Le traitement des regroupements de réponses lors d'enquêtes auprès des entreprises

Henri Bodet[1] & Theo Leroy[2]

[1] *Pôle Ingénierie Statistique d'Enquête, Insee Pays de la Loire, 105, rue des Français Libres - 44274 Nantes Cedex 2, henri.bodet@insee.fr*
[2] *Élève de l'Ensai, Rue Blaise Pascal, 35170 Bruz, theo.leroy@eleve.ensai.fr*

Résumé. Lors d'enquêtes auprès des entreprises portant sur des questions environnementales comme l'enquête annuelle sur les consommations d'énergie dans l'Industrie ou l'enquête sur les dépenses faites pour protéger l'environnement, on accepte parfois le regroupement de réponses. Il ne s'agit pas de permettre à plusieurs unités interrogées de transmettre une réponse simultanément mais de permettre à une unité de ne transmettre que la réponse agrégée.

Cette pratique permet, on l'espère, de réduire la charge statistique pour les entreprises mais également d'augmenter le taux de réponse. La contrepartie de l'obtention d'une donnée exacte à un niveau agrégé est une perte de précision de l'information à un niveau local ; ce qui semble satisfaisant, l'objectif de ces enquêtes étant d'obtenir des résultats nationaux.

L'objet de cette communication est de présenter les traitements mis en œuvre pour utiliser ces données agrégées et les conditions nécessaires pour que les estimations obtenues ne soient pas trop biaisées. On présentera une évaluation par simulation, à partir de données d'enquête, du biais obtenu en fonction de la part de ces regroupements dans l'échantillon ainsi qu'une évaluation de l'effet de cette pratique sur la charge statistique. Nous proposons des indicateurs pour repérer en cours de collecte les situations qui risquent de poser problème et nous envisageons l'application de la méthode généralisée de partage des poids.

Mots-clés.Méthodes de collecte, Enquêtes auprès des entreprises, Enquêtes sur l'environnement

Abstract. In business surveys on environment-related topics - such as Energy Consumption in Industry or Environmental Protection Expenditure Survey - grouped responses are sometimes accepted. This does not mean that individual responses are transmitted together but that a single response is transmitted for several business units – without any indication of their individual contribution to the total.

Doing so enables - or so we hope - to reduce the statistical burden of surveyed businesses but also to improve the response rate. The price to pay for it is a loss of accuracy at a local business unit level while we get an actual figure at an aggregated level. This seems to be a good deal since such surveys are designed to produce nationwide statistics.

In this paper, we will present the procedures which are used to deal with such data and which are the necessary conditions that have to be met in order to keep the bias at a low level. We will present an evaluation of the bias according to the share of grouped responses. The simulation was done from genuine business survey data. We will also evaluate the actual alleviation of the statistical burden enabled by the grouping of responses. We introduce a statistical indicator which can be used to spot grouping likely to provide a large bias. We look at the possibility to use a weighting sharing method.

Keywords. Data collection methods, Business surveys, Environment surveys

1 Les regroupements de réponses dans les enquêtes auprès des entreprises

1.1 Qu'est-ce-que c'est qu'un regroupement de réponses ?

La pratique du regroupement de réponses s'est développée principalement pour les enquêtes relatives à la protection de l'environnement. En particulier, pour les enquêtes suivantes :

- l'enquête Antipol destinée à connaître les dépenses des entreprises pour protéger l'environnement ;
- l'enquête annuelle sur les consommations d'énergie dans l'Industrie (dite EACEI) ;
- l'enquête sur les déchets non-dangereux de l'Industrie et du Commerce.

Cette pratique consiste à admettre qu'une unité réponde non pas pour elle-même mais pour un ensemble d'autres. Elle ne transmet alors pas la réponse pour chaque unité de ce "regroupement" mais bien pour le total. Il ne s'agit pas ici d'une réponse par grappe ou d'un vrai sondage indirect mais bien d'une réponse "dégradée" puisqu'on ne dispose que d'un total et non pas de la valeur pour chaque unité. De plus, on n'est pas sûr que les unités composant le regroupement soient bien dans l'échantillon - ni même dans le champ de l'enquête.

Originellement, cette tolérance trouve sa justification dans le fait que la donnée recherchée peut n'être pas connue au niveau de l'unité de collecte. C'est par exemple le cas de l'enquête sur les consommations d'énergie décrite ci-après.

Les avantages recherchés par cette tolérance sont les suivants :

- réduire la charge de réponse pour les entreprises ;
- améliorer le taux de réponse ;
- ne pas perturber l'estimation des agrégats car, même si on ne connait pas le détail, le total demeure exact.

Dans certaines enquêtes, les regroupements peuvent représenter presque 7 % de l'échantillon. Voici ce qu'il en est pour quelques enquêtes récentes :

Enquêtes	Taille de l'échantillon	Nombre d'unités regroupantes	Nombre d'unités regroupés
Développement Durable 2016	11 009	24	97
EACEI 2015 (extension régionale)	13 990	487	848
EACEI 2016	8 505	401	670
EACEI 2017	8 504	386	614
Antipol 2015	11 027	505	1 004

TABLE 1 – Quantification du regroupement

1.2 La méthode de proratisation des réponses

La méthode empirique pour les traiter consiste à répartir les agrégats entre les unités du regroupement au prorata d'une variable auxiliaire (en pratique, leur effectif salarié). On utilise ensuite les résultats obtenus pour les unités faisant partie de l'échantillon comme si c'était leur réponse. C'est une procédure raisonnable et simple à mettre en œuvre qui permet *a priori* de conserver le total. En particulier, une fois ce traitement effectué, les opérations post-collecte se déroulent normalement.

On peut être tenté d'inciter les services de collecte à développer cette pratique auprès des entreprises - puisque cela augmente le taux de réponse, réduit la charge en n'introduisant qu'un biais que l'on peut supposer léger voire nul.

2 L'enquête annuelle sur les consommations d'énergie

Dans cette partie, nous donnons quelques explications sur l'enquête qui nous sert à illustrer cet article : l'enquête annuelle sur les consommations d'énergie dans l'industrie dite EACEI.

Le questionnaire contient quatorze cadres qui correspondent à un type d'énergie (électricité, gaz naturel, houille, bois, etc.) et deux autres qui correspondent aux énergies non mentionnées dans les précédents (biogaz, combustion de déchets,..).

Le questionnaire de l'enquête est visible sur le site de l'Insee :
https://www.insee.fr/fr/statistiques/3364874documentation-sommaire

Les résultats sont disponibles sur le site de l'Insee :
https://www.insee.fr/fr/statistiques/3364874 L'objectif de cette enquête est de fournir des statistiques nationales ventilées par type d'activité.

L'échantillon est tiré suivant un plan de sondage stratifié. Tous les établissements industriels de plus de 250 salariés sont interrogés et les autres sont échantillonnés (certaines strates sont exhaustive du fait de l'optimisation de l'allocation).

La population-cible compte 22 000 établissements et l'échantillon en comporte 8 500. Le taux de réponse est de 90 %. La non-réponse n'est donc pas un problème de cette enquête - sauf si elle porte sur de grandes unités. Celles-ci étant interrogées tous les ans, on dispose d'assez d'information pour corriger leur non-réponse.

Pour les enquêtes auprès des entreprises, il y a toujours un choix à faire en matière d'unité de collecte qui est la plupart du temps l'établissement ou l'unité légale. Ici, l'établissement a été choisi comme unité de collecte car cela permet de diffuser quelques résultats régionaux mais surtout parce que l'information sur les consommations d'énergie est réputée être plus facilement mobilisable à ce niveau. Toutefois, il est parfois plus simple pour une entreprise d'avoir l'information à un autre niveau. Par exemple, si une énergie est achetée par le siège, elle ne dispose vraisemblablement que d'une seule facture.

Pour diminuer la charge de réponse, le service statistique a donc autorisé les entreprises à répondre pour un ensemble d'établissements. Offrir cette possibilité contribue certainement à l'excellent taux de réponse de cette enquête. De plus, l'information à un niveau détaillée n'est pas forcément disponible.

En 2015, l'enquête a bénéficié d'une extension exceptionnelle : 14 000 établissements ont été interrogés. L'objectif était de disposer de statistiques infra-régionales et d'information sur les énergies les plus rares. Ce taux de couverture exceptionnel a aussi été utilisé pour améliorer l'échantillonnage en intégrant des établissements *a priori* de petite taille mais recourant de façon importante à une énergie. Il fournit aussi un nombre important d'unités légales dont on a interrogé plusieurs établissements. Cette situation a été utilisée dans cet article pour évaluer l'impact des procédures de traitement des regroupements.

3 La réduction de la charge de réponse induite par le regroupement

Dans cette partie, nous cherchons, à travers le cas d'une enquête réelle, à évaluer un des avantage attendus du regroupement : réduire la charge qui pèse sur les entreprises.

En effet, si la possibilité de fournir des réponses groupées a été offerte pour gérer les situations où l'établissement interrogé ne dispose pas des données, c'est aussi une facilité qui peut permettre d'épargner à certains établissements d'avoir à répondre.

Nous nous proposons d'évaluer le gain en matière de charge de réponse. Pour cela, nous utiliserons la question - systématiquement posée - sur le temps consacré à la réponse au questionnaire. Ce temps de réponse est utilisé pour mesurer la charge que l'enquête fait peser sur l'entreprise.

Lorsqu'une entreprise omet de répondre à cette question, une valeur est estimée en fonction de son profil et du type de questions auxquelles elle a répondu. De la sorte, on tient compte de la charge de réponse de toutes les entreprises même celles qui n'ont pas répondu à cette question située à la toute fin du questionnaire.

3.1 Quels sont les déterminants du temps de réponse ?

Sur l'EACEI - comme sur d'autres enquêtes - le premier déterminant est le nombre de questions répondues qui correspond dans le cas de cette enquête au nombre de sources d'énergie que l'entreprise utilise. Une fois ce facteur contrôlé, la taille de l'établissement et le secteur ne sont pas significatifs.

Toutefois, même à nombre de sources d'énergie identiques, les regroupants ont tendance à mettre plus de temps à rechercher les données que les non-regroupants. Ceci peut s'expliquer par le fait que, même si une seule unité répond, le fait de coordonner les réponses prend du temps. En effet, la situation n'est sûrement pas la même pour toutes les énergies et si pour certaines, les renseignements sont disponibles au niveau de l'unité regroupante, pour d'autres, elle peut être amenée à rassembler de l'information au niveau des unités regroupées.

La table 2 donne les résultats sur l'EACEI 2016.

nombre de sources d'énergies	non regroupants	regroupants
aucune	26	18
1	46	62
2	56	77
3	80	94
4	110	168
5	92	///

TABLE 2 – EACEI 2016 -temps de réponse moyen en minutes suivant la situation de l'établissement

Ce tableau montre clairement deux choses : plus on est concerné par des sources d'énergies différentes, plus le temps de remplissage augmente (vingt à trente minutes supplémentaires pour un cadre supplémentaire) et, à nombre de sources d'énergies équivalent, le fait de regrouper des réponses augmente le temps de réponse de quinze à vingt minutes.

Si toutes les unités du regroupement sont dans l'échantillon, il y forcément un gain dans la mesure où le surcoût du regroupement en temps de remplissage est inférieur à celui d'une autre réponse. Malheureusement, en pratique et sur cette enquête, beaucoup d'établissement regroupés ne figurent pas dans l'échantillon.

Il n'est donc pas évident que le regroupement soit un si fort gain de temps pour les entreprises : répondre pour d'autres entraîne un surcroît de travail et un bon nombre d'unités incluses dans les regroupements ne sont pas dans l'échantillon.

3.2 Évaluation de l'impact du regroupement sur la charge statistique

Pour évaluer le gain, nous nous servirons de la procédure d'imputation existante pour obtenir un contrefactuel : le temps de réponse qu'il y aurait eu sans regroupement.

Nous avons imputé à chaque unité regroupée ou regroupante le temps de réponse d'une unité dans une configuration similaire non-regroupante.

La statistique d'intérêt sera la somme des temps de réponses sur l'échantillon : il représente la charge en minutes de l'enquête sur les entreprises. Nous prendrons en compte les valeurs imputées - en remplaçant les valeurs déclarées par les regroupantes par des valeurs que l'on aurait imputées à des établissements non-regroupants dans la même situation.

situation	nombre d'établissement	temps de réponse réel	temps de réponse sans les regroupements	effet du regroupement
ensemble	7 327	437 931	439 417	- 0,5 %
unités ni regroupantes ni regroupées	6 783	405 656	405 656	0 %
regroupantes	401	32 275	24 621	+ 30 %
regroupées	143	0	9 410	- 100 %

TABLE 3 – EACEI 2016 -simulation de l'impact du regroupement sur le temps de réponses total (en minutes)

Cet exercice - très simple et qui pourrait être approfondi - montre que l'impact général est négligeable et que le gain de temps que l'on estime à 9 000 minutes-personne est du même ordre de grandeur que la surcharge de travail pour les regroupants. Ceci s'explique notamment par le fait que les regroupés sont souvent situés en dehors de l'échantillon.

3.3 Impact du regroupement sur le service enquêteur

Dispenser les entreprises regroupées de répondre se paie par un surcroît de charge pour celles qui regroupe mais également pour le service enquêteur. Il ne nous a pas été possible

d'évaluer cette charge mais elle dépasse certainement 2 % du temps - alors qu'*in fine* cela ne dispense de répondre que 2 % des établissements.

L'expérience est, en effet, que ces regroupement occupent une bonne partie des échanges post-collecte et rajoutent à la complexité des traitements.

3.4 Conclusion sur l'impact du regroupement sur la charge statistique

L'expérience que nous avons faite est limitée pour les raisons suivantes :

— la procédure d'imputation présente une marge d'erreur qu'il aurait fallu évaluer, par exemple en faisant plusieurs simulations ;

— il y a certainement un biais de sélection : on peut supposer que les unités regroupées sont celles pour qui la fourniture de renseignements individuels serait plus difficile.

Toutefois, on ne peut pas nier que, si l'objectif du regroupement est d'alléger la charge de réponse des entreprises, il n'est pas évident qu'il soit atteint et il ne peut pas être un argument suffisant pour étendre encore plus cette méthode.

On voit aussi que ce résultat contre-intuitif (le regroupement de réponse ne réduit pas la charge statistique) s'explique surtout par le fait que la majorité des unités regroupées ne sont pas dans l'échantillon.

Cette rapide investigation devrait donc être complétée pour intégrer plusieurs dimensions :

— l'impact sur charge de réponse des les entreprises ;

— l'augmentation de la charge pour le service enquêteur ;

— la réduction de variance induite par l'augmentation du taux de réponse ;

— le biais introduit par le regroupement (que l'on est en mesure d'estimer en appliquant les développement présentés plus loin).

Ceci, et le biais qu'entraîne le traitement des regroupés absents de l'échantillon, milite pour faire en sorte que les regroupements soient concentrées sur des unités présentes dans l'échantillon.

4 L'évaluation numérique du biais induit par le regroupement sur des données réelles

Le biais introduit par la méthode de proratisation pour traiter le regroupement de réponses peut être estimé par simulation. Des regroupements de réponses sont reproduits en agrégeant des réponses individuelles issues d'une enquête. Le tirage des regroupement de réponses est réalisé de la manière suivante :

1. Choix du regroupant parmi tous les regroupants possibles (établissements dans l'échantillon non présent dans un groupe déjà construit)

2. Choix du nombre n_g d'établissement formant la réponse groupée
3. Choix des $n_g - 1$ unités regroupées par le regroupant

Les groupes constitués par simulation doivent imiter ceux effectivement rencontrés dans les enquêtes passées. Les établissements regroupés dans un groupe de réponse sont choisis au regard des caractéristiques du regroupant (unité légale, taille de l'effectif salarié, secteur d'activité...).

Une fois que tous les appariements souhaités ont été construits, on calcule des indicateurs statistiques (somme, proportion...) à partir des unités présentes dans un regroupement et des autres. Ces deux tables sont disjointes et leur union correspond à l'échantillon.

Les simulations ont été réalisées à l'aide du logiciel R à partir du fichier de mise à disposition de l'EACEI. L'implémentation de l'algorithme a été parallélisée afin d'accélérer les calculs. Cette parallélisation est possible car les simulations d'un nouveau fichier de réponses en regroupant puis dégroupant plusieurs réponses sont indépendantes.

Nous présentons ici des simulations de regroupement de réponses d'établissement appartement exclusivement à la même unité légale dans l'échantillon. Le regroupant est choisi selon une probabilité uniforme dans l'échantillon contenant exclusivement les réponses exploitables. Le nombre de regroupés est déterministe. On prend le nombre d'établissements de la même unité légale que le regroupant présent dans l'échantillon moins un (on enlève le regroupant). Les regroupés sont alors tous les établissements de la même unité légale que le regroupant présents dans l'échantillon excepté lui-même.

4.1 Impact sur une variable quantitative synthétique de l'enquête

La consommation brute d'énergie mesurée en tonne équivalent pétrole dans l'industrie est une statistique diffusée chaque année et synthétise les différentes consommations d'énergie (électricité, gaz, pétrole, bois...). On évalue le biais associé à cet agrégat causé par le regroupement de réponses.

Nous avons réalisé 400 000 simulations de regroupement. Le nombre de simulations par regroupé est toujours supérieur à 50. Les résultats obtenus sont résumés dans le graphique ci-dessous.

On note premièrement un biais positif créé par le regroupement de réponse. La valeur théorique est celle du fichier de mise à disposition est de 35 792 088 tandis que la valeur obtenue si toutes les unités légales répondaient de façon regroupée serait de 35 823 298 tep. L'erreur absolue est donc d'environ 31 000 tep mais est plutôt faible vue les unités en jeu de l'ordre de la dizaine de millions. L'erreur relative est seulement de 0,08 % dans ce cas extrêmes où il y a 2 111 établissements regroupés.

Le biais estimé par la simulation a lui-même une certaine variabilité qui vient de la sélection aléatoire des regroupements.

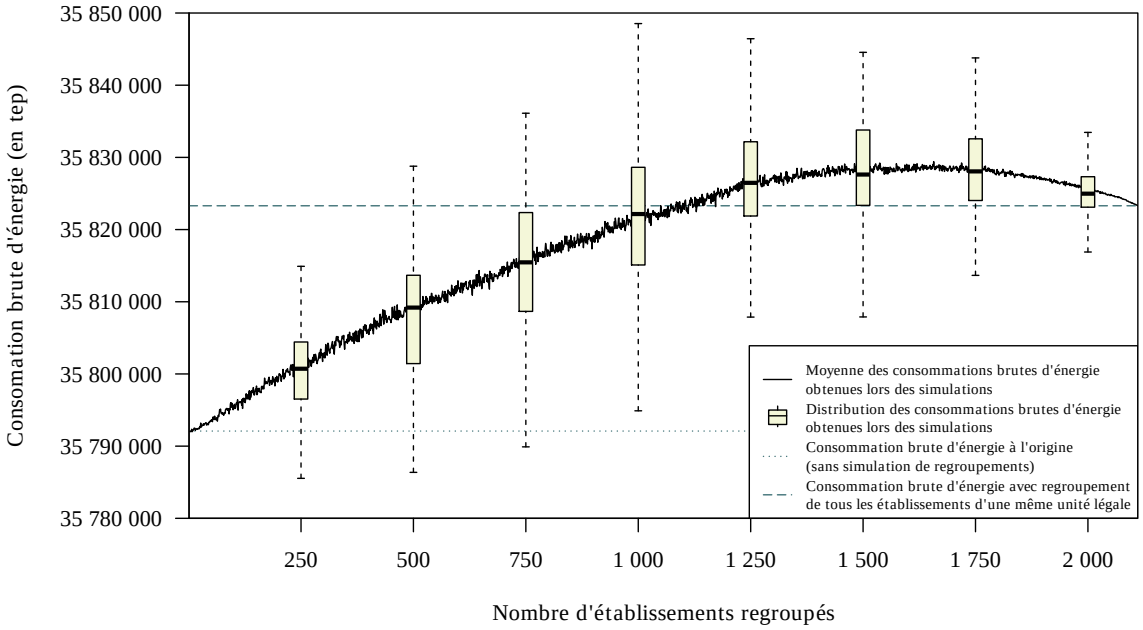

FIGURE 1 – Évolution de la consommation brute selon le nombre d'établissement regroupé

4.2 Impact sur des petits domaines de diffusion

Des chiffres sont aussi diffusés sur la consommation brute d'énergie sur des sous-domaines (par secteur d'activité, taille d'entreprises, localisation géographique). Ceux-ci ne conservent pas l'agrégat total lors des regroupements. En effet, si les établissements n'appartiennent pas à la même région, le prorata selon les effectifs entraînera une erreur sur la consommation brute d'énergie par région.

Nous avons réalisé 1 300 000 simulations de regroupement et nous avons calculé l'estimation que l'on aurait eue de la consommation brute d'énergie dans l'industrie en Île-de-France.

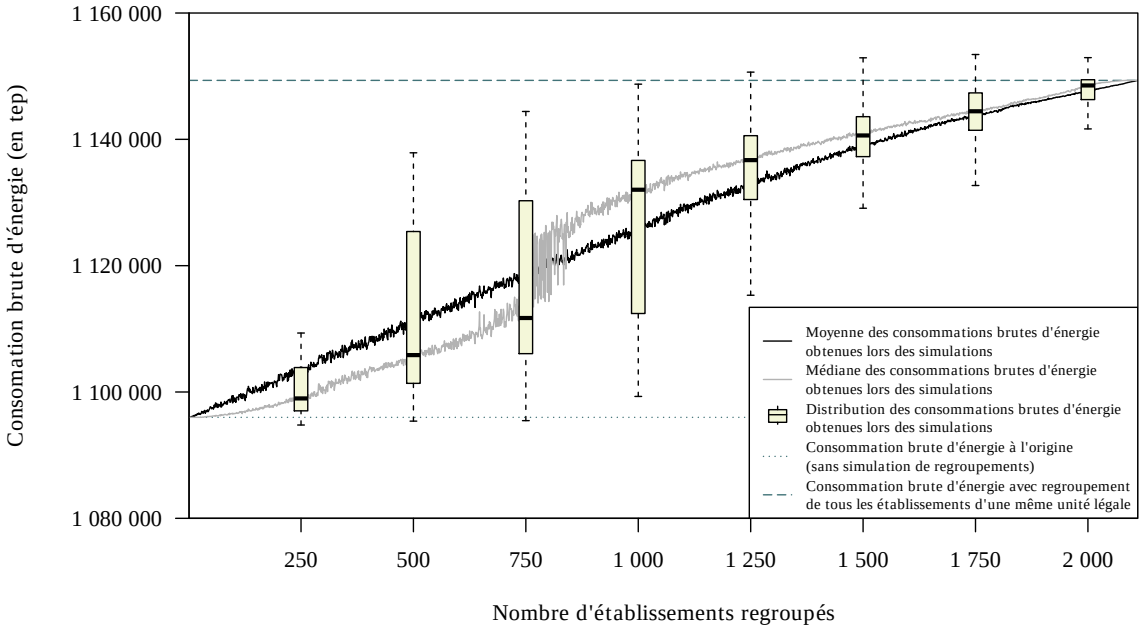

FIGURE 2 – Évolution de la consommation brute en Île-de-France selon le nombre d'établissement regroupé

On constate là encore un biais positif. L'erreur est de 25 000 tep pour 800 regroupés soit une erreur relative de 2,3 %. Le regroupement de réponses pourrait donc avoir des conséquences non négligeables sur des données locales. Ceci doit être nuancé : on admet généralement que les estimateurs aient une variance plus élevée à un niveau régional.

5 Expression du biais induit par le regroupement

Dans cette partie, nous nous attacherons à expliciter l'expression du biais induit par le regroupement si on applique la méthode de proratisation aux les unités de l'échantillon. Nous en déduirons des indicateurs qui permettent de repérer les regroupements susceptibles d'induire la plus grande erreur sur le total.

5.1 Expression générale du biais

Pour qu'il n'y ait pas d'erreur de mesure au niveau agrégé, nous allons montrer qu'il suffirait qu'une seule des deux conditions suivantes soit respectée :

— que toutes les unités d'un regroupement soient dans l'échantillon ;

— que les unités d'un regroupement aient le même poids de sondage.

Tout d'abord, fixons quelques notations :
— les indices i et g serviront respectivement pour les unités et les regroupements
— n_g : le nombre d'unités dans le regroupement g
— y_i : la variable d'intérêt pour l'unité i
— x_i : la variable auxiliaire utilisée pour le dégroupement (en pratique, l'effectif salarié)
— Y_g et X_g : les totaux des variables d'intérêt et auxiliaire sur le regroupement
— y_i^* : l'estimation de la variable issue de la procédure de dégroupement
— w_i : le poids de sondage de l'unité i
— s : l'échantillon sélectionné

Enfin, d'un point de vue de la "modélisation", on considérera que le regroupement n'a rien d'aléatoire : il y a des unités qui répondront individuellement, d'autres unités qui répondront ensemble. Si au moins une unité "d'un regroupement potentiel" est interrogée, on aura la réponse pour le regroupement.

L'estimateur de Horwitz-Thomson (celui que l'on aurait en l'absence de regroupement de réponse) est : $\hat{Y} = \sum_{i \in s} w_i y_i$. L'estimateur que l'on utilise en présence de regroupement est : $\hat{Y^*} = \sum_{i \in s} w_i y_i^*$ où $y_i^* = y_i$ si $i \notin G$ et $y_i^* = \frac{x_i}{X_{g(i)}} Y_{g(i)}$ si $i \in G$.

L'erreur que la procédure induit est :$\hat{Y} - \hat{Y^*}$. Cette erreur est la somme des erreurs E_g sur chaque regroupement : $E_g = \sum_{i \in s \cap g} w_i (y_i - \frac{x_i}{X_g} Y_g)$.

On voit donc qu'il y a deux conditions suffisantes pour que cette erreur soit nulle.

Première condition : que la procédure de dégroupement soit "exacte" (c'est-à-dire que $y_i = y_i^* = \frac{x_i}{X_g} Y_g$ pour toutes les unités.)

Deuxième condition : que toutes les unités du regroupement soient dans l'échantillon et aient le même poids w_g. En effet, on a alors : $E_g = w_g \sum_{i \in g} \Big(y_i - \frac{x_i}{X_g} Y_g \Big)$.

Or, par construction, $\sum_{i \in g} x_i = X_g$ et $\sum_{i \in g} y_i = Y_g$ d'où $E_g = 0$.

La première condition sera impossible à réaliser mais les erreurs doivent, dans une certaine mesure, se compenser s'il y a une relation linéaire "sans biais" - même si elle n'est pas exacte. Il est important pour que cela puisse fonctionner que l'ordre de grandeur des erreurs soit le même.

La seconde condition ne se réalise pas spontanément en pratique. Il y a toutefoi sun cas particulier qui peut se produire : si les deux unités regroupées appartiennent à la même strate de tirage, les poids sont identiques et donc le regroupement n'induit pas d'erreur..

Ceci nous conduira à examiner l'opportunité de méthodes alternatives pour traiter les regroupements en réduisant le biais théorique :

— inclure dans l'échantillon toutes les unités regroupées ;
— modifier le poids de sondage lors de la procédure d'estimation

.

La meilleure façon de le faire serait d'utiliser la méthode généralisée de partage des poids. Cette possibilité est développée page 14.

5.1.1 Cas des regroupements de deux unités

Dans un cas général

Explicitons ce qui se passe si le regroupement ne concerne que deux unités. Que nous noterons 1 et 2.

Si on note α_i le poids de l'unité i dans la variable auxiliaire x au sein du regroupement et β_i le poids de cette unité dans le total de la variable y, on obtient l'expression suivante :

$$E_{1+2} = (y_1 + y_2)(w_2 - w_1)(\beta_1 - \alpha_1) \tag{1}$$

Si on permute les rôles joués par les unités 1 et 2, cette relation reste inchangée. On peut donc supposer là encore que $w_2 \geq w_1$.

Le premier terme $(y_1 + y_2)(w_2 - w_1)$ est connu au moment des redressements et même de la collecte puisqu'il ne fait appel qu'aux poids de sondage et à $y_1 + y_2$ qui est l'information fournie par le regroupant.

En revanche, le second terme $\beta_1 - \alpha_1$ dépend d'une grandeur inobservée : y_1.

Il s'agit d'un terme sans dimension qui varie entre - 1 et + 1. Il reflète la proportionnalité effective entre la variable d'intérêt y et la variable auxiliaire x.

Si on note M le majorant de ce coefficient en valeur absolue, l'erreur maximale induite par le regroupement est donc majorée par

$$|E_{1,2}| \leq (y_1 + y_2)(w_2 - w_1)M \tag{2}$$

On peut bien sûr prendre $M = 1$ mais, si on dispose d'information sur le phénomène étudié, un majorant plus petit peut être utilisé.

Tous les éléments du terme $(y_1 + y_2)(w_2 - w_1)$ sont connus, on peut se servir de cette expression pour majorer le biais introduit par les regroupements. Il s'agit d'une majoration pessimiste dans la mesure où les erreurs ne sont pas de même signe.

On peut également s'en servir pour identifier les regroupement les plus "à risque". Si le terme $(y_1 + y_2)(w_2 - w_1)$ est élevé, il peut être utile de recontacter l'entreprise pour essayer d'avoir une meilleur information. S'il est faible, alors l'erreur induite par le regroupement au niveau agrégé sera faible. Il synthétise plusieurs facteurs qui jouent en sens contraire : pour les grandes unités, les poids w_i seront faibles mais $y_{1,2}$ sera grand.

Nous proposons donc d'utiliser un facteur de "risque" induit par le regroupement défini ainsi :

$$\text{Risque}_{1,2} = (y_1 + y_2)(w_2 - w_1) \tag{3}$$

Dans le cas où la variable d'intérêt est la consommation d'énergie

Nous avons testé ce que donnerait l'application de ces résultats en nous appuyant encore une fois sur l'EACEI 2015 pour laquelle nous disposons d'une information riche.

Le terme $|\beta_1 - \alpha_1|$ est majoré dans 95 % des cas par 0,36 et dans 99 % des cas par 0,6 (il s'agit d'une observation sur les données 2015).

Dans ce cas, on peut utiliser pour M la valeur 0,4 dans l'expression (2) sans prendre trop de risques.

Enfin, on peut voir ce que donnerait l'utilisation du facteur $(y_1 + y_2)(w_2 - w_1)$ pour repérer les regroupements à problème. On a profité de ce que de nombreux établissements avaient été interrogés et on s'est appuyé sur le cas des paires d'établissements appartenant à la même entreprise en calculant l'erreur qu'on aurait commise en proratisant leur réponse regroupée. En effet, on possède les vraies réponses et on peut se livrer à ce calcul.

La figure 3 montre comment l'erreur que l'on aurait commise varie en fonction du facteur $(y_1 + y_2)(w_2 - w_1)$.

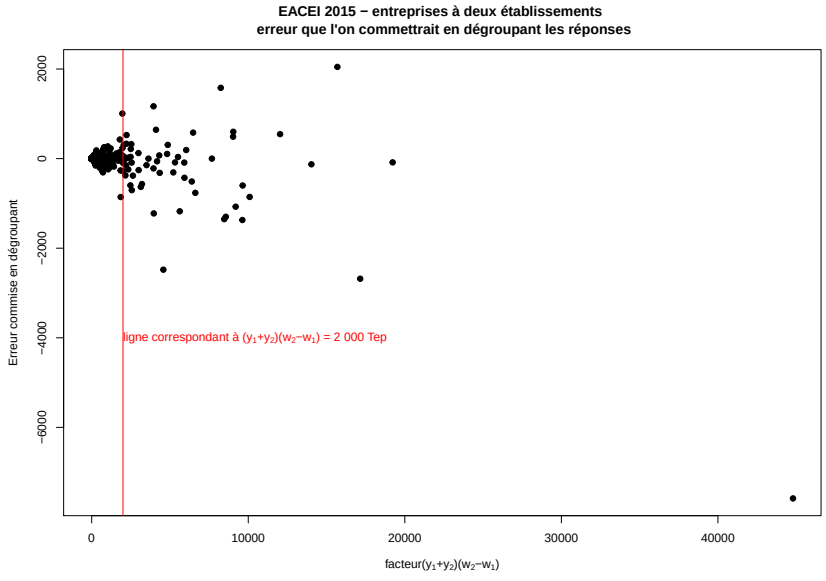

FIGURE 3 – EACEI 2015 - erreur potentielle en fonction de $(y_1 + y_2)(w_2 - w_1)$

On voit que les erreurs les plus grandes se produisent bien lorsque le facteur $(y_1 + y_2)(w_2 - w_1)$ est plus grand - même si dans certains cas ce facteur peut être grand et l'erreur faible.

Ce facteur est facile à calculer en cours de collecte. Si on demande de recontacter uniquement les entreprises regroupantes pour lesquelles ce facteur est grand, on peut éviter les plus grandes erreurs en ne contactant que peu d'entreprises.

Ainsi, les points situés à droite de la ligne rouge - ceux pour qui ce facteur dépasse 2 000 Tep - correspondent à un "saut" dans la distribution de ce facteur et ne rassemblent que 16 % des regroupements possibles. Dans la mesure où toutes les entreprises ne regrouperont pas leurs réponse,s il peut être raisonnable de surveiller ceux-ci uniquement. En tout cas, dès que le facteur $(y_1 + y_2)(w_2 - w_1)$ est important , cela vaut la peine de recontacter l'entreprise pour trouver une alternative à la procédure de proratisation.

5.1.2 Cas des regroupements de plus de deux unités

Dans le cas où le regroupement contient un nombre n_g d'unités plus grand que deux, la situation se complique.

On notera g le regroupement, Y_g la réponse groupée et on fixe dans ce regroupement une unité dont le numéro est 1. L'expression(1) se généralise de la façon suivante :

$$E_g = Y_g \sum_{i \in g, i \neq 1} (w_i - w_1)(\alpha_i - \beta_i) \tag{4}$$

Et donc, si M est un majorant des $|\alpha_i - \beta_i|$ la majoration (2) se généralise de la façon suivante : $E_g \leq Y_g(n_g - 1)(w_{\max} - w_{\min})M$

Où $w_{\max}$ et $w_{\min}$ sont respectivement les poids les plus élevés et les moins élevés sur regroupement. Ceci permet de définir un nouveau facteur de risque :

$$\text{Risque}_g = Y_g(n_g - 1)(w_{\max} - w_{\min})$$

Il synthétise trois informations : le niveau de l'agrégat représenté par le regroupement (Y_g), le nombre d'unités impliquées dans le regroupement $(n_g - 1)$ et la dispersion des poids dans le regroupement $(w_{\max} - w_{\min})$.

Il indique dans quelle mesure on s'éloigne de trois conditions idéales pour que le regroupement ne pose pas de problème : la quantité à répartir est nulle, il n'y a qu'une seule unité dans le regroupement et les poids sont identiques.

6 Utilisation de la méthode du partage des poids

Nous allons présenter dans cette article une piste très intéressante pour traiter les regroupements de réponse qui nous a été suggérée par un relecteur anonyme de l'article que nous remercions ici.

6.1 La méthode généralisée de partage des poids dite MGPP

Pour plus de détails sur cette méthode, le mieux est de se reporter à l'ouvrage de Pierre Lavallée cité en bibliographie. Nous rappelons ici les grandes lignes de la méthode.

Cette méthode permet de traiter les situations de "sondage indirect" où on sélectionne des unités dans une population U_A. Chaque unité sélectionnée fournit un ensembles d'unités d'une autre population U_B qui constitue l'échantillon *in fine*.

L'intérêt de la MGPP est qu'elle permet de traiter les cas où une unité de la population U_B peut-être liée à plusieurs unités de la population U_A. Dans ce cas, sa probabilité d'inclusion est plus grande que celle de chacune des unités de la population U_A auxquelles elle est liée.

La MGPP passe par le recueil d'une information auprès des unités de la population U_B : la liste des liens. La méthode consiste à donner à toutes les unités d'une grappe de U_B le même poids qui est déterminé par celui des unités de U_A liées à celles de cette grappe et par le nombre de liens.

On trouvera dans le livre de Pierre Lavallée des explications sur la mise en œuvre de cette méthode et ses propriétés.

6.2 Application au problème des regroupements de réponses

À première vue, ce problème ne relève pas du sondage indirect car on ne collecte pas les réponses des unités "indirectement échantillonnées" mais uniquement un total.

Toutefois, nous avons vu dans ce qui précède qu'il suffirait de réunir deux conditions pour que la procédure de dégroupement soit sans biais :

— que toutes les unités du regroupement soient dans l'échantillon ;

— que toutes les unités du regroupement aient le même poids.

La MGGP permet en fait de réaliser ces deux conditions : on peut rajouter à l'échantillon les unités regroupées, leur donner un poids qui sera le même sur l'ensemble des unités du regroupement.

Seulement, nous sommes dans un cas particulier où les populations U_A et U_B sont identiques. Ce qui ne pose aucun problème d'un point de vue calculatoire.

Pour pouvoir appliquer la MGPP, il nous suffit supposer que toutes les unités regroupées appartiennent au champ de l'enquête et de faire l'hypothèse de comportement suivante :

Hypothèse : Les regroupements existent "avant l'enquête" - même s'il ne sont révélés que par l'enquête - et, lorsqu'une unité du regroupement est sélectionnée, on reçoit une réponse agrégée quelque soit l'unité sélectionnée.

C'est une hypothèse qui peut paraître assez réaliste : si une information - mettons une facture d'électricité - n'est disponible que pour un ensemble d'unités, quelle que soit celle qu'on interroge, elle sera amenée à répondre pour l'ensemble.

Dans ce cas, on ajoutera à l'échantillon toutes les unités du regroupement et on affectera à toutes les unités - qu'elles figurent dans l'échantillon ou non - le poids résultant de l'application de la MGPP.

Dans ce cas particulier d'application de la MGPP, le nombre de liens du regroupement sera tout simplement le nombre d'unités qui le composent. Et les unités de la population U_A seront les unités du regroupements.

On prendrait donc comme poids w_g^* pour le regroupement g : $w_g^* = \frac{\sum_{i \in \cap g} w_i}{n_g}$

Ensuite, on pourrait répartir les agrégats au prorata des variables auxiliaires en conservant le total. C'est en outre une propriété de la MGPP que la somme des poids sera conservée.

6.3 Limites de l'application de la MGPP aux regroupements

Il semble que l'application de la MGPP puisse traiter de façon satisfaisante le problème. Toutefois, sa mise en œuvre poserait quelques difficultés :

- Cela ne résoudrait pas le problème d'unités regroupées "hors du champ de l'enquête"
  Certes mais cela ne l'aggraverait pas non plus. Il faudrait faire un premier traitement - avec les erreurs qu'il implique - pour estimer les agrégats en enlevant les unités "hors champ".

- L'hypothèse de comportement n'est pas forcément réaliste
  En effet, certaines unités regroupées répondent par elles-mêmes. Ce qui veut bien dire qu'il est faux que sélectionner une unité du regroupement entraîne la réponse de toutes. Ceci veut aussi dire que des entreprises fournissent des réponses groupées dans des cas où ce n'est pas nécessaire. Toutefois, cet inconvénient doit être nuancé : non seulement ces cas ne sont pas fréquents - même s'ils mobilisent beaucoup d'énergie - mais, à bien y réfléchir, ils doivent correspondre à des situations où le regroupement n'est pas nécessaire. Une restriction de cette possibilité aux situations où on ne peut pas faire autrement supprimerait ce problème.

- La modification des poids peut être intenable
  Dans les enquêtes auprès des entreprises, il y a de grandes unités qui sont systématiquement échantillonnées. Elles ont un poids de sondage unitaire et on veille parfois à ce qu'il ne soit pas transformé par les opérations de repondération. L'application de la MGPP peut conduire à augmenter le poids d'une grande unité en faisant la moyenne avec le poids de petites unités. En général, cela conduit à des systèmes de poids impossible à exploiter. Cet inconvénient serait réel mais, en fait,

cette situation correspond aussi à celle des regroupements à risque que l'on a identifié page 12. Limiter les regroupements à risque permettrait de limiter les cas où l'application de la MGPP est problématique.

6.4 Conclusion sur l'utilisation de la méthode généralisée de partage des poids

L'application de cette méthode semble prometteuse. Elle n'exclut pas un traitement rigoureux des regroupements en limitant le nombre de regroupements injustifiés, en surveillant les regroupement à risque et en traitant attentivement le cas des regroupements hors champ.

Nous n'avons pas encore expérimenté l'application de la MGPP aux regroupements en parallèle d'un traitement habituel en se contentant de proratiser les réponses mais c'est une piste que nous envisageons.

Bibliographie

Lavallée, P. (2002), *Le sondage indirect ou la méthode généralisée du partage des poids*, Édition de l'Université de Bruxelles, Bruxelles, Belgique