Prise en compte des covariables temporelles dans la modélisation des événements récurrents

Génia Babykina, Vincent Couallier, Yves Le Gat

Prise en compte des covariables temporelles dans la modélisation des événements récurrents

Génia Babykina[a,b], Vincent Couallier[b], Yves Le Gat[a]

[a] *Cemagref, UR "Réseaux, épuration et qualité des eaux". 50 avenue de Verdun 33612 Cestas Cedex*
genia.babykina@cemagref.fr
[b] *IMB, UMR 5251. Université Victor Segalen Bordeaux 2. 146, rue Léo-Saignat 33076 Bordeaux Cedex*

Résumé

L'étude porte sur la modélisation des événements récurrents d'un système réparable avec maintenances imparfaites, fondée sur la généralisation du modèle NHPP. L'effet des maintenances sur l'intensité est pris en compte par le nombre de défaillances précédentes. L'introduction des covariables dépendantes du temps dans le modèle a été détaillée ainsi que l'algorithme de simulation des données. L'application à la modélisation des défaillances d'un réseau d'eau est présentée.
Mots-clés : événements récurrents, NHPP, processus de naissance, variables dépendantes du temps, défaillances de conduites d'eau.

Abstract

The study aims at modelling recurrent events of a repairable system with imperfect maintenances based on the generalisation of the NHPP model. The effect of the maintenances on the intensity is taken into account by considering the number of previous failures. The introduction of time-dependent covariates into the model is detailed as well as the algorithm of data simulation. The application to failure modelling of the water distribution system is presented.
Key-words: recurrent events, NHPP, birth process, time-dependent covariates, water pipe failures.

1 Introduction

Il est naturel de modéliser les événements récurrents d'un système par le processus de comptage de type NHPP, Processus de Poisson Non-Homogène (Andersen *et al.* (1993)), dont l'intensité s'écrit

$$\begin{cases} \mathbb{E}\left[dN(t) \,|\, \mathcal{H}\right] = \mathbb{E}\left[dN(t)\right] = \lambda(t)\, dt \\ \lambda(t) = \lambda_0(t) e^{Z'\beta} \quad (Z' = Z^t) \end{cases}$$

avec $dN(t)$: le saut du processus à l'instant t, $\mathcal{H}$: l'historique, $\lambda_0(t)$: l'intensité de base, Z : vecteur des covariables, β : vecteur des paramètres à estimer.

En fiabilité le NHPP peut être utilisé dans la modélisation des défaillances d'un système réparable avec maintenances neutres (ABAO : *As Bad As Old*), i.e. après la réparation le système revient à l'état dans lequel il était juste avant la défaillance. En situation de maintenances imparfaites, chaque défaillance et/ou action de réparation déteriore l'état du système. Le Gat (2009) propose dans ce cas d'élargir le modèle de NHPP en y intégrant le facteur du nombre des défaillances précédentes. Le modèle ainsi construit est basé sur le processus de naissance (voir Ross (1983)) et son intensité s'écrit

$$\begin{cases} \mathbb{E}\left[dN(t) \,|\, N(t-)\right] = (1+\alpha j)\lambda(t), \;\; (\alpha > 0) \\ \lambda(t) = \delta t^{\delta-1} e^{Z'\beta} \end{cases} \tag{1}$$

avec j : le nombre d'événements précédents, t : l'âge du système, Z : vecteur p - dimensionnel des covariables fixes dans le temps, $\theta = \{\alpha, \delta, \beta_1, \ldots, \beta_p\}$: paramètres à estimer. Le modèle (1) est un NHPP pour $\alpha = 0$. On note $\Lambda(t) = \int_0^t \lambda(s)\, ds = t^\delta e^{Z'\beta}$.

Le Gat (2009) montre que le nombre de défaillances dans une période de temps donnée $[a, b]$, conditionnellement au nombre de défaillances précédentes suit la distribution *Binomiale Négative* :

$$\left[N(b) - N(a) \,|\, N(a-) = m\right] \sim \mathcal{NB}\left(\alpha^{-1} + m,\, e^{\alpha[\Lambda(b)-\Lambda(a)]}\right). \tag{2}$$

Cela nous permet d'écrire la vraisemblance du processus défini par le système (1) et d'estimer ses paramètres.

Notre étude s'applique à la modélisation du processus des casses de conduites dans un réseau d'eau, dont l'intensité est sujette à des fluctuations temporelles qui ne sont pas directement prises en compte par le modèle (1), ce qui introduit un biais dans les estimations. Nous proposons ici de modéliser les fluctuations d'intensité du processus en intégrant les variables dépendantes du temps dans la partie $Z'\beta$ du modèle. Les détails des calculs adaptés à ce cas sont présentés dans la Section 2. Dans la Section 3 nous décrivons la méthode de simulation des données suivant (1) en présence des variables temporelles.

2 Intégration des variables dépendantes du temps

Deux situations sont envisageables :

(a) Variable temporelle constante par morceaux, ce qui se traduit par une ou plusieurs période(s) marquée(s) dans le processus de défaillances (ex.: une année froide, campagne de recherche de fuites, augmentation de pression dans le réseau, etc.)

(*b*) Variable temporelle dont l'effet sur le processus est continu (ex.: la température moyenne journalière, l'humidité, etc.)

2.1 Variable temporelle constante par morceaux

Nous supposons dans ce cas que la période d'observation est découpée en k sous-intervalles, sur chacun desquels l'effet de la variable temporelle, noté $\tilde{\beta}_k$, est constant (cf. Figure 1 pour illustration).

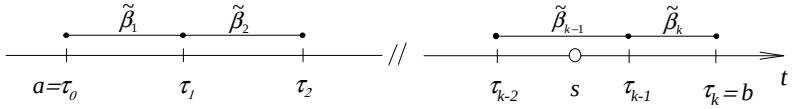

Figure 1: Variable temporelle discrète

Le vecteur p-dimensionnel des covariables fixes, Z, pour chaque individu est complété par k variables indicatrices $\tilde{Z}_l$, telles que :

$$\tilde{Z}_l(t) = \begin{cases} 1 \text{ si } t \in [\tau_{l-1}, \tau_l[\,, & l = 1, \ldots, k \\ 0 \text{ sinon} \end{cases}$$

Ainsi, le vecteur de covariables sera $\tilde{Z}_l = \{Z, \underbrace{0, \ldots, 0}_{(p+1,\ldots,p+(l-1))}, 1, \underbrace{0, \ldots, 0}_{(p+(l+1),\ldots,p+k)}\}$ et le vecteur des coefficients des covariables : $\tilde{\beta} = \{\beta, \tilde{\beta}_1, \ldots, \tilde{\beta}_k\}$.

Selon le modèle (1), l'écriture de la vraisemblance des événements survenus aux instants t_j dans l'intervalle $[a, b]$ nécessite le calcul des intégrales $\Lambda(s)$, $s \in \{a, b, t_j\}$. La modification de ces calculs en présence des covariables temporelles est detaillée ci-dessous.

Les covariables étant fixes sur chaque intervalle $\left[\tau_{l-1}, \tau_l\right[$, nous pouvons calculer l'intégrale :

$$\Lambda_l = \int_{\tau_{l-1}}^{\tau_l} \lambda(u)\, du = \int_{\tau_{l-1}}^{\tau_l} \delta t^{\delta-1}\, e^{\tilde{Z}_l'\tilde{\beta}} = \left(\tau_l^\delta - \tau_{l-1}^\delta\right) e^{\tilde{Z}_l'\tilde{\beta}}.$$

En fixant $s \in [\tau_h, \tau_{h+1}]$, les $\Lambda(s)$ dans la vraisemblance sont calculés selon :

$$\Lambda(s) = \sum_{l=0}^{h-1} \left(\tau_{l+1}^\delta - \tau_l^\delta\right) e^{\tilde{Z}_{l+1}'\tilde{\beta}} + \left(s^\delta - \tau_h^\delta\right) e^{\tilde{Z}_{h+1}'\tilde{\beta}} \tag{3}$$

Les inconvénients de la méthode présentée sont les suivants :

- le grand nombre de paramètres supplémentaires à estimer $\left(\tilde{\beta}_1, \ldots, \tilde{\beta}_k\right)$;
- la nécessité de connaître ou supposer connus le début et la fin de la période marquée, $[\tau_{l-1},\, \tau_l[$;
- la transformation des données pour l'estimation est compliquée.

En revanche, l'écriture explicite des intégrales $\Lambda(s)$ facilite les calculs et accelère la procédure d'optimisation de la vraisemblance.

2.2 Variable temporelle continue

Notons $X(t)$ la variable continue qui influence l'intensité du processus tout au long de la période d'observation. L'intensité instantanée s'écrit

$$\begin{cases} \mathbb{E}\left[dN(t) \,|\, N(t-)\right] = (1 + \alpha j)\lambda(t) \\ \lambda(t) = \delta t^{\delta-1} e^{Z'\beta + \gamma X(t)} \end{cases} \tag{4}$$

En supposant que la variable $X(t)$ est observée avec une certaine périodicité aux instants $\{\tau_1, \ldots, \tau_i, \ldots, \tau_m\}$, nous pouvons calculer $\lambda(t)$ à chacun de ces instants. Sur les intervalles $[\tau_i, \tau_{i+1}[$ entre les observations, $\lambda(t)$ peut être approchée par une droite $y(t)$.
L'intégrale $\Lambda(s)$ pour $s \in [\tau_h,\, \tau_{h+1}]$, s'écrit $\Lambda(s) = \sum_{i=0}^{h-1} \Lambda(\tau_i,\, \tau_{i+1}) + \Lambda(\tau_h,\, s)$ avec $\Lambda(\tau_i,\, \tau_{i+1}) = \int_{\tau_i}^{\tau_{i+1}} y(u)\, du$. Un exemple d'intégration approchée pour une covariable temporelle périodique est présenté sur la Figure 2.

Les avantages de la méthode sont les suivants :
- supposer continu l'effet de la variable sur l'intensité est plus réaliste ;
- l'approche ne nécessite l'estimation que d'un paramètre supplémentaire, γ.

Les limites de l'approche sont :
- il est nécessaire de supposer que la forme de $X(t)$ est connue ;
- l'intégration de $\lambda(t)$ approchée pour chaque individu ralentit le calcul de la vraisemblance à l'étape d'estimation.

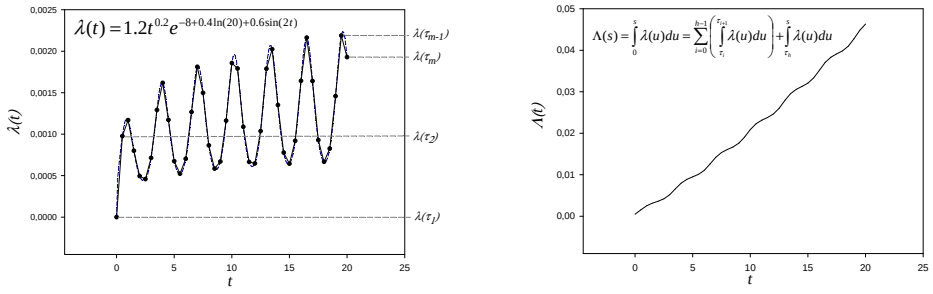

Figure 2: Calcul de $\Lambda(s)$ avec une variable temporelle continue observée ponctuellement

3 Simulation des données

Les temps des événements, t_j ($j \in \mathbb{N}$), sont simulés successivement par la méthode de la transformée inverse. La fonction utilisée est celle de la survie conditionnelle (notée $S(.)$) du temps (noté X_{j+1}) entre les défaillances j et $j+1$. Cette fonction est la probabilité conditionnelle de ne pas avoir d'événement dans l'intervalle $]t_j, t_j + x]$ sachant que j événements se sont produits sur $[0,\, t_j[$ $(N(t_j) = j)$. En utilisant (2) Le Gat (2009) montre que :

$$S\left(X_{j+1}\right) = \mathbb{P}\left[X_{j+1} > x \,|\, T_j = t_j\right] = e^{-(1+\alpha j)[\Lambda(t_j+x)-\Lambda(t_j)]}$$

Avec la méthode de la transformée inverse (voir par exemple Shih et Leems (1993)), nous avons : $e^{-(1+\alpha j)[\Lambda(t_{j+1})-\Lambda(t_j)]} = u_{j+1} \Rightarrow \Lambda\left(t_{j+1}\right) = \Lambda(t_j) - \frac{\ln(u_{j+1})}{1+\alpha j}$ et $t_{j+1} = \Lambda^{-1}(t_{j+1})$ avec u la réalisation de la variable aléatoire uniforme sur $[0,1]$ $(u \sim \mathcal{U}_{(0,1)})$ et t_{j+1} le temps de $(j+1)^{\text{ème}}$ événement.

Le problème qui se pose dans le cadre des simulations est la nécessité d'identifier l'intervalle de temps, $[\tau_h, \tau_{h+1}[$, dans lequel se trouve la prochaine défaillance avant de calculer son temps exact t_{j+1}. En effet, si $t_{j+1} \in [\tau_h,\, \tau_{h+1}[$, alors $\Lambda\left(t_{j+1}\right) = \Lambda(\tau_h) + \left(t_{j+1}^{\delta} - \tau_h^{\delta}\right) e^{Z'_{h+1}\beta}$ et d'après (3) nous avons :

$$t_{j+1} = \left(\frac{\Lambda(t_j) - \frac{\ln(u_{j+1})}{1+\alpha j} - \Lambda(\tau_h)}{e^{Z'_{h+1}\beta}} + \tau_h^{\delta}\right)^{1/\delta}$$

Il est donc nécessaire de connaître le vecteur des futures covariables, Z_{h+1}, afin de trouver le temps du futur événement, t_{j+1}.

L'algorithme itératif de simulation utilisé est le suivant (cf. Figure 3 pour l'illustration et un exemple) :

1. Calculer $\Lambda_l = \Lambda(\tau_l)$, $l = 0, \ldots, k$ du début et de la fin de chaque sous-intervalle où la variable temporelle est constante.
2. $j = 0$
3. Tirer $u_{j+1} \sim \mathcal{U}_{(0,1)}$. Calculer $E_{j+1} = -\frac{\ln(u_{j+1})}{(1+\alpha j)}$ et $\Lambda(t_{j+1}) = \Lambda(t_j) + E_{j+1}$ $(\Lambda(0) = 0)$.
4. Identifier la position de $\Lambda(t_{j+1})$. Si $\Lambda(t_{j+1}) \in [\Lambda(\tau_h),\, \Lambda(\tau_{h+1})[$ alors $t_{j+1} \in [\tau_h,\, \tau_{h+1}[$ et Z_{h+1} est utilisé dans l'inversion de Λ_{j+1}. Ainsi, le temps t_{j+1} peut être calculé.
5. $j = j + 1$
6. Répéter les étapes de 3 à 5 jusqu'à ce que l'on sorte de la période d'observation.

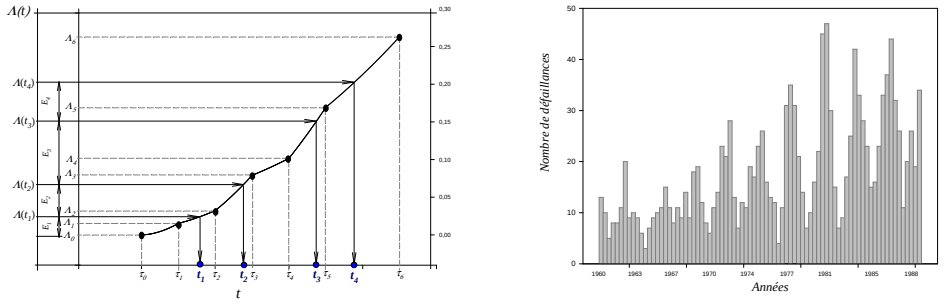

Figure 3: Simulation des données : schéma de l'algorithme et exemple de simulation

4 Analyse par la méthode de Monte-Carlo

Nous avons empiriquement analysé les propriétés de l'estimateur de maximum de vraisemblance pour les modèles suivants :

(a) le modèle (1) en présence d'une variable temporelle constante par morceaux avec $\alpha = 1.5$, $\delta = 1.2$, $\beta = \{-5,\ 0.2\}$, $\tilde{\beta} = 0.5$;

(b) le modèle (4) avec $\alpha = 1.3$, $\delta = 1.2$, $\beta = \{-8,\ 0.4\}$, $\gamma = 0.6$ et $X(t) = \sin(2t)$.

Les résultats de l'étude par simulations appliquée à la modelisation des défaillances de conduites d'eau seront présentés.

Bibliographie

[1] Andersen P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993) *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.
[2] Le Gat, Y. (2009) *Une extension du processus de Yule pour la modélisation stochastique des événements récurrents.* Thèse de doctorat, ENGREF, 2009.
[3] Ross, S. (1983) *Stochastic Processes*. John Wiley and Sons, New York.
[4] Shih, L.-H. et Leemis, L. (1993) Variate Generation for a Nonhomogeneous Poisson Process with Time Dependent Covariates. *Journal of Statistical Computation and Simulation*, **44**(3-4): 165-186.