MÉTHODES D'IMPUTATION DE VALEURS ABERRANTES POUR DES DONNÉES D'ENQUÊTE.

Ruilin REN (*)

(*) CREST / ENSAI

Cet article a pour objectif d'étudier des méthodes d'imputation de valeurs aberrantes représentatives pour des données issues d'une enquête par sondage. Par valeur aberrante représentative, nous désignons une valeur observée dans l'échantillon qui a un grand écart par rapport à sa valeur prévue, et qui existe aussi dans la partie non-observée de la population.

Il est évident que les méthodes d'imputation pour des données manquantes appliquent à ce cas, mais une valeur manquante est différente d'une valeur aberrante au sens statistique. Une valeur aberrante peut fournir de l'information utile en elle-même. Cette information peut être utilisée dans l'imputation pour lui imputer une valeur raisonnable tout en évitant la défaillance des méthodes standards en utilisant les valeurs imputées.

Pour arriver à ce but, nous étudions une méthode d'imputation basée sur un résultat d'estimation du total résistant aux valeurs aberrantes. En général, les estimateurs résistants ne sont pas élaborés pour une utilisation par le grand public. Les instituts de statistique souhaitent délivrer un fichier de données prêt à l'emploi par tous les publics et requièrent seulement l'utilisation des méthodes classiques. Il faut donc modifier ou imputer les valeurs aberrantes par des valeurs normales ou des valeurs moins aberrantes telles que l'estimation du total obtenue par une méthode résistante puisse être retrouvée par des méthodes classiques.

Cette procédure d'imputation est une procédure de calage sur marge. Elle est comparée avec les méthodes qu'on utilise pour l'imputation des valeurs manquantes, par exemple, les méthodes de *hot deck* et *cold deck*, la méthode du *plus proche voisin* et la méthode par *régression*. Les méthodes sont validées par des données d'enquête issues d'une enquête auprès des entreprises.