Recommandations quant à l'usage des pondérations de l'enquête TeO 2

Les redressements de l'enquête TeO 2 conduisent à la construction de poids relativement dispersés, en raison notamment de la complexité de son mode d'échantillonnage. Cette note, en rappelant brièvement les populations sur lesquelles chacune des étapes des redressements ont été réalisées, vise à aider les utilisateurs des données dans le choix des populations sur lesquelles diffuser des résultats.

1. Rappels sur l'échantillonnage et le redressement

1.a. Rappels sur l'échantillonnage

L'échantillon de l'enquête TeO 2 est subdivisé en cinq sous-échantillons :

- le sous-échantillon (01) des immigrés ;
- le sous-échantillon (02) des Domiens ;
- le sous-échantillon (03) des descendants d'immigrés ;
- le sous-échantillon (04) des descendants de Domiens ;
- le sous-échantillon (05) de la population générale.

Les immigrés et des Domiens sont présents dans les sous-échantillon (01), (02) et (05).
Les descendants d'immigrés et de Domiens sont présents dans les sous-échantillons (03), (04) et (05).
Les individus qui ne sont ni immigrés, ni Domiens, ni descendants d'immigrés ni descendants de Domiens ne sont présents que dans le sous-échantillon (05) de la population générale.

Pour la plupart des étapes des redressements, les opérations sont réalisées séparément sur 3 sous-populations : les immigrés et Domiens d'une part ; les descendants d'immigrés et de Domiens d'autre part ; enfin, la population générale.

Le tirage des immigrés et Domiens des sous-échantillons (01) et (02) est stratifié par groupes d'origines. Ces groupes d'origines sont donnés dans le tableau suivant :

Numéro	Groupe d'origines
1	Algérie
2	Maroc Tunisie
3	Afrique Sahélienne
4	Afrique centrale et du golfe de Guinée
5	Espagne, Italie, Portugal
6	Autres pays de l'UE 28
7	Turquie
8	Asie du Sud-Est
9	Chine
10	Pays avec de nombreux réfugiés
11	Autres pays du monde
12	DOM

Tab. 1 : *strates de tirage des sous-échantillons (01) et (02)*

Le tirage des descendants d'immigrés et de Domiens des sous-échantillons (03) et (04) a nécessité au préalable la construction d'une base de tirage. Cette base de tirage a été construite par une double opération :

- d’appariement entre l’Enquête Annuelle de Recensement 2018 (EAR) et des fichiers anonymisés d’État Civil (STATEC) ;
- de tirage de relevés en mairie.

À l’issue de ces opérations, deux bases de tirage ont finalement été construites :
- une base de tirage de 34 651 individus pour lesquels on connaît l’origine des parents de manière certaine ;
- une base de 3 385 individus pour lesquels on ne connaît pas l’origine des parents de manière certaine.

On distingue finalement les strates de tirage suivantes :

Numéro	Groupe d’origines
1	Algérie
2	Maroc Tunisie
3	Afrique Sahélienne
4	Afrique centrale et Guinéenne
5	Italie Espagne
6	Portugal
7	Autres pays de l’UE 28
8	Turquie
9	Vietnam-Laos
10	Cambodge
11	Autres pays
12	DOM
13	Origines incertaines – Afrique centrale et guinéenne
14	Origines incertaines – origines surreprésentées
15	Origines incertaines – origines sous-représentées

Tab. 2 : *Strates de tirage des sous-échantillons (03) et (04)*

Deux types de strates de tirage doivent être distinguées :
- Celles correspondant aux cibles de répondants par origine géographique ; il s’agit des strates suivantes : *Algérie, Maroc Tunisie, Afrique sahélienne, Afrique centrale et guinéenne, Italie Espagne, Portugal, Autres pays de l’UE 28, Turquie, Vietnam Laos , Cambodge, Autres pays, DOM* ;
- Celles correspondant à des individus pour lesquels on ne connaît pas l’origine des parents de manière certaine. Ces strates sont les suivantes : la strate exhaustive *Origines incertaines – Afrique centrale et guinéenne*, la strate *Origines incertaines – origines surreprésentées*, et enfin la strate *Origines incertaines – origines sous-représentées*.

1.b. Rappels sur les redressements

La chaîne aval a été construite de façon différenciée suivant les sous-échantillons.

i. Correction de la non-réponse

La correction de la non-réponse a consisté en l’application d’un modèle de régression logistique, suivi d’une construction de groupes de réponse homogènes par la méthode de Haziza-Beaumont.

Pour les immigrés et Domiens, et les descendants d'immigrés et de Domiens dans les sous-échantillons qui leur sont spécifiques (i.e. les sous-échantillons (01) à (04)), la correction de non-réponse au questionnaire de l'enquête a été faite en suivant une approche multimodèles, c'est-à-dire en construisant des modèles différenciés selon les groupes d'origines.
Pour les immigrés et Domiens, les groupes d'origines utilisés sont ceux décrits dans le tableau 1. Pour les descendants d'immigrés et de Domiens, les groupes d'origines (des parents) sont présentés dans le tableau 4 ci-dessous[1].

Par ailleurs, une étape spécifique de correction de la non-réponse pour les relevés mairie a été appliquée pour les descendants d'immigrés et de Domiens, en amont de la correction de la non-réponse au questionnaire.

Enfin, les individus du sous-échantillon (05) ont fait l'objet d'un modèle unique de correction de la non-réponse.

Numéro	Groupe d'origines
1	Algérie
2	Maroc Tunisie
3	Afrique Sahélienne
4	Afrique centrale et du golfe de Guinée
5	Espagne Italie
6	Portugal
7	Autres pays de l'UE 28
8	Turquie
9	Asie du Sud-Est
10	Autres pays du monde
11	DOM
12	Origines manquantes

Tab. 4 : *groupes d'origines pour la CNR des sous-échantillons (03) et (04)*

ii. Partage des poids

Le partage des poids intègre le fait que certains répondants peuvent théoriquement être captés via différents sous-échantillons. Plus précisément :

- les immigrés et Domiens sont à la fois dans le champ des sous-échantillons (01) et (02) (selon qu'ils sont immigrés ou Domiens) et (05) : ils possèdent donc deux liens avec la base de sondage ;
- les descendants d'immigrés et de Domiens sont à la fois dans le champ des sous-échantillons (03) et (04) et (05) : ils possèdent donc deux liens avec la base de sondage ;
- les individus qui ne sont dans aucun cas précédent ne peuvent être captés que par le sous-échantillon (05) : ils ne possèdent donc qu'un seul lien avec la base de sondage.

Par ailleurs, l'échantillonnage de l'enquête entraîne naturellement une **forte dispersion des poids**, due au fait que certaines origines sont surreprésentées (resp. sous-représentées), et que de plus ces surreprésentations (resp. sous-représentations) varient d'une origine à l'autre.

Ce problème a été traité en partie en **pondérant** les liens suivant les groupes d'origines.

1 Le groupe *Origines manquantes* regroupe ici les individus pour lesquels il n'a pas été possible de définir l'origine des parents, que ce soit par STATEC, les relevés mairie ou le questionnaire de l'enquête.

iii. Troncature des poids

Malgré l'utilisation de liens pondérés, les poids obtenus en sortie de l'opération précédente restent sujets à une forte dispersion, rendant nécessaire une ultime opération de troncature des poids.
Appliquée par groupe d'origines, celle-ci a consisté à tronquer les poids au 2e et au 98e percentile, i.e. à ramener au 2e percentile les poids situés en-dessous de ce dernier, et au 98e percentile les poids situés au-dessus de ce dernier. Cette opération a permis de considérablement réduire la dispersion des poids par groupe d'origines.

iv. Calage sur marges

Finalement, un calage sur marge a été réalisé à partir d'une source externe afin d'assurer une cohérence entre les totaux estimés de certaines variables (sexe, tranche d'âge, région de résidence, etc.) entre l'enquête TeO 2 et cette source. La source externe retenue pour le calage de TeO 2 est l'enquête annuelle de recensement (EAR) dont l'échantillon particulièrement important (environ 5 millions d'observations sur le champ de TeO 2) permet d'assurer une précision importante sur des populations précises.
En revanche, si le questionnaire de l'EAR renseigne sur le lieu et la nationalité de naissance des enquêtés, il ne permet pas d'estimer si leurs parents sont eux-mêmes immigrés ou Domiens. Ainsi, cette source externe ne permet d'effectuer un calage que pour 3 groupes distincts que sont les immigrés, les Domiens, et les « autres » – sans distinction pour ce dernier groupe entre les descendants d'immigrés, de Domiens, et les personnes sans ascendance migratoire. Ainsi, **le calage sur marge de l'enquête TeO 2 ne permet pas d'assurer une aussi bonne cohérence pour les descendants d'immigrés et de Domiens que pour les immigrés et Domiens.**

L'enquête emploi en continu (EEC) renseigne sur le lieu et la nationalité de naissance des parents des enquêtés. Si son effectif est trop faible pour permettre une utilisation comme source de calage, il est plus important que celui de TeO 2 et cette enquête peut être utilisée comme source de comparaison.
En comparant le nombre d'individus par statut migratoire et groupe d'origines estimé dans les deux enquêtes, il apparaît que **TeO 2 sur-estime, à champ équivalent, d'environ 10 % le nombre de descendants d'immigrés et de Domiens par rapport à l'EEC 2019-2020.**
Ces écarts pourraient s'expliquer notamment par des stratégies d'échantillonnage très différentes entre les deux enquêtes, ainsi que par l'inégale importance accordée à la question des origines dans ces deux enquêtes.
Enfin, il est utile de souligner que **ces taux de sur-estimation apparaissent comparables mais inférieurs à l'écart observé entre l'enquête TeO 1 et l'enquête emploi 2008.**

2. Recommandations quant aux tailles d'effectifs des domaines de diffusion

Du fait entre autres des fluctuations d'échantillonnage, les estimateurs calculés sont entachés d'une marge d'erreur. Afin d'assurer un niveau satisfaisant de précision, il est nécessaire que les domaines de diffusion sur lesquels portent ces estimateurs soient de taille suffisamment importante. Il est donc important de fixer et respecter un seuil d'effectif en dessous duquel il est préférable de ne pas diffuser de résultats.

Pour des raisons qui vont être brièvement détaillées en annexe, nous recommandons de ne diffuser d'indicateurs que sur des domaines d'effectifs au moins égaux à 400.

À titre d'information, le tableau ci-dessous présente les effectifs des répondants à l'enquête par groupes d'origines construits *ex-post* à partir de leurs réponses au questionnaire. **La plupart des groupes d'origines présentent ici des effectifs nettement supérieurs à 400, et ne posent donc pas de problème de diffusion.** En revanche, le groupe *Autres français nés hors France métropolitaine* a un effectif de 336 alors que le groupe *Descendants d'immigré(s) originaire(s) de Chine* n'a qu'un effectif de 31, rendant inenvisageable toute diffusion sur ce domaine.

Naturellement, même pour un groupe d'origines d'effectif initial élevé, tout croisement avec d'autres variables (comme des tranches d'âge, par exemple) nécessite de rester vigilant au fait de ne pas engendrer des domaines trop fins, dont les effectifs passeraient en dessous de ce seuil minimal.

Groupe d'origines	Effectif
Français nés en France métropolitaine	3 559
Autres français nés hors France métropolitaine	326
Descendants d'autres français nés hors France métropolitaine	3 153
Originaires d'un DOM	792
Descendants originaires d'un DOM	661
Immigrés d'Algérie	1 231
Descendants d'immigré(s) originaire(s) d'Algérie	1 500
Immigrés du Maroc/Tunisie	1 382
Descendants d'immigré(s) originaire(s) du Maroc ou de Tunisie	1 142
Immigrés d'Afrique sahélienne	836
Descendants d'immigré(s) originaire(s) d'Afrique sahélienne	554
Immigrés d'Afrique guinéenne ou centrale	1 066
Descendants d'immigré(s) originaire(s) d'Afrique guinéenne ou centrale	533
Immigrés d'Asie du Sud-Est	901
Descendants d'immigré(s) originaire(s) d'Asie du Sud-Est	745
Immigrés de Turquie	923
Descendants d'immigré(s) originaire(s) de Turquie	674
Immigrés de Chine	576
Descendants d'immigré(s) originaire(s) de Chine	31
Immigrés du Portugal	760
Descendants d'immigré(s) originaire(s) du Portugal	789
Immigrés d'Espagne ou d'Italie	339

Descendants d'immigré(s) originaire(s) d'Espagne ou d'Italie	913
Immigrés d'autres pays de l'UE28	856
Descendants d'immigrés originaires d'autres pays de l'UE 28	676
Immigrés d'autres pays	1 526
Descendants d'immigré(s) originaire(s) d'autres pays	737

Tab. 5 : *Effectifs de TeO 2 suivant les groupes d'origines*

3. Recommandations complémentaires pour l'étude des petits-enfants d'immigrés non-européens

Conformément aux recommandations du Comité du Label, l'enquête TeO 2 a été accompagnée d'une enquête complémentaire dédiée spécifiquement aux petits-enfants d'immigrés non-européens, qui dispose d'un caractère expérimental.
Ces deux enquêtes sont à considérer comme des enquêtes distinctes. Les différences d'échantillonnage entre les deux enquêtes créent notamment une différence de champ entre les petits-enfants d'immigrés non-européens interrogés – qui, pour l'enquête complémentaire, doivent avoir un parent entrant dans le champ de l'enquête principale.
Ainsi, les utilisateurs des données veilleront notamment à **ne pas ajouter l'ensemble des observations de l'enquête complémentaire à l'ensemble de celles de l'enquête principale**. Une telle opération augmenterait artificiellement la somme des poids des petits-enfants d'immigrés non-européens, et masquerait la différence de champ explicitée ci-dessus.

En revanche, pour une étude portant spécifiquement sur les petits-enfants d'immigrés non-européens, il est possible de concaténer les deux enquêtes en se restreignant au champ d'intérêt. Dans ce cas, les poids à utiliser pour les G3 de l'enquête principale (variable **poids_cales_g3**) sont différents des poids usuels de cette enquête (variable poids_cales).

Enfin, il est à noter que **la concaténation des deux sources ne permet d'atteindre qu'un nombre assez réduit de G3 non-européens.** La faiblesse de cet échantillon, ainsi que la différence de champ et le caractère expérimental de l'enquête complémentaire, incitent à des précautions dans la formulation des résultats obtenus sur cette population.

Annexe : Intervalle de confiance selon la taille d'échantillon

Pour un estimateur $\widehat{p}$ d'une proportion p inconnue, calculé par sondage aléatoire simple, un intervalle de confiance à 95 % avec une précision de 5 % est donné approximativement par la formule

$$\mathrm{IC} = \left[\widehat{p} - 2\sqrt{\frac{p(1-p)}{n}}\,,\ \ \widehat{p} + 2\sqrt{\frac{p(1-p)}{n}}\right] \quad (1)$$

où n est la taille de l'échantillon restreint au domaine[2].

Cet intervalle de confiance est à comprendre en ce sens : pour environ 95 % des échantillons de taille n tirés par sondage aléatoire simple, la vraie proportion p que l'on cherche à estimer est comprise entre $\widehat{p} - 2\sqrt{\frac{p(1-p)}{n}}$ et $\widehat{p} - 2\sqrt{\frac{p(1-p)}{n}}$, où $\widehat{p}$ est l'estimation calculée à partir de l'échantillon observé.
Le terme $2\sqrt{\frac{p(1-p)}{n}}$ est la précision de l'estimation. On la fixe en général au plus à 5 %, ce qui à p fixé contraint la taille n.

Une précision de 5 % est donc atteinte pour une taille n telle que

$$2\sqrt{\frac{p(1-p)}{n}} = 0,05$$

Dans la pratique, p n'est pas connue, mais on peut montrer que $p(1-p)$ est au plus égal à $0,25$, si bien qu'on cherche en général n sous l'hypothèse conservatrice que $p(1-p) = 0,25$. Autrement dit, on cherche n tel que

$$2\sqrt{\frac{0,25}{n}} = 0,05$$

soit

$$\frac{1}{\sqrt{n}} = 0,05$$

et donc **n=400**.

Notons que l'expression analytique de la précision des estimateurs de TeO 2 est en fait plus complexe que celle donnée dans la formule (1) puisque l'échantillon n'a pas été tiré par sondage aléatoire simple, mais en pratique on estime tout de même que cette dernière conduit à fixer une valeur raisonnable pour ce seuil.

2 Voir par exemple (Ardilly, 2006, p. 77).