Hackathon ”Les champs de Sirene”

Restitution de l’équipe des géocodeurs

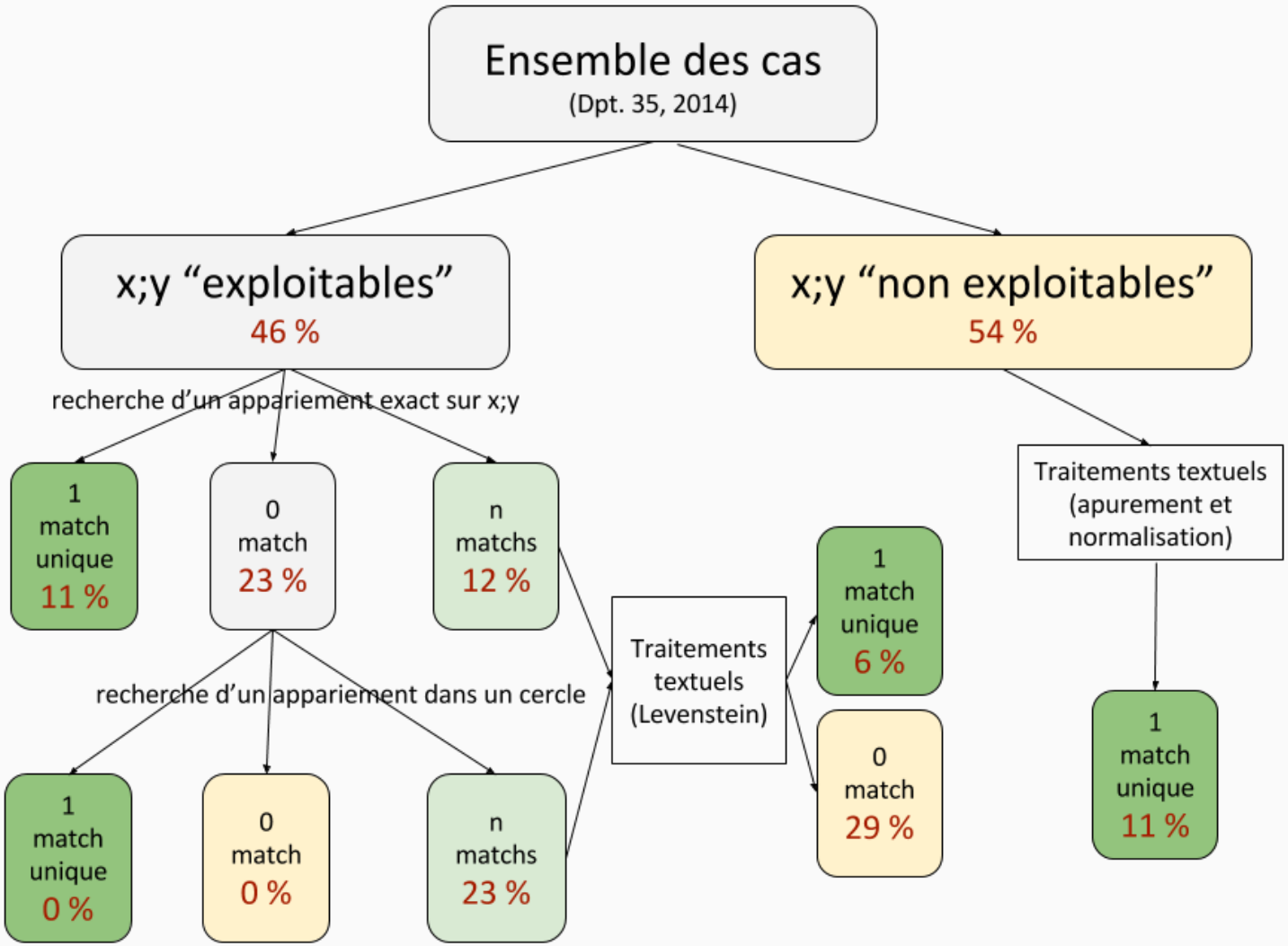
Cédric Couralet, Gaël de Peretti, Arlindo Dos Santos, Maëlle Fontaine, Juliette Fourcot, Joaquim Morais

Insee

Idées

- mobilisation en premier lieu des coordonnées géographiques
- prise en compte de la marge d'incertitude liée au processus de géocodage
- intérêt : limiter le nombre de cas à soumettre aux approches textuelles

Démarche


Ensemble des cas
(Dpt. 35, 2014)
x;y “exploitables”
46 %
x;y “non exploitables”
54 %
recherche d’un appariement exact sur x;y
1 match unique 11 %
0 match 23 %
n matchs 12 %
recherche d’un appariement dans un cercle
1 match unique 0 %
0 match 0 %
n matchs 23 %
Traitements textuels (Levenstein)
1 match unique 6 %
0 match 29 %
Traitements textuels (apurement et normalisation)
1 match unique 11 %

Démarche

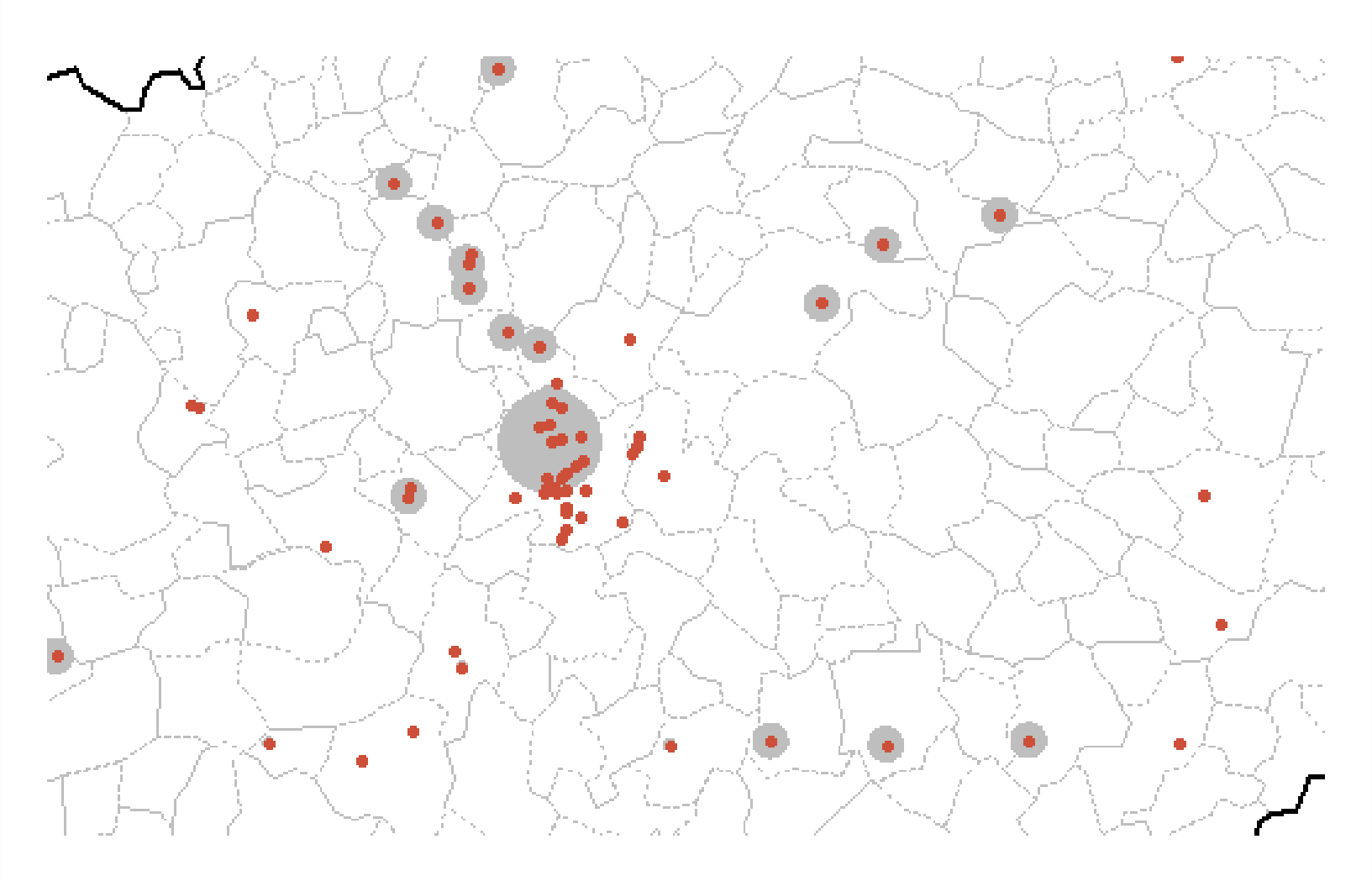


Figure 1: Exemples de cercles de rayons variables en fonction de la qualité du x;y dans Sirus

Traitements textuels

- Apurement de la raison sociale déclarée (mentions faites à la forme juridique, au nom de la commune) : permet d'augmenter les possibilités de matching sur cette variable
- Calcul de la distance de Damerau–Levenshtein entre RS_X et les champs Sirus pour accepter celle en dessous d'un certain seuil
  $\Rightarrow$ Accepte les fautes de frappes, et permet de trouver certaines entreprises individuelles (le nom de famille apparait dans l'adresse)
- Si plusieurs résultats, prise en compte du n° et libellé de voie
- Si pas de résultats, ne pas prendre en compte RS_X mais l'activité et l'adresse

Résultats

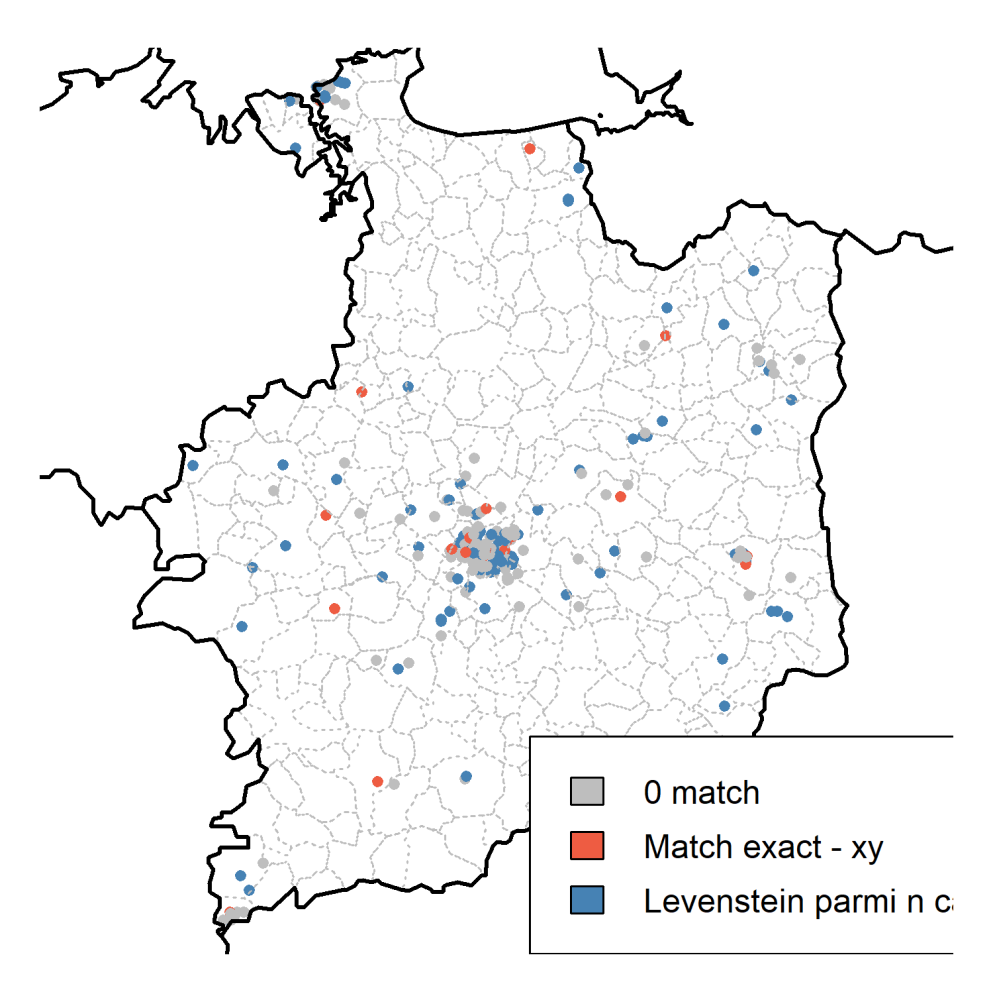


Figure 2: Exemples de cas à problème en Ille-et-Vilaine

Prolongements et limites

- Une mobilisation du x;y peut conduire à de nouveaux matchs et réduire fortement la combinatoire lors de l'identification textuelle ...
- ... mais cela repose sur une bonne qualité de celui-ci (ou une bonne caractérisation de la marge d'erreur).
- Prolongements possibles :
  - parmi les n matchs possibles, en éliminer en utilisant l'activité, la commune ...
  - recherche de x;y alternatifs pour les x;y "non-exploitables" ou ceux ne satisfaisant pas certains contrôles (sur la commune par exemple)