PANEL SOCIO-ECONOMIQUE

"Liewen Zu Lëtzebuerg"

Document PSELL n° 86 mai 1996

Procédure de pondération de l'étude relative aux employées privées

Bernard Gailly

CEPS/Instead
DIFFERDANGE
Grand-Duché de Luxembourg

1996

Document produit par le :

CEPS/Instead

CENTRE D'ETUDES DE POPULATIONS, DE PAUVRETE
ET DE POLITIQUES SOCIO-ECONOMIQUES

Rue Emile Mark, B.P. 48,
L- 4501 Differdange
tél. (352) 58 58 55-1 - Fax. 58 55 60

Président: Gaston Schaber

ISBN 2 - 87987 - 064 - X

Procédure de pondération de l'étude relative aux employées privées occupées au Luxembourg

La pondération d'un échantillon observé vise à améliorer sa capacité de fournir des estimations non biaisées des moyennes ou des répartitions de la population-cible, ici la population des employées privées occupées au Luxembourg en 1995.

Les biais susceptibles d'entacher les estimations des moyennes et des distributions de la population peuvent provenir de différentes sources. Quatre aspects ont été pris en compte dans le cadre de cette étude :

1. les erreurs contenues dans le fichier de la population elle-même
2. les difficultés à contacter un certain nombre de personnes pour les raisons les plus diverses
3. les difficultés à obtenir des réponses de la part des personnes qui ont pu être contactées
4. les erreurs qui proviennent soit de la composition de l'échantillon, soit d'autres sources d'erreurs qui n'ont pu être identifiées et corrigées au cours des étapes précédentes.

1. Le fichier initial : source d'inexactitudes

Il est tout à fait possible que l'échantillon observé ne reflète pas exactement la structure du fichier dont il est extrait même lorsque cet échantillon a été sélectionné selon les règles de l'art. Le fichier initial de la population peut, en effet, contenir un certain nombre « d'inexactitudes ».

Par exemple, il peut « dater » et comprendre des individus qui auraient dû disparaître de ce fichier parce qu'ils n'appartiennent plus à la population visée au moment où l'observation se déroule sur le terrain. On verra par la suite que ce décalage dans le temps se répercute inexorablement sur l'échantillon. Il rend impossible un ajustement absolument rigoureux de l'échantillon sur la structure de la population de référence. L'ajustement du profil de l'échantillon sur le profil de la population devra admettre certains compromis.

Le fichier initial des employées privées occupées au Luxembourg contenait trois types d'erreurs qui n'ont pas pu être corrigées :

1. un certain nombre de personnes sont décédées (0,1%)
2. un certain nombre de personnes ont quitté le pays (0,8%)
3. un certain nombre de personnes ont quitté le champs de l'enquête; elles ne travaillent plus (5,7%).

Tableau 1

Diagnostic Fichier total	Fréquence relative
1. Réponse	58,0
2. Refus	16,8
3. Mauvaise adresse	4,7
4. Personne absente	10,7
5. Ne travaille plus	5,7
6. Décédées	0,1
7. Emigrées	0,8
8. Non distribuées	3,2
Total = 4878	100

Nous ignorons totalement les conséquences réelles de ces inexactitudes. Il ne serait possible de connaître leurs conséquences sur la structure de la population qu'en procédant soit à une analyse au sein même du fichier initial soit à une observation systématique des personnes qui ne travaillent plus. Ces dernières représentent ,en effet, la grande majorité des décalages entre l'échantillon et la population (Tableau 1).

Il sera cependant possible d'en percevoir certaines manifestations lorsque toutes les erreurs attribuables au déroulement de l'enquête elle-même auront été corrigées.

Le fait que certaines adresses n'ont pas été distribuées dès lors que l'échantillon a atteint la taille voulue, n'a aucune influence sur la structure de l'échantillon. Les adresses distribuées ayant été sélectionnées de manière aléatoire, on peut admettre que le sous-échantillon composé des adresses restantes présente le même profil que les adresses distribuées.

Il existe donc un décalage de l'ordre de 6,6% entre la structure de la population et la structure de l'échantillon (total des modalités #5, #6 et #7). La comparaison entre la population et l'échantillon, avant et après pondération, devra admettre une marge de tolérance du même ordre.

2. L'échantillon et les source de biais

Nous appellerons « fichier pertinent » cette partie de l'échantillon sélectionné qui se prête à des interventions correctrices. Il est composé des personnes ayant répondu, des personnes ayant refusé explicitement de prêter leur collaboration à l'étude et des personnes qui n'ont pas été contactées pour des raisons diverses.

La procédure de pondération s'applique précisément à cette partie de l'échantillon parce qu'il est possible d'y diagnostiquer les différentes sources de biais systématiques qui affectent l'échantillon et de remédier à leurs effets. Cette procédure s'attaque à trois sources de biais : - les taux de contact - les taux de réponse - les erreurs d'échantillonnage.

1. Dans un premier temps, il s'agit de corriger les effets de l'inégalité des taux de contacts. Cette inégalité peut avoir contribué à sur-représenter dans l'échantillon des personnes qui pour des raisons précises sont plus aisément accessibles que d'autres. Il suffit que cette plus grande accessibilité soit systématiquement liée à certaines caractéristiques de la population que l'on souhaite précisément étudier pour que les résultats observés soient systématiquement biaisés.

2. Dans un second temps, il s'agit de corriger les effets de l'inégalité des probabilités de réponse des personnes qui ont été contactées. L'inégalité des probabilités de réponse peut avoir contribué à sur-représenter dans l'échantillon des personnes qui, pour des raisons très diverses, se sont révélées plus disponibles que d'autres. Si cette plus grande disponibilité est systématiquement liée à certaines catégories de population que l'on souhaite étudier, les résultats observés seront systématiquement biaisés et refléteront mal l'information relative à l'ensemble de la population.

Nous expliquerons pourquoi les taux de contact et les taux de réponse doivent être corrigés séparément.

3. Dans un troisième temps, on procède à une « post-stratification » de l'échantillon. Le profil de l'échantillon pondéré doit encore être redressé en fonction des disparités qui pourraient subsister entre la structure de la population et la structure de l'échantillon observé. Si les deux phases précédentes visaient à corriger des biais issus des conditions dans lesquelles l'enquête s'est déroulée, cette troisième étape corrige des erreurs qui proviennent en général de l'échantillon original extrait du fichier de la population.

2.1 Les taux de contacts : phénomène propre au déroulement de l'enquête

2.1.1 Distinguer probabilité de contact et probabilité de répondre

Le fichier « pertinent » est composé de trois grandes catégories de personnes :

⇒ les personnes qui ont répondu

⇒ les personnes qui ont refusé explicitement de répondre

⇒ les personnes qui n'ont pu être contactées.

Tableau 2

Diagnostic Fichier pertinent	Fréquence relative
1. Réponse	64,3
2. Refus	18,6
3. Mauvaise adresse	5,2
4. Personne absente	11,9
Total = 4396	100

Cette dernière catégorie est elle-même composée de deux types de situations rencontrées par les enquêteurs au cours de l'enquête : les adresses inexactes et les personnes « qui sont toujours absentes ».

Ces deux catégories représentent 17,1% des adresses distribuées soit une proportion pratiquement équivalente à celle des refus de répondre (Tableau 2).

Si la proportion des refus de répondre était négligeable par rapport à la proportion des non-contacts ou, inversement, si la proportion des non-contacts était négligeable par rapport à la proportion des non-réponses, il n'y aurait pas un grand intérêt à dissocier ces deux sources de biais. Par contre, plus ces deux facteurs représentent des proportions équivalentes du fichier « pertinent », plus il y est important de les distinguer. Pourquoi ?

Parce que la probabilité de répondre dépend d'une condition préalable : pour avoir la possibilité de répondre ou de refuser, il faut d'abord avoir été contacté. Autrement dit, la probabilité de répondre est une « probabilité conditionnelle » ou encore : la probabilité de répondre et la probabilité d'être contacté ne sont pas indépendantes l'une de l'autre.

Statistiquement, cette différence est importante. Si ces deux probabilités étaient *indépendantes*, les refus pourraient être traitées sans prendre en compte *les « non-contacts »*.

Or, ces deux probabilités sont *dépendantes.* Elles doivent donc être *multipliées.* C'est le cas dans cette étude. On peut illustrer le problème.

> *Soit deux enquêteurs. Le premier contacte une personne sur deux (0.5) et parmi ces « contacts », il obtient quatre réponses sur dix (0.4). Le second contacte toutes les personnes qui lui ont été confiées (1.0) mais n'obtient que deux réponses sur dix (0.2).*
>
> *Si le taux de réponse n'est* pas dépendant *du taux de contact, le taux de réponse du premier enquêteur est de 0.4 et le taux de réponse du second est de 0.2. Apparemment le premier enquêteur a été le plus performant.*
>
> *Si le taux de réponse et le taux de contact sont* multipliés, *le premier enquêteur obtient un score de 0.2 (0.5 * 0.4) et le second un score de 0.2 (1.0 * 0.2). Les deux enquêteurs ont donc atteint la même performance.*
>
> *Le premier enquêteur a rencontré plus de problèmes au moment de la recherche des personnes à interroger et le second au moment d'interroger les personnes qu'il a pu aisément contacter.*

Si ce ne sont pas les mêmes enquêteurs qui rencontrent des problèmes au moment de contacter les personnes à interroger et au moment d'obtenir leur collaboration et si, en outre, la probabilité d'obtenir un contact et la probabilité d'obtenir une réponse ne dépendent pas des mêmes facteurs, on réalise à quel point il est important de distinguer ces deux phénomènes.

Les effets de la sur-représentation (ou de la sous-représentation) de certaines catégories de ménages peuvent dépendre de manière distincte

* des facteurs qui déterminent la probabilité d'être contacté
* et, d'autre part, des facteurs qui déterminent la probabilité de répondre.

2.1.2 Dispositif d'analyse des taux de contact

Le « taux de contact » se mesure en rapportant le nombre de personnes contactées au nombre total de personnes qui auraient dû être contactées. Soit, ici, 88,1% (Tableau 2).

L'objectif de l'analyse des « taux de contact » est de vérifier s'il existe des personnes qui risquent d'être sur-représentées (ou sous-représentées) dans l'échantillon parce qu'elles appartiennent à une ou à plusieurs catégories de personnes qui se sont révélées systématiquement plus difficiles à contacter au cours de l'enquête.

Il existe différentes techniques permettant d'atteindre cet objectif :

* pondérations individuelles (régression logistique sur les caractéristiques des répondants/non-répondants)
* pondération fondée sur les seules caractéristiques de sondage
* pondération générale (s'il s'avère que la probabilité de contact est répartie aléatoirement à travers l'ensemble de l'échantillon ou si les informations disponibles ne suffisent pas pour isoler des zones de concentration des non-contacts)

* et, enfin, les pondérations par classes d'individus[1].

Nous aurons recours à la pondération par classes ou catégories d'individus parce qu'elle présente certains avantages statistiques (elle permet de contrôler et de limiter la variance des poids).

Les classes d'individus seront définies à partir des quelques variables dont il est possible de disposer avant toute observation des individus appartenant à l'échantillon (par exemple, il est impossible d'analyser les taux de contact des ménages en fonction du nombre d'adultes ou du nombre d'enfants dans le ménage puisque ces données sont inconnues pour les ménages non-contactés).

2.1.3 Les déterminants des taux de contact

On s'attend généralement à ce que la probabilité de contact dépende essentiellement de trois facteurs : un « effet enquêteur », un « effet localisation » ou « accessibilité géographique » et enfin à un effet « mobilité géographique des ménages ».

Le fichier de gestion de l'enquête permet d'attribuer à toute personne, y compris aux personnes non-contactées, le numéro code de l'enquêteur chargé d'effectuer l'interview, le pays de résidance des frontalières et le canton des employées qui résident au Luxembourg.

Les 4396 adresses ont été réparties entre 95 enquêteurs. Ce nombre élevé d'enquêteurs permet d'éviter qu'un ou deux d'entre eux mette la qualité de l'étude en péril en ne réalisant que de très faibles performances soit au moment de prendre contact avec les ménages soit au moment d'obtenir la collaboration des ménages contactés.

Il n'en reste pas moins que les performances des enquêteurs peuvent varier considérablement :

* 59 enquêteurs, soit 60% d'entre eux, obtiennent un taux de contact moyen de 90,6%

* 20 enquêteurs obtiennent un taux de contact moyen de 73,3%

* 16 enquêteurs n'obtiennent qu'un taux de contact de 57,5%.

Si la difficulté de contacter les ménages provenait d'une manière générale des ménages eux-mêmes, il serait difficile d'expliquer la grande disparité des résultats obtenus par les enquêteurs. Ces problèmes affecteraient les résultats de tous les enquêteurs sans créer de tels écarts.

La nature même des problèmes évoqués en cas de non-contact invite à penser qu'il s'agit bien d'un problème propre à certains enquêteurs : en cas de changement d'adresse, un enquêteur tenace pourrait rechercher la nouvelle adresse; en cas d'absences récurrentes de la personne l'enquêteur pourrait se montrer plus patient, plus insistant ou plus persévérant.

---

[1] La pondération individuelle est plus fondée en théorie mais la pondération par classes permet de mieux contrôler la variance des poids et d'éviter d'attribuer des poids excessifs à des sous-groupes de très petite taille.

Il s'avère, en effet, que la variance des taux de contact est largement absorbée par la variabilité des caractéristiques des enquêteurs.

Il semblait, à première vue, que la France et l'Allemagne avaient été des terrains particulièrement difficiles. En réalité, lorsque « l'effet enquêteur » est corrigé, il ne subsiste pratiquement plus de disparités entre les pays de résidence.

La correction de cet effet « enquêteur » s'effectue en attribuant à chaque personne interrogée un premier poids égal à l'inverse du taux de contact obtenu par l'enquêteur chargé de cette enquête. Plus le taux de contact obtenu par l'enquêteur est élevé, moins il est nécessaire de renforcer le poids des personnes qu'il a interrogées. Inversement, plus le taux de contact obtenu par l'enquêteur est faible, plus il est nécessaire de renforcer le poids des personnes qu'il a interrogées. Ces personnes risquent, en effet, d'être sous-représentées dans l'échantillon[2].

Chaque personne ou unité d'observation dispose dorénavant d'un poids relatif égal à une fraction ou à un multiple de l'unité selon qu'elle appartient à une catégorie sur-représentée ou sous-représentée. Seules les personnes appartenant à une catégorie correctement représentée conservent leur poids unitaire initial.

2.2 Les taux de réponse : phénomène propre au déroulement de l'enquête

2.2.1 Dispositif d'analyse des taux de réponse

Le « taux de réponse » se mesure par le rapport entre

* le nombre de personnes ayant répondu
* et le nombre de personnes ayant eu la possibilité de répondre parce qu'elles ont été contactées.

Tableau 3

Diagnostic Fichier "contactés"	Fréquence relative	Freq. rel. corrigée *
1. Réponse	77,6	77,3
2. Refus	22,4	22,7
Total = 3646	100	100

* corrigée des effets liés aux non-contacts

Sur un total de 3646 personnes contactées, 77,6% ont accepté de prêter leur concours à l'enquête. Soit un taux de réponse tout à fait acceptable (Tableau 3, 1ère col.).

Ce taux de réponse est calculé « comme si » toutes les personnes avaient pu être contactées et représentaient encore chacune « une » unité d'observation. Nous savons que ce n'est pas le cas. Toutes les personnes n'ont pas pu être contactées et la correction des effets de ce phénomène a eu pour conséquence que la majorité des personnes représente désormais une fraction ou un multiple de l'unité. Il s'en suit logiquement que la proportion des personnes ayant répondu est légèrement modifiée : le taux de réponse global s'établit à 77,3% des personnes contactées (Tableau 3, 2ème col.).

[2] On notera, ici, deux points. D'une part, le « taux de contact » d'un enquêteur est, en réalité, le taux de contact moyen de la classe d'enquêteurs à laquelle il a été affecté au cours de l'analyse. Cette procédure repose, d'autre part, sur un postulat auquel on peut réfléchir : elle présuppose que les ménages contactés présentent des caractéristiques analogues aux ménages non-contactés par les enquêteurs appartenant à une même classe. Ceci reste indémontrable.

L'objectif de l'analyse des « taux de réponse » est de vérifier s'il existe des personnes qui risquent d'être sur-représentées (ou sous-représentées) dans l'échantillon parce qu'elles appartiennent à une ou à plusieurs catégories de personnes qui se sont révélées systématiquement plus enclines que d'autres à prêter leur concours à l'enquête.

Nous aurons recours, ici encore, à la pondération par classes ou catégories d'individus. Cette procédure consiste à comparer les taux de réponse de différentes catégories de personnes.

Dans un premier temps, on choisit un certain nombre de variables qui permettent de classer tous les individus y compris les individus qui n'ont pas répondu. On compare la capacité de ces variables de classer l'ensemble des personnes contactées en deux catégories dont les taux de réponse présentent des contrastes significatifs (statistiquement). On choisit ensuite la variable qui offre le contraste le plus significatif entre les taux de réponse des deux catégories d'individus.

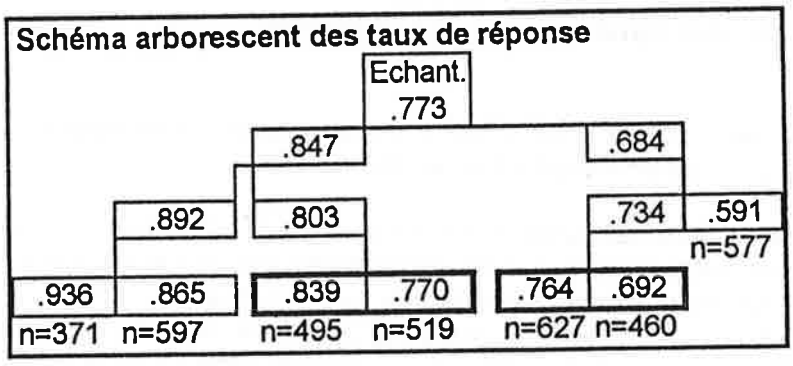
On se trouve ainsi en présence de deux sous-échantillons et on reprend la même procédure séparément pour chaque sous-échantillon. En poursuivant cette procédure, on construit progressivement un « arbre » qui décompose peu à peu l'échantillon en sous-groupes dont les taux de réponse sont de plus en plus contrastés et dont la taille est de plus en plus réduite.

La procédure est interrompue soit lorsqu'une nouvelle division des sous-groupes aboutirait à isoler des catégories dont la taille serait inférieure à 5% de la taille de l'échantillon total, soit lorsque les variables mises en concurrence n'arrivent plus à discerner des sous-groupes dont les taux de réponse sont significativement différents.

A chaque étape de la procédure, toutes les variables disponibles sont remises en concurrence. Une même variable peut donc intervenir plusieurs fois, à plusieurs stades de la division de l'échantillon, à des étapes différentes et dans des branches différentes de l'arbre qui se construit.

Dans cette étude nous ne disposons que de deux variables susceptibles de jouer ce rôle de division progressive de l'échantillon : les enquêteurs et la localisation des personnes.

ILLUSTRATION



Le schéma ci-joint illustre le type de résultat auquel on parvient en utilisant cette procédure.

Il n'est guère utile d'y insérer le nom des variables qui commandent les divisions successives de l'échantillon : **tout l'arbre repose sur des subdivisions progressives des groupes d'enquêteurs**. Seuls les sous-groupes encadrés en gras correspondent à des subdivisions des lieux de résidence des personnes contactées.

Chaque cellule correspond au taux de réponse obtenu par les enquêteurs auprès des personnes interrogées. La taille des sous-groupes (n=) n'est indiquée que pour les catégories finales auxquelles l'analyse aboutit.

La correction des effets liés à la variation des taux de réponse s'effectue en attribuant à chaque personne interrogée[3] un poids égal à l'inverse du taux de réponse de la sous-catégorie à laquelle elle appartient (pour mémoire : « pondération par classe »). Plus le taux de réponse obtenu par les enquêteurs est élevé, moins il est nécessaire de renforcer le poids des personnes interrogées. Inversement, plus le taux de réponse obtenu par l'enquêteur est faible, plus il est nécessaire de renforcer le poids des personnes interrogées : ces personnes risquent, en effet, d'être sous-représentées dans l'échantillon.

L'analyse des taux de réponse a été effectuée sur des unités qui ont été au préalable pondérées en raison inverse des taux de contacts (on aura observé dans le schéma que le taux de réponse de l'ensemble de l'échantillon s'établit à 77,3 % et non à 77,6 %).

Pour prendre en compte ce premier facteur de pondération, les poids attribués aux individus en raison de la variabilité des taux de réponse sont donc multipliés par les poids qui leur ont été attribués en raison de la variabilité des taux de contacts (cf. 2.1.1).

On obtient ainsi pour chaque personne effectivement interrogée un poids qui correspond à l'inverse de la probabilité de réponse du sous-groupe auquel elle appartient en prenant en compte le fait que, dans ce sous-groupe, toutes les personnes n'ont pas nécessairement bénéficié des mêmes chances d'être contactées.

A cette phase du processus de pondération, il faut considérer que l'essentiel des biais systématiques introduits par l'enquête est corrigé « au mieux ». Il y a certainement toujours moyen d'atteindre de meilleurs résultats mais la qualité des résultats obtenus dépend en partie de la variété des informations dont on peut disposer au moment de procéder à l'analyse des taux de contact et des taux de réponse.

Les variables disponibles pour l'ensemble de la population et pour l'ensemble de l'échantillon mis à notre disposition étaient relativement limitées. Sans doute est-ce dans cette voie que l'on peut envisager d'améliorer à l'avenir les conditions du contrôle de qualité des enquêtes et les conditions dans lesquelles les glissements de l'observation pourront être corrigés.

2.3 La post-stratification : dernier ajustement de l'échantillon

La post-stratification aussi appelée « calage aux marges » constitue la troisième et dernière phase de la procédure de correction des biais qui affectent l'échantillon.

En principe cette phase ne serait pas nécessairement indispensable si nous disposions d'emblée d'une information plus large concernant les caractéristiques de l'ensemble des personnes qui constituent l'échantillon extrait du fichier de la population (y compris les personnes dont on a pu observer par la suite qu'elles étaient décédées, émigrées ou qu'elles ne travaillaient plus). Si ces données avaient été disponibles, elles auraient pu être prises en compte au cours des phases précédentes.

[3] Il va de soi que seules les personnes ayant répondu bénéficient effectivement de ce poids, comme seules les personnes ayant été effectivement contactées ont bénéficié des poids destinés à redresser les biais liés à l'inégalité des taux de contact.

La post-stratification vise à rapprocher encore davantage la structure de l'échantillon observé et la structure de la population (telle qu'elle peut être connue à un moment donné). Elle se justifie donc par deux raisons essentielles. Elle sert, d'une part, à corriger la sous-représentation de certaines sous-populations dans l'échantillon. Elle améliore, d'autre part, la précision des estimations pour certaines classes de population (ou « post-strates »).

Envisageons un cas de figure. Une variable importante présente le même degré d'homogénéité dans différentes classes de la population. Si, parmi ces classes, certaines sont sur-représentées et d'autres sont sous-représentées, le degré de précision des estimations variera selon les classes[4]. Plus la taille d'une classe sera sous-estimée, moins précise sera l'estimation de la valeur dans cette classe.

La post-stratification est, toutefois, basée sur un postulat. On présuppose que les personnes se comportent différemment selon les classes de la population auxquelles elles appartiennent et que l'on a choisies pour procéder à la post-stratification.

Enfin, la post-stratification peut poser un problème. Par souci de précision, on peut chercher à multiplier les critères de stratification. Plus les strates sont nombreuses plus les estimations sont précises. Parallèlement, plus les strates sont nombreuses, plus il est probable qu'un grand nombre d'entre elles ne soient pas représentées dans l'échantillon. Ceci va nécessairement se répercuter sur le recalage de strates voisines.

Les variables à notre disposition sont : l'âge, l'état civil, le pays de résidence, la nationalité et le secteur d'activité des employées privées occupées au Luxembourg. Dans un premier temps, nous avons post-stratifié l'échantillon simultanément en fonction de

* 4 classes d'âges (moins de 30 ans - 30 à 39 ans - 40 à 49 ans - 50 ans et plus)
* 3 types d'état civil (célibataires - mariées - autres)
* et 4 pays de résidence (Luxembourg - Allemagne - Belgique - France).

Soit, au total, 48 strates dont

* 6 étaient vides et correspondaient à des strates de personnes âgées de plus de 50 ans
* et 16 correspondaient à de très petits effectifs dans l'échantillon (les femmes séparées, divorcées ou veuves des différentes classes d'âge dans les 4 pays).

Au terme de cette opération la distribution de ces quatre variables est correctement réajustée dans l'échantillon, mais la distribution des employées par secteur d'activité n'a pas bénéficié de ce dernier ajustement. Ceci signifie, en d'autres termes, que la répartition par secteur d'activité n'est corrélée avec aucune de ces trois variables. Un dernier réajustement par secteur d'activité pouvait donc être effectué sans influencer de manière significative le réajustement précédent.

Ces différentes opérations n'ont pas permis de corriger la sur-représentation des autochtones dans les trois pays frontaliers : ce phénomène ne peut être attribué qu'à la composition de l'échantillon extrait du fichier national.

---

[4] Pour mémoire : la précision de l'estimation dépend du rapport entre la variance et la taille de la classe. A variance (ou homogénéité) équivalente, la sous-représentation d'une classe diminue la précision des estimations dans cette classe.

3. Résultats

3.1 Evolution des variables de référence

Les tableaux 4 à 8 présentent l'évolution de 5 variables de référence.

La première colonne présente la distribution de chacune de ces variables dans la population-cible, c'est-à-dire dans le fichier national avant toute épuration des inexactitudes constatées au cours de l'enquête.

La deuxième colonne présente la distribution de ces variables telles qu'elles ont pu être observées dans l'échantillon des personnes contactées qui ont accepté de prêter leur concours à l'enquête. La comparaison entre la première et la deuxième colonne indique les catégories de la population qui ont eu à subir des *distorsions* soit en raison des *inexactitudes contenues dans le fichier national* soit en raison de *l'inégalité des taux de contacts enregistrée par les enquêteurs.*

La troisième colonne présente la distribution des variables après correction de l'inégalité des taux de réponse (conditionnellement à l'inégalité des taux de contacts). Cette troisième colonne *élimine donc la plus grande partie des biais liés aux conditions du déroulement de l'enquête.* La comparaison avec la deuxième colonne permet de mettre en évidence les zones de l'échantillon qui ont été affectées par ces biais. La comparaison avec la première colonne met en évidence les disparités qui subsistent entre l'échantillon et la population-cible (admise comme référence).

Rappelons ici qu'il nous est impossible d'affirmer que cet écart correspond à une « défaillance de l'échantillon qui nous a été fourni » ou à une « défaillance de l'enquête » ou encore à une « défaillance de la procédure de pondération ». Il n'est pas impossible que cet échantillon pondéré reflète parfois mieux la réalité du terrain que les données concernant « la population-cible » compte tenu du fait que ces dernières contiennent certaines inexactitudes qui sont apparues au cours de l'enquête.

La quatrième colonne présente la distribution des variables après « post-stratification » de l'échantillon. Par construction, ces distributions sont les plus conformes aux distributions de la population-cible (prise comme référence). Elles reflètent encore certaines distorsions que l'on peut néanmoins tolérer. En effet, le décalage observé entre la population-cible et la réalité du terrain porte pratiquement sur 6 % de la population et les écarts observés ici sont inférieurs à cette marge de tolérance.

Tableau 4 Répartition par pays de résidence (%)

Pays de résidence	Population col.1	Echantillon col.2	Ech. pondéré col.3	Post-stratific. col.4
Allemagne	5,6	4,6	5,2	5,9
Belgique	8,6	9,3	8,6	8,8
France	12,9	11,5	13,6	13,2
Luxembourg	72,9	**74,6**	72,6	72,1
	100	100	100	100

L'échantillon observé sur-représentait les jeunes femmes (moins de 30 ans), les Belges, les Luxembourgeoises et les femmes résidant au Luxembourg, les femmes mariées, séparées ou divorcées et les femmes occupées dans le secteur des services (social, santé).

Tableau 5	Répartition par âge (%)			
Age	Population	Echantillon	Ech. pondéré	Post-stratific.
	col.1	col.2	col.3	col.4
< 30 ans	41,7	44,4	44,7	42,7
30 - 39 ans	34,2	33,3	33,3	33,9
40 - 49 ans	17,3	16,6	16,4	17,1
50 et plus	6,9	5,7	5,6	6,3
	100	100	100	100

La procédure de pondération a permis de redresser la distribution des employées selon leur pays de résidence et de corriger ainsi leur distribution selon leur nationalité.

Tableau 6	Répartition par nationalité (%)			
Nationalité	Population	Echantillon	Ech. pondéré	Post-stratific.
	col.1	col.2	col.3	col.4
Allemande	7,6	6,6	7,2	8,1
Belge	11,3	12,5	11,8	12,0
Française	16,0	15,1	16,9	16,6
Italienne	3,7	3,8	3,6	3,7
Luxembourg	52,6	53,3	52,1	51,2
Portugaise	3,4	3,2	3,1	2,9
Autres	5,5	5,5	5,3	5,5
	100	100	100	100

La post-stratification selon l'âge, l'état civil et le pays de résidence a permis d'améliorer la précision des estimations selon l'âge et l'état civil. Par contre, elle est restée sans effet sur la distribution par secteur d'activité, ce qui tend à montrer que cette variable est effectivement indépendante des autres.

Tableau 7	Répartition par état civil (%)			
Etat civil	Population	Echantillon	Ech. pondéré	Post-stratific.
	col.1	col.2	col.3	col.4
Célibataire	43,7	36,5	36,7	44,4
Mariée	47,3	51,2	51,4	46,8
Séparée/Divorcée				
Veuve	9,0	12,3	11,9	8,8
	100	100	100	100

La seconde post-stratification, selon le secteur d'activité, a donc pu redresser la distribution de cette variable sans remettre en cause les résultats obtenus au préalable.

Tableau 8	Répartition par secteur d'activité (%)			
Secteur d'activité	Population	Echantillon	Ech. pondéré	Post-stratific.
	col.1	col.2	col.3	col.4
Primaire (1)	1,5	1,1	1,1	1,5
Industrie (2)	5,1	4,2	4,1	5,1
Batiment et G.civil	2,5	3,8	4,0	2,5
Commerce	23,3	21,3	21,2	23,3
Transports, comm.	3,4	3,5	3,6	3,4
Banque et Ass.	36,1	30,2	30,1	36,1
Services	28,1	35,9	35,9	28,1
	100	100	100	100

(1) Primaire : Agriculture, viticulture, énergie, eau
(2) Industrie : métallurgie et manufactures

3.2 Tests sur des variables croisées

Une première série de tests peut être effectuée en observant les écarts entre les observations de l'échantillon pondéré et la population de référence sur des variables « croisées » (par exemple, la distribution des classes d'âges selon les pays de résidence). Seul le résultat de l'ensemble de la procédure de pondération retiendra l'attention. Nous examinerons donc uniquement les distributions obtenues pour l'échantillon pondéré et post-stratifié.

Tableau 9 Répartition par âge selon le pays de résidence (%)

1. Luxembourg			
Age	Population	Echantillon	Post-stratific.
< 30 ans	38,2	**41,1**	39,1
30 - 39 ans	33,9	33,2	34,0
40 - 49 ans	19,2	18,5	18,8
50 - 59 ans	8,6	7,2	8,1
	100	100	100

2. Autres pays			
Age	Population	Echantillon	Post-stratific.
< 30 ans	51,0	**54,0**	52,2
30 - 39 ans	34,9	33,8	33,6
40 - 49 ans	11,9	10,7	12,7
50 et plus	2,2	1,5	1,5
	100	100	100

L'échantillon formé des personnes contactées et interrogées sur-représentait les jeunes femmes de moins de 30 ans tant au Luxembourg que dans les pays frontaliers (Tableau 9).

L'échantillon pondéré et post-stratifié fournit des distributions beaucoup plus proches du profil de l'ensemble de la population. Il n'arrive cependant pas à restituer rigoureusement la proportion de personnes âgées de 50 ans et plus parce ces personnes sont particulièrement peu nombreuses et mal représentées en Allemagne et en Belgique

Tableau 10 Répartition par âge selon la nationalité (%)

1. Luxembourgeoise			
Age	Population	Echantillon	Post-stratific.
< 30 ans	36,1	37,6	35,8
30 - 39 ans	33,8	33,6	34,0
40 - 49 ans	20,7	20,7	21,0
50 et plus	9,5	8,2	9,2
	100	100	100

2. Autres nationalités			
Age	Population	Echantillon	Post-stratific.
< 30 ans	47,9	52,2	50,1
30 - 39 ans	34,7	33,1	33,9
40 - 49 ans	13,4	11,8	12,9
50 et plus	4,0	3,0	3,2
	100	100	100

Ces jeunes femmes de moins de trente ans, sur-représentées dans l'échantillon, sont surtout des femmes de nationalité étrangère.

Cette sur-représentation n'atteint pas les femmes de nationalité luxembourgeoise (Tableau 10).

Les limites de la procédure de pondération et de post-stratification se manifeste donc essentiellement dans le sous-groupe des femmes de nationalité étrangère où les jeunes femmes de moins de trente ans restent légèrement en surnombre au détriment des trois autres classes d'âges.

Tableau 11	Nationalité selon le pays de résidence (%)		
	1. Luxembourg		
Nationalité	Population	Echantillon	Post-stratific.
Luxembourg.	71,3	70,9	70,4
Autre	28,7	29,1	29,6
	100	100	100
	2. Allemagne		
Nationalité	Population	Echantillon	Post-stratific.
Allemande	92,7	95,4	97,0
Autre	7,3	4,6	3,0
	100	100	100
	3. Belgique		
Nationalité	Population	Echantillon	Post-stratific.
Belge	89,8	92,4	92,6
Autre	10,2	7,6	7,4
	100	100	100
	4. France		
Nationalité	Population	Echantillon	Post-stratific.
Française	90,7	94,8	94,8
Autre	9,3	5,2	15,3
	100	100	100

Dans les trois pays frontaliers, les autochtones sont systématiquement sur-représentées (Tableau 11). Ce n'est pas le cas au Luxembourg. La pondération de l'échantillon et sa post-stratification ne parviennent pas à redresser ce biais. En Allemagne, cette procédure tend même à le renforcer.

Faute d'informations, il est impossible de savoir si ce biais provient de l'échantillon brut (tel qu'il a été extrait du fichier national) ou s'il résulte de la sélectivité des phénomènes de « non-contact » et de « non-réponse » en cours d'enquête.

Cette dernière hypothèse est *moins probable* que la première parce qu'il serait surprenant que des enquêteurs différents opérant dans des zones géographiques différentes produisent exactement le même biais et dans des proportions pratiquement identiques. Cette hypothèse ne peut pourtant pas être écartée. En outre, cette marge d'erreur est tolérable puisqu'il existe également un décalage de plus de 6% entre le fichier national et l'échantillon nettoyé de ses inexactitudes.

Les jeunes femmes de moins de trente ans sont sur-représentées dans les trois catégories d'états civils considérées (célibataires, mariées et autres) (Tableau 12).

La procédure de pondération et de post-stratification parvient à résorber en très grande partie ce phénomène.

Au total, on ne retiendra donc que la sur-représentation des autochtones dans les pays frontaliers.

On gardera ce fait présent à l'esprit au moment d'analyser les données tout en se souvenant aussi du fait que *ce phénomène ne peut influencer que les variables avec lesquelles il est fortement corrélé.*

Tableau 12	Répartition par âge selon l'état civil (%)		
	1. Célibataires		
Age	Population	Echantillon	Post-stratific.
< 30 ans	66,4	70,0	66,9
30 - 39 ans	24,8	22,1	24,7
40 - 49 ans	5,6	4,9	5,8
50 et plus	3,2	2,9	2,6
	100	100	100
	2. Mariées		
Age	Population	Echantillon	Post-stratific.
< 30 ans	25,8	33,9	26,8
30 - 39 ans	42,8	39,8	42,6
40 - 49 ans	23,8	20,6	23,7
50 et plus	7,6	5,7	6,9
	100	100	100
	3. Autres		
Age	Population	Echantillon	Post-stratific.
< 30 ans	5,2	12,1	5,6
30 - 39 ans	34,4	39,7	34,2
40 - 49 ans	39,6	34,2	39,0
50 et plus	20,7	14,1	21,2
	100	100	100

3.3 Tests sur des variables quelconques

La « pondération idéale » n'aurait aucun effet sur aucune variable de l'étude. Elle témoignerait ainsi du fait que l'échantillon ne souffrait d'aucun biais. On l'a vu, ce n'est pas le cas ici. Elle a permis de redresser certains biais liés aux enquêteurs et à la localisation des personnes. Elle a également permis d'améliorer la précision des estimations qui pourraient être faites à partir de certaines variables.

L'effet de l'estimateur des poids sur d'autres variables sera directement proportionnel à la corrélation qui existe entre ce facteur de pondération et les autres variables.

Entre autres tests, on peut donc envisager de comparer les distributions de fréquence de quelques variables observées dans le cadre de l'étude avant et après pondération.

Tableau 13	M. a un enfant de moins de douze ans	
	Echantillon	Post-stratific.
oui	30,0	26,1
non	70,0	73,9
total	100	100

Les quatre exemples présentés dans les tableaux 13 à 16 montrent que l'effet de la pondération varie selon les variables.

L'effet le plus sensible se manifeste par une réduction de la proportion de femmes ayant au moins un enfant de moins de 12 ans. Cet effet s'explique aisément par la correction systématique de la sur-proportion de femmes âgées de moins de trente ans.

Tableau 14	M. a suivi des post-formations	
	Echantillon	Post-stratific.
oui	62,4	63,3
non	37,6	36,7
total	100	100

L'effet de la pondération sera toujours plus fort sur des variables liées plus ou moins directement aux variables qui ont contribué à l'élaboration de l'estimateur des poids. La question relative aux « raisons de l'emploi au Luxembourg » n'a pas été posée aux personnes résidant au Luxembourg (Tableau 16) : il est logique que leur proportion diminue après pondération puisque leur proportion a été réduite dans le cadre la procédure de pondération (Tableau 4).

Tableau 15	Sans travail depuis plus de six mois	
	Echantillon	Post-stratific.
oui	24,2	23,3
non	75,8	76,7
total	100	100

Tableau 16	Raisons de l'emploi au Luxembourg	
	Echantillon	Post-stratific.
Non applicable	74,6	72,1
Pas travail dans pays de résidence	4,3	4,8
Meilleur salaire au Luxembourg	7,9	8,5
Dans pays de résidence travail ne convenait pas	0,8	1,0
Travail au Luxembourg convient mieux	1,8	1,9
Meilleures conditions de travail au Luxembrg.	0,8	1,0
Trajets plus courts	0,9	0,9
Opportunité	8,9	9,9
Total	100	100