

Analyse des données INSEE sur les quartiers

Etudiants :

Gernido HANAMPATRA

Maryem JEDID

Henry SAVIGNY

Sous l'encadrement de:

M.Fabien DUCHATEAU

M.Franck FAVETTA

Table des matières

1. Introduction.....	1
1.1. Contexte.....	1
1.2. Objectifs.....	2
1.2.1 Étude de la littérature scientifique.....	2
1.2.2 Mise à jour de la base de données "mongiris".....	2
1.2.3 Sélection d'indicateurs et développement d'un script Python.....	2
1.2.4 Analyse des résultats.....	3
2. Travail à réaliser.....	3
3. Planning prévisionnel.....	4

1. Introduction

Le présent cahier des charges décrit les objectifs, les méthodes, et les étapes du projet d'analyse des données INSEE sur les quartiers, qui vise à vérifier l'exactitude des données INSEE en se basant sur des applications cartographiques. Le projet est mené par M.Fabien Duchateau et M.Franck Favetta du laboratoire LIRIS.

1.1. Contexte

En France, l'INSEE divise le territoire en 50000 IRIS (Ilots Regroupés pour l'Information Statistique), qui peuvent être considérés comme des quartiers. Chaque quartier est caractérisé par environ 550 indicateurs, tels que le nombre de restaurants, de boulangeries et de piscines. Les données INSEE sont dispersées dans des fichiers Excel et ont été consolidées dans une base de données MongoDB appelée "mongiris", accessible via le lien <https://gitlab.liris.cnrs.fr/fduchate/mongiris>. Toutefois, il est important de noter que la mise à jour de ces données n'est pas systématique chaque année. De plus, certaines communes, souvent par manque de ressources, ne fournissent pas ces informations de manière régulière. Un autre enjeu notable concerne la représentation des données manquantes ou d'un comptage nul, les deux étant indistinctement représentés par la valeur zéro. Cette ambiguïté peut induire en erreur lors de l'analyse des données et nécessite une attention particulière pour une interprétation correcte.

L'objectif principal de ce projet est de vérifier la précision des données INSEE en utilisant des applications cartographiques telles que Bing Maps, Foursquare, Here, et OpenStreetMap. Les différents fournisseurs de cartes retournent des points d'intérêt (POI) potentiellement équivalents, mais représentés de manière différente, nécessitant une étape de comparaison (appariement spatial) pour détecter les doublons.

1.2. Objectifs

Le projet a pour objectifs principaux :

1.2.1 Étude de la littérature scientifique

Passer en revue la littérature scientifique récente sur l'appariement spatial des entités géographiques. Cette revue se basera sur des références.^{1 2 3 4}

1.2.2 Mise à jour de la base de données "mongiris"

Mettre à jour la base de données "mongiris" avec les dernières données INSEE pour les années 2022/2023. Cette mise à jour est essentielle pour assurer l'exactitude et la pertinence des informations utilisées dans les analyses suivantes.

1.2.3 Sélection d'indicateurs et développement d'un script Python

Choisir des indicateurs INSEE pertinents, en se concentrant sur ceux vérifiables via des fournisseurs cartographiques. Cette sélection exclut des indicateurs comme le taux de chômage par catégorie socio-professionnelle, qui ne sont pas mesurables cartographiquement. Développer un script Python permettant d'interroger au moins deux fournisseurs cartographiques pour dénombrer les POI d'une zone donnée, tout en évitant les doublons. Assurer la qualité de l'algorithme en termes de précision et de fiabilité dans le comptage des POI. L'évaluation se concentrera sur l'efficacité de l'algorithme à filtrer les doublons et les données non pertinentes.

¹ Kai Sun, Yunqiang Zhu, et Jia Song. **"Progress and challenges on entity alignment of geographic knowledge bases."** ISPRS International Journal of Geo-Information, 8(2):77, 2019.

² Emerson MA Xavier, Francisco J Ariza-López, et Manuel A Urena-Camara. **"A survey of measures and methods for matching geospatial vector datasets."** ACM Computing Surveys (CSUR), 49(2):1–34, 2016.

³ Suela Isaj, Esteban Zimányi, et Torben Bach Pedersen. **"Multi-source spatial entity linkage."** Dans Symposium on Spatial and Temporal Databases, pages 1–10, 2019.

⁴ Setu Shah, Vamsi Meduri, et Mohamed Sarwat. **"Gem: An efficient entity matching framework for geospatial data."** Dans Geographic Information Systems, pages 346–349, 2021.

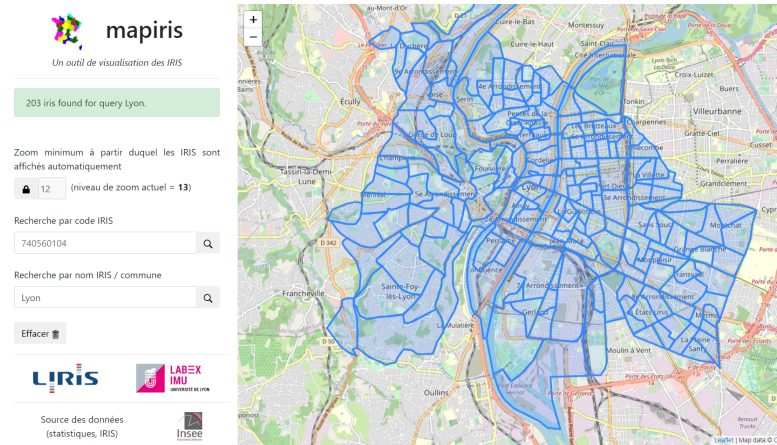


Image 1 - Visualisation du découpage de Lyon en IRIS. Comme les IRIS ne sont pas automatiquement délimitées par les fournisseurs cartographiques, nous devons approximer les délimitations.

1.2.4 Analyse des résultats

Choisir spécifiquement deux à trois zones pour l'analyse. Analyser les résultats de l'appariement avec les données INSEE pour ces zones. Évaluer la qualité des données INSEE en fonction de plusieurs critères, incluant la zone géographique, le type de quartier, et les indicateurs spécifiques. Identifier les quartiers avec des données obsolètes ou incomplètes.

2. Travail à réaliser

Pour atteindre nos objectifs, notre script Python utilisera les polygones déjà définis dans la base de données “mongiris” pour délimiter précisément les IRIS. Cette délimitation nous permettra d'effectuer des statistiques pertinentes sur chaque IRIS, comme le nombre d'hôtels, de restaurants, etc. Il est essentiel de récupérer avec précision les Points d'Intérêt (POI) de chaque IRIS.

Étant donné que nous extrayons des données de sources variées, ces données sont représentées de manière différente. Il est donc crucial de comparer ces informations pour évaluer la similarité entre chaque POI. Pour ce faire, nous tenterons de mettre en œuvre un algorithme configurable en termes de poids, seuils et coefficients. Dans la mesure où cela s'avère impossible, nous implémenterons, testerons et fournirons au moins deux algorithmes distincts de spatial entity matching, de sorte à ce que l'on puisse les comparer entre eux. Si le temps le permet et en fonction de notre satisfaction, nous ne nous limiterons pas à ces deux algorithmes seulement, afin d'avoir un aperçu un peu plus exhaustif des techniques de spatial entity matching.

Enfin, nous mesurerons la précision et le rappel de l'algorithme SEM pour déterminer son efficacité à identifier correctement les correspondances entre les POI et à minimiser les erreurs de correspondance. Ensuite, nous comparerons le nombre de POI identifiés par l'algorithme SEM avec les indicateurs INSEE correspondants. Cette comparaison nous permettra de juger si les données INSEE reflètent fidèlement la réalité ou si elles sont éloignées des observations actuelles.

3. Planning prévisionnel

Semaine	Tâches à réaliser
<i>Semaine 1</i> 23/10/2023 au 29/10/2023	Compréhension du sujet, définition du cahier des charges. Mise en place et configuration de la base de données MongoDB, et visualisation des données.
<i>Semaine 2</i> 30/10/2023 au 05/11/2023	Recherche des algorithmes de spatial entity matching (SEM), et choix des fournisseurs cartographiques (et documentation de leurs APIs). Synthèse des recherches.
<i>Semaine 3-7</i> 06/11/2023 au 10/12/2023	Effectuer des requêtes API aux fournisseurs pour collecter tous les Points d'Intérêt (POI) dans une zone élargie. Commencer le développement d'un algorithme spécifique pour filtrer les POI récupérés, en déterminant leur appartenance à un IRIS donné.
<i>Semaine 8-10</i> 11/12/2023 au 31/12/2023	Analyse des données pour faire des statistiques, et/ou évaluer la qualité des données (données pas à jour, ou erronées).
<i>Semaine 11</i> 01/01/2024 au 04/02/2023	Rédaction du rapport et tournage de la vidéo
04/02/2023 à la fin	Vidéo