LES MODÈLES DE DURÉE : APPLICATION A UNE COHORTE D'ENTREPRISES

Amel Gharbi

1. Introduction

L'étude des modèles de durées est un domaine récent, à l'origine appliqué à la démographie humaine ainsi qu'à la statistique médicale. Au début des années 80 cette approche a été élargie à l'analyse des durées du chômage et des durées de vie des entreprises.

Les études antérieures fondées sur les calculs de fréquences ont été élaborées dans un premier temps par M. J.-M. Callies, H. Viennet, D. Francoz et J. Bonneau. Ces études ont soulevé différents problèmes liés aux données incomplètes ou censurées. Il convient d'en citer deux:

1/ parmi le nombre total des cessations, 25 % sont connues lors des enquêtes d'amélioration du répertoire: ce sont les cessations d'origine direction régionale (DR), parmi lesquelles 50 % seulement ont une date de cessation renseignée. Alors que le reste est renseigné par une date fictive,

2/ à la fin de l'observation, un certain nombre d'entreprises sont considérées comme actives, parmi lesquelles certaines ont cessé leur activité avant la fin de l'observation alors qu'elles ne seront enregistrées que plus tard : problème des faux actifs qui est lié directement au problème du délai d'enregistrement.

Ces données conduisent à des valeurs non disponibles ou enregistrées avec erreur, et entraînant des biais dans les différentes statistiques qui ont été trouvées. L'une des méthodes statistiques permettant de contourner ces problèmes est l'application des modèles de durée. L'intérêt de ces méthodes est double : D'une part elles permettent de tenir compte de l'incomplétude de l'information et d'autre part elles rendent possible la modélisation de la durée de vie des entreprises. Ceci va nous permettre d'estimer les taux de survie des entreprises en réduisant l'ampleur du biais, d'en prévoir et éventuellement d'en contrôler l'évolution future.

Pour caractériser la variable aléatoire, durée de vie des entreprises, les indicateurs généralement utilisés sont le taux de survie; $S(t)=P(T>t)$ et le taux de cessation; $F(t)=P(T\leq t)$. Dans le cas des modèles de durée en plus du taux de survie et du taux

de cessation d'autres indicateurs peuvent caractériser la variable durée de vie, telle que le taux de cessation instantané conditionnellement au fait d'avoir survécu jusqu'à l'instant immédiatement antérieur; h(t), le cumul du taux instantané de cessation; H(t), la durée de vie moyenne restante, ... Ces indicateurs sont présentés en *Annexe1*.

Pour tenir compte de ces différents problèmes, j'ai mené une étude qui se décompose actuellement en trois étapes :

1/ Utilisation de la procédure SAS d'analyse des durées de vie (proc LIFETEST) afin d'obtenir :
- des strates homogènes vis-a-vis des variables retenues; origine de la création, tranche d'effectif salarié à la date de création, activité principale exercée et forme juridique (*Annexe2*). Ces variables semblent être les plus pertinentes dans le fichier pour expliquer la durée de vie des entreprises.
- des premières estimations des différents indicateurs; f(t), F(t), S(t), h(t)...

2/ Construction d'un modèle paramétrique compatible qualitativement avec les résultats obtenus dans la première étape ;

3/ Estimation de ce modèle, par la méthode du maximum de vraisemblance en utilisant la procédure SAS d'analyse des durées de vie (proc NLIN) ;

Seront représentés ci-dessous les résultats de l'étude disponibles à ce jour ainsi que la méthode adoptée pour la modélisation de la durée de vie des entreprises.

2. L'analyse descriptive (une approche non paramétrique)

Cette analyse consiste à mener une approche descriptive sans aucune composante économique structurelle expliquant le phénomène de la durée de vie des entreprises. C'est une analyse qui ne tient compte que de certaines variables qualitatives existant dans le fichier, tout en supposant que "toute chose est égale par ailleurs".

2.1. Description des données disponibles

L'étude de la durée de vie des entreprises est fondée sur l'exploitation du répertoire SIRENE (Système Informatique pour le Répertoire des ENtreprises et des Etablissement). Le fichier DEMO (comme démographie) est élaboré mensuellement à partir des mouvements enregistrés dans le répertoire SIRENE. Chaque mouvement (création, cessation...) se caractérise par deux dates :
- "date d'événement": date à laquelle est survenu l'événement ;
- "date de traitement": date d'enregistrement de l'événement dans le répertoire.

La "date de traitement" est toujours postérieure à la "date d'événement".

Les créations d'entreprises ainsi que leurs disparitions obéissent simultanément à un processus ponctuel dit de " naissance et de mort ", où la création peut s'interpréter comme "la naissance" et la cessation comme "la mort". La création ainsi que la cessation sont de plusieurs types :

- une création peut être soit une création *exnihilo* soit une création par reprise des moyens de production par une ou plusieurs unités (par fusion ou par absorption...) ;
- une cessation peut être soit une cessation totale d'activité soit une cessation par reprise des moyens de production (liquidation, fusion ou absorption, ...).

Dans le cas de la création de l'entreprise, d'une part il y a obligation de déclaration auprès d'un Centre de Formalité des Entreprises (CFE) dans les quinze jours qui suivent le début de l'activité, et cette obligation est bien respectée. D'autre part l'INSEE est tenu de fournir un numéro d'immatriculation à l'entreprise dans les deux jours après réception de son dossier. On peut conclure que l'enregistrement des créations est effectué dans un délai inférieur à 1 mois après le début de l'activité.

Dans le cas de la cessation d'activité d'une entreprise, il en est tout autrement. En effet même s'il y a une obligation réglementaire de déclaration auprès du CFE ; cette obligation n'est pas toujours respectée et lorsqu'elle l'est le respect des délais n'est pas forcément de mise. Par ailleurs l'information relative aux cessations d'activité peut être également acquise par les Directions Régionales (DR) au cours d'enquêtes spécifiques; qui sont les enquêtes d'amélioration du répertoire.

Les cessations d'origine Directions Régionales issues des enquêtes d'amélioration du répertoire représentent le 1/4 des cessations enregistrées. Parmi celles-ci 50 % ont une "date d'événement" renseignée correctement. On ne peut supprimer le faible effectif des dates d'événement inconnues, dans la mesure où cela pourrait créer un biais de sélection endogène dans le fichier.

Le champ de l'étude est la cohorte 87, soit l'ensemble des entreprises créées ou reprises entre le 01/01/1987 et le 31/12/1987. L'intérêt de ce choix est d'éviter le problème de manque d'information concernant les dates de créations, ceci fera l'objet d'une autre étude. Le nombre des entreprises existant dans cette cohorte est de 236 611. Le champ étudié porte sur l'industrie, le commerce et les services (champ ICS), il exclut les secteurs de l'agriculture et des activités financières. Les entreprises qui ont une durée de vie inférieure à un mois ne représentent que 2,5%. Elles ne seront pas prises en compte, car on considère qu'il s'agit soit d'une erreur d'enregistrement, soit que ces entreprises n'ont pas pu réaliser une activité économique réelle.

2.2. La stratification

Etant donné l'importance de la taille du fichier, ainsi que l'existence des variables explicatives qui semblent être pertinentes pour l'étude de la durée de vie des entreprises, ces dernières seront nos variables de stratifications. Des strates ont été déterminées afin de trouver des sous populations homogènes dont le comportement des entreprises était similaire en terme de durée de vie. La méthode appliquée est celle de la procédure LIFETEST : procédure SAS utilisant principalement les modèles de durée. Cette méthode a permis de stratifier le fichier tout en se basant sur le test de rang afin d'étudier l'homogénéité des strates. Le nombre de strates obtenus est de 41.

Les estimations obtenues par la procédure "LIFETEST" sont les estimateurs dit de **Kaplan-Meier**, dont le principe est présenté en *annexe3*.

2.3. l'analyse non paramétrique

Afin de réaliser cette étude j'ai choisi comme strate d'application, la strate regroupant l'ensemble des entreprises qui ont une tranche d'effectif compris entre 3 et 19 salariés, d'origine création *exnihilo*, regroupant toutes les activités sauf la construction et de forme juridique artisan-commerçant, (SARL et EURL), SA et société civile et autre (strat39), dont la loi des représentations graphique obtenues par la procédure "LIFETEST" est facile à identifier. Les tableaux qui illustrent d'une part la répartition des entreprises suivant leur état d'activité et d'autre part les estimations non paramétrique cumulés; le taux de survie; $\hat{S}(t)$ et le taux de cessation instantané cumulé sachant que l'entreprise a été active jusqu'à l'instant immédiatement antérieur; $\hat{H}(t)$. Ces taux seront présentés en année et en % (*tableau1*). Les représentations graphiques des différentes estimations appelées estimateurs Kplan-Meier; le taux de survie; $\hat{S}(t)$ et le taux instantané de cessation sachant que l'entreprise a été active jusqu'à l'instant immédiatement antérieur; $\hat{h}(t)$ et le taux de cessation; $\hat{f}(t)$. Ces estimateurs sont calculés en mois (*figure 1*).
Pour une meilleur répartition des différent état d'entreprises, j'ai scindé le tableau de répartition des entreprises selon leurs états d'activités et l'origine de la cessation en deux tableaux (Tableau1.a et Tableau1.b)

Tableau 1.a

Répartition des entreprises de la strate 39 suivant leurs états d'activités (encore active ou en cessation à la fin de l'observation)

Strate 39	Nombre d'entreprise	Proportion d'entreprise (en %)
Cessation	4835	54,46
Active	4043	45,54
Total	**8878**	**100**

On remarque qu'à 8 ans le nombre d'entreprises considéré comme encore actives est de 4 043, soit 45,54 % du total de la strate 39 et celles qui ont cessé leurs activités est de 4 835, soit 54,46 % (Tableau1.a).

Tableau 1.b

Répartition des entreprises de la strate 39 suivant l'origine de la cessation pendant la période d'observation

Strate 39	Nombre de cessation	Proportion des cessations / au nbre total des cessation (en %)	Proportion des cessations / au nbre total des entreprises (en %)
cessation d'origine			
- CFE[1]	3347	69,23	37,70
- DR[2] (date de cessation)	926	19,15	10,43
- DR (date fictive)	562	11,62	06,33
Total des cessation	**4835**	**100**	**54,46**

Parmi les cessations 3 347 sont d'origine CFE, soit 69,23 % du total des cessations et 37,70 % du total de la strate. Celles d'origine DR sont de 1 488 (926 + 562) soit 30,76 % du total des cessations et de 16,76 % du total de la strate. Parmi eux 926 cessations avec une date renseignée, soit 19,15 % du total des cessations et 10,43 % du total de la strate. Enfin celles d'origine DR avec une date fictive de cessation sont de 562, soit 11,62% du total des cessations et 06,33 du total de la strate (tableau1.b).

[1] CFE: Centre de Formalité des Entreprises.

[2] DR: Direction Régionale.

Tableau 2

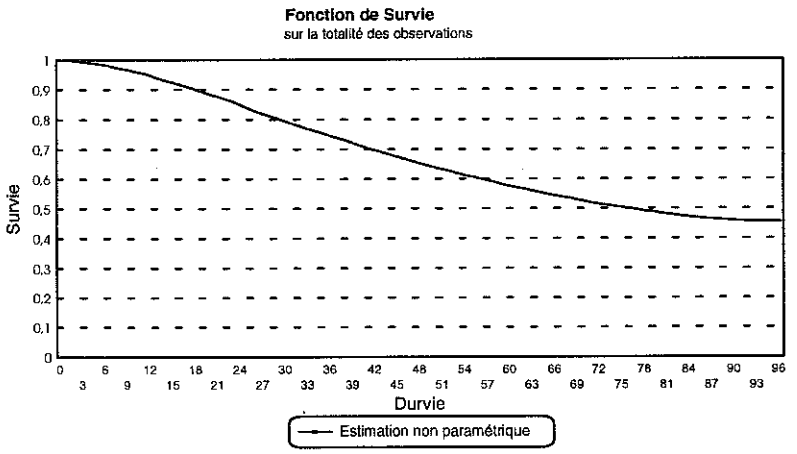
Les estimateurs Kaplan-Meier des taux de survie; $\hat{S}(t)$ ***et des taux de cessation instantané cumulé (hasard cumulé);*** $\hat{H}(t)$

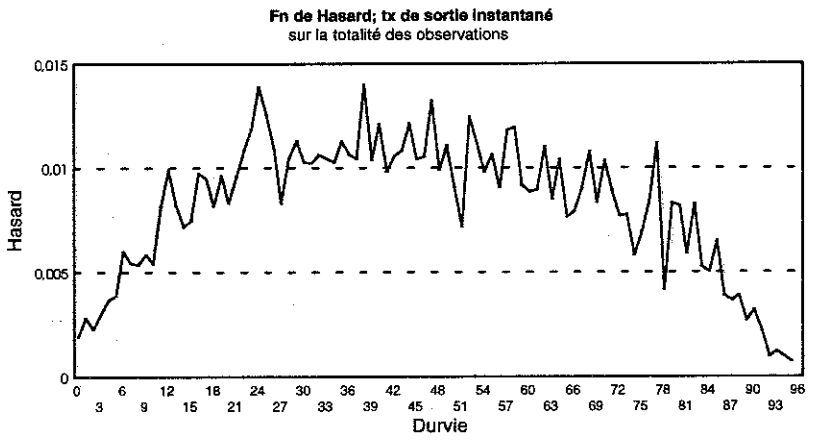
Durée(t)	$\hat{S}(t)$	$\hat{H}(t)$	Durée(t)	$\hat{S}(t)$	$\hat{H}(t)$
1 an	94,77	5,37	5 ans	57,51	55,32
2 ans	84,87	16,40	6 ans	51,48	66,40
3 ans	74,47	29,47	7 ans	47,16	75,16
4 ans	65,06	42,98	8 ans	45,54	78,65

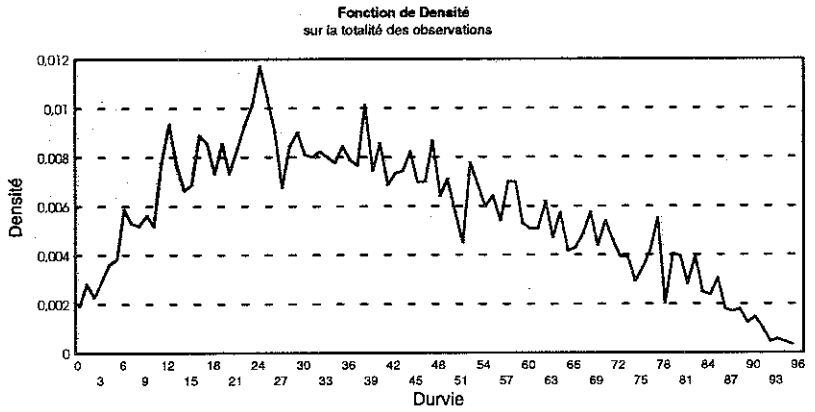
Parmi les indicateurs utilisés dans la méthode des modèles de durée, certains peuvent être des indicateurs annuel; le taux de survie; S(t), le taux de hasard cumulés ; H(t), ... En tenant compte du fait qu'à la fin de l'observation, certaines entreprises sont considérées comme active; ensemble d'entreprises censurées à droite, nous avons obtenus des estimations des taux de survie $\hat{S}(t)$ et des taux de cessation instantané cumulé $\hat{H}(t)$. Exemple à 5 ans : le taux de survie est de 57,51 %, alors que le taux de cessation instantané conditionnellement au fait d'avoir survécu jusqu'à l'instant immédiatement antérieur est de 55,32 %. Ce dernier ne représente pas une proportion des cessations produites à un instant précis, il s'agit d'un taux calculé avec plus de précision et dont l'intérêt majeur est de décrire l'évolution d'un comportement conditionnellement à son histoire.

Figure 1

Représentations graphiques de la fonction de Survie, de hasard et de densité obtenus par la 'PROC LIFETEST' sur l'ensemble des entreprises de la strate 39







Étant donné que nous n'avons pas tenu compte des problèmes de manque d'information existant dans le fichier, les taux de survie et les taux de cessations obtenus sont très proches de ceux obtenus par la méthode des fréquences empiriques, reste que les taux de cessation sont de plus en plus différents dans le temps. La comparaison des taux de survie et des taux de cessation obtenus par les deux méthodes sont illustrés dans les tableaux et les graphiques suivants :

Dans un premier tableau nous allons illustrer la comparaison des taux de survie (*Tableau 3.a*) et dans un deuxième tableau nous allons présenter la comparaison des taux de cessation (*Tableau 3.b*)

Tableau 3.a
Tableau de comparaison des taux de survie selon la méthode utilisée (fréquence et Kplan-Meier)

Durée	Taux de Survie (Fréquence)	Taux de Survie (Kplan-Meier)	Intervalle de confiance de S(t) (Kaplan-Meier)
1 an	94,77	94,77	[94,31; 95,23]
2 ans	84,88	84,87	[84,12; 85,62]
3 ans	74,49	74,47	[73,57; 75,38]
4 ans	65,07	65,06	[64,06; 66,05]
5 ans	57,52	57,51	[56,48; 58,54]
6 ans	51,5	51,48	[50,44; 52,52]
7 ans	47,19	47,16	[46,12; 48,20]
8 ans	45,56	45,54	[44,50; 46,57]

À partir des résultats illustrés dans se tableau on remarque que le simple fait de tenir compte du phénomène de la censure à droite, rend les taux de survie obtenus par l'estimation Kaplan-Meier légèrement inférieur à ceux obtenus par la méthode de fréquence; ces derniers sont surestimés. Ce qui conduit à une diminution de l'ampleur du biais obtenu par la méthode descriptive. Exemples: le taux de survie obtenu à 5 ans par la méthode de fréquence est de 57,52 % alors que celui obtenu par la méthode Kaplan-Meier est de 57,51 % d'où une diminution de 0,01 %. À 7 ans le taux de survie obtenu par la méthode de fréquence est de 47,19 % alors que celui obtenu par la méthode Kaplan-Meier est de 47,16% d'où une diminution de 0,03 %.

Dans cette étape malgré la légère diminution des taux de survie, nous avons remarqué que les taux obtenus par les deux méthodes appartiennent à l'intervale de confiance calculés par l'approche non paramétrique

Tableau 3.b

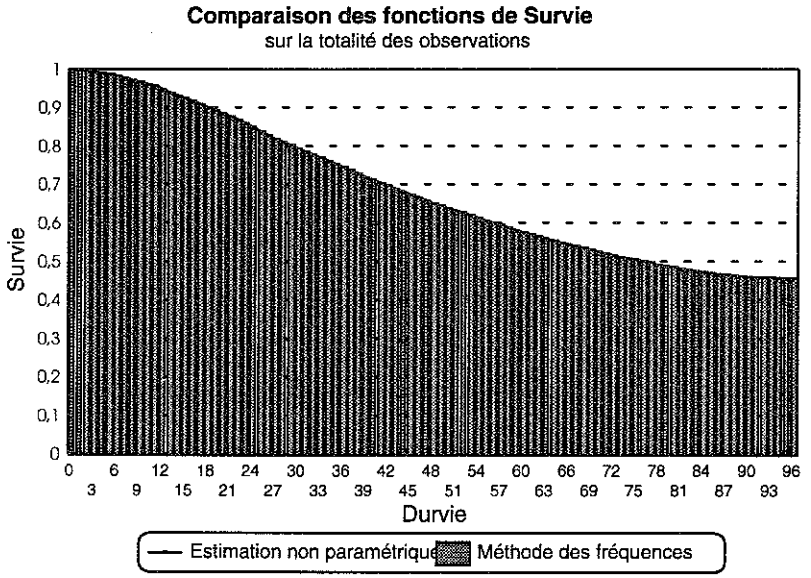
Tableau de comparaison des taux de cessations selon la méthode utilisée (fréquence et Kplan-Meier)

Durée	Taux de Cessation cumulé; F(t) (Fréquence)	Taux de Cessation instantané cumulé; H(t) (Kplan-Meier)
1 an	5,23	5,37
2 ans	15,12	16,40
3 ans	25,51	29,47
4 ans	34,93	42,98
5 ans	42,48	55,32
6 ans	48,5	66,40
7 ans	52,81	75,16
8 ans	54,44	78,65

Les taux de cessations obtenus par la méthode de fréquence son généralement sous estimés, étant donné la relation suivante (F(t)=1-S(t)), alors que ceux obtenus par l'estimation Kaplan-Meier sont plus précis. De ce fait les taux de cessation instantanés sont supérieurs à ceux obtenus par la méthode de fréquence. Exemple: à 5 ans le taux de cessation cumulé est de 42,48 %, alors que le taux de cessation instantané est de 55,32 %.

Figure 2

Représentation graphique des taux de survie selon les deux méthodes



Nous pouvons conclure à ce stade que l'application de la procédure d'analyse des durées de vie (proc LIFETEST) n'est pas compliquée, elle est de la même simplicité que la procédure de fréquence, mais seulement elle présente plus d'avantages. D'une part elle tient compte des données censurées ce qui nous permet d'avoir des résultats d'une meilleure précision et d'autre part elle nous donne plus d'indicateurs qui nous permettent de mieux analyser la durée de vie des entreprises ainsi que les l'intervalles de confiance de chacun des estimateurs. Ces indicateurs sont accompagnés de leurs représentations graphiques.

Nous ne pouvons pas nous contenter des résultats obtenus par cette approche, sachant d'avance qu'ils sont encore biaisés. Mais ils peuvent être considérés comme une étape préalable pour la modélisation de la durée de vie des entreprises. Nous facilitons ainsi le choix d'un modèle paramétrique raisonnable.

3. La modélisation

La méthode la plus simple pour estimer un modèle de durée est de procéder directement à l'estimation des paramètres de la loi de la variable aléatoire, par exemple estimer les paramètres de sa fonction de hasard. Cette modélisation va être effectuée sur des données individuelles, tous en considérant que tous les individus de la même strate se comportent de la même manière en terme de durée (homogénéité des individus de la strate).

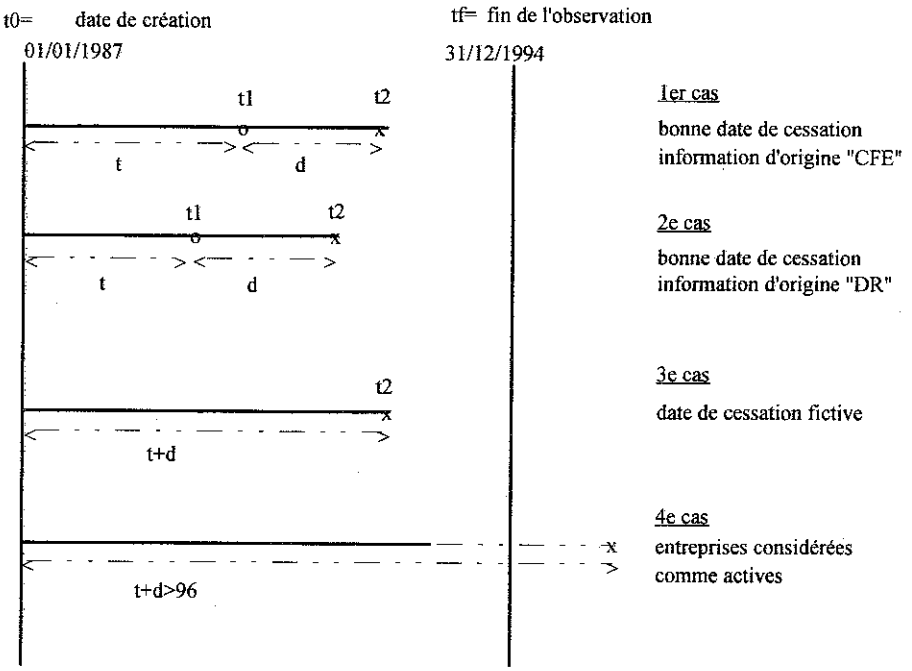
Une fonction de hasard peut s'interpréter comme une "probabilité instantanée de cessation conditionnellement au fait d'avoir survécu jusqu'à l'instant immédiatement antérieur". L'intérêt majeur des fonctions de hasard est de fournir un outil naturel pour une approche dynamique de la modélisation, à savoir décrire en probabilité l'évolution d'un système conditionnellement à son histoire. Elle est particulièrement utile dans l'étude des modèles de durées. Il est souvent très important au niveau de l'interprétation de connaître son évolution en fonction du temps. La fonction de hasard peut avoir diverses formes. Chacune pouvant être un indicateur de la famille des lois à la quelle appartient notre fonction.

Les méthodes habituelles de régression ne peuvent être appliquées pour l'estimation des modèles de durées, vu que ce sont généralement des modèles non linéaire (sauf dans certains cas particulier). La méthode la plus utilisée est celle du maximum de vraisemblance qui consiste à prendre comme estimation du paramètre (θ) de la loi suivit par les différentes observations, une valeur qui maximise la vraisemblance ; $\underset{\theta}{\text{Max}}\ l(T, D / \theta)$.

3.1. Les différents type d'observations

Pour expliciter les différents types de problèmes nous allons nous référer au graphique suivant illustrant les différents cas de figure dans le fichier;

Représentation des différents cas de figure



O ===> "date d'événement" de la cessation

X ===> "date de traitement" de la cessation

Les variables qui ont été créées pour les besoins de l'étude sont :

- la durée de vie des entreprises qui correspond à l'écart entre la date de création et la date de cessation d'activité (DURVIE), elle est calculée en mois ;

- le délai d'enregistrement de la cessation qui correspond à l'écart entre la date d'enregistrement de la cessation et la date réelle de la fin de l'activité (DELAI), de même cette variable est calculée en mois.

L'information disponible dans nos données relative à ces variables prend 4 formes possibles :

1er cas: l'entreprise a cessé son activité à la date t_1 et l'a fait enregistrer à la date t_2 par les CFE.

On connaît alors, par les CFE:

DURVIE = t_1-t_0

DELAI = t_2-t_1

Dans ce cas les données sont complètes.

2e cas: l'entreprise a cessé son activité à la date t_1, et ce sont les DR qui l'ont enregistrée à la date t_2.

On connaît alors, par les DR:

DURVIE = t_1-t_0

DELAI = t_2-t_1

De même, les données sont complètes suite à une relance.

3e cas: l'entreprise a cessé son activité à une date inconnue, et les DR ont fait l'enregistrement à une date t_2.

On connaît alors, par les DR:

DURVIE+DELAI=t_2-t_0

La durée observée n'est qu'un majorant de la durée réelle. On dit que la durée est observée avec une erreur de mesure.

4e cas: l'entreprise est considérée comme active étant donné qu'on n'a aucune information sur sa cessation à la fin de la période d'observation (31/12/94).

DURVIE+DELAI>96

Dans ce cas ni la durée de vie ni le délai d'enregistrement ne sont observés; cas d'information incomplète. La durée observée est inférieure a la durée réelle: on dit que les données sont censurées à droite

3.2. les contributions à la vraisemblance

À partir des données disponibles nous sommes amenés à modéliser un processus ponctuel tri-varié; la durée de vie de l'entreprise (DURVIE), le délai d'enregistrement de la cessation par l'entrepreneur (DELAI=D_1) et le délai d'enregistrement de la fin de l'activité et de l'enquête DR (DELAI=D_2). Ce délai dépend d'une part du degré de négligence des entrepreneurs et d'autre part de la date du déroulement de l'enquête. Plusieurs hypothèses sont alors émises selon les différents cas de figures :

Hypothèse 1 :

On considère que la variable durée de vie de l'entreprise T suit la même loi quel que soit le type d'observation.

Hypothèse 2 :

On suppose qu'il y a indépendance entre la variable durée T et respectivement les variables D_1 et D_2 tel que:

$$P(T = t_i, D_1 = d_i^{cfe}) = P(T = t_i) * P(D_1 = d_i^{cfe})$$

$$P(T = t_i, D_2 = d_i^{dr}) = P(T = t_i) * P(D_2 = d_i^{dr})$$

Hypothèse 3 :

Dans le cas des dates fictives, nous supposons que la loi du délai suit la même distribution que celles des D_2 dans le cas où on a la bonne "date d'événement" et l'origine de l'information de la cessation d'activité sont les DR.

Hypothèse 4 :

Dans le cas où l'entreprise est considérée comme active, nous supposons dans un premier temps que la loi du délai suit la même distribution que celle des D_1, étant donné que leur durée de vie est supérieure à 8 ans. On considère que la déclaration de la cessation sera spontanée (le degré de négligence de l'entrepreneur est presque nul).

Le modèle que nous allons traiter est présenté dans un premier temps en plusieurs étapes, selon les différents types d'observations:

1er cas:

L'entreprise a cessé son activité avant 31/12/94, l'information est d'origine CFE. L'information est donc complète; nous sommes en présence de données non censurées. La contribution à la vraisemblance est :

$$P_{T,\ D_1}((T = t_i, D_1 = d_i^{cfe}) / T + D_1 \leq 96 \text{ et } T \geq 1)$$

2e cas:

l'entreprise a cessé son activité avant 31/12/94, l'information est d'origine DR. Nous sommes aussi en présence de données non censurées. La contribution à la vraisemblance est:

$$P_{T,\ D_2}((T = t_i, D_2 = d_i^{dr}) / T + D_2 \leq 96 \text{ et } T \geq 1)$$

3e cas:

l'entreprise a cessé son activité avant le 31/12/94, l'information est d'origine DR et la date de cessation est une date fictive. La durée est une variable déterminée avec une erreur de mesure. La contribution à la vraisemblance est:

$$P_{T+D_2}((T = t_i + d_i^{dr}) / T + D_2 \leq 96 \text{ et } T \geq 1)$$

4e cas:

l'entreprise est considérée comme active alors qu'elle peut avoir cessé son activité avant la fin de l'observation, comme elle peut être encore active au 1/01/95, mais sa cessation sera enregistrée que plus tard. Dans ce cas la durée et le délai sont indéterminés. La contribution à la vraisemblance est:

$$P_{T+D}((T + D > 96) / T > 1 \text{ mois})$$

Les calculs détaillés de chacune des contributions sont présentés en *Annexe 4.*

4. L'approche paramétrique

Les données disponibles sur les durées de vie sont généralement incomplètes. Afin de résoudre les deux types d'erreurs: les dates fictives et le délai d'enregistrement des cessations, nous avons opté pour une modélisation en adoptant l'approche paramétrique. Les résultats de la procédure « LIFETEST » vont nous permettre de faire un choix raisonnable du modèle à adapter a nos variables.

La procédure utilisée pour l'estimation des modèles dans l'approche paramétrique est la procédure d'analyse des durées de vie; la « PROC NLIN ». Il s'agit d'une procédure utilisant principalement les modèles de durée. La méthode d'estimation utilisée dans cette procédure est celle du maximum du vraisemblance.

L'inconvénient de l'utilisation de cette procédure est double:

1/ Elle ne peut être directement appliquée dans certains cas; car c'est une procédure standard ;

2/ les modifications à apporter nécessitent une programmation, difficile à réaliser sous SAS.

Étant donné les difficultés rencontrées, nous sommes en train d'améliorer le programme qui nous permet de tenir compte des différentes modifications à apporter à la procédure NLIN. Les résultats de cette approche ne sont pas encore disponibles.

5. Conclusion

Les résultats obtenus par l'approche non paramétrique nous ont permis de confirmer l'importance du fait de tenir compte des entreprises qui sont encore active à la fin de l'observation; phénomène de censure à droite. Ceci nous a conduit à avoir des estimations des taux de survie légèrement inférieur à ceux obtenus par la méthode des fréquences. Étant donné que nous n'avons pas tenu compte des différents problèmes de manque d'information liés à l'enregistrement des cessations, nous ne pouvons pas nous contenter de ces estimations. Seulement ces derniers nous facilitent le choix d'un modèle dans le cadre d'une approche paramétrique. Ce qui fait l'objet de notre travail en cours.

Annexe 1

D'un point de vue théorique, afin de caractériser la variable aléatoire durée de vie des entreprises nous disposons de divers indicateurs:

1/ les indicateurs qui sont généralement utilisés :

- la fonction de densité:

f(t) peut s'interpréter comme la probabilité de cessation.

$$f(t) = \lim_{dt \to 0} \frac{1}{dt} P(t < T \leq t + dt)$$

- la fonction de répartition:

F(t) peut s'interpréter comme le cumul des probabilités de cessation

$$F(t) = P(T \leq t) = \int_0^t f(x)dx$$

F est continue, croissante, $F(0)=0$, $F(+\infty)=1$

2/ les indicateurs qui sont utilisés en plus en modèles de durée :

- la fonction de survie:

S(t) peut s'interpréter comme la probabilité que la durée de vie soit plus grande que t ;

$$S(t)=P(T>t)=1-F(t) = \int_t^{\infty} f(x)dx$$

S est continue, décroissante, $S(0) = 1$, $S(+\infty) = 0$;

- la fonction de hasard:

h(t) peut s'interpréter comme la probabilité instantané de cessation conditionnellement au fait d'avoir survécu jusqu'à l'instant immédiatement antérieur ; c'est le taux de sortie :

$$h(t) = \lim_{dt \to 0} \frac{1}{dt} P(t < T \leq t + dt / T > t)$$

$$h(t) = \lim_{dt \to 0} \frac{1}{dt} \frac{P(t < T \leq t + dt)}{P(T > t)}$$

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}.$$

- la fonction de hasard cumulé:

H(t) peut être interprété comme le taux de cessation cumulé. Cet indicateur est intéressant dans le cas où nous voulons calculer le taux de cessation pendant toute la période d'observation et non seulement à un instant précis (ce qui est le cas de la fonction de hasard; h(t)).

$$H(t) = \int_0^t h(x)\, dx$$

- la durée moyenne restante:

r(t) peut s'interpréter comme l'espérance de la durée qui reste sachant que l'on a déjà atteint t:

$$r(t) = E(T - t / T > t)$$

Les relations existant entre ces trois fonctions sont:

la fonction de survie est définie par:

$$S(t) = \exp\left[-\int_0^t h(x)dx \right] \quad t \in R^+$$

la fonction de répartition est définie par:

$$F(t) = 1 - S(t) = 1 - \exp\left\{ -\int_0^t h(x)\, dx \right\}$$

la fonction de hasard cumulé est définie par:

$$H(t) = \int_0^t h(x)\, dx = -\mathrm{Log}(S(t))$$

Annexe 2

Les variables explicatives sélectionnées pour les besoins de l'étude de la durée de vie des entreprises sont les suivantes :

- Origine de la création : variable qualitative en deux modalités :
  - création pure
  - reprise

- Date de création ;

- Date de cessation économique qui est la date d'événement

- Date de traitement de la cessation

- Tranche d'effectif salarié à la date de création : elle est représentée en 7 modalités:

0	effectif=0
1	1<=effectif<=2
2	3<=effectif<=5
3	6<=effectif<=9
4	10<=effectif<=19
5	20<=effectif<=49
6	effectif>=50

- Activité agrégée en 8 modalités suivant la variable ACTEN10
  - 0 : Services aux ménages
  - 2 : I.A.A (Industries agro-alimentaires)
  - 3 : Industrie hors I.A.A
  - 4 : Construction
  - 5 : Commerce
  - 6 : Transport
  - 8 : Hôtels cafés restaurants
  - 9 : Services aux entreprises

- Catégorie juridique : elle a été agrégée en 9 postes:
  - 01 : Artisan-commerçant
  - 02 : Commerçant
  - 03 : Artisan
  - 04 : Profession libérale
  - 20 : Société de fait
  - 52 : Société en nom collectif et société en commandite
  - 54 : SARL et EURL
  - 55 : SA
  - 60: Société civile et autre

Annexe 3

La procédure LIFETEST est une procédure SAS utilisant principalement des modèles de durée pour estimer la distribution de survie et calcul des tests de rang signés Wilcoxon qui sont des tests de comparaison d'échantillons; le test de rang se basant sur l'idée de mélanger deux ou plusieurs échantillons et qu'on ordonne le tout par ordre croissant, l'objectif de ce test est de voir si le mélange des sous-populations est homogène.

L'approche adoptée pour l'estimation de la distribution de survie est une approche non paramétrique appelé estimateur de Kaplan-Meier se basant sur les principes suivants:

soit $S(t) = P(x \geq t) = 1 - P(x < t) = 1 - F(t)$

Dans le cas des données non censurées la fonction de répartition empirique $\hat{F}_n(t)$ est la proportion des n variables t_1, t_2, t_3, ..., t_n qui sont inférieures à t, dont les réalisations sont des fonctions en escalier de saut égaux à 1/n:

$$\hat{F}_n(t) = \frac{\text{nombre d' observation} < t}{n} \quad \text{avec } t \geq 0$$

En supposant que les observations sont ordonnées par ordre croissant $t_1 < t_2 < t_3 < \ldots < t_n$

$$\begin{cases} \hat{F}_n(t) = 0 & \text{si } t < t_1; \\ \hat{F}_n(t) = \dfrac{i}{n} & \text{si } t_i \leq t < t_{i+1}; \\ \hat{F}_n(t) = 1 & \text{si } t > t_n; \end{cases}$$

À partir de la relation existant entre la fonction de survie et la fonction de répartition on peut en déduire la fonction de survie empirique :

$$\hat{S}_n(t) = \frac{\text{nombre d' observation} \geq t}{n}$$

$$\hat{S}_n(t) = 1 - \frac{\text{nombre d' observation} < t}{n}$$

$$\hat{S}_n(t) = 1 - \frac{i}{n} \quad \text{si } t_i \leq t < t_{i+1}$$

$$\hat{S}_n(t) = \frac{n-i}{n}$$

$$= \frac{(n-1)}{n} * \frac{(n-2)}{(n-1)} * \frac{(n-3)}{(n-2)} * \ldots * \frac{(n-i)}{(n-i+1)}$$

$$= (1-\frac{1}{n}) * (1-\frac{1}{n-1}) * (1-\frac{1}{n-2}) * \ldots * (1-\frac{1}{n-i+1})$$

$$= \prod_{i=1}^{n} 1_{\{t_i < t\}} * (1-\frac{1}{n-i+1})$$

$$= \prod_{i:t_i < t}^{n} (1-\frac{1}{n-i+1})$$

La fonction de survie empirique apparaît comme le produit de probabilités conditionnelle:

$$\hat{S}_n(t) = \hat{P}_n(x \geq t)$$

$$= \prod_{i:t_i < t} \hat{P}_n(x \geq t_i / x > t_{i-1})$$

avec $t_0 = 0$

Après l'instant t_{i-1}, on a i-1 cessations enregistrées, le nombre d'entreprises encore actives est de n-i+1 ce nombre est considéré comme l'ensemble des sujet à risque (c.a.d ni cessées ni censurées) a l'instant t_i. La probabilité pour qu'une entreprise appartenant à l'ensemble à risque ne cesse pas son activité à l'instant t_i est $1 - \frac{1}{n-i+1}$

En 1958 Kaplan-Meier a repris la même forme pour calculer l'estimation de la fonction de survie en tenant compte de la censure à droite:

Pour un échantillon de n observations on considère qu'on a k observations complètes $(t_1 < t_2 < t_3 < \ldots < t_k)$ et n-k observations censurées à droite.

Soit R_i le nombre de sujet à risque et M_i le nombre d'entreprises qui ont cessé leurs activités à l'instant t_i, l'estimateur de Kaplan-Meier est définit par le rapport entre le nombre d'entreprises cessé et le nombre de sujet à risque d'où :

$$\hat{S}_{K-M}(t) = \prod_{i:t_i<t} (1 - \frac{M_i}{R_i})$$

(Économétrie des modèles de durée; A. Moreau)

(The Lifetest procedure; Document SAS)

(Un aperçu des modèles de survie; C. Cases and S. Lollivier)

Annexe 4

1er cas:

l'entreprise a cessé son activité avant 31/12/94, l'information est d'origine CFE. L'information est donc complète; nous sommes en présence de données non censurées. La contribution à la vraisemblance est:

$$P_{T,\,D_1}((T = t_i, D_1 = d_i^{cfe}) / T + D_1 \leq 96 \text{ et } T \geq 1)$$

or $P_{T+D_1}(T + D_1 \leq 96 \text{ et } T \geq 1) = P_{T+D_1}(T + D_1 \leq 96 / T \geq 1) * P_T(T \geq 1)$

$$= \sum_{k=0}^{K} P_T(T \leq 96 - k / T \geq 1) * P(D_1 = k) * P_T(T \geq 1)$$

$$= \frac{\sum_{k=0}^{K} P_T(1 \leq T \leq 96 - k) * P_{D_1}(D_1 = k)}{P_T(T \geq 1)} * P_T(T \geq 1)$$

$$= \sum_{k=0}^{K} P_T(1 \leq T \leq 96 - k) * P_{D_1}(D_1 = k)$$

d'où

$$P_{T,\,D_1}((T = t_i, D_1 = d_i^{cfe}) / T + D_1 \leq 96 \text{ et } T \geq 1)$$

$$= \frac{P_T(T = t_i) * P_{D_1}(D_1 = d_i^{cfe})}{\sum_{k=0}^{K} P_T(1 \leq T \leq 96 - k) * P_{D_1}(D_1 = k)} * 1_{\{T + D_1 \leq 96 \text{ et } T \geq 1\}}$$

2e cas :

l'entreprise a cessé son activité avant 31/12/94, l'information est d'origine DR. Nous sommes aussi en présence de données non censurées. La contribution à la vraisemblance est :

$$P_{T,\,D_2}((T = t_i, D_2 = d_i^{dr}) / T + D_2 \leq 96 \text{ et } T \geq 1)$$

$$= \frac{P_T(T = t_i) * P_{D_2}(D_2 = d_i^{dr})}{\sum_{k=0}^{K} P_T(1 \leq T \leq 96 - k) * P_{D_2}(D_2 = k)} * 1_{\{T + D_2 \leq 96 \text{ et } T \geq 1\}}$$

3e cas :

l'entreprise a cessé son activité avant le 31/12/94, l'information est d'origine DR et la date de cessation est une date fictive. La durée est une variable déterminée avec une erreur de mesure. La contribution à la vraisemblance est :

$$P_{T+D_2}((T = t_i + d_i^{dr}) / T + D_2 \leq 96 \text{ et } T \geq 1)$$

$$= \frac{\sum_{k=0}^{K} P_T(t_i = T - k) * P_{D_2}(d_i^{dr} = k)}{\sum_{k=0}^{K} P_T(1 \leq T \leq 96 - k) * P_{D_2}(D_2 = k)} * 1_{\{T + D_2 \leq 96 \text{ et } T \geq 1\}}$$

4e cas :

l'entreprise est considérée comme active alors qu'elle peut avoir cessé son activité avant la fin de l'observation, comme elle peut être encore active au 1/01/95, mais sa cessation sera enregistrée que plus tard. Dans ce cas la durée et le délai sont indéterminés. La contribution à la vraisemblance est :

$$P_{T+D}((T + D > 96) / T > 1 \text{ mois}) = \sum_{k=0}^{K} P_T(T + k > 96) * P_D(D = k)$$

$$= \sum_{k=0}^{K} P_T(T > 96 - k) * P_D(D = k)$$

BIBLIOGRAPHIE

A. MOREAU (1989), "Econométrie des modèles de durée", *note du département de la recherche* N.123/G 305 (Insee).

D. FRANCOZ, J. BONNEAU (Avr 1994), "Le devenir des entreprises créées en 1987", *Insee Première*, n°312.

J.-J. DROESBEKE, B. FICHET, P. TASSI, *Analyse statistique des durées de vie-Modélisation des données censurées*, Economica, 1989.

J.P. FLORENS, D. FOUGERE, M. MOUCHART, "Modélisation des mobilités sur le marché du travail", *Discussion Paper* 9310, Institut de Statistique-Université Catholique de Louvain-la-neuve - Belgium, juin 1993.