Méthode de sirétisation ODR - 2023

Tifenn Corre, Thomas Pomeon, Julie Regolo

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

INRAE
la science pour la vie, l'humain, la terre

Observatoire du Developpement Rural
ODR

Méthode de sirétisation

Développée à partir du fichier des opérateurs habilités 2020 (INAO)

Notice des traitements de l'US-ODR

Mars 2023

Version	Date de mise à jour	Modification
v.1	2023-03-10	–

Table des matières

Table des figures

Liste des tableaux

1. Introduction et motivations

Le numéro SIRET (système d'identification du répertoire des établissements) est une variable clé dans les données individuelles mises à disposition pour l'US-ODR (unité de service Observatoire du Développement Rural, INRAE) par les différents fournisseurs (Institut national de l'origine et de la qualité (INAO), Agence de services et de paiement (ASP), Caisse centrale de la mutualité sociale agricole (CCMSA)). En effet, c'est aujourd'hui le seul identifiant d'entreprise commun à toutes les bases de données. Cependant il n'est pas toujours bien renseigné. Dans le but d'enrichir, à terme, les données en faisant des appariements individuels, nous travaillons en amont sur le numéro SIRET afin de compléter les valeurs manquantes. Nous appelons cette opération la **sirétisation**.

La méthode présentée dans ce document a été développée à partir des données des opérateurs habilités de l'INAO en 2020 (OH_2020). L'enrichissement des données de l'INAO pourra se faire avec les données de la CCMSA ou du recensement agricole (RA) par exemple, sous réserve de possibilité technique (amélioration du remplissage des SIRET) et juridique (accord des fournisseurs, respects des règles de la CNIL, déclaration des traitements).

De tels appariements permettront de mener des études sur le suivi des exploitations produisant des SIQO (signes officiels de la qualité) ou encore des études comparatives avec des exploitations sans SIQO (agriculture biologique / conventionnelle).

La carte ci-dessous présente la situation géographique des opérateurs habilités en 2020 pour lesquels le numéro SIRET n'est pas (ou mal) renseigné. Dans certains départements, le taux de non renseignement dépasse les 50%.

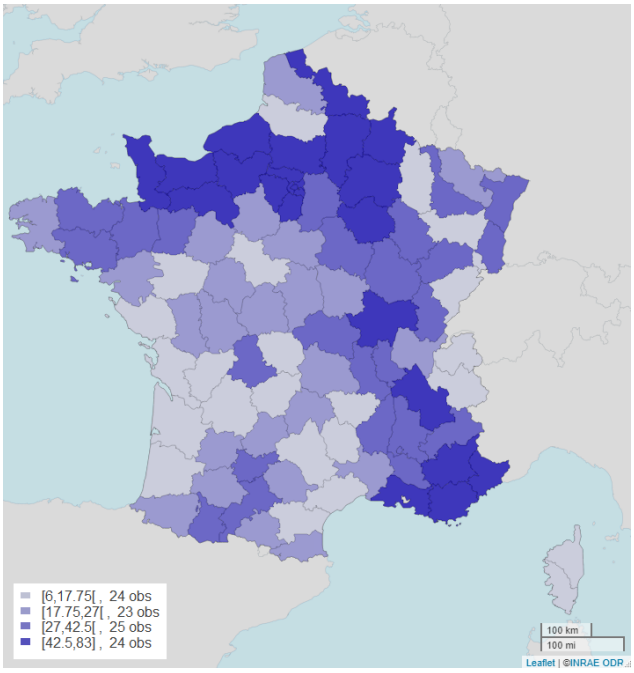


Figure 1 : Pourcentage d'opérateurs habilités où le numéro SIRET n'est pas renseigné

La méthode de sirétisation a vocation à être appliquée à d'autres jeux de données. Elle est ainsi conçue dans un esprit générique et non spécifique aux données des opérateurs habilités de l'INAO. Elle est développée avec le logiciel R[1].

1. R Core Team (2022). R : A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

2. Données

En vertu de la loi pour une République numérique, les données Sirene d'identité des entreprises et des établissements sont ouvertes en open data depuis début janvier 2017. Cette diffusion publique permet désormais de rechercher des entreprises dans la base de données des entreprises de France et de venir enrichir des données (ici celles des opérateurs habilités de l'INAO).

Dans ce cas précis, on ne s'intéresse qu'au rapatriement du numéro SIRET dans les données des opérateurs habilités de l'INAO lorsqu'il est manquant ou mal renseigné.

Après avoir étudié le contenu respectif des deux bases de données (Sirene au 1er janvier 2022 et opérateurs habilités en 2020), les champs communs apparaissant utiles pour la sirétisation sont les suivants :

- la raison sociale ;
- les noms et prénoms des dirigeants ;
- la commune du siège ;
- l'adresse principale et/ou secondaire de l'établissement.

Dans chaque jeu de données nous gardons également quelques variables supplémentaires qui aideront à la vérification des résultats comme le code d'activité NAF (Nomenclature d'Activités Françaises) par exemple.

2.1. Données de la base Sirene

Les données Sirene se présentent sous cinq fichiers stock mensuels :

- le fichier stock des entreprises (ensemble des entreprises actives et cessées dans leur état courant au répertoire) ;
- le fichier stock des valeurs historisées des entreprises (pour toutes les entreprises, ensemble des valeurs de certaines variables historisées dans le répertoire Sirene) ;
- le fichier stock des établissements (ensemble des établissements actifs et fermés dans leur état courant au répertoire) ;
- le fichier stock des valeurs historisées des établissements (pour tous les établissements, ensemble des valeurs de certaines variables historisées dans le répertoire Sirene) ;
- le fichier stock des liens de succession des établissements (prédécesseurs et successeurs des établissements).

Pour ce travail, nous utilisons les fichiers stock des établissements et des entreprises au 1er janvier 2022. Nous conservons toutes les lignes du fichier stock des établissements, pour ne pas se restreindre a priori sur les candidats potentiels, auxquelles nous rattachons les informations concernant leur unité légale disponibles dans le fichier stock des entreprises. En effet, certaines informations clés comme les noms des responsables ne sont disponibles qu'au niveau de l'entreprise et nécessitent d'être rattachées aux établissements correspondants. Le fichier des établissements étant lourd, il est découpé par département.

2.1.1. Fichier stock des établissements

Au 1er janvier 2022 :

- 32 221 920 lignes ;
- 48 variables ;
- une ligne par établissement (numéro SIRET).

Tableau 1 : Liste des champs du fichier stock des établissements

Nom	Libellé	Longueur	Type	Ordre
siren	Numéro SIREN	9	Texte	1
nic	Numéro interne de classement de l'établissement	5	Texte	2
siret	Numéro SIRET	14	Texte	3
statutDiffusionEtablissement	Statut de diffusion de l'établissement	1	Liste de codes	4
dateCreationEtablissement	Date de création de l'établissement	10	Date	5
trancheEffectifsEtablissement	Tranche d'effectif salarié de l'établissement	2	Liste de codes	6
anneeEffectifsEtablissement	Année de validité de la tranche d'effectif salarié de l'établissement	4	Date	7
activitePrincipaleRegistreMetiersEtablissement	Activité exercée par l'artisan inscrit au registre des métiers	6	Liste de codes	8
dateDernierTraitementEtablissement	Date de dernier traitement de l'établissement dans le répertoire Sirene	19	Date	9
etablissementSiege	Qualité de siège ou non de l'établissement	5	Texte	10
nombrePeriodesEtablissement	Nombre de périodes de l'établissement	2	Numérique	11
complementAdresseEtablissement	Complément d'adresse	38	Texte	12
numeroVoieEtablissement	Numéro de voie	4	Texte	13
indiceRepetitionEtablissement	Indice de répétition dans la voie	1	Texte	14
typeVoieEtablissement	Type de voie	4	Liste de codes	15
libelleVoieEtablissement	Libellé de voie	100	Texte	16

Nom	Libellé	Longueur	Type	Ordre
codePostalEtablissement	Code postal	5	Texte	17
libelleCommuneEtablissement	Libellé de la commune	100	Texte	18
libelleCommuneEtrangerEtablissement	Libellé de la commune pour un établissement situé à l'étranger	100	Texte	19
distributionSpecialeEtablissement	Distribution spéciale de l'établissement	26	Texte	20
codeCommuneEtablissement	Code commune de l'établissement	5	Liste de codes	21
codeCedexEtablissement	Code cedex	9	texte	22
libelleCedexEtablissement	Libellé du code cedex	100	Texte	23
codePaysEtrangerEtablissement	Code pays pour un établissement situé à l'étranger	5	Liste de codes	24
libellePaysEtrangerEtablissement	Libellé du pays pour un établissement situé à l'étranger	100	Texte	25
complementAdresse2Etablissement	Complément d'adresse secondaire	38	Texte	26
numeroVoie2Etablissementt	Numéro de la voie de l'adresse secondaire	4	Texte	27
indiceRepetition2Etablissement	Indice de répétition dans la voie pour l'adresse secondaire	1	Texte	28
typeVoie2Etablissement	Type de voie de l'adresse secondaire	4	Liste de codes	29
libelleVoie2Etablissement	Libellé de voie de l'adresse secondaire	100	Texte	30
codePostal2Etablissement	Code postal de l'adresse secondaire	5	Texte	31
libelleCommune2Etablissement	Libellé de la commune de l'adresse secondaire	100	Texte	32
libelleCommuneEtranger2Etablissement	Libellé de la commune de l'adresse secondaire pour un établissement situé à l'étranger	100	Texte	33
distributionSpeciale2Etablissement	Distribution spéciale de l'adresse secondaire de l'établissement	26	Texte	34
codeCommune2Etablissement	Code commune de l'adresse secondaire	5	Liste de codes	35

Nom	Libellé	Longueur	Type	Ordre
codeCedex2Etablissement	Code cedex de l'adresse secondaire	9	Texte	36
libelleCedex2Etablissement	Libellé du code cedex de l'adresse secondaire	100	Texte	37
codePaysEtranger2Etablissement	Code pays de l'adresse secondaire pour un établissement situé à l'étranger	5	Liste de codes	38
libellePaysEtranger2Etablissement	Libellé du pays de l'adresse secondaire pour un établissement situé à l'étranger	100	Texte	39
dateDebut	Date de début d'une période d'historique d'un établissement	10	Date	40
etatAdministratifEtablissement	État administratif de l'établissement	1	Liste de codes	41
enseigne1Etablissement	Première ligne d'enseigne de l'établissement	50	Texte	42
enseigne2Etablissement	Deuxième ligne d'enseigne de l'établissement	50	Texte	43
enseigne3Etablissement	Troisième ligne d'enseigne de l'établissement	50	Texte	44
denominationUsuelleEtablissement	Dénomination usuelle de l'établissement	100	texte	45
activitePrincipaleEtablissement	Activité principale de l'établissement pendant la période	6	Liste de codes	46
nomenclatureActivitePrincipaleEtablissement	Nomenclature d'activité de la variable activitePrincipaleEtablissement	8	Liste de codes	47
caractereEmployeurEtablissement	Caractère employeur de l'établissement	1	Liste de codes	48

Les champs retenus pour la sirétisation sont en fond gris

2.1.2. Fichier stock des entreprises

Au 1er janvier 2022 :

- 22 783 727 ;
- 33 variables ;
- une ligne par unité légale (numéro SIREN).

Tableau 2 : Liste des champs du fichier stock des unités légales

Nom	Libellé	Longueur	Type	Ordre
siren	Numéro SIREN	9	Texte	1
statutDiffusionUniteLegale	Statut de diffusion de l'unité légale	1	Liste de codes	2
unitePurgeeUniteLegale	Unité légale purgée	5	Texte	3
dateCreationUniteLegale	Date de création de l'unité légale	10	Date	4
sigleUniteLegale	Sigle de l'unité légale	20	Texte	5
sexeUniteLegale	Caractère féminin ou masculin de la personne physique	1	Liste de codes	6
prenom1UniteLegale	Premier prénom déclaré pour un personne physique	20	Texte	7
prenom2UniteLegale	Deuxième prénom déclaré pour un personne physique	20	Texte	8
prenom3UniteLegale	Troisième prénom déclaré pour un personne physique	20	Texte	9
prenom4UniteLegale	Quatrième prénom déclaré pour un personne physique	20	Texte	10
prenomUsuelUniteLegale	Prénom usuel de la personne physique	20	Texte	11
pseudonymeUniteLegale	Pseudonyme de la personne physique	100	Texte	12
identifiantAssociationUniteLegale	Numéro au Répertoire National des Associations	10	Texte	13
trancheEffectifsUniteLegale	Tranche d'effectif salarié de l'unité légale	2	Liste de codes	14
anneeEffectifsUniteLegale	Année de validité de la tranche d'effectif salarié de l'unité légale	4	Date	15
dateDernierTraitementUniteLegale	Date du dernier traitement de l'unité légale dans le répertoire Sirene	19	Date	16
nombrePeriodesUniteLegale	Nombre de périodes de l'unité légale	2	Numérique	17
categorieEntreprise	Catégorie à laquelle appartient l'entreprise	3	Liste de codes	18
anneeCategorieEntreprise	Année de validité de la catégorie d'entreprise	4	Date	19

Nom	Libellé	Longueur	Type	Ordre
dateDebut	Date de début d'une période d'historique d'une unité légale	10	Date	20
etatAdministratifUniteLegale	État administratif de l'unité légale	1	Liste de codes	21
nomUniteLegale	Nom de naissance de la personnes physique	100	texte	22
nomUsageUniteLegale	Nom d'usage de la personne physique	100	Texte	23
denominationUniteLegale	Dénomination de l'unité légale	120	Texte	24
denominationUsuelle1UniteLegale	Dénomination usuelle de l'unité légale	70	Texte	25
denominationUsuelle2UniteLegale	Dénomination usuelle de l'unité légale - deuxième champ	70	Texte	26
denominationUsuelle3UniteLegale	Dénomination usuelle de l'unité légale - troisième champ	70	Texte	27
categorieJuridiqueUniteLegale	Catégorie juridique de l'unité légale	4	Liste de codes	28
activitePrincipaleUniteLegale	Activité principale de l'unité légale	6	Liste de codes	29
nomenclatureActivitePrincipaleUniteLegale	Nomenclature d'activité de la variable activitePrincipaleUniteLegale	8	Liste de codes	30
nicSiegeUniteLegale	Numéro interne de classement (Nic) de l'unité légale	5	Texte	31
economieSocialeSolidaireUniteLegale	Appartenance au champ de l'économie sociale et solidaire	1	Liste de codes	32
caractereEmployeurUniteLegale	Caractère employeur de l'unité légale	1	Liste de codes	33

Les champs retenus pour la sirétisation sont en fond gris

2.2. Données des opérateurs habilités

Fichier des opérateurs habilités en 2020 :

- 837 863 lignes ;
- 37 variables ;
- une ligne par produit et activité.

Tableau 3 : Liste des champs retenus du fichier des opérateurs habilités

Nom	Libellé	Longueur	Type
libelle_OC	Libellé de l'organisme certificateur (OC)	200	texte
libelle_ODG	Libellé de l'organisme de défense et de gestion (ODG)	200	texte
ref_CDC	Référence du cahier des charges (si produit IGP ou Label rouge)	50	texte
ref_PC	Référence du plan de contrôle (PC)	100	texte
libelle_PC	Libellé du PC (non rempli)	100	texte
id_filiere_classe_inao	Identifiant de la filière classe INAO	11	texte
libelle_produit	Libellé du produit SIQO (selon le référentiel INAO)	300	texte
commune	Commune de l'opérateur	100	texte
code_insee	Code commune	6	Liste de codes
nom_operateur	Nom et/ou prénom et/ou dénomination sociale de l'opérateur	200	Texte
id_inao_odr	Identifiant unique de l'opérateur créé par l'US-ODR	40	Texte
statut_operateur	Statut d'habilitation de l'opérateur d'après les données OC	100	Texte
statut_simpl	Statut de l'opérateur simplifié	12	Texte
siret	Numéro SIRET de l'opérateur	14	Texte
no_cvi	Numéro de casier viticole informatisé (CVI)	20	Texte
no_cheptel	Numéro de cheptel	20	Texte
libelle_oc_activite	Libellé de l'activité d'origine (données OC)	200	Texte
libelle_odr_activite	Libellé de l'activité homogénéisé par l'US-ODR	100	Texte
code_activite_ref	Code d'activité homogénéisé par l'US-ODR	20	Texte
date_OC_extraction	Date d'extraction de la donnée OC	33	Texte
date_insertion	Date d'insertion de la donnée dans l'Observatoire Territorial des SIQO	10	Texte
observation	Observations éventuelles	300	Texte
IDP	Identifiant du produit (référentiel INAO)	11	Entier

Nom	Libellé	Longueur	Type
id_appellation	Identifiant de l'appellation (référentiel INAO)	12	Texte
signe_fr	Type de signe (classification française)	10	Texte
signe_ue	Type de signe (classification européenne)	10	Texte
filiere_classe_inao	Libellé de la filière classe INAO	100	Texte
date_decision	Date de la décision (si changement d'habilitation)	40	Texte
adresse	Adresse précise de l'opérateur	512	Texte
libelle_odr_activite_cumul	Identifiant du cumul des activités des opérateurs par produit	500	Texte
id_odr_activite_cumul	Code du cumul des activités des opérateurs par produit	3	Texte

Les champs retenus pour la sirétisation sont en fond gris

Sur les 837 863 lignes, 92.5% (774 831) sont renseignées comme habilitées [2].

2.3. Données complémentaires

Données Geofla®- communes 2016 : ce fichier contient une ligne par commune de France. Il permet de rapatrier les coordonnées géographiques du centroïde des communes, afin notamment de calculer des distances entre elles [3].

Tableau 4 : Liste des champs du fichier Geofla®- communes 2016

Nom	Libellé	Type
ID_GEOFLA	Identifiant	Texte
CODE_COM	Code géographique de la commune	Numérique
INSEE_COM	Numéro INSEE de la commune	Numérique
NOM_COM	Nom de la commune	Texte
STATUT	Statut administratif	Texte
X_CHF_LIEU	Abscisse du chef-lieu	Numérique
Y_CHF_LIEU	Ordonnée du chef-lieu	Numérique

2. Pour plus d'informations sur ces données, consulter la présentation de la base de données des opérateurs habilités à intervenir dans la production et la commercialisation des produits sous signe d'identification de la qualité et de l'origine (SIQO) : https://hal.inrae.fr/hal-03264972/document

3. Un premier travail de sirétisation a été effectué sur des données de l'Agence Bio 2017, d'où l'utilisation des données communales 2016 conservées ici. Environ 1500 lignes du fichier OH_2020 ne trouvent pas de correspondance dans ce référentiel communal mais il s'agit à 85% d'arrondissements de grandes villes pour lesquels on affecte le centroïde du premier arrondissement. Pour les prochaines opérations de sirétisation, il sera opportun d'adapter les coordonnées géographiques des communes aux données sirétisées.

Nom	Libellé	Type
X_CENTROID	Abscisse du centroïde	Numérique
Y_CENTROID	Ordonnée du centroïde	Numérique
Z_MOYEN	Altitude moyenne	Numérique
SUPERFICIE	Superficie	Numérique
POPULATION	Population	Numérique
CODE_CANT	Code géographique du canton	Numérique
CODE_ARR	Code géographique de l'arrondissement	Numérique
CODE_DEPT	Code géographique du département	Numérique
NOM_DEPT	Nom du département	Texte
CODE_REG	Code géographique de la région	Numérique
NOM_REG	Nom de la région	Texte

Les champs retenus pour la sirétisation sont en fond gris

3. Préparation des données

3.1. Fonctions d'uniformisation des champs texte

Dans un premier temps, les données Sirene et INAO sont nettoyées et uniformisées :

- remplacement des caractères spéciaux et caractères accentués ;
- champs mis en majuscules ;
- suppression des espaces en double ;
- remplacement des valeurs *NA* par des champs vides.

3.2. Données Sirene

Le fichier stock des établissements est découpé en fichiers départementaux selon le code commune de l'adresse principale pour faciliter les traitements avec le logiciel *R* :

- 97 fichiers départementaux de la France métropolitaine ;
- 2 fichiers DROM-COM ;
- 1 fichier des entreprises sans code commune renseigné ou situées à l'étranger.

Le même découpage est effectué sur le code commune de l'adresse secondaire quand elle existe.

Pour chaque fichier départemental créé :

- les champs de distribution spéciale (boîtes postales, cedex) sont uniformisés ;

- l'année de création de l'établissement est extraite de la date de création de l'établissement (4 premiers caractères), la date de création est ensuite supprimée ;
- les informations relative à l'entreprise de chaque établissement sont rapatriées du fichier stock des unités légales (appariement sur le numéro SIREN) ;
- les numéros SIREN et SIRET sont mis au bon format (ajout de « 0 » en début de champ si la longueur totale est inférieure à 9 et 14 caractères, respectivement) ;
- les coordonnées géographiques des codes communes des adresses principales et secondaires sont rapatriées du fichier Geofla (appariement sur le code commune).

  Pour les codes communes du fichier Sirene non retrouvés dans les données Geofla, on applique des corrections manuelles (recherche du code Insee Geofla existant le plus pertinent [4]).
- les champs de complément d'adresse sont uniformisés ;
- les 19 champs relatifs aux dénominations des établissements [5] sont concaténés dans un seul champ NOM_REGROUPE (séparateur « @ »), puis les 6 premières valeurs uniques sont extraites dans les champs NOM_REGROUPEi $(1 \leq i \leq 6)$;
- le terme « LIEU DIT » est supprimé des champs de complément d'adresse s'il débute le champ ;
- les champs de type de voie sont uniformisés (exemples RTE = ROUTE, IMP = IMPASSE, etc.) ;
- les champs numeroVoieEtablissement, indiceRepetitionEtablissement, typeVoieEtablissement, libelleVoieEtablissement pour l'adresse principale sont concaténés dans un seul champ ADRESSETEMP, de même pour l'adresse secondaire (ADRESSETEMP2), puis supprimés ;
- les champs relatifs aux adresses des établissements [6] sont concaténés dans un seul champ ADRESSE_REGROUPE (séparateur « @ »), puis les 3 premières valeurs uniques sont extraites dans les champs ADRESSE_REGROUPEi $(1 \leq i \leq 3)$

Le choix de conserver les 6 et 3 premières valeurs uniques de NOM_REGROUPE et ADRESSE_REGROUPE résulte d'une analyse complémentaire de dénombrement de champs non vides : moins de 0.1% des lignes ont plus de 6 valeurs distinctes de NOM ou 3 valeurs distinctes d'ADRESSE).

Le dessin d'enregistrement final pour chaque fichier départemental Sirene nettoyé est donné dans la table 5.

Tableau 5 : Liste des champs des fichiers Sirene départementaux nettoyés

Nom	Libellé	Type
NOM_SIRENE	Nom de l'établissement ou de ses dirigeants (éventuellement plusieurs valeurs possibles, séparées par des « @ »)	Texte

4. Lors d'une prochaine mise à jour, le package R COGugaison sera mobilisé : https://github.com/antuki/COGugaison

5. denominationUsuelleEtablissement, enseigne1Etablissement, enseigne2Etablissement, enseigne3Etablissement, denominationUniteLegale, nomUniteLegale, nomUsageUniteLegale, prenomUsuelUniteLegale, prenom1UniteLegale, prenom2UniteLegale, prenom3UniteLegale, prenom4UniteLegale, pseudonymeUniteLegale, complementAdresseEtablissement, complementAdresse2Etablissement, sigleUniteLegale, denominationUsuelle1UniteLegale, denominationUsuelle2UniteLegale, denominationUsuelle3UniteLegale

6. ADRESSETEMP, complementAdresseEtablissement, distributionSpecialeEtablissement, ADRESSETEMP2, complementAdresse2Etablissement, distributionSpeciale2Etablissement

Nom	Libellé	Type
NOM_1_SIRENE	Première valeur possible prise par NOM_SIRENE	Texte
NOM_2_SIRENE	Seconde valeur possible prise par NOM_SIRENE, « » si vide	Texte
NOM_3_SIRENE	Troisième valeur possible prise par NOM_SIRENE, « » si vide	Texte
NOM_4_SIRENE	Quatrième valeur possible prise par NOM_SIRENE, « » si vide	Texte
NOM_5_SIRENE	Cinquième valeur possible prise par NOM_SIRENE, « » si vide	Texte
NOM_6_SIRENE	Sixième valeur possible prise par NOM_SIRENE, « » si vide	Texte
ADRESSE_SIRENE	Adresse postale de l'établissement (éventuellement plusieurs valeurs possibles, séparées par des « @ »)	Texte
ADRESSE_1_SIRENE	Première valeur possible prise par ADRESSE_SIRENE	Texte
ADRESSE_2_SIRENE	Seconde valeur possible prise par ADRESSE_SIRENE, « » si vide	Texte
ADRESSE_3_SIRENE	Troisième valeur possible prise par ADRESSE_SIRENE, « » si vide	Texte
ANNEE_CREATION_ETAB_SIRENE	Année de création de l'établissement	Texte
ANNEE_CREATION_UNITLEG_SIRENE	Année de création de l'unité légale	Texte
NO_INSEE_SIRENE	Code commune de l'établissement (adresse principale)	Texte
NO_INSEE_SIRENE2	Code commune de l'établissement (adresse secondaire)	Texte
SIREN_SIRENE	Numéro SIREN de l'unité légale	Texte
SIRET_SIRENE	Numéro SIRET de l'établissement	Texte
PAYS_SIRENE	Pays de l'établissement (adresse principale)	Texte
PAYS_SIRENE2	Pays de l'établissement (adresse secondaire)	Texte
DEPCOMEN_SIRENE	Département de l'établissement (adresse principale)	Texte
DEPCOMEN_SIRENE2	Département de l'établissement (adresse secondaire)	Texte
ACTIVPPLE_ETAB_SIRENE	Activité principale de l'établissement	Texte

Nom	Libellé	Type
ACTIVPPLE_UNITLEG_SIRENE	Activité principale de l'unité légale	Texte
NOMENC_ACTIVPPLE_UNITLEG_SIRENE	Nomenclature de l'activité principale de l'unité légale	Texte
CATJUR_SIRENE	Catégorie juridique de l'unité légale	Texte
ETAT_ADMIN_ETAB_SIRENE	État administratif de l'établissement (actif/fermé)	Texte
ETAT_ADMIN_UNITLEG_SIRENE	État administratif de l'unité légale (actif/fermé)	Texte
X_CENTROID_SIRENE	Abscisse du centroïde de la commune (adresse principale)	Numérique
Y_CENTROID_SIRENE	Ordonnée du centroïde de la commune (adresse principale)	Numérique
X_CENTROID_SIRENE2	Abscisse du centroïde de la commune (adresse secondaire)	Numérique
Y_CENTROID_SIRENE2	Ordonnée du centroïde de la commune (adresse secondaire)	Numérique

3.3. Données des opérateurs habilités

Seuls les champs utiles à l'opération de sirétisation sont sélectionnés dans le fichier des opérateurs (cf. Table 3).

- Du fichier initial sont retirées les lignes sans code commune qui correspondent à des opérateurs étrangers.
- Les doublons sur les champs retenus sont retirés : on ne garde qu'une ligne par nom_operateur, code_insee, adresse, siret et statut_simpl.
- Tous les champs sont nettoyés.
- Le champ adresse est harmonisé dans un champ adresse_restr
  - suppression des codes postaux et villes ;
  - si l'adresse vaut ADRESSE1, elle est mise à vide ;
  - si l'adresse ne contient que des chiffres, elle est mise à vide ;
  - suppression des numéros de téléphone ;
  - ajout d'un espace entre des chiffres et des mots tels que BIS, TER, CHEMIN, ROUTE, etc. (exemple : 4BIS devient 4 BIS) ;
  - certaines adresses sont entre $<$ $>$: suppression des signes $<$ $>$;
  - suppression des termes du type « JOUR DE FERMETURE » (plusieurs variantes) ;
  - suppression des termes « à gauche », « à droite » et leurs variantes ;
  - suppression des termes « A XX KMS » ;
  - suppression du terme « ROUTE : $>$ » ;
  - pour les adresses contenant les termes « M2 », on ne garde que le texte après le dernier M2 ;

  - renommage des types de voie (exemples RTE = ROUTE, IMP = IMPASSE, etc.) ;
  - correction de caractères spéciaux (chiffres à la place de lettres comme par exemple 4 à la place de l'apostrophe) ;
  - pour les adresses contenant les termes « CHANGEMENT DE », « DONNEES BIO », « DE RUCHE », « : », « = » : extraction des chaînes de caractères de type « N° + type de voie + texte ». Si aucun remplacement n'est fait, l'adresse est mise à vide ;
  - suppression des termes MONSIEUR, MADAME et leurs variantes ;
  - pour les adresses de plus de 20 mots, extraction des chaînes de caractères de type « N° + type de voie + texte ». Si aucun remplacement n'est fait, l'adresse est mise à vide ;
  - suppression des doublons sur les champs retenus suite aux modifications apportées.
- Les coordonnées géographiques des codes communes sont rapatriées du fichier Geofla (appariement sur le code commune).

  Pour les codes communes du fichier des opérateurs habilités non retrouvés dans les données Geofla, on applique des corrections manuelles (recherche du code Insee Geofla existant le plus pertinent).
- Suppression de deux lignes de Monaco, absentes des données Sirene.
- Suppression des doublons sur les champs retenus suite aux modifications apportées.
- Suppression des numéros SIRET qui ne sont pas composés uniquement de chiffres, qui ne sont pas de longueur 9 (numéro SIREN) ou 14 (numéro SIRET) ou qui commencent par plusieurs fois le même chiffre (exemple : 00000000000000). Pour les numéros SIRET de 13 ou 15 caractères, il s'agit la plupart du temps d'un chiffre manquant, doublé ou bien d'un chiffre ajouté au début (comparaisons ponctuelles avec le site societe.com).

  Les corrections manuelles étant trop longues à faire, les numéros SIRET différents de 9 ou 14 caractères sont mis à vide.
- Suppression des doublons sur les champs retenus suite aux modifications apportées.
- Le champ nom_operateur est harmonisé
  - suppression des termes MONSIEUR, MADAME et leurs variantes ;
  - suppression des valeurs erronées telles que « #NOM » ou « !!!!!!!!EXPORT!!!!!! » ;
  - suppression des numéros de téléphone ;
  - suppression de termes comme « RESPONSABLE », « DIRECTEUR », etc. ;
  - suppression des doublons sur les champs retenus suite aux modifications apportées.
- On ne garde qu'une ligne par nom_operateur, code_insee, siret et adresse_restr.
- Suppression des lignes sans nom ni adresse renseignés.
- Création d'un identifiant ligne xxxind.

Le dessin d'enregistrement final pour le fichier OH_2020 nettoyé est donné dans la table 6.

Le numéro SIRET est manquant pour environ 32% des lignes du fichier (cf. Table 7).

Tableau 6 : Liste des champs du fichier OH_2020 nettoyé

Nom	Libellé	Type
xxxind	Identifiant de la ligne	Numérique
NOM_SOURCE	Nom de l'établissement ou de ses dirigeants (éventuellement plusieurs valeurs possibles, séparées par des « @ »)	Texte
ADRESSE_SOURCE	Adresse postale de l'établissement (éventuellement plusieurs valeurs possibles, séparées par des « @ »)	Texte
HAB_SOURCE	Statut de l'opérateur (habilité / non habilité)	Texte
NO_INSEE_SOURCE	Code commune de l'établissement	Texte
DEPARTEMENT_SOURCE	Département de l'établissement	Texte
SIREN_SOURCE	Numéro SIREN	Texte
SIRET_SOURCE	Numéro SIRET	Texte
X_CENTROID_SOURCE	Abscisse du centroïde de la commune	Numérique
Y_CENTROID_SOURCE	Ordonnée du centroïde de la commune	Numérique

Le suffixe _SOURCE permet de rendre les traitements génériques à d'autres jeux de données.

Tableau 7 : Répartition des opérateurs habilités selon les SIRET et adresse manquants

	Adresse renseignée	Adresse manquante	Total
SIRET renseigné	141 382	10 085	151 467
	63.8%	4.6%	68.4%
SIREN renseigné	172	12	184
	0.1%	0.0%	0.1%
SIRET manquant	58 104	11 709	69 813
	26.2%	5.3%	31.5%
Total	199 658	21 806	221 464
	90.1%	9.9%	100.0%

4. Méthodologie employée

L'idée générale de **l'algorithme de sirétisation** est que pour chaque ligne à sirétiser, on va :

- chercher toutes les entreprises de la même commune (ou du département) dans la base Sirene ;
- calculer des **indicateurs de dissimilarité/distance** sur les champs NOM (et ADRESSE si renseignée) ;
- appliquer un **modèle prédictif** sur ces indicateurs qui retourne une probabilité (un score), pour chaque candidat Sirene, d'être le bon numéro SIRET à retenir ;
- retenir le.s candidat.s Sirene qui dépasse.nt un certain seuil.

Pour ce faire, on se place dans le cadre de l'apprentissage statistique où une partie des données OH_2020 appariées avec les données Sirene va servir à définir les indicateurs et le meilleur modèle prédictif à retenir parmi une sélection, puis entrainer le modèle (échantillon d'apprentissage), tandis que le reste des données servira à valider la performance de la procédure complète une fois les indicateurs et modèles établis (échantillon test).

Un schéma simplifié de la procédure est donnée dans la figure 2 et le workflow du développement de l'algorithme est disponible en annexe 1.

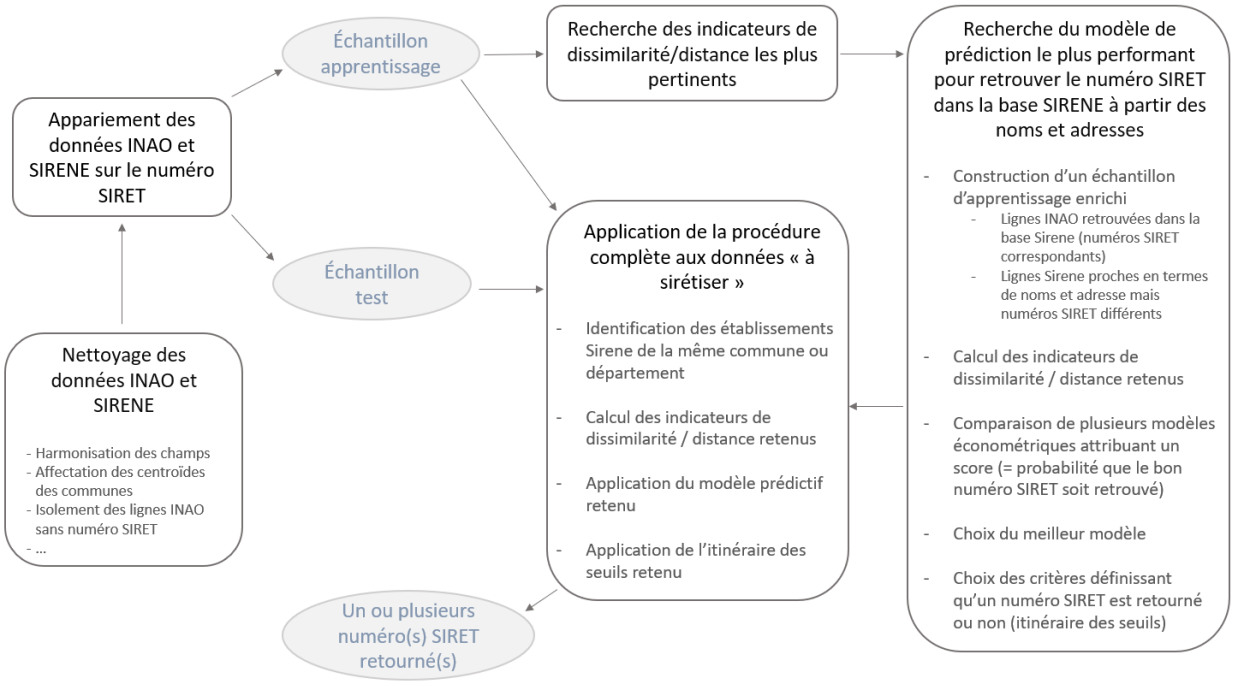


Figure 2 : Schéma simplifié de la procédure de sirétisation

4.1. Échantillonnage

Une fois chaque fichier de données nettoyé et préparé (variables mises au bon format et passées dans les fonctions de nettoyage), le fichier source, ici celui des opérateurs habilités en 2020, est apparié aux données Sirene, quand possible, sur le numéro SIRET pour compléter les données OH_2020 avec les données Sirene disponibles.

Les 69 813 lignes du fichier source sans numéro SIRET sont isolées dans un fichier *ech_siretvide.csv* et les 151 651 lignes restantes constituent la base de travail qui est ensuite découpée en deux échantillons :

- un échantillon d'apprentissage qui servira à étudier les distances et développer l'algorithme (70% des données) ;
- un échantillon test qui servira à vérifier les résultats de l'algorithme et s'assurer de sa neutralité vis-à-vis de l'échantillon d'apprentissage (30% des données).

Les résultats de l'appariement et du découpage sont donnés dans le tableau 8.

Tableau 8 : Résultats de l'appariement des fichiers OH_2020 et Sirene

Fichier	Lignes retrouvées dans Sirene	Lignes non retrouvées dans Sirene	Total
Apprentissage	99 084	7 071	106 155
Test	42 399	3 097	45 496
Total	141 483	10 168	151 651

4.2. Développement de l'algorithme

L'algorithme est développé sur l'échantillon d'apprentissage. Cet échantillon est lui-même réparti en deux sous-échantillons :

- les lignes dont le siret source renseigné a trouvé une correspondance dans la base Sirene (BASE_MERGE) ;
- les lignes dont le siret source renseigné n'a pas été retrouvé dans la base Sirene.

Sur la BASE_MERGE, on ne garde que les lignes où au moins le nom est renseigné, l'adresse pouvant, elle, être manquante.

Plusieurs indicateurs de distance et dissimilarité sur les chaînes de caractères des nom et adresse entre fichier source et Sirene sont calculés et étudiés par une analyse en composantes principales afin de déterminer lesquels sont les plus pertinents à mobiliser. Nous retenons quatre méthodes de calcul de dissimilarité (Q-GRAM, LCS, LCS_SIM, JACCARD) appliquées chacune aux champs NOM et ADRESSE, auxquelles s'ajoute la distance euclidienne entre les centroïdes des communes des fichiers source et Sirene, ce qui donne 9 indicateurs (cf. partie 5).

Afin de comparer plusieurs modèles prédictifs, il nous faut une base de travail avec des lignes INAO retrouvées dans la base Sirene (numéros SIRET identiques) et des lignes Sirene pouvant être des candidats plausibles (indicateurs de dissimilarité faibles) mais dont les numéros SIRET ne correspondent pas. Pour construire cette base de travail, on lance un algorithme simplifié sur la BASE_MERGE qui, pour chaque ligne du fichier source :

- extrait les entreprises Sirene de la même commune ou du même département [7] s'il n'y a pas d'entreprises Sirene dans la même commune (algorithme *commune + département*) ou si le choix est fait de lancer sur le département directement (algorithme *département*) ;

7. Lancer l'algorithme sur le département permet de prendre en compte les erreurs de code commune

- calcule les indicateurs de dissimilarité retenus entre la ligne source et les lignes Sirene sur les champs NOM (et ADRESSE si renseignée) ;
- calcule la moyenne arithmétique MEAN_DIST des indicateurs de dissimilarité calculés non manquants ;
- retourne les lignes Sirene correspondant aux 25 plus petites distances moyennes si l'algorithme est lancé sur les mêmes communes, 10 [8] s'il est lancé sur le département ;

Les meilleurs résultats retournés pour chaque ligne du fichier source sont ajoutés aux « vrais » résultats issus de l'appariement fichier source / fichier Sirene (BASE_MERGE) et constituent ainsi une base de travail pour entrainer l'algorithme.

On définit la variable OK = 1 (bon résultat : mêmes numéros SIRET source et Sirene) et OK = 0 (mauvais résultat : numéros SIRET source et Sirene différents mais indicateurs de dissimilarité faibles). Des critères supplémentaires à intégrer aux modèles sont déterminés afin de différencier au mieux la ligne de bon résultat des autres lignes. On retire les lignes où OK = 1 et MEAN_DIST élevée (les numéros SIRET correspondent mais les noms (et adresse) n'ont rien à voir) et les lignes où OK = 0 et MEAN_DIST = 0 (les numéros SIRET diffèrent mais les noms (et adresse) sont identiques) de la base de travail afin de ne pas entrainer les modèles sur des données fausses.

L'algorithme et les critères supplémentaires sont détaillés dans la partie 6.

Plusieurs modèles sont testés sur la base de travail (régressions et méthodes d'apprentissage) afin de sélectionner le plus performant en termes de prédiction. Les algorithmes sont lancés une fois en intégrant les champs NOM et ADRESSE non vides côté source et Sirene, et une fois en tenant compte seulement des champs NOM afin d'éviter le remplacement des distances manquantes d'adresse par des valeurs ad hoc. Ils sont présentés dans la partie 6.3.

Ces modèles renvoient, pour chaque ligne, un score entre 0 et 1 qui traduit la probabilité que la ligne Sirene corresponde à la ligne source d'après leurs champs NOM et ADRESSE respectifs, ainsi que leur commune. La distribution des scores prédits est étudiée afin de déterminer un seuil à partir duquel on considère que la prédiction est bonne.

Dans cette étude, la méthode des forêts aléatoires est celle qui minimise les erreurs de prédiction. Elle est donc retenue et appliquée à la base de travail pour fixer les paramètres des modèles qui sont ensuite sauvegardés et appelés par la procédure finale (cf. partie 6.4). La procédure complète est ensuite appliquée à l'échantillon d'apprentissage et les résultats sont étudiés (cf. partie 7).

4.3. Validation de l'algorithme

La procédure complète est appliquée à l'échantillon test, indépendant des données ayant servi à la construire, et les résultats sont étudiés (cf. partie 7). Les distributions des distances et des scores et le pourcentage de numéros SIRET retrouvés doivent être similaires à celles de l'échantillon d'apprentissage.

5. Distances et similarités

Dans cette section, nous présentons les mesures de distance/dissimilarité mobilisées pour la méthode de sirétisation.

8. Ces valeurs sont arbitraires et conditionnées par la limite de mémoire de calcul des traitements

Pour calculer les distances entre chaînes de caractères, nous avons utilisé le package *stringdist*[9] de R. Dans ce package, il est possible de calculer des distances ou des similarités. La similarité est calculée comme (1 - la distance entre les chaînes de caractères divisée par la distance maximale possible). Le résultat est un score entre 0 et 1, 1 correspondant à une similarité complète et 0 à une complète dissimilarité.

Lorsque nous travaillons sur les distances, on s'attache à ramener la valeur obtenue entre 0 et 1 afin de comparer les méthodes entre elles, soit en divisant la distance retournée par la fonction *R* entre les chaînes de caractères par la distance maximale possible, soit en faisant 1 - la similarité. Dans la plupart des cas abordés ici, les deux méthodes donnent la même valeur, à l'exception de la méthode *LCS*.

Une analyse en composantes principales sur ces distances et similarités a permis de sélectionner quatre méthodes : *q-gram*, *LCS* calculée par les distances et par les similarités et *Jaccard*.

5.1. Distance q-gram

La distance *q-gram* D entre deux chaînes s et t est définie par :

$$D_{qgram}(s;t;q) = \|v(s;q) - v(t;q)\|_1$$

Où un q-gramme est une chaîne de q caractères consécutifs,
$\|.\|_1$ indique la norme 1 (distance de Manhattan),
$v(s;q)$ est un vecteur d'entiers non négatifs de dimension $|\Sigma|^q$ dont les coefficients représentent le nombre d'occurences de tous les q-grammes possibles de s.

En d'autres termes, la fonction *stringdist(a, b, method='qgram', q=q)* donne le nombre de q-grammes présents dans l'une des deux chaînes mais pas dans l'autre.

On choisit de fixer $q = 1$ pour la sirétisation, ce qui sera 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', 'leila organa', method='qgram', q=1)
[1] 1 # un l en plus dans la seconde chaîne
> stringdist('leia organa', 'organa leia', method='qgram', q=1)
[1] 0 #les mêmes caractères
```

L'inconvénient de cette méthode est que pour $q = 1$, une parfaite similarité (ou distance nulle) peut signifier deux chaînes qui sont des anagrammes (deux chaînes différentes mais composées des mêmes caractères). C'est pourquoi nous utilisons d'autres méthodes en complément.

Pour pouvoir être comparable aux autres critères, cette distance D est ramenée entre 0 et 1 par la formule suivante :

$$d_{qgram1}(s;t) = \frac{D_{qgram}(s;t;1)}{|s| + |t|}$$

Où $|.|$ indique la longueur d'une chaîne de caractères.

9. Van der Loo M (2014). « The stringdist package for approximate string matching ». *The R Journal*, 6, 111-122. https://CRAN.R-project.org/package=stringdist.

La distance $d_{qgram1}(s;t)$ est ainsi égale à 0 si les deux chaînes s et t font la même taille et sont composées des mêmes lettres, 1 si aucune lettre n'est commune entre les deux chaînes s et t.

5.2. Distance LCS - Longest Common String

La fonction *stringdist(a, b, method='LCS')* retourne le nombre d'opérations (suppressions, insertions) nécessaires pour passer de la chaîne a à la chaîne b. Cette distance permet de calculer la longueur de la plus longue chaîne commune de caractères (LCS) qui peut par ailleurs être obtenue en comparant deux à deux les caractères des deux chaînes, en gardant le même ordre des caractères dans chaque chaîne, mais les caractères n'étant pas obligatoirement consécutifs. En effet, la distance LCS *stringdist(a, b, method='LCS')* donne le nombre de caractères non appariés.

Exemple :
Chaîne a : ADUZHD
Chaîne b : ARHUKNHD
Plus longue chaîne commune de caractères (LCS) : AUHD
Nombre d'opérations pour passer de la chaîne a à la chaîne b : *stringdist('ADUZHD', 'ARHUKNHD', method='lcs')* = 6 (suppression des caractères D et Z de la chaîne a et insertion des lettres R, H, K et N de la chaîne b)

La longueur de la LCS $|lcs(s,t)|$ entre deux chaînes de caractères s et t est donné par :

$$|lcs(s,t)| = \frac{|s| + |t| - D_{lcs}(s,t)}{2}$$

Où $|.|$ indique la longueur d'une chaîne de caractères,
$D_{lcs}(s,t)$ est le nombre d'opérations (suppressions, insertions) nécessaires pour passer de la chaîne s à la chaîne t retourné par la fonction *stringdist(s, t, method='LCS')*.

Pour pouvoir être comparable aux autres critères, la métrique est ramenée entre 0 et 1 par la formule suivante :

$$d_{lcs}(s;t) = 1 - \frac{|lcs(s,t)|}{min(|s|,|t|)}$$

Ainsi, la distance $d_{lcs}(s;t)$ est égale à 0 si l'une des deux chaînes est contenue dans l'autre puisque la longueur de la plus longue chaîne commune de caractères sera la même que la plus petite longueur des chaînes s et t. Cette distance, calculée par l'US-ODR est légèrement différente de celle retournée en utilisant la formule (1 - similarité donnée par la fonction *stringsim(s, t, method='LCS')*). Cette dernière ne considère pas une chaîne incluse dans une autre comme une similarité parfaite.

```
stringdist('leia', 'leela', method='lcs')
[1] 3 #3 opérations pour passer de leai à leela (1 suppression et 2 insertions)
stringdist('gaec du leia', 'du leia', method='lcs')
[1] 5 #5 opérations pour passer de l'une à l'autre
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
1 - stringsim('gaec du leia', 'du leia', method='lcs')
[1] 0.2631579  #alors que 1 - similarité renvoie une distance non nulle
```

Nous faisons donc le choix de conserver les deux distances, celle calculée par 1 - *stringsim(s, t, method='LCS')* (d_{lcssim}) et celle construite par l'US-ODR (d_{lcs}).

5.3. Distance Jaccard

Cette méthode se rapproche de la méthode *q-gram*. La distance *Jaccard* entre deux chaînes s et t est définie par :

$$d_{jaccard}(s;t;q) = 1 - (\frac{|v(s,q) \cap v(t,q)|}{|v(s,q) \cup v(t,q)|})$$

Où un q-gramme est une chaîne de q caractères consécutifs,
$v(s,q)$ est la liste des q-grammes de la chaîne s,
$|v(s,q) \bigcap v(t,q)|$ indique le nombre de q-grammes distincts communs aux deux chaînes de caractères s et t,
$|v(s,q) \bigcup v(t,q)|$ indique le nombre total de q-grammes distincts des deux chaînes de caractères s et t.

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, pour une complémentarité avec la distance *q-gram* ($q = 1$).

```
stringdist('leia', 'leela', method='jaccard', q=2)
[1] 0.8333333  #un q-gramme commun (le), 6 q-grammes distincts totaux, dist = 1 - (1/6)
stringdist('gaec leia', 'leia gaec', method='jaccard', q=2)
[1] 0.4  #6 q-grammes communs (le, ei, ia, ga, ae, ec), 10 q-grammes distincts totaux,
    ↪ dist = 1 - (6/10)
```

5.4. Distance euclidienne entre communes

La distance euclidienne $d_{com}(a,b)$ entre les centroïdes de deux communes a et b est donnée par :

$$\sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}$$

Où x_i est l'abscisse du centroïde de la commune i,
y_i est l'ordonnée du centroïde de la commune i

Compte tenu de l'unité des coordonnées des centroïdes du fichier GEOFLA, cette distance est divisée par 10^5 pour être comprise entre 0 et 1.

Lorsqu'une adresse secondaire existe pour un établissement dans les données Sirene, la distance euclidienne minimum entre Source/adresse principale Sirene et Source/adresse secondaire Sirene est retenue.

6. Recherche du meilleur modèle prédictif

A l'aide des distances retenues, nous entrainons différents modèles pour sélectionner le meilleur candidat à la sirétisation sur l'échantillon d'apprentissage. Nous étudions les résultats fournis par ces modèles afin de constituer un score qui reflète la qualité du résultat retourné et déterminer des critères qui permettent de décider si ce résultat est bon ou non.

6.1. Préparation de la base de travail

Dans un premier temps, nous constituons la base de travail contenant les bons résultats (lignes où les numéros SIRET correspondent entre les deux bases) et les faux résultats (entreprises relativement proches mais dont les numéros SIRET ne correspondent pas).

Pour ce faire, pour chaque ligne de l'échantillon d'apprentissage :

- ouvrir le fichier du département concerné ;
- ne garder que les entreprises de la même commune. S'il n'y a aucune entreprise dans la même commune ou si le choix est fait de travailler sur le département, garder toutes celles du département ;
- soient NOM_SIRENE_CONC et ADRESSE_SIRENE_CONC les champs NOM_SIRENE et ADRESSE_SIRENE où le caractère « @ » est remplacé par un espace, calculer les distances suivantes entre chaque couple d'entreprise source/Sirene :
  - QGRAM_NOM_ALGO : minimum de la distance d_{qgram1} entre les champs NOM pris 2 à 2 (i.e. minimum des distances $d_{qgram1}(NOM\_SOURCE, NOM\_SIRENE\_CONC)$, $d_{qgram1}(NOM\_SOURCE, NOM\_SIRENE\_1)$, $d_{qgram1}(NOM\_SOURCE, NOM\_SIRENE\_2)$, $d_{qgram1}(NOM\_SOURCE, NOM\_SIRENE\_3)$, $d_{qgram1}(NOM\_SOURCE, NOM\_SIRENE\_4)$, $d_{qgram1}(NOM\_SOURCE, NOM\_SIRENE\_5)$, $d_{qgram1}(NOM\_SOURCE, NOM\_SIRENE\_6)$) ;
  - QGRAM_ADRESSE_ALGO : minimum de la distance d_{qgram1} entre les champs ADRESSE pris 2 à 2 (i.e. minimum des distances $d_{qgram1}(ADRESSE\_SOURCE, ADRESSE\_SIRENE\_CONC)$, $d_{qgram1}(ADRESSE\_SOURCE, ADRESSE\_SIRENE\_1)$, $d_{qgram1}(ADRESSE\_SOURCE, ADRESSE\_SIRENE\_2)$, $d_{qgram1}(ADRESSE\_SOURCE, ADRESSE\_SIRENE\_3)$) ;
  - LCS_NOM_ALGO : minimum de la distance d_{lcs} entre les champs NOM pris 2 à 2 ;
  - LCS_ADRESSE_ALGO : minimum de la distance d_{lcs} entre les champs ADRESSE pris 2 à 2 ;
  - LCSSIM_NOM_ALGO : minimum de la distance d_{lcssim} entre les champs NOM pris 2 à 2 ;
  - LCSSIM_ADRESSE_ALGO : minimum de la distance d_{lcssim} entre les champs ADRESSE pris 2 à 2 ;
  - JAC_NOM_ALGO : minimum de la distance $d_{jaccard2}$ entre les champs NOM pris 2 à 2 ;
  - JAC_ADRESSE_ALGO : minimum de la distance $d_{jaccard2}$ entre les champs ADRESSE pris 2 à 2 ;
  - DISTANCE_COMMUNES_ALGO : distance d_{com} entre les centroïdes des communes ;
- calculer MEAN_DIST, la moyenne arithmétique des distances non manquantes ci-dessus ;
- retourner les entreprises dont la distance moyenne fait partie des 25 plus petites distances distinctes pour les entreprises dans la même commune, 10 pour les entreprises prises dans le même département.

Point d'attention : les champs NOM et ADRESSE manquants doivent être mis en *NA* et non pas en texte vide « » car la distance entre deux chaînes de caractères vides est 0.

Aux lignes retournées par cet algorithme, nous ajoutons les « bons » résultats s'ils ne sont pas partie des résultats retournés : si la bonne ligne du fichier Sirene (sur la comparaison des numéros SIRET) n'est pas retournée, elle est rajoutée à la table, ainsi que les variables de distances correspondantes. Cet algorithme simple nous donne la base de travail pour le choix de modèle, l'étude des seuils, l'écriture de l'algorithme final et le calcul du score.

6.2. Critères complémentaires

En plus des indicateurs de dissimilarité calculés, nous définissons des indicateurs supplémentaires pouvant aider au choix du bon résultat par l'algorithme :

- **nbdist0** : nombre de distances égales à 0 parmi celles listées ci-dessus (hors distance euclidienne).

  Cela permet d'avantager les lignes qui ont au moins une distance à zéro par rapport aux lignes n'ayant que des distances faibles ;

- **nbdistNOM0** : nombre de distances sur les champs NOM égales à 0 parmi celles listées ci-dessus (hors distance euclidienne).

  Cela permet d'avantager les lignes qui ont au moins une distance à zéro par rapport aux lignes n'ayant que des distances faibles lorsque le modèle n'est lancé que sur le champ NOM ;

- **NBMOTSCOMMUN_NOM** : nombre de mots de trois lettres ou plus communs dans les champs NOM des fichiers source et Sirene.

  Les mots de moins de trois lettres ne sont pas pris en compte pour ne pas comptabiliser des mots de liaison comme « ET », « LE », « LA », « DU », etc. En effet, on préférera apparier la raison sociale « LE CHATEAU DU CABREL » avec « CABREL » plutôt qu'avec « LE NOM DU TEMPS » qui auraient deux mots en commun si on ne faisait pas de restriction sur le nombre de caractères ;

  Point d'attention : si la raison sociale n'a que des mots de moins de trois lettres, il se peut qu'elle soit mal identifiée par l'algorithme mais en principe les autres distances sur les chaînes de caractères compensent cela.

- **NBMOTSCOMMUN_ADRESSE** : nombre de mots de trois lettres ou plus communs dans les champs ADRESSE des fichiers source et Sirene ;

- **NBMOTSCOMMUN1_NOM** : nombre de mots communs dans les champs NOM des fichiers source et Sirene sans restriction de nombre de caractères ;

- **NBMOTSCOMMUN1_ADRESSE** : nombre de mots communs dans les champs ADRESSE des fichiers source et Sirene sans restriction de nombre de caractères.

6.3. Modèles d'apprentissage

À partir des distances et indicateurs supplémentaires, nous lançons des simulations selon différentes méthodes sur la base de travail pour expliquer la variable OK $= 1$ si les SIRET source et Sirene correspondent, 0 sinon.

On commence par regarder la distribution de MEAN_DIST pour les lignes où OK $= 1$ et OK $= 0$ dans la base de travail (cf. figures 3 et 4). Les résultats sont sensiblement les mêmes pour la base de travail construite sur les entreprises du même département.

Pour certaines lignes où les numéros SIRET source et Sirene correspondent (OK $= 1$), certaines distances moyennes sont élevées (supérieures à 0.4). Ce sont les cas où les numéros SIRET correspondent mais les champs NOM voire ADRESSE n'ont rien à voir. Réciproquement, des lignes où les numéros SIRET ne correspondent pas (OK $=$ 0) ont une distance moyenne à 0. Une distance moyenne nulle implique que toutes les distances sont également nulles et que les champs NOM voire ADRESSE sont exactement les mêmes entre les données source et Sirene, on aimerait donc que l'algorithme retourne ces lignes, même si ici les numéros SIRET ne correspondent pas (OK $=$ 0). De même, les lignes OK $= 1$ avec une MEAN_DIST élevée peuvent tromper l'apprentissage de l'algorithme, on choisit de ne pas tenir compte des individus concernés par ces lignes pour développer les modèles.

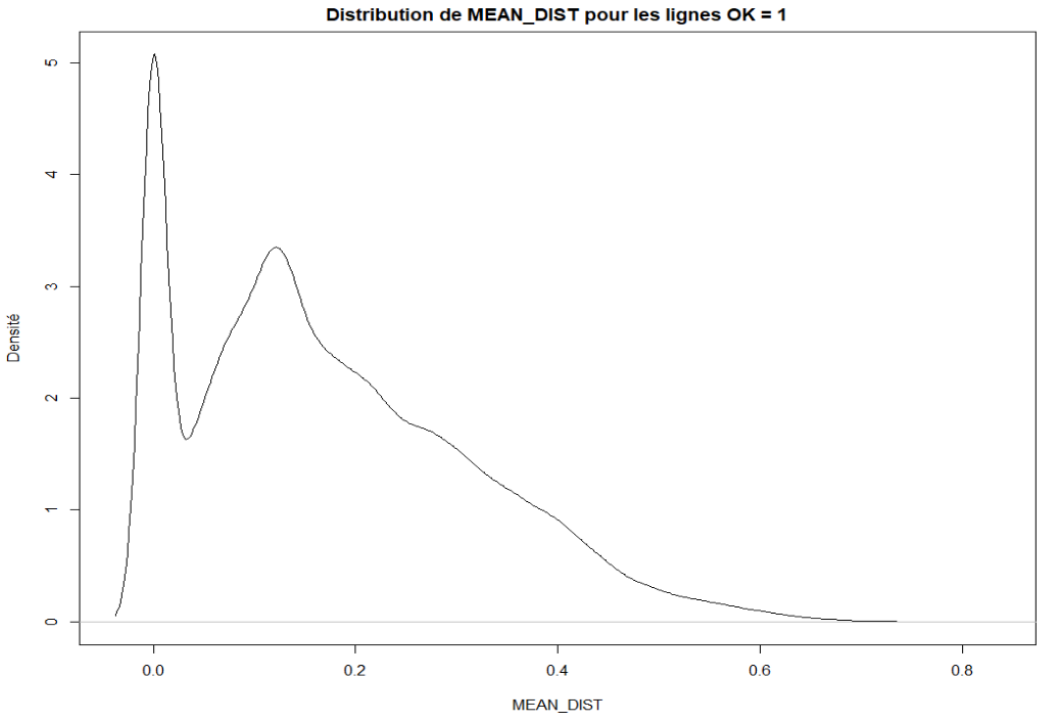


Figure 3 : Distribution de MEAN_DIST pour les lignes OK = 1

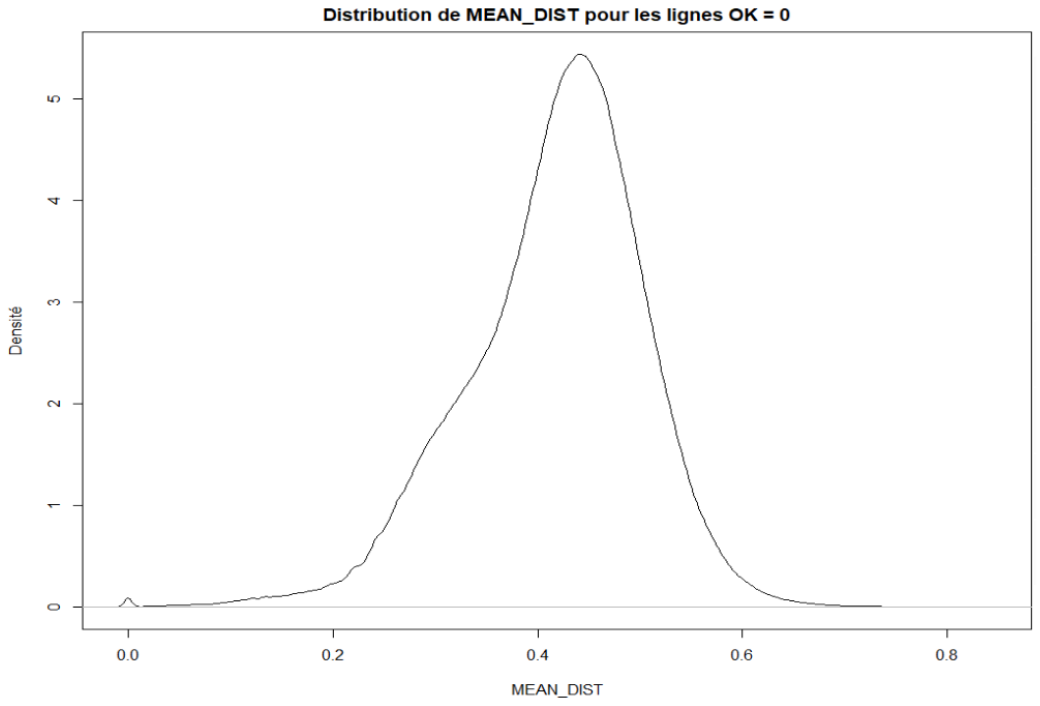


Figure 4 : Distribution de MEAN_DIST pour les lignes OK = 0

Une étude préalable a permis de sélectionner les modèles suivants à tester :

- régression pénalisée Ridge et Lasso (choix du paramètre λ par validation croisée 10-folds) ;
- régression sur composantes principales avec un seuil de 0.3 et 0.4 (pcr03, pcr04) ;
- régression sur moindres carrés partiels avec un seuil de 0.3 et 0.4 (pls03, pls04) ;
- analyse discriminante linéaire (LDA) ;
- forêts aléatoires avec 50 arbres (RandomForest) ;
- arbre de classification (Tree - choix du critère *cp* qui optimise l'arbre par validation croisée).

À noter que le modèle logit pour modéliser une sortie binomiale 0/1 n'est pas recommandé ici car les données ne sont pas indépendantes (plusieurs lignes pour un même individu), et surtout, il existe une importante colinéarité entre les distances choisies comme variables explicatives. Pour les modèles listés ci-dessus, la colinéarité entre variables explicatives n'est pas un problème.

La comparaison des modèles d'apprentissage est faite sur la base de travail des individus dont le jumeau est recherché dans la même commune, par validation croisée 10-folds répétée 5 fois :

- les individus de la base de travail sont répartis en 10 groupes distincts (variable *xxxind*) ;
- les lignes correspondant aux individus du groupe 1 constituent l'échantillon d'apprentissage tandis que les autres constituent l'échantillon test ;
- les modèles sont entrainés sur l'échantillon d'apprentissage et les prédictions sont faites sur l'échantillon test, puis un pourcentage d'observations bien classées est calculé pour ce premier cas de figure ;
- C'est ensuite les lignes des individus du groupe 2 qui constituent l'échantillon d'apprentissage, etc.

Et ce processus est répété 4 autres fois, ce qui donne 50 simulations par modèle. Ceci est fait par la fonction *vfold_cv* du package *rsample*.

Le fait de partitionner sur les individus (*xxxind*) et non sur les lignes permet de conserver, pour chaque individu, toutes les lignes qui lui ont été retournées par l'algorithme simple et ainsi entrainer l'algorithme à choisir entre elles.

Les modèles par validation croisée 10-folds répétée 5 fois sont lancés une première fois sur les lignes dont les champs NOM et ADRESSE sont renseignés côté source et Sirene, les modèles prennent alors comme variables explicatives les distances sur les champs NOM et ADRESSE, puis une fois sur les lignes dont au moins les champs NOM sont renseignés côté source et Sirene, et dans ce cas-là, les modèles intègrent seulement les distances sur les champs NOM. Cela évite le remplacement des distances manquantes par des valeurs ad hoc (comme 0.5 par exemple), ce qui peut biaiser les résultats.

Ensuite, les distributions des pourcentages d'observations bien classées par modèle (50 valeurs) sont représentées graphiquement, pour l'ensemble des observations (cf. figures 5 et 7), puis uniquement pour les observations dont le SIRET de la base INAO correspond au SIRET de la base Sirene (% de 1 bien classés, cf. figures 6 et 8). Compte tenu de la comparaison des résultats, nous retenons la méthode des forêts aléatoires qui donne la meilleure performance.

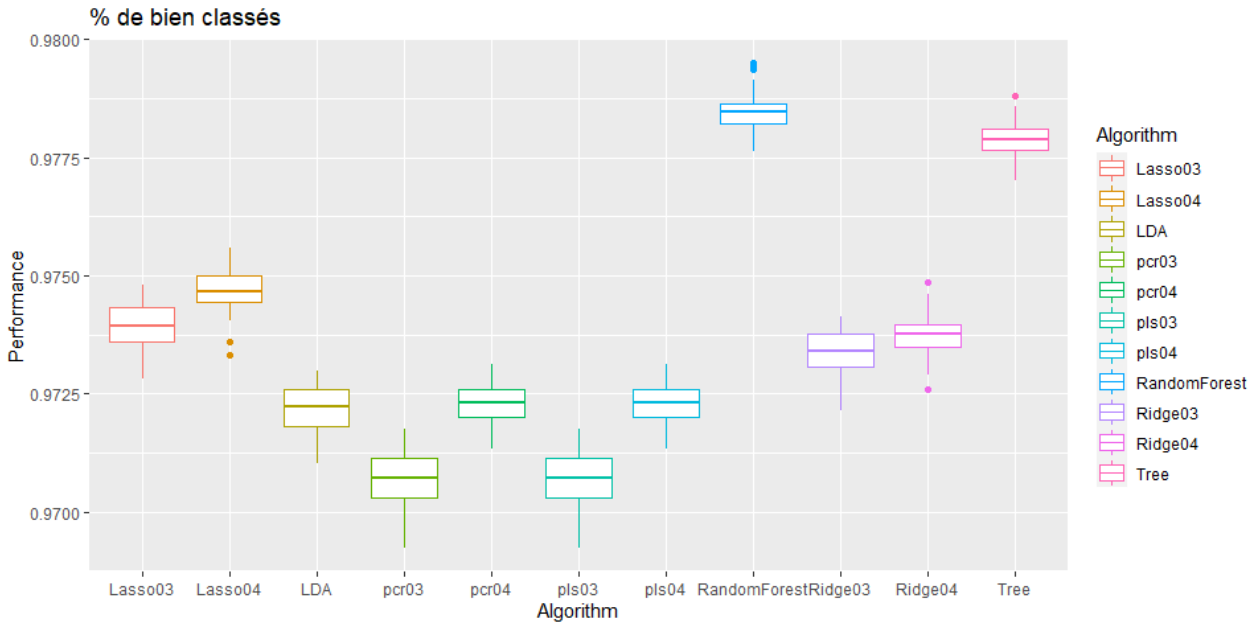


Figure 5 : Pourcentage de bien classés - Modèle sur champs NOM et ADRESSE dans la même commune

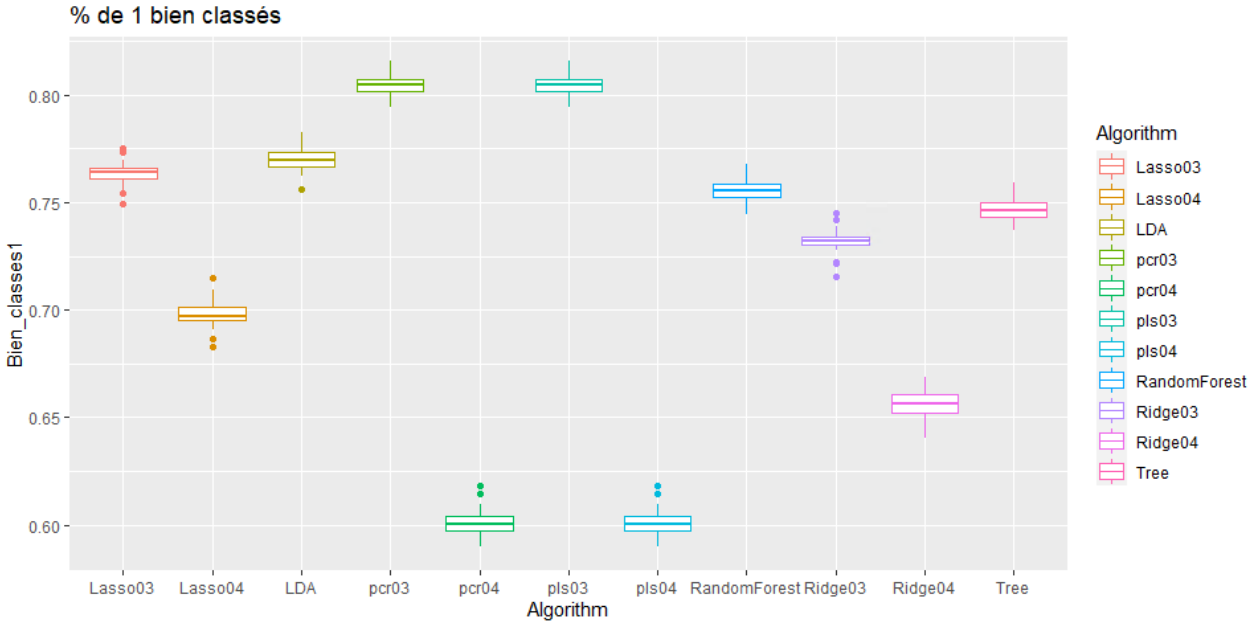


Figure 6 : Pourcentage de 1 bien classés - Modèle sur champs NOM et ADRESSE dans la même commune

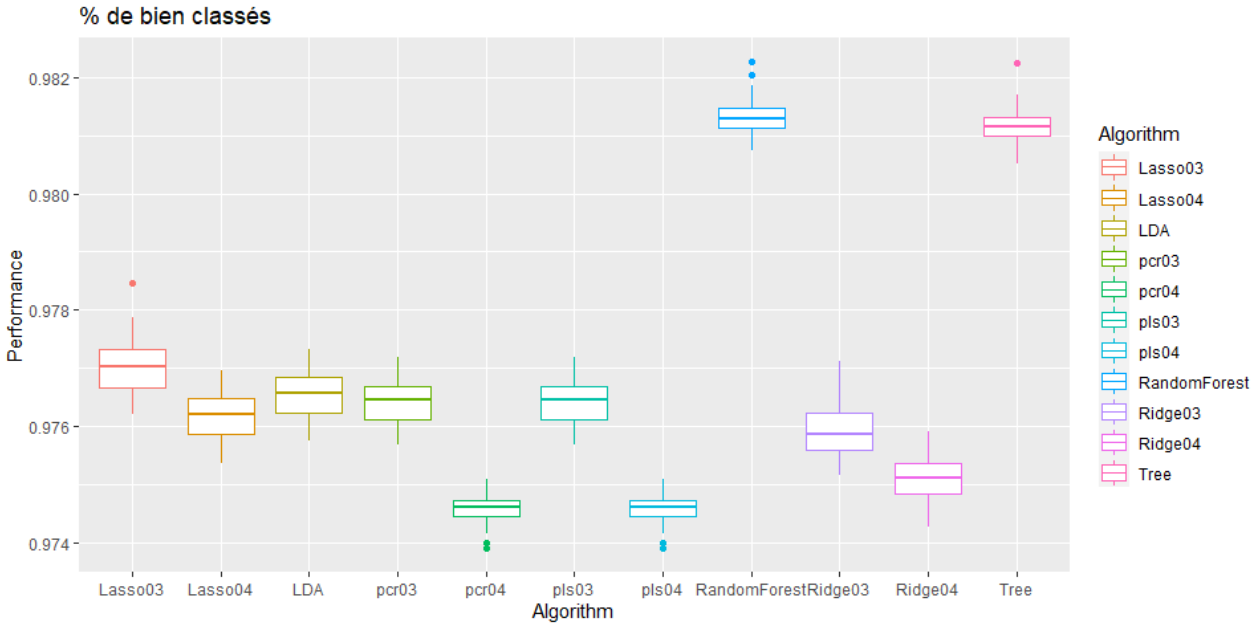


Figure 7 : Pourcentage de bien classés - Modèle sur champs NOM uniquement dans la même commune

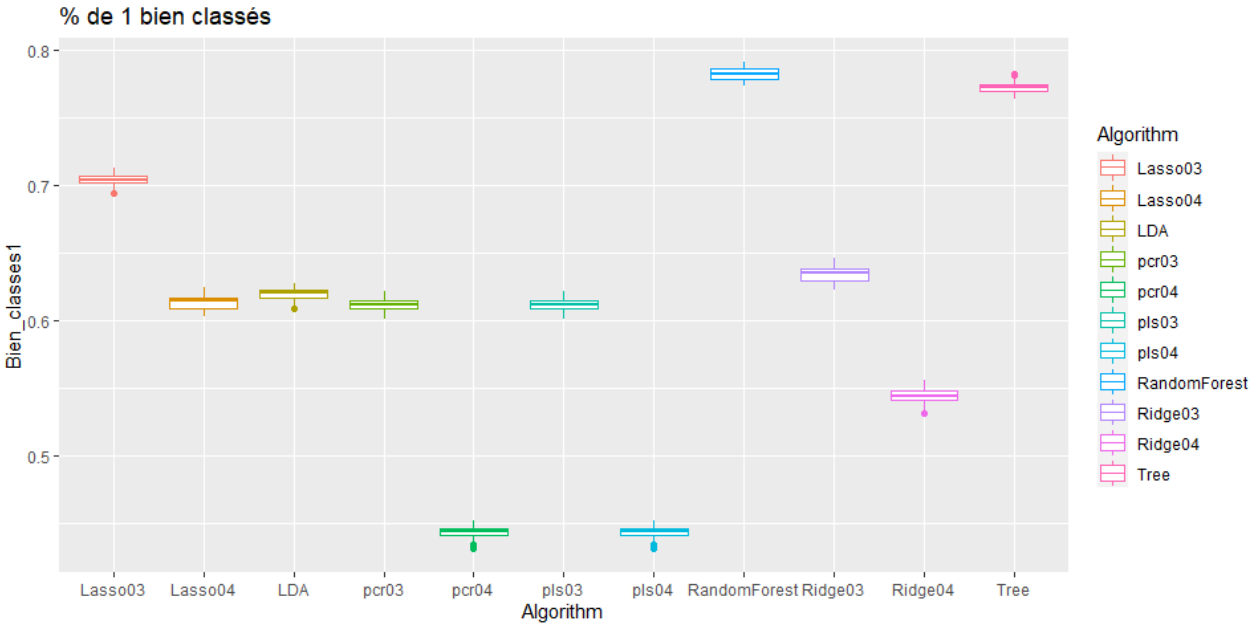


Figure 8 : Pourcentage de 1 bien classés - Modèle sur champs NOM uniquement dans la même commune

Une fois la méthode sélectionnée, nous l'appliquons à l'ensemble de la base de travail afin de fixer les paramètres des modèles qui seront utilisés par l'algorithme final. Nous choisissons de définir **quatre** modèles de forêts aléatoires :

- sur les entreprises dans la même commune, avec les distances et indicateurs complémentaires sur les champs NOM et ADRESSE comme variables explicatives (modèle **rfcna**, cf. partie 6.4.1.1) ;
- sur les entreprises dans la même commune, avec les distances et indicateurs complémentaires sur les champs NOM uniquement comme variables explicatives (modèle **rfcn**, cf. partie 6.4.1.2) ;
- sur les entreprises du même département, avec les distances et indicateurs complémentaires sur les champs NOM et ADRESSE, ainsi que la distance entre communes comme variables explicatives (modèle **rfdna**, cf. partie 6.4.2.1) ;
- sur les entreprises du même département, avec les distances et indicateurs complémentaires sur les champs NOM uniquement, ainsi que la distance entre communes comme variables explicatives (modèle **rfdn**, cf. partie 6.4.2.2).

6.4. Forêts aléatoires sur la base de travail

Pour les modèles décrits ci-après, 2/3 des individus de la base de travail servent à leur construction car beaucoup d'ajustement sont faits. Le 1/3 des individus restants sert de base test pour les modèles de forêts aléatoires. L'échantillon test créé au tout début de la méthodologie (cf. partie 4.1) servira, lui, à valider l'ensemble de la méthode de sirétisation.

6.4.1. Dans la même commune

On ne garde ici que les individus de la base de travail pour lesquels le jumeau est dans la même commune.

a. Modèle sur les champs NOM et ADRESSE (rfcna)

Seules les lignes dont les champs NOM et ADRESSE sont non vides côté source et Sirene sont conservées.

Nous lançons un modèle de forêts aléatoires (fonction *randomForest*) sur les variables suivantes :

- QGRAM_NOM_ALGO ;
- LCS_NOM_ALGO ;
- LCSSIM_NOM_ALGO ;
- JAC_NOM_ALGO ;
- QGRAM_ADRESSE_ALGO ;
- LCS_ADRESSE_ALGO ;
- LCSSIM_ADRESSE_ALGO ;
- JAC_ADRESSE_ALGO ;
- nbdist0 ;
- NBMOTSCOMMUN_NOM ;
- NBMOTSCOMMUN1_NOM ;
- NBMOTSCOMMUN_ADRESSE ;
- NBMOTSCOMMUN1_ADRESSE.

Les figures 9 et 10 montrent respectivement la fréquence d'utilisation des variables dans les forêts aléatoires et leur importance. On peut ainsi voir que les critères supplémentaires calculés comme le nombre de distances nulles ou les nombres de mots en communs, même s'ils sont moins utilisés que les distances, ont un pouvoir discriminant fort. Les distances sur l'adresse ont un pouvoir moins fort que le nom, ce qui correspond aux observations faites sur les données INAO appariées aux données Sirene : souvent, les adresses ne correspondent pas et il convient d'y accorder moins d'importance qu'à la raison sociale, ce qui semble être bien fait par les forêts aléatoires.

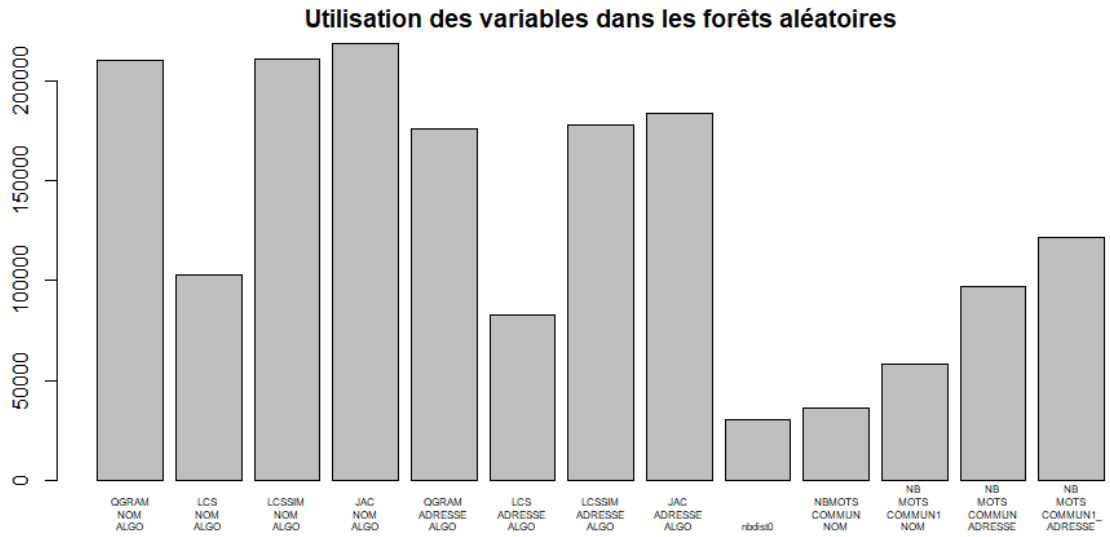


Figure 9 : Utilisation des variables - Modèle sur champs NOM et ADRESSE dans la même commune

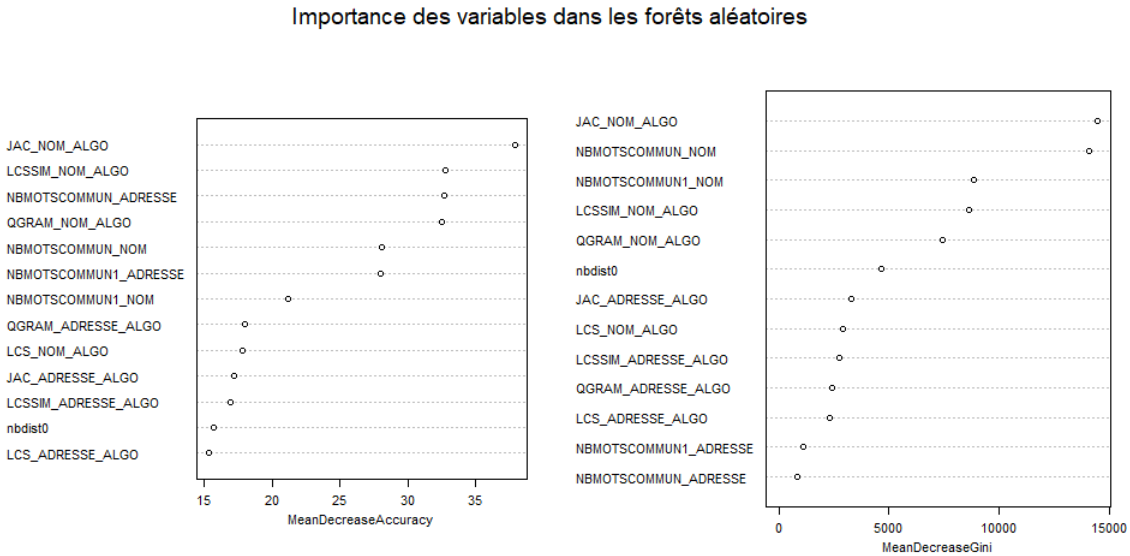


Figure 10 : Importance des variables - Modèle sur champs NOM et ADRESSE dans la même commune

Les matrices de confusion sur les échantillons d'apprentissage et test, à un seuil de 0.5, sont données dans les tableaux 9 et 10. On voit que les taux d'erreur sont sensiblement les mêmes entre les échantillons d'apprentissage et test. Le taux d'erreur des 0 est autour de 1% alors que celui des 1 est autour de 25%. On peut alors se poser la question de savoir si le seuil de 0.5 est pertinent.

Tableau 9 : Matrice de confusion au seuil 0.5 sur l'échantillon d'apprentissage - Modèle sur champs NOM et ADRESSE dans la même commune

Valeur observée	Valeur prédite 0	Valeur prédite 1	Erreur
0	948 319	9 484	0.99%
1	11 917	35 231	25.28%

Tableau 10 : Matrice de confusion au seuil 0.5 sur l'échantillon test - Modèle sur champs NOM et ADRESSE dans la même commune

Valeur observée	Valeur prédite 0	Valeur prédite 1	Erreur
0	472 910	5 236	1.10%
1	5 658	17 901	24.02%

Si on regarde la distribution des scores issus du modèle de forêt aléatoire sur l'échantillon test (figure 11), on se rend compte que la prédiction est globalement dichotomique et on préférera prendre uniquement les résultats prédits autour de 1 (à droite de la fonction de densité). En distinguant les lignes où OK = 0 des lignes où OK = 1, on se rend compte que pour les lignes où OK = 1, il y a des prédictions de score inférieures à 0.2 (cf. figures 12 et 13). Les statistiques descriptives des distributions sont données dans le tableau 11.

Tableau 11 : Statistiques descriptives des prédictions sur l'échantillon test - Modèle sur champs NOM et ADRESSE dans la même commune

OK	Min	Q1	Médiane	Moyenne	Q3	Max
0	0	0	0	0.02	0	1
1	0	0.53	0.81	0.70	0.97	1

Une analyse complémentaire des résultats montre que la plupart des mal classés sont en réalité bien classés par l'algorithme mais le choix du seuil les met dans le mauvais groupe. Il existe néanmoins des cas où le modèle de forêts aléatoires prédit un faible score malgré des distances entre champs NOM et ADRESSE faibles.

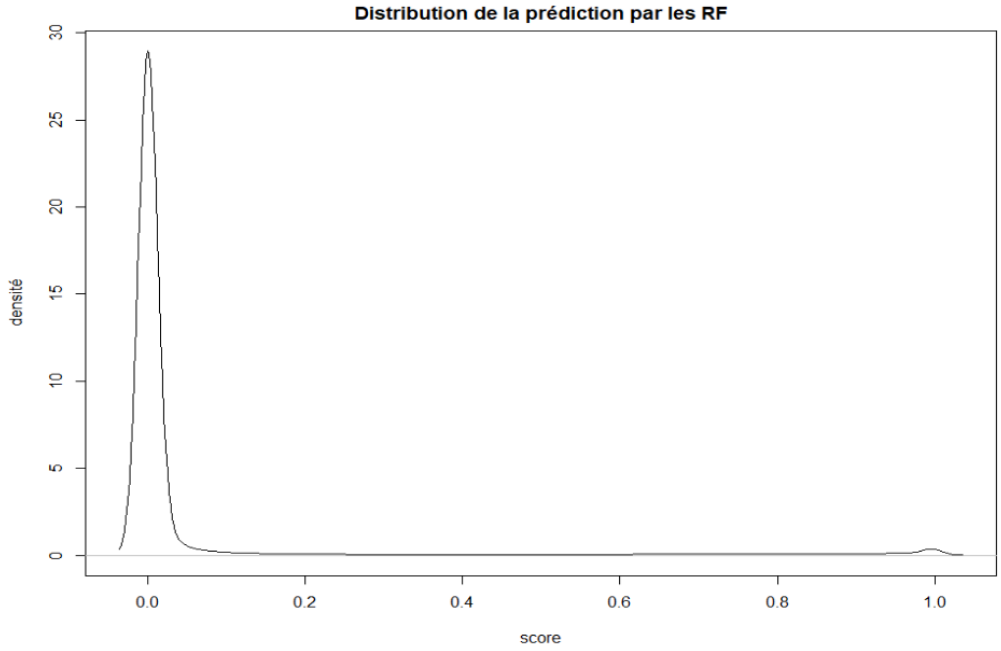


Figure 11 : Distribution de la prédiction sur l'échantillon test - Modèle sur champs NOM et ADRESSE dans la même commune

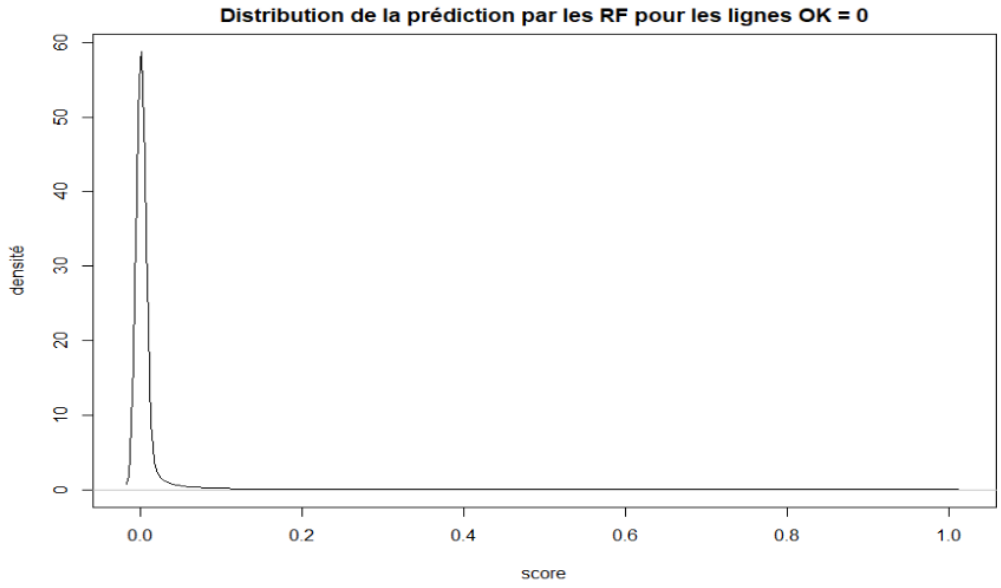


Figure 12 : Distribution de la prédiction sur l'échantillon test pour les lignes OK = 0 - Modèle sur champs NOM et ADRESSE dans la même commune

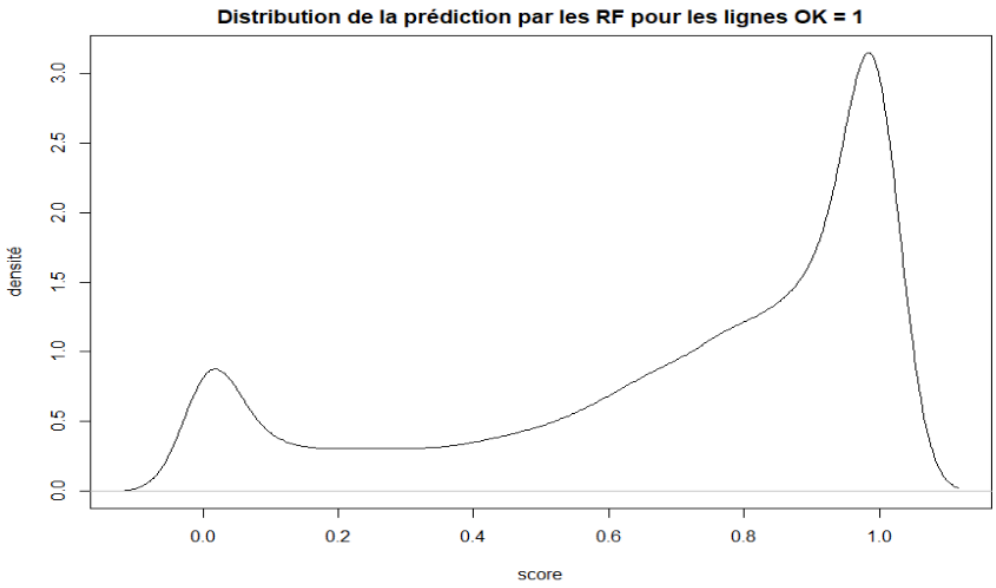


Figure 13 : Distribution de la prédiction sur l'échantillon test pour les lignes OK = 1 - Modèle sur champs NOM et ADRESSE dans la même commune

Le score prédit ne suffisant pas à lui seul à déterminer la validité d'une ligne, le choix est fait de retourner les résultats en plusieurs étapes : d'abord les lignes où le score est élevé (résultat sûr) et si aucune ligne n'est retournée, les lignes avec le(s) meilleur(s) score(s) par individu (cf. partie 6.5). Le score prédit par le modèle sur les champs NOM uniquement sera également pris en compte dans l'itinéraire des seuils.

b. Modèle sur les champs NOM uniquement (rfcn)

Toutes les lignes dont les champs NOM sont non vides côté source et Sirene sont conservées.

Nous lançons un modèle de forêts aléatoires sur les variables suivantes :

- QGRAM_NOM_ALGO ;
- LCS_NOM_ALGO ;
- LCSSIM_NOM_ALGO ;
- JAC_NOM_ALGO ;
- nbdistNOM0 ;
- NBMOTSCOMMUN_NOM ;
- NBMOTSCOMMUN1_NOM.

Les figures 14 et 15 montrent respectivement la fréquence d'utilisation des variables dans les forêts aléatoires et leur importance.

Les matrices de confusion sur les échantillons d'apprentissage et test, à un seuil de 0.5, sont données dans les tableaux 12 et 13. On voit que les taux d'erreur sont sensiblement les mêmes entre les échantillons d'apprentissage et test. On remarque que le taux d'erreur des 0 est autour de 1% alors que celui des 1 est autour de 22%, résultats similaires au modèle sur les champs NOM et ADRESSE.

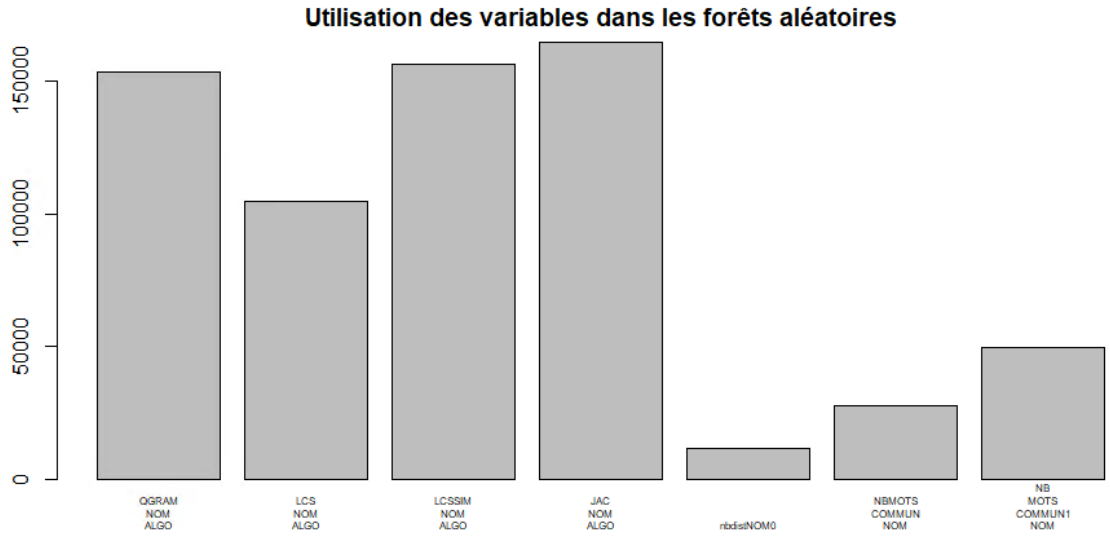


Figure 14 : Utilisation des variables - Modèle sur champs NOM dans la même commune

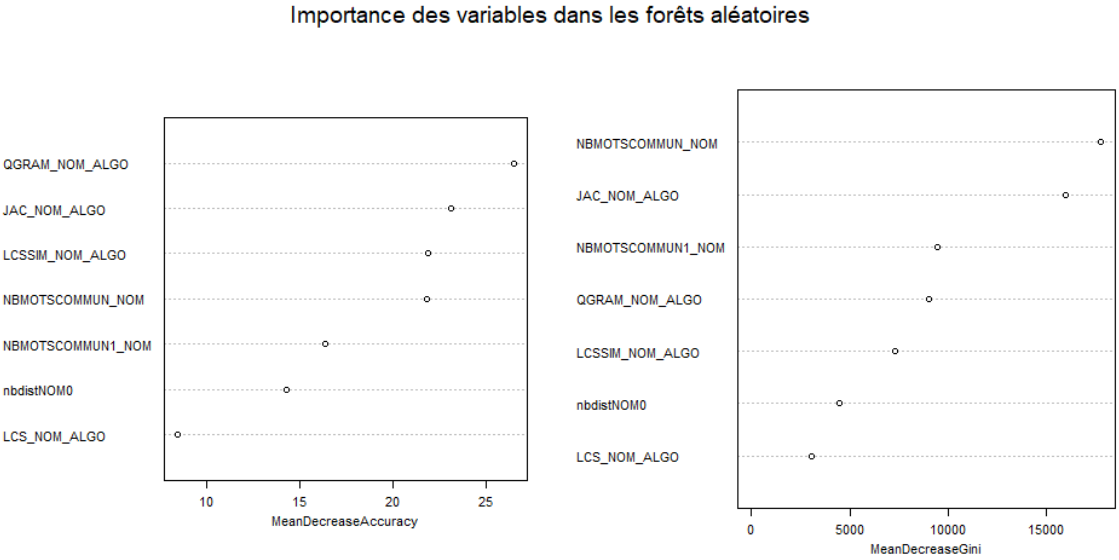


Figure 15 : Importance des variables - Modèle sur champs NOM dans la même commune

Tableau 12 : Matrice de confusion au seuil 0.5 sur l'échantillon d'apprentissage - Modèle sur champs NOM uniquement dans la même commune

Valeur observée	Valeur prédite 0	Valeur prédite 1	Erreur
0	1 323 357	14 304	1.07%
1	11 760	41 687	22.00%

Tableau 13 : Matrice de confusion au seuil 0.5 sur l'échantillon test - Modèle sur champs NOM uniquement dans la même commune

Valeur observée	Valeur prédite 0	Valeur prédite 1	Erreur
0	663 315	7 292	1.09%
1	5 788	20 961	21.64%

Tout comme dans le modèle avec les champs NOM et ADRESSE, la distribution des scores prédits est dichotomique et certaines lignes où OK = 1 ont un score faible (représentations graphiques similaires). Les statistiques descriptives des scores sont données dans le tableau 14.

Tableau 14 : Statistiques descriptives des prédictions sur l'échantillon test - Modèle sur champs NOM dans la même commune

OK	Min	Q1	Médiane	Moyenne	Q3	Max
0	0	0	0	0.01	0	1
1	0	0.66	0.99	0.76	1	1

6.4.2. Dans le même département

Ici, on ne fait plus de distinction sur la même commune ou non mais on prend en compte la distance entre communes.

a. Modèle sur les champs NOM et ADRESSE

Seules les lignes dont les champs NOM et ADRESSE sont non vides côté source et Sirene sont conservées.

Nous lançons un modèle de forêts aléatoires sur les variables suivantes :

- QGRAM_NOM_ALGO ;

- LCS_NOM_ALGO ;
- LCSSIM_NOM_ALGO ;
- JAC_NOM_ALGO ;
- QGRAM_ADRESSE_ALGO ;
- LCS_ADRESSE_ALGO ;
- LCSSIM_ADRESSE_ALGO ;
- JAC_ADRESSE_ALGO ;
- nbdist0 ;
- NBMOTSCOMMUN_NOM ;
- NBMOTSCOMMUN1_NOM ;
- NBMOTSCOMMUN_ADRESSE ;
- NBMOTSCOMMUN1_ADRESSE ;
- DISTANCE_COMMUNES_ALGO.

Les figures 16 et 17 montrent respectivement la fréquence d'utilisation des variables dans les forêts aléatoires et leur importance. La distance entre communes, bien que moins utilisée que d'autres variables, a tout de même un fort pouvoir discriminant.

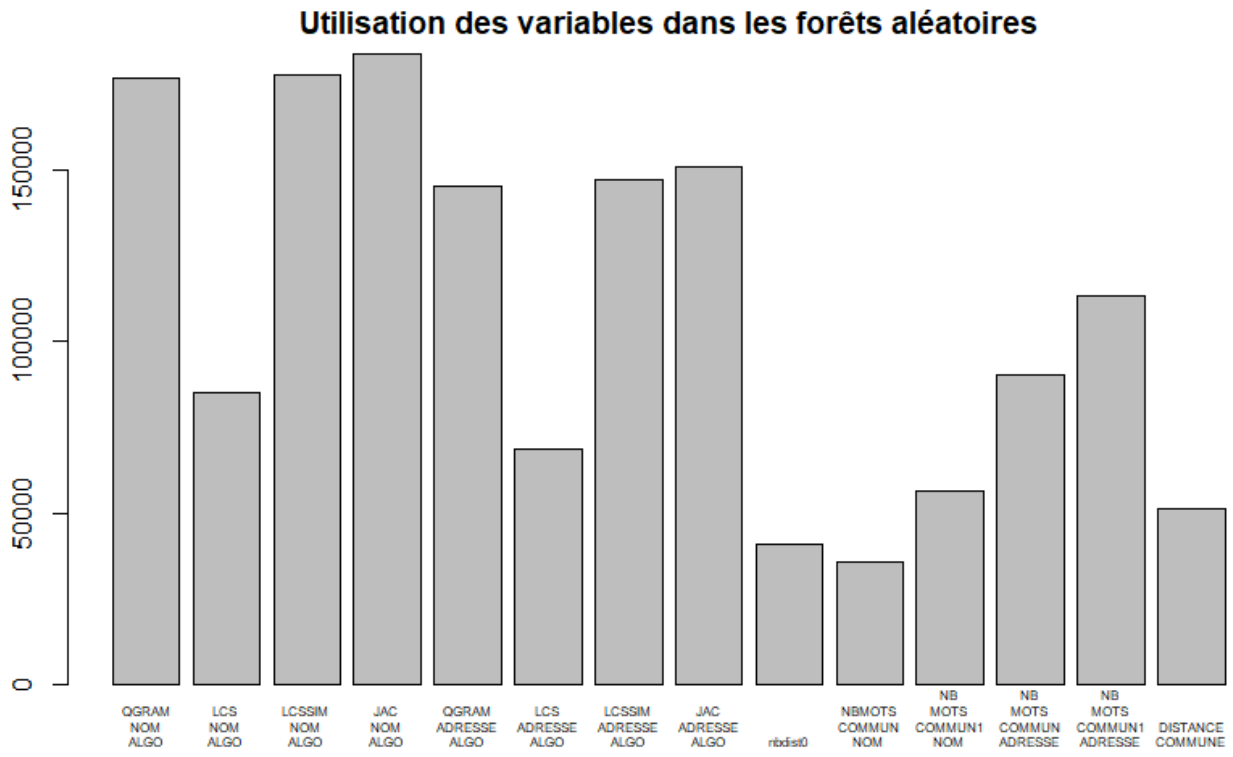


Figure 16 : Utilisation des variables - Modèle sur champs NOM et ADRESSE dans le même département

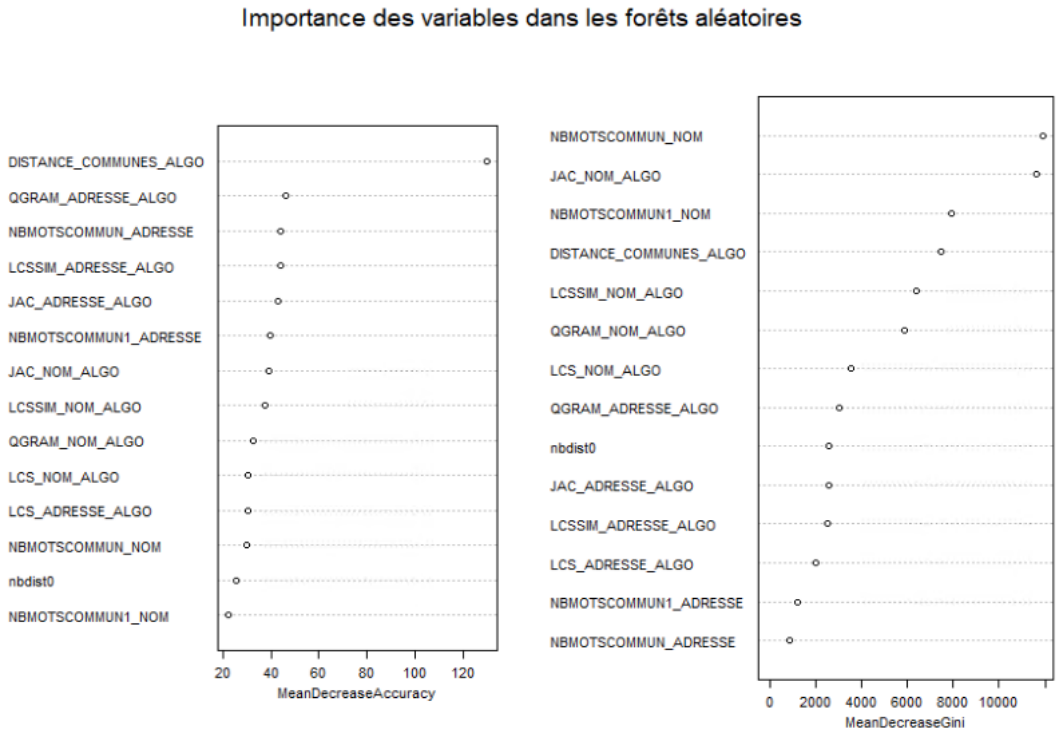


Figure 17 : Importance des variables - Modèle sur champs NOM et ADRESSE dans la même commune

Les matrices de confusion sur les échantillons d'apprentissage et test, à un seuil de 0.5, sont données dans les tableaux 15 et 16. On voit que les taux d'erreur sont sensiblement les mêmes entre les échantillons d'apprentissage et test. On remarque que le taux d'erreur des 0 est autour de 4%, quatre fois celui du modèle dans la même commune, alors que celui des 1 est autour de 20%, plus faible que dans le modèle dans la même commune.

Tableau 15 : Matrice de confusion au seuil 0.5 sur l'échantillon d'apprentissage - Modèle sur champs NOM et ADRESSE dans le même département

Valeur observée	Valeur prédite 0	Valeur prédite 1	Erreur
0	220 598	9 820	4.26%
1	10 327	41 225	20.03%

Tableau 16 : Matrice de confusion au seuil 0.5 sur l'échantillon test - Modèle sur champs NOM et ADRESSE dans le même département

Valeur observée	Valeur prédite 0	Valeur prédite 1	Erreur
0	109 834	5 377	4.67%
1	4 884	20 895	18.95%

Ici aussi, la distribution des scores prédits est dichotomique et certaines lignes où OK = 1 ont un score faible (cf. figures 18 et 19). Les statistiques descriptives des scores sont données dans le tableau 17.

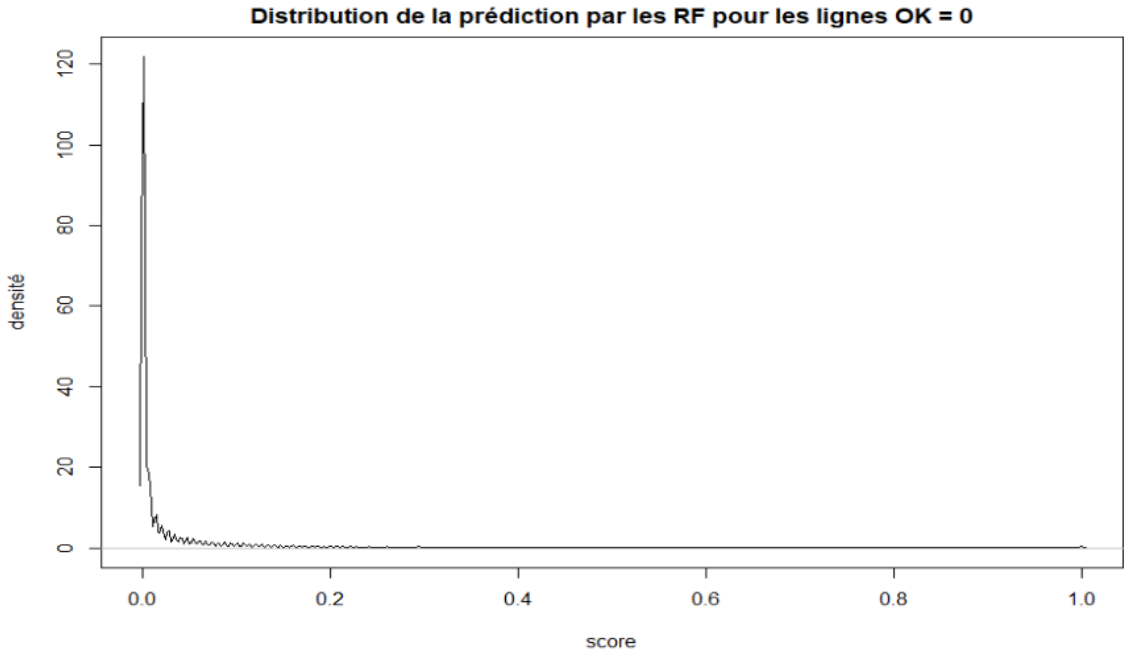


Figure 18 : Distribution de la prédiction sur l'échantillon test pour les lignes OK = 0 - Modèle sur champs NOM et ADRESSE dans le même département

Tableau 17 : Statistiques descriptives des prédictions sur l'échantillon test - Modèle sur champs NOM et ADRESSE dans le même département

OK	Min	Q1	Médiane	Moyenne	Q3	Max
0	0	0	0	0.06	0.02	1
1	0	0.63	0.85	0.74	0.98	1

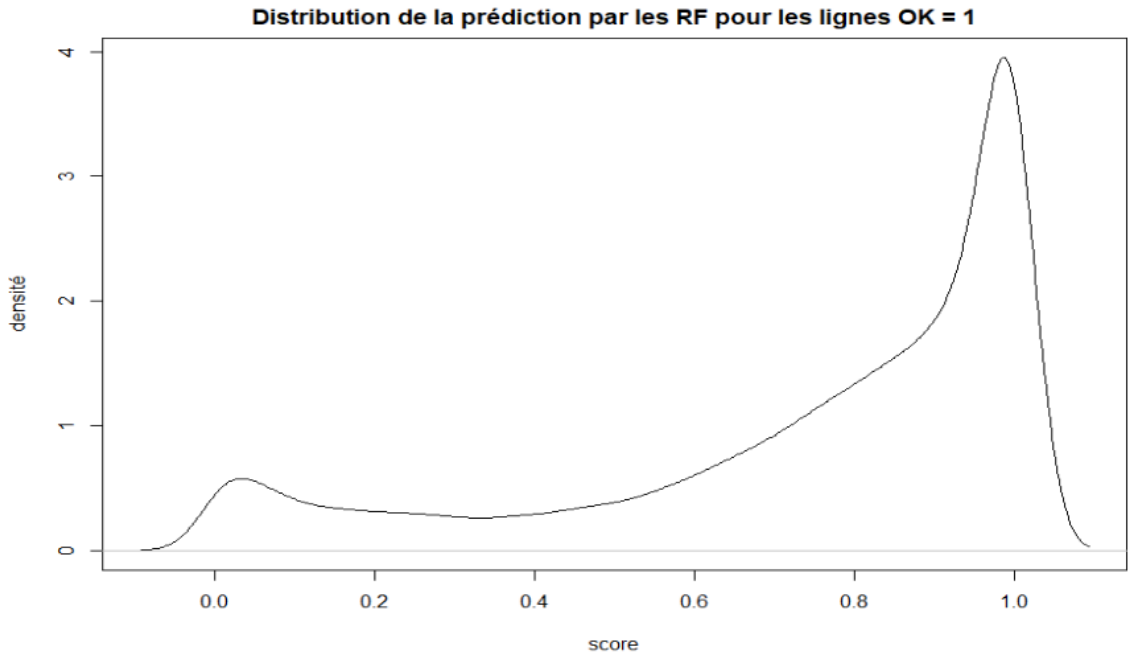


Figure 19 : Distribution de la prédiction sur l'échantillon test pour les lignes OK = 1 - Modèle sur champs NOM et ADRESSE dans le même département

b. Modèle sur les champs NOM uniquement

Toutes les lignes dont les champs NOM sont non vides côté source et Sirene sont conservées.

Nous lançons un modèle de forêts aléatoires sur les variables suivantes :

- QGRAM_NOM_ALGO ;
- LCS_NOM_ALGO ;
- LCSSIM_NOM_ALGO ;
- JAC_NOM_ALGO ;
- nbdistNOM0 ;
- NBMOTSCOMMUN_NOM ;
- NBMOTSCOMMUN1_NOM ;
- DISTANCE_COMMUNES_ALGO.

Les figures 20 et 21 montrent respectivement la fréquence d'utilisation des variables dans les forêts aléatoires et leur importance.

Les matrices de confusion sur les échantillons d'apprentissage et test, à un seuil de 0.5, sont données dans les tableaux 18 et 19. On voit que les taux d'erreur sont sensiblement les mêmes entre les échantillons d'apprentissage et test. On remarque que le taux d'erreur des 0 est autour de 2%, alors que celui des 1 est autour de 21%.

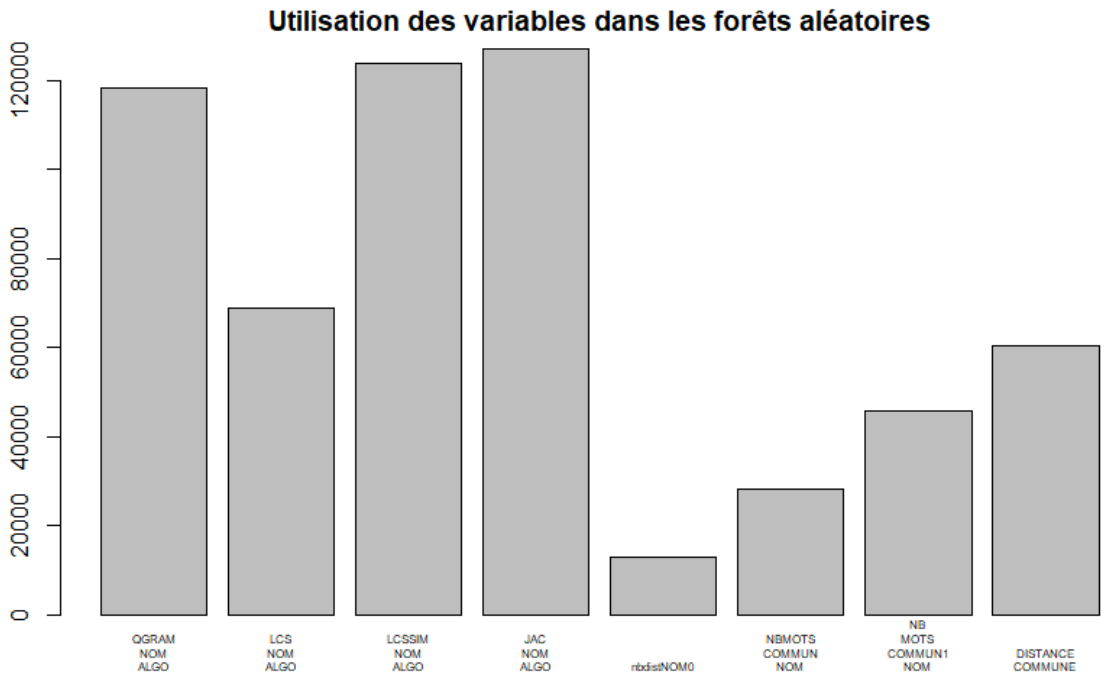


Figure 20 : Utilisation des variables - Modèle sur champs NOM dans le même département

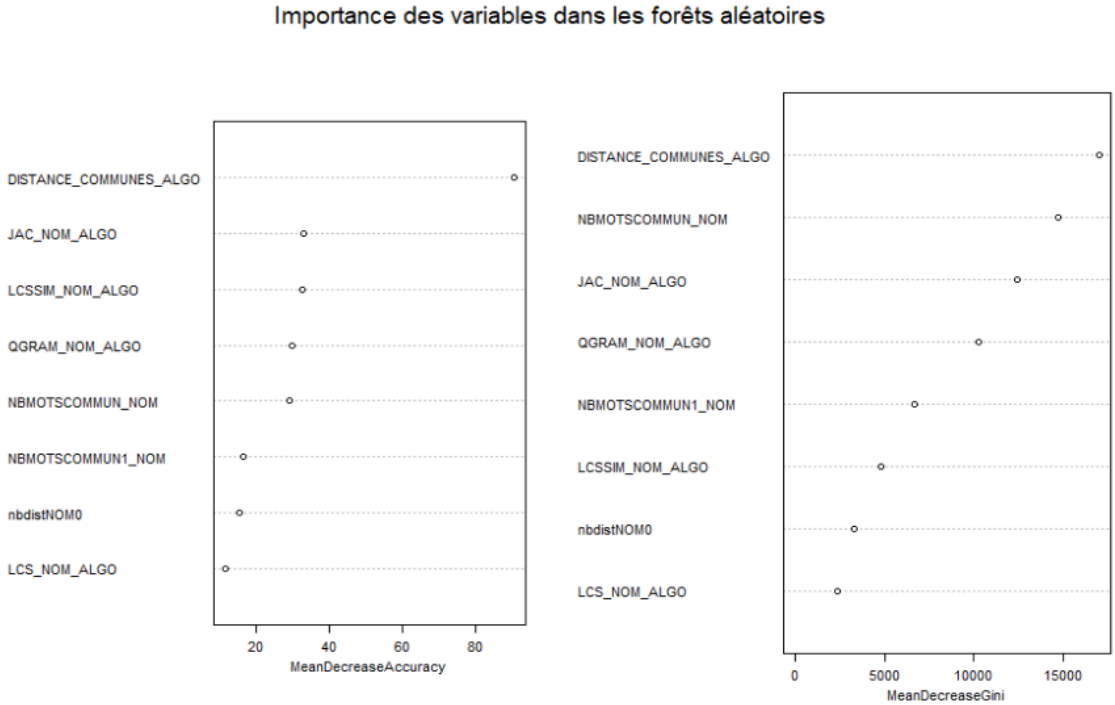


Figure 21 : Importance des variables - Modèle sur champs NOM dans le même département

Tableau 18 : Matrice de confusion au seuil 0.5 sur l'échantillon d'apprentissage - Modèle sur champs NOM uniquement dans le même département

Valeur observée	Valeur prédite 0	Valeur prédite 1	Erreur
0	554 231	12 680	2.24%
1	12 716	46 427	21.50%

Tableau 19 : Matrice de confusion au seuil 0.5 sur l'échantillon test - Modèle sur champs NOM uniquement dans le même département

Valeur observée	Valeur prédite 0	Valeur prédite 1	Erreur
0	276 955	6 704	2.36%
1	6 282	23 288	21.24%

Tout comme dans les modèles précédents, la distribution des scores prédits est dichotomique et certaines lignes où OK = 1 ont un score faible (cf. figures 22 et 23). Les statistiques descriptives des scores sont données dans le tableau 20.

Tableau 20 : Statistiques descriptives des prédictions sur l'échantillon test - Modèle sur champs NOM dans le même département

OK	Min	Q1	Médiane	Moyenne	Q3	Max
0	0	0	0	0.03	0	1
1	0	0.67	0.97	0.76	1	1

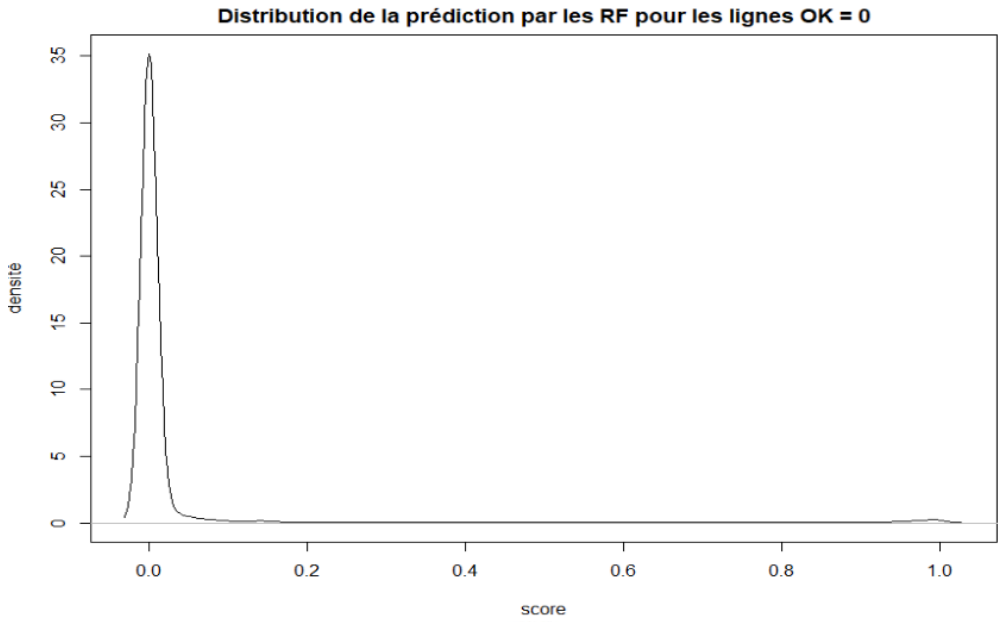


Figure 22 : Distribution de la prédiction sur l'échantillon test pour les lignes OK = 0 - Modèle sur champs NOM uniquement dans le même département

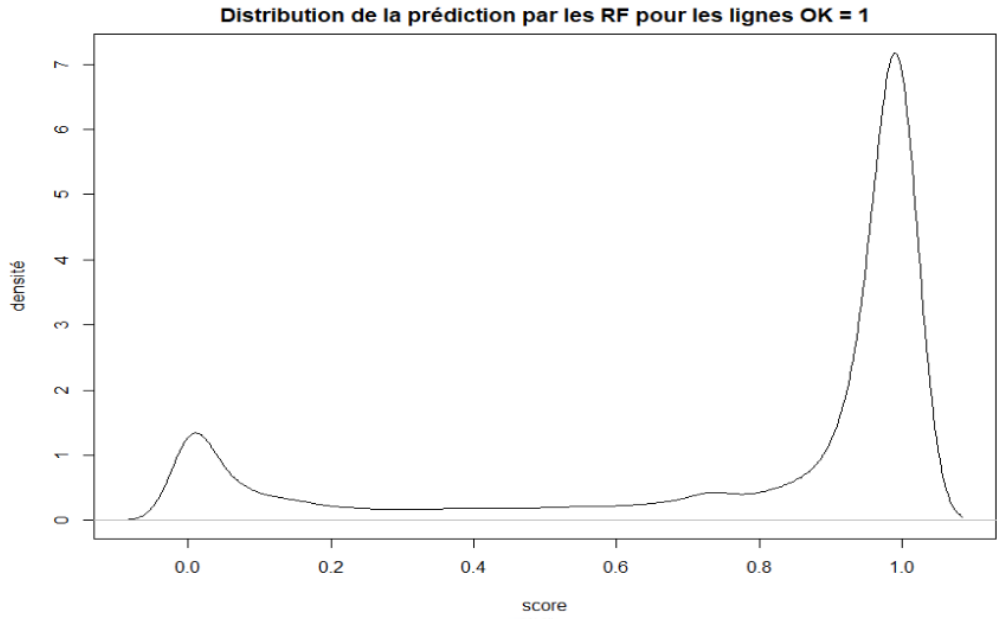


Figure 23 : Distribution de la prédiction sur l'échantillon test pour les lignes OK = 1 - Modèle sur champs NOM uniquement dans le même département

6.5. Choix des seuils

D'après les tableaux de statistiques descriptives 11, 14, 17 et 20, on choisit les itinéraires des seuils ci-dessous pour retourner les lignes Sirene. Dès qu'une ou plusieurs lignes Sirene remplissent les conditions, on sort de la boucle.

Si l'algorithme est d'abord lancé sur les entreprises de la même commune (algorithme *commune + département*) :

- même commune **et** score du modèle rfcna >= 0.6 **et** score du modèle rfcn >= 0.8 (étape retour 1) ;
- même commune **et** (score du modèle rfcna >= 0.7 **ou** score du modèle rfcn >= 0.8) (étape retour 2) ;
- même département **et** score du modèle rfdna >= 0.6 **et** score du modèle rfdn >= 0.8 (étape retour 3) ;
- même département **et** (score du modèle rfdna >= 0.7 **ou** score du modèle rfdn >= 0.8) (étape retour 4) ;
- même département **et** meilleur score du modèle rfdna non nul **et** plus petite distance moyenne (étape retour 5) ;
- même département **et** meilleur score du modèle rfdn non nul **et** plus petite distance moyenne (étape retour 6) ;
- même département **et** (meilleur score du modèle rfdna non nul **ou** meilleur score du modèle rfdn non nul **ou** plus petite distance moyenne) (étape retour 7).

Si l'algorithme est lancé directement sur les entreprises du même département (algorithme *département*) :

- même département **et** score du modèle rfdna > 0.8 (étape retour 1) ;
- même département **et** score du modèle rfdna > 0.6 **et** score du modèle rfdn > 0.8 (étape retour 2) ;
- même département **et** (score du modèle rfdna > 0.6 **ou** score du modèle rfdn > 0.8) (étape retour 3) ;
- même département **et** score du modèle rfdna parmi les 3 meilleurs et > 0.2 **et** score du modèle rfdn parmi les 3 meilleurs et > 0.2 (étape retour 4) ;
- même département **et** ((score du modèle rfdna parmi les 3 meilleurs et > 0.2) **ou** (score du modèle rfdn parmi les 3 meilleurs et > 0.2)) (étape retour 5) ;
- même département **et** meilleur score du modèle rfdna non nul (étape retour 6) ;
- même département **et** meilleur score du modèle rfdn non nul (étape retour 7) ;

7. Résultats de l'algorithme sur les échantillons d'apprentissage et test

La procédure complète est appliquée aux échantillons d'apprentissage et test définis en début de développement.

Pour chaque ligne de chaque échantillon :

- extraction des lignes Sirene de la même commune ou du même département (s'il n'y pas d'établissements dans la même commune ou si le choix est fait de lancer l'algorithme au niveau département directement) ;
- calcul des distances sur les champs NOM et ADRESSE entre la ligne source et les lignes Sirene ;
- calcul de la moyenne des distances non nulles ;
- restriction aux lignes Sirene correspondant aux 25 plus petites distances moyennes ;
- calcul des indicateurs complémentaires ;

- application des modèles de forêts aléatoires adéquats (même commune ou même département) sur les lignes retenues ;
- application de l'itinéraire des seuils aux scores obtenus selon l'algorithme *commune + département* ou *département*.

Les résultats sont ensuite comparés pour vérifier qu'il n'y a pas d'effet lié à l'échantillonnage.

Dans un premier temps, le nombre de résultats retournés par individu est étudié, pui on regarde à quelle étape de l'itinéraire des seuils sont retournées les résultats, ce qui donne une idée de la fiabilité des réponses. Enfin, le croisement nombre de résultats et étape retour est étudié. Dans chaque cas, le pourcentage de numéros SIRET correctement retrouvés est calculé.

Lorsque pour une même ligne source, plusieurs lignes Sirene sont retournées avec des scores ou étapes retour identiques, il est utile de mobiliser des variables complémentaires afin de privilégier un numéro SIRET parmi les autres (statut actif/fermé, mêmes codes NAF, etc.).

Les statistiques des tables 21, 23, 24, 26, 28 et 29 sont calculées uniquement sur les individus dont le numéro SIRET a été initialement retrouvé dans la base Sirene lors de l'appariement des deux bases (cf. partie 4.1).

7.1. Résultats de l'algorithme commune + département

Pour rappel, cet algorithme recherche dans un premier temps les lignes Sirene dans la même commune et retourne uniquement celles avec un score élevée (les candidats les plus probables). Si aucune ligne n'est retournée ou s'il n'existe pas d'établissements dans la même commune, la recherche est faite dans le département correspondant selon l'itinéraire des seuils décrit dans la partie 6.5.

7.1.1. Nombre de résultats retournés par individu

Tableau 21 : Nombre de résultats retournés par individu initialement retrouvé dans la base Sirene - Algorithme commune + département

Nombre de résultats	% individus		% numéros SIRET retrouvés	
	Apprentissage	Test	Apprentissage	Test
0	1.7	1.7	-	-
1	79.0	76.7	84.4	81.6
2	13.2	14.9	68.6	68.9
3	4.3	4.8	48.2	48.9
4	1.0	1.1	38.8	38.7
>=5	0.8	0.8	25.9	29.8
Tout	100.0	100.0	78.4	75.9

On note que plus le nombre de résultats retournés par individu est important, plus la probabilité que la ligne Sirene fasse partie des résultats est faible.

Dans les échantillons d'apprentissage et test, environ 2% des individus n'ont aucun résultat retourné alors que le numéro SIRET existe dans la base Sirene. Il s'agit en majorité d'individus dont les noms et adresses ne correspondent pas entre les deux bases pour un même numéro SIRET, la distance moyenne des indicateurs est donc trop élevée pour faire partie des lignes sélectionnées avant application du modèle prédictif. On trouve également des individus où la ligne Sirene ayant le bon numéro SIRET est sélectionnée dans la partie « commune » mais le score prédit est trop faible pour être retourné, et une fois l'algorithme lancé sur l'ensemble du département, la ligne n'est plus sélectionnée car il existe davantage de lignes Sirene avec des distances moyennes plus faibles. Pour ces lignes, le score des modèles prédictifs appliqués aux lignes sélectionnées est nul, aucune ligne n'est donc retournée.

Dans l'échantillon d'apprentissage (respectivement l'échantillon test), environ 79% (resp. 77%) des individus initialement retrouvés dans la base Sirene ont un seul résultat retourné et dans 84% (resp. 82%) des cas, il s'agit du bon numéro SIRET. Lorsque le numéro SIRET n'a pas été retrouvé mais un seul résultat a été retourné, il s'agit d'individus pour lesquels l'algortihme privilégie un numéro SIRET plus probable en termes de noms et adresses. Par exemple, le bon numéro SIRET peut avoir une adresse manquante dans la base Sirene alors qu'il existe une autre ligne Sirene avec l'adresse renseignée et exacte, c'est cette ligne qui sera retournée.

Dans la plupart des cas où un ou plusieurs résultats sont retournés mais pas le bon numéro SIRET, c'est parce que les champs NOM et ADRESSE sont trop différents entre les bases source et Sirene et qu'il existe des lignes Sirene plus probables. Les statistiques pour les individus retrouvés dans la base Sirene avec une MEAN_DIST inférieure à 0.4 sont données dans la table 22.

Tableau 22 : Nombre de résultats retournés par individu initialement retrouvé dans la base Sirene avec une MEAN_DIST $<$ 0.4 - Algorithme commune + département

Nombre de résultats	% individus		% numéros SIRET retrouvés	
	Apprentissage	Test	Apprentissage	Test
0	1.0	1.1	-	-
1	81.3	78.8	88.1	85.3
2	12.6	14.4	77.5	76.4
3	3.7	4.1	61.2	61.3
4	0.8	0.9	51.4	50.6
>=5	0.6	0.7	37.2	39.2
Tout	100.0	100.0	84.3	81.5

Une solution pour retrouver davantage de numéros SIRET dans les résultats retournés serait d'augmenter le nombre de lignes sélectionnées selon la distance moyenne (cf. partie 8).

7.1.2. Fiabilité des résultats retournés par individu

Tableau 23 : Etapes retour par individu initialement retrouvé dans la base Sirene - Algorithme commune + département

Etape retour	% individus		% numéros SIRET retrouvés	
	Apprentissage	Test	Apprentissage	Test
0	1.7	1.7	-	-
1	54.1	52.2	90.9	87.2
2	18.8	19.1	83.3	83.9
3	1.6	1.6	87.2	82.5
4	3.6	3.1	77.9	72.7
5	5.8	7.0	63.6	63.2
6	4.4	4.5	39.7	47.4
7	10.0	10.7	40.0	38.8
Tout	100.0	100.0	78.4	75.9

D'après le tableau 23, environ 78% des individus de l'échantillon d'apprentissage (76% de l'échantillon test) initialement retrouvés dans la base Sirene ont leurs résultats retournés aux étapes 1 à 4, correspondant à des scores des modèles prédictifs plus élevés. Pour ces individus, le bon numéro SIRET fait partie des résultats entre 78% et 91% (entre 73% et 87% pour l'échantillon test). Les résultats retournés aux étapes 5 à 7, correspondant à des scores plus faibles des modèles prédictifs, contiennent plus rarement le bon numéro SIRET que les lignes retournées aux étapes 1 à 4.

Dans le tableau 24, les résultats sont agrégés pour les étapes 1 à 4 et 5 à 7 selon le nombre de résultats. Pour 70% des individus de l'échantillon d'apprentissage initialement retrouvés dans la base Sirene (66% de l'échantillon test, respectivement), un seul résultat est retourné entre les étapes 1 à 4 (score $>= 0.6$) et pour 89% d'entre eux (86%, respectivement) le bon numéro SIRET est retrouvé.

En appliquant cet algorithme à un nouveau jeu de données, si un seul résultat est retourné entre les étapes 1 à 4, on peut dire que c'est le bon numéro SIRET à 86% (valeur correspondant à l'échantillon test). Attention à ne pas oublier que ces valeurs sont obtenues avec les données de l'INAO et les défauts/erreurs de renseignement des noms et adresses qui leur sont propres.

Par exemple, le nombre de résultats retournés et les étapes retour associées des individus des échantillons d'apprentissage et test ayant un numéro SIRET renseigné mais initialement **non retrouvés** dans la base Sirene sont donnés dans la table 25. Pour 7% des individus, aucun résultat n'est retourné, mais pour environ 45% des individus, un seul résultat est retourné avec un score élevé entre les étapes 1 à 4.

Tableau 24 : Nombre de résultats et étapes retournés par individu initialement retrouvé dans la base Sirene - Algorithme commune + département

Nombre de résultats	Etape retour	% individus		% numéros SIRET retrouvés	
		Apprentissage	Test	Apprentissage	Test
0	-	1.7	1.7	-	-
1	**1-4**	69.5	65.9	88.7	85.7
1	**5-7**	9.4	10.8	52.9	57.1
2	**1-4**	7.2	8.4	86.5	87.3
2	**5-7**	6.0	6.5	47.1	45.2
3	**1-4**	1.0	1.3	83.0	83.8
3	**5-7**	3.3	3.5	38.2	36.3
4	**1-4**	0.2	0.3	77.3	67.7
4	**5-7**	0.8	0.8	27.7	28.4
>=5	**1-4**	0.1	0.2	56.4	65.2
>=5	**5-7**	0.7	0.6	20.2	20.7
Tout	**Tout**	100.0	100.0	78.4	75.9

Tableau 25 : Nombre de résultats et étapes retournés par individu initialement non retrouvé dans la base Sirene - Algorithme commune + département

Nombre de résultats	Etape retour	% individus	
		Apprentissage	Test
0	-	6.6	6.5
1	**1-4**	45.6	44.9
1	**5-7**	16.2	15.9
2	**1-4**	6.2	6.4
2	**5-7**	11.0	12.4
3	**1-4**	0.9	0.9
3	**5-7**	9.1	7.7
4	**1-4**	0.3	0.3
4	**5-7**	2.3	2.7
>=5	**1-4**	0.2	0.2
>=5	**5-7**	1.6	2.1
Tout	**Tout**	100.0	100.0

7.2. Résultats de l'algorithme département

Pour rappel, cet algorithme recherche les lignes Sirene directement dans l'ensemble du département et pas dans la même commune en premier lieu. La distance euclidienne entre les communes est ajoutée au modèle prédictif.

7.2.1. Nombre de résultats retournés par individu

Tout comme dans l'algorithme *commune + département*, plus le nombre de résultats retournés par individu est important, plus la probabilité que la ligne Sirene fasse partie des résultats est faible. De même, aucun résultat n'est retourné pour environ 3% des individus (cf. tableau 26), pour lesquels la distance moyenne des indicateurs est trop élevée pour être sélectionnée parmi les lignes où le modèle prédictif sera appliqué.

Pour 85% des individus de l'échantillon d'apprentissage initialement retrouvés dans la base Sirene (84% pour l'échantillon test, respectivement), un seul résultat est retourné, ce qui est plus élevé qu'avec l'algorithme *commune+ département*. Dans 76% des cas (72%, respectivement), le bon numéro SIRET est retourné, ce qui est cependant plus faible qu'avec l'algorithme précédent. Cela vient principalement du fait, une fois de plus, de la distance moyenne trop élevée comparée à celles calculées sur l'ensemble du département. Les statistiques pour les individus retrouvés dans la base Sirene avec une MEAN_DIST inférieure à 0.4 sont données dans la table 27.

Tableau 26 : Nombre de résultats retournés par individu initialement retrouvé dans la base Sirene - Algorithme département

Nombre de résultats	% individus		% numéros SIRET retrouvés	
	Apprentissage	Test	Apprentissage	Test
0	3.1	3.1	-	-
1	85.2	84.0	76.0	72.2
2	9.3	10.2	70.7	70.5
3	1.6	1.9	56.7	60.3
4	0.4	0.5	50.5	48.5
>=5	0.4	0.3	35.5	45.5
Tout	100.0	100.0	72.6	69.4

Tableau 27 : Nombre de résultats retournés par individu initialement retrouvé dans la base Sirene avec une MEAN_DIST $<$ 0.4 - Algorithme département

Nombre de résultats	% individus		% numéros SIRET retrouvés	
	Apprentissage	Test	Apprentissage	Test
0	2.3	2.3	-	-
1	86.3	84.9	81.3	77.4
2	9.2	10.2	77.2	76.2
3	1.5	1.8	66.0	67.5
4	0.4	0.5	60.0	54.7
>=5	0.3	0.3	45.4	53.3
Tout	100.0	100.0	78.6	75.2

Ici aussi, une solution pour retrouver davantage de numéros SIRET dans les résultats retournés serait d'augmenter le nombre de lignes sélectionnées selon la distance moyenne (cf. partie 8).

7.2.2. Fiabilité des résultats retournés par individu

Tableau 28 : Etapes retour par individu initialement retrouvé dans la base Sirene - Algorithme département

Etape retour	% individus		% numéros SIRET retrouvés	
	Apprentissage	Test	Apprentissage	Test
0	3.1	3.1	-	-
1	49.6	44.4	90.9	85.8
2	7.5	11.1	80.1	82.7
3	15.4	15.2	81.5	82.7
4	4.7	6.1	62.5	67.6
5	7.3	6.5	62.7	52.2
6	8.3	9.4	11.9	17.0
7	4.1	4.2	9.7	10.4
Tout	100.0	100.0	72.6	69.4

D'après le tableau 28, environ 73% des individus de l'échantillon d'apprentissage (71% de l'échantillon test) initialement retrouvés dans la base Sirene ont leurs résultats retournés aux étapes 1 à 3, correspondant à des scores des modèles prédictifs plus élevés. Pour ces individus, le bon numéro SIRET fait partie des résultats entre 82% et 91% (entre 83% et 86% pour l'échantillon test). Les résultats retournés aux étapes 6 et 7, correspondant à des scores plus faibles des modèles prédictifs (meilleurs scores non nuls inférieurs à 0.2), contiennent plus rarement le bon numéro SIRET que les lignes retournées aux étapes 1 à 3 ou encore 4 et 5.

Dans le tableau 29, les résultats sont agrégés pour les étapes 1 à 5 et 6 à 7 selon le nombre de résultats. Pour 75% des individus de l'échantillon d'apprentissage initialement retrouvés dans la base Sirene (72% de l'échantillon test, respectivement), un seul résultat est retourné entre les étapes 1 à 5 (score $>= 0.6$) et pour 85% d'entre eux (81%, respectivement) le bon numéro SIRET est retrouvé.

En appliquant cet algorithme à un nouveau jeu de données, si un seul résultat est retourné entre les étapes 1 à 5, on peut dire que c'est le bon numéro SIRET à 81% (valeur correspondant à l'échantillon test). Attention à ne pas oublier que ces valeurs sont obtenues avec les données de l'INAO et les défauts/erreurs de renseignement des noms et adresses qui leur sont propres.

Par exemple, le nombre de résultats retournés et les étapes retour associées des individus des échantillons d'apprentissage et test ayant un numéro SIRET renseigné mais initialement **non retrouvés** dans la base Sirene sont donnés dans la table 30. Pour environ 8% des individus, aucun résultat n'est retourné, mais pour 53% des individus, un seul résultat est retourné avec un score élevé entre les étapes 1 à 5.

Tableau 29 : Nombre de résultats et étapes retournés par individu initialement retrouvé dans la base Sirene - Algorithme département

Nombre de résultats	Etape retour	% individus		% numéros SIRET retrouvés	
		Apprentissage	Test	Apprentissage	Test
0	-	3.1	3.1	-	-
1	**1-5**	74.7	72.3	85.0	81.4
1	**6-7**	10.5	11.7	11.7	15.4
2	**1-5**	8.0	8.8	80.5	79.6
2	**6-7**	1.3	1.4	9.0	13.3
3	**1-5**	1.3	1.6	68.7	68.8
3	**6-7**	0.3	0.3	8.5	13.1
4	**1-5**	0.3	0.4	69.9	64.2
4	**6-7**	0.1	0.1	4.0	3.8
>=5	**1-5**	0.2	0.2	60.2	62.6
>=5	**6-7**	0.1	0.1	2.1	6.8
Tout	**Tout**	100.0	100.0	72.6	69.4

Tableau 30 : Nombre de résultats et étapes retournés par individu initialement non retrouvé dans la base Sirene - Algorithme département

Nombre de résultats	Etape retour	% individus	
		Apprentissage	Test
0	-	7.6	7.5
1	**1-5**	53.6	53.1
1	**6-7**	23.7	24.0
2	**1-5**	7.8	8.1
2	**6-7**	3.1	3.5
3	**1-5**	2.0	1.9
3	**6-7**	0.8	0.8
4	**1-5**	0.6	0.5
4	**6-7**	0.2	0.2
>=5	**1-5**	0.3	0.3
>=5	**6-7**	0.3	0.1
Tout	**Tout**	100.0	100.0

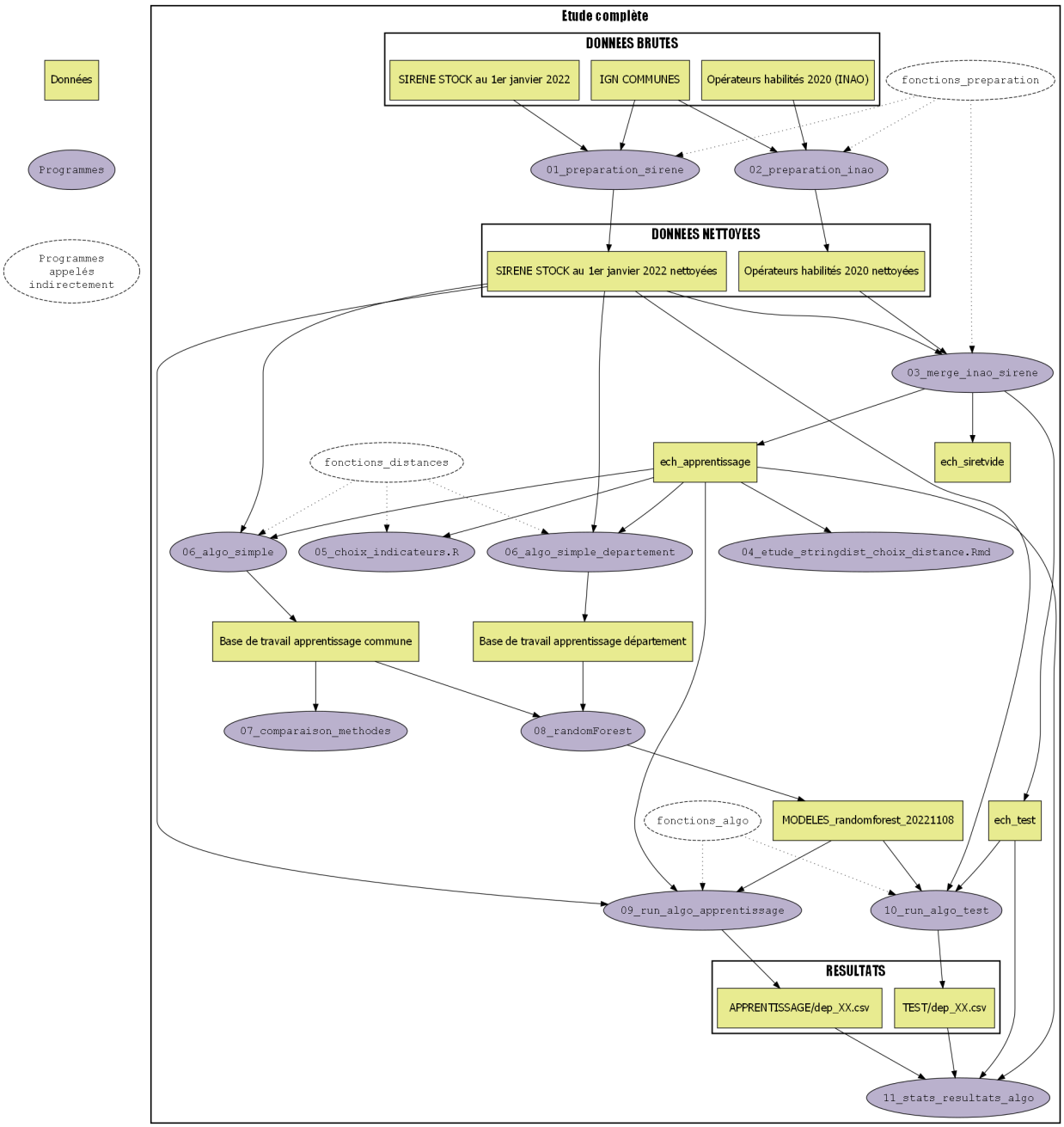
8. Suite des travaux

Plusieurs pistes d'amélioration de la procédure de sirétisation sont envisagées :

- augmenter le nombre de lignes Sirene retenues selon la distance moyenne (en cours) ;
- mieux identifier pourquoi certaines lignes retrouvées dans la base Sirene ne sont pas retournées ;
- pouvoir filtrer en amont les entreprises Sirene du même code NAF quand l'information est disponible dans les données source, ou selon le statut actif/fermé à une date donnée ;
- utiliser un modèle prédictif prenant en compte les données enflées en 0 (pour prendre en compte qu'il y a plus de 0 que de 1 lors de l'apprentissage) ;
- intégrer l'outil de vérification des numéros SIRET valides à la procédure (règle de calcul) : identifier les lignes du fichier source où le siret est renseigné, se compose de 14 chiffres mais n'est pas valide ;
- intégrer l'appariement avec les données Sirene non diffusibles (extraction des numéros SIRET possible par API) : identifier les lignes qu'on ne pourra pas retrouver dans la base Sirene ;
- tester la procédure sur les lignes INAO où le numéro SIRET était manquant (fichier *ech_siretvide.csv*) et avec d'autres jeux de données.

Annexes

Annexe 1. Workflow du développement de l'algorithme de sirétisation



Tifenn CORRE (tifenn.corre@inrae.fr)

Thomas POMÉON (thomas.pomeon@inrae.fr)

Julie REGOLO (julie.regolo@inrae.fr)

INRAE, Centre Occitanie-Toulouse
Unité de Service de l'Observatoire du Développement Rural (0685)
24 chemin de Borde Rouge, Auzeville – CS 52627
31326 Castanet-Tolosan cedex France

odr.inrae.fr

Observatoire du Developpement Rural
ODR