

Modélisation des événements du parcours de vie

Une introduction

Jean-Marie Le Goff¹

Décembre 2003

¹ Centre lémanique d'analyse des parcours et modes de vie (PaVie) et laboratoire de démographie et d'études familiales, Universités de Lausanne et de Genève. Nous remercions Gilbert Ritschard pour ses commentaires sur une première version de ce texte. Ce texte est en progrès. Tous commentaires et toutes suggestions sont les bienvenus. Contact : Jean-Marie.LeGoff@PaVie.Unil.ch.

1. INTRODUCTION	3
2. BUTS ET NOTIONS DE L'ANALYSE DES BIOGRAPHIES	4
2.1 LA NOTION DE RISQUE	4
2.1.1 <i>Evénements et processus</i>	4
2.1.2 <i>En temps continu</i>	5
2.1.3 <i>En temps discret</i>	7
2.2 UN MODELE EXPLICATIF	7
2.2.1 <i>Formulation générale</i>	7
2.2.2 <i>La dépendance du risque au temps</i>	8
2.2.3 <i>Les différences interindividuelles</i>	10
2.3 LES AUTRES NOTIONS PROBABILISTES DE LA MODELISATION DES PARCOURS DE VIE	11
2.3.1 <i>Fonction de séjour et proportion des individus ayant connu l'événement</i>	11
2.3.2 <i>Probabilité ou densité de connaître l'événement</i>	12
2.3.3 <i>Le risque cumulé</i>	13
3. ESTIMER UN MODÈLE	13
3.1 PREALABLES	13
3.1.1 <i>Données orientées événements</i>	14
3.1.2 <i>Censures et troncatures</i>	15
3.1.3 <i>Autres aspects importants à prendre en compte pour la spécification d'un modèle</i>	17
3.2. LES MODELES ET LEURS HYPOTHESES	18
3.2.1 <i>Les différentes approches</i>	18
3.2.2 <i>Tableau récapitulatif</i>	21
4. CONCLUSION	23
RÉFÉRENCES BIBLIOGRAPHIQUES	24

1. Introduction

Cet article a pour objet de présenter les méthodes d'analyse des biographies. Cette présentation est multiple, en ce sens qu'il s'agit d'exposer les buts et objectifs de l'analyse des biographies, les notions statistiques attachées à ces méthodes, les précautions à prendre avant de se lancer dans une analyse, ainsi que les différentes approches pouvant être développées.

Nous emploierons indistinctement les expressions d'analyse des biographies ou de modélisation des événements du parcours ou de l'histoire de vie pour désigner l'ensemble du corpus de techniques statistiques visant à analyser les événements d'histoire de vie. Le premier de ces termes a été proposé par Courgeau et Lelièvre (1989) dans les années quatre-vingt, alors que le second est une traduction assez libre de « *Event History Analysis* », dans le but d'insister sur l'usage de ces méthodes dans une optique d'analyse des parcours de vie (Tuma and Hannan 1984, Mayer and Tuma 1990). Nous éviterons, en revanche le terme d'analyse de survie (*Survival Analysis*) et plus encore celui d'analyse de « données de panne » (*Failure Time Data*)², malgré que les notions et techniques statistiques que recouvrent ces expressions soient fortement semblables à ceux de l'analyse des biographies (Cox and Oakes 1984, Kleinbaum 1996). Les événements analysés dans la modélisation des parcours de vie diffèrent de ceux de l'analyse de processus biologiques ou physiques en ce sens que, bien souvent, il s'agit d'événements qui résultent d'actions humaines (Blossfeld et Mills 2001). Selon les écoles de pensée dans les différentes disciplines des sciences sociales, ces actions seront interprétées, soit en termes de décisions rationnelles des individus, soit en termes d'obéissance à des normes collectives, soit en termes de mobilisations de différentes ressources (ressources psychologiques, capital social), soit encore, en termes de possibilités d'action qu'offre le contexte dans lequel se situent les individus.

On notera que l'expression « analyse d'histoire d'événements » (*Event History Analysis*) est aussi utilisée en Sciences Politiques, toujours pour désigner le même ensemble de techniques statistiques, mais en vue d'analyser des phénomènes à l'échelle d'organisations ou d'institutions collectives et non plus à l'échelle d'un individu humain (Box-Steffensmeier and Jones, 1997). Les objets de recherches vont ainsi porter sur l'analyse de la diffusion d'une grève d'une mine de charbon à une autre (Conell and Cohn 1995) ou sur l'étude de la décolonisation des pays africains dans les années soixante (Strang 1990). La particularité de ces travaux est que, bien souvent, leurs auteurs portent une attention particulière aux phénomènes de diffusion (Strang and Tuma, 1993). Néanmoins, il reste que le questionnement de recherche sous-jacent est celui de l'action humaine, celle-ci étant approchée de manière collective. L'usage du terme *Event History Analysis* se rapporte alors à l'histoire plutôt qu'au parcours ou à l'histoire de vie individuelle.

Dans cette présentation, nous resterons dans une optique d'usage des méthodes d'analyse des biographies en vue de l'analyse des parcours de vie. Dans le point suivant, nous présenterons ainsi les objectifs de la modélisation des événements du parcours de vie ainsi que les principales notions probabilistes attachées à ces méthodes. Dans le troisième point, nous nous pencherons sur les aspects se rapportant à l'estimation des modèles d'analyse des biographies, d'abord en nous interrogeant sur les données nécessaires et leur préparation en vue d'une exploitation statistique puis en présentant les différentes approches et leurs hypothèses.

² Vermunt (1997) recense d'autres termes utilisés dans les pays anglo-saxons tels que « *life-time models* » ou « *duration models* ».

2. Buts et notions de l'analyse des biographies

2.1 La notion de risque

2.1.1 Événements et processus

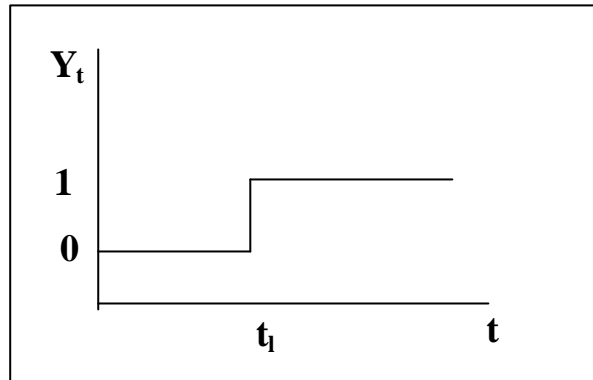
Selon Yamaguchi (1991), dans leur aspect le plus général, les techniques de modélisation des événements d'histoire de vie visent à analyser la durée dans un état ou dans une situation, par exemple, la durée de situation de célibat. Toutefois, cette analyse ne se limite pas à ce simple aspect de durée, mais aussi à la transition d'une situation à une autre, dans notre exemple, le passage de la situation de célibataire à la situation de marié. L'analyse des biographies est ainsi parfois appelée « analyse des données de transition » (Lancaster, 1992), ce dont témoigne le sigle du logiciel spécialisé dans l'usage de ces méthodes TDA (*Transition Data Analysis*, cf. Röhwer and Pötter, 2002).

La notion de transition implique un « événement », c'est-à-dire, un marqueur qui délimite dans le temps la situation d'origine de la situation d'arrivée (Coleman, 1981). Ainsi le mariage constituera l'événement séparant la situation de célibat de la situation de marié. Une telle définition de l'événement s'oppose à une autre approche couramment utilisée en sciences sociales dans laquelle ce sont les personnes interrogées dans une enquête qui définissent les événements ou les jalons importants de leur vie, ces événements étant alors interprétés comme des événements marquants du parcours de vie. La définition de l'événement en analyse des biographies est une définition factuelle. Un événement peut-être unique (ou rare), lorsque l'on se situe au niveau de l'individu. Il doit cependant être répétable lorsque l'on se situe au niveau d'une population. La modélisation des événements d'histoire de vie va ainsi s'intéresser à la distribution de l'occurrence d'un événement ou d'une transition au cours du temps. Ceci signifie que la population devant être prise en compte doit être composée d'individus *soumis au risque de connaître l'événement* (Yamaguchi, 1991). Ainsi, seules des personnes mariées peuvent être soumises au risque de divorcer, et une analyse du divorce doit exclure toute personne non-mariée. Un événement va ainsi correspondre à une transition entre deux états (*state*) (Ruspini, 2002).

D'un strict point de vue statistique, un événement peut être formalisé comme étant un changement dans une variable (Tuma et Hannan, 1984). Cette variable est toutefois particulière, en ce sens qu'il s'agit d'une variable ayant un nombre limité de modalités ou d'attributs qui sont mutuellement exclusifs. Cette variable peut être formalisée sous la forme d'un processus évoluant dans le temps avec un nombre discret d'états (Coleman, 1981). La figure 1 représente ainsi un processus Y_t variant au cours du temps entre deux états 0 et 1, ce processus étant observé chez un individu. Au temps t_1 , l'individu passe de l'état 0 à l'état 1 (de l'état de célibataire à l'état de marié). La suite du processus pourrait être un retour à l'état 0, si l'on ne faisait pas la distinction entre l'état de célibataire et l'état de divorcé.

Figure 1 : Processus à deux états

Erreur !



2.1.2 En temps continu

Dans l'ensemble de la population, le processus Y_t va constituer une variable aléatoire, en ce sens que le moment d'occurrence de l'événement variera selon les individus. L'intérêt porte alors sur la distribution de la probabilité d'un changement dans la variable Y_t au cours du temps. L'événement, ou la variation dans la variable Y_t , peut être symbolisé par $DY_t=1$ ($DY_t=0$ signifie alors que l'événement n'a pas eu lieu, c'est-à-dire, que les individus sont restés dans le même état). Dans le cas d'une variable Y_t à deux états, cette probabilité correspondra à la probabilité d'être passé dans la situation 1 au temps t' sachant que les individus étaient dans la situation 0 au temps t ($t' > t$). Il s'agit donc d'une probabilité conditionnelle, que l'on peut formaliser de la manière suivante :

$$P(\Delta Y_t = 1) = P(Y_{t'} = 1 / Y_t = 0) \quad (2.1)$$

Une formulation plus classique de cette probabilité est de considérer non pas le changement dans la variable Y_t , mais le moment au cours duquel les individus connaissent l'événement considéré. Ainsi, si T représente la durée observée dans l'état initial avant la transition vers le deuxième état³, alors cette probabilité sera la probabilité de connaître l'événement durant l'intervalle de temps $[t, t']$, sachant que les individus ne l'avaient pas encore connu à l'instant t . Cette probabilité s'écrira :

$$P(\Delta Y_t = 1) = P(t \leq T < t' / T \geq t) \quad (2.2)$$

Si la différence entre t' et t correspond à l'unité de temps utilisée pour mesurer les durées (par exemple, l'année) cette probabilité correspond à ce qui s'appelle le « quotient », tel que celui-ci est défini dans les tables classiques de démographie, notamment, la table de mortalité (Pressat 1983, Caselli et al. 2001). Ce quotient est souvent symbolisé par le symbole q_x où x symbolise l'âge. La valeur de cette probabilité dépend étroitement de la largeur de l'intervalle de temps $[t, t']$. Plus cet intervalle est court, plus le nombre d'événements durant cet intervalle de temps est petit et, en conséquence, plus la probabilité de connaître l'événement au cours de cet intervalle de temps est faible. Si l'on se situe en temps continu, on considérera alors un intervalle de temps Dt très petit. Dans ce cas la probabilité de connaître l'événement durant l'intervalle de temps $[t, t+Dt]$ est quasiment égal à 0 :

³ T est une variable aléatoire, c'est-à-dire, que cette « variable varie » selon les individus.

$$P(\Delta Y_t = 1) = P(t \leq T < t + \Delta t / T \geq t) \approx 0 \quad (2.3)$$

Cela n'est toutefois plus le cas si l'on divise cette probabilité par Δt . Cette division revient à estimer le nombre moyen d'événement que connaîtrait l'individu si les conditions dans lequel il se situe durant ce court instant se maintenaient tout au long d'une unité de temps (l'année ou le mois). Supposons ainsi que l'unité de temps est l'année et que l'intervalle de temps Δt est le mois (c'est-à-dire, un douzième d'année ce qui représente déjà un intervalle de temps considérable). Si la probabilité de connaître un événement durant le premier mois d'une année donnée est de $1/4$, alors, dans le cas d'un maintien des conditions pour lesquelles on obtient cette probabilité tout au long de l'année, les individus connaîtraient $(1/4) / (1/12) = 3$ événements en moyenne durant cette année-là. Ce nombre est ici une estimation du *risque* de connaître l'événement.

Le risque correspond ainsi au nombre moyen attendu d'événement si les conditions dans lesquels il est mesuré se maintenaient tout au long de l'unité de temps. Selon une métaphore proposée par Kleinbaum (1996), le risque est équivalent à une vitesse, mais qui serait mesurée de manière instantanée, comme cela est le cas lorsqu'elle est mesurée avec un compteur de vitesse qui est situé dans une voiture. La mesure à un moment donné d'une vitesse de 60 Km/heure indique que l'on ferait 60 Km dans l'heure qui suit si l'ensemble des conditions régissant la conduite de la voiture (état de la route, état d'esprit du conducteur) se maintenaient tout au long de l'heure qui suit. Le risque est ainsi une *grandeur latente*, qui mesure un *potentiel instantané d'occurrence d'un événement* (Yamaguchi 1991, Kleinbaum 1996, Mills 1999). La notion de risque peut être comprise de manière intuitive lorsque l'on est dans le cas d'un événement répétable (Allison 1995). Ainsi, un risque de 3 de contracter la grippe signifie que l'on risque de contracter trois fois la grippe au cours de l'année, en moyenne. Dans le cas d'événements non répétables, tel que le décès, ou un premier mariage, le risque peut sembler une notion plus difficile à appréhender, voire paraître absurde, puisqu'il peut être supérieur à 1. Il faut toutefois noter que l'inverse du risque va représenter la durée moyenne avant de contracter un événement (toujours dans le cas d'un maintien des conditions régissant l'occurrence des événements). Un risque de 3 signifie alors que la durée moyenne avant de connaître l'événement sera de $1/3$, c'est-à-dire, de quatre mois si l'unité de temps est l'année.

Lorsque le temps est continu, le risque est formulé de la manière suivante :

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(T < t + \Delta t / T \geq t)}{\Delta t} \quad (2.4)$$

Le symbole limite est ici utilisé, car lorsque Δt est très petit, on se situe dans le cas d'une indétermination en ce sens que le risque est égal à 0 divisé par 0. Il est à noter que le risque est parfois formalisé de la manière suivante dans la littérature :

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t / T \geq t)}{\Delta t} \quad (2.5)$$

Une autre formulation du risque proposée par Strang et Tuma (1993) est intéressante à mentionner, car elle reprend l'idée d'un événement en tant que changement dans une variable. Si le processus analysé est un processus à deux états 0 ou 1, le risque de passer de l'état 0 à l'état 1 s'écrira de la manière suivante :

$$h_i(t) = \lim_{\Delta t \rightarrow 0} \frac{P(Y(t + \Delta t) = 1 / Y(t) = 0)}{\Delta t} \quad (2.6)$$

D'autres termes que celui de risque peuvent être utilisés. Ainsi, les termes de quotient instantané ou de taux instantané ont été proposés en démographie (Courgeau et Lelièvre 1989, Toulemon 1996). Par ailleurs, dans les études sur la mortalité, le risque a souvent été appelé « force de mortalité » (Pressat 1983). Le terme d'intensité d'occurrence de l'événement peut aussi être utilisé. En revanche, en temps continu, *le terme de probabilité instantanée doit être banni*, car le risque n'est pas une probabilité, mais une densité conditionnelle, dont la valeur estimée peut être supérieure à 1. Les termes anglo-saxons couramment utilisés sont ceux de *hazard*, *instantaneous rate*, *conditional failure rate*, ou *intensity*. Un terme intéressant est celui de *propensity* (propensité), car il souligne le caractère latent du risque (Palloni 2001).

2.1.3 En temps discret

Le risque a été présenté dans le point précédent en considérant le temps de manière continue. On peut néanmoins considérer le temps de manière discrète (Cox 1972, Allison 1982, Yamaguchi 1991). En temps discret, le risque devient une probabilité. Il s'agit ainsi de la probabilité de connaître l'événement au temps t_1 sachant que l'on n'avait pas encore connu cet événement auparavant. Dans ce cas, le risque, que nous noterons ici $P(t_1)$ afin de souligner son caractère discret, sera formalisée par :

$$P(t_1) = P(T = t_1 / T \geq t_1) \quad (2.7)$$

Les approches en temps discret sont souvent intéressantes dès lors que l'on dispose de mesures de durées d'occurrence d'événements sur la base d'une unité de temps qui est longue (par exemple, l'année, dans le cas de données socio-démographiques) ou lorsque l'on observe de nombreuses échéances au cours d'un même intervalle de temps.

2.2 Un modèle explicatif

2.2.1 Formulation générale

Un des questionnements abordés au travers de l'utilisation des méthodes d'analyse des biographies est celui des différences interindividuelles dans les rythmes de transition d'une situation à une autre. Ces différences peuvent être de deux types. Le premier type correspond aux facteurs qui se rapportent à l'environnement dans lequel vit un individu, plus précisément, à son contexte socio-économique et culturel. Ainsi, on peut supposer que le risque pour une population de chômeurs de sortir du chômage dépendra de l'offre de travail à un moment donné. Le deuxième aspect se rapporte aux attributs des individus, tel que leur sexe, leur appartenance sociale, leur expérience professionnelle, etc. Certaines des caractéristiques prises en compte peuvent être des caractéristiques qui restent semblables tout au long du parcours de vie des individus. Il s'agit, par exemple, de caractéristiques qui se rapportent à l'origine sociale : on parle alors de *caractéristiques invariantes* au cours du temps (*fixed covariates*). Néanmoins, un grand nombre de caractéristiques peuvent se transformer au cours du temps. On parle dans ce cas de *variables dépendantes du temps* (*time varying covariates*).

Cet ensemble de variables, dépendantes du temps ou non, pourra être symbolisé mathématiquement par un vecteur de variables x_t dans lequel chaque variable correspondra à une caractéristique de l'individu⁴ :

$$x_t = (x_{1t}, x_{2t}, x_{3t}, \dots, x_{kt}) \quad (2.8)$$

Où k représente le nombre total de variables. x_{jt} , où $j=(1, 2, \dots, k)$ peut être une variable quantitative (par exemple, le revenu de l'individu) ou une variable qualitative binaire dans laquelle 0 signifie que l'individu ne possède pas la caractéristique donnée et 1 qu'il la possède (par exemple, avoir ou non un niveau d'études supérieures). Dans le cas où l'on aurait une variable de plus de deux modalités, la variable peut être décomposée en autant de variables binaires qu'il y a de modalités moins une. Ainsi, si l'on souhaite décrire le niveau d'éducation selon trois niveaux (bas, moyen, élevé), on aura deux variables où, par exemple, (0, 0) correspondra au niveau bas, (1, 0) au niveau moyen et (0,1) au niveau élevé. Les individus pour qui toutes les caractéristiques binaires sont égales à 0 constituent la population des individus de référence.

La relation la plus générale concernant la modélisation du risque consiste à considérer que celui-ci est *fonction* du temps d'une part, et des caractéristiques que l'individu possède au temps t d'autre part (Blossfeld and Rohwer, 2002).

$$h(t) = f^o(t, x_t) \quad (2.9)$$

f^o symbolise ici la notion de fonction. Malgré son aspect très général, cette relation est fondamentale en ce sens que tout «l'art» de la modélisation des méthodes d'analyse des biographies consiste à préciser la dépendance du risque par rapport au temps et aux caractéristiques des individus, voire l'interaction entre le temps et les caractéristiques individuelles (Yamaguchi, 1991). Cette relation peut avoir une autre interprétation (Courgeau et Lelièvre, 1989). Elle consiste à considérer que le risque de connaître un événement dépend du présent des individus, c'est-à-dire, du contexte dans lequel il se situe (la composante t du modèle) ainsi que de leur passé (la composante x_t du modèle), si par passé, on entend aussi bien des événements antérieurs du parcours de vie que les caractéristiques se rapportant à l'origine sociale ou au moment de la naissance. En vue de mieux détailler encore cette équation, il nous faut nous arrêter sur la dépendance du risque au temps et aux variables en présence.

2.2.2 La dépendance du risque au temps

Selon Blossfeld et Rohwer (2002), la dépendance du risque au temps peut avoir trois interprétations :

- *Proxy de variables dépendantes du temps* : la dépendance au temps peut exprimer un effet dû à une ou des caractéristiques qui se transforment au cours du temps, mais qui ne peuvent être mesurées. Le temps joue alors un rôle de «*proxy*» de ces facteurs non observés. Par exemple, l'accumulation d'expérience professionnelle spécifique dans l'exercice d'une activité professionnelle est peu facile à mesurer. Toutefois, on peut considérer que plus la durée de l'emploi s'allonge et plus il y a accumulation

⁴ L'indice t signifie ici que ce vecteur peut varier au cours du temps.

d'expérience professionnelle. Si le temps constitue un *proxy* ou un indicateur du cumul d'expérience professionnelle, alors on peut supposer que plus la durée de l'emploi s'allonge et plus un départ de l'emploi devient coûteux pour l'employé, mais aussi pour son employeur. En conséquence, le risque de départ de l'emploi devrait diminuer au cours du temps (Tuma et Hannan 1984, Blossfeld et Rohwer 2002)⁵. Des hypothèses similaires peuvent être posées dans le cas de l'analyse du divorce en fonction de la durée du mariage (Blossfeld and Rohwer, 2002). L'analyse de la dépendance au temps fait l'objet d'une grande attention de la part des économistes, notamment de ceux qui s'intéressent à la durée du chômage (Lancaster, 1979). Pour ces auteurs, le « risque » de retour en emploi est, par hypothèse, considéré dépendre de divers coûts et bénéfices qui évoluent au cours de l'allongement de la durée du chômage. De telles hypothèses sur la dépendance au temps sont néanmoins fragiles, en raison de la deuxième interprétation de celle-ci ;

- *Hétérogénéité non-observée (Unobserved Heterogeneity)* : la dépendance au temps peut en fait être fautive en ce sens qu'elle peut résulter de l'omission d'une ou plusieurs caractéristiques dans le modèle, bien souvent parce que ces caractéristiques ne sont pas mesurables. On parle dans ce cas d'un effet de dépendance au temps due à un facteur d'hétérogénéité non-observée. Ce cas de figure diffère du précédent en ce sens que la variable non mesurable était considérée dans le premier cas une variable dépendante du temps, mais qui était la même pour tous les individus de la population, ainsi que nous l'avons vu précédemment dans l'exemple de l'accumulation d'expérience au cours du temps. Dans le cas de figure de l'hétérogénéité non-observée, les variables absentes sont considérées être des variables invariantes au cours du temps, mais qui indiquent des différences entre les individus⁶. Les caractéristiques non-observées peuvent être introduites dans notre modèle général de la manière suivante (Vaupel et al., 1979) :

$$h(t) = f^o(t, x_i, z) \quad (2.10)$$

z représente les caractéristiques non observées d'un individu. Les premiers travaux visant à modéliser l'hétérogénéité non observée ont été initiés en démographie et en économie. En démographie, elle a été prise en compte dans l'analyse de la mortalité et le terme utilisé est alors celui de *fragilité* ou de *vulnérabilité (Frailty)* plutôt que celui de d'hétérogénéité. En économie, plus particulièrement dans les travaux qui portent sur la durée du chômage, l'absence de prise en compte de l'hétérogénéité non observée est considérée constituer un biais particulièrement important dans la recherche de la dépendance au temps, telle que nous l'avons décrite précédemment (Heckman et Singer, 1984). Signalons, enfin que les sociologues et politologues utilisant les modèles de l'analyse des biographies portent aussi un grand intérêt à la notion d'hétérogénéité non-observée (Greve et al, 1995) ;

- *Marqueur d'un processus de diffusion* : les modèles de diffusion s'appuient sur l'idée qu'il peut exister des processus d'imitation de comportement d'un individu sur celui d'un autre. Dans ce cas, on considère que le risque dépend de la proportion des

⁵ Nous reprendrons cette hypothèse dans les exemples d'analyse des biographies qui seront développés dans les articles consacrés à l'usage des méthodes non-paramétriques et à la mise en œuvre des modèles de régression logistique à temps discret avec SPSS.

⁶ Cf. quelques précisions sur la notion d'hétérogénéité non-observée dans l'article sur l'usage des méthodes non-paramétriques avec SPSS.

personnes qui ont déjà connu l'événement. Plus exactement, il est en fait supposé que le risque est d'abord croissant, puis décroissant (Diekmann 1989, Yamaguchi 1994). La spécificité de cette distribution dépendrait en fait d'un double phénomène : le premier correspondrait à un effet de pression sociale de la part des personnes ayant connu la transition (par exemple un mariage) sur celles n'ayant pas encore connu l'événement, les premières devenant de plus en plus nombreuses au fur et à mesure de l'écoulement du temps ; Le second correspondrait à une baisse de l'attractivité de l'événement (de la formation d'une union maritale) chez les personnes qui ne l'ont pas encore connu, voire, à une diminution des contacts sociaux entre personnes mariées et personnes non-mariées (Braun et Engelhart 2002, Blossfeld et Nazio 2002).

2.2.3 Les différences interindividuelles

Pour Yamaguchi (1991), il y a deux grands types de variables fixes :

- *Statuts attribués (ascribed status)* : ces variables concernent des caractéristiques que les individus ont acquies à leur naissance (sexe, origine sociale) ;
- *Statuts atteints avant le début (statuses attained prior to)* de l'entrée dans la période dans laquelle l'individu est soumis au risque de connaître l'événement, ces variables restant constantes par la suite. Il s'agit, par exemple, de l'âge au premier mariage dans l'analyse du divorce.

Les variables dépendantes du temps sont, quant à elles, de trois types (Blossfeld and Mills, 2001) :

- *Variables prédéfinies (defined time-dependent covariate)* : il s'agit de variable pour lesquelles l'évolution (leur course) au cours du temps est déterminée en avance pour l'ensemble des individus. Il s'agira typiquement des variables qui permettent de décrire chacune des différentes horloges pouvant être prises en compte dans une analyse. Ainsi, si l'intérêt porte sur l'analyse de départ d'un premier emploi, outre la durée écoulée depuis l'embauche, d'autres horloges pourront être introduites, tel l'avancement en âge, l'écoulement du temps dans le calendrier, etc. Ces variables sont dites prédéfinies, en ce sens que, *par définition*, elle évoluent de la même manière pour tous les individus et indépendamment des événements qu'ils vivent ;
- *Variables auxiliaires (ancillary time dependent covariate)* : dans cette catégorie entrent des variables se rapportant à l'évolution du contexte dans lequel se déroulent les existences individuelles. Typiquement, il s'agit ainsi de variables se rapportant aux transformations du contexte socio-économique. D'autres variables se rapportant à des infrastructures à proximité du lieu de résidence des individus peuvent être introduites. On peut citer ici les travaux développés par Hank (2002) sur la fécondité en Allemagne dans lesquels l'auteur porte son intérêt à des variables contextuelles, par exemple, le nombre de places de crèche dans une unité géographique. Ce type d'analyse connaît un certain nombre de développements visant à intégrer les modèles de l'analyse des biographies à ceux de *l'analyse multi-niveaux (Multi-level Analysis)* qui connaît actuellement un plein essor (Courgeau, 2003). Par définition, les variables auxiliaires sont des variables qui décrivent un processus qui est externe (ou presque) aux

individus, c'est-à-dire, que le processus individuel analysé ne peut avoir (quasiment) aucune influence sur ce processus externe ;

- *Variables internes (internal time-dependent covariate)* : ces variables se rapportent à des processus qui peuvent directement exercer une influence sur le processus analysé tout comme ce dernier peut exercer une influence réciproque sur le premier. On parle alors de processus interdépendants⁷. Typiquement, les variables internes dépendantes du temps se rapportent ainsi à des événements appartenant à d'autres domaines du parcours de vie, par exemple, aux événements de la vie familiale si l'intérêt porte sur un événement de la vie professionnelle. Néanmoins, ces variables internes peuvent correspondre à l'occurrence d'événements chez des personnes proches d'un individu, par exemple, les membres de sa famille ainsi que les personnes appartenant à son entourage ou à son réseau social (Lelièvre et al., 1997). Les dépendances et interdépendances du risque d'occurrence de l'événement étudié peuvent être ainsi classées en trois catégories : 1) dépendance à un événement du parcours de vie (par exemple, effet d'une conception hors-mariage sur le mariage ; 2) dépendance à une situation ou à un statut, qui peut changer au cours du temps (par exemple, niveau d'éducation sur le mariage) ; 3) dépendance à un ou plusieurs individus. Lillard (1993) propose toutefois une quatrième catégorie en considérant que le risque d'occurrence à un moment donné peut dépendre du risque d'occurrence d'un autre événement. Par exemple, la fécondité dans le mariage (plus exactement, le risque de concevoir un enfant) dépend du risque d'occurrence d'un divorce à un moment donné (Lillard and Waite 1993). Un tel modèle « risque dépendant » peut s'écrire :

$$h(t) = f^o(t, x, z, \mathbf{m}(t)) \quad (2.11)$$

Où z représente les caractéristiques non observées d'un individu et $\mathbf{m}(t)$ représente le risque d'occurrence d'un autre événement. Dans la suite du texte, nous en resterons au modèle le plus simple, c'est-à-dire, celui dans lequel l'hétérogénéité et la dépendance au risque d'occurrence à un autre événement n'interviennent pas.

2.3 Les autres notions probabilistes de la modélisation des parcours de vie

Si le risque constitue la notion probabiliste fondamentale de l'analyse des biographies, d'autres notions statistiques sont aussi indispensables pour décrire la distribution de l'occurrence d'un événement au cours du temps.

2.3.1 Fonction de séjour et proportion des individus ayant connu l'événement

La fonction de séjour est avec la distribution du risque la distribution la plus souvent utilisée dans la modélisation des événements de l'histoire de vie. A un moment donné du temps, c'est-

⁷ Plusieurs types d'approches des interdépendances entre événements ont été développées dans la littérature. Mentionnons ainsi les approches causales (Blossfeld and Mills 2001, Blossfeld et Rohwer 2001), les approches en termes d'interaction entre événements (Courceau et Lelièvre, 1986 et 1989) ou les approches en termes d'événements inter-reliés (Lillard and Waite 1993 ; Brien et al. 1999).

à-dire, au temps t , la probabilité de séjour correspond à la probabilité de ne pas avoir connu l'événement ou la transition étudiée. Ce qui s'écrit⁸ :

$$S(t) = P(T > t) \quad (2.12)$$

Avec $S(0)$ ou $S(t_0)=1$

La distribution au cours du temps de la probabilité de survie est toujours décroissante. Un autre terme souvent utilisé à la place de fonction de séjour est celui de fonction de survie, ce terme rappelant que les notions statistiques de l'analyse des événements de l'histoire de vie ont été importées de l'analyse de survie. Le complémentaire à 1 de la probabilité de séjour à l'instant t va représenter la proportion des individus ayant connu l'événement depuis l'instant origine. Si $F(t)$ représente cette proportion, alors :

$$F(t) = 1 - S(t) = P(T \leq t) \quad (2.13)$$

2.3.2 Probabilité ou densité de connaître l'événement

La distribution de T peut aussi être décrite par la probabilité (en temps discret) ou la densité (en temps continu) de connaître l'événement, souvent symbolisée par $f(t)$. En temps discret comme en temps continu, $f(t)$ possède un lien avec la distribution de la proportion $F(t)$ des individus ayant connu l'événement. Ainsi, en temps discret, $f(t_i)$ va correspondre à la probabilité de connaître l'événement en t_i :

$$f(t_i) = F(t_i + t_{i+1}) - F(t_i) = P(T = t_i) \quad (2.14)$$

La probabilité $f(t_i)$ diffère du risque $P(t_i)$ tel que nous l'avons décrit en temps discret en ceci que dans l'expression de $f(t_i)$ n'intervient pas la condition selon laquelle T doit être supérieur ou égal à t . Des équations 2.7 et 2.12, il résulte que :

$$P(t_i) = \frac{f(t_i)}{S(t_i)} \quad (2.15)$$

En temps continu, si t' est supérieur à t , alors (Blossfeld and Rohwer 2002) :

$$f(t) = \lim_{t' \rightarrow t} \frac{F(t') - F(t)}{t' - t} = \lim_{t' \rightarrow t} \frac{P(t \leq T < t')}{t' - t} \quad (2.16)$$

Ce qui plus conventionnellement s'écrit :

$$f(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} \quad (2.17)$$

et dans ce cas, le risque $h(t)$ devient :

⁸ Rappelons que T , variable aléatoire, représente la durée écoulée depuis l'instant initial et celui d'occurrence de l'événement.

$$h(t) = \frac{f(t)}{S(t)} \quad (2.18)$$

2.3.3 Le risque cumulé

En temps continu, l'expression 2.18 correspond à l'opposée de la dérivée du logarithme de $S(t)$:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d\text{Log}S(t)}{dt} \quad (2.19)$$

En conséquence, en inversant et retournant à l'exponentielle :

$$S(t) = \exp\left(-\int_0^t h(s)ds\right) = \exp(-H(t)) \quad (2.20)$$

$H(t)$ est appelé le risque cumulé (« *cumulated risk* » ou « *integrated risk* ») :

$$H(t) = \int_0^t h(s)ds \quad (2.21)$$

Il va ainsi mesurer le nombre d'événement que subirait un individu tout au long de la période d'observation s'il était constamment soumis au risque de connaître l'événement au cours du temps. Il est à noter que le risque cumulé $H(t)$ permet de faire le lien entre $S(t)$ et $h(t)$:

$$H(t) = -\text{Log}S(t) \quad (2.22)$$

Enfin :

$$H(t) \cong \sum h(t) \quad (2.23)$$

Cette dernière propriété est exploitée dans les estimations non-paramétriques telles que les méthodes actuarielles⁹ ainsi que les méthodes de Nelson-Aalen (Courgeau et Lelièvre 1989, Therneau et Grambsch, 2000). La représentation graphique de $H(t)$ donnera une idée de l'évolution du risque au cours du temps.

3. Estimer un modèle

3.1 Préalables

Dans ce point, nous allons porter plus particulièrement notre intérêt sur la phase préparatoire qui précède l'estimation d'un modèle de l'analyse des biographies. Disons-le sans détour, aboutir à un modèle bien spécifié et le moins biaisé possible nécessite de suivre une règle générale de conduite consistant à ne pas conditionner le futur des individus. Cette règle d'or, *ne pas conditionner le futur des individus*, doit diriger toute la phase qui va de la collecte des

⁹ Cf. chapitre sur l'usage des méthodes non-paramétriques avec SPSS pour des précisions sur le risque $H(t)$.

données à la préparation de la variable dépendante (durées avant transitions) et des variables indépendantes, qu'elles soient constantes ou dépendantes du temps. Autant dire qu'il s'agit d'une règle peu facile à suivre...

3.1.1 Données orientées événements

Bien que cet article ne soit pas le cadre approprié pour un développement approfondi des techniques de collecte des données, le rappel de quelques éléments essentiels peut servir d'introduction à notre propos sur les dangers d'un conditionnement des individus par leur futur¹⁰. Les données recueillies sur lesquelles doivent être appliquées les méthodes d'analyse des biographies doivent être des données «*orientées événements*» (*Event Oriented Design*) (Blossfeld et Rohwer, 2002). Ce terme signifie que les techniques de collectes doivent être conçues de manière à relever les deux informations essentielles que sont, d'une part les événements du parcours de vie, et d'autre part les dates de chacun de ces événements. Ces dates doivent être les plus précises possible. Notons que la seule information sur le fait que l'individu a connu l'événement sans que l'on sache la date d'occurrence de cet événement n'a aucune utilité d'un point de vue de l'analyse des biographies.

Un premier type de collecte de données orientées événement est la collecte de données d'archives ou de fichiers administratifs. De telles données présentent souvent l'inconvénient de ne pas être complètes. Ainsi, la source de données la plus simple pour les démographes est l'ensemble des données de l'état civil ou des registres (Wanner, 2001). Ces données contiennent des informations seulement sur les dates de naissance, de mariage(s), les dates de naissance des enfants, les dates de divorce(s) si il y a lieu. En conséquence, il n'est pas possible de faire un lien entre le parcours de vie familial et d'autres domaines de l'existence tel le cheminement professionnel, à moins de disposer de procédures permettant d'assembler ces données avec d'autres données longitudinales. Par ailleurs, ces archives ne sont pas, loin s'en faut, conçues pour une exploitation statistique, a fortiori pour un usage des modèles de l'analyse des biographies. Quoiqu'il en soit, ces données constituent souvent une matière première pour les historiens ou les démographes historiens, car elles offrent la possibilité de connaître les modalités du parcours de vie dans les populations du passé.

Un deuxième type de données correspond aux données d'enquêtes rétrospectives. Il s'agit d'enquêtes visant à retracer le parcours de vie des individus interrogés, ces derniers n'étant en général interviewés qu'une seule fois. Ces enquêtes permettent de collecter uniquement des données factuelles et reposent sur la bonne mémoire des personnes enquêtées. En outre, de telles enquêtes peuvent parfois mener à un biais de sélection des personnes enquêtées. Ce serait, par exemple, le cas d'une enquête qui interrogeraient des personnes âgées de soixante-dix ans ou plus, alors que ces personnes appartiennent à des générations pour lesquels les effets de la mortalité dus à l'âge commence à devenir important (Courgeau et Lelièvre 1989). Dans ce cas, sous l'hypothèse que la trajectoire au cours du parcours de vie joue un rôle sur la mort, le fait même de réaliser une enquête auprès des personnes âgées de 70 ans constitue un processus de sélection qui revient en fait à conditionner le futur des personnes enquêtées, c'est-à-dire, à sélectionner des survivants (Wunsch, 2001). En d'autres termes, la trajectoire de vie de ces personnes ne reflètera pas nécessairement l'ensemble des trajectoires de ces générations. Dans le même ordre d'idées, un autre phénomène est particulièrement

¹⁰ Il existe d'ailleurs sur le sujet de la collecte des données biographique une littérature prenant de plus en plus d'importance. Cf. Freedman et al. (1988), Brückner et Mayer (1998), Groupe de réflexion sur l'approche biographique (1999), Blossfeld et Rohwer (2001), Ruspini (2002).

indispensable à prendre en compte si l'on s'intéresse à retracer les parcours de vie des personnes ayant immigré en Suisse dans les années cinquante et soixante. Ce phénomène est celui du départ d'un grand nombre d'étrangers de Suisse lors des crises économiques du milieu des années soixante-dix. Ceux qui sont restés, et qui sont donc susceptibles d'être couverts par une enquête sur l'immigration, peuvent ne pas être représentatifs de l'ensemble des immigrants.

Un troisième et dernier type de données est composé des données des enquêtes par panel, c'est-à-dire, d'enquêtes à passages répétés. Dans de nombreux cas, les enquêtes par panel se limitent à questionner les individus sur leur situation au moment de l'enquête, ce qui offre alors une comparaison avec la situation qu'ils occupaient au moment de la vague d'enquête précédente. Blossfeld et Rohwer (2002) font toutefois remarquer qu'une telle comparaison peut-être erronée, en développant l'exemple fictif d'une personne qui déclarerait être mariée aussi bien au temps $t+1$ qu'au temps t , mais qui aurait en fait divorcé et se serait remariée durant l'intervalle de temps séparant les deux enquêtes : même si les procédures d'enquête permettront le plus souvent de montrer qu'il y a eu un changement de partenaire, aussi bien la date du divorce que celle du remariage seront manquantes. L'idéal serait donc de disposer d'enquêtes par panel dans lesquelles les personnes seraient interrogées rétrospectivement à chaque fois sur les événements de leur parcours de vie depuis la vague précédente.

3.1.2 Censures et troncatures

S'il est un problème qui se pose assez naturellement lorsque l'on veut mettre en œuvre une analyse des biographies, c'est celui des individus qui ne connaissent pas l'événement ou la transition que l'on veut étudier. Ecarter ces personnes reviendrait à conditionner le futur des individus sélectionnés, puisque ces personnes, au moment t_0 d'entrée dans la population soumise au risque sont « condamnés » à connaître l'événement. En outre, l'exclusion des personnes qui ne connaissent pas l'événement durant la période d'observation serait une opération particulièrement artificielle en ce sens que ces individus peuvent connaître l'événement en question après qu'ils aient été interrogés dans le cadre d'une enquête.

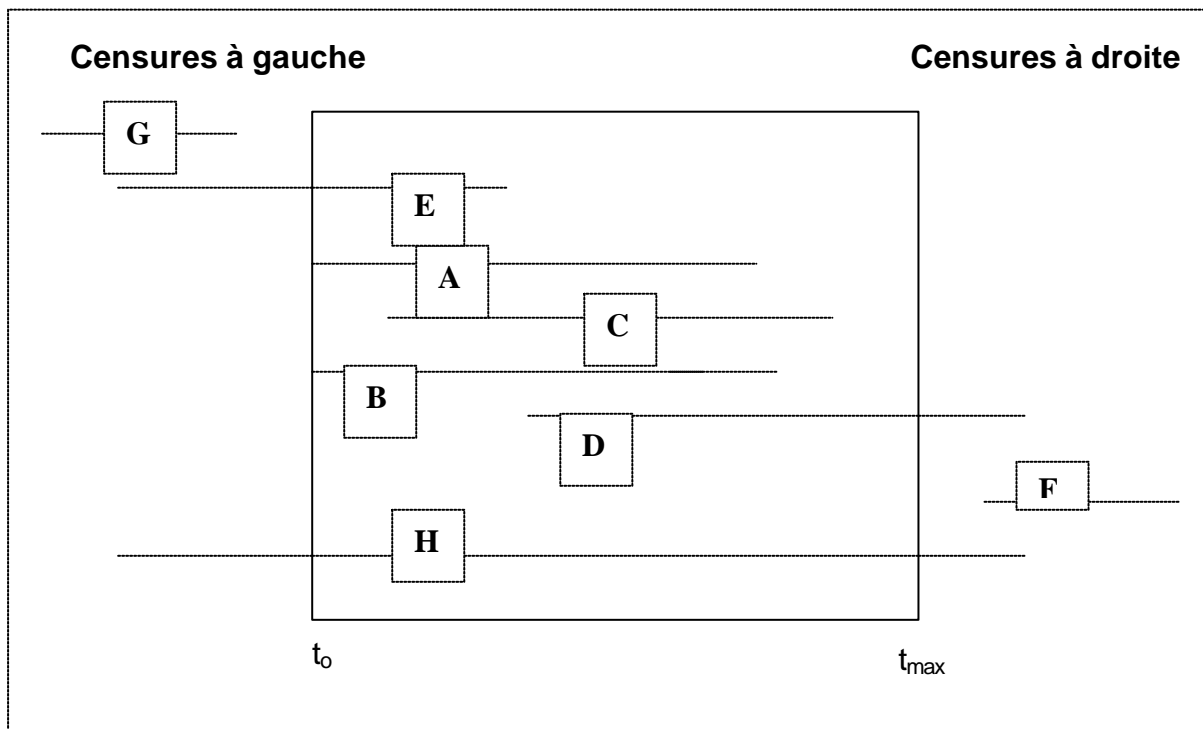
Le principe d'estimation d'un modèle d'analyse des biographies consiste plutôt à considérer que les individus pour qui n'est pas observée la transition étudiée *sont soumis au risque de connaître l'événement jusqu'au moment où il sortent d'observation*. On parle alors de durées qui sont *censurées*. Dans le cas particulier où l'on connaît le moment au cours duquel l'individu entre dans la population soumise au risque, mais où l'individu n'a pas connu l'événement au moment de l'enquête, on parle plus précisément de *censure à droite*. Le terme de *censure à gauche* est quand à lui employé dans le cas des individus pour qui on ne connaît pas le moment d'entrée dans la population soumise au risque ou dont le début d'observation est supérieur au moment d'entrée dans la population soumise au risque de connaître l'événement.

Un schéma permet de distinguer les différents cas de censure les plus souvent rencontrés (Yamaguchi 1991, Vermunt 1997, Blossfeld et Rohwer 2002). Dans ce schéma, un rectangle représente ce qui s'appelle une fenêtre d'observation, la largeur de cette fenêtre correspondant à l'intervalle de temps qui sépare le temps t_0 et le temps maximum t_{max} (moment de l'enquête) d'observation de la population soumise au risque de connaître l'événement (figure 2). Pour illustrer ce schéma, nous supposerons que l'événement étudié est le mariage de personnes célibataires et ne vivant pas en couples. Au temps t_0 (par exemple, à l'âge de 15 ans),

l'individu A et l'individu B sont soumis au risque de connaître l'événement dès cet instant t_0 . Sur ce schéma, C et D entrent un peu plus tard dans la population soumise au risque. Ainsi, si l'on s'intéresse au mariage en Suisse, et si C et D sont des personnes étrangères et célibataires arrivant en Suisse bien après qu'il aient eu 15 ans, alors ils entrent dans la population des personnes soumises au risque de se marier à l'âge qu'ils avaient lors de leur entrée en Suisse. A et C vont se marier durant la période d'observation alors que D quitte la fenêtre d'observation sans s'être marié. D est donc censuré à droite, même s'il se marie après le moment de l'enquête. On parle aussi dans le cas de D, d'une durée tronquée sur la droite (Yamaguchi, 1991). B, en revanche voit sa durée censurée avant qu'il sorte d'observation. Cela peut être dû au fait, par exemple, qu'il soit entré en union libre. Dans ce cas, il est sorti de la population des personnes vivant seules susceptibles de se marier. B est donc aussi censuré à droite. Un cas extrême est celui de l'individu F, puisque celui-ci entre dans la population soumise au risque de se marier après la période d'observation. Il n'interviendra donc pas même dans l'estimation des différents risques. On parle alors d'un individu « complètement censuré » (*fully censored*) sur la droite.

Dans la plupart des cas, les sorties d'observation interviennent de manière indépendante de l'occurrence de l'événement. Indépendance veut dire ici que ces sorties d'observation interviennent de façon aléatoire. En d'autre terme, les censures et les troncatures à droite ne constituent pas un biais dans l'estimation des paramètres d'un modèle d'analyse des biographies.

Figure 2 : censures et troncatures



Tel n'est pas le cas lorsqu'il y a censure à gauche. Ainsi, supposons que des individus étaient soumis au risque de se marier avant l'âge de 15 ans. Ce serait notamment le cas des individus E, G et H. En ce qui concerne E et H, ceux-ci entrent dans la population soumise au risque de connaître l'événement dès t_0 , mais on ne sait pas depuis quand (on parle alors de durées tronquées à gauche). E va connaître l'événement, alors que H va sortir d'observation sans

l'avoir connu. Les cas de troncatures à gauche posent problème en ce sens que l'on ne peut déterminer les durées d'observation. Mais le cas extrême reste celui de durées « *complètement censurées* » à gauche tel que le symbolise le parcours de G. Celui-ci connaît l'événement avant de pouvoir être comptabilisé dans la population soumise au risque. Ce cas de figure se rencontre, notamment, lorsque le temps t_0 correspond à un événement initial qui résulte d'un effet de sélection. Le cas typique est celui de l'analyse de la mortalité due au SIDA, dont la mesure se baserait uniquement sur les personnes révélant un test positif. Non seulement, on ne connaît pas le moment auquel les individus ont contracté le HIV, la conséquence étant que l'on dispose en fait uniquement de durées censurées à gauche, mais, en outre, sont oubliées toutes les personnes qui ne font pas de test de SIDA (Klein et Moeschberger 1997). En ce qui concerne les sciences sociales, des effets de censure et de troncature à gauche peuvent être rencontrés dans le cas de données d'enquêtes par panel.

3.1.3 Autres aspects importants à prendre en compte pour la spécification d'un modèle

Un autre aspect dans lequel le « risque » de conditionner le futur des individus est important se situe au niveau du choix des variables indépendantes x_{kt} lors de la phase de préparation des données (Yamaguchi 1991). Un exposé de ces mésusages est d'autant plus intéressant à détailler que ces erreurs n'empêchent pas l'estimation d'un modèle en ce sens qu'aucun message d'erreur ne vient signaler la mauvaise spécification d'un modèle au cours de la procédure d'estimation de celui-ci avec un logiciel statistique.

Un premier type d'erreur consiste ainsi à prendre en compte une variable normalement dépendante du temps en tant que variable fixe. Ce serait, par exemple, le cas d'une analyse de la naissance d'un premier enfant en relation avec la situation professionnelle des femmes au moment de l'enquête¹¹. Dans ce cas, l'utilisation de la situation professionnelle en tant que variable explicative du « risque » de mettre au monde un premier enfant conduirait inmanquablement à indiquer que la fécondité est plus élevée chez des femmes inactives que chez les femmes actives. Or, on sait qu'un grand nombre de femmes interrompent leur activité professionnelle pour une longue période lors de la naissance de leur premier enfant. Dans ces cas de figure, le mieux est de s'abstenir d'introduire ces variables dépendantes du temps que de les introduire sous la forme de variables fixes.

Un deuxième type d'erreur consiste, par exemple, à introduire la variable « s'est remarié » dans l'analyse d'un premier divorce, c'est-à-dire, non seulement d'anticiper le futur des individus, mais de le faire par une variable dont l'événement qui est analysé en est une condition.

Un troisième type d'erreur consiste, lorsque l'on dispose de parcours de vie incomplets, d'introduire une variable dépendante du temps relative à la dernière transition collectée et non à la première. Yamaguchi (1991) donne ainsi un exemple de mauvaise spécification d'un modèle d'analyse du mariage dans lequel est introduite la variable « dernière sortie d'une formation à plein-temps ». De manière similaire avec ce que nous avons vu concernant les censures à droite, une dernière sortie de formation observée ne signifie pas que les individus analysés ne sont pas retournés en formation après l'enquête. Dans la mesure du possible, la

¹¹ On se situe dans un cas fictif d'une enquête ayant relevé le parcours familial, mais seulement l'activité professionnelle au moment de l'enquête et non le parcours professionnel.

variable à introduire devrait être une variable se rapportant au moment où les individus sont sortis pour la première fois de formation.

3.2. Les modèles et leurs hypothèses

3.2.1 Les différentes approches

Dans le premier point de cet article, nous avons présenté le risque comme étant une fonction du temps et des caractéristiques des individus (cf. aussi Blossfeld et Rohwer, 2002).

$$h(t) = f^o(t, x_t) \quad (2.9)$$

Les différentes méthodes d'estimation reposent sur la manière d'aborder cette équation. Trois approches peuvent ainsi être distinguées :

- *Approche non-paramétrique* : les méthodes les plus courantes sont les méthodes de Kaplan-Meier et actuarielles, qui sont implémentées dans SPSS et qui font l'objet d'un article dans ce site internet. Une autre méthode un peu moins courante est la méthode de Nelson-Aalen, qui n'est pas implémentée dans SPSS (Therneau and Grambsch, 2000). Dans les méthodes non-paramétriques, on considère que le risque estimé au temps t est indépendant des risques qui ont été estimés aux temps précédents. En outre, on considère que la population est homogène, en ce sens que la distribution du risque est estimée pour l'ensemble de la population prise en compte (sans tenir compte des effets des caractéristiques des individus). Ceci revient donc à considérer que le risque ne dépend que de t . En outre, on cherche à connaître la distribution du risque au cours du temps. *Les modèles non-paramétriques sont donc des modèles essentiellement descriptifs*. Il est possible bien sûr de comparer deux ou plusieurs sous-populations, mais on ne peut distinguer la population de départ en tenant compte d'un grand nombre de caractéristiques pour des raisons d'effectifs. *Pour cette raison, les méthodes non-paramétriques répondent plutôt à des objectifs d'exploration des données*. L'analyse de résultats obtenus à partir de ces méthodes permet notamment de formuler des hypothèses sur la dépendance du risque au temps ainsi que sur les différences dans la distribution du risque en fonction des caractéristiques possédées par les individus (Courgeau et Lelièvre 1989, Blossfeld and Rohwer 2002). Il est à noter que ces méthodes permettent de prendre en compte les caractéristiques dépendantes du temps (cf., par exemple, Lelièvre 1987);
- *Approche semi-paramétrique* : avec cette approche, on entre dans la catégorie des *modèle explicatifs*. Cette approche tire son nom du fait qu'il n'est pas fait d'hypothèse sur la dépendance du risque au temps alors qu'il en est faite une sur les différences interindividuelles. Les estimations semi-paramétriques ont été proposées par Cox (1972), raison pour laquelle on les appelle aussi les *modèles de Cox*. Ce modèle repose sur l'*hypothèse de risques proportionnels*. Cette hypothèse signifie que le rapport des risques entre un individu qui possède une caractéristique donnée et un autre ne possédant pas cette caractéristique est constant, quel que soit l'instant t auquel on se situe. En d'autres termes, le risque d'un individu possédant une caractéristique donnée est multipliée par une constante, en comparaison avec les individus qui ne possèdent pas cette caractéristique. Cette hypothèse est généralisable à l'ensemble des

caractéristiques possédées par les individus. Ainsi si $h(t, x_t)$ représente le risque d'un individu possédant les caractéristiques x_t et si $h_0(t)$ représente le risque des individus pour qui toutes les caractéristiques x_t sont égales à 0 (individus de référence) alors le modèle s'écrit :

$$h(t, x_t) = h_0(t) \exp(x_{t1}b_1 + x_{t2}b_2 + \dots x_{tm}b_n) \quad (3.1)$$

b_1, b_2, \dots, b_n sont les coefficients du modèle à estimer. Ce modèle peut aussi s'écrire plus simplement :

$$h(t, x_t) = h_0(t) \exp x_t b \quad (3.2)$$

$$\text{où } x_t = (x_{1t}, x_{2t}, x_{3t}, \dots, x_{kt}) \text{ et } b = \begin{pmatrix} b_1 \\ b_2 \\ \dots \\ b_n \end{pmatrix}$$

La forme multiplicative du modèle permet d'éviter des estimations de coefficients conduisant à des risques négatifs, tels que cela pourrait l'être dans un modèle additif. Le passage par le logarithme permet de retrouver une forme linéaire plus classique :

$$\log h(t, x_t) = \log h_0(t) + x_t b \quad (3.3)$$

L'estimation des paramètres b repose sur la maximisation d'une équation de vraisemblance partielle (Cox, 1972). Il est à noter que des caractéristiques dépendantes du temps peuvent être introduites dans ce modèle. Cette possibilité constitue d'ailleurs un moyen de lever l'hypothèse de proportionnalité du risque si celle-ci ne s'avère pas être vérifiée. Les estimations obtenues sont robustes et constituent une bonne estimation des coefficients b , même en l'absence de spécification de la forme de la distribution du risque $h_0(t)$ au cours du temps (Courgeau et Lelièvre 1989). Pour cette raison, le modèle de Cox est un modèle très populaire (Kleinbaum, 1997). Il est sans doute le modèle le plus utilisé en sciences sociales ;

- *Approche paramétrique* : outre une hypothèse concernant le rôle joué par les caractéristiques, les modèles paramétriques reposent aussi sur une hypothèse qui concerne la forme de la distribution du risque au cours du temps. Cette hypothèse peut être posée à la suite de l'observation de la distribution du risque telle que celle-ci est obtenue à partir d'une estimation non-paramétrique. Elle peut toutefois découler d'une hypothèse théorique qui concerne le comportement des individus au fur et à mesure de l'écoulement du temps. Ainsi, l'hypothèse selon laquelle le risque de quitter un emploi diminue en raison de l'accumulation d'expérience professionnelle spécifique¹² peut être traduite par un modèle dans lequel la distribution du risque suit *une distribution de Gompertz*, c'est-à-dire, une distribution dans laquelle le logarithme du risque est une fonction linéaire du temps (Blossfeld et Rohwer, 2002). Avec l'hypothèse complémentaire de la proportionnalité des risques selon les caractéristiques possédées par les individus, le modèle s'écrit (Allison, 1984):

¹² Cf. premier point de cet article.

$$\log h(t, x_t) = at + bx_t + c \quad (3.4)$$

où c est une constante et a est un coefficient mesurant la pente du logarithme de la fonction de Gompertz. Les résultats de ce modèle, estimés à partir de la maximisation d'une équation de vraisemblance, iront dans le sens de l'hypothèse théorique si le coefficient c estimé est négatif. En revanche, un coefficient c positif indiquerait un risque croissant du temps et un coefficient c égal à 0 signifierait un risque constant au cours du temps, le modèle prenant alors le nom de *modèle exponentiel*, car la distribution $S(t)$ associée à un risque constant suit une distribution exponentielle.

L'hypothèse d'une diminution du risque de départ d'emploi pourrait être vérifiée à partir d'une autre modélisation paramétrique qui ferait intervenir une *distribution de Weibull* plutôt qu'une distribution de Gompertz. Dans ce deuxième cas, le logarithme du risque croît ou décroît de manière linéaire avec le logarithme du temps. Le modèle s'écrit ainsi :

$$\log h(t, x_t) = a \log t + bx_t + c \quad (3.5)$$

D'autres fonctions monotones du temps pourraient être utilisées pour modéliser la dépendance au temps, telle la fonction gamma. Toutefois, dans le cas d'autres hypothèses, d'autres distributions du risque peuvent être envisagées. C'est le cas, ainsi, de la distribution des risques lorsque l'événement étudié résulte d'un processus de diffusion d'un comportement au sein d'une population. Dans ce cas, la distribution est d'abord croissante, correspondant à la phase de dans laquelle de plus en plus d'individus adoptent le comportement, puis, est ensuite décroissante lorsque le comportement perd de son attractivité pour une partie de la population qui ne l'adoptera jamais. Une fonction monotone du temps ne peut être mobilisé dans ce cas précis, mais l'on pourra plutôt faire appel à une distribution log-logistique (Dieckmann 1990).

Pendant longtemps, les modèles paramétriques ont été critiqués. Il leur a ainsi été reproché une certaine rigidité due à la distribution statistique choisie. Un grand nombre de tests montraient que ces distributions n'étaient pas les plus adéquates. Même les modèles dont le choix de la distribution reposait sur une hypothèse théorique subissait la critique en raison de l'effet éventuel de l'hétérogénéité non observée. Un nouveau type de modèles paramétriques a toutefois connu un développement important dans les années récentes. Il s'agit des modèles *Piecewise* qui pourraient être traduits par « *modèles à risques coupés en morceau* ». Le principe de ces modèles est simple, puisqu'il consiste à estimer les paramètres d'une distribution seulement sur un court intervalle de temps et de recommencer sur l'intervalle de temps suivant et ainsi de suite. La distribution du risque est ainsi décomposée en une suite de distributions exponentielles (*modèle piecewise constant ou exponentiel*) ou de Gompertz (*modèle piecewise linéaire ou de Gompertz généralisé*). Ces modélisations paramétriques apparaissent ainsi extrêmement souples et peuvent donc s'adapter à n'importe quels types de données (Blossfeld et Rohwer 2002, Lillard and Panis 2003). Pour quelques auteurs, les modèles piecewise ne font plus partie de la catégorie des modèles paramétriques, mais plutôt de celle des modèles semi-paramétriques en raison de cette propriété de souplesse (Yamaguchi, 1991).

A ces trois approches, peut être ajoutée une quatrième :

- *Modèles à temps discret* (Allison 1982, 1984 et 1995, Yamaguchi 1991, Vermunt 1997) : les méthodes semi-paramétriques et paramétriques sont des approches dans lesquels le temps est considéré continu. Il peut se faire, néanmoins que les données dont on dispose ne soit pas assez fines pour mettre en œuvre ce type de méthode. Bien souvent cela est dû au fait que l'unité de temps prise en compte est trop grande, par exemple, l'année. Par ailleurs, il est à noter que l'estimation de modèles de Cox peut être erronée lorsqu'un nombre important d'individus connaît l'événement durant un même intervalle de temps (Cox 1972, Yamaguchi, 1991). Dans ce cas, les modèles à temps discret sont plus adéquats. Dans ces modèles, les hypothèses concernent aussi bien la distribution du risque que les différences interindividuelles, de façon similaire aux modèles paramétriques à temps continu. L'un des modèles les plus utilisés est le modèle *logit à temps discret* (*discrete time logit model*). Ce modèle consiste à estimer un modèle logit dans lequel la probabilité prise en compte est la probabilité de connaître l'événement à l'instant t sachant que l'on n'avait pas encore connu l'événement. Ce modèle s'écrit :

$$\log \frac{P(t, x_t)}{1 - P(t, x_t)} = a(t) + bx_t \quad (3.6)$$

$a(t)$ est une fonction du temps. Des spécifications de type Gompertz ou Weibull, c'est-à-dire, des spécifications dans lesquels il y a une relation linéaire entre d'une part, le logit du risque, et d'autre part, soit le temps, soit son logarithme peuvent être ajoutées à ce modèle (Allison, 1982). Toutefois, le temps peut aussi être introduit sous forme de variables binaires, chacune délimitant alors un intervalle de temps. Le modèle spécifié est alors analogue à un modèle *piecewise constant*.

3.2.2 Tableau récapitulatif

Le tableau 1 synthétise les différentes approches que nous venons de voir tout en indiquant les différentes procédures statistiques implémentées dans SPSS, lorsqu'elles existent, qui permettent de mettre en œuvre ces différentes approches. SPSS dispose d'un module SURVIVAL qui permet d'estimer des modèles non-paramétriques et semi-paramétriques. Il s'agit ainsi, en premier lieu, des procédures KM et SURVIVAL qui permettent respectivement de réaliser des estimations non-paramétriques de Kaplan-Meier et des estimations actuarielles. En second lieu, il s'agit de la procédure COX qui permet d'estimer des modèles de Cox, dans lesquels peuvent être introduites des variables dépendantes du temps.

SPSS n'offre pas la possibilité d'estimer des modèles paramétriques à temps continu. Ces modèles sont, en revanche implémentés dans un grand nombre de logiciels statistiques tels que SAS, Stata, S+, R ou TDA. A l'exception de TDA et Stata, les modèles reposent sur l'hypothèse de temps de sortie accéléré plutôt que sur l'hypothèse de risques proportionnels. Cette hypothèse repose sur l'idée d'une élasticité du temps, en ce sens que les différences interindividuelles sont analysées en termes de retardement ou d'avancement des échéances selon les individus et leurs caractéristiques. Il s'agit d'hypothèses peu utilisées dans le domaine des sciences humaines, même si les notions d'avancement et de retardement d'un phénomène, par exemple, de la fécondité, sont couramment évoquées dans les disciplines

démographiques (Kohler et al., 2002). Soulignons par ailleurs que SAS et Stata ne permettent pas d'estimer directement des modèles de type Piecewise¹³.

A ce module SURVIVAL s'ajoutent deux autres procédures pouvant être mobilisées en vue d'estimer des modèles à temps discret, bien qu'à l'origine elle n'aient pas été conçues pour la modélisation des événements de l'histoire de vie. Ces procédures sont d'une part la procédure LOGISTIC qui permet d'estimer des *modèles logit à temps discret* (Allison, 1984)¹⁴ et la procédure LOGLINEAR qui permet d'estimer des modèles log-linéaires du risque (*Log-Rate models*). Ces deux procédures offrent la possibilité de paramétrer des fonctions de Gompertz ou de Weibull en vue de tester une dépendance linéaire entre le logarithme du risque ou son logit et le temps ou son logarithme. La paramétrisation d'une fonction piecewise est également possible (Yamaguchi, 1991).

Tableau 1 : Les différentes méthodes et leur implémentation dans SPSS

Méthode	Non-paramétrique	Semi-paramétrique	Paramétrique	Temps discret
Hypothèse de l'effet d'une caractéristique sur le risque	-	- Risques proportionnels	- Risques proportionnels - Temps de sortie accéléré	- Risques (ou logits du risque) proportionnels
Hypothèse sur la distribution du risque au cours du temps	-	-	- Monotone - Croissant puis décroissant - Piecewise	Monotone - Croissant puis décroissant - Piecewise
Implémentation dans SPSS	- KM (Kaplan-Meier) - SURVIVAL (actuarielle)	- COX (possibilité d'introduire des variables dépendantes du temps)		- LOGISTIC - LOGLINEAR (modèles log-linéaires)

La préparation des données en vue de leur exploitation par l'une des procédures du module SURVIVAL est simple. Elle repose sur la création d'un *fichier épisode* (*spell*). Pour un individu, il y a autant de lignes qu'il y a d'événements appartenant à un même domaine. Ainsi dans le cas de l'analyse du divorce, par exemple, un individu sera caractérisé dans un fichier de données SPSS par autant de lignes qu'il a connu de mariages. Chaque épisode est caractérisé par un couple de données (T, c), c'est-à-dire, la durée observée T et l'indice de censure c . Cet indice sera ainsi égal à 1 si l'individu a connu l'événement et 0 sinon. A ce couple de données viennent s'ajouter l'ensemble des caractéristiques indépendantes que l'on souhaite introduire dans le modèle, que ces caractéristiques soient dépendantes du temps ou non. La préparation des données est un peu plus complexe dans le cas de l'estimation d'un modèle à temps discret (Mills 1999). S'il s'agit d'un modèle logistique, le fichier doit être un *fichier personne-période*. Un épisode sera ainsi décomposé en T lignes. Pour chacune des $T-1$ premières lignes sera associé un indice de censure égal à 0 alors que la dernière sera caractérisée par un indice de censure binaire indiquant si l'individu a connu l'événement ou non. Ce procédé peut aboutir à un fichier très long, notamment si l'unité de mesure du temps est le mois. La mise en œuvre d'un modèle log-linéaire exige plutôt de créer une autre

¹³ Dans le cas de SAS, Allison (1995) propose en fait d'estimer plusieurs modèles exponentiels, chacun d'entre eux étant estimé sur une intervalle de temps spécifique.

¹⁴ Notons que la procédure NORMREG permet aussi d'estimer des modèles à risques concurrents, c'est-à-dire, des modèles dans lesquels plusieurs échéances peuvent être prises en compte, par exemple, les différentes cause de mortalité ou l'entrée dans une union en distinguant s'il s'agit d'une union libre ou d'une union maritale.

construction des données visant à estimer à chaque instant et pour chacune des caractéristiques le nombre de personnes soumises au risque (Yamaguchi 1991, Mills 1999).

4. Conclusion

Les objectifs de l'analyse des biographies peuvent être résumés à partir des expressions-clés suivantes : analyse d'une transition au cours du temps pour une population soumise au risque ; distribution du risque d'occurrence des événements au cours du temps ; différenciation de cette distribution en fonction des caractéristiques individuelles, que ces caractéristiques soient invariantes ou dépendantes du temps. La définition de ces objectifs nous a permis d'introduire les différentes notions de la modélisation des événements du parcours de vie, notamment, les différentes notions probabilistes. La plus importante est la notion de risque. Dans sa formulation la plus générale, celui-ci est modélisé en tant que fonction du moment auquel on se situe d'une part, et des caractéristiques possédées par les individus ou qui décrivent le contexte dans lequel ils vivent, d'autre part.

A l'exception des approches non-paramétriques, l'estimation d'un modèle d'analyse des biographies fait appel à des techniques de régression reposant sur la maximisation d'une fonction de vraisemblance. Différentes procédures implémentées dans SPSS permettent d'estimer des modèles semi-paramétriques et des modèles à temps discret. Ce logiciel permet aussi de développer des approches non-paramétriques, ces dernières devant être utilisées en vue d'explorer les données, de tester des hypothèses concernant la dépendance du risque au temps ainsi que de tester des hypothèses sur les différences de risque entre les individus d'une population.

Les autres articles de ce dossier détaillent ces différentes procédures de modélisation des événements du parcours de vie.

Références bibliographiques

- Allison Paul (1982). "Discrete-Time Methods for the Analysis of Event Histories", *Sociological Methodology*, 13: 61-98.
- Allison Paul (1984), *Event History Analysis. Regression for Longitudinal Event Data*. Newbury Park: Sage.
- Allison Paul (1995). *Survival Analysis Using the SAS® System*. Cary: SAS Campus Drive.
- Blossfeld Hans-Peter and Mills Melinda (2001). « A Causal Approach of Interrelated Family Events. A Cross-National Comparison of Cohabitation, Nonmarital Conception and Marriage », *Canadian Studies in Population*, 28(2): 409-437.
- Blossfeld Hans-Peter and Nazio Tiziana (2003). "The diffusion of Cohabitation among Young Women in West Germany, East Germany and Italy", *European Journal of Population*, 19(1): 47-82.
- Blossfeld Hans-Peter and Rohwer Goetz (2002), *Techniques of Event History Modeling. New Approaches to Causal Analysis*. Mahwah, Lawrence Erlbaum Associates.
- Box-Steffensmeier Janet and Jones Bradford S. (1997). "Time is of the essence : Event History Models in Political Sciences". *American Journal of Political Science*, 41(4): 1414-1461.
- Braun Norman and Engelhardt Henriette (2002). *Diffusion Processes and Event History Analysis*. Rostock, Max Plank Institute for Demographic Research, MPIDR Working Paper WP 2002-07.
- Brien Michael J., Lillard Lee A., Waite Linda J (1999). "Interrelated Family-Building Behaviors: cohabitation, Marriage and Nonmarital Conception. *Demography*, 36(4): 535-551.
- Brückner Erika and Mayer Karl Ulrich (1998). « Collecting Life History Data: Experiences from the German Life History Study », in Giele Janet Z., Elder Glen H. Jr. (eds), *Methods of Life Course Research. Qualitative and Quantitative Approaches*. Thousand Oaks-London-New Dehli: Sage Publications: 152-181.
- Caselli Graziella, Valin Jacques et Wunsch Guillaume (2001). *Démographie : Analyse et synthèse. I. Dynamique des populations*. Paris : INED.
- Coleman James S. (1981). *Longitudinal Data Analysis*. New York: Basic Books, Inc., Publishers
- Conell Carol and Cohn Samuel (1995). "Learning from other People's Actions: Environmental Variations and Diffusion in French Coal Mining Strikes, 1890-1935". *American Journal of Sociology*, 101(2): 366-403.

- Courgeau Daniel (2003). "From the macro-micro opposition to multilevel analysis in demography". In Courgeau Daniel (ed.), *Methodology and Epistemology of Multilevel Analysis. Approaches from Different Social Sciences*. Dordrecht: Kluwer Academic Publishers: 43-91.
- Courgeau Daniel et Lelièvre Eva (1986). « Nuptialité et agriculture ». *Population*, 2: 303-326.
 Courgeau et Lelièvre Eva (1989). *Analyse démographique des biographies*. Paris INED.
- Cox David R. (1972). "Regression models and life-tables". *Journal of the Royal Statistical Society, Series B*, 34:187-202.
- Cox David R. and Oakes David R. (1984). *Analysis of Survival Data*, London: Chapman and Hall.
- Dieckmann Andreas (1990). "Diffusion and Survival Models for the Process of entry into Marriage" in Tuma Nancy and Mayer Karl-Ulrich (eds.), *Event History Analysis in Life course Research*. Madison: University of Wisconsin Press: 170-183.
- Freedman Deborah, Thornton Arland, Camburn Donald, Alwin Duane and Young-DeMarco Linda (1988). « The Life History Calendar: A technique for Collecting Retrospective Data ». *Sociological Methodology*, 18: 37-68.
- Greve Heinrich R., Strang David and Tuma Nancy B. (1995). "Specification and estimation of heterogeneous diffusion models". *Sociological Methodology*, P.V. Marsden ed., 25 : 377-420.
- Groupe de réflexion sur l'approche biographique (1999). *Biographies d'enquêtes. Bilan de 14 collectes biographiques*. Paris: INED-IRD-PUF/diffusion.
- Hank Karsten (2002). "Regional Social Contexts and Individual Fertility Decisions: A Multilevel Analysis of First and Second Births in Western Germany". *European Journal of Population*. 18(3): 281-299.
- Heckmann J.J. and Singer B. (1984). "A method for minimizing the impact of distributional assumptions in econometrics models for duration data". *Econometrica*. 52:271-320.
- Klein John P. and Moeschberger Melvin L. (1997). *Survival Analysis. Techniques for Censored and Truncated Data*. Berlin-New York. Springer Verlag.
- Kleinbaum David D.(1997). *Survival Analysis. A self-Learning Text*. New-York, Berlin : Springer Verlag.
- Kohler Hans-Peter, Billari Francesco and Ortega Jose (2002). "The Emergence of Lowest Low Fertility in Europe During the 1990s", *Population and Development Review*: 641-679.
- Lancaster Tony (1979). "Econometric Methods for the Duration of Unemployment", *Econometrica*, 939-956.

- Lancaster Tony (1992). *The Econometric Analysis of Transition Data*, Cambridge: Cambridge University Press.
- Lelièvre Eva (1987). «Activité professionnelle et fécondité: les choix et les déterminations des femmes françaises entre 1930 et 1960 ». *Cahier Québécois de Démographie*, 16(2) : 207-236.
- Lelièvre Eva, Bonvalet Catherine et Bry Xavier (1997). «Analyse biographique des groupes. Les avancées d'une recherche en cours ». *Population*, 52(4) : 803-830.
- Lillard Lee A. (1993). "Simultaneous Equations for Hazards: Marriage Duration and Fertility Timing". *Journal of econometrics*, 56:189-217.
- Lillard Lee A. and Waite Linda J. (1993). "A Joint Model of Marital Childbearing and marital Disruption". *Demography*, 30(4): 653-681.
- Lillard Lee L. and Panis Constantijn W.A. (2003). *aML User's Guide and Reference Manual*. Los Angeles: Econware.
- Mayer Karl Ulrich and Tuma Nancy Brandon (1990). "Life Course Research and Event History Analysis : An Overview". In Mayer Karl Ulrich and Tuma Nancy Brandon (eds), *Event History Analysis in Life Course Research*. Madison: The University of Wisconsin Press: 3-20.
- Mills Melinda (1999). *Construction of Input Data for Log-Linear Models of Event Histories*. Groningen : Population Research Centre-University of Groningen. Working Paper n° 3.
- Palloni Alberto (2001). "Diffusion in sociological Analysis". In Casterline John B. (eds), *Diffusion Processes and Fertility Transition. Selected Perspective*. Washington: National Academic Press: 66-114.
- Pressat Roland (1983). *L'analyse démographique*. Paris : PUF (3^e ed.).
- Rohwer Götz and Pötter Ulrich (2002). *TDA user's manual*. Bochum. Ruhr Universität Bochum. <http://steinhaus.stat.ruhr-uni-bochum.de/tman.html>
- Ruspini Elisabetta (2002). *Introduction to Longitudinal Research*. London : Routledge.
- Strang David (1990). "From Dependency to Sovereignty: An Event History Analysis of Decolonization 1870-1987". *American Sociological Review*, 55(6): 846-860.
- Strang David and Tuma Nancy B. (1993). « Spatial and Temporal Heterogeneity in Diffusion», *American Journal of Sociology*, 99(3): 614-639.
- Therneau Terry M. and Grambsch Patricia (2000). *Modeling Survival Data. Extending the Cox Model*. New-York, Berlin: Springer Verlag. Statistics for Biology and Health.
- Toulemon (1996). «La cohabitation hors-mariage s'installe dans la durée », *Population*, 3 : 675-715.
- Tuma Nancy B. and Hannan Michael T. (1984). *Social Dynamics. Models and Methods*. Orlando : Academic Press.

- Vaupel Jim, Manton Kenneth G. and Stallard Eric. (1979). «The impact of heterogeneity in individual frailty on the dynamics of mortality », *Demography* 16: 439-454.
- Vermunt Joeren K. (1997). *Log-Linear Models for Event Histories*. Newbury-Park: Sage publications.
- Wanner Philippe (2001). « Analyse biographique des événements familiaux. Un modèle d'application de la statistique du mouvement naturel de la population ». *Démos*, 3/2001: 1-22.
- Wunsch Guillaume (2001). « L'observation démographique longitudinale », in Caselli Graziella, Vallin Jacques et Wunsch Guillaume, *Démographie : Analyse et synthèse. I. Dynamique des populations*. Paris : INED : 149-163.
- Yamaguchi Kasuo (1991). *Event History Analysis*. Newbury-Park, London: Sage Publications.
- Yamaguchi Kasuo (1994). "Some accelerated Failure-Time Regression Models Derived from Diffusion Process Models: An Application to a Network Diffusion Analysis", *Sociological Methodology*, 24:267-300.