Le plan de sondage de l'enquête

L'échantillon de l'enquête TIC-TPE est tiré dans une base de sondage construite à partir du répertoire Sirus. La méthode d'échantillonnage est un **sondage aléatoire simple stratifié** selon le croisement entre le secteur d'activité de l'unité légale, sa tranche d'effectif et sa tranche de chiffre d'affaires.

Stratification

Les modalités des secteurs d'activité ont des niveaux d'agrégation très divers : de la classe au regroupement de sections, en passant par des regroupements de divisions ou de groupes.

Les tranches d'effectif sont découpées en trois modalités :
- 0 à 1 personne occupée ou effectif manquant ;
- 2 à 4 personnes occupées ;
- 5 à 9 personnes occupées.

Les tranches de chiffre d'affaires sont définies en quatre modalités, où « s » est un seuil d'exhaustivité :
- 0 à moins de 1 million d'euros ou chiffre d'affaires manquant ;
- 1 à moins de 2 millions d'euros ;
- 2 à s millions d'euros ;
- plus de s millions d'euros.

Les seuils d'exhaustivité dépendent du grand secteur auquel appartient l'unité. Ils sont déterminés de façon à retenir dans la partie exhaustive de l'échantillon les 40 unités ayant les plus grands chiffres d'affaires par grand secteur.

Nombre d'unités interrogées

Le nombre d'unités à échantillonner diffère selon les strates :
- Les unités dont le chiffre d'affaires dépasse le seuil fixé pour leur grand secteur d'activité sont interrogées exhaustivement.
- Pour les autres strates, le nombre d'unités à interroger est obtenu sur la base d'une allocation proportionnelle au nombre d'unités, *a priori* plus adaptée aux paramètres d'intérêt de type proportion. Afin d'améliorer l'estimation des montants, cette allocation a été modifiée pour interroger davantage d'unités dans les regroupements d'activités ayant une forte dispersion de montants de ventes web.
- Un nombre minimum d'unités est par ailleurs imposé dans les strates non exhaustives : il est de 10 par strate pour les tranches de chiffre d'affaires de moins de 2 millions d'euros, et de 5 par strate pour les chiffres d'affaires compris entre 2 et s millions d'euros. Lorsque les strates ne comportent pas suffisamment d'unités, on en tire le maximum.

In fine, **11 100** unités sont interrogées.

Le traitement de la non-réponse et le calage

Comme pour toutes les enquêtes, les résultats bruts de l'enquête TIC-TPE sont traités pour répondre à deux objectifs principaux :
- corriger le biais induit par les non-réponses totales et partielles ;
- améliorer la précision des estimations.

Les résultats de l'enquête seraient biaisés si l'on ne corrigeait pas la non-réponse, sauf dans l'hypothèse où les non-répondants auraient un comportement identique à celui des répondants, ce qui n'est pas le cas, puisque les non-répondants ne se répartissent *a priori* pas au hasard. Aussi, afin de compenser le biais, on effectue les traitements décrits ci-dessous.

Traitements préalables

Les unités enquêtées sont ventilées dans l'une des catégories suivantes :

- les unités **répondantes appartenant au champ de l'enquête** ;
- les unités **du champ reconnues comme non-répondantes « totales »** : cette catégorie englobe les unités du champ n'ayant pas retourné de questionnaire, ainsi que les unités du champ ayant renvoyé un questionnaire inexploitable ;
- les unités repérées comme **hors-champ** ;
- les unités **sans information** : il s'agit des unités pour lesquelles aucun questionnaire n'a été réceptionné, et dont on ne sait donc pas si elles sont hors-champ ou non-répondantes totales.

Traitement des unités sans information

Une des premières étapes du redressement de la non-réponse consiste à ventiler les unités sans information dans les autres catégories. Pour ce faire, on utilise les informations disponibles dans des sources externes (le répertoire Sirene notamment).

Si les recherches effectuées ne permettent pas de statuer, le poids des unités restant sans information est redistribué sur les autres unités. On construit pour cela des « groupes d'information homogènes » à l'aide d'une segmentation par arbre. Les variables retenues doivent être connues sur les unités répondantes et sur les autres, ce qui exclut le recours aux questions de l'enquête. Ces variables doivent permettre de prendre en considération les caractéristiques des unités et la probabilité d'avoir une information sur elles. Une fois les groupes constitués, les poids sont ensuite « transférés », au sein de chacun, des unités sans information aux autres unités.

Les unités pour lesquelles on est sûr d'avoir une information (via une source administrative) sont exclues de cette opération de repondération.

Redressement de la non-réponse totale

On considère que le fait de répondre ou pas à l'enquête est une nouvelle phase de sondage – la probabilité de sélection de chaque unité étant sa probabilité de répondre. On corrige donc la non-réponse en divisant le poids par la probabilité de réponse qu'il faut estimer.

On construit pour cela des « groupes de réponse homogènes » à l'aide d'une segmentation par arbre. Les variables retenues doivent être connues sur les unités répondantes et sur les autres, ce qui exclut le recours aux questions de l'enquête. Ces variables doivent être liées au comportement de réponse des unités.

Ont été retenus : le secteur d'activité de l'entreprise, son chiffre d'affaires, son ancienneté, son taux d'endettement, sa localisation et son caractère actif ou cessé.

Calage

Le calage permet d'améliorer la précision des estimations. L'idée générale est de retrouver, en utilisant l'échantillon et le poids associé à chaque unité, le bon nombre d'entreprises par domaine de diffusion (par secteur d'activité par exemple), ce qui n'est plus le cas après avoir modifié les poids pour traiter les unités sans information et pour corriger la non-réponse totale. Si les variables d'intérêt sont liées à ces domaines de diffusion, on réduira ce faisant la variance des estimateurs.

On modifie donc à nouveau le poids des entreprises à l'aide d'une méthode usuelle de « calage sur marges ». La méthode consiste à changer les poids des unités de telle sorte que le nombre d'unités total, estimé à partir des unités répondantes de l'échantillon, soit égal au total connu par ailleurs (la « marge », correspondant ici au nombre d'unités dans le champ de l'enquête), et ce tout en minimisant la déformation des poids par rapport aux poids d'origine.

L'échantillon est « calé » sur les marges suivantes, provenant du répertoire Sirene, les « grands secteurs » étant définis en trois modalités par industrie/commerce/services :

- nombre d'entreprises par secteur d'activité de stratification ;
- nombre d'entreprises par grand secteur x tranche d'effectif ;
- nombre d'entreprises par grand secteur x tranche de chiffre d'affaires ;

- nombre d'entreprises par catégorie juridique ;
- nombre de personnes occupées par grand secteur ;
- chiffre d'affaires par grand secteur.

Winsorisation

Après le calage, un dernier traitement de repondération est réalisé : la « winsorisation ». Cette étape vise à limiter l'impact de certaines unités qui pourraient influer de manière excessive sur l'estimation de certains agrégats, du fait de la valeur de la variable pour l'unité concernée. On calcule alors un nouveau poids en se fondant sur une valeur « acceptable » de la variable, obtenue à partir des valeurs déclarées par les autres unités de leur secteur d'activité. À l'issue de cette opération, la variable de poids est définitive.

Notons que cette étape doit être réalisée après la correction de la non-réponse partielle, car la winsorisation nécessite de connaître les valeurs finales des variables pour estimer les agrégats.

Redressement de la non-réponse partielle

Le redressement de la non-réponse partielle concerne les unités du champ ayant répondu à l'enquête, mais pas à toutes les questions auxquelles elles auraient dû répondre. Cette étape permet de compléter les items auxquels les entreprises n'ont pas répondu.

Certaines réponses peuvent se déduire de manière automatique. Dans ce cas, on parle de « règles déterministes ». Par exemple, si l'entreprise a déclaré avoir accès à internet, elle est supposée avoir un ordinateur.

Une fois ces règles appliquées, chaque variable du questionnaire est répartie, pour chaque unité répondante du champ, selon les trois situations suivantes :

- l'unité est concernée par la question et y a **répondu** ;
- l'unité est concernée par la question mais n'y a **pas répondu** ;
- l'unité n'est **pas concernée** par la question et n'a donc pas à y répondre (par exemple, si l'entreprise n'a pas d'ordinateur, elle n'a pas à répondre aux questions sur l'usage d'internet).

Les variables redressées dans le cadre de la correction de la non-réponse partielle relèvent du deuxième cas de figure. Après traitement, elles présentent une réponse.

En règle générale, on procède à une imputation aléatoire par donneur (hot-deck), c'est-à-dire qu'on utilise les renseignements donnés par les unités répondantes d'une catégorie (selon des critères de taille ou de secteur d'activité par exemple), pour compléter les questions manquantes des unités de la même catégorie.

Les caractéristiques prises en compte peuvent être des caractéristiques de l'unité connues avant l'enquête et des réponses à d'autres items du questionnaire. Ceci permet de conserver un lien statistique entre variables.

Certaines variables quantitatives sont imputées différemment.

- Les données de cadrage (chiffre d'affaires et effectif) sont imputées prioritairement à l'aide de sources externes (répertoire Sirene, fichier Fare).
- Dans le cas de pourcentages, les données sont d'abord discrétisées en dix tranches, l'imputation aléatoire portant sur la tranche de pourcentage. Le chiffre finalement affecté est la médiane de la tranche de pourcentage imputée.
- Pour les variables sur le montant des ventes web, pouvant être renseignées en valeur ou en pourcentage du chiffre d'affaires, l'imputation porte sur le pourcentage. Le montant en valeur est ensuite déduit par l'application de ce pourcentage au chiffre d'affaires.

Cas des unités non substituables

Les unités dites « non-substituables » correspondent à des unités particulières ne représentant qu'elles-mêmes. Elles ont ainsi un poids de 1 et font l'objet de traitements spécifiques : elles sont exclues des phases de repondération pour conserver leur poids unitaire et leur non-réponse totale est corrigée par imputation – au besoin en faisant intervenir des renseignements « à dire d'expert ».