DÉCOUPAGE OPTIMAL D'UNE VARIABLE QUANTITATIVE POUR LA STRATIFICATION : UNE EXPÉRIENCE SUR LES DONNÉES D'ENTREPRISES FRANÇAISES ET UNE MISE EN ŒUVRE POUR L'ÉCHANTILLONNAGE DES ENQUÊTES MENSUELLES DE BRANCHES

Laurent COSTA[1] (*), Arnaud FIZZALA[2] (**)

(*) Insee, Direction de la méthodologie et de la coordination statistique et internationale
(**)Direction de la recherche, des études, de l'évaluation et des statistiques, Sous-direction observation de la solidarité, Bureau handicap et dépendance.

Résumé

Le découpage en tranches d'une variable quantitative pour élaborer des strates fait l'objet de plusieurs articles scientifiques, notamment lorsque cette variable possède une distribution asymétrique, ce qui est le cas des variables d'effectif salarié et de chiffre d'affaires utilisées pour caractériser la taille des entreprises lors de l'élaboration de plans de sondage d'enquêtes auprès des entreprises. Toutefois de telles méthodes de stratification sont aujourd'hui rarement mises en œuvre à l'Insee où les tranches d'effectifs utilisées pour caractériser la taille des entreprises sont en général les mêmes d'une enquête à l'autre.

La première partie de l'article est consacrée à la présentation des trois méthodes de découpage optimal de strates implémentées dans le package R « *stratification* » à savoir, les méthodes de Dalenius, géométrique, Lavalée et Hidiroglou. Cette présentation est illustrée par une application sur les données correspondant au champ habituel des enquêtes françaises auprès d'entreprises. L'application consiste à découper la variable « effectif salarié » en sept tranches de façon à minimiser la taille d'échantillon nécessaire pour atteindre une précision fixée a priori de l'estimateur de l'effectif salarié de l'ensemble de la population. L'utilisation des méthodes de découpage optimal de strates peut permettre de réduire les tailles d'échantillon nécessaires pour atteindre un objectif de précision fixé a priori. Ces réductions sont d'autant plus prononcées que les objectifs de précision sont ambitieux, mais il faut veiller à ne pas rendre le plan de sondage trop « spécifique » au critère d'optimisation.

La deuxième partie de l'article s'intéresse à une application de ces méthodes de découpage, notamment celle de Dalenius, aux Enquêtes Mensuelles de Branches. En effet, dans le cadre du projet Ocapi (Observation conjecturelle de l'activité productive industrielle), une étude du plan de sondage des EMB a été demandée à la section échantillonnage. Il en a suivi une étude sur les différents types de stratification possibles selon le nombre d'entreprises fabriquant un produit ainsi que leur chiffre d'affaires. Le but de l'étude est d'améliorer la précision des estimations obtenues avec l'échantillon tout en conservant sa taille. Nous verrons ainsi qu'avec un taux de couverture en montant plus faible (50%) et une stratification optimisée associée à une allocation proportionnelle au chiffre d'affaires, nous pouvons améliorer la méthode actuelle.

[1] laurent.costa@insee.fr
[2] arnaud.fizzala@sante.gouv.fr

Abstract

The establishment of strata is an important topic in survey sampling, especially when the stratification variable has a skewed distribution, which is frequent in business surveys. The optimization of stratum boundaries is rarely used at Insee. We generally use the same boundaries for all the business surveys.
The first part of this paper presents three methods of optimization of stratum boundaries implemented in the R-package *Stratification*. We apply the methods on French businesses data. The results are good but we have to be careful to do not make the sample design too specific to the optimization criteria.
The second part of this paper presents is focused on an application of these methods on the monthly surveys of industry. This study made us improve the actual survey method by selecting an optimised stratification which gives us a better precision of the estimations without changing the size of the sample.

Mots-clés

Échantillonnage, stratification, précision.

Sommaire

Introduction

Le découpage en tranches d'une variable quantitative pour élaborer des strates fait l'objet de plusieurs articles scientifiques, notamment lorsque cette variable possède une distribution asymétrique, ce qui est le cas des variables d'effectif salarié et de chiffre d'affaires utilisées pour caractériser la taille des entreprises lors de l'élaboration de plans de sondage d'enquêtes auprès des entreprises.

Dans la première partie de cet article, nous présentons les trois méthodes de découpage optimal de strates implémentées dans le package R *stratification* mis à disposition sur internet par Sophie Baillargeon et Louis-Paul Rivest de l'université Laval au Canada. Nous les illustrons par une application sur les données correspondant au champ habituel des enquêtes françaises auprès d'entreprises. L'application consiste à découper la variable effectif salarié en sept tranches de façon à minimiser la taille d'échantillon nécessaire pour atteindre une précision fixée a priori de l'estimateur de l'effectif salarié de l'ensemble de la population.

Dans la deuxième partie de cet article, nous allons nous concentrer sur la mise en œuvre de ces théories de découpage de strate sur l'échantillonnage des Enquêtes Mensuelles de Branches. L'étude s'appuie uniquement sur les données concernant 4 produits ciblés ainsi que les 14 produits les plus importants des EMB. Le but est d'améliorer la précision des estimations obtenues avec l'échantillon tout en conservant sa taille. Nous verrons ainsi qu'avec une stratification optimisée associée à une allocation proportionnelle au chiffre d'affaires, nous pouvons améliorer la méthode actuelle.

1. Les méthodes de découpage optimal de strates : une expérience sur les données d'entreprises françaises

Les échantillons des enquêtes auprès des entreprises menées par l'Insee sont le plus souvent tirés selon des plans de sondage aléatoires simples stratifiés (*Demoly, Fizzala, Gros, 2014*). La population d'entreprises correspondant au champ de l'enquête est découpée en strates construites à partir de caractéristiques des entreprises. Le plus souvent, la population d'entreprises est stratifiée en croisant deux[3] critères : un critère d'activité (utilisant des agrégats de la nomenclature d'activités françaises) et un critère de taille (utilisant des tranches d'effectifs salariés et/ou des tranches de chiffres d'affaires). Les entreprises de plus grande taille possèdent un tel poids économique qu'elles sont généralement interrogées exhaustivement.

Le critère de taille le plus couramment utilisé à l'Insee est l'effectif salarié. Les tranches utilisées pour tirer les échantillons sont en général les mêmes d'une enquête à l'autre (1 à 9 salariés, 10 à 19, 20 à 49, 50 à 249, 250 à 499, 500 et plus). L'utilisation de ces tranches se justifie lorsque l'enquête a pour objectif la production de statistiques par tranche de tailles d'entreprises. En effet, stratifier selon ces tranches permet de maîtriser le nombre d'entreprises qui y seront échantillonnées et donc de contraindre le plan de sondage afin d'y viser une précision pour la publication de nos statistiques. Par contre, lorsque l'enquête n'a pas pour objectif de produire des statistiques sur ces tranches d'effectif, on peut se demander si d'autres tranches ne permettraient pas d'obtenir une meilleure précision pour les estimateurs finaux. Or, en pratique, ce sont les tranches d'effectifs citées ci-dessus qui sont le plus souvent utilisées à l'Insee.

La question plus générale du découpage optimal d'une variable X quantitative connue sur l'ensemble de la base de sondage afin d'élaborer une stratification fait l'objet de plusieurs articles scientifiques, notamment lorsque X possède une distribution asymétrique, ce qui est le cas ici[4]. Il n'existe pas de

[3] Un troisième critère de localisation géographique est également souvent utilisé pour le tirage mais il n'est, en général, pas pris en compte lors de l'optimisation du plan de sondage, il conduit en effet à des strates comportant un nombre d'unités trop faible pour calculer des statistiques robustes.

[4] Cas notamment des effectifs salariés ou des chiffres d'affaires pour les populations d'entreprises : les quelques très grandes entreprises d'une population ont une influence capitale sur le total de la variable à l'étude.

solution facilement calculable du fait de la dépendance entre chaque borne de strate. Ainsi, plusieurs auteurs proposent différentes méthodes servant à approcher la stratification optimale.

Dalenius et Hodges (*Dalenius, Hodges, 1959*) ont proposé la règle du cumul des racines carrées des fréquences. Il s'agit de la règle d'optimisation de limites de strates la plus connue, mais elle n'est que peu utilisée à l'Insee. Elle semble simple à mettre en œuvre mais elle se base sur plusieurs hypothèses (facteur correctif de population finie négligeable, distribution uniforme de la variable X dans chaque strate...) qui ne sont en général pas valables pour les populations d'entreprises.

Lavallée et Hidiroglou (*Lavallée, Hidiroglou, 1988*) ont proposé une méthode itérative, plus complexe à mettre en œuvre mais qui semble donner de meilleurs résultats. Cet algorithme permet notamment de prendre en compte une strate exhaustive, caractéristique importante des plans de sondage entreprises.

Gunning et Horgan (*Gunning, Horgan, 2004*) ont proposé la méthode géométrique, qui, d'après leur article, serait souvent plus performante que la règle des racines carrées cumulées des fréquences ainsi que l'algorithme de Lavallée et Hidiroglou, mais pas de façon systématique. On trouve en effet dans d'autres articles (*Kozak, Rad-Verma, 2006*) des résultats différents sur les comparaisons des différentes méthodes[5].

Dans cette partie de l'article, nous décrirons trois méthodes implémentées dans le package R *stratification (Baillargeon, Rivest, 2011)* :

- La méthode de Dalenius ;
- La méthode géométrique ;
- La méthode LH (Lavallée - Hidiroglou).

On illustrera l'application de chacune des méthodes à partir d'un exemple basé sur les données d'entreprises françaises correspondant au champ habituel des enquêtes entreprises. Dans cette illustration, nous nous placerons dans un cadre simplifié. En particulier, nous ne distinguerons pas les différents secteurs d'activité des entreprises[6], nous ne tiendrons pas compte des phénomènes de non réponse pour nos calculs et nous nous limiterons au cas où la variable de stratification est confondue avec la variable d'intérêt[7]. Nous discuterons ces hypothèses simplificatrices dans la partie 1.7.

1.1. Description des données utilisées

Pour illustrer l'application des trois méthodes de découpage optimal de strates, on utilisera les données d'entreprises françaises correspondant au champ habituel des enquêtes entreprises. Plus précisément, il s'agit de données extraites du référentiel[8] Sirus au 31/12/2012, correspondant aux unités légales :

- de 10 salariés ou plus ;
- marchandes ;
- exploitantes ;
- localisées en France (métropole ou l'un des cinq départements d'Outre-mer).

Cette population regroupe 209 089 unités.

Parmi elles, 129 très grandes unités emploient 5 000 salariés ou plus. Ces très grandes unités seront incluses d'office dans une strate exhaustive[9] dans nos calculs.

---

[5] Nous verrons d'ailleurs qu'avec la façon dont est implémentée la méthode itérative dans le package R *stratification*, cette dernière est systématiquement meilleure que la méthode géométrique.

[6] En pratique, la population est souvent stratifiée selon le secteur d'activité.

[7] On espère en pratique que le lien entre la variable de stratification et la variable d'intérêt est suffisamment fort pour que les conclusions restent valables. Certains liens particuliers entre la variable de stratification et la variable d'intérêt peuvent être pris en compte dans les méthodes de découpage (*Rivest 2002*) mais ne seront pas étudiés ici.

[8] Version du 7 juillet 2013

[9] En pratique, les seuils d'exhaustivité des enquêtes entreprises se situent souvent plus bas (500, 250, voire moins pour les plus « grandes » enquêtes), mais on verra que des seuils si bas ne sont pas forcément nécessaires dans le cadre de notre application. En revanche, instaurer un seuil, même très haut, permet d'utiliser plus facilement les méthodes de découpage optimal de Dalenius et géométrique en raison de leur dépendance aux valeurs extrêmes prises par la population. Le seuil de

Dans la suite, pour l'illustration de l'utilisation des méthodes de découpage optimal, on notera :

N = 209 089 : le nombre d'entreprises dans notre population ;
k : une entreprise de notre population ;
y_k : le nombre de salariés de l'unité k (c'est notre variable d'intérêt) ;

t_y : le nombre total de salariés dans la population (c'est notre paramètre d'intérêt). $t_y = \sum_{k=1}^{N} y_k$

$\hat{t}_y = \sum_{h=1}^{H} N_h \cdot \bar{y}_h$: l'estimateur de t_y

On s'intéressera à la précision de $\hat{t}_y$ mesurée à partir de son coefficient de variation : $Cv = \frac{\sqrt{V(\hat{t}_y)}}{t_y}$.

Les unités de notre population se répartissent, selon les tranches d'effectif habituelles de la façon suivante :

Tranche d'effectif	Nombre d'unités	Somme des effectifs	Dispersion[10] des effectifs
10 à 19	109 232	1 455 314	2,8
20 à 49	64 673	1 980 756	8,4
50 à 99	18 704	1 267 209	14,0
100 à 249	10 468	1 599 923	41,0
250 à 499	3 316	1 138 863	68,5
500 à 999	1 483	1 013 769	135,5
1 000 à 4 999	1 084	2 069 807	902,9
5 000 et plus	129	1 864 320	20 433,1
Total	**209 089**	**12 389 961**	**641,5**

Cette répartition selon les tranches d'effectif habituelles permet de se donner une première idée de l'asymétrie de la distribution de la variable effectif salarié dans notre population. Plus de la moitié des entreprises comportent entre 10 et 19 salariés et ne représentent pourtant « que » 10% de l'ensemble des salariés. 40% des salariés appartiennent aux entreprises de 500 salariés et plus alors que ces entreprises ne représentent qu'un peu plus de 1% des entreprises.

Les premières tranches d'effectif apparaissent peu dispersées par rapport aux suivantes et l'on imagine assez bien les gains de précision obtenus en stratifiant notre population avec ce premier découpage en tranches. Concrètement, il est possible de calculer la taille d'échantillon nécessaire à l'obtention d'un coefficient de variation de x% dans la situation d'un sondage aléatoire simple et dans celle d'un sondage stratifié[11] avec les tranches d'effectifs classiques. On obtient le tableau ci-dessous :

---

5 000 salariés semblait adapté puisqu'il concerne peu d'unités. Il correspond au seuil d'effectif définissant les grandes entreprises selon le décret d'application (n°2008-1354) de l'article 51 de la loi de modernisation de l'économie, relatif aux critères permettant de déterminer la catégorie d'appartenance d'une entreprise pour les besoins de l'analyse statistique et économique.

[10] on entend par dispersion d'une variable son écart-type

[11] sous l'hypothèse supplémentaire que la répartition de l'échantillon entre les strates est réalisée selon une allocation de Neyman

Tailles d'échantillons nécessaires pour obtenir différents coefficients de variation pour l'estimation de l'effectif total avec et sans stratification

Cv	Sans strate exhaustive		En sondant exhaustivement les entreprises de 5 000 salariés ou plus	
	Taille d'échantillon nécessaire avec sondage aléatoire simple	Taille d'échantillon nécessaire avec Sondage stratifié par tranche d'effectif	Taille d'échantillon nécessaire avec Sondage aléatoire simple	Taille d'échantillon nécessaire avec Sondage stratifié par tranche d'effectif
1%	177 435	1 468	57 922	666
2%	122 018	966	18 359	272
3%	80 247	615	8 644	195
4%	54 247	409	5 005	168
5%	38 295	285	3 276	156
6%	28 170	209	2 325	149
7%	21 464	159	1 747	144
8%	16 838	125	1 370	140
9%	13 533	100	1 111	140
10%	11 098	84	925	138

Comme on le présumait à la vue du tableau précédant, l'établissement de la stratification classique permet de réduire considérablement les tailles d'échantillon nécessaires pour atteindre différents coefficients de variation.

On peut constater dans les deux dernières colonnes que l'interrogation exhaustive des entreprises de 5 000 salariés ou plus améliore les résultats, sauf pour les coefficients de variation les plus forts.

Dans la suite, on va chercher à savoir s'il existe des bornes de stratification qui permettraient de réduire encore les tailles d'échantillon nécessaires pour atteindre les différents coefficients de variation en utilisant les méthodes optimales de découpage implémentées dans le package R *stratification*.

Afin de comparer convenablement les tranches d'effectif obtenues aux tranches d'effectifs classiques, on se placera dans le cas de 7 strates (L=7)[12] dont les limites sont à optimiser et d'une strate exhaustive constituée des entreprises de 5 000 salariés ou plus.

1.2. Formalisation du problème du découpage optimal des strates

Soit x_k la variable quantitative connue sur l'ensemble de la population et retenue pour la stratification et L le nombre de strates visées.

Soit y_k la variable d'intérêt dont on cherche à estimer le total t_y.

Nous supposons[13] dans la suite $y_k=x_k$.

Soit Y le vecteur composé des N valeurs y_k prises dans la population ordonnées par valeurs croissantes. On pose $b_0=y_0$ et $b_L=y_N$.

[12] Le nombre de tranches est discuté dans la partie 1.6.

[13] Comme dit précédemment, on espère en pratique que le lien entre x_k et y_k est suffisamment fort pour que les conclusions restent valables.

Une unité k appartient à la strate h si $y_k \in [b_{h-1}; b_h[$

On cherche les bornes $b_1, b_2, .., b_{L-1}$ qui minimisent $V(\hat{t}_y) = \sum_{h=1}^{L} N_h^2 (1 - f_h) \frac{S_{yh}^2}{n_h}$.

1.3. La méthode Dalenius

Dalenius et Hodges ont proposé cette méthode dans l'article *Minimum Variance Stratification* paru en mars 1959 dans l'American statistical association journal. Il s'agit de la méthode de découpage optimal de strate la plus connue. Elle est utilisée à l'Insee pour l'enquête mensuelle sur les grandes surfaces alimentaires[14] et pour l'enquête mensuelle de branche pour le secteur de la « mécanique industrielle ».

1.3.1. Conditions devant être vérifiées par les limites optimales des strates

Les raisonnements qui suivent s'appuient sur l'hypothèse que les taux de sondage dans chaque strate sont proches de zéro. Cette hypothèse est rarement vérifiée dans les enquêtes auprès des entreprises où les taux de sondage sont souvent élevés.

En supposant $f_h=0$, la formule de la variance de l'estimateur du total de la variable y devient :

$$V(\hat{t}_y) \approx \sum_{h=1}^{L} N_h^2 \frac{S_{yh}^2}{n_h}$$

Avec l'allocation de Neyman, on obtient :

$$V_{NEY}(\hat{t}_y) \approx \frac{1}{n} \left(\sum_{h=1}^{L} N_h S_{yh} \right)^2$$

Trouver le découpage en strates qui minimise $V_{NEY}(\hat{t}_y)$ est alors indépendant de la taille d'échantillon n et revient à trouver $b_1, ..., b_{L-1}$ minimisant $\sum_{h=1}^{L} N_h S_{yh}$.

Dalenius montre que chaque borne b_h doit vérifier :

$$\frac{(b_h - \mu_{yh})^2 + S_{yh}^2}{S_{yh}} = \frac{(b_h - \mu_{yh+1})^2 + S_{yh+1}^2}{S_{yh+1}} \quad (1)$$

Comme S_{yh} et u_{yh} dépendent de la borne b_h, cette équation ne peut pas se résoudre facilement.

Dalenius propose de l'approcher par un algorithme en introduisant des hypothèses supplémentaires.

1.3.2. Le cumul des racines carrées des fréquences

L'algorithme permettant d'approcher la solution de l'équation (1) le plus utilisé est basé sur le cumul des racines carrées des fréquences. Il implique l'hypothèse supplémentaire que les y_k sont repartis

[14] Dans une variante différente de celle implémentée dans le package R *stratification*

selon une loi uniforme sur la strate h. Dalenius montre que sous cette hypothèse, minimiser $\sum_{h=1}^{L} N_h S_{yh}$ revient à trouver des bornes b_h telles que $\sqrt{N_h}(b_h - b_{h-1}) = c$ où c est une constante.

1.3.3. L'algorithme en pratique

On connait la distribution des y_k selon un découpage fin en J classes de mêmes longueurs[15], et donc les fréquences N_j et les racines carrées des fréquences $\sqrt{N_j}$ associées à ce découpage.

On va alors déterminer la première strate en regroupant les J_1 premières classes de façon à ce que $\sum_{j=1}^{J_1} \sqrt{N_j} \approx \frac{\sum_{j=1}^{J} \sqrt{N_j}}{L}$ (avec L le nombre de strates cherchées).

La deuxième strate regroupera les J_2 classes suivantes de façon à ce que :

$$\sum_{j=J_1+1}^{J_1+J_2} \sqrt{N_j} \approx \frac{\sum_{j=J_1+1}^{J_1+J_2} \sqrt{N_j}}{L}$$

etc…

Lorsque les classes ne sont pas de mêmes longueurs, l'algorithme peut en théorie être adapté (*Cochran, 1977*), mais cette possibilité n'est pas implémentée dans le package *stratification*.

1.3.4. Le choix des J classes initiales

Le résultat final dépend à la fois du nombre et du choix des J classes initiales. Il n'y a pas de résultat théorique sur la meilleure valeur pour J. Notamment, on remarque qu'avec un nombre de classes infini (et des valeurs différentes pour chaque unité), l'algorithme reviendrait à créer des strates de tailles N_h égales, ce qui présente a priori peu d'intérêt.

Par défaut, le package *stratification* fixe :

- J=15L ou le nombre de valeurs distinctes pour y si ce dernier est inférieur à 15L.
- J classes de longueur égale $l = \frac{b_L - b_0}{J}$

Le nombre de classes J peut être modifié par l'utilisateur. En revanche, il n'est pas possible d'imposer des classes initiales en entrée du programme autrement qu'en jouant sur le nombre de classes. En particulier, il n'est pas possible d'imposer des classes de longueurs différentes.

1.3.5. Exemple d'application

En appliquant la méthode sur notre population d'entreprises[16] sur 500[17] classes initiales, on obtient les strates suivantes :

---

[15] On entend par longueur d'une strate h la quantité $l_h = b_h - b_{h-1}$

[16] La méthode est appliquée en dehors de la strate exhaustive a priori (celle contenant les entreprises de 5 000 salariés et plus).

[17] Avec le nombre de classes choisi par défaut par le package, les classes ne sont pas assez nombreuses (105) et l'algorithme ne fonctionne pas.

Tranche d'effectif	Nombre d'unités	Somme des effectifs	Dispersion des effectifs
10 à 19	109 232	1 455 314	2,8
20 à 39	52 250	1 431 870	5,7
40 à 89	29 230	1 637 237	13,4
90 à 219	11 402	1 554 005	35,9
220 à 528	4 441	1 446 701	82,9
529 à 1 355	1 707	1 380 481	229,1
1356 à 4 999	698	1 620 033	885,5
5 000 et plus	129	1 864 320	20 433,1
Total	**209 089**	**12 389 961**	**641,5**

Ces strates conduisent aux précisions suivantes :

Cv	Taille d'échantillon nécessaire		
	Sondage aléatoire simple	Sondage stratifié classique[18]	Sondage stratifié méthode Dalenius
1%	57 922	666	615
2%	18 359	272	257
3%	8 644	195	188
4%	5 005	168	164
5%	3 276	156	152
6%	2 325	149	146
7%	1 747	144	144
8%	1 370	140	141
9%	1 111	140	137
10%	925	138	137

L'optimisation des tranches d'effectif conduit à de légers (relativement à ceux obtenus en passant de sondage aléatoire simple à stratifié classique) gains de taille d'échantillon nécessaire pour atteindre les différents Cv. Plus l'on cherche à être précis (petites valeurs de Cv) et plus les gains obtenus par une optimisation des limites des strates sont importants.

Avec 500 classes, l'algorithme se déroule de la façon suivante.

On calcule tout d'abord la longueur des classes $l = \frac{4993-10}{500} = 9{,}966$ et on en déduit les bornes de classes initiales associées (19,966 ; 29,932 ; 39,898 ; 49,864 ; 59,830 … 4 983,034).

On aboutit ainsi aux 500 classes suivantes :

[18] On entend par stratification classique celle définie dans la partie 1.1 correspondant aux tranches 10 à 19 ; 20 à 49 ; 50 à 99 ; 100 à 249 ; 250 à 499 ; 500 à 999 ; 1 000 à 4 999 ; 5 000 et plus.

Classe	N_h	$\sqrt{N_h}$
10 à 19	109 232	331
20 à 29	33 979	184
30 à 39	18 271	135
40 à 49	12 423	111
50 à 59	6 937	83
60 à 69	4 296	66
70 à 79	3147	56
80 à 89	2427	49
90 à 99	1897	43
…		
4 984 à 4 993	1	1

En cumulant les racines carrées des fréquences par classe et en divisant par le nombre de strates, on obtient $\frac{\sum_{j=1}^{J}\sqrt{N_j}}{L} \approx 355{,}96$.

L'algorithme va maintenant chercher à regrouper les classes de façon à créer des strates telles que $\sum_{j\subset h}\sqrt{N_j} \approx 355.96$. Pour cela, il évalue les L (ici 7) valeurs $\sum_{j\subset h}\sqrt{N_j}$ pour chacune des 2^L possibilités de regroupements de classes tels que le cumul des racines carrées des fréquences associées aux classes avoisine 355,96.

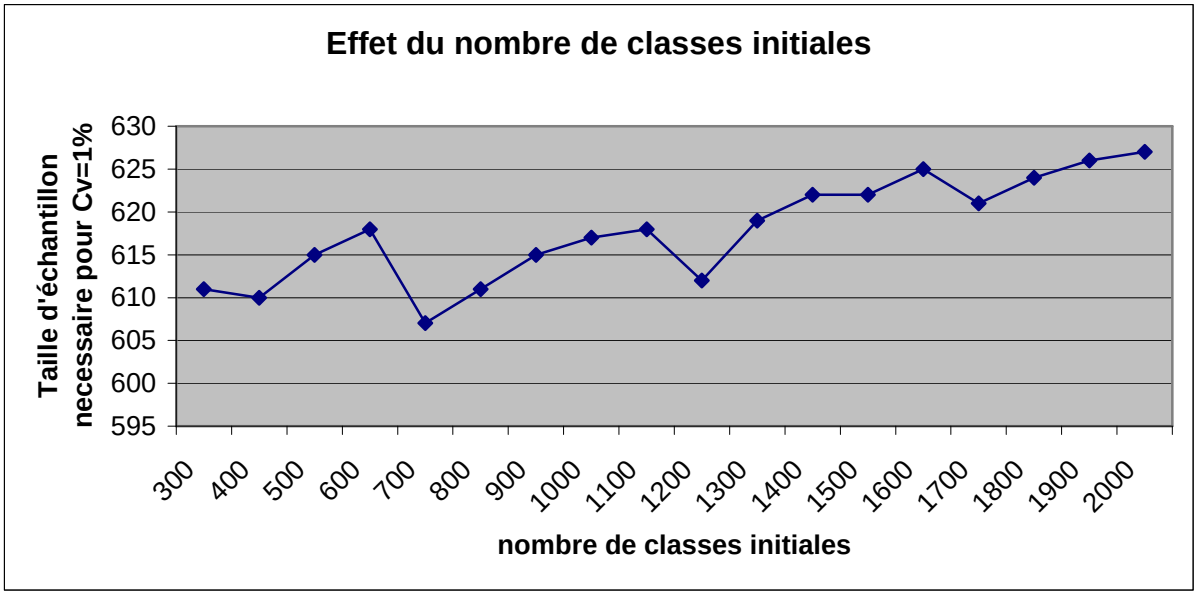
Par exemple, ici, pour les premières strates on a les possibilités suivantes :

Strate 1	**Strate 2**	**Strate 3**
classe 1 (331)	classes 2 et 3 (184+135=319)	…
	classes 2, 3 et 4 (184+135+111=430)	
classes 1 et 2 (331+184=515)	classes 3,4 et 5 (135+111+83=329)	
	classes 3,4,5 et 6 (135+111+83+66=395)	

Le programme retient ensuite le regroupement qui correspond au cumul de racines carrées des fréquences le plus proche de $\sum_{j\subset h}\sqrt{N_j} \approx 355.96$, c'est à dire celui qui minimise $\sum_{h=1}^{L}(\sum_{j\subset h}\sqrt{N_j} - 355.96)^2$.

Comme expliqué plus haut, le choix des classes initiales, donc, dans le cadre de ce package, du nombre de classes initiales, a un impact sur les résultats. Pour illustrer ceci, on peut tracer ce graphique qui correspond à la taille d'échantillon nécessaire à l'obtention d'un Cv de 1% avec 200, 300, 400,…,2000 classes initiales. On constate que, contrairement à ce que l'on pourrait penser,

l'augmentation du nombre de classes initiales n'améliore pas mécaniquement les précisions associées aux strates.



Il est donc conseillé, lors de l'utilisation de la méthode de Dalenius, de tester[19] plusieurs nombres de classes initiales.

1.4. La méthode géométrique

Gunning et Horgan ont proposé cette méthode dans l'article *Un nouvel algorithme pour la construction de bornes de stratification dans les populations asymétriques* paru en décembre 2004 dans la revue canadienne *Techniques d'enquête*. Cette méthode est peu[20] utilisée à l'Insee, pourtant l'algorithme est beaucoup plus simple que le précédent.

1.4.1. Conditions devant être vérifiées par les limites optimales des strates

D'après l'article de Gunning et Horgan, la méthode est fondée sur une observation de Cochran, selon laquelle, dans le cas de bornes quasi-optimales, les coefficients de variation empiriques sont souvent approximativement les mêmes pour toutes les strates. La méthode cherche donc à trouver les bornes $b_1,...,b_{L-1}$ telles que :

$$\frac{S_{yh}}{\mu_{yh}} = \frac{S_{yh+1}}{\mu_{yh+1}}$$

Calculer les écarts-types des strates et égaliser les coefficients de variation est un exercice trop compliqué pour être réalisable dans le cas général. Par contre, sous l'hypothèse supplémentaire que les y_k sont repartis selon une loi uniforme dans chaque strate h, on a :

$$\mu_{yh} \approx \frac{b_h + b_{h-1}}{2} \text{ et } S_{yh} \approx \frac{1}{\sqrt{12}}(b_h - b_{h-1})$$

[19] La mesure de l'effet du nombre de classes a été réalisée en fin de travail ce qui explique que nous n'ayons pas utilisé 700 classes initiales lors de l'illustration.
[20] Aucune utilisation à notre connaissance

Il s'ensuit :

$$\frac{S_{yh}}{\mu_{yh}} = \frac{S_{yh+1}}{\mu_{yh+1}} \Rightarrow b_h^2 = b_{h+1} b_{h-1}$$

D'où

$b_h = b_0 (\frac{b_L}{b_0})^{\frac{h}{L}}$ (b_0 doit être > 0) (2)

1.4.2. L'algorithme en pratique

L'algorithme est très simple à mettre en pratique. On connait l'ensemble des valeurs possibles y_k. On trouve la valeur maximale b_L et la valeur minimale b_0 des y_k et on construit nos bornes selon la formule (2).

1.4.3. Exemple d'application

En appliquant la méthode sur notre population d'entreprises dont les effectifs, hors partie exhaustive, varient entre 10 et 4 993, on obtient $b_1 = 10(\frac{4993}{10})^{\frac{1}{7}} \approx 24{,}29$, $b_2 = 10(\frac{4993}{10})^{\frac{2}{7}} \approx 59{,}01$, ...$b_7$=4 993.

Les strates correspondantes sont les suivantes :

Tranche d'effectif	Nombre d'unités	Somme des effectifs	Dispersion des effectifs
10 à 24	129 793	1 903 393	4,1
25 à 59	51 049	1 906 988	9,4
60 à 143	17 060	1 522 806	23,2
144 à 348	7 082	1 529 134	55,3
349 à 846	2 643	1 366 666	134,2
847 à 2 055	984	1 253 063	327,9
2 056 à 4 999	349	1 043 591	795,3
5 000 et plus	129	1 864 320	20 433,1
Total	**209 089**	**12 389 961**	**641,5**

Ces strates conduisent aux précisions suivantes :

Cv	Taille d'échantillon nécessaire			
	Sondage aléatoire simple	Sondage stratifié classique	Sondage stratifié méthode Dalenius	Sondage stratifié méthode géométrique
1%	57 922	666	615	611
2%	18 359	272	257	255
3%	8 644	195	188	187
4%	5 005	168	164	163
5%	3 276	156	152	151
6%	2 325	149	146	145
7%	1 747	144	144	143
8%	1 370	140	141	140
9%	1 111	140	137	138
10%	925	138	137	136

Les résultats obtenus avec la méthode géométrique sont très proches de ceux obtenus avec la méthode Dalenius. Pourtant, les strates correspondantes sont très différentes. Ainsi, on peut supposer qu'il existe d'autres bornes produisant de meilleurs résultats. La recherche de ces bornes optimales est l'objet de la méthode suivante.

1.5. La méthode LH (Lavallée et Hidiroglou)

La méthode LH correspond à une adaptation des premières méthodes itératives proposées par Lavallée et Hidiroglou en 1988. Kozak a proposé cette méthode itérative dans l'article[21] *Optimal stratification using random search method in agricultural surveys* paru en avril 2004 dans la revue *Statistics in transition*.

1.5.1. Conditions devant être vérifiées par les limites optimales des strates

Contrairement aux méthodes précédentes, qui cherchent à approcher l'équation (1) établie par Dalenius, cette méthode est basée sur des itérations. Elle compare un grand nombre de limites de strates et retient celles qui affichent la plus petite variance pour l'estimation du total de la variable de stratification.

On cherche donc à trouver les bornes $b_1,...,b_{L-1}$ qui minimisent la variance de l'estimateur du total de la variable y sans autres hypothèses.

Deux algorithmes implémentés dans le package *stratification* peuvent être utilisés pour remplir notre objectif. Nous étudierons ici uniquement la variante utilisant l'algorithme de recherche aléatoire de Kozak (*Kozak, 2004*), qui apparait plus souple, produit des résultats semblables (*Kozak et Verma, 2006*) et souffre moins de problèmes numériques (*Baillargeon, Rivest, Ferland 2007*) que celui de Sethi qui était mis en œuvre par Lavallée et Hidiroglou lorsqu'ils ont initié ce type de méthodes de découpage.

Le principe de l'algorithme de Kozak est simple : à chaque étape, une limite de strate est sélectionnée aléatoirement et modifiée aléatoirement. Si le nouvel ensemble de limites de strates est meilleur que le précédent, il remplace ce dernier. On recommence jusqu'à ne plus obtenir, après un certain nombre d'essais, de meilleures limites de strates.

[21] Voir aussi l'article de Kozak et Verma *Approche de la stratification par une méthode géométrique et par optimisation : Une comparaison de l'efficacité* paru en décembre 2006 dans la revue canadienne *Techniques d'enquête*.

1.5.2. L'algorithme en pratique

L'algorithme se déroule selon les étapes suivantes :

On trie la population en fonction des valeurs de la variable de stratification.

On calcule la valeur de la fonction d'optimisation (une précision pour une taille d'échantillon donnée ou une taille d'échantillon permettant d'atteindre une précision donnée) avec le vecteur de bornes initiales de stratification **b** (paramétrable) et on vérifie les contraintes (N_h>1, bon nombre de bornes,…). Si elles ne sont pas satisfaites, les points initiaux doivent être modifiés.

Pour *r* = 0,1,…, *R*[22] *;* on répète l'étape suivante :

On génère le point **b'** en tirant aléatoirement une limite de strate b_i puis en la modifiant comme il suit

$$b_i^{'} = b_i + j$$

où *j* est un nombre entier aléatoire[23] contenu dans l'intervalle [-*p*,*p*], *p* étant un nombre donné choisi d'après le nombre de valeurs distinctes dans la population (la valeur de *p* est d'autant plus élevée que le nombre de valeurs distinctes dans la population est grand).

Les autres bornes ne sont pas modifiées (pour $k \neq i$ on a $b_k^{'} = b_k$).

On vérifie les contraintes (N_h>1, $b_{i-1} < b_i^{'} < b_{i+1}$,...).

Si les contraintes sont satisfaites, on calcule la valeur de la fonction d'optimisation.

Si la valeur de la fonction d'optimisation sous le vecteur **b**' est plus petite que celle obtenue sous le vecteur **b**, on accepte le nouveau vecteur.

L'algorithme se termine lorsque la règle d'arrêt est satisfaite, c'est à dire si *r*=*R* ou que, lors des *m* (par exemple 50) dernières étapes, la valeur de la fonction d'optimisation ne s'est pas améliorée.

Contrairement aux méthodes précédentes, lancer plusieurs fois l'algorithme ne mènera probablement pas aux mêmes solutions. En pratique, on lance donc l'algorithme plusieurs fois, éventuellement en faisant varier des paramètres, et on retient le meilleur résultat parmi les répétitions de l'algorithme.

Enfin, on notera la présence de l'option takeall=1 permettant de demander la présence d'une strate exhaustive a posteriori, c'est à dire que la dernière strate (numéro L) sera exhaustive. Cette strate est différente de la strate exhaustive que l'on impose a priori (option certain=).

1.5.3. Exemple d'application

En appliquant la méthode, avec les paramètres par défaut du programme, sur notre population d'entreprises, on obtient les résultats suivants :

[22] R correspond au nombre maximal de répétitions de l'algorithme. Il vaut 10 000 par défaut dans le package et peut être modifié par l'utilisateur (paramètre maxiter).
[23] i et j sont recalculés à chaque itération.

Cv	Taille d'échantillon nécessaire				
	Sondage aléatoire simple	Sondage stratifié classique	Sondage stratifié méthode Dalenius	Sondage stratifié méthode géométrique	Sondage stratifié méthode LH
1%	57 922	666	615	611	602
2%	18 359	272	257	255	251
3%	8 644	195	188	187	184
4%	5 005	168	164	163	160
5%	3 276	156	152	151	150
6%	2 325	149	146	145	144
7%	1 747	144	144	143	140
8%	1 370	140	141	140	137
9%	1 111	140	137	138	137
10%	925	138	137	136	136

On réduit encore un peu nos tailles d'échantillon nécessaires pour atteindre les différents Cv.

Les bornes de stratification obtenues avec cette méthode dépendent du Cv visé, comme on peut le voir sur le tableau ci-dessous qui représente les 6 bornes obtenues pour nos 10 Cv visés :

Cv	b1	b2	b3	b4	b5	b6
1%	23,5	55,5	127,5	294,5	711,5	1758,5
2%	23,5	56,5	130,5	303,5	731,5	1811,5
3%	23,5	57,5	132,5	307,5	736,5	1831,5
4%	24,5	56,5	130,5	303,5	730,5	1819
5%	24,5	55,5	132,5	317,5	807,5	1840,5
6%	24,5	54,5	126,5	294,5	712,5	1782
7%	24,5	62,5	160,5	424,5	936	2080,5
8%	24,5	59,5	136,5	311,5	741	1822,5
9%	25,5	58,5	133,5	306,5	735,5	1822,5
10%	23,5	55,5	126,5	294,5	712,5	1782

Pour Cv=1%, l'algorithme se déroule de la façon suivante (seules les itérations où le critère d'optimisation[24] s'améliore apparaissent) :

[24] Ici la taille d'échantillon arrondie nécessaire pour obtenir un Cv de 1%, et à taille d'échantillon arrondie nécessaire égale, la taille d'échantillon non arrondie nécessaire.

b1	b2	b3	b4	b5	b6	Taille d'échantillon (arrondie) nécessaire pour Cv=1%	Taille d'échantillon (non arrondie) nécessaire pour Cv=1%	step	Numéro de l'itération
24,5	59,5	143,5	348,5	846,5	2048,5	611	607,4	0	0
24,5	59,5	143,5	326,5	846,5	2048,5	611	607,3	-22	10
24,5	59,5	143,5	326,5	846,5	1819	611	606,8	-69	17
24,5	59,5	143,5	337,5	846,5	1819	611	606,5	11	18
24,5	59,5	143,5	337,5	841,5	1819	610	606,0	-5	36
24,5	59,5	143,5	337,5	764,5	1819	607	603,9	-66	46
24,5	59,5	143,5	327,5	764,5	1819	607	603,2	-10	155
24,5	59,5	143,5	327,5	753,5	1819	607	603,1	-11	207
24,5	59,5	143,5	327,5	753,5	1796,5	606	603,1	-10	225
24,5	59,5	143,5	316,5	753,5	1796,5	606	603,0	-11	231
24,5	59,5	143,5	316,5	745	1796,5	606	602,9	-8	246
24,5	59,5	143,5	316,5	734,5	1796,5	606	602,8	-9	275
24,5	59,5	143,5	316,5	734,5	1774,5	606	602,8	-6	287
24,5	59,5	143,5	316,5	734,5	1782	606	602,8	2	294
24,5	59,5	142,5	316,5	734,5	1782	606	602,5	-1	310
24,5	59,5	142,5	316,5	734,5	1808,5	605	602,6	9	311
24,5	59,5	140,5	316,5	734,5	1808,5	603	602,3	-2	613
24,5	59,5	138,5	316,5	734,5	1808,5	603	602,1	-2	614
24,5	59,5	138,5	316,5	735,5	1808,5	603	602,1	1	628
24,5	59,5	137,5	316,5	735,5	1808,5	603	602,0	-1	633
24,5	57,5	137,5	316,5	735,5	1808,5	603	601,9	-2	671
24,5	57,5	135,5	316,5	735,5	1808,5	603	601,8	-2	719

L'algorithme démarre avec les bornes optimales obtenues avec la méthode géométrique[25]. Ensuite, il sélectionne au hasard un nombre (step) entre -100 et 100, ainsi qu'une des six bornes, modifie la borne sélectionnée en lui ajoutant step, et calcule la taille d'échantillon (arrondie et non arrondie) nécessaire pour obtenir un Cv de 1%.
L'algorithme répète cette opération neuf fois avant de trouver, à la dixième itération (deuxième ligne), en retranchant 22 à la borne b_4, une taille d'échantillon arrondie nécessaire pour obtenir un Cv de 1% égale (611) à celle obtenue avec les bornes initiales et une taille d'échantillon non arrondie nécessaire pour obtenir un Cv de 1% inférieure (607,3) à celle obtenue (607,4) avec les bornes initiales. L'algorithme conserve donc la nouvelle borne b_4 et recommence à chercher de meilleures bornes. Il modifie ainsi 21 fois les bornes en 719 itérations, puis il cherche 500 fois de meilleures bornes sans en obtenir, donc il s'arrête.

Afin de rendre plus robuste ce résultat, l'algorithme en entier est répété quinze fois avec des bornes initiales différentes[26] : les cinq premières répétitions sont réalisées avec comme bornes initiales celles obtenues par la méthode Dalenius, les cinq suivantes avec les bornes obtenues par la méthode géométrique et les cinq dernières avec des bornes dites robustes[27]). Ici, comme la méthode Dalenius ne fonctionne pas avec le nombre de classes initiales utilisé par défaut dans le package, les cinq répétitions correspondantes n'ont tout simplement pas lieu et l'on obtient les résultats finaux suivants :

[25] Pour fixer une borne, l'algorithme calcule une borne théorique puis utilise la valeur moyenne entre les deux valeurs les plus proches existant dans les données. La valeur 2055,3 issue de la méthode géométrique et associée à la borne 6 est ainsi transformée dès le début de l'algorithme en 2048,5 car les deux valeurs présentes dans les données l'avoisinant le plus sont 2 040 et 2 057.
[26] L'utilisateur peut aussi imposer des strates initiales (paramètre initbh).
[27] Les strates robustes sont constituées de façon à respecter les contraintes aussi souvent que possible. Elles contiennent approximativement le même nombre de valeurs uniques de y.

b_1	b_2	b_3	b_4	b_5	b_6	Taille d'échantillon (arrondie) nécessaire pour Cv=1%	Taille d'échantillon (non arrondie) nécessaire pour Cv=1%	Nombre d'itérations avant d'obtenir les bornes finales
24,5	57,5	135,5	316,5	735,5	1808,5	603	601,8	769
24,5	59,5	138,5	316,5	735,5	1763,5	603	602,0	815
23,5	56,5	129,5	296,5	711,5	1774,5	602	600,3	771
23,5	**55,5**	**127,5**	**294,5**	**711,5**	**1774,5**	**602**	**600,2**	**701**
24,5	59,5	138,5	324,5	785,5	1840,5	605	602,9	446
24,5	57,5	135,5	311,5	734,5	1796,5	603	601,6	1354
23,5	56,5	129,5	301,5	729,5	1782	603	600,5	1289
24,5	57,5	133,5	313,5	741	1808,5	603	601,6	1086
23,5	56,5	129,5	300,5	712,5	1763,5	603	600,4	963
24,5	58,5	140,5	338,5	801	1891	606	603,9	908

Enfin, le programme retient la répétition d'algorithme qui minimise la taille d'échantillon arrondie nécessaire pour obtenir un Cv de 1%, et, à taille d'échantillon arrondie nécessaire égale, la taille d'échantillon non arrondie nécessaire pour obtenir un Cv de 1%. Il s'agit ici de la quatrième répétition, surlignée en gras, qui mène à un échantillon de 602 unités.

L'échantillon est alors réparti selon une allocation de Neyman. Si l'on vise un Cv de 1%, on obtient :

Tranche d'effectif	Nombre d'unités	Somme des effectifs	Dispersion des effectifs	Allocation
10 à 23	129 793	1 903 393	4,1	83
24 à 55	48 821	1 779 032	8,5	80
56 à 127	17 749	1 442 704	19,9	61
128 à 294	7 746	1 457 125	45,5	61
295 à 711	3 206	1 405 453	113,0	63
712 à 1774	1193	1 297 890	291,3	60
1775 à 4 999	452	1 240 044	834,3	65
5 000 et plus	129	1 864 320	20 433,1	129
Total	**209 089**	**12 389 961**	**641,5**	**602**

1.5.4. Le paramètre de modification maximale d'une borne et le nombre d'itérations avant de considérer qu'il y a convergence

Étant donné la structure de l'algorithme, on voit que l'étendue des modifications de bornes aura un impact sur les résultats. Si elle n'est pas assez grande, on risque de ne pas atteindre les limites optimales mais un minimum local. Si elle est trop grande, on risque de passer à coté des limites optimales, ou alors il sera nécessaire d'augmenter le nombre d'itérations pour espérer converger. Pour éviter ces difficultés, le programme commence par tester des modifications importantes (correspondant à des valeurs j proches de p ou -p dans la description de l'algorithme) et, lorsqu'il n'a pas trouvé de meilleure solution au bout d'un certain nombre d'itérations, ce qui suppose que l'on doit être assez proche de la solution optimale, l'algorithme se limite au test de petites modifications (correspondant à des valeurs j proches de 0 dans la description de l'algorithme).
Par défaut, le programme prend comme modification maximale d'une borne (paramètre maxstep), la plus petite valeur entre 100 et le nombre de valeurs différentes divisé par 10. Puis, il prend

comme nombre d'itérations (paramètre maxstill) avant de considérer qu'il y a convergence, dix fois maxstep relevé à 50 ou diminué à 500 si nécessaire.

1.5.5. Le nombre de répétitions de l'algorithme

Étant donné que les recherches de bornes sont aléatoires, lancer deux fois l'algorithme peut mener à deux résultats différents. Le programme contient un paramètre (rep) permettant d'indiquer le nombre de fois que l'on souhaite lancer l'algorithme pour retenir la meilleure[28] solution. Par défaut, l'algorithme est lancé quinze fois : cinq fois avec les bornes obtenues avec la méthode de Dalenius comme bornes initiales, cinq fois avec les bornes obtenues avec la méthode géométrique comme bornes initiales, et cinq fois avec les bornes robustes comme bornes initiales.

Un moyen simple de vérifier que le paramètre rep est bien réglé consiste à lancer plusieurs fois l'instruction, ce qui revient à répéter les quinze répétitions de l'algorithme, et vérifier qu'on obtient bien à chaque fois le même résultat.

1.5.6. Imposer que la $L^{ème}$ strate soit exhaustive

Il est fréquent d'imposer une strate exhaustive dans les plans de sondage d'enquêtes auprès d'entreprises à partir d'un seuil d'exhaustivité sur l'effectif. Pour établir les résultats précédents, nous avons imposé que les entreprises de 5 000 salariés ou plus soient tirées d'office dans l'échantillon. Si l'on souhaite intégrer une strate exhaustive sans a priori sur le seuil d'exhaustivité, on peut, pour la méthode LH, imposer que la $L^{ème}$ strate soit tirée exhaustivement. L'algorithme fonctionne comme vu précédemment, mais au moment du calcul de l'allocation, il impose $n_L=N_L$.

Pour tester cette possibilité sur nos données, nous allons supprimer[29] l'exhaustivité des unités de 5 000 salariés et plus, et lancer l'algorithme avec L=8 en demandant l'exhaustivité pour la $L^{ème}$ strate. On obtient alors les résultats suivants :

Cv	Taille d'échantillon nécessaire					
	Sondage aléatoire simple	**Sondage stratifié classique**	**Sondage stratifié méthode Dalenius**	**Sondage stratifié méthode géométrique**	**Sondage stratifié méthode LH sans optimisation strate exhaustive**	**Sondage stratifié méthode LH avec optimisation strate exhaustive**
1%	57 922	666	615	611	602	601
2%	18 359	272	257	255	251	201
3%	8 644	195	188	187	184	105
4%	5 005	168	164	163	160	70
5%	3 276	156	152	151	150	51
6%	2 325	149	146	145	144	36
7%	1 747	144	144	143	140	28
8%	1 370	140	141	140	137	22
9%	1 111	140	137	138	137	18
10%	925	138	137	136	136	15

Le gain de taille d'échantillon devient très important notamment pour les grands Cv pour lesquels le seuil de 5 000 salariés était trop bas, comme on peut le voir dans le tableau ci-dessous qui présente les bornes optimales obtenues, la borne b_7 représentant le seuil d'exhaustivité optimal. On voit aussi

[28] Au sens de la fonction d'optimisation

[29] En pratique, il est possible de conserver une strate exhaustive a priori, mais ce n'est pas intéressant dans notre cas.

que le seuil d'exhaustivité optimal pour un Cv de 1% est de 4 713,5 salariés, ce qui est proche du seuil de 5 000 salariés que nous avions imposé précédemment et explique que les tailles d'échantillon nécessaires pour atteindre un Cv de 1% avec et sans optimisation du seuil d'exhaustivité soient si proches.

Cv	b_1	b_2	b_3	b_4	b_5	b_6	b_7
1%	23,5	56,5	127,5	293,5	704,0	1 710,5	4 760,0
2%	23,5	61,5	156,5	403,5	1 115,5	3 399,0	11 773,5
3%	24,5	63,5	164,5	423,5	1 186,0	3 633,0	16 055,0
4%	26,5	69,5	195,5	529,5	1 397,5	4 130,5	22 867,5
5%	25,5	66,5	178,5	497,5	1 480,5	4 889,5	32 711,5
6%	26,5	83,5	282,5	930,5	3 297,0	16 055,0	87 293,5
7%	27,5	82,5	266,5	913,5	3 312,0	16 055,0	87 293,5
8%	26,5	78,5	249,5	833,5	3 020,0	15 036,0	87 293,5
9%	25,5	86,5	312,5	965,5	3 166,5	16 055,0	87 293,5
10%	27,5	83,5	271,5	921,0	3 333,0	16 055,0	87 293,5

Pour information, Ces limites de strates conduisent aux populations suivantes par strate :

Cv	N_1	N_2	N_3	N_4	N_5	N_6	N_7	N_{exh}
1%	126 417	52 772	17 174	7 727	3 214	1 172	471	142
2%	126 417	55 433	17 016	6 798	2 354	852	187	32
3%	129 793	52 977	16 583	6 503	2 236	794	180	23
4%	135 754	49 384	15 995	5 433	1 728	619	160	16
5%	132 826	51 206	16 182	6 168	1 968	607	119	13
6%	135 754	53 564	14 560	3 909	1 076	203	21	2
7%	138 378	50 690	14 453	4 245	1 098	202	21	2
8%	135 754	52 237	15 086	4 523	1 228	237	22	2
9%	132 826	57 234	14 386	3 396	1 006	218	21	2
10%	138 378	50 940	14 324	4 135	1 088	201	21	2

1.6. Le nombre de strates : une marge de gain importante

Dans ce qui précède, nous nous sommes limités à trouver les sept strates permettant d'atteindre, avec la plus petite taille d'échantillon possible, un certain Cv pour l'estimation de l'effectif total. Dans cette partie on va chercher à jouer sur le nombre de strates et à évaluer les gains potentiels d'une hausse du nombre de strates.

La théorie des sondages nous enseigne que le nombre de strates doit théoriquement être le plus grand possible mais doit garantir, afin de préserver le caractère non biaisé de l'estimateur d'Horvitz Thompson, qu'au moins une[30] unité soit interrogée dans chaque strate.

Nous proposons donc la démarche suivante, pour une méthode donnée :

- Nous calculons tout d'abord les limites optimales de strates pour L (nombre de strates) variant de 2 à 50 ;

[30] Une taille d'échantillon seulement égale à 1 dans chaque (ou certaines) strate(s) est acceptable mais interdit les estimations sans biais de variance, sauf à faire appel à une technique de « collapse »… De plus, le phénomène de non-réponse accroit le risque de n'obtenir aucune information dans certaines strates où l'échantillon tiré à l'origine est trop petit (*Ardilly, 2006*).

- Nous calculons ensuite, pour chaque L, la taille d'échantillon minimale permettant d'atteindre un Cv de x% (x variera de 1% à 10%) en imposant[31] n_h>=1 pour tout h.
- Nous retenons au final le L_{opt} qui permet d'assurer un Cv de x% avec la plus petite taille d'échantillon.

On verra que L_{opt} dépend du niveau de précision à atteindre et n'est pas nécessairement très grand. En effet, comme on impose qu'au moins une unité soit tirée par strate, il arrive un moment où ajouter une strate supplémentaire implique d'augmenter la taille de l'échantillon.

1.6.1. Exemple avec Cv de 5% et la méthode géométrique

Le tableau qui suit recense les tailles d'échantillon nécessaires pour atteindre un Cv de 5% pour L variant de 2 à 50 en définissant les strates avec la méthode géométrique :

L	Taille d'échantillon arrondie nécessaire pour atteindre un Cv de 5%	Taille d'échantillon non arrondie nécessaire pour atteindre un Cv de 5%
2	444	443,3
3	247	245,7
4	196	194,3
5	171	168,9
6	159	156,0
7	151	148,6
8	148	144,3
9	146	141,1
10	142	138,6
11	**140**	**136,9**
12	141	135,7
13	142	134,6
14	143	133,9
15	144	133,3
...		
45	174	129,4

Pour atteindre un Cv de 5% et en réalisant une stratification avec la méthode géométrique, le nombre optimal de strates est de 11. Il permet d'atteindre un Cv de 5% avec un échantillon de 140 unités.

1.6.2. Résultats pour des Cv allant de 1% à 10% avec la méthode géométrique

En répétant le procédé ci-dessus pour des Cv allant de 1% à 10%, on obtient les résultats suivants :

[31] En pratique, on cherche souvent à imposer n_h>=10, mais ceci n'est pas paramétrable dans le package.

Cv	Taille d'échantillon nécessaire			
	Sondage stratifié classique	Sondage stratifié méthode géométrique (L=7)	Sondage stratifié méthode géométrique avec optimisation du nombre de strates	
			L_{opt}	n
1%	666	611	33	163
2%	272	255	22	151
3%	195	187	16	145
4%	168	163	14	143
5%	156	151	11	140
6%	149	145	10	139
7%	144	143	9	138
8%	140	140	8	137
9%	140	138	8	137
10%	138	136	7	136

En optimisant le nombre de strates, il est possible de réduire de façon importante les tailles d'échantillon nécessaires afin d'atteindre un Cv de x%. Ces calculs sont réalisés en imposant le seuil d'exhaustivité de 5 000 salariés. Essayons avec la méthode LH qui permet d'optimiser le seuil d'exhaustivité.

1.6.3. Résultats pour des Cv allant de 1% à 10% avec la méthode LH

Avec la méthode LH et en optimisant le seuil d'exhaustivité (takeall=1), on obtient :

Cv	Taille d'échantillon nécessaire			
	Sondage stratifié classique[32]	Sondage stratifié méthode LH avec optimisation du seuil d'exhaustivité et sans optimisation du nombre de strates (L=7)	Sondage stratifié méthode LH avec optimisation du nombre de strates et du seuil d'exhaustivité	
			L_{opt}	n
1%	666	601	39	60
2%	272	201	29	37
3%	195	105	18	26
4%	168	70	18	21
5%	156	51	16	18
6%	149	36	12	16
7%	144	28	11	14
8%	140	22	11	13
9%	140	18	10	12
10%	138	15	9	11

[32] Avec la stratification «classique», le seuil d'exhaustivité est fixé à 5 000 salariés. La strate exhaustive concerne alors 129 entreprises.

Avec la méthode LH avec optimisation du nombre de strates et du seuil d'exhaustivité, il est possible de beaucoup réduire les tailles d'échantillon nécessaires pour atteindre les différents Cv en augmentant le nombre de strates. Ces gains sont notamment importants pour les « petits » Cv correspondant aux meilleures précisions.

1.7. Discussion des résultats obtenus

En optimisant le nombre de strates, il semble qu'on soit allé au bout de l'exercice théorique consistant à minimiser la taille d'échantillon nécessaire à l'obtention d'un Cv de x% pour l'estimation du total de l'effectif salarié en stratifiant notre population d'entreprises selon l'effectif salarié. Les très (trop ?) bons résultats que nous obtenons peuvent être critiqués sur plusieurs points, dont ceux qui suivent.

1.7.1. En pratique, la variable de stratification n'est que corrélée à la variable d'intérêt (et non égale)

Dans ce qui précède, nous avons fait l'hypothèse que la variable de stratification et la variable d'intérêt étaient confondues. Cette hypothèse, qui est souvent formulée lors de conceptions de plan de sondage, n'est jamais vérifiée en pratique. Il est en effet inutile de réaliser une enquête si l'on connaît déjà la variable d'intérêt sur l'ensemble de la population. On compte donc sur un lien suffisamment « fort » entre notre variable de stratification et notre variable d'intérêt pour que les résultats obtenus lors de la conception du plan de sondage restent à peu près vrais lors de l'enquête réelle. Les résultats que nous avons obtenus ne sont-ils pas trop spécifiques à nos données ? Pour illustrer ce propos, nous avons calculé, pour les plans de sondage permettant d'atteindre un Cv de 1% correspondant à la stratification classique, la stratification LH avec 7 strates sans optimisation du seuil d'exhaustivité et la stratification LH avec optimisation du seuil d'exhaustivité et du nombre de strates, la précision de deux paramètres issus de variables d'intérêt que l'on considère habituellement liées à l'effectif salarié d'une année N :

- l'effectif salarié de l'année N-1[33] ;
- le chiffre d'affaires.

On obtient les résultats suivants :

Méthode de stratification	**Taille d'échantillon**	**Cv pour l'estimation de l'effectif total**	**Cv pour l'estimation de l'effectif N-1 total**	**Cv pour l'estimation du Chiffre d'affaires total**
Stratification classique (L=7)	666	1,0%	3,1%	16,1%
Stratification LH sans optimisation du seuil d'exhaustivité et du nombre de strates (L=7)	602	1,0%	2,8%	16,0%
Stratification LH avec optimisation du seuil d'exhaustivité et du nombre de strates (L=39)	60	1,0%	8,9%	61,8%

On constate que la première optimisation (stratification LH sans optimisation du seuil d'exhaustivité (L=7)) conduit à des résultats équivalents à ceux obtenus avec la stratification classique, pour l'estimation de l'effectif N-1 total et pour l'estimation du chiffre d'affaires total, tout en permettant une réduction de 10% de la taille d'échantillon. Cette optimisation semble donc robuste aux variables quantitatives liées à l'effectif et on devrait courir peu[34] de risques à l'adopter.

[33] En pratique, on serait plutôt intéressé par le lien avec l'effectif N+1, voire N+2, mais ces données ne sont pas disponibles et les liens devraient être comparables

[34] Il y a probablement des variables ou des domaines pour lesquels on va perdre un peu en précision par rapport à l'échantillon de 666 unités.

On constate aussi que la seconde optimisation (stratification LH avec optimisation du seuil d'exhaustivité et du nombre de strates) ne conserve pas de bonnes propriétés sur les variables que l'on considère liées à l'effectif salarié. Ainsi, cette optimisation est probablement trop liée aux données utilisées, et elle n'est optimale que pour l'estimation de l'effectif salarié total. Avant d'adopter cette optimisation, qui permet de diviser par 10 la taille d'échantillon par rapport à la stratification classique, il faut être conscient de la précision que l'on perd sur d'autres variables, même corrélées à l'effectif salarié, et être sûr que le lien entre les variables d'intérêt et la variable de stratification est suffisamment fort pour que les résultats soient exploitables.

1.7.2. En pratique, on souhaite souvent garantir qu'au moins 10 unités soient tirées dans chaque strate

Dans les parties précédentes, nous avons imposé qu'au moins une unité soit tirée dans chaque strate afin de garantir le caractère non biaisé des estimations. En pratique, on souhaite généralement imposer qu'environ[35] dix unités soient tirées dans chaque strate, en particulier pour permettre[36] des estimations de précision. Nous n'avons pas intégré une telle contrainte dans notre étude d'une part parce que le package *stratification* ne le permet pas et d'autre part parce qu'il était intéressant de pousser l'exercice théorique à son maximum. On voit bien ici que les solutions optimales trouvées avec la stratification LH en optimisant le nombre de strates et en optimisant le seuil d'exhaustivité ne seront pas applicables[37] en pratique car elles conduisent à un tirage de une ou deux unités par strate en moyenne. Ainsi, dans les cas pratiques, le nombre de strates devrait plutôt devenir une donnée en entrée du programme et que l'on n'optimisera pas ou peu. De plus, les calculs de précisions réalisés à partir du package *stratification* devront probablement être refaits avec les contraintes de tailles minimales d'échantillon par strate en vigueur à l'Insee.

1.7.3. En pratique, les objectifs de précision sont souvent multiples

Dans cette étude, on s'est intéressé à un unique paramètre d'intérêt (l'effectif total de la population) et on a cherché à optimiser le plan de sondage de façon à être le plus précis possible lors de l'estimation de ce paramètre. En pratique, les enquêtes n'ont que très rarement un unique objectif : plusieurs variables sont relevées lors de l'enquête et pour une même variable on produit plusieurs estimations en fonction de nos domaines d'intérêt. Lors de la réduction d'une taille d'échantillon, on perd a priori de la qualité sur certaines variables (les moins liées à notre variable d'optimisation notamment) ou certains domaines d'intérêt (les plus différents du domaine d'optimisation notamment). Il est donc important d'évaluer correctement, lors de l'optimisation d'un plan de sondage, à la fois les gains réalisés mais aussi les pertes et de s'assurer que ces dernières ne sont pas problématiques pour les utilisations futures des résultats de l'enquête.

Par exemple, si l'on reprend le cas évoqué lors de la discussion de la corrélation entre la variable de stratification et la variable d'intérêt, on peut s'interroger sur l'évaluation de variables qualitatives avec nos différentes stratégies d'échantillonnages. Notamment, est-ce que la stratification LH sans optimisation du seuil d'exhaustivité et du nombre de strates, qui semblait meilleure que la stratification classique pour toutes les variables liées aux effectifs, est aussi meilleure pour des estimations de proportions (a priori peu liées aux effectifs) ? En supposant $p_h = \frac{1}{2}$[38], on peut majorer la demi-longueur de l'intervalle de confiance associé à l'estimation d'une proportion sur l'ensemble du champ :

[35] Il s'agit ici de la pratique habituelle, on pourra se référer, pour plus de détails, au rapport de stage (rapport de stage des attachés-stagiaires issus du concours externe année scolaire 2012-2013) de Laurent Costa sur le sujet.

[36] Lorsque le nombre d'unités par strate est faible, il existe des techniques dites de collapse (*Ardilly 2006*) permettant de calculer des précisions, mais les calculs sont alors complexifiés et il n'existe pas encore d'outils permettant de contourner facilement ces difficultés à l'Insee.

[37] Cependant, on remarquera qu'en tirant 10 unités dans les 39 strates optimales trouvées avec la stratification LH en optimisant le nombre de strates et en optimisant le seuil d'exhaustivité, la taille d'échantillon (390) reste assez largement inférieure à 600.

[38] Cette hypothèse est souvent formulée dans les optimisations de plans de sondage. La dispersion (proportionnelle à $p_h(1-p_h)$) de la variable indicatrice associée atteint alors son maximum. Si la proportion réelle n'est pas de 50%, la précision associée à l'estimation sera meilleure.

Méthode de stratification	Taille d'échantillon	Cv pour l'estimation de l'effectif total	Cv pour l'estimation de l'effectif N-1 total	Cv pour l'estimation du Chiffre d'affaires total	Majoration de la demi-longueur de l'intervalle de confiance pour une proportion
Stratification classique (L=7)	666	1,0%	3,1%	16,1%	7,6%
Stratification LH sans optimisation du seuil d'exhaustivité et du nombre de strates (L=7)	602	1,0%	2,8%	16,0%	7,1%
Stratification LH avec optimisation du seuil d'exhaustivité et du nombre de strates (L=39)	60	1,0%	8,9%	61,8%	29,0%

Ici, la stratification LH sans optimisation du seuil d'exhaustivité et du nombre de strates reste équivalente, et même un peu meilleure, que la stratification classique pour l'estimation de proportions. Par contre, la stratification LH avec optimisation du seuil d'exhaustivité et du nombre de strates n'est pas acceptable si l'on souhaite évaluer des proportions proches de 50%.

1.7.4. En pratique, la présence de non-réponse rendra moins précises les estimations

Dans ce qui précède, nous avons supposé des taux de réponse de 100% dans nos calculs de précision. En pratique, la présence de non-réponse fera que les précisions obtenues seront moins bonnes.

La prise en compte de taux de réponse anticipés est possible dans le package (option rh=).

1.8. Conclusion sur cette première expérience d'utilisation du package

Les trois méthodes de découpage optimal de strates implémentées dans le package *stratification* sont assez simples à utiliser. Avec nos données, il est tout de même nécessaire de modifier les paramètres initiaux pour utiliser la méthode de Dalenius, et il est intéressant de tester la modification de plusieurs paramètres pour la méthode LH.

Selon nos calculs, en imposant un seuil d'exhaustivité à 5 000 salariés et en découpant en 7 strates le reste de notre population d'entreprises, il serait possible de réduire d'environ 10% la taille d'échantillon nécessaire pour atteindre un Cv de 1% pour l'estimation de l'effectif total de la population par rapport à la stratification classique. Le gain se réduit pour des objectifs de précision moins ambitieux et devient négligeable pour les Cv supérieurs à 3%. Les trois méthodes conduisent à des résultats assez proches tant que le seuil d'exhaustivité est fixé à 5 000 salariés.

Sans imposer une strate exhaustive à 5 000 salariés, il est possible de réaliser des gains supplémentaires de taille d'échantillon en utilisant la méthode LH avec recherche de seuil d'exhaustivité optimal.

En jouant sur le nombre de strates, on peut pousser l'exercice théorique de minimisation de taille d'échantillon pour atteindre un objectif de précision donné jusqu'au bout. On atteint alors les objectifs de précision avec des tailles d'échantillon beaucoup plus petites qu'avec la stratification classique. Cependant, lorsque l'on teste la robustesse du plan de sondage prometteur pour le Cv de 1% pour

l'estimation de l'effectif total de la population avec des variables censées être liées à la variable de stratification, on constate une nette dégradation de nos précisions par rapport à la stratification classique[39].

Il sera donc important dans les futures utilisation du package de contrôler l'optimisation des limites de strates par des calculs de précision sur d'autres variables que la variable de stratification afin de ne pas rendre trop spécifique le plan de sondage et de maitriser à la fois les gains et les pertes réalisés.

[39] De plus, la stratification des enquêtes auprès des entreprises repose habituellement sur plusieurs variables (activité principale exercée, éventuellement implantation géographique…). De ce fait, le découpage en tranches de la taille des entreprises ne devra pas comporter trop de tranches sous peine d'obtenir des strates comportant trop peu d'unités.

2. Application aux Enquêtes Mensuelles de Branches

Le but des enquêtes mensuelles de branche est de mesurer l'évolution d'un mois sur l'autre de la production industrielle française par produit. Pour ce faire, un échantillon d'entreprises[40] industrielles est sélectionné chaque année au mois de novembre de l'année *N-1*. Ces entreprises seront interrogées chaque mois de l'année *N* sur les évolutions de leur production pour certains des produits qu'elles fabriquent. Leurs réponses permettent par agrégation de constituer les séries témoin de l'Indice de la Production Industrielle (IPI).

Le plan de sondage des EMB doit ainsi permettre de sélectionner, pour chacun des produits intervenant dans les séries témoin de l'IPI, un échantillon d'entreprises fabriquant ce produit dont l'évolution de la production soit représentative des variations effectives dans l'économie française.

La base de sondage des EMB doit donc contenir, pour chaque produit, la liste des entreprises qui le fabriquent. Seule l'Enquête Annuelle de Production (EAP) et son extension l'EAbis apportent cette information. L'EAP permet en effet de recueillir chaque année, pour un échantillon d'entreprises métropolitaines des secteurs industriels hors agroalimentaire, la ventilation de leur production suivant leurs différentes activités et leurs différents produits. L'EAbis complète l'EAP en recueillant la même information pour les entreprises industrielles des DOM ou pour les entreprises non industrielles ayant des activités industrielles secondaires significatives.

La base de sondage actuelle est plus précisément formée par les strates exhaustives de l'EAP et l'EAbis. À la date de tirage des EMB pour une année *N* l'EAP *N-2* est disponible. Les informations de la base de sondage sont donc datées de deux ans.

L'unité d'échantillonnage est un produit d'une entreprise. Une entreprise peut donc être sélectionnée pour plusieurs des biens qu'elle produit.
La procédure d'échantillonnage actuelle dépend du nombre d'entreprises fabriquant un produit :

- pour les produits fabriqués par moins de 10 entreprises, toutes les entreprises sont interrogées ;
- pour les produits fabriqués par plus de 10 mais moins de 200 entreprises, l'échantillon est sélectionné par *cut-off* : pour chaque produit, les entreprises qui le fabriquent sont classées par ordre décroissant de chiffre d'affaires et sélectionnées dans l'échantillon jusqu'à ce que la somme de leur chiffre d'affaires pour le produit considéré dépasse 75 % de la production totale des entreprises de la base de sondage dans ce produit ;
- pour les 14 produits fabriqués par plus de 200 entreprises, l'échantillon est sélectionné suivant un échantillonnage aléatoire. Pour la mécanique industrielle, une strate exhaustive regroupe les 19 plus gros fabricants, le reste de l'échantillon étant constitué par un sondage stratifié en quatre strates, avec sondage proportionnel à la facturation dans chaque strate. Les bornes des strates ont été de plus calculées à partir de la facturation des entreprises, par application de la méthode des racines de fréquence cumulées. Pour les autres produits, les entreprises dont les productions dans le produit sont les plus importantes sont interrogées exhaustivement, jusqu'à ce que leurs productions cumulées représentent 50 % de la production du produit ; un échantillon aléatoire est sélectionné sur le reste de la base de sondage par un sondage aléatoire simple au 1/10ème.

Le plan de sondage des EMB sera amené à évoluer au cours des prochaines années, du fait de la baisse des ressources en gestionnaire au service statistique national d'enquêtes (SSNE) à Caen et de l'évolution du plan de sondage de l'EAP. En particulier, le champ de l'EAP augmente dès la campagne 2014 : les plus petites entreprises industrielles pourront être interrogées. De ce fait, il n'est plus possible pour l'EAP d'interroger l'ensemble des entreprises de son champ sur six ans comme actuellement.

[40] Dans cette note, nous utiliserons indifféremment les termes « entreprise » ou « unité légale » pour désigner les unités interrogées dans les enquêtes. Mais il s'agit bien à chaque fois d'unités légales. Les EMB, sauf exceptions, n'interrogent pas d'entreprises profilées. De même, nous utiliserons indifféremment les termes production, facturation et chiffre d'affaires pour désigner le montant de production réalisé par une entreprise dans un produit, *i.e.* le contenu de la variable MONTANTAA dans la base de sondage des EMB.

Vu ces décisions et afin d'harmoniser les champs entre enquêtes structurelles et conjoncturelles, la division Indices de chiffres d'affaires (ICA) a demandé à la division Sondages de réaliser une étude d'optimisation du plan de sondage des EMB avec un double objectif :

- optimiser, sur les 13 produits fabriqués par plus de 200 entreprises hors mécanique industrielle[41], le plan de sondage actuellement utilisé, pour améliorer la précision des estimations à taille d'échantillon fixée ;
- étudier l'effet d'une extension du champ sur les résultats des enquêtes de branche. Pour quatre produits fabriqués par moins de 200 entreprises, une collecte sur deux échantillons sera mise en place : le premier échantillon constitué suivant la procédure usuelle de *cut-off* dans les strates exhaustives de l'EAP et l'EAbis, le second avec un sondage aléatoire sur le champ formé des entreprises des strates exhaustives et des strates échantillonnées de l'EAP (EAP1+EAbis+EAP2).

L'étude porte donc sur la mise en place d'un plan de sondage optimisé pour 17 produits. Elle a été réalisée sur les données de l'EAP 2012, utilisées pour le tirage des EMB 2014.

Rappelons ici que l'EAP1 est la partie exhaustive de l'enquête annuelle de production. Il s'agit des entreprises industrielles de 20 salariés et plus ou avec un chiffre d'affaires supérieur à 5M€. Pour les secteurs dont le taux de couverture est inférieur à 85 %, la sélection est complétée, secteur par secteur, par les entreprises les plus significatives jusqu'à atteindre ce taux. Toutefois, pour le tirage des EMB, la base de sondage n'est constitué que des entreprises de 20 salariés et plus ou avec un chiffre d'affaires supérieur à 5M€.

L'EAP2 est un échantillon d'entreprises industrielles (hors entreprises sans salarié) tirées par sixième de la façon suivante :

- d'abord les entreprises créées pendant l'année précédant l'année d'interrogation ;
- puis celles n'ayant jamais été interrogées en EAP1 et EAP2 ;
- et enfin, les autres entreprises classées par date d'interrogation.

Enfin l'EAbis regroupe les entreprises industrielles des DOM ou celles dont l'APEN est non industrielle mais qui fabriquent des produits industriels.

2.1. Sur quel critère optimiser le plan de sondage ?

Pour optimiser le plan de sondage, nous avons besoin d'un critère permettant d'en évaluer la qualité. Nous avons donc choisi d'optimiser l'échantillon de façon à assurer le meilleur coefficient de variation possible dans l'estimation de la facturation totale de chaque produit. Pour estimer cette précision, nous avons réalisé, pour chaque plan de sondage étudié, 100 tirages indépendants d'un échantillon, et calculé l'estimateur et la variance de cet estimateur sur chacun des 100 échantillons.

Le coefficient de variation de l'estimateur du total t_y d'une variable y est alors estimé par :

$$CV_y = \frac{\frac{1}{100}\times\sum_{i=1}^{100}\sigma_i}{\frac{1}{100}\times\sum_{i=1}^{100}t_y^i}$$

Avec :

i représentant l'indice de l'échantillon $t_y^i = \sum_{k\in S^i} w_k y_k^i$ est l'estimateur du total de la variable y sur l'échantillon S^i sélectionné lors du i[ème] tirage et σ_i l'écart-type de t_y^i.

[41] pour laquelle le plan de sondage actuel a déjà été obtenu par une étude d'optimisation.

Les EMB ne sont cependant pas utilisées pour estimer le niveau de la production d'un produit, mais sa variation d'un mois sur l'autre. Nous proposons donc d'optimiser les plans de sondage sur une grandeur différente de la grandeur d'intérêt de l'enquête.

Cette démarche a cependant été retenue parce que

- l'optimisation du plan de sondage sur la précision de la mesure du taux de croissance de la facturation supposerait d'optimiser le plan de sondage sur la mesure de la précision du total d'une variable linéarisée. Or, nous ne savons pas si la distribution de ces variables linéarisées est stable d'une année sur l'autre et dans quelle mesure un plan de sondage plus précis qu'un autre une année le sera également une autre année ;
- pour les 13 produits fabriqués par plus de 200 entreprises, hors mécanique industrielle, le plan de sondage est déjà aléatoire : l'étude a pour but d'améliorer sa précision, en supposant que si un plan de sondage permet d'obtenir des estimateurs plus précis du total de la facturation d'un produit d'une année sur l'autre, il est peu probable qu'il soit moins précis pour estimer sa variation ;
- pour les quatre produits sur lesquels est testée l'extension de champ, nous avons essayé de comparer les précisions avec lesquelles nous estimons la variation de la production d'une année sur l'autre dans le cas d'un sondage par *cut-off* et avec un sondage aléatoire. Ainsi, nous comparons, pour les quatre produits, le biais avec lequel les échantillons cut-off nous permettent d'estimer la variation de la production annuelle avec l'intervalle de confiance que les échantillons aléatoires nous permettent de construire pour cette même grandeur compte-tenu de la variance due à l'échantillonnage.

Les plans de sondage que nous étudions sont stratifiés, avec allocation proportionnelle au chiffre d'affaires par strate.
Ainsi, pour estimer la précision que nous pouvons attendre d'un plan de sondage, nous avons décidé d'appliquer la formule résultant d'un sondage stratifié avec sondage aléatoire simple dans chaque strate : $V(t_y^i) = \sigma_i^2 = \sum_h N_h^2 (1 - \frac{n_h}{N_h}) \frac{S_h^2}{n_h}$ avec *h* les différentes strates de tirage, N_h le nombre d'observations de la strate dans la base de sondage, n_h la taille de l'échantillon dans la strate et S_h^2 la variance empirique de la variable d'intérêt dans la strate *h*, i.e. la variance empirique des facturations au cours de la période sur laquelle porte l'enquête.
Pour estimer celles-ci, nous utilisons les variances empiriques par strate observées dans la base de sondage, *i.e.* l'EAP 2012, en supposant que les distributions sont relativement stables au cours du temps.

La précision que nous obtenons ne tient pas compte du fait qu'en pratique ce n'est pas un tirage aléatoire simple mais un tirage proportionnel au chiffre d'affaires de la base de sondage qui est réalisé dans chaque strate. Le tirage proportionnel au chiffre d'affaires de la base de sondage, permet d'obtenir des estimateurs plus précis : nous surestimons ainsi la variance résultant des plans de sondage étudiés.

Nous ne tenons pas compte non plus de la non-réponse dans nos calculs de précision.

Pour chacun des 17 produits, nous avons ainsi pu comparer sur 100 simulations la performance de différents plans de sondage. Ces plans de sondage auront chaque fois la même forme. La taille de l'échantillon sélectionnée pour chaque produit est égale au nombre d'entreprises échantillonné pour le produit dans les EMB 2014. La base de sondage est d'abord divisée en une strate exhaustive, égale à l'échantillon qui aurait été obtenu avec un *cut-off* à x %, avec x variable ; le reste de l'échantillon est sélectionné suivant différentes procédures détaillées par la suite. On dira qu'une approche est plus précise qu'une autre lorsque le coefficient de variation de celle-ci est plus faible. Le but étant donc, à taille d'échantillon constante, d'obtenir le plus faible CV selon différentes approches.

2.2. Étude du plan de sondage

L'étude se basera sur l'ensemble des 13 produits présentant un nombre de siren supérieur à 200 ainsi que 4 produits ciblés. La base de sondage sera l'EAP exhaustive 2012 sauf pour les produits ci-dessous pour lesquels on constituera un échantillon issu des strates non exhaustives de l'EAP en incluant l'EAP2 afin de mettre en place un test pour comparer la série champ actuel et la série champ complémentaire.

DIVISION	UNITE	LIBELLE	PRODEMB 2014
A	A2	Fabrication de moules et modèles	2573A1R0
B	B1	Fabrication d'éléments en matière plastique pour la construction	2223Z2145000
C	C3	Instruments et appareils à usage thérapeutique	3250A1R2
D	D1	Fabrication d'autres produits minéraux non métalliques	2399z1131000

2.2.1. Approche simplifiée sans stratification

Dans cette première approche, on sélectionne un échantillon pour chaque produit de même taille que le nombre de siren appartenant à ce produit sélectionnés dans la base de sondage 2014. La strate exhaustive est donc constituée des entreprises les plus importantes jusqu'au seuil fixé puis on tire aléatoirement les autres unités de l'échantillon dans la base de sondage privée des unités sélectionnées dans la partie exhaustive. On réalise cette simulation sur 100 tirages d'échantillons puis on calcule le coefficient de variation pour chaque produit aux différents seuils.

Le tableau suivant qui s'intéresse à cette simulation sur les gros produits nous indique la précision actuelle du tirage sur ces éléments à la dernière colonne (taux de couverture de 50%).

serie	25%	50%
1624ZXR0	0,0307	0,0213
1812Z1199010	0,0404	0,0268
1812Z2125000	0,0172	0,0113
2229A2R1	0,0196	0,0140
2511Z1235040	0,0171	0,0119
2511Z2R010	0,0345	0,0232
2511Z3R011	0,0411	0,0268
2512Z1R0	0,0300	0,0194
2550B1R0	0,0261	0,0175
2561Z1R0	0,0261	0,0173
3312Z5R1	0,0555	0,0357
3320A111003A	0,0501	0,0323
3320C1R0	0,0740	0,0469

On s'aperçoit ainsi que l'on gagne en précision en augmentant le seuil[42], un *cut-off* sans stratification à 50 % apporte ici un gain de précision tout en conservant la taille de l'échantillon.

La même approche pour les quatre produits enquêtés sur une base exhaustive nous confirme les résultats sur des produits moins contributifs à l'EAP mais très représentatifs.

serie	25%	50%	65%
2223Z2145000	0,05132	0,02225	0,01614
2399Z1131000	0,02583	0,02008	
2573A1R0	0,02774	0,01950	0,01499
3250A1R2	0,04684	0,02574	

En effet, la précision s'améliore avec le taux de couverture et ce, même dans les cas où l'on peut monter à un seuil de 65 % en conservant la taille de l'échantillon. Seulement, vu le faible nombre d'entreprises à tirer aléatoirement, on considère ici que le seuil à 50 % est une meilleure approche de la méthode à adopter.

2.2.2. Approche avec stratification de la partie non exhaustive selon le montant

Dans cette seconde approche, nous avons stratifié la partie non exhaustive de chaque produit en fonction des quartiles de la distribution de ses facturations. Les allocations de l'échantillon dans chaque strate sont proportionnelles à la facturation. Un sondage aléatoire simple est ensuite réalisé dans chaque strate ; nous obtenons donc cinq strates en tout avec celle exhaustive pour chaque produit dont les tailles varient en fonction de la taille du produit dans l'échantillon et du poids de la strate dans la base.

La simulation est similaire avec la précédente dans la mesure où l'on tire aussi 100 échantillons pour chaque produit qui nous permettent de calculer un CV aux différents taux de couverture.

serie	25%	50%	65%
2223Z2145000	0,02187	0,01887	0,01583
2399Z1131000	0,02268	0,01954	
2573A1R0	0,01872	0,01643	0,01439
3250A1R2	0,03322	0,02302	

On peut voir que l'on garde la même tendance : plus on augmente la taille de strate exhaustive, plus on est précis. Seulement le choix du seuil à 50 % est encore préféré ici pour les mêmes raisons précédentes. De plus, on confirme qu'en stratifiant on gagne en précision car ces CV sont plus faibles que ceux obtenus dans l'approche précédente sans stratification sur ces produits.

[42] Ce résultat n'est pas surprenant dans la mesure où le *cut-off* favorise la précision de l'estimation d'un total France entière mais avec un biais qui n'est pas mesuré ici.

Le tableau suivant représentant la même simulation sur les produits les plus importants confirme ce point.

Serie	25%	50%
1624ZXR0	0,0246	0,0204
1812Z1199010	0,0293	0,0245
1812Z2125000	0,0127	0,0102
2229A2R1	0,0161	0,0133
2511Z1235040	0,0134	0,0113
2511Z2R010	0,0267	0,0218
2511Z3R011	0,0300	0,0244
2512Z1R0	0,0208	0,0175
2550B1R0	0,0200	0,0164
2561Z1R0	0,0208	0,0166
3312Z5R1	0,0363	0,0310
3320A111003A	0,0371	0,0299
3320C1R0	0,0542	0,0438

En effet, on améliore bien la précision en augmentant le seuil de couverture et on obtient de meilleurs résultats que dans l'approche précédente. Ainsi, le choix d'un taux de couverture à 50 % du total du montant d'un produit pour constituer une strate exhaustive et un tirage aléatoire stratifié selon les quartiles et avec une allocation proportionnelle au montant pour la partie non exhaustive paraît jusqu'ici l'approche la plus précise de l'étude.

2.2.3. Étude au niveau du seuil naturel

Dans le cas d'un sondage aléatoire proportionnel au montant, le seuil naturel représente le taux de couverture des entreprises d'un produit K qui ont un poids de tirage supérieur à 1 :

$n \frac{x_i}{\sum_{i \in K} x_i} > 1$ avec n la taille de l'échantillon et x_i le montant d'une entreprise i

Le seuil naturel est intéressant à étudier dans la mesure où les valeurs retenues pour la définition des taux de couverture de l'exhaustif nous paraissent arbitraires. Ainsi, on obtient la strate exhaustive la plus naturelle et on peut y effectuer la simulation précédente du calcul du CV sur 100 échantillons.

Les résultats sont obtenus dans le tableau suivant et nous permettent de comparer avec ceux obtenus avec des seuils arbitraires.

série	seuil naturel (%)	CV sans stratification	CV avec stratification	Taille strate exhaustive
gros produits				
1624ZXR0	16,2395	0,0343	0,0256	4
1812Z1199010	24,1260	0,0409	0,0292	6
1812Z2125000	18,3080	0,0186	0,0128	9
2229A2R1	15,1182	0,0221	0,0166	5
2511Z1235040	20,9834	0,0182	0,0140	10
2511Z2R010	24,6609	0,0347	0,0267	7
2511Z3R011	22,2377	0,0431	0,0299	5
2512Z1R0	25,8502	0,0289	0,0204	9
2550B1R0	28,4156	0,0244	0,0195	11
2561Z1R0	23,0589	0,0267	0,0209	8
3312Z5R1	33,6904	0,0465	0,0345	7
3320A111003A	31,7058	0,0436	0,0350	7
3320C1R0	26,5622	0,0687	0,0534	4
4 produits enquêtés				
2223Z2145000	30,1634	0,0304	0,0212	9
2399Z1131000	16,5333	0,0275	0,0233	4
2573A1R0	22,5012	0,0287	0,0188	7
3250A1R2	24,3507	0,0470	0,0331	7

On voit alors que la tendance est toujours la même : on est plus précis en stratifiant. Cependant, même si le seuil ici est naturel, on préférera le fixer à 50% (c'est-à-dire le choix de l'approche précédente) car les résultats obtenus sont meilleurs et le nombre d'entreprises à tirer dans la partie échantillonnée est raisonnable. De plus, cette méthode nous permettrait d'avoir le même seuil de couverture pour chaque produit.

2.2.4. Approche avec stratification optimisée

Dans cette nouvelle approche, nous allons stratifier chaque produit selon la méthode de Dalenius afin d'obtenir un découpage optimisé des strates. Il s'agit de la méthode de découpage optimal de strate la plus connue. Elle est utilisée à l'Insee pour l'enquête mensuelle sur les grandes surfaces alimentaires et pour l'enquête mensuelle de branche pour le secteur de la « mécanique industrielle » et l'on essaie ici de l'appliquer à nos différents produits.

Ainsi, après avoir trouvé les bornes des 4 strates optimales en fonction du seuil de couverture désiré, on effectue un tirage aléatoire proportionnel au montant sur ces strates pour obtenir le CV correspondant qui nous permettra d'évaluer la précision à nouveau.

serie	NAT	25%	50%
2223Z2145000	0,0208239	0,0205830	0,0185661
2399Z1131000	0,0232168	0,0229327	0,0194501
2573A1R0	0,0180270	0,0181649	0,0164007
3250A1R2	0,0263711	0,0263276	0,0222579

On peut directement s'apercevoir que les résultats sont meilleurs. En effet, même si la différence n'est pas significativement importante, le gain de précision nous permet d'envisager cette méthode un peu plus complexe par rapport à une stratification simple sur les quartiles du montant.

serie	NAT	25%	50%
1624ZXR0	0,024364	0,024431	0,020234
1812Z1199010	0,028073	0,028033	0,024320
1812Z2125000	0,012334	0,012314	0,010234
2229A2R1	0,016451	0,015983	0,013313
2511Z1235040	0,013240	0,013163	0,011258
2511Z2R010	0,024769	0,024691	0,021699
2511Z3R011	0,027428	0,027471	0,023645
2512Z1R0	0,019114	0,019173	0,017189
2550B1R0	0,018002	0,018204	0,016229
2561Z1R0	0,019520	0,019602	0,016511
3312Z5R1	0,031929	0,032311	0,030134
3320A111003A	0,032252	0,031108	0,029457
3320C1R0	0,047693	0,047133	0,043289

Le tableau sur les gros produits confirme la tendance : une amélioration en précision (même si elle est peu significative) qui nous pousse à retenir cette méthode comme l'approche la plus adéquate. C'est donc la précision affichée par la dernière colonne de ce tableau (taux de couverture à 50%) qui sera retenue pour la nouvelle méthode de tirage de ces produits. Nous avons alors décidé d'aller plus loin pour voir si nous obtenions une amélioration plus significative en augmentant le nombre de strates à 6 pour les quatre produits enquêtés. Il s'avère que les résultats sont à nouveau meilleurs mais la différence est encore très faible et surtout la faible taille des échantillons engendre des strates très faibles en effectif.

serie	NAT	25%	50%
2223Z2145000	0,020787	0,019900	0,018482
2399Z1131000	0,022969	0,022506	0,019206
2573A1R0	0,017771	0,017941	0,016409
3250A1R2	0,025903	0,025887	0,021987

2.3. Comparaison des précisions des méthodes

Pour les quatre produits fabriqués par moins de 200 entreprises sur lesquels va être testée une double collecte, nous avons souhaité comparer *a priori* les précisions des estimateurs issus des deux types d'échantillon. Nous cherchons ainsi à estimer la variation du chiffre d'affaires dans le produit déclaré par l'ensemble des entreprises de l'EAP1, l'EAbis et de l'EAP2 entre 2011 et 2012. Nous comparons deux estimateurs de cette variation :

1. le premier est obtenu si les productions du produit en 2011 et 2012 sont issues à chaque fois d'un échantillon sélectionné par *cut-off* ;

2. le second est obtenu si les facturations du produit chacune des deux années sont estimées par un échantillon sélectionné par sondage aléatoire simple suivant le plan de sondage proposé dans cette note.

L'échantillon sélectionné par *cut-off* n'est pas aléatoire, aussi les estimateurs qu'il permet d'obtenir n'ont aucune variance (vu qu'on ne prend pas en compte la non-réponse dans notre étude). En revanche, ils sont biaisés, en particulier si les petites entreprises, qui ne sont jamais interrogées, ont une variation de la production différente des grandes.

L'échantillon sélectionné aléatoirement permet à l'inverse d'obtenir des estimateurs sans biais, mais affectés d'une variance d'échantillonnage que nos simulations nous ont permis d'estimer. En supposant que les échantillons de deux années successives sont tirés indépendamment l'un de l'autre, nous pouvons approximer la variance avec laquelle nous estimons la variation de la facturation annuelle par 2*Variance de l'estimateur annuel. Remarquons que cette estimation de variance se place dans le cas le plus défavorable, où aucune stratégie de renouvellement, et donc de corrélation, des échantillons successifs n'est mise en place. Nous surestimons ainsi la variance obtenue avec deux échantillons aléatoires.

Ainsi, nous pouvons calculer la « vraie » valeur que nous cherchons à estimer, et la comparer d'une part à l'estimateur biaisé obtenu avec les échantillons *cut-off* et d'autre part à l'intervalle de confiance que nous pouvons construire avec les échantillons aléatoires.

On sait que l'écart quadratique moyen (EQM) est la somme du biais au carré et de la variance :

EQM = $biais^2$ + Var

Ainsi on voit que l'EQM de la méthode du *cut-off* sur les estimateurs des années d'enquête 2013 et 2014 prend en compte seulement le biais car la variance est nulle (dans l'hypothèse ou nous ne sommes pas en présence de non-réponse) par définition de la méthode ainsi, on obtient :

$$\text{biais} = ((t^{2012}_{y,exhaustive} - t^{2011}_{y,exhaustive}) - (t^{2012}_{y,cutoff} - t^{2011}_{y,cutoff}))^2$$

avec t représentant le total du montant d'un produit

Le biais étant vu ici comme la différence des deux différences entre les bases exhaustives des deux années et les bases *cut-off* (à 75% du montant de chaque produit).

Dans une autre approche, en considérant la méthode du sondage stratifié par quartiles, l'EQM ne présente pas de biais mais une variance de la différence des estimateurs 2012 et 2011. Les tirages seront supposés indépendants et les variances de l'estimateur environ égales sur les deux années, ainsi :

$$EQM = V(y_{2012} - y_{2011}) = V(y_{2012}) + V(y_{2011}) \approx 2\,V(y_{2012}) = 2\,(\hat{y}_{2012}\,CV)^2$$

avec $\hat{y}_{2012}$ l'estimateur du total du produit et CV les coefficients de variation calculés sur les produits dans l'approche avec stratification sur les montants.

On peut alors en dégager un intervalle de confiance à 95 % près de la précision qui nous permet de dire si le biais engendré par la méthode du cutoff est tolérable en comparaison aux résultats de la méthode par tirage stratifié non optimisé.

série	$t_{2012} - t_{2011}$	biais	intervalle de confiance
2223Z2145000	-220 272	-75 179	[-280 415,78 ; -160 128,22]
2399Z1131000	222 877	63 657	[181 261, 68 ; 264 492,32]
2573A1R0	-102 158	-26 262	[-118 093,08 ; -86 222,92]
3250A1R2	19 403	4 934	[-36 855,2 ; 75 661,2]

Dans les trois premiers produits, on peut voir que le biais de la méthode par *cut-off* est très élevé et que si on l'ajoute à la différence des totaux on voit que cette somme n'est pas comprise dans l'intervalle de confiance déterminé à partir des résultats sur la précision de la méthode aléatoire. Cependant, le dernier produit nous montre les limites dans la mesure où l'on obtient un biais beaucoup plus faible : cela pourrait s'expliquer par le grand effectif d'entreprises de l'EAP2 de ce produit. En effet, l'ajout massif d'entreprises pour ce produit pourrait conduire à une mauvaise approximation de l'EQM et de l'indépendance des tirages.

série	Taux de croissance	Biais du taux de croissance	Intervalle de confiance du taux de croissance
2223Z2145000	-0,10564	-0,04920	[-0,134483 ; -0,076797]
2399Z1131000	0,20077	0,08639	[0,163283 ; 0,238257]
2573A1R0	-0,15638	-0,05492	[-0,180773 ; -0,131987]
3250A1R2	0,01240	0,00445	[-0,023556 ; 0,048356]

Ici, nous avons décidé de représenter les différents estimateurs précédents sous forme d'évolution pour avoir des résultats plus abordables. Ainsi, on peut avoir une confirmation de nos résultats précédents : un taux de croissance très faible pour le dernier produit qui est également le seul dont le biais ajouté au taux de croissance est dans l'intervalle de confiance.

Conclusion

Les trois méthodes de découpage optimal de strates implémentées dans le package *stratification* sont assez simples à utiliser avec un minimum de connaissance du logiciel R. La notice du package comporte de nombreux exemples de code qu'il est possible de recopier et de modifier facilement.

Les méthodes de Dalenius et géométrique aboutissent à des bornes optimales indépendantes du niveau de précision recherché par l'utilisateur mais donneront systématiquement des résultats moins bons que la méthode LH qui utilise ces deux dernières comme points de départs à des recherches itératives de meilleures bornes.

Cependant les méthodes de Dalenius et géométrique ont leurs avantages. Par exemple, il n'est pas nécessaire de connaître la valeur exacte de y_k pour chaque observation avec la méthode de Dalenius : on peut se contenter d'un découpage fin en tranches. La méthode géométrique a l'avantage d'être très simple et nécessite très peu de temps de calcul. Elle peut facilement être introduite dans des processus itératifs comme nous l'avons fait par exemple pour optimiser le nombre de strates dans la partie 1.6. De plus, les résultats obtenus avec ces deux méthodes ne dépendent pas de la taille d'échantillon et le découpage obtenu peut donc servir pour plusieurs enquêtes. Enfin, ces deux méthodes sont de toute façon des étapes obligées pour pouvoir utiliser la méthode LH et apprécier l'efficacité de l'algorithme itératif.

La méthode LH a elle aussi ses avantages. Elle s'avère plus souple que les autres méthodes et on peut davantage la contraindre. Par exemple, il est possible d'imposer que les strates soient d'une certaine taille minimale (minNh), ce qui n'est pas possible avec la méthode géométrique et peu réalisable avec la façon dont est implémentée la méthode de Dalenius dans le package. Le gros avantage de cette méthode reste la possibilité d'une recherche de seuil d'exhaustivité optimal en imposant que la $L^{ème}$ strate soit tirée exhaustivement.

En jouant sur le nombre de strates, on peut pousser l'exercice théorique de minimisation de taille d'échantillon pour atteindre un objectif de précision donné jusqu'au bout. Les gains réalisés sont impressionnants, mais deviennent très liés aux données, ce qui les rend peu utilisables. En pratique, le nombre de strates devrait plutôt être un paramètre fixé par l'utilisateur suivant des contraintes du type respect d'un certain nombre d'unités tirées par strate etc…

Dans tous les cas, Il est important de contrôler l'optimisation des limites de strates par des calculs de précision sur d'autres variables que la variable de stratification elle-même afin de ne pas rendre trop spécifique le plan de sondage et de maitriser à la fois les gains et les pertes réalisés. La possibilité, non étudiée dans cet article, de tenir compte de certains types de liens entre la variable d'intérêt et la variable de stratification dans l'optimisation pourrait s'avérer utile dans cette optique.

Pour être pleinement utilisable à l'Insee, il serait intéressant de modifier le code du programme de façon à pouvoir imposer un nombre minimum d'unités à tirer par strate dans les calculs de variance et donc pour l'optimisation. Le fait que R soit un logiciel libre rend en théorie possible cette modification mais cela nécessiterait un investissement conséquent afin de comprendre tous les rouages du programme.

Des utilisations autres que l'optimisation du plan de sondage proprement dite pourraient être envisagées. Par exemple, la méthode LH en imposant l'exhaustivité de la $L^{ème}$ strate aurait pu être utilisée lors de la recherche de seuils d'exhaustivité pour le chiffre d'affaires réalisée pour l'enquête sur les technologies de l'information et de la communication 2012. Le package pourrait aussi être utilisé pour remplacer des tirages à probabilité proportionnelle à la taille par un tirage stratifié avec optimisation du nombre de strates ce qui permettrait d'utiliser des méthodes de coordination d'échantillons.

Dans l'application aux EMB, nous avons pu voir que le choix de la stratification jouait beaucoup sur la précision obtenue. Ainsi, en travaillant à taille d'échantillon constante, et en passant d'une absence de stratification jusqu'à une stratification optimisée par Dalenius, nous observons que les variations de

précision sont suffisamment importantes pour suggérer un changement de méthode de sondage pour cette enquête.

En effet, il est logique de penser que plus on stratifie plus on sera précis, ce qu'on l'on voit bien à nouveau ici. Cependant, le choix d'un nombre de 4 strates a du être effectué dans la mesure où la faible taille des échantillons entraînait un trop faible effectif par strate si on augmentait le nombre de strates même si l'on gagnait en précision. De plus, nous pouvons voir que le choix d'une stratification optimale est logique comparé à une stratification simple sur les quartiles car même si la différence n'est pas significativement importante, le gain de précision nous permet d'envisager cette méthode un peu plus complexe par rapport à une stratification simple sur les quartiles du montant.

Ce gain de précision par cette nouvelle méthode est traduit par les calculs des intervalles de confiance sur la variation des totaux du chiffre d'affaires entre les années 2011 et 2012.
Les résultats obtenus nous permettent de dégager que le plan de sondage qui conduit à stratifier la base de manière optimale selon Dalenius sur le chiffre d'affaires au sein de chaque produit avec une allocation proportionnelle au chiffre d'affaires (en respectant le cutoff à 50%) est le plus intéressant.

Ainsi, les travaux menés sur les trois méthodes de découpage optimal de strates ont permis d'élaborer un nouveau plan de sondage à une enquête qui manquait de précision. Pour l'année prochaine, une étude de la répartition de l'échantillon par produit pourra constituer une nouvelle source d'amélioration de la qualité de l'enquête à taille d'échantillon fixée.

Bibliographie

[1] Demoly E., Fizzala A., Gros E., « *Méthodes et pratiques des enquêtes entreprises à l'Insee* », Journal de la société française de statistiques N°155, 2014.

[2] Dalenius, T., et Hodges, J.L., Jr., « *Minimum variance stratification* », Journal of the American Statistical Association, 54, 88-101, 1959.

[3] Lavallée, P., et Hidiroglou, M., « *Sur la stratification de populations asymétriques* ». Techniques d'enquête, 14, 35-45, 1988.

[4] Gunning, P., et Horgan, J.M. « *Un nouvel algorithme pour la construction de bornes de stratification dans les populations asymétriques* ». Techniques d'enquête, 30, 177-185, 2004.

[5] Kozak, M., et Verma, M.R., « Approche de la stratification par une méthode géométrique et par optimisation : une comparaison de l'efficacité ». Techniques d'enquête, 32, 177-183, 2006.

[6] Baillargeon S. et Rivest L-P, « *Élaboration de plans stratifiés en R à l'aide du programme stratification* », Techniques d'enquête, 37, 59-72, 2011.

[7] Baillargeon S., Rivest L-P, Notice du package stratification, logiciel R, 2011.

[8] Rivest, L.-P., « *Une généralisation de l'algorithme de Lavallée et Hidiroglou pour la stratification dans les enquêtes auprès des entreprises* ». Techniques d'enquête, 28, 207-214, 2002.

[9] Cochran, W.G., « *Sampling Techniques* », troisième Édition. New York : John Wiley & Sons, Inc, 1977.

[10] Kozak, M. « *Optimal stratification using random search method in agricultural surveys* ». Statistics in Transition, 6(5), 797-806, 2004.

[11] Baillargeon S., Rivest L-P., Ferland M. « *Stratification en enquêtes entreprises : Une revue et quelques avancées* », Assemblée annuelle de la Société Statistique du Canada, 2007.

[12] Ardilly P., « *Les techniques de Sondage* », Éditions TECHNIP, 2006.

[13] Fizzala A., « *Découpage optimal d'une variable quantitative pour la stratification : comparaison de méthodes sur les données d'entreprises françaises* », Mémoire professionnel du master de Statistique publique de l'Ensai, 2013.