Redressements de l'enquête Trajectoires et Origines 2

Journées de Méthodologie Statistique

Olivier Guin, Pierre Tanneau, Willy Thao Khamsing

Insee

Mars 2022

Sommaire

Plan

L'enquête Trajectoires et Origines 2

- L'enquête Trajectoires et Origines 2 2019-2020 (TeO2) est une enquête réalisée par l'Institut national d'études démographiques (Ined) et l'Institut national de la statistique et des études économiques (Insee)
- La deuxième édition de l'enquête Trajectoires et Origines (TeO) de 2008-2009
- Étudie la diversité des populations en France et la situation des populations d'origine
- Objectif double :
  - produire des statistiques sur l'ensemble de la population
  - mais aussi sur certains groupes de populations parfois mal connues

Plan

L'échantillon TeO2

L'échantillon de l'enquête TeO2 est composé de 5 sous-échantillons, liés à des enjeux de diffusion :

- un sous-échantillon (01) d'individus immigrés (G1)
- un sous-échantillon (02) d'individus domiens (G1)
- un sous-échantillon (03) de descendants d'immigrés (G2)
- un sous-échantillon (04) de descendants de domiens (G2)
- un sous-échantillon (05) de la population générale (G1+G2+ni G1 ni G2)

Ceci va orienter chacun des maillons de la chaîne de redressements de l'enquête.

Plan

Correction de la non-réponse

- **CNR du sous-échantillon (05) :** un modèle commun à tous les individus pour corriger le non-réponse questionnaire
- **CNR des sous-échantillons (01) et (02) :** une approche **multimodèles** pour corriger la NR questionnaire, pour mieux adhérer aux enjeux de diffusion de l'enquête
- **CNR des sous-échantillons (03) et (04) :** Deux étapes de CNR différentes :
  - Une CNR pour les **relevés mairie**
  - Une CNR pour la **non-réponse au questionnaire** $\rightarrow$ approche multimodèles, là aussi

Correction de la non-réponse des individus des sous-échantillons 01 et 02

- Les effets des variables explicatives de la non-réponse peuvent varier d'un groupe d'origine à l'autre
- Par ailleurs, l'enquête TeO2 est soumise à des enjeux de diffusion par groupes d'origines
- Il a donc été décidé d'appliquer des modèles différents pour la CNR au questionnaire, suivant les (groupes de) groupes d'origines
- Différents scénarios ont été testés, chaque scénario étant défini par le choix :
  1. D'une stratification de la population en groupes d'origines
  2. Pour une strate donnée, par un choix de variables explicatives de la non-réponse questionnaire

Correction de la non-réponse des individus des sous-échantillons 03 et 04

- **CNR mairie :**
  - Pour certains individus, l'appariement de l'EAR 2018 avec Statec n'a pas suffi à déterminer leur appartenance ou non au champ des G2 → **opération de recherche en mairie**
  - 100 014 individus ont fait l'objet d'une opération mairie, laquelle n'a pas abouti pour 5 133 individus
  - Une correction spécifique du biais de sélection lié à la "non-réponse mairie" a été appliquée

- **CNR questionnaire :** appliquée aux individus des sous-échantillons 03 et 04, là encore par groupes d'origines

Résultats numériques (1/2)

Modèle	Nombre de GRH	Taille du plus petit GRH	Probabilité de réponse la plus faible	Probabilité de réponse la plus forte
Modèle sous-échantillon (05)	9	176	0,11	0,83

Tableau 1 : CNR des individus du sous-échantillon (05)

Modèle	Nombre de GRH	Taille[8] du plus petit GRH	Probabilité de réponse la plus faible	Probabilité de réponse la plus forte
Modèle 1	9	181	0,18	0,80
Modèle 2	9	198	0,14	0,78
Modèle 3	9	159	0,13	0,84
Modèle 4	10	139	0,18	0,83
Modèle 5	9	126	0,14	0,77

Tableau 2 : CNR des individus immigrés

Modèle	Nombre de GRH	Taille du plus petit GRH	Probabilité de réponse la plus faible	Probabilité de réponse la plus forte
Modèle sous-échantillon (02)	9	143	0,11	0,82

Tableau 3 : CNR des individus domiens

Résultats numériques (2/2)

Modèle	Nombre de GRH	Taille du plus petit GRH	Probabilité de réponse la plus faible	Probabilité de réponse la plus forte
Modèle mairie	6	2230	0,81	0,98

Tableau 4 : CNR "mairie" des individus G2

Modèle	Nombre de GRH	Taille[13] du plus petit GRH	Probabilité de réponse la plus faible	Probabilité de réponse la plus forte
Modèle 1	11	510	0,14	0,82
Modèle 2	10	151	0,07	0,68
Modèle 3	8	212	0,23	0,77
Modèle 4	9	312	0,15	0,79
Modèle 5	9	108	0,15	0,77

Tableau 5 : CNR "questionnaire" des descendants d'immigrés

Modèle	Nombre de GRH	Taille du plus petit GRH	Probabilité de réponse la plus faible	Probabilité de réponse la plus forte
(04) descendants de Domiens	6	134	0,22	0,71

Tableau 6 : CNR "questionnaire" des descendants de domiens

Plan

Pourquoi un partage des poids ?

- Les modèles de correction de la non-réponse (CNR) sont spécifiques à chaque croisement sous-échantillon croisé à une variable de regroupement de strate d'origines
- La concaténation des jeux de pondérations des croisements ne permet cependant pas d'assurer le caractère sans biais de nos estimateurs
- Le partage des poids intègre le fait que certains des répondants puissent théoriquement être captés via différents sous-échantillons dans la base de sondage (liens multiples)

Liens (1/2)

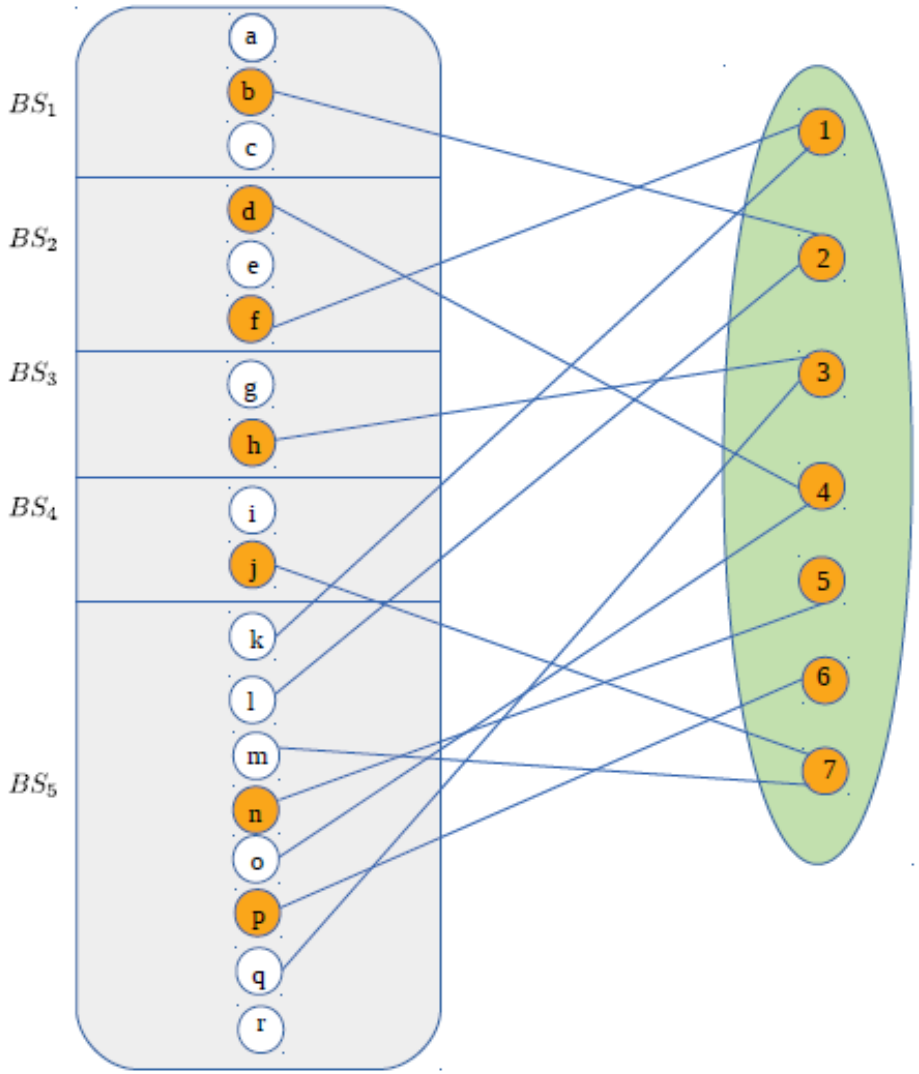

Fig.1 : liens entre les individus répondants (à droite) et la BDS (à gauche)

Liens (2/2)

- **Principe du partage des poids :**
  - Identifier les liens entre les individus répondants i et les unités j de la base de sondage
  - Utiliser ces liens pour construire les poids composites des individus i à partir des unités j

- **Notations :**
  - U^A désigne la base de sondage (formellement, les couples (individus, numéro de sous-échantillon)
  - $L_{j,i} \equiv \begin{cases} 1 & \text{si i et j ont un lien} \\ 0 & \text{sinon} \end{cases}$
  - $L_i^B = \sum_{j \in U^A} L_{j,i} = \begin{cases} 2 & \text{si i est immigré, domien,} \\ & \text{descendant d'immigrés ou de domiens} \\ 1 & \text{sinon} \end{cases}$

Poids

- Le partage des poids permet de passer d'un concept *couple* (individu, numéro de sous-échantillon) à un concept *individu*

- Le poids $\omega_i^{\mathrm{B,CNR}}$ de l'individu répondant i s'obtient à partir des poids $\omega_j^{\mathrm{A,CNR}}$ des unités répondantes $j \in r^A$ de la BDS par la formule

$$\omega_i^{\mathrm{B,CNR}} = \sum_{j \in r^A} \frac{L_{j,i}}{L_i^B} \omega_j^{\mathrm{A,CNR}}$$

Dispersion des poids

- **Problème :** l'application brute de cette formule conduit à une importante dispersion des poids, due :
  - au fait qu'un individu aura un poids très différent s'il est atteint par le sous-échantillon (05) sans surreprésentation ou par un des sous-échantillons (01) à (04) avec des surreprésentations élevées
  - à des surreprésentations différentes d'une origine à l'autre
- **Solution :** pondérer les liens, et faire dépendre ces pondérations des origines

Liens pondérés

- Les **liens pondérés** constituent une généralisation des liens $\frac{L_{j,i}}{L_i^B}$

- Le système $\{\widetilde{\theta}_{j,i}\}$ forme un jeu de liens pondérés si :
  - Pour tout $(i,j) \in r^B \times U^A$

$$0 \leq \widetilde{\theta}_{j,i} \leq 1 \text{ avec } \widetilde{\theta}_{j,i} = 0 \text{ ssi } L_{j,i} = 0$$

  - Pour tout $i \in r^B$:

$$\sum_{j \in U^A} \widetilde{\theta}_{j,i} = 1$$

- La formule de partage des poids devient alors :

$$\omega_i^{\text{B,CNR}} = \sum_{j \in r^A} \widetilde{\theta}_{j,i}\, \omega_j^{\text{A,CNR}}$$

Liens pondérés pour l'enquête TeO2 (1/2)

- Les sous-échantillons sont par construction **disjoints**, donc un individu i n'est en lien qu'avec **une seule unité répondante** $j = j_i$ de l'échantillon r^A, *même s'il peut être en lien avec deux unités de la base de sondage* U^A
- La formule de partage des poids s'écrit donc : $\omega_i^{\mathrm{B,CNR}} = \widetilde{\theta}_{j_i,i}\, \omega_{j_i}^{\mathrm{A,CNR}}$
- Les $\omega_{j_i}^{\mathrm{A,CNR}}$ sont connus (CNR), il reste donc à définir les $\widetilde{\theta}_{j_i,i}$

Liens pondérés pour l'enquête TeO2 (2/2)

- Le lien pondéré $\widetilde{\theta}_{j_i,i}$ dépend de la strate d'origines k_i de l'individu i et du sous-échantillon m_i dans lequel il a été tiré

- En notant n_m^k le nombre de répondants de la strate d'origines k dans le sous-échantillon m, le lien pondéré $\widetilde{\theta}_{j_i,i}$ est donné par :

	Immigré	Domien	Descendant d'immigré	Descendant de domien	Autre
$\widetilde{\theta_{j,i}}$	$\frac{n_{m_i}^{k_i}}{n_1^{k_i}+n_5^{k_i}}$	$\frac{n_{m_i}^{k_i}}{n_2^{k_i}+n_5^{k_i}}$	$\frac{n_{m_i}^{k_i}}{n_3^{k_i}+n_5^{k_i}}$	$\frac{n_{m_i}^{k_i}}{n_4^{k_i}+n_5^{k_i}}$	1

Tableau 1 : *Liens pondérés associés aux individus en fonction de la strate d'origines et du sous-échantillon*

Impact de la pondération des liens sur la dispersion des poids des immigrés et des domiens

	min	Q1	Med	Q3	max	moy	écart-type
Liens pondérés	9,30	109,44	209,95	358,29	8 040,58	310,25	389,98
Sans liens pondérés	8,10	56,06	110,99	197,83	34 224,75	274,45	968,04

Tableau 2 : Distribution des poids des individus répondants G1 après application du partage des poids, avec liens pondérés et sans liens pondérés.

Impact de la pondération des liens sur la dispersion des poids des descendants d'immigrés et de domiens

	min	Q1	Med	Q3	max	moy	écart-type
Liens pondérés	2,56	176,54	288,13	498,41	23 115,69	479,73	964,79
Sans liens pondérés	6,91	97,87	158,56	289,85	42 985,44	426,26	1 180,39

Tableau 3 : Distribution des poids des individus répondants G2 après application du partage des poids, avec liens pondérés et sans liens pondérés

Plan

Troncature des poids

- Une analyse préalable des poids après partage des poids montre que ceux-ci sont relativement dispersés
- Cette dispersion des poids est susceptible de générer de la variance
- Pour la limiter, il a été décidé de tronquer les poids au sein de chaque population (immigrés, Domiens, « autres ») avant même le calage
- Troncature par groupe d'origines : au sein d'un groupe d'origines donné, les poids ont été tronqués au 95^e percentile

Plan

Calage sur marges

- L'opération de calage permet d'estimer correctement les totaux de certaines variables, et de réduire la variance de nos estimateurs

- Cette opération a été appliquée sur trois populations : les immigrés, les domiens, les individus qui ne sont ni immigrés ni domiens

- **Marges de calage :** issues de l'empilement des Enquêtes Annuelles de Recensement (EAR) 2019 et 2020

- **Effet de mode :** l'EAR est pour l'essentiel auto-administrée, alors que TeO2 interroge les individus essentiellement en face-à-face
  - Certaines variables ont été écartées de la liste des variables de calage, car elles recouvrent des concepts différents entre l'EAR et TeO2
  - **Exemple :** le diplôme, qui fait l'objet d'une question dans l'EAR et d'un questionnaire détaillé dans TeO2

Plan

Conclusion

- La non-réponse a été modélisée de façon différenciée suivant les sous-échantillons (immigrés/domiens, descendants d'immigrés/de domiens, population générale) mais aussi suivant les groupes d'origines des individus ou de leurs parents
- La chaîne de redressements doit tenir compte de la multiplicité des liens possibles entre certains individus (immigrés, domiens, descendants d'immigrés, descendants de domiens) et la base de sondage
- Les redressements intègrent donc des opérations de partage des poids
- L'échantillon contient certaines valeurs influentes, en raison d'une importante dispersion des poids. Celle-ci a été traitée par troncature des poids
- Le calage a été appliqué sur trois populations : les immigrés, les domiens, les individus qui ne sont ni immigrés ni domiens. Les marges de calage sont issues du couplage des EAR 2019 et 2020.