Impact du nettoyage des données sur la qualité d'un appariement

Insee Mesurer pour comprendre

PROBLÉMATIQUE : LA CURIOSITÉ D'UNE BÉOTIENNE

« La préparation des données, c'est 80 % de la réussite d'un appariement / du temps de travail » (différentes personnes expérimentées dans le domaine, 2021)

- **80 % de quoi ? Comment mesure-t-on la réussite ?**
- **80 % c'est beaucoup !**

« Data cleaning made little difference to the overall linkage quality, with heavy cleaning leading to a decrease in quality. […] Data cleaning techniques have minimal effect on linkage quality. Care should be taken during the data cleaning process. » (Randall, 2013)

- **Minimal ?**

Essayer de mesurer l'impact d'un nettoyage plus ou moins poussé dans différents contextes sur le taux d'appariement et sur la précision

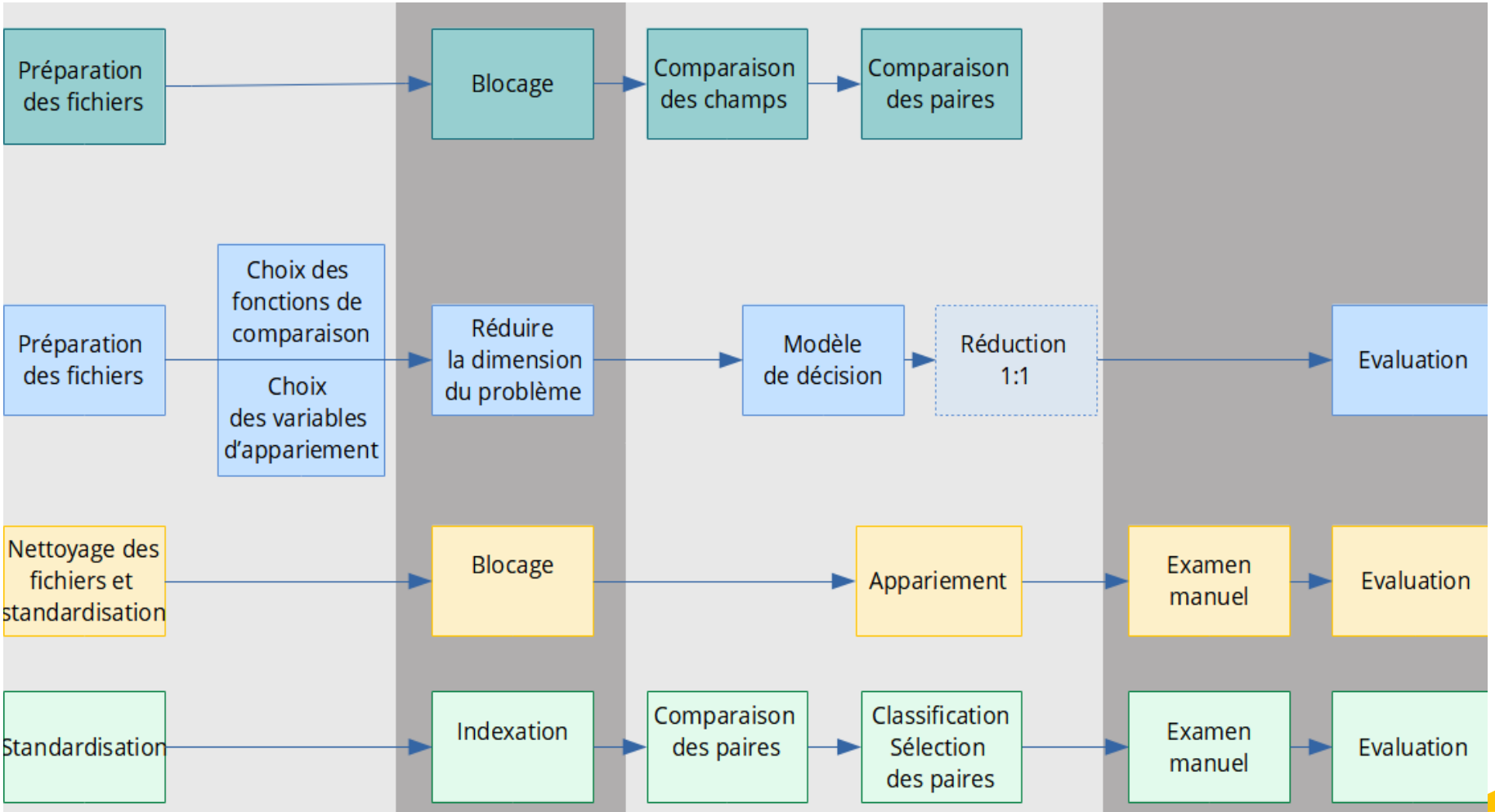
1 NETTOYAGE DES DONNÉES : DE QUOI PARLE-T-ON ?

2 DONNÉES UTILISÉES ET SCÉNARIOS DE TEST

3 RÉSULTATS

01 NETTOYAGE DES DONNÉES : DE QUOI PARLE-T-ON ?

ON S'INTÉRESSE À LA PREMIÈRE PARTIE DU PROCESSUS D'APPARIEMENT



OBJECTIFS DE LA PRÉPARATION DES DONNÉES

Constituer un ensemble de variables comparables

Nettoyer les données

Identifier les variables à utiliser lors du processus d'appariement

La préparation des données n'est pas uniquement une étape de modification des données. On pourrait proposer de la voir en quatre temps :

- Une phase d'observation et de qualification des données : concepts, format des variables, distributions, contrôles formels, conformité à une nomenclature
- Une phase de modification des données : formatage, traitement des valeurs manquantes, aberrantes, refuge ; codification etc.
- Une phase de prise de décision sur le processus d'appariement
- Une phase de documentation, et notamment la production d'indicateurs à l'issue des phases d'observation et de modification

On peut aussi décomposer cette étape en plusieurs types de tâches différentes liées au nettoyage des données

- Formatage : même format pour des données identiques (dates, sexe), gestion des valeurs manquantes ou partielles
- Codification : géographie, nomenclatures
- Segmentation : nom comportant civilité et relations, adresses
- Traitement des anomalies : valeurs conformes au format mais ne portant pas d'informations, ou informations erronées

02 DONNÉES UTILISÉES ET SCÉNARIOS DE TESTS

Enquête annuelle de recensement (EAR) 2019

- Enquête bi modale papier/internet en population générale

Fichier d'Imposition des personnes (FIP) 2019

- Pas d'informations nominatives sur les moins de 15 ans

Fichier Tous salariés (produit par le DERA à partir de plusieurs sources) 2019

- Personnes salariées

Les données pour les départements de la Lozère et des Hauts-de-Seine ont été utilisées

ÉLEMENTS SUR LES VARIABLES D'APPARIEMENT
NOMS ET PRÉNOMS

	EAR	FIP	Tous salariés
Nombre de mots de la variable nom ou nom de naissance			
0	0,7 %	0,0 %	0,0 %
1	80,3 %	90,0 %	88,5 %
2	**10,9 %**	7,9 %	9,5 %
3	**7,0 %**	1,5 %	1,6 %
4 ou plus	1,1 %	0,5 %	0,4 %
Proportion de nom marital non vide		**50,9 %**	**23,6 %**
Nombre de mots du prénom			
0	1,4 %	0,0 %	0,0 %
1	89,9 %	**73,3 %**	89,9 %
2	7,9 %	**17,5 %**	9,0 %
3	0,7 %	**8,1 %**	0,9 %
4 ou plus	0,1 %	1,0 %	0,1 %

ÉLEMENTS SUR LES VARIABLES D'APPARIEMENT
DATES

Valeur	EAR	FIP	Tous salariés
Date incomplète ou non valide			
0000		0,22 %	
Mois+00		Quelques valeurs	
Mois+99			Quelques valeurs
WWWW			0,06 %
..			0,30 %
Mois sans jour	Quelques valeurs		
Jour sans mois	Quelques valeurs		
Pics de distribution des dates			
1er janvier	1,09 %	1,47 %	0,88 %
31 décembre	0,45 %	0,45 %	0,4 %
jours « ordinaires »	De 0,22 % à 0,32 %	De 0,24 % à 0,32 %	De 0,23 % à 0,32 %

ÉLEMENTS SUR LES VARIABLES D'APPARIEMENT
LIEU DE NAISSANCE

	EAR	FIP	Tous salariés
Valeurs manquantes	33,0 %	0,0 %	2,2 %
Valeurs non conformes au COG	0,4 %	3,2 %	7,3 %

TROIS NIVEAUX DE NETTOYAGE DES DONNÉES

	Niveau 0	Niveau 1	Niveau 2
Formatage	Remplacement des caractères non alphabétiques dans les noms et prénoms (tous) Formatage identique pour les dates et le genre	Homogénisation des valeurs partielles de dates	
Codification		Lieu de naissance à l'étranger dans l'EAR	
Segmentation		Repérage des civilités, et relations (épouse, veuve, née...)	
Réduction du bruit		Suppression des mots de 1 lettre, sans voyelles, initiales Ajout du tiret dans certains prénoms composés	Utilisation d'un dictionnaire de noms et prénoms

Deux scénarios assez frustes sur le département de la Lozère (EAR X FIP, et FIP X Tous salariés)

- Appariement exact
  - sur nom de naissance, prénoms, date et lieu de naissance
  - Sur nom de naissance, premier prénom, date et département de naissance
- Appariement déterministe flou (inclusion de 3grams sur les variables nominatives)
  - sur nom de naissance, prénoms, date et lieu de naissance
  - Sur nom de naissance, premier prénom, date et département de naissance

Un scénario avec processus plus abouti sur le département des Hauts de Seine

- Identification au RNIPP (130M de personnes) par l'algorithme du CSNS (plusieurs étapes successives : une étape exacte, relâches simples, valeurs approchées)

03 RÉSULTATS

RÉSULTATS : TAUX D'APPARIEMENT POUR LES SCÉNARIOS SUR UN MÊME DÉPARTEMENT

		Jeux de variables utilisé					
		Tous les prénoms et lieux de naissance sur 5 positions (+ nom et date de naissance)			Premier prénom et lieu de naissance sur 2 positions (+ nom et date de naissance)		
		Niveaux de nettoyage des données					
	Fichiers	Niv. 0	Niv. 1	Niv. 2	Niv. 0	Niv. 1	Niv. 2
Appariement exact	EAR 48 / FIP 48	29,6 %	39,0 %	41,1 %	38,8 %	51,3 %	54,0 %
Appariement déterministe flou	EAR 48 / FIP 48	42,0 %	47,1 %	48,5 %	48,9 %	55,6 %	57,2 %
Appariement déterministe flou	**EAR 48 / FIP 48**	**42,0 %**	**47,1 %**	**48,5 %**	**48,9 %**	**55,6 %**	**57,2 %**
Appariement exact	Tous salariés 48 / FIP 48	56,1 %	57,7 %	57,9 %	71,4 %	72,7 %	72,8 %

RÉSULTATS : SCÉNARIOS D'APPARIEMENT SUR UN MÊME DÉPARTEMENT

Le nettoyage des données permet d'augmenter le taux d'appariement

Le choix des variables d'appariement a un effet d'ampleur similaire au nettoyage des données

L'effet est faible pour les fichiers administratifs, et plus important lorsqu'on utilise les données d'enquête

La relâche des contraintes a également un effet important

Les trois types d'action (nettoyage, choix des variables, relâche des contraintes) voient leurs effets atténués par les autres actions (avec un avantage au bon choix des variables)

Les premières analyses laissent penser que les faux positifs sont quasi nuls. Toutefois on est loin d'un taux d'appariement satisfaisant

RÉSULTATS : TAUX D'APPARIEMENT POUR LES DIFFÉRENTS SCÉNARIOS

		Jeux de variables utilisé		
		Tous les prénoms et lieux de naissance sur 5 positions (+ nom et date de naissance)		
		Niveaux de nettoyage des données		
	Fichiers	Niv. 0	Niv. 1	Niv. 2
Identification via le CSNS	EAR 92	94,2 %	94,4 %	93,7 %
Identification via le CSNS	FIP 92	99,4 %	99,3 %	98,5 %
Identification via le CSNS	Tous salariés 92	99,0 %	99,0 %	98,2 %

CHANGEMENT D'ÉTAPE D'IDENTIFICATION POUR LE FICHIER TOUS SALARIÉS 92

	Norme 0	Norme 1	Norme 2
Identifiés	888 382 (99,02 %)	888 244 (99,00 %)	880 706 (98,16%)
Non identifiés	8 811	8 949	16 487

Etape d'appariement exacte	32,1 %	**32,1 %**	**22,9 %**
Etapes par relâche simple	64,3 %	**64,2 %**	**69,3 %**
Etape par valeur approchée	2,7 %	**2,7 %**	**6,0 %**
Non identifiés	1,0 %	**1,0 %**	**1,8 %**

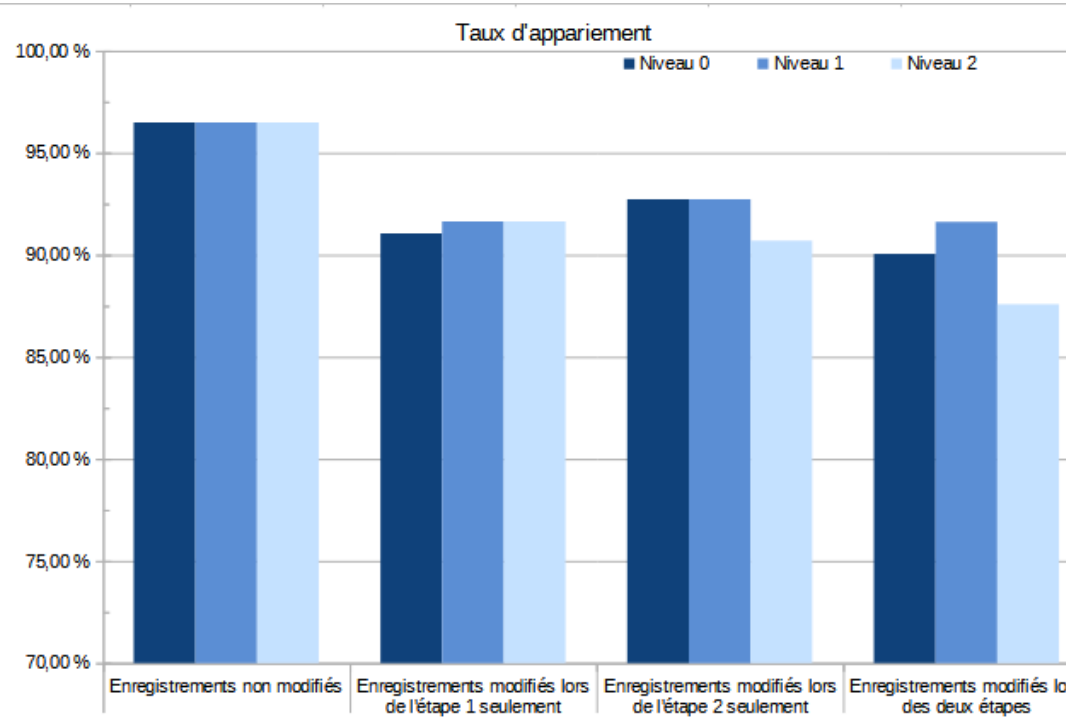
CHANGEMENT D'ÉTAPE D'IDENTIFICATION POUR LE FICHIER DE L'EAR 92

	Norme 0	Norme 1	Norme 2
Identifiés	118 630 (94,2 %)	118 994 (94,4 %)	118 103 (93,7 %)
Non identifiés	7 360	6 996	7 887
Etape d'appariement exacte	10,5 %	18,3 %	13,7 %
Etapes par relâche simple	66,9 %	63,3 %	65,3 %
Etape par valeur approchée	16,8 %	12,9 %	14,7 %
Non identifiés	5,8 %	5,5 %	6,3 %

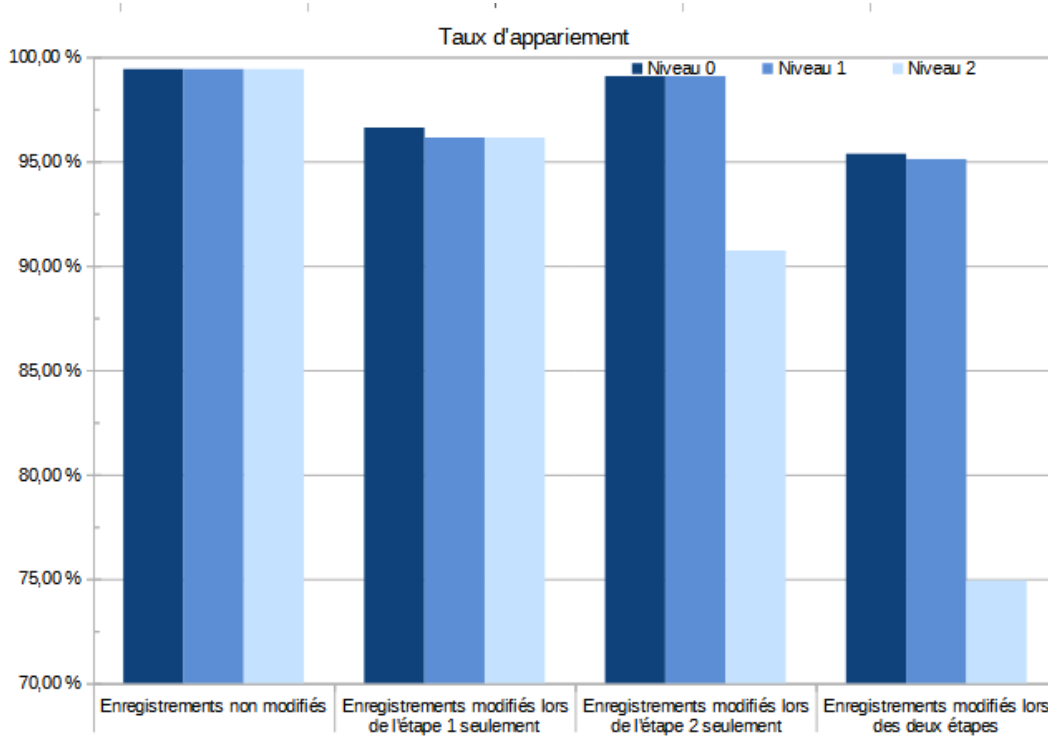
RÉSULTATS : PRÉCISION POUR LES DIFFÉRENTS SCÉNARIOS

		Jeux de variables utilisé		
		Tous les prénoms et lieux de naissance sur 5 positions (+ nom et date de naissance)		
		Niveaux de nettoyage des données		
	Fichiers	Niv. 0	Niv. 1	Niv. 2
Identification via le CSNS	EAR 92	Inconnu >2,23 %	Inconnu >1,12 %	Inconnu >4,96 %
Identification via le CSNS	FIP 92	Inconnu >0,08%	Inconnu >0,15 %	Inconnu >2,40 %
Identification via le CSNS	Tous salariés 92	0,43 %	0,42 %	1,43 %

ÉVOLUTION DU TAUX D'APPARIEMENT POUR LES ENREGISTREMENTS MODIFIÉS



EAR 92



Tous salariés 92

RÉSULTATS : PROCESSUS ABOUTI D'IDENTIFICATION À UN FICHIER NATIONAL

Dans un contexte de processus d'appariement abouti, le nettoyage des données joue très peu, et plutôt de façon négative sur le taux d'appariement

Les premières estimations indiquent une dégradation de la précision, quelle que soit la qualité des données en entrée (exception faite de la codification du lieu de naissance)

Un processus bien conçu surpasse largement le gain de qualité que l'on peut obtenir par le nettoyage des données

Les enregistrements ayant été nettoyés avaient un taux d'appariement avant normalisation plus faible que les autres

La préparation des données est une étape fondamentale pour :

- Qualifier les données
- Sélectionner les variables d'appariement, ainsi que leurs variantes utiles (peuvent être également indexées)
- Concevoir un processus d'appariement (correspondance exacte, approchée, calcul de scores...)

Le nettoyage des données peut :

- Améliorer le taux d'appariement, de façon très marginale si le processus est très élaboré
- Dégrader le taux de vrais positifs, notamment si le nombre de variables d'appariement est faible (absence d'informations localisantes par exemple)

Les premiers résultats de l'appariement (analyse de paires) sont utiles pour améliorer les résultats et peuvent conduire à faire évoluer le processus :

- Les interversions nom/nom marital, ou nom/prénom peuvent être supposées mais sont difficiles à évaluer de prime abord

Les pistes pour poursuivre les travaux :

- Décomposer les effets des différents traitements des données, pour identifier celles qui apportent le plus
- Regarder la corrélation des anomalies pour identifier les enregistrements les plus problématiques
- Tester l'apport d'un « dictionnaire » des prénoms amélioré, en ayant conscience que les apports seront faibles ?

Retrouvez-nous sur

insee.fr

Heidi Koumarianos
Insee – Département des Méthodes Statistiques

01 87 69 55 63
heidi.koumarianos@insee.fr

Insee Mesurer pour comprendre