



**Université d'Orléans**

**UFR Lettres Langues et Sciences Humaines**

**Département de Géographie**

**Master 1 GLET**

**UE : PROJET TUTEURE**

**Manuel d'utilisation :**

**Clustering Point (extension QGIS) : description, fonctionnement  
et application**

**Réalisé par :**

**BA Mamadou**

**BALDE Alassane**

**Sous la direction de :**

**Mr NEDJAI. R**

**Mr AZAROUAL. A**

**Année universitaire 2021/2022**

## **INTRODUCTION GÉNÉRALE :**

Un système d'information géographique est un outil informatique permettant de représenter et d'analyser toutes les choses qui existent sur terre ainsi que tous les évènements qui s'y produisent (selon le site ESRI France). Il permet d'acquérir, d'organiser, de gérer, de traiter, restituer et analyser des données à références spatiales sous différentes formes. C'est un système informatique permettant à partir de diverses sources de données, de rassembler, d'organiser, d'exploiter et présenter des informations localisées géographiquement dans l'objectif d'aménagement et de gestion des territoires. Il fait référence aux traitements d'informations liées à l'espace et au temps.

Le fait que la géographie constitue une science carrefour, cela offre aux SIG une diversité de domaines d'applications tels que l'environnement , le développement durable, la gestion des risques et catastrophes, l'agriculture, les réseaux, aménagement du territoire, la santé, la sécurité... Ils sont utilisés par une multitude d'acteurs tels que les collectivités territoriales, les administrations, les entreprises, le secteur public, les écoles, universités, les bureaux d'études...

Aujourd'hui il existe une diversité de logiciels utilisés dans les SIG dans lesquels certains sont en accès libre (Open Source : QGIS, GRASS GIS, SAGA, POSTGIS, DIVA-GIS...) et d'autres sont payants (ArcGIS, CARIS GIS, MapInfo, ENVI CLASSIC ...).

Notre travail concerne une extension du logiciel QGIS : le Clustering Point. L'objectif du travail constitue dans un premier temps de faire la description détaillée, expliquer son fonctionnement et enfin l'appliquer dans un domaine.

### **I°) L'EXTENSION CLUSTERING POINT**

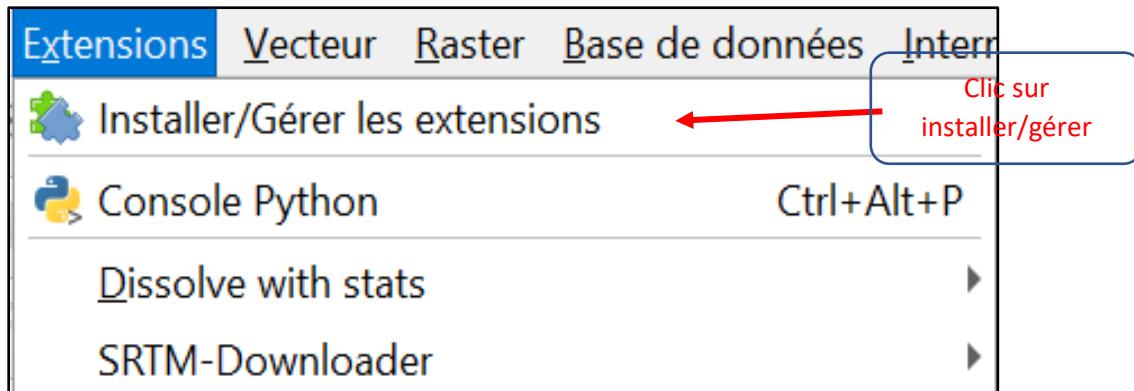
#### **1) Définition**

Clustering Point est une extension de QGIS qui offre un ensemble d'outils de cluster pour un regroupement automatique d'une couche (points, polygones, images) dans Quantum GIS minimisant ainsi la distance intra-groupe et maximisant la distance entre les groupes. Elle a été créé par Johannes Jenkner, il assure aussi sa maintenance. La dernière version de l'extension est 5.1 mise à jour le 5 mars 2022 à 15h 33' dont la version minimale de QGIS requise est la 3.16.0. Le clustering est une technique d'apprentissage automatique non supervisée de regroupement de points de données en fonction la distance, la similarité ou l'appartenance.

Il existe **trois algorithmes intrinsèquement différents** parmi lesquels l'utilisateur peut choisir. Premièrement nous avons les **algorithmes de classification non hiérarchique (K-Means et Fuzzy C-Means)**. Le premier (clustering K-Means) initialise de manière aléatoire les centres de cluster et réaffecte les membres du cluster jusqu'à ce que les centres cessent de bouger ; le deuxième (Fuzzy C-Means) génère en outre des probabilités d'appartenance au cluster pour des points individuels. Deuxièmement, il y a le clustering hiérarchique agglomératif qui commence avec autant de clusters qu'il y a de points et fusionne progressivement les clusters individuels selon une certaine fonction de lien.

## 2) Installation de l'extension sur QGIS

Dans l'interface de QGIS il existe un onglet dénommé « extension » dans lequel on fait clic droit pour dérouler le menu. Une fois faire ça, on clique « Installer/Gérer les extensions »



**Figure 1: Onglet pour l'installation de l'extension**

On va d'abord sur l'onglet « Non Installées » et sur la barre de recherche, on cherche le nom de l'extension : « ClusterPoints » et on lance l'installation en cliquant sur « installer le plugin »

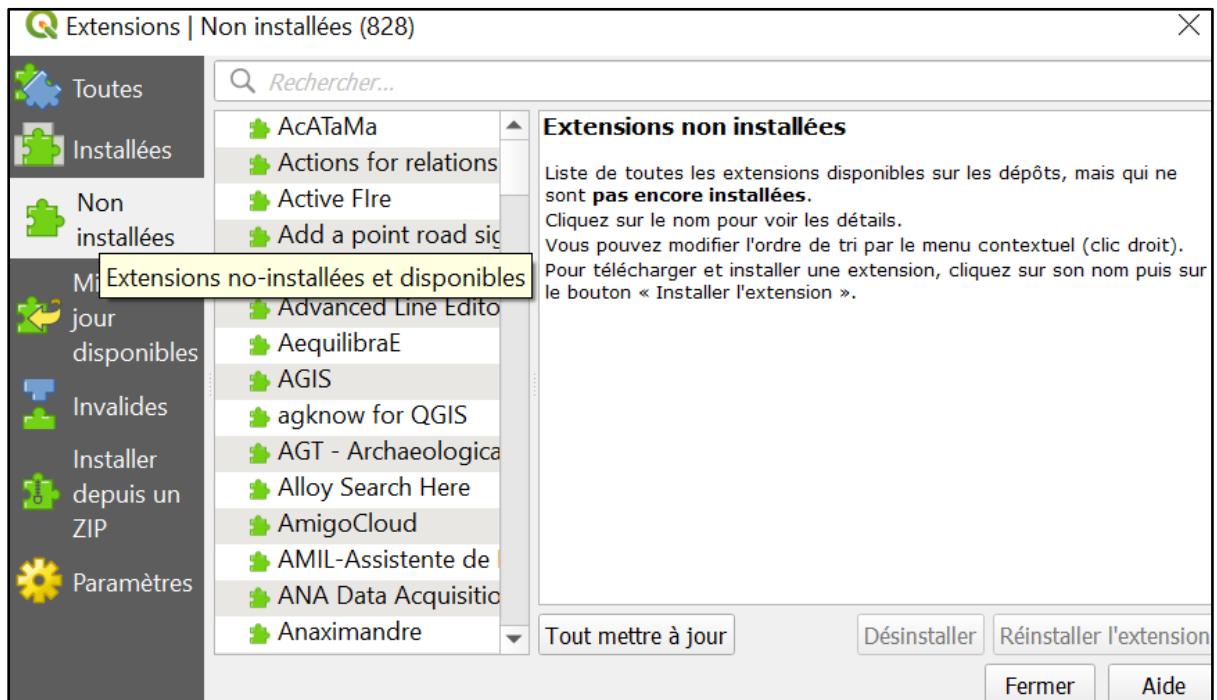


Figure 2 : fenêtre des extensions

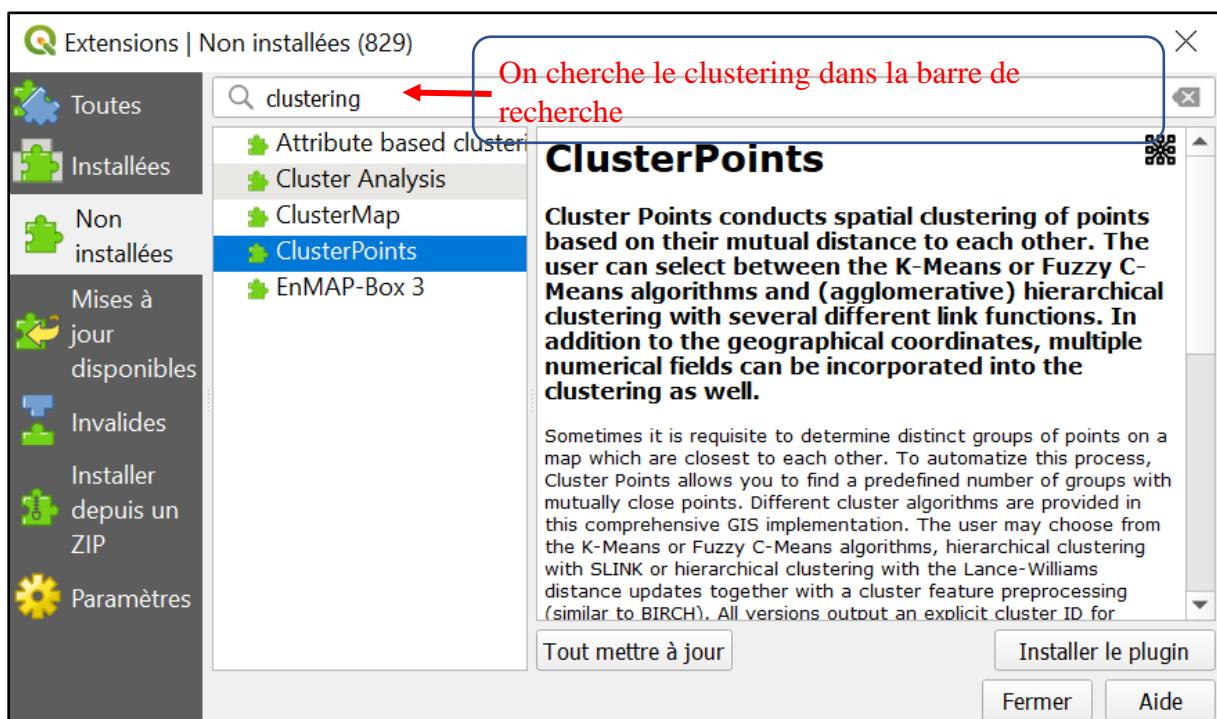
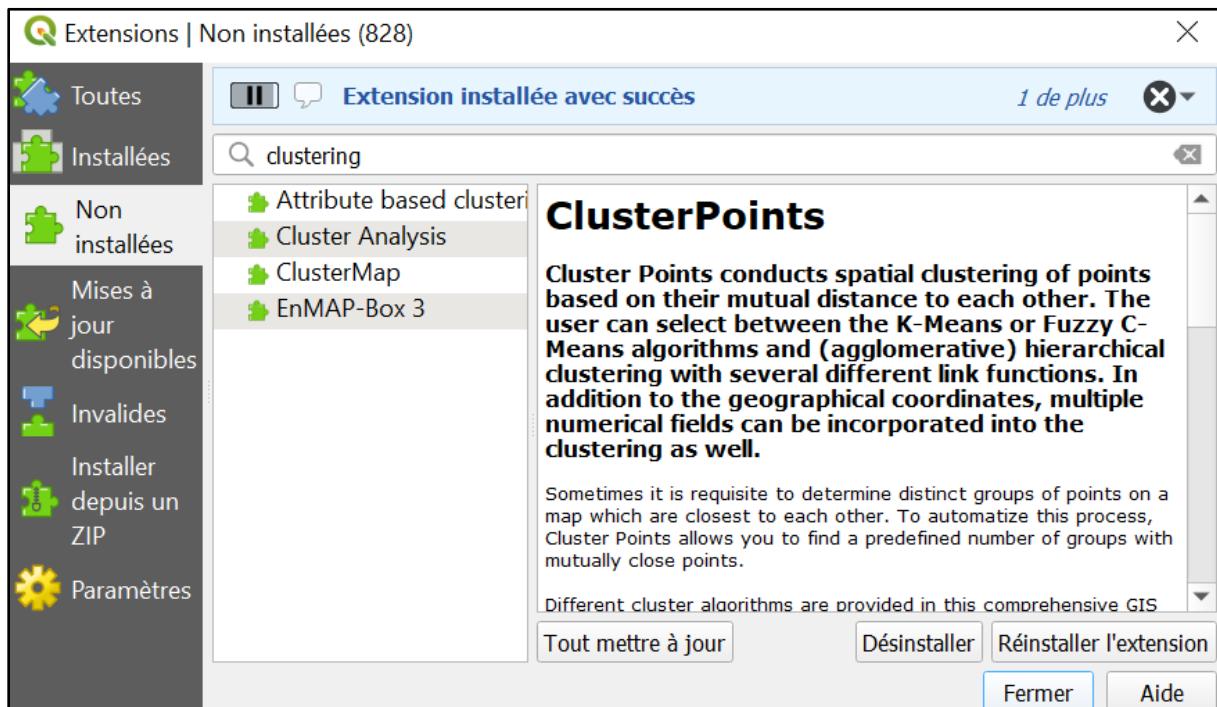


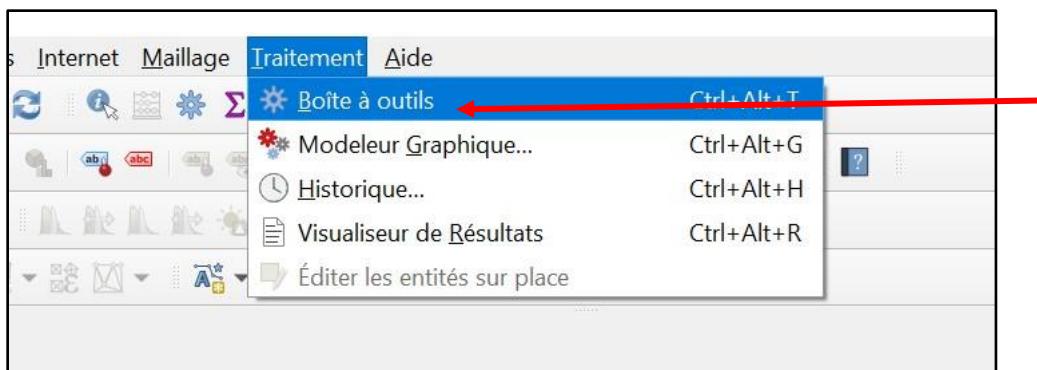
Figure 3 : Sélection de l'extension

On attend jusqu'à voir sur la barre de recherche : « Extension installée avec succès » et on ferme la fenêtre.



**Figure 4 : Installation de l'extension**

Une fois installer le plugin, l'extension se met directement sur la Boite à outils. Cette dernière se situe au niveau du menu « Traitement »



**Figure 5 : Boîte à outils pour accéder à l'extension**

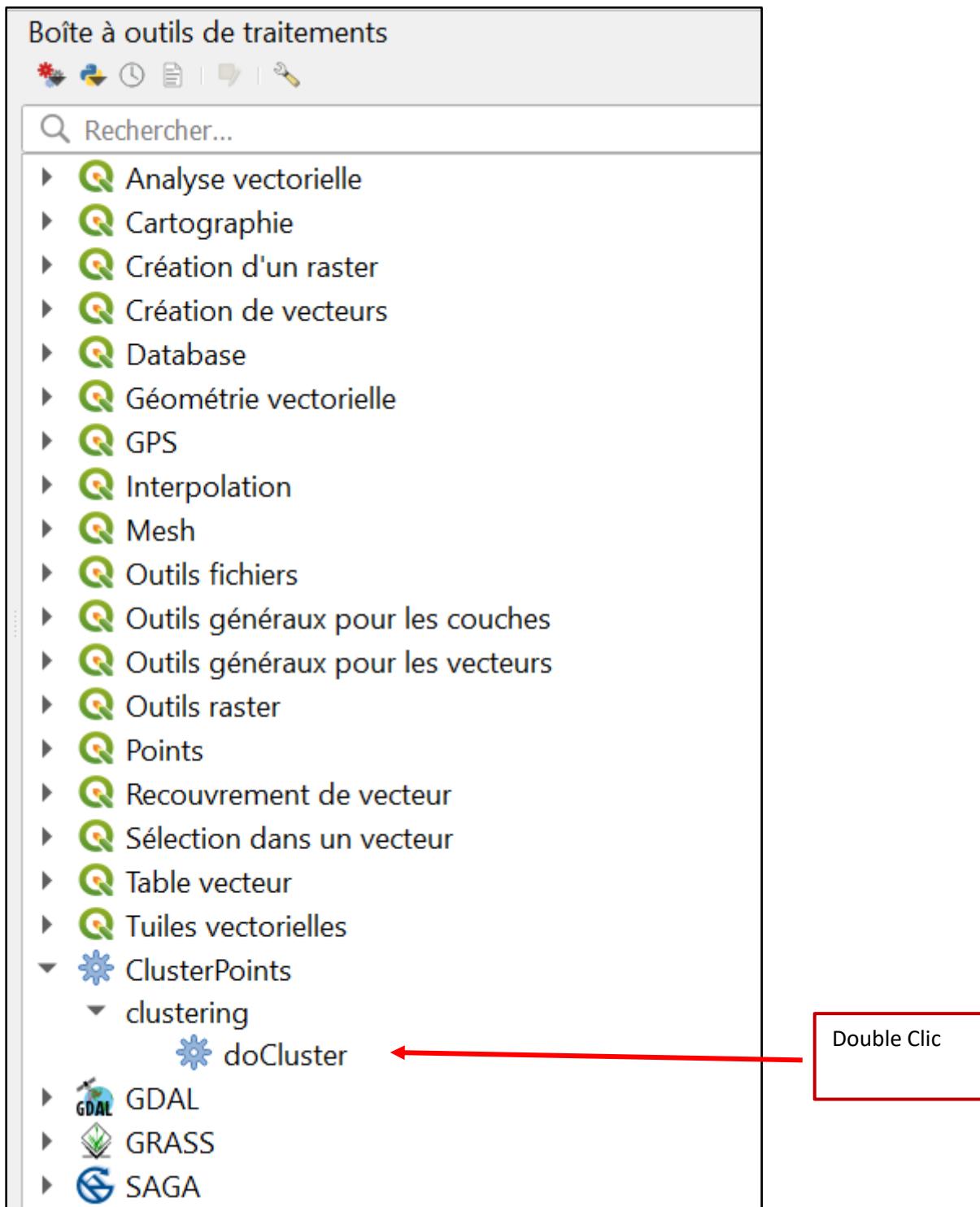
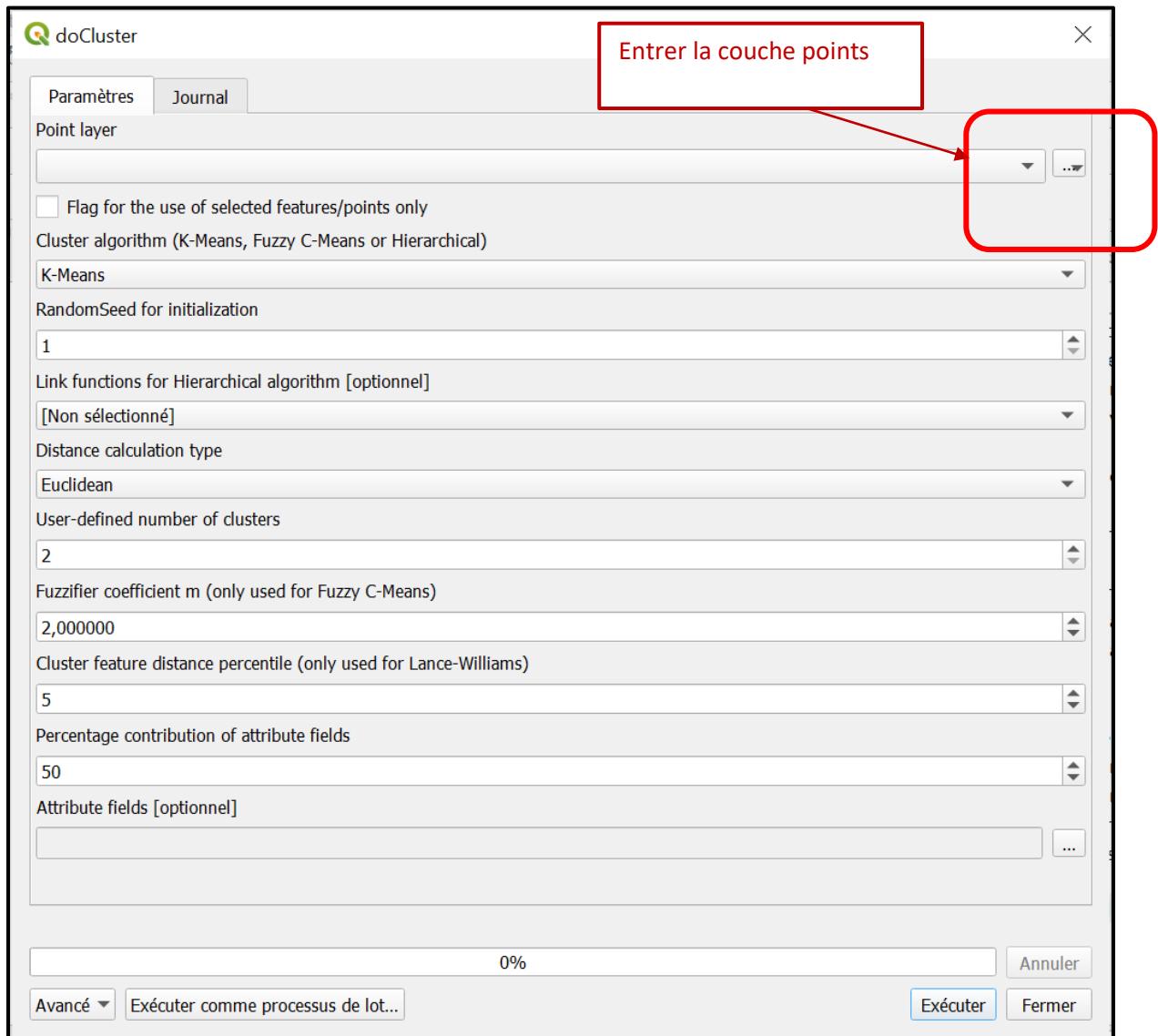


Figure 6 : Boite à traitement de QGIS

### 3) Interface de l'extension



**Figure 7 : Fenêtre de l'extension clustering pour le choix des paramètres**

## II/ DESCRIPTION ET FONCTIONNEMENT

La première chose à faire est d'ajouter la couche d'entrée (Point layer) c'est-à-dire la couche à partir de laquelle les clusters seront effectués.

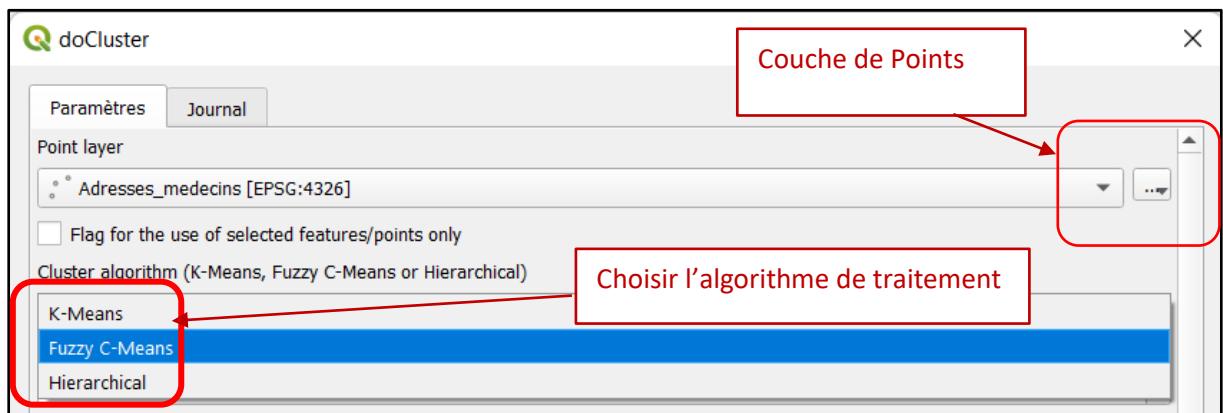
### 1/ Les algorithmes

Il faut ensuite choisir l'algorithme de classification qui sont au nombre de 03 :

**1.1 K-Means** : est un algorithme d'apprentissage non supervisé dédié au clustering non hiérarchique, cet algorithme itératif minimise la somme des distances entre les points. Il change les points de chaque cluster jusqu'à ce que la somme ne puisse plus diminuer. Le résultat est un ensemble de clusters compacts et clairement séparés.

**1.2 Fuzzy C-Means :** C'est un algorithme qui permet de générer les probabilités d'appartenances au cluster pour des points individuels. Il s'agit ici d'une amélioration et d'ajustements concernant la représentation d'appartenance, sur les techniques de fuzzification ou défuzzification ainsi que sur la fonction de distance entre les éléments.

**1.3 Hierarchical :** le clustering hiérarchique agglomératif ou divisif qui commence avec autant de clusters qu'il y a de points et fusionne progressivement les clusters individuels selon une certaine fonction de lien. Il consiste à créer une arborescence de cluster de manière ascendante ou descendante.



**Figure 8 : Zoom sur le choix de l'algorithme de clustering**

## 2) Les fonctions de liens

Elles permettent de modéliser le logarithme. Les modèles utilisant cette fonction sont des modèles log-linéaires, des rapports de chances. Pour structurer ce regroupement, on se base à partir des modèles divers en termes de fonctions d'organisations et de types de relations. Ces fonctions de lien sont :

**2.1 Single linkage (SLINK) :** On utilise cette fonction de lien pour effectuer une analyse de cluster à lien unique (plus proche voisin) sur un coefficient de dissemblance arbitraire.

**2.2 Single (Lance-Williams) :** C'est une méthode de liaison simple qui contrôle exclusivement la similitude des voisins les plus proches.

**2.3 Complète (Lance-Williams) :** elle est aussi dite méthode du voisin le plus éloigné, désigne un regroupement agglomératif qui définit les dissimilarités entre les clusters comme étant la distance maximale entre deux points quelconques des deux clusters

**2.4 Médian (Lance -Williams)** : ou méthode des centroïdes équilibrée, la distance entre deux clusters est égale à la distance médiane (centroïdes géométriques) d'un cluster et une variable de l'autre.

**2.5 Average (Lance-Williams)** : ou la méthode de liaison équilibrée, elle désigne la distance moyenne entre une variable d'un cluster et une variable de l'autre cluster.

**2.6 Ward's (Lance-Williams)** : Ici la distance entre deux cluster est égale à la somme des écarts quadratiques entre les points et les centres. Elle permet de regrouper deux cluster d'une partition pour obtenir une partition beaucoup plus agrégée. Cette fonction agglomère les deux classes en minimisant l'inertie intra classe

**2.7 Centroid (Lance-Williams)** : Également appelé méthode de groupe pondéré, elle met en évidence la distance entre les centres de gravité des paires de clusters. Cette distance séparant deux groupes est la distance entre les centres ou les moyennes de groupes.

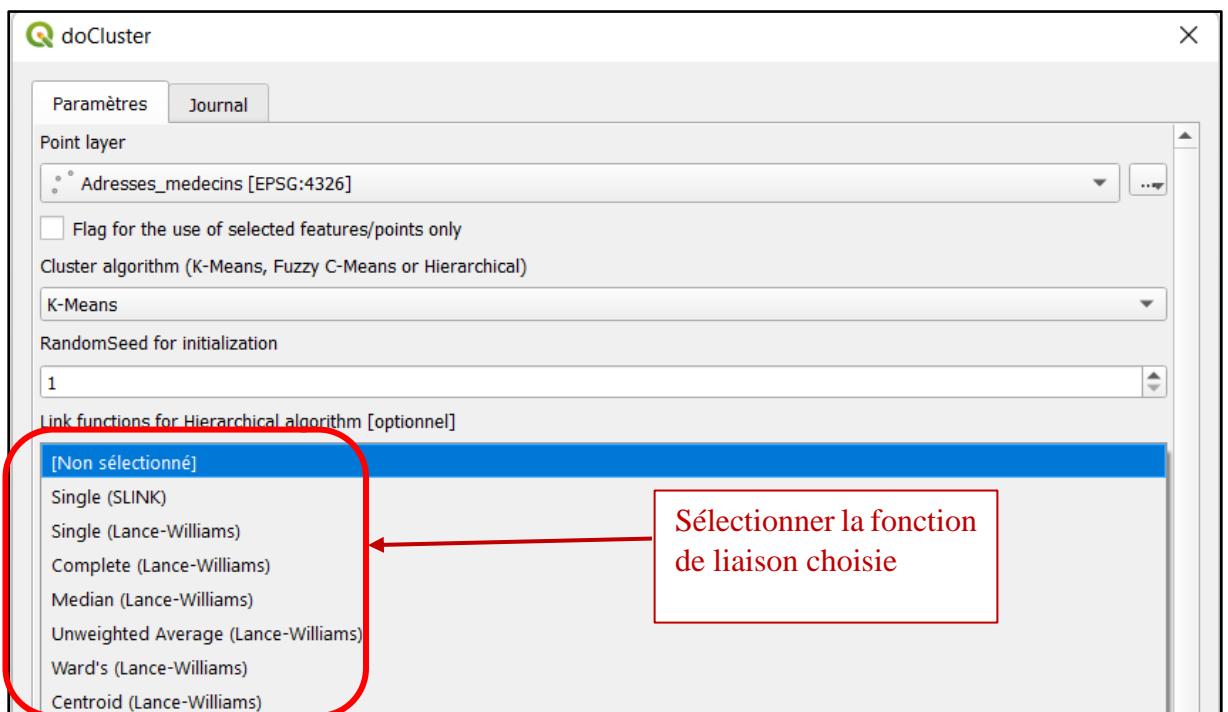


Figure 9 : Etape de choix de la fonction de lien

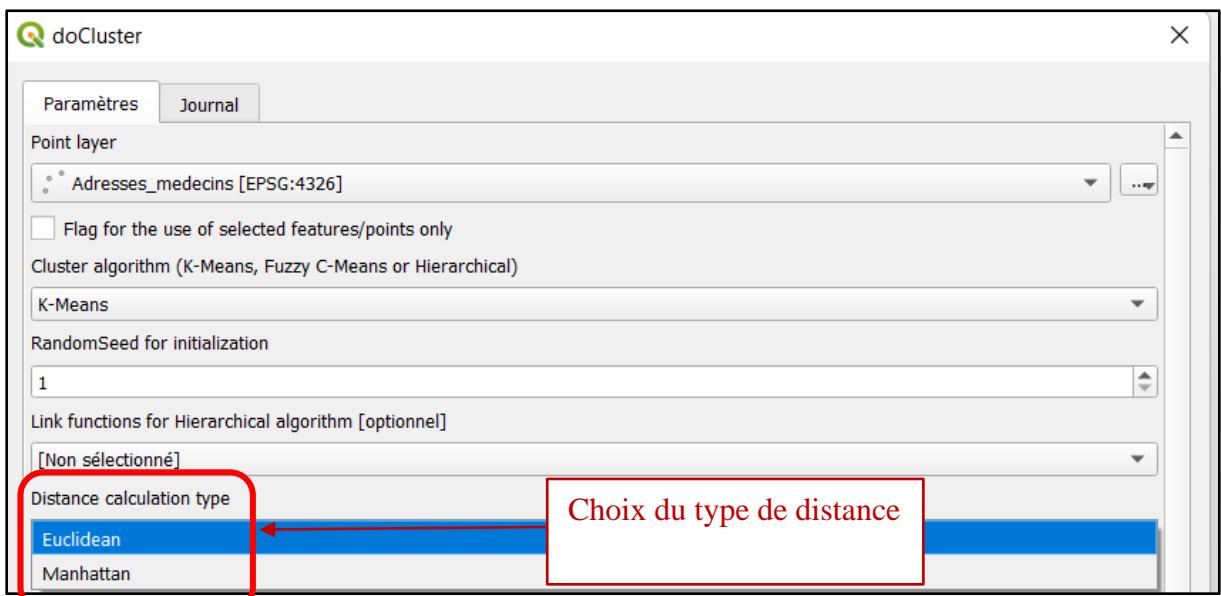
### 3) Type de calcul de la distance :

On a deux types de calculs de distance : distance euclidienne et distance de Manhattan.

**3.1 La distance euclidienne :** En mathématique la distance euclidienne entre deux points dans l'espace euclidien est la longueur d'un segment de ligne entre les deux points. Elle peut être calculée à partir des coordonnées cartésiennes, des points utilisant le théorème de Pythagore, donc parfois appelés distance de Pythagore. Cette distance est le plus couramment utilisée quand les variables sont continues. Elle désigne la distance ordinaire entre deux points. L'un des algorithmes qui utilise cette distance est le K-means.

**3.2 La distance de Manhattan :** appelée aussi distance-taxis est la distance entre deux points parcourue par un taxi lorsqu'il se déplace dans une ville où les rues sont agencées selon un réseau ou quadrillage. Cette distance fut définie par Hermann Minkowski, le trajet fait par un taxi lorsqu'il se déplace d'un nœud du réseau à un autre en utilisant les déplacements horizontaux et verticaux du réseau.

**La différence entre la distance Euclidienne et celle de Manhattan est que la première désigne le chemin le plus court entre deux points tandis celle que la seconde est la somme de toutes les autres distances réelles entre les deux points.**



**Figure 10 : Etape de choix de la distance**

#### 4°) Définition du nombre de cluster :

Ici on définit le nombre de cluster qu'on veut faire suivant notre projet ou la demande. Le nombre de clusters varie de 2 à 999.

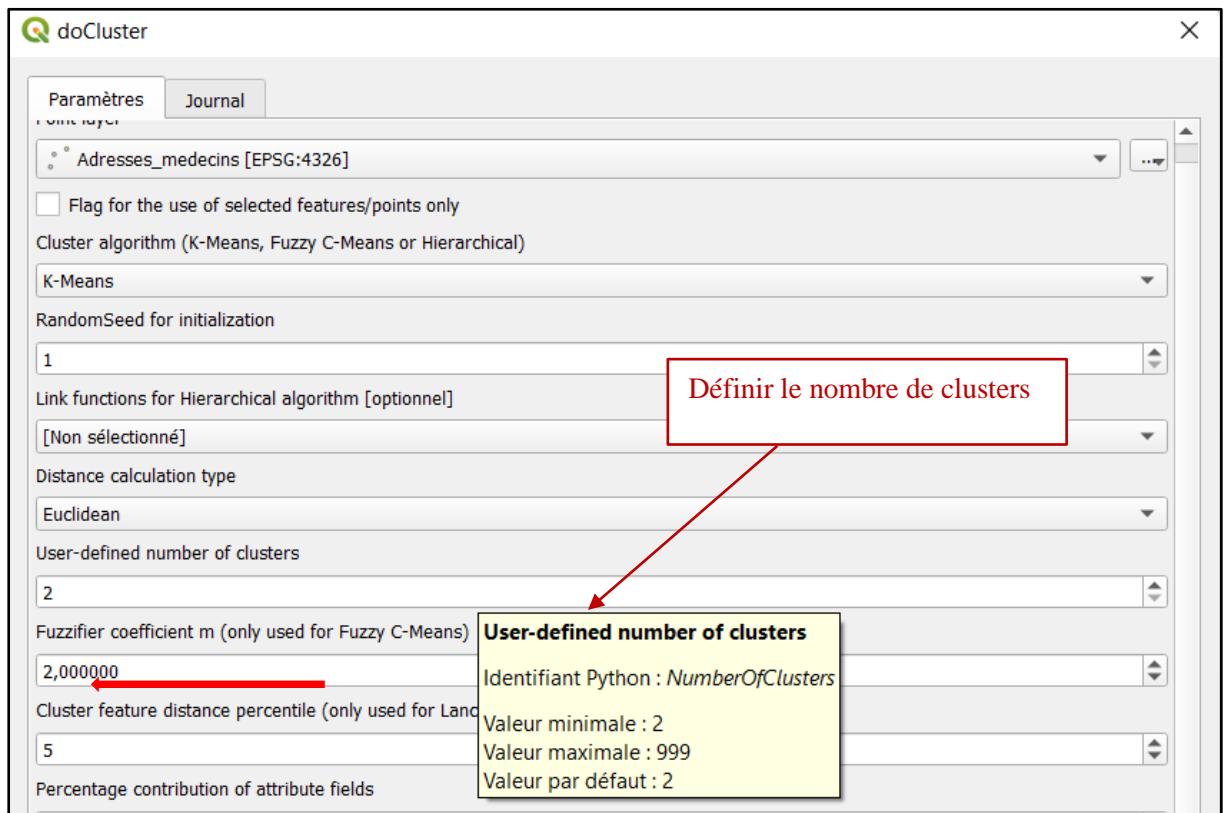
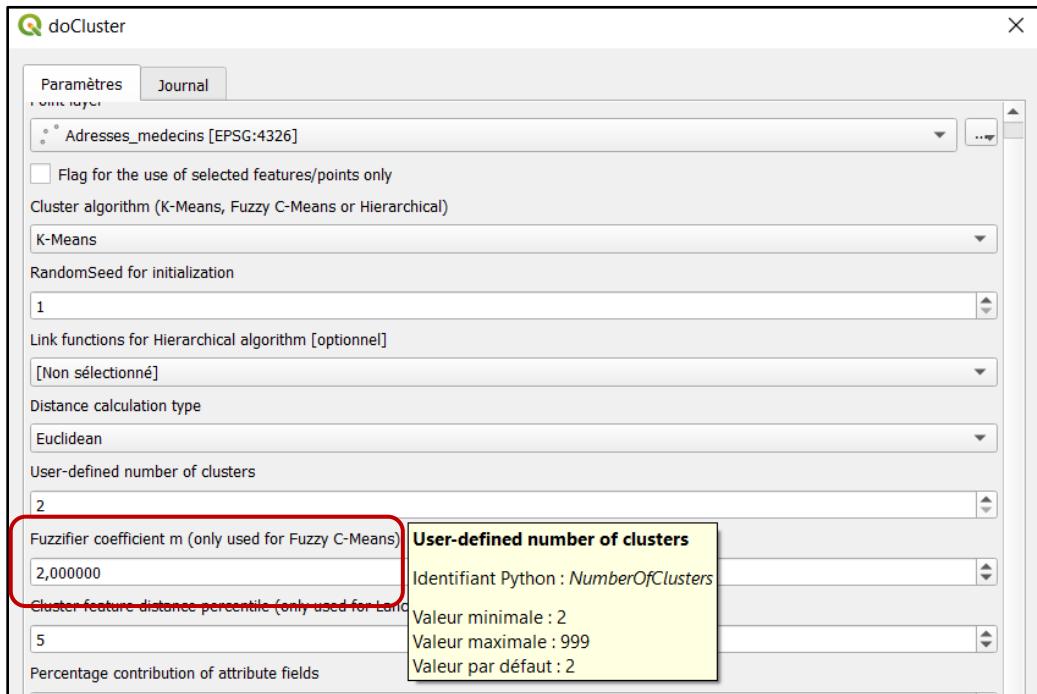


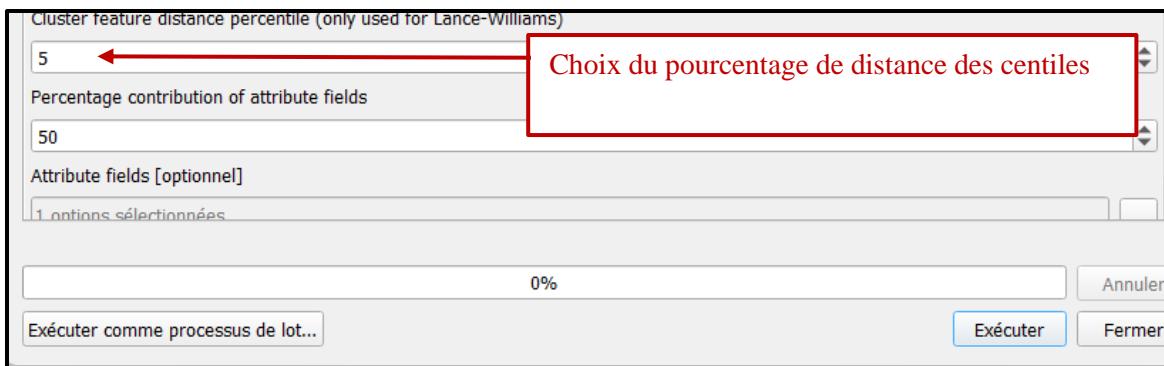
Figure 11 : Etape de choix du nombre de cluster.

**5°) Coefficient de Fuzzy :** Cette nouvelle fonctionnalité est utilisée seulement lorsqu'on choisi l'algorithme Fuzzy C-Means. La contribution majeure de cette fonctionnalité est la possibilité d'utiliser plusieurs coefficients de fuzzification (minimum 1.1) par opposition à un seul coefficient des autres algorithmes.



**Figure 12:** Étape de choix du coefficient de Fuzzy dans le cas où on utilise l'algorithme Fuzzy C-Means

**6°) Centiles de distance des caractéristiques du cluster :** Les centiles d'une variable désignent les valeurs seuils de cette variable, qui lorsqu'on ordonne la population selon les valeurs de cette variable, la partitionnent en 100 sous classes de tailles égales. On a donc de 0 à 99 centiles. La valeur par défaut est de 5.

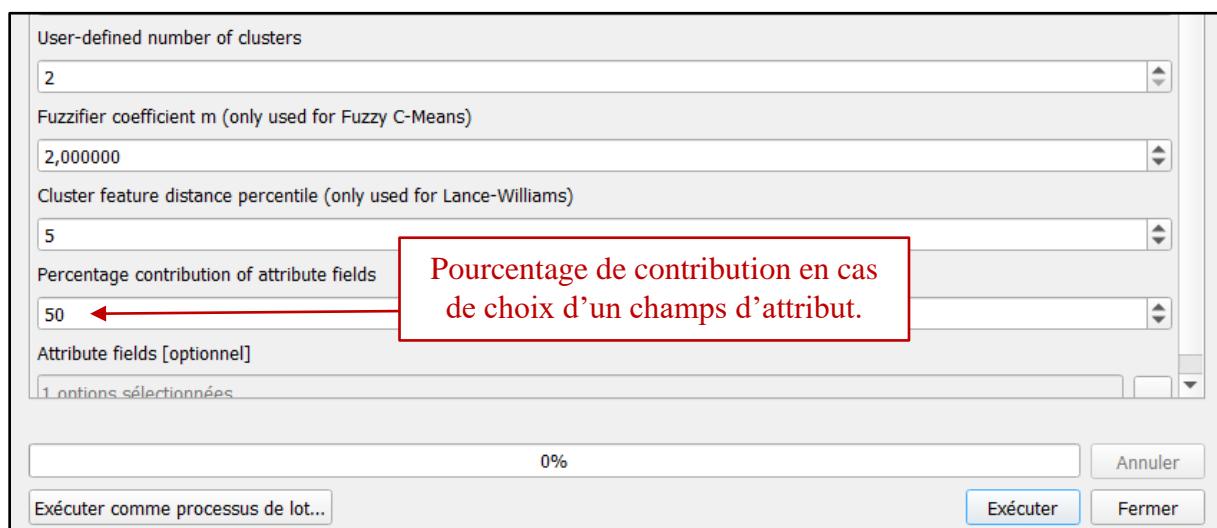


**Figure 13:** Étape de choix des centiles.

**7°) Pourcentage de contribution et champs d'attribut :** Selon l'auteur de cette extension : « si des informations autres que les coordonnées géographiques doivent être prises en compte, l'utilisateur peut sélectionner les champs numériques disponibles (champ d'attribut) de la couche d'entrée à incorporer

dans le processus de regroupement. L'utilisateur peut alors **spécifier le pourcentage de contribution** de ces champs dans leurs ensembles en complément ou en alternative aux coordonnées géographiques ».

On peut avoir trois situations différentes : si le pourcentage de contribution défini est 0, le regroupement ne prend en considération que les coordonnées géographiques (purement spatial) ; si il est défini à 100 donc le regroupement prend compte que le ou les champs sélectionnés ; enfin si le pourcentage est compris entre 0 et 100, il tient compte les coordonnées géographiques et les champs sélectionnés. Autrement dit, la valeur du pourcentage détermine si le regroupement est basé soit sur l'emplacement (cas du pourcentage faible : 0 à 49) soit sur les champs choisis (fort pourcentage : 51 à 100) ou bien la combinaison des deux (pourcentage = 50).



**Figure 14 : Etape de choix du pourcentage de contribution des champs d'attributs.**

### III°) C'EST QUOI QUANTUM GIS (QGIS) ?

Le logiciel QGIS est une application de système d'informations Géographiques (SIG) libre et open source qui permet de visualiser, de traiter, modifier, et analyser des données géospatiales.

#### 1/ Les points forts du logiciel sont :

La gratuité : téléchargement libre,

Complet : large possibilité d'ajout fonctionnel à travers de plugins gratuitement accessible,

Temps de réponse et performances : une architecture technique bien conçue, facilitant la réalisation de tâches,

Documentations : une riche documentation avec des explications et exemple détaillés,

Une grande communauté : composée d'experts passionnés disposant d'un bloc pour les échanges et le partage de connaissance,

Simplicité et utilisation : accessibilité à tous,

Compatibilité technique : logiciel compatible avec plusieurs environnement Windows, Linux, Mac iOS et connectable aux bases de données PostgreSQL, Post Gis, MySQL etc.

Évolutivité : la fréquence de nouvelles versions et évolution fonctionnelle pertinentes

## 2/ Domaines et thématiques d'application

Le logiciel est applicable dans les domaines suivant : environnement, gestion des ressources naturelles, urbanisme, aménagement du territoire, transport, télécommunication, santé, géomarketing, réseaux ( eau, assainissement, énergie), immobilier, géopolitique, sécurité, agriculture etc.

## 3/ Versionnement et téléchargement

La dernière version actuelle est le **3.24.0 'Tisler'**, elle est sortie le 18/02/2022. La version la plus stable est le **3.22.0 'Bialowieza'**

Pour le téléchargement, le logiciel est disponible sur le site :

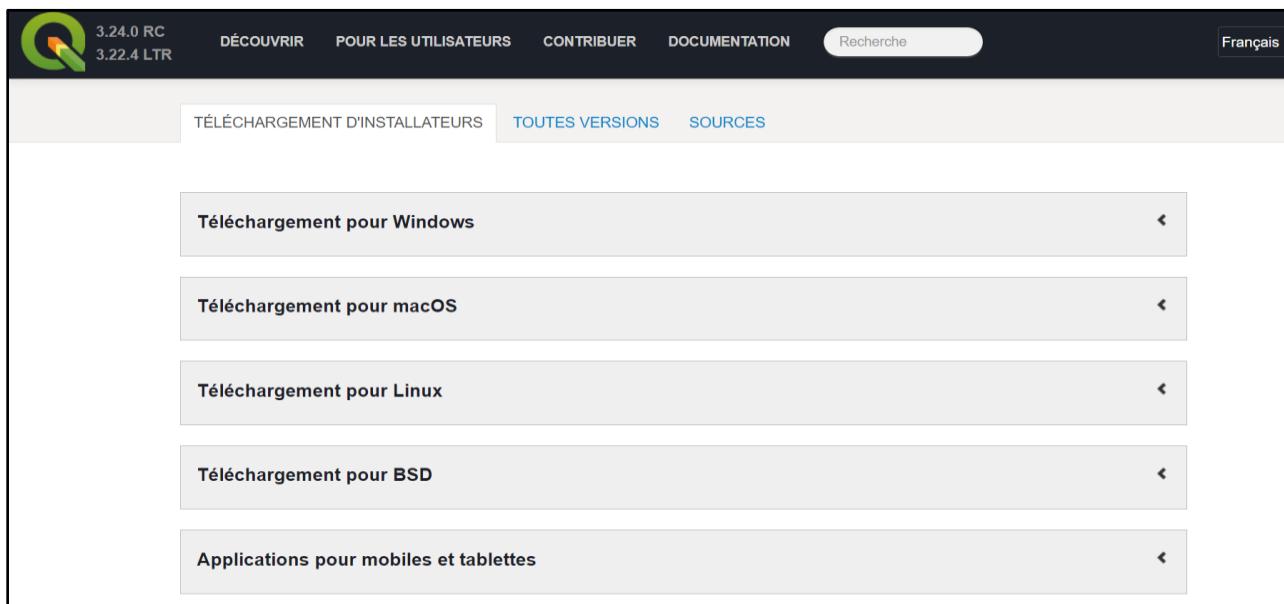
<https://www.qgis.org/fr/site/forusers/download.html>

Une fois dans ce site, il y'a trois menus sur l'interface : **Téléchargement d'installateurs**, **Toutes versions, et Sources** :

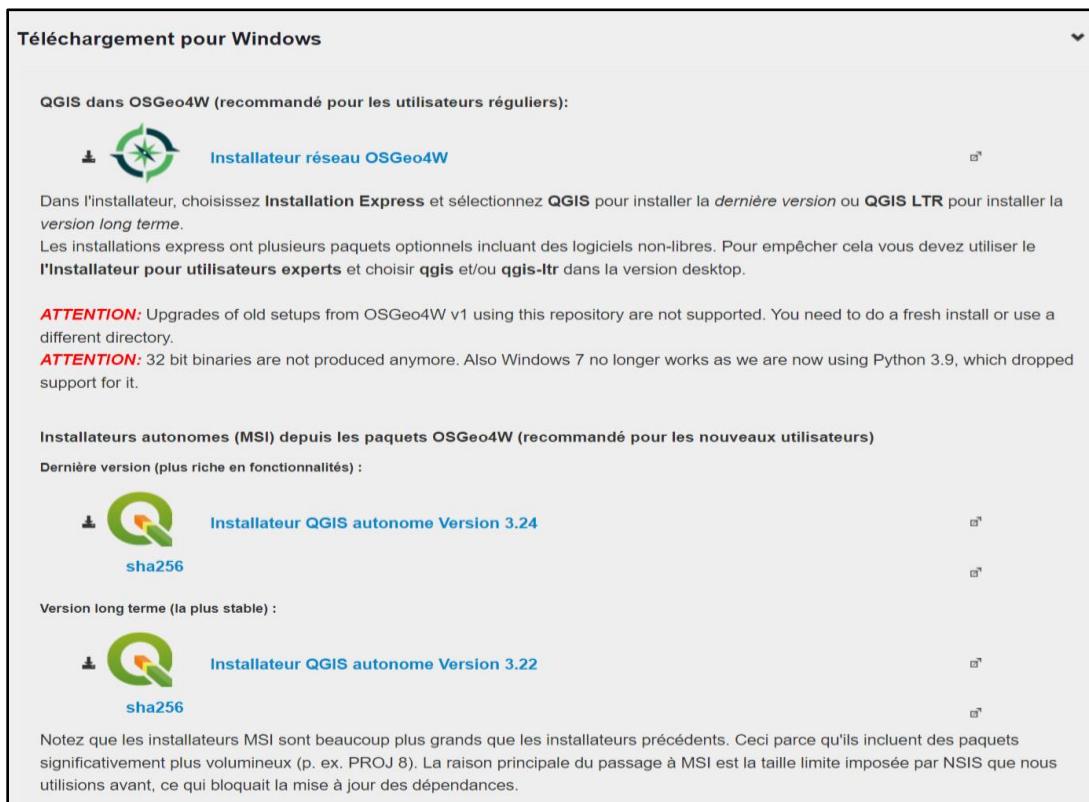


Figure 15 : fenêtre de téléchargement du logiciel QGIS

Ici on choisit d'abord notre système d'exploitation de notre machine ( Windows, Linux, MacOS, BSD, Applications pour mobiles et tablettes). Ensuite on peut choisir entre la dernière version et la version la plus stable.



**Figure 16 : Choix du système d'exploitation**



**Figure 17 : Choix de la version de QGIS**

**Toutes versions :** dans ce menu, on a la possibilité de consulter ou télécharger les versions antérieures de QGIS et des extensions.



**Figure 18 : fenêtre pour télécharger ou consulter d'autres versions antérieures ou extensions de QGIS**

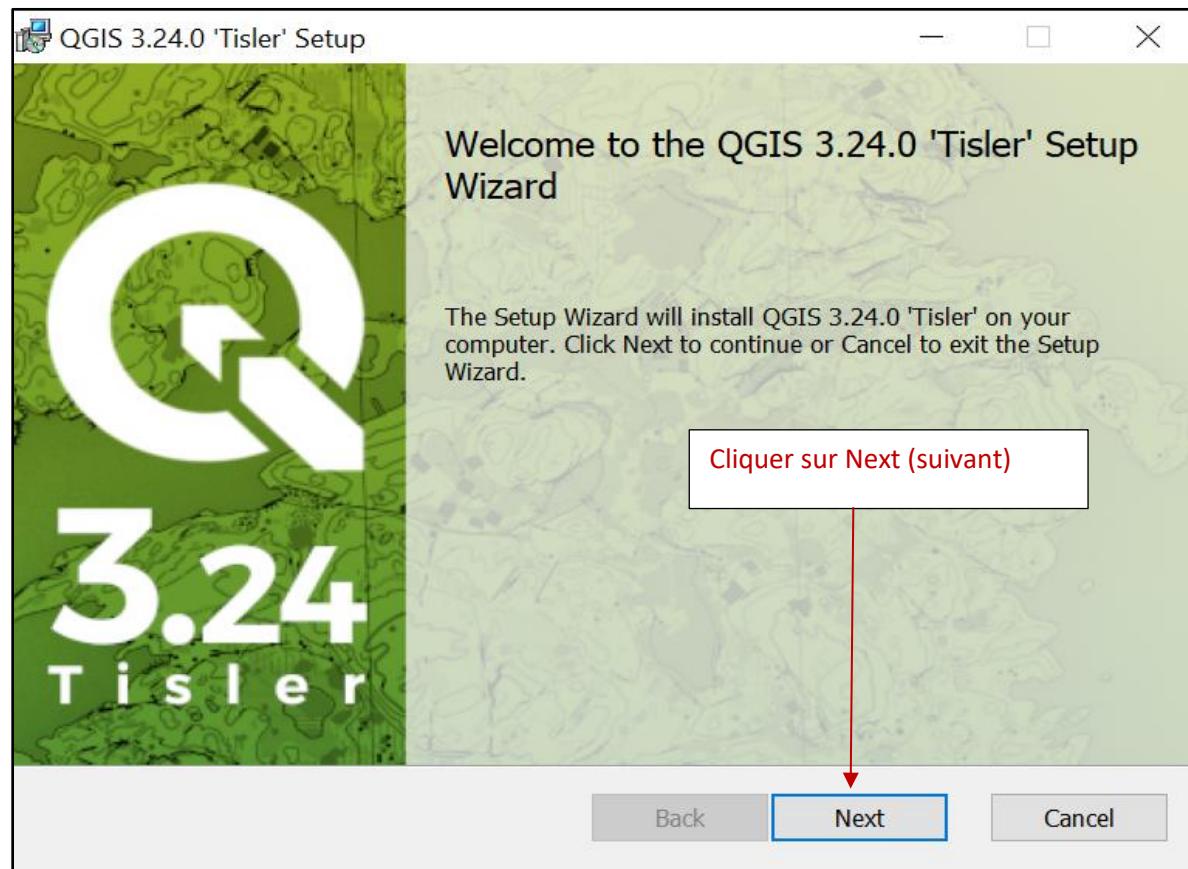
**Sources :** ici on la licence, le code source, le guide d'installation ; des extensions sont aussi disponibles ici.



**Figure 19 : Licence, accès au guide d'installation et à d'autres extensions**

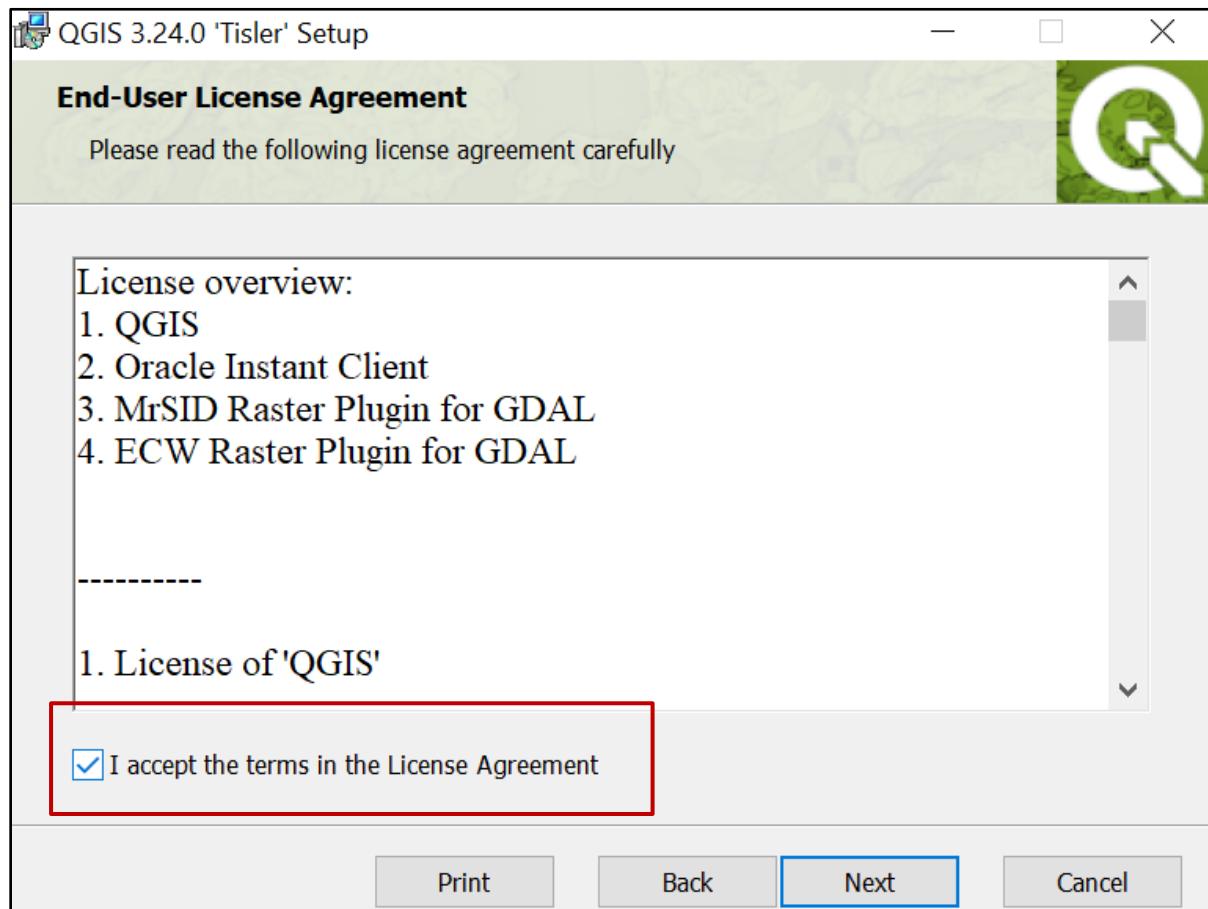
#### 4/ Installation

Après téléchargement du logiciel (le setup), on fait double clic pour lancer l'installation.



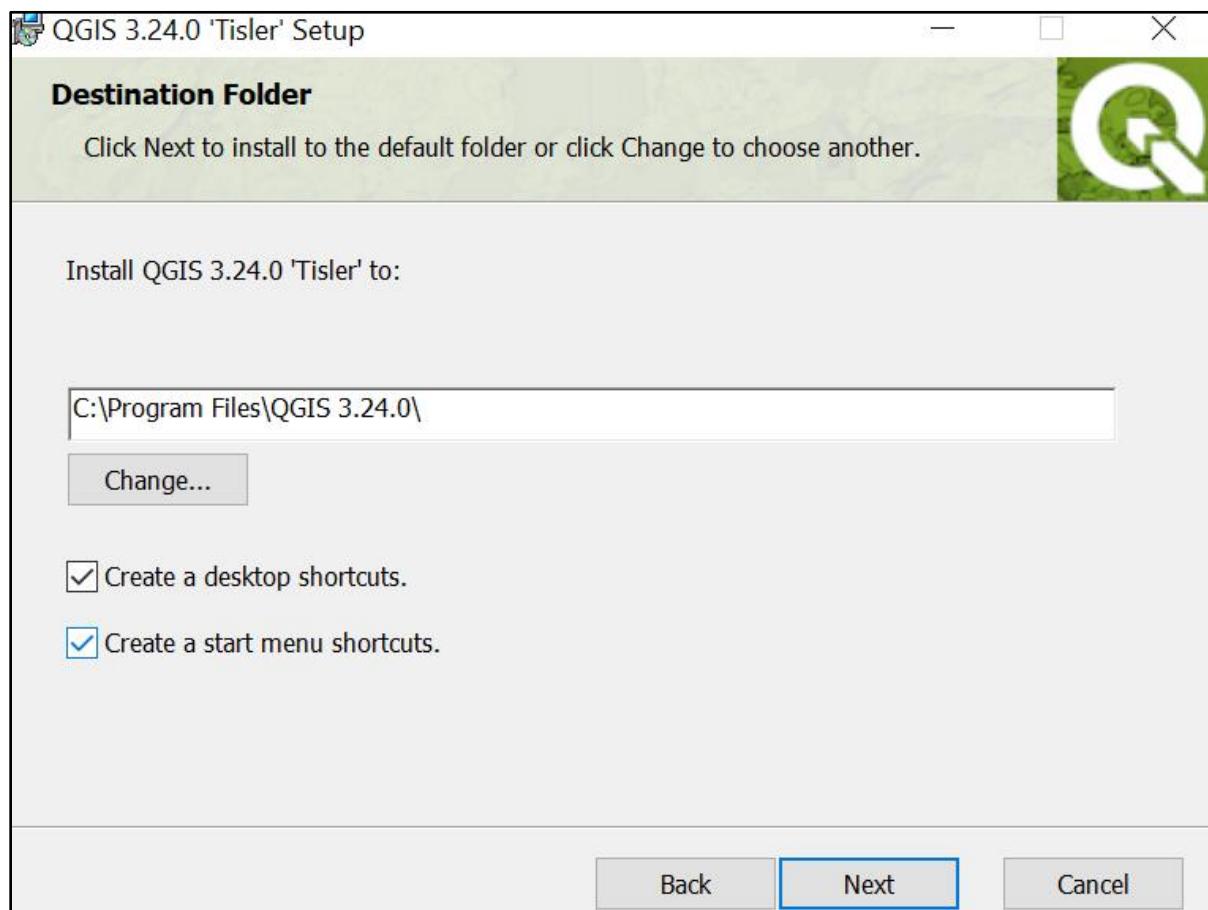
**Figure 20 : fenêtre d'installation**

Cliquer sur : Next (suivant)



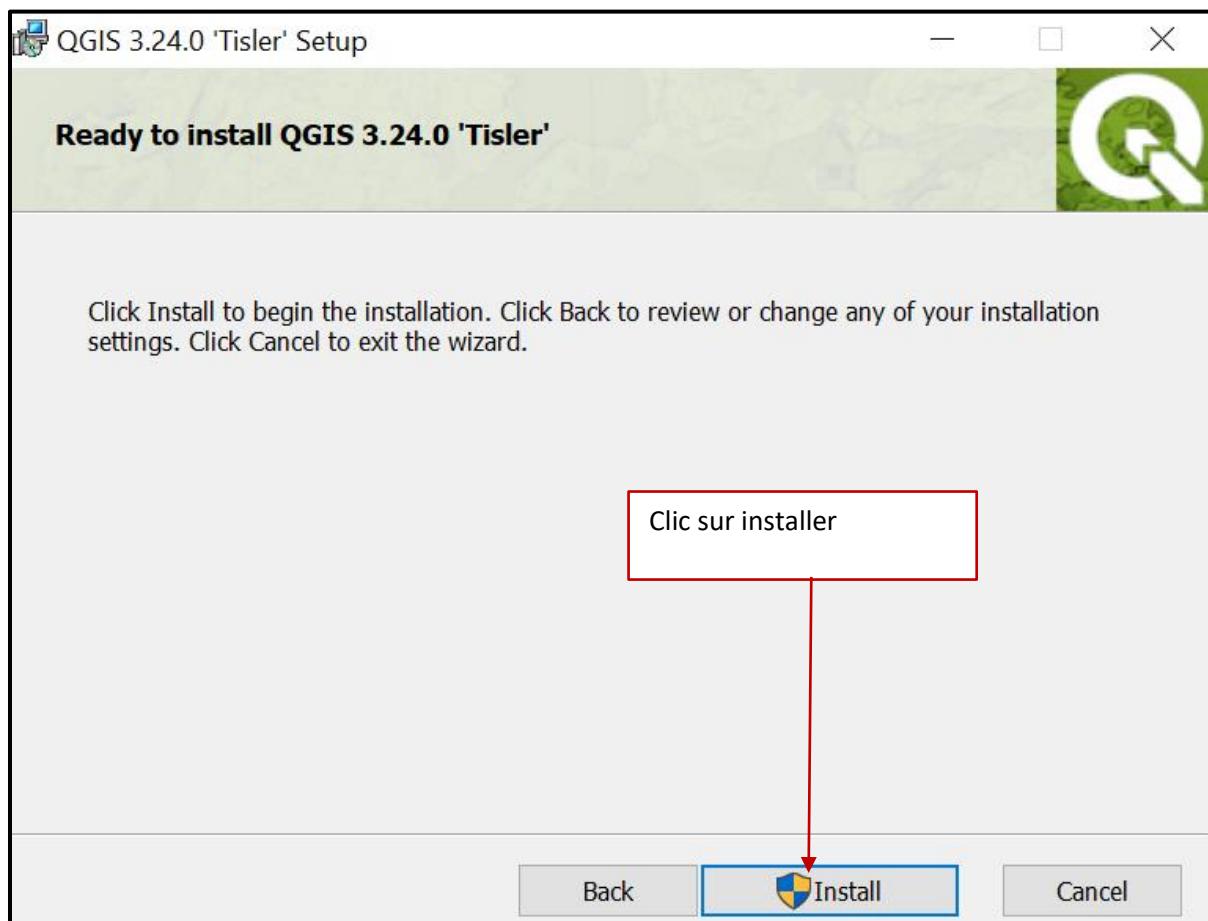
**Figure 21 : Acceptation des termes de licence**

On accepte les termes de la licence et on clic sur Next.



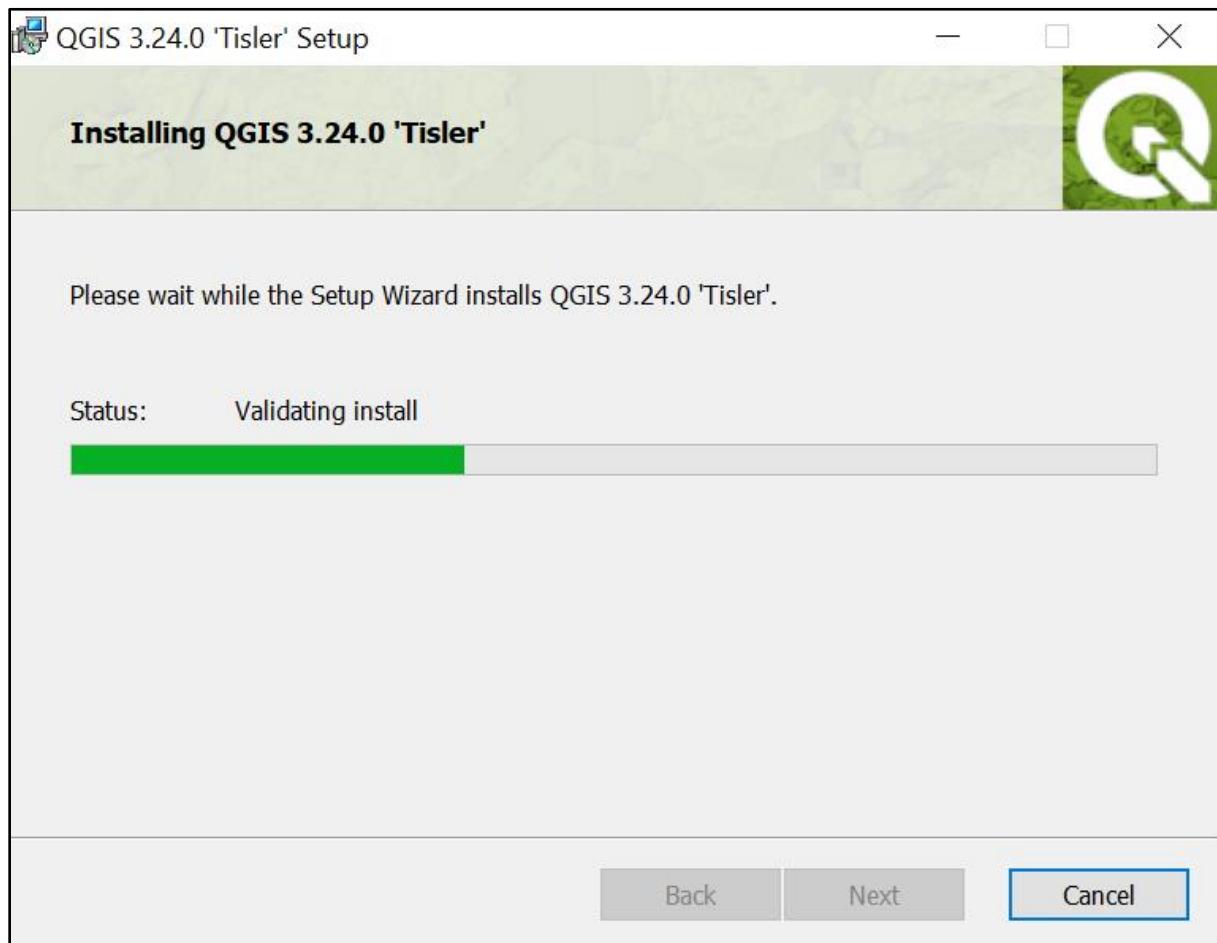
**Figure 22 : Chemin d'enregistrement du programme**

Là, il nous montre le chemin d'enregistrement du programme (par défaut il l'enregistre sur le disque C/ Programme files)



**Figure 23 : Début de l'installation**

Clic sur Install (installer) pour débuter l'installation.



**Figure 24 : Installation en cours**

Installation en cours.



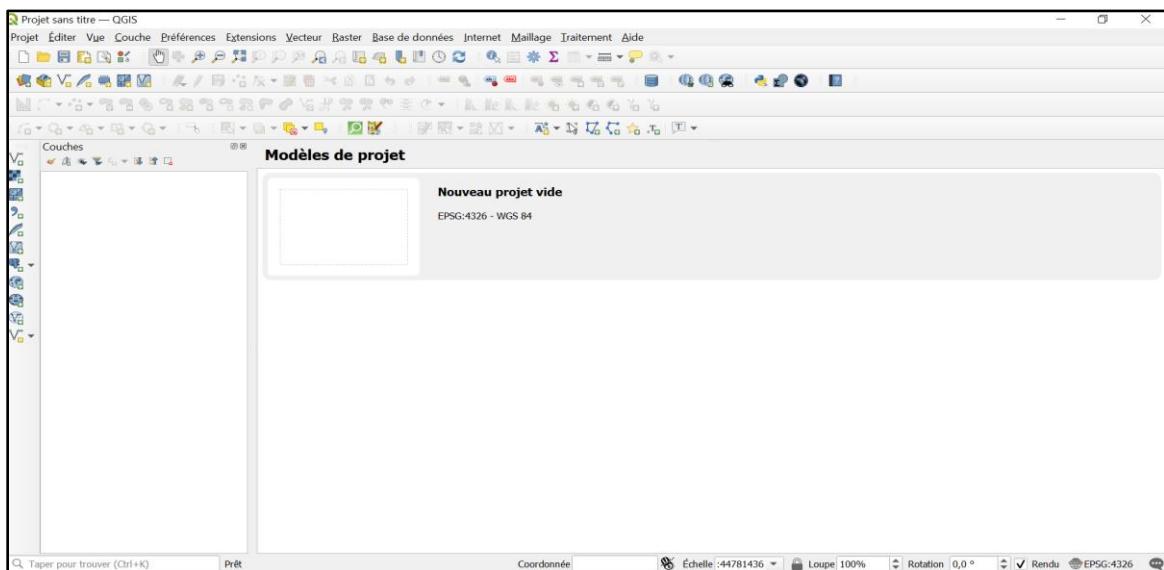
**Figure 25 : Finalisation de l'installation**

Clic sur Finish (finir) pour terminer l'installation.



**Figure 26 : Accès au dossier QGIS et ouverture**

Dans le menu Windows, on déroule le dossier QGIS 3.24.0 pour avoir accès aux différentes composantes du logiciels. Ainsi on lance QGIS Desktop 3.24.0 dont l'interface se présente comme suit :



**Figure 27 : interface d'accueil**

## IV/ APPLICATION DE L'EXTENSION

### 1/ DOMAINES D'APPLICATION

Dans ce manuel, nous avons choisi d'appliquer le clustering dans deux domaines :

- 1.1 La santé :** Répartition des médecins généralistes dans le département du Loiret. L'objectif est d'abord de voir comment est répartie les médecins généralistes à l'échelle du département à travers l'utilisation du clustering. Ensuite d'analyser la relation entre cette répartition avec le réseau routier et la population par Communes. Ce qui permettra de mettre en exergue les zones de fortes concentrations des médecins mais aussi les déserts médicaux en termes de personnels de santé.
- 1.2 Occupation du sol du département du Loiret :** Dans cette partie l'objectif visé est de classer à partir du clustering les types ou catégories d'occupation du sol dans le département du Loiret. Dans ce sillage, on partira d'une image satellitaire pour effectuer une carte d'occupation du sol à partir de l'algorithme K-Means Classification.

### 2/ DONNEES ET METHODES :

#### 2.1 Données de santé :

##### • Phase de préparation

Pour les données de santé, nous avons téléchargé les adresses des médecins généralistes du département dans le site [www.conseil-national.medecin.fr/annuaire/resultats](http://www.conseil-national.medecin.fr/annuaire/resultats) ;

Création d'un projet sur Google Earth avec comme sortie le format Kml

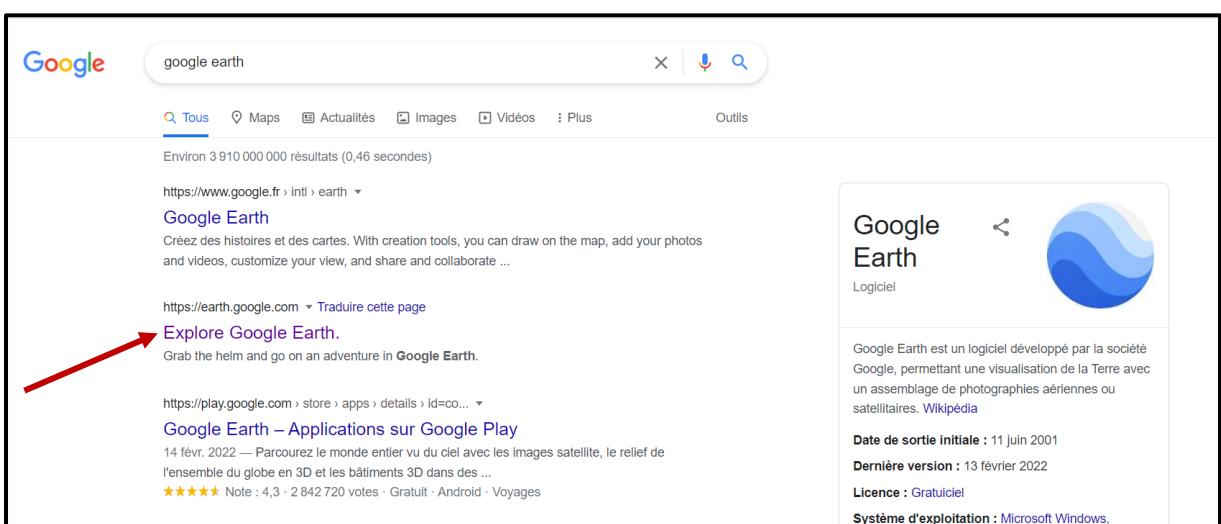


Figure 28 : recherche pour ouvrir Google Earth

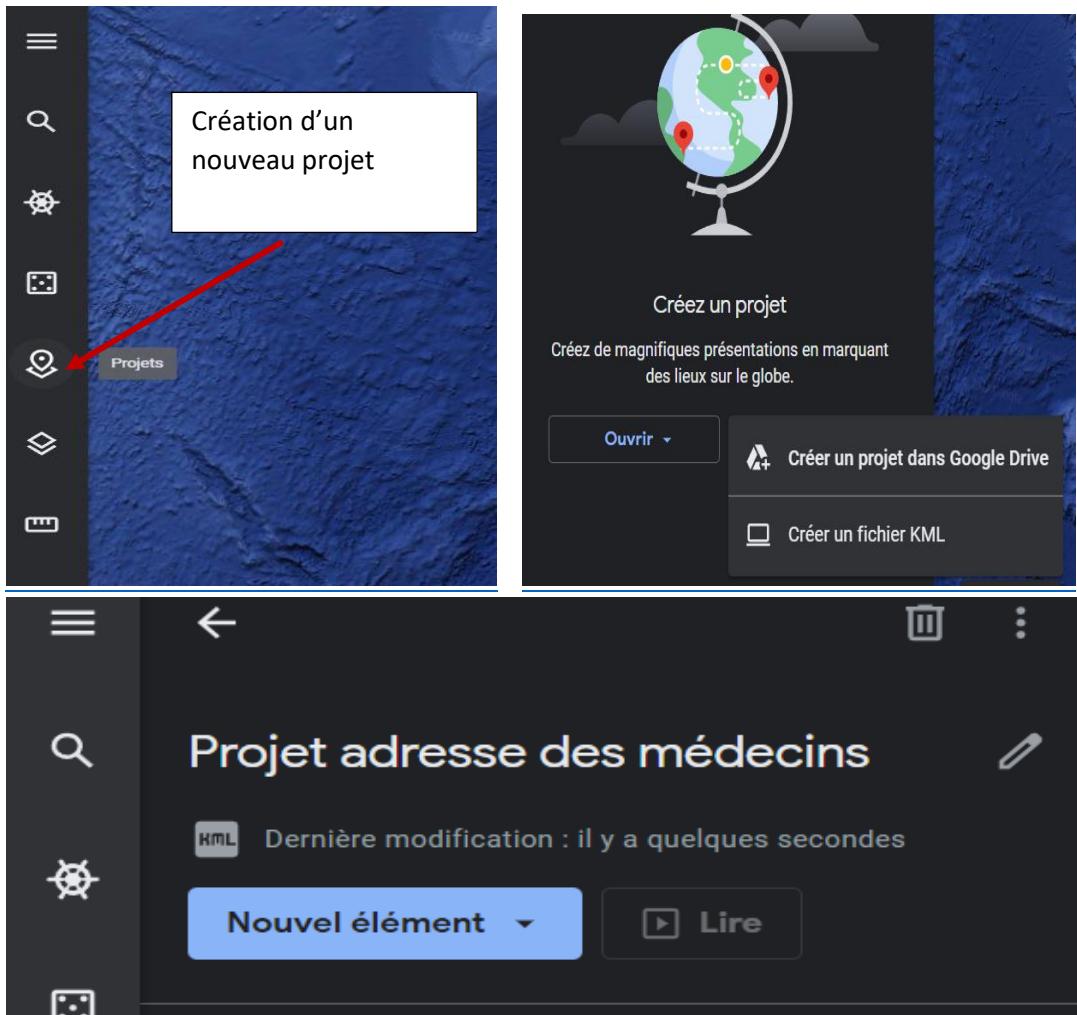


Figure 29 : fenêtre pour créer un projet sur Google Earth et l'enregistrer en format KML

Exportation et localisation des adresses sur Google Earth ;  
Enregistrement de l'ensemble des points (adresses) dans le projet

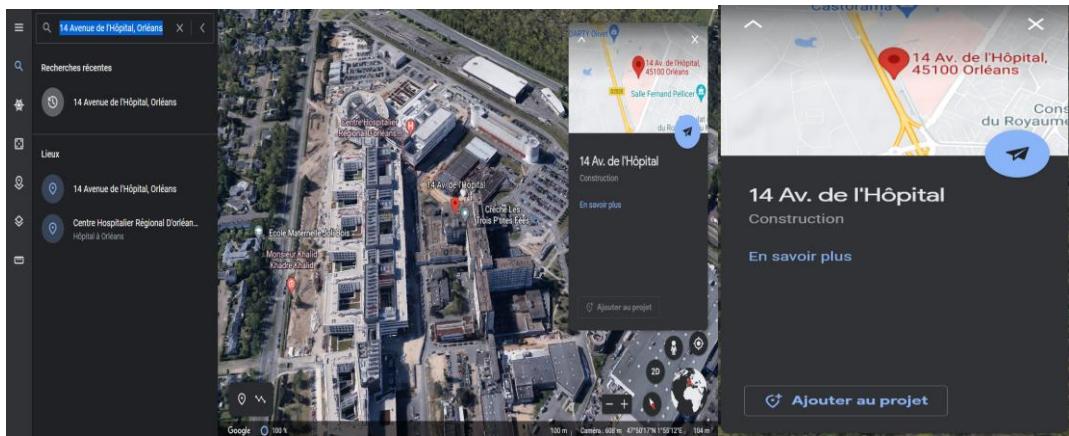
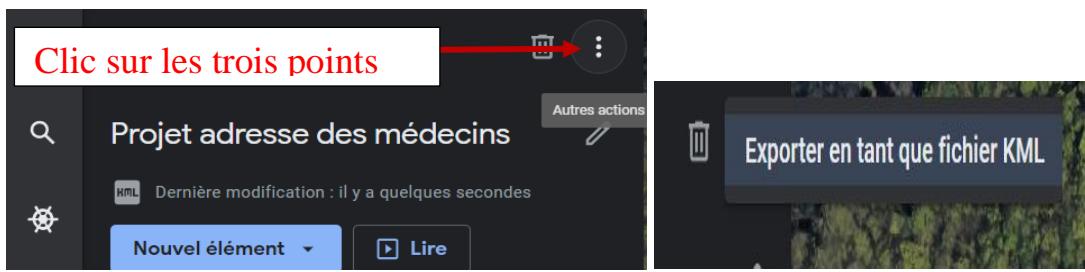


Figure 30 : Exemple de localisation d'une adresse et l'ajout dans le projet

Importation et enregistrement du projet (adresses des médecins) toujours au format kml dans un fichier ;

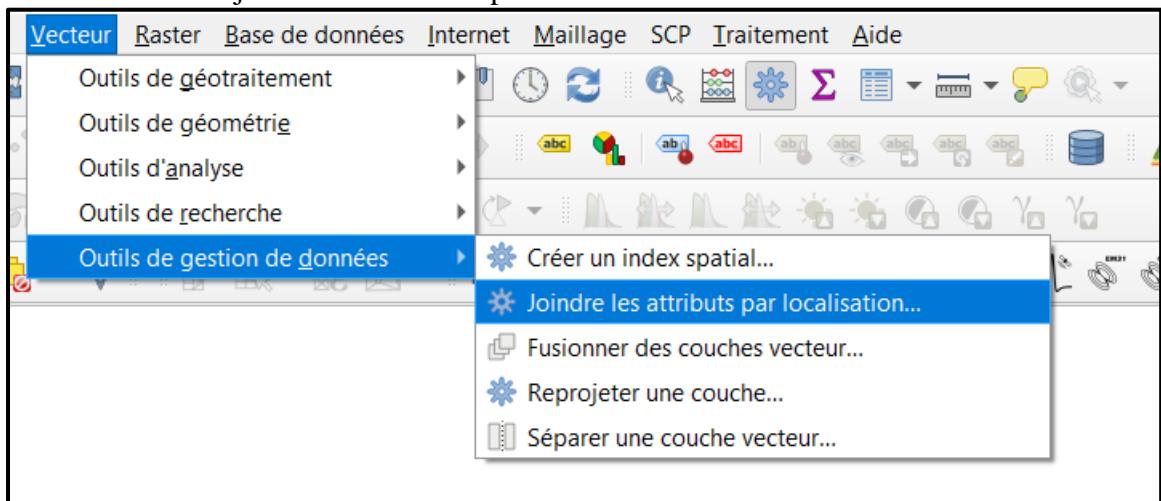


**Figure 31 : l'exportation du projet en format KML**

Lancement du logiciel QGIS, création d'un projet et paramétrage (système de projection : EPSG 2154-RGF93 v1/ Lambert-93 ;

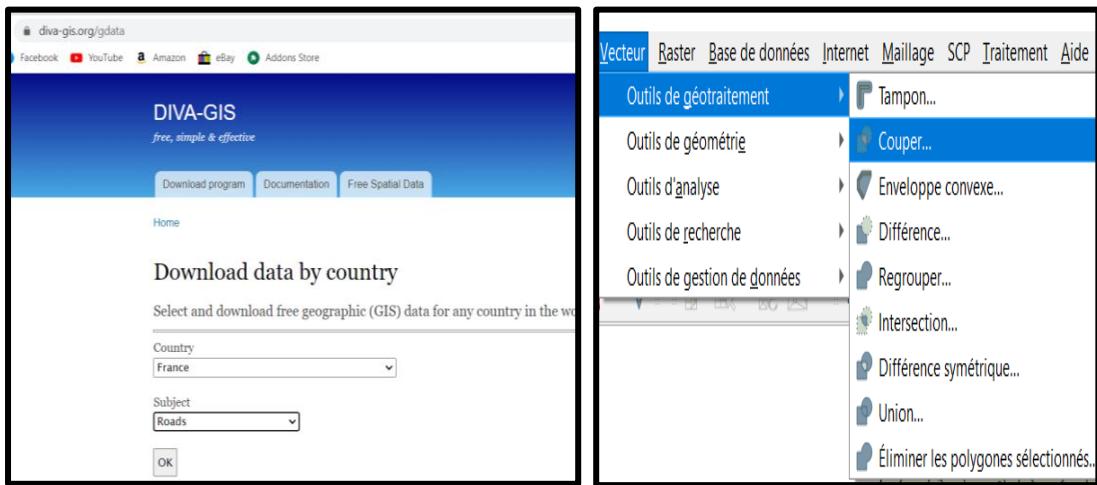
Importation du fichier (adresses des médecins) sur QGIS

Pour la population par communes: Nous avons utiliser une couche des communes du Loiret qui nous à été remise par Mr Bensaïd (dans le cadre des TD de SIG). Cette dernière comporte un champs population. Ainsi nous avons joins les attributs de ce champs avec la couche commune téléchargée sur diva-gis. L'outil utilisé est : menu vecteur → outil de gestion de données → joindre les attributs par localisation.



**Figure 32 : chemin de jointure des attributs**

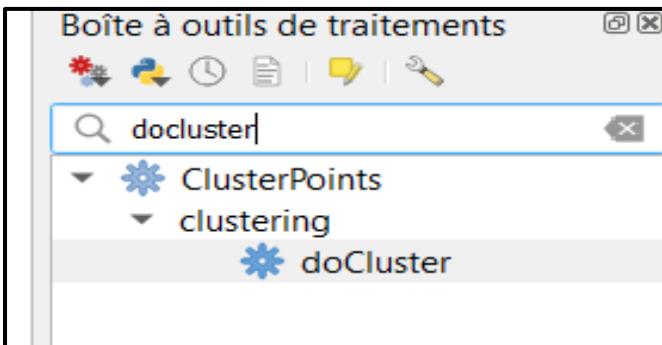
Les données : limites administratives (département et communes) et du réseaux routier : Nous les avons téléchargé sur le site : diva-gis.org/gdata ; à l'échelle de la France métropolitaine ainsi nous les avons découpés à l'échelle du département du Loiret en utilisant l'outil : Vecteur → outils de géotraitements → couper.



**Figure 33 : téléchargement de données sur Diva-GIS et découpage des routes avec la zone d'étude**

- **Phase de traitement**

Ouverture de l'extension : clustering point : menu traitement → Boite à outils → clusterPoints → clustering → doCluster (double clic pour ouvrir l'extension).



**Figure 34 : chemin pour ouvrir l'extension Cluster points**

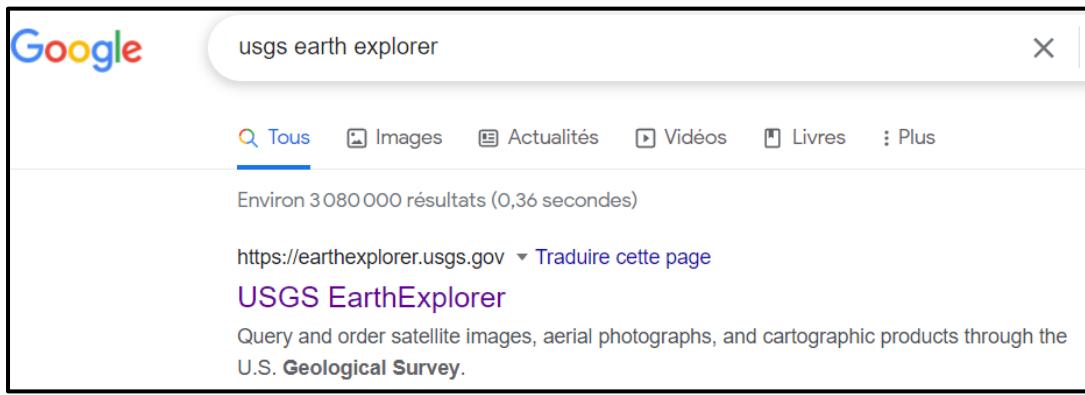
Choix et sélection des paramètres (algorithmes, distance, fonction de lien, nombre de cluster, ...) puis Ok

NB : Pour cette précédente, pour de mieux avoir la meilleure classification, c'est-à-dire la mieux adaptée, celle qui traduit mieux la réalité, nous avons effectuer plusieurs tests de classifications, en faisant diverses combinaisons des algorithmes, des distances et fonctions de lien mais aussi en variant le nombre de clusters afin de voir les meilleures combinaisons de paramètres et d'avoir un résultat de qualité.

## 2.2 Données occupation du sols :

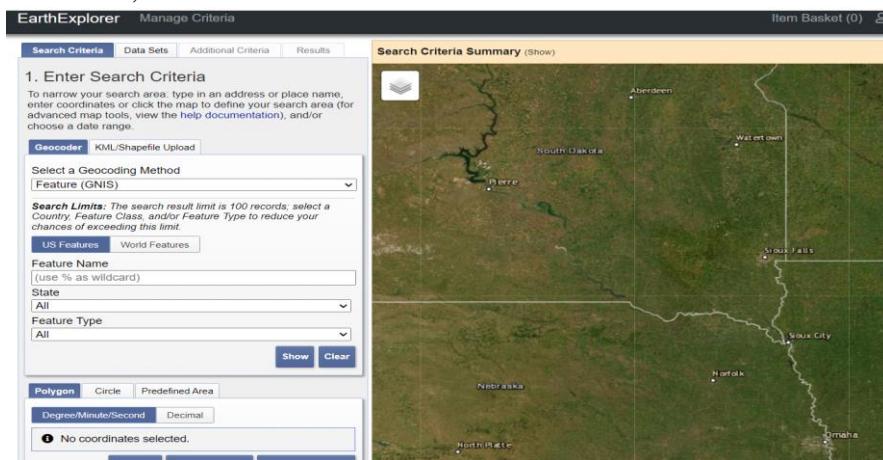
- **Préparation des données :**

Nous avons crée un compte sur le site : USGS ( earthexplorer.usgs.gov) ;



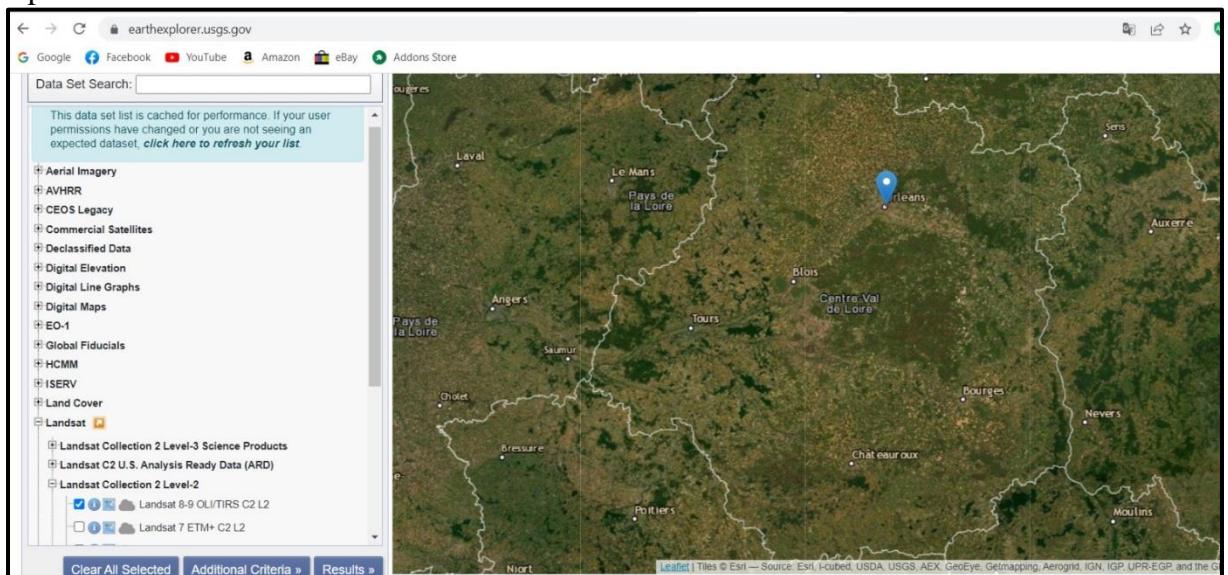
**Figure 35 : accès au site de USGS**

Connection sur le site pour procéder au téléchargement des images pour recouvrir toute la zone d'étude ;



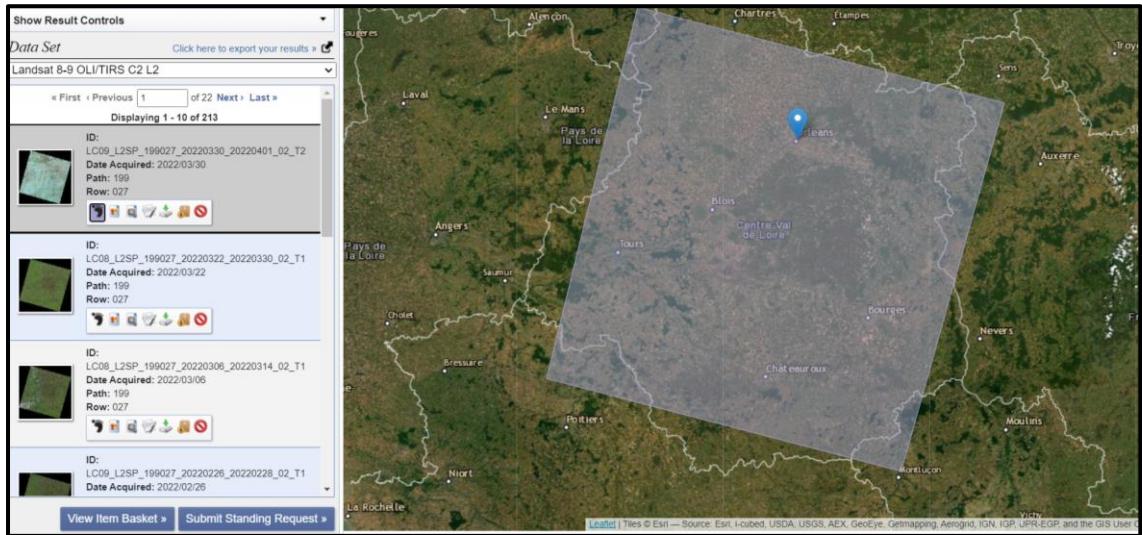
**Figure 36 : Fenêtre de choix des paramètres pour le téléchargement des images**

Zoom sur notre zone d'étude (département du Loiret) en précisant l'adresse et choix de la date → sur l'onglet : Cloud Cover, on mis couverture nuageuse à 10 % ; ensuite → sur l'onglet : data sets → Landsat → Landsat collection 2 level-2 → Landsat 8-9 OLI/TIRS C2 L2 → puis clic sur résultats.



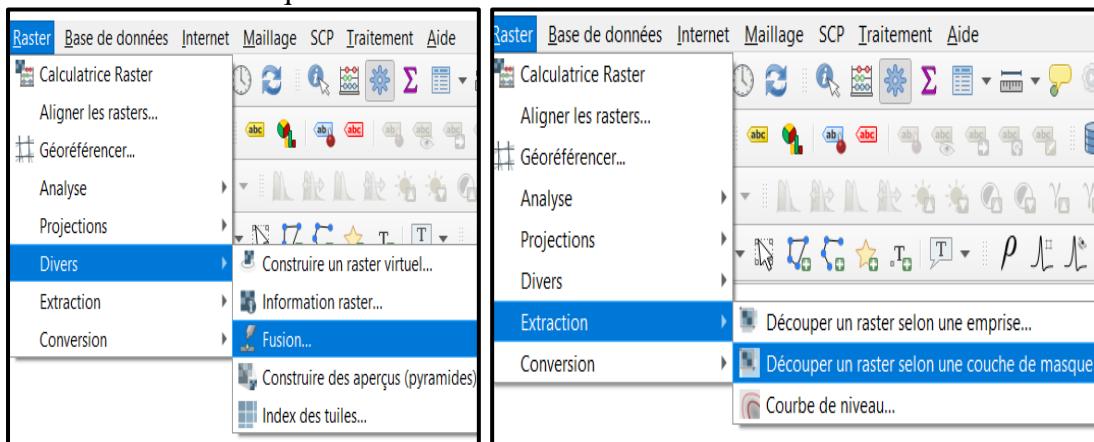
**Figure 37 : Zoom sur la zone d'étude et le choix du type d'images**

Visualisation et téléchargement des images.

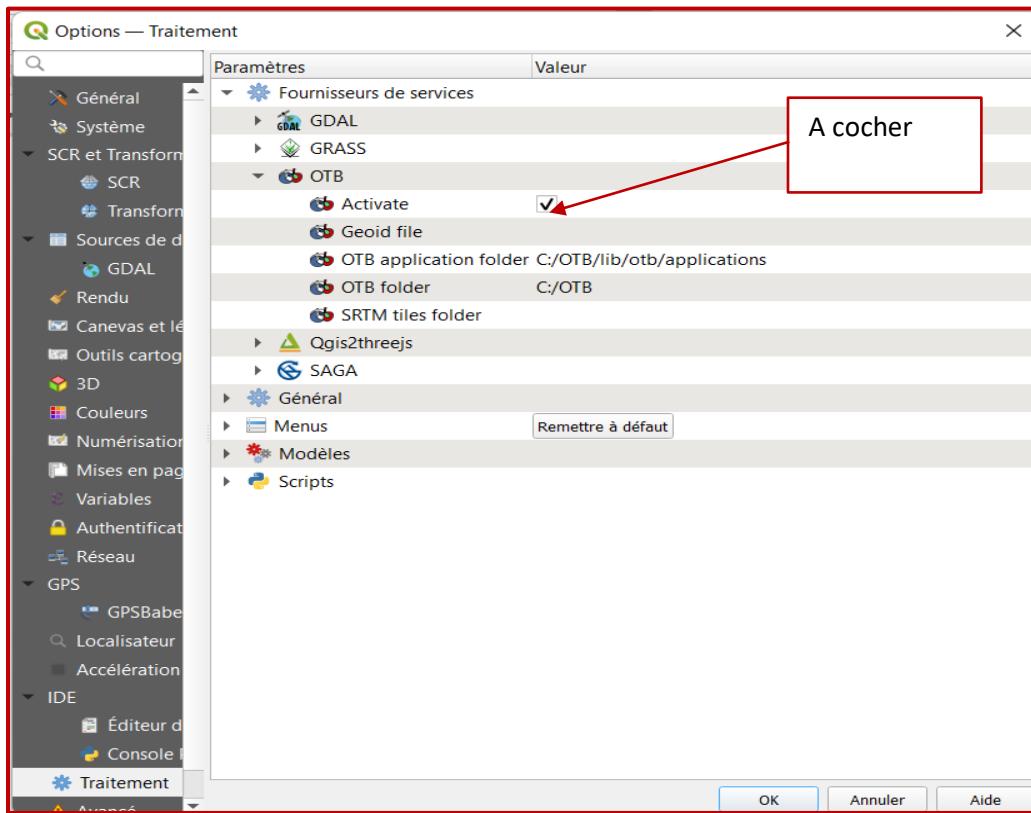
**Figure 38 : Visualisation des images et téléchargement des images**

- **Etape de traitement**

On a enfin ouvert les images sur QGIS puis procéder à leur fusionnements et l'extraction de la zone concernée. Les outils utilisés sont pour le fusionnement : menu raster → divers → fusion ; et pour l'extraction : menu raster → extraction → découper un raster selon une couche de masque.

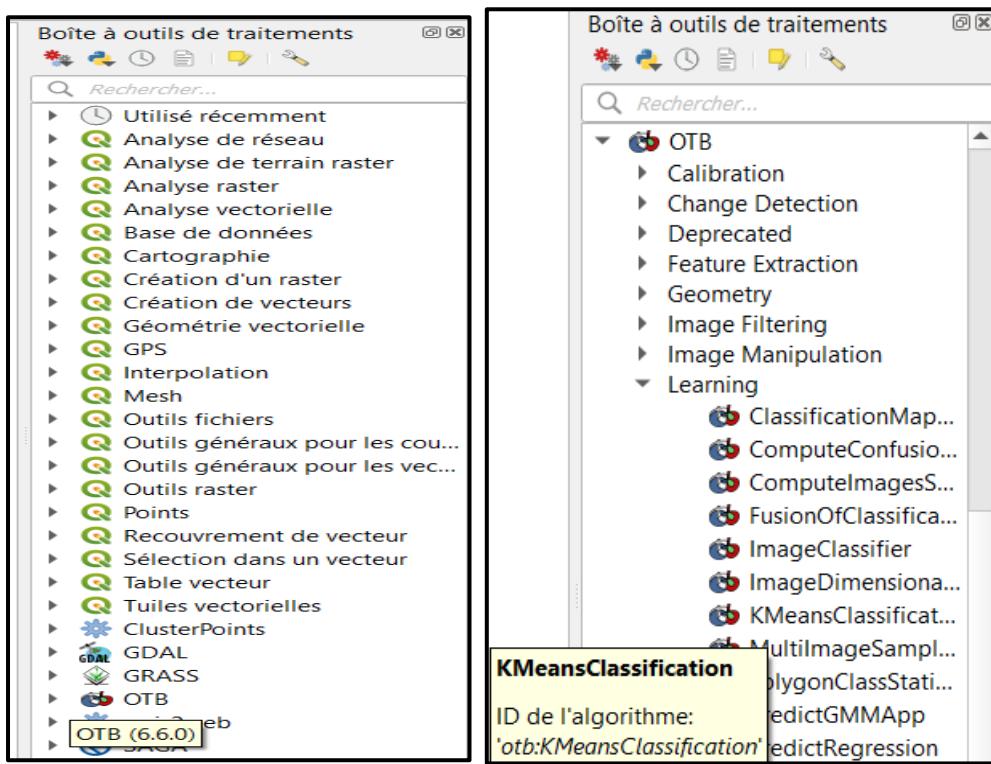
**Figure 39 : chemin du fusionnement et de celui du découpage des images**

Ainsi pour faire la classification de K-means, on a activé l'extension OTB (Orfeo Tool Box), qui prend en charge le format raster pour le clustering via K-means. Concernant l'activation de OTB, il faut passer par : menu préférences → options → traitement → fournisseurs de service → dérouler le menu OTB → puis cocher l'onglet : activé et → cliquer sur ok.



**Figure 40 : fenêtre d'activation de OTB**

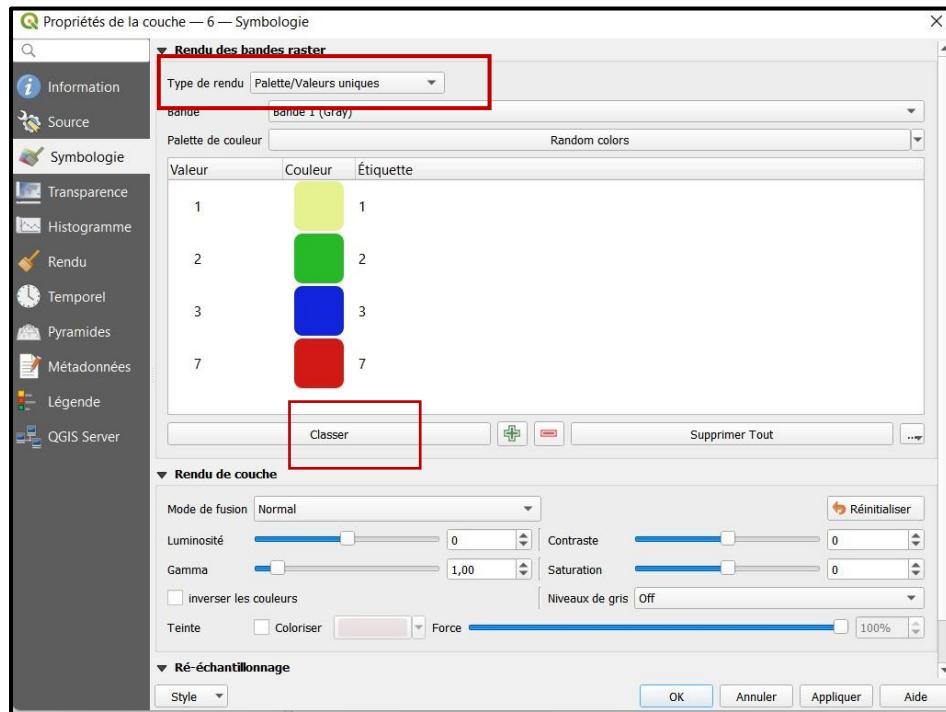
Après cela l'outil OTB va directement apparaître sur la boîte à outils de traitements. Ensuite on déroule le menu OTB → Learning → double clic sur l'outil K-means classification.



**Figure 41: chemin d'accès à l'outil OTB et de l'outil de classification K-means**

Après ouverture de l'outil → on sélectionne l'image → choix du nombre de cluster → ...

→ modifier la symbologie de la couche : palette à valeurs uniques → classer → puis ok.



**Figure 42 : modification de symbologie de la couche classifiée**

Pour avoir une idée ou un aperçu sur le résultat de la classification, on a à décider de la superposer avec la couche OSM et Google earth. Ce qui nous a permis de voir la pertinence des classes créées et ainsi envisager éventuellement une reclassification.

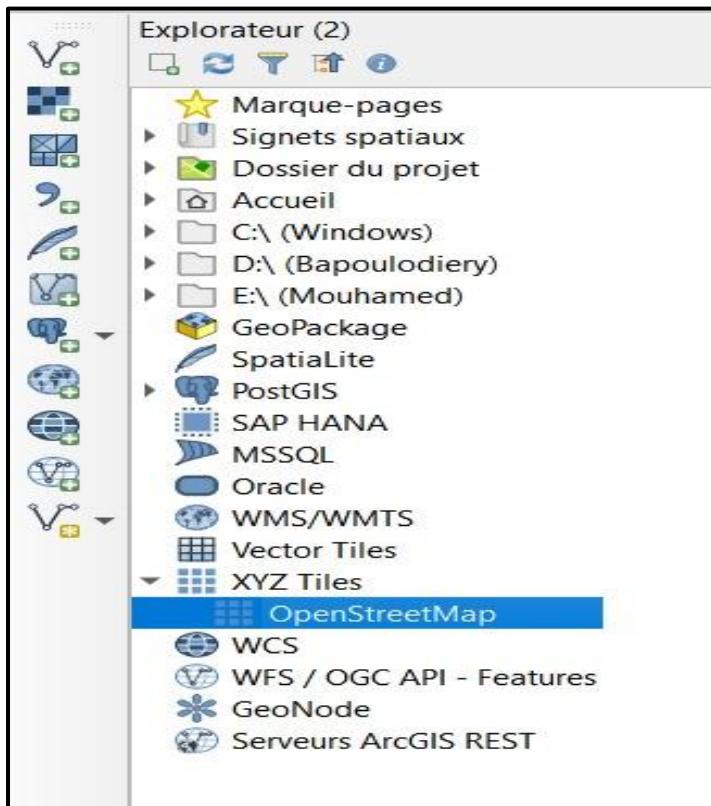


Figure 43 : chemin pour ouvrir OSM sur l'interface de QGIS

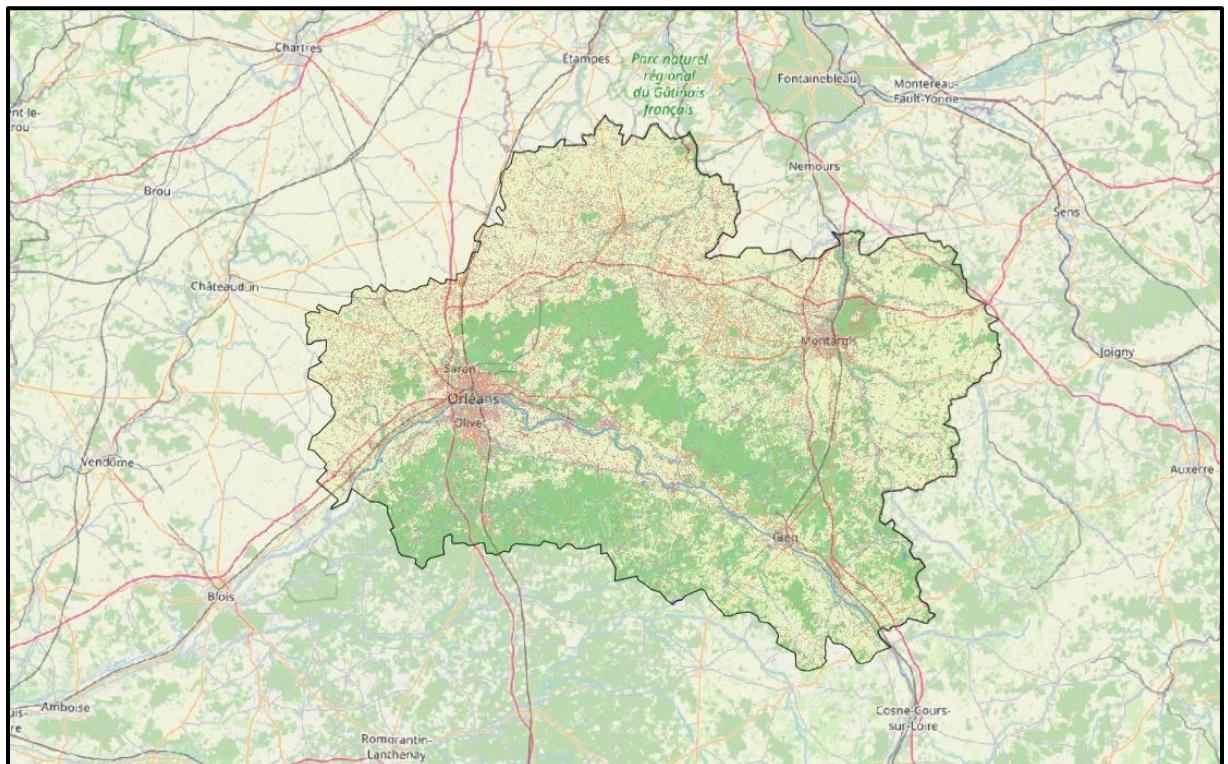


Figure 44 : superposition de OSM avec la couche classifiée

Nous avons ensuite installé une autre extension de QGIS : SCP (Semi-Automatic Classification Plugins) pour faire la reclassification de l'image : une fois dans cette extension il faut : dérouler le menu post-traitement → clic sur : reclassification → dans la

nouvelle fenêtre qui apparaît → sélectionner la classification (c'est-à-dire l'image raster concernée) → clic sur : calculer les valeurs uniques → on fait notre reclassification → clic sur : run pour exécuter.

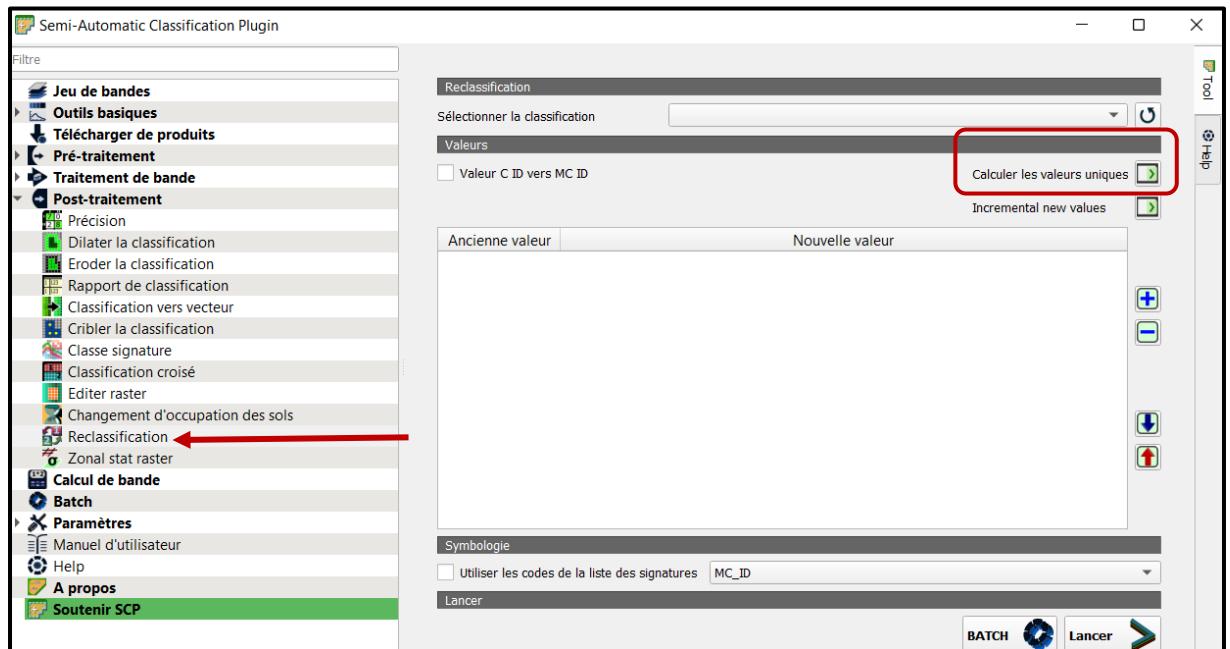


Figure 45 : fenêtre SCP et chemin pour ouvrir l'outil reclassification

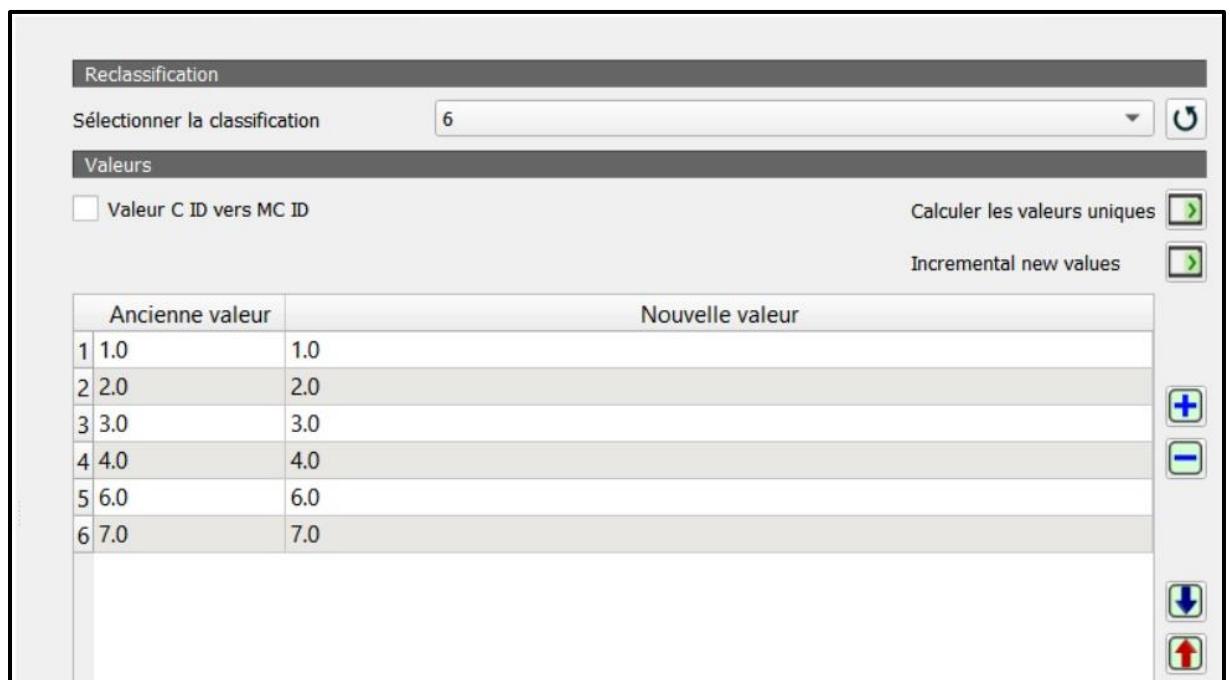
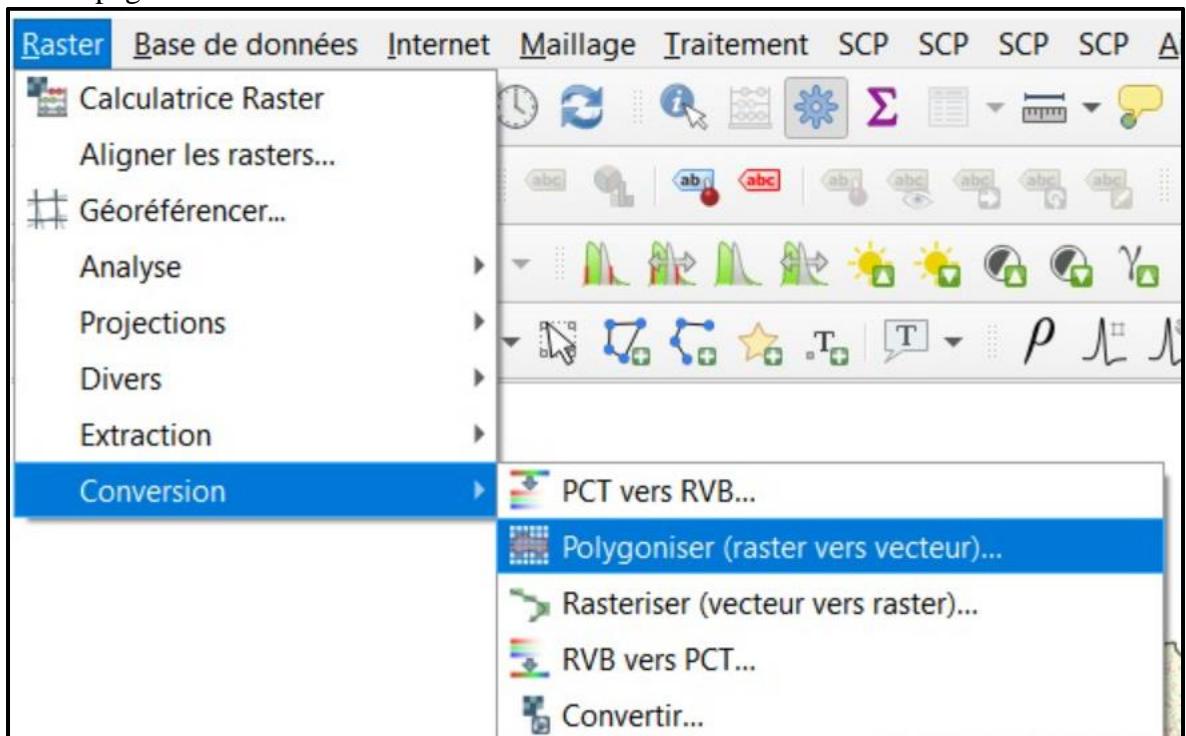


Figure 46 : fenêtre pour la reclassification

Après cela, il faudra sur l'onglet propriétés de la couche → symbologie → choisir pour le type de rendu : la palette à valeurs uniques → clic sur classer → puis sur Ok.

Ensuite nous avons changer les couleurs des classes et les réaffectant à des couleurs appropriées pour respecter les normes cartographiques : par exemple le bleu à tout ce qui hydrographie ; le vert à la végétation ; le rouge aux zones urbaines...

L'étape suivante à consister à vectoriser la couche pour avoir les classes sous formes polygonales : menu raster → conversion → polygoniser → (raster vers vecteur), puis réaliser la mise en page.

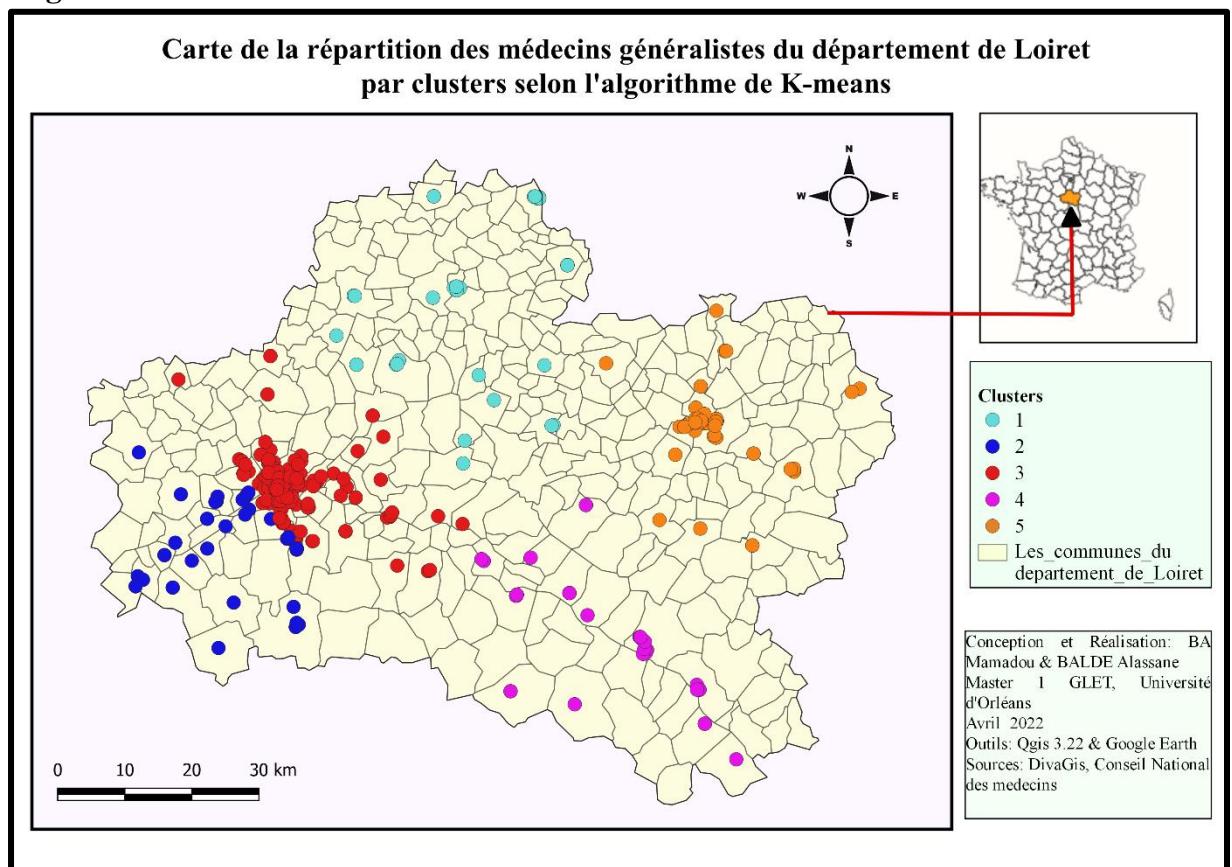


**Figure 47 : chemin pour la vectorisation des classes d'occupation des sols**

### 3/ APPLICATION, RESULTATS ET INTERPRETATION

#### 3.1 Domaine sanitaires

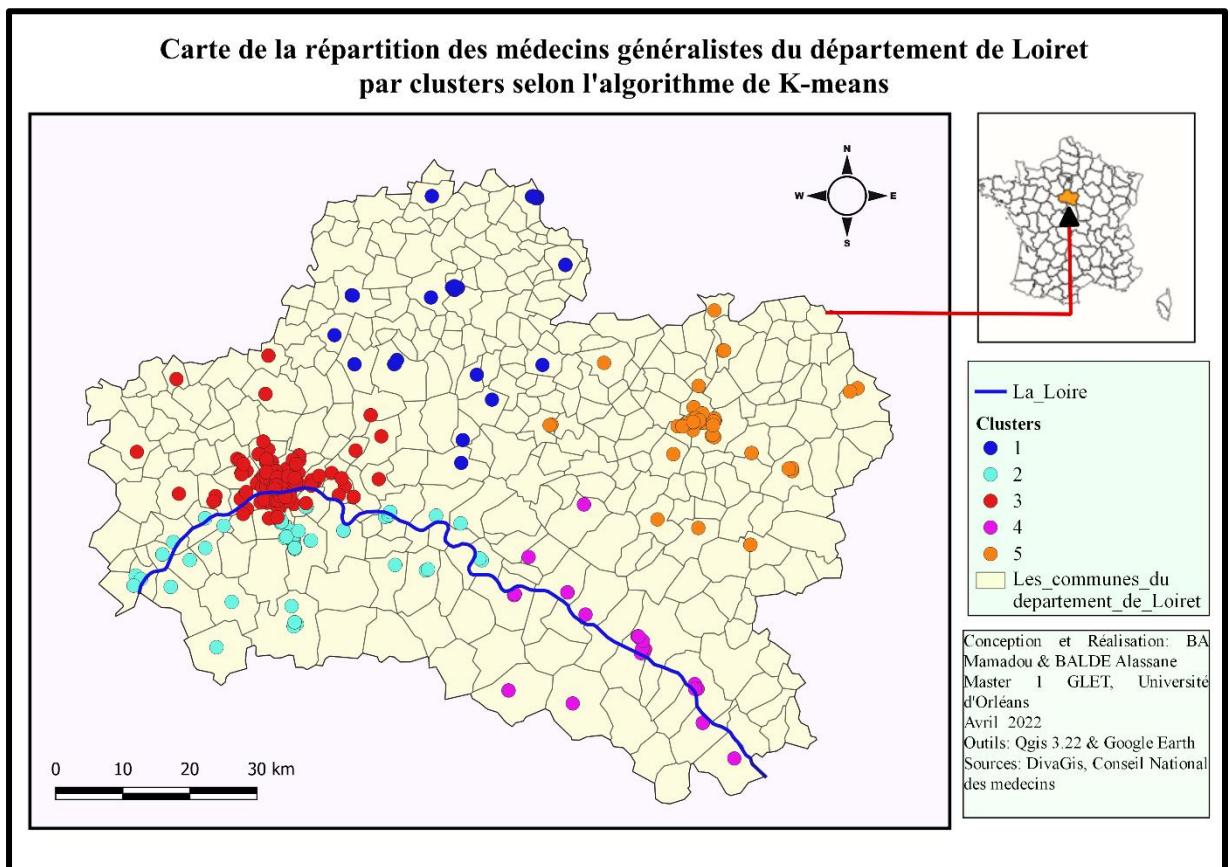
##### Algorithme K-means



**Carte 1 : Algorithmes de K-means, distance euclidienne ; fonction de lien : single link.**

Nous avons ici une carte qui représente la répartition des médecins généralistes dans le département du Loiret, classée en cluster selon l'algorithme de K-means avec comme fonction de lien : single Link et la distance euclidienne. On observe une répartition inégale de ces praticiens à l'échelle du département pratiquement concentrés dans les grandes centres urbaines tels que : l'agglomération d'Orléans, Montargis, Pithiviers, Gien, Beaugency et Meung sur Loire. D'après la classification de K-means (qui cherche à minimiser les distances entre les points), on voit les clusters suivants presque la même logique de la localisation des médecins. Ce clustering a permis de dégager cinq grandes classes : cluster 1 : qui correspond à Pithiviers et son aire urbaine d'influence ; cluster 2 : qui regroupe les villes telles que la Bretagne, Meung-sur-Loire et Beaugency localisées tout au long de la Loire dans la partie sud ouest du département ; le cluster 3 : qui concerne la zone d'Orléans et sa Métropole ; le cluster 4 : qui correspond à la Ville de Gien et son alentour et enfin le cluster 5 : qui regroupe le secteur de Montargis.

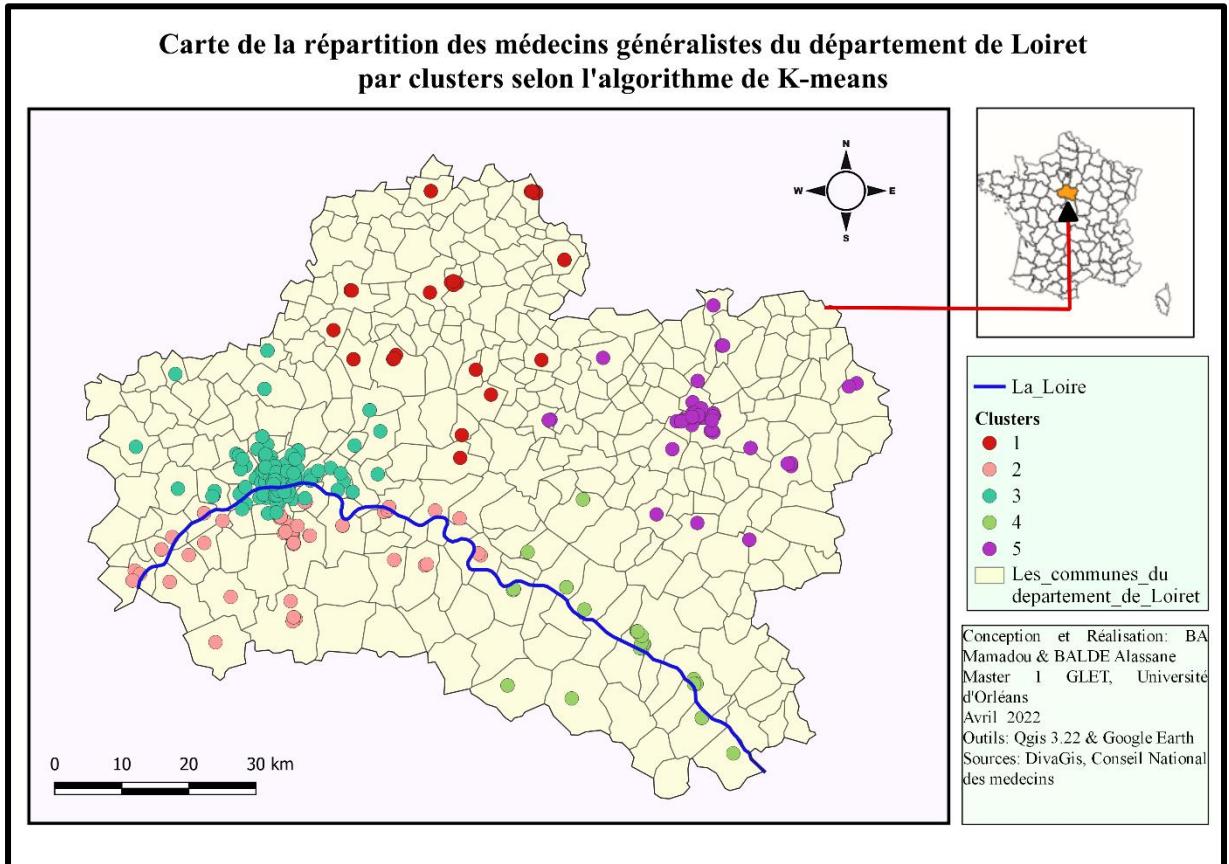
A ce stade, on peut dire cette répartition suit une logique d'urbanisation. Ce qui montre que les médecins s'installent dans les centres urbaines où l'accès aux biens et services est plus ou moins favorable que dans les espaces ruraux.



**Carte 2 : algorithmes de K-means, distance Manhattan ; fonction de lien : single Link.**

Dans cette deuxième carte qui représente aussi la répartition des médecins généralistes dans le département du Loiret avec la classification de K-means. Nous avons effectué un changement de distance en utilisant celle de Manhattan. Cette distance se base sur la somme de toutes les autres distances réelles entre deux points. C'est pourquoi cette carte montre une différence dans le regroupement notamment dans la partie sud ouest du département. Ce sont les clusters 2 et 3 qui ont connu des modifications car la zone de Jargeau, Saint-Denis de l'Hôtel et Sandillon et la Source se retrouvent dans le cluster 2 alors que avec la distance euclidienne (Carte 1) elles faisaient partie du cluster 3 c'est-à-dire Orléans et sa métropole. De même que Patay et la zone de Epieds en Beauce qui se trouvaient dans le cluster 2 de la carte 1 sont ici des éléments du cluster 3. Les autres clusters n'ont pas connu de changements.

En comparant ces deux cartes, on peut dire que celle réalisée avec la division euclidienne donne le meilleur résultat en termes de qualité des clusters car dans ce cas le calcul de distance est effectué par le plus proche voisin et ici la prise en compte de toutes les distances réelles possibles peut dans certains cas donner des formes de cluster avec des points très éloignés à l'intérieur.

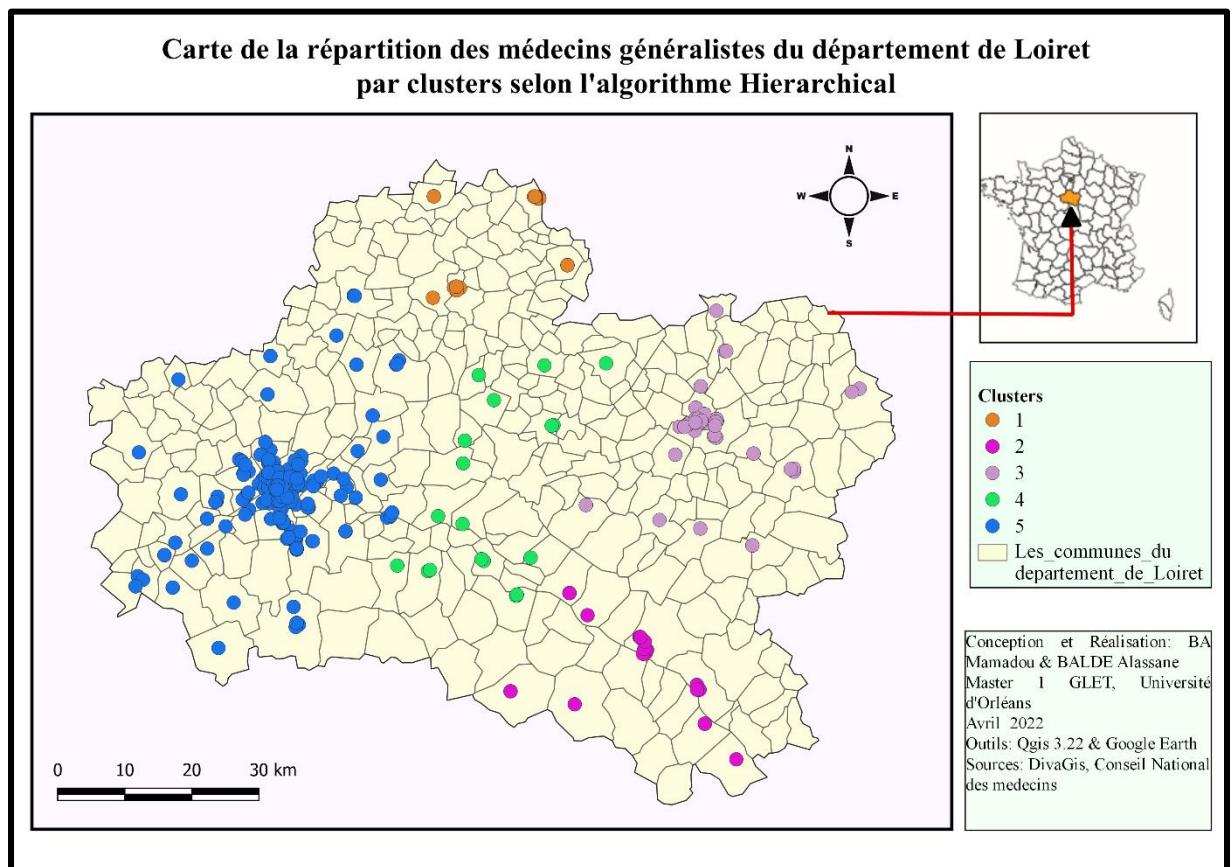


**Carte 3 : algorithmes de K-means, distance Manhattan ; fonction de lien : Wards (Lance-williams) :**

Dans cette troisième carte, où nous avons utilisé comme fonction de lien : Ward de la famille des Lance Williams pour voir quel résultats cela peut donner. Ainsi, on remarque c'est la même classification que celle de la distance de Manhattan avec comme fonction de lien Single Link. Après avoir réalisé plusieurs essayages avec la même distance en toutes les fonctions de Lance-Williams, le résultat reste le même.

Dans l'ensemble, on peut retenir que pour l'algorithme K-means, la distance euclidienne avec single Link comme fonction de liaison donne la meilleure classification (Carte 1).

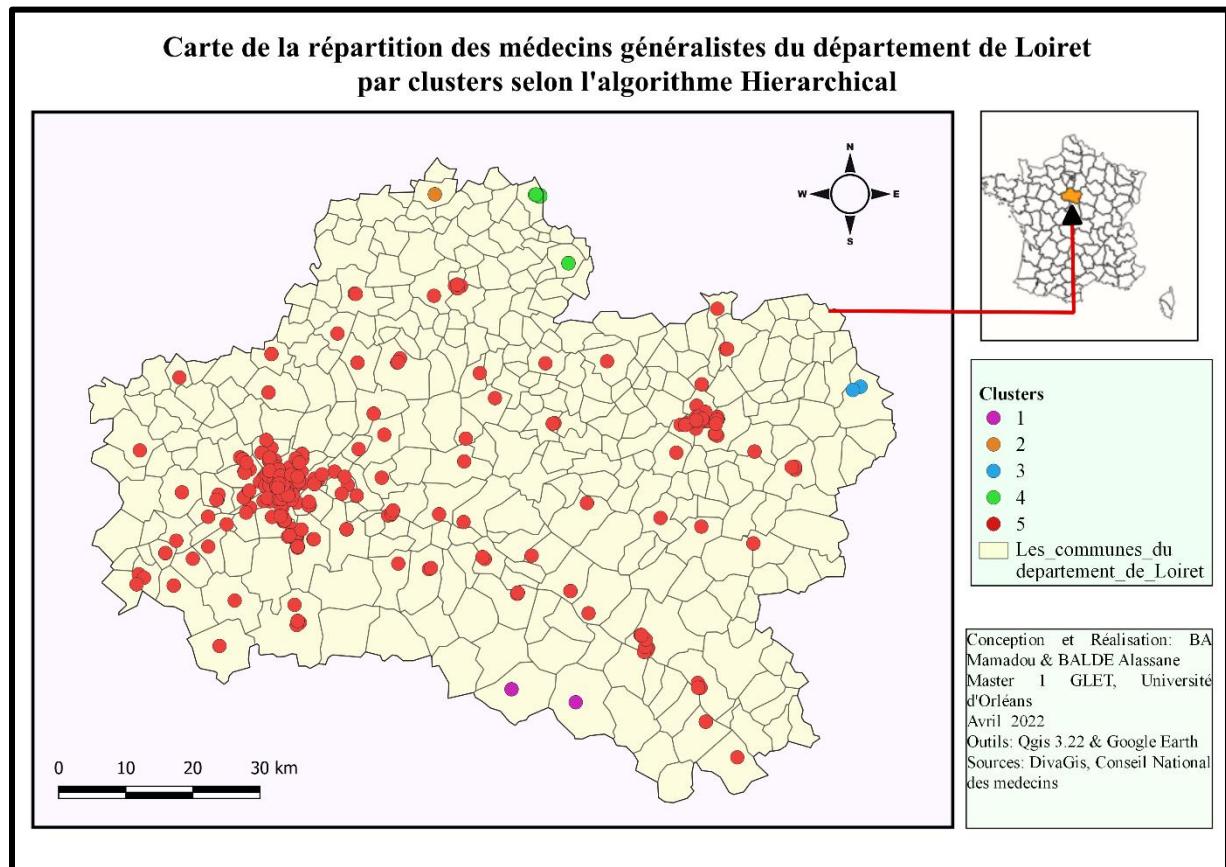
### Algorithmes de classification Hiérarchique



**Carte 4 : Algorithme de classification Hiérarchique, distance Euclidienne, fonction average ( moyenne)**

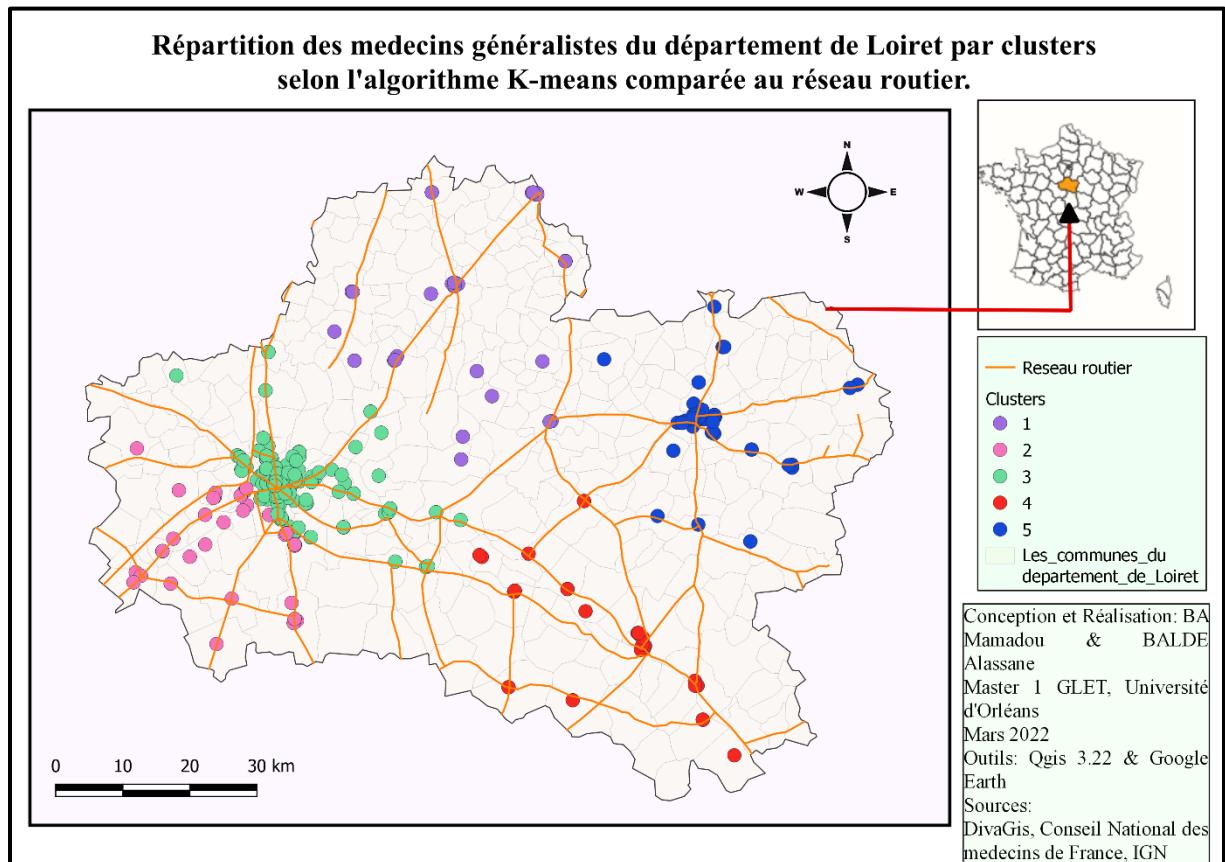
Dans cette quatrième carte, nous avons utilisé l'algorithme de la classification hiérarchique ascendante. Cette dernière qui part du principe agglomératif ou divisif, elle débute avec autant de clusters qu'il y a de points puis les agrège au fur et à mesure selon la fonction de lien choisie.

Concernant cette algorithme, nous avons fait des traitements avec toutes les fonctions de liens et les deux distances ; seule la combinaison de la distance euclidienne et la fonction de lien : Average (moyenne) présente le meilleur résultats (Carte 4). Les autres présentes des dis fonctionnalités par rapport à la distance des points à l'intérieur d'un cluster. C'est le cas de la carte 5.



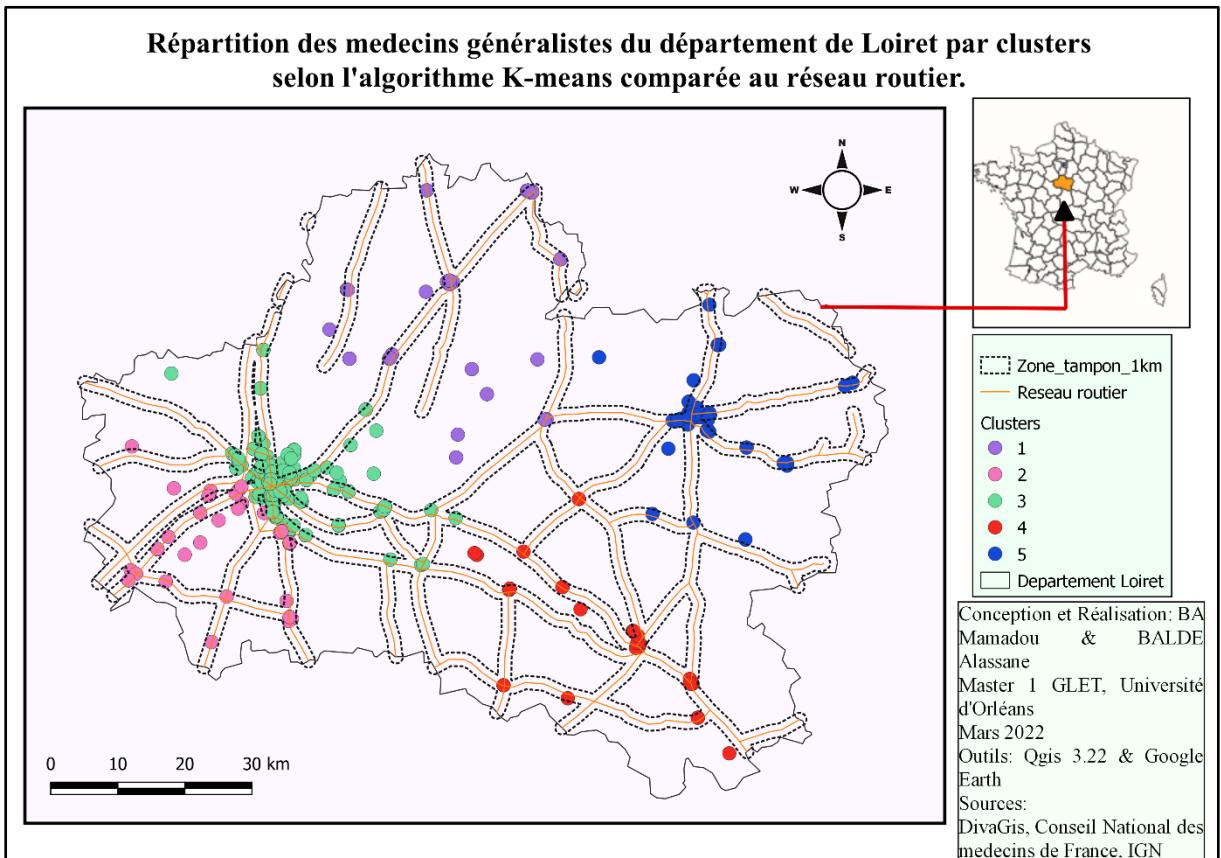
**Carte 5 : Algorithme hierarchical, distance euclidienne, fonction single Link**

Ici, on voit que le clustering est très loin de correspondre à la réalité. Presque on à tous les adresses dans un seul cluster (509 sur 517).



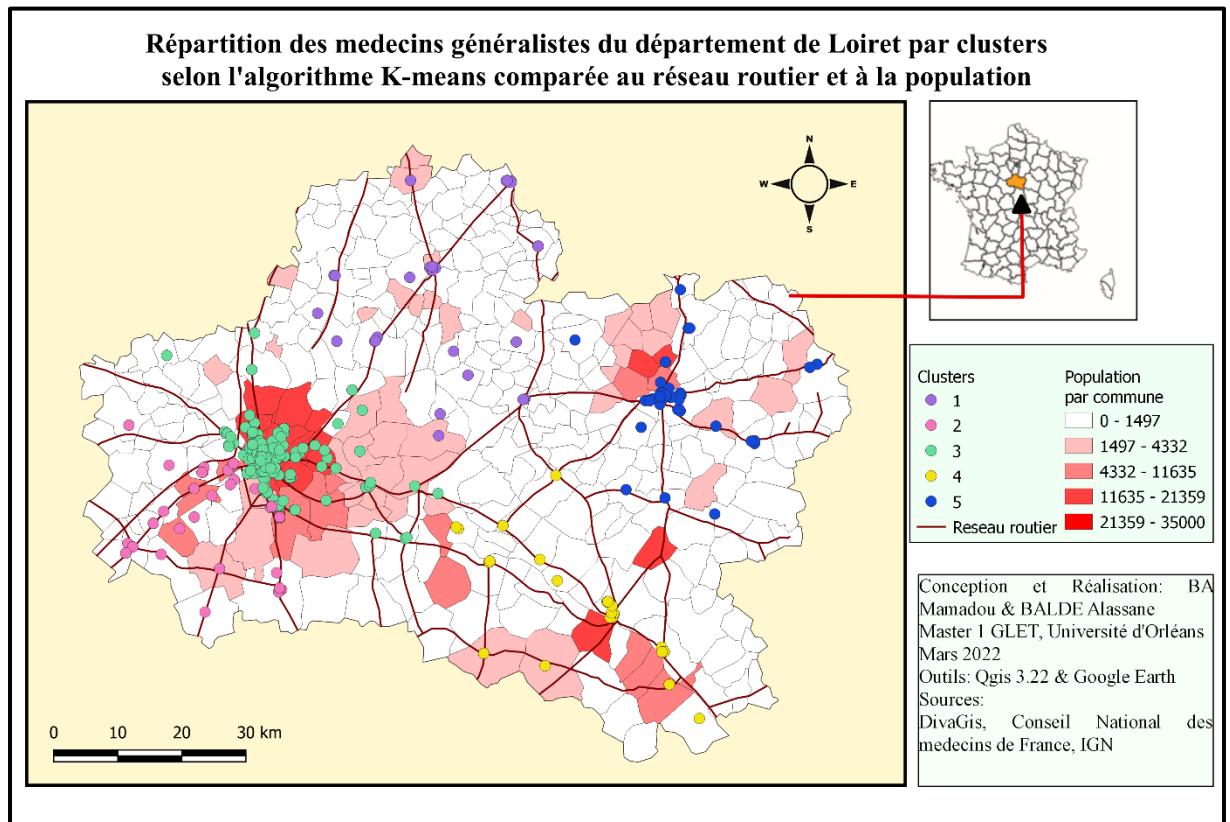
### Carte 6 : superposition des clusters avec les routes

Ici on a la carte des clusters avec (algorithme K-means, distance euclidienne, superposée avec du réseau routier. La première information qui saute à l'œil est que la localisation des médecins généralistes suit le réseau routier. La plus grande partie des médecins généralistes se sont logés à coté des voies routières. Ceci peut être expliquer par la facilité d'accessibilité géographique et de la mobilité aussi bien pour les patients que les médecins eux-mêmes. Tandis que dans les zones moins desservies en terme de réseau routier sont peu denses en terme de médecins généralistes. Par conséquent ces zones constituent des déserts médicaux comme les communes de Isdes, . La Bussière, Adon, Charmont-en-Beauce, Allainville-en-Beauce, Ervauville etc. Ce qui constitue un problème en termes d'accès rapide aux soins en cas d'urgence aux populations de ces localités.



### Carte 7 : Zone tampon de 1km autour du réseau routier

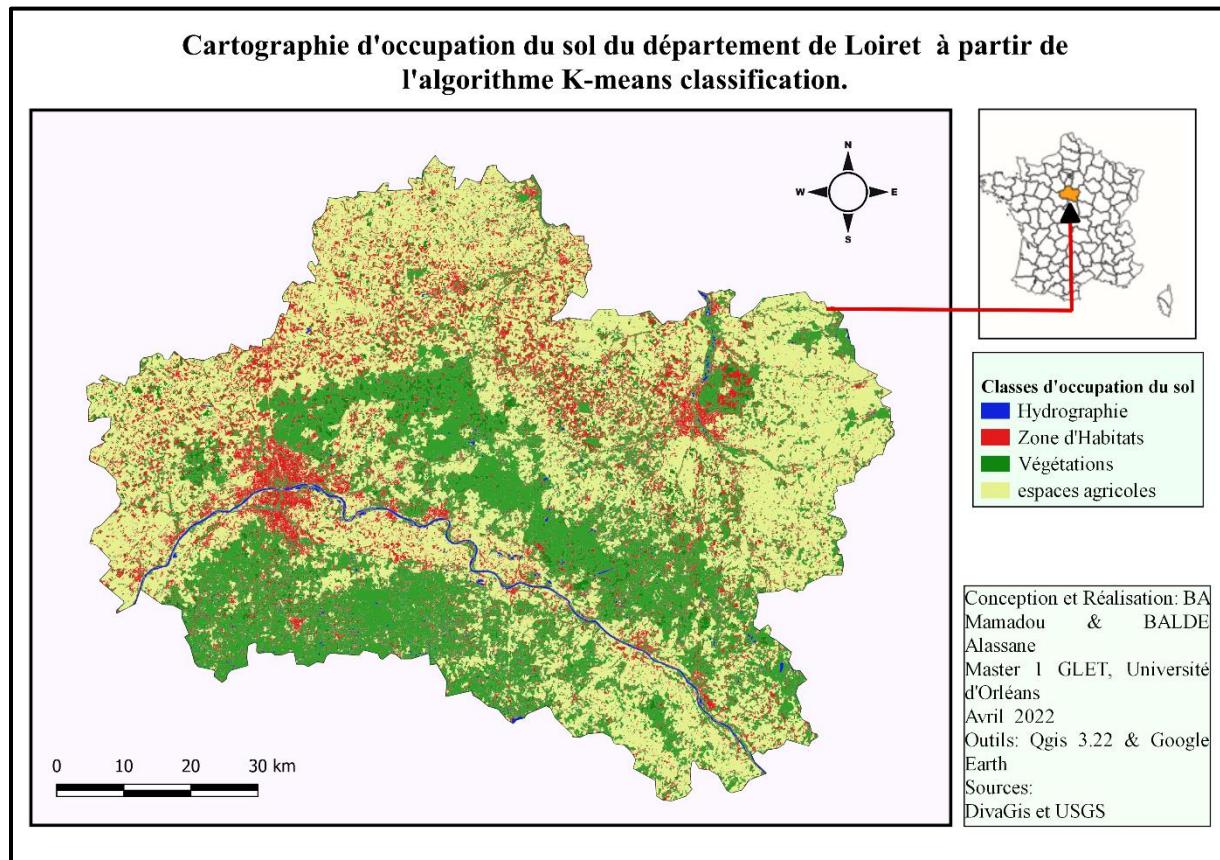
Dans cette carte on a choisi de faire une zone tampon de 1km autour des routes pour confirmer l'hypothèse comme quoi la majeure partie des médecins généralistes sont localisés tout près des axes routiers.



**Carte 8 : superposition des clusters avec la population par communes du département de Loiret**

Dans cette carte on ajouté une autre variable qui la répartition de la population par communes. On voit que les zones où on a de très forte concentration de la population renferme une présence importante de médecins généralistes par exemple dans les grands centres urbains (Orléans, Montargis, Gien, Pithiviers).

### 3.2 DOMAINE D'OCCUPATION DU SOL :



### Carte 9 : occupation du sol à l'échelle du département

Carte montre l'occupation du sol à l'échelle du département de Loiret que nous avons réalisé avec l'algorithme K-means classification. Quatre types de zones se dégagent : hydrographie, végétation, habitats et espaces agricoles. Les grandes zones urbaines apparaissent clairement en rouge (Orléans, Montargis, Gien et Pithiviers). Les cours d'eau et plan d'eau apparaissent en bleu (exemple de la Loire et du Loiret, étang du Pin, étang de Givarlet...). En vert on a la végétation (forêt domaniale d'Orléans, forêt domaniale de Lamotte-Beuvron etc.)

## **CONCLUSION**

L'extension clustering point est un outil qui permet la classification non supervisée et le partitionnement de données en clusters homogènes, présentant des caractéristiques similaires. Cette classification se fait à travers les choix de plusieurs paramètres tels que les algorithmes de classification, les types de distances, les fonctions de liens... L'objectif des algorithmes de clustering consiste à donner du sens aux données et d'en extraire de la valeur à partir d'un ensemble de données structurées et non structurées. Le choix des algorithmes et paramètres de traitements est fonction de l'objectif de l'étude visée, des propriétés des ces algorithmes... Le clustering est aujourd'hui utilisé dans plusieurs domaines notamment dans les domaines de l'économie et du marketing, des systèmes d'informations géographiques, l'environnement, dans le domaine sanitaires...

Dans ce manuel, le clustering est appliqué dans deux domaines : d'abord la répartition des médecins généralistes dans le département du Loiret, pour mieux comprendre leurs localisation, leurs facteurs explicatifs afin de ressortir les zones de fortes concentration et les déserts médicaux. Ce qui a donné des résultats satisfaisants car on voit une forte implantation des médecins dans grandes centres urbains (Orléans et sa métropole, Montargis et sa zone d'influence, Pithiviers et ses alentours, ainsi que Gien) près des grands axes routiers et tout au long de la Loire ; mais aussi des zones avec une faible présence de ces praticiens dans certaines localités périphériques comme Isdes, Vanne sur Cosson, Marcilly-en-villette, Griselles, La Selle-sur-le-Bied...

Ensuite en ce qui concerne l'occupation des sols, l'objectif était d'appliquer le clustering à des données de type raster particulièrement des images satellites afin de faire de la classification pour les comparées avec la réalité du terrain. Les résultats sont plus ou moins satisfaisant dans le sillage que les grands ensembles d'occupation du sols sont sortis tels que les zones habitats, la végétations, l'hydrographie, et les surfaces agricoles.

Malgré ces résultats, l'extension présente des défaillances dans la mesure où, on note de mauvais résultats en changeant les fonctions de liens, en plus actuellement, l'algorithme de Fuzzy- C-means intégré dans l'extension ne fonctionne pas. A cela s'ajout le fait qu'il faudra effectuer plusieurs tests en combinant plusieurs paramètres afin d'obtenir un résultats approprié.

**Bibliographie :**

ATIF J. Machine Learning et Data Mining : clustering, groupement, segmentation ; Université Paris Dauphine, certificat Data Science, 141p

GASSO G. & LERAY P. ; Clustering, INSA Rouen-Département ASI, Laboratoire LITIS, 64p

**Webographie**

<http://jenkner.com/ClusterPoints>

<https://github.com/jenkner/ClusterPoints>

<https://www.conseil-national.medecin.fr/annuaire/resultats>

<https://www.conseil-national.medecin.fr/annuaire/resultats>

[Clustering \(univ-lyon2.fr\)](#)

<https://support.minitab.com/fr-fr/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/cluster-variables/methods-and-formulas/linkage-methods/>

[https://fr.isecosmetic.com/wiki/Nearest-neighbor\\_chain\\_algorithm#Complete\\_linkage\\_and\\_average\\_distance](https://fr.isecosmetic.com/wiki/Nearest-neighbor_chain_algorithm#Complete_linkage_and_average_distance)

<https://analyticsinsights.io/le-clustering-definition-et-implementations/>

<https://dataanalyticspost.com/Lexique/clustering/>

<https://datacientest.com/dendrogramme>

<https://openclassrooms.com/fr/courses/4525281-realisez-une-analyse-exploratoire-de-donnees/5177936-effectuez-une-classification-hierarchique>

<https://mrmint.fr/algorithme-k-means>

## **LISTE DES FIGURES**

Figure 1: Onglet pour l'installation de l'extension .....	2
Figure 2 : fenêtre des extensions.....	3
Figure 3 : Sélection de l'extension.....	3
Figure 4 : Installation de l'extension.....	4
Figure 5 : Boite à outils pour accéder à l'extension.....	4
Figure 6 : Boite à traitement de QGIS.....	5
Figure 7 : Fenêtre de l'extension clustering pour le choix des paramètres.....	6
Figure 8 : Zoom sur le choix de l'algorithme de clustering .....	7
Figure 9 : Étape de choix de la fonction de lien.....	8
Figure 10 : Étape de choix de la distance.....	9
Figure 11 : Étape de choix du nombre de cluster.....	10
Figure 12: Étape de choix du coefficient de Fuzzy dans le cas où on utilise l'algorithme Fuzzy C-Means.....	10
Figure 13: Étape de choix des centiles.....	11
Figure 14 : Étape de choix du pourcentage de contribution des champs d'attributs.....	12
Figure 15 : fenêtre de téléchargement du logiciel QGIS.....	13
Figure 16 : Choix du système d'exploitation.....	14
Figure 17 : Choix de la version de QGIS.....	14
Figure 18 : fenêtre pour télécharger ou consulter d'autres versions antérieures ou extensions de QGIS.....	15
Figure 19 : Licence, accès au guide d'installation et à d'autres extensions.....	15
Figure 20 : fenêtre d'installation.....	16
Figure 21 : Acceptation des termes de licence.....	17
Figure 22 : Chemin d'enregistrement du programme.....	18

Figure 23 : Début de l'installation.....	19
Figure 24 : Installation en cours.....	20
Figure 25 : Finalisation del'installation.....	21
Figure 26 : Accès au dossier QGIS et ouverture.....	22
Figure 27 : interface d'accueil.....	22
Figure 28 : recherche pour ouvrir Google Earth.....	23
Figure 29 : fenêtre pour créer un projet sur Google Earth et l'enregistrer en format KML.....	24
Figure 30 : Exemple de localisation d'une adresse et l'ajout dans le projet.....	24
Figure 31 : l'exportation du projet en format KML.....	25
Figure 32 : chemin de jointure des attributs.....	25
Figure 33 : téléchargement de données sur Diva-GIS et découpage des routes avec la zone d'étude.....	26
Figure 34 : chemin pour ouvrir l'extension Cluster points.....	26
Figure 35 : accès au site de USGS .....	27
Figure 36 : Fenêtre de choix des paramètres pour le téléchargement des images.....	27
Figure 37 : Fenêtre de choix des paramètres pour le téléchargement des images.....	27
Figure 38 : Visualisation des images et téléchargement des images.....	28
Figure 39 : chemin du fusionnement et de celui du découpage des images .....	28
Figure 40 : fenêtre d'activation de OTB.....	29
Figure 41: chemin d'accès à l'outil OTB et de l'outil de classification K-means .....	29
Figure 42 : modification de symbologie de la couche classifiée.....	30
Figure 43 : chemin pour ouvrir OSM sur l'interface de QGIS.....	31
Figure 44 : superposition de OSM avec la couche classifiée .....	31
Figure 45 : fenêtre SCP et chemin pour ouvrir l'outil reclassification.....	32
Figure 46 : fenêtre pour la reclassification .....	32
Figure 47 : chemin pour la vectorisation des classes d'occupation des sols.....	33

**LISTE DES CARTES :**

Carte 1 : Algorithmes de K-means, distance euclidienne ; fonction de lien : single link.....	34
Carte 2 : algorithmes de K-means, distance Manhattan ; fonction de lien : single Link.....	35
Carte 3 : algorithmes de K-means, distance Manhattan ; fonction de lien : Wards (Lance-williams) .....	36
Carte 4 : Algorithme de classification Hiérarchique, distance Euclidienne, fonction average ( moyenne).....	37
Carte 5 : Algorithme hierarchical, distance euclidienne, fonction single Link.....	38
Carte 6 : superposition des clusters avec les routes .....	39
Carte 7 : Zone tampon de 1km autour du réseau routier.....	40
Carte 8 : superposition des clusters avec la population par communes du département de Loiret.....	41
Carte 9 : occupation du sol à l'échelle du département.....	42

## **TABLE DES MATIERES**

<b>INTRODUCTION GÉNÉRALE .....</b>	<b>1</b>
<b>I°) L'extension Clustering Point.....</b>	<b>1</b>
1) Définition.....	1
2) Installation de l'extension sur QGIS.....	2
3) Interface de l'extension.....	5
<b>II°) DESCRIPTION ET FONCTIONNEMENT.....</b>	<b>6</b>
1) Les algorithmes.....	6
1.1 Algorithme de K-means.....	6
1.2 Algorithme de Classification Hiérarchique.....	7
1.3 Algorithme de Fuzzy C-means.....	7
2) Les fonctions de liens.....	7
2.1 Single Link.....	7
2.2 Single Link ( Lance Williams).....	7
2.3 Complete ( Lance Williams).....	7
2.4 Median ( Lance Williams).....	8
2.5 Average ( Lance Williams).....	8
2.6 Wards ( Lance Williams).....	8
2.7 Centroide ( Lance Williams).....	8
3) Type de calcul de distance.....	8
3.1 Distance euclidienne.....	9
3.2 Distance Manhattan.....	9
4) Définition du nombre de Clusters.....	9
5) Coefficient de Fuzzy.....	10
6) Caractéristiques des centiles de distance du cluster.....	11
7) Pourcentage de contribution et champs attribut.....	11
<b>III°) C'EST QUOI QUANTUM GIS (QGIS) ?.....</b>	<b>12</b>
1) Les point forts du logiciel.....	12
2) Domaine et thématique d'application.....	12
3) Versionnement et téléchargement.....	13
4) Installation.....	15
<b>IV°) APPLICATION DE L'EXTENSION.....</b>	<b>23</b>
1) Thématique d'application.....	23
1.1 Santé.....	23
1.2 Cartographie de l'occupation du sol.....	23
2) Données et méthodes.....	23
2.1 Données de la santé.....	23
2.2 Données d'occupation du sol.....	26
3) Application, résultats et interprétation.....	34

3.1 Domaine sanitaire.....	34
3.2 Domaines d'occupation du sol.....	41
<b>CONCLUSION.....</b>	<b>43</b>
<b>BIBLIOGRAPHIE ET WÉBOGRAPHIE.....</b>	<b>44</b>
<b>Liste des Figures.....</b>	<b>45</b>
<b>Liste des cartes .....</b>	<b>47</b>
<b>TABLES DES MATIÈRES.....</b>	<b>48</b>