Enquête sur le milieu de travail – Imputation

Il existe trois types de non-réponse dans l'EMTE : la non-réponse des unités, la non-réponse à des questions et la non-réponse d'un cycle. Il y a non-réponse des unités lorsqu'il est impossible d'obtenir des données d'enquête pour toutes les variables d'une unité sélectionnée (milieu de travail ou employé) en raison d'un refus ou de l'impossibilité d'entrer en communication. Il y a non-réponse à des questions si nous sommes capables d'obtenir seulement une réponse partielle d'une unité sélectionnée, à la suite d'un refus, de l'impossibilité de répondre à certaines questions ou d'incohérences dans les données recueillies. Enfin, il y a non-réponse d'un cycle lorsque nous disposons au moins d'une réponse partielle provenant d'un cycle précédent pour une unité sélectionnée, sans détenir de renseignements pour le cycle actuel. La stratégie actuelle prévue pour le traitement des non-réponses prévoit le calcul d'une correction du poids pour les répondants dans le cas des non-réponses des unités, tandis que les non-réponses à des questions et les non-réponses d'un cycle sont traitées à l'aide de diverses méthodes d'imputation. Des versions transversales de ces méthodes sont utilisées pour les unités qui apparaissent pour la première fois dans le cycle actuel. Sinon, si des données historiques existent, des versions longitudinales sont utilisées.

Dans le cas de la non-réponse à des questions, on procède à un traitement et à une vérification avant l'imputation, afin de supprimer les incohérences dans les données recueillies. La vérification est fondée sur une série de règles qui doivent ou devraient être respectées. Ce processus mène la création de valeurs manquantes supplémentaires ou à l'imputation par déduction des valeurs qui auraient dû être déclarées. On utilise ce type d'imputation quand on peut déduire uniquement un seul champ manquant à partir de l'information fournie. Par exemple, si une composante d'une somme est manquante et si l'on dispose des autres composantes incluant la somme, on peut alors déterminer la composante manquante.

Une fois ce processus terminé, on impute les autres valeurs manquantes à l'aide d'une des quatre méthodes décrites ci dessous. Pour éviter de produire des incohérences à l'intérieur des données imputées, la plupart des champs interdépendants sont imputés sous forme de bloc. Puisque plusieurs questions font parties de cette catégorie, un système post-imputation a été élaboré pour conserver les relations entre les champs.

Quatre méthodes d'imputation sont présentement utilisées pour les portions employeurs et employés de l'EMTE : l'imputation par report, l'imputation par répartition, l'imputation « hot deck » pondérée et l'imputation par le plus proche voisin. On utilise l'imputation par report lorsque des données historiques sont disponibles. Elle consiste simplement à transférer la valeur du cycle précédent au cycle actuel. Pour les variables continues, la valeur peut être ajustée par une

tendance provenant d’une variable auxiliaire. Évidemment, il n’existe pas de version transversale de l’imputation par report.

On applique l’imputation par répartition lorsqu’on demande aux répondants de fournir un total et sa ventilation en plusieurs catégories dans les cas où il manque deux catégories ou plus. On calcule la répartition des catégories au macro-niveau et on l’applique au micro-niveau. Pour illustrer cette méthode, supposons qu’un répondant nous a donné l’emploi total, mais n’a pas été en mesure d’en fournir une ventilation par groupe professionnel. Nous appliquerions la répartition des groupes professionnels calculée au niveau industrie/taille au chiffre sur l’emploi total afin d’imputer les champs manquants. Cette méthode est uniquement transversale.

Dans le cas de la méthode « hot deck » pondérée, on impute un champ manquant à l’aide de la réponse d’un donneur choisi de façon aléatoire à l'intérieur d’une classe d’imputation; on impute directement la valeur du donneur pour la variable manquante ou le ratio établi pour le donneur entre la variable manquante et une variable auxiliaire. Dans ce dernier cas, le ratio est alors multiplié par la valeur auxiliaire du receveur. La méthode est longitudinale si la variable auxiliaire ou les classes d’imputation sont déterminées à partir des données du cycle précédent; sinon, elle est transversale. On sélectionne au hasard le donneur, la probabilité de sélection étant égale au ratio de son poids d’échantillonnage divisé par la somme des poids d’échantillonnage de toutes les unités contenues dans la classe d’imputation correspondante. On a adopté la méthode « hot deck » pondérée pour les deux principales raisons suivantes : i) la méthode est facile à appliquer et ii) elle donne des estimations ponctuelles approximativement sans biais pourvu que l’on puisse supposer que toutes les unités de chaque classe d’imputation ont la même propension à répondre (Rao, 1996).

Enfin, on utilise l’imputation par le plus proche voisin afin de préserver les liens qui existent entre certaines variables. Il s’agit d’une méthode d’imputation par donneur, tout comme la méthode « hot deck » pondérée. Cette méthode remplace les valeurs manquantes d’un receveur donné par les valeurs correspondantes du donneur qui se rapprochent le plus du receveur pour certaines variables d’appariement. Tout comme dans le cas de la méthode « hot deck » pondérée, un ratio du donneur peut être imputé pour le receveur, qui est alors multiplié par une variable auxiliaire du receveur. Encore une fois, la méthode est longitudinale si la variable auxiliaire, les classes d’imputation ou les variables d’appariement sont déterminées à l’aide des données du cycle précédent; sinon, elle est transversale. Cette méthode est très semblable à la méthode « hot deck » pondérée, sauf pour la sélection des donneurs. Dans le cas de la méthode « hot deck » pondérée, les donneurs sont choisis au hasard, tandis que, pour l’imputation par le plus proche voisin, ils font l’objet d’une sélection déterministe d’après certaines variables d’appariement.