Les estimations par capture-recapture ou par système multiple : quelques éléments théoriques

Document de travail

N° M2025-01 – Janvier 2025

Pascal ARDILLY
Heidi KOUMARIANOS

Insee
Mesurer pour comprendre

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES
Série des documents de travail « Méthodologie Statistique »
de la Direction de la Méthodologie et de la Coordination Statistique et Internationale

M 2025/01

Les estimations par capture-recapture ou par système multiple : quelques éléments théoriques

Pascal ARDILLY
Heidi KOUMARIANOS

Insee

JANVIER 2025

Remerciements :
Les auteurs remercient Olivier Haag, Eric Lesage et Corinne Prost pour leur relecture attentive du document, ainsi que les participants au groupe de travail méthodologique dédié au répertoire statistique des individus et des logements de l'Insee.

Direction de la méthodologie et de la coordination statistique et internationale
Département des Méthodes Statistiques -Timbre L001 -
88 Avenue Verdier - CS 70058 - 92541 Montrouge Cedex - France -
Tél. : 33 (1) 87 69 55 00 - E-mail : DG75-L001@insee.fr - Site Web Insee : http://www.insee.fr

Résumé

Le dénombrement des populations compte parmi les principaux objectifs de la statistique publique. Lorsqu'on ne peut pas pratiquer de recensement, mais que l'on dispose de plusieurs (au moins deux) sources offrant chacune une couverture partielle de la population dont il faut estimer la taille, il est possible d'utiliser une méthode dite de « capture-recapture ». La théorie associée à cette méthode considère que chaque individu de la population a, pour chacune des sources considérées, une probabilité spécifique de lui appartenir. Elle s'appuie à la base sur des hypothèses fortes et multiples. L'hypothèse absolument contraignante, sans échappatoire possible, est celle de la connaissance des intersections entre sources : sur elle, repose le fondement des méthodes de capture-recapture. Par ailleurs, ces sources ne doivent pas, après traitements éventuels, contenir d'enregistrements traduisant une sur-couverture de la population réelle. Côté individus, la technique de base – conduisant à l'estimateur de Lincoln-Petersen – suppose que l'éventuelle présence d'un individu dans une source donnée est indépendante de son éventuelle présence dans toute autre source, et qu'elle est par ailleurs indépendante de l'éventuelle présence de tout autre individu dans cette même source. Il faut enfin supposer que les probabilités d'appartenance à une source quelconque dépendent de la source mais ne dépendent pas de l'individu. Le développement des méthodes présentées dans ce document est structuré autour du relâchement de tout ou partie de ces hypothèses.

Mots-clé : Population, estimation, capture-recapture, sur-couverture, sous-couverture, modèle log-linéaire, sources multiples, appariements

Classification JEL : C13, C80

Les estimations par capture-recapture ou par système multiple

Quelques éléments théoriques

Table des matières

AVANT PROPOS en guise de guide de lecture

Ce document expose des méthodes d'estimation de taille de population, que l'on appelle communément « *méthodes de capture-recapture* ». On se place dans un contexte où on dispose de plusieurs sources, en nombre au moins égal à deux, chacune listant une partie de la population dont on cherche à estimer la taille. La situation se présente de telle manière qu'on ne peut hélas pas, au moyen d'appariements entre ces sources, reconstituer la population complète, ce qui revient à dire qu'il existe une partie de la population qui n'est présente dans aucune des sources disponibles. L'enjeu est donc d'estimer, d'une façon ou d'une autre, la taille de la sous-population « cachée ».

Différentes méthodes présentées dans ce document le permettent. Est-ce magique ? Évidemment non, car comme bien souvent lorsqu'on doit procéder à des estimations, les statisticiens contrebalancent leur ignorance d'une réalité infiniment complexe par des hypothèses… lesquelles sont précisément construites afin de contourner les obstacles. Le traitement mathématique de la question ne pose généralement pas de problème insurmontable, et la qualité de l'estimation produite est par conséquent dépendante de la pertinence de ces hypothèses initiales, lesquelles participent aux fondements de la méthode.

C'est en toute logique dans l'approche la plus simple que l'on pose le plus grand nombre d'hypothèses. La **partie I** traite ce premier contexte, qui produit un célèbre estimateur dit 'de Lincoln Petersen'. Cet estimateur, fondamentalement construit à partir d'un rapprochement des sources, apparaît dans un grand nombre d'applications pratiques, il est très facilement calculable et on peut en expliquer la logique à des non-spécialistes. Les différentes sous-parties de la partie I s'intéressent toutes à cet estimateur, mais elles apportent divers éclairages et des compléments à l'expression formelle de base. Ainsi, les sous-parties 1, 2 et 3 distinguent trois approches différentes pour justifier l'expression de l'estimateur lorsqu'on dispose de deux sources seulement, et toujours dans le cadre des hypothèses adoptées. La sous-partie 4 est centrée sur la qualité de l'estimation, évaluée au travers des indicateurs traditionnels de biais et de variance. La sous-partie 5 généralise l'estimateur en se plaçant dans une situation particulière où l'une des sources est une enquête de contrôle de type aréolaire. Quant à la sous-partie 6, elle permet de comprendre comment on peut adapter l'estimateur quand on dispose de plus de deux sources (ici trois).

La théorie de l'estimateur de Lincoln Petersen a été développée en considérant que chaque observation de chaque source en présence renvoie à un 'vrai' individu de la population. Il n'y a donc pas, par hypothèse, d'individu fictif dans les bases traitées. On présente cette hypothèse en disant qu'il n'y a pas de sur-couverture dans les bases. En pratique, c'est très fréquemment faux, car même les répertoires les mieux gérés comprennent une fraction d'observations obsolètes. La première sous-partie de la **partie II** propose une démarche pour réduire les effets de cette sur-couverture quand on utilise – encore et toujours – un estimateur de Lincoln Petersen. Par ailleurs, sans relation avec une éventuelle sur-couverture, l'estimateur de Lincoln Petersen n'apparaît théoriquement justifiable que si on est en mesure d'effectuer en amont un rapprochement parfait – un appariement - des sources considérées. Il faut donc être en capacité de constater que tel individu présent dans une source donnée est ou n'est pas présent dans chacune des autres sources. La seconde sous-partie de la partie II propose une adaptation de l'estimateur de Lincoln-Petersen dans un scénario où il existe des erreurs d'appariement… évidemment modélisées.

Les deux premières parties se focalisent donc sur l'estimateur de Lincoln Petersen, partant d'une situation qui s'avère largement idéalisée pour ce qui est de la version 'historique', et aboutissant à un estimateur modifié qui autorise la sur-couverture et certaines erreurs d'appariement. Au fil de l'eau, la théorie s'adapte jusqu'à un certain point, en relâchant certaines hypothèses d'origine pour qu'on puisse considérer que le nouveau contexte traduit la réalité de manière « acceptable ». Dit autrement, il subsiste toujours certaines

conditions, soit dans les données soit dans le processus de traitement, qu'il faut accepter pour qu'on puisse penser que l'estimation finale qui découle de la théorie est correcte. A la fin de la partie II, on est encore sous l'emprise de trois hypothèses fortes, mais on va pouvoir s'affranchir de deux d'entre elles : c'est tout l'objet de la **partie III.**

Il n'est plus possible de justifier l'expression de l'estimateur de Lincoln Petersen si on abandonne l'hypothèse d'indépendance entre les sources. Cette hypothèse considère que, pour tout individu donné de la population, son appartenance à une quelconque des sources ne préjuge en rien de son appartenance à n'importe quelle autre source prise parmi celles dont on dispose. Il est alors nécessaire de changer de paradigme et d'utiliser un tout autre outil pour procéder à l'estimation de taille. En la circonstance, le plus commun est la modélisation log-linéaire. La sous-partie 1 aborde cette méthode, après un développement introductif consacré à deux modèles spécifiques.

La seconde hypothèse fondamentale conditionnant l'expression de Lincoln Petersen est l'hypothèse d'homogénéité des probabilités d'appartenance des individus à une source donnée. Il faut comprendre par-là que dans la théorie de base, étant donnée une source tous les individus de la population ont la même chance d'être présents dans cette source (la probabilité commune varie néanmoins d'une source à l'autre). Relâcher cette hypothèse va faire perdre sa justification à l'estimateur de Lincoln Petersen. Dans ces conditions nouvelles, la sous-partie 2 apporte des éléments théoriques pour procéder à l'estimation de la taille de la population en formulant des hypothèses allégées sur les probabilités d'appartenance aux sources, tout en conservant par ailleurs l'hypothèse d'indépendance entre les sources.

Lorsqu'on dispose d'au moins trois sources, l'approche dite « *sample coverage* » propose une méthode pour traiter à la fois la dépendance entre sources et l'hétérogénéité des probabilités individuelles d'appartenance aux sources. Savante et audacieuse, elle semble être parmi les méthodes les plus générales dont on dispose. On la présente en sous-partie 3, avant de conclure.

Il y a une dernière hypothèse fondamentale qui résiste à tous les assauts : c'est celle de l'indépendance de comportement de capture entre individus. Cette hypothèse traduit le fait que le comportement en matière de capture d'un individu quelconque donné n'impacte aucun des comportements des autres individus de la population. Elle n'a jamais été remise en cause dans les modèles dont nous avons connaissance. Mais c'est peut-être aussi celle qui paraît la plus crédible en situation réelle...

Introduction

Le terme ‘capture-recapture’ évoque une expérience concrète de capture d’animaux dans un milieu sauvage. C'est effectivement le cas d'application historique de ce type de méthode, du moins c'est dans ce domaine d'application que le formalisme de ces méthodes a pris naissance.

La capture-recapture renvoie à un ensemble de techniques dont l’objectif est d’estimer une taille de population. Il s’agit donc d’une pure problématique de dénombrement, à l’exclusion de toute autre estimation. En particulier, ces méthodes ne permettent en aucun cas d’estimer des moyennes ou des proportions, ni tout autre paramètre plus ou moins complexe défini sur une population donnée, comme on a l’habitude de pratiquer dans les enquêtes par sondage traditionnelles. En revanche, la définition des populations concernées peut être basée sur des caractéristiques multiples : on n'est pas limité à la population complète constituant l’univers dans lequel on se place, on peut très bien s’intéresser à une sous-population quelconque. C’est ainsi qu’il est possible, avec la même méthodologie exactement que pour une population totale, d’estimer la taille d’une sous-population rare (difficile à joindre).

A la fin du 19ème siècle et au début du 20ème, plusieurs scientifiques ont cherché à estimer des tailles de populations animales en utilisant ces méthodes. On peut citer Petersen au Danemark en 1889 (estimation d’un effectif de poissons dans les pêcheries), Dahl en Norvège en 1917 (également sur des populations de poissons), puis Pearse et Lincoln aux Etats-Unis, en 1923 et 1930 (estimation de population de tortues et de gibiers d'eau). L'estimateur le plus classique, dit de Lincoln-Petersen, est ainsi nommé en référence à ces deux scientifiques, bien que d'autres scientifiques en aient fait usage (Goudie, Goudie 2007).

Mais ces méthodes peuvent s'appliquer plus largement au comptage de populations de toute nature, y compris humaines. On parle alors plutôt en anglais de *multiple-record system estimation* ou *dual system estimation* pour le cas de deux sources.
Le cadre d'obtention des données est différent : pour les populations animales, les données sont obtenues dans le cadre d'expériences (capture, marquage, recapture), alors que pour les populations humaines, il s'agit plus souvent de données déjà existantes qui sont mobilisées en exploitant des listes d'individus physiques. En toutes circonstances, on a la connaissance d’échantillons d’unités statistiques, mais pas de la population complète. Chao (2015) souligne deux principales différences entre le contexte du dénombrement animal et celui du dénombrement humain : dans le cas de populations animales, le nombre d'échantillons possibles est plus grand, et il existe naturellement un ordre séquentiel ou temporel quand on échantillonne des animaux, absent dans le cas de données déjà existantes sur les populations humaines.

Ces méthodes ont également été utilisées dans des champs d'application variés au sein des populations humaines : criminalité, usage de drogues, personnes sans abri, ou encore maladies rares (Chao 2015) (Bird, King 2018), ou encore contrôle de qualité du recensement aux Etats-Unis (voir infra).

Une application plus ancienne souvent citée, et relevant de l'estimation par système dual, est celle de Laplace en 1802. Laplace souhaitait estimer la taille de la population française. Il disposait pour cela d'un registre des naissances pour l'ensemble de la France. Laplace fit l'hypothèse qu'il existait, pour toute zone géographique, un rapport peu variable entre la taille de la population et le nombre de naissances consignées dans le registre. Il proposa d'effectuer un sondage pour évaluer ce rapport sur un échantillon de communes, et l'appliquer ensuite au registre national (ces travaux sont régulièrement évoqués pour souligner l'apport d'un sondage et celui de l'estimateur par le ratio, utilisé à cette occasion).

Une enquête fut réalisée dans trente départements au sein desquels plusieurs communes furent échantillonnées puis recensées. On releva dans ces communes, d'une part les identités des personnes figurant sur le registre national des naissances durant 3 années consécutives récentes (naissances enregistrées au cours des trois années), d'autre part les identités des personnes recensées, étant entendu que cette identification permettait de rapprocher les deux sous-populations en question. Les données dont on dispose à l'issue de l'opération sont les suivantes :

- le nombre total de naissances dans le registre national des naissances s'avère égal[1] à 1 million sur la période de 3 années considérées (Bru 1988);

- au sein de l'échantillon de communes : 2 037 615 personnes ont été recensées, dont 71 866 enfants figurant également sur le registre des naissances.

De ces données, on peut déduire qu'il y a 1 965 749 personnes recensées dans les communes, mais absentes du registre des naissances.

Tableau 1 : Effectifs de recensement et naissances publiés en 1803

		Enquête de recensement communale		
		Personnes recensées	Personnes non recensées	Ensemble
Registre national des naissances	Personnes figurant dans le registre national des naissances	**71 866**	928 134	**1 000 000**
	Personnes ne figurant pas dans le registre national des naissances	1 965 749	**???**	
	Ensemble	**2 037 615**		**???**

Source : (Bird, King 2018)

On dispose donc d'un tableau de contingence incomplet (effectif x_{ij} dans la cellule (i, j) ; effectifs marginaux x_{i+} et x_{+j}), dans lequel on connaît 3 effectifs sur 4.

1 Après un arrondi brutal qui contraste fortement avec la précision des autres dénombrements !

L'estimateur de la taille totale de population nationale N utilisé par Laplace est formé en multipliant le nombre total de naissances du registre x_{1+} par le rapport entre le nombre d'habitants x_{+1} et le nombre de naissances x_{11} dans l'enquête de recensement communale (un « multiplicateur de naissances »).

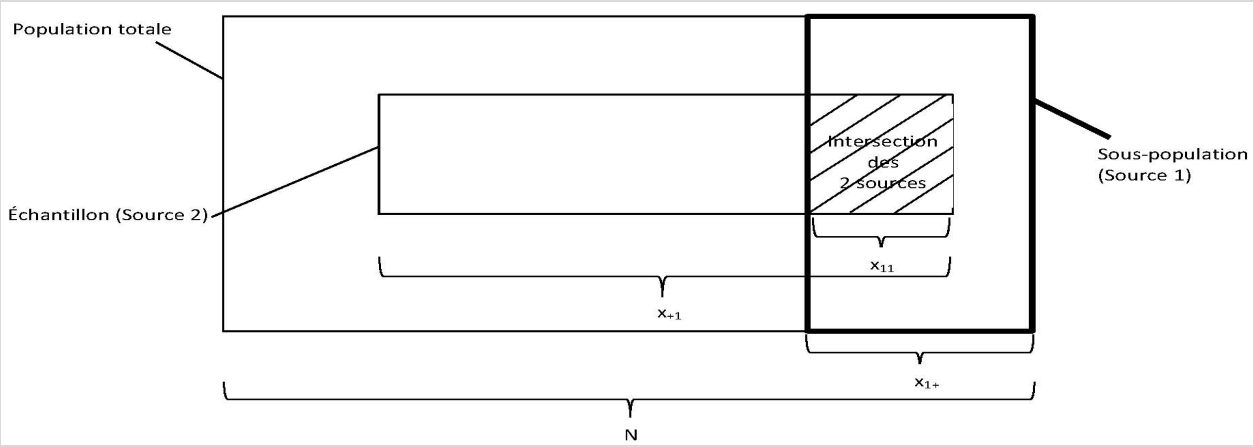
Soit

$$Taille\,de\,population\,totale \;=\; Nombre\,de\,naissances\,total \;\times\; \frac{Taille\,de\,population\,dans\,l'\,enquête}{Nombre\,de\,naissances\,dans\,l'\,enquête}$$

$$\hat{N}=\frac{x_{1+}\times x_{+1}}{x_{11}}=1\,000\,000\;\times\;\frac{2\,037\,615}{71866}=28\,352\,976.$$

L'estimation de Laplace s'est avérée bien meilleure que celle issue du recensement de 1801. L'estimateur dit 'par le ratio' obtenu dans ce contexte de deux sources de données différentes n'est autre que l'estimateur de Lincoln-Petersen, que nous allons détailler par la suite.

On peut au demeurant retrouver simplement cet estimateur lorsqu'on est familier de théorie des sondages. En effet, l'ensemble des personnes figurant dans le registre des naissances définit une sous-population de la population totale. Cette sous-population est en 'vraie' proportion $\frac{x_{1+}}{N}$. L'enquête, quand on la considère comme issue d'un sondage aléatoire simple de taille x_{+1} dans la population totale, fournit immédiatement un estimateur sans biais de cette vraie proportion si on forme la proportion que constitue la sous-population en question dans l'échantillon total, soit $\frac{x_{11}}{x_{+1}}$ Ainsi, 'en moyenne' $\frac{x_{11}}{x_{+1}}=\frac{x_{1+}}{N}$ et on en déduit l'estimateur de Laplace. Graphiquement :



L'application de Laplace est un cas « particulier » de l'estimation par système dual, puisque l'une des deux sources ne concerne qu'une partie de la population, ici les naissances au cours de trois années consécutives. Dans le cas le plus général, les mesures de capture-recapture s'effectuent en croisant des sources qui prétendent couvrir la population globale, et non sur un sous-ensemble.

Dans des contextes assez différents, on souligne ici que l'estimation par système dual, ou capture recapture, a pour objectif d'évaluer la taille d'une population en l'absence d'un dénombrement exhaustif, mais en mobilisant deux énumérations partielles (ou davantage) et en tirant parti de l'information disponible sur l'intersection de ces différentes énumérations.

Les méthodes de capture-recapture n'ont pas – ou très peu - été mises en œuvre par la Statistique publique en France jusqu'à présent. Elles ont par contre été appliquées pour l'estimation de populations rares ou pour effectuer des contrôles de qualité *a posteriori* du recensement dans d'autres pays. Deux exemples : aux Pays-Bas, une estimation nationale de personnes sans domicile a eu lieu pour les années 2009 à 2013 en exploitant trois sources de type registre; aux Etats-Unis, l'enquête CPS (*Current Population Survey*) a servi de seconde source au Census Bureau pour estimer le défaut de couverture du Recensement de 1980, et une enquête *ad hoc* de type aréolaire (*Post-Enumeration Survey*) a pris la suite pour estimer le défaut de couverture du Recensement de 1990.
Concernant la France, ce document de travail trouve sa justification essentielle dans les questions touchant à la qualité du futur répertoire Résil (REpertoire Statistique des Individus et des Logements), en particulier à l'appréciation de son éventuel défaut de couverture. Dans cette perspective, on pourra rapprocher Résil de l'enquête annuelle de Recensement (EAR).

Dans la suite de ce document, on s'attachera à formaliser cet estimateur en abordant les différentes approches qui peuvent le justifier. On précisera les hypothèses nécessaires, leur implication dans un contexte pratique, ainsi que les pistes de relâchement de ces hypothèses.

I. Estimation de la taille d'une population fondée sur un système d'enregistrements multiples : cadre général

Dans cette partie, on s'intéresse à la formalisation de l'estimation par système multiple, initiée par Lincoln au début du 20ème siècle, puis poursuivie notamment par Darroch (Darroch 1958).

Le terme 'liste' désigne un jeu de données qui a pour but d'énumérer des individus de la population cible, qu'il provienne d'un recensement ou d'une enquête par sondage[2]. Concrètement, la liste est un ensemble d'enregistrements.
Un enregistrement est supposé correspondre à une unité (un individu) de la population. En situation idéale, il y a concordance parfaite entre la liste et la population : les deux ensembles sont alors en bijection, tout individu de la population est représenté par un et un seul enregistrement de la liste et réciproquement.

2 Dans ce second contexte, un système de pondération supposé sans biais permet d'estimer la taille de population. Il y a certes une erreur d'échantillonnage, mais ce n'est pas à ce type d'inférence que s'intéresse la théorie de la capture-recapture.

En situation réelle, cette situation idéale n'existe pas. On parle alors de sous-couverture s'il existe des unités appartenant à la population pour lesquelles on ne dispose pas d'enregistrement dans la liste.
On parle de sur-couverture si certains enregistrements ne correspondent à aucun individu appartenant à la population ou s'ils sont en doublon avec au moins un autre enregistrement de la dite liste.

En matière de dénombrement, il est donc essentiel de bien distinguer les concepts en jeu : dans un fichier, une 'ligne', c'est-à-dire un enregistrement, ne correspond pas forcément à une unité de la population cible ayant une existence réelle. La sous-couverture se conçoit aisément et c'est plutôt à ce type de défaut que l'on pense en premier. La sur-couverture est un défaut plus insidieux. Ce peut être parce qu'un enregistrement est totalement fictif (par exemple suite à un problème de mise à jour, un répertoire contient des individus aujourd'hui disparus), ce peut être aussi un doublon : deux enregistrements correspondent bien à un même individu réel de la population- mais probablement ils se présentent sous des identifiants différents, ou avec des caractéristiques différentes suite à une erreur affectant certaines données individuelles – par exemple on a mal codé le sexe, ou l'adresse a changé. Les sur-couvertures sont souvent alimentées par des défauts de mise à jour de source, parce que les populations évoluent en continu et que les processus d'enregistrement et de traitement ne parviennent pas à en suivre le rythme d'évolution.

Il y a une vraie dissymétrie entre les problèmes posés respectivement par la sous-couverture et par la sur-couverture : la mesure et la correction de la sous-couverture est tout l'objet de la théorie ici abordée. En revanche, la sur-couverture est sensiblement plus redoutable à gérer, car compter des enregistrements n'est manifestement pas la même chose que compter des individus ! En pratique, on fait presque toujours l'hypothèse qu'il n'y a pas de sur-couverture. Néanmoins, une méthode – plus ou moins convaincante - est proposée en partie II pour traiter cette situation lorsqu'on soupçonne qu'elle se présente.

1. L'estimateur de Lincoln-Petersen : une approche basée sur des tables de contingence

Soit une population d'intérêt U, de taille inconnue N. On cherche à estimer N.

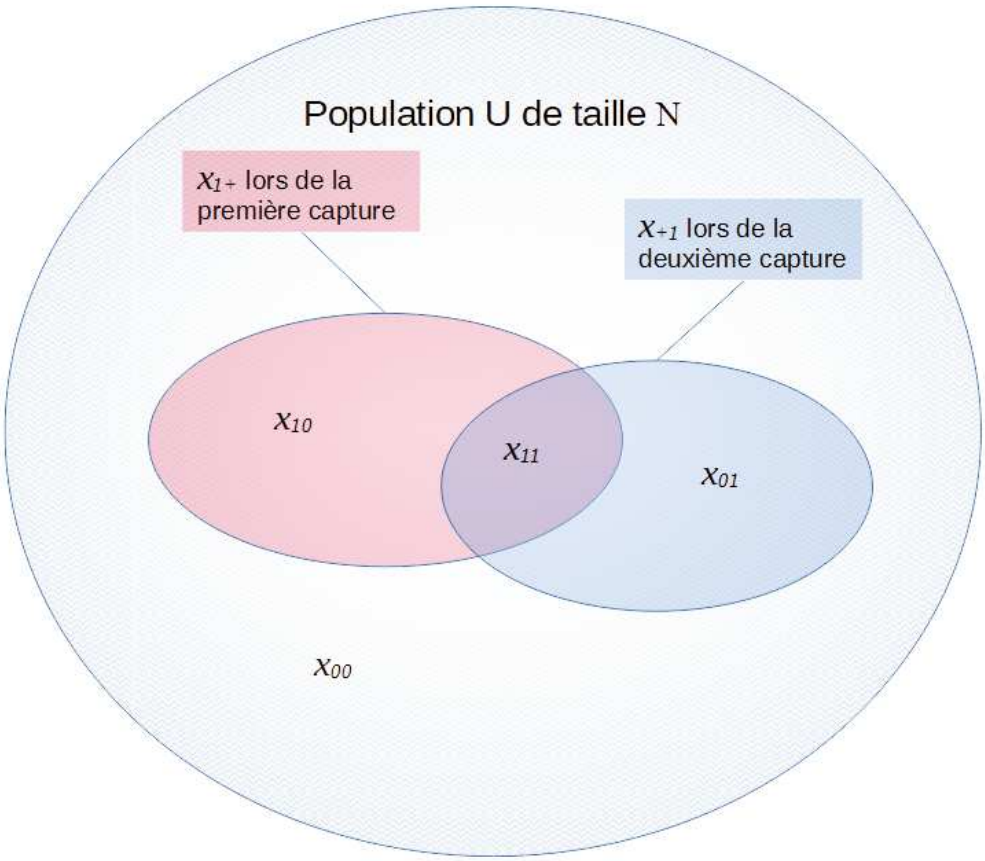
On dispose de deux énumérations incomplètes de U, que l'on peut appeler 'listes' (on considère donc qu'il n'y a pas de sur-couverture). Ces énumérations (que l'on peut faire correspondre à des 'captures') sont notées A et B. Dans le cas général, on peut avoir plus de deux énumérations. Les listes en question ne couvrent pas la population complète, parfois loin s'en faut, et il n'y a aucune sur-couverture. Néanmoins, il semble que presque toujours chaque individu de la population a une probabilité non nulle d'appartenir à la liste (on verra par la suite que c'est bien le cas, avec la première hypothèse). La littérature reste très discrète sur ce point, pourtant important : comme en matière de non-réponse, un individu est ou n'est pas présent dans une liste donnée, mais on considère toujours qu'il *pourrait* en faire partie. Une situation caricaturale – claire en termes probabilistes mais tout à fait particulière - est celle de l'enquête par sondage sans défaut de couverture, où le statisticien maîtrise les probabilités de sélection. Dans ce cas, la liste énumère l'échantillon sélectionné, qui en général est beaucoup plus petit que la population complète. Une situation moins évidente à percevoir est celle d'un recensement, qui a vocation à être exhaustif – mais qui en pratique ne l'est pas.

Après rapprochement des deux listes sur la base d'une identification commune, supposant que la jointure s'effectue sans aucune erreur[3], les statistiques intéressantes sont constituées par un ensemble de dénombrements x_{ij} où $i=0$ signifie la non-appartenance à A et $i=1$ signifie l'appartenance à A, et $j=0$ signifie la non-appartenance à B et $j=1$ signifie l'appartenance à B. Cela définit 4 effectifs, dont 3 seulement sont observées : en effet, par définition on ne connaît pas x_{00}.

3 Lorsqu'on dispose d'un identifiant commun, l'appariement est généralement très bon. Cependant, ce n'est pas toujours le cas.

On note $x_{1+} = x_{11} + x_{10}$ les individus capturés dans A, et $x_{+1} = x_{11} + x_{01}$ les individus capturés dans B. La taille totale de l'échantillon obtenu (c'est la réunion des deux listes, que l'on peut bien qualifier d'échantillon puisqu'il s'agit d'une sous-population de la population totale) est $n = x_{11} + x_{10} + x_{01}$. L'effectif non observé est $x_{00} = N - n$.

Schéma 1 : expérience de capture-recapture



Pour chaque individu K de la population ($1 \leq K \leq N$), on note $\delta_K(A)$ l'indicatrice de présence dans la sous-population A. C'est une variable aléatoire suivant une loi de Bernoulli. Sous une **première hypothèse H_1** (forte) que le paramètre de cette loi ne dépend pas de l'individu, ce qui traduit l'homogénéité des probabilités d'appartenance à la liste, cette loi n'est pas dégénérée et se note $B(1, p_A)$ où $E(\delta_K(A)) = p_A$. On a par définition

$$\sum_{K=1}^{N} \delta_K(A) = x_{1+} .$$

Sous la **seconde hypothèse H_2** que les $\delta_K(A)$ sont mutuellement indépendants, x_{1+} est une variable aléatoire suivant une loi binomiale[4], plus précisément la loi $B(N, p_A)$. On en déduit

$$E(x_{1+}) = N \cdot p_A.$$

De la même façon, on définit une probabilité d'appartenance à la liste B notée p_B et on parvient à

$$E(x_{+1}) = N \cdot p_B.$$

Enfin, on peut s'intéresser à la probabilité d'appartenir *à la fois* aux listes A et B : en introduisant l'indicatrice associée $\delta_K(AB)$, sous les mêmes hypothèses d'homogénéité et d'indépendance mutuelle entre individus, on obtient

$$E(x_{11}) = N \cdot p_{AB}$$

où $E(\delta_K(AB)) = p_{AB}$ désigne la probabilité qu'un individu quelconque de la population appartienne *à la fois* aux listes A et B.

Sous une **troisième hypothèse H_3** (forte), l'appartenance à l'une des deux listes n'apporte aucune information sur l'appartenance à l'autre liste. On parlera d'indépendance entre sources. Techniquement, pour tout individu K, $\delta_K(A)$ et $\delta_K(B)$ sont indépendantes. Cela est équivalent à

$$p_{AB} = p_A \cdot p_B.$$

En effet, puisque $\delta_K(AB) = \delta_K(A).\delta_K(B)$, il vient du fait de l'indépendance

$$p_{AB} = E\,\delta_K(AB) = E\,\delta_K(A).E\,\delta_K(B) = p_A \cdot p_B.$$

On a donc un tableau de contingence incomplet qui résume l'information disponible.

4 Une somme de N variables aléatoires indépendantes suivant la même loi de Bernoulli $B(1, p)$ suit une loi binomiale $B(N, p)$.

Tableau 2 : Tableau de contingence incomplet

		B		
		Présents	Absents	Ensemble
A	Présents	x_{11}	x_{10}	x_{1+}
	Absents	x_{01}	x_{00} (= ???)	
	Ensemble	x_{+1}		N (= ???)

On a l'égalité

$$\frac{E(x_{11})}{N} = p_{AB} = p_A . p_B = \frac{E(x_{1+})}{N} . \frac{E(x_{+1})}{N}, \text{ soit}$$

$$N = \frac{E(x_{1+}) . E(x_{+1})}{E(x_{11})}.$$

Dans l'esprit de l'estimateur des moments, on peut naturellement estimer N par

$$\hat{N} = \left[\frac{x_{1+} . x_{+1}}{x_{11}} \right] \quad (1)$$

où les crochets désignent la partie entière. Cet estimateur – célèbre – est dit « estimateur de Lincoln-Petersen ».

Il est essentiel de noter que dans cette présentation générale, l'aléa considéré n'est (en général) pas un aléa de sondage tel qu'on le rencontre dans les enquêtes classiques : il n'y a pas à proprement parler d'opération préalable de tirage d'échantillon (en tout cas pas nécessairement), et on n'a pas besoin de passer par une étape de sélection maîtrisée des sous-populations A et B en jeu. Il s'agit en réalité d'un aléa d'une nature plus générale : l'individu est ou n'est pas dans la liste, cela avec une certaine probabilité que l'on ne cherche pas à interpréter davantage. On se contente de considérer qu'elle existe et qu'elle concentre toute l'incertitude affectant le statut final de chaque individu, quelles que soient la nature et l'ampleur de cette incertitude.

L'exemple historique le plus célèbre est celui de l'estimation de la taille d'une population de poissons dans un lac : dans ce cas, A est un échantillon de poissons péchés à une date initiale dans le lac, puis marqués et remis à l'eau. La liste B est constituée par un nouvel échantillon de poissons, totalement indépendant du premier, péchés quelques jours plus tard (entre-temps, les poissons marqués la première fois se sont bien mélangés avec les autres !). Comme on repère les poissons marqués, on est en mesure d'obtenir x_{11}, x_{10}, x_{01}. On distingue bien les opérations d'échantillonnage, conçues de manière *ad hoc*.

Mais les listes A et B peuvent aussi être des fichiers quelconques, formés d'une manière que l'on ne maîtrise pas, dont l'existence est éventuellement ancienne et dans lesquels figurent – ou non – les individus de la population totale. Par exemple A peut être constituée par les logements d'une commune issus du recensement (exhaustif) de la population, et B par les logements de la même commune soumis à la taxe foncière. En la circonstance, ces deux listes sont incomplètes, même si elles se recoupent très largement : il y a - en toute généralité - des logements qui échappent à la fois à la taxe foncière et au recensement (effectif inconnu et inobservé x_{00}), et ce sont justement ceux-là que la méthode prétend dénombrer.

On trouve aussi la situation intermédiaire, où la liste A est un fichier qui prétend couvrir très largement la population complète (dans l'idéal, de façon exhaustive) et où la liste B est une liste beaucoup plus petite issue d'un échantillonnage au sens habituel. Typiquement, ce sont les situations où A est un recensement et B un échantillon d'enquête de contrôle du dit recensement (*post enumeration survey*). Ce cas de figure se rencontre en pratique, l'enquête de contrôle s'appuyant souvent sur un échantillon de grappes définies géographiquement.

Si une des deux sources est exhaustive, la seconde source est alors incluse dans cette source exhaustive, et on vérifie immédiatement qu'on aboutit toujours à $\hat{N} = N$.

Cette approche-cadre semble parfaitement acceptable, mais il faut garder en mémoire que trois hypothèses sous-tendent le processus d'estimation :

> H1 - chaque individu de la population a la même probabilité d'appartenance à une liste donnée – A ou B (*homogénéité* des probabilités de capture) ;
>
> H2 - l'affectation d'un individu à une liste donnée est indépendante de l'affectation des autres individus (*indépendance* des comportements entre individus).
>
> H3 - pour tout individu, l'appartenance à une des deux listes est indépendante de l'appartenance à l'autre (*indépendance* des occasions de capture) ;

Dans la suite de ce document, on appellera « hypothèses socles » l'ensemble de ces 3 hypothèses, portant respectivement sur l'homogénéité des probabilités individuelles, l'indépendance entre individus, et l'indépendance entre sources.

Les hypothèses socles sont (excessivement) fortes. L'hypothèse H2 d'indépendance des comportements entre individus semble pour sa part systématiquement adoptée dans la littérature : on considère qu'un individu donné n'en influence jamais un autre. Cette hypothèse est *a priori* acceptable dans bon nombre de situations pratiques, et semble moins contraignante que les deux autres hypothèses. Il y a néanmoins une situation où elle est difficile à accepter, c'est celle d'une enquête par sondage probabiliste où l'échantillonnage utilise, à un moment ou à un autre, un tirage de grappes. Dans ce cas, la sélection d'un individu d'une grappe donné apporte – évidemment - une information déterminante sur le sort des autres individus de la même grappe. Ce peut être aussi la conséquence d'un effet qui s'apparente à un effet de grappe sans qu'il y ait à proprement parler d'échantillonnage de grappe : par exemple dans une famille, si on

sait qu'un membre de la famille est recensé, cela augmente la probabilité des autres membres de la famille de l'être également.

Par ailleurs, on fait également deux hypothèses supplémentaires, qui ont trait cette fois non pas à l'estimation statistique mais à l'exactitude supposée des quantités x_{11}, x_{10}, x_{01}. Ainsi, on s'appuie sur les hypothèses suivantes, dites hypothèses d'exactitude :

> H4 - il n'y a pas de sur-couverture (présence de doublons, présence d'enregistrements erronés, présence d'éléments n'appartenant pas à la population U) : cela revient à dire que les effectifs x_{11}, x_{10}, x_{01} dénombrent bien des individus uniques appartenant à la population U ;
>
> H5 - on connaît de façon exacte l'appartenance des individus à une seule des deux listes, ou à l'intersection des deux listes. Dans un contexte d'expériences de capture et de marquage, cela signifie qu'il n'y a pas de perte de marquage. Lorsqu'il s'agit de données déjà collectées, cette hypothèse se traduit par un appariement parfait entre les différentes sources de données.

De plus, les articles citent souvent une hypothèse H6 de population fermée. Il s'agit de se placer dans un contexte où la population d'intérêt ne varie pas durant les périodes séparant les différentes occasions de capture. Cette hypothèse est vérifiée lorsqu'il y a homogénéité des probabilités de capture (et en faisant l'hypothèse que la population définie par l'union des individus existant à chaque occasion de capture est bien notre population d'intérêt), car alors il n'existe pas d'individus pour lesquels la probabilité d'échantillonnage vaudrait 0 pour l'une ou l'autre des occasions de capture. Cette hypothèse est importante dans le contexte d'opérations de capture d'animaux éloignées dans le temps, ou lorsque des listes d'individus sont constituées à des périodes sensiblement différentes (à moins que l'évolution démographique ne soit jugée particulièrement faible).

L'ensemble de ces hypothèses est extrêmement exigeant, et Cormack fait même mention d'un « défaut de confiance universel » dans la validité de ces hypothèses (Goudie et Goudie 2007).

A retenir :

Les méthodes d'estimation par système multiple (ou de capture-recapture) permettent d'estimer un total de population, en présence de plusieurs listes incomplètes dénombrant les individus.

Ces méthodes reposent sur plusieurs hypothèses fortes ;
- homogénéité des probabilités de capture ;
- indépendance des comportements entre individus ;
- indépendance des occasions de capture ;
- absence de sur-couverture ;
- appariement parfait entre les enregistrements des deux listes ;
- population fermée.

2. Une approche par maximum de vraisemblance

On peut formaliser l'estimateur de Lincoln-Petersen selon la variante suivante. On va considérer que tout individu K de la population a une probabilité $p_{K,ij}$ d'être en statut i pour la liste A et en statut j pour la liste B, où $i \in \{0,1\}$ et $j \in \{0,1\}$. La probabilité qu'a K d'être en statut i pour la liste A est notée $p_{K,i+}$ et la probabilité qu'a K d'être en statut j pour la liste B est notée $p_{K,+j}$.

L'hypothèse d'homogénéité des probabilités de capture conduit à considérer que $p_{K,ij}$ ne dépend pas de K , soit $p_{K,ij} = p_{ij}$. L'hypothèse d'indépendance mutuelle entre A et B pour chaque individu se traduit par

$$p_{ij} = p_{i+} \times p_{+j} \quad \forall (i,j)$$

En résumé, à chaque liste est associée une probabilité de capture qui lui est propre, mais elle est la même pour tous les individus ET les captures sont indépendantes d'une liste à l'autre. Les comportements restent également indépendants d'un individu à l'autre. Ce modèle appartient à une classe de modèles appelée M_t dans la littérature spécialisée (Wolter 1986) : dans un schéma qui relève plutôt de la capture animalière, cette notation signifie que les comportements des individus peuvent évoluer dans le temps (d'où l'indice t) mais pas dans l'espace (ce qui se traduit à tout instant par un comportement identique de tous les individus de la population).

Les variables aléatoires en jeu sont les effectifs par case, relevés à partir de l'intégralité des occasions de capture considérées, soit $x_{11}, x_{10}, x_{01}, x_{00}$ sur l'ensemble de la période.
Compte tenu des hypothèses faites, ce vecteur aléatoire suit une loi multinomiale $M(N ; p_{11}, p_{10}, p_{01}, p_{00})$, dont la densité s'écrit

$$L = \frac{N!}{x_{11}!\, x_{10}!\, x_{01}!\, x_{00}!} \cdot p_{11}^{x_{11}} \cdot p_{10}^{x_{10}} \cdot p_{01}^{x_{01}} \cdot p_{00}^{x_{00}} .$$

Du fait de l'indépendance postulée, il suffit de deux paramètres de probabilité pour écrire cette densité. On peut choisir $p_{1,+}$ et $p_{+,1}$. Après quelques calculs :

$$L = \frac{N!}{x_{11}!\, x_{10}!\, x_{01}!\, (N - x_{11} - x_{10} - x_{01})!} \cdot p_{1+}^{x_{1+}} \cdot p_{+1}^{x_{+1}} \cdot (1 - p_{1+})^{N - x_{1+}} \cdot (1 - p_{+1})^{N - x_{+1}}$$

La taille de population à estimer N se présente comme un paramètre de la densité – au même titre que les deux probabilités. Une méthode d'estimation efficace est celle du maximum de vraisemblance. La maximisation de la densité L conduit à[5]

$\hat{N}=\left[\frac{x_{1+}\,.\,x_{+1}}{x_{11}}\right]$ (les crochets désignent la partie entière).

$\hat{p}_{1+}=\frac{x_{11}}{x_{+1}}$ = proportion d'unités de la liste B qui sont dans la liste A.

$\hat{p}_{+1}=\frac{x_{11}}{x_{1+}}$ = proportion d'unités de la liste A qui sont dans la liste B.

$\hat{N}$ est bien l'estimateur de Lincoln-Petersen. C'est aussi $\hat{N}=x_{11}+x_{10}+x_{01}+\left[\frac{x_{10}\,.\,x_{01}}{x_{11}}\right]$, ce qui revient à estimer l'effectif inconnu x_{00} par $\hat{x}_{00}=\left[\frac{x_{10}\,.\,x_{01}}{x_{11}}\right]$.

Dans certaines circonstances, on peut encore simplifier le modèle en considérant *a priori* qu'on a en sus $p_{1+}=p_{+1}$, probabilité commune notée p - ce qui correspond à un modèle de capture équiprobable (ou 'uniforme'). La vraisemblance se simplifie puisqu'il n'y a plus qu'un seul paramètre de probabilité et sa maximisation conduit à l'estimateur $\hat{N}=\frac{\left(2.x_{11}+x_{10}+x_{01}\right)^2}{4.x_{11}}$ (ce qui n'est pas vraiment intuitif…). Au passage, on estime p par $\hat{p}=\frac{2.x_{11}}{2.x_{11}+x_{10}+x_{01}}$.

Cette méthodologie reste simple – la suite proposera des modèles plus complexes. Mais quelle que soit l'approche, il n'y a pas de miracle, il faut d'une façon ou d'une autre postuler un mécanisme (ici c'est le mécanisme d'indépendance qui est utilisé) qui 'construit' le paramètre clé p_{00} de la même façon qu'il 'construit' tous les autres paramètres p_{ij}. L'objectif profond du mécanisme est de réduire la dimension du problème, donc le nombre de grandeurs 'libres', c'est-à-dire de grandeurs à estimer. Sous réserve qu'elles soient en nombre suffisant, les données que constituent les comptages connus permettent d'estimer les déterminants de ce mécanisme, donc toutes ces probabilités – en particulier p_{00}.

5 L'estimation $\hat{N}$ n'est pas formellement rigoureuse : après dérivation de la Log vraisemblance en N, il faut approximer $\sum_{i=1}^{n}\frac{1}{i}$ par $\log n+\mu$ où μ désigne la constante d'Euler (environ 0,577). Le développement suppose donc implicitement que n (qui prendra successivement les valeurs N et x_{00}) est suffisamment grand, disons quelques dizaines au moins (l'erreur relative d'approximation est de 1,7 % avec n=10 et de 0,2 % avec n=50). Dans notre contexte, on peut considérer que c'est toujours le cas.

Ce qu'il faut retenir :

L'approche par maximum de vraisemblance :

Sous condition du respect des hypothèses mentionnées plus haut, le vecteur des effectifs $x_{11}, x_{10}, x_{01}, x_{00}$ suit une loi multinomiale $M(N; p_{11}, p_{10}, p_{01}, p_{00})$.

En maximisant la vraisemblance associée, on estime x_{00} par $\hat{x}_{00} = [\frac{x_{10} \cdot x_{01}}{x_{11}}]$ et par conséquent N par $\hat{N} = [\frac{x_{1+} \cdot x_{+1}}{x_{11}}]$ .Quelle que soit l'approche, il est nécessaire de formaliser des hypothèses sur le mécanisme statistique, afin notamment de réduire la dimension du problème et de pouvoir formaliser une estimation.

3. La vision '*odds ratio*'

Le lecteur familier de l'utilisation des *odds ratios* dans les tables de contingence trouvera naturelle l'expression de l'estimateur de Lincoln-Petersen (voir encadré sur les *odds ratios*). En effet, dans un contexte d'homogénéité des probabilités individuelles, l'indépendance qui conditionne l'utilisation de cet estimateur se traduit par un *odds ratio* égal à 1, soit

$$\text{Odds ratio} = \frac{p_{11} \cdot p_{00}}{p_{10} \cdot p_{01}} = 1 .$$

Le vecteur des effectifs par case suivant une loi multinomiale, on estime sans biais p_{ij} par x_{ij}/N , ce qui conduit à poser $\frac{x_{11} \cdot x_{00}}{x_{10} \cdot x_{01}} = 1$. On en déduit immédiatement $\hat{x}_{00} = [\frac{x_{10} \cdot x_{01}}{x_{11}}]$, puis $\hat{N} = x_{11} + x_{01} + x_{10} + \hat{x}_{00}$.

Encadré : Les odds ratio

On considère une table de contingence croisant deux variables qualitatives A et B. L'effectif en ligne i et colonne j est noté x_{ij} .

Le *odds* (une 'cote' en français) est un rapport de deux probabilités. On fixe déjà une modalité d'une des deux variables qualitatives, au choix - par exemple la modalité 1 de la variable A - et on considère uniquement les individus de cette sous-population. Par construction, on se place donc conditionnellement à l'appartenance à la sous-population de référence choisie. Le *odds* (1,A) est alors le rapport entre la probabilité de vérifier une des deux modalités (au choix) de la seconde variable et la probabilité de vérifier l'autre modalité de cette même seconde variable. L'ordre des modalités est donc encore au choix, par exemple on retiendra la modalité 1 de la variable B.
Avec ces choix, le *odds* (1,A) s'énonce ainsi : rapport de la probabilité de vérifier la modalité 1 de B et de la probabilité de vérifier la modalité 0 de B conditionnellement au fait de vérifier la modalité 1 de A. Soit encore, avec des notations naturelles :

$$\textit{odds}\ (1,A) = \frac{p_{1\,B|1\,A}}{p_{0\,B|1\,A}} = \frac{p_{1\,B\ et\ 1\,A}}{p_{0\,B\ et\ 1\,A}} = \frac{p_{11}}{p_{10}}$$

De manière tout à fait parallèle, on définit le *odds* (0,A) où on prend cette fois comme référence l'autre modalité de la variable A (et on ne change surtout pas l'ordre des modalités de B).

$$\textit{odds}\ (0,A) = \frac{p_{1\,B|0\,A}}{p_{0\,B|0\,A}} = \frac{p_{1\,B\ et\ 0\,A}}{p_{0\,B\ et\ 0\,A}} = \frac{p_{01}}{p_{00}}$$

Le ratio de ces deux *odds* – dit '*odds ratio'* - s'écrit :

$$\textit{odds ratio} = \frac{odds(1,A)}{odds(0,A)} = \frac{p_{11}\cdot p_{00}}{p_{10}\cdot p_{01}}$$

L'indépendance entre A et B se traduit par le fait qu'on peut supprimer tous les conditionnements sans rien changer aux probabilités : toute probabilité de B conditionnelle à A est égale à la probabilité de B. Sur les *odds ratio*, comme il n'y a que 2 modalités pour chaque variable qualitative, cela se traduit de manière équivalente par odds (1,A) = odds (0,A), soit *odds ratio* = 1 :

$$\text{Indépendance} \Leftrightarrow \frac{p_{11}\cdot p_{00}}{p_{10}\cdot p_{01}} = 1$$

4. Appréciation de la qualité

On se place toujours dans le cas où les hypothèses socles (citées précédemment) sont vérifiées. L'aléa en jeu est (exclusivement) celui qui conduit à la présence ou à l'absence de chaque individu dans chacune des listes. Cet aléa est *a priori* source de biais et de variance.

S'agissant du biais, on peut considérer qu'il est asymptotiquement (N très grand) négligeable, donc considéré comme nul en pratique dès lors que les listes en jeu A et B sont de (très) grande taille. Cela résulte en particulier du comportement asymptotique du maximum de vraisemblance. Plus précisément, on montre au premier ordre (Wolter 1986)

$$\frac{E(\hat{N})-N}{N} \approx \frac{1}{N} . \frac{p_{0+} \cdot p_{+0}}{p_{1+} \cdot p_{+1}} .$$

On constate qu'en toute rigueur le biais est positif, c'est-à-dire que $\hat{N}$ surestime N . Cela étant, dans tout contexte où les sources en jeu sont des sources administratives prétendant à une sous-couverture modérée, les probabilités p_{1+} et p_{+1} seront suffisamment éloignées de zéro pour que le ratio $\frac{p_{0+} \cdot p_{+0}}{p_{1+} \cdot p_{+1}}$ reste numériquement (très) faible devant N . Dans ces conditions, certes le biais n'est pas rigoureusement nul, mais numériquement il devrait être la plupart du temps (très) petit[6].

Pour ce qui est de la variance, on peut montrer qu'elle est approximée au premier ordre par

$$V(\hat{N}) \approx N . \frac{p_{0+} \cdot p_{+0}}{p_{1+} \cdot p_{+1}}$$

et qu'elle peut être estimée par

$$\hat{V}(\hat{N}) = x_{1+} \cdot x_{+1} . \frac{x_{10} \cdot x_{01}}{x_{11}^3} .$$

L'écart-type de $\hat{N}$ est fonction de $\sqrt{N}$, le coefficient de variation de $\hat{N}$ varie donc en $\frac{1}{\sqrt{N}}$.

On peut même considérer que $\hat{N}$ suit une loi de Gauss, et sauf situation très exceptionnelle on peut négliger le (carré du) biais devant la variance dans l'expression de l'erreur quadratique moyenne puisque

$$E(\hat{N}-N)^2 \approx \left(\frac{p_{0+} \cdot p_{+0}}{p_{1+} \cdot p_{+1}} \right)^2 + N . \frac{p_{0+} \cdot p_{+0}}{p_{1+} \cdot p_{+1}} \approx N . \frac{p_{0+} \cdot p_{+0}}{p_{1+} \cdot p_{+1}} .$$

6 Par exemple si chaque source couvre la moitié de la population, $p_{0+} = p_{+0} = p_{1+} = p_{+1} = 0,5$, le ratio vaut 1 et il est bien évidemment totalement négligeable devant N dès que N . va par exemple dépasser 100. En pratique, il n'y a pas vraiment de sujet...

En négligeant le biais, on peut alors former les intervalles de confiance associés, ce qui permet d'apprécier la qualité de l'estimation (on pourra toujours augmenter la limite inférieure de l'intervalle estimé jusqu'à la valeur $x_1+x_{10}+x_{01}$ si la borne théorique est inférieure à ce seuil). On rappelle que tout ceci est conditionné par le respect des hypothèses socles.

Il y a une évidence qu'il est utile de rappeler : de manière générale, $\hat{N}$ est (bien entendu) biaisé lorsque le modèle est faux, en particulier lorsque les hypothèses socle ne sont plus vérifiées, mais quelle que soit la nature de l'erreur et son origine, l'ampleur numérique du biais (dont on ne contrôle plus le signe) est nécessairement très liée à l'importance numérique de x_{00} : quand x_{00} est petit (parce que $p_{1,+}$ ou/et $p_{+,1}$ est proche de 1), $\hat{N}$ se rapproche de N et le biais sera faible en valeur absolue (et encore davantage en valeur relative). C'est le cas chaque fois qu'au moins une des sources est une source administrative pseudo-exhaustive, et c'est encore plus vrai lorsqu'il en est ainsi des deux sources en présence.

Ce qu'il faut retenir :

À distance finie et en toute rigueur, l'estimateur de Lincoln-Petersen est biaisé et surestime la taille de population. Le biais relatif est néanmoins asymptotiquement nul, et on vérifie qu'en pratique les sources et les tailles de population en jeu sont telles que ce biais est numériquement négligeable.

Le coefficient de variation de $\hat{N}$ varie en $1/\sqrt{N}$.

On peut considérer $\hat{N}$ comme Gaussien et on dispose d'un estimateur de sa variance.

5. Le cas particulier des enquêtes de contrôle de type aréolaire

Dans cette partie, sont concernées les situations concrètes où on n'observe pas les effectifs de la table de contingence croisant les deux listes, mais seulement une fraction de ces effectifs. En pratique, on va à peu près systématiquement rencontrer ce cas lorsqu'on évalue la qualité d'un recensement (Fienberg 1992) : A est la liste des individus recensés et B est la population enquêtée par un échantillon de contrôle, par exemple de type aréolaire (tirage de grappes de logements). Le contrôle porte sur un échantillon de logements pour des raisons de coût, mais la liste B est définie comme celle que l'on obtiendrait si le contrôle portait sur l'intégralité des logements du territoire recensé. En contrepartie, il faudra introduire des poids de sondage pour se raccrocher à la théorie générale de l'estimation duale à partir de données échantillonnées, l'échantillon étant heureusement tiré de manière contrôlée.

Dans un tel contexte, on pourrait imaginer modifier la définition de la liste B et se contenter de la limiter à l'échantillon de grappes. Mais dans ce nouveau scénario, il y aurait par construction des effets de grappe dans la formation de B, si bien qu'on ne pourrait plus raisonnablement considérer comme indépendantes les appartenances – ou non - à B des individus de la population complète : dit autrement, inclure l'aléa de sondage des grappes dans l'aléa de Bernoulli n'est pas conceptuellement acceptable, car l'hypothèse H2 des hypothèses socle (indépendance des comportements entre individus) ne serait plus vérifiée. Sur le plan pratique, l'appariement préalable au calcul de l'estimateur s'effectue pour sa part sur les seules grappes échantillonnées.

On va donc traiter les grappes formant un échantillon dans B comme des unités d'échantillonnage permettant de produire une inférence au moyen d'un système de pondérations : donc une approche classique pour un statisticien d'enquête. Il n'y a pas de difficulté technique particulière mais plus de lourdeur formelle parce qu'il faut prendre en considération deux aléas de natures différentes : l'aléa d'appartenance aux listes, et l'aléa de sondage ayant produit l'échantillon à enquêter au sein de B. Ces deux aléas sont bien distincts. En effet considérons un individu donné quelconque de la population complète, qui se trouve donc dans une grappe donnée avant échantillonnage (par exemple les grappes sont les îlots découpés au sein des communes). L'aléa de sondage fait que la grappe (et donc l'individu) est sélectionnée ou pas dans l'échantillon de contrôle. L'aléa d'appartenance à B est le processus qui fait que lors d'une opération de contrôle, après l'opération de terrain, un individu de la population qui devrait être soumis au contrôle est ou n'est pas dans le fichier des individus ayant participé à l'enquête de contrôle[7] (pour de multiples raisons, et d'ailleurs peu importe ces raisons). En situation extrême, si le contrôle prétend porter sur l'intégralité de la population (en somme si on refait tout le recensement…), il n'y aura pas d'aléa de sondage mais il y aura toujours un aléa d'appartenance à la liste B.

Dans toute grappe échantillonnée g , après un travail d'appariement que l'on suppose toujours parfait, on peut observer les effectifs associés à cette grappe $x_{g,11}$ et $x_{g,01}$ - et donc $x_{g,+1}$. La règle conventionnelle ici adoptée consiste à comptabiliser l'individu dans la grappe où il a effectivement été recensé, quitte à ce qu'il y ait *après* le recensement une erreur de localisation. En pratique, on peut considérer qu'on n'observe pas (comme il faudrait) $x_{g,10}$ ni $x_{g,1+}$, pour une raison liée à la présence d'erreurs de localisation des logements au recensement. En effet, si l'échantillon de contrôle gère (*a priori*) correctement la localisation (il en dépend fondamentalement), on n'a plus cette garantie au niveau du recensement (si la grappe est l'îlot par exemple, le recensement peut affecter un code d'îlot erroné à un individu recensé). Au niveau de la grappe, on compte donc correctement les individus de l'échantillon de contrôle, mais pas forcement ceux du recensement. Pour cette raison d'ailleurs, le calcul des $x_{g,11}$ et $x_{g,01}$ n'est lui-même pas immédiat car il faut rapprocher la partie contrôlée dans la grappe de l'ensemble du recensement[8]. Bien évidemment on n'observe pas non plus $x_{g,00}$.

7 De même que l'on ne parvient pas à recenser tous les individus exhaustivement sur un territoire, on ne parvient pas non plus à contrôler tous les individus du territoire en question : dans les deux cas les raisons sont les mêmes puisque l'opération de contrôle n'est jamais qu'une seconde opération de recensement, soumise sur le fond aux mêmes faiblesses...

8 Noter qu'on pourrait imaginer ne pas corriger les erreurs géographiques et donc considérer la liste A avec ses erreurs (dit autrement, on adopterait une règle de comptabilisation là où l'individu apparaît dans le fichier du recensement – sans corriger l'erreur éventuelle de localisation). On risquerait alors le double compte dans N. Soit un individu vivant dans l'îlot I qui est effectivement recensé et contrôlé, mais qui est affecté par le recensement dans l'îlot J, la grappe g étant ici l'îlot I. Supposons que l'îlot J soit lui aussi contrôlé (pour simplifier). Si d'une façon ou d'une autre on ne détecte pas l'erreur sur l'îlot, l'individu comptera pour 1 dans $x_{I,01}$, et aussi pour 1 dans $x_{J,10}$. Après inférence, il va compter pour M/m dans x_{01} et aussi dans x_{10} , alors que ces effectifs respectifs ne souffrent, par construction, aucune intersection !

On connaît x_{1+} puisqu'il s'agit de l'effectif global recensé ; ce n'est pas incompatible avec la non-connaissance des $x_{g,1+}$ – tout simplement parce qu'on n'a pas besoin d'une localisation correcte 'à la grappe' pour connaître l'effectif global.
On ne connaît ni x_{11} ni x_{01} - ni par conséquent x_{+1} puisqu'on n'observe rien dans les grappes non échantillonnées.

Si on tire par sondage aléatoire simple m grappes parmi M, on estimera naturellement l'effectif inconnu x_{11} par $\hat{x}_{11}=\frac{M}{m}.\sum_{g=1}^{m} x_{g,11}$ et l'effectif inconnu x_{01} par $\hat{x}_{01}=\frac{M}{m}.\sum_{g=1}^{m} x_{g,01}$. Puis on forme $\hat{x}_{+1}=\hat{x}_{11}+\hat{x}_{01}$. Finalement, la taille de population est estimée par

$$\hat{\hat{N}}=\left[\frac{x_{1,+} \times \hat{x}_{+,1}}{\hat{x}_{1,1}}\right].$$

On sait exprimer les espérances, variances et variances estimées de $\hat{\hat{N}}=\left[\frac{x_{1+}\ \hat{x}_{+1}}{\hat{x}_{11}}\right]$ par rapport, soit à l'aléa de Bernoulli, soit à l'aléa de sondage, soit aux deux aléas combinés (Wolter 1986). Avec le sondage aléatoire simple considéré ici (par exemple), on vérifie qu'en prenant en compte les deux aléas, on a approximativement

$$\frac{E(\hat{\hat{N}})-N}{N} \approx \frac{1}{N}.\left\{\frac{p_{0+}\cdot p_{+0}}{p_{1+}\cdot p_{+1}} + \left(\frac{1-f}{f}\right).\frac{p_{0+}}{p_{1+}\cdot p_{+1}}\right\}$$

où f désigne le taux de sondage m/M, et

$$V(\hat{\hat{N}}) = N.\frac{p_{0+}\cdot p_{+0}}{p_{1+}\cdot p_{+1}} + N.\left(\frac{1-f}{f}\right).\frac{p_{0+}}{p_{1+}\cdot p_{+1}}.$$

Il est normal que le biais et la variance de $\hat{\hat{N}}$ soient systématiquement plus grands que ceux de $\hat{N}$, et qu'on retrouve les valeurs relatives à $\hat{N}$ lorsque $f=1$.

L'estimation de cette variance (toujours avec les deux aléas combinés) offre au moins 3 alternatives (Wolter, 1986) :

- soit $\hat{V}_1(\hat{\hat{N}})$ =

$$x_{1+}\cdot\hat{x}_{+1}.(\hat{x}_{+1}-\hat{x}_{11}).\frac{(x_{1+}-\hat{x}_{11})}{\hat{x}^3_{11}} + \left(\frac{1-f}{f}\right).\frac{(x^2_{1+}\cdot\hat{x}_{+1}.(\hat{x}_{+1}-\hat{x}_{11}))}{\hat{x}^3_{11}}$$

- soit $\hat{V}_2(\hat{\hat{N}})$ =

$$x_{1+}\,.\,\hat{x}_{+1}.(\hat{x}_{+1}-\hat{x}_{11}).\frac{(x_{1+}-\hat{x}_{11})}{\hat{x}_{11}^3} \;+\; M.\left(\frac{1-f}{f}\right).\frac{1}{m-1}.\sum_{g=1}^{m} d_g^2$$

où $d_g=\frac{x_{1+}}{\hat{x}_{11}}.\,x_{g+1}-\frac{x_{1+}\,.\,\hat{x}_{+1}}{\hat{x}^{2}_{11}}.\,x_{g11}$.

- soit l'utilisation des méthodes de réplication habituelles – *jackknife* ou *bootstrap* des grappes.

Le choix entre les deux estimateurs analytiques n'est pas du tout évident à faire. L'estimateur $\hat{V}_1(\hat{\hat{N}})$ a l'avantage d'être convergent. L'estimateur $\hat{V}_2(\hat{\hat{N}})$ serait pour sa part plus robuste au non-respect des hypothèses socles. Enfin, la différence entre $\hat{V}_1(\hat{\hat{N}})$ et $\hat{V}_2(\hat{\hat{N}})$ est d'ordre $\frac{M}{\sqrt{(m)}}$, ce qui n'est pas particulièrement sympathique parce que ce terme devient grand quand le nombre de grappes - définies aussi bien que tirées - augmente[9].

On peut étendre cette méthodologie à un échantillonnage de la liste B selon un plan complexe quelconque. Les principes ne changeront pas, mais les expressions de biais et de variance sont à reprendre entièrement.

Pour terminer cette partie, il faut signaler que l'on peut facilement rencontrer en pratique des distributions de $\hat{\hat{N}}$ qui s'éloignent significativement d'une loi de Gauss. Dans ce cas, même en considérant les biais asymptotiques comme négligeables devant les variances, l'utilisation d'intervalles de confiance produira des erreurs potentiellement fortes. Ce contexte est dû en grande partie à l'échantillonnage, alors même que les praticiens n'appliquent pas le processus d'estimation directement sur l'ensemble de la population mais « stratifient » celle-ci au préalable, afin de pouvoir bénéficier de l'hypothèse de comportement homogène des individus (il faut comprendre ici : homogène en termes de probabilité d'appartenance respectivement aux sources A et B). En effet, plus on se place sur des sous-populations découpées de manière *ad hoc*, plus on a de chances de satisfaire cette hypothèse d'homogénéité qui, on le rappelle, est une composante essentielle des hypothèses socles. Dans ces circonstances, les tailles de sous-population par grappe peuvent être (parfois) faibles et très dispersées, ce qui affecte la stabilité de $\hat{x}_{11}$. En outre, si on a la malchance d'avoir affaire à des listes dont l'intersection n'est pas très fréquente – typiquement le cas d'un recensement de mauvaise qualité car alors beaucoup d'individus contrôlés en liste B ne seront pas retrouvés en liste A – certains effectifs $x_{g,11}$ seront particulièrement faibles, et on trouvera donc une certaine proportion d'estimations particulièrement grandes. La pratique des contrôles du recensement (au moins) tend à produire des distributions de $\hat{\hat{N}}$ dissymétriques, avec asymétrie (*skewness*) positive.

Ce développement prend pour exemple un plan de sondage particulier, qui est ici un tirage aléatoire simple de grappes. Avec d'autres échantillonnages, les expressions formelles d'estimation, de calcul de biais et

9 En principe, le cas asymptotique correspond à $m\to\infty$ et $M\to\infty$, mais on est plutôt dans le contexte où soit $f=\frac{m}{M}\to 0$, soit $f\to constante$. Dans les deux cas, $\frac{M}{\sqrt{(m)}}\to\infty$.

d'erreur évolueront et seront certes plus complexes, mais sur le principe on pourra toujours adapter la théorie et procéder, dans le même esprit, à une estimation de la taille de la population totale.

6. On peut étendre le mécanisme au cas de 3 sources

Supposons que l'on dispose de 3 sources et que les hypothèses socles s'appliquent. En particulier il y a indépendance mutuelle de capture entre les 3 occasions respectives.

Tableau 3 : Tableau de contingence dans le cas de 3 sources

Présence dans les différentes sources			
Source A	**Source B**	**Source C**	**Notation**
1	1	1	x_{111}
1	1	0	x_{110}
1	0	1	x_{101}
1	0	0	x_{100}
0	1	1	x_{011}
0	1	0	x_{010}
0	0	1	x_{001}
0	0	0	x_{000}

On remarque que si on se place conditionnellement à la capture dans l'une des 3 sources (peu importe laquelle), alors on dispose de tous les dénombrements croisant les 2 autres sources. Une idée simple consiste alors à dénombrer les individus capturés dans la source 'conditionnelle' et à estimer la taille de la population non capturée dans cette source (on se ramène alors au cas à deux sources pour cette sous-population).

Si la source conditionnelle est la source A, on observe complètement le statut de l'effectif capturé par cette source, ce qui permet de calculer $x_{100}+x_{101}+x_{110}+x_{111}$. Dans la partie complémentaire, donc les individus non capturés par la source A, on peut utiliser l'estimation de Lijncoln-Petersen dans le cas de deux sources, soit $x_{001}+x_{010}+x_{011}+\left(\frac{x_{001}\cdot x_{010}}{x_{011}}\right)$, le tout dernier terme étant l'estimation de n_{000} - qui est le seul effectif qui soit inconnu (et qui justifie donc la mise en œuvre de toute cette procédure). La taille d'échantillon observée est notée n, et on a $n=x_{100}+x_{101}+x_{110}+x_{111}+x_{001}+x_{010}+x_{011}$. L'estimation de la taille de population N devient donc

$$\hat{N}=n+\left(\frac{x_{001}\cdot x_{010}}{x_{011}}\right).$$

Comme on peut effectuer l'exercice en conditionnant en amont par n'importe laquelle des 3 sources, une stratégie naturelle consiste à moyenner les 3 estimations obtenues, soit

$$\hat{N} = n + \frac{1}{3}.\left(\frac{x_{001}.x_{010}}{x_{011}} + \frac{x_{100}.x_{001}}{x_{101}} + \frac{x_{100}.x_{010}}{x_{110}} \right).$$

Un point important à noter : l'accès à l'information complète conditionnelle à la capture observée par une source ne permet pas de tirer de conclusion rigoureuse sur la pertinence de l'hypothèse d'indépendance qui sous-tend le processus d'estimation ci-dessus. On pourrait certes tester l'indépendance entre les 2 sources non conditionnelles par un banal test du khi-2 en se restreignant à la sous-population appartenant à la source conditionnelle. Il est alors vraisemblable qu'un rejet d'une hypothèse d'indépendance sur l'une des tables de contingence partielle soit fatal pour accepter l'indépendance mutuelle, mais il n'y a pas de garantie : d'une part on reste sans possibilité de test sur les sous-populations qui ne sont pas dans la source conditionnelle (puisqu'on ne les observe pas complètement, pas définition), et d'autre part il est illusoire de vouloir produire un test sur une table de contingence agrégée à partir des tables de contingence qui la composent : c'est un principe découlant de l'effet de structure – ou paradoxe de Simpson – bien connu : l'agrégation de 2 tables de contingence pour lesquelles le test a conclu à l'indépendance des variables qualitatives n'entraîne pas l'indépendance de ces mêmes variables au niveau de la table agrégée !

Lorsqu'on utilise trois listes A, B et C, la source additionnelle C permet néanmoins d'apprécier le degré de dépendance entre les sources A et B (ce qui est inenvisageable sans la présence de cette source C). En effet, le Odds ratio mesurant un degré de dépendance entre deux variables, on peut toujours le calculer quand on conditionne par rapport à une des sources observées : dans ce cas, aucune case de la table de contingence 2-2 n'est d'effectif inconnu, par construction. Avec 2 listes au contraire (estimateur DSE), on a toujours une case d'effectif inconnu, celle qui regroupe les individus n'appartenant à aucune des deux listes.

Par exemple voici l'Odds ratio conditionnel à l'appartenance à la liste C - les indices étant rangés dans l'ordre A, B, C :

$$\frac{x_{111}.x_{001}}{x_{101}.x_{011}}.$$

Si la source C relève d'une méthodologie de constitution complètement différente des 2 autres sources et si B et C sont de tailles comparables, l'agrégation de C à B devrait donc, de façon générale, apporter plus d'indépendance dans la relation avec la source A. L'opération d'agrégation préalable permet donc parfois de rendre plus acceptable (plus raisonnable...) l'hypothèse d'indépendance qui sous-tend l'utilisation de l'estimateur de Lincoln- Petersen.

En présence de 3 sources, on peut estimer x_{000} de deux façons supplémentaires (au moins) :

i) En ignorant tout simplement la source C, et en conservant les hypothèses socles.

$\hat{x}_{00+} = \frac{x_{10+}.x_{01+}}{x_{11+}}$ puis $\hat{x}_{000} = \hat{x}_{00+} - x_{001}$.

Il est possible que $\hat{x}_{000}<0$, situation traduisant une sous-estimation marquée de la taille de population totale (la taille manquante x_{000} est forcément positive ou – de manière extrême – nulle). Vouloir utiliser de manière explicite la liste C est, dans ce cas, mal venu pour l'opération d'estimation.

ii) La source C sert à renforcer la source B avec laquelle on la regroupe – on considère donc qu'on a 2 listes, qui sont A et $B \cup C$:

L'effectif de la source A qui n'est pas dans $B \cup C$ est : x_{100}
L'effectif de la source $B \cup C$ qui n'est pas A : $x_{010}+x_{011}+x_{001}$
L'effectif de la source A qui est également dans $B \cup C$ est : $x_{111}+x_{110}+x_{101}$

Alors sous les hypothèses habituelles d'indépendance entre A d'une part et $B \cup C$ d'autre part, $\hat{x}_{000}=\frac{x_{100}.(x_{010}+x_{011}+x_{001})}{x_{111}+x_{110}+x_{101}}$ puis finalement
$\hat{N}=x_{100}+x_{101}+x_{110}+x_{111}+x_{001}+x_{010}+x_{011}+\hat{x}_{000}$.

On peut opter pour le regroupement de 2 listes le plus adapté : soit C renforce A (la seconde liste étant B), soit B renforce A (la seconde liste étant C). Un éclairage de la meilleure option se fait en examinant les *odds ratios (OR)* conditionnels. Par exemple si l'OR conditionnel à A, défini par $\frac{x_{111}.x_{100}}{x_{101}.x_{110}}$, est proche de 1 – suggérant l'indépendance (ou une faible dépendance) entre B et C - on aura tendance à grouper prioritairement A avec B ou A avec C (plutôt que B avec C) en espérant profiter d'une « certaine indépendance » entre $A \cup B$ et C ou entre $A \cup C$ et B .

II. Comment estimer lorsque les hypothèses d'exactitude ne sont pas vérifiées ?

Les deux hypothèses que l'on a nommées 'hypothèses d'exactitude' (hypothèses H4 et H5) portent sur la qualité des données collectées. Le non-respect de ces hypothèses conduit à observer des valeurs entachées d'erreur, qui induisent un biais dans les estimateurs produits.
Elles ont également un impact sur la qualité des données utilisées par le processus d'estimation, par exemple sur la probabilité estimée d'être dans l'intersection observée entre A et B.

Le relâchement de ces hypothèses, séparément ou de façon conjointe, est discuté en détail dans l'article de Zhang et Dunne (2017).

1. Les estimations par système multiple en présence d'erreurs d'énumération (sur-couverture, doublons, enregistrements erronés)

Lorsqu'une des listes contient de la sur-couverture, à savoir des enregistrements qui ne correspondent pas à des unités réelles de la population ou des doublons, l'estimateur de Lincoln-Petersen est biaisé.

Dans l'approche qui suit, on va faire « comme si » l'indicateur de présence / absence d'une unité de U dans A était la résultante d'un processus déterministe. Pour développer cette approche de manière techniquement propre, on va utiliser un conditionnement. En conséquence, la variable

$$\delta_K(A)=1 \ \ si \ \ K\in A\cap U, \ \ 0 \ \ sinon$$

est considérée comme prenant des valeurs fixées, c'est-à-dire que l'on va conditionner par les valeurs $\delta_K(A)$. En revanche, on conserve son caractère aléatoire à la variable

$$\delta_K(B)=1 \ \ si \ \ K\in B\cap U, \ \ 0 \ \ sinon$$

considérant par ailleurs les hypothèses suivantes :

- on suppose qu'il y a l'homogénéité de capture dans B ;
- on suppose que A et B ont de la sous-couverture (c'est l'hypothèse générale faite jusqu'à présent) ;
- on suppose que B est exempt de sur-couverture, et que A seulement contient de la sur-couverture
on note r le nombre d'enregistrements dans A qui n'appartiennent pas à la population d'intérêt U ;
- on suppose vraie l'hypothèse d'appariement parfait.

$$\sum_{K=1}^{N}\delta_K(B)=x_{+1}$$

$$\sum_{K=1}^{N}\delta_K(A)=x_{1+}-r$$

L'homogénéité de capture dans B nous permet d'écrire :

$$\forall K\in U, \ E\delta_K(B)=P(\delta_K(B)=1)=P(\delta_K(B)=1|\delta_K(A)=1)=P(\delta_K(B)=1|\delta_K(A)=0)=\pi$$

$$E(x_{+1})=N\pi \qquad (1)$$

$$\sum_{K=1}^{N}\delta_K(A).\delta_K(B)=x_{11} \Rightarrow E(x_{11}|\delta(A))=\sum_{K=1}^{N}\delta_K(A).E(\delta_K(B))=\sum_{K=1}^{N}\delta_K(A).\pi$$

$$E(x_{11}|\delta(A))=(x_{1+}-r).\pi \qquad (2)$$

où $\delta(A)=(\delta_1(A),\delta_2(A),...,\delta_N(A))$

(1) et (2) nous permettent d'écrire

$$\hat{N}=[\frac{(x_{1+}-r).x_{+1}}{x_{11}}]$$

L'estimateur se fonde alors sur la méthode des moments, et non sur le maximum de vraisemblance.

Les conséquences de la sur-couverture sur le tableau de contingence et l'estimateur DSE

Dans le tableau de contingence, l'erreur d'énumération portera donc sur x_{10}, puisque si B ne contient pas de sur-couverture, alors les individus dénombrés en x_{11} et x_{01} appartiennent bien à la population d'intérêt.

Si on note r la sur-couverture, c'est à dire le nombre d'enregistrements surnuméraires dans la liste tirée de A, on a l'estimateur classique de Lincoln-Petersen qui s'écrit :

$$\hat{N}_{LP}=[\frac{x_{1+}.x_{+1}}{x_{11}}]$$

soit

$$\hat{N}_{LP}=[\frac{(x_{10}+x_{11}).x_{+1}}{x_{11}}]=[\frac{(x_{10}-r+r+x_{11}).x_{+1}}{x_{11}}]=[\frac{(x_{10}-r+x_{11}).x_{+1}}{x_{11}}+r.\frac{x_{+1}}{x_{11}}]$$

et finalement $\hat{N}_{LP}=[\frac{(x_{1+}-r).x_{+1}}{x_{11}}+r.\frac{x_{+1}}{x_{11}}]$.

Comparant à l'expression précédente issue de la méthode des moments, qui était sans biais (ou presque), il apparaît que le biais induit par une sur-couverture qui n'est pas explicitement prise en compte est

$$\hat{B}=\hat{N}_{LP}-\frac{(x_{1+}-r).x_{+1}}{x_{11}}$$

$$\hat{B}=r.\frac{x_{+1}}{x_{11}}=r.\frac{(x_{11}+x_{01})}{x_{11}}=r.(1+\frac{x_{01}}{x_{11}})$$

L'estimateur de Lincoln-Petersen surestime donc la taille de la population, le biais étant supérieur ou égal à la sur-couverture réelle r.

Zhang et Dunne (Zhang, Dunne 2017) parlent d'estimateur « naïf » pour l'estimateur de Lincoln-Petersen dans ce contexte, et parlent d'estimateur idéal pour l'estimateur corrigé de la sur-couverture.

$$\hat{N}_{naïf} = [\frac{x_{1+} \cdot x_{+1}}{x_{11}}]$$

$$\hat{N}_{idéal} = [\frac{(x_{1+} - r) \cdot x_{+1}}{x_{11}}]$$

$$\hat{N}_{naïf} > \hat{N}_{idéal}.$$

Estimateur ajusté, s'appuyant sur l'identification des données erronées

La difficulté est de disposer ici d'une estimation de cette sur-couverture. Elle est parfois obtenue grâce à une enquête de sur-couverture (dans le cadre d'enquêtes post censitaires par exemple) qui réinterroge des individus présents dans la liste A, pour évaluer leur appartenance à la population d'intérêt.

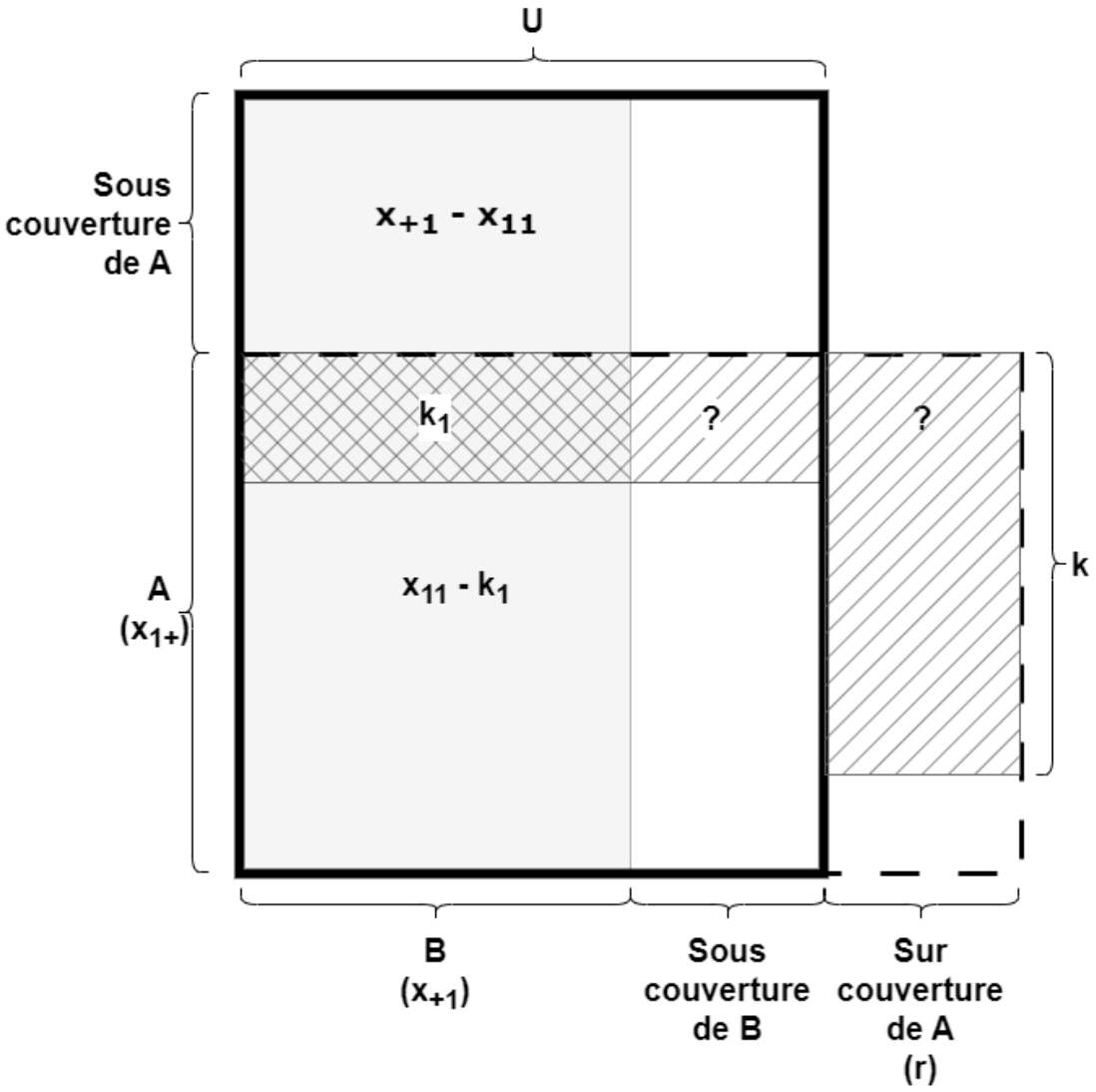
Dans de nombreux contextes, il n'est pas possible de réinterroger les individus. Cela peut être le cas par exemple lors de l'utilisation de données administratives.

Zhang et Dunne proposent un estimateur tronqué, s'appuyant sur la capacité à établir pour chaque individu un score corrélé à son appartenance réelle à la population.
Les méthodes de calcul du score ne sont pas discutées, mais on suppose que l'on dispose pour chaque individu de la source A d'un score de propension à appartenir à la population d'intérêt.
On suppose que ce score est au moins « un peu corrélé » avec l'appartenance réelle à la population d'intérêt. Cela suppose que l'on dispose d'informations individuelles sur les enregistrements présents dans la liste A, susceptibles d'être corrélées à l'appartenance à la population d'intérêt.
En fonction de ce score de propension, on va ôter (de manière progressive) les enregistrements que l'on considère comme erronés (c'est-à-dire qui ne sont en fait pas dans U – typiquement des doublons, des enregistrements qui ont eu par le passé une réalité mais qui sont maintenant périmés, ou même des enregistrements fictifs qui n'ont jamais eu aucune réalité), avant de réaliser une estimation DSE. L'hypothèse fondamentale consiste à considérer que l'amputation des k enregistrements de A supprime totalement la sur-couverture (ou autrement dit que les r enregistrements sont bien inclus dans les k enregistrements tronqués).

On note k le nombre d'enregistrements considérés ***a priori*** comme erronés et (donc) supprimés de A. À ce stade, la liste A a une taille $x_{1+} - k$. Comme le dénominateur de l'estimateur est formé du nombre d'individus présents dans les deux listes, l'amputation de A conduit naturellement à une amputation de cette intersection. On note k_1 le nombre d'enregistrements supprimés qui appartiennent aussi à B. La taille de l'intersection est alors $x_{11} - k_1$. En effet, comme B n'a pas de sur-couverture, ces k_1 enregistrements appartenant à A et à B ne sont pas erronés, et il est légitime de tous les soustraire de x_{11}.
On ne sait pas statuer sur le caractère réellement erroné ou non pour les $k - k_1$ unités non appariées avec B : comme B est affecté de sous-couverture, une partie de ces unités peut toujours être dans U (sans être dans B).

Schéma 2 : Présence de sur-couverture

(en situation où la sur-couverture n'est pas encore totalement supprimée)



Le système de *scoring* ne pouvant être parfait, on risque évidemment d'enlever de A des enregistrements qui sont dans U. Du point de vue théorique, la phase de troncature de A est sans conséquence, car une telle situation revient ni plus ni moins qu'à aggraver l'éventuelle sous-couverture de A toutes choses égales par ailleurs – on rappelle à l'occasion qu'on raisonne conditionnellement à la composition de A. Quelle que soit la façon dont les k enregistrements ont été sélectionnés – même si c'est à probabilités fortement inégales – on peut reprendre intégralement le processus de constitution de l'estimateur DSE à partir de la situation après amputation : les hypothèses précédentes restent parfaitement vérifiées si on part de la liste A amputée et de la liste B intégrale. Ainsi, on remplace mécaniquement x_{1+} par la nouvelle taille de A qui est $x_{1+} - k$, on remplace x_{11} par la nouvelle taille d'intersection qui est $x_{11} - k_1$, tandis que la taille de la liste B reste égale à x_{+1}. Puisqu'il n'y a plus de sur-couverture après l'amputation (hypothèse fondamentale), on peut appliquer l'expression classique de l'estimateur DSE.

On note maintenant $\hat{N}_0$ l'estimateur naïf, et $\hat{N}_k = \frac{(x_{1+} - k).x_{+1}}{(x_{11} - k_1)}$ l'estimateur pour lequel k enregistrements ont été supprimés.

La façon dont on sélectionne les k enregistrements devient désormais le principal enjeu, en particulier la façon dont k se positionne par rapport à la taille de la sous-couverture r. Dans le cas spécifique où la troncature se fait totalement au hasard, on peut dire que tout individu de A a une probabilité d'être éliminée égale à $\frac{k}{x_{1+}}$, ce qui fait que si on raisonne conditionnellement à B, $k_1 = \sum_{i=1}^{x_{11}} 1_{i \text{ éliminé}}$ suit une loi de Bernoulli $B\left(x_{11}, \frac{k}{x_{1+}}\right)$, et donc en moyenne k_1 vaut $x_{11} . \frac{k}{x_{1+}}$.

On peut aussi considérer que la sélection des k enregistrements se fait *ex ante* dans A (sans se référer à B) et que seulement ensuite la liste B est constituée. Chacun de ces k enregistrements se trouve – ou non - capturé dans la liste B, chacun indépendamment des autres avec une probabilité qui est la probabilité commune de capture dans B, estimée à partir de A par $\frac{x_{11}}{x_{1+}}$. Dans cette optique $k_1 = \sum_{i=1}^{k} 1_{i \in B}$ suit une loi de Bernoulli $B\left(k, \frac{x_{11}}{x_{1+}}\right)$. On aboutit – heureusement - à la même moyenne...

La situation optimale semble intuitivement celle où on ampute exactement la sur-couverture, ni plus ni moins, c'est-à-dire celle où $k = r$ et $k_1 = 0$.

Résultat 1

$k_1/x_{11} < k/x_{1+} \Rightarrow \hat{N}_k < \hat{N}_0$
$k_1/x_{11} = k/x_{1+} \Rightarrow \hat{N}_k = \hat{N}_0$

Si la détection des enregistrements erronés est plus efficace qu'une détection au hasard (on pourrait dire à l'aveugle), alors la proportion d'enregistrements supprimés dans l'intersection de A et B sera plus faible que la proportion d'enregistrements supprimés dans l'ensemble de A : le résultat 1 dit que l'estimateur ajusté sera alors plus petit que l'estimateur naïf, si bien qu'on ajuste l'estimateur dans la bonne direction (on réduit le biais).

Si on supprime les enregistrements au hasard, l'estimateur final a la qualité de l'estimateur naïf, ni plus ni moins.

La 'bonne' référence reste en permanence celle de l'estimateur idéal $\hat{N}_{idéal}$, qui est sans biais. On vise donc à s'en approcher au maximum, en conservant en mémoire deux éléments :
- on ne connaît pas r ;
- quand on choisit d'augmenter k, alors k_1 augmente, mais d'une manière non contrôlée, parce qu'on ne sait pas comment se partagent les k enregistrements entre U et la sur-couverture de A.

$$\hat{N}_{idéal}=[\frac{(x_{1+}-r).x_{+1}}{x_{11}}]$$
$$\hat{N}_k=[\frac{(x_{1+}-k).x_{+1}}{x_{11}-k_1}]$$

Résultat 2

$k<r \Rightarrow \hat{N}_{idéal}<\hat{N}_k$

Si le nombre d'enregistrements supprimés est plus petit que le nombre d'enregistrements erronés, alors l'estimateur ajusté reste supérieur à l'estimateur idéal.

Mais on ne connaît pas r …. Si on a une estimation du taux de sur-couverture au sein de la liste A, on peut avoir une approximation de la valeur maximale de k pour laquelle on est sûr de ne pas inverser le sens du biais (donc de maintenir une surestimation de la taille N).

Résultat 3

Si on supprime les enregistrements d'une façon suffisamment efficace pour éliminer tous les enregistrements erronés, alors on a $E(\hat{N}_k)=\hat{N}_{idéal}$, où E est l'espérance par rapport à l'aléa de constitution de B, conditionnellement à A[10].

En effet, ce scénario conduit à $k-r>0$ et il existe donc nécessairement des individus de U éliminés. Ces individus peuvent être captés dans B, avec une probabilité constante naturellement estimée par $\frac{x_{11}}{x_{1+}-r}$, parce que ce ratio est la proportion des individus de $A \cap U$ captés dans B[11]. On peut ainsi considérer que chacun des $k-r$ individus de U éliminés est capté dans B avec la probabilité $\frac{x_{11}}{x_{1+}-r}$. Donc $k_1=\sum_{i=1}^{k-r} 1_{i\in B}$ suit une loi de Bernoulli $B\left(k-r, \frac{x_{11}}{x_{1+}-r}\right)$. On en tire $E(k_1)=\frac{x_{11}}{x_{1+}-r}.(k-r)$ et finalement $E(\hat{N}_k)=\hat{N}_{idéal}$. Ainsi, si les r enregistrements erronés figurent parmi les k enregistrements supprimés, k étant alors plus grand que r, l'estimateur $\hat{N}_k$ vaut 'en moyenne' l'estimateur idéal.

10 L'élimination des k enregistrements ne s'effectue plus totalement au hasard (voir supra), mais de manière empirique, dans un esprit de sélection 'à choix raisonné'. Par conséquent, il n'est pas possible de préciser une loi de k_1 . Dans cette logique, il ne semble pas acceptable de traiter la liste B comme issue d'un processus aléatoire contrôlé - ce qui exclut au final tout raisonnement conditionnel à B.

Contrairement à ce qui était possible dans le cas d'une élimination totalement au hasard des k enregistrements (voir supra), il nous paraît difficile d'adopter un raisonnement conditionnel à B parce qu'on autorise une sélection empirique des enregistrements à éliminer. Il n'y a pas d'échantillonnage contrôlé – en pratique il serait plutôt à choix raisonné - et il est sans espoir de trouver une quelconque loi de k_1 dans ces conditions.

11 Donc aussi $P(B|A\cap U)$, qui vaut $P(A\cap B|A\cap U)$.

On tire de ce résultat 3 que l'objectif essentiel est de faire en sorte d'éliminer tous les enregistrements erronés ($k \geq r$) sans qu'il n'y ait à s'inquiéter de l'élimination concomitante de 'bons' enregistrements, et cela quelle que soit la méthode de sélection de ces enregistrements (on rappelle qu'on est toujours sous couvert des hypothèses socles). Dans ce cas idéal, on se débarrasse du biais.
Les résultats 1 et 2 disent que si on n'y parvient pas $(k<r)$ – donc si on conserve quand même dans A certains enregistrements erronés – on n'arrive pas à éliminer le biais mais on le réduit par rapport à l'estimateur naïf (qui est celui qui s'impose si on ne fait rien).

Estimation de variance

A est traitée comme fixe.

En linéarisant l'estimateur $\hat{N}_{idéal}$ on parvient à l'estimateur de variance de l'estimateur idéal :

$$\hat{V}(\hat{N}_{idéal}) = \frac{x_{+1}(x_{+1}-x_{11})(x_{1+}-r)(x_{1+}-r-x_{11})}{x_{11}^3} .$$

En faisant l'hypothèse que $E(\hat{N}_k) \approx N$, on obtient une estimation de la variance de l'estimateur tronqué

$$\hat{V}(\hat{N}_k) = \frac{x_{+1}(x_{+1}-(x_{11}-k_1))(x_{1+}-k)((x_{1+}-k-(x_{11}-k_1)))}{(x_{11}-k_1)^3} .$$

Critères d'arrêt de la troncature

On porte un intérêt tout particulier à la façon dont k_1 varie en fonction de k . On ne peut pas en dire grand-chose dans le cas le plus général, si ce n'est que k_1 est une fonction croissante de k . Certainement, la stratégie de *scoring* est déterminante, car l'élimination d'enregistrements erronés (en nombre r) n'a aucun effet sur k_1 alors que l'élimination de n'importe quel enregistrement de U peut augmenter k_1 : en effet, si l'enregistrement éliminé est dans la liste B, k_1 augmente de 1 (mais reste à sa valeur sinon).

Pour trouver une règle raisonnable reliant k_1 et k , il faut modéliser le processus. On va considérer que la liste A est constituée d'une sous liste A_r contenant les enregistrements les plus suspects dont l'intégralité de la sur-couverture (mettons que $r \geq 1$) et d'une sous-liste A_0 ne contenant donc que des enregistrements dans U . Ces deux listes n'ont évidemment aucune réalité, elles constituent seulement un cadre simple pour modéliser le processus d'élimination. Au sein de A_r , on n'a pas la capacité de faire des différences de qualité entre enregistrements, si bien que l'élimination peut être assimilée à une élimination complètement au hasard. Notons k_r la taille de A_r . On a $k_r \geq r$ et la proportion d'enregistrements erronés que contient cette liste est $p = \frac{r}{k_r}$, si bien que lorsqu'on sélectionne un enregistrement dans A_r pour l'éliminer, on a une probabilité p qu'il fasse partie des r enregistrements de la sur-couverture. Si on tire

cette fois k enregistrements (en restant dans A_r ce qui, impose $k<k_r$), il y en a en moyenne $k.(1-p)$ dans U . Si on note π la probabilité de capture dans B de n'importe quel individu de U , on se trouve avec (en moyenne toujours) $k.(1-p).\pi$ enregistrements dans B : c'est par définition la valeur de k_1 . On peut écrire, pour résumer

$$E(k_1|k, k \leq k_r) = k(1-p)\pi .$$

Si maintenant on élimine l'intégralité de A_r en ponctionnant aussi A_0 , alors $k>k_r$ et le nombre d'enregistrements éliminés dans A_0 est par définition $k-k_r$. Or, dans A_0 , tous les enregistrements sont dans U . Chacun a une probabilité π d'être aussi dans B. Le bilan conduit à éliminer en moyenne dans B un nombre d'enregistrements égal à $k_r.(1-p).\pi$ dans A_r (reprendre le résultat ci-dessus avec $k=k_r$) et égal à $(k-k_r).\pi$ dans A_0 . Au total, cela fait $(k-p.k_r).\pi = (k-r).\pi$. C'est aussi la valeur moyenne de k_1 . D'où

$$E(k_1|k, k>k_r) = (k-r)\pi .$$

Intuitivement, la qualité de l'estimation devrait dépendre de l'efficacité du calibrage de A_r . Il est probablement préférable d'isoler au plus près les r enregistrements de la sur-couverture, donc d'avoir k_r proche de r (p proche de 1). Mais cela nécessite de l'information *a priori* pour pouvoir affecter des scores pertinents. A noter que le plus 'mauvais' *scoring* est le *scoring* aléatoire[12], correspondant au cas où on n'est pas en mesure de distinguer une sous-population A_r pertinente, et que dans ce cas $p=\frac{r}{x_{1+}}$: c'est la situation où n'importe quel enregistrement de A, qu'il soit dans A_r ou dans A_0 , a la même probabilité d'être erroné et cette probabilité commune n'est autre que la proportion d'enregistrements erronés dans A. On peut donc partir du principe qu'on a toujours $\frac{r}{x_{1+}} \leq p \leq 1$.

Zhang et Dunne proposent plusieurs critères d'arrêt de la troncature lors des opérations de correction de la sur-couverture.

Premier critère

Un premier scénario est celui où on parvient à mettre en place un *scoring* suffisamment efficace pour éliminer rapidement les r enregistrements erronés. Dans ce cas, l'ajustement (par suppression d'enregistrements erronés) est meilleur que la suppression d'enregistrements au hasard et la suite $\hat{N}_k$ est décroissante. Selon le résultat (2), on a $\hat{N}_k > \hat{N}_{idéal}$, idéalement tant que $k<r$ mais cela se traduit dans notre modélisation par $k<k_r$, et (résultat 3) la suite $\hat{N}_k$ se stabilise à la valeur $\hat{N}_{idéal}$ dès que k contient l'intégralité des r erreurs, soit $k \geq k_r$, puis conserve cette valeur si éventuellement k continue à croître. Ce scénario est celui où on élimine l'intégralité de A_r avant de commencer à ponctionner A_0 . Le changement de régime se fait quand $k=k_r$.

12 On ne va pas jusqu'à imaginer un processus de *scoring* totalement contre-productif qui consisterait à éliminer en priorité des enregistrements appartenant à U … Un minimum de connaissance de la source devrait permettre d'éviter ce désastre !

Un second scénario est celui où ne parvient pas à éliminer 'rapidement' tous les enregistrements erronés (*scoring* insuffisamment efficace). On imagine mieux ce cas si p est petit, donc A_r grand. Après une phase initiale où on aura pu éventuellement éliminer de manière plus ciblée, on en arrive à une élimination quasi-aléatoire, Dans ces circonstances, l'estimateur $\hat{N}_k$ va également se stabiliser, mais à un niveau supérieur à $\hat{N}_{idéal}$.

Il convient de stopper le processus d'élimination dès qu'on constate que $\hat{N}_k$ n'évolue plus malgré l'augmentation de k , puisque l'estimation ne gagne plus en qualité. Dans le premier scénario (celui qu'on espère !) on a alors atteint le seuil $k=k_r$ et tous les enregistrements erronés ont été éliminés.

Second critère

On a vu que l'espérance de k_1 est une fonction affine de k , qui change de pente lorsque k dépasse k_r .
Il convient donc de stopper le processus d'élimination dès qu'on constate ce changement de pente.
Plus p est proche de 1, plus le changement de pente sera important.

Il est donc possible de détecter un changement dans le ratio $\frac{k_1}{k}$, signalant de fait qu'on arrive à la valeur critique $k=k_r$.

Dans la réalité, la probabilité p n'est pas constante, elle est vraisemblablement décroissante avec k (le *scoring*, même imparfait, permet d'éliminer les unités les plus probablement erronées avant les autres) : Reprenant l'égalité $E(k_1|k,k\le k_r)=k(1-p)\pi$, cela signifie que le ratio k_1/k prend des valeurs successives $(1-p)\pi$ où p est décroissante au fur et à mesure que k augmente[13] . On peut considérer que k_1/k se représente par une succession continue de droites dont la pente augmente quand k augmente, ce qui conduit à une fonction d'allure convexe, et un point d'arrêt de l'élimination peut raisonnablement se situer à l'endroit où le rayon de courbure de la fonction k_1/k est le plus fort..

Troisième critère

Comme le ratio k_1/k change au seuil k_r , on peut aussi s'attendre à ce que l'estimateur de variance $\hat{V}(\hat{N}_k)$ se comporte différemment avant et après k_r , ce qui nous donne un troisième critère d'arrêt.

Il est intéressant de tracer les 3 fonctions de k que sont $\hat{N}_k$, k_1 et $\hat{V}(\hat{N}_k)$ afin de détecter le seuil k_r qui marquera l'arrêt de l'ajustement.

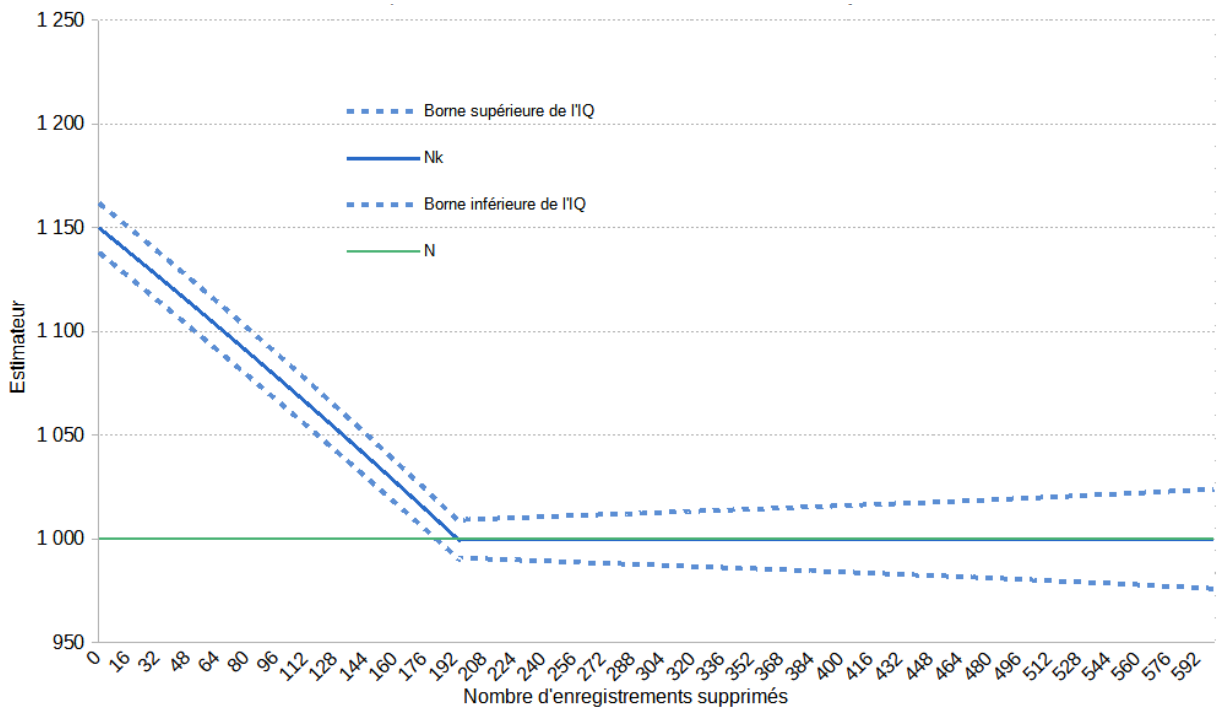
La performance du processus de suppression des enregistrements supposés erronés est cruciale.

Zhang et Dunne présentent différentes simulations, s'appuyant sur des variantes de π, de p, et de taux de sur-couverture.
Les représentations graphiques sont intéressantes pour évaluer le critère d'arrêt. Plus on est performant pour identifier les individus en sur-couverture, plus l'estimateur convergera rapidement vers sa valeur idéale.

13 On balaie une liste où le scoring est supposé être 'assez efficace' : en début de liste on trouve une forte proportion d'erreurs, et au fur et à mesure qu'on descend dans la liste, il y en a de moins en moins.

Le graphique ci-dessous illustre une simulation du comportement de l'estimateur $\hat{N}_k$ et de son intervalle de confiance (voir expression de la variance estimée supra), en fonction du nombre d'enregistrements supprimés.

Schéma 3 : Comportement de l'estimateur au fur et à mesure de l'ajustement



Notation

N	Taille réelle de la population U	1 000
π	Probabilité de capture dans B	0,90
x_{+1}	Taille de l'énumération B	900
r	Nombre d'individus en sur-couverture dans A	135
p	Probabilité qu'un individu enlevé soit erroné	0,70
k_r	Rang k_r où le comportement change ($k_r = r/p$)	192,86
X_{1+} - r	Nombre d'individus dans A appartenant à la population U	900
x_{1+}	Taille de l'énumération de A	1 035
$r/(x_{1+}-r)$	Proportion de sur-couverture dans A (par rapport à la vraie population)	0,15
x_{11}	Taille de l'intersection A et B	810

Cas où les deux listes contiennent de la sur-couverture

On conserve les hypothèses précédentes, et on supprime seulement l'hypothèse d'absence de sur-couverture de B. On note respectivement r_1, r_2 et r_{12} les nombres d'enregistrements erronés présents respectivement dans A, dans B, et dans l'intersection de A et B. On vérifie que les estimateurs évoluent ainsi :

$$\hat{N}_{naïf} = [\frac{x_{1+} \,.\, x_{+1}}{x_{11}}]$$

$$\hat{N}_{idéal} = [\frac{(x_{1+} - r_1).(x_{+1} - r_2)}{(x_{11} - r_{12})}]$$

Le biais de l'estimateur naïf sera du même signe que λ avec

$$\lambda = (\frac{r_1}{x_{1+}} + \frac{r_2}{x_{+1}} - \frac{r_1}{x_{1+}} \frac{r_2}{x_{+1}}) - \frac{r_{12}}{x_{11}} .$$

Si la proportion d'enregistrements erronés dans l'intersection $A \cap B$ est plus faible que le minimum des proportions d'enregistrements erronés dans chacune des sources A et B, alors l'estimateur naïf surestimera en moyenne la vraie taille de population.

Pour ajuster l'estimateur, il sera nécessaire cette fois de scorer les enregistrements de A et de B, et de les supprimer progressivement. On aboutit à

$$\hat{N}_k = \frac{(x_{1+} - k_1).(x_{+1} - k_2)}{(x_{11} - k_{12})} \text{ , avec } k = (k_1, k_2, k_{12}).$$

On retrouve l'estimateur idéal si on choisit la combinaison optimale (mais inconnue !) $k = (r_1, r_2, r_{12})$. On montre que la différence $\hat{N}_{naïf} - \hat{N}_k$ a le même signe que λ_k.

avec $\lambda_k = (\frac{k_1}{x_{1+}} + \frac{k_2}{x_{+1}} - \frac{k_1}{x_{1+}} \frac{k_2}{x_{+1}}) - \frac{k_{12}}{x_{11}}$.

Il n'y a pas de relation simple entre les valeurs λ et λ_k, mais il apparaît clairement qu'il faut éviter d'avoir $\lambda . \lambda_k \leq 0$, situation qui reviendrait à éloigner $\hat{N}_k$ de $\hat{N}_{idéal}$ encore davantage que ne l'est $\hat{N}_{naïf}$.
Comme on peut s'attendre à $\lambda > 0$, la théorie incite à concevoir le *scoring* afin que $\lambda_k > 0$ (mais la mise en pratique paraît peu réaliste….).

A retenir :

Alors que les hypothèses socles concernent la nature aléatoire des phénomènes d'appartenance aux listes, l'hypothèse d'absence de sur-couverture a trait à l'exactitude des données dont on dispose.
Le non-respect de cette hypothèse (présence de sur-couverture dans une source) entraîne un biais positif de l'estimateur DSE (surestimation de la taille de population).

Afin de corriger ou *a minima* de réduire ce biais, la solution proposée par Zhang et Dunne consiste à tenter de supprimer la sur-couverture en s'appuyant sur une modélisation de l'appartenance à la population d'intérêt U. La qualité de cette modélisation est bien sûr cruciale.
La suppression progressive des enregistrements supposés en sur-couverture permet de détecter le seuil d'arrêt.

Cette méthode ne garantit pas la suppression complète de la sur-couverture (et donc du biais induit), notamment si on ne dispose d'aucune information corrélée avec la probabilité d'appartenance à la population concernée.

2. Les estimations par système multiple en présence d'erreurs d'appariement

Dans un contexte de mesure de population animale, l'absence d'erreurs d'appariement revient à dire que les marquages effectués en première capture peuvent être observés sans erreur lors des captures ultérieures (il n'y a pas eu de perte de marquage, ni présence de marquage ne provenant pas de la première capture).
Ici, on s'intéressera au cas d'utilisation de sources de données déjà collectées. L'enjeu d'appariement devient dans ce cas crucial. En effet, l'hypothèse d'absence d'erreurs d'appariement n'est généralement pas valide. Les erreurs d'appariement entre les différentes sources se traduisent par des effectifs erronés dans le tableau de contingence.

Quelques rappels sur les appariements

Usuellement, on considère que le résultat d'un appariement est la classification de l'ensemble des paires d'enregistrements issues du produit cartésien des deux sources de données.
Chaque paire est considérée
- comme liée, ou retenue, si on considère que les deux enregistrements concernent le même individu ;
- comme non liée ou rejetée, si on considère que les deux enregistrements concernent deux individus différents.

On peut se tromper de deux manières différentes lors d'un appariement :
- en acceptant une paire à tort, on parle alors de *faux positif* ;
- en rejetant une paire à tort, on parle alors de *faux négatif*.
Ces erreurs d'appariement peuvent impacter le tableau de contingence de différentes manières.

Distinguer différents types d'erreur

Ding et Fienberg (Ding, Fienberg 1994) proposent de distinguer les appariements à tort (faux positifs) en quatre sous-cas.

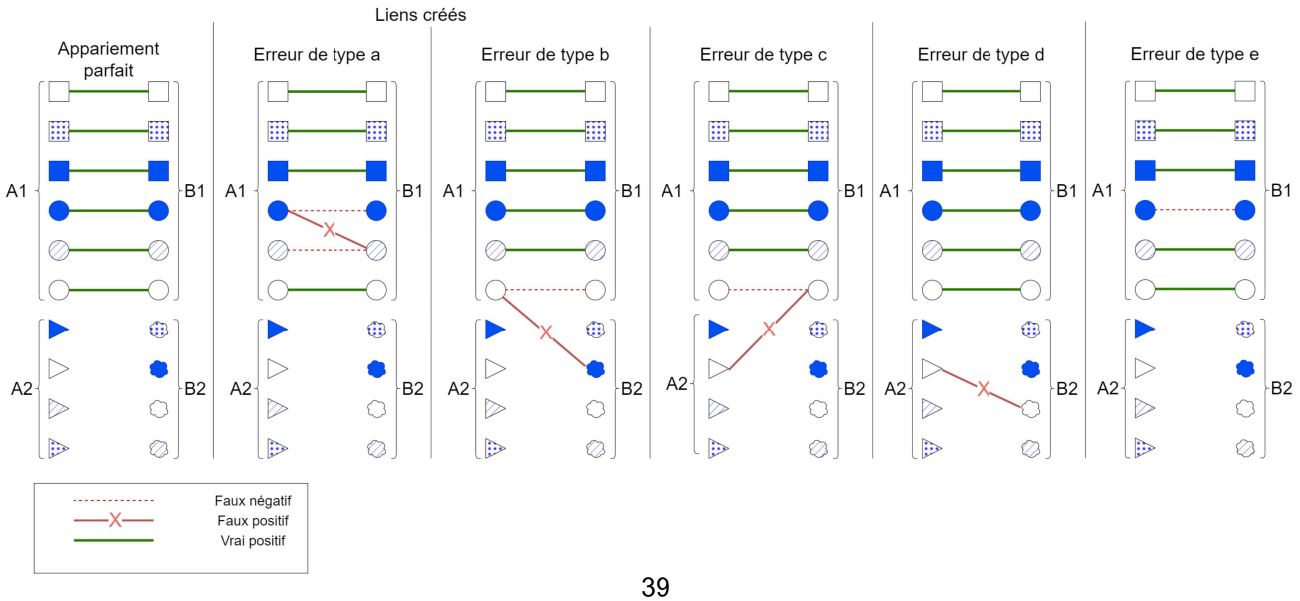
On décompose de façon théorique les listes A et B en deux sous-listes, afin de mieux se représenter les erreurs d'appariement : les sous-listes A_1 et B_1 ont la même taille, et à chaque enregistrement de A_1 correspond un enregistrement de B_1 représentant le même individu (les paires (a_i, b_i) appartenant à A_1 X B_1 sont les vrais positifs, mais elles ne sont pas connues).
Les sous-listes A_2 et B_2 contiennent des enregistrements pour lesquels il n'existe pas d'enregistrement correspondant au même individu dans l'autre source.
Autrement dit, les enregistrements de A_1 correspondent à des individus appartenant à l'intersection de A et B. Les enregistrements de A_2 correspondent à des individus appartenant seulement à la liste A. Il s'agit ici de l'appartenance réelle des individus aux listes. Mais lorsqu'on réalise un appariement, on peut se tromper, et apparier à tort des enregistrements qui représentent des individus différents. Lorsqu'on raisonne sur les erreurs d'appariement, on raisonne sur les liens : on doit classer l'ensemble des paires théoriques.

Ces sous-listes nous permettent de décrire plus précisément les faux positifs, ou appariements acceptés à tort, en décomposant les erreurs en différents types :

- erreur de type a : appariement à tort d'un individu de A_1 avec un individu de B_1 ; autrement on se trompe, mais au sein des sous populations A_1 et B_1 ; dans l'erreur de type a, on se trompe plusieurs fois : le lien faux positif masque en fait deux liens faux négatifs ;
- erreur de type b : appariement à tort d'un individu de A_1 avec un individu de B_2 ;
- erreur de type c : appariement à tort d'un individu de A_2 avec un individu de B_1 ;
- erreur de type d : appariement à tort d'un individu de A_2 avec un individu de B_2.

Schéma 4 : typologie des situations d'appariement

On suppose dans ce modèle qu'on ne se trompe généralement pas « plusieurs fois » pour un même enregistrement, et que les trois premières erreurs sont négligeables. On ne considère alors que les appariements à tort de type **d**.
Dans la suite, on nomme également erreur de type **e** les faux négatifs pour lesquels il y a uniquement oubli d'appariement.



Impact des erreurs d'appariement sur le tableau de contingence

Si l'on ne conserve que ces deux types d'erreur (« d » et « e »), on s'aperçoit que celles-ci affectent les effectifs du tableau de contingence de manière opposée, et que le biais sera donc lié à la prépondérance d'un défaut ou de l'autre dans les appariements :
- dans le cas d'un faux positif (erreur de type **d**), pour lesquels les individus ne sont présents que dans une seule des deux sources (autrement dit, on ne se trompe qu'une seule fois en liant à tort ces deux individus) => x_{11} sera sur estimé de 1, et x_{01} et x_{10} seront sous estimés de 1 ;
- dans le cas d'un faux négatif (erreur de type **e**), à savoir que l'on **n'a pas** apparié un enregistrement de A et un enregistrement de B qui concernait le même individu => x_{11} sera sous-estimé de 1, et x_{10} et x_{01} seront sur estimés de 1.

Il est important de souligner que les marges x_{1+} et x_{+1} ne sont pas affectées par les erreurs d'appariement. C'est bien la taille de l'intersection x_{11} et les tailles des sous-populations hors intersection x_{10} et x_{01} qui sont affectées.

Ainsi, dans l'estimateur classique de Lincoln-Petersen, c'est le dénominateur seulement qui est impacté par les erreurs d'appariements.

Modélisation DSE en présence d'appariement de Ding et Fienberg

Dans le paragraphe qui suit, on fait l'hypothèse qu'il n'y a pas de sur-couverture (doublons ou individus n'appartenant pas à la population), et que l'appariement réalisé est un appariement 1-1 (à chaque individu de A , on appariera au plus un seul individu de B et inversement).

Le modèle étudié autorise des faux négatifs (oubli d'appariement, type **e** sur le schéma). Deux enregistrements concernant le même individu (de fait, présent à la fois dans A_1 et B_1) ont une probabilité α d'être réellement appariés (autrement dit une probabilité 1- α d'être de faux négatifs). Il autorise également des faux positifs, avec la seule erreur de type **d**. On suppose alors que chaque enregistrement de A (en fait de A_2) a une probabilité β d'être apparié à tort à un enregistrement de B (en fait de B_2). Ces valeurs α et β sont les mêmes pour chaque individu.

On note p_{10}, p_{01} et p_{11} les probabilités respectives d'être présent dans la liste A seulement, la liste B seulement ou dans les deux listes.

Les hypothèses faites permettent de considérer - comme dans la partie I.2. (p15) - que le vecteur $(x_{11},x_{10},x_{01},x_{00})$ suit une loi multinomiale M(N ; $p_{11},p_{10},p_{01},p_{00}$), et que n (= $x_{11}+x_{01}+x_{10}$) suit une distribution binomiale de taille N et de probabilité $p_{11}+p_{10}+p_{01}$.

Dans un contexte qui n'impose pas de contrainte particulière sur les probabilités p_{ij} , on a alors deux fonctions de vraisemblance :

$$L_1(p_{11},p_{10},p_{01}) = \frac{(x_{11}+x_{10}+x_{01})!}{x_{11}!\,x_{10}!\,x_{01}!} \cdot \frac{{p_{11}}^{x_{11}}{p_{10}}^{x_{10}}{p_{01}}^{x_{01}}}{(p_{11}+p_{10}+p_{01})^{x_{11}+x_{10}+x_{01}}}$$

qui relève d'une approche conditionnelle à l'observation des effectifs d'individus x_{11},x_{10},x_{01} (les probabilités impliquées sont des probabilités conditionnelles à cette observation), et

$$L_2(N)=\frac{N!}{n!(N-n)!}(p_{11}+p_{10}+p_{01})^n[1-(p_{11}+p_{10}+p_{01})]^{N-n}$$

qui traduit la loi binomiale de n, avec $n=x_{11}+x_{10}+x_{01}$.

La maximisation de la fonction L_1 permet d'obtenir les estimateurs des trois probabilités p_{11}, p_{10}, p_{01} , et on utilise ensuite ces estimations pour la maximisation de L_2 afin de trouver l'estimation $\hat{N}$.

Dans un cadre de modélisation plus contraignante, en particulier du fait de l'hypothèse d'indépendance entre sources, les paramètres impliqués dans la modélisation sont : les probabilités individuelles d'appartenir aux deux sources respectives – notées p_A et p_B , ainsi que les probabilités α et β qui gouvernent les appariements (on pourrait tout aussi bien dire les erreurs d'appariement…).

En situation d'appariement parfait, on a (hypothèse socles) $p_{11}=p_A.p_B$. Cependant, en présence d'erreurs d'appariement, la probabilité d'être observé dans la case (1,1) du tableau n'est pas la probabilité réelle d'appartenance simultanée aux deux populations. Si l'on observe un individu dans cette case (1,1), deux cas se présentent :

- il appartient réellement aux deux listes, et l'appariement a été correctement réalisé ; cela s'effectue avec la probabilité $\alpha.p_A.p_B$;

- il appartient à A seulement (probabilité $p_A.(1-p_B)$), et il a été apparié à tort à un individu de B ; cela s'effectue avec la probabilité $\beta.p_A.(1-p_B)$. Noter qu'il n'y a volontairement pas de symétrie dans cette approche – pas de terme supplémentaire de type $\beta.p_B.(1-p_A)$ - car le parti est pris de considérer l'appariement comme une opération s'appuyant seulement sur la liste A_2 pour aller ensuite chercher un éventuel correspondant dans B_2[14].

On a alors $p_{11}=\alpha.p_A.p_B+\beta.p_A.(1-p_B)$ et on en tire $p_{10}=p_{1+}-p_{11}$ et $p_{01}=p_{+1}-p_{11}$.

Lorsqu'on injecte ces expressions dans les deux vraisemblances produites supra, on obtient d'abord des estimations $\hat{\alpha},\hat{\beta}$ et $\hat{p_A},\hat{p_B}$ à partir de L_1 , puis à partir de L_2 l'estimateur du maximum de vraisemblance de N, qui peut être exprimé par

$$\hat{N}=\frac{x_{11}+x_{10}+x_{01}}{\hat{p_A}+\hat{p_B}-(\hat{\alpha}-\hat{\beta})\hat{p_A}\hat{p_B}-\hat{\beta}\hat{p_A}} .$$

On peut appliquer cette méthode en distinguant des strates au sein desquelles les erreurs d'appariement sont homogènes (mais dépendent de la strate).

14 Ce point est discutable. Il ne nous semble pas facile à accepter, voire même à comprendre. Il y a une difficulté à raisonner avec des individus alors que l'appariement (et ses caractéristiques probabilistes) porte sur des couples. Cela suppose en particulier que les échecs d'appariement proviennent presque exclusivement des caractéristiques propres à l'individu présent dans A_2 et non de celles des autres individus – sinon il faudrait au minimum faire dépendre β de caractéristiques de la sous-population B_2 (de sa taille par exemple).

Discussion sur la prise en compte des erreurs d'appariement en présence de sur-couverture

Zhang et Dunne proposent aussi des éléments de prise en compte des erreurs d'appariement. Le cadre est celui proposé plus haut pour la prise en compte de sur-couverture dans une des deux listes seulement (ici la liste A). Le nombre d'individus appariés est noté x_{11L}. Ce nombre masque en fait des erreurs d'appariement. On note x_e le nombre de liens ratés (faux négatif de type **e)** et x_d le nombre de faux liens (faux positif de type **d**). Le nombre d'individus appartenant réellement à l'intersection est donc $x_{11}=x_{11L}+x_e-x_d$.

On note Γ l'ensemble $\{x_{11}, x_{10}, x_{01}, r\}$. On conditionne par Γ, ce qui revient à restreindre l'aléa au processus d'appariement – qui de fait est considéré comme un mécanisme aléatoire. On introduit deux paramètres pour prendre en compte les erreurs d'appariement, qui sont

$$\mu_L=E(x_{11L}|\Gamma)=x_{11}-E(x_e|\Gamma)+E(x_d|\Gamma)$$
$$\mu_L=x_{11}-x_{11}\frac{E(x_e|\Gamma)}{x_{11}}+E(x_d|\Gamma)\frac{E(x_{11L}|\Gamma)}{E(x_{11L}|\Gamma)}$$
$$\mu_L=x_{11}-x_{11}f+\mu_L q$$

$$soit\ \mu_L(1-q)=x_{11}(1-f)$$

avec

$f=\frac{E(x_e|\Gamma)}{x_{11}}$ et $q=\frac{E(x_d|\Gamma)}{E(x_{11L}|\Gamma)}$ qui sont respectivement les taux de faux négatifs parmi les vraies paires et de faux positifs parmi les paires constituées (ces 2 paramètres s'interprètent bien comme des taux d'erreur).

En remplaçant μ_L par x_{11L}, comme on le fait dans la 'méthode des moments', on obtient

$\widetilde{N}_L=\frac{x_{+1}\ (x_{1+}-r)}{\xi\, x_{11L}}$ en posant $\xi = \frac{(1-q)}{(1-f)}$.

Donc l'estimateur corrigé des erreurs d'appariement et de sur-couverture est

$$\widetilde{N}_L=\frac{x_{+1}\ (x_{1+}-r)}{\xi\, x_{11L}}.$$

Les valeurs x_{+1} et x_{1+} sont connues, s'agissant des tailles constatées des listes A et B (les erreurs d'appariement ne les affectent pas – étant par ailleurs rappelé qu'il n'y a pas d'enregistrement erroné dans B et que tout enregistrement erroné dans A est systématiquement comptabilisé dans r, si bien que $(x_{1+}-r)$ est exactement la taille de la vraie population $A\cap U$). La valeur x_{11L} est également constatée. La valeur de r est en revanche inconnue. Idem pour celle de ξ.

$\widetilde{N}_L$ est l'équivalent de l'estimateur appelé 'idéal' dans un contexte sans erreur d'appariement.

Si l'on compare cet estimateur à l'estimateur naïf $\hat{N}_{naïf}=\frac{x_{+1}\ x_{1+}}{x_{11L}}$, on peut écrire :

$$\frac{\widetilde{N}_L}{\hat{N}_{naïf}} = \frac{(x_{1+} - r)}{x_{1+}\ \xi}$$

$$\frac{\widetilde{N}_L}{\hat{N}_{naïf}} = (1 - \frac{r}{x_{1+}})(1 + \frac{q-f}{1-q})$$

L'estimateur idéal $\widetilde{N}_L$ étant considéré comme sans biais, le biais de l'estimateur naïf est donc fonction simultanément du taux de faux positifs q, du taux de faux négatifs f, et du taux de sur-couverture $\frac{r}{x_{1+}}$.

On constate que tous les scénarios sont possibles, selon le positionnement de q par rapport à f. Il est clair que la présence du taux de sur-couverture joue systématiquement dans le sens d'une surestimation de $\hat{N}_{naïf}$. Si $f > q$, le biais de surestimation de $\hat{N}_{naïf}$ ne peut que s'aggraver. En revanche, si $f < q$, il peut y avoir sous estimation de N par $\hat{N}_{naïf}$. On le vérifie par exemple si $f = 0$ et $\frac{r}{x_{1+}} < q$.

Dans la pratique, le taux de faux positifs q peut être estimé en analysant des paires acceptées, qui est un ensemble de taille réduite, puisque $O(N)$. Comme la volumétrie reste quand même importante, on pourra procéder par échantillonnage de paires, dont on va apprécier au cas par cas la pertinence de l'appariement, pour estimer au final un taux d'erreur.

A l'inverse, le taux de faux négatifs f doit être estimé parmi l'ensemble des vraies paires, que l'on ne connaît pas *a priori* et que l'on doit par conséquent considérer comme étant l'ensemble des paires possibles : c'est un ensemble de taille $O(x_{1+}\ .\ x_{+1})$, donc de taille gigantesque dès lors que l'une des sources au moins prétend être exhaustive. Dans ces conditions, procéder par échantillonnage est théoriquement possible, mais on est face à un problème typique d'estimation de taille d'une population rare (la proportion de faux négatifs est supposée être modeste), ce qui exige des échantillons de taille énorme pour obtenir une estimation de taux d'erreur ayant une précision correcte. Il est donc généralement très compliqué d'avoir une estimation de ce taux.

Les opérations d'apurement statistique portent donc, en pratique, sur la réduction de q beaucoup plus que sur celle de f, si bien que la situation la plus probable est celle où $f > q$, ce qui va dans le sens d'une aggravation du biais de $\hat{N}_{naïf}$.

Dans ce contexte, Zhang et Dunne proposent de corriger les phénomènes de sur-couverture et d'erreurs d'appariement, en adaptant l'estimateur ajusté :

$$\widetilde{N}_k = \frac{x_{+1}\ (x_{1+} - k)}{\hat{\xi}\ (x_{11L} - k_{1L})}\ .$$

La correction proposée ici tient compte des erreurs d'appariement. Il est nécessaire d'estimer celles-ci, par les méthodes usuelles d'estimation (annotation de paires, fichier étalon or[15]), avec les limites mentionnées pour l'estimation de f.

15 On appelle *gold standard* ou fichier étalon un échantillon de paires, représentatif des fichiers à apparier (en matière de variables d'appariement et de distribution des erreurs) et dont le vrai statut est connu. Cependant, dans la majorité des cas, il n'existe pas de *gold standard* et il faut donc ajouter une étape d'annotation manuelle pour qualifier un ensemble de paires, en faisant intervenir un observateur humain : la paire proposée par le processus est-elle une « vraie paire » ?, ou le processus a-t-il apparié à tort ?, ou bien encore est-il impossible de trancher ?

S'appuyer sur une estimation unique de ces deux taux revient à faire l'hypothèse que les taux d'erreurs d'appariements sont homogènes sur l'ensemble de la population (notamment pour la sous-population des enregistrements éliminés versus ceux qui sont conservés). Cette nouvelle hypothèse est extrêmement forte, et rarement valide ; aussi, il est fréquent de recourir à des estimations stratifiées, pour limiter la trop forte dépendance à cette hypothèse d'homogénéité.

Cependant, si les variables explicatives de l'hétérogénéité des erreurs d'appariement diffèrent de celles expliquant l'hétérogénéité des probabilités de capture, il devient difficile de procéder simultanément à tous les ajustements nécessaires.

En ce qui concerne l'effet sur le biais de l'estimateur naïf, en général, les erreurs d'appariement agissent dans le même sens que la sur-couverture, mais faire une hypothèse abusive d'indépendance des captures agit pour sa part dans le sens opposé. L'un dans l'autre, on peut donc dire qualitativement qu'il y a une forme de compensation, mais on ne peut pas en apprécier l'ampleur, et au final il est clair que l'estimateur naïf est biaisé.

On retiendra que la prise en compte des erreurs d'appariement dans le modèle DSE pose plusieurs difficultés :

- on a raisonné ici en simplifiant les erreurs d'appariement possibles (notamment pas d'erreurs multiples) ; il existe des méthodes permettant d'aller plus loin dans la prise en compte des erreurs d'appariement, au prix d'une plus grande complexité méthodologique (Zult et al. 2021; de Wolf 2019) ;

- l'estimation ajustée repose sur une évaluation des taux de faux positifs et de faux négatifs : ce second taux est beaucoup plus difficile à estimer de façon précise ;

- cet estimateur fait également l'hypothèse d'homogénéité dans les erreurs d'appariement, ce qui n'est certainement pas le cas. Cela implique qu'il sera nécessaire dans ce cadre d'effectuer des estimations post-stratifiées.

A retenir :

La présence d'erreurs d'appariement peut entraîner un biais dans l'estimateur DSE. Le signe du biais sera lié à la prépondérance d'un défaut d'appariement (la prépondérance de faux positifs entraîne un biais négatif de l'estimateur, soit une sous-estimation de la taille de population).

Les méthodes de correction de ce biais reposent sur des coefficients correctifs, fondés sur des estimations de taux de faux positifs et de taux de faux négatifs.

III. Relâchement des hypothèses socles pour l'estimation

Jusqu'ici, seul l'estimateur de Lincoln Petersen a été considéré. Dans la partie I on se plaçait en situation idéale en supposant que les hypothèses socles et les hypothèses d'exactitude étaient vérifiées. Dans la partie II, on a donné une méthode pour adapter l'estimation en présence de sur-couverture et de certains types d'erreur d'appariement, c'est-à-dire en relâchant successivement les deux hypothèses d'exactitude - mais l'expression de Lincoln Petersen demeurait à la base de la méthodologie d'estimation.

Au-delà du problème posé par la sur-couverture, on est constamment resté sous la dépendance des hypothèses socles, dont les plus contestables sont d'une part l'indépendance entre les sources et d'autre part l'homogénéité des probabilités d'appartenance aux sources. Lorsqu'on va plus avant dans l'abandon de ces hypothèses qui conditionnent la théorie de l'estimation, ce qui est aussi une façon de se rapprocher de la réalité des situations rencontrées en pratique, on doit envisager d'autres estimateurs que l'estimateur de Lincoln Petersen.

Avant d'aborder des approches plus complexes, il faut mentionner une piste facile et efficace qui permet, dans certaines circonstances, de conserver l'avantage de simplicité qu'apporte l'estimateur de Lincoln Petersen. Si les hypothèses socles ne sont pas vérifiées quand on considère la population cible dans son ensemble, il est en revanche possible qu'elles soient acceptables sur des sous-populations qui la partitionnent. On peut alors raisonner sous-population par sous-population, et additionner au final les estimations de Lincoln Petersen obtenues sur chacune de ces sous-populations. Cette technique, dite de 'post stratification' dans la littérature, doit être systématiquement envisagée avant de mettre en œuvre les méthodes présentées dans cette partie III.

A noter que si la (post) stratification permet de traiter des problèmes d'hétérogénéité et/ou de dépendance, la somme des estimations dans chaque strate diffèrera de l'estimation du total obtenue sans stratification, ce qui permet de conforter l'utilisation de la stratification. Toutefois, il convient de mentionner quelques limites :
- la stratification nécessite la connaissance « parfaite » des variables de stratification pour chaque individu considéré au sein de chacune des listes ;
- on ne peut jamais être certain d'avoir traité entièrement les problèmes d'hétérogénéité et de dépendance, notamment il y a toujours un risque d'hétérogénéité et de dépendance causées par des variables inobservées ;
- lorsqu'on dispose d'au moins trois sources, on peut détecter une éventuelle dépendance des sources deux à deux.

Enfin, la stratification peut aussi être intéressante pour corriger les erreurs d'appariement, comme cela est précisé dans la partie précédente.

Dans toute la suite, on maintient les hypothèses d'exactitude.

1. Relâchement de l'hypothèse d'indépendance entre sources

Tout ce qui précède se justifie, entre autres, sous l'hypothèse d'indépendance H3 : on considère que la probabilité d'appartenance conjointe (respectivement de non-appartenance conjointe) à un ensemble de sources est systématiquement égale au produit des probabilités d'appartenance (respectivement de non-appartenance) à chacune des sources prise isolément - cela en considérant toutes les combinaisons possibles d'appartenance/non-appartenance[16].

S'agissant de capture d'animaux, on peut s'attendre à ce que ces derniers soient sensibles à une forme d'apprentissage, et que leur probabilité de capture à la seconde occasion soit non seulement différente de leur probabilité de capture à la première, mais surtout qu'elle soit influencée par l'évènement de capture éventuelle lors de la première occasion, et dans ce cas il n'y a clairement plus d'indépendance. S'il y a capture lors d'une première occasion, il peut en découler une diminution de la probabilité de capture à la seconde occasion (situation dite '*shy trap*') ou ce peut être le contraire[17] (situation '*happy trap*').

S'agissant de fichiers administratifs, il peut exister des corrélations entre les présences dans les fichiers : par exemple, il y a une corrélation forte entre l'appartenance à la source DSN (Déclarations sociales nominatives) et l'appartenance à la source fiscale, car les DSN sont utiles pour le calcul du prélèvement à la source. Certains individus absents d'une liste donnée auront plus de chances d'être également absents des autres listes, du fait de leurs caractéristiques propres ou du fait des organismes qui constituent ces listes - pour des raisons méthodologiques par exemple. On peut penser que ce phénomène affectera davantage des sous-populations bien particulières, par exemple des personnes en situation de grande précarité, qui ont une propension plus forte à échapper aux enregistrements administratifs.

Lorsqu'il y a dépendance, en conservant l'hypothèse que les probabilités de capture ne dépendent pas de l'individu, l'estimateur de Lincoln-Petersen est biaisé, avec un biais qui – contrairement à la situation standard (partie I) – n'est plus négligeable *a priori.* En situation *happy trap*, il sous-estime la vraie taille N, et en situation *shy trap* il surestime N. Pour le percevoir, considérons x_{1+} le nombre d'individus capturés parmi N à la première occasion. À la seconde occasion, x_{+1} désigne le nombre total de captures et x_{11} le nombre d'individus de nouveau capturés parmi les x_{1+} individus capturés à la première occasion (ce qui nécessite un système d'identification performant, sous forme d'un marquage permanent lors de la première capture pour le monde animal). Dans un cas *happy trap* on peut s'attendre à une probabilité d'être capturé une seconde fois conditionnellement à une première capture (rapport $x_{11} \,/\, x_{1+}$) qui soit supérieure à la probabilité de seconde capture (rapport $x_{+1} \,/\, N$). D'où $\frac{x_{11}}{x_{1+}} > \frac{x_{+1}}{N}$, impliquant $N > x_{+1}.\frac{x_{1+}}{x_{11}}$: l'estimateur de Lincoln-Petersen sous-estime la taille de population.

Cette même conclusion peut être obtenue au travers de considérations impliquant le *odds ratio*. De façon générale, le *odds ratio* peut quantifier l'écart à l'indépendance. Reprenant l'encadré sur les *odds ratios*, la corrélation entre les sources A et B s'avère d'autant plus positive que l'information sur l'appartenance à A fait croître la probabilité d'appartenance à B, c'est-à-dire que $P_{1B|1A}$ est grand devant $P_{1B|0A}$, ou

16 $P(A \cap B) = P(A).P(B)$ et $P(\bar{A} \cap \bar{B}) = P(\bar{A}).P(\bar{B})$. La première égalité entraîne la seconde.

17 Par exemple s'il est attiré dans le piège par une nourriture appétissante et qu'il est relâché délicatement, il refera volontiers l'expérience !.

encore que *odds(1,A)* est grand devant *odds(0,A)*. Cela se traduira au final par un *odds ratio* plus grand, et en tout cas plus grand que 1. S'il vaut θ , on estime x_{00} par $\hat{x}_{00}(\theta)=\theta.\frac{x_{01}.x_{10}}{x_{11}}=\theta.\hat{x}_{00}$. L'estimateur final 'correct' de la taille totale de la population devient $\hat{N}(\theta)=x_{11}+x_{01}+x_{10}+\theta.\hat{x}_{00}$. Si on postule une indépendance entre les listes alors qu'il y a en réalité dépendance, on va utiliser l'estimateur $\hat{N}=x_{11}+x_{01}+x_{10}+\hat{x}_{00}$ au lieu de $\hat{N}(\theta)$. De ce fait, on introduit un biais (asymptotique) : si $\theta<1$, traduisant une corrélation négative entre sources qui est la situation *shy trap,* $\hat{N}/\hat{N}(\theta)>1$ et on surestimera N ; si $\theta>1$, (situation de corrélation positive *happy trap*) on sous estimera N .

On peut apprécier numériquement la sensibilité (robustesse) du ratio $\hat{N}/\hat{N}(\theta)$ - donc le risque de sur/sous estimation de N - à la valeur de θ en faisant varier θ , par exemple en considérant différentes valeurs[18] situées entre 0,5 et 2. Soit A la source qui couvre le mieux la population complète (B est donc l'autre source). On peut vérifier que cette sensibilité est numériquement d'autant plus importante que A couvre mal B (c'est-à-dire que $\frac{x_{11}}{x_{+1}}$ est petit) : ce cas correspond à des petites valeurs de x_{11} et donc à de grands $\hat{x}_{00}(\theta)$. Il semble donc qu'il y ait plus de risque de biais lorsqu'on utilise des sources corrélées négativement.

Un premier modèle simple avec deux sources

L'introduction la plus naturelle d'une dépendance entre sources passe par l'influence de l'apprentissage. En effet, l'apprentissage génère par nature de la dépendance : il peut engendrer soit une situation '*Happy trap*' soit une situation '*Shy trap*'. On va considérer deux opérations de piégeage, notées A et B.

Le modèle d'apprentissage peut fonctionner ainsi : chaque piège a sa propre probabilité de capture, qui est la même pour tous les animaux (on maintient l'hypothèse d'homogénéité des probabilités de capture, qui est l'une des composantes des hypothèses socle), et on considère que le piège A capture tout animal avec une probabilité p tandis que le piège B capture avec cette même probabilité p tout animal non capturé la première fois (donc non attrapé au piège A). Selon ce modèle, on peut dire que pour tout animal, la « première capture » se fait avec la même probabilité à chaque occasion.

Les hypothèses se traduisent ainsi : $p(A)=p_{1+}=p$ et $p(B|\bar{A})=\frac{p_{01}}{p_{0+}}=p$, en notant $\bar{A}$ l'évènement complémentaire de A . Par ailleurs, on note $p(B|A)=\frac{p_{11}}{p_{1+}}=c$, où c joue un rôle de paramètre.

18 On peut construire d'innombrables scénarii de listes A et B plus ou moins corrélées (= avec une intersection plus ou moins grande) à partir d'une population donnée de taille N , puisque la seule contrainte est d'assurer $x_{11}+x_{01}+x_{10}+x_{00}=N$. A chaque scénario, correspond une valeur de θ . Si on se fixe x_{11} , alors θ peut encore prendre autant de valeurs qu'on forme de couples (x_{01},x_{10}) .

La situation *'happy trap'* – première situation concrète où l'indépendance a disparu - survient, par définition, lorsque $p(B|A) > p(B)$, donc lorsque $c > c.p + p.(1-p)$, soit finalement $c > p$. La mesure de dépendance peut être résumée par le Odds ratio en revenant à sa définition

$$\theta = \frac{p(B|A)}{1-p(B|A)} \times \left(\frac{p(B|\bar{A})}{1-p(B|\bar{A})}\right)^{-1},$$

qui est aussi en toute circonstance et après quelques lignes de calcul $\theta = \frac{p_{11} \cdot p_{00}}{p_{10} \cdot p_{01}}$

Sous ce modèle, $\theta = \frac{c(1-p)}{p(1-c)}$. Cette expression a une interprétation très simple : si $c > p$ alors $\theta > 1$ et on est en situation '*happy trap'* (corrélation positive entre les deux listes), si $c < p$ alors $\theta < 1$ et on est en situation *'shy trap'* (corrélation négative entre les deux listes). Ces deux situations sont des situations de dépendance entre sources : seul le cas $\theta = 1$ traduit une indépendance.

Partant de là, obtenir $\hat{N}$ résulte d'une opération purement technique. La densité de probabilité associée à ce contexte est celle d'une loi multinomiale[19], avec trois observations x_{11}, x_{10}, x_{01} et cinq paramètres, à savoir 4 probabilités plus la taille totale de population N. Au niveau des probabilités, on implique en réalité 3 probabilités inconnues puisque la somme des 4 probabilités initiales vaut 1. Par ailleurs, le modèle $p(A) = p(B|\bar{A})$ introduit une contrainte supplémentaire qui réduit à 2 le nombre de paramètres de type probabilité. La densité peut s'écrire en utilisant seulement les 3 paramètres N, c, p, à estimer à partir de 3 observations. Le maximum de vraisemblance peut donc être obtenu sans qu'il y ait des relations imposées entre les paramètres estimés. La maximisation de la densité conduit aux estimations $\hat{N}$, $\hat{p}$ et $\hat{c}$ suivantes (Wolter 1986) :

$$\hat{N} = \frac{(x_{11}+x_{10})^2}{x_{11}+x_{10}-x_{01}},$$

$\hat{p} = 1 - \frac{x_{01}}{x_{11}+x_{10}}$ et $\hat{c} = \frac{x_{11}}{x_{11}+x_{10}}$ - sous réserve que $\hat{N} \geq 0$.

À distance finie (mais ce sera une situation en pratique très rare[20]), les propriétés de biais et de variance peuvent s'obtenir (de manière certes un peu compliquée…) à partir des moments des lois binomiales suivies

19 La densité associée aux N individus est celle d'une multinomiale car les comportements de ces individus sont mutuellement indépendants. Le schéma probabiliste reste celui de tirages 'indépendants et identiquement distribués' de N valeurs dans une loi commune qui affecte à chaque individu une modalité et une seule parmi les 4 situations possibles croisant l'appartenance ou non à A (1er piégeage) et l'appartenance ou non à B (2nd piégeage). La densité élémentaire associée à tout individu K est $p_{11}^{1_{K\in 11}} . p_{10}^{1_{K\in 10}} . p_{01}^{1_{K\in 01}} . p_{00}^{1_{K\in 00}}$. Il reste à exprimer chaque p_{ij} en fonction des paramètres N, c, p, et à multiplier ces N densités élémentaires.

20 Pour l'Insee, les plus petites populations considérées seront des populations communales – et à ce niveau vraisemblablement il s'agira de communes de grande taille.

par les composantes x_{11}, x_{10}, x_{01} . En situation asymptotique (plus vraisemblable), les propriétés de biais et de variance sont celles – bien connues - du maximum de vraisemblance,

Un second modèle en présence d'une troisième source

Lorsqu'on dispose de trois sources, l'une des sources – mettons la source C – peut être exploitée pour quantifier le degré de dépendance entre les deux autres sources – soit A et B. L'intensité de la dépendance est alors mesurée parmi la sous-population présente dans la liste C (c'est une évaluation conditionnelle à C). Un modèle envisageable consiste à appliquer cette dépendance à la population entière.

Le degré de dépendance entre A et B est mesuré par le *Odds ratio* conditionnel à l'appartenance à C, soit

$$R = \frac{x_{111} \cdot x_{001}}{x_{101} \cdot x_{011}}$$

et on fait ensuite l'hypothèse que cette intensité est la même dans la population complète, ce qui conduit à

$R = \frac{x_{11+} \cdot \hat{x}_{00+}}{x_{10+} \cdot x_{01+}}$, puis finalement $\hat{x}_{00+} = \frac{x_{111} \cdot x_{001}}{x_{101} \cdot x_{011}} \cdot \frac{x_{10+} \cdot x_{01+}}{x_{11+}}$ dont on tire $\hat{x}_{000} = \hat{x}_{00+} - x_{001}$, puis $\hat{N}$.

Cette approche par modèle (simple) est une façon de prendre en compte la dépendance entre listes et peut être étendue au cas où il y a plus de 2 listes.

L'appel aux modèles log-linéaires, en présence de 3 sources

L'approche par modélisation de la probabilité de capture est la bonne façon de traiter la dépendance entre sources. L'utilisation d'un modèle individuel distinguant un effet individu et un effet source est probablement l'approche la plus naturelle de la question. Mais décrire des comportements de nature individuelle (on parle de « modèle écologique ») à ce niveau de finesse est une approche exigeante. Au contraire, les modèles log-linéaires à venir assurent des modélisations au niveau 'macro' et sont plus faciles à concevoir et à utiliser.

On va considérer maintenant 3 sources A, B et C, et on va comme précédemment utiliser la notation i, j, k pour caractériser l'appartenance ou non d'un individu à ces listes respectives. Ainsi, on peut considérer le jeu de probabilités de type p_{ijk} où chaque indice prend la valeur 1 lorsque l'individu appartient à la source associée, et la valeur 0 sinon. Ainsi, p_{101} par exemple désigne la probabilité qu'un individu quelconque de la population soit présent dans les sources A et C, et absent de la source B. On définit ainsi 8 probabilités, vérifiant

$$\sum_{\substack{i\in\{0,1\}\\ j\in\{0,1\}\\ k\in\{0,1\}}} p_{ijk}=1 .$$

On utilisera des signes + pour définir les probabilités marginales, selon le principe $p_{i++}=\sum_{\substack{j\in\{0,1\}\\ k\in\{0,1\}}} p_{ijk}$.

Le cas le plus général est celui où on ne met aucune contrainte sur les p_{ijk} - hormis évidemment la sommation à 1. Cela conduit à laisser libres 7 valeurs de p_{ijk} .

L'indépendance entre les 3 sources, postulée jusqu'à présent comme hypothèse fondamentale, revient à imposer des relations extrêmement contraignantes aux p_{ijk} , puisque ces paramètres doivent alors vérifier

$$\forall(i,j,k) \qquad p_{ijk}=p_{i++}\cdot p_{+j+}\cdot p_{++k} .$$

Sous ces conditions, les 8 probabilités initiales p_{ijk} sont décrites par 6 paramètres. Du fait des propriétés de sommation à 1, au final 7 probabilités sont décrites par 3 paramètres – par exemple on peut choisir les paramètres p_{1++} , p_{+1+} et p_{++1} .

Ces deux configurations – d'une part celle où les probabilités sont sans contrainte, d'autre part celle où il y a indépendance entre les trois sources - sont les configurations extrêmes en termes de quantité d'information mobilisée *a priori*. On peut construire une grande variété de contextes en se plaçant dans des situations intermédiaires, c'est-à-dire exploiter plus de paramètres qu'il n'en faut pour décrire l'indépendance mutuelle, mais moins que ne requiert la situation la plus générale. On est alors résolument dans l'univers des modèles, s'agissant dans tous les cas de formuler des hypothèses simplificatrices de la réalité. La sémantique utilisée est au demeurant troublante : on relie des probabilités à des probabilités, les variables à gauche du signe d'égalité – qui jouent le rôle de variables expliquées – sont de même nature que les variables situées à droite de l'égalité, lesquelles jouent plutôt le rôle de variables explicatives.

Pour pouvoir tirer des propriétés de cette approche, on l'a inscrite dans un cadre formel dit des modèles linéaires généralisés (*Generalized Linear Models - GLM*), et plus spécifiquement ici des modèles log-linéaires, qui constituent un cas particulier des GLM. On part de l'information observée, ici les effectifs x_{ijk} constatés pour chaque croisement i,j,k associé aux listes (A,B,C) - chacun de ces trois indices prenant les valeurs exclusivement 0 (pas d'appartenance à la liste) ou 1 (appartenance à la liste). On mobilise toujours deux hypothèses jusqu'ici jamais remises en cause, qui sont *primo* l'indépendance mutuelle de comportement entre les individus de la population, c'est-à-dire que le comportement d'un individu quelconque (son appartenance ou non à une liste…) n'a aucune relation avec celui de tout autre individu, et *secundo* l'homogénéité des probabilités combinées $p_{K,ijk}$ donnant la probabilité que l'individu K se trouve en situation ijk . Cette dernière hypothèse fait que l'on traite des p_{ijk} et non des $p_{K,ijk}$.

Ces conditions sont exactement celles d'une distribution multinomiale des x_{ijk} (tirages indépendants de N individus – c'est-à-dire avec remise - selon une distribution de probabilités donnée). Soit

$$\{x_{ijk}\}_{ijk} \rightarrow Multinomiale(N ; \{p_{ijk}\}_{ijk}) .$$

Chaque composante x_{ijk} suit une loi binomiale $x_{ijk} \rightarrow Binomiale(N ; p_{ijk})$ et on obtient

$$Ex_{ijk} = N . p_{ijk} \qquad Vx_{ijk} = N . p_{ijk} . (1 - p_{ijk}) \qquad Cov(x_{ijk}, x_{i'j'k'}) = -N . p_{ijk} . p_{i'j'k'} .$$

On peut aussi traiter les x_{ij} comme les réalisations de variables aléatoires suivant une loi de Poisson, que l'on peut considérer comme une limite de loi binomiale. Les lois de Poisson sont plus souvent utilisées que les lois multinomiales car leur support est l'ensemble $\mathbb{N}$, ainsi il n'y a *a priori* pas à considérer qu'il y a une quelconque limite aux tailles des sous-populations en jeu. On considère par ailleurs que les variables aléatoires sont mutuellement indépendantes (ce qui n'était pas vrai dans l'approche multinomiale) mais les paramètres de leurs lois respectives vont être reliés par un modèle. Ainsi, la densité associée à l'ensemble de l'information collectée est toujours un produit de densités élémentaires de Poisson, avec des paramètres qui, par contre, vérifient certaines relations qui les lient les uns aux autres.

$$\forall (i, j, k) \qquad x_{ijk} \rightarrow Poisson(N . p_{ijk})$$

$$Ex_{ijk} = N . p_{ijk} \qquad Vx_{ijk} = N . p_{ijk} \qquad Cov(x_{ijk}, x_{i'j'k'}) = 0 .$$

Une modélisation linéaire 'standard' consiste, sur le principe, à paramétrer sous forme linéaire l'espérance des variables observée, ici Ex_{ijk} , mais une modélisation GLM généralise (comme son nom l'indique) cette approche en modélisant de façon linéaire une *fonction* de l'espérance Ex_{ijk} . La fonction est au choix du statisticien, *a priori* simple et ayant de bonnes propriétés mathématiques, en particulier la dérivabilité C_∞ .

Le modèle log linéaire est le cas spécifique où cette fonction est le logarithme. Il s'agit donc d'imposer une formalisation paramétrée de type linéaire à $\log(Ex_{ijk}) = \log(N) + \log(p_{ijk})$ - c'est-à-dire aussi à $\log(p_{ijk})$.

Il est naturel de décomposer $\log(p_{ijk})$ en une somme d'effets propres aux différentes sources et d'effets croisés – comme cela se fait dans les modèles linéaires d'analyse de variance (ANAVAR), soit :

$$\log p_{ijk} = \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}$$

Ici, λ_i^A sera l'effet causé par la modalité i associée à la source A, λ_{ij}^{AB} sera un effet à ajouter lorsque les modalités i de la source A et j de la source B seront conjointement satisfaites (et idem pour les autres termes de même nature), et enfin λ_{ijk}^{ABC} traduit un nouveau supplément algébrique caractérisant le croisement des modalités i, j, k de A,B,C. La forme de cette décomposition est clairement linéaire selon les différents paramètres *lambda*.

On rajoute un paramètre réel λ (se substituant à $\log(N)$) quand on considère la variable $\log(Ex_{ijk})$:

$$\log Ex_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC} .$$

Si on adopte une approche vectorielle, les vecteurs auxquels s'appliquent ces paramètres sont composés de 0 et de 1 uniquement. L'écriture en l'état est évidemment très largement surparamétrée : le vecteur à gauche de l'égalité – celui des $\log(Ex_{ijk})$ - a 8 composantes alors qu'à droite de l'égalité on a le vecteur constitué uniquement de 1, plus 6 vecteurs associés à une unique source, 12 vecteurs croisant deux sources et 8 vecteurs pour traduire les effets triples. Tous les vecteurs considérés sont définis dans $\mathbb{R}^8$.

Pour former le modèle définitif, il faut distinguer deux étapes :

- *ETAPE 1* (identifiabilité) : révision du paramétrage pour supprimer le surparamétrage. C'est une question de dimension : il faut réduire le nombre de paramètres dans l'expression linéaire, concrètement forcer à zéro un certain nombre de paramètres. L'annexe 2 fournit un éclairage sur cette phase. Avec 3 sources, on vérifie qu'il faut imposer très exactement 19 contraintes aux paramètres impliqués dans la décomposition des $\log Ex_{ijk}$ selon les effets introduits ci-dessus. A la fin de cette étape, on a ré écrit les $\log Ex_{ijk}$ en fonction d'un jeu de paramètres *lambda*. Mais aucune hypothèse n'a été faite à ce stade, il s'agit seulement d'une manière alternative (et parcimonieuse en termes de nombre de paramètres impliqués) d'écrire les $\log Ex_{ijk}$ (ou les $\log p_{ijk}$) sans valeur ajoutée particulière, si ce n'est en termes d'interprétation.

- *ETAPE 2 (*modélisation) : nouvelle réduction du nombre de paramètres, au-delà de la simple ré écriture. En pratique, il s'agit d'annuler certains paramètres parmi ceux qui ont été conservés à l'issue de l'étape 1. C'est en effectuant ce type d'opération que l'on commence à contraindre les p_{ijk} à respecter certaines propriétés.

Revenons sur chacune de ces étapes plus en détail.

L'étape 1 permet de produire un paramétrage «non-informatif». Une façon simple de le faire consiste à annuler tous les paramètres pour lesquels l'un des indices est maximum (voir annexe 2), ou de manière alternative minimum. Le domaine de capture-recapture offre un contexte original, car les modalités associées à chaque liste (en tant que variable qualitative) ne prennent que 2 valeurs : 1 pour l'appartenance à la liste et 0 pour la non-appartenance à la liste. Ainsi, le principe de formation du paramétrage initial est simple : on annule tous les paramètres pour lesquels au moins un indice prend sa valeur maximale, donc 1. Cela revient à dire que l'on conserve uniquement les paramètres où tous les indices valent 0, qui sont aussi ceux qui traduisent la non-appartenance à chacun des croisements pris en compte dans le jeu complet de paramètres. Pour un modèle avec 3 sources, sont *a priori* non nuls les paramètres $\lambda, \lambda_0^A, \lambda_0^B, \lambda_0^C, \lambda_{00}^{AB}, \lambda_{00}^{AC}, \lambda_{00}^{BC}, \lambda_{000}^{ABC}$. Il est possible aussi, si c'est plus pratique, d'annuler les paramètres où l'un des indices est minimum – en la circonstance il s'agit de la valeur zéro : la stratégie serait alors inverse de celle qui a été retenue puisqu'on ne conserverait que les paramètres où tous les indices sont égaux à 1 pour chaque croisement de sources.

Dans ces circonstances, l'écriture paramétrée la plus générale des $\log(Ex_{ijk})$ est donc :

$$\log Ex_{111} = \lambda$$
$$\log Ex_{011} = \lambda + \lambda_0^A$$
$$\log Ex_{101} = \lambda + \lambda_0^B$$
$$\log Ex_{001} = \lambda + \lambda_0^A + \lambda_0^B + \lambda_{00}^{AB}$$
$$\log Ex_{110} = \lambda + \lambda_0^C$$
$$\log Ex_{010} = \lambda + \lambda_0^A + \lambda_0^C + \lambda_{00}^{AC}$$
$$\log Ex_{100} = \lambda + \lambda_0^B + \lambda_0^C + \lambda_{00}^{BC}$$
$$\log Ex_{000} = \lambda + \lambda_0^A + \lambda_0^B + \lambda_{00}^{AB} + \lambda_0^C + \lambda_{00}^{AC} + \lambda_{00}^{BC} + \lambda_{000}^{ABC}$$

Cette écriture se décrit par une bijection entre les $\log(Ex_{ijk})$ et la famille des paramètres *lambda* conservés. À ce stade, une éventuelle opération d'estimation des paramètres *lambda* n'aurait aucun intérêt : l'ajustement serait parfait, puisque basé sur autant d'observations que de données (8 de chaque – en considérant que l'information x_{000} est connue) !

Partant de là, on enchaîne avec l'étape 2 – qui est totalement ouverte. L'idée générale consiste à réduire le nombre de paramètres impliqués – donc à construire un modèle – afin d'estimer chacun des paramètres conservés puis d'en tirer une prédiction de x_{000}. La phase ultime de l'opération est bel et bien de prédire x_{000} par une valeur $\hat{x}_{000}$ car c'est là *le seul dénombrement qui n'est pas observé*. On achève l'estimation de N par $\hat{N} = \hat{x}_{000} + x_{100} + x_{010} + x_{110} + x_{001} + x_{101} + x_{011} + x_{111}$.

La réduction de la dimension du problème est la botte secrète qui permet de contourner le blocage causé par la non-connaissance de x_{000} : lorsque 8 effectifs sont reliés à 8 paramètres mais qu'on ne dispose que de 7 informations (effectifs observés), on ne peut pas aboutir. Par contre, si les 8 effectifs sont reliés à 7 paramètres, puisqu'on dispose de 7 informations alors on va pouvoir estimer les 7 paramètres et prédire en retour l'effectif inconnu.
Le contexte particulier de capture-recapture croise des sources en nombre (généralement) limité, et pour chaque source on ne distingue que 2 modalités (appartenance ou non), ce qui fait que le nombre de données x_{ijk} est par nature (généralement) petit : les modèles sont ajustés sur 4 -1 = 3 données quand on a 2 sources, 8 -1 = 7 données quand on a en 3, 16 – 1 = 15 données avec 4 sources… il faut 7 sources pour dépasser les 100 données.

La pratique classique des modèles log linéaires s'applique dans un contexte où on connaît tous les effectifs croisés x_{ijk}, et où l'objectif est de décrire la façon dont un ensemble de variables qualitatives influent sur la structure d'une population. Dans le cas présent, le contexte et l'objectif changent : l'un des effectifs n'est pas observé (un seul dans ce cas certes - mais ça suffit pour changer la problématique) et l'objectif premier devient précisément d'estimer cet effectif manquant… en exploitant les hypothèses d'association entre variables que traduisent le modèle (ou de structure de probabilités d'appartenance combinée aux sources, ce qui revient au même). Une autre caractéristique de cette modélisation log linéaire, mais elle n'est cette fois en rien spécifique au contexte de capture recapture, c'est qu'il n'y a pas vraiment de 'variables explicatives' à opposer à des 'variables expliquées', ou plus exactement il n'y a pas de différence entre ces deux concepts : toutes les variables qualitatives sont à considérer au même niveau, et soit on les perçoit

comme expliquées parce qu'elles définissent les observations x_{ijk}, soit on les considère comme explicatives parce que c'est la façon dont les x_{ijk} en dépendent qui est intéressante.

Pour définir le modèle, tout type de contrainte est possible – on a vu qu'il ne fallait pas utiliser plus de 7 paramètres mais c'est la seule vraie contrainte. Cela étant, il paraît hautement souhaitable de pouvoir interpréter les contraintes imposées et cet aspect est difficile lorsqu'on imagine des contraintes compliquées.

Une position très simple mais extrême – que l'on ne rencontrera pas en pratique – consiste à annuler tous les paramètres à l'exception de la constante λ : c'est le modèle dans lequel les p_{ijk} sont constants (égaux de fait à 1/8). C'est évidemment déraisonnable.

En revanche, il est très fréquent que l'on fasse les hypothèses suivantes :

$$\lambda_{00}^{AB} = \lambda_{00}^{AC} = \lambda_{00}^{BC} = \lambda_{000}^{ABC} = 0 .$$

On conserve alors 4 paramètres, valeur qui est largement en dessous du seuil critique. Le modèle obtenu sous cette hypothèse occupe une place centrale en pratique puisqu'il correspond très exactement à la situation d'indépendance entre les sources – donc à l'hypothèse que l'on a fait jusqu'à présent pour justifier l'estimateur de Lincoln-Petersen. L'explication est simple : en situation d'indépendance, par définition on a $p_{ijk} = p_{i++} \cdot p_{+j+} \cdot p_{++k}$, $\forall (i,j,k)$, soit $\log p_{ijk} = \log p_{i++} + \log p_{+j+} + \log p_{++k}$. C'est une relation de type $\log p_{ijk} = \lambda_i^A + \lambda_j^B + \lambda_k^C$, qui cadre parfaitement avec le modèle $\log Ex_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C$.

A contrario, les paramètres non nuls $\lambda_{00}^{AB}, \lambda_{00}^{AC}, \lambda_{00}^{BC}$ et λ_{000}^{ABC} traduisent l'écart à l'indépendance. *Laisser dans la modélisation l'un – au moins – de ces paramètres est donc équivalent à abandonner l'une des hypothèses fondamentales justifiant l'estimateur de Lincoln-Petersen.*

Les effets croisés sont donc à introduire quand on pense que l'hypothèse d'indépendance n'est pas vérifiée. Il y a toujours un « acte de foi » à l'origine d'un tel choix, car contrairement au cas des tables de contingence complètes, il n'est pas possible de tester cette hypothèse à cause de la valeur manquante x_{000}. Bien qu'il n'y ait pas de fatalité, on peut considérer qu'on impose en pratique

$$\lambda_{000}^{ABC} = 0 ,$$

parce que c'est peut-être la contrainte la plus naturelle à imposer pour sortir du contexte de simple reparamétrage des $\log(Ex_{ijk})$.

De plus, il semble que λ_{000}^{ABC} est le paramètre qui a l'interprétation la plus complexe. Pour la construction du modèle, la question de l'interprétation des paramètres a son importance, d'autant plus qu'on ne peut pas tester la nullité des paramètres que l'on a choisi d'annuler par nous-même en amont pour rendre l'estimation possible. Lorsqu'on dispose de 3 sources (ou davantage) , on peut interpréter plus facilement les effets doubles $\lambda_{ij}^{AB}, \lambda_{ik}^{AC}$ et λ_{jk}^{BC}. Pour cela, il faut revenir aux *odds ratios* (voir encadré en I.3), qui non seulement permettent de caractériser l'indépendance éventuelle entre variables qualitatives, mais encore quantifient l'intensité de la dépendance lorsqu'elle existe.

Considérons 2 sources quelconques parmi les 3, par exemple A et B. Le *odds ratio* associé à la table de contingence croisant A et B a une expression formelle très compliquée dont on ne tirera rien, en revanche si on conditionne par rapport à une modalité quelconque k de la troisième source C, le *odds ratio* de la table de contingence conditionnelle, noté ici $\theta_{(k)}^{AB}$ vérifie

$$\log \theta_{(k)}^{AB} = \log\left(\frac{Ex_{11k}.Ex_{00k}}{Ex_{10k}.Ex_{01k}}\right) = \lambda_{11}^{AB} + \lambda_{00}^{AB} + \lambda_{10}^{AB} + \lambda_{01}^{AB} .$$

Compte tenu des contraintes identifiantes adoptées, on aboutit à $\lambda_{00}^{AB} = \log \theta_{(k)}^{AB}$, et cela est vrai pour tout k.

On en tire deux conséquences.

D'une part, une interprétation facile : on peut immédiatement conclure que s'il y a une quelconque dépendance entre A et B, soit dans la sous-population des individus présents dans la liste C (donc pour $k=1$), soit dans la sous-population des individus absents de la liste C (pour $k=0$), alors λ_{00}^{AB} est non nul et il faut l'inclure dans le modèle. Les effets croisés non nuls sont bien caractéristiques de relations de dépendance entre les couples de listes.

D'autre part, une méthode d'estimation : adoptant une approche de type estimation par les moments, on va naturellement estimer λ_{00}^{AB} en posant :

$$\hat{\lambda}_{00}^{AB} = \log\left(\frac{x_{111}.x_{001}}{x_{101}.x_{011}}\right) .$$

Par ailleurs, de manière immédiate et dans le même esprit : $\log Ex_{111} = \lambda$ entraîne $\hat{\lambda} = \log x_{111}$, puis $\hat{\lambda}_0^A = \log x_{011} - \lambda = \log \frac{x_{011}}{x_{111}}$. On obtient ainsi de suite l'ensemble des estimations des *lambda,* puis finalement, en utilisant la 8ème équation du système complet paramétré :

$$\hat{x}_{000} = \frac{x_{111}.x_{100}.x_{010}.x_{001}}{x_{110}.x_{101}.x_{011}} .$$

Il n'est d'ailleurs pas nécessaire d'enchaîner toutes ces opérations d'estimation : l'estimation $\hat{x}_{000}$ s'obtient immédiatement en exploitant l'égalité établie ci-dessus : $\log \theta_{(0)}^{AB} = \log \theta_{(1)}^{AB}$, qui vaut aussi λ_{00}^{AB} et qui se traduit par :

$$\log\left(\frac{Ex_{110}.Ex_{000}}{Ex_{100}.Ex_{010}}\right) = \log\left(\frac{Ex_{111}.Ex_{001}}{Ex_{101}.Ex_{011}}\right).$$

Nous faisons remarquer que le processus d'estimation repose sur des modèles « expliquant » l'effectif attendu Ex_{ijk} et non les probabilités p_{ijk}, malgré la très grande proximité des deux concepts. La décomposition de la probabilité ne sert qu'une fois – mais c'est une étape essentielle - pour justifier l'indépendance entre les sources si (et seulement si, de fait) il n'y a aucun effet croisé dans le modèle.

Afin de simplifier éventuellement le modèle, on pourrait mettre en place des tests de nullité des coefficients *lambda* que l'on a choisi de conserver, puisqu'on dispose d'estimateurs de ces coefficients.

Comme $\log(Ex_{ijk}) = \log(N) + \log(p_{ijk})$, on peut, si cela présente un intérêt, obtenir un estimateur des probabilités élémentaires par cellule selon

$$\hat{p}_{ijk} = \frac{\hat{x}_{ijk}}{\hat{N}}.$$

Pour tout triplet $(i,j,k) \neq (0,0,0)$, on choisira naturellement $\hat{x}_{ijk} = x_{ijk}$ et on adoptera par ailleurs l'expression $\hat{x}_{000}$ obtenue ci-dessus. En formant $\hat{N} = \hat{x}_{000} + x_{100} + x_{010} + x_{110} + x_{001} + x_{101} + x_{011} + x_{111}$, on assure que la somme des $\hat{p}_{ijk}$ vaut bien 1.

Voici une interprétation intéressante de ce résultat.

Définissons le coefficient

$$\rho = \frac{p_{111} \cdot p_{100} \cdot p_{010} \cdot p_{001}}{p_{000} \cdot p_{110} \cdot p_{101} \cdot p_{011}}.$$

Partant de l'expression

$$\log p_{ijk} = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}$$

et une fois l'identification des paramètres assurée en annulant tous ceux dont au moins un des indices vaut 1 (par exemple), on vérifie qu'en toute généralité on obtient[21] $\lambda_{000}^{ABC} = -\log\rho$: c'est le terme dit « d'interaction du second ordre » - les $\lambda_{ij}^{AB}, \lambda_{ik}^{AC}$ et λ_{jk}^{BC} étant les « interactions du premier ordre ».

L'expression de $\hat{x}_{000}$ obtenue ci-dessus découle de l'hypothèse $\lambda_{000}^{ABC} = 0$. On est par conséquent dans la cadre d'un modèle où il n'y a pas d'interaction du second ordre. Imaginons que l'on constitue 2 tables de contingence 2 X 2, l'une (table 1) à partir de l'échantillon vérifiant $k = 1$ - donc tous les individus présents

21 Si on choisit d'annuler les paramètres avec un indice égal à 0, on obtient $\lambda_{111}^{ABC} = \log\rho$, ce qui est un peu plus joli...

dans la source C, l'autre (table 2) à partir de l'échantillon vérifiant $k=0$ - donc tous les individus absents de la source C. Les 2 tables se présentent comme suit, avec les notations traditionnelles :

Table1

x_{111}	x_{101}
x_{011}	x_{001}

Table 2

x_{110}	x_{100}
x_{010}	x_{000}

La table 1 présente un degré de corrélation mesurée par le *odds ratio* $\theta_1=\frac{x_{111}.x_{001}}{x_{101}.x_{011}}$. On peut formuler l'hypothèse que l'intensité de cette corrélation est la même dans la table 2, c'est-à-dire que le *odds ratio* de la table 2, $\theta_2=\frac{x_{110}.x_{000}}{x_{100}.x_{010}}$ est tel que $\theta_1=\theta_2$. Alors $\frac{x_{111}.x_{001}}{x_{101}.x_{011}}=\frac{x_{110}.x_{000}}{x_{100}.x_{010}}$ et on retrouve l'expression de $\hat{x}_{000}$.

Noter qu'on ne peut hélas pas tester l'hypothèse $\rho=1$ car on ne dispose pas de la valeur x_{000}.

Cela donne une interprétation simple de l'estimateur $\hat{x}_{000}$: on l'obtient lorsque l'appartenance – ou non - à la source C n'impacte en rien la structure de corrélation entre les tables A et B (on peut – évidemment - présenter cette propriété de manière équivalente en considérant toute permutation des sources A, B et C). C'est la situation d'indépendance entre une source et les autres, qui se traduit donc par l'absence d'interaction du second ordre.

Accessoirement, on pourra retenir de cette approche que lorsqu'on est en présence de 3 sources, le degré de dépendance de 2 sources parmi 3 conditionnellement à l'appartenance (ou non) à la troisième est traduit par le coefficient ρ défini ci-dessus et que l'égalité $\rho=1$ correspond à l'absence d'interaction du second ordre.

Cette méthodologie d'estimation utilise le cadre, les concepts et le vocabulaire des modèles log linéaires que l'on utilise dans nombre d'opérations statistiques où l'objectif est d'étudier les liaisons entre des variables qualitatives. Néanmoins, le rapprochement entre ce qui est fait ici et ce qu'on trouve dans l'approche traditionnelle trouve rapidement ses limites, car si on rencontre en amont les mêmes problèmes d'identifiabilité, le contexte, les outils et surtout l'objectif de la théorie traditionnelle diffèrent clairement du nôtre. En particulier, on ne cherche pas à estimer des effectifs pour chaque cellule dans le cadre du modèle adopté, on ne le fait que pour la cellule non observée, c'est-à-dire x_{000}. L'estimation ici pratiquée n'est finalement que la résultante d'un système d'équations simples partant d'une hypothèse invérifiable et permettant d'aboutir mathématiquement par une technique en cascade à une estimation ponctuelle. Elle n'a

pas besoin d'utiliser les techniques de maximum de vraisemblance de la théorie classique. En particulier on ne parle pas de qualité de l'ajustement du modèle. Tout cela tient au fait que l'objectif n'est pas de résumer une information complexe en réduisant la dimension d'un espace de paramètres mais à prédire une unique valeur inconnue en exploitant une structure de dépendances que l'on a postulée. D'une certaine façon, on pourrait aussi dire que c'est une approche plus descriptive qu'inférentielle.

Modèles log-linéaires : étude de robustesse à l'hypothèse d'indépendance, dans le cas de deux sources

Considérons deux sources A et B. La séquence d'appartenance (ou non) respectivement à A et à B est traduite par un couple pouvant prendre quatre valeurs : 11, 10, 01 ou 00. Soit x_{ij} le nombre d'individus de la population totale dans la situation (i,j) et p_{ij} la probabilité pour un individu de vérifier la séquence (i,j). On a $N = x_{11} + x_{10} + x_{01} + x_{00}$.

Le modèle complet (surparamétré) portant sur les effectifs attendus Ex_{ij} s'écrit :

$$\log Ex_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}$$

Compte tenu des lois possibles de x_{ij} (multinomiale ou Poisson), $Ex_{ij} = N \,.\, p_{ij}$ dans tous les cas.
Les questions d'identification (voir annexe 1) amènent à imposer au total $I + J + 1 = 5$ contraintes. Si on procède comme dans le cas de 3 sources, on impose : $\lambda_1^A = \lambda_1^B = 0$ ainsi que $\lambda_{11}^{AB} = \lambda_{10}^{AB} = \lambda_{01}^{AB} = 0$

L'indépendance entre sources se traduit par les égalités $p_{ij} = p_{i\cdot}\, p_{j}$ soit $\log Ex_{ij} = \log N + \log p_{i+} + \log p_{+j}$. Le cas de l'indépendance conduit donc à annuler les interactions λ_{ij}^{AB} et à poser

$$\log Ex_{ij} = \lambda + \lambda_i^A + \lambda_j^B$$

avec toujours $\lambda_1^A = \lambda_1^B = 0$.

On dispose de 3 valeurs observées x_{11}, x_{10}, x_{01} et le modèle implique 3 paramètres $\lambda, \lambda_0^A, \lambda_0^B$ pour « expliquer » 4 valeurs : $x_{11}, x_{10}, x_{01}, x_{00}$ - en réalité 3 puisque x_{00} est inconnu. Il y a donc moyen d'aboutir. C'est au demeurant très simple avec 2 sources : comme $Ex_{00} = \exp(\lambda + \lambda_0^A + \lambda_0^B)$, on choisira $\hat{x}_{00} = \exp(\hat{\lambda} + \hat{\lambda}_0^A + \hat{\lambda}_0^B)$. Un système de 3 équations linéaires à 3 inconnues ne pose aucun problème, on résout et on en tire – sans surprise - l'estimation de Lincoln-Petersen $\hat{x}_{00} = \left[\frac{x_{10} \cdot x_{01}}{x_{11}}\right]$.

Partant de la vraisemblance du vecteur $x_{11}, x_{10}, x_{01}, x_{00}$ (cas multinomial) ou du vecteur x_{11}, x_{10}, x_{01} (loi de Poisson) et considérant N comme un paramètre, on peut vérifier facilement que l'utilisation de la méthode du maximum de vraisemblance (comme alternative à la méthode de type moment ici mise en œuvre) donne exactement les mêmes estimations (mais les calculs sont plus compliqués).

L'indépendance peut ne pas être acceptable au niveau de l'ensemble de la population mais être plus raisonnablement envisageable sur des sous-populations partitionnant la population totale. Par exemple il peut y avoir indépendance pour les hommes, indépendance pour les femmes, mais pas indépendance sans distinction de genre. Numériquement, le *odds ratio* pour les hommes peut être égal à 1, celui pour les femmes également, sans que le *odds ratio* global le soit (facile à vérifier). Cela parce que l'indépendance conditionnelle n'entraîne pas l'indépendance globale : $P(AB|Z=z)=P(A|Z=z).P(B|Z=z)$ pour chaque genre z n'entraîne pas $P(AB)=P(A).P(B)$. En vertu de cette remarque de fond, on peut envisager d'utiliser des modèles log linéaires impliquant des variables tenant lieu de variables explicatives. Ainsi, en introduisant une variable qualitative Z à 2 modalités notées respectivement 0 et 1 :

$$\log Ex_{ijz} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_z^Z + \lambda_{iz}^{AZ} + \lambda_{jz}^{BZ} + \lambda_{ij}^{AB} + \lambda_{ijz}^{ABZ}$$

avec les contraintes identifiantes suivantes : les λ_s - où s est une séquence d'indices - sont nuls lorsque l'un des indices composant s est nul (on ne conserve donc que les paramètres où tous les indices valent 1). Ceci est le modèle dit 'saturé', qui ne traduit rien d'autre qu'une reparamétrisation des Ex_{ijz} ; l'indépendance conditionnelle à $Z=z$ entre A et B s'obtient en posant[22] $\lambda_{ij}^{AB} = \lambda_{ijz}^{ABZ} = 0$ pour tout (i,j,z). On peut effectuer une étude de robustesse à l'hypothèse d'indépendance en considérant des *odds ratios* différents de 1 et en regardant comment varient les deux effectifs non observés que sont $\hat{x}_{000} = \exp(\hat{\lambda})$ et $\hat{x}_{001} = \exp(\hat{\lambda} + \hat{\lambda}_1^Z)$, dont on déduit immédiatement $\hat{N}$. Techniquement, on y parvient en considérant λ_{ij}^{AB} et λ_{ijz}^{ABZ} comme fixes, au sens 'non estimés' (on appelle de telles grandeurs des '*offsets*') et on envisage différentes valeurs $\tilde{\lambda}_{ij}^{AB}$ et $\tilde{\lambda}_{ijz}^{ABZ}$, ce qui permet de fixer directement les *odds ratios*

$$\theta_z = \frac{Ex_{11z}.Ex_{00z}}{Ex_{10z}.Ex_{01z}} = \exp(\tilde{\lambda}_{00}^{AB} + \tilde{\lambda}_{00z}^{ABZ}).$$

Pour une étude plus simple, on peut se limiter à faire varier un seul paramètre θ en imposant une égalité entre les différents θ_z : $\theta = \theta_z = \exp(\tilde{\lambda}_{00}^{AB})$ pour tout θ considéré, soit $\tilde{\lambda}_{00z}^{ABZ} = 0$, ce qui revient à décréter que la dépendance est la même au sein des différentes sous-populations associées aux différentes modalités de Z. On rappelle que tous les autres $\tilde{\lambda}_{ijz}^{ABZ}$ sont nuls pour des raisons d'identification.

Pour récapituler la démarche permettant de mener l'étude de robustesse :
- on impose $\tilde{\lambda}_{ijz}^{ABZ} = 0$ pour tout (i,j,z), on se fixe $\tilde{\lambda}_{00}^{AB}$ et on en déduit θ ;

22 Il faut et il suffit que les probabilités p_{ij} soient exprimables sous forme du produit d'un terme en i et d'un terme en j, ce qui impose d'annuler à la fois les λ_{ij}^{AB} et les λ_{ijz}^{ABZ}.

- on estime les différents λ_s qui ne sont pas rendus nuls pour des raisons d'identification (méthode des moments ou méthode de maximum de vraisemblance) ;
- on en déduit $\hat{x}_{000}$, $\hat{x}_{001}$, et finalement $\hat{N}$ en tant que fonction de θ ;
- on fait varier θ (au travers de $\widetilde{\lambda}_{00}^{AB}$), sans manquer évidemment le cas central où $\widetilde{\lambda}_{00}^{AB}=0$ (soit $\theta=1$) qui correspond à la situation d'indépendance conditionnelle ;
- on regarde comment $\hat{N}$, en tant que fonction de θ , est sensible à la valeur de θ . Une forte sensibilité (qui ne peut s'apprécier que subjectivement il faut l'avouer…) témoigne d'un manque de robustesse à l'hypothèse d'indépendance, c'est-dire que si on postule une indépendance entre sources alors qu'en réalité ce n'est pas vrai, on prend un grand risque de produire une estimation éloignée de la vraie valeur (c'est un peu l'équivalent d'une variance d'estimateur dans un sondage). Une conjecture citée dans la littérature : si les x_{11z} sont faibles devant les x_{10z}, x_{01z} , c'est-à-dire si les sources ont un faible recouvrement (corrélation négative), alors il y aura manque de robustesse.

La variable Z est supposée présente dans les deux sources. Au prix d'une certaine complication technique, on sait traiter les cas où les informations auxiliaires permettant l'indépendance conditionnelle ne sont pas systématiquement observées dans les 2 sources (cas de variables dites 'partiellement observées').

2. Relâchement de l'hypothèse d'homogénéité des probabilités individuelles de capture

Dans le monde réel, l'hétérogénéité individuelle est probablement la règle dans la plupart des contextes de capture-recapture. En particulier dans le cas des fichiers administratifs, il est difficile de croire que tous les individus de la population aient la même probabilité d'appartenir à un fichier donné : au-delà des évènements aléatoires, la probabilité d'appartenance est *a priori* fonction de certaines caractéristiques de l'individu. Le processus de constitution des listes est également en cause, privilégiant éventuellement certaines sous-populations par rapport à d'autres. Par exemple, le processus fiscal GMBI (Gérer Mes Biens Immobiliers) sert à la collecte de la taxe d'habitation sur les résidences secondaires et vacantes. On peut alors imaginer que l'effort administratif sera moins important sur les résidences principales, cette pratique pouvantentraîner une hétérogénéité des probabilités de prise en compte des logements dans la source en fonction de leur catégorie.

Reprenant les notations du I,2, soit $p_{K,1+}$ (respectivement $p_{K,+1}$) la probabilité pour l'individu K d'être en liste A (respectivement en liste B). Si les probabilités de capture dépendent de l'individu K (hétérogénéité inter individus), c'est-à-dire s'il faut considérer des $p_{K,1+}$ et des $p_{K,+1}$ variant avec K , alors $\hat{N}$ (estimateur de Lincoln-Petersen) est biaisé, non convergent[23], et le terme principal du biais vaut $-N\dfrac{cov(p_{1+},p_{+1})}{cov(p_{1+},p_{+1})+\overline{p_{1+}}\cdot\overline{p_{+1}}}$ où cov désigne la vraie covariance entre les $p_{K,1+}$ et les $p_{K,+1}$, et $\overline{p_{1+}}$ et $\overline{p_{+1}}$ sont les vraies moyennes respectivement des $p_{K,1+}$ et des $p_{K,+1}$. On peut aussi écrire le biais selon

23 $\hat{N}$ ne converge pas en probabilité vers N quand le nombre d'observations $x_{00}+x_{01}+x_{10}$ devient très grand.

$$-N\left(1-\frac{\overline{p_{1+}}\ .\ \overline{p_{+1}}}{\overline{p_{1+}\ .p_{+1}}}\right)$$

en notant $\overline{p_{1+}\ .p_{+1}}$ la vraie moyenne des produits $p_{K,1+}\ .p_{K,+1}$.

Rigoureusement, le premier ordre du biais[24] s'annule si et seulement si les vecteurs des probabilités d'appartenance respective aux listes A et B sont indépendants. La configuration la plus radicale pour y parvenir (en fait la seule qui puisse correspondre à une situation concrète réaliste) est celle où l'une des probabilités est constante – voire les deux, comme dans les propriétés socle. *A priori*, les situations réelles traduisent des covariances positives : $cov\left(p_{1,+}\ p_{+,1}\right)>0$ si bien que souvent on sous-estime N. On réduit ce biais significativement en post-stratifiant afin de rendre au moins une des probabilités constante (ou, disons, peu variable). Il est intéressant de noter que la propriété d'inclusion constante suffisante pour annuler le biais ne concerne qu'une seule des listes, « au choix ».

En pratique on a souvent une source au moins qui prétend à l'exhaustivité. Dans ce cas, si c'est par exemple la première source, les $p_{K,1+}$ seront majoritairement très proches de 1 (ce qui n'empêche pas que ces probabilités soient éventuellement significativement éloignées de 1 sur des sous-populations d'effectif plutôt réduit) : on sera de fait dans la situation très favorable d'un jeu de probabilités 'presque constantes' et le biais sera numériquement (très) petit. On retrouve bien le contexte décrit en partie I.4.

L'hétérogénéité des probabilités individuelles va entraîner presque sûrement une dépendance entre les listes, c'est-à-dire que l'indépendance postulée au niveau individu est perdue lorsqu'on agrège les individus et qu'on se place au niveau de la population entière. Cela se perçoit facilement, car l'indépendance individuelle signifie que les égalités suivantes sont vérifiées (1 désigne toujours l'appartenance à une liste, 0 la non-appartenance) :

$$p\left(K\in 11\right)=p_{K,1+}\ .\ p_{K,+1}$$
$$p\left(K\in 10\right)=p_{K,1+}\ .\left(1\ -\ p_{K,+1}\right)$$
$$p\left(K\in 01\right)=\left(1\ -\ p_{K,1+}\right).p_{K,+1}$$
$$p\left(K\in 00\right)=\left(1\ -\ p_{K,1+}\right).\left(1\ -\ p_{K,+1}\right)$$

D'où $Ex_{11}=\sum_{K=1}^{N} p_{K,1+}\ .\ p_{K,+1}$,..., $Ex_{00}=\sum_{K=1}^{N}\left(1\ -\ p_{K,1+}\right).\left(1\ -\ p_{K,+1}\right)$.

La satisfaction des propriétés d'indépendance au niveau individuel ne préjuge en rien de l'ampleur de la corrélation entre les valeurs de ces probabilités attachées aux deux sources respectives. Il s'agit bien de deux phénomènes sans relation. Mais on perçoit au moins qualitativement que si les $p_{K,1+}$ et les $p_{K,+1}$ sont corrélés positivement, alors Ex_{11} et (symétriquement) Ex_{00} seront grands[25] : la diagonale de la

24 Tout cela n'est qu'une résultante d'approximations mathématiques… on ne dispose pas d'une expression donnant la valeur exacte du biais.

25 On se fixe les masses totales $\sum_{K=1}^{N} p_{K,1+}$ et $\sum_{K=1}^{N} p_{K,+1}$ et on identifie les individus par K selon les valeurs décroissantes de $p_{K,1+}+p_{K,+1}$. Si on enlève une valeur ϵ à $p_{K,1+}$ et à $p_{K,+1}$ de tout individu $K>1$ où ces probabilités sont toutes deux non nulles (et le restent), pour la reporter simultanément dans $p_{1,1+}$ et $p_{1,+1}$, alors on reste à masse constante et on vérifie

table de contingence sera chargée, et il y aura dépendance entre les variables caractérisant les appartenances aux deux listes respectives. C'est le même phénomène (appelé aussi paradoxe de Simpson) qui fait que l'agrégation de 2 tables de contingence indépendantes peut produire une table de contingence avec dépendance. On retrouve également ce type de mécanisme en traitant les effets de structure.

Des modèles ont été développés pour traduire ces situations – la classe des modèles concernés est appelée M_h (h pour *heterogeneity*) dans la littérature.

Une façon commune de procéder consiste à modéliser les probabilités de capture individuelles – car il faut nécessairement réduire la dimension du problème à quelques paramètres (laissés libres).

La perspective d'une modélisation explicite prenant en compte une hétérogénéité individuelle

Si on est en mesure d'ordonner les occasions de capture, on spécifie la probabilité $P_{K,s}$ de capturer l'individu K dans la source s au travers des

$$P\left(capturer\ \ K\ \ dans\ \ l'échantillon\ \ s\ \mid\ la\ \ composition\ \ des\ \ échantillons\ \ 1\ \ à\ \ s-1\right)$$

On va rencontrer ce type de contexte favorable quand par exemple on pose un piège chaque jour pendant un mois dans une forêt pour capturer des animaux, les échantillons étant repérés par le jour concerné. Il paraît beaucoup moins adapté au cas de présence/absence dans un ensemble de sources provenant d'horizons variés (parce qu'il n'y a pas d'ordre 'naturel' dans l'ensemble des sources). Néanmoins, cette approche retrouve tout son intérêt si les listes sont produites avec une certaine régularité au cours du temps, par exemple des fichiers trimestriels ou, mieux, mensuels. Partant de là, une option consiste à poser :

$P_{K,s}=p_K.e_s$ jusqu'à la première capture ;
$P_{K,s}=\phi.p_K.e_s$ lors de toutes les autres occasions de (re)captures.

On pourrait noter plus lourdement $P_{K,s|K\notin 1\cup 2\cup\ldots\cup(s-1)}=p_K.e_s$ et $P_{K,s}=P_{K,s|K\in 1\cup,2\cup\ldots\cup(s-1)}=\phi.p_K.e_s$.

Les e_s traduisent un effet 'occasion de capture' et les p_K traduisent l'hétérogénéité entre les individus (effet 'individu'), lors de toute capture : l'hypothèse H1 n'est donc plus vérifiée. Dans sa forme simplifiée, le paramètre ϕ (qualifiable d'« effet comportemental ») porte la dépendance entre les différentes occasions pour un individu donné : on s'en convainc facilement en comparant les probabilités de capture d'un individu quelconque à la seconde occasion, d'une part dès lors qu'il a été tiré à la première occasion, et d'autre part quand ce n'est pas le cas. Ainsi, $P_{K\in 2|K\notin 1}=p_K.e_2$ et $P_{K\in 2|K\in 1}=\phi.p_K.e_2$. Si ϕ diffère de 1,

très facilement qu'on augmente $\sum_{K=1}^{N} p_{K,1+}.p_{K,+1}$. Ainsi, avec cette méthode de fabrication d'une corrélation de plus en plus grande, Ex_{11} a tendance à augmenter – on pourra dire par extension que Ex_{11} augmente quand ses composantes $p_{K,1+}$ et $p_{K,+1}$ augmentent conjointement.

alors $P_{K\in 2|K\in 1} \neq P_{K\in 2|K\notin 1}$ et il y a de fait dépendance entre les deux occasions . L'hypothèse H3 n'est donc plus vérifiée.

En posant $\alpha_K = \log(p_K)$; $\mu_s = \log(e_s)$; $\gamma = \log(\phi)$ et $y_{K,s} = 1$ si i a été capturé avant l'occasion s et 0 sinon :

$$\log P_{K,s} = \alpha_K + \mu_s + \gamma . y_{K,s}$$

La présence des α_K traduit une hétérogénéité des probabilités de capture entre individus (mise en défaut de la composante H1 des hypothèses socle). La présence de γ traduit la dépendance entre sources du processus de capture (mise en défaut de la composante H3 des hypothèses socle). Pour qu'il y ait indépendance entre les différents échantillons constituant les différentes sources, il faut annuler γ .

On peut étendre le modèle en ajoutant des variables individuelles (*covariates*) :

$$\log P_{K,s} = \mu_s + \gamma Y_{K,s} + \beta^t . Z_K$$

où Z_K est un vecteur de caractéristiques individuelles qui vont expliciter le α_K .

Tous ces modèles peuvent être écrits en version *Logit*, de façon totalement analogue :

$$Logit(P_{K,s}) = \log \frac{P_{K,s}}{1-P_{K,s}} = \alpha_K + \mu_s + \gamma . Y_{K,s}$$

en partant des expressions :

$P_{K,s} = \dfrac{p_K . e_s}{1 + p_K . e_s}$ jusqu'à la première capture

$P_{K,s} = \phi . \dfrac{p_K . e_s}{1 + p_K . e_s}$ lors de toutes les autres (re)captures

Le cas spécifique $Logit(P_{K,s}) = \alpha_K + \mu_s$ correspondant à l'absence de dépendance entre sources ($\gamma = 0$) est connu sous le nom de « modèle de Rasch ». Il sera repris dans la partie suivante. Cela étant, à ce stade, on ne peut que constater le surparamétrage du modèle individuel en l'état : il y a un paramètre par individu, ce qui est beaucoup trop pour que l'approche soit opérationnelle.

Le modèle de Rasch

La complexité liée à l'hétérogénéité impose de fait que l'on fasse par ailleurs une hypothèse d'indépendance des captures successives au niveau individuel (et il s'agit bien du niveau individuel – cette précision est essentielle – on parle dans ce cas d'indépendance entre individus ou d'indépendance 'locale'). Restons

dans le cas où on dispose de 3 sources A, B et C. Soit $p_{K,ijk}$ la probabilité que l'individu K soit caractérisé par l'état ijk où i vaut 0 si K n'est pas capturé dans la source A et 1 s'il est capturé (notations identiques pour les sources B et C, impliquant respectivement les indices j et k). On impose donc

$$p_{K,ijk} = p_{K,i++} \cdot p_{K,+j+} \cdot p_{K,++k} \qquad \forall (i,j,k)$$

Supposons en outre que pour tout couple d'individus (K, K') on ait

$$\frac{\dfrac{p_{K',1++}}{p_{K',0++}}}{\dfrac{p_{K,1++}}{p_{K,0++}}} = \frac{\dfrac{p_{K',+1+}}{p_{K',+0+}}}{\dfrac{p_{K,+1+}}{p_{K,+0+}}} = \frac{\dfrac{p_{K',++1}}{p_{K',++0}}}{\dfrac{p_{K,++1}}{p_{K,++0}}}.$$

L'hypothèse traduit le fait qu'un *odds ratio* impliquant tout couple d'individus ne dépend pas de la source. Il y a donc une hétérogénéité des comportements individuels, mais elle n'est pas totalement quelconque, elle reste 'encadrée' par ces égalités, vérifiées pour chacun des C_N^2 couples que l'on peut former. On vérifie (facilement) que ces contraintes se traduisent sous une forme plus manipulable, qui est : $\forall K$

$$\log \frac{p_{K,1++}}{p_{K,0++}} = t(K) + \beta_A \text{ et } \log \frac{p_{K,+1+}}{p_{K,+0+}} = t(K) + \beta_B \text{ et } \log \frac{p_{K,++1}}{p_{K,++0}} = t(K) + \beta_C$$

Dit autrement : tout *Odds* impliquant un individu et une source est une fonction de la somme d'un effet individu et d'un effet source. A ce stade on ne précise pas de valeur pour $t(K)$, on peut simplement affirmer qu'il s'agit d'une valeur réelle qui varie avec l'individu concerné.

On introduit maintenant une notation plus simple : $p_{K,s}$ est la probabilité que l'individu K soit présent dans la source s (par exemple $p_{K,A} = p_{K,1++}$, $p_{K,B} = p_{K,+1+}$ …). Le jeu d'hypothèses (le 'modèle') se traduit par :

$$p_{K,s} = \frac{e^{t(K)+\beta_s}}{1+e^{t(K)+\beta_s}}.$$

Ce modèle est connu sous le nom de « modèle de Rasch ». Il est souvent présenté comme le modèle qui « assure une égalité des *odds ratios* comparant les distributions marginales des variables concernées » pour tout couple d'individus. Dans le cadre du rapprochement de deux sources, par exemple Résil et le RP, cette présentation savante pourrait se traduire ainsi en langage commun: «si un individu a (par exemple) 2 fois plus de chances[26] qu'un autre d'être recensé, alors il a aussi 2 fois plus de chances d'être dans Résil ». En pratique, on va trouver ce type de modèle quand on compare des individus ayant subi des évaluations indépendantes. Par exemple des élèves notés indépendamment par 2 professeurs : si l'élève Dupont a 3

26 Cette façon de parler est plus évocatrice, mais elle serait plus exacte si on considérait les rapports de probabilités « de succès » et non les rapports des *odds* – qui sont (un peu) plus compliqués à interpréter.

fois plus de chances que l'élève Durand d'avoir au moins 10/20 quand il est noté par le professeur 1, il a également 3 fois plus de chances d'avoir au moins la moyenne quand il est noté par le professeur 2.

Sous réserve que l'hétérogénéité des $p_{K,ijk}$ reste « modérée » (K varie et (i,j,k) est fixé), on peut considérer que les effectifs x_{ijk} croisant les statuts (i,j,k) suivent une loi multinomiale paramétrée par des p_{ijk} … qui ne dépendent évidemment plus de K puisqu'on se place à un niveau agrégé. L'approximation résulte d'un calcul formel assez lourd, mais elle est naturelle à partir du moment où la loi des x_{ijk} est effectivement multinomiale s'il y a parfaite homogénéité. On s'attend donc à ce que cette approximation reste valable si on ne s'éloigne « pas trop » de l'homogénéité.

Si on considère maintenant une combinaison quelconque de modalités (i,j,k), en conditionnant successivement par les N individus de la liste – chacun recevant la masse $\frac{1}{N}$, on construit ainsi la probabilité 'macro' p_{ijk} :

$$p_{ijk} = \sum_{K=1}^{N} p_{K,ijk} \cdot \frac{1}{N} = \frac{1}{N} \sum_{K=1}^{N} p_{K,i++} \cdot p_{K,+j+} \cdot p_{K,++k} \cdot$$

En utilisant la forme des $p_{K,s}$ du modèle de Rasch, il vient, après quelques lignes de calcul

$$p_{ijk} = \frac{1}{N} \exp\{i.\beta_a + j.\beta_b + k.\beta_c\} . \sum_{K=1}^{N} \exp(t(K).\omega) . p_{K,000}$$

en posant $\omega = i + j + k$. Ainsi, $\omega \in \{0,1,2,3\}$.

La probabilité $p_{K,000}$ varie (par définition) avec l'individu K . La masse affectée à K conditionnellement au fait que l'on est situé dans la sous-population qui n'est captée dans aucune des 3 sources, masse notée $p_{K|000}$, est proportionnelle à $p_{K,000}$, soit $p_{K|000} = \frac{p_{K,000} \cdot \frac{1}{N}}{p_{000}}$, soit encore

$p_{K|000} = \frac{1}{N.p_{000}} . p_{K,000} = \frac{p_{K,000}}{\sum_{h=1}^{N} p_{h,000}}$ (et on a bien $\sum_{K=1}^{N} p_{K|000} = 1$).

Cela conduit à considérer que p_{ijk} est proportionnel à

$\exp\{i.\beta_a + j.\beta_b + k.\beta_c\} . \sum_{K=1}^{N} \exp(t(K).\omega) . p_{K|000}$, d'où

$$\log p_{ijk} = \alpha + i.\beta_a + j.\beta_b + k.\beta_c + \log \sum_{K=1}^{N} \{\exp(t(K).\omega).p_{K|000}\} .$$

Le terme 'compliqué' est $\sum_{K=1}^{N} \exp(t(K).\omega).p_{K|000}$. Dans un espace probabilisé où les éventualités sont les individus formant la population d'intérêt, si on affecte la masse de probabilité $p_{K|000}$ à un individu K donné, ce terme n'est autre qu'une espérance mathématique. La variable aléatoire en jeu est $t(K)$, ou plus exactement $\exp(t(K).\omega)$, étant entendu que ω est un entier positif fixé et connu. La distribution des probabilités est de nature conditionnelle, puisqu'on se place conditionnellement à l'évènement 000, donc conditionnellement à une situation d'absence de capture dans les 3 sources considérées. On note T la variable aléatoire $t(K)$:

$$\sum_{K=1}^{N} \exp(t(K).\omega).p_{K|000} = E[\exp(T.\omega) \mid (i,j,k) = (0,0,0)] .$$

On peut considérer le log de cette expression comme une fonction de ω, notée

$$\gamma(\omega) = \log E[\exp(T.\omega) \mid (i,j,k) = (0,0,0)] .$$

Il est bienvenu d'avoir pu éliminer les effets individuels : ce sont les paramètres associés aux sources, qui traduisent des 'effets sources' qui sont intéressants et qu'il faut donc conserver – puis estimer. Si on résume, on a établi, à ce stade

$$\log p_{ijk} = \alpha + i.\beta_a + j.\beta_b + k.\beta_c + \gamma(\omega) .$$

On parle aussi de modèle de « quasi symétrie » parce que la fonction 'gamma' (qui n'est qu'une des composantes de la probabilité) ne change pas de valeurs quand on permute les indices de source i, j, k. L'expression de $\log p_{ijk}$ comprend une partie additive où chaque source contribue de manière propre et un terme complémentaire traduit par la fonction 'gamma'. La partie additive est caractéristique des situations d'indépendance entre sources (voir développements sur le modèle log linéaire). Le complément 'gamma' est donc motivé par une forme de dépendance entre les sources impliquées. Il remplace, en quelque sorte, les effets croisés $\lambda_{ij}^{AB}, \lambda_{ik}^{AC}, \lambda_{jk}^{BC}$ que l'on rencontre classiquement lorsqu'on utilise les modèles log-linéaires pour traiter les situations de dépendance entre sources. Au demeurant, en utilisant les notations du II.1, le modèle devient

$$\log p_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \gamma(i+j+k)$$

en imposant (par exemple) les contraintes identifiantes particulières $\lambda_0^A = \lambda_0^B = \lambda_0^C = 0$. La convention adoptée ici est contraire à celle de la partie III.1 parce que le modèle développé ci-dessus affecte un effet source lorsque l'individu est présent dans la source en question (par exemple $i.\beta_a \neq 0$ si et seulement si $i = 1$, ce qui incite à retenir) $\lambda_1^A \neq 0$ et donc $\lambda_0^A = 0$).

Par rapport à l'écriture traditionnelle des modèles caractérisant l'indépendance mutuelle des sources, on a rajouté 3 paramètres $\gamma(1)$, $\gamma(2)$ et $\gamma(3)$, étant entendu que par construction $\gamma(0)=\log 1=0$. A ce stade, l'expression de la fonction ' gamma' reste au choix. Si on veut un modèle offrant seulement une interaction d'ordre 1 et si on veut que les différentes interactions soient identiques quel que soit le couple de sources concernées, on pose $\gamma(i+j+k)=1 \Leftrightarrow i+j+k \geq 2$ (valeur 0 dans tout autre cas).

En théorie, on peut préciser la fonction 'gamma', qui dépend directement de la distribution de l'effet individuel $t(K)$ (variable aléatoire T) dans la sous population qui n'a jamais été captée par aucune source. Le modèle de Rasch est ici un modèle à effet 'individu' aléatoire[27] (mais effet 'source' fixe) et on doit faire une hypothèse portant sur la loi de T pour en tirer $\gamma(\omega)$. Une difficulté mathématique est due à l'existence d'une série (infinie) de contraintes portant sur les valeurs de $\gamma(\omega)$ pour différentes valeurs de ω. Ces contraintes reflètent celles qui affectent les moments d'une distribution quelconque. Par exemple, puisque la variance de $\exp(T)$ est positive, il en découle obligatoirement $\gamma(2) \geq 2.\gamma(1)$. De même $\gamma(3)+\gamma(1) \geq 2.\gamma(2)$, etc.

La technique d'estimation va mobiliser quatre éléments:

- le modèle (pseudo) log-linéaire introduit ci-dessus ;
- une hypothèse sur la distribution de T ;
- une hypothèse de nullité de l'interaction du second ordre ;
- l'estimateur du maximum de vraisemblance en approche conditionnelle.

La vraisemblance en approche conditionnelle est celle de la loi multinomiale conditionnellement à l'observation des effectifs. Cela signifie que parmi les 8 cases (i,j,k), on exclut la case $(0,0,0)$, qui est la seule pour laquelle il n'y a pas d'observation. Les probabilités considérées ne sont plus les $p_{K,ijk}$ mais les $p(K \in (i,j,k) | K \notin (0,0,0))$, égales à $\frac{p_{K,ijk}}{1-p_{K,000}}$. L'effectif total observé est $n=x_{001}+x_{010}+x_{001}+x_{110}+x_{101}+x_{01}+x_{111}$, si bien que la densité à maximiser est

$$\frac{n!}{\prod_{ijk \neq 000} x_{ijk}!} . \left[\frac{p_{ijk}}{1-p_{000}} \right]^{x_{ijk}} .$$

Les probabilités p_{ijk} et p_{000} sont fonctions des paramètres $\alpha, \beta_a, \beta_b, \beta_c$ et des paramètres intervenant dans la fonction 'gamma'. Noter qu'on parvient au même résultat si on fait une hypothèse de loi de Poisson portant sur les effectifs observés x_{ijk}.

27 Point technique : considérer les paramètres individuels 'initiaux' $t(K)$ comme des effets (= des variables) aléatoires est bien pratique à ce niveau. En effet, si on les traite comme des effets fixes, les estimateurs de maximum de vraisemblance des effets source $\beta_a, \beta_b, \beta_c$ perdent leur propriété (essentielle) de convergence, car il y a N paramètres, valeur tendant vers l'infini. Si on veut traiter les effets individuels comme fixes, c'est possible mais il faut accepter une acrobatie qui aboutit à ce que les estimateurs des β aient un statut d'estimateurs du maximum de vraisemblance conditionnels.

La fonction à maximiser est éminemment complexe. En toute généralité, la loi de T (conditionnelle) a pour densité $\phi(t)$, si bien que $\gamma(\omega)=\log\left(\int \exp(t.\omega).\phi(t)\,dt\right)$. Les paramètres impliqués sont les β plus ceux qui interviennent dans la fonction $\phi(t)$. La (log) densité à maximiser est proportionnelle à $\sum_{i,j,k} x_{ijk}.\log\frac{p_{ijk}}{1-p_{000}}$. Pour la résolution numérique, on approxime la valeur de l'intégrale.

Pour expliciter 'gamma', il faut faire une hypothèse portant sur la loi de T (conditionnelle à $(i,j,k)=(0,0,0)$), soit $\phi(t)$. Pour cela, une loi naturelle est la loi de Gauss $N(0,\sigma^2)$. Dans ce cas, on peut vérifier qu'il existe deux réels δ et γ tels que $\gamma(\omega)=\delta.\omega+\gamma.\omega^2=\delta.(i+j+k)+\gamma.(i+j+k)^2$, ce qui fait que le modèle pseudo log-linéaire se simplifie selon[28] :

$$\log p_{ijk}=\alpha+i.\beta_a+j.\beta_b+k.\beta_c+\gamma.(i+j+k)^2$$

où γ désigne désormais un paramètre réel inconnu.

En injectant les p_{ijk} paramétrés dans la fonction de densité, on maximise une fonction complexe, ce qui permet d'estimer les paramètres du modèle. A l'issue de la maximisation, si on utilise une densité gaussienne on obtient les estimateurs EMV (conditionnels) de $\alpha,\beta_a,\beta_b,\beta_c$ et γ, tous réels. On obtient ensuite les estimations d'effectifs par case $\hat{x}_{ijk}$ selon $\hat{x}_{ijk}=\hat{E}(x_{ijk})=n.\frac{\hat{p}_{ijk}}{1-\hat{p}_{000}}$ (espérance d'une loi binomiale). Mais, à ce stade, il nous manque toujours l'estimation essentielle $\hat{x}_{000}$.

Postuler une loi normale pour T revient à fixer à la valeur 1 le coefficient ρ introduit à la fin de la partie « *L'appel aux modèles Log-linéaires, en présence de 3 sources* ». En effet, on vérifie que sous le modèle précédent on a (en toute généralité)

$$\rho=\frac{e^{\gamma(3)}.e^{3.\gamma(1)}}{e^{\gamma(0)}.e^{3.\gamma(2)}}$$

et qu'avec l'expression de γ découlant de la loi normale, on aboutit à $\gamma(3)+3.\gamma(1)=\gamma(0)+3.\gamma(2)$, c'est-à-dire à $\rho=1$. Il n'y a donc pas d'interaction de second ordre (*rappel* : cela signifie que les deux *odds ratios* impliquant 2 sources conditionnellement à la troisième, sont égaux). Ce contexte particulier, ajoute une contrainte qui lie les 8 probabilités qui ont été distinguées, dont p_{000}, ce qui doit mécaniquement permettre d'estimer le paramètre manquant x_{000}. En se calquant sur la contrainte $\rho=1$, on peut donc produire un estimateur de x_{000} selon

28 Pour simplifier, on a repris la notation β_a pour désigner $\beta_a+\delta$ - idem pour β_b et β_c.

$$\hat{x}_{000}=\frac{\hat{x}_{111}.\hat{x}_{100}.\hat{x}_{010}.\hat{x}_{001}}{\hat{x}_{110}.\hat{x}_{101}.\hat{x}_{011}},$$

où les valeurs des $\hat{x}_{ijk}$ ne sont pas les valeurs observées mais les estimations du maximum de vraisemblance obtenues ci-dessus. On termine en formant $\hat{N}=n+\hat{x}_{000}$.

$\hat{N}$ dépend fondamentalement d'une hypothèse $\rho=1$ qu'on ne peut pas vérifier[29].

En toute généralité, les expressions de l'estimation $\hat{x}_{000}$ puis au final de $\hat{N}$ dépendent de la forme retenue pour la modélisation des $\log p_{ijk}$. La forme précédente est celle qui correspond à l'hypothèse $\rho=1$, qui elle-même découle de la loi T gaussienne formulée pour décrire la distribution des effets individuels du modèle. Le résultat est certes assuré avec une loi de Gauss, mais on l'obtient également si $\exp(T)$ suit une loi gamma de paramètres ν et η parce qu'alors, après calculs, on obtient $\exp(\gamma(\omega))=\Gamma(\omega+\nu)/\eta^{\omega}.\Gamma(\nu)$ et $\rho=\frac{\nu.(\nu+2)}{(\nu+1)^2}$, qui est proche de 1 si on opte pour un grand ν. L'approximation $\rho=1$ est encore valable dès que $\exp(T)$ suit une loi inverse 'gamma' ou une loi de Weibull. On peut retenir que si on fait confiance au modèle de Rasch, considérer que ρ est proche de 1 (pas d'interaction du second ordre) est assez plausible en situation courante, ce qui robustifie quelque peu l'expression $\hat{x}_{000}$ obtenue avec le modèle attaché à la loi de Gauss[30].

La pertinence du modèle de Rasch est intimement liée à la pertinence du modèle portant sur les $\log p_{ijk}$. Ce dernier fait intervenir une fonction 'gamma' opérant sur des fonctions symétriques (puisqu'il s'agit d'une fonction de $i+j+k$). On vérifie que ce modèle, de par cette composante symétrique (ou, de manière équivalente, l'égalité de 3 *odds ratios* par couple d'individus – voir supra), revient à formuler l'hypothèse suivante[31] :

$$p_{011}.p_{100} = p_{101}.p_{010} = p_{110}.p_{001}. \quad (*)$$

Cela donne l'opportunité de tester la pertinence du modèle de Rasch (contrairement au cas évoqué ci-dessus du test de l'absence d'interaction d'ordre 2). Déjà 'à vue' en comparant les produits des effectifs par case associés aux probabilités respectives, car en situation favorable les trois valeurs $x_{011}.x_{100}$, $x_{101}.x_{010}$ et $x_{110}.x_{001}$ devraient être numériquement proches. Ensuite de manière plus rigoureuse, car il existe des tests statistiques construits à partir de ratios de ces produits et permettant d'apprécier la significativité de la différence à la valeur 1. On rappelle que la méthodologie générale utilisée permet d'obtenir des estimateurs du maximum de vraisemblance de tous les paramètres impliqués dans le

29 Il faut toujours faire une hypothèse sur la forme de la densité $\phi(t)$, donc implicitement sur ρ : Si ρ n'est pas rendu formellement nul par cette hypothèse, on ne connaît pas la valeur x_{000} donc on ne peut pas tester $\rho=1$. On ne peut pas non plus tester en amont l'adéquation de la loi de l'effet individuel aléatoire à une loi donnée (Gauss par exemple) car on ne dispose d'aucune estimation de ces effets individuels.

30 Un cas notoire où le modèle ne sera pas valable est celui où T suit une loi bimodale, par exemple un mélange de 2 lois normales.

31 Très simple : partir de $p_{ijk}=\frac{1}{N}\exp\{i.\beta_a+j.\beta_b+k.\beta_c\}.\sum_{K=1}^{N}\exp(t(K).\omega).p_{K,000}$ où (rappel) $\omega=i+j+k$ et remplacer les indices i,j,k par les 0 et les 1 qu'il faut.

modèle, donc au final des 8 probabilités p_{ijk} . Cela conduit à des effectifs estimés pour chacune des 7 cellules où x_{ijk} est observée, soit $\hat{x}_{ijk} = n.\frac{\hat{p}_{ijk}}{1-\hat{p}_{000}}$. Il est ainsi possible de produire des statistiques de qualité de l'ajustement des modèles utilisant la déviance G^2 , concept attaché aux modèles linéaires généralisés qui constitue le pendant d'une somme de carrés de résidus pour un modèle linéaire classique en établissant une distance entre les 7 effectifs x_{ijk} observés et leurs prédictions $\hat{x}_{ijk}$.

$$G^2 = 2.\sum_{(i,j,k)\neq(0,0,0)} x_{ijk}.\log\frac{x_{ijk}}{\hat{x}_{ijk}}$$

Ainsi, plus le modèle est proche des données (donc moins il sera contraint), plus la déviance sera faible. Le test peut directement porter sur la comparaison de 2 modèles « emboités[32] ». Dans ce cas, l'un des deux modèles est le modèle le moins contraint (indice 0), celui qui s'ajuste le mieux aux données, l'autre est le modèle le plus contraint (avec un minimum de paramètres - indice 1 – ici prise en compte de la contrainte (*) indiquée ci-dessus) et la statistique testant la validité du modèle 1 par rapport au modèle 0 est :

$$G^2_{1/0} = 2.\sum_c \hat{x}_{c1}.\log\frac{\hat{x}_{c1}}{\hat{x}_{c0}}$$

où c désigne la case courante de la table de contingence et $\hat{x}_{c0}$ (respectivement $\hat{x}_{c1}$) est l'espérance de l'effectif en case c associé au modèle 0 (respectivement 1).

La double contrainte (*) portant sur les probabilités de capture successives est exigeante – et si on parvient à faire un test, peut-être sera-telle rejetée. Il est possible de l'assouplir en ne retenant qu'une des 2 égalités. Par exemple on peut se contenter de poser l'hypothèse $p_{011}.p_{100} = p_{101}.p_{010}$. Cet assouplissement a un équivalent au niveau du modèle de Rasch, qui se trouve lui aussi moins contraint et devient $\forall K$

$$\log\frac{p_{K,1++}}{p_{K,0++}} = t(K)+\beta_A \text{ et } \log\frac{p_{K,+1+}}{p_{K,+0+}} = t(K)+\beta_B \text{ et } \log\frac{p_{K,++1}}{p_{K,++0}} = s(K)+\beta_C$$

c'est-à-dire pour résumer que l'effet individu associé à la source 3 diffère de celui qui est associé aux sources 1 et 2 (là où il reste en revanche le même). Cela peut avoir une grande utilité en pratique lorsque l'appartenance à la source C (par exemple) est peu (ou pas) corrélée à l'appartenance aux sources A et B, lesquels sont pour leur part plutôt liées entre elles. Par exemple si A est le recensement, B un échantillon de contrôle du recensement et C le fichier des titulaires du permis de conduire, on peut penser être dans cette situation.

Si on traduit l'effet individuel $t(K)$ sous forme de variable aléatoire T et l'effet individuel $s(K)$ sous forme de variable aléatoire S , on vérifie que plus la covariance entre T et S est faible[33], plus le rapport

32 Les paramètres de l'un sont un sous-ensemble des paramètres de l'autre.
33 C'est-à-dire plus les effets individuels associés ont tendance à être différents.

$\frac{p_{110} \cdot p_{001}}{p_{011} \cdot p_{100}}$ - qui est aussi $\frac{p_{110} \cdot p_{001}}{p_{101} \cdot p_{010}}$ - est grand. Ce rapport, comparé à 1, mesure jusqu'à quel point on s'éloigne de la seconde égalité de (*).

Cette configuration dégradée est intéressante quand on peut ainsi distinguer des « groupes de listes », avec des corrélations d'appartenance au sein de chaque groupe, mais des faibles corrélations d'appartenance d'un groupe à l'autre. Dans le cas de 3 listes, l'expression mathématique du modèle explicitant les p_{ijk} évolue ainsi :

$$\log p_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \gamma(i+j,k)$$

où $\gamma(\omega,\omega') = \log E[\exp(T.\omega).\exp(S.\omega') \mid (i,j,k)=(0,0,0)]$. Il faut alors faire des hypothèses sur la loi conjointe (T,S) pour finaliser.

Le but ultime de ces opérations complexes reste l'estimation de x_{000} par $\hat{x}_{000}$: dans tous les cas (modèle de Rasch 'standard', ou modèle partiel, ou modèle d'indépendance complète entre sources, ou encore modèle avec interaction du second ordre) on dispose d'une théorie (fondée sur les développements limités) pour estimer un écart-type de $\hat{x}_{000}$.

Sur le plan pratique, les contextes présentés dans la littérature sont parfois peu robustes aux hypothèses posées, surtout lorsqu'il s'agit de populations peu couvertes par les listes disponibles. En particulier l'introduction de corrélations entre sources augmente significativement les estimations de taille de population (de manière équivalente, on peut dire que l'estimation supposant une indépendance qui n'existe pas sous-estime la taille de population - on retrouve là le résultat énoncé au tout début de cette partie). C'est pourquoi, avec un même jeu de données et lorsque la part cachée x_{00} est « grande » en proportion de la population totale, on trouve hélas plus souvent qu'il ne faudrait des estimations très différentes selon que l'on utilise tel ou tel modèle (par exemple dans un rapport de 1 à 2, voire même de 1 à 3 !!!).

Comme dans la plupart des cas pratiques, il est conseillé de stratifier *a priori* la population dès lors qu'on a connaissance de variables expliquant l'hétérogénéité des probabilités individuelles : le modèle de Rasch, ou toute autre modélisation individuelle équivalente, ne repose pas sur un jeu de variables pré identifiées qui expliquerait une telle hétérogénéité ; c'est pourquoi il est préférable de l'appliquer à des sous-populations au sein desquelles on ne sait plus distinguer explicitement ce qui cause l'hétérogénéité des probabilités de capture, autrement dit à des sous-populations au sein desquelles on soupçonne de l'hétérogénéité exclusivement résiduelle.

L'introduction de variables latentes

On peut utiliser des modèles qui prennent en compte des variables dites 'latentes'. Par définition, il s'agit de variables qualitatives *inobservables* qui interviennent dans l'explication de la structure de corrélation des variables caractérisant l'appartenance aux sources (Agresti, Lang 1993 ; Bartolucci, Forcina 2017) .

La variable latente est notée X et sa modalité courante l, variant de 1 à L. On choisit X et les modalités associées, mais on ne les observe pas. On considère que la probabilité d'appartenir à la source s vérifie $p_{K,s}=\phi_{ls1}$ (l'indice 1 caractérise toujours l'appartenance à la source) pour tout individu K dès lors qu'il vérifie la modalité l de X. La distinction entre les probabilités individuelles est donc entièrement et exclusivement expliquée par la valeur de X. On notera que le modèle de Rasch est compatible avec cette approche puisqu'il s'appuie sur des effets source (qui demeurent) et des effets individus – que l'on peut pour la circonstance considérer comme identiques lorsque ces individus appartiennent à certaines classes latentes. Ainsi, pour un tel modèle, la probabilité de présence dans la source s de tout individu K vérifiant la modalité l de X aurait la forme

$$p_{K,s}=\phi_{ls1}=\frac{e^{\alpha_l+\beta_s}}{1+e^{\alpha_l+\beta_s}}.$$

Quelle que soit la formulation de la probabilité individuelle, la propriété d'indépendance entre sources au niveau individu étant maintenue, il en résulte

$$\forall(i,j,k) \quad \text{et} \quad \forall K\in l \qquad p_{K,ijk}=p_{K,i++}\cdot p_{K,+j+}\cdot p_{K,++k}=\phi_{lAi}\cdot\phi_{lBj}\cdot\phi_{lCk}$$

puis $p_{ijk}=\sum_{K=1}^{N} p_{K,ijk}\cdot\frac{1}{N}=\frac{1}{N}\sum_{l=1}^{L} N_l\phi_{lAi}\cdot\phi_{lBj}\cdot\phi_{lCk}=\sum_{l=1}^{L} p_l\phi_{lAi}\cdot\phi_{lBj}\cdot\phi_{lCk}$ où $p_l=\frac{N_l}{N}$ est la probabilité pour un individu quelconque d'être en catégorie l. On peut toujours présenter p_{ijk} en disant qu'il s'agit d'une probabilité de capture relative à un individu 'pris au hasard'. On constate que l'agrégation des classes latentes a fait perdre l'indépendance entre les sources au niveau agrégé : il n'y a en effet aucune raison pour que les p_{ijk} s'écrivent comme un produit d'un terme (même très compliqué) en i, d'un terme en j et d'un terme en k. Cette expression suggère d'utiliser le modèle log linéaire suivant portant sur les effectifs attendus $Ex_{ijk}=N.p_{ijk}$:

$$\log Ex_{ijk}=\lambda+\lambda_i^A+\lambda_j^B+\lambda_k^C+\log\left[\sum_{l=1}^{L}\exp\left(\lambda_l^X+\lambda_{il}^{AX}+\lambda_{jl}^{BX}+\lambda_{kl}^{CX}\right)\right].$$

la partie de droite faisant intervenir un terme complexe qui mobilise des interactions entre la variable latente X et les 3 sources respectives. Comme dans le cas général (qui précède), on aboutit à un modèle log linéaire qui à l'évidence n'est pas caractéristique d'une indépendance entre les sources – bien que cette indépendance existe à l'origine au niveau individu. Mais l'hétérogénéité entre individus est d'une certaine façon 'moins forte' puisqu'elle ne provient que de la différence de modalité de la variable latente. On voit d'ailleurs immédiatement que si $L=1$ on retombe sur le modèle d'indépendance.

La propriété de quasi-symétrie (pas de modification de la probabilité par une permutation quelconque des 3 sources) est recherchée (bien que non nécessaire) : elle correspond aux égalités $\lambda_{1l}^{AX}=\lambda_{1l}^{BX}=\lambda_{1l}^{CX}$ et $\lambda_{0l}^{AX}=\lambda_{0l}^{BX}=\lambda_{0l}^{CX}$ pour toute modalité l de la variable latente. C'est une contrainte forte, qui porte sur les interactions entre la variable latente et chacune des sources.

En ce qui concerne la procédure d'estimation des paramètres, la littérature semble privilégier l'utilisation d'une densité des x_{ijk} sous forme multinomiale ou sous forme de produit de lois de Poisson indépendantes, l'estimation mobilisant un algorithme EM (Espérance-Maximisation ; Dempster, Laird, Rubin 1977). L'algorithme EM est une procédure itérative qui est précisément conçue pour produire de l'estimation de maximum de vraisemblance dans un cadre où il y des valeurs manquantes (très schématiquement, on impute les valeurs manquantes dans la densité – étape E - , on maximise la vraisemblance – étape M - , on produit des estimateurs des paramètres qui servent ensuite à refaire l'imputation des valeurs manquantes en l'améliorant – étape E - et on relance la boucle ainsi jusqu'à convergence). En la circonstance, les valeurs manquantes (prédites) en question sont des x_{ijkl} c'est-à-dire des effectifs $\widetilde{x}_{ijkl}$ reconstitués dans des cases définies à partir des modalités de la variable latente, lesquelles valeurs sont ensuite mobilisées dans un cadre classique mieux adapté au modèle log linéaire présenté ci-avant. Le modèle portant initialement sur les Ex_{ijk} (ci-dessus) fait en effet apparaître explicitement des paramètres impliquant les niveaux de la variable latente : partant de là, puisque $Ex_{ijk}=\sum_{l=1}^{L} E\widetilde{x}_{ijkl}$ on peut en déduire

$\log \sum_{l=1}^{L} E\widetilde{x}_{ijkl}=\log\left[\sum_{l=1}^{L}\exp\left(\lambda+\lambda_i^A+\lambda_j^B+\lambda_k^C+\lambda_l^X+\lambda_{il}^{AX}+\lambda_{jl}^{BX}+\lambda_{kl}^{CX}\right)\right]$ et il paraît naturel de poser $\log E\widetilde{x}_{ijkl}=\lambda+\lambda_i^A+\lambda_j^B+\lambda_k^C+\lambda_l^X+\lambda_{il}^{AX}+\lambda_{jl}^{BX}+\lambda_{kl}^{CX}$.

L'affaire se termine par l'estimation de tous les paramètres '*lambda*' (faisant apparaître ou non les niveaux l), puis de x_{000} selon $\log \hat{x}_{000}=\hat{\lambda}+\hat{\lambda}_0^A+\hat{\lambda}_0^B+\hat{\lambda}_0^C+\log\left[\sum_{l=1}^{L}\exp\left(\hat{\lambda}_l^X+\hat{\lambda}_{0l}^{AX}+\hat{\lambda}_{0l}^{BX}+\hat{\lambda}_{0l}^{CX}\right)\right]$.

Le nombre de classes latentes doit être fixé. Cette étape préalable semble quelque peu arbitraire, néanmoins on peut utiliser le critère BIC[34] pour faire ce choix, s'agissant bien d'une question de choix de modèle (critère à minimiser). L'appréciation de l'efficacité du modèle doit prendre en considération la déviance, et on peut aussi utiliser de nouveau le BIC.

On peut relâcher l'hypothèse d'indépendance au niveau individuel

Lorsqu'on connaît, pour chaque individu capturé au moins une fois, l'intégralité de son 'histoire' relative à ses captures-recaptures successives (mettons qu'il y en a T), on peut abandonner l'hypothèse H3 d'indépendance individuelle entre ces captures successives. Cette hypothèse est sur le principe assez forte, car elle nie tout phénomène d'apprentissage. La donnée élémentaire est constituée par la réalisation de la variable aléatoire Y_t relative à la source t , valant par exemple 1 si capture et 0 sinon. Dans le cadre le plus général, la loi de Y_t dépend de l'individu, mais il faut réduire la dimension du problème pour être en mesure de procéder à une estimation de N . Une méthode consiste à postuler l'existence de variables latentes X , définissant L classes au sein desquelles cette probabilité reste constante. Il peut y avoir beaucoup de classes, qu'il faut définir explicitement, en particulier il faut en fixer *a priori* le nombre total. L'approche standard de la modélisation avec variables latentes suppose que les comportements de capture

34 Critère d'inspiration bayésienne qui effectue un arbitrage entre la qualité d'ajustement d'un modèle et le nombre de paramètres impliqués, étant entendu que l'augmentation – non souhaitable - du nombre de paramètres va dans le sens d'un meilleur ajustement.

sont indépendants conditionnellement à l'appartenance aux classes latentes. Mais on peut sortir de ce scénario de base en introduisant des contraintes portant sur les probabilités jointes d'appartenance à plusieurs sources. Par exemple si on s'en tient à deux sources A et B, les probabilités à modéliser sont

$$P\left(Y_A=i, Y_B=j \mid X=l\right)$$

où $i\in\{0,1\}$, $j\in\{0,1\}$ et $l=1,2,...,L$.

L'idée générale consiste à faire porter différentes hypothèses – le choix est très ouvert ! - sur ces probabilités en leur imposant un système de contraintes paramétrées par de nouveaux paramètres β . Ainsi, on n'a plus l'indépendance *individuelle* (conditionnelle à X) entre sources, c'est-à-dire que

$$P\left(Y_A=i, Y_B=j \mid X=l\right) \neq P\left(Y_A=i \mid X=l\right) \ . \ P\left(Y_B=j \mid X=l\right) \ .$$

Concrètement, on part d'une expression pratique impliquant les probabilités conditionnelles marginales plus les probabilités jointes « simples » - sur laquelle on peut fonder des interprétations. On note η le vecteur constitué par les probabilités sur lesquelles on fait porter les hypothèses ; ce sont toujours des probabilités conditionnelles aux modalités de la variable latente.

Le cas le plus simple consiste à utiliser seulement $\eta_{A|l}=\log\frac{p\left(Y_A=1 \mid X=l\right)}{1-p\left(Y_A=1 \mid X=l\right)}$, qui intervient naturellement en situation d'indépendance entre individus.

Si on veut introduire des lois jointes, en conservant une approche assez simple (donc en autorisant des corrélations entre couples de sources – mais pas davantage), on utilisera par exemple une expression de type *odds-ratio* (en log), soit

$$\eta_{A,B|l}=\log\frac{p\left(Y_A=1, Y_B=1 \mid X=l\right).p\left(Y_A=0, Y_B=0 \mid X=l\right)}{p\left(Y_A=1, Y_B=0 \mid X=l\right).p\left(Y_A=0, Y_B=1 \mid X=l\right)} \ .$$

Lorsqu'il s'agit de captures successives au cours du temps t , une autre piste consiste à envisager des liaisons de type Markov, c'est-à-dire des probabilités de transition entre deux occasions de capture successives (voire plus que deux).

$$\eta_{t|l,t-1}=\log\frac{p\left(Y_t=1 \mid X=l, Y_{t-1}=y_{t-1}\right)}{1-p\left(Y_t=1 \mid X=l, Y_{t-1}=y_{t-1}\right)}$$

où $t=2,\ 3,...,\ T$ et $y_{t-1}\in\{0,1\}$.

On inclut dans le vecteur η les probabilités marginales de la variable latente, c'est-à-dire les $P\left(X=l\right)$, qui sont inconnues (par construction) et les lois conditionnelles des Y_s source par source (indice s pour la

source courante), sous la forme $\log \frac{p(Y_s=1|X=l)}{1-p(Y_s=1|X=l)}$, ainsi que toutes les lois jointes conditionnelles de type $\eta_{s,t|l}$ que l'on souhaite prendre en compte. On fait ensuite les hypothèses simplificatrices qui constituent le modèle, lesquelles consistent à poser

$$\eta = Z.\beta .$$

Si on a affaire au modèle à variable latente 'de base' qui postule l'indépendance conditionnelle des appartenances aux sources, Z est l'identité. Sinon, Z s'adapte pour que les contraintes que l'on a posées soient satisfaites : une matrice qui n'est plus l'identité va conduire à imposer des relations entre les coordonnées du vecteur η. Par exemple on peut adopter le modèle de Rasch à variables latentes (par construction, l'effet individu est constant au sein des classes latentes), auquel cas les contraintes portent sur les probabilités marginales - Z est alors constituée de 0 et de 1 placés dans certaines cases. Toujours en s'en tenant aux lois marginales, si les captures se font au cours du temps, on peut imposer que l'effet source β_t soit une fonction linéaire du temps, du type $\beta_t = \phi_0 + t.\phi_1$. S'il y a des lois jointes, la structure de Z va dépendre de la façon dont les lois jointes sont définies en fonction des probabilités marginales.

Par ailleurs, on relie le vecteur η aux probabilités jointes $P(Y=y, X=l)$, où $y \in \{0,1\}^T$, puisque ces dernières fournissent l'information maximale à partir de laquelle on peut exprimer toutes les probabilités conditionnelles et leurs transformées. Le vecteur $y=(y_1, y_2, ..., y_T)$ est composé de 0 et de 1, et décrit la succession des captures (ou non) d'un individu.

Le processus se déroule ensuite comme dans le cas où il y a indépendance entre individus (partie précédente) : on forme la vraisemblance en considérant la loi des effectifs observés (une loi multinomiale ou de Poisson, conditionnelle à des effectifs non nuls – donc observés), qui est une fonction des $P(Y=y|X=l)$ [35] et des $P(X=l)$, c'est-à-dire *in fine* des coefficients β. La maximisation de la densité en tant que fonction très complexe des β (y compris les $P(X=l)$) peut utiliser l'algorithme EM pour estimer les $P(X=l)$. On obtient une estimation de $P(Y=0)$ et on conclut par $\hat{N} = \frac{n}{1-\hat{P}(Y=0)}$.

On sait produire une incertitude, donc un intervalle de confiance si on néglige le biais devant l'écart-type, pour $\hat{N}$ (il existe une statistique encadrant la vraie valeur N dont la loi est un chi-2).

3. Les modèles « *sample coverage* »

Cette partie propose une méthode qui semble un peu à part dans l'ensemble des méthodes qui sont exposées dans la littérature, mais qui a un double avantage : d'une part elle permet de relâcher complètement deux des hypothèses socles, à savoir la H1 et la H3, et d'autre part elle produit des estimateurs de taille $\hat{N}$ qui ne nécessitent aucune mise en œuvre d'un algorithme savant. Ce dernier

35 Comme la densité est conditionnelle à l'observation, les probabilités sont au départ des probabilités conditionnelles à l'évènement $Y>0$ (au moins une des coordonnées de Y vaut 1).

argument est fort : dans les parties précédentes, dès lors qu'il y avait un modèle relâchant H1 ou H3, il fallait le plus souvent activer une procédure de maximum de vraisemblance : c'est loin d'être immédiat, cela est exigeant et place le niveau opérationnel assez haut. Ce qui suit relève d'un cheminement méthodologique certes complexe – au point parfois d'apparaître bien audacieux voire fragile - mais au final ne nécessite que des calculs algébriques à la portée d'un programmeur modeste (pour peu qu'on ne se laisse pas impressionner par la longueur de la formule obtenue !).

On trouvera davantage d'éléments dans (Chao, Tsay, 1998 ; Chao, Tsay 2001 ; Chao 2001, Chao 2015).

Préliminaires

La situation est celle d'une population de taille N (inconnue) et de T listes disponibles. Une méthodologie d'estimation de N utilise des modèles portant sur les relations entre les indicatrices $X_{K,s}=1_{K\ est\ présent\ dans\ la\ liste\ s}$, lesquelles sont techniquement des variables aléatoires de Bernoulli. Pour cela (cas de la modélisation dite « à effet fixe»), on définit pour tout couple de listes (s,t) donné, les coefficients

$$\gamma_{s,t}=\frac{1}{\mu_s.\mu_t}.\frac{1}{N}.\sum_{K=1}^{N} E\left(X_{K,s}-\mu_s\right).\left(X_{K,t}-\mu_t\right)$$

où $\mu_s=\frac{1}{N}.\sum_{K=1}^{N} P_{K,s}=\frac{1}{N}.\sum_{K=1}^{N} EX_{K,s}=E\,\bar{X}_s$ (probabilité d'inclusion moyenne dans la liste s).

En présence de T listes, on généralise ce coefficient ainsi

$$\gamma_{s_1,s_2,\ldots,s_T}=\frac{1}{\mu_{s_1}.\mu_{s_2}.\ldots..\mu_{s_T}}.\frac{1}{N}.\sum_{K=1}^{N} E\left(X_{K,s_1}-\mu_{s_1}\right).\left(X_{K,s_2}-\mu_{s_2}\right).\ldots..\left(X_{K,s_T}-\mu_{s_T}\right)$$

Ces coefficients (2 listes) s'écrivent $\frac{Ecov\left(X_s,X_t\right)}{E\,\bar{X}_s.E\,\bar{X}_t}$ et mesurent donc une association entre les échantillons capturés (un coefficient fort signifie qu'il faut s'attendre à récupérer deux échantillons à large intersection). Ils seront utiles pour l'estimation de N .

Ces concepts sont tout à fait généraux. En pratique, on peut trouver au moins deux situations (modèles) simplificatrices pour caractériser la relation entre les $X_{K,s}$.

La première situation postule l'indépendance[36] entre les appartenances aux diverses sources (H3 vérifiée, mais pas H1), pour chaque individu de la population (indépendance 'locale'). Toute l'information est alors contenue dans les lois marginales $P_{K,s}=Proba\left(X_{K,s}=1\right)$ puisque par hypothèse les lois jointes sont les produits des probabilités marginales. Les coefficients 'gamma' quantifient la corrélation entre ces probabilités, puisque ces coefficients deviennent

36 Il semble que l'on puisse lever cette hypothèse et développer une théorie qui s'en affranchisse.

$$\gamma_{s,t} = \frac{1}{\mu_s . \mu_t} . \frac{1}{N} . \sum_{K=1}^{N} \left(p_{K,s} - \mu_s \right) . \left(p_{K,t} - \mu_t \right) .$$

Il est essentiel de bien différencier la propriété d'indépendance locale et les propriétés de corrélation des lois marginales : la construction des probabilités jointes est déconnectée en général des relations pouvant exister entre les lois marginales. Les lois jointes sont les $Proba\left(X_{K,s} = 1 \;\; ET \;\; X_{K,t} = 1 \right)$. Par l'hypothèse d'indépendance 'locale', elles sont ici définies comme le produit des lois marginales $Proba\left(X_{K,s} = 1 \right)$ et $Proba\left(X_{K,t} = 1 \right)$. Cela ne préjuge pas de la relation entre ces deux lois marginales, en ce sens où les vecteurs dont les coordonnées sont respectivement les $Proba\left(X_{K,s} = 1 \right)$ et les $Proba\left(X_{K,t} = 1 \right)$ lorsque K décrit la population, peuvent par ailleurs être très corrélés, moyennement corrélés ou, à l'extrême, indépendants. Dit autrement, il ne faut pas mélanger une corrélation qui se conçoit plutôt 'dans le temps' et une corrélation qui se conçoit plutôt 'dans l'espace'.

La seconde situation est celle où il y a une dépendance entre sources, mais pas d'hétérogénéité des probabilités individuelles $Proba\left(X_{K,s} = 1 \right)$, qui sont toutes égales à P_s. Les lois jointes ne dépendent pas non plus de l'individu, et on les note P_{st} si les deux sources impliquées sont s et t. Ainsi, H1 serait vérifiée mais pas H3. Dans ce cas, on vérifie très facilement que

$$\gamma_{s,t} = \frac{P_{st}}{P_s . P_t} - 1$$

Si on a par exemple 3 sources

$$\gamma_{s,t,u} = \frac{P_{stu}}{P_s . P_t . P_u} - \gamma_{st} - \gamma_{su} - \gamma_{tu} - 1 .$$

Il est possible d'obtenir des coefficients 'gamma' de même nature – mais définis différemment - en adoptant une approche parallèle (modélisation dite « à effet aléatoire »). En effet, on exploite l'indépendance entre les N comportements individuels (les individus ne s'influencent pas les uns les autres, ce qui n'a rien à voir avec l'indépendance entre les T sources – ou indépendance locale) pour considérer les N vecteurs rassemblant les T comportements étudiés comme des réalisations indépendantes de vecteurs aléatoires identiquement distribués. La loi commune de ces vecteurs est *a priori* quelconque, mais son intérêt est de définir explicitement la structure de corrélation entre les T probabilités marginales. Ainsi $\left(P_{K,1}, P_{K,2}, \ldots, P_{K,T} \right)$ - que l'on notera plus simplement $\left(P_1, P_2, \ldots, P_T \right)$ - suit une certaine loi multidimensionnelle paramétrée (que l'on postule !). Alors on définit - *primo* pour toute source s : $\mu_s = EP_s$ et *secundo* pour tout couple de sources $\left(s, t \right)$:

$$\gamma_{s,t} = \frac{E\left(P_s - \mu_s \right) . \left(P_t - \mu_t \right)}{\mu_s . \mu_t}$$

également généralisable au cas de T listes. Cette approche (effets aléatoires) produit *in fine* les mêmes estimations et les mêmes résultats que la précédente (effets fixes) .

Quelle que soit l'approche, c'est cette corrélation 'gamma' qui est à l'origine de la dépendance entre échantillons successifs, en créant une structure particulière de répartition des effectifs parmi les différents échantillons. En particulier il y aura des accumulations de présence conjointe d'individus parmi les couples d'échantillons pour lesquels $\gamma_{s,t}$ est grand. Avec le modèle de Rasch, en présence de deux sources, il est toujours vrai que $\gamma_{s,t} \geq 0$.

i) Cas de 2 sources :

Le « *sample coverage* » (SC) pour un échantillon s donné est défini de manière traditionnelle à partir d'un jeu de probabilités d'inclusion p_K associées à s : c'est la moyenne pondérée par ces probabilités de la variable indicatrice d'appartenance à s soit

$$SC = \frac{\sum_{K=1}^{N} p_K . 1_{K \in s}}{\sum_{K=1}^{N} p_K} .$$

On remarquera que si les probabilités individuelles sont constantes, il s'agit du taux de sondage – de façon plus générale, on peut donc dire que c'est un taux de sondage pondéré. On peut aussi, de manière plus cohérente avec le vocabulaire employé pour désigner cet indicateur, le voir comme un taux de couverture (pondéré), puisqu'après tout un taux de sondage s'interprète aussi comme la proportion de la population totale qui se trouve présente dans ('couverte par') l'échantillon. En présence de 2 listes indépendantes A et B, on introduit le SC de la liste-échantillon A par rapport à la liste-échantillon B:

$$SC(A;B) = \frac{\sum_{K=1}^{N} p_{K,B} . 1_{K \in A}}{\sum_{K=1}^{N} p_{K,B}} = \frac{\sum_{K=1}^{N} p_{K,B} . X_{K,A}}{\sum_{K=1}^{N} p_{K,B}}$$

puisque $1_{K \in A} = 1(X_{K,A} > 0) = X_{K,A}$. Cet indicateur quantifie le recouvrement entre les deux sources. La pondération rend sa perception plus compliquée, mais en supposant un instant que $p_{K,B}$ est constant, l'indicateur est égal au rapport de la taille de la liste A à la taille de la population totale – et on retrouve le taux de couverture de la population par la liste A. L'ajout de poids individuels $p_{K,B}$ conserve l'interprétation d'un taux de couverture « en moyenne », en considérant qu'il s'agirait plutôt de la couverture de la sous-population de la liste B assurée par la liste A : en prenant une grande liberté avec les calculs d'espérance,

on peut en effet considérer que $SC(A;B)$ est l'espérance de $\frac{\sum_{K=1}^{N} 1_{K\in B}.X_{K,A}}{\sum_{K=1}^{N} 1_{K\in B}}$, qui est la part de la sous-population de la liste B que l'on retrouve dans la liste A.

Pour tenir compte maintenant de la dépendance entre sources (hypothèse H3 non vérifiée) et se placer dans le cas le plus général qui soit (puisque par ailleurs H1 n'est pas vérifiée non plus), on remplace $p_{K,B}=E(X_{K,B})$ par $E(X_{K,B}|X_{K,A})=p(K\in B|X_{K,A})$. Pour symétriser la question, on utilise finalement :

$$SC=\frac{1}{2}.\left\{\frac{\sum_{K=1}^{N} E(X_{K,B}|X_{K,A}).X_{K,A}}{\sum_{K=1}^{N} E(X_{K,B}|X_{K,A})} + \frac{\sum_{K=1}^{N} E(X_{K,A}|X_{K,B}).X_{K,B}}{\sum_{K=1}^{N} E(X_{K,A}|X_{K,B})}\right\}.$$

On interprète cet indicateur comme une mesure de l'information apportée par le recouvrement des deux listes. Intuitivement, plus ce taux est élevé, moins il y aura d'individus qui ne sont couverts que par une seule des sources, par extension plus l'intersection des deux sources sera grande dans l'ensemble constitué par leur réunion. Le conditionnement des espérances ne modifie pas l'interprétation.

Le principe essentiel est que le *sample coverage* peut être bien estimé dans un contexte de dépendance locale (*a fortiori* de dépendance 'globale'), donc dans un cadre général où l'hypothèse H3 n'est pas vérifiée. On cherche donc à obtenir une relation entre le *sample coverage* SC et la taille de population N, et à en tirer une estimation de N à partir d'une estimation du SC.

Si on assimile l'espérance d'un ratio au ratio des espérances (ce qui est bien sûr faux en toute rigueur, mais cela permet de produire un estimateur par les moments qui soit raisonnable), on a

$$E\left\{\frac{\sum_{K=1}^{N} E(X_{K,B}|X_{K,A}).X_{K,A}}{\sum_{K=1}^{N} E(X_{K,B}|X_{K,A})}\right\}\approx\frac{\sum_{K=1}^{N} E(E(X_{K,B}|X_{K,A}).X_{K,A})}{\sum_{K=1}^{N} E(X_{K,B})}$$

$$E\left\{\frac{\sum_{K=1}^{N} E(X_{K,B}|X_{K,A}).X_{K,A}}{\sum_{K=1}^{N} E(X_{K,B}|X_{K,A})}\right\}=\frac{\sum_{K=1}^{N} E(E(X_{K,A}.X_{K,B}|X_{K,A}))}{E(\sum_{K=1}^{N} X_{K,B})}$$

$$E\left\{\frac{\sum_{K=1}^{N} E\left(X_{K,B}|X_{K,A}\right).X_{K,A}}{\sum_{K=1}^{N} E\left(X_{K,B}|X_{K,A}\right)}\right\}=\frac{\sum_{K=1}^{N} E\left(X_{K,A}.X_{K,B}\right)}{x_{+1}}$$

en remarquant que $\sum_{K=1}^{N} X_{K,B}$ est le nombre total d'individus dans la liste B, que nous avons toujours noté x_{+1} .

C'est aussi (par définition de x_{11}) :

$$\frac{E\sum_{K=1}^{N} X_{K,A}.X_{K,B}}{x_{+1}}=\frac{E\,x_{11}}{x_{+1}} .$$

Finalement, selon l'approche de l'estimation par les moments, SC sera estimé (de manière *a priori* 'pas trop biaisée') par $\hat{SC}=\frac{1}{2}.\left\{\frac{x_{11}}{x_{+1}} + \frac{x_{11}}{x_{1+}}\right\}$.

Par ailleurs, en toutes circonstances (donc en situation générale de dépendance locale), c'est-à-dire dans un contexte où ni H1 ni H3 ne sont vraies :

$$\gamma_{A,B}=\frac{1}{\mu_A.\mu_B}.\frac{1}{N}.\sum_{K=1}^{N} E\left(X_{K,A}.X_{K,B}\right) - 1$$

Or $\mu_A=\frac{1}{N}.\sum_{K=1}^{N} P_{K,A}=\frac{E\left(x_{1+}\right)}{N}$ et $\sum_{K=1}^{N} X_{K,A}.X_{K,B}=x_{11}$ si bien que

$\gamma_{A,B}=\frac{N}{E\left(x_{+1}\right).E\left(x_{1+}\right)}.E\left(x_{11}\right) - 1$ et

$$N=\frac{E\left(x_{+1}\right).E\left(x_{1+}\right)}{E\left(x_{11}\right)}.\left(1+\gamma_{A,B}\right) .$$

Dans le cas particulier où les hypothèses-socle sont vérifiées, on constate immédiatement que $\gamma_{A,B}=0$. La vraie valeur $N=\frac{E\left(x_{+1}\right).E\left(x_{1+}\right)}{E\left(x_{11}\right)}$ est naturellement estimée par $\hat{N}=\frac{x_{+1}.x_{1+}}{x_{11}}$, qui n'est autre que l'estimateur de Lincoln-Petersen.

Mais de façon générale, $\gamma_{A,B} \neq 0$. L'estimation de $\gamma_{A,B}$ n'est alors malheureusement pas possible car on doit estimer 4 paramètres alors que l'on ne dispose que de 3 données : les paramètres sont $\gamma_{A,B}, N, \mu_A, \mu_B$ et les données sont x_{11}, x_{12}, x_{21} . Par ailleurs, on ne peut tester aucune hypothèse concernant $\gamma_{A,B}$. Et si, dans l'hypothèse très favorable où on disposerait d'une estimation *a priori* $\widetilde{N}$ obtenue grâce à des données externes, on formait $\hat{\gamma}_{A,B} = \frac{\widetilde{N}}{x_{+1}.x_{1+}}.x_{11} - 1$, il en sortirait au final $\hat{N} = \widetilde{N}$ - donc la qualité serait entièrement dépendante de celle de $\widetilde{N}$.

Cette situation débouche donc sur une impasse. C'est parce qu'on mobilise trop peu de sources : avec 2 sources, la méthode n'aboutit pas. Par contre, la partie suivante montre que l'on peut conclure à partir de 3 sources.

ii) Cas de 3 sources :

Cette configuration va permettre d'estimer N , contrairement au cas de 2 sources.

La notion de couverture définie ici concerne cette fois les réunions de 2 listes, la pondération étant apportée par la troisième liste. On introduit 3 composantes du *sample coverage SC*, et considérant les trois couples de listes (A,B), puis (A,C) puis (B,C) et en généralisant la construction du cas précédent (2 échantillons). Pour le couple de listes (A,B) par exemple, on forme

$$SC(A,B) = \frac{\sum_{K=1}^{N} E(X_{K,C} | X_{K,A}, X_{K,B}).1(X_{K,A} + X_{K,B} > 0)}{\sum_{K=1}^{N} E(X_{K,C} | X_{K,A}, X_{K,B})} \quad .$$

S'il n'y a pas de dépendance locale, alors $E(X_{K,C} | X_{K,A}, X_{K,B}) = P_{K,C}$ et le dénominateur vaut Ex_{++1} .

On notera que l'indicatrice $1(X_{K,A} + X_{K,B} > 0)$ vaut 1 exactement lorsque K apparaît dans la liste A ou/et dans la liste B, ce qui concerne 3 configurations : $X_{K,A} = 1 \ et \ X_{K,B} = 0$ ou $X_{K,A} = 0 \ et \ X_{K,B} = 1$ ou $X_{K,A} = 1 \ et \ X_{K,B} = 1$. Raisonnant comme dans la partie précédente, le dénominateur est proche de la taille de la liste C et le numérateur est proche de la taille de l'intersection entre la liste C (traduite par l'espérance conditionnelle) et la réunion des listes A et B (traduite par l'indicatrice). Cet indicateur est normalement proche de la part de la liste C qui se trouve également dans la réunion des listes A et B. Finalement, c'est donc un taux de couverture de la liste C par l'union des listes A et B.

On forme l'indicateur SC final selon :

$$SC(A,B,C)=\frac{1}{3}.(SC(A,B)+SC(A,C)+SC(B,C)) \ .$$

C'est la couverture moyenne d'une liste par la réunion des deux autres. On peut vérifier que, sans hypothèse particulière, l'espérance de SC est approximée[37] par

$$E(SC)\approx 1 \ - \ \frac{1}{3}.\left\{\frac{\sum_K P[X_{K,A}=1,X_{K,B}=0,X_{K,C}=0]}{\sum_K E(X_{K,A})}+\frac{\sum_K P[X_{K,A}=0,X_{K,B}=1,X_{K,C}=0]}{\sum_K E(X_{K,B})}+\frac{\sum_K P[X_{K,A}=0,X_{K,B}=0,X_{K,C}=1]}{\sum_K E(X_{K,C})}\right\}$$

On en tire immédiatement un estimateur de SC, dans l'esprit d'un estimateur des moments :

$$\hat{SC}(A,B,C)=1-\frac{1}{3}.(\frac{x_{100}}{x_{1++}}+\frac{x_{010}}{x_{+1+}}+\frac{x_{001}}{x_{++1}}) \ .$$

Un atout très appréciable de cette méthode vient du fait que ce taux de couverture SC est estimable dans un cadre très général, sans qu'il n'y ait besoin de formuler aucune des hypothèses fortes que sont l'homogénéité et l'indépendance locale. De plus, on constate que $\hat{SC}(A,B,C)$ s'exprime très simplement à partir des données disponibles.

Cet avantage ne signifie pas que l'indicateur en lui-même soit exprimable de manière simple dans le cadre général. En revanche, on vérifie qu'en l'absence de toute dépendance, locale comme globale - situation d'homogénéité des probabilités de capture individuelles et pas de corrélation entre les appartenances aux différents échantillons - on obtient une expression simple de $SC(A,B,C)$ puisque :

$$SC(A,B,C)=\frac{\Delta}{N}$$

où

$$\Delta=\frac{1}{3}.\left\{\sum_{K=1}^{N}1(X_{K,A}+X_{K,B}>0)+\sum_{K=1}^{N}1(X_{K,A}+X_{K,C}>0)+\sum_{K=1}^{N}1(X_{K,B}+X_{K,C}>0)\right\} \ .$$

Cela conduit à l'estimateur naturel de N

$$\hat{N}=\frac{\Delta}{\hat{SC}(A,B,C)}$$

Les 3 composantes de Δ sont très facilement calculables puisqu'elles dénombrent chacune les individus qui sont dans l'une au moins des 2 listes considérées (en considérant respectivement tous les couples de listes). Chaque somme dénombre une réunion de 2 ensembles et c'est pourquoi Δ/N s'interprète

37 On n'hésite pas à approximer l'espérance d'un ratio par le ratio des espérances.

naturellement comme un taux de recouvrement. Avec les notations standards, on peut écrire $\Delta = \frac{1}{3}.\left\{2.\left(x_{1++} + x_{+1+} - x_{++1}\right) - x_{+11} - x_{1+1} - x_{11+}\right\}$.

Finalement, sous la condition essentielle qu'il y ait indépendance (locale) complète et homogénéité des probabilités de capture individuelles (donc respect des hypothèses socle) :

$$\hat{N} = \frac{\left\{2.\left(x_{1++} + x_{+1+} - x_{++1}\right) - x_{+11} - x_{1+1} - x_{11+}\right\}}{3 - \left(\frac{x_{100}}{x_{1++}} + \frac{x_{010}}{x_{+1+}} + \frac{x_{001}}{x_{++1}}\right)} .$$

Lorsqu'il y a dépendance locale et hétérogénéité des probabilités d'inclusion, donc dans un cadre général, on étend la définition des coefficients γ au cas de 3 listes

$$\gamma_{A,B,C} = \frac{1}{\mu_A.\mu_B.\mu_C}.\frac{1}{N}.\sum_{K=1}^{N} E\left(X_{K,A} - \mu_A\right).\left(X_{K,B} - \mu_B\right).\left(X_{K,C} - \mu_C\right)$$

Sans toucher à la définition de Δ et en partant du développement de son espérance $E\Delta$ dans le contexte de dépendance générale (mêlant par conséquent l'hétérogénéité des probabilités individuelles et les dépendances entre sources), on peut montrer la relation[38] :

$$N = \frac{E(\Delta)}{E(SC)} + \frac{N}{3.E(SC)}.\left(\Gamma_1 + \Gamma_2\right) ,$$

avec

$$\Gamma_1 = \left(\mu_A + \mu_B\right).\left(1 - \mu_C\right).\gamma_{A,B} + \left(\mu_A + \mu_C\right).\left(1 - \mu_B\right).\gamma_{A,C} + \left(\mu_B + \mu_C\right).\left(1 - \mu_A\right).\gamma_{B,C}$$

$$\Gamma_2 = -\left(\mu_A.\mu_B + \mu_A.\mu_C + \mu_B.\mu_C\right).\gamma_{A,B,C} .$$

$E(SC)$ conserve l'expression (valable en situation générale) qu'elle a ci-dessus.

Cette équation fort compliquée reliant N et SC constitue la colonne vertébrale de l'opération d'estimation. Elle relie plus exactement la taille de population N à l'espérance du SC , aux probabilités moyennes (les μ), et aux mesures de dépendance introduites ci-avant (les γ). Intervient aussi l'espérance $E(\Delta)$, qu'on ne cherche pas à développer davantage (parce que ce n'est pas utile pour la suite) et qui dépend de manière très compliquée des probabilités individuelles d'inclusion jointe. Noter que si les hypothèses socles sont vérifiées, tous les γ sont nuls, et que si on postule un modèle de Rasch, contrairement au cas de 2 échantillons, on peut avoir des $\gamma_{A,B,C}$ négatifs.

38 C'est en réalité une égalité approchée, valide dès lors que l'espérance d'un ratio est assimilable au ratio des espérances – les variables aléatoires sont ici des indicatrices d'appartenance aux listes. On ne peut y croire que si la population est de grande taille.

Après transformation de l'expression précédente, on aboutit (en toute généralité toujours) à une forme plus opérationnelle :

$$N = \frac{E(\Delta)}{E(SC)} + \frac{\Gamma_3}{3.E(SC)} + \frac{R.N}{3E(SC)}$$

$$\Gamma_3 = E(x_{1+0} + x_{+10}).\gamma_{A,B} + E(x_{10+} + x_{+01}).\gamma_{A,C} + E(x_{01+} + x_{0+1}).\gamma_{B,C}$$

$$R = \mu_A \mu_B.\{\gamma_{AB}(\gamma_{AC} + \gamma_{BC}) - \gamma_{ABC}\} + \mu_A \mu_C.\{\gamma_{AC}(\gamma_{AB} + \gamma_{BC}) - \gamma_{ABC}\} + \mu_B \mu_C.\{\gamma_{BC}(\gamma_{AB} + \gamma_{AC}) - \gamma_{ABC}\}$$

A ce stade, on fait remarquer que dans de nombreuses circonstances pratiques, on a $R = 0$:

- c'est vrai s'il n'y a pas de dépendance entre les sources et qu'il y a hétérogénéité des probabilités $P_{K,s}$ dans une seule des 3 sources (c'est donc en particulier vrai si les hypothèses socles sont vérifiées) ;

- c'est vrai également s'il y a homogénéité des probabilités mais que la dépendance ne concerne qu'un seul couple de sources sur les trois couples possibles ;

- c'est tout aussi vrai sous le modèle $P(X_{Ks} = 1) = h_K.e_s$ où les $(h_1, h_2, h_3, ..., h_N)$ sont considérés comme N réalisations indépendantes et identiquement distribuées d'un tirage dans une loi $gamma(\alpha, \beta)$. La classe 'gamma' est une classe très large recouvrant de nombreuses lois usuelles. Les γ peuvent s'écrire comme des fonctions de α, donc on exprime facilement R en fonction de α et on vérifie qu'il est nul.

La littérature considère qu'il est justifié également d'annuler R dans le cas d'un modèle de Rasch, et dans d'autres situations pratiques encore.

On considère donc désormais que $R = 0$. Cette considération est essentielle et permet à ce processus d'estimation d'aboutir, car alors il est bien clair que cette condition conduit à ce que $\gamma_{A,B,C}$ devienne fonction des autres γ et des μ : le fait d'économiser l'estimation de $\gamma_{A,B,C}$ sauve la mise puisque le nouveau bilan comptable conduit à 7 paramètres ($\mu_A, \mu_B, \mu_C, \gamma_{AB}, \gamma_{AC}, \gamma_{BC}, N$) qu'il faut estimer à partir de 7 données (tous les x_{stu} croisant les 3 sources – sauf x_{000} qui est la seule valeur non observée) – ce qui devient faisable. D'où l'estimateur de type moments défini par :

$$\hat{N} = \frac{\Delta}{\hat{SC}} + \frac{1}{3.\hat{SC}}.\{(x_{1+0} + x_{+10}).\hat{\gamma}_{A,B} + (x_{10+} + x_{+01}).\hat{\gamma}_{A,C} + (x_{01+} + x_{0+1}).\hat{\gamma}_{B,C}\}$$

où (rappel)

$$\hat{SC}(1,2,3)=1-\frac{1}{3}.\left(\frac{x_{100}}{x_{1++}}+\frac{x_{010}}{x_{+1+}}+\frac{x_{001}}{x_{++1}}\right)$$

avait été obtenu dans le cadre le plus général, et (rappel également)

$$\Delta=\frac{1}{3}.\left\{2.\left(x_{1++}+x_{+1+}-x_{++1}\right)-x_{+11}-x_{1+1}-x_{11+}\right\} .$$

On rappelle que, sans qu'aucune condition ne soit requise, on a toujours

$\gamma_{A,B}=\frac{N}{E(x_{1++}).E(x_{+1+})}.E(x_{11+})-1$ (voir partie précédente – cas de 2 sources) et donc

$\hat{\gamma}_{A,B}=\frac{\hat{N}}{x_{1++}.x_{+1+}}.x_{11+}-1$ (idem pour $\hat{\gamma}_{A,C}$ et $\hat{\gamma}_{B,C}$).

Par des simulations, on a vérifié que cet estimateur a un « bon comportement » dès lors que le *sample coverage* SC est « suffisamment grand » (la littérature se risque à mentionner un seuil de 55 % ...). Si cet indicateur est trop faible, la variance de $\hat{N}$ sera grande. Il y a donc des conditions d'utilisation. Cela étant, si le SC est petit et que la corrélation entre les appartenances aux échantillons est positive (respectivement négative), on pourra considérer $\hat{N}$ comme une borne inférieure (respectivement supérieure) de la vraie valeur N .

A la fin, on obtient l'estimation

$$\hat{N}=\frac{x_{+11}+x_{1+1}+x_{11+}}{3\hat{SC}}.\left\{1-\frac{1}{3\hat{SC}}\left[\frac{(x_{1+0}+x_{+10}).x_{11+}}{x_{1++}.x_{+1+}}+\frac{(x_{10+}+x_{+01}).x_{1+1}}{x_{1++}.x_{++1}}+\frac{(x_{0+1}+x_{01+}).x_{+11}}{x_{+1+}.x_{++1}}\right]\right\}^{-1}$$

A ce stade, il s'agit de l'expression obtenue dans le cadre le plus général parmi ceux qui ont été abordés dans ce document : cet estimateur accepte l'hétérogénéité des probabilités individuelles (H1 non vérifiée), et il accepte aussi toute structure de dépendance entre les échantillons (H3 non vérifiée). Il faut néanmoins que les comportements des individus restent indépendants d'un individu à l'autre – H2 vérifiée - mais c'est la seule des 3 composantes des hypothèses socle qui subsiste. Il faut néanmoins garder en mémoire qu'une 'acrobatie', plus ou moins audacieuse, conduit à considérer le coefficient R comme nul et que cette condition est nécessaire pour que l'estimateur ci-dessus soit valable

Quand N tend vers $+\infty$, alors $\frac{\hat{N}}{N} \rightarrow 1-\frac{R/N}{A+R/N}$ (convergence en probabilité), où $A=\mu_A\mu_B.\{\gamma_{AB}+1\}+\mu_A\mu_C.\{\gamma_{AC}+1\}+\mu_B\mu_C.\{\gamma_{BC}+1\}$. Le biais asymptotique de $\hat{N}$ est nul si et seulement si $R=0$ et il est d'autant plus faible que R est faible. Donc R joue clairement un rôle central pour la qualité de $\hat{N}$.

Un estimateur de variance de $\hat{N}$ par *bootstrap* est possible, en formant des réplications $\{x^{\epsilon}_{000}, x^{\epsilon}_{100}, \ldots, x^{\epsilon}_{111}\}$ issues d'une loi multinomiale de paramètres $\hat{N}$ et $\{\frac{\hat{x}_{000}}{\hat{N}}, \frac{x_{100}}{\hat{N}}, \ldots, \frac{x_{111}}{\hat{N}}\}$ pour ce qui concerne les probabilités associées. Noter qu'on a obtenu $\hat{x}_{000}$ en calculant $\hat{N} - \sum_{ijk} x_{ijk} 1_{[i+j+k>0]}$. Pour chaque réplication, on déroule le processus d'estimation décrit ci-dessus et on obtient une estimation $\hat{N}^{\epsilon}$. Les statistiques empiriques habituelles permettent d'apprécier la variance de $\hat{N}^{\epsilon}$ - et donc de $\hat{N}$.

Une généralisation au cas d'un nombre quelconque de sources existe, au prix d'expressions formelles des estimateurs très compliquées.

Conclusion

L'univers des méthodes statistiques développées pour estimer des tailles de population est extrêmement vaste, et la recherche sur ce sujet est très active. La littérature est abondante et on en trouve trace dans des périodes anciennes – des articles datant d'avant guerre figurant encore dans les références bibliographiques (par exemple '*The Estimation of the Total Fish Population of a Lake* » publié par Z. Schnabel en 1938). Aux Etats-Unis, le *Census Bureau* a fait un usage abondant de ces techniques au début des années 80 – peut-être même avant - pour estimer la sous-couverture des recensements.
Les méthodes utilisées sont souvent astucieuses, mais parfois aussi très compliquées. Voici quelques obstacles, réserves, ou plus modestement des remarques – disparates - que l'on pourrait formuler pour qualifier cet univers :

- les méthodologies s'insèrent bien dans le paradigme dominant, celui de la statistique dépendant de modèles : les modèles, que l'on peut qualifier de manière très générale comme des hypothèses simplificatrices de la réalité, sont partout présents, la vérification – ou non - des hypothèses socles structurant par ailleurs les développements théoriques.

- le champ d'application intéressant concerne les cas où on dispose de trois sources ou davantage. Avec 'seulement' deux sources, les outils exploitables s'avèrent finalement assez pauvres.

- autant la sous-couverture des sources disponibles constitue le fondement même de tous ces développements, qui s'en nourrissent, autant la sur-couverture est une problématique complexe qui doit être traitée en amont et qui ne peut que dégrader la qualité des estimations des tailles de population. C'est l'occasion de rappeler qu'il ne faut jamais oublier la différence entre deux concepts bien différents : un enregistrement dans un fichier et un individu dans une population.

- dans la série « hypothèse non valide », il serait intéressant d'ajouter un point sur les appariements, cruciaux dans nos processus comme Résil, et sur le travail soigneux à mener pour en estimer les défauts.

- les expériences reproduites dans la littérature semblent globalement montrer un manque de robustesse à la méthode assez marqué – ce qui n'est pas *a priori* rassurant.
L'estimateur de variance est lié à la taille de l'intersection entre les deux populations. Dans le cas général, c'est donc les taux de couverture des deux listes, et de leur intersection qui va déterminer la variance de l'estimateur.
Dans le cas où ces taux sont très élevés, la variance de l'estimateur sera faible, et l'erreur quadratique aussi, si toutefois les hypothèses sont respectées.
Si les hypothèses ne sont pas respectées, alors le biais induit pourra très vite dépasser ce qu'on pourrait estimer comme une amélioration de l'estimation.
Il faut probablement voir ces techniques comme des façons de valider – ou non - les estimations de tailles de population produites par les opérations statistiques sur lesquelles plane le doute. Il n'est pas du tout sûr qu'elles soient appropriées pour déceler avec finesse des sous-couvertures associées à des sources quasi exhaustives, ou même seulement qui prétendent l'être. Concrètement, si on envisage de les exploiter pour corriger les dénombrements de sources pseudo exhaustives - comme le recensement ou Résil - il pourrait être prudent de se limiter à des sous-populations ciblées et indiscutablement sous-estimées., c'est-à-dire des sous populations pour lesquelles l'erreur d'estimation en l'absence de mobilisation de DSE est supérieure aux biais liés aux hypothèses non respectées.
Autrement dit, l'apport des méthodes DSE doit être évalué à l'aune du respect supposé des hypothèses.

- Exactement comme pour le traitement de la non-réponse en théorie des sondages, il y a un effort à faire pour comprendre – et accepter ! - la notion d'inclusion aléatoire dans une liste. C'est certes le fondement de

ces méthodes mais c'est aussi une vision abstraite, difficile à expliquer à des non statisticiens parce qu'elle n'est pas vraiment naturelle : qu'un individu soit ou ne soit pas dans une liste est évidemment facile à constater, mais il est beaucoup moins clair que ce soit un phénomène résultant d'une expérience aléatoire gérable par une probabilité (ce qui n'empêche pas que, mathématiquement, ce paradigme soit très adapté !). L'approche aléatoire a aussi l'avantage d'offrir un cadre pour exprimer un éventuel biais théorique de l'estimateur de la taille de population et pour calculer sa variance. Néanmoins, elle crée aussi quelques paradoxes. Par exemple, le sens à donner à la propriété d'indépendance entre sources (une des trois composantes fondamentales des hypothèses socle) lorsque deux (ou plus) sources sont quasi exhaustives est assez déstabilisateur : pour de très nombreux individus, la probabilité d'appartenance à chaque source est 'presque' 1, la probabilité d'appartenance conjointe est 'presque' 1, par conséquent on se trouve de manière mécanique en situation de (quasi) indépendance ! Dans le même style : lorsque les probabilités individuelles d'appartenance aux sources sont (très) majoritairement proches de 1, ne faut-il pas considérer qu'on se trouve de fait en situation de probabilités homogènes (une autre des composantes fondamentales des hypothèses socle) ?
La difficulté dans la mise en œuvre concrète sera donc plutôt de disposer de variables de stratification permettant de distinguer des sous-populations pour lesquelles les hypothèses seraient bien respectées des sous-populations pour lesquelles on a plus de doute.

- L'existence d'un éventuel échantillonnage au sein des listes – en tout ou partie - ne devrait jamais constituer un obstacle infranchissable. Un échantillonnage complique certes le contexte et les développements théoriques, mais ce n'est qu'une 'couche' d'aléa supplémentaire, qui se cumule avec l'aléa 'primitif' - pour ne pas dire fondateur – qui sous-tend toutes ces méthodes d'estimation par capture-recapture. On n'en trouve cependant que peu de traces dans la littérature, hormis dans le contexte d'enquêtes post-censitaires, car les statisticiens d'enquête se sont manifestement (très) peu intéressés à ces méthodes d'estimation par système multiple – ce qui est par ailleurs bien naturel puisqu'il ne s'agit pas de techniques d'enquête mais de techniques de dénombrement.

Ce document propose un aperçu de quelques méthodes d'estimation de tailles de population, sans même prétendre être exhaustif en matière de types de méthodes présentes dans la littérature. Il reste beaucoup à faire, en particulier il demande à être complété par les offres 'logiciel', *a priori* celles accessibles sous forme de *package R.* Il serait également éclairant d'apprécier la robustesse des estimations aux différentes hypothèses socles en appliquant un protocole de simulations sur des données artificielles dont on contrôle parfaitement la vraie structure[39].

* * * * *

39 Un exemple simple : on simule une liste A à partir de valeurs $P(A)$ et $P(\bar{A})$ et une liste B à partir de valeurs $P(B|A)$ et $P(B|\bar{A})$ qui diffèrent – à voir alors le comportement de l'estimateur de Lincoln-Petersen et celui des estimateurs plus savants qui tiennent compte de la dépendance.

ANNEXE : les modèles log-linéaires

i) Cas de 2 variables qualitatives

On distingue deux variables qualitatives A et B prenant respectivement I et J modalités (dans cette annexe, pour plus de généralité, on les considérera quelconques). En conséquence, la population est partitionnée en $I.J$ sous-populations. Les individus sont classés dans un ordre qui permet de distinguer les sous-populations caractérisées par les modalités (i,j). On veut ré écrire un vecteur l de coordonnées l_{ij}, vecteur de $\mathbb{R}^{IJ}$, en posant dans la sous-population (i,j) :

$$l_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}.$$

Cette réécriture des l_{ij} - dont les valeurs sont absolument quelconques - suppose que les paramètres $\lambda, \lambda_i^A, \lambda_j^B$ et λ_{ij}^{AB} soient définissables, ce qui signifie qu'ils peuvent s'obtenir de manière unique à partir des l_{ij}. Une façon simple de concevoir la question consiste à considérer le vecteur l comme un vecteur quelconque de l'espace vectoriel $\mathbb{R}^{IJ}$ qu'il faut décomposer dans une base de vecteurs de ce même espace, puisque dans toute base la décomposition est unique.

Dans $\mathbb{R}^{IJ}$, toute base est constituée d'exactement $I.J$ vecteurs formant un système libre. Le problème se transforme de fait en problème de dénombrement, car les vecteurs constituant la base sont formés par des 0 et des 1 et lorsqu'ils seront en nombre convenable, ils n'auront aucune raison *a priori* d'être liés.

Le paramètre λ est associé au vecteur composé entièrement de 1, cela $I.J$ fois.
Le jeu des I paramètres λ_i^A est associé à I vecteurs, la coordonnée numéro α $(1 \le \alpha \le IJ)$ du vecteur associé à λ_i^A valant 1 si et seulement si la coordonnée numéro α du vecteur l s'écrit l_{ij}, où j peut être quelconque.
Le jeu de paramètres μ_j est associé à J vecteurs, le jeu de paramètres δ_{ij} est associé à $I.J$ vecteurs, et ils sont construits selon la même logique.

La somme des vecteurs associés aux λ_i^A est égale au vecteur formé partout et seulement de 1. Idem pour les vecteurs λ_j^B. Il n'y a pas d'autres relations perturbatrices, et il est donc naturel de faire en sorte que ces deux relations ne soient plus vérifiées.

Pour que le nombre de vecteurs associés aux λ_i^A diminue de 1 (c'est-à-dire passe de I à $I-1$), il y a *grosso modo* deux méthodes. Soit on annule un (seul) coefficient λ_i^A ce qui fait tout simplement disparaître le vecteur associé. Soit on impose une (seule) relation entre les I coefficients λ_i^A, ce qui revient à

distinguer seulement $I-1$ coefficients indépendants, et donc $I-1$ vecteurs au final. Le cas 1 se traduit souvent par $\lambda_I^A=0$, le cas 2 se traduit par $\sum_{i=1}^{I}\lambda_i^A=0$.

De même, on réduit de 1 le nombre de vecteurs associés aux λ_j^B .

Le nombre de vecteurs libres associés aux paramètres $\lambda,\lambda_i^A,\lambda_j^B$ est égal à $1+(I-1)+(J-1)=I+J-1$, qui représente donc le nombre de vecteurs associés aux λ_{ij}^{AB} qui doivent disparaître (puisqu'ils sont eux-mêmes en nombre $I.J$). C'est aussi le nombre de contraintes à faire peser sur les λ_{ij}^{AB} , cela afin de parvenir *in fine* à un jeu de $I.J$ vecteurs libres. Un système de contraintes peut être

$$\lambda_{iJ}^{AB}=0 \quad \forall i=1,\ldots,I \text{ et } \lambda_{Ij}^{AB}=0 \quad \forall j=1,\ldots,J$$

ou comme alternative

$$\sum_{i=1}^{I}\lambda_{ij}^{AB}=0 \quad \forall j=1,\ldots,J \text{ et } \sum_{j=1}^{J}\lambda_{ij}^{AB}=0 \quad \forall i=1,\ldots,I .$$

Chaque cas traduit bien $I+J-1$ contraintes – en non $I+J$ - parce qu'il y a une (seule) redondance pour chaque alternative : $\lambda_{IJ}^{AB}=0$ dans le premier cas et $\sum_{i=1}^{I}\left(\sum_{j=1}^{J}\lambda_{ij}^{AB}\right)=\sum_{j=1}^{J}\left(\sum_{i=1}^{I}\lambda_{ij}^{AB}\right)$ dans le second.

Le 'bilan comptable' du nombre de vecteurs libres utilisés pour décomposer le vecteur l est donc :

$$1+(I-1)+(J-1)+[I.J-(I+J-1)]$$

qui vaut $I.J$ - ce que l'on souhaitait. Le nombre total de contraintes imposées (dites 'contraintes identifiantes') est égal à

$$1+1+(I+J-1)=I+J+1$$

ii) Cas de 3 variables qualitatives

Avec 3 variables qualitatives A, B et C, le paramétrage devient

$$l_{ijk}=\lambda+\lambda_i^A+\lambda_j^B+\lambda_k^C+\lambda_{ij}^{AB}+\lambda_{ik}^{AC}+\lambda_{jk}^{BC}+\lambda_{ijk}^{ABC}$$

En adaptant très fidèlement le raisonnement développé dans le cas de 2 variables – tous les principes restant les mêmes - on obtient des contraintes identifiantes en imposant pour chaque composante la nullité du coefficient dont l'un des indices prend sa valeur maximale, soit

$\lambda_I^A=0$; $\lambda_J^B=0$; $\lambda_K^C=0$

$\lambda_{Ij}^{AB}=\lambda_{iJ}^{AB}=0$ pour tout (i,j)

$\lambda_{Ik}^{AC}=\lambda_{iK}^{AC}=0$ pour tout (i,k)

$\lambda_{Jk}^{BC}=\lambda_{jK}^{BC}=0$ pour tout (j,k)

$\lambda_{Ijk}^{ABC}=\lambda_{iJk}^{ABC}=\lambda_{ijK}^{ABC}=0$ pour tout (i,j,k)

Vérifions que, dans ces conditions, le nombre de paramètres laissés 'libres' est correct :

i) nombre de paramètres à 1 seul indice laissés libres (y compris λ , qui compte pour 1)

$$1+(I-1)+(J-1)+(K-1)$$

ii) nombre de paramètres à 2 indices laissés libres

Adoptons une vision géométrique dans l'espace : pour les λ_{ij}^{AB} par exemple, on veut dénombrer les cases d'un rectangle de dimension $I.J$ dont on a retiré intégralement les 2 arrêtes. La première arrête contient I cases, la seconde J cases. Mais si on retire $I+J$ cases, on aura enlevé 2 fois la case constituant le coin du rectangle. Il faut donc rajouter cette case. Le dénombrement final est $I.J-I-J+1$.

Idem pour les autres paramètres, ce qui donne un dénombrement égal à

$$(I.J-I-J+1)+(I.K-I-K+1)+(J.K-J-K+1)$$

iii) nombre de paramètres à 3 indices laissés libres

On veut enlever les 3 faces (complètes) d'un parallélépipède. Un effort de vision dans l'espace permet de se convaincre d'emblée qu'il reste alors un parallélépipède de $(I-1).(J-1).(K-1)$ cases.

Au final, le nombre de paramètres laissés libres est égal à

$1+(I-1)+(J-1)+(K-1)+(I.J-I-J+1)+(I.K-I-K+1)+(J.K-J-K+1)+$
$(I-1).(J-1).(K-1)$, qui vaut $I.J.K$.

C’est exactement le nombre de valeurs observées composant le vecteur l, ce qui montre que les contraintes initiales sont en nombre convenable.

Noter que le nombre total de contraintes identifiantes est égal à $1+I+J+K+IJ+IK+JK$.

Bibliographie

AGRESTI, A, LANG, J, 1993, Quasi-symetric latent class models, with application to rater agreement, *Biometrics*, 1993, N° 49

BARTOLUCCI, F, FORCINA, A, 2017, In : BÖHNING, Dankmar, *Capture-Recapture Methods for the Social and Medical Sciences* [en ligne]. 1. Chapman and Hall/CRC. pp. 237-257. ISBN 978-1-315-15193-9. [Consulté le 11 février 2021].

BIRD, Sheila M. et KING, Ruth, 2018. Multiple Systems Estimation (or Capture-Recapture Estimation) to Inform Public Policy. *Annual Review of Statistics and Its Application* [en ligne]. 7 mars 2018. Vol. 5, n° 1, pp. 95-118. DOI 10.1146/annurev-statistics-031017-100641. [Consulté le 13 septembre 2022].

BRU, Bernard, 1988. Estimations laplaciennes. Un exemple : la recherche de la population d'un grand empire, 1785-1812. *Journal de la société statistique de Paris*. 1988. Vol. tome 129, n° no 1-2.

CHAO, A, TSAY, P, 1998, A sample coverage approach to multiple-system estimation with application to census undercount, *Journal of the American Statistical Association*. 1998. vol 93, N° 441

CHAO, A, TSAY, P, 2001, Population size estimation for capture-recapture models with applications to epidemiological data, *Journal of Applied Statistics*. 2001. vol 28, N° 1

CHAO, A, 2001, An overview of closed capture-recapture models, *Journal of Agricultural, Biological, and Environmental Statistics*, 2001, Vol 6 , N° 2

CHAO, Anne, 2015. Capture-Recapture for Human Populations. In : BALAKRISHNAN, N., COLTON, Theodore, EVERITT, Brian, PIEGORSCH, Walter, RUGGERI, Fabrizio et TEUGELS, Jozef L. (éd.), *Wiley StatsRef: Statistics Reference Online* [en ligne]. 1. Wiley. pp. 1-16. ISBN 978-1-118-44511-2. [Consulté le 20 septembre 2022].

CORMACK,R, 1989, Log-linear models for capture-recapture, *Biometrics*, 1989, vol 45, N° 2

DARROCH, J N, 1958. The Multiple-Recapture Census: I. Estimation of a Closed Population. *Biometrika*. 1958. pp. 18.

DARROCH, J , FINBERG, S, GLONEK, G, JUNKER, B, 1993, A three-sample multiple-recapture approach to Census population estimation with heterogeneous catchability, *Journal of the American Statistical Association*. 1993. vol 88, N° 423

DEMPSTER, A, LAIRD, N, RUBIN, D, 1977, Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society ; Series B 1977*. Vol. 39, n° 1.

DE WOLF, Peter-Paul, 2019. Connecting Correction Methods for Linkage Error in Capture-Recapture. *Journal of Official Statistics*. 2019. pp. 21.

DING, Ye et FIENBERG, Stephen E., 1994. Dual System Estimation of Census undercount in the Presence of Matching Errors. *Survey Methodology*. 1994.

FIENBERG, Stephen, 1992. Bibliography on capture-recapture modelling with application to census undercount adjustment. *Survey Methodology* [en ligne]. 1992. Disponible à l'adresse : https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1992001/article/14494-eng.pdf?st=ir95w5bQ [Consulté le 12 septembre 2022].

GERRITSE, S, VAN DER HEIJDEN, P , BAKKER, B, 2015, Sensitivity of population size estimation for violating parametric assumptions in log-linear models, *Journal of Official Statistics*, 2015, Vol 31, N°3

GOODMAN, L, 1974, Exploratory latent structure analysis using both identifiable and unidentifiable models, *Biometrika*, 1974, N° 61 - 2

GOUDIE, I. B. J. et GOUDIE, M., 2007. Who captures the marks for the Petersen estimator? *Journal of the Royal Statistical Society: Series A (Statistics in Society)* [en ligne]. juillet 2007. Vol. 170, n° 3, pp. 825-839. DOI 10.1111/j.1467-985X.2007.00479.x. [Consulté le 19 septembre 2022].

WOLTER, Kirk M, 1986. Some Coverage Error Models for Census Data. *Journal of the American Statistical Association*. 1986. pp. 10.

ZHANG, Li-Chun, 2015, On modelling register coverage errors, *Journal of Official Statistics*, 2015, Vol 31, N°3

ZHANG, Li-Chun et DUNNE, John, 2017. Trimmed dual system estimation. In : BÖHNING, Dankmar, *Capture-Recapture Methods for the Social and Medical Sciences* [en ligne]. 1. Chapman and Hall/CRC. pp. 237-257. ISBN 978-1-315-15193-9. [Consulté le 11 février 2021].

ZHANG, Li-Chun, 2019, A note on dual system population size estimator, *Journal of Official Statistics*, 2019, Vol 35, N°1

ZULT, Daan, DE WOLF, Peter-Paul, BAKKER, Bart F. M. et VAN DER HEIJDEN, Peter, 2021. A General Framework for Multiple-Recapture Estimation that Incorporates Linkage Error Correction. *Journal of Official Statistics* [en ligne]. 1 septembre 2021. Vol. 37, n° 3, pp. 699-718. DOI 10.2478/jos-2021-0031. [Consulté le 15 septembre 2021].

Série des Documents de Travail « Méthodologie Statistique »

9601 : Une méthode synthétique, robuste et efficace pour réaliser des estimations locales de population.
G. DECAUDIN, J.-C. LABAT

9602 : Estimation de la précision d'un solde dans les enquêtes de conjoncture auprès des entreprises.
N. CARON, P. RAVALET, O. SAUTORY

9603 : La procédure FREQ de SAS - Tests d'indépendance et mesures d'association dans un tableau de contingence.
J. CONFAIS, Y. GRELET, M. LE GUEN

9604 : Les principales techniques de correction de la non-réponse et les modèles associés.
N. CARON

9605 : L'estimation du taux d'évolution des dépenses d'équipement dans l'enquête de conjoncture : analyse et voies d'amélioration.
P. RAVALET

9606 : L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT).
S. LOLLIVIER, M. MARPSAT, D. VERGER

9607 : Enquêtes régionales sur les déplacements des ménages : l'expérience de Rhône-Alpes.
N. CARON, D. LE BLANC

9701 : Une bonne petite enquête vaut-elle mieux qu'un mauvais recensement ?
J.-C. DEVILLE

9702 : Modèles univariés et modèles de durée sur données individuelles.
S. LOLLIVIER

9703 : Comparaison de deux estimateurs par le ratio stratifiés et application aux enquêtes auprès des entreprises.
N. CARON, J.-C. DEVILLE

9704 : La faisabilité d'une enquête auprès des ménages.
1. au mois d'août.
2. à un rythme hebdomadaire

C. LAGARENNE, C. THIESSET

9705 : Méthodologie de l'enquête sur les déplacements dans l'agglomération toulousaine.
P. GIRARD.

9801 : Les logiciels de désaisonnalisation TRAMO & SEATS : philosophie, principes et mise en œuvre sous SAS.
K. ATTAL-TOUBERT, D. LADIRAY

9802 : Estimation de variance pour des statistiques complexes : technique des résidus et de linéarisation.
J.-C. DEVILLE

9803 : Pour essayer d'en finir avec l'individu Kish.
J.-C. DEVILLE

9804 : Une nouvelle (encore une !) méthode de tirage à probabilités inégales.
J.-C. DEVILLE

9805 : Variance et estimation de variance en cas d'erreurs de mesure non corrélées ou de l'intrusion d'un individu Kish.
J.-C. DEVILLE

9806 : Estimation de précision de données issues d'enquêtes : document méthodologique sur le logiciel POULPE.
N. CARON, J.-C. DEVILLE, O. SAUTORY

9807 : Estimation de données régionales à l'aide de techniques d'analyse multidimentionnelle.
K. ATTAL-TOUBERT, O. SAUTORY

9808 : Matrices de mobilité et calcul de la précision associée.
N. CARON, C. CHAMBAZ

9809 : Échantillonnage et stratification : une étude empirique des gains de précision.
J. LE GUENNEC

9810 : Le Kish : les problèmes de réalisation du tirage et de son extrapolation.
C. BERTHIER, N. CARON, B. NEROS

9901 : Perte de précision liée au tirage d'un ou plusieurs individus Kish.
N. CARON

9902 : Estimation de variance en présence de données imputées : un exemple à partir de l'enquête Panel Européen.
N. CARON

0001 : L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT) (version actualisée).
S. LOLLIVIER, M. MARPSAT, D. VERGER

0002 : Modèles structurels et variables explicatives endogènes.
J.-M. ROBIN

0003 : L'enquête 1997-1998 sur le devenir des personnes sorties du RMI - Une présentation de son déroulement.
D. ENEAU, D. GUILLEMOT

0004 : Plus d'amis, plus proches ? Essai de comparaison de deux enquêtes peu comparables.
O. GODECHOT

0005 : Estimation dans les enquêtes répétées : application à l'Enquête Emploi en Continu.
N. CARON, P. RAVALET

0006 : Non-parametric approach to the cost-of-living index.
F. MAGNIEN, J. POUGNARD

0101 : Diverses macros SAS : Analyse exploratoire des données, Analyse des séries temporelles.
D. LADIRAY

0102 : Économétrie linéaire des panels : une introduction.
T. MAGNAC

0201 : Application des méthodes de calages à l'enquête EAE-Commerce.
N. CARON

C 0201 : Comportement face au risque et à l'avenir et accumulation patrimoniale - Bilan d'une expérimentation.
L. ARRONDEL, A. MASSON, D. VERGER

C 0202 : Enquête Méthodologique Information et Vie Quotidienne - Tome 1 : bilan du test 1, novembre 2002.
J.-A. VALLET, G. BONNET, J.-C. EMIN, J. LEVASSEUR, T. ROCHER, P. VRIGNAUD, X. D'HAULTFOEUILLE, F. MURAT, D. VERGER, P. ZAMORA

0203 : General principles for data editing in business surveys and how to optimise it.
P. RIVIERE

0301 : Les modèles logit polytomiques non ordonnés : théories et applications.
C. AFSA ESSAFI

0401 : Enquête sur le patrimoine des ménages - Synthèse des entretiens monographiques.
V. COHEN, C. DEMMER

0402 : La macro SAS CUBE d'échantillonnage équilibré
S. ROUSSEAU, F. TARDIEU

0501 : Correction de la non-réponse et calage de l'enquêtes Santé 2002
N. CARON, S. ROUSSEAU

0502 : Correction de la non-réponse par répondération et par imputation
N. CARON

0503 : Introduction à la pratique des indices statistiques - notes de cours
J-P BERTHIER

0601 : La difficile mesure des pratiques dans le domaine du sport et de la culture - bilan d'une opération méthodologique
C. LANDRE, D. VERGER

0801 : Rapport du groupe de réflexion sur la qualité des enquêtes auprès des ménages
D. VERGER

M2013/01 : La régression quantile en pratique
P. GIVORD,
X. D'HAULTFOEUILLE

M2014/01 : La microsimulation dynamique : principes généraux et exemples en langage R
D. BLANCHET

M2015/01 : la collecte multimode et le paradigme de l'erreur d'enquête totale
T. RAZAFINDROVONA

M2015/02 : Les méthodes de Pseudo-Panel
M. GUILLERM

M2015/03 : Les méthodes d'estimation de la précision pour les enquêtes ménages de l'Insee tirées dans Octopusse
E. GROS
K. MOUSSALAM

M2016/01 : Le modèle Logit Théorie et application.
C. AFSA

M2016/02 : Les méthodes d'estimation de la précision de l'Enquête Emploi en Continu
E. GROS
K.MOUSSALAM

M2016/03 : Exploitation de l'enquête expérimentale Vols, violence et sécurité.
T. RAZAFINDROVONA

M2016/04 : Savoir compter, savoir coder. Bonnes pratiques du statisticien en programmation.
E. L'HOUR
R. LE SAOUT
B. ROUPPERT

M2016/05 : Les modèles multiniveaux
P. GIVORD
M. GUILLERM

M2016/06 : Econométrie spatiale : une introduction pratique
P. GIVORD
R. LE SAOUT

M2016/07 : La gestion de la confidentialité pour les données individuelles
M. BERGEAT

M2016/08 : Exploitation de l'enquête expérimentale Logement internet-papier
T. RAZAFINDROVONA

M2017/01: Exploitation de l'enquête expérimentale Qualité de vie au travail
T. RAZAFINDROVONA

M2018/01: Estimation avec le score de propension sous R
S. QUANTIN

M2018/02: Modèles semi-paramétriques de survie en temps continu sous R
S. QUANTIN

M2019/01 : Les méthodes de décomposition appliquées à l'analyse des inégalités
B. BOUTCHENIK
E. COUDIN
S. MAILLARD

M2020/01 : L'économétrie en grande dimension
J. L'HOUR

M2021/01 : R Tools for JDemetra+ - Seasonal adjustment made easier
A. SMYK
A. TCHANG

M2021/02 : Le traitement du biais de sélection endogène dans les enquêtes auprès des ménages par modèle de Heckman
L. CASTELL
P. SILLARD

M2021/03 :
Conception de questionnaires auto-administrés
H. KOUMARIANOS
A. SCHREIBER

M2022/01 : Introduction à la géomatique pour le statisticien : quelques concepts et outils innovants de gestion, traitement et diffusion de l'information spatiale
F. SEMECURBE
E. COUDIN

M2022/02 : Le zonage en unites urbaines 2020
V. COSTEMALLE
S. OUJIA
C. GUILLO
A. CHAUVET

M2023/01 : Les réseaux de neurones appliqués à la statistique publique : méthodes et cas d'usages
D. BABET
Q. DELTOUR
T. FARIA
S. HIMPENS

M2023/02 : Redressements de la première vague de l'enquête epicov : un exemple de correction des effets de sélection dans les enquêtes multimodes
L.CASTELL
C. FAVRE-MARTINOZ
N. PALIOD
P. SILLARD

M2023/03 : Appariements de données individuelles : concepts, méthodes, conseils
L.MALHERBE

M2023/04 : Victimations déclarées et effets de mode : enseignements de l'expérimentation panel multimode de l'enquête cadre de vie et sécurité
L. CASTELL
M. CLERC
D. CROZE
S. LEGLEYE
A. NOUGARET

M2024/01 : Estimation en temps réel de la tendance-cycle : apport de l'utilisation des filtres asymétriques dans la détection des points de retournement
A.QUARTIER-LA-TENTE

M2024/02 : La disponibilité des coordonnées de contact dans fidéli-nautile - quels enseignements pour les protocoles de collecte ?
G. CHARRANCE (INED)

M2024/03 : Discuter l'existence d'un effet de sélection dans un cadre multimode grâce à une analyse de sensibilité - Application aux enquêtes annuelles de recensement
L. COURT
S. QUANTIN

M2024/04 : Vers une désaisonnalisation des séries temporelles infra-mensuelles avec JDemetra+
A. SMYK
K. WEBEL

M2025/01 : Les estimations par capture-recapture ou par système multiple : quelques éléments théoriques
P. ARDILLY
H. KOUMARIANOS