REVUE DE STATISTIQUE APPLIQUÉE

E. COURILLEAU
J. M. MARION

Comparaison de modèles d'estimation de la fonction de survie appliquée à des données routières

Revue de statistique appliquée, tome 47, nº 1 (1999), p. 81-97
<http://www.numdam.org/item?id=RSA_1999__47_1_81_0>

Rev. Statistique Appliquée, 1999, **XLVII** (1), 81-97

COMPARAISON DE MODÈLES D'ESTIMATION DE LA FONCTION DE SURVIE APPLIQUÉE À DES DONNÉES ROUTIÈRES

E. Courilleau[1], J.M. Marion[2]

[1] *Laboratoire Central des Ponts et Chaussées de Nantes, BP 19, 44340 Bouguenais*

[2] *Institut de Mathématiques Appliquées – U.C.O. 44, rue Rabelais, BP 808, 49008 Angers cedex 01*

RÉSUMÉ

Dans cet article, les lois de survie et le modèle de Cox sont utilisés pour analyser des données sur les dégradations des chaussées. Partant du modèle de Cox, une méthode basée sur des courbes enveloppes permet d'estimer la fonction de survie lorsqu'il y a des données présentant simultanément des censures à droite et à gauche. Une analyse des qualités prédictives sur ces données permet de comparer les méthodes.

Mots-clés : *Modèle de Cox, lois de survie, données censurées, maximum de vraisemblance, méthode de Newton-Raphson, dégradation des chaussées.*

ABSTRACT

In this paper, survival distributions and Cox model are used to analyse road deterioration data. Starting with Cox model, we introduce a method with upper and lower curves to estimate survival function in presence of simultaneous right and left censored data. The different methods are compared with analysis of prediction performance on road data.

Keywords : *Cox model, survival distributions, censored data, maximum likelihood, Newton-Raphson method, road deterioration.*

1. Introduction

Les lois de survie [3,6] et le modèle de Cox [1] peuvent être utilisés dans de nombreux domaines. Nous présentons dans cet article une comparaison de ces deux approches sur des données routières. Le modèle de Cox est principalement développé, et notamment une méthode permettant de prendre en compte simultanément les données censurées à droite et les données censurées à gauche.

L'entretien des chaussées françaises représente de grands enjeux économiques. L'objectif de notre étude est de montrer comment les lois de survie et le modèle de Cox ont été utilisés pour répondre à la problématique de l'évolution des dégradations de chaussées. Ces deux méthodes peuvent prendre en compte différentes variables exogènes comme le trafic ou l'épaisseur des couches de chaussées par exemple. Quand les lois d'évolution des dégradations seront connues, on pourra faire de la gestion pluriannuelle de l'entretien routier. Ces lois permettront de comparer à moyen terme les différentes stratégies d'entretien de chaussées et de choisir celle qui améliore le plus l'état du réseau au moindre coût.

Tout d'abord, le modèle de Cox et la méthode des courbes enveloppes sont présentés ainsi qu'un bref rappel sur l'estimation dans les modèles de survie paramétriques en utilisant la procédure LIFEREG du logiciel SAS [7]. Enfin, les résultats obtenus sur des données routières avec le modèle de Cox et les lois de survie sont comparés.

2. Modèle de Cox

Le modèle de Cox, modèle semi-paramétrique, permet de considérer des fonctions de survie sans leur donner des formes précises tout en introduisant un certain nombre de variables exogènes dans le modèle.

2.1 Théorie

a) Généralités

La fonction de hasard est définie par :

$$h(t) = \frac{f(t)}{S(t)} \tag{1}$$

avec f fonction de densité, S fonction de survie. Puisque $S(0) = 1$, on a :

$$S(t) = \exp\left(-\int_0^t h(u)du\right)$$

Le modèle à hasard proportionnel ou modèle de Cox est défini par :

$$h(t, x) = h_0(t)\exp(\widetilde{x}\beta) \tag{2}$$

où :

– $h(t, x)$, fonction de hasard au temps t,

– $h_0(t)$, fonction de hasard arbitraire non spécifiée,

– $\widetilde{x}$, transposé du vecteur des variables exogènes,

– β, vecteur des paramètres à estimer.

La fonction de densité connaissant x se déduit de (1) :

$$f(t,x) = h_0(t) \cdot \exp(\widetilde{x}\beta) \cdot \exp\left(-\exp(\widetilde{x}\beta) \cdot \int_0^t h_0(u)du\right)$$

et

$$S(t,x) = (S_0(t))^{\exp(\tilde{x}\beta)} \tag{3}$$

Avec

$$S_0(t) = \exp\left(-\int_0^t h_0(u)du\right) \tag{4}$$

Dans un premier temps, l'estimation des paramètres du vecteur β dans le cas où il n'y a pas de données censurées est étudiée.

b) Estimation de β sans données censurées

La méthode de la vraisemblance partielle de Cox est utilisée. Dans le cas de données non censurées, tous les âges de franchissement du phénomène étudié (en l'occurrence, le taux de fissuration des chaussées τ), sont connus. Soit $0 \leqslant \xi_1 < \xi_2 < \cdots < \xi_n$, les âges d'observations de franchissement du taux de fissuration des chaussées τ pour les différents tronçons de l'échantillon. Considérons, sans perte de généralité, que le taux de fissuration τ est franchi par un seul tronçon, noté i, à l'instant ξ_i. Nous définissons l'ensemble $R(t)$ des tronçons qui n'ont pas encore franchi le taux de fissuration τ à l'instant $(t-0)$:

$$R(t) = \{i; \xi_i \geqslant t\}$$

Les éléments de cet ensemble (tronçons de chaussées) sont dits «à risque à l'instant t».

Sachant qu'à l'instant t, le taux de fissuration des chaussées τ est franchi sur un tronçon (t est donc choisi dans l'ensemble des ξ), la probabilité $P_i(t)$ que ce soit sur le tronçon i appartenant à $R(t)$ est donnée par :

$$P_i(t) = \frac{h(t,x_i)}{\sum\limits_{k \in R(t)} h(t,x_k)}$$

puisque $h(t,x)$ s'interprète comme densité conditionnelle du temps d'atteinte de τ,sachant que ce temps est supérieur ou égal à t.

D'après (2), on déduit :

$$P_i(t) = \frac{\exp(\widetilde{x}_i\beta)}{\sum\limits_{k \in R(t)} \exp(\widetilde{x}_k\beta)} \tag{5}$$

La fonction de hasard de base h_0 disparaît dans (5). Pour estimer les paramètres du vecteur β, la fonction de vraisemblance, appelée aussi vraisemblance partielle de Cox est calculée :

$$L(\beta) = \prod_{i=1}^{n} P_i(\xi_i) = \prod_{i=1}^{n} \frac{\exp(\widetilde{x}_i\beta)}{\sum\limits_{k\in R(\xi_i)} \exp(\widetilde{x}_k\beta)}$$

Les estimateurs des composants β_j du vecteur β sont solutions des équations de vraisemblance :

$$\frac{\partial \ln(L(\beta))}{\partial\beta_j} = \sum_{i=1}^{n}\left(x_{ij} - \frac{\sum\limits_{k\in R(\xi_i)} x_{kj}\cdot\exp(\widetilde{x}_k\beta)}{\sum\limits_{k\in R(\xi_i)} \exp(\widetilde{x}_k\beta)}\right) = 0 \quad \text{pour } 1 \leqslant j \leqslant q \quad (6)$$

où

– q, nombre de composantes du vecteur β,

– x_{ij}, $j^{\text{ième}}$ composante du vecteur des variables exogènes pour le tronçon i.

Les solutions en β_j sont calculées par la méthode de Newton-Raphson [6].

Le test du score, par exemple, permet de tester la significativité des β_j.

c) Estimation de β avec des données censurées à droite

La présence de données censurées à droite ne modifie pas la valeur de $L(\beta)$ [2], l'ensemble des éléments à risque étant connu à chaque instant. L'introduction de données censurées à droite ne change pas la valeur de l'estimateur de β.

d) Estimation de la fonction de survie

Les composantes du vecteur β étant maintenant estimées, la fonction de survie de base S_0 définie dans (4) va être calculée. Soit :

– D_j : ensemble des tronçons pour lesquels le taux de fissuration des chaussées étudié a été franchi à l'âge ξ_j,

– C_j : ensemble des tronçons censurés à droite à l'âge ξ_j,

– z : nombre d'âges distincts de franchissement du taux de fissuration des chaussées.

La contribution à la fonction de survie d'un tronçon, ayant x comme vecteur de variables exogènes, et sur lequel le taux de fissuration des chaussées est franchi à l'instant ξ_j est d'après (3) :

$$S(\xi_j, x) - S(\xi_j + 0, x) = (S_0(\xi_j))^{\exp(\tilde{x}\beta)} - (S_0(\xi_j + 0))^{\exp(\tilde{x}\beta)}$$

La contribution d'un tronçon censuré à l'âge ξ_j est :

$$S_0(\xi_j + 0)^{\exp(\tilde{x}\beta)}$$

La fonction de vraisemblance peut donc s'écrire (en posant $\xi_0 = 0$ et $D_0 = \varnothing$) :

$$L(\beta) = \prod_{j=0}^{z} \left\{ \prod_{l \in D_j} \left[S_0(\xi_j)^{\exp(\tilde{x}_l\beta)} - S_0(\xi_j + 0)^{\exp(\tilde{x}_l\beta)} \right] \prod_{l \in C_j} S_0(\xi_j + 0)^{\exp(\tilde{x}_l\beta)} \right\} \tag{7}$$

Comme dans [4], on constate que la solution $S_0(\xi)$ qui maximise $L(\beta)$ est discontinue aux points ξ_j puisque autrement $L(\beta) = 0$.

On est conduit à considérer un modèle discret avec un taux de hasard noté $1 - v_j$ en chaque point ξ_j (v_j est donc la probabilité conditionnelle que le taux de fissuration ne soit pas atteint en ξ_j pour un tronçon à risque en ξ_j).

Alors, l'estimation de la fonction de survie S_0 s'écrit en tout point ξ : $S_0(\xi) = \prod_{j;\xi_j<\xi} \widehat{v}_j$ avec les $\widehat{v}_j$ solutions des équations de vraisemblance suivantes (voir Annexe) :

$$\sum_{l \in D_j} \frac{\exp(\widetilde{x}_l\widehat{\beta})}{1 - v_j^{\exp(\tilde{x}_l\beta)}} = \sum_{l \in R(\xi_j)} \exp(\widetilde{x}_l\widehat{\beta}) \qquad \text{pour } 1 \leqslant j \leqslant z \tag{8}$$

Dès que le nombre de tronçons franchissant le taux de fissuration des chaussées est supérieur à un, il faut trouver une solution itérative à l'équation (8). Pour cela, on calcule :

$$\widehat{v}_j^{\exp(\tilde{x}_l\hat{\beta})} = \exp\left(\exp(\widetilde{x}_l\widehat{\beta}) * \ln(\widehat{v}_j)\right)$$

Puisque $\widehat{v}_j^{\exp(\tilde{x}_l\hat{\beta})} \simeq 1 + \exp(\widetilde{x}_l\widehat{\beta}) * \ln(\widehat{v}_j)$, on obtient :

$$\widehat{v}_j = \exp\left(- \frac{d_j}{\sum_{m \in R(\xi_j)} \exp(\widetilde{x}_m\widehat{\beta})} \right)$$

où d_j représente le cardinal de l'ensemble D_j.

On en déduit une estimation de la fonction de survie :

$$S(\xi, x) = \prod_{j/\xi_j<\xi} \widehat{v}_j^{\exp(\tilde{x}\hat{\beta})} \tag{9}$$

Ceci permet de calculer la fonction de survie, et donc la fonction de hasard, dans le cas où les données peuvent être censurées à droite.

e) Cas des données censurées à gauche

L'introduction de données censurées à gauche [5] entraîne quelques difficultés. En effet, il n'est pas possible de connaître l'ensemble à risque à tout instant ξ. Pour résoudre ce problème, il suffit d'inverser l'échelle de temps [9], [10]. La borne supérieure est fixée à Max et on calcule pour chaque tronçon $s = \text{Max} - \xi$, on se retrouve ainsi dans le cas de données censurées à droite. La valeur des estimations des composantes du vecteur β est obtenue par les équations (6). On en déduit ainsi $S(s, x)$:

$$S(s,x) = \prod_{j/s_j<s} \widehat{v_j}^{\exp(\tilde{x}\hat{\beta})}$$

La fonction de survie de départ se retrouve en calculant :

$$S(\max - s, x) = \prod_{j/s_j<\text{Max}-s} \widehat{v_j}^{\exp(\tilde{x}\hat{\beta})}$$

d'où : $S(\xi, x) = \prod_{j/\text{Max}-\xi_j<\xi} \widehat{v_j}^{\exp(\tilde{x}\hat{\beta})}$.

Les paragraphes précédents montrent que l'on peut calculer les fonctions de survie dans le cas où il y a des données censurées à droite et dans le cas où il y a des données censurées à gauche. Il faut déterminer une méthode permettant de prendre en compte à la fois les données censurées à gauche et les données censurées à droite.

2.2 Méthode développée avec les données censurées à droite et à gauche

Une méthode permettant de tracer des courbes enveloppes de la fonction de survie est proposée.

Une étude, notée *env* sup, a été réalisée en ne prenant que les tronçons non censurés et les tronçons censurés à droite. Une seconde étude, notée *env* inf, a été réalisée en ne prenant que les tronçons non censurés et les tronçons censurés à gauche. Pour *env* inf, il faut inverser l'échelle des temps afin de pouvoir utiliser le modèle de Cox. Ceci permet de résoudre le problème d'ensemble à risque dans le cas des données censurées à gauche.

Pour chaque taux de fissuration des chaussées étudié, la courbe de survie de l'ensemble de l'échantillon doit se trouver entre les courbes obtenues pour *env* sup et *env* inf. Pour cela, il suffit que la fonction de survie soit croissante avec les données censurées à droite et décroissante avec les données censurées à gauche. Soit un échantillon, ayant comme fonction de survie $S_{\text{échan}}$ définie comme (*cf.* (9)) :

$$S_{\text{échan}}(\xi, x) = \prod_{j/\xi_j<\xi} \widehat{v}_j^{\exp(\tilde{x}\hat{\beta})} = \prod_{j/\xi_j<\xi} \exp\left(-\frac{d_j}{\sum_{m\in R(\xi_j)} \exp(\widetilde{x}_m\widehat{\beta})}\right)^{\exp(\tilde{x}\hat{\beta})} \tag{10}$$

Posons :

$$A = \prod_{j/\xi_c<\xi_j<\xi} \exp\left(-\frac{d_j}{\sum_{m\in R(\xi_j)} \exp(\widetilde{x}_m\widehat{\beta})}\right)^{\exp(\tilde{x}\hat{\beta})}$$

Supposons que l'on ajoute à cet échantillon, un tronçon censuré à droite à l'instant ξ_c inférieur à ξ, on obtient une nouvelle fonction de survie S_1 :

$$S_1(\xi, x) = \prod_{j/\xi_j<\xi_c} \exp\left(-\frac{d_j}{\sum_{m\in R(\xi_j)} \exp(\widetilde{x}_m\widehat{\beta}) + \exp(\widetilde{x}_{\xi_c}\widehat{\beta})}\right)^{\exp(\tilde{x}\hat{\beta})}$$
$$\cdot\exp\left(-\frac{d_{\xi_c}}{\sum_{m\in R(\xi_c)} \exp(\widetilde{x}_m\widehat{\beta}) + \exp(\widetilde{x}_{\xi_c}\widehat{\beta})}\right)^{\exp(\tilde{x}\hat{\beta})} \cdot A$$

La fonction $S_{\text{échan}}$ (10) peut s'écrire :

$$S_{\text{échan}}(\xi, x) = \prod_{j/\xi_j<\xi_c} \exp\left(-\frac{d_j}{\sum_{m\in R(\xi_j)} \exp(\widetilde{x}_m\widehat{\beta})}\right)^{\exp(\tilde{x}\hat{\beta})}$$
$$\cdot\exp\left(-\frac{d_{\xi_c}}{\sum_{m\in R(\xi_c)} \exp(\widetilde{x}_m\widehat{\beta})}\right)^{\exp(\tilde{x}\hat{\beta})} \cdot A$$

La fonction $S_{\text{échan}}$ est donc inférieure à la fonction S_1 quel que soit ξ. La fonction de survie est donc croissante avec les données censurées à droite. Une étude analogue permet de montrer que la fonction de survie est décroissante avec les données censurées à gauche.

Dans notre étude, la fonction de survie de *env* sup constitue donc une borne supérieure de la fonction de l'échantillon total. À l'inverse, *env* inf constitue une borne inférieure pour la fonction de survie de l'échantillon total.

La fonction de survie de l'échantillon est donc comprise entre les fonctions de survie de *env* inf et celles de *env* sup.

Dans le cas où les courbes obtenues pour *env* inf et *env* sup sont suffisamment proches, le problème sera résolu. La procédure PHREG [8] du logiciel SAS est utilisée pour la détermination des paramètres. La courbe de survie dépend du vecteur de variables exogènes.

Il faut maintenant déterminer une loi de probabilité qui permette de lisser les fonctions de survie empiriques obtenues pour *env* inf et *env* sup. D'après (3) et un vecteur de variables exogènes choisi, tel que $\exp(\widehat{\widetilde{x}\beta}) = 1$, les fonctions $S_0^{env\,\inf}$ et $S_0^{env\,\sup}$ peuvent être estimées. Il sera possible d'obtenir les fonctions de survie pour un autre vecteur de variables exogènes.

Les fonctions de survie étant déterminées pour les deux études *env* sup et *env* inf, les courbes enveloppes de la fonction de survie de l'échantillon global peuvent être tracées. La fonction de survie de l'échantillon total est pondérée par les effectifs.

La fonction de survie S_0 de l'échantillon total est définie comme :

$$S_0(\xi) = \frac{n_{env\,\inf} \cdot S_0^{env\,\inf}(\xi) + n_{env\,\sup} \cdot S_0^{env\,\sup}(\xi)}{n_{env\,\inf} + n_{env\,\sup}} \tag{11}$$

où $n_{env\,\inf}$ et $n_{env\,\sup}$ représentent les effectifs ainsi que $S^{env\,\inf}$ et $S^{env\,\sup}$ les fonctions de survie associées à chaque étude *env* inf et *env* sup. Dans le cas où $\exp(\widehat{\widetilde{x}\beta}) \neq 1$, on ne peut pas utiliser la relation (11) pour la somme pondérée car la valeur des paramètres de β est inconnue pour la totalité de notre échantillon. La procédure suivante est adoptée :

$$S(\xi) = \frac{n_{env\,\inf} \cdot S^{env\,\inf}(\xi) + n_{env\,\sup} \cdot S^{env\,\sup}(\xi)}{n_{env\,\inf} + n_{env\,\sup}}$$

d'où :

$$S(\xi, x) = \frac{n_{env\,\inf} \cdot (S_0^{env\,\inf}(\xi))^{\exp(\tilde{x}\beta_{env\,\inf})} + n_{env\,\sup} \cdot (S_0^{env\,\sup}(\xi))^{\exp(\widetilde{x\beta}_{env\,\sup})}}{n_{env\,\inf} + n_{env\,\sup}} \tag{12}$$

On peut donc construire pour n'importe quel vecteur de variables exogènes les fonctions de survie pour chaque taux de fissuration des chaussées étudié.

3. Modèle de survie paramétrique

Dans un tel modèle, la fonction de hasard s'exprime de façon paramétrique en fonction du temps, faisant éventuellement intervenir des variables exogènes.

Soit T la variable aléatoire indiquant l'âge d'apparition du phénomène étudié (ici, le franchissement d'un taux de fissuration des chaussées noté τ). La loi de probabilité de T a une distribution de type connu, fonction d'un paramètre vectoriel θ.

Soit ξ_i l'instant d'observation du $i$ème tronçon $(1 \leqslant i \leqslant n)$, nous noterons :

D_{1_i} l'indicatrice de l'événement «franchissement du taux de fissuration τ avant ξ_i» (censure à gauche).

D_{2_i} l'indicatrice de l'événement «franchissement du taux de fissuration τ après ξ_i» (censure à droite).

D_{3_i} l'indicatrice de l'événement «franchissement du taux de fissuration τ lors de l'observation ξ_i».

La fonction de vraisemblance, correspondant à l'échantillon des instants d'observations $\xi_1, \xi_2, \cdots, \xi_n$ des n tronçons, s'écrit alors :

$$L(\theta) = \prod_{i=1}^{n} \left\{ F(\xi_i)^{D_{1i}} S(\xi_i)^{D_{2i}} f(\xi_i)^{D_{3i}} \right\}$$

avec F, S et f respectivement la fonction de répartition, la fonction de survie et la densité de probabilité de la distribution de T.

On obtient les estimateurs des composantes θ_j du vecteur θ comme solutions des équations de vraisemblance :

$$\frac{\partial \ln(L(\theta))}{\partial \theta_j} = 0 \qquad \text{pour } 1 \leqslant j \leqslant p$$

Quand le système des équations ci-dessus est un système d'équations linéaires, les paramètres du vecteur θ sont déterminés analytiquement. Dans le cas d'un système d'équations non linéaires, une méthode itérative d'optimisation, par exemple la méthode de Newton-Raphson [6] est utilisée pour la détermination des paramètres du vecteur θ.

La procédure LIFEREG de SAS [7] a été utilisée ici, elle consiste à ajuster un modèle paramétrique sur des données qui peuvent être censurées.

Le modèle de départ est de la forme :

$$y_i = \widetilde{x}_i \beta + \eta \varepsilon_i \qquad \text{pour } 1 \leqslant i \leqslant n$$

avec :

y_i représentant le logarithme de l'âge d'observation du franchissement du phénomène étudié.

$\widetilde{x}_i$ est le transposé du vecteur des variables exogènes du tronçon i.

β est le vecteur des paramètres à estimer.

η est un paramètre d' échelle, par défaut sa valeur est prise égale à 1.

ε_i est une variable aléatoire dont la distribution est connue (exponentielle, Weibull, loglogistique...)

Pour une estimation des valeurs initiales des paramètres η et β, la procédure LIFEREG utilise la méthode des moindres carrés en ignorant les données censurées.

À partir de ces valeurs initiales, on calcule la log-vraisemblance associée à l'échantillon des instants d'observations et on obtient :

$$\ln(L) = \sum_{i=1}^{n} \{D_{1_i} \ln(F_\varepsilon(\zeta_i)) + D_{2_i} \ln(S_\varepsilon(\zeta_i)) + D_{3_i} \ln(f_\varepsilon(\zeta_i))\}$$

avec $\zeta_i = \dfrac{\ln \xi_i - \widetilde{x}_i\beta}{\eta}$ et $F_\varepsilon, S_\varepsilon, f_\varepsilon$ désignant respectivement la fonction de répartition, la fonction de survie et la densité de probabilité des ε_i.

Les paramètres η, β et ceux de la loi des ε_i sont estimés par la méthode du maximum de vraisemblance et la méthode de Newton-Raphson.

Connaissant la loi des ε_i, on en déduit la fonction de survie de T correspondant à l'âge de franchissement du phénomène étudié.

4. Application – Comparaison modèles de Cox et lois de survie

4.1. Échantillon et modèles

L'échantillon utilisé pour la comparaison entre les méthodes statistiques est issu d'une base de données constituée d'environ 700 tronçons de sections de routes nationales. Ces tronçons, dont la longueur est de 50 mètres, pouvant être relevés plusieurs fois, forment un ensemble de 3919 relevés. Pour chaque relevé, plusieurs variables exogènes sont disponibles.

Trois variables exogènes ont été utilisées dans cette étude :

– Le trafic, qui est pris en compte par la variable :

$$\text{trafic} = \frac{\text{tr}_{\text{national}} - \text{tr}_{\text{tronçon}}}{\text{tr}_{\text{national}}}$$

où $\text{tr}_{\text{tronçon}}$ représente la progression du trafic sur un tronçon donné et $\text{tr}_{\text{national}}$ la progression moyenne du trafic sur l'ensemble des tronçons de la base. Sur la base de données étudiée, cette variable trafic prend ses valeurs dans l'intervalle $[-3; +3]$.

– La technique de construction, définie par une variable discrète à deux modalités. Par la suite, nous donnerons la valeur 0 aux tronçons en grave ciment et la valeur 1 aux tronçons en grave laitier.

– L'«épaisseur équivalente» prenant en compte entre autre l'épaisseur des couches des chaussées [11].

Les méthodes utilisées vont être illustrées dans le cas de la fissuration transversale sur des chaussées «semi-rigides» (ce sont des chaussées qui, de par leur composition, ont une faible capacité à se déformer mais sont sujettes à des fissurations transversales sous l'effet de l'accumulation de contraintes provoquées par le trafic des poids lourds, la nature et la qualité des matériaux traités, le collage entre la couche de roulement et l'assise...)

Le taux de fissuration τ est défini relativement au fait que le nombre maximal potentiel de fissurations transversales sur un tronçon de cinquante mètres est de vingt et un.

Par exemple, dans l'échantillon pour un taux de fissuration de 5% on dénombre : 146 données non censurées, 489 censurées à gauche, 165 censurées à droite.

Pour la recherche des *modèles de survie paramétrique*, La procédure LIFEREG de SAS [7] a été utilisée. Plusieurs lois ont ainsi été testées : la loi exponentielle, la loi de Weibull, la loi lognormale, la loi loglogistique... Les lois de survie permettant d'introduire des variables exogènes, préalablement définies dans les modèles. En utilisant le critère de la vraisemblance maximale, la loi loglogistique est apparue comme étant la mieux adaptée pour la modélisation de l'évolution de la fissuration des chaussées. Elle a donc été choisie.

La figure 1 montre un exemple de courbe de survie obtenue pour le taux de fissuration des chaussées de 5% pour un vecteur de variables exogènes donné (trafic $= -1$; technique $= 1$; épaisseur $= 45,17$).

Le vecteur estimé $\widehat{\beta}$ ayant comme composantes : ($\widehat{\beta}_1 = 0.551$; $\widehat{\beta}_2 = 0.458$; $\widehat{\beta}_3 = -0.009$)

La loi loglogistique étant définie par sa densité $f(t) = \dfrac{\alpha\gamma t^{\gamma-1}}{(1+\alpha t^{\gamma})^2}$ $\quad (t \geqslant 0)$ avec comme estimation des paramètres : $\widehat{\gamma} = 2.125$; $\widehat{\alpha} = 0.034$.

Pour la recherche *du modèle de Cox*, la pocédure PHREG de SAS [8] a été utilisée pour la détermination des paramètres avec les variables exogènes définies précédemment.

La figure 2 présente un exemple de courbes enveloppes de la fonction de survie ainsi que la courbe de survie pondérée par les effectifs, dans le cas où le taux de fissuration des chaussées est de 5% pour un vecteur de variables exogènes (trafic = 1; technique = 0; épaisseur = 96).

Les composantes estimées du vecteur β étant données dans le tableau suivant :

TABLEAU 1

Composantes du vecteur β obtenu par le modèle de Cox.

	Trafic	Technique	Epaisseur
$\widehat{\beta}$ *env* sup.	0.909	−0.078	−0.082
$\widehat{\beta}$ *env* inf.	0.042	−1.107	0.014

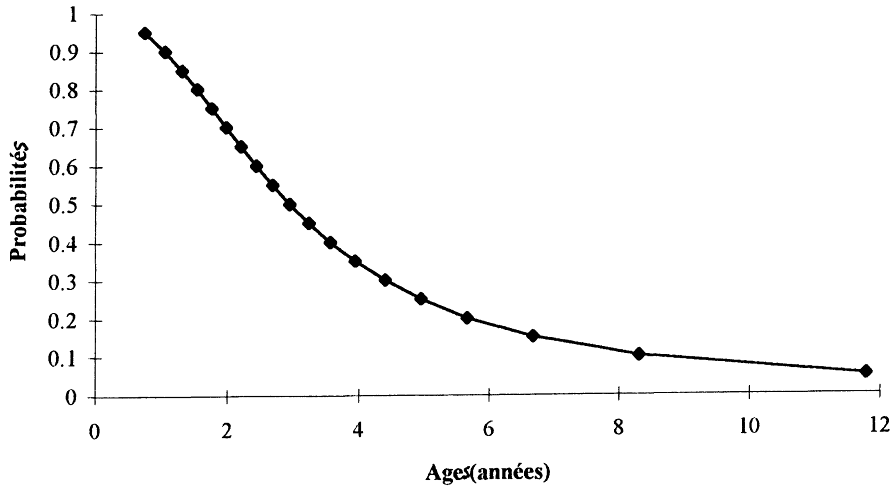

FIGURE 1
Exemple de courbe de survie loglogistique.

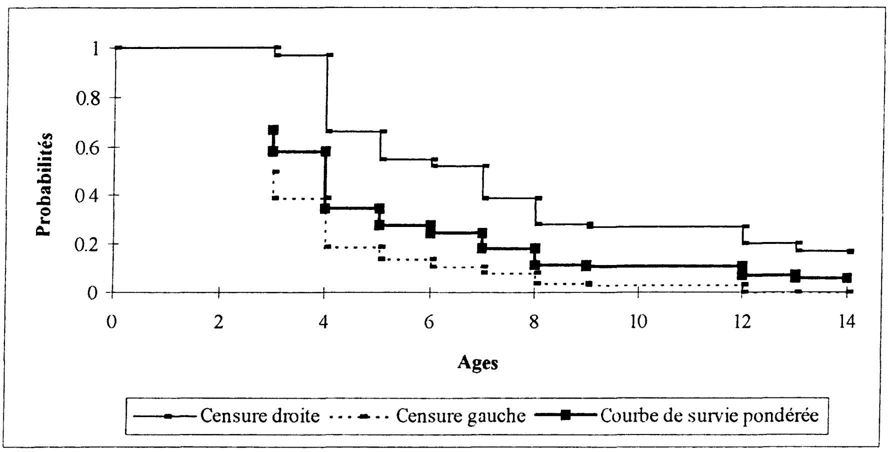

FIGURE 2
Exemple de courbes enveloppes.

Pour lisser les courbes de survie empiriques, plusieurs lois de probabilité ont été testées (Weibull, lognormale, log logistique...). On a donc estimé les paramètres de ces différentes lois pour chaque taux de fissuration étudié et pour les courbes envinf et envsup. Pour cela, la procédure NLIN du logiciel SAS [7] a été utilisée.

Les meilleurs résultats, au sens différence entre observations et valeurs ajustées sont obtenus avec la loi log logistique.

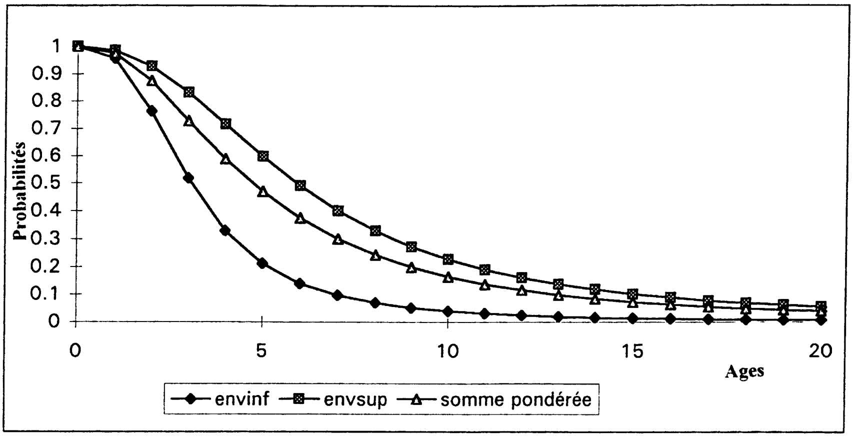

FIGURE 3
Exemple de lissage des courbes enveloppes.

La figure 3 présente les courbes qui lissent les courbes de survie enveloppes et la courbe de survie pondérée.

4.2. Comparaison entre les modèles

Pour comparer les lois d'évolutions obtenues avec la méthode des courbes enveloppes et celles obtenues avec les lois de survie, l'écart quadratique moyen entre observations et estimations a été choisi.

Ce critère a été calculé pour les prévisions à un an, à deux ans, à trois ans, à quatre ans, à cinq ans, à dix ans pour la population totale et les prévisions sur l'ensemble de la population à partir de la première observation.

a) Calcul des écarts moyens pour les prévisions à 1, 2, 3, 4, 5 et 10 ans

Pour un tronçon donné, observé à l'âge n, la courbe d'évolution du taux de fissuration des chaussées est calculée par la méthode des lois de survie puis par la méthode des courbes enveloppes. Le taux de fissuration des chaussées *observé* à l'âge $n+1$ (resp. $n+2$, $n+3$, $n+4$, $n+5$ et $n+10$) sur ce tronçon est comparé avec le taux de fissuration des chaussées *estimé* par la courbe d'évolution, aux mêmes âges.

Soit un tronçon donné, observé plusieurs fois, par exemple :

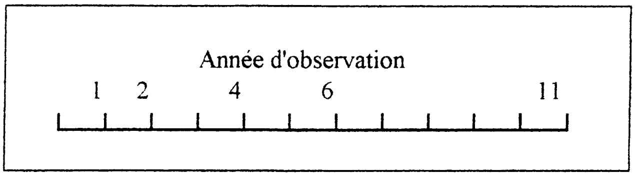

FIGURE 4
Âge d'observation d'un tronçon.

Pour calculer les prévisions à un an, le taux de fissuration des chaussées estimé à deux ans à partir de l'observation à un an est comparé avec le taux de fissuration des chaussées observé à deux ans. Pour les prévisions à deux ans, le taux estimé à quatre ans à partir de l'observation à deux ans, est comparé avec le taux observé à quatre ans, et ainsi de suite.

Les écarts moyens obtenus par les différentes méthodes, pour les différentes durées de prévisions, sont donnés dans le tableau 2 :

TABLEAU 2
Ecarts moyens entre observations et estimations

	Durée de la prévision	**1 an**	**2 ans**	**3 ans**	**4 ans**	**5 ans**	**10 ans**
	Effectif	1487	1596	1134	954	598	96
Lois de survie	EM	38.2	51.8	63.5	52.8	79.6	219.6
Courbes enveloppes	EM	66.1	75.8	82.4	95.6	134.1	224.4

La méthode des courbes enveloppes donne de moins bons résultats au niveau de l'écart moyen pour l'ensemble des prévisions, même si les écarts sont relativement faibles. La taille modeste de l'échantillon peut expliquer la meilleure performance du modèle paramétrique par rapport au modèle semi-paramétrique.

b) Calcul des prévisions à partir de la première observation de chaque tronçon

Une étude a été réalisée en déterminant les paramètres des lois d'évolution à partir de la première observation de chaque tronçon. Nous avons ensuite comparé les écarts entre les estimations du taux de fissuration des chaussées obtenues par les deux méthodes et le taux de fissuration des chaussées effectivement observé pour toutes les observations suivantes.

Pour l'exemple de la figure 2, les taux de fissuration des chaussées à deux, quatre, six et onze ans estimés à partir de l'observation à un an sont comparés aux taux de fissuration des chaussées observés à deux, quatre, six et onze ans.

Le tableau 3 représente l'écart moyen entre le taux de fissuration observé et le taux de fissuration estimé :

TABLEAU 3
Ecarts moyens entre observations et estimations

		Total
Lois de survie	Ecarts moyens	44.5
Courbes enveloppes	Ecarts moyens	106.5

La méthode des lois de survie est la meilleure des deux méthodes pour les prévisions à partir de la première observation. Cependant les écarts entre les deux méthodes ne sont pas très importants. Un meilleur système de pondération pour les courbes enveloppes pourrait améliorer les résultats. De même, un échantillon plus important avantagerait la méthode semi-paramétrique.

5. Conclusion

Cette étude illustre une utilisation du modèle de Cox. Ce dernier permet de traiter indépendamment les individus censurés à gauche et les individus censurés à droite, ce qui était le cas pour les données routières. Pour résoudre ce problème, la méthode des courbes enveloppes a été développée. La taille de l'échantillon étant relativement faible, ceci permet de justifier l'avantage des lois de survie sur une méthode semi-paramétrique du type modèle de Cox.

6. Annexe
Estimation de la fonction de survie

Dans la recherche de la solution $S_0(\xi)$ qui maximise $L(\beta)$ avec :

$$L(\beta) = \prod_{j=0}^{z} \left\{ \prod_{l \in D_j} \left[S_0(\xi_j)^{\exp(\tilde{x}_l \beta)} - S_0(\xi_j + 0)^{\exp(\tilde{x}_l \beta)} \right] \prod_{l \in C_j} S_0(\xi_j + 0)^{\exp(\tilde{x}_l \beta)} \right\} \quad \text{(cf (7))}$$

nous sommes conduits à considérer $S_0(\xi) = \prod_{j; \xi_j < \xi} v_j$.

Pour $1 \leqslant j \leqslant z$ on pose $S_0(\xi_j) = S_0(\xi_{j-1} + 0) = \prod_{m=0}^{m=j-1} v_m$ avec $v_0 = 1$.

En substituant dans (7), et en remarquant que $S_0(\xi_0) = S_0(0) = 1$

$$\begin{aligned} L(\beta) &= \prod_{j=1}^{z} \left\{ \prod_{l \in D_j} \left(\prod_{m=0}^{j-1} v_m^{\exp(\tilde{x}_l \beta)} - \prod_{m=0}^{j} v_m^{\exp(\tilde{x}_l \beta)} \right) \prod_{l \in C_j} \left(\prod_{m=0}^{j} v_m^{\exp(\tilde{x}_l \beta)} \right) \right\} \\ &= \prod_{j=1}^{z} \left\{ \prod_{l \in D_j} \left(\prod_{m=0}^{j-1} v_m^{\exp(\tilde{x}_l \beta)} (1 - v_j^{\exp(\tilde{x}_l \beta)}) \right) \prod_{l \in C_j} \left(\prod_{m=0}^{j} v_m^{\exp(\tilde{x}_l \beta)} \right) \right\} \end{aligned}$$

Le terme en v_j s'écrit :

$$\prod_{l \in D_j} \left[1 - v_j^{\exp(\tilde{x}_l \beta)} \right] \prod_{l \in C_j} v_j^{\exp(\tilde{x}_l \beta)} \prod_{k > j} \left(\prod_{l \in D_k} v_j^{\exp(\tilde{x}_l \beta)} \prod_{l \in C_k} v_j^{\exp(\tilde{x}_l \beta)} \right)$$

D'où :

$$L(\beta)=\prod_{j=1}^{z}\left[\prod_{l\in D_j}\left(1-v_j^{\exp(\tilde{x}_l\beta)}\right)\prod_{l\in C_j}v_j^{\exp(\tilde{x}_l\beta)}\prod_{l\in\bigcup\limits_{k>j}(D_k\cup C_k)}v_j^{\exp(\tilde{x}_l\beta)}\right]$$

Soit,

$$L(\beta)=\prod_{j=1}^{z}\left[\prod_{l\in D_j}\left(1-v_j^{\exp(\tilde{x}_l\beta)}\right)\prod_{m\in R(\xi_j)-D_j}v_j^{\exp(\tilde{x}_m\beta)}\right]$$

avec : $R(\xi_j)=\bigcup\limits_{k\geqslant j}(C_k\cup D_k)$ ensemble à risque à ξ_j.

On remplace β par son estimateur $\widehat{\beta}$ obtenu à partir des équations (6), puis on recherche les estimateurs des v_j solutions des équations de vraisemblance :

$$\frac{\partial\ln(L(\widehat{\beta}))}{\partial v_j}=0\qquad\text{pour } 1\leqslant j\leqslant z$$

Donc :

$$\sum_{l\in D_j}\frac{\exp(\widetilde{x}_l\widehat{\beta})v_j^{\exp(\tilde{x}_l\hat{\beta})}}{1-v_j^{\exp(\tilde{x}_l\hat{\beta})}}=\sum_{m\in R(\xi_j)}\exp(\widetilde{x}_m\widehat{\beta})-\sum_{m\in D_j}\exp(\widetilde{x}_m\widehat{\beta})\quad\text{pour } 1\leqslant j\leqslant z$$

Soit :

$$\sum_{l\in D_j}\frac{\exp(\widetilde{x}_l\widehat{\beta})}{1-v_j^{\exp(\tilde{x}_l\hat{\beta})}}=\sum_{l\in R(\xi_j)}\exp(\widetilde{x}_l\widehat{\beta})\tag{13}$$

On en déduit une estimation de la fonction de survie :

$$S(\xi,x)=\prod_{j;\xi_j<\xi}\widehat{v}_j^{\exp(\tilde{x}\hat{\beta})}$$

avec $\widehat{v}_j$ solution de (13).

7. Références

[1] COX D.R. (1972), *Regression models and life-tables*, J. Roy. Statist. Soc. B 34, pp 187–220.

[2] COX D.R et OAKES D. (1984), *Analysis of Survival Data*, London, Edition Chapman and Hall.

[3] DROESBEKE J.J., FICHET B., TASSI P., éditeurs. (1989), *Analyse statistique de durées de vie : Modélisation des données censurées*, Economica.

[4] KALBFLEISH J.D. et PRENTICE R.L. (1980), *The statistical analysis of failure time data*, New York : Wiley and Sons, Inc.

[5] KAPLAN E.L. et MEIER P. (1958), *Non parametric estimation from incomplete observations*, J. Amer. Statist. Assoc. 53, pp 457–481.

[6] LAWLESS J.F. (1982), *Statistical Models and Methods for lifetime Data*, Edition Wiley.

[7] SAS/STAT, (1989), User's guide, version 6, fourth edition, Volume 1 et 2, Cary, NC : SAS Institute Inc.

[8] SAS Technical Report P-229, (1992), SAS/STAT Software : Changes and enhancements, Release 6.07, Cary, NC : SAS Institute Inc. The PHREG procedure pp 433–479

[9] TURNBULL B.W. (1974), *Non parametric estimation of a survivorship function with doubly censored data*, J. Amer. Statist. Assoc. 69, pp 169–173.

[10] TURNBULL B.W. (1976), *The empirical distribution function with arbitrarily grouped, censored and truncated data*, J. Roy. Statist. Soc. B 38, pp 290–295.

[11] ULLIDTZ P. (1987), *Pavement analysis*, Elsevier.