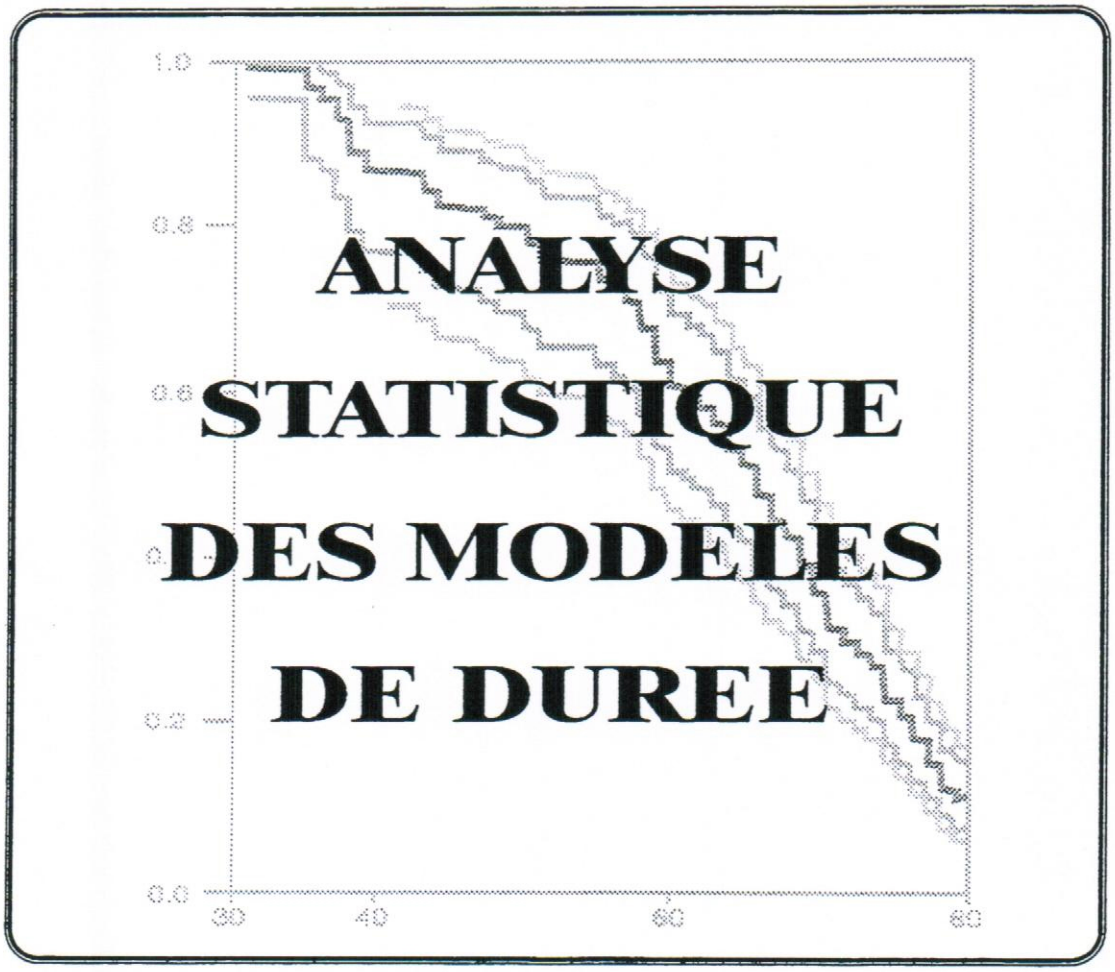
Université de Marne-la-Vallée

Equipe d'Analyse et de Mathématiques Appliquées

Rencontres Franco-Belges de Statisticiens

XVII^{ème} Rencontre Franco-Belge de Statisticiens



ANALYSE STATISTIQUE DES MODELES DE DUREE

Equipe d'Analyse et de Mathématiques Appliquées
UNIVERSITE DE MARNE-LA-VALLEE
21 et 22 Novembre 1996

LISTE DES POSTERS

Applications médicales

Applications économiques

Applications industrielles

Trois autres posters, portants sur des résultats théoriques et non sur des applications, ont été par ailleurs présentés lors de la rencontre.

M. BOUATOU (Université Joseph Fourier, Grenoble)
Ondelettes et analyse de survie.

D. DROUET-MARI (Université Bretagne-Sud, IUT de Vannes)
Mesures d'association - Tests d'indépendance en survie bivariée.

E. OULD-SAID (Université du Littoral, Calais)
On the regression function estimation under random censoring.

INFORMATION ET VALEURS PRONOSTIQUES DE CERTAINES VARIABLES DANS LE CANCER DU SEIN

Daniel HANS* Jérome MANUCEAU**
Marylène TROUPÉ** Jean VAILLANT**
* Centre médical Val Redon
317 Bd du Redon
13009 Marseille

**Université des Antilles-Guyane
UFR Sciences exactes et naturelles-Campus de Fouillole
Département de Mathématiques-Informatique
97159 Pointe-à-Pitre cedex

Abstract: In the context of survival analysis, mutual information theory between σ-algebras can be used for selecting qualitative or quantitative cavariables with a strong predictive value. The information rate carried out by covariables can be tested by means of a decomposition similar to the analysis of variance. Statistical units with some missing data are not excluded from the analysis so that a loss of information is avoided. These results are applied to survival data from 1304 patients with breast cancer followed over a period of ten years.

Résumé: Dans le cadre de l'analyse de durées de vie, la notion d'information mutuelle entre σ-algèbres est utilisée pour sélectionner un ensemble de covariables qualitatives ou quantitatives à fortes valeurs prédictives. Le taux d'incertitude expliqué est testé sur la base d'une décomposition similaire [5] à celle de l'analyse de variance. Les données manquantes sur certaines covariables ne conduisent pas à l'élimination des unités statistiques concernées. On évite ainsi la perte d'information qui en découlerait. Ces résultats sont appliqués à l'étude de la durée de vie de 1304 malades atteints du cancer du sein suivis sur une période de 10 ans. La technique utilisée permet de trouver un nombre limité de facteurs ayant une valeur prédictive intéressante.

1 Introduction

La notion d'entropie d'une loi de probabilité P sur un espace Ω est un outil intéressant pour caractériser l'incertitude associée à P [6]. Quand Ω est dénombrable, l'entropie de P notée $H(P)$ est définie par :

$$H(P) = -\sum_{\omega\in\Omega} P(\omega)Ln(P(\omega))$$

avec la convention $0\mathrm{Ln}0 = 0$ qui revient, en fait à écrire

$$H(P) = -\sum_{\omega\in\Omega'} P(\omega)Ln(P(\omega))$$

où $\Omega' = \{\omega \in \Omega \mid P(\omega) > 0\}$.

Pour une variable aléatoire X sur Ω dénombrable, on définit alors son entropie notée $H(X)$ par celle de sa loi image P_X :

$$H(X) = -\sum_{x \in X(\Omega)} P\Big(X^{-1}(\{x\})\Big) Ln\Big(P(X^{-1}(\{x\}))\Big).$$

$H(X)$ mesure l'incertitude sur la valeur de $X(\omega)$ pour un choix (ou réalisation) d'un élément ω de Ω selon la loi P. Ainsi, si $\mathcal{A} = (A_i)_{i \in I}$ est une partition de Ω, on définit l'entropie de cette partition par $H(\mathcal{A}) = H(X_{\mathcal{A}})$ où $X_{\mathcal{A}}$ est la variable aléatoire définie de la façon suivante :

$$\begin{array}{rccc} X_{\mathcal{A}} : & \Omega & \longrightarrow & \{0,1\}^I \\ & \omega & \longmapsto & \big(\mathbb{I}_{A_i}(\omega)\big)_{i \in I} \end{array}$$

Par conséquent,

$$H(X_{\mathcal{A}}) = -\sum_{i \in I} P(A_i) Ln(P(A_i)).$$

$H(X_{\mathcal{A}})$ mesure l'incertitude sur l'appartenance à l'une des parties A_i avant l'observation d'un élément ω de Ω.

Si on observe deux variables aléatoires X et Y sur Ω, l'entropie de X conditionnellement à une valeur y prise par Y est l'entropie de la loi de X conditionnelle à l'événement $Y = y$, que l'on note $H(X \mid Y = y)$. On définit alors l'entropie de X conditionnelle à Y que l'on note $H(X \mid Y)$ par

$$H(X \mid Y) = \sum_{y \in Y(\Omega)} P(Y = y) H(X \mid Y = y)$$

qui n'est autre que l'espérance de la fonction aléatoire qui à y fait correspondre $H(X \mid Y = y)$.

Si on considère maintenant une loi de probabilité P sur Ω cette fois non dénombrable, quelques précautions doivent être prises pour la généralisation formelle de la notion d'entropie vue précédemment. Ceci conduit à l'entropie dite généralisée $H(P;\mu)$ relative à une mesure de référence μ sur Ω ([2],[7]) et à l'entropie d'une σ-algèbre $\mathcal{A}$ de type séparable [1] que l'on notera par la suite $H(\mathcal{A})$. Le lien formel entre entropie de Shannon et entropie de Kullback est donné, par exemple, dans [7] ainsi que leurs propriétés respectives.

2 Information entre σ-algèbres

Soient $\mathcal{A}$ et $\mathcal{B}$ deux σ-algèbres, on note $\mathcal{A} \vee \mathcal{B} = \sigma(\mathcal{A} \cup \mathcal{B})$ la σ-algèbre engendrée par $\mathcal{A} \cup \mathcal{B}$.

Soient $\mathcal{A}$ et $\mathcal{B}$ deux σ-algèbres de type séparable, On définit l'entropie de $\mathcal{A}$ conditionnelle à $\mathcal{B}$ par

$$H(\mathcal{A} \mid \mathcal{B}) = H(\mathcal{A} \vee \mathcal{B}) - H(\mathcal{B}).$$

C'est l'incertitude restant sur $\mathcal{A}$ quand on connait $\mathcal{B}$.

On rappelle les propriétés suivantes

1) $\mathcal{A} \subset \mathcal{B} \Rightarrow H(\mathcal{A}) \leq H(\mathcal{B})$.

2) $H(\mathcal{A} \vee \mathcal{B}) \leq H(\mathcal{A}) + H(\mathcal{B})$ et $H(\mathcal{A} \mid \mathcal{B}) \leq H(\mathcal{A})$.

Si $\mathcal{A}$ et $\mathcal{B}$ sont d'entropies finies, on a l'égalité dans les expressions précédentes si et seulement si $\mathcal{A}$ et $\mathcal{B}$ sont indépendants (On écrit alors $\mathcal{A} \perp \mathcal{B}$.).

Pour une bonne quantification $I(\mathcal{A}, \mathcal{B})$ de l'information entre σ-algèbres $\mathcal{A}$ et $\mathcal{B}$, il est nécessaire de poser en postulats quelques propriétés exigibles de $I(\mathcal{A}, \mathcal{B})$. Il nous parait souhaitable que $I(\mathcal{A}, \mathcal{B})$ vérifie les postulats suivants :

1) $\mathcal{A} \perp \mathcal{B} \quad \Leftrightarrow I(\mathcal{A}, \mathcal{B}) = 0$.
2) $I(\mathcal{A}, \mathcal{B}) = I(\mathcal{B}, \mathcal{A})$.
3) $\mathcal{A} \subset \mathcal{C}) \Rightarrow I(\mathcal{A}, \mathcal{B}) \leq I(\mathcal{C}, \mathcal{B})$.

En choisissant $I(\mathcal{A}, \mathcal{B}) = H(\mathcal{A}) + H(\mathcal{B}) - H(\mathcal{A} \vee \mathcal{B})$, tous ces postulats sont vérifiés et on a, en plus, $I(\mathcal{A}, \mathcal{A}) = H(\mathcal{A})$.

Quand on n'a pas de modèles paramétriques suffisamment raisonnables par rapport au phénomène observé (variable d'intérêt Y et k covariables X_1, ...,X_k), une démarche non paramétrique est envisageable à partir de l'information fournie par les tables de contingences associées à ces observations de façon analogue à celle utilisée par [4] et [5]. Nous utilisons ici un principe de maximum d'information par rapport à une σ-algèbre de référence [3], cette dernière correspondant dans le cas présent aux observations de la variable d'intérêt Y.

3 Méthode de sélection de covariables

La méthode employée est basée sur l'estimation des entropies conditionnelles et non conditionnelles de la variable Y et des covariables $X_1, ..., X_k$. Ces estimations sont calculées à partir des fréquences empiriques fournies par des tables de contingences à plusieurs entrées. Pour les covariables continues, une discrétisation est donc nécessaire.

3.1 Préparation des tables de contingences

Pour diminuer l'occurrence de faibles effectifs dans les tables de contingences, le nombre de modalités d'une covariable est parfois restreint en utilisant un principe de maximum d'entropie, c'est-à-dire l'équiprobabilité des modalités.

3.2 Méthode de sélection

Soient Y la variable d'intérêt, X une covariable quelconque, (n_{ij}) le tableau de contingence qui leur est associé, et $p = (p_{ij})$ la matrice des espérances des fréquences relatives.

L'information relative de X sur Y est

$$I_r(X,Y) = \frac{I(X,Y)}{H(Y)}$$

dont un estimateur convergent est

$$T = \frac{\sum_{i,j} \widehat{p_{ij}} Ln\left(\frac{\widehat{p_{ij}}}{\widehat{p_{i.}}\widehat{p_{.j}}}\right)}{\sum_{j} \widehat{p_{.j}} Ln(\widehat{p_{.j}})}$$

où $\widehat{p_{ij}}$, $\widehat{p_{i.}}$ et $\widehat{p_{.j}}$ sont les notations classiques pour les fréquences relatives empiriques.

La variance asymptotique de T est

$$Var(T) = \frac{1}{n} Var\left(\sum_{i,j} \frac{\partial I_r(X,Y)}{\partial p_{ij}}(p) \mathbb{I}_{ij}\right)$$

où $n = \sum_{i,j} n_{ij}$ est le nombre d'unités statistiques considérées, et les $\mathbb{I}_{ij}$ sont les variables indicatrices associées au tableau (n_{ij}).

On peut de manière analogue calculer l'information relative combinée de plusieurs covariables $X_1, ..., X_k$ sur Y. L'évaluation de la variance d'information permet alors d'établir des régions de confiance et d'éliminer les covariables les moins pertinentes.

Lorsque Y est la variable de survie, $I_\tau(X, Y)$ est appelé valeur pronostique de X et noté $I_\tau(X)$.

4 Hiérarchisation des modalités d'une covariable

Dans le cas où Y est la variable de survie, on appellera taux de guérison d'une sous-population le taux de survie à la fin de l'expérimentation.

Chaque covariable divise la population de malades en une partition correspondant aux différentes modalités. Ces dernières seront ordonnées en fonction de leur taux de guérison.

Si deux modalités ont même taux de guérison, on fera appel aux taux de survie moyens.

5 Application au cancer du sein

5.1 Présentation des données

Les données proviennent d'une enquête réalisée à Marseille par les professeurs J.-M. SPITALIER et D. HANS. Elles concernent 1304 patients suivis sur une période de 10 ans. Les variables observées sur chaque patient sont au nombre de 45. Elles sont quantitatives et qualitatives et un prétraitement a permis de ne retenir qu'une quinzaine d'entre elles à savoir:
1-âge, 2-classe clinique, 3-classe thermographique, 4-classe sénographique, 5-classe échographique, 6-PEV clinique, 7-diamètre clinique, 8-allure clinique, 9-côté, 10-histologie, 11-histologie N, 12-nombre de ganglions envahis, 13-récepteurs œustradiol, 14-récepteurs progestérone, 15-stade UICC.

5.2 Sélection des covariables

Les calculs montrent que les meilleures covariables sont le stade UICC, le nombre de ganglions envahis et l'histologie (Fig. 1).

La meilleure combinaison de 4 covariables est celle comprenant les trois précédentes plus la covariable Age (Fig. 2).

5.3 Hiérarchisation des modalités d'une covariable

Sachant qu'une covariable a une bonne valeur pronostique, il est intéressant de pouvoir affecter aux différentes valeurs possibles de cette covariable une courbe théorique de survie (Fig. 3). Ces courbes représentent la survie conditionnelle à la valeur prise par cette covariable chez un malade et sont des outils de pronostic médical.

6 Conclusions

Nous avons présenté une méthode asymptotique qui ne s'applique que dans le cas d'un nombre important d'unités statistiques. Elle permet la sélection de covariables apportant le plus d'information sur la variable d'intérêt. Son application au cas de malades atteints du cancer du sein a permis de mettre en évidence quatre covariables principales.

Fig. 1. Analyse de données de survie

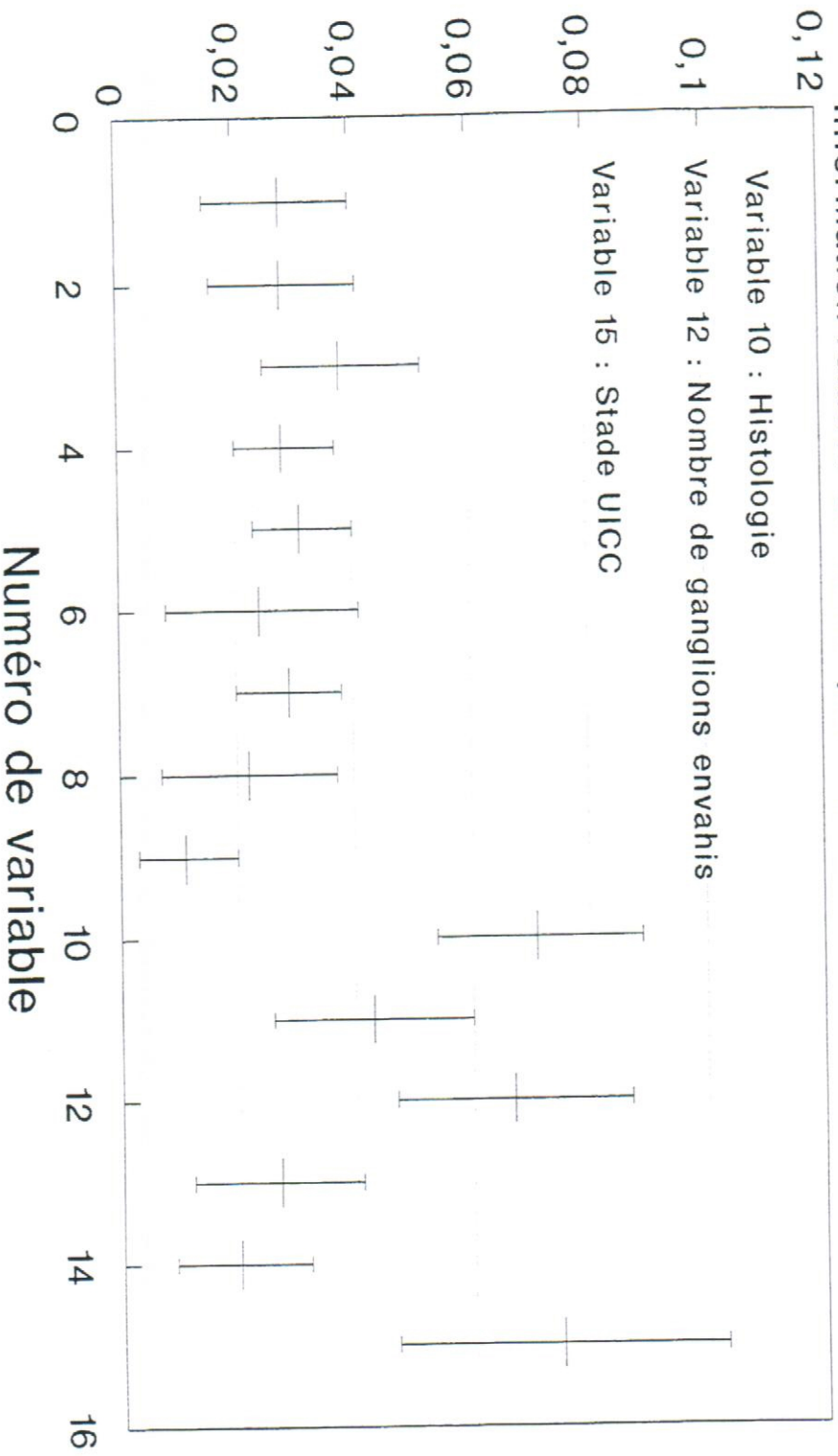


Fig. 2. Analyse de données de survie

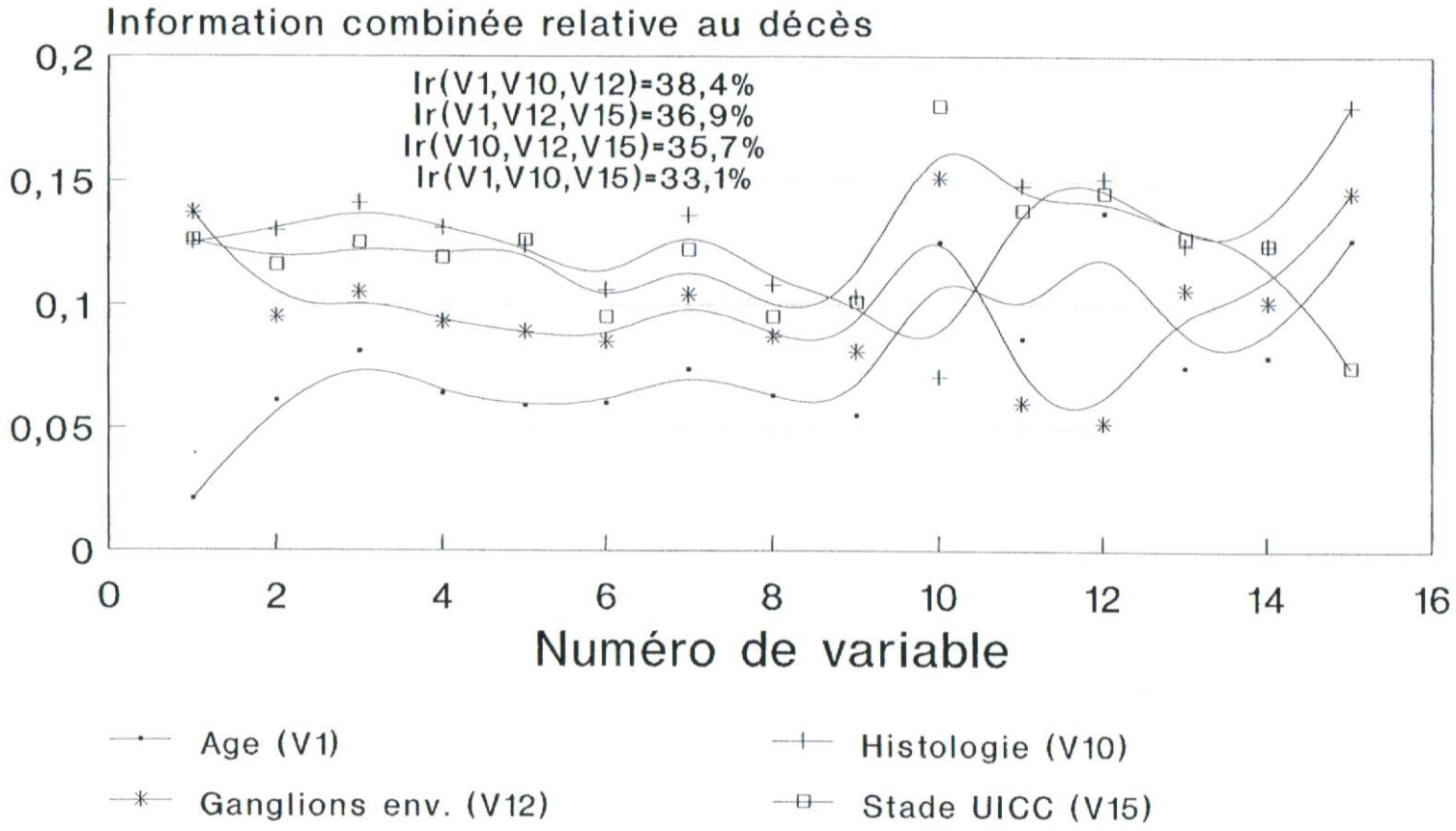
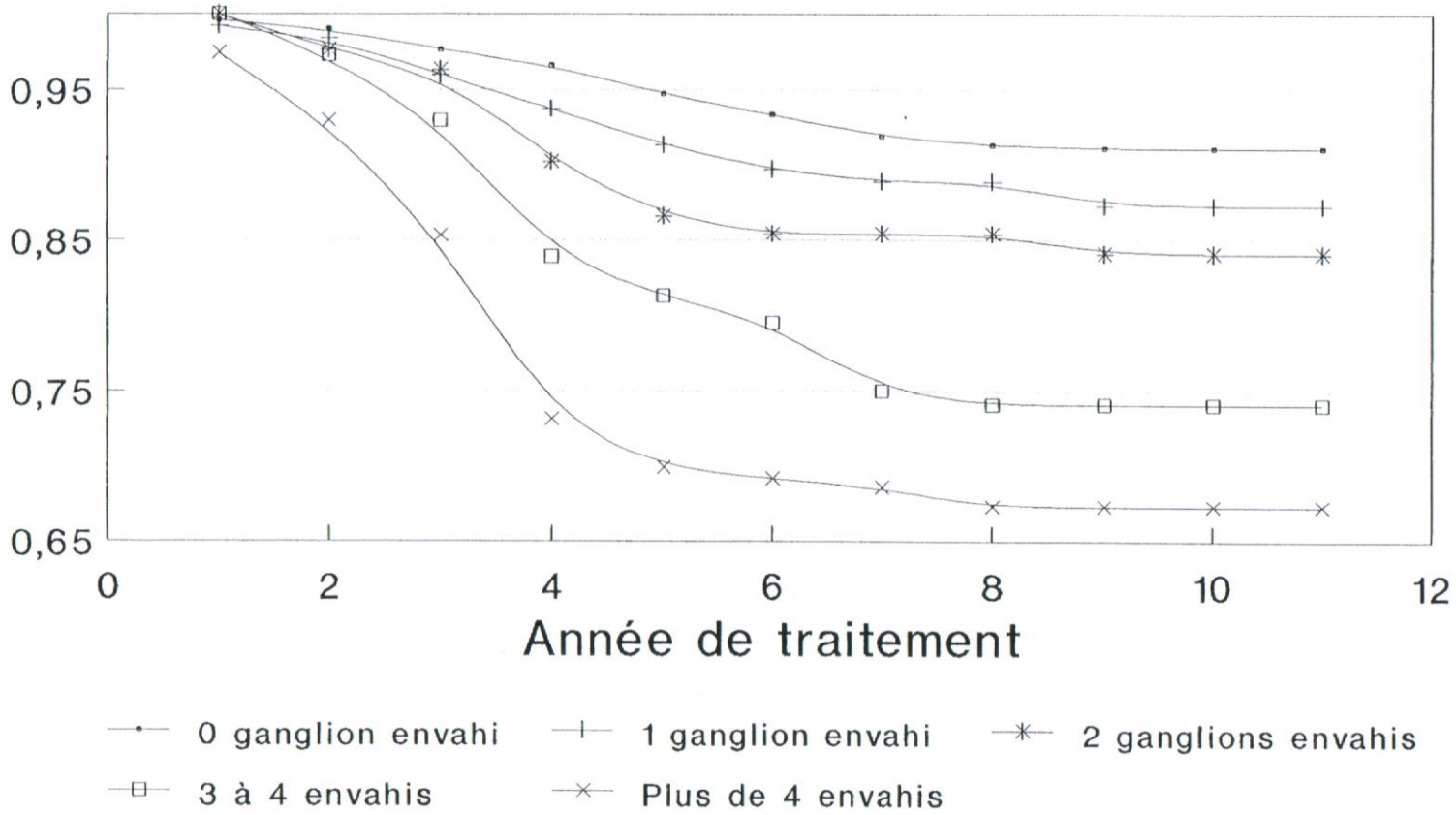


Fig. 3. Courbes de survie
(associées à la covariable V12)



References

[1] ARCONTE,A. KHATTABI, J. MANUCEAU, C. MARTIAS, Mutual Information between σ-algèbras, *Prépublication UAG*, **95/07**,17p, 1995.

[2] DALEY & VERES-JONES, An introduction to the theory of point processes, *Springer-Verlag*, New-York, 1988.

[3] MANUCEAU, J., TROUPÉ, M., VAILLANT, J., Information between σ-algebras: applications to selection of pronostic variables. *To appear.*

[4] RAO, C.R., Diversity: Its measurement, decomposition, apportionment and analysis. *Sankhya Series A*, **44**,1-22, 1982.

[5] RAO, C.R., Generalization of ANOVA through entropy and cross entropy functions. In *Probability theory and mathematical statistics*, **vol. 2**,477-494, Sciences Press, Utrecht, 1986.

[6] RÉNYI, A., On measures of entropy and information. *Proceedings of the fourth Berkeley Symposium on Mathematical Statistics and Porbability* **1**, 547-561, 1961.

[7] ROBERT, C., An entropy concentration theorem: applications in artificial intelligence and descriptive statistics. *J. Appl. Prob.* **27**,303-313, 1990.