ETUDE DES DONNÉES MANQUANTES EN SÉRO-ÉPIDÉMIOLOGIE

Oumy Niass[1] & Abdou Kâ Diongue[2] & Aïssatou Touré[3]

[1] *LERSTAD-UGB, UIM-IPD, LSTA-Paris VI-France, niass_oumy@yahoo.fr , oniass@pasteur.sn*
[2] *LERSTAD-Université Gaston Berger, Saint-ouis-Sénégal, abdou.ka.diongue@edu-ugb.sn*
[3] *Unité d'Immunologie, Institut Pasteur de Dakar, Dakar-Sénégal, atoure@pasteur.sn*

Résumé. Les données manquantes représentent un problème récurrent en biologie, en particulier dans les études séro-épidémiologiques. La méthode la plus couramment utilisée est la suppression des sujets ayant des observations manquantes. Ce qui peut induire à une perte d'information. Le but de cette étude est de comparer un ensemble de techniques statistiques élaborées dans ce sens et aussi de montrer que la suppression n'est pas toujours une méthode efficace. Pour ce faire un échantillon de 300 observations a été collecté sur des enfants vivant dans 8 villages dans le but d'étudier la relation entre les réponses d'anticorps dirigées contre différents antigènes du *P.falciparum*. A partir de cet échantillon complet, 10 bases incomplètes ont été crée aléatoirement avec des proportions valeurs manquantes variant de 5 à 50%. Six méthodes de traitement des données manquantes ont été appliquées: par la moyennes, des k-plus proches voisins, la régression simples, les imputations multiples avec l'algorithme EM et la prédictive mean-matching. La comparaison a été faite en termes d'erreur moyenne (RMSE, MAE, R-square), de p. value, des critères AIC et BIC dans la sélection de modèle. Les résultats montrent qu'au-delà de 5% de données manquantes il est préférable d'estimer les valeurs manquantes que de les supprimer. Concernant les méthodes d'imputation, l'imputation multiple et les k-plus proches voisins sont préférables si la proportion de données manquante est énorme.

Mots-clés. Données manquantes, imputation, séro-épidémiologie, Paludisme, *P.falciparum* …

Abstract. Treatment of missing data represents a recurrent problem in biology, in particular in the sero-epidemiological studies. Indeed, the most common method used to deal with missing data is to restrict the analysis to subjects having complete information for the set of variables of interest, which can lead to a drop-out and/or introduce some slants in the evaluation. The aim of this paper is to compare some missing techniques and demonstrate that estimating missing data is sometimes more efficient than deleting them. Cross-sectional data was obtained inspecting by the relationship between different malaria antibody responses against some antigens of *P.falciparum* in a sample of 300 children from eight villages in rural area of Senegal (West Africa). The complete dataset was used to create, by simulation, incomplete dataset with percentage missing value varying between 5 to 50\%. Six methods for dealing with missing value: Complete-case (CC) analysis so-called listwise deletion, mean substitution, the k-nearest neighbors (knn), multiple imputation using the expectation-maximization (EM), predictive mean matching (pmm) and using regression were applied to all ten incomplete dataset for the same missing position. Root mean square errors (RMSE), mean square errors (MAE), p.value, multiple R-square, AIC and BIC criterion were using to compare these missing data approaches. Results demonstrate that multiple imputation using predictive mean matching (MI.pmm) and k-nearest neighbors (knn) were preferred to other missing data methods when missing data percentage was great (over 5 percent). Listwise deletion produces the most inaccurate results. Based on these results it is preferable to estimate missing value than to restrict the analysis with subject who had complete observation.

Keywords. Missing data, imputation, sero-epidemiology, malaria, *P.falciparum* …

1 Contexte

L'appréhension des données manquantes est un problème délicat. Non pas à cause de sa gestion informatique mais plutôt à cause des conséquences de leur traitement (suppression des individus ayant une mesure manquante ; ou remplacement par une valeur plausible à partir des observations disponibles : On parle d'imputation) sur les résultats d'analyse ou sur les paramètres d'intérêt. Les données manquantes peuvent se retrouver dans les variables à expliquer ou les variables indépendantes. Il existe plusieurs solutions aux problèmes des données manquantes. La méthode d'élimination est le mode de gestion le plus couramment utilisée (c'est la méthode par défaut de tous les logiciels statistiques usuels). Cette technique est raisonnable pour une proportion de données manquantes au moins égale à 5%, sinon elle peut, induire à une perte d'information ou introduire des biais dans les conclusions tirées des résultats d'analyse ou même empêcher la convergence statistique du modèle souhaité comme l'ont bien montré Molenberghs G et Kenward M (2007) et tout récemment Vergouw D et al.(2010).
Little et Rubin (1987) ont montré qu'il existe de nombreuses méthodes alternatives de traitement de données manquantes, en passant par l'analyse cas-complet, les méthodes d'imputation dites simples à l'imputation multiple.
L'objectif principal de cette étude est de comparer les estimations de différentes méthodes d'imputation pour contourner le problème des données manquantes. Pour ce faire nous utilisons une base de données réelles dans laquelle nous allons appliquer les méthodes dites simples qui consistent à remplacer une valeur manquante par une seule valeur plausible et les méthodes d'imputation multiple.

2 Méthodes de traitement de données manquantes

Little et Rubin (1987, 2002) mettent en évidence trois hypothèses distinctes sur l'origine du mécanisme de non réponse : MCAR (Missing Completely At Random), si la probabilité de non réponse pour une variable donnée ne dépend pas de celle-ci, mais uniquement des paramètres extérieurs indépendants de cette variable ; MAR (Missing At Random), si la probabilité de non-réponse peut dépendre des observations mais pas des DM (Données Manquantes); et MNAR (Missing Not At Random), lorsque la probabilité de non-réponse est liée aux valeurs prises par la variable ayant des DM.

2.1 Méthodes d'imputation simples

Imputation par la moyenne ou médiane (I.Mean)

On attribue à chaque valeur manquante la moyenne ou la valeur médiane de la même variable. L'inconvénient de cette méthode est qu'elle conduit à une réduction systématique de la dispersion de chacune des variables et risque de briser d'éventuelles relations multidimensionnelles sous-jacentes entre les variables.

Imputation utilisant la régression (I.Reg)

Le principe de cette méthode est d'utiliser les observations renseignées pour créer un modèle de régression et ensuite utiliser le modèle pour prédire les données manquantes. Les variables avec données manquantes sont considérées comme des variables dépendantes. Les valeurs manquantes seront remplacées par les valeurs prédites selon le modèle.

Explicitement, soient $Y = (Y_{Obs}, Y_{mis})$ une variable aléatoire avec des valeurs manquantes (Y_{mis}) et un modèle de régression linéaire simple basé sur les valeurs observées :

$$Y_{obs} = X\beta + \mu \quad \text{où} \quad \mu \sim N(0, \sigma) \text{ avec}$$

$\beta = (\beta_1, \beta_2, \ldots, \beta_p)$, $X = (X_1, X_2, \ldots, X_p)$ et $\mu = (\mu_1, \mu_2, \ldots, \mu_p)$

X représente la matrice des covariables (On suppose que X est complète) et β le paramètre d'intérêt. Nous avons β^* et σ^* les estimations respectives de β et σ obtenues à partir du modèles sur les données observées. Les valeurs manquantes sont prédites par :

$$Y_{mis} = X\beta^* + \mu^* \text{ , avec } \mu^* \sim N(0, \sigma^*)$$

Little et Rubin (1987) considère cette approche comme étant conditionnelle car dépendant des prédicteurs. Elle est plus sophistiquée que la méthode d'imputation par la moyenne, mais peut conduire à une surestimation des relations entre variables explicatives et dépendantes (Schafer JL et Graham JW, 2002).

Imputation par la méthode du plus proche voisin (I.KNN)

On attribue à l'individu qui a une mesure manquante sur une variable donnée, la valeur de celui qui est le plus proche qui présente une mesure pour cette même variable. Cette similarité est habituellement définie par une fonction de distance entre les variables.

Soient Y la variable d'intérêt, X la matrice des covariables et j, l'identifiant de l'individu n'ayant pas une mesure pour la variable Y (Y_j est manquant). On cherche parmi tous les individus i ayant une mesure à l'ensemble des variables l'individu j_0 qui minimise une certaine distance entre l'individu j et i :

$$j_0 = \text{Arg min}_{1 \le i \le p}\{d(i,j)\}$$

d est une mesure de distance, par exemple la distance euclidienne $d(i,j) = \sqrt{\sum_{k=1}^{p}(X_i^k - X_j^k)^2}$.

Une fois que j_0 est déterminé, la valeur Y_{j_0} est attribuée à Y_j.

$$Y_j^* = Y_{j_0}$$

Imputation par l'algorithme EM (Espérance-Maximisation)

Initialement développé par A. P. Dempster, N.M. Laird et D. Rubin (1977), l'algorithme EM est un algorithme itératif de calcul d'estimateur de vraisemblance par des modèles paramétriques lorsque les données sont observées. Dans le cadre du traitement des données manquantes, Il permet de compléter les valeurs manquante en se basant sur la vraisemblance maximale (maximum-likelihood estimation) de l'ensemble des données disponibles.

L'algorithme EM est une succession de deux étapes : une étape (E) où on évalue l'espérance de la log-vraisemblance pour la valeur courante du paramètre puis une autre étape (M) où on actualise le paramètre en maximisant cette nouvelle fonction du paramètre. L'estimation ainsi obtenue est celle qui maximise la probabilité d'observer ce qui a été réellement observé (Allison, 2001). L'algorithme converge vers un point stationnaire sous des hypothèses de régularité.

2.2 Méthode par Imputation multiple

Elle a été proposée par Rubin en 1978 puis développée et décrite en détail par Rubin (1987) et Schafer (1997). Elle consiste à remplacer une valeur manquante par m (m> 1) valeurs plausibles au sens d'un modèle statistique.

Rubin décrit la méthode comme une succession de 3 étapes. D'abord on attribue des valeurs aux données manquantes en utilisant un modèle aléatoire adapté. Ensuite répéter m fois cette étape afin

d'obtenir les m tableaux de données complétées. Et enfin analyser ces m tableaux en utilisant les méthodes statistiques standards pour l'analyse des données complètes.

$$\beta_i^* = \frac{1}{m}\sum_{j=1}^{m} \beta_{i,j}^*$$

Plus le nombre m d'imputation est grand, plus les estimations seront précises. Cependant, Rubin (1987, 1996) a montré qu'en pratique à partir d'un faible nombre d'imputations (par exemple m=5) on a de bons résultats.
Dans ce travail nous allons appliquer cette méthode en utilisant :

- un algorithme basé sur le bootstrap, approchant des résultats de l'algorithme EM (IM.EM)
- l'approche "Predictive Mean Matching (IM.pmm)" (Zio Di.M. et Guarnera. U, 2009)

3 Méthodologie

Comparaison et technique d'évaluation des méthodes

Dans le but de comparer les différentes méthodes d'imputation, nous disposons, une base contenant des données immunologiques du paludisme. Ces données vont constituer notre matrice de référence (MR) où il n'y a aucune valeur manquante. Nous avons choisi de manière aléatoire sur les n observations de la matrice MR des représentants de DM (données manquantes) au taux de T variant entre 5 et 50%. Ainsi nous avons créé des matrices avec valeurs manquantes (MVM). Pour chaque MVM les valeurs manquante simulées sont ré-estimées par la méthode de remplacement. La matrice MVM est ainsi complétée et nommée ME (Matrice Estimée). Enfin à chaque position des valeurs manquantes, nous avons calculé la différence entre la valeur réelle et la valeur estimée.
Pour évaluer l'estimation, la moyenne des carrés d'erreur notée RMSE (Root Mean Square Error), l'erreur absolue moyenne, les moyennes et écart-types sont calculés pour chaque pourcentage de données manquantes.

$$RMSE_t = \sqrt{\frac{\sum_{i=1}^{M}(R_i - E_i)^2}{M}} \qquad MAE_t = \sqrt{\frac{\sum_{i=1}^{M}(R_i - E_i)}{M}}$$

Avec R_i la valeur réelle observée à la position où une valeur manquante a été insérée, E_i la valeur estimée par la méthode à cette même position et M correspond au nombre de valeurs manquantes dans la matrice MVM. Nous avons utilisé ces mesures d'erreur ainsi que la moyenne résiduelle et les critères de sélection de modèle (AIC, BIC) pour la comparaison des méthodes étudiées.

4 Résultats

Nos données sont issues d'un suivi longitudinal fait sur 1448 enfants de moins de 10ans vivant dans huit villages (Aïdara, Daga Ndoup, Keur Ndianko, Keur saloly Bouya, Keur Samba Gueye, Némé Nding et Touba Nding) dans l'arrondissement de Toubacouta, dans la région de Fatick (Sénégal), dans le cadre du projet d'EDCTP. Sur un échantillon complet de 300 enfants, nous avons créé aléatoirement 10 bases incomplètes de taux de valeurs manquantes variant entre 5% et 50% et 290 bases complétées.
L'ensemble des méthodes produisent des estimations avec un taux d'erreur moyen assez faible. Le taux d'erreur varie faiblement avec le pourcentage de valeur manquante pour les méthodes IM.pmm et I.KNN. Pour les méthodes I.Mean et I.Reg les taux d'erreur sont plus ou moins élevés (Figure 1). La méthode d'imputation multiple par « predictivees-mean-matching » est la méthode plus fiable

dans ce sens.
Les résultats montrent que la suppression des données manquantes est la méthode la moins efficace aussi bien pour l'estimation des moyennes et variances que pour les critères de

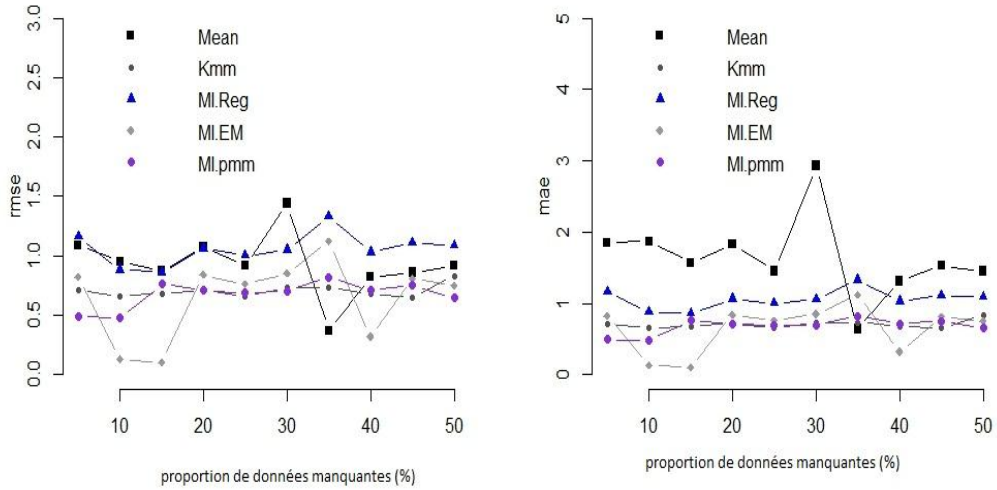

Fig1 : Evolution des RMSE et MAE en fonction du taux de données manquantes

sélection de modèle. A 5% de valeur manquantes, toutes les méthodes donnent des résultats probants. Au-delà de 5%, la méthode de la restriction sous-estime les différents paramètres (Figure 2).

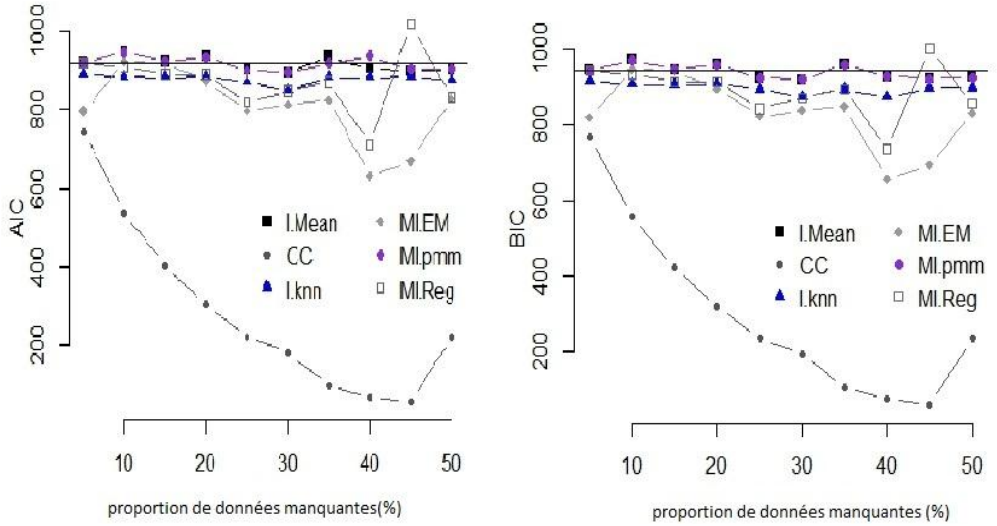

Figure 2 : Critères AIC et BIC en fonction du taux de données manquantes : La ligne noire horizontale représente la valeur de l'AIC (BIC) obtenue sur les données complètes (Données de référence) .

5 Conclusion

Nos résultats montrent qu'il est préférable d'estimer les données manquantes que de les supprimer. S'agissant des méthodes d'imputation, les méthodes basées sur l'imputation multiple sont les meilleurs dans l'ensemble. Ces résultats semblent être en conformité avec d'autres études publiées récemment (Zio Di.M. et Guarnera. U, 2009). La méthode basée sur les k plus proches voisins donne aussi des résultats satisfaisants.

Cette étude montre aussi que certaines méthodes d'imputation influencent l'adéquation des modèles multivariés lorsque la proportion de valeurs manquantes est assez élevée. Ceci a été mis en évidence tout récemment par Héraud-Bousquet V (2012). Il semble intéressant de faire des études plus approfondies afin d'évaluer l'impact de ces imputations sur les paramètres d'autres modèles multivariés.

Bibliographie

[1] Molenberghs G, Kenward M (2007), *Missing data in clinical studies,* Wiley series in Probability and Statistics, Chichester.

[2] Vergouw D, Heymans MW, Peat GM, Kuijpers T, Croft PR, de Vet HC, et al. (2012), *The search for stable prognostic models in multiple imputed data sets,* BMC Med Res Methodol; 10:81.

[3] Little R.J.A., Rubin D.B. (1987), *Statistical Analysis with Missing Data,* New York:John Wiley.

[4] Little RJA, Rubin DB (2002), Statistical *analysis with missing data,* Wiley series in Probability and Statistics. 2nd ed. New York: Wiley.

[5] Dempster, A. P., Laird, N. M. et Rubin, D. B. (1977). *Maximum likelihood estimation from Incomplete data via the EM algorithm (with Discussion). J. R. Statist. Soc.* B 39, 1-38.

[6] Schafer JL (1997), *Analysis of incomplete multivariate data*, Monographs on Statistics and Applied Probability 72 Chapman & Hall, London.

[7] Allison P. D. (2000), *Multiple Imputation for Missing Data: A Cautionary Tale*, Sociological Methods Research, 28(3), 301–309.

[8] Zio Di.M., Guarnera. U (2009). Semiparametric predictive mean matching: An empirical evaluation, AStA Advances in Statistical Analysis, 93(2).

[9] Héraud-Bousquet V. (2012), *Traitement de données manquantes en épidémiologie : Application de l'imputation multiple à des données de surveillance et d'enquêtes,* PHD, tel-00713926.

[10] Schafer JL and Graham JW (2002), *Missing data: Our view of the state of the art*, Psychological Methods.