




Hackathon

“Les champs de Sirene”



Mesurer la qualité de
l'interrogation web



Combiner trois approches

Web-scraping

- Échantillon de 2000 individus aléatoires.
- Interrogation moteurs de recherche (Qwant, Google, Bing)
- Limite aux dix premiers résultats sur le site societe.com.

Calculs de distance

- Appariement avec fichier SIRUS.
- Distance “textuelle” entre 5 variables du fichier Sirus et RS de recensement.
- Distance euclidienne entre les x,y “RP” et ceux de Sirus.

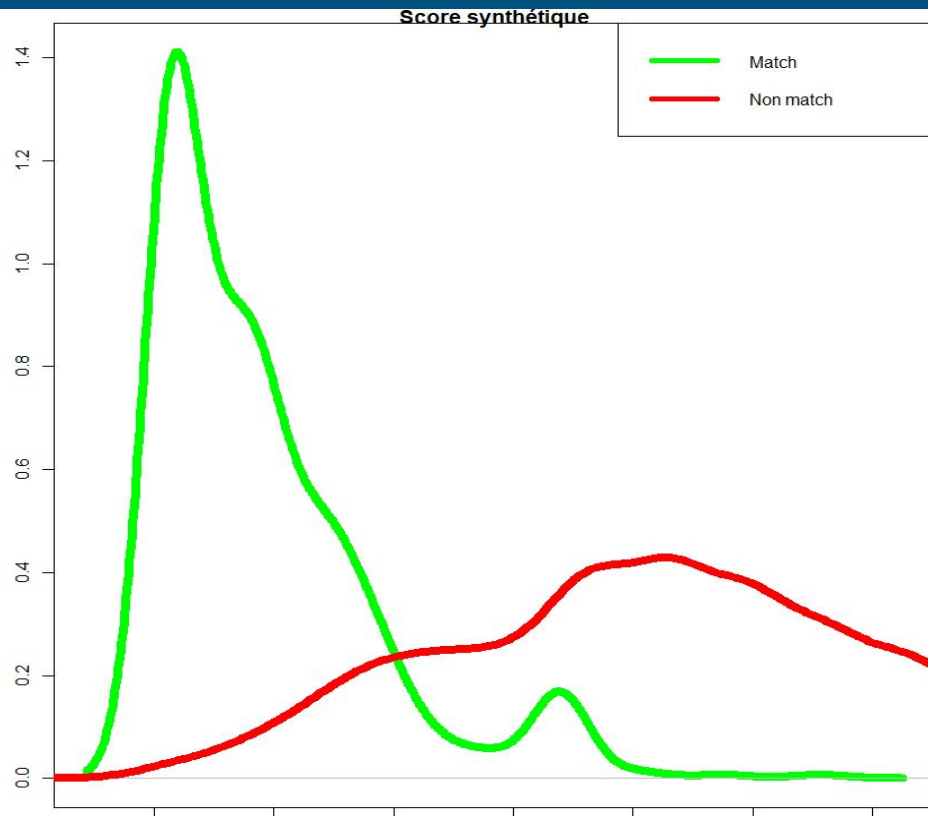
Mesure de la qualité

ACP sur les variables pour avoir un “score” synthétique de la qualité: choix entre les plusieurs SIRET.

Résultats - Concordance selon la méthode utilisée

Méthode	Concordance
1er résultat scraping	43%
Scraping + matching raison sociale	45%
Scraping + géographie	35%
Scraping + indicateur synthétique (ACP)	47%
Scraping + contrôle qualité + indicateur synthétique (ACP)	79%

Contrôle qualité, utile pour utiliser l'approche scraping



Améliorations possibles

- Passage à l'échelle
 - Questions computationnelles ?
 - Faisabilité du scraping ?
 - Questions légales
- Que faire des bulletins où on ne trouve aucun résultat sur un moteur de recherche (~45 % des BI) ?
- Améliorations possibles du contrôle qualité

Merci !

EklekGeek

