aline :

Le datamining m'a rappelé une science inventée par Isaac Asimov dans son livre Fondation. Cette science, nommée psychohistoire, se prétend capable de déterminer les actions futures de l'humanité grâce à des formules mathématiques issues de son histoire passée. Ces théories ne peuvent se prétendre justes, d'après son créateur, que si elles concernent des masses de populations suffisamment importantes pour ne pas être perturbées par les comportements individuels.

De même nous avons pu constater en essayant d'utiliser les méthodes de clustering que la fouille de donnée n'a de sens que si elle concerne un nombre suffisant de données, et qu'avec une fouille efficace il est possible de prévoir, avec un certain pourcentage de chance, des comportements.

Un jour, je suis tombée par hasard sur un article parlant d'une famille américaine qui aurait porté plainte contre une chaîne de magasin car leur technique de fouille de donnée utilisée pour envoyer de la publicité hyperspecialisée avait amené les parents à découvrir que leur fille était enceinte. Quelques jours après j'ai reçu une lettre du magasin auquel j'étais fidélisée me proposant des réductions sur des produits de consommation courante. J'ai été rassurée de voir qu'aucun des produits proposés ne me correspondait.

Tout cela m'a incitée à me poser des questions sur l'éthique de la fouille de données. Peut-on vraiment tirer toutes les informations des données qui nous sont présentées ou ne faut-il pas, à un moment, savoir laisser de la vie privée dans son étude et ne pas tenter le diable ? Je pense que cette question doit certainement concerner les hommes et les femmes chercheurs en datamining et que, comme dans toute autre science, la fouille de données doit être régulée par des règles d'éthique.