Développement d'un outil de sirétisation et application aux données de l'INAO

Tifenn Corre[1], Elise Maigné[2], Thomas Poméon[1], Julie Régolo[1]

(1) INRAE, US-ODR, 31320 Auzeville-Tolosane

(2) INRAE, UR MIAT, 31320 Auzeville-Tolosane

Plan

Introduction

- US-ODR : unité de service INRAE (département Ecosocio)
- Gère et développe le **SISPA** : système d'information multipartenaires (INRAE, MASA, ASP, CCMSA, INAO, ARF, MTE) sur les systèmes et politiques agricoles
- Rassemble, traite des bases de données et met à disposition des indicateurs via des tableaux, fiches ou cartes dynamiques
  - ✓ Travaux d'évaluation des programmes de développement rural (financés par le Fonds européen agricole de développement rural (FEADER))
  - ✓ Observatoire des signes d'identification de la qualité et de l'origine (SIQO)
  - ✓ Tableaux de bord de l'emploi agricole
  - ✓ Occupation du sol (RPG complété)
  - ✓ Pratiques culturales (spatialisation des ventes de produits phytosanitaires)
  - ✓ …

Introduction

- Un des objectifs du SISPA : mobiliser les différentes sources de données disponibles pour répondre à des questions de recherche

- Exemples
  - ✓ ANR FAST (Faciliter l'action publique pour sortir des pesticides) : analyse des déterminants de l'usage de produits phytosanitaires
  - ✓ RA 2020 x SIQO : demande du SSP pour enrichir les données du RA 2020 (motivation du développement de l'outil présenté)

- Identifiant commun à ces sources de données : le **numéro SIRET**

- Peut parfois être manquant, erroné ou peu fiable

- Intérêt d'avoir une **procédure générique** de vérification et d'attribution des numéros SIRET

1- Sirétisation

Sirétisation

- **Sirétisation** : opération consistant à affecter les numéros SIRET à un jeu de données à partir des champs noms (et adresses si renseignées)
- Appariement sur correspondance exacte des noms (et adresses) limité (fautes de frappe, articles, abréviations…)
- Idée générale de la méthode de sirétisation :
  - ✓ Calculer des **distances** entre les entreprises d'un jeu de données et de la base Sirene à partir des champs noms et adresses renseignés
  - ✓ Appliquer un **modèle prédictif** sur ces distances qui retourne une probabilité (score) qu'une ligne Sirene donnée est la bonne
  - ✓ Retourner les établissements Sirene les plus probables

2- Données

Base Sirene

au 1er janvier 2022

- Fichier stock des établissements (ensemble des établissements actifs et fermés dans leur état courant au répertoire)
  - ✓ 32 221 920 lignes
  - ✓ 48 variables (en particulier l'adresse et l'enseigne)
  - ✓ Une ligne par établissement (numéro SIRET)
- Fichier stock des unités légales (ensemble des entreprises actives et cessées dans leur état courant au répertoire)
  - ✓ 22 783 727 lignes
  - ✓ 33 variables (en particulier les noms, prénoms et dénominations sociales)
  - ✓ Une ligne par unité légale (numéro SIREN)
- Appariement des fichiers établissements et unités légales sur le numéro SIREN pour affecter les informations de l'entreprise à chaque établissement

Source données Sirene

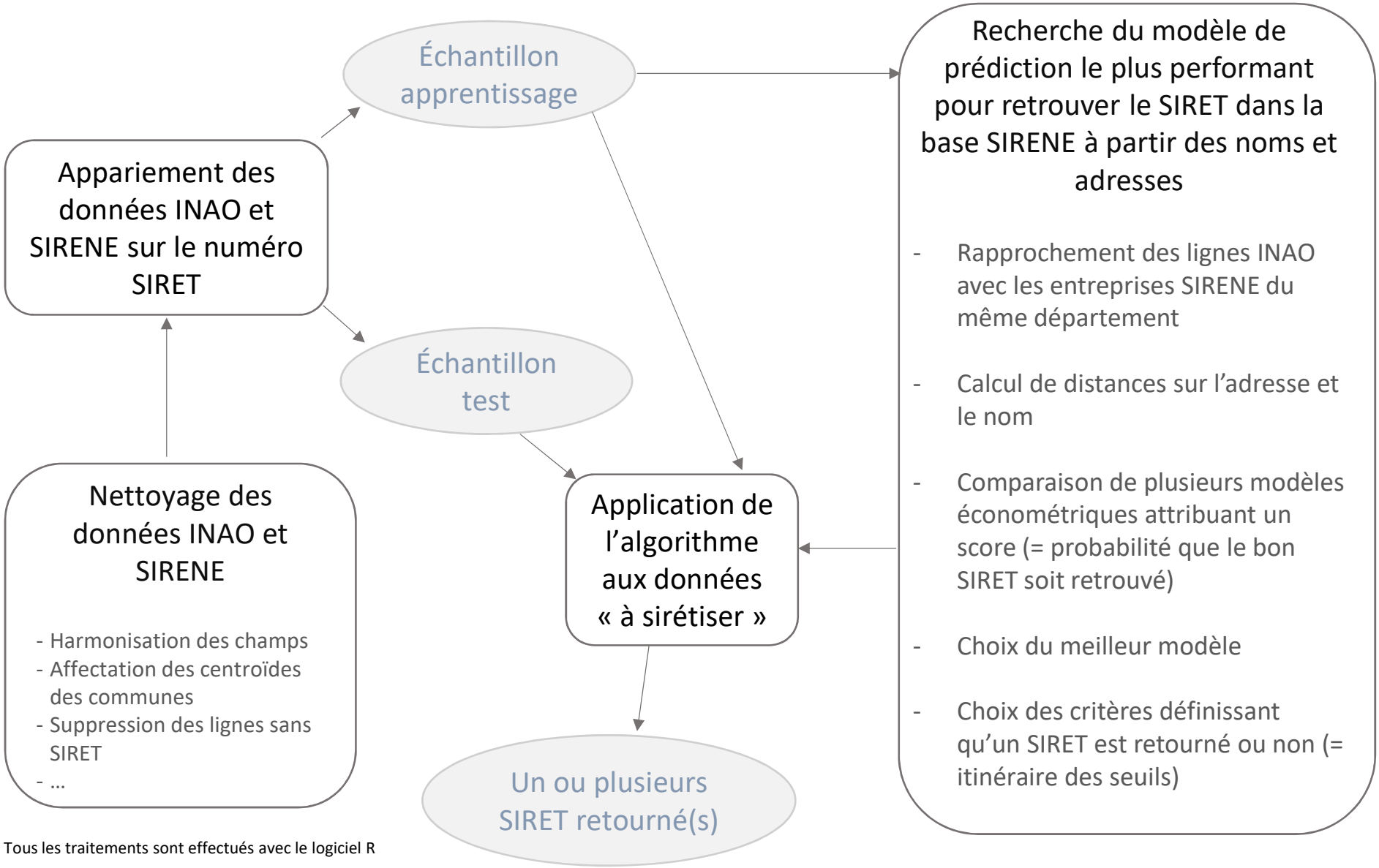
Opérateurs habilités à intervenir dans la production et la commercialisation des produits sous SIQO en 2020

- Données de l'INAO traitées par l'US-ODR
  - ✓ Partenariat dans le cadre de l'observatoire territorial des SIQO (OT-SIQO) depuis 2011
  - ✓ Liste consolidée annuelle depuis 2011 des opérateurs habilités pour des produits sous SIQO
- On garde une ligne par nom opérateur, code Insee, adresse et numéro SIRET
  - ✓ 221 464 lignes
  - ✓ **70%** de numéros SIRET renseignés valides
- Exemples de SIRET non valides observés :
  - ✓ 13 ou 15 caractères
  - ✓ Commencent par plusieurs fois le même chiffre (ex : 4444444xxxxxxxx)

Note méthodologique des données OH

2- Méthodologie

Workflow des traitements



Tous les traitements sont effectués avec le logiciel R

Apprentissage

Choix des distances

- Fichier de travail MERGE : échantillon d'apprentissage où les lignes INAO et Sirene ont mergé (= même SIRET)

- Champs utiles
  - ✓ Numéro SIRET
  - ✓ Nom
  - ✓ Adresse
  - ✓ Code commune
  - ✓ Autres variables d'aide à la décision (statut actif/fermé, code NAF, etc.)

- Calcul des distances
  - ✓ Package *stringdist* (R) : 10 méthodes différentes
  - ✓ Distances ramenées entre 0 (similarité complète) et 1 (complète dissimilarité) pour les comparer entre elles (distance / distance max)
  - ✓ Analyse multivariée des distances → 3 distances retenues

Apprentissage

Distances

❑ Distance Q-GRAM

La fonction *stringdist(a, b, method='qgram', q=q)* donne le nombre de q-gramme présents dans l'une des deux chaînes mais pas dans l'autre

On choisit de fixer **q = 1** pour la sirétisation, efficace en cas de fautes de frappe (une lettre d'écart) ou de deux chaînes avec les mêmes mots mais placés dans un ordre différent

```
stringdist('leia', 'leela', method='qgram', q=1)
[1] 3 # lettres l, e et a communes ; lettres e, i et l non communes

stringdist('leia organa', 'organa leia', method='qgram', q=1)
[1] 0 #les mêmes caractères
```

Inconvénient avec q=1, une parfaite similarité (distance nulle) peut signifier que ce sont deux anagrammes (mêmes lettres dans un ordre différent)

La distance est ramenée entre 0 et 1 par division avec la somme totale des longueurs des deux chaînes

Apprentissage
Distances

❑ Distance Jaccard

La fonction *stringdist(a, b, method='jaccard', q=q)* renvoie

$$1 - \frac{\text{Nb de q−gramme distincts communs à } a \text{ et } b}{\text{Nb total de q−gramme distincts des deux chaînes}}$$

Cette distance, comprise entre 0 et 1, permet de capter des permutations de lettres ou de syllabes

On choisit de fixer ***q* = 2** pour la sirétisation, en complément de la distance *q-gram* (*q* = 1).

```
stringdist('leia', 'leela', method='jaccard', q=2)
[1] 0.8333333 #un q-gramme commun (le), 6 q-gramme distincts totaux, dist = 1 - (1/6)

stringdist('gaec leia', 'leia gaec', method='jaccard', q=2)
[1] 0.4 #6 q-gramme communs (le, ei, ia, ga, ae, ec), 10 q-gramme distincts totaux, dist
→ = 1 - (6/10)
```

Apprentissage

Distances

❑ Distance LCS (Longest Common String)

La fonction *stringdist(a, b, method='LCS')* retourne le nombre d'opérations *D* (suppressions, insertions) nécessaires pour passer de la chaîne *a* à la chaîne *b*

A partir de cette valeur, on construit la LCS (plus longue chaîne commune)

$$LCS = \frac{|a| + |b| - D}{2}$$

|.| longueur de la chaîne de caractères

La métrique est ensuite ramenée entre 0 et 1 par la formule

$$1 - \frac{LCS}{\min(|a|, |b|)}$$

La distance est égale à 0 si l'une des deux chaînes est contenue dans l'autre

```
stringdist('gaec du leia', 'du leia', method='lcs')
[1] 5 #5 opérations pour passer de l'une à l'autre, LCS = (12 + 7 - 5)/2 = 7

fct_lcs01('gaec du leia', 'du leia')
[1] 0 #fonction de l'US-ODR retourne distance de 0 car une chaîne incluse dans l'autre
```

Apprentissage

Recherche du meilleur modèle prédictif

- Constitution de l'échantillon pour tester des modèles
  - ✓ Lignes où SIRET INAO = SIRET SIRENE (fichier de travail MERGE)
  - ✓ Lignes où SIRET INAO != SIRET SIRENE mais distances sur noms et/ou adresse faibles (candidats potentiels plausibles)

- Pour chaque ligne *i* de l'échantillon d'apprentissage :
  - ✓ On récupère les établissements du fichier SIRENE qui sont dans le **même département** que la ligne *i*
  - ✓ On calcule pour chaque établissement les différents indicateurs de distance retenus (*cf.* slides précédentes) entre le nom SIRENE et le nom INAO de la ligne *i*, et entre l'adresse SIRENE et l'adresse INAO de la ligne *i*
  - ✓ On ne garde que les établissements dont la moyenne des distances est inférieure à la 10ème plus petite moyenne des distances (valeur arbitraire)

- Suppression des lignes où les SIRET concordent mais la distance moyenne est élevée (noms et adresses n'ont rien à voir) et des lignes où les SIRET diffèrent mais la distance moyenne est nulle (mêmes noms et adresses)

Apprentissage

Recherche du meilleur modèle prédictif

- Variable à prédire : OK = 1 si SIRET INAO et SIRENE sont les mêmes, 0 sinon

- Variables explicatives (en plus des distances)
  - ✓ Nbdist0 : nombre de distances nulles parmi celles calculées
  - ✓ NBMOTSCOMMUN_NOM : nombre de mots en commun de plus de 2 lettres entre le nom INAO et le nom SIRENE
  - ✓ NBMOTSCOMMUN_ADRESSE : nombre de mots en commun de plus de 2 lettres entre l'adresse INAO et l'adresse SIRENE
  - ✓ Distance euclidienne entre les centroïdes des deux communes

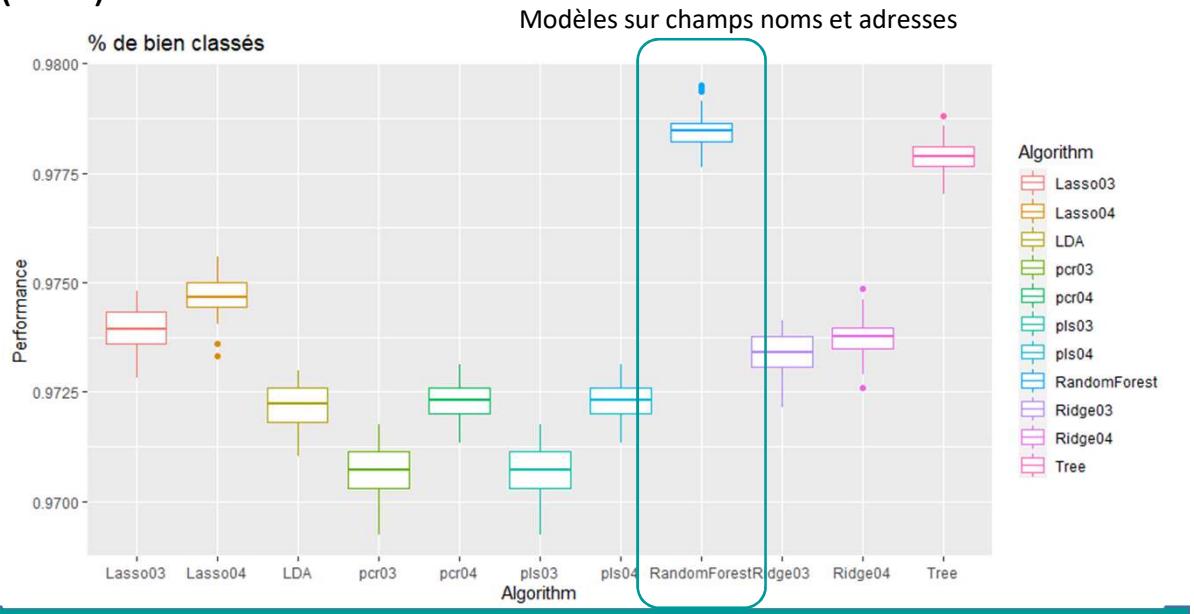
- Certains modèles renvoient des probabilités, il faut déterminer un seuil à partir duquel on considère que la prédiction est 1 (0,5 par défaut)

- 2 modèles :
  - ✓ sur les champs nom et adresse
  - ✓ sur les noms uniquement (respectivement 10% et 2% de lignes sans adresse dans les données INAO et SIRENE)

Apprentissage

Comparaison de modèles en validation croisée

- But : identifier la méthode qui minimise les erreurs de prédiction
  - ✓ Régression pénalisée Ridge
  - ✓ Régression pénalisée Lasso
  - ✓ Régression sur composantes principales (pcr)
  - ✓ Régression sur moindres carrés partiels (pls)
  - ✓ Analyse discriminante linéaire (LDA)
  - ✓ Arbre de classification (tree)
  - ✓ **Forêts aléatoires**



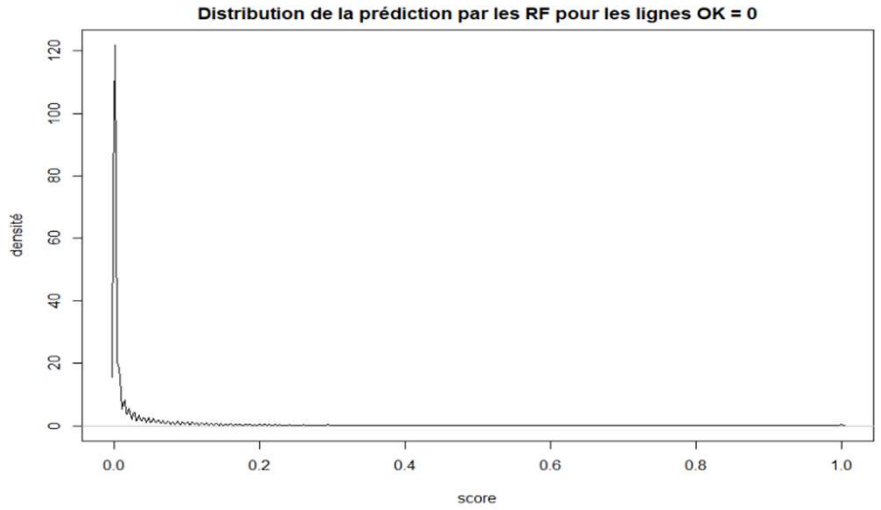
Apprentissage

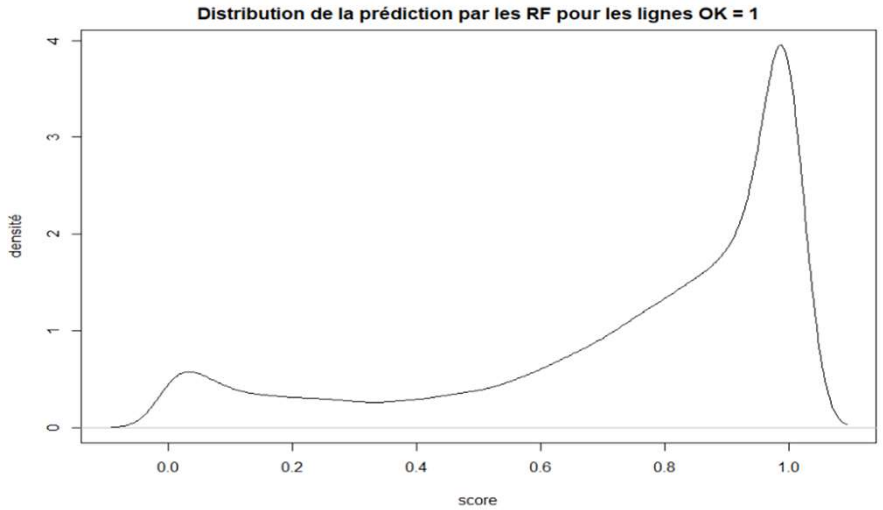
Forêts aléatoires, itinéraire des seuils

- 2 modèles lancés sur l'ensemble du fichier de travail
  - ✓ Sur les champs noms et adresses
  - ✓ Sur les champs noms uniquement

- Étude des scores

  Exemple sur le modèle noms et adresses :





- Construction d'un itinéraire des seuils pour retourner les résultats

Apprentissage

Forêts aléatoires, itinéraire des seuils

- Étape 1 : score modèle nom et adresse > 0.8
- Étape 2 : score modèle nom et adresse > 0.6 **et** score modèle nom > 0.8
- Étape 3 : score modèle nom et adresse > 0.6 **ou** score modèle nom > 0.8
- Étape 4 : score modèle nom et adresse parmi les trois meilleurs **et** > 0.2 **et** score modèle nom parmi les trois meilleurs **et** > 0.2
- Étape 5 : (score modèle nom et adresse parmi les trois meilleurs **et** > 0.2) **ou** (score modèle nom parmi les trois meilleurs **et** > 0.2)
- Étape 6 : meilleur score non nul du modèle nom et adresse
- Étape 7 : meilleur score non nul du modèle nom

Application de la méthode aux échantillons d'apprentissage et test

- Pour chaque ligne des fichiers
  - ✓ Identification des entreprises SIRENE du même département
  - ✓ Calcul des distances retenues sur les noms et adresse
  - ✓ On garde les lignes où la distance moyenne < 10 plus petites distances moyennes
  - ✓ Calcul des nombres de mots en commun sur les noms et adresses
  - ✓ Application des modèles RF nom + adresse et nom uniquement → scores
  - ✓ Itinéraire des seuils

4- Résultats

Résultats

Nombre de résultats retournés par individu

Nb résultats retournés	% individus		% SIRET retrouvés	
	Apprentissage	Test	Apprentissage	Test
0	6.8	6.6	NA	NA
1	82.2	77.2	**80.4**	**77.3**
2	8.8	12.5	**76.4**	**75.3**
3	1.4	2.0	**69.1**	**69.5**
4	0.4	1.0	**59.5**	**72.6**
>=5	0.3	0.7	**40.1**	**63.1**
Tout	**100.0**	**100.0**	**75.8**	**73.1**

% SIRET retrouvés : exclusion des individus dont le SIRET n'existe pas dans la base Sirene et des individus retrouvés mais avec une distance moyenne élevée (les noms et/adresse n'ont rien à voir)

Résultats

Étapes retournées par individu

Etape retour	% individus		% SIRET retrouvés	
	Apprentissage	Test	Apprentissage	Test
0	6.8	6.6	NA	NA
1	45.4	40.8	**93.2**	**88.3**
2	7.1	10.5	**84.4**	**84.8**
3	14.1	13.8	**84.7**	**87.0**
4	4.9	6.1	**72.6**	**76.4**
5	7.1	6.4	**76.0**	**67.2**
6	8.8	9.9	19.5	25.9
7	5.8	5.9	12.7	13.6
Tout	**100.0**	**100.0**	**75.8**	**73.1**

% SIRET retrouvés : exclusion des individus dont le SIRET n'existe pas dans la base Sirene et des individus retrouvés mais avec une distance moyenne élevée (les noms et/adresse n'ont rien à voir)

Résultats

Nombre de résultats et étapes retournés par individu

Nb résultats retournés	Etape retour	% individus		% SIRET retrouvés	
		Apprentissage	Test	Apprentissage	Test
0	**NA**	6.8	6.6	NA	NA
1	**1-5**	69.7	64.5	88.7	85.4
1	**6-7**	12.6	12.7	17.4	22.1
2	**1-5**	7.3	10.2	85.9	84.6
2	**6-7**	1.5	2.2	13.7	19.2
3	**1-5**	1.1	1.6	82.5	79.7
3	**6-7**	0.3	0.4	11.8	16.8
…	**…**	**…**	…	**…**	…
Tout	**Tout**	**100.0**	100.0	**75.8**	73.1

% SIRET retrouvés : exclusion des individus dont le SIRET n'existe pas dans la base Sirene et des individus retrouvés mais avec une distance moyenne élevée (les noms et/adresse n'ont rien à voir)

5- Exemple d’application

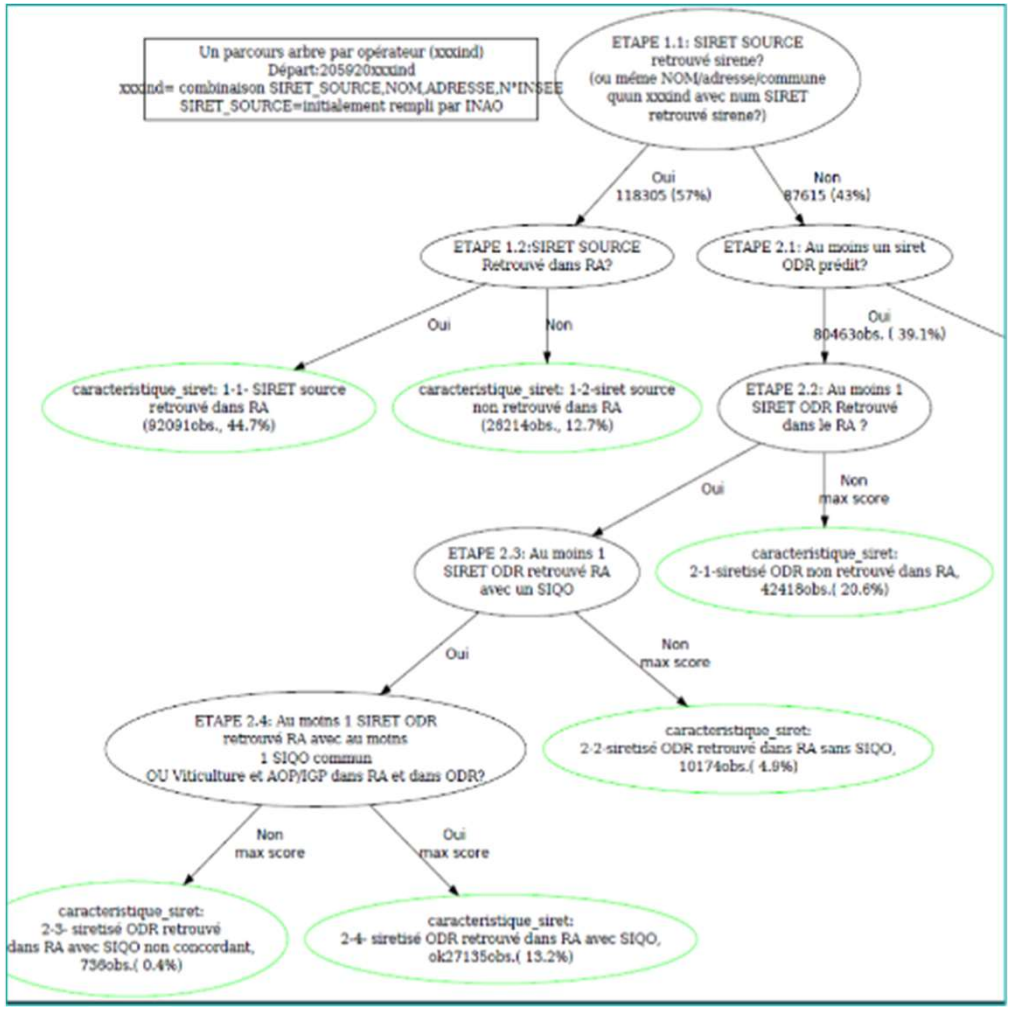
Appariement INAO avec le RA 2020

- Sirétisation du fichier INAO 2020 en partenariat avec le SSP sur tous les opérateurs habilités 2020

- Processus de sirétisation sur 207 403 opérateurs
  - ✓ Conservation des SIRET existants et actifs en 2020 : par fusion avec la base SIRENE (y compris non diffusibles) : **57%**
  - ✓ Deux méthodes de sirétisation pour compléter les manquants :
    - SSP : appariements exacts successifs (nom, adresse...) avec SIRENE + prestataire externe (manuel)
    - ODR : distances nom et adresse, Random Forest, itinéraire des seuils. Niveau communal et départemental
  - ✓ Combinaisons des résultats et appariement avec le RA 2020

Appariement INAO avec le RA 2020

- Application de la méthode ODR avec les SIRET actifs en 2020 uniquement
- En cas de résultats multi SIRET : filtres successifs :
  - Dans le RA ?
  - Déclare un SIQO dans le RA ?
  - Même SIQO ?
  - APET agricole ou agro-alimentaire ?
- Résultats : 97% avec siret. Deux rangs de précision : 93% / 33%



Appariement INAO avec le RA 2020

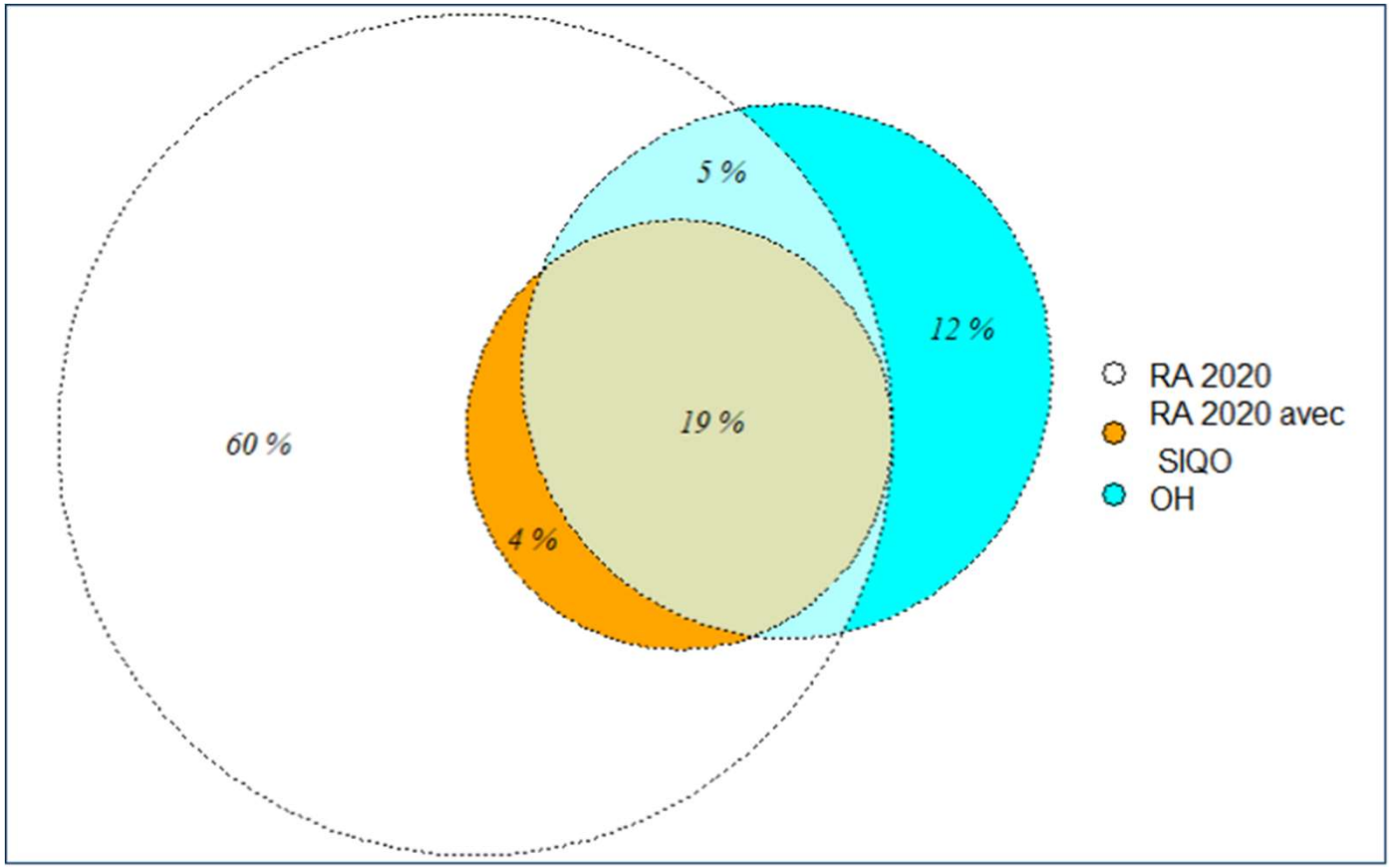
- Comparaison avec la siretisation du SSP : 15% de SIRET différents de l'ODR

- Accord sur une méthode de sélection entre SIRET SSP et ODR :
  - ✓ On privilégie le SIRET qui est dans le RA2020
  - ✓ On privilégie le successeur (utilisation base INSEE prédécesseur/successeur)
  - ✓ On privilégie le SIRET qui est retrouvé dans les Entreprises Viti-Vinicoles
  - ✓ SIRET ODR si fiabilité 93%, sinon SIRET du SSP

- Résultats, sur 207 403 opérateurs :
  - ✓ **57% seulement avaient un siret actif en 2020** et valide (conservés),
  - ✓ 40% ont été siretisés (10% issus du SSP, 30% de l'ODR)
  - ✓ 3% restent sans siret

Appariement INAO avec le RA 2020

Résultats de l'appariement (% SIRET)



Appariement Imparfait. Problèmes de SIRET ? Décalage dans le temps ? Surtout filière viticole.

Appariement INAO avec le RA 2020

Dimension	Nombre de répondants	% déclarant un SIQO dans le RA	% retrouvés dans OH2020	% déclarant un SIQO et retrouvés dans OH2020
1-micros	118 401	14	14	10
2-petites	109 223	24	24	19
3-moyennes	102 156	33	34	28
4-grandes	77 985	38	41	34
TOTAL	**407 766**	**26**	**27**	**21**

Conclusion

Pistes d'amélioration

- Élargir à la France entière si résultats insatisfaisants ? (problème de mémoire de calcul)
- Être plus sélectif sur l'échantillon d'apprentissage
- Intégrer l'appariement avec les données non diffusibles dans l'outil générique pour identifier les lignes qu'on ne pourra pas retrouver dans la base Sirene (pour vérification de SIRET existants seulement)
- Utiliser un modèle de forêts aléatoires avec une distribution de probabilité enflée en zéro pour indiquer qu'il y a beaucoup plus de 0 que de 1
- Appliquer la méthode à de nouveaux jeux de données (données bio)

Merci de votre attention

Pour plus d'information :
odr.inrae.fr
tifenn.corre@inrae.fr
julie.regolo@inrae.fr