# RAPPORT DU MINI-PROJET : Médias Géo-localisés

FORMATION :

MIAGE MASTER 2 BDDA

COURS :

FSY

PROFESSEURS:

Sébastien FERRE

Simon MALINOWSKI

BINOMES DU GROUPE :

Apéfa Kékéli Renée AFONOUVI

Steicy Dona THIAW

ANNEE SCOLAIRE : 2022 - 2023

## Sommaire



## Table des figures

# I. Préparation des données géo-localisées

Afin de nettoyer proprement les données et de supprimer les anomalies dans la base de données, nous avons affectué les différents opérations que nous résumons de la facon suivante:

Etape 1 : Elimination des doublons avec GroupBy

Nous avons constaté qu'après cette étape, nous sommes passés de **54800** lignes de données à seulement **4195** lignes

Etape 2 : Filtrage des données selon l'année

Cette étape a eu lieu après qu'on constaté avec la réalisation des statistiques que certaines photos dans les données datait des années 1900. Nous avons choisi de ne prendre en compte que les photos datant de 2000 et plus.

Ce qui a réduit notre base de données à **4164** lignes.

Etape 3 : Choix des zones géographiques de forte concentration avec Geo-Coordinate Row Filter

Cette sélection nous a permis de nous concentrer sur les zones à forte concentration d'actvité qui est délimité par les quartiers présentés sur la figure 1. Le nombre de lignes de données est maintenant passé à **4124**.

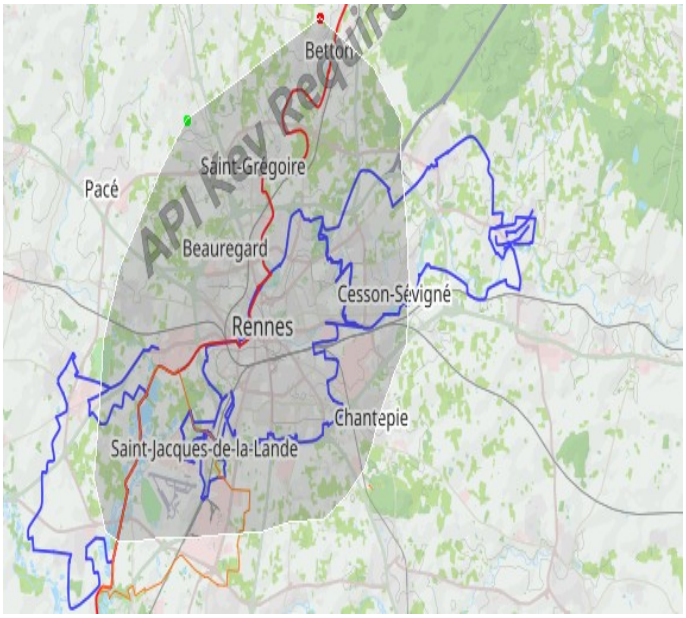


*Figure 2: Sélection du périmètre des données*

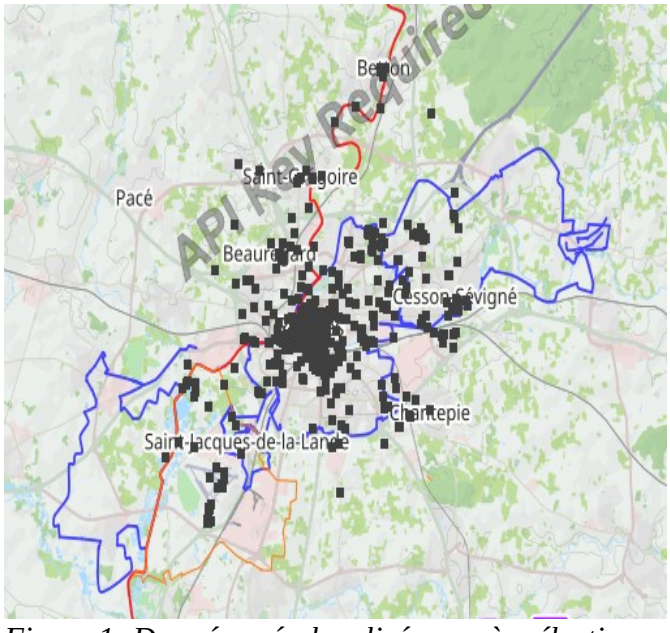


*Figure 1: Données géo-localisées après sélection*

**Etape 4** : Pour finir cette première partie nous avons supprimé les colonnes que nous avons considéré comme redondants car n'apportant aucune information supplémentaire à l'aide du nœud Column Filter( Voir Figure3).

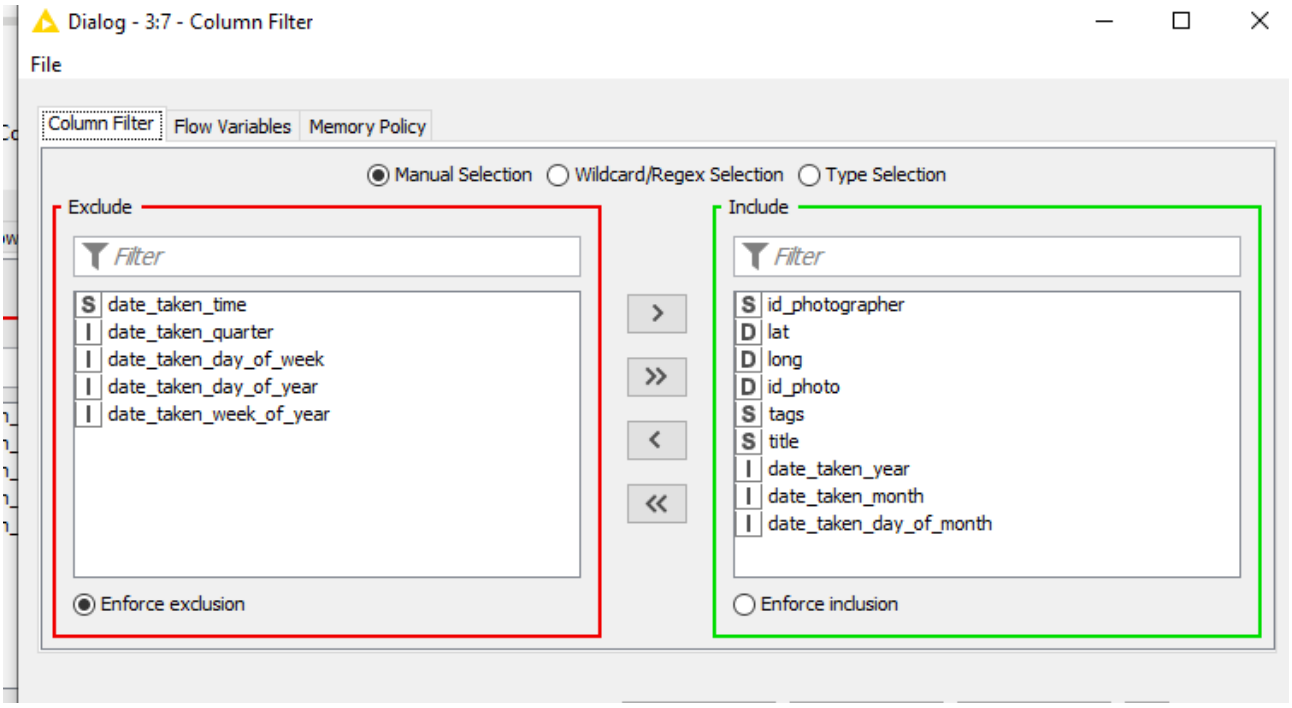


*Figure 3: A gauche les colonnes redondants de date et à droite les colonnes préservés*

*Nous avons obtenus à cette dernière étape de néttoyage des données, **4124** lignes de données, ce qui représente 4.5% de nos données de départ qui étaient au nombre de **54800.***

# II. Utilisation du clustering pour découvrir les points d'intérêts

**Etape 1** : Choix des colonnes et choix du nombre de clusters

Nous avons testé en premier lieu de considerer les dates comme critère de regroupement afin d'identifier les lieux les plus fréquentés de Rennes par période et choisi 8 comme nombre de clusters initiales . Cet essai nous a montré que les points d'intérêt ne dépendait en rien de la période de l'année.

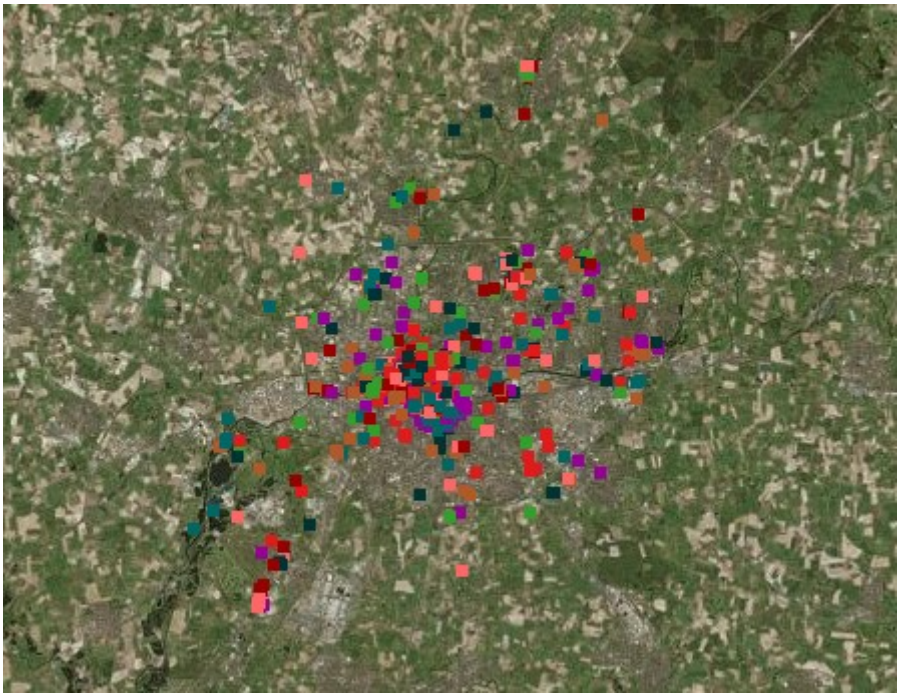

*Figure 4: Clusters indistinguables avec le clustering basé sur le mois*

Dans un second temps nous avons plutôt considéré la latitude et la longitude pour établir les clusters. Nous avons par ailleurs variées le nombre de clusters entre 8 et 24, ce qui nous a permis de déterminer la valeur optimale de clusters qui est 12. La dernière manipulation que nous avons faites concerne l'inititatialisation des centroïdes. Le choix de centroïdes aléatoires nous a fourni un meilleur clustering que le choix des k premières lignes avec k le nombre de clusters.

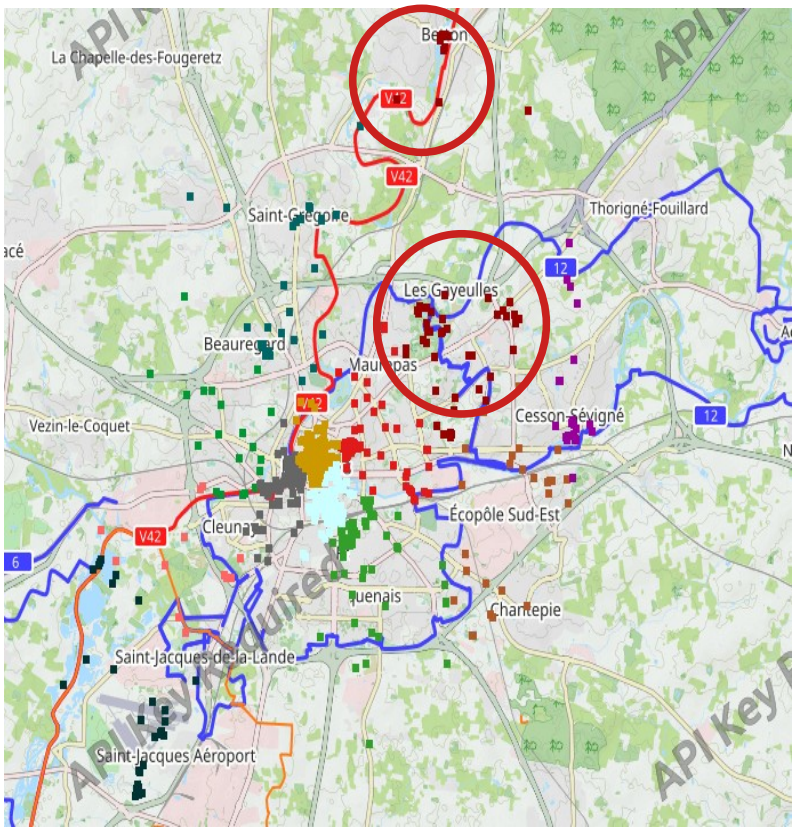


*Figure 5: Résultats avec les k-premières lignes, On constate par exemple que les photos localisées à Betton et à Les Gayeulles sont dans le même cluster*

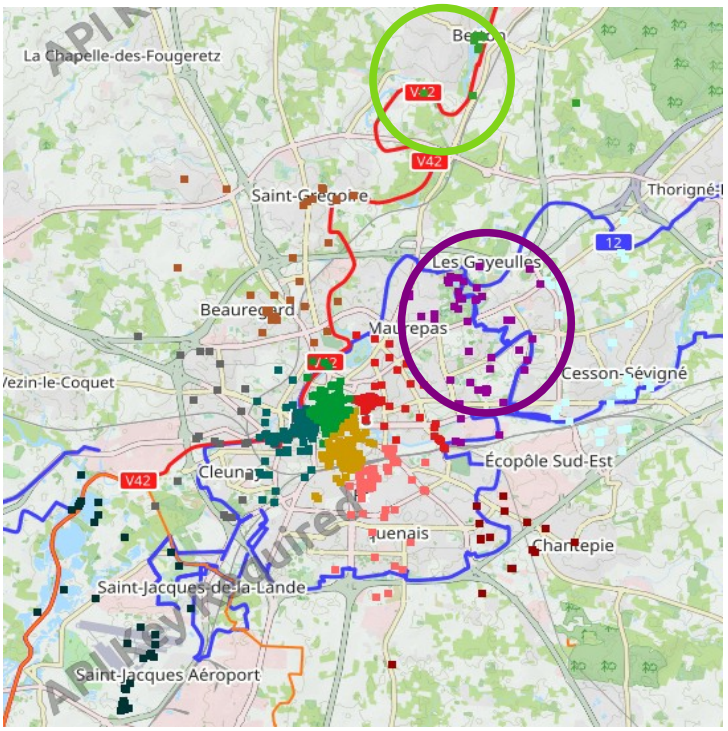


*Figure 6: Résultats avec les centroïdes aléatoires, Les photos prises à Betton sont distingués des photos prises à Les Gayeulles*

Etape 2 : Analyse des résultats

| Régions identifiant les clusters | Nombre de photos | Pourcentage de photos | Commentaires |
|---|---|---|---|
| Beauregard | 85 | 2.5% | Concentration constaté autour de Louis Chouinard et Linon |
| Les Gayeulles | 1594 | 39% | Le parc des Gayeulles fait l'objet de l'intèressement dans cette zone |
| Jeanne d'Arc | 192 | 5% | Le prinicipal point d'intérêt observé est le parc Thabor |
| Betton | 19 | 0.5% | Le point d'intérêt observé est l'étang de Betton |
| Longs Champs | 69 | 2 % | Les photos géo-localisées se retrouvent sur le campus de beaulieu et sur Cesson ViaSilva |
| Cesson-sevigné | 210 | 5% | Les photos ont été prises principalement au parc de Champagné et le jardin de Bourchevreuil. |
| Jacques Quartier | 717 | 17% | Aucun commentaire particulier |
| Charles de Gaulle | 308 | 7% | La concentration dans cette région est due aux photos prises autour de la gare et du Champ de Mars |
| Moulin du Comte | 103 | 3% | Aucun commentaire particulier |
| Bourg L'Evesque | 262 | 6% | La plupart des photos ont été prise autour du Quai Saint-Cast, Saint-Cyr et la cathédrale Saint-pierre |
| Rennes -centre | 455 | 11% | En tant que région avec la plus forte concentration, on observe que les photos ont été prises à 3 endroits principalement: Saint-Anne, République et le centre-ville |
| Saint-jacques-de-la-lande | 81 | 2% | L'avenue de l'aéroport, le moulin d'Apigné et la rue Jules Vales ont des regroupements de photos considérables |

# III. Description des points d'intérêts avec les motifs ensemblistes

Etape 1 : Création des documents et nettoyage des données texte

Nous avons en premier lieu supprimés les données qui avaient une valeur null pour la colonne ***tags*** . Ce qui réduit notre base de données à **2051** lignes.

Comme énoncé dans le projet nous avons ensuite suivi les étapes de nettoyage des données texte en fixant les différens paramètres nécessaires:

- la création des documents avec *String To Document* avec le titre de la photo comme titre du document, les tags comme contenu et le cluster correspondant à la photo comme categorie du document
- la suppression des ponctuations dans le contenu des documents avec *Punstuation Erasure.*
- La suppression de caractères de **moins de 3 lettres** avec N Chars Filter.
- La suppression des mots contenant des caractères numériques avec *Number Filter*.
- La conversion de tous les mots des contenus des différents document en miniscules avec *Case Converter*.
- La suppression successive des mots vides en langue francaise et en langue anglaise avec *Stop Word Filter*.
- La suppression des documents vides après ses traitements avec *Row Filter* . Nos lignes de données sont maintenant au nombre de **1923**
- La stemmatisation des différents mots avec *Snowball Stemmer et le choix de du Stemmer de Porter*
- La création du vocabulaire des mots contenus dans la base de données avec *Bag Of Words Creators* qui a également sorti en résultat des lignes de données contenant les différents mots retrouvées dans les différents documents

Etape 2 : Vectorization des données texte

Nous avons utilisé le nœud *Document Vector* qui s'est chargé n'ont seulement de propositionaliser les valeurs des termes du vocabulaires mais aussi de les utiliser pour représenter chaque document sous forme de vecteur.

Ensuite nous avons pu récupérer l'information de la catégorie de chaque document avec *Catefory To Class* avant de réaliser des vecteurs compressés des documents en une seule colonne avec *Create Bit Vector*.

Etape 3 : Création de motifs ensemblistes

Afin d'accomplir l'objectif de cette tâche, nous avons réalisés *12 nœuds Row Filter* afin de segmenter la base de données en 12 en fonction du cluster auquel appartient le document vectorisé. Nous avons ensuite utilisé le nœud Item Set Finder pour chaque cluster.

***Le support fixé pour chaque cluster est de 20%.***

Nous avons donc ci-dessous un exemple qui illustre les résultats obtenus dans le workflow et qui présente les motifs assemblistes pour des clusters correspondants aux régions Betton et les Gayeulles.

Item Sets - 3:81 - Item Set Finder (Borgelt)

File Edit Hilite Navigation View

Table "default" - Rows: 4 | Spec - Columns: 4 | Properties | Flow Variables

| Row ID | ItemSet | ItemSe... | ItemSe... | Relativ... |
|---|---|---|---|---|
| Row0 | [eau] | 1 | 3 | 23.077 |
| Row1 | [penséenomad] | 1 | 3 | 23.077 |
| Row2 | [betton,illeetvilain] | 2 | 4 | 30.769 |
| Row3 | [betton,bretagn] | 2 | 4 | 30.769 |

*Figure 7: Motifs ensemblistes du cluster caractérisant la région "Betton"*

Item Sets - 3:86 - Item Set Finder (Borgelt)

File Edit Hilite Navigation View

Table "default" - Rows: 5 | Spec - Columns: 4 | Properties | Flow Variables

| Row ID | ItemSet | ItemSetSize | ItemSetSupport | RelativeItemSetSupport% |
|---|---|---|---|---|
| Row0 | [eau] | 1 | 4 | 20 |
| Row1 | [franc,renn] | 2 | 4 | 20 |
| Row2 | [fleur] | 1 | 6 | 30 |
| Row3 | [gayeul,parc,renn] | 3 | 5 | 25 |
| Row4 | [bretagn,renn] | 2 | 8 | 40 |

*Figure 8: Motifs ensemblistes du cluster caractérisant la région "Les Gayeulles"*

# IV. Informations complémentaires

Nous avons pu tester le « hiérarchical clustering » que nous avons pas considéré au final en raison de sa faible performance par rapport au k-means dans le projet.