DÉMOGRAPHIE ET CULTURES

Colloque international de Québec
(Canada, 25-29 août 2008)

ASSOCIATION INTERNATIONALE DES DÉMOGRAPHES DE LANGUE FRANÇAISE
A I D E L F – *133, boulevard Davout – 75980 Paris Cedex 20 (France) – http://www.aidelf.org*

AIDELF. 2008. DÉMOGRAPHIE ET CULTURES - Actes du colloque de Québec (Canada, 25-29 août 2008),
Association internationale des démographes de langue française, 1 212 pages.

Identification indirecte des sous-groupes culturels et risques d'erreurs

Ceren INAN
Université de Bordeaux 4, IEDUB

Introduction

Quel que soit l'objectif de l'étude, l'identification des individus appartenant à un sous-groupe culturel peut être problématique en raison des frontières vagues entres les sous-populations, des appartenances multiple des individus et de leur implication, plus ou moins forte, variante dans le temps. Ainsi, contrairement à des catégories couramment utilisées en démographie, la définition d'un sous-groupe culturel peut nécessiter une élaboration plus complexe et être plus fragile. Les démographes peuvent porter leur analyse sur des groupes définie par les caractéristiques culturelles les plus significatives, ou pensées comme telle, des sous-groupes culturels. Ainsi, par exemple, la langue maternelle, l'appartenance ethnique, le pays d'origine, la religion ou le nom de famille peuvent être utilisés pour classer les individus comme membre d'un sous-groupe. Dans certains cas, la classification des individus à partir de ces caractéristiques culturelles peut faire défaut et même être contraire à l'éthique. Par conséquent, à défaut des caractéristiques culturelles, on peut tenter de définir l'appartenance à un groupe par un comportement différentiel. Or, un classement de ce genre, avec des caractéristiques culturelles ou comportementales, ne garantit pas que les individus classés dans un groupe soient réellement des membres de ce groupe et que les individus appartenant à ce groupe soient tous classés comme membre. Notre article tente d'attirer l'attention sur les erreurs possibles encourues par l'usage de telles pratiques d'identification indirecte des individus appartenant à un sous-groupe de culture. Notre présentation se décline en deux parties. Dans un premier temps, nous présenterons un cas d'école avec une seule variable d'identification, un premier exemple basé sur une variable dichotomique, correspondant à une caractéristique culturelle supposée exclusive, suivie d'une autre discrète à plusieurs modalités, correspondant à une caractéristique de comportement spécifique. Enfin, nous présenterons les cas où on augmente le nombre de variable d'identification dans des cadres de constructions plus complexes. Chaque fois, nous essayerons de présenter les divers facteurs jouant sur le biais, tels, la structure de la population selon les sous-groupes culturels, l'intensité de la variable d'identification à l'intérieur des sous-groupes et, si on souhaite effectuer des mesures différentielles, de l'intensité de la variable étudiée à l'intérieur des sous-groupes.

1. Cas simple - variable étudiée et variable d'identification.

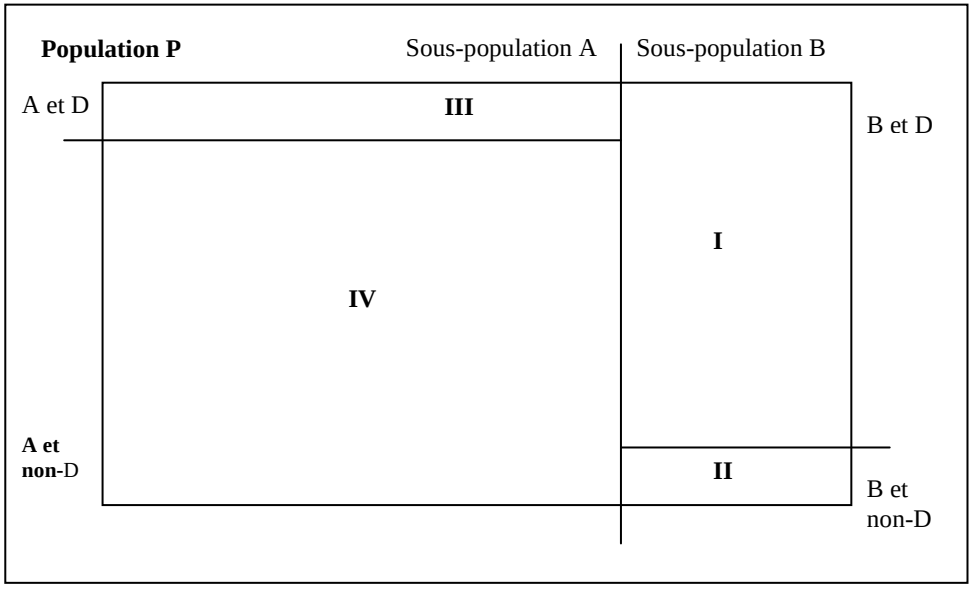
Supposons une population P divisée en deux sous-populations, A et B. Nous voulons étudier une variable C à l'intérieur de la sous-population B que nous ne pouvons pas directement identifier pour une raison quelconque. De ce fait, nous utilisons une variable d'identification D, qui partagera la population en deux sous-ensembles : ceux qui ont D et ceux qui n'ont pas D (le schéma n°1). En somme, nous avons quatre sous-ensembles, I, II, III et IV, qui regroupent respectivement, les individus qui appartiennent à la sous-population B et qui ont la caractéristique D, les individus qui appartiennent à la sous-population B et qui n'ont pas la caractéristique D, les individus qui n'appartiennent pas à la sous-population B et qui ont la caractéristique D et enfin les individus qui n'appartiennent pas à la sous-population B et qui n'ont pas la caractéristique D.

AIDELF. 2008. DÉMOGRAPHIE ET CULTURES - Actes du colloque de Québec (Canada, 25-29 août 2008), Association internationale des démographes de langue française, 1 212 pages.

Une mesure de la variable étudiée à l'intérieur de la sous-population D dépendra de l'occurrence de la variable à l'intérieur des sous-ensembles I et III et du poids de chacun de ces sous-ensembles relatif à la sous-population D. Or, la valeur recherchée à l'intérieur de la sous-population B dépendra de l'occurrence de la variable à l'intérieur des sous-ensembles I et II et du poids de chacun de ces sous-ensembles relatif à la sous-population B. Ainsi, l'écart entre ce que nous voulons mesurer en B et ce que nous mesurerons avec D dépendra non seulement des individus exclus à tort mais aussi des individus inclus à tort.

Il est évident qu'une variable d'identification qui écarte totalement les individus appartenant à une sous-population que nous voulons isoler mais qui ne possèdent pas la caractéristique d'identification utilisée et les individus qui possèdent cette caractéristique d'identification sans appartenir à cette sous-population est une variable d'identification de qualité. En classant les individus correctement, elle nous fournira non seulement une bonne estimation de la variable étudiée mais aussi une bonne estimation de l'effectif de la sous-population B. Dans ce cas, le rapport du sous-ensemble qui représente les individus appartenant et ayant la caractéristique d'identification avec les individus appartenant ou ayant la caractéristique[1] sera proche de 1 et signifiera une superposition quasi totale.

SCHÉMA N°1 : LA POPULATION P STRUCTURÉE ENTRE DEUX SOUS-POPULATIONS ET SELON LA CARACTÉRISTIQUE D.



Cependant, il est faux de penser que toutes les autres variables d'identification, différentes de celles qui minimisent à la fois le sous-ensemble II et à la fois le sous-ensemble III, provoquent un biais et sont dénuées d'intérêt. Par exemple, une variable d'identification qui minimise seulement les individus inclus à tort et qui n'est pas corrélée avec la variable étudiée au niveau des individus que nous voulons isoler nous procurera aussi une bonne mesure. Quels que soit leur poids, les individus identifiés seront représentatifs de l'ensemble de la sous-population B.

[1] $P(D \cap B)/P(DUB) \in [0 ;1]$

AIDELF. 2008. DÉMOGRAPHIE ET CULTURES - Actes du colloque de Québec (Canada, 25-29 août 2008), Association internationale des démographes de langue française, 1 212 pages.

Enfin, certaines variables (fondamentales) peuvent paraître évidentes à utiliser alors que leur niveau de superposition est très médiocre. Par exemple, pour isoler les immigrés (au sens de « né(e) de nationalité étrangère et à l'étranger ») on peut se contenter d'une variable « lieu de naissance », dans ce cas en France, cette variable sera un très mauvais estimateur des immigrés provenant d'Algérie qui ne représentent que 44% de ceux qui sont nés en Algérie (Tableau n°1). On peut aussi vouloir utiliser la variable « langue utilisée pour communiquer avec l'entourage ». Dans ce cas toujours en France, l'utilisation de l'anglais sera problématique (Tableau n°2).

TABLEAU N°1 : LIEU DE NAISSANCE ET QUALITÉ D'IMMIGRÉ (AU SENS DE "NÉ(E) DE NATIONALITÉ ÉTRANGÈRE ET À L'ÉTRANGER").

Lieu de naissance	Qualité d'immigré	
	Immigré	Non immigré
Algérie	512 420	663 790
Ailleurs	3 342 420	37 982 680

Source : Étude de l'Histoire Familiale de 1999, population âgée de 18 ans et plus, CP.

TABLEAU N°2 : LA PRATIQUE DE L'ANGLAIS AVEC DES PROCHES SELON LA LANGUE TRANSMISE À L'ENQUÊTÉ PAR SES PARENTS CHEZ LES NON IMMIGRÉS.

l'enquêté pratique avec des proches	L'anglais est transmis par			
	ni mère, ni père	que la mère	que le père	père et mère
l'anglais	1 860 430	18 260	18 190	20 680
autre langue	36 685 580	16 400	18 070	8 860

Source : Étude de l'Histoire Familiale de 1999, population âgée de 18 ans et plus, CP.

1.1. Transmission de l'alsacien par la mère et le taux de premier départ à 20-24 ans parmi les femmes du groupe de génération 1950-1954.

Ce premier exemple illustre les biais possibles encourus, causé par une variable d'identification exclusive au sous-groupe à étudier[2]. Supposons que nous voulons étudier le premier départ du foyer parental[3] chez des femmes alsaciennes et que nous utilisons la transmission de l'alsacien en tant que la première autre langue que le français transmise par la mère comme variable d'identification. Nous supposons que la transmission de l'alsacien est une caractéristique exclusive (aucune mère de non-alsacienne ne transmet l'alsacien) mais non commune (il existe des mères d'alsaciennes qui ne transmet pas l'alsacien) chez les alsaciennes. Nous avons estimé un taux de premier départ du foyer parental à 20-24 ans pour les femmes du groupe de génération 1950-1954 à qui l'alsacien était transmis par la mère. Ce dernier s'élève à 547,44 pour mille. Nous avons ensuite modélisé l'écart entre cette valeur mesurée et la valeur pour l'ensemble des femmes alsaciennes de cette génération en fonction des différentes valeurs possibles[4] de ce taux chez des femmes alsaciennes à qui l'alsacien n'était pas transmis par la mère et des différents poids possibles[5] de celles-ci dans l'ensemble des femmes alsaciennes de ce groupe de générations (graphique n°1).

2 P(B/D) = 1.

3 Pour rendre notre exemple introductif simple, nous avons choisie un événement non renouvelable et non fatal.

4 Nous avons limité les valeurs possibles entre 500 pour mille et 610 pour mille.

5 Nous avons limité les poids possibles entre 0,05 et 0,95.

Revenons au schéma n°1 pour cet exemple. Nous avons supposé qu'il n'existe aucun individu qui possède la caractéristique D (transmission de l'alsacien) sans appartenir à la sous-population B (les femmes alsaciennes), donc le sous-ensemble III est vide. Dans ce cas, si la sous-population B est homogène quant à la variable étudiée C (le premier départ à 20-24 ans) et/ou quant à la variable d'identification D, ou si la sous-population B est hétérogène quant à la variable étudiée C et quant à la variable d'identification D mais qu'il n'y a pas de corrélation entre C et D, la mesure de la variable C effectuée à partir des individus du sous-ensemble I sera identique à la mesure effectuée à partir des individus de la sous-ensemble II, et donc, sera représentative de l'ensemble de la sous-population B. Si ces conditions ne sont pas remplies, la mesure de la variable C à partir des individus du sous-ensemble I peut différer de l'occurrence réelle de cette variable à l'intérieur de la sous-population B. Et s'il y a une différence, étendre l'information issue du sous-ensemble I à l'ensemble des individus de la sous-population B constituera un biais.

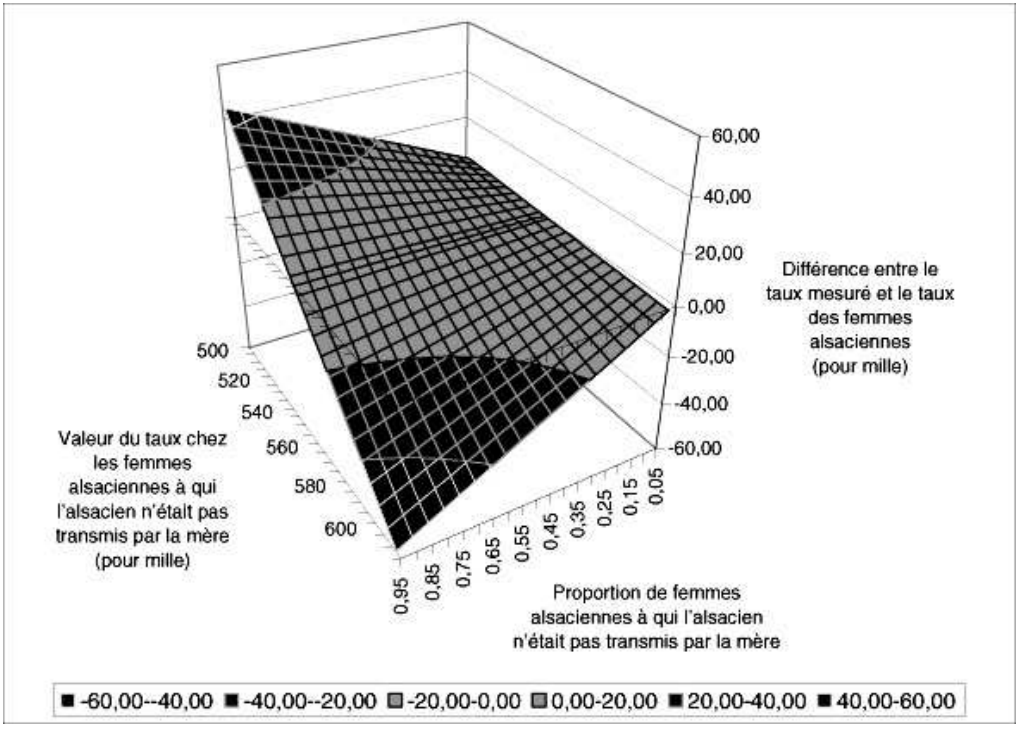
Le biais sera fonction de la dépendance qui existe entre B et D d'une part et C et D d'autre part. Si la corrélation entre B et D est assez forte et/ou si la corrélation entre C et D est assez faible, le biais peut être négligé. Au contraire, si la corrélation entre B et D n'est pas assez forte et si la corrélation entre C et D n'est pas assez faible, le biais peut être relativement élevé. Ainsi, la différence entre la valeur de C mesurée dans le sous-ensemble I et la sous-population B peut être posée comme étant égale à la différence de la valeur de C mesurée dans le sous-ensemble I et celle inconnue du sous-ensemble II, pondérée par le poids du sous-ensemble II dans la sous-population B, lui aussi inconnu. Une corrélation forte entre B et D assure un poids faible au sous-ensemble II alors qu'une corrélation faible entre C et D assure des valeurs de C proches, au niveau des deux sous-ensembles.

Si la transmission de la langue par la mère et l'intensité du premier départ du foyer parental à 20-24 ans ne sont pas corrélées (graphique n°1), quelle que soit la proportion des femmes alsaciennes à qui l'alsacien n'était pas transmis par la mère, le taux mesuré à travers la variable d'identification se confond avec le taux de l'ensemble des femmes alsaciennes. Si le fait d'être une femme alsacienne et la transmission de la langue par la mère sont très fortement corrélés, quel que soit le taux de premier départ du foyer parental à 20-24 ans chez les femmes alsaciennes à qui l'alsacien n'était pas transmis par la mère, le taux mesuré et recherché se confondent. Dans la majorité des cas simulés le biais se trouve entre -20 pour mille et 20 pour mille et il ne dépasse plus ou moins 40 pour mille que si la différence entre les valeurs des taux chez les femmes alsaciennes à qui l'alsacien était transmis et celles à qui elle ne l'était pas est relativement élevée et si la transmission de l'alsacien par la mère ne concerne qu'une minorité de femmes alsaciennes[6].

[6] Dans cet exemple issu d'une enquête, un écart inférieur à plus ou à moins 40 pour mille peut être l'effet de l'aléa. Seul un écart supérieur à plus ou à moins 40 pour mille permet de distinguer deux taux avec un risque inférieur à 5%.

AIDELF. 2008. DÉMOGRAPHIE ET CULTURES - Actes du colloque de Québec (Canada, 25-29 août 2008), Association internationale des démographes de langue française, 1 212 pages.

GRAPHIQUE 1 : MODÉLISATION DE LA DIFFÉRENCE ENTRE LE TAUX MESURÉ ET LE TAUX RECHERCHÉ.



Source : Étude de l'Histoire Familiale de 1999, CP.

En augmentant la quantité d'information, nous remarquons que les écarts simulés entre le taux mesuré et le taux recherché ne sont pas tous possibles et certains sont moins vraisemblables que d'autres. Nous savons que l'effectif des femmes des génération 1950-1954 est estimé à 2 166 480 individus dont 41 210[7] à qui l'alsacien était transmis par la mère et que le taux du premier départ du foyer parental à 20-24 ans des femmes du groupe de générations 1950-1954 est de 507,44 pour mille pour l'ensemble des femmes interrogées dans le cadre de l'EHS 1999.

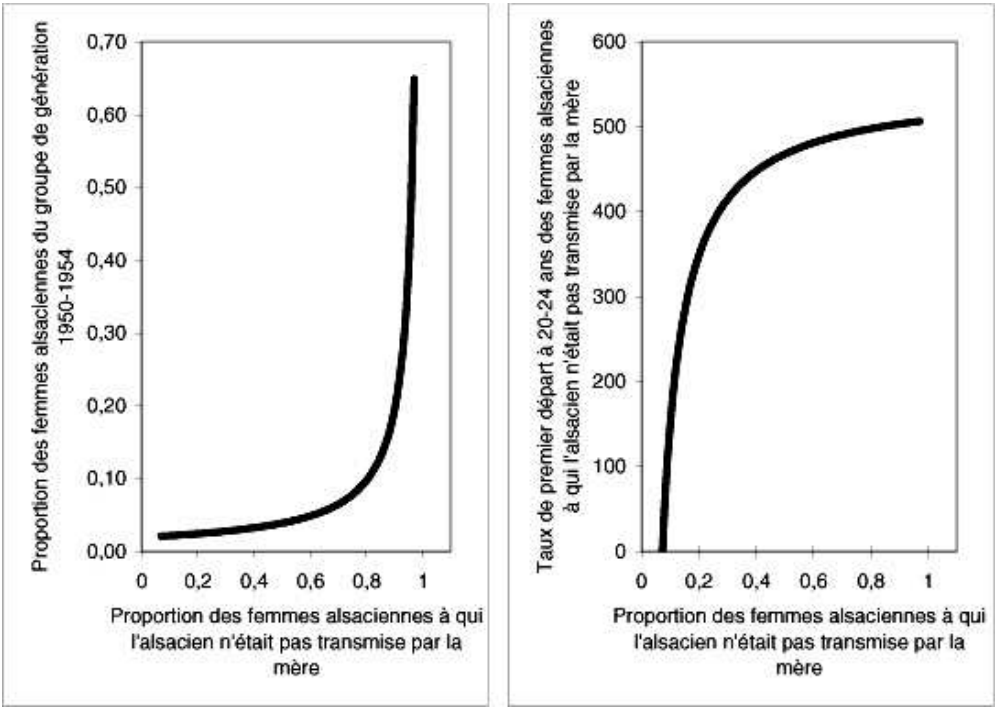
Sur le graphique n°2, nous avons représenté la proportion de femmes alsaciennes parmi les femmes de ces générations interrogées en 1999 comme une fonction de la proportion de femmes alsaciennes à qui l'alsacien n'était pas transmis par la mère parmi les femmes alsaciennes. Nous avons rapporté sur le graphique n°3 la valeur du taux de premier départ à 20-24 ans des femmes alsacienne à qui l'alsacien n'était pas transmis par la mère avec la condition que le taux à 20-24 ans de l'ensemble des femmes alsaciennes de ce groupe de générations soit au moins égale au taux à 20-24 ans de l'ensemble des femmes de cette groupe de génération, alsacienne ou non.

La lecture de ces deux graphiques nous informe que la structuration de la population entre alsaciennes et non-alsacienne et la différence de comportement quant au premier départ entre ces deux sous-populations ont une influence sur le biais, même si la variable d'identification est

[7] Soit 1,9%.

exclusive[8]. Si le taux à 20-24 ans des alsaciennes est au moins égal à celui de l'ensemble des femmes interrogées, il existe une corrélation négative entre le premier départ à 20-24 ans et la (non)transmission de l'alsacien par la mère, qui est d'autant plus soulignée que la proportion de femmes alsaciennes à qui l'alsacien n'était pas transmis par la mère décroît (atteignant même des valeurs aberrantes pour une proportion inférieur à 25%). La proportion des femmes alsaciennes à qui l'alsacien n'était pas transmis par la mère ne peut être supérieure à 80%, à moins que les femmes alsaciennes représentent plus de 10% des femmes de ce groupe de générations.

GRAPHIQUE°2 ET 3 : LA PROPORTION DES FEMMES ALSACIENNES ET LE TAUX DE PREMIER DÉPART (CONDITIONNÉ) EN FONCTION DE L'INTENSITÉ DE LA TRANSMISSION PAR LA MÈRE.



Source : Étude de l'Histoire Familiale de 1999, CP.

Dans l'ensemble deux cas sont possibles, soit le taux de premier départ du foyer parental à 20-24 ans des femmes alsaciennes du groupe de génération 1950-1954 est supérieur de façon significative à celui de l'ensemble des femmes de ces générations, soit le taux n'est pas supérieur (au mieux identique sinon supérieur de façon non significatif) et il existe une dépendance entre le premier départ à 20-24 ans et la transmission de l'alsacien par la mère qui est d'autant plus forte que l'intensité de la transmission est élevée. Dans cette situation, s'il existe en plus une corrélation entre la transmission et un phénomène perturbateur (mortalité,

[8] Même si P(B/D) = 1, $P(D/B) = \frac{1}{1+(P(B/\overline{D})*P(\overline{D})/P(D))} = \frac{1}{1+(\frac{P(B\cap\overline{D})}{P(D)})}$

migration, etc.), il sera préférable d'étudier le premier départ de façon séparé entre les femmes à qui l'alsacien a été transmis et pour qui elle ne l'a pas été[9].

1.2. Identification des femmes du groupe de génération 1945-1954 à qui l'arabe ou les langues berbères a été transmis par la mère à partir du nombre d'enfants eus.

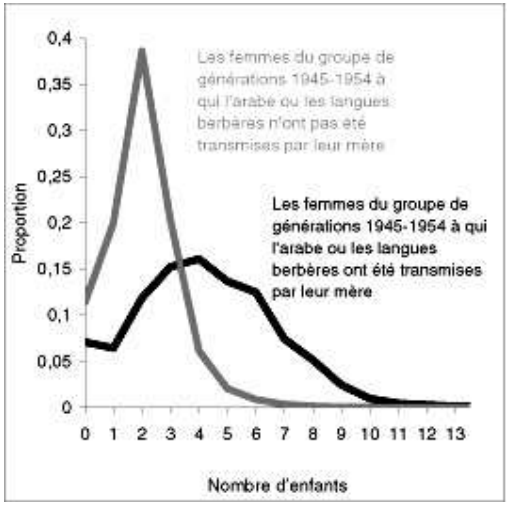
Une sous-population peut être caractérisée par un comportement différent du reste de la population mais cette différence peut ne pas suffire à identifier indirectement les individus appartenant à cette sous-population. En plus de l'objectif d'identification, d'estimation du poids ou de vérification d'autres comportements différentiels, la qualité de la variable d'identification dépendra conjointement du poids de la sous-population en question et de l'homogénéité du comportement par rapport à la variable d'identification à l'intérieur des deux sous-populations. Les conditions que nous avons présenté plus haut dans le cas simple restent valables dans ce cas plus complexe. Mais à ces conditions s'ajoute, concernant le sous-ensemble III (voir schéma n°1), la nécessité d'avoir une incidence de D assez faible dans la sous-population A pour un poids donné de la sous-population B. L'absence de vérification de cette condition ne se traduit pas forcement par un biais[10]. D'abord, les sous-ensembles I, II et III peuvent constituer une entité homogène, dans ce cas, quelque soit leur poids respectifs il n'y aura pas de biais. Toutefois cette homogénéité peut provenir d'une absence de comportement différentiels entre les sous-populations ou d'une dépendance forte entre l'appartenance au sous-ensemble III, via D, et la variable C. Enfin, le biais provenant du sous-ensemble III peut s'annuler avec le biais engendré par le sous-ensemble II. Dans les deux cas, même avec une superposition médiocre, on peut avoir et une bonne estimation de la valeur de C dans la sous-population B et une bonne estimation du poids de B dans l'ensemble.

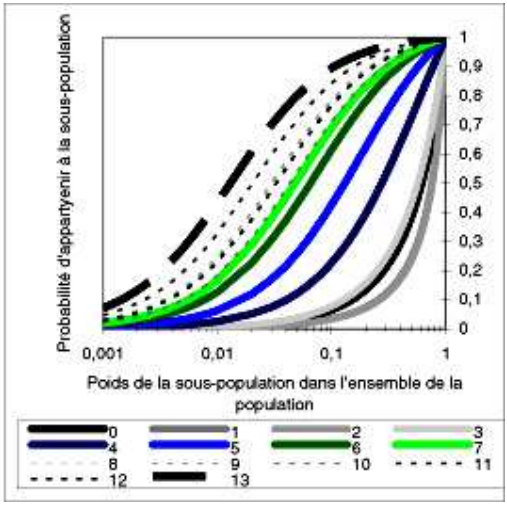
9 C'est à dire, éclater la sous-population B à deux sous-ensembles, ce qui possèdent la caractéristique d'identification et ce qui ne la possèdent pas. Les conditions d'élaboration d'un indice synthétique non biaisé reste valables pour un indice synthétique issue d'une estimation indirecte. Les conditions que nous présentons s'ajoute à ces conditions déjà connues. L'absence d'hétérogénéité relative à la variable d'identification ne signifie pas que la population est homogène face au phénomène étudié ou aux différents phénomènes perturbateurs. Nous n'allons pas approfondir d'avantage le cas de l'hétérogénéité d'une sous-population face à une variable d'étude et de perturbation. Pour plus de détail sur la condition nécessaire pour élaborer un indice synthétique non biaisé voire BLAYO (1991).

10 Voire aussi plus loin, « II-1) Modélisation de l'appartenance avec des variables d'identification ».

AIDELF. 2008. DÉMOGRAPHIE ET CULTURES - Actes du colloque de Québec (Canada, 25-29 août 2008), Association internationale des démographes de langue française, 1 212 pages.

GRAPHIQUE N°4 ET 5 : RÉPARTITION DES FEMMES PAR NOMBRE D'ENFANTS.
PROBABILITÉS D'APPARTENIR À LA SOUS-POPULATION.





Source : Étude de l'Histoire Familiale de 1999, CP.

Dans quelle mesure le nombre de naissance eus d'une génération qui a terminé sa vie féconde peut être utilisé comme identifiant indirect ? Nous avons représenté sur le graphique n°4 la distribution de deux groupes de femmes des générations 1945-1954, celles à qui l'arabe ou les langues berbères ont été transmis par leur mère (sous-population B) et les autres, selon le nombre d'enfants eus[11] déclaré à l'Étude de l'Histoire Familiale de 1999. En utilisant les deux distributions du graphique n°5, nous avons estimé des probabilités d'appartenir à la sous-population B pour un nombre d'enfants eus donné en fonction du poids de celles-ci dans l'ensemble des femmes du groupe de générations 1945-1954.

Dans ce cas, indépendamment de la finalité de l'identification indirecte, même si les deux distributions paraissent différentes, l'utilisation du nombre d'enfants eus comme variable d'identification est peu satisfaisant. La réduction de la part des individus inclus à tort dans la sous-population B se traduit par une augmentation plus importante de la part des individus exclus à tort, et inversement, la réduction de la part des individus exclus à tort de la sous-population B se traduit par une augmentation encore plus forte des individus inclus à tort.

Dans le cas où nous recourons à une variable d'identification afin de comparer les deux sous populations selon une variable étudiée quelconque, le nombre d'enfants eus devient un identifiant d'autant plus médiocre que le poids de la sous-population est faible. Ce qui signifie que seules les différences de comportement assez soulignées entre les deux sous-populations seront remarquées, à condition que la variable d'identification et les variables étudiées ne soient pas corrélées. Dans notre cas, les femmes à qui l'arabe ou les langues berbères a été transmises par leur mère représentent 3% du groupe de génération 1945-1954. Classer les femmes qui ont eu 8 enfants ou plus comme appartenant à la sous-population B se traduit par 50% de femmes inclues à tort et 90% de femmes exclues à tort, tandis que le meilleure compromis entre les exclues et inclues à tort c'est à dire le classement des femmes qui ont eu 6 enfants ou plus donne 60% d'inclues à tort et 70% d'exclues à tort.

11 Sans les enfants adoptés.

De même manière, « le nombre d'enfant eus » est une mauvaise variable pour l'identification au niveau individuel, les deux distributions sont trop proches et la distribution au niveau de la sous-population B est trop étalée pour une telle tentative. Les variables d'identification doivent générer des distributions assez distinguées dans les sous-populations pour pouvoir identifier nettement les individus.

Néanmoins, elle peut permettre une estimation de la proportion des femmes appartenant à la sous-population B parmi le groupe de génération 1945-1954. La distribution de l'ensemble des femmes du groupe de générations 1945-1954 selon le nombre d'enfants eus dépend d'une part de la distribution des deux sous-populations et d'autre part du poids respectif de chacune d'entre elles. Dans le cas où on connaît les distributions des deux sous-populations et la distribution de l'ensemble, il est aisé d'estimer le poids respectif des deux sous-populations. Cependant, l'exercice est périlleux si les distributions utilisées ne sont pas appropriées et le biais sera d'autant plus conséquent que la proportion de la sous-population que nous voulons estimer est faible.

En somme, la qualité d'une distribution qui nous permet d'identifier une sous-population dépend tant de l'écart du comportement moyen que de la moyenne des écarts de comportements et du poids de la sous-population à identifier. La distribution de la première sous population se traduit par une descendance finale de 4,23 enfants par femmes, celle de la seconde de 2,03 enfants par femmes. Et bien que la différence soit importante, les différences des deux distributions sont insatisfaisantes.

2. Augmentation du nombre de variable d'identification

Augmentation du nombre de variable d'identification ne se traduit pas forcement par une amélioration de la qualité de l'estimation. En plus des conditions citées ci haut, l'augmentation du nombre de variable ne garantie une hausse de la superposition que si chaque nouvelle variable élimine un peu plus de ceux qui sont inclus à tord ou réhabilite d'avantage de ceux qui sont exclus à tord jusqu'à là. La façon d'imbriquer les variables peut influencer la superposition. Par exemple, augmenter les intersections peut augmenter le nombre d'exclus à tort, alors qu'augmenter le nombre d'union peut augmenter le nombre d'inclus à tort[12], ce qui nécessite une bonne combinaison des « et » et des « ou ». Enfin, quand on utilise une seule caractéristique comme variable d'identification, la mesure effectuée peut être attribuée au sous-ensemble (par exemple les femmes à qui l'alsacien était transmise par la mère) issus de la variable d'identification au lieu de la sous-population que nous voulons isoler (par exemple les femmes alsaciennes), ce qui peut être moins évident si on augmente le nombre de variable d'identification et surtout si l'appartenance est donnée à travers un model.

2.1. Modélisation de l'appartenance avec des variables d'identification.

La sous-population que nous voulons étudier peut se distinguer du restant de la population selon certaines caractéristiques (démographiques, socio-économiques ou autres) mais ces caractéristiques peuvent ne pas suffire à identifier les individus appartenant à cette sous-population, à relever une mesure correcte de la variable étudié ou à estimer la proportion de la sous-population en question.

Pour illustrer les biais découlant de l'augmentation du nombre de variable d'identification nous avons recourue à une modélisation simple de la probabilité d'appartenance des individus et de leur classement à un sous-groupe définie par « la transmission d'une langue autre que le français par leur mère[13] ». Nous avons élaboré huit modèles[14]. L'évaluation de la qualité du

12 Puisque $\forall$ P(K) et $\forall$ P(L), $P(K \cap L) \leq P(KUL)$.

13 Y compris les langues régionales de France.

modèle est basée sur l'estimation de la proportion de la sous-population que nous voulons isoler (qui est de 25,33 pour cent), l'estimation du variable que nous voulons mesurer à l'intérieur de cette sous-population (ici l'âge moyenne qui est de 50,25 ans) et le potentiel du modèle à identifier les individus appartenant à la sous population[15]. Les trois premier modèles abritent une variable fortement corrélée avec « la transmission de la langue par la mère » qui est « la transmission de la langue par le père », 90% des individus qui ont une mère qui à transmise une autre langue ont aussi reçu de leur père une autre langue et 93% des individus qui ont un père « transmetteur » ont aussi reçu une autre langue par leur mère. Classement des individus à partir de leur probabilité conditionnelle est effectué de deux manières, déterminé[16] ou aléatoire[17].

Les trois meilleurs modèles sont ceux qui incluent une variable fortement corrélé avec l'appartenance (modèles n°1, 2 et 3), quel que soit le mode de classement des individus, même si la qualité de l'estimation est supérieure avec un classement déterminé. L'augmentation de la superposition entre le modèle n°1 et 2 est liée à l'introduction de la variable «transmission de la langue aux enfants » qui discrimine les faux exclus du modèle n°1, des individus avec une mère transmetteur et un père non transmetteur. Le modèle n°3 dispose d'une qualité intermédiaire entre le n°1 et le n°2, les variables sexe, C.S.P., diplôme, région de résidence et taille de l'unité urbaine de résidence, ayant un apport plus faible que la variable «transmission de la langue aux enfants ».

Les modèles n°4, 5, 6, 7 et 8, avec un mode de classement déterminé, sont tous de qualité médiocre, avec des superpositions faibles, ils sous estiment la proportion et surestiment l'âge moyen, à l'exception de l'estimation de l'âge moyen issus du modèle n°4 et 6. Le classement aléatoire assure une estimation non biaisée des proportions, même quand le modèle est de qualité très médiocre (modèles n°5, 6, 7 et 8). L'âge moyen est sous estimé mais plus proche de la valeur au sein de la sous-population à isoler que de la valeur dans le reste de la population (46,08 ans). La superposition est faible mais supérieur à ce qu'on aurait observé[18] si on classait les individus de façon aléatoire avec une probabilité individuelle identique de 0,2533. Quel que soit le mode de classement des individus, la qualité du modèle n°5 est nettement inférieure du modèle n°4, ce qui est engendré par l'élimination des variables région de résidence et taille de l'unité urbaine de résidence. L'apport en quantité d'information de ces deux dernières variables, affaiblie face à la variable «transmission de la langue aux enfants » dans le modèle n°3, est mieux saisi. La structuration du modèle (n°6) ou croiser les variables pour tenir compte des effets conjointes (modèle n°8) peuvent augmenter la qualité du modèle.

---

14 Voir l'annexe pour plus d'explication sur les modèles.

15 Reflété par le rapport entre les membres identifiés et l'ensemble des membres ou des identifiés {P(D∩B)/P(DUB)}.

16 Tout individu ayant une probabilité supérieure à 50% est classé comme appartenant à la sous-population que nous voulons isoler (il est possible de choisir un autre seuil de classement, par exemple 90% ou la probabilité d'appartenance mesurée au niveau de l'ensemble de la population).

17 On tire au hasard un nombre entre 0 et 1 (distribution uniforme), si le nombre est inférieur à la probabilité conditionnelle, individu est classé comme appartenant à la sous-population que nous voulons isoler (il est possible d'utiliser une distribution autre que uniforme).

18 Qui est de 14,5 % ($\frac{(0{,}2533)^2}{(0{,}2533)^2 + 2*(0{,}2533*(1-0{,}2533))}$).

TABLEAU°3 : QUALITÉ D'ESTIMATION DES DIFFÉRENTS MODÈLES.

	Estimation de la proportion de personnes à qui la mère a transmise une langue autre que le français (pour cent). Valeur mesurée : 25,33%		Estimation de l'âge moyen (âge révolu). Valeur mesurée : 50,25 ans		Superposition du modèle {P(D∩B)/P(DUB)} (pour cent).	
Modèle n°	Déterminé	Aléatoire	Déterminé	Aléatoire	Déterminé	Aléatoire
1	25,62	25,35	50,25	49,93	82,91	71,05
2	25,81	25,31	50,25	49,87	83,48	71,82
3	25,71	25,34	50,23	50,10	83,05	71,50
4	6,81	25,35	51,02	49,01	16,59	20,82
5	1,14	25,33	69,61	48,91	2,31	16,55
6	3,68	25,31	51,36	48,97	8,15	17,73
7	0,20	25,27	72,86	48,24	0,32	16,03
8	3,19	25,26	57,78	48,21	6,48	16,38

Source : Étude de l'Histoire Familiale de 1999, CP.

Le modèle n°6 avec un classement déterminé propose une estimation acceptable de l'âge moyen, légèrement surestimé, puisque l'âge moyen des individus classés correctement, ceux inclut à tord et ceux exclus à tord (le sous-ensemble qui a le plus de poids) sont quasi-identique (tableaux n°4). Le modèle n°6 avec classement aléatoire propose aussi une estimation pas trop biaisé, cette fois sous estimé. Le poids des exclus et inclus à tord sont identiques et plus élevé que les individus correctement classés. Le biais est engendré par un âge moyen des individus classé à tord qui se positionne entre la valeur au sein de la sous-population à identifier et le reste de la population. Et si le modèle n°8 avec classement déterminé surestime largement l'âge moyen, c'est parce qu'il amplifie, contrairement aux deux classements du model n°6 et le modèle n°8 avec classement aléatoire, la corrélation existante entre l'âge et la probabilité d'appartenir à la sous-population à isoler (tableau n°4 et graphique n°6).

TABLEAU 4 : PROPORTION ET ÂGE MOYEN DES PERSONNES SELON LEUR APPARTENANCE OBSERVÉ ET ESTIMÉ.

	Déterminé		Aléatoire	
Proportion/Âge moyen	Classés comme ayant une mère qui n'a transmise que le français	Classés comme ayant une mère qui a transmise une langue autre que le français	Classés comme ayant une mère qui n'a transmise que le français	Classés comme ayant une mère qui a transmise une langue autre que le français
Observé comme	Modèle n°6			
Une personne à qui la mère n'a transmise que le français	73,18 / 45,96	1,49 / 51,55	56,98 / 45,55	17,69 / 47,77
Une personne à qui la mère a transmise une autre langue	23,14 / 50,15	2,19 / 51,23	17,70 / 49,60	7,62 / 51,76
Observé comme	Modèle n°8			
Une personne à qui la mère n'a transmise que le français	73,21 / 45,84	1,49 / 57,95	56,53 / 45,76	18,14 / 47,06
Une personne à qui la mère a transmise une autre langue	23,59 / 49,70	1,74 / 57,63	18,21 / 49,90	7,12 / 51,14

Source : Étude de l'Histoire Familiale de 1999, CP.

TABLEAU 5 : MODÈLES N°3, 4 ET 4BIS APPLIQUÉS AUX SOUS-ENSEMBLES IMMIGRÉS ET NON IMMIGRÉS.

		Estimation de la proportion (pour cent). Valeur mesurée : Immigrés 93,15% Natif 18,63%		Estimation de l'âge moyen (âge révolu). Valeur mesurée : Immigrés 47,54 Natif 51,64		Superposition du modèle {P(D∩B)/P(DUB)} (pour cent).	
Modèle	Statut	Déterminé	Aléatoire	Déterminé	Aléatoire	Déterminé	Aléatoire
3	Immigré	91,56	84,1	47,59	47,66	96,97	88,8
3	Non Immigré	19,21	19,52	51,52	51,18	77,03	64,65
4	Immigré	15,4	33,57	48,32	48,69	16,08	33,16
4	Non Immigré	6,08	24,36	51,87	49,26	17,15	17,73
4b	Immigré	100	93,29	47,66	47,54	93,15	87,65
4b	Non Immigré	4,45	18,69	53,78	50,52	15,52	16,97

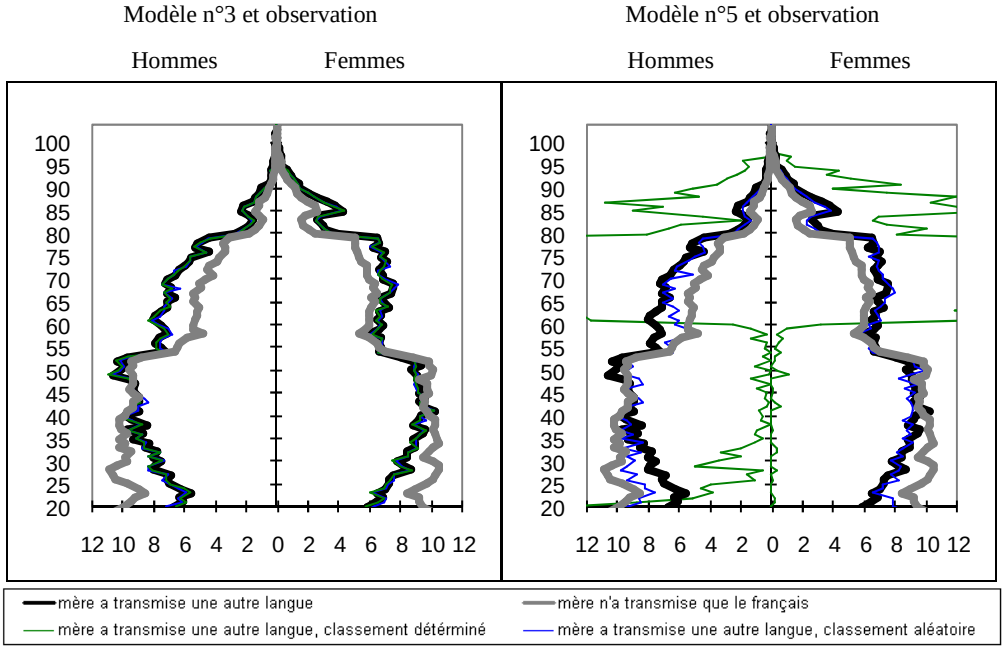
Source : Étude de l'Histoire Familiale de 1999, CP.

AIDELF. 2008. DÉMOGRAPHIE ET CULTURES - Actes du colloque de Québec (Canada, 25-29 août 2008), Association internationale des démographes de langue française, 1 212 pages.

GRAPHIQUE 6 : PROBABILITÉ D'APPARTENIR À LA SOUS-POPULATION DES INDIVIDUS À QUI LA MÈRE A TRANSMISE UNE LANGUE AUTRE QUE LE FRANÇAIS SELON L'ÂGE

Source : Étude de l'Histoire Familiale de 1999, CP.

GRAPHIQUES 7 ET 8 : PYRAMIDES OBSERVÉ ET ISSUS DES MODÈLES N°3 ET 5, SELON LA MÉTHODE DE CLASSEMENT.



Source : Étude de l'Histoire Familiale de 1999, CP.

Appliquer le modèle à une date ou à un niveau géographique différents nécessite des probabilités d'appartenance similaires selon les variables d'identification retenues[19], sans quoi il y a un risque de biais (tableau n°5), d'autant plus souligné que quand la sous-population à isoler est hétérogène. Les personnes à qui leur mère a transmise une langue autre que le français sont constituées de différents groupes de personnes, par exemple, immigrés[20] et non immigrés. Ces sous-groupes peuvent avoir des caractéristiques différentes, selon les variables d'identification ou selon les variables de mesure. Le modèle incorpore cette hétérogénéité au niveau national, ce qui nécessite que le lieu ou le moment d'application du modèle soient représentatif de l'hétérogénéité du niveau national et du date de collecte des données à partir des quelles le modèle à été élaboré, au risque de biais (modèle n°4). Ce qui peut être évité, si la sous-population que nous voulons isoler est un ensemble homogène quant aux variables d'identification (ce qui est assuré par la variable « type de langue transmise par le père » au niveau du modèle n°3) ou quant aux variables à mesurer. Éclater le modèle entre immigré et natifs est aussi une solution (modèle n°4b).

L'existence des variables fortement corrélés permet aussi une analyse plus raffinée (Graphique n°6). Cependant, si les variables d'identification sont faiblement corrélés, le classement déterministe est à rejeter alors que le classement aléatoire rend possible une analyse raffinée jusqu'à un certain point, comme c'est le cas sur le graphique n°7 chez les femmes.

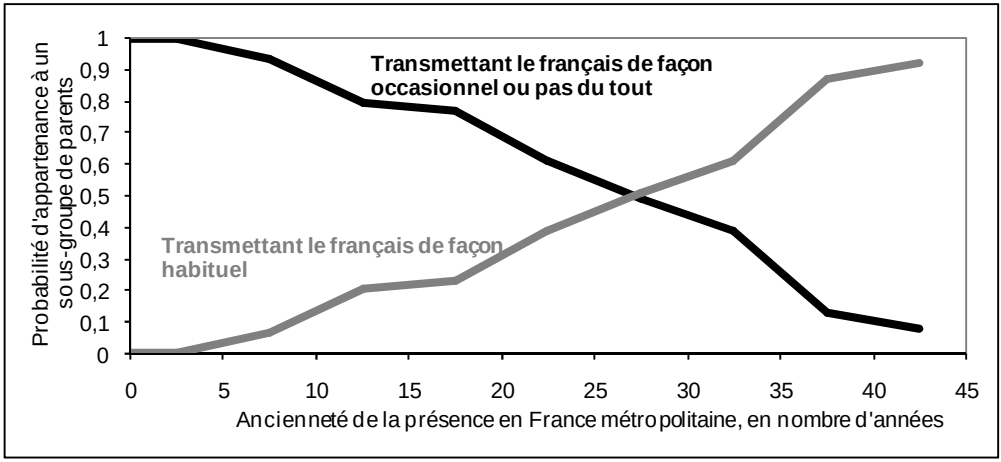
2.2. Approches alternatives : l'appartenance et le degré d'implication.

Jusqu'ici l'appartenance est abordée dans une optique binaire, où les individus appartiennent ou pas à une sous-population. Or, en ce qui concerne les sous-groupes culturels, l'appartenance individuelle peut se faire à des degrés d'implication différents et c'est le degré de cette implication qui peut expliquer un comportement différentiel vis-à-vis du reste de la population. Certaines méthodes hétérodoxes[21] permettent d'aller au-delà du cadre dualiste, tout comme certaines analyses déjà existantes, comme par exemple quand on tienne compte de la date d'entrée en France chez les immigrés, ou un autre événement constituant d'une cohorte, du moment où la durée passée entre cet événement constituant et la date de l'occurrence de l'événement étudié reflète aussi une variation du degré d'implication.

[19] Les probabilités conditionnelles.
[20] Au sens de "né(e) de nationalité étrangère et à l'étranger".
[21] Le raisonnement bayésien ou la logique floue.

AIDELF. 2008. DÉMOGRAPHIE ET CULTURES - Actes du colloque de Québec (Canada, 25-29 août 2008), Association internationale des démographes de langue française, 1 212 pages.

GRAPHIQUE 9 : PROBABILITÉ D'APPARTENANCE À UN SOUS-GROUPE DE PARENTS, LES IMMIGRANTS D'ORIGINAIRE D'ITALIE D'ESPAGNE ET DU PORTUGAL DU GROUPE DE GÉNÉRATION 1950-1959.



Source : Étude de l'Histoire Familiale de 1999, CP.

Le graphique n°9, nous laisse croire que plus l'ancienneté est élevé plus on a de probabilité de transmettre aux plus jeunes des enfants le français de façon habituel (avec ou sans la langue du pays d'origine) quand ils avaient 5 ans. Or, l'information est récoltée de façon rétrospective, avec certainement des effets de sélection s'il y a une corrélation entre l'émigration (de retour) et une mauvaise implication linguistique. Puis, Les typologies des migrants peuvent être corrélées avec la date d'entrée en France et, à durée de séjour égale, l'histoire génésique des individus peut être très variée. Enfin, le plus important, ce n'est pas l'ancienneté du séjour individuelle au moment de l'enquête qui va nous informer de l'implication linguistique et sa transmission mais la durée écoulé entre l'entré en France et les $5^{\text{ièmes}}$ anniversaires de chacun des enfants (pour qui l'enquêté a répondu) de façon séparé, dans des cohortes d'entrées pour ce groupe de générations 1950-1959.

Conclusion

Seule une variable d'identification qui se superpose quasi-totalement avec l'appartenance à un sous-groupe culturels est exempte des risques d'erreurs liées à l'identification indirecte. Cependant, il sera faut de rejeter tous les autres variables d'identification en les considérant comme source de biais. La qualité d'une variable d'identification dépend de la qualité et de la quantité d'information que le chercheur possède et sera jugé par une réflexion sur les conditions de son application. Elle dépend entre autre de l'objectif (estimation de leur proportion, mesure différentielle d'une variable étudié ou identification individuelle) et de l'étendue souhaitée de l'identification (autant que possible même s'il y a des individus inclus à tord ou une partie mais avec moindre d'inclus à tord possible) ainsi que du poids de la sous-population à identifier.

L'augmentation du nombre de variable assure une meilleure identification que si les variables ou leur imbrication assure une bonne corrélation. Si la corrélation est faible, les classements aléatoires peuvent au moins servir à estimer correctement la proportion du sous-groupe ou la mesure différentielle d'une variable étudié, sans pour autant identifier les individus, mais ils ont aussi leur limite face à des variables très faiblement corrélé. Pour utiliser les probabilités d'appartenance conditionnelles (selon divers caractéristiques) à un niveau ou

une époque différente il faut que les probabilités conditionnelles restent peu changé, à risque de biais graves.

Interroger les individus selon leurs caractéristiques constituantes peut ne pas suffire pour une identification des appartenants d'un sous-ensemble culturels. Seule l'interrogation directe et indirecte des individus sur leurs appartenances et leur degré d'implication, mise en parallèle avec diverses caractéristiques culturelles et comportementales, peut nous informer de la solidité des variables d'identification. Même dans ce cas, application des variables d'identification à un niveau et / ou une époque différents peut être problématiques.

BIBLIOGRAPHIE

BLAYO, Chantal. « Choix des cohortes et des sous-cohortes : règles générales et application à l'avortement », Population, n° 6, 1991.

PARANT, Éric ; BERNIER, Jacques. « Le raisonnement bayésien. Modélisation et inférence », Springer, 2007.

AIDELF. 2008. DÉMOGRAPHIE ET CULTURES - Actes du colloque de Québec (Canada, 25-29 août 2008), Association internationale des démographes de langue française, 1 212 pages.

ANNEXE

MODÈLES DU TABLEAU N°3

Modélisation des probabilités d'être une personne à qui la mère a transmise une autre langue que le français, sachant un ensemble de caractéristiques données.

Distribution binomiale avec une fonction de lien (link function) logit.

Modèle n°	Variables explicatives	Structuration	Log Likelihood	Type de la régression
1	Transmission de la langue par le père	Aucune	-7 324,67	Régression linéaire généralisée
2	Transmission de la langue par le père Transmission aux enfants	Aucune	-6 967,58	Régression linéaire généralisée
3	Transmission de la langue par le père Sexe C.S.P. (25 classes) Diplôme (10 classes) Région de résidence Taille de l'unité urbaine de résidence (9 classes)	Aucune	-7 097,24	Régression linéaire généralisée
4	3 sans la variable Transmission de la langue par le père	Aucune	-21 455,15	Régression linéaire généralisée
4b	3 sans la variable Transmission de la langue par le père	Qualité d'immigré	2 sous-modèles Immigrés : -846,30 Natifs : -15 809,15	Régression linéaire généralisée
5	4 sans les variables Région de résidence Taille de l'unité urbaine de résidence (9 classes)	Aucune	-23 057,60	Régression linéaire généralisée
6	4 sans les variables Région de résidence Taille de l'unité urbaine de résidence (9 classes)	Taille de l'unité urbaine de la résidence de l'ego	9 sous-modèles ; entre -5 478 et -857	Régression linéaire généralisée
7	C.S.P. (25 classes) Taille de l'unité urbaine de résidence (9 classes)	Aucune	-23 279,82	Régression linéaire généralisée
8	C.S.P. (25 classes) croisée avec Taille de l'unité urbaine de résidence (9 classes)	Aucune	-23 062,13	Régression linéaire généralisée