# L' imputation des données manquantes, la technique de l'imputation multiple, les conséquences sur l'analyse des données: l'enquête 1999 KOF/ ETHZ sur l'innovation

**Eidgenössische Technische Hochschule Zürich**

*Ecole polytechnique fédérale de Zurich*
*Politecnico federale di Zurigo*
*Swiss Federal Institute of Technology Zurich*

*Konjunkturforschungsstelle*
*Centre de recherches conjoncturelles*
*Centro di ricerche congiunturali*
*Center for Business Cycle Research*

*Weinbergstrasse 35, ETH Zentrum*
*CH-8092 Zürich*
*Phone 01 / 632 42 39*
*Fax 01 / 632 12 18*

Courriel: ***laurent.donze@kof.gess.ethz.ch***
URL: **http://www.dplanet.ch/users/ldonze**

# *L'imputation des données manquantes, la technique de l'imputation multiple, les conséquences sur l'analyse des données : l'enquête 1999 KOF/ETHZ sur l'innovation*

par

## *Laurent Donzé*

Zurich, février 2001

**Résumé**

L'imputation des données manquantes est généralement utilisée pour la correction de la non-réponse partielle d'une enquête. Pour l'enquête 1999 sur l'innovation du KOF/ETHZ, nous utilisons la technique de l'imputation multiple. Nous présentons d'abord quelques aspects théoriques du problème de l'imputation et de la méthode de l'imputation multiple. Nous illustrons cette dernière de plusieurs exemples tirés de l'enquête sur l'innovation. Nous comparons brièvement notre approche avec celle utilisée par Eurostat dans le cadre du « Second Community Innovation Survey » (CIS 2).

**Mots clefs**

Non-réponse partielle, imputation, imputation multiple, enquête sur l'innovation, Community Innovation Survey (CIS)

**Codes JEL**

C42, C81, O31

Exposé présenté au Congrès annuel de la Société suisse d'économie et de statistique, les 15 et 16 mars 2001, à Genève.

# L’imputation des données manquantes, la technique de l’imputation multiple, les conséquences sur l’analyse des données : l’enquête 1999 KOF/ETHZ sur l’innovation.

Tous les trois ans, le Centre de recherches conjoncturelles de l’École polytechnique fédérale de Zurich (KOF) mène auprès des entreprises de son panel une enquête sur les activités d’innovation dans les secteurs suisses de l’industrie, la construction et des services. L’objet de la présente recherche est la correction de la non-réponse partielle apparaissant dans le cadre de cette enquête. On parle de non-réponse partielle (”item nonresponse”) lorsque l’enquêté n’a pas répondu à l’une ou l’autre des questions de l’enquête. Il est de plus en plus fréquent, en effet, de tenir compte et/ou dans la mesure du possible de corriger ce type de non-réponse. Plusieurs techniques ont été développées dans ce but dont le dénominateur commun est l’imputation des données manquantes. En particulier, celle-ci nous permet d’obtenir des séries de données complètes ce qui dans bien des analyses est profitable à plus d’un titre. Revers de la médaille, des techniques d’estimation et d’analyse adéquates doivent alors être utilisées pour exploiter ces nouvelles données.

Dans une première partie théorique, nous montrons quelles sont les conditions d’une bonne imputation. Nous nous devons de définir quelques concepts théoriques : types de données manquantes, modèles informatifs / non informatifs, modèles implicites / explicites. Nous présentons les procédures les plus courantes d’imputation et nous défendons la technique d’imputation multiple. Nous décrivons comment imputer de manière ”appropriée” les valeurs manquantes et opérer avec une base de données résultant d’une imputation multiple. Un point important de l’imputation est la construction de cellules d’ajustement qui respectent l’hypothèse dite MCAR (”missing completely at random”). Nous montrons alors comment cette hypothèse peut être testée.

Nous décrivons dans la seconde partie les étapes de l’imputation des données manquantes de l’enquête en question. Nous montrons en outre comment l’imputation multiple peut facilement être utilisée dans ce genre d’enquête. Quelques exemples illustrent notre propos. Nous comparons brièvement notre façon de faire avec celle d’Eurostat. Notre approche apparaît simple et efficace. Les résultats qui se dégagent confirment la robustesse de la démarche utilisée.

## 1. Quelques notions théoriques

Les principales notions présentées ci-dessous ont été développées par **Rubin** et/ou **Little**. Le lecteur intéressé se référera avec profit aux références données en fin d'article.

### *1.1 Types de données manquantes*

Définissons d'abord quelques éléments. Soit $X = \{x_{ij}\}$, une matrice de données d'ordre $(N \times V)$ d'éléments $x_{ij}$ où $N$ est le nombre d'observations, $V$ le nombre de variables et $x_{ij}$ est la valeur de la variable $j$ pour l'observation $i$, $i = 1,\ldots,N$, $j = 1,\ldots,V$. Soit $M = \{m_{ij}\}$, une matrice indicatrice de données manquantes d'éléments $m_{ij}$, telle que $m_{ij} = 1$ si $x_{ij}$ manque, et $m_{ij} = 0$ sinon. La matrice $M$ décrit la structure des données manquantes. Il est utile de traiter $M$ comme une matrice stochastique.

Souvent, la structure des données manquantes dépend des variables considérées. On dira que les données manquantes sont de type MCAR ("missing completely at random") si le fait de ne pas avoir de valeur est totalement indépendant des variables $X$. Formellement, si $p(M \mid X, f)$ est la distribution conditionnelle de $M$ étant donné $X$, de paramètres $f$ qui caractérisent les taux de réponse, les données manquantes sont de type MCAR si $p(M \mid X, f) = p(M \mid f)$ pour tout $X$.

La caractéristique MCAR est importante puisqu'elle conduit à des estimateurs aux propriétés statistiques intéressantes, par exemple convergents.

Si les données manquantes ne sont pas MCAR, il s'agira de savoir si les différences dans les caractéristiques des non-répondants et des répondants peuvent être expliquées par des variables communes aux répondants et non-répondants. Posons $X_{obs}$ pour la partie observée des données $X$ et $X_{mis}$, la partie manquante. On dira que les données sont de type MAR ("missing at random") si $p(M \mid X_{obs}, X_{mis}, f) = p(M \mid X_{obs}, f)$ pour tout $X_{mis}$, c'est-à-dire que la distribution de $M$ étant donné $X$, ne dépend seulement que des variables $X_{obs}$ enregistrées dans la base de données.

### *1.2 Types de modèles*

Une technique répandue pour gérer la non-réponse partielle est d'imputer une valeur à chaque donnée manquante. Quelques principes doivent cependant être respectés (cf. **Little** et **Schenker** (1995) par exemple). Premièrement, les imputations devraient être basées sur la distribution prédictive des valeurs manquantes étant donné les valeurs observées. Idéalement, tous les items observés devraient être pris en compte en créant les imputations. En général, cela s'avère impossible. Mais la connaissance de la matière permet de sélectionner fort heureusement les variables. Un deuxième principe important de l'imputation est d'utiliser des tirages aléatoires et non pas les prévisions les meilleures, cela afin de préserver la distribution des variables dans la base complétée des données.

Les procédures d'imputation peuvent être basées sur des modèles qui peuvent être "explicites", "implicites" ou une combinaison des deux. Les modèles explicites sont issus en général de la théorie statistique : régression linéaire, modèle linéaire généralisé, etc. Par

exemple, une imputation stochastique par régression, utilise un modèle explicite. Les modèles implicites sont ceux qu'on retrouve dans les procédures permettant de régler pratiquement des problèmes de structures de données. Ils sont souvent de type non paramétrique. À titre d'exemple, on peut citer les procédures "hot-deck" qui reposent sur une modélisation implicite.

Une distinction supplémentaire est également utile : les modèles sur lesquels reposent les méthodes d'imputation, qu'ils soient implicites ou explicites, peuvent en outre être classés en modèles "informatifs" ou "non informatifs" ("ignorable / nonignorable model"). On dira qu'un modèle est non informatif ("ignorable model") si une valeur $X_V$ d'un non-répondant est seulement stochastiquement différente de celle d'un répondant qui a les mêmes valeurs $X_1,\ldots,X_{V-1}$. Un modèle est dit par contre informatif ("nonignorable model") si on admet que même si un répondant et un non-répondant sont identiques par rapport à $X_1,\ldots,X_{V-1}$, leurs valeurs $X_V$ diffèrent systématiquement. Autant les modèles informatifs que non informatifs peuvent être utilisés comme base d'une procédure d'imputation. Il faut cependant souligner qu'il n'existe pas pour un modèle informatif d'évidences directes dans les données pour certifier la validité des hypothèses sous-jacentes. Il est donc conseillé de considérer plusieurs modèles informatifs et d'en inspecter la sensibilité.

*1.3 Les techniques d'imputation*

L'imputation des données manquantes n'est pas nouvelle et il existe à ce jour une foule de techniques développées à cet effet (cf. par exemple la revue de **Caron**). Le grand avantage de l'imputation est qu'elle permet de créer des bases de données complètes. Cependant, cela n'est pas sans conséquence dans le calcul de certains estimateurs, en particulier de la variance. Chaque technique d'imputation conduit à une formule de variance ainsi qu'à une estimation de variance particulière.

On distingue les groupes de méthodes suivants :

- les méthodes déductives : la donnée manquante est déduite des réponses aux autres questions ;
- les méthodes de type "cold-deck" : elles utilisent l'information obtenue à partir des répondants d'une autre enquête ;
- les méthodes utilisant la prévision par un modèle de régression ;
- les méthodes de type "hot-deck" : elles donnent pour la valeur manquante la valeur observée d'un individu répondant (=le donneur) choisi selon une procédure adéquate. Il existe différentes procédures connues, et régulièrement utilisées en pratique, de choix des données ; citons par exemple le hot-deck aléatoire, séquentiel hiérarchisé, métrique, etc.

*1.4 L'imputation multiple*

La correction de la non-réponse partielle par une imputation unique, à savoir pour chaque valeur manquante imputer une seule valeur, présente un défaut majeur. En effet, du fait qu'une unique valeur imputée ne peut pas représenter toute l'incertitude à propos de la valeur à imputer, les analyses qui considèrent les valeurs imputées de manière équivalente aux valeurs observées en général sous-estiment l'incertitude, même si la non-réponse est correctement modélisée et des imputations aléatoires sont générées. Cet handicap, s'il n'est

pas correctement appréhendé, peut conduire entre autres à des variances nettement sous-estimées.

La technique de l'imputation multiple développée principalement par **Rubin**, peut remédier à ces inconvénients. Elle remplace chaque donnée manquante par deux valeurs ou plus tirées d'une distribution appropriée pour les valeurs manquantes sous les hypothèses postulées à propos de la non-réponse. Le résultat est deux bases de données ou plus. Chacune est analysée en utilisant une même méthode standard. Les analyses sont ensuite combinées de telle manière à refléter la variabilité supplémentaire due aux données manquantes. Un autre avantage de l'imputation multiple est qu'elle permet de trouver des estimateurs ponctuels plus efficaces. Enfin, d'un point de vue théorique, on peut motiver la méthode d'imputation multiple par une approche bayesienne.

Plus le nombre $K$ d'imputations est grand, plus les estimateurs seront précis. Cependant, en pratique, on constate qu'on a de bons résultats déjà à partir d'un petit nombre d'imputations (par exemple $K = 5$).

*1.5 Procédures d'imputation multiple "appropriée" ("proper imputation")*

On dit d'une procédure d'imputation multiple, basée ou non sur des modèles implicites ou explicites, ou sur des modèles informatifs ou non informatifs, qu'elle est "appropriée" si elle incorpore la variabilité adéquate parmi les $K$ ensembles d'imputations dans le cadre d'un modèle (cf. définition exacte in **Rubin** (1987), chap. 4). Une procédure d'imputation "appropriée" garanti des inférences qui soient valables.

Toutes les procédures ne sont pas "appropriées". La procédure d'imputation ABB ("Approximate Bayesian Bootstrap") l'est. C'est une procédure de type "hot-deck" qui incorpore les idées des méthodes bootstrap.

*1.6 La procédure d'imputation ABB*

On peut décrire la procédure d'imputation ABB de la manière suivante. Soit une collection de $n$ unités de même valeurs $X_1,\ldots,X_{V-1}$, où l'on trouve pour la variable $X_V$, $n_R$ répondants et $n_{NR} = n - n_R$ non-répondants. Pour chacun des $K$ ensembles d'imputations, on tire aléatoirement avec remise dans l'ensemble des répondants d'abord $n$ valeurs possibles de $X_V$. Ensuite, on impute les $n_{NR}$ valeurs manquantes de $X_V$ en tirant aléatoirement avec remise de l'ensemble des $n$ valeurs possibles.

Notons que le tirage de $n_{NR}$ valeurs manquantes d'un échantillon possible de $n$ valeurs plutôt que de l'échantillon observé de $n_R$ valeurs génère la variabilité appropriée entre les imputations, du moins en supposant de grands échantillons aléatoires simples. Ensuite, supposer une collection de $n$ unités de mêmes valeurs $X_1,\ldots,X_{V-1}$, permet de classer les répondants et les non-répondants dans un même ensemble homogène. Nous verrons ci-dessous en traitant de la construction de cellules d'ajustement (ou d'imputation), comment on peut y parvenir.

*1.7 Estimation ponctuelle et inférence dans un système d'imputation multiple*

Nous allons présenter le calcul d'un estimateur et de sa variance dans le cadre d'un système d'imputation multiple, tout d'abord en envisageant le cas scalaire puis le cas vectoriel.

a) Le cas scalaire

Soient $\hat{q}_m$, $\hat{U}_m$, $m = 1,\ldots,M$, les $M$ estimateurs respectivement du paramètre $q$ et de sa variance $U$, calculés à partir de chacun des $M$ ensembles de données complétés par imputation. L'estimateur final $\bar{q}$ de $q$ s'écrit :

$$\bar{q} = M^{-1}\sum_{m=1}^{M}\hat{q}_m ,$$

à savoir comme la moyenne des $M$ estimateurs obtenus à partir des données complétées.

Soient également $\overline{U}$, la moyenne des $M$ variances, et $B$, la variance des estimateurs calculés. On a :

$$\overline{U} = M^{-1}\sum_{m=1}^{M}\hat{U}_m ,$$

et

$$B = (M-1)^{-1}\sum_{m=1}^{M}\left(\hat{q}_m - \bar{q}\right)^2 .$$

La variance totale $T$ de $(q - \bar{q})$ est donnée par la somme de la composante intra-imputation $\overline{U}$ et de la composante inter-imputation multipliée par le facteur de correction de petit échantillon $\left(1 + M^{-1}\right)$ :

$$T = \overline{U} + \left(1 + M^{-1}\right)B .$$

Les estimations d'intervalles et de niveaux de signification sont obtenus en utilisant une distribution $t$ centrée sur $\bar{q}$, de facteur $T^{1/2}$ et de degrés de liberté $n = (M-1)\left(1 + r^{-1}\right)^2$, où $r = (1 + M^{-1})B/\overline{U}$. On a $(q - \bar{q})T^{-1/2} \sim t_n$. En général, $r$ estime la quantité $g/(1-g)$, où $g$ est la fraction de l'information au sujet de $q$ qui est due à la non-réponse.

b) Le cas vectoriel

Le paramètre estimé $\hat{q}_m$ est un vecteur. $\hat{U}_m$ est sa matrice de variances-covariances estimée. Les expressions développées ci-dessus pour $\bar{q}$, $\overline{U}$ et $T$ sont identiques au cas scalaire, tandis que $B$ s'écrit comme :

$$B = (M-1)^{-1}\sum_{m=1}^{M}(\hat{q}_m - \bar{q})(\hat{q}_m - \bar{q})^T .$$

On peut calculer une statistique $D$ de type Wald comme :

$$D = \frac{\bar{q}^T \bar{U}^{-1} \bar{q}}{k(1+r)},$$

où $r = (1+M^{-1})tr(B\bar{U}^{-1})/k$, $k$ est le nombre d'éléments du vecteur $q$, $r$ est le ratio moyen estimé inverse de la fraction de l'information manquante. La statistique $D$ permet de tester l'hypothèse nulle globale $q = 0$ et suit approximativement une loi de Fischer $F_{k,w}$ avec $w$ choisi comme :

$$w = \begin{cases} 4 + (v-4)\left\{1 + (1-\frac{1}{2}v)r^{-1}\right\}^2 & \text{si} \quad v = k(M-1) > 4; \\ \frac{1}{2}v(1+k^{-1})(1+r^{-1})^2 & \text{sinon.} \end{cases}$$

*1.8 La construction de cellules d'ajustement*

La procédure d'imputation ABB décrite ci-dessus a pour principe d'attribuer à une donnée manquante une valeur observée chez un répondant. Il s'agit donc de trouver parmi les répondants quels sont les donneurs potentiels. Une façon simple de procéder est de classer les observations en groupes homogènes. On donnera à un non-répondant la donnée d'un répondant appartenant au même groupe. Les classes regroupant les observations d'un même groupe sont appelées cellules d'ajustement ou d'imputation.

Nous nous intéressons à la construction de ces cellules d'ajustement. On peut tout simplement utiliser par exemple les variables de stratification, ou effectuer des croisements de différentes variables de classes (sexe, classes d'âge, etc.). Dans ce dernier cas, plus on effectue de croisements, "meilleures" sont les cellules d'ajustement, mais moins nombreux sont les donneurs potentiels. Une autre méthode, simple de conception également, de construction de ces cellules d'ajustement, est basée sur les "propensity scores" (cf. **Rosenbaum** et **Rubin** (1983, 1985)). Cette méthode offre le grand avantage qu'elle génère des cellules d'ajustement dans lesquelles les données manquantes sont en principe MCAR, autorisant ainsi l'emploi de modèles non informatifs d'imputation.

La mise en œuvre de la méthode consiste à modéliser la probabilité de répondre en fonction d'un certain nombre de caractéristiques. On estime le modèle par une équation de régression de type probit ou logit. Pour chaque observation, on calcule sur la base du modèle estimé la probabilité de répondre. On regroupe ensuite les probabilités obtenues en 5 à 6 classes, par exemple selon les quintiles. Ces classes forment les cellules d'ajustement.

*1.9 Le test de l'hypothèse MCAR*

Plusieurs analyses statistiques de données avec des valeurs manquantes font l'hypothèse que les données manquent de manière complètement aléatoire (MCAR). Il n'existe que très peu de tests de cette hypothèse. **Little** (1988) propose un test statistique global de cette hypothèse et montre que sous l'hypothèse nulle la distribution de la statistique développée suit une loi $c^2$. Il faut remarquer que l'hypothèse testée (MCAR), est plus forte que l'hypothèse MAR. Le test proposé par **Little** (1988) est asymptotiquement valable sous les hypothèses de normalité. Cependant, le test semble plus approprié quand les variables sont de type quantitatif.

## 2. L'enquête 1999 KOF/ETHZ sur l'innovation

### *2.1 Généralités*

Avant d'entreprendre l'imputation des données manquantes, les données des questionnaires collectés subissent un intense travail de vérification et de préparation (tests de cohérence logique, suppression des valeurs aberrantes, correction ou ajout de certaines valeurs, etc.). Une part importante de ce travail préliminaire est la mise en œuvre d'une pondération adéquate destinée à corriger la non-réponse globale. Nous devons souligner, à propos de pondération, que tous les travaux d'imputation que nous décrivons ci-après sont effectués sans pondération.

L'imputation d'une variable ou d'un groupe de variables (cf. ci-dessous) suit les deux étapes suivantes. D'abord, la construction de cellules d'ajustement. Ensuite, l'imputation des données manquantes par la méthode ABB. On répète la seconde étape 5 fois de manière à constituer pour une variable $X$ à valeurs manquantes, 5 variables $XI1,\ldots,XI5$ à valeurs complètes.

Bien souvent, dans des enquêtes complètes, les variables relatives à une question doivent être traîtées non pas séparément mais simultanément. C'est le cas notamment lorsqu'on interroge à propos de part relative de certaines composantes ou lorsque l'enquêté a le choix de répondre à l'une ou l'autre des questions d'un thème précis. Par exemple, dans le questionnaire destiné à l'industrie, la question 2.2 porte sur la part du chiffre d'affaires consacré aux produits nouveaux, aux produits sensiblement améliorés et aux produits pas ou peu améliorés, le total de ces trois parts étant de 100%. Les variables correspondant à ces trois parts se réfèrent à un groupe de questions et seront imputées en bloc (cf. ci-dessous).

### *2.2 Les cellules d'ajustement*

Le principe adopté dans la construction des cellules d'ajustement est le suivant. Dans la mesure du possible, on utilise la méthode des "propensity scores". Si la modélisation de la probabilité de la non-réponse n'est pas convaincante (ajustement insatisfaisant du modèle, variables peu significatives, etc.), on utilise la variable de stratification de l'enquête. Celle-ci est construite à partir des variables de tailles de l'entreprise (3 classes) et de branches économiques (28 classes). Si le nombre de valeurs observées par strate sur la variable à imputer est insuffisant, on procède au regroupement de deux ou plusieurs strates contiguës, c'est-à-dire au contenu plus ou moins similaire.

Pour modéliser la probabilité de réponse à une question $x$ (variable $X$), on construit une variable indicatrice $IX$ qui prend la valeur 1 si $X$ a une valeur observée, 0 sinon. Cette probabilité, à savoir $P[IX = 1]$, est alors estimée par un modèle de régression de type logit. Les variables indépendantes du modèle sont les variables structurelles de l'entreprise. Celles-ci, des variables qualitatives, sont sélectionnées par une procédure d'élimination "backward". Le choix de ces variables est fort limité. Nous n'avons en effet que peu de possibilités de trouver d'autres variables, sans observations manquantes, qui soient pertinentes pour la modélisation de cette probabilité. Nous sommes en mesure de construire les variables structurelles suivantes :

- 3 variables de tailles de l'entreprise :

  GR_K : petite ;
  GR_M : moyenne ;
  GR_G : grande ;

- 8 variables de branches économiques :

  IND_1 : alimentation, textile, habillement, bois ;
  IND_2 : papier, graphisme, chimie, matières plastiques et caoutchouc, produits minéraux non métalliques ;
  IND_3 : métallurgie, travaux des métaux, industrie des machines ;
  IND_4 : électrotechnique, électronique, horlogerie, véhicules ;
  IND_5 : autres industries, énergie ;
  BAU_1 : construction ;
  DL_1 : commerce de gros, commerce de détail, hôtellerie, services personnels ;
  DL_2 : transport et télécommunications, banques et assurances, immobilier et location, informatique et R&D ;

- 7 variables régionales :

  REG_1 : région lémanique ;
  REG_2 : espace Mittelland ;
  REG_3 : Suisse du Nord-Ouest ;
  REG_4 : Zürich ;
  REG_5 : Suisse orientale ;
  REG_6 : Suisse centrale ;
  REG_7 : Tessin.

Ces variables binaires prennent la valeur 1 si la propriété est rencontrée, 0 sinon. Il est clair que toutes les variables ne peuvent figurer à la fois dans l'équation de régression (problème de multicollinéarité).

Lorsque nous avons affaire à un groupe de questions (ou de variables), nous construisons les cellules d'ajustement de la manière suivante. Si le groupe de questions est formé par exemple des variables $X1$, $X2$ et $X3$, nous générons la variable $IX$ qui prendra la valeur 0 si simultanément les valeurs de $X1$, $X2$ et $X3$ sont manquantes, et 1 sinon. Si $IX = 0$, on jugera que l'enquêté n'a pas répondu. La probabilité de réponse est ensuite estimée comme dans le cas d'une seule variable.

*2.3 Contrôles préliminaires à l'imputation*

Un contrôle préliminaire à l'imputation doit porter sur le nombre de valeurs observées et manquantes par cellule d'ajustement ainsi que sur la vérification de l'hypothèse MCAR. A propos de ce dernier test, nous adoptons la démarche suivante. Nous avons vu qu'il s'agit d'un test global. Ce test porte donc sur un ensemble de variables. Celui-ci est composé de variables de base auxquelles nous ajoutons la variable à imputer ou la variable générée dans le cas de l'imputation d'un bloc de variables avec la valeur 0 redéfinie comme valeur manquante. Les variables de bases que nous avons choisies sont une partie des variables quantitatives du début du questionnaire, à savoir :

- F14a : nombre d'employés en Suisse à fin 1998 ;
- F16a : chiffre d'affaires de l'entreprise réalisé par le site suisse en 1998 ;

- F17b : part des exportations au chiffre d'affaires de 1998 ;
- F18 : part des frais de personnel au chiffre d'affaires de 1998 ;
- F19 : part de la consommation intermédiaire au chiffre d'affaires de 1998.

*2.4 L'imputation proprement dite*

La procédure d'imputation est effectuée pour chaque cellule d'imputation. Le cas des groupes de variables est traité comme suit. On assigne aux variables du groupe désigné comme manquant, les valeurs des variables d'un groupe reconnu non manquant. Par exemple, pour la question 2.2 du questionnaire destiné à l'industrie, si les variables F22in_1, F22in_2 et F22in_3 ont toutes l'observation $i$ manquante, on trouvera une observation $j$ telle que pour cette observation F22in_1 ou F22in_2 ou F22in_3 aient une valeur.

*2.5 Deux exemples d'imputation*

A titre d'exemple, nous montrons comment nous avons imputé la variable F26A_1 et le groupe de variables constitué des variables F42_1, F42_2, F42_3, F42_4 de la question 4.2 de l'enquête. Ces questions étant les mêmes pour les trois secteurs (industrie, construction, services), nous travaillons à partir de la totalité des questionnaires rentrés. L'influence éventuelle de la branche économique interviendra dans la modélisation de la probabilité de réponse à ces questions. Par contre, ces questions n'étant posées qu'aux entreprises se déclarant innovatrices, nous nous restreignons bien évidemment à cet ensemble de firmes.

a) Imputation de la variable F26A_1

La variable F26A_1 concerne la question 2.6 du questionnaire. Celle-ci s'énonce ainsi : « *Si votre entreprise a introduit des innovations-produits pendant la période 1997-1999, veuillez en évaluer l'importance (sur une échelle de 1 à 5 ; 1 = très faible et 5 = très grande) en ce qui concerne l'état de la technique ?* ». La variable F26A_1 prend les valeurs 1, 2, …,5. Sur les 1355 observations (= nombre d'innovateurs), 337 firmes n'ont pas répondu à la question (=25%). La répartition des valeurs manquantes par strate est inégale. Certaines strates n'ont aucune valeur manquante tandis que d'autres, au contraire, ont toutes les données manquantes. Choisir les strates comme cellules d'ajustement ne serait pas adéquat. Nous nous aiderons donc de la méthode des "propensity scores".

Soit IF26A_1 la variable qui prend la valeur 1 si F26A_1 a une valeur et 0 sinon. Nous pouvons estimer de manière satisfaisante la probabilité de réponse ($P[\text{IF26A\_1} = 1]$) par logit. Le tableau N°1 résume l'estimation du modèle.

**Tableau N°1 :**
**Modélisation par logit de la probabilité de réponse à la question 2.6**
**(Variable dépendante : IF26A_1)**

| Variables du modèle | Paramètres estimés | Ecarts-types |
|---|---|---|
| **Constante** | 0.7737** | 0.0849 |
| **IND_1** | 0.7698** | 0.2490 |
| **IND_3** | 0.6676** | 0.1796 |
| **IND_4** | 0.9298** | 0.2282 |
| **IND_5** | 0.8356* | 0.3961 |
| **REG_3** | 0.4156* | 0.1934 |
| **REG_7** | -0.5686* | 0.2740 |
| **Nbr. obs.** | 1355 | |
| **-2 Log ?** | 43.404 | |

*Notes :* *1)* "Nbr. obs" est le nombre d'observations, "-2 Log ?" est la statistique du quotient de vraisemblance pour tester la dépendance globale.

*2)* "**" significatif à 1%, "*" significatif à 5%.

L'estimation des probabilités nous permet de construire 4 cellules d'ajustement. Le Tableau N°2 présente les principales caractéristiques de ces cellules.

**Tableau N°2 :**
**Caractéristiques des cellules d'ajustement pour l'imputation de la variable F26A_1**

| N° de cellule | Observations totales | Observations non manquantes | Observations manquantes | $d^2$ |
|---|---|---|---|---|
| **1** | 633 | 429 | 204 | $21.18 < c^2_{0.95;82}$ |
| **2** | 366 | 295 | 71 | $12.89 < c^2_{0.95;69}$ |
| **3** | 102 | 83 | 19 | $22.60 < c^2_{0.95;38}$ |
| **4** | 254 | 211 | 43 | $18.89 < c^2_{0.95;60}$ |
| **Total** | 1355 | 1018 | 337 | |

*Notes* : On rejette l'hypothèse MCAR si la statistique $d^2 > c^2_{(1-a);df}$ , où $c^2_{(1-a);df}$ est une loi de chi-carré à $df$ degrés de liberté et seuil de signification $a$ . Pour le calcul de la statistique $d^2$ et de $df$ se référer à **Little** (1988).

On constate que pour les 4 cellules d'ajustement, l'hypothèse MCAR ne peut pas être rejetée. D'autre part, il y a suffisamment de "donneurs" par cellules pour procéder à l'imputation. La méthode ABB sera donc appliquée pour les observations de chacune des quatre cellules d'ajustement.

a) Imputation des variables de la question 4.2

La question 4.2 est la suivante : « *Quelle est l'importance (sur une échelle de 1 à 5 ; 1 = aucune et 5 = très grande) des dépenses liées à l'innovation pour des investissements dans 1) des machines et moyens spécialisés (F42_1) ; 2) l'acquisition de savoir externe (licences, etc.) (F42_2) ; 3) la formation / le perfectionnement des collaborateurs (F42_3) ; 4) la mise sur le marché de produits nouveaux ou sensiblement améliorés (F42_4) ?* ». Nous construisons d'abord la variable IF42 qui prend la valeur 0 si F42_1, …, F42_4 n'ont pas de valeurs, 1

sinon. La variable IF42 nous montre qu'il y a 147 (=10.8%) non-réponses à la question 4.2 sur 1355 observations. La répartition par strate est inégalitaire. Nous procédons donc comme auparavant par la méthode des "propensity scores". Le tableau N°3 résume l'estimation du modèle.

**Tableau N°3 :**
**Modélisation par logit de la probabilité de réponse à la question 4.2**
**(Variable dépendante : IF42)**

| Variables du modèle | Paramètres estimés | Ecarts-types |
|---|---|---|
| **Constante** | 2.0238** | 0.1249 |
| **IND_4** | 0.7745* | 0.3785 |
| **DL_1** | -0.7551** | 0.2014 |
| **GR_M** | 0.4494* | 0.1851 |
| **Nbr. obs.** | 1355 | |
| **-2 Log ?** | 28.128 | |

*Notes :* *1)* "Nbr. obs" est le nombre d'observations, "-2 Log ?" est la statistique du quotient de vraisemblance pour tester la dépendance globale.

*3)* "**" significatif à 1%, "*" significatif à 5%.

Nous pouvons construire trois cellules d'ajustement. Le Tableau N°4 présente les principales caractéristiques de ces cellules.

**Tableau N°4 :**
**Caractéristiques des cellules d'ajustement pour l'imputation des variables F42_1, F42_2, F42_3 et F42_4**

| N° de cellule | Observations totales | Observations non manquantes | Observations manquantes | $d^2$ |
|---|---|---|---|---|
| **1** | 776 | 666 | 110 | $28.09 < c^2_{0.95;76}$ |
| **2** | 409 | 380 | 29 | $17.10 < c^2_{0.95;66}$ |
| **3** | 170 | 162 | 8 | $5.17 < c^2_{0.95;40}$ |
| **Total** | 1355 | 1208 | 147 | |

*Notes* : On rejette l'hypothèse MCAR si la statistique $d^2 > c^2_{(1-a);df}$ , où $c^2_{(1-a);df}$ est une loi de chi-carré à $df$ degrés de liberté et seuil de signification $a$ . Pour le calcul de la statistique $d^2$ et de $df$ se référer à **Little** (1988).

La lecture du Tableau N°4 nous enseigne que nous pouvons procéder sans crainte à l'imputation des données selon la démarche que nous avons énoncée.

*2.6 Exemples d'estimation à partir d'une imputation multiple*

Comme nous l'avons décrit ci-dessus, la technique de l'imputation multiple revient à imputer un certain nombre de fois les variables manquantes d'une variable. Nous prenons comme règle, qui en pratique s'avère judicieuse, de générer 5 imputations. Nous répéterons donc les analyses avec chacune des 5 bases de données complétées. Illustrons cela d'abord par le cas de l'estimation des fréquences relatives d'une variable, puis par l'estimation d'un modèle décrivant les facteurs de l'innovation-produit.

a) Fréquences relatives de la variable F26A_1

Dans le cas de la variable F26A_1 (cf. ci-dessus), nous avons construit les 5 variables suivantes: F26A_1I1, F26A_1I2, …, F26A_1I5. Nous calculons les fréquences relatives par classe (5 valeurs possibles) de ces variables. Le résultat final est donné par la moyenne par classe de ces 5 variables. Le Tableau N°5 résume les résultats. La ligne "F26A_1 imputé" est l'estimation finale de la variable F26A_1 après imputation. On remarquera la relative stabilité des différentes imputations ainsi que du résultat final par rapport à la variable initiale.

**Tableau N°5 :**
**Fréquences relatives (en %) de la variable F26A_1, des variables imputées ainsi que de l'estimation finale après imputation.**

| Variables | Classe 1 | Classe 2 | Classe 3 | Classe 4 | Classe 5 | Total | Nbr. obs. |
|---|---|---|---|---|---|---|---|
| **F26A_1** | 4.90 | 12.50 | 33.00 | 37.60 | 12.00 | 100.00 | 1018 |
| **F26A_1I1** | 5.50 | 12.30 | 32.30 | 37.90 | 12.10 | 100.00 | 1355 |
| **F26A_1I2** | 4.90 | 12.40 | 32.90 | 37.80 | 12.00 | 100.00 | 1355 |
| **F26A_1I3** | 4.90 | 13.00 | 33.10 | 37.50 | 11.50 | 100.00 | 1355 |
| **F26A_1I4** | 5.10 | 11.80 | 33.70 | 36.80 | 12.50 | 100.00 | 1355 |
| **F26A_1I5** | 5.80 | 12.30 | 32.50 | 37.10 | 12.30 | 100.00 | 1355 |
| **F26A_1 imputé** | 5.23 | 12.40 | 32.90 | 37.40 | 12.10 | 100.00 | 1355 |

b) Modélisation de l'innovation-produit pour l'industrie

L'une des graves conséquences de la non-réponse partielle est de perdre un grand nombre d'observations si nous tentons de combiner plusieurs variables. C'est le cas notamment des tableaux croisés de variables ou lors d'estimations de modèles de régression. L'imputation permet de combler cette lacune. Prenons l'exemple de la modélisation de l'innovation-produit dans le secteur économique de l'industrie. Sans commenter plus à fond la spécification du modèle, mentionnons simplement que nous estimons par un modèle de type logit l'introduction ou non par la firme d'une innovation de type produit. Un certain nombre de variables caractérisent le modèle de base (cf. Tableau N° 6). Pour la spécification du modèle figurant dans le Tableau N° 6, nous pouvons compter sur 705 observations sur un total de 1049 si nous travaillons avec les données non imputées, soit une perte de quelque 33% des données. Cela est grandement dommageable tant du point de vue de la fiabilité des résultats que de la pertinence de l'analyse. Le Tableau N° 6 nous donne les estimations du modèle obtenues sans et avec imputations. Le résultat final des 5 imputations (estimation des paramètres et leurs écarts-types) est calculé selon les formules présentées en section 1.7.

**Tableau N° 6:**
**Modélisation par logit de l'innovation-produit dans l'industrie sans et avec imputations (Variable dépendante : INNOPD)**

| Variables du modèle | Sans imputation | Avec imputation N° 1 | Avec imputation N° 2 | Avec imputation N° 3 | Avec imputation N° 4 | Avec imputation N° 5 | Résultat final des 5 imputations |
|---|---|---|---|---|---|---|---|
| **Constante** | -3.1876** *0.6858* | -2.8179** *0.5059* | -2.8363** *0.5046* | -2.5864** *0.5017* | -2.7431** *0.5017* | -2.8204** *0.5027* | -2.7608** *0.5160* |
| **NFPERS** | 0.2285* *0.1124* | 0.2082* *0.0855* | 0.1913* *0.0854* | 0.2176* *0.0852* | 0.1994* *0.0858* | 0.2046* *0.0848* | 0.2042* *0.0860* |
| **PREIK** | -0.0097 *0.0860* | 0.0636 *0.0653* | 0.0830 *0.0647* | 0.0343 *0.0657* | 0.0181 *0.0647* | 0.0202 *0.0646* | 0.0438 *0.0721* |
| **NPREIK** | 0.1143 *0.0936* | 0.0389 *0.0685* | 0.0434 *0.0686* | 0.0591 *0.0695* | 0.0788 *0.0682* | 0.0165 *0.0682* | 0.0473 *0.0732* |
| **AKONK2** | -0.0459 *0.2371* | 0.1210 *0.1822* | 0.1275 *0.1818* | 0.1953 *0.1795* | 0.0848 *0.1829* | 0.1641 *0.1802* | 0.1385 *0.1872* |
| **AKONK3** | -0.0929 *0.3393* | -0.2543 *0.2440* | -0.2074 *0.2465* | 0.0325 *0.2498* | -0.1503 *0.2471* | 0.0134 *0.2489* | -0.1132 *0.2852* |
| **AKONK4** | -0.6179* *0.3055* | -0.6537** *0.2217* | -0.7292** *0.2213* | -0.5567* *0.2179* | -0.6230** *0.2237* | -0.5738** *0.2204* | -0.6273** *0.2335* |
| **ISCHPD** | 0.2676** *0.0876* | 0.1540* *0.0627* | 0.1796** *0.0621* | 0.1372* *0.0635* | 0.1520* *0.0631* | 0.2465** *0.0636* | 0.1739* *0.0789* |
| **TECHP** | 0.2915** *0.0950* | 0.2390** *0.0680* | 0.1916** *0.0670* | 0.1590* *0.0680* | 0.1587* *0.0671* | 0.1755** *0.0680* | 0.1848* *0.0768* |
| **WQ1PD** | 0.3547** *0.0874* | 0.3222** *0.0642* | 0.3027** *0.0637* | 0.3037** *0.0639* | 0.3474** *0.0640* | 0.3275** *0.0641* | 0.3207** *0.0671* |
| **WQ4PD** | -0.1282 *0.1146* | -0.0576 *0.0838* | -0.0296 *0.0832* | -0.1192 *0.0834* | -0.0797 *0.0832* | -0.0169 *0.0828* | -0.0606 *0.0945* |
| **WQ5PD** | -0.3004** *0.0955* | -0.1501* *0.0702* | -0.1682* *0.0694* | -0.1663* *0.0693* | -0.1642* *0.0706* | -0.1569* *0.0699* | -0.1612* *0.0704* |
| **WQ6PD** | 0.0847 *0.0694* | 0.0453 *0.0537* | 0.0617 *0.0528* | 0.0592 *0.0531* | 0.0178 *0.0532* | 0.0530 *0.0530* | 0.0474 *0.0566* |
| **WQ7PD** | -0.0167 *0.0848* | -0.0664 *0.0646* | -0.0217 *0.0651* | -0.0332 *0.0655* | -0.0044 *0.0648* | -0.0405 *0.0648* | -0.0332 *0.0697* |
| **WQ10PD** | 0.3821** *0.1110* | 0.2594** *0.0836* | 0.2466** *0.0856* | 0.3353** *0.0868* | 0.3051** *0.0840* | 0.2150** *0.0832* | 0.2723** *0.0995* |
| **WQ11PD** | 0.2228* *0.0927* | 0.1595* *0.0670* | 0.1570* *0.0672* | 0.1692* *0.0658* | 0.1942** *0.0666* | 0.1729** *0.0671* | 0.1706* *0.0687* |
| **Nbr. obs.** | 705 | 1049 | 1049 | 1049 | 1049 | 1049 | |
| **-2 Log ?** | 148.692** | 160.205** | 160.618** | 156.708** | 163.198** | 157.848** | |
| ***D*** | | | | | | | 9.5006** |

*Notes :*

1) La variable dépendante du modèle est INNOPD qui prend la valeur 1 si la firme a introduit une innovation-produit et 0 sinon. Les variables indépendantes sont NFPERS (évolution de la demande), PREIK (concurrence par les prix), NPREIK (concurrence autre que les prix), AKONK2 à AKONK4 (nbr. de concurrents), ISHCPD (efficacité de la protection des avantages concurrentiels pour les innovations-produits), TECHP (potentiel technologique), WQ1PD à WQ11PD (sources externes de connaissance : clients, fournisseurs, concurrents, entreprises du même groupe, universités et hautes écoles, brevets, foires et expositions);

2) "Nbr. obs" est le nombre d'observations, "-2 Log ?" est la statistique du quotient de vraisemblance pour tester la dépendance globale et "D" est la statistique de type Wald également pour tester la dépendance globale dans le cas d'une imputation multiple;

3) L'écart-type du coefficient est indiqué en italique sous ce dernier;

4) "**" significatif à 1%, "*" significatif à 5%.

## 3. Comparaison avec Eurostat

Dans le cadre du programme CIS 2 (The Second Community Innovation Survey), **Eurostat** a développé également sa propre méthodologie d'imputation des valeurs manquantes. Quoique ne participant pas directement au programme CIS, nous avons été fortement intéressé par ce qu'Eurostat proposait pour une enquête similaire à la nôtre. Nous décrivons brièvement ci-après leur façon de faire.

Eurostat préconise pour la correction de la non-réponse partielle de son questionnaire, une imputation unique des valeurs manquantes par une méthode de type "hot-deck séquentiel" en utilisant le concept d'entropie. Le principe de la méthode est de classer dans un certain ordre les observations (grâce au calcul de l'entropie). On attribue alors à chaque valeur manquante la valeur du répondant qui la précède. L'imputation est effectuée par cellule d'ajustement, définie en l'occurrence simplement par la branche économique et la taille de l'entreprise.

L'entropie, ou l'information espérée d'une distribution (cf. par exemple **Theil** (1967)), peut être considérée comme une mesure de désordre et donc aussi comme une mesure d'incertitude. La difficulté principale consiste à calculer l'entropie pour deux observations. La mesure de l'entropie utilisée est une construction ad hoc qui dépend du type de variables rencontrées. Les variables métriques sont converties en variables ordinales. L'entropie est calculée par rapport à un vecteur de variables et pour deux observations, celle ou la valeur de la variable à imputer est manquante et une autre ou la valeur est observée. On assignera à la valeur manquante, la valeur de l'observation conduisant à l'entropie minimale.

La méthode proposée par Eurostat est assez laborieuse. Elle a comme principal défaut que cette procédure non aléatoire conduit à l'utilisation répétée du même donneur. En outre, elle a le défaut de toutes les méthodes basées sur une imputation unique, à savoir qu'elle sous-estime la variabilité dans les imputations. Enfin, soulignons que le choix des variables participant au calcul de l'entropie peut être assez arbitraire.

## 4. Conclusion

L'imputation de données manquantes n'est pas une affaire banale. On constate que les méthodes sont nombreuses et qu'il n'existe pas de recettes définitives, le statisticien devant agir de cas en cas. Cependant, nous sommes à même d'affirmer que la démarche que nous proposons conduit à des imputations satisfaisantes tant d'un point de vue théorique qu'empirique. En outre, la méthode d'imputation multiple, étant donné sa relative simplicité, fournit des avantages non négligeables dans l'analyse des données. Cette méthode reste cependant peu répandue. Nous ne saurons assez en recommander son utilisation.

## 5. Références


**Caron, N.** : *Les principales techniques de correction de la non-réponse, et les modèles associés*, INSEE, Série des Documents de Travail « Méthodologie Statistique », N° 9604.

**Donzé, L**. (1998) : *Développement et entretien du « panel d'entreprises » du KOF/ETHZ. Une étude méthodologique*, Zurich.

**Donzé, L.** (2000) : *Le « panel d'entreprises » du KOF/ETHZ : échantillon, enquêtes, non-réponse*, Rapport final rédigé à l'attention du Fonds national suisse de la recherche scientifique, Programme prioritaire « Demain la Suisse », Zurich.

**Eurostat** : *The Second Community Innovation Survey. Documentation for Estimation of missing data (metric, ordinal, nominal* variables), Annexes II.4, II.5, II.6.

**Little, R. J.** (1988) :"*A Test of Missing Completely at Random for Multivariate Data With Missing Values*", Journal of the American Statistical Association, December 1988, Vol. 83, No. 404, Theory and Methods, pp. 1198-1202.

**Little, R. J. and Rubin, D. B**. (1987) : *Statistical Analysis with Missing Data*, John Wiley & Sons.

**Little, R. J. and Schenker, N**. (1995) : *Missing Data*, in **Arminger, G. ; Clogg, C. C. and Sobel, M. E.** : *Handbook of Statistical Modeling for the Social and Behavioral Sciences* ; Plenum Press, New York and London, 1995, chapter 2, pp. 39-75.

**Rosenbaum, P**. **R.** (1983) : "*Then central role of the propensity score in observational studies for causal effects*", Biometrika, 70, 1, pp. 41-55.

**Rosenbaum, P**. R. (1985) : "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score", The American Statistician, February 1985, Vol. 39, No. 1, pp. 33-38.

**Rubin, D. B.** (1987) : *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons.

**Rubin, D. B**. (1988) : "An overview of multiple imputation", *American Statistical Association, Proceedings of the Section on Survey Research Methods,* pp 79-84

**Rubin, D. B.** (1996): "Multiple Imputation after18+ Years*", Journal of the American Statistical Association*, June 1996, Vol. 91, No. 434, pp 473-489.

**Theil, H.** (1967) : *Economics and information theory*, North-Holland publishing company, Amsterdam.

**Huisman, M**. (1999) : *Item Nonresponse : Occurrence, Causes and Imputation of Missing Answers to Test Items*, DSWO Press, Leiden University, The Netherlands.