

# IMPORTANCE D'UNE BONNE MÉTHODE D'IMPUTATION DES QUESTIONS D'ÉDUCATION DANS LE CADRE DE L'EPA

Caroline Pelletier<sup>1</sup>

## RÉSUMÉ

L'Enquête sur la population active (EPA), comme toute enquête, doit faire face au traitement de la non-réponse. L'étude porte sur le traitement de la non-réponse pour les questions d'éducation servant à dériver le plus haut niveau d'éducation atteint par une personne. Il s'agit d'une information auxiliaire utilisée dans la formation des classes d'imputation du système hot-deck développé pour la composante principale de l'Enquête. La méthode proposée améliorera la qualité de la variable dérivée et contribuera à améliorer l'imputation de données manquantes dans la composante principale tout en remédiant au biais subsistant dans les données actuellement imputées manuellement par des codeurs expérimentés.

MOTS CLÉS : Imputation cold-deck; imputation déterministe; imputation historique; non-réponse.

## ABSTRACT

The Labour Force Survey (LFS), like other surveys, has to deal with the treatment of nonresponse. This study is about the treatment of nonresponse for the education questions used to derive the highest level of education ever attained by a person. It is used in the creation of imputation classes for the hot-deck imputation system developed for the principal component of the Survey. The proposed method will increase the quality of the derived variable and contribute to the improvement of the imputation of missing data in the principal component while removing the bias in the data currently imputed manually by experienced coders.

KEY WORDS: Carry-forward imputation, Cold-deck imputation, Deterministic imputation, Nonresponse

## 1. INTRODUCTION

L'Enquête sur la population active (EPA) est une enquête mensuelle, par panel avec 6 groupes de rotation qui recueille des données permettant de produire des estimations dont le taux de chômage pour différents niveaux géographiques et facteurs démographiques. À tous les mois, près de 54 000 ménages sont contactés, ce qui représente un peu plus de 100 000 personnes. Le questionnaire soumis à chaque ménage comprend plusieurs composantes dont la composante démographique. Les questions se rapportant à la composante démographique sont posées uniquement lors de la première entrevue à moins que la composition du ménage ne change. La composante démographique inclut des questions sur l'âge, l'état matrimonial et notamment sur l'éducation.

Le traitement des données des questions sur l'éducation constitue le sujet principal de la présente étude. Cet article vise à donner une meilleure compréhension du processus courant de vérification et de traitement des données et à proposer une méthode plus objective et automatique pour l'imputation des questions reliées à l'éducation. Il comporte trois parties, à savoir la description des questions et de la non-réponse reliées à l'éducation, la description du processus courant de vérification et de traitement des données et la présentation d'une nouvelle méthode.

---

<sup>1</sup> Caroline Pelletier, Immeuble R.H. Coats, 16-P, Tunney's Pasture, Ottawa, Ontario, K1A 0T6, Canada, caroline.pelletier@statcan.ca.

## 2. QUESTIONS ET NON-RÉPONSE LIÉES À L'ÉDUCATION

### 2.1 Questions

Dans la composante démographique du questionnaire de l'EPA, quatre questions sur l'éducation sont posées. Les réponses à ces questions sont subséquemment utilisées pour dériver le plus haut niveau d'éducation atteint par chaque personne de l'échantillon. Le plus haut niveau d'éducation est une des variables utilisées dans la formation des classes d'imputation du système d'imputation hot-deck servant à imputer les données manquantes de la composante sur les activités sur le marché du travail de l'EPA, ce qui implique comme contrainte de résoudre la non-réponse aux questions sur l'éducation avant l'application du processus d'imputation hot-deck.

### 2.2 Non-réponse

Entre les mois de janvier et de juin 2002, les taux mensuels de non-réponse aux quatre questions reliées à l'éducation n'ont pas excédé 1,0 %. Ces taux correspondent au ratio du nombre de refus et de « ne sait pas » sur le nombre de fois où la question a été posée ou confirmée. Ces taux n'incluent pas les cas où les réponses ont été effacées suite à la détection d'incohérences. La non-réponse est dite totale si l'individu n'a fourni aucune réponse aux quatre questions d'éducation et partielle s'il a fourni certaines réponses à ces mêmes questions. Dans les deux cas, la non-réponse nécessite une imputation qui peut se faire de manière déterministe, par transfert de données ou manuelle dans le processus courant de traitement des données.

## 3. PROCESSUS COURANT DE VÉRIFICATION ET DE TRAITEMENT DES DONNÉES

Selon qu'il s'agisse d'une première entrevue ou d'une entrevue subséquente, le déroulement des opérations de vérification et de traitement diffère quelque peu. Une étape supplémentaire dans le cas des entrevues subséquentes consiste à utiliser l'imputation historique, ou « carry-forward », c'est-à-dire l'imputation par transfert des données du mois précédent.

### 3.1 Règles de vérification

Des règles de vérification sont utilisées à plusieurs reprises au cours du processus de traitement des données afin d'assurer que les données répondent à des critères pré-établis pour la suite des opérations. Lorsqu'un enregistrement échoue ces règles, il peut être corrigé de différentes façons : l'imputation déterministe, l'imputation par transfert de données (dans le cas des entrevues subséquentes seulement) et finalement l'imputation manuelle effectuée par un codeur.

### 3.2 Imputation déterministe

Lorsque le bureau central reçoit les données des répondants, elles sont soumises à des règles de vérification. On effectue alors une imputation déterministe à l'aide d'un processus automatisé. Par exemple, si un répondant n'a pas étudié au niveau post-secondaire, alors il n'a pas obtenu un diplôme à ce niveau.

### 3.3 Imputation historique ou « carry-forward »

L'imputation « carry-forward » ou par transfert de données est présentement utilisée à l'EPA. Si la personne a fourni des réponses plausibles le mois précédent, les réponses du mois précédent sont transférées pour pallier la non-réponse du mois courant. Les données ne sont jamais transférées plus d'une fois. C'est donc dire que cette méthode n'est pas appliquée automatiquement pendant deux mois consécutifs.

### 3.4 Imputation manuelle

L'imputation manuelle est faite par un codeur en utilisant l'information auxiliaire disponible (principalement l'âge et la description de l'emploi actuel ou du dernier emploi occupé) et ses propres connaissances de l'enquête. Après avoir consulté l'information disponible, le codeur évalue ensuite les divers scénarios possibles, choisit celui qu'il croit le plus plausible et attribue les réponses correspondantes aux questions. Cette méthode fait donc appel au jugement du codeur. Comme aucune valeur manquante ou incohérente n'est admise dans la suite du processus pour les quatre questions d'éducation, le codeur doit imputer tous les non-répondants, peu importe que la non-réponse soit partielle ou totale.

Deux codeurs se partagent le travail et doivent effectuer plus de 1000 transactions à chaque mois, ce qui représente près d'une journée de travail par codeur. Une transaction correspond à l'imputation d'une question pour un individu. On peut noter que si un cas est renvoyé au codeur, deux transactions sont comptées. La plupart des cas traités par les codeurs dans la composante démographique sont reliés aux questions sur l'éducation. Le codeur peut devoir imputer de une à quatre questions, ce qui représente plus d'une transaction pour une personne. Par exemple, en juin 2002, 1016 questions ont dû être imputées. Ces questions étaient associées à 403 personnes. C'est donc dire qu'en moyenne, 2,5 questions ont dû être imputées par personne.

#### **4. PROPOSITION D'UNE MÉTHODE AUTOMATIQUE**

Suite à l'observation du processus d'imputation manuelle sur les questions liées à l'éducation, un projet visant à automatiser ce processus a été entamé. Le but principal de ce projet visait à développer un processus automatisé qui reproduirait le travail du codeur d'une façon plus efficace et objective et ce :

- en utilisant le plus possible la même information dont disposent les codeurs;
- en conservant les réponses fournies par le non-répondant lorsque la non-réponse est partielle;
- en développant un outil simple qui puisse s'intégrer facilement au flux actuel des opérations de traitement des données.

La nouvelle méthode d'imputation pour traiter les cas de non-réponse aux questions sur l'éducation qui sont soumis aux codeurs chaque mois comporte trois parties : l'imputation déterministe, l'imputation par transfert de données (historique ou carry-forward) et l'imputation cold-deck basée sur l'âge et la profession. L'application de ces méthodes d'imputation se fait séquentiellement de sorte que les enregistrements traités par la première méthode sont considérés comme complétés et ne sont pas soumis aux deux autres méthodes. De cette façon, tous les enregistrements pourront être traités par l'une ou l'autre des trois méthodes. Après chaque étape, on s'assure que les valeurs imputées ne contiennent pas d'incohérences en appliquant des règles de vérification pré-établies.

##### **4.1 Imputation déterministe**

Nous proposons d'ajouter de nouvelles règles de vérification en émettant les deux hypothèses suivantes :

- 1 – Si un répondant n'a pas terminé son secondaire, alors il n'a pas étudié au niveau post-secondaire.
- 2 – Si un répondant a étudié au niveau post-secondaire, alors il a terminé son secondaire.

Notons que l'on ne peut pas appliquer l'imputation déterministe si la non-réponse est totale car les règles supposent qu'au moins certaines réponses ont été fournies

##### **4.2 Imputation historique ou « carry-forward »**

L'imputation par transfert de données est actuellement appliquée une seule fois et donc pour un seul mois. On suggère de l'étendre à plus d'un mois. Dans le cas d'une non-réponse totale, on ne fait que copier les valeurs du mois précédent au mois courant. Dans le cas d'une non-réponse partielle, on doit d'abord s'assurer que les réponses données le mois courant sont identiques aux valeurs du mois précédent avant de procéder au transfert des données. L'hypothèse sous-jacente formulée est que le niveau d'éducation d'une personne reste constant au cours d'une période de six mois. On évitera également qu'il soit différent entre deux mois à cause de l'application d'une méthode d'imputation différente. Cette méthode ne peut évidemment pas s'appliquer s'il s'agit du premier mois dans l'enquête pour le non-répondant.

##### **4.3 Imputation cold-deck basée sur l'âge et la profession**

La dernière méthode, l'imputation cold-deck basée sur l'âge et la profession du non-répondant, permet d'imputer tous les autres enregistrements. On crée d'abord les classes d'âge suivantes : 15-19 ans, 20-24 ans, 25-29 ans, 30-39 ans, 40-54 ans et 55 ans et plus. Il s'agit des mêmes classes d'âge que celles que l'on retrouve dans la publication de l'EPA.

Les donneurs sont choisis parmi les enregistrements du mois précédent pour lesquels les réponses données aux quatre questions d'éducation ont été conservées telles quelles ou bien imputées par transfert de données. On utilise les répondants

du mois précédent parce que l'on n'a pas accès à tous les enregistrements du mois courant lorsque le traitement manuel est effectué. Il a été vérifié que la distribution du plus haut niveau d'éducation chez les répondants est semblable d'un mois à l'autre. Donc, le choix d'utiliser les répondants du mois précédent pour imputer les non-répondants du mois courant est une option valable.

On crée des classes d'imputation pour trouver un donneur qui servira à imputer un non-répondant. Trois dimensions sont utilisées pour créer les classes d'imputation, à savoir les réponses données par le non-répondant à l'une ou l'autre des quatre questions reliées à l'éducation (dans le cas de non-réponse partielle), la profession et l'âge. Si on ne trouve aucun donneur dans la classe d'imputation du non-répondant, on laisse tomber l'âge et on conserve les deux autres dimensions. Si le code de profession n'est pas disponible, alors on cherche parmi les répondants pour lesquels ce code n'est pas disponible et qui sont dans la même classe d'âge. Finalement, si on n'a ni l'âge, ni la profession, on utilise alors tous les répondants du mois précédent et ce, peu importe leur âge et leur profession. Dans tous les cas, on s'assure de conserver les réponses qui ont pu être données.

Une classe d'imputation contient tous les donneurs potentiels pour un non-répondant. Soit  $n$ , le nombre de donneurs potentiels pour un non-répondant. La combinaison des réponses d'un donneur potentiel aux quatre questions d'éducation correspond à un scénario de réponse. Soit  $k$ , le nombre de scénarios possibles pour un non-répondant. On peut également définir  $n_i$ , avec  $1 \leq i \leq k$ , le nombre de donneurs potentiels pour le scénario  $i$ . On a  $n_1 + n_2 + \dots + n_k = n$ .

À chacun des scénarios, on peut assigner un intervalle. Le premier intervalle s'étend de 1 à  $n_1$ , le deuxième de  $n_1 + 1$  à  $n_1 + n_2$  alors que l'intervalle  $k$  s'étend de  $n_{k-1} + 1$  à  $n$ .

On obtient alors une certaine distribution et on assigne au non-répondant un nombre aléatoire entre 1 et  $n$  en se basant sur cette distribution. On impute au non-répondant les réponses provenant du scénario pour lequel le nombre aléatoire choisi tombe dans l'intervalle correspondant.

## 5. RÉSULTATS

Dans le but de valider l'application de la nouvelle méthode et de comparer les résultats qu'elle produit, cette méthode a été utilisée pour imputer tous les non-répondants aux questions d'éducation pour la période de six mois allant d'avril à septembre 2002. Au total, 2553 individus ont été imputés en utilisant l'imputation déterministe, « carry-forward » (ou historique) ou cold-deck. Pour chacune de ces méthodes d'imputation, on compare la distribution des valeurs imputées par le codeur avec celle du système automatisé (la nouvelle méthode). Pour l'imputation cold-deck, on compare également les distributions des valeurs imputées par le système et le codeur avec la distribution des répondants du mois précédent. L'ensemble des répondants du mois précédent comprend les individus qui ont répondu aux quatre questions d'éducation ou dont les valeurs ont été transférées du mois précédent. On en compte 642 833 pour la période allant d'avril à septembre 2002. Tous les résultats présentés se rapportent à la période de six mois à l'étude.

### 5.1 Imputation déterministe

L'imputation déterministe a permis de traiter 309 individus, soit 12 % des non-répondants entre avril et septembre 2002. Différentes règles ont été bâties à partir de l'hypothèse selon laquelle un répondant qui a étudié au niveau post-secondaire a obtenu son diplôme d'études secondaires. Les règles en question ont permis d'imputer 188 individus. Dans tous les cas, les valeurs imputées en utilisant l'imputation déterministe sont identiques à celles imputées par le codeur. Cette hypothèse mérite donc d'être conservée.

Quant à la seconde hypothèse selon laquelle un répondant qui n'a pas obtenu son diplôme d'études secondaires n'a pas étudié au niveau post-secondaire, elle a permis de bâtir des règles qui ont servi à imputer 121 individus. Le système impute une valeur identique à celle mise par le codeur pour 89 individus (74 %). C'est donc dire que les valeurs imputées par la méthode déterministe diffèrent de celles imputées par le codeur pour 32 individus.

### 5.2 Imputation historique ou « carry-forward »

L'idée d'étendre l'imputation par transfert de données permet de traiter 100 individus entre les mois d'avril et septembre 2002, ce qui représente 4 % des non-répondants au cours de cette période. En étendant le « carry-forward », on émet

l'hypothèse que le niveau d'éducation d'un individu reste inchangé au cours des six mois qu'il participe à l'enquête. On peut noter que les valeurs imputées par la méthode « carry-forward » sont identiques à celles imputées par le codeur pour 70 individus et différentes pour 30 individus. Cependant, tous les résultats auraient pu être identiques puisque le codeur disposait des valeurs du mois précédent et pouvait identifier si les réponses données avaient été fournies par le répondant, transférées par « carry-forward » ou imputées par un codeur.

### 5.3 Imputation cold-deck basée sur l'âge et la profession

Les 2144 enregistrements, soit 84 % des non-répondants, qui n'ont pas été traités par l'une ou l'autre des deux premières méthodes l'ont été par la méthode d'imputation cold-deck. Comme pour les deux autres méthodes, on a comparé les valeurs imputées par le système à celles imputées par le codeur<sup>2</sup>.

Le tableau 1 compare la distribution du plus haut niveau d'éducation complété entre les enregistrements traités par le codeur ainsi que par le système avec celle des répondants du mois précédent.

**Tableau 1 – Distribution du plus haut niveau d'éducation complété pour les enregistrements traités par l'imputation cold-deck basée sur l'âge et la profession**

Plus haut niveau d'éducation complété	Répondants du mois précédent		Valeurs imputées par le codeur		Valeurs imputées par l'imputation cold-deck	
	Nombre	%	Nombre	%	Nombre	%
Secondaire 2 ou moins	69 211	11	39	2	282	13
Secondaire 3 ou 4	78 682	12	295	14	217	10
Secondaire 5 non terminé	43 316	7	140	7	139	6
Secondaire 5 terminé	121 760	19	767	36	529	25
Aucun diplôme post-secondaire	55 049	9	494	23	194	9
Diplôme d'une école de métiers	74 032	12	168	8	251	12
CÉGEP	98 129	15	84	4	276	13
Certificat universitaire	16 037	2	75	3	42	2
Baccalauréat	59 101	9	59	3	161	8
Diplôme supérieur au baccalauréat	27 516	4	23	1	53	2
Total	642 833	100	2 144	100	2 144	100

Comme la détermination des valeurs à imputer repose sur des scénarios observés auprès des répondants du mois précédent, on s'attend à ce que la distribution du plus haut niveau d'éducation complété des enregistrements imputés par la méthode cold-deck soit similaire à celle des répondants du mois précédent. C'est effectivement le cas, à l'exception de ceux ayant terminé un « secondaire 5 » qui sont sur-représentés par le système par rapport à la distribution des répondants du mois précédent (25 % versus 19 %).

Quand on compare la distribution des valeurs assignées par le codeur à celle des répondants du mois précédent, le codeur semble sous-représenter les gens ayant un « secondaire 2 ou moins » (2 % versus 11 %) de même que ceux ayant obtenu un « diplôme au niveau post-secondaire » (11 % versus 30 %). Par ricochet, il y a une nette sur-représentation des gens ayant terminé un « secondaire 5 » et de ceux n'ayant « aucun diplôme post-secondaire ». En effet, 59 % des enregistrements imputés par le codeur tombent dans l'une ou l'autre de ces deux catégories alors que c'était le cas pour 28 % des répondants du mois précédent. Quelques grandes différences sont concentrées dans des groupes d'âge et de profession spécifiques. Par exemple, chez les 65 ans et plus, les gens ayant un « secondaire 2 ou moins » sont sous-représentés par rapport aux répondants du mois précédent (4 % versus 34 %). De même, dans le groupe de profession des

<sup>2</sup> En production, le codeur ne dispose pas de tous les codes de profession car certains codes peuvent être assignés après l'imputation manuelle. Par contre, pour la simulation, on disposait de tous les codes de profession qui avaient été attribués. Ce contexte a probablement favorisé une meilleure imputation par la méthode cold-deck par rapport à celle du codeur pour certains cas. Toutefois, on a démontré que cette méthode produit les résultats escomptés, que le code de profession soit disponible ou non.

affaires, finance et administration, les personnes ayant un « diplôme de niveau baccalauréat et plus » sont sous-estimées par le codeur (3 % versus 15 %).

Puisque le nombre de non-répondants est peu élevé comparé au nombre de répondants pour un mois donné, les valeurs imputées par le codeur et par la méthode cold-deck n'ont globalement aucun impact sur la distribution finale du plus haut niveau d'éducation complété. Peu importe que les non-répondants aient été imputés par le codeur ou le système, les résultats ont au plus un impact de  $\pm 0,2$  % sur la distribution du plus haut niveau d'éducation complété par groupe d'âge.

Parce que l'EPA possède un taux de non-réponse relativement bas, l'étude a démontré que la stratégie d'imputation utilisée (courante ou nouvelle) n'a pas vraiment d'impact sur la distribution finale du plus haut niveau d'éducation, qui inclut les répondants et les non-répondants. Cependant, l'impact pourrait être plus grand à un niveau d'analyse plus détaillé ou avec un taux de non-réponse plus élevé.

On peut finalement noter que la nouvelle méthode permet de traiter plus rapidement les non-répondants aux questions d'éducation. Si l'on suppose que pour un mois donné il y a 400 non-répondants, le système impute tous ces enregistrements par l'une ou l'autre des trois méthodes d'imputation en 20 minutes. Il s'agit d'une nette amélioration par rapport au fait qu'à chaque mois, deux codeurs consacrent près d'une journée pour imputer manuellement le même nombre de non-répondants aux questions d'éducation.

## 6. CONCLUSION

Puisque les données sur l'éducation sont utilisées dans la formation des classes d'imputation dans le système hot-deck servant à imputer des valeurs manquantes de la composante principale de l'EPA, elles ont donc un impact indirect sur la qualité de d'autres résultats imputés, ce qui renchérit l'importance d'avoir une bonne méthode d'imputation.

La nouvelle approche proposée suppose que, étant donné l'âge et la profession, la population des non-répondants possède la même distribution que celle des répondants en ce qui a trait au niveau d'éducation. Basée sur cette hypothèse, les résultats imputés par la nouvelle stratégie d'imputation reflètent la distribution visée. D'un autre côté, cet exercice a clairement démontré qu'il y avait de grandes différences entre la distribution du plus haut niveau d'éducation résultant de l'imputation effectuée par les codeurs et celle des répondants du mois précédent. À moins qu'il y ait une forte conviction qu'on ne devrait pas supposer ce genre de relation entre les non-répondants et les répondants, nous sommes enclins à conclure que l'imputation effectuée par les codeurs est entachée d'un certain biais. À partir des résultats de l'étude, une tendance nette émerge chez les codeurs qui imputent plus de valeurs se trouvant au milieu de la distribution plutôt qu'à chaque extrémité. En plus de l'objectivité et d'une meilleure qualité des données de l'enquête, la nouvelle méthode sera plus efficace puisqu'elle permettra d'épargner sur les ressources allouées aux opérations manuelles et sur le temps de traitement. Finalement, elle peut être reproduite.