Comment passer des anciennetés aux durées ? illustration à partir de l'enquête famille et logement de 2011.

Vianney COSTEMALLE[1]*(*)*

(*) Insee, Unité des Études Démographiques et Sociales

Résumé

L'ancienneté d'une situation mesure depuis combien de temps une personne est dans cette situation tandis que la durée mesure combien de temps la personne va rester dans cette situation. Beaucoup d'enquêtes transversales donnent accès aux anciennetés des répondants au moment de l'enquête, mais souvent elles ne permettent pas de connaître les durées. À l'aide de simulations on explore les liens entre ces deux notions, montrant qu'il faut être prudent lorsqu'on passe d'un discours sur les anciennetés à un discours sur les durées. On indique en particulier que deux effets antagonistes, l'un de censure, l'autre de sélection, font que les anciennetés donnent des estimations biaisées des durées, et que ces deux biais peuvent s'annuler l'un l'autre dans des cas particuliers. On montre aussi à partir d'une modélisation développée par S. Nickell qu'on peut inférer les durées à partir des anciennetés si on connaît par ailleurs les flux d'entrée dans la situation étudiée. On introduit en plus des risques instantanés proportionnelles comme dans le modèle de Cox et on modélise le risque instantané de base par une fonction constante par morceaux. On estime le modèle ainsi obtenu par maximum de vraisemblance et avec des simulations, on montre que cela permet de rectifier les deux biais. On estime alors notre modèle sur l'Enquête Famille et Logement afin de déterminer combien de temps les adultes restent à la tête d'une famille monoparentale.

Mots-clés

Modèle de durée, simulations, échantillonage dans le stock, anciennetés, familles monoparentales.

1. vianney.costemalle@insee.fr

Introduction

L'ancienneté d'une situation est le temps qui s'est écoulé entre le début de cette situation et le moment où elle est observée, tandis que la durée est le temps total de cette situation, entre le début et la fin de la situation. Les anciennetés et les durées sont donc a priori deux concepts différents qui répondent respectivement aux questions «Depuis combien de temps est-on dans cette situation ?» et «Combien de temps cette situation dure-t-elle ?». L'origine du problème présenté ici vient de l'Enquête Famille et Logement (EFL) de 2011 dans laquelle seules les anciennetés des familles monoparentales [2] au moment de l'enquête sont connues. Dès lors, comment montrer que des différentiels sociaux par exemple ont un impact sur la durée de ces familles monoparentales ? Est-ce que ceux qui sont depuis plus longtemps en famille monoparentale sont également ceux qui vont rester plus de temps dans cette situation ? Le problème est de savoir s'il est possible d'inférer des informations sur les durées à partir des anciennetés au moment de l'enquête, et si cela n'est pas possible de savoir quelles informations supplémentaires il est nécessaire de connaître pour mener cette inférence.

Dans toute la suite, on s'interessera de façon générale à des situations à durée limitée qu'on appellera *situations d'intérêts*, et on utilisera les familles monoparentales comme une illustration de telles situations. Les méthodes et résultats présentées ici pourront alors aussi s'appliquer à d'autres situations d'intérêts, comme les périodes de chômage ou les périodes de mariages par exemple. L'enquête EFL a été conduite parallelement au recensement de 2011 sur 360000 personnes et permet de compter le nombre de familles monoparentales à la date du recensement. De plus, pour chacune de ces familles monoparentales, on connaît la cause d'entrée dans cette situation (séparation d'avec le conjoint, décès du conjoint ou naissance d'un enfant hors relation de couple cohabitant) ainsi que l'année d'entrée. On peut donc en déduire les anciennetés avec la précision de l'année. On sera par ailleurs amené à utiliser une autre enquête, l'Étude des Relations Familiales et Intergénérationnelles (ERFI), qui a été menée par l'Ined et l'Insee en trois vagues successives, en 2005, 2008 et 2011, sur 10000 répondants, en France métropolitaine. Cette enquête permet, grâce à ses questions retrospectives, de connaître tout le passé conjugal des répondants lorsqu'ils ont vécu plusieurs unions. Lorsqu'on fait le rapprochement entre le passé conjugal et la présence des enfants mineurs, on peut alors déterminer les périodes de monoparentalité pour chaque adulte répondant d'ERFI. On a ainsi accès à des durées (et non des anciennetés) de monoparentalité, censurées sur la droite. Mais le désavantage d'une telle enquête est l'effectif limité (il y a environ 1700 répondants qui ont été au moins une fois à la tête d'une famille monoparentale) et les problèmes, liés à la mémoire, de détermination des dates par les répondants. ERFI permettra d'apporter des informations supplémentaires nécessaires à l'estimation du modèle. De plus on pourra aussi comparer les résultats obtenus à partir de l'EFL et ceux obtenus à partir d'ERFI.

Les modèles de durées classiques sont basés sur des *échantillonages dans le flux*, c'est-à-dire dire d'un échantillonage sur l'ensemble des personnes entrant dans la situation pendant une période donnée. Au contraire, les données issues de l'EFL résultent d'un *échantilonnage dans le stock*, c'est-à-dire qu'on observe les personnes qui sont dans la situation à un instant donné (le moment de l'enquête). De plus, toutes les données sont censurées

2. Une famille monoparentale est constituée d'un parent vivant sans conjoint dans le logement et ayant au moins un enfant mineur vivant dans le même logement.

sur la droite, car les anciennetés observées sont en fait des durées censurées à droite. On développe ici une modélisation qui avait déjà été proposée par Nickell [1] pour estimer la probabilité conditionnelle de quitter une situation de chômage en fonction du temps passé dans cette situation. Néanmoins, comme on le verra par la suite, cette modélisation diffère légèrement de celle de Nickell en ce qui concerne le risque instantané, ce qui lui donne plus de souplesse et permet d'estimer des situations diverses, notamment le cas des durées des familles monoparentales. Si on compare les anciennetés aux durées, alors par définition les anciennetés sont plus courtes que les durées. Toutefois, on observe les anciennetés de ceux qui sont dans la situation d'intérêt au moment de l'enquête, et qui ne sont qu'une partie de tous ceux qui vivent à un moment ou un autre de leur vie cette situation d'intérêt. Mécaniquement, ceux qui passent moins de temps dans la situation vont avoir moins de chance d'être enquêté lorsqu'ils sont dans cette situation. Ainsi, ceux qui sont dans la situation au moment de l'enquête ont en réalité des durées en moyenne plus grandes que l'ensemble des personnes qui vivent la situation à un moment ou un autre de leur vie.

Dans une première partie on modélise le problème et on calcul la vraisemblance des observations. Puis dans une seconde partie, on illustre à l'aide de simulations les différences entre anciennetés et durées ainsi que la robustesse du modèle. Dans une troisième partie on présente les résultats obtenus à partir de l'EFL qu'on compare aux résultats issus d'ERFI. Enfin dans une dernière partie, on discute des hypothèses du modèle et de ses limites.

1 Modélisation du problème

On considère un groupe de m personnes qui vont chacune vivre une et une seule fois une situation de monoparentalité en tant qu'adulte au cours de leur vie. On note alors pour chaque personne i, D_i l'année où sa situation de monoparentalité a débutée et F_i l'année où cette situation s'est terminée. On définit alors $T_i = F_i - D_i$ comme le temps passé en famille monoparentale. On a donc affaire ici à une variable de durée T prenant des valeurs discrètes [3] $(0, 1, 2, ...)$. Afin de simplifier le problème on considère ici l'hypothèse suivante :

Hypothèse 1 : *On ne vit au plus qu'une seule situation d'intérêt au cours de sa vie.*

Cette hypothèse n'est pas contraignante, car on peut toujours considérer une personne ayant vécu deux périodes de monoparentalité comme deux personnes différentes. Comme le montre la figure 1 tous ceux qui vivent une situation de monoparentalité ne sont pas forcément dans cette situation au moment de l'enquête. Dans l'exemple donné, seul l'année de début de monoparentalité de l'individu A (2009) est observée. Pour les individus B et C, on n'observe rien, soit parce qu'ils ont vécu leur période de monoparentalité avant l'enquête, soit parqu'ils l'ont commencée après. On observe donc un sous-échantillon de taille n des m individus ($n \leq m$) et pour chaque personne de ce sous-échantillon on connaît les années de début de monoparentalité $\{d_1, ..., d_n\}$, mais pas les années de fin de monoparentalité. Il s'agit donc d'un échantillonnage dans le stock dont *toutes* les durées sont censurées sur la droite.

3. Dans la suite, on considérera parfois le cas où T est continue et ce sera alors explicitement dit.

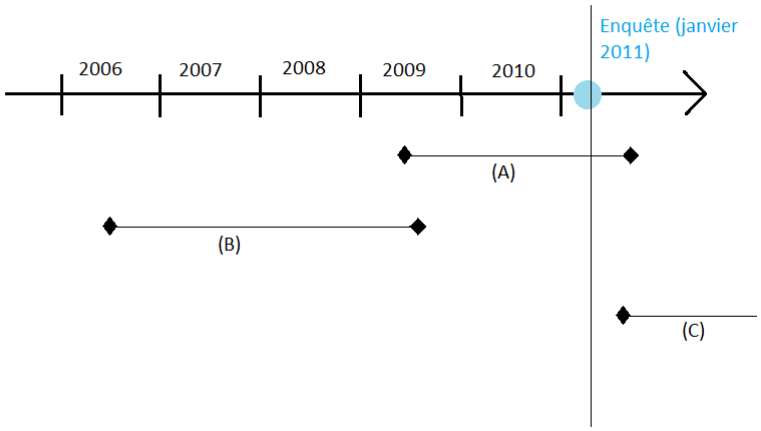

FIGURE 1: Illustration de l'échantillonage dans le stock résultant de l'enquête EFL.

Avant d'aller plus loin, on rappelle quelques résultats sur les variables aléatoires discrètes modélisant des durées. On définit en particulier la **survie** au temps t ($t \in \mathbb{N}$) comme étant la probabilité que la situation dure au moins t unité de temps, qu'on note alors $S_T(t) = \mathbb{P}(T \geq t)$. La **densité** de T correspond ici à la probabilité que la situation se termine au temps t et on note $f(t) = \mathbb{P}(T = t)$. Enfin, le **risque instantané** au temps t est définit comme étant la probabilité que la situation se termine au temps t sachant qu'elle a duré jusqu'au temps t. Autrement dit, c'est la proportion de personnes parmi ceux qui ont survécu jusqu'au temps t qui quittent la situation au temps t. On écrit $h(t) = \mathbb{P}(T = t|T \geq t) = \frac{\mathbb{P}(T=t)}{\mathbb{P}'(T \geq t)} = \frac{f(t)}{S_T(t)}$. On peut alors montrer qu'il y a une relation entre la survie au temps t et le risque instantané aux temps $\tau < t$ (voir [2]) :

$$\forall t \in \mathbb{N}^*, \quad S_T(t) = \prod_{\tau=0}^{t-1}(1 - h(\tau)) \quad et \quad S_T(0) = 1. \tag{1}$$

Ainsi, chacune des trois fonctions S_T, f et h permet de caractériser entièrement la loi de la variable aléatoire T.

1.1 Calcul de la vraisemblance

On cherche ici à calculer la vraisemblance des observations $\{d_1, ..., d_n\}$. On suppose de plus qu'on connaît pour chaque individu des caractéristiques individuelles qu'on note sous la forme de la covariable X. La participation à la vraisemblance d'un individu i dont les caractéristiques sont X_i s'écrit alors à l'aide de la formule de Bayes :

$$\begin{aligned} & \mathbb{P}(D_i = d_i | \text{i est en famille monoparentale au moment de l'enquête et } X_i) \\ = \; & \mathbb{P}(D_i = d_i | D_i \leq 2010 < D_i + T_i,\, X_i) \\ = \; & \frac{\mathbb{P}(D_i \leq 2010 < D_i + T_i | D_i = d_i,\, X_i)\mathbb{P}(D_i = d_i | X_i)}{\mathbb{P}(D_i \leq 2010 < D_i + T_i | X_i)} \end{aligned}$$

Sous l'hypothèse que T_i et D_i sont indépendantes (hypothèse 2), on peut alors calculer la probabilité d'être en famille monoparentale au moment de l'enquête sachant l'année de début de monoparentalité $\mathbb{P}(D_i \leq 2010 < D_i + T_i | D_i = d_i,\, X_i) = \mathbf{1}_{\{d_i \leq 2010\}} \mathbb{P}(T_i > 2010 - d_i | X_i) = \mathbf{1}_{\{d_i \leq 2010\}} S_T(2011 - d_i, X_i)$, avec $S_T(., X)$ la fonction de survie de T qui dépend des caractéristiques individuelles X [4].

Hypothèse 2 : *La variable de durée T est indépendante de la variable de flux D.*

De même, on calcul la probabilité d'être en famille monoparentale au moment de l'enquête, $\mathbb{P}(D_i \leq 2010 < D_i + T_i | X_i) = \sum_u \mathbb{P}(D_i \leq 2010 < D_i + T_i | D_i = u,\, X_i) \mathbb{P}(D_i = u | X_i) = \sum_{u \leq 2010} S_T(2011 - u, X_i) \mathbb{P}_D(u | X_i)$ où $\mathbb{P}_D(u|X)$ est la probabilité d'entrer en famille monoparentale l'année u, pour les individus dont les caractéristiques individuelles sont X. La vraisemblance devient alors :

$$\mathcal{L}(\{d_1, ..., d_n\}, \{x_1, ..., x_n\}) = \prod_{i=1}^{n} \left[\frac{S_T(2011 - d_i, x_i) \mathbb{P}_D(d_i | X = x_i)}{\sum_{u \leq 2010} S_T(2011 - u, x_i) \mathbb{P}_D(u | X = x_i)} \right] \tag{2}$$

Il s'agit maintenant de trouver la loi de la variable de durée T qui maximise la vraisemblance. Pour caractériser cette durée, plusieurs approches sont possibles. Toute l'information sur la loi de T est contenue soit dans la fonction de survie, soit dans la fonction de risque instantané, soit dans la fonction de densité. On choisit ici, après avoir tester les différentes possibilités, d'estimer le risque instantané, puis dans déduire les caractéristiques de la durée T. De plus, il faut noter qu'il est nécessaire de connaitre $\mathbb{P}_D$ à une constante près afin de pouvoir calculer effectivement la vraisemblance.

1.2 Modélisation du risque instantané

La fonction de survie qui apparaît dans l'équation de vraisemblance (2) peut se réécrire en fonction du risque instantané à l'aide de la relation (1). Dans l'article de Nickell, le risque instantané h est modélisé comme une fonction du temps et des caractéristiques individuelles X selon une loi logit : $logit(h(t, X)) = X\beta + a_1 t + a_2 t^2$. Cette paramétrisation qui a l'avantage d'être simple n'est pas assez flexible pour pouvoir rendre compte de situations diverses. Afin d'avoir plus de liberté dans l'estimation du risque instantané, on définit comme dans le modèle de Cox [3] un modèle semi-paramétrique avec risques instantanés proportionnels :

$$h(t, x, \theta) = \begin{cases} h_0(t, \alpha) e^{\beta x} \\ 1 \qquad \text{si } h_0(t, \alpha) e^{\beta x} \geq 1 \end{cases} \tag{3}$$

où $\theta = (\alpha, \beta)$ est le vecteur des paramètres du modèle.

Hypothèse 3 : *Le rapport des risques instantanés ne dépend que des covariables X et pas du temps (hypothèse de proportionnalité).*

On modélise alors le risque instantané de base h_0 par une fonction constante par morceaux. Pour cela, on découpe l'échelle du temps en petites périodes de même durée d_h

4. $S_T(t, x) = \mathbb{P}(T \geq t | X = x)$.

sur lesquelles le risque instantané va être constant. Pour tenir compte du fait que pour les longues périodes il y a très peu de données, on fait en sorte que le risque instantané soit constant à partir de la k^{me} période. Comme la modélisation est discrète, le risque instantané doit forcément être compris entre 0 et 1. On écrit alors le risque instantané de base :

$$h_0(t, \alpha) = \frac{1}{1 + e^{\alpha_{j(t)}}}$$

où $j(t)$ indique le numéro de la période dans laquelle se trouve t :

$$j(t) = \begin{cases} r & \text{si } (r-1)d_h \leq t < rd_h \text{ et } r \leq k \\ k & \text{si } (k-1)dp_h \leq t \end{cases} .$$

Un exemple d'une telle fonction est donnée sur la figure 2.

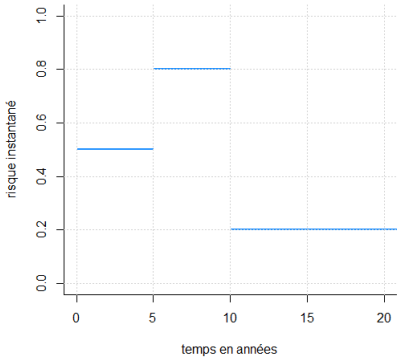

FIGURE 2: Exemple de risque instantané constant par morceau, avec $d_h = 5$ et $k = 3$.

Le paramètre $\beta = (\beta_1, ..., \beta_p)$ correspond au paramètre de régression du logarithme du risque instantané sur les covariables X et permet de mesurer l'effet d'une caractéristique individuelle sur le risque instantané. Si $\beta_j > 0$ cela signifie que le risque instantané de quitter la situation augmente et que la survie diminue lorsque X^j augmente. C'est l'inverse si $\beta_j < 0$. Si la covariable X_j est binaire, alors β_j mesure environ le taux de variation du risque instantané lorsque X_j passe de 0 à 1. Il y a donc $k + p$ paramètres à estimer $(\alpha_1, ..., \alpha_k$ et $\beta_1, ..., \beta_p)$ à l'aide de la méthode du maximum de vraisemblance. À première vue, le modèle ainsi spécifié semble être identifiable. Il reste néanmoins à pouvoir donner une estimation pour la probabilité d'entrée en famille monoparentale une année u, donnée par $\mathbb{P}_D(u)$.

1.3 Estimation de la probabilité d'entrée dans la situation d'intérêt

La probabilité de débuter la situation d'intérêt l'année u est proportionnelle aux flux d'entrées dans la situation et dépend des caractéristiques individuelles X. $\mathbb{P}_D$ apparraissant au numérateur et au dénominateur de la vraisemblance, la constante de proportionnalité n'est pas utile pour calculer la vraisemblance. Il suffit donc d'avoir une estimation du flux d'entrée, à un constante de proportionalité près, en fonction des covariables X. Il faut en

général une source annexe pour pouvoir le déterminer. Si X est continue, il est possible d'estimer la probabilité d'entrée une année donnée en fonction de X à l'aide d'une régression logistique multinomiale. Si X prend un nombre fini (et petit) de valeurs alors il faut estimer le flux d'entrée pour chaque sous-population correspondante. On estime ici ce flux à l'aide de l'Étude des relations familiales et intergénérationnelles (ERFI), en lissant les valeurs à l'aide d'une moyenne mobile pour supprimer le bruit aléatoire. Ainsi, on dispose de $\hat{\mathbb{P}}_D$ (voir figure 3).

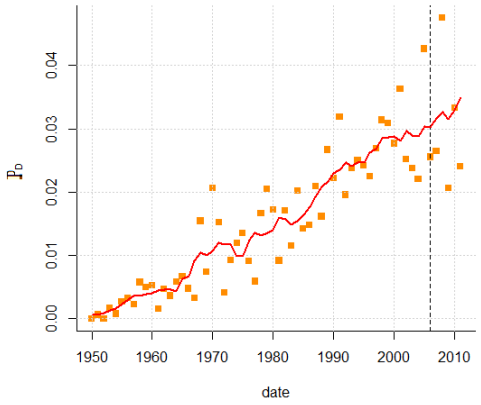

FIGURE 3: Estimation de $\mathbb{P}_D$ lorsqu'on ne considère aucune covariable. Les points oranges correspondent aux flux d'entrée en famille monoparentale et la ligne rouge correspond au lissage.
Champ : personnes entrées en famille monoparentale avant 2011 et étant âgées de 18 à 73 ans en 2005, France métropolitaine. Source : Ined-Insee, ERFI, 2005,2011.

La vraisemblance à maximiser s'écrit donc :

$$\mathcal{L}(d_1,...,d_n|x_1,...,x_n,\theta) = \prod_{i=1}^{n}\left[\frac{\hat{\mathbb{P}}_D(d_i,x_i)\prod_{\tau=0}^{2010-d_i}(1-\frac{e^{\beta x_i}}{1+e^{\alpha_{j(\tau)}}})}{\sum_{u\leq 2010}\hat{\mathbb{P}}_D(u,x_i)\prod_{\tau=0}^{2010-u}(1-\frac{e^{\beta x_i}}{1+e^{\alpha_{j(\tau)}}})}\right] \quad (4)$$

On fera de plus l'hypothèse suivante :

Hypothèse 4 : *Les covariables X ne varient pas avec le temps.*

Cette hypothèse n'est pas nécessaire au calcul de la vraisemblance, mais doit être faite compte tenu des informations dont on dispose. En effet, on ne connait les variables explicatives qu'à un instant donné, qui est le moment de l'enquête. On supposera donc que les variables considérées gardent toujours la valeur qu'elles prennent au moment de l'observation, c'est-à-dire de l'enquête.

2 Illustration à partir de simulations

Dans toute cette partie on utilise des simulations afin d'illustrer différents résultats et de tester des hypothèses. Bien que la modélisation proposée précédemment repose sur

des lois discrètes, on simulera ici des variables aléatoires continues, puis on passera aux observations en ne gardant que la partie entière. De plus, on considerera l'ancienneté dans la situation d'intérêt [5] qu'on note $A = 2011 - D$.

2.1 Effet de censure et effet de sélection

Si on considère l'ancienneté comme une durée censurée sur la droite, alors cette ancienneté donne une estimation doublement biaisée de la durée. D'une part, la probabilité d'être enquêté alors qu'on est dans la situation d'intérêt augmente avec la durée passée dans cette situation, ce qui fait que les personnes dans la situation au moment de l'enquête vont avoir un moyenne des durées plus grandes. Ce biais est appelé *biais de sélection*. D'autre part, les anciennetés sont par définition plus courtes que les durées, c'est ce qu'on appelle le *biais de censure*. Ces deux biais étant de sens contraire ont donc tendance à se neutraliser. Peut-on savoir si le biais de sélection est plus fort que le biais de censure, ou si c'est le contraire ? A priori, aucun des deux n'a de raison de l'emporter sur l'autre. En fait cela dépend, comme on va le voir par la suite, fortement de la loi de la variable de durée T.

2.1.1 Illustration

Pour simplifier, on se place ici dans le cas particulier où le flux d'entrée dans la situation est constant, c'est-à-dire que $\mathbb{P}_D$ est constant. On simule D selon une loi uniforme et T selon une loi de Weilbull de paramètre de forme k et de paramètre d'échelle λ. Lorsque $k = 1$, T suit une loi exponentielle, et plus k est grand, plus T est piquée autour de sa moyenne (sa variance diminue). En comparant la survie de l'ancienneté ($\mathbb{P}(A \geq t)$) à la survie de la durée réelle ($\mathbb{P}(T \geq t)$), on peut trouver des cas où l'effet censure est le plus fort et des cas où au contraire l'effet sélection l'emporte (figure 4). Dans le cas particulier où $k = 1$, les deux effets se neutralisent entièrement et la répartition des anciennetés est la même que celle des durées. En fait, il semble ici que l'effet censure soit plus fort lorsque le risque instantané de T augmente (ce qui est le cas quand $k > 1$) et que l'effet sélection soit plus fort lorsque le risque instantané diminue ($k < 1$).

5. Cette variable n'est définie pour ceux qui sont dans la situation d'intérêt au moment de l'enquête, c'est-à-dire lorsque $D \leq 2010 < D + T$.

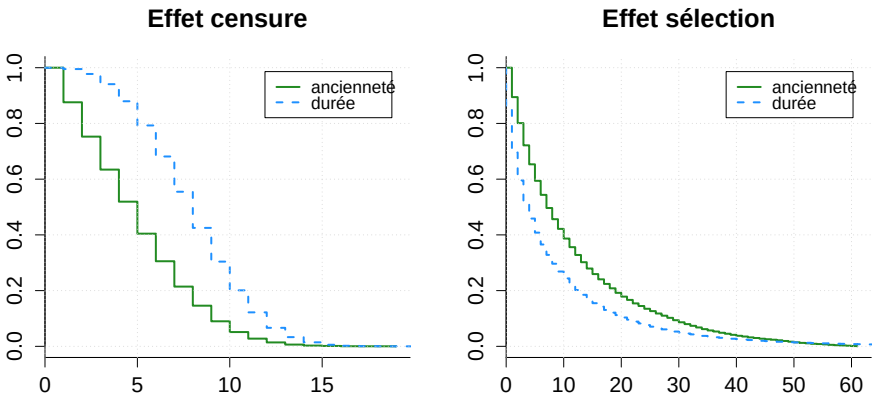

FIGURE 4: Comparaison entre la fonction de survie de l'ancienneté et celle de la durée, à l'aide de simulations. À gauche où $T \sim \text{Weibull}(k = 3, \lambda = 8.96)$ c'est l'effet censure qui est le plus fort tandis qu'à droite où $T \sim \text{Weibull}(k = 0.7, \lambda = 6.32)$ c'est l'effet sélection qui est le plus fort (dans les deux cas $D \sim$ Uniforme).

2.1.2 Conséquences sur la comparaison des anciennetés

Si l'effet sélection est plus fort, les anciennetés observées seront donc en moyenne plus grandes que les durées, tandis que si l'effet censure est plus fort, les anciennetés seront en moyenne plus petite que les durées. Si donc on compare deux groupes, et qu'on trouve qu'en moyenne les ancienntés du premier groupe sont plus faible que celles du deuxième, cela n'implique pas que les durées sous-jacentes du premier groupe sont en moyenne plus petite que les durées du deuxième groupe. Cela est en effet possible si dans le premier groupe, l'effet censure est très fort et dans le deuxième l'effet sélection est très fort, comme l'illustre la figure 5.

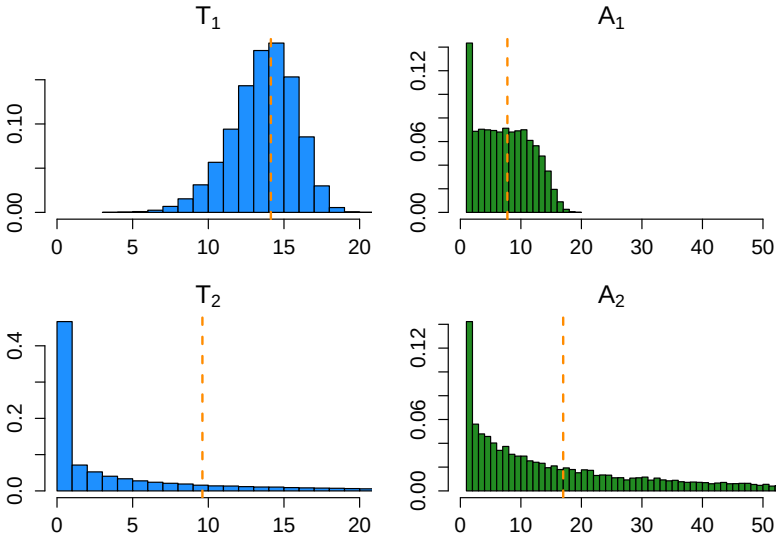

FIGURE 5: Densités de la variable de durée (à gauche) et de la variable d'ancienneté (à droite) pour deux groupes différents. Cela illustre le fait qu'on peut avoir $\mathbb{E}[T_1] > \mathbb{E}[T_2]$ et pourtant $\mathbb{E}[A_1] < \mathbb{E}[A_2]$, où $T_1 \sim \text{Weibull}(k = 8, \lambda = 15)$, $T_2 \sim \text{Weibull}(k = 0.5, \lambda = 5)$ et A_1 et A_2 sont les anciennetés observées (dans les deux cas, $D \sim$ Uniforme). En pointillés oranges sont indiquées les espérances de chaque variable aléatoire.

2.1.3 Le modèle permet de rectifier ces biais

Afin de tester la robustesse du modèle, plusieurs situations ont été simulées puis estimées à l'aide du modèle décrit pour voir si le modèle donne des résultats correctes ou au contraire biaisés.
La première situation est lorsque l'effet censure est le plus fort, ce qui peut être obtenu en simulant la durée T selon une loi de Weibull de paramètre de forme supérieur à 1. Les résultats des estimations sont donnés sur la figure 6

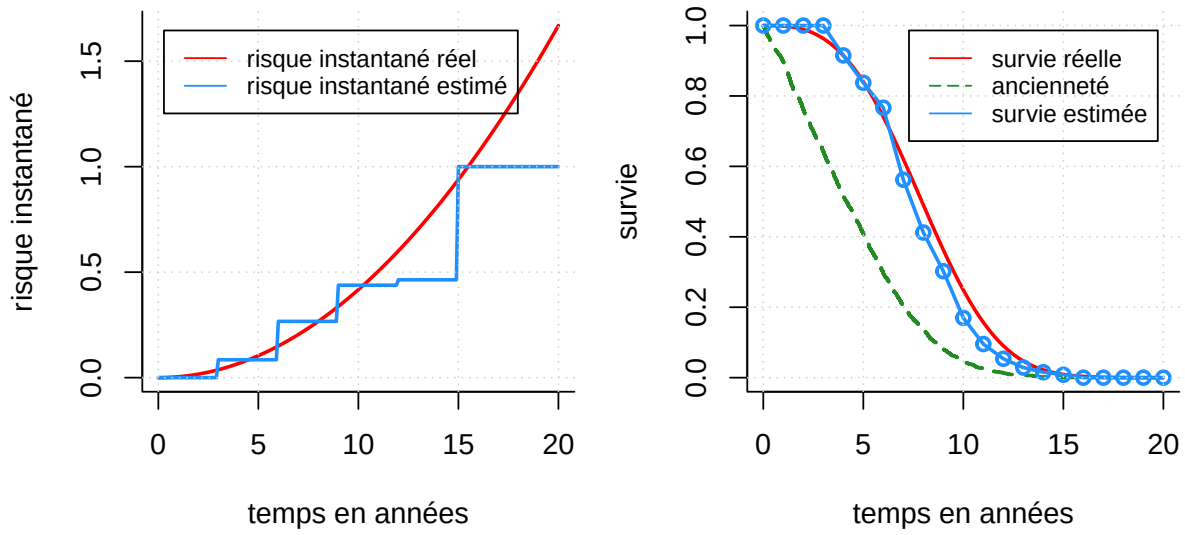

FIGURE 6: Comparaison entre le risque instantané simulé et celui estimé (à gauche) et la survie simulée et celle estimée (à droite). Simulations : $T \sim \text{Weibull}(k = 3, \lambda = 8.3)$.

La figure 7 montre les résultats des estimations lorsque l'effet sélection est plus important.

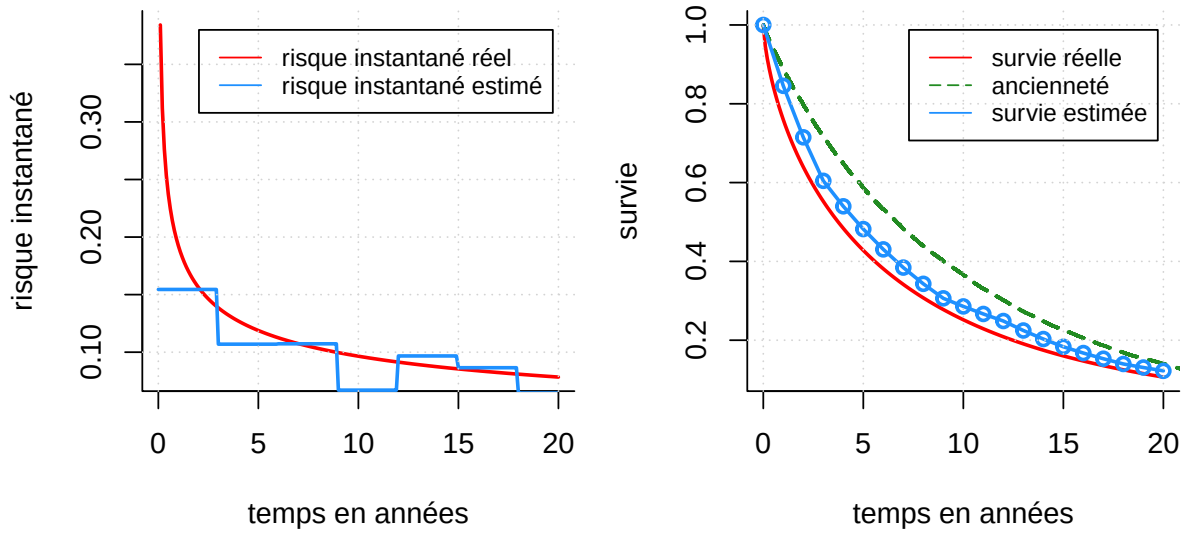

FIGURE 7: Comparaison entre le risque instantané simulé et celui estimé (à gauche) et la survie simulée et celle estimée (à droite). Simulations : $T \sim \text{Weibull}(k = 0.7, \lambda = 6.3)$.

On teste maintenant le modèle lorqu'on rajoute une variable explicative catégorielle pouvant prendre trois valeurs distinctes, 1, 2 ou 3. Le risque instantané s'écrit alors : $h(t,x) = h_0(t)exp(\beta_2 \mathbb{1}_{\{x=2\}} + \beta_3 \mathbb{1}_{\{x=3\}})$. On estime bien les bons paramètres β_2 et β_3 comme le montre la figure 8 qui montre que les fonctions de survie estimées pour les trois groupes correspondent parfaitement à leur surive réelle.

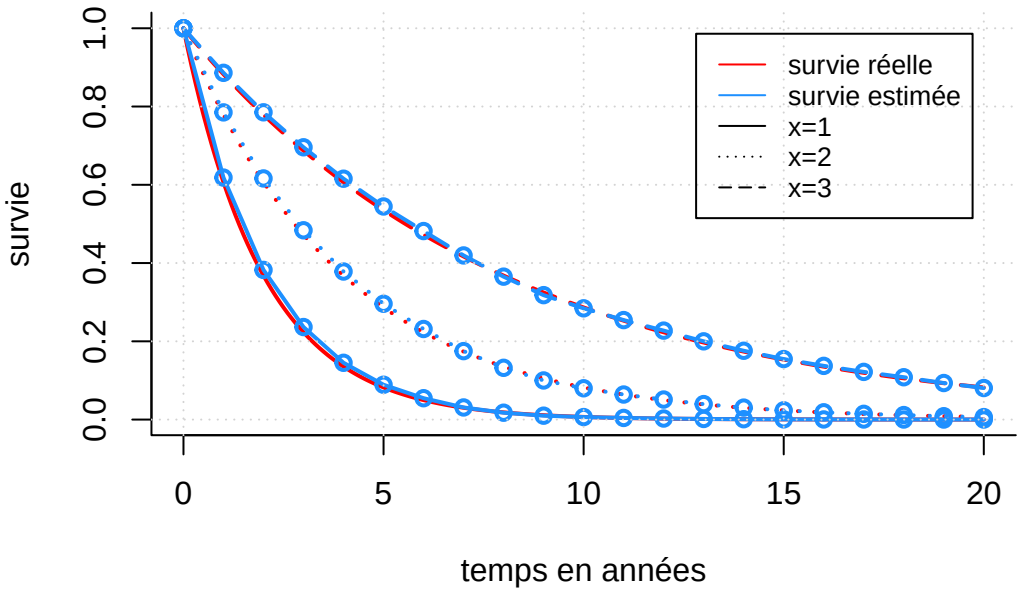

FIGURE 8: Comparaison entre les survies réelles et estimées lorque le risque instantané dépend d'une covariable. Simulations : $\beta_2 = -0.7$ et $\beta_3 = -1.4$ et $h_0(t) = 0.5$

Sur trois exemples simples, on a ainsi pu constater que le modèle permet de rectifier le biais de sélection, le biais de censure et d'estimer les paramètres de régression correspondant à des risques instantanés proportionnels.

2.2 Influence de la variable de flux

2.2.1 Cas d'un flux constant

Le cas particulier où le flux d'entrée dans la situation d'intérêt est constant mérite d'être discuté. Si on note f_A la densité de l'ancienneté $A = 2011 - D$ (qui n'est ici définit que pour ceux qui sont dans la situation d'intérêt au moment de l'enquête), alors pour $x \in \mathbb{N}$, $f_A(x) = \mathbb{P}(D = 2011 - x | D \leq 2010 < D + T) = \frac{\mathbf{1}_{x\geq 1} S_T(x)}{\sum\limits_{u\geq 1} S_T(u)}$. Or, $\sum\limits_{u\geq 1} S_T(u)$ correspond précisément à l'espérance de la variable T[6]. D'où :

$$f_A(x) = \frac{\mathbf{1}_{x\geq 1} S_T(x)}{\mathbb{E}[T]} \tag{5}$$

On remarque donc que la densité de l'ancienneté est proportionnelle à la survie de la durée T. Dans le cas où D et T sont continues et non plus discrètes, on a toujours

6. En remarquant que $f(t) = S_T(t) - S_T(t+1)$ et en remplaçant dant $\mathbb{E}[T] = \sum\limits_{t\geq 0} t f(t)$, on obtient ce résultat.

$f_A(x) = \frac{S_T(x)}{\mathbb{E}[T]}$ pour $x \geq 0$. Si T suit alors une loi exponentielle de paramètre λ, l'ancienneté suit exactement la même loi exponentielle. En effet, dans ce cas, $f_A(x) = S_T(x)/\lambda = f_T(x)$. Ce résultat est bien connue dès qu'on s'intéresse aux processus de renouvellement, comme le montre Lancaster [4]. De plus, Lancaster montre que si T est une variable aléatoire de durée quelconque, on a $\mathbb{E}[A] = \frac{1}{2}(\mathbb{E}[T] + \mathbb{V}[T]/\mathbb{E}[T])$. Donc, l'effet de censure est plus grand lorsque $\mathbb{E}[T] > \mathbb{V}[T]$ tandis que l'effet sélection l'emporte lorsque $\mathbb{E}[T] < \mathbb{V}[T]$ .C'est ce qu'on avait déjà remarqué sur les simulations précédentes, où l'effet censure est d'autant plus fort que la variable T est piquée autour de son espérance, c'est-à-dire lorsque la variance de T est petite devant son espérance.

2.2.2 Cas d'un flux croissant ou décroissant

On simule ici une loi exponentielle d'espérance 10 pour la durée et on regarde l'impacte d'un flux d'entrée croissant ou décroissant sur les anciennetés. On simule donc des flux d'entrée dans la situation d'intérêt, entre 1950 et 2011 pour être comme dans le cas de l'EFL, ce qui rend le cadre plus concert. On rappelle que si le flux était constant, on aurait la même répartition des anciennetés et des durées, car on est dans le cas particulier de la loi exponentielle.

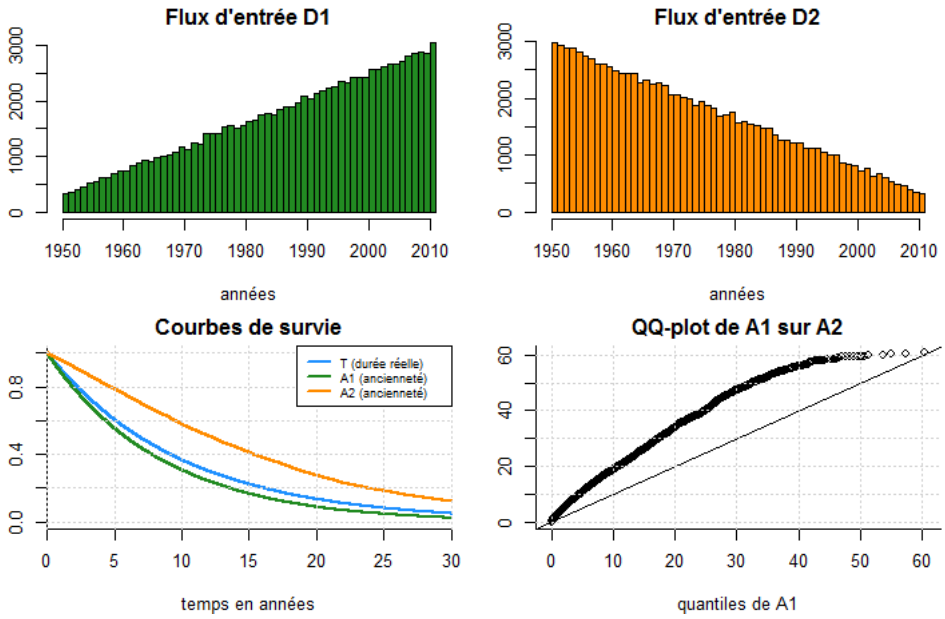

FIGURE 9: Illustration de l'effet d'un flux croissant (D_1, en haut à gauche) et d'un flux décroissant (D_2, en haut à droite) sur les anciennetés observées (A_1 et A_2). La figure en bas à gauche compare les survies de T, A_1 et A_2 et la figure d'en bas à droite compare les quantiles de A_1 aux quantiles de A_2. $T \sim \mathcal{E}xp(10)$

On voit sur la figure 9 qu'un flux croissant (ici le flux D_1 est multiplié par 10 en 60 ans) entraîne des anciennetés plus petites que les durées et favorise donc l'effet censure, tandis qu'un flux décroissant implique des anciennetés plus grandes que les durées (effet sélection plus fort). Alors que les groupes 1 et 2 ont exactement les mêmes durées, leurs anciennetés n'ont pas la même répartition (voir le qq-plot de la figure 9), car les flux d'entrée dans la situation sont très différents.

3 Résultats à partir de l'Enquête Famille et Logement

Parmi les 359770 répondants à l'EFL, 12519 sont en situation de monoparentalité au moment de l'enquête [7], dont 1073 hommes et 11446 femmes. L'ancienneté mesurée la plus grande est de 41 ans.

3.1 Sans variables explicatives

On présente ici les résultats des estimations du risque instantané sans prendre en compte de covariables. La figure 10 montre que le risque instantané global n'est pas monotone : initialement il diminue, puis se stabilise et ré-augmente. Cette forme en "U" suggère que soit on sort rapidement de la monoparentalité, soit on y reste longtemps. En effet, la probabilité de sortir de cette situation est la plus faible entre 3 et 8 ans. La courbe de survie obtenue à partir du risque instantané montre qu'au bout de 3 ans la moitié des personnes en famille monoparentale sont sorties de cette situation, qu'au bout de 8 ans il en reste encore 28%, au bout de 12 ans plus que 14%, et que seulement 3% restent plus de 18 ans.

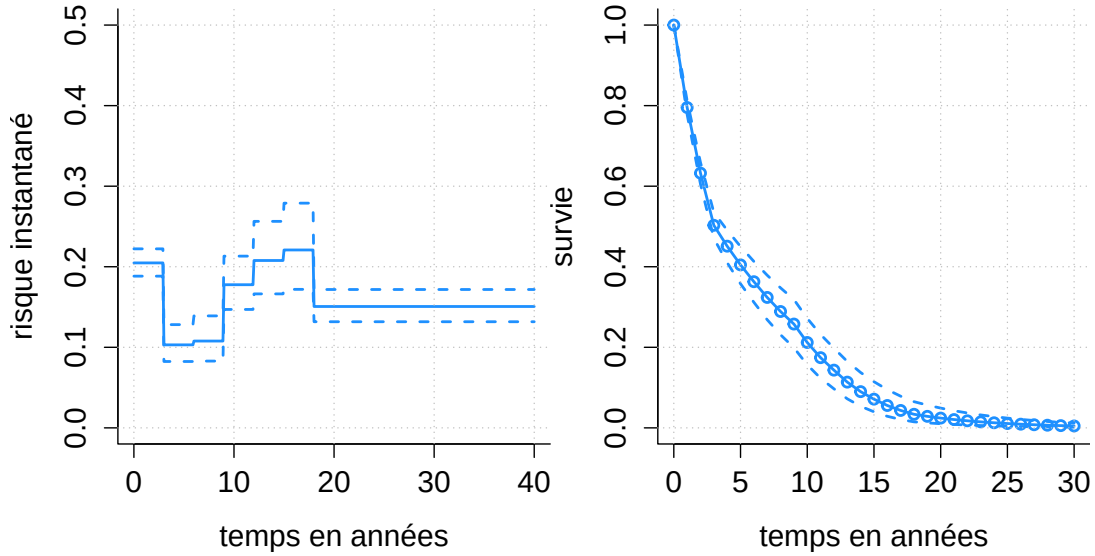

FIGURE 10: Estimations du risque instantané (à gauche) et de la survie (à droite) du temps passé en famille monoparentale.
Champ : France métropolitaine. Source : Insee, EFL, 2011.

La figure 11 compare les quantiles estimés de la variable de durée T aux quantiles de la variable d'ancienneté. L'ancienneté donne la même réponse pour le quantile d'ordre 0.6 que l'estimation de la durée, mais pour les quantiles inférieurs, l'ancienneté fournit des quantiles plus petit que ceux estimés pour la durée réelle, et pour les quantiles supérieurs à 0.6, l'ancienneté donne des quantiles plus grands que les estimations. Ceci suggère donc que l'ancienneté à tendance à surestimer la survie pour les temps courts et à la sous-estimer pour les temps longs.

7. On a retiré les 205 personnes étant entrée en famille monoparentale l'année de l'enquête, en 2011, car elles n'apportent pas d'information à notre modèle.

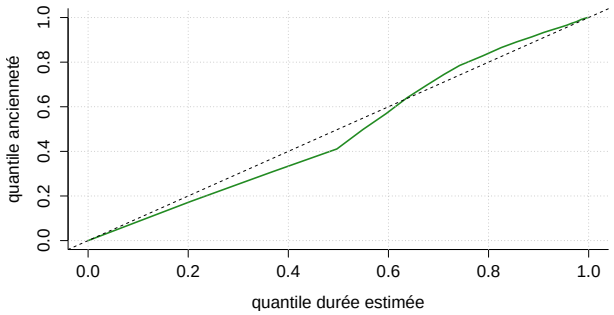

FIGURE 11: Comparaison entre les quantiles de l'ancienneté et les quantiles de la durée estimée.
Champ : France métropolitaine. Source : Insee, EFL, 2011.

En estimant séparement le risque instantané pour les femmes et les hommes, on constate que ces risques instantanés n'ont pas la même forme (voire figure 12) et que par conséquent l'hypothèse de proportionnalité des risques instantanés n'est pas justifiée pour la comparaison par sexe. Le risque instantané des femmes à également une forme en "U", tandis que celui des hommes est plus fluctuant et plus élevé en moyenne. Cette fluctuation laisse penser qu'il n'y a pas assez de données chez les hommes pour faire converger les estimations par maximum de vraisemblance. Par la suite, on séparera l'analyse entre les hommes et les femmes en s'attachant plus à ces dernières pour lesquelles les résultats seront plus robustes.

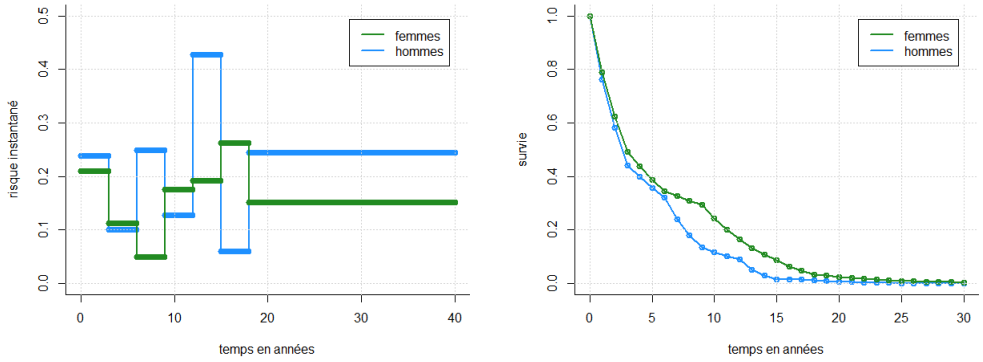

FIGURE 12: Estimations du risque instantané (à gauche) et de la survie (à droite) du temps passé en famille monoparentale selon le sexe.
Champ : France métropolitaine. Source : Insee, EFL, 2011.

3.2 Avec variables explicatives

Dans cette partie, on décrit les résultats obtenus en introduisant des covariables indépendantes du temps, à savoir le niveau de diplôme atteint au moment de l'enquête, la catégorie sociale au moment de l'enquête et la cause d'entrée en famille monoparentale. Le niveau de diplôme atteint permet de distinguer quatre sous-populations : les personnes n'ayant aucun diplôme (19.8%), les personnes ayant un diplôme de niveau inférieur au baccalauréat (34.5%), les personnes ayant un diplôme de niveau équivalent au baccalauréat

(18.6%) et les personnes ayant un diplôme de niveau strictement supérieur au baccalauréat (27.1%) [8]. On distingue ensuite huit catégories sociales [9] : les agriculteurs (0.5%), les artisans, commerçants ou chefs d'entreprise (3.4%), les cadres (8.6%), les professions intermédiaires (20.3%), les employés (41.3%), les ouvriers (12.5%), les chômeurs n'ayant jamais travaillé et inactifs de moins de 60 ans (11.3%) et les autres (2.1%). Enfin on considère trois principales causes d'entrée en famille monoparentale : l'entrée pour cause de séparation d'avec le conjoint (79.1%), l'entrée pour cause de décès du conjoint (6.2%) et l'entrée pour avoir eu un enfant hors couple (14.7%).

Des différences de durées selon le diplôme es résultats de l'estimation des paramètres de régression β sont donnés dans le tableau 1. Si β est positif, cela signifique que le risque instantané est plus élevé que celui du groupe de référence, donc que la survie est plus courte que celle du groupe de référence. On remarque que les p-valeurs associées aux estimations des paramètres pour les hommes indiquent que ces paramètres ne sont pas significativement différents de 0 (à 5%) ce qui n'est pas le cas pour les femmes. Pour les hommes comme pour les femmes, ceux qui ont un diplôme équivalent ou inférieur au baccalauréat restent moins longtemps en situation de monoparentalité que les personnes n'ayant aucun diplôme. Par contre, d'après les estimations de ce modèle, il n'y a pas de différence significative entre ceux qui n'ont aucun diplôme et les plus diplômés ("bac +"), comme si le fait d'être très diplômé était un handicap pour sortir de la monoparentalité.

	Hommes		Femmes	
Variable	Coefficient	Valeur-p	Coefficient	Valeur-p
β_{bac-}	-0.023	0.8	0.12	3.7×10^{-5}
β_{bac}	0.058	0.6	0.25	1.1×10^{-12}
β_{bac+}	-0.007	0.94	-5.4×10^{-4}	0.99

TABLE 1: Estimations des coefficients de régression β lorsque la variable explicative est le diplôme. La catégorie de référence est "aucun diplôme".
Champ : France métropolitaine. Source : Insee, EFL, 2011.

Le cadre gauche de la figure 13 montre les risques instantanés de base h_0 estimés à l'aide d'une fonction constante par morceaux. Ces risques ont la même forme que ceux de la figure 12 ce qui laisse penser que l'hypothèse de proportionnalité des risques instantanés n'est pas absurde. Le cadre de droite indique que même les femmes les plus diplômées restent plus longtemps en famille monoparentale que les hommes les moins diplômés. On voit aussi l'écart important qui existe en terme de durée passée en famille monoparentale entre les femmes qui n'ont aucun diplôme et les hommes ayant un diplôme de niveau supérieur au baccalauréat. D'après ces estimations, les femmes ayant un diplôme de niveau équivalent au baccalauréat restent en moyenne 4.6 ans en situation de monoparentalité, contre 6.2 ans pour celle qui n'ont aucun diplôme ou un diplôme supérieur au baccalauréat et 5.4 pour celles qui ont un diplôme de niveau inférieur au baccalauréat.

8. Par souci de concision, on nommera par la suite ces quatre catégories "aucun diplôme", "bac -", "bac" et "bac +".

9. Les retraités (0.9% des familles monoparentales) sont reclassés dans leur ancienne catégorie sociale.

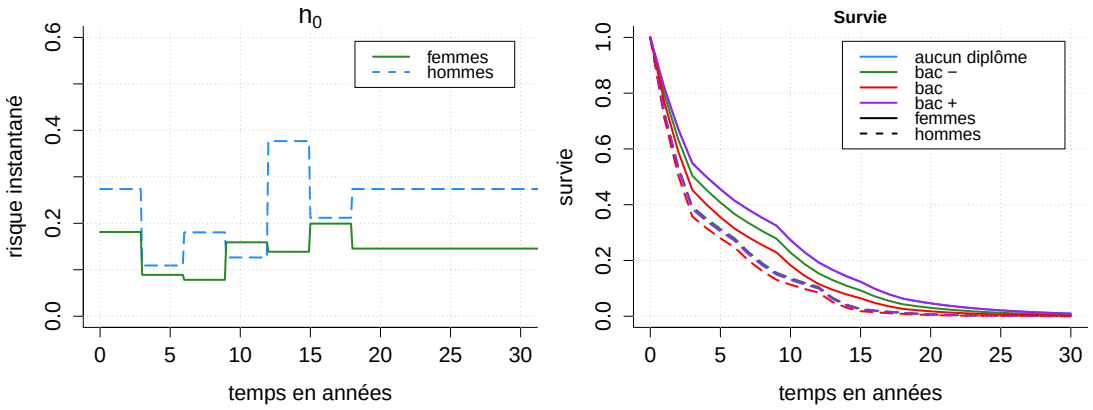

FIGURE 13: Estimations du risque instantané de base (à gauche) et de la survie (à droite) du temps passé en famille monoparentale en fonction du diplôme.
Champ : France métropolitaine. Source : Insee, EFL, 2011.

Au contraire, la catégorie sociale semble ne pas avoir d'effet sur la durée de monoparentalité Comme le montre le tableau 2, aucun coefficient n'est significatif à 5% mis à part celui correspondant aux femmes de profession intermédiaire dont la durée de monoparentalité est un peu plus longue que les ouvrières. Pour ces estimations, nous avons supprimés les personnes n'entrant pas dans une des six premières catégories sociales [10].

	Hommes		Femmes	
Variable	Coefficient	Valeur-p	Coefficient	Valeur-p
$\beta_{agriculteur}$	-0.056	0.81	-0.23	0.38
$\beta_{a.c.c.}$	-0.047	0.71	-0.0013	0.98
β_{cadre}	-0.15	0.15	-0.11	0.066
$\beta_{prof.int}$	-0.086	0.34	-0.15	0.0016
$\beta_{employé}$	0.042	0.68	-0.0089	0.82

TABLE 2: Estimations des coefficients de régression β lorsque la variable explicative est la catégorie sociale. La catégorie de référence est "ouvrier".
Champ : France métropolitaine, personnes actives ou retraitées. Source : Insee, EFL, 2011.

Cet absence d'effet de la catégorie sociale peut venir du fait qu'à la fois la catégorie sociale n'est pas figée au cours du temps, mais peut évoluer, et que ceux qui sont inactifs peuvent provenir de catégories sociales différentes.

Les causes d'entrées en monoparentalité : principales sources des différences de durées. Les différences de durées les plus importantes sont observées lorsqu'on classe les personnes selon leur cause d'entrée en famille monoparentale. Chez les hommes comme chez les femmes, ceux qui ont vécu une situation de monoparentalité à l'issu d'une séparation sont ceux qui y passent le moins de temps. Au contraire les personnes entrées pour avoir eu un enfant en dehors d'un couple cohabitant restent le plus longtemps en famille

10. "agriculteur","artisan-commerçant-chef d'entreprise", "cadre", "profession intermédiare", "employé" et "ouvrier"

monoparentale : pour une femme, le risque instantané de sortir de la monoparentalité est 1.9 plus élevé lorsqu'elle est séparée que lorsqu'elle a eu un enfant hors couple.

	Hommes		Femmes	
Variable	Coefficient	Valeur-p	Coefficient	Valeur-p
$\beta_{séparation}$	1.2	1.9×10^{-7}	0.64	5.6×10^{-85}
$\beta_{décès}$	1.2	1.2×10^{-6}	0.54	2.8×10^{-30}

TABLE 3: Estimations des coefficients de régression β lorsque la variable explicative est la cause d'entrée en fmaille monoparentale. La catégorie de référence est "enfant".
Champ : France métropolitaine. Source : Insee, EFL, 2011.

Les veuves et veufs passent également moins de temps en famille monoparentale (5 ans pour les femmes et 3.1 ans pour les hommes en moyenne) que ceux qui ont eu un enfant hors couple (8.9 ans pour les femmes et 10.6 ans pour les hommes), mais plus de temps que les séparés (4.4 ans pour les femmes et 3.1 ans pour les hommes). Chez les hommes, il n'y a pas de différence de durée entre les veufs et les séparés.

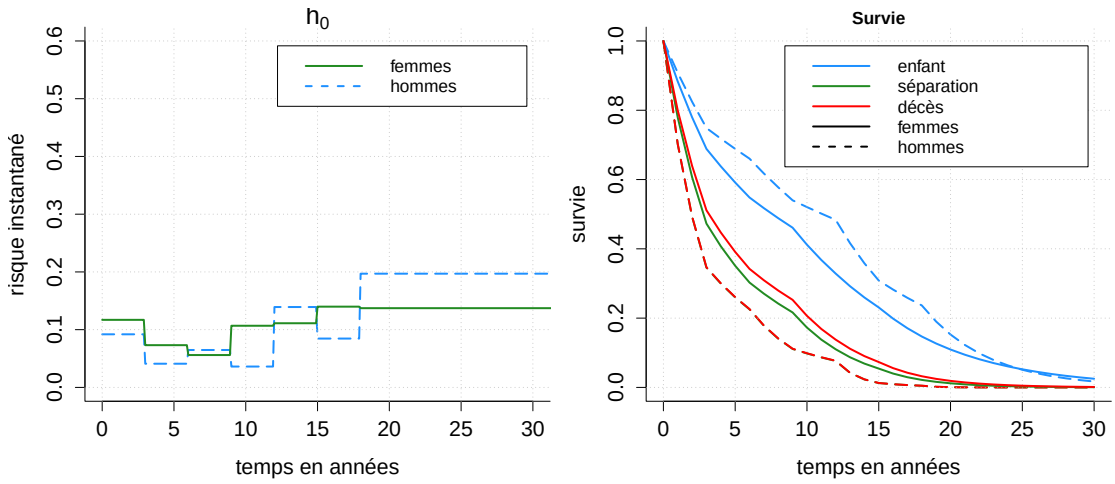

FIGURE 14: Estimations du risque instantané de base (à gauche) et de la survie (à droite) du temps passé en famille monoparentale en fonction de la cause d'entrée.
Champ : France métropolitaine. Source : Insee, EFL, 2011.

3.3 Comparaison avec ERFI

Avec les données d'ERFI on remarque qu'il y a une corrélation négative entre le temps passé en famille monoparentale et l'année d'entrée dans celle-ci. Les périodes de monoparentalité sont donc de plus en plus courtes. Pour pouvoir mener une comparaison plus juste entre ERFI et l'EFL on choisit donc parmi les enquêtés d'ERFI ceux qui sont rentrés pour la première fois en monoparentalité après 1985 [11].
À l'aide d'ERFI, on retrouve la même forme en "U" de risque instantané de quitter la monoparentalité (figure 15). Les deux risques instantanés sont superposés ce qui indique

11. 99.5% des enquêtés de l'EFL en famille monoparentale au moment de l'enquête sont entrés dans cette situation après 1985.

une concordance des résultats, même si le risque instantané estimé à partir de l'EFL augmente plus tôt que celui estimé par ERFI.

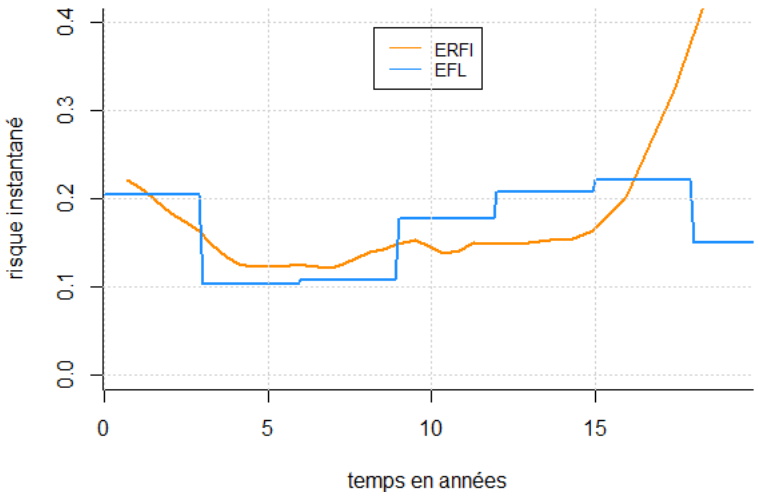

FIGURE 15: Comparaison du risque instantané de sortir de la monoparentalité estimé à partir d'ERFI et à partir de l'EFL.
Champ : France métropolitaine. Sources : Ined-Insee, ERFI, 2005, 2011 et Insee, EFL, 2011.

Pourtant, en examinant les résultats plus en détail, on s'aperçoit qu'il y a des différences lors de l'estimation par sexe : l'EFL estime des durées plus courtes pour les femmes qu'ERFI, tandis que c'est le contraire pour les hommes (figure 16). Il apparaît que ce sont surtout pour les courtes durées (≤ 7 ans) que les divergences sont le plus marquées.

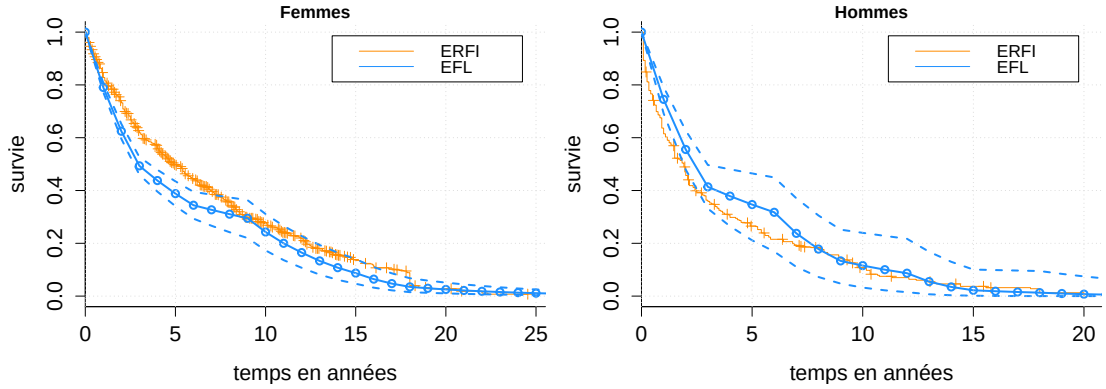

FIGURE 16: Comparaison de la survie du temps passé en famille monoparentale estimée à partir d'ERFI et à partir de l'EFL, en fonction du sexe. En pointillées sont indiqués les intervalles de confiance à 95%.
Champ : France métropolitaine. Sources : Ined-Insee, ERFI, 2005, 2011 et Insee, EFL, 2011.

D'après les résultats obtenus à partir d'ERFI, il n'y a pas d'effet du diplôme sur la durée de la monoparentalité. Pour les autres variables explicatives, les résultats sont semblables à ceux obtenus avec l'EFL, à savoir que la catégorie sociale n'est pas un facteur explicatif du temps passé en famille monoparentale, et qu'il y a un fort écart de durée entre les séparés

et veufs d'une part et ceux qui sont entrés en monoparentalité pour avoir eu un enfant hors couple d'autre part. Néanmoins, cet écart est moins marqué dans les résultats d'ERFI, qui de plus suggère que les risques instantanés ne sont pas proportionnels lorsqu'on prend la cause d'entrée comme covariable.

4 Limites et perspectives

Quatre hypothèses ont été formulées dans le cadre de la modélisation des données afin de simplifier le problème et de permettre de calculer les paramètres. La première hypothèse postule que chaque individu ne vit au plus qu'une seule période de monoparentalité. Cette hypothèse s'avère être fausse lorsqu'on regarde les résultats issus d'ERFI car la proportion de monoparents ayant vécu deux situation de monoparentalité ou plus augmente, passant de 9% pour ceux qui sont nés entre 1933 et 1942 à 22% pour ceux qui sont nés entre 1953 et 1962. Cela n'a pas de conséquence sur les estimations du modèle si chaque période de monoparentalité (que ce soit première, deuxième ou troisième) suit la même loi de durée.

La deuxième hypothèse porte sur l'indépendance entre la durée de monoparentalité T et la date d'entrée dans cette situation D. ERFI montre qu'il y a une corrélation négative entre ces deux variables aléatoires, ce qui correspond au fait qu'on reste en moyenne moins longtemps en famille monoparentale que par le passé. Les simulations montrent qu'une corrélation négative entraîne une sous-estimation de la survie. Cela n'est pas tellement gênant, puisque dans le cas d'une corrélation négative, la survie devient de plus en plus faible, et une sous-estimation de celle-ci est donc en quelque sorte une anticipation de ce qu'elle sera dans un futur proche. Les simulations ont également révélé qu'il semble très difficile d'estimer cette corrélation à l'aide des seules observations des anciennetés.

La troisième hypothèse suppose que les risques intantanés sont proportionnels entre eux. On a vu que cette hypothèse ne tient pas si on veut comparer les hommes aux femmes, ce qui est confirmé par les résultats d'ERFI. ERFI montre également que les risques instantanés ne sont pas proportionnels lorsqu'on s'intéresse à la cause d'entrée comme variable explicative. Ceci a pour conséquence qu'il ne faut pas considérer les estimations comme donnant des résultats indiscutables, comme par exemple que 80% des séparés sortent de la monoparentalité avant 10 ans contre 60% pour ceux qui ont eu un enfant hors couple. L'interprétation doit être que les séparés restent moins longtemps que ceux qui ont eu un enfant hors couple.

La quatrième et dernière hypothèse admettant la non variabilité dans le temps des variables explicatives est discutable pour certaines de ces covariables retenues. Par exemple, rien n'affirme que le diplôme ou la catégorie sociale après 18 ans n'évolue pas. En revanche la cause d'entrée en famille monoparentale est déterminée de manière non ambigue pour tous. Néanmoins, la mesure de ces covariable se fait au moment où l'individu est en situation de monoparentalité, et on peut raisonablement penser que ces covariables ne changent pas, ou peu, durant cette situation. C'est bien la valeur de ces variables pendant la monoparentalité qui nous intéresse et non les valeurs qu'elles peuvent prendre avant ou après. On peut donc penser que même si cette hypothèse n'est sans doute pas tout à fait respecter, elle n'engendre pas de biais significatif sur les résultats.

Une autre source de difficulté est le fait d'estimer les flux d'entrées en famille monopa-

rentale selon des caractéristiques individuelles données. Cela nécessite une source annexe et relativise donc l'intérêt de la méthode développée ici à être généralisée à d'autres enquêtes. Toutefois, il n'est pas nécessaire de connaître précisément les flux, mais seulement de déterminer leur tendance : est-ce qu'ils augmentent, diminuent ou restent constants ? Une telle estimation a ici été possible grâce à ERFI. Dans le cas où aucune information n'est disponible sur les flux, il est alors possible d'envisager plusieurs scénarios d'évolution de flux, et d'estimer les durées selon ces différents scénarios.

Bibliographie

[1] Nickell S., "Estimating the probability of leaving unemployment", Econometrica, vol 47, n°5, pp 1249-1266, septembre 1979.
[2] Florens J.-P., Fougère D., Mouchart M., "Duration models and point processes", Insee, Document de travail n°2007-37.
[3] Cox D. ,"Regression models and life-tables", Journal of the Royal Statistical Society, Series B, vol 34, n°2, pp 187 :220, 1972.
[4] Lancaster T., "The econometric analysis of transition data", Cambridge university press edition, 1990.
[5] Buisson G., Costemalle V., Daguet F., "Depuis combien de temps est-on parent de famille monoparentale ?", Insee, Insee Première n°1539, 2015.
[6] David O., Eydoux L., Martin C., Millar J., Séchet R., "Les familles monoparentales en Europe", Caisse nationale d'allocations familiales, Dossiers d'études 54, 2004.
[7] Chardon O., Daguet F., Vivas E., "Les familles monoparentales - des difficultés à travailler et à se loger", Insee, Insee Première 1195, juin 2008.