

***PROBABILISTES OU DÉTERMINISTES,  
DES MÉTHODES D'APPARIEMENTS AU  
BANC D'ESSAI DU PROGRAMME RÉSIL***



# 01 POURQUOI LES APPARIEMENTS SONT-ILS UN ENJEU POUR RÉSIL

« Le programme RESIL vise à construire un système de répertoires d'individus, de ménages et de locaux d'habitation, durable et évolutif, mis à jour à partir de sources administratives diverses. »

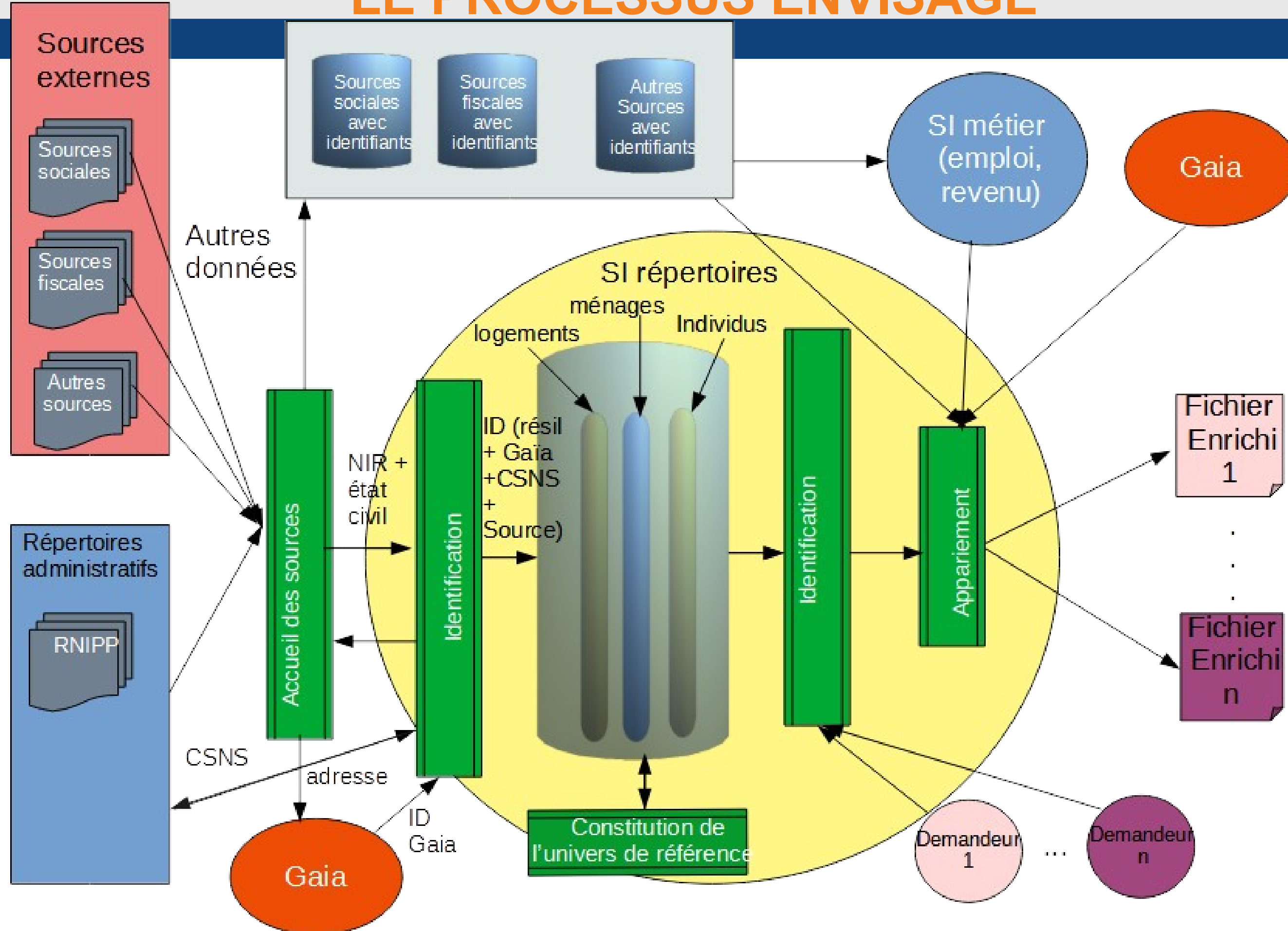
*(source : Compte rendu du CD Insee du 5 octobre 2020)*

- 2 répertoires (individus et logements) mis à jour au fil de l'eau,
- un univers de référence consistant en trois bases annuelles (individus, logements et ménages)

cet univers de référence peut notamment servir à produire des bases de sondage utiles pour l'échantillonnage et le calage, et à vérifier la qualité de couverture des sources administratives ou à faciliter les appariements de sources ;

- et trois services (accueil des sources, production d'univers de référence et production de fichiers enrichis par appariement) ;
- dont la gestion des informations localisantes est assurée par le référentiel Gaïa.

# LE PROCESSUS ENVISAGÉ



# 02 LES DONNÉES APPARIEÉS

---

- Les individus de plus de 15 ans des départements d'Île et Vilaine et de Lozère
  - de l'EAR 2019
  - du Fichier d'Imposition des Personnes (FIP)

- Les variables appariées
  - noms (marital et de naissance) ;
  - prénoms ;
  - date et département de naissance ;
  - adresse de résidence (2 premiers mots directeurs)

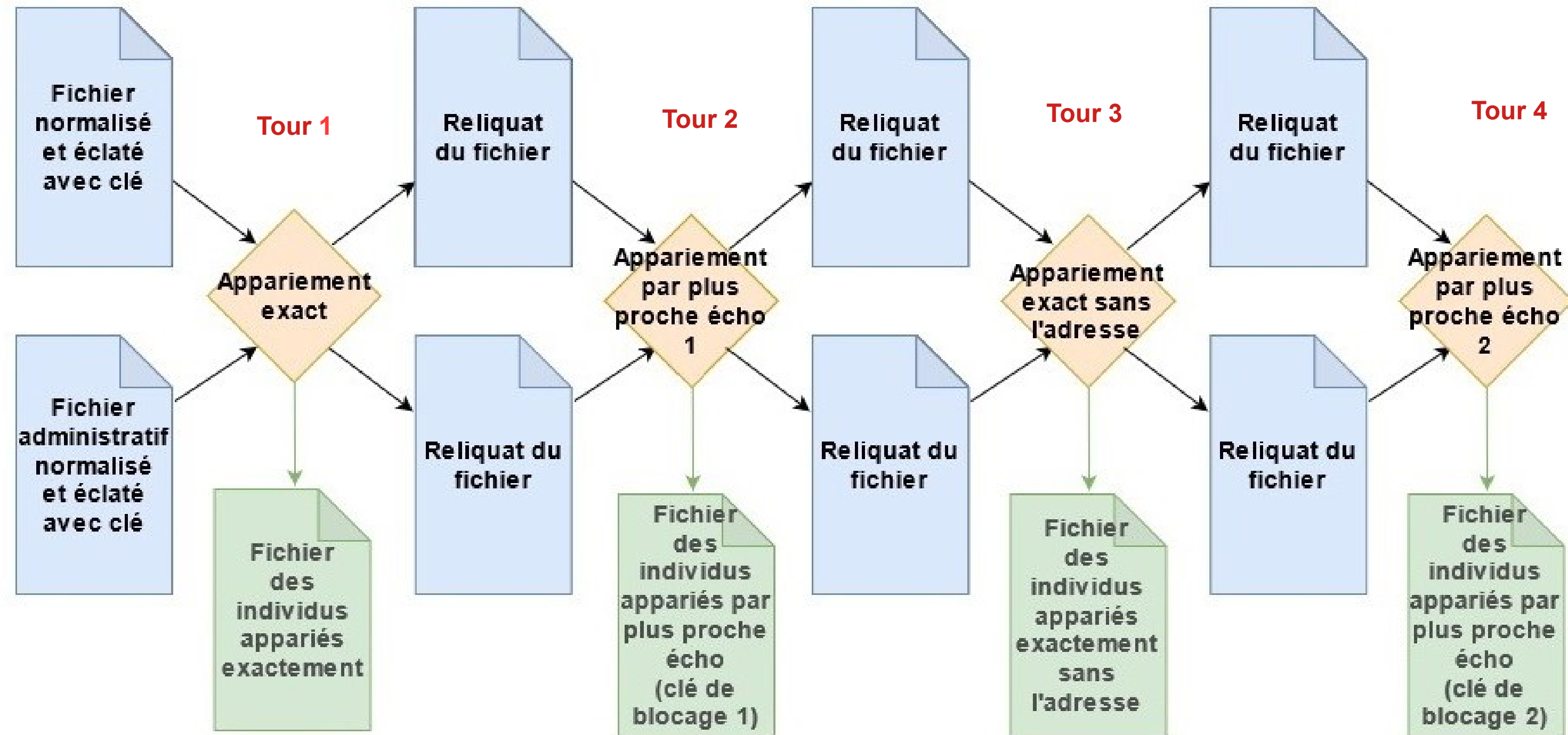


# 03

## LES MÉTHODES

---

# D'APPARIEMENTS TESTÉES

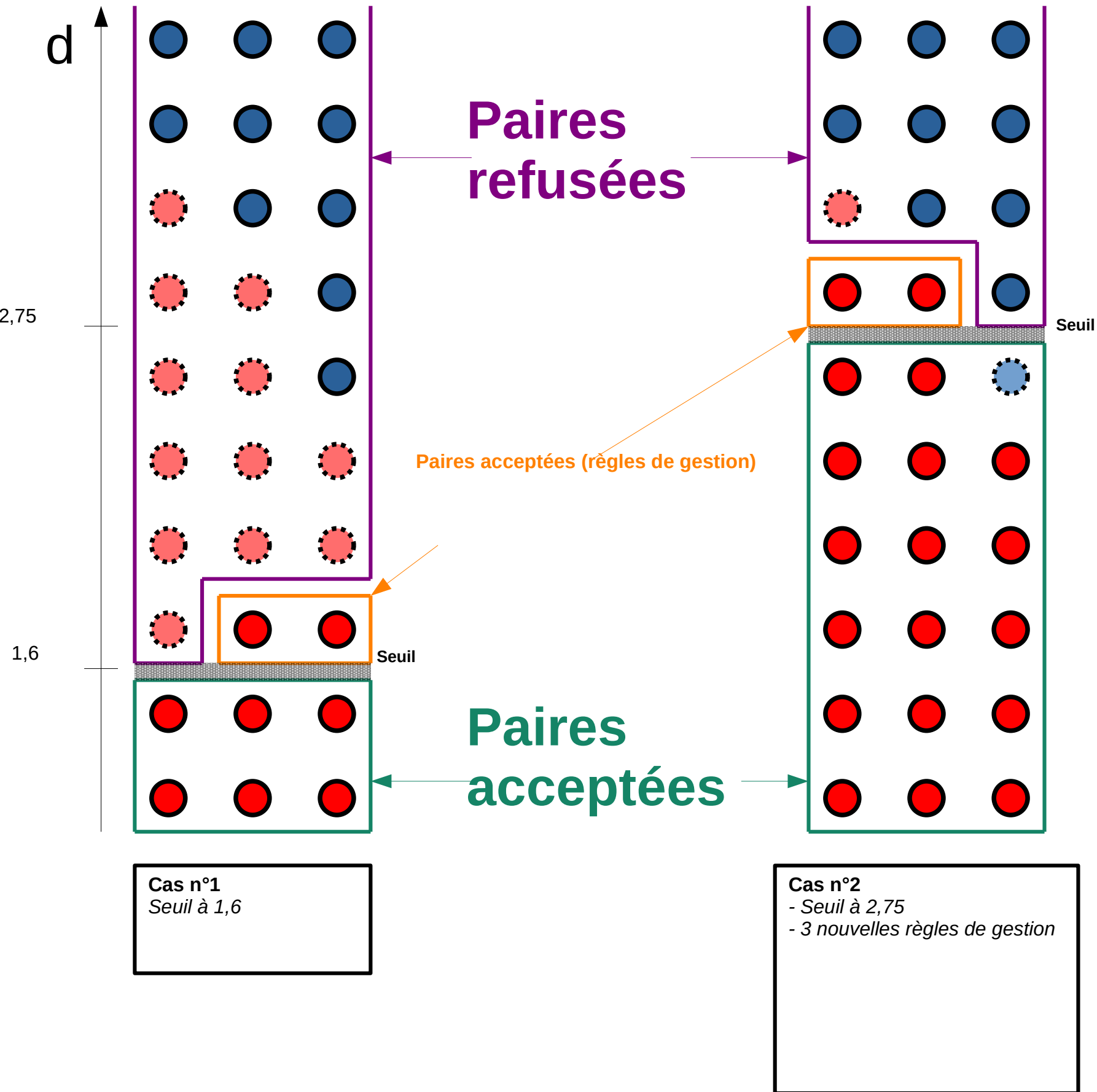






- Définition la distance entre deux individus a et b

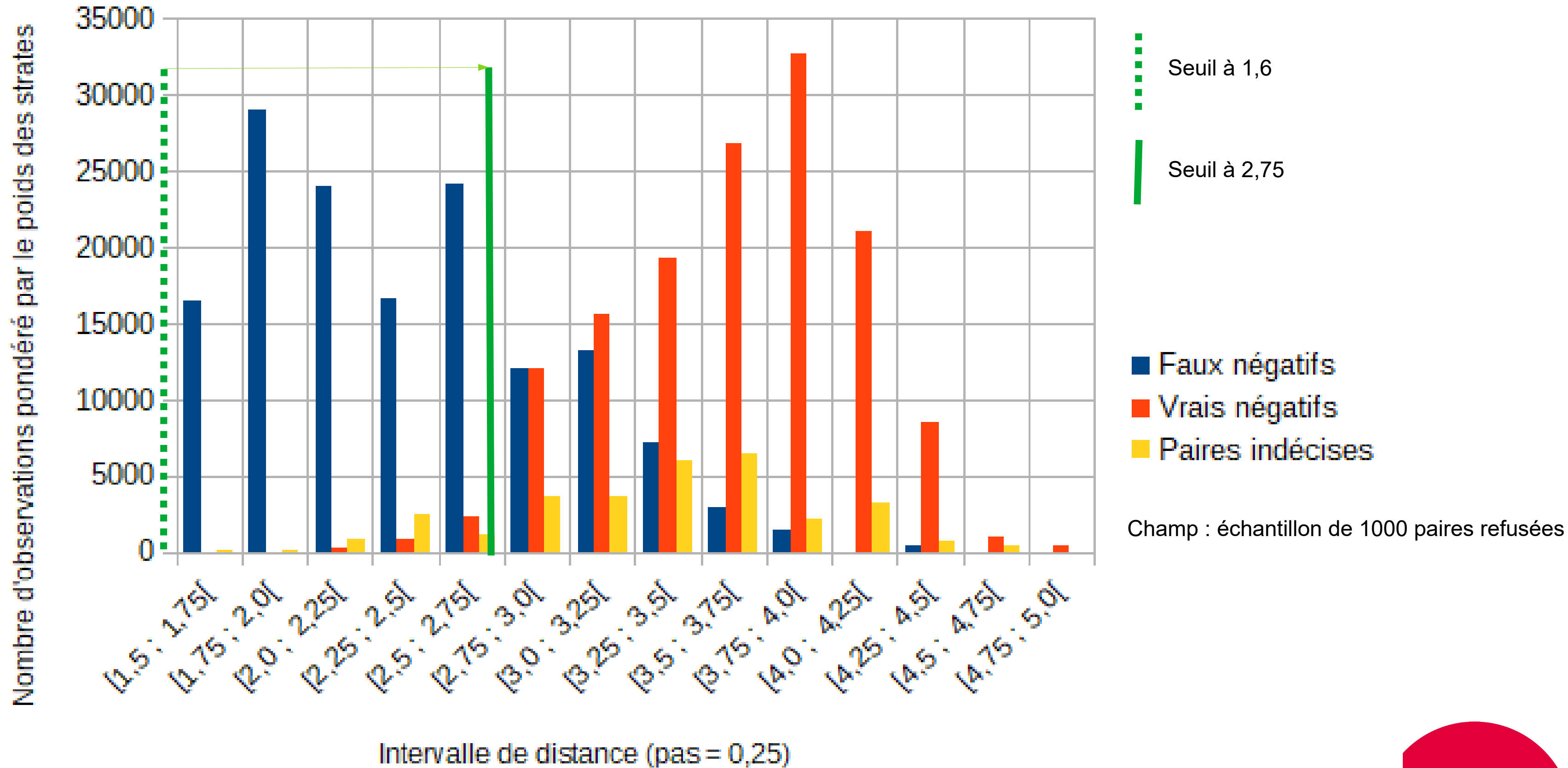
$$d(a,b) = W_1 * dNom (Noma, Nomb) + \dots + W_k * dAdresse (Adressea, Adresseb)$$

Chaque sous-distance est comprise entre 0 et 1

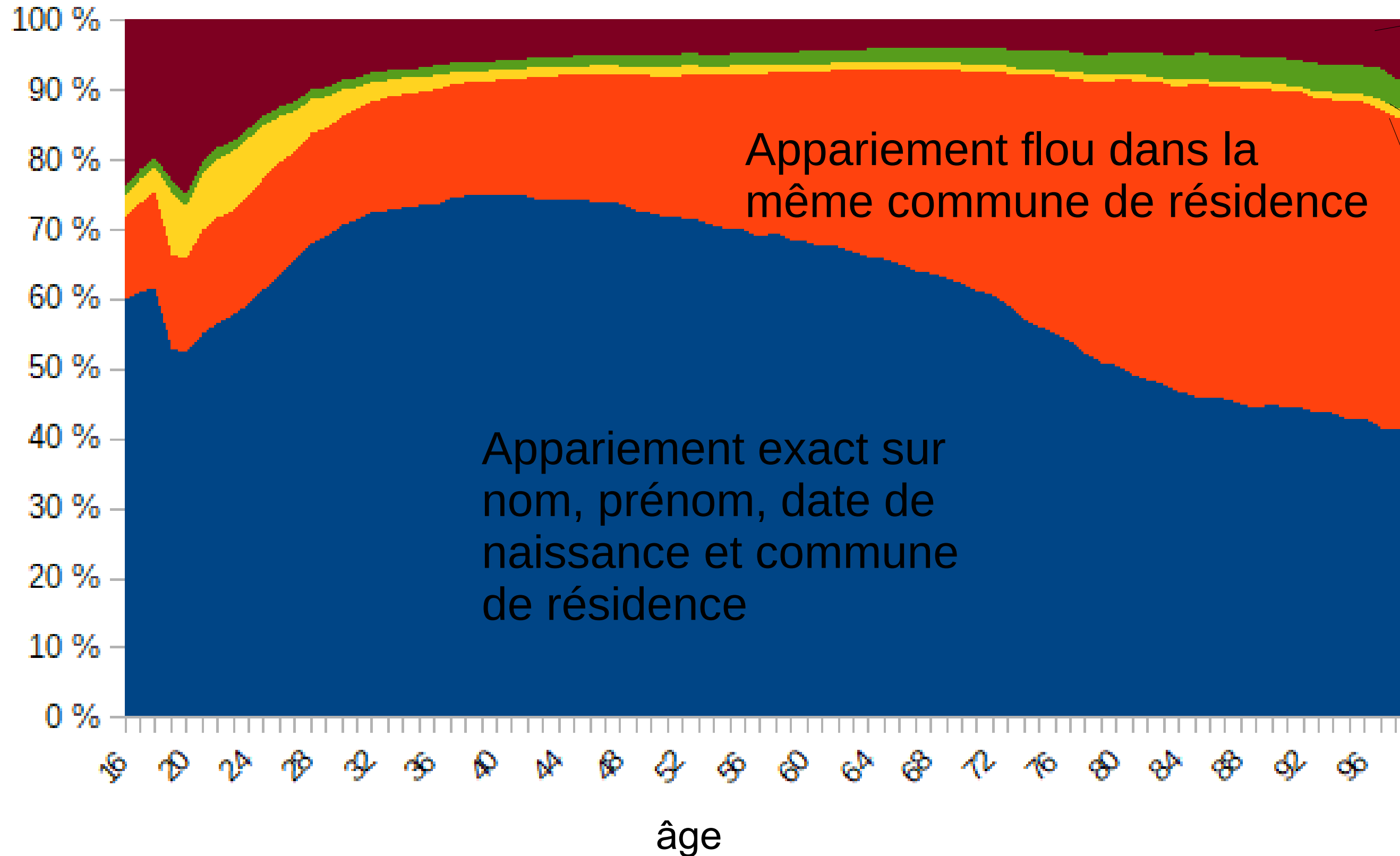
- Les variables utilisées : le nom, le prénom, la date de naissance, le département de naissance, le sexe, la commune et les mots directeurs de l'adresse.
- Basé principalement sur la distance de Levenshtein
- La paire est acceptée si sa distance est inférieure à un seuil et qu'il y a pas une paire contenant un des 2 individus avec une distance inférieure



	Paire mêmes individus	Paire individus différents
Paire acceptée	 Vrai positif	 Faux positif
Paire refusée	 Faux négatif	 Vrai négatif



## Taux d'appariement



Individus non appariés

Appariement flou optimisé à la suite du contrôle visuel

Appariement exact sur nom, prénom et date de naissance (France entière)

- **Outil développé par Istat (Open source)**
- **Pour ce test : utilisation des méthodes probabilistes (Fellegi – Sunter)**
- **Appariement exact**
- **Clé de blocage :**
  - **Commune de résidence (48)**
  - **Commune de résidence + année de naissance pour Rennes (35)**
- **Distance de Levensthein (pour les libellés), égalité pour les nombres**
- **Le seuil d'acceptation des paires est une probabilité de 0,9**

- Indicateurs qualité calculés à partir des probabilités du modèle :

- $$P = \frac{\text{Nombre de vraies paires acceptées}}{\text{Nombre de paires acceptées}} = \frac{\text{Vrais positifs}}{\text{Vrais positifs} + \text{Faux positifs}} = 0,998$$

- $$R = \frac{\text{Nombre de vraies paires acceptées}}{\text{Nombre de vraies paires}} = \frac{\text{Vrais positifs}}{\text{Vrais positifs} + \text{Faux négatifs}} = 0,927$$



- **Méthode probabiliste (Fellegi-Sunter)**
- **Pour la Lozère**
  - **Appariement exact**
  - **Blocage commune**
  - **Blocage Année de naissance**
- **Pour Île et Vilaine (400 milliards de paires potentielles)**
  - **Appariement exact**
  - **Blocage commune et année de naissance**
  - **Blocage commune**
  - **Blocage Année de naissance**
- **Distance : similarité de Jaro-Winkler avec un seuil à 0,92 pour les libellés et égalité pour les nombres**
- **Seuil d'acceptabilité des paires : 0,5**

# 04 LES RÉSULTATS

- **MESURE DE LA QUALITÉ DES FICHIERS EN ENTRÉE**
- **MESURE DE LA QUALITÉ DU PROCESSUS**
  - Taux d'appariement
  - Taux de faux positifs et faux négatifs
- **MESURE DE LA QUALITÉ DE LA POPULATION APPARIÉE**
  - Comparaison de la distribution de variables d'intérêt

	Département 48	Département 35
Population de plus de 15 ans EAR	10 127	130 950
Population de plus de 15 ans FIP	62 823	874 304
Appariés de façon exacte	6 239	96 697
Taux d'appariement « Rapsodie » (en %)	91,3	94,8
Taux appariement « Relais » (en %)	92,1	93,7
Taux d'appariement « Python » (en %)	92,4	95,1
Faux positif « Rapsodie » (en %)	0,02	0,05
Faux positif « Relais » (en %)	0,3	0,03
Faux positif « Python » (en %)	0,3	0,7

Paire retenue par			Paires correctes (en %)		Paires incorrectes (en %)		Paires indéçises (en %)		Total	
Rapsodie	Relais	Python	Dép. 48	Dép. 35	Dép. 48	Dép. 35	Dép. 48	Dep. 35	Dép. 48	Dep. 35
Oui	Oui	Non	80.0	90.0	20.0	3.0	0.0	7.0	5	227
Oui	Non	Oui	100.0	98.7	0.0	0.5	0.0	0.8	44	1155
Oui	Non	Non	97.7	84.0	2.3	6.5	0.0	9.5	43	851
Non	Oui	Oui	76,2	93.0	11,9	5.0	11,9	2.0	143	424
Non	Non	Oui	52.6	26.3	31.6	61.0	15.8	12.7	19	1465
Non	Oui	Non	72.0	82.3	20.0	9.4	8.0	8.3	25	113
Oui	Oui	Oui	100.0	100.0	0.0	0.0	0.0	0.0	8591	126 151

# 05 CONCLUSIONS

---

- **Des résultats connus confirmés :**
  - **Problèmes de volume pour les méthodes probabilistes (taille du fichier et taille maximale de sous-population)**
  - **Intérêt de l'analyse visuelle pour optimiser les résultats**
  - **Un algorithme déterministe bien paramétré donne des résultats satisfaisants**
- **Complémentarité possible des méthodes**

- **Mise en production d'un outil déterministe**
- **Utilisation d'un outil de type probabiliste sur échantillon à des fins de contrôle qualité**
- **Mise en place d'une interface de mesure de la qualité**
  - Estimation d'indicateurs qualité
  - Optimisation du moteur (nouvelles règles)
  - Disponibilité de paires étiquetées (utile pour un modèle de type machine learning supervisé)



**Merci de votre attention**

