ESTIMATEURS EN COURS D'ENQUÊTE

ET PRIORITÉS DE RELANCES

Benoît BUISSON (*)

(*) Insee, Pôle ingénierie statistique entreprises

Introduction

Cette contribution vise simplement à décrire la pratique du pôle ingénierie statistique entreprises en terme de suivi de la collecte des enquêtes thématiques entreprises, notamment de l'enquête TIC. Nous reprenons le terme enquête « thématique » pour désigner des enquêtes spécifiques sur des thèmes précis, hors le dispositif récurrent d'informations sur les statistiques structurelles et conjoncturelles. De plus en plus toutefois les enquêtes « thématiques » peuvent avoir un caractère régulier : par exemple l'enquête TIC (tous les ans) ou encore l'enquête innovation (tous les deux ans). Au départ ce type d'enquête était apériodique. Cette contribution a pour objectif de montrer l'intérêt de pouvoir disposer d'estimateurs de variables cibles en cours d'enquête, ce qui revient à dire qu'il **faut assurer la correction de la non-réponse totale et partielle en cours d'enquête**, et non plus en fin d'enquête comme cela peut être fait dans certains cas. Cette contribution détaille également la mise en application **d'une méthode de priorités de relance pour non-réponse totale**, méthode définie et appliquée par l'Australian Bureau of Statistics.

1. Estimateurs et calcul de précision en cours d'enquête

1.1. Pourquoi calculer des estimateurs en cours d'enquête ?

Lors du déroulement d'une enquête entreprises, par voie postale éventuellement complétée par une collecte par internet, il est très important **d'obtenir le maximum d'informations** de la part des entreprises à ce moment précis du processus de production. Traditionnellement dès que la phase de gestion-collecte est terminée, c'est à dire lorsque les gestionnaires sont passés à une autre opération, il est très rare de retourner vers les entreprises notamment dans les phases de mise en cohérence finale des réponses, redressement et calage. Durant la collecte, il est donc primordial d'obtenir le maximum d'informations tout en ciblant les contacts pour **concentrer les moyens** là où ils sont les plus utiles.

De manière générale pour cibler les contacts, il va être primordial de **calculer des estimateurs de certaines variables cibles** assez tôt dans le processus de collecte pour notamment juger de

l'influence des données individuelles sur ces estimateurs agrégés. Un des mérites de cette approche réside également dans **la définition précoce des variables cibles**, idéalement avant l'échantillonnage mais tout au moins au début de collecte, par la maitrise d'ouvrage de l'enquête. Pour calculer ces estimateurs en cours de collecte, il faut essayer « à chaque instant » de faire comme si nous avions à diffuser des estimateurs « définitifs », et donc de faire comme si en fait la collecte s'arrêtait là. Cela imposera notamment de **corriger la non-réponse partielle**, du moins sur les variables jugées cibles, de **corriger la non-réponse totale** et éventuellement de caler l'enquête. Bien sûr plus nous sommes avancés dans le processus de collecte, plus ces estimations deviennent fiables.

L'intérêt de calculer des estimateurs en cours d'enquête se révèle ainsi multiple, cela permet notamment de :

- Cibler les contrôles sur les entreprises répondantes qui influencent le plus l'estimation de l'agrégat de la ou des variables cibles ;
- Repérer les données hors-norme ou « aberrantes » ;
- Mieux cerner l'influence des données redressées sur l'estimation de l'agrégat ;
- Calculer des indicateurs de précision associés aux estimateurs, ce qui peut permettre d'avoir une démarche du type « critère d'arrêt de collecte ».

Ainsi pour l'enquête TIC 2007-2008, la priorité de la maitrise d'ouvrage (l'ex Département des Activités Tertiaires) était clairement de fiabiliser les estimations du commerce électronique, notamment les ventes par internet de la part des entreprises. Les estimations, en cours de collecte, du montant des ventes et achats par internet a permis notamment de cibler les contrôles, les relances pour non-réponse totale tout en mettant à disposition chaque quinzaine **un tableau de bord** du commerce électronique (estimateurs des variables dites cibles, indicateurs de précision).

1.2. Comment calculer des estimateurs en cours d'enquête ?

Ces estimateurs seront calculés à partir de données, par définition, partiellement validées. Traditionnellement les contrôles de saisie ou automatiques ont déjà été réalisés, par contre les contrôles manuels qui peuvent nécessiter des rappels sont en cours. Il faut donc très vite réaliser, sur les données répondantes, **l'arbre de calcul** pour conduire aux estimateurs. Bien que d'intérêt « méthodologique » qui peut être jugé mineur, cette phase est importante et peut s'avérer complexe car il faut anticiper tous les cas. Ainsi tout simplement une entreprise qui déclare faire des ventes par internet sans en préciser le montant doit être redressée pour ce même montant contrairement à celle qui ne déclare pas faire de vente. Autre exemple, pour traduire une réponse en pourcentage en valeur nous avons besoin de grandeurs de référence (chiffre d'affaires ou achat) qui selon les cas peuvent provenir du questionnaire, des données de lancement, de sources externes ou de données redressées. De ce fait, en calculant ces estimateurs, nous anticipons sur les phases ultérieures d'apurement-mise en cohérence en détectant éventuellement des cas complexes.

Il est impératif de **corriger la non-réponse** dès ce stade car cette correction va profondément influencer la valeur des estimateurs. Comment corriger **la non-réponse partielle** précocement dans le processus d'enquête d'une manière la plus fiable possible ? Pour les enquêtes répétitives dans le temps, comme pour l'enquête TIC, **il est tout simplement possible d'utiliser le modèle de correction de l'année passée** (déterminer sur l'échantillon précédent complet). Ainsi pour l'enquête TIC 2006-2007, le pourcentage du montant des ventes par internet avait été corrigé par imputation aléatoire, en fonction du secteur et de la taille de l'entreprise, après passage en classe de cette même variable. Le processus était le suivant : constitution de 4-5 classes après analyse des réponses[1], recherche de variables explicatives sur les répondants (secteur et taille dans le cas présent), imputation aléatoire, passage d'une variable en classe à une variable quantitative (médiane de la classe observée sur les répondants). Nous reconduisons tout simplement ce processus sur les répondants 2007-2008. Au début de la collecte, peu de répondants, cette démarche peut être jugée fragile d'autant plus que dans le cas présent cela introduit beaucoup de « points d'accumulation » (qui s'observent toutefois chez la sous-population des répondants à cette question). Pour fiabiliser la démarche, nous pourrions introduire également le « sur-échantillon » des répondants de l'an passé, cela présentera toutefois l'inconvénient de s'éloigner de l'optique « fin de collecte ». **La situation est plus complexe pour les enquêtes non-répétitives** ou pour la première génération d'une enquête répétitive. Pour l'enquête déchets par exemple, une autre approche a été proposée. La variable d'intérêt était le tonnage de déchets non dangereux émis par les établissements commerciaux. A partir de l'analyse des répondants, il a été mis en évidence des relations linéaires (par groupe d'activités économiques) entre ce même tonnage et la taille de l'établissement, la taille étant une variable de lancement incluse dans la base de sondage. Ces relations linéaires, qui se sont bien évidemment faites plus précises au fur et à mesure de la collecte, ont été appliquées aux établissements non-répondants.

Il est également impératif de **corriger la non-réponse totale**. Traditionnellement pour les enquêtes thématiques entreprises, la non-réponse totale est corrigée par **repondération** après mise en évidence de **groupes de réponses homogènes (GRH)**. Pour calculer des estimateurs du commerce électronique dans l'enquête TIC 2007-2008, nous avons procédé de la sorte … en utilisant les groupes de réponses homogènes mis en évidence l'année précédente. Ces GRH faisait notamment intervenir le comportement de réponse à l'enquête précédente, le secteur de l'entreprise, sa taille et sa localisation géographique. Bien que non complètement identiques une année sur l'autre, les GRH mis en évidence sont assez proches d'une génération d'enquête à l'autre. Ce processus apparaît plus fiable que de rechercher des GRH en cours d'enquête, comme cela peut s'imposer pour des enquêtes non-répétitives. Pour l'estimation de variables cibles en cours d'enquête, nous assimilons de fait non-retour à non-réponse. Ce n'est qu'après la fin de la collecte, sur les non-retours « définitifs » que nous effectuerons des recherches dans des sources externes (source TVA notamment) pour repérer des entreprises cessées ou hors-champ (moins de 10 salariés par exemple dans le cas de l'enquête TIC).

[1] Les entreprises répondent rarement réaliser 26,85 % de leur ventes par internet mais plutôt 25 %…

Cela peut être considéré comme une limite de cette approche, mais il est jugé trop coûteux de réaliser des vérifications sur sources externes pour un grand nombre d'entreprises.

Dans les enquêtes thématiques entreprises, nous adoptons la notion d'entreprises « **non-substituables** ». il s'agit d'entreprises très importantes par rapport au thème de l'enquête, que nous jugeons suffisamment hors-normes pour d'une part ne pas les utiliser (lorsqu'elles sont répondantes) parmi les entreprises qui servent à imputer les non-répondantes et pour d'autre part rechercher le maximum d'informations dans des sources externes pour remplir leur questionnaire (elles ne sont jamais repondérées). En cours d'enquête, nous bloquons bien leur poids à un en cas de réponse. Par contre, nous les redressons comme les autres en cas de non-réponses totale. Nous ne disposons pas suffisamment de temps là aussi pour faire des recherches externes en début de collecte sur l'ensemble des entreprises non substituables (200 entreprises environ pour un échantillon de 12500 entreprises dans l'enquête TIC). Il s'agit d'une deuxième différence relativement au traitement qui s'opérera en fin d'enquête.

1.3. Un exemple à partir de l'enquête TIC

Le tableau ci-dessous reproduit **le tableau de bord** transmis à la maitrise d'ouvrage et à la maitrise d'œuvre de l'enquête en cours de collecte pour suivre l'avancement de celle-ci :

montant en milliards d'euros

	07-déc	20-déc	04-janv	17-janv	01-fev	15-fev	29-fev	14-mars	28-mars	apuré 2	RDR
vente par internet	49,9	51,2	57,6	51,0	52,7	50,3	50,1	52,8	53,7	57,6	59,8
vente par EDI	189,3	275,2	258,3	244,4	234,0	247,7	235,9	238,0	239,9	256,8	243,8
vente électronique	239,2	326,4	315,9	295,5	286,8	298,0	286,0	290,9	293,6	314,4	303,6
achat par internet	45,1	46,0	65,1	39,2	38,6	40,6	43,5	43,4	48,0	51,9	57,6
achat par edi	101,3	121,9	111,0	185,4	134,0	141,2	145,7	145,9	149,6	152,3	155,1
achat électronique	146,5	167,9	176,1	224,7	172,6	181,8	189,2	189,3	197,7	204,2	212,7
CV en %											
vente par internet	21,6	20,1	16,4	15,7	13,9	12,2	10,0	9,3	9,1	8,1	
vente par EDI	10,5	25,4	21,4	14,8	11,1	11,8	4,4	4,3	4,2	4,0	
vente électronique	10,9	22,1	18,2	12,8	9,6	10,2	4,2	4,0	3,9	3,6	
achat par internet	24,9	21,3	34,1	11,0	8,5	5,9	9,9	9,8	9,7	6,1	
achat par edi	33,0	20,0	16,2	32,5	11,7	9,7	6,4	6,3	6,0	5,9	
achat électronique	29,3	18,2	17,1	27,2	10,2	7,8	5,9	5,9	5,6	4,8	
taux de réponse	36%	46%	57%	65%	76%	81%	83%	84%	84%	85%	

Pour information, le calcul de précision a été effectué avec **la macro SAS CALKER**, élaborée par l'ex division H2E et traditionnellement utilisée dans les enquêtes entreprises. Le calcul de la variance prend donc en compte la non-réponse (totale uniquement) en faisant toutefois l'hypothèse

simplificatrice que les GRH correspondent aux strates de tirage. De ce point de vue, le calcul de la variance via CALKER surestime probablement la variance des estimateurs.

L'utilisation d'un tel tableau de bord, outre le suivi de la collecte, permet de **repérer assez vite les valeurs hors-normes** qui doivent vite être corrigées : par exemple début janvier pour les achats par internet ou encore mi-janvier pour les achats par EDI. Ainsi l'estimateur des achats par EDI passe de 111 milliards au 4 janvier à 185 milliards au 17 janvier, le coefficient de variation passant de 16 à 32 sur la même période. En parallèle à ce tableau de bord, nous avons calculé des « **contributions** » de chaque entreprise « répondante » (sous-entendu au questionnaire) à chaque agrégat du commerce électronique. Ce calcul de contribution est très simpliste : il s'agit de rapporter la donnée de l'entreprise multipliée par son poids (suite à la correction de la non-réponse totale) sur l'agrégat concerné. Nous répondons donc à la question : qu'est ce qui se serait passé si l'entreprise en question avait répondu une valeur nulle pour la variable cible concernée. Cette approche n'est pas équivalente à savoir ce qui se serait passé si l'entreprise n'avait pas répondu au questionnaire, car dans ce cas elle aurait été corrigée par repondération. Une possibilité d'amélioration serait de calculer systématiquement ce type d'indicateur. Le calcul des contributions avait deux objectifs. D'une part de repérer et d'alerter la maitrise d'œuvre sur les déclarations brutes des entreprises qui avaient une forte influence, pour avoir notamment une confirmation de la déclaration. Ce calcul de contribution était auparavant opéré en fin de course (après le redressement-calage), lorsque qu'il était très délicat de joindre de nouveau les entreprises par téléphone. D'autre part de repérer l'influence de la correction de la non-réponse partielle sur l'agrégat en question. Si une donnée corrigée par imputation aléatoire a une grande contribution, l'idée est de rappeler l'entreprise pour obtenir la réponse à la question donnée. Il s'agit d'un signal d'alerte également pour nous pour fiabiliser au maximum l'opération de redressement pour ce type d'entreprise, si en fin de course sa réponse venait toujours à manquer.

L'intérêt de calculer des estimateurs en cours d'enquête apparaît donc multiple et permet notamment de cibler les rappels et les relances pour non-réponse partielle. La partie 2.23 va détailler une méthode de priorité de relances pour non réponse totale, qui se base indirectement sur les calculs d'estimateurs de variables cibles. Dans un premier temps, nous rappellerons dans une partie 2.1 quelques approches déjà utilisées pour définir des priorités de relance pour non-réponse totale.

2. Définition de priorités de relance pour non-réponse totale

La définition de priorités de relance est devenue **un enjeu important** dans la gestion des enquêtes notamment entreprises. Un calendrier de gestion de plus en plus tendu, associé à des moyens humains contraints pour effectuer les contrôles et relances, militent pour uniquement contacter les entités jugées « prioritaires ». Nous parlerons dans la suite de priorités de relance téléphoniques. Car toutes les entreprises non-répondantes sont via des procédures automatiques par courrier, procédures plus ou moins « denses » selon le caractère obligatoire ou non de l'enquête.

2.1. Différentes approches pour définir des priorités de relance

L'approche la plus rudimentaire consiste à **relancer toutes les entreprises par téléphone ou aucune** ! Bien sûr cette approche n'est guère employée, par contre elle peut être ancrée dans le comportement des gestionnaires. Il faut en effet expliquer pourquoi il est nécessaire de cibler les relances téléphoniques sur certaines entreprises bien qu'il soit nécessaire d'interroger toutes les entreprises de l'échantillon. La pratique de certaines enquêtes montre qu'il faut faire preuve de pédagogie sur le sujet.

Une approche plus courante consiste à **définir un ordre de priorités de relance en fonction de variables auxiliaires**, généralement des variables issues de la base de sondage. Il peut être décidé ou non de pondérer ces valeurs de variables auxiliaires en fonction du poids de lancement des entreprises. Ainsi pour les enquêtes entreprises, nous pouvons utiliser le chiffre d'affaires ou les effectifs multipliés par le poids de sondage. L'ordre de priorité établi est donc statique, il est défini une fois pour toutes en début d'enquête et est uniquement mis à jour avec la liste des entreprises répondantes. L'idée est bien sûr que la variable auxiliaire choisie soit en relation très étroite avec la variable d'intérêt du questionnaire, ce qui n'est pas toujours le cas, par exemple pour le commerce électronique. En effet des structures moyennes (filiales d'un groupe par exemple) peuvent effectuer 100% de leur vente ou de leur achat par voie électronique, contrairement à des très grandes entreprises. Quelle que soit la nature du lien entre variable auxiliaire et variable d'intérêt, **cette approche ne tient ni compte de l'avancée de la collecte ni de la correction de la non-réponse totale**. Par exemple une entreprise appartenant à un GRH où le taux de réponse est faible peut se révéler plus importante à relancer qu'une entreprise plus « grande » au sens de la variable auxiliaire pondéré. Cette approche est perfectible notamment du fait qu'elle ignore l'avancée et les résultats issus de la collecte.

Une approche nettement plus intéressante de ce point de vue est celle de « **la perte de variance anticipée** », développée par Philippe BRION. Nous résumons et reprenons ci-dessous une note de Philippe BRION (cf. bibliographie) parue en 2003 sur le sujet. L'idée est de cibler les relances sur les strates pour lesquelles on s'attend à une **forte réduction de la variance** entre la situation actuelle en cours de collecte et la situation idéale sans non-réponse.
Pour un sondage aléatoire simple stratifié (ce qui est le cas de beaucoup d'enquêtes auprès des entreprises), la variance de l'estimation d'un total d'une variable Y vaut :

$$V\left(\hat{T}(\mathrm{Y})\right)=\sum_{h=1}^{k} N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{1}{n_h}S_h^2$$

où h est l'indice de strate ; N_h est le nombre d'entreprises de la strate h ; n_h est le nombre d'entreprises enquêtées dans la strate h ; S_h^2 est la dispersion de la variable Y au sein de la strate h

Si dans la strate h on n'a récupéré que r_h questionnaires au lieu des n_h attendus on peut procéder à de la répondération : on montre que si l'on repondère de manière uniforme à l'intérieur de chaque strate de tirage et conditionnellement à l'hypothèse que les non répondants n'ont rien de spécifique à l'intérieur de chaque strate, la variance de l'estimation du total, pour sa partie relative à la strate h, devient $V_h\left(\hat{T}(Y)\right) = \mathrm{N}_h^2\left(1-\frac{r_h}{N_h}\right)\frac{1}{r_h}S_h^2$

Si $x_h = \frac{r_h}{n_h}$ est le taux de questionnaires utilisés dans la strate h, on a donc une perte de variance de : $\Delta V_h = \frac{N_h^2 S_h^2}{n_h}\left(\frac{1}{x_h}-1\right)$ dans le cas où on procède par repondération. On peut ainsi calculer, pour chaque variable d'intérêt, la quantité : $\sum_h \Delta V_h / V\left(\hat{T}(Y)\right)$, où $\mathrm{V}\left(\hat{T}(Y)\right)$ est la variance attendue avec, dans chaque strate, les n_h questionnaires enquêtés ; **la quantité calculée est la « perte de variance anticipée ».** Au fur et à mesure que l'enquête se déroule, elle diminue (pour arriver à zéro si on pouvait récupérer et traiter l'ensemble des questionnaires). Afin d'établir des priorités de relance, on peut regarder la part de la perte de variance anticipée due à chaque strate par l'intermédiaire des différentes grandeurs ΔV_h : ceci donne des indications sur les **priorités** à donner dans le traitement des différentes states (il faut se consacrer d'abord aux strates pour lesquelles ΔV_h est la plus importante). Nous pouvons donc obtenir **un ordre de priorité par strate**, selon les strates de tirage. La méthode peut être affinée en utilisant des groupes de réponses homogènes différents des strates de tirage, ce qui complexifie la formalisation. Il reste à choisir, au sein d'une strate donnée, les entreprises à relancer. En regard uniquement de la méthode, ces entreprises peuvent être choisies aléatoirement par exemple (ce qui peut être troublant pour les gestionnaires, surtout si on pratique un nouveau tirage aléatoire à chaque période de relance). Il faut également interclasser les strates, faut-il par exemple mieux relancer 3 entreprises de la strate qui arrive en première position selon ce critère et deux de la strate en deuxième position ou 5 entreprises de la strate 1 ? Pour cela, nous pouvons faire l'hypothèse simplificatrice que la probabilité de réponse suite à relance téléphonique est de 1 et donc de réitérer les calculs. Cette approche est bien **dynamique.** Elle complète et va plus loin que l'approche, souvent appliquée, qui consiste à **relancer en priorité les strates avec un taux de réponse faible**, notamment du fait qu'elle prend en compte la dispersion de la variable d'intérêt au sein de la strate. Il est nécessaire d'estimer cette dispersion, ce qui peut s'avérer délicat surtout en début de collecte. Il peut donc être utile, dans le cas d'une enquête répétée, d'utiliser les S_h^2 calculés l'année précédente (s'ils ont été calculés !) au moins au début du traitement de l'enquête. Pour une enquête ponctuelle, S_h^2 va être estimé sur les questionnaires déjà reçus et traités. Cependant, cette estimation peut être relativement instable. Philippe BRION signale que cette approche se révèle surtout pertinente pour les petites et moyennes entreprises, le traitement des très grandes unités

manquantes relevant d'une autre approche : visite d'un enquêteur, utilisation de sources externes (nous pouvons ici retrouver le concept d'entreprises non substituables).

2.2. La méthode utilisée pour les enquêtes TIC et déchets

Nous avons vu que la méthode « perte de variance anticipée » se focalise sur la variance de l'estimateur final. La méthode proposée ici se base sur l'estimateur lui-même et non plus sa variance, elle reprend en l'appliquant dans le cadre de la repondération et en utilisant des hypothèses simplificatrices l'approche développée par Richard MAC KENZIE dans son article cité en bibliographie.

Soit Y la variable d'intérêt de type quantitatif. En absence de non-réponse, nous obtenons Y_1 comme estimateur de Y. Dans ce cas, nous supposons que toutes les entreprises échantillonnées ont répondu et que leur réponse est jugée « valide ». En présence de non-réponse, avant les opérations de relances téléphoniques ciblées, nous obtenons Y_2 comme estimateur de Y. Cet estimateur tient compte de la correction de la non-réponse, que nous supposerons faite par repondération, les strates utilisées pour la correction de la non-réponse ne sont pas définies a priori. Dans le cas d'une enquête répétitive, cela peut-être les groupes de réponses homogènes mis en évidence pour la vague précédente. Pour une enquête non-répétitive, cela peut être les strates de tirage, en analysant toutefois en cours d'enquête le comportement de réponse des entreprises pour actualiser ces strates. En présence de non-réponse et après une relance téléphonique ciblée sur une entreprise (relance que nous supposons aboutir avec une probabilité de 1), nous obtenons Y_2^R comme estimateur de Y. Nous supposons également que les unités qui n'ont pas été relancées n'ont pas répondu.

Nous pouvons considérer que nous choisissons de **relancer l'entreprise qui minimise la distance** $F_R = \left(Y_2^R - Y_1\right)^2$ **.** Posons : $F_2 = (Y_2 - Y_1)^2$, nous faisons de plus l'hypothèse que la valeur absolue de l'écart entre Y_2 et Y_1 est plus grand que la valeur absolue de l'écart entre Y_2 et Y_2^R. Cette hypothèse semble peu contraignante, elle pose éventuellement problème en tout début de collecte et en toute fin de collecte. Toutefois ce n'est pas lors de ces phases que la notion de « priorités de relance » est cruciale.

Du fait de cette hypothèse, minimiser la distance F_R revient à maximiser la grandeur $(Y_2 - Y_2^R)^2$ sous la contrainte que $F_2 - F_R \succ 0$. Il faut donc s'éloigner le plus possible de Y_2 tout en se rapprochant de Y_1 (une analyse graphique est immédiate). **La contrainte pose problème** car elle fait intervenir Y_1, grandeur inconnue et qui peut poser des problèmes d'estimation. Comment faire dès lors pour vérifier la condition qui porte sur la contrainte ? Dans le cas d'une enquête répétitive, nous pouvons utiliser l'estimation de Y de l'enquête précédente. Cela peut servir de garde-fou à condition que cette estimation soit fiable. Or, dans certains cas, les priorités de relance sont justement définies dans le but

d'améliorer une estimation que nous supposons fragile. Autre possibilité : choisir une variable auxiliaire, que nous savons liée à la variable d'intérêt et pour laquelle nous disposons d'une estimation fiable ou mieux de sa « vraie » valeur sur la population. Il est important de disposer d'un garde-fou en la matière pour une raison simple : nous allons choisir de relancer l'entreprise qui fait le plus varier l'estimateur. Il faut donc s'assurer que cette variation s'effectue dans le « bon » sens. Il est probablement rare, une fois que le taux de réponse atteint un niveau minimal acceptable, que la qualité de l'estimation se dégrade. Toutefois il vaut mieux rester prudent, car ici nous cherchons précisément à relancer les unités qui font varier l'estimateur. Une autre possibilité de « garde-fou » peut consister en l'analyse de l'évolution de Y_2 au cours du temps, au fur et à mesure des relances et des arrivées de questionnaires.

Nous souhaitons donc maximiser la valeur absolue de $\left(Y_2 - Y_2^R\right)$ sous contrainte**.** Nous avons :

$$Y_2 = \sum_{rep} w_i^* Y_i = \sum_{rep} \sum_{strate} w_i^* Y_i \text{ et } Y_2^R = \sum_{rep+1} \sum_{strate} w_i^{**} Y_i$$

Ici les strates font référence aux strates de correction de la non-réponse - ou groupes de réponse homogène. Pour calculer la différence $\left(Y_2 - Y_2^R\right)$, il faut bien voir que cette différence est nulle sur toutes les strates qui ne contiennent pas l'unité à relancer. En effet dans ces strates, les répondants sont les mêmes avant et après relance tout comme les poids. La différence est donc nulle.

Soit la strate s qui contient l'unité à relancer, nous avons :

$$\left(Y_2 - Y_2^R\right) = \sum_{rep}^{s} w_i^* Y_i - \sum_{rep+1}^{s} w_i^{**} Y_i = -w_{rel}^{**} Y^{rel} + \sum_{rep}^{s} Y_i \left(w_i^* - w_i^{**}\right) \text{ ; avec}$$

$$w_i^* = w_i * \frac{n_s}{r_s}$$

$$w_i^{**} = w_i * \frac{n_s}{r_s + 1}$$

r_s désigne le nombre de répondants dans la strate (GRH) de l'unité à relancer ; n_s le nombre d'unités échantillonnées dans cette même strate ; w_{rel} le poids initial de l'unité à relancer.

On obtient ainsi :

$$\left(Y_2 - Y_2^R\right) = -w_{rel} \frac{n_s}{r_s + 1} Y^{rel} + \frac{n_s}{r_s (r_s + 1)} \sum_{rep}^{s} w_i Y_i = \frac{n_s}{r_s + 1} \left[\left(\frac{1}{r_s} \sum_{rep}^{s} w_i Y_i \right) - w_{rel} Y^{rel} \right] \quad (\mathbf{A})$$

Il faut bien voir ici qu'au moment de cibler la relance téléphonique, Y^{rel} est inconnue. **Il faut donc pour établir les priorités de relance estimer Y sur les unités non répondantes**. Dans le cas d'une enquête répétitive, nous pouvons estimer Y pour les unités à relancer à partir de la réponse à la précédente enquête (si l'unité était répondante à la précédente enquête). Dans le cas d'une enquête non-répétitive (ou dans le cas d'une unité non-répondante aux vagues précédentes), il faut estimer à

partir d'une approche modèle, dans un premier temps, la valeur de la variable d'intérêt pour les unités non-répondantes.

Comment traiter les unités qui ont répondu au questionnaire, mais pour lesquelles il nous manque la réponse à la variable d'intérêt (cas de la non-réponse partielle sur la variable d'intérêt) ? La première idée demeure bien sûr de relancer ces unités, dans le cadre des contrôles/relances pour non réponse partielle cette fois. Ce type d'unités n'apparaitra pas dans les priorités de relances définie dans cette partie, qui sont donc bien **des priorités de relance sur la variable cible lorsque l'unité n'a pas répondu du tout au questionnaire**. Il reste néanmoins un problème à régler : dans la formule précédente il est nécessaire d'avoir une valeur (ou estimation pour les non-répondants partiels) sur la variable cible pour toutes les unités qui ont répondu. De ce fait, nous proposons de corriger la non-réponse partielle à ce niveau, selon un mécanisme d'imputation aléatoire. Pour le cas des enquêtes répétitives, il peut s'agir de reprendre le modèle de correction de la non-réponse partielle adopté lors de l'enquête précédente. Pour les enquêtes non-répétitives, il faut donc proposer - puis affiner - un modèle de correction de la non-réponse partielle sur la variable cible en cours d'enquête. Cela peut s'avérer délicat, notamment au début du processus. Il est bon de souligner que ce modèle de correction de la non-réponse partielle doit s'éloigner le moins possible de la méthode utilisée pour estimer la variable cible sur les non-répondants totaux. Même si la finalité n'est pas la même, ce modèle peut être le même, ce qui peut assurer un gain de temps.

Un cas particulier intéressant est celui où **la correction de la non-réponse s'effectue sur les strates de tirage de l'échantillon**. Cela peut être le cas en début de processus pour les enquêtes non-répétitives, même si nous préconisons de déterminer rapidement en cours d'enquête les GRH sur lesquels se basera la correction de la non-réponse. Dans ce cas, comme les strates de correction de la non-réponse sont confondues avec les strates de tirage, les w_i sont égaux dans la formule (A) : il s'agit de l'inverse du taux de sondage de la strate. La formule (A) devient donc :

$$\left(Y_2 - Y_2^R\right) = \frac{n_s w_s}{r_s + 1}\left[\left(\frac{1}{r_s}\sum_{rep}^{s} Y_i\right) - Y^{rel}\right]$$

Il est fréquent, pour une enquête entreprise, d'avoir **plusieurs variables cibles quantitatives**. Au-delà d'avoir un rang de priorité pour chaque variable cible, il peut être intéressant de fournir une seule liste de priorités qui mixe les différentes variables cibles. Cela peut se révéler intéressant notamment d'un point de vue pratique par rapport aux gestionnaires de l'enquête. Dans le cas d'enquêtes répétitives où l'estimation des variables cibles est réputée fiable, nous avons à notre disposition une estimation de référence pour la variable cible. Il est ainsi possible de « **normer** », c'est à dire de diviser la formule (A) par l'estimation de la variable cible. De fait nous pouvons sommer les n valeurs issues de A et définir des priorités sur la somme. En cas d'enquête non-répétitive ou lorsque l'estimation d'une ou plusieurs variables est jugée peu fiable, il est possible et sans doute préférable d'agir sur **les rangs**. Ainsi un rang de priorité de relance est défini pour chaque variable cible. Nous pouvons faire la somme des rangs et définir ainsi un ordre de priorité sur cette somme de rang.

Pour définir les priorités de relances, nous avons juste retenu ici comme critère : se rapprocher le plus possible de l'estimateur obtenu en l'absence de non-réponse, qui est lui la plupart du temps sans biais. De ce fait, dans la définition des priorités, **les entreprises à relancer en priorité sont celles qui, dans leur strate, sont le plus éloignées de la moyenne**. Nous voyons bien la justification de cela par rapport au critère retenu mais cela pose un problème relativement à l'estimateur de la variance. En effet aller « chercher les valeurs extrêmes » revient à augmenter la variance sur l'échantillon, et donc probablement à **surestimer la variance sur la population**. Il y a donc un risque que la variance sur la population, et donc les intervalles de confiance, soient mal estimées ou exagérément pessimistes. De plus, nous voyons bien que si l'étude de la distribution de la variable (et non plus uniquement l'estimation de cette variable) est une priorité forte du concepteur, ce type de méthode peut poser problème. **Comment remédier à ces problèmes ?** Deux pistes peuvent être envisagées. La première piste consisterait à appliquer la méthode précédante uniquement aux « très grandes unités ». Par exemple dans les enquêtes entreprises, on limiterait l'application de cette méthode aux unités des strates exhaustives. Pour les autres unités, on pourrait appliquer la méthode « minimisation de la perte de variance anticipée ». **La seconde piste consisterait à définir un critère unique qui combine les deux rapproches** : se rapprocher de l'estimateur en l'absence de non-réponse d'une part, se rapprocher de la variance en l'absence de non-réponse d'autre part.

2.3. La mise en application pour les enquêtes TIC et déchets

En termes **de mode opératoire**, il faut donc permettre le calcul du terme (A) pour les unités non-répondantes c'est à dire :

- Calculer le nombre de répondants dans les GRH de correction de la non-réponse ;
- Estimer la variable cible pour toutes les unités non-répondantes (phase sans doute la plus délicate) ;
- Estimer la variable cible pour les unités avec non-réponse partielle.

De fait, par variable cible, une fois l'ordre de priorité de relance calculé et transmis aux gestionnaires de l'enquête il faut **suivre l'estimation en cours d'enquête des variables cibles**, en étant notamment attentif aux variations importantes. Pour chaque unité potentiellement à relancer, nous pouvons déterminer **la variation de l'estimateur si l'unité répondait en phase avec la donnée prévue**. L'examen de la variation de l'estimateur, par exemple pour la première unité à relancer, au fil du temps est un renseignement précieux. Nous nous attendons bien sûr à une diminution de cette variation.

Cette méthode a été appliquée la première fois dans le cadre de **l'enquête déchets début 2007**, enquête dont l'objectif principal est de mesurer le tonnage de déchets non dangereux généré par les établissements commerciaux. Cette enquête se déroulait pour la première fois en France, ce qui a posé **deux difficultés** pour l'application de la méthode décrite ci-dessus. Premièrement, il était nécessaire d'estimer, pour les non-répondants, le total des déchets non dangereux. Pour cela, un

modèle a été établi sur les répondants entre l'effectif de l'établissement et le tonnage de déchets émis, par type d'activité. Ce modèle a été jugé plus robuste pour certaines activités que d'autres. La mise en application de ce modèle n'a pas posé de difficultés mais s'est avéré coûteuse en temps. Les listes étant actualisées toutes les deux semaines, il était nécessaire d'actualiser les coefficients du modèle, qui se sont avérés stables au-delà d'un certain seuil de réponse. Deuxièmement, il a été nécessaire de « **supposer » un modèle de correction de la non-réponse a priori en début d'enquête**. Nous avons donc supposé que les groupes de réponses homogènes étaient identiques aux strates de tirage. Faute de temps, nous avons gardé cette hypothèse tout au long de l'application de l'enquête. Les strates de tirages résultaient d'un croisement entre l'activité de l'établissement et sa taille, en regard des effectifs. Après la fin de gestion de l'enquête, lors de la correction de la non-réponse, nous nous sommes aperçus que le critère géographique était à prendre en compte dans la définition des groupes de réponses homogènes. Notamment du fait d'un faible taux de réponse dans les grandes villes, surtout pour le commerce de détail. L'écart entre GRH retenus suite à la correction de non-réponse et GRH retenus a priori a été de nature à **fragiliser la méthode**. Pratiquement des listes de priorités étaient transmises à chaque gestionnaire toutes les semaines durant la phase de gestion de l'enquête. Cette opération est bien passée auprès des gestionnaires. **Toutefois ceux-ci se sont révélés « troublés » par l'aspect dynamique de la méthode**. Pour les établissements qui restaient non-répondants, l'ordre de priorité pouvait changer au fil du temps, ce qui n'était pas toujours facile à interpréter par les gestionnaires.

La méthode a été appliquée une deuxième fois sur **l'enquête TIC fin 2007/début 2008**. Cette enquête peut être jugée répétitive car la plupart des questions (2/3 du questionnaire) sont posées tous les ans C'est un plus indéniable par rapport à l'exemple précédent. La variable d'intérêt retenu ici est le montant des ventes électroniques par internet. En effet lors des précédentes enquêtes TIC, nous nous sommes aperçus que l'estimation des volumes de ventes et d'achats électroniques étaient particulièrement fragiles. Du fait que cette enquête avait eu lieu l'année passée sur le même thème, plusieurs points méritent d'être soulignés. Pour les entreprises qui sont présentes dans les deux échantillons (TIC 2006/2007 d'une part et TIC 2007/2008 d'autre part), nous prenons comme estimations des ventes, l'estimation retenue pour l'an dernier. Il peut s'agir des valeurs brutes de l'an dernier ou des valeurs retenues suite à la correction de la non-réponse. Ces estimations servent lorsque l'unité concernée est potentiellement à relancer (non-réponse totale) ou lorsqu'elle est n'a pas répondu aux variables cibles (non-réponse partielle). Pour **corriger la non-réponse partielle sur les ventes électroniques** sur les autres entreprises, nous avons choisi le modèle de correction de non-réponse partielle adopté pour cette même variable lors de la précédente enquête (TIC 2006/2007). En effet mettre en place un modèle de correction de non-réponse partielle fondé sur les répondants à TIC 2007/2008 apparaît fragile notamment en début de processus. Pour **estimer les ventes électroniques sur les non-répondants totaux** non interrogés lors de la vague précédente, nous proposons d'appliquer ce même modèle de correction de la non-réponse partielle mis en évidence sur la précédente enquête. C'est sans doute une des fragilités de la méthode, mais nous ne voyons pas trop actuellement comment faire mieux. Lors de ce processus de mise en évidence de priorités de

relances, **les estimateurs des ventes seront analysés** fréquemment. L'impact de la première unité à relancer sur les estimateurs sera également analysé. Ainsi des listes de priorités de relances ont été transmises aux gestionnaires de cette enquête toutes les deux semaines. Par rapport à l'enquête déchets, nous avons essayé de faciliter cette opération **en reportant sur les listes des renseignements importants pour les gestionnaires**, par entreprise : entreprise déjà contactée ou non par le gestionnaire, délai éventuel de réponse par exemple. En pratique cette opération s'est (très) bien passée. Elle a permis d'améliorer la qualité de l'estimateur des ventes électroniques par internet, en ciblant les contrôles et relances, en dialoguant d'avantage avec les entreprises sur le sujet en cours de collecte. Il est par contre nécessaire **d'accompagner ce type d'opération auprès des gestionnaires**. A titre d'exemple, en période de collecte, les gestionnaires se réunissaient tous les lundis matins. Nous avons pu lors de ce type de réunions présenter les grandes lignes de la méthode, bien détailler le mode opératoire et l'aménager en fonction des remarques pratiques utiles des gestionnaires. Il est notamment apparu important de présenter le tableau de bord du commerce électronique aux gestionnaires, le fait que les estimateurs devenaient de plus en plus précis était aussi perçu comme une traduction immédiate de l'effort porté par les gestionnaires sur les variables du commerce électronique.

Cette approche est en cours d'application pour **l'enquête TIC 2009**. Deux modifications importantes peuvent être signalées. D'une part, en ce qui concerne les priorités de relance pour non-réponse totale. En premier lieu les entreprises à relancer en priorité sont les non-substituables, en deuxième lieu les entreprises connues pour faire du commerce électronique et repérées en tant que tel par le syndicat professionnel FEVAD, et en troisième lieu selon le critère de priorités de relance défini précédemment sur les ventes par internet. D'autre part, et surtout, le mode opératoire a changé. Jusqu'à présent la maitrise d'œuvre de l'enquête était assuré par l'ex pôle ESSS, qui était situé tout comme le pôle Ingénierie Statistique Entreprises à la direction régionale des Pays de la Loire. Le mode de fonctionnement choisi était donc établi en fonction de cette proximité. En pratique, pour les priorités de relance, le pôle ISE mettait à disposition toutes les deux semaines des listes de priorités par gestionnaire sur un disque partagé. A partir de 2009, en lien avec le projet ESANE, la maitrise d'œuvre de l'enquête est assurée par le pôle Enquêtes Entreprises de la Direction Régionale de Midi-Pyrénées. En accord avec les maitrises d'œuvre et d'ouvrage un fonctionnement différent à été décidé. En pratique le pôle ISE a conçu une macro SAS qui permet de développer le mode opératoire retenu selon les critères de priorités de relance. Cette macro SAS fonctionne à partir de la table issue de la collecte de l'enquête et d'une table qui comprend les données estimées pour les variables cibles pour toutes les entreprises (dans ce cas on se situe de fait avant le début de la collecte). Ainsi le pôle ESE est autonome pour la sortie des priorités de relance (fréquence notamment). Cette macro est paramétrable, un des paramètres est le nombre d'unités à relancer fourni par gestionnaire. Au moment de l'écriture de cet article, la macro SAS était en phase de test par le pôle ESE.

L'objectif de cette contribution était surtout de montrer **l'importance de disposer d'un tableau de bord de suivi sur les variables cibles, avec indicateur de précision, en cours de collecte**. Bien que d'idée assez simple, il n'est pas certain que cette pratique soit généralisée pour toutes les enquêtes. Comme nous l'avons vu, la mise en place de ce tableau de bord et de sorties dérivées permet de mieux repérer les valeurs hors-normes, de cibler les contrôles et de cibler les relances pour non-réponse totale. En ce qui concerne **les priorités de relance pour non-réponse totale, des progrès et des expérimentations sont encore à faire**. Une idée séduisante serait de mixer l'approche décrite précédemment avec l'approche de type variance à gagner. Pour fiabiliser les estimateurs, nous pouvons aussi penser jouer sur l'échantillonnage en appliquant des notions comme les allocations de Neyman dans les sondages stratifiés, ce qui n'est pas fait dans le cas de TIC et du commerce électronique. Une deuxième idée serait donc de simuler ce type d'approche pour voir les modifications engendrées par rapport au plan de sondage actuel. En terme de niveau d'application, le pôle ISE doit définir une méthode de priorités de relance pour l'enquête thématique innovation, dans le cadre différent des extensions régionales.

Bibliographie

[1] Philippe BRION, "Critère d'arrêt de traitement d'enquête" note interne INSEE n°44/E210 du 11 avril 2003. Cette note a été reprise, sous forme résumée, dans « échantilonnage et méthodes d'enquêtes », Dunod 2004, pages 113 à 118.

[2] Phlippe BRION, « Arbitrages entre délais et précision dans les enquêtes statistiques », document de travail de l'INSEE E2007/018

[3] Richard MAC-KENZIE, « A framework for priority contact of non respondents », document de l'Australian Bureau of Statistics (document disponible sur le site de l'OCDE http://www.oecd.org/dataoecd/60/16/30890652.pdf)

[4] Nathalie CARON « la correction de la non-réponse par repondération et par imputation », document de travail de l'INSEE n°M0502