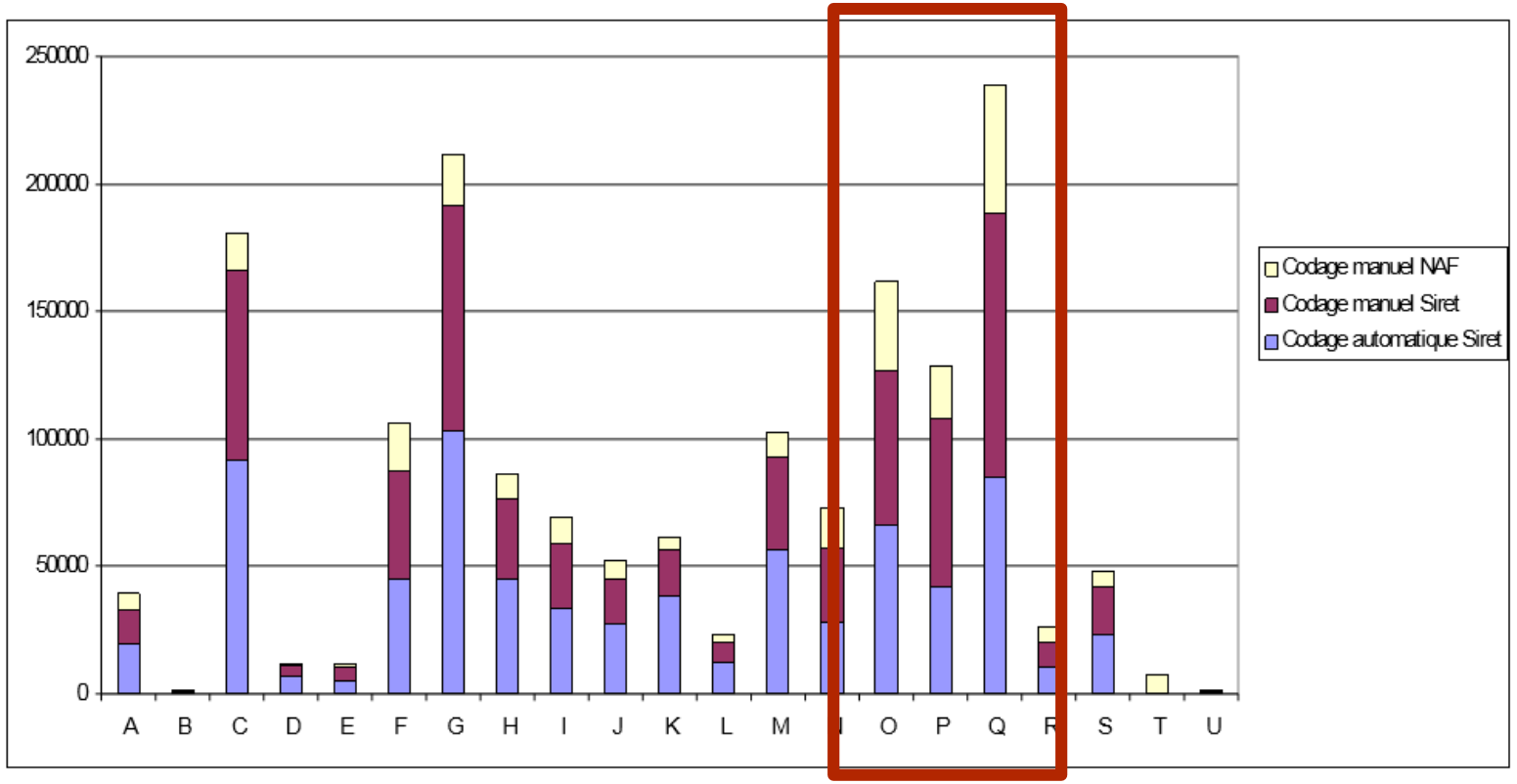
Identification de l'établissement employeur dans les secteurs O, P et Q

HACKATHON INSEE 2018

Introduction : que font les QOPS ?



Répartition des différents types de codage en fonction des secteurs d'activité
Source : EAR 2017

Au 31 décembre 2002, l'emploi public regroupait **5,2 millions de personnes.**

Au sens de l'Observatoire de l'emploi public, ie qui dépendent d'administrations ou d'organismes dans lesquels le recrutement de droit commun de l'agent relève des titres II, III et IV du statut de la Fonction publique.

Introduction : que font les QOPS ?

Spécificité des QOPS :

- **Etablissement aussi points d'intérêt** dans leurs villes. Ex : nom_voie est « Collège Victor Hugo »
- **Employeur mal (ou pas) identifié**. Ex : raison sociale déclarée « Education nationale (collège Victor Hugo) »
- **Raisons sociales vagues**, peu informatives… Ex : « Education nationale »

➢ **Un secteur de taille significative avec des spécificités propres**

Champ : 563 295 personnes du RP 2017 (sur 1 643 901 au total).

Restriction de champ par mots-clefs.

Pistes à améliorer

Ce qui a marché, et ce qui n’a pas marché…

Idées principales

- **Nettoyage des champs spécifique aux secteurs O, P et Q**
  - Des adresses : repérage des raisons sociales contenues dans les adresses
  - Des raisons sociales : information importante souvent contenue entre parenthèses
- **Recherche de proximités :**
  - Lexicographiques sur la raison sociale (similarité cosinus)
  - Géographiques sur les coordonnées (plus proches voisins)
- **Exploitation des complémentarités entre les API** :
  - SIRENE
  - BAN
  - Addok

Processus final

1/5 Nettoyage des champs

- **Extraction de l'information pertinente dans les raisons sociales** (ex : « Education nationale (Collège Victor Hugo) »).

- **Récupération de la raison sociale lorsqu'elle est contenue dans l'adresse** (ex : l'adresse est « Collège Victor Hugo »)

2/5 Requêtage de Sirene.Addok

Utilisation d'une source d'information complémentaire, qui permet de résoudre des cas « triviaux ».

API créée par Etalab, combine plusieurs sources d'information (OpenStreetMap, base SIRENE, BAN et BANO)

Moteur de recherche :

- Input : à code commune exact, recherche sur la raison sociale
- **Output : SIRET**

✓ **30 % de réponses acceptées (ie où le score est supérieur à 0.5)**

Limite : seules 2/3 sont correctes (arbitrage précision - nombre de retours)

3/5 Appariement des raisons sociales avec SIRENE

Pour chaque individu, on cherche dans sa commune de travail *et les communes adjacentes* la **raison sociale la plus ressemblante à celle qu'il a déclarée**.

→ **Similarité cosinus entre les raisons sociales du RP et celles de Sirene.**

- Ne pénalise pas la disparition de déterminants
- Ne pénalise pas les petits changements
- Peu sensible à l'ordre des mots

Limite : choix d'une métrique plus adaptée aux libellés que l'on trouve dans le secteur.

4/5 Requêtage de l’API Sirene sur l’adresse

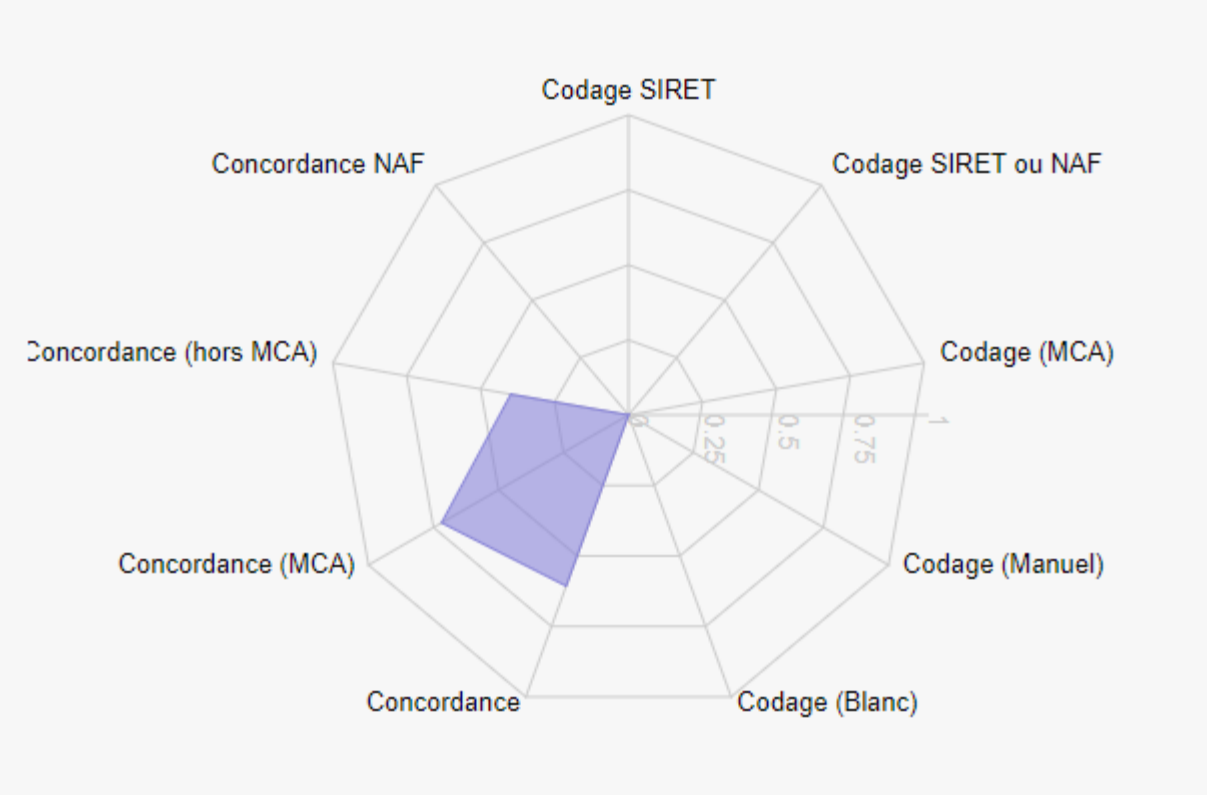
Quand la raison sociale n’est pas informative mais que l’on a une adresse.

- Entrée : adresse
- Sortie : SIRET

5/5 Appariement par les plus proches voisins

- Entrée :
  - **Quand la géolocalisation est de bonne qualité**, coordonnées géographiques du RP
  - **Quand la géolocalisation est de mauvaise qualité** (8.5 et 9), coordonnées géographiques récupérées dans la BAN à partir de l'adresse du RP (15 % de nouvelles géolocalisations)
- Recherche par plus proche voisin dans l'ensemble commune x secteur (O, P ou Q)
- **Sortie : SIRET du plus proche voisin**

Résultats


Codage SIRET
Codage SIRET ou NAF
Codage (MCA)
Codage (Manuel)
Codage (Blanc)
Concordance
Concordance (MCA)
Concordance (hors MCA)
Concordance NAF
0.25
0.5
0.75
1

BADQOPS

Santé humaine et action sociale

Administration publique

Enseignement

Merci pour votre attention