

ESTIMATION ET PREVISION RELATIVES A LA PREVENTION DE LA TRANSMISSION DU VIH DE LA MERE A L'ENFANT .

Par KENGNE William Charky

Master de Statistique Appliquée
ENSP - Université de Yaoundé 1

22 Octobre 2007

Généralités

Dans les études empiriques, il arrive fréquemment que l'ensemble des données avec lesquelles on doit travailler ne soit pas complet. C'est-à-dire qu'il comporte des valeurs manquantes (v.m.) . Lorsque les données sont obtenues à l'aide d'une enquête par sondage, il est fréquent d'être confronté au problème de la non-réponse. Un autre problème auquel sont confrontés les statisticiens est la prévision future d'un phénomène dont on a observé son comportement dans le passé. Il s'agit donc de supposer que les mêmes causes produisent les mêmes effets.

Généralités

Solution facile à implémenter pour les v.m.

Éliminer toutes les unités ou observations qui comportent au moins une variable à valeur manquante.

Conséquences :

- perte d'information ;
- risque d'introduction de biais dans les analyses statistiques.

Autre solution

Chercher à construire des valeurs artificielles pour remplacer les valeurs manquantes : c'est l'imputation.

Généralités

Solution facile à implémenter pour les v.m.

Éliminer toutes les unités ou observations qui comportent au moins une variable à valeur manquante.

Conséquences :

- perte d'information ;
- risque d'introduction de biais dans les analyses statistiques.

Autre solution

Chercher à construire des valeurs artificielles pour remplacer les valeurs manquantes : c'est l'imputation.

Généralités

Solution facile à implémenter pour les v.m.

Éliminer toutes les unités ou observations qui comportent au moins une variable à valeur manquante.

Conséquences :

- perte d'information ;
- risque d'introduction de biais dans les analyses statistiques.

Autre solution

Chercher à construire des valeurs artificielles pour remplacer les valeurs manquantes : c'est l'imputation.

Contexte

Données sur la prévision de la transmission du VIH de la mère à enfant qui comporte plus de 69% de données manquantes (DM).

Problématique

- Estimer à partir des valeurs observées, les valeurs manquantes sur la base des données PTME.
- Prévoir les réalisations des variables de la PTME pour le second semestre 2007.

Plan de l'exposé

- I- Présentation des données PTME
- II- Objectifs
- III- Méthode utilisées
- IV- Résultats
- V- Conclusion

Description des données PTME

- Source : GTC.
- Données de panel.
- Données mensuelles recueillies sur 18 mois (Janvier 2006 - Juin 2007).
- 596 sites au total dissiminés à travers le territoire national.
- 11 variables.

Mécanisme b'obtention des données et origines des DM

Obtention:

Données recueillies auprès des sites PTME et transférées au GTP puis au GTC pour compilation.

Origine des DM

- Les données PTME sont censées raportées les activités auprès des sites PTME. Il arrive que les activités effectuées au sein de certains sites ne sont pas enregistrées.
- Il arrive aussi que les données transférées ne parviennent pas au GTP : problème de transfère de données.

Ce sont là les deux principales orines des DM dans les données PTME.

Mécanisme b'obtention des données et origines des DM

Obtention:

Données recueillies auprès des sites PTME et transférées au GTP puis au GTC pour compilation.

Origine des DM

- Les données PTME sont censées raportées les activités auprès des sites PTME. Il arrive que les activités effectuées au sein de certains sites ne sont pas enregistrées.
- Il arrive aussi que les données transférées ne parviennent pas au GTP : problème de transfère de données.

Ce sont là les deux principales orines des DM dans les données PTME.

Mécanisme b'obtention des données et origines des DM

Obtention:

Données recueillies auprès des sites PTME et transférées au GTP puis au GTC pour compilation.

Origine des DM

- Les données PTME sont censées raportées les activités auprès des sites PTME. Il arrive que les activités effectuées au sein de certains sites ne sont pas enregistrées.
- Il arrive aussi que les données transférées ne parviennent pas au GTP : problème de transfère de données.

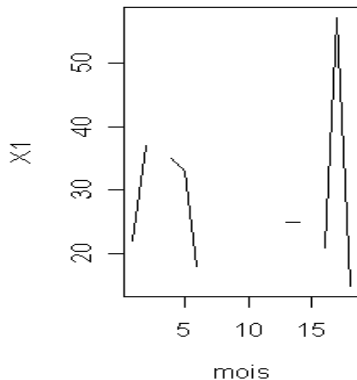
Ce sont là les deux principales orines des DM dans les données PTME.

Qualité des DM

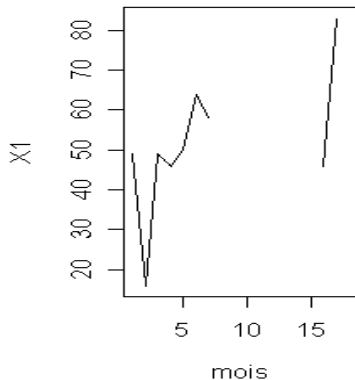
- Sous une vue transversale, on a en principe les données manquantes totales de part leurs origines. i.e. pour un mois données, les enregistrements sont soit présentes soit absentes sur tous les 11 variables.
- Mais pour des raisons diverses, on note aussi la présence des données manquantes partielles dans les données PTME.

Problèmes posés par les DM dans les données PTME

X1 serie.n°2 avec DM



X1 serie.n°30 avec DM



Objectifs

- Réduire (Voir même amener à 0) le taux de DM dans les données PTME par imputation simple.
- Prévoir les réalisations des variables de la PTME pour le second semestre 2007.
- Mettre en place un modèle dynamique de prévision (fiable) des variables de la PTME

Méthodes utilisées pour l'imputation

- *La méthode déductive* : on déduit la valeur manquante des valeurs des autres variables ou des informations disponibles sur l'individu.
- *La méthode de la moyenne transversale ou par période*: on remplace de façon transversale la variable non observée chez certains individus par la moyenne de la même variable observée chez les autres individus.
- *La méthode k-nn déterministe* : soit $\vec{X} = (X_1, \dots, X_p)$ un vecteur aléatoire réel qui doit être observé sur un individu i . On a observé X_1, \dots, X_{p-1} et que X_p est manquante sur l'individu.
Parmi les individus dont \vec{X} a été entièrement observée, on prend les k plus proche de l'individu i , et on affecte à $X_{i,p}$,

Méthodes utilisées pour l'imputation

Problème d'application de la méthode $k - nn$

- Choix de k .
- estimation de l'erreur de prédiction.

Méthodes utilisées pour l'imputation

La data augmentation

- $Y = (Y_{obs}, Y_{mis})$.
- **Etape1** : $Y_{mis}^{(k+1)} \sim \mathbb{P}(Y_{mis}/Y_{obs}, \theta^{(k)})$
- **Etape2** : $\theta^{(k+1)} \sim \mathbb{P}(\theta, Y_{obs}/Y_{obs})$

Partant d'une valeur initiale $\theta^{(0)}$, on obtient une séquence $(Y_{MIS}^{(1)}, \theta^{(1)}) ; (Y_{MIS}^{(2)}, \theta^{(2)}), \dots$ dont la distribution stationnaire est $\mathbb{P}(\theta, Y_{mis}/Y_{obs})$

L'idée de la data augmentation (DA) est d'imputer à Y_{mis} une valeur produite par le même mécanisme qui devrait produire Y_{mis}

Méthodes utilisées pour la prévision

Modélisation $ARIMA(p, d, q)$

- Estimation de d
- Estimation de p
- Estimation de q

On utilise le plus souvent la méthodologie de Box et Jenkins (1970) pour estimer ces trois paramètres.

Après l'étape d'identification, on estime les paramètres du modèle soit :

Méthodes utilisées pour la prévision

Modélisation $ARIMA(p, d, q)$ (suite)

- 1 en maximisant la vraisemblance

$$\mathcal{L}(x_1, \dots, x_T, \theta, \phi, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{T/2}} \frac{1}{(\det((XX')^{-1}))^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(X\theta - y)^T (XX')^{-1} (X\theta - y)\right)$$

- 2 en minimisant la somme des carrés des résidus

$$S(\phi, \theta) = \sum_{i=1-Q}^T \hat{\varepsilon}_i^2.$$

Les prévisions s'obtiennent en utilisant la représentation AR de la serie.

Méthodes utilisées pour la prévision

Le lissage exponentiel simple

$$\hat{X}_T(h) = (1 - \beta) \sum_{j=0}^{T-1} \beta^j X_{T-j}. \quad (1)$$

Application de la méthode déductive

Variables	Taux de V. M. avant imp en (%)	Taux de V. M. après imp en (%)	Taux d'imp en (%)
X_1	52.60	52.04	0.56
X_2	58.00	52.13	5.87
X_3	54.40	52.04	2.36
X_4	84.20	58.82	25.38
X_5	74.20	52.04	22.16
X_6	71.24	52.04	19.20
X_7	72.10	52.04	20.06
X_8	74.40	52.04	22.36
X_9	81.25	52.04	29.21
X_{10}	78.90	52.16	26.74
X_{11}	59.10	52.04	7.06

Application de la méthode $k - nn$

Application de la V.C. pour X_2

Pour la variable X_2 , la plus petite erreur estimée, obtenue avec la moyenne des k voisins ou la modalité majoritaire des k voisins est 71% au mois de Mai avec $k = 5$. Il apparaît donc que la méthode k -nn n'est pas pertinente pour imputer X_2 car cette erreur est trop importante.

Application de la V.C. pour X_4

Pour la variable X_4 le k -nn avec modalité majoritaire des k voisins prédit mieux la variable X_4 que le k -nn avec moyenne des k voisins. On a pour près de 9 mois une erreur de prévision qui ne dépasse pas 9%.

Application de la méthode $k - nn$ (suite)

Application de la V.C. pour X_{10}

Pour la variable X_4 le k -nn avec modalité majoritaire prédit mieux la variable X_{10} que le k -nn avec la moyenne des k voisins. La prédiction est assez bonne, car pour près de 75% de mois, l'erreur de prédiction estimée n'atteint pas 6,5%.

Application de la méthode $k - nn$

Pour la variable X_2 , on a moins de 1% de DM et cela nous a suggéré une imputation par moyenne transversale.

Variables	Taux de V. M. avant imp en (%)	Taux de V. M. après imp en (%)	Taux d'imp en (%)
X_2	52.13	52.04	0.09
X_4	58.82	52.04	6.78
X_{10}	52.16	52.04	0.12

Table: Résultat de l'imputation par le k -nn et par la moyenne.

Application de la data augmentation

Principe

Dans la province de l'Adamaoua, on a 10 sites (sur 42) dont la variable X_1 a été observée pendant tous les 18 mois. Nous allons d'abord modéliser les 10 séries temporelles de la variable X_1 pour chacun des 10 sites. Ensuite, nous allons nous appuyer sur ces modèles pour avoir les renseignements sur la variable X_1 des autres sites de cette province. Les sites où X_1 a été observée sur tous les 18 mois sont les sites N° 14, 15, 21, 22, 23, 24, 26, 27, 29 et 42. Nous avons adopté la méthodologie de de Box-Jenkin.

Résultats d'identification des modèles

Il ressort de l'application de la méthodologie de Box-Jenkin qu'on peut supposer que les 10 séries ci-dessus sont nées d'un processus $ARIMA(1, 1, 1)$.

Application de la data augmentation

Principe

Dans la province de l'Adamaoua, on a 10 sites (sur 42) dont la variable X_1 a été observée pendant tous les 18 mois. Nous allons d'abord modéliser les 10 séries temporelles de la variable X_1 pour chacun des 10 sites. Ensuite, nous allons nous appuyer sur ces modèles pour avoir les renseignements sur la variable X_1 des autres sites de cette province. Les sites où X_1 a été observée sur tous les 18 mois sont les sites N° 14, 15, 21, 22, 23, 24, 26, 27, 29 et 42. Nous avons adopté la méthodologie de de Box-Jenkin.

Résultats d'identification des modèles

Il ressort de l'application de la méthodologie de Box-Jenkin qu'on peut supposer que les 10 séries ci-dessus sont nées d'un processus $ARIMA(1, 1, 1)$.

Application de la data augmentation

Application de la data augmentation pour l'imputation

Le modèle étant identifié, il reste à estimer les coefficients $\theta = (\varphi_1, \theta_1)$ du modèle $ARIMA(1, 1, 1)$ pour chacune des séries de la variable X_1 dans l'Adamaoua. Les données de certains sites étant incomplètes, nous allons procéder comme suit:

Partant d'une valeur $\theta^{(0)}$, on va générer une suite de paramètre $(\theta^k)_k$.

On se donne une valeur tol et l'on s'arrête lorsque $\|\theta^k - \theta^{k-1}\| < tol$ et on dit que la séquence $(\theta^k)_k$ a convergé pour tol .

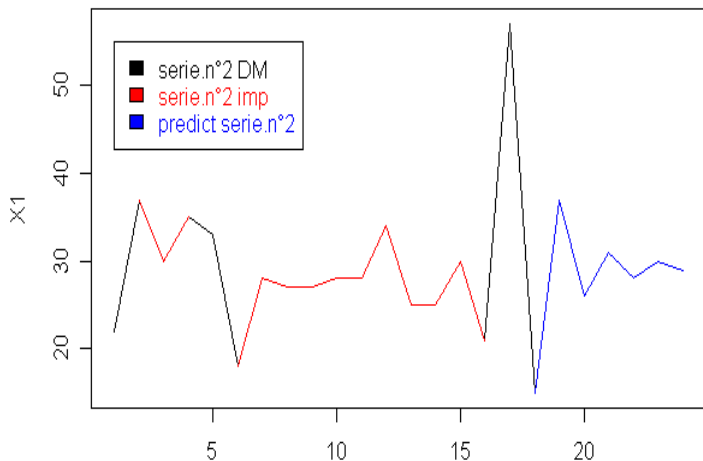
Exemple

itération $n^{\circ}k$	$\theta^k = (ar_1^k, ma_1^k)$	$\ \theta^k - \theta^{k-1} \ $
itération $n^{\circ}0$	$\theta^0 = (-0.5854, -0.9999)$	
itération $n^{\circ}1$	$\theta^1 = (-0.7606, -0.5213)$	0.5097
itération $n^{\circ}2$	$\theta^2 = (-0.5224, -0.9999)$	0.5346
itération $n^{\circ}3$	$\theta^3 = (-0.5331, -0.9999)$	0.0107
⋮	⋮	⋮
itération $n^{\circ}11$	$\theta^{11} = (-0.4509, -0.9999)$	
itération $n^{\circ}12$	$\theta^{12} = (-0.7672, -0.1920)$	0.8676
itération $n^{\circ}13$	$\theta^{13} = (-0.4604, -0.9999)$	0.8642
itération $n^{\circ}14$	$\theta^{14} = (-0.4614, -0.9999)$	0.0009
itération $n^{\circ}15$	$\theta^{15} = (-0.4613, -0.9999)$	0.0001

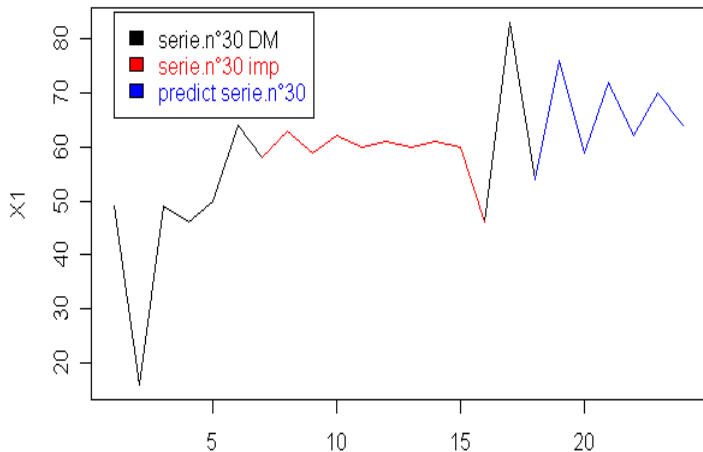
Mois	Juil-07	Août-07	Sept-07	Oct-07	Nov-07
Prévision Série N° 2	37	26	31	28	30
Prévision série N° 30	76	59	72	62	70
Prévision série N° 35	20	15	17	16	17
Prévision série N° 40	38	39	39	39	40

Table: Prévisions de X_1 des sites N° 2, N° 30, N° 35 et N° 40 pour le second semestre 2007.

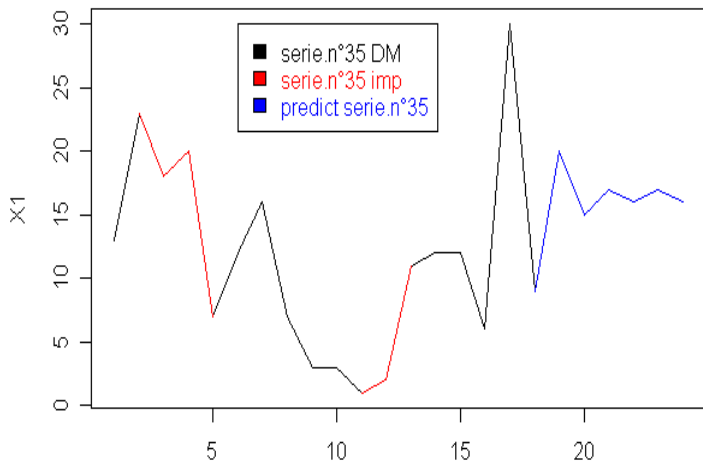
evolution de la serie n°2



évolution de la serie n°30



évolution de la serie n°35



évolution de la serie n°40

