Mardi 15 mars 2005 : 15h30-15h55 *(en parallèle avec la session 9)*

Optimiser la production en statistique d'entreprise : lien entre coût minimal de vérification et finesse de diffusion.

Pascal RIVIERE
INSEE-département des applications et des projets

Lors d'une enquête auprès d'entreprises, mais aussi dans les sources administratives que l'on traite sur cette même population, les données que l'Insee reçoit sont loin d'être parfaites : on observe en pratique de très nombreuses anomalies, certaines correspondant à de réelles erreurs, d'autres non. On a donc toujours, d'une manière ou d'une autre, un programme qui traite automatiquement les données pour déterminer quels sont ces questionnaires (ou formulaires) « douteux ». Ceux-ci sont ensuite vérifiés manuellement, et la plupart du temps on en vérifie trop, car de nombreux « douteux » (ainsi jugés par le programme) se révèlent corrects dans les faits. On peut donc peut-être éviter de tout vérifier.

Ce qui soulève deux questions : en fonction de quels critères optimiser la vérification ? Comment estimer a priori le coût minimal de cette opération ?

L'optimisation des vérifications devrait idéalement s'appuyer sur un critère de qualité, qui ne peut être qu'un critère de précision : il s'agirait d'estimer l'erreur quadratique moyenne, en prenant en compte, idéalement, l'ensemble des composantes de l'erreur. Un tel indicateur présenterait un double intérêt. Il permettrait d'abord de prioriser les opérations, en fonction du "gain potentiel en qualité" apporté par la vérification-apurement, c'est-à-dire l'écart entre l'indicateur actuel et l'espérance de l'indicateur après intervention humaine. Il s'agirait donc toujours du gain procuré par une tâche élémentaire du gestionnaire, comme vérifier une donnée douteuse, vérifier une donnée jugée acceptable par le programme (car rien ne dit qu'il ait raison), rappeler un non-répondant, ... L'indicateur de qualité fournirait en second lieu la possibilité de définir un critère d'arrêt : dès qu'il atteint une valeur-cible que l'on considère comme satisfaisante, on peut arrêter les contrôles manuels.

Dans la pratique, construire de tels indicateurs de qualité pour piloter la production se révèle très délicat. La première raison, la plus classique, réside dans la difficulté qu'il y a à estimer une erreur prenant en compte de multiples facteurs. Mais même si l'on suppose que la méthodologie existe, cette précision devra être calculée non seulement par variable, mais aussi par domaine de diffusion. En effet, le but n'est pas de produire précisément le total d'une variable pour la France entière mais bien de produire toute une batterie de statistiques, sur des variables et domaines très divers (secteurs d'activité économique, régions, ...). Ainsi l'optimisation de la production statistique s'effectue-t-elle en fonction de ces domaines de diffusion.

Intuitivement, il est évident que plus la finesse de diffusion en importante, plus le coût de vérification sera élevé : on aura plus de travail si l'on diffuse au niveau zone d'emploi que si l'on reste au niveau régional, par exemple. Mais il est difficile d'évaluer **a priori** le lien qui existe entre ce coût minimal de vérification et le nombre de "cases" de diffusion, entre autres parce que l'on ne sait pas, avant de vérifier les données, si elles seront correctes ou non.

Pour arriver à faire cette évaluation a priori, nous sommes obligés de simplifier le problème de manière drastique.

On considère que l'objectif est de fournir des statistiques sur un certain nombre de domaines de diffusion. On notera $\boldsymbol{H}$ le nombre de domaines, $\boldsymbol{N}$ la taille de l'échantillon considéré (nombre total de *questionnaires envoyés* dans le cas d'une enquête, nombre de formulaires dans le cas d'une

source administrative). Dans un domaine de diffusion donné, le travail de production va consister à vérifier les questionnaires douteux (en proportion f) un par un ; on suppose que l'intervention humaine les fait passer de « douteux » à « acceptables ».). On suppose aussi que le nombre de cas à vérifier est équiréparti entre les domaines. Notre critère d'arrêt sera très simple : on décide que l'on arrête la vérification dès que le taux de questionnaires erronés est, avec une probabilité supérieure à 1 - a, au-dessous d'un taux-limite r que l'on se fixe. Nous sommes loin ici du raisonnement fondé sur l'erreur quadratique moyenne, mais en même temps cette façon de procéder peut se rapprocher d'une certaine pratique, où les gestionnaires qui effectuent la production supportent mal l'idée d'un taux d'erreur important.

En modélisant cela (ce qui constitue le centre du papier), on peut mettre en évidence, par approximation et en vertu d'hypothèses simplificatrices, un nombre minimal n^* de questionnaires à vérifier, défini par la formule suivante :

$$\frac{1}{n^*} = \frac{1}{fN} + \frac{a}{H},$$

où $$a = \frac{\frac{r}{f}}{g_a^2\left(1 - 2\frac{r}{f}\right)}$$

En d'autres termes, l'inverse du coût minimal est fonction affine de l'inverse du nombre de domaines de diffusion.

La conclusion à en tirer, c'est que si l'on se dote de critères de qualité (nécessairement liés à ce que l'on diffuse, donc aux domaines de diffusion), il existe bien un lien direct entre coût de la vérification des questionnaires et nombre de domaines de diffusion. On s'aperçoit que la production devient inévitablement très coûteuse lorsque les tailles d'échantillons par domaine sont faibles. La priorisation des vérifications, fondée sur des indicateurs de précision et des mesures de gain en précision, est en cela une façon d'optimiser.