

**Thème: Méthodes d'exploitation de  
données incomplètes: application aux  
données du programme national de  
lutte contre le paludisme**

Par  
GABFOUBE Victor  
*25 Octobre 2007*

# Plan

- Introduction
- Méthodes
- Résultats
- Conclusion

# Introduction

- **Le paludisme** (plasmodium, anophèle, transfusion sanguine)
- 4 espèces de plasmodium: falciparum (90<sup>o</sup>/o), oval, malariae, vivax.
- Endémie parasitaire depuis des siècles  $\implies$  dégâts irréparables en Afrique - Cameroun.

# Introduction

## ● **Ministère de la santé publique** - PNLN -

Objectif: réduire la morbidité et la mortalité imputables au paludisme au plus bas niveau.

**Données:** SSE recueille les données sur le terrain et les analyse.

### **Difficultés:**

- L'enclavement des services périphériques;
- absence d'un système efficace de communication;
- faible intérêt du personnel de santé.

# Introduction

⇒ Données manquantes et erronées.

**Problématique:** morbidité? mortalité? extrapolation (48<sup>o</sup>/o)? fiabilité?

# Méthodes

# Méthodes

- Disposition de données de 2006 en panel (coupes transversales et coupes temporelles).  
⇒ tableau de  $120 \times 27$ .
- Unité = province,
- variable: indexes de morbidité et mortalité.

Valeurs manquantes =  $62,53\%$ .

# Méthodes

## Méthode de traitement des données manquantes:

- **Elimination par liste**

- ★ **Avantage:** données réelles
- ★ **Inconvénient:** risque de biaiser la représentativité de l'échantillon.

# Méthodes

- *Imputation*

- ★ **Avantage:** crée une base de données complète, conserve la représentativité de l'échantillon.

- **Remplacement par la moyenne de la variable**

- ★ **Inconvénient:** Déformation de la distribution marginale, variance, corrélation avec d'autres variables.

# Méthodes

- Simulation suivant une loi de probabilité
- Transformation Box-Cox

$$Y = P_\lambda(X) = \begin{cases} \frac{X^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0 \\ \ln(X) & \text{si } \lambda = 0 \end{cases} \quad (1)$$

$$\implies X = \begin{cases} \exp\left(\frac{\ln(\lambda Y + 1)}{\lambda}\right) & \text{si } \lambda \neq 0 \\ \exp(Y) & \text{si } \lambda = 0 \end{cases} \quad (2)$$

# Méthodes

- Méthode de type hot-deck (procédure d'imputation ABB)
  - ★ cellule d'imputation ou cellule d'ajustement (*propensity score, la méthode du plus proche voisin ...*).
  - ⇒ Imputation multiple.

# Méthodes

## Estimation ponctuelle et inférence dans un système d'imputation multiple

$\delta_m, U_m, m = 1, \dots, M, M$  estimateurs respectifs du paramètre  $\delta$  et de sa variance  $U$ .

★ L'estimateur final de  $\delta$  est

$$\bar{\delta} = \frac{1}{M} \sum_{m=1}^M \delta_m \quad (3)$$

# Méthodes

★ variance intra-imputation:

$$U = \frac{1}{M} \sum_{m=1}^M U_m \quad (4)$$

★ variance inter-imputation:

$$B = \frac{1}{M-1} \sum_{m=1}^M (\delta_m - \bar{\delta})^2 \quad (5)$$

★ variance totale:

$$T = U + \left(1 + \frac{1}{M}\right) B. \quad (6)$$

# Méthodes

- ★ Intervalle de confiance du multiple imputation de  $\delta$  :

$$I.C_{MI}(\delta) = \bar{\delta} \pm q_{ddl} \sqrt{T} \quad (7)$$

où  $q_{ddl}$  = quantile de la distribution de Student de degré de liberté

$$ddl = (M - 1) \left( 1 + \frac{MU}{(M + 1)} \right)^2. \quad (8)$$

# Méthodes

## ★ Taux d'informations manquantes

$$\gamma = \frac{r + \frac{2}{ddl+3}}{r + 1} \quad (9)$$

avec

$$r = \frac{T - U}{M} \quad (10)$$

est l'accroissement relatif de la variance due aux données manquantes.

# Méthodes

- ★ Efficacité d'une estimation basée sur  $M$  imputations est

$$\left(1 + \frac{\gamma}{M}\right)^{-1} \quad (11)$$

# Résultats

# Résultats

- **Elimination par liste:**

**Morbidité:**  $55,19\%$   $I.C_{95\%} = [36,43 ; 73,95]$

**Taux d'hospitalisation:**  $81,77\%$   $I.C_{95\%} = [66,42 ; 95,92]$

**Mortalité:**  $41,42\%$   $I.C_{95\%} = [22,83 ; 60,00]$

★ Femmes enceintes:  $64,30\%$  et  $70,83\%$

★ Enfants de moins de 5 ans:  $72,58\%$  et  $55,17\%$

# Résultats

## Limite

- Réduction du tableau à 15 lignes
- Provinces représentées: EST:1 ; NO:4 ; NW: 1 ; OU:2 ; SW:7.

# Résultats

- Remplacement par la moyenne:

**Morbidité:**  $54,79\%$   $I.C_{95\%} = [36,02 ; 73,57]$

**Taux d'hospitalisation:**  $67,99\%$   $I.C_{95\%} = [50,39 ; 85,59]$

**Mortalité:**  $35,84\%$   $I.C_{95\%} = [17,75 ; 53,93]$

★ Femmes enceintes:  $82,91\%$  et  $57,87\%$

★ Enfants de moins de 5 ans:  $74,26\%$  et  $33,83\%$

# Résultats

- **Simulation suivant une loi de probabilité et mode de regression:**

★ *Transformation Box-Cox* → test de Shapiro-Wilk

**exception:** DP.p, DF.cc et DF.p.

**Imputation:** transformation inverse des valeurs simulées

**correlation:** DP.p - E.cf; DF.cc - DP.cc et DF.p - HE.cl

**Nombre d'imputations:**  $M = 15$ .

# Résultats

- **Morbidité:**  $47,26\%$  ;  $I.C_{95\%} = [36,76 ; 57,75]$  ;  $\gamma = 99,93\%$  ; **Efficacité:**  $93,75\%$
- **Taux de hospitalisation:**  $60,58\%$  ;  $I.C_{95\%} = [47,56 ; 73,67]$  ;  $\gamma = 99,78\%$  ; **Efficacité:**  $93,76\%$
- **Mortalité:**  $30,93\%$  ;  $I.C_{95\%} = [19,78 ; 42,07]$  ;  $\gamma = 99,87\%$  ; **Efficacité:**  $99,76\%$

Femmes enceintes:  $65,06\%$ ;  $47,01\%$

Enfants de moins de 5 ans:  $57,03\%$ ;  $37,03\%$

# Résultats

- **Procédure ABB**

Taux de données manquantes très élevé, et les données étant manquantes complètement au hasard MCAH  $\implies$  Cellule d'ajustement d'imputation = 1.

Nombre d'imputation:  $M = 15$ .

# Résultats

● **Morbidité:**  $52,94\%$  ;  $I.C_{95\%} = [42,17 ; 63,70]$  ;

$\gamma = 99,91\%$  ; **Efficacité:**  $93,75\%$

● **Taux de hospitalisation:**  $69,17\%$  ;

$I.C_{95\%} = [57,02 ; 81,27]$  ;  $\gamma = 99,81\%$  ;

**Efficacité:**  $96,16\%$

● **Mortalité:**  $33,57\%$  ;  $I.C_{95\%} = [26,46 ; 43,97]$  ;

$\gamma = 99,92\%$  ; **Efficacité:**  $96,16\%$

Femmes enceintes:  $69,73\%$  ;  $62,03\%$

Enfants de moins de 5 ans:  $66,36\%$  ;  $38,32\%$

# Résultats

## Comparaison

★ Par méthode:

Pour mesurer la précision de l'imputation, on calcule l'indice d'erreur quadratique moyenne normalisé (*Mean Square Error*)

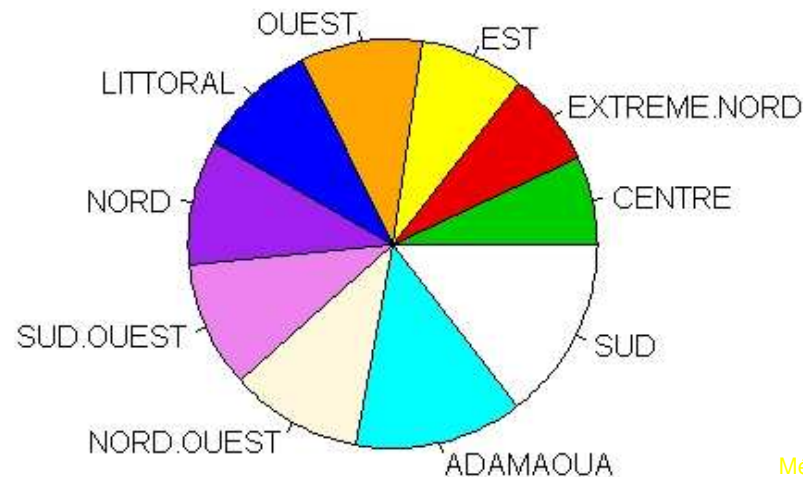
$$MSE = \frac{1}{n \times p \times p_m} \sum_{j=1}^p \left[ \frac{\sum_{i=1}^n (x_{ij} - \hat{x}_{ij})^2}{\sum_{i=1}^n (x_{ij} - \widehat{x}_{ij})^2} \right] \quad (12)$$

# Résultats

La méthode de remplacement par la moyenne bénéficie d'une légère faveur.

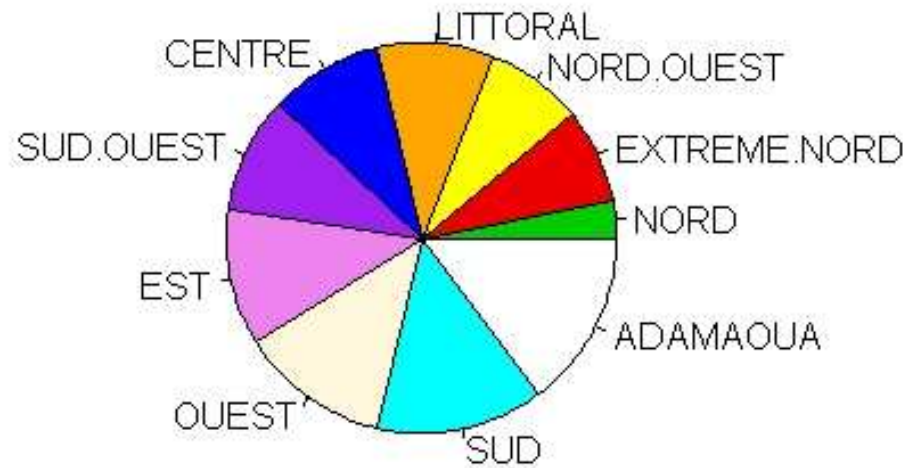
★ Par province:

**Taux de consultations liées au paludisme par province**



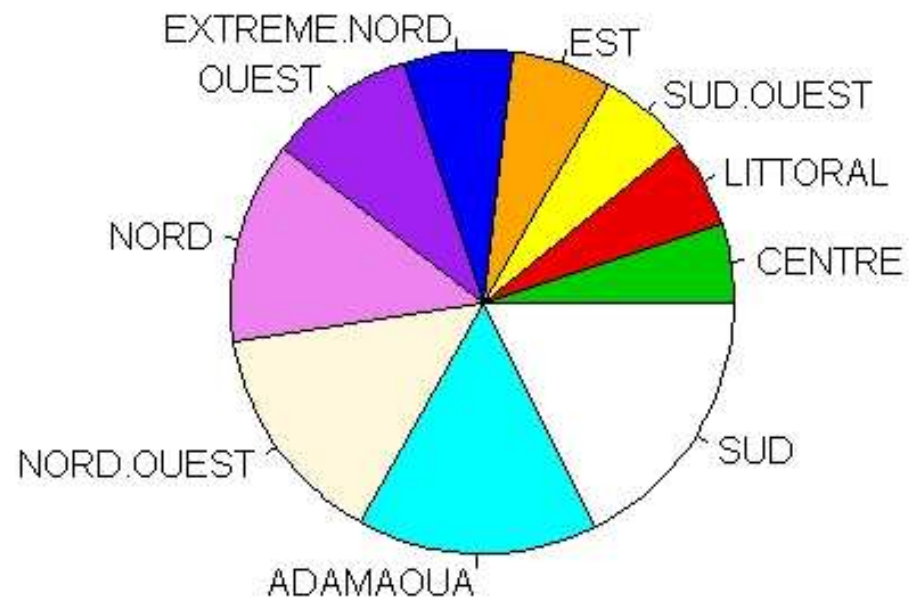
# Résultats

## Taux d'hospitalisations liées au paludisme par province



# Résultats

**Taux de décès liés au paludisme par province**



# Résultats

## Extrapolation:

Sur l'ensemble de la population du Cameroun

**Morbidité:**  $26,47\text{‰}$   $I.C_{95\text{‰}} = [18,58 ; 34,39]$

**Mortalité:**  $5,21\text{‰‰}$   $I.C_{95\text{‰}} = [0,00 ; 46,03]$

# Conclusion

# Conclusion

Au terme de nos analyses, le PNLN, par ricochet le Ministère de la santé publique, a du pain sur la planche vis-à-vis des objectifs qu'il s'est fixé.

En effet, la morbidité et la mortalité imputables au paludisme restent encore très élevées, surtout chez les enfants de moins de 5 ans et les femmes enceintes, qui représentent 22% de la population camerounaise.

MERCI POUR VOTRE  
BIEN AIMABLE  
ATTENTION